An introductory overview of the current state of statistical genetics

Cavan Reilly

Division of Biostatistics,

University of Minnesota,

e-mail: cavanr@biostat.umn.edu

DNA

DNA is a molecule that exists in the nucleus of most eukaryotic (e.g. human) cells-we refer to the DNA of an organism as its *genome*.

DNA is a molecule that has a sequence of sugar molecules (and a phosphate group) with 1 of 4 distinct *bases*: adenine (A), cytosine (C), guanine (G) and thymine (T). We think of DNA sequences in terms of sequences of bases.

A base bound to a sugar and phosphate group is called a *nucleotide*.

Each base has a complementary base that it will naturally form a hydrogen bond: the complement pairs are A-T and G-C.

Chromosomes

In its naturally occurring form, a single stranded DNA molecule binds to another single stranded DNA molecule so that each base on one strand binds to its complement.

The human genome has approximately 3 billion basepairs and about 99.9% of these are identical among humans.

In healthy humans, DNA exists as 23 pairs of distinct molecules known as *chromosomes*. Other organisms have different numbers of chromosomes and some even have more than just a pair of each.

One pair, the sex chromosomes, determine the sex of the person.

Gene expression

In its naturally occurring state DNA is bound to a variety of molecules, and the manner in which these other molecules bind to DNA impacts the functionality of DNA.

Which molecules are bound to a DNA molecule varies over time and in response to stimuli.

Which molecules are bound to the molecule impacts how the cell uses the DNA.

Cells use DNA as a template for making proteins.

Proteins are molecules generated by cells that allow the cell to fulfill a function.

Gene expression (cont.)

A subsequence of the genome that is used to make a specific protein is called a *gene coding region*, or sometimes just a *gene*.

The term *gene* is used somewhat loosely: we might mean the subsequence of bases at the gene coding region or the location in the genome of the subsequence.

The term *locus* is used to refer to a location in the genome (there may or may not be a gene at a locus).

Although the number of genes isn't fully known, there are about 21,000-23,000 human genes.

Gene expression (cont.)

Transcription is the process by which alterations in the collection of proteins bound to a DNA molecule leads to copying a gene coding region to create an RNA molecule.

This RNA molecule has information about the particular sequence of bases at this gene coding region.

Translation is the process by which a cell uses the information in a modified version of the RNA molecule (called mRNA) to create a protein.

Gene expression is a term that is used to describe the entire process of translation and transcription of a gene.

Mendelian genetics

Since there is some variation among DNA sequences in humans, there are genes that have different subsequences among people.

Genes (or loci) that exhibit differences among members of a species are called *polymorphic*.

The possible subsequences at a gene (or a locus) are called *alleles*.

Since all humans have 2 copies of each chromosome, every gene has 2 alleles. One is inherited from the mother (the maternal allele) and the other from the father (the paternal allele).

Mendelian genetics (cont.)

The alleles that a subject has at a locus (or set of loci) is called that subject's *genotype*.

If, for some human, the 2 alleles at a locus are distinct we say that subject is *heterozygous* otherwise the subject is *homozygous*.

An observed feature of an organism is referred to as a *phenotype*.

Mendelian genetics (cont.)

If a trait is at least partially genetically transmitted, the extent to which it is hereditary is called its *heritability*.

There are methods for estimating this just using observable features of organisms-in particular you don't need to know which genes are responsible for the trait.

Note: Heritability is used in a technical sense that only applies to continuous traits.

The manner in which a genotype impacts a hereditary phenotype depends on the trait.

Mendelian genetics (cont.)

In Mendelian genetics, individuals with homozygous genotypes display the phenotype corresponding to the common allele, but heterozygous genotypes can display either phenotype.

If an organism has a heterozygous genotype and displays a certain trait, then that trait is said to be dominant and the other trait is recessive.

Traits that are controlled by one gene in a recessive or dominant fashion are called *Mendelian traits* otherwise *complex traits*.

Cystic fibrosis is a well documented Mendelian traits: currently 12,000 genes have been implicated in Mendelian traits.

Haplotype

When we consider 2 loci simultaneously, if know which pairs of alleles are on the same chromosome then we know the *haplotype*.

For example, consider 2 genes (denoted by a letter) where each gene has 2 alleles (which we distinguish by case).

If someone has genotype AA and Bb, then we know that 1 chromosome must have alleles AB and the other chromosome must have alleles Ab-in this case we know the haplotype.

Haplotype (cont.)

In contrast, if someone has genotype Aa and Bb then 2 configurations are possible: one chromosome has alleles AB while the other has ab or one chromosome has alleles Ab while the other has alleles aB-here haplotype is unknown.

Conventional assays for genotyping a subject (i.e. determining the genotype for many loci simultaneously) do not give information about haplotypes, they only provide information about genotypes for each locus separately.

We will see this presents a fundamental challenge for determining which genes are involved in disease processes.

Meiosis

One important function of some cells is that they go through the *cell cycle* and replicate-a process called mitosis.

During the process of replication, the cell makes copies of its chromosomes and then passes the copied chromosomes on to the next generation-this can lead to errors, or *mutations*.

