Bio-computing approaches and tools for identification of single nucleotide polymorphisms: a comprehensive review

Neelofar Sohi* and Amardeep Singh

Department of Computer Science and Engineering,
University Patiala,
Punjab, India
Email: neelofarsohi7@gmail.com
Email: amardeepsingh@pbi.ac.in
*Corresponding author

Abstract: Single nucleotide polymorphisms (SNPs) are the most common form of genetic sequence variations. SNPs involve substitution of one nucleotide for another in the DNA sequence. SNPs play an important role in phenomenon of genetic diversity, evolution of species, trait differences, individual's response to a particular drug and they might lead to complex multi-factorial and common diseases such as cancer, inflammatory disease, cardiovascular disease and so on. Identifying and analysing these SNPs is highly important for finding their role in causing diseases and for prevention of diseases. This paper presents a comprehensive review on single nucleotide polymorphisms (SNPs), association of SNPs with diseases, identification of SNPs, tools and methods for SNP identification and SNP identification approaches.

Keywords: single nucleotide polymorphisms; SNPs; SNP identification tools; SNP identification approaches; genetic variations.

Reference to this paper should be made as follows: Sohi, N. and Singh, A. (2021) 'Bio-computing approaches and tools for identification of single nucleotide polymorphisms: a comprehensive review', *Int. J. Computer Aided Engineering and Technology*, Vol. 15, No. 4, pp.427–457.

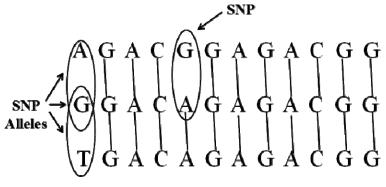
Biographical notes: Neelofar Sohi is an Assistant Professor in the Department of Computer Science and Engineering, Punjabi University, Patiala (Punjab, India). She is pursuing PhD in the area of Biocomputing. Her areas of interest are biocomputing, bioinformatics, DNA computing and quantum computing.

Amardeep Singh is a Professor in the Department of Computer Science and Engineering, Punjabi University, Patiala (Punjab, India). He has a teaching experience of more than 20 years. His areas of interest are biocomputing, DNA and quantum computing, soft computing and software engineering. He has supervised more than 12 PhD thesis and around 40 MTech dissertations. He has two books to his credit. He has been the coordinator of three major projects under Ministry of Human Resource Development, Govt. of India.

1 Introduction

With completion of Human Genome Project (HGP), in 2003, enormous amount of genetic data was generated as 20,500 human genes were revealed and 3.3 billion base pairs of human genome were sequenced. In the past two decades, various endeavours and progress in genome sequencing led to generation of massive data on genetic variants, in particular, SNPs (NHGRI, 2015). Projects like Genome-Wide Association Studies (GWAS) and The Cancer Genome Atlas (TCGA) revealed a number of sequence variations. Now, it is a big challenge to distinguish between harmless and disease-causing variations. It is established from the conducted studies that 99.5% of human genome sequences are identical with a difference of 0.5% only (Sachidanandam et al., 2001). Various forms of sequence variations are there viz. variable number of tandem repeats, insertions ordeletions(indels) and Single nucleotide polymorphisms (SNPs). SNPs are the most common form of genetic sequence variations (Collins et al., 1997; Robert and Pelletier, 2018). SNPs involve change of single nucleotide, i.e., adenine (A), guanine (G), cytosine (C) or thymine (T) in the genome sequence (Vallejos-Vidal et al., 2020). SNPs involve substitution of one nucleotide for another in the DNA sequence (Collins et al., 1997). Figure 1 presents an example of SNP.

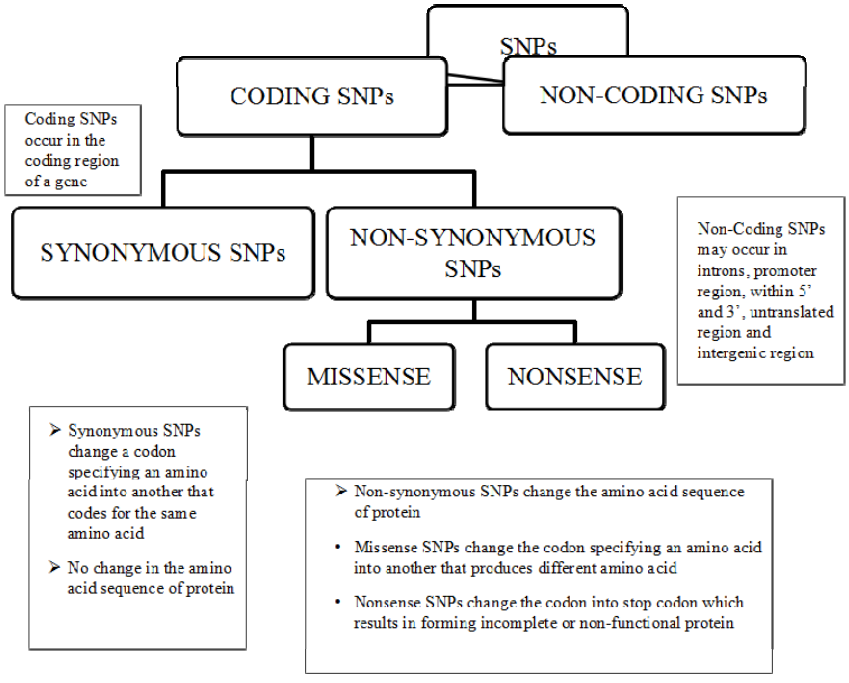
Figure 1 Example of SNP



Mostly, SNPs are bi-allelic as the rate of mutation at a base-pair position is quite low and it is least probable that two point mutations occur at same base-pair position. This makes them a perfect choice as a marker for mapping the history of populations and for other tasks (Borsting and Morling, 2013). SNPs are found to occur at every 100–300 bp and allele frequency is higher than 1% (Wang et al., 1998). Under SNPs, if the change of nucleotide occurs in this way: A <-->G or C <-->T, it is termed as transition and if it occurs in this way: A <-->C or T or G <-->C or T, it is termed as transversion. Transitions occur more commonly than transversions (Robert and Pelletier, 2018). There are different types of SNPs based on their genomic location hence responsible for different kind of functional effects as presented in Figure 2. SNPs occurring in the coding region of SNPs, i.e., exons are called coding SNPs whereas those occurring in the non-coding regions, i.e., introns, promoter regions, untranslated regions or intergenic regions are called non-coding SNPs. Coding SNPs are of two types: synonymous and non-synonymous SNPs. Synonymous SNPs are those which change a codon into another codon that codes for same amino acid. Hence, there occurs no change in the protein whereas in non-synonymous SNP, a codon is changed into another codon forming a different amino acid. Non-synonymous SNPs are further of two types: mis-sense and non-sense. Mis-sense SNPs occur when codon specifying an amino acid changes into codon forming different amino acid. Non-sense SNPs occur when codon specifying an

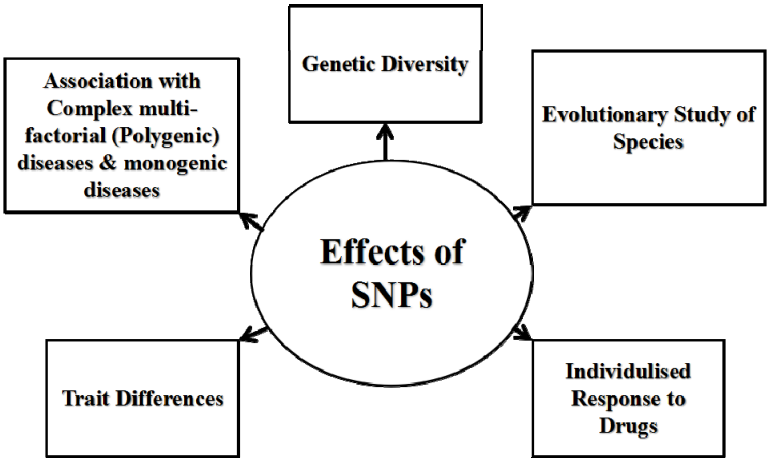
amino acid changes into stop codon leading to formation of incomplete or non-functional protein (Robert and Pelletier, 2018).

Figure 2 Different types of SNPs



There are so many studies which unravel the functional impact of SNPs and other mutations on genes (as shown in Figure 3).

Figure 3 Various effects of SNPs

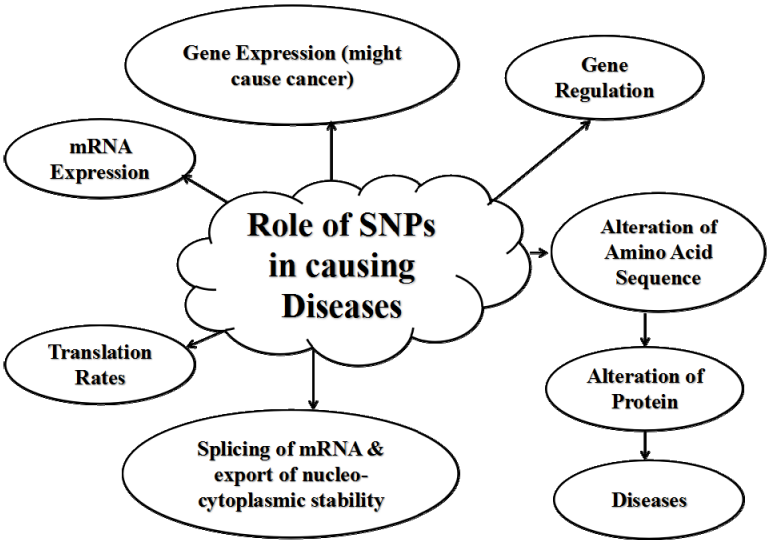


These are excellent biomarkers for disease diagnosis and prognosis (Srinivasan et al., 2016). SNPs might be responsible for genetic diversity, evolution of species, trait

differences, individual's response to a particular drug (pharmacogenomics) and complex multi-factorial and common diseases like cancer, inflammatory disease, cardiovascular disease and so on (Hassan et al., 2016). SNPs might alter the amino acid sequence hence altering the function of protein. Altered protein might lead to abnormalities and diseases in humans. SNPS affect an individual's response immune-response towards pathogens and diseases hence they increase susceptibility of a human to develop some disease (Wijmenga and Zhernakova, 2018). Understanding the association of sequence variations with diseases can enable prevention and early diagnosis and treatment of diseases. Humans having presence of these disease-causing SNPs are genetically predisposed to the risk of developing that disease (Collins et al., 1997; Sohi and Singh, 2018).

Hence, identifying and analysing these SNPs is highly important for finding their role in causing diseases (as shown in Figure 4) (Wijmenga and Zhernakova, 2018). Sequence variations like SNPs can affect the gene expression in different ways depending upon genomic location of the variant. SNPs may occur in the coding region (exons), intergenic region or non-coding region (introns) (Ahmad et al., 2018). If the SNP occurs in transcriptional regulatory element, it might affect mRNA expression. If the SNP occurs in genes, it might affect splicing of mRNA, export of nucleo-cytoplasmic, stability, and translation. If the SNP lies within a coding sequence, it might lead to formation of a different amino acid. This is termed as non-synonymous SNP (Robert and Pelletier, 2018).

Figure 4 Various ways in which SNPs cause diseases



Non-synonymous single nucleotide polymorphisms (nsSNPs) might alter gene regulation and function of proteins. It is observed that non-synonymous SNPs constitute around 2% of total known SNPs which have association with genetic diseases (Hassan et al., 2016). If the mutation is synonymous then same amino acid is encoded. It affects the translation rates or half-life of mRNA. If the SNP forms such an amino acid sequence that produces a premature stop codon, it might produce truncated protein (Robert and Pelletier, 2018). SNPs lying in non-coding regions viz. 3'UTR and 5'UTR, and SNPs lying in microRNA

(miRNAs or mRNA) binding sites which are called mirSNPs can affect miRNA's function and also gene expression. These might lead to cancer. Rigorous testing is required to establish association of a SNP with some disease (Hassan et al., 2016).

