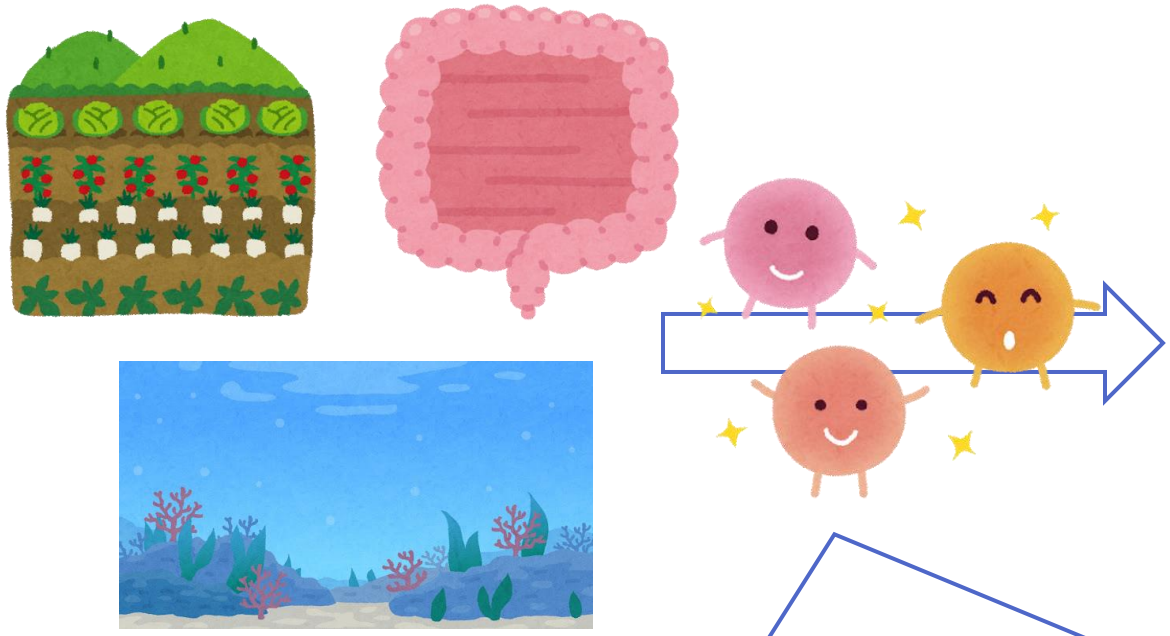
Development of data driven metatranscriptomic analysis

Hiroshima University

Ryo Mameda, Hidemasa Bono

Outline

Complex Microbiome



Benefits

- Plant growth (soil microbiome)
- Immune regulation (gut microbiome)
- Ecosystem maintenance (water microbiome)

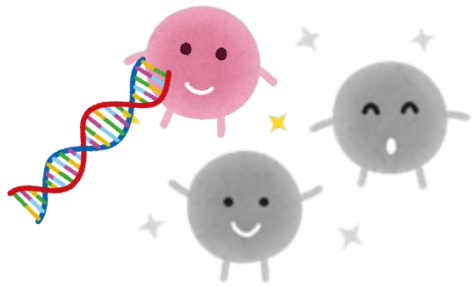
→ Evaluating microbial activity and maintaining microbiome are essential.
= Gene expression analysis

Specific bacterial activity
(nitrification, anti-microbial products…)

Comprehensive bacterial activity
(carbon metabolism…)

Many activities are based on microbial (and host) interactions.
→ It makes complicated to analyze its activity.
= Focusing on comprehensive activity.

