

A Roadmap to the De-novo Assembly of the Banana Slug Genome

Stefan Prost ¹

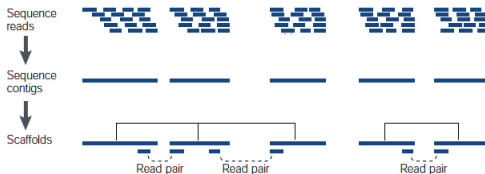
¹Department of Integrative Biology, University of California, Berkeley, United
States of America

April 6th-10th, 2015

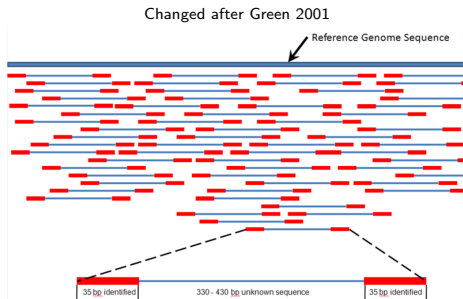


- 1 *A priori* Information about the Genome
- 2 Sequencing Strategies and Platforms
- 3 Sequencing Libraries
- 4 Raw Data Processing and Quality Assessment
- 5 Assembly Strategies and Tools
- 6 Assembly Quality Assessment
- 7 Further Improvement of the Assembly
- 8 What is a finished Assembly?
- 9 Downstream Processing

De-novo Assembly



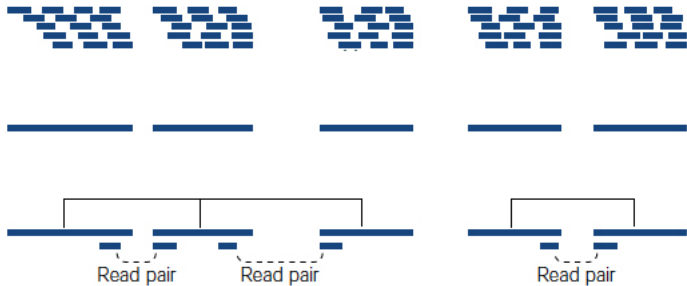
Reference-based Mapping



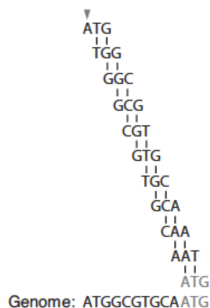
Wikipedia

Sequence
readsSequence
contigs

Scaffolds



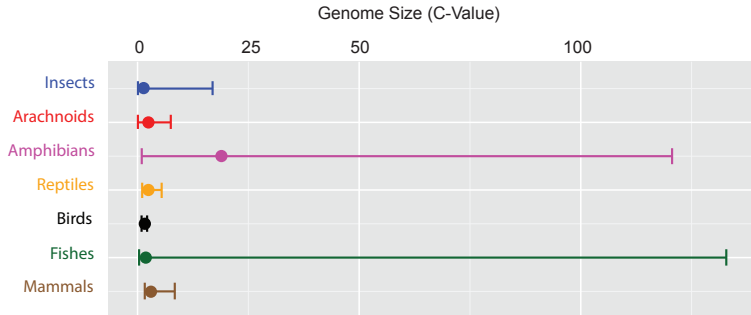
Changed after Green 2001



Definition: Kmer

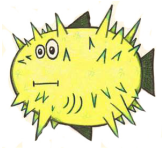
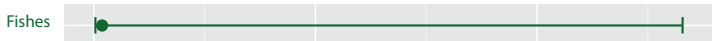
Short, unique element of DNA sequence of length n

- Expected Genome Size
- Expected Repeat Content
- Expected Heterozygosity
- Diploid or Polyploid?



