European Patent Office

Principal Directorate Tools/Documentation

Trilateral Technical conference

Washington May 17-20, 2004

Subject: **Study on Single Nucleotide Polymorphism (SNP) / Haplotype Databases and Search Tools for Examiners** (Status Report)

Project: DR2

Author: EPO

EPO Responsible: Gérard Giroud

Prepared by: Ana Richart de la Torre & Stéphane Nauche

1. Introduction

During the last few years the development of pharmacogenetics, based on how variation in human genes leads to variation in our response to drugs, has gained a great importance.
This, together with the development of the "Human genome project", focused on determining the sequences of the chemical base pairs that make up human DNA, has lead to the discovery of multiple variations of genomic DNA, such as single nucleotide polymorphisms (SNPs) (1) and haplotypes (2), a combination of those. Therefore it has sparked a great interest in patent protection resulting in an increase in the number of patent applications claiming SNPs and haplotypes as disease markers, as well as corresponding methods of use (figure 1).
Furthermore, patent applications disclosing SNPs and haplotypes claim hundreds or thousands of related nucleic acid molecules. Claims to SNPs and haplotypes do present special search and examination challenges.

Currently examiners, do not have optimized tools to perform exhaustive searches, therefore the analysis of claimed SNPs or haplotypes comprises multiple searches where a plurality of sequences and keywords needs to be entered for a search to be complete. The search for prior art relating to the specific use of SNPs or haplotypes should be done using automated tools, whereby a limited number of sequences and keywords suffice for the whole search (Trilateral Project DR2).

The EPO presented a report on SNPs/Haplotypes databases during the last trilateral working group. The Trilateral Offices agreed to further study Single Nucleotide Polymorphism (SNP)/Haplotype Databases and Search Tools for Examiners.
The EPO provided a questionnaire to its partners to identify most relevant SNP/Haplotype Databases. This document provides a further status report

(1) SNPs are single base pair positions in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals in some population(s), wherein the least frequent allele has an abundance of 1% or greater. "SNPs what they are & what might they tell us", Anthony Brookes Research Group, available at http://www.cgr.ki.se/cgb/groups/brookes/snps.htm.

(2) The term 'haplotype' refers to a combination of SNPs on a chromosome, usually within the context of a particular gene. "Haplotype identification" at http://www.variagenics.com/articles/haplotypeid.html.

2. Patent applications disclosing SNPs/Haplotypes

2.1 Filing figures

As a consequence of the great development of research in this field, an increase in the number of patent applications has also taken place during the last few years. It is reflected by the analysis of published and incoming applications disclosing SNPs and haplotypes, recorded at WPI, EPODOC, and DOSYS databases, as shown in figure 1.

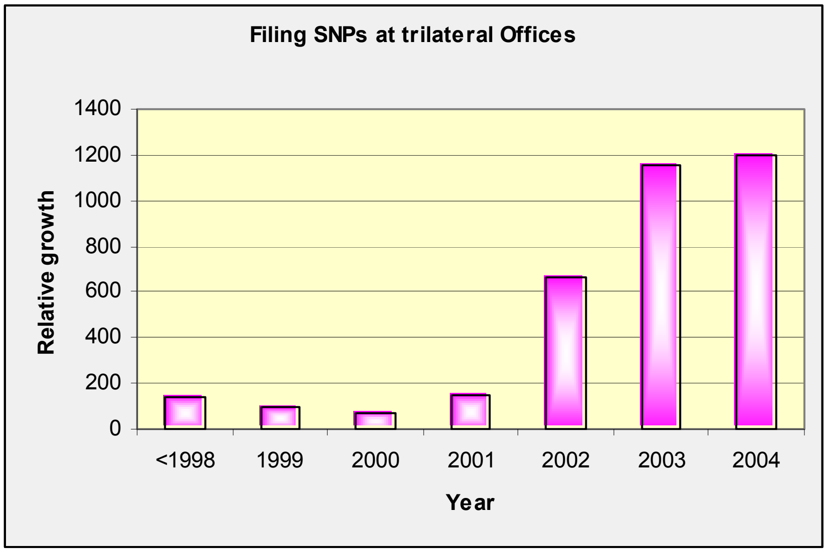

Figure 1: Number of patent applications claiming SNPs and haplotypes in the last years
(To avoid different interpretations data provided by the trilateral partners are provided according to a base100 for 1999)

2.2 Concealed workload

A duplication (or triplication, or more) of the work needed under one single filing on SNPs and haplotypes field, can be expected in the form of non-unitary patents. Roughly one in four applications in the field is non-unitary, each of these resulting in **at least** a duplication of the work, as can be seen in the next figure.

