

Databases

Sequence Databases



**Release/
Updates**

Nucleotide Databases:

EMBL: European Molecular Biology Laboratory

Genbank

DDBJ: DNA Data Bank of Japan

Current Release: ~ 65 Million entries

**International
repository for all
nucleotide
sequences
submitted by
researchers**

Accession numbers are unique to each entry.

**One alphabetical character is followed by five digits, or
two alphabetical characters are followed by six digits.**

Sequence Databases

Nucleotide Databases:

RefSeq: Reference Sequence

NC_123456

Complete Prokaryote Genome

Complete Eukaryote Chromosome

NG_123456

Homo sapiens Genomic Region

A database of non-redundant reference sequences standards, including genomic DNA contigs, mRNAs and proteins for known genes. Contributions are taken from the NCBI and collaborative sequencing efforts

NM_123456

mRNA of several organisms, including *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*

Those accession numbers beginning with X indicate model entries produced as a result of the Genome Annotation process.

Sequence Databases

Protein Databases:

SwissProt + TrEMBL + PIR = UNIPROT

<http://www.uniprot.org>

Current Release: ~2.3 Million entries

UniParc – protein archive database

UniRef – reference clusters

Released as of January 2005
without restriction

Sequence Databases

Protein Databases:

RefSeqP: Reference Sequence Proteins

RefSeqP provides a protein reference standard for the central dogma. It is used, as is RefSeq, to provide a foundation for the functional annotation of the human genome.

Accession numbers for all proteins are of the format: NP_123456

Sequence Databases



Searching for a sequence:

Text Search:

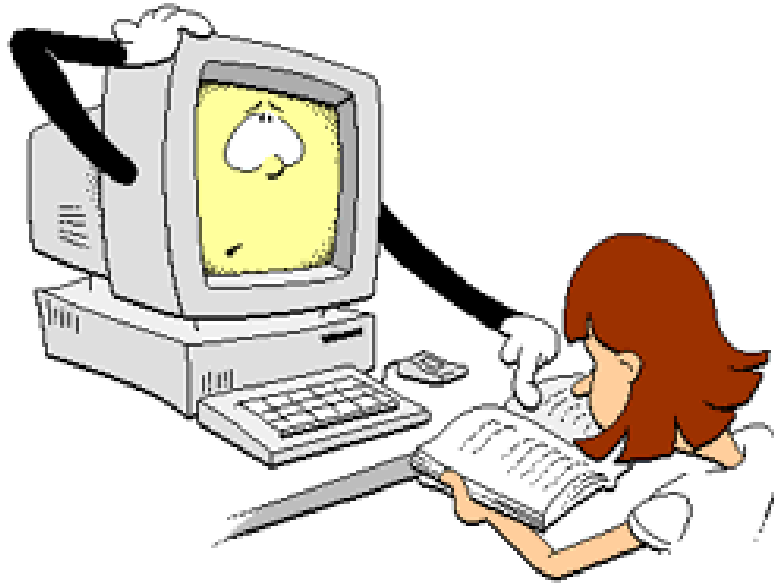
Use text with a boolean operator

BRCA1 & BRCA2 – searches for BRCA1 **AND** BRCA2

BRCA1 | BRCA2 – searches for one gene **OR** the other

BRCA1 ! BRCA2 – searches for BRCA1 **BUT NOT** BRCA2

Computers are **THICK!**



Database entries often presented as **flatfiles**

Each piece of information is on a separate line, distinguished by a code. Computers index this code, so they can search for the relevant entry.

EMBL entry for a sequence fragment implicated in Human Breast Cancer

Identification	ID	AY144588 standard; DNA; HUM; 68 BP.
Accession	AC	AY144588;
Sequence Version	SV	AY144588.1
Date	DT	23-SEP-2002 (Rel. 73, Created)
	DT	23-SEP-2002 (Rel. 73, Last updated, Version 1)
Description	DE	Homo sapiens truncated breast and ovarian cancer susceptibility protein
Keyword	DE	(BRCA1) gene, partial cds.
Organism Source	KW	.
Organism Classification	OS	Homo sapiens (human)
	OC	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia;
	OC	Eutheria; Primates; Catarrhini; Hominidae; Homo.

