
DTU

DTU Health Technology
Bioinformatics

22126: Next Generation Sequencing Analysis
DTU - January 2021
Gabriel Renaud

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

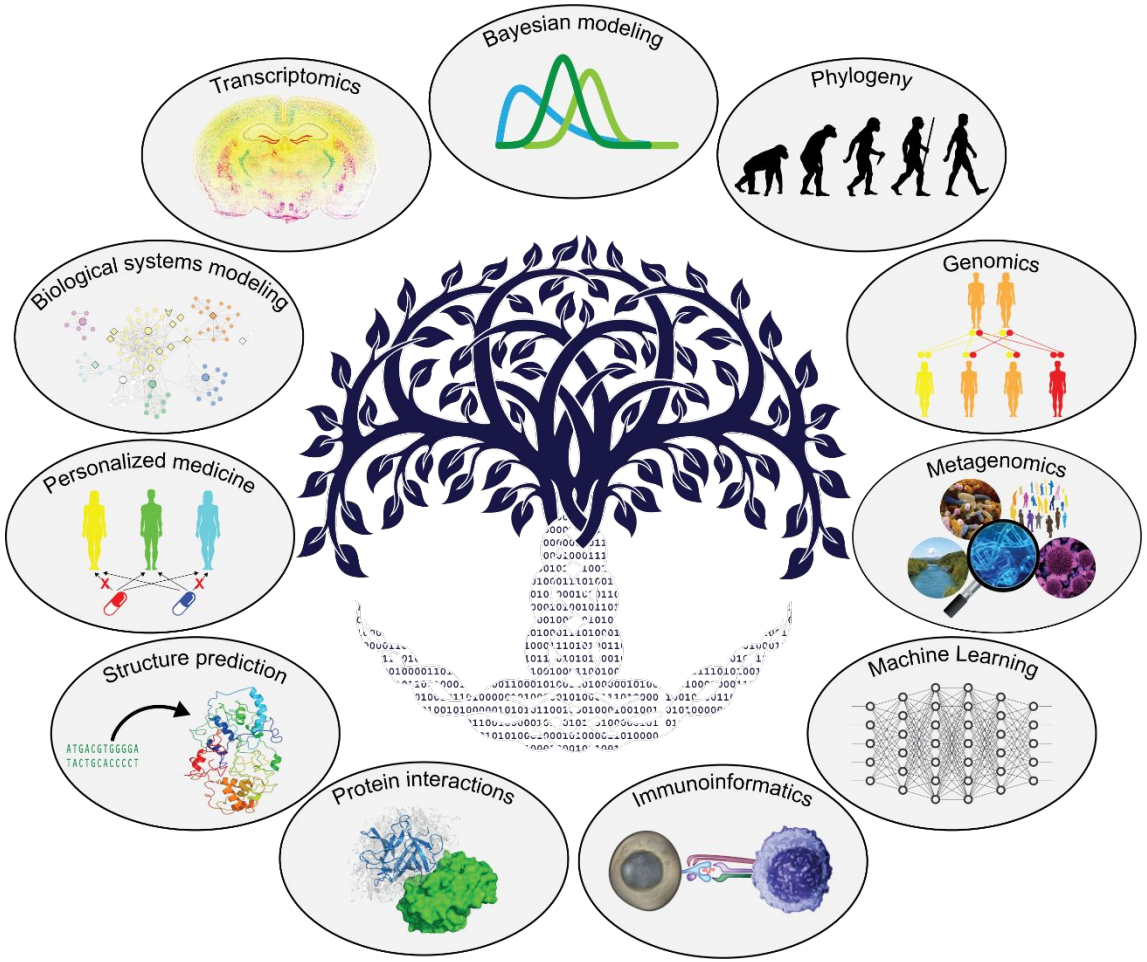
Who am I?

- 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 are we?