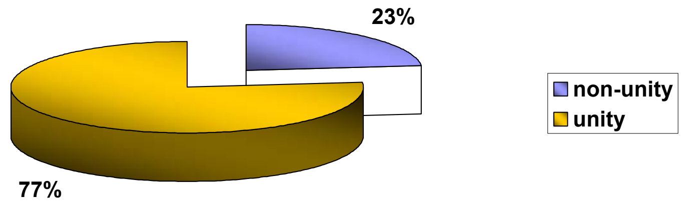

Figure 2: Share of non-unitary applications over the total number of filings (source: EPO)

2.3 Origin of patent applications

The country of origin of patent applications disclosing SNPs and Haplotypes was analyzed to assess where major developments in the art arise. This has twofold implications: Firstly, as a partial indication of the biotechnology workload that Trilateral Offices will be facing, and secondly, as to where sources of information such as SNPs and haplotypes databases or patent databases, are likely to develop.

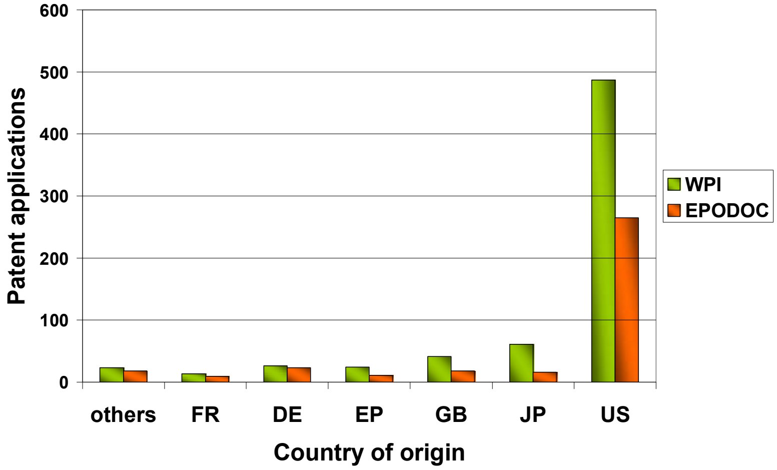

Figure 3: Representation of the country of origin of applications disclosing SNPs and Haplotypes contained in WPI and EPODOC databases (Source: EPO)

As shown in figure 3, the vast majority of SNPs and haplotypes patent applications stem from the US, all other countries having a significantly lower number of filings.

2.4 Source of claimed SNPs

Since a set of available databases reporting SNPs and haplotypes should be identified (Trilateral project DR2), an analysis regarding what species the sequences claimed in the applications are associated with, is necessary. This way, it is possible to focus on those databases that show the best species coverage according to the patent applications.

The vast majority of developments in the field of SNPs and haplotypes are being carried out by United States research groups, and as consequence, most of the patent applications in the field have US priority (section 2.3).

The search in full-text databases of published patent(s)/patent applications allows the selection of applications claiming SNPs and Haplotypes, ruling out those claiming methods of detection instead sequences. Since the aim of this project is to tackle the search of applications claiming sequences containing SNPs and Haplotypes, a cluster of full text English databases was used to assess which species the sequences are associated with in these applications.

