Karola Kirsanow

December 2, 2019

I regularly present my work to different audiences. These are the 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 [note that a SNP is a position in the genome that can differ between individuals].

Bioinformatic challenges associated with ancient DNA

Title Slide

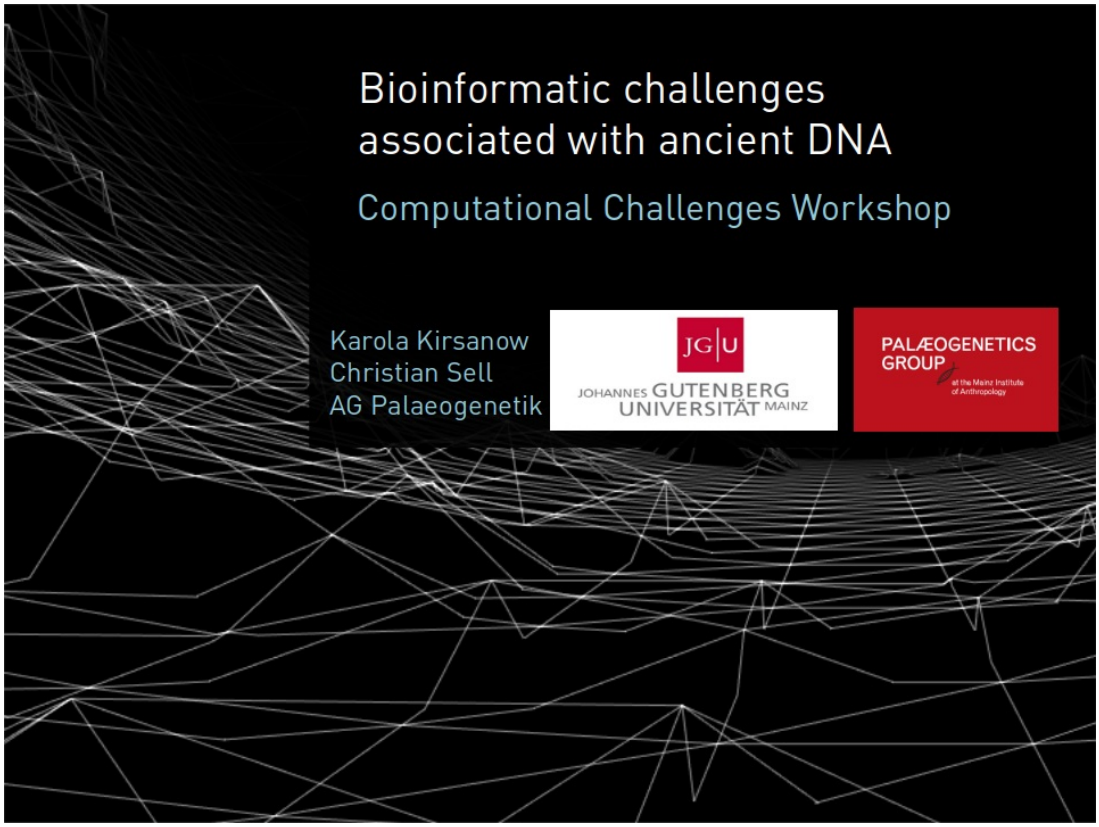


Figure 1: Title slide

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

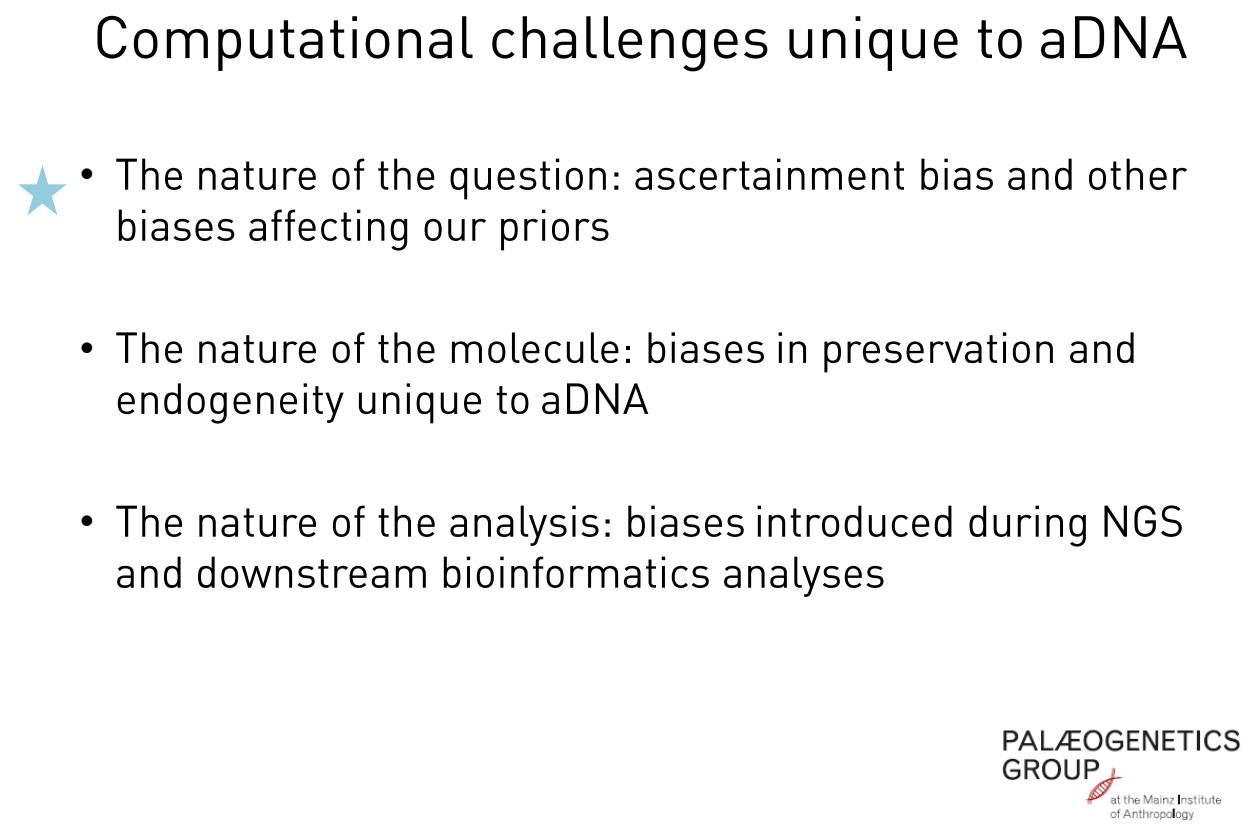


Figure 2: Slide 2

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

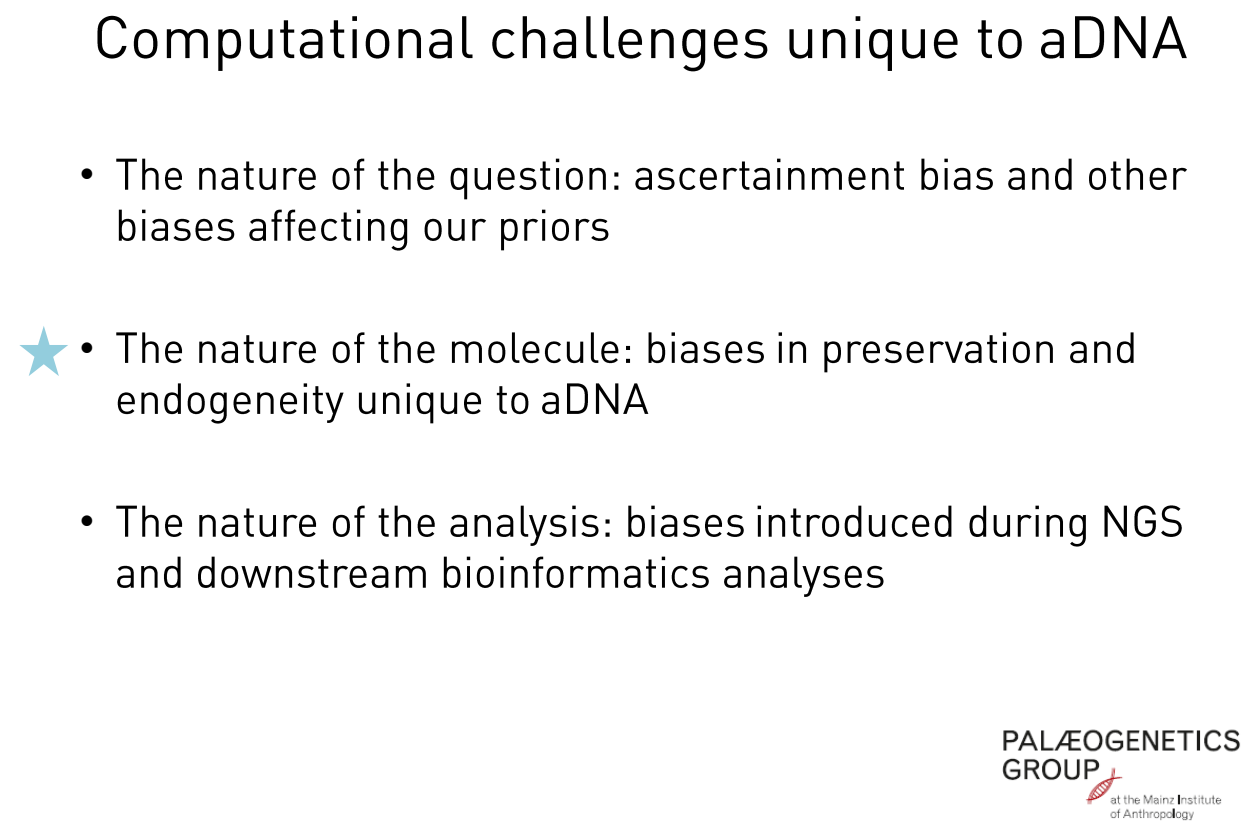


Figure 3: Slide 3

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

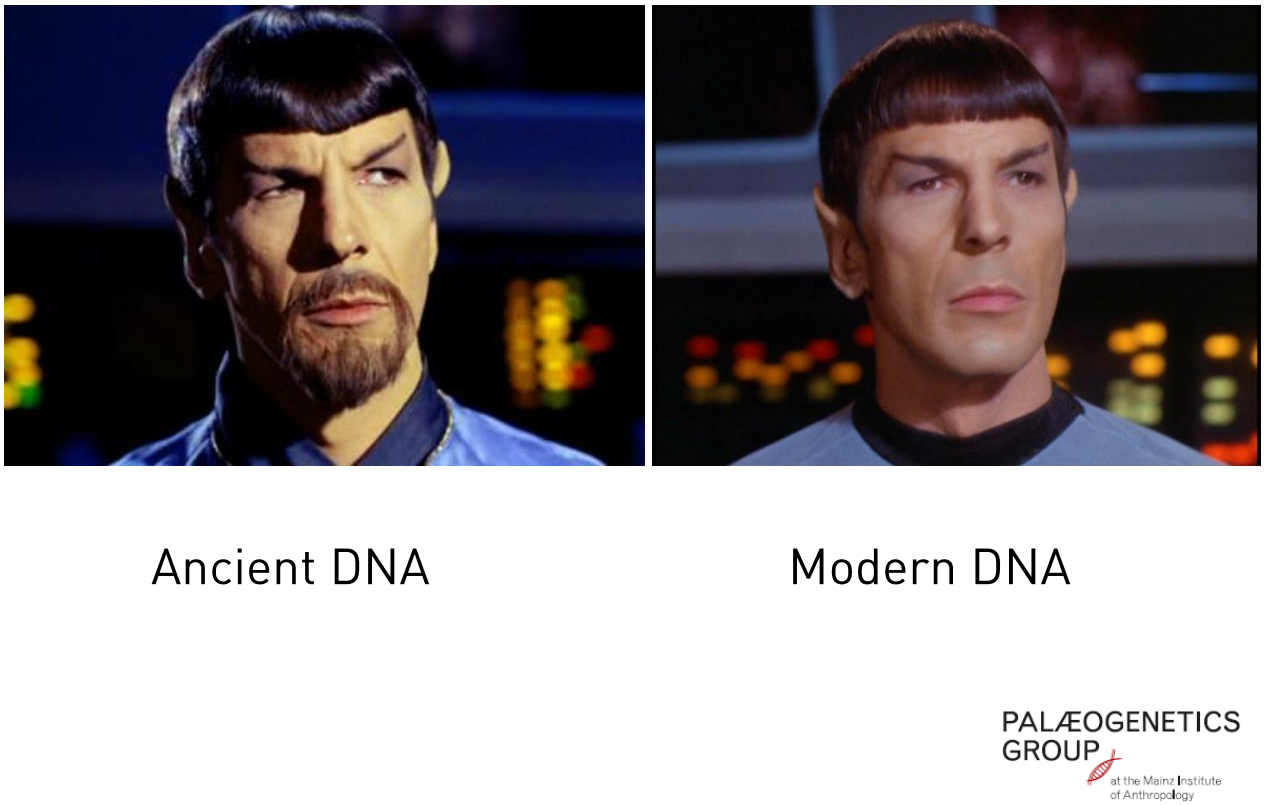


Figure 4: Slide 4

