Review on Single Nucleotide Polymorphism Analysis Methods

Nusrath A
Asst. Professor,
Dept. of Computer Science,
Farook College, Calicut, Kerala.

Beema Raiza P T
PG Student,
Department of Computer Science,
Farook College, Calicut, Kerala.

Abstract- **Single Nucleotide Polymorphism (SNP) variants represent a prevalent form of genetic variation. Mutation in the coding regions like SNPs are frequently associated with the development of various genetic diseases. Identifying Single Nucleotide Polymorphisms (SNPs) that underlie complex disease is expected to enable early diagnosis, effective treatment and ultimately prevention of target disease. Also there are a tremendous number of SNPs on the human genome, estimated at over 10 million average. So computational tools are required for prioritizing SNPs according to their potentially deleterious effects to human health. This paper is a review of various methods and tools available for identification of deleterious SNPs.**

Keywords — Single Nucleotide Polymorphism (SNP), Human Genome.

I. INTRODUCTION

The advancement of sequencing technology provided the whole genome information of many organisms. This made the greatest challenges of storing these huge data and analysing and interpreting these data to get practical solutions for many problems like diagnosis of inherited diseases. Bioinformatics databases provide facilities for storing different types of biological data and bioinformatics tools are very popular for analysing and interpreting these data.

The identification of genetic risk factors underlying human inherited diseases has long been a goal in human and medical genetics. Since genetic variation is believed to be the major factor that stimulates the diversity between individuals, considerable efforts have been taken to understand associations between human genetic variants and their phenotypic effects . A number of successful stories have shown that such efforts are helpful in capturing the causative variants which affect human inherited diseases, providing important information for grasping genetic bases of complex diseases, and further promoting the prevention diagnosis, and treatment of these diseases [20].

Genetic variants can typically be classified into several categories, including single-nucleotide polymorphisms (SNPs), small insertions and deletions, and structural variants [20]. A single-nucleotide polymorphism (SNPs) occur normally throughout a humans DNA sequence. They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNPs in the human genome. Most commonly, these variations are found in the DNA between genes. They can act as biological markers, helping scientists locate genes that are associated with disease. When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting the gene's function.

Most SNPs have no effect on health or development. Some of these genetic differences, however, have proven to be very important in the study of human health. Researchers have found SNPs that may help predict an individual's response to certain drugs, susceptibility to environmental factors such as toxins, and risk of developing particular diseases. SNPs can also be used to track the inheritance of disease genes within families. Future studies will work to identify SNPs associated with complex diseases such as heart disease, diabetes, and cancer.

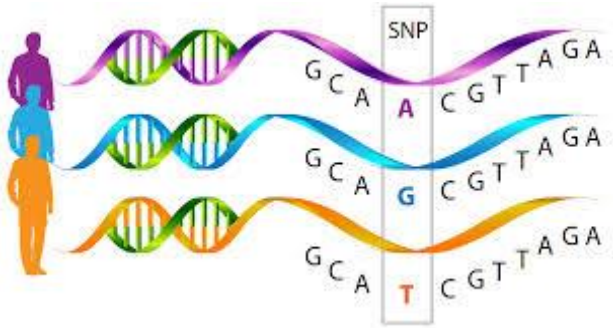

Fig 1: Example for SNP

According to the genomic region each SNP is examined for its possible deleterious effects with respect to the corresponding bio-molecular functional category as follows:

1) Protein Coding: SNPs in exonic regions may alter protein structure and function by creating a new start or stop codon (i.e. nonsense SNPs) or a deleterious amino acid substitution (i.e. missense SNPs).

2) Splicing Regulation: SNPs in (canonical) splice sites may disrupt splicing regulation, resulting in exon skipping or intron retention. SNPs in exonic splice sites may interfere with alternative splicing regulation by changing exonic splicing enhancers or silencers.

3) Transcriptional Regulation: SNPs in transcription regulatory regions (e.g. transcription factor binding sites, CpG islands, microRNAs, etc.) can alter binding sites, and thus disrupt proper gene regulation.

4) Post-Translational Modification: SNPs in protein-coding regions may alter post-translational modification sites, interfering with proper posttranslational modification[21].

This paper is divided into 4 sections. Section 2 describes some important databases for retrieving SNPs. Section 3 discusses various SNP tools and Section 4 describes the SNPs related to APOE gene of Alzheimer's disease and lastly section 5 concludes this paper.

II. SNP DATABASES

There are number of databases that gives information about SNPs and their characteristics. It includes single nucleotide polymorphism database (dbSNP)[2], Online Mendelian Inheritance in Man (OMIM) database[3], the Human Gene Mutation Database (HGMD)[4], the UniProt/Swiss-Prot database[5], Human Genome Variation database (HGVbase)[6], the Protein Mutant Database (PMD)[7] , and the database for nonsynonymous SNP's function prediction (dbNSFP)[8].