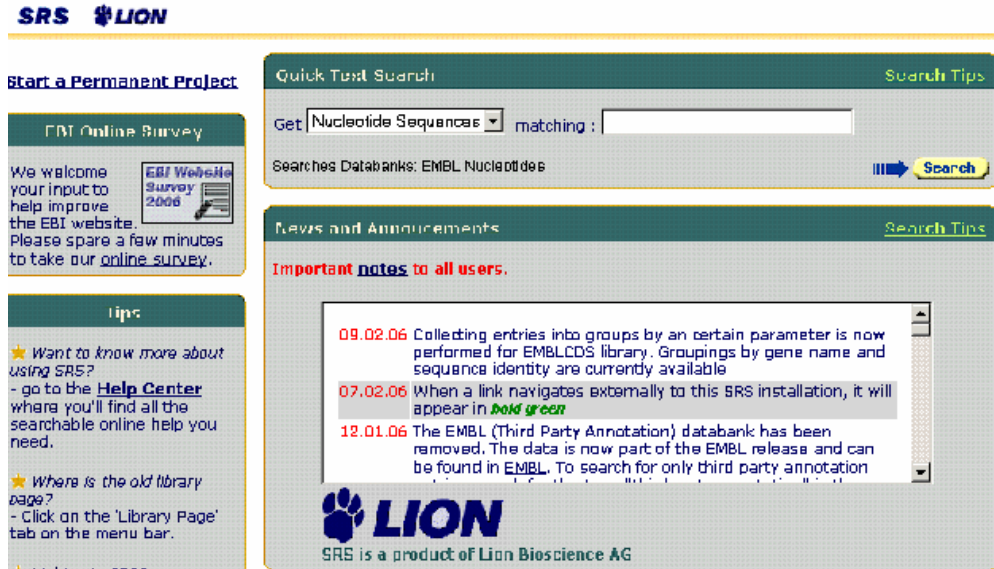
Feature Table
Header

Feature Table
Data

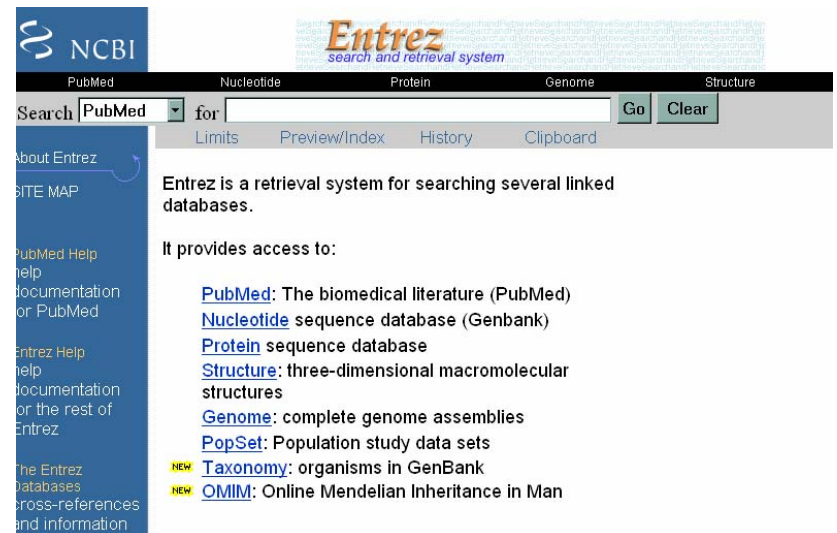
FH	Key	Location/Qualifiers
FH		
FT	source	1..68
FT	/country="	India: South India"
FT	/db_xref="	taxon:9606"
FT	/note="	identical sequence found in daughter with breast
FT	cancer"	
FT	/sex="	female"
FT	/organism="	Homo sapiens"
FT	/isolation_source="	mother with breast cancer"
FT	/dev_stage="	adult"
FT	/mRNA	68
FT	/gene="	BRCA1"
FT	/product="	truncated breast and ovarian cancer
FT	susceptibility protein"	

Searching the databases with a “search engine”:

The Sequence Retrieval System (SRS) from LION Bioscience AG is a very common search tool



The NCBI in the USA has its own search engine called Entrez.