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

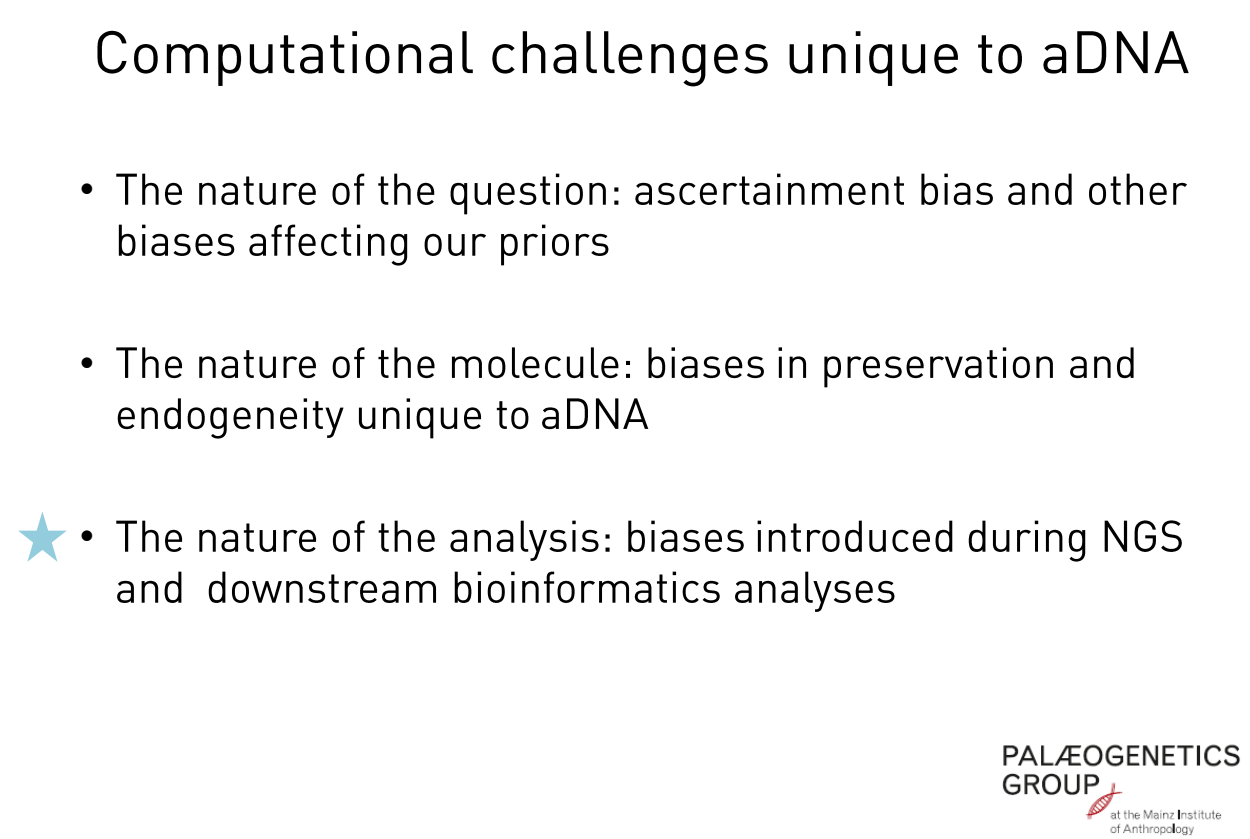


Figure 5: Slide 5

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

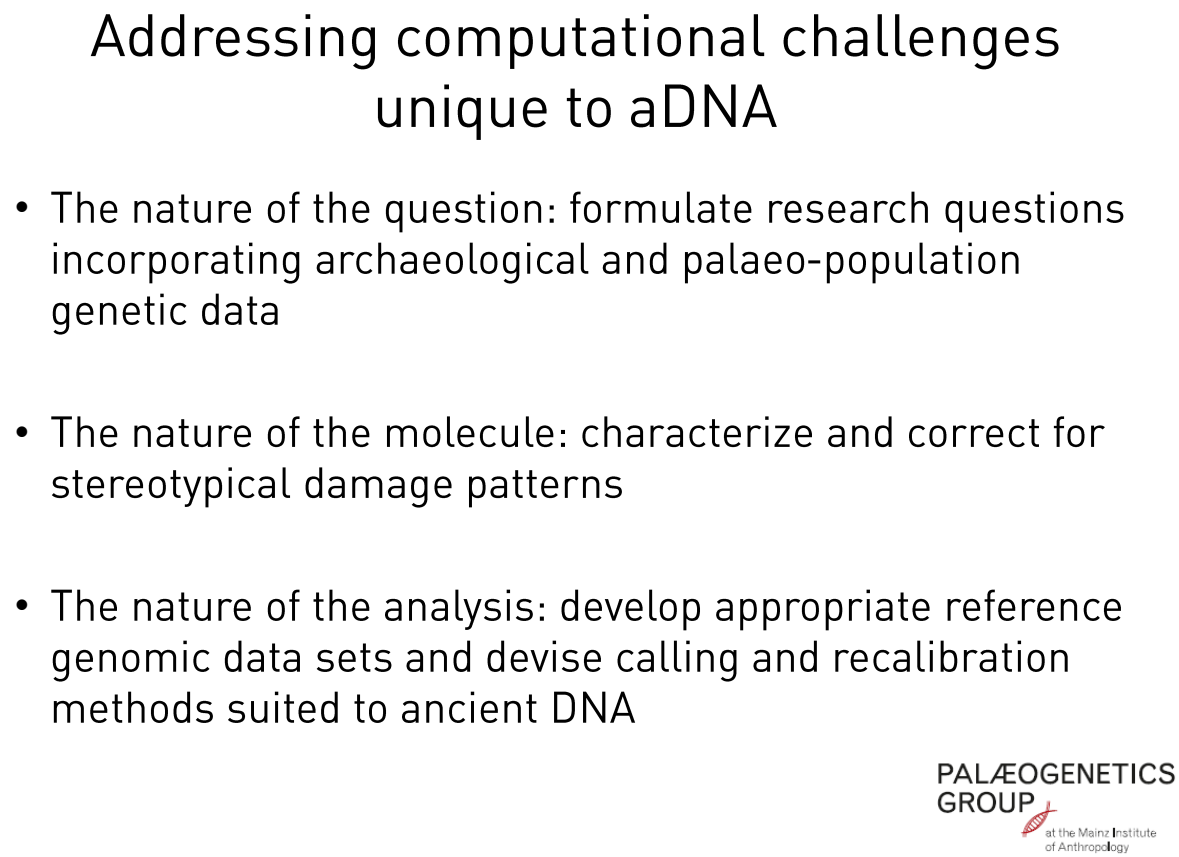


Figure 6: Slide 6

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

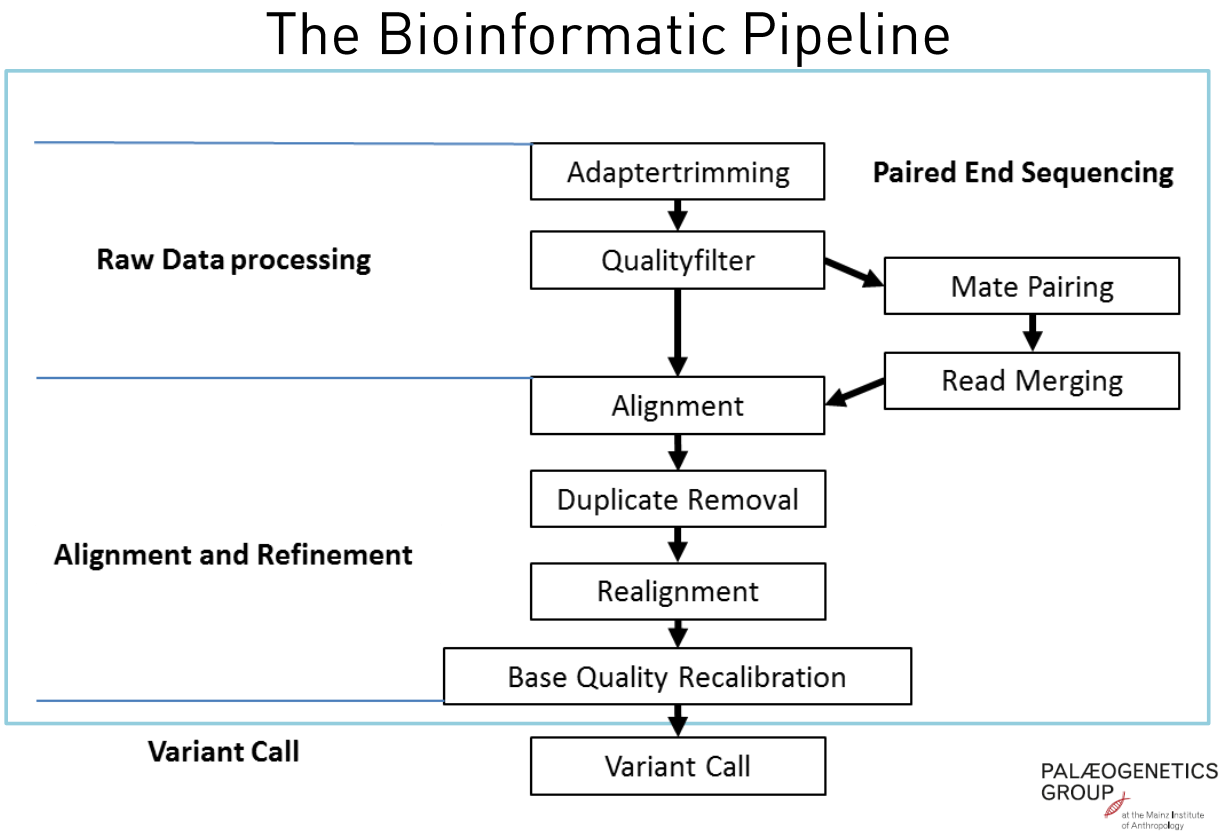


Figure 7: Slide 7

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

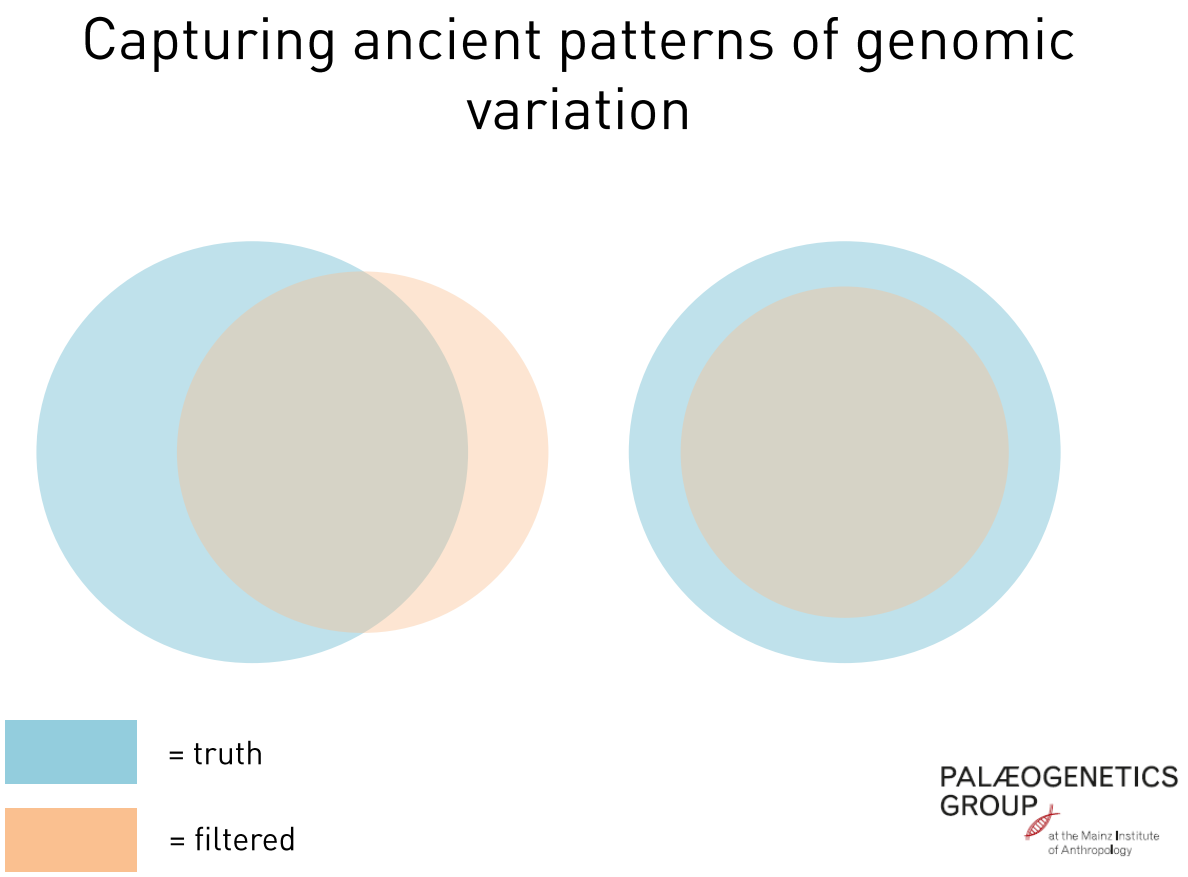


Figure 8: Slide 8

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

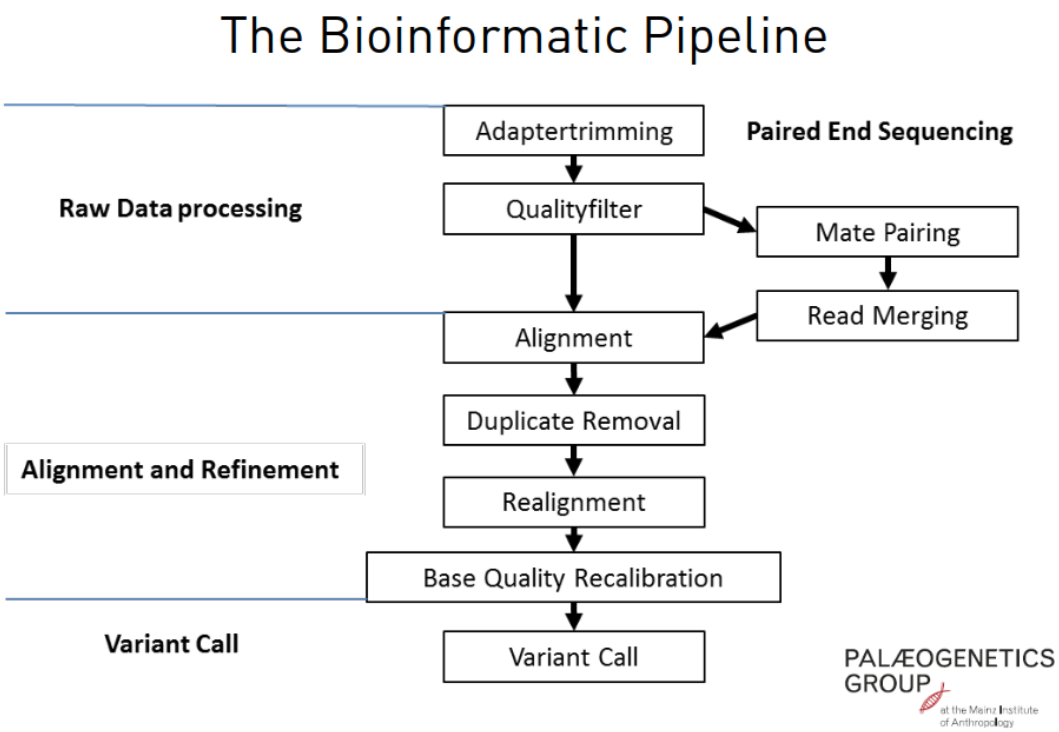


Figure 9: Slide 9

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

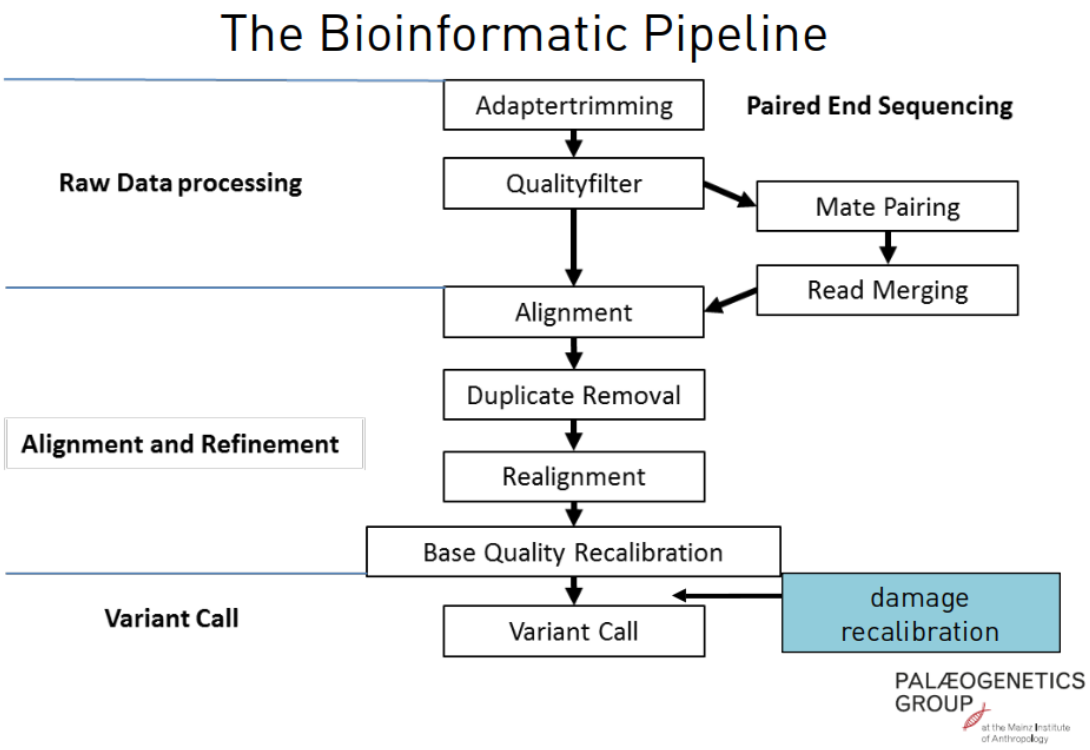


Figure 10: Slide 10

