

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 for Evolutionary Anthropology 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
- Pangenomes

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

Who are we?

- Organizer:
 - Gabriel Renaud
 - Kristoffer Vitting-Seerup
 - Asker Brejnrod
 - Josh Rubin
 - Nicola Vogel
 - Louis Kraft
 - DTU Bioinformatics
 - Peter Wad Sackett
- DTU Food
 - Pimlapas Leekitechaoenphon (Shinny)
- Copenhagen University:
 - Martin Sikora
 - Shilpa Garg

Main teaching assistants

Josh Rubin <jdru@dtu.dk>

Nicola Vogel <navo@food.dtu.dk>

Louis Kraft <lokraf@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

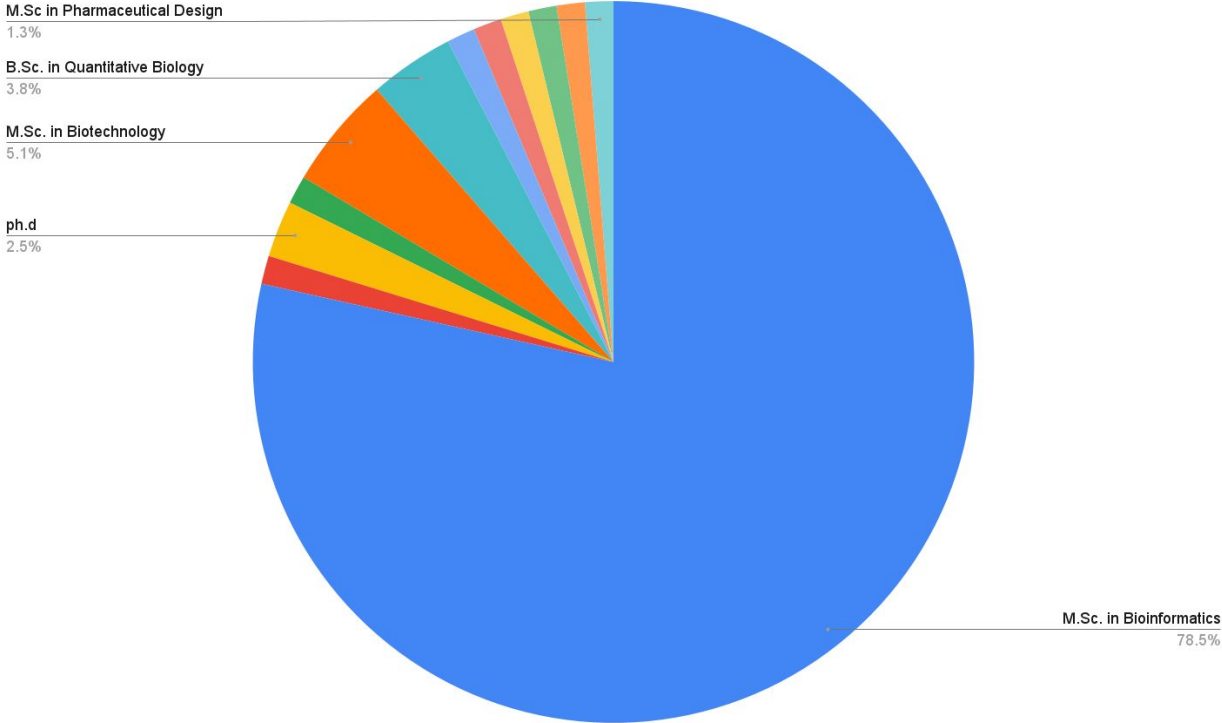
If my internet connection drops, please stay! I will come back

Schedule, exercises, general plan:

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

Who are you?

January 2023



Feedback

- My 4th time! 2nd time in person.
- 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?

Article | [Open Access](#) | [Published: 09 February 2022](#)

Signatures of TOP1 transcription-associated mutagenesis in cancer and germline

[Martin A. M. Reijns](#) ✉, [David A. Parry](#), [Thomas C. Williams](#), [Ferran Nadeu](#), [Rebecca L. Hindshaw](#), [Diana O. Rios Szwed](#), [Michael D. Nicholson](#), [Paula Carroll](#), [Shelagh Boyle](#), [Romina Royo](#), [Alex J. Cornish](#), [Hang Xiang](#), [Kate Ridout](#), [The Genomics England Research Consortium](#), [Colorectal Cancer Domain UK 100,000 Genomes Project](#), [Anna Schuh](#), [Konrad Aden](#), [Claire Palles](#), [Elias Campo](#), [Tatjana Stankovic](#), [Martin S. Taylor](#) ✉ & [Andrew P. Jackson](#) ✉

[Nature](#) **602**, 623–631 (2022) | [Cite this article](#)

21k Accesses | **6** Citations | **192** Altmetric | [Metrics](#)

Findings:

A deletion found in cancer and elsewhere is due to a specific protein TOP1

Published: 03 May 2022

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 tools⁶⁴. 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.

