BIOINFORMATICS AND SYSTEMS MODELLING

BIOINF 3000 / BIOTECH 7005

WEEK 12

ANCIENT DNA PRACTICAL

Bastien Llamas
Tuesday 25 October 2022

Email: bastien.llamas@adelaide.edu.au
Phone: 08 8313 0262
Address: Darling Building, Level 2, Room 206

A bit of context

You are an ancient DNA researcher!

You work with archaeologists and another team of ancient DNA researchers on a project about early human migrations out of Africa. The project is now completed, and the two ancient DNA laboratories have decided to replicate some of their results in order to validate their respective work.

You received a number of anonymised tooth samples from the other ancient DNA laboratory, for which you do not know the geographic origin: it could come from Africa, Europe, Asia, or the Americas. You extracted DNA in your specialist ancient DNA facility and prepared DNA libraries for high throughput sequencing (Figure 1). You sent the DNA libraries for sequencing on an Illumina HiSeq 2500, and you requested 2 x 100 sequencing—i.e. the machine will sequence 100 nucleotides from the start of the DNA molecules, and 100 nucleotides from the end of the same molecules (Figure 1).

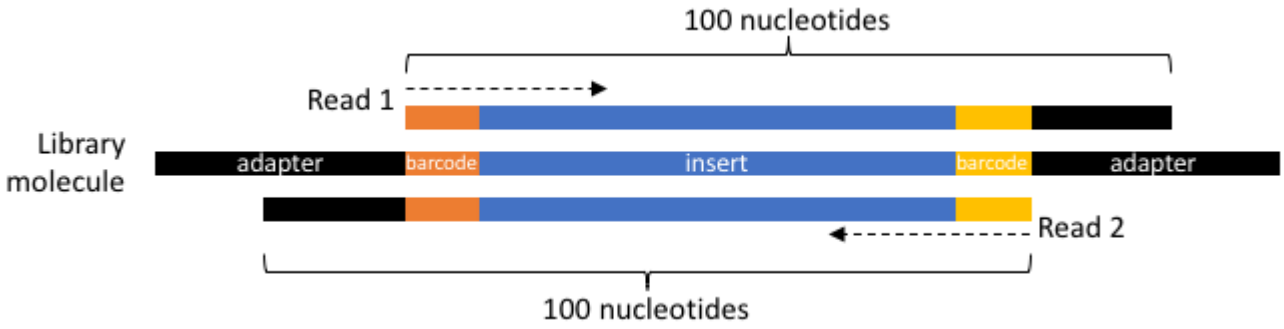

Figure 1: A sequencing library molecule is made of an insert (an ancient DNA fragment) flanked by sequencing adapters. In this particular experiment, each individual library is identified by a unique 5-mer barcode added between the adapters and the insert to allow for demultiplexing (i.e., to divide sequencing reads into separate files for each unique barcode combination). The library molecules are sequenced in two steps. First, sequencing starts from one end of the library molecules and goes for 100 nucleotides (Read 1). Second, sequencing starts from the other end and also goes for 100 nucleotides (Read 2). Because ancient DNA inserts are short, sequencing outputs (a.k.a "reads") are likely to contain the insert as well as some barcode and/or adapter sequences.

You just received the results from the sequencing service provider. It is now time to process the data and determine the geographic origin of the samples.

Below are some basic information and hands-on exercises to show you how to process high throughput sequencing data in ancient DNA research.

The raw sequencing data

The raw sequencing data are called "reads", and they are in a FASTQ format. More information about this particular data format can be found at https://en.wikipedia.org/wiki/FASTQ_format.

Briefly, the FASTQ format uses four lines per sequence:

- Line 1 always begins with `@`. It contains a sequence identifier generated by the sequencing machine and a description (optional).
- Line 2 is the raw sequence (or "read").
- Line 3 always begins with `+`. It usually does not contain any other information.
- Line 4 encodes the quality scores for the sequence in Line 2.

