European Journal of Bioinformatics Vol. 2, 2015, 1-8

Whites Science
A Scientific Erudition

COMPUTATIONAL TOOLS TO DETECT SINGLE NUCLEOTIDE POLYMORPHISM (SNP) IN NUCLEOTIDE SEQUENCES: A REVIEW

Raghunath Satpathy*, Rashmiranjan Behera
Department of **Biotechnology, MIRC Lab, MITS Engineering College, Rayagada, Odisha,India**

ABSTRACT

Single nucleotide polymorphisms (SNPs) are basically single base pair alterations present in the genomic DNA. SNPs is usually treated as one of the most common genetic markers in case of plants, animals as well as the human genome to study the complex genetic traits and evolutionary status of the genome. SNPs are widely used as popular markers due to their continuous presence in the genome, highly reproducible, relatively easy to score. In addition to this, SNPs in coding sequences are used to directly examine the genetics of expressing genes and to study various polymorphic functional traits. Specifically the non-synonymous SNPs are more attractive because they alter the amino acid that ultimately affecting the protein functions. The direct application of SNP exists with pharmacogenomics study and crop improvement. Various strategies have been used for SNP discovery that comes from both observational and computational techniques. SNPs can be detected by laboratory based experimental methods, which are time consuming and expensive also the development costs are high. The implementations of Bioinformatics approach reduce the development cost of SNPs as it uses publicly available sequences from databases like expressed sequence tags (ESTS) that cause the development of SNP markers rapid and less expensive.

Keywords Single nucleotide polymorphism, computational methods, genome evolution, genetic traits, sequence analysis

INTRODUCTION

In nature any two human genomes show about 0.1% variation (99.9% identical) in their genomic features that makes different from one to the other. One of the important types of genomic variation that falls within this 0.1% is single nucleotide polymorphisms (SNP), represents a single base change between two individuals at a particular position [1]. Based on variation patterns, SNPs comprises changes like transitions (purine to purine change), transversions (purine to pyrimidine change or vice versa) and tiny insertions/deletions (indels) , which are most frequently observed as biallelic in nature[2]. The majority of these SNPs found in individuals only a fraction of the substitutions having its functional significance and it is the basis for the diversity [3]. The cytogenetic basis of uses of the SNPs is, they can be used as genetic markers, hence regarded as a powerful tool in the field of genetic factors associated with diseases [4]. Also the Single nucleotide polymorphism (SNP) detection broadly applied to mine the new polymorphisms in a given nucleotide sequence. However the experimental basis of SNP detection is performed with both difficulty and expensive way [5]. Comparative genomics study based SNP discovery indicates that, on average, the presence of one SNP in every 1,000 bp of DNA sequences. The traditional method which is adapted to mine the SNPs randomly in the human genome by observing the alterations in restriction sites in the genome by identification of restriction fragment length polymorphisms (RFLPs) [6]. The method include isolation of the high quality genomic DNA from multiple individuals, followed by digestion with a number of restriction enzymes, separated by gel electrophoresis and transferring it to the nylon filters (Southern Blotting). These Southern blots were then probed with random genomic clones to identify variations in the restriction fragment lengths. Because for SNP mining very small amounts of DNA are to be detected, the

Corresponding author:

