SINGLE NUCLEOTIDE POLYMORPHISM

SNP

COMPARISON OF HUMAN AND MOUSE

Fall 2004 - Foundations of Computational Biology 601

December 6th, 2004

Professor: Michael Recce

Group Members:

Amira Elserougy

Shweta Bhargava

Rahul Patil

Jhelum Naik

Contents

- **Data – Attached at the end of the report – By: All Group Members**
  - SNPs in 484 genes
  - Filtered Data – SNPs in 82 genes
  - Parser Code
  - Filtered Data with Gene Lengths

- **Presentations**
  1) **Presentation 1 – "Introduction to the topic"- By: Amira Elsrougy**
  2) **Presentation 2 – "Review of three reference papers"- By: Jhelum Naik**
  3) **Presentation 3 – "Initial Results"- By: Shweta Bhargava**
  4) **Presentation 4 – "Final Results & Conclusion"- By: Rahul Patil**

- **Website – www.angelfire.com/sk3/compbio601 - By: Amira Elserougy**
  Project website includes:
  1) **3 reference papers**
  2) **additional references**
  3) **articles**
  4) **4 presentations – in powerpoint, and PDF file**
  5) **data and code**
  6) **final report in PDF file**

Abstract

The goal of this project is to compare the distribution of SNPs over the 484 genes of humans and mice. The data of the human and the mouse SNPs will be obtained from the public SNP database (dbSNP) homepage. The genes that were used are the homologous genes between human and mouse. Out of the 484 genes, 82 genes were filtered out and those were the ones that had SNPs in their coding regions. The gene lengths and the mRNA size of each for human and mouse are also obtained.
Then the data was analyzed in regards to the average SNP density, and the density of the exon and intron regions. The distribution of SNPs was calculated according to the percentage of SNPs in the intron and exon regions. The distribution of alleles in the human and mouse data were also obtained. Next, a test for SNP conservation in Human and mouse was carried out.

Web resources and tools were used for exploring the human and mouse genome variations. Statistical analysis and Aligning methods (using LAlign software) were applied to the data to gather an amount of conclusions that can be useful in future work. The conclusions were represented in graphs in charts as seen in the results section. The comparison of the SNP patterns in both organisms would create a major impact on the level of understanding of human disease, human population genetics, and human evolution. It is predicted that eventually, doctors will be able to have individual SNP profiles of their patients that will help them to organize patients into groups and find correlations between certain SNP profiles that will help with providing patients with more individualized drug therapy.

Introduction

Single Nucleotide Polymorphisms (also known as SNPs, pronounced as ‘snips’) are defined to be genetic variations that occur within a specific DNA sequence.
A genetic variation is considered to be the situation where a single nucleotide (Adenine, Thymine, Cytosine, or Guanine) replaces one of the other three nucleotides. SNPs make up an approximation of 90% of all observed human variations. They occur every 100 to 300 base pairs on the human genome. Variations can be distinguished to be SNPs and not mutations if they occur in at least 1% of the population.
The following Table summarizes the differences between SNPs and mutations

SNP	**MUTATION**
1 Base Pair Change	1 + Base Pair Changes (deletions)
Occur in nature	Occur naturally or in the laboratory
Present in populations > 1%	Present in populations < 1%
No necessary phenotype manifestation	Phenotype manifestation
Affects any region of the genome	Affects genes
Germline inheritance	Germline or somatic inheritance

Single Nucleotide Polymorphisms occur in both coding regions and non-coding regions of the human genome. Since only 3 – 5% of the DNA sequences codes for proteins, most SNPs are located outside of the coding regions. Many of the SNPs present in the human genome have no function or effect on the cell, but researchers have observed that they play a role in predisposing people to disease and influence people’s

response to certain drugs. Researches have been mostly interested in the SNPs found in coding regions because they are most likely to be involved with the biological functions of proteins. Approximately, 50% of all SNPs in exons are considered to be biologically silent meaning they do not have any effect on the function of the gene or on any inherited trait. A SNP is found by aligning overlapping DNA sequences and identifying variable positions as seen in the following diagram:

GCATGCA**A**GCAGATA

GCATGCA**C**GCAGATA

GCATGCA**A**GCAGATA

GCATGCA**A**GCAGATA

The frequency, stability, and even distribution of SNPs in the human genomes cause them to be very valuable genetic markers that locate a disease on the human genome map. Since SNPs are usually found near a gene that is associated with a specific disease, they can be used to search for and eventually isolate the disease-causing gene. SNPs can be applied in the study of evolution in tracing evolutionary history of different populations. They can be helpful in DNA fingerprinting, usage in criminal or parental verification. SNPs can also be used as marker for mapping of polygenic traits and in prescribing genotype-specific medications.

Based on sequencing from individuals of different ethnic origins, about 1 out of every 1,250 base pairs encodes single nucleotide polymorphisms. Meaning, out of 2.9 billion base pairs there are about 10 million specific SNPs that account for all genetic variation encoded in the human population. The SNP frequency which is the fraction of individuals in a population expressing a particular SNP is of great importance in studying

SNPs. The following table demonstrates the likelihood of finding SNPs of various frequencies as a function of the number of individuals screened.

Number of Individuals	**SNP Frequency**				
	> 1%	**> 2%**	**> 5%**	**> 10%**	**> 20%**
2	4%	8%	19%	34%	59%
5	10%	18%	40%	65%	89%
10	18%	33%	64%	88%	99%
20	33%	55%	87%	99%	> 99%
40	55%	80%	98%	> 99%	> 99%

Finding the similarities and differences between two SNPs of two organisms provides more insight about the diseases and their expressions that those SNPs are markers for. The close relationships between the two organisms (Human and Mouse) minimizes the risk of having multiple substitutions at the same site making the results unclear. The decision of comparing the SNPs between Humans and Mouse was made according to the literature proving that 90% of the mouse genome can be lined up with a region of the human genome. Comparing the polymorphism rate between human and mouse highlight the species where one species has a different level of diversity than the other.

The comparison of both genomes was shown to be useful if studying drug development in humans. In the future, an appropriate drug for a certain individual can be determined in advance of treatment by analyzing a patient's SNP profile. This would allow pharmaceutical companies to present drugs that would allow doctors to prescribe more individualized therapies that are more particular to the patient's needs.

