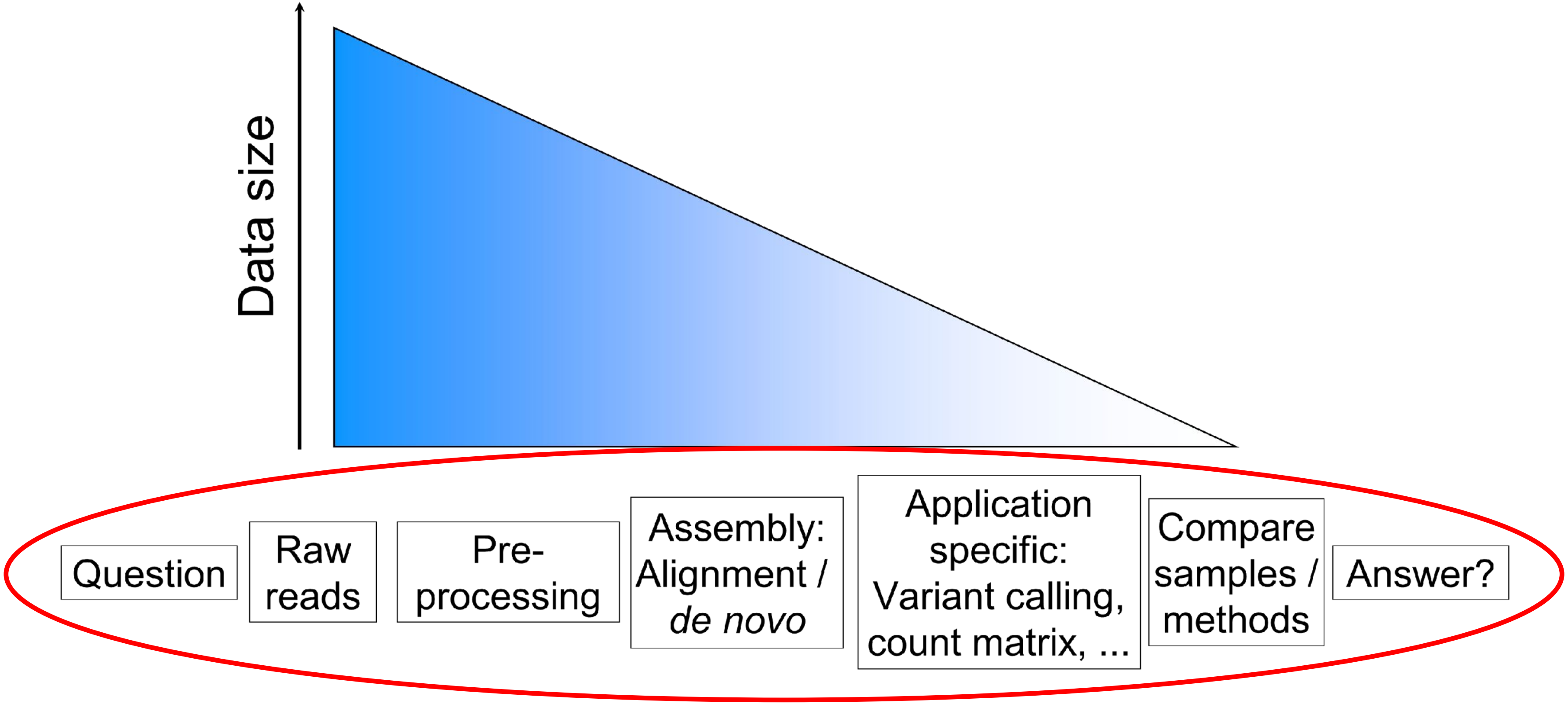
DTU

DTU Health Technology
Bioinformatics

Projects

Gabriel Renaud
Associate Professor
Section of Bioinformatics
Technical University of Denmark
gabre@dtu.dk

