Introduction to de Novo Assembly
de novo Assembly

Definition

de novo is a Latin expression meaning "from the beginning," "afresh," "anew," "beginning again". So, in our application, it is the process of building a genome from scratch, or, without a reference genome to guide us. In terms of complexity and time requirements, de-novo assemblies are orders of magnitude slower and more memory intensive than mapping, or “reference” assemblies. It is similar to putting together a very large jigsaw puzzle with the pieces flipped face down and no picture on the box to guide you!
de novo Assembly

Why do de novo?

- Golden age of genomics as far as cost of sequencing and tools to assemble! Therefore, still not a lot of finished genomes out there to use as references or maps for guided assembly.
- Golden age of genomics as far as cost of sequencing and tools to assemble! Therefore, the "finished" genomes out there available may well end up being rubbish as, with every new breakthrough, we find substantial improvements that can be made to existing models.
de novo Assembly

Project Overview

- Acquire and isolate DNA from specimen and prep for sequencing.
- Decide on amount and type of sequencing required to build genome.
- Send samples to sequencer center.............W-A-I-T !
- Begin to look at storage and compute needs.....W-O-R-R-Y !
- Switch to selection of software stack for processing....C-O-N-F-U-S-I-O-N!
- Publication deadline approaching.........P-A-N-I-C !
de novo Assembly
The 10+ Commandments of Assembly

1. Do I really need to assemble?
2. Good data is more important than choice of assembler.
3. Have a specific goal.
4. An assembly is a hypothesis to be tested.
5. Assembly programs are not haplotype aware.
6. More data may help.
7. If you haven’t found contamination in your data you haven’t looked hard enough.
8. A different assembler may help.
9. Make sure the assembly agrees with the reads that were used to put it together.
10. N50 is not a measure of quality.
11. But we don’t have a measure of quality.
13. Trust contigs more than scaffolds more than gap filling.
14. The answer to your question may not be in your data.
15. A bad assembly that completes, is better than a good assembly that doesn’t.
de novo Assembly

Sequencing Technologies & Platforms

- Illumina MiSeq
- Read Length: Up to 300bp
- # of Reads: Up to 25 million/flowcell
- Throughput per run: Up to 15GB
- Run Time: ~65 Hours
- ~$100,000
de novo Assembly

Sequencing Technologies & Platforms

- Illumina HiSeq
- Read Length: Up to 100bp
- # of Reads: Up to 1.5 billion/flowcell
- Throughput per run: Up to 300GB
- Run Time: ~11 Days
- ~$750,000
de novo Assembly
Sequencing Technologies & Platforms

- Pacific Biosciences (PacBio) RS II
- Read Length: Average of 8.5kbp
- # of Reads: 50,000 / SMRT cell
- Throughput per run: ~375GB
- Run Time: 180 minutes
- ~$700,000
de novo Assembly

Sequencing Technologies & Platforms

- Oxford Nanopore MinION
- Read Length: Average of 5.4kbp
- # of Reads:
- Throughput per run:
- Run Time:
- ~$1000*
de novo Assembly

Sequencing Technologies & Platforms
de novo Assembly

Types of sequencer output

- Single Read
- Pair Ended
- Mate Pair
- Long Reads
de novo Assembly

Types of sequencer data

Figure 4. Paired-End Sequencing and Alignment

Paired-End Reads

Alignment to the Reference Sequence

Read 1

Read 2

Reference

Repeats

Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.
de novo Assembly

Types of sequencer data

Figure 5. De Novo Assembly with Mate Pairs

Using a combination of short and long insert sizes with paired-end sequencing results in maximal coverage of the genome for de novo assembly. Because larger inserts can pair reads across greater distances, they provide a better ability to read through highly repetitive sequences and regions where large structural rearrangements have occurred. Shorter inserts sequenced at higher depths can fill in gaps missed by larger inserts sequenced at lower depths. Thus a diverse library of short and long inserts results in better de novo assembly, leading to fewer gaps, larger contigs, and greater accuracy of the final consensus sequence.
de novo Assembly

Computer Hardware (Option #1)

- Apple Macintosh Workstation
- OSX is Unix based (app compatibility)
- Intel multi-core CPU
- 32GB Memory
- 2TB Disk Storage
- ~$3500.00
de novo Assembly

Computer Hardware (Option #1)

- Dell “Fat Node“
- Linux O.S.
- 4x Intel multi-core CPU's
- 768GB Memory
- Remote Disk Storage
- ~$27,000.00
de novo Assembly

Computer Hardware (Option #1)

- SGI UV2000
- Linux O.S. with NUMA Extensions
- Up to 256 Intel multi-core CPU's (2048 cores)
- 64TB Memory
- Remote Disk Storage
- ~$10M
de novo Assembly

So...let's get to work!!!

