



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: ٠
  - Gabriel Renaud
  - Shyam Gopalakrishnan
    Aarhus University
  - Gisle Vestergaard
  - Josh Rubin
  - Nicola Vogel
  - DTU Bioinformatics
  - Peter Wad Sackett
- DTU Aqua
  - Francesca Bertolini
- DTU Food
  - Pimlapas Leekitecharoenphon (Shinny)

- Copenhagen University:
  - Martin Sikora
- - 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





medicine



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

Martin W. Breuss<sup>1,2</sup>, Danny Antaki<sup>3,4,5,6</sup>, Renee D. George<sup>1,2</sup>, Morgan Kleiber<sup>3,4,5</sup>, Kiely N. James<sup>1,2</sup>, Laurel L. Ball<sup>1,2</sup>, Oanh Hong<sup>3,4,5,6</sup>, Ileena Mitra<sup>7,8</sup>, Xiaoxu Yang<sup>1,2</sup>, Sara A. Wirth<sup>1,2</sup>, Jing Gu<sup>1,2</sup>, Camila A. B. Garcia<sup>1,2</sup>, Madhusudan Gujral<sup>3,4,5,6</sup>, William M. Brandler<sup>3,4,5,6</sup>, Damir Musaev<sup>1,2</sup>, An Nguyen<sup>1,2</sup>, Jennifer McEvoy-Venneri<sup>1,2</sup>, Renatta Knox<sup>1,2,9</sup>, Evan Sticca<sup>1,2</sup>, Martha Cristina Cancino Botello<sup>10</sup>, Javiera Uribe Fenner<sup>10</sup>, Maria Cárcel Pérez<sup>11</sup>, Maria Arranz<sup>11</sup>, Andrea B. Moffitt<sup>12</sup>, Zihua Wang<sup>12</sup>, Amaia Hervás<sup>13</sup>, Orrin Devinsky<sup>10,14</sup>, Melissa Gymrek<sup>7,8</sup>, Jonathan Sebat<sup>0,3,4,5,6\*</sup> and Joseph G. Gleeson<sup>0,1,2\*</sup>

Denovo 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



WES and WGS trio analysis. WGS sequencing and analysis for F01–08 and F13-20 were performed as described previously<sup>13,37</sup>. 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)<sup>38</sup>. 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



"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<sup>™</sup> 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



"... 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."





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:





#### The shell terminal



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

```
command argument_1 argument_2
```

• Examples:



#### The shell terminal

	gabriel@desktop: /tmp	00
w Search Terminal Help		
ktop:/tmpS samtools view NE0221.bam  head -n 20		
NNW3BBXX:8:2205:18142:9016 0 mtref 1 250 7556M * 8	823 0	AAFFFJJJJJJJJ
NNW3BBXX:8:1124:22343:30222 16 mtref 1 250 7853M * 8	823 0 ATCCGACATCTGGTTCCTACTTCAGGGCCCATAAAGCCTAAATAACCCACACGTTCCCCTTAAATAAGACATCACGATGGAT	222222222222
33333333333333333333333333333333333333	Z0:1:12201 Z1:1:12201 NM:1:2	
NNW3BBXX:8:2204:27336:16506 16 mtref 1 250 7556M * 8	823 0 TGATATCTGGTTCCTATTTTAGGGTTATAAAGCCTAAATAGCCCACACATTCCCCTTAAATAAGACATCACGATGGATCAC	222222222222
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	Z0:1:34976 Z1:1:34976 NM:1:6	
NNW3BBXX:8:2201:11972:21430 16 mtref 1 250 7457M * 8	823 0 GACATETGGETECTACTAGGGECATAAAGGECTAAATAGECCACACGTTECEETTAAATAAGACATEACGATGGATCACA	
AJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ		
NNWSDDAA.6.1217.10230.11530 10 Milet 1 250 /1520 - 6	523 0 ATCTGGTTCCTACTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACAATGAA	
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ		
11111111111111111111111111111111111111	70:1:23589 71:1:23589 NM:1:4	5555555555555555
NNW3BBXX:8:1203:3772:45097 16 mtref 1 250 67511M * 8	823 Ø GGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGGATGGAT	FJFJJFJFAJJFFA <fjja< td=""></fjja<>
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	20:1:6477 Z1:1:6477 NM:1:1	
NNW3BBXX:8:1205:5335:30538 16 mtref 1 250 62S8M * 8	823 0 CTACTTCAGGGCCATAAAGCTTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAG FJJJJFA <j< td=""><td>JJFJJJJJJJJJJJJJJFJJJJJ</td></j<>	JJFJJJJJJJJJJJJJJFJJJJJ
JJJJJFFA-J7JJJFJJFFJJJFJJFFFFAA MD:Z:8 AS:1:650 Z0:1:12090 Z	Z1:1:12090 NM:1:2	
NNW3BBXX:8:2222:10815:11864 16 mtref 1 250 6159M * 8	823 0 TACTTTAGAGCCATAAAGCCTAAATAGCCCACACATTCCCCTTAAATAAGACATCACGATGGATCACAGA JJJJJJJJJ	
NNW3BBXX:8:1219:12083:44517 16 mtref 1 250 56525M * 8	823 0 CAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCTTTAAATAAGACATCACGATGGATCACAGATCTATCACCCTATTAA	00000F000000
33333333333333333333333333333333333333	AS:i:735 Z0:i:17895 Z1:i:17895 NM:i:3	
NNW3BBXX:8:1225:16102:17122 16 mtref 1 250 55517M * 8	823 0 AGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTATCAC	111111111111111111111111111111111111111
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	7 Z1:1:6417 NN:1:1	
NNW3BBXX:8:212/:239//:20/23 10 MTFet 1 250 53528M * 8	823 0 GOLCATAAAGCCTAAATAGCCCACACATTCCCCTTAAATAAGACATCACGATGAATCACAGGTCTATCACCCTATTAACCA	11111111111
NNW3DDAA.8812201/415114555 10 MLTET 1 250 52519M - 6	525 5 GEELATMAAGTETMAATAGEECAAACGTTEECETTMAATMAGACATEACGATGAATAACGACTTACACGE 55555555	
JJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFMM MD.2./1011 M3.(.035 20.(.1//5	74 21.0.1/774 NT.0.3	
11111111111111111111111111111111111111		
NNW3BBXX:8:2226:16691:22819 16 mtref 1 250 51530M * 8	823 Ø CCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGGTGGATCACAGGTCTATCACCCTATTAACCACT	111111
333333FF333333333333333333333333333333	Z0:1:6507 Z1:1:6507 NM:1:1	
NNW3BBXX:8:2224:13453:21149 16 mtref 1 250 50531M * 8	823 0 CATAAAGCTTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATTACA <u>GGTCTATCACCCTATTAACCACT</u>	F00000000000
33333333333333333333333333333333333333	AS:1:760 Z0:1:12201 Z1:1:2201 NM:1:2	
NNW3BBXX:8:2222:13088:14783 16 mtref 1 250 49532M * 8	823 0	222222222222
NNW3BBXX:8:2125:1438:7433 16 mtref 1 250 41521M * 8	823 0 TAAATAGTCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAAGTCTATCACCCTA JJJJJJJJJJJJJJJJJJJJJJJJJJJ	000000000000000000F00F
JJFFJJJJJJJFFAAA MD:Z:7G13 AS:i:570 Z0:i:12010 Z1:i:1201	10 NM:1:2	
NNW3BBXX:8:1206:26575:44166 16 mtref 1 250 36545M * 8	823 0 AGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTATCACCCCTATTAACCACTCACAAGAGGCTCTCCA	3333F3333333
JJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	AS:1:760 20:1:12201 21:1:12201 NM:1:2	100000000000000000000000000000000000000
NNW3BBXX:8:1114:2412:16260 10 MTref 1 250 35546M * 8	823 B GECCACACGITECECITAAATAAGACATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCAT	-JFJFA-AA-FJ
FAJJFF-AJJJ <jjjjjjjjjjfjjffajjjjf<fjfjjjfffjjjfaf-jfaf-< td=""><td>20:1:814 21:1:814 NM:1:0</td><td></td></jjjjjjjjjjfjjffajjjjf<fjfjjjfffjjjfaf-jfaf-<>	20:1:814 21:1:814 NM:1:0	
ktop://mps		
ktop:/tmps		
Kcop. / caps		

