Karola Kirsanow

November 24, 2019

I regularly present my research to different audiences. These are the first ten slides and associated notes of a presentation I delivered to a group of computational biologists. Here I am describing the bioinformatic peculiarities of ancient DNA work to an audience that works with modern human DNA.

Bioinformatic challenges associated with ancient DNA

Title Slide

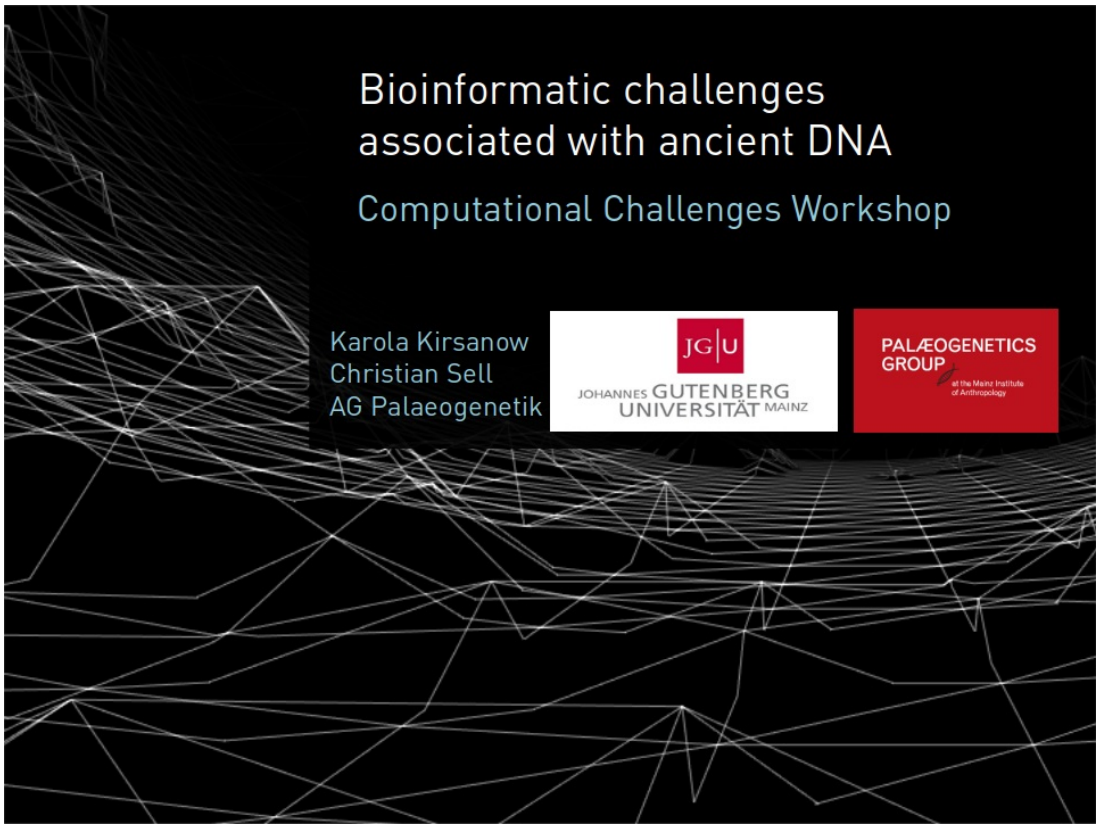

Figure 1: Title slide

I'm glad to be here today to describe how we relate the sparse and often compromised ancient human DNA record to much larger and more comprehensive modern genomic datasets.

