# Databehandling fra sekventering

## Indhold

Databehandling fra sekventering	1
Tolkning af Workflow-rapporten i Epi2Me	2
Efterbehandling af data	4
Hvad har I fundet?	4
Videre muligheder for databehanding	5
FASTO-filer	5

Data fra MinION-devicen oversættes i programmet minKNOW til en FASTQ-fil. FASTQ-filen gemmes på computeren i den mappe (directory), som man har angivet i minKNOW.

Data kan derefter bearbejdet i programmet Epi2Me:

**Epi2Me i skyen:** Aligner (dvs, sammenligner) de uploadede sekvenser i FASTQ-fil med databasen og identificerer bakterierne. Det sker i et online workflow.



**Epi2Me-desktopagent på computeren:** Henter FASTQ-filen fra computeren og uploader den til skyen.



Figur 1. Programmer og databehandling.

*Vi laver databehandlingen i Epi2Me fælles for klassen. Derefter tolker I på resultaterne i grupper.* 

#### Tolkning af Workflow-rapporten i Epi2Me

Første side indeholder oplysninger om sekventeringens forløb, antal succesfulde og mislykkede reads og deres gennemsnitlige nøjagtighed i forhold til databasens sekvenser.

🖻 🖅 M Indi	oakke (942) - kreste 🧢 NNF-Workshop-Deltag 📘	Rediger Opgave - Lecti 🛛 📄 Færdigmalet Træbeklæ	🔘 Community - Dashboai 🍦 Fastq 16	S [Instance $ imes$ + $ imes$	- 1
$\leftarrow \   \rightarrow \   \heartsuit$		flow_instance/221879		□ ☆	t≡ ll_
		FOR RESEARCH USE ONLY			
Fastq 16S	[Instance ID: 221879]				
STATE	stopped	WORKFLOW	Fasto 165 r3.2.2		
DURATION	34d 3h 52m 27s		[Workflow ID: 1760]		
STARTED	26.11.2019 15.15.33	DATA REGION	eu-west-1		
STOPPED	30.12.2019 19.08.00				
Attributes: 🙃 ani2n	erzelik eize: 200000 🙆 epi2mermex eize: 50000000 🛱 eni2	mesolit.reade: 2000		~	
	•	Calogory			
QC and Barcod	ing [rev. 3.10.4] → 16S Classification [rev. 3.0.1]				
				CSV	A Share
	9% CLASSED % UNCLASSED	READS CLASSIFIED 91	reads unclassified 1	ave. accuracy 84%	
Taxa at Ra	nk:Genus → her for at søge Ħ	SHOWING TAXA WITH READS	i 📕 øj 🐠 🔒	^ 🕳 📾 🌾 ላን 🕏	DAN 15:3

Længere nede på siden vises de samlede resultater fra alignment med databasen:

🛱 🖅 M Indbakke (942) - krest	e 🦱 NNF-Workshop-Deltag	L Rediger Opgave - Lecti	📄 Færdigmalet Træbeklæ	O Community - Dashboai	🍦 Fastq 16S [Instance $ imes$	+ ~	-	ø ×
$\leftarrow$ $\rightarrow$ $\circlearrowright$ $\textcircled{a}$ h	tps://epi2me.nanoporetech.com/w	orkflow_instance/221879				□ ☆	¢	e
Taxa at Rank: Genus	Cumulative Reads +	SHOWING TAXA WITH NCBI Taxono	reads omy Tree				HIDE FILTERS	r
Leuconostro	34		MINIMUM ABUNDANCE CUTOFF		SHOW TOP	N TAXA		
Serratia	10		3% 1% 0.5% 0.1% 0%		10 20 30	100 200		
Lactobacillus	5							
Lactococcus	13					() EXPOR	RT PNG ENABLE ZO	CM
Enterobacter	2		Enterobi	acter				
Erwinia	2		Klebsiell	a				
Weissella	2							
Klebsiella	La .		• Erwinia					
Pantoea	11		Pantoea Tatumal					
Tatumella	11		• Tatumen	a				
Xenorhabdus	11		Xenorha	ibdus				
Pectobacterium	1							
			Pectoba	cterium				
		root • • •	• Serratia					
			• La	tobacillus				
i≣ Reads per Barcode ID			w w	eissella				
ID 💙 Read Count	EXCLUDE /	Lineage Taxa - Not Classified						
BC03 27	EXCLU	Taxa - Classified	• La	telococcus				
BC07 53	DICLU	æ						
No Barcode 📃 12	EXCLU	xe.						
🕂 🔎 Skriv her for at søge	Ħ	😑 🚍 🟦	🖻 🖉 🦉 🧃	🗧 💆 🤞	🔒 ^ 👄 •	📾 🦽 🗘 🗮	DAN 15:38	020 😼

Figur 10 viser hvad de forskellige elementer på siden angiver:

•	🔄 M Indbakke (942) - kreste	lacktrice NNF-Workshop-Deltag	diger Opgave - Lecti	Færdigmalet Træbeklæ	O Community - Dashboa	Fastq 16S [Instance $ imes$ + $ imes$	- 0	×
$\leftarrow$	$\rightarrow$ O $\textcircled{a}$ http:	://epi2me.nanoporetech.com/workflow	_instance/221879			□ ☆	te la le	
			NCBI Taxono	my Tree			HIDE FILTERS	^
- 1	Taxon ¢	Cumulative Reads -						
	Leuconostoo	30	1 1	MINIMUM ABUNDANCE CUTOFF		SHOW TOP N TAXA		
	Lactobacilus	5	I I	3% 1% 0.5% 0.1% 0%		10 20 30 100 200		
	Lactococcus	12				() ED	PORT PNG ENABLE ZOOM	
	Erwinia	12						
	Weissella	2		Enteroba	icter	Filtre ift nøjagtig	thed	
	Serratia	1				There in the Jug the	lica	
	Enterobacter	1		Erwinia				
	Pantoea	14		Pantoea				
	Pectobacterium	11		•				
	identificered	e reads	root Caracteria	• Serratia				Н
		Valg af		• La	ctobacillus	Ved at føre cursere	n over	
		barcode			uconostor	novnot visos onlych	ingor om	
						navnet vises opiysn	inger om	
ſ	Reads per Barcode ID	\		1 - "	cissolia	nøjagtighed.		
	ID M Read Count		Lineage			Vod at klikko på pa	upot vicos	
		EXCLUDE FAL	Taxa - Classified	•• <u>La</u>	ctococcus	veu at klikke på hav	met vises	
	8003 27	The second	-		genus	uddybende forklari	nger om	
	BC07 33	EACCODE .		2 CUMUL	ATIVE READ COUNT	haktorian	0	
<u> </u>	No Barcode 12	NCLUE		85.5	AVG. ACCURACY	pakterien		
). 🗳	$\mathcal P$ Skriv her for at søge	E C	è 🗖 🕯	🖻 🖷 🧶 🧃	ş 🔼 🛷 🔹	👷 🔨 📥 🦽 🕼	to 16:54 16:54 22-02-2020	5

#### Efterbehandling af data

#### Hvad har I fundet?

Del de identificerede bakterier imellem jer og undersøg dem nærmere. Vær opmærksom på om de oplysninger I finder er på arts-, slægts- eller familie-niveau. Indenfor samme slægt kan der være stor forskel på enkelte bakterietypers biologi og eventuelle patologi (evne til at fremkalde sygdomme).

Navn på bakterie			
Hvordan ser den ud?			
Form: kok, stav, skrueformet?			
$\circ$ $\Box$ $\sim$			
Lejring: -enkelt (monokok), -dobbelt (diplokok, diplobacillus), -kæde (streptokok /streptobacillus), -stak (staphylokok)?			
0 00 9 80 0088			
Gram+ eller gram-?			
Hvilke egenskaber og			
stofskifteprocesser er			
karakteristiske for bakterien?			
Kan bakterien være patogen for			
planter, dyr eller mennesker?			
Hvor forekommer den normalt?			
Hvilke abiotiske forhold			
karakteriserer dens niche?			
Lever den associeret til andre			
organismer, fx planter eller dyr?			
Ville du forvente at finde den, der			
hvor I har jeres prøve fra?			

Fremlæg jeres bakterier for klassen.

Sammenlign de bakterier der forekommer på de fire lokaliteter. Hvad viser de om lokaliteterne?

#### Videre muligheder for databehanding

De enkelte sekvenser i FASTQ-filerne kan kopieres og sammenlignes med indrapporterede sekvenser i genbank. Det gøres ved Nucleotid BLAST.

- 1. Åben hjemmesiden: https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGETYPE=BLASTHome
- 2. Vælg Nucleotide BLAST
- 3. Kopier en sekvens fra FASTQ-filen og indsæt den i søgefeltet 'Enter... FASTA sequence'.
- 4. Tryk på BLAST.

Nu fremkommer en side med de indrapportede sekvenser, som matcher sekvensen bedst.

- 5. Hvilke oplysninger kan ses på hjemmesiden?
- 6. Hvilke bakterier matcher bedst?
- 7. Hvor godt matcher sekvensen med de indrapporterede?

### FASTQ-filer

Resultaterne fra sekventeringen vil ligge på computeren som en FASTQ-fil. I FASTQ-filen er hver DNAsekvens repræsenteret af 4 linjer.

- Unikt ID starter med et '@'. Linien kan fx indeholde information om prøven og flowcellen.
- Selve DNA-sekvensen (ATGC....)
- Et '+', som markerer at sekvensen slutter
- En linje med en kvalitetskode for hver af baserne i linie 2. Kvalitetskoden er angivet med ASCII-symboler<sup>1</sup>.
- 1. Åben en af FASTQ-filerne i en text-editor.
- 2. Identificer de fire linjer for hver DNA-sekvens. Hvilke oplysninger kan du finde i linierne?

<sup>&</sup>lt;sup>1</sup> Værdien er angivet med ASCII-symboler, hvis betydning kan findes ved søgning på nettet. Værdien angiver  $Q = -log_{10}P$ , hvor P er sandsynligheden for at en base er fejlbestemt. Dvs. at jo højere Q-værdi, jo mere sikker er bestemmelsen.