

DTU





**DTU Health Technology
Bioinformatics**

**22126: Next Generation Sequencing Analysis
DTU - January 2022
Gabriel Renaud**

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

Who am I?

- PhD in Bioinformatics from Max Planck Institute in Leipzig
- Postdoc at KU
- Associate Professor at DTU in Dec. 2019
- Worked since 2006 with NGS
- slow response: gabre [at] dtu [dot] dk
- fast response: gabriel [dot] reno [at] gmail [dot] com

Who am I?

What keeps me busy:

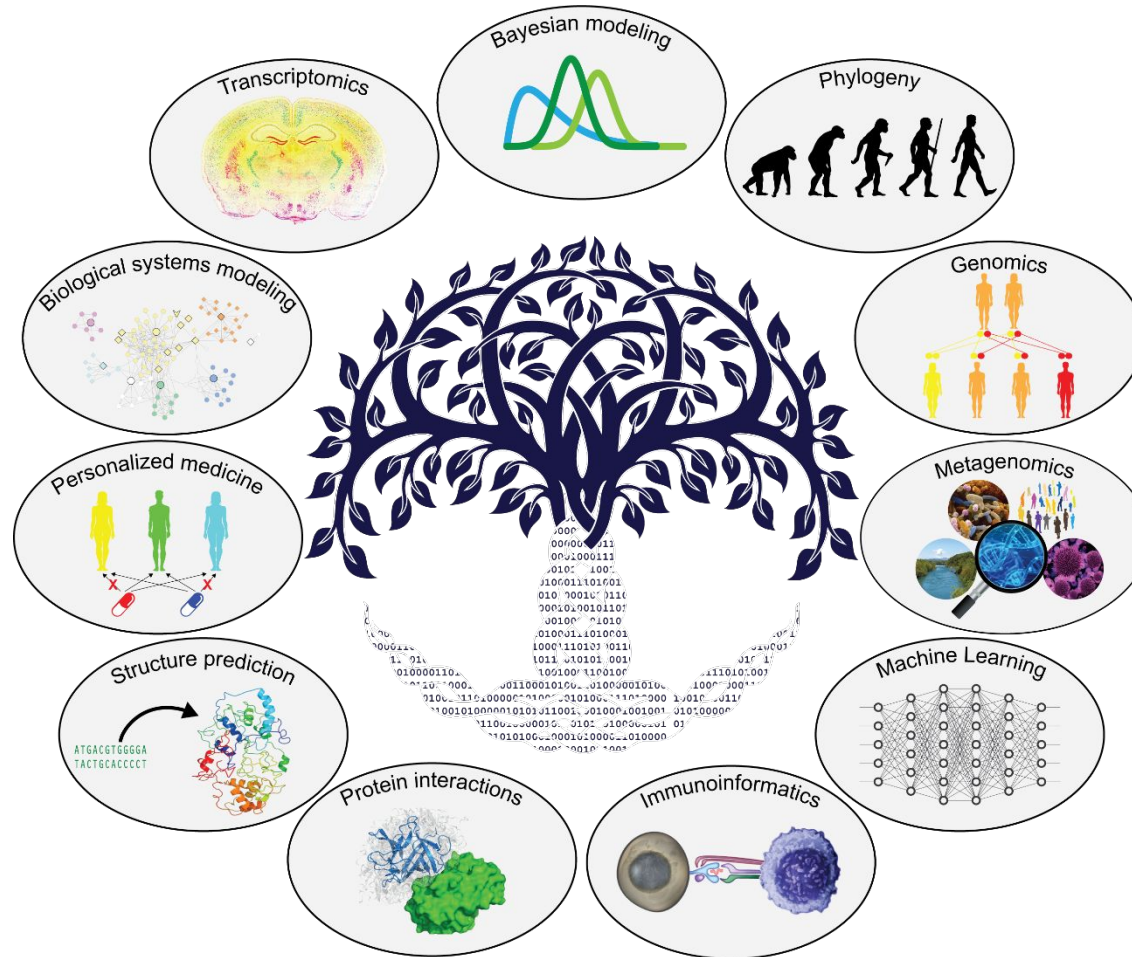
- Methods for NGS analysis
- Ancient DNA and modern samples
- Large sets of genotypes

Looking to do a masters' project dealing with NGS, email me!

Who are we?

