A Roadmap to *De-Novo* Genome Assembly of Higher Eukaryotes.

Stefan Prost [1,2]

[1]Department of Biology, Stanford University, Palo Alto, United States of America

[2]Department of Integrative Biology, University of California, Berkeley, United States of America

Sept. 25th-26th, 2017

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

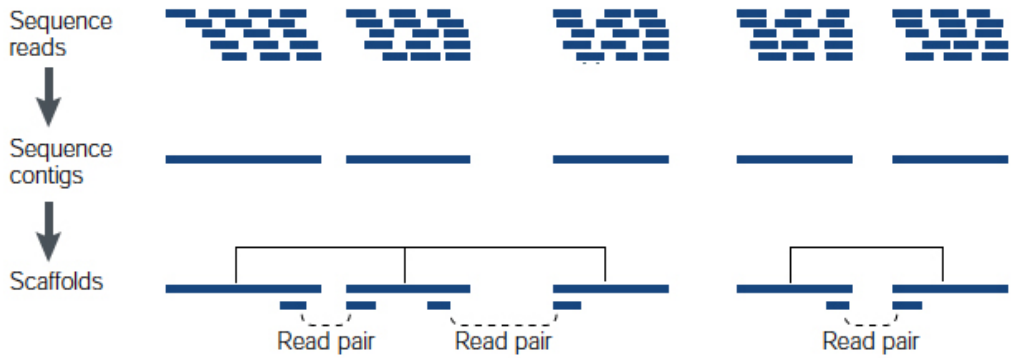
7 Further Improvement of the Assembly - Computational Methods

8 Further Improvement of the Assembly - Laboratory Methods

9 Mind the Gap! Or not??

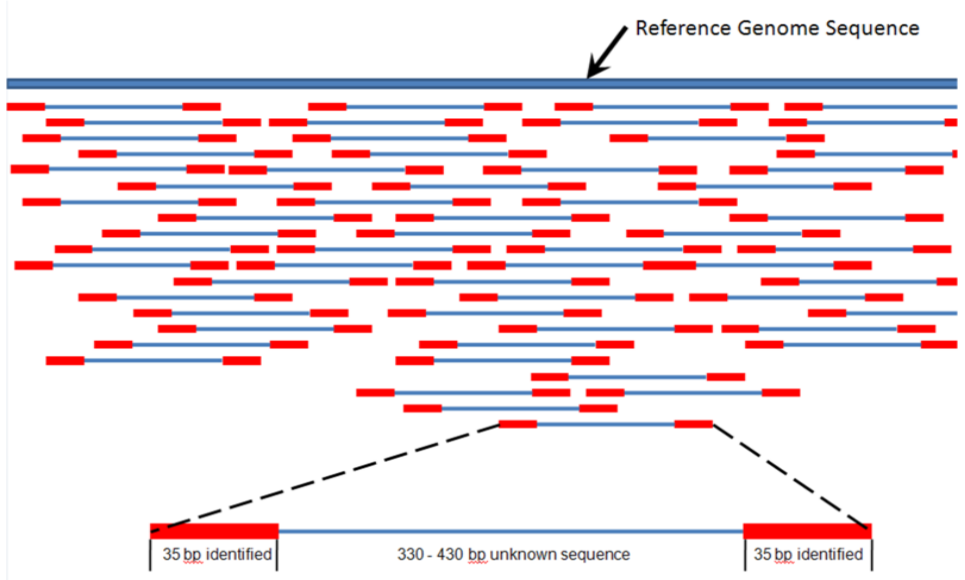
10 Downstream Processing

De-novo Assembly



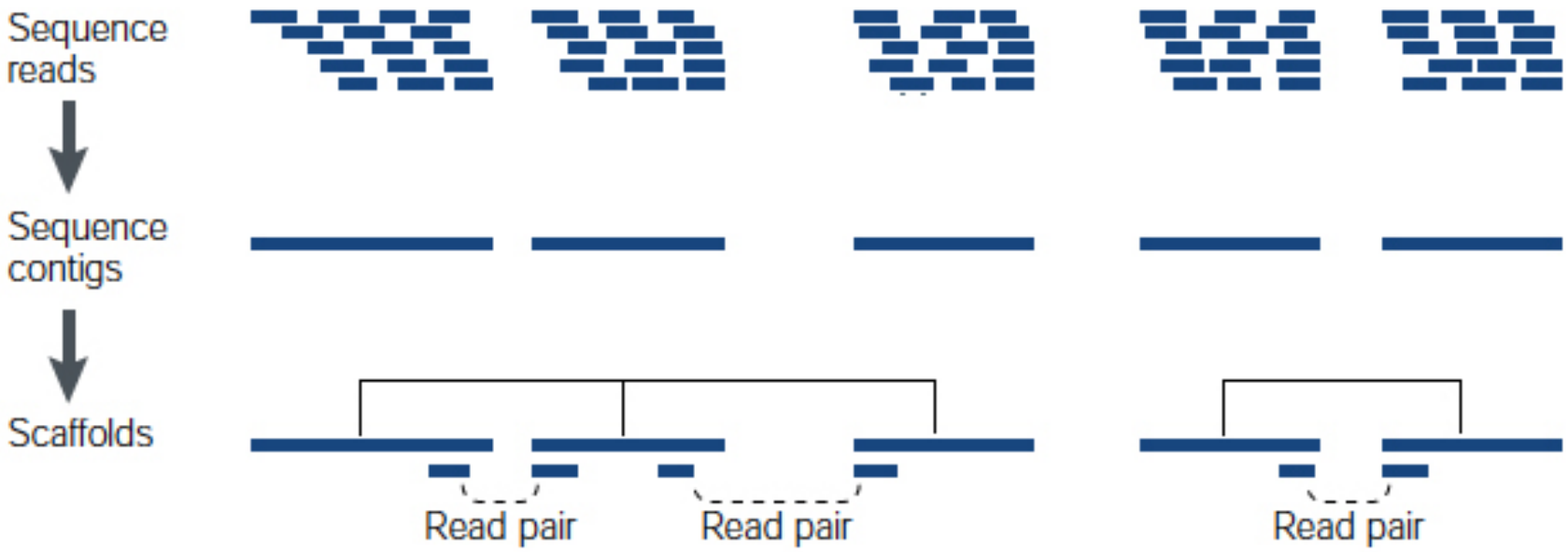
Changed after Green 2001

Reference-based Mapping



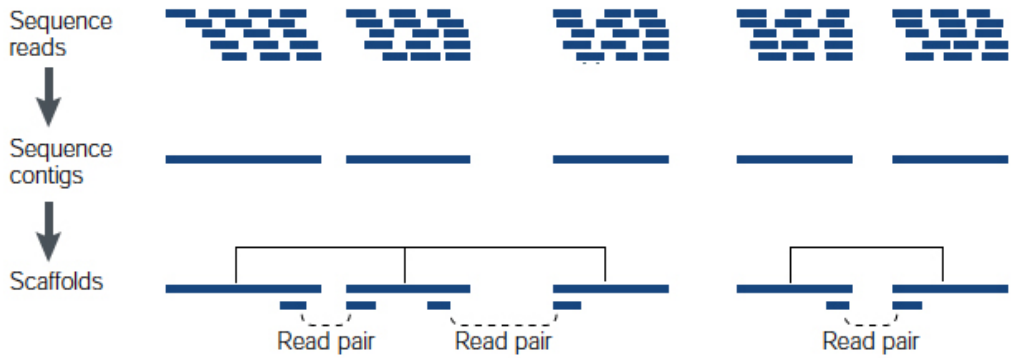
Wikipedia

De-novo Assembly



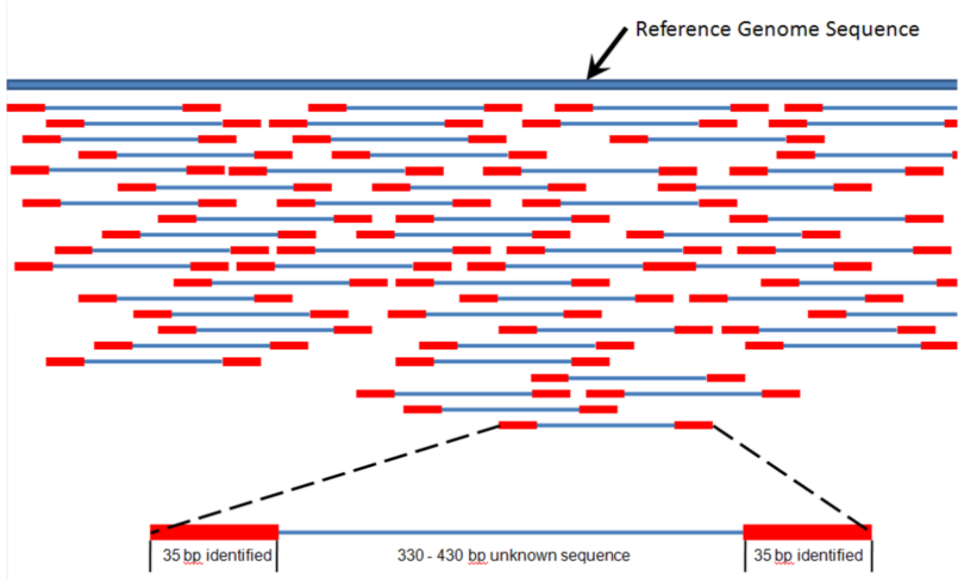
Changed after Green 2001

De-novo Assembly

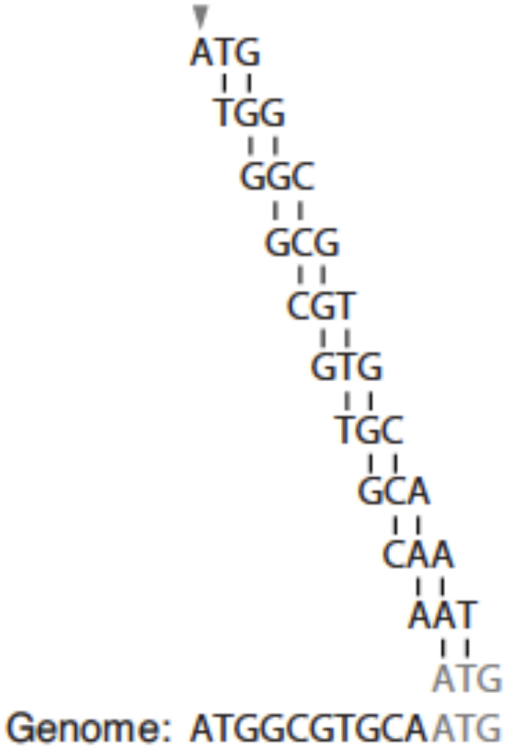


Changed after Green 2001

Reference-based Mapping



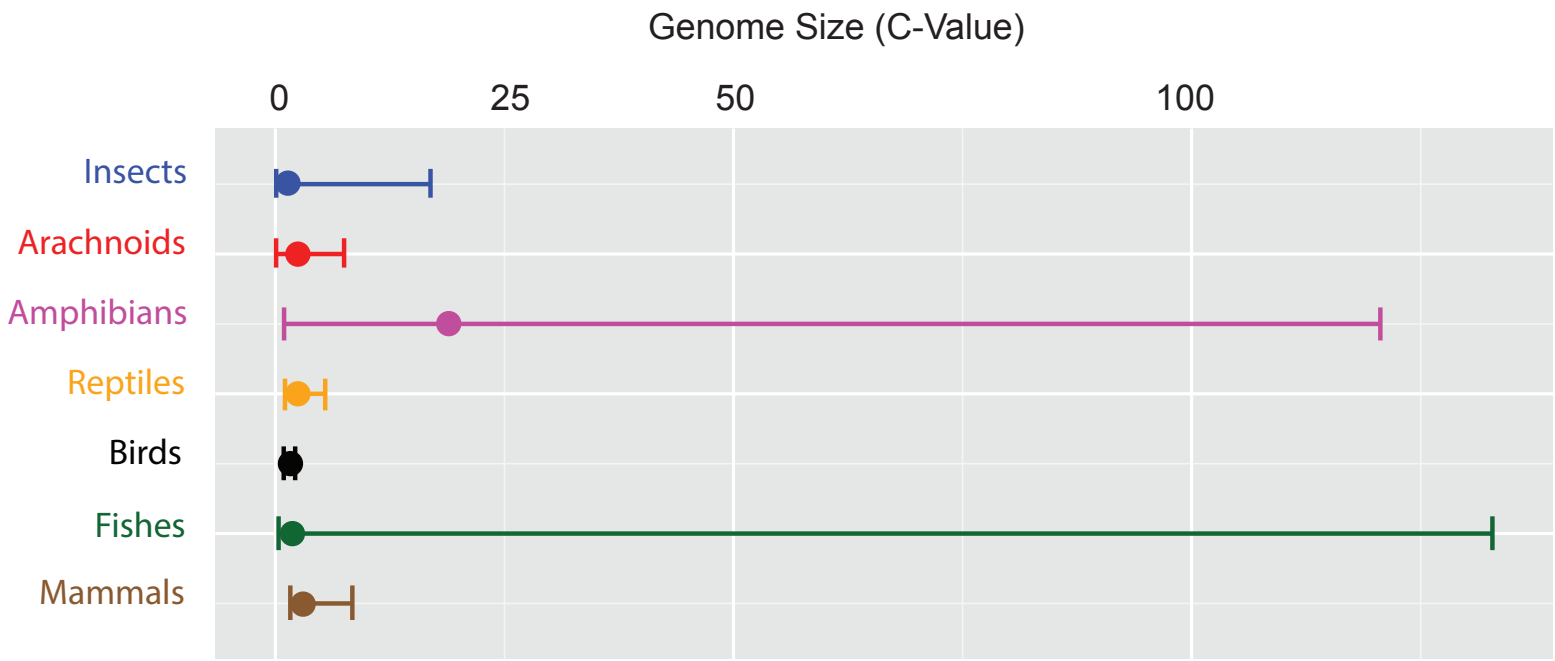
Wikipedia



Definition: Kmer

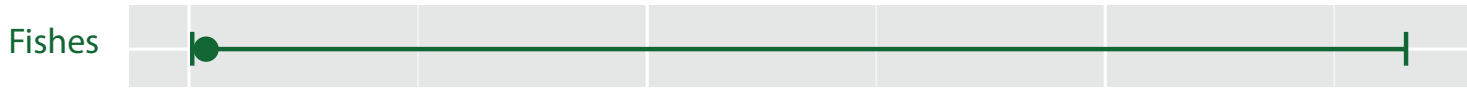
Short, unique element of DNA sequence of length n

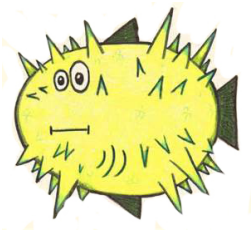
- Expected Genome Size
- Expected Repeat Content
- Expected Heterozygosity
- Diploid or Polyploid?
- DNA Amount per Individual



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

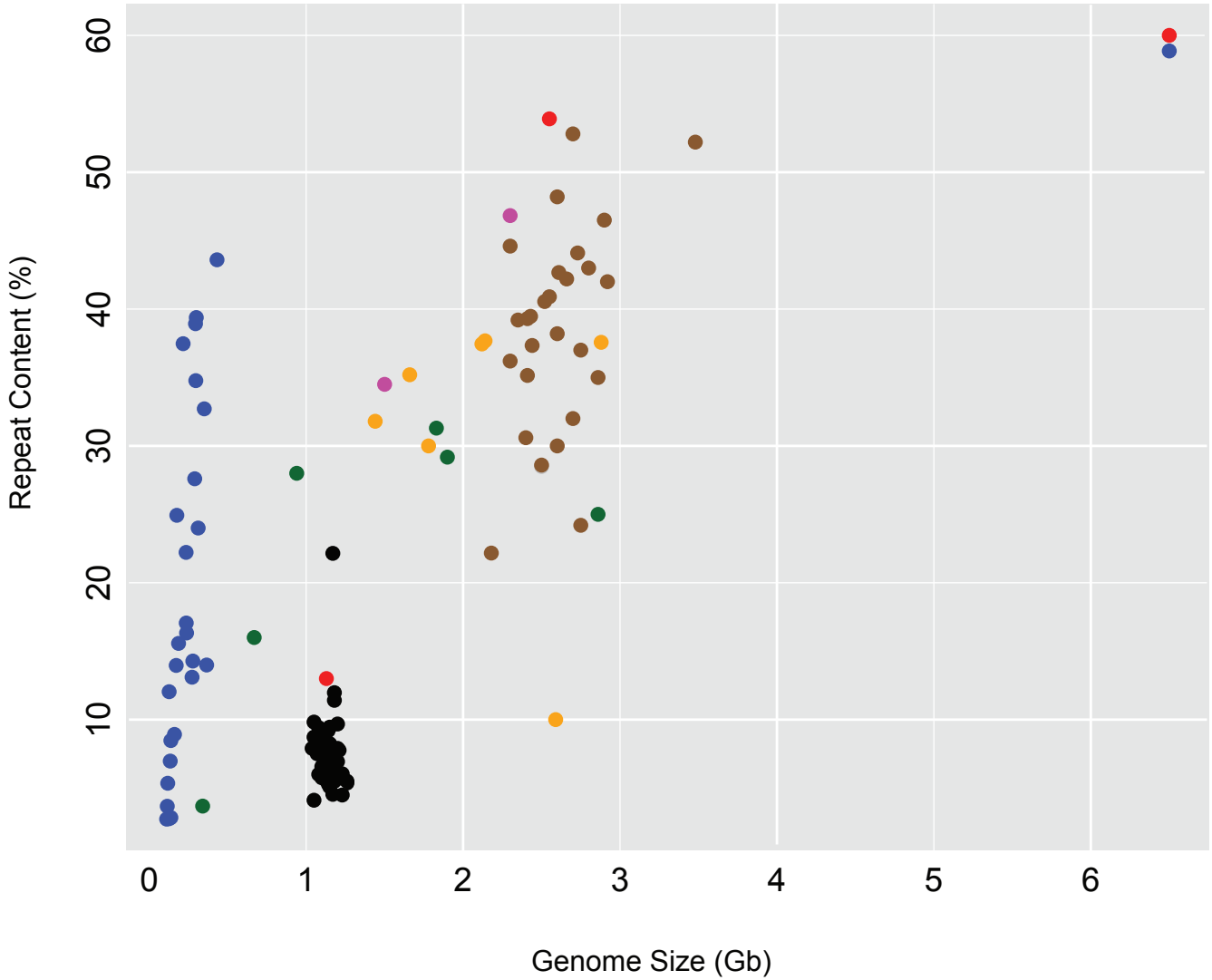
Genome Size Range in Fishes



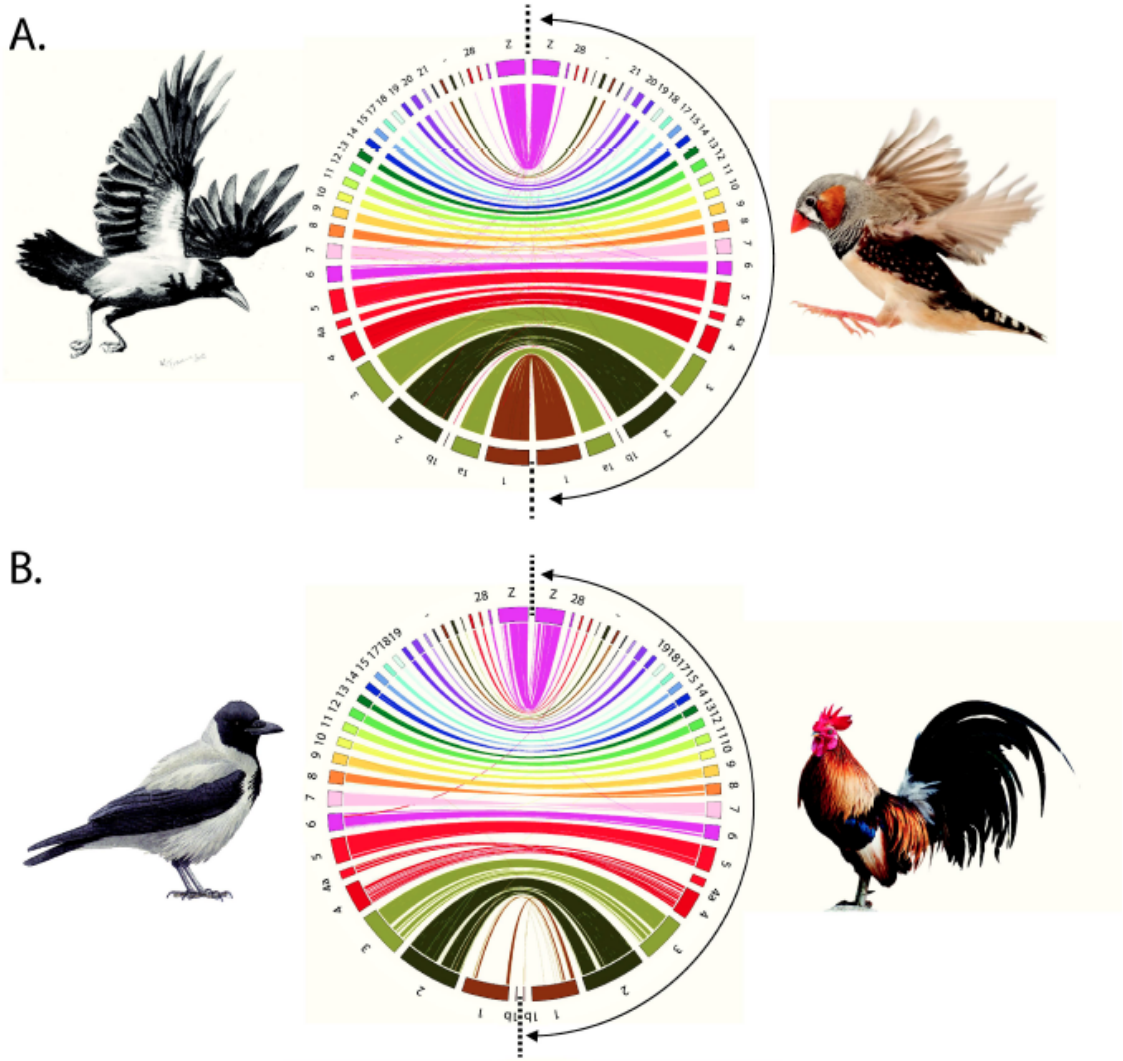


- Helpful Online Databases:
  - Genomes
    - www.ncbi.nlm.nih.gov/genome/browse
    - www.gigadb.org
    - https://phytozome.jgi.doe.gov/pz/portal.html
  - Genome Sizes (C-values)
    - www.genomesize.com