The dbSNP (Single Nucleotide Polymorphism database) is a public-domain archive for a broad collection of simple genetic polymorphisms. This collection of polymorphisms includes single-base nucleotide substitutions (also known as single nucleotide polymorphisms or SNPs), deletion insertion polymorphisms(DIPs) and short tandem repeats(STRs). The dbSNP has been designed to support submissions and research into a broad range of biological problems like physical mapping, functional analysis, pharmacogenomics, association studies, and evolutionary studies. Because dbSNP was developed to complement GenBank, it may contain nucleotide sequences from any organism[2]..

OMIM (Online Mendelian Inheritance in Man) is a comprehensive and freely available database. In 1995, OMIM was developed for the World Wide Web by NCBI,(National Center for Biotechnology Information).It is mainly focuses on the relationship between phenotype and genotype and also collecting molecular relations between genetic variations. OMIM contain information about all known Mendelian disorders and their relative genes[3].

HGMD (Human Gene Mutation Database) is mainly used for genetics and genomic research and analysing the genomes of organisms. It includes germ-line disease-causing mutations and deleterious polymorphisms. HGMD provides two versions of databases, one is for academic (non-profit users) and the other is for professional usage[4].

.The UniProt dataset is a comprehensive, high quality and computational analysis for evidenced-based associations between terms from the Gene Ontology resource and UniProtKB proteins.it provides protein function descriptions and domain structures[5].

HGVbase(Human Genome Variation database) is an non redundant database for comprehensive catalog of normal human gene and genome variation,especially SNP variation. it is more accurate and high-quality HGVbase provides both neutral polymorphisms and disease-related mutations[6].

Protein Mutant Database is mainly depended on literature not on protein based. It includes protein mutant data, providing information on functional and structural influences for amino acid mutations at specific region of a protein. Each entry in the database corresponds to one article, which may contain several or a number of protein mutants[7].

dbNSFP is a database for all potential reference sequence in the human genome and their functional predictions. Among the analysis steps, functional prediction (being deleterious) plays an important role in filtering or prioritizing nonsynonymous SNP (NS). the information about nsSNPs and prediction scores from four popular tools (SIFT , PolyPhen-2 ,LRT , and Mutation Taster, along with a conservation score we can use PhyloP[8].

The following table summarizes various database from which we can take information about SNPs.

Table 1: Databases for SNP

Database and website	Features
dbSNP http://www.ncbi.nlm.nih.gov/snp	comprehensive repository for single-nucleotide substitutions, short deletion, and insertion polymorphisms
OMIM http:/www.omim.org/	Powerful, comprehensive and widely used database. collecting molecular relations between genetic variations and phenotypes.
HGMD http://www.hgmd.cf.ac.uk/ac/index.php	provides academic or nonprofit users. record all germ-line disease-causing mutations and deleterious polymorphisms.
UniPROT Database http://www.uniport.org/	high quality,manually curated. provides convincing protein sequences and annotations
HGVbase http://hgvbase.cgb.ki.se	accurate, high-quality, and nonredundant database. provides both neutral polymorphisms and disease-related mutations.
PMD http://pmd.ddbj.nig.ac.jp	for protein mutants provides information about amino acid mutations at specific positions of proteins and the structural alterations.
dbNSFP http://sites.google.com/site/jpopgen/dbNSFP	provides information about nsSNPs and prediction scores from four popular tools(SIFT,PolyPhen-2,LRT, and MutationTaster)

III. SNP TOOLS

There are some important computational tools for the identification of deleterious or neutral snp based on different parameters according to their different features.. Snp detection tools like SIFT[9][10], PolyPhen2[11], SNAP[12], MSRV[13], LTR[14], Mutation Taster[15], KGGseq[16][19], SInBaD[17] and GERP[18] are valuable for the analysis of SNP and their prioritization for characterizing deleterious ones. The input data for many tools includes protein sequence or protein ID, the amino acid substitution, position of the substitution, chromosome sequence or alignment sequence, then the tools can run automatically with their corresponding predictive scores.

SIFT takes a query sequence and uses multiple alignment information to predict tolerated and deleterious substitutions for every position of the query sequence. SIFT

is a multistep procedure that (1) searches for similar sequences, (2) chooses closely related sequences that may share similar function to the query sequence , (3) obtains the alignment of these chosen sequences, and (4) calculates normalized probabilities for all possible substitutions from the alignment. Positions with normalized probabilities less than 0.05 are predicted to be deleterious, those greater than or equal to 0.05 are predicted to be tolerated.

