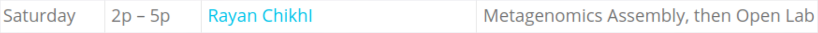
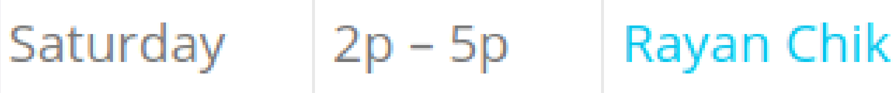
Schedule



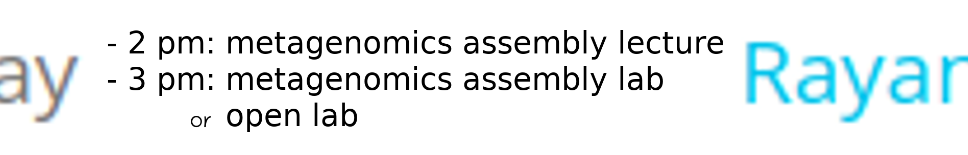
Saturday	2p – 5p	Rayan Chikhl	Metagenomics Assembly, then Open Lab

Schedule



Saturday	2p – 5p	Rayan Chik

Schedule

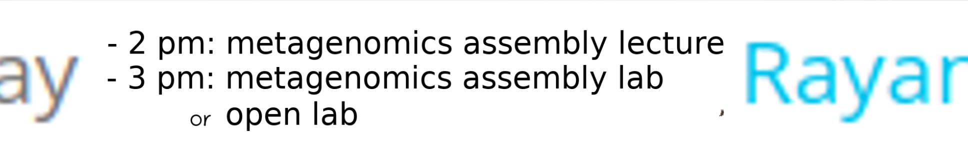


- 2 pm: metagenomics assembly lecture
- 3 pm: metagenomics assembly lab

or open lab

Also at 4 pm: optional Metagenomics ’faculty ~~lunch~~ coffee’

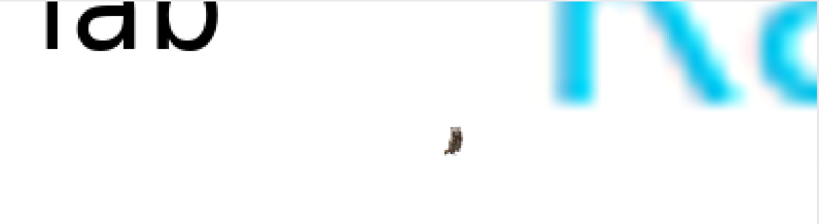
Schedule



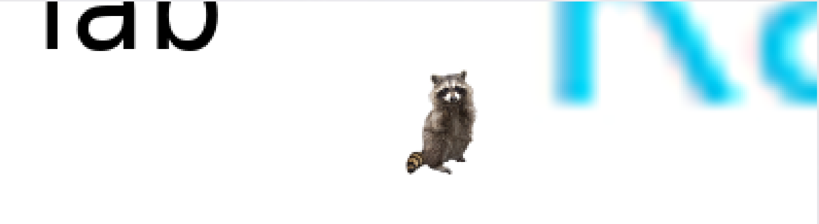
- 2 pm: metagenomics assembly lecture
- 3 pm: metagenomics assembly lab

or open lab

?


lab

?


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

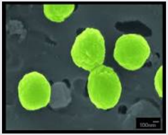
Workshop on Genomics 2020

🤒 !

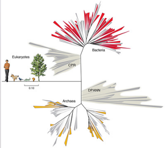
🤒 ! 🧟 ?

I wanted participants to know about..

