Next-Generation Sequencing applied to ancient DNA

Aurélien Ginolhac & Mikkel Schubert
16th June 2014
What is ancient DNA?

DNA extracted from fossil remains of every kind

<table>
<thead>
<tr>
<th>Extinct species</th>
<th>Tooth</th>
<th>Coprolites</th>
<th>Bones</th>
<th>Hair</th>
<th>Museum samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denisova</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-clovis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neanderthal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mammoth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quagga</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Extant species, mainly human | | | |
|-------------------------------|-------------|
| Otzi                          | Aborigine   |
| Plague, Y. pestis             |             |
Why is ancient DNA interesting?

Provides access to extinct species

Not 2 species but sexual dimorphism, Bunce et al. 2003
Why is ancient DNA interesting?

Provide access to evolutionary points from the past

Rasmussen et al. 2011
Why use NGS?

800 mg, Neanderthal
379 bp mtDNA, HVR-I

38 mg, Altai Neanderthal
52X nuclear genome

Sanger

Illumina


Which protocol?

Stoneking and Krause, 2011
How ancient DNA looks like?

DNA depurination leads to fragmentation
How ancient DNA looks like?

Read length distribution

Middle Pleistocene horse from Thistle Creek
Orlando et al. 2013
How ancient DNA looks like?

Consequence of short read length

Read 1

ADAPTER INSERT ADAPTER

Read 2

Kmer Content

Position in read (bp)
Trimming adapter sequences and collapsing overlapping reads

PE, full overlap

Read 1

Read 2

PE, partial overlap

Read 1

Read 2

AdapterRemoval: easy cleaning of next-generation sequencing reads

Stinus Lindgreen
Trimming adapter sequences
and collapsing overlapping reads

before

after
Trimming adapter sequences

tricky when few bases of adapter are sequenced
How ancient DNA looks like?

Cytosine deamination
How ancient DNA looks like?

5’-ends C>T

G>A 3’-ends

Sequence analysis:
mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters
Håkon Jónsson1,2, Aurelien Girod1, Mikel Schubert1, Philip L. F. Johnson2 and Ludovic Orlando1
Typical steps in aDNA analyses

- Trimming of adapter sequences and low quality bases
- Mapping against one or more genomes
- Remove PCR duplicates, per library
- Merging, tagging of BAM files
- Coverage and depth summary statistics
- mapDamage model and rescaling of putative damage
- Genotyping of the BAMs
- Phylogenetic inference
- Metagenomic analyses
PALEOMIX, an automated pipeline from FASTQ to phylogeny

https://github.com/MikkelSchubert/paleomix

Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX

Mikkel Schubert¹, Luca Ermini¹, Clo Der Sarkissian¹, Håkon Jónsson¹, Aurélien Ginnolhac², Robert Schaefer², Michael D Martin¹, Ruth Fernández¹, Martin Kircher³, Molly McCue³, Eske Willerslev¹ & Ludovic Orlando¹

¹Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark. ²Divisional Information and Computational Biology Graduate Program, University of Minnesota, Rochester, Rochester, Minnesota, USA. ³Department of Genome Sciences, University of Washington, Seattle, Washington, USA. *College of Veterinary Medicine, University of Minnesota, St. Paul, Minnesota, USA. Correspondence should be addressed to L.O. (Lorlando@umn.edu).

Nature Protocols, May 2014
PALEOMIX

Metagenomic pipeline, not described today

BAM pipeline includes mapDamage

Phylogenetic pipeline

Centre for Geogenetics
PALEOMIX, usage

PALEOMIX - pipelines and tools for NGS data analyses.
Version: v1.0.1

Usage: paleomix <command> [options]

Commands:
  paleomix help                          -- Displays this message.

Pipelines:
  paleomix bam_pipeline                  -- Pipeline for trimming and mapping of NGS reads.
  paleomix phylo_pipeline                -- Pipeline for genotyping and phylogenetic inference from BAMs.

