CAMI tutorial

assessing metagenomics software with the CAMI benchmarking toolkit

Jie Zhu
zhujie@genomics.cn

2021-03-15 Mon

BGI Research

Outline

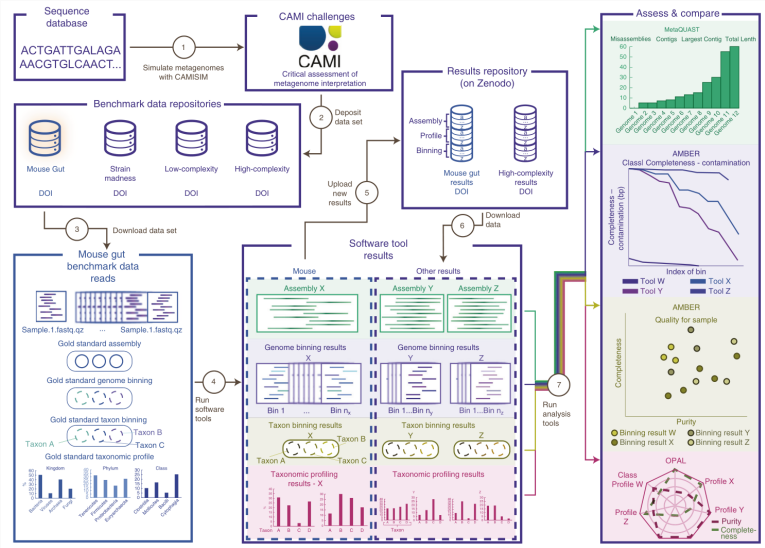
Topic

What is CAMI?

In just over a decade, metagenomics has developed into a powerful and productive method in microbiology and microbial ecology. The ability to retrieve and organize bits and pieces of genomic DNA from any natural context has opened a window into the vast universe of uncultivated microbes. Tremendous progress has been made in computational approaches to interpret this sequence data but none can completely recover the complex information encoded in metagenomes. A number of challenges stand in the way. Simplifying assumptions are needed and lead to strong limitations and

potential inaccuracies in practice. Critically, methodological improvements are difficult to gauge due to the lack of a general standard for comparison. Developers face a substantial burden to individually evaluate existing approaches, which consumes time and computational resources, and may introduce unintended biases. The Critical Assessment of Metagenome Interpretation (CAMI) is a new community-led initiative designed to help tackle these problems by aiming for an independent, comprehensive and bias-free evaluation of methods

Overview


Sequence database
ACTGATTGALAGA
AACGTGLCAACT...
1
Simulate metagenomes with CAMISIM
CAMI challenges
CAMI
Critical assessment of metagenome interpretation
Benchmark data repositories
Mouse Gut
Strain madness
Low-complexity
High-complexity
DOI
DOI
DOI
DOI
2
Deposit data set
Results repository (on Zenodo)
Assembly
Profile
Binning
Mouse gut results
DOI
High-complexity results
DOI
Upload new results
5
3
Download data set
Mouse gut benchmark data reads
Sample.1.fastq.gz
...
Sample.1.fastq.gz
Gold standard assembly
Gold standard genome binning
Gold standard taxon binning
Taxon A
Taxon B
Taxon C
Gold standard taxonomic profile
Kingdom
Phylum
Class
4
Run software tools
Software tool results
Mouse
Other results
Assembly X
Assembly Y
Assembly Z
Genome binning results
X
Y
Z
Bin 1
...
Bin n_X
Bin 1...Bin n_Y
Bin 1...Bin n_Z
Taxon binning results
Taxon A
Taxon B
Taxon C
Taxonomic profiling results - X
Taxonomic profiling results
Taxon
6
Download data
7
Run analysis tools
Assess & compare
MetaQUAST
Misassemblies
Contigs
Largest Contig
Total Length
Genome 1
Genome 2
Genome 3
Genome 4
Genome 5
Genome 6
Genome 7
Genome 8
Genome 9
Genome 10
Genome 11
Genome 12
AMBER
Classl Completeness - contamination
Completeness – contamination (bp)
Index of bin
Tool W
Tool X
Tool Y
Tool Z
AMBER
Qualify for sample
Completeness
Purity
Binning result W
Binning result X
Binning result Y
Binning result Z
OPAL
Class Profile W
Profile X
Profile Y
Profile Z
Purity
Complete-ness

Assembly challenge

What is an assembly method?