(make sure you are connected to wireless on 2G or 5G, not guest!)
de novo Assembly

Our data set for today is...

Saccharomyces cerevisiae
“yeast“
Simplest eukaryotic genome (1.2 × 10^7 base pairs of DNA)
6,275 genes, compactly organized on 16 chromosomes.
Only about 5,800 of these genes are believed to be functional.
Estimated ~31% of yeast genes have homologs in the human genome
de novo Assembly

Organization of Data

Do the following:

ssh <username>@razor.uark.edu
Password: **********

cd /scratch/<username>

cp -r /storage/jpummi/Workshop .

cd Workshop

ls
BBMap.pbs data FastQC.pbs Quast.pbs SOAP.config SOAP.pbs SPades.pbs Velvet.pbs
de novo Assembly

Organization of Data

cd data

ls -lh

```
total 50M
-rw-r--r-- 1 drhoads drhoads 13M Apr  7 13:41 GCA_000773925.1_ASM77392v1_genomic.fna
-rw-r--r-- 1 drhoads drhoads 19M Apr  7 13:41 PE-350.1.fastq
-rw-r--r-- 1 drhoads drhoads 19M Apr  7 13:41 PE-350.2.fastq
```
# de novo Assembly

## Organization of Data

(Example)

<table>
<thead>
<tr>
<th>Current Filename</th>
<th>File Size</th>
<th>Quality Encoding</th>
<th>Avg Read Length</th>
<th>Insert Size</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>L001_R1_MiSeq.fastq</td>
<td>16GB</td>
<td>Illumina 1.9</td>
<td>300bp</td>
<td>479</td>
<td>Pair Ended</td>
</tr>
<tr>
<td>L001_R2_MiSeq.fastq</td>
<td>16GB</td>
<td>Illumina 1.9</td>
<td>300bp</td>
<td>479</td>
<td>Pair Ended</td>
</tr>
<tr>
<td>L003_R1_MP.fastq</td>
<td>21GB</td>
<td>Illumina 1.9</td>
<td>101bp</td>
<td>6000</td>
<td>Mate Pairs</td>
</tr>
<tr>
<td>L003_R2_MP.fastq</td>
<td>21GB</td>
<td>Illumina 1.9</td>
<td>101bp</td>
<td>6000</td>
<td>Mate Pairs</td>
</tr>
<tr>
<td>s1_R1_PE.fastq</td>
<td>27GB</td>
<td>Illumina 1.5</td>
<td>101bp</td>
<td>150</td>
<td>Pair Ended</td>
</tr>
<tr>
<td>s1_R2_PE.fastq</td>
<td>27GB</td>
<td>Illumina 1.5</td>
<td>101bp</td>
<td>150</td>
<td>Pair Ended</td>
</tr>
<tr>
<td>s2_R1_PE.fastq</td>
<td>24GB</td>
<td>Illumina 1.5</td>
<td>101bp</td>
<td>150</td>
<td>Pair Ended</td>
</tr>
<tr>
<td>s2_R2_PE.fastq</td>
<td>24GB</td>
<td>Illumina 1.5</td>
<td>101bp</td>
<td>150</td>
<td>Pair Ended</td>
</tr>
</tbody>
</table>
de novo Assembly

Coverage

Coverage
Easy calculation:
(# reads x avg read length) / genome size
So, for haploid human genome:
30m reads x 100 bp = 3 bn
“1x” doesn’t mean every DNA sequence is read once. It means that, if sampling were systematic, it would be. Sampling isn’t systematic, it’s random!
de novo Assembly

Coverage

How much data do I need?!?

Often, a simple organism with little or no “repeats“, 20X – 60X of Illumina is enough.

