308–311 *Nucleic Acids Research, 2001, Vol. 29, No. 1*

dbSNP: the NCBI database of genetic variation

S. T. Sherry*, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski[1] and K. Sirotkin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA and [1]National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA

Received October 3, 2000; Accepted October 4, 2000

ABSTRACT

In response to a need for a general catalog of genome variation to address the large-scale sampling designs required by association studies, gene mapping and evolutionary biology, the National Center for Biotechnology Information (NCBI) has established the dbSNP database [S.T.Sherry, M.Ward and K.Sirotkin (1999) *Genome Res.*, 9, 677–679]. Submissions to dbSNP will be integrated with other sources of information at NCBI such as GenBank, PubMed, LocusLink and the Human Genome Project data. The complete contents of dbSNP are available to the public at website: http://www.ncbi.nlm.nih.gov/SNP. The complete contents of dbSNP can also be downloaded in multiple formats via anonymous FTP at ftp://ncbi.nlm.nih.gov/snp/.

BACKGROUND

A key aspect of research in genetics is the association of sequence variation with heritable phenotypes. Occurring roughly every 1200 bp in comparisons of a pair of human chromosomes, single nucleotide polymorphisms (SNPs) are among the most common genetic variation. There is currently great interest in SNP discovery since a dense catalog of SNPs is expected to facilitate large-scale studies in association genetics (1), functional and pharmaco-genomics (2), population genetics and evolutionary biology (3), and positional cloning and physical mapping (4). To serve this need for such a general catalog, the National Center for Biotechnology Information (NCBI) established the Single Nucleotide Polymorphism Database (http://www.ncbi.nlm.nih.gov/SNP) in collaboration with the National Human Genome Research Institute (NHGRI).

Since its inception in September 1998, the dbSNP database has served as a central, public repository for genetic variation. Once such variations are identified and cataloged in the database, additional laboratories can use the sequence information around the polymorphism and the specific experimental conditions for further research applications. As with all NCBI resources, the data within dbSNP is available freely and in a variety of forms.

SCOPE

dbSNP currently classifies nucleotide sequence variations with the following types and percentage composition of the database: (i) single nucleotide substitutions, 99.77%; (ii) small insertion/deletion polymorphisms, 0.21%; (iii) invariant regions of sequence, 0.02%; (iv) microsatellite repeats, 0.001%; (v) named variants, <0.001%; and (vi) uncharacterized heterozygous assays, <0.001%. There is no requirement or assumption about minimum allele frequencies or functional neutrality for the polymorphisms in the database. Thus, the scope of dbSNP includes disease-causing clinical mutations as well as neutral polymorphisms. In addition to the record identifiers assigned by both the submitter and NCBI, dbSNP entries record the sequence information around the polymorphism, the specific experimental conditions necessary to perform an experiment, descriptions of the population containing the variation and frequency information by population or individual genotype.

The current level of activity in the discovery of general sequence variation suggests that SNP markers with unknown selective effects will be the majority of submitted records. Although most submissions are currently for *Homo sapiens*, dbSNP already has submissions for *Mus musculus*, and in general the database can accept variation information from any species and from any part of a particular genome. dbSNP is currently integrated with other large public variation databases such as the NCI CGAP-GAI database of EST-derived SNPs (6), the TSC (The SNP Consortium, Ltd) variation initiative (6) and HGBASE (7). Links to these and other future public databases are established by the LinkOut scheme discussed below.

UTILITY

dbSNP links variations (polymorphisms and clinical mutations) to other NCBI sequence resources via BLAST and E-PCR analysis of the flanking sequence that immediately surrounds the variation. Links to the literature databases are made with the citation information provided at submission time. This integration process makes dbSNP part of the NCBI 'discovery space' as illustrated in Figure 1. In this model, dbSNP serves dual roles as both a 'first point of entry' into the resource network for query and retrieval of specific variation records, and as an information server for searches that start in other resources such as GenBank, PubMed, LocusLink or the genome sequence databases.

