Manual for http://denovoassembler.sf.net/
Sbastien Boisvert

```
Software version:

Racine du dpt : svn+ssh://globe/home/boiseb01/SVN-ASSEMBLER-CODE
UUID du dpt : 58b31dbf-21b7-474f-bf1d-173c48c58bce
r290 | boiseb01 | 2009-03-15 20:09:01 -0400 (dim, 15 mar 2009) | 1 line
```

Contents

1 Installation

The dependencies are a POSIX (GNU/Linux) system, ruby, g++, make, coreutils, etc.

To install the software, uncompress the archive. Then, type ./configure;make in the directory. After that, add the src directory to your path. You need to install the ruby interpreter.

2 Overview

The assembler is implemented in various independant modules. To use the assembler, just use the wrapper DNA.rb, which will do almost everything for you.

3 Input files

Input files are sequences. They can be formatted in at least 3 formats: fasta, fastq, and sff. To use paired-end with sff files, you must extract relevant information using 'dna_ConvertSffToFasta'.

4 Command line

```
[boiseb01@ls30 unamed-assembly]$ DNA.rb
Welcome to DNA, the de novo assembler
DNA.rb is an assembler that calls the c++ modules
usage:
DNA.rb [-wordSize 21] [-minimumCoverage 2] -directory output <sequence files (fasta or sff or fastq)
Note: if your sff files contain  read pairs,  you must utilize dna_ConvertSffToFasta to extract pair
DNA.rb will detect fastq or 454 paired information, provided files are named the same except a '_1.*
```

The '-wordSize' argument refers to the length of sequences in vertices. The '-minimumCoverage' is the most important argument. It is a threshold, the mers in reads need at least a coverage equal to it to be considered.

5 Paired-end reads

To detect paired information, the assembler assumes that paired reads, for a particular library, are stored in 2 files with the same name, except that left reads are in '*_1.*' and that right reads are in '*_2.'. Furthermore, the default insert size for 454 is 2500 whereas it is 200 for Illumina/Solexa. Please note that the distance for each paired-end library, is the distance between the beginning of the left read to the beginning of the right read. So, if you have an Illumina paired-end dataset with fragment size of 200 and read length of 36, the distance is 200-36=164. When you run DNA.rb, the 'dna_BuildGraph' module will build the graph. While this is being done, you should edit the 'PairedReads.txt' file in your assembly directory to see if you need to change the default values.

Example for SRA001125:

```
2
./SRA001125/sdata/SRR001665_1.fastq ./SRA001125/sdata/SRR001665_2.fastq 164
./SRA001125/sdata/SRR001666_1.fastq ./SRA001125/sdata/SRR001666_2.fastq 164
```

6 Output files

The directory (via '-directory'), you will find these files. The name of the files is self-explanatory.

```
START
Parameters.txt
InputFiles.txt
PairedReads.txt
dna_GetPairedInformation.START
dna_GetPairedInformation.rb.log
dna_BuildGraph.START
dna_BuildGraph.log
CoverageDistribution.txt
Edges.txt
graph                            <---------- the whole graph, with some annotations
dna_ExtractContigs.START
dna_ExtractContigs.log
dna_JoinContigs.log
dna_KeepLargeContigs.log
dna_MergeContigs.log
contigs-amos.afg
contigs-coverage.txt
contigs.fasta
contigs-repeats.txt
2LargeContigs.fasta
3MergedContigs.fasta
4JoinedContigs.fasta     <------ final assembly
END
```

7 Data analysis

The assembler offers many ways to analyse your sequence data.

7.1 Coverage distribution

The coverage distribution must be inspected in order to set the '-minimumCoverage' parameter correctly. Usually, '-minimumCoverage 2' do the job for 454. For Illumina/Solexa, you can either set '-minimumCoverage auto' or inspect the curve in 'CoverageDistribution.txt' and select what you see fit.

