<p>| Saturday | 2p – 5p | Rayan Chikhla | Metagenomics Assembly, then Open Lab |</p>
<table>
<thead>
<tr>
<th>Day</th>
<th>Time</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saturday</td>
<td>2p – 5p</td>
<td>Rayan Chikli</td>
</tr>
</tbody>
</table>
Schedule

- 2 pm: metagenomics assembly lecture
- 3 pm: metagenomics assembly lab or open lab

Also at 4 pm: optional Metagenomics 'faculty lunch coffee'
Schedule

- 2 pm: metagenomics assembly lecture
- 3 pm: metagenomics assembly lab
  or open lab
Congratulations to

1. Forrest Walker
2. Alena di Primio
3. ? you?

for completing the hidden *raccoon facts* challenge
Metagenomics assembly

Rayan Chikhi
with some help from Dag Ahren and Sergey Nurk

Institut Pasteur

Workshop on Genomics 2020
I wanted participants to know about..

The discovery of Asgard archea

[Takei and Horikoshi, 1999]

Analysis of single cells of a super-abundant ocean bacteria

[Kasshtar et al, 2014]

Newfound groups of bacteria

[Brown et al, 2015]
Metagenomics

What?
- Term coined by Jo Emily Handelsman *et al* (1998)
- *the application of modern genomics technique without the need for isolation and lab cultivation of individual species* (Chen, Pachter 2005)

Why?
- Most microorganisms are not possible to culture and hence the only way to investigate their genome is to use metagenomics.
**Metagenomics vs metataxonomics**

**Metataxonomics** (will be on Microbiome day)
- 16S or 18S rRNA sequencing
- Fast and cost-effective
- Limited (no gene content, no viruses)
- Applications: taxonomic profiling, rRNA phylogeny, ..

**Metagenomics**
- Shotgun sequencing of DNA
- Versatile, enables assembly
- Applications: functional genome analyses, whole genome phylogeny, pathogen detection, ..

Source: Breitwieser et al, Briefings in Bioinformatics 2017
Metagenomics analysis scenarios

**Assembly** route
1. *de novo* assembly
2. contigs binning
3. taxonomic assignment

**Species identification** route
- Taxonomic assignment of reads
  - Kraken2 (minimizers), Kaiju, Centrifuge, etc

**Direct comparison** route
- direct comparison of experiments (e.g. similarity matrix)
  - Mash, Sourmash, Simka, etc
- (won’t be covered here)
## Elements of choice

<table>
<thead>
<tr>
<th></th>
<th>selection</th>
<th>all reads</th>
<th>assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Biological question</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>presence/absence of known species</td>
<td>***</td>
<td>***</td>
<td>*</td>
</tr>
<tr>
<td>discovery of novel species</td>
<td>*</td>
<td></td>
<td>***</td>
</tr>
<tr>
<td>functional analysis</td>
<td></td>
<td>*</td>
<td>**</td>
</tr>
<tr>
<td><strong>Complexity of the community</strong></td>
<td>H/M/L</td>
<td>M/L</td>
<td>L</td>
</tr>
<tr>
<td><strong>Requirements</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>computational time</td>
<td>++</td>
<td>+</td>
<td>++++</td>
</tr>
<tr>
<td>sequencing depth</td>
<td>+</td>
<td>+</td>
<td>++++</td>
</tr>
<tr>
<td>bioinformatics skills</td>
<td>+</td>
<td>+</td>
<td>++++</td>
</tr>
</tbody>
</table>

Computational time: from a few minutes to a few days/weeks
Read-based approaches: web servers or pipelines

Credit: H. Touzet, CNRS
Metagenome-Assembled Genomes (MAGs)

A MAG is **one bin** selected out of an assembled metagenome.