As the final results of various genome projects accumulate, it is intended that all variations will be associated with a nucleotide sequence record and/or physical map contig. In the soon approaching post-sequencing phase of the human genome project, annotation of the sequence with features such as new genes or regulatory regions will provide new functional contexts for currently 'anonymous' variations that have been

*To whom correspondence should be addressed. Tel: +1 301 435 7799; Fax: +1 301 480 9241; Email: sherry@ncbi.nlm.nih.gov

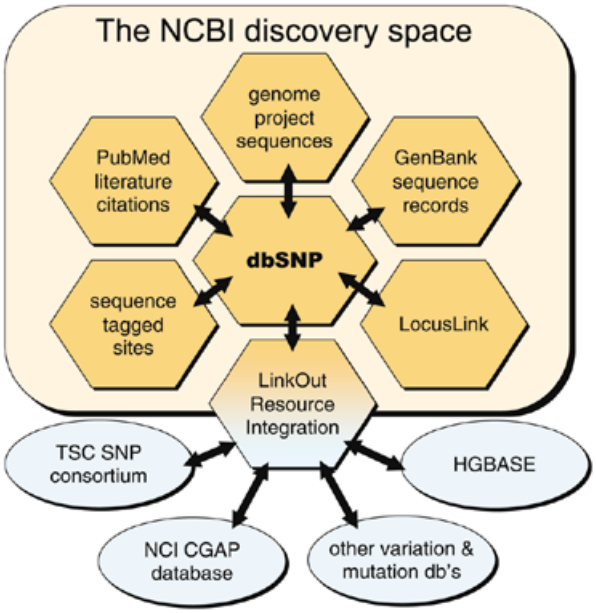

Figure 1. Records in dbSNP are cross-annotated within other internal information resources such as PubMed, genome project sequences, GenBank records, the LocusLink nomenclature/sequence database and the dbSTS database of sequence tagged sites. Users may query dbSNP directly, or start a search in any part of the NCBI discovery space to construct a set of dbSNP records that satisfy their search conditions. Records are also integrated with external information resources through hypertext URLs that dbSNP users can follow to explore the detailed information that is beyond the scope of dbSNP curation.

found on random sequence. As records appear for these new genes, links to dbSNP variations will be automatically annotated on the appropriate Reference Sequence or UniGene cluster. Resource integration of variation records extends beyond NCBI by use of 'LinkOut URLs' that refer to external databases with further information about the variation. This integration is important when one considers the general task of effectively annotating a complete genome for variation and its consequences for the organism (Fig. 2). NCBI has adopted the model in which variations are cataloged in dbSNP while functional descriptions of the local sequence region are noted as GenBank, dbSTS, Reference Sequence, LocusLink or UniGene records.

Since genes and their component nucleotides are potentially involved in multiple pathways and hence multiple downstream phenotypes, NCBI does not annotate the detailed biochemical or phenotypic consequences of variation directly on the sequence. Rather, links are maintained in dbSNP to external databases that each characterize particular axes of phenotypic variation, in much the same way that LocusLink maintains a current set of sequence accessions and nomenclature information for genes. In this fashion, dbSNP records can be linked to more complete descriptions of individual variations in locus-specific mutation databases. A federation of such databases can be found online at http://ariel.ucs.unimelb.edu.au:80/~cotton/guide1.htm.

We are designing dbSNP to facilitate searches along five major axes of information: (i) sequence location, (ii) function, (iii) cross-species homology, (iv) SNP quality or validation status and (v) degree of heterozygosity (degree of population variation). By setting thresholds of inclusion on one or more of these axes, users can extract the subset of records that are best suited to their research needs.