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. Today I would like to discuss how we evaluated the available methods for aDNA base quality recalibration.

Slide 11

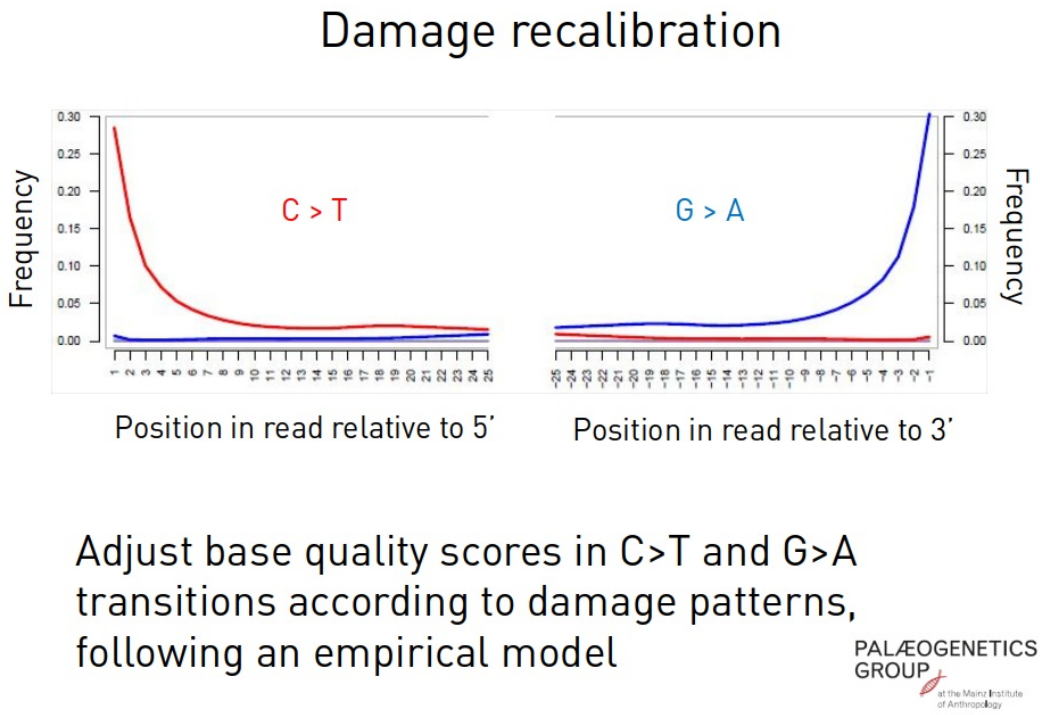


Figure 11: Slide 11

aDNA damage is characterized by a consistent pattern of post-mortem de-aminations which mimic genuine A>G and C>T transition mutations, and which occur at predictable relative positions along a read. Our damage-correction script is based on an empirical model describing the distribution of damage throughout the read, and adjusting base quality scores relative to the likelihood of damage at each site.

Slide 12

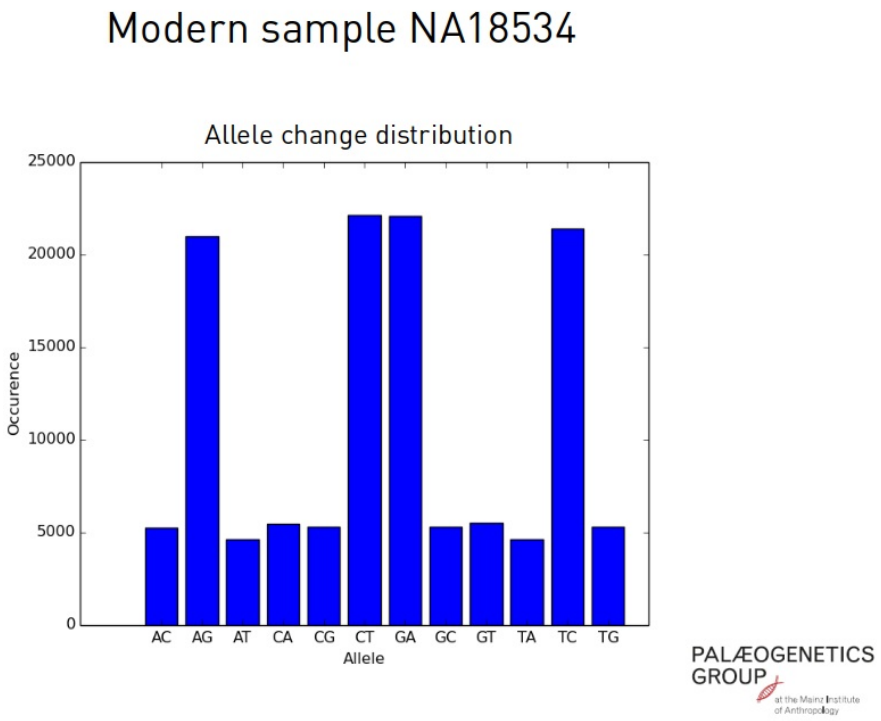


Figure 12: Slide 12