Generalized NGS analysis



Question

Raw reads

Pre-processing

Assembly: Alignment / *de novo*

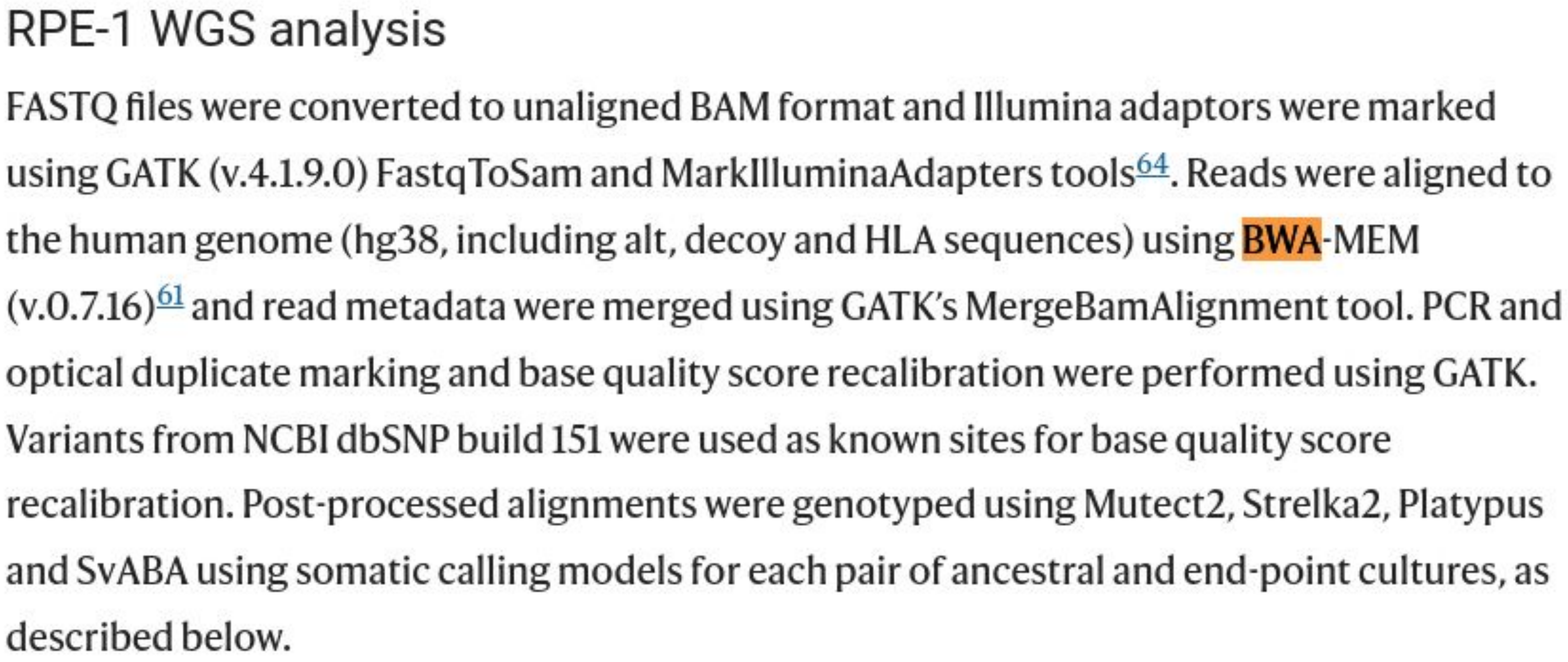
Application specific: Variant calling, count matrix, ...

Compare samples / methods

Answer?

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

Why are we here?



RPE-1 WGS analysis

FASTQ files were converted to unaligned BAM format and Illumina adaptors were marked using GATK (v.4.1.9.0) FastqToSam and MarkIlluminaAdapters tools[64]. Reads were aligned to the human genome (hg38, including alt, decoy and HLA sequences) using BWA-MEM (v.0.7.16)[61] and read metadata were merged using GATK's MergeBamAlignment tool. PCR and optical duplicate marking and base quality score recalibration were performed using GATK. Variants from NCBI dbSNP build 151 were used as known sites for base quality score recalibration. Post-processed alignments were genotyped using Mutect2, Strelka2, Platypus and SvABA using somatic calling models for each pair of ancestral and end-point cultures, as described below.