Such mutations are thought to be an important source of genetic diversity.

However there are much stronger forces that lead to genetic diversity in organisms that utilize sexual reproduction for creating further generations.

Meiosis (cont.)

Meiosis is the process that leads to the formation of sex cells (i.e. sperm and egg), also known as *gametes*.

Meiosis starts in a manner similar to mitosis: all copies of all of the chromosomes are created and a pair of each is transmitted to the pair of resulting cells.

Each of these daughter cells has 2 copies of each chromosome-one from each of the organism's parents.

Meiosis (cont.)

In the next step of meiosis the daughter cells split again to create a total of 4 cells where each of these cells only has one of each chromosome (rather than a pair).

When the daughter cells divide, each chromosome from a pair is equally likely to be transmitted to the resulting cells.

Thus each gamete has a mix of chromosomes-some come from the maternal source and some from the paternal source.

Meiosis (cont.)

When 2 gametes fuse the resulting cell will thereby have a pair of each of the chromosomes and the resulting organism will have a unique genome.

This process leads to substantial genetic diversity.

Recombination

While the random assortment of chromosomes to gametes generates new genomes, there is another mechanism that increases genetic diversity: *recombination*.

During meiosis, before the pairs of chromosomes split up and go to different cells, they sometimes will swap sections of chromosomes between pairs to create entirely new chromosomes.

This process is called recombination.

Recombination (cont.)

As an example, suppose we consider the fate of chromosome 2 after a recombination. If we use letters to denote loci, case to distinguish alleles and suppose that the maternal and paternal alleles are all distinct then:

Before recombination:
chromosome 2 (maternal): `ABCDEFG`
chromosome 2 (paternal): `abcdefg`

After recombination:
chromosome 2 (one copy): `ABCdefG`
chromosome 2 (the other copy): `abcDEFg`

This process creates even further genetic diversity.

Recombination rate

The *recombination rate* between 2 loci is the probability of the 2 loci having alleles from different parental sources.

If 2 loci are very close then it is unlikely that a recombination will separate them, consequently the recombination rate is low.

Conversely, if 2 loci are on distinct chromosomes then the probability that the alleles come from different parental sources is $\frac{1}{2}$ since chromosomes pass to gametes independently of one another.

Hence for all pairs of loci, the recombination rate is in the interval $\left(0, \frac{1}{2}\right]$

References

There are many great texts on molecular biology and genetics-some I have liked are

Molecular Cell Biology, Lodish et al.

Human Molecular Genetics, Strachan and Read

I also cover this material in my text:

Statistics in Human Genetics and Molecular Biology, Reilly

Markers

A *marker* is some known feature of a genome.

For example, the center of a chromosome can be observed under a microscope (during certain stages of the cell cycle) hence this is a marker.

A *polymorphic marker* is a marker that varies among members of a species-all (normal) chromosomes have a center so the center is not a polymorphic marker.

A *single nucleotide polymorphism* (SNP) is a single nucleotide that is known to vary among individuals.

While older studies used other types of markers, SNPs are the markers of choice today.

SNPs

Millions of SNPs have been identified in the human genome and low cost assays are available for genotyping subjects at these SNPs.

One of the technologies for such large scale genotyping is similar to the technology used for measuring gene expression (i.e. microarrays) and some are still conducting research on optimizing these sorts of assays.

Another avenue of research in this area deals with genotyping errors.

With DNA from related subjects, genotyping errors can be detected since certain configurations of alleles are not possible.

Genetic linkage

Linkage analysis is a set of techniques whose goal is to estimate the recombination fraction between loci where (at least) 1 locus alters risk for disease.

If we know the recombination rate between a marker (whose location on the genome is known) and a locus that alters disease risk is near zero, then we have found the location of a locus that alters disease risk (i.e. we have found the “disease gene”).

By using many markers that cover the genome we can therefore go fishing for the location of the disease gene-*genome scans*.

Genetic linkage (cont.)

By knowing which genes have alleles that alter risk for some disease we can gain insight into the disease and perhaps develop better treatments for patients with the disease.

We can also provide information for pregnant mothers regarding the probability that their offspring will have a particular disease.

Such genetic counciling is very important for some human subpopulations as debilitating recessive traits circulate in some subpopulations.

A simple example

To understand the difficulties of linkage analysis, we will consider a simple example.

Suppose we assume that there is a gene with certain alleles that alter the risk for some disease.

Suppose further that we are willing to assume that this disease is a dominant trait.

linkage example (cont.)

Consider a mating where one parent is affected by this trait and the other is not (and some offspring are affected and some are normal).

Note that by assuming the disease has a genetic source and acts in a dominant fashion we know the genotype for the disease gene:

all affecteds (i.e. those with the disease) are heterozygous with a copy of the disease allele

all normal subjects are “homozygous” for the “normal“ allele (there could be multiple normal alleles).

linkage example (cont.)

Suppose we genotype both parents at some marker and discover that both parents are homozygous, but with different alleles (say the diseased parent is A_1, A_1 and the other is A_2, A_2).

Use D to represent the disease allele and N to represent the normal allele at the disease locus.

This implies that all offspring will be heterozygous at this marker allele (with genotype A_1, A_2).

