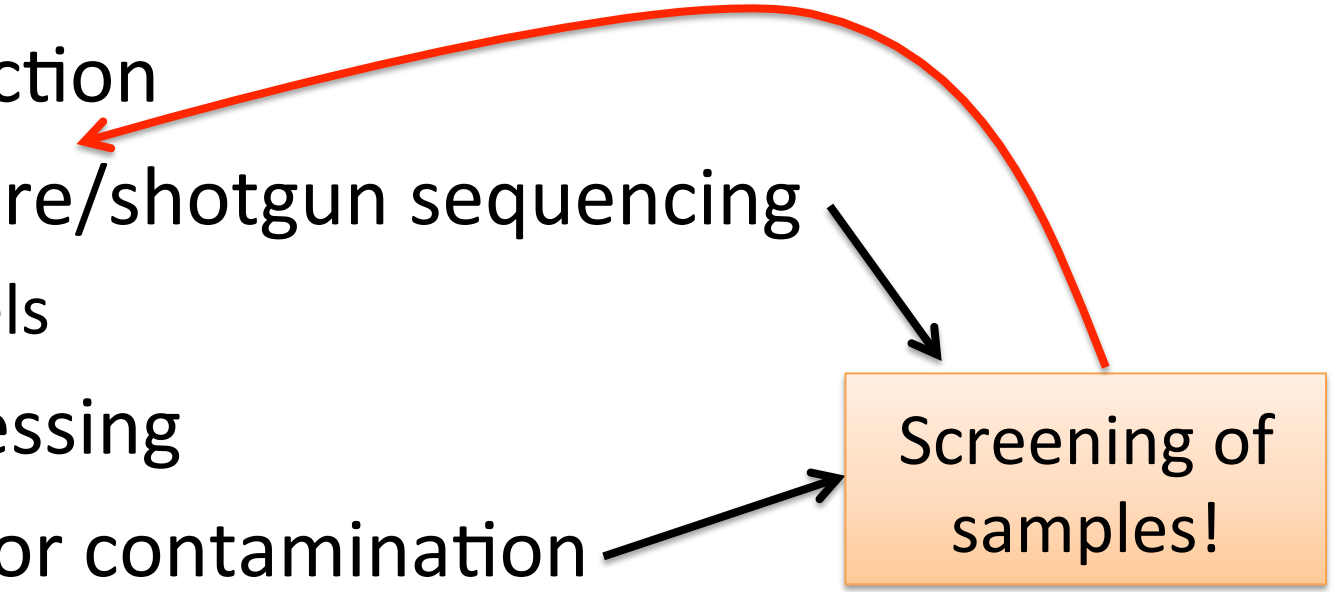
Lesson 3: From bones to (pseudohaploid) genotypes

Wednesday July 11: 2:00 – 4:30 pm

Many parts leading up to dataset used for analysis

- Sampling of bone material
- DNA extraction
- DNA capture/shotgun sequencing
  - SNP Panels
- Data Processing
- Checking for contamination
- Final dataset!