1.1 Motivation

This paper presents a comprehensive review on SNPs, association of SNPs with diseases, identification of SNPs, tools and methods for SNP identification and SNP identification approaches. There are a number of review papers on SNPs in the past years. Following are some of the motivating factors for this review paper:

1 Lack of a comprehensive review paper covering majority of tools and approaches for SNP identification.

2 Lack of a paper with discussion on background and importance of SNPs along with discussion on tools and approaches for their identification.

3 Lack of a paper covering all these aspects since 2000s up to 2020.

This paper aims to cover all the relevant literature related to the problem under study beginning from Chakravarti's (1999) pioneering work published in 1999 up to 2020. In Section 2, findings of other researchers, as presented in their review work are summarised. Section 3 covers various tools and approaches for SNP identification. In this section, underlying principle, workflow and findings of each tool are discussed. Next, findings of conducting this review are discussed and future scope of this study is also described. At the end of the paper, a table is presented giving brief overview of various state-of-the-art tools for SNP identification.

2 Studies reviewing bio-computing approaches and tools for SNP analysis

Various researchers have reviewed existing tools, algorithms and databases catering to identification, analysis and annotation of SNPs, and other genetic variations. A journal named Nucleic Acids Research (NAR) comes up with an article on annual basis that covers new biology databases. A total of 92 online biology databases were reported in this journal's 2012 29th edition. There were 1380 databases categorised into 14 categories and 41 subcategories by NAR online Molecular Biology Database collection. This collection is available at http://www.oxfordjournals.org/nar/database/a/. Table 1 presents summary of some of the prominent tools available for SNP identification and analysis.

Chakravarti (1999) have discussed about different sequence variations in his paper. Authors discussed SNPs with special focus stating that SNPs are the single base differences between sequences. These are the most common form of sequence variation. SNPs are abundant in the human genome which occur at a density of around 1 per kilobase (kb) of DNA. Next, author discussed various approaches to understand genetic variations that underlie complex diseases (Chakravarti, 1999).

Table 1 Summary of some prominent tools

SNP calling tool	*Type of tool*	*Language/platform used*	*Technique used*	*Year*	*Funding*	*Homepage*
LD annot (Prunier et al., 2019)	Pipeline	Supported on Unix or Windows in bash developer mode	Finds LD and finds annotation from all those regions which have genetic association with the candidate polymorphisms. LD-annot calculates an average distance (in bp) between two SNPs based on specified r2 threshold in entire dataset.	2019	-	https://github.com/ArnandDroitLab/LD-annot
Arshad's technique (Arshad et al., 2018)	Mining of SNPs from dbSNP database (for TAGAP gene)	-	Fivein silico tools namely SIFT, PROVEAN, PolyPhen-2, PhD-SNP and SNPs and GO used for predicting effects of nsSNPs on structure or function of TAGAP protein.	2018	-	SIFT: http://sift.jcvi.org/www/SIFT_seq_submit2.html PROVEAN: http://provean.jcvi.org/index.php PolyPhen-2: http://genetics.bwh.harvard.edu/pph2/ PhD-SNP: http://snps.biofold.org/phd-snp/phd-snp.html SNPs&GO: http://snps.biofold.org/snps-and-go/snps-and-go.html
Ensembl variant effect predictor (VEP) (McLaren et al., 2016)	Open-source, free online, platform independent tool	Written in Perl language	Used for analysis, annotation and prioritisation of sequence variations in both coding and non-coding regions using GENCODE and Reference Sequence (RefSeq) of National Center for Biotechnology Information (NCBI)	2016	Funded project	https://www.ensembl.org/vep
Mohamed M. Hassan study (Hassan et al., 2016)	Set of tools used	-	Different tools used for extracting the SNPs from SNP databases and predicting their effects are SIFT, polyphen-2, SNAP2 server, I-Mutant suite, CPH models, UCSF Chimera Model Software, automatic protein structural analysis and information using HOPE server, PolymiRTS database (3'UTR), effect of SNPs within 5' UTR on transcription factor binding sites and effect of 3'/5' splice sites SNPs/Indels (HSF tool). SNPs are extracted from dbSNP, UniProt, HapMap, 1,000 genomes project, gene bank, and ClinVar	2016	-	-
In Silico Genotyper (ISG) (Sahl et al., 2015)	Pipeline	Written in Java language	MUMmer and BWA for alignment and GATK for SNP calling	2015	Funded project	https://github.com/TGenNorth/ISGPipeline
kSNPv3 (Gardner et al., 2015)	-	Supported on Linux/MAC OS	Faster and provides flexible annotation in two modes. Its input file consists of paths to genome files instead of genome sequences	2015	Funded project	http://sourceforge.net/projects/kSNP/files/
Pers technique (Pers et al., 2015)	A tool named 'SNPsnap server' is developed based on this approach	-	A set of query SNPs associated with some disease are matched with known set of SNPs using some parameters such as minor allele frequency, linkage disequilibrium, distance to nearest gene and gene density	2014	Funded project	http://www.broadinstitute.org/mpg/snpsnap/

Table 1 Summary of some prominent tools (continued)

SNP calling tool	*Type of tool*	*Language/platform used*	*Technique used*	*Year*	*Funding*	*Homepage*
kSNPv2 (Gardner and Hall, 2013)	-	Supported on Linux and Mac OS	Step I: Enumerate k-meroligos using open source code jellyfish Step II: Next, it searches for SNP from list of central base variants Step III: Identified SNPs are verified using MUMmer Step IV: Maximum likelihood, trees are formed based upon SNP locations Step V: Annotation of SNPs is done using Genbank files taken from NCBI	2013	-	https://sourceforge.net/ projects/ksnp
snpTree (Leekitcharoenphon et al., 2012)	Web-server	-	BWA and SAM tools (for raw reads)	2012	Funded project	No longer available
SNiPlay (Dereeper et al., 2011)	Web-based application and pipeline	Web interface is coded in Perl CGI scripts on an Apache web server. JavaScript and Ajax technologies enable the interaction	Polymor-find program combines Phred/Phrap/Consed software suite with Polyscan program	2011	Funded project	http://sniplay.cirad.fr/
SAM tools (Li et al., 2009)	Package	Written in C language	Useful for indexing, variant calling and as an alignment viewer	2009	Funded project	http://samtools.sourceforge.net/
BCF tools (Li et al., 2009.	-	Written in C language	Useful for manipulating variant calls in VCF or BCF formats	2009	Funded project	http://www.htslib.org/
PineSAP (Wegrzyn et al., 2009)	Pipeline platform	Written in Perl language; Supported on LINUX/UNIX	Combination of Phred for base-calling; Phrap and Probcons RNA for sequence alignment; Polybayes and Polyphred for SNP and indel identification	2009	Funded project	http://dendrome.ucdavis.edu/adept2/resequencing.html
VarDetect (Ngamphiw et al., 2008)	-	Compatible with Microsoft Windows, Linux and Mac OSX	Base-calling is improved by partitioning and re-sampling. Primary and secondary peaks are identified depicting primary and secondary alleles respectively. Difference between observed and predicted peak intensity ratio is used to predict the SNP	2008	Funded project	http://www.biotec.or.th/GI/tools/vardetect
AlleleID (AlleleID: design qPCR and microarray assays for related organisms)	-	-	Based upon Clustal-W for MSASNPs fetched from Genbank	-	For commercial use (price: rs. 253,520)	http://www.premierbiosoft.com/

Table 1 Summary of some prominent tools (continued)

SNP calling tool	*Type of tool*	*Language/platform used*	*Technique used*	*Year*	*Funding*	*Homepage*
PolyScan (Chen et al., 2007)	-	-	Combination of Phred for base calling, cross-match for sequence alignment, noise reduction to distinguish between true peaks and background noise and SNP identification based on concept of drop in peak height	2007	Funded project	http://genome.wustl.edu/tools/software/polyscan.cgi
QualitySNPng (Tang et al., 2006)	Standalone app with GUI	Written in C, Perl and PHP4S	Alignment with CAP3 and variant calling with 3 filters	2006	Funded project	http://www.bioinfomatics.nl/tools/snpweb/
SNP3D (Yue et al., 2006)	Web resource and database	Supported on LINUX through Apache software	Module I: identifies candidate genes for a specific disease. Module II: Provides information about relationship between set of candidate genes. Module III: Predicts effect of non-synonymous SNPs on function of proteins	2006	Funded project	http://www.SNP3D.org
SNPHunter (Wang et al., 2005)	Web-client (dependent on dbSNP for obtaining new SNP and for updation)	Implemented in Microsoft VB.NET	Step I: SNP search module-searches for SNPs from NCBI dbSNP Step II. SNP management: manages information retrieved by SNP search module Step III. Locus link module: takes a list of locus link gene Ids and conducts batch-mode SNP search Step IV. Filter SNP module: automatic and manual filtering to filter out SNPs	2005	Funded project	http://www.hsph.harvard.edu/ppg/
Zhang's tool (Zhang et al., 2005)	-	-	Combination of Phred for base calling, SIM for sequence alignment &and neighbourhood Quality Standard for assessing whether a variation is a true variation or not	2005	Funded project	http://lpg.nci.nih.gov
InSNP (Manaster et al., 2005)	-	Supported on Windows/ LINUX	Step I. InSNP: aligns the given sequence against the reference sequence using a simple word match Step II. Base-calling Step III. SNP Identification by identifying positions in the sequence that are different from the reference sequence. Step IV. Verification of the SNPs is done by human experts by visually inspecting them.	2005	Funded project	www.mucosa.de/insnp/
PupaSNP (Conde et al., 2004)	Web-based searching tool	-	This tool takes a list of genes or chromosomal coordinates and retrieves SNPs using web-based tools. It extracts their functional information using OMIM and gene ontology	2004	Funded project	http://pupasnp. bioinfo.cnio.es.

Table 1 Summary of some prominent tools (continued)

SNP calling tool	*Type of tool*	*Language/platform used*	*Technique used*	*Year*	*Funding*	*Homepage*
SNPbox (Weckx et al., 2005)	Web-server	-	Blast	2004	Funded project	No longer available
SNPicker (Niu and Hu, 2004)	Web-client (part of web-based SeqVISTA suite)	-	REBASE database used for design of primers and SNP detection	2004	Funded project	No longer available
EnsMart (Kasprzyk et al., 2004)	Online retrieval system/database	-	Query based retrieval system from Ensembl database	2004	Funded project	No longer available
viewGene (Kashuk et al., 2001)	-	Java application	Fasta, Micropeats, RepeatMasker, cross match, blast used for sequence alignment; UCSC genome browser used for SNP detection	2002	Funded project	No longer available
SNPper (Riva and Kohane, 2002)	Web-server	-	Web-based application extracts data from public databases	2002	-	No longer available
Sachidanandam et al. (2001)	A map with 1.42 million SNPs distributed across entire human genome	-	This map integrates SNPs made available by a number of projects viz. The SNP Consortium (TSC), Human Genome Project (HGP), White head Institute, Sanger Centre and Washington University. Detection of SNPs was done using two algorithms viz. polybayes and neighbourhood quality standard.	2001	Funded project	-
Genotools SNP manager	SNP analysis software	-	-		Funded project	No longer available
NCBI tools	Hosted online	-	Set of tools used for SNP calling: dbSNP, Variation Viewer and OMIM/ medgen	-	Funded project	https://www.ncbi.nlm.nih.gov/
Newberg's technique (Picoult-Newberg et al., 1999)	Mining procedure	-	Step I: Phred for base-calling II.Phrap for sequence alignment III. Four-filter proceure: involves identifying sequence mismatch by its type; removing low quality SNPs, assessing quality of base calling	1999	-	Results can be found in NCBI SNP database
PolyPhred (Nickerson et al.,1997)	Automated program	-	Combination of Phred, Phrap and Consed Based on a four-phase procedure drop in peak height indicates SNP at a location	1997	Funded project	http://snp.ims.u-tokyo.ac.jp

It is reported in the review that 60,000 within-gene SNPs i.e. coding SNPs have been found by International SNP Map Working Group. There is a coding SNP in every 1.08 kilobases of gene sequence. A SNP is found to be present in 93% of genes out of which 98% lie in the range of 5 kilobases of some SNP. Nucleotide Diversity is defined as the number of base differences between two genomes divided by total number of base pairs. It is a good measure of variation. This measure captures factors of history and biology that the human genome underwent. The purpose of human SNP map is to find out role of genes in causing multi-gene diseases. This whole research and understanding about SNPs opens up new research path. First is to understand natural selection in human population. Here, SNPs prove helpful to find areas with very low variation levels to find out genomic areas with beneficial mutations. Second, biggest area of research becomes to understand molecular basis of variations which cause diseases. Molecular Anthropology is an important area to understand molecules and knowledge that molecular diversity gives about the evolution of humans.

