Genome-wide haplotype analysis

Dr. Vajira Samanthi Weerasekara MBBS, MSc
Medical Officer (Health Informatics), Ministry of Health, Colombo, Sri Lanka
E-Mail address: vsamanthi@yahoo.com

Sri Lanka Journal of Bio-Medical Informatics 2012;**3**(1):20-24
DOI: http://dx.doi.org/10.4038/sljbmi.v3i1.2564

Abstract

This review was done with the objective of reviewing genome-wide haplotype analyses which have been already carried out globally. After the completion of the human genome project, several studies on haplotype analysis on various aspects have been published. Such knowledge may provide valuable information on human evolutionary history and lead to indentifying genetic variants that are responsible for occurrence of various diseases. In addition to the review on International HapMap project, this review emphasises on the current understanding of the haplotype structure, diversity and distribution of haplotypes, as well as current understanding on linkage disequilibrium with its relationship to the haplotypes. It further reviews the tools available for haplotype analysis with their statistical applicability.

Keywords - Genome-wide Haplotype Analysis; HapMap Project

Introduction

The human genome consists of 23 pairs of chromosomes derived from each parent. Each chromosome is a single piece of coiled DNA containing many genes. DNA from each parent is a mosaic of the preceding ancestor. The genomic sequences of any two individuals are 99.9 percent identical. Sites of DNA where individuals differ at a single base pair are called Single Nucleotide Polymorphisms (SNP). It has been found that approximately 10 million common SNPs are responsible for the majority of the variations between DNA sequences of different people.

Either sets of nearby SNPs on the same chromosomes that are inherited together and statistically associated; or a combination of alleles at different loci on the chromosomes that are transmitted together is a haplotype. Recent studies have shown that the human genome has a haplotype block structure[1,2] such that it can be divided into discrete blocks of limited haplotype diversity. A block may contain a different number of SNPs, out of which only few SNPs are sufficient to identify the relevant block. In each block, the SNPs which can be used to distinguish a large fraction of the haplotypes are called tag SNPs. It is thought that these associations, and the identification of a few alleles of a haplotype block can unambiguously identify all other polymorphic sites in its region. Such information is very valuable for investigating the genetic basis of common diseases. Haplotype structures may provide critical information on human evolutionary history and the identification of genetic variants underlying various human traits. The phenomenon called Linkage Disequilibrium (LD) or the allelic association explains the linkage of these haplotypes.

The haplotype map of the human genome which describes the common patterns of human DNA sequence variation is available in the international HapMap project. The HapMap is expected to

be a key resource for researchers to find genes affecting health, disease and responses to drugs and environmental factors.

Linkage Disequilibrium

In population genetics, LD is the non-random association of alleles at two or more loci, not necessarily on the same chromosome. It is not the same as linkage which describes the association of two or more loci on a chromosome with limited recombination between them. A study of haplotypes consisting of a short tandem repeat polymorphism (STRP) and an Alu deletion polymorphism at the CD4 locus in 42 worldwide populations demonstrated the usefulness of comparing LD patterns of different populations for inferring population history. This study showed a common and recent African origin for all non-African human populations[3]. In a study to show relationship of the sequence features and the degree of LD in the genome, it has been found that the variation in LD is broadly similar in a genome. This study has been done obtaining genotype data from the international HapMap Project. It has been found that the LD is generally low within approximately 15 Mb of telomeres of each chromosome noticeably elevated in large duplicated regions of genome as well as within approximately 5 Mb of centromeres and other heterochromatic regions. It showed that regions of strong LD are typically GC poor and have reduced polymorphism. In addition, these regions are enriched for LINE repeats, but have fewer SINE, DNA, and simple repeats than the rest of the genome[4].

International HapMap Project and Disease Association

HapMap database has approximately five million tag SNPs. This makes genome scan approaches to finding regions with genes that affect diseases much more efficient and comprehensive.

Further, the HapMap is a powerful resource for studying the genetic factors contributing to variation in response to environmental factors, in susceptibility to infection, and in the effectiveness of adverse responses to drugs and vaccines. These studies are based on the assumptions that there are higher frequencies of the contributing genetic components in a group of people with the studying component than in a group of similar people without the studying component. Using these tag SNPs, researchers can find chromosome regions having different haplotype distributions in the two groups of people, those with a disease or response and those without. Each region is then studied in more detail to discover which variants in which genes in the region contribute to the disease or response, leading to more effective interventions. This also allows the development of tests to predict which drugs or vaccines would be most effective in individuals with particular genotypes for genes affecting drug metabolism.

All HapMap data are freely available to the public through the database dbSNP. In 2005, the International HapMap Consortium released the Phase I HapMap. It consists of over a million accurate and complete SNP genotypes generated in 269 individuals from four geographically diverse populations. These samples were selected from Yoruba in Ibadan (Nigeria), Japanese in Tokyo (Japan), Han Chinese in Beijing (China), and the CEPH (U.S. Utah residents with ancestry from northern and western Europe).

In 2007, the International HapMap Consortium released the Phase II HapMap, adding over 2.1 million SNPs to the original map using the same 269 individuals. The Phase II HapMap consists of an improved choice of tag SNPs, a better understanding of how well studies capture patterns of genetic variation and the potential to increase the power of association experiments using fixed marker sets through imputation. It also reveals novel aspects of the structure of linkage disequilibrium, including the importance of recent co-ancestry among individuals and the distribution and causes of untaggable SNPs.

Further analysis was done in HapMap three having samples from seven additional populations namely Maasai in Kinyawa (Kenya), Luhya in Webuye (Kenya), Chinese in metropolitan Denver, Gujarati Indians in Houston, Toscani in Italia (Italy), African ancestry in the Southwest USA, and Mexican ancestry in Los Angeles (USA).