An assembly method returns longer nucleotide sequences derived by puzzling together individual sequencing reads. These sequences are assumed to represent contiguous stretches from one genome included in the microbiome sample that was sequenced.

Profiling challenge

What is a profiling method?

A profiling method returns an estimate for the frequencies of different taxa in a sequenced microbial community based on analysis of the sequence sample. The main output is a vector with relative abundances for the different sample taxa. The relative abundances of taxa from the same 'rank' of the taxonomy (e.g. superkingdom, including archaea, bacteria and eukaryotes) cannot sum up to more than 1.

Binning challenge

What is a binning method?

A binning method assigns an identifier to every sequence in a sequence sample, where the total number of identifiers is ideally less than the total number of sequences. Thus the act of binning places the sequences into broader categories. A bin includes all the sequences with the same identifier. If these identifiers identify taxa from a taxonomy, the method is a taxonomic binning method.

Topic

1st CAMI Challenge Dataset 1: CAMI low dieversity

summary

Platform	Hiseq
Number of samples	1
Total Size	15 Gbp
Read length	2x150 bp
Insert size mean	270 bp
Insert size stddev	27 bp

list

```
java -jar ~/.local/bin/camiClient.jar -l \
https://openstack.cebitec.uni-bielefeld.de:8080/swift/v1/CAMI_I_L

CAMI_low_RL_S001__insert_270_GoldStandardAssembly.fasta.gz
RL_S001__insert_270.fq.gz
gold_standard_low_single.fasta.gz
goldstandard_low_1.filtered.profile
goldstandard_low_1.profile
gs_read_mapping.binning.gz
gsa_mapping.binning
novelty_complete.tsv
source_genomes_low.tar.gz
taxonomy.tar.gz
unique_common.tsv
```

1st CAMI Challenge Dataset 2: CAMI medium diversity

summary

Platform	Hiseq
Number of samples	2
Total Size	40 Gbp
Read length	2x150 bp
Insert size mean	270 bp and 5kbp
Insert size stddev	10%

list

```
cami_list https://openstack.cebitec.uni-bielefeld.de:8080/swift/v1/CAMI_I_MEDIUM
```

1st CAMI Challenge Dataset 3: CAMI high diversity

summary

Platform	Hiseq
Number of samples	5
Total Size	75 Gbp
Read length	2x150 bp
Insert size mean	270 bp
Insert size stddev	10%
Diversity	High

list

```
cami_list https://openstack.cebitec.uni-bielefeld.de:8080/swift/v1/CAMI_I_HIGH
```

CAMI1 Toy Test Dataset: Low Complexity

summary

Platform	simulated from public genomes(30)
Number of samples	
Total Size	15 Gbp
Read length	2x100 bp
Insert size mean	180 bp
Insert size stddev	10%
Diversity	Low

list

```
cami_list https://openstack.cebitec.uni-bielefeld.de:8080/swift/v1/CAMI_I_TOY_LOW
```

CAMI1 Toy Test Dataset: Medium Complexity

summary

Platform	simmulater from public genomes(225)
Number of samples	2
Total Size	
Read length	
Insert size mean	180 bp and 5kb
Insert size stddev	
Diversity	Medium

list

```
cami_list https://openstack.cebitec.uni-bielefeld.de:8080/swift/v1/CAMI_I_TOY_MEDIUM
```

CAMI1 Toy Test Dataset: High Complexity

summary

Platform	simulated from public genomes(450)
Number of samples	5
Total Size	75Gbp
Read length	2x100 bp
Insert size mean	180bp
Insert size stddev	18bp
Diversity	High

list

```
cami_list https://openstack.cebitec.uni-bielefeld.de:8080/swift/v1/CAMI_I_TOY_HIGH
```

CAMI2: Rhizosphere challenge

short reads

Platform	simulated Illumina HiSeq metagenome data
Number of samples	21
Total size	105 Gb
Read length	2x150 bp
Insert size mean	270 bp
Insert size s.d.	20 bp

long reads

- Pacific Bioscience

Platform	simulated Pacific Bioscience metagenome data
Number of samples	21
Total size	105 Gb
Average read length	3,000 bp
Read length s.d.	1,000 bp

- Oxford Nanopore

Platform	simulated Oxford Nanopore metagenome data
Number of sampels	21
Total size	105 Gb
Average read length	1,610 bp
Read length s.d.	~3,000 bp

taxonomy database

`https://openstack.cebitec.uni-bielefeld.de:`
`8080/swift/v1/CAMI_DATABASES/taxdump_cami2_toy.tar.gz`

