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

by Adrienne Kitts and Stephen Sherry

Summary

Sequence variations exist at defined positions within genomes and are responsible for individual phenotypic characteristics, including a person's propensity toward complex disorders such as heart disease and cancer. As tools for understanding human variation and molecular genetics, sequence variations can be used for gene mapping, definition of population structure, and performance of functional studies.

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.

The 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.

Introduction

The dbSNP has been designed to support submissions and research into a broad range of biological problems. These include physical mapping, functional analysis and pharmacogenomics, association studies, and evolutionary studies. Because dbSNP was developed to complement GenBank, it may contain nucleotide sequences (Figure 1) from any organism; currently, the majority of the data is for human and mouse.

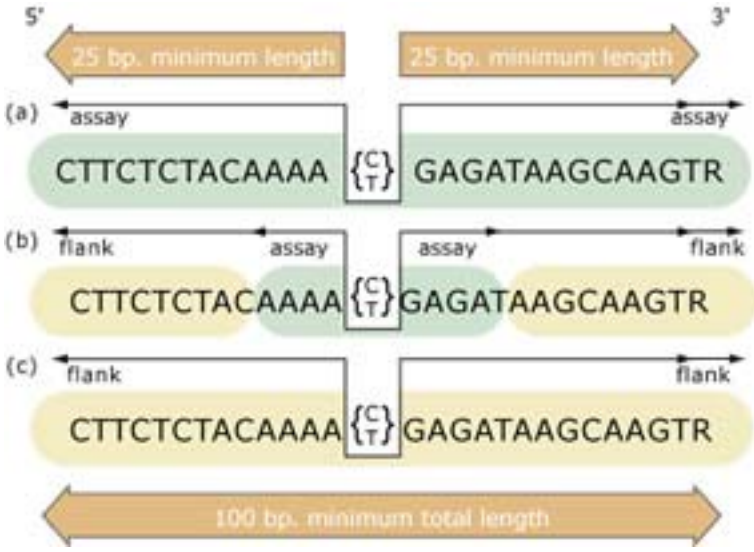


Figure 1: The structure of the flanking sequence.
The structure of the flanking sequence in dbSNP is a composite of bases either assayed for variation or included from published sequence. We make the distinction to distinguish regions of sequence that have been experimentally surveyed for variation (*assay*) from those regions that have not been surveyed (*flank*). The minimum sequence length for a variation definition (SNPassay) is 25 bp for both the 5′ and 3′ flanks and 100 bp overall to ensure an adequate sequence for accurate mapping of the variation on reference genome sequence. (*a*) Flanking sequence completely surveyed for variation. Both 5′ and 3′ flanking sequence can be defined with 5′_assay and 3′_assay fields, respectively, when all flanking sequence was examined for variation. This can occur in both experimental contexts (e.g., denaturing high-pressure liquid chromatography or DNA sequencing) and computational contexts (e.g., analysis of BAC overlap sequence). (*b*) Partial survey of flanking sequence can occur when detection methods examine only a region of sequence surrounding the variation that is shorter than either the 25 bp per flank rule or the 100 bp overall length rule. In these experimental designs (e.g., chip hybridization, enzymatic cleavage), we designate the experimental sequence 5′_assay or 3′_assay, and you can insert published sequence (usually from a gene reference sequence) as 5′_flank or 3′_flank to construct a sequence definition that will satisfy the length rules. (*c*) Unknown or no survey of flanking sequence can occur when variations are captured from published literature without an indication of survey conditions. In these cases, the entire flanking sequence is defined as 5′_flank and 3′_flank.

Physical Mapping

In the physical mapping of nucleotide sequences, variations are used as positional markers. When mapped to a unique location in a genome, variation markers work with the same logic as Sequence Tagged Sites (STSs) or framework microsatellite markers. As is the case for STSs, the position of a variation is defined by its unique flanking sequence, and hence, variations can serve as stable landmarks in the genome, even if the variation is fixed for one allele in a sample. When multiple alleles are observed in a sample pedigree, pedigree members can be tested for variation genotypes as in traditional physical mapping studies.

Functional Analysis

Variations that occur in functional regions of genes or in conserved non-coding regions can cause significant changes in the complement of transcribed sequences, leading to changes in protein expression that can affect aspects of the phenotype such as metabolism or cell signaling. We note possible functional implications of DNA sequence variations in dbSNP in terms of how the variation alters mRNA transcripts.

Association Studies

The associations between variations and complex genetic traits are more ambiguous than simple, single-gene mutations that lead to a phenotypic change. When multiple genes are involved in a trait, then the identification of the genetic causes of the trait requires the identification of the chromosomal segment combinations, or haplotypes, that carry the putative gene variants.

Evolutionary Studies

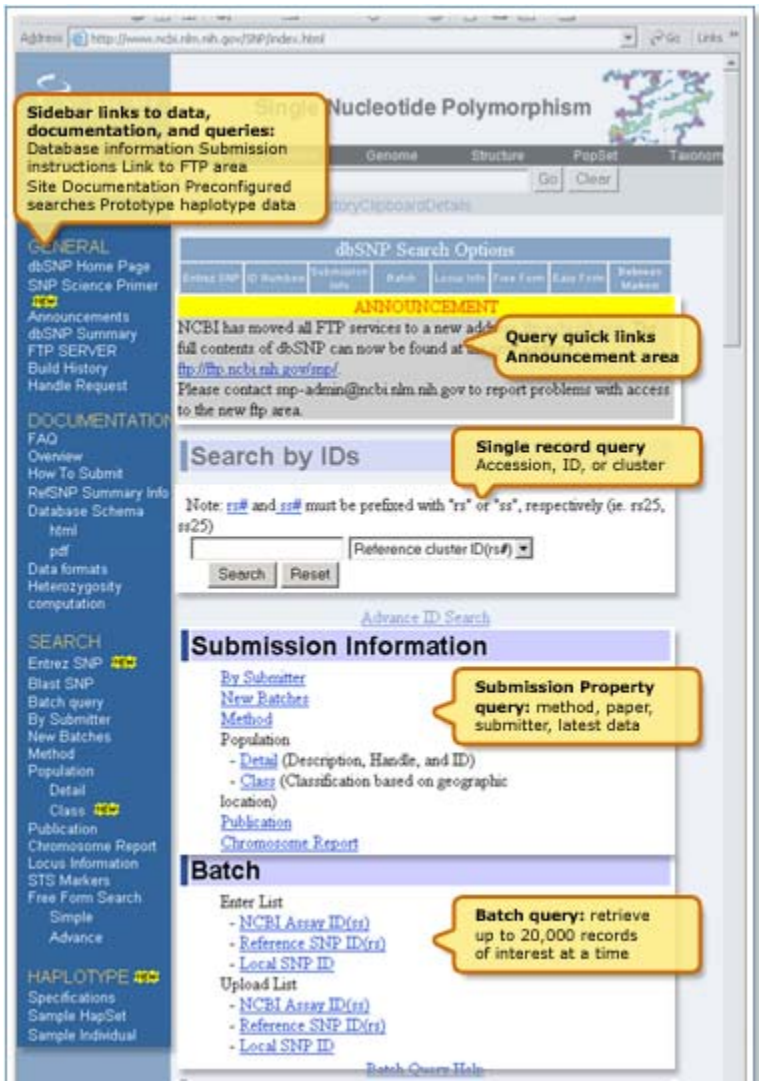
The variations in dbSNP currently represent an uneven but large sampling of genome diversity. The human data in dbSNP include submissions from the SNP Consortium, variations mined from genome sequence as part of the human genome project, and individual lab contributions of variations in specific genes, mRNAs, ESTs, or genomic regions.

