BioStats 710 Final Exam

May 2, 2016

(Chapter 5 through Chapter 7)

Jeff Du

1 Chapter Five

• Linkage Analysis: Use pedigrees with diseased individuals with genetic marker (polymorphism, detectable location), to access whether the DSL (Disease Susceptibility Loci) is in "linkage" with any of the markers.

• Recombinant gamete: Two types of gametes are possible when following genes on the same chromosomes. If crossing over does not occur, the products are parental gametes. If crossing over occurs, the products are recombinant gametes.

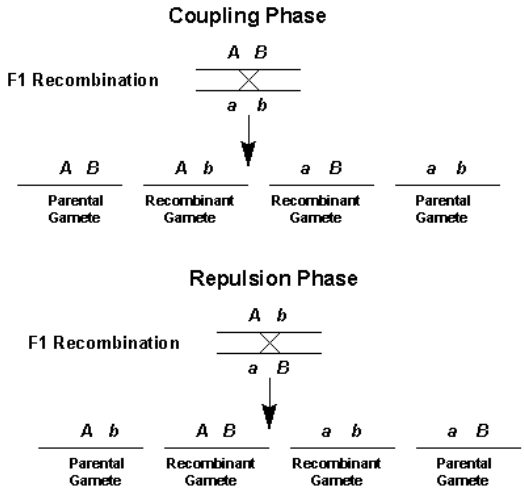

Figure 1: A sample of Recombination Gametes

• map function: Devised by Haldane, map function transform θ (recombination fraction) into adistance measure by using the assumption that the number of crossovers follows a Poisson distribution with mean 2L.

$L = E(\sharp crossovers)/2$
and $E(\sharp crossovers) = 2L$

$$P(X = k) = \frac{e^{-2L} * (2L)^k}{k!}$$

$$L = -[Ln(1 - 2\theta)]/2$$

Here L is measure by Morgans, 1 Morgan = distance between 2 loci.

• Morgan: Morgans or centimorgans (cM) measures the expected number of crossovers between two loci per gamete.

• physical map: A physical map gives the locations of identifiable landmarks on DNA, distance by base pairs. 1,000,000 Bp $\approx$ 1 centiMorgan.

• association analysis: Basically look at whether marker alleles are associated with the trait.
Very different from linkage; It does not care which allele is congregating with DSL and maybe different from family to family.

• linkage disequilibrium: The genetic association will also be visible between the marker and the phenotype if a particular allele at the genetic marker tends to appear together on the same gamete with the disease allele at the disease locus. This latter concept, the association of alleles at two loci, is referred to as linkage disequilibrium (LD) – refers to association between alleles at different Loci. Also, recombination breaks down LD, more recombination, faster LD decays.

• haplotype: The set of alleles lying on the same chromosome is called the haplotype. (i.e. AB and ab are two haplotypes)

• D prime (D'): The coefficient of linkage disequilibrium D is not always a convenient measure of linkage disequilibrium because its range of possible values depends on the frequencies of the alleles it refers to.
Lewontin's suggested normalising D by dividing it by the theoretical maximum for the observed allele frequencies as follows: $D'\ (D\ prime) = D/D_{max}$.

$$\text{If } D_{AB} < 0,\ D'_{AB} = \frac{D_{AB}}{min(P_A P_B, P_a P_b)}$$

$$\text{If } D_{AB} > 0,\ D'_{AB} = \frac{D_{AB}}{min(P_A P_b, P_a P_B)}$$

D' ranges from -1 to +1.
more likely to take extreme values when allele frequency are small.
± 1 implies that at least one of the possible hapletypes was not observed.

• phase: The order or sequence of several alleles or genes. For a set of al-

leles or genes, the sequence of these correlated alleles are called phase in genetic.

• $\triangle$: Also called r^2, standard squared correlation coefficient for allele freq at two markers.
range from 0 to 1.
$r^2 = 1$ implies markers provide the same statistical information.
population geneticists, like r^2 value.
measures the loss in efficiency when marker A is replaced with marker B, in association study.

