In Silico RFLP Simulation, incorporated with SNPs

By

MUHAMMAD ASIF

National Center for Bioinformatics
Faculty of Biological Sciences
Quaid-i-Azam University
Islamabad, Pakistan
2013

In Silico RFLP Simulation, incorporated with SNPs

By

MUHAMMAD ASIF

Thesis submitted in the partial fulfillment of the requirements for the degree of
MASTER OF PHILOSOPHY
In
Bioinformatics
Supervisor
Dr. Sajid Rashid

National Center for Bioinformatics
Faculty of Biological Sciences
Quaid-i-Azam University
Islamabad, Pakistan
2013

Acknowledgements

Thanks to my **ALLAH** (Azza wa jal), who is Rehman, Raheem,Ghafoor ,Malik, Sattar and infact He is Zoljalal-e-wal Ikram. One who is Most Perfect, free from all deficiencies, worthy of all Praise, the Most Majestic and the Most High. He is the Self-Sufficient Master Who is not in need of anything whilst all are in need of Him. My Allah always gave me safe place even in critical situations in my life. I ever realized only one support and hope which was of my Allah. No doubt, He is the Most deserving of being remembered, the Most deserving of being worshipped, the Most deserving of being thanked.

All praises and respect are for the **Holy Prophet Hazrat Muhammad (Sallallah-o-Allah-e-Waallehe Wassalam)**. In His high moral qualities, spiritual power, high resolve, the excellence and perfection of His teaching is perfect example of the manifestation of Divine qualities. His teachings did not allow me to diffract from my aim. Also lead me in my career. His name is peace for heart and body.

I would like to express my appreciation to the Dean biological sciences and chairman of National center for Bioinformatics, **Dr. Asghari Bano**. She has been a great source of inspiration and a genuine role model to me. I am very grate full for her efforts to facilitate the NCB students in research.

I acknowledge my deepest sense of gratitude and very sincere thanks to my supervisor, **Dr. Sajid Rashid** for the absolute support to make this thesis possible. There are no words to give reward to him. He deserves more than customary acknowledgment for his invaluable guidance, continuous advice and encouragement throughout the course of this thesis. He has been patient enough to accept my inadeduencies which he might not be accustomed. His precious suggestions helped me refining my concepts and improving my work. I really feel honored that I had this opportunity to work with such a learned scholar as my supervisor. In short, he is role model for me in all aspects of life.

I also would like to acknowledge my other teachers **Dr. Amir Ali Abbasi**, **Dr. Syed Sikander Azam**, and **Dr. Muhammad Faisal** at NCB. Their dynamic personalities really change my ideology towards research.

I am thankful to my all colleagues **Miss. Nosheen**, **Miss. Zahida Parveen**, **Miss. Samina kausar**, **Miss. Madia wali**, **Miss Arooba Khan**, **Miss. Rabia**, **Miss. Kanwal**, **Mr. Muhammad**

Aqeel and **Miss Zohra bibi** for their sincere cooperation. I would like to thank my other class mates **Miss. Ayesha Yousaf**, **Mr. Saad Raza** and **Mr. Muhammad Jan** for their support during my study period. I am also grate full to my dearest friend **Mr. Muhammad Sohail Raza** for his corporation. I would like to thank my friends, **Mr. Mirza Hamaad** and **Mr. Asif Manzoor**.
One person who has always been ready to help me was my roommate, **Muhammad Atif**. Throughout the years of my M Phil study, he has heart fully supported, encouraged, advised, and guided me. Thank you **Atif** for all your support!

I would like to thank administrative and technical staff members of the NCB specially **Ali** and **Naseer bahie** who have been kind enough to advise and help in their respective roles.

Thanks to my brothers **Mr. Shahid Hussain** and **Muhammad Umair** and sisters who gave me good thinking, love and care, and prepare me to proceed in my studies. They lead me in every matter of life. Treated me like a friend, help me to make important decisions of life. I specially want to pay honor to my dear brother "**Mr. Zahid Hussain**", whose efforts for my admission, studies and moral support is illustration of his love for me.

Finally, I take this opportunity to express the profound gratitude from my deep heart to my beloved parents, grandparents, a great blessing and gift of Allah, for their love and continuous support both spiritually and materially. I especially want to say thanks to my kind mother and father, **Mr. Ghulam Hussain** for accepting the many days, nights, and weekends apart during this time of my M Phil studies. No words are enough to thank them for their sacrifices and hardships, they endure for me. Thank you and love you again for always being with me.

May Allah bless and protect all the above people. (Ameen)

MUHAMMAD ASIF

Dedicated

To

My Grand Father and Family

Table of Contents

List of Figure

List of Tables

List of Abbreviation

ALFRED	Allele Frequency Database
CAPS	Cleaved Amplified Polymorphic Sequence
CGI	Common Gate Interface
CGP	Cancer Genome Project
COSMIC	The Catalogue of Somatic Mutations in Cancer
DNA	Deoxyribonucleic Acid
EPIC	Eclipse PlugIn
GUI	Graphical User Interface
HapMap	Haplotype Map
HGC	Human Genome Center
HGP	Human Genome Projects
HGVbase	Human Genome Variability Database
HTML	Hyper Text Markup Language
IMS	Institute of Medical Science
JSNP	Japanese SNP information
JST	Japan Science and Technology Corporation
KMP	Knuth Morris Pratt algorithm
Linux	Linus' Unix
MAF	Minor Allele Frequency
miRNAs	microRNAs
NCBI	National Center for Biotechnology Information
OMIM	Online Mendelian Inheritance in Man
PCR	Polymerase Chain Reaction
PERL	Practical Extraction and Report Language
PHP	PHP: Hypertext Preprocessor
PineSAP	The Pine Alignment and SNP Identification Pipeline

PMMP	Prime Minister’s Millennium Project
RFLP	Restriction Fragment Length Polymorphism
SARS	Severe Acute Respiratory Syndrome
SDE	Software Development Environment
SNPs	Single Nucleotide Polymorphism
SPR Opt	SNP and PCR-RFLP Optimization
STRs	Short Tandem Repeats
TSC	The SNP Consortium
UCSC	University of California Santa Cruz
Unix	Uniplexed Information and Computing
VRSAP	Visual Representation of SNPs And restriction fragment length Polymorphism
WAMP	Windows Apache MySQL PHP

Abstract

The completion of human and other genome sequencing projects through high throughput sequencing approaches provides an opportunity to explore their byproduct, SNPs (single nucleotide polymorphism). Among the short variants that occurred in human genome, SNPs are more prominent than other variants. It is estimated that in human genome, there are more than four million SNPs, which act as molecular markers to comprehend the genome. SNPs are also used in agriculture for crop breeding and help in tracing the genetic disease and its susceptibility. Furthermore, SNPs are precise and effective source to understand evolutionary mechanisms. In this context, RFLP (Restriction Fragment Length Polymorphism), a commonly used technique to genotype the SNPs, is effectively used to study the genetic disorders, DNA finger printing and genome mapping. In this study, we developed a server VRSAP (Visual Representation of SNPs and Polymorphism), which facilitates in combining SNPs and RFLP based data in a single platform and largely helps in designing novel markers for genetic study. VRSAP mainly comprises three modules namely, SNP, RFLP and SNP-RFLP. All these modules work together for string and graphical visualization of SNPs and restriction sites. VRSAP generates a hypothetical gel image to investigate the unique pattern of markers. VRSAP functionality includes multiple species thereby aiding to understand the complexity among orthologs or paralogs. This pipeline is also applicable to human population to uncover the common SNPs with respect to their drug response. By using these features, a comparative study of disease causing genes can be performed to design a specific drug. Finally, VRSAP functionality was validated by the inclusion of transmembrane 106 protein gene family, which clearly indicated the effect of SNPs on restriction pattern change. Taken together, the comparative view generated by VRSAP is helpful in understanding the genome, gene family or even gene evolution.

1. Introduction

Biological sciences have undergone a continuous process of reshaping since 1990, initiative of first sequencing project. The aim of these sequencing projects was sequence assembly, annotation and finally to develop archives for further analysis. This fascinating step results into exponential and rapid growth that has boosted our knowledge, concerning the genetic and genomic variations (polymorphisms). The key element behind this accelerating velocity is the rapidly reducing cost of genome sequencing and it seems that there will be no need of justification for “It is likely that we shall have all sequenced genome in the near future”. Considerable efforts made for genomes understanding brought an era of genetic polymorphism (DNA sequence differences among the genome i.e. individuals, groups or populations). This distinctness is produced by sequence repeats, insertion, deletions and recombinational events (Parvanov *et al.*, 2009; Paigen *et al.*, 2010). Differences in eyes color and hair structure are simple examples of polymorphism. Any disturbance induced in a process due to internal or external factors contribute to polymorphism. The external inducers may be viruses and radiations.

1.1 Mutation and polymorphism

Repeats, insertions, and deletions which created the diversity among individuals are referred as genetic mutation. A phenomenon of a change in DNA sequence occurred because of external agent which may results into disease and also not present in most of individuals of specie is termed as mutation. Nucleotide substitutions are also responsible for emanation of SNPs. Two types of nucleotide substitutions are transition and transversion. First one, compelled for the substitution occurring in between the double ringed purines or single ringed pyrimidines; second one is linked with inter substitution between purine and pyraminidine. It is noted that transition counts for two third of all SNPs which were occurred in a gene.

A nucleotide modification which caused a change in the ancestral DNA sequence of a gene is known as SNP. To be an allelic, a substitution must reach a level where minor allele frequency (MAF) becomes greater than 1%, then it may be referred as the true SNP. SNP is a self explanatory term which covers the sub terms like insertions and deletions, collectively known as the indels. SNPs might be the result of two processes: first one is the incorrect base incorporation during the DNA replication, an extremely rare event where a misincorporation of a base occurs during the DNA synthesis. The frequency of this process has been estimated to be approximately 10^{-9} -10^{-10} per nucleotide (Nachman and Crowell, 2000). The second one source for SNPs is *in situ* chemical modification of a base.

SNPs can affect the gene at transcriptional, proteome and at genome level (Figure 1.1). For example any change in the promotor provides an opportunity to control the gene expression.

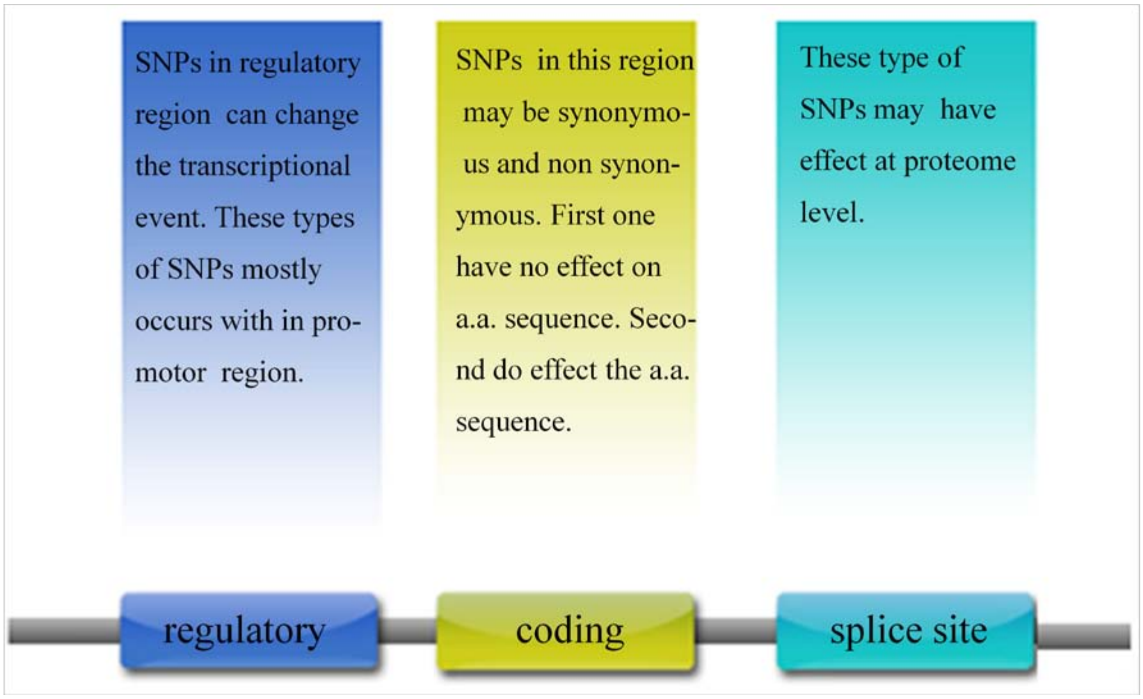

SNPs in regulatory region can change the transcriptional event. These types of SNPs mostly occurs with in pro- motor region.
SNPs in this region may be synonymo- us and non synon- ymous. First one have no effect on a.a. sequence. Seco- nd do effect the a.a. sequence.
These type of SNPs may have effect at proteome level.
regulatory
coding
splice site

Figure 1.1: SNPs types. A gene consists of coding and non-coding region. Non-coding region further comprises regulatory, splice site etc. SNPs occur in every segment of genes. Mostly these are found in the non-coding region which might be the regulatory or splice site region. In figure common types of SNPs are shown.

Dispersion of SNPs does not obey a uniform pattern as observed in human genome, among chromosomes or even within a single chromosome. It had also been observed that sex chromosome shows more inflexibly toward SNPs as compared to other chromosomes. Currently fashioned approaches indicate that SNPs concentrated about a specific region or within a single chromosome usually means a region of medical or research interest. On the basis of precise and improved sequencing projects it is argued that SNPs occur with an average of every 300 base pairs (Brookes, 1999). SNPs also contribute in the phenomenon of eukaryotic genome plasticity (Fradin *et al*., 2003; Hube, 2004; Forche *et al.*, 2009). SNPs have a wide spectrum of applications that can be specifically reduced to qualitative markers, to solve the mystery of mixed quantitative samples, dealing with genomic DNA and RNA transcripts (Chunming, 2007).

Human genome mapping project (HGP) had cropped up with sanction, authorize and legal set of human single nucleotide polymorphisms (SNPs). This refined product with emergent significance was unexpected and consequence of several genome (cellular organisms, viral etc) sequencing tasks. It may be termed as the secondary product of those comprehensive projects. In depth knowledge of locus and sighting of SNPs by using the resequencing technologies, results into first SNP map (International Human Genome Sequencing Consortium, 2001*;* Sachidanandam, 2001*).* Furthermore; our awareness towards SNPs provided a chance to understand the structure and role of genome or its gene content. The SNPs present the most absolute and tightly spaced system of genome landmarks, available at the moment. Moreover, SNPs also permit the researchers to draft the linkage maps, actual and specific mapping of genes engrossed in multifaceted disorders. Besides improved mapping, knowledge of SNPs also facilitates us with a new and captivating layer to increase our understanding towards human variability and thus providing an opportunity to catch on disease susceptibility, common population variations and evolution. Above all, low mutational rate as compared to other polymorphic markers makes SNPs an ultimate choice in exploring the evolutionary history of populations.

Moreover, SNPs are under investigation to understand human disease gene mapping (Wang *et al.*, 1998), human evolution (Stoneking 2001; Bamshad *et al.*, 2004; McCarthy *et al.*, 2008), pharmacogenomics, functional proteomics, germ line gene therapy, somatic cell genetic mutations and somatic cell gene therapy. SNPs are the crux essence of genome that may provide a way to answer the enigma nature of human genetics: why comparable number of genes in species at opposite ends of the evolutionary scale can create such extremely different levels of complication which are still remains to be explored.

1.2 Experimental methods for genotyping the SNPs

1.2.1 PCR-RFLP

With an aim of genetic association study, one of the methods which can be the ultimate choice for SNPs genotyping is PCR-RFLP. SNPs can be differentiated by utilizing RFLP where the target is achieved by analyzing the patterns, derived from the cleavage of amplified DNA. One short-coming of RFLP is its requirement of definite DNA amount. However, this problem is overcome by exploiting the DNA amplification property of PCR (Polymerase chain reaction) in shorter period of time which also reduces the time required for RFLP analysis (Kwok, 2001; Fan *et al*., 2006). Other experimental methods include single strand conformational polymorphism (Orita *et al.,* 1989), enzymatic mutation detection (Youil *et al.,* 1995), microarray or variant detector arrays (Wang *et al.*, 1998; Marshall and Hodgson, 1998; Ramsay, 1998; Hacia *et al.*, 1999; Hacia and Collins, 1999; Dong *et al.,* 2001; Qi *et al.*, 2001; Yoshino *et al.*, 2001), heteroduplex analysis (Lichten and Fox, 1983), MALDI-TOF (Griffin and Smith, 2000), pyrosequencing (Ahmadian *et al.*, 2006) and Invader assay (Olivier, 2005) are being used to locate the SNPs.

1.3 Data servers for SNP

The problem of data management, generated from the extensive experimental work is being resolved by creating the data servers.

1.3.1 COSMIC

COSMIC (The Catalogue of Somatic Mutations in Cancer) (Forbes *et al.*, 2008) is probably the most extensive worldwide source of information about somatic mutations which were reported inside individual tumor. It merges curation from the scientific literature together with tumor resequencing information, coming from the particular cancer genome project at the Sanger Institute, U.K. Practically 4800 genes of different families and about 250,000 tumors have recently been reviewed thus, providing more than 50,000 mutations. All these information are accessible for exploration and further analysis. The actual COSMIC website receives information through three unique sub projects: explicitly technological materials (scientific literature), CGP resequencing task, and the cancer cell line jobs. Access the main COSMIC home page at *http://www.sanger.ac.uk/cosmic.*

1.3.2 JSNP

JSNP (Hirakawa *et al.*, 2002) is a database associated with Japanese Single Nucleotide Polymorphism (SNP) information. It was started in early 2000 with the Prime Minister's Millennium Project (PMMP). The purpose of JSNP was to determine up to 150,000 SNPs through the Japan populace, situated in gene or within surrounding regions that may affect the code series of genetics. The task was completed by the cooperation between Human Genome Center (HGC) within the Institute of Medical Science (IMS) and the The Japan science and Technology Corporation (JST). Connected documents are obtainable at *http://snp.ims.u-tokyo.ac.jp/*.