Gray et al. (2000) have presented a comprehensive review on SNPs. The study projects that SNPs are the markers which can potentially be used to test the association of a genetic variant and a disease. In 1980s, SNPs were the most favoured variant for finding association with diseases. In 1990s, simple tandem repeats (STRs) became popular marker for finding association of a gene with disease. STRs especially became popular because they showed high variation in alleles in the number of repeat units. Second, they were evenly and widely distributed across human genome. Their typing was possible using PCR amplification. STRs were especially useful in linkage studies. In linkage studies, STRs were used for pedigree analysis to identify genes responsible for a monogenic disease. Then the focus of research drifted from single-gene diseases towards complex multifactorial diseases such as osteoporosis, diabetes, cancer, cardiovascular and inflammatory diseases involving multiple genes. These diseases are a big social burden. STRs are helpful in pharmacogenetics (or pharmacogenomics) in order to design the treatment for an individual based on individual drug response. In multifactorial diseases there is a combined effect of multiple genes all having small contribution towards causing the disease. If linkage analysis studies are applied for identifying those causative genes in multifactorial diseases, sometimes there is very little difference in frequency of some polymorphism in group of cases – with disease and group of controls-unaffected individuals. Hence, linkage studies bear less fruits for multifactorial diseases. Association studies began to be increasingly used for multifactorial studies. Therefore, for population based, case-control studies, STRs are not a good choice as markers because they have high mutation rates. In comparison to STRs, SNPs are highly abundant and stable marker owing to low rates of mutation. Also, STRs are surrogate markers which means that an STR might not be associated with a disease but some other STR in its neighbourhood can be. SNPs too exhibit this surrogate nature but they are many times responsible for some disease and have functional effects. This review focuses upon identification of SNPs, SNP based association studies, frequency of SNPs across human genome and SNP genotyping methods.

There are two methods of performing association studies. First, a SNP causing a functional effect can be tested for association with a disease. Second, study of linkage disequilibrium (LD) using a SNP as marker. LD measures the degree of association of between two genetic markers further identifying genomic region associated with a disease. Next, SNP identification is discussed in the review. It suggests that SNPs emerge as primarily important tool to be used as a marker for association studies. Therefore, SNP

identification becomes very important. Four methods are suggested for this: Detection of single strand conformation polymorphism, heteroduplex analysis, direct DNA sequencing and the latest variant detector arrays. Next part of the review discusses frequency of SNPs across genome. SNP frequency within a gene and the pattern of occurrence of SNPs has been investigated in a number of studies. By the end of 2001, it was aimed by TSC to broaden the range of SNPs found in genome up to 300,000. Next part of the review suggests SNP genotyping methods that can be used for their identification. The technologies suggested were laboratory based such as fluorescent micro-array based systems (affymetrix), fluorescent bead-based system, automated enzyme-linked immunosorbent assays and so on. It is concluded in the review that with SNPs as markers and association studies, new avenues for understanding of genetic diseases and traits will open in future (Gray et al., 2000).

Mooney (2005) presented a comprehensive review on SNPs. In this review all the resources and tools are reviewed which can be used for predicting functional effects of variants. All prominent web resources for annotating and analysing SNPs are reviewed. Variants are categorised according to their association with genes. Presence of variants in a particular region of gene viz. exonic region, intronic region, upstream or downstream of the gene reveals great information. NCBI dbSNP, OMIM and Ensembl are few resources which annotate the variants. SNPs are identified and annotated using resources such as GoldenPath, UCSCGenome Browser and genome assembly. A prominent resource containing Ensembl is a primary resource of genomic information followed by SNPper. This provides good quality variant calling and annotation. National Cancer Institute as a part of Cancer Genome Anatomy Project's Genetic Annotation Initiative(CGAP-GAI) came up with new tools for SNP identification and analysis. SNP annotation is done using two major tools viz. PolyPhen and SIFT. nsSNPs might impair the normal functioning of genes. It has been established from the findings obtained from sorting intolerant from tolerant (SIFT), that 25% of nsSNPs as reported in dbSNP cause modification of protein function. A good review and classification of non-synonymous SNPs is performed by Sunyaev et al. (2001) where he finds 20% of the nsSNPs to be altering the protein function. The study conducted by Chasman and Adams (xxxx) has found 26% to 32% nsSNPs to be affecting the protein function. A comparative study of disease-causing variants found by HGMD and dbSNP was conducted by Wang and Moult (2001). Functional nsSNPs were analysed by Ng and Henikoff (2006) in their study. Saunders and Baker (2002) adopted machine-learning approach for analysis. It is quite complex to understand effect of variants on gene expression (Mooney, 2005).

Kim et al. (2006) performed a pilot study of LD and haplotype structure of 200 kb region of 22q13.2 in Korean population. In this study, 165 SNPs were identified from the region using direct sequencing. This data was compared with dbSNP database (build 124). 76 SNPs were used to analyse the patterns of LD, haplotype diversity and recombination rates from Korean study population and compared with HapMap database. LD and haplotype frequencies found from Korean and Japanese population were found to be highly similar with high degree of correlation between high LD and low recombination frequency. Patterns of LD were found to be similar in Han Chinese, Japanese, Korean and CEPH population. Haplotype frequencies were, however, significantly different between them. Purpose of International Haplotype Mapping Project (HapMap) is to find out LD in the human genome. In the 200 kb region belonging to chromosome 22, LD patterns were analysed using D′ and r2 between pairwise

combination of markers using Haploview version 3.2 (http://www.broad.mit.edu/mpg/haploview/) (Kim et al., 2006).

Sripichai and Fucharoen (2007) in their paper have discussed about basic concept of SNPs and importance of SNPs. SNPs are understood to enable reconstruction of history of genome owing to their inheritability from one generation to the next. They can also be used in study of evolution of species and in population studies. Next, authors discuss about annotation study based on SNPs. They define an association study as the one conducted to establish relationship of a disease to the region of genome. Next, concept of LD is discussed in the paper. It is stated that if a factor leads to enhanced risk of some disease this implies higher frequency of that factor in group of cases – ones with disease than the group of controls. The study suggests candidate-gene approach as another approach to perform association study where genes with pre-known associations or linkages are picked for testing association. The authors describe SNPs as genetic variations which cause alteration of one of the nucleotides viz. adenine (A), thymine (T), cytosine (C) or guanine (G). SNPs are biallelic form of polymorphisms. SNPs are the most common source of genetic variation, They are responsible for 90% of total human genetic variations. SNPs are found in both coding and non-coding genomic regions. They are highly important owing to their role in reconstructing genome, finding evolution of species, population study and their role in predisposing the person to diseases. They act as good markers for building high-density genetic maps which are further used in association studies. SNPs are good markers as they are abundant, inheritable, evolutionary stable. A study conducted to find a relationship between a phenotype and one or more regions of the genome is known as association study. Frequency of SNPs is found in the two given populations which vary for presence of phenotype. It is based on the fact that if a variant has association with a disease then individuals suffering from that disease will have higher frequency of that variant in comparison to individual with absence of that disease. The authors describe this kind of association as linkage equilibrium (LD). Another approach to check the association is candidate-gene approach. Genetic factors leading to risk of diseases have been revealed through large-scale studies such as Alzheimer's disease (APOE), type 1 diabetes(human leukocyte antigen (HLA), type 2 diabetes(PPARG), deep vein thrombosis (factor V), myocardial infarction (LTA), stroke (PDE4D) and asthma (ADAM33) (Sripichai and Fucharoen, 2007).

Altshuler et al. (2008) have presented all the relevant concepts related to genetic mapping of Mendelian and multi-gene, complex diseases. The study illustrates that genetic mapping is localisation of genes associated with phenotype. Linkage Analysis is suggested for finding genes associated with single-gene (mendelian) diseases. Linkage Analysis states that markers present on genes showing similar association with some trait lie close to each other in genome. Genome Wide Association Studies (GWAS) are suggested for finding genes associated with complex, multi-gene diseases. Here, a catalogue of common human genetic variations is created and these variations are tested for association with diseases. It was Sturtevant who proposed the linkage analysis method for fruit flies in 1913. Linkage analysis is performed in three main steps: First step is to perform a genome-wide search to find the location; second step is to perform case-control study to find responsible mutation and third step is to identify molecular and cellular functions of the found genes. Several studies of genes associated with Mendelian diseases revealed that approach based on candidate genes is not so successful. Some responsible genes are initially completely unknown. Second, it was concluded that there occur alterations in proteins due to disease-causing mutations and there are many

disease-causing alleles present. Also, there is heterogeneity, incomplete penetrance and variable expressivity related to Mendelian diseases. Genetic mapping was used for common diseases such as hypertension, breast cancer and diabetes but this approach failed to yield fruitful results. Studies on population genetics and genomics led to new approach of localisation of genes through association study. This study compares frequency of a genetic variant in the group of affected and group of unaffected individuals. A new genome-wide approach for performing association study came up in the middle of 1990s. It was proposed to create a catalogue of common genetic variants and then test them for association to diseases. It was in 2002 that the International HapMap Project was started. The idea was to find out frequency of SNPs and LD pattern in human genome in 270 samples taken from Europe, Asia and West Africa. Around 1 million SNPs were identified by 2005 and around 3 million by 2007. It was found that the common SNPs have correlation to one or more close proxies (Altshuler et al., 2008).

Johnson (2009) has presented a comprehensive review on SNPs. The study reports that there has been enormous research on genes which contribute to monogenic and complex polygenic diseases. Also, there exist wide variety of bioinformatic databases, software and resources for storage and analysis of genetic data. This review focuses on SNPs, issues related to SNPs and bioinformatic resources including tools and databases containing information about SNPs. SNPs emerge as a marker which are easiest to identify in comparison to other variants such as indels, microsatellites, copy number variants and epigenetic markers, which are also associated with diseases. It reports that there are more than 800 databases of information about human genetic variation where there are few which are most prominently used. Such data resources are classified into three major categories viz. Common genetic variation databases, rare genetic variation databases and databases of variation where there is addition of some functional information about the variation. The largest database of common genetic variation is NCBI's dbSNP. This resource is a repository to variants identified through HGP discovered a significant number of common variants. This database is the primary freely available resource that caters to various tasks related to variations such as mapping of known variants to the human genome, providing identifiers for known and novel variants, identifying known variation within a gene, identifying functional effects of variants, design of assays for measuring a particular variants and calculating frequency of an allele of a variant in some population. The dbSNP variants are added to NCBI,UCSC and EMBL in order to provide integration of SNP data with other genome data. Querying of SNP data in batches and downloading of information from dbSNP is also available. International HapMap Project is another database containing SNP data. The HapMap project aimed at estimating allele frequencies and LD patterns among common human genetic variations. The data can be downloaded and can be viewed using HapMap browser. The 1000 Genomes Project is another endeavour that released first lot of data in 2009 and is a good source of human genetic variation including both common and rare variation. A number of databases are there for identifying variation within and across human population such as Japanese SNP database (JSNP), Thai SNP database, Taiwan-Han Chinese SNP database, SNP@ethnos8, CEPH genotype and ALFRED. There are few more resources such as dbGAP, OMIM, Human Genome Epidemiology (HuGE) and Genetic Association Database (GAD). Many of these databases are based upon information in dbSNP and HapMap (Johnson, 2009).

