

**DTU**



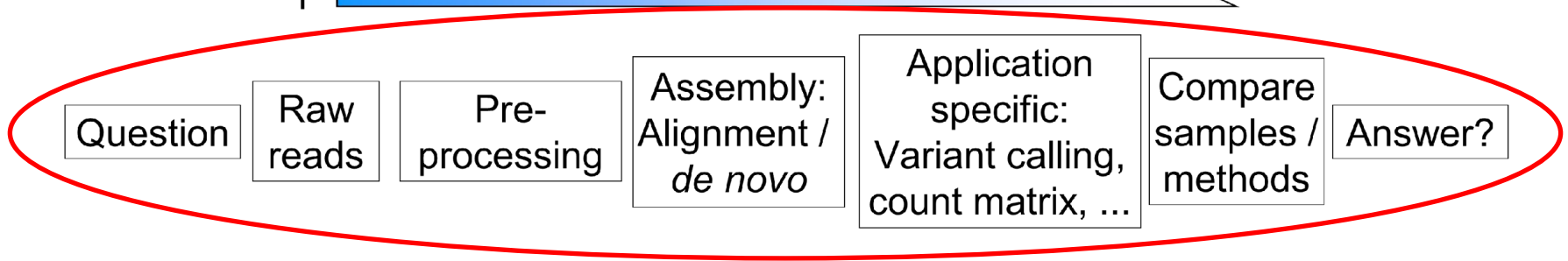
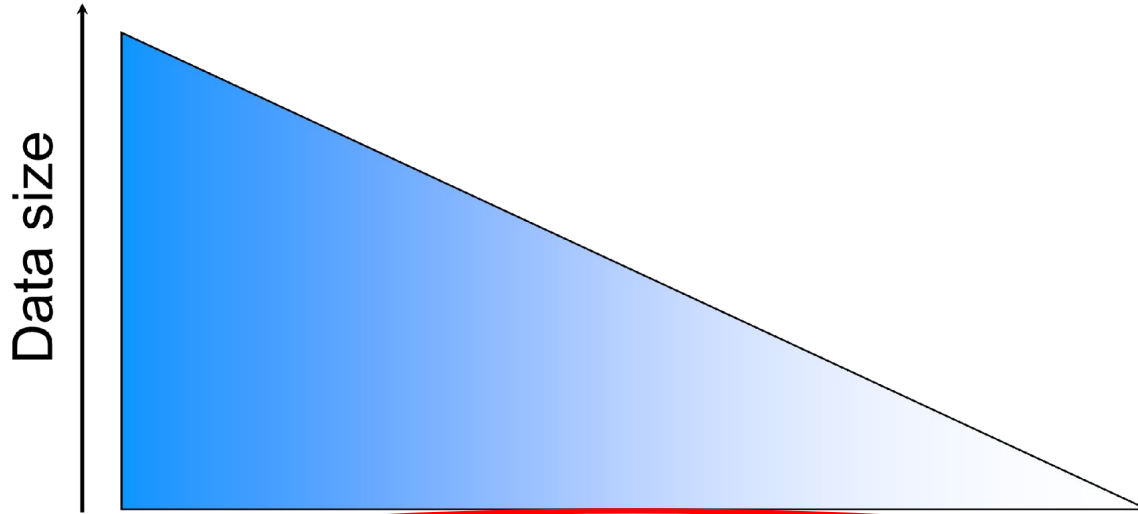


**DTU Health Technology  
Bioinformatics**

## *Projects*

*Gabriel Renaud  
Associate Professor  
Section of Bioinformatics  
Technical University of Denmark  
gabriel.reno@gmail.com*

# Generalized NGS analysis



Remember the slide from day 1? About the paragraph from a scientific paper?

# Why are we here?

FASTQ files were aligned to the hg19 reference genome using BWA-mem (v0.78). Duplicates were marked with Picard (v1.1.08), and base recalibration and realignment was performed using GATK (v2.8.1). Merged in silico bulk sequencing BAMs were generated by processing together all WGS FASTQ files from multiple regions to generate a single BAM file for each tumor. BAM files generated from individual tumor regions as well as in silico merged BAMs were processed for variant calling and filtering as described below. Substitution and indel calls were made using MuTect2 from GATK (v3.4.0). Structural variants (deletions, duplications, inversions and translocations) were called using delly<sup>29</sup> with a minimum of 4 discordant reads in the tumor required to call each SV (v0.7.1). Clonal and subclonal CNVs were called using Battenberg v3.2.2. All mutation calls (SNVs and SVs) were filtered as previously described using in house pipelines<sup>30</sup>. To reiterate, we required a minimum depth of 10X in the tumor and normal with 0 reads supporting the variant in the matched normal.