- Organizer:
  - Gabriel Renaud
  - Shyam Gopalakrishnan
  - Gisle Vestergaard
  - Trine Zachariasen
  - DTU Bioinformatics
  - Peter Wad Sackett
- DTU Aqua
  - Francesca Bertolini
- DTU Food
  - Pimlapas Leekitecharoenphon (Shinny)
- Copenhagen University:
  - Martin Sikora
- Aarhus University
  - Søren Besenbacher

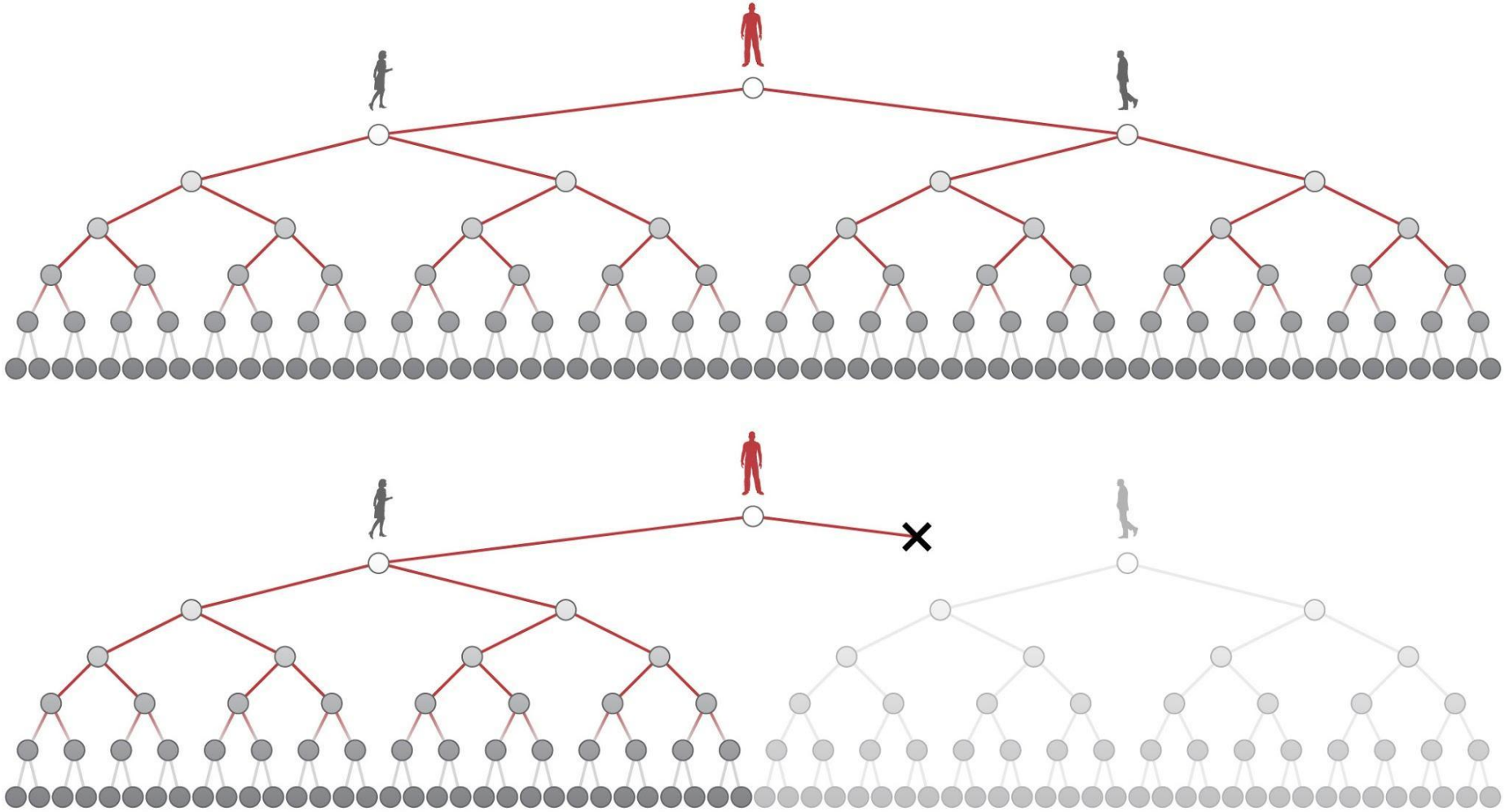
What do we do?