Email: rnsatpathy@gmail.com

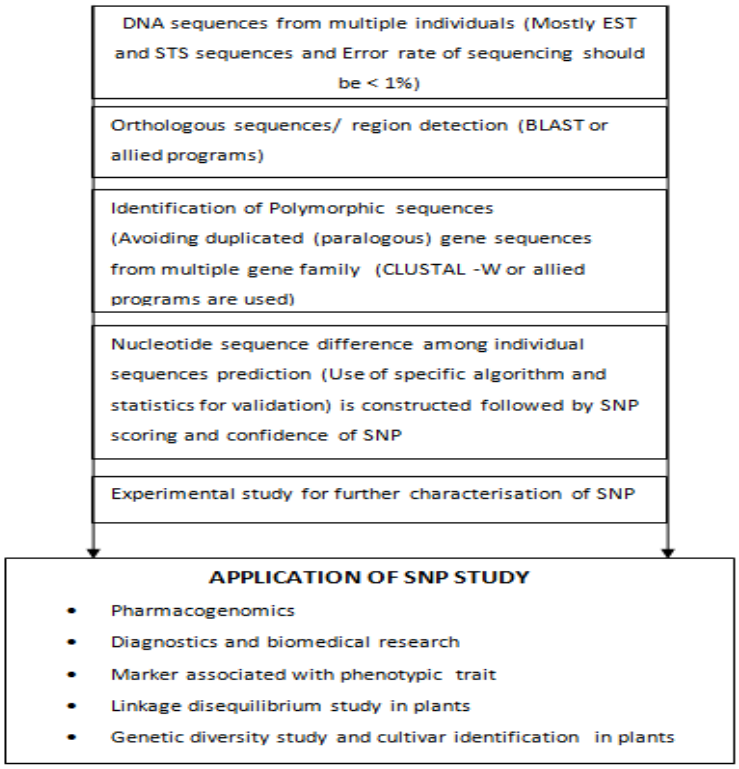
radioactive labels are applied with the probes for this purpose [7-8]. One of the other challenges in the global SNP discovery is to find certain density of genetic marker in the whole genome from the regions of interest. There are also several SNP detection methods are followed at present and mostly these method find out the difference between the mismatched heteroduplex DNA from the perfectly matched homoduplex DNA. Many other techniques including PCR are used to detect the polymorphisms, subsequently used for SNP discovery. The DNA sequencing methods hold good for SNP discovery [9]. The DNA sequencing techniques for SNP discovery is the direct sequencing of PCR products and there is a clear recognition that sequencing data from a heterozygote and homozygote [10]. But in practical basis of DNA sequencing methods are quite costly and requires expertise in the field. Also to probe DNA fragments for SNPs by this method, highly successful polymorphism scanning methods are required [11].

In this paper, discussion about the importance and major challenges in computational basis SNP discovery has been made along with a brief sketch of various existing computational methods implemented in different commonly used Bioinformatics tools and their implementation.

COMPUTATIONAL BASIS OF SNP DISCOVERY PROTOCOLS

The frequency calculation of SNP indicates that, the SNPs do not occur randomly. Whenever SNPs are generated by, the mutation in an ancestor in thousand years ago, that SNP will be inherited by a lot of different people, but not by all. Mining the SNP from the whole genome is a difficult task. A major challenge in computational SNP discovery is distinguishing allelic variation from the sequence variation between non-homologous sequences along with the method for recognizing the sequencing errors. For the majority of the public EST sequences, the unavailability of quality files makes difficult for detection of reliable SNPs as it has to rely on improved sequence comparisons only. In silico methods for SNP discovery are now being adopted, providing cheap and efficient methods for easy mining. Large quantities of sequence data have been generated internationally through Expressed Sequence Tag (EST) or genome sequencing projects and these provide a valuable resource for the mining of molecular markers. The detailed overview of computational methods for SNP prediction is

Fig. 1. Schema for computational SNP prediction in gene sequences and its applications in various fields of study



described in figure 1.

The computational methods of SNP discovery have been proved to be the cheapest and efficient method that most often take help of various databases, on/off line tools and genome browsers [12]. Since huge quantities of sequence data have been generated globally by genome sequencing

1. PolyBayes

PolyBayes uses a Bayesian-statistical model to check out the differences within assembled sequences, further suppoted with the depth of coverage values, the base quality values and the expected rate of polymorphic sites in the region. Base quality values can be obtained by running the

Table 1. Some representative databases that contain SNP data in different group of organism

