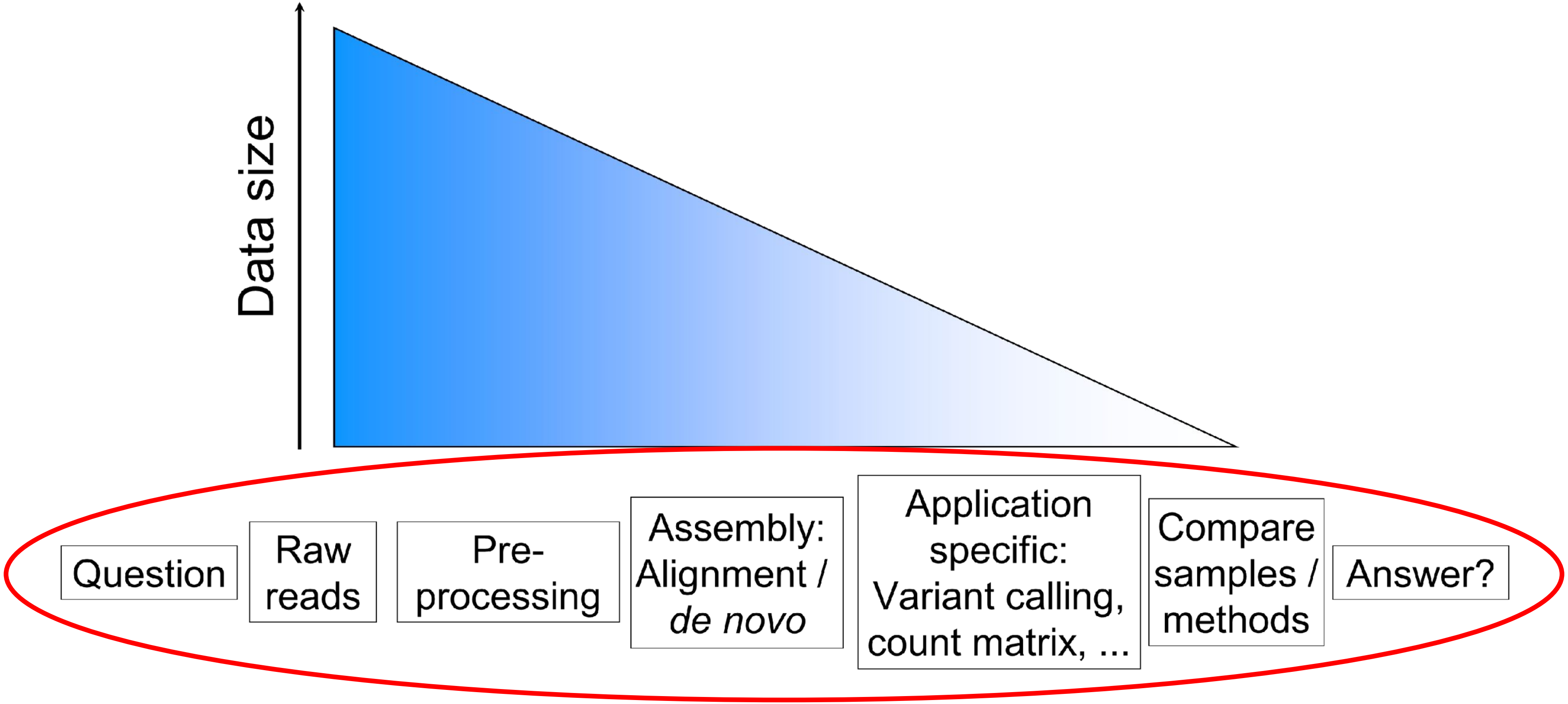
DTU

DTU Health Technology
Bioinformatics

Projects

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

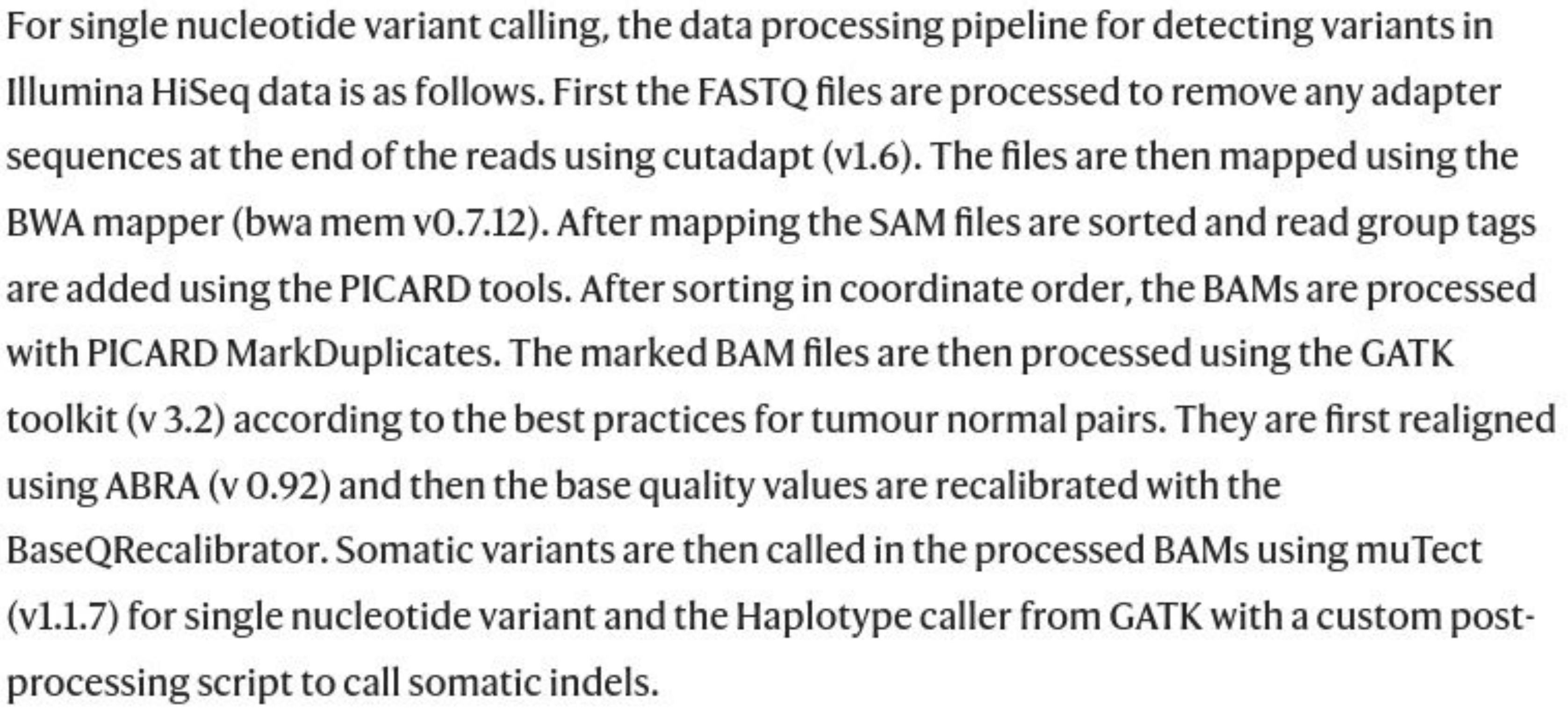
Generalized NGS analysis



- Question
- Raw reads
- Pre-processing
- Assembly: Alignment / *de novo*
- Application specific: Variant calling, count matrix, ...
- Compare samples / methods
- Answer?

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

Why are we here?



For single nucleotide variant calling, the data processing pipeline for detecting variants in Illumina HiSeq data is as follows. First the FASTQ files are processed to remove any adapter sequences at the end of the reads using cutadapt (v1.6). The files are then mapped using the BWA mapper (bwa mem v0.7.12). After mapping the SAM files are sorted and read group tags are added using the PICARD tools. After