- Organizer:
 - Gabriel Renaud
 - Shyam Gopalakrishnan
 - Gisle Vestergaard
 - Josh Rubin
 - Nicola Vogel
 - DTU Bioinformatics
 - Peter Wad Sackett
- DTU Aqua
 - Francesca Bertolini
- DTU Food
 - Pimlapas Leekitechaoenphon (Shinny)
- Copenhagen University:
 - Martin Sikora
- Aarhus University
 - Søren Besenbacher

What do we do?



Main teaching assistants

Josh Rubin <jdru@dtu.dk>

Nicola Vogel <navo@food.dtu.dk>

Online class this year

Discord/Zoom:

- Feel free to turn off your cam when you need
- But I do like seeing faces :-)
- Evaluations: we need to see you
- I conduct polls
- Ask questions please:
 - unmute and start talking
 - raise your hand
 - type in the chat
- work in teams
- office hours on Discord



Online class this year

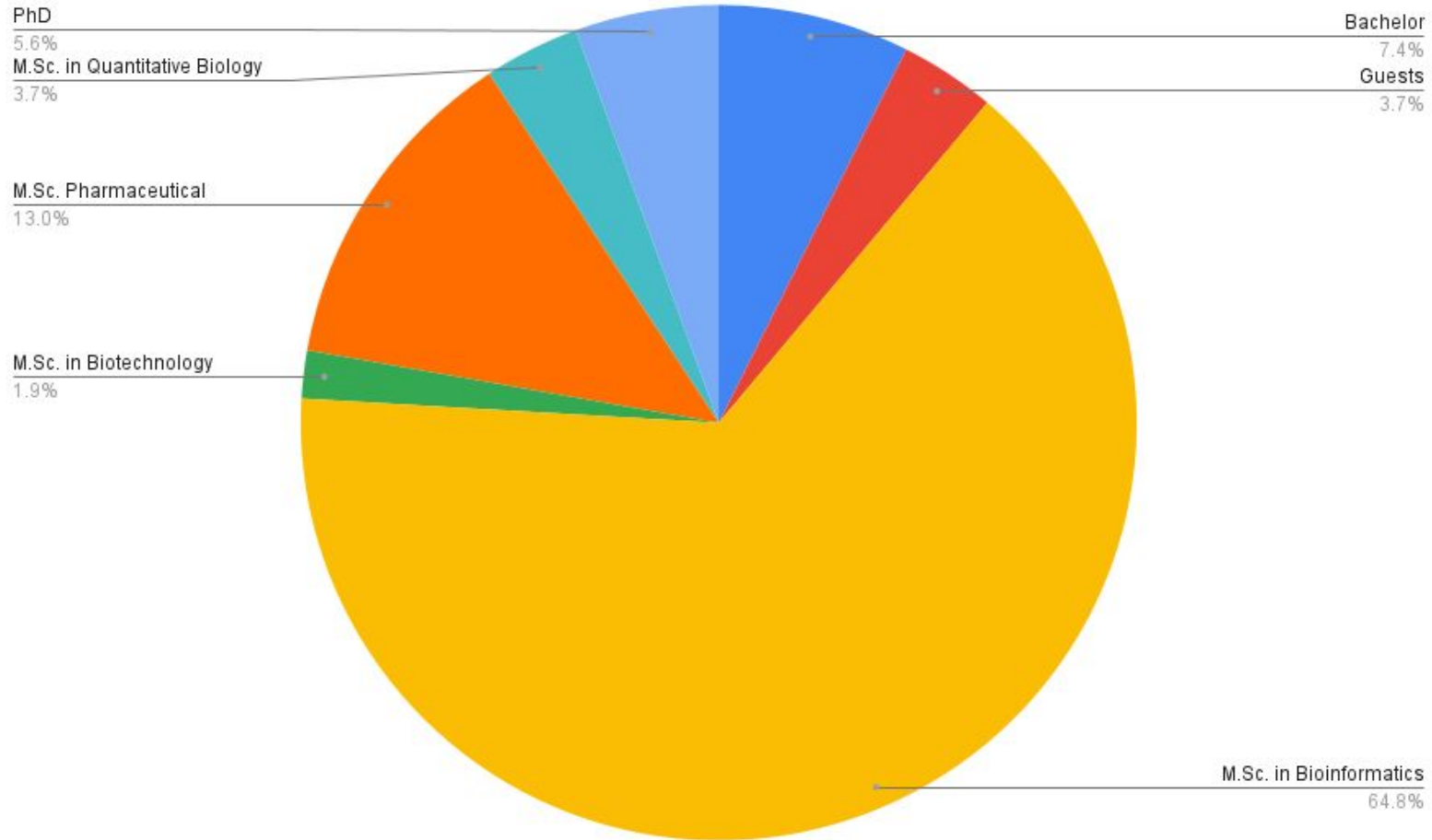
If my internet connection drops, please stay! I will come back