Bayesian modeling
Transcriptomics
Phylogeny
Biological systems modeling
Genomics
Personalized medicine
Metagenomics
Structure prediction
ATGACGTGGGGA
TACTGCACCCCT
Machine Learning
Protein interactions
Immunoinformatics

Main teaching assistants

Trine Zachariasen <trizac@dtu.dk>

Online class this year



source: NYTimes

Online class this year

Discord:

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

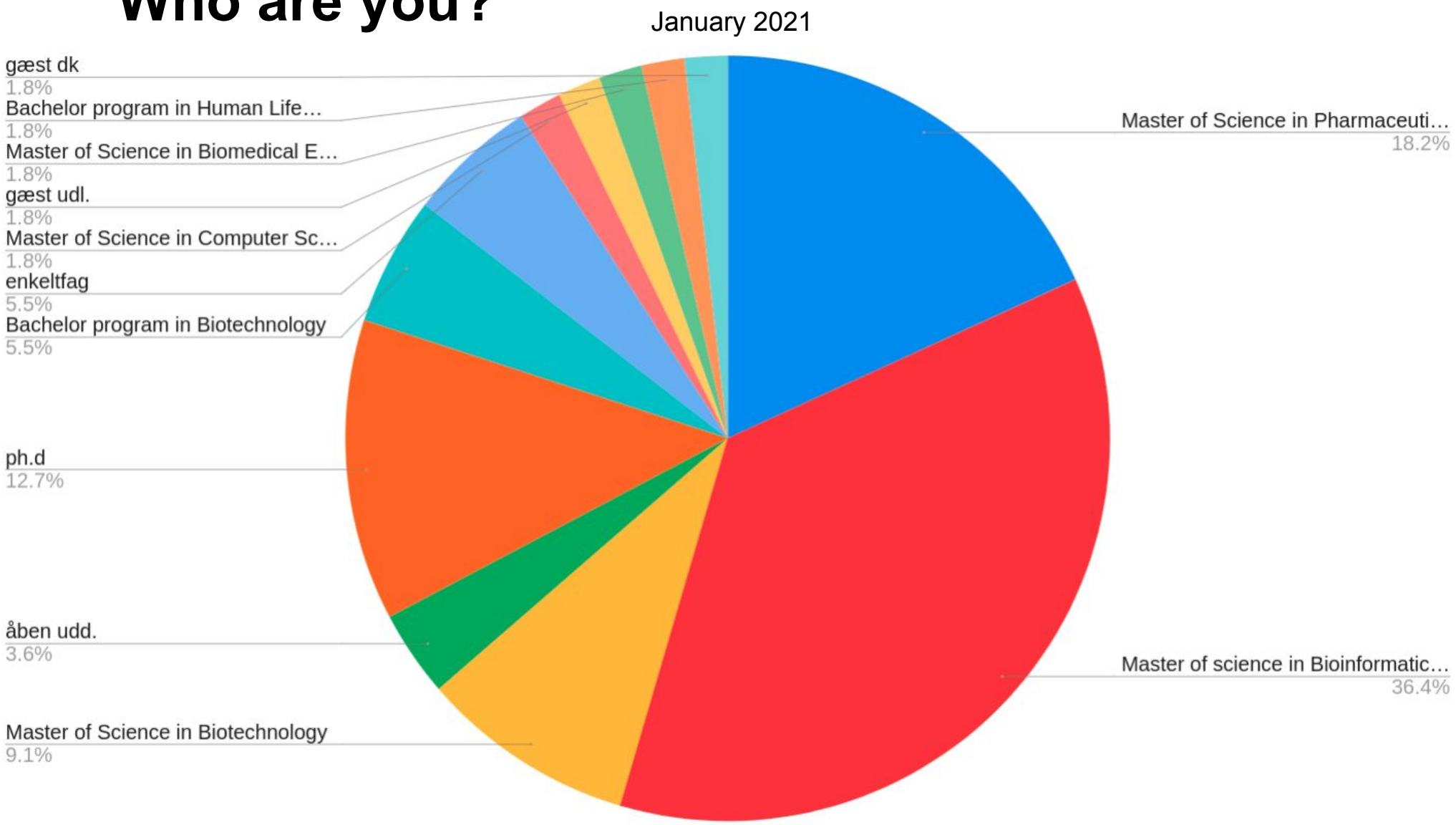
Online class this year

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 2021
gæst dk
1.8%
Bachelor program in Human Life…
1.8%
Master of Science in Biomedical E…
1.8%
gæst udl.
1.8%
Master of Science in Computer Sc…
1.8%
enkeltfag
5.5%
Bachelor program in Biotechnology
5.5%
ph.d
12.7%
åben udd.
3.6%
Master of Science in Biotechnology
9.1%
Master of Science in Pharmaceuti…
18.2%
Master of science in Bioinformatic…
36.4%

