CS262 Lecture 18: Human Population Genomics Continued

Scribe: Max Drach

March 3, 2016

Motivation

This lecture we discuss more advanced methods for tracking complex heritable traits.

The multiple sequence alignment algorithms that we discussed last lecture, $MUSCLE$ and $PROBCONS$ help us analyze protein sequences and other short segments of DNA. However, because of time constraints we need other methods in order to analyze large expanses of the genome to allow us to understand traits influenced by many different parts of the genome.

Today we will discuss Heritable Complex Traits, Standard Principle Components Analysis, Variable-Length Markov Chains, and other methods to explain heritability of complex traits.

Heritable Complex Traits

Height, taste, curly hair, immunity to diseases – these are all complex traits that are highly heritable. We know this from studies that compare identical twins reared apart and unrelated individuals reared together. See Figure 1.

From these, we can tell how much is explained by genetics, how much from environment.

Let's discuss some of the traits shown in Figure 2.

Diabetes: Type 1 diabetes is highly heritable, however type 2 diabetes is not. Type 1 is caused by mutations in the genome that inhibit proper insulin regulation, while type 2 is caused by environmental stresses on the body, like obesity or lack of exercise.

Longevity: for the most part, tied to the environment. Interestingly, though, extreme longevity – living 100+ years – is highly heritable.

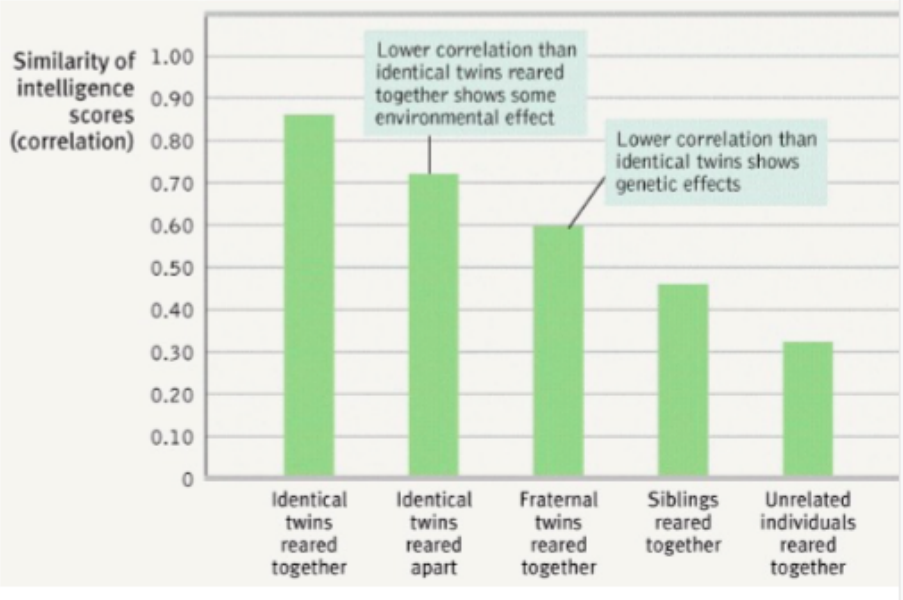

Figure 1:

Trait/Disease	Estimated heritability
Alcoholism	50-60%
Alzheimers	58-79%
Asthma	30%
Bipolar Disorder	70%
Depression	50%
Hair Curliness	85-95%
Lung Cancer	8%
Height	81%
Obesity	70%
Longetivity	26%
Sexual Orientation	60%
Schizophrenia	81%
Type 1 diabetes	88%
Type 2 diabetes	26%

Figure 2:

Height: is very heritable, as you know if you look at your parents and siblings. Shorter people have shorter parents/siblings and taller people have taller parents/siblings.

