The Sri Lankan Personal Genome Project: an overview

Prof. Vajira H. W. Dissanayake MBBS, PhD
Professor, Department of Anatomy; Medical Geneticist, Human Genetics Unit, Faculty of Medicine, University of Colombo, Sri Lanka
E-Mail address: vajirahwd@hotmail.com

Pubudu S. Samarakoon BSc, MSc
Tutor, Biomedical Informatics Course, Postgraduate Institute of Medicine, University of Colombo, Sri Lanka.
Current address: Research Fellow, Department of Medical Genetics, University of Oslo, Norway.
E-Mail address: saneth.samarakoon@gmail.com

Dr. Vinod Scaria MBBS
Scientist, CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Delhi, India
E-Mail address: vinods@igib.res.in

Ashok Patowary BSc, MSc
Scientist, CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Delhi, India
E-Mail address: s.sivasubbu@igib.res.in

Dr. Sridhar Sivasubbu PhD
Scientist, CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Delhi, India
E-Mail address: s.sivasubbu@igib.res.in

Dr. Rajesh S. Gokhale PhD
Director, CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Delhi, India
E-mail address: director@igib.res.in

Sri Lanka Journal of Bio-Medical Informatics 2011;2(1):4-8
DOI: http://dx.doi.org/10.4038/sljbmi.v2i1.3711

Abstract

The first Sri Lankan Personal Genome was sequenced heralding the entry of Sri Lanka into the new era of whole genome sequencing. This paper explains the background and the rationale for the project, gives a brief overview of what was found in the Sri Lankan Personal Genome, and discusses the future directions of the project.

Keywords - Sri Lankan Personal Genome Project; Sri Lanka; Genome; Sinhalese; Whole Genome Sequencing

Background

With the completion of the Human Genome Project (HGP)[1,2] and a decade of human genomics, we are at an interesting juncture in the history of mankind. New technologies have enabled sequencing of complete human genomes at a fraction of the original cost of what was spent on the Human Genome Project (HGP). At the same time, these technologies have significantly improved the scale and the ease of sequencing, and as a result it is now possible to sequence entire human genomes and understand the genomic make-up of the individual, with widespread potential applications in healthcare[3].

Major projects worldwide which followed the Human Genome Project, including the HapMap project[4] and the 1000 genomes project[5], have been cataloguing human genetic variations at rapid speed. In addition, they have fuelled the growth of technologies and analytical methods to scan entire genomes for informative genetic markers, aimed at

understanding the differences and similarities between individuals. This is the first step towards understanding genotype-phenotype correlations. Such large studies have been undertaken by multiple groups from around the world[6]. This has resulted in the identification of a large number markers associated with complex disorders and drug-response[7]. This is just the tip of the iceberg, and many new associations continue to be reported in scientific literature on a daily basis.

In addition to understanding genetic variations and how they contribute to disease, there has been a large body of work aimed at understanding other important aspects such as epigenetic mechanisms and genomic regulation. This was made possible by new genomic tools which enabled researchers to address questions at a genomic level and developments in bioinformatics, made possible by the availability of cheaper and faster computers which made it possible to do large-scale data analysis, and the development of robust algorithms to mine data and model biological phenomenon on a genomic scale.

The Sri Lankan personal genome

Today any country aspiring to provide its people access to state of the art healthcare cannot ignore the rapid advances in the field of human genomics. It is imperative at this juncture for every country to acquire the much-required tools and know-how. In addition, it is also necessary to create the baseline data for understanding the genetic diversity of its population.

The Sri Lankan Personal Genome Project is the first step in this direction in Sri Lanka, and marks the entry of Sri Lanka to the exciting field of whole genome sequencing. Sri Lanka is home to over 20 million people with rich racial, cultural and linguistic diversity[8]. The earliest evidence (34,000 BP) of anatomically modern man in South Asia is found in Sri Lanka[9]. The rich diversity of human populations in the island has been influenced by migration from the mainland India. Sri Lanka also has a rich heritage in organised medical care. The hospital at Mihintale (437 – 367 BC) is the most ancient hospital to be discovered in the World[10]. The Sri Lankan population consists presently of six major populations, the Sinhalese, Sri Lankan Tamils, Indian Tamils, Moors, Burghers, and Malays[8]. It is also home to other smaller diverse populations like Vaddhas, the descendents of the original inhabitants of the island who were geographically isolated from other populations, and Kaffirs, descendents of African slaves brought to the island over 500 years ago.