Oeveren and Janssen (2009) have presented a review of SNP mining tools. They have categorised the tools as de novo tools and reference sequence based tools. de novo tools

are AutoSNP, QualitySNP and MAVIANT. Reference based tools include PolyBAyes, PolyPhred, NovoSNP and SNPdetector. Also, a general procedure for SNP mining is described (Oeveren and Janssen, 2009).

Mooney et al. (2010) have presented an excellent comprehensive review on bioinformatic tools for identifying disease gens and SNPs. Firstly, the authors discuss online databases which contain SNP data and disease data such as db SNP, HGMD, OMIM, PharmGKB, dbGAP and so on. Second, authors have presented available tools for predicting genes associated with diseases such as Fit SNPs, Endeavour Algorithm, GeneSeeker, Gene2Disease, SUSPECTS, PROSPECTR, TOM, Prioritiser, KEGG, BIND, HPRD, Gentrepid, Phenophred and so on. Third, authors have discussed tools for SNP searching, visualisation and annotation such as UCSC Genome Browser and Ensembl. A large number of databases are available that contain SNP and other variation data. The SNPs database, dbSNP (dbSNP, http://www.ncbi.nlm.nih.gov/SNP/) is the most prominent source of SNP data. Other databases are Human Gene Mutation Database (HGMD, http://www.hgmd.cf.ac.uk/ac/index.php), Online Mendelian Inheritance in Man (OMIM, http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim), Pharmacogenetics Knowledge Base (PharmGKB, http://www.pharmgkb.org/) anddatabase of Genotype and Phenotype (dbGAP, http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap). There are so many databases, tools and web resources containing genomic data that it leads to rich wealth of data. There are a number of tools available for predicting genes associated with diseases such as FitSNPs (Functionally interpolated SNPs; http://fitsnps.stanford.edu/), Endeavour algorithm (http://homes.esat.kuleuven.be/~bioiuser/endeavour/), GeneSeeker (http://www.cmbi.ru.nl/geneseeker/), Gene2Disease (G2D, http://www.ogic.ca/projects/g2d_2/), Gene Ontology (GO, http://www.geneontology.org/), SUSPECTS (http://www.genetics.med.ed.ac.uk/suspects/), PROSPECTR (http://www.genetics.med.ed.ac.uk/prospectr/), OMIM (TOM, http://www-micrel.deis.unibo.it/~tom/), PRIORITIZER (http://pcdoeglas.med.rug.nl/prioritizer/), Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.ad.jp/kegg/), Biomolecular Interaction Network Database (BIND, http://binddb.org/), Human Protein Reference Database (HPRD, http://www.hprd.org/), Gentrepid (https://www.gentrepid.org/), PhenoPred (http://www.phenopred.org/) (Mooney et al., 2010).

Borsting and Morling (2013) have thrown light on SNPs and short tandem repeats (STRs) as markers where STRs are better choice to be used as markers in criminal investigations because for over past more than two decades, STR records of criminals and victims have been stored in national and international databases. SNPs can be better markers for finding relationships as their mutation rate is less of the order of 0.00000001 where as that of STRs is 0.001–0.003. In this paper, authors have thrown light on various applications of SNPs with special focus on advantage of SNP identification for identifying a human. This study does not involve any computer based software or algorithm for solving the problem (Borsting and Morling, 2013).

Bianco et al. (2013) have presented a comprehensive review describing databases useful for biomedical research of genetic diseases, sequences, mutations, gene expression and protein expression data. The paper summarises prominent databases for biomedical literature such as PubMed; database containing genetic data such as NBI, Ensembl, UCSC Genome Browser, GWAS database. Authors have presented the usage of various databases for disease information by studying inflammatory bowel disease using three databases such as PubMed, OMIM and GWASdb. The prominent journal NAR annually comes up with a research paper covering newly created biological databases and

resources. The 19th edition got published in 2012. It lists 92 online databases containing biological data and also listed 100 more research papers on this topic. There is an online database collection, NAR online Molecular Biology Database collection. It lists 1,380 databases and is available at http://www.oxfordjournals.org/nar/database/a/. There are a number of NCBI databases including BioProject database (www.ncbi.nlm.nih.gov/bioproject/), BioSample database (http://www.ncbi.nlm.nih.gov/biosample/), PopSet database (http://www.ncbi.nlm.nih.gov/popset/), Clone database (CloneDB) (http://www.ncbi.nlm.nih.gov/clone/), molecular modelling database (MMDB) (http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml), database of expressed sequence tags (dbEST), database of genomic structural variation (dbVar), Entrez, databases of genotypes and phenotypes (dbGaP), database of major histocompatibility complex (dbMHC), database of short genetic variations (dbSNP), Ensembl and Genome Browser of the University of CaliforniaSanta Cruz and GenScan. There are a number of gene regulation databases including Decipher, Histone, Starbase and microsniper database (Bianco et al., 2013).

3 Bio-computing approaches and tools for SNP analysis

There are a number of bio-computing approaches and tools available for SNP identification, analysis and annotation. Also, there exists variety of tools for analysis of genetic variations. It goes back to 1997 when Collins and team did a breakthrough about sequence variations with their pioneering work highlighting the significance of sequence variations for their association with diseases. There are a number of automated software and tools for SNP identification and analysis. Most of them are pipelines consisting of similar kind of popular tools yielding results of SNP identification. Many of them yield good results. Some of the prominent tools and software are discussed in this section.

Nickerson et al. (1997) have proposed an automated program, PolyPhred, for identification of SNPs. Polyphred is used in combination with Phred, Phrap and Consed. When compared to ABI sequencing software, Phred is found to have higher accuracy and low error rate. There is an error probability associated with each of the bases (Ewing et al., 1998). The principle for identifying polymorphism is to compare sequence traces of homozygotes and heterozygotes where two changes are looked for:

a a significant drop in normalised peak height at a polymorphic site when traces from homozygous and heterozygous individuals are compared

b a second underlying peak at the position.

It uses a four phase procedure to determine a sequence of base calls from processed trace. The first phase involves determining the idealised peak locations, i.e., predicted peak locations. Second phase is to identify the observed peaks in the trace. Third phase involves matching the observed peak locations to the predicted peak locations where some peaks are omitted and some others are split apart. Final step is to check out the unmatched observed peaks, to determine if these actually represent some base but they were not assigned to a predicted peak. Hence, the corresponding base is then inserted into the read sequence. Phrap is used for sequence alignment for aligning the query sequence against the reference sequence. It takes as input the normalised peak areas and quality values retrieved from Phred for each location in the sequence. If another peak gets

detected at a base and the peak height gets reduced, then it is considered heterozygous by PolyPhred. Consed is a tool used for editing, evaluating and viewing the traces (Nickerson et al., 1997). In this Phred/Phrap/Consed setup, monitoring of data quality takes place and then it is displayed as a main component of the system. There is a relationship between sequence quality as found by Phred and performance of PolyPhred. Factors including peak spacing, the relative size of the uncalled and called peaks and the dip in signal between called peaks are used to generate quality measures in Phred. When sequence quality is low (Phred quality = 20), the signal-to-noise ratio is obtained as the ratio of PolyPhred true positives to false positives and is low. The signal-to-noise ratio, i.e., true positive to false positives is enhanced as the scanning window for PolyPhred is set for analysing data at raised quality thresholds. As a part of Japanese Millennium Genome Project, Haga et al. (2002) identified a total of 190,562 genetic variations consisting of 174,269 SNPs and 16,293 insertion/deletions from DNA samples of 24 Japanese individuals using PolyPhred for SNP identification. Data and methods of the study are available at http://snp.ims.u-tokyo.ac.jp (Haga et al., 2002).

Picoult-Newberg et al. (1999) have proposed a strategy to extract SNPs from public expressed sequence tag (EST) databases. 300,000 distinct sequences were taken from ESTs obtained from 19 different cDNA libraries and identification of 850 mismatches was done from contiguous EST data sets without de novo sequencing. Genetic bit analysis (GBA) was the technique used to confirm presence of a subset of these candidate SNPs and estimation of allele frequencies was done in three human populations from different ethnic origins. GBA is a polymerase-mediated, single-base, primer extension technique. This is an approach used for rapid and efficient SNP identification in specific regions and genome-wide. For SNP discovery, it gathers data from sequences from different libraries of cDNAs. The SNPs found in the study were submitted to the National Center of Biotechnology (NCBI) SNP database under submitter handles ORCHID (SNPS-981210-A) and debnick (SNPS-981209-A and SNPS-981209-B).

Authors have proposed a three-step procedure for SNP calling. Firstly, Phred is used for base-calling, identification of bases in the sequence. Next, Phrap is used for sequence alignment to align the query sequence against the reference sequence. It classifies Sequences which do not have enough of similarity information as singlets are categorised by Phrap, also they are removed from the set of assembled contigs. Finally, SNP identification is done through four filters. Filter 1 segregates and removes cluster of mismatches occurring in low quality trace data regions. Filter 1 looks out for window sizes of 5, 10 or 20 bp around location where single base-pair mismatch occurs. Identification of the type of sequence mismatch as base substitution or insertion/deletion is done by filter 2. Function of filter 3 and filter 4 is to keep track of the quality of each base call with respect to its position and frequency in a contig. Filter 3 ignores mismatches occurring in first 100 bases because there is susceptibility of errors in the starting portion of the reads. Filter 4 calls a mismatch as a high quality candidate SNP, if it is found to occur in more than one sequence in a contig. Therefore, mismatches resulting from copying errors are discarded. This approach is very cost efficient as it uses existing sequence resources used for discovery of genes instead of finding markers leading to useful EST SNPs. Number of SNPs in Washington University EST database is doubling in every 6 months and 5,000 new EST sequences are added to it every week. SNP markers are highly important in large-scale analysis of human genotype-phenotype relationships owing to their significance in genetic diseases, as well as their density in the genome and low mutation rate (Picoult-Newberg et al., 1999).

Zhang et al. (2005) have proposed an automated SNP detection tool. The tool has been developed to mimic the human visual inspection process. Firstly, Phred is used for base calling, calculating quality scores and generating primary and secondary peak information for each trace file. Next, alignment of the query sequence against a reference sequence is done using SIM which is based upon Smith Waterman algorithm. There is some sequence variation but still there is an optimal alignment of PCR reads. The alignments are trimmed from the ends.

In the end, variation site is inspected using neighbourhood quality standard. Also each base in its flanking window is checked to exceed a user-defined quality threshold. Identified SNPs are then validated. A base is termed as a true variation if this base and each base in its 4bp flanking region have a Phred quality score greater than or equal to 25 and their sequence similarity has to be greater than or equal to 95%. Height of the Secondary Peak for heterozygous allele must be at least 30% of the height of Primary Peak. If the peak height is lesser than 20% then it is considered as noise. The proposed tool has been compared with the human visual inspection, PolyPhred and NovoSNP and has been found to be superior than them. The tool has been found to have low false positive and false negative rates.