1.3.3 OMIM

OMIM: Online Mendelian Inheritance in Man (Hamosh *et al.*, 2005) is the NCBI phenotype data source, which catalogues the characteristics related to diseases and disorders. Some time, it presents the results without any reference to the gene, a case when no link has presently been explained for that specific query. For each entry in OMIM, a distinctive number which were frequently used in literature is assigned. The obvious distinction from other data sources is that, OMIM is actually managed through NCBI and created individually in John Hopkins University. OMIM is accessible from *http://www.ncbi.nlm.nih.gov/omim.*

1.3.4 The dbSNP database

NCBI dbSNP (Sherry *et al.*, 1999; Sherry *et al.*, 2011) is the primary data source associated with SNP information which was fetched from the HGP (Human genome project). dbSNP remained very active in order to gather all the information with respect to SNPs from The SNP Consortium (Thorisson *et al.*, 2003), Perlegen SNP genotyping and HapMap (Clark *et al.*, 2005; Goldstein *et al.*, 2005; Hinds *et al.*, 2005). The actual SNP information tends to be frequently up-to-date within synchrony along with genome rebuilds and makes sure the high quality associated with SNP locus mapping. *http://www.ncbi.nlm.nih.gov/SNP/* is the address where dbSNP can be found.

1.3.5 The HapMap project

The actual worldwide HapMap project (Clark *et al.*, 2005; Goldstein *et al.*, 2005; Hinds *et al.*, 2005) premiered at the end of October 2002 using the mentioned purpose of identifying the haplotype framework i.e. structure of human genome. HapMap objectives can be estimated from their personal statement, "all typical human sequence variation, offering information required to channelize the genetic research associated with clinical phenotypes." HapMap project can be found at *http://hapmap.ncbi.nlm.nih.gov/*.

1.3.6 Ensembl

Ensembl (McLaren *et al.*, 2011; Kinsella *et al.*, 2011; Flicek *et al.*, 2012) is the genome informational databases along with substantial directories and tools, coming through the vital part of the Sanger Center and EMBL within HGP. Ensembl is constantly on a way of supplying the most annotated genome information coupled with the largest selection of species along with genome evaluation. In case of human SNP data and assisting information, the Ensembl information is extremely closely related to the data which is present in dbSNP. Therefore there might exists a same selection methods to access the NCBI or Ensembl, mostly resulting into exactly the same information, utilizing similar frames with regard to searching and cross-referencing information for any research activity. The web address for Ensemble is *http://www.ensembl.org/index.html.*

1.3.7 The SNP consortium

The SNP Consortium (Thorisson *et al.*, 2003) is actually operated by the Cold Spring Harbor Laboratory resulting from general public relationship associated with seventeen organizations. All data of this particular database includes more than 1.8 million loci, which are classified in dbSNP. Each directory of data lean to be completely cross-referenced along with hyperlink outs, however, TSC (The SNP Consortium) also utilized a different SNP locus recognition program. A primary objective of the consortium was to build the very first compact SNP linkage map. This particular task came up with the most important functions from the TSC data source to genotype with respect to SNPs, across the Western (termed Caucasian), Africa as well as Chinese language populace. The SNP Consortium can be traced at *http://snp.cshl.org/*.

1.3.8 HGVbase

Human Genome Variability Database (Fredman *et al.,* 2002) includes almost more than 10 million records of human genome variations such as SNPs, Indels, as well as STRs. HGVbase utilizes its own approach for locus recognition: the nine-digit number prefixed along with SNP. Currently it is working with a vision of cataloging phenotype/genotype at a larger scale. *http://hgvbase.cgb.ki.se/* is the web address to reach at HGVbase.

1.3.9 ALFRED

Another substantial and significant Data source which reviews allele frequency with respect to polymorphic markers is Allele Frequency Database (Cheung *et al.*, 2000). It is composed of more than 1501 loci, 475 populations, and 41,980 frequency tables. Regrettably, the actual SNP information, kept in ALFRED is patchy and irregular (only 841 rs- numbers).

http://alfred.med.yale.edu/alfred/index.asp is the address where ALFRED can be traced.

1.4 Computational tools

Availability of data servers results into a number of computational tools and softwares. Various computational tools are:

1.4.1 SNP cutter

Zhang and his colleagues (Zhang *et al.*, 2005) explained an online application, 'SNP Cutter' that styles PCR-RFLP assays on a set allied with SNPs from the human genome. NCBI, dbSNP rs IDs or formatted SNPs are likely to be pasted in to the SNP cutter as inputs which in turn utilize restriction enzymes from the pre-selected listing for enzyme digestion. This program has the ability of creating primers with regard to possibly natural PCR-RFLP or mismatch PCR-RFLP, based on the SNP sequence information. SNP Cutter provokes the information, required to assess, review and to carry out genotyping experiments. The actual SNP Cutter is accessible at

http://bioinfo.bsd.uchicago.edu/SNP_cutter.htm.

1.4.2 SNP-RFLPing 2

The main purpose of SNP-RFLPing 2 (Chang et *al.*, 2010) is to offer extensive PCR-RFLP information along with functionality regarding to SNPs. The information used by SNP-RFLPing 2 is assessed from HapMap project, tag SNPs, gene ontology- dependent searching, miRNAs, and SNP500 Cancer. For an accessed SNP with an objective of multiple specie analysis gene centric investigations are performed. The RFLP restriction enzymes and the correlated PCR primers for natural as well as the mutagenic types for every SNP tend to be examined at the same time.

1.4.3 SNP2CAPS

SNP2CAPS (Bikandi *et al.*, 2004) helps to understand the actual computational transformation associated with SNP markers into CAPS (cleaved amplified polymorphic sequence) marker (Parsons *et al.*, 1997). An unmarked computer application called SNP2CAPS functions on a simple algorithm which requires the actual verified multiply aligned sequences for restriction sites detection. After that a selection pipeline that allows the deduction of CAPS candidates by the identification of putative alternative restriction patterns. While implementing this algorithm any primer pair flanking the SNP site may be suited for CAPS marker analysis and evaluation. SNP2CAPS will not measure the specialized effectiveness and usefulness of the recommended restriction patterns, which means that in some instances the actual analysis of restriction sites tend to be (i) crowded towards the edges from the PCR fragments or even (ii) crowded to some 2nd restriction site to permit resolution of the size of polymorphic DNA fragments within agarose gel. This particular tool was actually created within Perl and it can take multiple alignments as inputs because of AlignIO Bioperl module's components.

1.4.4 PineSAP

The Pine Alignment and SNP Identification Pipeline (Wegrzyn *et al.*, 2009) offers a high-throughput treatment for SNP conjecture, utilizing multiple sequence alignments by using re-sequencing information. This particular pipeline combines the crossbreed associated with server scripting, current resources and machine learning approach which boosts the speed and precision associated with SNP calls.

This pipeline introduced by the PineSAP worked on Unix/Linux system. This tool was implemented by using Perl. PineSAP uses Phred (Ewing *et al.*, 1998) and Phrap (Machado *et al.*, 2006) in order to effectively and precisely deal with call bases and then aligns re-sequencing reads. Subsequent alignment, SNPs and indels tend to be recognized with the Polyphred (Nickerson *et al.*, 1997) and Polybayes (Marth *et al.*, 1999) deals. Sequence-based info is actually pulled out and prepared via a supervised machine learning algorithm to accept or reject the actual SNP predictions. The address to locate the PineSAP is *http://dendrome.ucdavis.edu/adept2/resequencing.html*.

1.4.5 SPR Opt

SNP and PCR-RFLP Optimization (Gardner *et al.*, 2005) functions extensively, to provide whole-genome investigation to be able to forensically discriminate numerous sequences. SPR Opt computes almost all SNPs or even PCR-RFLP variants found in the actual sequences and organizes all of them into haplotypes based on their own co-segregation throughout the sequences. It works in combinatory fashion to find out which sets are associated with haplotypes and offer maximum discrimination from entered sequences. Phylogenetic trees are developed depending on SNPs, PCR-RFLPs, and entire genomes, in comparison with respect to SARS virus. It proposed that phylogenies supported by SNP or PCR-RFLP variants usually do not complement with those depending on multiple sequence alignment of the full genomes. The software and source code for SPR Opt is publicly available and free for non-profit use at *http://www.llnl.gov/IPandC/technology/software/softwaretitles/spropt.php*

1.4.6 SNPselector

A web-based application, SNPselector (Xu *et al.*, 2005) requires a set of gene names or a set of genomic locations as a query and then it finds the respective obtainable SNPs in Ensembl. This categorizes the type of SNPs on the basis of tagging with respect to linkage disequilibrium, SNP allele frequencies, resource, functionality, regulating potential and repeat status. SNPselector results can be downloaded as a compacted Excel spreadsheet file. SNPselector is freely available at *http://primer.duhs.duke.edu/.*

SNPselector is actually applied within object-oriented PERL. The primary component could be operated by UNIX command-line tools. Additionally, it gives a special CGI package. A local MySQL data source (*http://www.mysql.com/*) stores all the SNPs as well as associated genome-annotated information. SNPselector uses the downloaded information from HapMap project (*http://www.hapmap.org*), The SNP Consortium (Thorisson *et al.*, 2003) (*http://snp.cshl.org/*), JSNP (*http://snp.ims.u-tokyo.aiconditioning.jp/*), Affymetrix (*http://www.Affymetrix.com/*), Perlegen, Ensembl and UCSC genome browser.

1.4.7 SNPPicker

SNPPicker (Sicotte *et al.*, 2011) improves the selection of tag SNPs through typical bin-tagging applications to develop and customized genotyping panels. This application form utilizes a multi-step lookup technique in conjunction with the statistical model to increase the actual genotyping achievement from the chosen tag SNPs. SNPPicker was created within Perl and java. It is available at *http://mayoresearch.mayo.edu/mayo/research/bio stat/software.cfm.*

1.4.8 SNP-PHAGE

SNP-PHAGE (Matukumalli *et al.*, 2006) is a SNP discovery pipeline with additional features for identification of common haplotypes within a sequence tagged site

(Haplotype Analysis) and GenBank (dbSNP) submission. It was created within Perl and utilized a MySQL with a purpose of database development. In order to boost the actual effectiveness associated with a SNP prediction, a machine learning tool which was produced by the same team is incorporated as part of this particular bundle as an optionally available function. The SNP-PHAGE software program pipeline is an amalgam of UNIX/Linux command line and internet browser interface. It was practically implemented within Perl and utilized regular free modules for example Bioperl. For full functionality, SNP-PHAGE still needs some other deals, for example Phred (Ewing *et al.*, 1998), Phrap, CrossMatch (*Phrap 2006)*, PolyPhred (Nickerson *et al.*, 1997), PolyBayes (Marth *et al.*, 1999) and C4.5 (Quinlan et *al.*, 1993). All these applications are openly available for educational purpose utilization, provided at SNP-PHAGE web site. The SNP-PHAGE package is being made available as an open source at *http://bfgl.anri.barc.usda.gov/ML/snp-phage/.*

1.4.9 SNPLogic

SNPLogic (Pico *et al.*, 2009) combines information about the genetic events associated with SNPs (gene, chromosomal region, functional location, haplotype tags along with transcription factor binding sites, splicing sites, miRNAs as well as evolutionarily conserved regions), genotypic information (allele frequencies for each per population and validation method), coverage of commercial arrays (ParAllele, Affymetrix and Illumina), practical forecasts (modeled upon structure and sequence frameworks) and associations or set up links (biological pathways, gene ontology conditions as well as OMIM disease terms). Therefore, SNPLogic may be used to trace, to prioritize applicant SNPs and evaluate the customized and commercial arrays panel. It enables you to recognize and discover the SNPs annotation by using the pathway and functional prediction scores information. Features at the front end of website with respect to SNPLogic were created within PHP and powerful dynamic HTML/JavaScript while the back end was maintained through the Mysql5. Accessible address for SNPLogic is *http://www.snplogic.org.*

1.5 Aim and objectives

All the experimental methods mentioned above have certain limitations and always there arises a question mark on their performance and accuracy. Experimental methods using the DNA sequencing methodologies are although rapid in their processing but at the same time they are highly expensive. Other methods like single strand conformational polymorphism have shortcoming with their fragment length and the temperature variation. Moreover, pyrosequencing approaches require the read length improvement. MALDI-TOF (Griffin and Smith 2000) like techniques require high expertises and established lab as well. To compensate these drawbacks of experimental techniques, various computational efforts have been made, also discussed in 1.3 sections. All the data servers, covered earlier can easily be classified into two categories, First one is general, dealing with SNPs data coming from all populations e.g. Ensembl, HapMap project and dbSNP and the second one is specific, accumulates the data from a specific population like JSNP database. Although they also have other information but their primary objective involves the specific population. Same trend had been observed in the current tool development trends such as SNP Cutter, SNP-RFLPing 2 and SPR opt are specialized for the PCR-RFLP analysis while other, SNPSelecter are for genetic study across the population. Currently, according to our knowledge there is not a single tool or a server which exploited the SNPs and RFLP across the orthologs and paralogs with incorporation of hypothetical gel image.

To fill this gap we have developed a server, VRSAP (visual representation of SNPs and Restriction Fragment Length Polymorphism) which generates the sequence string map and displays graphical image for SNPs and RFLP results. In addition to SNP and RFLP modules, VRSAP also combines both SNP and RFLP modules. SNP-RFLP module's string map highlights the SNPs which occurred in restriction sites. This module also provides the graphical view of string map, coupled with the hypothetical gel image, concerning with multiple sequences. We believe that VRSAP can be a conclusive and informative tool in the following areas: genetic studies (involving the genetic markers and disease causing gene), evolution, gene families evolution, mutational rate across different genes or species, paralogs evolution, tracing the evolutionary conserved restriction sites, locating the recently governed restriction sites, creating SNPs and RFLP map, finding the origin of different species and primer designing.

2. Materials and Methods

The web application, VRSAP was implemented within PERL and JavaScript along with HTML (Hyper Text Markup Language) support. It contains applications for computational evaluation of SNPs, RFLP and both at species and populations levels. Flowchart for this tool is shown in figure 2.1.

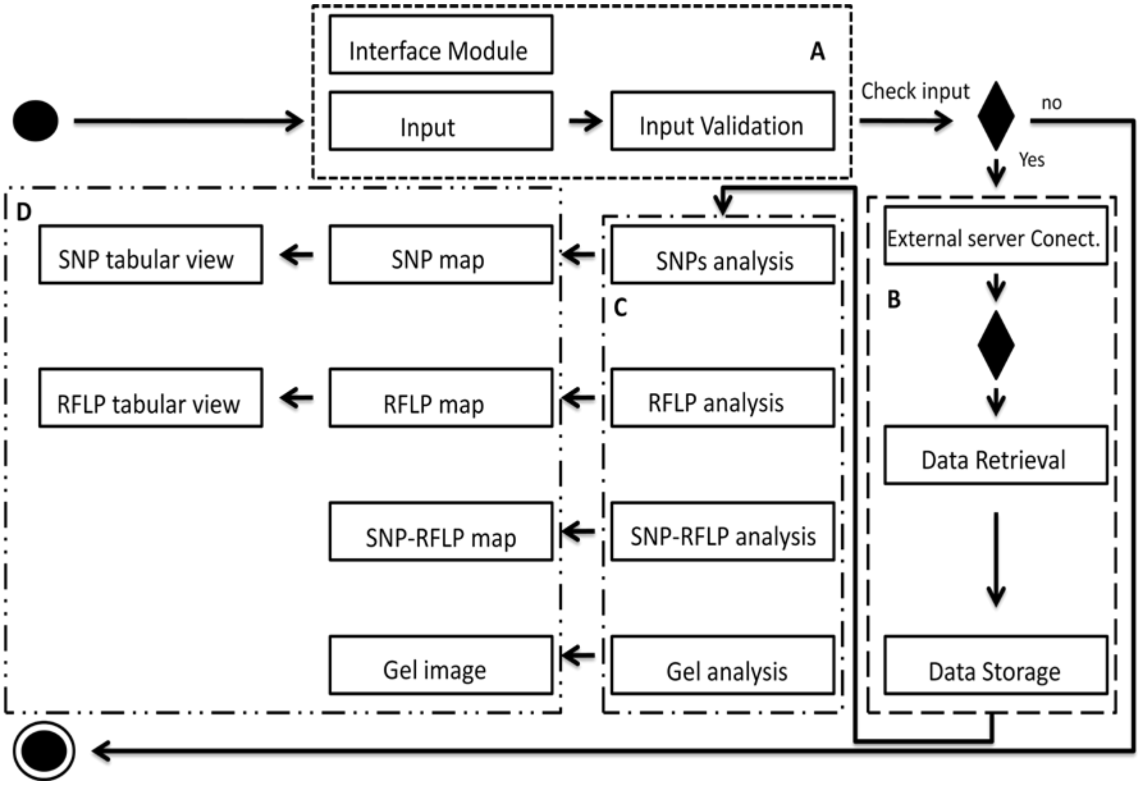


Figure 2.1: Schematic overview of VRSAP working. A. Input of gene names and their species. The input module is validated by input validation module in interaction layer (See 2.1). B. Depending upon the user's input, a connection with external server is created and the retrieved data is stored in database and files. C. This is a processing phase where SNP, RFLP and SNP-RFLP analyses are applied, depending upon the request made by the user in input module. D. This layer represents the output from the processing layer.

In order to effectively cope with actual required job, numerous segments or modules associated with VRSAP, created within Perl as well as CGI, lying down within three levels. These levels are composed of modules which tend to be developed whilst maintaining their own self-employed behavior and reusability.

Eclipse (*http://www.eclipse.org/*), ActivePerl (*http://www.activestate.com/activeperl*), WAMP: Windows Apache MySQL PHP (*http://www.wampserver.com*) and Mozilla (www.mozilla.org/) were used to create and test the VRSAP. Whole working environment is shown in figure 2.2.

Eclipse is an SDE (software advancement environment) mainly implemented within the Java. It also offers a program with flexible working environment, which allows a greater number of alternative plug-ins. This particular element of Eclipse has the ability to perform development along with different languages besides java for example Perl, PHP, Python, R, Ruby and others. Presently we utilized eclipse classic 3.7.2. To operate Perl scripts a plug-in, named EPIC was installed.