Lung cancer: is not heritable. Risk of lung cancer is almost completely due to smoking. This is because tobacco smoke is a mutagen. Researchers estimate that every 15 cigarettes causes a new a mutation to be fixed into the genome. So, people who smoke a pack a day get a new mutation each day. This also means that if you smoke at one time in your life, it can affect you later in life, because mutations are permanent. *Fun fact* Serafim used to smoke, but he stopped after college.

Skin cancer: Similar to smoking, it is largely do to environmental factors (sun exposure). Skin regenerates quickly and is prone to mutations.

Examining Complex Traits in Genome

Take extreme height as an example. Height is obviously a heritable trait, yet we find it extremely difficult to explain height with genetics. Why can't we just sequence genomes of a group of tall people and map the places in the genome that are linked to increased height (e.g. find the haplotypes)? This is because height differences are caused by a variety of factors throughout the genome, and these can't be analyzed with traditional methods. We require a Genome Wide Association Study – and there are barriers to this as explained below.

Variations in the human genome fall into four primary categories.
See Figure 3, a plot of all types of alleles.

The y-axis of the plot denotes how strongly the allele affects an expressed trait. For example, if a person is 50 times more likely to have a disease given the allele, it is high on the y-axis. The x-axis shows how common the alleles are. A gene that appears very infrequently, like fewer than 1000 people, would be far left on this axis.

The top-right corner

The top-right corner contains very common alleles that have a very high effect on traits expressed in individual. These include alleles that impact eye color, curly hair, skin color, or hairiness. These alleles often code for traits that are make a population more fit in a new environment and sweep the population really quickly.

For example, one of these alleles is skin color – if you are in sub-Saharan Africa and have very white skin, you will get cancer, and may have progeny at a 5-10% lower than other people. Even if you are only 2% less likely to have kids, that is huge effect across generations. In 50 generations – that gene would be completely

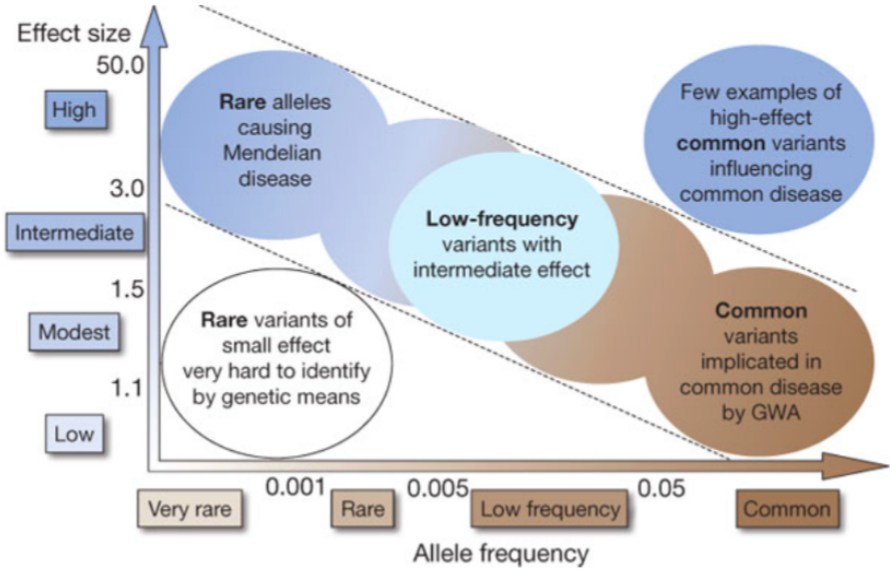

Figure 3:

wiped out. *Fun fact:* skin color can change in a population in a mere 1000 years. If you put a white population in a sunny climate, within 1000 years their skin will be dark, because so few alleles must be changed to accomplish this.

The top-left cornner

We know about alleles in the top-left corner – the rare alleles of high effect. These are rare, typically because they are harmful. These genes often code for inherited Mendelian diseases, like cystic fibrosis. If genes this corner were beneficial, they would eventually become more common, and shift right in the graph.

