Genome-wide SNP discovery in associating with human diseases phenotypes

Dr. P. H. Chaminda Kumanayake BDS, MSc
Dental Officer (Health Informatics), Ministry of Health, Colombo, Sri Lanka
E-Mail address: ckumanayake@yahoo.com

Sri Lanka Journal of Bio-Medical Informatics 2012;**3**(1):25-31
DOI: http://dx.doi.org/10.4038/sljbmi.v3i1.2451

Abstract

Single Nucleotide Polymorphisms or SNPs are a most abundant, stable and simple base pair changes that occur in the genome. It is an important variation that can be used to describe many unsolved problems in modern medicine such as individual variation to disease response, differences in response to treatment, allergies to drug treatment, etc. Monogenic Mendalian diseases are very rear and most of the time the disease has complex multi-genetic involvement. With the advancement of sequencing technologies SNP discovery is becoming fast, accurate and less expensive. As a result the availability of SNP data has become more abundant and is used to create SNPmap and SNPprofile. This SNP map and SNP profile helps to locate the genes that involve some complex diseases like diabetes, vascular diseases, and mental disorders and to describe individual variation in response to treatment as well as finding a drug target in pharmacogenomics. With such developments in Bio-informatics, the dream of "individualized treatment" is becoming a reality.

Keywords - Monogenic Mendalian; SNPprofile; SNPmap

Background

With the advancement of new DNA sequencing technologies, 99% of the human genome is sequenced and now interest is turning to the evaluation of variation which can be seen in the human genome[(1)]. These variations can be single base-pair changes (SNP), insertion and deletion (Indel) or copy number variation (CNV). Variations of single nucleotides (A,T,C or G) at specific positions in the genome sequence that differs between members of the species or paired chromosomes in an individual are called Single Nucleotide Polymorphisms, or SNPs. These alterations of single nucleotides are considered as SNPs when they occur in at least 1% of the population. SNPs, which make up about 90% of all human genetic variation occurs every 100 to 300 bases along the three billion base pairs of human genome. SNPs are also evolutionarily stable, making them easier to follow in population studies[(1)].

Although more than 99.9% of genomes of different individuals are identical, variations in these genomes are responsible for all the differences in a human. Variations in DNA sequence have a major impact on how humans respond to diseases; environmental factors such as bacteria, viruses, toxins and chemicals. Further, SNPs are directly associated with human response to drugs and other therapies. This makes SNPs a valuable tool for various biomedical research such as developing pharmaceutical products and medical diagnostics.

Due to the significance of SNPs, SNP discovery is one of the major fields in genomics where scientists have worked and created a SNP map of human genome. The Human Genome Project[(2)] The SNP Consortium or TCS project[(3)], International HapMap project[(4)] and The Human Variome project[(5)] are worth studying. Scientists believe SNP maps will help them identify the multiple genes associated with complex condition such as cancer, diabetes, vascular disease, and some forms of mental illness[(6)]. These associations are difficult to establish with conventional gene hunting methods because a single altered gene may make

only a small contribution to the disease. In addition to pharmacogenomic, diagnostic and biomedical research implications, SNP maps help identify thousands of additional markers in the genome[7]. These markers simplify navigation of the much larger genome map generated by HGP researchers.

Medical relevance of SNPs in the human genome

Identification of SNPs that contribute to susceptibility to common diseases will provide highly accurate diagnostic information that will facilitate early diagnosis, prevention, and treatment of human diseases. The biological and medical considerations are important, if not critical, in identifying high-risk individuals that are very likely to bear genetic variants which will predispose them to a particular disease, whereas, the low risk individuals may bear allelse that may protect them from disease. Therefore, it is important that such individuals with "extreme phenotypes" should be sequenced preferentially to identify phenotypically or medically relevant allelic variants. Therefore, various risks and phenotypes may have to be taken into account when modeling the SNP analysis. It should be certain that such biologically and medically informed approach would reduce the number of individuals needed to be sequenced based on their risks or protection, leading to an increased chance of identifying important variants.

Researchers have found that most SNPs are not responsible for a disease state because they are intergenic SNPs. Instead, they serve as biological markers for pinpointing a disease on the human genome map, as they are located near a gene found to be associated with a certain disease[7].

It is a fact that, single gene disorders inherited according to Mendalian law are rare and most common diseases like diabetes are caused by multiple genes. Finding all of these genes is difficult. Recently, there has been focus on the idea that all of the genes involved can be traced by using SNPs[7]. By comparing the SNP patterns in affected and non-affected individuals, patients with diabetes and healthy controls can catalog the specific DNA variation that underlie susceptibility for diabetes.

