SHOTGUN METAGENOMICS WORKFLOW USING OXFORD NANOPORE PLATFORM

Presenter: PhD Student Quynh Nhu Nguyen, Heidelberg University Hospital, Germany

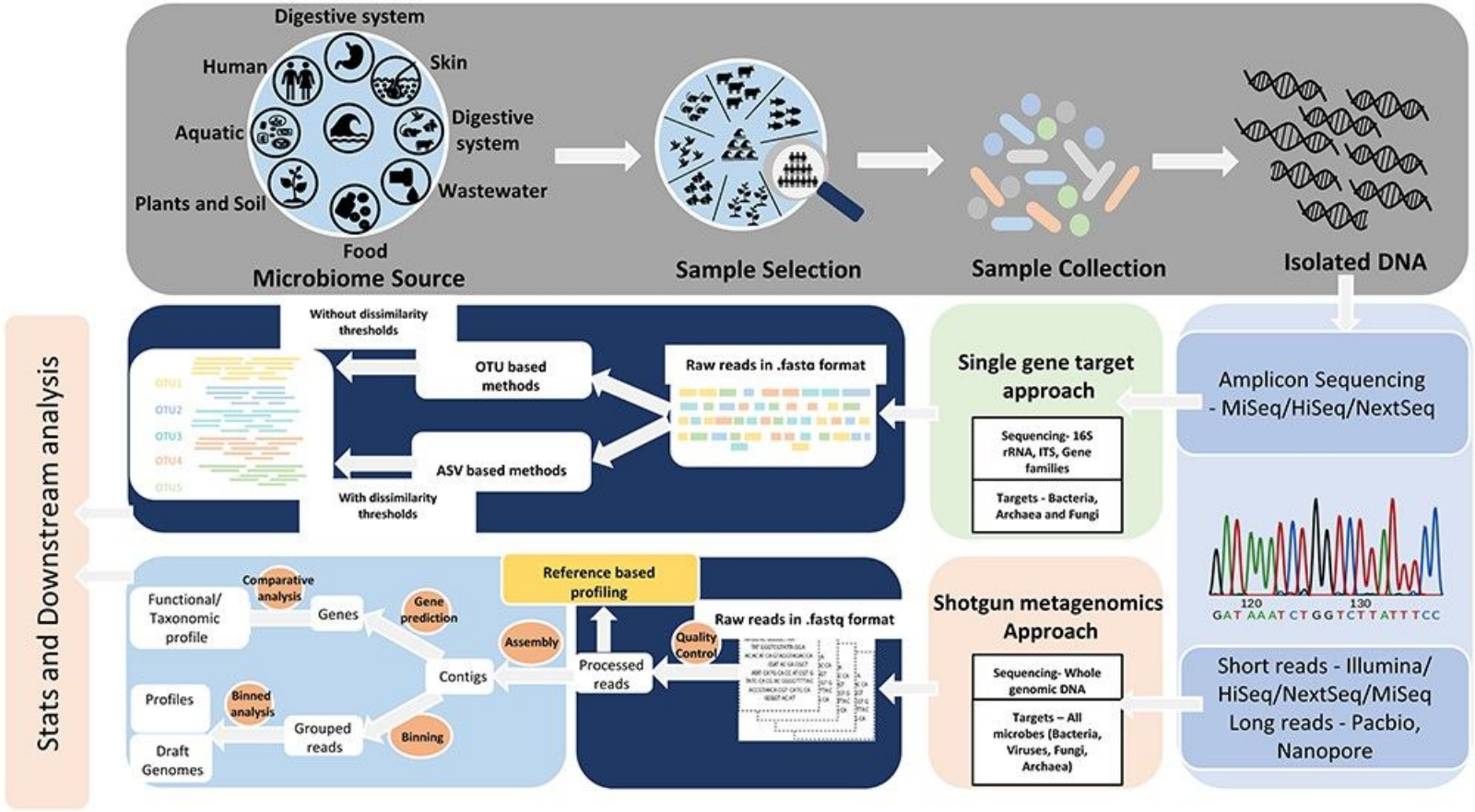
Phu Quy Ho, Can Tho University, Vietnam

13.10.2024

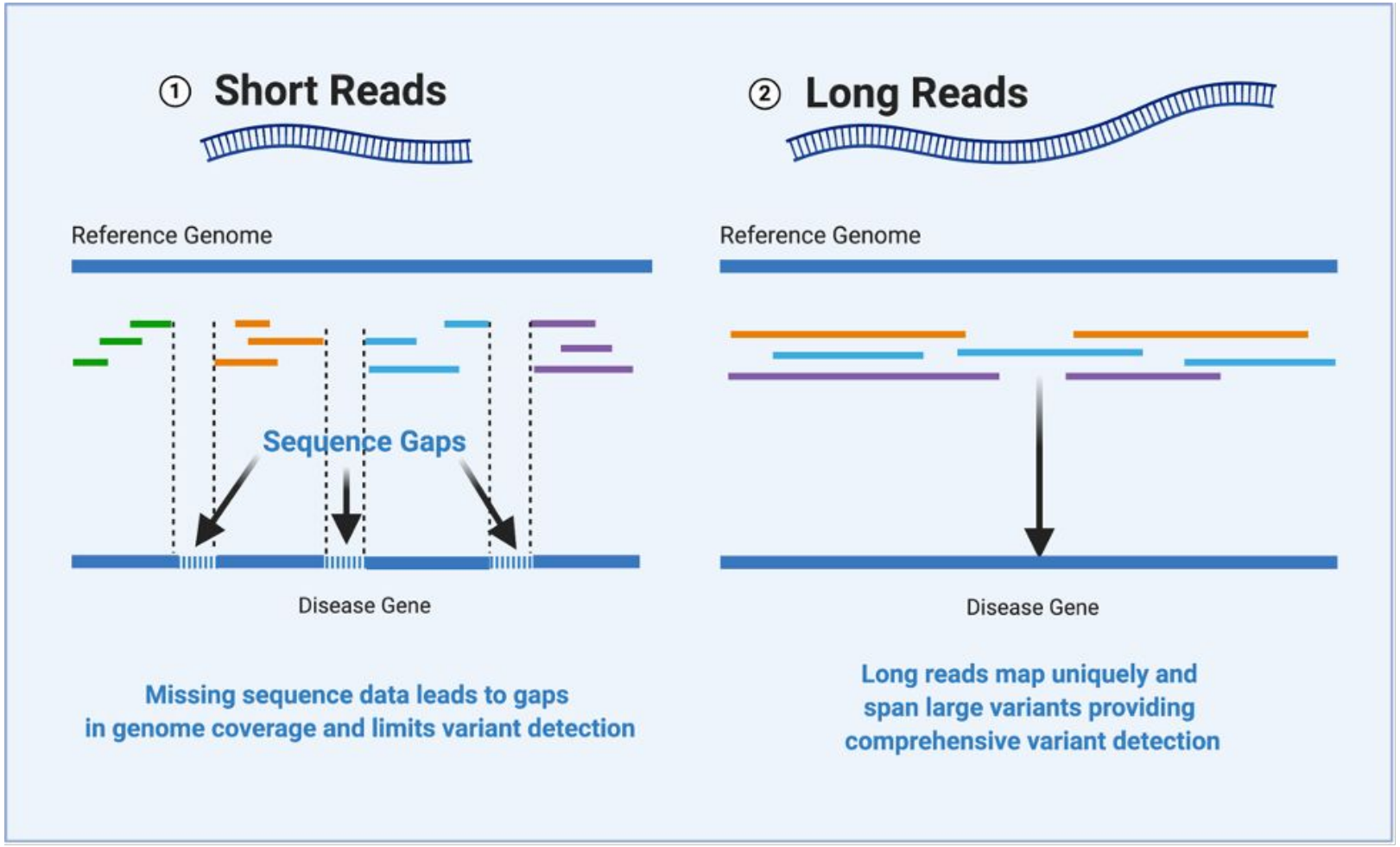
CONTENTS

INTRODUCTION

Targeted amplicon and shotgun metagenomic sequencing



Richa Bharti, Dominik G Grimm, Current challenges and best-practice protocols for microbiome analysis, *Briefings in Bioinformatics*, Volume 22, Issue 1, January 2021, Pages 178–193, https://doi.org/10.1093/bib/bbz155



https://hudsonalpha.org/wp-content/uploads/2021/02/short-read-2.png

Basic Statistics

Measure	Value
Filename	TD_92_run0_raw_1.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	19513006
Total Bases	2.9 Gbp
Sequences flagged as poor quality	0
Sequence length	19-150
%GC	46

Quality control after trimming

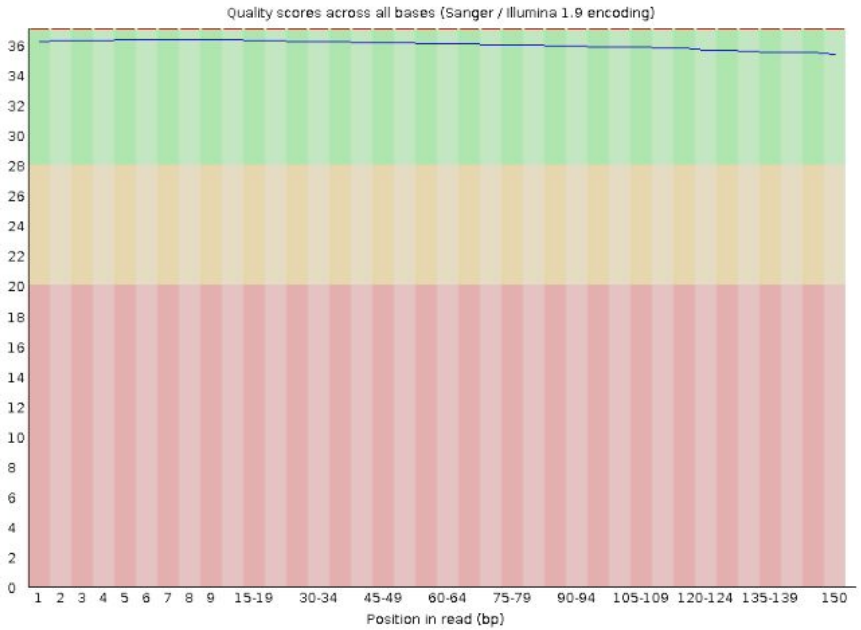
Basic Statistics

Measure	Value
Filename	TD_92_long.unmapped.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	584180
Total Bases	2.8 Gbp
Sequences flagged as poor quality	0
Sequence length	1000-45642
%GC	43

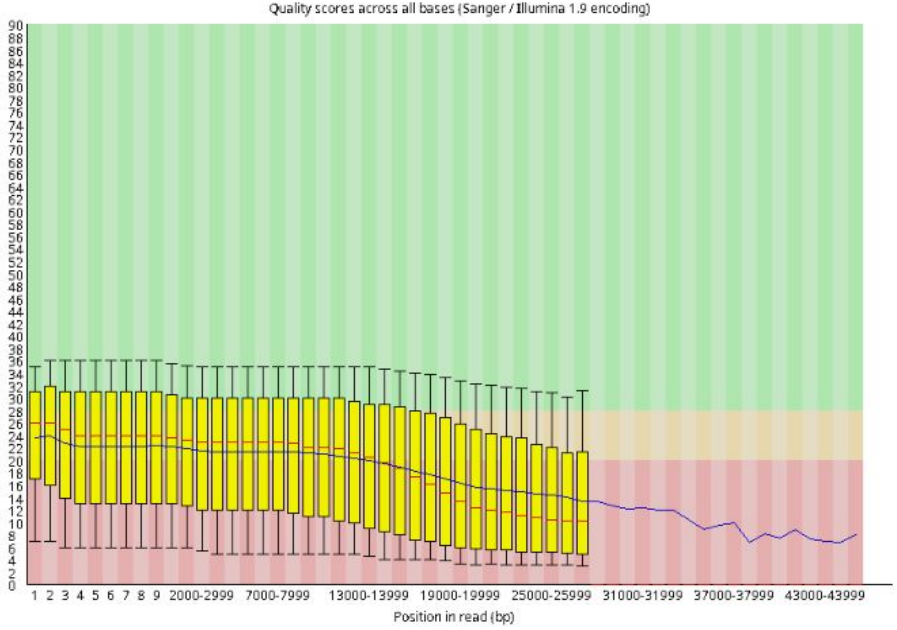
Short reads

Long reads

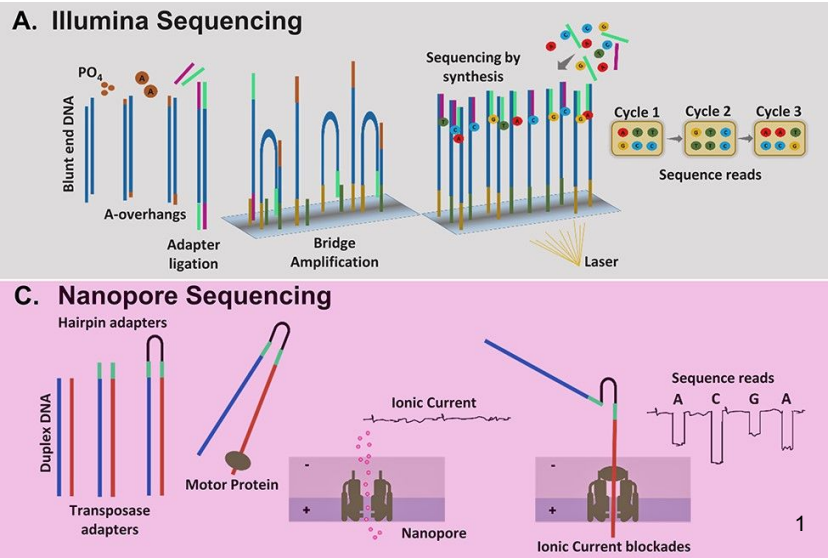
Per base sequence quality

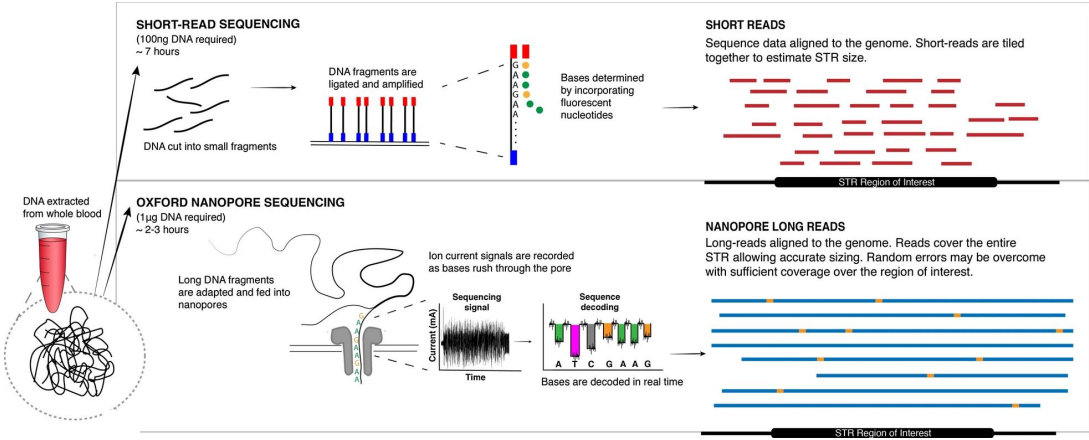


Per base sequence quality



Illumina Sequencing and Nanopore Sequencing





1.Richa Bharti, Dominik G Grimm, Current challenges and best-practice protocols for microbiome analysis, *Briefings in Bioinformatics*, Volume 22, Issue 1, January 2021, Pages 178–193, https://doi.org/10.1093/bib/bbz155

