Getting complete genomes from complex samples using nanopore sequencing

Rasmus H. Kirkegaard, Søren M. Karst, and Mads Albertsen

Albertsen Lab, Center for Microbial Communities, Aalborg University

Introduction

Most of the DNA sequencing data today is produced with short read sequencing. However, it is unable to resolve repeat structure even in pure culture genomes and often repetitive elements cannot be linked to their respective genomes in metagenome data. Long read nanopore sequencing has the potential to close the gaps and produce circular genome assemblies from complex systems.

Aim

To investigate if long read DNA sequencing can fix strain and repeat assembly problems

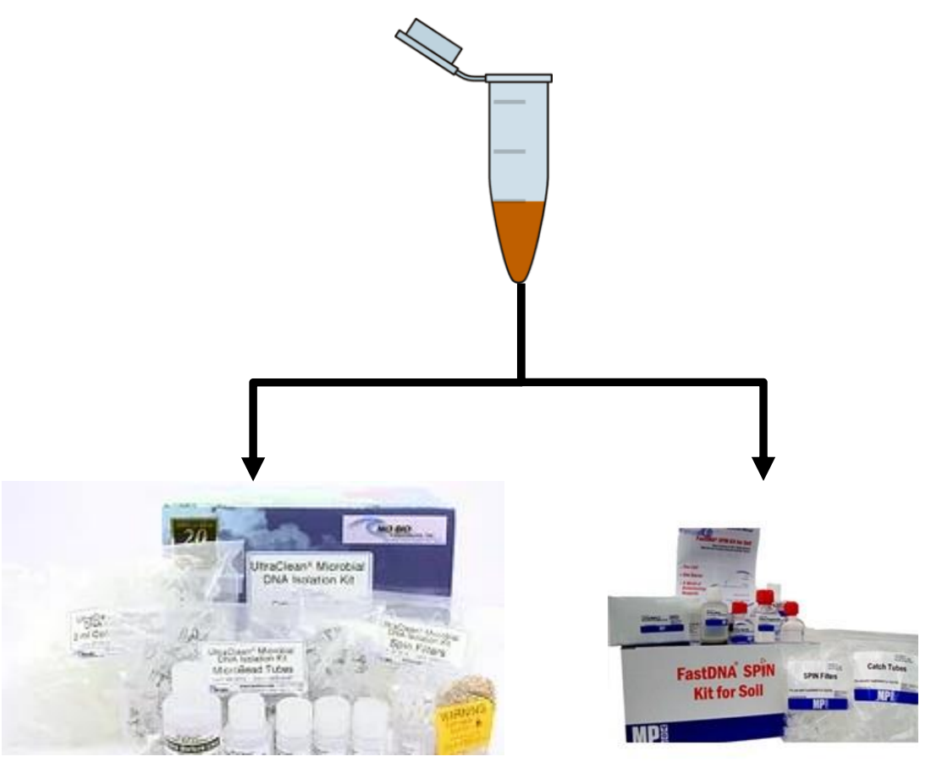
Methods

sampling

- Sludge samples from a full-scale anaerobic digester in Fredericia

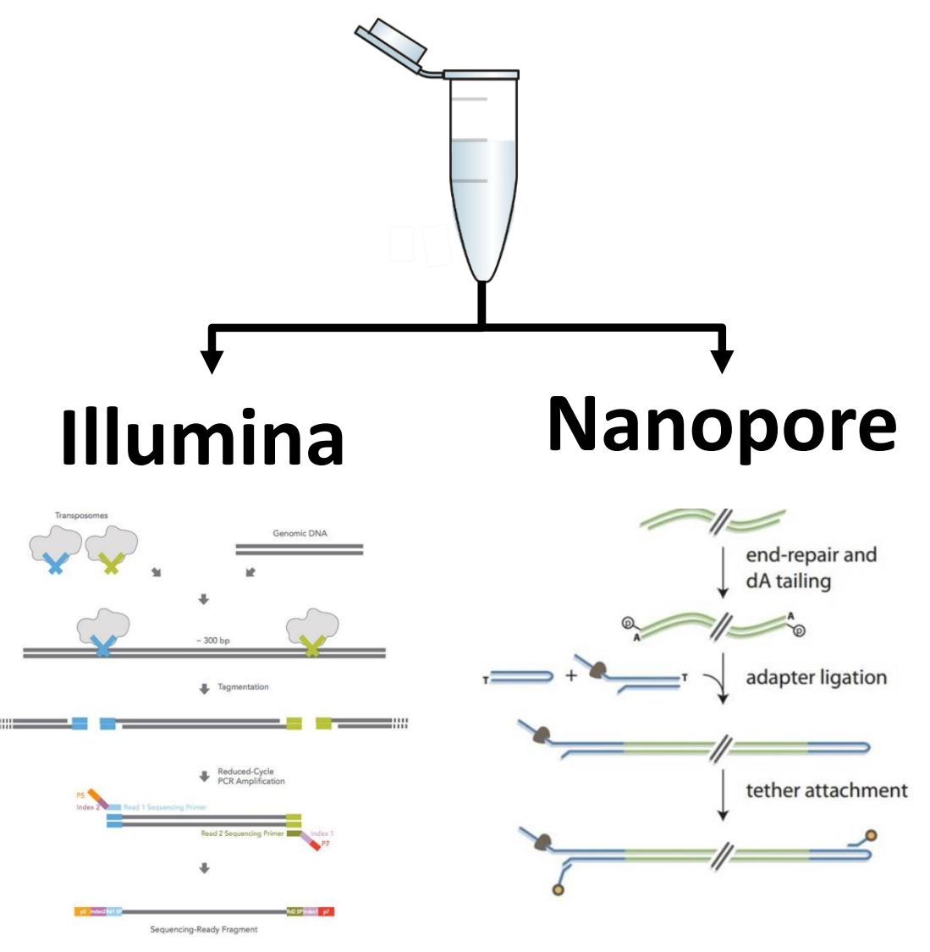
extraction

- FastDNA SPIN kit for soil
- PowerMicrobial Maxi DNA Isolation kit



preparation

- Libraries prepared for illumina short read DNA sequencing
- Library prepared for nanopore long read DNA sequencing

