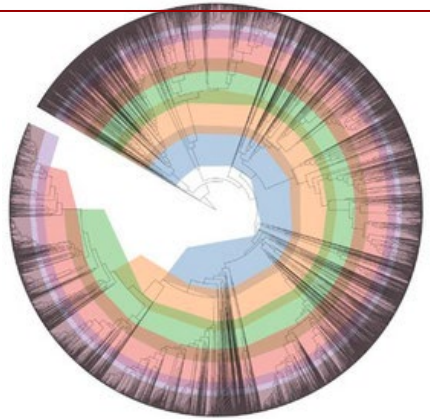
ENVIRONMENTAL METAGENOMICS

Physalia course, online, 11-15 November 2024

Long read assembly

Nikolay Oskolkov, Lund University, NBIS SciLifeLab
Samuel Aroney, Queensland University of Technology



Physalia Courses

NB: original course material courtesy:
Dr. Antti Karkman, University of Helsinki
Dr. Igor Pessi, Finnish Environment Institute (SYKE)
As. Prof. Luis Pedro Coelho

A bit about me

- Bachelor at University of Queenslar
- DPhil at Oxford University, UK
  - Internship with Zooniverse – Software development
- Post-doc at Queensland University of Technology, Australia
  - In the Woodcroft group → →
  - sandpiper.qut.edu.au

