Introduction to Quantitative Genomics / Genetics

BTRY 7210: Topics in Quantitative Genomics and Genetics

September 10, 2008

Jason G. Mezey
Outline

• History and Intuition.
• Statistical Framework.
• Current Approaches.
• Current Challenges.
History

- **Quantitative Genomics / Genetics** may be loosely defined as the field concerned with the statistical modeling of the genetics of 'complex' phenotypes.

- Relevant history:
  - 1900-1980: statistical analysis of the patterns of inheritance (i.e. the resemblance between relatives).
  - 1980-2002: mapping (= identification) of the genetic loci responsible for most Mendelian diseases (e.g. diseases where alleles at a 'single' genetic locus determines disease).
  - 2002-present: 'age of genomics' first convincing mapping of genetic loci for complex traits (i.e. cases where genotype cannot be inferred directly from the phenotype).

- Techniques come from three relatively distinct fields: evolutionary biology, agricultural sciences, medical sciences.

- All three share a common goal: mapping genetic loci.
• The genome of an individual is an important determinant of individual phenotype.
• As a consequence, specific polymorphisms (SNPs, INDELs, transposable elements, chromosome aberations, etc.) can produce differences among individuals in a population.
• This is because different states of a polymorphism can result in differences in the biological processes that produce a phenotype, i.e. such polymorphisms are causal.
• The goal of quantitative genomics is to identify such causal polymorphisms, the genetic locus in which they are present, or their general genomic location.
Populations

- We analyze a population (more specifically a sample) of individuals to map causal polymorphisms or genetic loci.
- At the minimum, data for mapping includes: measurements of phenotypes (= traits) and genotypes (= set of genomic positions where individuals have different polymorphism ‘states’ or genetic loci where there are different ‘alleles’).
- Since genomes are organized into chromosomes, genotypes in close physical proximity will be correlated. This is called Linkage Disequilibrium and it means we need not have measured the causal polymorphism to map a genetic locus (e.g. polymorphisms analyzed are ‘genetic markers’).
- Currently, the data for mapping experiments typically include both phenotypic measurements and genotypes (> 500K SNPs) for 100’s to 10K+ individuals.
- The challenge: using these data to identify individual causal polymorphisms or genetic loci.
The Statistical Model

Our objective is to draw conclusions about a population from a sample.

The sample space in this case is:

$$\Omega = \{\phi \times \kappa\}$$

where $\phi$ is the set of possible phenotypes and $\kappa$ is the set of possible genomes.

On this sample space we define a probability function $p(\Omega)$ and the random variables $Y(\phi)$ and $X(\kappa)$ where:

$Y$ maps to a $k$-dimensional vector $Y$ of phenotypic measurements (each of which may be continuous or discrete).

$X$ maps to an $m$-dimensional vector $X$ where each element is a dummy variable representing the genotype at genomic position $j$ (usually taking values -1, 0, 1).
The Statistical Model

We assume that $Y$ and a causal subset $q < m$ of the genotypes $X$ are related by the following:

$$Y = g(X) + \epsilon$$

where $g(X) = E(Y|X)$ and $\epsilon$ is a random variable which accounts for the difference of the phenotypic vector of an individual $i$ from the expected value given the genotype.

For the purposes of quantitative genomic inference it is convenient to use Generalized Linear Models (GLM) to represent this relationship. These have the following properties (for a single phenotype $Y$):

1. A random component $\epsilon$ of the variable $Y$ has a distribution in the exponential family.
2. A link function relates the random vector $X$ and parameter vector $\beta$ to $Y$: $E(Y|X) = \gamma^{-1}(X\beta)$.
3. The variance of $Y|X$ is a function of $E(Y|X)$. 

A Recognizable Example...

Assume the following for the GLM:

1. The random component $\epsilon$ follows a normal distribution with mean zero and unknown variance $\sigma^2_\epsilon$.

2. The link function is the identity function: $\gamma^{-1}(X\beta) = X\beta$

3. The variance of $Y \mid X$ is a constant: $\mathbb{V}(Y \mid X) = \sigma^2_\epsilon$.

In this case, the GLM is the simple linear regression model, which can be written (for a single phenotype and a single polymorphism $X$) as follows:

$$Y = \mu + \beta X + \epsilon$$

$$\epsilon \sim N(0, \sigma^2_\epsilon)$$
Our goal is inference concerning GLM parameters using a sample.

For our purposes, we are interested in estimation and testing hypotheses concerning GLM parameters (generally the \( \beta \)'s).

There are two broad inference approaches:

- **Frequentist:** do not assume parameters are random variables.
  - Example, Maximum Likelihood Estimation: 
    \[
    \hat{\theta}_{MLE} = \sup_{\theta \in \Theta} L(\theta | Y).
    \]
  - Example, Likelihood Ratio Tests: \( \Lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta | Y)}{\sup_{\theta \in \Theta} L(\theta | Y)} \).