An example of your raw FASTQ data looks like:

```
@HWI-ST1359:56:C4EE8ACXX:6:1101:8215:1988 1:N:0:CCGGTAC
TAGCTAATTGAGATGGAAGAGCACACGTCTGAACTCCAGTCACCCGGTACATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAACAAAAACAACACAACAC
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJJJIJJJGHHIJJJIHHIHHJJJJJIJJJJGHFFFFECECECCCCDDD####################
@HWI-ST1359:56:C4EE8ACXX:6:1101:8082:1996 1:N:0:CCGGTAC
GTGGATCCTATCGGTTCTCGACTCGCTTCAGATCTACTTTGAATCTACTTTAGATCTATCGTAACGACTTAACTCGGAGATCGGAAGAGCACACGTCTGAA
+
CCCFFFFFHHHHHJJJJJJJJJJJJJIIIIJJJJJJJJJJIJJJJJJJJJJIIJJJJJJJJIIIHHHFFFFEEDDDDDDDCDDDBDDBDDDDDDDDDDA
@HWI-ST1359:56:C4EE8ACXX:6:1101:8352:1972 1:N:0:CCGGTAC
NTCTCCGATGCGTATCCAATGGGGAATCATCATACTACCACACGATGCACATATGAGATAGATGAGATCGGAAGCACACGTCTGAACTCCAGTCACCCGGT
+
#4=DFFFFHHHHFHJJJJJJJJJJJIIJJJJJJJJJJJJJIJJJJJJIJJIJJJJJJJJJHHFGEHEFEFDDDDACCABABDDDDDDDD@@ACCDDDDD0
@HWI-ST1359:56:C4EE8ACXX:6:1101:8415:1994 1:N:0:CCGGTAC
ATACGGTCATCGGGCGATCAGCTAGTCCTTTCTCGTTCGACTTTCGTACAGATGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCCGGTACATCTCG
+
CCCFFFDFHHHHHIIJJIJJJJIJJFHJIJJJJJGHJHIIJJJJJIHHIHHHGHFFFFEFDDDDD@DCDDDBBDDBCCDDDDDDDDCDCDB559BCBACDA
```

Quality control with FastQC

The first thing you do when you receive sequencing data is perform a quality control. The program **`FastQC`** (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is very easy to use and produces html reports that can be visualised in any internet browser.

🖳 Unzip the file `C4EE8ACXX-6-ACGV_GAII_INDEX10_CCGGTAC_L006_R1_001_fastqc.zip`, then open the file `fastqc_report.html` in your internet browser. Have a look at the different sections in the report and try and answer the below questions.

🖳 *Basic Statistics*
❓ How many reads have been generated? What is the sequence length?

🖳 *Per base sequence quality*
❓ Is the per base sequence quality good or bad?

🖳 *Per base sequence content*
❓ How do you explain the pattern for the first 5 bases?

🖳 *Sequence Duplication Levels*
❓ Why is there sequence replicates (36% duplicates, 18% triplicates, 9% quadruplicates, etc)?

🖳 *Overrepresented sequences* and *Kmer Content*
❓ What are the overrepresented sequences? Why do we observe so many Kmers? (Hint: ancient DNA molecules are short, but the sequencing effort is 2 x 100 nucleotides)

Demultiplexing

There was ~200 DNA libraries pooled together in that single sequencing effort. The use of unique combinations of 5-mer barcodes at the start and end of the ancient DNA molecules during the sequencing library preparation (Figure 1) allows to demultiplex the data. Several software have been developed to perform demultiplexing, to choose one is often more a question of personal preference than performance of the program.

After demultiplexing, you obtained ~200 individual files (one for each sample) that you can further process. You will focus on one particular sample dataset.

Adapter trimming and collapsing of reads

Ancient DNA molecules are typically less than 100 bp in length, meaning the 100-nucleotide-long reads may include some of the barcode and/or sequencing adapter sequences (Figure 1). It is therefore necessary to trim the barcode/adapter sequences from the data in order to only analyse the inserts.

A program has been developed for ancient DNA, where the barcode/adapter sequences are trimmed, and the overlapping reads are merged together (or collapsed). The program is **AdapterRemoval** (https://github.com/MikkelSchubert/adapterremoval).

Typically, more than 95% of the read pairs (Read 1 and Read 2) can be collapsed in ancient DNA datasets, and researchers will preferably use these collapsed reads for further processing.

The file `Unknown.fastq.gz` contains your sample sequencing data after demultiplexing, trimming, and collapsing. A sanity check with **FastQC** was performed to assess the quality of the data.

Unzip the files `Unknown_fastqc.zip`, then open the file `fastqc_report.html` in your internet browser. Have a look at the different sections in the report.

Do sequence duplication levels look good?