Null Results Are Important

Systematic surveys of sequence variation will undoubtedly reveal sequences that are invariant in the sample. These observations can be submitted to dbSNP as NoVariation records that record the sequence, the population, and the sample size that were used in the survey.

Searching dbSNP

The SNP database can be queried from the dbSNP homepage http://www.ncbi.nlm.nih.gov/SNP/ (Figure 2) by using the Entrez SNP **Search** box at the top of the page or by using the links to eight basic dbSNP search options (located just below the Entrez SNP **Search** box). Each search option is described below.



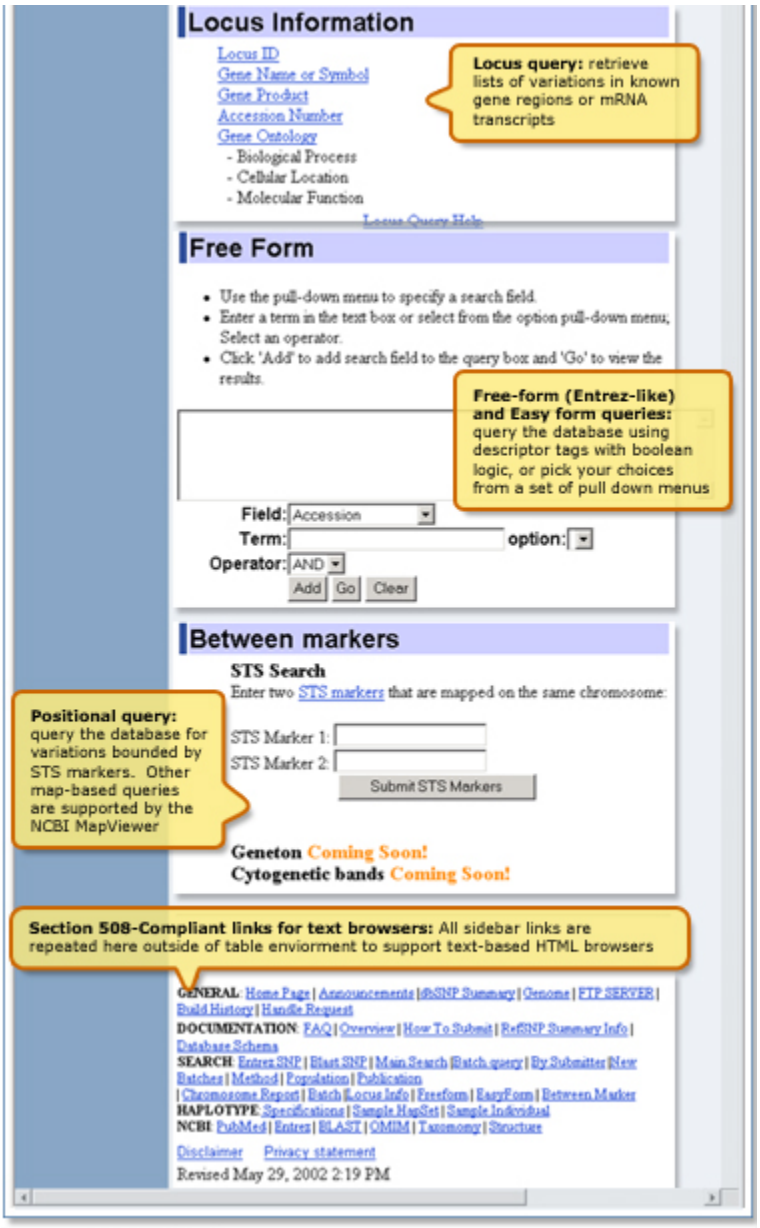


Figure 2: The dbSNP homepage.
We organized the dbSNP homepage with links to documentation, FTP, and sub-query pages on the *left sidebar* and a selection of query modules on the *right sidebar*.

Entrez SNP

The dbSNP is now a part of the Entrez integrated information retrieval system (Chapter 14) and may be searched using either qualifiers (aliases) or a combination of 28 different search fields. A complete list of the qualifiers and search fields can be found on the Entrez SNP site.

Single Record (Search by IDs) Query in dbSNP

Use this query module to select SNPs based on dbSNP record identifiers. These include reference SNP (refSNP) cluster ID numbers (rs#), submitted SNP Accession numbers (ss#), and local (or submitter) IDs for the same variations.

SNP Submission Information Queries

Use this module to construct a query that will select SNPs based on submission records by laboratory (submitter), new data (called "new batches", this query limitation is more recent than a user-specified date), methods used to assay for variation (Table 1), populations of interest (Table 2), or publication information.

Table 1. Method classes organize submissions by a general methodological or experimental approach to assaying for variation in the DNA sequence.

Method class	Class code in Sybase, ASN.1, and XML
Denaturing high pressure liquid chromatography (DHPLC)	1
DNA hybridization	2
Computational analysis	3
Single-stranded conformational polymorphism (SSCP)	5
Other	6
Unknown	7
Restriction fragment length polymorphism (RFLP)	8
Direct DNA sequencing	9

Table 2. Population classes organize population samples by geographic region.

Population class	Description	Population class in Sybase, ASN.1, and XML
Central Asia	Samples from Russia and its satellite Republics and from nations bordering the Indian Ocean between East Asia and the Persian Gulf regions.	8
Central/South Africa	Samples from nations south of the Equator, Madagascar, and neighboring island nations.	4
Central/South America	Samples from Mainland Central and South America and island nations of the western Atlantic, Gulf of Mexico, and Eastern Pacific.	10
East Asia	Samples from eastern and south eastern Mainland Asia and from Northern Pacific island nations.	6
Europe	Samples from Europe north and west of Caucasus Mountains, Scandinavia, and Atlantic islands.	5
Multi-National	Samples that were designated to maximize measures of heterogeneity or sample human diversity in a global fashion. Examples include OEFNER|GLOBAL and the CEPH repository.	1
North America	All samples north of the Tropic of Cancer, including defined samples of United States Caucasians, African Americans, Hispanic Americans, and the NHGRI polymorphism discovery resource (NCBI|NIHPDR).	9
North/East Africa and Middle East	Samples collected from North Africa (including the Sahara desert), East Africa (south to the Equator), Levant, and the Persian Gulf.	2
Pacific	Samples from Australia, New Zealand, Central and Southern Pacific Islands, and Southeast Asian peninsular/ island nations.	7
Unknown	Samples with unknown geographic provinces that are not global in nature.	11
West Africa	Sub-Saharan nations bordering the Atlantic north of the Congo River and central/southern Atlantic island nations.	3

Population class	Description	Population class in Sybase, ASN.1, and XML

dbSNP Batch Query

Use sets of variation IDs collected from other queries to generate a variety of SNP reports.

Locus Information Query

The links in this section all point to the NCBILocusLink resource. This resource permits users to query for records by gene name, gene symbol, gene product, gene ontology, or Accession number. Records in LocusLink will have a pointer back to dbSNP if you have associated one or more variations with a gene feature.

