Long-reads Binning For Microbial Metagenomics Considering Multi-kingdoms

Group 11

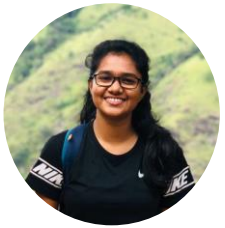
Supervisors

Dr. Damayanthi Herath
Senior Lecturer
Department of Computer Engineering

Dr. Vijini Mallawaarachchi
Postdoctoral Research Associate
Flinders University, Australia

Team Members

E/18/030
Sathsarani



E/18/282
Nethmi

E/18/283
Jayathri

Outline

- Domain and Background
- Existing Tools for Metagenomic Binning
- Identifying the Research Gap
- Proposed Workflow
- Impact
- Use of AI Tools
- Demonstration

Understanding Metagenomics

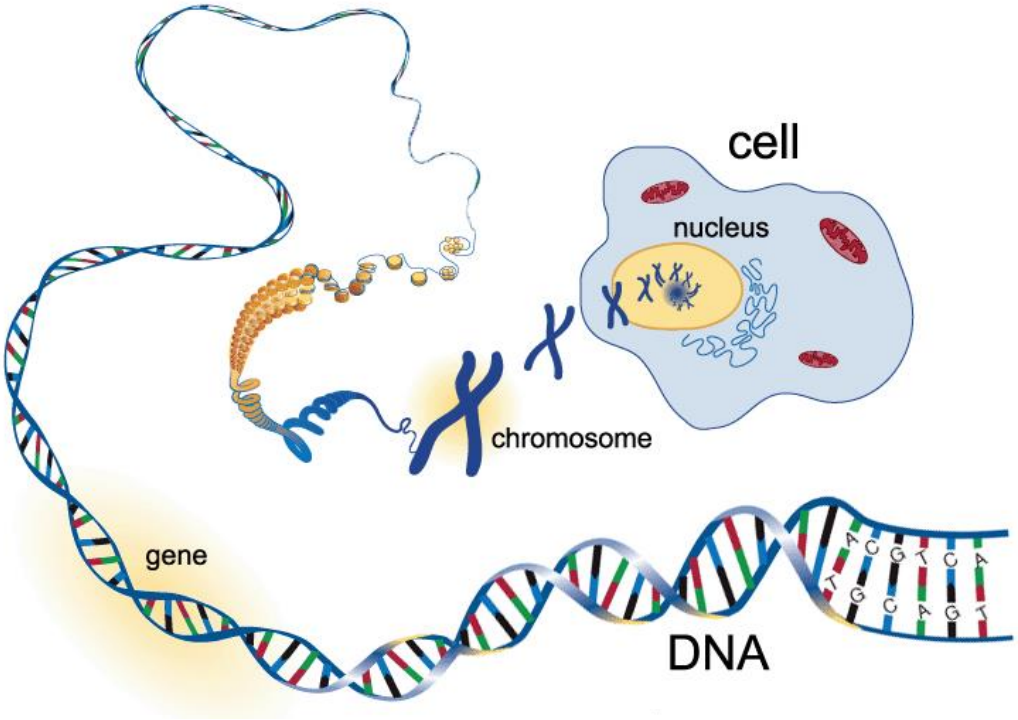
Genome - The complete set of genetic material present in an organism
Ex:- Human genome, Bacterial genome

4 nucleotide bases:

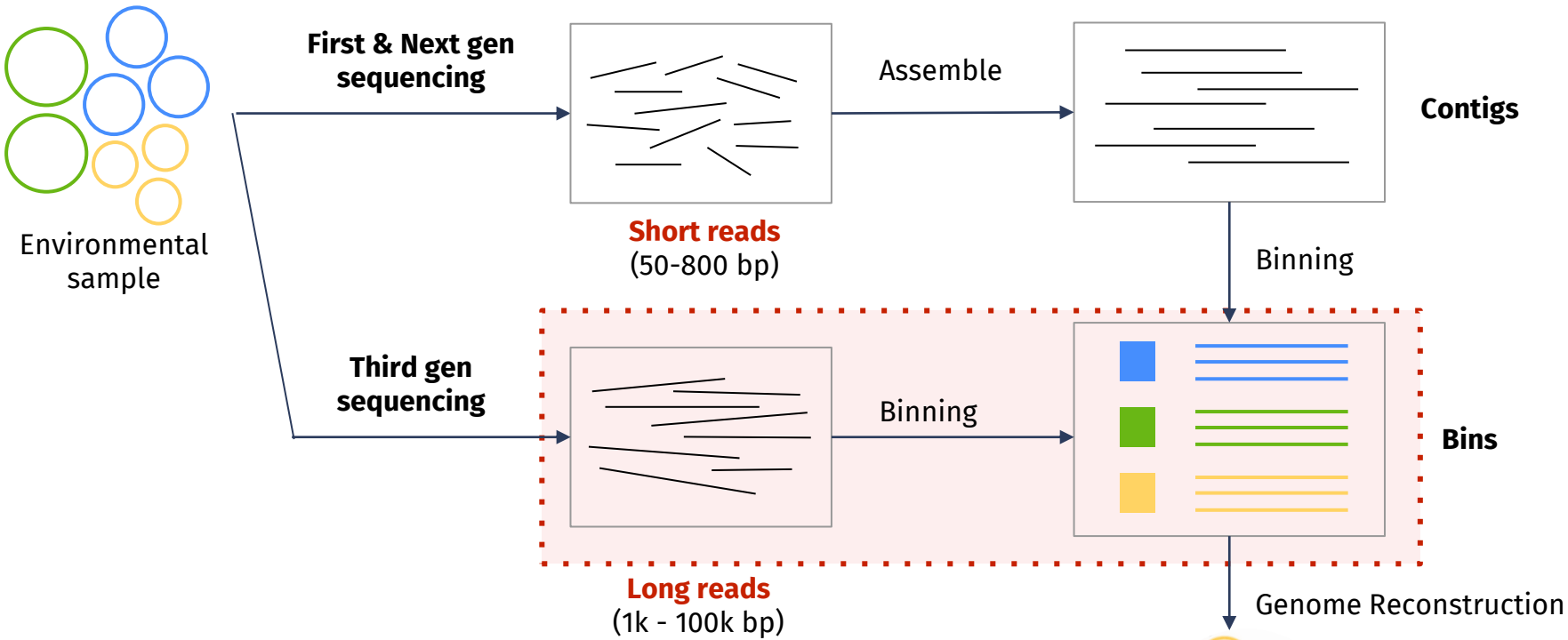
Adenine, Guanine, Cytosine, Thymine

DNA sequencing - Extracting the long strings of genetic material into readable lengths

...CCTTACTTATAATGCTCATGCTA...



Background



Binning - Grouping reads with similar species together

Binning in Metagenomics

Features considered to cluster the reads in to bins;

1) **Composition** - normalized frequency of short substrings of a particular read

Oligonucleotide Frequency

AATCGGCAATCGAATGCCG

Trinucleotide Composition of AAT = $\frac{\text{No. of specific k-mer}}{\text{Total no. of k-mers}}$

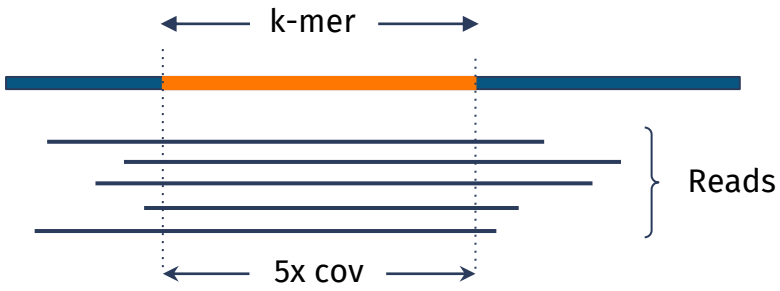
= 3/14

Binning in Metagenomics

2) **Coverage** - number of reads that overlaps with a specific region in a reference genome

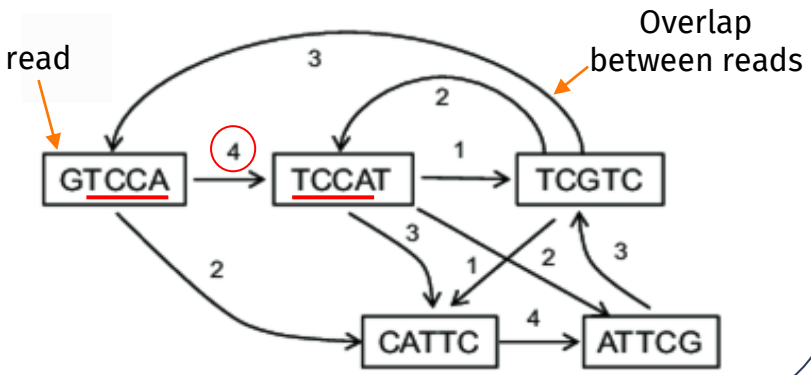
K-mer Coverage

- K-mer is a substring of length k



Read Overlap Graph

- Node degree represents the coverage of respective reads



Existing Long-reads Binning Tools

Tool	Feature Extraction Methods		Clustering Algorithms
	Composition	Coverage	
MetaBCC-LR	Trinucleotide frequency profiles	K-mer coverage histogram	DBSCAN (density based ML clustering algorithm)

Wickramarachchi, A., Mallawaarachchi, V., Rajan, V., & Lin, Y. (2020). Metabcc-lr: meta genomics b inning by c overage and c omposition for l ong r eads. Bioinformatics, 36(Supplement_1), i3-i11.