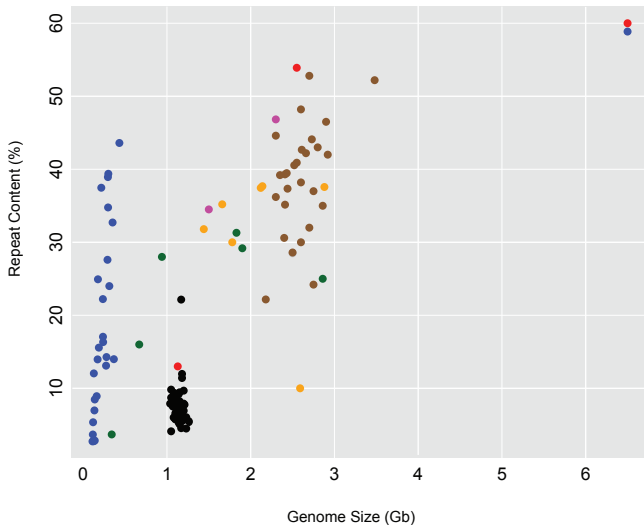
C-values obtained from: www.genomesize.com/

Genome Size Range in Fishes

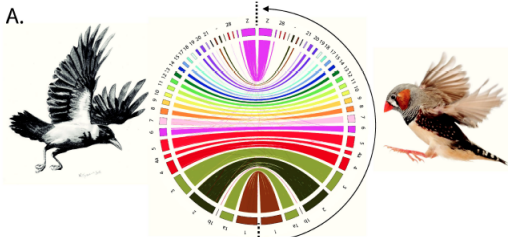


(1) A priori Information about the Genome

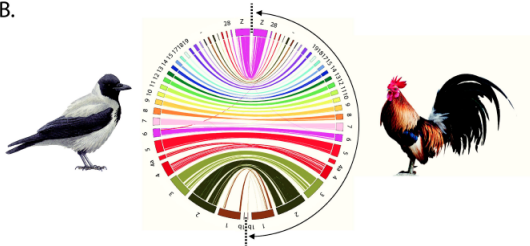
Genome Size and Repeat Content



A.



B.



Poelstra et al. 2014

■ Helpful Online Databases:

■ Genomes

- www.ncbi.nlm.nih.gov/genome/browse
- www.gigadb.org

■ Genome Sizes (C-values)

- www.genomesize.com

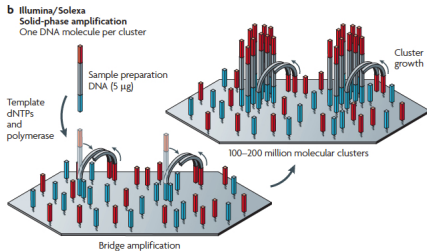
- Expected Genome Size
- Expected Repeat Content
- Expected Heterozygosity
- Diploid or Polyploid?

- First Generation Sequencing
 - Sanger Sequencing

- Second Generation Sequencing (*PCR Needed*)
 - **Illumina**: MiSeq & HiSeq
 - Roche: 454
 - Life Science: IONtorrent & IONproton
 - ABI: SOLiD

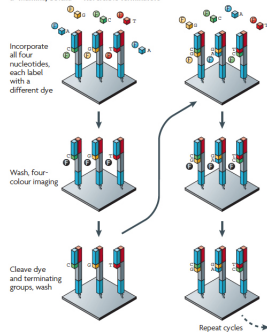
- Third Generation Sequencing (*Single Molecule Sequencing*)
 - Helicos Biosciences: Heliscope
 - **Pacific Biosciences**: PacBio RS II
 - **Oxford Nanopore**: MinION & GridION

Illumina Sequencing



Metzker 2010

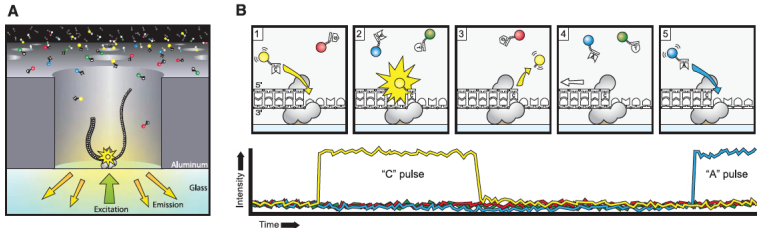
a Illumina/Solexa — Reversible terminators