Data and Methods

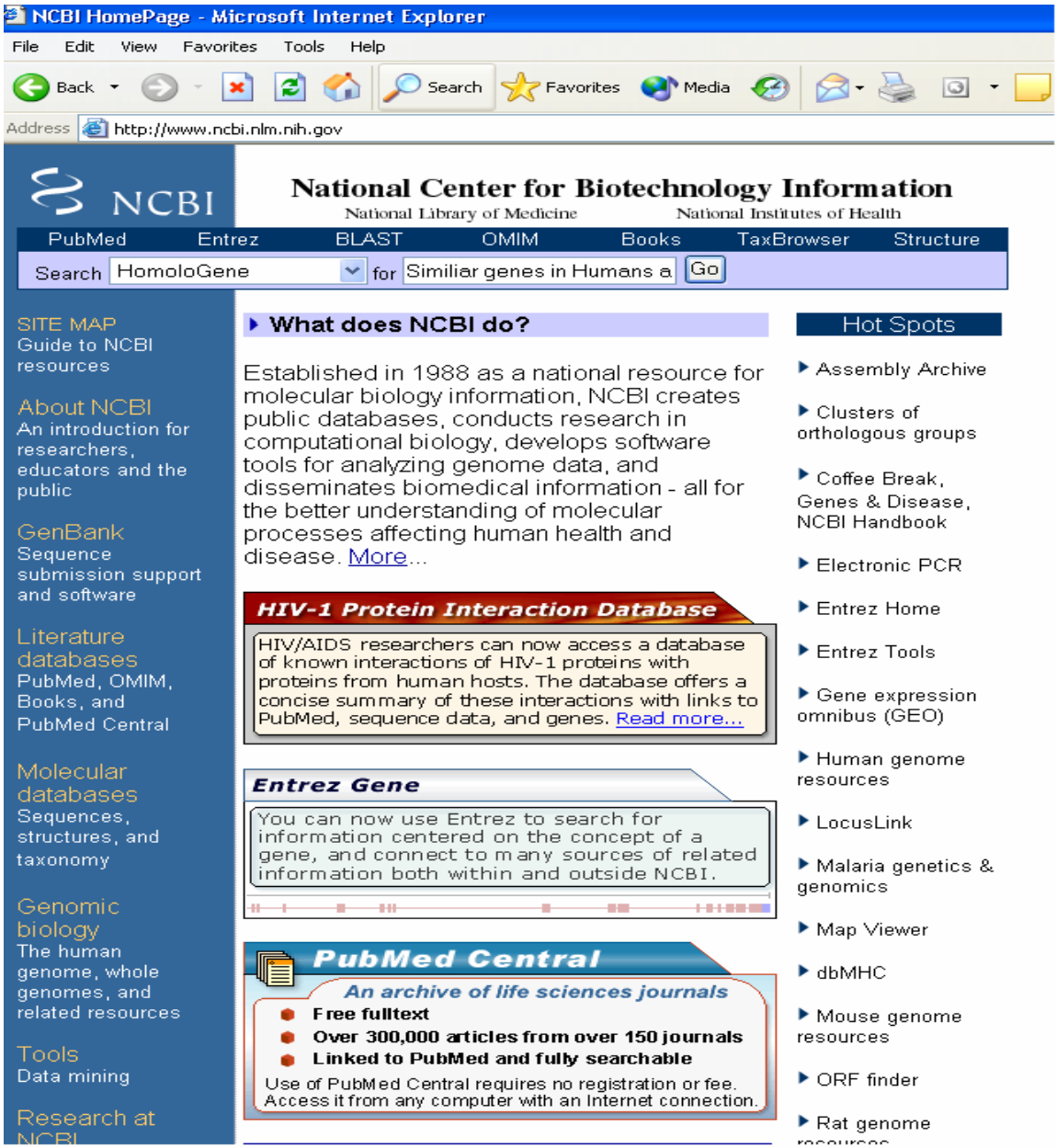
Data Required

The data that we require for our project is SNPs on homologous(similar) genes between humans and mouse.

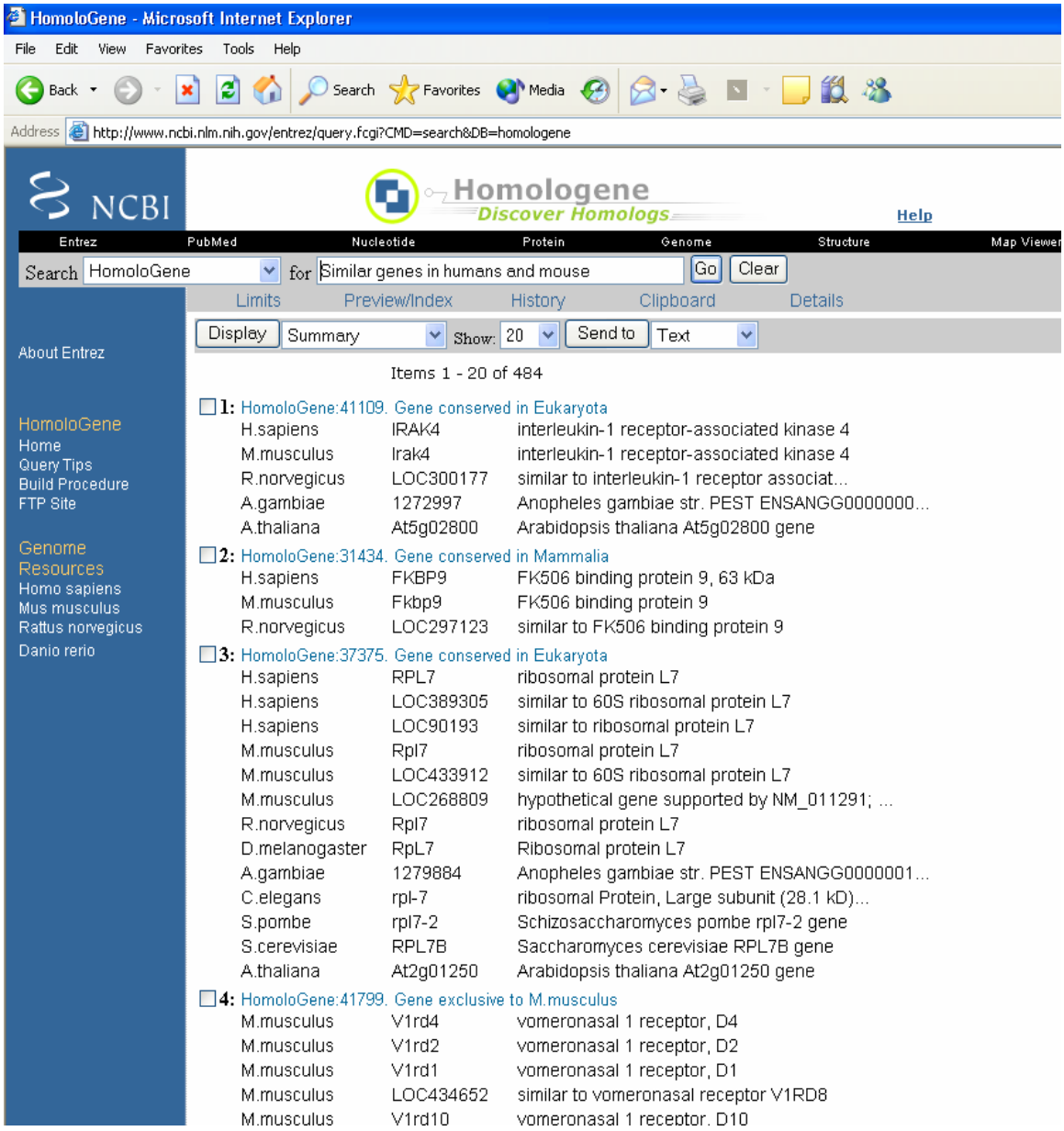
Obtaining Homologous Genes between Human and Mouse

The NCBI homologene database was used to query for “Similar Genes In Humans and Mouse”.