[Reset](#) [Quick Search](#)

Search Options

1. Select the **databanks** you want to search

2. Enter your query in the **Quick Search** or choose from the **Available Databanks**

[Standard](#)
[Extended](#)

You can browse all the **entire databank**. First, select the **databank** you want to click:

Available Databanks

[Expand all](#) [Collapse all](#) Show databanks tooltips:

[Reset](#) search

Search Options

Combine search terms with:

Use wildcards

Get results of type:

Fields you can search | **Your search terms**

In a single field, you can separate multiple values by &, |, !

[Search](#)

SRS 

Result Display Options

View results using:

or

Create a view

Show results per page

Quick Launch

Launch analysis tool:

[Launch](#)

Packages Information

[BLAST](#)
[FASTA](#)
[CLUSTAL](#)
[OTHER](#)
[EMBOSS](#)

Available Analysis Tools - listed by type

[Expand all](#) [Collapse all](#)

- [Alignment Tools](#)
- [Display Tools](#)
- [Edit Tools](#)
- [Information Tools](#)
- [Nucleic Tools](#)
- [Protein Tools](#)
- [Phylogeny Tools](#)
- [Similarity Search Tools](#)

[-] **Mutation and SNP databases**

- all**
- | | | |
|---|---|---|
| <input type="checkbox"/> CFTR | <input type="checkbox"/> PAH | <input type="checkbox"/> HAEMA |
| <input type="checkbox"/> HAEMB | <input type="checkbox"/> LDLR | <input type="checkbox"/> PAX6 |
| <input type="checkbox"/> EMD | <input type="checkbox"/> MUTRES | <input type="checkbox"/> CD40LBASE |
| <input type="checkbox"/> G6PD | <input type="checkbox"/> ANDROGENR | <input type="checkbox"/> OMIMALLELE |
| <input type="checkbox"/> EMBLCHANGE | <input type="checkbox"/> XCGDBASE | <input type="checkbox"/> DMD |
| <input type="checkbox"/> ATM | <input type="checkbox"/> P16 | <input type="checkbox"/> GAA |
| <input type="checkbox"/> IL2RGBASE | <input type="checkbox"/> OTC | <input type="checkbox"/> BIOMDB |
| <input type="checkbox"/> HUMUT | <input type="checkbox"/> HBVARS | <input type="checkbox"/> P53B |
| <input type="checkbox"/> HUMAN_MITBASE | <input type="checkbox"/> P53 | <input type="checkbox"/> APC |
| <input type="checkbox"/> P53LINK | <input type="checkbox"/> RB1 | <input type="checkbox"/> NCL |
| <input type="checkbox"/> ARCGD | <input type="checkbox"/> GD | <input type="checkbox"/> GPI |
| <input type="checkbox"/> BTKBASE | <input type="checkbox"/> VWF | <input type="checkbox"/> HEHP |
| <input type="checkbox"/> HEREDSPHERO | <input type="checkbox"/> PGK | <input type="checkbox"/> PK |
| <input type="checkbox"/> TPI | <input type="checkbox"/> RDS | <input type="checkbox"/> RHODOPSIN |
| <input type="checkbox"/> FANCONI | <input type="checkbox"/> HEXA | <input type="checkbox"/> FVII |
| <input type="checkbox"/> IGF2R | <input type="checkbox"/> OMIMOFFSET | <input type="checkbox"/> HGVBASE |
| <input type="checkbox"/> HGBASE_SUBMITTER | <input type="checkbox"/> HGBASE_HAPLOTYPE | |