HapMap provides a fine-scale genetic map and location of recombination hotspots. Further, it provides new information about the influence of natural selection on protein-changing variants and downloadable data on Genotypes, SNP Frequencies, LD data, Phasing Data, Allocated SNPs, CNV Genotypes, Recombination rates and Hotspots, SNP assays, information on Inferred genotypes, Mitochondrial and chromosome Y haplogroups. This combination of genotyping and sequencing allowed comparison of genome-wide patterns of variation.

A graphic representation of the complete set of human chromosomes or karyogram is also available in the HapMap project. This may visualize diseases or traits which are in the National Human Genome Research Institute's Catalog of Published Genome-Wide Association Studies with at least one replication study as of September 2008[(5),(6),(7),(8),(9)].

Haplotype detection techniques

According to the published studies, researches have studied haplotypes either by dividing the whole chromosome into smaller regions for analysis or by grouping haplotypes into smaller number before analysis. In the first group where the chromosomes were divided into smaller number generally have a sliding window on the searching area and assess evidence for haplotypes within each window[(2)]. Use of window is done as the number of haplotype patterns in each window is less than that of the whole chromosome. Therefore, the analysis involves fewer parameters if there is an association between haplotypes or between haplotypes and diseases.

The second approach uses the assumption that an unknown mutation occurred at some point in the evolutionary history become embedded within the historical structure called a cladogram[(10)].

The haplotypes can be searched by haplotype resolution or haplotype phasing techniques. These methods work by applying the observation that certain haplotypes are common in certain genomic regions. Therefore, given a set of possible haplotype resolutions, these methods choose those that use fewer different haplotypes overall. The specifics of these methods are of varying types. Some of them are based on combinatorial approaches such as parsimony, whereas others use likelihood functions based on different models and assumptions such as the Hardy-Weinberg principle, the coalescent theory model, or perfect phylogeny. These models are combined with

optimisation algorithms such as expectation-maximisation algorithm (EM), Marco chain Monte Carlo (MCMC), or hidden markov models (HMM)[(11),(12)].

Currently, there exist several tools to facilitate haplotype block inference and tag SNP selection. For example, the International HapMap Project website not only provides bulk download of genotype and frequency data from the International HapMap Project, but also interactive access to visualise the distribution of SNPs for any genomic region of interest. Haploview is a standalone application that performs LD and haplotype block analysis on publicly available or user supplied genotype data. HtSNPer1.0 and HaploBlock can also be used to analyse genotype data supplied by users. HaploBlock- Finder is a web based tool that allows for the inference of haplotype blocks and tag SNP selection from genotype data uploaded by users. The Genome Variation Server (GVS)] and PupaSuite are other tools with several online analysis utilities for accessing human genotype data. More recently, TAMAL was developed adopting a pre-processing strategy to facilitate the selection of potential genotyping targets.

Although there are many available computer programmes for haplotype analysis applicable to samples of unrelated individuals, many of these programmes have limitations or very specific uses. Programmes for haplotype analysis were identified through keyword searches and through various internet search engines. The available tools were considered with the view of following factors. The algorithms used, algorithm accuracy, assumptions, the accommodation of genotyping error, implementation of hypothesis testing, handling of missing data, software characteristics and web based implementations.

Highlights

1. Haplotype structure will provide valuable information on identification of genetic variants underlying various human traits as well as information on human evolutionary history. LD patterns in different regions and different populations help in the same.

2. Empirical studies have shown that the chromosomes are structured in such a way that each chromosome can be divided into many blocks. In each block there is limited haplotype diversity.

3. Several computational methods have been developed to identify haplotypes. Although there is no method to be claimed as uniformly best, more comparative methods can be used for haplotype analysis.

4. International HapMap project provides a good repository as well as a set of tools to be used for haplotype analysis. Because of the most important issue in haplotype analysis is to incorporate knowledge of haplotype patterns and to reduce the number of haplotype the existence of such database and tools make researchers' work more productive.

References

1. Smith AV, Thomas DJ, Munro HM, Abecasis GR. Genome Research, Sequence features in regions of weak and strong linkage disequilibrium 2005; **15**: 1519-34.

2. Integrating Ethics and Science in the International HapMap Project. The International HapMap Consortium. Nature Reviews Genetics 2004; **5:** 467-75.

3. A second generation human haplotype map of over 3.1 millon SNPs. The Internatinal HapMap consortium. Nature 2007; **449:** 851-61.

4. A Haplotype Map of the Human Genome. The International HapMap Consortium. Nature 2005; **437**(7063)**:** 1299-1320.

5. The International HapMap Project. The International HapMap Consortium. Nature 2003; **426:** 789-96.

6. Thorisson GA, Smith AV, Krishnan L, Stein LD. The International HapMap Project Web site. Genome Research 2005; **15:** 1591-93.

7. Tishkoff SA, Dietzsch E, Speed W. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Sciense, 1996; **271:** 1380-87.

8. Eskin E, Heralperin E, Sharan R. Optimally Phasing Long Genomic Regions using Local Haplotype. Nature Genetics, 2001; **29**(2): 229-32.

9. Zhang P, Sheng H, Morabia A, Gilium TC. Optimal step Length EM algorithm (OSLEM) for the estimation of haplotype frequency and its application in lipoprotein lipase genotyping. BMC Bioinformatics. 2003; **4**: 3.

10. Wang J, Wang W, Li R, Yingrui L, Tian G, Goodman L, Fan W. The diploid genome sequence of an Asian. Nature 2008; **456**.

11. Finishing the euchromatic sequence of the human genome. Consortium, International Human Genome Sequencing. Nature 2004; **431:** 931-45.

12. Initial sequencing and analysis of the human genome. Consortium, International Human Genome Sequencing. Nature 2001; **409:** 860-921.