Hence an affected offspring will be heterozygous at the marker locus and the disease locus, i.e. this offspring will be *doubly heterozygous*.

linkage example (cont.)

Suppose these parents have an affected offspring. Note that we will also know the haplotype here which we denote $D, A_1|N, A_2$.

Now suppose the affected offspring mated with someone who is not affected by this disorder and was homozygous at the marker locus.

If this mating resulted in an affected offspring then by examining the alleles at the marker locus we are able to determine if a recombination occurred between the marker locus and the disease locus.

linkage example (cont.)

Suppose the unaffected parent in the second mating had genotype A_1, A_1.

Then the marker genotype of the affected offspring will be either A_1, A_2 or A_1, A_1.

In the former case, the affected parent transmitted the disease allele and the allele A_2 hence there was a recombination.

In the latter case, the affected parent transmitted the disease allele and the allele A_1 hence there was not a recombination.

linkage example (cont.)

Now suppose the unaffected parent in the second mating had genotype A_1, A_2.

Then the genotype of the affected offspring will be either A_1, A_1 (no recombination), A_1, A_2 (can't tell if recombination occurred) or A_2, A_2 (a recombination).

If we had many matings where we could deduce if a recombination occurred then we could use the sample proportion (i.e. the MLE) of recombinations to estimate the recombination fraction.

If we found a marker that was close to the disease locus (in terms of the recombination fraction) then we would know the location of the disease gene.

linkage example (cont.)

We could then use tools available online to find if there are any genes known to be at that location.

Typically there will be many genes (hundreds) that are “close” to most markers when we use data from related subjects.

This is because related subjects typically share large chromosomal regions so that the ability to localize a disease gene is poor.

LOD scores

Rather than estimate the recombination fraction, geneticists test the null hypothesis that the recombination fraction is $\frac{1}{2}$ against the alternative that it is less than $\frac{1}{2}$.

If the null is rejected than the 2 loci are said to be *linked*.

Rather than use the usual likelihood ratio test, geneticists have traditionally computed the *LOD score*.

LOD scores (cont.)

The LOD score is the base 10 log of the likelihood ratio, and the null is rejected if the LOD score exceeds 3.

Using the χ^2 approximation to the likelihood ratio test we can determine that if the LOD score exceeds 3, the p-value is less than 0.0002.

One way to interpret this small significance level is in terms of a Bonferroni correction: 0.05/250=0.0002.

While this may have been reasonable when there were only a few hundred markers available, the current use of millions of markers makes this less relevant.

LOD scores (cont.)

For large, multigeneration families computing the likelihood can be extremely complex.

Even for our simple example, there are configurations of marker alleles that result in noninformative matings.

Aside from the noninformative matings, the likelihood is just the binomial likelihood since recombinations are Bernoulli experiments.

Elston Stewart algorithm

The first algorithm developed for computing the likelihood for a general, simple pedigree was the Elston Stewart algorithm.

Let $y_{i,.}$ represent the phenotypic data for subject i and y the collection of all the $y_{i,.}$.

Note: here phenotypic data is everything we have measurements for, including the reported genotypes that are reported from the genotyping method.

Elston Stewart algorithm (cont.)

Let g_i be the genotype for subject i and use g to represent the collection of these.

Note: here genotypes are unobservable-so the measured genotype can deviate from the unobserved genotype.

Assuming phenotypes are conditionally independent given genotypes we can write the likelihood as follows

$$\begin{aligned} L(y) &= \sum_g p(y|g)p(g) \\ &= \sum_g \Big(\prod_i p(y_{i,.}|g_i) \Big) p(g) \end{aligned}$$

Elston Stewart algorithm (cont.)

If we let $\mathcal{G}_t$ represent the set of subjects in generation t for $t = 1, \ldots T$, then we get

$$p(g) = p(g_i \text{ for } i \in \mathcal{G}_1) \prod_{t=2}^{T} p(g_i \text{ for } i \in \mathcal{G}_t | g_i \text{ for } i \in \mathcal{G}_{t-1})$$

The computational complexity scales exponentially in the number of markers due to the sum, hence this isn't practical for contemporary applications.

Lander Green algorithm

The Lander Green algorithm scales linearly in the number of markers.

A meiosis indicator, $s_{i,j}$, for meiosis i at locus j is 0 if the allele comes from the maternal source and 1 if from the paternal source.

Note that these will be unobservable and unknown for some subjects and loci.

If we let s represent the collection of these then, as in the Elston Stewart algorithm, the phenotypes are conditionally independent given these indicators, hence

$$L(y) = \sum_s p(y|s)p(s)$$

Lander Green algorithm (cont.)

If we let

$y_{.,j}$ represent the data for the j^{th} marker,

$s_{.,j}$ represent the meiosis indicators at the j^{th} marker $j = 1, \ldots, L$

$s_{i,.}$ represent the meiosis indicators for subject i

then if the data are for known loci and are ordered along the chromosome in j we have

$$L(y) = \sum_{s} \left(\prod_{j} p(y_{.,j} | s_{.,j}) \right) \left(\prod_{i} p(s_{i,.}) \right).$$

Lander Green algorithm (cont.)