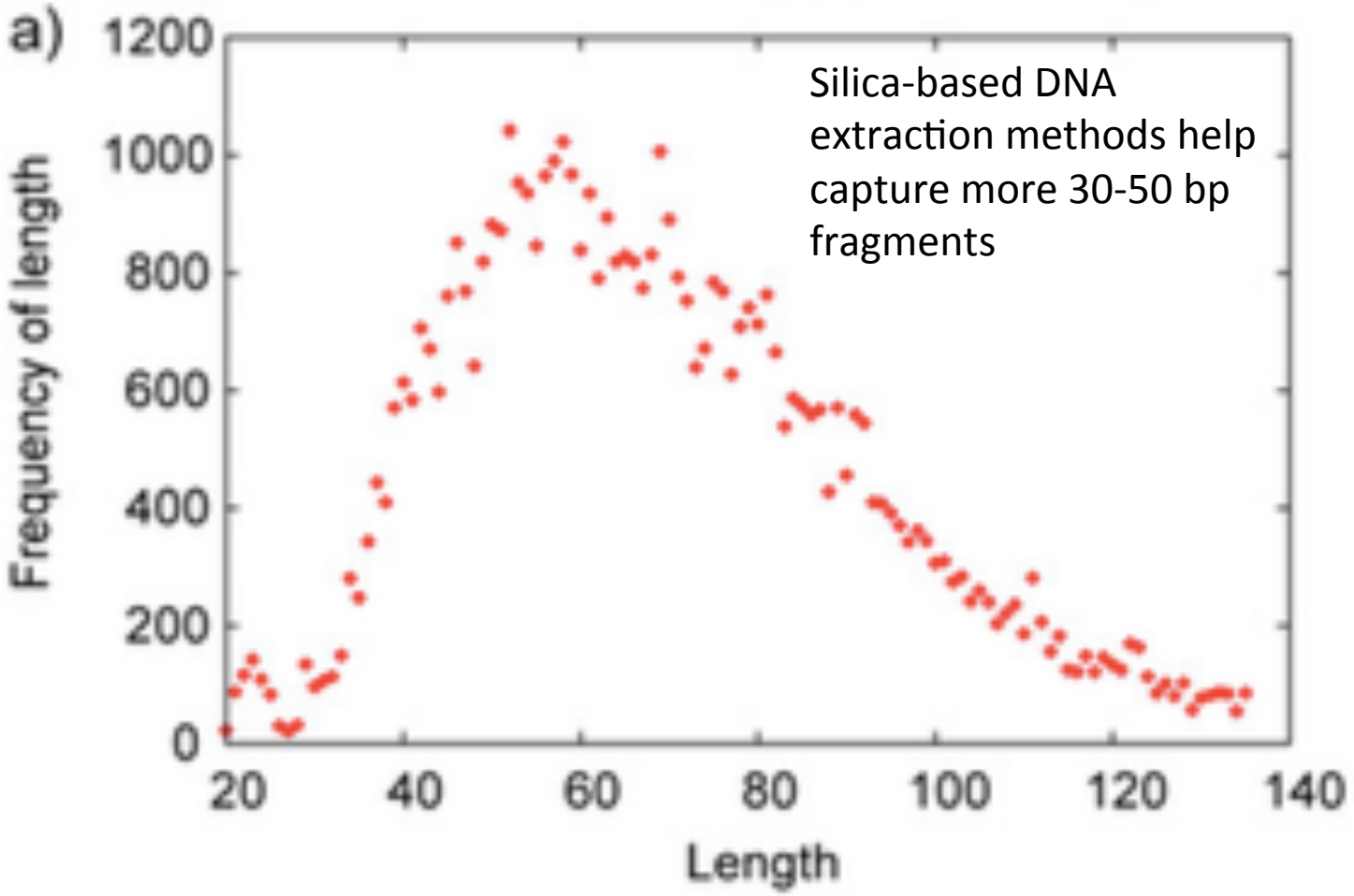
Bone material

- Mostly from bony material
- Best preservation, lowest microbial seems to be in petrous bone, so preferred choice
- Strict lab procedures to avoid contamination, damage as little of skeletal remains as possible, and not waste materials (very little to begin with!)

Complications of aDNA

- Short fragment length
- Ancient DNA Damage
- Environmental microbial DNA
- Modern human contamination

Short fragment length



Fu et al. 2013, Figure S1A

Fragmentation occurs more often at purines (R)

Damage more likely at overhangs (single-stranded parts)