Advancement of new sequencing technologies enables to sequence genomes of different individuals at low cost and these sequences are known as Personal Genomes. Today number of personal genomes are available and some of them are:

- The complete genome of an individual Human[8]
- The Sequence of the Human Genome[9]
- The Diploid Genome Sequence of an Individual Human[10]
- The First Korean Genome[11]
- The Korean Individual Genome (AKI)[12]
- Chinese Genome[13]
- African Genome[14]

Analysis of association between SNPs and diseases

Genetic polymorphisms contribute to variations in phenotypes, risk to certain diseases and response to drugs and the environment. Genome wide linkage analysis and positional cloning have been tremendously successful for mapping human disease genes that underlie

monogenic Mendelian diseases[15]. But most common diseases (such as diabetes, cardiovascular diseases, and cancer) and clinically important quantitative traits have complex genetic architectures; a combination of multiple genes and interactions with environmental factors is believed to determine these phenotypes.

Genome wide association studies offer a promising approach for mapping associated loci. The completion of the human genome sequence[17] enabled the identification of millions of Single Nucleotide Polymorphisms (SNPs)[18] and the construction of a high density haplotype map[19]. These advances have set the stage for large scale genome wide SNP surveys for seeking genetic variations associated with or causative of a wide variety of human diseases.

SNP detection methodologies

Next generation massively parallel sequencing technologies provide ultra high thorughput at two orders of magnitude lower unit cost than capillary Sanger sequencing technology. One of the key applications of next generation sequencing is studying genetic variation between individuals using whole genome or target region re-sequencing[20].

For more than two decades, Sanger sequencing and fluorescence based electrophoresis technologies have dominated the DNA sequencing field. DNA sequencing is the method of choice for novel SNP detection, using either a random shotgun strategy or PCR amplification of regions of interest. Most of the SNPs deposited in dbSNP were identified by these methods[21]. A key advantage of the utility of traditional Sanger sequencing is the availability of the universal standard of phred scores[22] for defining SNP detection accuracy in which the phred programme assigns a score to each base of the raw sequence to estimate an error probability.

With high thoroughput clone sequencing of shotgun libraries, a standard method for SNP detection[23] is to align the reads onto a reference genome and filter low quality mismatches according to their phred score known as the "neighborhood quality standard" (NQS)[24]. With direct sequencing of PCR amplified sequences from diploid samples, software, including SNPdetector[25], novoSNP[26], PolyPhred[27], and PolyScan[28] have been developed to examine chromatogram files to detect heterozygous polymorphisms.

New DNA sequencing technologies which have recently been developed and implemented, such as the Illumina Genome Analyser (GA), Roche/454 FLX system, and AB SOLiD system, have significantly improved and dramatically reduced the cost as compared to capillary based electrophoresis systems[29]. In a single experiment using one Illumina GA, the sequence of approximately 100 million reads of up to 50 bases in length can be determined. This ultrahigh thoroughput makes next generation sequencing technologies particularly suitable for carrying out genetic variation studies by using large scale re-sequencing of siseable cohorts of individuals with a known reference[30]. Three humans have been sequenced using these technologies: James Watson's genome by 454 Life Sciences (Roche) FLX sequencing technology[31], an Asian genome[32], and an African genome[33] sequenced by Illumina GA technology. Additionally, given such sequencing advances, an international research consortium has formed to sequence the genomes of at least 1000 individuals to create the most detailed human genetic variation map to date.

SNP detection methods for standard sequencing technologies are well developed. However, given distinct differences in the raw output data format from different next generation

sequencing, novel methods for accurate SNP detection are essential. To meet these needs, the massively parallel Illumina GA technology, a method of consensus calling and SNP detection have developed. The Illumina platform uses a 'phred' like quality score system to measure the accuracy of each sequenced base pair. Using this, we calculated the likelihood of each genotype at each site based on the alignment of short reads to a reference genome together with the corresponding sequencing quality scores. The statistical analysis is carried out for genotype with the highest posterior probability at each site using a Bayesian method. The Bayesian method has been used for SNP calling for traditional Sanger sequencing technology[34] and has also been introduced for the analysis of next generation sequencing data[35]. In the method presented here, we have taken into account the intrinsic bias or errors that are common in Illumina GA sequencing data and recalibrated the quality values for use in inferring consensus sequence.