Serial number	Name	Availability	Remark
1	HGBASE(Human GenicBi-Allelic SEquences))	http://hgbase.cgr.ki.se	Human
2	Db SNP	(http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi	Human
3	SNP DATA BASE	http://genome.wustl.edu/genome_data/c._elegans_single_nucleotide_polymorphism_data	*C.elegans*
4	FESD II	http://sysbio.kribb.re.kr:8080/fesd/	Human
5	snpdb/ Haplotype Polymorphism	http://probes.pw.usda.gov:8080/snpworld/Search	wheat
7	Japanese SNP Database	http://snp.ims.u-tokyo.ac.jp/	Human
8	SeattleSNPs	http://pga.gs.washington.edu/	Human
9	IBISS (Interactive Bovine In Silico SNP (IBISS) database)	http://www.livestockgenomics.csiro.au/ibiss/	Based on bovine EST sequences
10	Mouse SNP Query Form	http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=snpQF	Mouse
11	Hypertension Candidate Gene SNPS	http://cmbi.bjmu.edu.cn/genome/candidates/snps.html	Human

projects and these data provide a valuable resource for the mining and discovery of important molecular markers like SNP [13]. Sequence data from the experimental genome projects are being generated largely due to the implementation of next generation sequencing (NGS) technologies. These powerful technologies have the ability for rapid sequencing representative samples of genomes even also the whole eukaryotic genomes.

While dealing with large amount of sequence data it is expected that there must be a clear differentiation between true SNPs and sequence error [14]. In one of the first examples of this application, a total of 36,000 maize SNPs were identified in data from a single run of the Roche 454 GS20 DNA sequencer [15]. Various common tools and data bases available for mining SNP is described as below:

sequence trace files through the PHRED base-calling program. [16-17], and repeats can be removed from sequences using Repeat Masker [18]. The output can be viewed through the Consed alignment viewer. Recent studies using PolyBayes include SNP discovery for and bird species [19]. Fully probabilistic SNP detection algorithm that calculates the probability (SNP score) that discrepancies at a given location of a multiple alignment represent true sequence variations as opposed to sequencing error.

2. PolyPhred

PolyPhred compares a sequence trace files from different people to spot the heterozygous sites [20]. The sequence trace files are used to find SNPs, so can identify positions in the sequence based on the peaks occurs. The quality of an SNP is assigned based on the spacing between peaks; the relative size of called and uncalled peaks; and the dip between

peaks. PolyPhred only analyses nucleotides that have a minimum quality as determined by Phred [21]

3. Polyfreq

PolyFreq is a tool that allows an efficient mining for SNPs from the sequence alignments [22-23]. It contains basically five programs that contain its own set of parameters. From the alignment, highly similar regions are found and screened for the candidate SNPs. Finally the SNP discovery is based on a minimum a priori polymorphic rate, a minimum depth and a minimum percent of identity having the

Table 2. Resources of the SNP related tool related to the disease in human

Name of the database /tools	Type of data	Availability	Application
DACS -DB	Disease associated with cytokine SNP data base	http://www.iupui.edu/~cytosnp/	(SNPs) associated with cytokine genes can act as markers for identification of diseases, traits and phenotypes
VnD	Disease related SNP and drug	http://vnd.kobic.re.kr:8080/VnD/	It is crucial to understand the trilateral relationship between genomic variations, diseases and drugs
miRdSNP	A database of disease-associated SNPs and MicroRNA target sites on 3'UTRs of human genes	http://mirdsnp.ccr.buffalo.edu	Identifying the SNPs effect on destroying, or modify the efficiency of miRNA binding to the 3'UTR of a gene, resulting in gene dysregulation. Causes diseases
SNPranker 2.0	Interpretation of SNPs associated with disease	http://www.itb.cnr.it/snpranker	The capability of correlating specific genotypes with human diseases ultimately correlated to the pathology
F-SNP	The functional effects of SNPs obtained from 16 Bioinformatics tools and databases.	http://compbio.cs.queensu.ca/F-SNP/	Effect on human disease at the splicing, transcriptional, translational, and post-translational level.
SNPnexus	Multiple mutation	http://snp-nexus.org/about.html	Selection of functionally relevant Single Nucleotide Polymorphisms (SNP) for large-scale genotyping studies of multifactorial disorders
SNPs&GO	Information derived from protein sequence, 3D structure and function	http://snps.biofold.org/snps-and-go/pages/method.html	Predict whether a given variation can be classified disease-related or neutral.
SNPeffect	A database for phenotyping human single nucleotide polymorphisms (SNPs).	http://snpeffect.switchlab.org/about#Welcome_to_SNPeffect_4.0	Primarily focuses on the molecular characterization and annotation of disease and polymorphism variants in the human proteome.
(PhD-SNP)	Disease related neutral polymorphisms	http://snps.biofold.org/phd-snp/phd-snp.html	It predicts the pathological effect based on the local sequence environment of the mutation.
Screening for Non Acceptable Polymorphisms (SNAP)	Position in a protein as gain or loss in protein function.	https://www.rostlab.org/services/SNAP/	Based on neural network and advanced machine-learning approach to predict the functional change of SNPs on the protein's function.