Ancient DNA Damage

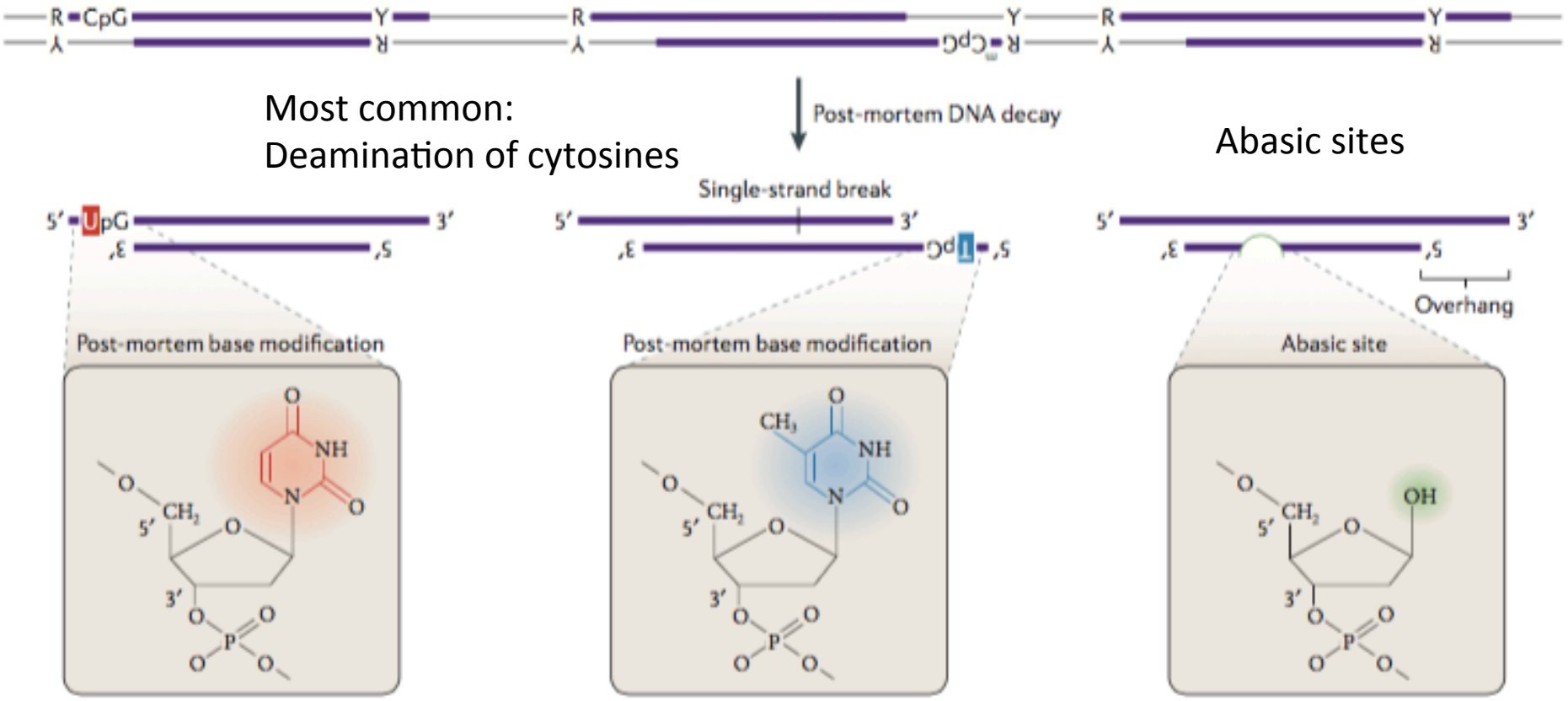
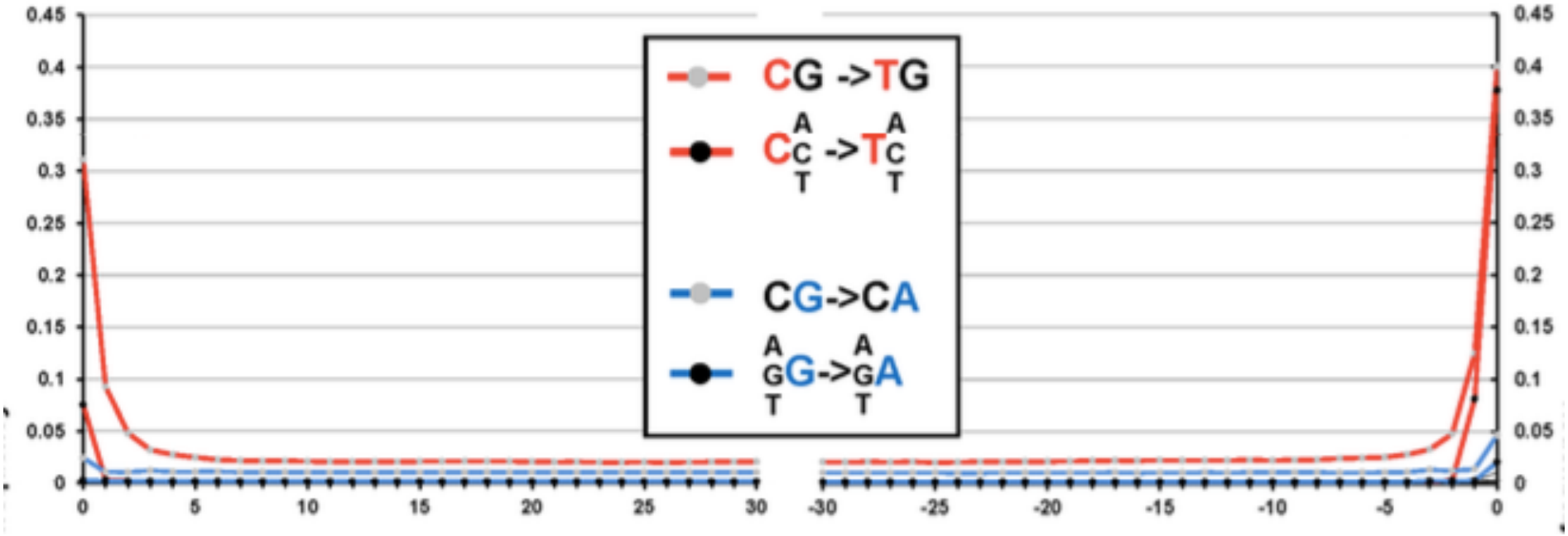


Figure 2 | **Typical ancient DNA molecules.** A diverse range of degradation reactions affect DNA post-mortem and result in extensive fragmentation (preferentially at purine nucleotides) and base modifications. The most common base modification identified in high-throughput sequencing data sets is deamination of cytosines into uracils (red), or thymines (blue) when cytosines were methylated (^{m}C). Such deaminations occur much faster at overhanging ends. Other modifications include abasic sites (green) and single-strand breaks (vertical lines). The chemical structures of three damage by-products (uracils, thymines and abasic sites) are shown. R, purine; Y, pyrimidine.

