

**DTU**





**DTU Health Technology  
Bioinformatics**

**22126: Next Generation Sequencing Analysis  
DTU - January 2025  
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?

How to contact me?

- slow response: gabre [at] dtu [dot] dk
- medium response: gabriel [dot] reno [at] gmail [dot] com
- fastest response: Discord

# 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
  - Amanda Gammelby Qvesel
  - Mads Hartmann
  - Kristoffer Vitting-Seerup
  - Frederikke Pedersen
  - Peter Wad Sackett
- DTU Food
  - Pimlapas Leekitecharoenphon (Shinny)
- Copenhagen University:
  - Martin Sikora

# Main teaching assistants

- Amanda Gammelby Qvesel
- Mads Hartmann

# Disclaimers

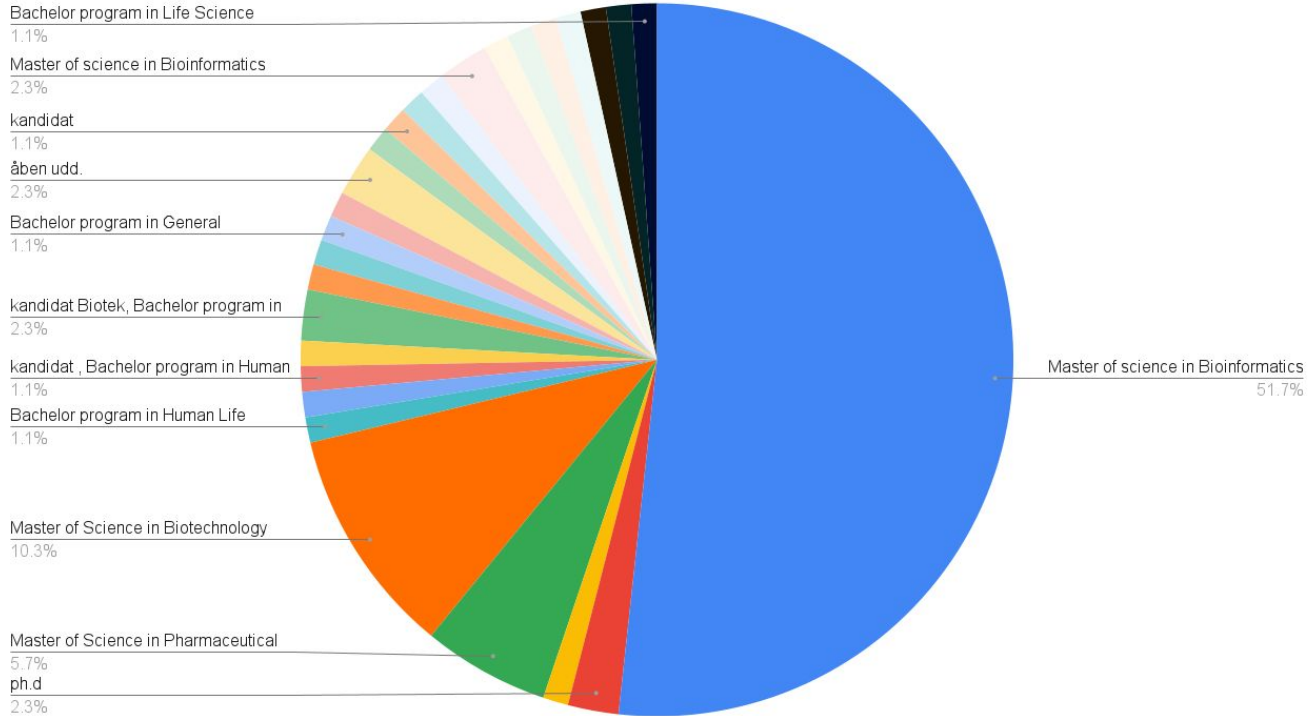
- Conflict of interest: none to declare, I do not own any stocks or consulting for any sequencing company
- I (over)used images generated by Dall-E+Midjourney and images from Wikipedia



# Who are you?

January 2025

## Background



# Feedback

- My 6th time! 4rd 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?