The bottom-left corner

On the bottom left, we have rare alleles of small effect. If an allele is rare and has small effect, most likely we have not seen it. It is extremely difficult to detect these kinds of alleles as it required a lot of data (which we do not have; gene sequencing was extremely expensive until very recently. These are also difficult to detect because techniques, like genotyping microarrays, don't necessarily capture the alleles at the few specific positions that they test.

The bottom-right corner

Finally, the bottom right corner includes alleles of little effect. Many common diseases stem from combinations of many of these low-effect alleles.

Difficulties Tracking Heritable Complex Traits

This brings us back to our discussion of height, which may be determined by:

1) **Many variants of small effect** genes in the bottom-right corner that are difficult to detect individually despite the strong impact of their combined effect.
2) **Rare variants** not captured in genotyping microarrays.
3) **Structural variants** not captured in short read sequencing.
4) **Epistatic effects** non-linear gene-gene interactions.

This brings us to the concept of a **haplotype** – a combination of alleles that code for a specific trait. We actually don't have common and rare haplotypes like we have common and rare alleles. "Rare" haplotypes are merely rare *combinations* of common alleles. If two rare variants are accumulated nearby on the genome and have an unusually low probability of being recombined, these are said to have linkage disequilibrium.

Lets say this rare varient casues type 1 diabetes (say this variant causes insulin to spike 3x when eating sugar). Let's do a **Genome-Wide Association Study (GWAS)** that probes the common allele variants of the individual's genome. What will the study show? We will discover all of the common variants linked to rare allele variants. These are not the actual alleles that cause diabetes, but common genomes that are linked to these genes. GWAS allow data from many individuals to be analyzed and collected without needing to consider the entire genome.

Shortcomings and Future Solutions

Problems: Genome-wide association studies cannot give us detailed information about traits resulting from many gene variants with a small effect. Also, because of epistatic effects, any number of alleles alone cannot tell us what the affect is, we need to know the combination of them (non-linear gene-gene interactions).

Solution: Whole genome sequencing of many individuals (more than a few thousand) can solve these problems. The price of whole-genome sequencing drops every year, and as prices drop the abundance of data increases.

Problem: Structural variants are not captured by short readings.

Solution: Alignment algorithms also continue to increase, so structural variants will become less of a problem.

Problem: Phenotypes are triggered by the environment. For example, imagine someone has a gene that makes them extremely excitable by the color red. They become rich, they get a red Ferarri, drive fast, crash, and die. This is environmental. This would not be easy to pick in genome study.

Solution: These kinds of genes are difficult to detect but better simulations could help.

Global Ancestry Inference

Standard Principal Components Analysis

Europe

Seven years ago some researchers that wrote a paper published in Nature Magazine gave talk at Stanford about this result.

This is what they did: They took several hundred Europeans and genotyped them. They created a vector for every European corresponding to each genotype that was taken (each square was 0, 1, or 2 depending on number of copies of gene).

They then projected these N-dimensional vectors based on the variance of each component in order to create a useful visualization. Specifically, they performed Principal Component Analysis, removed dimensions that had little variance, which was all except two. In other words, despite having many-dimensioned vectors, the vectors formed a disk in the N-dimensional space that they were plotted in so not much information was lost by projecting them into a two-dimensional space. This is called Standard Principal Components Analysis (PCA).

They colored each dot in this projection according to the person's country of origin that the genotype data came from. They got this plot (rotated a bit and stretched 5x).

They got a squished version of a map of Europe! The genes aligned to countries of origin surprisingly well. The dots, on average, had 300km of geographic error. After the talk, Serafim was amazed. But the results make sense. The data correlated well with the two principal axes of migration – 1 from Africa going north, and the other to the east and west. Gene distribution can tell us a lot about our history.