“In theory, there is no difference between theory and practice  
but in practice there is a huge difference between theory and practice.”

-Yogi Berra?  
-Richard Feynman?  
Someone else?

# Problems with NGS...

1. Reanalyzing previous data is part of the job
2. Impossible/difficult to reproduce results
3. “in house scripts” do not get me started...
4. Getting the data in the first place
5. Getting the metadata to match the data

# Learning objectives

1. Are you able to:
  - a. work in group and delegate tasks?
  - b. set realistic objectives?
  - c. use the command line?
  - d. understand the strength and weakness of each tool?
  - e. explain key steps in a critical manner?



# Projects

- Try to analyze an empirical dataset and present results on poster
  - Either replicate some results or ask your own question
- Aim for at least 1 figure, 1 table or 2 figures **of results**
- 5-6 pr. group
- You can find a dataset on SRA/ENA
- Try to find raw data, untrimmed
  - If not, please contact us

# Projects

- You can use your own data if everyone in the group agrees ***and*** it can be presented on a poster
- Subset! Do not analyze very large datasets (time, resources)
- Subset! Do not replicate every figure/table!

# Pitfalls

- Beware of esoteric tools or overuse of “in-house” scripts or programs
- No, you do not need to use exactly the same tools they used
- Some data (esp. human patient data) is not available due to privacy issues
- You can download subsets using SRA

# Pitfalls

- Limit the number of samples
- You will **not** get penalized if you gave a *bona fide* attempt at replicating results and cannot
- You can ask your own questions!

# Group formation

- Try to create groups with multiple competences
- Chose a group based on eg. field of interest
- Do not bite off more than you can chew:
  - Downloading the data, preprocessing, aligning will take several days

# Group formation

- Everyone is expected to put in work
- If there are freeloaders, let me know



# Posters

- Each group will create a 1 page poster (not a report)
- You have to send it to:
  - gabriel.reno@gmail.com

Before Thursday 18th at noon (12PM)

It is 15% of the grade.



# Posters

- I cannot read long paragraphs, please use very  
concise prose
- Bullet points are great!
- **Important text in bold**

# Posters

Little text, bullet points are amazing!



# Grouping & Guidance

- Fill in group information in Google doc
- 3 min chat tomorrow at 13
  - What do you plan to do?
  - How much data?
- Project assistance: every day
  - Teachers+TA
  - Write on Discord

# Grouping & Guidance

- Data goes here:

```
/home/projects/22126_NGS/projects/groupX
```

- Put the parsing scripts+plots there.
- Do not fabricate results.

# Pupil servers

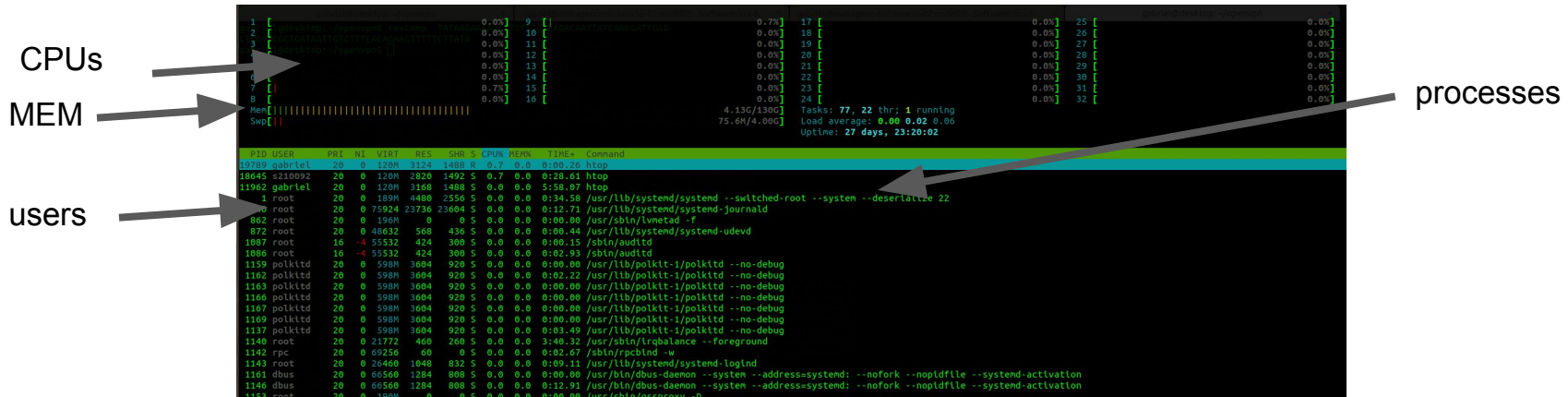
name	CPUs	RAM
pupil1	40	252G
pupil2	24	110G
pupil3	24	94G
pupil4	16	63G
pupil5	16	63G