SCALE

As of this writing, dbSNP contains 1 463 178 submissions from 97 registered groups describing variation in five species (human, mouse, rat, chimpanzee and the malaria parasite). Submissions can be divided into four general categories with the following percentages of the total database size: (i) SNP mining from the human genome project sequences, 65%; (ii) private investigator/corporate experimental results, 28%; (iii) mined from EST databases, e.g. (6), 6%; (iv) continuing results of the NHGRI SNP discovery RFA, 1%. Public and private initiatives to discover new SNPs in humans identified over 306 000 variations in the period 1999–2000 (7,9). An additional 1 095 945 submissions were received from groups that mined the BAC end sequences and clone overlaps that were used to construct the human genome sequence contigs. These data mining projects conducted several rounds of data mining and submission to dbSNP for the June, July and September data freezes that were held during the summer and fall of 2000 for analysis of the working draft genome sequence. These SNPs will be clustered within an estimated 30 000 sequence overlap regions that are dispersed throughout the genome.

SUBMISSION

Submissions are welcomed from all sources, public and private. These groups are working on a variety of aspects of variation discovery, new technologies for detection and rapid genotyping in large samples. Data can be submitted directly to NCBI via instructions on the dbSNP 'How to Submit' Web page (http://www.ncbi.nlm.nih.gov/SNP/get_html.cgi?whichHtml=how_to_submit). Required submission information includes the observed alleles at a particular locus, the flanking sequence that surrounds the mutation, the experimental methods used and a pointer to a companion STS or GenBank record. Each individual laboratory is assigned a 'handle' to serve as a unique identifier, which will allow submissions to be associated with a specific laboratory. NCBI will also assign dbSNP accessioning, i.e. ss#, to each submitted variation. A reference identifier, i.e. rs#, will also be assigned to each unique variation in an organism reference genome. These will be used to map the variations to external resources or databases, including other NCBI databases.

SEARCHING

dbSNP can be searched directly or via other NCBI resources that comprise the NCBI discovery space as illustrated in Figure 1. Direct searching can be done by submitter handle (laboratory), new batches of submissions, identification method used, population type studied, publication title, level of population variation or STS mapping information. As an integrated part of NCBI, the contents of dbSNP are cross-linked to records in other information resources such as GenBank, LocusLink, the human genome sequence and PubMed. The result sets from