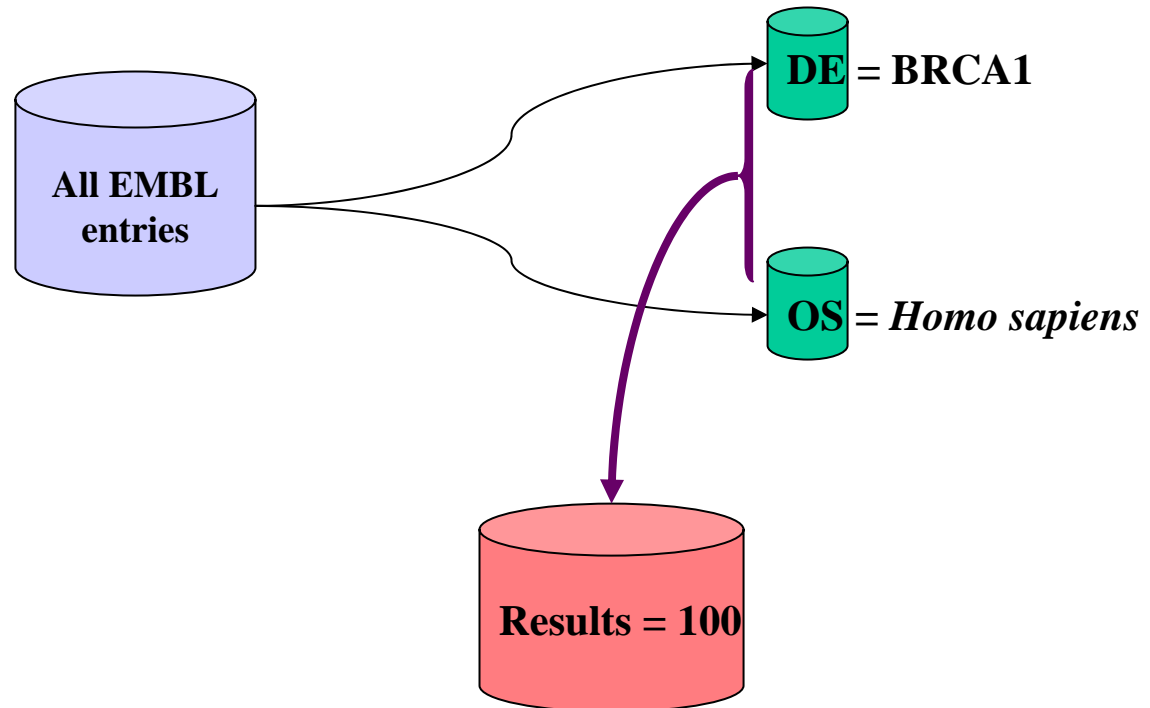
- | | |
|---|--|
| <input type="checkbox"/> P53 | <input type="checkbox"/> APC |
| <input type="checkbox"/> RB1 | |
| <input type="checkbox"/> GD | |
| <input type="checkbox"/> VWF | |
| <input type="checkbox"/> PGK | |
| <input type="checkbox"/> RDS | |
| <input type="checkbox"/> HEXA | |

The IARC Somatic p53 Mutations Database is a compilation of somatic p53 mutations in human tumor cells and cell lines from a systematic search of reports published in the literature

To obtain comprehensive information on this databank, click the link

To search for the BRCA1 gene in Homo sapiens in the EMBL database:

BRCA1 [DE] & Human [OS]



A couple more search tips:

Wildcards

? – specific wildcard replacing one position

e.g. `embl:BRCA?` will allow BRCA1 and BRCA2

* – random wildcard replacing many positions

e.g. `embl:BR*` will allow BRCA1 and BRCA2

Wildcards can be placed generally anywhere in the query,

e.g. `embl:*`; `embl: *805`; `embl: hs?1280`

A couple more search tips:

Spelling

anaemia; colour – UK spelling convention

anemia; color – US spelling convention

plus.....random typos and bad spelling

A couple more search tips:

Small and Capital letters

BRCA1 should be the same as brca1 as far as a text search in the sequence database is concerned.