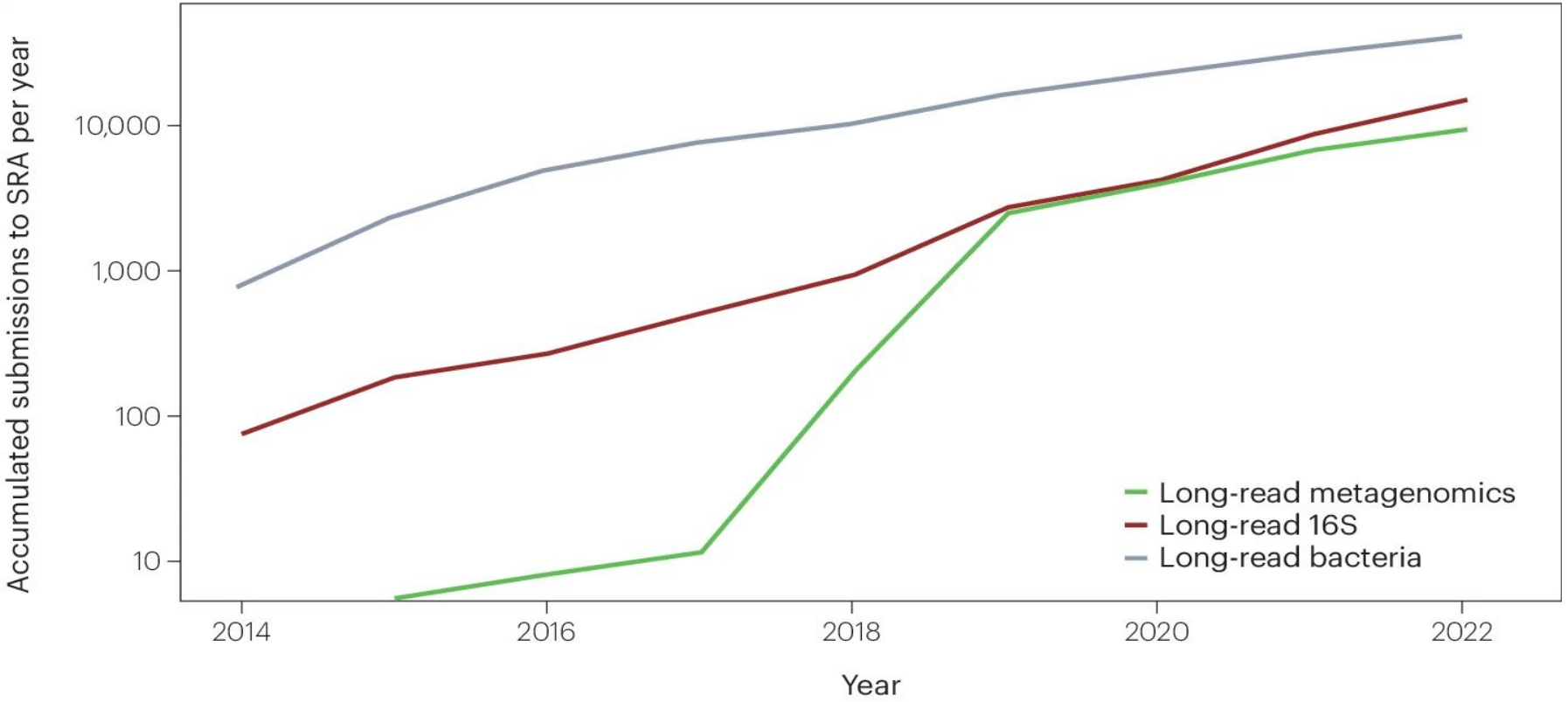
2.Chintalaphani, S.R., Pineda, S.S., Deveson, I.W. et al. An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *acta neuropathol commun 9*, 98 (2021). https://doi.org/10.1186/s40478-021-01201-x

Comparison between short-read and long-read technologies

Platform	MiSeq	NovaSeq 6000	Sequel II	Revio	Flongle	MinION	GridION	PromethION
Company	Illumina	Illumina	PacBio	PacBio	Nanopore	Nanopore	Nanopore	Nanopore
Read length (average)[a]	Up to 2 × 300 bp	Up to 2 × 250 bp[b]	~13.5–20 kb[52]	15–18 kb	20 kb	20 kb	20 kb	20 kb
Yield per cell (Gb)	Up to 15	~350	25 (HiFi)	90	1–1.5	15–20	15–20	~120
Runtime (h)	4–56	13–44	30	24	16–24	72	72	72
Read accuracy (Q20+)	99.75%[128]	99.75%[128]	99.18–99.8%[52]	>99.5%[129]	97–99%[12]	97–99%[12]	97–99%[12]	97–99%[12]
Can perform Direct RNA-seq	No	No	No	No	Yes	Yes	Yes	Yes
DNA input needed	1–500 ng[18]	1–500 ng[18]	150 ng–1 µg[7] [c]	150 ng–1 µg[7] [c]	150 ng–1 µg[7]	150 ng–1 µg[7]	150 ng–1 µg[7]	150 ng–1 µg[7]
Estimated sequencing costs per Gb[d]	US$178–1,705[130,131]	US$3.95[131]	US$30–43[131,132]	US$8–11[132]	US$118–437[131,132]	US$21–51[131,132]	US$29–51[132]	US$6–12[132]

Agustinho, D.P., Fu, Y., Menon, V.K. et al. Unveiling microbial diversity: harnessing long-read sequencing technology. *Nat Methods* 21, 954–966 (2024). https://doi.org/10.1038/s41592-024-02262-1

The growth of long-read-related submission to the Sequence Read Archive (SRA) in recent years

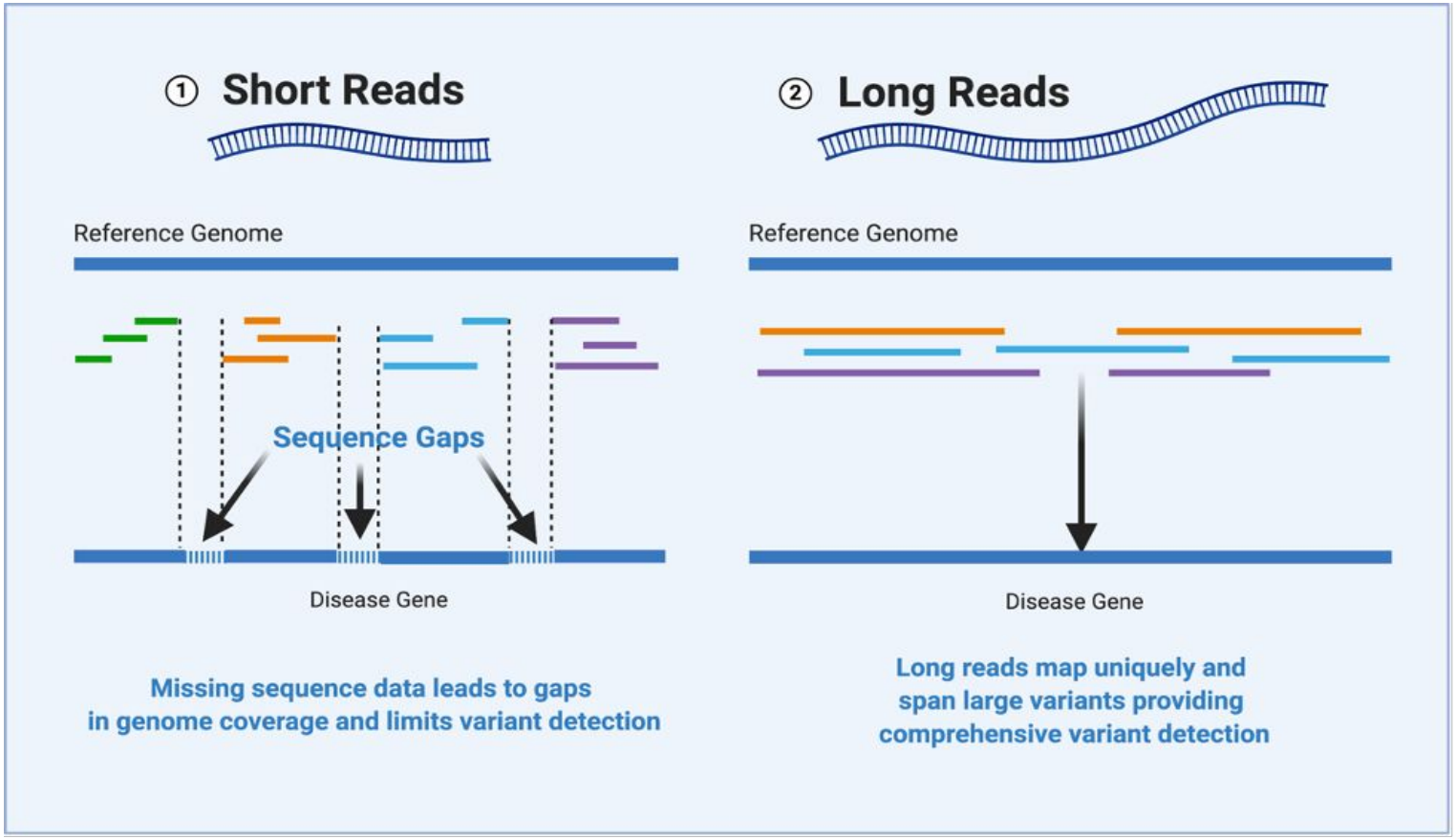


Agustinho, D.P., Fu, Y., Menon, V.K. et al. Unveiling microbial diversity: harnessing long-read sequencing technology. *Nat Methods* 21, 954–966 (2024). https://doi.org/10.1038/s41592-024-02262-1

APPLICATION

1. De novo genome assembly and Metagenomics
2. Structural variant detection
3. Epigenetics studies
4. Portable and real-time sequencing device

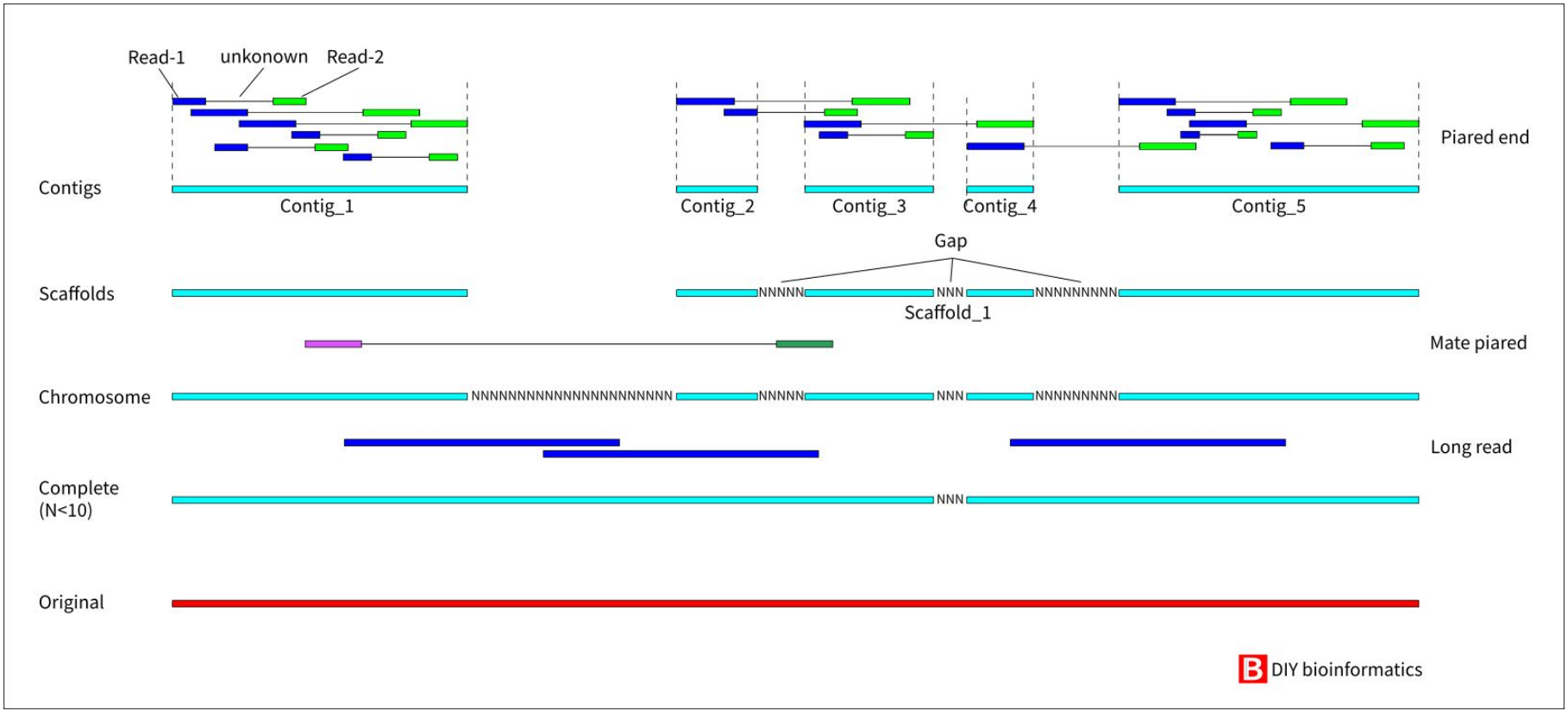
1. DE NOVO GENOME ASSEMBLY AND METAGENOMICS



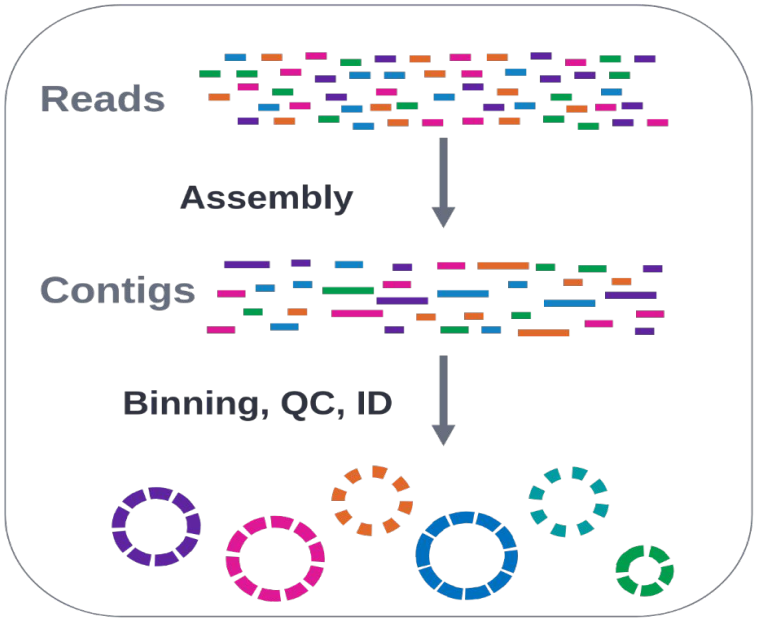
https://hudsonalpha.org/wp-content/uploads/2021/02/short-read-2.png

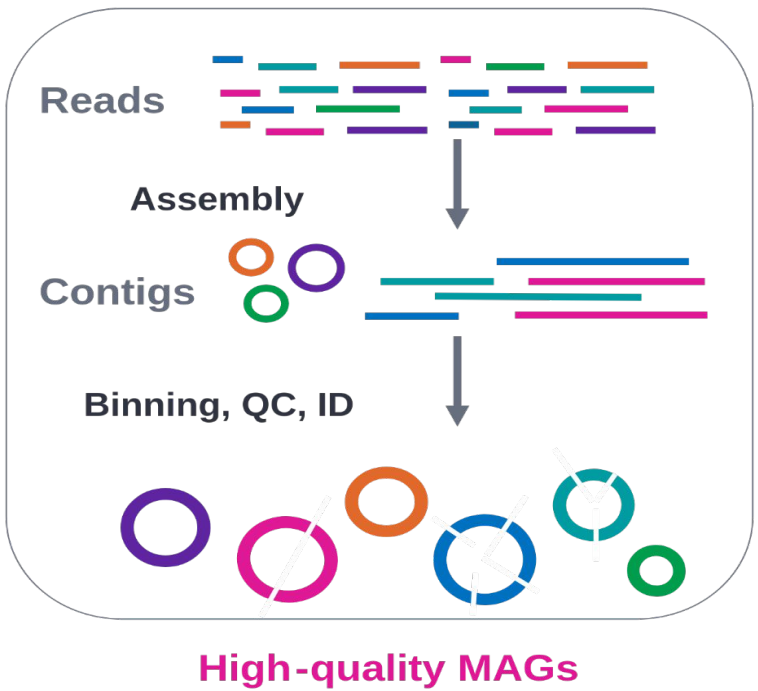
https://www.youtube.com/watch?v=LmUQ4f2f5nc

Contigs, Scaffolds



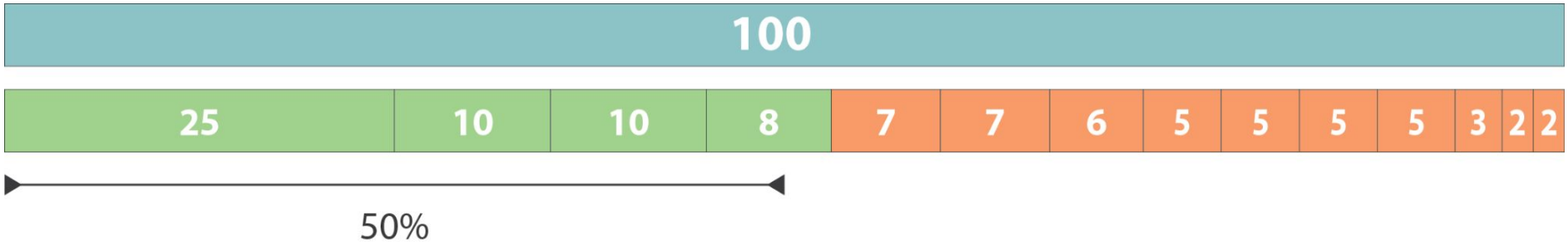
https://diy-bioinformatics.com/contig-scaffold-chromosome-complete_assembly-level/





https://www.pacb.com/blog/sequencing-101-metagenome-assembled-genomes/

N50 and L50



Genome size = 100

Sequence sorted by size list L = (25, 10, 10, 8 , 7, 7 , 6 , 5, 5, 5, 5, 3, 2, 2) = 100

50% of the total length is contained within sequences of at least 8bp: 25 + 10 + 10 + 8 ≥ 50

N50 = 8 and L50 = 4

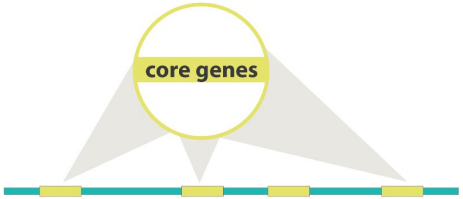
Metrics considered for assessing the performance of the different sequencing strategies

Metric	Description
Assembly length	The total number of nucleotides of the unfiltered (no minimum contig size threshold applied) metagenomic assembly.
Assembly contiguity	The N50 metric of assembly contiguity, defined as the length of the shortest contig for which longer and equal-length contigs cover at least 50% of the assembly.
Assembly coverage	Percentage of short-read sequences (a random subset of 20 Gbp) mapping to the assembly.
MAG number	Number of metagenome-assembled genomes with >70% CheckM completeness and <10% redundancy recovered from the sequencing data.
MAG contiguity	Average number of contigs per MAG.
MAG completeness	Average completeness of the MAGs in terms of CheckM assessment based on the presence of single-copy core genes.
MAG contamination	Inverse to the average contamination of the MAGs in terms of CheckM assessment based on the redundancy of single-copy core genes.