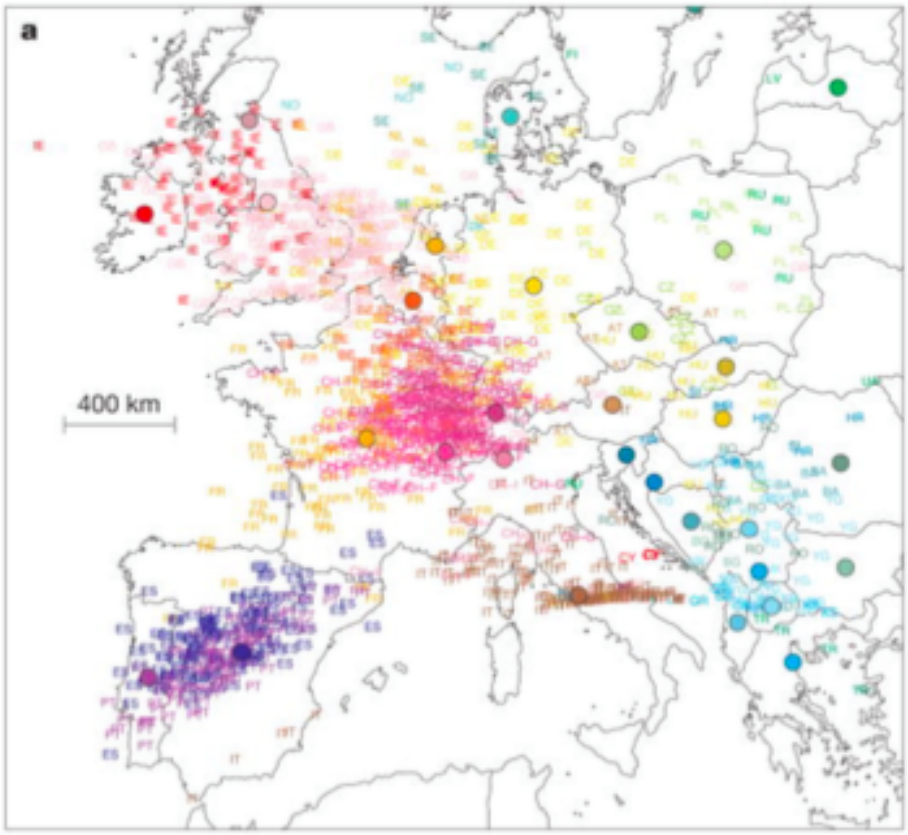

Figure 4:

This study did not translate well to studies over a wider area. When the researchers tried to include Sub-Saharan Africa in the study, it would be way off of the map, just because the genomes of people from those two locations are so different. This will only work properly in a place where people have high genetic similarity and the distance in genes is reflected by the distance in geography.

Mexico

Inspired by the results of the European study, Stanford researchers performed a similar analysis on different indigenous groups in Mexico. They found that these groups roughly correlated with a map of Mexico. They also found that divergent indigenous groups in Mexico differed in DNA as much as Europeans and East Asians.

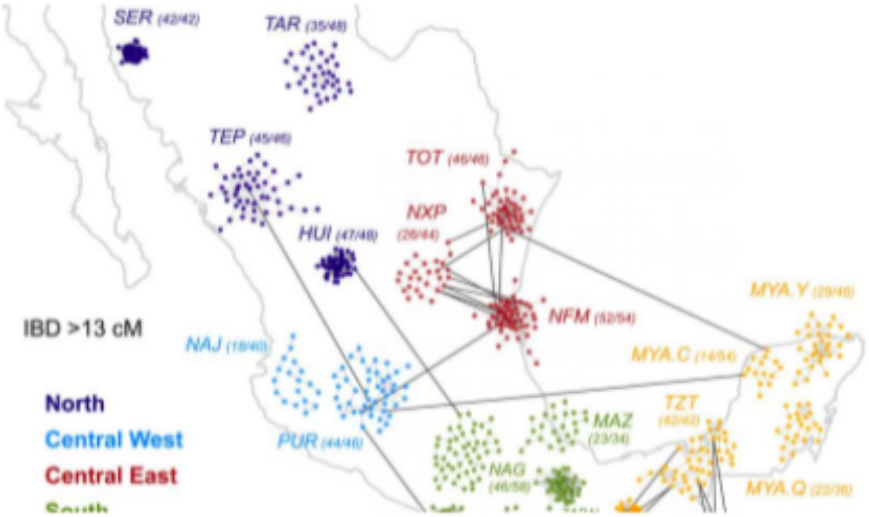

Figure 5:

It is amazing how much we can learn about history by just looking at genetics.

