1. **Source Code**

The source code of SparseAssembler, DBG2OLC and Sparc can be found here:

<https://github.com/yechengxi/SparseAssembler>

<https://github.com/yechengxi/DBG2OLC>

<https://github.com/yechengxi/Sparc>

To compile, download the code into separate folders and use:

g++ -O3 -o SparseAssebmler \*.cpp

g++ -O3 -o DBG2OLC \*.cpp

g++ -O3 -o Sparc \*.cpp

1. **Datasets used in the paper**

Illumina datasets used in the paper:

|  |  |
| --- | --- |
| Datasets | Illumina Data Source |
| *S. cer* w303 | 50x Miseq, <http://schatzlab.cshl.edu/data/ectools/> |
| *A. thaliana* ler-0 | 50x Miseq, <http://schatzlab.cshl.edu/data/ectools/> |
| *H. sapiens* | 80x Hiseq, Accession No.: SRR1283824 |
| *E.coli* K12 | 50x Miseq, Accession No.: SRR826442, SRR826444, SRR826446, SRR826450 |

PacBio/Nanopore datasets used in the paper:

|  |  |
| --- | --- |
| Datasets | PacBio Data Source |
| *S. cer* w303 | <http://schatzlab.cshl.edu/data/ectools/> |
| *A. thaliana* ler-0 | <http://schatzlab.cshl.edu/data/ectools/> |
| *H. sapiens* | <http://datasets.pacb.com/2014/Human54x/fast.html> |
| *E. coli* K12 | <http://gigadb.org/dataset/100102> |

Reference genomes used in the paper:

|  |  |
| --- | --- |
| Datasets | Reference Data Source |
| *S. cer* w303 | <http://www.cbcb.umd.edu/software/PBcR/mhap/asm/yeast.quiver.all.fasta> Accession No.: GCA\_000292815.1 |
| *A. thaliana* ler-0 | <http://www.cbcb.umd.edu/software/PBcR/mhap/asm/athal.quiver.all.fasta> |
| *H. sapiens* | Accession No.: GCA\_000772585.3 |
| *E. coli* K12 | Accession No.: NC\_000913 |

1. **Exemplary Assembly Commands**

**Commands for hybrid assembly:**

***Step0.*** [Optional] Preparations:

We have provided code to help you select a subset of the reads:

<https://github.com/yechengxi/AssemblyUtility>

After compilation, you can use the following command to select a subset of reads from fasta/fastq files. Note that longest 0 is used here, if you set it to 1 it will select the longest reads.

*./SelectLongestReads sum 600000000 longest 0 o Illumina\_50x.fastq f Illumina\_500bp\_2x300\_R1.fastq*

./SelectLongestReads sum 260000000 longest 0 o Pacbio\_20x.fasta f Pacbio.fasta

And you can use the following command to evaluate an assembly.

*./AssemblyStatistics contigs YourAssembly.fasta*

The program will generate two txt files containing essential statistics about your assembly.

***Step1****.* Use an accurate DBG-assembler to construct short but accurate contigs. Please make sure they are the **raw** DBG contigs **without** using repeat resolving techniques such as gap closing or scaffolding. Otherwise you may have poor final results due to the errors introduced by the heuristics used in short read assembly pipelines.

SparseAssembler command format:

*./SparseAssembler GS GENOME\_SIZE NodeCovTh FALSE\_KMER\_THRESHOLD EdgeCovTh FALSE\_EDGE\_THRESHOLD k KMER\_SIZE g SKIP\_SIZE f YOUR\_FASTA\_OR\_FASTQ\_FILE1 f YOUR\_FASTA\_OR\_FASTQ\_FILE2 f YOUR\_FASTA\_OR\_FASTQ\_FILE3\_ETC*

A complete example on the *S.cer* w303 dataset:

Download the Illumina reads from

<ftp://qb.cshl.edu/schatz/ectools/w303/Illumina_500bp_2x300_R1.fastq.gz>

Normally with ~50x coverage, NodeCovTh 1 EdgeCovTh 0 can produce nice results.

*./SparseAssembler LD 0 k 51 g 15 NodeCovTh 1 EdgeCovTh 0 GS 12000000 f ../Illumina\_data/Illumina\_50x.fastq*

In this test run, the N50 is 29 kbp. As we have selected the beginning part of the sequencing file, which can be of lower quality, the next step may help to improve the assembly quality.

[Miscellaneous]

For other more complex genomes or a different coverage, the first run may not generate the best result. The previous computations can be loaded and two parameters can be fine-tuned to construct a cleaner de Bruijn/ k-mer graph:

*./SparseAssembler* ***LD 1 NodeCovTh 2 EdgeCovTh 1*** *k 51 g 15 GS 12000000 f ../Illumina\_data/Illumina\_50x.fastq*

The N50 is improved to 32kbp in my run.

The output Contigs.txt will be used by DBG2OLC.

*Step2.* Overlap and layout. Feed DBG2OLC with the contig file in fasta format from the previous step (Contigs.txt in this example).