Schedule, exercises, general plan:

https://teaching.healthtech.dtu.dk/22126/index.php/Program_2021

Who are you?

January 2022



Feedback

- 10th time we are running the course
- My 3rd time!
- Second time online
- We are still improving
- It is very difficult to keep up with new tech...
- NGS is very broad now, no one masters everything
- Please give us feedback !
 - Please do the evaluation at DTU Inside



Why are we here?



Autism risk in offspring can be assessed through quantification of male sperm mosaicism

Martin W. Breuss^{1,2}, Danny Antaki^{3,4,5,6}, Renee D. George^{1,2}, Morgan Kleiber^{3,4,5}, Kiely N. James^{1,2}, Laurel L. Ball^{1,2}, Oanh Hong^{3,4,5,6}, Ileena Mitra^{7,8}, Xiaoxu Yang ^{1,2}, Sara A. Wirth^{1,2}, Jing Gu^{1,2}, Camila A. B. Garcia^{1,2}, Madhusudan Gujral^{3,4,5,6}, William M. Brandler^{3,4,5,6}, Damir Musaev^{1,2}, An Nguyen^{1,2}, Jennifer McEvoy-Venneri^{1,2}, Renatta Knox^{1,2,9}, Evan Sticca^{1,2}, Martha Cristina Cancino Botello¹⁰, Javiera Uribe Fenner¹⁰, Maria Cárcel Pérez¹¹, Maria Arranz¹¹, Andrea B. Moffitt¹², Zihua Wang¹², Amaia Hervás¹³, Orrin Devinsky ¹⁴, Melissa Gymrek^{7,8}, Jonathan Sebat ^{3,4,5,6*} and Joseph G. Gleeson ^{1,2*}

De novo mutations arising on the paternal chromosome make the largest known contribution to autism risk, and correlate with paternal age at the time of conception. The recurrence risk for autism spectrum disorders is substantial, leading many families to decline future pregnancies, but the potential impact of assessing parental gonadal mosaicism has not been considered. We measured sperm mosaicism using deep-whole-genome sequencing, for variants both present in an offspring and evident only in father's sperm, and identified single-nucleotide, structural and short tandem-repeat variants. We found that mosaicism quantification can stratify autism spectrum disorders recurrence risk due to de novo mutations into a vast majority with near 0% recurrence and a small fraction with a substantially higher and quantifiable risk, and we identify novel mosaic variants at risk for transmission to a future offspring. This suggests, therefore, that genetic counseling would benefit from the addition of sperm mosaicism assessment.

Published: 23 December 2019

Why are we here?

WES and WGS trio analysis. WGS sequencing and analysis for F01–08 and F13–20 were performed as described previously^{13,37}. Exome capture and sequencing of F09–12 were performed at the New York Genome Center (Agilent Human All Exon 50 Mb kit, Illumina HiSeq 2000, paired-end, 2 × 100) and the Broad Institute (Agilent Sure-Select Human All Exon v.2.0, 44-Mb baited target, Illumina HiSeq 2000, paired-end, 2 × 76). Sequencing reads were aligned to the hg19 reference genome using BWA (v.0.7.8). Duplicates were marked using Picard's MarkDuplicates (v.1.83, <http://broadinstitute.github.io/picard>) and reads were realigned around insertion/deletions (InDels) with GATK's IndelRealigner. Variant calling for SNVs and InDels was performed according to GATK's best practices by first calling variants in each sample with HaplotypeCaller and then jointly genotyping them across the entire cohort using CombineGVCFs and GenotypeGVCFs. Variants were annotated with SnpEff (v.4.2) and SnpSift (v.4.2), and allele frequencies from the 1000 Genomes Project and the Exome Aggregation Consortium (ExAC)³⁸. De novo variants were called for probands using Triodenovo (v.0.06) with a minimum de novo quality score of 2.0 and subjected to manual inspection. Variants from F01–F08 were further

Why are we here?