SNP detector was used in the HapMap project. At the stage of discovering SNPs, 48 individuals were selected from four populations viz. Centre d'Etude du Polymorphisme Humain collection, Yoruba individuals from Ibadan, Nigeria, Japanese individuals from Tokyo, Japan and Han Chinese individuals from Beijing, China. Cell lines of the individuals are available from the Coriell Institute for Medical Research (http://locus.umdnj.edu/nigms/products/hapmap.html). Out of a total 11,241 candidate SNPs found in the region, half (51.9%) were novel in comparison to data in build 121 of dbSNP. Out of those SNPs, 80% of the were having a minor allele frequency greater than 0.05. The error rate of SNP detector in comparison to PolyPhred and NovoSNP is quite low. SNP detector is accurate enough to identify SNPs from PCR templates with low false negative rates (2%-6%) and moderate false positive rates (1%-9%) (Zhang et al., 2005).

Chen et al. (2007) have proposed an automated software, PolyScan for indel and SNP detection. It provides de novo detection of heterozygousindels with high sensitivity and enhanced specificity. PolyScan improves the accuracy of SNP identification accuracy as it combines the results of existing SNP detection programs. Most variant identification pipelines are sequential, multi-program approaches such asphred/phrap/PolyPhred or phred/SIM/SNP detector. In such pipelines, errors propagate from one stage to the next. PolyScan integrates base calling, alignment, statistical sequence analysis and indel and SNP identification into one program. An ace file and the PhD file consisting of called bases, positions and quality scores are input into PolyScan. Firstly, the chromatograms are re-analysed based on the called base positions which act as initial conditions and boundaries for locating additional peaks in the four fluorescence channels. Phred is used for base-calling to identify the bases in the sequence. Next, sequence alignment is done in order to align the sequence against the reference sequences using cross-match program. Next, Noise reduction is done to distinguish between true peaks and background noise. The program runs through each fluorescent channel using overlapping 30-bp windows, tracking height, sharpness, and regularity of each peak. Then indel identification is done based on identified indel signatures. Finally, SNPs are identified as doublet peaks whose heights are half of those for homozygous individuals. Drop in the peak height is the indicator of presence of variation. Trace statistics are calculated from the individual channels through integral base re-calling and noise-reduction. Two procedures namely

horizontal and vertical scan are performed. In horizontal scan, distance metrics are computed within reads. It generates significance estimates of observed heterozygous trace patterns based on these computed distance metrics. In vertical scan, peak height is computed. It generates information about heterozygous peak height variation in the individuals. In the study, performance comparison of PolyScan with PolyPhred and mutation surveyor is done using parameters as specificity and sensitivity (Chen et al., 2007).

Ngamphiw et al. (2008) have proposed VarDetect for identification of SNPs and other forms of genetic variation. This is an algorithm for interpreting fluorescence based chromatograms and detect the corresponding nucleotide variations in an automatic mode. In this tool, SNPs are identified from the chromatograms obtained for different nucleotides. Base calling is improved by partitioning and re-sampling technique. Primary peak depicting the primary allele and secondary peak depicting the secondary allele is to be identified. If multiple peaks occur at a particular position, it is indication of presence of SNP. Next, observed peak intensity ratio; Q_o^i is calculated from peak intensities of various peaks at i^{th} location as Q_o = highest peak intensity / (sum of all intensities). Vicinity peak intensity ratio; Q_v^i is calculated relative to two bases to the left and two bases to the right of base call location as Q_v^3 = (I1 + I2 + I 4 + I5) / 4. The difference between observed peak intensity ratio; Q_o^i and vicinity peak intensity ratio; Q_v^i; $Q_v^i - Q_o^i = \delta$ called detection value is calculated which is used to predict SNP. This difference above a defined threshold value is considered significant and is termed as SNP. Chromatogram traces are converted to numeric codes using CodeMap technique. Homozygous bases are converted to 0 and 2 codes and heterozygous base is converted to 1. VarDetect has been compared with PolyPhred, novoSNP, Genalys and Mutation Surveyor using fluorescence-based chromatograms and has been found to be most efficient among them all.

Comparison of VarDetect with PolyPhred (version 6.11 beta), Genalys (version3.3.23a), novoSNP (version 2.0.3) and Mutation Surveyor (trial version 3.23) was done using parameters such as features and accuracy. False positive (FP) and false negative (FN) SNP counts are the main parameters used for comparison. True positive count (TP) conveys the number of predicted SNPs which correspond to the actual and verified SNPs in the given area. FP conveys the number of SNPs predicted by the software which are not actually present. FN conveys the number of true SNPs which were missed from being identified by the software. Efficiency is calculated in terms of precision (TP / (TP + FP)), recall (TP / (TP + FN)) and F-score (2 × (precision × recall) / (precision + recall)). Performance comparison of VarDetect was done with other tools using chromatogram traces utilised in past SNP identification from 15 candidate genes. In a past study, 171 SNPs were validated from a set of 77 reads. These validated SNPs acted as a TP count. VarDetect is compatible with most of the operating systems such as Microsoft Windows, Linux and Mac OSX. VarDetect program available at http://www.biotec.or.th/GI/tools/vardetect (Ngamphiw et al., 2008).

Wegrzyn et al. (2009) have proposed a SNP identification pipeline called PineSAP which also provides multiple sequence alignments. It provides a high-throughput solution for analysis of re-sequencing data consisting of chromatogram files for forward and reverse reads of different individuals. This pipeline is supported for Unix/Linux platform and has been written in Perl. PineSAP is a combination of Phred, Phrap and Probcons

RNA used for base-calling and alignment. Phred is used for Base-calling is done using Phred and two tools namely Phrap and Probcons RNA are used for sequence alignment. Next, Polybayes and PolyPhred techniques are used for SNP and indel identification. Supervised machine learning algorithms are used to extract and process the sequence information to validate or reject the predicted SNPs. Its code is available at http://dendrome.ucdavis.edu/adept2/resequencing.html (Wegrzyn et al., 2009).

Li et al. (2009) came up with a software package called SAM tools used for manipulating alignments in SAM or BAM format. It provides conversion from other alignment formats, removal of PCR duplicates, per-position information generation in pileup format, short indel and SNP identification and presents the alignments in text based viewer. This software package is available in both C and Java languages with little difference in their functionality. SAM tools software package is available at http://samtools.sourceforge.net/ and can be downloaded from https://github.com/samtools/samtools (Li et al., 2009).

BWA is based upon Burrows wheeler transform (BWT). BWT is useful for pattern matching and compression. BWA is suitable for mapping of less-divergent sequences with a large sized reference genome like the whole human genome. Source code of BWA is available at http://bio-bwa.sourceforge.net/. It has four different algorithms namely BWA-backtrack used for sequence reads up to 100 bp, BWA-SW used for sequence reads from 70 bp–1 Mbp, BWA-MEM used for sequence reads from 70 bp–1 Mbp and BWA aln/SAMSE/SAMPE (Li and Durbin, 2009).

BCFtools are used to view the BCF file which contains all the variants. Its source code is available at http://www.htslib.org/ (Li et al., 2009).

AlleleID is a commercially available comprehensive desktop tool specifically designed for bacterial identification, pathogen detection or species identification. It uses ClustalW algorithm for multiple sequence alignment. The product is available at http://www.premierbiosoft.com/. In this study, a demo version of the product, AlleleID 7.84 is used for evaluation. Its market price is 3227 EUR (Rs. 2,53,520) ('AlleleID: design qPCR and microarray assays for related organisms', http://www.premierbiosoft.com).

In Silico Genotyper (ISG) is a parallel, open source tool for identification and annotation of SNPs and insertions/ deletions. ISG is written in Java and has BWA-MEM, MUMmer and GATK as its main components. ISG works on the Queue pipeline system of the BroadInstitute(http://www.broadinstitute.org/gatk/auth?package=Queue). Data can be input in formats of .txt', '.fastq', or '.fastq.gz'for raw reads, '.fasta' format for genome assembly, '.gbk' format for genome annotation, '.bam' format for binary alignment map or 'vcf' format for variant call. If single or paired reads are provided, BWA-MEM is used for alignment of reads against the reference genomein '.fasta' format. SNPs and indels are called with the Unified Genotyper method using thresholds like minimum depth of coverage and allele proportion variation defined by the user.For a specific application, GATK variables can be altered by user. Show-snps function of MUMmer can be used for SNP calling. The position where a SNP is found in one genome, is queried in rest of the genomes. If a location in query sequence does not qualify one user-defined filter then 'N' is placed at that location. If a non-SNP location qualifies all the filters but it is not a SNP then GATK is used for measuring the base quality and coverage at that location. A reference state is used for this purpose which has to be a location of sufficient quality. Otherwise, a '.' or 'N' is placed at that location depending on sufficient coverage or quality, respectively. Its source code is available at https://github.com/TGenNorth/

ISGPipeline. In this paper, performance comparison of ISG with other popular tools such as snpTree, SPANDx, kSNP, CO-Phylog and ParSNP has been carried out establishing ISG as the most efficient tool. snpTree method is similar to ISG but cannot handle hundreds and thousands of sequences. SPANDx is a reference dependent approach which does not work for genome assemblies. kSNP is similar to ISG in terms of provided function whereas different in terms of methodology. It works for both raw reads and genome assemblies but it takes time while concatenating raw reads. CO-Phylog does not produce SNP annotation. ParSNP is not able to handle raw reads, for that it needs an additional step of assembly. 1,000 Escherichia coli and Shigella Genomes retrieved from Genbank were used to compare the speed of ISG with other counterparts. It was found that ISG scales linearly with increasing number of genomes (Sahl et al., 2015).

Tang et al. (2006) have come up with a tool for identification of SNPs and indels called QualitySNPng. It provides a data storage and retrieval system for SNPs, haplotypes and alignments. QualitySNPng is a haplotype-based strategy for detection of synonymous and non-synonymous SNPs from public EST data. There is no need of trace or quality files or the genome sequence data. Haplotypes are the different alleles of a gene. QualitySNPng has been used for SNP identification in potato, chicken and humans. This pipeline program consists of five steps. Firstly, ESTs are assembled using cross_match for removing vectors and CAP3is used for clustering of sequences. Second, clusters are formed with at least four members by analysing alignment information. Third, detection of SNPs is done and differentiation of variations between and within genotypes is done. Fourth, non-synonymous are distinguished from synonymous SNPs using FASTY. Fifth, results are stored into database. The pipeline program is written in standard C-Shell script for a Linux workstation. The individual program is written in the C programming language. At third step, there are three filters for identifying SNPs. First filter finds potential SNPs and separates out variations between and within genotypes. Second filter finds clusters consisting of variations produced as a result of sequencing errors and paralogous sequences. Third filter identifies unreliable SNPs by assigning confidence scores to the SNPs on the basis of sequence redundancy and quality. QualitySNPng is as good as its counterparts and at some points outperforms them. It needs to provide a genomic sequence or sequence quality files. This program identifies SNPs and haplotypes as well. It can also be used for EST based genotyping. One big advantage of this program is a retrieval system which can produce output in different formats. The source code of the tool is available at http://www.bioinfomatics.nl/tools/snpweb/ (Tang et al., 2006).

Wang et al. (2005) have developed a SNP data extraction tool known as SNPHunter. This tool extracts the SNP data such as physical location, flanking sequence and functional class of the SNP from NCBI dbSNP database for the provided input genes. Microsoft Visual Basic .NET has been used for implementing SNPHunter. User queries are delegated by HTTP parser to the databases such as dbSNP, MapViewer, LocusLink and AceView at NCBI. It also does parsing of data. SNPSearch, SNP Management, and LocusLink SNP are its three major components. The SNPSearch module takes gene symbol as input and selects SNPs on the basis of heterozygosity, chromosomal position and functional class. If upstream or downstream sequence of gene are to be included, this can also be specified by the user. The SNP Management module lets the user to retrieve and then manage various kinds of information about SNPs fetched by search module. The LocusLink module takes a list of LocusLink gene IDs (i.e., Entrez Gene IDs) as input and then it conducts batch-mode SNP search through LocusLink. It helps to fetch SNP data

for a set of genes. This tool prepares a summary of SNPs. There comes a new 'Filter SNP' panel. A user defined criteria can be specified for filtering of SNPs. The local filter filters out the SNPs without need of web. It can do filtering in both manual and automatic mode. SNPs can be stored in local disk space. This tool performs batch-mode SNP selection. This process is really fast. A case-control study was conducted involving ten biological candidate genes for type 2 diabetes mellitus. Some of the genes included in study are CAPN10, FABP4, IL6, NOS3, PPARG, TNF, UCP2, CRP, ESR1 and AR. It was a practical example of applying SNPHunter for extracting SNPs present on these genes from dbSNP. All these ten genes were selected based upon their biochemical and physiological functions. SNPs were selected using four criteria.