[nature](#) > [articles](#) > [article](#)

Article | [Open access](#) | Published: 08 April 2024

## Tumour-selective activity of RAS-GTP inhibition in pancreatic cancer

[Urszula N. Wasko](#), [Jingjing Jiang](#), [Tanner C. Dalton](#), [Alvaro Curiel-Garcia](#), [A. Cole Edwards](#), [Yingyun Wang](#), [Bianca Lee](#), [Margo Orlen](#), [Sha Tian](#), [Clint A. Stalnecker](#), [Kristina Drizyte-Miller](#), [Marie Menard](#), [Julien Dilly](#), [Stephen A. Sastra](#), [Carmine F. Palermo](#), [Marie C. Hasselluhn](#), [Amanda R. Decker-Farrell](#), [Stephanie Chang](#), [Lingyan Jiang](#), [Xing Wei](#), [Yu C. Yang](#), [Ciara Helland](#), [Haley Courtney](#), [Yevgeniy Gindin](#), [Karl Muonio](#), [Ruiping Zhao](#), [Samantha B. Kemp](#), [Cynthia Clendenin](#), [Rina Sor](#), [William P. Vostrejs](#), [Priya S. Hibshman](#), [Amber M. Amparo](#), [Connor Hennessey](#), [Matthew G. Rees](#), [Melissa M. Ronan](#), [Jennifer A. Roth](#), [Jens Brodbeck](#), [Lorenzo Tomassoni](#), [Basil Bakir](#), [Nicholas D. Socci](#), [Laura E. Herring](#), [Natalie K. Barker](#), [Junning Wang](#), [James M. Cleary](#), [Brian M. Wolpin](#), [John A. Chabot](#), [Michael D. Kluger](#), [Gulam A. Manji](#), [Kenneth Y. Tsai](#), [Miroslav Sekulic](#), [Stephen M. Lagana](#), [Andrea Califano](#), [Elsa Quintana](#), [Zhengping Wang](#), [Jacqueline A. M. Smith](#), [Matthew Holderfield](#), [David Wildes](#), [Scott W. Lowe](#), [Michael A. Badgley](#), [Andrew J. Aguirre](#), [Robert H. Vonderheide](#), [Ben Z. Stanger](#), [Timour Baslan](#), [Channing J. Der](#), [Mallika Singh](#) ✉ & [Kenneth P. Olive](#) ✉

— Show fewer authors

[Nature](#) **629**, 927–936 (2024) | [Cite this article](#)

### Findings:

- A drug RMC-7977 effectively targets a gene causing pancreatic cancer, causing tumor cell death and halting growth, while sparing normal tissues from significant harm.
- In mouse models, the drug greatly extended survival, though some tumors developed resistance linked to increased Myc gene levels.
- RMC-7977 is a potential treatment strategy for pancreatic cancer.

Published: April 2024

## 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.

# 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

**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

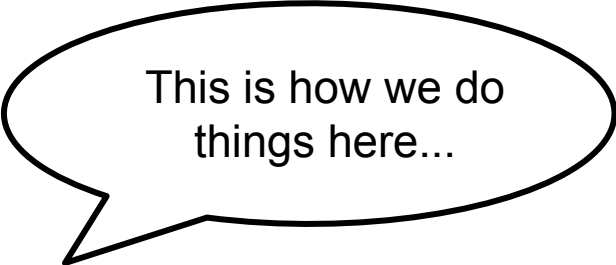


## **Tips for NGS in general**

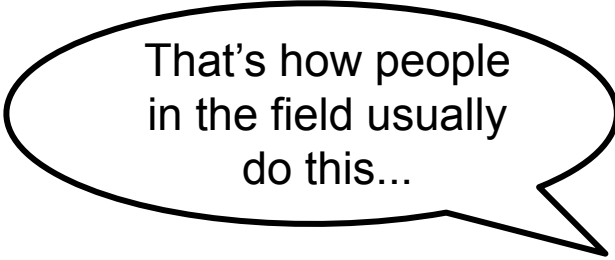
- New tools or procedures get released all the time
- The best tool/format/pipeline now may not be the best in 2034
- Understand how they work, in which cases they perform well

## 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...

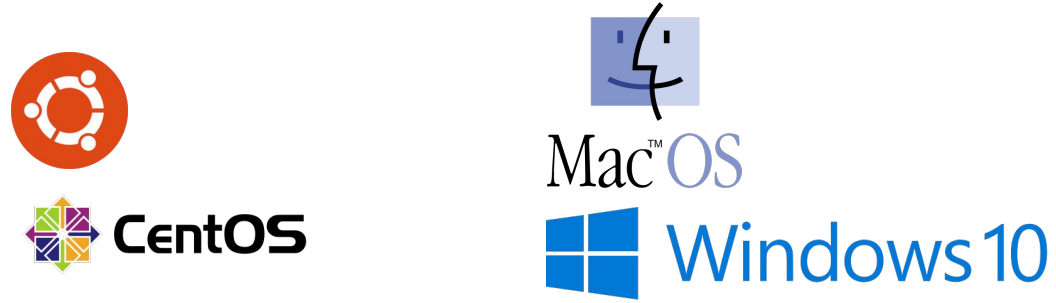


# The shell terminal

```
gabriele@desktop: ~/mp
gabriele@desktop:/mp$ sudo nmap -v 10.0.2.15
Nmap scan report for 10.0.2.15
Host: 10.0.2.15
OS: Linux 3.10