Existing Long-reads Binning Tools

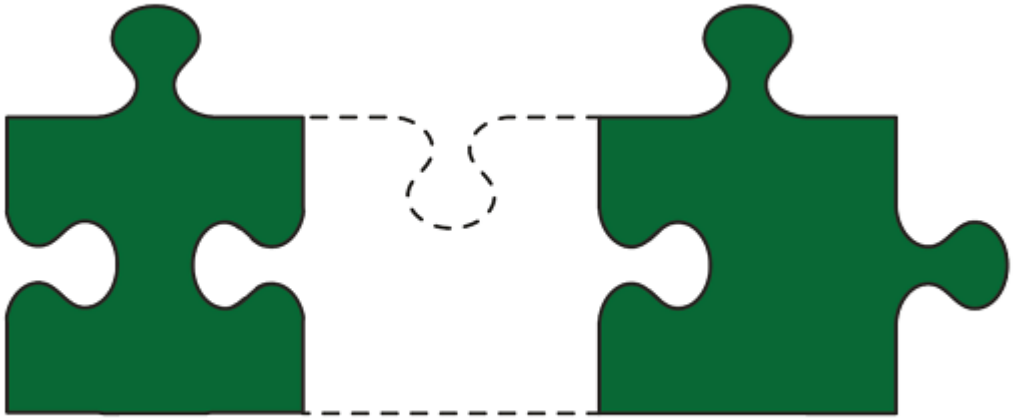
Tool	Feature Extraction Methods		Clustering Algorithms
	Composition	Coverage	
MetaBCC-LR	Trinucleotide frequency profiles	K-mer coverage histogram	DBSCAN (density based ML clustering algorithm)
LRBinner	Trinucleotide composition vector	k-mer coverage vector	Distance based statistical grouping technique

Wickramarachchi, A., & Lin, Y. (2021). LRBinner: binning long reads in metagenomics datasets. In *21st International Workshop on Algorithms in Bioinformatics (WABI 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Existing Long-reads Binning Tools

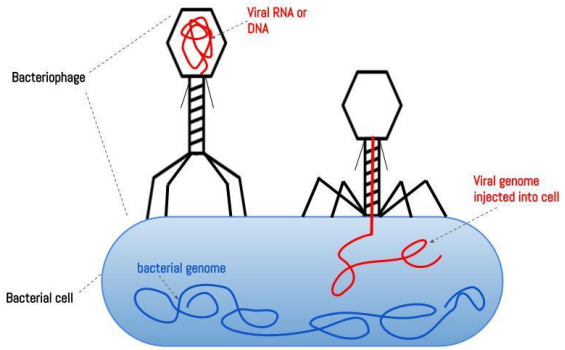
Tool	Feature Extraction Methods		Clustering Algorithms
	Composition	Coverage	
MetaBCC-LR	Trinucleotide frequency profiles	K-mer coverage histogram	DBSCAN (density based ML clustering algorithm)
LRBinner	Trinucleotide composition vector	k-mer coverage vector	Distance based statistical grouping technique
OBLR	Tetranucleotide frequency vector	Node degree of the Read overlap graph	HDBSCAN (density based hierarchical ML clustering algorithm)

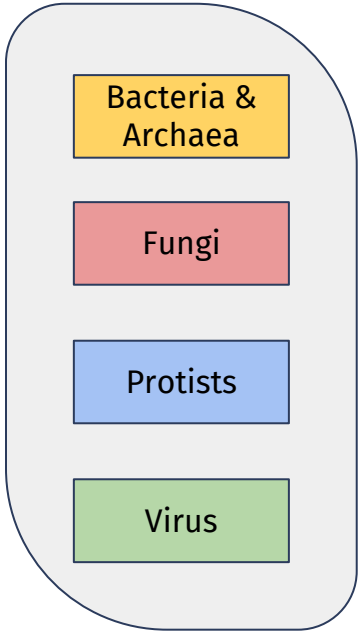
Wickramarachchi, A., & Lin, Y. (2022, May). Metagenomics binning of long reads using read-overlap graphs. In *RECOMB International Workshop on Comparative Genomics* (pp. 260-278). Cham: Springer International Publishing.



Bridging the Gap

Consideration of Multi-kingdom





Species of different kingdoms can present in the same sample.

