CS 262: Human Population Genomics

Maeva Fincker- March 10th 2015

Heritability & Environment: the missing heritability problem

- We want to quantify how much the genotype affect the phenotype, and to predict phenotype from genotype.
- Scientists have been comparing different traits in twins and less related persons raised together or in a different environment in order to identify which traits are more influence by the environment or by the genotype.
- They are many traits, such as height, that are highly heritable. However, we still cannot predict height from genotype only.
- More complex traits (diseases, diabetes, mental illnesses) are even more difficult to predict. Genes only explain a small % of the disease occurrence.

Missing heritability problem: individual genes cannot account for much of the heritability of diseases, behaviors, and other phenotypes.

Graph:

- shows the effect size of an allele (how much more likely you are to have a trait, given the allele) varying with the allele frequency.
  - common allele: around 5% of the population have the allele
  - rare allele: down to a single person having the allele
- *Genome-Wide Association Study* (GWAS): compare alleles between a cohort of a ~1000 people exhibiting a trait and a control group of ~1000 people.
  - The idea is to correlate the occurrence of certain alleles with the occurrence of a trait/ disease. GWAS use genotyping arrays to look at SNPs.
  - GWAS generally sample alleles with high frequencies (>1% of the population).
  - GWAS associate small effect size alleles with common ones, which is expected: if an allele is likely to give you a disease, it will not be selected and become common.

The key to the missing heritability problem is probably to be found in the low frequency alleles with medium effect size.

Global ancestry inference i.e. finding out the geographical origin of a group of individuals

- Find SNPs using genotyping array to map common alleles:
  - genotyping arrays use DNA probes on the chip that are complimentary to the SNPs of interest in the genome.
  - an array can map around 10ˆ6 SNPs (for 23&me for example)
- An individual is then described by a vector or 10ˆ6 SNPs.

- Run the genotyping of a 1000 individuals and give these 1000 vectors to the computer without specifying the geographical origin of these people.
- Using PCA on the SNPs vectors, the authors identified the first 2 components of highest variance, scaled and rotated them.
- By projecting each individual on the 2 PCA axes, the authors were able to recover the location of each individual (within 300km of their place of birth).
- This works in Europe. It doesn't work in places like the US (too much mixed ancestry, and mix of ancestry is not uniform across the genome) or Africa (much higher variation).

Modeling haplotype with VLMC (Variable Length Markov Chain Linkage Model)

- See Beagle program
- Goal: get a compact model of the haplotype of a given group of individuals
- Given a table of occurrences of haplotypes, one way to represent this table is a tree, where the number of leaves is the number of observed haplotypes.
- Compression of the tree in linear time:
    - for any given node, the program finds nodes that are similar enough (i.e. children trees are identical with respect to branch number and proportion)
    - ex: nodes 3.1 and 3.3
    - The compression of the tree leaves a network / graph from which the linkage disequilibrium can be computed.

Why is it useful ?

Phasing:
Given a set of individuals G_1, … G_N where each $G_i = H_1 + H_2$ (alleles coming from the 2 parents), how do you determine the two haplotypes of origin for each individual ?
We infer the individual genotype by going through the states of the graph using DP, and infer the pair of paths that maximize the probability of getting the right genotype.
Locally accurate, switch state error globally

Identity by descent (IBD)
Tracing the origin of pieces of DNA that you share with your relatives.
Same idea as phasing using the Beagle graph: each individual is modeled as a pair of paths in the graph.
Individuals are identical by descent when some of their paths are the same.

Mexican ancestry

- Very complex ancestry with a population with mixed origin: Europe and indigenous populations.
- Understanding ancestry and geographical distribution of ancestry in Mexico requires combining PCA, IBD and clustering studies.

Population sequencing

- When we sequence an individual, we want enough reads (x30) to cover all alleles and get rid of sequencing errors. *Expensive: $1500 / person*
- To sequence 10^6 to 10^9 people, we need a different approach that will bring the cost down: *we use the haplotype information.*
-

Method:

- sequence each individual with a low coverage (x2 - x6)
- given a set of people $G_1, \ldots G_N$ where $G_i = g_{i1}, \ldots g_{in}$ (haplotype)
  and $P_1, \ldots, P_N$ where $P_i = [p_{ijk} = P(g_{ij} = k \mid data)]$
- we cannot tease apart error from rare allele using only the coverage of each position:

Given 100k individuals, let's sequence each at 5x coverage

==> coverage of 500,000x for each positions in the genome: 500,000 reads per position

Assuming 0.4% error rate: 2000 reads are false per position. For example, we'll get 800 C, 400 G and 800 T.If an allele occurs in 0.1% of the population, we won't be able to tease it apart from the error reads.

To be able to tell which reads is false and which has a rare allele, we are gonna use the haplotype information and linkage disequilibrium.

Maximization of the probability of having true reads by modeling linkage disequilibrium using nearest neighbors:

- If the change in a position is a real minor allele, it is likely to be in linkage disequilibrium with other alleles.
- Given a change in position i and a set of individuals where the change appears, we look for positions j that are roughly in the same set of people.
- A candidate polymorphic site is considered if it has neighbors with the same linkage disequilibrium.
- Calculate p_{ijk} and g_{ij} with the reveal algorithm, and iterate until convergence.

The algorithm finds minor alleles that are in 10 people or more give 1000 genomes.

Fixation, positive and negative selection

3 ways for an allele to propagate:

- drift, aka neutral selection: P(rare allele) = P(common allele)
- positive selection: P(rare allele) > P(common allele)
- negative selection: P(rare allele) < P(common allele)

Finding positive selection:

Different tests target selection at different timescales.

Ka/Ks test:

- one of the oldest test
- used to find very strong positive selection over very long timeframe
- for a given gene, count up the number of synonymous mutation and expected number of synonymous mutations.
- if # count / # expected < 1: strong positive selection
- this test was first used on immune response genes.