$\star$ explain how the Haldane map function is derived
If we have three genes (alleles) A, B, and C arranged in the same order on a chromosome (gene). Then we have: $r_{AC} = r_{AB} * (1 - r_{AC}) + r_{BC} * (1 - r_{BC})$
Haldane pointed out that this relationship implies another, namely that the probability that there are k recombination events between two loci m map units apart is given by the Poisson distribution:

$$p(m, k) = \frac{e^{-m} m^k}{k!}$$

a recombination event between A and C requires that there be an odd number of recombination events between them (1, 3, 5, ...), i.e.,

$$r = \sum_{k=0}^{\infty} \frac{e^{-m} m^{(2k+1)}}{(2k+1)!} = \frac{1 - e^{-2m}}{2}$$

We could have a map unit as: $m = -ln(1 - 2r)/2$

$\star$ name the rule of thumb that relates base pairs to Morgans
The overall rule of thumb is that one centimorgan of genetic distance is about one million base pairs of physical distance. However, this comparison can vary dramatically across certain parts of chromosomes.

2 Chapter Six

• linkage analysis: Mapping disease locus to position relative to known markers. The goal is to make inference about cosegregation of 2 (or more) loci in family. these loci could be marker loci. or at least one could be a candidate disease locus. in general, the more variable the marker, the more useful it is for linkage analysis. Because, each founder will bring a different copy to the pedigree and it will be possible to determine exactly which chromosomal segments were transmitted.

• ordered genotype: Genotype plus phase information.
$A_1B_2/A_2B_1 \rightarrow Genotypes A_1A2$ *and* B_1B_2
A_1B_2 are on the same chromosome, and therefore comprise a haplotype.
A_2B_1 are on the same chromosome, and therefore comprise a haplotype.

• LOD score: In genetics, the LOD score is a statistical estimate of whether two genes, or a gene and a disease gene, are likely to be located near each other on a chromosome and are therefore likely to be inherited.
The LOD score is a measure of support for an arbitrary value of θ in the range (0, 1/2), which is maximized when θ is the maximum likelihood estimate.
$LOD - score = log_{10}(LR(\theta))$
In likelyhood analysis, Inference is often based on the lokelyhood ration: $LOD(\theta) = log_{10}\frac{L(\theta)}{L(1/2)}$

• incomplete penetrance: the presence of a gene that is not phenotypically expressed in all members of a family with the gene.

• parametric linkage: Assuming a given genetic model (parametric part), we write down a likelyhood (or the observed) for the pedigree data (phenotypes, genotypes, over covariates, phase), by generally making a number of simplifying assumption that allow us to express it in turns of a few unknown parameters.
Establish a pedigree
Make a number of estimates of recombination frequency
Calculate a LOD score for each estimate
The estimate with the highest LOD score will be considered the best estimate

• non-parametric linkage analysis: based on idea that if a locus is associated with a disease, the relatives with similar phenotypes should share more alleles IBD at that locus than (would be) expected by their relatedness alone.

note: non-parametric linkage analysis is actually full parametric assumptions. it is nonparametric in that it does not assume a model of inheritance (autosomal dominant etc).
note: non-parametric linkage depends on marker loci that are close to disease locus also showing an increase in IBD relative to expectation.

3 Chapter Seven

• family-based controls: We use family information as 'controls'.
these approaches are somewhat in terminated between linkage and association in that they require some pedigree information, typically from parents.

However, they retain the main feature of association. Studies in that the relationship between disease status and genotype is compared.

Note: we will consider
1. study design in which an affect individual and his/her parents are sampled and genotyped.
2. Alleles that were transmitted to the affected offspring are compared to the alleles that were not transmitted.

• trio: The most popular family study design involves parents and their affected offspring, commonly called trios, in other words: nuclear families with one offspring and both parents.

• manhattan plot: A Manhattan plot is a type of scatter plot, usually used to display data with a large number of data-points - many of non-zero amplitude, and with a distribution of higher-magnitude values, for instance in genome-wide association studies (GWAS).[1] In GWAS Manhattan plots, genomic coordinates are displayed along the X-axis, with the negative logarithm of the association P-value for each single nucleotide polymorphism (SNP) displayed on the Y-axis, meaning that each dot on the Manhattan plot signifies a SNP. Because the strongest associations have the smallest P-values (e.g., 1015), their negative logarithms will be the greatest.