“Around 2 a.m. on Jan. 5, after working over 40 hours straight, Dr. Zhang and his team at the Shanghai Public Health Clinical Center sequenced the unknown virus on the NovaSeq™ 6000 System. They published its genome on **Jan. 10th 2020.**”

<https://www.illumina.com/company/news-center/blog/2020-in-genomics.html>



Yong-Zhen Zhang

Why are we here?

“... Moderna’s mRNA-1273, which reported a 94.5 percent efficacy rate on November 16, had been designed by **January 13th 2020**. This was just **two days** after the genetic sequence had been made public

...

It was completed [...] **more than a week before** the first confirmed coronavirus case in the United States.”



Yong-Zhen Zhang

<https://nymag.com/intelligencer/2020/12/moderna-covid-19-vaccine-design.html>

Not a wet lab course...



...it's a computational one



Tips

Tip: Do not memorize the name of the tools/procedure, they come and go



Tips

Tip: Understand the problem and how various tools work

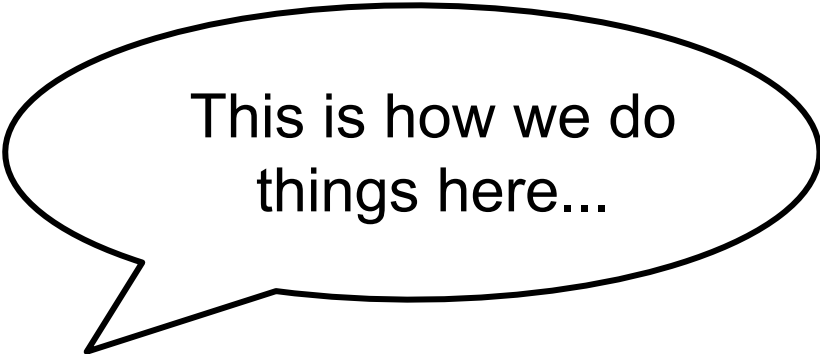


Tips for NGS in general

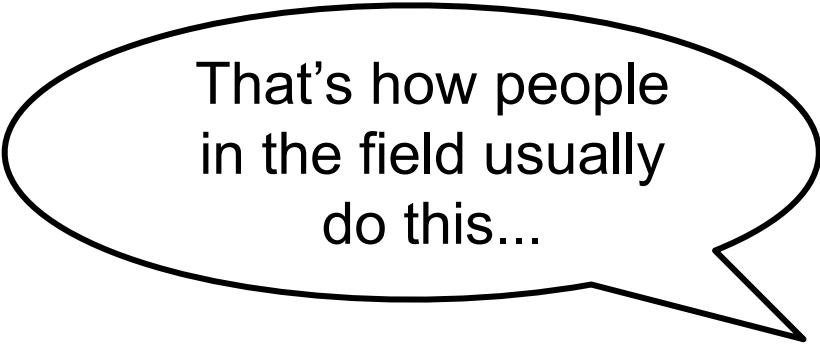
- New tools or procedures get released all the time
- The best tool/format/pipeline in 2022 may not be the best in 2032
- Understand how they work, in which cases they perform well

Tips for NGS in general

- Read benchmarking papers and reviews
- Beware of:



This is how we do things here...