1

Contiguity

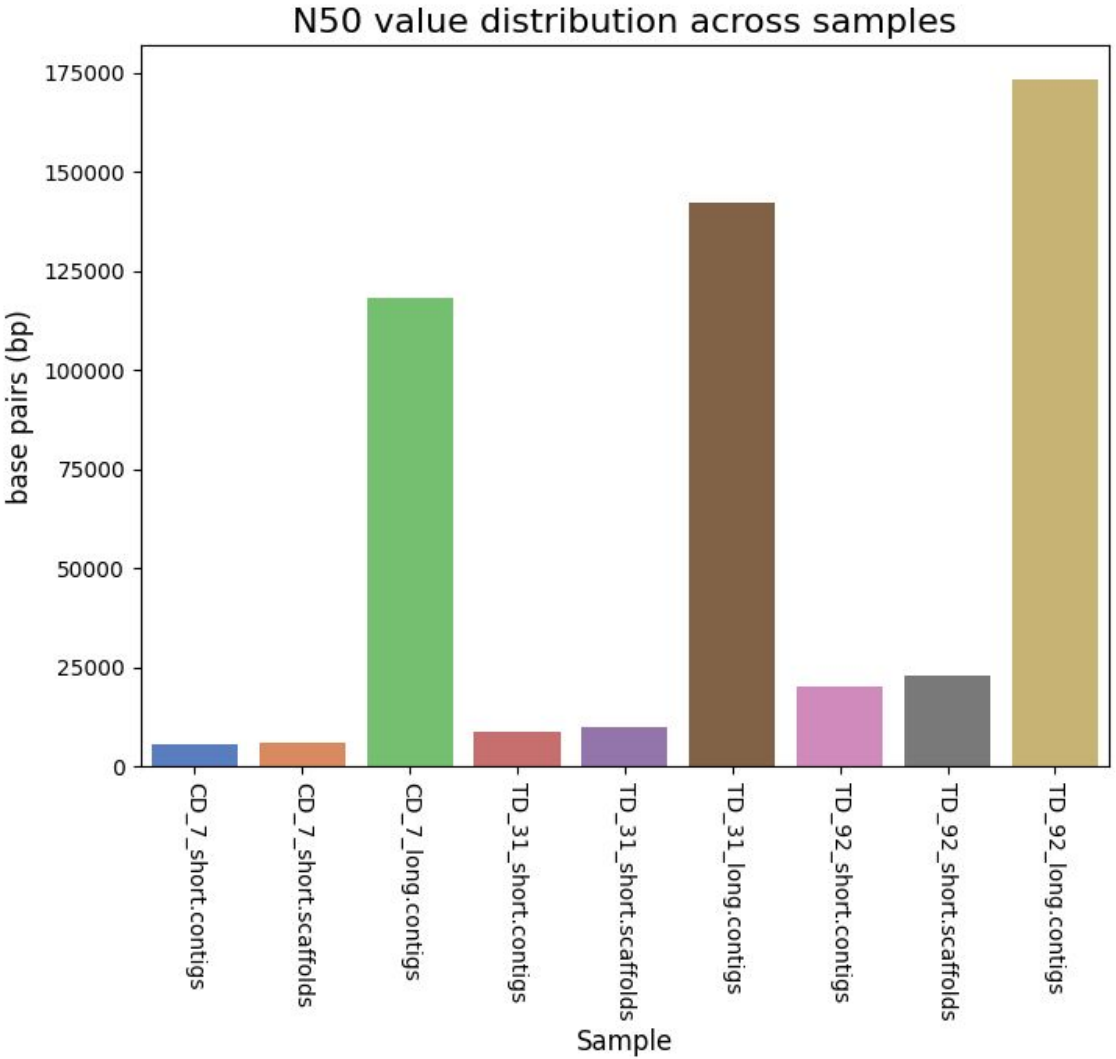
Completeness



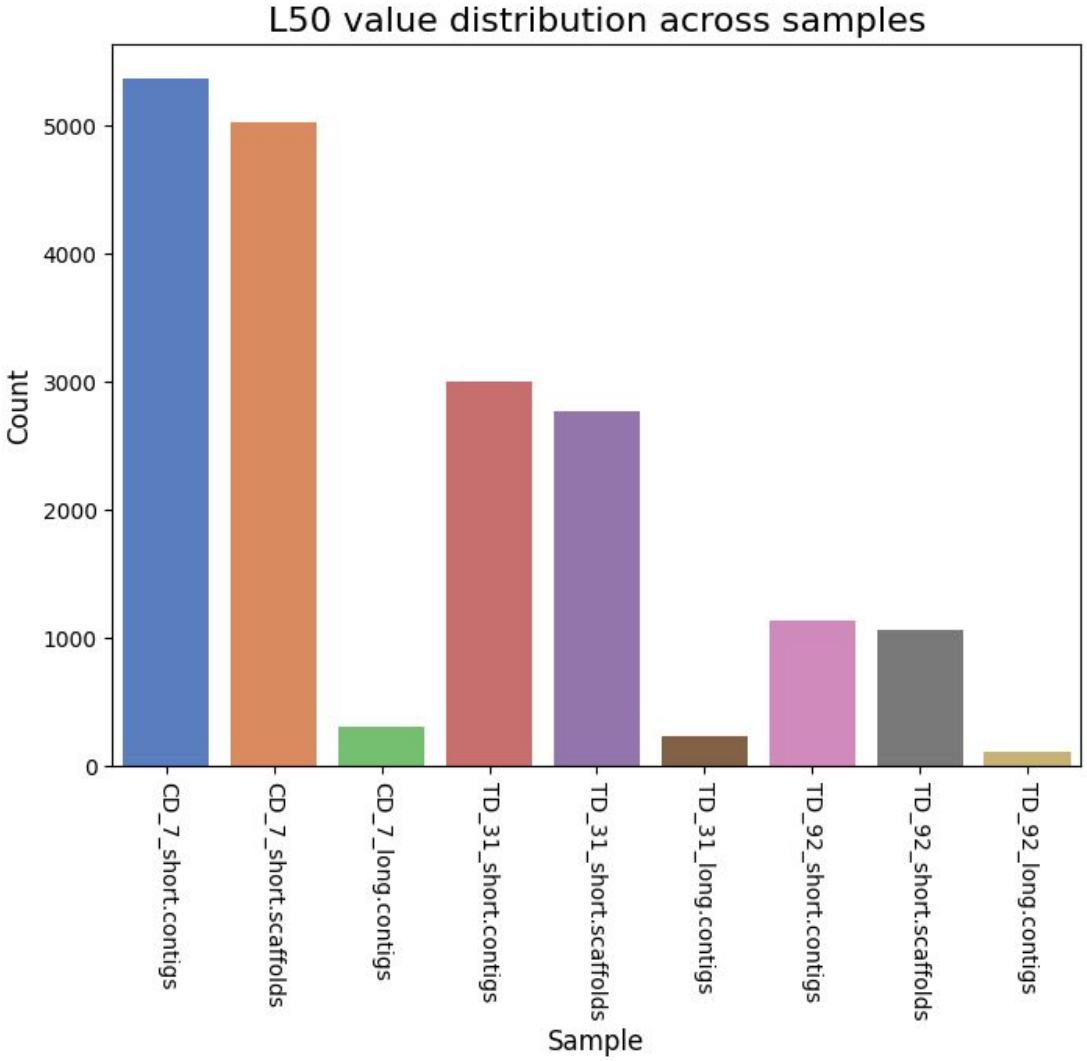
Correctness

2

1.Eisenhofer R, Nesme J, Santos-Bay L, Koziol A, Sørensen SJ, Alberdi A, Aizpurua O. 2024. A comparison of short-read, HiFi long-read, and hybrid strategies for genome-resolved metagenomics. Microbiol Spectr 12:e03590-23. https://doi.org/10.1128/spectrum.03590-23
2. https://training.galaxyproject.org/training-material/topics/assembly/tutorials/assembly-quality-control/slides-plain.html



Long-read sequencing provides a **higher N50** value, which indicates longer contiguous sequences in genome assemblies.



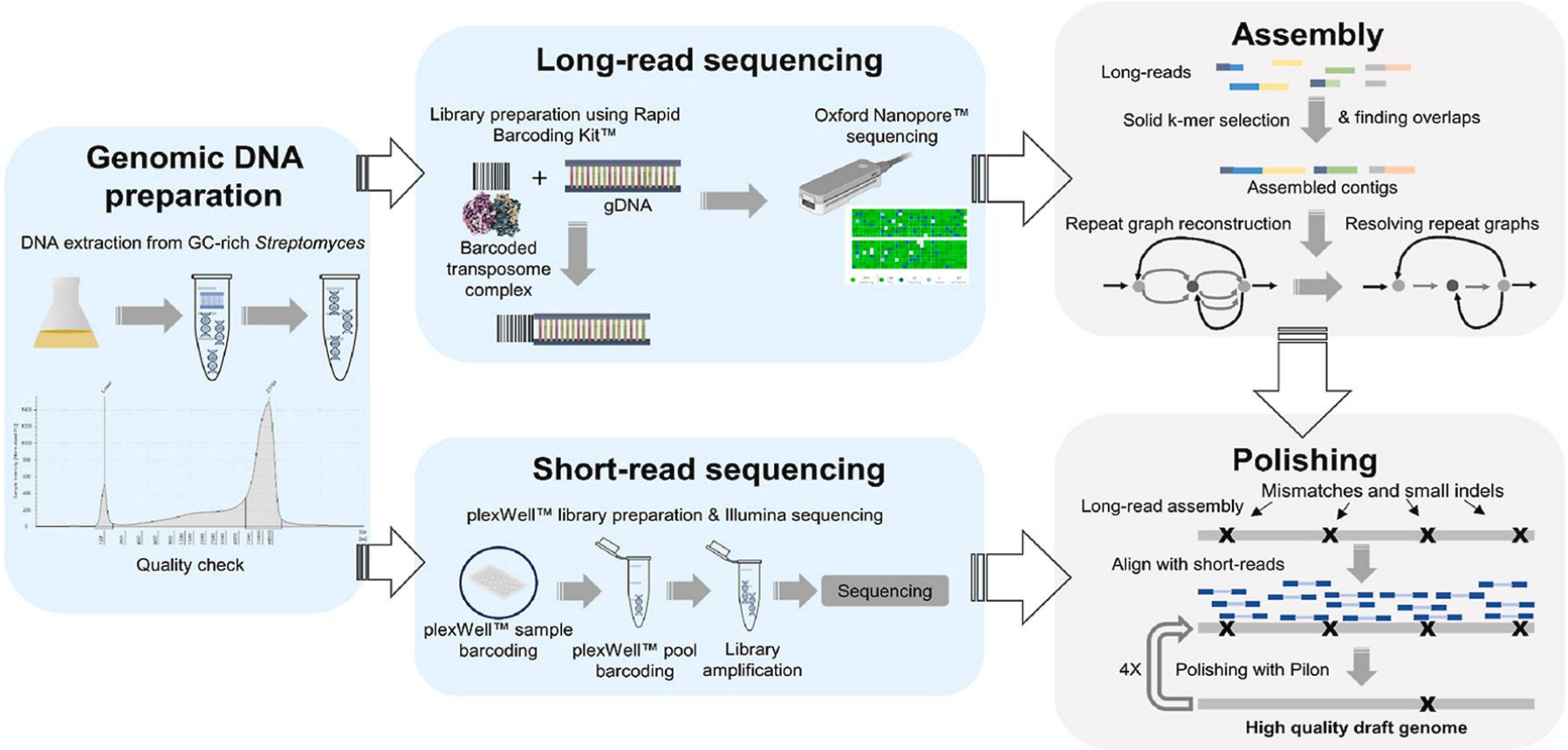
Long-read sequencing offers a **lower L50** value, meaning fewer contigs are required to cover 50% of the genome, contributing to more continuous assemblies.