sorting in coordinate order, the BAMs are processed with PICARD MarkDuplicates. The marked BAM files are then processed using the GATK toolkit (v 3.2) according to the best practices for tumour normal pairs. They are first realigned using ABRA (v 0.92) and then the base quality values are recalibrated with the BaseQRecalibrator. Somatic variants are then called in the processed BAMs using muTect (v1.1.7) for single nucleotide variant and the Haplotype caller from GATK with a custom post-processing script to call somatic indels.

“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 Feyman?
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
- No single-cell data

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

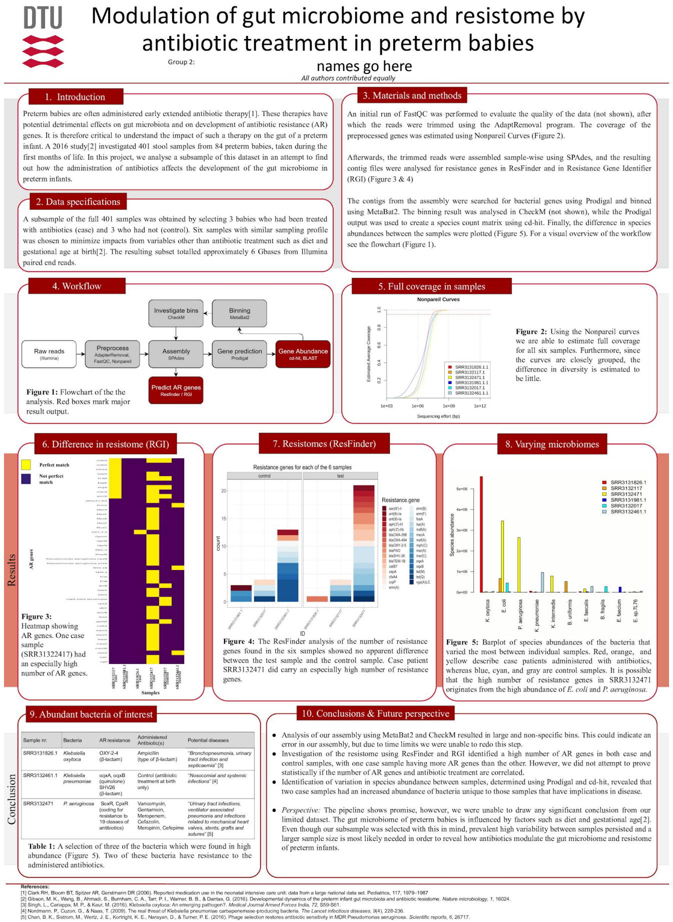
- Try to create groups with multiple competences
- Chose a group based on eg. field of interest