P51587 should be the same as p51587 as far as an accession number search is concerned.

Not always the case! Sometimes on purpose – more often than not a result of bad programming!

Single Nucleotide Polymorphisms

A single base change occurring in a population at a frequency of $> 1\%$

Gene Variation – physical mapping; functional analysis; association studies; evolutionary studies

Occurrence – average 0.3 – 1 kb apart

base pair frequency of approx. 0.1%

exonic incidence lower

approx. half exonic SNPS non-synonymous

Current maps not exhaustive

dbSNP, Bethesda, USA:

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

The largest database for single nucleotide polymorphisms.
Accession numbers used in dbSNP are not the same as other SNP databases.

ss# - submitted SNPs

rs# - reference SNPs

Clusters of submitted SNPs

NCBI Assay ID	Handle Submitter ID	Validation Status	Orientation /Strand	Alleles
ss2420164	HGBASE SNP000002770		fwd/B	C/T
ss24458917	PERLEGEN afd0523962	<input checked="" type="checkbox"/>	fwd/B	C/T
ss24796350	SEQUENOM sqnm139497	<input checked="" type="checkbox"/>	fwd/B	C/T

Bold type indicates longest cluster

Population Diversity

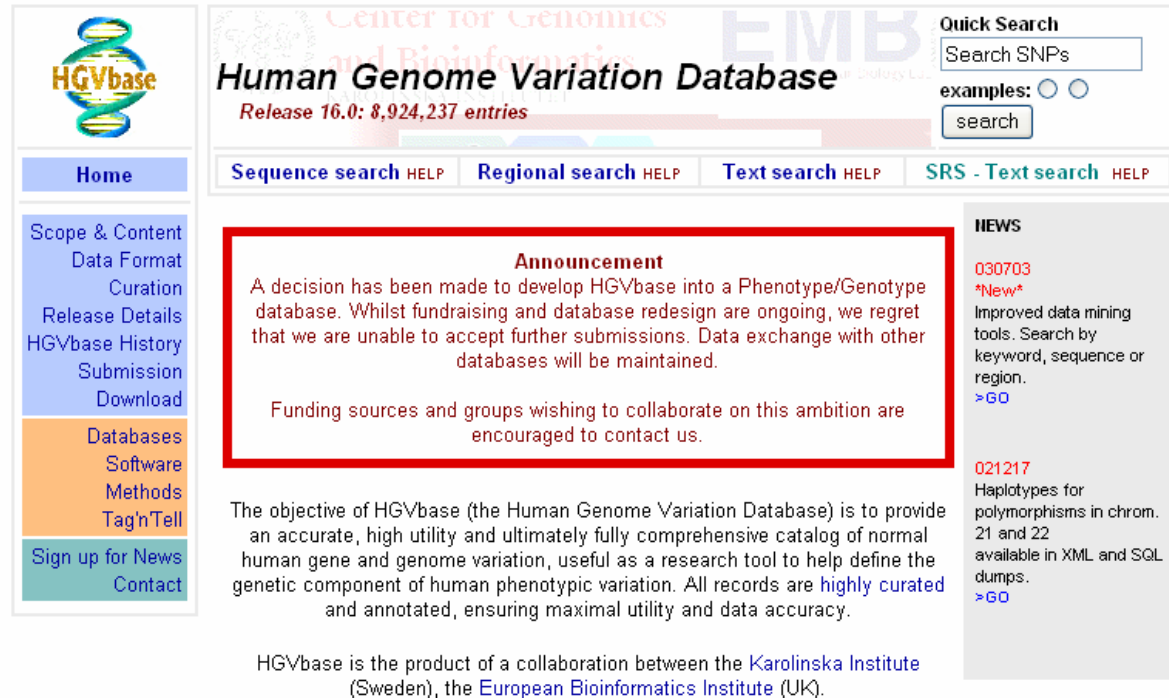
ss#	Sample Ascertainment					Genotypes			Allele	
	Population	Individual Group	Sample (2N)	Founder (N)	Source	C/C	C/T	HWP	C	T
ss24458917	AFD EUR PANEL	European	48	24	IG	0.917	0.083	1.000	0.958	0.042
	AFD AFR PANEL	African American	46	23	IG	1.000			1.000	
	AFD CHN PANEL	Asian	48	24	IG	1.000			1.000	
	HapMap-CEU	European	120	60	IG	0.983	0.017		0.992	0.008
	HapMap-HCB	Asian	90	45	IG	1.000			1.000	
	HapMap-JPT	Asian	88	44	IG	1.000			1.000	
	HapMap-YRI	Sub-Saharan African	120	60	IG	1.000			1.000	
ss24796350	CEPH		184		AF				0.980	0.020