If we then assume *no interference*, i.e. that a recombination at one location doesn't impact the probability of a recombination at a nearby location, then we can use the Markov property (applied to the markers) to get

$$L(y) = \sum_s p(s_{\cdot,1}) \prod_{j=2}^{L} p(s_{\cdot,j}|s_{\cdot,j-1}) \prod_{j=1}^{L} p(y_{\cdot,j}|s_{\cdot,j}).$$

We can treat the $s_{\cdot,j}$ as values of a hidden Markov process and use the associated tools to maximize this likelihood.

MCMC approaches

Note that both of these algorithms have the form

$$L(y) = \sum_x p(y|x)p(x)$$

where x is either

the set of genotypes (for the Elston Stewart algorithm)

or the set of meiosis indicators (for the Lander Green algorithm)

MCMC approaches (cont.)

In MCMC based approaches one treats x as a latent variable and samples this from its posterior.

Then given these, one samples the model parameters (such as the recombination fraction).

If one samples loci, i.e. g, then the algorithm is referred to as an L-sampler.

If one samples the meioses then the algorithm is referred to as an M-sampler.

Current approaches combine the 2 types of moves.

The software LOKI can be used to conduct these computations

Tree based methods

There are also sparse binary tree methods for computing the likelihood.

The basic idea is that many of the terms in the Lander Green algorithm are just zero (since many genotypes are impossible due to constraints across generations), hence if we carefully keep track of these zero terms we can greatly speed up the computation.

The software MERLIN can be used to do these computations

Nonparametric linkage analysis

In genetics, nonparametric linkage analysis refers to methods that don't make any assumptions regarding the genetic mode of transmission (e.g. dominance).

Central to these methods is the idea of *identity by descent* (IBD).

Consider 2 siblings and the alleles they have at some locus.

If both siblings receive the same alleles from both parents, then those siblings are said to share 2 alleles identical by descent.

Other possible values for the number of alleles IBD are 0 and 1.

IBD example

Consider some locus, and suppose both parents are heterozygous with completely different alleles: denote the alleles of the mother as A_1, A_2 and denote the alleles the father has by A_3, A_4.

Suppose one offspring receives A_1 from the mother and A_3 from the father.

Suppose the other offspring receives A_1 from the mother but A_4 from the father.

In this case, the siblings share 1 allele and both received this from the mother, so they have 1 allele IBD.

IBS example

Now suppose the mother has genotype A_1, A_2 while the father has genotype A_1, A_3

Suppose one offspring receives A_1 from the mother and A_3 from the father.

Suppose the other offspring receives A_2 from the mother but A_1 from the father.

In this case, the siblings share 1 allele but received this allele from different parents-in this case we say the siblings share 0 alleles IBD but 1 allele identical by state (IBS).

Nonparametric linkage analysis (cont.)

Suppose we have a set of sibships that are all affecteds.

If we can find a region of the genome with extensive allele sharing among all pairs of siblings then we would be inclined to think that this shared region contains the gene(s) that lead to the disorder.

To this end, we need to know the probability distribution of the number of alleles shared IBD by an affected sibpair for a locus unrelated to the disorder.

If we let $z_i = P(i \text{ alleles are IBD})$ then it is not hard to determine that

$$z_0 = 1/4 \quad z_1 = 1/2 \quad z_2 = 1/4$$

Nonparametric linkage analysis (cont.)

Methods that utilize IBD sharing then look for departures from these probabilities at a set of loci.

A variety of tests of this form exist: e.g. the mean test.

Some approaches examine the extent of IBD sharing over intervals of the genome-this is called interval mapping.

This is most natural in the context of studying quantitative trait loci (QTLs), which are loci that influence a quantitative trait.

References

A “classic” text for much of this material is

Analysis of Human Genetic Linkage, Ott

and a nice, challenging book is

Mathematical and Statistical Methods for Genetic Analysis.

QTLs

Polygenes are genes that work in a coordinated fashion to influence some trait.

Such genes each have a small impact on the quantitative trait, but there is some evidence that they are in spatial proximity on the genome (e.g. fruit flies and bristle number).

Another advantage of interval mapping is that the distance of a marker to a QTL is confounded with the effect size of the QTL on the quantitative trait.

Finally, there is some evidence that such an approach has greater power than a marker by marker approach.

Haseman Elston regression

Haseman Elston regression is a simple method for testing if a quantitative trait is linked to a set of markers.

Suppose we measure some quantitative trait for all members of a set of sibships, denote the pairs (y_{1i}, y_{2i}).

Also suppose we have a measure of the proportion of alleles shared IBD at some locus (or in some interval) for each pair, x_i.

If we regress $(y_{1i} - y_{2i})^2$ on x_i then it transpires that the usual hypothesis test of no association tests if the marker is near a QTL.

Variance components methods

Haseman Elston regression has been generalized to deal with general pedigrees for extended families.

If we let y_{ij} denote the outcome variable for subject i in family j then we suppose

$$y_{ij} = \mu + \alpha_j + \alpha_j^* + \epsilon_{ij},$$

where
α_j represents the additive genetic contribution for family j from the interval being examined
α_j^* represents the contribution from QTLs outside this interval
and ϵ_{ij} are iid errors (usually assumed to be normal).