default values are 0.001, 100 and 0.97 respectively. Aligned nucleotides must have a quality value equal or greater than the default value (the default value is 20). Mismatches parts are then considered as the candidate SNPs if the quality of the five base pairs flanking them has a good score value.

4. SNPDetector

SNPDetector uses thread to call bases and determine quality scores from trace files, and then aligns reads to a reference sequence using a Smith-Waterman algorithm [24]. SNPs are identified where there is a sequence difference and the flanking sequence is of high quality. SNPDetec-tor has been used to find SNPs in 454 data and has been included within a comprehensive SNP discovery pipeline.

5. NovoSNP

NovoSNP requires both trace files and a reference sequence as input. The trace files are base-called using Phred and quality clipped, then aligned to a reference sequence using BLAST [25]. An SNP confidence score is calculated for each predicted SNP. NovoSNP is written in Tcl with a graphical user interface written in Tk and runs on Linux and Windows. NovoSNP has been used in a study of genotype-phenotype correlation of human disease [26].

6. **AutoSNP**

Redundancy is the principle means of differentiating between sequence errors and real SNPs. While this approach ignores potential SNPs that are poorly represented in the sequence data and that can be used directly from GenBank [27]. AutoSNP is therefore applicable to any species for which the nucleotide sequence data are available in the data bank. A co-segregation score is calculated based on whether multiple SNPs, that basically defines a haplotype, which is used as a second, independent measure of confidence. AutoSNP is written in Perl and is run from the Linux command line with a FASTA file of sequences as input. The output is presented as linked HTML to the index page presenting a summary of the results. AutoSNP has been applied to several species including maize and fishes [28-29].

7. SNPServer

SNPServer is a real time implementation of the autoSNP method, accessed via a web server. A single FASTA sequence is pasted into the interface and similar sequences are retrieved from a nucleotide sequence database using BLAST [30]. The results are presented as HTML. Alternatively, a list of FASTA sequences may be input for assembly or apreassembled ACE format file may be analyzed. SNPServer has been used in studies including sea anemones and human [31-32].

8. InSNP

InSNP is a specialized package for the detection of targeted mutation detection. InSNP provides a user friendly interface with better functionality for mutation detection than commonly available sequence handling software. It provides similar SNP detection sensitivity and specificity as the public domain and commercial alternatives in the investigated dataset. InSNP lowers the barriers to the use of automated mutation detection software and aids in the improvement of the efficiency of such experiments. The software installer can be obtained from **www.mucosa.de/insnp** [33].

APPLICATION OF IN SILICO BASED SNP PREDICTION METHODS IN DISEASE IDENTIFICATION

The DNA sequences variation in case of humans is directly related to the effect in terms of critical human genetic diseases, hence the SNP knowledge are also used in the personalized medicine area for clinical treatment purposes [34]. However, their greatest importance in biomedical research is for comparing specific regions of the genomes. Not only in case of human, also the study of SNP data are successfully applied in in crop and livestock breeding programs.SNPs are usually biallelic in nature and thus can be easily assayed [35]. A single SNP may cause a complex genetic disease. The fact for genetic diseases creation by SNPs is observed as , it does not usually function individually; rather it work co-ordinately with other SNPs to manifest a disease condition as has been seen in Osteoporosis disease [36]. SNPs have been also used in genome-wide association studies (GWAS), e.g. as high-resolution markers in gene mapping related to diseases or normal traits. The mining of SNPs will also help to understand the effects like pharmacokinetics (PK) or pharmacodynamics, i.e. how drugs act on individuals with different genetic variants. Other example of wide range human disease includes Sickle–cell anemia, β Thalassemia and Cystic fibrosis, which are the result from SNP study [37-39].