- Expected Genome Size
- Expected Repeat Content
- Expected Heterozygosity
- Diploid or Polyploid?
- DNA Amount per Individual


Repeat Content (%)
10
20
30
40
50
60
0
1
2
3
4
5
6
Genome Size (Gb)

- Expected Genome Size
- Expected Repeat Content
- Expected Heterozygosity
- Diploid or Polyploid?
- DNA Amount per Individual



Poelstra et al. 2014

- Expected Genome Size
- Expected Repeat Content
- Expected Heterozygosity
- Diploid or Polyploid?
- DNA Amount per Individual

1st Generation

ABI (Sanger)

2nd Generation

Roche 454

Illumina HiSeq

Life Technologies IonTorrent

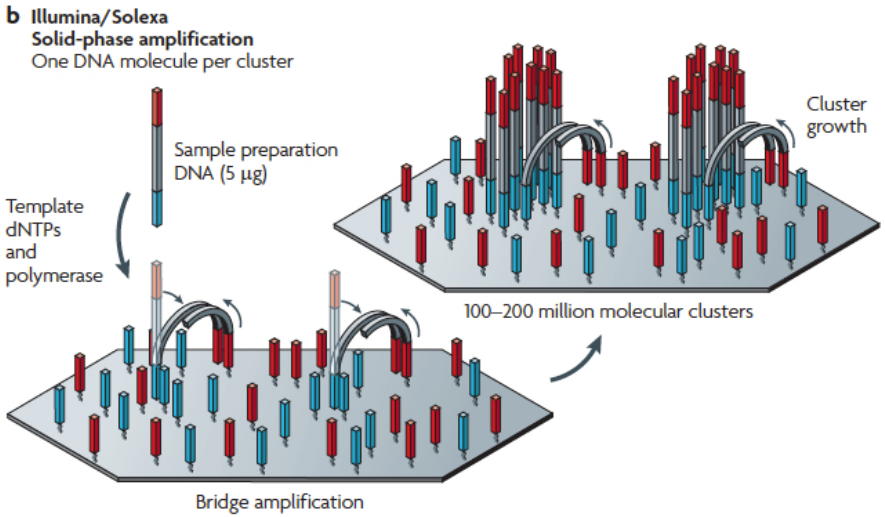
3rd Generation

Oxford Nanopore MinIon

Pacific Biosciences RSII

- First Generation Sequencing
  - Sanger Sequencing
- Second Generation Sequencing (*PCR Needed*)
  - Illumina: MiSeq & HiSeq
  - Roche: 454
  - Life Science: IONtorrent & IONproton
  - ABI: SOLiD
  - BGIseq-500
  - Qiagen GeneReader
- Third Generation Sequencing (*Single Molecule Sequencing*)
  - Helicos Biosciences: Heliscope
  - Pacific Biosciences: PacBio RS II
  - Oxford Nanopore: MinION & GridION

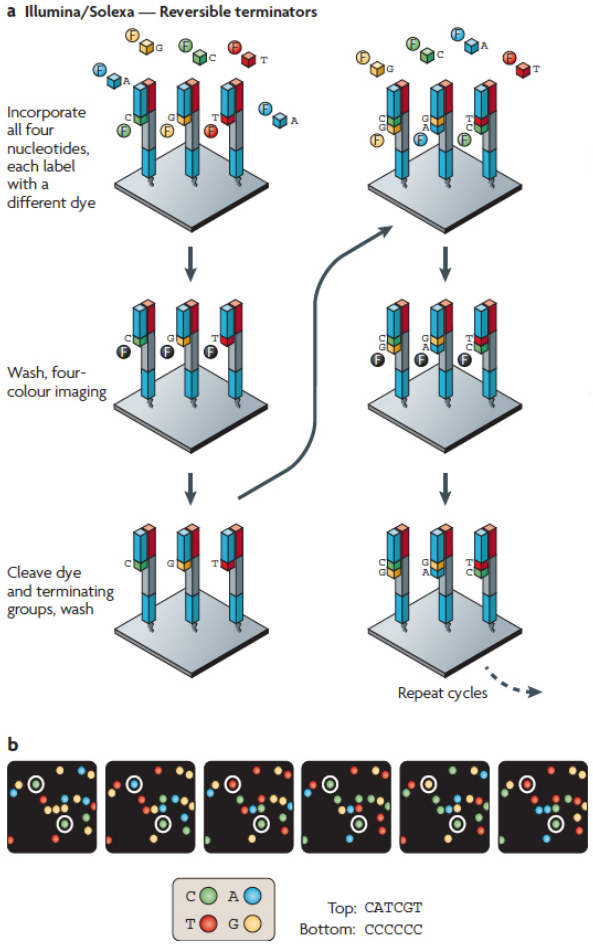
Illumina Sequencing



Metzker 2010

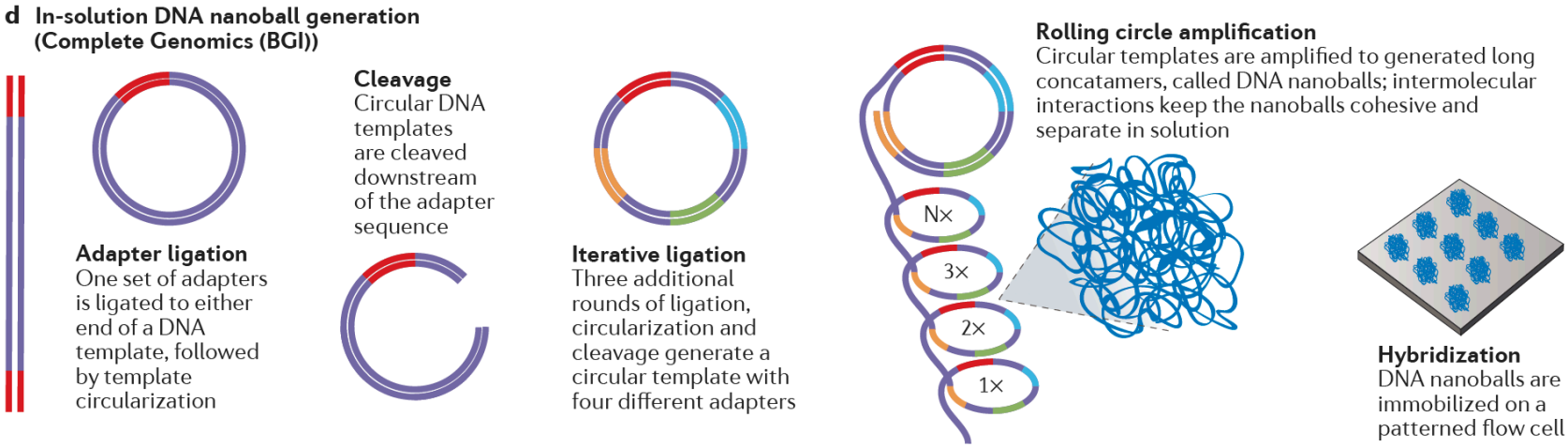
Basic amplification technology:
https://www.youtube.com/watch?v=HMyCqWhwB8E

Exclusion PCR Cluster Formation:
https://youtu.be/pfZp5Vgsbw0

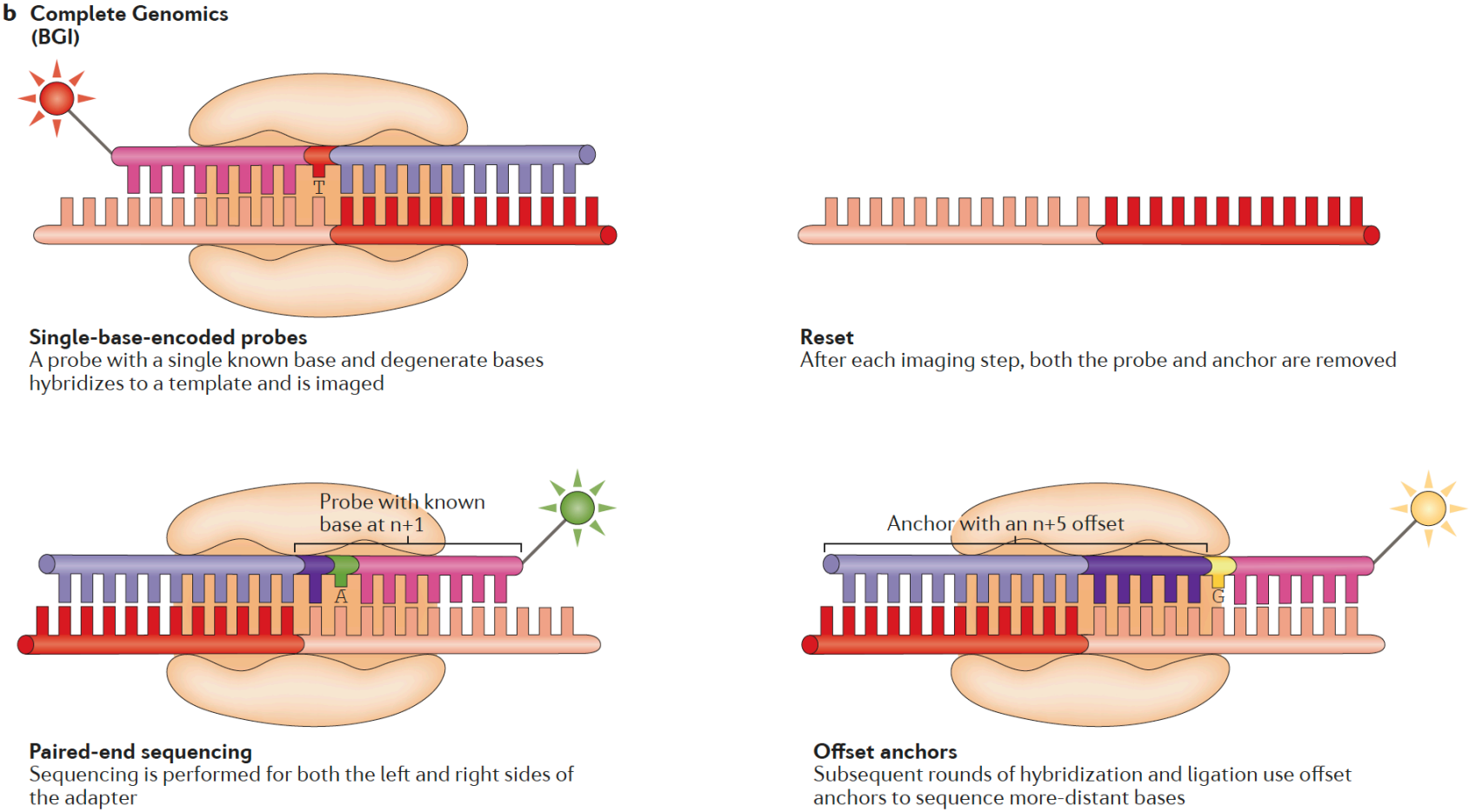


Metzker 2010

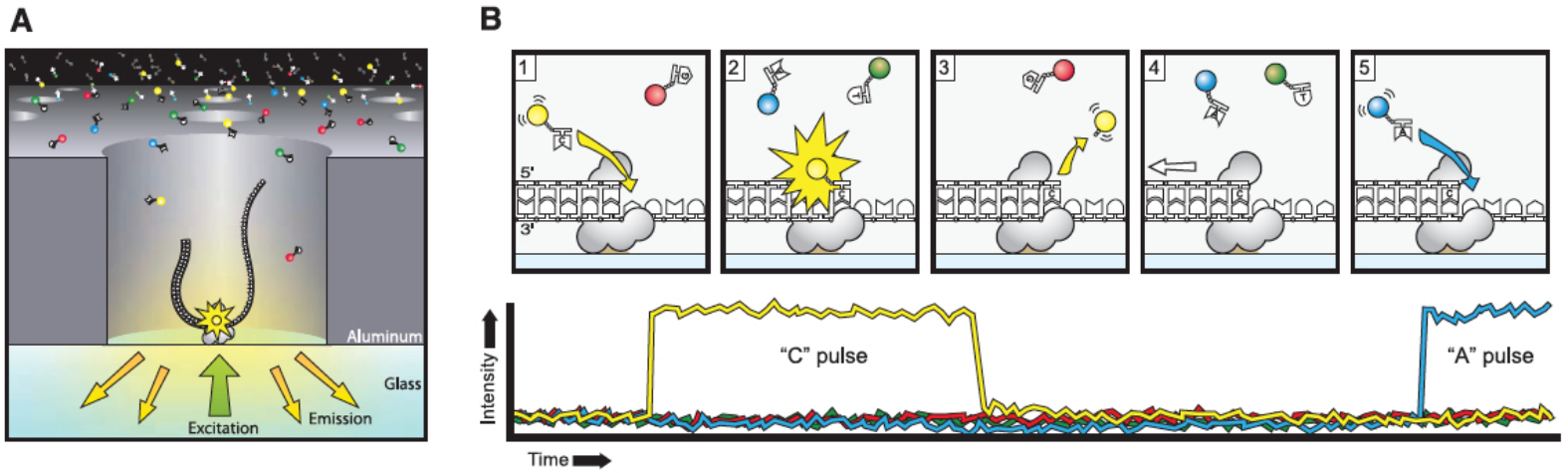
NovaSeq 5000/6000



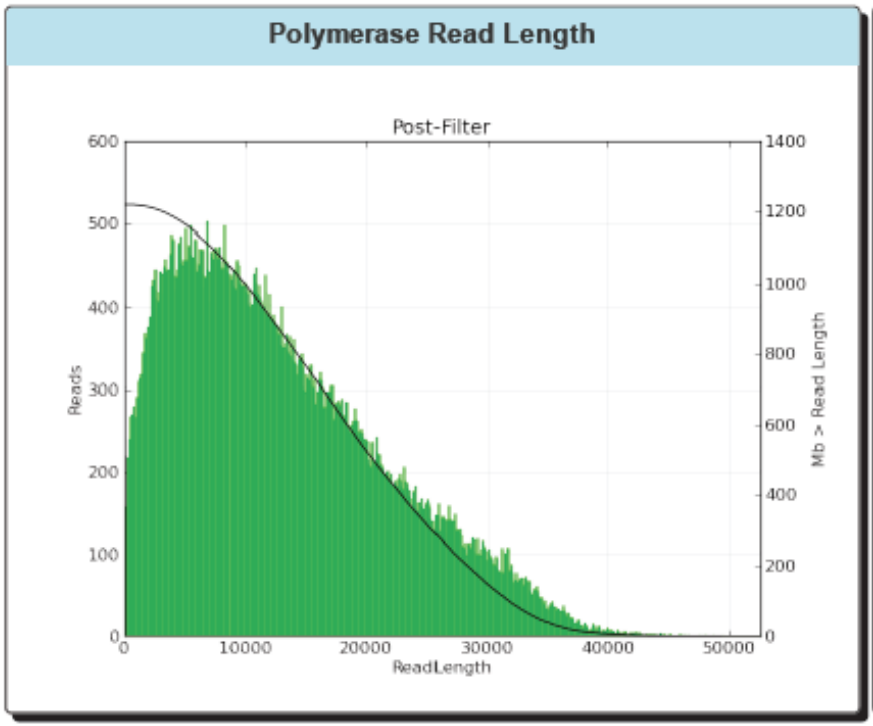
BGIseq-500 (Goodwin, McPherson and McCombie et al. 2016)

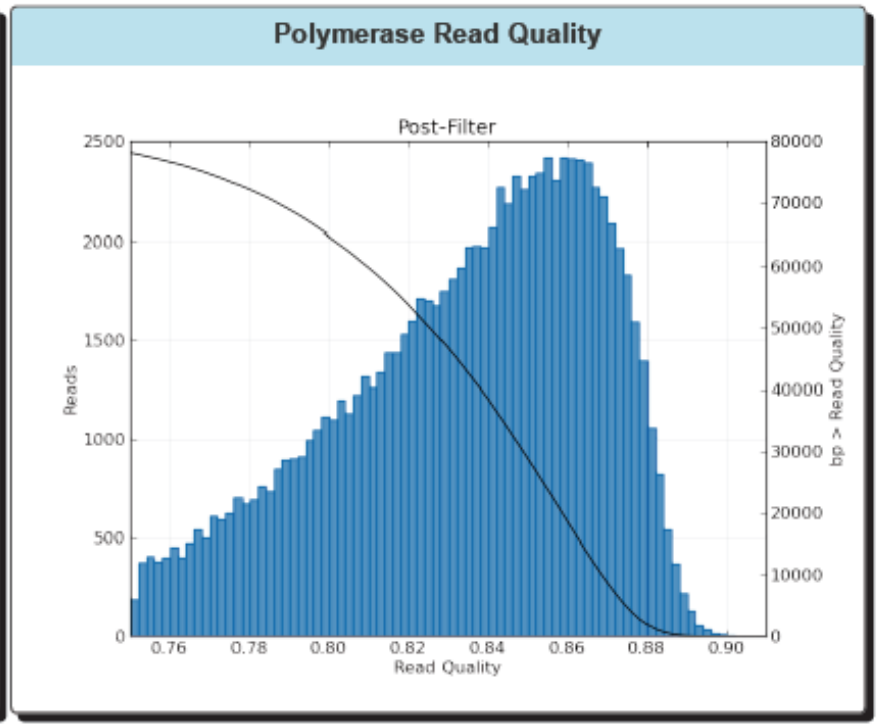


BGIseq-500 (Goodwin, McPherson and McCombie et al. 2016)



Eid et al. 2009





Subread Filtering

Mean Subread length	10,660	N50	16,247
Total Number of Bases	1,222,648,053	Number of Reads	114,688

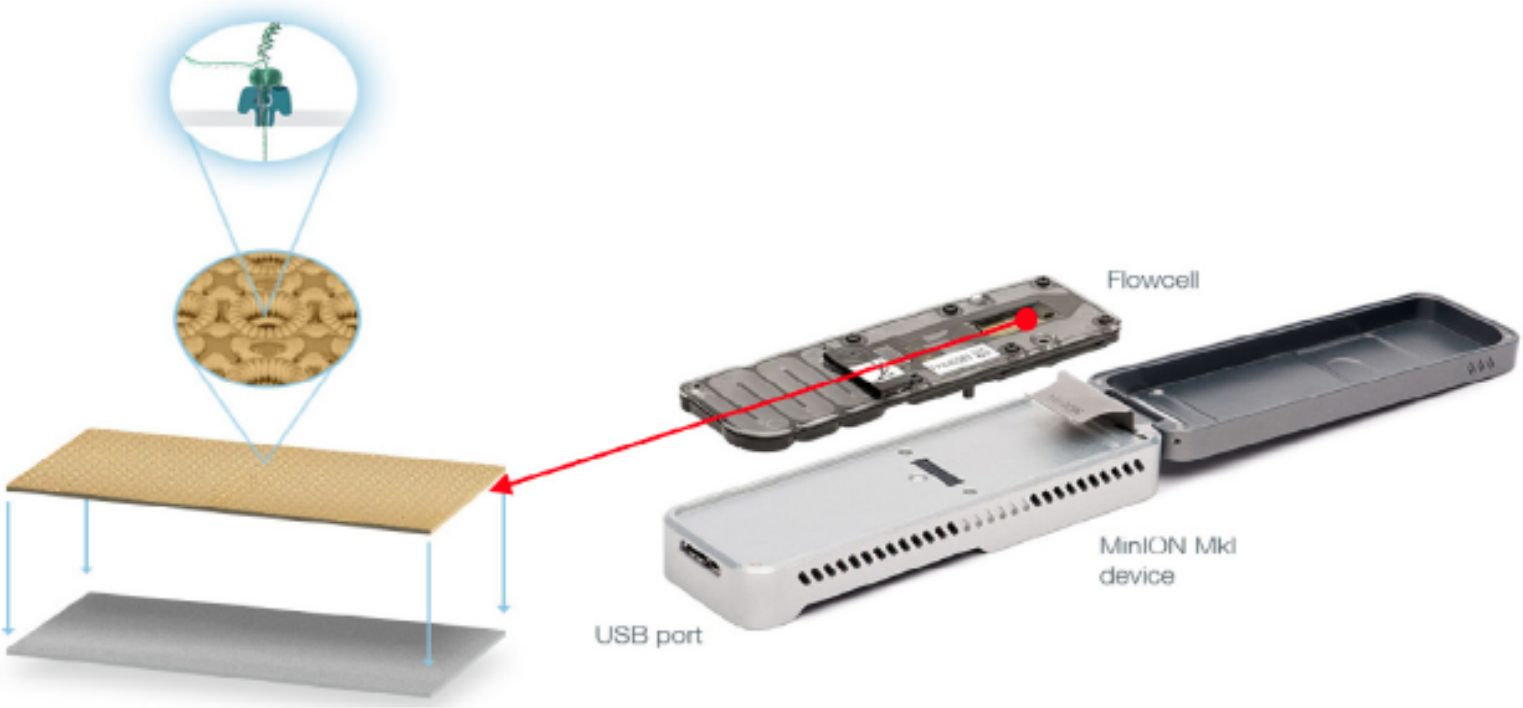
Choanoflagellate (Salpingoeca rosetta) Genome Assembly:

	n	**N50**	**max**	**length**
original	154	1.4Mb	2.6Mb	52Mb
scaffold	125	1.8Mb	4.2Mb	52Mb
pacbio 1	948	147kb	1.4Mb	59Mb (sprai, 1 cell)
pacbio 2	458	640kb	3.3Mb	60Mb (sprai, 5 cells)
scafpac2	122	1.3Mb	3.7Mb	51Mb (merged)
pac2scaf	67	1.6Mb	6.2Mb	57Mb (merged)

