SNP Comparison

Group Members

Amira

Jhelum

Rahul

Shweta

Chimpanzee Or Mouse

- The earlier goal of the project was to compare SNP distribution over certain genes in human and chimpanzee
- But due to unavailability of sufficient amount of data on chimpanzee, we had to change our focus to mouse

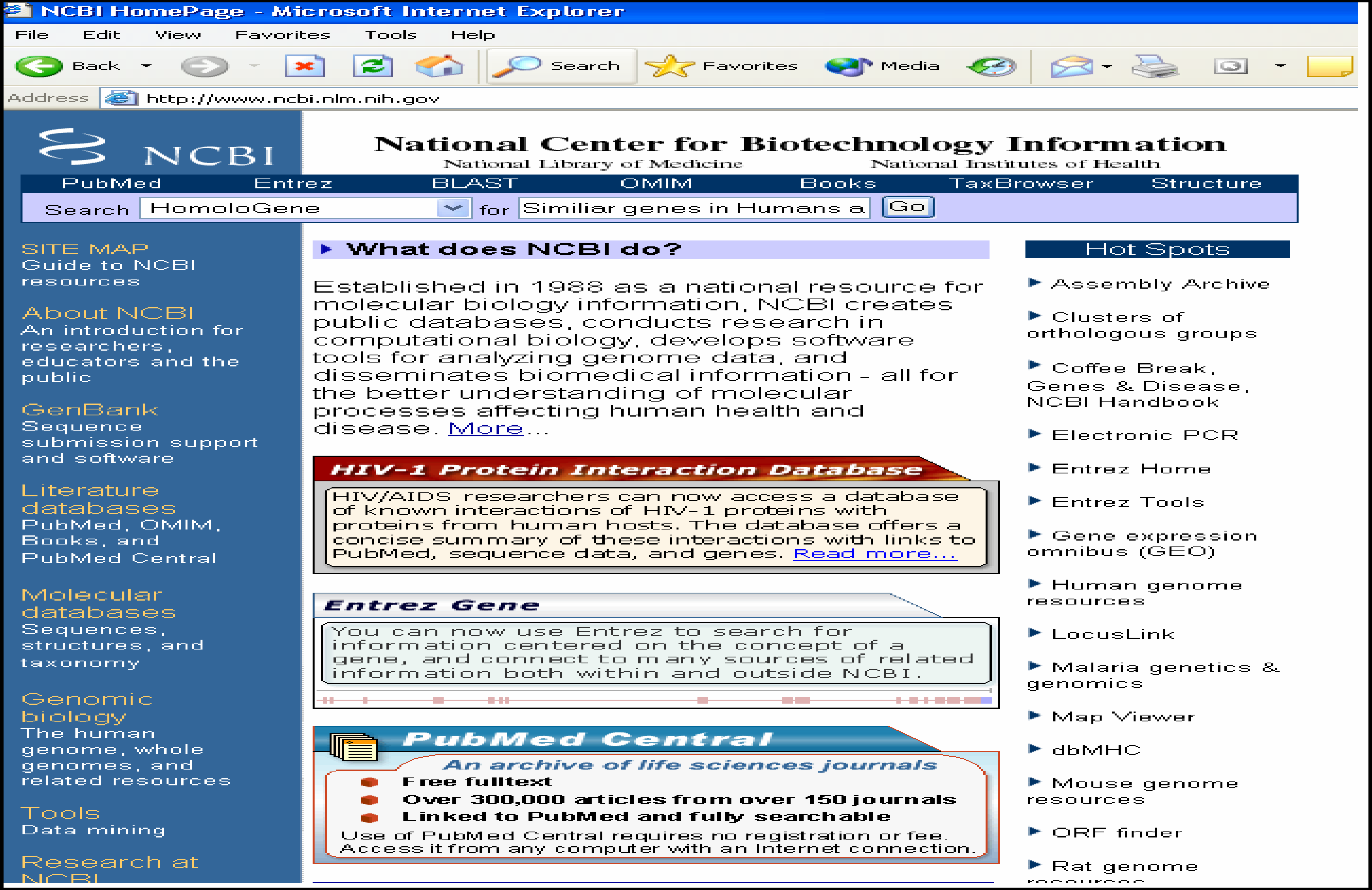
Reasons for choosing Mouse

- Mouse Genome and SNP data is readily available on the NCBI website
- Mouse is the most important animal model and is widely used in the study of human diseases.
- Mouse carries virtually the same set of genes as the human and more then 90% of the mouse genome can be lined up with a region on the human genome.

Accomplishments Till Date

- Homology searching for Human and Mouse Genes on the NCBI website.
- List of potential 484 genes
- Analyzed these homologous genes for SNIPS
  - Total number of SNIPS
  - SNIPS in the coding region
- Tabulated data
  - Unfiltered data
  - Filtered data

NCBI Website



National Center for Biotechnology Information

National Library of Medicine National Institutes of Health

SITE MAP
Guide to NCBI resources

About NCBI
An introduction for researchers, educators and the public

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases
Sequences, structures, and taxonomy

Genomic biology
The human genome, whole genomes, and related resources

Tools
Data mining

Research at NCBI

▶ What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

HIV-1 Protein Interaction Database

HIV/AIDS researchers can now access a database of known interactions of HIV-1 proteins with proteins from human hosts. The database offers a concise summary of these interactions with links to PubMed, sequence data, and genes. Read more...

Entrez Gene

You can now use Entrez to search for information centered on the concept of a gene, and connect to many sources of related information both within and outside NCBI.

