OLLI SG 497
Ancient DNA

Session 2 - October 5, 2022

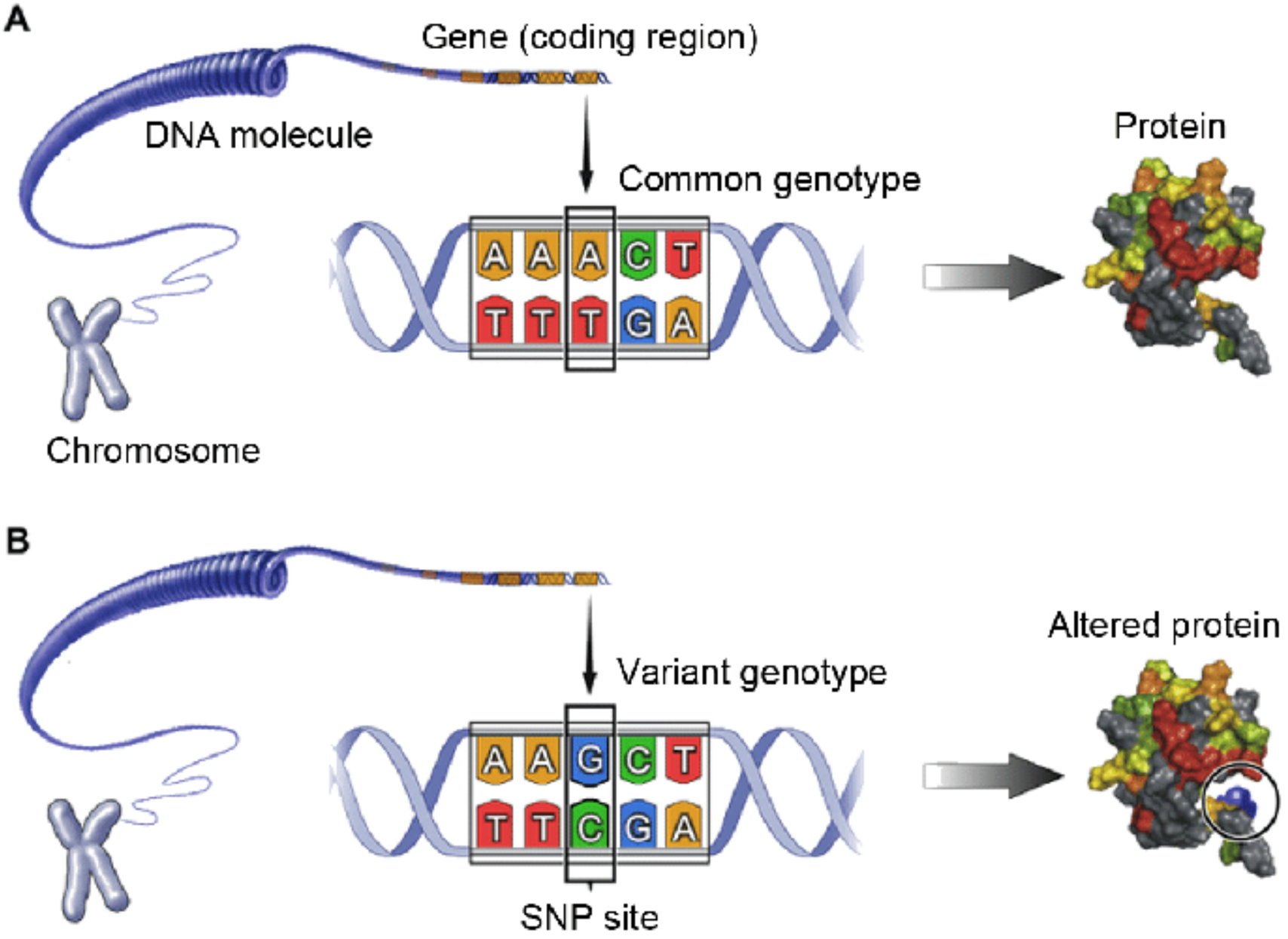
Recap

- Introduced the main topics of the science of ancient DNA.
- Posed two questions about the methodology.
- Gave a thumbnail sketch of DNA, chromosomes, genes, SNPs, and meiosis.

