Lecture 12 – Human Genome Diversity
CS 262 – Computation Genomics
12 February 2015
Scribe: Isabel Bush
Images from CS 262 Lecture Slides

How similar are we and why?

Can attempt to answer through sequencing of individual genomes

- Resequencing (compare to a reference) opposed to de Novo sequencing for new species
- Prof Batzoglou expects 100,000 genomes will be sequenced by 2015

Human Evolution

- Humans and other species evolved through phylogenic tree
    - Humans diverged with Old World monkeys about 25 – 35 million years ago
        - Chimpanzees & bonobos are most similar to us but have very different behavior – chimps have patriarchal societies and resolve conflict through fights, bonobos have matriarchal societies and resolve differences through sex – human behavior somewhere in between
    - Mammalian line diverged about 100 million years ago
- Originally two migrational theories
    - Out of Africa ~50,000 yrs ago – humans from Africa completely replaced others elsewhere
    - Multiregional evolution – theory that humans developed separately in different regions (theory was debunked)
- Europeans & Asians are ~5% Neanderthal
    - Likely mating of Homo sapiens and Neanderthals after left Africa
    - Determined ~5% Neanderthal by comparing alignment of modern European individuals' genomes with Neanderthal to African individuals' alignment with Neanderthal
    - Why did the Neanderthal's disappear?
        - Prof Batzoglou thinks Neanderthals just interbred with Homo sapiens enough to slowly disappear (small mating preference over time)
- Coalescence: Defining "Adam & Eve"
    - Adam -> Y-chromosome coalescence
        - Since Y-chromosomes passed from father to all sons, only a subset of the Y-chromosomes of men alive today were present in the previous generation (all brothers have same Y-chromosome from their dad)
        - Trace further and further back, must coalesce to Y-chromosome of a single individual
        - By measuring differences between Y-chromosomes of men today and estimating the number of changes per generation, can calculate "Adam" existed about 120,000 - 340,000 yrs ago

- Eve -> Mitochondrial DNA coalescence
  - Since mitochondrial DNA is passed from mother to child through the egg, can trace mitochondrial DNA back through female generations in the same way as the Y-chromosome for males
  - "Eve" existed about 99,000 – 150,000 yrs ago

Genetic Definitions

- Alleles – Different versions of a given gene (or position in the gene)
  - Major alleles – Most common allele
  - Minor alleles – Other alleles
- Heterozygosity – Probability that two alleles chosen at random in a population will be different
  - $H = \frac{4N\mu}{1+4N\mu}$, where N = population size and μ = mutation rate
  - Mutation rate introduces variation, larger populations have a higher chance of passing on those variations
  - In humans (& many other species), $4N\mu$ is small and thus $H \approx 4N\mu$
- Recombination – Rearrangement of DNA due to crossing over in chromosomes
  - Rare (about 1/100 Mbp)
  - Need at least one per chromosome or miosis wouldn't work
- Linkage disequilibrium – Degree of correlation between two SNP locations
  - Association of alleles from grandparents is not random (correlated in large blocks between recombination sites)
- F_{ST} is a measure of genetic diversity in a population
  - $F_{ST} = \frac{H-H_{pop}}{H}$, where H_{POP} = heterozygosity of specific population, H = heterozygosity of the entire species
  - F_{ST} is usually between 0 and 1
  - High F_{ST} indicates low population diversity (more similar individuals)
  - Genetic diversity decreases linearly with population distance from sub-Saharan Africa (as can be seen in Figure 1). Humans have left behind genetic diversity as they've migrated out of sub-Saharan Africa, and new mutations are not enough to compensate

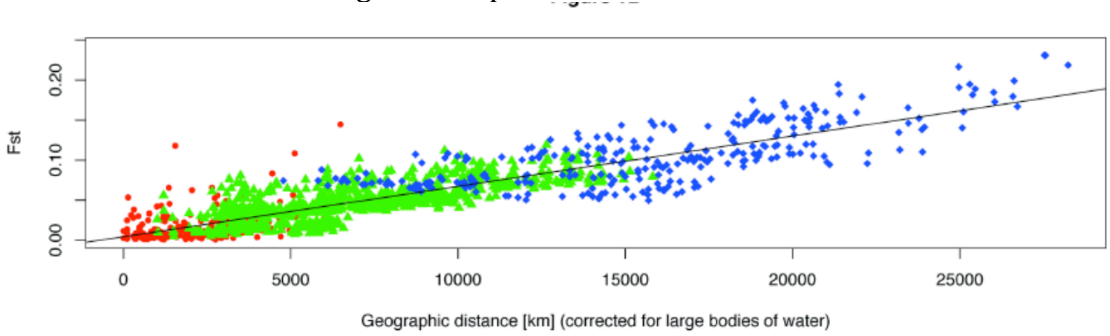

Figure 1: Fall in Heterozygosity: Genetic diversity decreases with distance from sub-Saharan Africa (each dot on the plot represents the F_{ST} for specific local population group at some distance from Africa)