Feedback

- 9th time we are running the course
- My second time!
- My first online class
- 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?

nature medicine

ARTICLES

https://doi.org/10.1038/s41591-019-0711-0

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 Botello[10], Javiera Uribe Fenner[10], Maria Cárcel Pérez[11], Maria Arranz[11], Andrea B. Moffitt[12], Zihua Wang[12], Amaia Hervás[13], 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

Tips

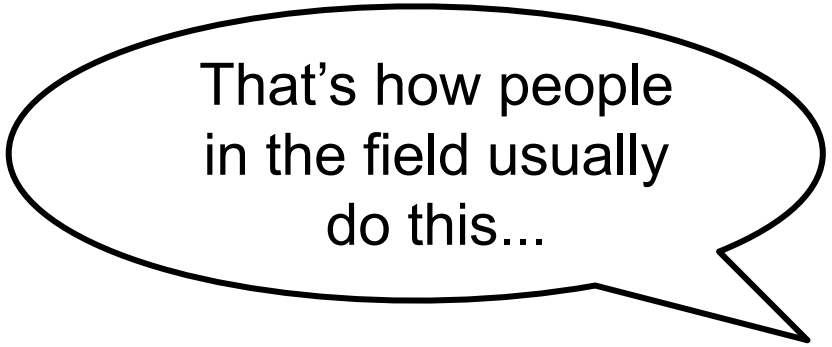
Tip: Do not memorize the name of the tools/procedure, they come and go

Tips

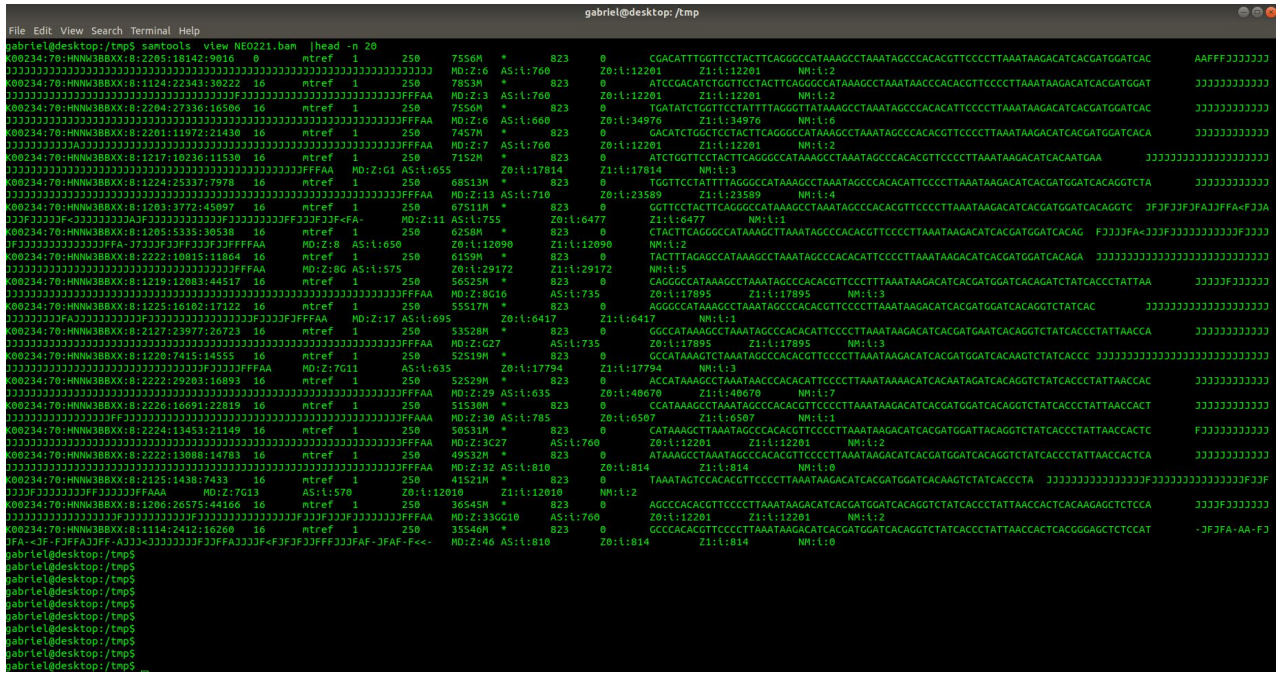
Tip: Understand the problem and how various tools work