Different read lengths and insert sizes in the mix help as well.

As things get larger and repeat regions increase in number, long reads help a lot!

A generally accepted “rule of thumb“ seems to be ~80X Illumina (PE + MP of varying lengths and insert sizes) plus 20-40X PacBio for a Hybrid approach.
de novo Assembly

Coverage

[jpummil@razor-l3 data]$ more PE-350.1.fastq
@DRR001841.41/1
AAAAGAATGGAAATCTATGTTTTTATTATTACAAGTTTTTGAGATTGCAGACTGAAAATCAAATGATTTTCTGAGATTGAATTGAAGTCATCGGGTC
+
CCCCBBBBBBBCCCCCCCCCBBBBBBBBBBBBBBBBBBBBBBBBBBBBCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCMaryLeeC1@BCC

@DRR001841.45/1
ACGACTTTGATCATTGTCCTCAATAACTTTATTATTAGATCTGCTCACGGAAGTGCAGCTATTCTACGAAGACGAAAAGTCTGCCTAA
+
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCAC3B5B=
de novo Assembly

Coverage

Genome size = 12,000,000

wc -l PE-350.1.fastq
381536 PE-350.1.fastq

wc -l PE-350.2.fastq
381536 PE-350.2.fastq

(381536+381536)/4 = 190768 reads

AND...

head -n 1998 PE-350.1.fastq | tail -n 1 | wc -c
91 avg read length

SO...

(190768*91)/12,000,000 = 1.44x coverage
de novo Assembly

Example Software Stack

- FastQC – Quality Assessment tool for sequencer data
- BBMap – short read aligner and other bioinformatic tools
- SOAPdenovo2 – assembler for de novo assembly of NGS data
- SPAdes – assembler for de novo assembly of NGS data
- Velvet – assembler for short read NGS data
- Quast – Quality assessment tool for genome assemblies
de novo Assembly

Data Quality Assessment

Do the following:

Make sure you are in...
/storage/<username>/Workshop

Open and modify as appropriate the file called FastQC.pbs
#!/bin/bash

#PBS -N FastQC
#PBS -j oe
#PBS -m abe
#PBS -M <username>@gmail.com
#PBS -o FastQC.$PBS_JOBID
#PBS -q XXX
#PBS -l nodes=1:ppn=2
#PBS -l walltime=00:10:00

cd $PBS_O_WORKDIR

module purge
module load fastqc

export DATA=/scratch/<username>/Workshop/data

mkdir FastQC

fastqc $DATA/PE-350.1.fastq $DATA/PE-350.2.fastq -t 2 -o FastQC/
Do the following:

qsub FastQC.pbs

Wait for email response to signal job complete...

cd FastQC

ls

PE-350.1_fastqc.html  PE-350.1_fastqc.zip  PE-350.2_fastqc.html  PE-350.2_fastqc.zip
de novo Assembly

Data Quality Assessment

Do the following:

Transfer the files you just created back to your local machine for viewing

Mac or Linux (from local machine):
scp <username>@ razor.uark.edu:/scratch/<username>/Workshop/FastQC/*.

Windows:
Filezilla or Putty PSCP (http://the.earth.li/~sgtatham/putty/latest/x86/pscp.exe)
De novo Assembly

Data Quality Assessment

Do the following:

On your local system, double click one of the files ending in .html that you downloaded...
de novo Assembly

BBMap Toolkit

A Bioinformatics utility suite written and maintained by Brian Bushnell at Joint Genome Institute (DoE)

This package includes BBMap, a short read aligner, as well as various other bioinformatic tools. It is written in pure Java, can run on any platform, and has no dependencies other than Java being installed (compiled for Java 6 and higher). All tools are efficient and multithreaded.
de novo Assembly

BBMap Toolkit

To look at some of the unique tools, first do:

module load bbmap

Then, type the full name of any command below to get a description and usage info:
de novo Assembly

BBMap Toolkit

Do the following:

Make sure you are in...
/storage/<username>/Workshop

Open and modify as appropriate the file called BBMap.pbs
de novo Assembly

BBMap Toolkit
(removing adapters)