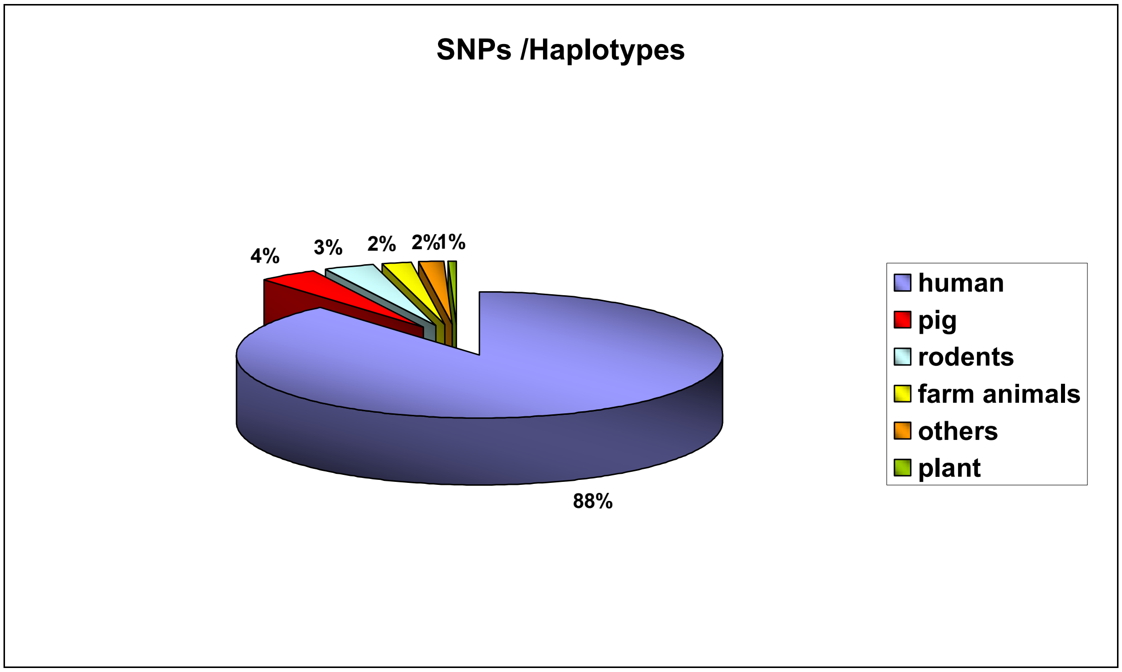

Figure 4: Assessment of the species the sequences are associated with, in the applications disclosing SNPs or Haplotypes (Source: EPO)

As shown in figure 4, 88% of the applications disclosing SNPs or Haplotypes are related to human, followed by 4% claiming pig sequences and 3% dealing with rodents (rat and mouse). According to USPTO, more than 95% of their SNPs applications disclosed are of human origin.

In view of these percentages, the selection of a set of databases reporting SNPs and Haplotypes, should be focused as a first step, on those covering human (*Homo sapiens*) sequences, and also pig (*Sus scrofa*) and rodents (*Mus musculus, Rattus norvegicus*) sequences.

Databases

The set of databases that the examiners at EPO are currently using to perform a haplotype or SNP search is the following:

- EMBL (EBI): constitutes Europe's primary source for DNA and RNA sequences. Wild type sequences can be queried by running FAST programs, retrieving the results from a set of databases included in its search environment such as, EMBL ALL, or naGeneSeq.

-HGVbase: summarizes all known variations in the human genome, reporting SNPs in a percentage greater than 99% of its entries. It has recently been included in the search environment of EMBL, and sequences can be queried by running FAST programs.

-Registry: contains over 14.5 million unique substance records identified by CAS (Chemical Abstracts Service). All types of organic and inorganic substances are covered, including protein and nucleic acid sequences. It offers the possibility of performing searches of sequences with some variation (alternative residues), where the query string can be embedded in a larger sequence.

- The SNP Consortium ltd: database reporting more than 1,8 million of single nucleotide polymorphisms. A SNP search can be performed by introducing the name of a given gene and retrieving all the SNPs reported on it.

-dbSNP: NCBI's SNPs database, reporting data from several species. It also offers the possibility of retrieving all known SNPs of a gene, and visualizing the span of the sequence where they are located.

-GeneCards: is a database of human genes, their products, functions and their involvement in diseases. It reports the set of alternative symbols the genes are known under, what allows the examiners to face the fact that a non-standardized naming of the genes has been established in the prior art. Among many other features, a table containing known variants of the query gene is reported, and links to SNPdb are provided for each variant.