CONCLUDING REMARK

SNP detection is very much important as it is having the clinical significance. But the experimental technologies that have evolved to trace SNP from nucleotide sequences are multistep process labor-intensive and require technical expertise. In this context the Bioinformatics is widely used to tackle the problem. These methods are used for sample tracking, sequence comparisons, and data analysis. The DNA sequence data can be stored, analyzed by using different methods ultimately the results obtained can be further processed and reported automatically by sophisticated computers at the end. From many years the bioinformatics has proved to be an integrated part of both genetics and genomics for annotation purpose of the genome sequences. Due to the exponential growth in size of the databases, successful implementation of Bioinformatics resources will result new opportunities for data annotation like SNP mining in whole genome.

REFERENCES

[1] Cooper DN, Smith BA, Cooke HJ, Niemann S, Schmidtke J. An estimate of unique DNA sequence heterozygosity in the human genome. Hum Genet 1985; 69: 201-205.

[2] Doveri S, Lee D, Maheswaran M, Powell W. Molecular markers: History, features and applications. In Principles and Practices of Plant Genomics. Enfield, USA, Science Publishers 2008, pp 23-68.

[3] Collins FS, Guyer MS, Charkravarti A. Variations on a theme: cataloging human DNA sequence variation. Science 1997; 278: 1580-1581.

[4] Johnson GC, Todd JA. Strategies in complex disease mapping. Curr. Opin Genet Dev 2000; 10: 330-334.

[5] Kwok PY, Chen X. Detection of single nucleotide polymorphisms. Current issues in molecular biology 2003; 5: 43-60.

[6] Haliassos A, Chomel JC, Tesson L, Baudis M, Kruh J, Kaplan JC, Kitzis A. Modification of enzymatically amplified DNA for the detection of point mutations. Nucleic Acids Res 1989; 17:3606.

[7] Pandey S, Ranjan R, Pandey S, Mishra RM, Seth T, Saxena R. Effect of ANXA2 gene single nucleotide polymorphism (SNP) on the development of osteonecrosis in Indian sickle cell patient: a PCR-RFLP approach. Indian J Exp Biol. 2012 ; 50: 455-458.

[8] Bortolotti D, Gentili V, Melchiorri L, Rotola A, Rizzo R. An accurate and reliable real time SNP genotyping assay for the HLA-G +3142 bp C>G polymorphism. Tissue Antigens. 2012 ; 80: 259-262.

[9] Alderborn A, Kristofferson A, Hammerling U. Determination of Single-Nucleotide Polymorphisms by Real-time Pyrophosphate DNA Sequencing. Genome Res 2000; 10: 1249-1258.

[10] Li H, Ruan J, Durbin J. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008; 18: 1851-1858.

[11] Reed GH, Wittwer CT. Sensitivity and Specificity of Single-Nucleotide Polymorphism Scanning by High-Resolution Melting Analysis. Clinical Chemistry 2004; 10: 1748-1754.

[12] Brodzik AK, Francoeur J. A new approach to in silico SNP detection and some new SNPs in the Bacillus anthracis genome. BMC Res Notes. 2011; 4: 114.

[13] van Oeveren J, Janssen A. Mining SNPs from DNA sequence data; computational approaches to SNP discovery and analysis. Methods Mol Biol. 2009; 578: 73-91.

[14] Hayes BJ, Nilsen K, Berg PR, Grindflek E, Lien S. SNP detection exploiting multiple sources of redundancy in large EST collections improves validation rates. Bioinformatics 2007; 23: 1692-1693.

[15] Duran C, Appleby N, Edwards D, Batley B. Molecular Genetic Markers: Discovery, Applications, Data Storage and Visualisation. Current Bioinformatics 2009; 4: 16-27.

[16] Marth GT, Korf I, Yandell MD. A general approach to single-nucleotide polymorphism discovery. Nat Genet 1999; 23: 452-456.

[17] Ewing B, Hillier L, Wendl MC, Green P. Base-calling of auto-mated sequencer traces using phred. I. Accuracy assessment. Genome Res 1998; 8: 175-185.