• Available on various platforms







#### Why the shell terminal

gabriel@desktop: /tmp	e e e
e Edit View Search Terminal Help	
priel@desktop:/tmp\$ samtools view NEO221.bam  head -n 20	
)234:70:HNNW3BBXX:8:2205:18142:9016 0 mtref 1 250 7556M * 823 0 CGACATTTGGTTCCTACT	TCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCAC AAFFFJJJJJJJ
Z1:i:12201 MD:Z:6 AS:i:760 Z0:i:12201 Z1:i:12201	NM:i:2
)234:70:HNNW3BBXX:8:1124:22343:30222 16 mtref 1 250 7853M * 823 0 ATCCGACATCTGGTTCCT/	ACTTCAGGGCCATAAAGCCTAAATAACCCACACGTTCCCCTTAAATAAGACATCACGATGGAT JJJJJJJJJJJJ
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	NM:1:2
2234:70:HNNW3BBXX:8:2204:27336:16506 16 mtref 1 250 7556M * 823 0 TGATATCTGGTTCCTATT	TTAGGGTTATAAAGCCTAAATAGCCCACACATTCCCCTTAAATAAGACATCACGATGGATCAC JJJJJJJJJJJJJ
13333333333333333333333333333333333333	NM:1:6
2234:70:HNNW3BBXX:8:2201:11972:21430 16 mtref 1 250 7457M * 823 0 GACATCIGGCICCIACTI	CAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACA JJJJJJJJJJJJ
HILLING 20:1:12201 21:12	NM:1:2
22347/05/HNNW3865X181/1/120236111530 10 PETET 1 250 /152M * 823 0 ATLIGUTULIALITAG	
C:::// MILLIULULULULULULULULULULULULULULULULULU	
7234/10/INNIW 305A.6.1224.2337/1776 10 MLET 1 230 00315M ~ 023 0 UUTICLEHTTHAUUUC	
1334-70-MNNJ3REYY-9-1283-3772-45907 16 mtraf 1 258 67511M * 823 A COTTOTATTCACCOC	NO. C.4
1 1 1 1 2 5 7 7 1 1 1 2 5 7 7 7 1 1 6 7 7 1 1 2 5 7 7 7 1 1 6 7 7 7 7 1 1 6 7 7 1 1 1 5 1 7 7 5 7 6 1 7 6 4 7 7 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	
134-70-HNNW388XX-8:1205:5335:30538 16 mtref 1 250 6258M * 823 0 CTACTTCAGGCCATAAA	CTTANATAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAG F111FA<111F1111111111111111111111111111
1)]]]]]]]]]]]]]]FA-]7]]]FJ]FF]]FFFFFAA MD:7:8 AS:1:650 Z0:1:1:2090 Z1:1:12090 NM:1:2	
0234:70:HNNW3BBXX:8:2222:10815:11864 16 mtref 1 250 6159M * 823 0 TACTITAGAGCCATAAAG	
DJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	
0234:70:HNNW3BBXX:8:1219:12083:44517 16 mtref 1 250 56525M * 823 0 CAGGGCCATAAAGCCTAA	ATAGCCCACACGTTCCCTTTAAATAAGACATCACGATGGATCACAGATCTATCACCCTATTAA JJJJJFJJJJJJ
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	:i:17895 NM:i:3
)234:70:HNNW3BBXX:8:1225:16102:17122 16 mtref 1 250 55517M * 823 0 AGGGCCATAAAGCCTAAA	TAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTATCAC JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
DJJJJJFAJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	
)234:70:HNNW3BBXX:8:2127:23977:26723 16 mtref 1 250 53S28M * 823 0 GGCCATAAAGCCTAAATA	GCCCACACATTCCCCTTAAATAAGACATCACGATGAATCACAGGTCTATCACCCTATTAACCA JJJJJJJJJJJJ
13333333333333333333333333333333333333	:1:17895 NM:1:3
0234:70:HNNW3BBXX:8:1220:7415:14555 16 mtref 1 250 52S19M * 823 0 GCCATAAAGTCTAAATAG	CCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAAGTCTATCACCC JJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	
2234:70:HNNW3BBXX:8:2222:29203:16893 16 mtref 1 250 52529M * 823 0 ACCATAAAGCCTAAATAA	CCCACACATTCCCCTTAAATAAAACATCACAATAGATCACAGGTCTATCACCCCTATTAACCAC JJJJJJJJJJJJJ
13JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	NM:1:7
1234;70:HNNW3BBXX:8:2226:16691:22819 16 MTret 1 250 51530M * 823 0 CCATAAAGCCTAAATAGC	CCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTATCACCCTATTAACCACT JJJJJJJJJJJJ
10501/12 10001/12 10001/12 10001/12 10001/12 10001/12 10000	NM:1:1: SACASSTTANATAASASASASASSASSASSASSASSASSASSASSASSAS
2234/70100003005X1812224139453121149 10 MTPET 1 250 50510 ° 825 0 CATAAAUCTTAAATAUCT	
1334-70-HNNU3BRYY-9-222-13899-14783 16 mt-ef 1 258 49537M * 823 8 ATAACCTTAAACCTTAAATACCC	
7234-10-INNIW 2007A-0-2222-13000-14703 10 PILET 1 230 47532F - 023 0 ALAMANCULARAMINUCUL	
234-70:HNNW388XX:8:2125:1438:7433 16 mtref 1 250 41521M * 823 0 TAAATAGTCCACACGTTC	CCCTTABATABGACATCACGATGGATCACAAGTCTATCACCCTA 111111111111111111111111111
DFJJJJJJJFFJJJJJJJFFAAA MD:7:7G13 AS:1:570 70:1:12010 71:1:12010 NM:1:2	
234:70:HNNW388XX:8:1206:26575:44166 16 mtref 1 250 36545M * 823 0 AGCCACACGTTCCCCTT	AAATAAGACATCACGATGGATCACAGGTCTATCACCCCTATTAACCACTCACAAGAGCTCTCCA JJJJEJJJJJ
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	:1:12201 NM:1:2
0234:70:HNNW3BBXX:8:1114:2412:16260 16 mtref 1 250 35546M * 823 0 GCCCACACGTTCCCCTTA	AATAAGACATCACGATGGATCACAGGTCTATCACCCCTATTAACCACTCACGGGAGCTCTCCAT - JFJFA-AA-FJ
A- <jf-fjffajjff-ajjj<jjjjjjjjjjjjjfjjffajjjjf<fjfjjjfaf-jfaf-< td=""><td>NM:i:0</td></jf-fjffajjff-ajjj<jjjjjjjjjjjjjfjjffajjjjf<fjfjjjfaf-jfaf-<>	NM:i:0
priel@desktop:/tmp\$	
priel@desktop:/tmp\$	
priel@desktop:/tmp\$	
oriel@desktop:/tmp\$	
oriel@desktop:/tmp\$	
oriel@desktop:/tmps	
CTALMAACVTAAL/TMAS	