Orlando et al. 2017 Figure 2

Ancient DNA Damage



Meyer et al. 2012, Figure S5

Characteristic high mutation frequency at read ends
See high deamination at fragment termini

Ancient DNA Damage

USER mix:
Uracil DNA glycosylase
Endonuclease VIII

- Replaces deaminated cytosines with abasic sites and cleaves out abasic sites
- Full USER-treatment – no damage left
- Can also remove DNA damage except in some number of first and last positions of the reads, so possible to test for presence of characteristic DNA damage patterns, while making DNA libraries robust for popgen analysis

Transversions Only:
Remove transitions
A<->G, C<->T

- Looking at biallelic sites where know alleles
- Since deamination is a C->(U)T mutation, uncertain whether allele call is truly derived mutation or due to damage for A/G and C/T biallelic sites.

Two main library types

- Double-stranded vs. single-stranded
  - Double-stranded is standard protocol using modern DNA
  - Single-stranded allows access to single-stranded fragments, which provides access to more endogenous material.

Below is the standard protocol currently used in the Fu lab for the differently prepared libraries

Box S2.1. Strategy used to retain damaged fragments for contaminated libraries

ss UDG-treated libraries: Restrict to fragments with C→T substitutions in the first position at the 5'-end and the last two positions at the 3'-end.

ss noUDG-treated libraries: Restrict to fragments with C→T substitutions in the first three positions at the 5'-end and the last three positions bases at the 3'-end.

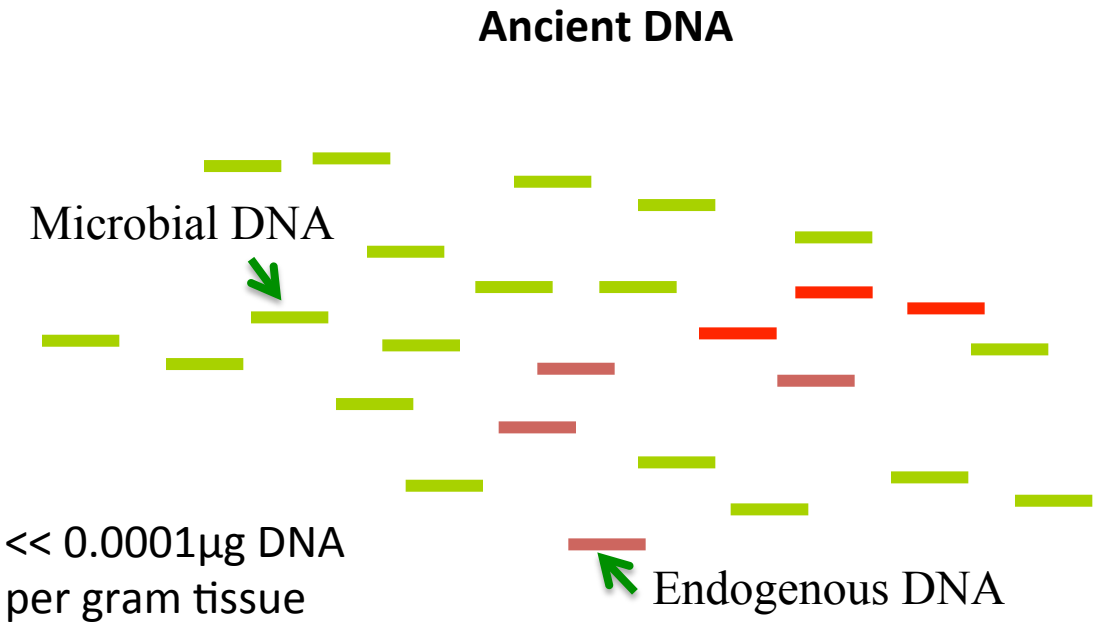
ds UDG- partial treated libraries: Restrict to fragments with C→T substitutions in the first position at the 5'-end and G→A substitutions in the last position at the 3'-end.