Variance components methods (cont.)

We can then write the covariance between 2 measurements from subjects in the same family as

$$\mathrm{Cov}(y_{ij}, y_{i'j}) = R_{ii'}(j)\sigma^2_\alpha + 2\Theta_{ii'}(j)\sigma^2_{\alpha^*},$$

where
$R_{ii'}(j)$=the proportion of the chromosomal region under investigation shared IBD between individuals i and i' in family j

$2\Theta_{ii'}(j)$=the coefficient of coancestry between these same 2 individuals-a number that indicates how much genetic material would be shared between these 2 individuals at random based on relatedness.

Variance components methods (cont.)

Since the covariance matrix is a function of σ^2_α we can test the hypothesis that $\sigma^2_\alpha = 0$ using the usual likelihood ratio test.

Advantages:

1. easy to compute the likelihood-however you still need $R_{ii'}(j)$.
2. easy to extend to allow for covariates-just have the mean function depend on them.
3. it's "nonparametric" in that no mode of transmission is assumed, although it's parametric in that normality is assumed.

Nonparametric linkage analysis (cont.)

On the other hand, for some study designs (sib-pair data) variance components methods have inflated type I error rates.

If the data arise from the extreme sib-pair design then normality is unlikely (due to bimodality).

An alternative, proposed by Sham et al. (2002), in which IBD sharing is regressed on the squared difference (and squared sum) of the trait values seems to have comparable power and correct type I error rates.

This was based on a recent large simulation study of the most popular QTL identification methods (Kleensang (2010)).

References

Some standard texts on QTLs are

Genetics and Analysis of Quantitative Traits, Lynch and Walsh

Statistical Genomics, Liu

Genetic associations

So far, we've focused on related subjects (families and affected sib pairs), but in fact all humans are related at some level.

This fact manifests itself in the genome via a phenomenon called *linkage disequilibrium* (LD).

Linkage disequilibrium exists when 2 (or more) alleles tend to be present in the same haplotype more than we expect by chance.

It arises due to recombination: as sequences of meioses go over the generations, ancestral chromosomes are continually cut up and reassembled due to recombination.

Genetic associations (cont.)

Within a family we noted that recombination will tend to produce shared large pieces of chromosomes-this is a problem for linkage analysis since 10-20 Mb (i.e. megabase) sections will be found.

When we look at unrelated subjects, much shorter segments of chromosomes will be retained since many more generations of meioses will separate these 2 unrelated individuals.

This phenomenon used to be exploited for the purpose of “fine mapping”: after linkage is detected using family data, unrelated cases and controls would be used to test for an association between the presence of a particular allele and case status.

Genetic associations (cont.)

Such a test could just be conducted using Pearson's χ^2 for contingency tables.

Note that in linkage analysis, if a marker is linked to a disease allele then subjects in that family with the disease would tend to have the same allele at the marker, but different families could have different marker alleles that are transmitted with the disease allele.

In contrast, in association studies, we expect to find the same allele at the marker locus that is associated with the disease allele.

Admixture and confounding

Unfortunately, genetic associations can arise for reasons other than linkage between the 2 loci in question.

This can arise due to the existence of subpopulations that are genetically similar in a heterogeneous population (admixture).

For example, African-Americans are at greater risk for sickle cell anemia and have alleles at marker loci that have different frequencies than the general US population.

Admixture and confounding (cont.)

Hence if one tests for an association between having sickle cell anemia and the presence of a certain allele at some marker, one will tend to find such an association due to the confounding effects of ethnicity.

This has led to several different strategies for attempting to correct for this source of bias.

Family based controls

The idea behind family based controls was originally developed for the *trio* study design in which one has access to a collection of parents and their affected offspring.

In this context, the family based control genotype is the pair of alleles that were not transmitted to the affected offspring.

As an example, suppose the alleles at some marker locus for one parent are A_1, A_2 and for the other parent they are A_3, A_4.

If the genotype of the affected offspring is A_1, A_3 then that is the case genotype and the control genotype is just A_2, A_4.

Family based controls (cont.)

One can then make the following table and use McNemar's χ^2 test to test for an association.

	control A_1	control $\overline{A}_1$	total
case A_1	a	b	w
case $\overline{A}_1$	c	d	x
total	y	z	n

This is known as the transmission disequilibrium test (TDT) and it has been extended to general pedigrees and quantitative traits.

The family based association test (FBAT) is a generalization of the TDT and many of its extensions (Horvath, Xu, Laird, 2001).

Accounting for population structure

If one is to use unrelated subjects, there are a number of approaches available.

One approach, called genomic control, attempts to correct the null distribution of the test statistic for genetic relatedness.

The other approach entails attempting to find clusters of subjects based on a large collection of SNPs.

Accounting for population structure (cont.)

Once clusters are found, these are accounted for in the test for an association either by:

including the indicators in the linear model along with genotype

or by “correcting” the phenotypic data for these variables by regressing on the indicators then using the residuals to test for association.