```

- Available on various platforms



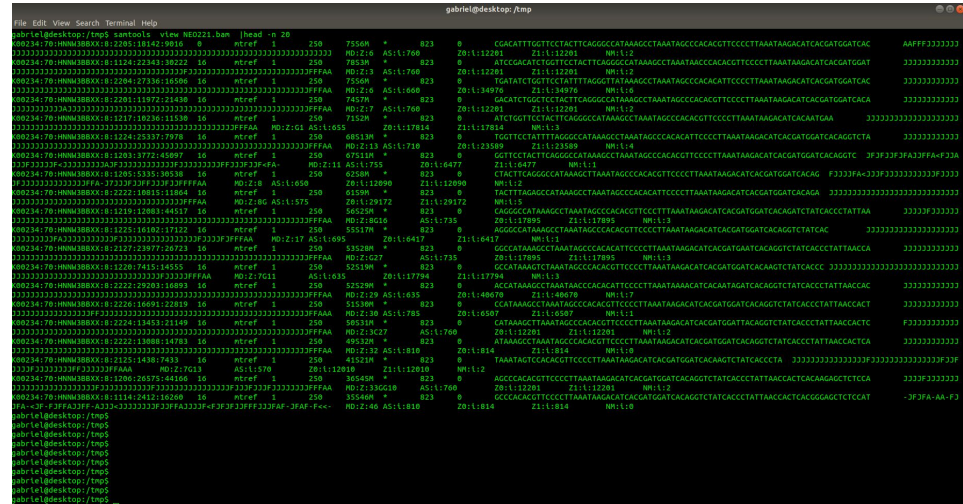








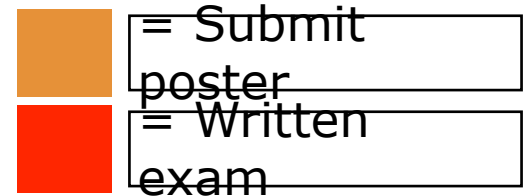
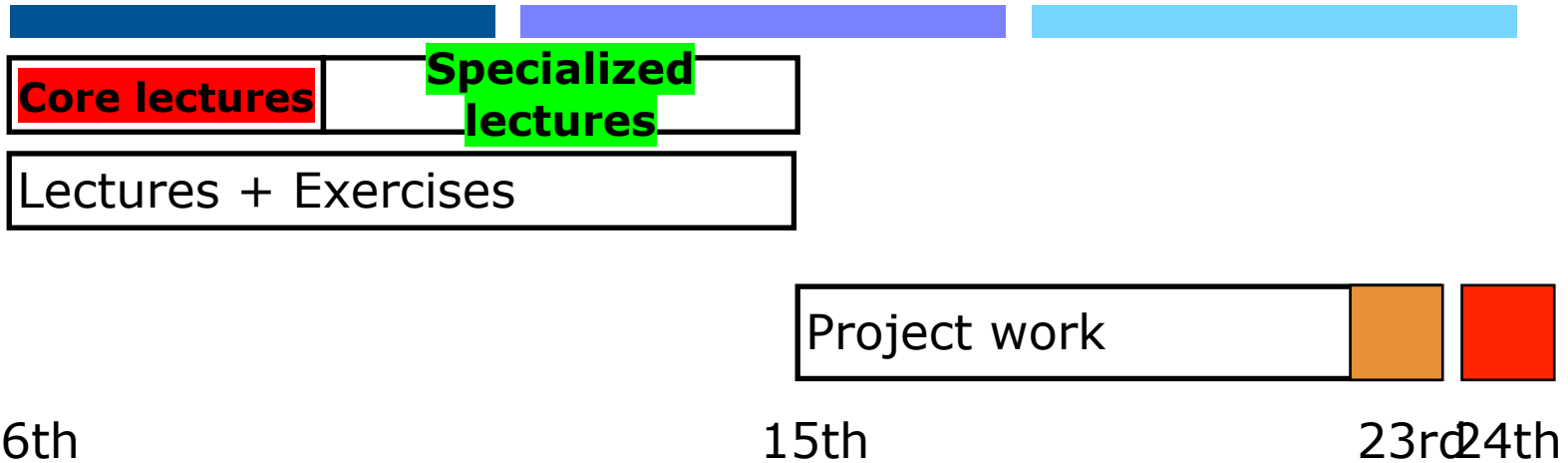
# Why learn to use UNIX/Linux? (in general)