- 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

Previous projects



Modulation of gut microbiome and resistome by antibiotic treatment in preterm babies

Group 2:

names go here

All authors contributed equally

1. Introduction

Preterm babies are often administered early extended antibiotic therapy[1]. These therapies have potential detrimental effects on gut microbiota and on development of antibiotic resistance (AR) genes. It is therefore critical to understand the impact of such a therapy on the gut of a preterm infant. A 2016 study[2] investigated 401 stool samples from 84 preterm babies, taken during the first months of life. In this project, we analyse a subsample of this dataset in an attempt to find out how the administration of antibiotics affects the development of the gut microbiome in preterm infants.

2. Data specifications

A subsample of the full 401 samples was obtained by selecting 3 babies who had been treated with antibiotics (case) and 3 who had not (control). Six samples with similar sampling profile was chosen to minimize impacts from variables other than antibiotic treatment such as diet and gestational age at birth[2]. The resulting subset totalled approximately 6 Gbases from Illumina paired end reads.

3. Materials and methods

An initial run of FastQC was performed to evaluate the quality of the data (not shown), after which the reads were trimmed using the AdaptRemoval program. The coverage of the preprocessed genes was estimated using Nonpareil Curves (Figure 2).

Afterwards, the trimmed reads were assembled sample-wise using SPAdes, and the resulting contig files were analysed for resistance genes in ResFinder and in Resistance Gene Identifier (RGI) (Figure 3 & 4)

The contigs from the assembly were searched for bacterial genes using Prodigal and binned using MetaBat2. The binning result was analysed in CheckM (not shown), while the Prodigal output was used to create a species count matrix using cd-hit. Finally, the difference in species abundances between the samples were plotted (Figure 5). For a visual overview of the workflow see the flowchart (Figure 1).

4. Workflow

Figure 1: Flowchart of the the analysis. Red boxes mark major result output.

5. Full coverage in samples

Figure 2: Using the Nonpareil curves we are able to estimate full coverage for all six samples. Furthermore, since the curves are closely grouped, the difference in diversity is estimated to be little.

Results

6. Difference in resistome (RGI)

Figure 3: Heatmap showing AR genes. One case sample (SRR31322417) had an especially high number of AR genes.

7. Resistomes (ResFinder)

Figure 4: The ResFinder analysis of the number of resistance genes found in the six samples showed no apparent difference between the test sample and the control sample. Case patient SRR3132471 did carry an especially high number of resistance genes.

8. Varying microbiomes

Figure 5: Barplot of species abundances of the bacteria that varied the most between individual samples. Red, orange, and yellow describe case patients administered with antibiotics, whereas blue, cyan, and gray are control samples. It is possible that the high number of resistance genes in SRR3132471 originates from the high abundance of *E. coli* and *P. aeruginosa.*

Conclusion

9. Abundant bacteria of interest

Sample nr.	Bacteria	AR resistance	Administered Antibiotic(s)	Potential diseases
SRR3131826.1	*Klebsiella oxytoca*	OXY-2-4 (β-lactam)	Ampicillin (type of β-lactam)	*"Bronchopneumonia, urinary tract infection and septicaemia"* [3]
SRR3132461.1	*Klebsiella pneumoniae*	oqxA, oqxB (quinolone) SHV26 (β-lactam)	Control (antibiotic treatment at birth only)	*"Nosocomial and systemic infections"* [4]
SRR3132471	*P. aeruginosa*	SoxR, CpxR (coding for resistance to 19 classes of antibiotics)	Vancomycin, Gentamicin, Meropenem, Cefazolin, Meropinin, Cefepime	*"Urinary tract infections, ventilator associated pneumonia and infections related to mechanical heart valves, stents, grafts and sutures"* [5]

Table 1: A selection of three of the bacteria which were found in high abundance (Figure 5). Two of these bacteria have resistance to the administered antibiotics.

10. Conclusions & Future perspective

- Analysis of our assembly using MetaBat2 and CheckM resulted in large and non-specific bins. This could indicate an error in our assembly, but due to time limits we were unable to redo this step.
- Investigation of the resistome using ResFinder and RGI identified a high number of AR genes in both case and control samples, with one case sample having more AR genes than the other. However, we did not attempt to prove statistically if the number of AR genes and antibiotic treatment are correlated.
- Identification of variation in species abundance between samples, determined using Prodigal and cd-hit, revealed that two case samples had an increased abundance of bacteria unique to those samples that have implications in disease.

- *Perspective:* The pipeline shows promise, however, we were unable to draw any significant conclusion from our limited dataset. The gut microbiome of preterm babies is influenced by factors such as diet and gestational age[2]. Even though our subsample was selected with this in mind, prevalent high variability between samples persisted and a larger sample size is most likely needed in order to reveal how antibiotics modulate the gut microbiome and resistome of preterm infants.