The discovery of Asgard archea [Takai and Horikoshi, 1999]



Analysis of single cells of a super-abundant ocean bacteria [Kashtar *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

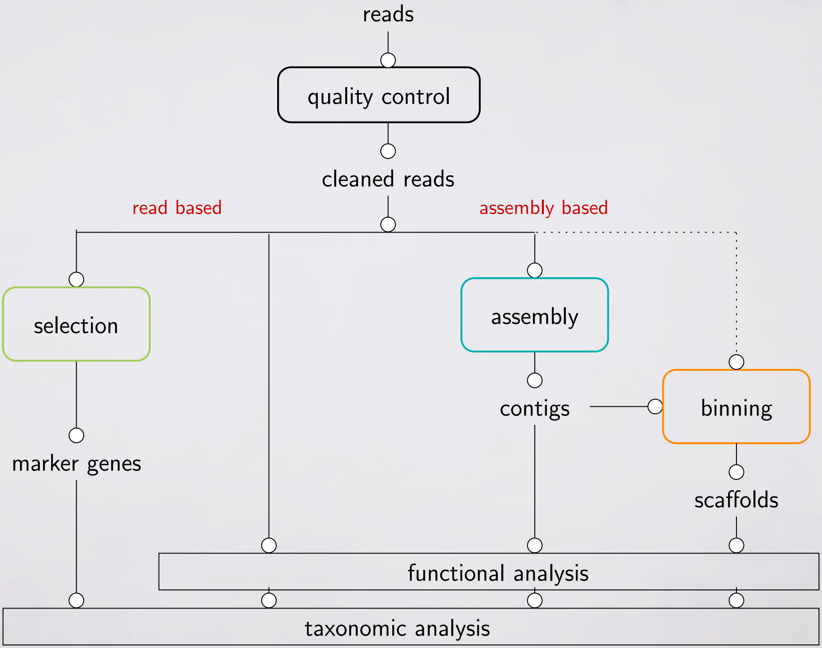
1. *de novo* assembly
2. contigs binning
3. taxonomic assignment

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)



Credit: H. Touzet, CNRS

Elements of choice

	selection	all reads	assembly
Biological question			
presence/absence of known species	⋆⋆⋆	⋆⋆⋆	⋆
discovery of novel species	⋆		⋆⋆⋆
functional analysis		⋆	⋆⋆
Complexity of the community	H/M/L	M/L	L
Requirements			
computational time	++	+	+++
sequencing depth	+	+	+++
bioinformatics skills	+	+	+++

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

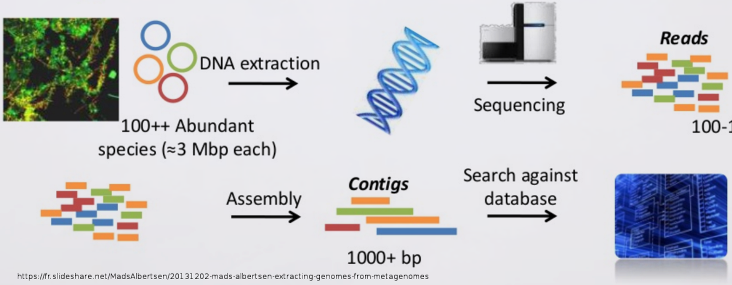
- Minimise the risk of false hybrid assembly
- It is possible to select which cells to sequence

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

no hits?

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)

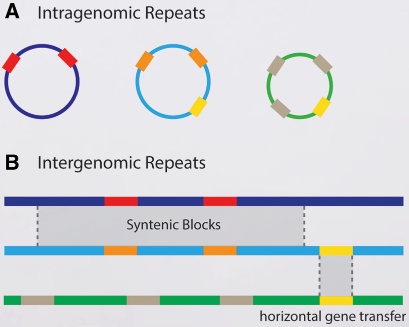
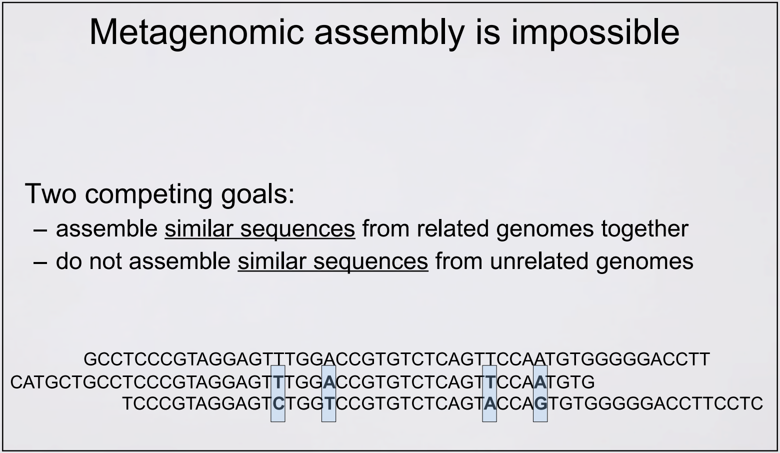