- **Bayesian:** \( p(\theta | Y) \propto p(Y | \theta)p(\theta) \).
  - Example, estimation using median of the posterior \( p(\theta | Y) \).
  - Example, we can ‘test’ using Bayes factor or a credible interval of the posterior.
Quantitative Genomic Inference

- The most basic problem in quantitative genomics is determining which of the \( q < m \) genotypes in \( \mathbf{X} \) are in linkage disequilibrium with polymorphisms which have causal effects on a phenotype \( \mathbf{Y} \).

- Intuitively, these are cases where experimentally substituting one allele for another would produce a change in the expected phenotype:

  \[
  A_i A_j \overset{ij \neq kl}{\rightarrow} A_k A_l \Rightarrow \Delta E(Y)
  \]

- The simplest approach for identifying such cases is to fit a GLM for each genotype individually.
Quantitative Genomic Inference

- For example, if we have measured a normally distributed phenotype, such as human height, for \( m \) genotypes we can fit \( m \) simple linear regression models of the form:

\[
Y = \mu + \beta X + \epsilon
\]

and perform the following hypothesis test in each case:

\[
H_0 : \beta = 0 \quad H_A : \beta \neq 0
\]

- For runs of genotypes in linkage disequilibrium where we reject \( H_0 \), we consider this reasonable evidence that we have mapped the causal polymorphism to a physical location in the genome.

- This approach of applying individual tests for every genotype is currently the most commonly applied approach for mapping loci.
Alternative Individual Genotype Tests

- The GLM is the quantitative genetic model that is the foundation of quantitative genomics / genetic analysis (mapping loci, additive genetic variance, etc.).
- However (intuitively) the parameterized GLM approaches maps by testing for differences among means of groups in a sample partitioned according to genotype (continuous traits) or differences in the frequency of genotypes in (discrete) trait categories.
- This means that any statistical approach which tests for such differences may be used for mapping genetic loci on a marker by marker basis.
- An incomplete list of common approaches:
  - Continuous traits: Parametric (GLMs, ANOVAs, t-tests), Non-Parametric (Kruskal-Wallis, permutation-based).
  - Discrete: GLM, $\chi^2$, Cochran-Armitage, Fisher’s Exact.
  - Pedigree Based: Transmission-Disequilibrium Test (TDT), sib-pair test, QFAM.
Experimental Mapping Designs

Association Analysis / Linkage Disequilibrium Mapping refers to designs where the individuals in the sample are not highly related.

Linkage Analysis / Perdigree Analysis refer to designs where individuals in the sample are highly related and the relationships are generally known.

Controlled Breeding Designs / QTL Analysis (F2, RILs, NILs) refer to cases where the relationship among individuals are both known and controlled.
Challenges: Multiple Tests

- The problem: If we perform tests for 500K markers this is a severe multiple testing problem, i.e. we expect significant results by chance.

- Question: How to control Type I error (false-positives) without sacrificing power? The problem is made more difficult because our tests are correlated.

- A few approaches: False Discovery Rates (FDR), Bonferroni, in combination with Principal Component Analysis (PCA), Permutation, Hidden Markov Model (HMM) approaches.
Challenges: Haplotypes

- A haplotype is a combination of alleles transmitted together. Tests based on haplotypes (instead of individual polymorphisms) can sometimes be more powerful.

- Questions: How to infer haplotype structures that produce the ‘best’ tests?

- A few approaches: phasing using unrelated individuals and using family structure, haplotype testing in combination with linkage analysis.
Challenges: Population Structure

- The problem: If two populations differ in their mean value or frequency of a discrete phenotype (disease) then every genotype that varies between them will produce a positive result.

- Question: how to identify cryptic population structure and correct for it when testing?

- A few approaches: STRUCTURE, PCA, incorporating population structure as a co-factor in GLM, tests which are robust to structure such as TDT.
Challenges: Shared Ancestry

- The problem: If members of the population are related this can lead to reduction of power.

- Question: How to estimate ancestry and incorporate this into our models?

- A few approaches: Pedigree based estimation, haplotype based estimates, Mixed models with co-ancestry factor, direct modeling via pedigrees.
Challenges: Multi-Locus Approaches

- The problem: If there are more than one contributing genotype, this can lead to less power and to false positives when using individual marker testing approaches.

- Question: How to fit a model with $p$ parameters when sample size $N$ is small without over-fitting (the large $p$, small $N$ problem)?

- A few approaches: Bayesian hierarchical models, algorithmic searches and model selection (AIC, DIC, penalized likelihood).
Challenges: Epistasis

- The problem: The complete genetic model includes interactions among polymorphisms, in fact there are $3^q$ possible parameters.

- Question: How to identify pair-wise effects when power is low? How to account for these effects overall, i.e. treating them as nuisance parameters?

- A few approaches: exhaustive pairwise tests, hierarchical Bayesian modeling, Kernel approaches.
Other Challenges

- Genetic architecture (Common Disease-Common Variant, Many Rare Variants) and determining the most powerful approaches for different architectures.
- Imputation and testing approaches for missing genotypes.
- Experimental designs for controlled breeding approaches.
- Computationally efficient approaches for linkage analysis.
- Coalescent based approaches.
- etc.
I will lead discussion for the following paper:

*Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies;* Hoggart et al. 2008; (PLOS Genetics).

This will include a quick introduction to the basics of MLE search algorithms and Bayesian MCMC approaches.