CS262 – Computation Genomics – Winter 2015
Lecture 15 – Human Population Genomics (02/24/2015)

Scribed by: Junjie (Jason) Zhu
Image Source: Lecture Notes

Introduction

As the cost of sequencing individuals is significantly decreasing, it seems that it is just a matter of time before we can all get our genomes sequenced. Following the 1000 Genome project, we will expect on the order of hundreds of thousands of individuals to be sequenced in the next few years under various projects, such as UK 10K, the Million Human Genome Project and etc. These massive-scale projects lead to applications that are unforeseen. With genomics and computational biology, we are starting to build a much deeper understand of ourselves as a population. This lecture focuses on two main currently popular topics: human evolution and genome-wide association (GWA) studies.

Revisiting Human Genome Diversity

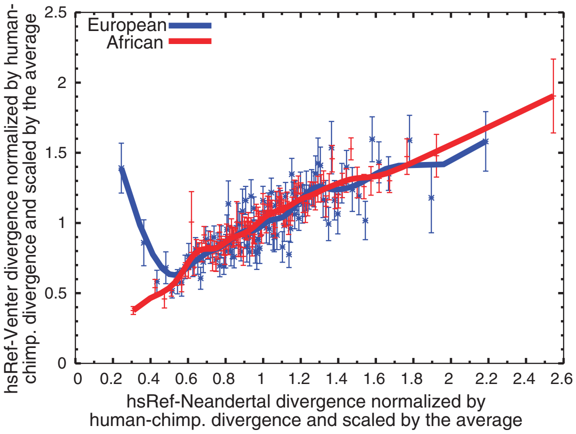
Two lectures ago, we studied evolution and phylogenetic trees. We introduced the idea of comparing the mutations in two individual genomes to trace a common ancestor. It is possible to estimate the time (with respect to the molecular clock) when the common ancestor existed by knowing the number of mutations accumulated per generation (mutation rate). This approach enables us to build the theory that humans originated in Africa and migrated out of Africa about 50,000 years ago. Other interesting things we can do with this approach is finding "Adam" and "Eve" by Y chromosome coalescence and mitochondrial chromosome coalescence respectively.

Human Evolution Involving Other Populations

The Neandertal Genome

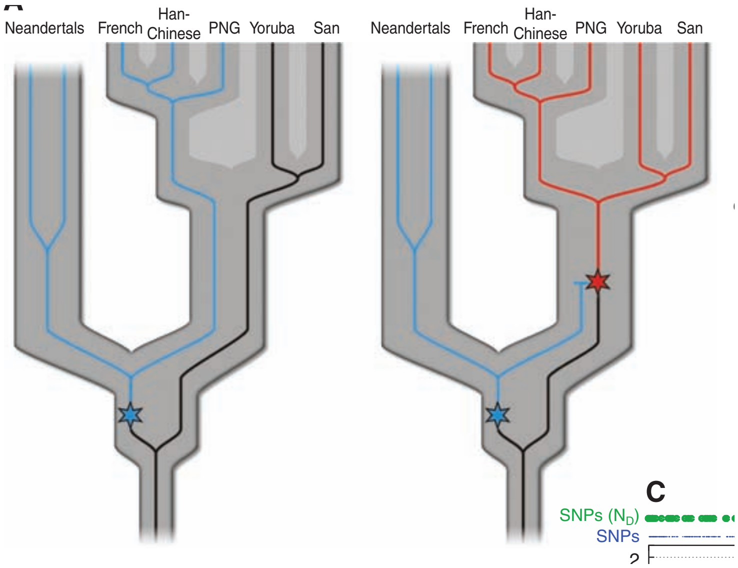
It was also mentioned briefly that Europeans and Asians are approximately 5% Neanderthal, and that there was gene flow from Neandertals into ancestors of non-Africans before they diverged. Here we go into more details on how this theory is suggested by genomic data. In 2010, the draft sequence of the Neandertal Genome published in *Science* by Green et al. using the bones of three different Neandertals. [1] The Neandertal genome was sequenced and compared with five modern humans from different regions in the world. From this paper, a couple of figures come to our attention.