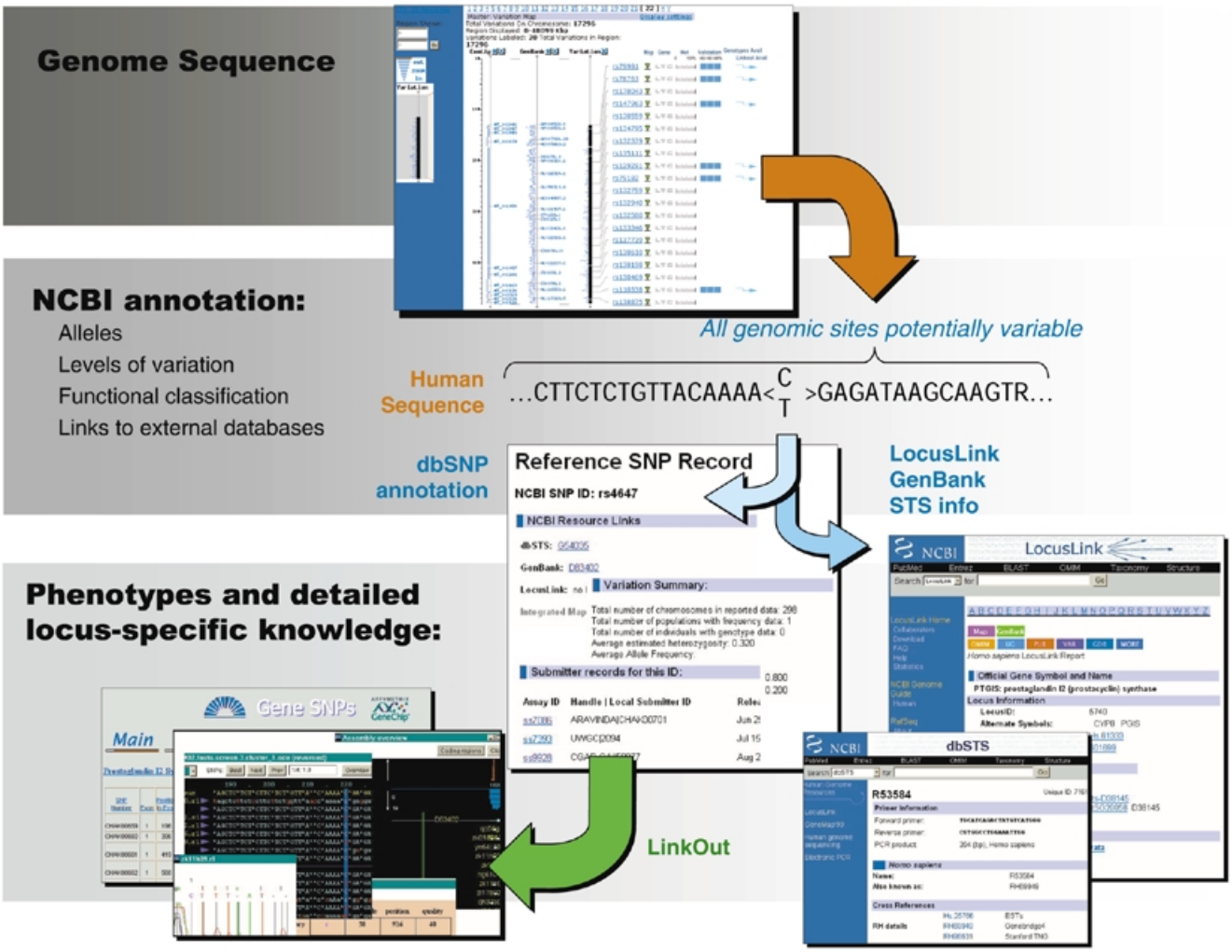

Figure 2. Annotation of the multiple biochemical or phenotypic consequences of individual variations in genomic sequence is accomplished through (i) the annotation of a single dbSNP record on the genome sequence to indicate the presence and extent of variation, and (ii) the maintenance of a list of accession links and URLs within the dbSNP record to other information resources. In this way a single variation can be easily represented in multiple biochemical pathways or phenotypic backgrounds.

queries in any of these resources will point the user back to the relevant records in dbSNP.

BLAST. dbSNP can be searched with the standard BLAST algorithm that will compare a user-submitted sequence against all flanking sequence records in dbSNP. The BLAST service is provided on the dbSNP homepage, rather than the general NCBI BLAST page.

LocusLink. dbSNP can also be queried by integrating it with other NCBI resources. Via LocusLink, queries can be done by gene name or nomenclature association. Query results from the LocusLink database will show a purple 'V' button in SNP records have been mapped to the gene. Clicking on the 'V' will lead to a list of the reference SNP records for any gene in the LocusLink database.

Entrez. Selecting the 'Linkout' display format from the pull-down menu beneath the query bar will provide a hypertext link to dbSNP (labeled as 'NCBI variation database') when dbSNP submissions have been mapped to the queried sequence. Following the link will lead to the list of reference SNP records similar to those for LocusLink.

Genome sequence. The NCBI genome viewer can be set to show 'variations' as a sequenced-based map. This map is aligned in common sequence-based coordinates with other sequence features such as gene regions, STS markers, reference sequence contigs and clone sequences. By clicking 'display options' the user may select the variation map as the 'master' map and 'show verbose'. In this configuration, the full annotation of variations on the genome sequence are visible. The extended display shows 'at-a-glance' indicators of: quality of the mapped location of the variation; functional classification (locus-region, transcript, or coding region) if variation is in a gene region; 95% confidence interval for average heterozygosity; quality of marker (validation or success rate probability); genotypes available; and submitter linkouts available.