#!/bin/bash
#
#PBS -N BBNorm
#PBS -q XXX
#PBS -j oe
#PBS -m abe
#PBS -M <username>@gmail.com
#PBS -o BBNorm.$PBS_JOBID
#PBS -l nodes=1:ppn=2
#PBS -l walltime=00:30:00

module purge
module load bbmap

cd $PBS_O_WORKDIR

export DATA=/storage/<username>/Workshop/data

bbduk.sh in1=$DATA/PE-350.1.fastq in2=$DATA/PE-350.2.fastq out1=$DATA/PE-350.1.adap.fastq out2=$DATA/PE-350.2.adap.fastq ref=adapters.fasta ktrim=r mink=10
de novo Assembly

BBMap Toolkit
(removing adapters)
adapters.fasta

You can put them in a properly-formatted fasta file, like this:

>1
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
>2
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
>3
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
>4
AGATCGGAAGAGC

Also, some stored in:
ls /share/apps/bbmap/bbmap/resources/
nextera.fa.gz primes.txt.gz sample2.fq.gz
phix174_ill.ref.fa.gz sample1.fq.gz truseq.fa.gz
de novo Assembly

BBMap Toolkit
(removing adapters)

Note that BBDuk uses kmers, and the default kmer length is 28; it will not find adapters shorter than kmer length. You can change it to, say, 13 with the "k=13" flag, but the shorter it is the more false positives will be found, particularly if you allow mismatches.
#!/bin/bash

#PBS -N BBNorm
#PBS -q XXX
#PBS -j oe
#PBS -m abe
#PBS -M <username>@gmail.com
#PBS -o BBNorm.$PBS_JOBID
#PBS -l nodes=1:ppn=2
#PBS -l walltime=00:30:00

module purge
module load bbmap

cd $PBS_O_WORKDIR

export DATA=/storage/<username>/Workshop/data

reformat.sh in1=$DATA/PE-350.1.fastq in2=$DATA/PE-350.2.fastq out1=$DATA/PE-350.1.trim.fastq out2=$DATA/PE-350.2.trim.fastq outsingle=singletons.fq qtrim=rl trimq=10 minlength=50
de novo Assembly

#!/bin/bash
#
#PBS -N BBNorm
#PBS -q XXX
#PBS -j oe
#PBS -m abe
#PBS -M <username>@gmail.com
#PBS -o BBNorm.$PBS_JOBID
#PBS -l nodes=1:ppn=2
#PBS -l walltime=00:30:00

module purge
module load bbmap

cd $PBS_O_WORKDIR

export DATA=/storage/<username>/Workshop/data

bbnorm.sh in1=$DATA/PE-350.1.fastq in2=$DATA/PE-350.2.fastq out1=$DATA/PE-350.1.norm.fastq out2=$DATA/PE-350.2.norm.fastq target=99999999 min=9 passes=1
de novo Assembly

Assembler Software...decisions, decisions
de novo Assembly

Assembler Software...decisions, decisions

A survey of popular ‘omics’ assembly tools

Keith Bradnam
UC Davis Genome Center
July 2014
de novo Assembly

Assembler Software...decisions, decisions

Table 1: Primary assembly tools for eukaryotic genome assembly

<table>
<thead>
<tr>
<th>Assembler</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALLPATHS-LG (2010)</td>
<td>16</td>
</tr>
<tr>
<td>ABySS (2008)</td>
<td>13</td>
</tr>
<tr>
<td>SOAPdenovo (2009)</td>
<td>10</td>
</tr>
<tr>
<td>Velvet (2007)</td>
<td>10</td>
</tr>
<tr>
<td>SPAdes (2012)</td>
<td>7</td>
</tr>
<tr>
<td>Ray (2010)</td>
<td>6</td>
</tr>
<tr>
<td>Celera (2004)</td>
<td>5</td>
</tr>
<tr>
<td>CLC (2008)</td>
<td>5</td>
</tr>
<tr>
<td>MaSuRCA (2012)</td>
<td>5</td>
</tr>
</tbody>
</table>
de novo Assembly