3. Provision of SNP information in Patent Applications: Representation of SNPs in Sequence Listings.

A new standard for the presentation of nucleotide and amino acid sequence listings in patent applications has been developed. Controlled vocabulary in English substitutes the numerical codes of the previous standard, to remain synchronized with the worldwide database providers.

Among many other improvements in relation with the previous standard in use, (ST 25), the new standard provides specific mechanism for the annotation and representation of mutations and polymorphisms.

The representation of variants in the sequence listing shall be mandatory, and therefore said variants will be disclosed either as separate SEQ ID's or each position has to be described accurately in the Features of the sequence.

Each sequence will be accompanied by a minimal set of characteristics composed of lines. The FH (feature table header) and FT (feature table data) lines provide mandatory information on the sequence such as: coding sequences (CDS), origin of sequences, and **variants and mutations**. The FH lines are present to increase understandability of the FT lines, and both of them will be formed by controlled vocabulary, as used in the major sequence databases, such as DDBJ, EMBL and GenBank.

Regarding the information reported about variants and mutations in the FH and FT lines:

- For nucleotide sequences:

  -the 'Location/Qualifiers' key shall be used on the FH line;
  -the 'variation' Feature key shall be used on the FT line.

- For amino acid sequences:

  -the 'Location/Qualifiers' key shall be used on the FH line;
  -the 'variant' Feature key shall be used on the FT line.

Examples of Variation key feature lines for a nucleotide sequence:

An adenine replaces the nucleotide given in position 4 of the sequence.

```
FH   Key               Location/Qualifiers
FH
FT   variation         4
FT                     /replace = “a”
```

Examples of Variant key feature lines for an amino acid sequence:

An Isoleucine replaces the amino acid given in position 13 of the sequence.

```
FH   Key               Location/Qualifiers
FH
FT   Variant           13
FT                     /replace = “I”
```

4. Representation of SNPs and Haplotypes information in General Sequence Databases

The three main nucleic acid sequence databases (EMBL, Genbank and DDBJ), collaborate to produce and exchange sequences daily, holding the same data. The collaboration includes a common feature table which describes biologically interesting regions in the sequence entries, a common set of unique sequence identification numbers, and facilitates the exchange of sequences among the three databases. Only slight stylistic differences can be found in how the databases present the information.

Since EMBL, GenBank and DDBJ show a common feature table, EMBL entries will be given as examples of formats and contents of an SNP or haplotype release in a General Sequence Database.

The terms SNP and haplotype are included in two different feature keys being “variation” the key comprising SNPs (not exclusively) amongst other types of genetic alterations, and “source” is the key for haplotype, which appears as an “optional qualifier” of this feature.

4.1 Example of variation release:

Query: IGF2AS and variant
Release:

```
ID   AY375532_3; parent: AY375532
AC   AY375532;
FT   variation        1145
FT                    /frequency=".42"
FT                    /replace="a"
SQ   Sequence      1 BP;
     g                                                             1
//
```

A SNP is reported on position 1145 of IGF2AS gene sequence (insulin-like growth factor 2), where the base G of the parent sequence is replaced by the base A in 42% of the cases.
A link to the complete entry is furnished (AY375532) allowing the visualization of the whole sequence. In most of genes queried the parent sequence is extremely long and in this particular case, it consists of 20229 base pairs. Additionally, the concrete positions where variants are located on the sequence, do not present an special enhancement on visualization. On top of these, there is not consistency in the numbering of base pairs and nomenclature of genes between the prior art and the applications.
These difficulties make the search of SNPs slow and cumbersome for examiners. It would be useful to improve visualization on bases where SNPs are located in the parent sequence, and

alignments of query sequences with sequences available in the databases, should be done in order to make coincide the different positions of the base pairs.