To understand the genetic diversity of the Sri Lankan populations, and to create the baseline data for disease association studies, we had earlier created the Sri Lankan Genome Variation Database[11]. This database contains information on Single Nucleotide Polymorphisms (SNPs) found in Sinhalese, Sri Lankan Tamils and Moors, the three major ethnic groups in the Sri Lankan population. The database presently contains information including genotype frequencies of 34 genomic variations encompassing 14 medically important genes. The database has been designed keeping in mind international standards for describing and annotating variations, including those of the Human Genome Variation Society (HGVS)[12]. In the true spirit of collaborations and open access to data, the database also accepts submissions from the research community and thus offers a standard access point to the spectrum of genetic variations in the population to researchers and clinicians. The resource is accessible at URL: http://www.hgucolombo.net/slgv/home.htm

As a proof of concept towards the goal of interpreting and analysing complete genome data, we sequenced a complete genome of an anonymous Sinhalese male of Sri Lankan origin with both upcountry and low country descent. Sequencing was performed using next-generation

sequencing technology, with over 20x coverage of the genome. Analysis of the genome resulted in the identification of 2,811,918 SNPs, of which 222,739 were novel in comparison to dbSNP[13] build 131. This accounted for almost 7.9% of entire set of variations in the genome, pointing to the necessity of having more complete genomes to have a more comprehensive picture of the spectrum of genetic variations in humans. Analysis also resulted in the identification of 489,921 insertion-deletion (InDel) events in the genome.

Future directions

The immediate strategic goals of the Sri Lankan Personal Genome Project for 2011-2012 are to understand in depth, the genetic variations and their potential phenotypic consequences. Thus, for the years 2011-2012 we articulate our research in terms of three main streams – annotating the genetic variations unique to Sri Lankans, studying the interactions between genes in relation to disease phenotypes, visualising the annotation of the Sri Lankan genome via "Sri Lankan Genome Browser" and the "Sri Lankan Genome Variation Database".

The first Sri Lankan Personal Genome has revealed over 2.8 million single nucleotide variations of which over 200,000 are unique variations which have hitherto not been identified in other populations as revealed by comparison with variations collected in the dbSNP database build 131. We hope that in depth annotation of these variations will provide crucial insights into some phenotypes which could be specific for the Sri Lankan population. Coupling this knowledge with associated clinical phenotypes and traits will potentially enable scientists to generate new hypothesis on the given association. Consequently, these hypothesis can be validated by specifically genotyping these unique variations in the Sri Lankan population. Recent advances in the field of bioinformatics and data mining offers the tools required for the annotation and functional interpretation of SNP data. [For example see the article by Harendra *et al*. in this issue of the Journal[14]]. The value of this information can be further enhanced by comparative studies with data coming from other projects including the 1000 Genomes Project and other population specific personal genome projects.

Recent advances in genomic technologies have enabled researchers to unravel many a biological pathway and process at molecular detail. It would be imperative to exploit this data and perform integrative analysis so as to understand the biological context and functional consequence of genomic variations. This would include (i) understanding biological interaction networks including genetic interaction networks and protein interaction networks from public databases like OMIM and HuGe Navigator, (ii) curation and integration of the interaction network so as to understand molecular processes of diseases and drug metabolic pathways, including integrating association data from public repositories and resources to understand the molecular pathways of biological processes, (iii) integration of the variation data with the gene interaction network to understand the potential consequences of the genomic variations which could be modelled and validated in model systems.

To ensure the wide use and ease of interpretation, we have made available the genomic variations and annotations of the Sri Lankan Personal Genome on the Sri Lankan Genome Browser, an online genome browser built on the Generic Model Organisms (GMOD) Gbrowse[15]. This would serve as the central hub for exchange of data, visualisation of genomic variations and their annotations including data that would come out of the Sri Lankan Personal Genome Project in the future (Figure 1). The resource is freely accessible online at www.srilankangenome.net.