PubMed Central

An archive of life sciences journals

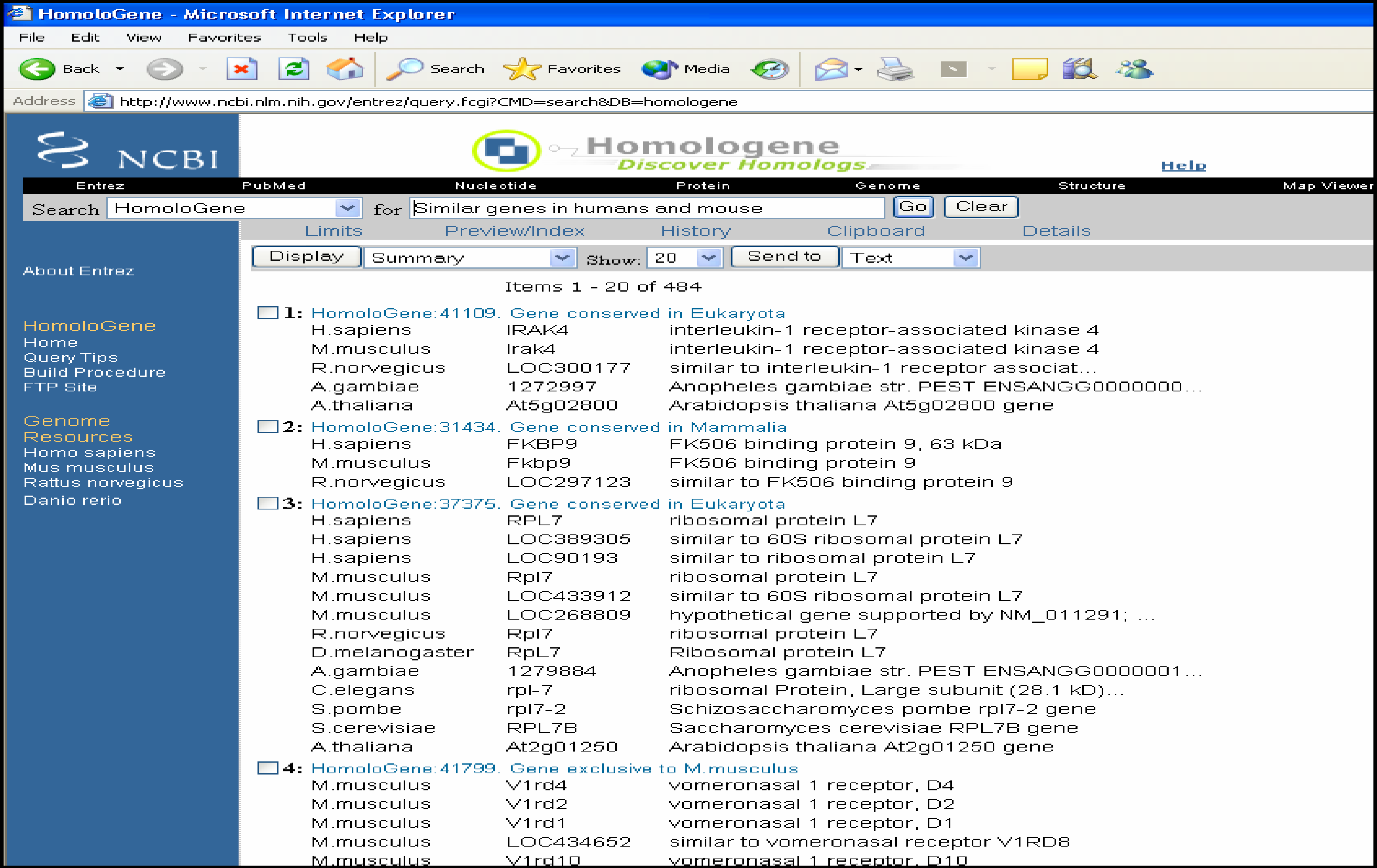
- **Free fulltext**
- **Over 300,000 articles from over 150 journals**
- **Linked to PubMed and fully searchable**

Use of PubMed Central requires no registration or fee. Access it from any computer with an Internet connection.

Hot Spots

- Assembly Archive
- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools
- Gene expression omnibus (GEO)
- Human genome resources
- LocusLink
- Malaria genetics & genomics
- Map Viewer
- dbMHC
- Mouse genome resources
- ORF finder
- Rat genome resources

List of Homologenes



Address: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=homologene

Search: HomoloGene for: Similar genes in humans and mouse

Display: Summary Show: 20 Send to: Text

Items 1 - 20 of 484

□ **1:** HomoloGene:41109. Gene conserved in Eukaryota

H.sapiens	IRAK4	interleukin-1 receptor-associated kinase 4
M.musculus	Irak4	interleukin-1 receptor-associated kinase 4
R.norvegicus	LOC300177	similar to interleukin-1 receptor associat...
A.gambiae	1272997	Anopheles gambiae str. PEST ENSANGG0000000...
A.thaliana	At5g02800	Arabidopsis thaliana At5g02800 gene

□ **2:** HomoloGene:31434. Gene conserved in Mammalia

H.sapiens	FKBP9	FK506 binding protein 9, 63 kDa
M.musculus	Fkbp9	FK506 binding protein 9
R.norvegicus	LOC297123	similar to FK506 binding protein 9