4.2 Example of haplotype release:

Query: DRD4 and haplotype.
Release:

```
ID   AB107017_1; parent: AB107017
AC   AB107017;
FT   source          1..217
FT                   /country="Japan"
FT                   /db_xref="taxon:9606"
FT                   /haplotype="Hsa108"
FT                   /mol_type="genomic DNA"
FT                   /organism="Homo sapiens"
SQ   Sequence   217 BP;
     gtgcgccgcc ctccccgccc gcgcccgcgc cccggcgccc ccgcgccccg cccgccgccc        60
     tcaccgcggc ctgtgcgctg tccggcgccc cctcggcgct ccccgcaggt tcgtggccgt       120
     ggccgtgccg ctgcgctaca accggcaggg tgggagccgc cggcagctgc tgctcatcgg       180
     cgccacgtgg ctgctgtccg cggcggtggc ggcgccc                                217
//
```

This release reports the sequence of the second intron of dopamine receptor D4 (DRD4), and additionally, the presence of an haplotype ("Hsa 108") in this span of the gene sequence. Since, only the name of such an haplotype is given, but not further details about the positions and replacements of base pairs are reported, it does not yield sufficient information to carry out an haplotype search in a complete way.

5. Specific Databases

A set of databases containing SNPs and/or haplotypes has been selected from the existing available mutation and variation databases (around 300 databases including LSDBs). The factors considered when carrying out this selection have been:

i) coverage of the major species dealt with in the incoming patent applications;
ii) frequency of updates in each database;
iii) number of records and quality of their sources (journals, LSDBs, research groups, and others);
iv) data redundancy;
v) links to other databases;

A selection of relevant databases is given as Annex 1

The following databases has been selected so far as the most valuable ones.:

Database	Keyword search	Sequence Search
■ STN Database (CAS Registry)	X	X
■ dbSNP	X	X
■ HGVBase	X	X
■ Genecards	X	
■ OMIM	X	
■ HAP^{TM} Database	X	

Some offices do have the databases available in their search environments, other use access them via the internet. Depending on the database parsing and engine used to query the data, some databases can be queried either for sequences or for keywords or for both.

Conclusion:

As previously correctly stated by the USPTO (Trilateral pre-conference, Tokyo, November 2003) expanding SNP and haplotype technology has resulted in an increase in the number of patent applications claiming SNPs and haplotypes, and claims to SNPs and haplotypes do indeed present special search challenges.

The search for a SNP/haplotype is complex by nature. In addition, selection of appropriate databases is far from trivial; ad hoc search tools have not yet been investigated in details.

Taking into account the difficulties encountered in having a clear and exhaustive view in the databases and tools needed for searching claims directed to SNPs and haplotypes, it is proposed to continue the SNP/haplotype study by providing detailed information the seven databases regarded as the most valuable and investigate on search engines and interfaces to improve searching efficiencies.

ANNEX 1: Selection of relevant databases

Human Genome Mutation Database (HGMD)

CONTENTS	Comprises various types of mutations, including polymorphisms in genes causing inherited disease. HGMD does not include mutations lacking obvious phenotypic consequences.
URL	http://www.hgmd.org
SPECIES COVERAGE	Human
ENTRIES	38177, total number of entries; about 800 entries of SNPs
REDUNDANCY	non redundant
UPDATES	monthly
SOURCES	>250 journals. Contains also links to unpublished mutation data available in online public locus-specific mutation databases.
LINKS (Xref)	The records are cross-referenced to different LSDBs. Other useful links to core databases home pages are provided.
SRS-EMBL (Availability)	NO
QUERY SEQUENCES	NO

SNPdb

CONTENTS	NCBI's SNPs database, reporting data from several species. .
URL	http://www.ncbi.nih.gov/SNP/
SPECIES COVERAGE	It covers amongst many other species, human, rodents, pig, cattle sps, and plant species.
ENTRIES	11,805,698 RefSNPs
REDUNDANCY	no answer
UPDATES	every 4-8 weeks
SOURCES	SNPs derived from ~300 sources. The major contributors to the database are laboratories associated with the National Human Genome Research Institute (NHGRI) grants program.
LINKS (Xref)	GenBank eventually
SRS-EMBL (Availability)	NO
QUERY SEQUENCES	NO

Human Genome Variation Database (HGVbase)