Slide 02

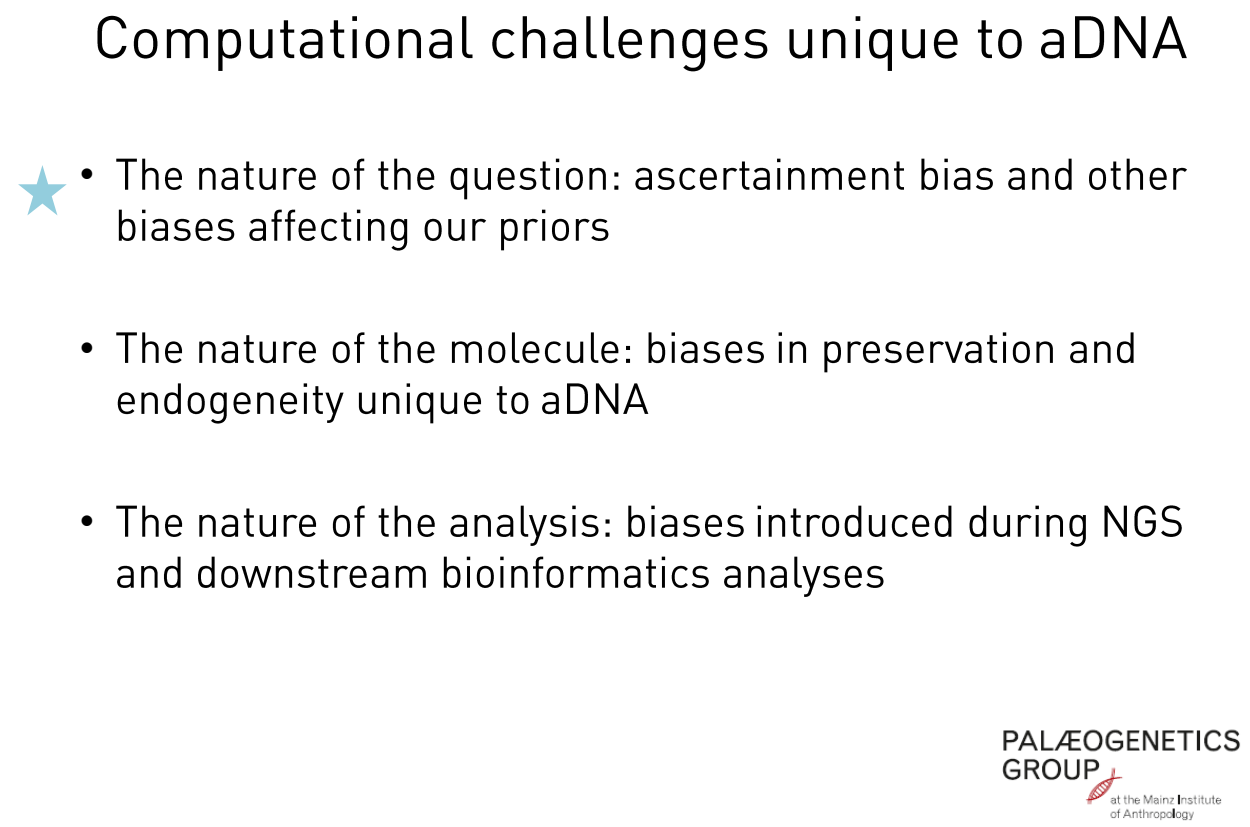

Figure 2: Slide 2

We develop our research questions based largely on data gathered from modern human individuals, particularly modern Europeans. So there are several levels of potential bias here:

First, 'horizontally' speaking, we are assuming our ancient population resembles HapMap Eurasians, and

Second, 'vertically' speaking, we are assuming that patterns of variation in modern populations are comparable with those in ancient populations.

We run the risk of begging the question when the aim of our analysis is to compare ancient and modern variation, or to identify now-extinct patterns of ancient variation, if we rely too heavily on our modern sample as a reference set.