Physalia Courses

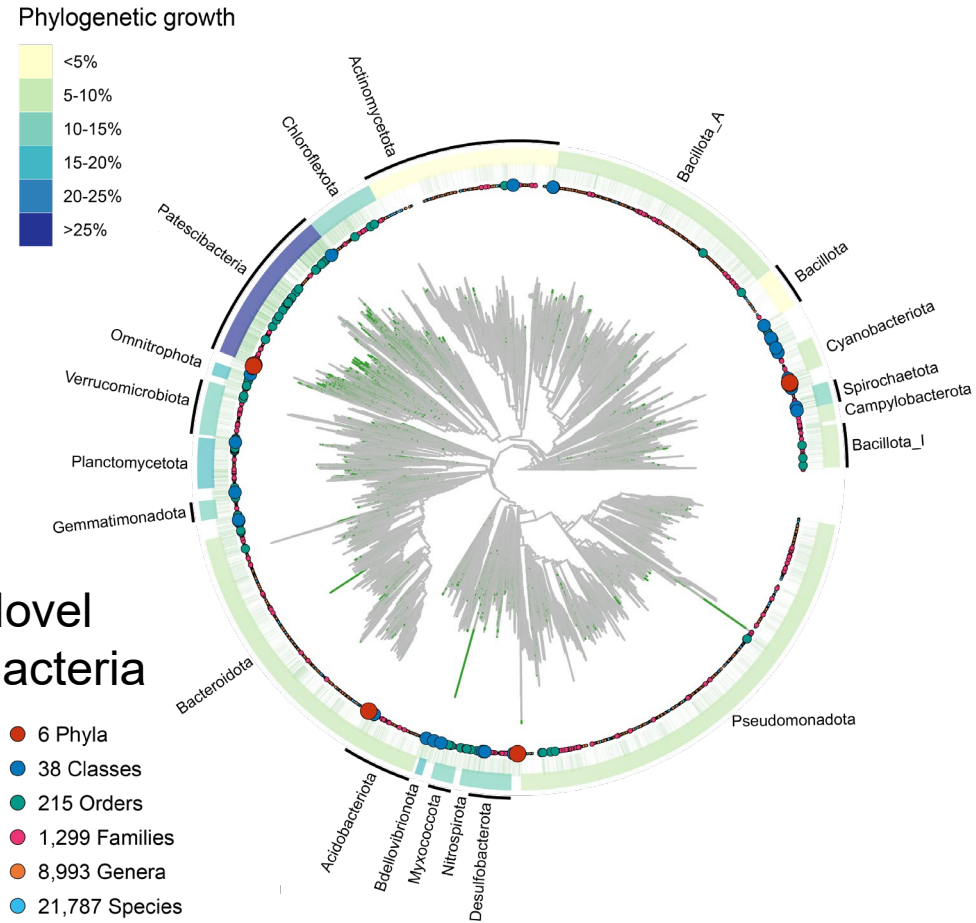
Sandpiper

A bit about my work

- Method development
- Global-scale genome recovery
- Permafrost thaw

CoverM

Aviary



Novel Bacteria

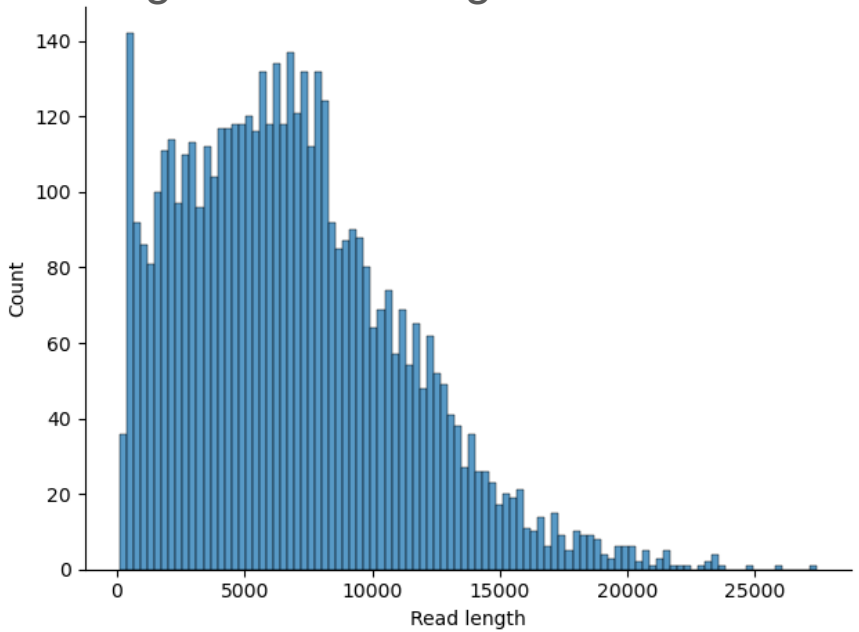
- 6 Phyla
- 38 Classes
- 215 Orders
- 1,299 Families
- 8,993 Genera
- 21,787 Species

Long reads are longer than short reads

When short reads first appeared, they were 35bps!

Nowadays, short reads are 150-300bp and long reads are longer

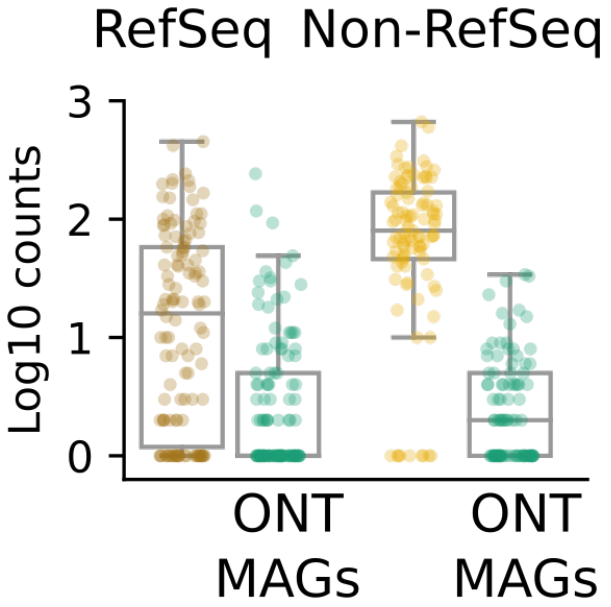
PacBio



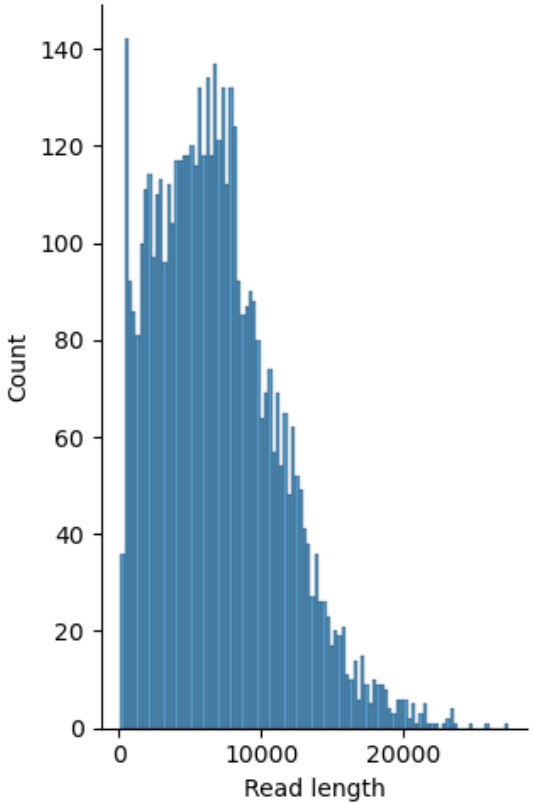
We can get better assemblies with long reads!