Free Form (Entrez-like) and Easy Form Queries

Free Form is the most flexible query structure in dbSNP. Modeled on the NCBI Entrez retrieval system, queries can be conducted using multiple database field values to restrict a query to specific subsets of data. The Easy Form query is identical to the Free Form query, with the exception that the Easy Form query has a series of pull-down menus from which a value can be selected for the most popular query fields.

Between-Markers Positional Query

Use this query approach if you are interested in retrieving variations that have been mapped to a specific region of the genome bounded by two STS markers. Other map-based queries are available through the NCBI Map Viewer tool.

ADA Section 508-compliant Link

All links located on the left sidebar of the dbSNP homepage are also provided in text format at the bottom of the page to support browsing by text-based web browsers. Suggestions for improving database access by disabled persons should be sent to: snp-admin@ncbi.nlm.nih.gov.

Submitted Content

The SNP database has two major classes of content: the first class is submitted data, i.e., original observations of sequence variation (Figure 3); 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.

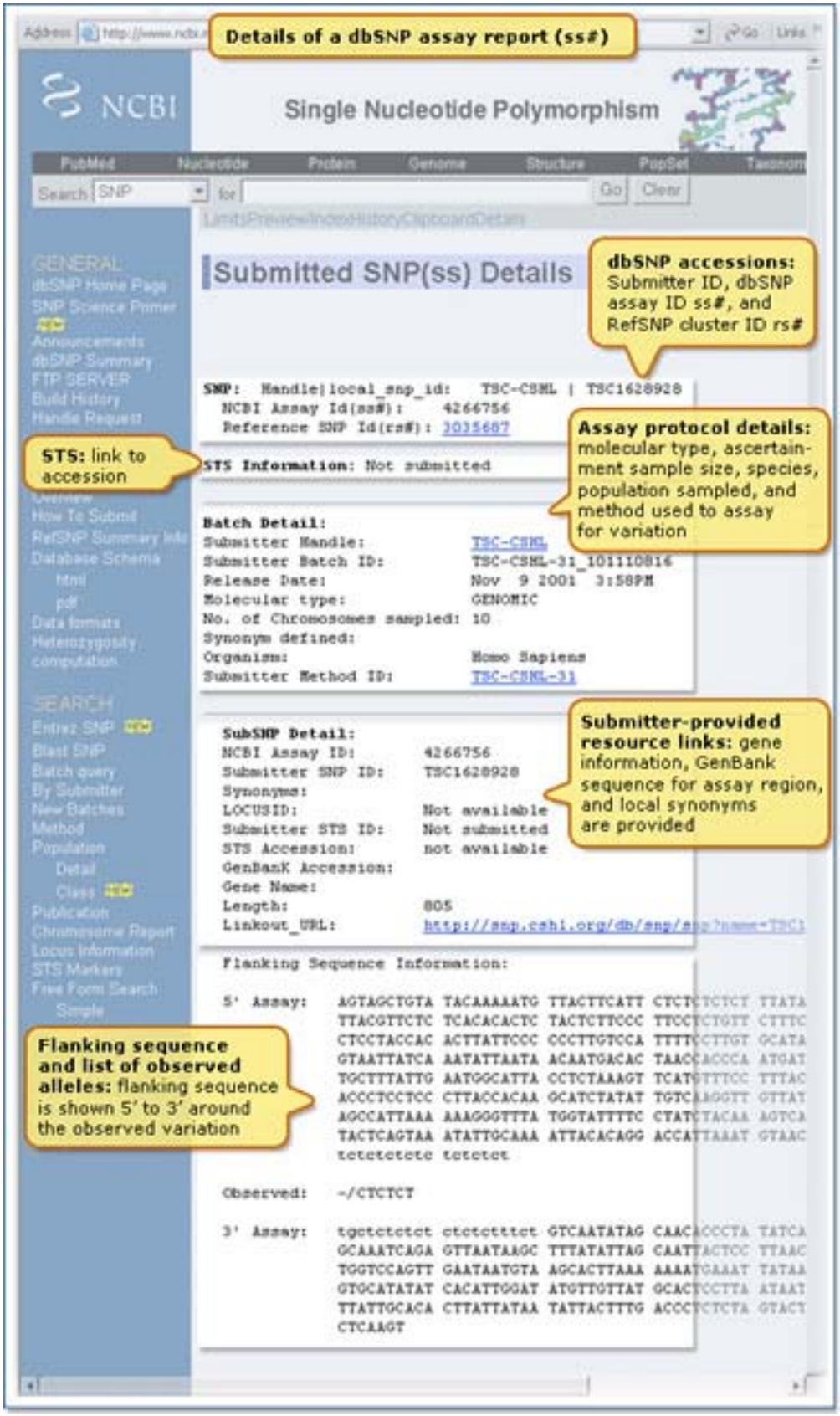


Figure 3: dbSNP submission details for an assay report.

The major sections of the report are described in the *right sidebar*.

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.

Alleles

Alleles define variation class (Table 3). In the dbSNP submission scheme, we define single-nucleotide variants as G, A, T, or C. We do not permit ambiguous IUPAC codes, such as N, in the allele definition of a variation. In cases where variants occur in close proximity to one another, we permit IUPAC codes such as N, and in the flanking sequence of a variation, we actually encourage them. See Table 3 for the rules that guide dbSNP post-submission processing in assigning allele classes to each variation.

Table 3. Allele definitions define the class of the variation in dbSNP.

dbSNP variation class[a, b]	Rules for assigning allele classes	Sample allele definition	Class code in Sybase, ASN.1, and XML[c]
Single Nucleotide Polymorphisms (SNPs)[a]	Strictly defined as single base substitutions involving A, T, C, or G.	A/T	1
Deletion/Insertion Polymorphisms (DIPs)[a]	Designated using the full sequence of the insertion as one allele, and either a fully defined string for the variant allele or a "-" character to specify the deleted allele. This class will be assigned to a variation if the variation alleles are of different lengths or if one of the alleles is deleted ("-").	T/-CCTA/G	2
Heterozygous sequence[a]	The term heterozygous is used to specify a region detected by certain methods that do not resolve the polymorphism into a specific sequence motif. In these cases, a unique flanking sequence must be provided to define a sequence context for the variation.	(heterozygous)	3
Microsatellite or short tandem repeat (STR)[a]	Alleles are designated by providing the repeat motif and the copy number for each allele. Expansion of the allele repeat motif designated in dbSNP into full-length sequence will be only an approximation of the true genomic	(CAC)8/9/10/11	4

dbSNP variation class[a, b]	Rules for assigning allele classes	Sample allele definition	Class code in Sybase, ASN.1, and XML[c]
	sequence because many microsatellite markers are not fully sequenced and are resolved as size variants only.		
Named variant[a]	Applies to insertion/deletion polymorphisms of longer sequence features, such as retroposon dimorphism for Alu or line elements. These variations frequently include a deletion "-" indicator for the absent allele.	(alu) / -	5
No-variation[a]	Reports may be submitted for segments of sequence that are assayed and determined to be invariant in the sample.	(NoVariation)	6
Mixed[b]		Mix of other classes	7
Multi-Nucleotide Polymorphism (MNP)[a]	Assigned to variations that are multi-base variations of a single, common length.	GGA/AGT	8

[a] Seven of the classes apply to both submissions of variations (submitted SNP assay, or ss#) and the non-redundant refSNP clusters (rs#'s) created in dbSNP.

