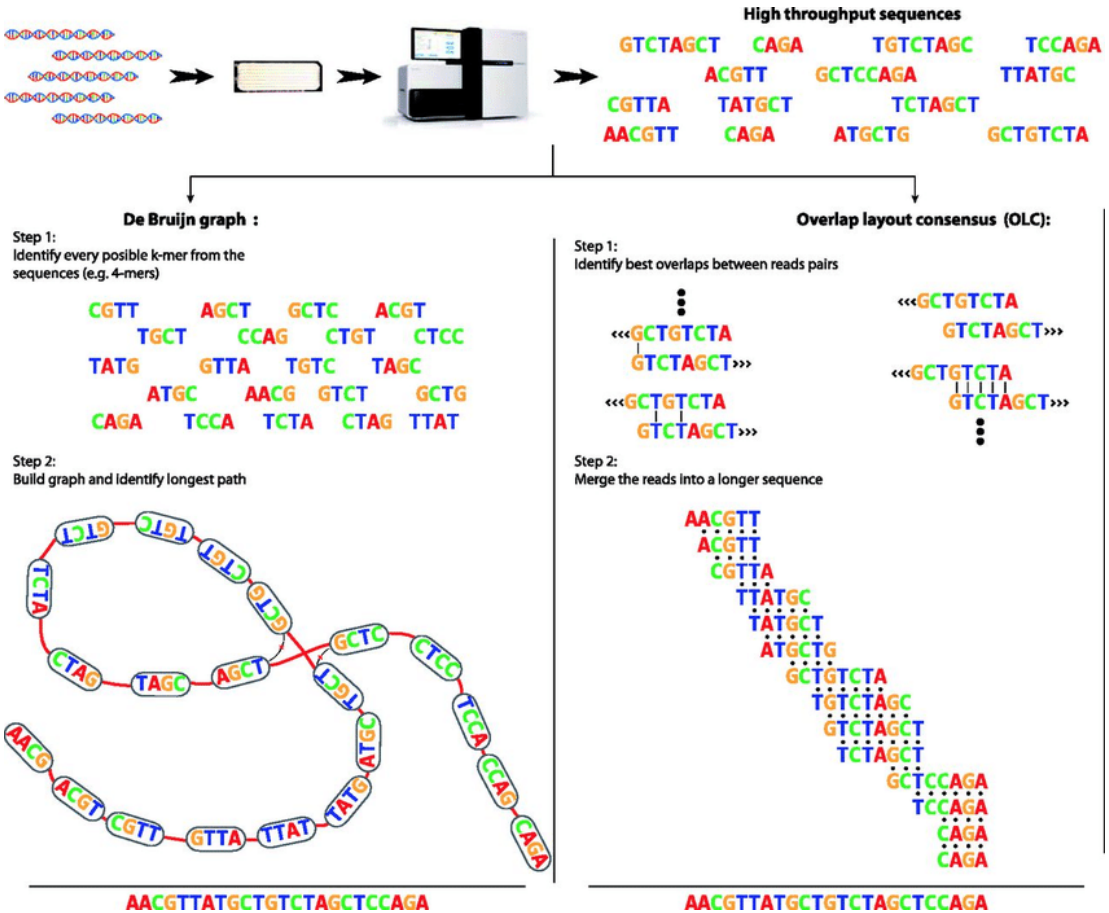
Genome assembly with 2nd and 3rd WGS data