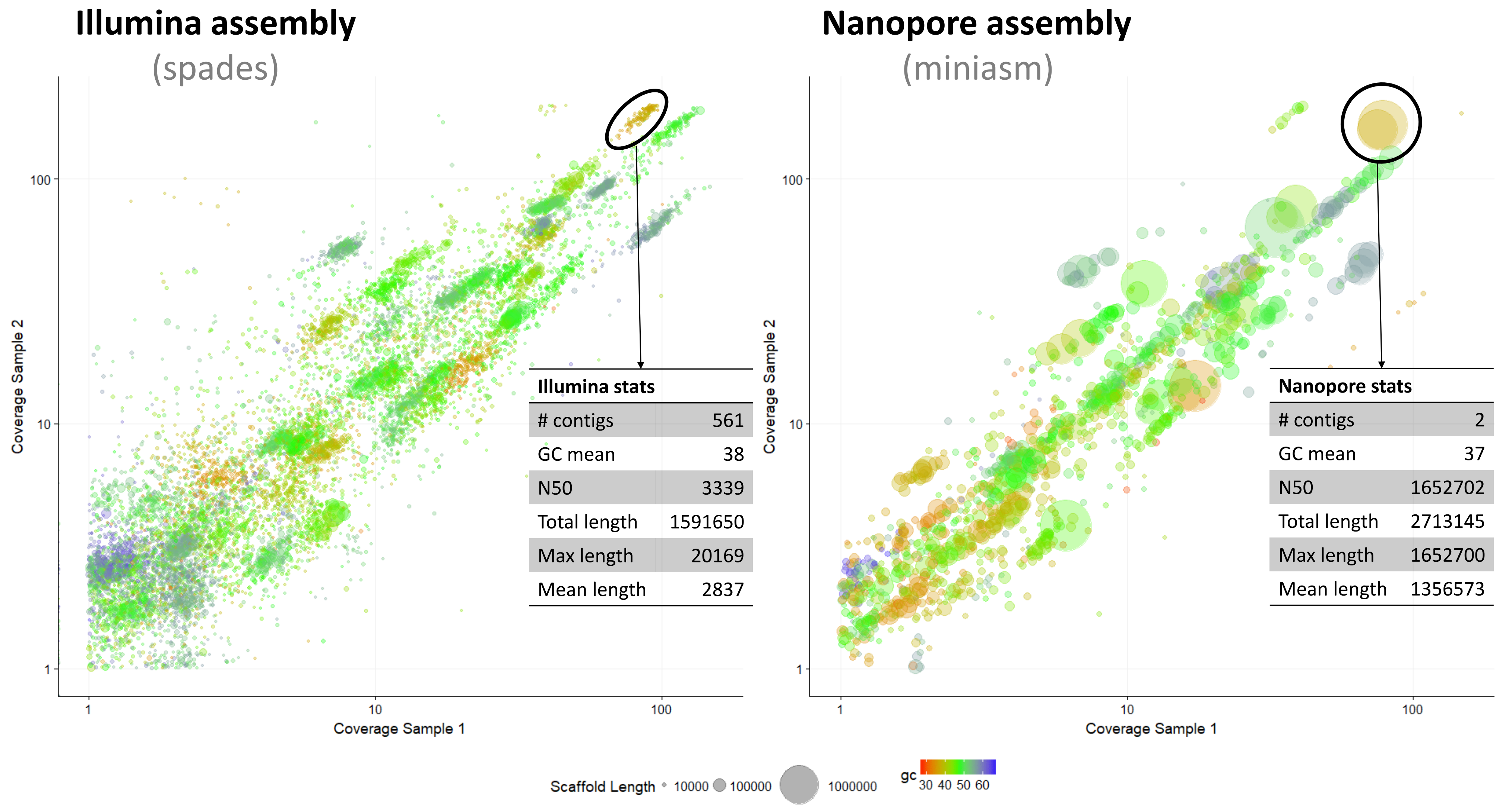


sequence

Conclusions

- Long read assemblies are much more contiguous than short read assemblies as the long reads can span the repetitive elements
