Genome-wide scan with SNPs

58305112 Research Seminars on Data Analysis for Bioinformatics
Seminar # 8

Presentation: Tero Hiekkalinna
Scribe: Sarish Talikota

March 11, 2005

Introduction

There has been considerable interest in the use of single nucleotide polymorphism for understanding the genetics of complex human diseases. In contrast to application of SNPs in genetic studies, the analysis of SNP data pose a number of challenges. Abundance of SNPs is generally seen as an advantage over other markers but insufficient in term of information. It has been estimated that for a genome-scan over 500,000 SNPs are required, and hence more specialised databases are used for these studies.

Basic genetics

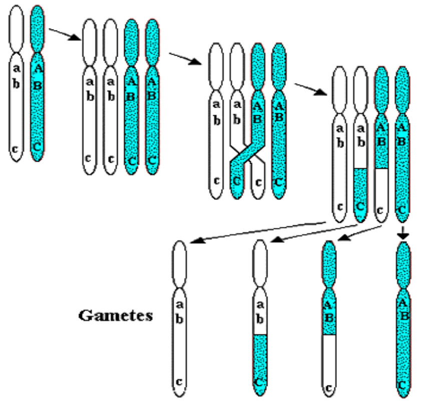
Mendel's Laws

The principles of heredity were written by the Augustinian monk Gregor Mendel in 1865. Mendel summarized his findings in two laws, the *Law of Segregation* and the *Law of Independent Assortment*. The Law of Segregation states that the members of each pair of alleles separate when gametes are formed and the gamete will receive one allele or the other. The Law of Independent assortment states that each member of a pair of homologous chromosomes segregates during meiosis independently of the members of other pairs, so that alleles carried on different chromosomes are distributed randomly to the gametes.

Meiosis

Meiosis is the process of reduction in the amount of genetic material. This comprises two successive nuclear divisions with only one round of DNA replication. There are two possibilities of arrangement of chromosomes. One possibility is that out of four gametes two will have maternal chromosomes and the remaining two will have paternal chromosomes and the other possibility could be that all the four chromosomes will share one paternal and one maternal chromosome.

Crossing-over, recombination and Linkage



Crossing-over and recombination during meiosis

During the early stages of cell division in meiosis, two chromosomes of a homologous pair may exchange segments producing genetic variation. This exchange of genetic material is called *recombination* or *crossing-over*. The point of cross-over is called *chiasma*, at such chiasmata, bits of crossed over chromatids can swap with one another. Genes that tend to stay together after recombination are said to be *linked*. A recombination in meiosis will always lead to 2 recombinants and 2 parental types (haploid gametes). Recombination will rarely separate loci that lie very close together on the chromosome because of no cross-over located precisely in the small space between the loci. Hence these alleles that are close together co-segregate more often than expected by independent assortment and likewise, if loci are far apart then crossing-over will separate those forming recombinants. *Recombination fraction* (**θ**) is the measure of genetic distances between two loci or the probability that a recombinant gamete is transmitted. If two loci are on different chromosomes, they segregate independently, hence **θ=0.5** and if they are on the same chromosome and right next to each other they will segregate together during meiosis and hence **θ=0**. If two loci are linked then **θ<0.5** and if they are not then **θ=0.5**.

Human genetics

Nuclear genome comprises approximately 3200 Mb nucleotides of DNA of which only 1.5% is functional with over 20000 genes. This large difference in coding region and the complete genome is because of large amount of repetitive elements spanning over 50% of the genome. Most of the cells in the humans are diploid and contains two copies of genome. Information is contained in 24 chromosomes, of which 22 pairs are autosomes and two sex chromosomes. Each pair contains maternal and paternal copies. At each locus there are two copies of alleles, one inherited from father and one from mother, which is called as genotype. Alleles are alternative forms of a gene and are present at the specific locus, which is at specific position on the genome. Variation in sequence at this position or locus will lead to a phenomenon called polymorphism. Polymorphism is a change in the DNA sequence or repeat element at a specific location, these are called markers. Every individual might change at this location and since that they are spanned all over the genome they are useful in mapping human disease genes as they are close to the disease gene. Many such markers have been identified of which some are RFLP (Random Fragment Length Polymorphism), RAPD (Random Amplification of Polymorphic DNA), AFLP (Amplified Fragment Length Polymorphism), Microsatellites and recently SNPs (Single Nucleotide Polymorphism). These markers have revolutionised the field of human genetics in understanding the genetic basis of diseases.

Genetic Markers

Difference in the genome from one individual is about 0.1% and this difference has the potential to effect the function of the gene and hence the phenotype of the individual. But not all markers are associated with visual phenotype. Most commonly used genetic markers these days are microsatellites and SNPs because of their advantageous over first generation DNA markers (RFLPs, RAPDs etc.).