Slide 03

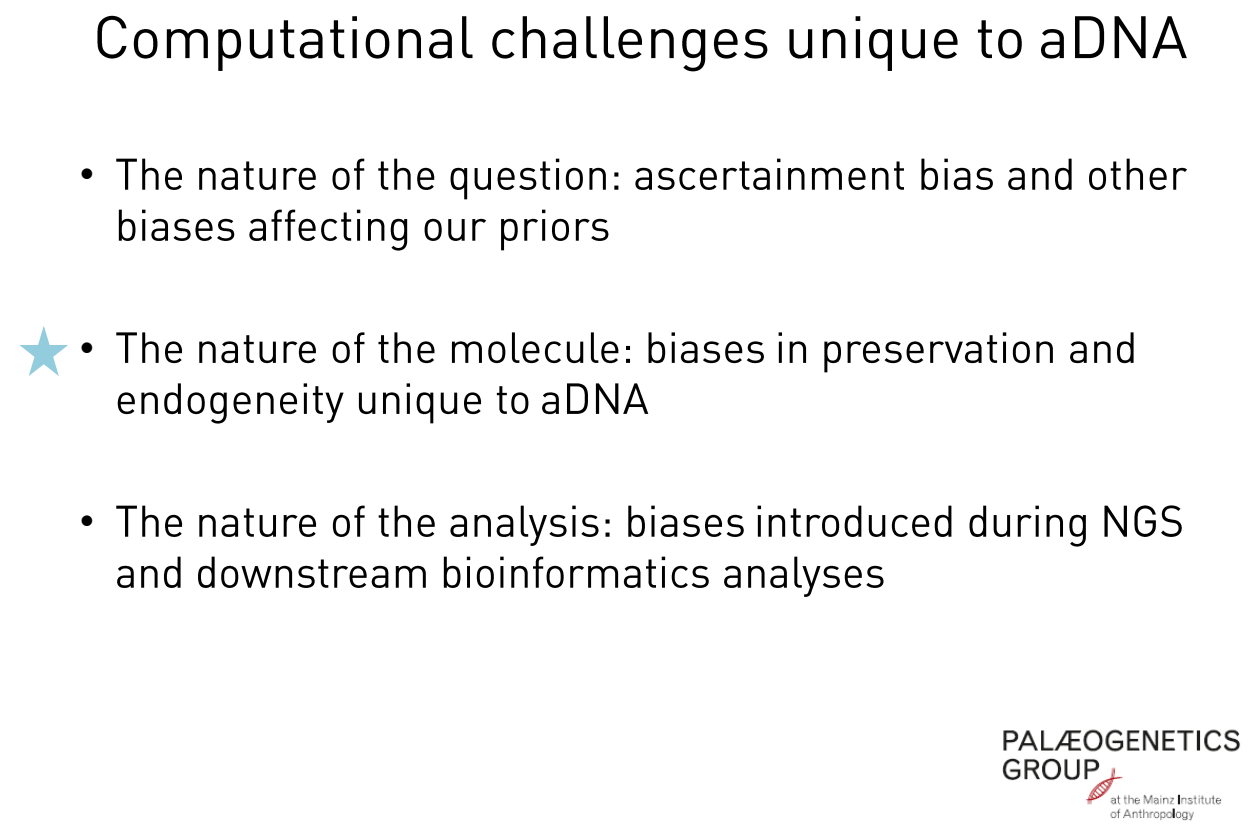

Figure 3: Slide 3

Certain physical characteristics of ancient DNA can be simultaneously problematic for interpretation and useful for data validation. For example, aDNA has relatively shorter fragment lengths and characteristic patterns of damage (missing bases, de-aminations) relative to modern DNA. Indeed, undamaged contaminant molecules are better suited for enzymatic manipulation, and can thus be over-represented in the product pool.