References:
[1] Clark RH, Bloom BT, Spitzer AR, Gerstmann DR (2006). Reported medication use in the neonatal intensive care unit: data from a large national data set. Pediatrics, 117, 1979–1987
[2] Gibson, M. K., Wang, B., Ahmadi, S., Burnham, C. A., Tarr, P. I., Warner, B. B., & Dantas, G. (2016). Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. *Nature microbiology*, *1*, 16024.
[3] Singh, L., Cariappa, M. P., & Kaur, M. (2016). Klebsiella oxytoca: An emerging pathogen?. *Medical Journal Armed Forces India*, *72*, S59-S61.
[4] Nordmann, P., Cuzon, G., & Naas, T. (2009). The real threat of Klebsiella pneumoniae carbapenemase-producing bacteria. *The Lancet infectious diseases*, *9*(4), 228-236.
[5] Chan, B. K., Sistrom, M., Wertz, J. E., Kortright, K. E., Narayan, D., & Turner, P. E. (2016). Phage selection restores antibiotic sensitivity in MDR Pseudomonas aeruginosa. *Scientific reports*, *6*, 26717.

Posters

- Each group will create a 1 page poster (not a report)

- You have to send it to:
    - gabriel.reno@gmail.com

Before Thursday 23rd at noon (12PM)

It is 25% 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

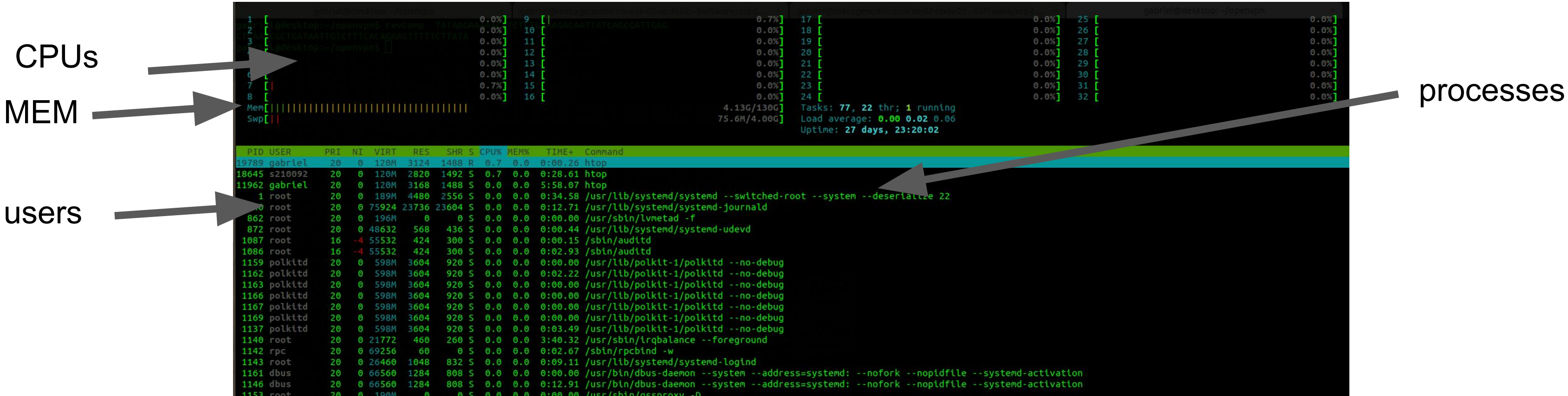
Be nice

- Run larger programs on the servers using nice eg.

```
nice -n 19 blastall -i alldatainthegalaxy -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?



```
  1  [                          0.0%]   9  [|                         0.7%]   17 [                          0.0%]   25 [                          0.0%]
  2  [                          0.0%]   10 [                          0.0%]   18 [                          0.0%]   26 [                          0.0%]
  3  [                          0.0%]   11 [                          0.0%]   19 [                          0.0%]   27 [                          0.0%]
                                0.0%]   12 [                          0.0%]   20 [                          0.0%]   28 [                          0.0%]
                                0.0%]   13 [                          0.0%]   21 [                          0.0%]   29 [                          0.0%]
  6  [                          0.0%]   14 [                          0.0%]   22 [                          0.0%]   30 [                          0.0%]
  7  [|                         0.7%]   15 [                          0.0%]   23 [                          0.0%]   31 [                          0.0%]
  8  [                          0.0%]   16 [                          0.0%]   24 [                          0.0%]   32 [                          0.0%]
  Mem[|||||||||||||||||||||||||||||||||||                  4.13G/130G]   Tasks: 77, 22 thr; 1 running
  Swp[||                                                   75.6M/4.00G]   Load average: 0.00 0.02 0.06
                                                                          Uptime: 27 days, 23:20:02