Figure 2.2: Working environment of VRSAP

To read the Perl script, active Perl from the ActiveState was utilized to apply and implement the application. There are lots of active Perl versions available from ActiveState like: 5.12.1.1201, 5.8.9.826, and 5.10.1.1007, we used the 5.10.1.1007 because of its compatibility with the Bioperl. WAMP (stands for the Apache, MySQL and one of PHP, Perl or Python) tend to be individually acting program system that makes use of the actual Microsoft windows operating systems. WAMP server has a bundle of features which supplies the actual GUI for MySQL data source for management. It also provide a server option for scripting different languages like Python or even Perl, referred to as phpMyAdmin. Firefox is a web browser which was continuously utilized to trace the outcome from our scripts.

VRSAP works in three layers (all the layers and their module's functionality are shown in Figure 2.3). Its first layer, known as the Input layer controls the factors and objects that count the overall representation, inputs, validation and lastly transferring the submitted values to the second layer. Processing involved in VRSAP is made possible by the second layer, known as the processing layer. All processes concerning SNPs, RFLP and SNPs-RFLP analysis are fulfilled in this layer. Processing layer is also called as the active layer of VRSAP. Output layer, a third layer is used to generate the results and their representations (Figure 2.3).

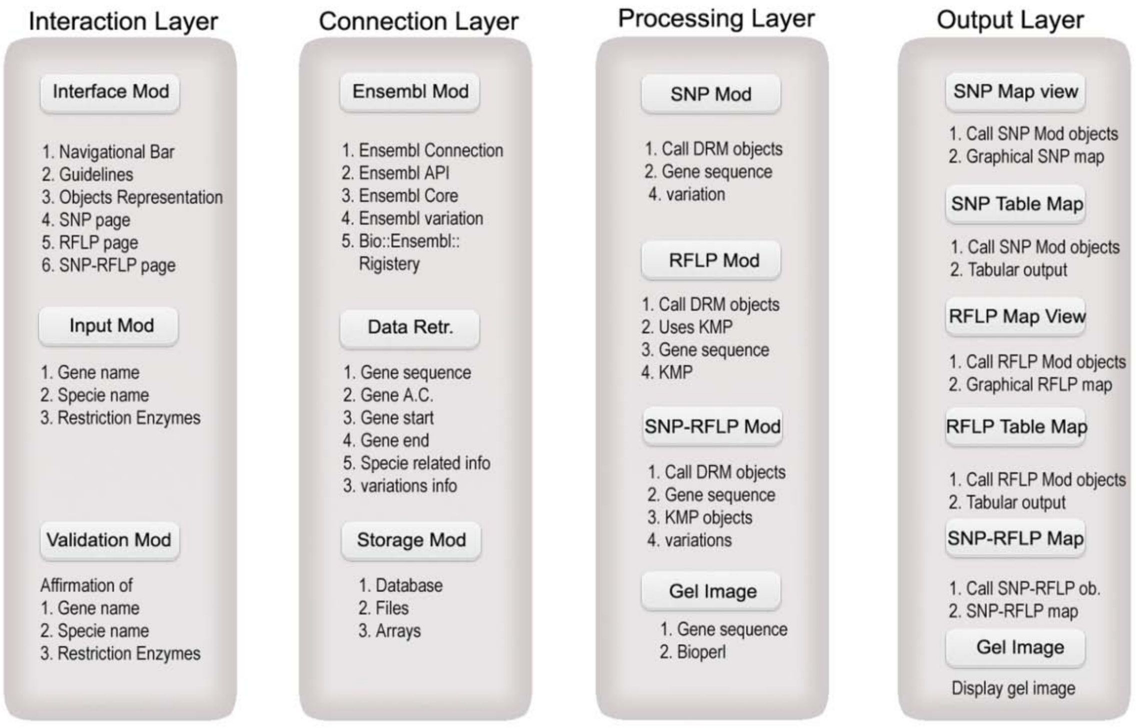


Figure 2.3: VRSAP layers and their corresponding modules

2.1 Interaction layer

Interaction layer consists of three modules, specifically User interface module, Input module and an affirmation module known as Validation module. A primary objective associated with the interaction layer is to interact with the active user after which it confirms the validation of forwarded inquiries.

2.1.1 Interface module

Primary or index web page, symbolizing the web application along with minor explanation and keeping the actual navigational bar, hyperlinks to some other supplied options is created and supported through the user interface module. Every object and its attributes (color, textual content and the last one (body table), which presents VRSAP entire look) is designed in this module. The primary goal associated with user interface module is to help the user through illustrations along with the graceful images. The navigational bar provides the user an entry to some other web pages within short time. Furthermore, textual content by means of recommendations, guidelines, information and functionality circulation for rendering the results can be quite ideal for the users. User interface module components are created and applied within the HTML and graphics are developed by utilizing the Photoshop.

2.1.2 Input module

The predicaments and jobs aimed by the users were dealt with the second module from the interaction layer that inlayed within the user interface module, known as input module. Submission choices consist of gene and specie names. These types of demands tend to be pasted within the textual content form and several multiple select boxes. Both the gene and the specie names are used for the SNP analysis, whereas for RFLP and SNP-RFLP analysis, an additional effort is required that is to check mark the interested restriction enzymes; enlisted nearer to other input options. In short, VRSAP was made interactive by introducing the input module (written in the HTML and closely connected to CGI and JavaScript scripts).

2.1.3 Input validation module

It is a process of checking and validation of inputs. Input validation module, third and the last module of interaction layer is responsible for the dealing with illegalness of submitted queries. Input validation module was written in the JavaScript and then incorporated in the HTML. This module first checks whether the gene name is present or not and also checks either the specie name is selected from the select box for a gene. In case of RFLP and a complete SNP-RFLP analysis, input validation module also confirms that the restriction enzymes are marked (chosen) or unmarked. After the validation of queries the values are passed to the next layer, connection layer (section 2.2) and the processing layer (Section 2.3).

2.2 Connection layer

VRSAP whole activity is primarily dependent on the data retrieved from the external sources such as Ensembl. For this purpose, a connection was established with external server, Ensembl by utilizing the Ensembl API. Whole process occurring in ensembl connection module was connection layer, second layer of VRSAP development.

2.2.1 Ensembl connection module

To deal with multiple queries, submitted in input module, there is a need of data which is obtained from the external server by creating a link, using Ensembl API. This electronic communication was controlled by the ensembl connection module which utilizes the Bio::Ensembl::Rigistery module of Ensembl API followed by a method call, load_registery_form_db. This method first makes sure either user and host name is provided or not. As Ensembl API came into functional form by using the Perl, to approach it Perl language is used to write the ensembl connection module. In order to fetch the data for the SNP and SNP-RFLP analysis, ensembl connection module first connects with the core source of Ensembl for the retrieval of genes related data, specified in input module then it moves towards the variation sources of Ensembl with a gene name query to rescue the variation records. For RFLP analysis, it only ties the Ensembl core source as there is no need to ally with variation source. Every time when a new input submission takes place in input module, this module creates a new connection with Ensembl.

2.2.2 Data retrieval module

After the satisfactory validation and successful connection to Ensembl, the requested and addressed records are fetched by the data retrieval module. Data retrieval module operates by using Perl and it works on ensembl connection module by creating the gene_ adopter objects and making a call to different methods of Ensembl API module through the ensemble connection module objects. For example to get the requested gene data with gene name inquiry this module invokes the external_name method with gene_ adopter objects which have the information from the ensembl connection module. Other inquires also includes variation requirements which are gained by introducing the slice_adopter. These slice adopters also depends on ensembl connection for their proper functioning. Accomplishing the job, data fetching assigned in the input module all the data and records from the Emsembl core and variation sources are now ready for storage. For the sake of this purpose various objects are transfer into the storage module.

2.2.3 Storage module

Third and last module of connection layer is storage module. Here all the objects and instances having the data as properties from the data retrieval and ensembl connection module undergo a process of storage. Storage module has different modes for storage: one of them is the creation of relation database. A relational database comprises the tables designed for data storage in columns and rows. In case of VRSAP, a database with a name “ensemble” was created in which three tables exit namely restriction, gene and variation. The restriction table contains all the information about restriction enzymes. Rebase (Roberts *et al*., 2010) is the main source of restriction enzymes for the restriction table. Restriction table stores different information such as the enzyme name, its recognition site, and its source, rebase ID etc.

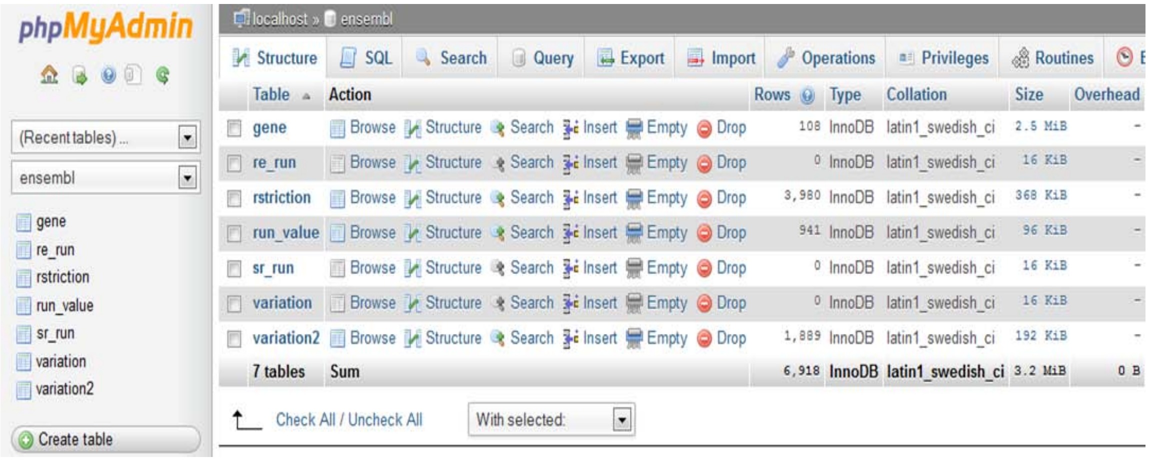


	Table	Action	Rows	Type	Collation	Size	Overhead
□	gene	Browse Structure Search Insert Empty Drop	108	InnoDB	latin1_swedish_ci	2.5 MiB	-
□	re_run	Browse Structure Search Insert Empty Drop	0	InnoDB	latin1_swedish_ci	16 KiB	-
□	rstriction	Browse Structure Search Insert Empty Drop	3,980	InnoDB	latin1_swedish_ci	368 KiB	-
□	run_value	Browse Structure Search Insert Empty Drop	941	InnoDB	latin1_swedish_ci	96 KiB	-
□	sr_run	Browse Structure Search Insert Empty Drop	0	InnoDB	latin1_swedish_ci	16 KiB	-
□	variation	Browse Structure Search Insert Empty Drop	0	InnoDB	latin1_swedish_ci	16 KiB	-
□	variation2	Browse Structure Search Insert Empty Drop	1,889	InnoDB	latin1_swedish_ci	192 KiB	-
	7 tables	Sum	6,918	InnoDB	latin1_swedish_ci	3.2 MiB	0 B

Figure 2.4: Database created with name "ensembl". Tables with names gene, re_run, restriction, run_value, sr_run, variation and variation2 are shown on left. Their structure and action are shown on right side of the image.

Second table of ensembl data base, gene is created with an objective to provide a container for all gene related data i.e. gene name, specie and gene sequence etc. Variation table stores the variation properties of gene objects. This table contains the allele and position of allele etc. WAMP server was taken into account for this database development as shown in figure 2.4.

Second mode of storage was file. Various attributes of gene object like sequence, ID and variation related objects were also stored in the file to facilitate the end user to download the data in future for further analysis. As the data retrieval module works on ensembl connection module in the same way, connection layer modules provide a platform for the processing layer modules.

2.3 Processing layer

This active layer performs a number of processing on data obtained from the connection layer to treat with SNP, RFLP and SNP-RFLP. For these purposes, sub processing module are designed which work in a well defined division of labor. For SNP analysis, a sub processing domain was designed in which different objects, created in data retrieval and storage module comparisons were made available with a purpose to mark the SNPs on the gene sequence. The comparison of slice and gene adopter's resulted into the SNP map.

Another sub processing module utilizes the KMP algorithm (Knuth *et al.*, 1977) with intentions of tracing recognition sites of restriction enzymes, marked in input module. Two functions of algorithm i.e. preprocessor and KMP chase the location of recognition site on gene sequences and store them in an array of locations. Third sub processing module joins both above mentioned sub-processing modules to achieve the SNP-RFLP related jobs with an additional constraint, SNPs in restriction sites.

Moreover, a graphical intend sub-processing module, known as gel sub processing module uses the bioperl modules to generate a graphically interactive and attractive representation of gel image. It shows the bands on a gel and DNA size makers, which user can select from the multiple options, provided in the select box.

2.4 Output layer

A series of actions from the interactive layer to processing layer resulted in the outcomes, based up on the actual inquiries and questions from the active users. Entire group of results tend to be introduced in the last layer of VRSAP that is the output layer. As in the processing layer, different sub-processing modules, depending upon the requests (made in input module) perform multiple actions on them and these efforts end up with different results. Results from sub-processing modules are presented in the output layer by output modules. Output layer have SNP Map viewer module, SNP tabular view module, RFLP Map viewer Module, RFLP tabular view module, SNP-RFLP Map viewer module and Gel display module.

2.4.1 SNP map viewer module

A module that recalls the objects which were created in the SNP sub-processing module to generate the strings associated with the gene sequences is referred as the SNP map viewer module. SNPs within sequences were differentiated by colored background. The functionality of this module was allocated through the Perl, HTML and the JavaScript. Perl statements claim of performing the required comparison of objects while HTML <span> tag colored the background where SNP occurred. JavaScript scripts result into a horizontal scroll bar which encapsulates the entire SNP map. Horizontal scroll bar prevents the whole screen from being moved during the analysis

2.4.2 SNP tabular view module

All the details explored in SNP map viewer module were represented by means of figures and names in a table through the SNP tabular view module from the output layer. Much like the SNP map, this particular module also calls the objects of SNP sub processes module and then list them in a table created by using the create table command of the HTML. The HTML table was made dynamic i.e. no matter how many records are there to show, it accommodates all. The number and the names with in the fields were printed by the Perl statements and these values comes from the SNP map objects.

2.4.3 RFLP map viewer module

RFLP sub processing module in the processing layer forms the objects, containing the RFLP information. Pattern locations are being called in this module by a Perl script to generate the RFLP map. Like SNP map viewer module, the actual power of the component is really a combination of HTML, JavaScript and Perl. Exactly the horizontal slide bar was also added in this module to relieve the string visualization.

2.4.4 RFLP tabular view module

Analogous to SNP tabular view module, RFLP tabular view module signifies the entire RFLP map information's into number and names within a table, produced from the combined endeavor of Perl and HTML by calling the objects of RFLP sub processing module.

2.4.5 SNP-RFLP map viewer module

SNP-RFLP Map viewer module makes use of the exact approach as discussed in the SNP map viewer and RFLP map viewer. It calls the objects SNP-RFLP module in order to fabricate the SNP-RFLP map. This module is yet a blend associated with Perl and HTML.

2.4.6 Gel output module

Gel output module displays the gel image developed within Bimode having an extension of PNG. The Gel image is created using the Perl at the server end and is displayed by the HTML <img> tag.

3. Results

Present study resulted into VRSAP, which is capable of conducting three different types of analyses. It contains independent applications which include SNP, RFLP and SNP-RFLP analyses. A graphical and hypothetical gel image is also incorporated in RFLP and SNP-RFLP. Overview of VRSAP output is shown in figure 3.1.

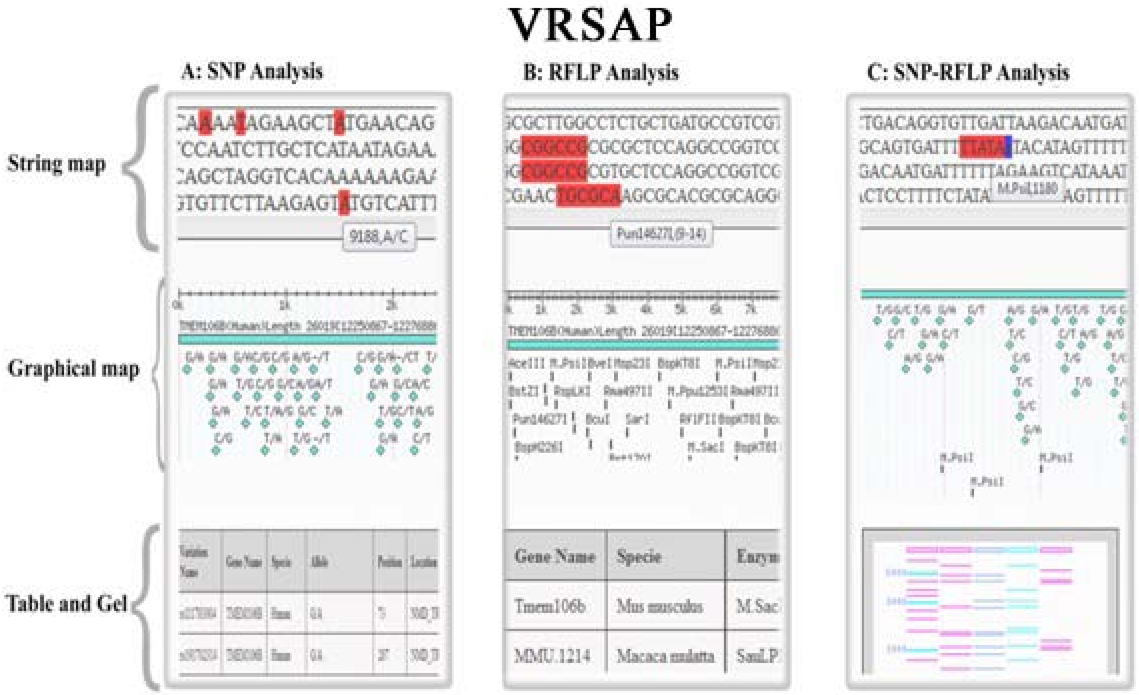


Figure 3.1: VRSAP output overview, enlists all the analysis output to depict a picture of results. (A) SNP analysis shows the components of SNP module which includes the string map, graphical map (variant image) and table, enlisting the gene and its variation. (B) RFLP analysis reveals the segments for the RFLP module which contains the string map, graphical map (restriction image) and the restriction table which stores the data related to restriction enzymes. (C) SNP-RFLP analysis represents the SNP-RFLP module that also comprises string and graphical map (VAR image). Third segment of SNP-RFLP module is a gel image.