- Genome coverage: Selected SNPs should span the region of gene and the 30 kb 5' upstream and 30 kb 3' downstream area.
- Priority based on function: There has to be an order of priorities for selecting SNPs where non-synonymous SNPs are top priority followed by synonymous SNPs, ssSNPs, 5'upstream SNPs, 3' downstream SNPs and then intronic SNPs.
- HET-based priority: HET is a parameter that can be calculated using POLYMORPHISM Software. HET threshold is set for intronic and 5' upstreamor 3' downstream region SNPs where SNPs having threshold greater than 0.095 are given high priority whereas no such threshold is set for cSNPs and ssSNPs. Threshold of 0.095 is same as that of MAF greater than or equal to 5%.
- SNP density: Even distribution of SNPs is desirable in gene region, 30 kb 5' upstream and 30 kb 3' downstream regions. The density of SNPs should be 5–50 SNPs/Kb depending upon the size of gene. Genes having size less than 10 kb should have a density of 50 SNPs/Kb. Genes having size from 10 to 100 Kb should have a density of 10 SNPs/Kb and genes having size more than 100kbshould be having a density of 5 SNPs/Kb.

Authors have compared the tool to its other counterparts such as SNPper, SNPicker, SNPBox and viewGene. Comparison has established that SNPHunter works in both ad hoc-mode and batch-mode, works well for a set of genes rather than just one gene and provides both automatic and manual selection of SNPs. The project is available at http://www.hsph.harvard.edu/ppg/. This tool completely relies upon dbSNP database for extracting SNPs. Although this database is the primary source of SNP data still some of other research projects like Seattle SNPs and SNP500 cancer having more SNP data need to be considered for the purpose (Wang et al., 2005).

Ensembl variant effect predictor is used for analysis, annotation and prioritisation of sequence variations in both coding and non-coding regions. This is an open-source, free online, platform independent tool for interpreting the found sequence variations. VEP enables automated annotation of variations reducing the manual efforts and time. Sequence variations which are responsible for diseases, if analysed can lead to better prevention and treatment planning for those diseases. In order to interpret the genetic variants, effect of a variant on a transcript or a protein has to be considered. It is associated with annotation of transcripts and categorisation of variants as coding or non-coding. GENCODE and reference sequence (RefSeq) of National Center for Biotechnology Information (NCBI) are primary source of annotation for humans. If there occurs any change in version or updates of these then annotation gets changed too.

Transcript isoforms and versions should be taken care of for proper interpretation. There is another system of nomenclature for reporting of variants termed as Human Genome Variation Society (HGVS) which relies upon transcripts and proteins too. This problem with annotation leads to ambiguities. A highly robust tool is the need of the hour to overcome the challenges of annotating variants and handling ever-increasing sequencing data. Ensembl variant effect predictor (VEP) came up as an excellent tool to provide annotation and analysis of genetic variants in both coding and non-coding genomic regions. It has been used for analysing features of farm animals, diagnosing humans clinically and for performing research on Genome Wide Association Studies. This tool has also been used for analysis in research projects such as 1,000 Genomes and Exome Aggregation Consortium (ExAC).Tools like Gemini take output of VEP as input. This annotation and analysis tool offers flexibility to a high degree. The Ensembl Variant Effect Predictor software offers a systematic approach for annotation and prioritisation of genetic variants in studies involving sequencing and analysis. Analysis of SNPs, short indels, copy number variants and structural variants is usually quite time-consuming. VEP offers a faster approach for annotation. There is a broad range of sources used for annotations by VEP such as transcripts, regulatory regions, frequencies from previously observed variants, citations, information about clinical significance and predictions of biophysical consequences of variants. How much stable and of how much good quality an annotation is going to be, relies upon the used transcript set. A variety of variant annotations and transcript isoforms are used by VEP for prioritisation of results and reduce manual intervention. VEP takes input in VCF format. Variant Call Format(VCF) is the standard format used in sequencing programs. VEP can also take in the variant identifiers from dbSNP and HGVS nomenclature notations. Variants can be mapped from cDNA or protein coordinates to the genome and vice versa. Output of VEP is in an HTML format, text file, tab-delimited, VCF, GVF, or JSON format. The tab-delimited form is the default format for output. VCF format of output follows a standard form for cross-platform comparison and benchmarking of results (McLaren et al., 2016).VEP is available at https://asia.ensembl.org/Tools/VEP.

Yue et al. (2006) have developed a web resource cum database which provides information about disease-gene relationship. SNPs3D has three modules. One module identifies candidate genes for a specific disease. Second module provides information about relationship between set of candidate genes. Third module predicts effect of non-synonymous SNPs on function of proteins. This software enables searching for genes associated with diseases. Authors have also compiled candidate genes lists for 76 diseases taken from NCBI (Yue et al., 2006).

Gardner and Hall (2013) came up with kSNPv2, an improved version of kSNPv1. kSNPv2 provides SNP gene annotation, better scaling of draft genomes available as assembled contigs or raw reads, estimation of optimal value of k, distribution of packages of executables for Linux and Mac OS and a user guide. The program kSNP is used for finished and draft sequences to identify SNPs and phylogenies. Generally, Multiple Sequence Alignment or a set of pairwise alignments are needed to further locate SNPs. With kSNP, there is no need of a reference genome and multiple sequence alignment. It can scale well for upto gigabases of sequences in single run. kSNPv2 is the second version of kSNPv1 which has been proved to be faster, memory-efficient and provides better SNP annotation. A tool named Mauve performs whole genome multiple alignment with huge memory requirements restricting its usage upto 30 bacterial genomes only. It takes around 70 hours to align 25 genomes. Two tools namely Gegens and BopGenomes

carry out comparison of complete genomes without performing their alignment. The comparison is done based on presence or absence of DNA segments where SNPs are not identified. Analysis of hundreds of bacterial or viral genomes is possible with kSNPv2 in few hours. It can handle raw reads, finished genomes and genome assemblies.

Hence, kSNP is a reference free method. For SNP identification, firstly the user inputs a fast a file of target genomes and also specifies k, length of the flanking sequence including SNP. For example, k = 15 means that each SNP will be flanked by 7 bases on its either side. The SNP lies at the central base of the k-mer, and flanking (k – 1) / 2 bases are there on each side of the SNP defining the SNP locus. First step is to enumerate k-meroligos and their count for each input genome through open source code jellyfish. For raw reads, singleton k-mers that occur only once are dropped as they can probably be sequencing errors. If the SNP conflict occurs that means if there is more than one central base, then that k-mer is also dropped leading to a SNP conflict. Sorting of deleted k-mers is done using merge sorting method. Next, it searches for SNP from list of central base variants. Next, a comparison of found SNP location is done with deleted k-mers due to conflict. Next, allele, i.e., the central base and flanking sequence in the genome are identified. Identified SNPs are verified using MUMmer. Output is available in various formats viz. SNP allelefasta alignment, SNP matrix, and list of SNPs with allele locations. There are a number of methods for calculating phylogeny of SNPs such as Parsimonator, neighbour joining of pairwise distances from SNP difference counts and maximum likelihood. Using maximum likelihood, trees are formed based upon SNP locations present in all of the genome or user-specified genome. Next, mapping of SNP locations to the trees is done. Annotation of SNPs is done using Genbank files taken from NCBI (Gardner and Hall, 2013).

Gardner et al. (2015) proposed kSNPv3, an improved version of kSNPv2. It is faster, provides flexible annotation, its input file consists of paths to genome files instead of genome sequences. Authors have compared k SNPv3 with its counterparts such as ParSNP and RealPhy. Project is available at http://sourceforge.net/projects/kSNP/files/ (Gardner et al., 2015).

kSNP3.0 is an enhanced version of kSNP v2. Genomes can be retrieved using various tools available in kSNP3.0 suite, in order to carry out analysis. This program also provides annotation of SNPs whereas kSNP v2 provided annotation of SNPs within the chromosomes only. In the input file of kSNP3.0, paths of original genome files are included and not genome itself hence providing ease of addition or removal of genome from the list. It offers higher flexibility in annotating SNPs. This process of annotation is usually quite time-consuming. Hence, kSNP3.0 offers two different modes viz. Standard annotation where annotation of some SNP is done based upon the first genome in the list. Second, full annotation where annotation of some SNP is done based upon every genome where it is present. Speed of standard annotation process is quite high. In terms of accuracy, kSNP 3.0 and kSNP v2 stand equal. kSNP 3.0 was compared with other tools like Parsnp, RealPhy and Epstein et al approach.

Parsnp can only be used for aligning finished or assembled genomes and not for unassembled raw reads. RealPhy, at first performs the alignment of raw reads to the reference genome therefore working of RealPhy greatly relies upon accuracy of alignment done. The approach proposed by Epstein et al is for mapping raw reads to complete genome.

Conde et al. (2004) came up with a web-based searching tool for searching SNPs called PupaSNP. This tool takes a list of genes or chromosomal coordinates and retrieves

SNPs using web-based tools. It extracts their functional information using OMIM and Gene Ontology. The PupaSNP web interface is available at http://pupasnp.bioinfo.cnio.es.

A list of genes is given as input to the PupaSNP. Gene list can be fed as the list of gene identifiers such as gene Ensembl IDs,GenBank IDs, Swissprot/TrEMBL IDs. Chromosome location for the genes can also be provided where all genes in that region are selected. Tools can be used for finding functional effects of SNPs. In the output a text file is generated enlisting descriptions of identified SNPs in tab-delimited format. The file has major columns as name of the SNP, ID of the gene, starting location for translation and alleles. Chromosomal position is used to indicate the gene as cytoband units or absolute chromosome position. There is a provision of uploading the list of genes or can be pasted in the box.SNPs having adverse functional effects are identified. Information about functions of various genes can be found from OMIM and Gene Ontology. The resources also contain information about homologous genes. The best way is to describe SNPs by indicating their relative location with respect to gene and not absolute coordinates. SNPs occurring in various regions of genes are found such as promoter region, intronic region, exonic region and coding region. Identified SNPs and genes are linked to Ensembl Genome Browser. It is highly important to validate the found SNPs. Authors found at the time of this study, that only 2,359,534 out of 5,798,183 SNPs, i.e., 40% of SNPs were validated in dbSNP build 118. Out of these SNPs, population frequencies could be found only for 160,466 and 94,867 SNPs are associated with diseases. For finding allele frequencies, a study was conducted by taking 48 persons from the Spanish population. The fragments of interest were amplified by polymerase chain reaction (PCR) using specific primers designed with OLIGO 4.1 program. Initially, the SNP screening was done using a denaturing high-performance liquid chromatograph (dHPLC) system (WAVE, Transgenomics Limited, Crewe, UK). Data handling and optimisation of dHPLC was done using Navigation software. Sequence analysis of every PCR product showing change of chromatogram profile was done next. E.Z.N.A. Cycle-Pure Kit (Omega Bio-tek, USA) was used for purification of PCR products and an automatic sequencer ABI PRISMTM 3700 (Applied Biosystems. PerkinElmer, USA) was used for its sequencing. Fourml of a Big Dyeterminator cycle sequencing Kit (Perkin Elmer, USA), 10 pmolof the sense/antisense primer, 5% DMSO and 6 to 12 ng of amplified DNA were used for carrying out the reaction. Out of a total of 28 SNPs which were identified , there were 24 SNPs which were authenticated and were proved to be polymorphic in Spanish population chosen for conducting study (Conde et al., 2004).