Ex: Bacteriophage viral Infection

Marker genes

A gene or DNA sequence with a known location on a chromosome that can be used to identify individuals or species

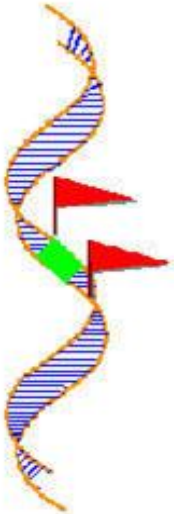
Single copy marker gene: marker genes that occur only once in almost every genome

Bacteria & Archaea

Fungi

Protists

} Single-copy marker genes

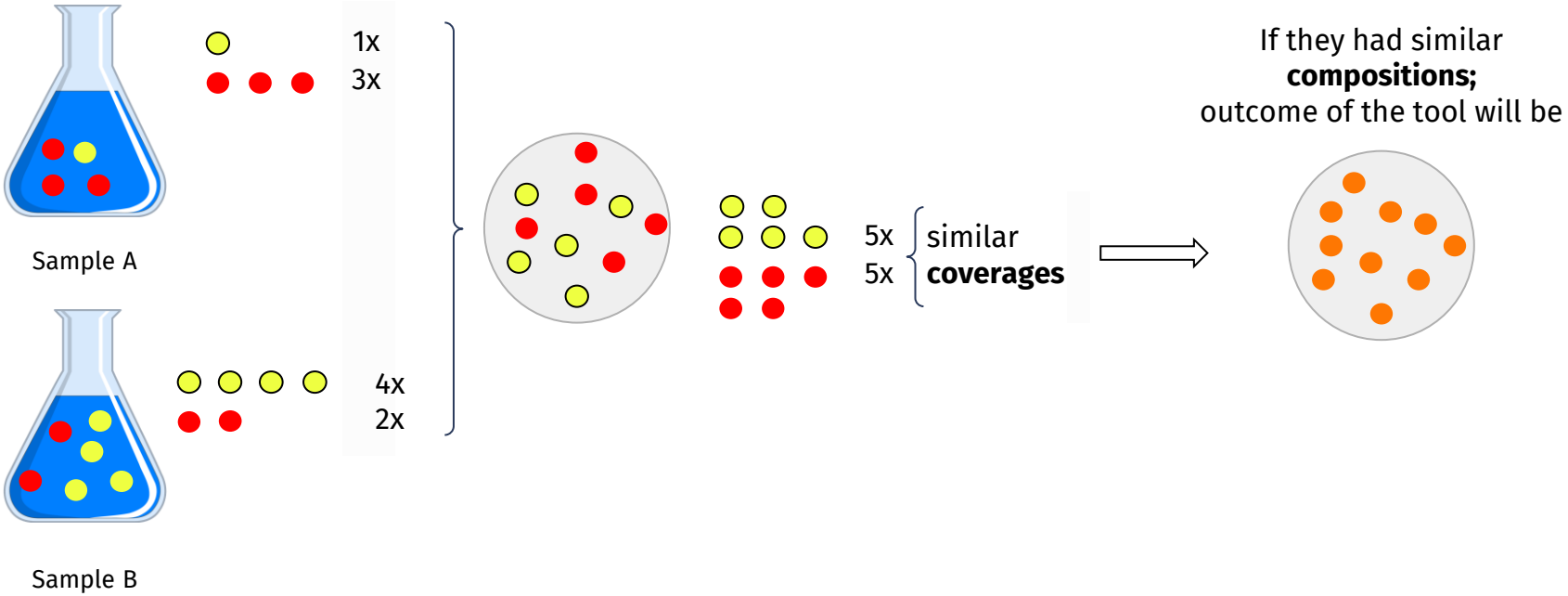


Viruses: Orthologous gene sequences
specialized databases:

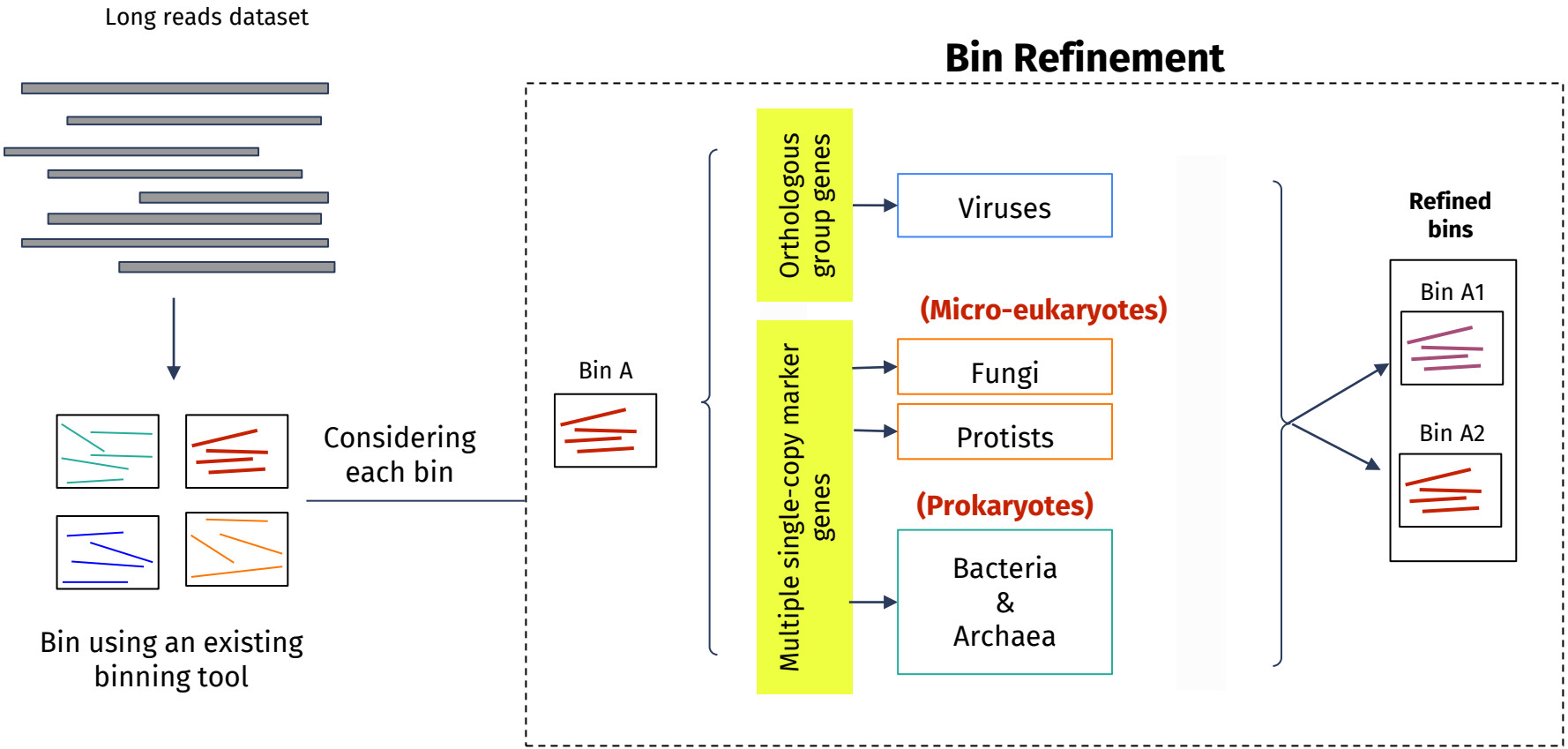
- VOG (for all viruses)
- PHROG (for proviruses)

Consideration of Multiple Samples

★ **Differential abundance**: variation in abundance of different species across different samples



Proposed Workflow



Impact

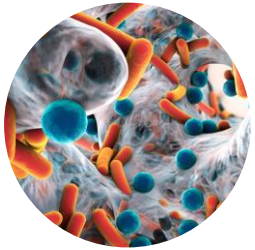
Refined, accurate metagenomic bins

Improve the quality of the assembly process

Better genome reconstruction

Novel Medicine and disease intervention

Precision Agriculture and food security



Understanding microbial biodiversity

Advanced research

Use of AI tools

Gather information on the domain and existing work.

- Generating bias and inaccurate information.
- Knowledge was not up to date.

To search research articles.

- Most of the available features were not for free.

To assist in the process of writing and to generate citations.

- Limited understanding on the context.

Demonstration

Structure of the datasets

Fastq format

```
@m4/6272/CCS Read=4;length=6272bp;startpos=2033937;chromosome=Staphylococcus_aureus_chromosome;numberOfErrors=118;totalErrorProb=0.0177;passes=3.980537
TTCAACTTTATTTCGCAGTTGGTAGATTGTACACTTTCTTCGTTTCTATTAATTTATTCCCATTGAAGTCGTTTCATTACCTTGAGATGGATACCATTTCTTTTTGATTATCATTTTTAGTATTGTCTTCTTGATTTAGTTGCTGCATATC
+
11111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111133333333333333333333333333333
```

Line 1: Begins with @ followed by read ID

Line2: Nucleotide sequence

Line3: Begins with + and is often optional.It is the separator between sequence and the quality score

Line4: Quality scores corresponding to each base in the sequence

Fasta format

- No quality score data

Simulating reads data

	Bacillus_subtilis_complete_genome.fasta	FASTA File	3,951 KB
	Enterococcus_faecalis_complete_genome.fasta	FASTA File	2,779 KB
	Escherichia_coli_complete_genome.fasta	FASTA File	4,762 KB
	Listeria_monocytogenes_complete_genome.fasta	FASTA File	2,923 KB
	Pseudomonas_aeruginosa_complete_genome.fasta	FASTA File	6,634 KB
Fungi	Saccharomyces_cerevisiae_draft_genome.fasta	FASTA File	12,674 KB
	Salmonella_enterica_complete_genome.fasta	FASTA File	4,649 KB
	Staphylococcus_aureus_complete_genome.fasta	FASTA File	2,667 KB

```
1163160979 Feb  9 16:41 testDataset.fastq
```

```
simlord --read-reference reference.fasta -n 10000 -fl 5000 -pi 0.12 -pd 0.12 -ps 0.02 testDataset
```