Microsatellites

Microsatellites are DNA regions with variable number of short tandem repeats flanked by a unique sequence. The tandem repeats are usually simple dinucleotides ($(CA)_n$) with dinucleotide repeated about ten times. They are highly polymorphic as there could be many genotype classes an allele (10 alleles could have 55 possible genotype classes). Human genome has highly polymorphic mono, tri and tetra or bigger repeat elements and the high degree of polymorphism in the repeats make them marker of choice for mapping studies etc.

Some of the advantages of using Microsatellites as useful markers are that they are locus specific, codominance of alleles (heterozygotes can be distinguished from homozygotes), PCR based, random distribution throughout the genome. Sometime information can be misleading, like the heterozygotes can be misclassified as homozygotes when null alleles occur due to mutation in primer annealing sites. Microsatellites have many synonyms, like SSLP (Singel Sequence Length

Polymorphism), SSR (Simple sequence Repeats), STMS (Sequence Tagged Microsatellites).

SNPs

SNP (Single Nucleotide Polymorphism) pronounced as *snip* is a small genetic variation in the DNA sequence. SNP variation occurs when a single nucleotide, say **A**, replaces one of the three nucleotides (**T, G, C)**.

Seq 1 ATT $\boxed{\mathbf{A}}$ AATCCA
Seq 2 ATT $\boxed{\mathbf{T}}$ AATCCA

SNPs occur more than *1%* in the human population as the coding part of the DNA in humans is *~5%*. SNPs occur mostly out side the coding regions but SNPs in the coding regions would be interesting as they add to variation in the function of the protein. If DNA to be considered as a SNP, it is considered that atleast frequent allele should have a frequency of 1% or greater. SNPs are bi-allelic in practice and the reasons being, low frequency single nucleotide substitution at the origin of SNPs and bias in mutations.

99.9% of one individual DNA sequences will be identical to that of another person. Of the 0.1% difference, over 80% will be single nucleotide polymorphisms (SNPs).

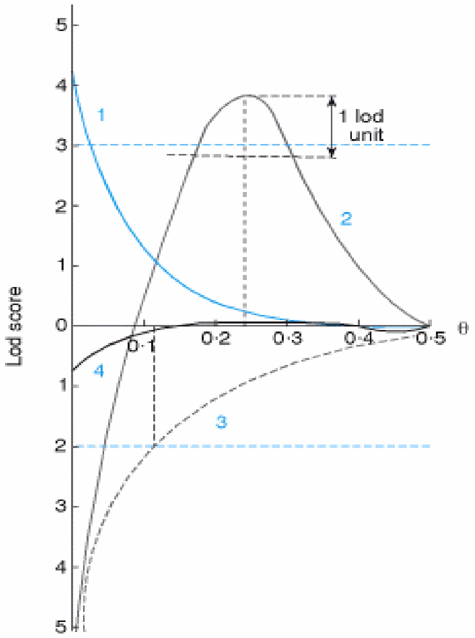
Mutations results in either *transitions* (Transitions are Interchanges of purines (A ←→ G) or of pyrimdines (C ←→ T), which therefore involve bases of similar shape) or *transversions* (Transversions are interchanges between purine and pyrmidine bases, which therefore involve exchange of one-ring and two-ring structures). In nature we see 2 transitions and 4 transversions and hence the transitions and transversions ratio should be 0.5. But most of the data available today shows high bias towards transitions. One probable explanation would be *high spontaneous rate of deamination of 5-methyl cytosine (5mC) to thymidine* (C←→T) SNPs and
G←→A) on the other strand.

Human genome contains about 10-30 million SNPs with an average of SNP every 100-300 bases. More than 4 million SNPs have been identified and the information is publicly available. They are stable from evolutionary stand point by not changing much from generation to generation making them easier for population studies.

Linkage Disequilibrium (LD)

Linkage disequilibrium (LD) is a phenomenon that when two chromosome locations (two loci, two markers) are so close to each, that there is a lack of ancestral recombination event in between. Since LD reflects ancestral recombination events, it provides a starting point to the current recombination events, something like the initial condition. LD can be summarised with several statistical measure such as **D** (linkage disequilibrium coefficient) (Kaplan and Weir, 1992), **D′** (divide **D** by its maximum possible value, given the allele frequencies at two loci) and $\mathbf{r^2}$ (measure of choice for quantifying and comparing LD in the context of gene mapping). **D′** is one of the most commonly used measures of the extent of gametic disequilibrium between multi-allele loci but in case of low allele frequencies $\mathbf{r^2}$ has more reliable sample properties than **D′**. Some of the other measure that were proposed other than those mentioned above are **Δ=r** (Hill and Wier, 1992), **Ψ** (Edwards, 1963).