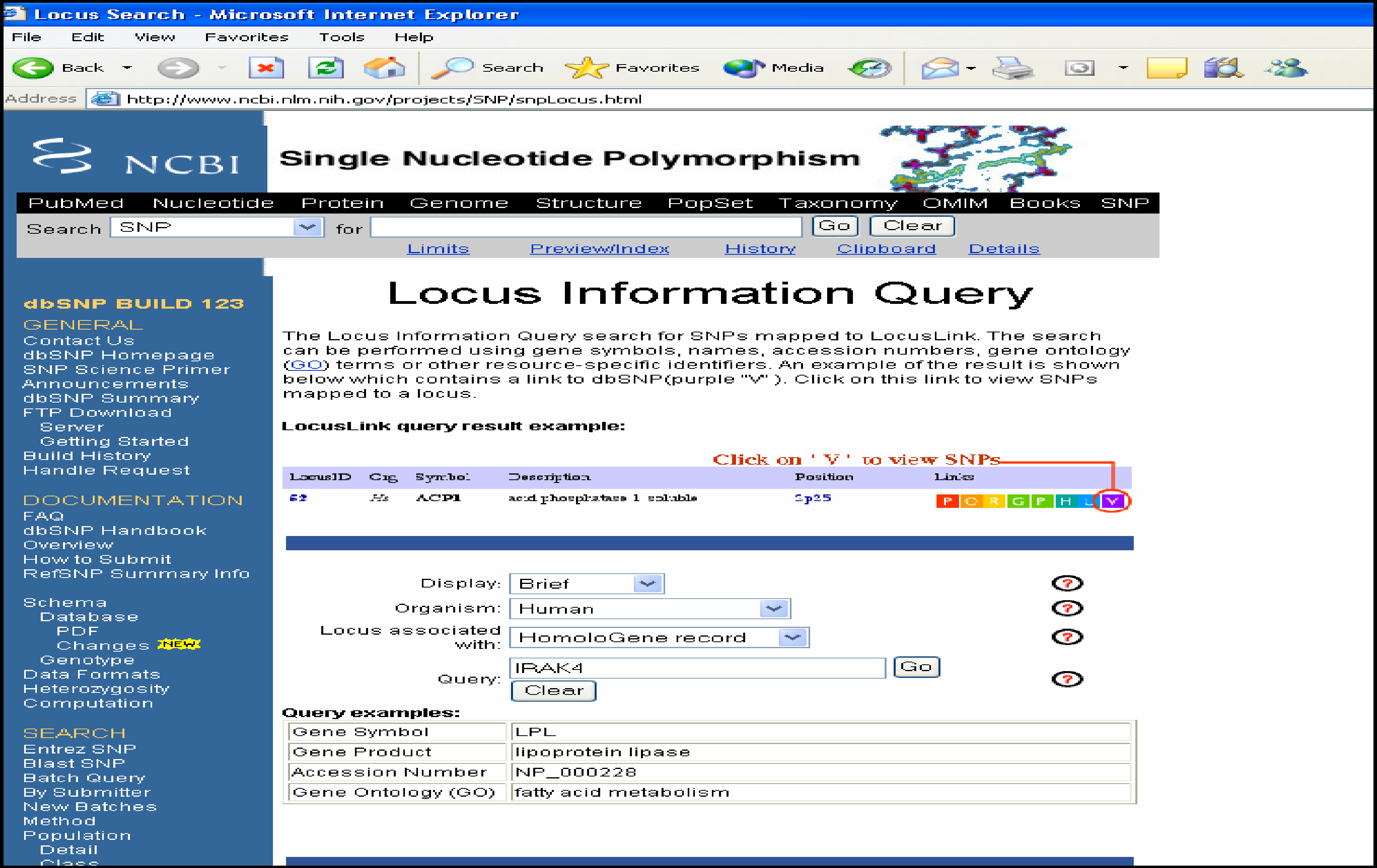
□ **3:** HomoloGene:37375. Gene conserved in Eukaryota

H.sapiens	RPL7	ribosomal protein L7
H.sapiens	LOC389305	similar to 60S ribosomal protein L7
H.sapiens	LOC90193	similar to ribosomal protein L7
M.musculus	Rpl7	ribosomal protein L7
M.musculus	LOC433912	similar to 60S ribosomal protein L7
M.musculus	LOC268809	hypothetical gene supported by NM_011291; ...
R.norvegicus	Rpl7	ribosomal protein L7
D.melanogaster	RpL7	Ribosomal protein L7
A.gambiae	1279884	Anopheles gambiae str. PEST ENSANGG0000001...
C.elegans	rpl-7	ribosomal Protein, Large subunit (28.1 kD)...
S.pombe	rpl7-2	Schizosaccharomyces pombe rpl7-2 gene
S.cerevisiae	RPL7B	Saccharomyces cerevisiae RPL7B gene
A.thaliana	At2g01250	Arabidopsis thaliana At2g01250 gene

□ **4:** HomoloGene:41799. Gene exclusive to M.musculus

M.musculus	V1rd4	vomeronasal 1 receptor, D4
M.musculus	V1rd2	vomeronasal 1 receptor, D2
M.musculus	V1rd1	vomeronasal 1 receptor, D1
M.musculus	LOC434652	similar to vomeronasal receptor V1RD8
M.musculus	V1rd10	vomeronasal 1 receptor, D10

Searching for SNIPS



Locus Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
NCBI
PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP
Search
for
Go
Clear
Limits Preview/Index History Clipboard Details

http://www.ncbi.nlm.nih.gov/projects/SNP/snpLocus.html

Single Nucleotide Polymorphism

Search: SNP

dbSNP BUILD 123

GENERAL
Contact Us
dbSNP Homepage
SNP Science Primer
Announcements
dbSNP Summary
FTP Download
Server
Getting Started
Build History
Handle Request

DOCUMENTATION
FAQ
dbSNP Handbook
Overview
How to Submit
RefSNP Summary Info

Schema
Database
PDF
Changes NEW
Genotype
Data Formats
Heterozygosity
Computation

SEARCH
Entrez SNP
Blast SNP
Batch Query
By Submitter
New Batches
Method
Population
Detail
Class

Locus Information Query

The Locus Information Query search for SNPs mapped to LocusLink. The search can be performed using gene symbols, names, accession numbers, gene ontology (GO) terms or other resource-specific identifiers. An example of the result is shown below which contains a link to dbSNP(purple "V"). Click on this link to view SNPs mapped to a locus.