Assembly visualization

short reads TD_92

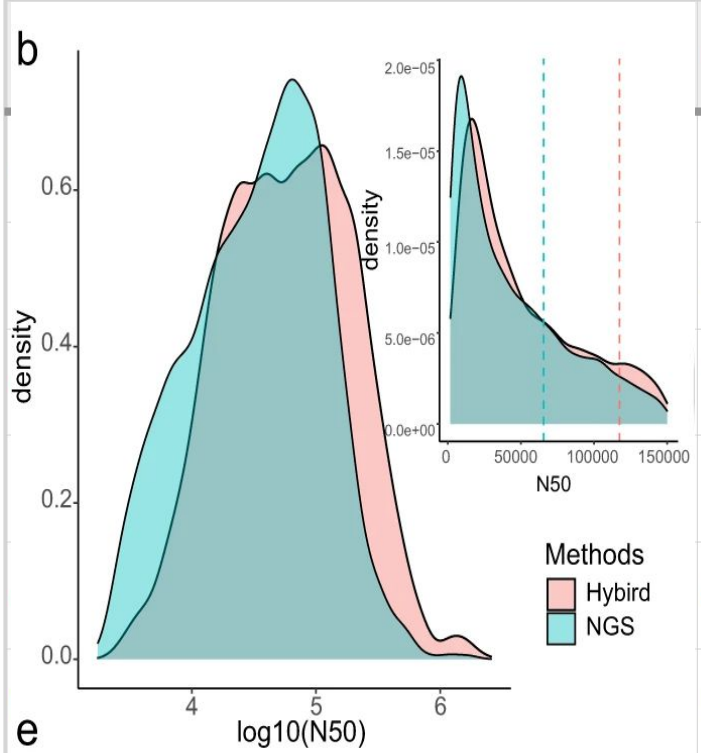
long reads TD_92

Hybrid sequencing



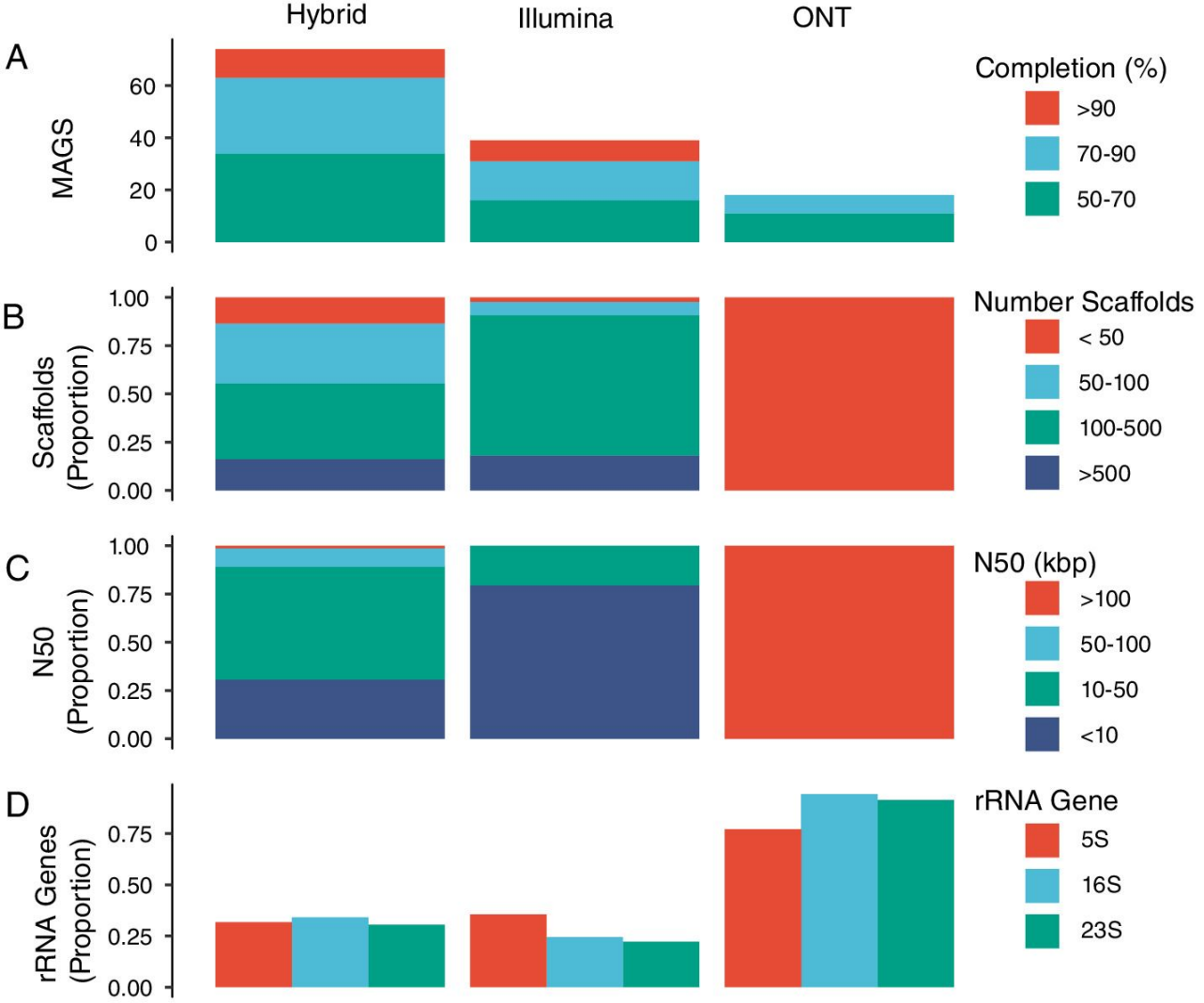
Elena Heng, Lee Ling Tan, Dillon W.P. Tay, Yee Hwee Lim, Lay-Kien Yang, Deborah C.S. Seow, Chung Yan Leong, Veronica Ng, Siew Bee Ng, Yoganathan Kanagasundaram, Fong Tian Wong, Lokanand Koduru. (2023). Cost-effective hybrid long-short read assembly delineates alternative GC-rich Streptomyces hosts for natural product discovery. Synthetic and Systems Biotechnology, 8(2), 253-261. ISSN 2405-805X. https://doi.org/10.1016/j.synbio.2023.03.001.

Hybrid sequencing improves the quality of metagenome assembly



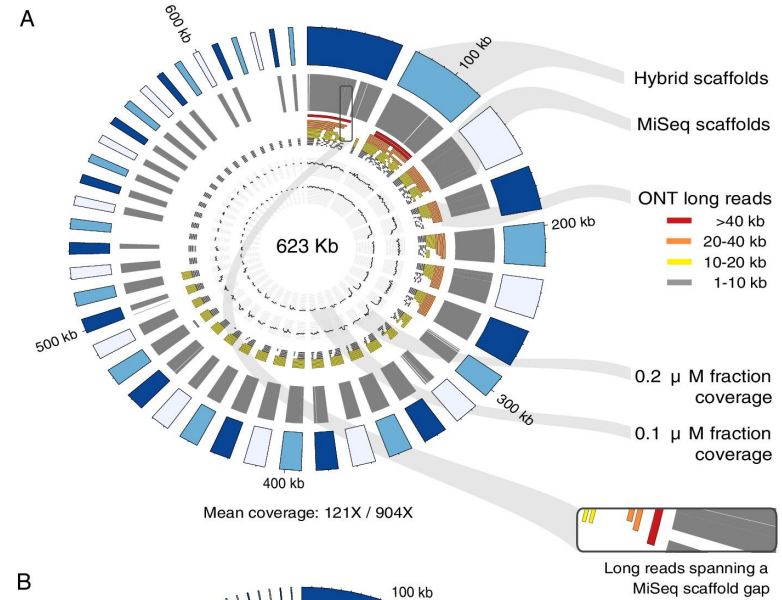
Hybrid assembly		Illumina-only	
1.47E6 (3.97E5–9.23E6)[a]		5.64E7 (3.41E7–1.07E8)	
6.08E9 (1.72E9–2.08E10)[a]		8.46E9 (5.12E9–1.61E10)	
3.55E5 (1.12E5–7.77E5)		4.29E5 (1.35E5–8.63E5)	
3.82E8 (1.69E8–7.05E8)	↑	3.64E8 (1.60E8–6.63E8)	↓
9283 (2002-31,606)	↑	2962 (745-11,932)	↓
50 (21-196)	↑	44 (14-74)	↓

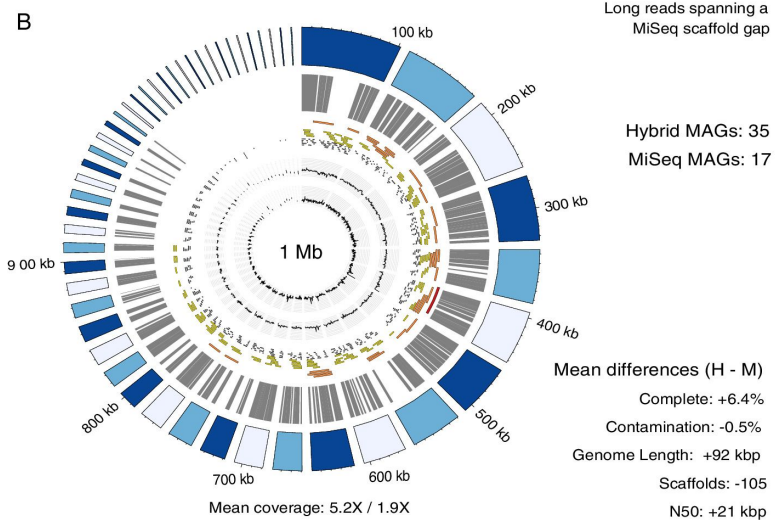
Chen, L., Zhao, N., Cao, J. *et al.* Short- and long-read metagenomics expand individualized structural variations in gut microbiomes. *Nat Commun* 13, 3175 (2022). https://doi.org/10.1038/s41467-022-30857-9



Hybrid sequencing improves the quality of metagenome assembly

Overholt WA, Hölzer M, Geesink P, Diezel C, Marz M, Küsel K. Inclusion of Oxford Nanopore long reads improves all microbial and viral metagenome-assembled genomes from a complex aquifer system. *Environ Microbiol.* 2020;22(9):4000-4013. doi:10.1111/1462-2920.15186





Genome circos plots for the most (A) and least (B) covered Patescibacteria MAGs retrieved by the hybrid and Illumina-only assembly methods

Overholt WA, Hölzer M, Geesink P, Diezel C, Marz M, Küsel K. Inclusion of Oxford Nanopore long reads improves all microbial and viral metagenome-assembled genomes from a complex aquifer system. *Environ Microbiol.* 2020;22(9):4000-4013. doi:10.1111/1462-2920.15186

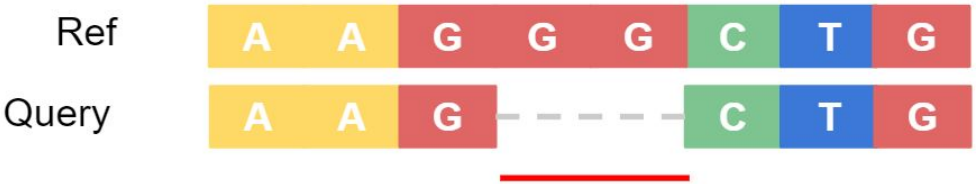
2. STRUCTURAL VARIANT DETECTION

Structural variants (SVs) are large-scale events (>50 bp) where entire sections of genetic material have changed.

Sequence Variants

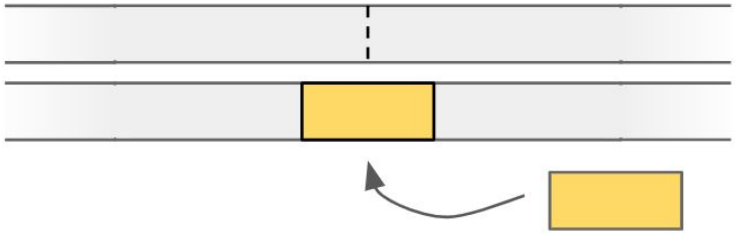
SNV (**S**ingle **N**ucleotide **V**ariant)

INDEL (**In**sertion or **Del**etion)

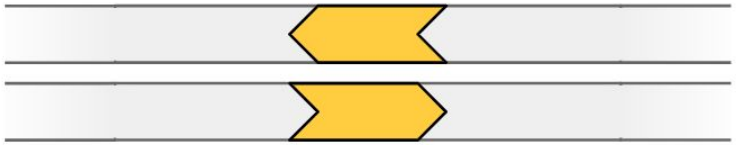


Structural Variants

Insertion

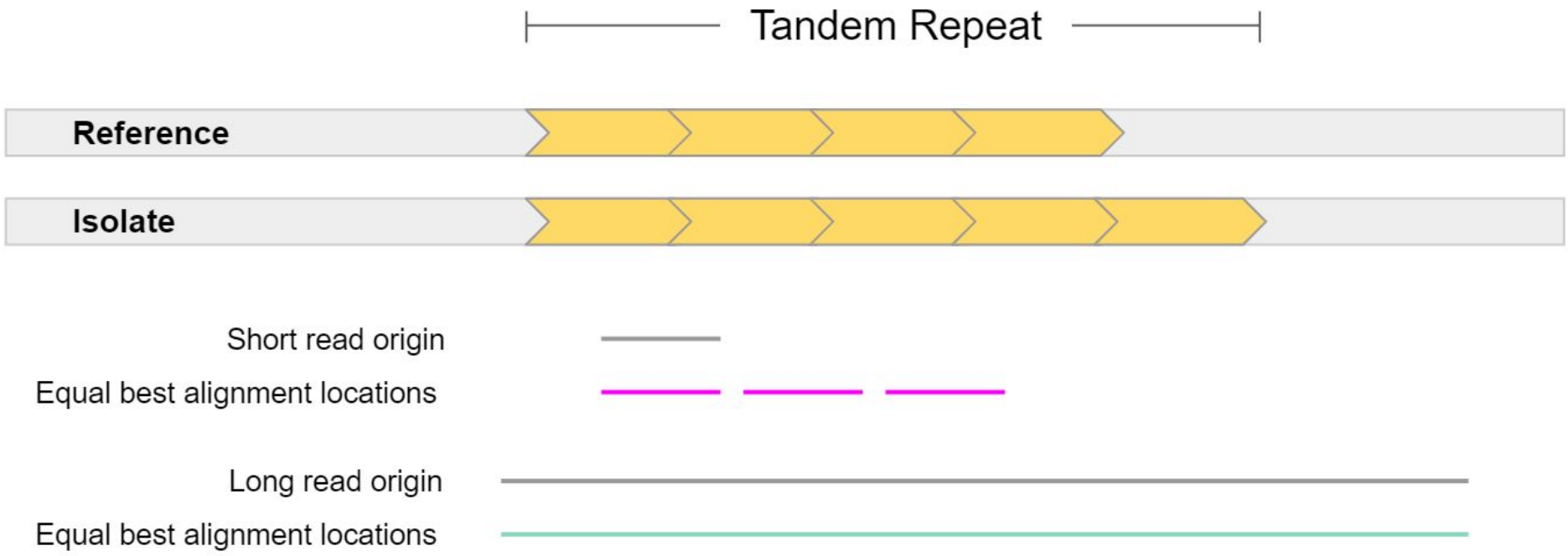


Inversion



https://www.melbournebioinformatics.org.au/tutorials/tutorials/longread_sv_calling/longread_sv_calling/

SVs are difficult to detect due to their association with repeat elements, but **long reads improve mappability, allowing for more accurate detection.**

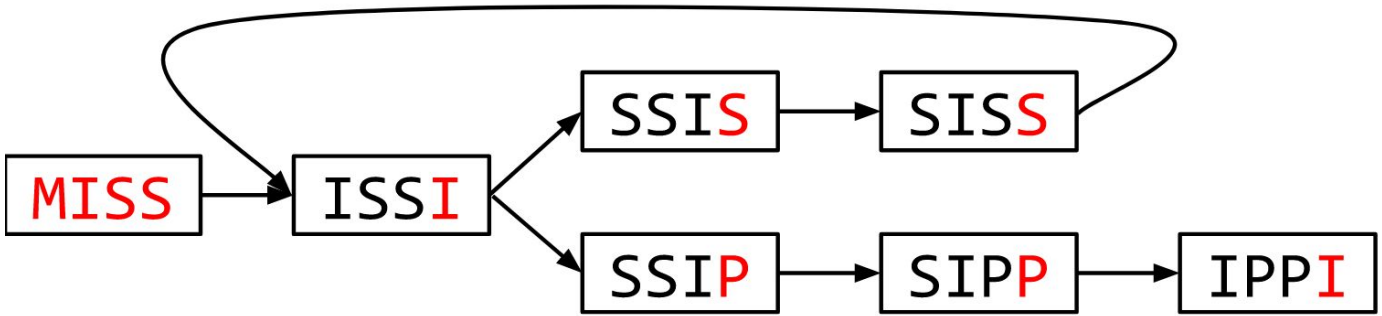


https://www.melbournebioinformatics.org.au/tutorials/tutorials/longread_sv_calling/longread_sv_calling/

Reference: MISSISSIPPI

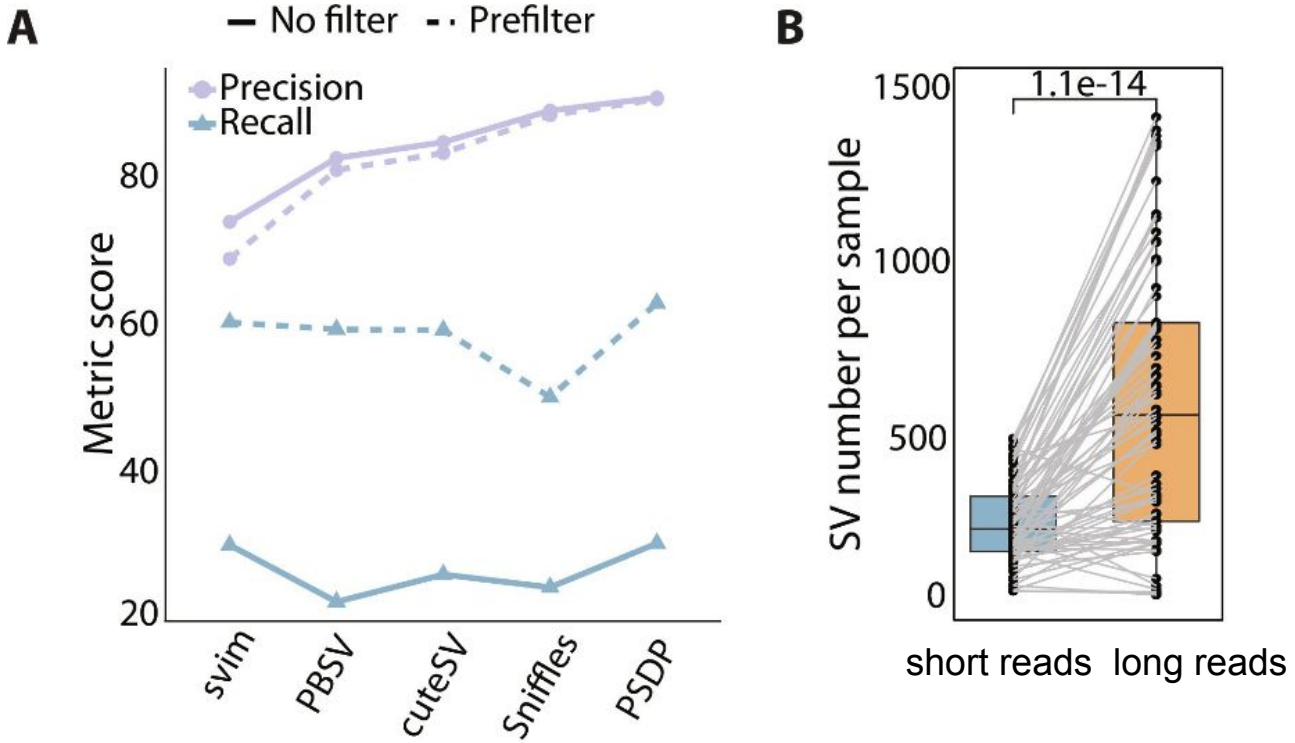
Reads:

MISS ISSI SSIS SISS SSIP SIPP IPPI



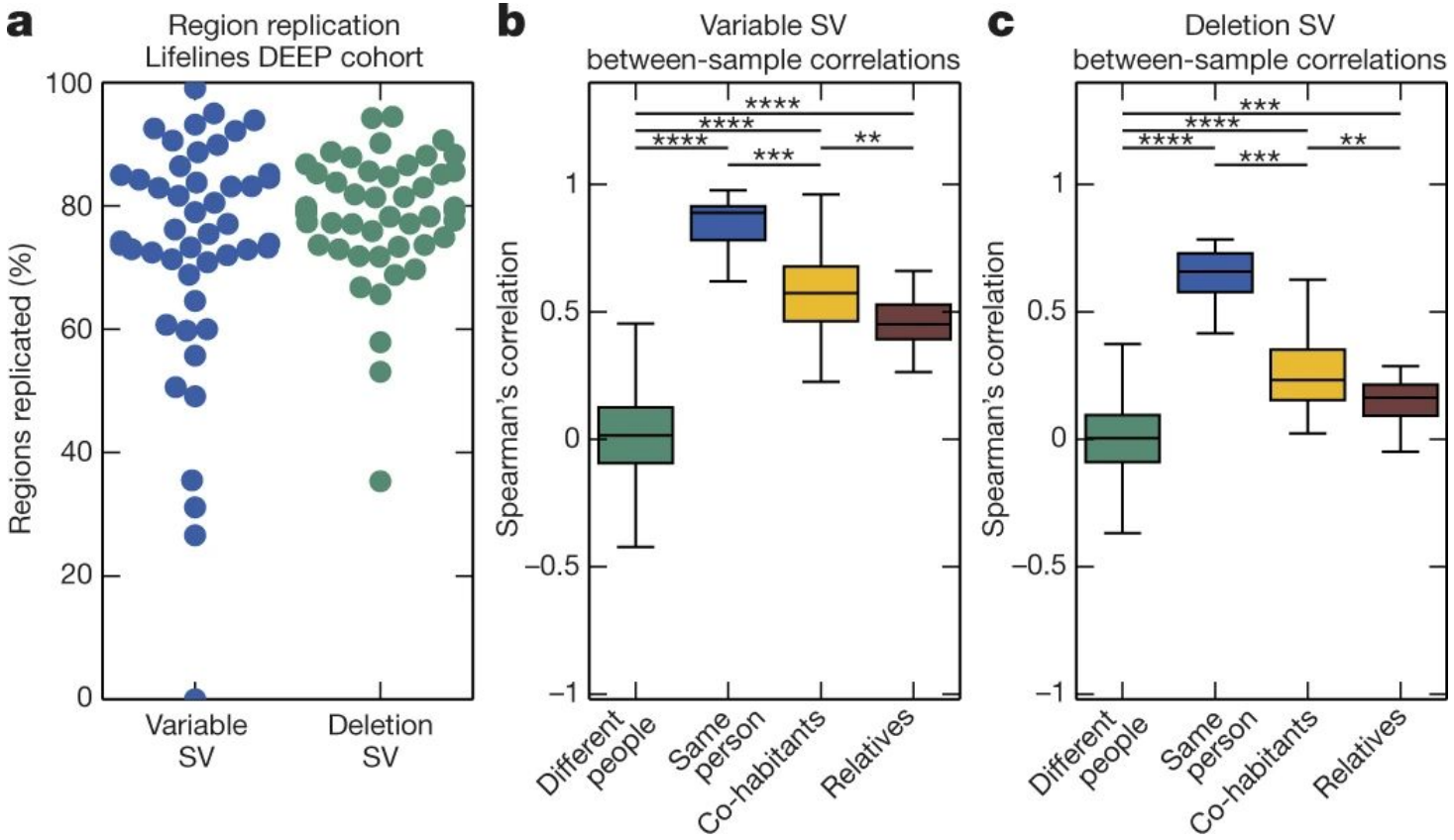
MISSISSIPPI or MISSISSISSISSIPPI or ...

Long reads enhance the accuracy and resolution in identifying large insertions, deletions, inversions, and translocations.



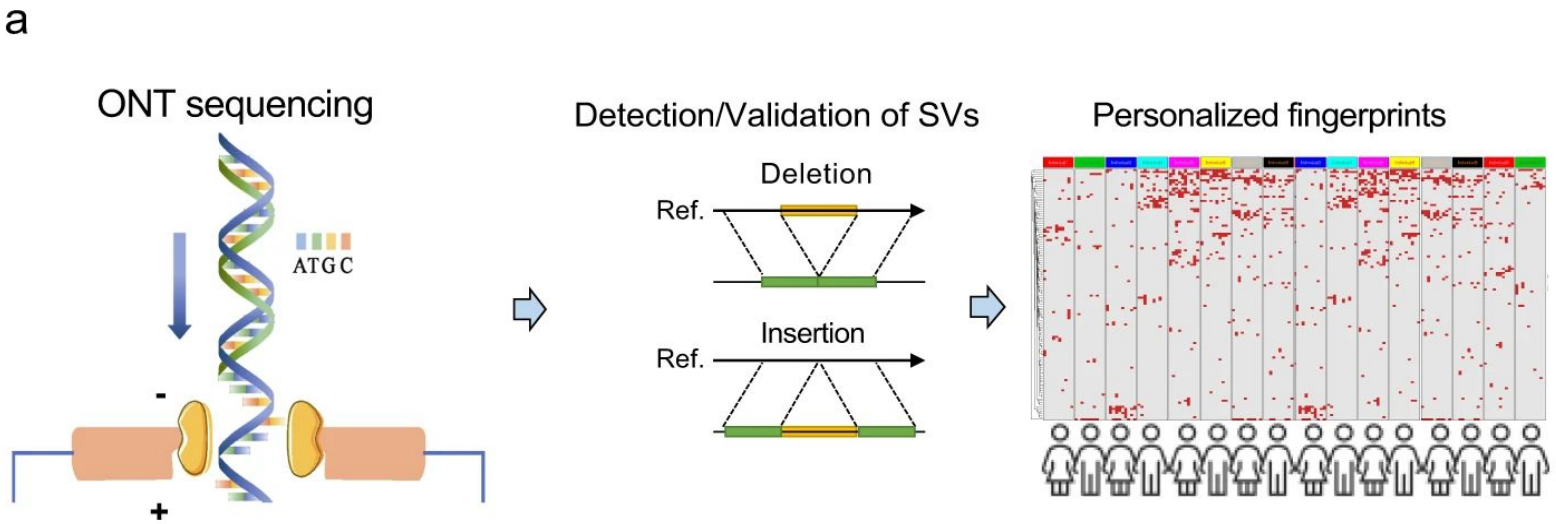
Senying Lai et al. ,Long-read sequencing reveals extensive gut phageome structural variations driven by genetic exchange with bacterial hosts. *Sci. Adv.*10,eadn3316(2024).DOI:10.1126/sciadv.adn3316

SVs are prevalent across distinct populations



Zeevi, D., Korem, T., Godneva, A. et al. Structural variation in the gut microbiome associates with host health. *Nature* 568, 43–48 (2019). https://doi.org/10.1038/s41586-019-1065-y

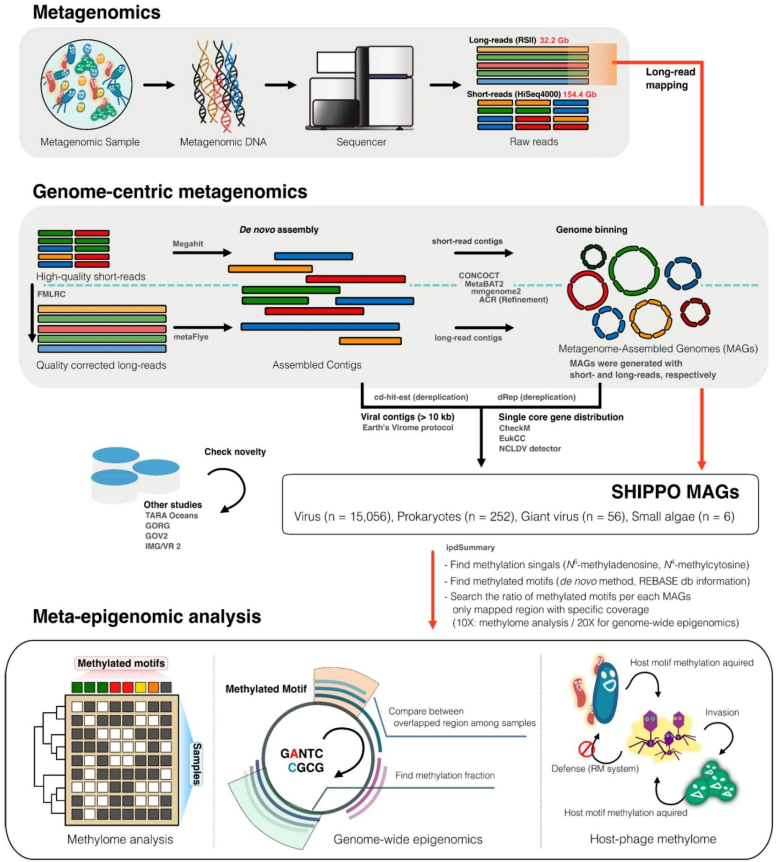
Hybrid sequencing expands SVs like insertions and inversions that **represent personalized microbial signatures**, complicating correlations with metabolites and host health indicators



Chen, L., Zhao, N., Cao, J. *et al.* Short- and long-read metagenomics expand individualized structural variations in gut microbiomes. *Nat Commun* 13, 3175 (2022). https://doi.org/10.1038/s41467-022-30857-9

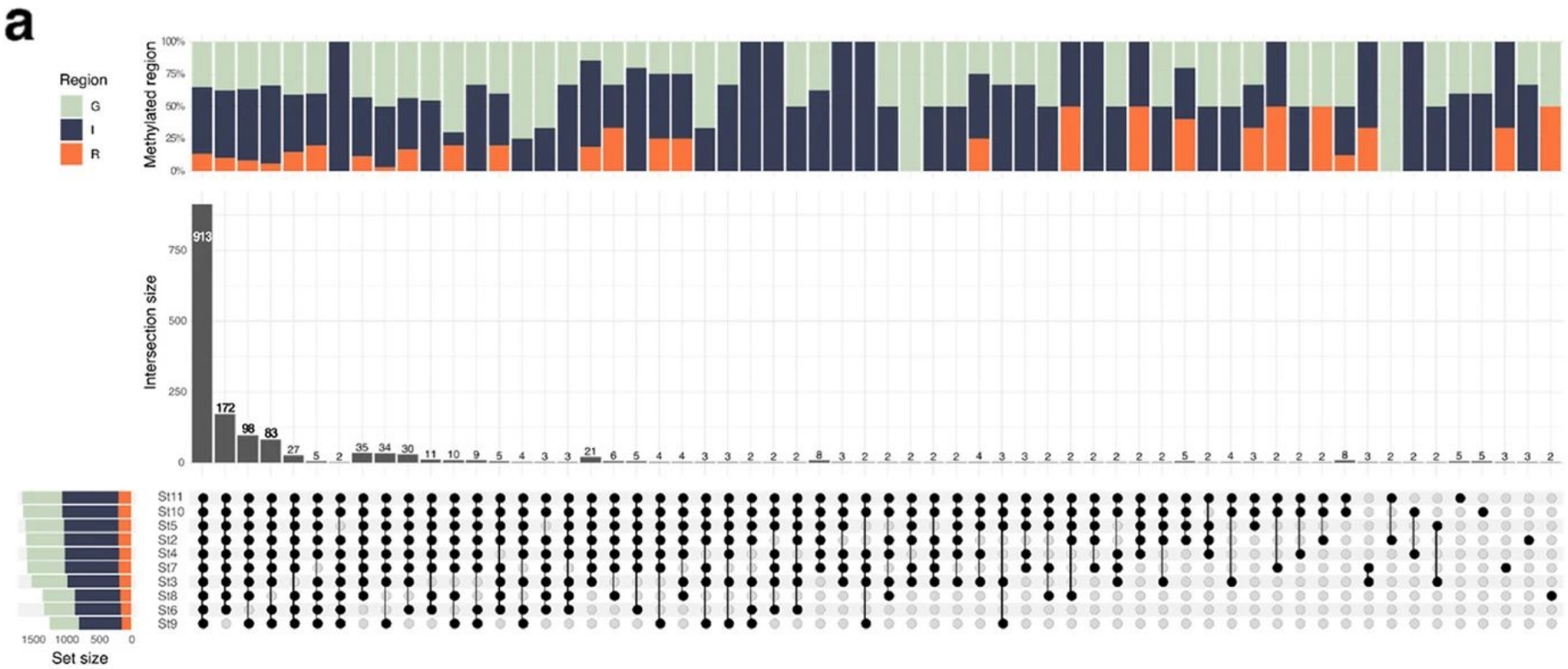
3. EPIGENETICS STUDIES

Meta-epigenome analysis benefits from long-read sequencing by offering comprehensive insights into DNA methylation patterns and structural variations across diverse microbial communities, even at **sub-strain levels.**



Seong, H.J., Roux, S., Hwang, C.Y. et al. Marine DNA methylation patterns are associated with microbial community composition and inform virus-host dynamics. Microbiome 10, 157 (2022). https://doi.org/10.1186/s40168-022-01340-w

Compares methylation at each genome position on a Pelagibacter MAG

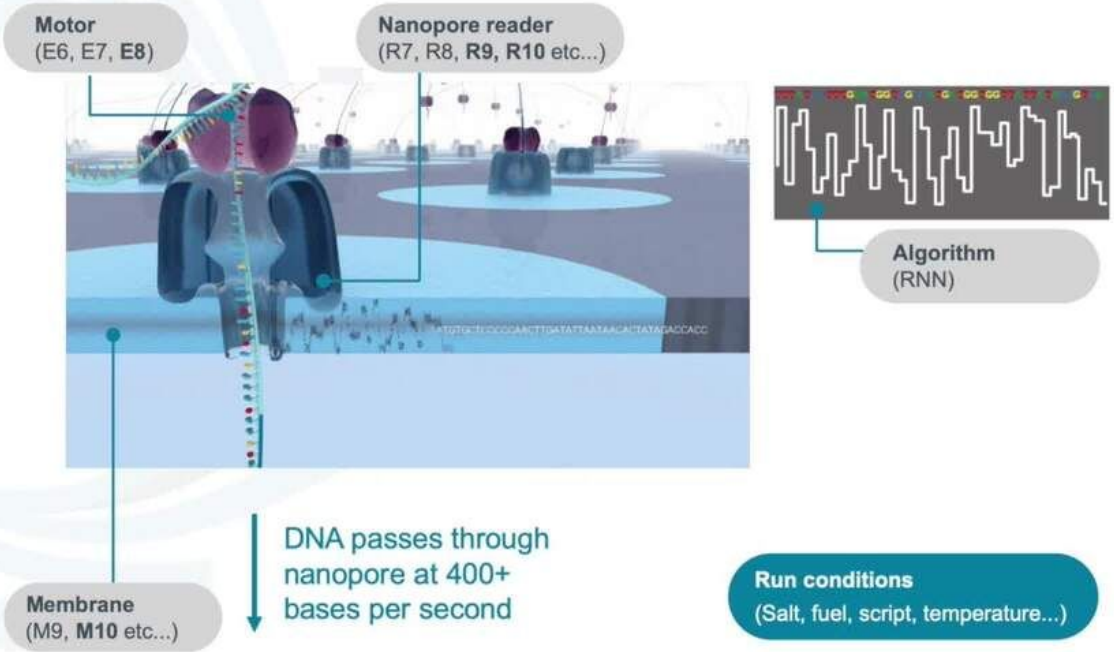


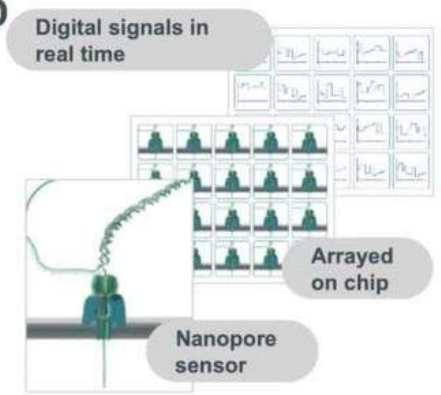
Seong, H.J., Roux, S., Hwang, C.Y. et al. Marine DNA methylation patterns are associated with microbial community composition and inform virus-host dynamics. Microbiome 10, 157 (2022). https://doi.org/10.1186/s40168-022-01340-w