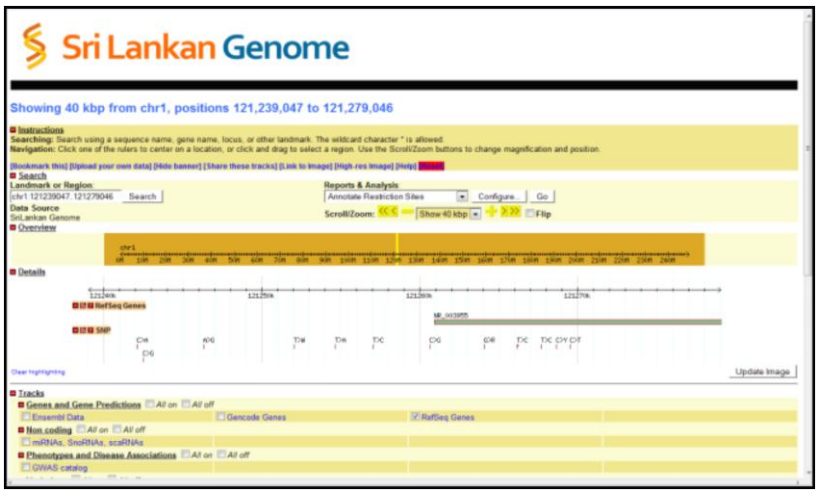

Figure 1. The Sri Lankan Genome Browser provides quick visualization of genomic variations and would ease the annotation and interpretation of genetic variations.

In the future, we hope to unravel the genetic diversity of Sri Lankan populations by sequencing more individuals from different racial groups. We also hope to collaborate both nationally and internationally to assimilate knowledge and expertise and possibly co-create resources which would enable the interpretation of data and its application in healthcare. This includes participation in co-creating open resources like OpenPGx (www.openpgx.org) for interpreting genomic variations and participation in collaborative initiatives aimed at understanding the diversity of Asian populations. We also recognise that application of genomics in healthcare would not be possible without educating and involving medical professionals in genomics research and that this would include educating medical professionals on analysing and interpreting genomic information and using such information in their clinical practice.

Declaration of conflicts of interest

The authors declare that they have no conflicts of interest.

Acknowledgement

The Sri Lankan Personal Genome Project was funded by the NOMA Project of the Postgraduate Institute of Medicine, University of Colombo, Sri Lanka.

References

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409: 860–921. doi:10.1038/35057062

2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. The sequence of the human genome. Science 291: 1304–1351. doi:10.1126/science.1058040

3. Green ED, Guyer MS; Charting a course for genomic medicine from base pairs to bedside. Nature. 2011; 470: 204-13. doi:10.1038/nature09764

4. The International HapMap Consortium. The International HapMap Project. Nature 2003; 426:789-796.

5. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature 2010; 467(7319):1061-73

6. http://www.gwascentral.org/index. Visited on 1/3/2011.

7. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106: 9362-7. doi: 10.1073/pnas.0903103106

8. http://www.statistics.gov.lk/PopHouSat/PDF/p7%20population%20and%20Housing%20Text-11-12-06.pdf. Visited on 1/3/2011.

9. S.U. Deraniyagala. Pre and Proto-historic Settlement in Sri Lanka. Proceedings of the XIIIth International Congress of the International Union of Prehistoric and Protohistoric Sciences. 1998:277-285.

10. Müller-Dietz HE, Die Krankenhaus-ruinen in Mihintale (Ceylon). Historia Hospitalium 1975:10;6-71.

11. Samarakoon PS, Jayasekara RW, Dissanayake VHW. The Sri Lankan Genome Variation Database. Sri Lanka Journal of Biomedical Informatics 2011;2(1):10-21. DOI: http://dx.doi.org/10.4038/sljbmi.v2i1.2861

12. Scriver CR, Nowacki PM, Lehvaslaiho H. Guidelines and recommendations for content, structure, and deployment of mutation databases. Hum Mutat 1999; 13:344-50. PMID: 10612816

13. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29:308-11.

14. Harendra GG, Jayasekara RW, Dissanayake VHW. In silico analysis of Single Nucleotide Polymorphisms (SNPs) in the Heparin-Binding EGF-like Growth Factor (HBEGF) gene and their allelic profiles in the Sri Lankan population: a comprehensive approach to prioritise SNPs for candidate gene studies. Sri Lanka Journal of Bio-Medical Informatics 2011;2(1):22-38. DOI: http://dx.doi.org/10.4038/sljbmi.v2i1.2926

15. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. The generic genome browser: a building block for a model organism system database. Genome Res. 2002; 12:1599-610. doi: 10.1101/gr.403602