LocusLink query result example:

Click on ' V ' to view SNPs

LocusID	Org	Symbol	Description	Position	Links
52	Hs	ACP1	acid phosphatase 1, soluble	2p25	P C R G P H L V

Display: Brief

Organism: Human

Locus associated with: HomoloGene record

Query: IRAK4

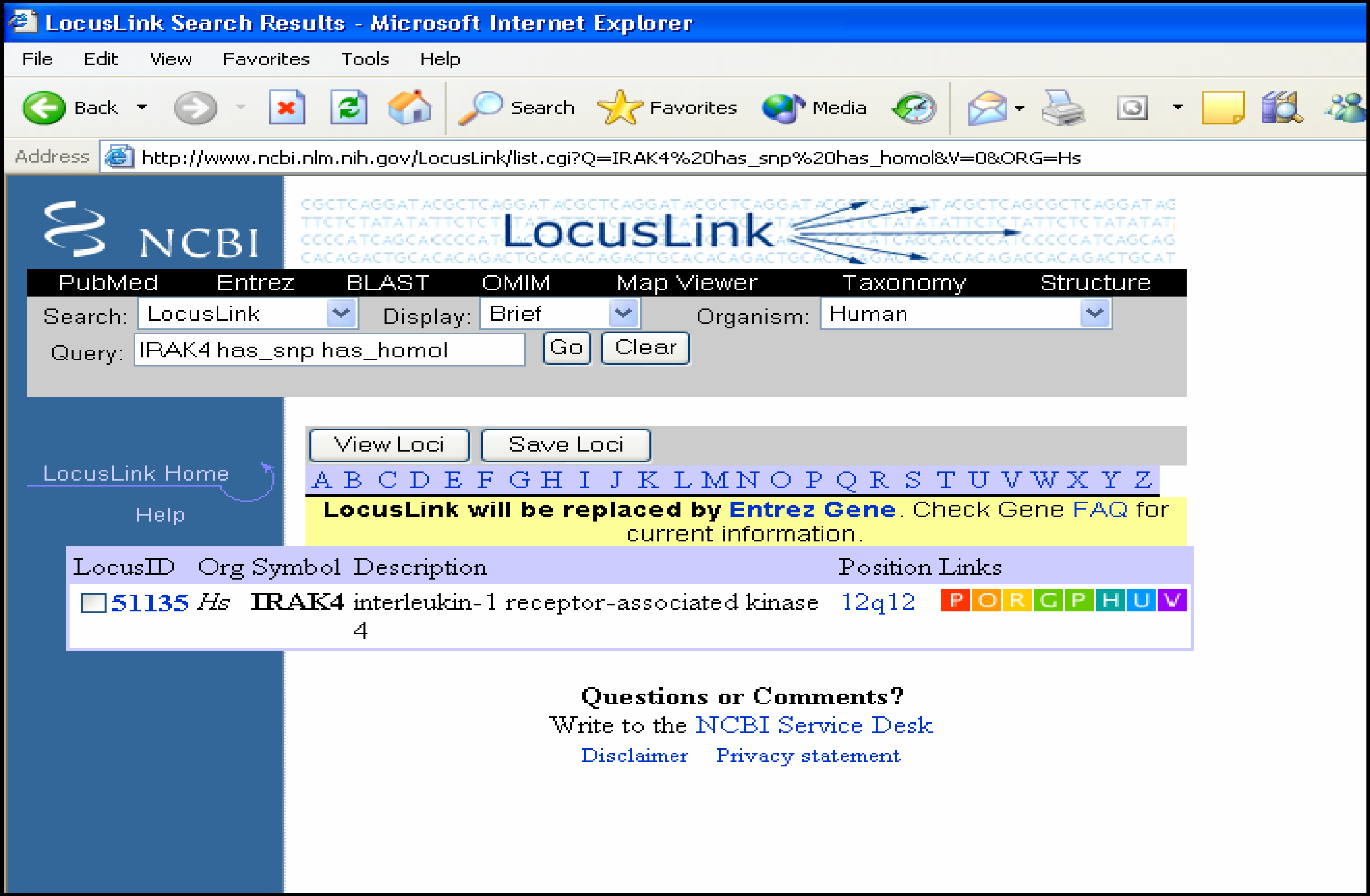
Go

Clear

Query examples:

Gene Symbol	LPL
Gene Product	lipoprotein lipase
Accession Number	NP_000228
Gene Ontology (GO)	fatty acid metabolism

Brief Description of IRAK4 Gene



LocusLink Search Results - Microsoft Internet Explorer

Address: http://www.ncbi.nlm.nih.gov/LocusLink/list.cgi?Q=IRAK4%20has_snp%20has_homol&V=0&ORG=Hs

Search: LocusLink Display: Brief Organism: Human

Query: IRAK4 has_snp has_homol

LocusLink will be replaced by Entrez Gene. Check Gene FAQ for current information.