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
19789 gabriel    20   0  120M  3124  1488 R  0.7  0.0  0:00.26 htop
18645 s210092    20   0  120M  2820  1492 S  0.7  0.0  0:28.61 htop
11962 gabriel    20   0  120M  3168  1488 S  0.0  0.0  5:58.07 htop
    1 root       20   0  189M  4480  2556 S  0.0  0.0  0:34.58 /usr/lib/systemd/systemd --switched-root --system --deserialize 22
    0 root       20   0 75924 23736 23604 S  0.0  0.0  0:12.71 /usr/lib/systemd/systemd-journald
  862 root       20   0  196M     0     0 S  0.0  0.0  0:00.00 /usr/sbin/lvmetad -f
  872 root       20   0 48632   568   436 S  0.0  0.0  0:00.44 /usr/lib/systemd/systemd-udevd
 1087 root       16  -4 55532   424   300 S  0.0  0.0  0:00.15 /sbin/auditd
 1086 root       16  -4 55532   424   300 S  0.0  0.0  0:02.93 /sbin/auditd
 1159 polkitd    20   0  598M  3604   920 S  0.0  0.0  0:00.00 /usr/lib/polkit-1/polkitd --no-debug
 1162 polkitd    20   0  598M  3604   920 S  0.0  0.0  0:02.22 /usr/lib/polkit-1/polkitd --no-debug
 1163 polkitd    20   0  598M  3604   920 S  0.0  0.0  0:00.00 /usr/lib/polkit-1/polkitd --no-debug
 1166 polkitd    20   0  598M  3604   920 S  0.0  0.0  0:00.00 /usr/lib/polkit-1/polkitd --no-debug
 1167 polkitd    20   0  598M  3604   920 S  0.0  0.0  0:00.00 /usr/lib/polkit-1/polkitd --no-debug
 1169 polkitd    20   0  598M  3604   920 S  0.0  0.0  0:00.00 /usr/lib/polkit-1/polkitd --no-debug
 1137 polkitd    20   0  598M  3604   920 S  0.0  0.0  0:03.49 /usr/lib/polkit-1/polkitd --no-debug
 1140 root       20   0 21772   460   260 S  0.0  0.0  3:40.32 /usr/sbin/irqbalance --foreground
 1142 rpc        20   0 69256    60     0 S  0.0  0.0  0:02.67 /sbin/rpcbind -w
 1143 root       20   0 26460  1048   832 S  0.0  0.0  0:09.11 /usr/lib/systemd/systemd-logind
 1161 dbus       20   0 66560  1284   808 S  0.0  0.0  0:00.00 /usr/bin/dbus-daemon --system --address=systemd: --nofork --nopidfile --systemd-activation
 1146 dbus       20   0 66560  1284   808 S  0.0  0.0  0:12.91 /usr/bin/dbus-daemon --system --address=systemd: --nofork --nopidfile --systemd-activation
```

Thou shall keep your files zipped

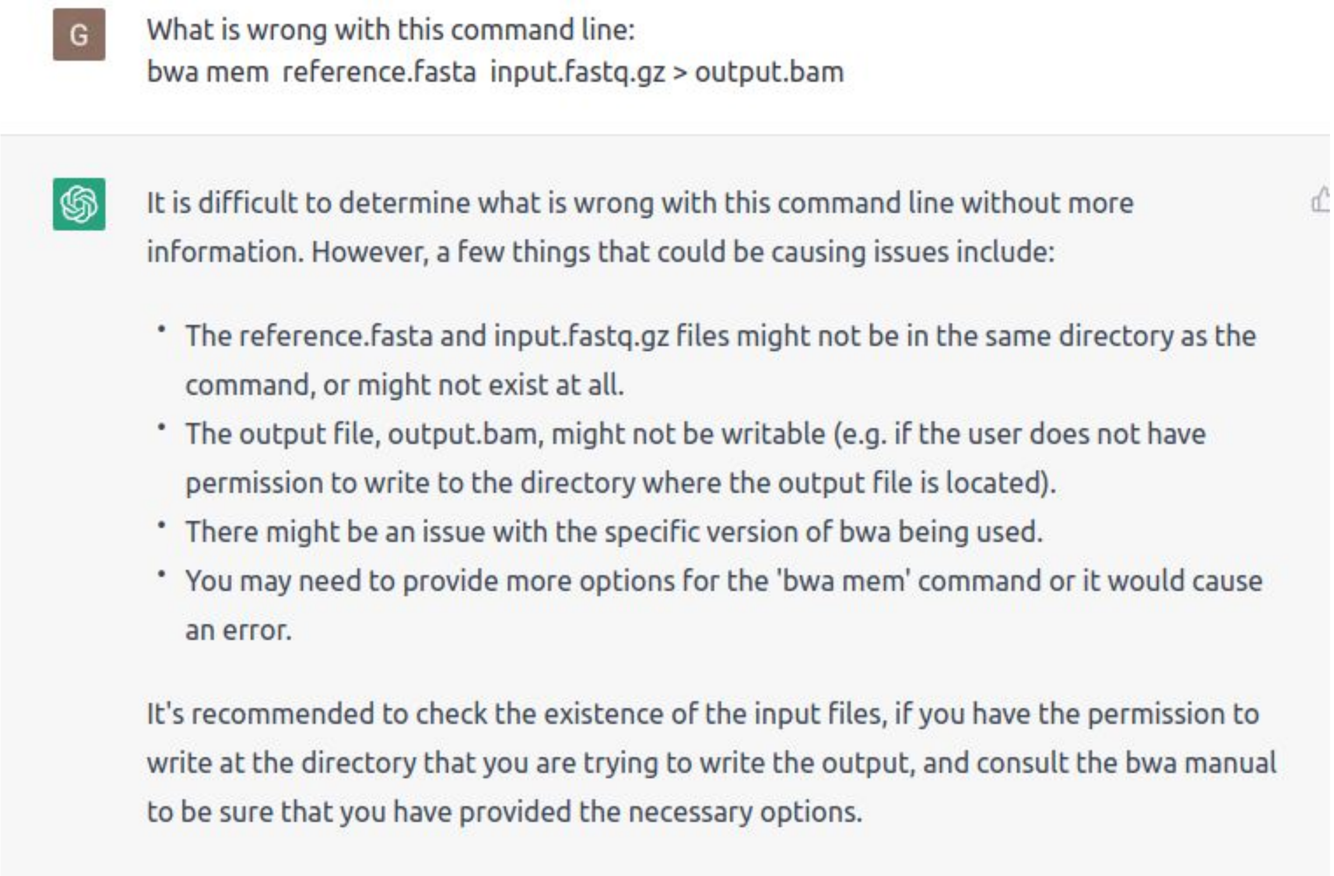
- Zip your vcf, text whatever files
    - there are tools to work with zipped files (zcat, zgrep, zless)
- Use BAM/CRAM/vcf.gz **never sam/vcf**
- Use symlinks instead of copying (no wasted space!):