PolyPhen-2 is an automatic tool for prediction of possible impact of an amino acid substitution on the structure and function of a human protein. This prediction is based on a number of features comprising the sequence, phylogenetic and structural information characterizing the substitution. For a given amino acid substitution in a protein, PolyPhen-2 extracts various sequence and structure-based features of the substitution site and feeds them to a probabilistic classifier.

Sequence Information Based Decision model(SInBaD) provides a quantitative measure for every single nucleotide mutation within a human gene evaluating whether it is likely to be functional or not. SInBaD allows the selection of candidate nucleotide variations or genes in individual genomes. With the advent of new generation sequencing technologies in conjunction with human genome re-sequencing projects on their way, SInBaD can be used to study human dataset.

KGGSeq is comprehensive and efficient framework used to filter and prioritize genetic variants from whole exome sequencing data. KGGSeq is a software platform in Bioinformatics and statistical genetics functions making use of valuable biologic resources. And knowledge for sequencing-based genetic mapping of variants/genes responsible for human diseases/traits.

The following table summerise different tools for SNP analysis.

Table 2: SNP Analysis Tools

Tool and website	Features
SIFT http://blocks.fhcrc.org/sift/SIFT.html.	For analysing sequence
Polyphen-2 http://genetics.bwh.harvard.edu/pph2/	For analysing sequence and structure
SNAP http://www.broadinstitute.org/mpg/snap/	For analysing sequence and annotation based
MSRV http://bioinfo.au.tsinghua.edu.cn/member/ruijiang/english/software.html	For analysing sequence
LTR http://www.genetics.wustl.edu/jflab/lrtquery.html	For analysing sequence
Mutation Taster http://www.mutationtaster.org/	For analysing sequence and annotation based
KGG seq http://statgenpro.psychiatry.hku.hk/limx/kggseq/	For analysing sequence and annotation based
SInBaD http://tingchenlab.cmb.usc.edu/sinbad/	For analysing sequence
GERP http://mendel.stanford.edu/sidowlab/downloads/gerp/index.html	For analysing sequence

IV. AN EXAMPLE: SNPS RELATED TO APOE GENE OF ALZHEIMER'S DISEASE

The involvement of the SNPs of Apo lipoprotein E(ApoE) causing Alzheimer's disease. Alzheimer's disease are major areas of study in neurobiology along with APP, PSEN1,and PSEN2. Four leading genes (APP, PS1, PS2, and APOE) have been determined, as causative elements of this disorder. It has been seen that the mutations in APP, PS1, and PS2 cause early onset AD while APOE is the only gene that has been always marked as a risk factor for late-onset disease. ApoE gene is located in the chromosome 19 on long arm and base pairs from 50,100,878 to 50,104,489 and cytogenetic location 19q13.2. With a total of 3,597 bases, the gene comprises of four exons and three introns. ApoE is polyallelic in nature with the alleles ε1, ε2, ε3, and ε4, out of which the gene ApoE ε4 is found to be the major risk factor for early onset of AD. The human apoE gene contains several SNPs distributed across the gene. The ApoE gene responsible for the Alzheimer's disease has been examined to identify functional consequences of single-nucleotide polymorphisms (SNPs).

The SNPs listed by dbsnp related to APOE gene of Alzheimer's gene is given in figure below.

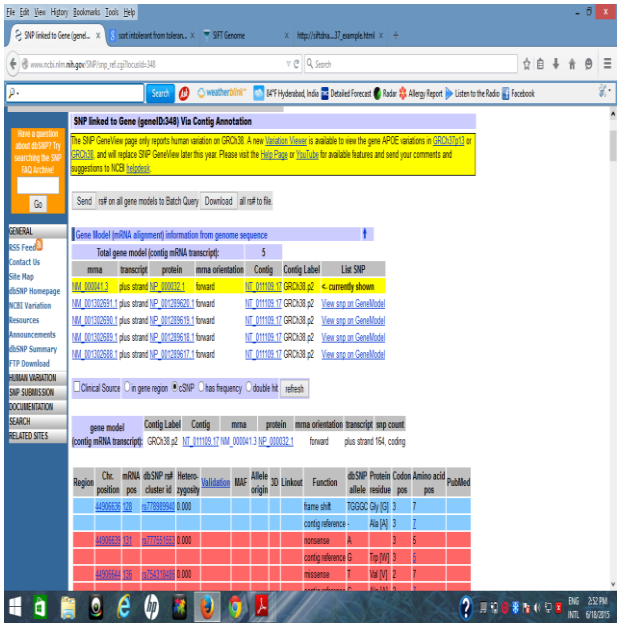

Fig 1: Description of SNPs related to APOE gene

V CONCLUSION

This review paper is an effort to study about different methods and algorithms used in various SNP analysis tools for identifying the deleterious effect of different SNPs .This paper not only describes the features of various tools, It also describes various databases for SNPs. During this era of next generation sequencing and personalized medicine, it is very important to find out very early SNPs which are going to cause a genetic disorder and hence can prevent that disorder or disease. So more and more efficient tools which integrates more efficient methods are the needs of this age.

