Experiment receipts

This is an introduction for reproducing the assembly results mentioned in the manuscript of SOAPdenovo2.

1. YH genome assemblies

1) Decompress the file YH_pipeline.tgz: tar xvfz YH_pipeline.tgz.
2) There are four command shells, representing four assemblies:

- run_1.0.sh: Use SOAPdenovo1 pipeline to assemble the YH genome.
- run_2.0.sh: Use SOAPdenovo2standard pipeline to assemble the YH genome, you can enable multi-*k*-mer mode in this pipeline.
- run_sparse.sh: Enable SOAPdenovo2 sparse DBG mode to assemble YH genome.
- run_sparse_muti_test.sh: Enable both the SOAPdenovo2 sparse DBG mode and multi-*k*-mer mode to assemble the YH genome.

3) For each assembly, there are 9 steps including data download, read filter, error correction, assembly and gap closure. So you need to check whether each step has finished correctly.

4) To get the final assembly results, if you only proceeded to the scaffolding ('scaff') step in SOAPdenovo, you can find the genome sequence with a file name: "assembly/*.scafSeq". If you have finished all the steps including GapCloser, you can find the final genome sequences file named as "gapcloser/*.scafSeq.GC".

2. To assembly GAGE data:

There are three datasets of GAGE assembled by SOAPdenovo, so we have three pipelines that should be decompressed respectively:

- Bombus_impatiens_pipeline.tgz
- Rhodobacter_sphaeroides_pipeline.tgz
- Staphylococcus_aureus_pipeline.tgz

Before running the assembly, you need to download the reads and put them into the correct locations:
1) Download the three compressed datasets from http://gage.cbcb.umd.edu/data/.
2) Decompress the datasets and put the reads into the corresponding Data folder.

Each genome has two command shells:

- SOAPdenovo1_pipeline.sh: Use SOAPdenovo1 pipeline to assemble the data in working directory.

- SOAPdenovo2_pipeline.sh: Use SOAPdenovo2 pipeline to assemble the data in working directory.

The final assembly sequence file is named as "*.scafSeq.GC.fa";

3. To assembly Assemblathon1 data:

Decompress the file: Assemblathon1_pipeline.tgz.

The Assemblathon1 assembly including two datasets:

- Bacteria (contamination) sequence database: you can download it from ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.fna.tar.gz
- Assembly data: the pipeline will download the data automatically, if you have these data already, please read the README carefully in the Assemblathon1_pipeline.

After all the datasets are ready, please run the command shell:
run.sh: Use SOAPdenovo1 and SOAPdenovo2 pipeline in serial to assemble the Assemblathon1 reads.

The final assembly sequence file is named as "consensus/*.scafSeq.GC.filter";

For more details, please refer to the README in each pipeline.