```
ln -s /path/to/largefile .
```

- Beware, what is wrong with this?:

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

2023:



What is wrong with this command line:

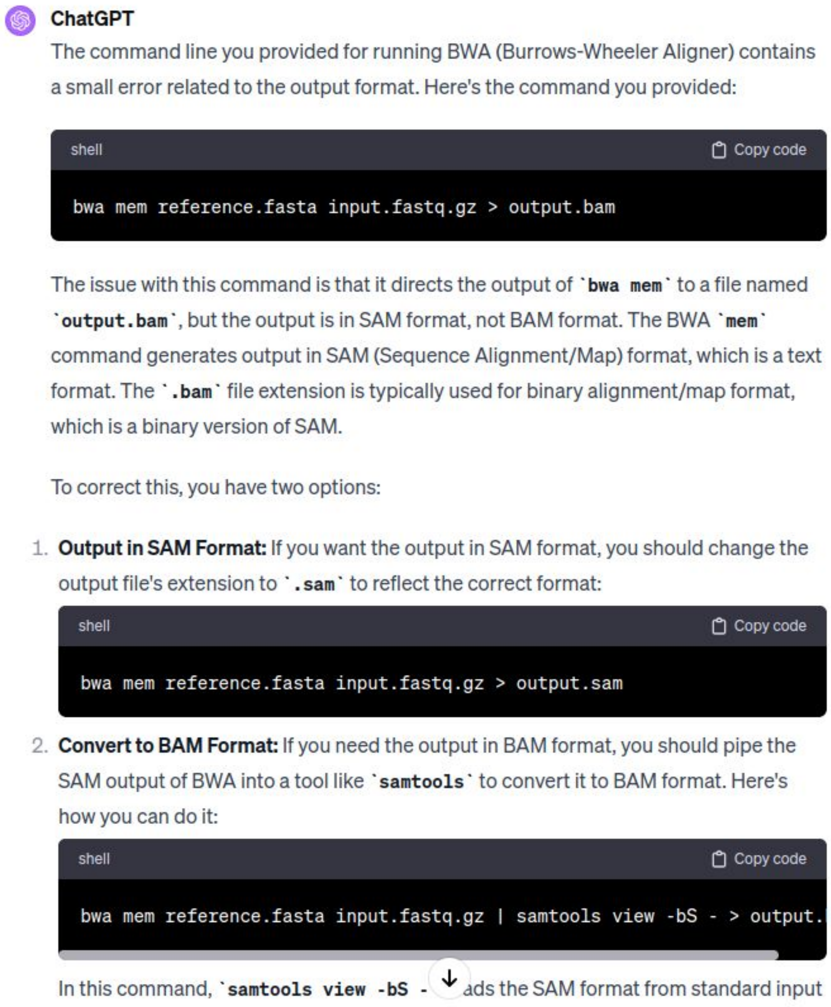
```
bwa mem  reference.fasta  input.fastq.gz > output.bam
```

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:



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

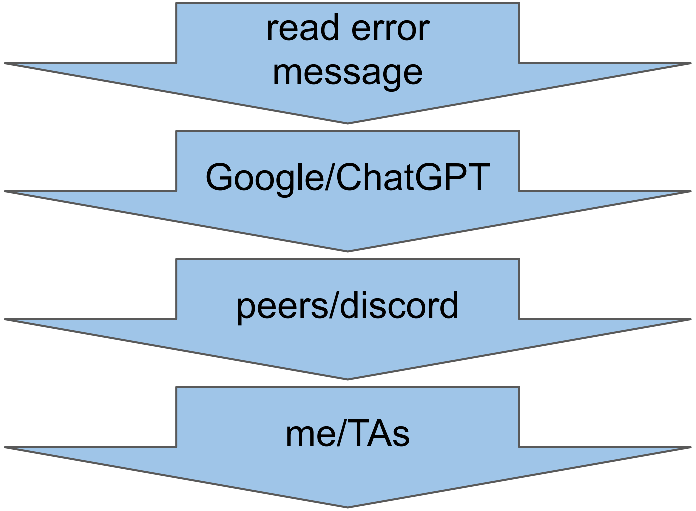
2. **Convert to BAM Format:** If you need the output in BAM format, you should pipe the SAM output of BWA into a tool like `samtools` to convert it to BAM format. Here's how you can do it:

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

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

How to get help?

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



read error message

Google/ChatGPT

peers/discord

me/TAs

The final exam

- Friday, Jan 24th at 9am
- I should be around
- Multiple choice
- 75% of the grade

The final exam

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

Things to remember

- I am not big on memorizing but…
- … please remember
  - Names of sequencing techs: e.g. Illumina, ONT
  - Their idiosyncrasies
  - Some basic formats we saw (e.g. FASTA, BAM, BED…)

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