The steps are as shown



The result of the above search yielded a list of **484 similar genes** between humans and mouse.

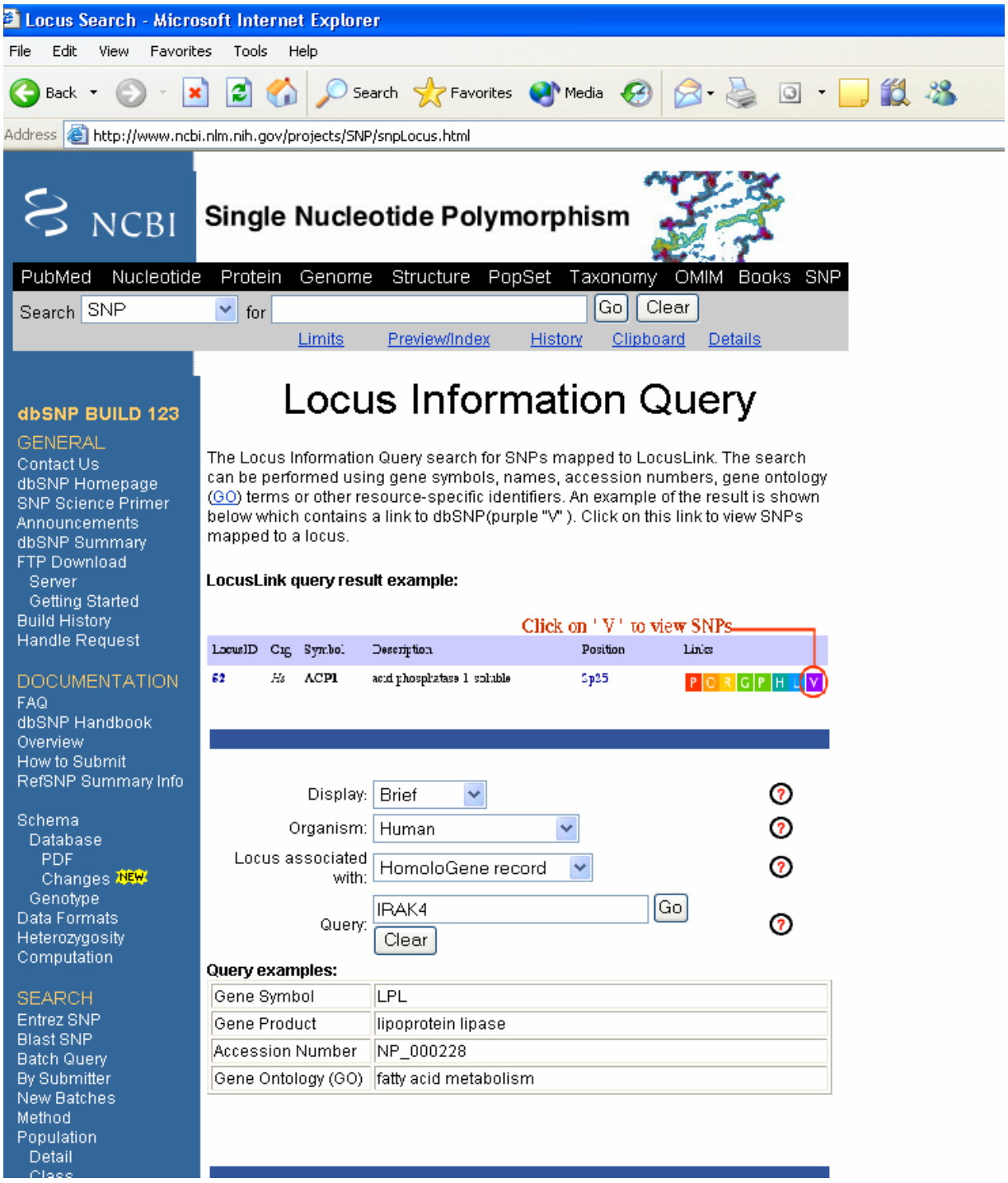


Searching for SNPs

The NCBI dbSNP database was used to search for SNPs in the above obtained 484 genes for both human and mouse.

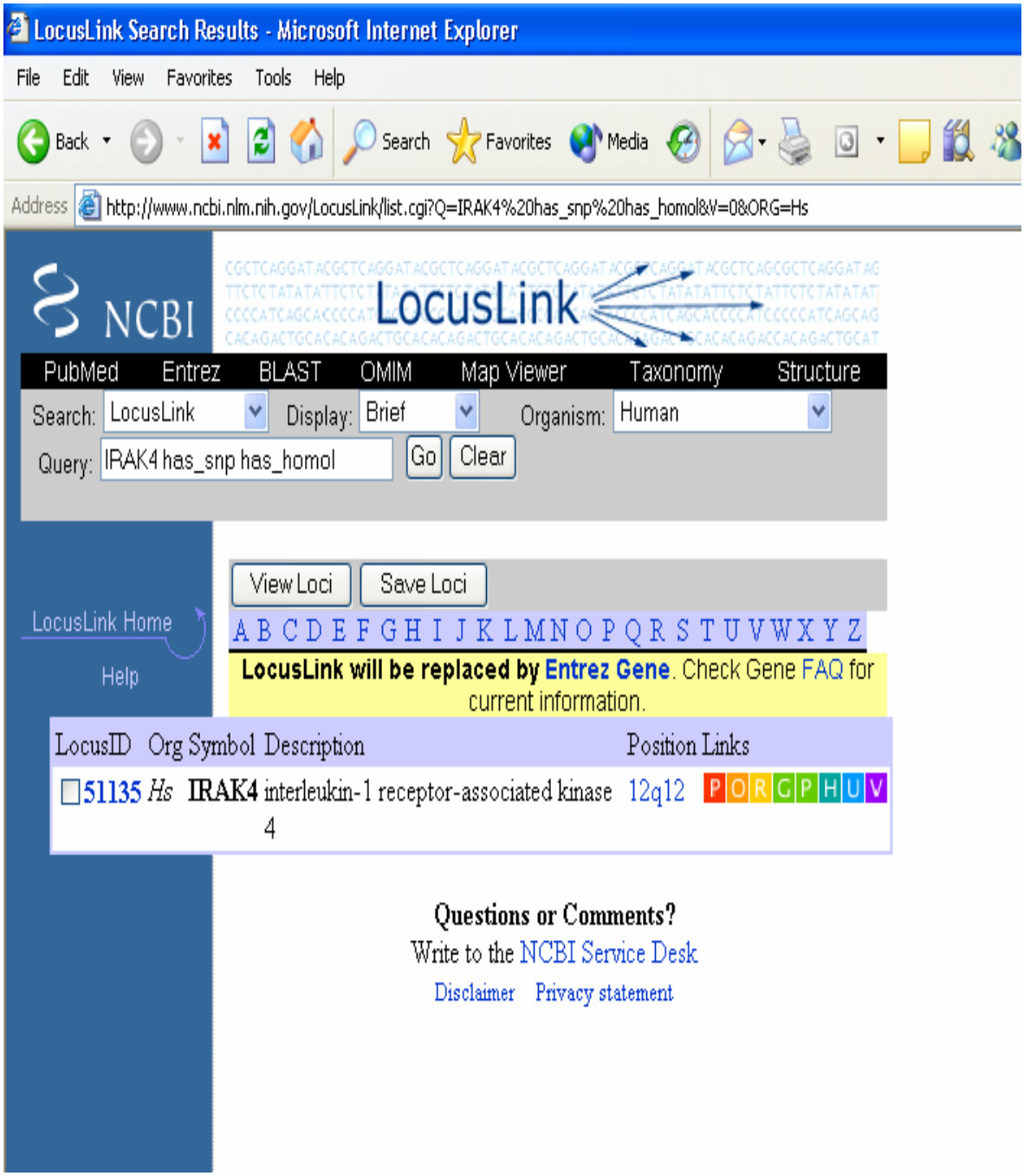
The next set of figures show the flow of control using the NCBI database for obtaining SNPs in the coding and the total gene region for both human and mouse corresponding to the above 484 similar homologenes.