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
- 4-5 pr. group
- You can find a dataset on SRA/ENA
- Try to find raw data, untrimmed
  - If not, please contact us

Projects

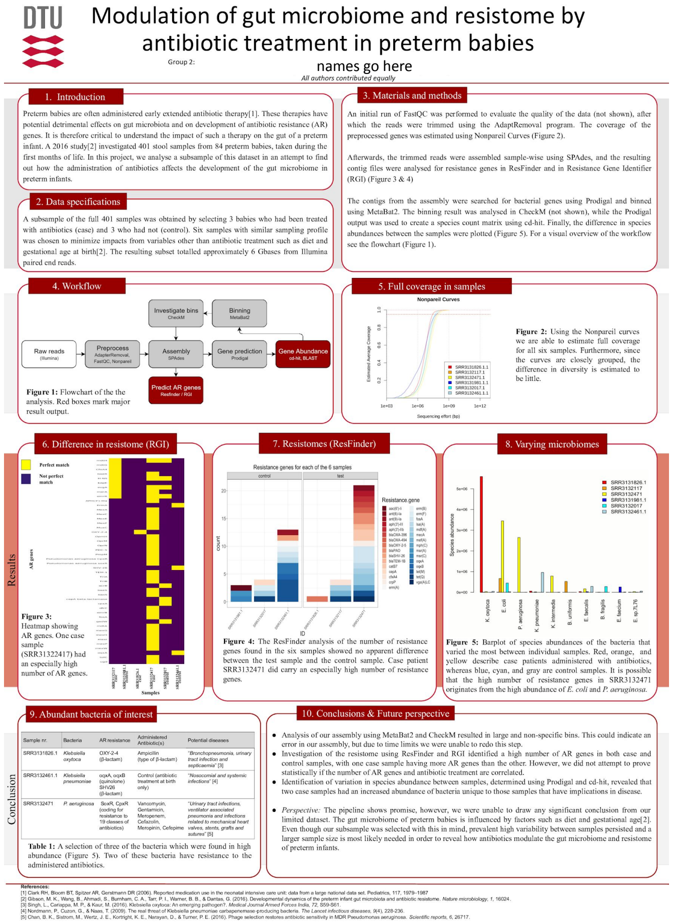
- You can use your own data if everyone in the group agrees ***and*** it can be presented on a poster
- Subset! Do not analyze very large datasets (time, resources)
- Subset! Do not replicate every figure/table!

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

Previous projects



DTU

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

Raw reads (Illumina)

Preprocess AdapterRemoval, FastQC, Nonpareil

Assembly SPAdes

Gene prediction Prodigal

Gene Abundance cd-hit, BLAST

Binning MetaBat2

Investigate bins CheckM

Predict AR genes Resfinder / RGI

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 poster

- You can print posters at the DTU library for 20-30kr

Posters

- Each group will create a poster

- ~~You can print posters at the DTU library for 20-30kr~~

Online this year: send us a high resolution PDF!

Posters

1) The group number, student names and student numbers of all group members, must be stated on the poster

2) The poster must specify the individual students contribution to the project.
It is allowed to state that everyone contributed equally

~~3) The poster must not extend the poster board (160 cm high, 120 cm wide)~~
~~**Note**, If you print through [illegible] the poster dimensions are: 1189mm x 841mm~~

4) Guide for making an good poster
http://wiki.bio.dtu.dk/teaching/index.php/Poster

Grouping & Guidance

- Fill in group information in Google doc
- 5 min presentation tomorrow at 13
  - What do you plan to do?
  - How much data?
- Project assistance: every day
  - Teachers+TA via Discord
- Data goes here:

```
/data/shared/groups/group_X
```