4. PORTABLE AND REAL-TIME SEQUENCING DEVICE

NANOPORE DNA/RNA SEQUENCING: A NOVEL, SCALABLE METHOD

DNA/RNA strand passes through pore ➔ signal interpreted into sequence data





https://www.youtube.com/watch?v=GUb1TZvMWsw

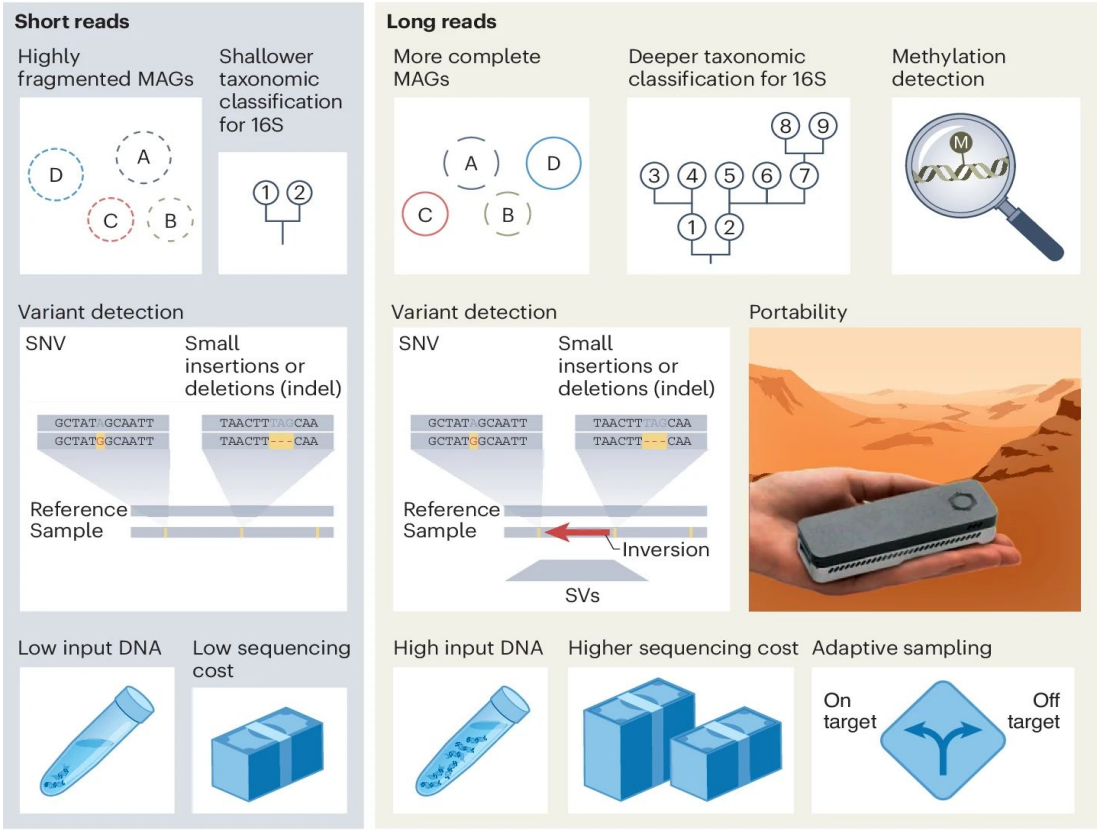
Oxford NANOPORE Technologies

CHALLENGES

- Accurate and unbiased species identification
- Lack of universally accepted benchmark samples
- Difficulty in discerning sub-strain variations
- Computer resources

Agustinho, D.P., Fu, Y., Menon, V.K. et al. Unveiling microbial diversity: harnessing long-read sequencing technology. *Nat Methods* 21, 954–966 (2024). https://doi.org/10.1038/s41592-024-02262-1

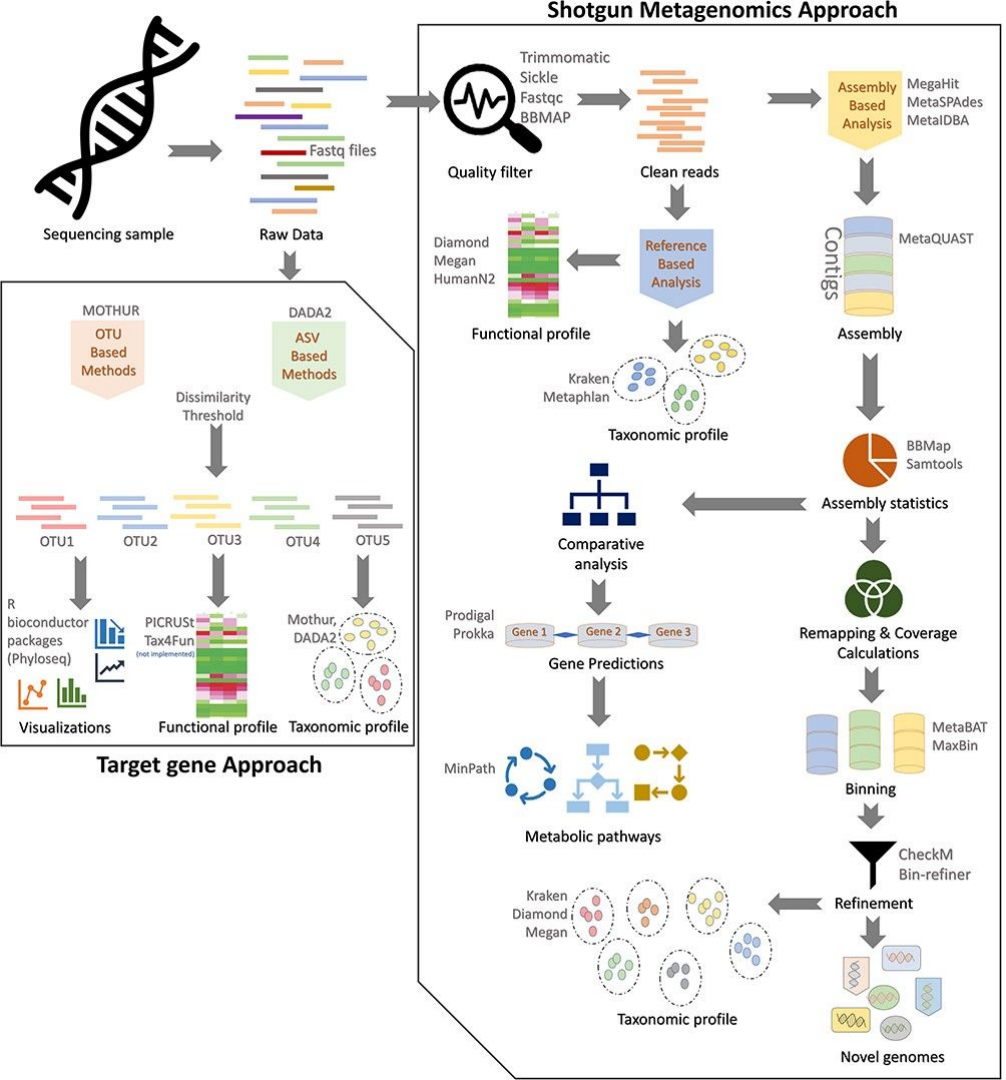
SUMMARY



Long reads can

- generate less fragmented genome assemblies
- lower-level (species/strain) taxonomic characterization
- DNA/RNA methylation pattern identification
- large SV detection
- highly portable sequencers and realtime

Agustinho, D.P., Fu, Y., Menon, V.K. et al. Unveiling microbial diversity: harnessing long-read sequencing technology. *Nat Methods* 21, 954–966 (2024). https://doi.org/10.1038/s41592-024-02262-1



WORKFLOW

Richa Bharti, Dominik G Grimm, Current challenges and best-practice protocols for microbiome analysis, *Briefings in Bioinformatics*, Volume 22, Issue 1, January 2021, Pages 178–193, https://doi.org/10.1093/bib/bbz155

Examples of Software

	Short reads	Long reads	Hybrid	EPI2ME
Adapter / quality trimming	fastp	porechop	fastp, porechop filtlong	No
Host read removal	Bowtie2	Minimap2	Bowtie2, Minimap2	Minimap2
Quality control	FastQC	Fastcat, NanoPlot, FastQC	FastQC, NanoPlot	Fastcat
Taxonomic classification	Reads: Kraken2, Bracken, Centrifuge Contigs, MAGs: CAT, GTDB-Tk			Kraken2, Bracken
Assembly	SPAdes, MEGAHIT QUAST (for evaluation)	Canu, MetaFlye, hifiasm-meta and pipelines such as Lathe QUAST (for evaluation)	SPADES Hybrid	No
Binning	MetaBAT2, MaxBin2, CONCONCT, MyCC, GroopM, MetaWRAP, Anvi'o, SemiBin QUAST (for evaluation)	MetaBAT2, MaxBin2, CONCONCT, MetaBCC-LR, LRBinner QUAST (for evaluation)	MetaBAT2, MaxBin2, CONCONCT, MyCC, GroopM, MetaWRAP, Anvi'o, SemiBin QUAST (for evaluation)	No
Binning refinement	DAS Tool, CheckM			No
Statistics	R, python packages			

ONT Metagenomics workflow

epi2me-labs/**wf-metagenomics**

Metagenomic classification of long-read sequencing data

8 Contributors

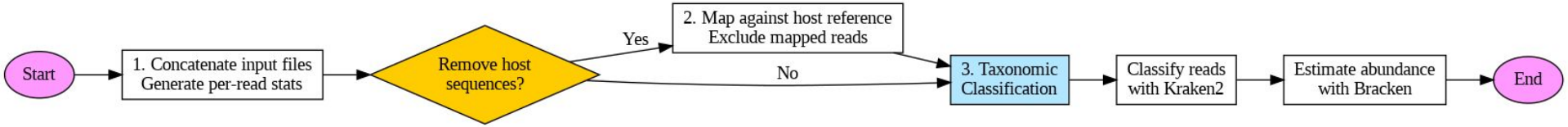
2 Issues

55 Stars

23 Forks

https://github.com/epi2me-labs/wf-metagenomics

ONT Metagenomics workflow



1. **Concatenate fastq(.gz) files whilst creating a summary of the sequences**
   - Use fastcat to concatenate FASTQ files
   - Output per-read stats (lengths, average qualities)

2. **Remove host sequences**
   - Use Minimap2 to map against provided host reference
   - Exclude mapped reads from further analysis

3. **Taxonomic Classification by kraken2 (faster method)**
   - Classify reads taxonomically
   - Use Bracken for species abundance estimation
   - Supports real-time analysis

A SIMULATION PIPELINE

Metadata

- BioProject: PRJNA820119
- Strategy: SHOTGUN
- Host: Homo sapiens
- Isolation source: fecal samples
- Collection date: November, 2019
- Geographic location: China: Beijing
- latitude and longitude: 39.3 N 116.194 E
- Instrument: Illumina NovaSeq 6000 and PromethION

Chen, L., Zhao, N., Cao, J. *et al.* Short- and long-read metagenomics expand individualized structural variations in gut microbiomes. *Nat Commun* 13, 3175 (2022). https://doi.org/10.1038/s41467-022-30857-9

Metadata

SRR18491279	CD_7_short	(19.6M spots, 5.9G bases, 1.8Gb) x 2
SRR18491202	CD_7_long	511,893 spots, 3.5G bases, 2.9Gb
SRR18491059	TD_31_short	(18.8M spots, 5.6G bases, 1.7Gb) x 2
SRR18490949	TD_31_long	476,473 spots, 2.9G bases, 2.5Gb
SRR18491219	TD_92_short	(19.7M spots, 5.9G bases, 1.8Gb) x 2
SRR18491134	TD_92_long	582,734 spots, 2.9G bases, 2.5Gb

Chen, L., Zhao, N., Cao, J. *et al*. Short- and long-read metagenomics expand individualized structural variations in gut microbiomes. *Nat Commun* 13, 3175 (2022). https://doi.org/10.1038/s41467-022-30857-9

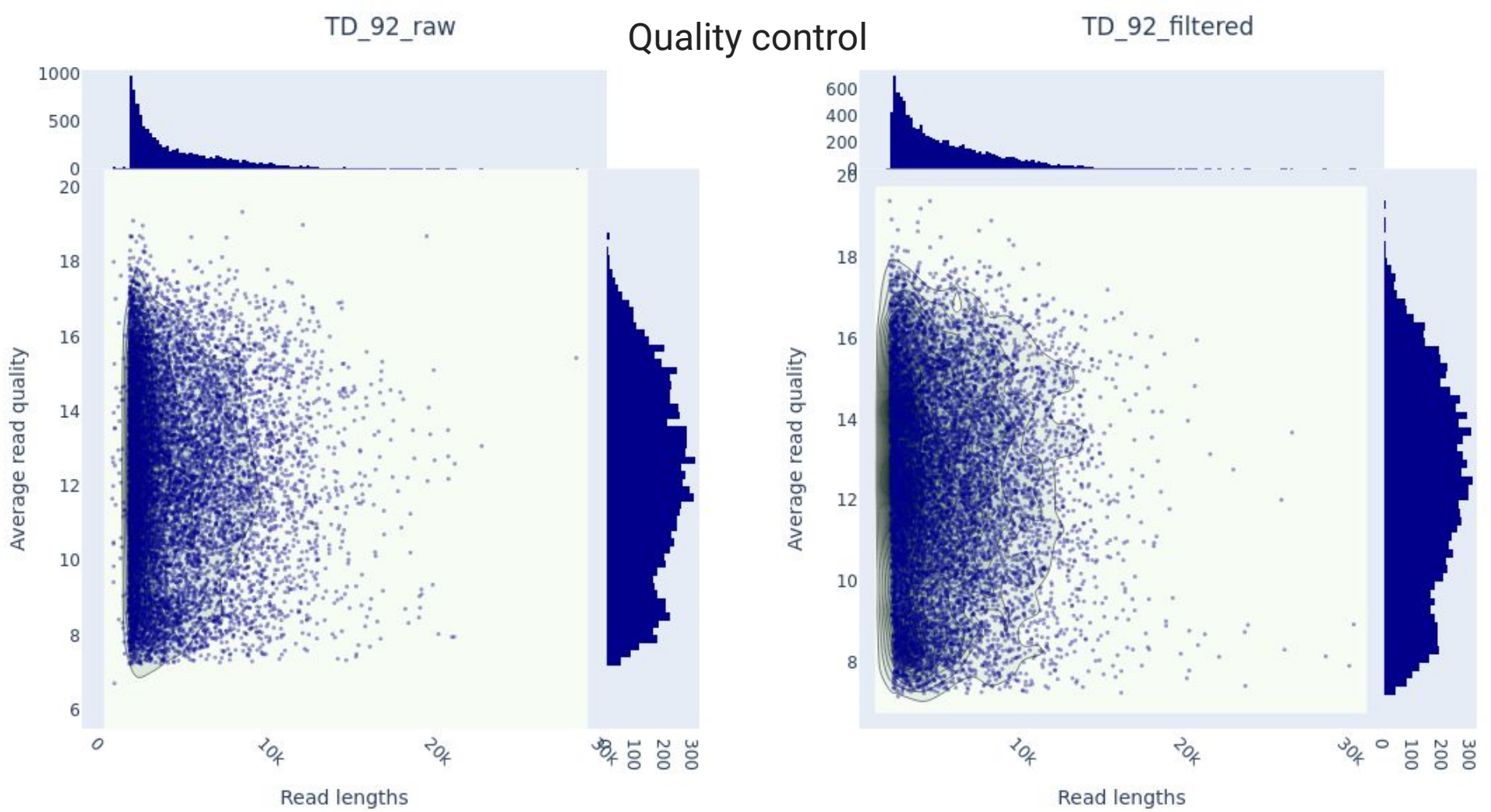
WORKFLOW

	Short reads	Long reads
Adapter / quality trimming	fastp	porechop
Host read removal	Bowtie2	Minimap2
Quality control	FastQC	NanoPlot
Taxonomic classification	Kraken2, Bracken	
Assembly	SPAdes QUAST (for evaluation)	MetaFlye QUAST (for evaluation)
Binning	MetaWRAP - MetaBAT2	
Binning evaluation, refinement and annotation	MetaWRAP - CheckM	
Statistics: relative abundance, alpha diversity, beta diversity	R, python packages	