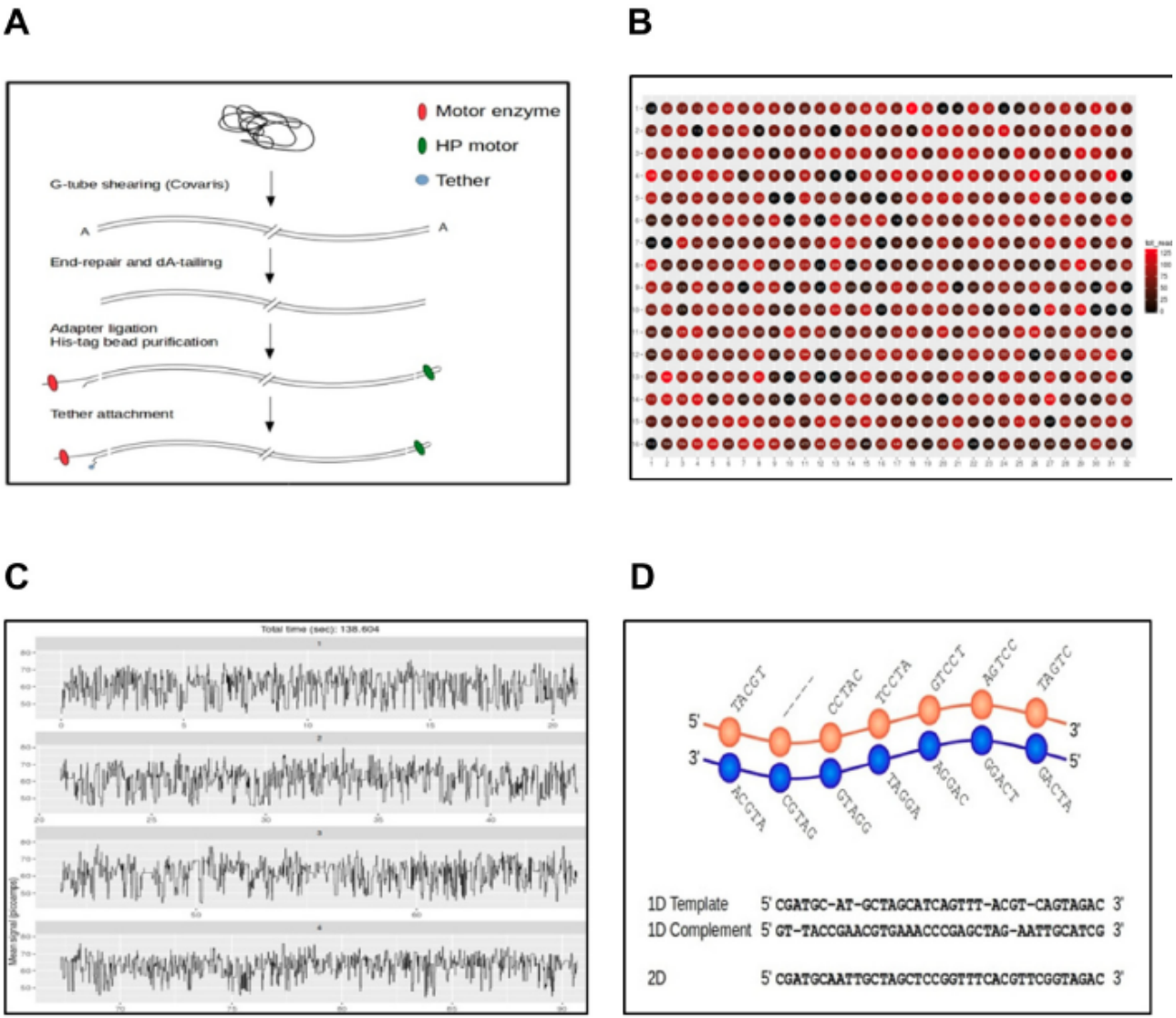
In collaboration with Laura Wetzel and Nicole King (UCB).

Suggested reading: Lu, Giaordano and Ning, 2016 (Genomics Proteomics Bioinformatics)

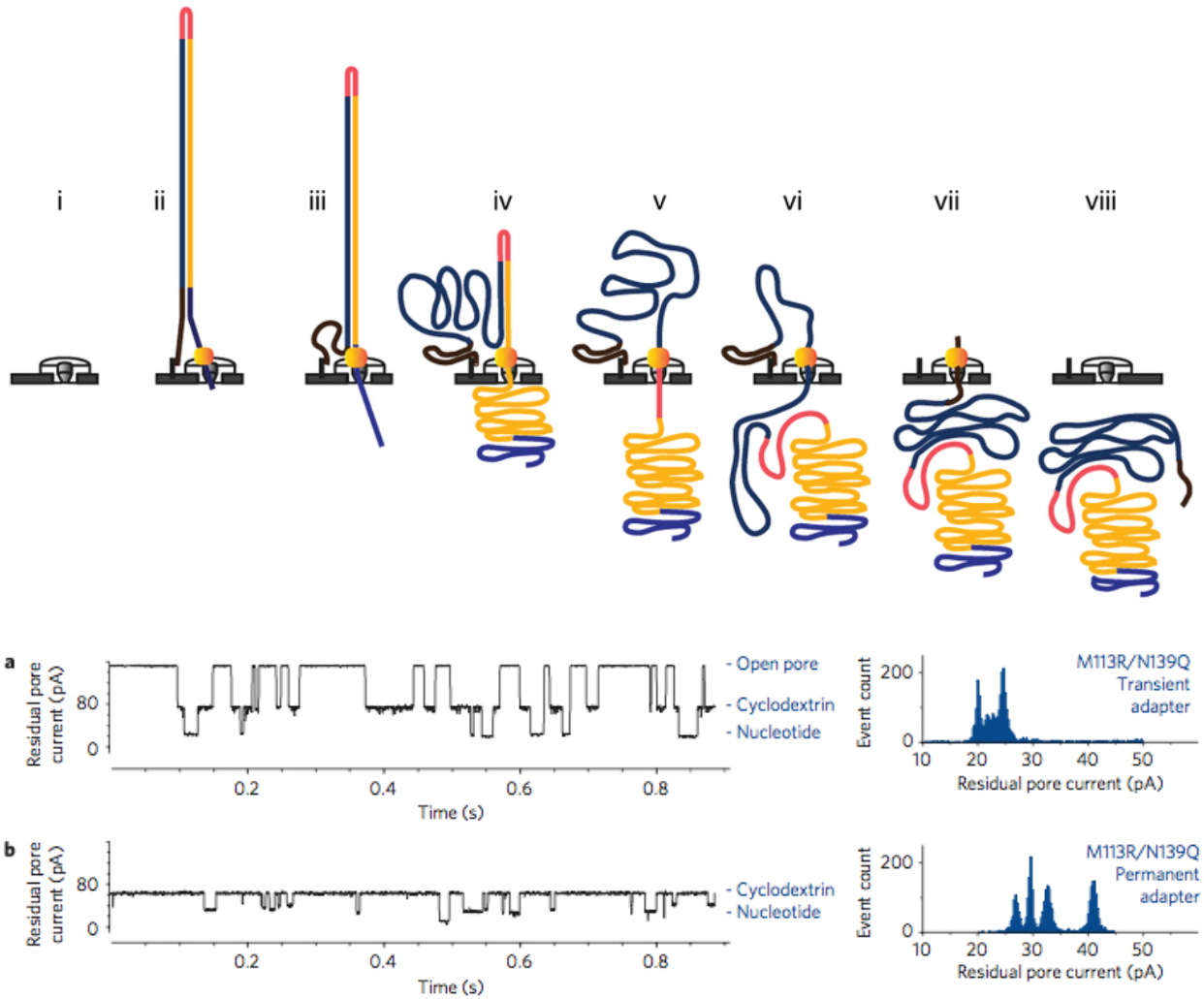
SmidgIon



Lu, Giaordano and Ning, 2016



Lu, Giaordano and Ning, 2016

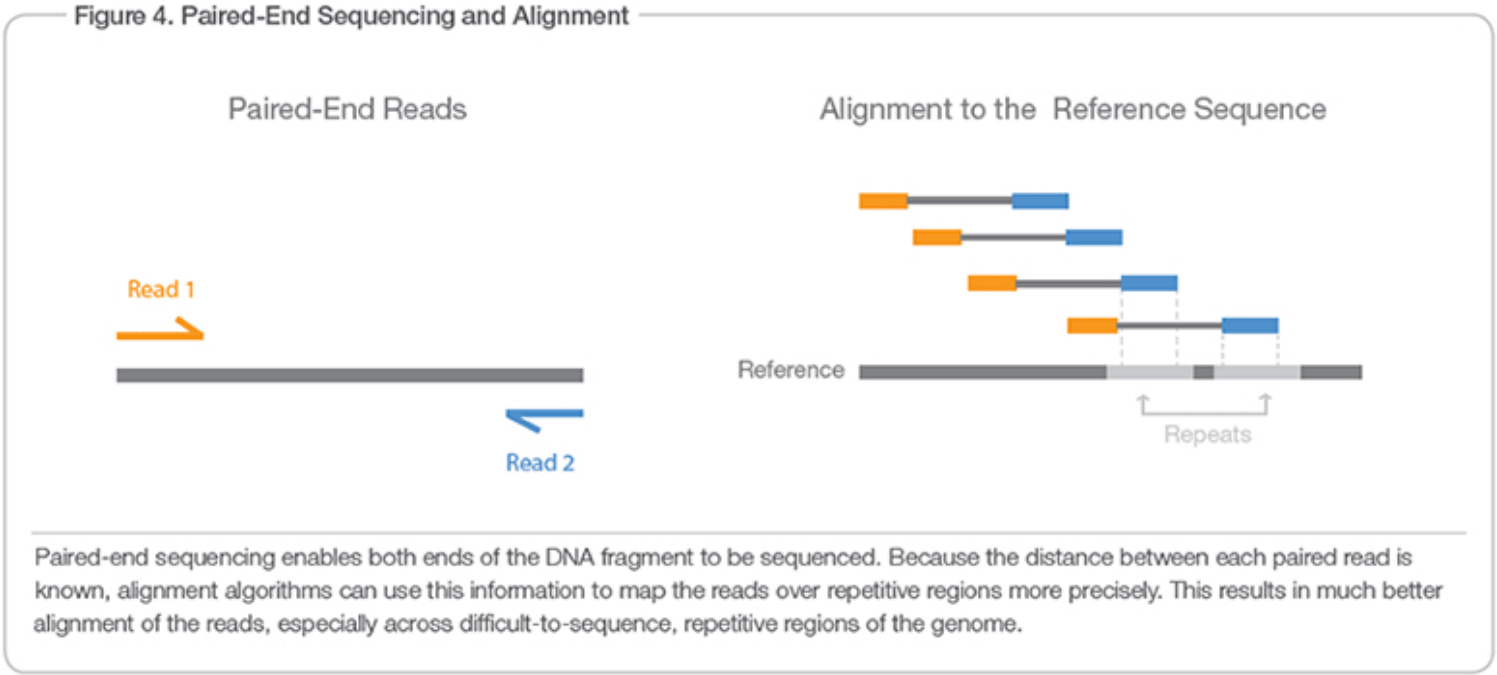


Jain et al. 2014; Clarke et al. 2009

Illumina Sequencing Libraries

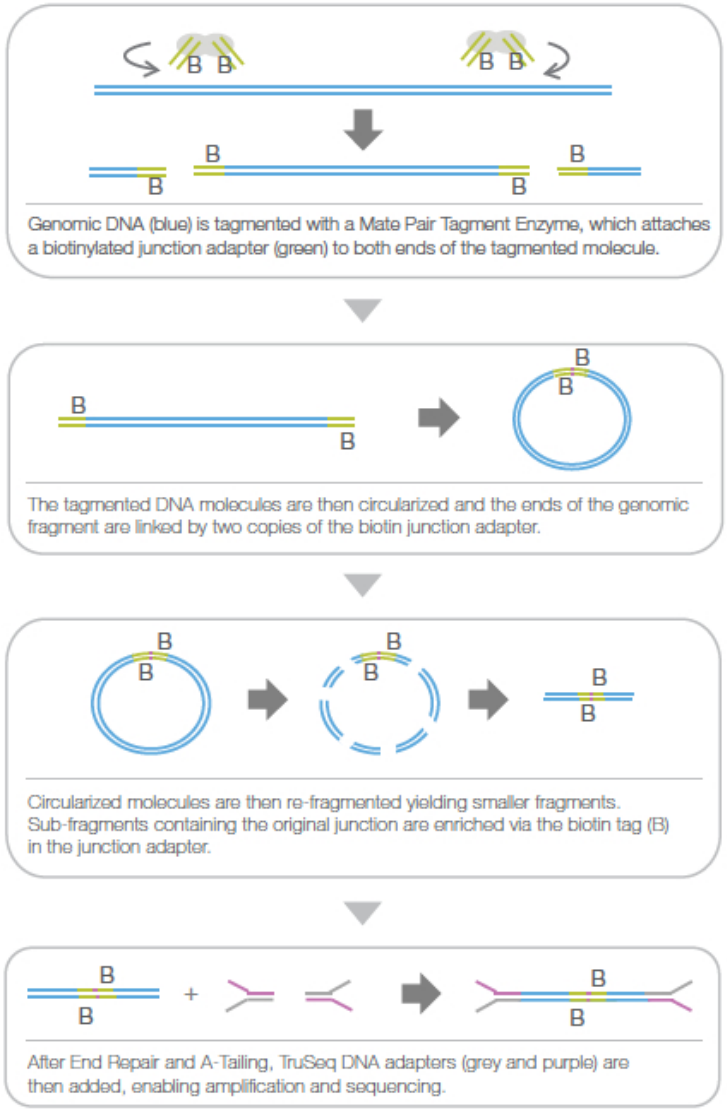
- Paired-end Libraries
- Circularization-based Libraries
  - Mate-pair
  - Fosmid/BAC
- Dilution-based Libraries
  - 10X Genomics
- 3D Structure-based Libraries
  - Dovetail Genomics Chicago
  - Hi-C

Illumina Paired-end Sequencing Libraries



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**

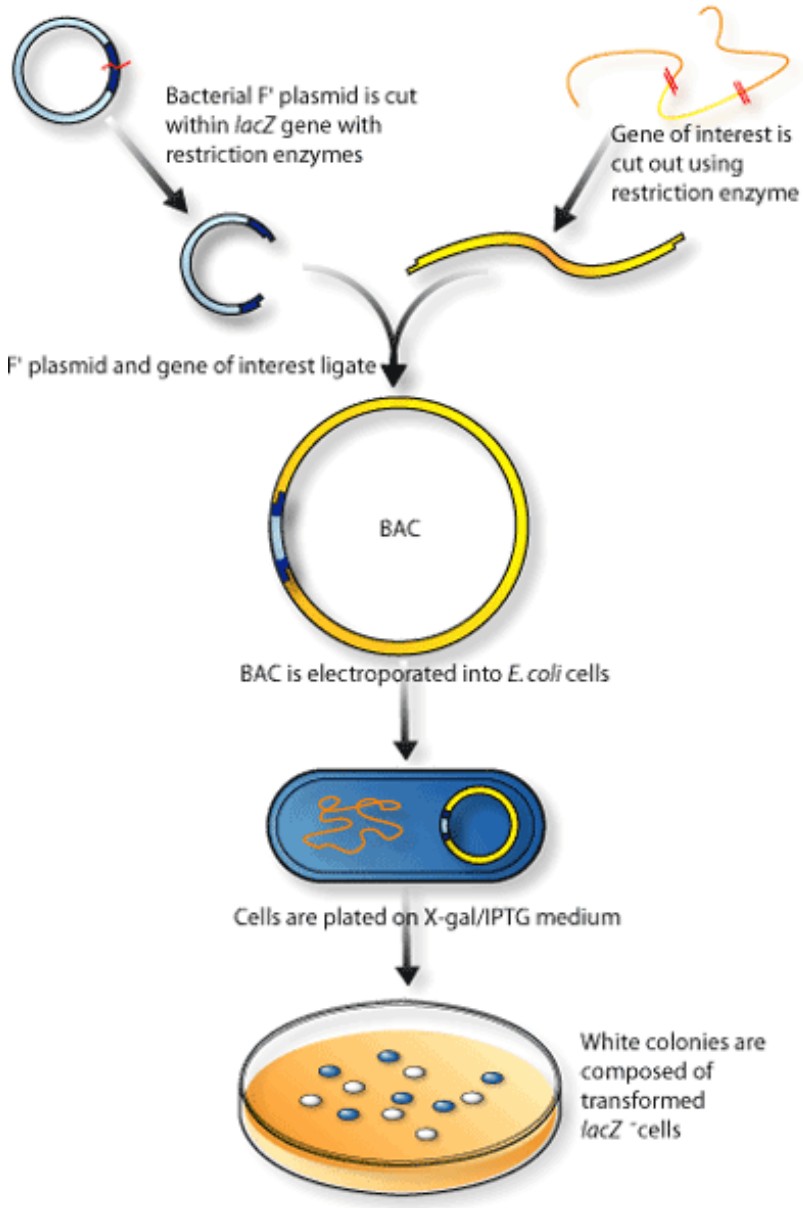
Custom Recipe for 1.3Gb Bird Genome (<15% Repeats)

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

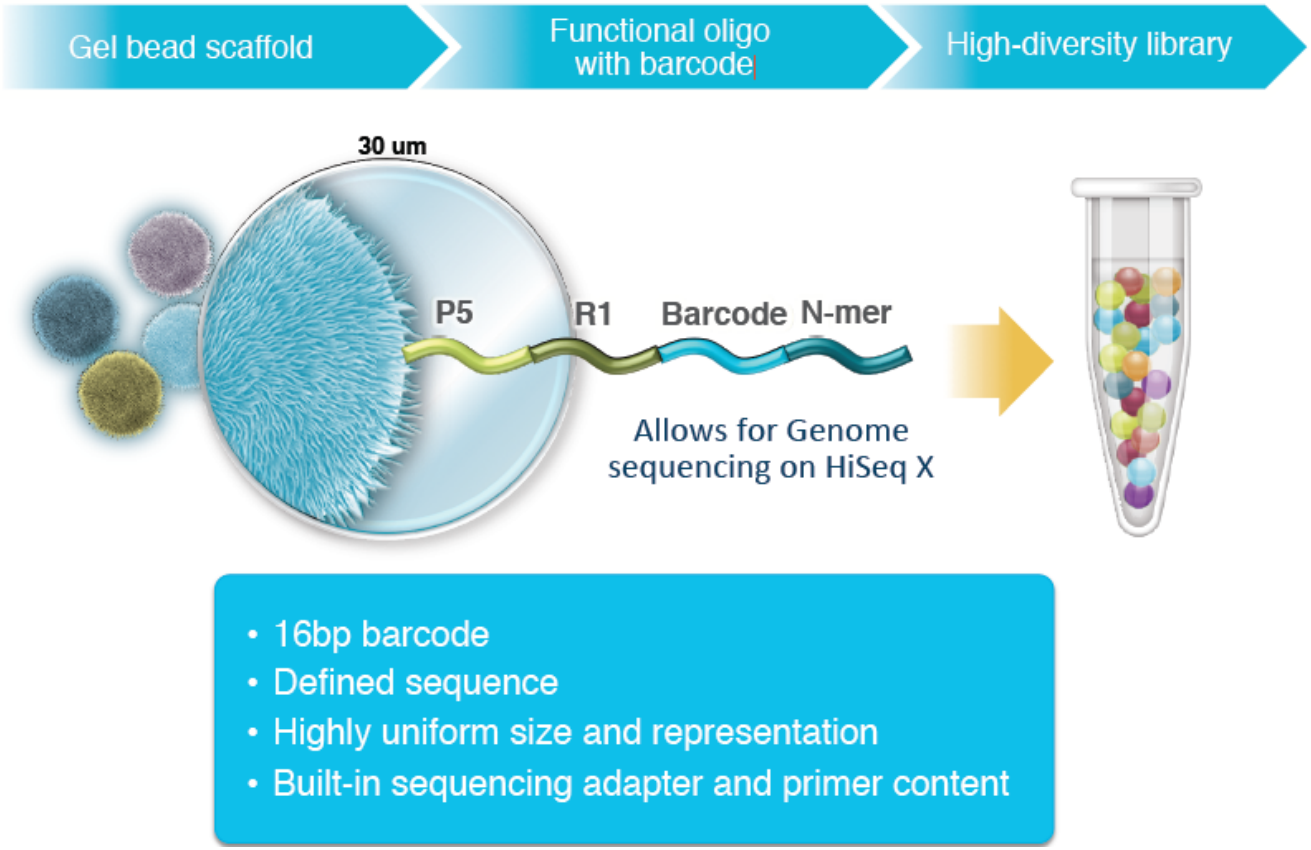
Custom Recipe for 2.5Gb Mammal Genome (<50% Repeats)

- 180bp PE - 4 lanes HiSeq (50x coverage)
- 3kb MP - 4 lanes HiSeq (40x coverage)
- 8kb/20kb MP - 1 lane (10x coverage)