Today's Meeting

- Finish up on recombination and meiosis.
- Statistical genetics and Principal Component Analysis.
- Overview of the laboratory method for extracting, purifying and sequencing Ancient DNA.
- Issues raised in Chapter 1: How the Genome Explains Who We Are.

Single Nucleotide Polymorphism



Meiosis and Recombination

- During cell division of germline cells (eggs and sperm), random shuffling of the DNA contributed by both parents occurs.
- Chromosomes swap segments (crossing over or recombination), and independently align.
- This results in multiple, unique gametes with 23 chromosomes (actually, for eggs it is somewhat more complicated than this, but the net result is the contribution of 23 chromosomes to the zygote).
- Meiosis.

Statistical Genetics

Considerations

- Luca Cavalli-Sforza - Reich's assessment:
  - Limited data.
  - Faulty mathematical technique.
- Molly Przeworski - **2006** study:
  - Genome scans of present-day human genetic variation will miss most instances of natural selection because they will not have the **statistical power** needed to detect it.
  - **180 independent genetic changes** that are more common in shorter people.

Statistical Genetics

Reich's Approach

- More data:
  - Genome-wide Association Studies (GWAS) on present-day human genomes. Hundreds of thousands of samples.
  - Rapidly increasing genome-wide studies of ancient DNA.
- Better mathematical/statistical technique:
  - Principal Component Analysis (over large data sets).

Statistical Genetics

Principal Component Analysis

- Primary use of Principal Component Analysis:
  - Used to reduce the dimensionality of large data sets by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
  - Reduce the number of variables of a data set, while preserving as much information as possible.

Statistical Genetics

Principal Component Analysis

HOW DO YOU DO A PRINCIPAL COMPONENT ANALYSIS?

- Standardize the range of continuous initial variables
- Compute the covariance matrix to identify correlations
- Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
- Create a feature vector to decide which principal components to keep
- Recast the data along the principal components axes

Statistical Genetics

Principal Component Analysis - A Caution

From an article on PCA on Wikipedia:

In 1978 Cavalli-Sforza and others pioneered the use of principal components analysis (PCA) to summarize data on variation in human gene frequencies across regions. The components showed distinctive patterns, including gradients and sinusoidal waves. They interpreted these patterns as resulting from specific ancient migration events.

Since then, PCA has been ubiquitous in population genetics, with thousands of papers using PCA as a display mechanism. **Genetics varies largely according to proximity,** so the first two principal components actually show spatial distribution and may be used to map the relative geographical location of different population groups, thereby showing individuals who have wandered from their original locations.

PCA in genetics has been technically controversial, in that the technique has been performed on discrete non-normal variables and often on binary allele markers. The lack of any measures of standard error in PCA are also an impediment to more consistent usage. In August 2022, the molecular biologist Eran Elhaik published a theoretical paper in Scientific Reports analyzing 12 PCA applications. He concluded that it was easy to manipulate the method, which, in his view, generated results that were 'erroneous, contradictory, and absurd.' Specifically, he argued, the results achieved in population genetics were characterized by cherry-picking and circular reasoning.

Statistical Genetics

Principal Component Analysis - But On the Other Hand…

From: Ancient DNA Analysis, published February 11, 2021, in *Nature Reviews Methods Primer*

Principal component analysis (PCA) is a classic, exploratory statistical method that is used to represent genetic affinities among individuals within simple graphs. It summarizes genetic variation measured from individual genotypes in hundreds to thousands of individuals for thousands to millions of SNP loci into a reduced number of dimensions that are shaped by ancestry. Those dimensions represent the principal components and provide the main axes of the graphical representation of genetic affinities. aDNA data from an individual are often projected onto present-day genetic variation, but when sufficient data are available, ancient specimens can be included in the principal component computation itself…. It is notable that PCA clustering is sensitive to the sample size of different ancestries and also to their amount of genetic drift, which exaggerates principal component distances. Therefore, individuals from the same population but distant in time may be misleadingly separated in PCA space. Multidimensional scaling has been proposed as an alternative to PCA in cases where only minimal sequence coverage is available, typically 0.001–0.1× for human data. For both methods, accurate clustering can be achieved even in the absence of sequence overlap between different ancient individuals, as long as sufficient sequence data enable the estimation of genetic distance to a predefined panel of relevant ancestries.