We take the IRAK4 human gene as an example:

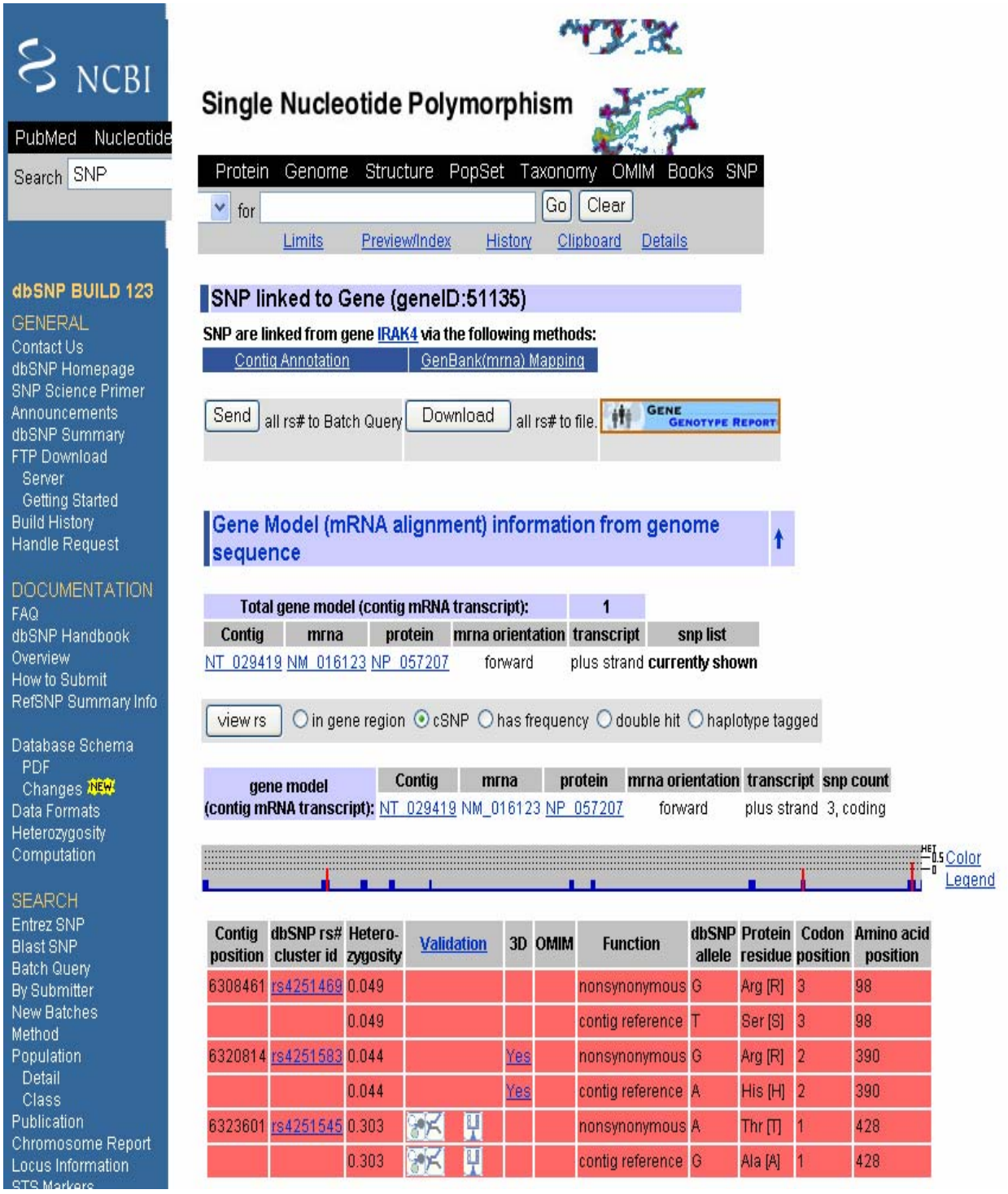


We performed a locus link query in the dbSNP for each gene for both human and mouse based on the homologene record. The example shown above is for the Human IRAK4

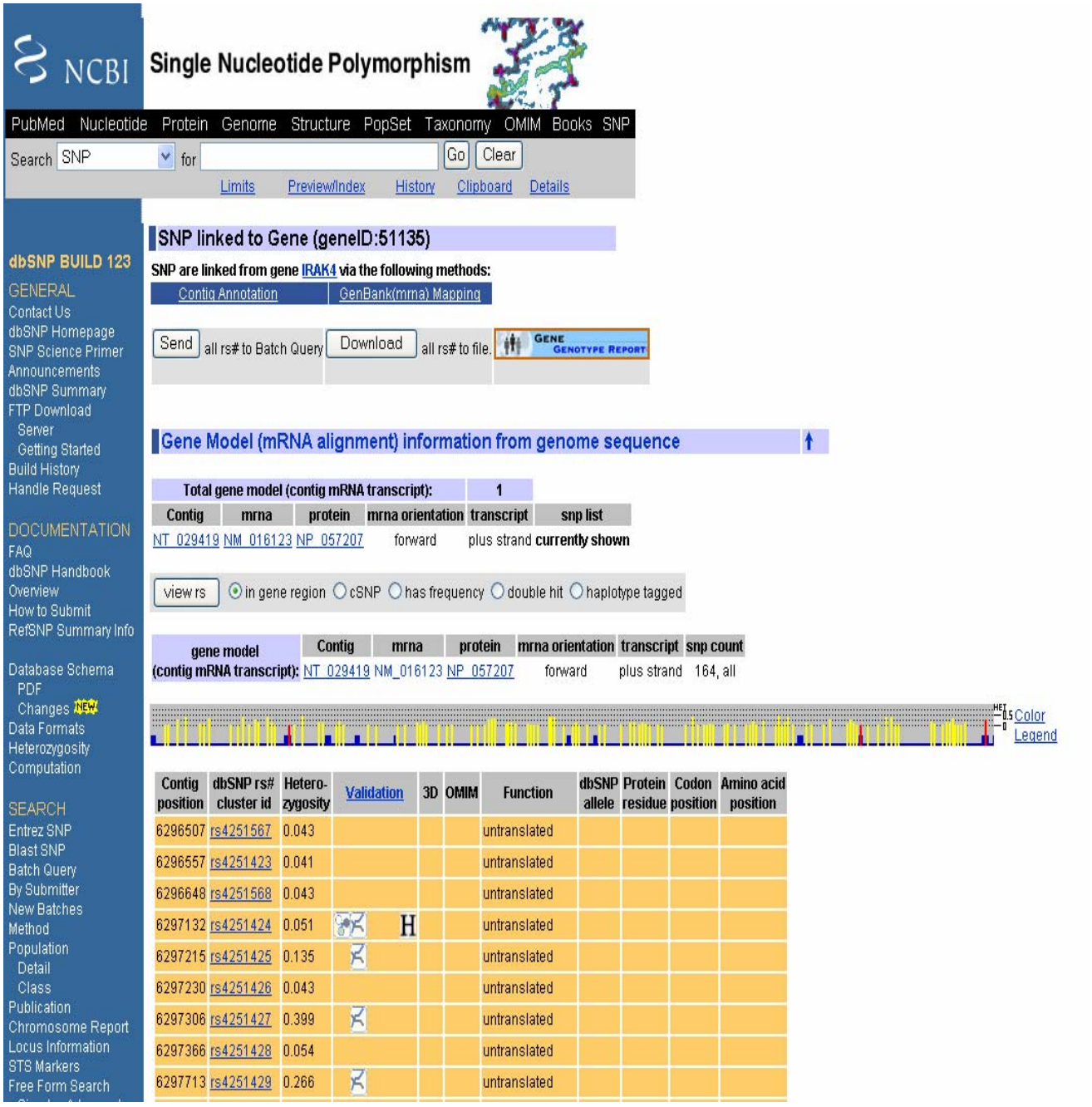
gene.