Tips

- New tools or procedures get released all the time
- The best tool/format/pipeline in 2020 may not be the best in 2030
- Understand how they work, in which cases they perform well
- Read benchmarking papers and reviews
- Beware of:



The shell terminal



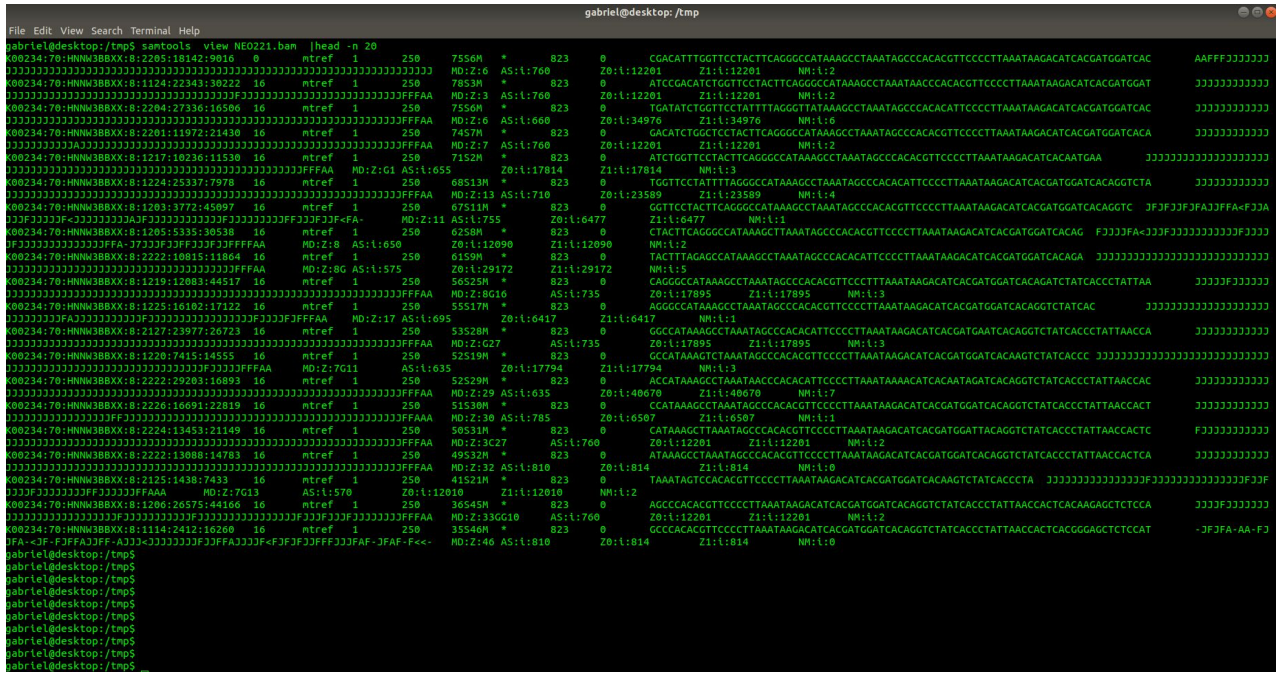
- Terminal allows users to interact with the computer using commands in the format:

```
command argument_1 argument_2
```