Linkage analysis



Linkage analysis is to infer the positions of two or more loci by examining patterns of allele transmissions from parent to offspring, or patterns of allele sharing by relatives. Linkage is resulted from the recombination events in the last 2-3 generations, which differs from LD which is resulted much earlier, ancestral recombination event. Linkage analysis is statistical test to look if the genetic marker co-segregates with the loci (disease) in the families. Result of a linkage analysis is some LOD scores at various recombination fractions with positive (3) scores for linkage and negative (-2) against linkage. The *LOD score* is the decimal log likelihood ratio,

$$Z(\theta) = \log_{10} L(\theta) / L(\theta{=}0.5)$$

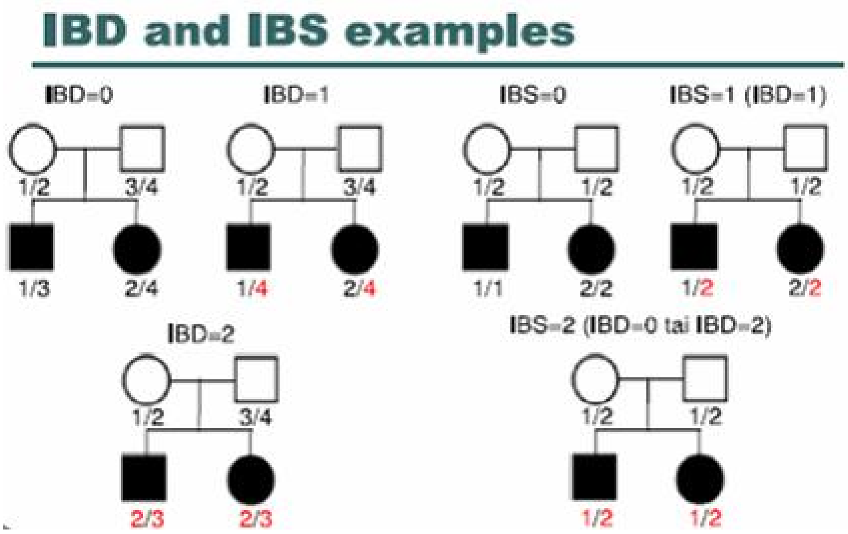
LOD score of 3 shows a significant evidence for linkage, with a 5% change of error. Linkage can be rejected if Z is < -0.2, but a LOD score between -2.0 and 3.0 is considered inconclusive and warrants additional study.

$$-2.0 \leq Z(\theta) \geq 3.0 \text{ (Insufficient Data)}$$

Parametric linkage analysis requires mode if inheritance to be specified prior to analysis. Standard LOD score analysis is called parametric because it requires precise genetic model, detailing disease allele frequency, maker allele frequency, and penetrance of each genotype. Where as non-parametric linkage analysis (ASP – Affected SIB Pair method) is based on allele sharing by relatives. These methods ignore unaffected people, and look for alleles or chromosomal segments that are shared by affected individuals.

IBD & IBS

There are two different measures of allele-sharing, identity-by-state (IBS) or identity-by-descent (IBD). Two alleles of the same form (i.e. having the same DNA sequence) are said to be IBS. If, in addition to being IBS, two alleles are descended from (and are therefore replicates of) the same ancestral allele, then they are said to be IBD. The general idea of non-parametric linkage analysis is that, in the vicinity of a disease locus, sib-pairs who are concordant for disease status (i.e. both affected or both unaffected) should show an increase in allele-sharing, and those who are discordant for disease status (i.e. one affected and one unaffected) should show a decrease in allele-sharing, from the level of allele-sharing expected of sib-pairs. Of the two measures of



allele-sharing, IBS is easier to define and determine, but IBD is more directly relevant to linkage.

Affected sib-pairs are often far more informative for linkage than unaffected sib-pairs, or sib-pairs with one affected and one unaffected member, under many plausible models of complex inheritance. This method is based on comparing observed IBD distribution of the effected sib-pairs to the expected distribution (¼, ½, ¼) under the null hypothesis of 'no' linkage. The greater the deviation of the expected IBD distribution from the null hypothesis values of 1/4, 1/2, 1/4, the greater is the power for detecting linkage.

