The command lines for simulation dataset creation

This file contains all the command lines we used to generate the simulation datasets listed in the paper.

Simulation by NeSSM:

1) Three simulated datasets according to Dataset A, B, and C. These three datasets are mapped back to their reference genomes. For every dataset, two types of reads are simulated.
For Dataset A (for a Low complexity metagenome, LC), the command for the first type is:

```
./NeSSM -index index(#1) -list lc.txt(#2) -m 454 -l 100 -r 150000 -o LC-100
```

The parameters for the second type:

```
./NeSSM -index index(#1) -list lc.txt(#2) -m 454 -l 250 -r 60000 -o LC-250
```

The simulations of Dataset B (Medium complexity dataset, MC) and Dataset C (High complexity dataset, HC) are similar to dataset A's. The only difference is the metagenome community composition structure files ("`-list mc.txt`" and "`-list hc.txt`" for dataset B and C respectively).

2) Simulated dataset according to Dataset D to assess the error models and the distribution of quality values in NeSSM.
The config file (self-simulation.config) is estimated from the Dataset D using the script supplied by NeSSM. The community composition structure can be arbitrary. The command is:

```
./NeSSM -index index(#1) -list lc.txt(#2) -m illumina -l 120 -r 2000000 -c self-simulation.config -o name(#3)
```

3) Simulated dataset according to Dataset F in the Illumina platform.
The config file (self-simulation.config), the information of sequencing coverage bias (coverage.txt) and the community composition structure (percentage.txt) are estimated from Dataset F using the scripts supplied by NeSSM. The command is:

```
./NeSSM -list percentage.txt -index index(#1) -m illumina -l 75 -r 2975345 -c self-simulation.config -b coverage.txt -o name(#3)
```

4) Simulate two datasets to assess the speed of NeSSM.
The GPU version and CPU version of NeSSM are used respectively. The command for 454 platform is:

```
./NeSSM -list hc.txt(#2) -index index(#1) -m 454 -l 250 -r 90000000 -o name(#3)
```

The command for Illumina platform is:

```
./NeSSM -list hc.txt(#2) -index index(#1) -m illumina -l 36 -r 90000000 -o name(#3)
```

5) Two simulated datasets according to Dataset E in 454 platform.
The config file (self-simulation.config) and the community composition structure (percentage.txt) are estimated from Dataset E using scripts supplied by NeSSM. The command is:

```
./NeSSM -index index(#1) -list percentage.txt -m 454 -r 475694 -exact 1 -o name(#3) -c self-simulation.config
```

Another dataset is simulated with the community composition structure (morgan.txt) supplied by the Morgan *et al.* The simulation parameters are similar except the "-list morgan.txt".

6) Simulated three different datasets to assess different assemble softwares
Low Complexity:

```
./NeSSM -index index(#1) -list lc.txt(#2) -m illumina -l 36 -r 3000000 -o LC
```

Medium Complexity:

```
./NeSSM -index index(#1) -list mc.txt(#2) -m illumina -l 36 -r 24000000 -o MC
```

High Complexity:

```
./NeSSM -index index(#1) -list hc.txt(#2) -m illumina -l 36 -r 30000000 -o HC
```

#1: the index file is created by users from the reference genome data
#2: the information of lc.txt, mc.txt and hc.txt file is in table S3
#3: the name is defined by user

Simulation by MetaSim:

1) Two simulated datasets to assess the speed of MetaSim.
For 454 platform, the length of reads is 250 bps, the number of reads is 90,000,000, and the community composition structure is hc.txt (in table S3). Other parameters are used by default.
For Illumina platform, the length of reads is 36 bps, the number of reads is 90,000,000, and the community composition structure is hc.txt (in table S3). Other parameters are used by default. The error model is downloaded from http://ab.inf.uni-tuebingen.de/software/metasim/errormodel-80bp.mconf/view.

2) Two simulated datasets according to Dataset E in 454 platform.
The length of reads is 193 bps, the number of reads is 475,694. The community composition structures re-estimated by NeSSM and supplied by Morgan *et al.* are used respectively. Other parameters are used with default values.

3) A simulated dataset according to Dataset F in Illumina platform.
The length of reads is 75 bps and the number of reads is 2,975,345. The community composition structure estimated by NeSSM is used. Other parameters are used with default values.