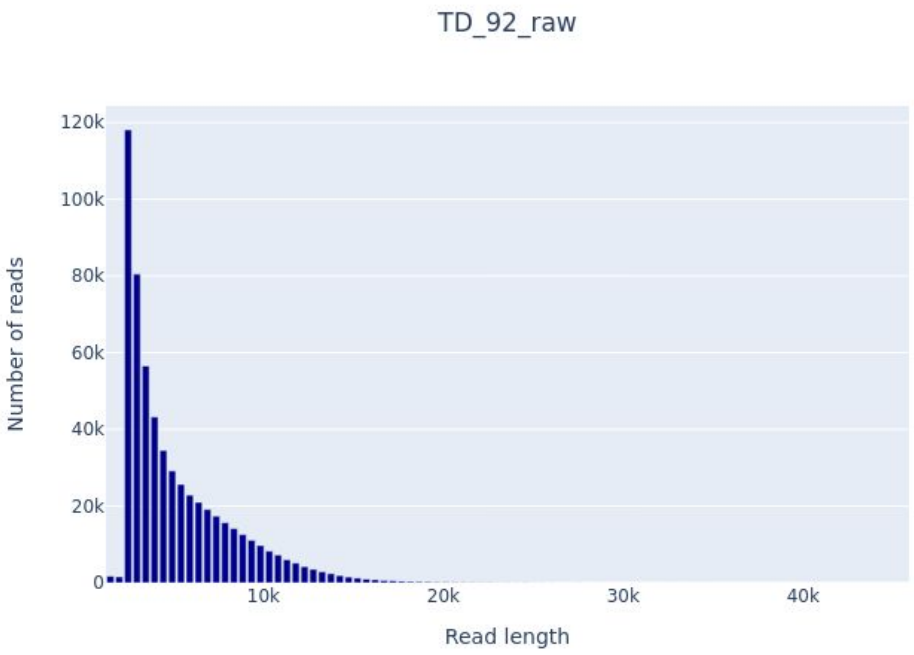
1. QUALITY CONTROL

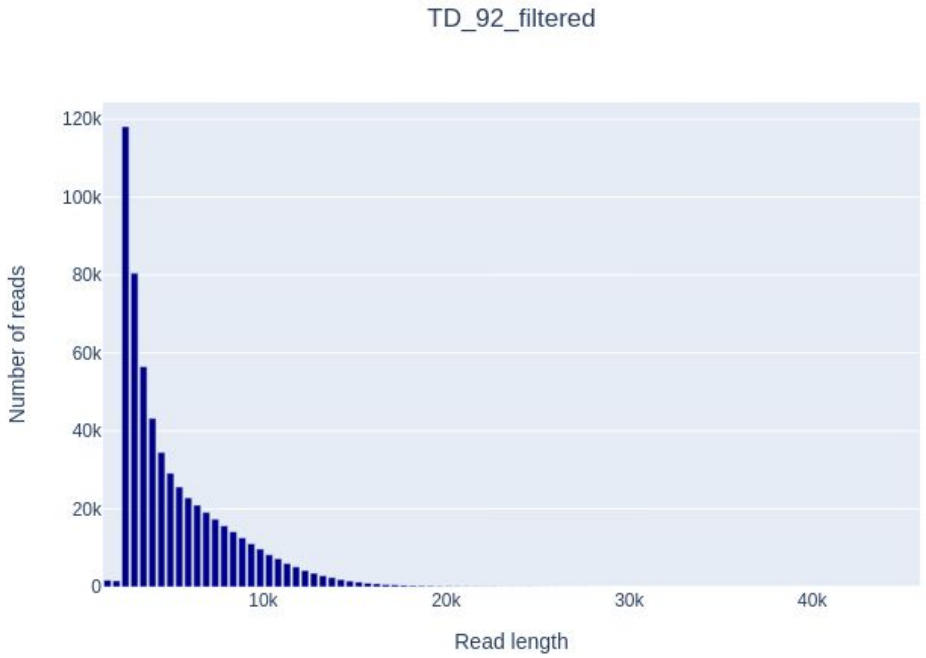
- Quality Assessment
  - Before Trimming: Assess raw read quality (e.g., Phred scores).
  - After Trimming: Re-evaluate trimmed reads for quality improvements.
- Trimming and Filtering
  - Remove Adapters: Trim adapter sequences from reads.
  - Filter Low-Quality Bases: Remove bases with low-quality scores.
  - Filter Short Reads: Discard reads below a specified length.
- Decontamination
  - Remove Host DNA: Eliminate host DNA sequences.
  - Correct Sequencing Errors: Improve read accuracy by correcting miscalled bases.

Quality control



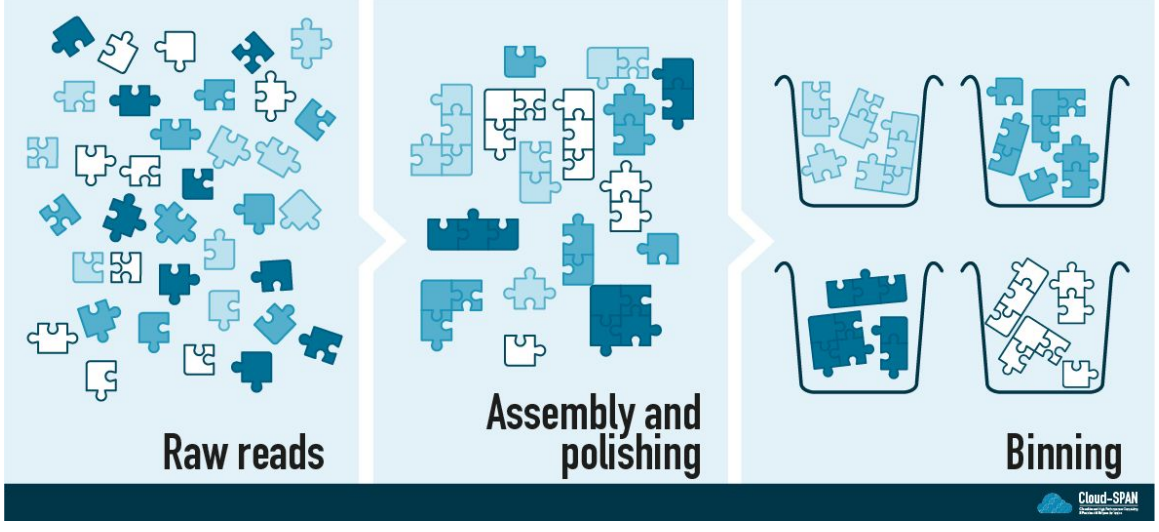
Quality control



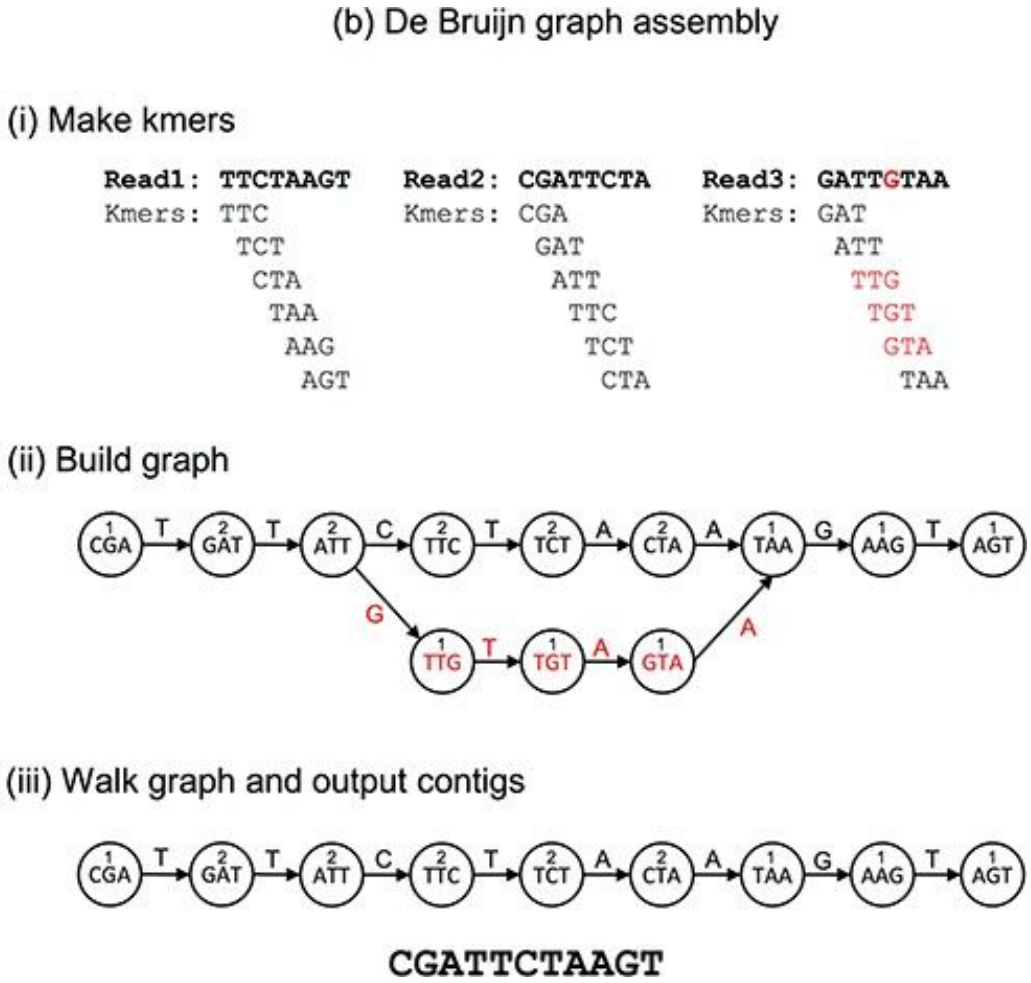


2. ASSEMBLY: Reconstruct longer genomic fragments

- Challenges in Metagenomic Assembly:
  - Uneven Coverage: Variability in coverage across different species; conserved regions complicate assembly.
  - Strain-Level Variations: Differences between strains of the same species.
  - Repetitive Elements: Presence of repetitive sequences can hinder accurate assembly.
- Assembly Strategies
  - De Novo Assembly
  - Assembly Quality Assessment
  - Binning: Group contigs into putative genomes



SPAdes Genome Assembler



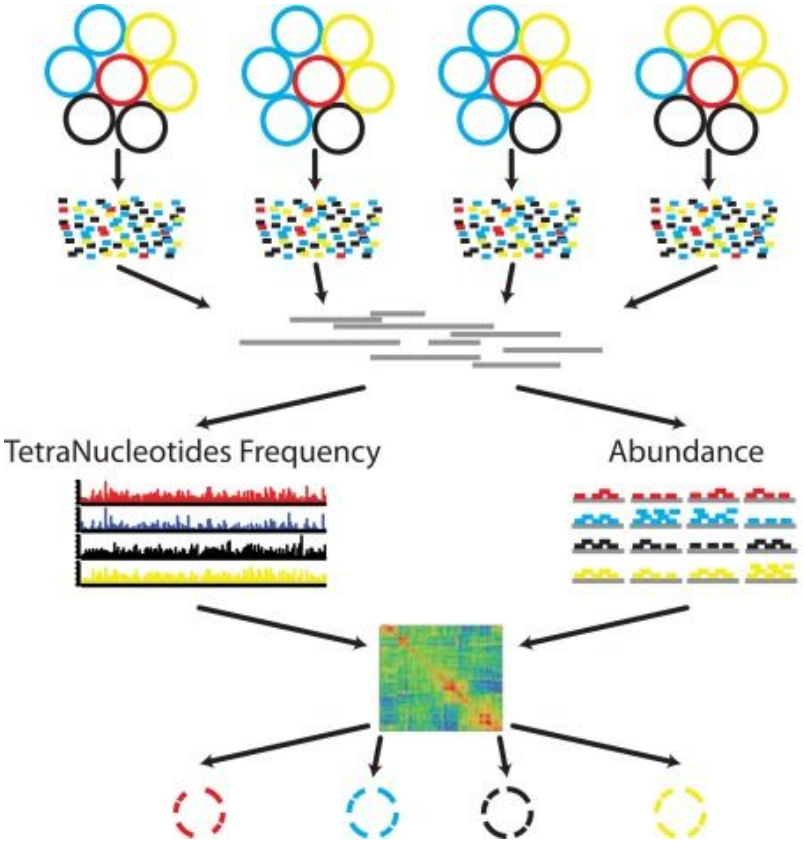
Martin Ayling, Matthew D Clark, Richard M Leggett, New approaches for metagenome assembly with short reads, *Briefings in Bioinformatics,* Volume 21, Issue 2, March 2020, Pages 584–594, https://doi.org/10.1093/bib/bbz020

Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19(5):455-477. doi:10.1089/cmb.2012.0021

QUAST for contigs

Assembly	CD_7_short.contigs	CD_7_short.scaffolds	CD_7_long	TD_31_short	TD_31_short.scaffolds	TD_31_long	TD_92_short	TD_92_short.scaffolds	TD_92_long
# contigs (>= 0 bp)	352497	348091	3430	229053	225253	2996	136094	133897	1750
# contigs (>= 1000 bp)	35521	34893	3408	28789	28175	2977	19763	19094	1742
# contigs (>= 5000 bp)	5869	5924	3320	5413	5360	2839	3879	3841	1651
# contigs (>= 10000 bp)	2600	2661	2934	2660	2716	2407	2048	2084	1391
# contigs (>= 25000 bp)	787	835	1738	910	939	1293	940	968	798
# contigs (>= 50000 bp)	276	293	877	351	368	666	490	503	443
Total length (>= 0 bp)	265543681	265641333	194584883	220188265	220277443	159179728	169341946	169411224	106053939
Total length (>= 1000 bp)	155166798	156498690	194571080	147726318	148887068	159166732	128183543	128793310	106048669
Total length (>= 5000 bp)	97131210	99950938	194245324	102441551	104838575	158686661	96413741	98505521	105725045
Total length (>= 10000 bp)	74532639	77396318	191274569	83385709	86547617	155381648	83799568	86358980	103729294
Total length (>= 25000 bp)	46818895	49422186	171171254	56895322	59488372	137283925	66570189	68911734	93837056
Total length (>= 50000 bp)	29151419	30713587	140563593	37653832	39851956	115167055	50485398	52341712	81276403
# contigs	85334	84489	3426	65996	65090	2995	39368	38498	1750
Largest contig	529830	755834	2584321	565327	816179	2342557	1314428	1314428	2229045
Total length	189046883	190225872	194583122	173202607	174179833	159179245	141689831	142166149	106053939
GC (%)	46.63	46.64	45.46	45.92	45.92	45.11	43.9	43.9	42.74
N50	5405	5788	118255	8863	9820	142056	20197	22907	173207
N90	715	718	22293	787	794	19835	1035	1045	22330
auN	33120.8	35692.2	368224.9	41572.5	47205.6	321501.9	65353.1	68282.9	402776.2
L50	5367	5024	307	3002	2771	237	1131	1060	114
L90	53364	52378	1905	38006	37035	1562	19112	18269	867
# N's per 100 kbp	0	51.28	0	0	51.05	0	0	49.25	0
# predicted rRNA genes	38 + 17 part	39 + 18 part	732 + 25 part	31 + 26 part	31 + 26 part	456 + 42 part	36 + 20 part	36 + 20 part	319 + 26 part

MetaBAT



Preprocessing

1. Samples from multiple sites or times
2. Metagenome libraries
3. Initial de-novo assembly using the combined library

MetaBAT

4. Calculate TNF for each contig
5. Calculate Abundance per library for each contig
6. Calculate the pairwise distance matrix using pre-trained probabilistic models
7. Forming genome bins iteratively

Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ.* 2015;3:e1165. Published 2015 Aug 27. doi:10.7717/peerj.1165

CheckM - assessing the quality of genome bins

TD_92_long/binning/bin_refinement/binsA.stats

bin	completeness		contamination		GC	lineage	N50	size	binner
bin.10	0.0	0.0	0.268	root	37144	248280	binsA		
bin.11	73.66	0.0	0.464	Bacteroidales	1031723	3054580	binsA		
bin.12	97.60	0.227	0.593	Bifidobacteriaceae	2229045	2229045	binsA		
bin.1	33.01	0.0	0.416	Clostridiales	1104936	1647912	binsA		
bin.2	79.95	2.230	0.435	Bacteroidales	1008497	4241802	binsA		
bin.3	74.58	4.474	0.373	Clostridiales	654883	2711616	binsA		
bin.4	65.82	0.158	0.412	Clostridiales	660293	2538921	binsA		
bin.5	43.10	8.620	0.429	Bacteria	467667	3921431	binsA		
bin.6	81.89	102.5	0.419	Bacteria	152868	15164013	binsA		
bin.7	0.0	0.0	0.422	root	371188	395268	binsA		
bin.8	0.0	0.0	0.437	Bacteria	306292	354129	binsA		
bin.9	90.75	0.309	0.462	Bacteroidales	604225	3654929	binsA		
bin.unbinned	5.199	0.458	0.418	Bacteroides	41611	1037190	binsA		

3. TAXONOMIC CLASSIFICATION

Kraken: K-mer-based taxonomic classification that is fast and highly accurate.