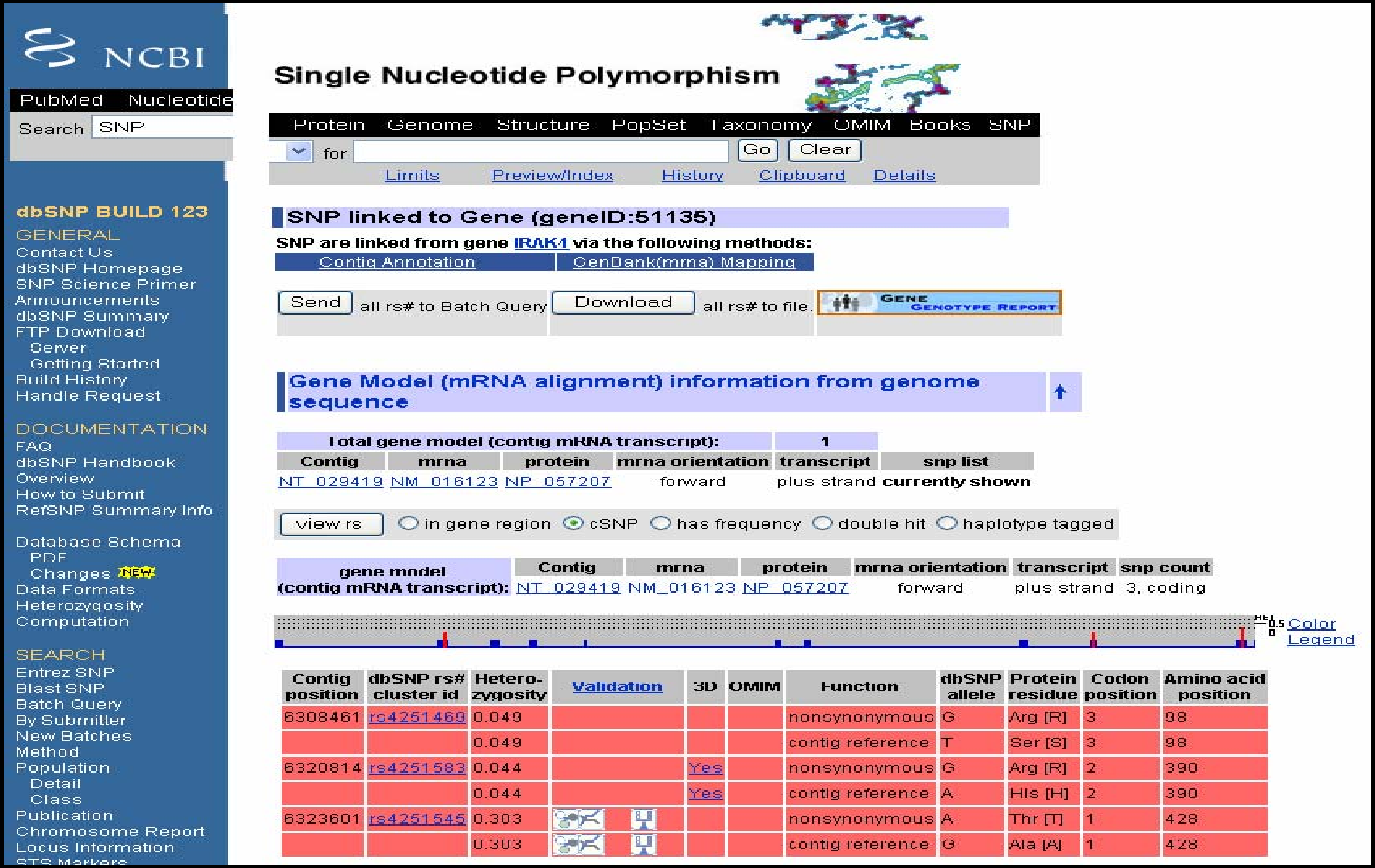
LocusID	Org	Symbol	Description	Position	Links
□ 51135	*Hs*	IRAK4	interleukin-1 receptor-associated kinase 4	12q12	P O R G P H U V

Questions or Comments?

Write to the NCBI Service Desk

Disclaimer Privacy statement

SNIPS for IRAK4 Gene



SNP linked to Gene (geneID:51135)

SNP are linked from gene IRAK4 via the following methods:

Contig Annotation | GenBank(mrna) Mapping

Send all rs# to Batch Query | Download all rs# to file. | GENE GENOTYPE REPORT

Gene Model (mRNA alignment) information from genome sequence

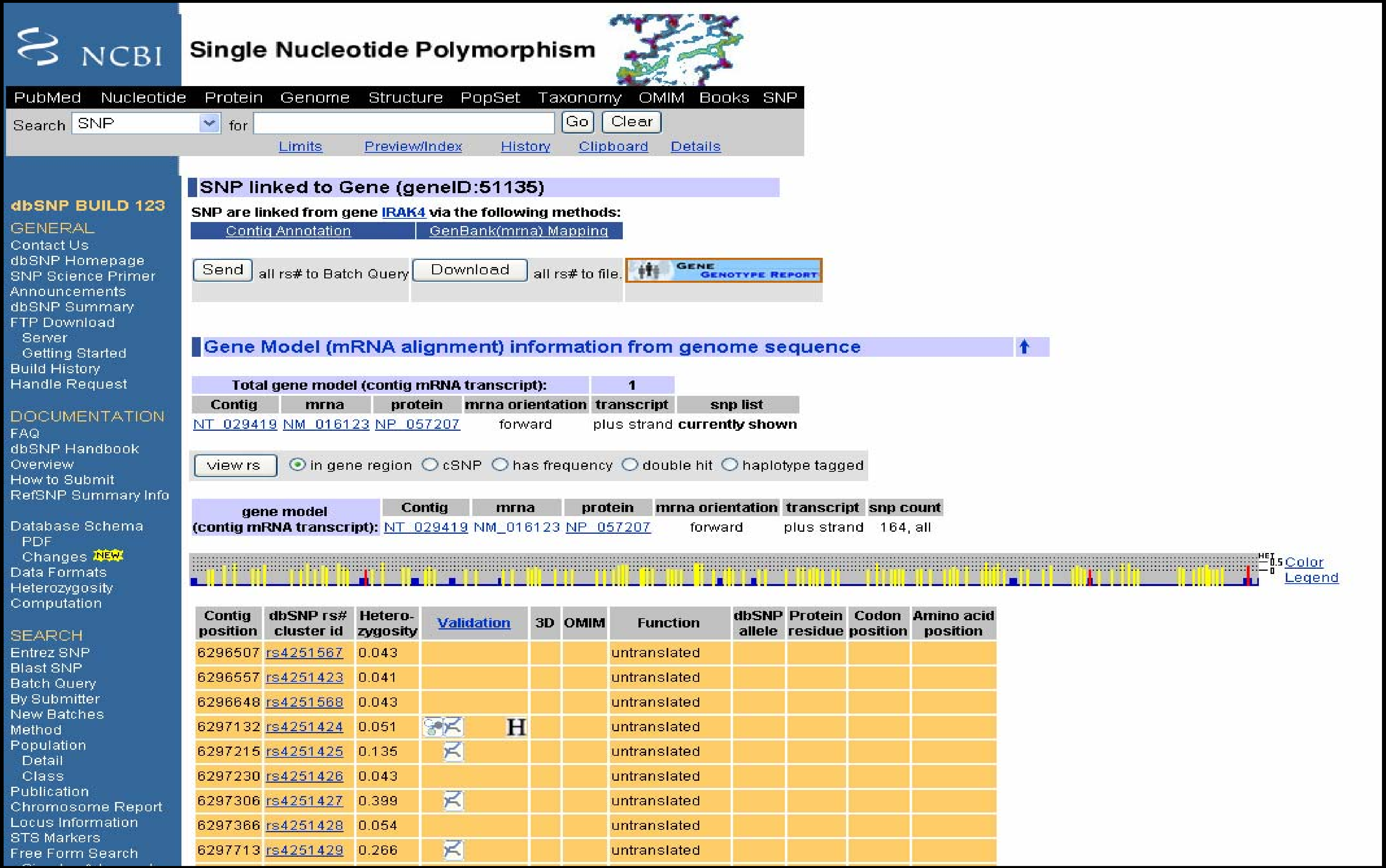
Total gene model (contig mRNA transcript):				1	
Contig	mrna	protein	mrna orientation	transcript	snp list
NT_029419	NM_016123	NP_057207	forward	plus strand	currently shown

view rs ○ in gene region ● cSNP ○ has frequency ○ double hit ○ haplotype tagged

gene model	Contig	mrna	protein	mrna orientation	transcript	snp count
(contig mRNA transcript):	NT_029419	NM_016123	NP_057207	forward	plus strand	3, coding

Contig position	dbSNP rs# cluster id	Hetero-zygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
6308461	rs4251469	0.049				nonsynonymous	G	Arg [R]	3	98
		0.049				contig reference	T	Ser [S]	3	98
6320814	rs4251583	0.044		Yes		nonsynonymous	G	Arg [R]	2	390
		0.044		Yes		contig reference	A	His [H]	2	390
6323601	rs4251545	0.303				nonsynonymous	A	Thr [T]	1	428
		0.303				contig reference	G	Ala [A]	1	428

Total SNIPS for IRAK4 Gene



SNP linked to Gene (geneID:51135)

SNP are linked from gene IRAK4 via the following methods:

Contig Annotation | GenBank(mrna) Mapping

Send all rs# to Batch Query | Download all rs# to file. | GENE GENOTYPE REPORT

Gene Model (mRNA alignment) information from genome sequence

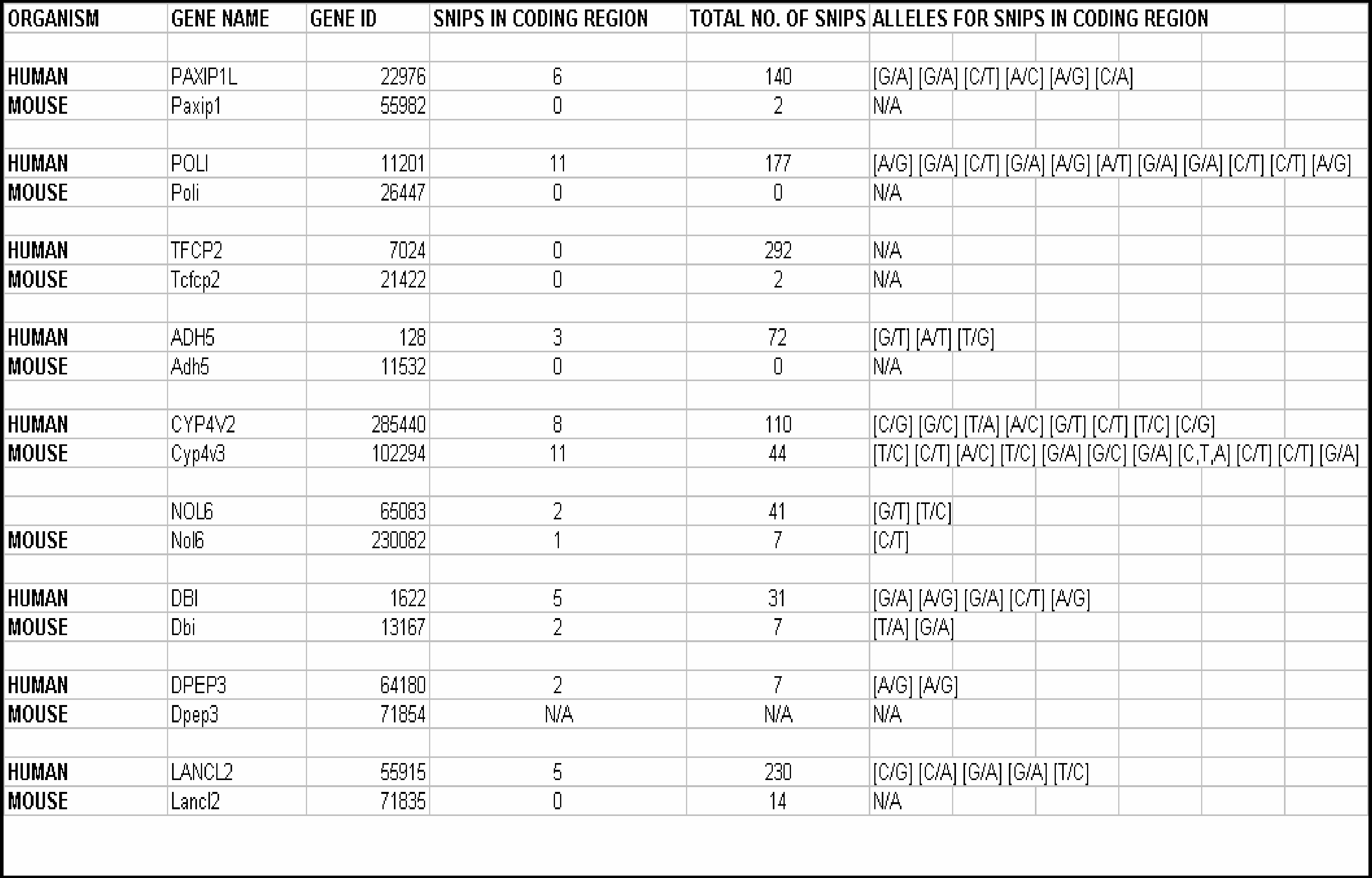
Total gene model (contig mRNA transcript):				1	
Contig	mrna	protein	mrna orientation	transcript	snp list
NT_029419	NM_016123	NP_057207	forward	plus strand	**currently shown**

view rs ● in gene region ○ cSNP ○ has frequency ○ double hit ○ haplotype tagged

gene model (contig mRNA transcript):	Contig	mrna	protein	mrna orientation	transcript	snp count
	NT_029419	NM_016123	NP_057207	forward	plus strand	164, all

Contig position	dbSNP rs# cluster id	Hetero-zygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
6296507	rs4251567	0.043				untranslated				
6296557	rs4251423	0.041				untranslated				
6296648	rs4251568	0.043				untranslated				
6297132	rs4251424	0.051	H			untranslated				
6297215	rs4251425	0.135				untranslated				
6297230	rs4251426	0.043				untranslated				
6297306	rs4251427	0.399				untranslated				
6297366	rs4251428	0.054				untranslated				
6297713	rs4251429	0.266				untranslated				