They evaluated this SNP detection method using the Asian genome sequence, which has 36× high quality data[32]. The evaluation demonstrated that their method has a very low false call rate at any sequencing depth, and excellent genome coverage in high depth data, making it very useful for SNP detection in Illumina GA re-sequencing data at any sequencing depth. This methodology and the developed software described in this report have been integrated into the Short Oligonucleotide Alignment Programme (SOAP) package[35] and named "SOAPsnp" to indicate its functionality for SNP detection using SOAP short read alignment results as input.

SOAPsnp: A de novo short reads assembler

SOAPsnp is a member of the SOAP (Short Oligonucleotide Analysis Package)[36]. The programme is a re-sequencing utility that can assemble consensus sequence for the genome of a newly sequenced individual based on the alignment of the raw sequencing reads on the known reference. The SNPs can then be identified on the consensus sequence through the comparison with the reference. In the first Asian genome re-sequencing project, evalution of SOAPsnp result on Illumina HapMap 1M BeadChip Duo genotyping sites shows great accuracy. Over 99% of the genotyping sites are covered at over 99.9% consistency. Further, PCR plus Sanger sequencing of the inconsistent SNP sites confirmed majority of the SOAPsnp results.

Other methods

Polybayes Software

PolyBayes is a computer programme for the automated analysis of single nucleotide polymorphism (SNP) discovery in redundant DNA sequences. The primary motivation for its development is to provide a general and reliable tool for the discovery of genetic variation and automated analysis of SNP discovery in redundant DNA sequences [37].

ssahaSNP

This is a software package which can detect homozygous SNPs and indels on a eukaryotic genome scale from millions of shotgun reads. Matching seeds of a few kmer words are found to locate the position of the read on the genome. Full sequence alignment is performed to detect base variations. Quality values of both variation bases and neighbouring bases are checked to exclude possible sequence base errors. To analyse polymorphism level in the

genome, they used the package to detect indels from 20 million WGS reads against the draft WGS assembly. From the dataset, a total of 663,660 indels have been detected, giving an estimated average indel density at about one indel every 2.48 Kb. Distribution of indels length and variation of indel mapped times were also analysed [(38)].

PbShort

PbShort is the working name for a new short read SNP and short INDEL discovery programme, a re-incarnation of the PolyBayes SNP discovery tool developed by Gabor Marth at Washington University. This version is specifically optimised for the analysis of large numbers (millions) of high throughput next generation sequencer reads, aligned to whole chromosomes of model organism or mammalian genomes.

The software takes advantage of "random" access to DNA read data and assembly information stored in binary files. Computer programmes for converting text format sequence files and assembly formats are packaged together with the SNP caller software[(39)].

However, in the process of finding association between SNPs and human disease the exom – sequencing has become more prominent [(39)]. Exome sequencing is more affordable than Whole Genome Sequencing (WGS) as well as several other advantages such as that by exome - sequencing can detect additional small variants missed by WGS.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Human Genome. Nature 2001; **409**(6822): 860-921.

2. Gudmundur A, Thorisson, Stein LD. The SNP Consortium website: past, present and future. Nucleic Acids Research 2003; **31**(1).

3. The International HapMap Project. Consortium, The International HapMap. Nature 2003; **426**(6968): 789-96.

4 Cotton RGH, Auerbach AD, Axton M, Barash CI, Berkovic SF. The Human Variome Project. Science 2008; **322**(5903): 861-2.
Doi: 10.1126/science.1167363.

5. Altshuler D, Pollara VJ, Cowles CR. An SNPmap of the Human Genome. Nature 2000; **407**.

6. Voronko OE, Bodoev NV, Archakov AI. The use of SNP markers for estimation of individual genetic predisposition to diabetes mellitus type 1 and 2. **2**(2)

7. SNP Fact Sheet. www.ornl.gov/hgmis. [Online] U.S. Department of Energy Office of Science, September 2008. Retrieved in November 2010.
Available from: http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml

8. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L *et al*. The complete genome of an individual by massively parallel DNA sequencing Nature 2008; **452(**7189): 872-6.
Doi: 10.1038/nature06884.

9. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, *et al.* The Sequence of the Human Genome. Science 2001; **291**(5507): 1304-51.

10. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, *et al.* The Diploid Genome Sequence of an Individual Human. PLoS Biology 2007; **5**(10) 5(10):e254.

11. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, *et al.* The first Korean genome sequence and analysis:Full genome sequencing for a socio-ethnic group. Genome Research 2009; **19**(9): 1622-9.
Doi: 10.1101/gr.092197.109

12. Kim JI, Ju YS, Park H, Kim S, Lee S, *et al.* A highly annotated whole-genome sequence of a Korean individual. Nature 2009; **460**(7258): 1011-5
Doi:10.1038/nature08211.

