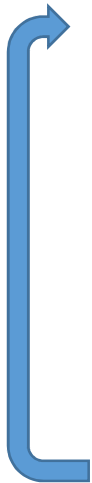


Workflow of de novo assembly

- Experimental Design
- Clean sequencing data
- Run assembly software for contiging and scaffolding
- Evaluation of assembly
- Gap closing
- Anchor to chromosome (optional)

Iterations



Experimental design using Illumina Platform

Estimate genome size:

500 mb

Platform: Illumina Hiseq

Paired-end library: 150bp x 2 ; 2 lanes; >100x coverage;

Mate pair library: three libraries (5kb, 10kb, 15kb), run on 2 lanes

Software:

Soap denovo

Abyss

AllPath-LG (require overlapping reads)

Platanus (heterozygous genome)

MaSuRCA (hybrid, computationally demanding)

Large memory computer

BioHPC lab large memory server: 512mb RAM 64-core

Data cleaning

- Trim low quality data (quality score based trimming)
- Clip sequencing adapters (alignment to adapter sequence)

```
java -jar /programs/trimmomatic/trimmomatic-0.32.jar PE -phred33 \  
SRR1554178_1.fastq SRR1554178_2.fastq \  
r1.fastq u1.fastq r2.fastq u2.fastq \  
ILLUMINACLIP:/programs/trimmomatic/adapters/TruSeq3-PE-2.fa:2:30:10  
LEADING:10 \  
TRAILING:10 \  
SLIDINGWINDOW:4:15 \  
MINLEN:50
```

Trimmomatic

Input

Output

R1.fastq R2.fastq

Paired1 Unpaired1 Paired2 Unpaired2



Palindrome clip mode

r1.fastq	417503786
r2.fastq	411903328
u1.fastq	62712666
u2.fastq	2776034

Kmer coverage based read error correction

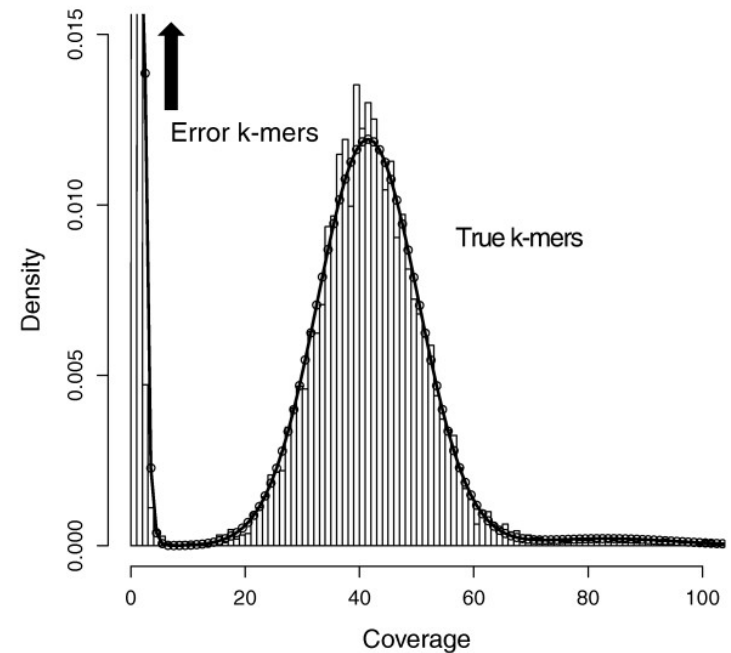
- Identify reads containing untrusted k -mers
- Either correct reads with errors so that all k -mers are trusted or simply discard these reads

Quake

<http://www.cbcbl.umd.edu/software/quake/>

SOAPec_v2.01

<http://soap.genomics.org.cn/soapdenovo.html>



Running assembly software

Testing different size kmers and assembly software.

Not always possible, as assembly of large genomes takes very long time on a large memory computer.

SOADdenovo config file

```
#maximal read length
max_rd_len=101
[LIB]
#average insert size
avg_ins=300
#if sequence needs to be reversed
reverse_seq=0
#in which part(s) the reads are used
asm_flags=3
#in which order the reads are used while scaffolding
rank=1
# cutoff of pair number for a reliable connection (at least 3 for short insert size)
pair_num_cutoff=3
#minimum aligned length to contigs for a reliable read location (at least 32 for short insert s
map_len=32
#a pair of fastq file, read 1 file should always be followed by read 2 file
q1=r1.fastq
q2=r2.fastq
```