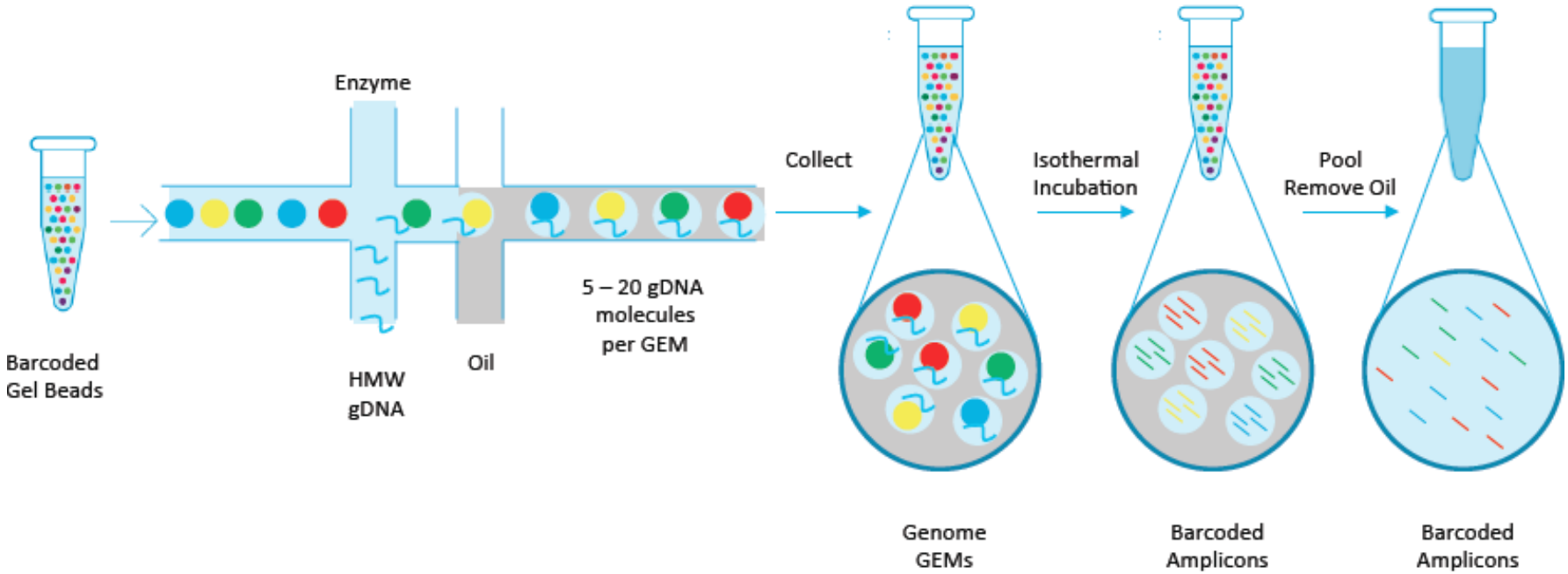
- Fosmid Vector
  - Bacterial F-plasmid
  - <40kb Insert Size
- Bacterial Artificial Chromosome (BAC)
  - <300kb Insert size



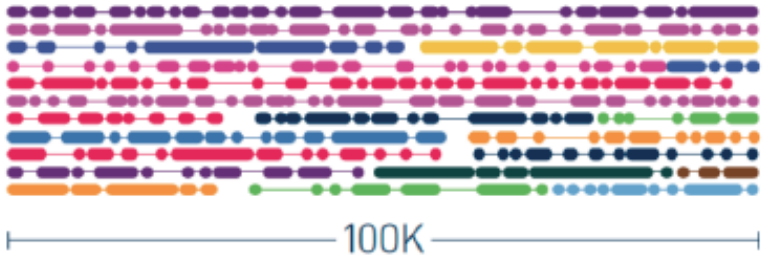
4,000,000 Barcodes



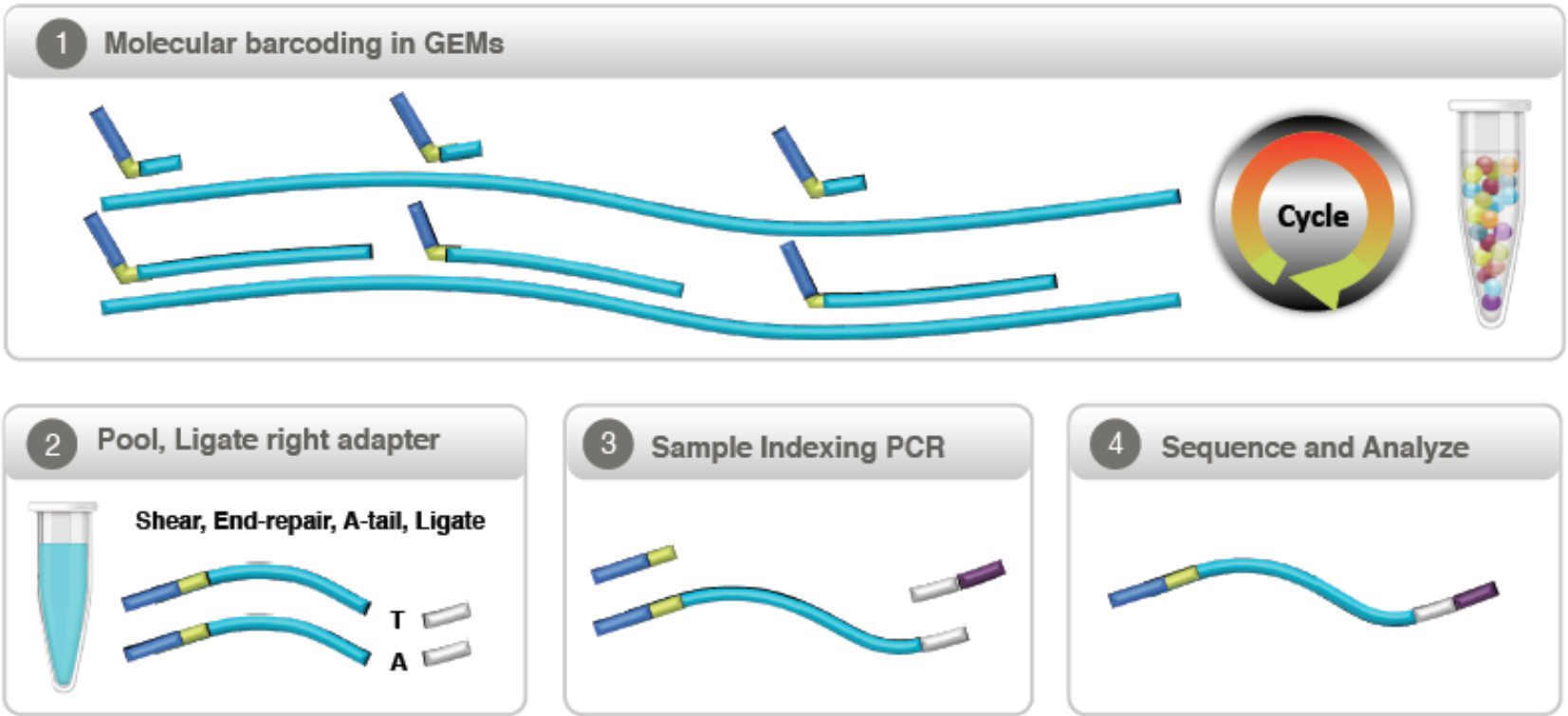
- 16bp barcode
- Defined sequence
- Highly uniform size and representation
- Built-in sequencing adapter and primer content

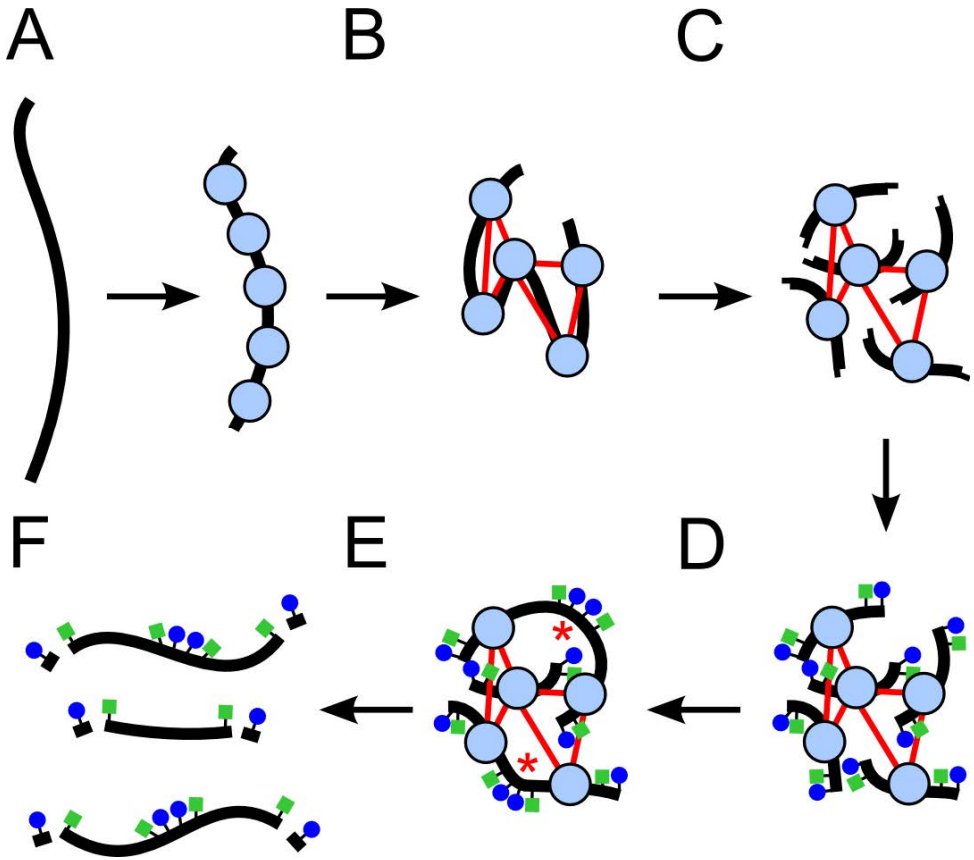


Linked-Reads



10X linked Long-Read Libraries





Putnam et al. 2015 (ArXiv)

Tuatara Assembly Statistics

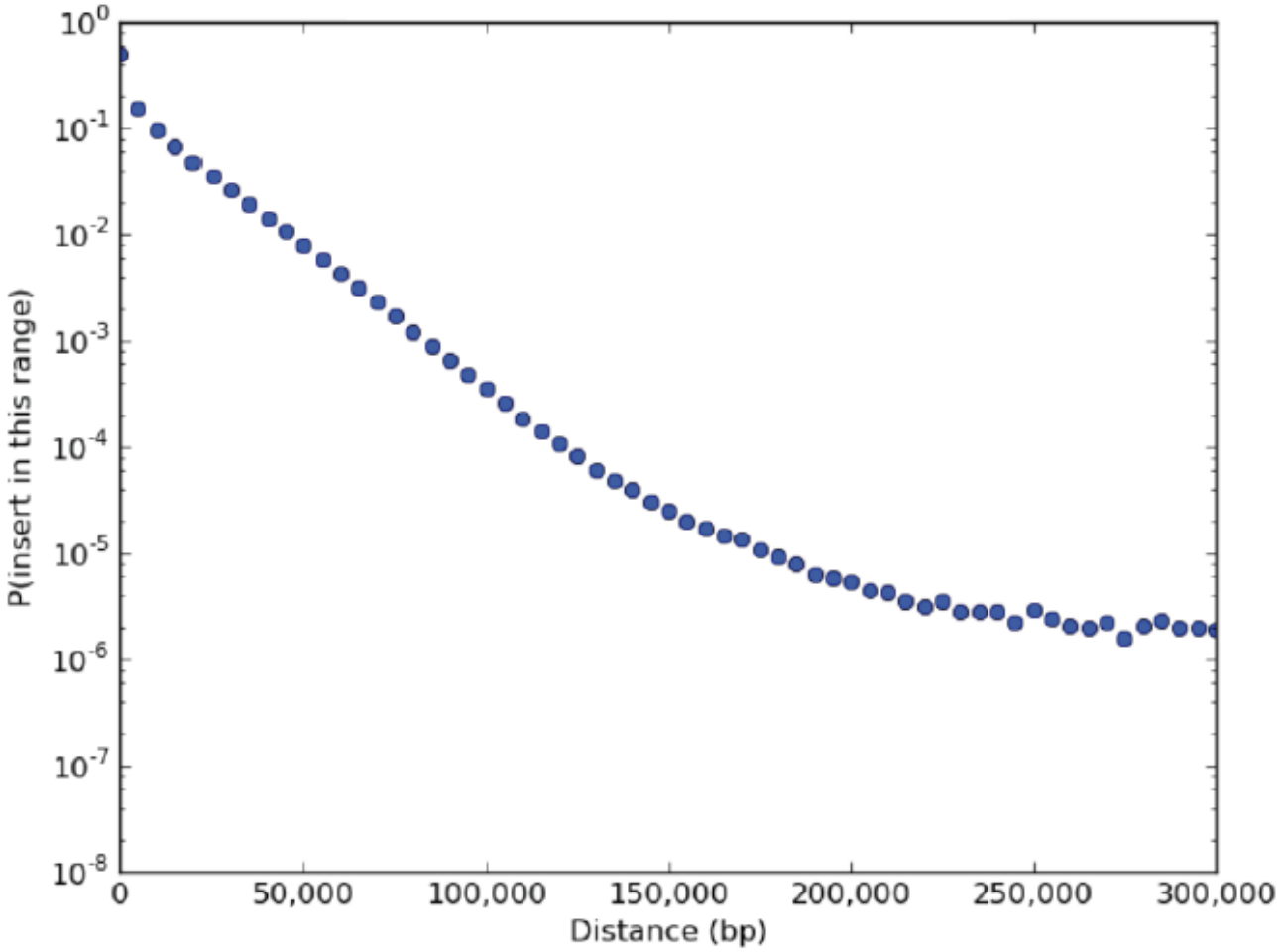
	Starting Assembly	Final Assembly	Fold Increase
Genome Size	4.2691 Gb	4.2718 Gb	-
N50	348 Kb	2.23 Mb	6.4X
N90	69 Kb	438 Kb	6.3X

Estimated Dovetail coverage (1-50 Kb pairs): 43.5X

Dovetail GENOMICS LLC

	Meraculous Human NA12878 N50 (Mbp)
Fragment only (84X coverage)	0.033
Frag+Mate	0.45
Frag+Mate+Fos	9.1
Fragment+ cHiCago (1 HiSeq lane)	20.9

Dovetail Library Insert Distributions



Dovetail GENOMICS LLC

Read Quality Assessment

- Base Quality

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

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

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

Software Tools

- FastQC (Andrews 2010)
- Preqc (Simpson 2014)
- GenomeScope (Vurture et al. 2017)

Fastq Format

```
@HISEQ:119:C42B3ACXX:7:1101:2009:2249 2:N:0:CGACCTG
TCTTGGGGACAGGGAATTTCATTTCCAAATGAAAATCCTCAAAGAACGCCTTTTATTTTACAGGAGGCTGTATATCTTAGCCAAAGTGGTAGATCGGAAGA
+
BB<BFFFBFB<BFF7BBF<BF<FBB<FFF<FFBFF<BFFFFBF'BF<7BB<BFBFB<BFFFFF<FFFFFF<BBFB<BB<BBBBBBB7<B<BB<77<BBB77
```

Quality Score Encoding

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 .........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 .............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ...............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`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

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Read Quality Assessment

- Base Quality

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

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

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

Software Tools

- FastQC (Andrews 2010)
- Preqc (Simpson 2014)
- GenomeScope (Vurture et al. 2017)

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

G = Genome Size

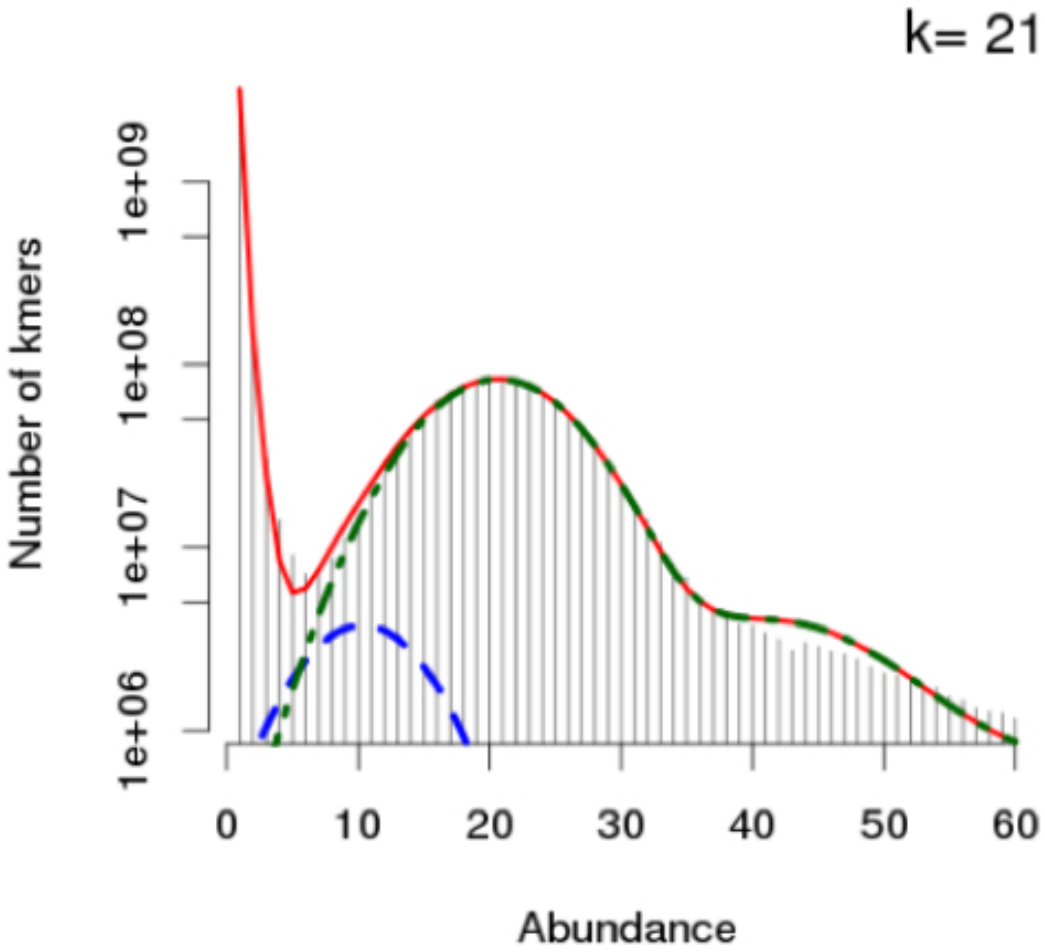
pn = proportion of correct reads

l = read length

k = kmer length

λ_k = mode of the k-mer count histogram

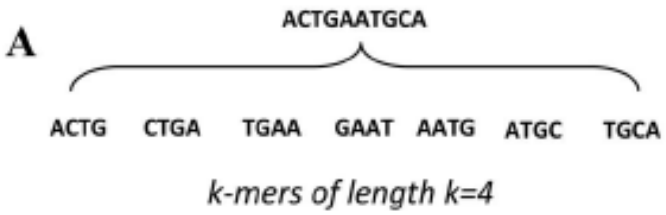
Simpson 2013, arXiv

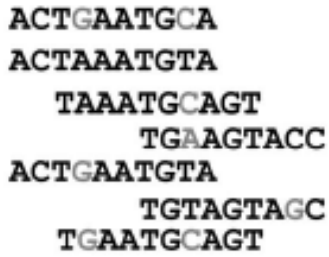

k= 21
Number of kmers
1e+06
1e+07
1e+08
1e+09
0
10
20
30
40
50
60
Abundance

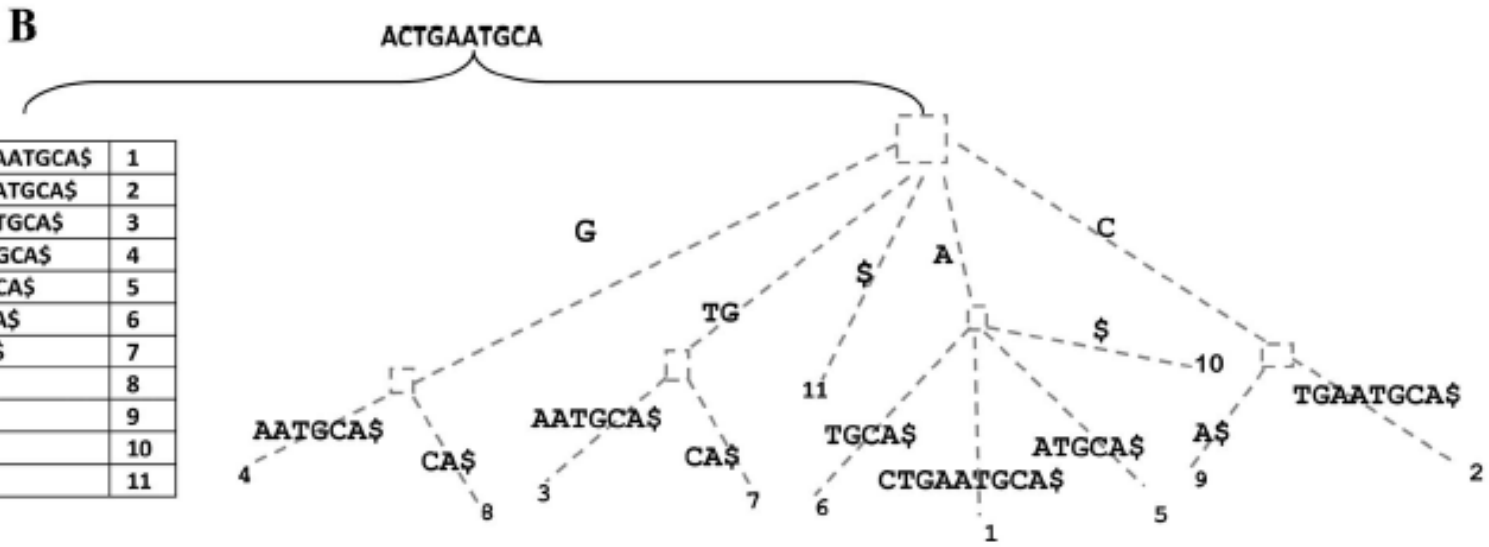
- Adapter and Low Quality Base Trimming
  - Skewer (Jiang et al. 2014)
  - AdapterRemoval (Lindgreen 2012)
  - Trimmomatic (Bolger et al. 2014)
  - Cutadapt (Martin 2011)
- Contamination Filtering
  - Kraken
  - Blast
  - Allpaths-LG
    - Removing Low Frequency k-mers
    - Discarding Scaffolds Shorter than 1kb