- 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

			d	abriel@deskt	op: /tmp					e e e
le Edit View Search Terminal Help										
briel@desktop:/tmpS_samtools_view_NE0221_bam_	lbead on 20									
0234:70:HNNW3BBXX:8:2205:18142:9016 0	mtref 1	50 7556M	* 823		ACATTTGGTTCCT	CTTCAGGGCCATAA	AGCCTAAATAGCCCACACGTTCCC	CTTAAATAAGACATCACGA	IGGATCAC	AAFFF3333333
		MD:Z:6	AS:1:760	Z0:1:1220	Z1:1:12	01 NM:1:2				
0234:70:HNNW3BBXX:8:1124:22343:30222 16	mtref 1	250 78S3M	* 823		CCGACATCTGGTTG	CTACTTCAGGGCCA	TAAAGCCTAAATAACCCACACGTT	CCCCTTAAATAAGACATCAG	GATGGAT	222222222222
00000000000000000000000000000000000000	וכנכנכנכנכנכנכנכנכ	FFAA MD:Z:3	AS:1:760	Z0:1:1220:	Z1:1:122	01 NM:1:2				
0234:70:HNNW3BBXX:8:2204:27336:16506 16	mtref 1	250 75S6M			GATATCTGGTTCCT	TTTTAGGGTTATAA	AGCCTAAATAGCCCACACATTCCC	CTTAAATAAGACATCACGA	FGGATCAC	3333333333333
				Z0:1:34970	5 Z1:1:349					
0234:70:HNNW3BBXX:8:2201:11972:21430 16	mtref 1			0 G/	CATCTGGCTCCTA	TTCAGGGCCATAAA	GCCTAAATAGCCCACACGTTCCCC	TTAAATAAGACATCACGAT	GGATCACA	222222222222
				Z0:i:1220						
0234:70:HNNW3BBXX:8:1217:10236:11530 16	mtref 1				CTGGTTCCTACTTC	AGGGCCATAAAGCC	TAAATAGCCCACACGTTCCCCTTA	AATAAGACATCACAATGAA		100000000000000000000000000000000000000
	FFFAA MD:Z:G1 /	\S:i:655	Z0:i:17814	Z1:i:17814	NM:i:3					
0234:70:HNNW3BBXX:8:1224:25337:7978 16	mtref 1	250 68S13M	* 823	0 T(	GTTCCTATTTTAG	GCCATAAAGCCTAA	ATAGCCCACACATTCCCCTTAAAT	AAGACATCACGATGGATCA	CAGGTCTA	222222222222
	111111111111111111111111111111111111111	FFAA MD:Z:1	3 AS:1:710	20:1:23589	21:1:23	89 NM:1:4				
10234:70:HNNW3BBXX:8:1203:3772:45097 16	mtret 1	50 67511M	823	0 60	TICCTACTICAGG	CCATAAAGCCTAAA	TAGECCACACGTTECCETTAAATA	AGACATCACGATGGATCAC	AGGIC JF.	JFJJFJFAJJFFA <fjja< td=""></fjja<>
JFJJJJJF <jjjjjjjjjjjjjjjjjjjjjjjjjjjjj< td=""><td>JJFJJF<fa- i<="" td=""><td>D:Z:II AS:U:/</td><td>55 20:1:0</td><td>+// Z.</td><td></td><td>NMILLI AACCTTAAATACCC</td><td></td><td>TOACCATCOATCACAC</td><td></td><td></td></fa-></td></jjjjjjjjjjjjjjjjjjjjjjjjjjjjj<>	JJFJJF <fa- i<="" td=""><td>D:Z:II AS:U:/</td><td>55 20:1:0</td><td>+// Z.</td><td></td><td>NMILLI AACCTTAAATACCC</td><td></td><td>TOACCATCOATCACAC</td><td></td><td></td></fa->	D:Z:II AS:U:/	55 20:1:0	+// Z.		NMILLI AACCTTAAATACCC		TOACCATCOATCACAC		
10 10 10 10 10 10 10 10 10 10 10 10 10 1	MD:7:9 AS:1:650	70.1.1	7000 71+1+1	2000 N	1+1+2	MAGETTMAATAGEE	сасасоттессствааталоваса	TEACOATOGATEACAO F.	5555FAC55.	36333333333563333
0234+70+HNNW3BBXX+8+2222+10815+11864 16	mtref 1	50 6150M	* 823	A T	CTTTAGAGCCATA	AGCCTAAATAGCCC	ACACATTCCCCTTAAATAACACAT	CACGATGGATCACAGA 1		
	MD:Z:86 AS:1:575	Z0:1:2	9172 Z1:1:2	9172 N	1:1:5					
0234:70:HNNW3BBXX:8:1219:12083:44517 16	mtref 1	50 56525M	* 823	6 C/	GGGCCATAAAGCCT	AAATAGCCCACACG	TTCCCTTTAAATAAGACATCACGA	TGGATCACAGATCTATCAC	CTATTAA	33333F333333