Fig: Olsen *et al, 2017*



Metagenomic assembly is impossible

Two competing goals:

- assemble similar sequences from related genomes together
- do not assemble similar sequences from unrelated genomes

```
      GCCTCCCGTAGGAGTTTGGACCGTGTCTCAGTTCCAATGTGGGGGACCTT
CATGCTGCCTCCCGTAGGAGTTTGGACCGTGTCTCAGTTCCAATGTG
         TCCCGTAGGAGTCTGGTCCGTGTCTCAGTACCAGTGTGGGGGACCTTCCTC
```

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

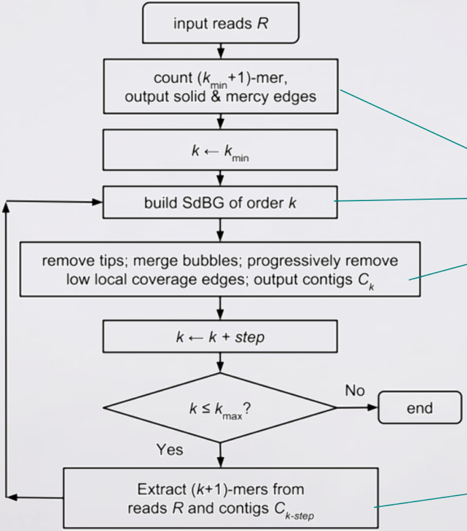
anvi’o pipeline

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

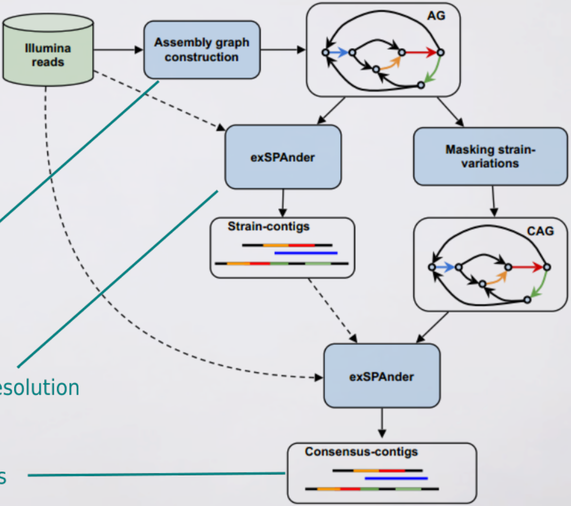

input reads *R*
count (k_{min}+1)-mer, output solid & mercy edges
$k \leftarrow k_{min}$
build SdBG of order *k*
remove tips; merge bubbles; progressively remove low local coverage edges; output contigs C_k
$k \leftarrow k + step$
$k \leq k_{max}$?
No
end
Yes
Extract (*k*+1)-mers from reads *R* and contigs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

graph construction

simplifications (errors & variants)

output contigs

metaSPAdes

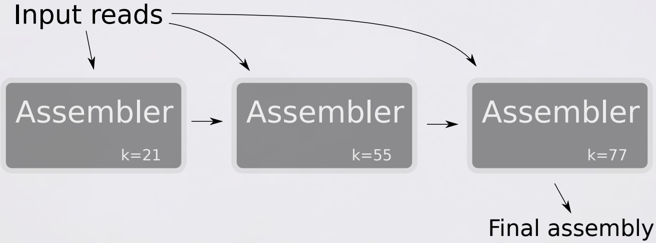


Graph construction & simplifications

Further repeat-resolution

Output contigs

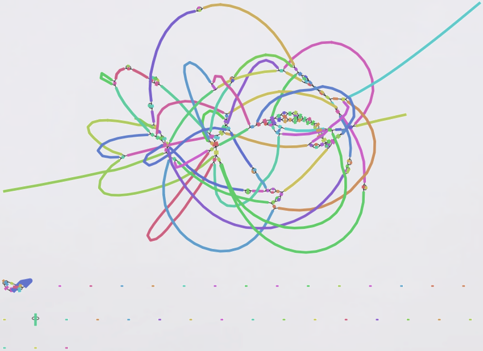
Multi-k



In principle, **better** than single-k assembly.