Yan Hui
Postdoc
Department of Food Science
E-mail: huiyan@food.ku.dk

UNIVERSITY OF COPENHAGEN

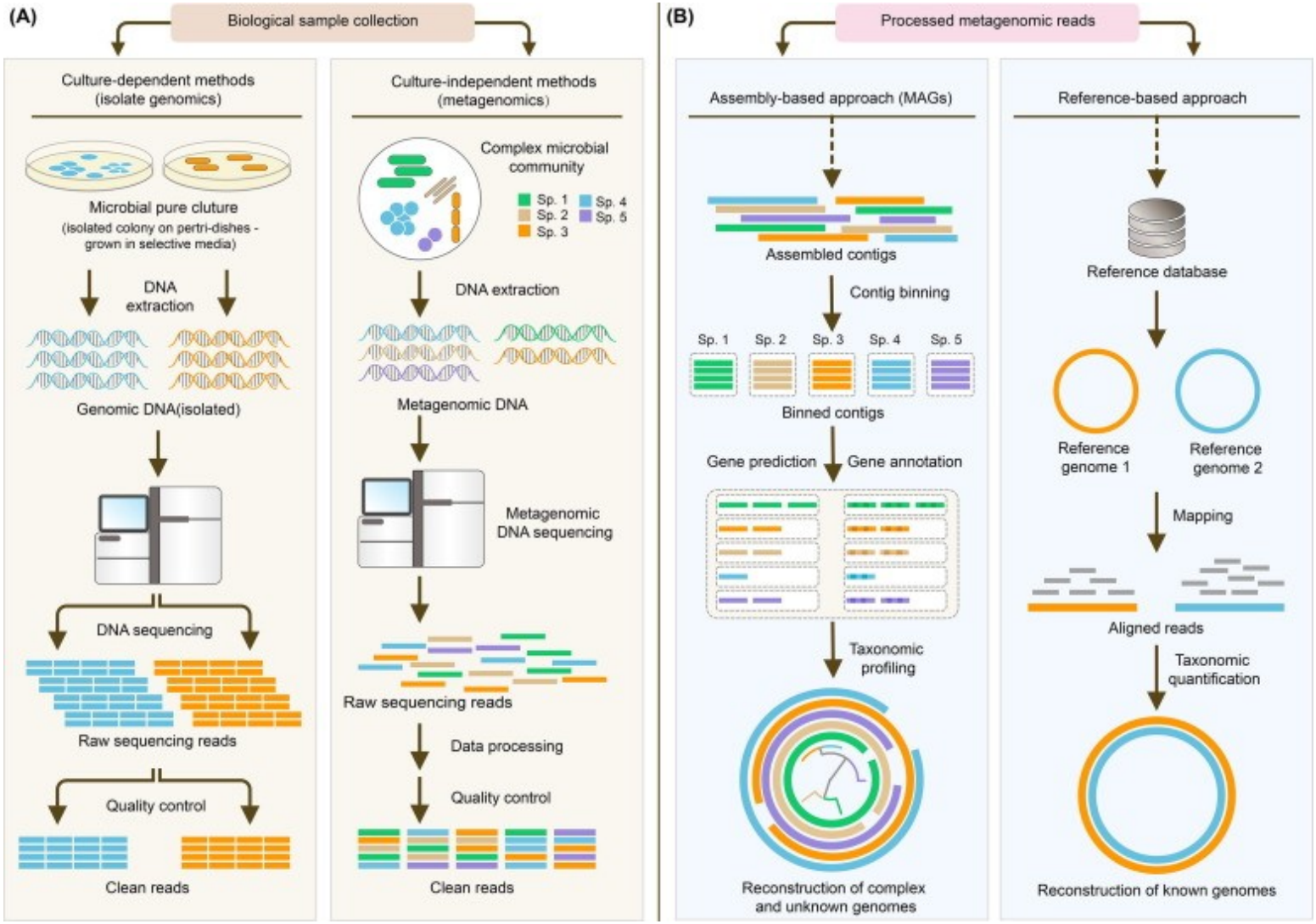
Challenges in assembly

- 2nd sequencing: short reads > long fragment
- 3rd sequencing: errors in reads



https://academic.oup.com/ve/article/2/2/vew022/2797616

WGS & Metagenomics

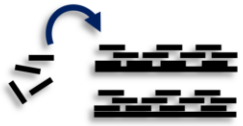


https://pubmed.ncbi.nlm.nih.gov/34900140/

Assembly-based metagenomics

1 Quality Control
- PCR duplicates removal
- Quality trimming
- Host removal
- Common contaminant removal

➢ **QC reads**

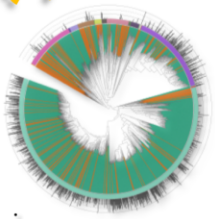


2 Assembly
- Error correction
- Paired-end merging
- Assembly (metaSpades/megahit)
- Post-filtering

➢ **High-quality Scaffolds**

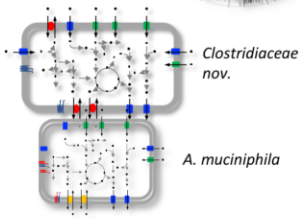
3 Genomic Binning
- Binning (metabat, maxbin2)
- Quality Assessment (checkM)
- Bin refining (DAS Tool)
- Dereplication (dRep)
- Quantification
- Robust taxonomic classification (GTDB)