- Error Correction (EC) using the k-mer spectrum
  - BLESS
  - SGA
  - CUDA
  - DecGPU
  - Euler
  - Musket
  - Quake
  - Reptile
- EC using a Suffix Tree/Array Approach
  - RACER
  - SHREC
  - HiTEC
  - HSHREC
  - PSAEC
- EC using Multiple Sequence Alignment
  - Coral
  - Echo
  - MyHybrid

Error Correction Strategies

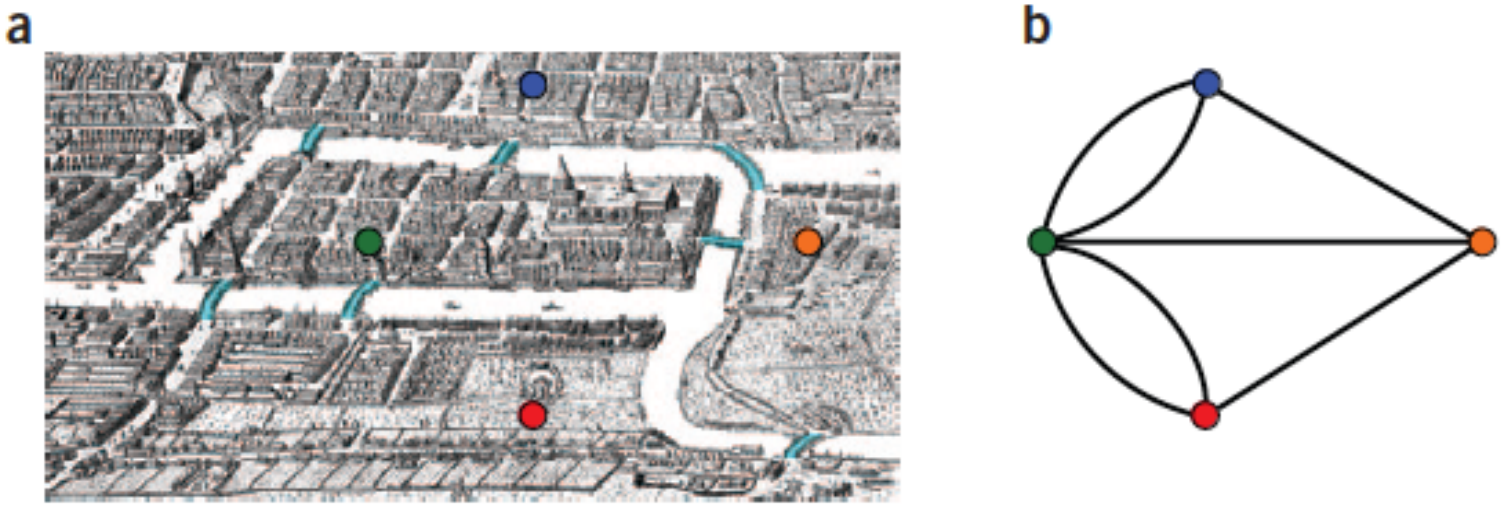




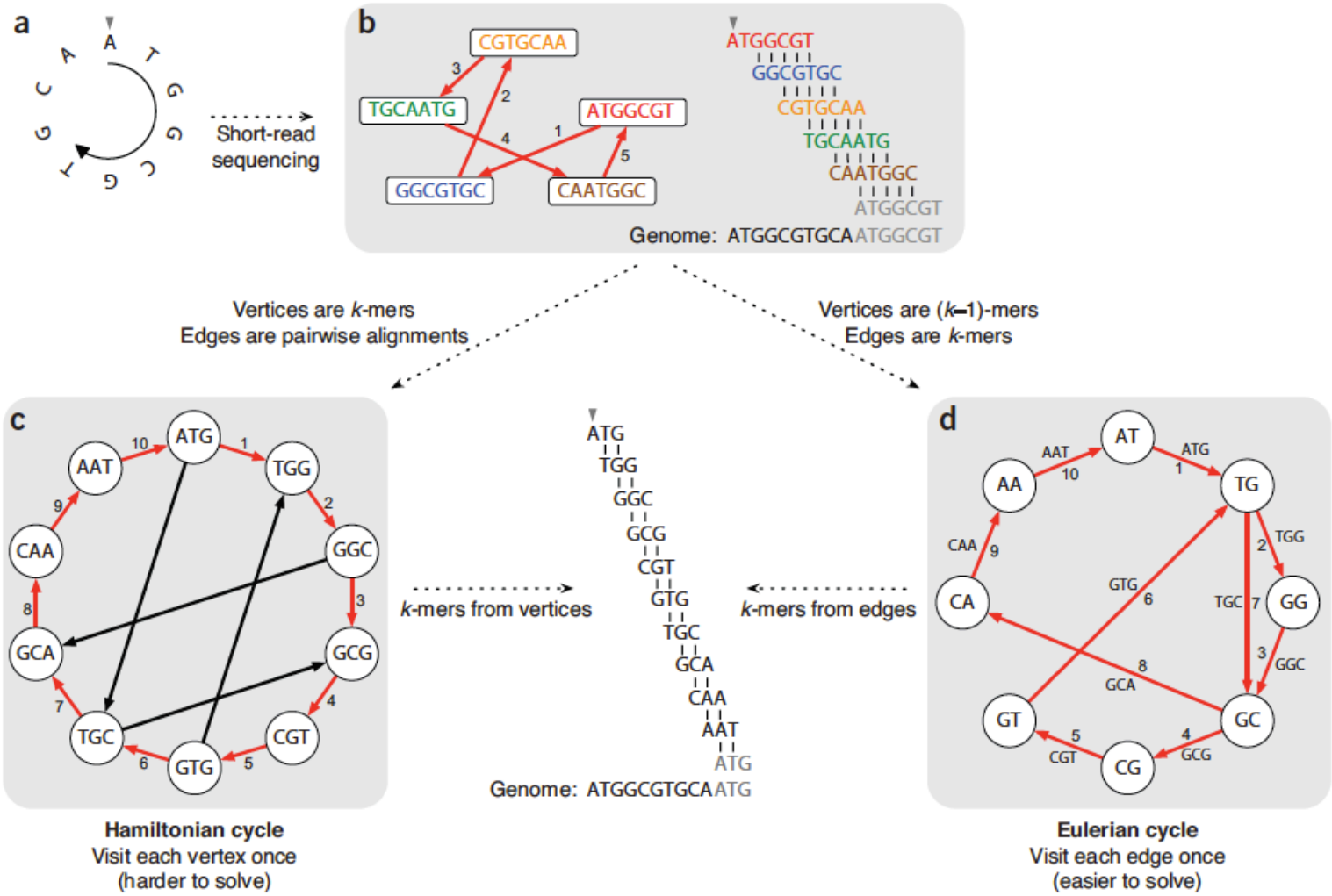


El-Metwally et al. 2013

Bridges of Koenigsberg problem (Graph Theory by Euler)

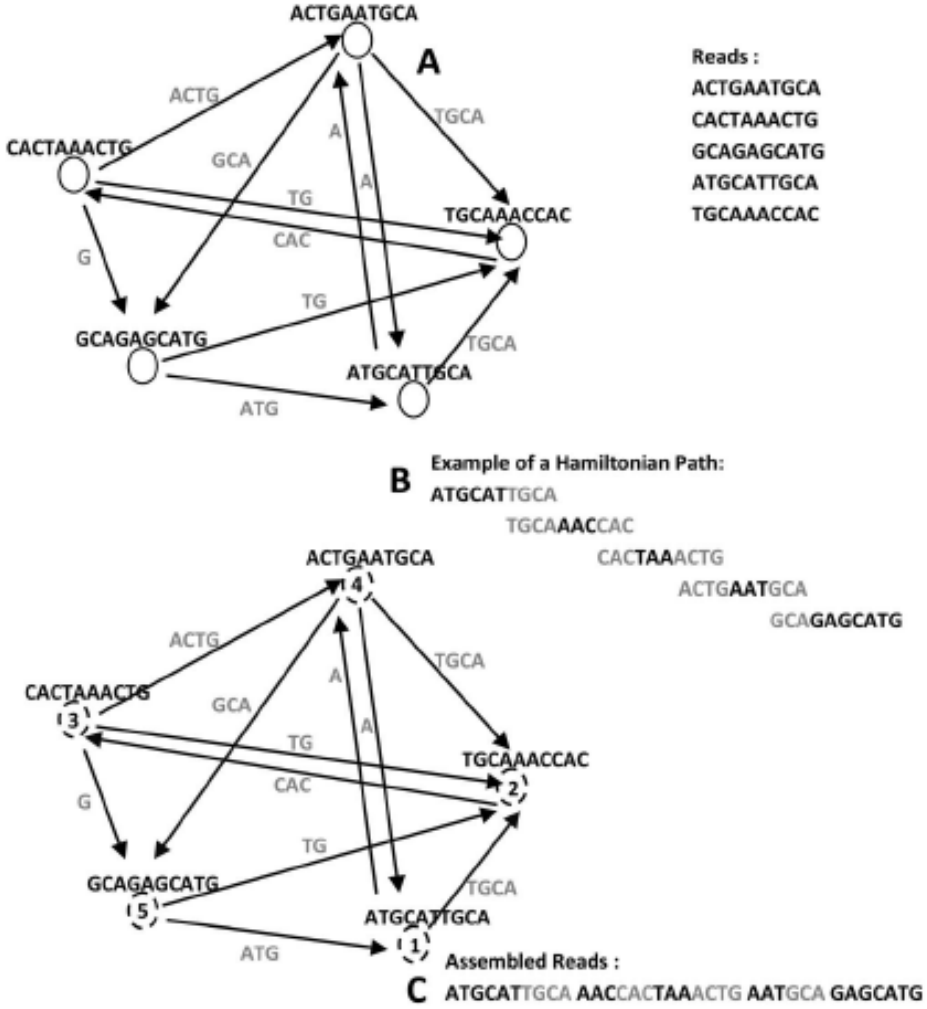


Compeau et al. 2011

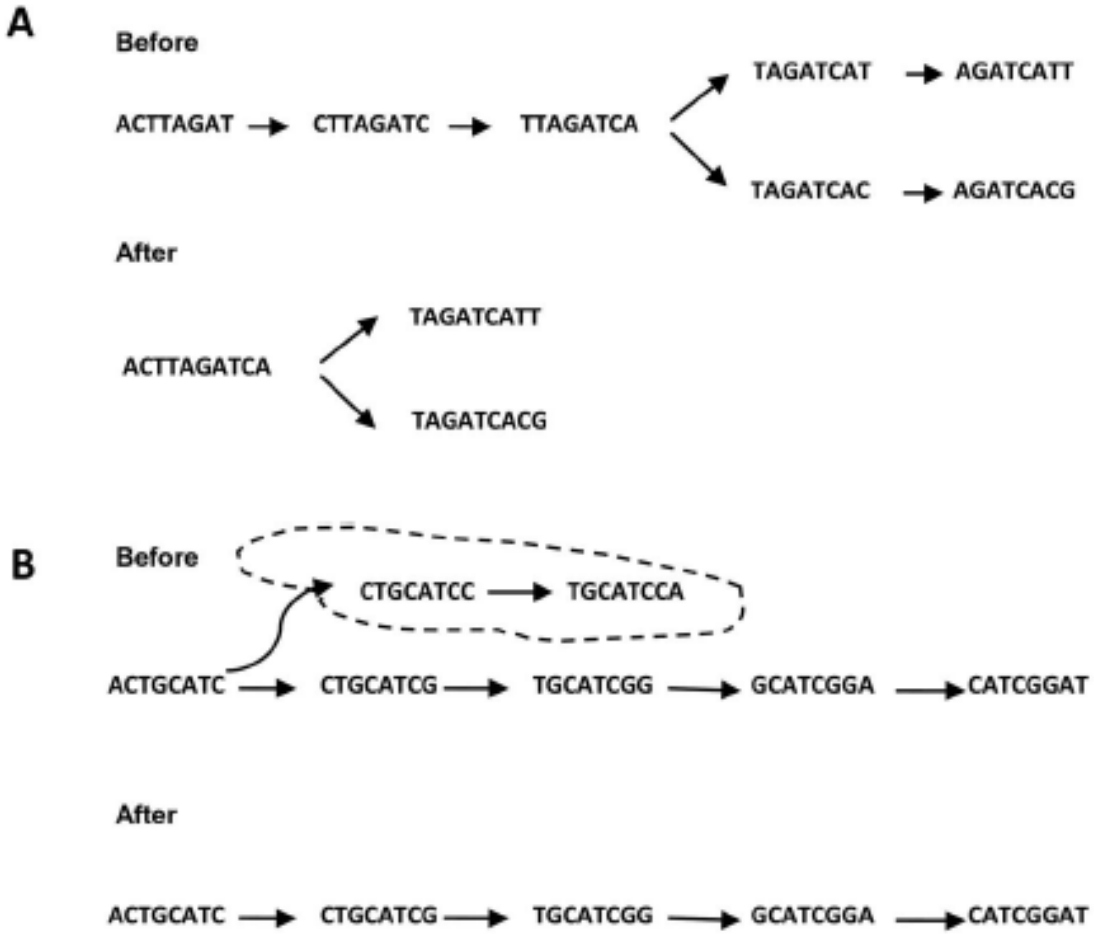


Compeau et al. 2011

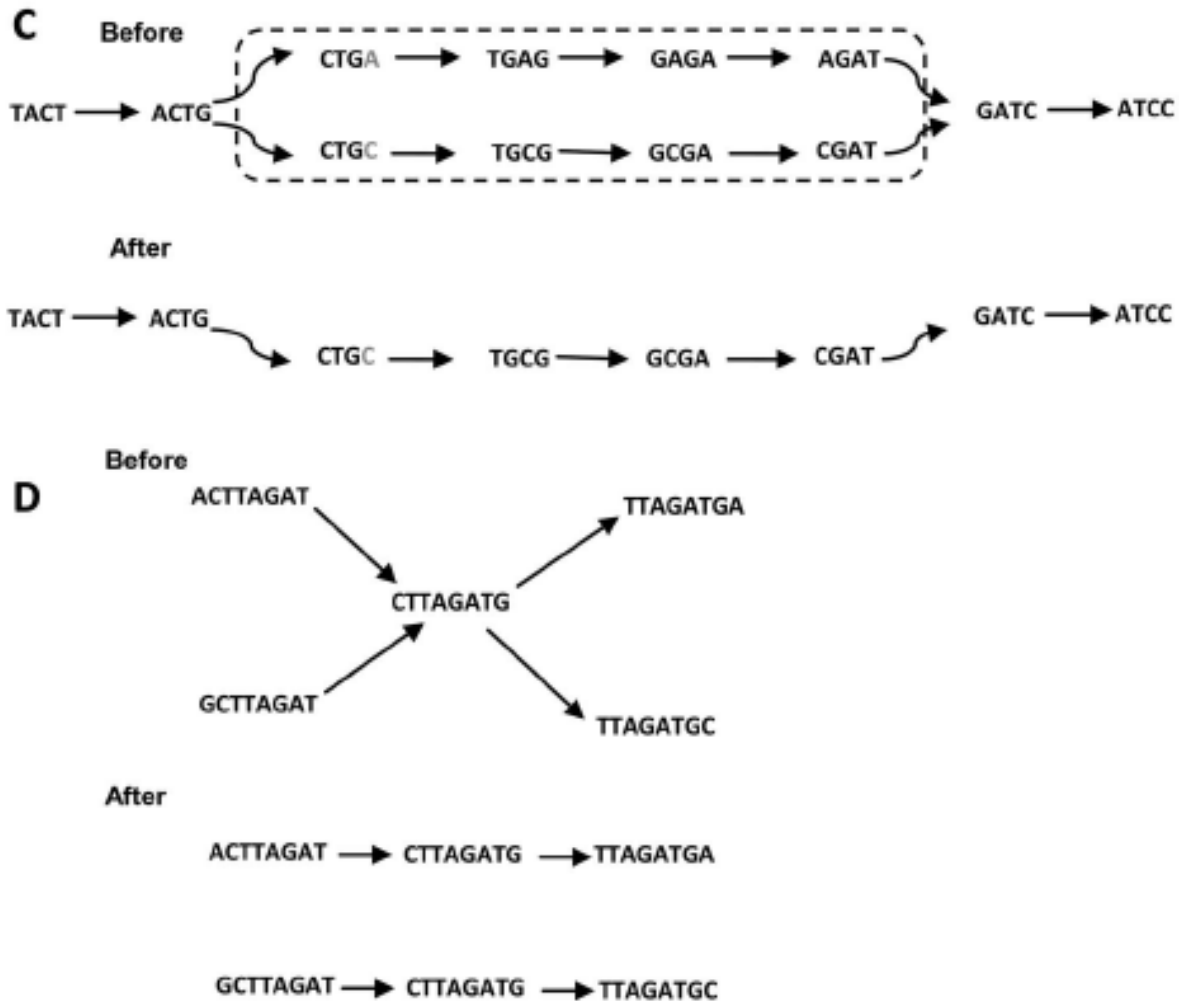
Overlap-layout-consensus Graph



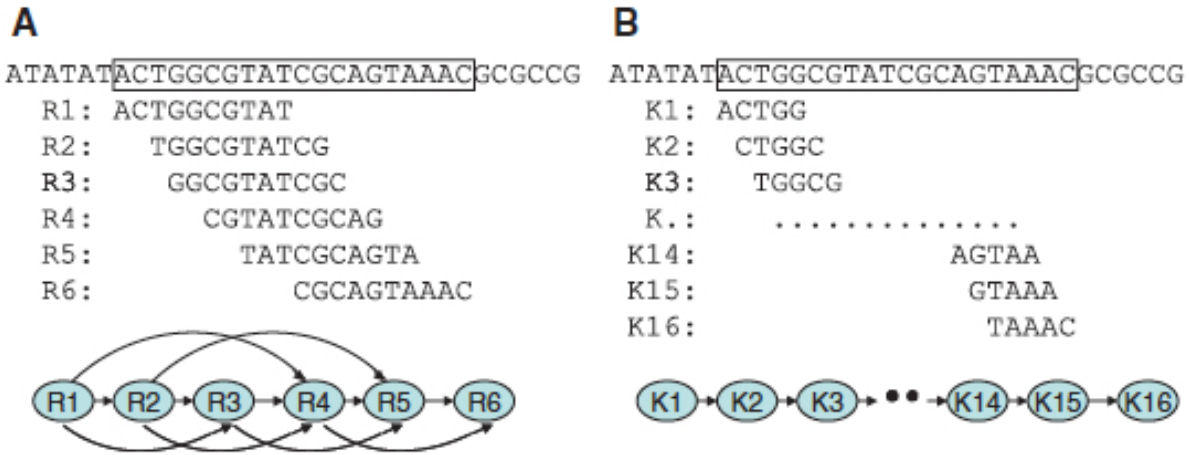
Graph Simplification



Graph Simplification



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



Li et al. 2011, Briefings in Functional Genomcis

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

- *De Bruijn* Graph Assembler
  - Allpaths-LG (Gnerre et al. 2011)
  - Abyss (Simpson et al. 2009, 2016)
  - SOAPdenovo (Luo et al. 2012)
  - Platanus (Kajitani et al. 2014)
  - Discovar (Weisenfeld et al. 2014)
- Overlap-Layout-Consenus Assembler
  - SGA (Simpson and Durbin 2012)
  - Celera assembler (Myers et al. 2002)
- Greedy-Based Assembler
  - SSAKE (Warren et al. 2007)
- 10X Genomics
  - SuperNova (10X Genomics)

Short Read Scaffolder - Illumina

- SSPACE (Boetzer et al. 2011)
- BESST (Sahlin et a. 2014)

Short Read Scaffolder - 10X

- Architect (Kuleshov et al. 2016)
- ARCS
- fragscaff (Adey, et al. 2014)
- links (Warren et al. 2015)

RNA based scaffolding

- L_RNA_Scaffolder (Xue et al. 2013)
- Rascaf (Song et al. 2016)
- AGOUTI (Zheng et al. 2016)

Reference-assisted Chromosome Assembly

- RACA (Kim et al. 2013)
- Ragout (Kolmogorov 2013)

Long Read Assembler - Pacbio

- Falcon
- Sprai
- PBcR - Celera Assembler (Berlin et al. 2014)
- Canu (Koren at al. 2017)
- HINGE (Kamath et al. 2017)