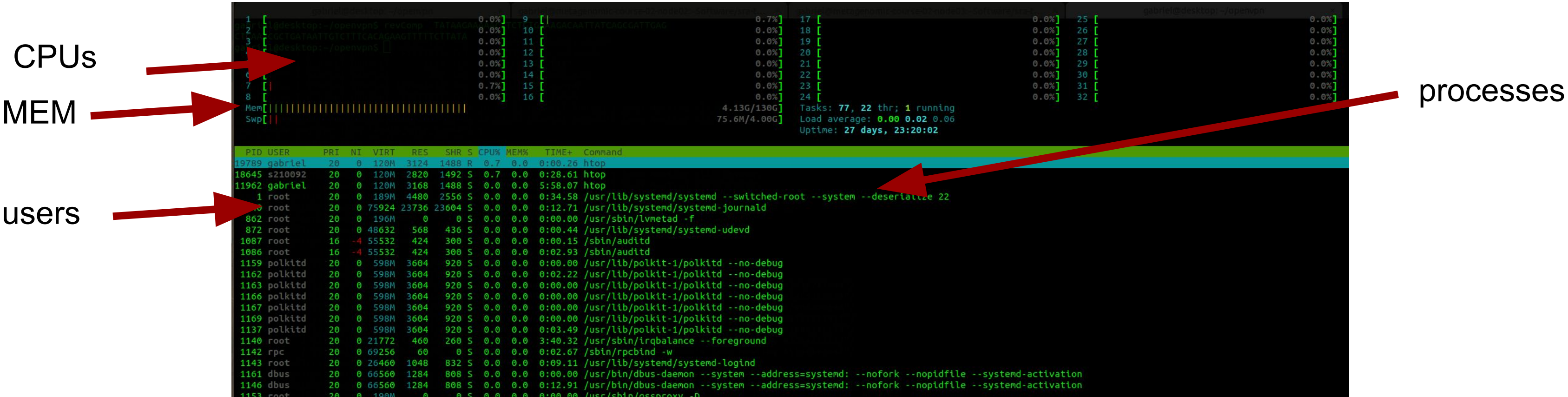
Be nice!

- Run larger programs on the servers using nice eg.

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

- How much memory am I using?

```
htop
```



CPUs

MEM

users

processes

```
  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%]
  4  [                                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
 1153 root       20   0  190M     0     0 S  0.0  0.0  0:00.00 /usr/sbin/gssproxy -D
```