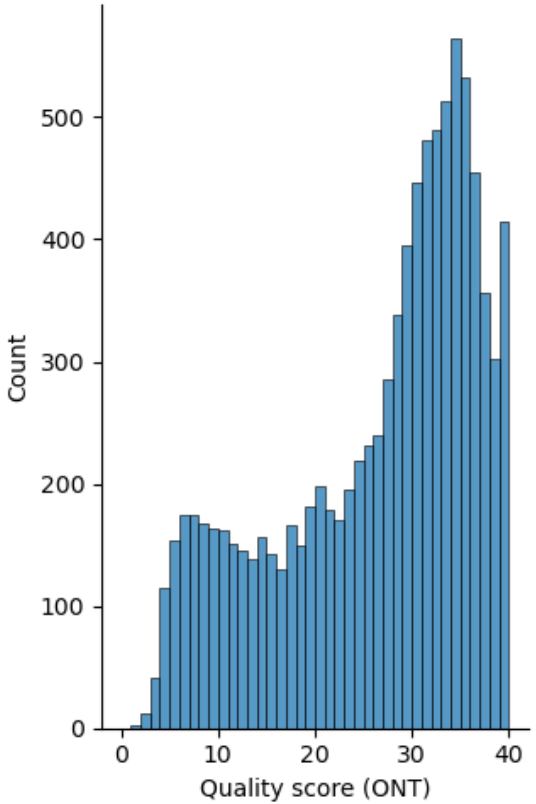
Some individual reads are longer than contigs from short-read assembly

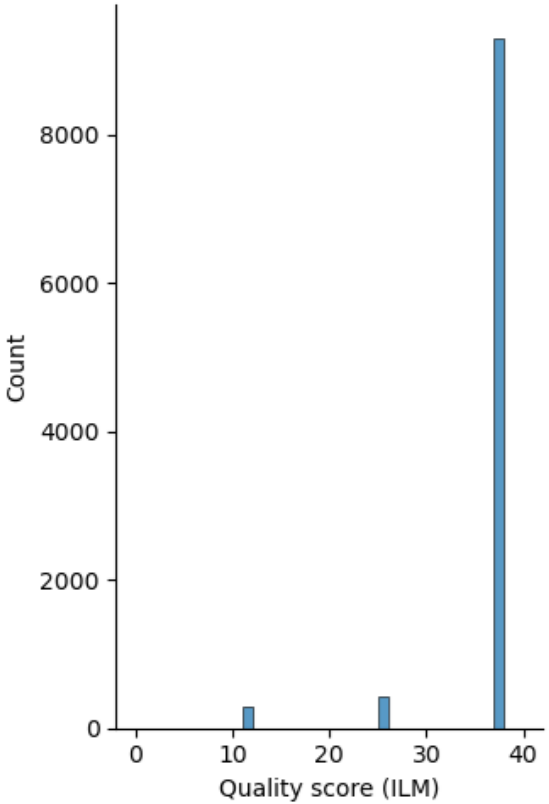
It can cover repeats and other hard to assemble regions



Why do we even use short reads then?





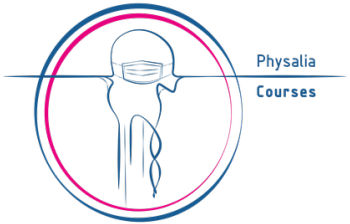


Best of both worlds: do both!

In fact, the data I showed before was from hybrid assembly!

How do you handle hybrid assembly?

1. Polish the long-reads with Illumina
2. Use a hybrid assembler (that takes ONT + ILM)
3. Polish the output of a long-read assembler with ILM
4. Use a short read assembler, then scaffold with the long reads
5. …?