/programs/SOAPdenovo2/SOAPdenovo-127mer all -s config.txt -K 127 -R -o assembly

avg_ins=2000 reverse_seq=1

Multiple libraries can be mixed in one assembly

```
[LIB]
avg_ins=450
reverse_seq=0
asm_flags=3
q1=r1.fastq
q2=r2.fastq
[LIB]
asm_flags=1
q=u1.fastq
[LIB]
avg_ins=2000
reverse_seq=1
asm_flags=3
q1=r1.fastq
q2=r2.fastq
```


Using different kmer size or mixed kmer size

SOAPdenovo-127mer all -s config.txt -K 101 -R -o assembly

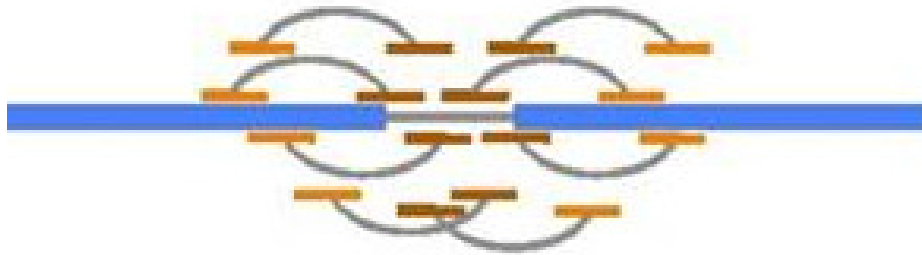
SOAPdenovo-127mer all -s config.txt -K 127 -R -o assembly

SOAPdenovo-127mer all -s config.txt -K 101 -m 127 -R -o assembly

ABySS

```
abyss-pe k=127 \  
  name=abyss_contig \  
  lib='pe1 pe2' \  
  mp='mp1 mp2' \  
  pe1='pe1_1.fq pe1_2.fq' \  
  pe2='pe2_1.fq pe2_2.fq' \  
  mp1='mp1_1.fa mp1_2.fa' \  
  mp2='mp2_1.fa mp2_2.fa'
```

Running Gap closing software



Software:

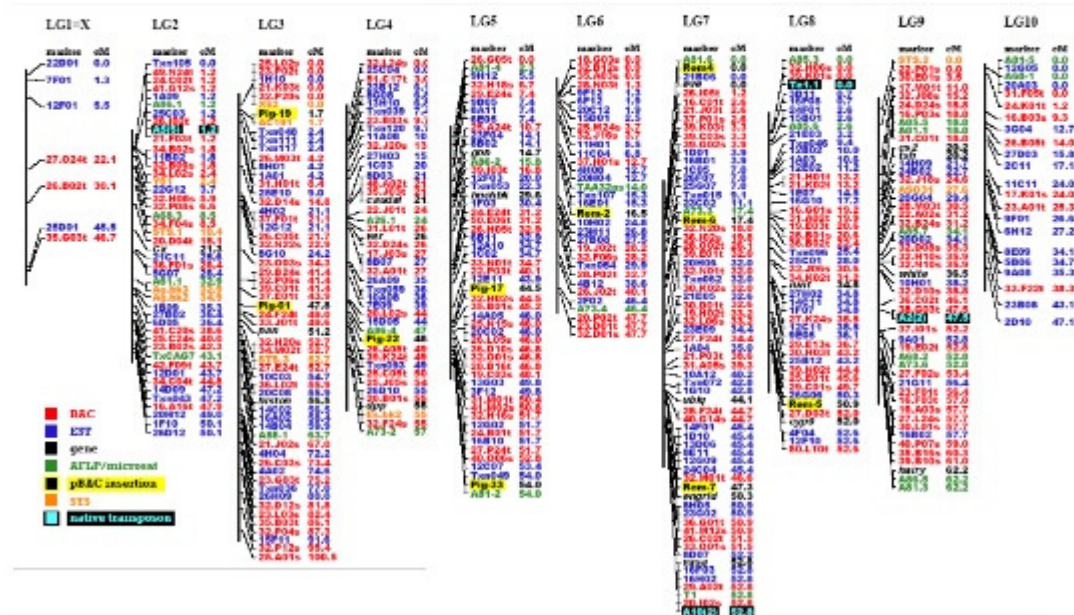
SOAPdenovo GapCloser: <http://sourceforge.net/projects/soapdenovo2/files/GapCloser/>

IMAGE: <http://sourceforge.net/projects/image2/files/>

Can be run multiple iterations to close the gap.

Using Physical Map to Anchor Scaffold to Chromosome

Molecular map markers used to anchor scaffolds to Chromosome builds



Few X markers, no Y, variable marker density

BioNano Map

<http://www.slideshare.net/kstatebioinformatics/using-bionano-maps-to-improve-an-insect-genome-assembly>

Evaluation of Genome assembly 1

Metrics for contig length

N50 and L50 *

N50 scaffold/contig length is calculated by summing lengths of scaffolds/contigs from the longest to the shortest and determining at what point you reach 50% of the total assembly size. The length of the scaffold/contig at that point is the N50 length.