[b] The "Mixed" class is assigned to refSNP clusters that group submissions from different variation classes.

[c] Class codes have a numeric representation in the database itself and in the export versions of the data (ASN.1 and XML).

Method

Each submitter defines the methods in their submission as either the techniques used to assay variation or the techniques used to estimate allele frequencies. We group methods by method class (Table 1) to facilitate queries using general experimental technique as a query field. The submitter provides all other details of the techniques in a free-text description of the method. Submitters can also use the **METHOD_EXCEPTION** field to describe changes to a general protocol for particular sets of data (batch-specific details). Submitters generally define methods only once in the database.

Population

Each submitter defines population samples either as the group used to initially identify variations or as the group used to identify population-specific measures of allele frequencies. These populations may be one and the same in some experimental designs. We assign populations a population class (Table 2) based on the geographic origin of the sample. These broad categories provide a general framework for organizing the approximately 700 (as of this writing) sample descriptions in dbSNP. Similar to method descriptions, population descriptions minimally require the submitter to provide a Population ID and a free-text description of the sample.

Sample Size

There are two sample-size fields in dbSNP. One field is called the **SNPASSAY SAMPLE SIZE**, and it reports the number of chromosomes in the sample used to initially ascertain or discover the variation. The other sample size field is called **SNPPOPUSE SAMPLE SIZE**, and it reports the number of chromosomes used as the denominator in computing estimates of allele frequencies. These two measures need not be the same.

Population-specific Allele Frequencies

Alleles typically exist at different frequencies in different populations; a very common allele in one population may be quite rare in another population. Also, allelic variants can emerge as private polymorphisms when particular populations have been reproductively isolated from neighboring groups, as is the case with religious isolates or island

populations. Frequency data are submitted to dbSNP as allele counts or binned frequency intervals, depending on the precision of the experimental method used to make the measurement. dbSNP contains records of allele frequencies for specific population samples defined by each submitter (Table 4).

Table 4. Validation status codes summarize the available validation data in assay reports and refSNP clusters.

Validation evidence	Description	Code in database for ss#	Code in FTP dumps for ss#	Code in database for rs#	Code in FTP dumps for rs#
Not validated	For ss#, no batch update or validation data, no frequency data (or frequency is 0 or 1). rs# status code is OR'd from the ss# codes.	0	Not present	0[a]	Not present
Multiple reporting	Status = 1 for an rs# with at least two ss# numbers; having at least one ss# is validated by a non-computational method. For a ss#, status = 1 if the method is non-computational.	1	1[b]	1,0[b]	1
With frequency	Frequency data is present with a value between 0 and 1.	2	2	2	2
Both frequency	For ss#, the method is non-computational and frequency data is present. If the ss# is a single cluster member, then the rs# code is set to 2.	3	3	3/2	3
Submitter validation	Submission of a batch update or validation section that reports a second validation method on the assay.	4	4	4	4

[a] If the rs# has a single ss# with code 1, then rs# is set to code 0.

[b] For a single member rs where the ss# validation status = 1, the rs# validation status is set to 0.

Population-specific Genotype Frequencies

Similar to alleles, genotypes have frequencies in populations that can be submitted to dbSNP.

Population-specific Heterozygosity Estimates

Some methods for detection of variation (e.g., denaturing high-pressure liquid chromatography or DHPLC) can recognize when DNA fragments contain a variation without resolving the precise nature of the sequence change. These data define an empirical measure of heterozygosity when submitted to dbSNP.

Individual Genotypes

dbSNP accepts individual genotypes for samples from publicly available repositories such as CEPH or Coriell. Genotypes reported in dbSNP contain links to population and method descriptions as shown in Figure 3. General genotype data provide the foundation for individual haplotype definitions and are useful for selecting positive and negative control reagents in new experiments.

Validation Information

dbSNP accepts individual assay records (ss# numbers) without validation evidence. When possible, however, we try to distinguish high-quality validated data from unconfirmed (usually computational) variation reports. Assays validated directly by the

submitter through the **VALIDATION** section show the type of evidence used to confirm the variation. Additionally, dbSNP will flag an assay as validated (Table 4) when we observe frequency or genotype data for the record.

Computed Content (The dbSNP Build Cycle)

We release the content of dbSNP to the public in periodic "builds" that we synchronize with the release of new genome assemblies (Chapter 13). During each build, we cluster the data submitted since the last build into existing refSNPs and form new refSNPs when necessary. The following 12 tasks define the sequence of steps in the dbSNP build cycle (Figure 4).