We can use modern data from 1000 genomes to develop an initial idea about what the genomic distribution of mutations should look like. We can expect that the relative numbers of the different *bona fide* transition and transversion mutations (in contrast to allele changes induced by post mortem damage) should be similar between ancient and modern samples.

So here we have a modern British individual...

Slide 13

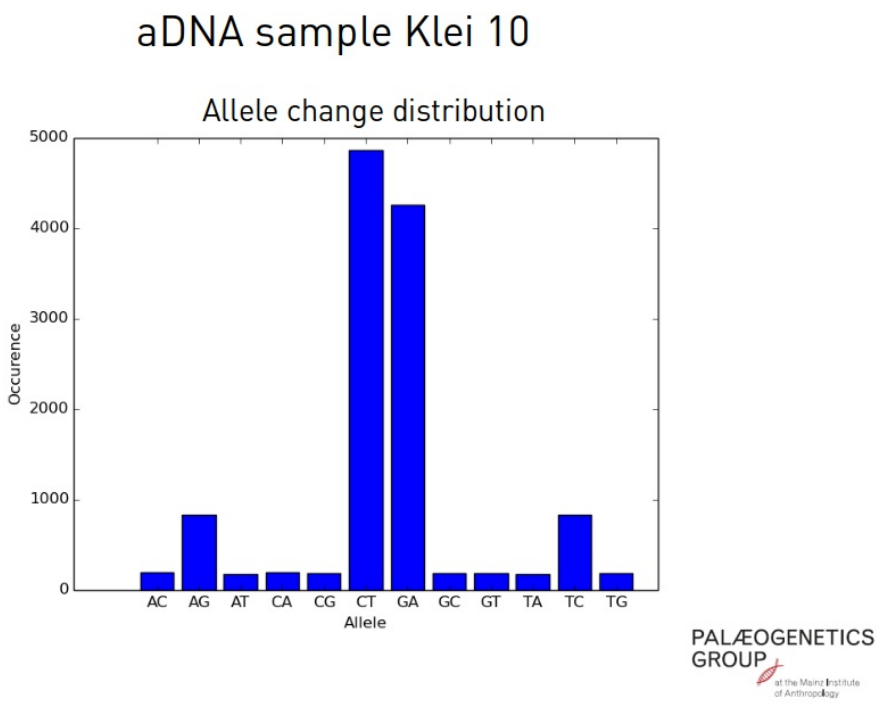


Figure 13: Slide 13

... And here we have the same distribution for a Neolithic individual from Greece, about 6000 years before present. Note the sharply elevated numbers of C>T and G> A transitions: this is precisely what we would expect to find in a sample affected by postmortem deamination reactions.