Visualization of multi-k graphs

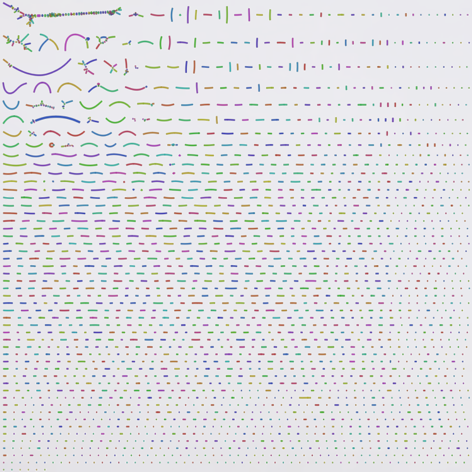
Salmonella genome, SPAdes assembly



k = 99

In contrast, with single-k

Salmonella genome, Velvet assembly



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

https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size

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

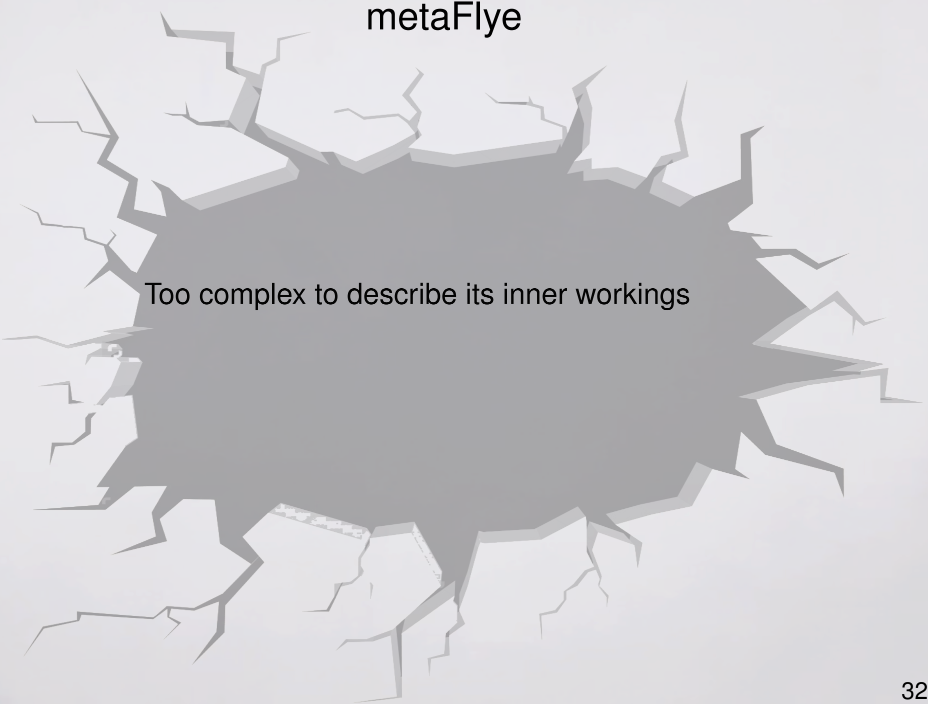
Oxford Nanopore: **needs polishing**

Alternative route: HiC, linked reads

metaFlye

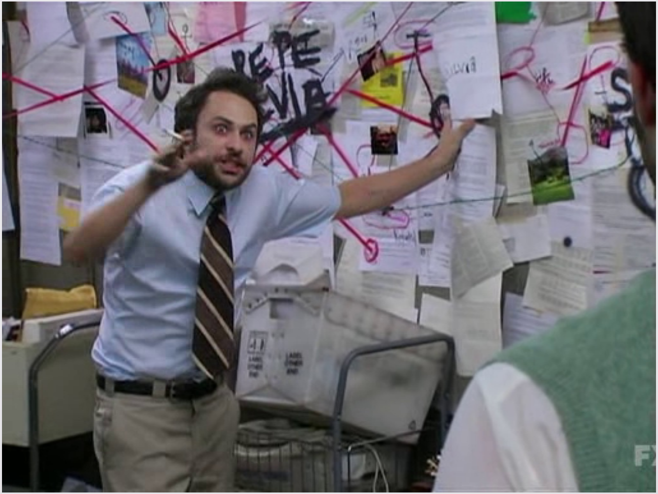
Too complex to describe its inner workings