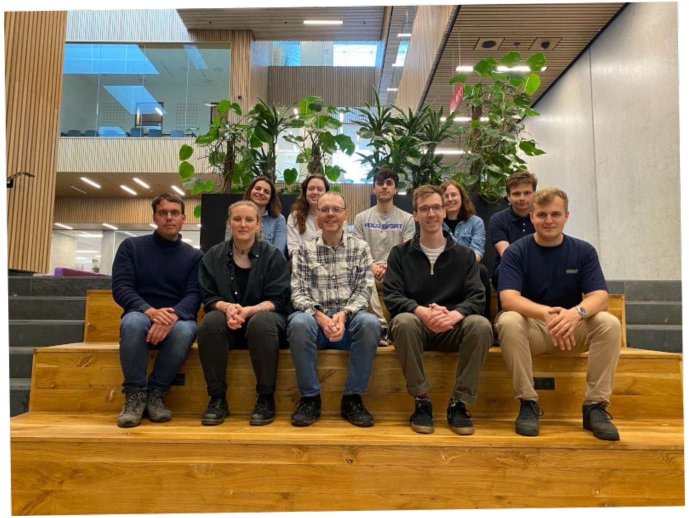
- Know the history of the seq. tech you are using as you may work with older data

- 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

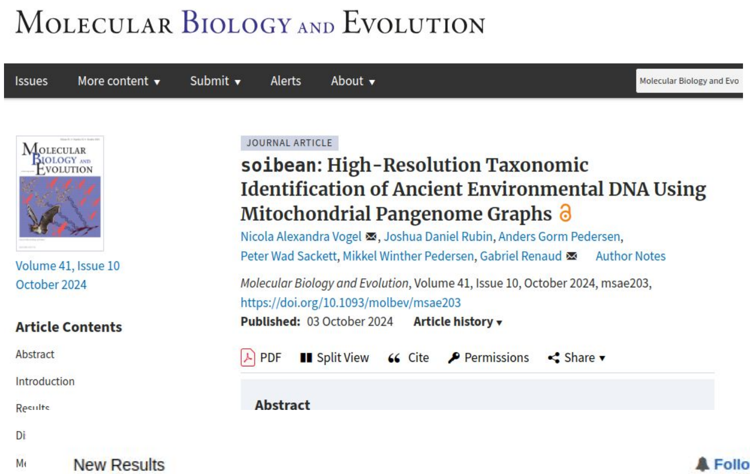
- Like NGS? Genotyping? Population genetics? Ancient DNA? pangenome graphs? sedimentary DNA?



The Modern and Ancient Genomes Group

- me
- 1 Postdoc
- 2 PhDs
- 4 Master's

Articles in 2024



MOLECULAR BIOLOGY AND EVOLUTION

Volume 41, Issue 10
October 2024

Article Contents

JOURNAL ARTICLE

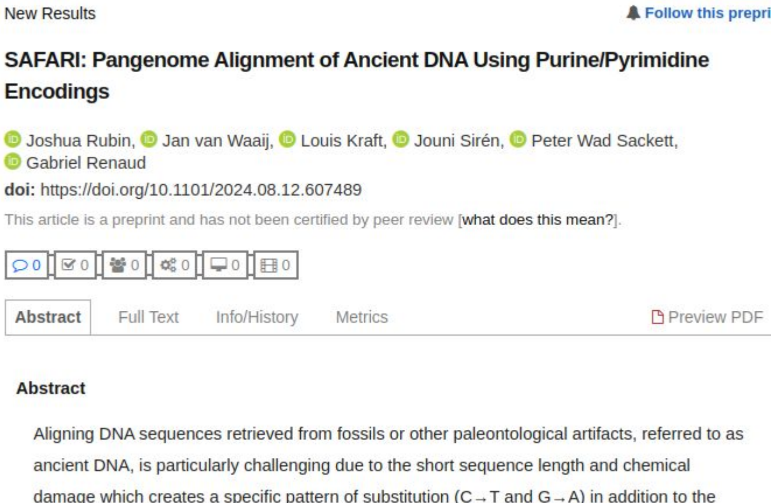
soibean: High-Resolution Taxonomic Identification of Ancient Environmental DNA Using Mitochondrial Pangenome Graphs

Nicola Alexandra Vogel, Joshua Daniel Rubin, Anders Gorm Pedersen, Peter Wad Sackett, Mikkel Winther Pedersen, Gabriel Renaud Author Notes

Molecular Biology and Evolution, Volume 41, Issue 10, October 2024, msae203, https://doi.org/10.1093/molbev/msae203

Published: 03 October 2024 **Article history** ▾

Abstract



New Results

SAFARI: Pangenome Alignment of Ancient DNA Using Purine/Pyrimidine Encodings

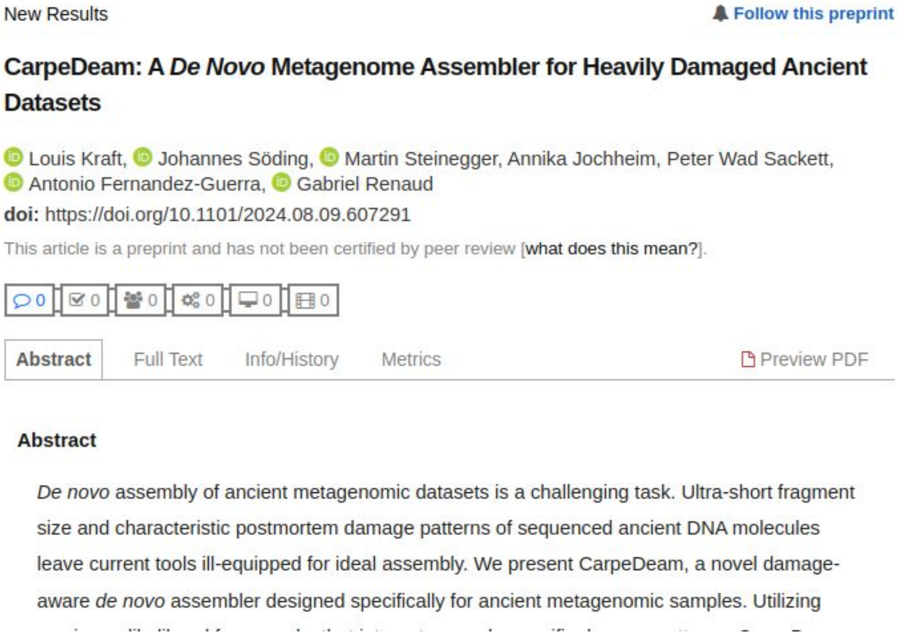
Joshua Rubin, Jan van Waaij, Louis Kraft, Jouni Sirén, Peter Wad Sackett, Gabriel Renaud

doi: https://doi.org/10.1101/2024.08.12.607489