# Be nice

- Run larger programs on the servers using nice eg.

```
nice -n 19 blastall -i alldataainthegalaxy -db  
everythingeversequenced
```

- If the server is too full come talk to me :-)
- You can now all nodes ! But check available CPU/RAM. How to check?



# Thou shall keep your files zipped

- Zip your vcf, text whatever files
  - there are tools to work with zipped files (zcat, zgrep, zless)
- Use BAM/CRAM **never sam**
- Beware, what is wrong with this?:

```
bwa mem reference.fasta input.fastq.gz > output.bam
```

2023:



What is wrong with this command line:

```
bwa mem reference.fasta input.fastq.gz > output.bam
```



It is difficult to determine what is wrong with this command line without more information. However, a few things that could be causing issues include:

- The reference.fasta and input.fastq.gz files might not be in the same directory as the command, or might not exist at all.
- The output file, output.bam, might not be writable (e.g. if the user does not have permission to write to the directory where the output file is located).
- There might be an issue with the specific version of bwa being used.
- You may need to provide more options for the 'bwa mem' command or it would cause an error.

It's recommended to check the existence of the input files, if you have the permission to write at the directory that you are trying to write the output, and consult the bwa manual to be sure that you have provided the necessary options.



2024:

### ChatGPT

The command line you provided for running BWA (Burrows-Wheeler Aligner) contains a small error related to the output format. Here's the command you provided:

```
shell Copy code  
bwa mem reference.fasta input.fastq.gz > output.bam
```

The issue with this command is that it directs the output of `bwa mem` to a file named `output.bam`, but the output is in SAM format, not BAM format. The BWA `mem` command generates output in SAM (Sequence Alignment/Map) format, which is a text format. The `.bam` file extension is typically used for binary alignment/map format, which is a binary version of SAM.

To correct this, you have two options:

1. **Output in SAM Format:** If you want the output in SAM format, you should change the output file's extension to `.sam` to reflect the correct format:

```
shell Copy code  
bwa mem reference.fasta input.fastq.gz > output.sam
```

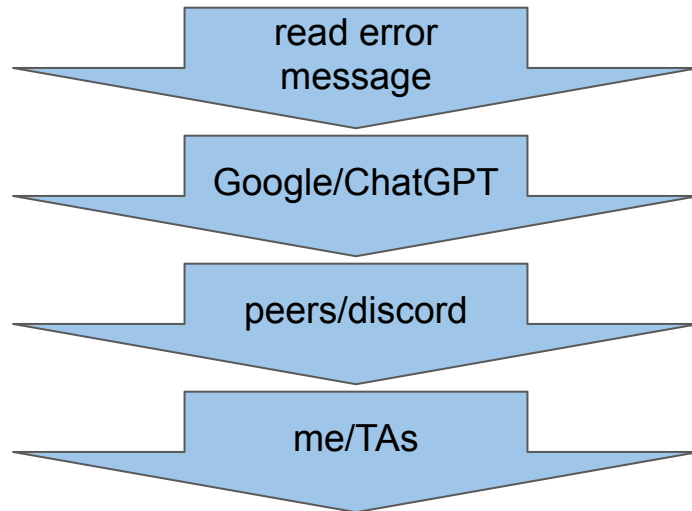
2. **Convert to BAM Format:** If you need the output in BAM format, you should pipe the SAM output of BWA into a tool like `samtools` to convert it to BAM format. Here's how you can do it:

```
shell Copy code  
bwa mem reference.fasta input.fastq.gz | samtools view -bS - > output.
```

In this command, `samtools view -bS -`  adds the SAM format from standard input

# How to get help?

- You want software installed, ask me or TAs
- Help with NGS software:



# The final exam

- Friday, Jan 19th at 9am
- I should be there
- Multiple choice
- 85% of the grade

# The final exam

- Understand the material
- Focused on the core lectures
- 3-4 general questions on the specialized lectures

# Parting words

- Do not trust your data, use:
  - quality controls
  - visualizations
- No one size fits all solution for everything
  - How to genotype, population geneticists vs medical field
- Every tool shown in this class may/will be outdated in 5 years
  - Sorry for no textbook but it would be outdated soon
  - Read recent papers, reviews, benchmarking studies
  - bioRxiv is great but not peer-reviewed

# Parting words

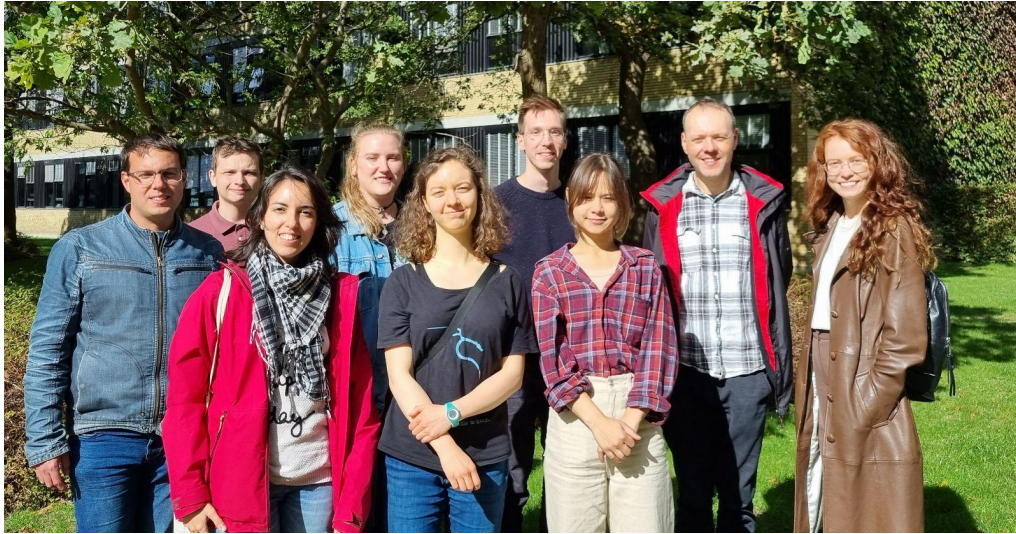
- Question existing methods, pipelines, be wary of:
  - “This is how we do things around here”
  - “This is the standard pipeline for this kind of data”
- Understand how tools work, test
- Do not trust your code, test

# Parting words

- Know the history of the seq. tech you are using as you may work with older data
- Do your literature search, use existing tools when possible
  - Google Scholar
  - Twitter (yes unfortunately)
- Talk to lab techs, molecular biologists, try to learn as much as possible about the biology and biotech behind the data being generated.

# Special projects/Master's projects

- Like NGS? Genotyping? Population genetics? Ancient DNA? pangenome graphs? sedimentary DNA?



- me
- 2 Postdocs
- 3 PhDs
- 3 Master's
- 1 undergrad

The Modern and Ancient Genomes Group



# Articles in 2023

Methods in Ecology and Evolution



APPLICATION | Open Access |

## euka: Robust tetrapodic and arthropodic taxa detection from modern and ancient environmental DNA using pangenomic reference graphs

Nicola Alexandra Vogel , Joshua Daniel Rubin, Mikkel Swartz, Juliette Vlieghe, Peter Wad Sackett, Anders Gorm Pedersen, Mikkel Winther Pedersen, Gabriel Renaud

First published: 23 September 2023 | <https://doi.org/10.1111/2041-210X.14214>

Handling Editor: Aline Lee

SECTIONS

### Abstract

1. Ancient environmental DNA (aeDNA) is a crucial tool for environmental reconstruction. However, the construction of aeDNA involves the inherited challenges of ancient DNA, such as the low abundance of eDNA samples, such as taxonomic identification of identified taxonomic groups. Current methods for aeDNA involve performing mapping followed by taxonomic identification, which do abundance estimation. The former leaves about 50% of the reads unassigned, and the latter is not designed for large