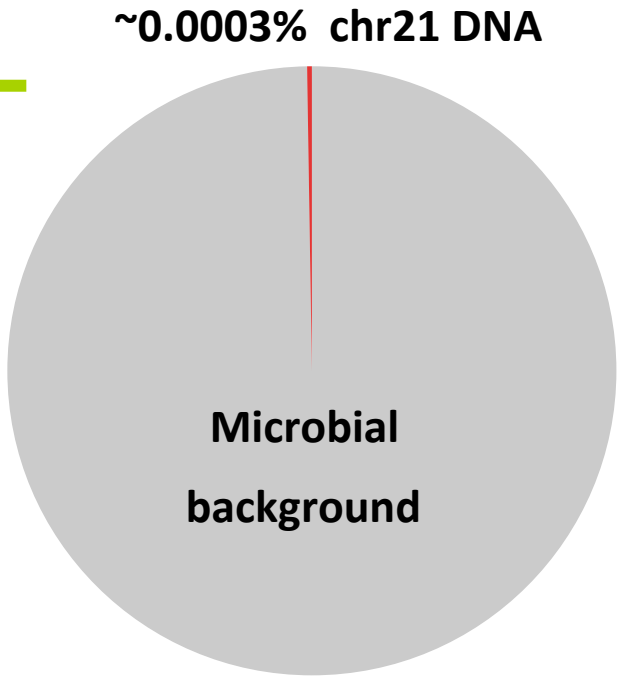
ds noUDG-treated libraries: Restrict to fragments with C→T substitutions in the first three positions at the 5'-end, and G→A substitutions in the last three positions on the 3'-end.

ds UDG-treated libraries: Cannot restrict to damaged fragments so do not use.

Fu et al. 2016, SI 2

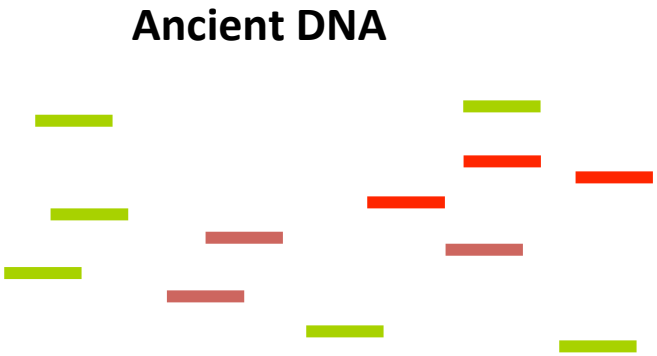
Environmental Microbial DNA





Shotgun sequencing very inexpensive
and waste of resources!

Target-probe hybridization: DNA Capture



<< 0.0001µg DNA per gram tissue

Endogenous DNA

What is being captured?
mtDNA
Bacterial plasmids
Short nuclear loci – SNP panels!

New techniques for whole genome capture

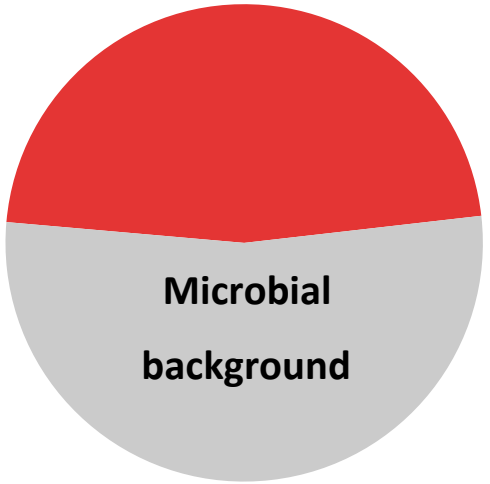
Knowledge of probe
- need to know well, so usually present-day sample
- has been successful using probes that are 10-13% different from target

Customized biotinylated probes
- best for small regions (hundreds of kb to few Mb)

Microarrays
- can hold ~1 million probes

Known Adaptor Sequence
- Can amplify over and over - reusable

46.8% chr21 DNA



Tianyuan, Fu et al. 2013

Common SNP Panels

- Fu et al. (2015, Oase paper)
  - **“390k”: 394,577 SNPs, ~90% on HO array (main set in Haak et al. 2015)**
  - “840k”: 842,630 SNPs – rest on HO array, all SNPs on Illumina 610-Quad Array, all on Affymetrix 50k array, select other SNPs
  - “1000k”: 997,780 SNPs of all transversion polymorphisms in two Yoruba males or in the Altai Neanderthal (with MQ>99%) with chimp alleles available
  - “Archaic”: SNPs where Yoruba (1KG) are high frequency in one allele, with at least one archaic carrying alternative allele (~1.7M SNPs)
- HO array: ~600k SNPs (Patterson et al. 2012)
- 1.2M: ‘390k’ + ‘840k’
- 2.2M: ‘390k’ + ‘840k’ + ’1000k’

bwa → SAMtools → Genotyping

- Major Pipeline
  - Index and align fastq to reference (fasta)
  - Convert resulting SAM file to BAM file, including GQ and MQ filters as needed.
  - Convert BAM file to preferred data format (pseudohaploid calls, EIGENSTRAT?)

- http://www.htslib.org/doc/samtools.html

Relevant File Formats

- FASTA – reference genome
- FASTQ – raw read data, contains base quality scores
- SAM – Sequence Alignment/Map Format – TAB delimited text format showing alignment of reads in FASTQ to reference.
- BAM – Binary form of SAM, use samtools view to see information.