➢ **Genomes**
➢ **Abundances**



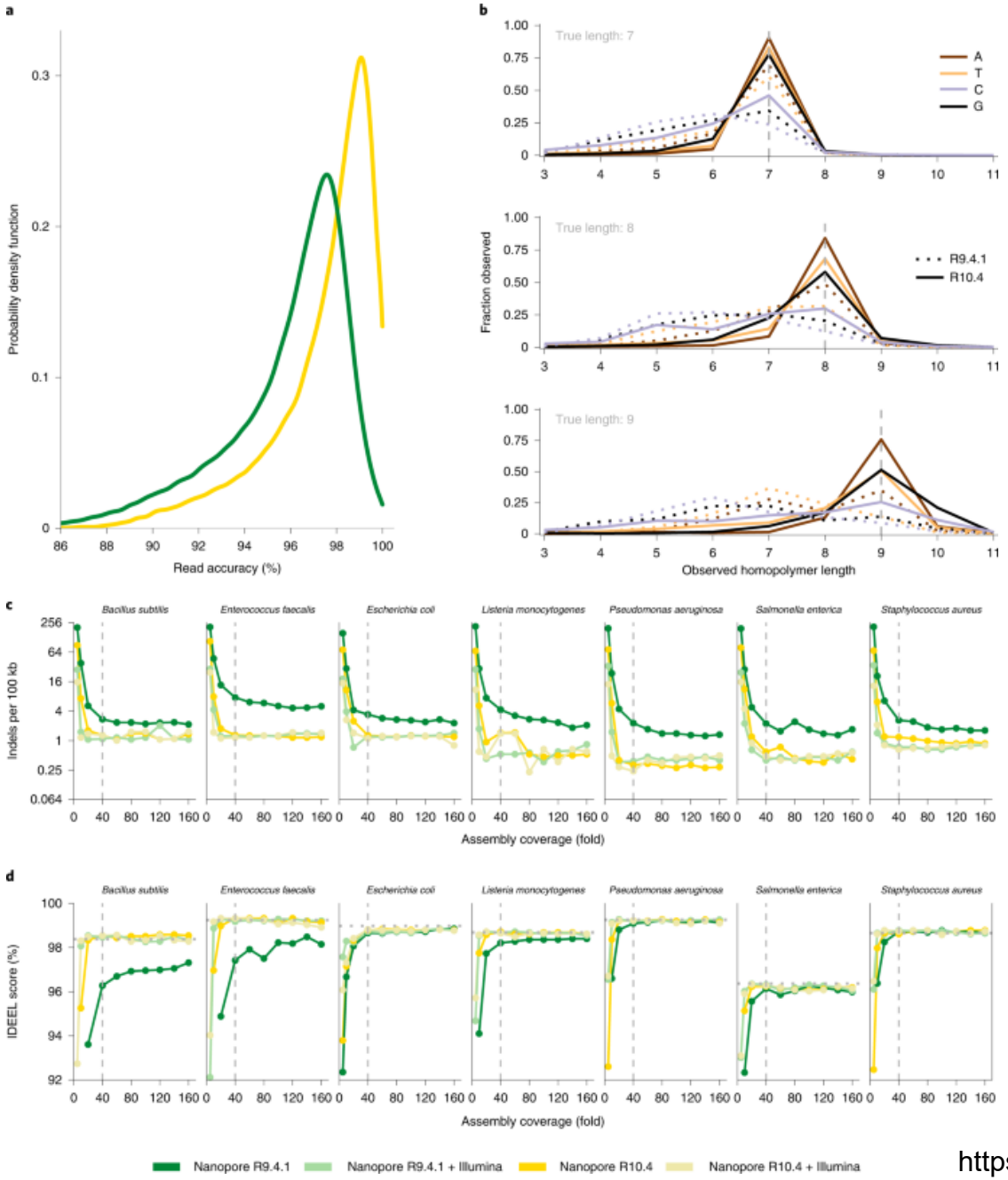
4 Annotation
- Gene prediction (prodigal)
- Cluster redundant genes (linclust)
- Annotation (eggNOG)