Reference genome

Number of Reads

Fixed length of read

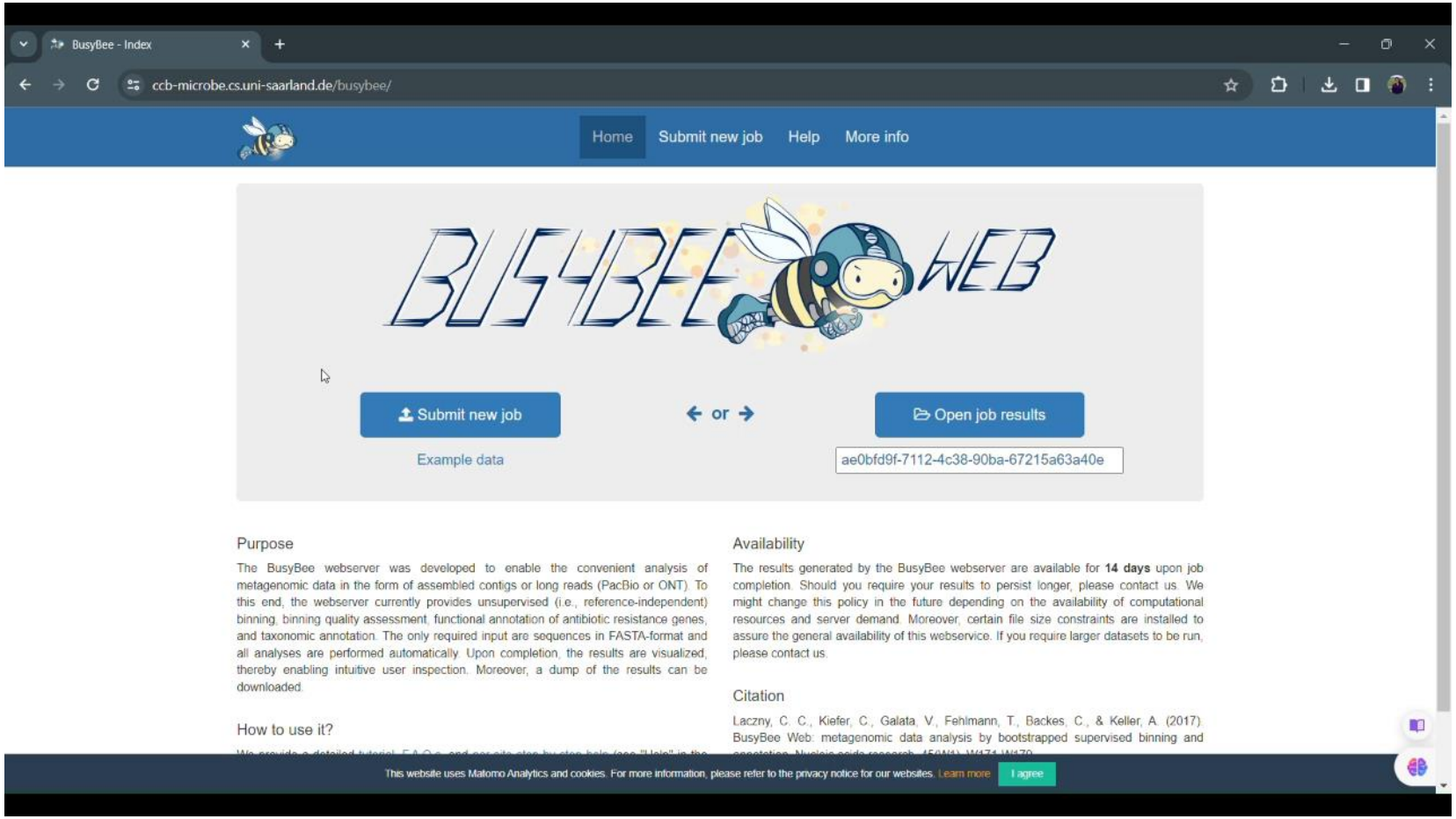
Custom subread error probabilities

Running long-reads binning tools

1. BusyBee Web - Web based tool
2. MetaBCC-LR
3. LRBinner

} Command line based python tools

1) BusyBee Web



2) MetaBCC-LR

Step 1: Count K-mers

```
(metabcc-lr) e18030@ampere:/storage/scratch1/e18-4yp-multi-kingdom-binning/binningTools/MetaBCC-LR$ python mbcclr --resume -r ../../dataset
s/zymo_hmw_r941.fastq -o MetaBCC_output -e umap -c 25000 -bs 10 -bc 10 -k 4
2024-02-09 21:19:49,693 - INFO - Command mbcclr --resume -r ../../datasets/zymo_hmw_r941.fastq -o MetaBCC_output -e umap -c 25000 -bs 10 -b
c 10 -k 4
2024-02-09 21:19:49,693 - INFO - Resuming the program from previous checkpoints
2024-02-09 21:19:49,693 - INFO - Counting K-mers
INPUT FILE ../../datasets/zymo_hmw_r941.fastq
OUTPUT FILE MetaBCC_output/profiles/3mers
K SIZE 4
THREADS 8
Profile Size 136
Total 4-mers 256
Loaded Reads 8851918
2024-02-10 01:07:12,778 - INFO - Counting K-mers complete
```