Genome-wide scan

Genome-wide linkage scans have traditionally employed panels of microsatellite markers spaced at intervals of approximately 10 cM across the genome. However, there is a growing realization that a map of closely spaced single-nucleotide polymorphisms (SNPs) may offer equal or superior power to detect linkage, compared with low-density microsatellite maps. Genome-scans with microsatellites will result ~350 markers with one marker every 5-10cM (1cM ~ 1 000 000 bases) where with SNPs 3000 – 10000 markers can be obtained with one SNP every 5Kb. Linkage studies are more often performed on a genome-wide scale, as they have benefits of assessing the entire genome unbiased. Recently genome-wide analysis was performed using microarrays to look for natural variation in the expression profiles from the members of CEPH Utah pedigrees. Information pertaining to genotypes and phenotypes are available for research scientist at (http://www.cephb.fr and http://locus.umdnj.edu/nigms/cehp/cehp.html).

Genetic analysis of genome-wide variation in human gene expression (Michael Morley et al., Nature 2004)

Genetic variation is a common phenomenon that occurs in species from yeast to humans. The relative amount of variation within and between populations varies from species to species, depending on history and environment. This natural variation is primarily because of change in the expression levels of the genes. Microarrays methods are used these days to look the differential expression of the genes and to understand the significance behind this expression pattern. In this paper, microarrays were used to measure baseline expression level of genes in immortalized B cells from members of (Centre d'Etude du Polymophisme Humain) CEPH/Utah pedigrees. In this analysis they estimated the variance in expression levels of ~8500 genes among unrelated individuals (94 CEPH grandparents) and mean of variance of array replicates (2 per individual). They obtained SNP markers from SNP consortium (http://snp.cshl.org/) and performed the genome wide linkage analysis using S.A.G.E (Statistical Application of Genetic Epidemiology) computer program (http://darwin.cwru.edu/sage/index.php). About 3554 most variable expression phenotypes in 14 families were selected for analysis. Further analysis was done to check he evidence for linkage with varying stringency levels (*t*-value). For most stringent *t*-value corresponded to <4.2 x 10^{-7} (genome-wide significance level = 0.001) and <3.7 x 10^{-5} for relaxed stringency (genome-wide significance level = 0.05 with low t-value). They found 142 expression phenotype with evidence of linkage found over *p*-value threshold and by relaxing the stringency (t>4) it generated 984 expression phenotypes, which is 7 times more prone to false positive generation. Considering the regions that are liked to the expression levels to be regulatory regions, they categorised 142 expression phenotype as either *cis* or *trans* regulators. They could distribute the phenotype with 19% being *cis*-acting, 77.5% *trans*- acting and remaining having two regulators (out of 984 obtained with low stringency, there were 16% with two or more regulators). Hence the correlation between the expression pattern and linkage is evident from these analyses. Apart from genomic regions with regulators that effect single phenotypes in *cis* or *trans*, they also found transcriptional regulators that influence multiple expression phenotypes which are called master regulators. Where the closely linked genes are influenced by same *cis* or *trans* regulators. It was a nice finding that showed correlation between linkage and gene expression phenotypes but the linkage analysis part is not convincing

Problems using SNPs

▪ SNPs are generally less informative than microsatellites because of their biallelic nature. This may be a disadvantage for association and linkage analysis. Polymorphism information content (PIC) which is important in linkage analysis is low for SNPs (0.375) when compared to microsatellites (0.7-0.8).
▪ closely linked SNP markers with strong LD will generate false positives.
▪ Reduced ability to detect genotyping errors, as they are biallelic.

Problems in statistical analysis

SIBPAL which is used for multipoint genome-wide linkage analysis does not use grandparent genotype and phenotype information. It also assumes Linkage equilibrium between marker loci in multipoint analysis. Understanding patterns of LD between markers will help interpretation of results of association studies.

Conclusions

■ Compared to other markers SNPs are popular due to cheap genotyping costs.
■ SNPs are useful for association studies (LD) due to low mutation rate. It is important to know LD between SNPs. Generation of false positives is more when LD between markers is ignored.
■ Genome-wide scans are becoming more informative because of the methodologies used. Microarray techniques and linkage analysis in the paper by Michael Morley et al. have showed determinants that contribute to variation in the human gene expression.
■ SNPs have some drawbacks but when compared to other markers they are more efficient and SNP consortium is growing to meet the requirements of genome-wide scans.

References

Edwards AWF (1963) The measure of association in a 2x2 table. J. Roy. *Stat. Soc. A*. 126:109-114.

Hill W and Weir B (1994) Maximum likelihood estimation of gene location by linkage disequilibrium. *Am. J. Hum. Genet.* 54: 705-714.

Kaplan, N. and B.S. Weir. 1992. Expected behavior of conditional linkage disequilibrium. *Am. J. Hum. Genet.* 51:333-343.

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. Genetic analysis of genome-wide variation in human gene expression. *Nature*. 2004 Aug 12;430(7001):733-4.

Tom Strachan, Andrew P. Read. Human Molecular Genetics 3rd Edition. Garland Science (2003).