Long Read Assembler - Oxford Nanopore

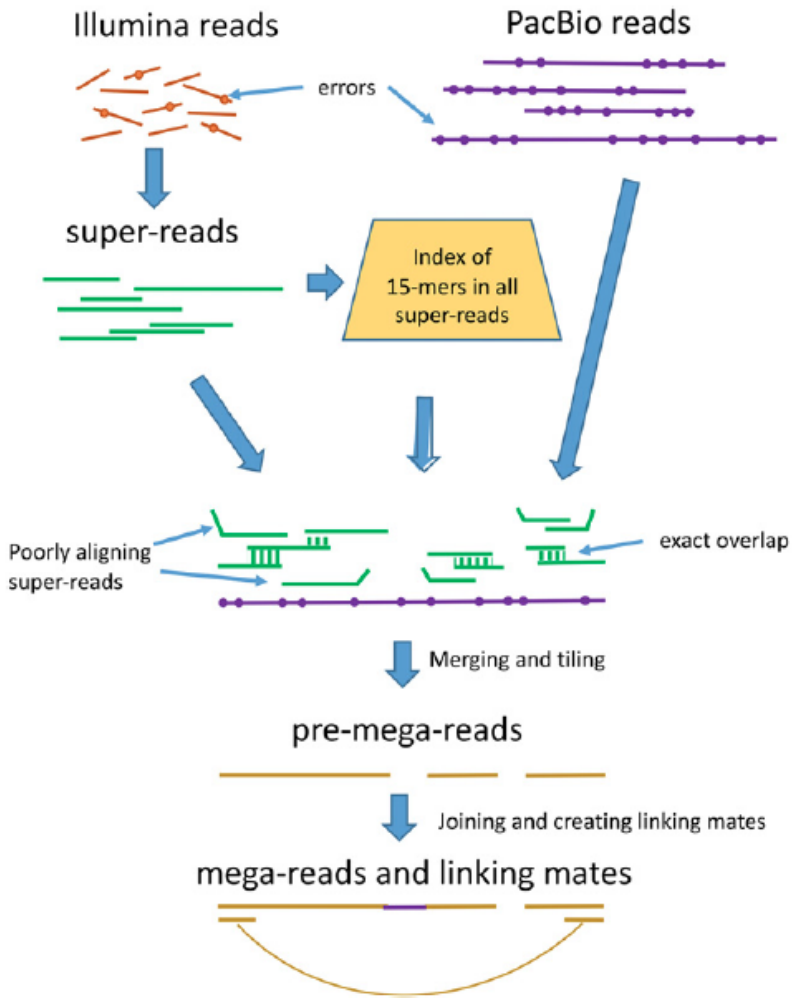
- LQS (Loman, Quick and Simpson 2015)
- Miniasm (Li 2016)
- Falcon
- PBcR - Celera Assembler (Berlin et al. 2014)
- Canu (Koren at al. 2017)
- HINGE (Kamath et al. 2017)

Hybrid Assembler

- Allpaths-LG (Gnerre et al. 2011)
- pacBioToCA
- SPAdes (Bankevich et al. 2012)
- Masurca (Zimin et al. 2017)

Long Read Scaffolder

- SSPACE-Longread (Boetzer and Pirovano 2014)
- PBjelly (English et al. 2012)
- LINKS
- npScarf (Cao et al. 2017)

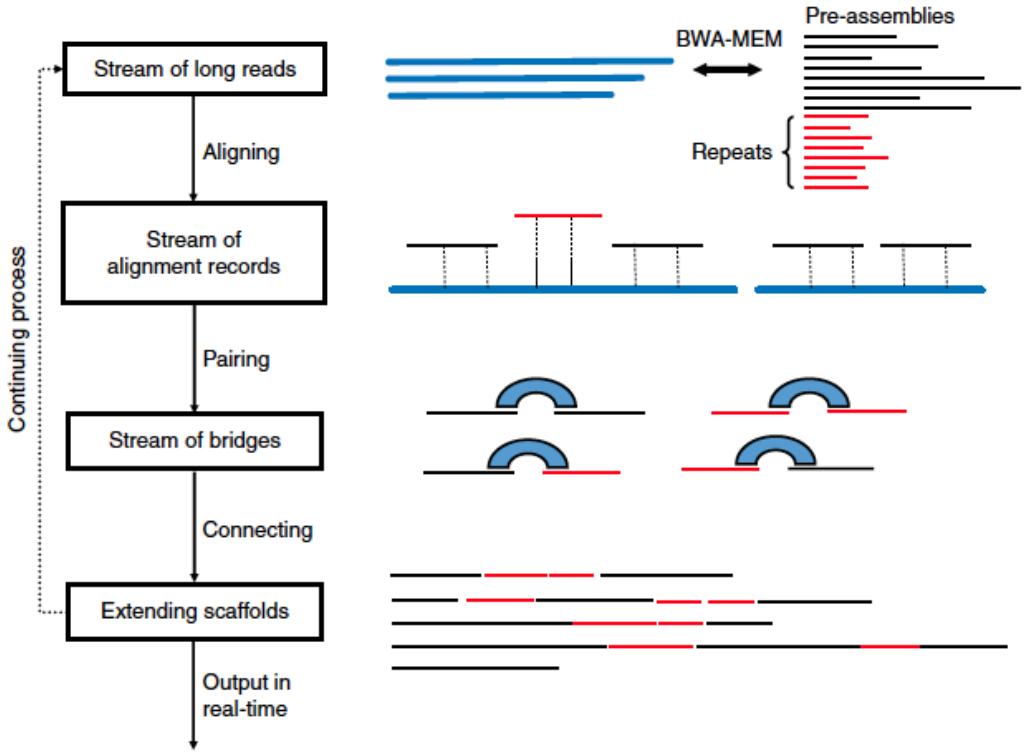


Hybrid Assembler

- Allpaths-LG (Gnerre et al. 2011)
- pacBioToCA
- SPAdes (Bankevich et al. 2012)
- Masurca (Zimin et al. 2017)

Long Read Scaffolder

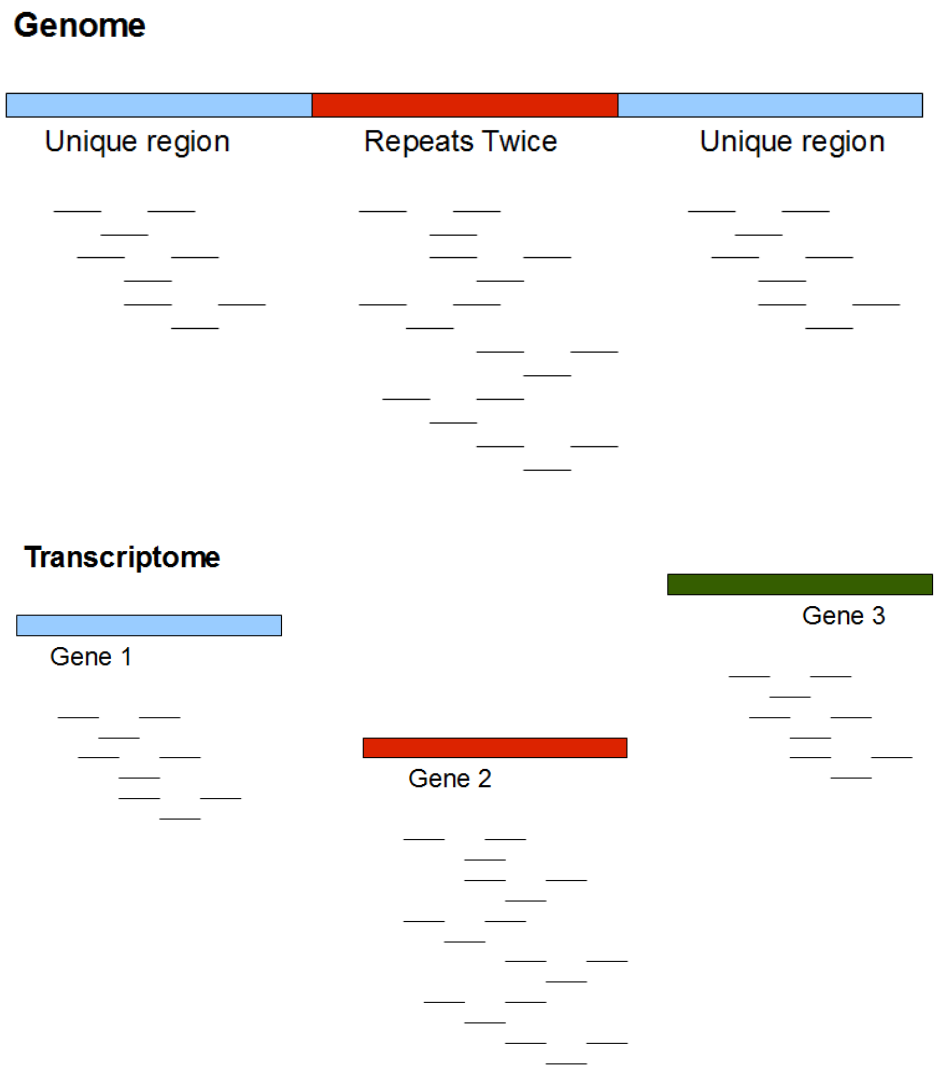
- SSPACE-Longread (Boetzer and Pirovano 2014)
- PBjelly (English et al. 2012)
- LINKS
- npScarf (Cao et al. 2017)



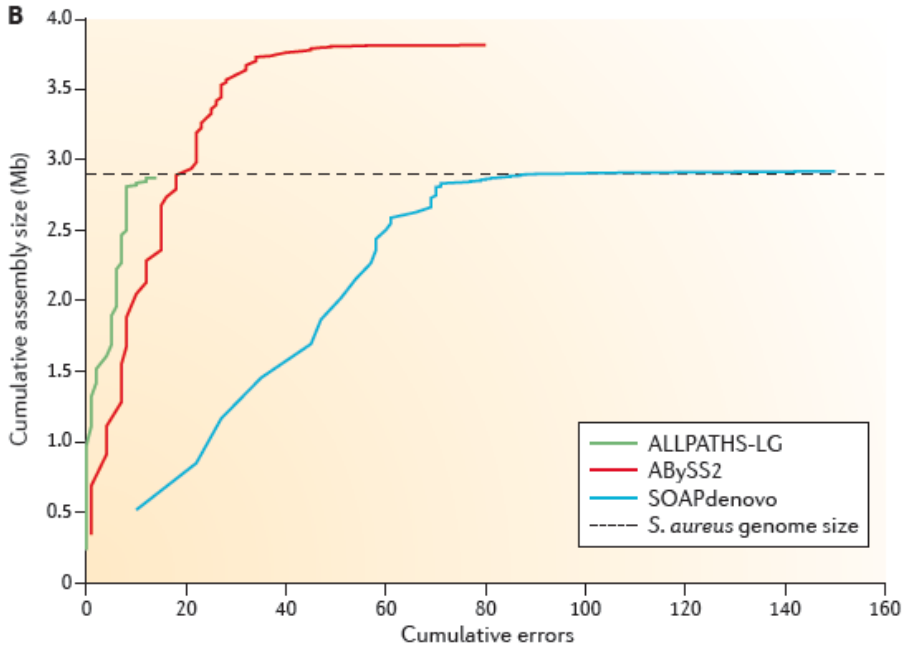
Short Read Assembler II

- Microbe Assemblers
  - Spades (Bankevich et al. 2012)
  - Newbler (http://www.454.com/products/analysis-software/)
  - Mira3 (Miller et al. 2008)
  - Velvet (Zerbino and Birney 2008)

- Transcriptome Assemblers
  - Trinity (Grabherr et al. 2011)
  - Bridger (Chang et al. 2015)
  - Trans-Abyss (Robertson et al. 2010)
  - Oase (Schulz et al. 2012)
  - HiSat2 + Stringtie

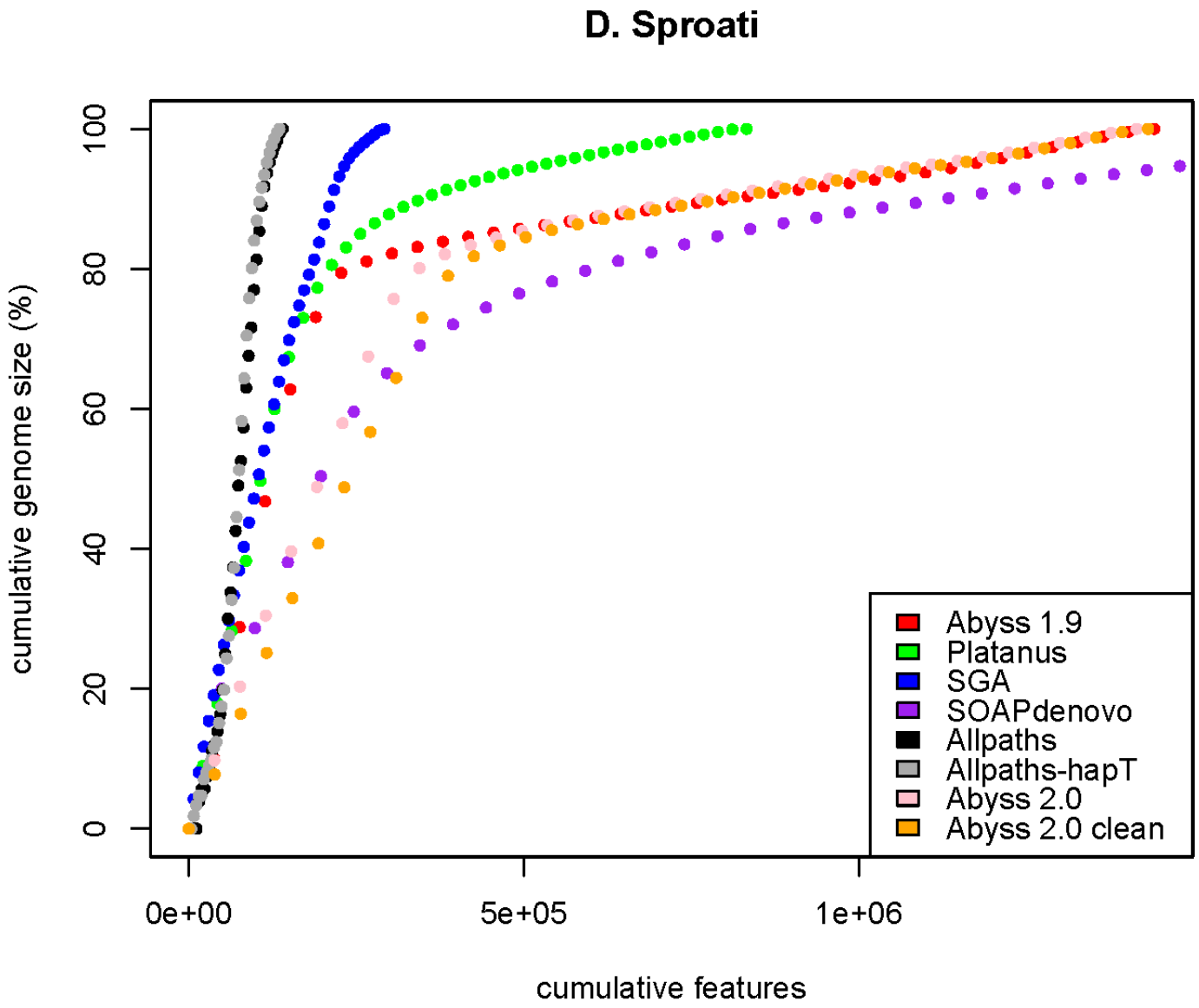

Genome
Unique region
Repeats Twice
Unique region
Transcriptome
Gene 1
Gene 2
Gene 3

- By annotation: CEGMA (Parra et al. 2007)
- By annotation: BUSCO (Simao et al. 2015)
- Using RNA transcripts: Baa.pl (Ryan 2013/4, Arxiv), GMAP (Wu et al. 2005)
- *De novo* likelihood-based measures (LAP; Ghodsi et al. 2013)
- Feature Response Curve (FRC; Vezzi et al. 2012)
- Assembly Statistics



Features:

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


D. Sproati
cumulative genome size (%)
0
20
40
60
80
100
0e+00
5e+05
1e+06
cumulative features
Abyss 1.9
Platanus
SGA
SOAPdenovo
Allpaths
Allpaths-hapT
Abyss 2.0
Abyss 2.0 clean

Gap Filling

- Sealer (Paulino et al. 2015)
- GapCloser (Luo et al. 2012)
- GapFiller (Nadalin et al. 2011)

Resolving Misassemblies

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

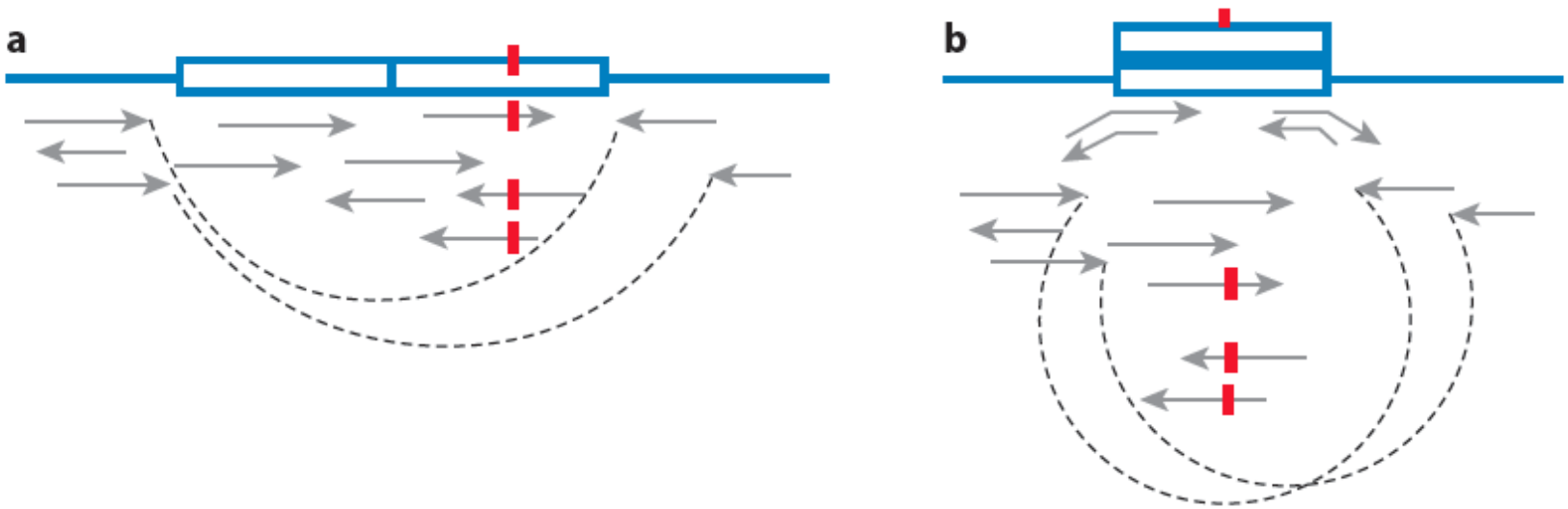
Genome Merging

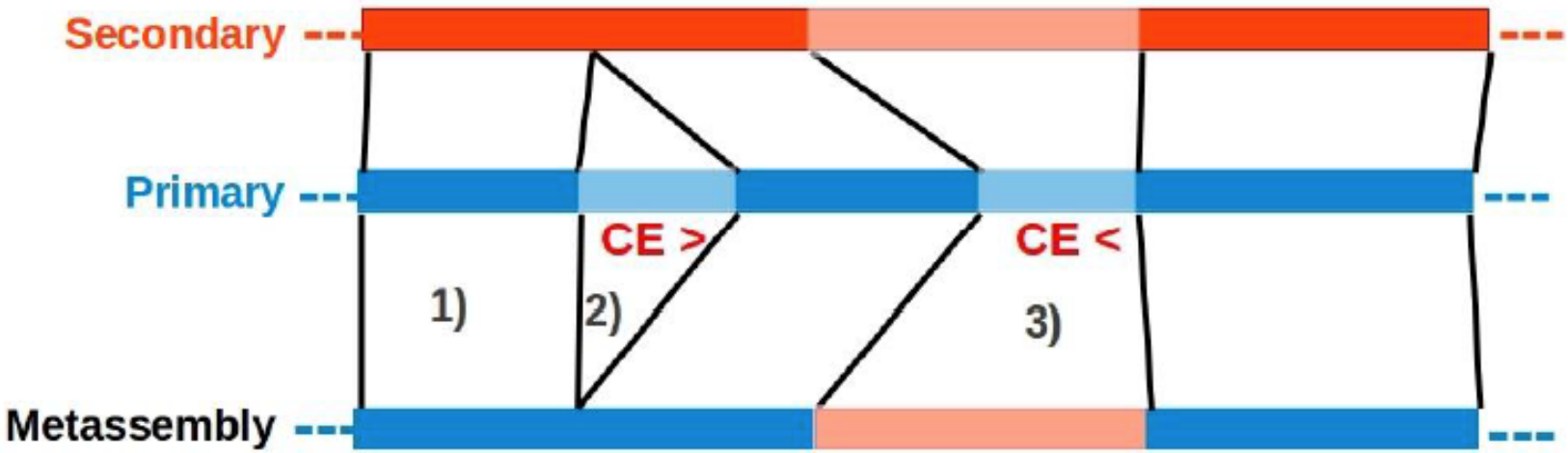
- Metassembler (Wences and Schatz 2015, Arxiv)
- GAM-NGS (Vicedomini et al. 2013)
- Quickmerge (Chakraborty et al. 2015 bioRxiv)

Consensus Polishing for Long Read Data

- Quiver
- Arrow
- Nanopolish (Loman et al. 2015)

Finding Misassemblies





(Wences and Schatz 2015, Arxiv)