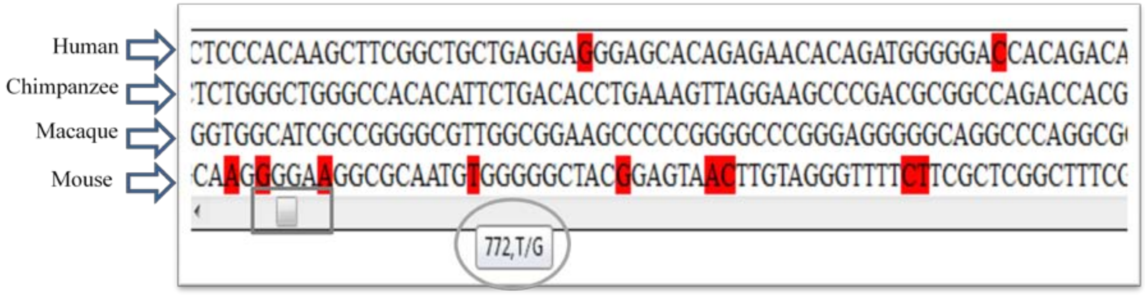
Each type of analysis comprises different sub segments. These segments are string map, graphical image and table or gel. These analyses components work independently; therefore their results are discussed separately.

3.1 SNP analysis

With an objective to provide a visual representation of SNPs among paralogs, orthologs or population, SNP analysis module was created. Results from the SNP analysis module comprise three sub segments: (1) String map for all sequences, (2) Graphical map and (3) Table, which enlists the gene and its variation attributes. All these three segments (Figure 3.1A) are elaborated below.

3.1.1 String map

String map explores the SNPs, which are located in gene sequences by highlighting the target (SNPs) nucleotides in red background. The generated string map is a collection of gene sequences which were submitted by the user in the input module of interacting layer. In string map SNPs are represented by red color. Figure 3.2 demonstrates the string map for human *HR* gene orthologs in chimpanzee (*HR*), macaque (*HR*) and Fruitfly (*JHDM2*). One distinguishing characteristic of string map is that it enlists the whole gene sequences, which could be viewed from the very first nucleotide to the last one by using scroll bar to make a comparative analysis of SNPs. Other distinctiveness linked with the string map is that on mouse over at SNP location, it pops up a small box, listing the SNP position and the replaced allele string (as shown in figure 3.2 by a light grey circle which clearly illustrates the position 772 and the SNP T/G). Nucleotides other than the SNPs in gene sequence are popped up, on mouse over their genomic locations. Only SNPs exhibit colored background, which prevent the designed application from being so messy. As the gene length is a variable factor among gene families, for example human *Gli* gene length is 12,127 nucleotides, while the *Gli3* gene is up to 2, 76,921 nucleotides in length.



```
CTCCCACAAGCTTCGGCTGCTGAGGAGGGAGCACAGAGAACACAGATGGGGACCACAGACA
TCTGGGCTGGGCCACACATTCTGACACCTGAAAGTTAGGAAGCCCGACGCGGCCAGACCACG
GGTGGCATCGCCGGGGCGTTGGCGGAAGCCCCCGGGGCCCGGGAGGGGGCAGGCCCAGGCG
CAAGGGGAAGGCGCAATGTGGGGGCTACGGAGTAACTTGTAGGGTTTTCTTCGCTCGGCTTTCC
```

Figure 3.2: String map for human *HR* orthologs. First lane of string is for the human *HR* gene sequence, second: *HR* gene sequence of chimpanzee, third: *HR* gene sequence for macaque and forth lane accounts for fruitefly (*JHDM2)* gene sequence. Red colored nucleotides are SNPs and mouse over effect shows the position and the SNP which is circled by grey circle. Scroll bar is enclosed by dark grey rectangle.

A normal computer screen is able to show up to 120 nucleotides in a row with screen resolution of 1280×800. As the sequence length may be up to hundred thousand, the inquirer has to scroll the screen towards right and left to fully examine the variation patterns. To avoid the full screen movement, we created a scroll bar specifically for the string map viewer to move it towards both sides (Figure 3.2).

3.1.2 Graphical map

The second component of SNP analysis is the graphical representation of all the SNPs occurring in the gene sequences (Figure 3.3). This graphically intended segment displays an image created at the run time in accordance to the user's queries. The image was named as variant image which can be sub-divided into three different components: namely (a) arrow or scaling track, (b) sequence track and (c) SNPs track. Scaling track (shown by the light blue arrow in Figure 3.3), scales the whole image according to the largest gene sequence length and it is mostly scaled in the kilo base pairs. Whole scaling was marked up by an arrow which consisted of two terminals.

The first one is head and second is tail. Arrow’s head symbolizes the 3-prime end of the gene sequence while the 5-prime end sequence was made noticeable by an arrow tail. Moreover, arrow or scaling segment was padded from right and left margins to define the boundaries for variant image (illustrated by a sky blue arrow in figure 3.3). Second segment of variant image is its sequence bar (symbolized by purple arrow in figure 3.3), which is a box equal to the sequence length (shown in cyan color in figure 3.3, just beneath the scaling track for the first sequence). In addition to graphical representation of gene sequence, another feature attached with this segment is to display the gene name, specie name, gene sequence length and starting and ending positions in the genome above the sequence box bar (Figure 3.3). The last track of graphically proposed image is SNPs track (signified by light reddish arrow in figure 3.3D), illustrating the SNP as diamond shape entitled with the allele string, which replaces the sequence nucleotide at the specific location.

To make interpretation easy of variant image; in addition to scaling, sequence and SNP track, a grid as a background was provided (grey arrow in the figure 3.3). The number of tracks in the variant image depends upon the input module of the interacting layer. The scaling, sequence, and SNP tracks were coupled with the cyan colored grid counts for each sequence. Moreover, all the above mentioned tracks repeated themselves depending upon the number of genes used as a query from the active user (Figure 3.3). The number of gene names, as query can be increased while keeping an eye on the gene sequence length and external server data fetching limit.

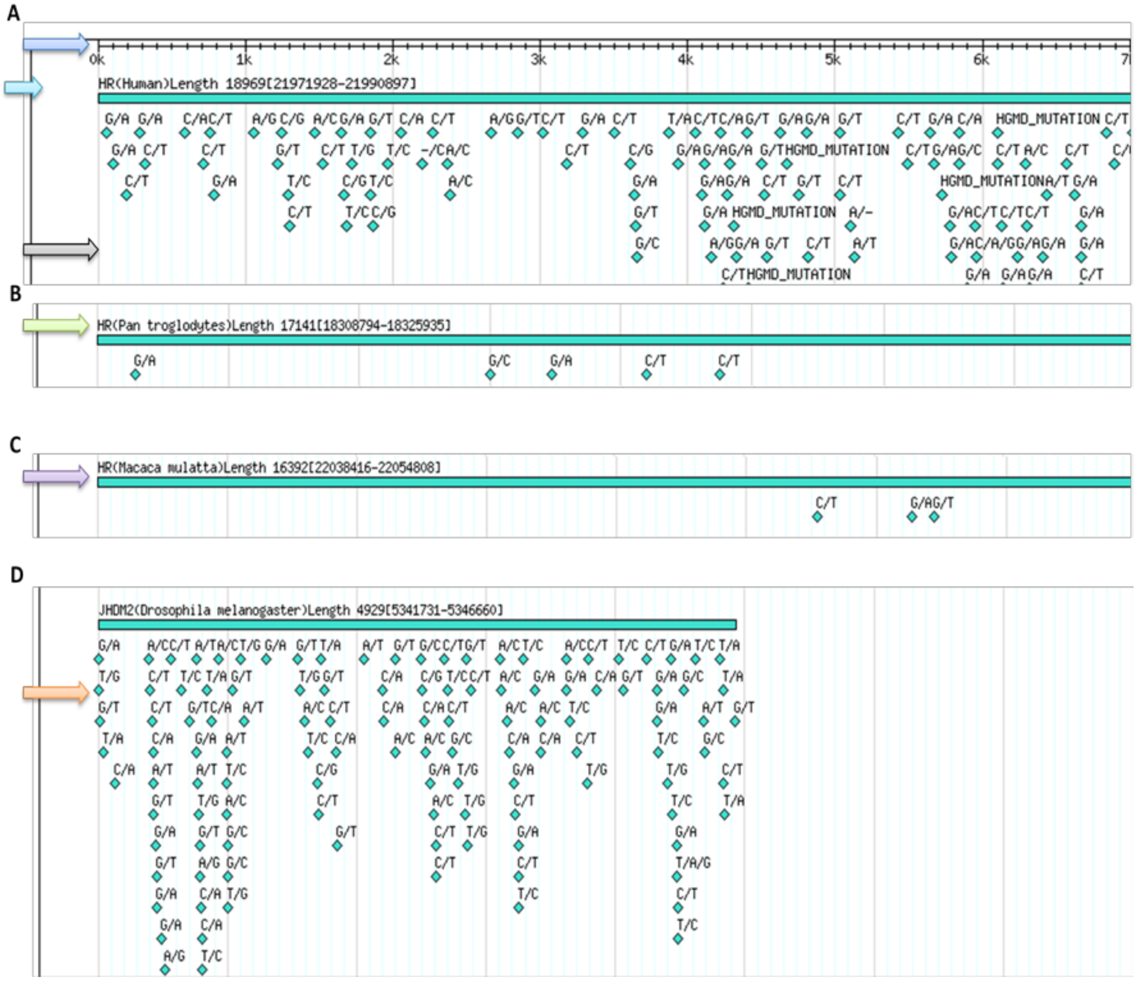
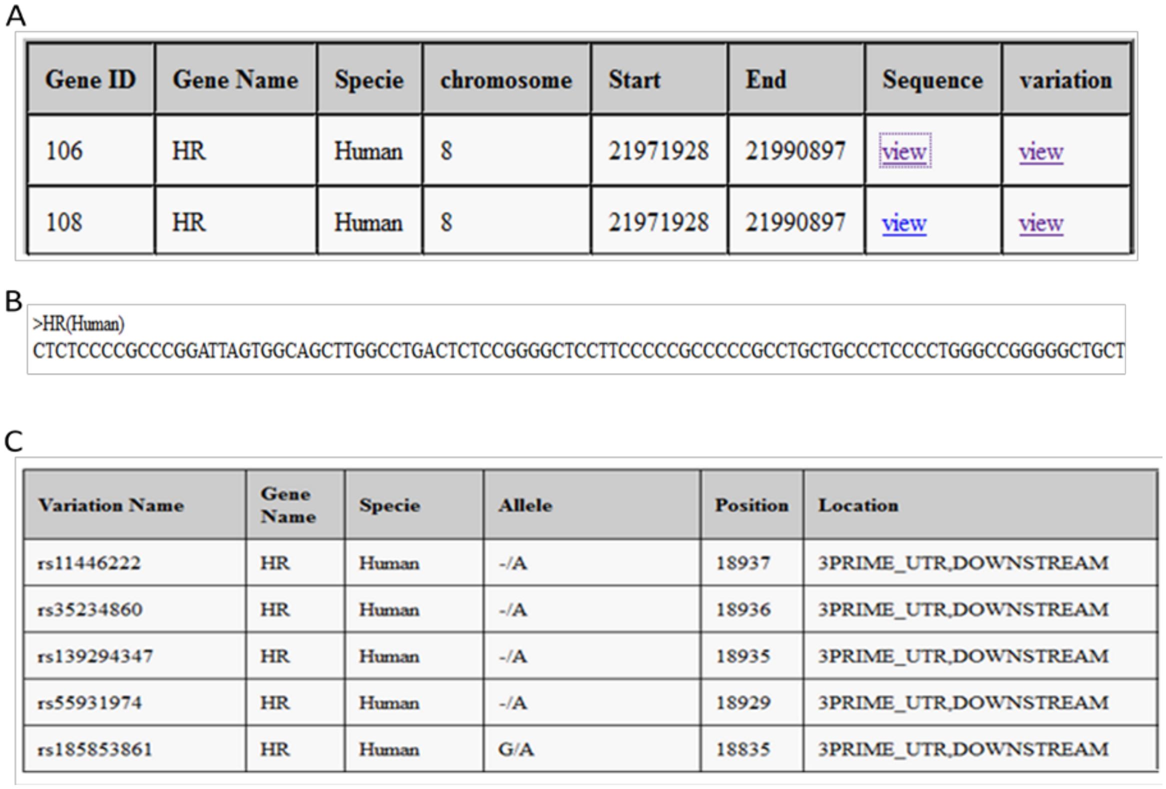

A
0k
1k
2k
3k
4k
5k
6k
7
HR(Human)Length 18969[21971928-21990897]
HGMD_MUTATION
B
HR(Pan troglodytes)Length 17141[18308794-18325935]
G/A
G/C
G/A
C/T
C/T
C
HR(Macaca mulatta)Length 16392[22038416-22054808]
C/T
G/AG/T
D
JHDM2(Drosophila melanogaster)Length 4929[5341731-5346660]

Figure 3.3: SNP map for *HR* orthologous. (A) Shows the scaling, padding and grid pointed by light blue, sky blue and grey arrows. (B) Presents the gene name by light green. (C) Depict the gene sequence (graphical representation), marked by purple arrow. (D) Illustrates the SNPs tracks, made noticeable by light reddish and arrow.

3.1.3 Table, enlisting the gene sequence and variation data

The table fields contain the gene sequence and variation associated facts and figures, as demonstrated in string map (Figure 3.2) and variant image (Figure 3.3).

Table, (Figure 3.4A), enlists the gene name, specie name, gene length, gene start, gene end and variation related information. Hyperlinks created for the Sequence column are used to fetch the gene sequence in the FASTA format (Figure 3.4B). Similarly, the hyperlinks in the Variation column could be utilized to explore the variation attributes of a gene (Figure 3.4C). The clear and full representation of SNP table, sequence and variation table is shown in figure 3.4.



Gene ID	Gene Name	Specie	chromosome	Start	End	Sequence	variation
106	HR	Human	8	21971928	21990897	view	view
108	HR	Human	8	21971928	21990897	view	view

```
>HR(Human)
CTCTCCCCGCCCGGATTAGTGGCAGCTTGGCCTGACTCTCCGGGGCTCCTTCCCCCGCCCCCGCCTGCTGCCCTCCCCTGGCCGGGGGCTGCT
```

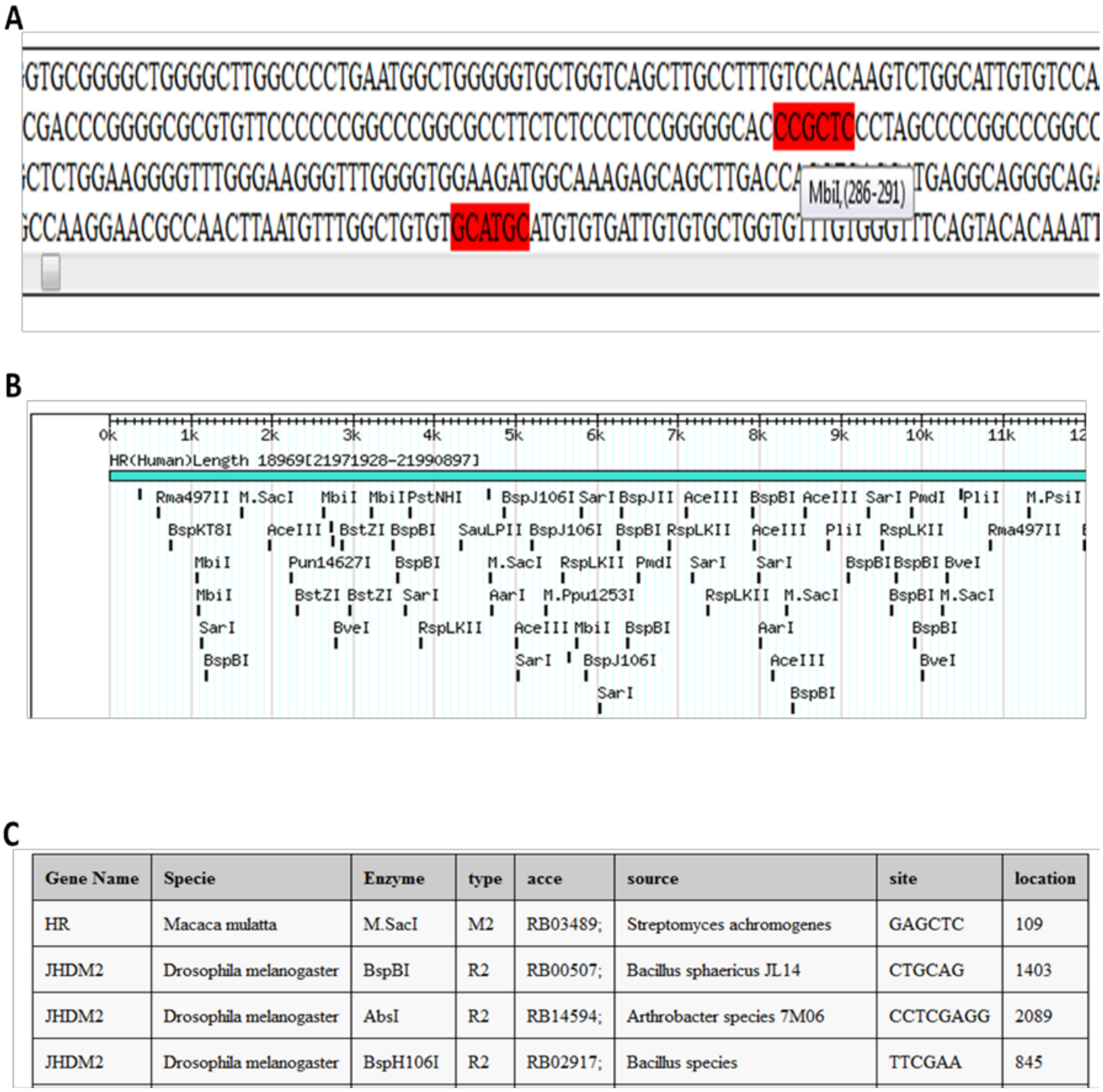
Variation Name	Gene Name	Specie	Allele	Position	Location
rs11446222	HR	Human	-/A	18937	3PRIME_UTR,DOWNSTREAM
rs35234860	HR	Human	-/A	18936	3PRIME_UTR,DOWNSTREAM
rs139294347	HR	Human	-/A	18935	3PRIME_UTR,DOWNSTREAM
rs55931974	HR	Human	-/A	18929	3PRIME_UTR,DOWNSTREAM
rs185853861	HR	Human	G/A	18835	3PRIME_UTR,DOWNSTREAM

Figure 3.4: Table representing the gene and its variations. (A) Shows the primary table which contains the gene related information like gene name, specie name, gene length, gene start, gene end. By clicking the view in Sequence column as shown in section A, a new file opened which gives the gene sequence in FASTA format as shown in B section. (B) Exemplify the FASTA format sequence for human *HR* gene. (C) Shows the table which contains the variation or SNPs relevant to a specific gene which can be viewed by clicking on view, written in variation column of primary table as shown in section A of this image. This table enlists the variation attributes like variation name, allele, allele position and location.