L50 measure is the *number* of scaffolds/contigs that are greater than, or equal to, the N50 length.

NG50 and LG50

The **NG50** and **LG50** measures are the same as the N50 and L50 measures except that rather than compare against the total assembly size

- This is the definition from Assemblathon 2. There is a growing trend to switch the N50 and L50 definition.

Standalone tools for generating metrics

(Most assembly software provides N50/L50 metrics in the report)

1. Quast (<http://bioinf.spbau.ru/quast>)

- Contig size
- Comparison with a reference genome.
 - Structure variation/misasassembly.
 - Genome fraction: % represented reference genome
 - Duplication ration: copy number ratio between assembly and reference in aligned region.
 - Reference gene representation .

2. REAPR: Scoring each base of the assembly based on alignment of paired-end reads

Input:

BAM file from alignment of reads to the assembly (independent alignment of paired ends)

Metrics reported by REAPR:

- Scaffold errors
- % of error free bases
- Corrected N50

3. Evaluate by gene content

CEGMA

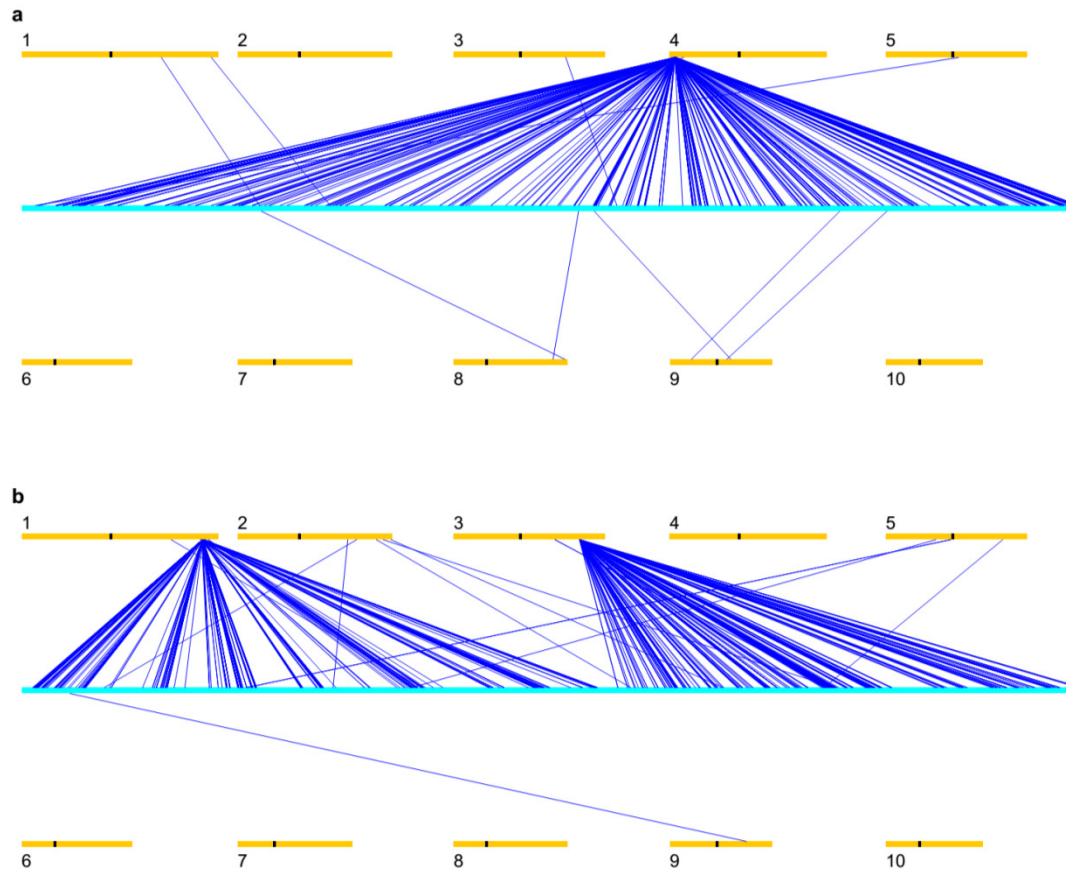
- A pre-define a set of 458 conserved core eucaryoptic proteins that are present in a wide range of taxa;
- HMMER and BLAST+ based identification of the core gene set in the newly assembled genome.

QUAST

- Compare with a closely related reference genome.

Evaluate based on genetic mapping

Use mapped GBS sequence tags to evaluate each contig



Fei Lu, Buckler lab

<http://www.nature.com/ncomms/2015/150416/ncomms7914/full/ncomms7914.html>