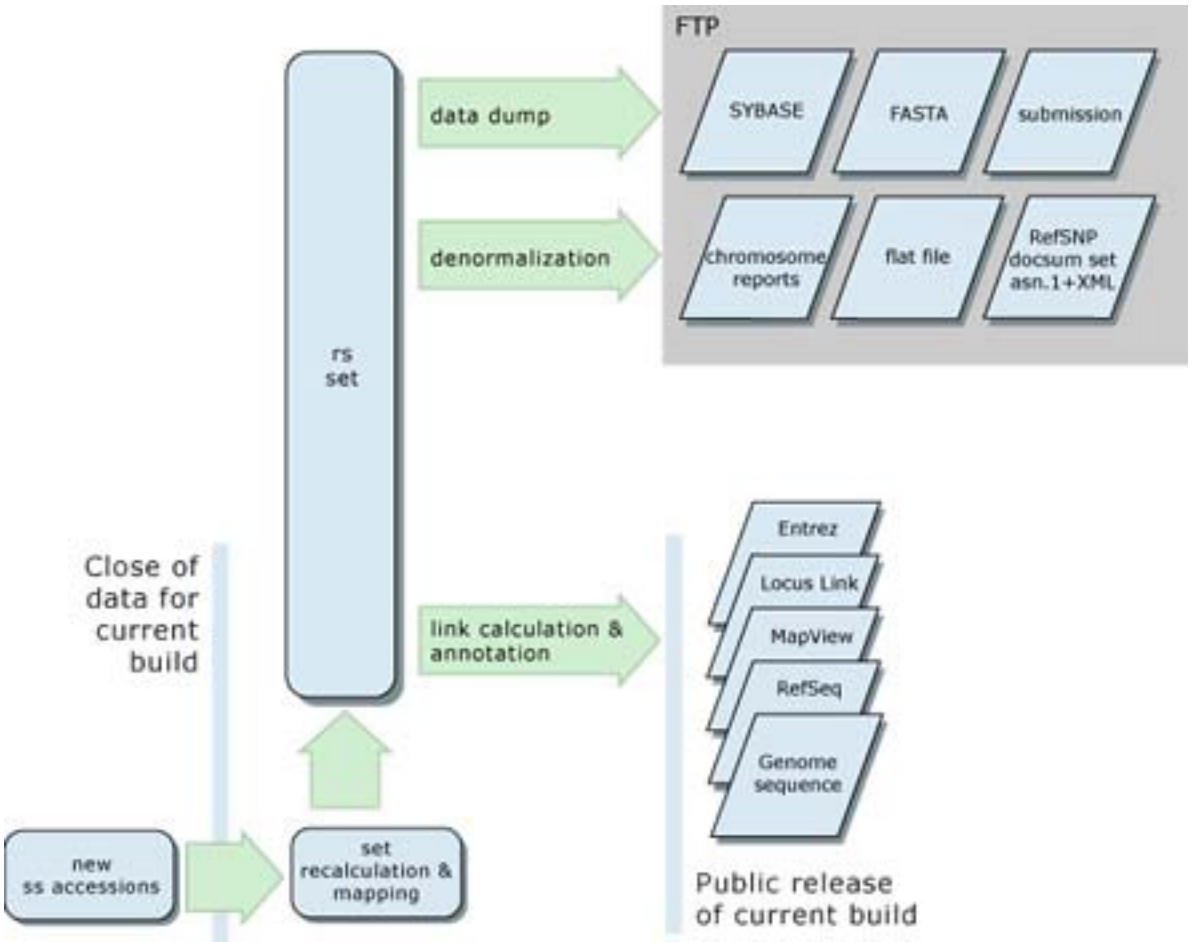


Figure 4: The dbSNP build cycle.
The dbSNP build cycle starts with close of data for new submissions. We map all data, including existing refSNP clusters and new submissions, to reference genome sequence if available for the organism. Otherwise, we map them to non-redundant DNA sequences from GenBank. We then use map data on co-occurrence of hit locations to either merge submissions into existing clusters or to create new clusters. We then annotate the new non-redundant refSNP (rs) set on reference sequences and dump the contents of dbSNP in a variety of comprehensive and denormalized formats on the dbSNP FTP site for release with the online build of the database.

Mapping and Reclustering New Submissions

Each build starts with a "close of data" that defines the set of new submissions that will be mapped to genome sequence by MegaBLAST for subsequent reclustering and annotation. The set of new data entering each build typically includes all submissions received since the close of data in the previous build.

Resource Integration

We annotate the non-redundant set of variations (refSNP cluster set) on reference genome sequence contigs, chromosomes, mRNAs, and proteins as part of the NCBI RefSeq project (Chapter 17). We compute summary properties for each refSNP cluster, which we then use to build fresh indexes for dbSNP in Entrez and to update the variation map in the NCBI Map Viewer. Finally, we update links between dbSNP and dbMHC, UniSTS, LocusLink, PubMed, and UniGene.

Public Release

Public release of a new build involves an update to the public database and the production of a new set of files on the dbSNP FTP site. We make an announcement to the dbsnp-announce mailing list when the new build is publicly available.

refSNP Cluster Assignment for Non-Redundant Datasets

Data submitted to dbSNP are clustered and provide a non-redundant set of variations for each organism in the database. We maintain these clusters as refSNPs in dbSNP in parallel to the underlying submitted data. We distinguish refSNPs from assay submissions by using an **rs**-prefixed (**r**ef**S**NP) Accession number instead of the **ss**-prefixed (**s**ubmitted **S**NP) Accession number assigned to individual submissions.

refSNPs are compact sets of identifiers that are used to annotate variations on other NCBI resources. A refSNP has a number of summary properties that are computed over all cluster members (Figure 5). We export the entire refSNP set in many report formats on the FTP site and as sets of results through a dbSNP batch query. We maintain both refSNPs and submitted SNPs as FASTA databases for BLAST searches of dbSNP.

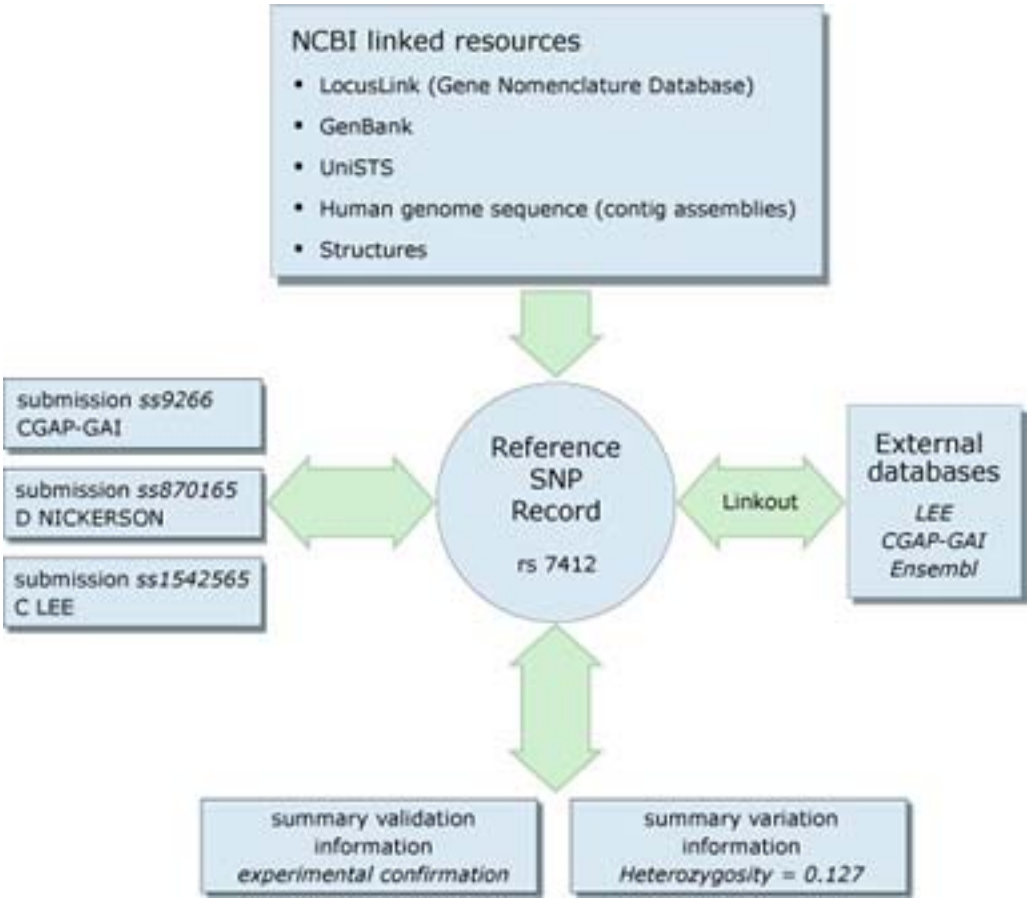


Figure 5: Schematic view of refSNP rs7412 and its connections to underlying submissions ss9266, ss870165, and ss1542565.
rs7412 has an average heterozygosity of 12.7% based on the frequency data provided by the three submissions, and the cluster as a whole is validated because one of the underlying submissions has been experimentally validated. rs7412 is annotated as a variation feature on RefSeq contigs, mRNAs, and proteins. Pointers in the refSNP summary record direct the user to additional information on the three submitter web sites, through the linkout URLs supplied in each submission. These websites may contain additional data that were used in the initial variation call, or it may be additional phenotype or molecular data that indicate the function of the variation.

Summary Data Measures

We compute summary measures for each refSNP to integrate data provided by each independent submitter.

The refSNP Clustering Process

Submitters can arbitrarily define variations on either strand of DNA sequence; therefore, submissions in a refSNP cluster can be reported on the forward or reverse strand. The orientation of the refSNP and, hence, its sequence and allele string, is set by a cluster exemplar. By convention, the clustering process (Figure 6) picks a cluster exemplar as that member of a cluster with the longest sequence. In subsequent builds, this sequence may

be in reverse orientation to the current orientation of the refSNP. When this occurs, we try to preserve the orientation of the refSNP, if possible, by using the reverse complement of the cluster exemplar to set the orientation of the refSNP sequence.