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

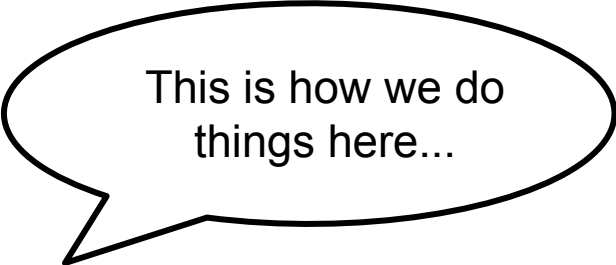


Tips for NGS in general

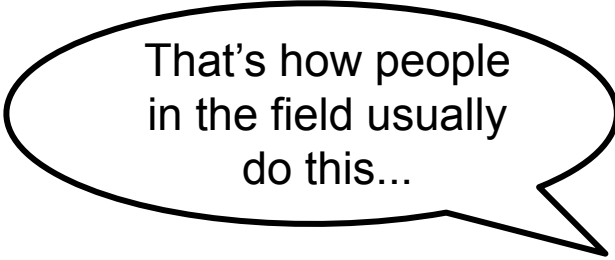
- New tools or procedures get released all the time
- The best tool/format/pipeline in 2023 may not be the best in 2033
- Understand how they work, in which cases they perform well

Tips for NGS in general

- Read benchmarking papers and reviews
- Beware of:

A black-outlined speech bubble with a tail pointing towards the bottom-left.

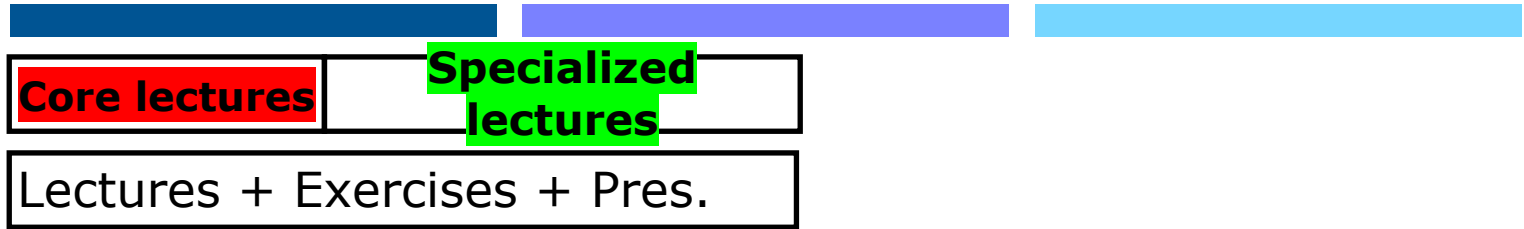
This is how we do things here...

A black-outlined speech bubble with a tail pointing towards the bottom-right.

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

Course structure

- 3 weeks, 2 tracks



Date: 2n
d

11th

20th



Course breakdown I

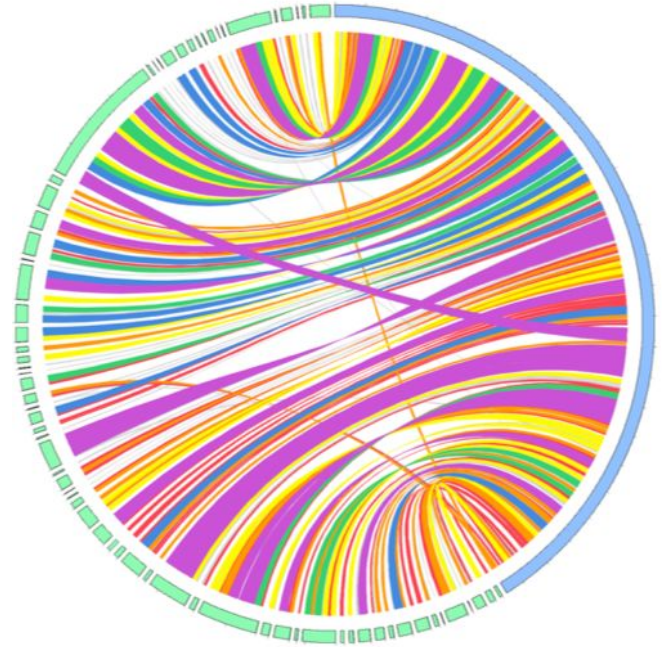
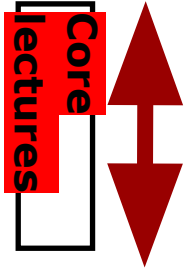
- Monday 2nd January
 - Introduction NGS technology
 - Tech talk groups
 - Unix and first look at data
- Tuesday 3rd January
 - Data basics & preprocessing
 - Alignment