	וכככככככככככככככ	FFAA MD:Z:8	G16 AS:1:7	35 Z(	:i:17895	Z1:i:17895	NM:1:3			
0234:70:HNNW3BBXX:8:1225:16102:17122 16	mtref 1				GGCCATAAAGCCTA	AATAGCCCACACGT	TCCCCTTAAATAAGACATCACGAT	GGATCACAGGTCTATCAC		
000000FA00000000F000000000000000F0000F	FFAA MD:Z:17		Z0:1:6417	Z1:i:6417						
0234:70:HNNW3BBXX:8:2127:23977:26723 16	mtref 1	50 53S28M			GCCATAAAGCCTAAA	TAGCCCACACATTC	CCCTTAAATAAGACATCACGATGA	ATCACAGGTCTATCACCCT	ATTAACCA	22222222222
		FFFAA MD:Z:G				Z1:1:17895	NM:1:3			
0234:70:HNNW3BBXX:8:1220:7415:14555 16	mtref 1	250 52S19M	* 823		CATAAAGTCTAAA	AGCCCACACGTTCC	CCTTAAATAAGACATCACGATGGA	TCACAAGTCTATCACCC J	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
13333333333333333333333333333333333333	MD:Z:7G11	\S:1:635	Z0:1:17794	Z1:1:1779	NM:1:3					
0234:70:HNNW3BBXX:8:2222:29203:16893 16	mtref 1	250 52S29M	* 823	0 A(	CATAAAGCCTAAA	AACCCACACATTCC	CCTTAAATAAAACATCACAATAGA	TCACAGGTCTATCACCCTA	TTAACCAC	222222222222
	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	FFAA MUIZIZ	9 AS:1:035	20:1:40670	21:1:400	70 NM:1:7				
0234:70:HNNW3BBXX:8:2220:10091:22819 10	mtrer 1	50 51530M	823	70.1.6507	ATAAAGCCTAAATA	GCCCACACGTTCCC	CTTAAATAAGACATCACGATGGAT	CACAGGICIAICACCCIAI	TAACCACT	
13333333333333377333333333333333333333	ntcof 1	FRAM MUIZIS	* 022	20:1:0507	TAAACCTTAAATA		TTAAATAACACATCACCATCCATT	ACACCTCTATCACCCTATT	ACCACTO	6333333333333
0234.70.00000000000.0.2224.13433.21149 10		EEAA MD-7-2	COT AC+4+7	6 7/	++++12201	71+1+12201	MM+1+2	ACAGGICIATCACCCIATIS	ACCACIC	F3333333333
0234:70:HNNW3BBXX:8:2222:13088:14783 16	mtref 1	50 49532M	* 823	A 4	AAAGCCTAAATAG	CCACACGTTCCCCT	TAAATAAGACATCACGATGGATCA	CAGGTCTATCACCCTATTA	ACCACTCA	1111111111111
		FEAA MD:Z:3	2 AS:1:810	70:1:814	71:1:814	NM:1:0				
0234:70:HNNW3BBXX:8:2125:1438:7433 16	mtref 1	250 41S21M	* 823		AATAGTCCACACG	TCCCCTTAAATAAG	ACATCACGATGGATCACAAGTCTA	TCACCCTA JJJJJJJ	000000FD.	3333333333333333533F33F
JJFJJJJJJJJFFJJJJJJJFFAAA MD:Z:7G13	AS:1:570	0:i:12010	Z1:i:12010	NM:1:2						
0234:70:HNNW3BBXX:8:1206:26575:44166 16	mtref 1	250 36S45M			GCCCACACGTTCCC	TTAAATAAGACATC	ACGATGGATCACAGGTCTATCACC	CTATTAACCACTCACAAGA		3333F3333333
00000000000000000000000000000000000000	JJJFJJJFJJJJJJJJJJJJ		3GG10 AS:1:7							
0234:70:HNNW3BBXX:8:1114:2412:16260 16	mtref 1	250 35S46M		0 G(	CCACACGTTCCCCT	TAAATAAGACATCA	CGATGGATCACAGGTCTATCACCC	TATTAACCACTCACGGGAG		- JFJFA-AA-FJ
A- <jf-fjffajjff-ajjj<jjjjjjjjjjfjjffajjjjjf<fjfjf.< td=""><td></td><td>F&lt;&lt;- MD:Z:4</td><td>6 AS:i:810</td><td>Z0:i:814</td><td></td><td>NM:i:0</td><td></td><td></td><td></td><td></td></jf-fjffajjff-ajjj<jjjjjjjjjjfjjffajjjjjf<fjfjf.<>		F<<- MD:Z:4	6 AS:i:810	Z0:i:814		NM:i:0				
briel@desktop:/tmp\$										
briel@desktop:/tmp\$										
briel@desktop:/tmp\$										
briel@desktop:/tmps										
briel@desktop:/tmp5										
briel@desktop:/tmp\$										
in ceremesteroh:/ ruhs										