The result for the locus link query displayed the above information about the human IRAK4 gene. Clicking on the “V” symbol gives the information regarding variations in the above gene



The above view displays the number of SNPs in the coding region for the Human IRAK4 gene along with the alleles for each of the SNPs



The above view displays the total number of SNPs in the complete gene region of the IRAK4 gene.

Tabulating Data

The results obtained from the data mining were tabulated as follows.

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

The complete table is attached at the end of the report.

Final Data

We wanted to concentrate our study on genes which had SNPs in their coding region for both human and mouse so that further analysis of biological significance can be performed. Hence we reduced the above obtained data from 484 genes to 82 genes by selecting the genes which had SNPs in their coding region.

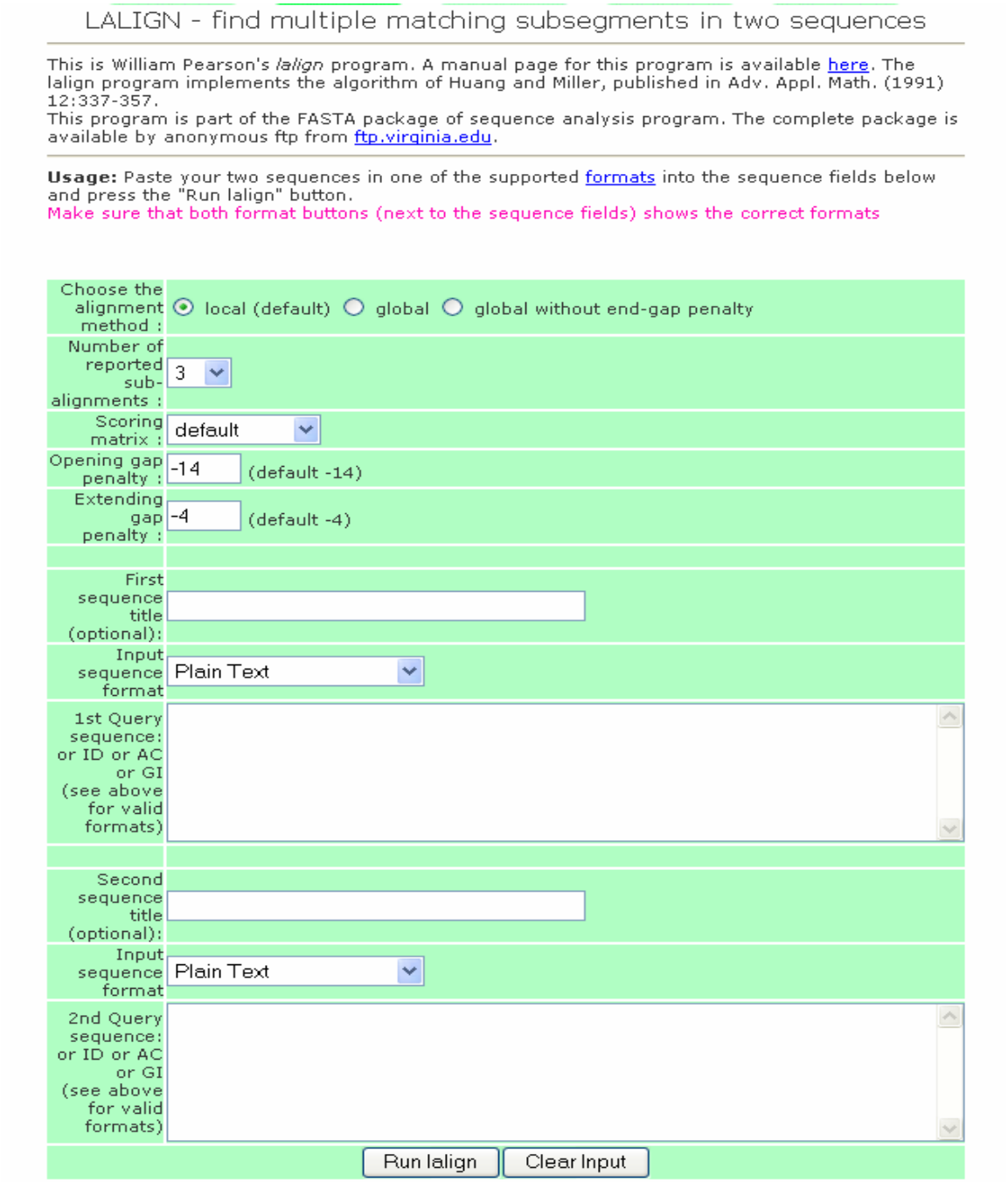
The complete table is attached at the end of the report.

Methodology for SNP Comparison

The following methods were used to analyze our data to obtain the desired result.

Alignment Of Human and Mouse mRNA

The LAlign software was used to align the human and the mouse mRNA as shown in figure



LALIGN - find multiple matching subsegments in two sequences

This is William Pearson's *lalign* program. A manual page for this program is available here. The lalign program implements the algorithm of Huang and Miller, published in Adv. Appl. Math. (1991) 12:337-357.
This program is part of the FASTA package of sequence analysis program. The complete package is available by anonymous ftp from ftp.virginia.edu.