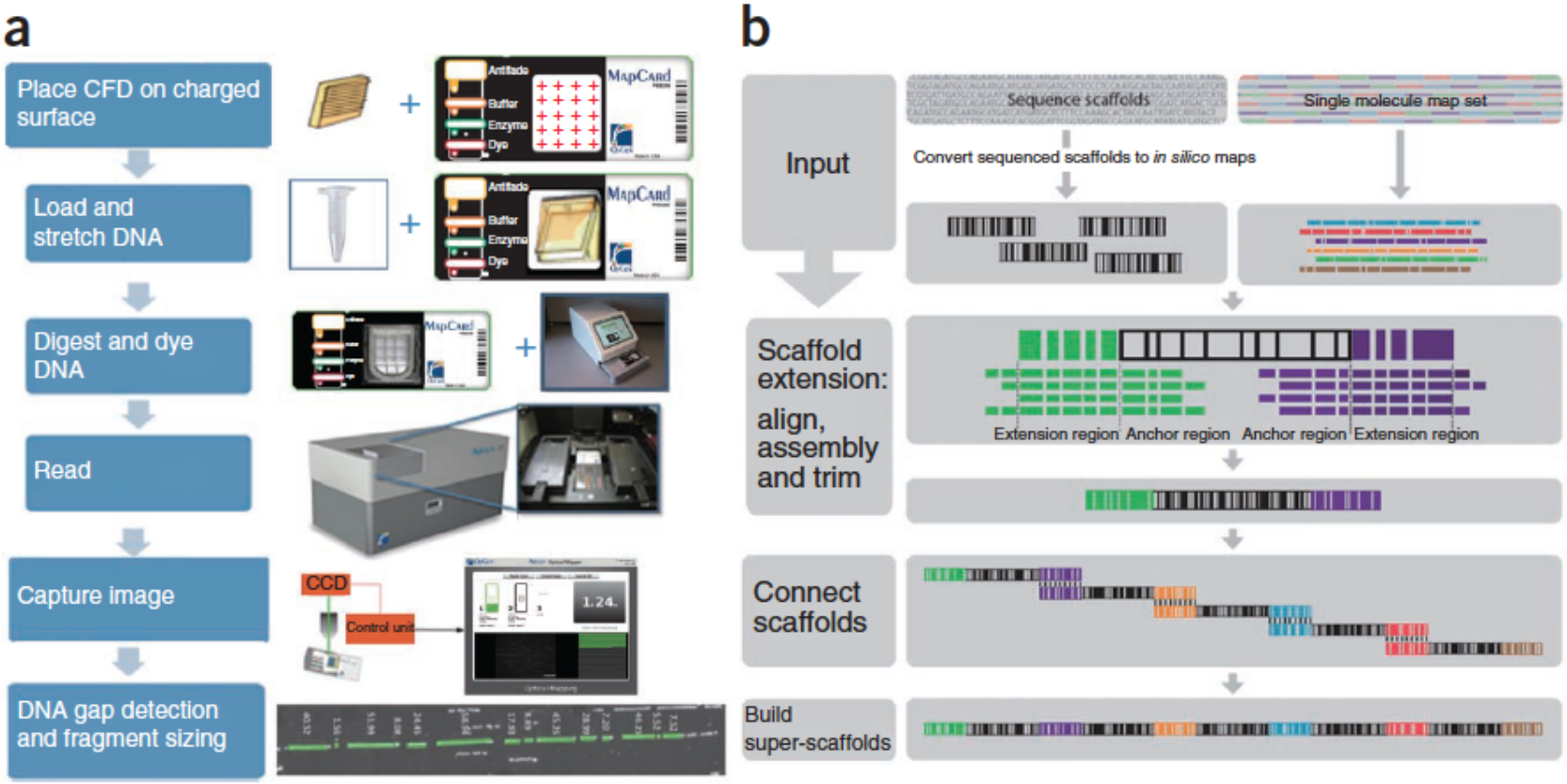
Consensus Polishing - PacBio

- Quiver
- Arrow

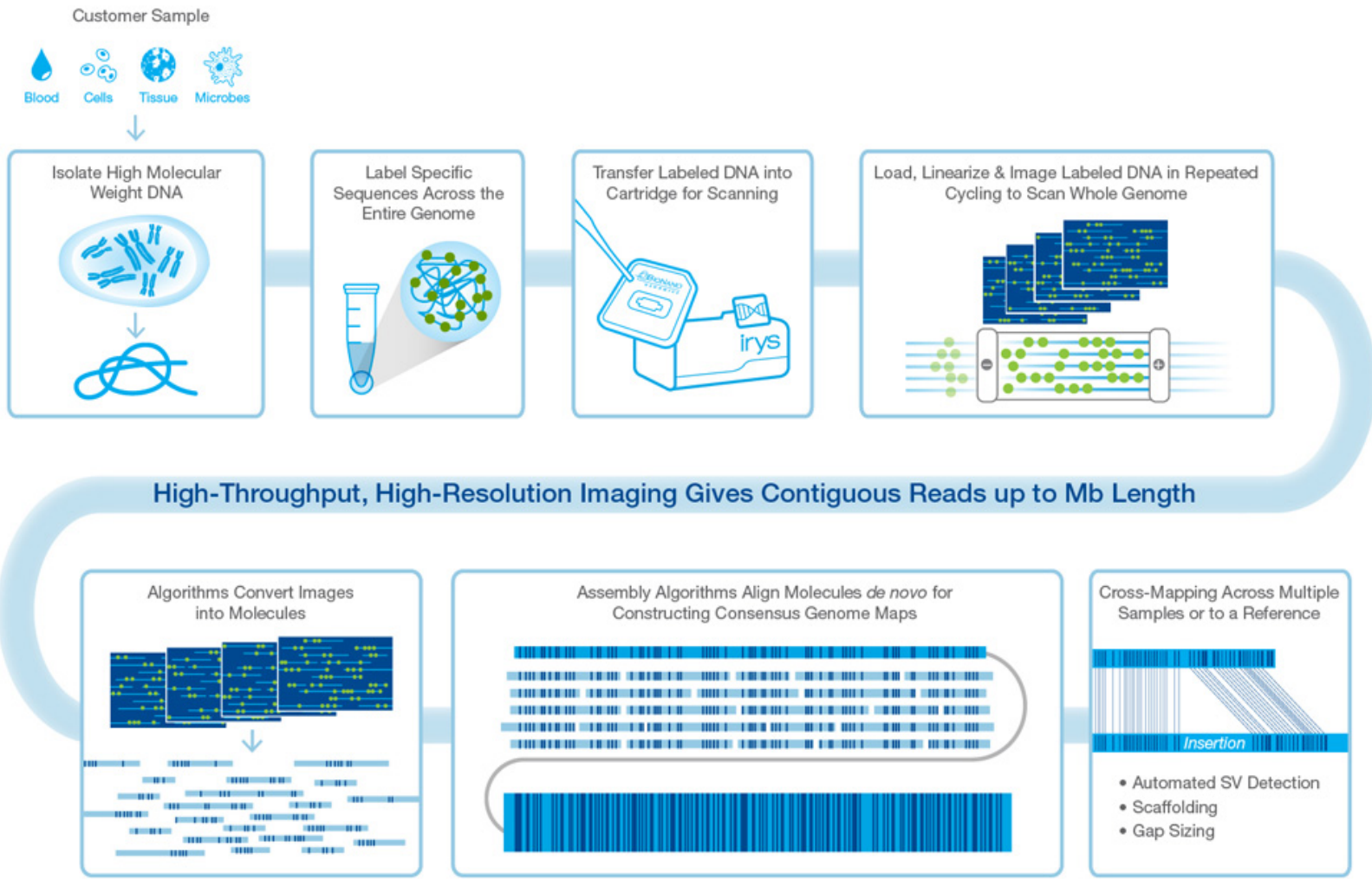
Consensus Polishing - Nanopore

- Nanopolish
- polishing/contig extension: FinisherSC (Lam et a. bioRxiv)

Optical Genome Mapping (OpGen)



Nanochannel-based Genome Mapping (Bionano Irys)

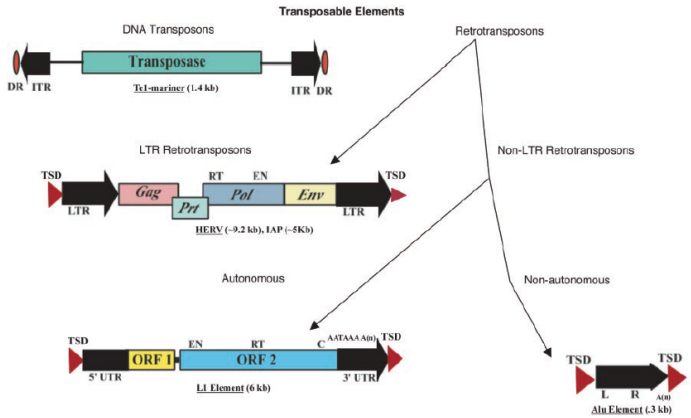


What is a good assembly and when is it finished?

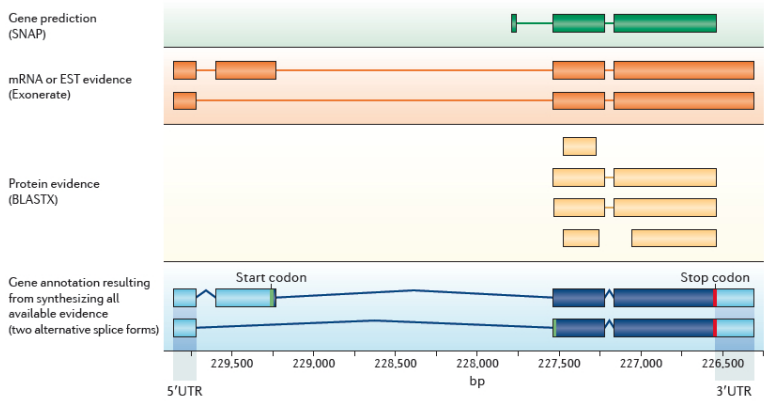
- No Finished Assemblies
  - >100 Euchromatic Gaps in Human Genome
  - *Drosophila melongaster* release 6.1 (More Centric Heterochromatic Sequences Added)
- Scaffold N50 Close to or Bigger than Average Gene Size
- Number of Scaffolds Should be Close to Number of Chromosomes

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 - Computational Methods
8. Further Improvement of the Assembly - Laboratory Methods
9. Mind the Gap! Or not??
10. Downstream Processing

1 Repeat Annotation



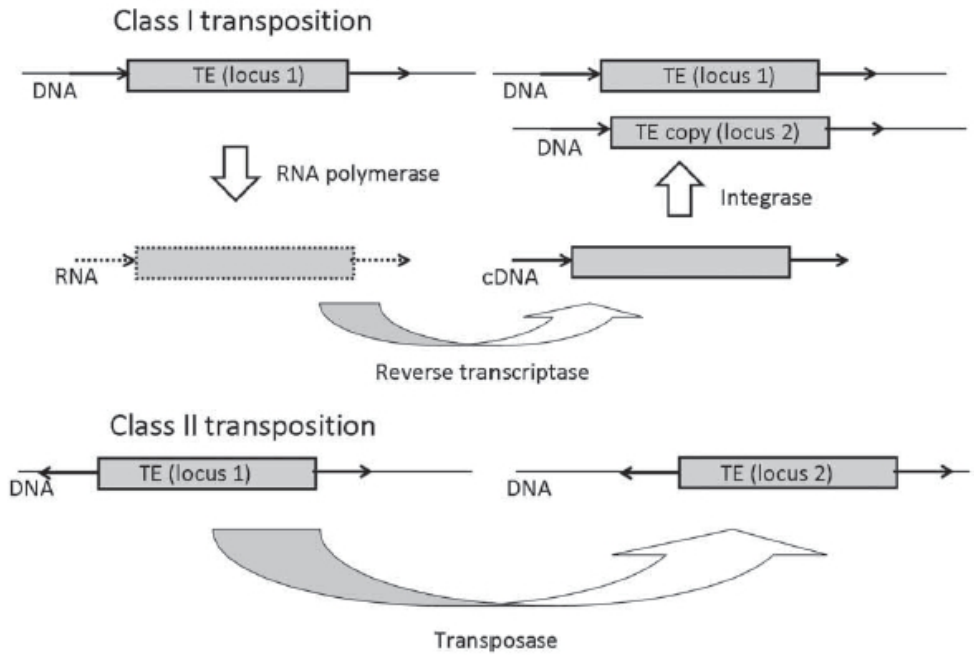
2 Gene Annotation



Repeat Annotation

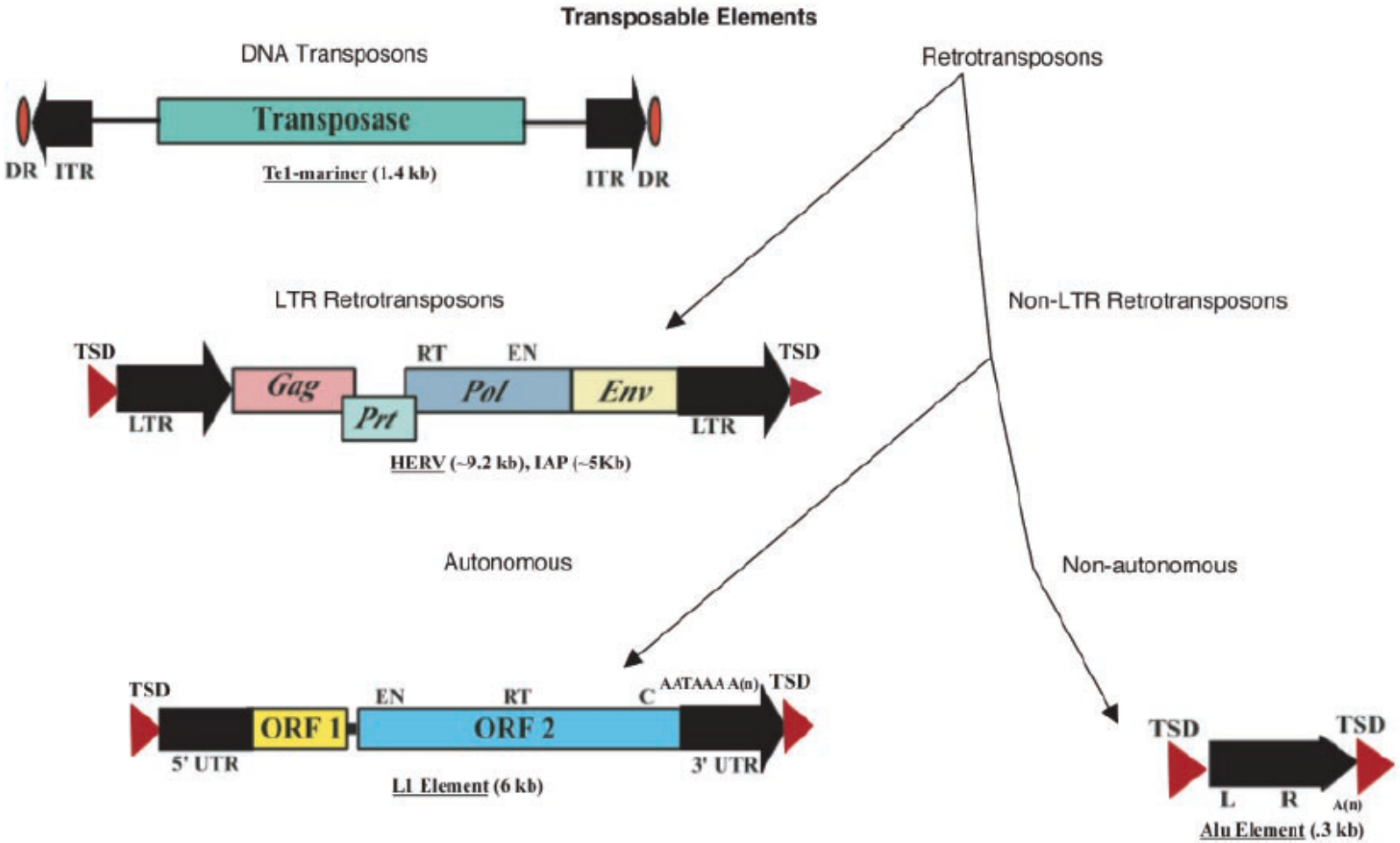
1. Mask Repeats for Gene Annotation
2. Study Repeat Content and Evolution

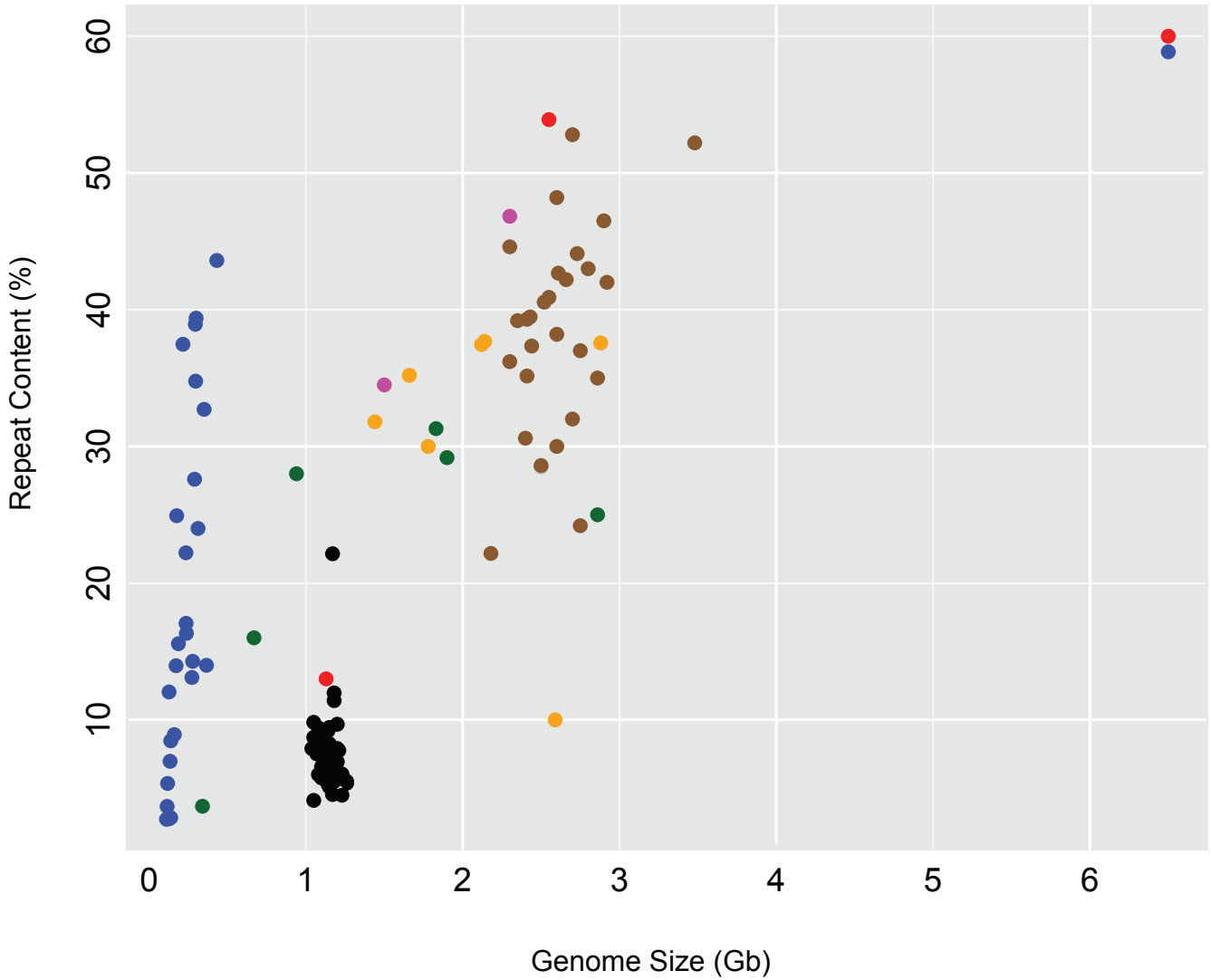
- **Low Complexity Sequences**
  - Simple Repeats and Satellites
- **Transposable Elements**
  - Class I: Retrotransposons *Copy and Paste*
  - Class II: DNA Transposons *Cut and Paste*