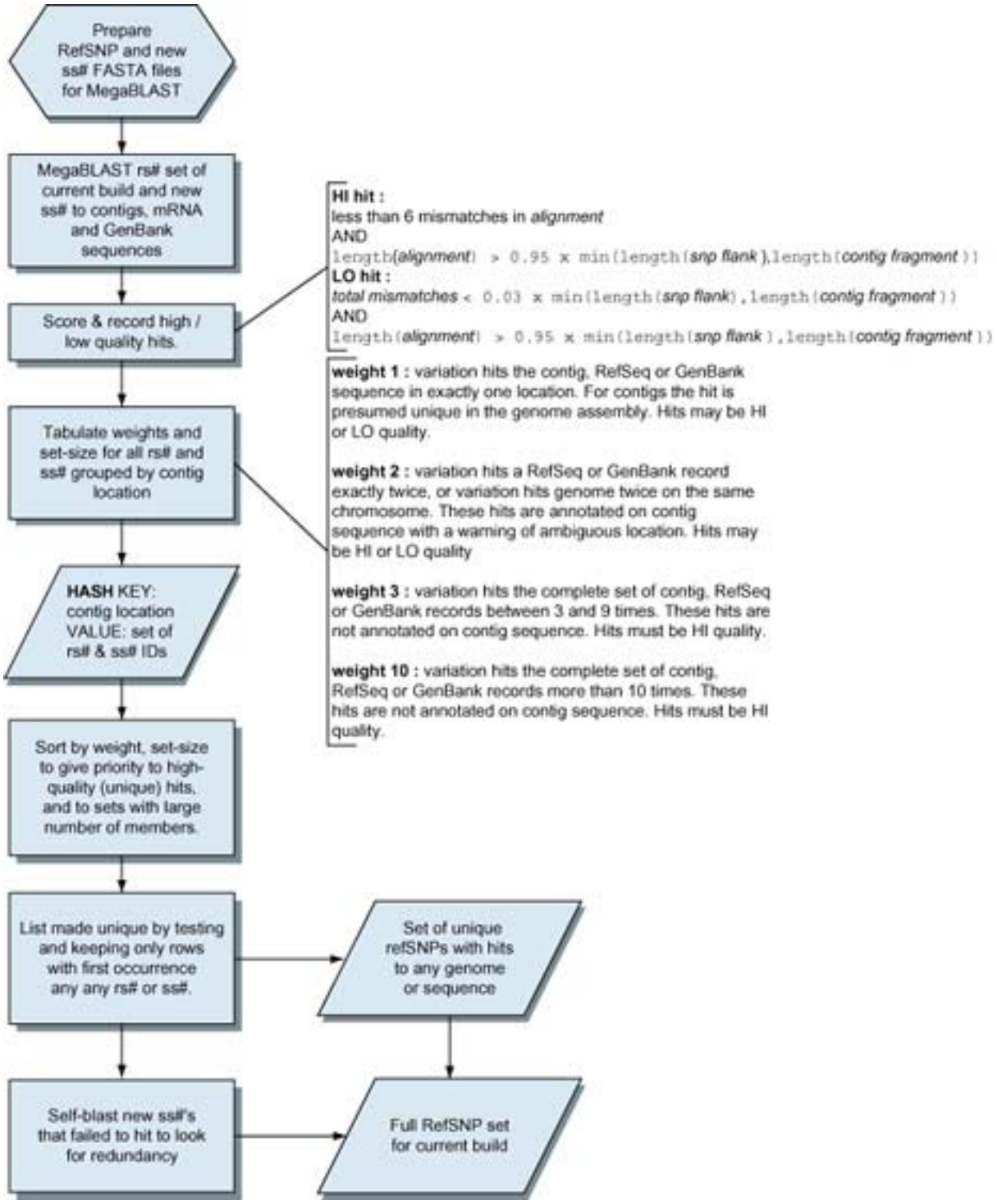


Figure 6: The dbSNP reclustering process.

We define clusters on shared locations (refSNPs) when we BLAST all existing refSNPs against contig sequence. In cases where contig sequence is not available or the variation is defined in an mRNA flanking sequence that will not map to a contig, we compute the refSNP set based on hits to the RefSeq or GenBank sequences for the organism. We rank map hits as either LO or HI quality and parse the hits to assign a weight to each refSNP. We make the set of all hits unique by dropping contig location and retaining only the first occurrence of each rs#, ss#, id, and the cluster in which each id number appears. The resulting data include all refSNPs for the current build that have at least one hit to contig, RefSeq, or GenBank sequence. We then compare the set of new submissions that fail to hit these sequence sets against each other using BLAST. This removes any potential redundancy in the incoming data, and unmapped refSNPs are instantiated for these data as well. This final merged set of data constitutes the refSNP set for the current build.

Once the clustering process determines the orientation of all member sequences in a cluster, it will gather a comprehensive set of alleles for a refSNP cluster.
Hint:When the alleles of a submission appear to be different from the alleles of its parent refSNP, check the orientation of the submission for reverse orientation.

Summary Measures of Variation

The best single measure of a variation's diversity in different populations is its average heterozygosity. This measure serves as the general probability that both alleles are in a diploid individual or in a sample of two chromosomes. Estimates of average heterozygosity have an accompanying standard error based on the sample sizes of the underlying data, which reflects the overall uncertainty of the estimate.

Additional summary measures of variation include counts of populations and individuals sampled for this variation.

Mapping to Reference Genome Sequence

When reference genome assemblies are available, we use them as anchor sequence to place refSNP clusters into a genomic context. We clean dbSNP flanking sequence with RepeatMasker and then re-map them to the most current build of each genome using MegaBLAST. The mapping results then define a new non-redundant set of variations for the genome.

Reclustering

We define a refSNP operationally as a variation at a location on an ideal reference chromosome. Such reference chromosomes are the goal of genome assemblies. However, work is still in progress in cases such as the human genome project; therefore, we must currently define a refSNP as a variation in the interim reference contig sequence. Every time there is a genomic assembly update, the interim reference contig sequence changes, and refSNPs must be updated or reclustered.

The reclustering process begins when NCBI updates the genomic assembly. We BLAST all existing refSNPs as well as any newly submitted SNPs (not yet bound to a refSNP cluster) against the genome assembly. Then, we cluster SNPs that co-locate at the same place on the genome into a single refSNP. Usually, new clusters are composed entirely of new submitted SNPs, or else the newly submitted SNPs cluster to an already existing refSNP. When newly submitted SNPs cluster among themselves, they are assigned to a new refSNP ID#, and when they cluster with an already existing refSNP, they are assigned to the cluster for that refSNP.

Sometimes a refSNP will co-locate with another existing refSNP. In this case, the refSNP with a higher ID number is retired, and all the submitted SNPs in its cluster are reassigned to the refSNP with the lower ID number.

Once the clusters are formed, the variation of a refSNP is the union of all possible alleles defined in the set of submitted SNPs that composed the cluster. Figure 6 is a detailed flow chart of the reclustering process.

NCBI Contig Annotation

We annotate weight 1 and 2 refSNP variations on NCBI RefSeq chromosomes, contig sequences, mRNAs, and proteins as variation features with multiple allele qualifiers (one per allele). Weight 2 records receive an additional warning note to indicate the ambiguous nature of the mapping result.
Hint:The two hits that define a weight 2 variation may not reflect paralogy in the genome. Sequence assemblies are imperfect, and some regions of unique genome sequence are potentially reflected in two or more contig sequence fragments. Because we are currently unable to distinguish such cases from true paralogy, we annotate the variation in both locations with a warning and leave the assessment of the flanking sequence to the user.