Example for SRA000156:

```
1 5782104
2 595905
3 168034
4 66344
5 42680
6 36858
7 43654
8 64228
9 96290
10 136554
11 195462
12 254512
13 316102
14 376884
15 433406
16 489936
17 527132
18 561421
19 564022
20 553822
21 534380
22 500523
23 470108
24 421381
25 377744
26 339936
27 297552
```

```
28 252017
29 216590
30 181573
31 153574
32 127675
33 104926
34 84307
35 67006
36 55424
37 45892
38 38520
39 31254
40 25453
41 19438
...
```

7.2 Topology distribution

You will find this distribution in the file 'dna_ExtractContigs.log'. A majority of vertices have exactly one parent and one child. These are usually easy to assemble.

Example for SRA000156:

```
0 parents, 1 children: 29732 vertices
0 parents, 2 children: 11 vertices
0 parents, 3 children: 1 vertices
0 parents, 4 children: 2 vertices
1 parents, 0 children: 29732 vertices
1 parents, 1 children: 9824894 vertices          <----------- the majority of vertices are '1-1'
1 parents, 2 children: 46438 vertices
1 parents, 3 children: 228 vertices
1 parents, 4 children: 39 vertices
2 parents, 0 children: 11 vertices
2 parents, 1 children: 46438 vertices
2 parents, 2 children: 970 vertices
2 parents, 3 children: 64 vertices
2 parents, 4 children: 18 vertices
3 parents, 0 children: 1 vertices
3 parents, 1 children: 228 vertices
3 parents, 2 children: 64 vertices
3 parents, 3 children: 18 vertices
3 parents, 4 children: 11 vertices
4 parents, 0 children: 2 vertices
4 parents, 1 children: 39 vertices
4 parents, 2 children: 18 vertices
4 parents, 3 children: 11 vertices
```

7.3 Repeats

The 'contigs-repeats.txt' file contains repeat information from the assembly. A vertex is considered as a repeat if it has "too many annotations".

7.4 Coverage

The 'contigs-coverage.txt' file contains contig coverage information.

7.5 AMOS messages

The http://amos.sf.net/ project is a set of tools to handle assemblies. The AMOS message format is the best specification for formatting assemblies. The hawkeye utility allows the display of AMOS banks.

8 Technologies

8.1 Roche/454

The assembler supports paired-end reads, and any version of the instrument. The assembler natively support the sff format, but it these files contain paired-end information, you will need to use the 'dna_ConvertSffToFasta' module before running the assembler.

8.2 Illumina/Solexa

The assembler supports paired-end reads, and any length of Illumina reads. Note that you will usually need to select the minimum coverage by plotting the 'CoverageDistribution' file.

Type 'R –vanilla', then

```
r=read.table('CoverageDistribution.txt')
plot(r[[1]],log(r[[2]]),xlab='Coverage',ylab='Count',main='Coverage distribution',xlim=c(1,400))
```

8.3 ABI SOLiD

The software was not tested with this technology, but it should *work*.

8.4 Hybrid assemblies

Because Illumina and 454 don't have the same sequencing errors and bias, it is obvious that an hybrid assembler would be very good. Consequently, our assembler permits that.

Example:

```
DNA.rb -directory HybridAssembly -minimumCoverage 10 454_1.fasta 454_2.fasta 454_shotgun.fasta \
 Illumina_fragments.fastq Illumina_1.fastq Illumina_2.fastq
```

9 Copying

The assembler is licensed under the terms of the general public license, version 3 or later.

10 Implementation

All the core modules are in C++. The mers are represented with 'uint64_t' on 64 bits. They are collected along with their count from reads. Don't bother trying to use the assembler on 32-bits architectures, it probably won't work. Also, it is currently only tested on 'x86_64' architecture (http://amd.com/, http://intel.com/). All assembly operations are performed in 'uint64_t'-space, using mostly bit shifts.

The DNA.rb wrapper eases assembly processes although you might want to fully understand how the whole thing works at some point.