Diversity Estimation of Metagenomics Samples

Abstract

Author Summary

Introduction

Comparison of metagenomics samples
Coverage estimation of metagenomics reads
Diversity evaluation of metagenomics samples
reference-free, assembly-free annotation-free, binning-free

Results

Comparison of metagenomics samples

Theoretical analysis

Synthetic data

same number of species (100), different composition
same coverage(20X, 1X)
same error rate (no error, illumina error profile)
Coverage matters, as expected
after saturation, it can give correct number. if too low coverage, it is not accurate. But there should be a way to figure out the relationship.
with 1X coverage, 50% of real coverage.
next to do:
1. figure out the relationship between coverage and overlap accuracy
2. synthetic data with real bacterial genomes.
3.

Discussion

Methods

Code and data set availability

synthetic data

We built 4 series of synthetic data sets:
Each series include four sampels with specific composition:
SampleA: 100 species with 80 common to B
SampleB: 100 species with 80 common to A
SampleC: 100 species with 20 common to A/B, and 60 common to D
SampleD: 100 species with 20 common to A/B, and 60 common to D

4 Series with different coverage and different error rate:

1. high coverage(20X) without error
2. low coverage(1X) without error
3. high coverage(20X) with error, illumina error profile
3. low coverage(1X) without error, illumina error profile

Figure Legends

Tables

20X, no error

	sampleA	sampleB	sampleC	sampleD
sampleA		80%	20%	20%
sampleB	80%		20%	20%
sampleC	20%	20%		60%
sampleD	20%	20%	60%	

N% of reads in X are covered in Y

1X, no error

	sampleA	sampleB	sampleC	sampleD
sampleA		44.%	11%	11%
sampleB	44%		11%	11%
sampleC	11%	11%		33%
sampleD	11%	11%	33%	

N% of reads in X are covered in Y

20X, with error

	sampleA	sampleB	sampleC	sampleD
sampleA		74.3%	18..6%	18..6%
sampleB	74.3%		18..6%	18..6%
sampleC	18..6%	18..6%		55.7%
sampleD	18.5%	18.5%	55.8%	

N% of reads in X are covered in Y

1X, with error

	sampleA	sampleB	sampleC	sampleD
sampleA		30.2%	7.5%	7.5%
sampleB	30.2%		7.6%	7.6%
sampleC	7.2%	7.3%		22.7%
sampleD	7.2%	7.3%	22.7%	

N% of reads in X are covered in Y

Figure 1. % of reads in sampleX are covered in sampleY