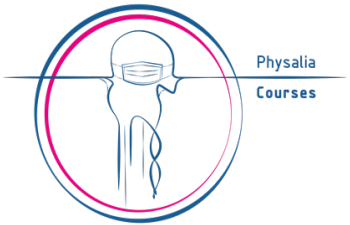


ONT data is in FASTQ format (like Illumina)

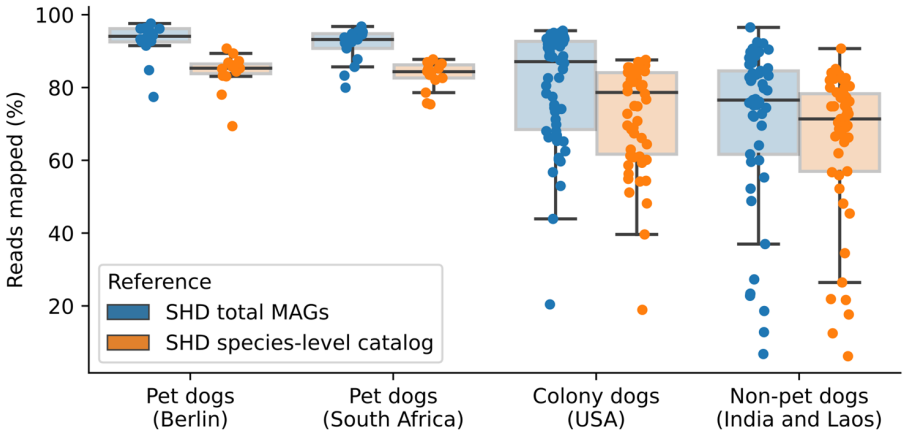
```
@abbc5672-3510-4be1-a028-488f0aa1db18
TACGTATTGCTAAGGTTAAAACGAGTCTCTTGGGACCCATAGACAGCACCTTGCGCAAGATCGATGTGAGCGGGTATGGCAGAGGTTCCGCCGGATGAACCTGGTTTCAATTCGACGATGACAGCATCAAACGTGCGATTCACGATG
+
+./:<;::>><@><??AEA7436<=F{GEA<=;=<;;<@????@AA?@BD=889:ACCA>=>@AABBD@@??BDJEEFFEBDCELFAA@@@@C;;;:DCDECCDC778?=CHHGECDDDFBAAAFAAA?=<<=?AADEF=;;:===B
@7f579801-edde-4458-90f9-9136b7a3159d
CGTATTGCTAAGGTTTAAAACGAGTTCTTGGGACCCATAGACAGCACCTCGCCATCAGGATTTTCGTCACGACCTTCCCTGGGCTCACTATCAGGGATCACCACCAAATCGGCCGCTGATGTCGCGCATCAACTGGAAACGGATGTT
+
'+*+++322;>?B@<:==;;44('*,-/----,,,,==@ADC<;;9***+BED@=<<?>;DC>>=CD3333BAEB@666>CAABECB@AAA@BBAECB@?>?---1..217666?B@EGCCC?B???D>EIFDDLFD@@@ACDDFBB
@cb88a4bb-e5b0-47e1-adee-9238348a1f04
GCTTATGTAACCTATTCGTTCAGTTACGTATTGCTAAGGTTAAAACGAGTCTCTTTTGGGACCCATAGACAGCACCTGCCCTTGCCGGGATGCTCGTCCTTCAGATCGTGAACGGCCGCAACGACATCTCGTCGGAACCAGATATAG
+
#####$$$#"""""#%&'.88CA?=76@@AADBA:;:;@BDEG@@@>----138{547356@9:>6:102>AHGCA@@EFLEGGFGJCBBCEDBC@+***7.())1&&BABJILEC@<;;=FH66636571))))*5A=<4335/.))
@2fae7030-0ca6-439a-8ea8-895dc050ea35
AAAACGAGTCTCTTGGGTCCCATAGACAGCACCTTGCGATAGGGCGCTGCAGCAATGGTTTTGCTTTCTGCGCGATGGCCAGCGTTGTAGCGCGGTCATCGTTTTCCACCGCGAACGGAATAATCTTCACGGTCGCGATCATCTCGC
+
.74*+,=;=>?<2-,**+,,+('&'**+.38755+*)*..10-'%%&'.79:?CBA<DLC>?@DD?@?@1201=DDCCCFFDCCB@?=<={>7ABFG@@?CIPE==>=>@,,,?;:0//.4::>BKDA?>?BB?<<;:9885567::
@a6b78936-61be-41ad-a802-9321d04e69bd