Usage: Paste your two sequences in one of the supported formats into the sequence fields below and press the "Run lalign" button.
Make sure that both format buttons (next to the sequence fields) shows the correct formats

Choose the alignment method :	local (default) global global without end-gap penalty
Number of reported sub-alignments :	3
Scoring matrix :	default
Opening gap penalty :	-14 (default -14)
Extending gap penalty :	-4 (default -4)
First sequence title (optional):	
Input sequence format	Plain Text
1st Query sequence: or ID or AC or GI (see above for valid formats)	
Second sequence title (optional):	
Input sequence format	Plain Text
2nd Query sequence: or ID or AC or GI (see above for valid formats)	

Run lalign | Clear Input

The LAlign software returns the alignment of the two sequences along with the nucleotide positions of the alignment . A portion of the output is as shown.

```
89.2% identity in 240 nt overlap; score:  946 E(10,000):  1e-71

               10        20        30        40        50        60
human  CGCTGCTCCTGCTGCTGCTCTGGGTGACCGGGCAGGCAGCGCCCGTGGCGGGCCTGGGCT
       :::::::::::::::::::::::::::::::::::::::: :: :::  :::::::: ::
mouse  CGCTGCTCCTGCTGCTGCTCTGGGTGACCGGGCAGGCAGCCCCGGTGTTGGGCCTGG-CT
               10        20        30        40        50

                70        80        90       100       110
human  CCGA-CGCGGAGCTGCAGATCGAGCGGCGCTTCGTGCCCGACGAGTGCCCGCGCACCGTG
         :: : :::: :: :::::: ::: : :::::::::: :: ::::: :::::::: :::
mouse  GTGAGCTCGGAACTTCAGATCCAGCAGAGCTTCGTGCCTGATGAGTGTCCGCGCACGGTG
      60        70        80        90       100       110

     120       130       140       150       160       170
human  CGCAGCGGCGACTTCGTGCGCTACCACTACGTGGGGACGTTCCCCGACGGCCAGAAGTTC
       : ::: :::::::::::::::::::::::::::::::: :::: ::::::::::::::
mouse  CACAGTGGCGACTTCGTGCGCTACCACTACGTGGGGACTTTCCTCGACGGCCAGAAGTTC
     120       130       140       150       160       170
```

Comparison of SNP position on Human and Mouse mRNA

We developed a parser which compared the position of SNP on both the human and the mouse aligned segments of the mRNA to identify if the SNPs are conserved between the two species.

The input to the parsing algorithm is a file which contains

- Start point of the alignment as obtained from the LAlign software
- Number of SNP's and their positions for both the human and the mouse mRNA.
- Alleles for the SNP's for both the human and the mouse.

The algorithm is as follows

- Read in the start point (nucleotide position) of the alignments from the given input file.
- Read in the SNPs locations for both human and the mouse from the given input file.
- Read in the number of total coding SNPs in human and mouse from the given input file.
- Find the relative distance between the starting point of the alignment and the SNP.
- Compare if the SNPs in humans and mouse occur at the same relative position.
- If they occur at the same position and their alleles are the same, then the SNP is said to be conserved.

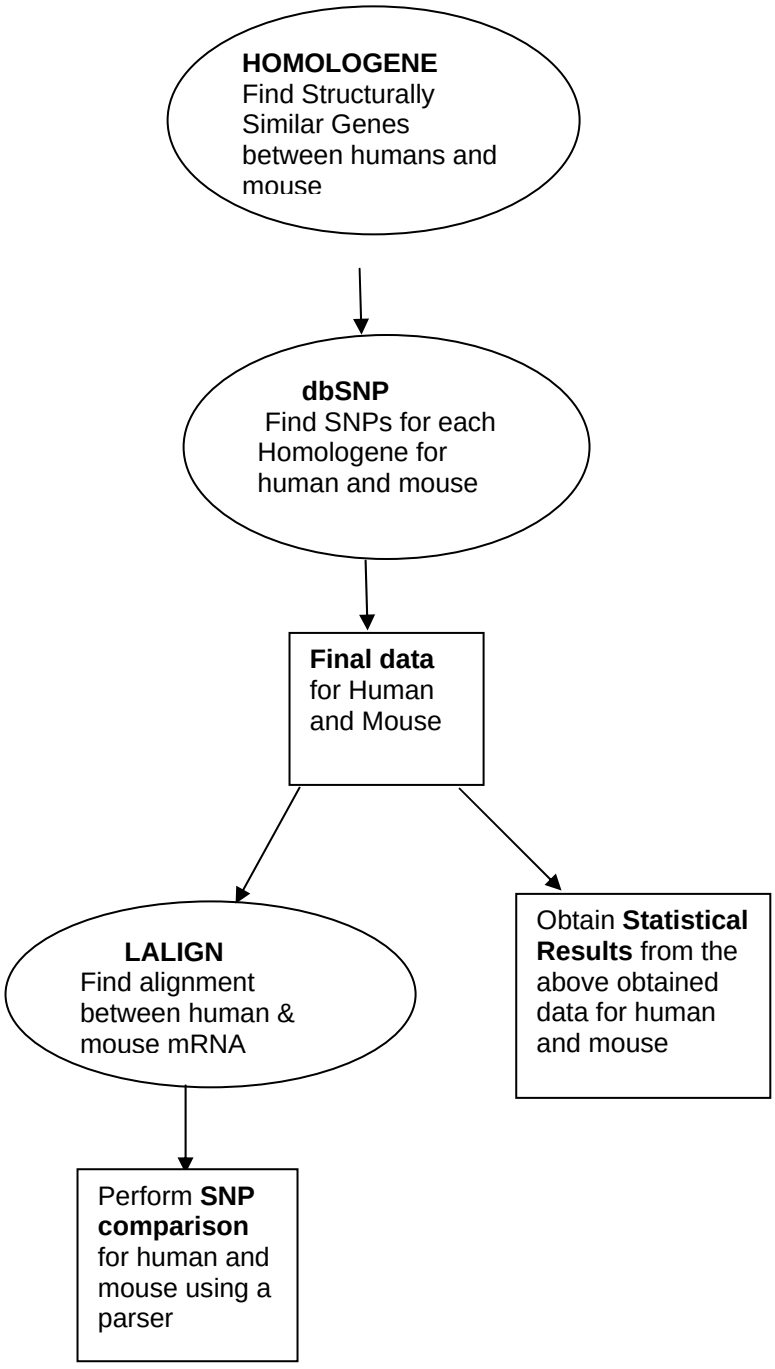
The code for the parser is attached at the end of the report.

Statistical Analysis

We used statistics to obtain results pertaining to SNP percentage, SNP density, Number of SNP in Exons and Introns and CT,GA,TG,GT transitions.

The results obtained are explained in details in the results section.

The following is a **diagrammatic representation** of how the data was obtained and how the methods were applied on the final data to obtain results.



Results

Our analysis in involves 2 steps.

A: Calculating the simple statistics for the Human and Mouse data.

We analyze the data according to the average SNP density in Human and Mouse as well as the density in the exon and the Intron region. We have also calculated the distribution of SNP in both Human and Mouse according to the percentage of SNPs in the Intron and Exon.

B: Test for SNP conservation in Human and Mouse.