BAM/SAM tools:
  paleomix cleanup                      -- Reads SAM file from STDIN, and outputs sorted, tagged, and filtered BAM, for which NM and MD tags have been updated.
  paleomix coverage                     -- Calculate coverage across reference sequences or regions of interest.
  paleomix depths                       -- Calculate depth histograms across reference sequences or regions of interest.
  paleomix duphist                      -- Generates PCR duplicate histogram; for use with the 'Preseq' tool.
  paleomix rmdupCollapsed               -- Filters PCR duplicates for collapsed paired-end reads generated by the AdapterRemoval tool.

VCF/GTF/BED/Pileup tools:
  paleomix create_pileup                -- Creates tabix-indexed pileup for a (sparse) set of BED coordinates.
  paleomix gtf_to_bed                    -- Convert GTF file to BED files grouped by feature (coding, RNA, etc).
  paleomix sample_pileup                -- Randomly sample sites in a pileup to generate a FASTA sequence.
  paleomix vcf_filter                    -- Quality filters for VCF records, similar to 'vcftools.pl varFilter'.
  paleomix vcf_to_fasta                 -- Create most likely FASTA sequence from tabix-indexed VCF file.

Misc tools:
  paleomix cat                          -- Generalized cat command for gz, bz2 and uncompressed files.
PALEOMIX, scheduling

Config file with default values can be created
~/.pypeline/bam_pipeline.ini

Uses all cores by default; easy to set to use e.g. half of the cores on a per-server basis

Scheduling:
--bwtie2-max-threads=BWTIE2_MAX_THREADS
    Maximum number of threads to use per BWA instance [4]
--bwa-max-threads=BWA_MAX_THREADS
    Maximum number of threads to use per BWA instance [4]
--max-threads=MAX_THREADS
    Maximum number of threads to use in total [48]
--dry-run
    If passed, only a dry-run is performed, the dependency tree is printed, and no tasks are executed.
Hands On with PALEOMIX

Two exercises,

1) run an example on simulated reads

Makefile (YAML format)

Options:
- Platform: Illumina
- QualityOffset: 33
- SplitLanesByFilenames: yes
- CompressionFormat: b22
- Aligners:
  - Program: BWA
  - BWA:
    - MinQuality: 23
    - FilterUnmappedReads: yes
    - UseSeed: yes
  - Bowtie2:
    - MinQuality: 22
    - FilterUnmappedReads: yes
      --very-sensitive:
  - PCRDuplicates: filter
  - RescaleQualities: yes
  - mapDamage:
    - --downsample: 100000
- Features:
  - Realigned BAM # Generate indel-realigned BAM using the GATK Indel realigner
  - mapDamage # Generate mapDamage plot for each (unrealigned) library
  - Coverage # Generate coverage information for the raw BAM (w/ indel realignment)
  - Depths # Generate histogram of number of sites with a given read-depth
  - Summary # Generate target summary (uses statistics from raw BAM)
- Prefixes:
  - rCRS:
    - Path: 000_prefixes/rCRS.fasta
  - Label: "mitochondrial"
- ExampleProject:
  - Synthetic_Sample_1:
    - ACGATA:
      - Lane 1: 000_data/ACGATA_L1_R{Pair}_*.fastq.gz
    - GCTCTG:
      - Lane 1: 000_data/GCTCTG_L1_R1_*.fastq.gz

options

reference, FASTA

data, FASTQ
2) create a Makefile for analysing real horse data and ancient DNA
Hands On session!

Fetch the tutorial, open a terminal and type:

```
cp /home/local/27626/exercises/paleomix/PALEOMIX_handson.pdf .
```

The slides are available here:

```
/home/local/27626/exercises/paleomix/PALEOMIX_slides.pdf
```
Acknowledgments

Ludovic Orlando, Hákon Jónsson & the paleomix group
Danish High-Throughput Sequencing Centre

Stinus Lindgreen, Anders Krogh
Biocentre, København University

Bent Petersen, Josef Vogt, Thomas Sicheritz-Ponten
CBS, Technical University of Denmark

http://geogenetics.ku.dk/