REFERENCES

[1] Google search " http://ghr.nlm.nih.gov/handbook/genomicresearch/snp/"

[2] S. T. Sherry, M. H. Ward, M. Kholodov et al., "DbSNP: the NCBI database of genetic variation," Nucleic Acids Research, vol. 29, no. 1, pp. 308–311, 2001

[3] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A.McKusick, "Online Mendelian Inheritance in Man (OMIM), aknowledgebase of human genes and genetic disorders," Nucleic Acids Research, vol. 33, pp. D514–D517, 2005.

[4] P. D. Stenson, E. V. Ball, M. Mort et al., "Human Gene Mutation Database (HGMD): 2003 update," Human Mutation, vol. 21, no. 6, pp. 577–581, 2003.

[5] T. U. Consortium, "廄e universal protein resource (UniProt) in 2010," Nucleic Acids Research, vol. 38, pp. D142–D148, 2010.

[6] D. Fredman, M. Siegfried, Y. P. Yuan, P. Bork, H. Lehväslaiho, and A. J. Brookes, "HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources," Nucleic Acids Research, vol. 30, no. 1, pp. 387–391,2002.

[7] T. Kawabata, M. Ota, and K. Nishikawa, "廄e protein mutant database," Nucleic Acids Research, vol. 27, no. 1, pp. 355–357, 1999.

[8] X. Liu, X. Jian, and E. Boerwinkle, "dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions," Human Mutation, vol. 32, no. 8, pp. 894–899, 2011.

[9] Pauline C. Ng and Steven Henikoff," SIFT: predicting amino acid changes that affect protein function" Fred Hutchinson Cancer Research Center, 3812–3814 Nucleic Acids Research, 2003, Vol. 31, No. 13

[10] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," Nature Protocols, vol. 4, no. 7, pp. 1073–1081, 2009.

[11] V. Ramensky, P. Bork, and S. Sunyaev, "Human nonsynonymous SNPs: server and survey," Nucleic Acids Research, vol. 30, no. 17, pp. 3894–3900, 2002.

[12] Y. Bromberg and B. Rost, "SNAP: predict effect of nonsynonymous polymorphisms on function," Nucleic Acids Research, vol. 35, no. 11, pp. 3823–3835, 2007.

[13] R. Jiang, H. Yang, L. Zhou, C. C. J. Kuo, F. Sun, and T. Chen,"Sequence-based prioritization of nonsynonymous singlenucleotide polymorphisms for the study of disease mutations," American Journal of Human Genetics, vol. 81, no. 2, pp. 346–36

[14] S. Chun and J. C. Fay, "Identi□cation of deleterious mutations within three human genomes," Genome Research, vol. 19, no. 9, pp. 1553–1561, 2009.

[15] J. M. Schwarz, C. Rödelsperger, M. Schuelke, and D. Seelow, "MutationTaster evaluates disease-causing potential of sequence alterations," Nature Methods, vol. 7, no. 8, pp. 575–576, 2010.

[16] M. X. Li, H. S. Gui, J. S. H. Kwan et al., "A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases," Nucleic Acids Research, vol. 40, no. 7, p. e53, 2012.

[17] L. Kjong-Van and C. Ting, "Exploring functional variant discovery in non-coding regions with SInBaD," Nucleic Acids Research, vol. 41, no. 1, p. e7, 2013

[18] L. Kjong-Van and C. Ting, "Exploring functional variant discovery in non-coding regions with SInBaD," Nucleic Acids Research, vol. 41, no. 1, p. e7, 2013.

[19] G. M. Cooper, D. L. Goode, S. B. Ng et al., "Single-nucleotide evolutionary constraint scores highlight disease-causing mutations," Nature Methods, vol. 7, no. 4, pp. 250–251, 2010.

[20] Jiaxin Wu and Rui Jiang, "Prediction of Deleterious Nonsynonymous Single-NucleotidePolymorphism for Human Diseases" Hindawi Publishing Corporation,the Scientific World Journal, Article ID 675851, 10 pages,2013

[21] Phil Hyoun Lee and Hagit Shatka**,"** An integrative scoring system for ranking SNPs by their potential deleterious effects", Bioinformatics,Volume 25, Issue 8Pp. 1048-1055,2009

[22] P. K. Krishnan Namboori & K. V. Vineeth & V. Rohith &Ibnul Hassan & Lekshmi Sekhar & Akhila Sekhar &M. Nidheesh," The ApoE gene of Alzheimer's disease (AD)" Funct Integr Genomics DOI 10.1007/s10142-011-0238-,2011