1. Segments of Neandertal ancestry in the human reference genome.



To search for segments where Neandertals and modern humans differ little, haploid human DNA sequences were used instead of diploid sequences, as the latter require both alleles to derive from Neandertals to produce a strong signal. Thus, the human reference genome was used for the purpose of comparison. It was found that European genome segments that are very similar to that in Neandertals are very different from those in present-day humans, whereas this phenomenon is not observed in African segments that are very similar to that in Neandertals. This implies that interbreeding between human and Neandertals occurred after the time point of “Out of Africa”. It is further discussed in the paper from an alternative approach that “non-Africans haplotypes match Neandertals unexpectedly often”. [1]

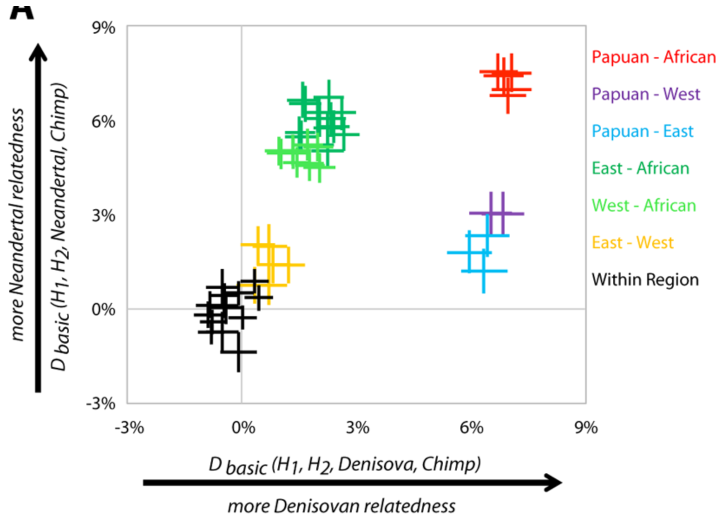
2. Selective Sweep Screen



> The Neandertal genome shares a lot of derived alleles (shown as the blue star in the figure above) in common with modern humans. Thus, it is reasonable to look for alleles where modern humans share a common ancestor *after* diverging from Neandertals. In genomic regions where an advantageous mutation arises (shown as red star) and sweeps to high frequency or fixation in modern humans, we should not expect the Neandertals to have such alleles. Evolution with selection will be discussed in more details later in this class. [1]

The Denisovan Genome

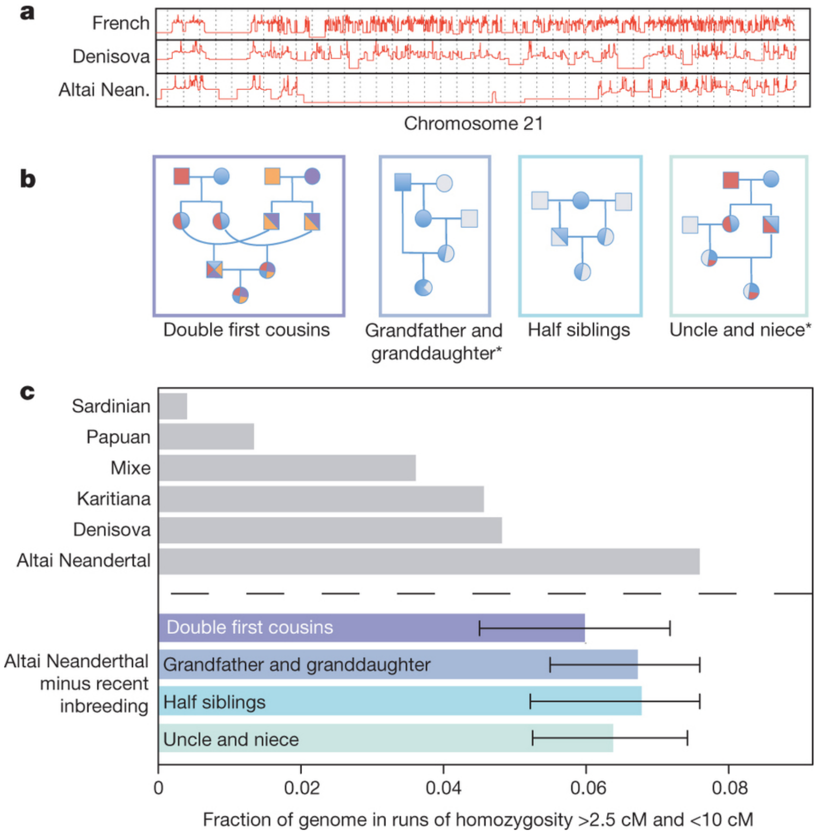
Another population found very far from us is the Denisovan. Also published in 2010 [2], it was reported that a complete mitochondrial DNA sequence retrieved from a bone excavated in 2008 in Denisova Cave. It suggested the existence of another population that lived close in time and space with Neanderthals and modern humans. Here, we are also interested in the question of how much of this Denisovan genome is shared with us.