metaFlye



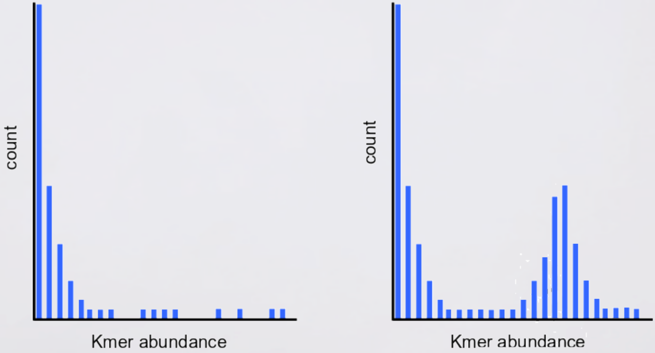
Too complex to describe its inner workings

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:

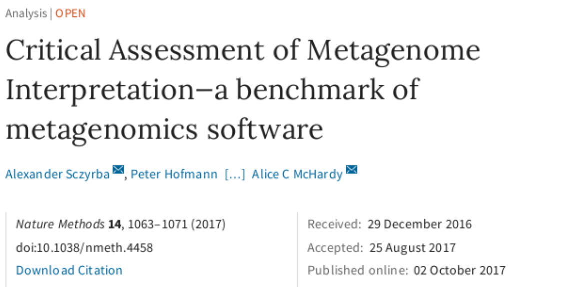
- N50, NG50
- Total size
- % of reads mapping correctly back to the assembly
- Number of predicted genes
- % of contigs matching some known references

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, ..



Analysis | OPEN

Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software

Alexander Sczyrba ✉, Peter Hofmann [...] Alice C McHardy ✉

Nature Methods **14**, 1063–1071 (2017)