You are now going to process the data using the program **Geneious** (already installed on your computer).

Duplicate reads removal

In **Geneious**, create a new folder called “Unknown ancient human”. Import or drag the file `Unknown.fastq.gz`. In the menu, go to `Sequence`, then select `Remove Duplicate Reads`. Use the default parameters and click `OK`.

Compare the number of reads before and after duplicate removal.

The document `Unknown (without duplicates)` was exported as compressed fastq file and analysed with **FastQC**.

💻 Unzip the files `Unknown (without duplicates)_fastqc.zip`, then open the file `fastqc_report.html` in your internet browser.

❓ Have a look at the different sections in the report. Do sequence duplication levels look good?

Reads mapping

The reads are now ready for mapping. We will use the RSRS (Reconstructed Sapiens Reference Sequence) reference mitochondrial genome sequence as a template to align the reads.

💻 In **Geneious**, import or drag the file `RSRS.fasta`. Still in **Geneious**, select (using ctrl) the two documents `RSRS` and `Unknown (without duplicates)`.

💻 In the ribbon, go to `Align/Assemble` and select `Map to Reference`.

💻 In the `Sensitivity` drop-down menu, select `Medium Sensitivity / Fast` (Figure 2).

💻 Tick the box `Save consensus sequences` (Figure 2), click the `Options` button, and tick `Trim to reference sequence` (Figure 3).

💻 Click `OK`.

💻 Click `OK` again.

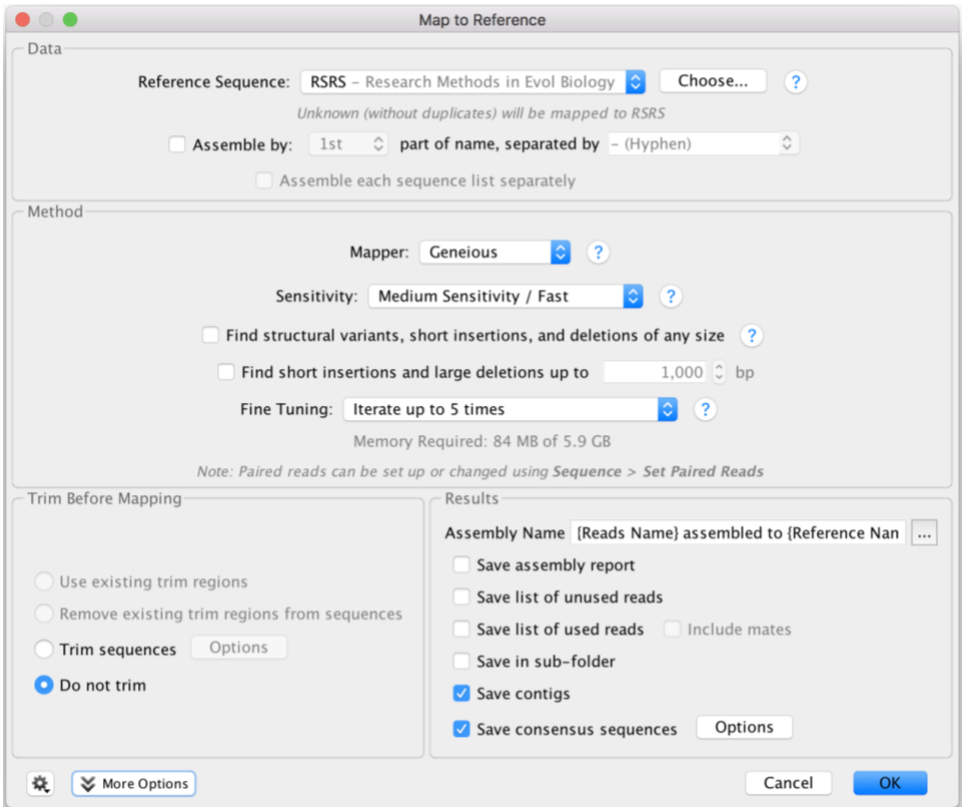

Figure 2: `Map to Reference` options.