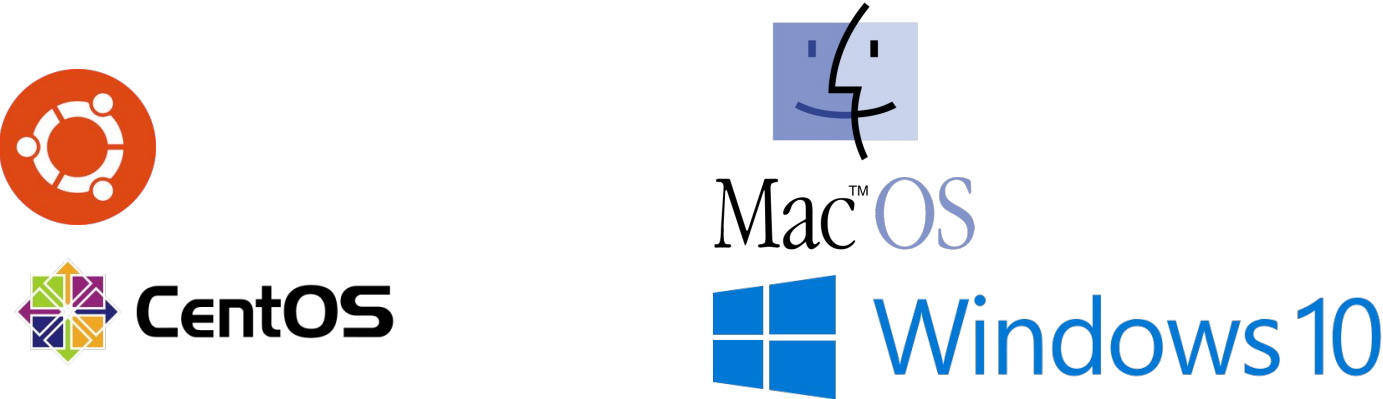


That's how people in the field usually do this...