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
```

Evaluation: presentation and oral exam

- You will give a group presentation about your poster (4-5 minutes)
  - each person should speak at least once.
  - what the study was about
  - what you have done
  - results you got:
    - Quality of data, replicate certain results, pitfalls

Evaluation: presentation and oral exam

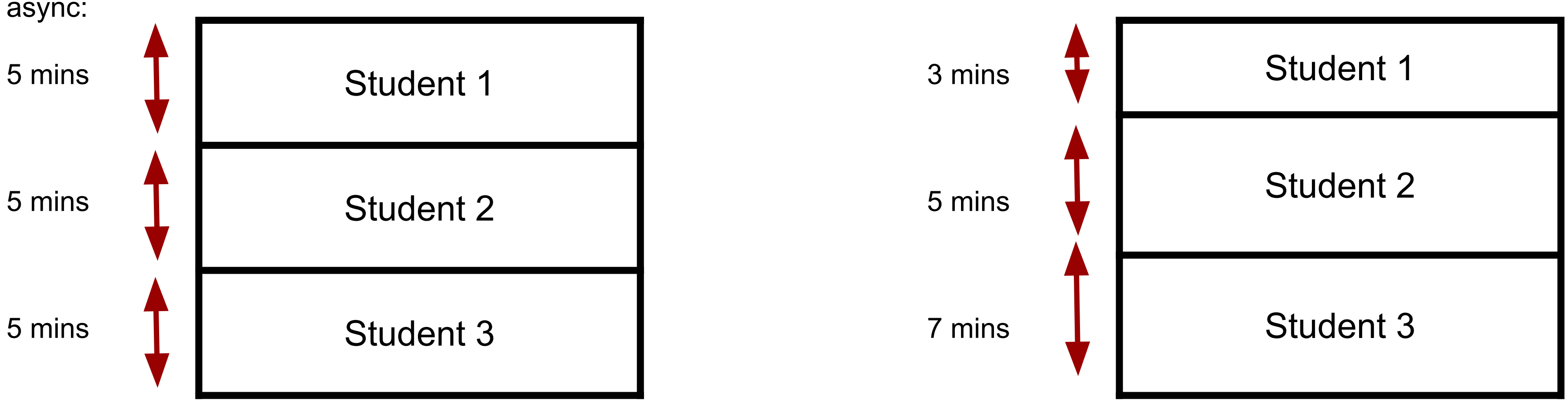
- We will ask one person at a time to come and we will ask you about 4-5 questions about the project:
  - The goal: did you understand what we taught in class and what you did
  - We can quiz you on your project and can have notions of what we saw in class
  - Do not memorize, **understand**
  - Do not communicate with others in your group

- 2 evaluators will meet and the final mark will be a blend of your oral exam, group performance (minor tech talks) can help us distinguish between a 10 or 12.

Evaluation: presentation and oral exam

2 models:

sync:
async:



5 mins Student 1
5 mins Student 2
5 mins Student 3

3 mins Student 1
5 mins Student 2
7 mins Student 3

Evaluation: presentation and oral exam

- I favor async because students:
  - are nervous
  - not native English speakers
  - understand but need time to gather their thoughts
- Disadvantages: You do not know when you will go
- How to survive the exam:
  - Get water+snacks
  - Ask if you want to go for lunch

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
  - 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
- Do your literature search, use existing tools when possible

Special projects/Master's projects

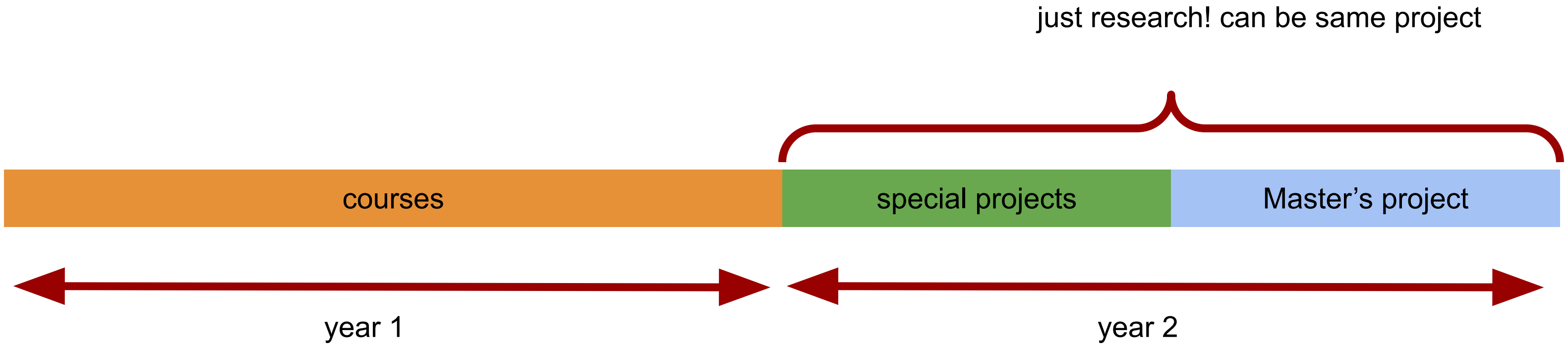
- Like NGS? Genotyping? Population genetics? Ancient DNA?

The Modern and Ancient Genomes Group

- me
- 2 Postdocs
- 3 PhDs
- 3 Master's
- 2 undergrads

Special projects/Master's projects

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



Thanks!