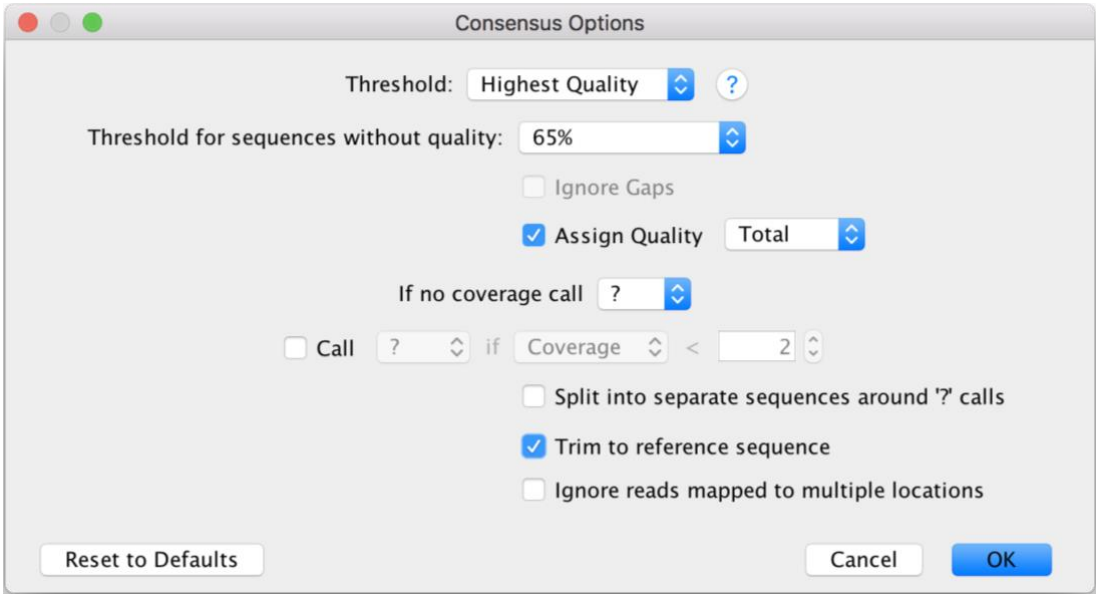

Figure 3: `Consensus` options.

💻 Explore the document `Unknown (without duplicates) assembled to RSRS`. Select the `%` tab in the sidebar and find the mean coverage.

❓ Is the coverage reasonably good?

💻 Select the home tab in the sidebar, click on the drop-down `Colors` menu, and select `MacClade`.

❓ What are the most frequent nucleotide substitutions? Do you notice any pattern?

One characteristic post-mortem damage that can be detected in ancient DNA is the deamination of C into U, which will be amplified and sequenced as T. Deamination of C occurs primarily at the extremities of ancient DNA molecules, and the resulting 5' C-to-T substitutions (and complementary 3' G-to-A substitutions) can be summarised using the program **`mapDamage`** (https://ginolhac.github.io/mapDamage/). Your data have the characteristic profile of 5' C-to-T and 3' G-to-A substitutions (Figure 4).

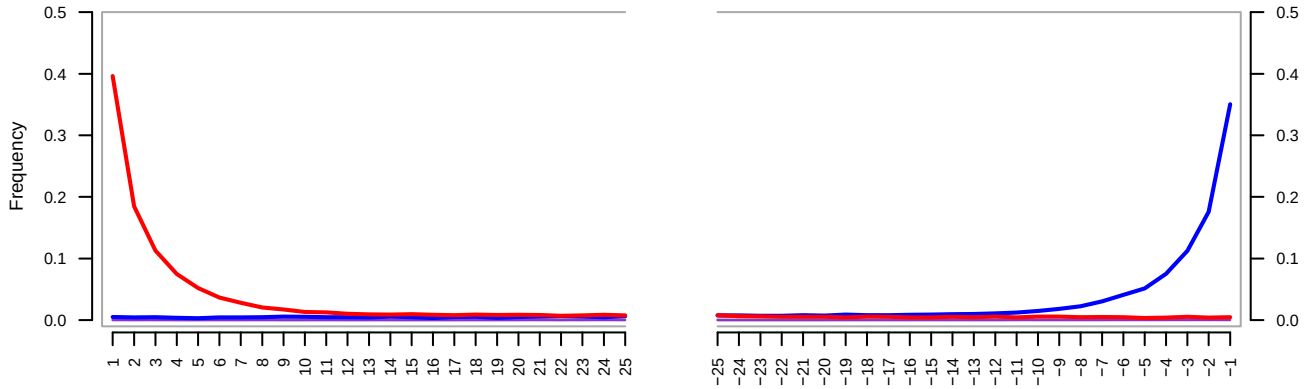

Figure 4: Characteristic accumulation of 5' C-to-T (red) and 3' G-to-A (blue) substitutions observed in ancient DNA datasets.