Perlegen Sciences Inc. – Therapeutics and diagnostics company. Analyses genetic variants as part of clinical trials.

HapMap project – Identification and cataloguing of genetic similarities and differences in Human populations

Summarize all known human variations

Genotype-phenotype association analysis



HGVbase

Center for Genomics and Bioinformatics
Human Genome Variation Database
 Release 16.0: 8,924,237 entries

Quick Search
 Search SNPs
 examples:

Sequence search HELP Regional search HELP Text search HELP SRS - Text search HELP

Announcement
 A decision has been made to develop HGVbase into a Phenotype/Genotype database. Whilst fundraising and database redesign are ongoing, we regret that we are unable to accept further submissions. Data exchange with other databases will be maintained.
 Funding sources and groups wishing to collaborate on this ambition are encouraged to contact us.

The objective of HGVbase (the Human Genome Variation Database) is to provide an accurate, high utility and ultimately fully comprehensive catalog of normal human gene and genome variation, useful as a research tool to help define the genetic component of human phenotypic variation. All records are **highly curated** and annotated, ensuring maximal utility and data accuracy.

HGVbase is the product of a collaboration between the [Karolinska Institute](#) (Sweden), the [European Bioinformatics Institute](#) (UK).

NEWS
 030703
 New
 Improved data mining tools. Search by keyword, sequence or region.
[>GO](#)
 021217
 Haplotypes for polymorphisms in chrom. 21 and 22 available in XML and SQL dumps.
[>GO](#)

HGVbase is supported by



<http://www.hgmd.org/>



Welcome to the Human Gene Mutation Database at the Institute of Medical Genetics in Cardiff.

Copyright © Cardiff University 2006. All Rights Reserved.



[HGMD Search](#)

[Statistics](#)

[What's new](#)

[HGMD Background](#)

Gene Symbol	Chromosomal location	Gene name	cDNA sequence	Splice junctions	Mutations
PAX6	11p13	Paired box homeotic gene 6	<input type="button" value="Get cDNA"/>	Splice junctions	Mutations

V. Ball, P.D. Stenson, A.D. Phillips, N.S.T. Thomas.

Mutation type	Total number of mutations	Mutation data
Missense/nonsense	79	<input type="button" value="Get mutations"/>

2006. You can read the latest news on the

Regulatory	Accession Number	Codon change	Amino acid change	Codon number	Phenotype	Reference
Small deletions	CM992377	ATG-AAG	Met-Lys	1	Aniridia	Gronskov (1999) Eur J Hum Genet 7, 274
Small insertions	CM993962	cATG-GTG	Met-Val	1	Aniridia	Wildhardt (1999) PAX6 Locus-specific database Unpublished 0000000167
Small indels	CM981463	AAC-AGC	Asn-Ser	17	Aniridia	Azuma (1998) Invest Ophthalmol Vis Sci 39, 2524
Gross deletions	CM981464	cGGG-TGG	Gly-Trp	18	Cataract, secondary glaucoma	Wolf (1998) Hum Mutat 12, 304
Gross insertions	CM030485	CGG-CCG	Arg-Pro	19	Aniridia	Vincent (2003) Eur J Hum Genet 11, 163
Complex rearrangements	CM941142	cCGG-GGG	Arg-Gly	26	Peters' anomaly	Hanson (1994) Nat Genet 6, 168
	CM993963	ATT-AGT	Ile-Ser	29	Aniridia	Wildhardt (1999) PAX6 Locus-specific database Unpublished 0000000172
	CM981465	gATT-GTT	Ile-Val	29	Aniridia	Azuma (1998) Invest Ophthalmol Vis Sci 39, 2524
	CM991010	aGCT-CCT	Ala-Pro	33	Aniridia	Hanson (1999) Hum Mol Genet 8, 165
	CM983675	TGCg-TGA	Cys-Term	40	Aniridia	Vincent (2003) Eur J Hum Genet 11, 163
	CM992378	ATT-AGT	Ile-Ser	42	Aniridia	Gronskov (1999) Eur J Hum Genet 7, 274
	CM991011	tTCC-CCC	Ser-Pro	43	Aniridia	Hanson (1999) Hum Mol Genet 8, 165
	CM971138	cCGA-TGA	Arg-Term	44	Aniridia	Neuner-Jehle (1997) Hum Mutat 12, 138