➢ **Functional annotations**



Workflows

metaWRAP: https://github.com/bxlab/metaWRAP
Metagenome-Atlas: https://github.com/metagenome-atlas/atlas

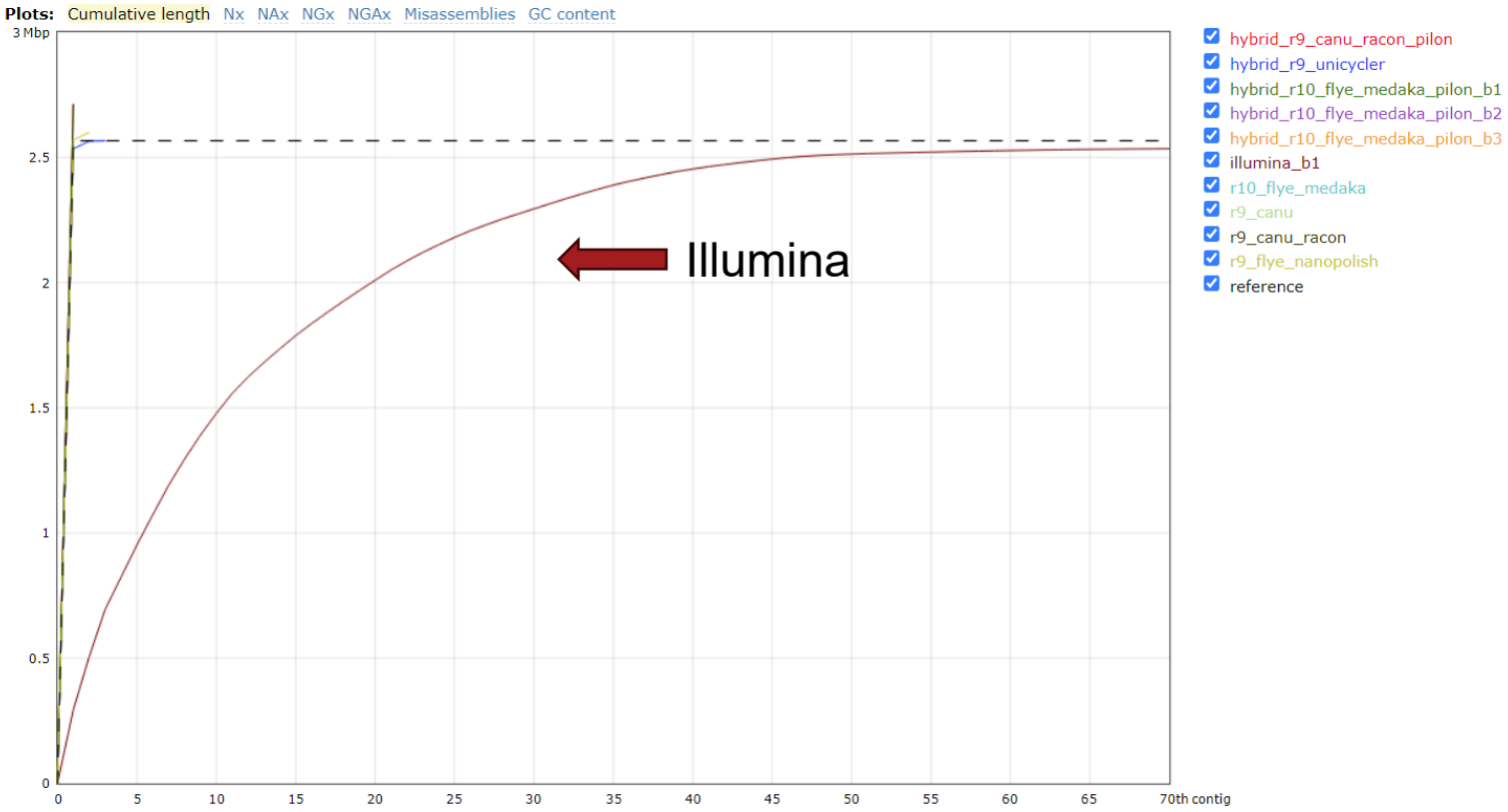


ONT R10.4 improvements

Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing

https://www.nature.com/articles/s41592-022-01539-7

A *Propionibacterium freudenreichii* genome assembled with 2nd and 3rd WGS data



Contigs are ordered from largest (contig #1) to smallest.