3.2 RFLP analysis

RFLP (Restriction fragment length polymorphism) analysis (Figure 3.5) across the population, species or in the paralogs is another outstanding feature of VRSAP to facilitate the user. RFLP analysis patterns show analogy towards the SNP analysis as it also explains the string, graphical image and the tabular output with an addition of gel image. In RFLP analysis, the restriction sites are traced.

Similar to SNP analysis string map, RFLP string map exemplifies the gene sequence nucleotides by highlighting the region of restriction sites (Figure 3.5A). These restriction sites were colored red. Mouse over effect for colored regions of RFLP string map, popped up the restriction enzyme name and restriction site starting and ending position. A scroll bar, identical to SNP analysis encapsulates the string map. As far as RFLP image is concerned it shows full resemblance with variant image with respect to scaling, and sequence track; however, SNP track in RFLP image is replaced by the RFLP track, small rectangles instead of diamonds tagged with the enzyme name (Figure 3.5B). RFLP scaling pattern, padding, and reiteration of all tracks pursued the same roles as implemented in the variant image. RFLP table includes the entities like gene, species, enzyme name and enzyme attributes such as enzyme type, rebase I.D., source, restriction site and location (Figure 3.5C). A cross reference with rebase (Roberts *et al.*, 2010) database was created through hyperlinks. Along with RFLP string map, image and tabular output, an additional feature was included in the form of a hypothetical gel image which was sketched and introduced by the RFLP analysis (Figure 3.6). The left most gel image lane represents ladder. To make the gel image more users friendly, a number of ladders were incorporated which can be opted from the drop down select box. Other lanes of gel image are gene sequences whose numbers are equivalent to the number of gene names used as input.



A

GTGCGGGGCTGGGGCTTGGCCCCTGAATGGCTGGGGGTGCTGGTCAGCTTGCCTTTGTCCACAAGTCTGGCATTGTGTCCA
CGACCCGGGGCGCGTGTTCCCCCCGGCCCGGCGCCTTCTCTCCCTCCGGGGGCACCCGCTCCCTAGCCCCGGCCCGGCC
CTCTGGAAGGGGTTTGGGAAGGGTTTGGGGTGGAAGATGGCAAAGAGCAGCTTGACCA MbiI,(286-291) TGAGGCAGGGCAGA
CCAAGGAACGCCAACTTAATGTTTGGCTGTGTGCATGCATGTGTGATTGTGTGCTGGTG TTCAGTACACAAAT

B
0k 1k 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k 12
HR(Human)Length 18969[21971928-21990897]
Rma497II M.SacI MbiI MbiIPstNHI BspJ106I SarI BspJII AceIII BspBI AceIII SarI PmdI PliI M.PsiI
BspKT8I AceIII BstZI BspBI SauLPII BspJ106I BspBI RspLKII AceIII PliI RspLKII Rma497II
MbiI Pun14627I BspBI M.SacI RspLKII PmdI SarI SarI BspBI BspBI BveI
MbiI BstZI BstZI SarI AarI M.Ppu1253I RspLKII M.SacI BspBI M.SacI
SarI BveI RspLKII AceIII MbiI BspBI AarI BspBI
BspBI SarI BspJ106I AceIII BveI
SarI BspBI
C

Gene Name	Specie	Enzyme	type	acce	source	site	location
HR	Macaca mulatta	M.SacI	M2	RB03489;	Streptomyces achromogenes	GAGCTC	109
JHDM2	Drosophila melanogaster	BspBI	R2	RB00507;	Bacillus sphaericus JL14	CTGCAG	1403
JHDM2	Drosophila melanogaster	AbsI	R2	RB14594;	Arthrobacter species 7M06	CCTCGAGG	2089
JHDM2	Drosophila melanogaster	BspH106I	R2	RB02917;	Bacillus species	TTCGAA	845

Figure 3.5: RFLP analysis for *HR* gene. (A) String map for human *HR* gene orthologous including chimpanzee, macaque and fruitefly. First lane is the human *HR* gene sequence. Second lane shows the chimpanzee *HR* gene sequence. Third and fourth lanes depict the macaque (*HR*) and fruitefly (*JHDM2*) gene sequences. Two restriction site with red colored background, their enzyme name and location are shown. (B) Shows the restriction image which is the graphical representation of RFLP analysis for *HR* gene. (C) Tabular output, listing the enzyme and gene related attributes. The acce column of table presents the values for rebase ID.

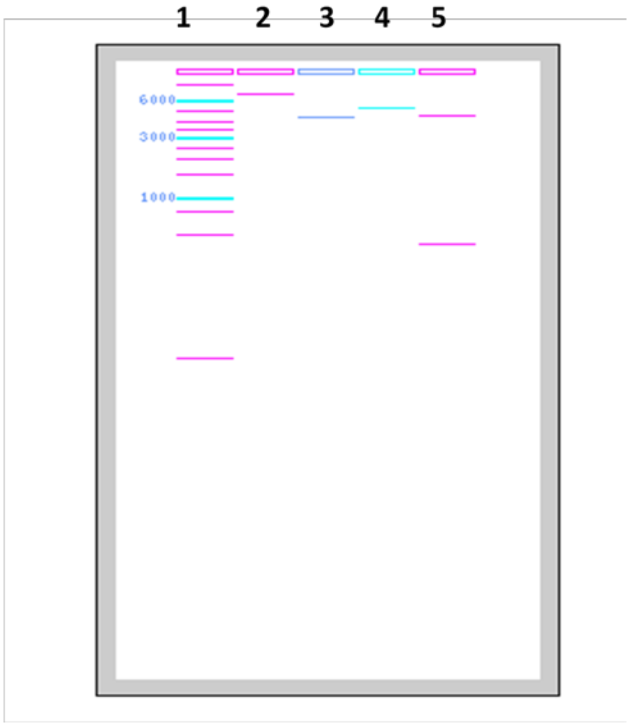


Figure 3.6: Gel image of RFLP analysis for *HR* gene. First lane of gel image represents the ladder. 2. Is for human *HR* gene. 3, 4 and 5 are for chimpanzee, macaque and mouse. Bands for human *HR* and its orthologoues genes are dispersed without any symmetry.

3.3 SNP-RFLP analysis

Final and the third analysis option provided by the VRSAP combine both SNP and RFLP analysis into a single, called the SNP-RFLP analysis (Figure 3.7). This section is complied with the same rules for patterning the tracks i.e. string map, var image (Figure 3.7B) and gel image (an analogy with the RFLP analysis section). However, the SNP-RFLP analysis is different from the rest of scrutiny segments in many ways. For example string map focuses at SNPs that occur in the restriction sites of different restriction enzymes. This type of inspection aims for two targets: first for the restriction sites and second for the SNPs in those restriction sites. In the first step, SNP-RFLP colored the nucleotides background red where it found the restriction site. Coloring the nucleotide in red is similar to RFLP analysis.

In the second step, it locates the SNPs in these reported restriction sites. The traced SNPs in the SNP-RFLP were colored blue (Figure 3.7). Thus, the approach to position the SNPs is similar to the one that was used in SNP analysis. It's pop up box combines the traits from both SNP and RFLP analyses, staging the location, restriction enzyme name and the allele string for blue back grounded nucleotide (Figure 3.7A). This section is also confined by horizontal scroll bar. This amalgam module revealed some deviation in case of graphical image over the already discussed sections. It showed all the SNPs occurring in the gene sequences and the accounted the restriction sites. The reason behind this minor deviation is that it presents the overview of density of SNPs and restriction sites in gene regions. Other arrangements of SNP-RFLP image settle the same criteria as discussed in SNP and RFLF analysis. The gel image in this case was inherited from the RFLP analysis.

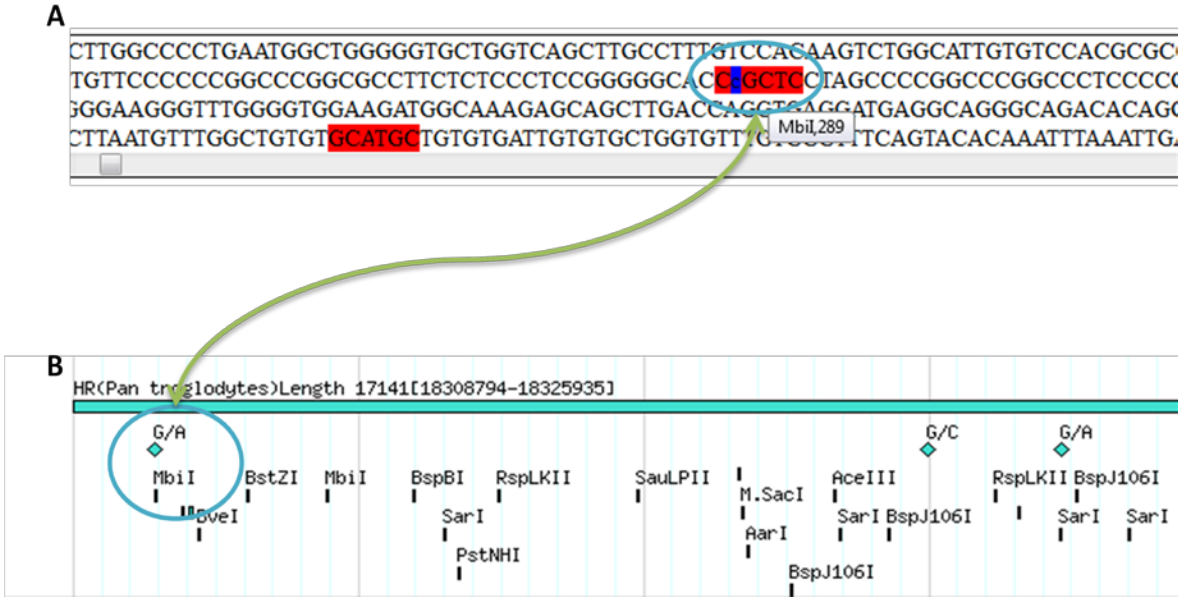

A
CTTGGCCCCTGAATGGCTGGGGGTGCTGGTCAGCTTGCCTTTGTCCACAAGTCTGGCATTGTGTCCACGCGC
TGTTCCCCCCGGCCCGGCGCCTTCTCTCCCTCCGGGGGCACCcGCTCCTAGCCCCGGCCCGGCCCTCCCCC
GGAAGGGTTTGGGGTGGAAGATGGCAAAGAGCAGCTTGACCAGGTGAGGATGAGGCAGGGCAGACACAGC
CTTAATGTTTGGCTGTGTGCATGCTGTGTGATTGTGTGCTGGTGTT
TCAGTACACAAATTTAAATTG
MbiI,289
B
HR(Pan troglodytes)Length 17141[18308794-18325935]
G/A
G/C
G/A
MbiI
BstZI
MbiI
BspBI
RspLKII
SauLPII
M.SacI
AceIII
RspLKII
BspJ106I
BveI
SarI
PstNHI
AarI
SarI
BspJ106I
BspJ106I
SarI
SarI

Figure 3.7: SNP-RFLP analysis for *HR* gene. (A) String map for *HR* gene orthologous where a SNP, G/A creates a new restriction site in chimpanzee which is encircled by the blue circle. (B) Presents the restriction image. Blue circle shows the graphical representation of newly created restriction site. A and B are linked by an green arrow

3.4 Analysis of transmembrane protein 106

All the above discussed analyses were performed by the VRSAP on the transmembrane protein 106B family which contains three members namely *TMEM106A, TMEM106B and TMEM106C* in human. All members of this family were also included in analysis in other species which including chimpanzee, macaque and mouse (Table 3.1).

3.4.1 SNP map for *TMEM 106B*

SNPs for human *TMEM106B* gene and its orthologous *TMEM106B, MMU and TMEM106B* in *pan troglodytes* (chimpanzee), *macaca mulatta* (macaque) and mouse (*mus musculus*) respectively were mapped. The respective length and SNPs of these genes are shown in table 3.1. It can be verified that mutational rate is much greater in human *TMEM106B* as compared to its orthologous. This is also depicted in (Figure 3.8). This high mutational rate in human might be due to the huge intergenic DNA sequences among coding regions. The mouse (*mus musculus*) *TMEM106B* gene showed more variation patterns than *maca mulatta* and *pan troglodytes* but lesser than human as mentioned earlier.

Specie	**Gene**	**Length**	**SNPs**	**Avg.**
Human	*TMEM106A*	7695	99	78
Chimpanzee	*TMEM106A*	7641	0	0
Macaque	*TMEM106A*	5056	1	5056
Mouse	*TMEM106A*	9546	46	207
Human	*TMEM106B*	26019	618	42
Chimpanzee	*TMEM106B*	21792	8	2724
Macaque	*MMU.1214*	22012	6	3668
Mouse	*TMEM106B*	19510	84	232
Human	*TMEM106C*	5309	103	52
Chimpanzee	*TMEM106C*	5339	2	2670
Macaque	*TMEM106A*	5056	1	5056
Mouse	*TMEM106A*	9546	46	207

Table 3.1: Listing the gene names, length, SNPs and their average in corresponding species

Macaque (*Macaca mulatta*) *TMEM106B* behaved more strictly towards the SNP variation pattern as compared to the other primates and mouse (*mus musculus*). In human *TMEM106B*, every region of 1kbp contained the SNPs. In case of *TMEM106B*, an orthologous of human *TMEM106B* in chimpanzee (*pan troglodytes*), the region 1-1369 nucleotides have no SNPs where as human gene accumulated more than thirty SNPs in this gene segment. In mouse, only three SNPs occurred in 1-1369 nucleotide regions. For the above mentioned region macaque had no SNPs. Macaque *TMEM106B* reported its first SNP at 3957 nucleotide. From 12 to 14 kbp, human *TMEM106B* showed a cluster of SNPs while chimpanzee and macaque genes exhibited one and mouse showed only three SNPs.

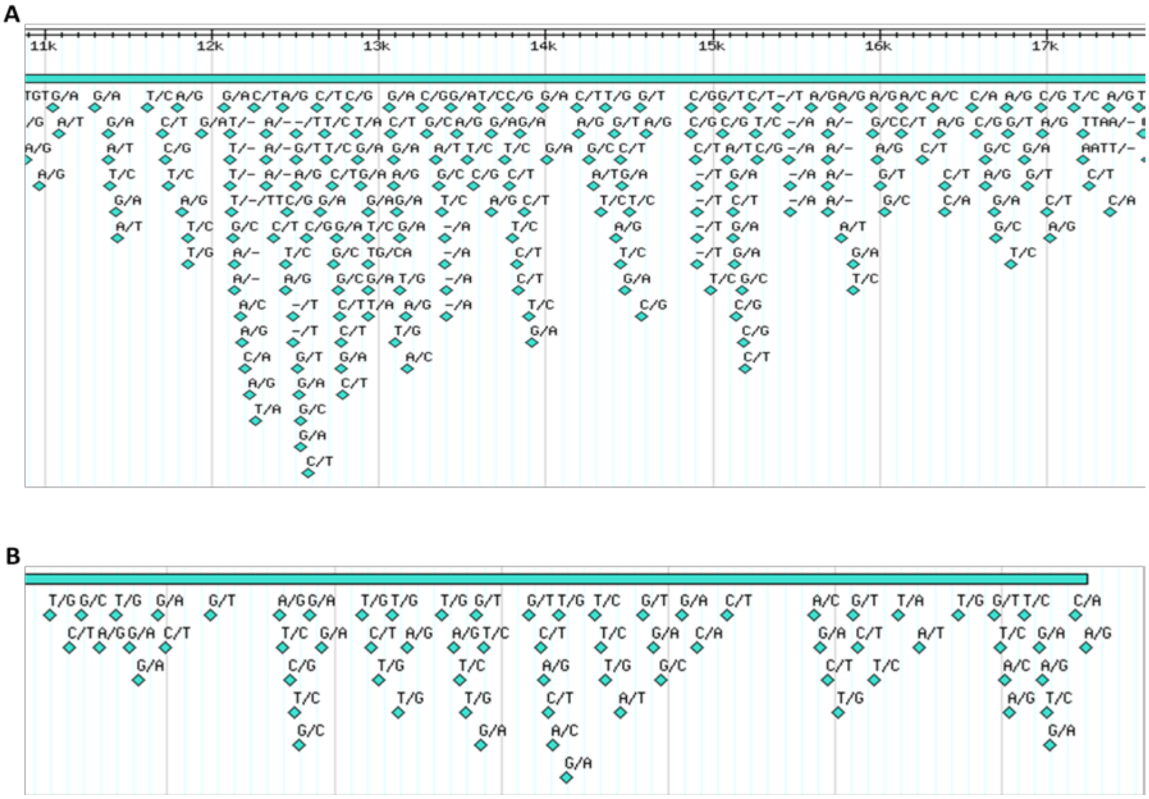
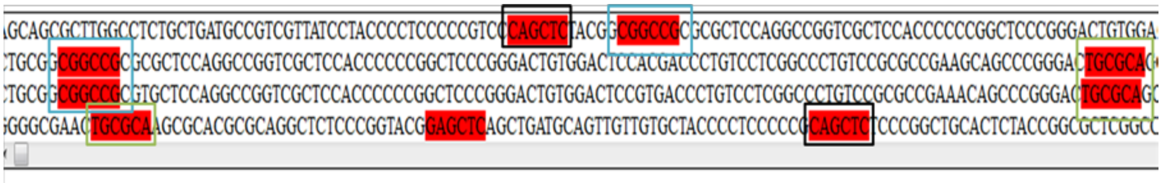


Figure 3.8: Variant image for *TMEM106B* gene. (A) Shows the SNPs high density region for human while (B). Shows the mouse region where SNPs occurred in clusters.