We do not believe that weight 3 and weight 10 variations have sufficient utility to warrant their annotation, but the mapping results for these variations are still available in dbSNP.

We annotate NoVariation records on NCBI RefSeq chromosomes, contig sequences, mRNAs, and proteins as a miscellaneous feature, or misc_feat. All dbSNP annotations also include a db_Xref cross-reference pointer back to dbSNP that uses the refSNP ID number.

Annotating GenBank and Other RefSeq Records

GenBank records can be annotated only by their original authors. Therefore, when we find high-quality hits of refSNP records to the HTGS and non-redundant divisions of GenBank, we connect them using LinkOut (Chapter 16).

We annotate RefSeq mRNAs with variation features when the refSNP has a high-quality hit to the mRNA sequence. If the variation is in the coding region of the transcript and has a non-synonymous allele that changes the protein sequence, we also annotate the variation on the protein translation of the mRNA. The alleles in protein annotations are the amino acid translations of the affected codons.

NCBI Map Viewer Variation and Linkage Maps

The Map Viewer (Chapter 19) can show multiple maps of sequence features in common chromosome coordinates. The variation map shows all variation features that we annotate on the current genome assembly. There are two ways to see the variation data. The default graphic mode shows the data as tick marks on the vertical coordinate scale. When **variation** is selected as the master map, a summary of map quality, quality warning, functional relationships to genes, average heterozygosity with standard error, and validation information are provided. If genotype, haplotype, or LinkOut data are available, the master map will also contain links to this information.
Hint:The summary values can be viewed or downloaded directly as a tab-delimited table if you select the **Show Data as Table** option from the left sidebar.

Functional Analysis

Variation Functional Class

We compute a functional context for sequence variations by inspecting the flanking sequence for gene features during the contig annotation process. We are also currently developing a method to do the same analysis on RefSeq/GenBank mRNAs.

Table 5 defines variation functional classes. We base class on the relationship between a variation and any local gene features. When a variation is near a transcript or in a transcript interval but not in the coding region, then we define the functional class by the position of the variation relative to the structure of the aligned transcript. In other words, a variation may be near a gene (locus region), in a UTR (mrna-utr), in an intron (intron), or in a splice site (splice site). If the variation is in a coding region, then the functional class of the variation depends on how each allele may affect the translated peptide sequence.

Table 5. Function codes for refSNPs in gene features.[a]

Functional class	Description	Database code
Locus Region	Variation is within 2 Kb 5′ or 500 bp 3′ of a gene feature (on either strand), but the variation is not in the transcript for the gene. This class is indicated with an L in graphical summaries.	1
Coding	Variation is in the coding region of the gene. This class is assigned if the allele-specific class is unknown. This class is indicated with a C in graphical summaries.	2
Coding-synon	The variation allele is synonymous with the contig codon in a gene. An allele receives this class when substitution and translation of the allele into the codon makes no change to the amino acid specified by the reference sequence. A variation is a synonymous substitution if all alleles are classified as contig reference or coding-synon. This class is indicated with a C in graphical summaries.	3
Coding-nonsynon	The variation allele is nonsynonymous for the contig codon in a gene. An allele receives this class when substitution and translation of the allele into the codon changes the amino acid specified by the reference sequence. A variation is a nonsynonymous substitution if any alleles are classified as coding-nonsynon. This class is indicated with a C or N in graphical summaries.	4
mRNA-UTR	The variation is in the transcript of a gene but not in the coding region of the transcript. This class is indicated by a T in graphical summaries.	5
Intron	The variation is in the intron of a gene but not in the first two or last two bases of the intron. This class is indicated by an L in graphical summaries.	6
Splice-site	The variation is in the first two or last two bases of the intron. This class is indicated by a T in graphical summaries.	7
Contig-reference	The variation allele is identical to the contig nucleotide. Typically, one allele of a variation is the same as the reference genome. The letter used to indicate the variation is a C or N, depending on the state of the alternative allele for the variation.	8
Coding-exception	The variation is in the coding region of a gene, but the precise location cannot be resolved because of an error in the alignment of the exon. The class is indicated by a C in graphical summaries.	9

[a] Most gene features are defined by the location of the variation with respect to transcript exon boundaries. Variations in coding regions, however, have a functional class assigned to each allele for the variation because these classes depend on allele sequence.

Typically, one allele of a variation will be the same as the contig (contig reference), and the other allele will be either a synonymous change or a nonsynonymous change. In some cases, one allele will be a synonymous change, and the other allele will be a nonsynonymous change. If any allele is a nonsynonymous change, then the variation is classified as a nonsynonymous variation. Otherwise, the variation is classified as a synonymous variation.

- The allele is the same as the contig (contig reference) and hence causes no change to the translated sequence.
- The allele, when substituted for the reference sequence, yields a new codon that encodes the same amino acid. This is termed a synonymous substitution.
- The allele, when substituted for the reference sequence, yields a new codon that encodes a different amino acid. This is termed a nonsynonymous substitution.

- A problem with the annotated coding region feature prohibits conceptual translation. In this case, we note the variation class as coding, based solely on position.

Because functional classification is defined by positional and sequence parameters, two facts emerge: (*a*) if a gene has multiple transcripts because of alternative splicing, then a variation may have several different functional relationships to the gene; and (*b*) if multiple genes are densely packed in a contig region, then a variation at a single location in the genome may have multiple, potentially different, relationships to its local gene neighbors.

SNP Position in 3D Structure

When a SNP results in amino acid sequence change, knowing where that amino acid lies in the protein structure is valuable. We provide this information using the following procedure. To find the location of a SNP within a particular protein, we attempt to identify similar proteins whose structure is known by comparing the protein sequence against proteins from the PDB database of known protein structures using BLAST. Then, if we find matches, we use the BLAST alignment to identify the amino acid in the protein of known structure that corresponds to the amino acid containing the SNP. We store the position of the amino acid on the 3D structure that corresponds to the amino acid containing the SNP in the dbSNP table SNP3D.

Resource Integration

Links from SNP Records to Submitter Websites

The SNP database supports and encourages connections between assay records (ss#'s) and supplementary data on the submitter's website. This connection is made using the **LINKOUT** field in the SNPassay batch header. LinkOut URLs are base URLs to which dbSNP can append the submitter's ID for the variation to construct a complete URL to the specific data for the record. We provide LinkOut pointers in the batch header section of SNP detail reports and in the refSNP report cluster membership section.

Links within NCBI

We make the following connections between refSNP clusters and other NCBI resources during the contig annotation process:

LocusLink

There are two methods by which we localize variations to known genes: (*a*) if a variation is mapped to the genome, we note the variation/gene relationship (Table 5) during functional classification and store the locus_id of the gene in the dbSNP table SNPContigLocusId; and (*b*) if the variation does not map to the genome, we look for high-quality blast hits for the variation against mRNA sequence. We note these hits with the protein_ID (PID) of the protein (the conceptual translation of the mRNA transcript). LocusLink scans this table nightly and updates the table MapLinkPID with the locus_id for the gene when the protein is a known product of a gene.