The shell terminal

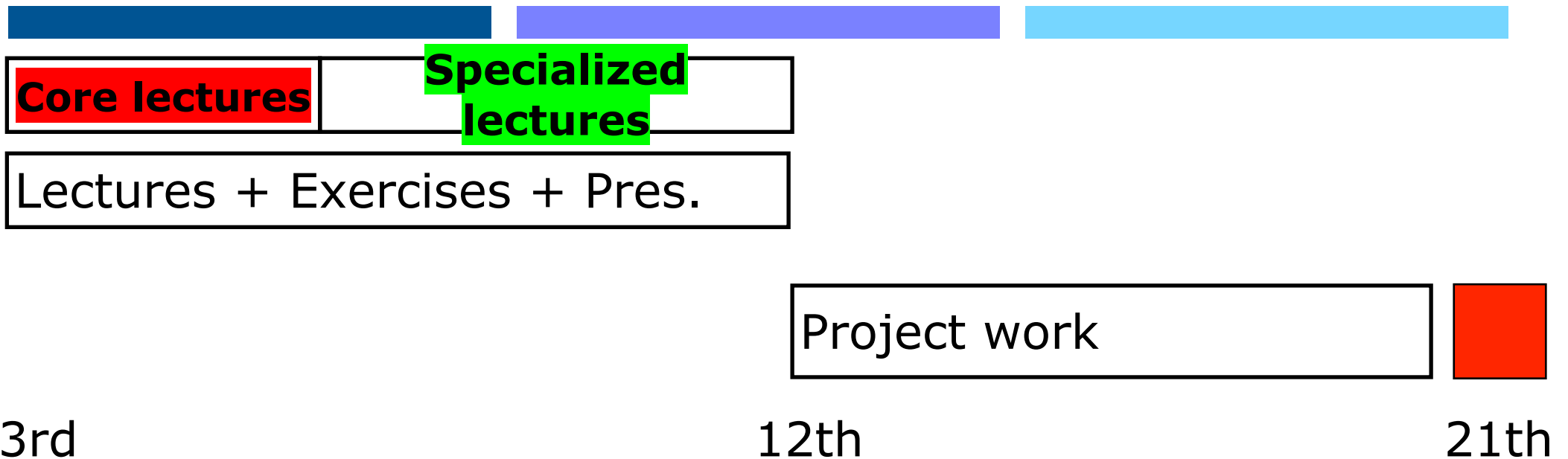
```
gabriel@desktop: /tmp$ cat NE021.ban | head -n 20
K00234:70:HNNW3B8XX:8:1285:18142:9084  0      ntrfref  1      250      7550M *      823  0      GCACATTTGGTTCCTACTTCAGGGCCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCAC
K00234:70:HNNW3B8XX:8:1285:18142:9084  0      ntrfref  1      250      7550M *      823  0      GCACATTTGGTTCCTACTTCAGGGCCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCAC
K00234:70:HNNW3B8XX:8:1124:22343:30222  16     ntrfref  1      250      7853M *      823  0      ATCCGCACATCTGGTTCCTACTTCAGGGCCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGAT
K00234:70:HNNW3B8XX:8:1284:27336:16306  16     ntrfref  1      250      7550M *      823  0      TGATATCTGGTTCCTACTTCAGGGCCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCAC
K00234:70:HNNW3B8XX:8:1284:25337:7978   16     ntrfref  1      250      88513M *     823  0      TGGTTCCTATTTTAGGGCCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCACAGGTCTA
K00234:70:HNNW3B8XX:8:1281:11972:21430  16     ntrfref  1      250      7457M *      823  0      GCACATCTGGTTCCTACTTCAGGGCCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCAC
K00234:70:HNNW3B8XX:8:1217:10236:11530  16     ntrfref  1      250      7152M *      823  0      ATCTGGTTCCTACTTCAGGGCCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGAA
K00234:70:HNNW3B8XX:8:1285:18142:9084  0      ntrfref  1      250      6250M *      823  0      CTACTTCAGGGCCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCACAGGTCTA
K00234:70:HNNW3B8XX:8:1283:3772:45097   16     ntrfref  1      250      67511M *     823  0      GGTTCTACTTCAGGGCCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCACAGGTC
K00234:70:HNNW3B8XX:8:1284:25337:7978   16     ntrfref  1      250      88513M *     823  0      TGGTTCCTATTTTAGGGCCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCACAGGTCTA
K00234:70:HNNW3B8XX:8:2222:10815:11804  16     ntrfref  1      250      6159M *      823  0      TACTTTAGAGCCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCACAGGTC
K00234:70:HNNW3B8XX:8:1225:16102:17122  16     ntrfref  1      250      55517M *     823  0      AGGGCCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCACAGGTCTATCAC
K00234:70:HNNW3B8XX:8:1222:29203:16893   16     ntrfref  1      250      52529M *     823  0      ACCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAAGACATCAGCATGGATCACAGGTCATCACCCCTATTAA
K00234:70:HNNW3B8XX:8:1227:23977:26725  16     ntrfref  1      250      53520M *     823  0      GGCCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCACAGGTCATCACCCCTATTAA
K00234:70:HNNW3B8XX:8:1220:7415:14555   16     ntrfref  1      250      52519M *     823  0      GCCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCACAGGTCATCACCC
K00234:70:HNNW3B8XX:8:2226:16691:22819  16     ntrfref  1      250      51530M *     823  0      CCATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCACAGGTCATCACCCCTATTAA
K00234:70:HNNW3B8XX:8:2224:13453:21149  16     ntrfref  1      250      50531M *     823  0      CATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCACAGGTCATCACCCCTATTAA
K00234:70:HNNW3B8XX:8:2222:13808:14783   16     ntrfref  1      250      49532M *     823  0      ATAAAGCCCTAAATAGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCACAGGTCATCACCCCTATTAA
K00234:70:HNNW3B8XX:8:2125:1438:7433    16     ntrfref  1      250      41521M *     823  0      TAATATGCACAGCTCCCTTAAATAAGACATCAGCATGGATCACAGGTCATCACCCCTA
K00234:70:HNNW3B8XX:8:1286:26575:44166  16     ntrfref  1      250      36548M *     823  0      AGCCACAGCTCCCTTAAATAAGACATCAGCATGGATCACAGGTCATCACCCCTATTAA
K00234:70:HNNW3B8XX:8:1114:2412:16260   16     ntrfref  1      250      35548M *     823  0      GCCACAGCTCCCTTAAATAAGACATCAGCATGGATCACAGGTCATCACCCCTATTAA
JFA<-JF-FJFJAJFF-AJJK<JJJJJJJFJFJAJJJJF<-JFJFJJFFJJJFAF-JFAF-F<-<
gabriel@desktop: /tmp$
gabriel@desktop: /tmp$
gabriel@desktop: /tmp$
gabriel@desktop: /tmp$
gabriel@desktop: /tmp$
gabriel@desktop: /tmp$
gabriel@desktop: /tmp$
gabriel@desktop: /tmp$
```