**Advantages**
- Well-established sequencing (Illumina)
- Cheap

**Disadvantages**
- In complex communities:
  - Only the most abundant taxa are likely to be "well" assembled
  - High computational requirements
SAGs (Single-Amplified Genomes)

Relies on recent techniques that allows for isolation of single cells followed by single cell amplification

**Advantages**
- Minimise the risk of false hybrid assembly
- It is possible to select which cells to sequence

**Disadvantages**
- Complex laboratory protocols
- Contamination (even from kits/reagents)
- Amplification is biased (new protocols are under development - spoiler alert: they’re still biased)
Metagenomic assembly

Reconstruct genomes of species, possibly even strains, from short read sequencing data of an environment

https://fr.slideshare.net/MadsAlbertsen/20131202-mads-albertsen-extracting-genomes-from-metagenomes
Challenges

1. closely related strains
2. uneven depths, & low depths
3. inter-species repeats
4. size of datasets
5. lack of long reads

(adapted from A. Korobeynikov’s talk)

Fig: Olsen et al, 2017
Metagenomic assembly is impossible

Two competing goals:
- assemble similar sequences from related genomes together
- do not assemble similar sequences from unrelated genomes

Mihai Pop, Sergey Koren, Dan Sommer

Credit: H. Touzet, CNRS
What comes after assembly

Contigs binning
- CONCOCT
- MetaBAT2
- MaxBin2

Taxonomic identification
- CAT/BAT
- ProPhyle
- PhyloPythiaS

anvi’o pipeline
Metagenome assembly software

- metaSPAdes  [Nurk et al, Genome Res., 2017]
- MEGAHIT   [Li et al, Methods, 2016]
- metaFlye   [Kolmogorov et al, bioRxiv, 2019]
- Minia-pipeline  [me!]
- IDBA-UD
- Ray-meta
- SOAPdenovo2
- metaVelvet/-SL
- Omega
- InteMAP
- Meraga
- Velour
- A∗
Under the hood of metagenome assemblers
MEGAHIT $< v1.0$

1. **input reads $R$**
2. **count $(k_{\min}+1)$-mer, output solid & mercy edges**
3. $k \leftarrow k_{\min}$
4. **build SdBG of order $k$**
5. **remove tips; merge bubbles; progressively remove low local coverage edges; output contigs $C_k$**
6. $k \leftarrow k + \text{step}$
7. **$k \leq k_{\text{max}}$?**
   - **Yes**: Extract $(k+1)$-mers from reads $R$ and contigs $C_{k-\text{step}}$
   - **No**: **end**
metaSPAdes

Graph construction & simplifications

Further repeat-resolution

Output contigs
Multi-k

In principle, better than single-k assembly.
Visualization of multi-k graphs

*Salmonella* genome, SPAdes assembly

\[ k = 99 \]
In contrast, with single-k

Salmonella genome, Velvet assembly

$k = 91$ (too high, but shown for comparison)

Metagenomics with long reads

1. metaFlye  [Kolmogorov et al, 2019]
2. wtdbg2  [Nicholls et al, GigaScience, 2019]
3. Canu  [see wtdbg2 article]
4. miniasm + Racon

Oxford Nanopore: **needs polishing**

Alternative route: HiC, linked reads
Too complex to describe its inner workings
metaFlye

Too complex to describe its inner workings
metaFlye
When *can* you assemble

Look at *k*-mer histograms of the reads! (KMC, DSK tools)

Credit: www.cmbi.ru.nl/~dutilh/metagenomics/course_HAN_2014/Speth.pdf
Digital normalization

https://github.com/dib-lab/khmer
- Reduce dataset size
- Facilitates assembly

Potential drawbacks:
- assembly fragmentation
- low-coverage variant loss

Why you shouldn’t use digital normalization
http://ivory.idyll.org/blog/
why-you-shouldnt-use-diginorm.html
Evaluation metrics

Same as regular assembly:
- N50, NG50
- Total size
- % of reads mapping correctly back to the assembly
- Number of predicted genes
- % of contigs matching some known references

Metagenome-specific:
- metaQUAST
- CheckM, marker genes, [Parks et al, Genome Res. 2015]
- VALET, internal consistency, [Olson et al, BFB 2017]
CAMI benchmark

- 3 artificial communities
  ▶ low, medium, high complexity (600 genomes, 5x15 Gbp)
- 6 assemblers evaluated: MEGAHIT, Minia, Ray-meta, ..
Quality of metagenome assembly

a: all genomes,  b: genomes with ANI \(\geq\) 95%,  c: genomes with ANI < 95%

No assembler could reconstruct close strains.