Top: CATCGT
Bottom: CCCCCC

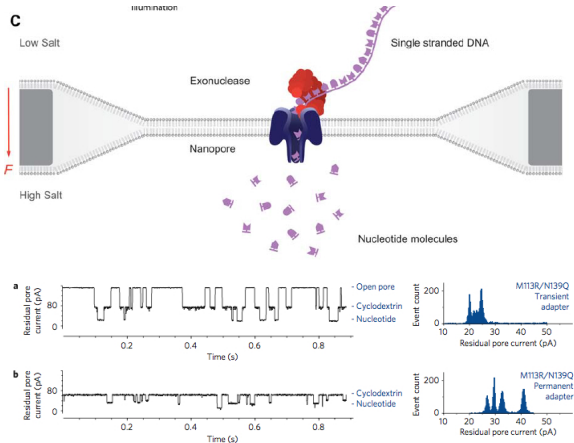
Metzker 2010

Pacific Biosciences



Eid et al. 2009

Oxford Nanopore



Schadt et al. 2010; Clarke et al. 2009

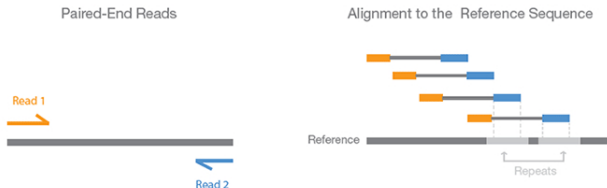
- First Generation Sequencing
 - Sanger Sequencing

- Second Generation Sequencing (*PCR Needed*)
 - Illumina: MiSeq & HiSeq
 - Roche: 454
 - Life Science: IONtorrent & IONproton
 - ABI: SOLiD

- Third Generation Sequencing (*Single Molecule Sequencing*)
 - Helicos Biosciences: Heliscope
 - Pacific Biosciences: PacBio RS II
 - Oxford Nanopore: MinION & GridION

Illumina Paired-end Sequencing Libraries

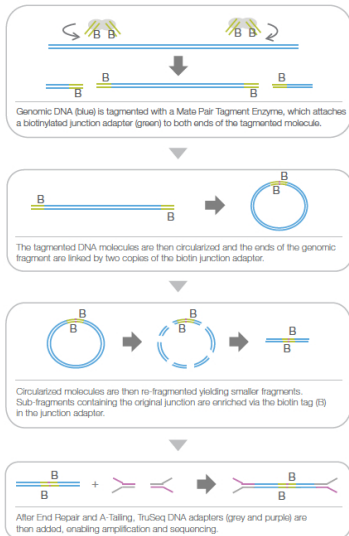
Figure 4. Paired-End Sequencing and Alignment



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.

[Illumina Homepage](#)

Illumina Mate Pair Sequencing Libraries



Allpaths-LG Recipe

Libraries (insert types)	Fragment size (bp)	Read length (bases)	Sequence coverage (x)	Required
Fragment	180*	≥ 100	45	yes
Short jump	3,000	≥ 100 preferable	45	yes
Long jump	6,000	≥ 100 preferable	5	no**
Fosmid jump	40,000	≥ 26	1	no**

Custom Recipe for 1.3Gb Bird Genome (<15% Repeats)

- 180bp PE - 1 lane HiSeq (30x coverage)
- 650bp PE - 1 lane HiSeq (30x coverage)
- 5kb/8kb MP - 1 lane HiSeq (20x coverage)

Custom Recipe for 2.5Gb Mammal Genome (<50% Repeats)

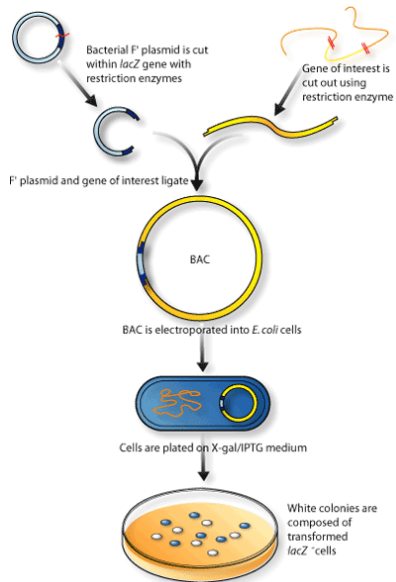
- 180bp PE - 4 lanes HiSeq (50x coverage)
- 3kb MP - 4 lanes HiSeq (40x coverage)
- 8kb/15kb MP - 1 lane (10x coverage)

Fosmid Vector

- Bacterial F-plasmid
- <40kb Insert Size

Bacterial Artificial Chromosome (BAC)

- <300kb Insert size



Read Quality Assessment

- Base Quality

Phred Score: $Q_{Phred} = -10 \log_{10} P(\text{error})$

e.g. a Phred score of 20 translates to a 1% error rate

- GC Content
- Sequence Duplication Levels
- Kmer Content
- ...