- Accurate short reads are still needed for polishing indel errors in long read assemblies to allow gene calling

Results



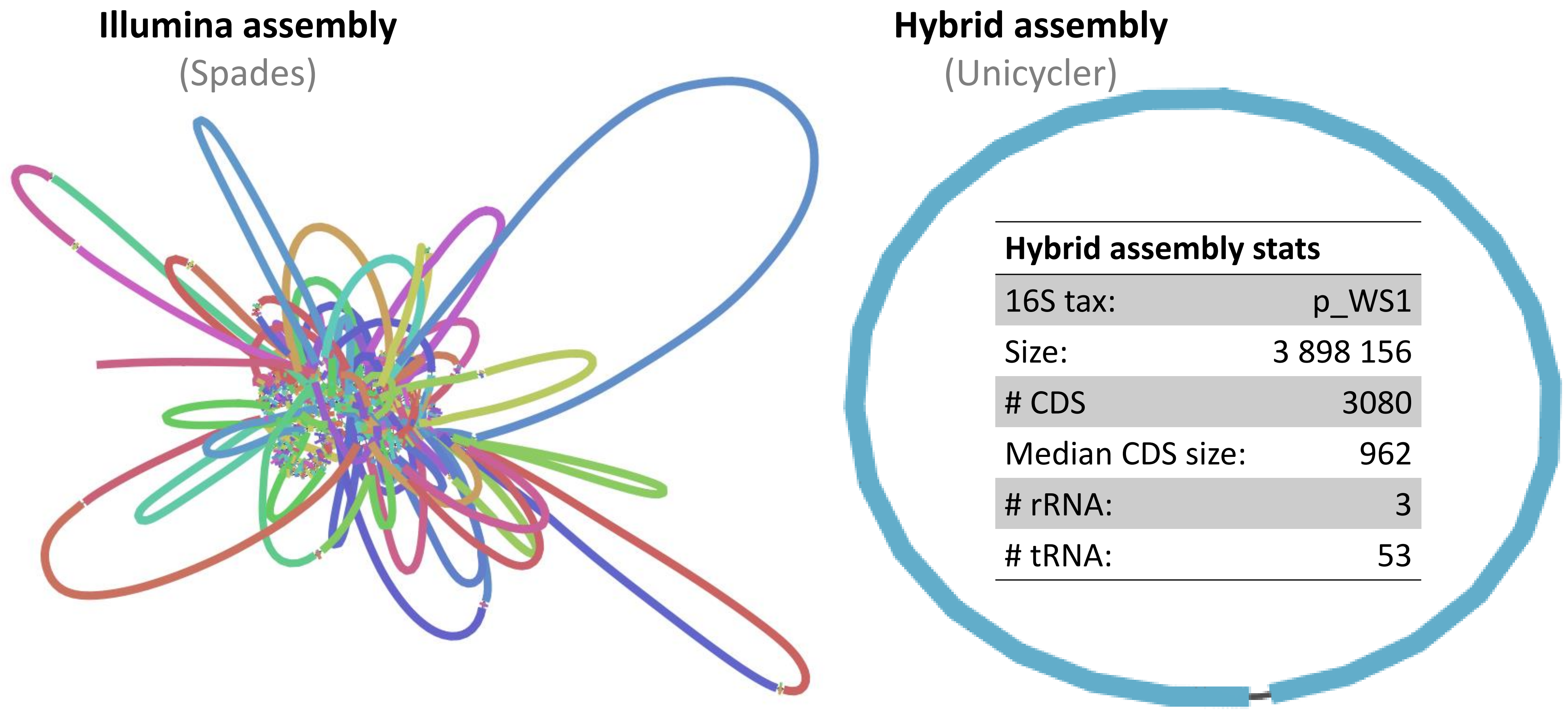
Illumina assembly
(spades)