Slide 14

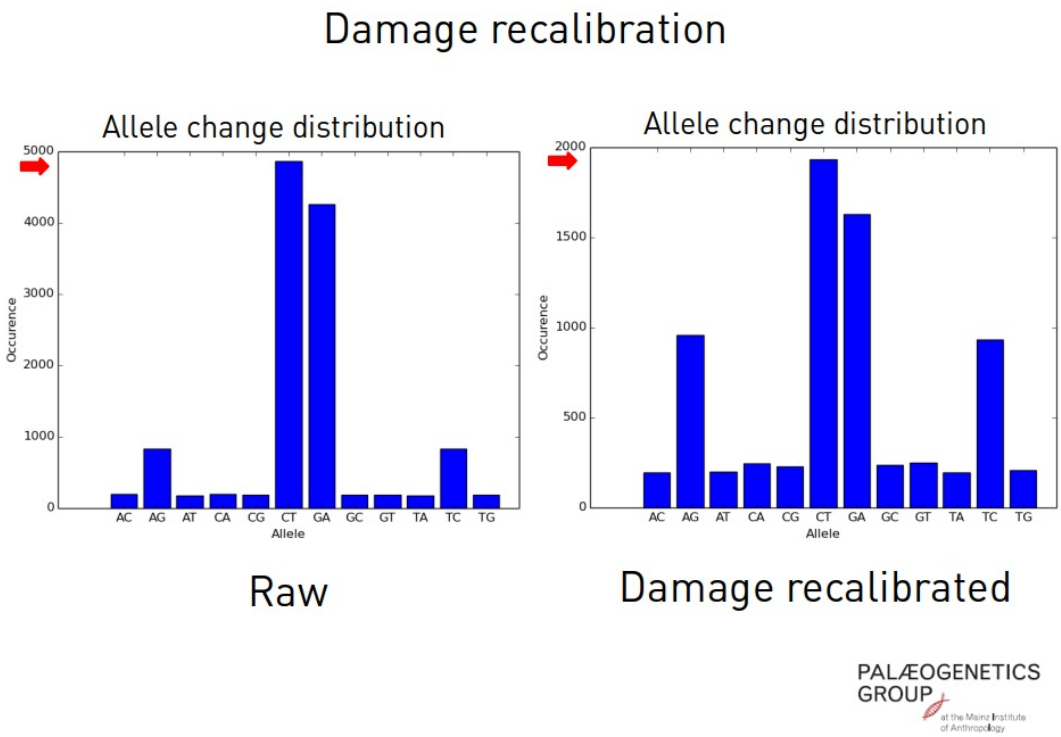
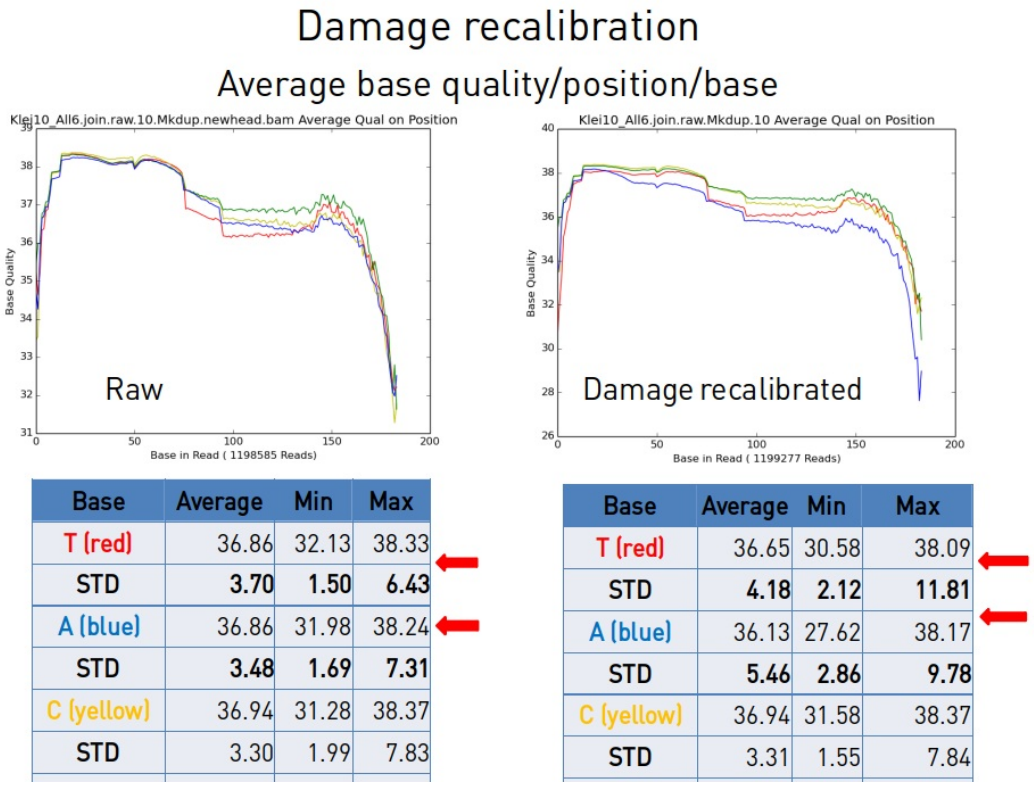


Figure 14: Slide 14

When we apply our damage correction script, we can see that the overdominance of C>T and A>G allele changes is reduced, along with the total number of identified variants.

Slide 15



Base	Average	Min	Max
T (red)	36.86	32.13	38.33
STD	3.70	1.50	6.43
A (blue)	36.86	31.98	38.24
STD	3.48	1.69	7.31
C (yellow)	36.94	31.28	38.37
STD	3.30	1.99	7.83

Base	Average	Min	Max
T (red)	36.65	30.58	38.09
STD	4.18	2.12	11.81
A (blue)	36.13	27.62	38.17
STD	5.46	2.86	9.78
C (yellow)	36.94	31.58	38.37
STD	3.31	1.55	7.84

Figure 15: Slide 15

In terms of average base quality per position per base, our damage recalibration appears to be effectively implementing the damage model developed from observing actual damage patterns in ancient DNA: the base-quality- score variance observed on the alleles most likely to be endpoints from de-amination reactions (A T) is greater as base qualities are adjusted throughout the read. The base quality curve is depressed and smoothed overall.

Slide 16

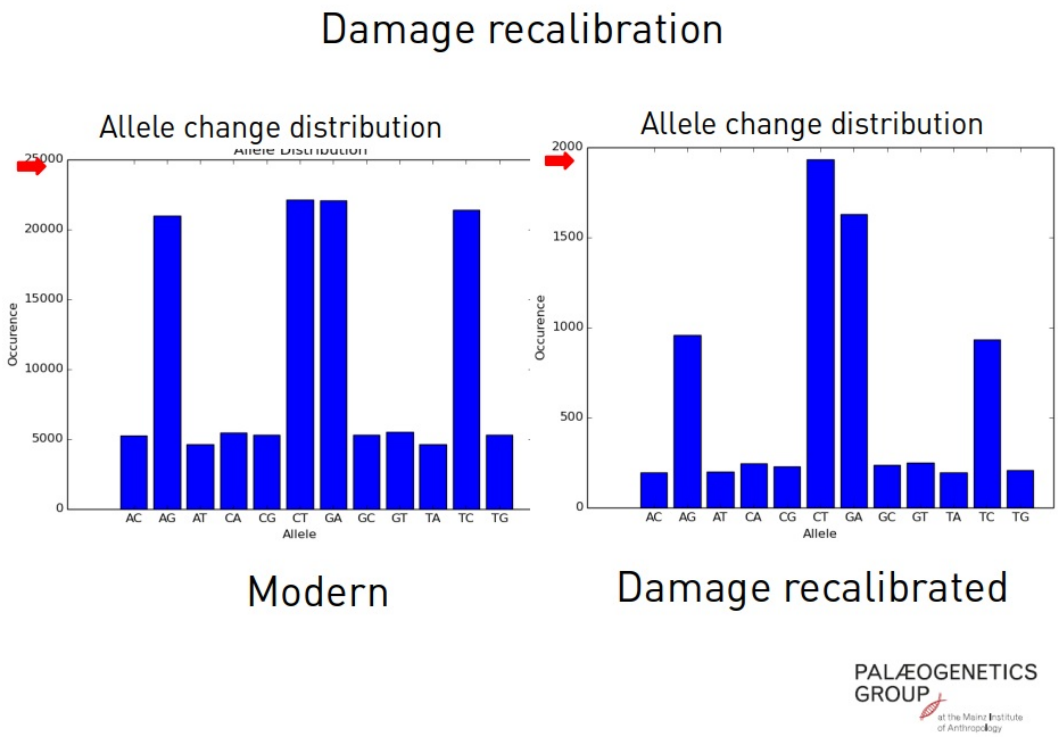


Figure 16: Slide 16

But we are still differentiated from the modern sample (note that the difference in the total number of variants modern/ancient is exaggerated here: the ancient Greek sample has fewer sites present in this particular analysis)

Slide 17

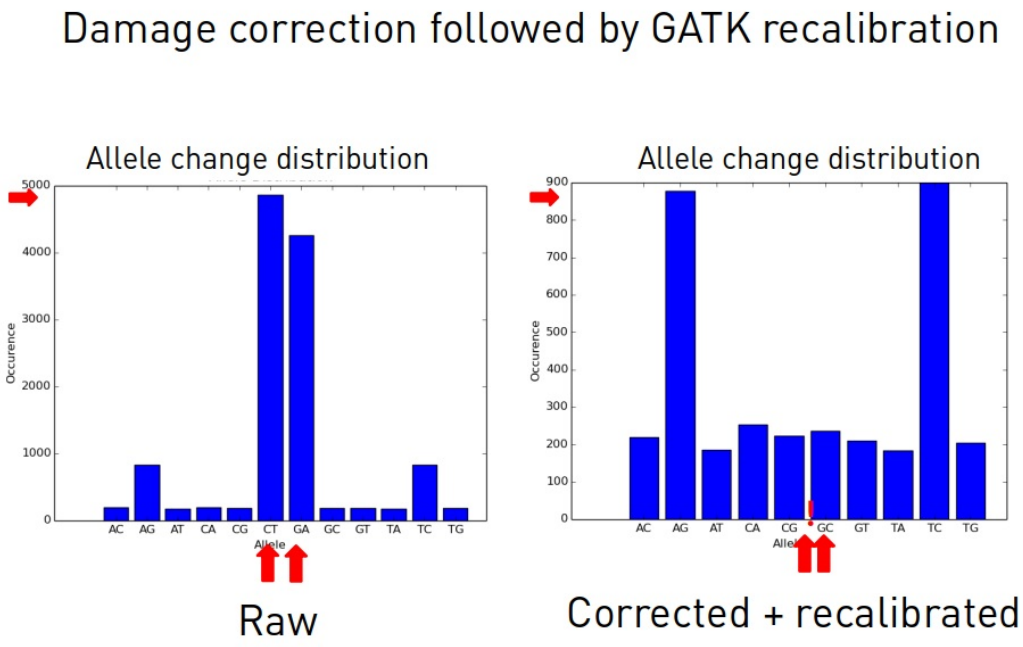


Figure 17: Slide 17

Our testing has shown that applying a damage recalibration script prior to GATK realignment and base quality recalibration over-corrects the damage patterns in our ancient genomes, resulting in huge losses of C>T and G>A mutations relative to SNP calls made on alignments processed first by GATK before applying the damage recalibration script.