Software Tools

- FastQC
- Preqc

@HISEQ:119:C42B3ACXX:7:1101:2009:2249 2:N:0:CGACCTG
TCTTGGGGACAGGGAATTCATTCCAAATGAAATCCTCAAAGAACGCCTTTTATTTTACAGGAGGCTGTATATCTTAGCCAAAGTGGTAGATCGGAAGA
+
BB<BBBBBFB<BFF7BBF<BF<FBB<FFF<FFBFF<BFFFFBFBF<7BB<BFBFB<BFFFF<FFFFFF<BBFB<BB<BBBBBBB7<B<BB<77<BBB77

```

SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....
..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....
!"#%&'()*+,-./0123456789;<=>?@ABCDEFGHIJKLMNopqrstuVwxyz{\}|~
|                                     |               |
33                               59   64           73                   104             126
0.....26...31.....40
          -5...0.....9.....40
              0.....9.....40
                  3....9.....40
0.2.....26...31.....41

```

Read Quality Assessment

- Base Quality

Phred Score: $Q_{Phred} = -10 \log_{10} P(\text{error})$

e.g. a Phred score of 20 translates to a 1% error rate

- GC Content
- Sequence Duplication Levels
- Kmer Content
- ...

Software Tools

- FastQC
- Preqc

$$G = \frac{pn(l-k+1)}{\lambda_k}$$

G = Genome Size

pn = proportion of correct reads

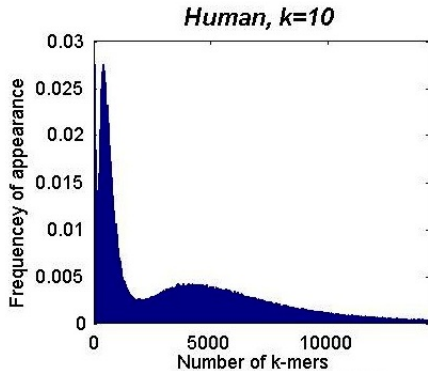
l = read length

k = kmer length

λ_k = mode of the k-mer count histogram

Simpson 2013, arXiv

- Error Correction (EC) using the k-mer spectrum
 - BLESS
 - SGA
 - CUDA
 - DecGPU
 - Euler
 - *Musket*
 - Quake
 - Reptile
- EC using kmer counts
 - **RACER**
 - SHREC
 - HiTEC
 - HSHREC
 - PSAEC
- EC using Multiple Sequence Alignment
 - Coral
 - Echo
 - MyHybrid



- Error Correction (EC) using the k-mer spectrum
 - BLESS
 - SGA
 - CUDA
 - DecGPU
 - Euler
 - *Musket*
 - Quake
 - Reptile
- EC using kmer counts
 - **RACER**
 - SHREC
 - HiTEC
 - HSHREC
 - PSAEC
- EC using Multiple Sequence Alignment
 - Coral
 - Echo
 - MyHybrid

■ Adapter and Low Quality Base Trimming

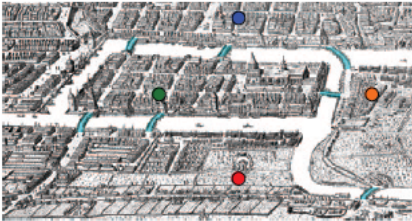
- Skewer (Jiang et al. 2014)
- AdapterRemoval (Lindgreen 2012)
- Trimmomatic (Bolger et al. 2014)

■ Contamination Filtering

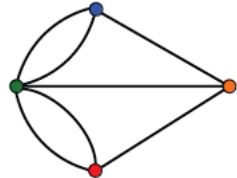
- Blast
- Allpaths-LG
 - Removing Low Frequency k-mers
 - Discarding Scaffolds Shorter than 1kb