Step 2: Count 15-mers

```
2024-02-10 01:07:13,253 - INFO - Counting 15-mers
INPUT FILE ../../datasets/zymo_hmw_r941.fastq
OUTPUT FILE MetaBCC_output/profiles/15mers-counts
THREADS 8
Loaded Reads 8851918
WRITING TO FILE
COMPLETED : Output at - MetaBCC_output/profiles/15mers-counts
2024-02-10 07:19:20,530 - INFO - Counting 15-mers complete
```

2) MetaBCC-LR

Step 3: Generate 15-mer profiles

```
2024-02-10 07:19:20,906 - INFO - Generating 15-mer profiles
K-Mer file MetaBCC_output/profiles/15mers-counts
LOADING KMERS TO RAM
FINISHED LOADING KMERS TO RAM
INPUT FILE ../../datasets/zymo_hmw_r941.fastq
OUTPUT FILE MetaBCC_output/profiles/15mers
THREADS 8
BIN WIDTH 10
BINS IN HIST 10
Loaded Reads 8851918
COMPLETED : Output at - MetaBCC_output/profiles/15mers
2024-02-10 10:37:23,247 - INFO - Generating 15-mer profiles complete
```

Step 4: Sampling reads

```
2024-02-10 10:37:23,342 - INFO - Sampling Reads
2024-02-10 11:43:33,794 - DEBUG - 3mer data shape (8851918, 136)
2024-02-10 11:43:33,805 - DEBUG - 15mer data shape (8851918, 10)
2024-02-10 11:43:33,805 - DEBUG - Sampling count 25000
2024-02-10 11:43:35,086 - INFO - Sampling reads complete
```

2) MetaBCC-LR

Step 5: Clustering using coverage and then composition (Small clusters are discarded)

```
2024-02-10 11:43:35,087 - INFO - Binning sampled reads
2024-02-10 11:43:35,376 - DEBUG - Clustering using coverage
2024-02-10 11:44:43,894 - DEBUG - Identified number of coverage clusters - 1
2024-02-10 11:44:43,895 - DEBUG - Clustering using composition
2024-02-10 11:44:43,895 - DEBUG - Discarding small clusters (< 500 reads in the sampled set)
2024-02-10 11:46:38,201 - DEBUG - Identified number of coverage and composition clusters - 4
2024-02-10 11:46:38,201 - DEBUG - Discarding small clusters
2024-02-10 11:46:38,212 - INFO - Binning sampled reads complete
```

Step 6: Predict read bins

```
2024-02-10 11:46:38,212 - INFO - Predict read bins
3 Mers MetaBCC_output/profiles/3mers
15 Mers MetaBCC_output/profiles/15mers
Stats MetaBCC_output/misc/cluster-stats.txt
Threads 8
Bins size = 3
2024-02-10 11:52:05,026 - INFO - Predict read bins complete
2024-02-10 11:52:05,026 - INFO - Program Finished!. Please find the output in MetaBCC_output/final.txt
```

3) LRBinner

Step 1: Counting the reads

```
2024-02-10 09:04:41,009 - INFO - Command lrbinner.py reads -r ./../../datasets/zymo_hmw_r941.fastq -bc 10 -bs 32 -o ./../results/lrb_result --resume
--cuda -mbs 5000 --ae-dims 4 --ae-epochs 200 -bit 0 -t 32
2024-02-10 09:04:41,021 - INFO - CUDA found in system
2024-02-10 09:04:41,022 - INFO - Resuming the program from previous checkpoints
2024-02-10 09:04:41,022 - INFO - Counting k-mers
INPUT FILE ./../../datasets/zymo_hmw_r941.fastq
OUTPUT FILE ./../results/lrb_result/profiles/com_profs
K_SIZE 3
THREADS 32
Profile Size 32
Total 3-mers 64
Loaded Reads 8851918
```

Step 2: Counting k-mers (k=3) and 15-mers

```
2024-02-10 11:46:45,359 - INFO - Counting k-mers complete
2024-02-10 11:46:45,363 - INFO - Counting 15-mers
INPUT FILE ./../../datasets/zymo_hmw_r941.fastq
OUTPUT FILE ./../results/lrb_result/profiles/15mers-counts
THREADS 32
Loaded Reads 8851918
WRITING TO FILE
COMPLETED : Output at - ./../results/lrb_result/profiles/15mers-counts
2024-02-10 11:56:10,694 - INFO - Counting 15-mers complete
2024-02-10 11:56:10,703 - INFO - Computing 15-mer profiles
K-Mer file ./../results/lrb_result/profiles/15mers-counts
LOADING KMERS TO RAM
FINISHED LOADING KMERS TO RAM
INPUT FILE ./../../datasets/zymo_hmw_r941.fastq
OUTPUT FILE ./../results/lrb_result/profiles/cov_profs
```