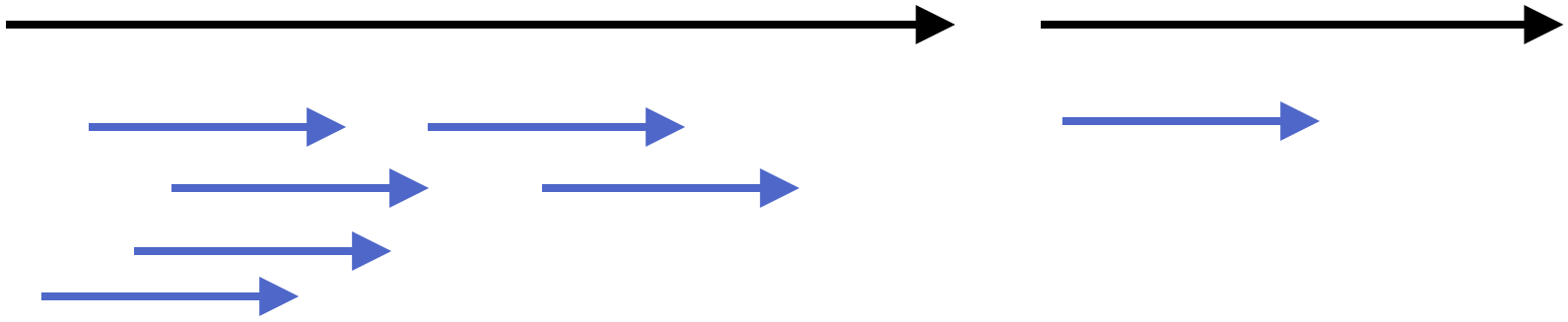
Gene Expression Analysis



Quantification of gene expression requires genome data.
However, only 2% of environmental bacterial genome are revealed.
→ Metagenomic contigs are widely used.
Predicted protein coding sequences are references for gene quantification.

Reference sequences
(metagenomic predicted protein coding sequences)

NGS reads
(metagenomic or metatranscriptomic reads)