The prerequisite for this kind of analysis is the availability of high quality data in terms of both completeness and the annotation.

Intron/Exon Distribution:

In Human data 3 percent of the SNP were found in exon while 97 % in Intron.

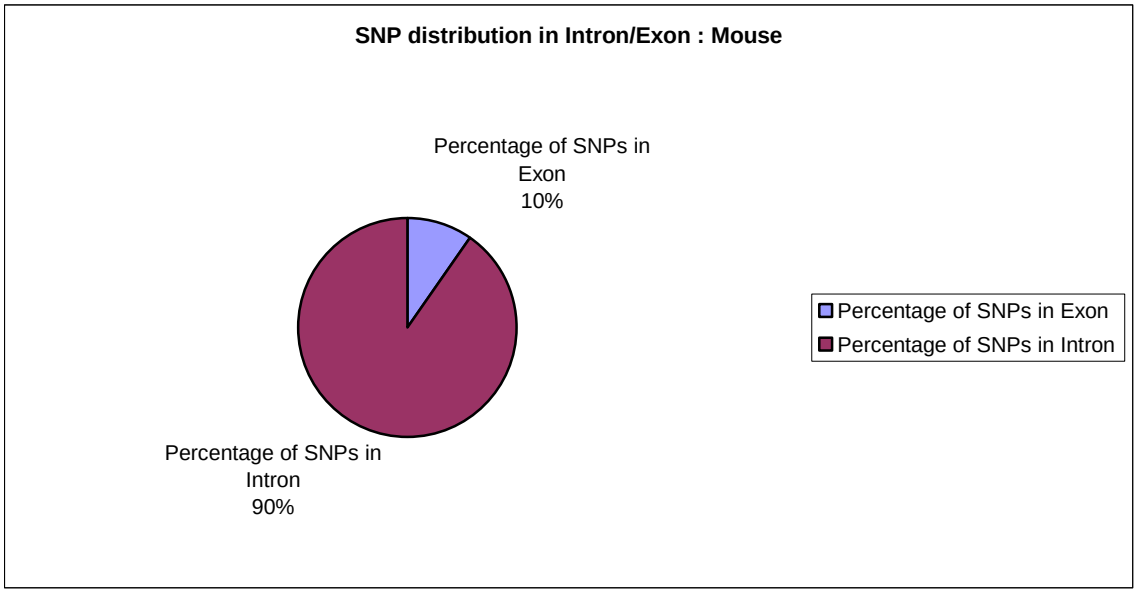


Fig 1 SNP distribution in Mouse

The comparative figure for Mouse is 9.7 and 90.3% respectively.

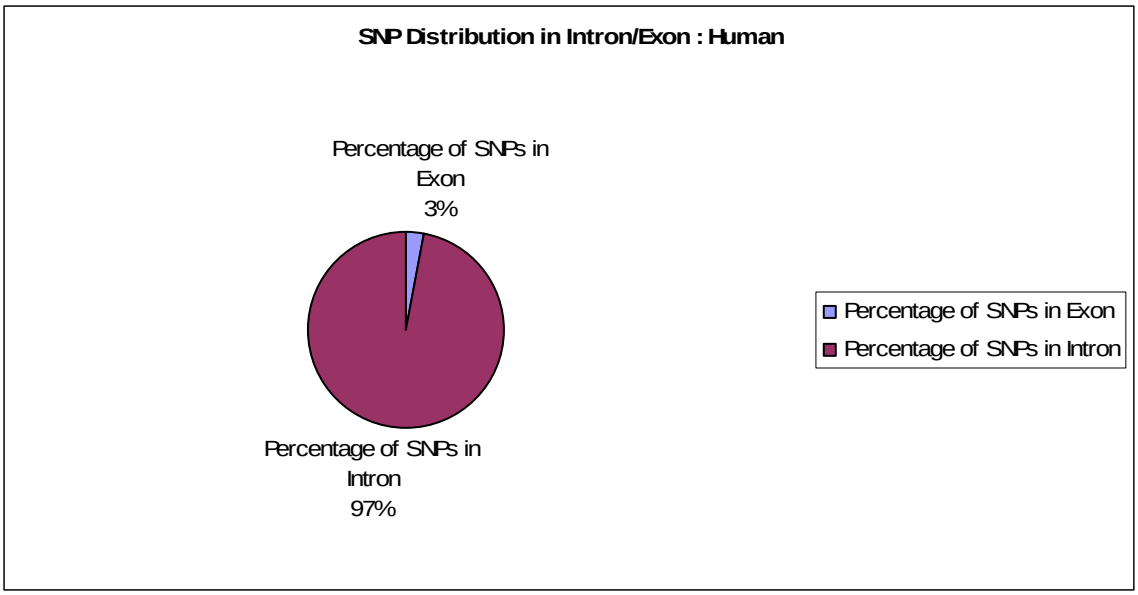


Fig 2 SNP distribution in Mouse

SNP Density: In the human data on average we find a SNP every 335 bp while in mouse the comparative figure is 1087 bp.

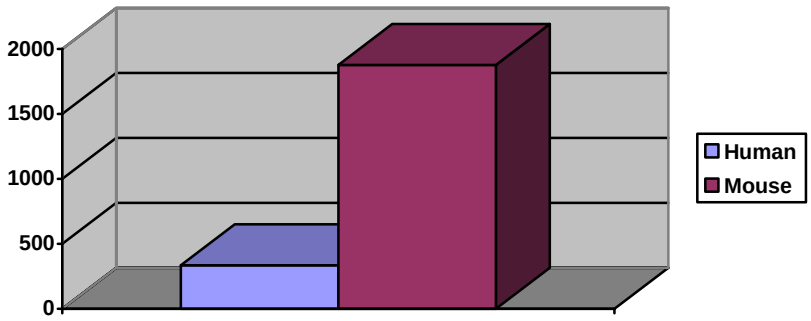


Fig 3 SNP density in Human and Mouse

SNP Density in Exon:

In human the SNP is found every 612 bp on average in the Exonic region. In the mouse exonic data the SNP density was found to be 1877bp.

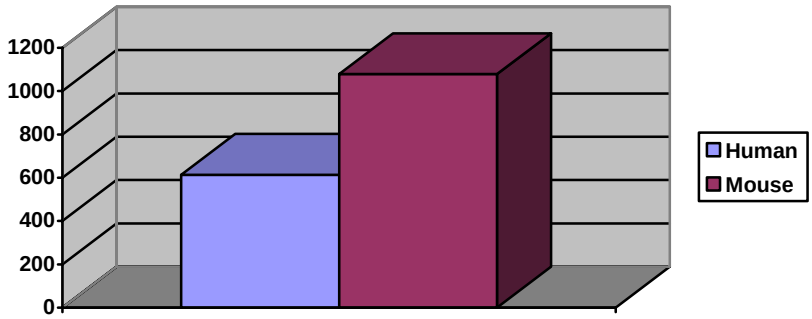


Fig 4 SNP density in Exon : Human and Mouse

SNP Density in Intron:

The intronic density in human data is 328bp while in mouse it it 1962bp.