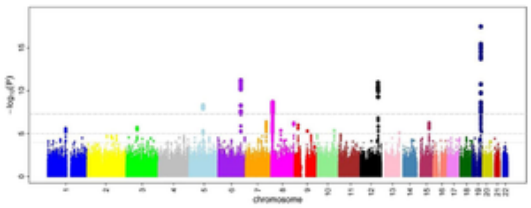

Figure 2: A sample of Manhattan Plot

• genomic control: Genomic control makes the assumption that the population substructure at the marker loci of interest is the same as for the non-functional markers, the selection of the non-functional markers is critical for the validity of the approach.
Genomic control (Devlin and Roeder) addresses the effects that population-

admixture has on the variance of the test statistic. Rather than relying on the theoretical variance, which may or may not be correct, the approach simply estimates empirically the variance in the χ^2 statistics computed as the null markers. Avariance inflation factor is estimated by comparing the empirical variance to the variance of the χ^2 distribution.

• qqplot: In GWAS, a quantile-quantile (Q-Q) plot can be used to characterize the extent to which the observed distribution of the test statistic follows the expected (null) distribution. Generates a Quantile-Quantile plot for -log10 p-values from genome wide association tests.

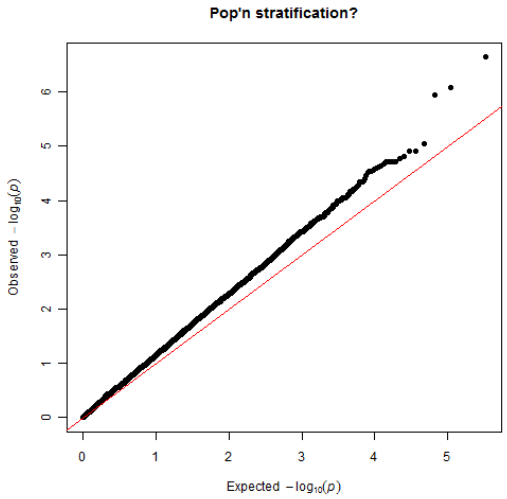

Figure 3: A sample of QQ-Plot

• genome-wide association study: In genetic epidemiology, a genome-wide association study (GWA study, or GWAS), also known as whole genome association study (WGA study, or WGAS), is an examination of many common genetic variants in different individuals to see if any variant is associated with a trait.

• whole exome study: a technique for sequencing all the expressed genes in a genome (known as the exome). It consists of first selecting only the subset of DNA that encodes proteins (known as exons), and then sequencing that DNA using any high throughput DNA sequencing technology. There are 180,000 exons, which constitute about 1 percent of the human genome, or approximately 30 million base pairs.[1] The goal of this approach is to identify genetic variation that is responsible for both Mendelian and common diseases such as Miller syndrome and Alzheimer's disease without the high costs associated with whole-

genome sequencing.

•λ used in genomic control:

The genomic inflation factor λ is defined as the ratio of the median of the empirically observed distribution of the test statistic to the expected median, thus quantifying the extent of the bulk inflation and the excess false positive rate.

$\lambda = median(\chi^2)/0.456$

$\chi^2_{adjusted} = \chi^2/\lambda$

⋆ under what conditions will it be possible to observe an association between disease and a marker allele?
Answer: linkage disequilibrium?

⋆ List the general conditions for confunding in a case-control study.
Answer: 1 Expossure (dd or dD U DD) is associated with confounder (P) in controls. 2 Disease is associated with confounder (P), among the unexposed (dd).

Note: both of these conditions must hold, for a study to reach the wrong result due to confunding. For example, if allele frequencies where the source in P1 as P2 -¿ a pooled analysis would reach correct conclusion.

⋆ Explain the basic idea behind a trio-based genetic association study.
Answer: 1 study design in which an affect individual and his/her parents are sampled and genotyped. 2 alleles that were transmitted to the affected offsprings are compared to the alleles that were not transmitted.

⋆ Derive the TDT. **Answer:** TDT: Transmission Disequilibrium Test, a test for independence of allele and transmission using McNewars

⋆ Why is TDT robust to population stratification?
Answer: The TDT (transmission-disequilibrium test) is a family-based test thus avoiding problems of population stratification - that uses association thus allowing fine mapping.

⋆ Describe 3 approaches for correcting for population stratification using marker data. explain their pros and cons.
Answer: 1 Linkage studies. 2 Association study. 3 case-control study.