Issue

① Mapping tool optimization
② Effects of abundant genes

Outline

Introduction

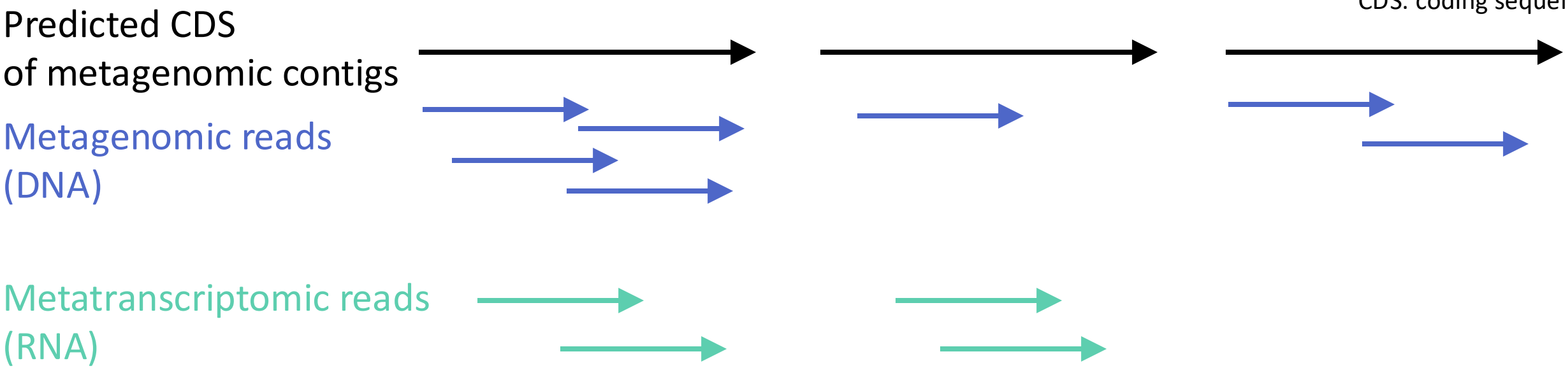
Isuue1: Mapping Tool Optimization

Isuue2: Effects of Abundant Genes

Conclusion

Mapping DNA and RNA

CDS: coding sequences



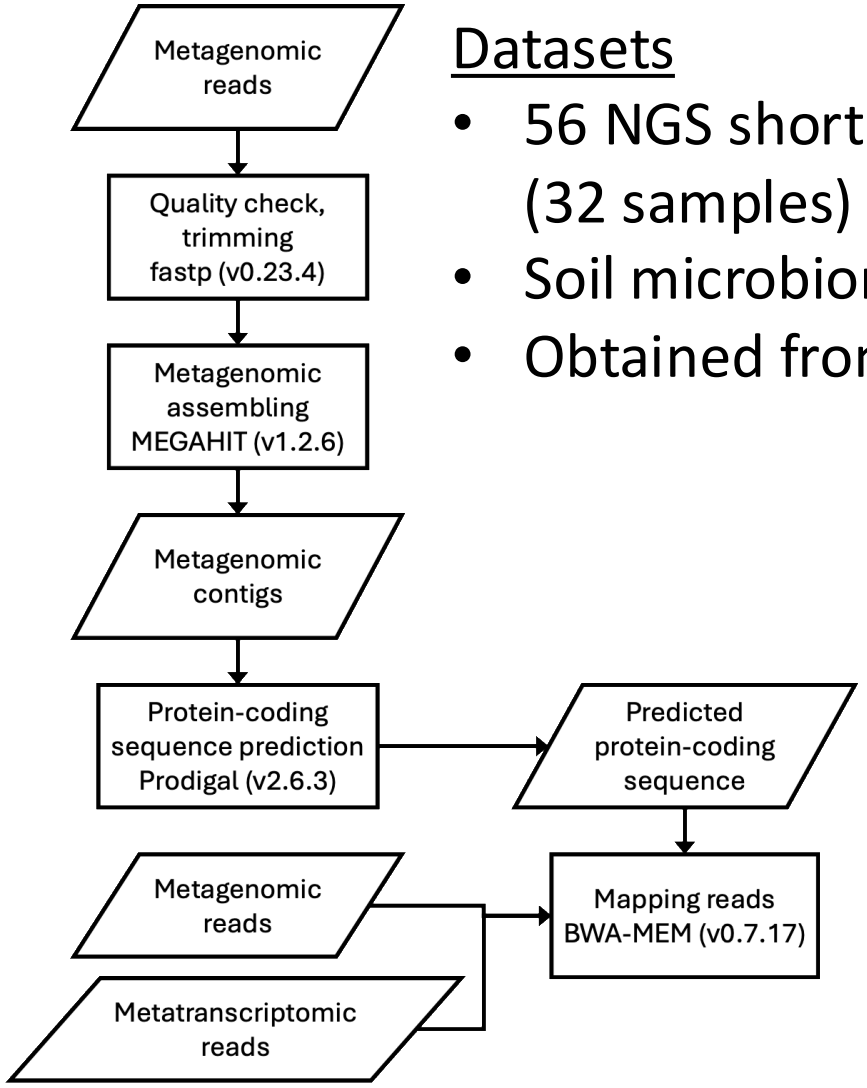
Expressed, but metagenomic reads are abundant.
= Low expression level

Similar RNA abundance, but few metagenomic reads are detected.
= High expression level