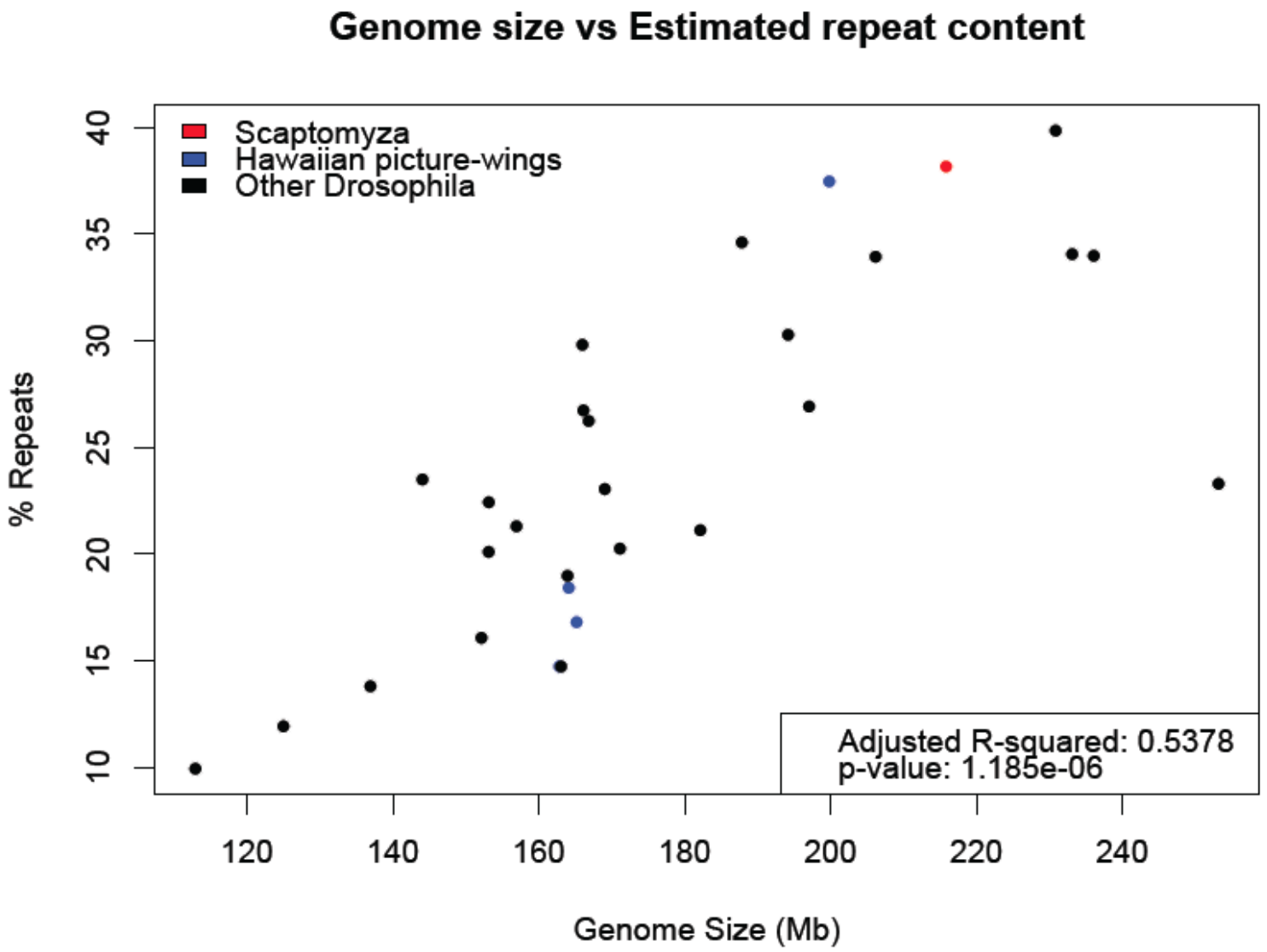


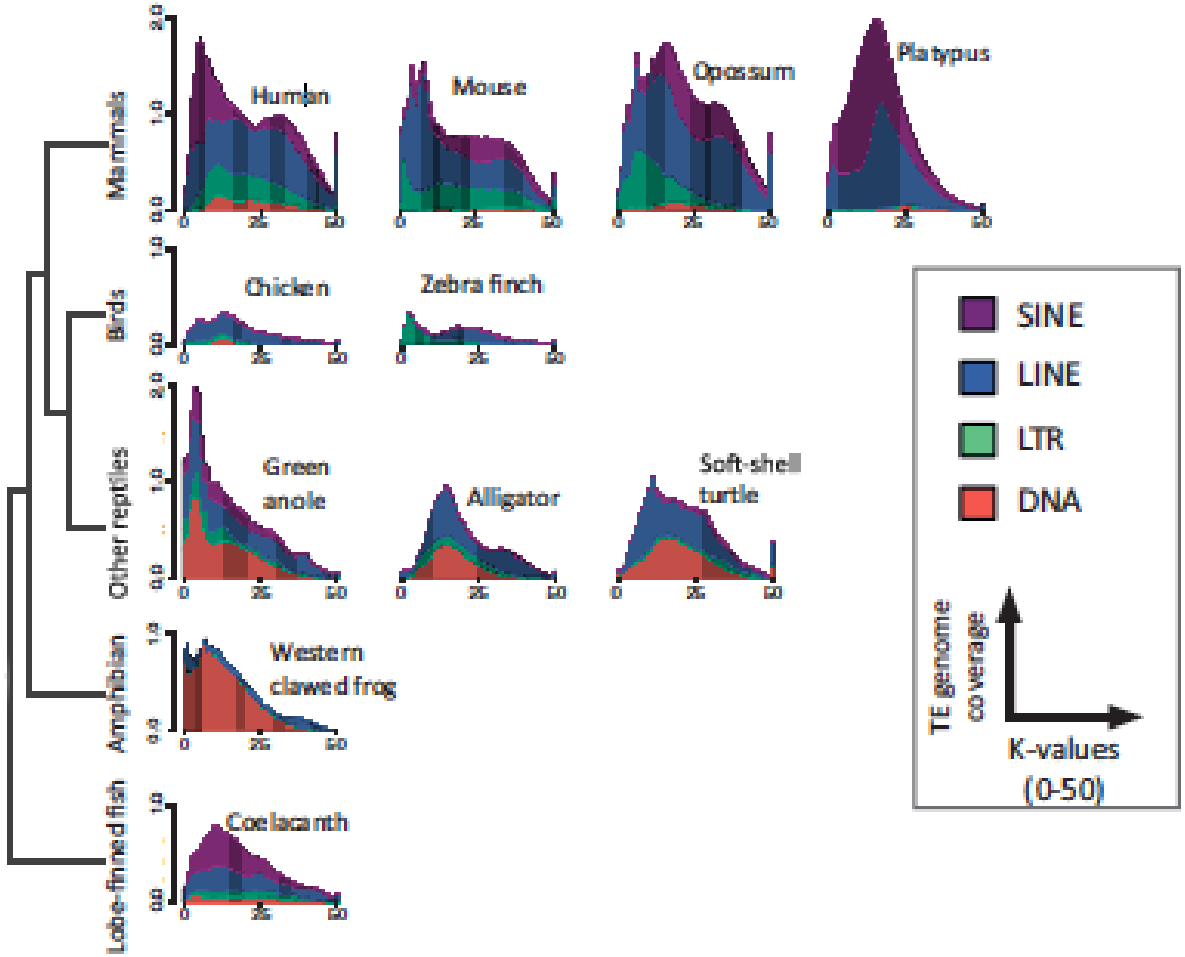
Hermann et al. 2013

- **Class I TEs**
  - Long Terminal Repeats (LTR)
    - Endogenous Retrovirus
  - Non-LTR
    - Long Interspersed Nuclear Elements (LINEs)
    - Short Interspersed Nuclear Elements (SINEs)
- **Class II TEs**
  - Subclass I (both DNA strands are cleaved)
  - Subclass II (Rolling-circle, self-synthesizing)


Transposable Elements
DNA Transposons
Retrotransposons
Transposase
DR ITR
Tc1-mariner (1.4 kb)
ITR DR
LTR Retrotransposons
Non-LTR Retrotransposons
TSD
RT EN
TSD
Gag
Prt
Pol
Env
LTR
LTR
HERV (~9.2 kb), IAP (~5Kb)
Autonomous
Non-autonomous
TSD
EN
RT
C
AATAAA A(n)
TSD
ORF 1
ORF 2
5' UTR
3' UTR
L1 Element (6 kb)
TSD
TSD
L
R
A(n)
Alu Element (.3 kb)


Repeat Content (%)
10
20
30
40
50
60
0
1
2
3
4
5
6
Genome Size (Gb)


Genome size vs Estimated repeat content
Scaptomyza
Hawaiian picture-wings
Other Drosophila
Adjusted R-squared: 0.5378
p-value: 1.185e-06
% Repeats
10
15
20
25
30
35
40
120
140
160
180
200
220
240
Genome Size (Mb)



Chalopin et al. 2015

- **Homology**
  - RepeatMasker (http://www.repeatmasker.org/)
- **De novo**
  - RepeatModeler (http://www.repeatmasker.org/)
  - WindowMasker (Morgulis et al., 2005)
  - RepeatScout (Price et al., 2005)
  - Piler (Edgar and Myers, 2005)
- **De novo from reads**
  - REPdenovo (github)
  - Tedna (Zytnicki et al., 2014)

CAVE: De novo tools can wrongly identify highly conserved protein-coding genes

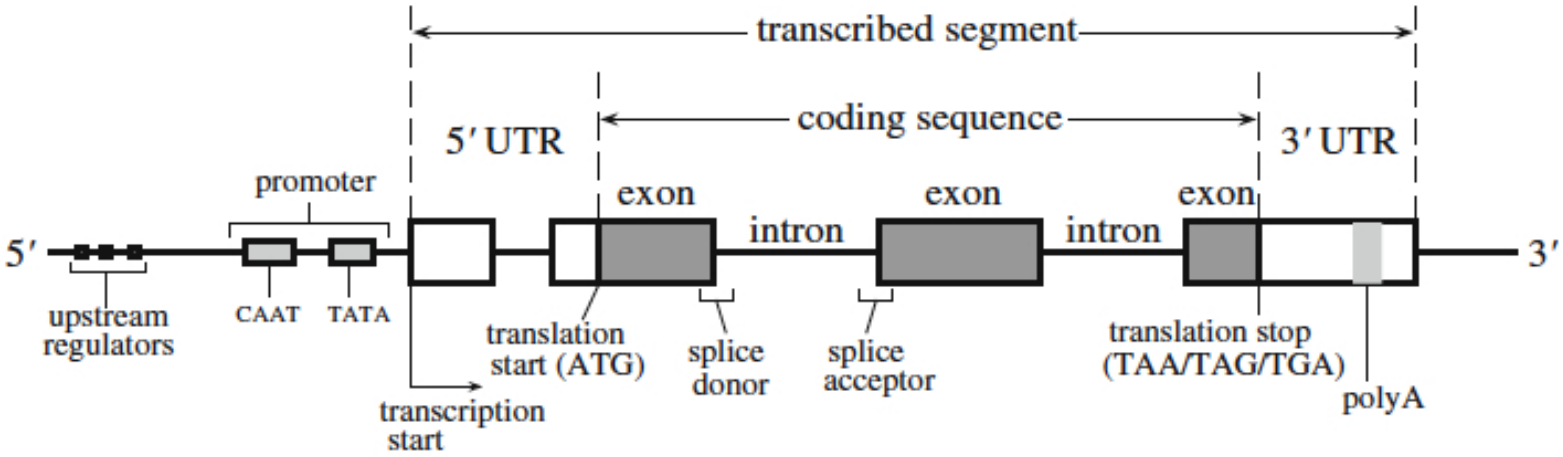
For Gene Annotation

1. BuildDatabase -name genome genome.fasta
2. RepeatModeler -pa xx -database genome
3. Merge *de-novo* and RepBase library
4. RepeatMasker -pa xx -gccalc -nolow -lib [consensi.fa.classified and RepBase] genome.fasta

For Repeat Annotation

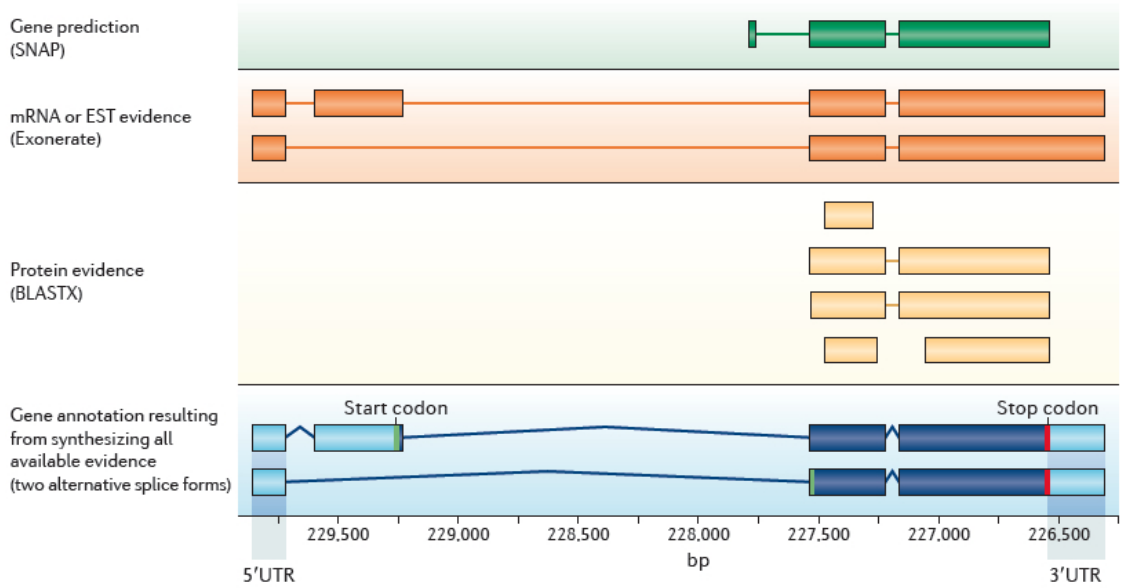
1. BuildDatabase -name genome genome.fasta
2. RepeatModeler -pa xx -database genome
3. Merge *de-novo* and RepBase library
4. RepeatMasker -pa xx -a -gccalc -lib [consensi.fa.classified and RepBase] genome.fasta

Gene Structure



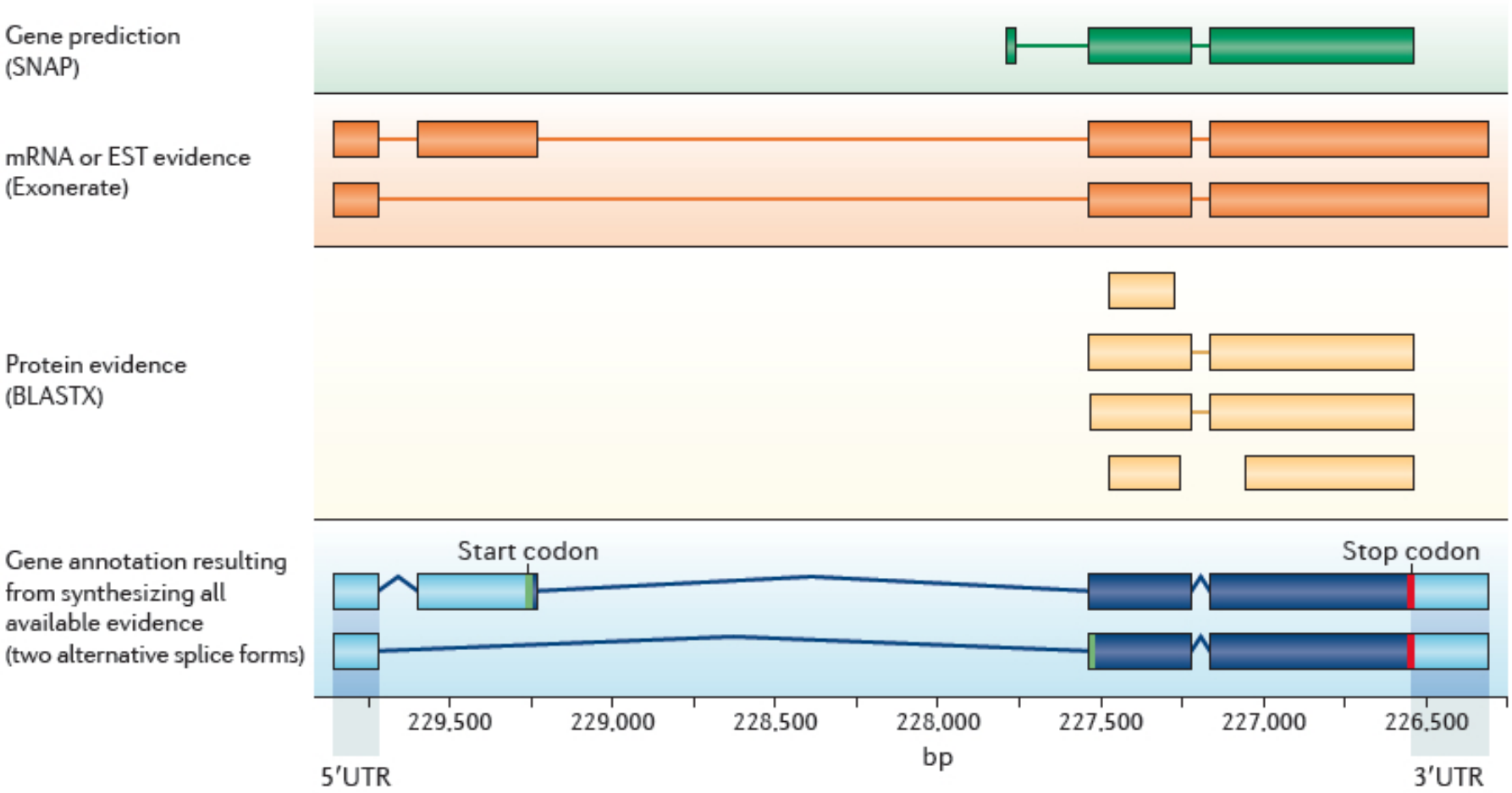
(Marina Axelson-Fisk, Springer-Verlag London, 2015)

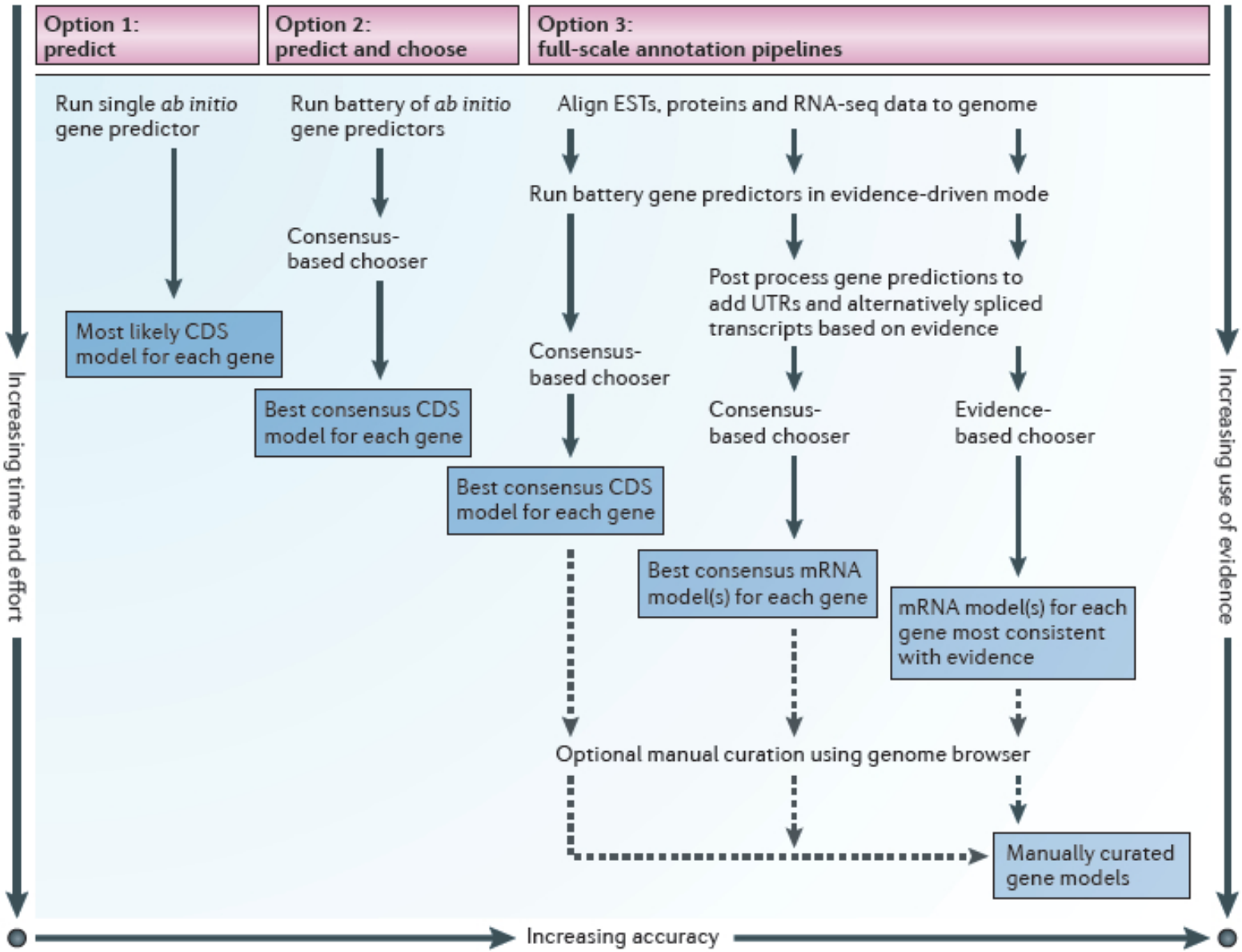
1. **Evidence Alignment**
   - Protein
   - EST
   - RNA-seq
2. **Ab initio Prediction**



Ab Initio Gene Prediction

- Do not need evidence data
- Need to be trained for the organism (codon frequencies, distribution of exon-intron length, etc.)
- Most find single most likely coding sequence (CDS)
- Do not report untranslated regions (UTRs)
- Cannot deal with alternative splicing
- Accuracy usually 60-70%
- With training can be up to 100%
- Need a very good genome assembly


Gene prediction
(SNAP)
mRNA or EST evidence
(Exonerate)
Protein evidence
(BLASTX)
Gene annotation resulting
from synthesizing all
available evidence
(two alternative splice forms)
Start codon
Stop codon
229,500
229,000
228,500
228,000
227,500
227,000
226,500
bp
5′UTR
3′UTR


Option 1: predict
Option 2: predict and choose
Option 3: full-scale annotation pipelines
Run single ab initio gene predictor
Run battery of ab initio gene predictors
Align ESTs, proteins and RNA-seq data to genome
Run battery gene predictors in evidence-driven mode
Consensus-based chooser
Post process gene predictions to add UTRs and alternatively spliced transcripts based on evidence
Most likely CDS model for each gene
Consensus-based chooser
Best consensus CDS model for each gene
Consensus-based chooser
Evidence-based chooser
Best consensus CDS model for each gene
Best consensus mRNA model(s) for each gene
mRNA model(s) for each gene most consistent with evidence
Optional manual curation using genome browser
Manually curated gene models
Increasing time and effort
Increasing use of evidence
Increasing accuracy