3.4.2 RFLP map for *TMEM106B*

RFLP analysis of *TMEM106B, TMEM106B, MMU.1214* and *TMEM106B* of human, chimpanzee, macaque and mouse respectively was carried out and their string map is shown in figure 3.9. Irrespective to strict behavior of these genes towards the SNPs, all of them represented the clusters of restriction sites for restriction enzyme (blue rectangle in figure 3.9). Chimpanzee and macaque exhibited a unique mode for restriction enzymes by having almost the same RFLP map, which can also be analyzed from the gel image (Figure 3.10).

The restriction site of *BStZI* restriction enzyme was common in all primates including human, chimpanzee and macaque. *BStZI* site occupied the same locations in chimpanzee and macaque; while in human, it was found at position 63, after the *ACeIII* enzyme restriction site. Almost a similar pattern of restriction sites was observed in human, chimpanzee and macaque as compared to mouse.



```
AGCAGCGCTTGGCCTCTGCTGATGCCGTCGTTATCCTACCCCTCCCCCGTCCCAGCTCTACGGCGGCCGCGCGCTCCAGGCCGGTCGCTCCACCCCCCGGCTCCCGGGACTGTGGA
CTGCGGCGGCCGCGCGCTCCAGGCCGGTCGCTCCACCCCCCGGCTCCCGGGACTGTGGACTCCACGACCCTGTCCTCGGCCCTGTCCGCGCCGAAGCAGCCCGGGACTGCGCAG
CTGCGGCGGCCGCGTGCTCCAGGCCGGTCGCTCCACCCCCCGGCTCCCGGGACTGTGGACTCCGTGACCCTGTCCTCGGCCCTGTCCGCGCCGAAACAGCCCGGGACTGCGCAGC
GGGGCGAACTGCGCAAGCGCACGCGCAGGCTCTCCCGGTACGGAGCTCAGCTGATGCAGTTGTTGTGCTACCCCTCCCCCGCAGCTCTCCCGGCTGCACTCTACCGGCGCTCGGCC
```

Figure 3.9: String map for *TMEM106B*. Blue colored rectangle shows the restriction sites which are common in human, chimpanzee and macaque. Green colored rectangles presents the restriction sites common in chimpanzee, macaque and mouse. Black colored rectangles show the restriction sites common in human and mouse.

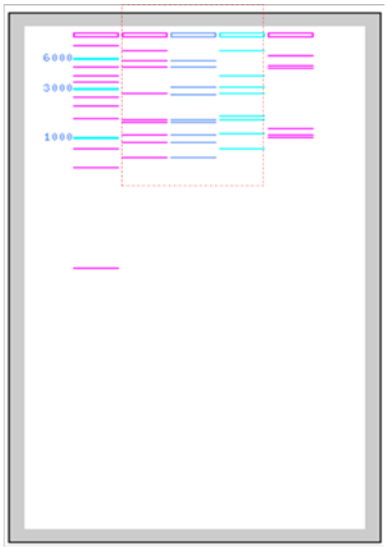


Figure 3.10: Gel image for *TMEM106B*. The red rectangle enclosed the human, chimpanzee and macaque gene bands (from left to right). First lane is ladder and fifth lane depicts mouse gene bands. Within the red colored rectangle symmetry of band patterns was observed.

Table generated in response to queries (gene and specie name and restriction enzymes) showed that restriction enzyme of type R_2 cuts more frequently than others. M_2 types of restriction enzymes cut all the sequences only five times. A hypothetical gel image for *TMEM106B* orthologous depicted the bands ranged from 800bp to more than 6000bp (Figure 3.10). The ladder used for this gel image was 1kbp. In gel image, the first lane is for the ladder. Red colored soft rectangle (Figure 3.10) shows the human, chimpanzee and macaque gene bands. Because of analogous restriction map, human and chimpanzee showed almost similar patterning of gene sequence bands. Gel bands for mouse *TMEM106B* exhibited a dispersion pattern as no symmetry was observed in the string map. Thus conservation of restriction sites in different species that were separated out in about 7 million years ago through speciation, open a new research window and requires attentions to answer the question like; why these restriction sites are conserved? What parameters are involved in such conservation?

3.4.3 SNP-RFLP map for *TMEM106B*

After localizing the SNPs and restriction sites in *TMEM106B*, SNP-RFLP analysis was carried out. *TMEM106B* analysis exhibited a more conclusive and unique patterning. The restriction enzyme with name *M.Psil* (query for SNP-RFLP analysis) was capable to cut *TMEM106B*, orthologous, when it found TTATAA (restriction site). During the inspection of SNP-RFLP for *TMEM106B*, it was observed that the restriction pattern TTATAA for *M.Psil* was restored by a SNP (an allele string A/G). In chimpanzee, same restriction site was also presented at a position, upstream to the newly created restriction site in human. String map and var image for *TMEM106B* are shown in figure 3.11, while hypothetical gel image is depicted in figure 3.12. From the gel (Figure 3.12) it was clear that human and chimpanzee displayed the same gene sequence band patterns. It can be argued that to maintain this patterning a new site was created in human. Thus this data clearly nullified the statement that mutation is a random and uncertain phenomenon.

The gel image showed that the similarity in band patterning was due to the restoration of a restriction pattern (TTATAA) in human. TTATAA is the recognition site for the *M.Psil* restriction enzyme. From the gel, it was observed that the similar patterning of gel bands between human and chimpanzee was maintained by the SNP (A/G).

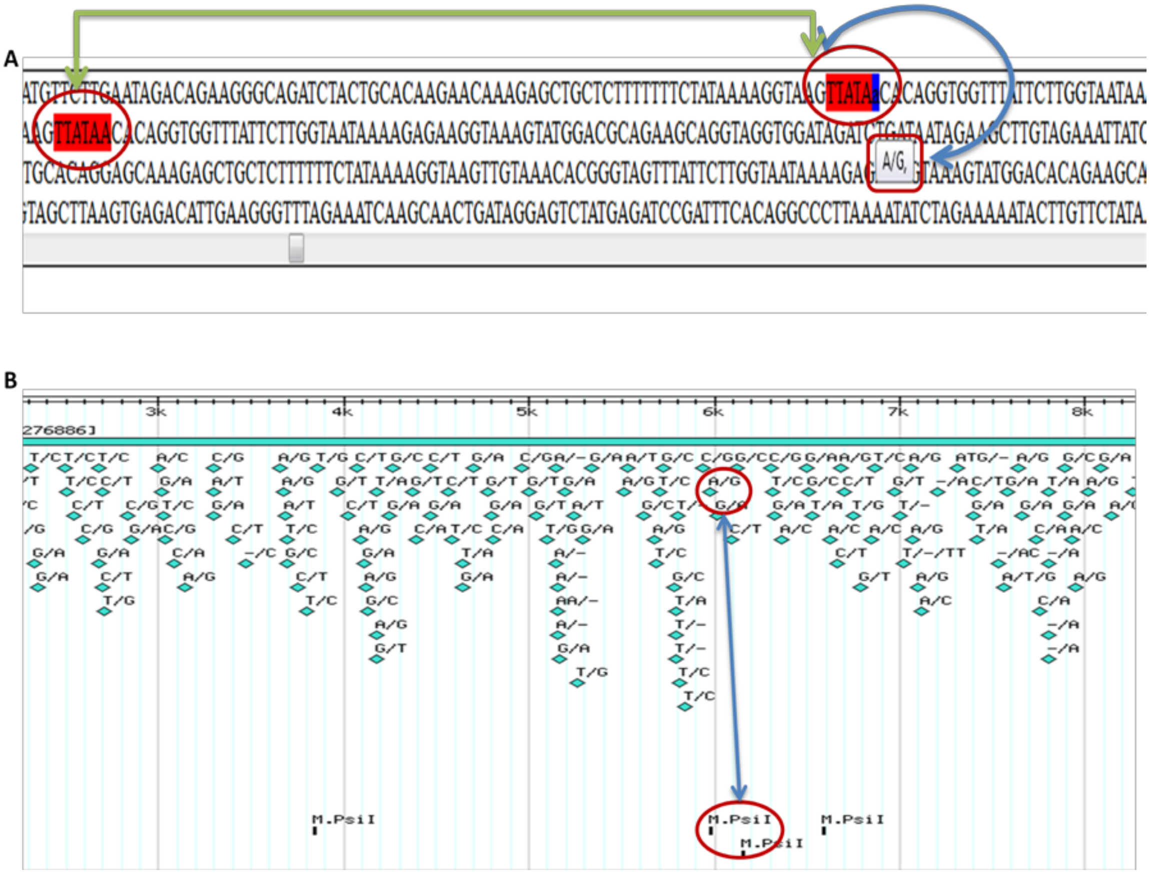

A
ATGTTCTTGAATAGACAGAAGGGCAGATCTACTGCACAAGAACAAAGAGCTGCTCTTTTTTTCTATAAAAGGTAAGTTATAaCACAGGTGGTTTATTCTTGGTAATAA
AAGTTATAACACAGGTGGTTTATTCTTGGTAATAAAAGAGAAGGTAAAGTATGGACGCAGAAGCAGGTAGGTGGATAGATCTGATAATAGAAGCTTGTAGAAATTATC
TGCACAGGAGCAAAGAGCTGCTCTTTTTTCTATAAAAGGTAAGTTGTAAACACGGGTAGTTTATTCTTGGTAATAAAAGAG
A/G,
GTAAAGTATGGACACAGAAGCA
GTAGCTTAAGTGAGACATTGAAGGGTTTAGAAATCAAGCAACTGATAGGAGTCTATGAGATCCGATTTCACAGGCCCTTAAAATATCTAGAAAAATACTTGTTCTATA
B
3k
4k
5k
6k
7k
8k
276886]
A/G
M.PsiI
M.PsiI
M.PsiI
M.PsiI

Figure 3.11: SNP-RFLP map for *TMEM106B*. (A) String map shows where a restriction site is created in human, SNP and the restriction site in chimpanzee. Both restriction sites and the SNP are enclosed by red circles. Link between the restriction sites are shown with green arrow while blue arrow connects the replaced SNP and the newly created restriction site. (B) Var image provides the graphical representation, where the SNP and the enzyme are circled by red circles, connected by a blue color arrow.

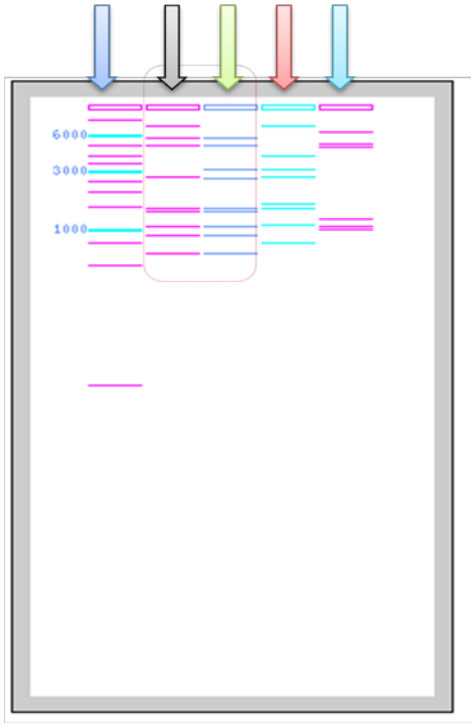


Figure 3.12: Gel image for *TMEM106B.* Light blue arrow indicates the marker lane while grey, light green, light reddish and cyan arrows represent the human, chimpanzee, macaque and mouse sequence bands respectively. A soft cornered rectangle enclosed the human and chimpanzee genes which showed symmetry in their band patterning.

3.4.4 SNP- map for *TMEM106A*

Only a single SNP was observed in macaque, while chimpanzee showed even worst and inflexible attitude against SNPs, as not a single SNP was found. In human *TMEM106A*, which contained almost equal gene length (7,995 nucleotides) with *TMEM106A* of chimpanzee was exposed more to SNPs, thereby forming clusters. These SNPs might be due to the insertion or deletion in human gene sequence, however equal gene length requires some attention. Mouse *TMEM106A* represented more SNPs prone at a region from 1bp to 1kb and from 82 to 95 kbp. Although mouse has greater sequence length than human, the SNPs observed in human were greater in numbers.

3.4.5 RFLP map for *TMEM106A*

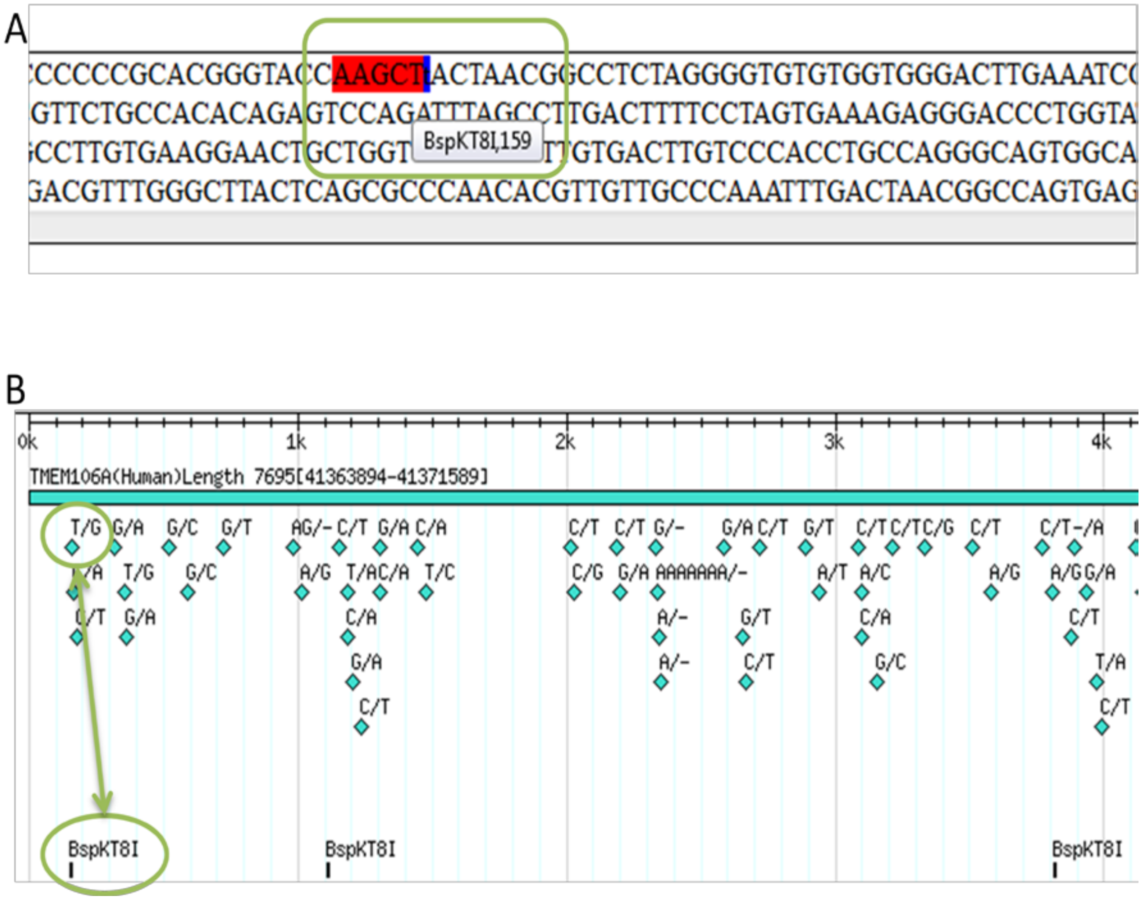
In contrast to RFLP map that was generated for *TMEM106B* orthologous, the *TMEM106A* RFLP did not follow the restriction sites conservation phenomenon.

In contrast to human and chimpanzee *TMEM106B* gene, there was not a uniform symmetry in restriction site patterns. Another contrary fact for *TMEM106A* was that it exhibited the restriction sites only for R2 types of restriction enzymes. Not a single restriction enzyme other than the R_2 (Type II restriction enzymes) type cut this gene, even for a single time. Gel image created on the basis of *TMEM106A* showed irregular pattern, when a comparative analysis effort was made.

3.4.6 SNP-RFLP map for *TMEM106A*

As discussed in RFLP map of *TMEM106A* that it did not follow any symmetry for restriction enzymes. Approximately similar behavior was analyzed in SNP-RFLP map. The restriction enzyme *M.Psil* did not cut *TMEM106A* even not for a single time, although this enzyme created new restriction sites in *TMEM106B* and *TMEM106C*. But in case of *TMEM106A* a new restriction site was created by replacing the last nucleotide of pattern, AAGCTT, a restriction pattern which is recognized by BspKT81 restriction enzyme. Nearby this restriction site, no other pattern of restriction site was found, even in orthologous. The string map and var image for *TMEM106A* is described in figure 3.13.

To trace the effect of T/G variation, a graphically rich gel image was provided (Figure 3.14). From the gel image, as it was already discussed that this SNP is not a restoring one due to which the pattern of gel bands was random.