CONTENTS	Summarizes all known variations in the human genome, facilitating genotype-phenotype association analyses that explore how SNPs and other sequence variations may influence phenotypes. .
URL	http://hgvbase.cgb.ki.se/
SPECIES COVERAGE	Human
ENTRIES	2,859,130 records (99% reporting SNPs)
REDUNDANCY	non redundant
UPDATES	Last update on 23-July-2003. This activity has ceased for the moment due to limitations of funding. HGVbase is presently more focused into a phenotype-genotype project.
SOURCES	SNPs derived from nearly 800 sources. HGVbase data is harvested (with permission) or submitted from all major public genome databases and extracted from published literature. Individual or bulk submissions from research groups are also received.
LINKS (Xref)	EMBL, Ensembl, GenBank, dbSNP, OMIM, PubMed, PolyPhen.
SRS-EMBL (Availability)	YES
QUERY SEQUENCES	Direct DNA sequence searches use the BLAST program. It is possible to enter a query sequence up to 25,000 bases in raw format as well as DNA sequences of the allelic variations plus their flanking domains.

ALFRED

CONTENTS	Focused on allele frequencies, comprising DNA polymorphisms and other sequence variations, sufficiently defined and studied in at least 6 human populations. .
URL	http://alfred.med.yale.edu/alfred/index.asp
SPECIES COVERAGE	Human
ENTRIES	932 polymorphisms, including SNPs, STRPs, VNTRs, INDELs, and Haplotypes. It's estimated that at least half of them are SNPs.
REDUNDANCY	Some redundancy
UPDATES	Daily
SOURCES	Data from the published literature, and directly submitted from researchers.
LINKS (Xref)	PubMed, Gene Bank, dbSNP, OMIM, GDB, CHLC, CEPH, and LSDBs.
SRS-EMBL (Availability)	NO
QUERY SEQUENCES	NO

The SNP Consortium ltd

CONTENTS	Database reporting more than 1,8 million of single nucleotide polymorphisms.
URL	http://snp.cshl.org/
SPECIES COVERAGE	Human
ENTRIES	~1.8 million SNPs.
REDUNDANCY	no answer
UPDATES	Last update on October 24th, 2002: final ~400,000 previously unreleased TSC SNPs made available via website. These SNPs have also been submitted to dbSNP.
SOURCES	TSC allele frequency/genotype project member laboratories (Celera, Motorola, Sanger, WICGR)
LINKS (Xref)	no answer
SRS-EMBL (Availability)	NO
QUERY SEQUENCES	NO

JSNP database

CONTENTS	A database of common gene variations in the Japanese population.
URL	http://snp.ims.u-tokyo.ac.jp/
SPECIES COVERAGE	Human
ENTRIES	195,059 SNPs; 84,566 SNPs with allele frequency
REDUNDANCY	non- redundant
UPDATES	bimonthly
SOURCES	NCBI, Laboratory for Genotyping, the SNP Research Center, the Institute of Physical and Chemical Research (RIKEN), JBIC, dbSNP, UniSTS, UniGene, Model mRNA, RefSeq.
LINKS (Xref)	Search through HOWDY (human organized whole genome database) linked to GenBank, OMIM, dbSNP, GDB. Search by Blast SNP, linked to dbSNP, GenBank, and HGVbase.
SRS-EMBL (Availability)	NO
QUERY SEQUENCES	YES, they use BLAT (BLAST-Like Alignment Tool) search against NCBI build 34 of the human genome.

HAP^{TM} database

CONTENTS	Database reporting both SNPs and gene-based haplotypes, discovered in pharmaceutically relevant genes. Includes information for ~9,000 genes reporting frequency values in each population group. It's a subscription-based database.
URL	http://www.dna.com/products_services/hapdatabase.html
SPECIES COVERAGE	Human
ENTRIES	Aprox. 180,000 unique SNPs. Aprox. 180,000 unique haplotypes.
REDUNDANCY	non-redundant
UPDATES	It varies depending on internal research programs necessities.
SOURCES	Data generated by re-sequencing of 93 individuals from various ethnic backgrounds. (Geinanssance Pharmaceuticals research groups) Data from the two major public databases: dbSNP and HGVbase.
LINKS (Xref)	Public databases for comparison purposes.
SRS-EMBL (Availability)	NO
QUERY SEQUENCES	Could be possible, they use BLAST algorithm.