- Examples:

```
ls -al
pwd
```

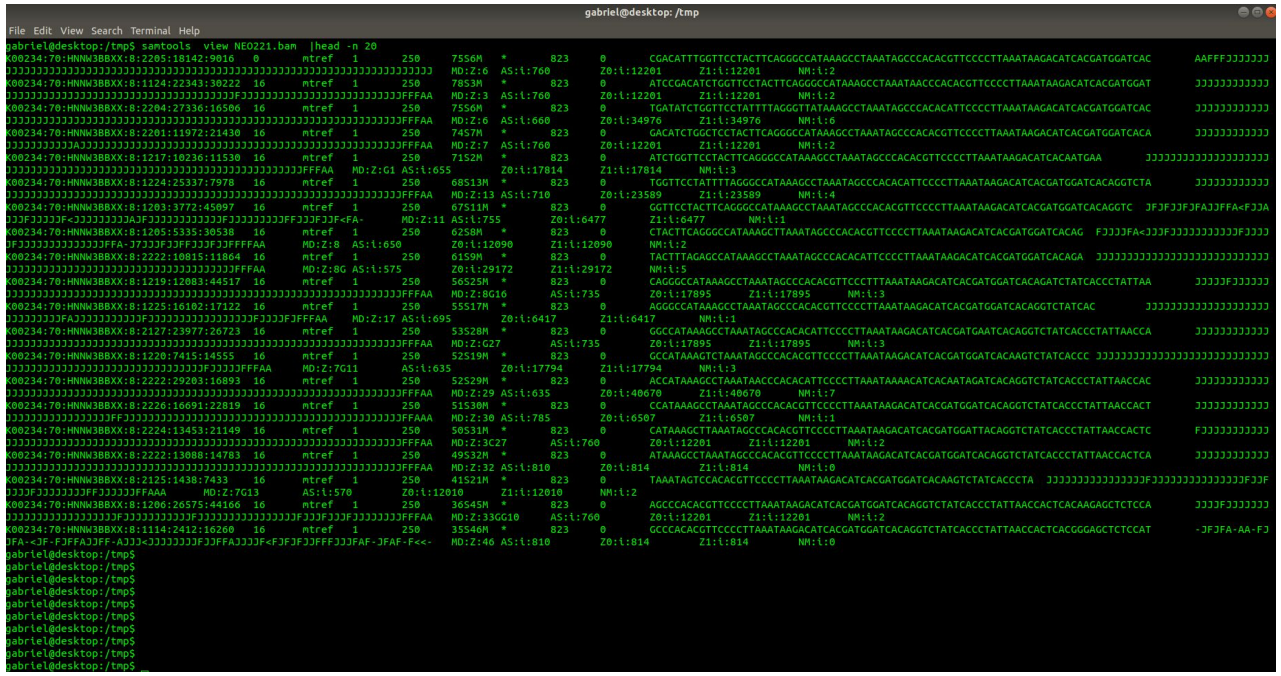
The shell terminal



- Available on various platforms

Mac™ OS

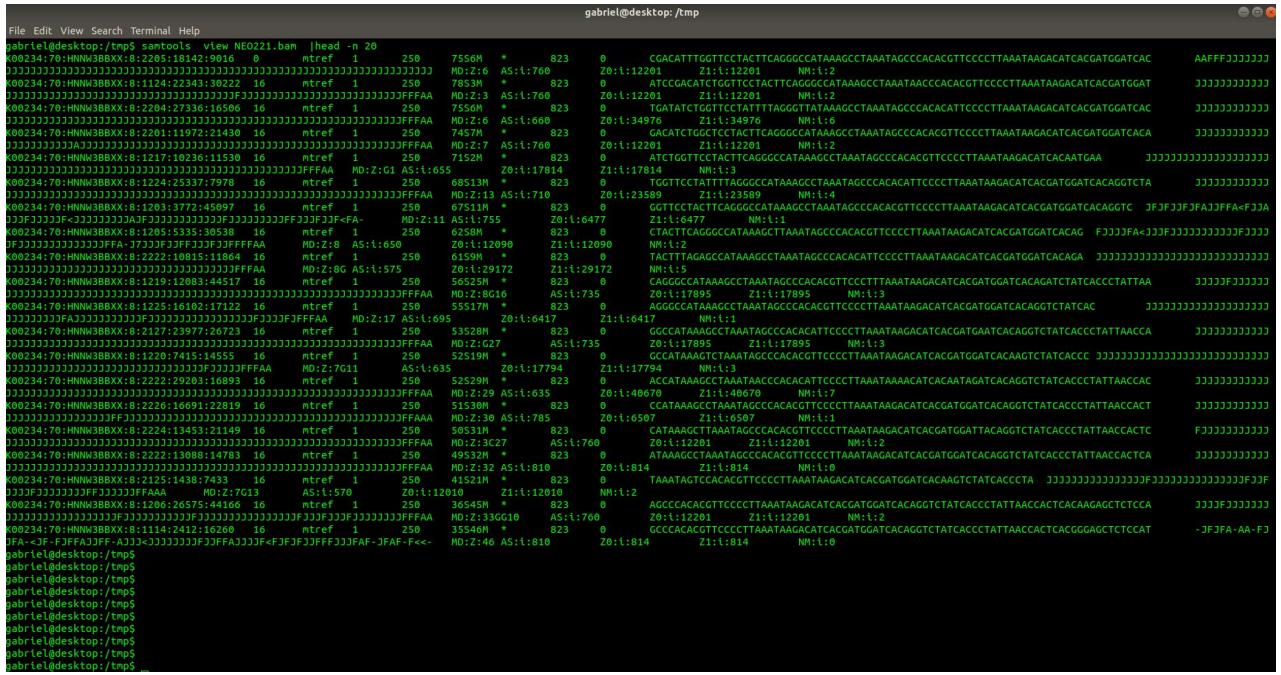
Why the shell terminal



- Almost every tool for NGS analysis are command line only
- Generally more efficient/flexible, you can play around with the tools/data:
  - ex: put all text files with a specific string in a zipped archive

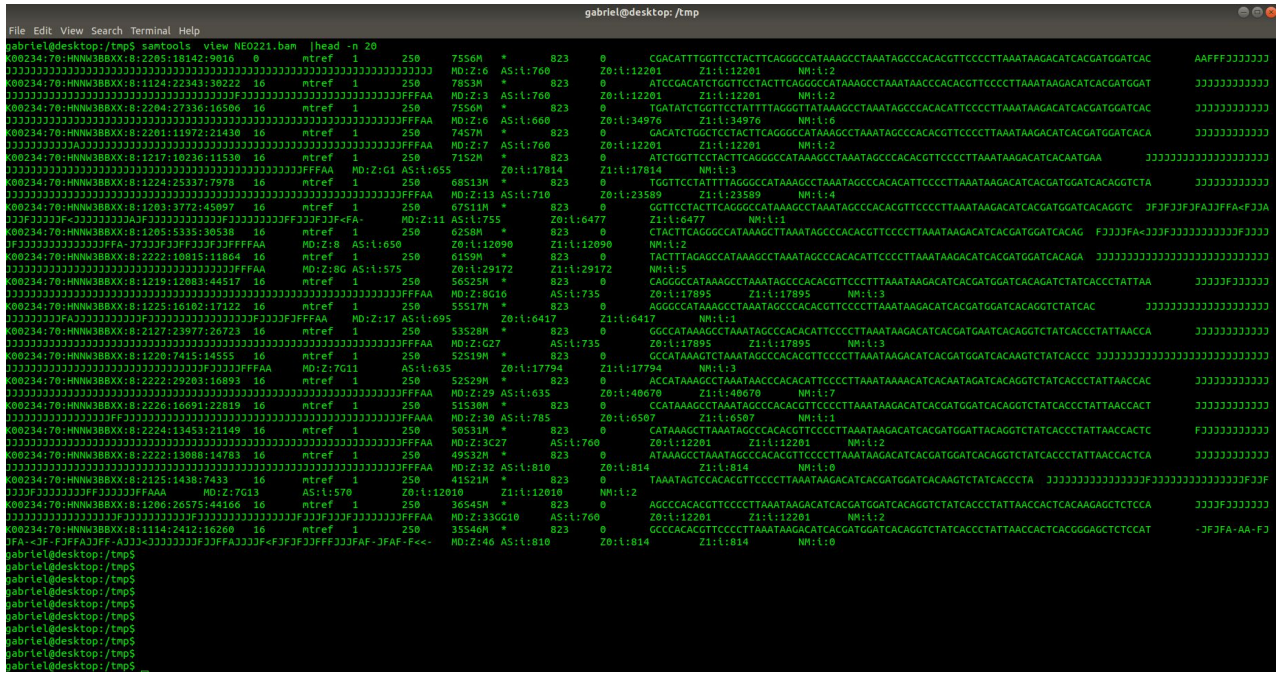
    a complete pain in a point-and-click windows environment, a breeze for the terminal

Why the shell terminal



- They can be pipelined, i.e. analyzing 100 files in windowed mode is a pain ...
- Alternative approaches: Galaxy, CLC-workbench, Geneious

Why learn to use UNIX/Linux? (in general)



- Contains several little programs (sed, cut, grep, paste) that can be combined to make really powerful queries
- File descriptors and pipe can be used to spare you a lot of time/disk space
- Make/Snakemake can automate workflows
- Open source tools
- You can basically finish a PhD in computational bio. without knowing how to code

Course structure

- 3 weeks, 2 tracks

Lectures + Exercises + Pres.