A *Propionibacterium freudenreichii* genome assembled with 2nd and 3rd WGS data

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Aligned to "ncbi_pacbio_TL110" | 2 566 312 bp | 1 fragment | 67.33 % G+C

Worst Median Best ☑ Show heatmap

Fragmented assemblies

Genome statistics	hybrid_r9_canu_racon_pilon	hybrid_r9_unicycler	hybrid_r10_flye_medaka_pilon_...	hybrid_r10_flye_medaka_pilon_...	hybrid_r10_flye_medaka_pilon_...	illumina_b1	r10_flye_medaka	r9_canu	r9_canu_racon	r9_flye_nanopolis
Genome fraction (%)	100	99.946	99.968	100	100	98.591	100	100	100	100
Duplication ratio	1.058	1	1	1	1	1.001	1	1.051	1.056	1.013
Largest alignment	2 566 818	2 534 145	2 565 436	2 566 281	2 566 279	293 988	2 565 347	2 549 265	2 562 649	2 569 292
Total aligned length	2 714 629	2 566 051	2 565 436	2 566 281	2 566 279	2 532 212	2 565 347	2 696 060	2 710 312	2 598 585
NGA50	2 566 818	2 534 145	2 565 436	2 566 281	2 566 279	104 535	2 565 347	2 549 265	2 562 649	2 569 292
LGA50	1	1	1	1	1	8	1	1	1	1
Misassemblies										
# misassemblies	2	1	0	0	0	2	0	2	2	1
Misassembled contigs length	2 714 629	30 625	0	0	0	58 625	0	2 696 061	2 710 312	30 664
Mismatches										
# mismatches per 100 kbp	21.81	0.9	0.66	0.66	0.7	0.2	11.73	93.8	195.59	162.47
# indels per 100 kbp	42.11	2.46	2.07	1.83	2.07	0.36	40.77	403.59	278.68	200.49
# N's per 100 kbp	0	0	0	0	0	0	0	0	0	0
Statistics without reference										
# contigs	1	3	1	1	1	70	1	1	1	2
Largest contig	2 714 629	2 534 145	2 565 436	2 566 281	2 566 279	293 988	2 565 347	2 696 061	2 710 312	2 569 292
Total length	2 714 629	2 566 051	2 565 436	2 566 281	2 566 279	2 534 467	2 565 347	2 696 061	2 710 312	2 599 956
Total length (>= 1000 bp)	2 714 629	2 566 051	2 565 436	2 566 281	2 566 279	2 528 342	2 565 347	2 696 061	2 710 312	2 599 956
Total length (>= 10000 bp)	2 714 629	2 564 770	2 565 436	2 566 281	2 566 279	2 453 508	2 565 347	2 696 061	2 710 312	2 599 956
Total length (>= 50000 bp)	2 714 629	2 534 145	2 565 436	2 566 281	2 566 279	1 788 339	2 565 347	2 696 061	2 710 312	2 569 292

Mismatches										
# mismatches per 100 kbp	21.81	0.9	0.66	0.66	0.7	0.2	11.73	93.8	195.59	162.47
# mismatches	592	23	17	17	18	5	301	2529	5301	4222
# indels per 100 kbp	42.11	2.46	2.07	1.83	2.07	0.36	40.77	403.59	278.68	200.49
# indels	1143	63	53	47	53	9	1046	10 881	7553	5210
# indels (<= 5 bp)	1117	61	52	46	52	7	1046	10 621	7477	5150
# indels (> 5 bp)	26	2	1	1	1	2	0	260	76	60
Indels length	1772	211	106	103	105	111	1189	18 495	10 583	7167
# N's per 100 kbp	0	0	0	0	0	0	0	0	0	0
# N's	0	0	0	0	0	0	0	0	0	0

ONT: Indels > Mismatches

ONT R9 mismatches

Exercises

UNIVERSITY OF COPENHAGEN

Exercise

- MAC2023: https://yanhui09.github.io/MAC2023/

Cross-platform support, incl. MacOS

Linux/amd64 platform

Thanks