CAMI2: Clinical pathogen detection challenge

CAMI2: Marine Dataset

CAMI2: Strain Madness Dataset

CAMI2: Toy Human Microbiome Project Dataset

Simulated metagenome data from five different body sites of the human host, namely gastrointestinal tract, oral cavity, airways, skin and urogenital tract.

short reads

Platform	simulated Illumina HiSeq metagenome data
Number of samples	49 (12 GI tract, 10 oral cavity, 12 airways, 13 skin, 9 urogenital tract)
Total size	246 Gbp
Read length	2x150 bp
Insert size mean	270 bp
Insert size stddev	25 bp

long reads

Platform	simulated Pacific Biosciences metagenome data
Number of samples	49 (12 GI tract, 10 oral cavity, 12 airways, 13 skin, 9 urogenital tract)
Total size	255 Gbp
Average read length	3500 bp
Read length	1000 bp

list

```
Airways: https://openstack.cebitec.uni-bielefeld.de:8080/cami/v1/CAMI_Airways/
Gastrointestinal Tract: https://openstack.cebitec.uni-bielefeld.de:8080/cami/v1/CAMI_Gastrointestinal_tract/
Oral Cavity: https://openstack.cebitec.uni-bielefeld.de:8080/cami/v1/CAMI_Oral_cavity/
Skin: https://openstack.cebitec.uni-bielefeld.de:8080/cami/v1/CAMI_Skin/
Urogenital Tract: https://openstack.cebitec.uni-bielefeld.de:8080/cami/v1/CAMI_Urogenital_tract/
```

download

```
# java -jar cami4cami.jar -d URL -p PATTERN

# download just fasta files
java -jar cami4cami.jar -d URL -p fq.gz

# download just the gold standard assembly, including the mapping of contigs to reference genomes:
java -jar cami4cami.jar -d URL -p gsa

# download just sample_0 from this data category:
java -jar cami4cami.jar -d https://openstack.cebitec.uni-bielefeld.de:8080/cami/v1/CAMI_Skin -p sample_0
```

explanation

Folder structure:

Sample folders start with the date of creation and end with the sample number: YYYY-MM-$dd_{0000\text{-}0049\ \text{samples}}$. @

In every sample folder there are three subfolders, bam, contigs and reads.

- bam

The bam folder contains the mapping of all the created reads to the input genomes. Inside this folder is a bam file for every genome for which at least one read was produced, which is uniquely indicated by a combination of GTU and a running ID counter for the number of genomes included in that GTU in the sample: GTU_{bam}.

- contigs

The contigs folder contains the gold standard assembly for that particular sample It contains two files, the gold standard in fasta format: $anonymous_{gsa\text{-}fasta.fa.gz}$. And the mapping for each contigs to its genome/taxon id and position in this genome: $gsamapping.tsv.gz$.

- reads

The reads folder contain the created reads for that sample. It contains two files, one with the fq reads themselves, containing both reads for paired end sequencing and with anonymized names: $anonymous_{reads.fq.gz}$. And the second one is a mapping of every single read to the genome it originated from and the original read ID (pre anonymization) $reads_{mapping.tsv.gz}$.

- abundance

every data set contains one abundance file per sample mapping GTUs to genomes: abundance@.tsv

every data set contains the pooled gold standard assembly over all samples in the folder: $anonymous_{pooled\_fasta.fa}$

config file used for creating the data set at hand (can be used as input to CAMISIM for re-creating the data set): config.txt

mapping from the original (BIOM) GTU name to the genome fasta file $genome_{mod.tsv}$.

- genomes

genomes folder containing all the reference genomes used over all samples (using the mapping from $genomes_{mod.tsv}$) genomes This folder contains all the fasta files of the downloaded genomes: $genome_{fasta.fa}$.

Since the contigs are anonymized, a file mapping each contig to its genome/taxon id and position in the respective genome is provided $gsamappingbwa.tsv.gz$.

To each input GTU (from the BIOM file), new tax IDs are mapped. One of the level on which the GTU was mapped in the NCBI and one to the specifically downloaded genome, contains a $novel_{species}$ column in case new genomes are provided, otherwise this column is "unassoc". and can be ignored. metadata.tsv

- hybrid

In the top folder "hybrid" there are assembly and binning gold standards created from both the short and long read data sets. For every sample as well as all samples pooled, the bam files of short and long read simulators (as described for the "bam" subfolder above) are merged and the gold standards calculated the same way as for the individual short or long read samples.

CAMI2: Toy Mouse Gut Dataset

Topic