TACGTATTGCTAAGGTTAAAACGAGTCTCTTGGGACCCATAGACAGCACCTTCTCCCCGCGACGGCCGCGAACTGAGCGCCTCCTTCGCCCGCTTTCCGGAGAGCTTCGGGCACCCGTGGCAATCGGTCGTGCTGACCCCCACCGAT
+
'++53222<;3:756@@L<:99=?C@H?<9;;;=<<<<?>@ABCD@AACJKKDD96::AA?@@B?==;<ABDDB@>>?@F?@AF@@>=:;:=LNLBAAAQFCDDB8>?>DCDICCBABCDFBA@?A@=<;<<<<=ACBAB>89456<
@a0f65575-ca1b-4c9b-8c5d-fbaadc2bd609
ATGTATTGCTAAGGTTAAAACGAGTCTCTTGTACTGATAGACAGCACCTGACTTGGGTTGTCGTGTGTGCGCATGCGCATGGTGCCCTGCTTGATCGCCTTTTCCAGCGTCCTGCTGGCGCGCTCGTAGACGTCCTCTTCGGTGAAG
+
((@CA>>@@9=;;>CFJ>===CEDOH@644...../77:>BBBAA?@@BA@@@?@CCED=<<@{{L<888<?::99=>;:99===<>=??>C?A?>?AJKLHLEC>=>=@AH??><=B@::BBDDGR?==<@AACFH>=181--2''
@60a4d5ec-e51b-4aa5-b9f6-618e9c03293c
ACGTATTGCTAAGGTTAAAACGAGTCTCTTGGGACCCATAGACAGCACCTGAACAGCTTTCTGCTTGACCCGTTCTCTGAGCCGATAGAGATCAACGAACACCGAGACCTTATTTTTCAAAATTTCCGGCACGACCGGTTTCACCAG
+
+-95433>?:A;;AEHN>:::@ABFGGF?>?=;;;;>?>@>>>>=>>>DFBC<<==?RGHDAAAEEB@:61232,+)'&''((,,,-/,,-.+*))7=>?EDBCC=<@=EGCNLQPFLRLLLXJABACHAA@???AELNRJHF=;;=
@348e43b0-a765-4c7b-af2f-6b92b03e6116
CGTATTGCTCAAGGTTAACAAAACGAGTCTTGGGACCCGATACACAGCACCTATTATGCCGGTTACCGAGAATCCGATTCACGTGCGCGGACCTATTTCCGACCGCGAACAGAAAATGATGGCACTGGGCTACCTGCACGAAGCCTG
+
$%(,))$%%&/,11-%%%*0/-,-020.,-*,--0(&&*((++78:;@<;<=>>=A><<=@ACA3203/.1-(''(,+&'''()+(%$$$)---223667;<=>?-**+,,,5/...2;<LB=::9<-=A@AE{IE@66?@GCGHFG
@c90e871f-5912-4415-a705-c212fc55f2df
TCAAATATCGCTAAAGGTTAAAACTAGTTCTTGTACTCATAGACAGCACCTGTTCGTACCGGAAAAGCGTGAAACCAAAGGCAATCGCACGGTCCAGTATCTCCGGCGCGGAGGACGTATCGGCGATCTGCTGAGGATTCTGCTTCT
+
$&()*))((**)-767764/0&''(3*('(&)*(&&*,::;=>==>=///05446>B@;+.29A??BC;;>?)&&&2DABB=-,,-C?99:;;;21149;A:8-,,0>?@@B??AACF@??>BCEFFDKPHEDCCEBFEEIKIJ==>
```