Statistical Genetics

Principal Component Analysis

- Reich’s use of PCA - Simon Foundation Lecture:
  - Video clip 1 - from 22:25 to 25:00
  - Video clip 2 - from 26:50 to 31:20

Sequencing Ancient DNA

Mathias Meyer and Qiaomei Fu's Method

- Synthesized 52 nucleotide long segments of DNA (single-stranded).
- The sequence of nucleotides in these segments overlapped, so that in total they covered all the nucleotides in human chromosome 21.
- Using techniques developed for producing microchips, they attached millions of these segments to wafers (microarrays).
- Fragments of DNA (single-stranded) extracted from the ancient bone sample were separated out from the other molecules in the sample.
- The solution containing the DNA fragments was washed over the wafers.

Sequencing Ancient DNA

Mathias Meyer and Qiaomei Fu's Method

- A sequence of DNA nucleotides will tightly bind to the complementary sequence.
- The synthesized 52 nucleotide long sequences served as "bait" to "fish out" (extract) human DNA from the bone sample.
- They were then able to sequence all of the fished out sequences.

Sequencing Ancient DNA

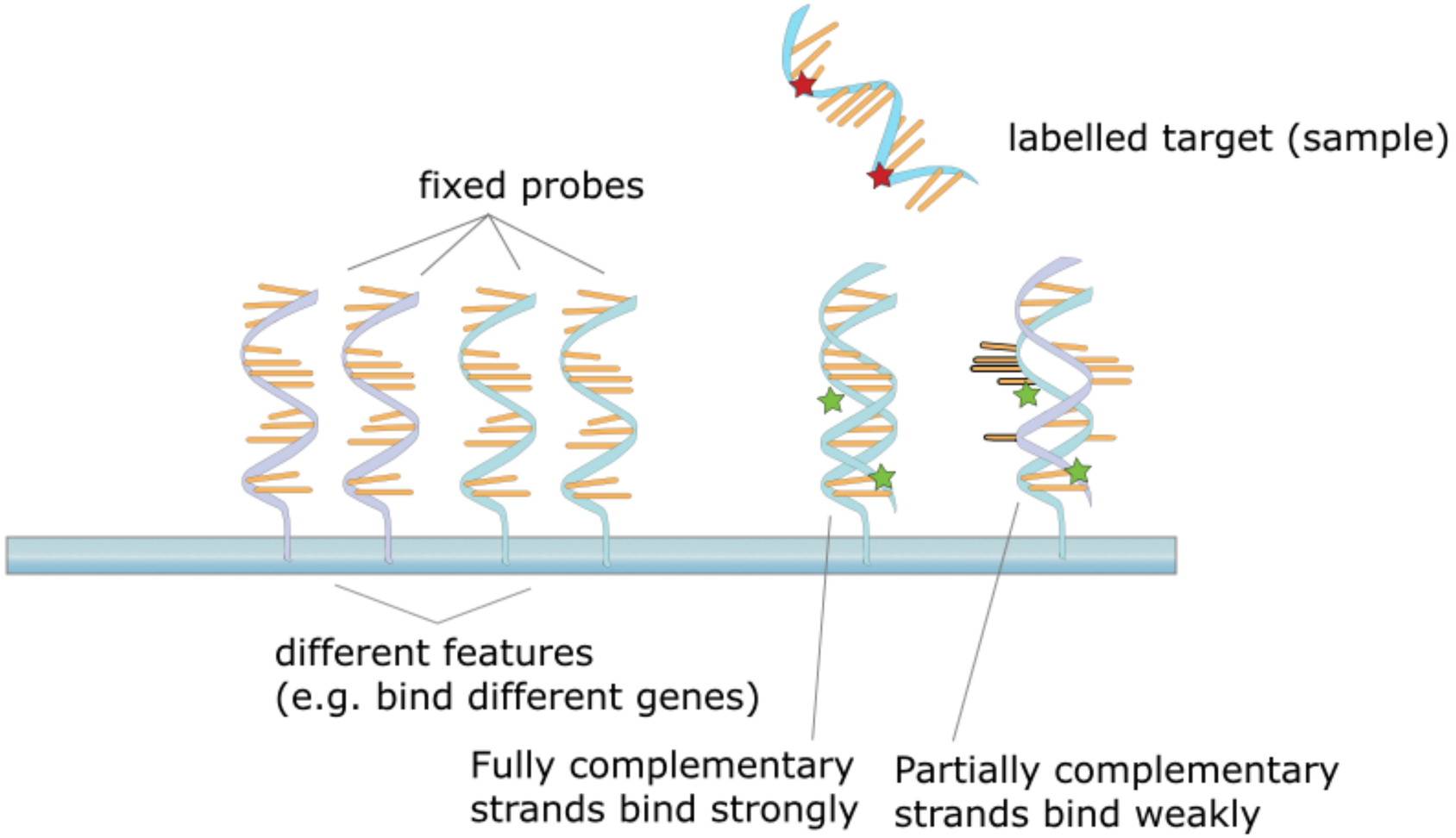
DNA Microarrays

From: Wikipedia article on DNA microarray

The core principle behind microarrays is hybridization between two DNA strands, the property of complementary nucleic acid sequences to specifically pair with each other by forming hydrogen bonds between complementary nucleotide base pairs. A high number of complementary base pairs in a nucleotide sequence means tighter non-covalent bonding between the two strands. After washing off non-specific bonding sequences, only strongly paired strands will remain hybridized. Fluorescently labeled target sequences that bind to a probe sequence generate a signal that depends on the hybridization conditions (such as temperature), and washing after hybridization. Total strength of the signal, from a spot (feature), depends upon the amount of target sample binding to the probes present on that spot. Microarrays use relative quantitation in which the intensity of a feature is compared to the intensity of the same feature under a different condition, and the identity of the feature is known by its position.

Sequencing Ancient DNA