- Available on various platforms



Course structure

- 3 weeks, 2 tracks



 = Poster exam

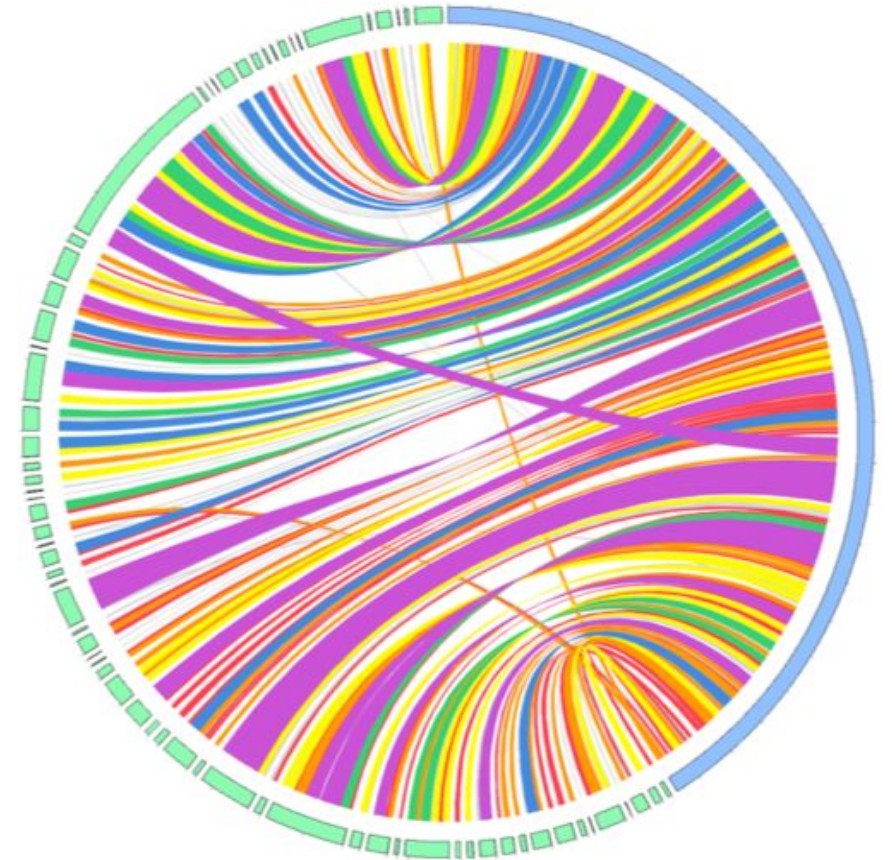
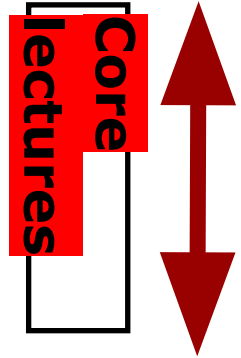
Course breakdown I

- Monday 3rd January
 - Introduction NGS technology
 - Tech talk groups
 - Unix and first look at data
- Tuesday 4th January
 - Data basics & preprocessing
 - Alignment



Course breakdown II

- Wednesday 5th January
 - Functional Human Variation
 - Alignment processing
 - *de novo* assembly
- Thursday 6th January
 - *de novo* metagenomics
 - Quantitative metagenomics



Course breakdown III

- Friday 9th January
 - cell free DNA
 - Recap test (after lunch)
- Monday 10th January
 - RNAseq
 - Cancer-seq
- Tuesday 11th January
 - Genomic Epidemiology
 - Tech talk work & Presentations

Course breakdown IV

- *Wednesday 12th January*
 - Ancient DNA
 - Project work
 - Prepare presentations for tomorrow
- *Thursday, 13th January*
 - Short project presentations
 - Project work
- *Friday 14th - Thursday 20th*
 - Project work
- *Friday 21st*
 - *Poster Exam*



Tech Talks

- More on this later...
- 4-5 pr. group
- Describe a sequencing protocol
- Prepare a short presentation

Projects

- Try to analyze an empirical dataset and present results on poster
- 4-5 pr. group
- You can find a dataset on SRA/ENA
- You can use your own data if everyone in the group agrees **and** it can be presented on a poster
- Do **not** analyze very large datasets (time, resources)

Points to remember

- Understand principles of the analysis
- The exercises will be useful for your projects and hopefully also later
- Have an exercise buddy and do them as a team, preferably on each individuals laptop so everyone gets to learn the command-line
- Please **just ask** questions at any time !