Bridges of Königsberg problem (Graph Theory by Euler)

a



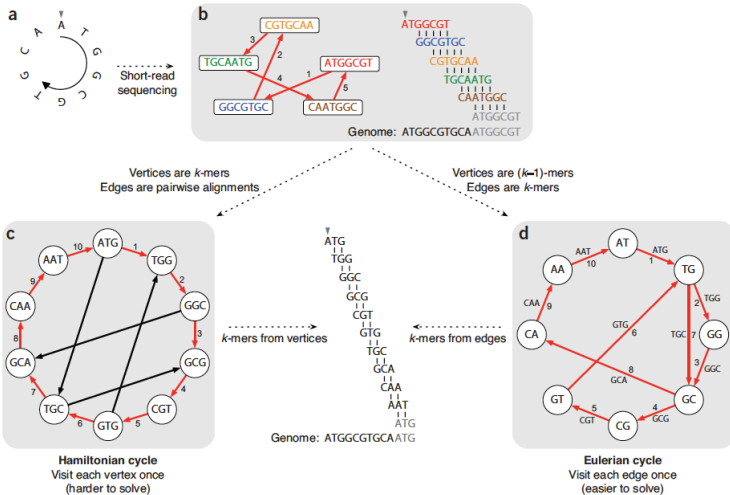
b



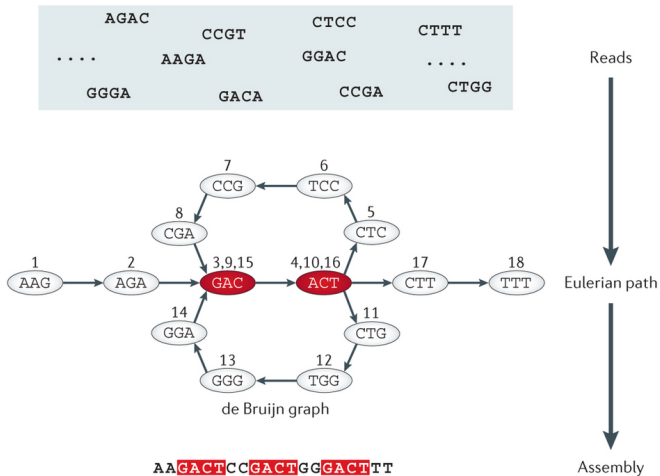
Compeau et al. 2011

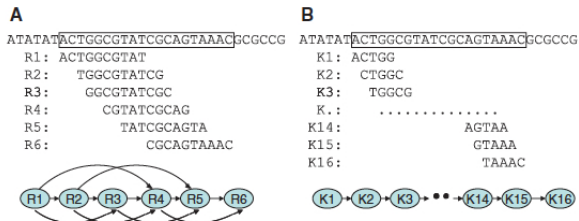
(5) Assembly Strategies and Tools

de Bruijn graph



Compeau et al. 2011



Overlap-layout-consensus vs. *de Bruijn* graph

Li et al. 2011, Briefings in Functional Genomics

Overlap-layout-consensus based Tools:

SGA (Simpson and Durbin 2012), Celera assembler (Myers et al. 2002)

de Bruijn graph based Tools:

Allpaths-LG (Gnerre et al. 2011), Abyss (Simpson et al. 2009), SOAPdenovo (Luo et al. 2012)

Short Read Assembler

- Allpaths-LG (Gnerre et al. 2011)
- Abyss (Simpson et al. 2009)
- SOAPdenovo (Luo et al. 2012)
- Platanus (Kajitani et al. 2014)
- SSAKE (Warren et al. 2007)
- SGA (Simpson and Durbin 2012)
- Celera assembler (Myers et al. 2002)
- MaSuRCA (Zimin et al. 2013)

Long Read Assembler

- Allpaths-LG (Gnerre et al. 2011)
- DBG2OLC (Ye et al. 2014)
- PBcR - Celera Assembler (Berlin et al. 2014)

Short or Long Read Scaffolder

- SSPACE (Boetzer et al. 2011)
- PBjelly (English et al. 2012)