Modern DNA contamination

- Test for DNA damage on reads (before USER treatment)
- Check for multiple allele calls in mtDNA
- Check for multiple allele calls in non-recombining portion of X-chromosome

- Pre-processing: damage-restricted fragments
- Post-processing: likelihood methods
  - Using modern reference panels

Modern DNA contamination

- Checking mtDNA
  - Mapping reads to mtDNA genome
  - ContamMix v1.0-5
    - Fraction fragments matching reconstructed consensus better than pool of worldwide mtDNA sequences is low (<95%)
    - Lower bound of above value <85%
- Checking X-chromosome in males
  - Base discordance between overlapping reads
  - ANGSD software: generates contamination estimate taking into account local sequence or alignment error from included reads.
  - http://popgen.dk/angsd/index.php/Contamination

What’s happening in Haak et al. (2015)?

Haak et al. (2015) – 1st page

We generated genome-wide data from 69 Europeans who lived between 8,000–3,000 years ago by enriching ancient DNA libraries for a target set of almost 400,000 polymorphisms. Enrichment of these positions decreases the sequencing required for genome-wide ancient DNA analysis by a median of around 250-fold, allowing us to study an order of magnitude more individuals than previous studies[1–8] and to obtain new insights about the past. We show that the populations of Western and Far Eastern Europe followed opposite trajectories between 8,000–5,000 years ago. At the beginning of the Neolithic period in Europe, ~8,000–7,000 years ago, closely related groups of early farmers appeared in Germany, Hungary and Spain, different from indigenous hunter-gatherers, whereas Russia was inhabited by a distinctive population of hunter-gatherers with high affinity to a ~24,000-year-old Siberian[6]. By ~6,000–5,000 years ago, farmers throughout much of Europe had more hunter-gatherer ancestry than their predecessors, but in Russia, the Yamnaya steppe herders of this time were descended not only from the preceding eastern European hunter-gatherers, but also from a population of Near Eastern ancestry. Western and Eastern Europe came into contact ~4,500 years ago, as the Late Neolithic Corded Ware people from Germany traced ~75% of their ancestry to the Yamnaya, documenting a massive migration into the heartland of Europe from its eastern periphery. This steppe ancestry persisted in all sampled central Europeans until at least ~3,000 years ago, and is ubiquitous in present-day Europeans. These results provide support for a steppe origin[9] of at least some of the Indo-European languages of Europe.

Genome-wide analysis of ancient DNA has emerged as a transformative technology for studying prehistory, providing information that is comparable in power to archaeology and linguistics. Realizing its promise, however, requires collecting genome-wide data from an adequate number of individuals to characterize population changes over time, which means not only sampling a succession of archaeological cultures[2], but also multiple individuals per culture. To make analysis of large numbers of ancient DNA samples practical, we used in-solution hybridization capture[10,11] to enrich next generation sequencing libraries for a target set of 394,577 single nucleotide polymorphisms (SNPs) ('390k capture'), 354,212 of which are autosomal SNPs that have also been genotyped using the Affymetrix Human Origins array in 2,345 humans from 203 populations[4,12]. This reduces the amount of sequencing required to obtain genome-wide data by a minimum of 45-fold and a median of 262-fold (Supplementary Data 1). This strategy allows us to report genomic scale data on more than twice the number of ancient Eurasians as has been presented in the entire preceding literature[1–8] (Extended Data Table 1).

We used this technology to study population transformations in Europe. We began by preparing 212 DNA libraries from 119 ancient samples in dedicated clean rooms, and testing these by light shotgun sequencing and mitochondrial genome capture (Supplementary Information section 1, Supplementary Data 1). We restricted the analysis to libraries with molecular signatures of authentic ancient DNA (elevated damage in the terminal nucleotide), negligible evidence of contamination based on mismatches to the mitochondrial consensus[13] and, where available, a mitochondrial DNA haplogroup that matched previous results using PCR[4,14,15] (Supplementary Information section 2). For 123 libraries prepared in the presence of uracil-DNA-glycosylase[16] to reduce errors due to ancient DNA damage[17], we performed 390k capture, carried out paired-end sequencing and mapped the data to the human genome. We restricted analysis to 94 libraries from 69 samples that had at least 0.06-fold average target coverage (average of 3.8-fold) and used majority rule to call an allele at each SNP covered at least once (Supplementary Data 1). After combining our data (Supplementary Information section 3) with 25 ancient samples from the literature — three Upper Paleolithic samples from Russia[1,6,7], seven people of European hunter-gatherer ancestry[2,4,5,8], and fifteen European farmers[2,3,4,8] — we had data from 94 ancient Europeans. Geographically, these came from Germany ($n = 41$), Spain ($n = 10$), Russia ($n = 14$), Sweden ($n = 12$), Hungary ($n = 15$), Italy ($n = 1$) and Luxembourg ($n = 1$) (Extended Data Table 2). Following the central European chronology, these included 19 hunter-gatherers (~43,000–2,600 BC), 28 Early Neolithic farmers (~6,000–4,000 BC), 11 Middle Neolithic farmers (~4,000–3,000 BC) including