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract

Aligning DNA sequences retrieved from fossils or other paleontological artifacts, referred to as ancient DNA, is particularly challenging due to the short sequence length and chemical damage which creates a specific pattern of substitution (C→T and G→A) in addition to the



New Results

CarpeDeam: A *De Novo* Metagenome Assembler for Heavily Damaged Ancient Datasets

Louis Kraft, Johannes Söding, Martin Steinegger, Annika Jochheim, Peter Wad Sackett, Antonio Fernandez-Guerra, Gabriel Renaud

doi: https://doi.org/10.1101/2024.08.09.607291

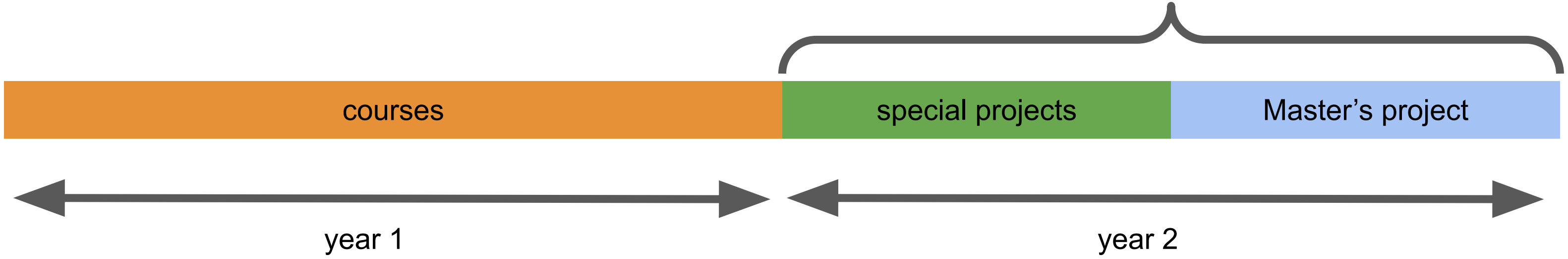
This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract

De novo assembly of ancient metagenomic datasets is a challenging task. Ultra-short fragment size and characteristic postmortem damage patterns of sequenced ancient DNA molecules leave current tools ill-equipped for ideal assembly. We present CarpeDeam, a novel damage-aware *de novo* assembler designed specifically for ancient metagenomic samples. Utilizing

Special projects/Master’s projects

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



just research! can be same project

courses

special projects

Master’s project

year 1

year 2

Thanks!