ABYSS command

```
abyss-pe k=60 np=20 name=little_eagle_k60 lib='pe1' mp='mp1 mp2'
pe1='D23CCACXX_NZGL00439_Little_Eagle_200bp_L006_R1_001.fastq.gz
D23CCACXX_NZGL00439_Little_Eagle_200bp_L006_R2_001.fastq.gz'
mp1='C1WMWACXX_NZGL00439_Little_Eagle_MP5kb_L007_R1_001.fastq.gz
C1WMWACXX_NZGL00439_Little_Eagle_MP5kb_L007_R2_001.fastq.gz'
mp2='HOMTRADXX_NZGL00439_Little_Eagle_MP5kb_L001_R1_001.fastq.gz
HOMTRADXX_NZGL00439_Little_Eagle_MP5kb_L001_R2_001.fastq.gz'
```

Output Files

- little_eagle-unitigs.fa
- little_eagle-contigs.fa
- little_eagle-scaffolds.fa

Get Assembly Stats:

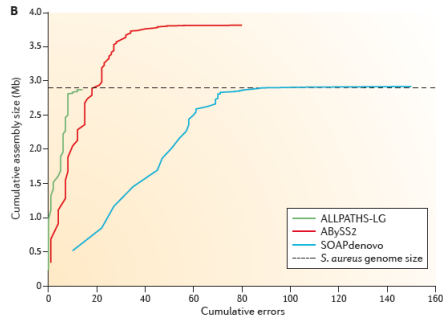
```
abyss-fac little_eagle-scaffolds.fa
```

n	n:500	n:N50	min	N80	N50	N20	max	sum	
577130	20211	526	500	167980	573013	1234088	7692772	1.13e9	little_eagle-scaffolds.fa

- By annotation: CEGMA (Parra et al. 2007)
- Using RNA transcripts: Baa.pl (Ryan 2013/4, Arxiv)
- *De novo* likelihood-based measures (LAP; Ghodsi et al. 2013)
- Feature Response Curve (FRC; Vezzi et al. 2012)
- Assembly Statistics

Definition: **N50** contig length also called **L50**

A contig N50 is calculated by first ordering every contig by length from longest to shortest. Next, starting from the longest contig, the lengths of each contig are summed, until this running sum equals one-half of the total length of all contigs in the assembly. The contig N50 of the assembly is the length of the shortest contig in this list. (Yandel and Ence 2012)



Features:

- Mate-pair Orientations and Separations
- Repeat Content by k-mer Analysis
- Depth-of-coverage
- Correlated Polymorphism in the Read Alignments
- Read Alignment Breakpoints

- By annotation: CEGMA (Parra et al. 2007)
- Using RNA transcripts: Baa.pl (Ryan 2013/4, Arxiv)
- *De novo* likelihood-based measures (LAP; Ghodsi et al. 2013)
- Feature Response Curve (FRC; Vezzi et al. 2012)
- Assembly Statistics

Definition: **N50** contig length also called **L50**

A contig N50 is calculated by first ordering every contig by length from longest to shortest. Next, starting from the longest contig, the lengths of each contig are summed, until this running sum equals one-half of the total length of all contigs in the assembly. The contig N50 of the assembly is the length of the shortest contig in this list. (Yandel and Ence 2012)

Gap Filling

- GapCloser (Luo et al. 2012)
- GapFiller (Nadalin et al. 2011)

Resolving Mis-Assemblies

- REAPR (Hunt et al. 2013)
- NxRepair (Murphy et al. 2014)

Genome Merging

- Metassembler (Wences and Schatz 2015, Arxiv)
- GAM-NGS (Vicedomini et al. 2013)

(7) Further Improvement of the Assembly

REAPR

a Map read pairs to assembly



b Compute per-base statistics

i read coverage



ii type of read coverage, on each strand



iii read clipping



iv fragment coverage



v FCD error



c Score each base



Break assembly



Compute fragment coverage distribution (FCD) error at a given base

No gap present



FCD error



If the base of interest lies in a gap



FCD error



Gap Filling

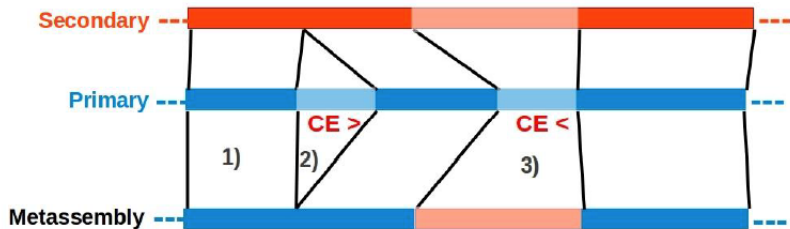
- GapCloser (Luo et al. 2012)
- GapFiller (Nadalin et al. 2011)