- 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)

		gabriel@desktop: /tmp	00
File Edit View Search Terminal Help			
abriel@desktop:/tmpS_samtools_view_NE0221.bam	lhead -n 20		
00234:70:HNNW3BBXX:8:2205:18142:9016 0	mtref 1 250	7556M * 823 0 CGACATTTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCAC	AAFFFJJJJJJ
		J MD:Z:6 AS:i:760 Z0:i:12201 Z1:i:12201 NM:i:2	
00234:70:HNNW3BBXX:8:1124:22343:30222 16	mtref 1 250	78S3M * 823 Ø ATCCGACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAACCCACACGTTCCCCTTAAATAAGACATCACGATGGAT	
12222222222222222222222222222222222222	JJJJJJJJJJJJJJJJJFFFAA	A MD:Z:3 AS:1:760 Z0:1:12201 Z1:1:12201 NM:1:2	
30234:70:HNNW3BBXX:8:2204:27336:16506 16	mtref 1 250	7556M * 823 0 TGATATCTGGTTCCTATTTAGGGTTATAAAGCCTAAATAGCCCACACATTCCCCTTAAATAAGCATCACGATGGATCAC	
111111111111111111111111111111111111111	JJJJJJJJJJJJJJJJJFFFAA	A MD:Z:0 AS:1:060 Z0:1:34976 Z1:1:34976 NM1:6	
00234:70:HNNW3BBXX:8:2201:11972:21430 16	MTFEF 1 250	745/M = 823 0 GACATELIGUELIACHTAGGGCCALAAGGCCLAAAAGCCCACACGTECCCCTAAATAGGCCACAGGTGGATGACAC	
00224+70+UNNU2DDVV+0+1217+10226+11520 16	stoof 1 250	4 NU.2.7 MS.C.700 20.C.12201 21.C.12201 NN.C.2 74530 * 923 A ATTECTTECTECTECTAAAACCECTAAATACCECTAAATACCECTTAAATAACACATACAAATAA	
10234.70.11110280.7.0.1217.10230.111330 10	FEFAA MD:7:G1 AS:1:6	-655 70:1:17814 71:1:17814 NN:1:3	
00234:70:HNNW3BBXX:8:1224:25337:7978 16	mtref 1 250	68513M * 823 0 TGGTTCCTATTTTAGGGCCATAAAGCCCACACACACTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTA	
	JJJJJJJJJJJJJJJJJFFFAA	MD:Z:13 AS:1:710 Z0:1:23589 Z1:1:23589 NM:1:4	
00234:70:HNNW3BBXX:8:1203:3772:45097 16	mtref 1 250	67S11M * 823 Ø GGTTCCTACTTCAGGGCCATAAAGGCCTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTC JFJFJJ	FJFAJJFFA <fj< td=""></fj<>
		:11 AS:1:755 Z0:1:6477 Z1:1:6477 NM:1:1	
00234:70:HNNW3BBXX:8:1205:5335:30538 16	mtref 1 250	62S8M * 823 0 CTACTTCAGGGCCATAAAGCTTAAATAGCCCACACGTTCCCCTTAAATAAGCACACGATGGATCACAG FJJJJFA <jjjfjj< td=""><td>000000000F00:</td></jjjfjj<>	000000000F00:
	MD:Z:8 AS:1:650	Z0:1:12090 Z1:1:12090 NM:1:2	
00234:70:HNNW3BBXX:8:2222:10815:11864 16	mtref 1 250	61S9M * 823 0 TACTTTAGAGCCATAAAGCCTAAATAGCCCACACATTCCCCTTAAATAAGACATCACGATGGATCACAGA JJJJJJJJJJJJJJJ	
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA	MD:Z:8G AS:1:575	Z0:i:29172 Z1:i:29172 NM:i:5	
00234:70:HNNW3BBXX:8:1219:12083:44517 16	mtref 1 250	56525M * 823 0 CAGGGCCATAAAGGCCTAAATAGCCCACACGTTCCCTTTAAATAAGACATCACGATGGATCACAGATCTATCACCCTATTAA	00000F00000
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	JJJJJJJJJJJJJJJJJJJFFFAA	A MD:Z:8G16 AS:t:735 Z0:t:17895 Z1:t:17895 NM:t:3	