Variable-Length Markov Chain (VLMC)

A Variable-Length Markov Chain is a common way of modeling variation in a population.

We see a number of different haplotypes in Figure 6.

And we know how many individuals have each of the haplotypes. This represents the variation in a population. One way to write this table succinctly is create a

Haplotype	Count		
	Total	Case	Control
0000	21	12	9
0001	79	43	36
0011	95	43	52
0110	116	59	57
1000	25	14	11
1001	112	60	52
1011	152	69	83

A

Figure 6:

tree with two splits at each node, into two child nodes, one corresponding to a 1 in the next position of the haplotype, the second corresponding to a 0. There is a different leaf node for every haplotype. See Figure 7.

The way we compress this is to find nodes whose sub-trees look similar (for example the node 3.1 at the top of the graph has roughly half have 0 (95), and half 1 (100). We can merge with the 3.3 node at the bottom of the tree, which also has 0 in about half of its positions and 1 at the other half. Continue to merge similar subtrees to compress our representation of the data.

One of the uses of a VLMC is phasing. Phasing is finding corresponding sequences from both chromosomes of an individual. As it turns out, with a VLMC, this is is a very simple dynamic programming problem. Find the most likely paths through the model – at every position, an individual could be in a pair of states. Transition from a pair of states to another pair of states based on probability. The dynamic programming matrix will store the likelihood of each state transition. We can use this to find the maximum path through all optimal states.

$$V_{ab}(i) = max_{c,d} V_{c,d}(i-1) + trans(c,d \rightarrow a,b)$$

In addition to phasing, VLMCs have been used to analyze identity by descent – to see if two individuals inherited a DNA sequence from a common ancestor.

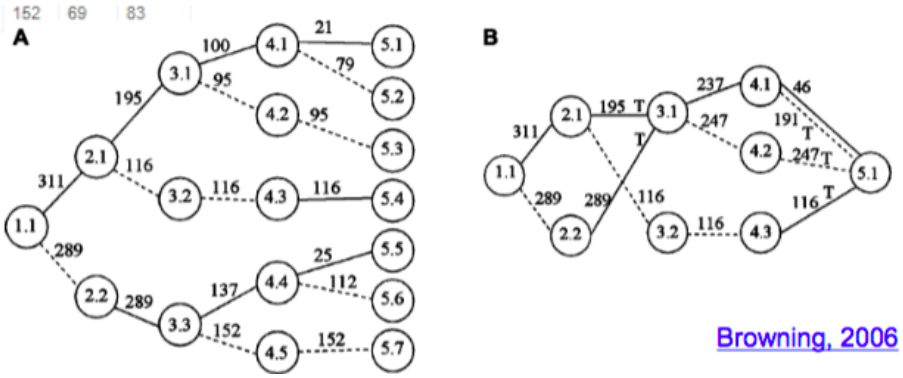

Figure 7:

If we have the haplotypes of two individuals, it is very easy to find segments that are in common, as this is a simple string similarity search. However, if individuals are not unphased, you have the haplotypes, not the genotypes.

One such algorithm that uses VLMCs to solve the identity by descent problem is BEAGLE (FastIBD) – essentially, each of the two sets of genotype data is passed throught BEAGLE. BEAGLE tells the most likely 10 or so parses of the two haplotypes by following the most likely paths of the VLMC graph. IBD finds the path that both individuals share – the likely IBD path.

Next Lecture

Tomorrow we will discuss Positive Selection, Negative Selection, and Fixation.