Sachidanandam et al. (2001) developed a map with 1.42 million SNPs distributed across entire human genome. This map integrates SNPs made available by a number of projects viz. The SNP Consortium (TSC), HGP, White head Institute, Sanger Centre and Washington University. Detection of SNPs was done using two algorithms viz. Polybayes and Neighbourhood Quality Standard. This map acted as a good resource for understanding haplotype variation across the genome and to identify genes significant in diagnosis and therapy. This map contains SNPs available up to November, 2000. With 95% contribution from The SNP Consortium (TSC) and the public HGP, 1,023,950 candidate SNPs (http:// snp.cshl.org) came from TSC found using shotgun sequencing of genomic fragments. Other institutions which found SNPs are Whitehead Institute, Sanger Centre and Washington University. The method adopted involved following steps: firstly, alignment of reads to each other and to the genome was done. Next, single-base mutations were identified using Polybayes and the neighbourhood quality standard

(NQS). Then identification of Candidate SNPs was done. A number of method were employed to identify SNPs in EST overlaps. Found SNPs area available at their individual dbSNP entries (http://www.ncbi.nlm.nih.gov/SNP/) (Sachidanandam et al., 2001).

Prunier et al. (2019) came up with a tool named LD annot to find annotations of polymorphisms based on phenomenon of LD. Authors define that a polymorphism may not be directly responsible for some phenotype but there can be some other polymorphism in neighbouring area with short physical distance which might be causative of some disease. LD-annot calculates an average distance (in bp) between two SNPs based on specified r2 threshold in entire dataset. This distance is considered on both sides of a SNP in order to consider SNPs in its LD. LD-annot tool is available at https://github.com/ArnandDroitLab/LD-annot. There have been numerous research endeavours based upon variant identification approaches which aim to identify genomic basis for the variations. The tool developed in this study finds LD and finds annotation from all those regions which have genetic association with the candidate polymorphisms. They can then be prioritised and analysed. Research is done to establish association of genetic markers with quantitative traits. One approach is to do testing and genotyping of individuals selected from the population. Then they are studied to find candidate polymorphisms in controlled and uniform conditions using GWAS. Second approach is to genotype offsprings of a controlled cross between two individuals. The two individuals vary substantially for the given trait. This is again done in controlled and uniform conditions. These approaches pave way for SNP identification. The developed annotation tool was tested on different datasets of that of a domesticated animal, a domesticated plant and a wild insect. The sampling size, number of tested SNPs and candidate SNPs were also different for these datasets. This enabled performance evaluation of the tool (Prunier et al., 2019).

SNiPlay is a web-based application offering simplicity and robustness for extraction and analysis of polymorphisms in genomic data. This pipeline generates nicely formatted output which can be further analysed by other tools. There is a pipeline of programs and a relational database in SNiPlay. This database is based on MySQL. SNPs are identified from alignments using a home-made module. The web interface is coded in Perl CGI scripts on an Apache web server. JavaScript and Ajax technologies enable the interaction. Polymor-find program combines Phred/Phrap/Consed software suite with Polyscan program (Dereeper et al., 2011). SniPlay is available at http://sniplay.cirad.fr/ (Dereeper et al., 2011).

Hassan et al. (2016) have carried out a study of BRAF gene of RAF family located on chromosome 7 (7q34) consisting of 18 exons and translated into protein named 'B-RAF prot-oncogene serine/ threonine protein kinase'. In this study, SNPs&Indels present in coding&non-coding regions of BRAF gene are identified. Various state-of-the-art bioinformatics tools are used to identify SNPs and indels. A total of 111 SNPs identified from coding regions are found to be high damage causing and six SNPs as less damage causing. Different tools used for extracting the SNPs from SNP databases and predicting their effects are SIFT, Polyphen-2, SNAP2 server, I-Mutant suite, CPH Models, UCSF Chimera Model Software, Automatic Protein Structural Analysis and Information Using HOPE Server, PolymiRTS Database (3'UTR), Effect of SNPs within 5' UTR on Transcription Factor Binding Sites and Effect of 3'/5'Splice Sites SNPs/Indels (HSF Tool). SNPs are extracted from dbSNP, UniProt, HapMap, 1000 Genomes Project, gene bank, and ClinVar (Hassan et al., 2016).

Pers et al. (2015) have come up with a web based SNP identification and annotation tool. The principle behind this approach is to identify the SNPs matching with the associated query SNPs. First a set of SNPs is taken randomly. Then a set of query SNPs is taken as input. These SNPs are the ones associated with some disease. Now, various parameters are calculated for both sets of SNPs for matching the extracted SNPs with query SNPs. First, Minor allele frequency is calculated and SNPs are classified into minorallele frequency bins (1–2, 2–3, …, 49–50% strata). Second, LD buddies are found for every SNP, these are the SNPs in LD at different thresholds. Third, distance to nearest gene is calculated. If the SNP lies within a gene then distance to the starting site of gene is calculated. Fourth, Gene density, number of genes near the location of SNP is calculated using concept of LD and physical distance like 100 kb, 200 kb, …, 1,000 kb and so on for defining physical distance of location/loci. This approach has been used by a number of researchers in their study (Gamazon et al., 2010, 2013; Allen et al., 2010; Nicolae et al., 2010; Maurano et al., 2012; Schaub et al., 2012; Wood et al., 2014). Novelty of this study is development of a tool based on this approach. SNPsnap server is available athttp://www.broadinstitute.org/mpg/snpsnap/.

Arshad et al. (2018) conducted a study on SNPs of TAGAP gene from the world's most extensive, dbSNP database. 1721 SNPs were retrieved out of which 275 were nsSNPs, 147 lied in 5`UTR, 162 lied in 3`UTR region and remaining ones were of some other category. The nsSNPs were selected for this study. There are fivein silico tools namelySIFT, PROVEAN, PolyPhen-2, PhD-SNP and SNPs&GO which are used for predicting effects of these nsSNPs on structure or function of TAGAP protein. Results produced by SIFT indicates that nsSNPs which scored tolerance index (TI) of 0.05 lie in the 'intolerant' category. In the tool named PROVEAN, a variant is marked as 'disease-causing' if final score is less that the threshold value of –2.5 and variant is marked as 'neutral' if the final score is above this threshold value. The tool PolyPhen-2 categorises the nsSNPs into three categories viz. probably damaging, possibly damaging and benign nsSNPs. Here marking of a nsSNP as 'probably damaging' is the most confident one. A total of 95 nsSNPs were found to be diseased by the tool PhD-SNP whereas only 18 nsSNPs were found to be diseased by the tool SNPs&GO. Those nsSNPs which were selected by at least four of these tools were termed as 'high-risk' nsSNPs (Arshad et al., 2018).

Manaster et al. (2005) have developed a specialised software called InSNP for automated detection of SNPs and Indels. First of all, InSNP aligns the given sequence against the reference sequence using a simple word match where a window of 20 bases beginning from each primer site is used. Next, base-calling from the sequencing files is done. In order to determine mutation frequency, the allele whose corresponding peak has the greatest area under the curve is termed as primary allele. The allele is termed as the other allele, if the area under its peak is more than 30% of the total area under all peaks at that location else that base is termed as homozygous. InSNP automatically identifies possible SNPs by identifying positions in the sequence that are different from the reference sequence. Further verification of the SNPs is done by human experts by visually inspecting them. This is beginning of good sequence. Results of InSNP were compared with those of PolyPhred and Mutation Surveyor (Manaster et al., 2005).

4 Conclusions and future scope

There are a number of approaches and tools available for SNP identification and analysis. The working principles of various techniques are discussed in Section 3. There are certain research gaps identified during the course of study based on the extensive literature survey:

- Majority of the tools require huge memory space and longer time for generating output. Hence, these tools need a specific configuration of the system and cannot be used on a simple personal computer.
- Most of the tools which detect the SNPs satisfactorily work well for small sequences and their performance deteriorates as the sequences become longer.
- Most of these tools are pipelines consisting of similar kind of popular tools applied in sequence. Here, error generated at one stage propagates to the next stage.
- There is no standard format for output from various tools. Hence, there is absence of one benchmark or a set of parameters which can be used to evaluate and compare performance of state-of-the-art techniques.
- Several methods exist for identifying SNPs though there is no global approach to identify all types of SNPs. Most of the methods focus at coding region SNPs.

There have been various endeavours and projects in the area of genome sequencing such as HGP, GWAS, and TCGA, whose outcomes reveal that genetic sequence variations have strong association with diseases. SNPs are found to be the most common genetic sequence variation which constitute 90% of sequence variations and are found to be responsible for about half of the known human inherited diseases. Therefore, there is great significance and scope of this study as summarised below:

- Identification and analysis of SNPs is highly important for early prevention, diagnosis and treatment of diseases.
- SNPs act as excellent markers for locating candidate genes associated with a disease.
- Identification of SNPs can enable individualised drug treatment as they are responsible for particular response of a person towards a drug.
- Identification and analysis of SNPs plays vital role in studying genetic diversity, evolutionary study of a species and differences of traits among individuals.
- An automated tool can identify SNP in shorter time and with lesser memory requirement for the task.

Disclaimer

This study is not part of any funded project.

References

'AlleleID: design qPCR and microarray assays for related organisms' [online] http://www.premierbiosoft.com.

Ahmad, T., Valentovic, M.A. and Rankin, G.O. (2018) 'Effects of cytochrome P450 single nucleotide polymorphisms on methadone metabolism and pharmacodynamics', *Biochemical Pharmacology*, Vol. 153, pp.196–204.

Allen, H.L. et al. (2010) 'Hundreds of variants clustered in genomic loci and biological LD-annot: a bioinformatics tool to automatically provide candidate SNPs with annotations for genetically linked genes', *Frontiers in Genetics*, Vol. 10, No. 7317, p.1192.

Altshuler, D., Daly, M.J. and Lander, E.S. (2008) 'Genetic mapping in human disease', *Science*, Vol. 322, No. 5903, pp.881–888.

Arshad, M., Bhatti, A. and John, P. (2018) 'Identification and in silico analysis of functional SNPs of human TAGAP protein: a comprehensive study', *PLoS ONE*, Vol. 13, No. 1, p.e0188143.

Bianco, A.M., Marcuzzi, A., Zanin, Z., Girardelli, M., Vuch, J. and Crovella, S. (2013) 'Database tools in genetic diseases research', *Elsevier Genomics*, Vol. 101, No. 2, pp.75–85.

Borsting, C. and Morling, N. (2013) 'Single-nucleotide polymorphisms', in *Encyclopedia of Forensic Sciences*, 2nd ed., Elsevier Ltd., University of Copenhagen, Denmark.

Chakravarti, A. (1999) 'Population Genetics-making sense out of sequence', *Nature Genetics*, Supplement, Vol. 21, Supplement 1, pp.56–60.

Chasman, D. and Adams, R.M. (2001) 'Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation', *Journal of Molecular Biology*, Vol. 307, No. 2, pp.683–706.

Chen, K., McLellan, M.D., Ding, L., Wendl, M.C., Kasai, Y., Wilson, R.K. and Mardis, E.R. (2007) 'PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data', *Genome Research*, Vol. 17, No. 5, pp.659–666.

Collins, F.S. et al. (1997) 'Variations on a theme: cataloging human DNA sequence variation', *Science*, Vol. 278, No. 5343, p.1580.

Conde, L., Vaquerizas, J.M., Santoyo, J., Al-Shahrour, F., Ruiz-Llorente, S., Robledo, M. and Dopazo, J. (2004) 'PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level', *Nucleic Acids Research*, Vol. 32: pp.W242–W248.