ORIGINAL RESEARCH article

Front. Bioinform., 07 December 2023

Sec. Genomic Analysis

Volume 3 - 2023 | <https://doi.org/10.3389/fbinf.2023.1260486>

## Benchmarking software tools for trimming adapters and merging next-generation sequencing data for ancient DNA

Annette Lien<sup>1</sup> Leonardo Pestana Legori<sup>2</sup> Louis Kraft<sup>1</sup> Peter Wad Sackett<sup>1</sup> Gabriel Renaud<sup>1\*</sup>

<sup>1</sup> Department of Health Technology, Section for Bioinformatics, Technical University of Denmark, Kongens Lyngby, Denmark

<sup>2</sup> University of Debrecen, Debrecen, Hungary

Ancient DNA is highly degraded, resulting in very short sequences. Reads generated with modern high-throughput sequencing machines are generally longer than ancient DNA molecules, therefore the reads often contain some portion of the sequencing adaptors. It is crucial to remove those adaptors, as they can interfere with downstream analysis. Furthermore, overlapping portions when DNA has been read forward and backward (paired-end) can be merged to correct sequencing errors and improve read quality. Several merging tools have been developed, however, no one has attempted to benchmark them. This study compares the effect of different merging tools on downstream analyses. Through the analysis of ancient DNA samples, the results show that different tools were analyzed in their ability to reconstruct ancient DNA. The results show that different merging tools exhibit notable differences in their

## HaploCart: Human mtDNA haplogroup classification using a pangenomic reference graph

Joshua Daniel Rubin , Nicola Alexandra Vogel, Shyam Gopalakrishnan, Peter Wad Sackett, Gabriel Renaud

Version 2 Published: June 7, 2023 • <https://doi.org/10.1371/journal.pcbi.1011148>

Article	Authors	Metrics	Comments	Media Coverage
---------	---------	---------	----------	----------------

### Abstract

Author summary

1 Introduction

2 Results

3 Discussion

4 Methods

Supporting information

Acknowledgments

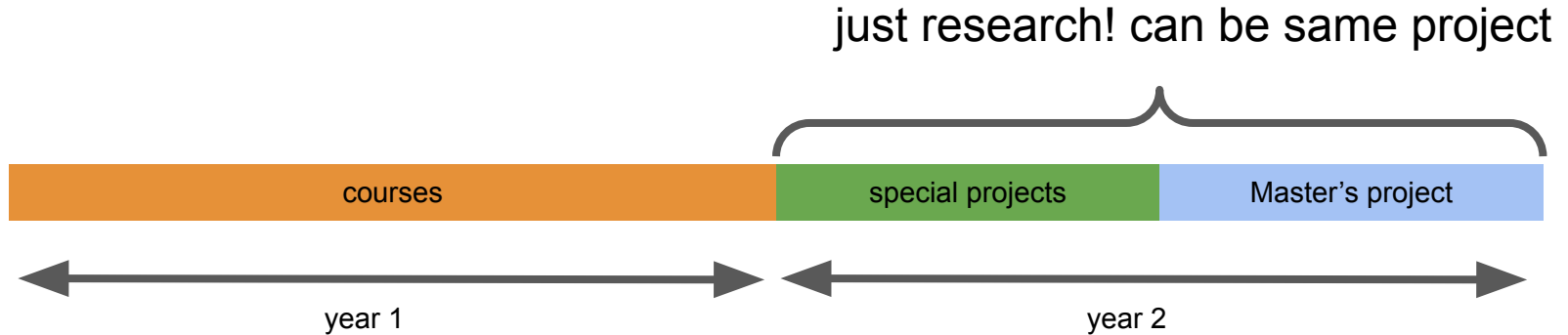
References

### Abstract

Current mitochondrial DNA (mtDNA) haplogroup classification tools map reads to a single reference genome and perform inference based on the detected mutations to this reference. This approach biases haplogroup assignments towards the reference and prohibits accurate calculations of the uncertainty in assignment. We present HaploCart, a probabilistic mtDNA haplogroup classifier which uses a pangenomic reference graph framework together with principles of Bayesian inference. We demonstrate that our approach significantly outperforms available tools by being more robust to lower coverage or incomplete consensus sequences and producing phylogenetically-aware confidence scores that are unbiased towards any haplogroup. HaploCart is available both as a command-line tool and through a user-friendly web interface. The C++ program accepts as input consensus FASTA, FASTQ, or GAM files,

# Special projects/Master's projects

- Like research? Thinking about going to PhD? want to write a paper?



Thanks!