So we now know that GATK 'learns' from our damage-correction adjustment, and compounds it by further adjusting quality scores downward until it kicks out most transitions.

Slide 18

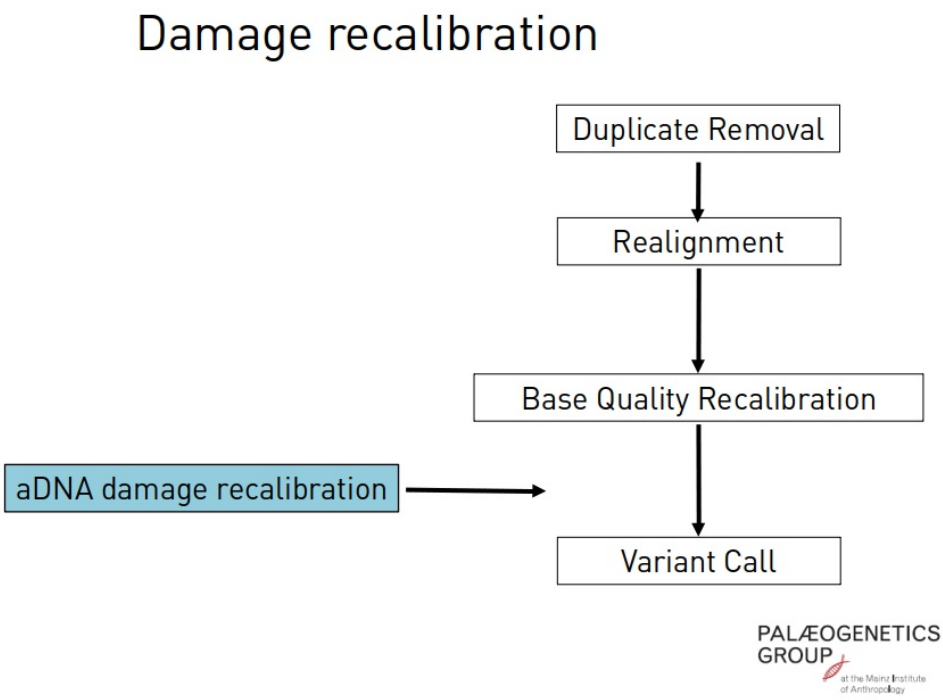


Figure 18: Slide 18

We have therefore concluded that the damage correction script should be the last step before variant calling.

Slide 19

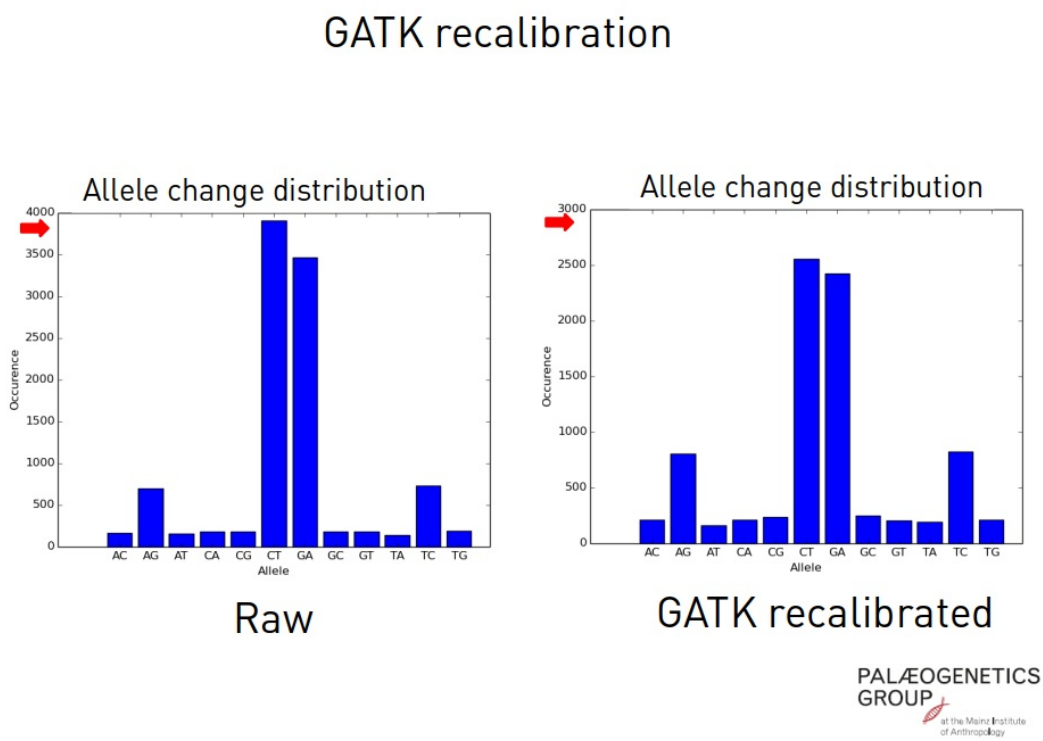


Figure 19: Slide 19

Let's discuss the step immediately prior to the damage recalibration step, where we use the GATK base quality score recalibration tool developed by the Broad Institute.

The BQSR program walks over the reads in a BAM file and tabulates data about each base such as the read group the read belongs to, its assigned quality score, the machine cycle producing the base, and the states of the current base + previous base (the dinucleotide)

BQSR creates bins for each category of data, and for nonvariant sites within within each bin, counts the number of bases and how often they mismatch the reference base. This model is then use to calibrate the base quality scores for all bases in the BAM according to their particular characteristics.

Slide 20

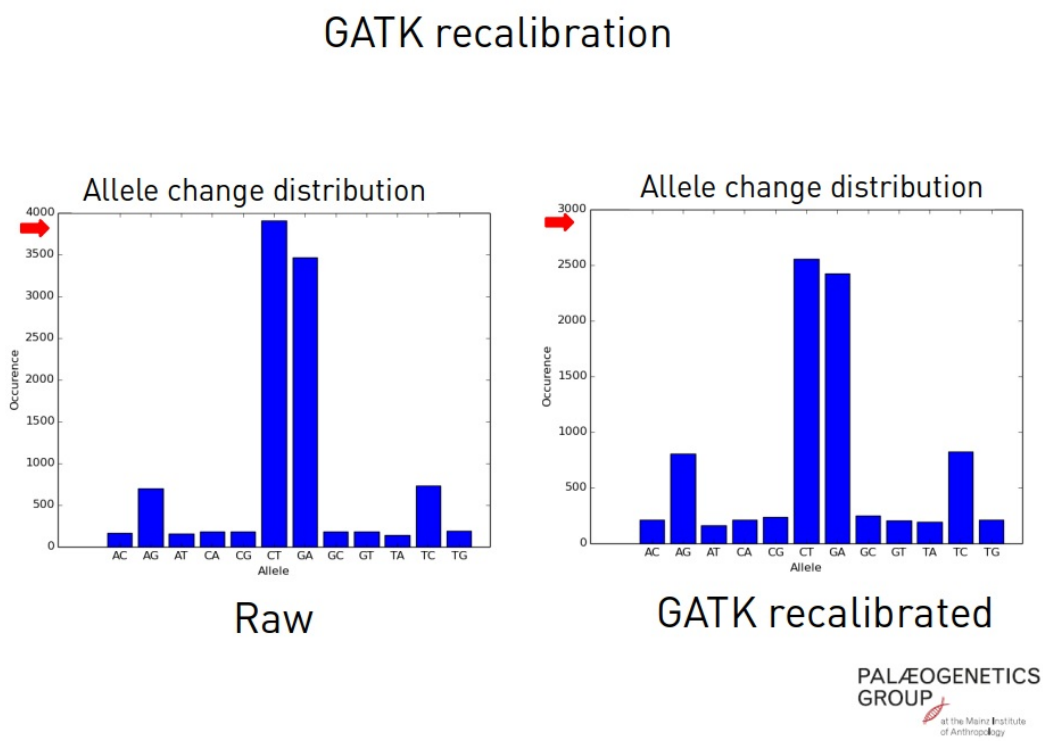
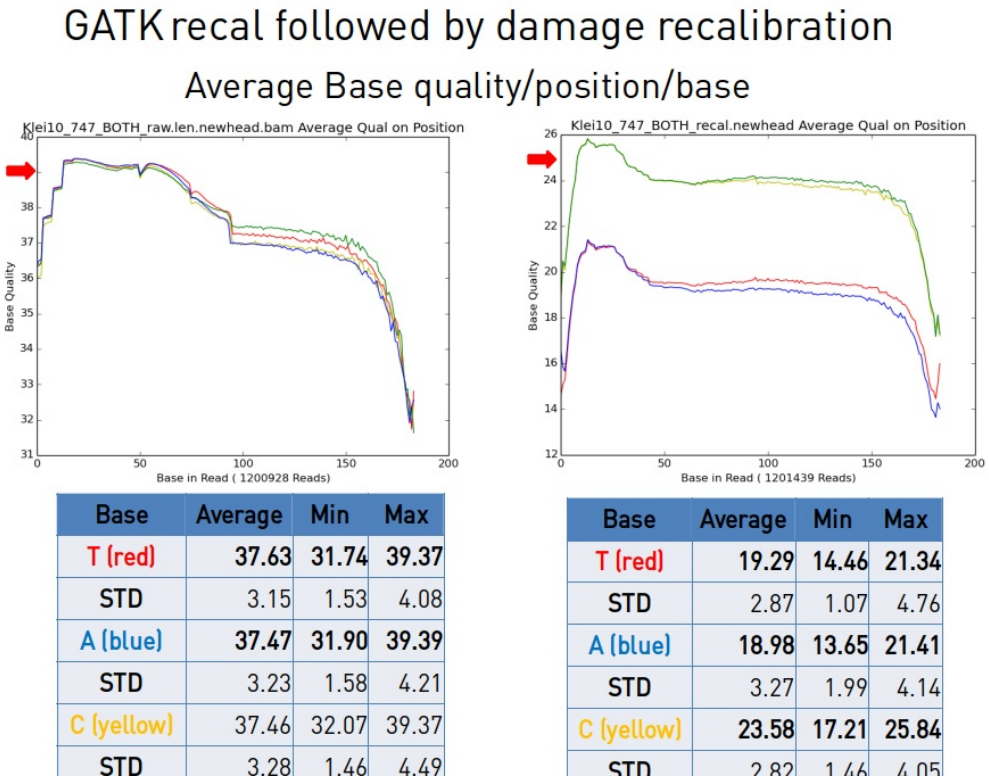