DNA are detected, but RNA are absent.
= Unexpressed genes

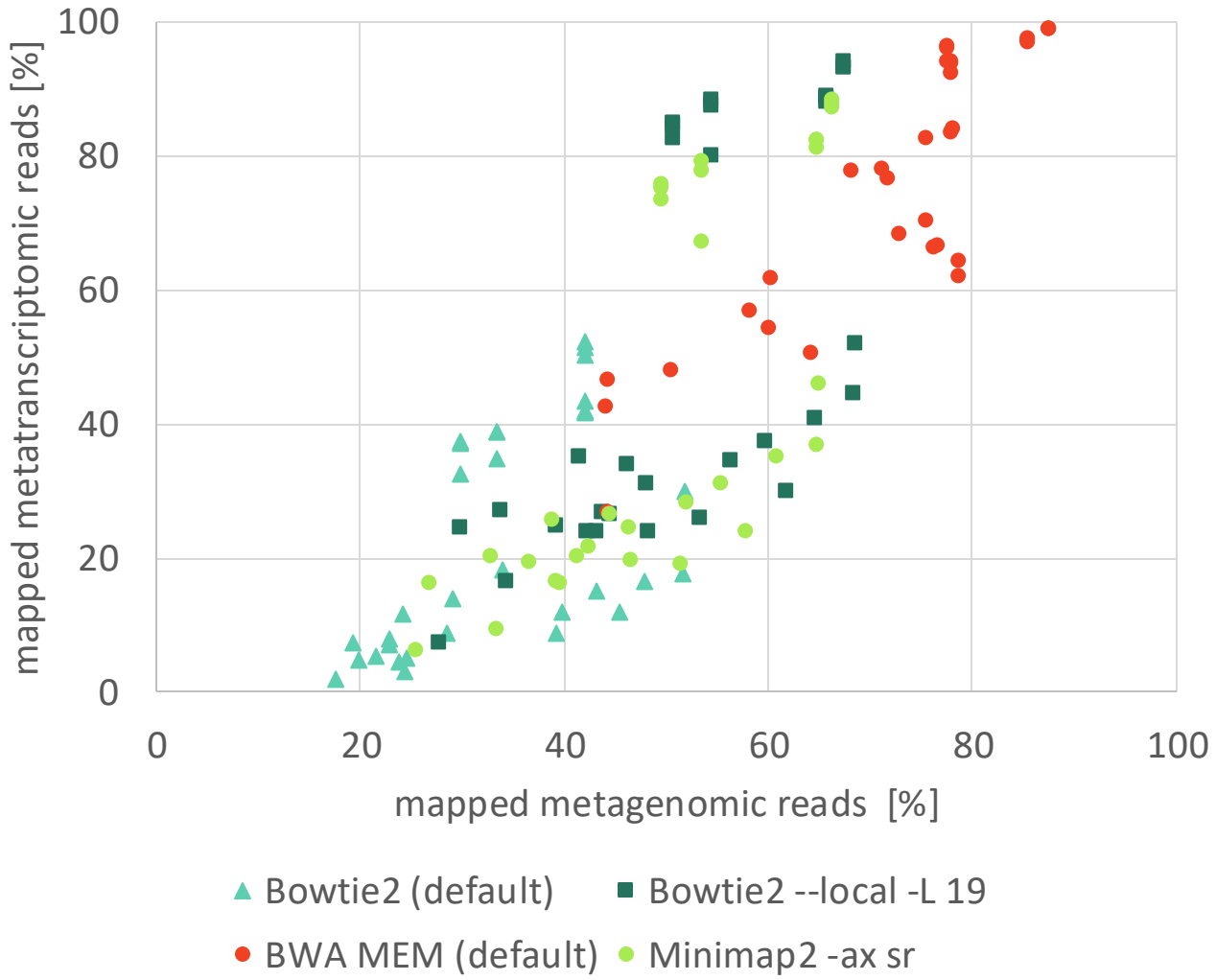
Both reads should be mapped efficiently.