Slide 04

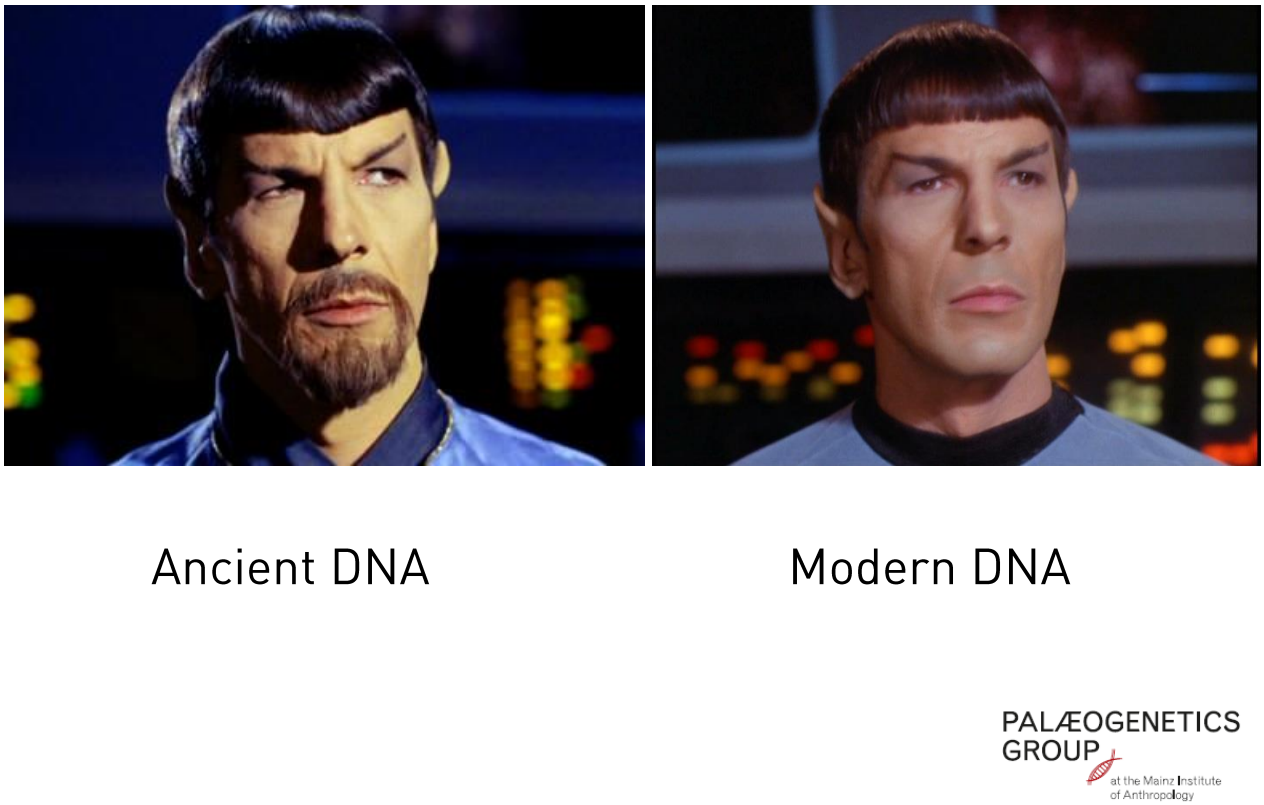

Figure 4: Slide 4

Long pristine stretches of DNA are suspicious — it is kind of like doing genomic work in the universe where Spock has a beard. When we see a short sequence with this diagnostic nucleotide misincorporation pattern, our first thought is 'Great, endogenous DNA!' and our second is 'Crap, now I have to identify and account for the damaged nucleotides'.

Slide 05

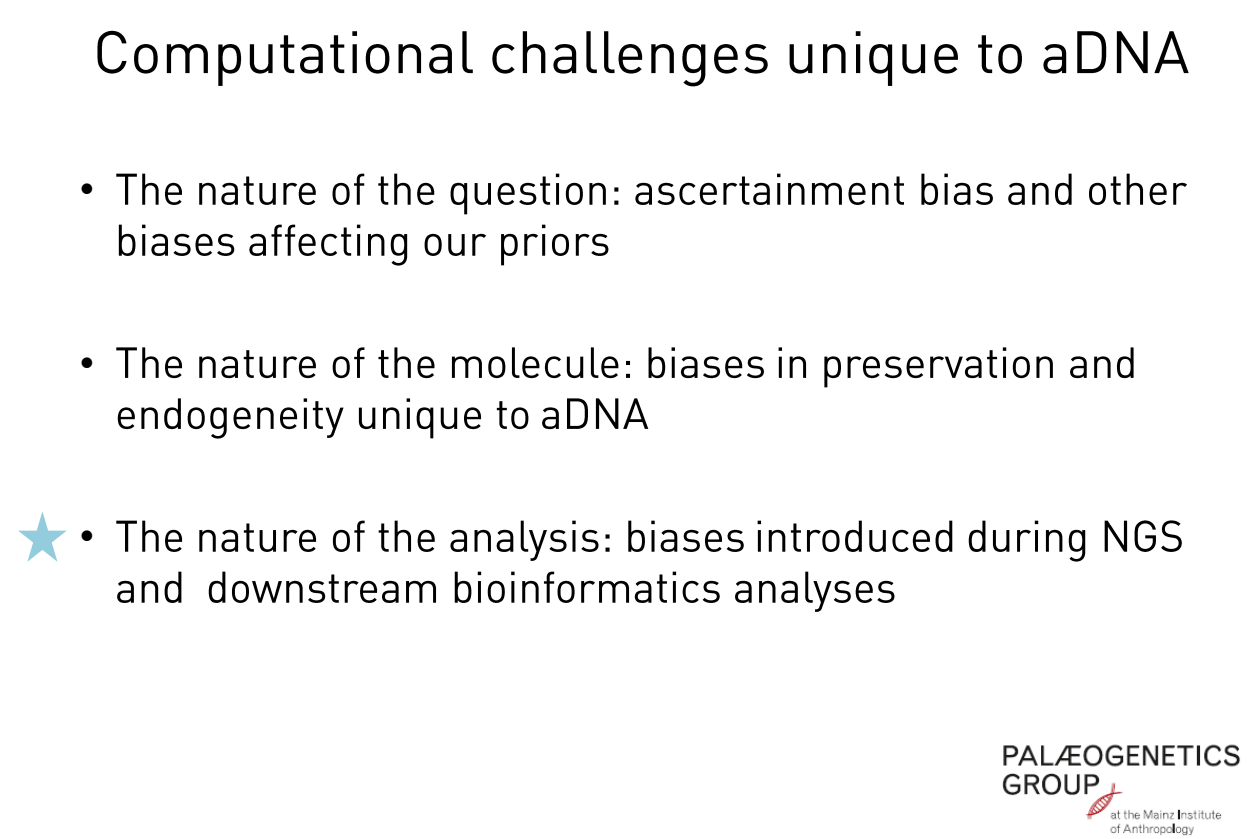

Figure 5: Slide 5

There are also more straightforward issues associated with the physical realities of ancient DNA, such as the problem of calling accurate genotypes from low coverage data, and difficulties phasing genomic data with missing sites and an excess of rare variants.