Dereeper, A., Nicolas. S., Cunff, L.L., Bacilieri, R., Doligez, A., Peros, J.P., Ruiz, M. and This, P. (2011) 'SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grape vine diversity projects', *BMC Bioinformatics*, Vol. 12, p.134.

Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998) 'Base-calling of automated sequencer traces using Phred I. Accuracy assessment', *Genome Research*, Vol. 8, No. 3, pp.175–185.

Gamazon, E.R. et al. (2013) 'Enrichment of CIS-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants', *Molecular Psychiatry*, Vol. 18, No. 3, pp.340–346.

Gamazon, E.R., Huang, R.S., Cox, N.J. and Dolan, M.E. (2010) 'Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci', *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 107, pp.9287–9292.

Gardner, S.N. and Hall, B.G. (2013) 'When whole-genome alignments just won't work: kSNPv2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes', *PLOS ONE*, Vol. 8, No. 12, p.e81760.

Gardner, S.N., Slezak, T. and Hall, B.G. (2015) 'kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome', *Bioinformatics Advance Access*, Vol. 31, No. 17, pp.2877–2878.

Gray, I.C., Campbell, D.A. and Spurr, N.K. (2000) 'Single nucleotide polymorphisms as tools in human genetics', *Human Molecular Genetics*, Vol. 9, No. 16, pp.2403–2408.

Haga, H., Yamada, R., Nakamura, Y. and Tanaka, T. (2002) 'Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190 562 genetic variations in the human genome', *Journal of Human Genetics*, Vol. 47, No. 11, pp.605–610.

Hassan, M.M., Omer, S.E., Khalf-allah, R.M., Mustafa, R.Y., Ali, I.S. and Mohamed, S.B. (2016) 'Bioinformatics approach for prediction of functional coding/noncoding simple polymorphisms (SNPs/Indels) in human BRAF gene', *Advances in Bioinformatics 2016*, p.2632917.

Johnson, A.D. (2009) 'SNP bioinformatics: a comprehensive review of resources', *Circulation: Cardiovascular Genetics*, Vol. 2, No. 5, pp.530–536.

Kashuk, C., Gupta, S., Eichler, E. and Chakravarti, A. (2001) 'viewGene: a graphical tool for polymorphism visualization and characterization', *Genome Research*, Vol. 12, No. 2, pp.333–338.

Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T. and Birney, E. (2004) 'EnsMart: a generic system for fast and flexible access to biological data', *Genome Research*, Vol. 14, No. 1, pp.160–169.

Kim, K.J., Lee, H.J., Park, M.H., Cha, S.H., Kim, K.S., Kim, H.T., Kimm, K., Oh, B. and Lee, J.Y. (2006) 'SNP identification, linkage disequilibrium, and haplotype analysis for a 200-kb genomic region in a Korean population', *Genomics*, Vol. 88, No. 5, pp.535–540.

Leekitcharoenphon, P. et al. (2012) 'snpTree – a web-server to identify and construct SNP trees from whole genome sequence data', *BMC Genomics*, Vol. 13, Suppl. 7, p.S6.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, Vol. 25, No. 14, pp.1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) '1000 genome project data processing subgroup: the sequence alignment/map (SAM) format and SAM tools', *Bioinformatics*, Vol. 25, No. 16, pp.2078–2079.

Manaster, C., Zheng, W., Teuber, M., Wachter, S., Doring, F., Schreiber, S. and Hampe, J. (2005) 'InSNP: a tool for automated detection and visualization of SNPs and InDels', *Human Mutation*, Vol. 26, No. 1, pp.11–19.

Maurano, M.T. et al. (2012) 'Systematic localization of common disease-associated variation in regulatory DNA', *Science*, Vol. 337, No. 6099, pp.1190–1195.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P. and Cunningham, F. (2016) 'The Ensembl variant effect predictor', *Genome Biology*, Vol. 17, p.122.

Mooney, S. (2005) 'Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis', *Briefings in Bioinformatics*, Vol. 6, No. 1, pp.44–56.

Mooney, S.D., Krishnan, V.G. and Evani, U.S. (2010) 'Bioinformatic tools for identifying disease gene and SNP candidates', *Methods Molecular Biology*, Vol. 628, pp.307–319.

National Human Genome Research Institute (NHGRI) (2015) *An Overview of the Human Genome Project* [online] http://www.genome.gov/12011238 (accessed 10 February 2021).

Ng, P.C. and Henikoff, S. (2006) 'Predicting the effects of amino acid substitutions on protein function', *Annual Review of Genomics and human Genetics*, Vol. 7, pp.61–80.

Ngamphiw, C., Kulawonganunchai, S., Assawamakin, A., Jenwitheesuk, E. and Tongsima, S. (2008) 'VarDetect: a nucleotide sequence variation exploratory tool', *BMC Bioinformatics*, Vol. 9, No. 12, p.S9.

Nickerson, D.A., Tobe, V.O. and Taylor, S.L. (1997) 'PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescense-based resequencing', *Nucleic Acids Research*, Vol. 25, No. 14, pp.2745–2751.

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E. and Cox, N.J. (2010) 'Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS', *PLoS Genetics*, Vol. 6, No. 4, p.e1000888.

Niu, T. and Hu, Z. (2004) 'SNPicker: a graphical tool for primer picking in designing mutagenic endonuclease restriction assays', *Bioinformatics*, Vol. 20, No. 17, pp.3263–3265.

Oeveren, J.V. and Janssen, A. (2009) 'Mining SNPs from DNA sequence data; computational approaches to SNP discovery and analysis', in Komar, A.A. (Ed.): *Single Nucleotide Polymorphisms: Methods in Molecular Biology*, p.578, Humana Press.

Pers, T.H., Timshel, P. and Hirschhorn, J.N. (2015) 'SNPsnap: a web-based tool for identification and annotation of matched SNPs', *Bioinformatics*, Vol. 31, No. 3, pp.418–442.

Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A. and Boyce-Jacino, M. (1999) 'Mining SNPs from EST databases', *Genome Research*, Vol. 9, No. 1, pp.167–174.

Prunier, J., Lemaçon, A., Bastien, A., Jafarikia, M., Porth, I., Robert, C. and Droit, A. (2019) 'LD-annot: a bioinformatics tool to automatically provide candidate SNPs with annotations for genetically linked genes', *Frontiers in Genetics*, Vol. 10, p.1192.

Pusch, W., Kraeuter, K.O., Froehlich, T., Stalgies, Y. and Kostrzewa, M. (2001)' Genotools SNP MANAGER: a new software for automated high-throughput MALDI-TOF Mass spectrometry SNP genotyping', *BioTechniques*, Vol. 30, pp.210–215.

Riva, A. and Kohane, I.S. (2002) 'SNPper: retrieval and analysis of human SNPs', *Bioinformatics*, Vol. 18, No. 12, pp.1681–1685.

Robert, F. and Pelletier, J. (2018) 'Exploring the impact of single-nucleotide polymorphismson translation', *Frontiers in Genetics*, Vol. 9, p.507.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., Hunt, S.E., Cole, C.G., Coggill, P.C., Rice, C.M., Ning, Z., Rogers, J., Bentley, D.R., Kwok, P.Y., Mardis, E.R., Yeh, R.T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R.H., McPherson, J.D., Gilman, B., Schaffner, S., Van Etten, W.J., Reich, D., Higgins, J., Daly, M.J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M.C., Linton, L., Lander, E.S., Altshuler, D. and International SNP Map Working Group (2001) 'A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms', *Nature*, Vol. 409, No. 6822, pp.928–933.

Sahl, J.W., Beckstrom-Sternberg, S.M., Hepp, C.M., Auerbach, R.K., Tembe, W., Wagner, D.M., Keim, P.S. and Pearson, T. (2015) 'The in silico genotyper (ISG): an open-source pipeline to rapidly identify and annotate nucleotide variants for comparative genomics applications', *Biorxiv: The Preprint Server for Biology*, http://dx.doi.org/10.1101/015578.

Saunders, C.T. and Baker, D. (2002) 'Evaluation of structural and evolutionary contributions to deleterious mutation prediction', *Journal of Molecular Biology*, Vol. 322, pp.891–901.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. and Snyder, M. (2012) 'Linking disease associations with regulatory information in the human genome', *Genome Research*, Vol. 22, No. 9, pp.1748–1759.

Sohi, N. and Singh, A. (2018) 'Single nucleotide polymorphisms: identification and association with breast cancer using biocomputing approach', Presented at the *IEEE Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 14–15 June.

Srinivasan, S., Clements, J.A. and Batra, J. (2016) 'Single nucleotide polymorphisms in clinics: fantasy or reality for cancer?', *Critical Reviews in Clinical Laboratory Sciences*, Vol. 53, No. 56, pp.29–39.

Sripichai, O. and Fucharoen, S. (2007) 'Genetic polymorphisms and implications for human diseases', *Journal of the Medical Association of Thailand*, Vol. 90, No. 2, pp.394–398.

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., Kondrashov, A.S. and Bork, P. (2001) 'Prediction of deleterious human alleles', *Human Molecular Genetics*, Vol. 10, No. 6, pp.591–597.

Tang, J., Vosman, B., Voorrips, R.E., van der Linden, C.G. and Leunissen, J.A. (2006) 'Quality SNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species', *BMC Bioinformatics*, Vol. 7, p.438.

Vallejos-Vidal, E., Reyes-Cerpa, S., Rivas-Pardo, J.A., Maisey, K., Yanez, J.M., Valenzuela, H., Cea, P.A., Castro-Fernandez, V., Tort, L., Sandino, A.M., Imarai, M. and Reyes-Lopez, F.E. (2020) 'Single-nucleotide polymorphisms (SNP) mining and their effect on the tridimensional protein structure prediction in a set of immunity-related expressed sequence tags (EST) in Atlantic Salmon (Salmo salar)', *Frontiers in Genetics*, Vol. 10, p.1406.

Wang, D.G., Fan, J., Siao, C., Berno, A., Young, P., Sapolsky, R. et al. (1998) 'Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome', *Science*, Vol. 280, No. 5366, pp.1077–1082.

Wang, P. et al, (2005) 'SNP function portal: a web database for exploring the function implication of SNP alleles', *Bioinformatics*, Vol. 22, No. 14, pp.e523–e529.

Wang, Z. and Moult, J. (2001) 'SNPs, protein structure and disease', *Human Mutation*, Vol. 17, 263270.

Weckx, S., De Rijk, P., Van Broeckhoven, C. and Del-Favero, J. (2005) 'SNPbox: a modular software package for large-scale primer design', *Bioinformatics*, Vol. 21, No. 3, pp.385–387.

Wegrzyn, J.L., Lee, J.M., Liechty, J. and Neale, D.B. (2009) 'PineSAP-sequence alignment and SNP identification pipeline', *Bioinformatics*, Vol. 25, No. 19, pp.2609–2610.

Wijmenga, C. and Zhernakova, A. (2018) 'The importance of cohort studies in the post-GWAS era', *Nature Genetics*, Vol. 50, No. 3, pp.322–328.

Wood, A. et al. (2014) 'Defining the role of common variation in the genomic and biological architecture of adult human height', *Nature Genetics*, Vol. 46, No. 11, pp.1173–1186.

Yue, P., Melamud, E. and Moult, J. (2006) 'SNPs3D: candidate gene and SNP selection for association studies', *BMC Bioinformatics*, Vol. 7, p.166.

Zhang, J., Wheeler, D.A., Yakub, I., Wei, S., Sood, R., Rowe, W., Liu, P.P., Gibbs, R.A. and Buetow, K.H. (2005) 'SNPdetector: a software tool for sensitive and accurate SNP detection', *PLoS Computational Biology*, Vol. 1, No. 5, pp.0395–0404.