The figure above focuses on the sharing of derived alleles among modern humans, Denisovans and Neandertals. Pairs of different human population are compared in terms of the "D-statistics", which is a measure of the rate at which the pairs of populations share derived alleles with Denisovans and Neandertals. It is found that Denisovans share more alleles with Papauns than with Europeans and Asians. Notice that for population within regions, the derived alleles they share in common are typically specific to their own population and is not that related to Neandertals and Denisovans. [2]

Runs of Homozygosity

"Runs of homozygosity" are regions of the genome where the copies inherited from our parents are identical because of a common ancestor they had. It does not specifically refer to situations such as marriage between cousins, as we are all related if we go far enough back in the phylogeny tree. However, we can infer the history of intermarriage among a population from homozygosity. In the figure below, we see the fraction of genome in runs of homozygosity for different populations which implies that Neanderthals had a high level of inbreeding which might have been due to the population being segregated and confined to a small region. Modern humans on the other have a much lower level of run of homozygosity as inbreeding in modern society is less common. [1]

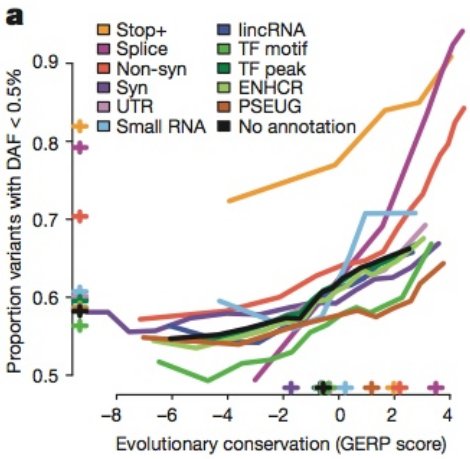


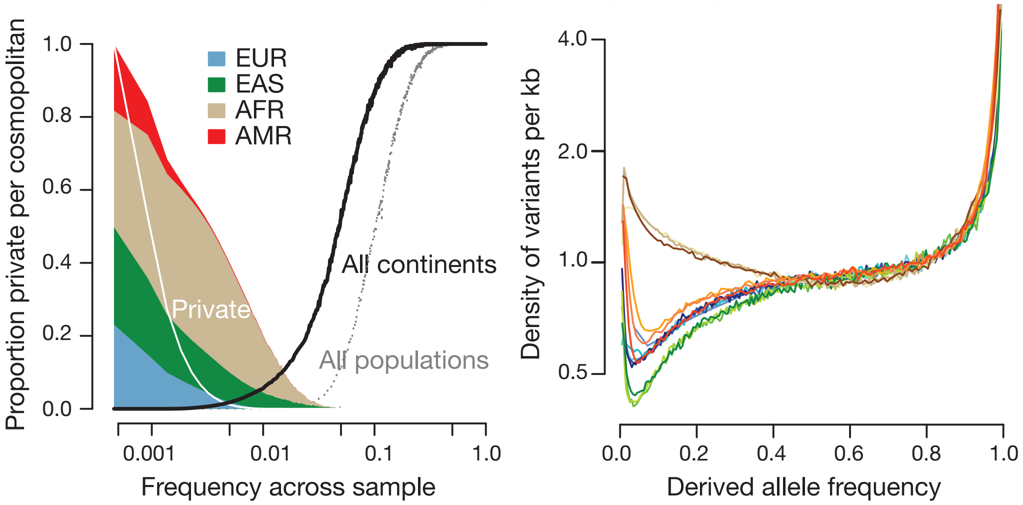
Population Sequencing and Association Studies

1000 Genomes Project

The goal of the 1000 Genomes Project is to find human genetic variation among humans that can be used for association studies. The samples, whose whole genomes are sequenced, are chosen from very different ethnic groups to capture the population diversity. In our previous lecture on multiple-sequence alignment, we discussed how genes are conserved among mammalians due to purifying selection, so we expect the sites where we observe low allele frequency of functional variants to also be conserved across mammals.