💻 Have a look at the consensus sequence at the top of the alignment.

❓ Does the ancient DNA damage have an impact on consensus nucleotide calls?

Alignment of the consensus sequence to human mitochondrial genomes

In **Geneious**, import or drag the file `human_mtDNA_alignment.phy`. Still in **Geneious**, select (using ctrl) the two documents `human_mtDNA_alignment` and `Unknown (without duplicates) assembled to RSRS 2`. In the ribbon, go to `Align/Assemble` and select `Multiple Align`. Use the default parameters and click `OK`.

Your unknown sequence is now aligned to some human mitochondrial genomes from Africa, Europe, Asia, and the Americas.

In **Geneious**, select the document `Nucleotide alignment`. In the ribbon, go to `Tree`.

In the `Outgroup` drop-down menu, select `Africa_L0a1a_EU092665` (Figure 5).

Tick the box `Resample tree`, and write `1,000` in the box `Number of Replicates` (Figure 5).

Click `OK`.

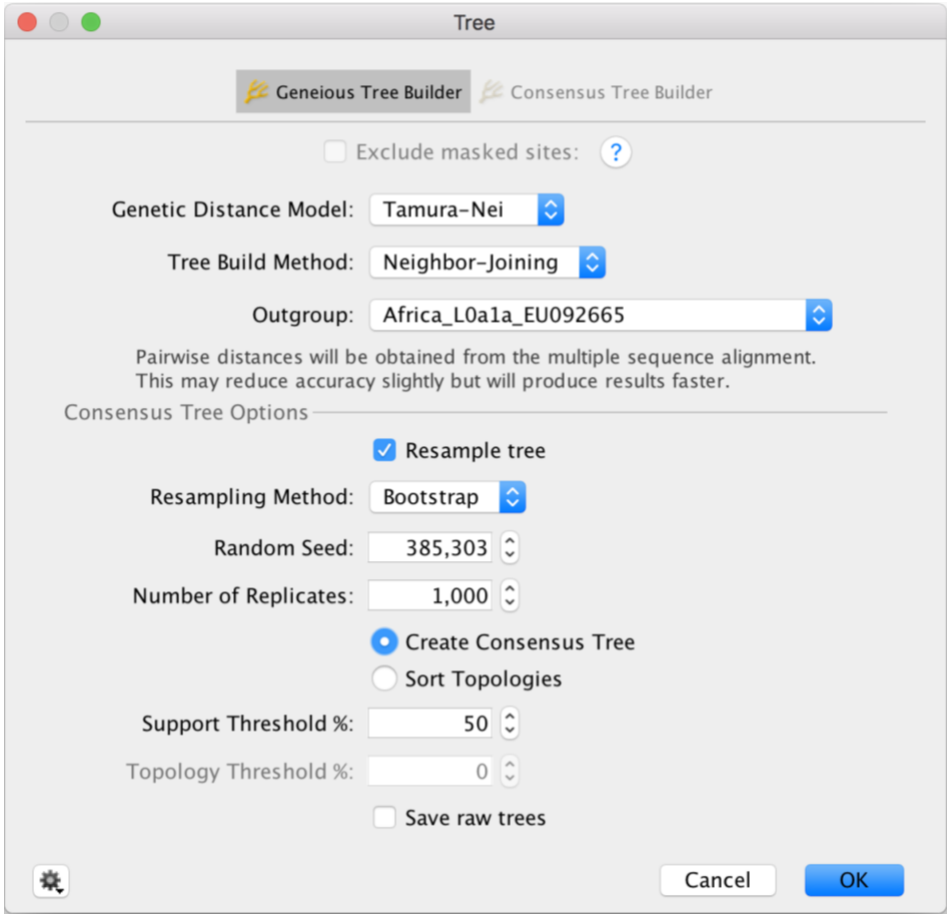

Figure 5: `Tree` options.

Select the output document `Nucleotide alignment consensus tree`.

What human mitochondrial genome sequences are most closely related to your unknown sample? What do you conclude about the geographic origin of your unknown sample?