13. Wang J, Ruiqiang WW, Li R, Li Y, Tian G, et al. The diploid genome sequence of an Asian individual. Nature 2008; **456**(7218):60-5
Doi:10.1038/nature07484

14. Schuster SC, Miller W, Ratan1 A, Tomsho LP, Giardine B, et al Complete Khoisan and Bantu genomes from Southern Africa. Nature 2010; **463**(7283): 943-7.
Doi: 10.1038/nature08795.

15 Jimenez-Sanchez J, Childs B, Valle D. Human disease genes. Nature 2001; **409.** (6822):853-5.

16. Wang WYS, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. Nature Reviews 2005; **6**.

17. Lander ES, Linton LM, Birren B Nusbaum C, Zody MC, *et al.* Initial sequencing and analysis of the human genome. Nature 2001; **409** (6822): 860-921.

18. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001; **409**(6822): 928-33.

19. A haplotype map of the human genome. Consortium, International HapMap. Nature 2005; **437**(7063): 1299-320.

20. Li R, Li Y, Kristiansen K. SOAP: short oligonucleotide alignment program. Bio-informatics 2008; **24**(5).

21. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, *et al.* dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 2001; **29**(1): 308-11..

22. Ewing B, Green P. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. Genome Research 1998; 8(3):186-94.

23. Ning Z, Cox AJ, Mullikin JC. SSAHA: A fast search method for large DNA databases. Genome Research 2000; (10):1725-9.
Doi: 10.1101/gr.194201 .

24. Altshuler D, Pollara VJ, Cowles CR. Van Etten WJ, Baldwin J, *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 2000; **407**(6803): 513-6.

25. Zhang J, Wheeler DA, Yakub I, Wei S, Sood R, *et al.* SNPdetector: A Software Tool for Sensitive and Accurate SNP Detection. PLoS Computational Biology 2005; **1**(5).

26. Weckx S, Del-Favero J, Rademakers R, Claes L, Cruts M, et al. novoSNP, a novel computational tool for sequence variation discovery. Genome Research, 2005; **15**(3): 436-42.
Doi: 10.1101/gr.2754005

27. Stephens M, Sloan JS, Robertson PD, Scheet P, Nickerson DA. Automating sequence-based detection and genotyping of SNPs from diploid samples. Nature Genetics 2006; **38**(3): 375-81.
Doi: http://dx.doi.org/10.1038/ng1746

28. Chen K, McLellan MD, Ding L, Wendl MC, Kasai Y, *et al.* PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. Genome Research 2007; **17**(5): 659-66.
Doi:10.1101/gr.6151507

29. Shendure J, Mitra RD, Varma C. Advanced sequencing technologies: methods and goals. Nature Reviews Genetics 2000; **5**.
Doi:10.1038/nrg1325

30. Bentley DR.Whole-genome re-sequencing. Current Opinion in Genetics & Development. 2006; **16**(6): 545-52.

31. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 2008; **452**(7189): 872-6.
Doi: 10.1038/nature06884.

32. Wang J, Wang W, Li R, Li Y, Tian G, *et al.* The diploid genome sequence of an Asian individual. Nature 2008; **456**(7218):60-5.
Doi: 10.1038/nature07484.

33. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008; **456**(7218): 53-9.
Doi: 10.1038/nature07517.

34. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, *et al.* A general approach to single-nucleotide polymorphism discovery. Nature Genetics 1999; **23**(4):452-6.
Doi:10.1038/70570

35. Li R, Li Y, Fang X, Yang H, Wang J, *et al.* SNP detection for massively parallel whole-genome resequencing. Genome Research 2009; 19(6):1124-32. Doi: 10.1101/gr.088013.108.

36. SOAPsnp. SOAP-Short Oligonucleotide Analysis Package. Beijing Genomics Institute, 2007. Retrieved in December 2010. Available from: http://soap.genomics.org.cn/soapsnp.html

37. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, *et al.* A general approach to single-nucleotide polymorphism discovery. Nature Genetics 1999; **23**(4):452-6.

38. ssahaSNP - A Polymorphism Detection Tool on a Whole Genome Scale (2008). Scientific Commentary. Available from: http://en.scientificcommons.org/42384856

39. The MarthLab: PbShort. Boston College. Retrieved in December 2010. Available from: http://bioinformatics.bc.edu/marthlab/PbShort