Resolving Mis-Assemblies

- REAPR (Hunt et al. 2013)
- NxRepair (Murphy et al. 2014)

Genome Merging

- Metassembler (Wences and Schatz 2015, Arxiv)
- GAM-NGS (Vicedomini et al. 2013)



(Wences and Schatz 2015, Arxiv)

Gap Filling

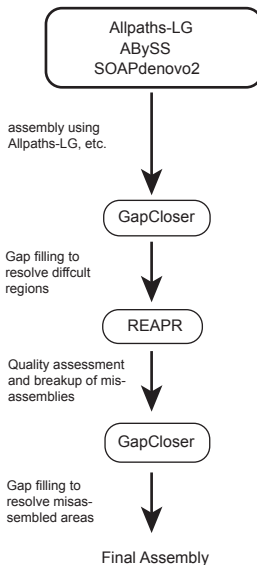
- GapCloser (Luo et al. 2012)
- GapFiller (Nadalin et al. 2011)

Resolving Mis-Assemblies

- REAPR (Hunt et al. 2013)
- NxRepair (Murphy et al. 2014)

Genome Merging

- Metassembler (Wences and Schatz 2015, Arxiv)
- GAM-NGS (Vicedomini et al. 2013)

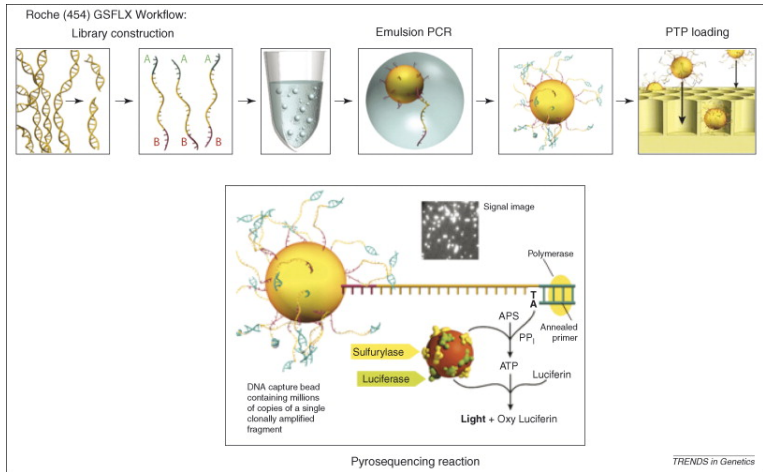


What is a good assembly and when is it finished?

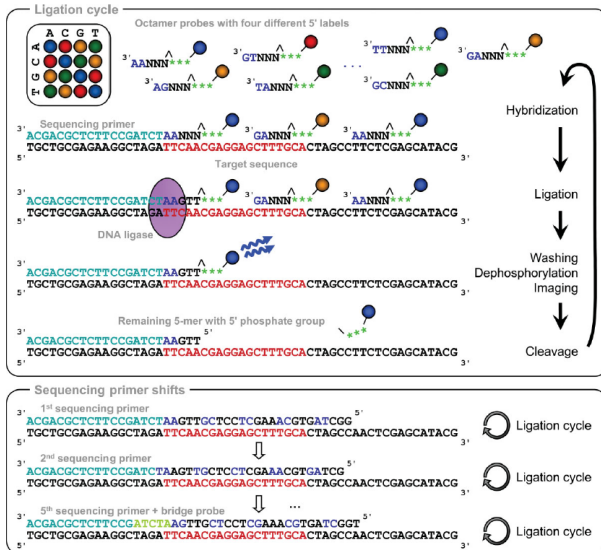
- No Finished Assemblies
 - >100 Euchromatic Gaps in Human Genome
 - *Drosophila melanogaster* release 6.1 (More Centric Heterochromatic Sequences Added)
- Contig L50 Close to or Bigger than Average Gene Size
- Number of Scaffolds Should be Close to Number of Chromosomes