Core
lectures

Course breakdown II

- Wednesday 4th January
 - Functional Human Variation
 - Alignment processing
 - *de novo* assembly
- Thursday 5th January
 - *de novo* metagenomics
 - Quantitative metagenomics



Course breakdown III

- Friday 6th January
 - Long read technologies
 - Recap test (after lunch)
- Monday 9th January
 - RNAseq
 - Ancient DNA
- Tuesday 10th January
 - Genomic Epidemiology
 - Tech talk work & Presentations

Course breakdown IV

- Wednesday 11th January
 - Cancer-seq
 - Project work
 - Prepare presentations for tomorrow
- Thursday, 12th January
 - Short project presentations
 - Project work
- Friday 13th - Thursday 19th
 - Project work
- Friday 20th
 - 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)

Piazza

- Teachers and TAs will be available to help with your projects
- Office hours during project period: 10-14
- Use Piazza as a platform to communicate with your peers, TAs and teachers
 - Collective knowledge
 - Access @ piazza.com/danish_technical_university/summer2019/1

The logo for Piazza, featuring the word "piazza" in a blue, lowercase, sans-serif font. The letter 'p' is significantly larger than the other letters, and the 'a's have a distinctive shape with a small gap at the top.

Piazza

- Teachers and TAs will be available to help with your projects
- Office hours during project period: 1pm-3pm
- Use Piazza as a platform to communicate with your peers, TAs and teachers
 - Collective knowledge
 - Access @
piazza.com/danish_technical_university/winter2021/2212622176/home
 - Access code: 2212622176

The logo for Piazza, featuring the word "piazza" in a blue, lowercase, sans-serif font. The letter 'p' is significantly larger than the other letters, and the 'a's have a distinctive shape with a small gap at the top.

Points to remember

- **Understand** principles of the analysis
- The exercises will be useful for your projects and hopefully also later
- You don't need to do all the exercises but the ones from the core lectures are important
- 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

- Pupil cluster
- We have 3 nodes
 - pupil1 40 cores 252G RAM
 - pupil2 24 cores 110G RAM
 - pupil 24 cores 94G RAM
- Be careful with disk space
- Limited computational power
- If you want software installed, ask me!



Exam

- Each group will create a poster

~~• You can print posters at the BFC for 20-30 kr~~

Online this year: send us a high resolution PDF!

- Each group will present 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).