doi:10.1038/nmeth.4458

Download Citation

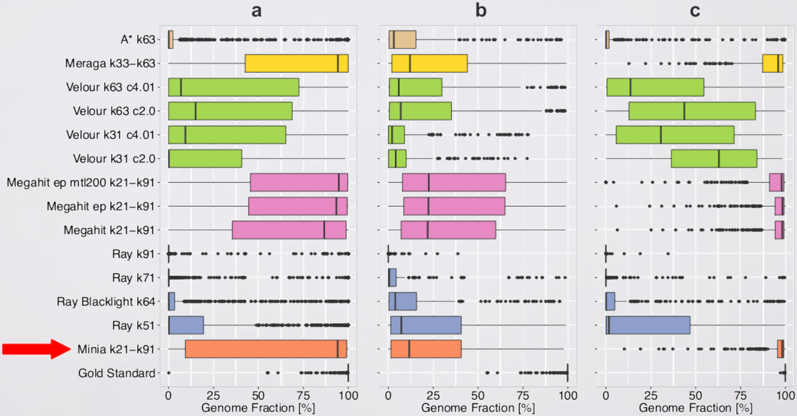
Received: 29 December 2016

Accepted: 25 August 2017

Published online: 02 October 2017

Quality of metagenome assembly

a: all genomes, b: genomes with ANI >= 95%, c: genomes with ANI < 95%

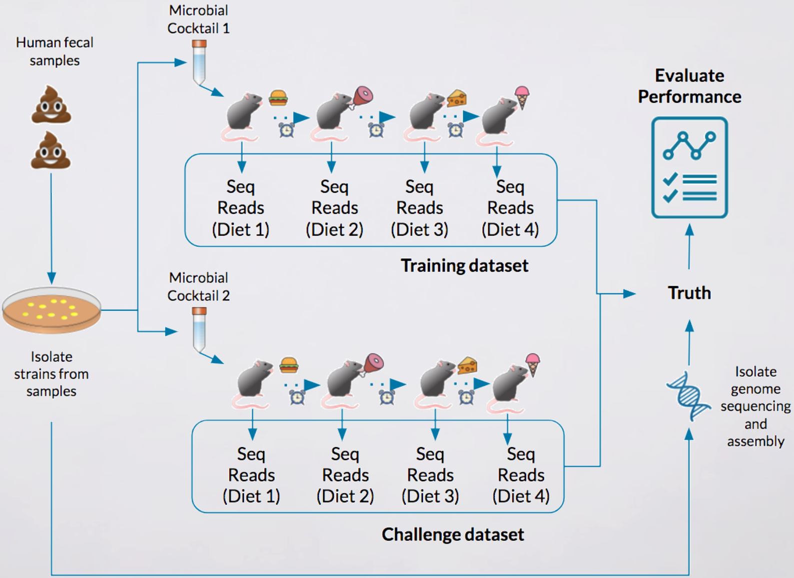


[Sczyrba, Nat Meth 2018]

No assembler could reconstruct **close strains**.

Metagenomics software is still immature, story time..

Mosaic DNANexus Challenge 2018



Mosaic DNANexus Challenge 2018

Focus on **strains** assembly

Evaluation metrics:

- Genome Fraction
- misassemblies

Mosaic DNANexus Challenge 2018

Focus on **strains** assembly

mosaic

Evaluation metrics:

- Genome Fraction
- misassemblies

Method	**N50**	**Genome Fraction**	**# misassemblies**
What a regular assembler would give	7.1 Kbp	84.1%	1998

Mosaic DNANexus Challenge 2018

Focus on **strains** assembly

mosaic

Evaluation metrics:

- Genome Fraction
- misassemblies

Method	**N50**	**Genome Fraction**	**# misassemblies**
What a regular assembler would give	7.1 Kbp	84.1%	1998
Initial step (BCALM)	0.5 Kbp	**95.3%**	**23**

Mosaic DNANexus Challenge 2018

Focus on **strains** assembly

Evaluation metrics:

- Genome Fraction
- misassemblies

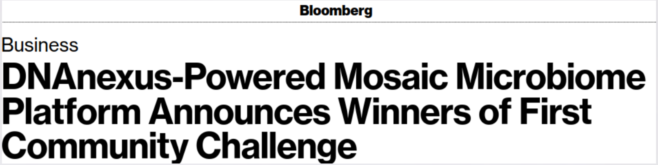
Method	**N50**	**Genome Fraction**	**# misassemblies**
What a regular assembler would give	7.1 Kbp	84.1%	1998
Initial step (BCALM)	0.5 Kbp	**95.3%**	**23**

don't do it

Bloomberg

Business

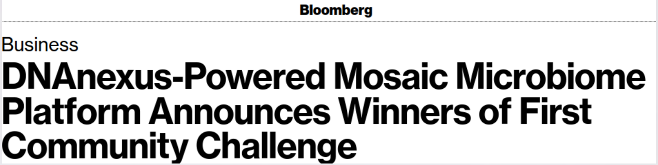
DNAnexus-Powered Mosaic Microbiome Platform Announces Winners of First Community Challenge



Bloomberg

Business

DNAnexus-Powered Mosaic Microbiome Platform Announces Winners of First Community Challenge



Bloomberg

Business

DNAnexus-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
- out of RAM? `https://github.com/GATB/minia-pipeline`

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

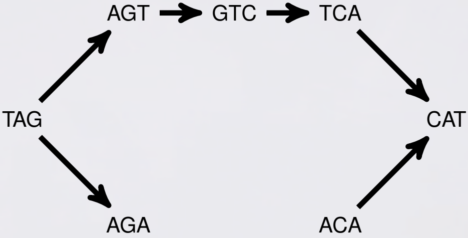
Exercice

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

Exercice: solution



- Discard TTG (connected to nothing)
- Observe a *k*-mer was missing (GAC)
- Two strains: TAGTCAT, TAGACAT