Assembler Software...decisions, decisions

Table 2: Primary assembly tools for bacterial/archaeal genome assembly

<table>
<thead>
<tr>
<th>Assembler</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPAdes (2012)</td>
<td>29</td>
</tr>
<tr>
<td>Velvet (2007)</td>
<td>14</td>
</tr>
<tr>
<td>ABySS (2008)</td>
<td>6</td>
</tr>
<tr>
<td>ALLPATHS-LG (2010)</td>
<td>6</td>
</tr>
<tr>
<td>HGAN (2013)(^1)</td>
<td>6</td>
</tr>
<tr>
<td>MIRA (1998?)(^2)</td>
<td>6</td>
</tr>
<tr>
<td>SOAPdenovo (2009)</td>
<td>4</td>
</tr>
</tbody>
</table>
de novo Assembly

Assembler Software...decisions, decisions

Table 3: Primary assembly tools for transcriptome genome assembly

<table>
<thead>
<tr>
<th>Assembler</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trinity (2011)</td>
<td>57</td>
</tr>
<tr>
<td>SOAPdenovo-Trans (2011)</td>
<td>5</td>
</tr>
<tr>
<td>Trans-ABySS (2010)</td>
<td>4</td>
</tr>
<tr>
<td>CLC (2008)</td>
<td>3</td>
</tr>
<tr>
<td>MIRA</td>
<td>2</td>
</tr>
<tr>
<td>Oases (2010)</td>
<td>2</td>
</tr>
</tbody>
</table>
Figure 1: Frequency of assemblies generated by survey respondents
de novo Assembly

We'll be using the following assemblers in the workshop:

- SOAPdenovo – Short Oligonucleotide Analysis Package
  - [http://soap.genomics.org.cn/soapdenovo.html](http://soap.genomics.org.cn/soapdenovo.html)

- SPAdes – St. Petersburg genome Assembler

- Velvet – Sequence assembler for very short reads
  - [https://www.ebi.ac.uk/~zerbino/velvet/](https://www.ebi.ac.uk/~zerbino/velvet/)
The term k-mer typically refers to all the possible substrings, of length k, that are contained in a string. In Computational genomics, k-mers refer to all the possible subsequences (of length k) from a read obtained through DNA Sequencing. The amount of k-mers possible given a string of length, L, is L-k+1 whilst the amount of possible k-mers given n possibilities (4 in the case of DNA eg. ACTG) is n^k.
de novo Assembly

Assembly

Do the following:

Make sure you are in...
/storage/<username>/Workshop

Open and modify as appropriate the file called SOAP.config
de novo Assembly

SOAPdenovo2

[jpummil@razor-l3 Workshop]$ more soap.config
#maximal read length
max_rd_len=91
[LIB]
#average insert size
avg_ins=157
#if sequence needs to be reversed
reverse_seq=0
#if in which part(s) the reads are used
asm_flags=3
#if use only first 100 bps of each read
rd_len_cutoff=90
#if in which order the reads are used while scaffolding
rank=1
#if cutoff of pair number for a reliable connection (at least 3 for short insert size)
pair_num_cutoff=3
#if minimum aligned length to contigs for a reliable read location (at least 32 for short insert size)
map_len=32
#if a pair of fastq file, read 1 file should always be followed by read 2 file
q1=/scratch/<username>/Workshop/data/L001_R1_001_Sub.fastq
q2=/scratch/<username>/Workshop/data/L001_R2_001_Sub.fastq
de novo Assembly

Assembly

Do the following:

Make sure you are in...
/storage/<username>/Workshop

Open and modify as appropriate the file called SOAP.pbs
#PBS -N SOAPdenovo2
#PBS -q XXX
#PBS -j oe
#PBS -m abe
#PBS -M <username>@gmail.com
#PBS -o SOAP.$PBS_JOBID
#PBS -l nodes=1:ppn=2
#PBS -l walltime=00:30:00

module purge
module load gcc/4.6.3
module load soapdenovo2

cd $PBS_O_WORKDIR

mkdir SOAP-27

mkdir SOAP-27

SOAPdenovo-63mer all -F -p 2 -s SOAP.config -o SOAP-27/test -K 27
de novo Assembly

Assembly

Do the following:

Make sure you are in...
/storage/<username>/Workshop

Open and modify as appropriate the file called SPAdes.pbs
#!/bin/bash

#PBS -N SPAdes
#PBS -q XXX
#PBS -j oe
#PBS -m abe
#PBS -M <username>@gmail.com
#PBS -o SPades.$PBS_JOBID
#PBS -l nodes=1:ppn=2
#PBS -l walltime=00:30:00

module purge
module load gcc/4.8.2
module load spades

cd $PBS_O_WORKDIR

export DATA=/scratch/<username>/Workshop/data/

export OMP_NUM_THREADS=2

spades.py -t 2 -k 27 --sc --pe1-1 $DATA/PE-350.1.fastq --pe1-2 $DATA/PE-350.2.fastq -o SPAdes-27
de novo Assembly

Assembly

Do the following:

Make sure you are in...
/storage/<username>/Workshop

Open and modify as appropriate the file called Velvet.pbs
de novo Assembly

```bash
#!/bin/bash
#PBS -N Velvet
#PBS -q XXX
#PBS -j oe
#PBS -m abe
#PBS -M <username>@gmail.com
#PBS -o Velvet.$PBS_JOBID
#PBS -l nodes=1:ppn=2
#PBS -l walltime=00:30:00

module purge
module load gcc/4.6.3
module load velvet

cd $PBS_O_WORKDIR

export DATA=/scratch/<username>/Workshop/data/

export OMP_NUM_THREADS=2


velvetg Velvet-27 -cov_cutoff auto -exp_cov auto -cov_cutoff 5 -exp_cov 40 -ins_length 157 -min_contig_lgth 90
```
de novo Assembly

Command Line Variation

SOAPdenovo-63mer all -F -p 4 -s SOAP.config -o SOAP-27/test -K 27

VS

spades.py -t 4 -k 27 --sc --pe1-1 $DATA/PE-350.1.fastq --pe1-2 $DATA/PE-350.2.fastq -o SPAdes-27

VS

&
velvetg Velvet-27 -cov_cutoff auto -exp_cov auto -cov_cutoff 5 -exp_cov 40 -ins_length 157 -min_contig_lgth 90
de novo Assembly

Quality Assessment of Assemblies

So, NOW we have three unique assemblies generated from three different assemblers:

[jpummil@razor-l3 Workshop]$ ls -l SOAP-27/test.contig
-rw-rw-- 1 jpummil jpummil 661287 Apr  7 13:33 SOAP-27/test.contig

[jpummil@razor-l3 Workshop]$ ls -l SPAdes-27/contigs.fasta
-rw-rw-- 1 jpummil jpummil 577338 Apr  7 13:33 SPAdes-27/contigs.fasta

[jpummil@razor-l3 Workshop]$ ls -l Velvet-27/contigs.fa
-rw-rw-- 1 jpummil jpummil 576655 Apr  7 13:33 Velvet-27/contigs.fa

Which one's the “best”?!?!
de novo Assembly

Quality Assessment of Assemblies

QUAST
QUality ASsessment Tool for Genome Assemblies
Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler
St. Petersburg Academic University of the Russian Academy of Sciences
de novo Assembly

Quality Assessment of Assemblies

#!/bin/bash
#
#PBS -N Quast
#PBS -j oe
#PBS -m abe
#PBS -M <username>@gmail.com
#PBS -o Quast.$PBS_JOBID
#PBS -q XXX
#PBS -l nodes=1:ppn=12
#PBS -l walltime=00:30:00

module purge
module load gcc/4.6.3 python/2.7.5 mkl/13.1.0
module load quast

cd $PBS_O_WORKDIR
export DATA=/scratch/<username>/Workshop

de novo Assembly

Quality Assessment of Assemblies

Do the following:

Transfer the files you just created back to your local machine for viewing

Mac or Linux (from local machine):
scp <username>@razor.uark.edu:/scratch/<username>/Workshop/quast_results/latest/*.pdf .