00234:70:HNNW3BBXX:8:1225:10102:17122 10	Mtret 1 250	SSS1/M = 823 0 AGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTATCAC JJJJJJ.	111111111111111111111111111111111111111
JJJJJJJJFAJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	FFAA MD:2:17 AS:1:0		
00234:/0:MNNW3BBAX:8:212/:239//:20/23 10	MUTET 1 250	33528M * 823 0 GOCCATAAAGCCTAAATAGCCCACACATTCCCCTTAAATAAGACATCACGATGAATCACGATGAATCACGCCTATTAACCA	
90234 • 70 • HNNU3BBXX • 8 • 1220 • 7415 • 14555 16	mtrof 1 250	55510M * 823 A GEGATABAGETABATAGEGAGEGETETBABATAGEGATGAGEGATGAGATGA	
11111111111111111111111111111111111111	MD:7:7611 AS:1:6	-635 20-11-17794 21-11-17794 NM-11-3	
00234:70:HNNW3BBXX:8:2222:29203:16893 16	mtref 1 250	52S29M * 823 0 ACCATAAAGCCTAAATAACCCACACATTCCCCTTAAATAAA	
	JJJJJJJJJJJJJJJJJJJJFFFAA	A MD:Z:29 AS:1:635 Z0:1:40670 Z1:1:40670 NM:1:7	
00234:70:HNNW3BBXX:8:2226:16691:22819 16	mtref 1 250	51S30M * 823 Ø CCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTATCACCCTATTAACCACT	
		A MD:Z:30 AS:1:785 Z0:1:6507 Z1:1:6507 NM:1:1	
00234:70:HNNW3BBXX:8:2224:13453:21149 16	mtref 1 250	50S31M * 823 0 CATAAAGCTTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGATTACAGGTCTATCACCCCTATTAACCACTC	
		A MD:Z:3C27 AS:i:760 Z0:i:12201 Z1:i:12201 NM:i:2	
00234:70:HNNW3BBXX:8:2222:13088:14783 16	mtref 1 250	49S32M * 823 0 ATAAAGCCTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTATCACCCCTATTAACCACTCA	
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	JJJJJJJJJJJJJJJJJJFFFAA	A MD:Z:32 A5:1:810 Z0:1:814 Z1:1:814 NM:1:0	
00234:70:HNNW3BBXX:8:2125:1438:7433 16	mtref 1 250	41S21M * 823 0 TAAATAGTCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAAGTCTATCACCCTA JJJJJJJJJJJJJJJJJJJJJJJJJJJ	10000100001F01
JJJFJJJJJJJJFFJJJJJJJFFAAA MD:Z:7G13	AS:1:570 Z0:1:1	12010 Z1:1:12010 NA:1:2	
00234:70:HNNW3BBXX:8:1200:20575:44166 16	mtref 1 250	36545M 823 0 AGCCCACACGTTCCCCTTAAATAAGACATCACGATGGGTTCACAGGTCTATCACCCTATTAACCACTCACAAGAGCTCTCCCA	JJJJFJJJJJ.
JJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJ	JJJFJJJFJJJJJJJJJFFFAA	A MD:2:33GG10 AS:1:760 20:112201 21:112201 NMT:2	
00234:70:HNNW3BBAX:8:1114:2412:10200 10	115551115A5-15A5-5-4-	355400 * 823 0 GCCCACACUTECCETARATARAGACATICACATIGUATCACAGUTETATCACCETATTAACCACTCACGUGAGUETETCAT	-JEJEA-AA-E
abciel@deckton:/tmn\$		- HD.2.40 AS.C.810 20.C.814 21.C.814 NH.C.0	
abriel@desktop:/tmp5			
abriel@desktop:/tmpS			
abriel@desktop:/tmpS			
abriel@desktop:/tmp\$			
nabriel@desktop:/tmpS			

- 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



Date: 3rd

Project work

12th

21th





## 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.
- <u>http://teaching.healthtech.dtu.dk/22126</u>
- 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