Attention: In MetaSim, the community composition table is the abundance of species, not the abundance of reads estimated by NeSSM. So the abundance of every species should be converted by dividing its genome length when the estimated result by NeSSM is used in MetaSim. In GemSIM and Grinder, the community composition table is also the abundance of species.

Simulation by GemSIM:

1) Two simulated datasets to assess the speed of GemSIM.
For 454 platform, the length of reads is 250 bps, the number of reads is 10,000, 50,000, 100,000 and 150,000 respectively to assess the time when 90,000,000 reads are simulated, and the community composition structure is hc.txt (in table S3). The parameters are:
./GemReads.py -R directory_of_reference_genomes -a hc.txt -n reads_number -m models/r454ti_s.gzip -q 33 -o GemSIM
For Illumina platform, the length of reads is 36 bps, the number of reads is 10,000, 50,000, 100,000 and 150,000 respectively to assess the time when 90,000,000 reads are simulated, and the community composition structure is hc.txt (in table S3). The parameters are:
./GemReads.py -R directory_of_reference_genomes -a hc.txt -n reads_number -m models/ ill100v4_s.gzip -q 64 -o output_filename

2) Two simulated datasets according to Dataset E in 454 platform.
The number of reads is 475,694. The community composition structures (percentage.txt) re-estimated by NeSSM. The error mode (model.gzip) is estimated form Dataset E by GemSIM. The parameters used in simulation are:
./GemReads.py -R directory_of_reference_genomes -a percentage.txt -n 475,694 -m model.gzip -ld -q 33 -o GemSIM-DatasetE-self
Another data is simulated similarly except the community composition structure table supplied by Morgan *et al*.

3) A simulated dataset according to Dataset F in Illumina platform.
The length of reads is 75 bps and the number of reads is 2,975,345. The community composition structure (percentage.txt) estimated by NeSSM is used. The error mode (model.gzip) is estimated form Dataset E by GemSIM. The parameters used in simulation are:
./GemReads.py -R directory_of_reference_genomes -a percentage.txt -n 2,975,345 -m model.gzip -ld -q 64 -o GemSIM-DatasetF

Simulation by Grinder:

1) Two simulated datasets to assess the speed of Grinder.
For 454 platform, the length of reads is 250 bps, the number of reads is 1,000, 5,000, 10,000 and 15,000 respectively to assess the time when 90,000,000 reads are simulated, and the community composition structure is hc.txt (in table S3). The parameters are:

./grinder -rf reference_genomes -af hc.txt -tr reads_number -rd 250 -md linear 1 2 –fq 1 –ql 30 10 –hd Balzer –bn Grinder

For Illumina platform, the length of reads is 36 bps, the number of reads is 1,000, 5,000, 10,000 and 15,000 respectively to assess the time when 90,000,000 reads are simulated, and the community composition structure is hc.txt (in table S3). The parameters are:

./grinder -rf reference_genomes -af hc.txt -tr reads_number -rd 75 -md poly4 3e-3 3.3e-8 –fq 1 –ql 30 10 –bn Grinder

2) Two simulated datasets according to Dataset E in 454 platform.
The number of reads is 475,694. The community composition structures (percentage.txt) re-estimated by NeSSM. The parameters used in simulation are:

./grinder -rf reference_genomes -af percentage.txt -tr 475,694 -mr 17 83 -md linear 1 2 -rd 193 normal 60 –fq 1 –ql 30 10 –hd Balzer –bn Grinder-DatasetE-self

Another data is simulated similarly except the community composition structure table supplied by Morgan *et al*.

3) A simulated dataset according to Dataset F in Illumina platform.
The length of reads is 75 bps and the number of reads is 2,975,345. The community composition structure (percentage.txt) estimated by NeSSM is used. The parameters used in simulation are:

./grinder -rf reference_genomes -af percentage.txt -tr 2,975,345 -rd 75 -md poly4 3e-3 3.3e-8 -mr 95 5 –fq 1 –ql 30 10 –bn Grinder

Simulation by pIRS:

The system of pIRS can simulate data with sequencing coverage bias. The pIRS is used to compare with NeSSM. Because pIRS can only simulate single genome, so only the Acinetobacter baumannii (ATCC 17978) is simulated. The error model (model.dat) is estimated from Dataset F using pIRS and 606,771 reads are simulated. The parameters are:

./pirs simulate -i reference_genome -d model.dat -l 75 -o pIRS

In order to generate model.dat file by pIRS, the “-w” is set “100” when run the script of gc_coverage_bias.