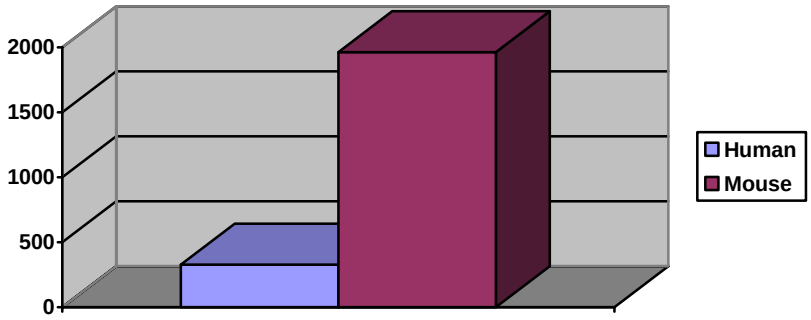


Fig 5 SNP density in Intron : Human and Mouse

Distribution of Alleles in Human and Mouse:

Following graph depicts distribution of the alleles in Human and mouse data.

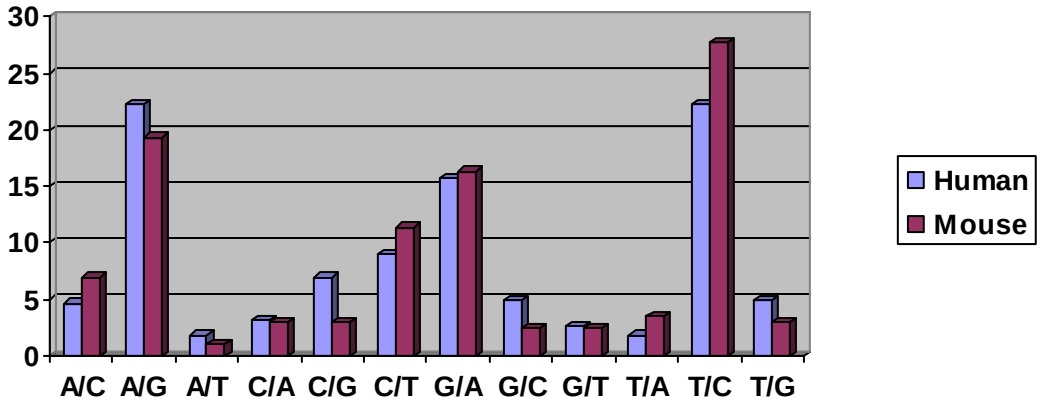


Fig 6 Alleles distribution: Human and Mouse

Conservation of SNP:

We used the data from 10 homologous genes in Human and mouse to test for the conservation of SNP in human according to methodology discussed previously.

We found that none of the SNPs in the 10 mouse genes are conserved in corresponding homologous human genes.

Discussion

As mentioned earlier in the report, we had started with the aim of comparing chimpanzee and human SNPs for certain genes. Due to the unavailability of data about the chimpanzee SNPs (the dbSNP has only 2 SNPs recorded for chimpanzee) we shifted our focus to another organism, mouse (Mus musculus).

"**The mouse that roared**"

The laboratory mouse has been an indispensable tool for investigators in biomedical research. There is scarcely any major area in mammalian biology or medicine in which mouse studies haven't contributed as surrogates for human studies. In all fields from genetics and development, for immunology and pharmacology, for cancer and heart disease, even for behavior, learning and memory and psychiatric disorders the laboratory mouse has become an indispensable tool.

"**The human and mouse genomes sequences can be viewed as two decks of cards obtained by re-shuffling from a master deck – an ancestral mammalian genome".(Pavel Pevzner)**

In the Nature paper, scientists comparing human and mouse genomes found that more than 90 percent of the mouse genome could be lined up with a region on the human genome. That is because the gene order in the two genomes is often preserved over large stretches, called 'conserved synteny. At the nucleotide level approximately 40% of the human genome can be aligned with the mouse genome.

For all the above mentioned reasons, we decided to proceed with mouse as the organism to perform a SNP comparison over genes.

Conclusion

We applied the techniques, mentioned in the Data and Methods Section of our report to obtain the results.

From the results obtained we reached to the following conclusions,

1) **SNPs are not conserved between Homo sapiens and Mus musculus for our experimental dataset.**

   The greater the evolutionary distance between two species, the SNPs are lost through recombination .The human and the mouse lineages diverged about 75 million years ago and hence this immense evolutionary distance results in the absence of the conservation of SNPs between humans and mouse. If a particular SNP was so critical that it had to be conserved through such a large evolutionary distance, then it can be assumed it could undergo mutation and be passed to the humans as a positive selection.

2) **In our experimental data, for both humans and mouse, the number of SNPs present in the introns is greater than that in the exons.** The introns are the non-coding part of the gene and hence such variations may not result in any significant changes in gene expressions.
3) **The percentage of SNPs in the exons of the mouse in the dataset is greater than that in the humans.**
4) **The overall SNP density in humans is greater than that in the mouse in the dataset.**

   This may be because of the difference in the amount of SNP data of the two species. The humans have a large quantity of SNP data documented in the dbSNP database compared to the mouse. Also the laboratory mouse is inbred to some extent and hence the single nucleotide divergence within this mouse species is less.
5) **The C->T transitions (G->A) is high than any other transitions, in both human and mouse, in the experimental dataset.**

   This high level of C->T SNPs is related to the 5-methylcytosine deamination reaction which occurs frequently in the CpG nucleotides.

Thus we performed a statistical analysis of the SNPs in both the humans and mouse came to the above conclusions. Though the conclusions are not spectacular they lay down the foundation for the further research along similar lines.

Future Prospects

Our project can be further improved by covering all possible genes between the humans and mouse as and when more SNP data corresponding to the mouse becomes available.

We propose that SNPs comparison between closely related organisms will help in differentiating the organisms on a genetic level and establish the reasons for the importance of every species. When SNP data of chimpanzee becomes available, the SNP comparison between humans and chimpanzee, will lead to interesting discoveries regarding unique skills present in humans. The human and chimpanzee genome sequences are 98% different from each other. So this close relationship between the sequences minimizes the risk that multiple substitutions at the same sites will obscure the results. Analysis of polymorphism within species and divergence between species will shed light on the evolutionary constraints on the genes.

References

1. "Comparative genomics: The mouse that roared" - Mark S. Boguski
2. "Initial sequencing and comparative analysis of the mouse genome" –Mouse Genome Sequencing Consortium.
3. "Genome wide Comparison of DNA sequences between Humans and Chimpanzees". – Ebersberger, Metzler, Schwarz and Pääbo.
4. Human Genome Project Information. "SNP Fact Sheet". Website: www.ornl.gov/sci/techresources/Human_Genome/faq/snps.html
5. National Cancer Institute. "SNPs and Cancer". Website: http://press2.nci.nih.gov/sciencebehind/snps_cancer/snps_cancer.html
6. National Center for Biotechnology Information. "SNPs: Variations On A Theme". Website: http://www.ncbi.nlm.nih.gov/About/primer/snps.html
7. Alzheimer's Association. "Of Mice and Humans: Comparing the Mouse and Human Genome". Website: www.alz.org/News/02Q4/mice.asp