<http://snp.cshl.org>



[Home](#) :: [Frequency/Genotype](#) :: [Linkage Maps](#) :: [Protocols](#)
[Search](#) :: [News](#) :: [About](#) :: [Help](#) :: [Download data](#) :: [Feedback](#)

Single Nucleotide Polymorphisms for Biomedical Research

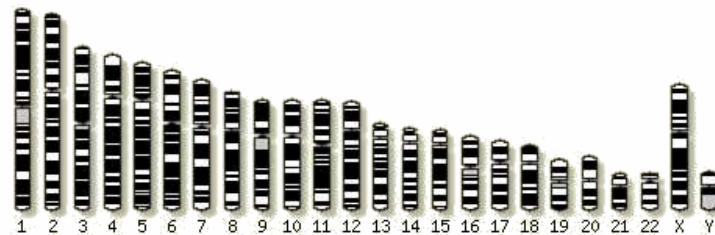


Image courtesy of <http://www.ensembl.org>, slightly modified from the original

Search: ([advanced](#) | [help](#))

Single nucleotide polymorphisms (SNPs) are common DNA sequence variations among individuals. They promise to significantly advance our ability to understand and treat human disease. The SNP Consortium (TSC) is a public/private collaboration that has to date discovered and characterized nearly 1.8 million SNPs ([more](#))

<http://human.genelynx.org>

Gene Lynx

A portal to mammalian genomes

Human 32 696 records
 Mouse 54 728 records
 Rat 5 255 records

GeneLynx guide

GeneLynx info

Genomic resources

Genomic sequences	NCBI EBI DBS Z83307.1	NCBI EBI DBS S70304.1	NCBI EBI DBS S70305.1
	NCBI EBI DBS S70307.1		
GDB	118997		
GenAtlas	PAX6		
Ensembl gene	ENSG00000007372		
UCSC Genome Browser	chr11:31775792-31797188		
Vista Genome Browser	Align sequence to [Rat Jun03] [Rat Jan03] [Mouse Feb03] [Multiple: Mouse Feb03 & Rat Jun03]		
AceView	PAX6		
dbSNP	rs1800427 rs3026375 rs3026367 rs3026370	rs3026383 rs5790867 rs3026368	rs3026371 rs694617 rs3026369

Transcripts

<http://www.ebi.ac.uk/asd/>

Alternative Splicing Database Project

Search ASD
Search: ALL ASD products

QUERY SUMMARY:
Dataset : Database(s): ALL ASD

Exon Isoform Event

EXON ISOFORMS	EVENT TYPE	AEDB ASSOCIATION	CONSERVATION
8126..8240 8126..8249	Alternative donor		

ALTSPLICE Human 26803..270

ID	SNP
ENSG00000009830	POI 8241..1211
	21824..268
	40276..442
ENSG00000174371	EXON HE
	HE
	MUT
	7060..7179

Chr. 1
Length 47.51 Kb

AltSplice SNPs

AltSplice patterns

Ensembl trans.

DNA(contigs)