A
B
0k
1k
2k
3k
4k
TMEM106A(Human)Length 7695[41363894-41371589]
T/G G/A G/C G/T AG/- C/T G/A C/A C/T C/T G/- G/A C/T G/T C/T C/T C/G C/T C/T -/A
A/A T/G G/C A/G T/A C/A T/C C/G G/A AAAAAAA/- A/T A/C A/G A/G G/A
C/T G/A C/A A/- G/T C/A C/T
G/A A/- C/T G/C T/A
C/T C/T
BspKT8I
BspKT8I
BspKT8I

```
CCCCCGCACGGGTACCAAGCTACTAACGGCCTCTAGGGGTGTGTGGTGGGACTTGAAATCC
GTTCTGCCACACAGAGTCCAGATTTAGCCTTGACTTTTCCTAGTGAAAGAGGGACCCTGGTA
CCTTGTGAAGGAACTGCTGGT BspKT8I,159 TTGTGACTTGTCCCACCTGCCAGGGCAGTGGCA
GACGTTTGGGCTTACTCAGCGCCCAACACGTTGTTGCCCAAATTTGACTAACGGCCAGTGAG
```

Figure 3.13: SNP-RFLP map for *TMEM106A*. (A) String map. In string map restriction pattern and the mouse over effect shows the enzyme and pattern occurrence position which is encapsulated by green soft rectangle. (B) Graphical image shows the graphical representation of string map. It encircles the SNP and restriction enzyme by green circles on var image.

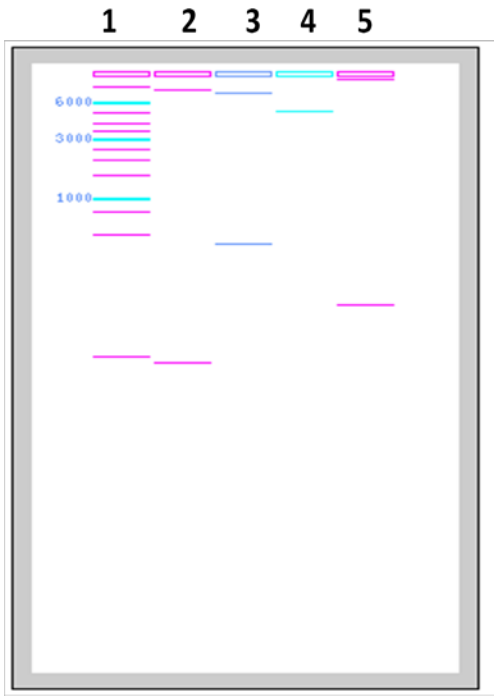


Figure 3.14: Gel image for *TMEM106A*. 1, 2, 3, 4, 5 lanes shows ladder, human, chimpanzee, macaque and mouse gene bands. From the image it is clear that the bands did not follow any symmetry.

3.4.7 SNP- map for *TMEM106C*

Likewise to other genes of this family, SNP map for human *TMEM106C* orthologous also showed that it contained more SNPs as compared to others. In macaque, only a single SNP with allele string T/A was observed; while in chimpanzee, two SNPs with allele string T/C and G/C were reported up to now. In contrast to these two primates, macaque and chimpanzee mouse exhibited more variations with respect to SNPs. High density of SNPs in 400 to 500 base pair region was reported in human and mouse *TMEM106C*. Mouse exhibited more SNPs in region 500-600 base pairs as compared to human.

3.4.8 RFLP map for *TMEM106C*

During the RFLP analysis of *TMEM106C*, as symmetry of restriction sites were observed in human *TMEM106C* and chimpanzee *TMEM106C*. This fact was also confirmed by the gel image. It was clear that human and chimpanzee *TMEM106C* genes demonstrated the similar band pattern. Similar to *TMEM106A*, all the restriction enzymes were of R_2 type.

3.4.9 SNP-RFLP map for *TMEM106C*

String map coupled with the SNP-RFLP map for *TMEM106C* was illustrated in figure 3.15. For *TMEM106C,* a new restriction site was created by a SNP A/G (similar SNP that also contributed in creating new restriction site in case of human *TMEM106B*). As far as the restriction pattern of *TMEM106C* is concerned, it was created by replacing the third nucleotide from the 5-prime end where as in *TMEM106B*; it was at the last position. This restriction activity is shown by blue circles in figure 3.15B. Output for *TMEM106C* in the form of gel is shown in figure 3.16, where a green dotted rectangle indicated the effect of this pattern restoration. The graphical representation of var image is shown in figure 3.16. Similar banding behavior was observed for human and chimpanzee in gel view (Figure 3.17). An additional and novel feature that was observed in the string map was that the distance between the two consecutive restriction patterns in human was similar to the distance between the same restriction patterns in chimpanzee (thirteen nucleotides).

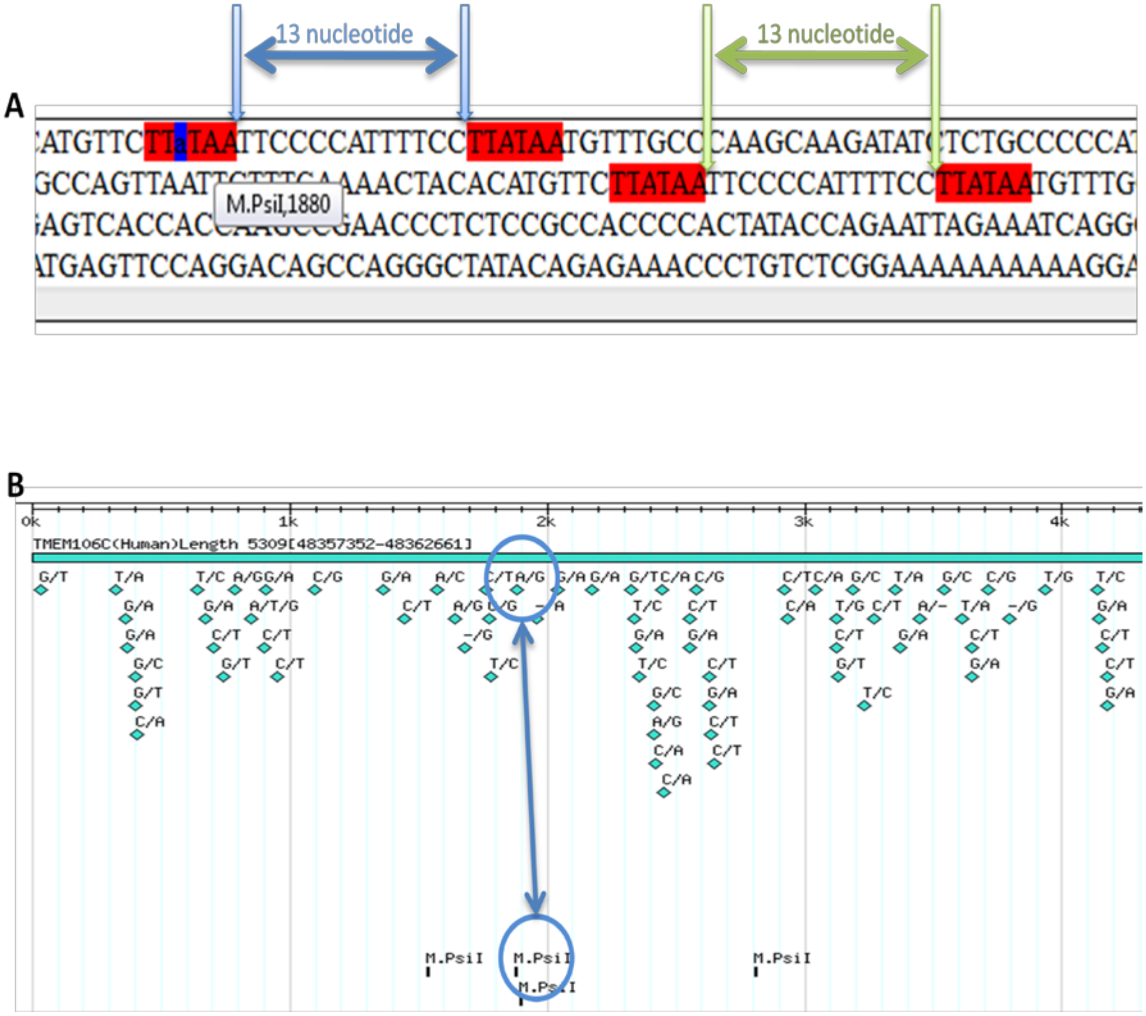

A
13 nucleotide
13 nucleotide
ATGTTCTTaTAATTCCCCATTTTCCTTATAATGTTTGCCCAAGCAAGATATCTCTGCCCCCA
GCCAGTTAATTCTTTCAAAACTACACATGTTCTTATAATTCCCCATTTTCCTTATAATGTTTG
AGTCACCACCAAGCCGAACCCTCTCCGCCACCCCACTATACCAGAATTAGAAATCAGG
TGAGTTCCAGGACAGCCAGGGCTATACAGAGAAACCCTGTCTCGGAAAAAAAAAAGGA
M.PsiI,1880
B
0k
1k
2k
3k
4k
TMEM106C(Human)Length 5309[48357352-48362661]
M.PsiI
M.PsiI
M.PsiI
M.PsiI

Figure 3.15: SNP-RFLP map for *TMEM106C.* A. shows the string map for *TMEM106C.* In human gene sequence the blue colored arrows shows the distance between the human restriction sites. Green colored arrows shoes the distance between the chimpanzee restriction sites. B. depicts the graphical representation of this newly created restriction and the SNP, shown by blue circles.

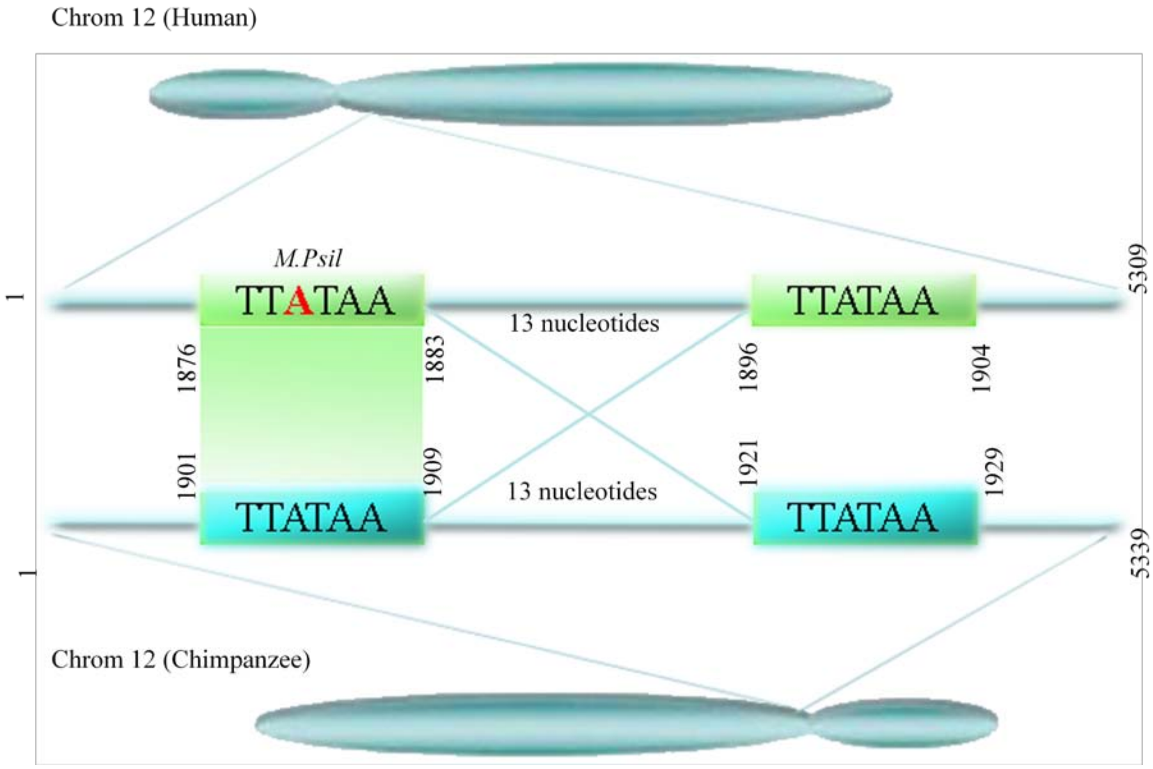


Figure 3.16: Graphical representation of var image for TMEM106C. The restriction pattern for restriction enzyme *M.PsiI* occurs at 1877-1882 in human which is also found at 1902-1909 in chimpanzee. Both these genes are located at chromosome 12 in their respective species. In figure a segment of human TMEM106C (1-5309) and chimpanzee TMEM106C (1-5339) are shown. The distance between the restriction pattern in human and in chimpanzee is shown by the cross. This distance is equal to 13 nucleotides, common in both. Red colored A (SNP) in pattern TTATAA (1877-1882) depicts that this site is newly created. Green colored background shows the relationship in between the newly created restriction sites of human and chimpanzee (old one).

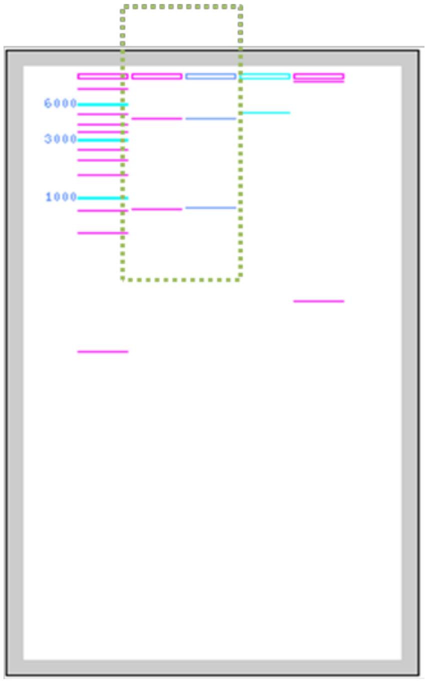


Figure 3.17: Gel image *TMEM106C*. The green rectangle enclosed the human and chimpanzee gene sequence bands. Both human and chimpanzee shows similar pattern of bands. Next to chimpanzee lane is the lane of macaque. Last lane represents the mouse.

3.5 Conclusion of analysis of transmembrane protein 106B family

We used transmembrane protein 106B family for investigation and validation of our web server. As mentioned earlier that this family comprises of three members, TMEM106A, TMEM106B and TMEM106C in human. From our analysis it is obvious that TMEM106B contained highest number of SNPs in human, TMEM106C is second followed by the TMEM106A (Table 3.1). Therefore it is stated that TMEM106B have high mutational rate however for this statement, additional study and factors are required. By looking at the table 3.1, one can easily observe that evolution of different gene family's members and ultimately the evolution of gene family can be understood. It can also be used to discover the relationship among paralogs based on variation data. As TMEM106B and TMEM106C are closer, they almost have same RFLP and SNP-RFLP map as compared to TMEM106A. Although this data set is small, yet it undoubtedly and evidently exposed many novel happenings that were not reported to date. By analyzing more data, VRSAP can provide a benchmark, where the terms like “mutational rate” and “random mutation” can be challenged.

4. Discussion

Implementation of high throughput approaches to the completion of various genome projects generated a new enthusiasm and favorable circumstances for the researchers. In this very potential study, we highlighted and expanded our understanding towards genetic studies like gene evolution, gene family evolution, genetic markers, investigating diseases caused by the genetic factors and identification of the true paralogs and orthologs by making decisions based on the common characteristics among the individuals or species. Our main motivation was to design a server/tool to assess the already existing larger projects like UCSE, Ensembl and Hapmap and to analyze the assessed information to generate the convincing results in the form of SNPs, RFLP and SNP-RFLP maps. In addition to these maps, VRSAP results also included a hypothetical gel image, which represented gene sequence bands. In comparison to other data servers for SNPs (Hirakawa *et al.*, 2002; Clark *et al.*, 2005; Forbes *et al.*, 2008; Sherry *et al.*, 2011), VRSAP is the only server which provides the SNP map along-with the RFLP. VRSAP uses Ensembl as a source of data, which gives it an edge over the other servers as there is no need to update its dataset. Ensembl contains gene reference sequences and the variation data. The variation data in Ensembl was retrieved from HapMap via dbSNP. VRSAP also takes the advantage of this pipeline as it indirectly connects to three servers during its processing.

For SNPs, SNP-RFLP and RFLP analyses, VRSAP inputs include the gene name, species name and restriction enzymes after that rest of work is done by VRSAP itself thus, facilitating the users when comparing with other computational tools (Bikandi *et al.*, 2004; Zhang *et al.*, 2005; Wegrzyn *et al.*, 2009; Chang *et al.*, 2010). Another distinctive characteristic of VRSAP is that it provides the pattern of restriction sites among different species to point out common restriction sites. Currently, for tracing the restriction sites in gene sequence, a number of tools are available e.g. NEBcutter (Vincze *et al.*, 2003). VRSAP provides the RFLP analysis for multiple sequences which is the first advantage over NEBcutter. Secondly it uses KMP (Knuth *et al.*, 1977) algorithm which is faster than the Brute Force algorithm, which is implemented by NEBcutter. In case of gel image, VRSAP makes it possible to compare gel bands of multiple sequences on the same gel in contrast to NEBcutter which provides gel image for only one sequence.

Another striking feature of VRSAP, which makes it suitable for RFLP analysis, is its recognition of additional restriction sites created by SNPs, which can be considered as unique genetic markers. By making comparative analysis of RFLP, the restriction sites which are under strong natural selections can be observed. VRSAP also enlists the restriction sites which are restored by SNPs. This raises many questions by exploring the nature at molecular level such that why specific restriction sites are restored? What are the effects of SNPs at restriction sites? Why some restriction sites are still conserved in human and other primates?

Flanking sequences of SNPs can be viewed in the string map. As the string map shows whole gene sequence up to the requested length, the flanking sequences of the SNP can be traced, a facility provided by VRSAP when compared with SNP-Flankplus (Yang *et al.*, 2008). Moreover, this information can be used for primer designing.

Genetic diversity, novelty and human diseases are the primary products of mutations that occurred in many folds of genes. Therefore, knowledge of mutational rate can be a convincing step for understanding the above mentioned phenomenon. One very basic application of mutational rates that can be observed among genes and lineages is to estimate the divergence time among closely related species, which can be utilized for selection by comparative sequence analysis, testing coalescent time and finally the mutational processes, which are key for genome evolution (Kumar and Subramanian, 2002; Kehrer and Cooper, 2007; Marques *et al.*, 2009). The complete knowledge of genes synonymous and non-synonymous mutations in the coding regions of different sequences can be further utilized for the rate analysis calculation, a very important factor to analyze the positive and negative natural selection on that specific dataset. Gene’s synonymous and non-synonymous mutations can easily be visualized from the table, created in SNPs analysis module of VRSAP, which enlists all the SNPs occurring in a gene. With the information of coding region variations obtained from the VRSAP, user can extend its analysis for natural selection detection to make a population genetics study.