Figure 20: Slide 19

When we apply our damage correction script at the appropriate point in our pipeline— after a first pass at base quality recalibration by GATK— we find an allele change distribution similar to that observed in modern DNA. At this point we have discarded about a third of the observed variant positions as likely to be compromised. Any further recalibration at this point would probably be over-extending our current knowledge of external factors compromising aDNA.

Slide 21



Base	Average	Min	Max
T (red)	37.63	31.74	39.37
STD	3.15	1.53	4.08
A (blue)	37.47	31.90	39.39
STD	3.23	1.58	4.21
C (yellow)	37.46	32.07	39.37
STD	3.28	1.46	4.49

Base	Average	Min	Max
T (red)	19.29	14.46	21.34
STD	2.87	1.07	4.76
A (blue)	18.98	13.65	21.41
STD	3.27	1.99	4.14
C (yellow)	23.58	17.21	25.84
STD	2.82	1.46	4.05

Figure 21: Slide 21

Here we can see how the combination of damage correction and GATK base recalibration deflates the qualities assigned to A and T bases.

Slide 22

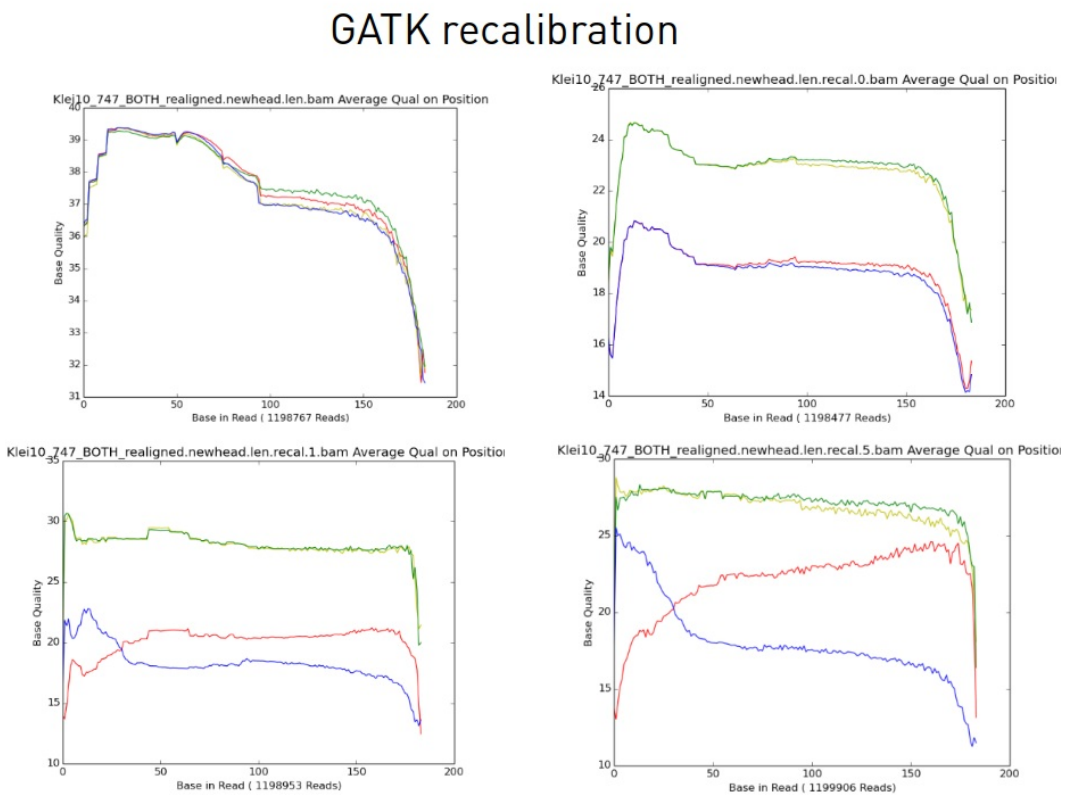


Figure 22: Slide 22

Self-recalibration quality distribution: If we iteratively recalibrate corrected files,we can see how GATK learns more about the damage pattern in each iteration —particularly in the last iteration, where A + T qualities are generally lower than those of C + G; and T bases increase in quality from a trough at the 5' end, while A bases reverse the pattern from the 3' end.

Slide 23

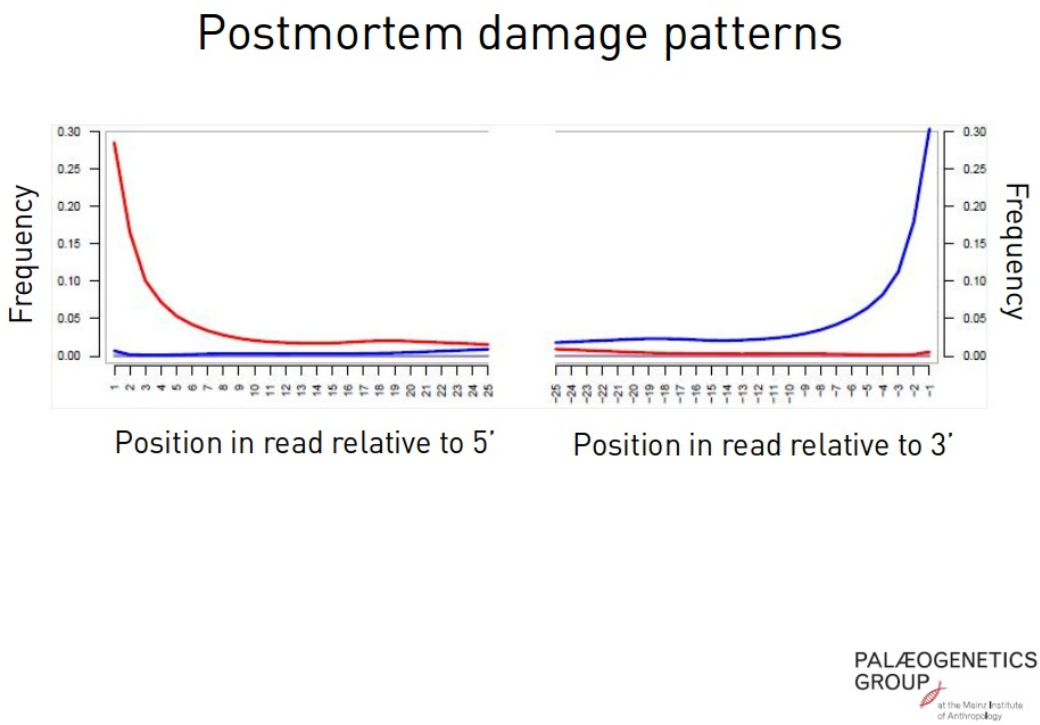


Figure 23: Slide 23

...WHICH IS WHAT WE WOULD EXPECT when correcting for statistically-validated patterns of post-mortem damage.

Slide 24

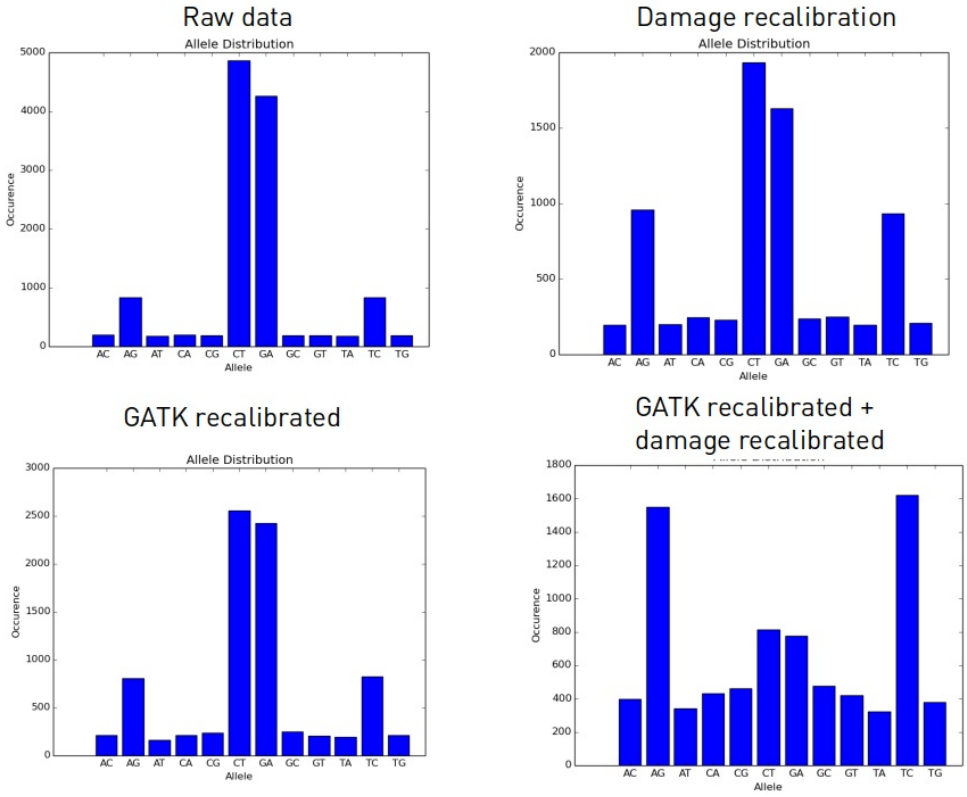


Figure 24: Slide 24

HOWEVER, as this 4-way comparison illustrates, we can still observe an over-representation of C>T and G>A changes: each recalibration method reduces those two mutation types almost exclusively.