[18] Mallon AM, Strivens M. DNA sequence analysis and comparative sequencing. Methods 1998; 14: 160-178.

[19] Sironi L, Lazzari B, Ramelli P, Gorni C, Mariani P. Single nucleo-tide polymorphism discovery in the avian Tapasin gene. Poult Sci 2006; 85: 606-612.

[20] Nickerson DA, Tobe VO, Taylor SL. PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res 1997; 25: 2745-2751.

[21] Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 1998; 8: 186-194.

[22] Wang J, Huang X. A method for finding single-nucleotide polymorphisms with allele frequencies in sequences of deep coverage. BMC Bioinformatics 2005; 6.

[23] Souche E L , Hellemans B, Van Houdt JKJ, Canario A, Klages S, Reinhardt R, Volckaert FAM. Mining for Single Nucleotide Polymorphisms in Expressed Sequence Tags. J Integr Bio inform. 2007; 4:73.

[24] Zhang J, Wheeler DA, Yakub I, Wei S, Sood R, Rowe W, Liu PP, Gibbs RA, Buetow KH. SNPdetector: a software tool for sensitive and accurate SNP detection. PLoS Comput Biol. 2005; 1: 53.

[25] Ewing B, Hillier LD, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 1998; 8: 175-185.

[26] Weck S, Del-Favero J, Rademakers R, Claes L, Cruts M, De Jonghe P, Broeckhoven CV, Rijk PD. NovoSNP, a novel computational tool for sequence variation discovery. Genome Res. 2005; 15: 436-42.

[27] Barker G, Batley J, O'Sullivan H, Edwards KJ, Edwards D. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics 2003; 19: 421–422.

[28] Barker G, Batley J, O'Sullivan H, Edwards KJ, Edwards D. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant physiology 2003; 132: 84-91.

[29] Vasemägi A, Gross R, Palm D, Paaver T, Primmer CR. Discovery and application of insertion-deletion (INDEL) polymorphisms for QTL mapping of early life-history traits in Atlantic salmon. BMC genomics 2010; 11: 156.

[30] Savage D, Batley J, Erwin T, Logan E., Love CG, Lim GA, Mongin E, Barker G, Spangenberg GC, Edwards D. SNPServer: a real-time SNP discovery tool. Nucleic acids research 2005; 33: 493-495.

[31] Sullivan JC, Reitzel, AM, Finnerty JR. Upgrades to StellaBase facilitate medical and genetic studies on the starlet sea anemone, Nematostella vectensis. Nucleic acids research 2008; 36: 607-611.

[32] Batley J, Edwards D. Mining for SNPs and SSRs using SNPServer, dbSNP and SSR taxonomy tree. Bioinformatics for DNA Sequence Analysis. Humana Press, 2009; 303-321.

[33] Manaster C, Zheng W, Teuber M, Wächter S, Döring F, Schreiber S, Hampe J. InSNP: a tool for automated detection and visualization of SNPs and InDels. Hum Mutat. 2005; 26: 11-19.

[34] Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB.Bioinformatics challenges for personalized medicine.Bioinformatics. 2011 ;27:1741-1748.

[35] Sachidanandam R, Weissman D, Schmidt SC, Kakol J M, Stein LD, Marth G, Sherry S,Mullikin JC, Altshuler D. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001; 409: 928–933.

[36] Singh M, Singh P, Juneja PK, Singh S, Kaur T. SNP–SNP interactions within APOE gene influence plasma lipids in postmenopausal

osteoporosis. Rheumatology International 2010; 31: 421–423.

[37] Bakir-Gungor B, Sezerman OU. A New Methodology to Associate SNPs with Human Diseases According to Their Pathway Related Context. PLoS ONE 2011; 6: e26277. doi:10.1371/journal.pone.0026277.

[38] Lai E. Application of SNP technologies in medicine: lessons learned and future challenges. Genome Res. 2001; 11: 927-9.

[39] Conde L, Bracci PM, Richardson R, Montgomery SB, Skibola CF. Integrating GWAS and expression data for functional characterization of disease-associated SNPs: an application to follicular lymphoma. Am J Hum Genet. 2013 10; 92: 126-30.