In short, VRSAP results can be exercised to outline the mutational rate of certain data sets.

Mostly, the genes or gene family based comparative genomic studies rely on the coding regions, ignoring the expense of examining the regulator of sequences (Carroll, 2003). Functional non coding elements such as promoters, enhancers, and flanking sequences and most importantly the introns can have a major role in the regulation of gene expression (Wray *et al*., 2003). These functional, some time evolutionary conserved regions play a crucial role in human and other genome evolution. In contrast to other approaches (Gardner *et al.*, 2005; Xu *et al.*, 2005; Pico *et al.*, 2009; Sicotte *et al.*, 2011), VRSAP explores both coding and non-coding regions to highlight the SNPs, RFLP and SNP-RFLP; thereby covering the functional and the regulators at the same time. These variation differences or the similarities become a vital source in investigating a disease in human when similar genes are detected in other species. For example, rheumatoid arthritis and multiple types of cancers like breast and colon cancers have been rarely reported in chimpanzee and more commonly reported in humans (Shastry, 2002). Another example is heart disease, common in human and its closest relative chimpanzee however, it follows different pathogenically processes in chimpanzee (Varki *et al.*, 2009). The answers for these questions will clearly be a valuable tool and have an impact on medical care of human. Additionally, we can also judge human specific unique and key features which play a fundamental role in human evolution. Thus, VRSAP provides a comparative view of genes along with SNPs and exploring the mechanism of variations among genes by showing the string map and graphical display, thereby facilitating the curious researches by providing an opportunity where human or other specie specific features (common in all primates or in mammals) can be judged. Moreover, VRSAP can play an essential role in marker designing along with SNPs and RFLP which can be a conclusive evidence for disease studies (Roses *et al.*, 2010), revealing origins (Huang *et al.*, 2012) of a species and for the identification and mapping of conserved orthologs (Kuhn *et al.*, 2012). For example in cacao genetic mapping studies, SNP markers already successful in west Africa were largely replaced RAPDs, AFLPs, microsatellites or candidate genes (Lanaud *et al.*, 1999; Kuhn *et al*., 2005; Brown e*t al*., 2008). Another example of utilization of SNP map as marker is QTL mapping (Dirlewanger *et al.*, 2004; Minvielle *et al.*, 2006; Beraldi *et al.*,2007; Dawson *et al.*,2007*;* Cabrera *et al.*, 2012). Although genetic approaches successfully dealt with many diseases, still several diseases associated with heritability require detailed identification. Sequencing of multiple genomes such as chimpanzee, mouse and their comparison with human to reveal the genetic basis of diseases and evolution coupled with self human population's comparison provided an opportunity to answer the genetically linked diseases and syndrome.

Taken together, VRSAP as a web based application is more informative with easy to use and without involving installation steps, which made the SNPs, RFLP and SNPs-RFLP analysis more comprehensive and reasonable in multiple species. Moreover by providing graphical display VRSAP can serve as a valuable tool in Pharmcogenomics, evolution, population genetics and molecular studies.

5. References

1. Ahmadian A, Ehn M and Hober S: **Pyrosequencing: history, biochemistry and future.** *Clin Chim Acta* 2006, 363(1-2):83-94.
2. Bamshad M, Wooding S, Salisbury BA and Stephens JC: **Deconstructing the relationship between genetics and race.** *Nature reviews* 2004, 5(8):598-609.
3. Beraldi D, Mcrae AF, Gratten J, Slate J, Visscher PM and Pemberton JM: **Mapping quantitative trait loci underlying fitness related traits in a free-living sheep population.** *Evolution* 2007, 61(6):1403-1416.
4. Bikandi J, San MR, Rementeria A and Garaizar J: **In silico analysis of complete bacterial genomes: PCR, AFLP-PCR and endonuclease restrictionn.** *Bioinformatics* 2004, 20(5):798-799.
5. Brookes AJ: **The essence of SNPs.** *Gene* 1999, 234:177–186.
6. Brown J, Sautter R, Olano C, Borrone J, Kuhn D, Motamayor J and Schnell R: **A composite linkage map from three crosses between commercial clones of cacao, Theobroma cacao L.** *Trop Plant Biol* 2008, 1(2):120–130.
7. Cabrera A, Rosyara UR, Franceschi PD, Sebolt A, Sooriyapathirana SS, Dirlewanger E, Garcia JQ, Schuster M, Iezzoni AF and Knaap EVD: **Rosaceae conserved orthologoues sequences marker polymorphism in sweet cherry germplasm and construction of a SNP-based map.** *Tree Genetics & Genomes* 2012, 8:237–247.
8. Carroll SB: **Genetics and the making of Homo sapiens.** *Nature* 2003, 422:849-857.
9. Chang HW, Cheng YHU, Chuang LY and Yang CH: **SNP-RFLPing 2: an updated and integrated PCR-RFLP tool for SNP genotyping.** *BMC Bioinformatics* 2010, 11:173.
10. Cheung KH, Osier MV, Kidd JR, Pakstis AJ, Miller PL, and Kidd K: **ALFRED: An allele frequency database for diverse populations and DNA polymorphisms.** *Nucleic Acids Res* 2000, 28:361–363.
11. Chunming D: **Other applications of single nucleotide polymorphisms.** *Trends in Biotechnology* 2007, 25:7.

12. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH and Nielsen R: **Ascertainment bias in studies of human genome wide polymorphism.** *Genome Research* 2005, 15:1496-1502.
13. Dawson DA, Akesson M, Burke T, Pemberton JM, Slate J and Hansson B: **Gene order and recombination rate in homologous chromosome regions of the chicken and a passerine bird.** *Mol Bio Evo.* 2007, 24(7):1537-1552.
14. Dirlewanger E, Graziano E, Joobeur T, Garriga CF, Cosson P, Howad W, and Arus P: **Comparative mapping and marker assisted selection in rosaceae fruit crops.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, 101(26):9891-9896.
15. Dong SL, Wang E, Hsie L, Cao YX, Chen XG and Gingeras TR: **Flexible use of high density oligonucleotide arrays for single nucleotide polymorphism discovery and validation.** *Genome Res* 2001, 11:1418–1424.
16. Ewing B and Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, 8:186-194.
17. Fan JB, Chee MS and Gunderson KL: **Highly parallel genomic assays.** *Nat Rev Genet* 2006, 7(8):632-644.
18. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Silva DC, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GRS, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Suarez XMF, Harrow J, Herrero J, Hubbard TJP, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A and Searle SMJ: **Ensembl.** *Nucleic Acids Res* 2012**,** 40:D84-D90.
19. Forbes SA , Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA and Stratton MR: **The Catalogue of Somatic Mutations in Cancer (COSMIC).** *Curr Protoc Hum Genet* 2008, 10:11.

20. Forche A, Magee PT, Selmecki A, Berman J and May G: **Evolution in *Candida albicans* populations during single passage through a mouse host.** *Genetics* 2009, 182(3):799–811.
21. Fradin C, Kretschmar M, Nichterlein T, Gaillardin C, Enfert CD and Hube B: **Stage specific gene expression of *Candida albicans* in human blood.** *Molecular Microbiology* 2003, 47:1523–1543.
22. Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H and Brookes AJ: **HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources.** *Nucleic Acids Res* 2002, 30:387–391.
23. Gardner SN and Wagner MC: **Software for optimization of SNP and PCR-RFLP genotyping to discriminate many genomes with the fewest assays.** *BMC Genomics* 2005, 6(1):1-73.
24. Goldstein DB and Cavalleri GL: **Genomics: Understanding human diversity.** *Nature* 2005, 437:1241-1242.
25. Griffin TJ and Smith LM: **Single nucleotide polymorphism analysis by MALDI-TOF mass spectrometry.** *Trends Biotechnolgy* 2000, 18(2):77- 84.
26. Hacia JG and Collins FS: **Mutational analysis using oligonucleotide microarrays.** *J Med Genet* 1999, 36:730–736.
27. Hacia JG, Fan JB, Ryder O, Jin L, Edgemon K, Ghandour G, Mayer RA, Sun B, Hsie L, Robbins CM, Brody LC, Wang D, Lander ES, Lipshutz R, Fodor SPA and Collins FS: **Determination of ancestral alleles for human single nucleotide polymorphisms using high density oligonucleotide arrays.** *Nat Genet* 1999, 22:164–167.
28. Hamosh A, Scott AF, Amberger JS, Bocchini CA and McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2005, 33:D514–D517.
29. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA and Cox DR: **Whole genome patterns of common DNA variation in three human populations.** *Science* 2005, 307:1072-1079.

30. Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T and Nakamura Y: **JSNP, a database of common gene variations in the Japanese population.** *Nucleic Acids Res* 2002, 30:1.
31. Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, Guo Y, Lu Y, Zhou C, Fan D, Weng Q, Zhu C, Huang T, Zhang L, Wang Y, Feng L, Furuumi H, Kubo T, Miyabayashi T, Yuan X, Xu Q, Dong G, Zhan Q, Li C, Fujiyama A, Toyoda A, Lu T, Feng Q, Qian Q, Li J and Han H: **A map of rice genome variation reveals the origin of cultivated rice.** *Nature* 2012, 1-5.
32. Hube B: **From commensal to pathogen: stage and tissue specific gene expression of *Candida albicans*.** *Curr Opin Microbiol* 2004, 7:336–341.
33. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, 409:860–921.
34. Kehrer SH and Cooper DN: **Understanding the recent evolution of the human genome: insights from human chimpanzee genome comparisons.** *Hum Mutat* 2007, 28:99–130.
35. Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, King AJ, Staines D, Derwent P, Kerhornou A, Kersey P and Flicek P. **Ensembl BioMarts: a hub for data retrieval across taxonomic space.** *Database (Oxford)* 2011.
36. Knuth DE, Morris JH and Pratt VR: **Fast pattern matching in strings.** *Siam J. Comput* 1977, 6:323- 349.
37. Kuhn DN, Borrone J, Meerow AW, Motamayor JC, Brown JS and Schnell RJ: **Single-strand conformation polymorphism analysis of candidate genes for reliable identification of alleles by capillary array electrophoresis.** *Electrophoresis* 2005, 26(1):112–125.
38. Kuhn DN, Livings DT, Main D, Zheng P, Saski C, Feltus FA, Mockaitis K, Farmer AD, May GD, Schnell RJ and Motamayor JC: **Identification and mapping of conserved ortholog set (COS) II sequences of cacao and their conversion to SNP markers for marker assisted selection in Theobroma cacao and comparative genomics studies.** *Tree Genetics & Genomes* 2012, 8:97 –111.

39. Kumar S and Subramanian S: **Mutation rates in mammalian genomes.** *PNAS* 2002, 99(2):803– 808.
40. Kwok PY: **Methods for genotyping single nucleotide polymorphisms.** *Annu Rev Genomics Hum Genet* 2001, 2:235-258.
41. Lanaud C, Risterucci AM, Pieretti I, Falque M, Bouet A and Lagoda PJL: **Isolation and characterization of microsatellites in Theobroma cacao L.** *Mol Ecol* 1999, 8(12):2141–2143.
42. Lichten MJ and Fox MS: **Detection of non-homology containing heteroduplex molecules.** *Nucleic Acids Res* 1983, 11:3959–3971.
43. Machado M, Magalhaes WCS, Sene A, Araujo B, Campos ACF, Chanock SJ, Scott L, Oliveira G, Santos ET and Rodrigues EMR: **Phred-Phrap package to analyses tools: a pipeline to facilitate population genetics re-sequencing studies.** *Investigative Genetics* 2011, 2:3.
44. Marques BT, Ryder OA, and Eichler EE: **Sequencing primate genomes: what have we learned?** *Annu Rev Genomics Hum Genet* 2009, 10:355–386.
45. Marshall A and Hodgson J: **DNA chips: an array of possibilities.** *Nat Biotechnol* 1998, 16:27–31.
46. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY and Gish WR: **A general approach to single nucleotide polymorphism discovery.** *Nat Genet* 1999, 23:452-456.
47. Matukumalli LK, Grefenstette JJ, Hyten DL, Choi IkY, Cregan PB and Tassell CPV: **SNP-PHAGE: High throughput SNP discovery pipeline.** *BMC Bioinformatics* 2006, 7:468.
48. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP and Hirschhorn JN: **Genome-wide association studies for complex traits: consensus, uncertainty and challenges.** *Nature reviews* 2008, 9(5):356-369.
49. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P and Cunningham F: **Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.** *BMC Bioinformatics* 2010, 26(16):2069-70.

50. Minvielle F, Kayang BB, Inoue-Murayama M, Miwa M, Vignal A, Gourichon D, Neau A, Monvoisin JL and Ito S: **Search for QTL affecting the shape of the egg laying curve of the Japanese quail.** *BMC Genetics* 2006, 7:26.
51. Nachman MW and Crowell SL: **Estimate of the mutation rate per nucleotide in humans.** *Genetics* 2000, 156:297–304.
52. Nickerson DA, Tobe VO and Taylor SL: **PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence based resequencing.** *Nucleic Acids Res* 1997, 25:2745-2751.
53. Olivier M: **The Invader assay for SNP genotyping.** *Mutat Res* 2005, 573 (1-2):103-110.
54. Orita M, Iwahana H, Hayashi K and Sekiya T: **Detection of polymorphism of human DNA by gel electrophoresis as single strand conformation polymorphisms.** *Proc Natl Acad Sci USA* 1989, 86:2766–2770.
55. Paigen K and Petkov P: **Mammalian recombination hot spots: properties, control and evolution.** *Nat Rev Genet* 2010, 11:221–233.
56. Parsons BL and Heflich RH: **Genotypic selection methods for the direct analysis of point mutations.** *Mutat Res* 1997, 387:97-121.
57. Parvanov ED, Petkov PM and Paigen K: **Prdm9 Controls Activation of Mammalian Recombination Hotspots.** *Science* 2009, 327:835.
58. Pico AR, Smirnov IV, Chang JS, Yeh RF, Wiemels JL, Wiencke JK, Tihan T, Conklin BR and Wrensch M: **SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system.** *Nucleic Acids Res* 2009, 37:803–D809.
59. Qi XQ, Bakht S, Devos KM, Gale MD and Osbourn A: **L-RCA (Ligation rolling circle amplification): a general method for genotyping of single nucleotide polymorphism (SNPs).** *Nucleic Acids Res* 2001, 29:U68–U74.
60. Quinlan RJ: **C4.5: Programs for Machine Learning.** *Morgan Kaufmann* 1993.
61. Ramsay G: **DNA chips: state of the art.** *Nat Biotechnol* 1998, 16:40–44.
62. Roberts, Richard J, Tamas V, Janos P and Dana M**: REBASE-a database for DNA restriction and modification: enzymes, genes and genomes.** *Nucleic Acids Re* 2010, 38:D234-6.

63. Roses AD, Lutz MW, Madsen HA, Saunders AM, Huentelman MJ, Bohmer KAW and Reiman EM: **A TOMM40 variable length polymorphism predicts the age of late-onset Alzheimer's disease.** *The Pharmacogenomics Journal* 2010, 10:375-384.
64. Sachidanandam R: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.** *Nature* 2001, 409:928–933.
65. Shastry BS: **SNP alleles in human disease and evolution.** *J Hum Genet* 2002, 47:561-566.
66. Sherry ST, Ward MH and Sirotkin K: **dbSNP Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation.** *Genome Res* 1999, 9:677-679.
67. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM and Sirotkin K: **dbSNP the NCBI database of genetic variation.** *Nucleic Acids Res* 2011, 29:308–311.
68. Sicotte H, Rider DN, Poland GA, Dhiman N and Kocher JPA: **SNPPicker: High quality tag SNP selection across multiple populations.** *BMC Bioinformatics* 2011, 12:129.
69. Stoneking M: **Single nucleotide polymorphisms: from the evolutionary past.** *Nature* 2001, 409:821–822.
70. Thorisson GA and Stein LD: **The SNP Consortium website: past, present and future.** *Nucleic Acids Res* 2003, 31:124–127.
71. Varki N, Anderson D, Herndon JG, Pham T, Gregg CJ, Cheriyan M, Murphy J, Strobert E, Fritz J, Else JG and Varki A: **Heart disease is common in humans and chimpanzees, but is caused by different pathological processes.** *Evolutionary Applications* 2009, 2:101–112.
72. Vincze T, Posfai J and Roberts RJ: **NEBcutter: a program to cleave DNA with restriction enzymes.** *Nucleic Acids Research* 2003, 31(13):3688–3691.
73. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen NP, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M and Lander ES: **Large scale**

identification, mapping, and genotyping of single nucleotide polymorphisms in the human genome. *Science* 1998, 280:1077–1082.

74. Wegrzyn JL, Lee JM and Neale JLDB: **PineSAP, sequence alignment and SNP identification pipeline.** *Bioinformatics Applications Note* 2009, 25: 2609–2610.
75. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV and Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol. Biol. Evol*. 2003, 20:1377–1419.
76. Xu H, Gregory SG, Hauser ER, Stenger JE, Vance MAP, Vance JM, Zuchner S and Hauser MA: **SNPselector: a web tool for selecting SNPs for genetic association studies.** *Bioinformatics* 2005, 21(22):4181–4186.
77. Yang CH, Cheng YH, Chuang LY and Chang HW: **SNP-Flankplus: SNP ID centric retrieval for SNP flanking sequences.** *Bioinformation* 2008, 3(4):147-149.
78. Yoshino T, Takeyama H and Matsunaga T: **Single nucleotide polymorphism analysis using a bacterial magnetic particle microarray.** *Electrochemistry* 2001, 69:1008–1012.
79. Youil R, Kemper BW and Cotton RGH: **Screening for mutation by enzyme mismatch cleavage with T4 endonuclease VII.** *Proc Natl Acad Sci USA* 1995, 92:87–91.
80. Zhang R, Zhu Z, Zhu H, Nguyen T, Yao F, Xia K, Liang D and Liu C: **SNP Cutter: A comprehensive tool for SNP PCR-RFLP assay design.** *Nucleic Acids Res* 2005, 33:W489-492.