Tabulation Of the Data

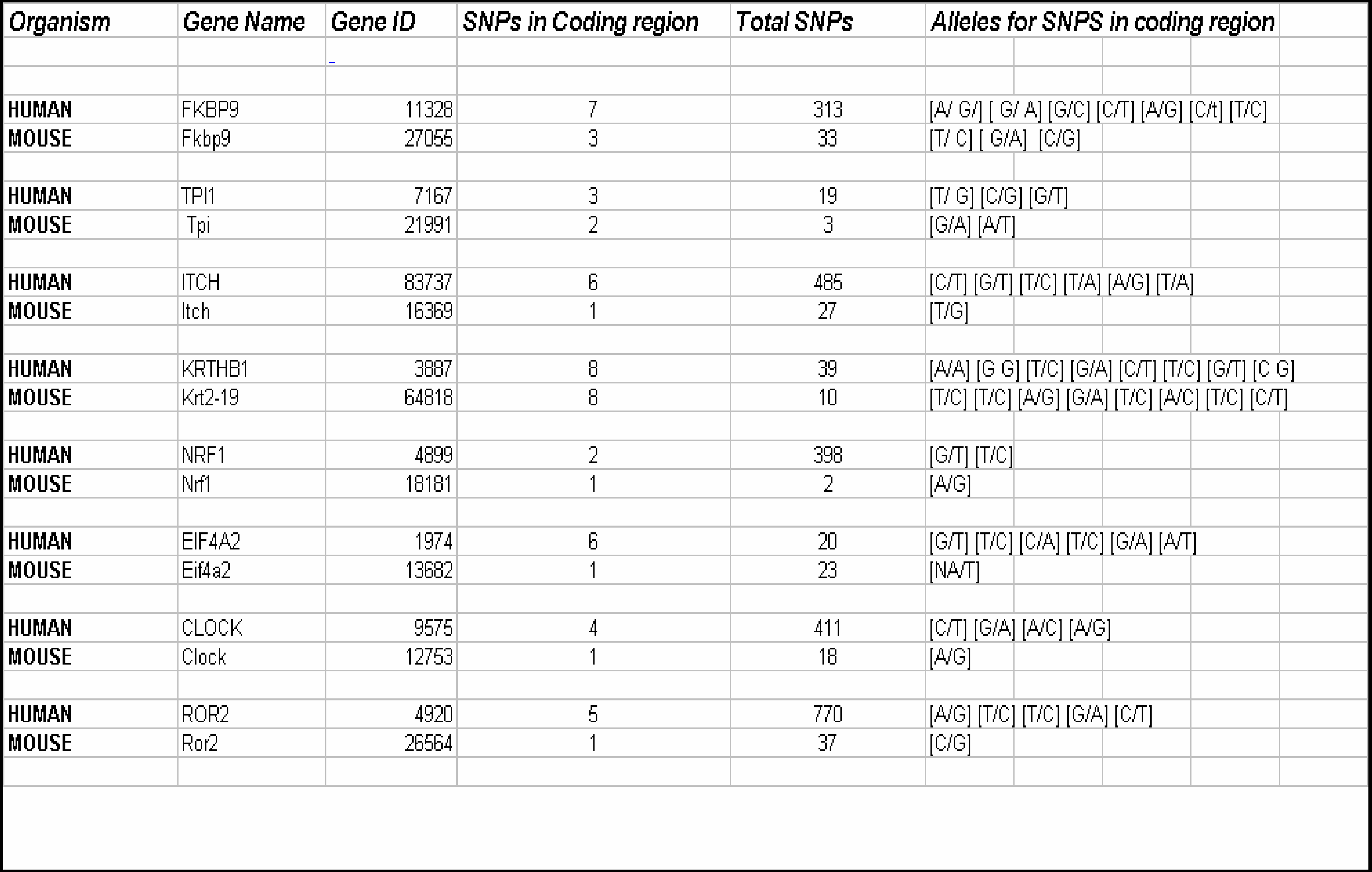


ORGANISM	GENE NAME	GENE ID	SNIPS IN CODING REGION	TOTAL NO. OF SNIPS	ALLELES FOR SNIPS IN CODING REGION
HUMAN	PAXIP1L	22976	6	140	[G/A] [G/A] [C/T] [A/C] [A/G] [C/A]
MOUSE	Paxip1	55982	0	2	N/A
HUMAN	POLI	11201	11	177	[A/G] [G/A] [C/T] [G/A] [A/G] [A/T] [G/A] [G/A] [C/T] [C/T] [A/G]
MOUSE	Poli	26447	0	0	N/A
HUMAN	TFCP2	7024	0	292	N/A
MOUSE	Tcfcp2	21422	0	2	N/A
HUMAN	ADH5	128	3	72	[G/T] [A/T] [T/G]
MOUSE	Adh5	11532	0	0	N/A
HUMAN	CYP4V2	285440	8	110	[C/G] [G/C] [T/A] [A/C] [G/T] [C/T] [T/C] [C/G]
MOUSE	Cyp4v3	102294	11	44	[T/C] [C/T] [A/C] [T/C] [G/A] [G/C] [G/A] [C,T,A] [C/T] [C/T] [G/A]
	NOL6	65083	2	41	[G/T] [T/C]
MOUSE	Nol6	230082	1	7	[C/T]
HUMAN	DBI	1622	5	31	[G/A] [A/G] [G/A] [C/T] [A/G]
MOUSE	Dbi	13167	2	7	[T/A] [G/A]
HUMAN	DPEP3	64180	2	7	[A/G] [A/G]
MOUSE	Dpep3	71854	N/A	N/A	N/A
HUMAN	LANCL2	55915	5	230	[C/G] [C/A] [G/A] [G/A] [T/C]
MOUSE	Lancl2	71835	0	14	N/A

Data Analysis

- Concentrated our study on genes which have SNIPS in their coding region for both mouse and human.
- After first round of analysis, we reduced our data from 484 genes to 82 genes based on number of SNIPS in the coding region.

Filtered Data



Organism	*Gene Name*	*Gene ID*	*SNPs in Coding region*	*Total SNPs*	*Alleles for SNPS in coding region*
HUMAN	FKBP9	11328	7	313	[A/ G/] [G/ A] [G/C] [C/T] [A/G] [C/t] [T/C]
MOUSE	Fkbp9	27055	3	33	[T/ C] [G/A] [C/G]
HUMAN	TPI1	7167	3	19	[T/ G] [C/G] [G/T]
MOUSE	Tpi	21991	2	3	[G/A] [A/T]
HUMAN	ITCH	83737	6	485	[C/T] [G/T] [T/C] [T/A] [A/G] [T/A]
MOUSE	Itch	16369	1	27	[T/G]
HUMAN	KRTHB1	3887	8	39	[A/A] [G G] [T/C] [G/A] [C/T] [T/C] [G/T] [C G]
MOUSE	Krt2-19	64818	8	10	[T/C] [T/C] [A/G] [G/A] [T/C] [A/C] [T/C] [C/T]
HUMAN	NRF1	4899	2	398	[G/T] [T/C]
MOUSE	Nrf1	18181	1	2	[A/G]
HUMAN	EIF4A2	1974	6	20	[G/T] [T/C] [C/A] [T/C] [G/A] [A/T]
MOUSE	Eif4a2	13682	1	23	[NA/T]
HUMAN	CLOCK	9575	4	411	[C/T] [G/A] [A/C] [A/G]
MOUSE	Clock	12753	1	18	[A/G]
HUMAN	ROR2	4920	5	770	[A/G] [T/C] [T/C] [G/A] [C/T]
MOUSE	Ror2	26564	1	37	[C/G]

Interim Results

- Consolidated data from Human and Mouse
  - Based on Homologous genes between the two species.
  - Distribution of SNIPS on the homologenes.
- Analyzed the data to select the genes with SNIPS in the coding region for both the species.

Next Steps

- Analyze the selected 82 genes and draw statistical conclusions of biological significance from the above data.

- Further filter the data
  - To study the distribution of SNIPS on potential genes for both the species using a parser.

- The complete data for all 484 genes and the selected 82 genes is available on our website
  - http://www.angelfire.com/sk3/compbio601/