For example, under 5X coverage it is quite possible that only one chromosome of a diploid individual has been sampled. And conversely, the high raw error rates of next-generation sequencing may cause a significant number of homozygous genotypes to be wrongly inferred as heterozygous if the call is based on the presence/absence of a non-ref allele. In most NGS, the error rate is at least 0.1% even after stringent filtering based on quality scores, in 5X data an error will appear in 0.5% of all homozygotes, which is above the Minor Allele Frequency cutoff for calling a SNP. In multisample calls, then, most SNPs will be errors. So there is a tradeoff between including too many 'SNPs' and under-calling heterozygotes.

Slide 06

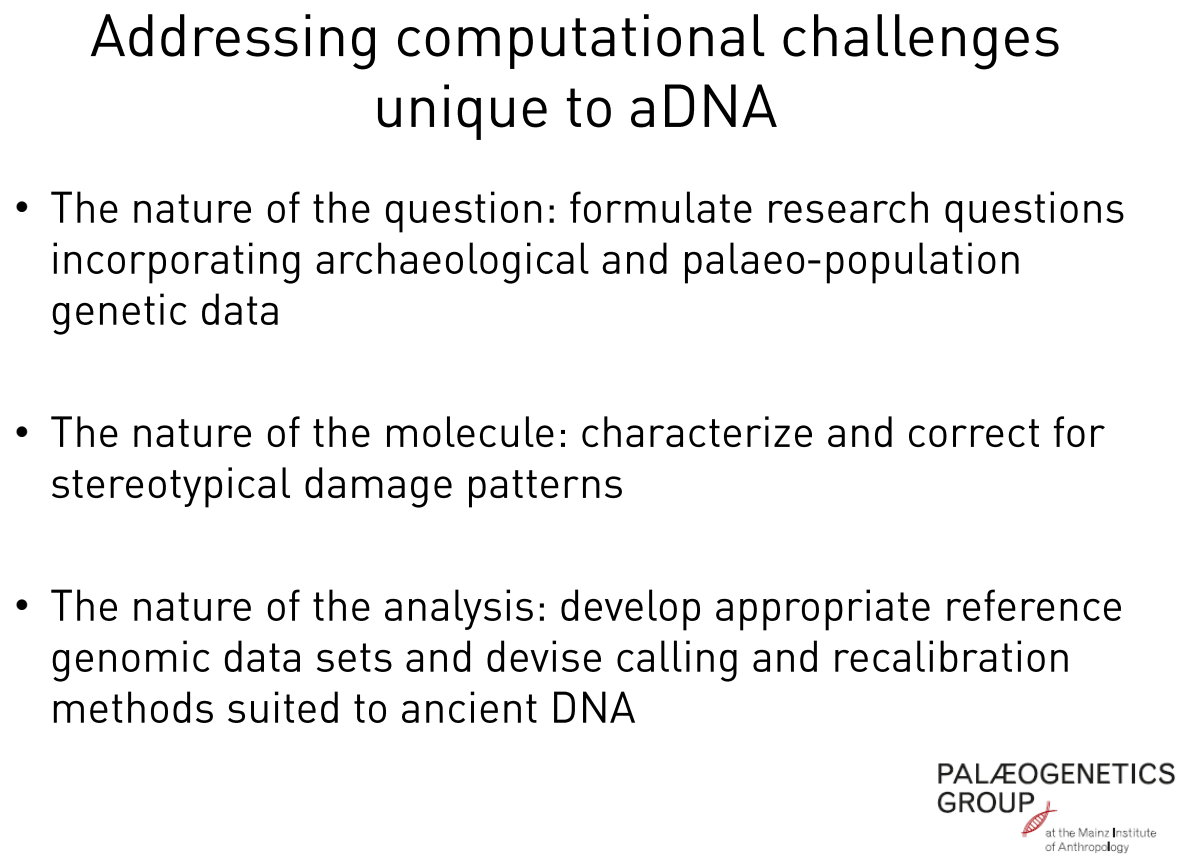

Figure 6: Slide 6

The third category of challenges is conceptually related to the first — any downstream analysis that requires population level observations as a prior can be complicated by our limited dataset concerning ancient population genetics. SNP and genotype calling algorithms use probabilistic frameworks with expectations based on modern empirical data. For example, certain widely-used genotype callers can incorporate population frequency data as a weak prior when calculating genotype quality scores. Since the true frequency of variants in our ancient populations is often unknown at the outset of an analysis, we have to be careful to use flat prior so as to avoid circular reasoning when we estimate posterior probabilities. This problem also attaches to any attempts to use unsupervised imputation methods on ancient genomic data.