[Sczyrba, Nat Meth 2018]
Metagenomics software is still immature, story time..
Mosaic DNANexus Challenge 2018

Focus on **strains** assembly

**Evaluation** metrics:
- Genome Fraction
- misassemblies
Mosaic DNANexus Challenge 2018

Focus on **strains** assembly

**Evaluation** metrics:
- Genome Fraction
- misassemblies

<table>
<thead>
<tr>
<th>Method</th>
<th>N50</th>
<th>Genome Fraction</th>
<th># misassemblies</th>
</tr>
</thead>
<tbody>
<tr>
<td>What a regular assembler would give</td>
<td>7.1 Kbp</td>
<td>84.1%</td>
<td>1998</td>
</tr>
</tbody>
</table>
Mosaic DNANexus Challenge 2018

Focus on **strains** assembly

**Evaluation** metrics:
- Genome Fraction
- misassemblies

<table>
<thead>
<tr>
<th>Method</th>
<th>N50</th>
<th>Genome Fraction</th>
<th># misassemblies</th>
</tr>
</thead>
<tbody>
<tr>
<td>What a regular assembler would give</td>
<td>7.1 Kbp</td>
<td>84.1%</td>
<td>1998</td>
</tr>
<tr>
<td>Initial step (BCALM)</td>
<td>0.5 Kbp</td>
<td>95.3%</td>
<td>23</td>
</tr>
</tbody>
</table>

Initial step (BCALM)
Focus on **strains** assembly

**Evaluation** metrics:
- Genome Fraction
- misassemblies

<table>
<thead>
<tr>
<th>Method</th>
<th>N50</th>
<th>Genome Fraction</th>
<th># misassemblies</th>
</tr>
</thead>
<tbody>
<tr>
<td>What a regular assembler would give</td>
<td>7.1 Kbp</td>
<td>84.1%</td>
<td>1998</td>
</tr>
<tr>
<td>Initial step (BCALM)</td>
<td>0.5 Kbp</td>
<td>95.3%</td>
<td>23</td>
</tr>
</tbody>
</table>

*don’t do it*
Business

DNA nexus-Powered Mosaic Microbiome Platform Announces Winners of First Community Challenge
Business

DNA nexus-Powered Mosaic Microbiome Platform Announces Winners of First Community Challenge
Evaluating metagenome assemblies is hard
Conclusion

- Metagenome assembly is a hard problem
- Due to strains & low-abundance species, mostly
- Trade-off between contiguity, and genome fraction/misassemblies. Questions on assemblies ranking.
- So far, limited availability of: long reads, Hi-C, linked-reads

References:

- Ayling et al, New approaches for metagenome assembly with short reads, 2019
- metaFlye article

Acknowledgments: Dag Ahren, Sergey Nurk, Camille Marchet, Antoine Limasset, the fantastic team of the Workshop on Genomics 2020, Chris Quince, Aaron Darling, Guillaume Rizk, Claire Lemaitre, Pierre Peterlongo, Charles Deltel, Paul Medvedev, Dominique Lavenier
Exercice

$k$-mers:

1. ACA
2. AGA
3. AGT
4. CAT
5. GTC
6. TAG
7. TCA
8. TTG

Two strains of a short genome are in this dataset, please assemble them. Ignore reverse-complements.
- Discard TTG (connected to nothing)
- Observe a $k$-mer was missing (GAC)
- Two strains: TAGTCAT, TAGACAT