UniSTS

When an original submitted SNP record shows a relationship between a SNP and a STS, we share the data with dbSTS and establish a link between the SNP and the STS record. We also examine refSNPs for proximity to STS features during contig annotation. When we determine that a variation needs to be placed within an STS feature, we note the relationship in the dbSNP table SnpInSts.

UniGene
The contig annotation pipeline relates refSNPs to UniGene EST clusters based on shared chromosomal location. We store Variation/UniGene cluster relationships in the dbSNP table UnigeneSnp.

PubMed
We connect individual submissions to PubMed record(s) of publications cited at the time of submission. If you want to view links from PubMed to dbSNP, select **LinkOuts** as a PubMed query result.

dbMHC
dbSNP stores the underlying variation data that define HLA alleles at the nucleotide level. The combinations of alleles that define specific HLA alleles are stored in dbMHC. dbSNP points to dbMHC at the haplotype level, and dbMHC points to dbSNP at both the haplotype and variation level.

How to Create a Local Copy of dbSNP

dbSNP is a relational database with about 100 tables. NCBI deploys dbSNP in both MSSQL and Sybase environments, and the public can download the full contents of the database from the dbSNP FTP site. The following sections will guide you in this process.

Schema: The dbSNP Physical Model

A schema is a necessary part of constructing your own copy of dbSNP because it is a visual representation of dbSNP and shows the logical relationship between data in dbSNP. It is available as a printable PDF file from the dbSNP FTP site.

Data in dbSNP are organized into "zones" or boxes, depending on the nature of the data. Each zone is color coded to allow the viewer to find the data more easily. The current color groupings are available online. The data dictionary currently includes a description of all the tables in dbSNP, tables of columns and their properties, and tables of foreign keys in the conceptual model. Foreign keys are not enforced in the physical model because they make it harder to load table data asynchronously. In the future, we will add descriptions of individual columns. The data dictionary described above is available in rich text format in the file dbSNPdataDictionary.rtf.gz. The data are also available online from the dbSNP website.

Resources Required for Creating a Local Copy of dbSNP

Software:

- **Relational database software**. If you are planning to create a local copy of dbSNP, you must first have a relational database server, such as Sybase, Microsoft SQL server, or Oracle. dbSNP at NCBI runs on both Sybase and MSSQL servers, but we know of users who have successfully created their local copy of dbSNP on Oracle.
- **Data loading tool**. Loading data from the dbSNP FTP site into a database requires a bulk data-loading tool, which usually comes with a database installation. For example, we use the bcp (bulk-copy) utility that comes with Sybase.
- **winzip/gzip to decompress FTP files**. Complete instructions on how to uncompress *.gz and *.Z files can be found here [http://www.ncbi.nlm.nih.gov/Ftp/uncompress.html].

- **Perl binaries (optional)**. There is a sample Perl script that validates whether all rows in the bcp file are loaded successfully into the table. See **validating data loading** in the *Stepwise Procedure for Creating a Copy of dbSNP* for more details.

Hardware:

- **Computer platforms/OS**. Databases can be maintained on any PC, Mac, or UNIX with an Internet connection.
- **Disk space**. Currently, a complete dbSNP database needs 20 GB. You need to first create a database of at least 20 GB in size. Aside from the database device size of 20 GB, you will also need another 15 GB to store downloaded data files before loading them into your database. Allowing space for other miscellaneous uses, we recommend free disk space of 40 GB.
- **Internet connection**. We recommend a high-speed connection to download such large database files.

dbSNP Data Location

The FTP directory contains the schema, data, and SQL statements to create the tables and indices for dbSNP. The /data subdirectory contains all data files of dbSNP tables, organized as one file per table. The file name convention is: <tablename>.bcp.gz. We use the Sybase bcp tool to generate the data files was one line per table row. Columns of data in each line are tab delimited.

The /schema subdirectory contains the files dbSNP_table.atx.gz and dbSNP_index.atx.gz. These files use standard SQL DDL language to create tables and indexes.

There are many utilities available to generate table/index creation statements from a database. We use a tool called *atxtract* to generate the files.

Stepwise Procedure for Creating a Local Copy of dbSNP

1. Prepare the local area. (check available space, etc.)

2. Download schema files.dbSNP_table.atx.gz and dbSNP_index.atx.gz. Save the file in your local directory and decompress the files.

Hint:For example, on UNIX operating systems, use gunzip to decompress the files: gunzip dbSNP_table.atx.gz gunzip dbSNP_index.atx.gz to get the files dbSNP_table.atx and dbSNP_index.atx.

3. Create the tables. Open dbSNP_table.atx file with a text editor. Replace 'snp_sherry' with the name of the database you have created on your server.

Hint:If you are using Sybase and the SQL server, use the command `isql -S <servername> -U username -P password -i dbSNP_table.atx`

4. Download the data. Table data files can be found at: ftp://ftp.ncbi.nih.gov/snp/Sybase/data. There are many pieces of FTP client software available that will facilitate downloading large number of files.

Hint:Use “ftp -i” to turn off interactive prompting during multiple file transfers to avoid having to hit “yes” to confirm transfer hundreds of times.

5. Sample FTP protocol. Type `ftp -i ftp.ncbi.nih.gov` (use “anonymous” as the user name and your email as a password.) Type `cd snp/Sybase/data`. Type `ls` (if you see a list of <tablename>.bcp.gz files, you are in the right directory). Type `binary` (to set binary transfer mode). Type `mget *.gz` (to initiate transfer. Depending on the speed of connection, this may take more than 1 hour because the total transfer size currently is 1.5 GB and growing). To decompress the *.gz files, type `gunzip *.gz` (currently, the total size of the uncompressed bcp files is more than 10 GB).

6. Load data. Use the data-loading tool of your database server (e.g., bcp for Sybase).

Hint:Because of the sheer volume of data, our tests to load data and then create indexes on a Sun Sparc Sybase server using "fast bcp" took about 8 hours.

7. Create indices. Open the dbSNP_index.atx file with a text editor and replace "snp_sherry" with the local name of the database as you have created it on your server. If you are using Sybase and the SQL server, use the command `isql -S servername -U username -P password -i dbSNP_index.atx`

8. Use scripts to automate data loading. In the dbSNP FTP snp/Sybase area, there is a sample UNIX C shell script cmd.loaddata that will combine tasks of creating tables, loading data, and creating indexes.

9. Validate data loading. In the dbSNP FTP snp/Sybase area, there is a sample Perl script that validates whether all data are loaded successfully. For each table, it compares the number of rows in the bcp file with the number of rows in the newly loaded table and reports discrepancies.

10. Data integrity (creating a partial local copy of dbSNP). dbSNP is a relational database. Each table has either a unique index or a primary key. Foreign keys are not reinforced. There are advantages and a disadvantage to this approach. The advantages are: it makes it easy to drop and re-create the table using dbSNP_table.atx, which makes it possible to create a partial local copy of dbSNP. For example, if you are interested only in the original submitted SNP and their population frequencies and not in their map locations on NCBI genome contigs or GenBank Accession numbers (both are huge tables), then these tables can be skipped (i.e., SNPContigLoc and MapLink). Of course, to select tables for a particular query, the contents of each table and the dbSNP entity relationship (ER) diagram need to be understood. The disadvantage of unreinforced references is that either the stored procedures or the external code needs to be written to ensure the referential integrity.