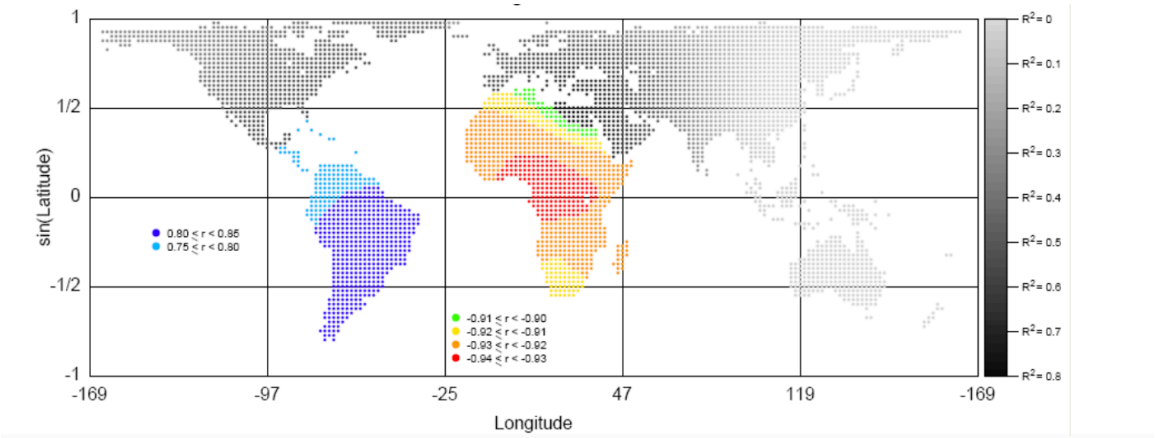

Figure 2: Map indicates the fit correlation if assume humans originated in that location. If sub-Saharan Africa is assumed to be the origin, the fit is good; if assume Europe, there is no correlation; and if assume South America there is an inverse correlation (essentially reading the above plot in reverse).

HapMap Project

- To characterize genetic diversity across populations
- Wished to find high-frequency minor alleles across populations of humans
- Use microarrays with probes having complementary strands to major and minor alleles & determine if your DNA sticks to the major allele probe, minor allele probe, or both
    - This is the idea between 23andMe and other genetic ancestry companies

Linkage Disequilibrium

- Linkage disequilibrium between two alleles can be calculated as the probability of finding them together less their single-nucleotide probabilities

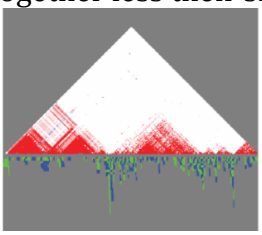

Figure 3: In charts such as above, each point on the x-axis is a different position along the DNA. A red dot forming an equilateral triangle between two positions indicates those two positions are correlated.

- Recombination hotspots
    - Predictable locations on the chromosome for recombination to occur
    - Although human & chimpanzees are 98-99% similar and each have very predictable recombination hotspots, the hotspots are very different between the two species
    - Hotspots are “1-use only”
        - If your mom recombined your grandparents DNA at a hotspot, when you recombine your parents, it will be at a different hotspot
        - Result is that hotspots change throughout generations

Human Genome Resequencing

- Align to a reference genome using BWA/Bowtie
- Types of SNVs
    - Germline (SNPS) – inherited & present in all cells
    - Somatic – present in some cells due to mutations as cells divided
- Phasing – determine haplotypes from genotypes
    - For example, if determine through sequencing that a person's genotype is A/C at one position and A/T at another, phasing allows to determine if the two haplotypes are A…A and C…T or A…T and C…A.
    - Phasing is needed because any single read is not likely to have more than one heterozygous region
        - Since ~3 billion base pairs and 3 million heterozygous regions, ~1000 bps between heterozygous alleles
        - Reads are only ~150 bps long
    - 3 ways to find phasing
        1. Statistical phasing (over a population)
            - Maximize likelihood of all data given all haplotypes
            - Error rate is relatively high
        2. Pedigree analysis
            - Accept phase that is consistent with parents
        3. Molecular phasing
            - Cut DNA into 10-kbp segments
            - Dilute (add water)
            - Put segments into wells (each well has a small fraction of DNA)
            - Amplify (PCR) and attach primers (unique to each well)
            - Perform DNA sequencing to obtain read clouds for each well
            - Each read cloud comes from either paternal or maternal DNA