DOCUMENTATION

The database is updated after each new data submission. A regularly updated database summary documents the number of SNP identified, submitters, publications cited, and methods and populations defined. Complete submission guidelines are available on the dbSNP website. A FAQ page lists frequently asked questions derived from user inquiries. The complete contents of dbSNP are available via anonymous ftp from the dbSNP ftp site in the following formats: (i) submission format in which we receive the data; (ii) FASTA format for users who wish to maintain a local dbSNP blast database; (iii) Sybase table dumps of all tables in dbSNP; (iv) refSNP document summaries providing summary information on each refSNP cluster in ASN.1-binary, ASN.1-text and flatfile; and (v) a database dump in XML format for interdatabase exchange. A complete description of the dbSNP ftp area and these report formats can be found at ftp://ncbi.nlm.nih.gov/snp/00readme.

FUTURE PLANS

dbSNP continues to make enhancements to the user interface to improve searching and data submission. The new main query interface will soon accept boolean operators and fielded queries so that the user may simultaneously constrain a search by degree of validation, map location, functional class, frequency level, mapping quality or strings in textual annotation. This expanded query facility will also permit structured queries and batch retrieval of results. Online Web data submission will complement the established batch submission process.

dbSNP is a rapidly maturing database. Although many contributors submit data, the majority of data has been received from large, genome-oriented data mining and discovery projects. For this reason, dbSNP is expected to enjoy continued growth over the next few years. Continued data exchange with other public variation and mutation databases and the extension of the database to support haplotype data objects will also increase the amount of in data dbSNP and enhance its utility.

REFERENCING dbSNP

We suggest that dbSNP be referenced as follows: Sherry,S.T., Ward,M. and Sirotkin,K. (1999) dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. *Genome Res.*, **9**, 677–679.

We suggest that the abbreviation dbSNP be used for this database.

ADDRESSES

For assistance in using dbSNP, please write to info@ncbi.nlm.nih.gov. For other questions regarding dbSNP, please contact the dbSNP support staff at snp-admin@ncbi.nlm.nih.gov. Mail may addressed to Steve Sherry, National Center for Biotechnology Information, National Library of Medicine, Building 38A, Room 8N805, Bethesda, MD 20894, USA. Tel: +1 301 435 7799; Fax: +1 301 480 9241.

ACKNOWLEDGEMENTS

This research was supported in part by an appointment to the NLM Associate Fellowship Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education.

REFERENCES

1. Kruglyak,L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.*, **22**, 139–144.
2. Carulli,J.P., Artinger,M., Swain,P.M., Root,C.D., Chee,L., Tulig,C., Guerin,J., Osborne,M., Stein,G., Lian,J. and Lomedico,P.T. (1998) High throughput analysis of differential gene expression. *J. Cell. Biochem.*, **30–31** (Suppl.), 286–296.
3. Cavalli-Sforza,L.L. (1998) The DNA revolution in population genetics. *Trends Genet.*, **14**, 60–65.
4. Collins,F.S. (1999) Shattuck lecture–medical and societal consequences of the Human Genome Project. *N. Engl. J. Med.*, **341**, 28–37.
5. Buetow,K.H., Edmonson,M.N. and Cassidy,A.B. (1999) Reliable identification of large numbers of candidate SNPs from public EST data. *Nature Genet.*, **21**, 323–325.
6. Masood,E. (1999) As consortium plans free SNP map of human genome. *Nature*, **398**, 545–546.
7. Brookes,A.J., Lehväslaiho,H., Siegfried,M., Boehm,J.G., Yuan,Y.P., Sarkar,C.M., Bork,P. and Ortigao,F. (2000) HGBASE A Database of SNPs and other variations in and around human genes. *Nucleic Acids Res.*, **28**, 356–360.