OMIM

CONTENTS	OMIM, Online Mendelian Inheritance in Man. This database is a catalog of human genes and genetic disorders The database contains textual information and references. It also contains links to MEDLINE an other databases.
URL	http://www3.ncbi.nlm.nih.gov/omim/
SPECIES COVERAGE	Human
ENTRIES	15 325
REDUNDANCY	No
UPDATES	once per month
SOURCES	NCBI
LINKS (Xref)	LSDBs, MEDLINE, Entrez, related resources at NCBI.
SRS-EMBL (Availability)	Yes
QUERY SEQUENCES	NO

Celera Human SNP reference database

CONTENTS	It reports mapped genetic variations to help further the understanding of genetic basis of disease. Thus, the polymorphisms present at Celera database are correlated with genes, gene structure, conserved and regulatory regions, protein changes and disease. It's a subscription-based database.
URL	http://www.celeradiscoverysystem.com/index.cfm
SPECIES COVERAGE	Human
ENTRIES	3.5 million mapped genetic variations
REDUNDANCY	non-redundant
UPDATES	4 times a year.
SOURCES	Celera discovery system, OMIM, TSC, dbSNP, HGVbase, HGMDB
LINKS (Xref)	OMIM, TSC, dbSNP, HGVbase, HGMDB
SRS-EMBL (Availability)	NO
QUERY SEQUENCES	YES, NCBI BLAST version 2,2,5,

Celera Mouse SNP reference database

CONTENTS	It reports over 3 million non-redundantly mapped variations distributed throughout the mouse genome. These variations are correlated with gene structure and protein changes to aid in the identification or validation of potential disease genes. It's a subscription-based database.
URL	http://www.celeradiscoverysystem.com/index.cfm
SPECIES COVERAGE	Mouse strains (129x1/SvJ, DAB/2J, A/J, C57BL/6J, 129s1/SvImJ). Human-mouse synthetic gene regions?
ENTRIES	Over 3,1 million mouse SNPs
REDUNDANCY	non-redundant
UPDATES	every 4 months, corrections in annotations, but new SNPs are registered every 1-2 years
SOURCES	Celera discovery system for the following four strains of mouse: 129s1/svimj, 129x1/svj, a/j and dba/2j. C57bl/6j data, imported from the publicly available sequence.
LINKS (Xref)	They do not cross-reference to other sources.
SRS-EMBL (Availability)	NO
QUERY SEQUENCES	It's possible to BLAST a query sequence and retrieve the corresponding SNP IDs.

Integrated Information Databases

DATABASE	***GeneCards^{TM}***	***HOWDY***
URL	http://bioinfo.weizmann.ac.il/cards/	http://gdb.jst.go.jp/HOWDY/
CONTENTS	GeneCards^{TM} is a database of human genes, their products and their involvement in diseases.	HOWDY is a database system to retrieve human genome information of most of the important public data sources in the world.
SPECIES COVERAGE	Human	Human
SOURCES (SNPs information)	SNPdb	SNPdb, JSNP
LINKS (SNPs repositories)	SNPdb	SNPdb, JSNP

Minor SNP repositories

DATABASE	***URL***	***CONTENTS***	***COVERAGE***
GenesSNP	http://www.genome.utah.edu/genesnps/	Integrates gene, sequence and polymorphism data into individually annotated gene models. The human genes included are related to DNA repair and cell cycle pathways; these genes are though to play a role in susceptibility to environmental exposure.	Human
Leelab SNP database	http://www.bioinformatics.ucla.edu/snp/	Leelab has developed some programs such as, PHRAP, BRO and POA, to identify in coding regions (cSNPs) from publicly available expressed sequence tag (EST) databases. All data has been deposited in dbSNP.	human EST data
rSNP guide	http://util.bionet.nsc.ru/databases/rsnp.html	SNPs in regulatory gene regions onto their interaction with nuclear proteins.	Human
topoSNP	http://gila.bioengr.uic.edu/snp/toposnp/	Visualization of non-synonymous SNPs. Online resource for analyzing nsSNPs that can be mapped onto known 3D structures of proteins.	Human