In the figure below, the proportion of variants with derived allele frequencies less than 0.5%, which is a measure of purifying selection, is plotted against the GERP score, which is a measure of evolutionary conservation. Variations related to stop codons and splicing have been very infrequent since humans diverged from other primates. [3]

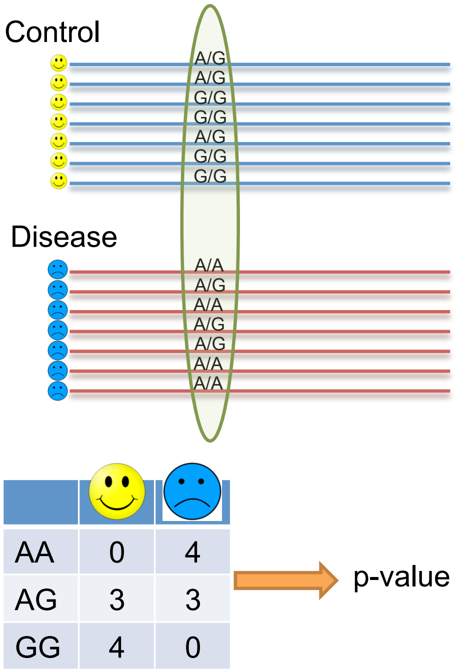




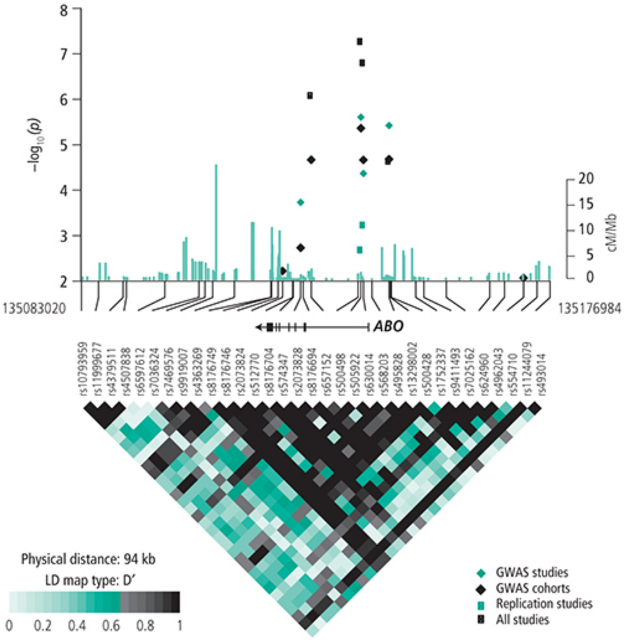
In the left figure above, we observe the fraction of alleles specific to different ancestry groups. Africans have accumulated relatively more of low-frequency variants, whereas Americans have accumulated the least, which is related to how long these populations have been around. In the right figure above, we are interested in the expected number of derived alleles across humans. However, one may point out that if the population is constant, then the plot should be flat. The reason why we see low density of variants at low derived allele frequency and high density of variants at high derived allele frequency can be explained by two main reasons: 1) A lot of derived alleles got fixed in a small population, and 2) Recent population expansion has led to accumulation of a larger number of rare alleles.

GWA Studies

Whole genome information from many individuals allows us to conduct studies to identify how common genetic variants related to certain traits, and typically the focus is on single-nucleotide polymorphisms (SNPs) and traits like common diseases where we can get a large enough sample size. For a given disease, we compare two groups of population: one with disease (cases) and one without the disease (controls). Then the goal is to use the genotypes of these individuals to see which loci are associated with the disease as shown in the left figure below. The metric of statistical significance is typically the p-value. The smaller it is for a SNP, the more likely that this SNP is associated. Thus, we can plot the negative log of the p-values of al SNPs in a Manhattan plot as shown in the right figure below, and find the peaks that indicate strong correlation with the disease. Notice that there are multiple close-by SNPs that have small p-values due to linkage disequilibrium, so that SNPs that are linked also show some correlation with the trait.