- 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/Nextflow 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



# Course breakdown I

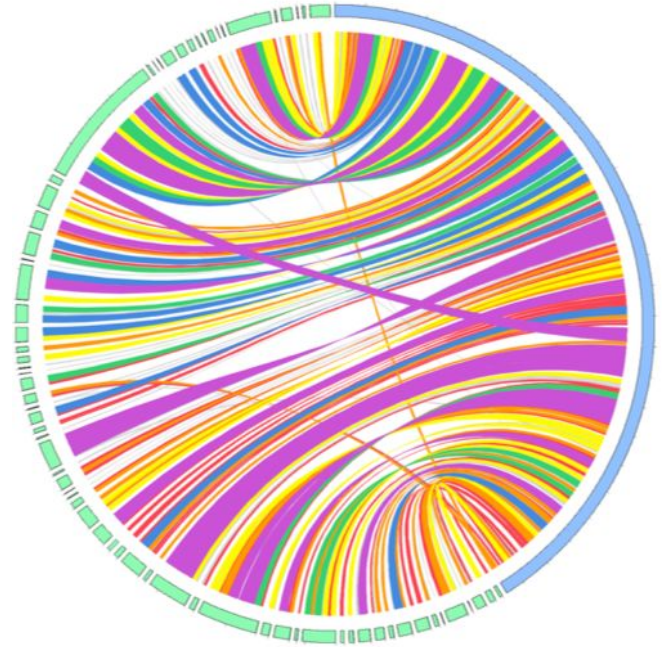
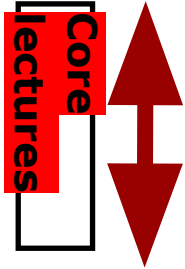
- Day 1
  - Introduction NGS technology
  - Unix and first look at data
- Day 2
  - Data basics & preprocessing
  - Alignment
  - Reminder about Bayesian



Core  
lectures

## Course breakdown II

- Day 3
  - Variant calling
- Day 4
  - *de novo* assembly
  - Long read sequencing



## Course breakdown III

- Friday 10th January
  - Ancient DNA
  - Recap test (after lunch)
  - RNAseq
  - Catch up
- Monday 13th January
  - Metagenomics
- Tuesday 14th January
  - Genomic Epidemiology
  - Group formation project

# Course breakdown IV

- Wednesday 15th January
  - Cancer-seq
  - Project work
- Thursday, 12th - Thursday 22nd
  - Project work
- Thursday 23th
  - Submit 1 page poster
- Friday 24th
  - Written Exam



# Projects

- Try to analyze an empirical dataset and present results on poster
- 4-6 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)
- 2024: You have 2 days, a weekend and a week to finish



download+align data here!!

# 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 !



# Points to remember

- You get the solutions for the exercises but **do not copy-paste!!**
- You will not get to copy-paste for the project

# Cloud computing

- Pupil cluster
- We have 5 nodes
  - pupil1 40 cores 252G RAM
  - pupil2 24 cores 110G RAM
  - pupil3 24 cores 94G RAM
  - pupil4 48 cores 126G RAM
  - pupil5 48 cores 126G RAM
- Be careful with disk space
- Limited computational power
- If you want software installed, ask me!



# Poster

- Each group will create a poster
- The goal of the project is:
  - Do not memorize, **understand** what you are doing during the project
  - Understand the concepts taught in class
  - Learn NGS from firsthand experience
- Please send the PDF before noon on Thursday the 23rd