Illumina stats	
# contigs	561
GC mean	38
N50	3339
Total length	1591650
Max length	20169
Mean length	2837

Nanopore assembly
(miniasm)

Nanopore stats	
# contigs	2
GC mean	37
N50	1652702
Total length	2713145
Max length	1652700
Mean length	1356573

Short reads vs. long reads metagenome assemblies for differential coverage binning. The short read assembly produces much smaller contigs than the nanopore based assembly (ovals). Visualised with the mmgenome package.



Illumina assembly
(Spades)

Hybrid assembly
(Unicycler)

Hybrid assembly stats	
16S tax:	p_WS1
Size:	3 898 156
# CDS	3080
Median CDS size:	962
# rRNA:	3
# tRNA:	53

Short reads vs hybrid approach reassembly. Assembly graphs visualised with Bandage. The short read assembly is unable to resolve the repeat structure within the genome whereas the hybrid approach can create finished level genomes as circular assemblies.

Genome assembly stats for multiple E. Coli assemblies. Long read assemblies are much more contiguous than short read assemblies. However, the inherent indel errors within the long read data necessitates a hybrid approach using the strengths of both data types to produce high quality genomes.

	Type	Contigs	Size (bp)	Rel. Size	ANI (%)	CheckM %	# CDS	rRNA	Median CDS size	MM 100kb	Indels 100kb
Spades	Short	84	4546220	0.98	100.00	99.9	4235	11	810	0.4	0.1
Spades-hybrid	Hybrid	1	4620377	1.00	100.00	99.9	4282	22	815	5.9	0.4
Miniasm	Long	1	4410101	0.95	83.95	0.0	1124	16	206	NA	NA
Miniasm+1xRacon	Long	1	4619818	1.00	99.01	70.9	10192	22	287	239.9	541.4
Miniasm+2xRacon	Long	1	4622021	1.00	99.23	75.0	9626	22	308	222.5	450.0
Miniasm+2xRacon+Pilon	Hybrid	1	4622021	1.00	99.96	98.5	4509	22	761	10.3	19.2
CANU	Long	1	4533574	0.97	98.69	50.0	10478	21	263	113.5	662.9
CANU+Nanopolish	Long	1	4563152	0.98	99.27	71.8	9664	22	296	146.7	468.0
CANU+Nanopolish+Pilon	Hybrid	1	4567411	0.98	99.97	98.4	4415	22	770	6.1	15.8
Unicycler	Hybrid	1	4633976	1.00	100.00	99.9	4305	22	815	4.3	2.3
Reference U00096.2		1	4639675	1.00	100.00	100.0	4300	22	818	0.0	0.0

rhk@bio.aau.dk

www.AlbertsenLab.org

kirk3gaard