How do results look like with long-reads?

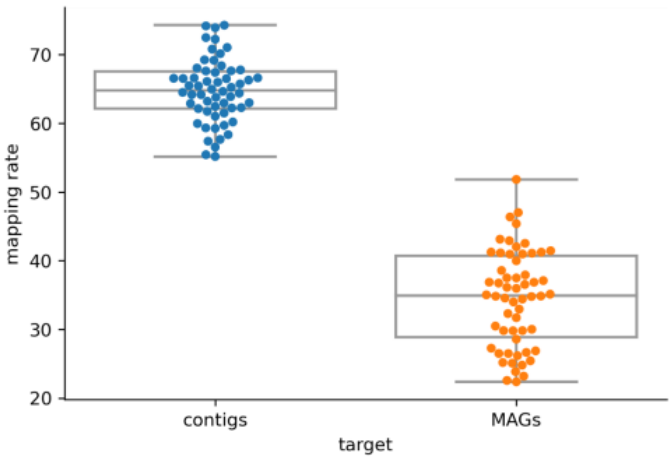
		#contig	Average length	N50
Sample A	mNGS	277,884	1,865	8,778
	PacBio-HiFi	2,575	99,842	269,406
	Nanopore	3,414	65,685	199,639
Sample B	mNGS	146,857	2,402	28,310
	PacBio-HiFi	1,447	157,228	1,081,788
	Nanopore	1,795	123,512	658,841
Sample C	mNGS	170,813	1,956	18,485
	PacBio-HiFi	822	171,215	1,270,126
	Nanopore	1,370	126,533	891,411



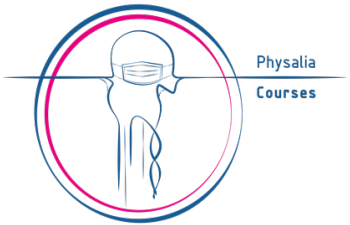
Results depend on the habitat!



Dog gut
(20 Gbp + 20 Gbp) x 50 samples



Soil
(40 Gbp + 40 Gbp) x 52 samples

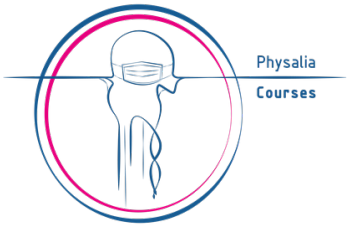


In dog guts, 20 Gbp was enough to get excellent representation of the community, but in soil even double that was not enough! — could also be Euks!

From Anna Cuscó (dog) & Yiqian Duan (soil)

A few other notes on long-reads

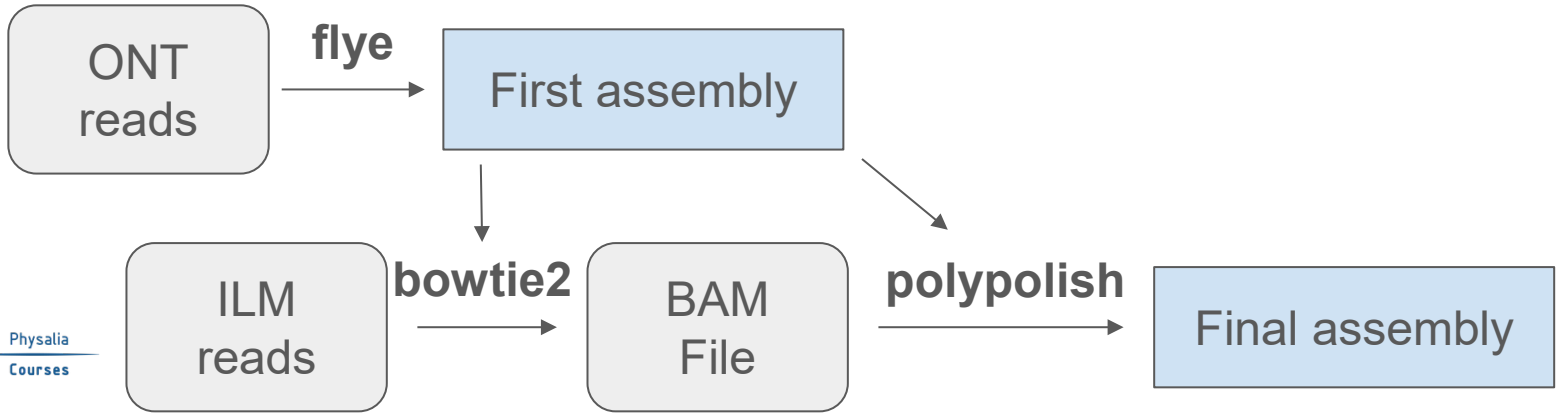
- Still a very rapidly evolving field (both in the wetlab & in the drylab)
- Differences between ONT versions matter a lot
- We might be able to get even more information in the future
  - Methylation
  - Non-canonical nucleotides
- **Long-reads are the future**