- Collects k-mers from reference genomes and associates them with taxonomic IDs (LCA).
- Breaks query sequences into k-mers and looks them up in the database.
- Uses a weighted voting scheme for taxonomy assignment.
- Output: Provides taxonomic classifications and a summary report of taxonomic composition.

Bracken: Bayesian Reestimation of Abundance with Kraken

- Estimates species abundance in metagenomic samples, addressing Kraken's limitations.
- Improves species-level abundance estimates using a Bayesian algorithm to redistribute reads.
- Fast and efficient; works directly with Kraken output.

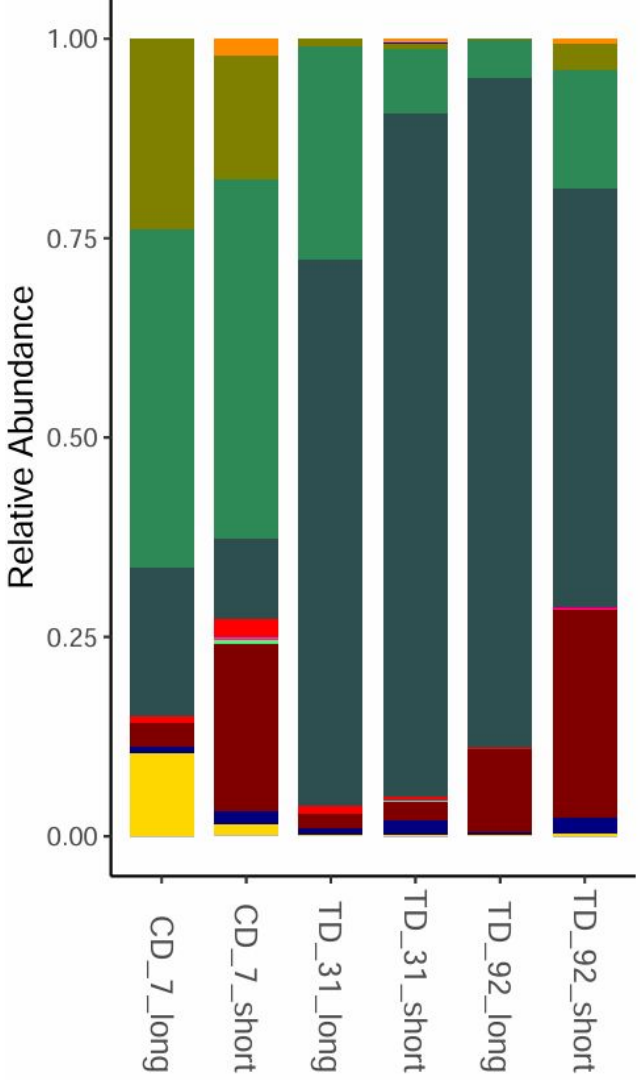
Collection	Contains	Date	Archive size (GB)	Index size (GB)	HTTPS URL	Inspect	Library	MD5
Viral	Refeq viral	6/5/2024	0.5	0.6	.tar.gz	.txt	.tsv	.md5
MinusB	Refeq archaea, viral, plasmid, human[1], UniVec_Core	6/5/2024	7.1	10.2	.tar.gz	.txt	.tsv	.md5
Standard	Refeq archaea, bacteria, viral, plasmid, human[1], UniVec_Core	6/5/2024	60	78	.tar.gz	.txt	.tsv	.md5
Standard-8	Standard with DB capped at 8 GB	6/5/2024	5.5	7.5	.tar.gz	.txt	.tsv	.md5
Standard-16	Standard with DB capped at 16 GB	6/5/2024	11	15	.tar.gz	.txt	.tsv	.md5
PlusPF	Standard plus Refeq protozoa & fungi	6/5/2024	64	83	.tar.gz	.txt	.tsv	.md5
PlusPF-8	PlusPF with DB capped at 8 GB	6/5/2024	5.5	7.5	.tar.gz	.txt	.tsv	.md5
PlusPF-16	PlusPF with DB capped at 16 GB	6/5/2024	11	15	.tar.gz	.txt	.tsv	.md5
PlusPFP	Standard plus Refeq protozoa, fungi & plant	6/5/2024	135	182	.tar.gz	.txt	.tsv	.md5
PlusPFP-8	PlusPFP with DB capped at 8 GB	6/5/2024	5.1	7.5	.tar.gz	.txt	.tsv	.md5
PlusPFP-16	PlusPFP with DB capped at 16 GB	6/5/2024	11	15	.tar.gz	.txt	.tsv	.md5
EuPathDB46[2]	Eukaryotic pathogen genomes with contaminants removed	4/18/2023	8.4	11	.tar.gz	.txt	N/A	N/A
nt Database	Very large collection, inclusive of GenBank, RefSeq, TPA and PDB	5/30/2024	684	889	.tar.gz	.txt	.tsv	.md5

Kraken 2 / Bracken database

https://benlangmead.github.io/aws-indexes/k2

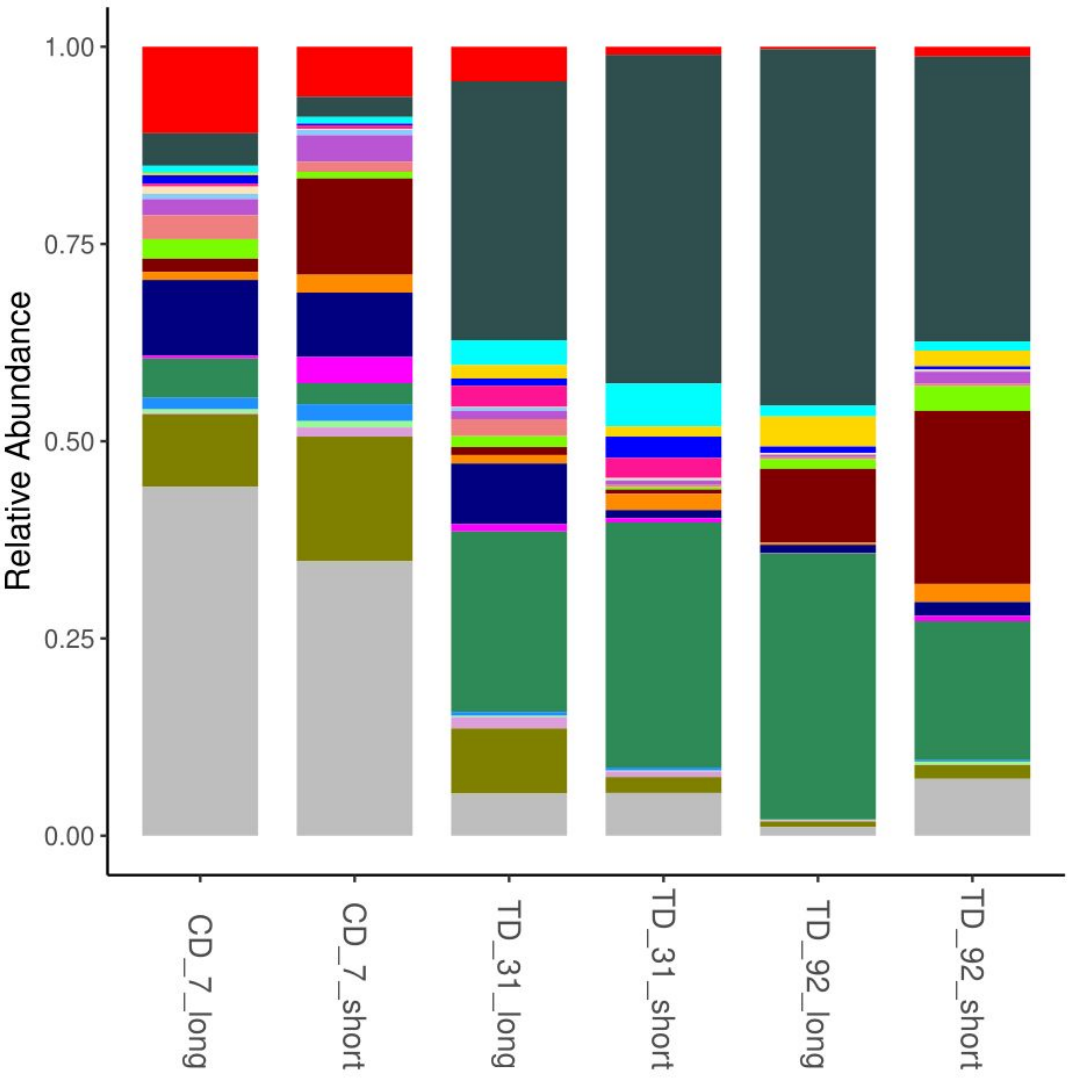
Lu J, Rincon N, Wood DE, Breitwieser FP, Pockrandt C, Langmead B, Salzberg SL, Steinegger M. (2022) Metagenome analysis using the Kraken software suite.. Nature Protocolsdoi: 10.1038/s41596-022-00738-y

Relative abundance by Phylum



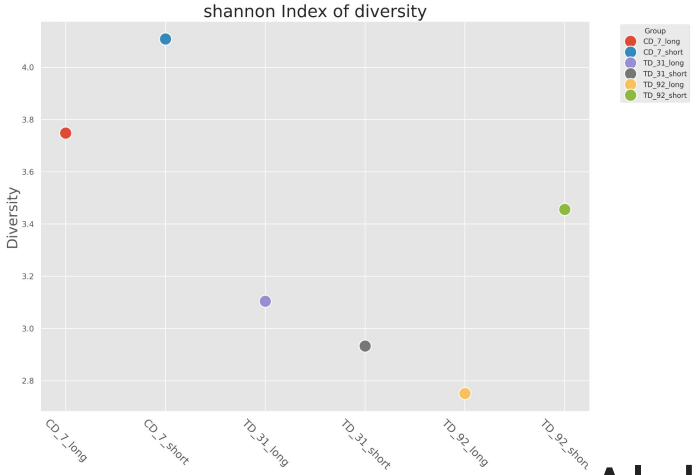
- d__Archaea|p__Candidatus_Lokiarchaeota
- d__Archaea|p__Candidatus_Thermoplasmatota
- d__Archaea|p__Euryarchaeota
- d__Archaea|p__Nitrososphaerota
- d__Archaea|p__Thermoproteota
- d__Bacteria|p__Actinomycetota
- d__Bacteria|p__Bacillota
- d__Bacteria|p__Bacteroidota
- d__Bacteria|p__Campylobacterota
- d__Bacteria|p__Cyanobacteriota
- d__Bacteria|p__Deinococcota
- d__Bacteria|p__Fusobacteriota
- d__Bacteria|p__Mycoplasmatota
- d__Bacteria|p__Planctomycetota
- d__Bacteria|p__Pseudomonadota
- d__Bacteria|p__Spirochaetota
- d__Bacteria|p__Thermodesulfobacteriota
- d__Eukaryota|k__Metazoa|p__Chordata
- d__Viruses|k__Bamfordvirae|p__Nucleocytoviricota
- d__Viruses|k__Heunggongvirae|p__Uroviricota
- other

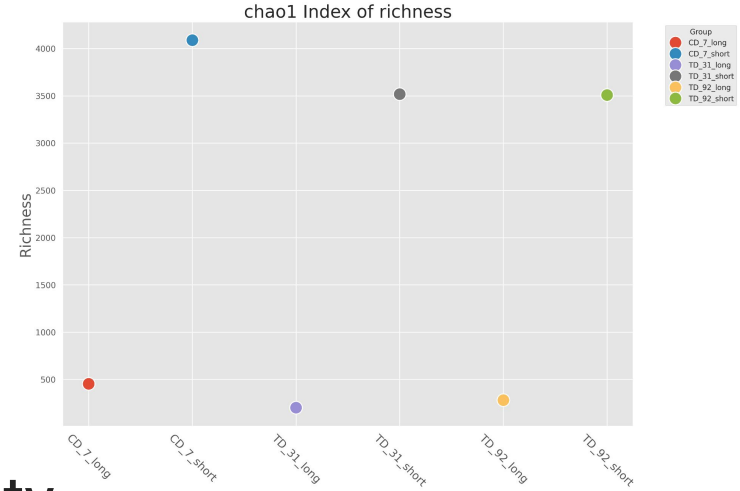
Relative abundance by Species

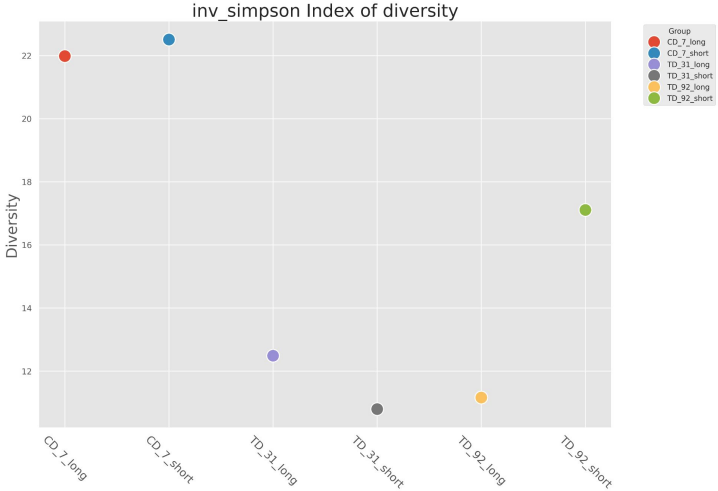


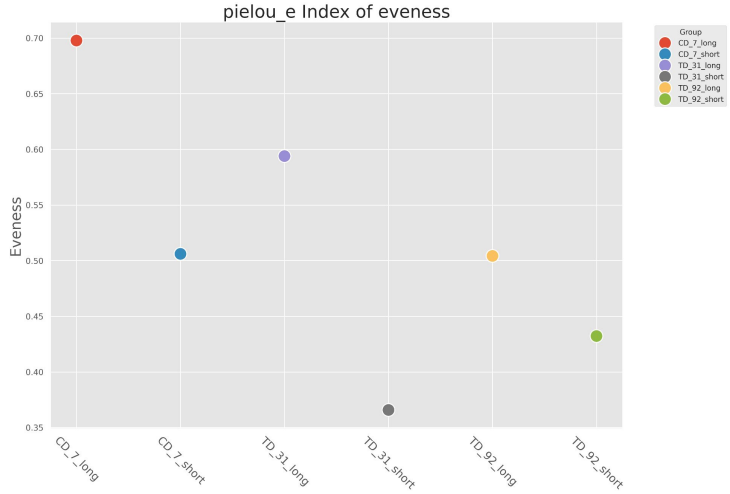
- Agathobacter_rectalis
- Bacteroides_finegoldii
- Bacteroides_fragilis
- Bacteroides_ovatus
- Bacteroides_sp._CACC_737
- Bacteroides_thetaiotaomicron
- Bacteroides_uniformis
- Campylobacter_coli
- Clostridioides_difficile
- Clostridium_perfringens
- Enterococcus_faecium
- Escherichia_coli
- Homo_sapiens
- Lachnospira_eligens
- Monoglobus_pectinilyticus
- Parabacteroides_distasonis
- Pusillibacter_faecalis
- Roseburia_sp._831b
- Segatella_copri
- Vescimonas_fastidiosa
- other

Alpha diversity









Beta diversity