DNA Microarrays



Sequencing Ancient DNA

Nadin Rohland and David Reich's Enhancements

- Expanded the method to include more of the genome, and "industrialized" the method.
- Instead of focusing on just the nucleotide sequence of chromosome 21, they synthesized 52 nucleotide sequences that corresponded to more than a million positions that are known to be human and known to vary among humans.
- This dramatically increased the amount of human DNA from the sample.
- They added the use of robots to speed up the process, and enhanced computer systems to analyze the resulting sequence data.

How the Genome Explains Who We Are

Two Approaches

- Differences/mutations in DNA sequences tell us how long ago we shared a common ancestor. It tells us about human history.
    - “The higher the density of differences separating two genomes on any segment, the longer it has been since the segments shared a common ancestor.” - Reich
- Differences/mutations in DNA sequences explain how we differ from other animals and our human predecessors, … what makes us unique.
    - Recognizably modern human behavior arose fifty thousand years ago as a result of a rise in the frequency of a single mutation of a gene affecting the biology of the brain. - Klein

How the Genome Explains Who We Are

Reich on “Simple Explanations”

- FOXP2
  - Three mutations in this gene that are present only in modern humans.
  - However, all human populations today are capable of complex cognition, even though some lineages separated out more than 200,000 YA.
  - The common ancestor of everyone alive today, and who had the mutated FOXP2 gene, lived more than one million years ago.
  - The mutated FOXP2 gene cannot explain the change in modern human behavior that occurred 50,000 YA.

How the Genome Explains Who We Are

Reich on “Simple Explanations”

- The genome revolution has had the most success in explaining human migrations, rather than in explaining human biology.
- Reich’s focus will be on population migrations, population mixtures, and population replacements.

Next Up

- Finish up issues from the Introduction and Chapter 1.
  - No “Mitochondrial Eve”, no “Y Chromosome Adam.”
  - Early non-African population bottleneck.
  - The effect of natural selection.
- Neanderthals!