Exercises overview

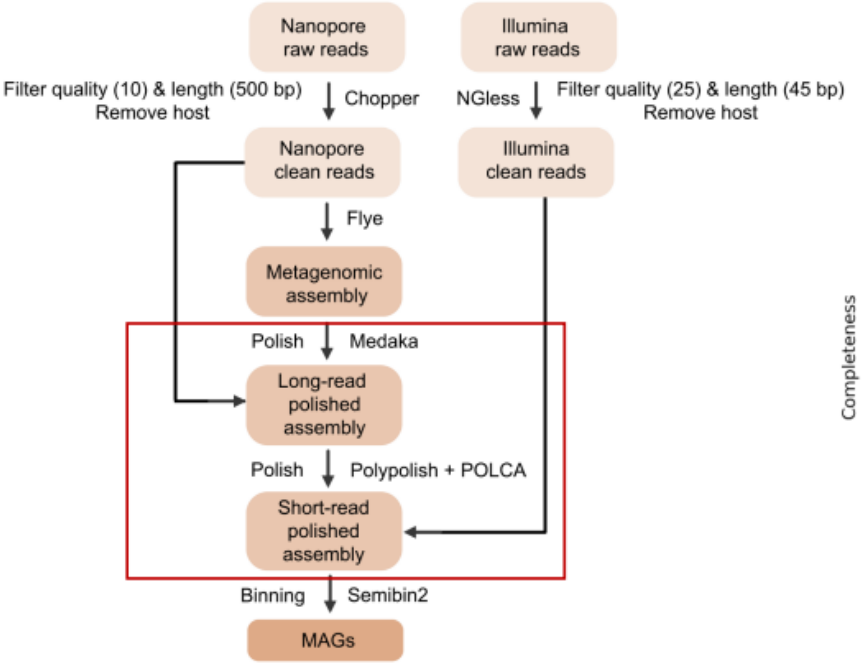
1. Assemble with **flye** (long-read only)
2. Polish with **polypolish** (using the short reads)

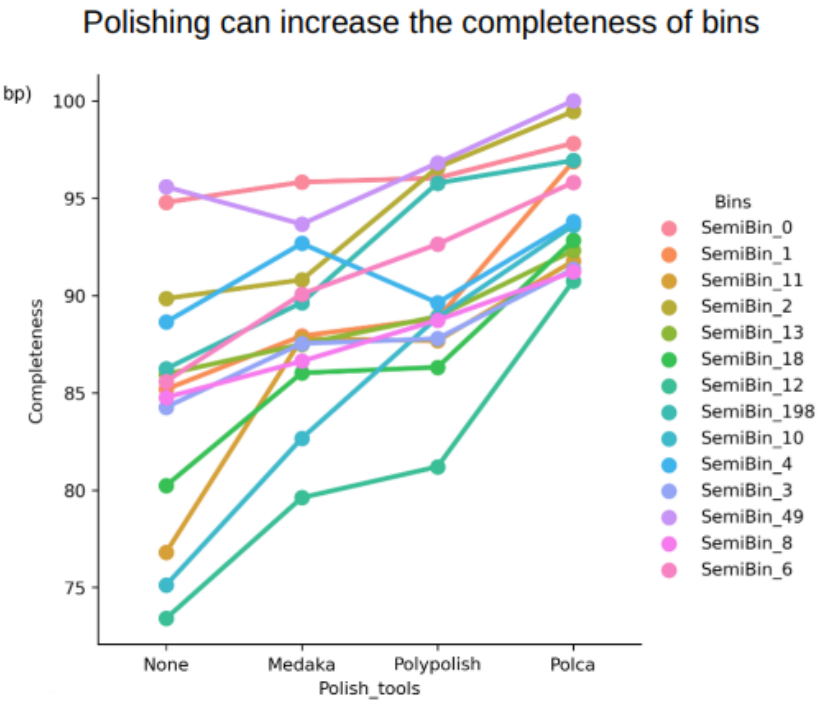
In order to polish, we need to align the short reads to the contigs from the **flye** assembly



Physalia Courses

Real workflow





From Yiqian Duan