Slide 07

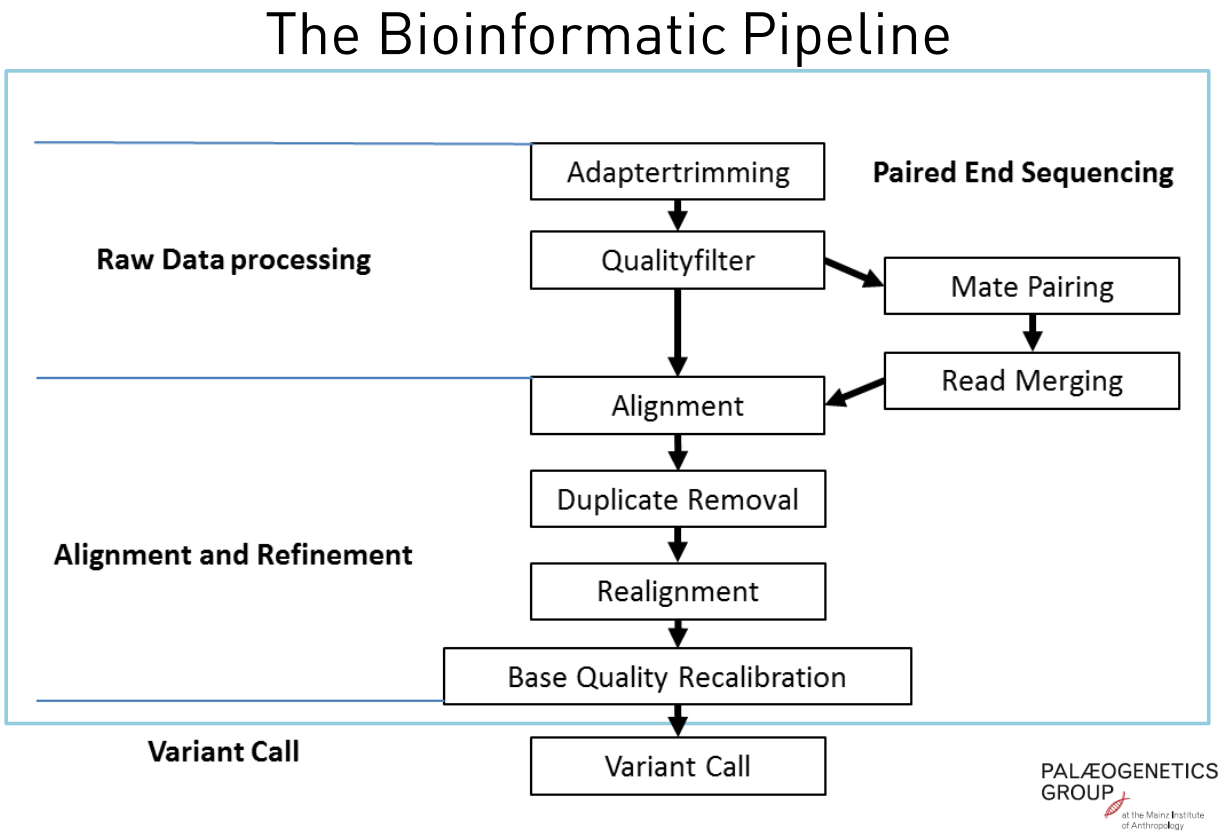

Figure 7: Slide 7

We employ 3 primary approaches to address these problems:

1. Exploit other data streams outside of the circularity problem
2. Deploy a damage control model to tackle the problem of postmortem damage
3. Create ancient genomic reference sets comparable to those for select modern populations and genotype them using the same variant calling framework so that at least relative intra-sample comparisons can be made (navigating the distance between ancient and modern populations on, say, a PCA will be trickier)