Windows:
Filezilla or Putty PSCP (http://the.earth.li/~sgtatham/putty/latest/x86/pscp.exe)
**de novo Assembly**

**Quality Assessment of Assemblies**

### Report

<table>
<thead>
<tr>
<th></th>
<th>test.contig</th>
<th>SPAdes-27.contigs</th>
<th>Velvet-27_contigs</th>
</tr>
</thead>
<tbody>
<tr>
<td># contigs (&gt;= 0 bp)</td>
<td>1118</td>
<td>48</td>
<td>66</td>
</tr>
<tr>
<td># contigs (&gt;= 1000 bp)</td>
<td>153</td>
<td>20</td>
<td>38</td>
</tr>
<tr>
<td>Total length (&gt;= 0 bp)</td>
<td>620734</td>
<td>566085</td>
<td>564907</td>
</tr>
<tr>
<td>Total length (&gt;= 1000 bp)</td>
<td>521715</td>
<td>561438</td>
<td>557119</td>
</tr>
<tr>
<td># contigs</td>
<td>185</td>
<td>22</td>
<td>41</td>
</tr>
<tr>
<td>Largest contig</td>
<td>14104</td>
<td>350161</td>
<td>65325</td>
</tr>
<tr>
<td>Total length</td>
<td>546696</td>
<td>562919</td>
<td>559466</td>
</tr>
<tr>
<td>GC (%)</td>
<td>38.49</td>
<td>38.44</td>
<td>38.47</td>
</tr>
<tr>
<td>N50</td>
<td>4137</td>
<td>90982</td>
<td>22511</td>
</tr>
<tr>
<td>N75</td>
<td>2389</td>
<td>42859</td>
<td>13982</td>
</tr>
<tr>
<td>L50</td>
<td>43</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>L75</td>
<td>86</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td># N's per 100 kbp</td>
<td>0.00</td>
<td>0.00</td>
<td>83.65</td>
</tr>
</tbody>
</table>
de novo Assembly

Quality Assessment of Assemblies

N50

Given a set of contigs, each with its own length...
N50 is defined as the length for which the collection
Of all the contigs of that length or longer contains
At least half of the sum of the length of all the contigs.
# de novo Assembly

## Quality Assessment of Assemblies

<table>
<thead>
<tr>
<th>Method</th>
<th>N50</th>
<th># of Contigs</th>
<th>Longest Contig</th>
<th>Overall Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOAPdenovo2 (k=27)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOAPdenovo2 (k= )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPAdes (k=27)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPAdes (k= )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Velvet (k=27)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Velvet (k= )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
de novo Assembly

Quality Assessment of Assemblies

Mauve Alignment
Continuity is not the only way of looking at assembly quality; it's also useful to map the input reads to the assembly to determine the percent mapped (higher is better) and number of mismatches/indels (lower is better). Also, running gene prediction to try to find broken genes can sometimes help indicate assembly quality.
de novo Assembly
Quality Assessment of Assemblies

BLAST Search

Saccharomyces Genome Database
S. cerevisiae WU-BLAST2 Search

http://www.yeastgenome.org/cgi-bin/blast-sgd.pl
de novo Assembly

Quality Assessment of Assemblies

BLAST Search
de novo Assembly

General Obstacles Encountered

- Data sets of this size are cumbersome and tedious to work with given current tools available
- Storage of the various stages of modified data is a problem due to system disk capacity
- Verifying data integrity can be difficult as there is no 'reference' to compare to.
- Some software tools require specialized machines due to extreme demands for memory
- The age old data question...“What to keep, what to discard?”
**de novo Assembly**

**Additional Help - Forums**

- SEQAnswers – The Next Generation Sequencing Community
  - [http://seqanswers.com](http://seqanswers.com)

- Biostars – Bioinformatics Explained
  - [http://www.biostars.org](http://www.biostars.org)
de novo Assembly

Additional Help - Twitter

- Aaron Quinlan @aaronquinlan
- Robert Davey @froggleston
- Keith Bradnam @kbradnam
- Mick Watson @biomickwatson
- Michael Schatz @mike_schatz
- Shaun Jackman @sjackman
- Tracy Teal @tracykteal
- Adam Phillippy @aphillippy
- Eugene Myers @TheGeneMyers
- Jared Simpson @jaredtsimpson
- Nick Loman @pathogenomenick
- Torsten Seemann @torstenseemann
- Ewan Birney @ewanbirney
- Titus Brown @ctitusbrown
de novo Assembly

Jeff Pummill
Director – AHPCC
XSEDE Staff
Arkansas High Performance Computing Center
University of Arkansas, Fayetteville, AR 72701
jpummil@uark.edu
@jpummil on Twitter