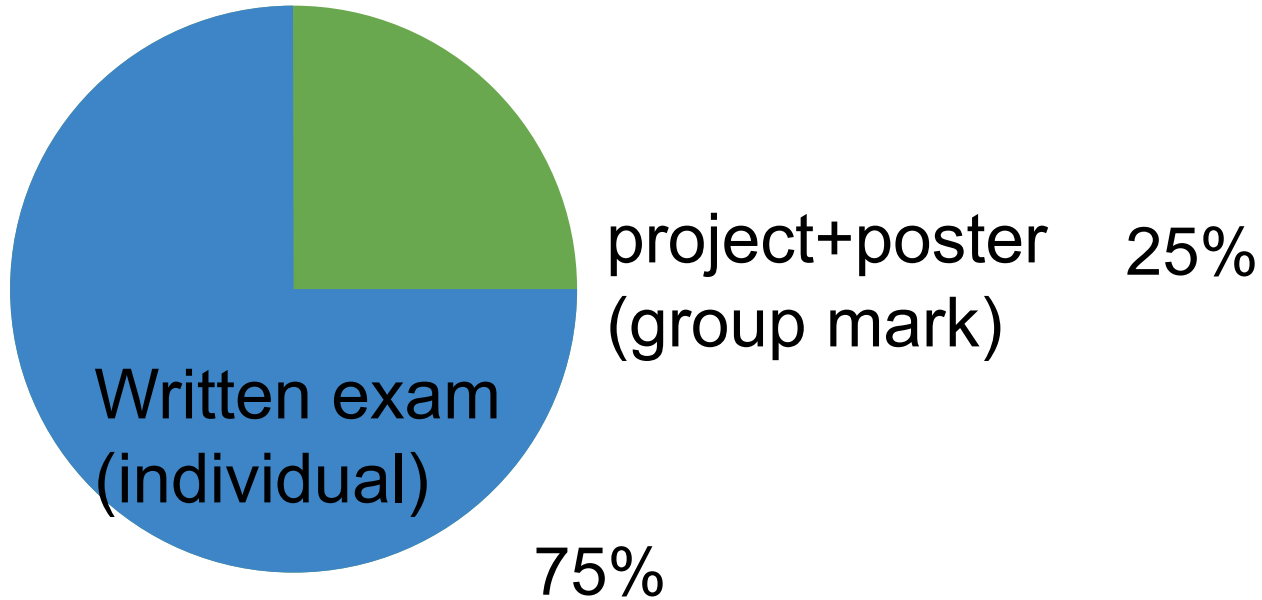
# Written Exam

- **You cannot write the exam if you have not submitted the poster**
- Multiple choice exam
- Focuses on the core lectures
- Will have 1 very basic question per specialized lecture

## 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 4 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

# Disclaimer

- 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!

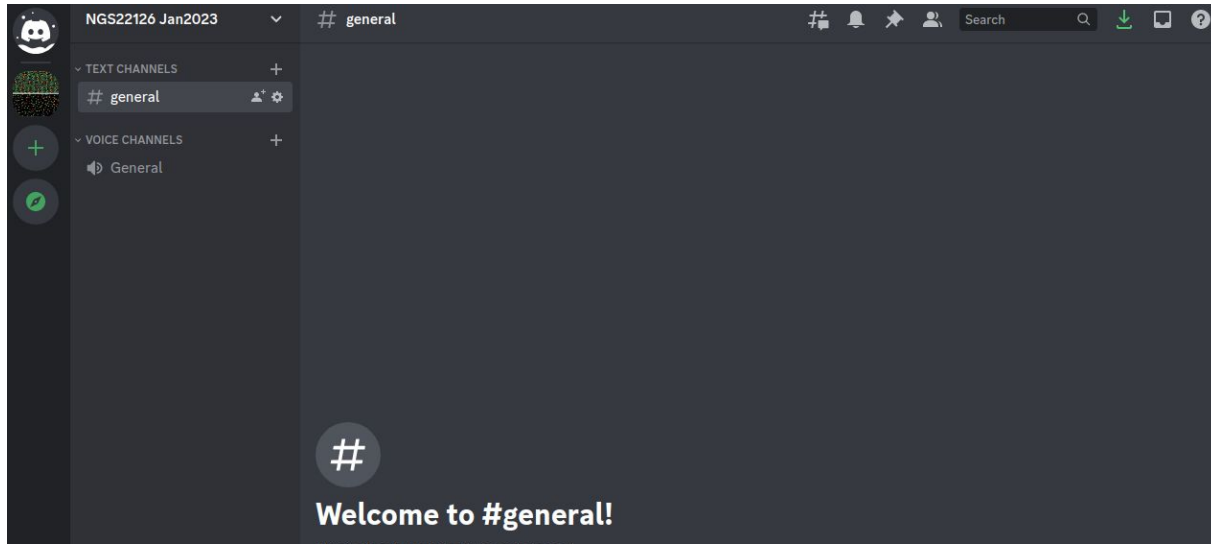
# Discord

- 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