Haak et al. (2015) – 1st page

We generated genome-wide data from 69 Europeans who lived between 8,000–3,000 years ago by enriching ancient DNA libraries for a target set of almost 400,000 polymorphisms. Enrichment of these positions decreases the sequencing required for genome-wide ancient DNA analysis by a median of around 250-fold, allowing us to study an order of magnitude more individuals than previous studies[1–8] and to obtain new insights about the past. We show that the populations of Western and Far Eastern Europe followed opposite trajectories between 8,000–5,000 years ago. At the beginning of the Neolithic period in Europe, ~8,000–7,000 years ago, closely related groups of early farmers appeared in Germany, Hungary and Spain, different from indigenous hunter-gatherers, whereas Russia was inhabited by a distinctive population of hunter-gatherers with high affinity to a ~24,000-year-old Siberian[6]. By ~6,000–5,000 years ago, farmers throughout much of Europe had more hunter-gatherer ancestry than their predecessors, but in Russia, the Yamnaya steppe herders of this time were descended not only from the preceding eastern European hunter-gatherers, but also from a population of Near Eastern ancestry. Western and Eastern Europe came into contact ~4,500 years ago, as the Late Neolithic Corded Ware people from Germany traced ~75% of their ancestry to the Yamnaya, documenting a massive migration into the heartland of Europe from its eastern periphery. This steppe ancestry persisted in all sampled central Europeans until at least ~3,000 years ago, and is ubiquitous in present-day Europeans. These results provide support for a steppe origin[9] of at least some of the Indo-European languages of Europe.

Genome-wide analysis of ancient DNA has emerged as a transformative technology for studying prehistory, providing information that is comparable in power to archaeology and linguistics. Realizing its promise, however, requires collecting genome-wide data from an adequate number of individuals to characterize population changes over time, which means not only sampling a succession of archaeological cultures[2], but also multiple individuals per culture. To make analysis of large numbers of ancient DNA samples practical, we used in-solution hybridization capture[10,11] to enrich next generation sequencing libraries for a target set of 394,577 single nucleotide polymorphisms (SNPs) ('390k capture'), 354,212 of which are autosomal SNPs that have also been genotyped using the Affymetrix Human Origins array in 2,345 humans from 203 populations[4,12]. This reduces the amount of sequencing required to obtain genome-wide data by a minimum of 45-fold and a median of 262-fold (Supplementary Data 1). This strategy allows us to report genomic scale data on more than twice the number of ancient Eurasians as has been presented in the entire preceding literature[1–8] (Extended Data Table 1).

We used this technology to study population transformations in Europe. We began by preparing 212 DNA libraries from 119 ancient samples in dedicated clean rooms, and testing these by light shotgun sequencing and mitochondrial genome capture (Supplementary Information section 1, Supplementary Data 1). We restricted the analysis to libraries with molecular signatures of authentic ancient DNA (elevated damage in the terminal nucleotide), negligible evidence of contamination based on mismatches to the mitochondrial consensus[13] and, where available, a mitochondrial DNA haplogroup that matched previous results using PCR[4,14,15] (Supplementary Information section 2). For 123 libraries prepared in the presence of uracil-DNA-glycosylase[16] to reduce errors due to ancient DNA damage[17], we performed 390k capture, carried out paired-end sequencing and mapped the data to the human genome. We restricted analysis to 94 libraries from 69 samples that had at least 0.06-fold average target coverage (average of 3.8-fold) and used majority rule to call an allele at each SNP covered at least once (Supplementary Data 1). After combining our data (Supplementary Information section 3) with 25 ancient samples from the literature — three Upper Paleolithic samples from Russia[1,6,7], seven people of European hunter-gatherer ancestry[2,4,5,8], and fifteen European farmers[2,3,4,8] — we had data from 94 ancient Europeans. Geographically, these came from Germany ($n = 41$), Spain ($n = 10$), Russia ($n = 14$), Sweden ($n = 12$), Hungary ($n = 15$), Italy ($n = 1$) and Luxembourg ($n = 1$) (Extended Data Table 2). Following the central European chronology, these included 19 hunter-gatherers (~43,000–2,600 BC), 28 Early Neolithic farmers (~6,000–4,000 BC), 11 Middle Neolithic farmers (~4,000–3,000 BC) including

Haak et al. (2015) Strategy

- Silica-based DNA extraction
- Double-stranded library preparation with truncated barcoded Illumina adapters
  - No DNA damage repair – for screening
    - Shallow shotgun sequencing
    - Mitochondrial DNA capture and sequencing
    - Above can be used to test authenticity – do we find characteristic damage of aDNA? Tests of contamination using mtDNA
  - Full USER-treatment and partial UDG repair
    - SNP Capture on 390k dataset

DNA Processing

We assigned read pairs to libraries by searching for matches to the expected index and barcode sequences (if present, as for the Adelaide and Boston libraries). We allowed no more than 1 mismatch per index or barcode, and zero mismatches if there was ambiguity in sequence assignment or if barcodes of 5 bp length were used (Adelaide libraries).