Cloud computing

- Virtual machines (you cannot break them!)
- Danish National Supercomputer for Life Science (Computerome) located at DTU Risø
- 16048 cores, 92 Tb RAM an 3Pb storage
- We have 3 nodes
 - Each has 28 cores and 128 Gb RAM



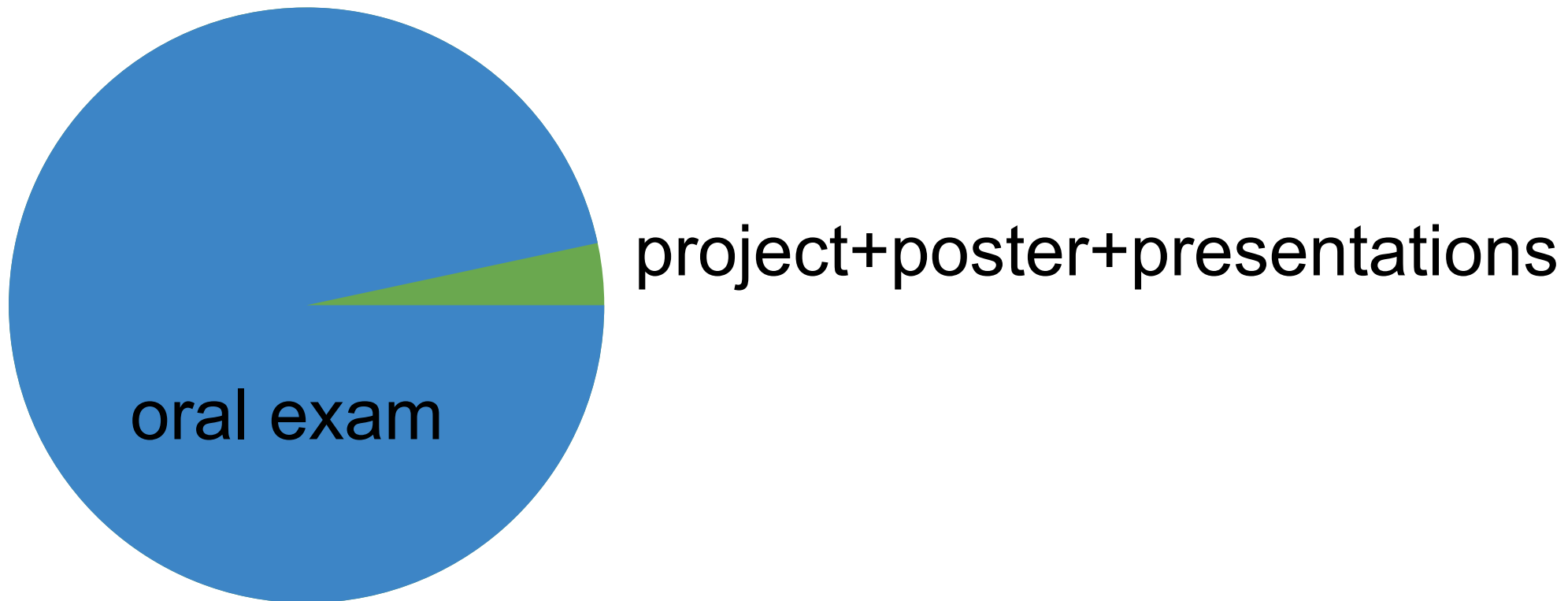
Exam

- Each group will create a poster
- Each group will present the poster for the examiners
- Then each individual in the group will one-by-one be asked questions on the core concepts and your project (5-10 min).
 - Do not memorize, **understand** what you are doing during the project
 - Understand the concepts taught in class

Tips for this class

- Do not memorize definitions, **understand** concepts
- The core lectures are especially crucial
- The final exam is an oral one which will evaluate your understanding, not whether you can parroting definitions
- Do the exercises! Understand what you are doing:
 - inspect the input
 - inspect the output
 - play with parameters

Marking scheme



Disclaimer

- Sequencing technology changes very rapidly!
- We will dive into many areas and you will not learn to master everything
- However, we hope that the building blocks we provide will allow you to see new opportunities

Be adventurous!

You do not have the ability to do anything destructive

Unless you physically destroy our computers!

The worst that can happen is that you lose your own data

Course webpage

- Course program, slides, handouts, exercises etc.
- <http://teaching.healthtech.dtu.dk/22126>
- We want the course page to be a repository for you!

Reading + wifi

- There are no textbooks for the course, it changes too rapidly
- Wireless networks
 - Use “dtu” and your dtu/campusnet login to get access to wireless
 - Eduroam
 - Alternative wifi: “You can haz wifi”

Pre-test

- Test your knowledge before we start
- Not used for grading or exam
- Used to understand where you are and what you need