Each method is insufficient in isolation, and the combination of both calibration methods introduces a bias toward the overrepresentation of A>G and T>C mutations.

Slide 25

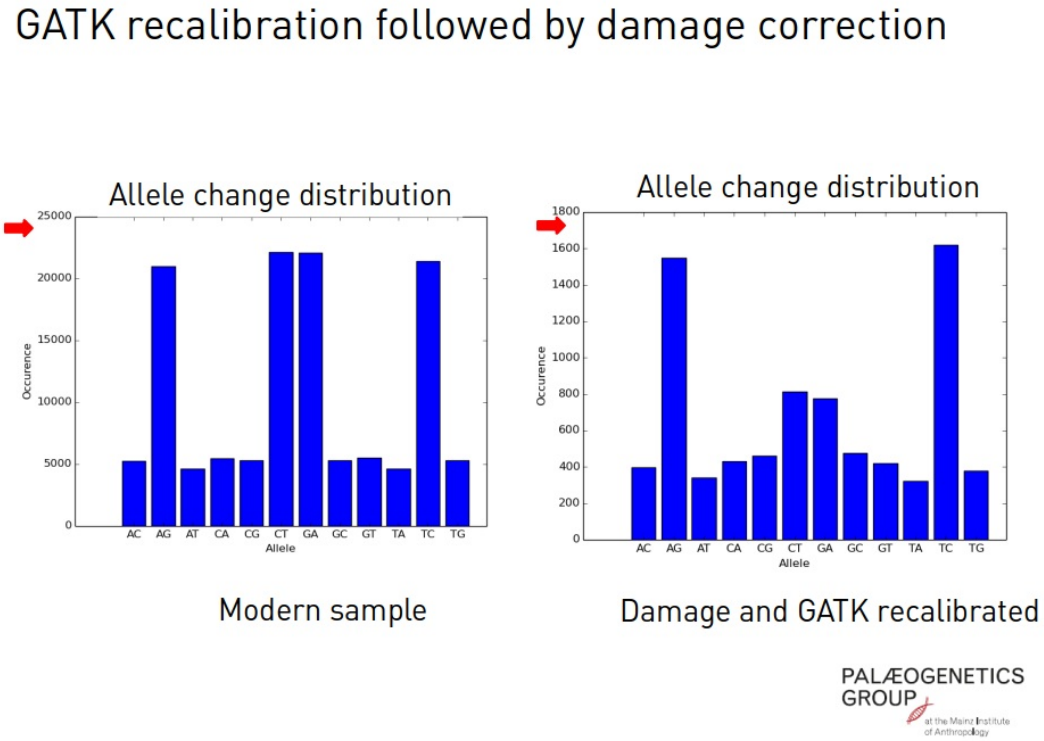


Figure 25: Slide 25

A DIRECT COMPARISON of a modern and an ancient sample demonstrates that damage correction plus GATK recalibration still over-corrects for C>T and G> A allele changes (again ignoring the total number of variants).

Slide 26

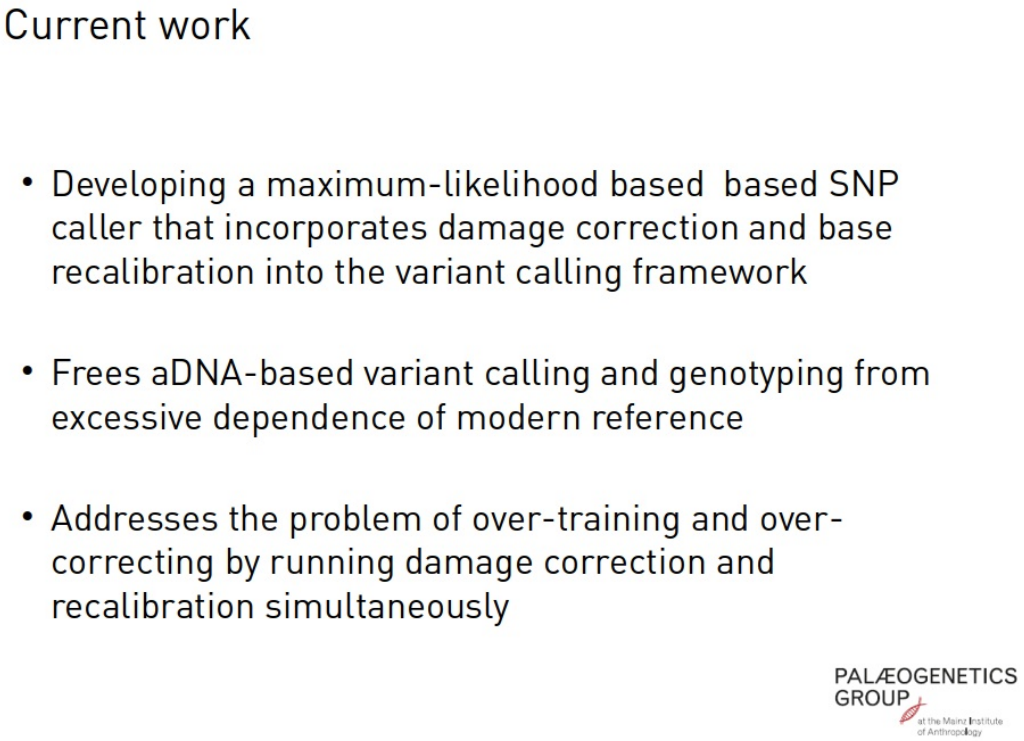


Figure 26: Slide 26

It's clear from this data that assembling our damage recalibration pipeline from off-the-rack methods is a suboptimal solution.

Instead, we are working with our collaborators at Stony Brook University in New York and the University of Fribourg in Switzerland on a new SNP caller designed specifically for the peculiar properties of aDNA.

Postscript

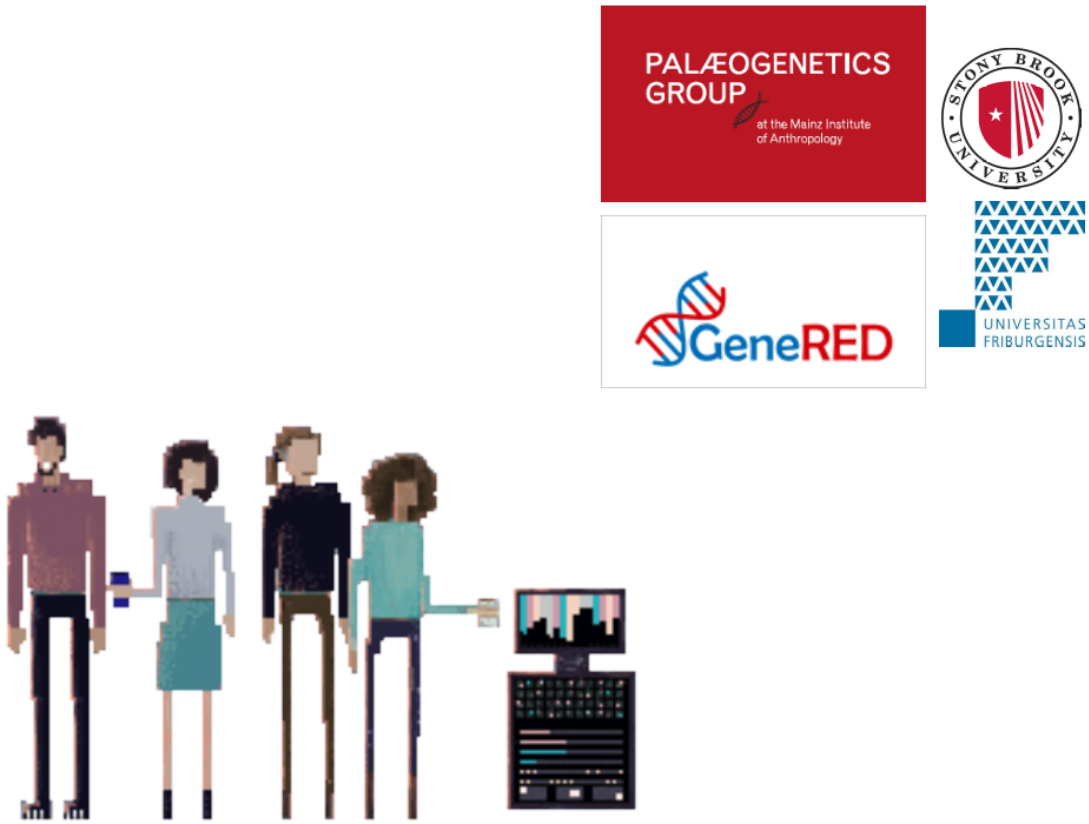


Figure 27: Postscript

The SNP-calling program described in this talk has been completed and successfully implemented on several very interesting ancient DNA datasets, delivering important contributions to our understanding of palaeo-population genetics.

I like this talk because it is an example of communicating with a specialist audience in an adjacent discipline. This can be more challenging that communicating research to a lay audience, because adjacent-experts often harbor unchecked assumptions coloring their perception of a problem.