Download the PacBio reads from:

<ftp://qb.cshl.edu/schatz/ectools/w303/Pacbio.fasta.gz>

The basic command format of DBG2OLC is:

*./DBG2OLC k KmerSize AdaptiveTh THRESH\_VALUE1 KmerCovTh THRESH\_VALUE2 MinOverlap THRESH\_VALUE3 Contigs NGS\_CONTIG\_FILE f LONG\_READS.FASTA RemoveChimera 1*

In the following example, the first 20x PacBio reads are extracted from the abovementioned file and we can assemble with:

*./DBG2OLC k 17 AdaptiveTh 0.0001 KmerCovTh 2 MinOverlap 20 RemoveChimera 1 Contigs Contigs.txt f ../Pacbio\_data/Pacbio \_20x.fasta*

In our test run, the N50 is 583kbp.

There are three major parameters that affect the assembly quality:

M = matched k-mers between a contig and a long read.

**AdaptiveTh**: adaptive *k*-mer matching threshold. If M < AdaptiveTh\* Contig\_Length, this contig cannot be used as an anchor to the long read.

**KmerCovTh**: fixed k-mer matching threshold. If M < KmerCovTh, this contig cannot be used as an anchor to the long read.

**MinOverlap**: minimum overlap score between a pair of long reads.

For each pair of long reads, an overlap score is calculated by aligning the compressed reads and score with the matching *k*-mers.

[Miscellaneous]

At this point, the parameters may be fine-tuned to get better performance. As with SparseAssembler, LD 1 can be used to load the compressed reads/anchored reads.

Suggested tuning range is provided here:

For 10x/20x PacBio data: KmerCovTh 2-5, MinOverlap 10-30, AdaptiveTh 0.001~0.01.

For 50x-100x PacBio data: KmerCovTh 2-10, MinOverlap 50-150, AdaptiveTh 0.01-0.02.

Some other less flexible or less important parameters:

k: *k*-mer size, 17 works well.

Contigs: the fasta contigs file from existing assembly.

MinLen: minimum read length.

RemoveChimera: remove chimeric reads in the dataset, suggest 1 if you have >10x coverage.

For high coverage data (100x), there are two other parameters:

ChimeraTh: default: 1, set to 2 if coverage is ~100x.

ContigTh: default: 1, set to 2 if coverage is ~100x.

These two are used in multiple alignment to remove problematic reads and false contig anchors. When we have high coverage, some more stringent conditions shall be applied as with the suggested parameters.

*Step 3.* Call consensus. Install blasr and the consensus module (sparc/pbdagcon). Make sure they are in your path variable.

The input files for consensus are:

(1) backbone\_raw.fasta by DBG2OLC

(2) DBG2OLC\_Consensus\_info.txt by DBG2OLC

(3) DBG contigs (in fasta format)

(4) PacBio reads (in fasta format)

You can check the N50 of (1) to see if you are satisfied, otherwise keep tuning and don’t proceed.

# this is to concatenate the contigs and the raw reads for consensus

*cat Contigs.txt pb\_reads.fasta > ctg\_pb.fasta*

# we need to open a lot of files to distribute the above file into lots of smaller files

*ulimit -n unlimited*

#run the consensus scripts

*sh ./split\_and\_run\_sparc.sh backbone\_raw.fasta DBG2OLC\_Consensus\_info.txt ctg\_reads.fasta ./consensus\_dir 2 >cns\_log.txt*

Commands used to assemble other genomes:

The *A. thaliana* Ler-0 dataset:

20x PacBio reads:

*./DBG2OLC KmerCovTh 2 AdaptiveTh 0.005 MinOverlap 20 RemoveChimera 1 Contigs Contigs.txt k 17 f ../PacBio/20x.fasta*

40x PacBio reads:

*./DBG2OLC KmerCovTh 2 AdaptiveTh 0.01 MinOverlap 20 RemoveChimera 1 Contigs Contigs.txt k 17 f ../PacBio/40x.fasta*

The *H. sapiens* dataset:

Longest 30x PacBio reads:

*./DBG2OLC k 17 KmerCovTh 2 MinOverlap 20 AdaptiveTh 0.01 RemoveChimera 1 Contigs Contigs.txt f 30x.fasta >DBG2OLC\_LOG.txt*

**Commands for purely Illumina reads:**

The program command is slightly different.

Example command:

./DBG2OLC LD 0 MinOverlap 70 PathCovTh 3 Contigs Contigs.txt k 31 KmerCovTh 0 f ReadsFile1.fa f ReadsFile2.fq f MoreFiles.xxx

There are four critical parameters:

k: k-mer length (max size: 31).

**KmerCovTh**: # k-mer matches for a contig to be regarded as a genuine anchor, suggest 0-1.

**MinOverlap**: # ‘consistent’ k-mers between each pair of reads to be considered to overlap.

**PathCovTh**: the minimum occurrence for a compressed read for a compressed read to be used, suggest 1-3.

Assembly is reported as DBG2OLC\_Consensus.fasta.

The command we used for E. coli Illumina Miseq dataset:

./DBG2OLC k 31 PathCovTh 2 MinLen 50 MinOverlap 31 Contigs Contigs.txt KmerCovTh 0 f reads.fasta