- Downstream Processing
 - Repeat Annotation
 - Gene Annotation
 - Mapping to get Diploid Genome

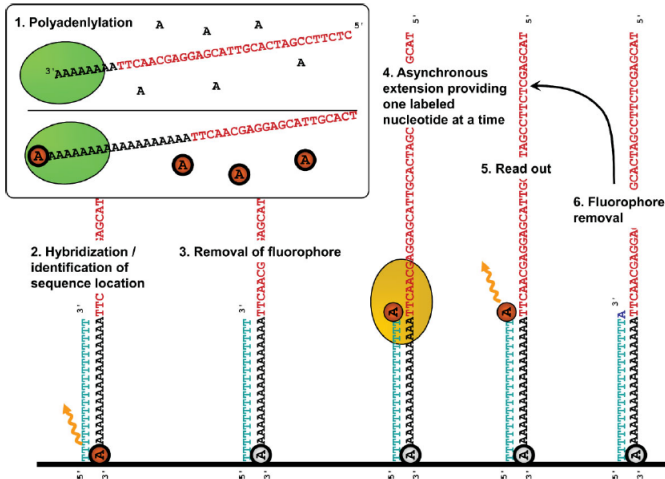
Roche 454



ABI SOLiD

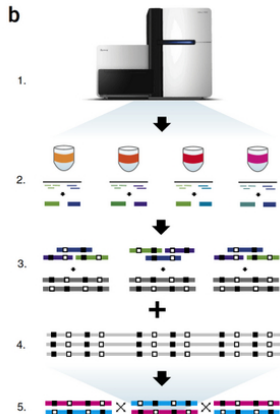
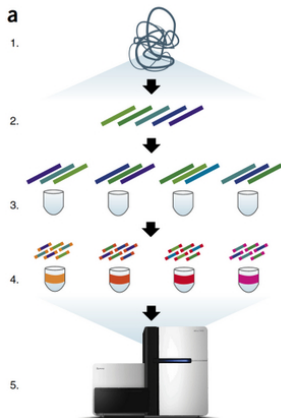


Helicos

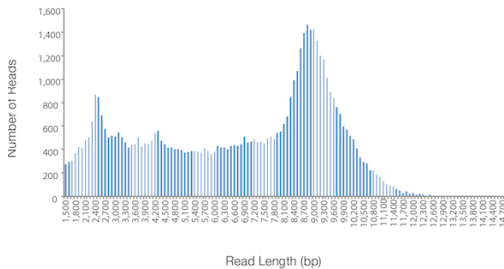


Kircher and Kelso 2010

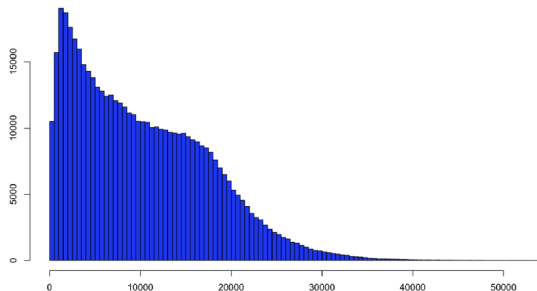
TruSeq Synthetic Long-Read Libraries (Moleculo)

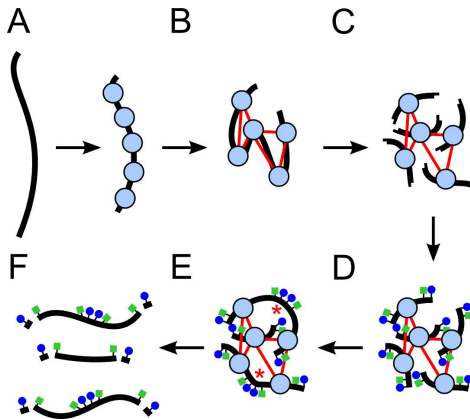


Molecule



PacBio





Putnam et al. 2015 (ArXiv)