3) LRBinner

Step 3: Training the VAE

```
2024-02-10 12:07:01,872 - INFO - Profiles saving as numpy arrays
2024-02-10 12:10:50,337 - INFO - Profiles saving as numpy arrays complete
2024-02-10 12:10:50,353 - INFO - VAE training information
2024-02-10 12:10:50,353 - INFO -        Dimensions 4
2024-02-10 12:10:50,354 - INFO -        Hidden Layers [128, 128]
2024-02-10 12:10:50,354 - INFO -        Epochs 200
Training VAE:  45%|                                                                                  | 90/200 [2:25:10<1:43:39, 56.54s/it]
```

Step 4: Clustering the reads

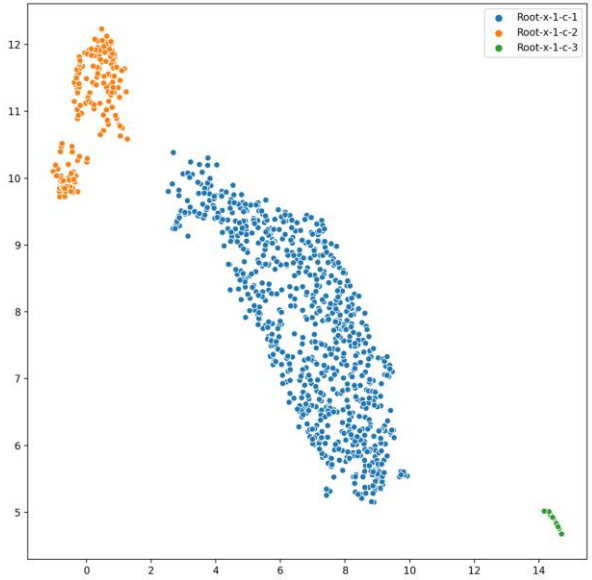
```
Training VAE: 100%|                                                                                  | 200/200 [3:52:32<00:00, 69.76s/it]
2024-02-10 16:04:42,895 - INFO - VAE training complete
2024-02-10 16:04:42,991 - INFO - Clustering algorithm running
Performing exhaustive search!: 100%|                                                                | 1280/1280 [00:02<00:00, 516.68it/s]
2024-02-10 16:06:47,957 - INFO - Detected 9 clusters
2024-02-10 16:06:48,678 - INFO - Detected 7 clusters with more than 5000 points
2024-02-10 16:06:48,678 - INFO - Building profiles
2024-02-10 16:08:47,412 - INFO - Binning unclassified reads
100%|                                                                                                | 6062/6062 [00:01<00:00, 5627.41it/s]
2024-02-10 16:08:48,490 - INFO - Binning complete with 7 bins
```

Performance comparison

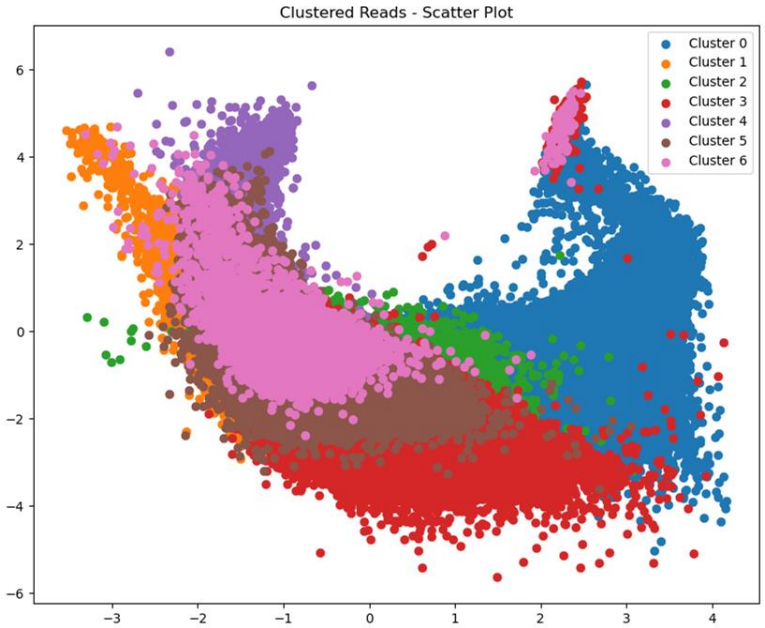
Tool	Result bin count	Time consumed (Hours)	Evaluation criteria			
			Precision	Recall	F1-score	ARI
MetaBCC-LR	3	1.07	47.52	92.23	62.73	28.29
LRBinner	7	4	67.74	94.17	78.8	63.64

Cluster Images

MetaBCC-LR



LRBinner



Thank You!

Q & A