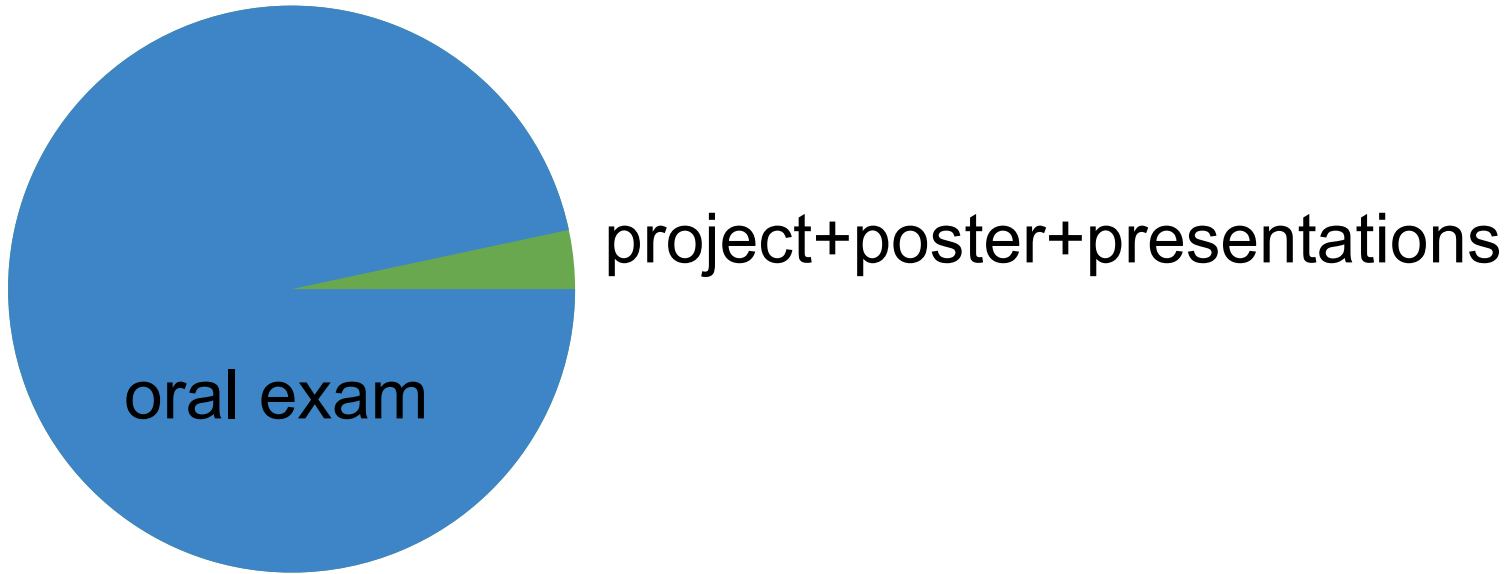
Exam

- 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 (esp. the first 3 days).
- Understand what you are doing:
 - inspect the input
 - inspect the output
 - play with parameters

Marking scheme



Disclaimer

- Sequencing technologies change 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
- We will talk about old techs, working with NGS means working with older datasets from previous studies

Be adventurous!

You do not have the ability to do anything
destructive

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!

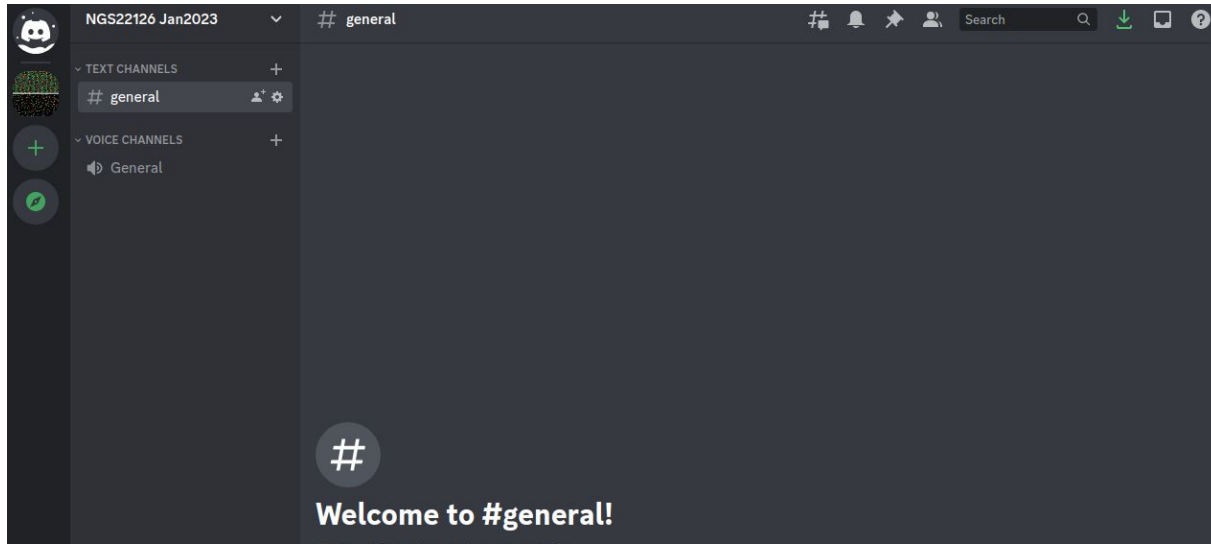
Discord

- Chat with others during off-hours. Create channels!
- Please use your real name:

Jan Jansen



n00b_0wner_18



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

Pre-test

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