These methods assume that none of the markers are associated with the disease and are all in linkage equilibrium-some recent work has attempted to deal with linkage disequilibrium assumption (Zou et al. 2010).

GWAS

A genome wide association study (GWAS) is a study design in which

many unrelated cases and controls are recruited into a study

each subject is genotyped for hundreds of thousands (or millions) of SNPs.

These have become the most popular tool for discovering genes that alter risk for complex diseases because

1. patient recruitment is easy
2. they are perhaps more powerful than linkage studies
3. they give more precise estimates of location.

GWAS (cont.)

However there are a number of problems:

1. test multiplicity
2. confounding due to ethnicity
3. some SNPs are in complete linkage disequilibrium (see 1).

Much work has been done on these problems.

As for the last point: HAPMAP.

GWAS (cont.)

Common variants versus rare variants.

In GWAS, only SNPs that are found above some threshold of frequency (e.g. 1%) in the population are tested for an association to a trait.

This is based on the “common trait-common variant” hypothesis: that common diseases are attributable to common variants.

Unfortunately, in most GWAS, SNPs with only a very modest effect have been detected-there is a problem of “missing heritability”.

GWAS (cont.)

For example, in a recent study a genetic variant that alters risk for obesity had an odds ratio of 1.2.

Compared to a risk factor like “low education”, with an odds ratio of 3.8, there is a sense of disappointment.

This has led many to abandon the “common trait-common variant” hypothesis and recently researchers have been looking for rare variants that alter disease risk.

GWAS (cont.)

New technology has allowed more systematic investigation of the impact of rare variants (i.e. next generation sequencing, more later).

Some work has been done on using SNP arrays to construct haplotypes which are rare variants (e.g. the algorithm DASH).

GWAS (cont.)

Nonetheless, with many existing GWAS data sets available, many questions about optimal analysis of GWAS data sets are still of interest.

For instance, should we really test every SNP independently of the rest?

Should we group them by gene, or maybe by pathways the genes belong to?

GWAS (cont.)

If we let x_{ij} denote the SNP allele for subject i at marker j and y_i is the phenotype for subject i, then testing each SNP independently implies the following model

$$\mathrm{E}y_i = \beta_0 + \beta_j x_{ij}$$

for all j.

This will lead to a loss of power due to the need to correct for the large number of tests.

GWAS (cont.)

However, specifying models for the relationship between a trait and multiple loci typically involves models with many parameters-leads to tests with high degrees of freedom (and loss of power).

Here

$$\mathrm{E}y_i = \beta_0 + \sum_{j=1}^{L} \beta_j x_{ij}$$

so the test of no association would have L degrees of freedom.

GWAS (cont.)

As an alternative to these approaches, some have assumed that there is a single parameter that determines the strength of the association between all SNPs and the trait.

$$\mathrm{E}y_i = \beta_0 + \beta_c \sum_{j=1}^{L} x_{ij}$$

So here the test of $\mathrm{H}_0 : \beta_c = 0$ just has one degree of freedom.

GWAS (cont.)

Problem: some SNPs may be positively associated and some negatively associated with the phenotype.

Hence some have proposed “flipping” the coding of the x_{ij} prior to testing so that the direction of the marginal associations is the same for all SNPs.

Other tests have been constructed by squaring the individual effect estimates before combining them into one test statistic.

There doesn’t appear to be a uniformly best test among the ones that have been investigated.

Test multiplicity

Controlling the FWER makes for some very weak tests: if we use the Bonferroni correction we will have little power to detect associations.

As an alternative, many have investigated controlling the false discovery rate (FDR) (or more recently, the false discovery proportion, FDP).

FDP has become of more interest since we are interested in controlling the number of false positives for a given experiment, not the average FDP.

Test multiplicity (cont.)

Consider the following table:

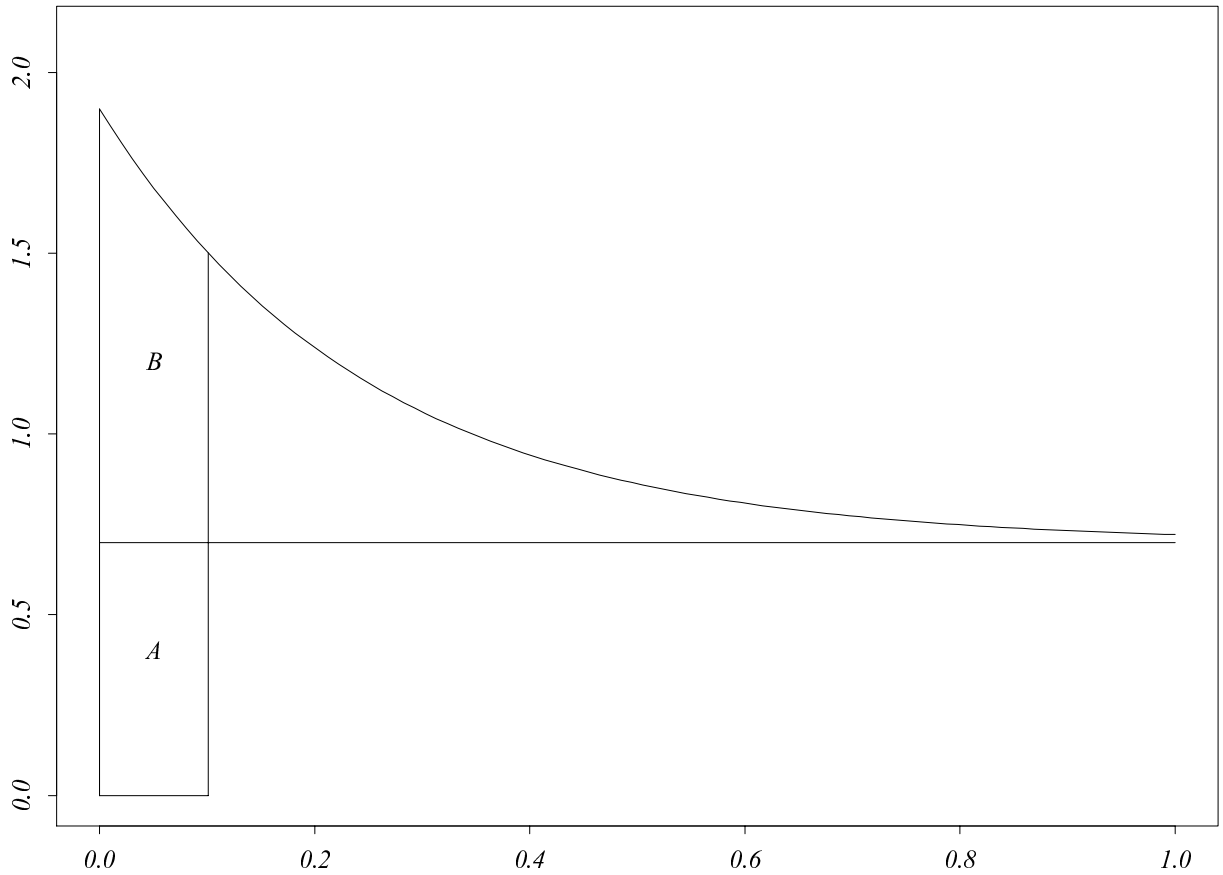
	number not rejected	number rejected	
true null	U	V	L_0
false null	T	S	L_1
total	$p-R$	R	L

If
$V(t)$ is the number of true null p-values less than t
$S(t)$ is the number of false null p-values less than t
$R(t)$ is the number of p-values less than t

then the $\mathrm{FDP}(t) = V(t)/R(t)$ and $\mathrm{FDR}(t) = \mathrm{E}[V(t)/R(t)]$

Test multiplicity (cont.)

Intuitively, the FDR can be thought of as $A/(A+B)$:



Test multiplicity (cont.)

Early methods for estimating the FDR assumed that all tests were independent-this is clearly not the case for nearby SNPs in LD.

A number of papers have focused on examining how much correlation (or what type) existing methods of FDR estimation can handle.

More recently the focus has shifted to approximating the FDP as $L_0 \to \infty$ and estimating this approximation.

That is, current approaches assume that only a few SNPs will be associated with the disease.

Test multiplicity (cont.)

Here the basic idea is if Z_i represents one of the test statistics $i = 1 \ldots, L$ then it is assumed

$$(Z_1, \ldots, Z_L) \sim \mathrm{N}((\mu_1, \ldots, \mu_L), \Sigma)$$

for some Σ.

One then uses the spectral decomposition of Σ to account for dependence among tests.

Phenotype definitions and heterogeneity

Traditionally disease states have been distinguished based on physical symptoms and simple assays that are widely used (e.g. lipid profiles).

From the perspective of clinical management, disease states should be distinguished based on effective therapeutic strategies not pathogenesis.

However it is widely suspected that the disease states that are currently distinguished based on treatment strategies are not sufficiently fine.

Phenotype definitions and heterogeneity (cont.)

With more carefully defined phenotypes we can more accurately distinguish between disease states.

This has led to the use of more sophisticated assays that measure certain molecular characteristics and attempts to find linkage with these molecular characteristics: e.g. e-QTLs.

As these molecular characteristics are frequently measured in a high throughput context, test multiplicity becomes a severe issue.

Phenotype definitions and heterogeneity (cont.)

Related to poor phenotype definition is the notion of *heterogeneity*: the idea that there are different genetic sources for the same phenomenon.

Mixture models and latent class models have been employed to investigate linkage and association in this context.

Here is an example of linkage of a QTL to a heterogeneous and poorly defined phenotype (here, asthma) using mixture models.

Phenotype definitions and heterogeneity (cont.)

Assume the distribution for those with the disease in family j has mean μ_j,

further assume that these μ_j arose from another mixture model, reflecting the different genetic sources of the disorder in the population.

let y_{ij} represents the p-vector of quantitative variables for subject i in family j

x_{ij} is a set of covariates for this subject indicating if the subject is likely affected-a poorly defined phenotype

β is a vector of regression coefficients giving the effect of each element of x_{ij}

Phenotype definitions and heterogeneity (cont.)

let $f_j(\mu_j)$, $h_k(\eta_k)$ and g_j be densities on a Euclidean space of dimension p with means μ_j and η_k for f_j and h_k.

let λ_{ij} be an indicator variable which is one when subject i in family j is affected

then we can write the model in the following form:

$$\begin{aligned} y_{ij} &\sim \lambda_{ij} f_j(\mu_j) + (1 - \lambda_{ij}) g_j \\ \mu_j &\sim \sum_{k=1}^{K} \phi_k h_k(\eta_k) \\ \lambda_{ij} &\sim \mathrm{Ber}(x'_{ij}\beta). \end{aligned}$$

Phenotype definitions and heterogeneity (cont.)