Slide 08

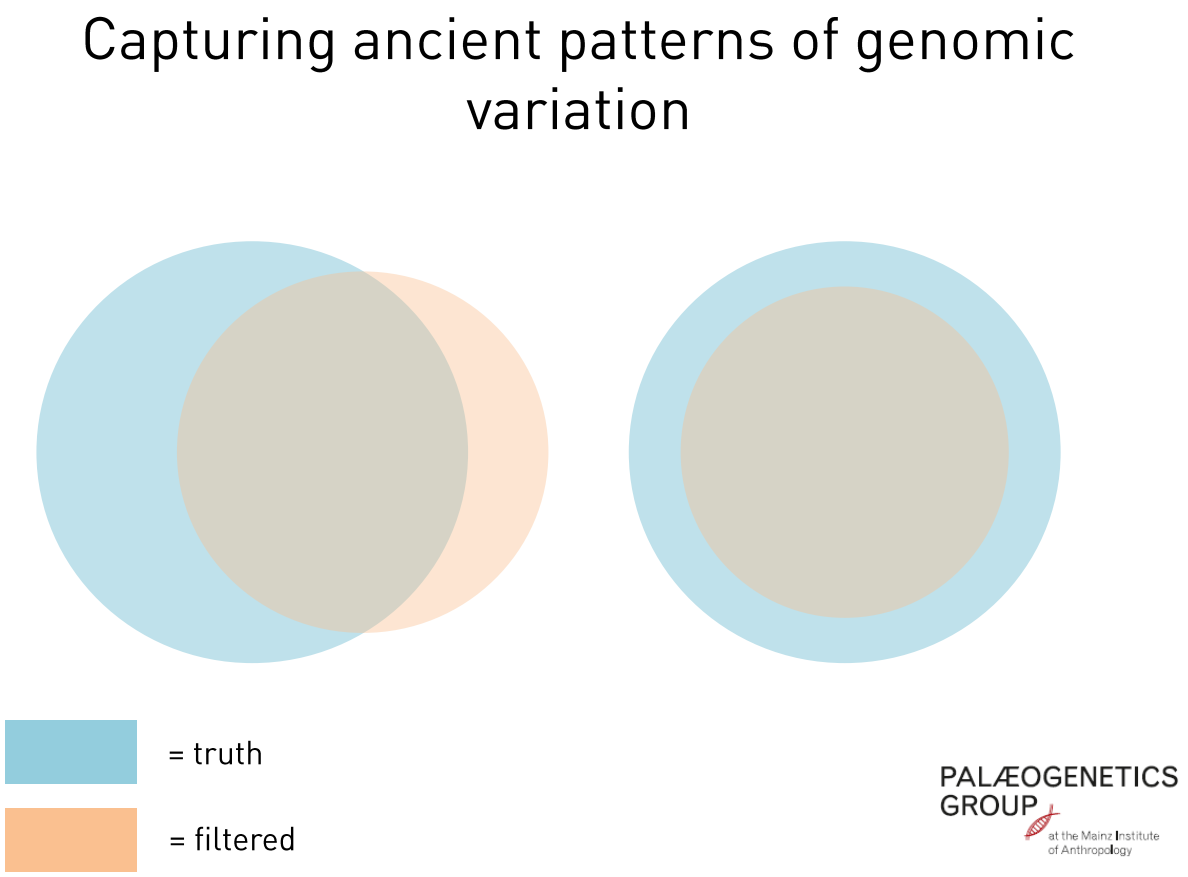

Figure 8: Slide 8

Our initial approach was to attempt to bioinformatically transform our ancient genomic data so that it is statistically comparable to modern NGS data before entering the variant calling portion of the pipeline. This method introduces its own host of assumptions and necessarily reduces the number of reads available for analysis.