Optimization of Mapping Tools for DNA and RNA



Datasets

- 56 NGS short reads (32 samples)
- Soil microbiome
- Obtained from NCBI SRA

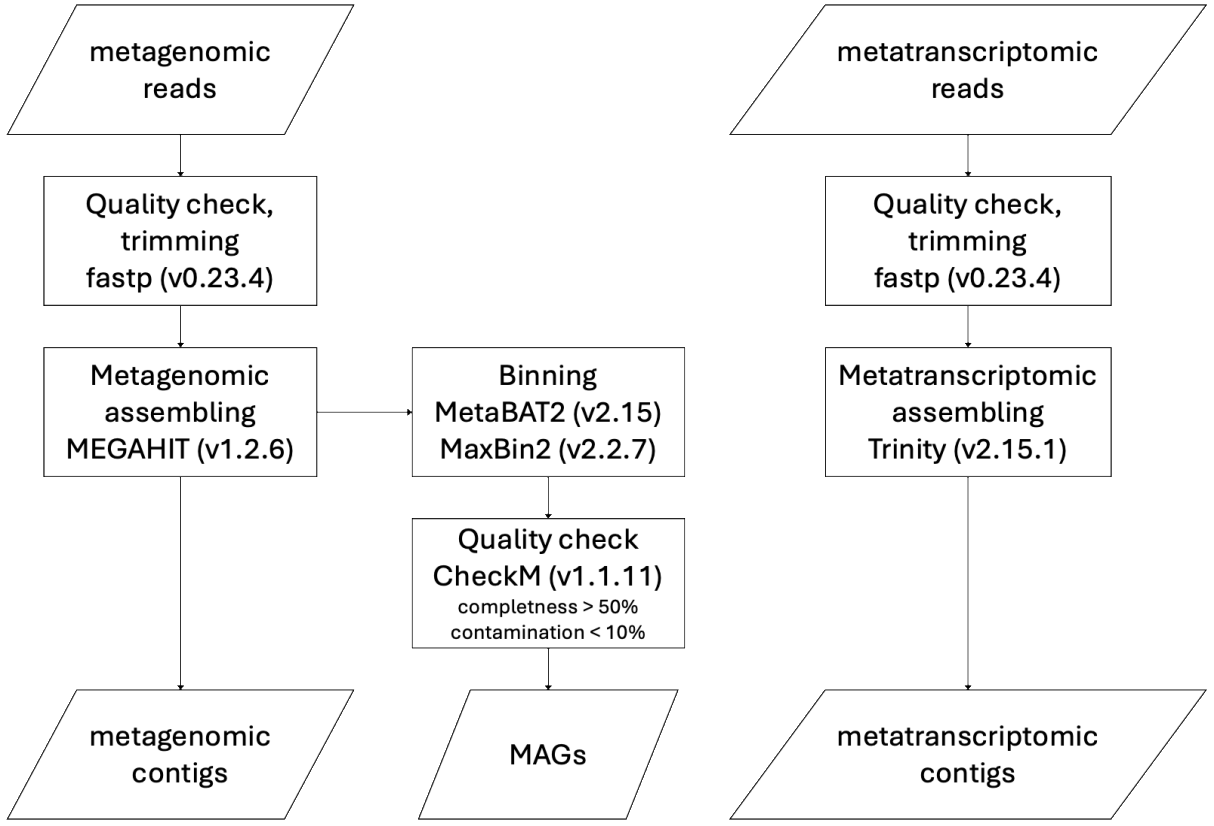


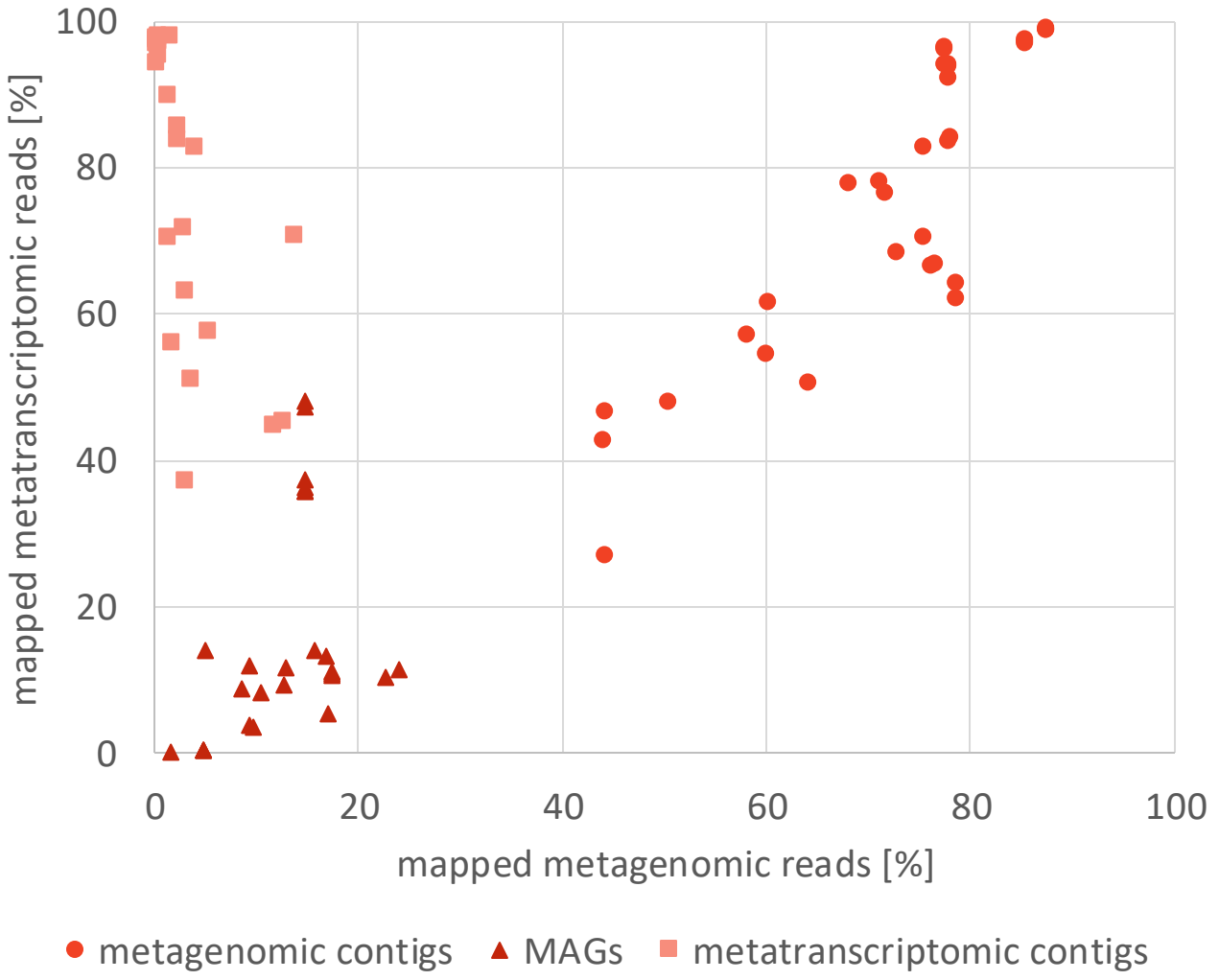
SRA: Sequence Read Archive

BWA MEM is the best choice for metagenomic and metatranscriptomic reads.

Metagenomic Contigs

Although metagenomic contigs are widely used, three types of mapping references are compared.





MAGs: Metagenome-Assembled Genomes

Metagenomic contigs are effective in mapping both reads for the datasets.

Outline

Introduction

Isuue1: Mapping Tool Optimization

Isuue2: Effects of Abundant Genes

Conclusion

Effects of Abundant Gene

After mapping step, mapped read counts can be used for calculation of TPM (transcripts per million) and GPM (genes per million).
Gene expression = TPM / GPM
→ It can be affected by abundant expressed genes, such as ribosomal RNA.

Mapping References	**Mapped counts of metagenomic reads**	**rRNA depletion *in vitro***	**Mapped counts of metatranscriptomic reads**
CDS of metagenomic contig (SRR24888648)	0.16% (155,043/98,847,988)	depletion QIAseq FastSelect 5S/16S/23S Kits	36.0% (23,079,523/64,026,082)
CDS of metagenomic contig (SRR22507541)	0.46% (206,915/44,755,462)	no depletion	95.1% (34,017,387/35,774,766)

Even though rRNA depletion was performed,
rRNA is remained in metatranscriptomic reads.