	🙂	🙁
AA	0	4
AG	3	3
GG	4	0

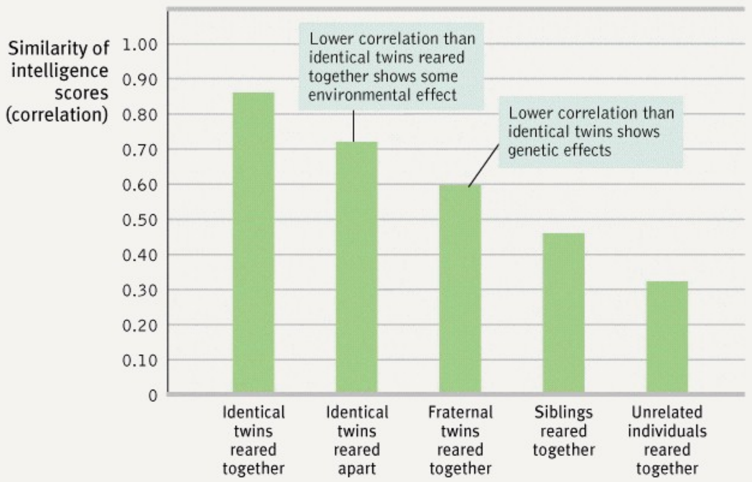


Traditionally, there have been several limitations in GWA studies, such as insufficient sample size and false positive results. However, high throughput whole genome sequencing technologies can provide a better alternative to overcome the related shortcomings.

Heritability and Environment.

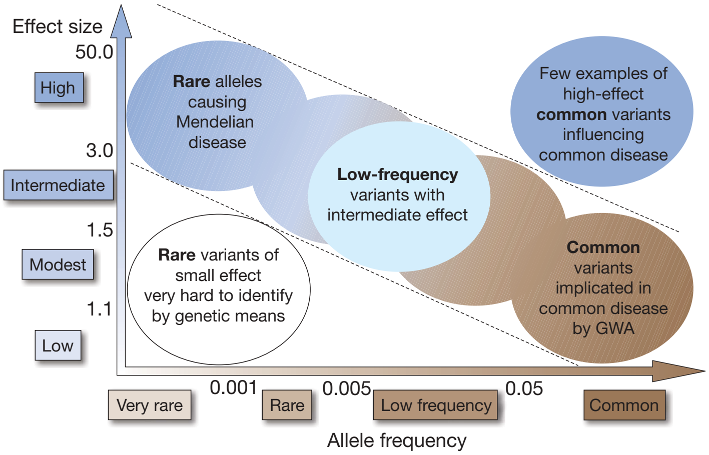
Taking a step back to the big-picture question, we are interested in how much of the variance of certain traits across populations can be explained by the variance of genes, in other words, whether certain traits are heritable.

Twin studies are popular methods to understand the genetic influences on certain traits in a population sample. (Intuitively, a given trait in only one member of the pair of twins provides a strong signal of environmental effects.) It was found that there is a higher correlation of intelligence among identical twins (who share almost all genomes) reared apart than fraternal twins (just like normal siblings) reared together. Thus, this comparison suggests that intelligence has a strong dependency on genetics. [4]



What is the "Missing Heritability"?

Hundreds of genetic variants have been identified to be associated with complex diseases and traits through GWA studies, but most variants only explain a small portion of familial clustering. This leads to the concept of "missing" heritability where SNPs cannot explain the remaining risk or variance. Even traits such as the human height, which is a classical complex trait found to be associated with 40 loci, cannot be explained well by the genotype information. [5]


Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common
Allele frequency

The figure above captures the feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect. The diagonal components are where most emphasis in identifying associations lies. Certain variants may not be sufficient (with minor allele frequency <0.5%) in the sample to be detected by GWA genotyping arrays.

Importantly, low frequency variants (with minor allele frequency between 1% and 5%) in the middle of the plot can have significant effect sizes too low to support Mendelian segregation and to be detected by traditional linkage approaches. Meanwhile, the minor allele frequency is also too low to be clearly identified by GWA studies. Thus, the 1,000 Genomes Project and other large-scale projects are targeted towards finding these variants with low or even rare minor allele frequencies.

References

[1] Green, Richard E., et al. "A draft sequence of the Neandertal genome." science 328.5979 (2010): 710-722.
[2] Krause, Johannes, et al. "The complete mitochondrial DNA genome of an unknown hominin from southern Siberia." Nature 464.7290 (2010): 894-897.
[3] 1000 Genomes Project Consortium. "An integrated map of genetic variation from 1,092 human genomes." Nature 491.7422 (2012): 56-65.
[4] Bienvenu, O. J., D. S. Davydow, and K. S. Kendler. "Psychiatric 'diseases' versus behavioral disorders and degree of genetic influence." Psychological medicine 41.1 (2011): 33.
[5] Manolio, Teri A., et al. "Finding the missing heritability of complex diseases." Nature 461.7265 (2009): 747-753.