We used Seqprep (https://github.com/jstjohn/SeqPrep) to search for overlapping sequence between the forward and reverse read, and restricted to molecules where we could identify a minimum of 15 bp of overlap. We collapsed the two reads into a single sequence, using the consensus nucleotide if both reads agreed, and the read with higher base quality in the case of disagreement. For each merged nucleotide, we assigned the base quality to be the higher of the two reads. We further used Seqprep to search for the expected adaptor sequences at either ends of the merged sequence, and to produce a trimmed sequence for alignment.

Assign reads to libraries (using barcodes).

Determine sequence using forward and reverse reads with enough overlap. (Consider Base Quality)

DNA Processing

We assigned read pairs to libraries by searching for matches to the expected index and barcode sequences (if present, as for the Adelaide and Boston libraries). We allowed no more than 1 mismatch per index or barcode, and zero mismatches if there was ambiguity in sequence assignment or if barcodes of 5 bp length were used (Adelaide libraries).

We used Seqprep (https://github.com/jstjohn/SeqPrep) to search for overlapping sequence between the forward and reverse read, and restricted to molecules where we could identify a minimum of 15 bp of overlap. We collapsed the two reads into a single sequence, using the consensus nucleotide if both reads agreed, and the read with higher base quality in the case of disagreement. For each merged nucleotide, we assigned the base quality to be the higher of the two reads. We further used Seqprep to search for the expected adaptor sequences at either ends of the merged sequence, and to produce a trimmed sequence for alignment.

We mapped all sequences using BWA-0.6.1 (ref. 35). For mitochondrial analysis we mapped to the mitochondrial genome RSRS[36]. For whole-genome analysis we mapped to the human reference genome hg19. We restricted all analyses to sequences that had a mapping quality of MAPQ $\geq$ 37.

We sorted all mapped sequences by position, and used a custom script to search for mapped sequences that had the same orientation and start and stop positions. We stripped all but one of these sequences (keeping the best quality one) as duplicates.

Assign reads to libraries (using barcodes).

Determine sequence using forward and reverse reads with enough overlap. (Consider Base Quality)

Map sequence using bwa to reference genome. (Consider Mapping Quality)

Remove duplicates.

Preparing the dataset

390k capture, sequence analysis and quality control. For 390k analysis, we restricted to reads that not only mapped to the human reference genome hg19 but that also overlapped the 354,212 autosomal SNPs genotyped on the Human Origins array[4]. We trimmed the last two nucleotides from each sequence because we found that these are highly enriched in ancient DNA damage even for UDG-treated libraries. We further restricted analyses to sites with base quality ≥ 30.

We made no attempt to determine a diploid genotype at each SNP in each sample. Instead, we used a single allele—randomly drawn from the two alleles in the individual—to represent the individual at that site[20,39]. Specifically, we made an allele call at each target SNP using majority rule over all sequences overlapping the SNP. When each of the possible alleles was supported by an equal number of sequences, we picked an allele at random. We set the allele to 'no call' for SNPs at which there was no read coverage.

Two things to note:

1. Is our **sample high or low coverage**? Difficult to call heterozygotes without high coverage (best if 30x, but at least >10x)
2. Average coverage does not mean we have every targeted SNP – if no fragment with high enough quality covers a SNP, then cannot call allele.

'Pseudohaploid' genotyping or allele calls – each individual is treated as haploid – one allele is chosen at random at the selected biallelic sites.

Exploring the Haak et al. data

- EHG: Karelia, Samara
- SHG: Motala
- EN: LBK, LBKT, Starcevo, Spain_EN
- MN/CA: Yamnaya, Esperstedt, Baalberge, Spain_MN
- LN: Alberstedt, BenzigerodeHeimburg, Bell Beaker, Karsdorf, Corded Ware
- BA/IA: Halberstadt, Unetice

Activity

- Pick new groups from the Haak et al. data
- Use Supplementary Data 1 in (https://www.nature.com/articles/nature14317) to determine the average coverage for the 340k data per library, the fraction of libraries that were included for a set, mean contamination estimates, etc. If you have more time, look at other columns and see if you can understand what they mean and if there's other interesting facts to share.
- Should we add any of this information to our info file?