Effects of Ribosomal RNA Contamination

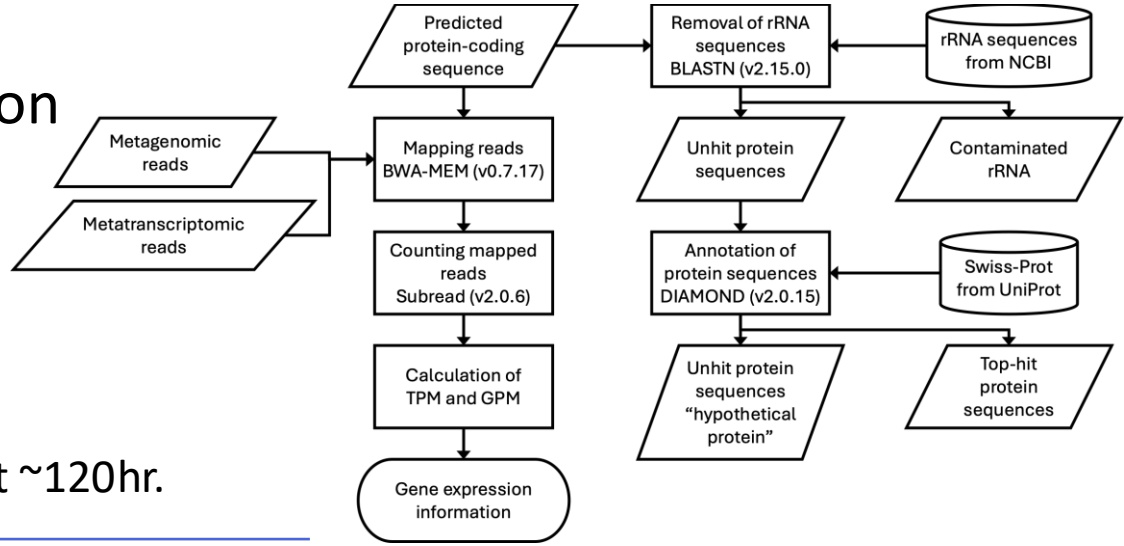
Calculation TPM/GPM for each metagenomic CDS with/without rRNA genes

↓

Sum TPM/GPM values with the same UniProt annotation

↓

Analysis between samples by DESeq2 (v1.46.0)



Sample information https://doi.org/10.1186/s40168-023-01739-z

Soil complex microbiome were anaerobically incubated with rice straw at ~120hr.

Incubation time	Metagenomic reads	Metatranscriptomic reads
14hr	SRR22507544	SRR22506304, SRR22506327, SRR22506328
21hr	SRR22507543	SRR22506324, SRR22506325, SRR22506326
28hr	SRR22507542	SRR22506321, SRR22506322, SRR22506323
35hr	SRR22507541	SRR22506317, SRR22506319, SRR22506320

Differentially Expressed Genes Analysis

Calculation with rRNA

Time shift	Upregulated genes (p<0.05)	Downregulated genes (p<0.05)
14hr→21hr	3.3% 2236/68494	3.1% 2141/68494
14hr→28hr	5.6% 3701/66320	3.6% 2393/66320
14hr→35hr	6.8% 4465/66103	3.7% 2471/66103

Calculation without rRNA

Time shift	Upregulated genes (p<0.05)	Downregulated genes (p<0.05)
14hr→21hr	10.8% 9509/88434	2.0% 1799/88434
14hr→28hr	13.8% 9245/67127	3.6% 3319/67127
14hr→35hr	14.8% 9904/66796	5.2% 3464/66796

rRNA contamination can be supposed to cause inconsistencies.

Outline

Introduction

Isuue1: Mapping Tool Optimization

Isuue2: Effects of Abundant Genes

Conclusion

Conclusion

Results

- BWA-MEM is the best tool for mapping metagenomic and metatranscriptomic reads to predicted protein coding sequences of metagenomic contigs.
- rRNA contamination can be lead inconsistencies for gene expression analysis.
- These results were published. https://doi.org/10.3390/microorganisms13050995

Research Goal

- Maintaining complex microbiome effectively.