Slide 09

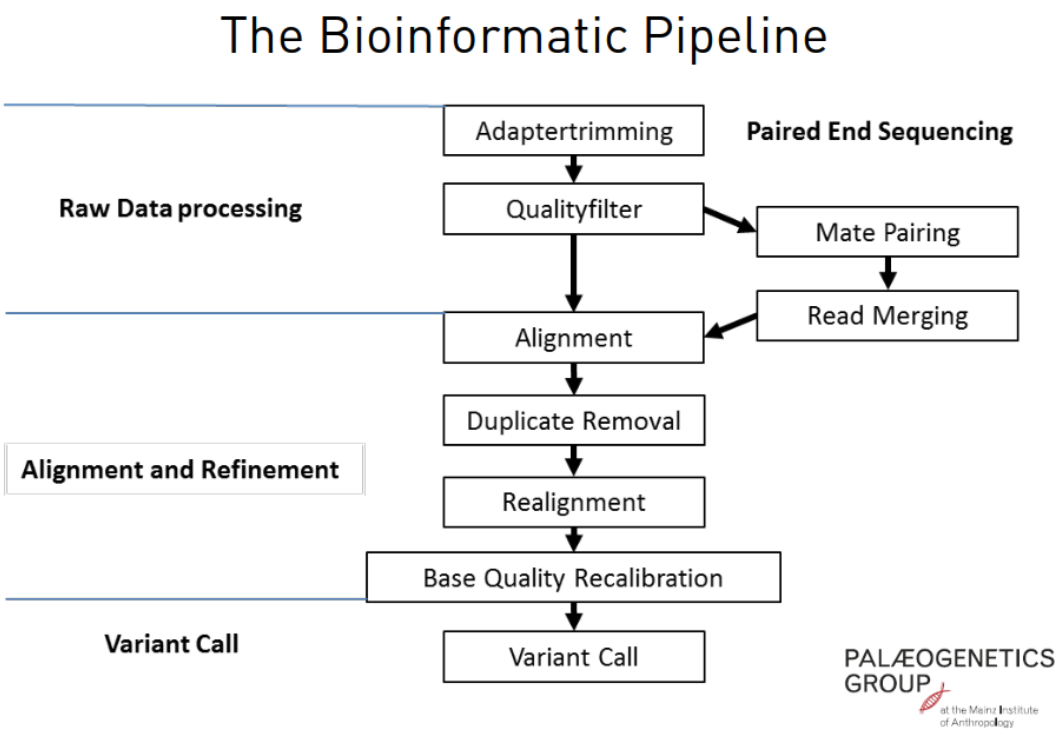

Figure 9: Slide 9

We were also not entirely sure of the mapping between the variation remaining in our filtered aDNA data set and the true variation extant in the ancient individuals or populations.

Slide 10

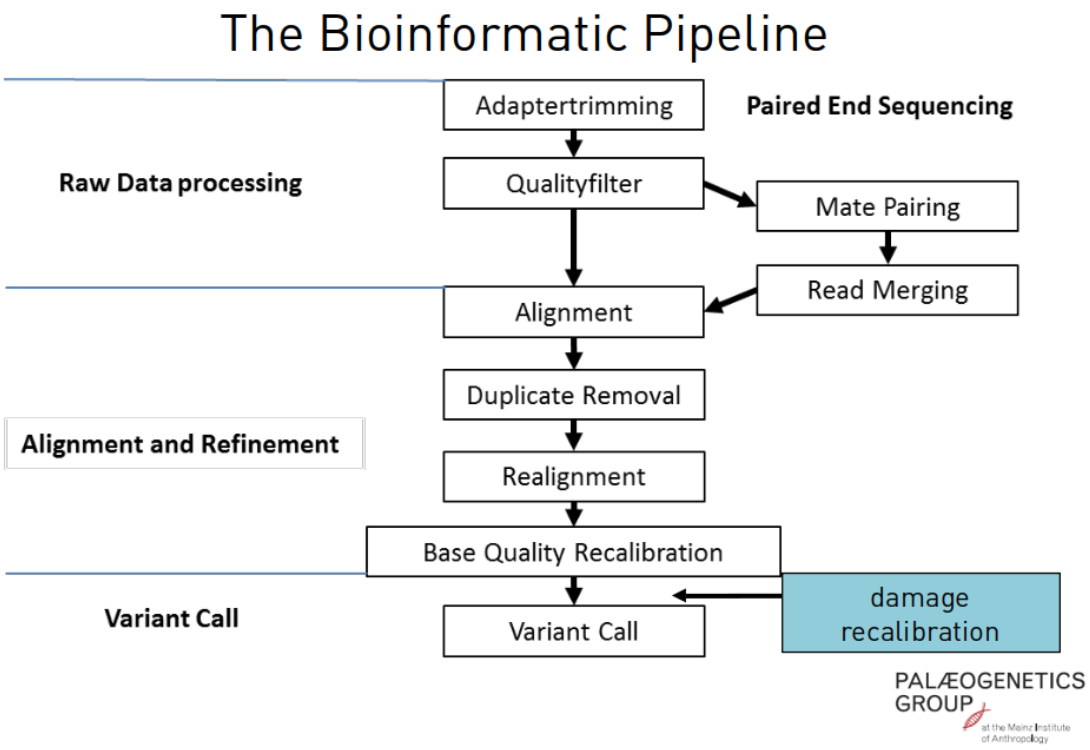

Figure 10: Slide 10

INSTEAD, we have decided to focus on the unique characteristics of ancient DNA itself to develop appropriate pipelines and priors for our analyses. Specifically, we are looking at improving damage recalibration and variant calling algorithms. I would like to specifically discuss base quality recalibration today.