Diversity Estimation of Metagenomics Samples

Abstract

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

Introduction

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.
1x coverage , 63% of genome is covered
next to do:
1. figure out the relationship between coverage and overlap accuracy
2. synthetic data with real bacterial genomes.

two experiments to do
figure out the ecology of this two levels of diversity difference
3.

Discussion

Methods

Code and data set availability

synthetic data Try 1

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

synthetic data Try 2

We built 3 series of synthetic data sets:
10 species

Sample1A:
species IDs:
1,2,3,4,5,6,7,8,9,10
relative abundance:
20:18:16:4:3:2:2:2:2:2

Sample1B:
species IDs:
1,2,3,14,15,16,17,18,19,20
relative abundance:
20:18:16:4:3:2:2:2:2:2

Sample1C:
species IDs:
21,22,3,4,5,6,7,8,9,10
relative abundance:
2:2:2:2:2:3:4:16:18:20

A and B high overlap on individual level, low overlap on species level
A and C high overlap on species level, low overlap on individual level
B and C low overlap on species level and low overlap on individual level

With error
And without error

1X 10:9:8:2:1.5:1:1:1:1:1

And

6X 60:54:48:12:9:6:6:6:6:6

Relative abundance matters.
We may do diginorm to eliminate the relative abundance firstly then use this method to evaluate richness...

Figure Legends

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

6X, no error

	sample1A	sample1B	sample1C
sample1A		76.2%	94.2%
sample1B	76.2%		2.8%
sample1C	46.2%	22.5%	

N% of reads in X are covered in Y

1X, no error

	sample1A	sample1B	sample1C
sample1A		77.3%	54.3%
sample1B	77.5%		2.9%
sample1C	29.0%	12.3%	

N% of reads in X are covered in Y

6X, with error

	sample1A	sample1B	sample1C
sample1A		71.4%	85.6%
sample1B	71.4%		2.7%
sample1C	42.3%	20.4%	

N% of reads in X are covered in Y

1X, with error

	sample1A	sample1B	sample1C
sample1A		71.3%	38.4%
sample1B	71.3%		2.6%
sample1C	21.8%	8.5%	

N% of reads in X are covered in Y

Figure 2. Test 2 % of reads in sampleX are covered in sampleY

Tables