Project work

Date: 4th 13th 22th

= Poster exam

Course breakdown I

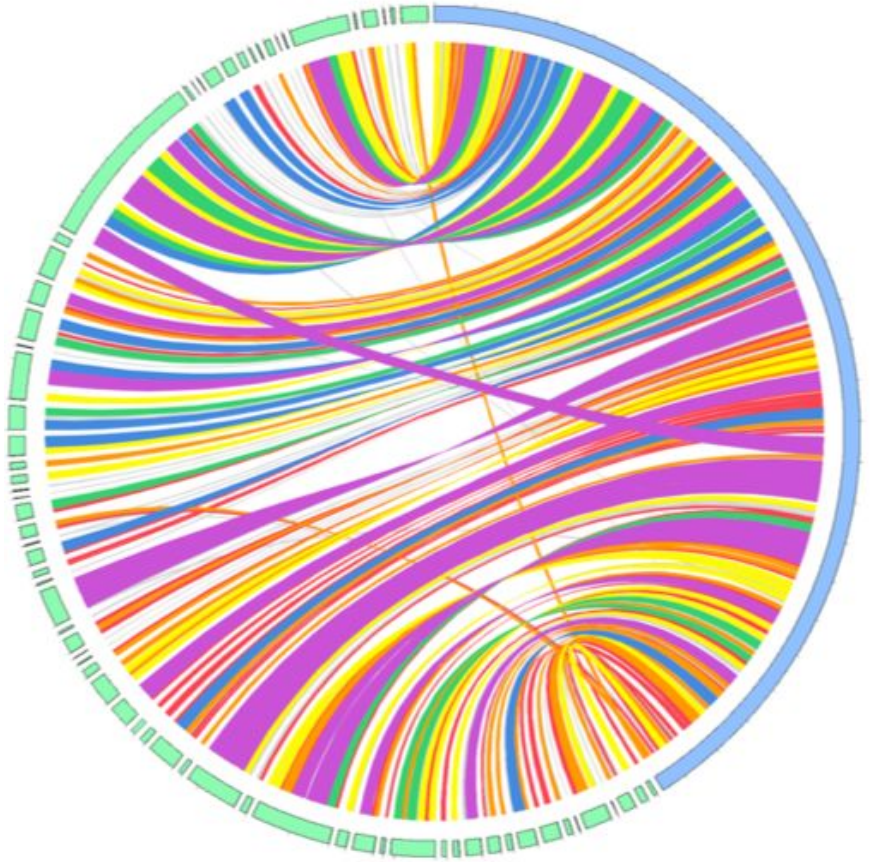
- Monday 4th January
  - Introduction NGS technology
  - Tech talk groups
  - Unix and first look at data

- Tuesday 5th January
  - Data basics & preprocessing
  - Alignment

Course breakdown II

- Wednesday 6th January
  - Functional Human Variation
  - Alignment processing
  - *de novo* assembly

- Thursday 7th January
  - *de novo* metagenomics
  - Quantitative metagenomics



Course breakdown III

- Friday 8th January
  - cell free DNA
  - Recap test (after lunch)

- Monday 11th January
  - RNAseq
  - Cancer-seq

- Tuesday 12th January
  - Genomic Epidemiology
  - Tech talk work & Presentations

Course breakdown IV

- *Wednesday 13th* January
  - Ancient DNA
  - Project work
  - Prepare presentations for tomorrow
- Thursday, 14th January
  - Short project presentations
  - Project work
- *Friday 15th - Thursday 21th*
  - Project work
- *Friday 22th*
  - *Poster Exam* (online)

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
- Don’t analyze very large datasets (time, resources)

Points to remember

- Learn 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

- The course has moved to the Cloud!
- Danish National Supercomputer for Life Science (Computerome) located at DTU Risø
- 16048 cores, 92 Tb RAM an 3Pb storage
- We have 2 nodes
  - Each has 28 cores and 128 Gb RAM

Exam

- Each group will create a poster
- Due to COVID19, please produce a large, high resolution PDF
- Each group will present online the poster for the examiners
- Then each individual in the group will one-by-one be asked questions on the learning objectives and your project (5-10 min).
  - Do not memorize, **understand** what you are doing during the project
  - Understand the concepts taught in class

Marking scheme

project+poster+presentations

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 text-book for the course
- There are papers uploaded to DTU inside that you can read for more information
- 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
- This year, please copy-paste the questions in a text file and fill it and send it to me via email (see addresses in previous slides)