Given this framework, we define our quantitative variable as

$$z_{ij} = \| \, y_{ij} - \eta_k \, \|_2,$$

where k is the index of the mixture component to which family j belongs.

Estimation can be conducted using clustering tools if we estimate the mean of a cluster with the centroid of its elements.

This approach detected a novel locus that increases risk of developing asthma in a collection of 27 large pedigrees.

Gene-gene Interactions

While the use of linear models for testing for interactions between markers is straightforward in principal:

huge number of possible interactions (even just pairwise) in contemporary GWAS-computing is prohibitive.

use of many predictors in regression leads to large standard errors

penalty for multiple hypothesis testing is becomes severe.

Gene-gene Interactions (cont.)

There have been multiple reports of GWAS with SNPs that don’t have significant marginal main effects but do have significant interactions.

This likely arises due to genes working in pathways: if 2 genes must have normal alleles for some biological process then one would expect interactions between SNPs in LD with these genes.

Gene-gene Interactions (cont.)

Some of these interactions have been high order:

Ritchie et al. (2001) reported an interaction between 4 polymorphisms in 3 different genes in a case control breast cancer study.

Moreover these 3 genes are known to be regulators of the catechol-estrogen pathway.

Parallel computing strategies have been employed to overcome the computational burden due to the large number of tests in a parametric context.

Gene-gene Interactions (cont.)

Dimension reduction strategies have been commonly employed to deal with the explosion in the number of tests.

The use of penalized regression approaches (e.g. the lasso) has been less popular due to the potential complexity of gene-gene interactions.

One popular dimension reduction technique is called multifactor dimensionality reduction (MDR).

For simplicity, suppose we have case control data with equal numbers of subjects in each group.

Gene-gene Interactions (cont.)

MDR is an algorithm that consists of 4 steps:

1. select a group of polymorphisms

2. create a table based on these polymorphisms and compute the ratio of cases to controls in each cell

3. if the ratio in a cell exceeds some threshold, then label that cell “high risk” and label other cells “low risk”-so now you have a univariate summary based on the set of polymorphisms

4. Use cross validation to compute prediction errors and select the model with the lowest error rate.

Gene-gene Interactions (cont.)

When there are many polymorphisms the table will become high dimensional and sparse, thus MDR has seen refinements to deal with this.

Other approaches have been based on comparing genetic similarity at a set of markers to trait similarity.

Much of this discussion carries over to gene-environment interactions, except in that context the “environment” variables may be continuous (so MDR would need to discretize such variables).

Next generation sequencing

Over the last several years next generation sequencing (NGS) has started to become widespread: many areas of current interest.

In NGS, a DNA molecule is cut into many short (25-150, depending on the manufacturer) segments, these segments are then sequenced, and then assembled to obtain the original DNA sequence.

This is similar to strategies that were used to sequence the human genome, except now we have the assembled human genome to use as a reference for assembly.

NGS (cont.)

Currently it costs $5,000-$10,000 to generate a human genome and it only takes a few days.

The 1000 genomes project is currently sequencing the genomes of many humans.

Much effort thus far has gone towards alignment of the short sequences that these technologies report to a reference genome.

These methods use the same sorts of approaches that have been used for years to align nucleotide or amino acid sequences to each other (i.e. BLAST type searches).

NGS (cont.)

There are many potential applications of NGS, only one of which is identifying polymorphisms, such as:

1. mapping new genomes
2. identifying new transcripts
3. RNA sequencing (RNA-seq)
4. identification of transcription factor binding sites (CHiP-seq, i.e. chromatin immunoprecipitation instead of CHip-chip)
5. methylation analysis

NGS (cont.)

But even in the context of just studying polymorphisms, NGS will allow us to examine more than just newly identified SNPs: also small indels (i.e. insertions and deletions) and large structural differences.

A challenge of these data sets is that since we are looking for rare variants, some grouping of these variants must be conducted.

NGS (cont.)

A common approach is to define a variable that just indicates if any rare variant occurs in some functional unit (e.g. a gene or a part of a gene)-*burden tests* (e.g. the cohort allelic sum test (CAST)).

Alternatively one can also use information on the number of rare alleles in a region (e.g. the weighted sum test (WST)).

Drawbacks of these methods are:

not all variants will be functional

some could have effects that go in different directions.

NGS (cont.)

The C-alpha test compares the expected variance to the actual variance of the distribution of allele frequencies, so it avoids the perils of burden tests.

SKAT is another recent method that avoids the pitfalls of burden tests-it uses a random effects model with random effects for each rare variant in a region and tests if the variance component is zero.