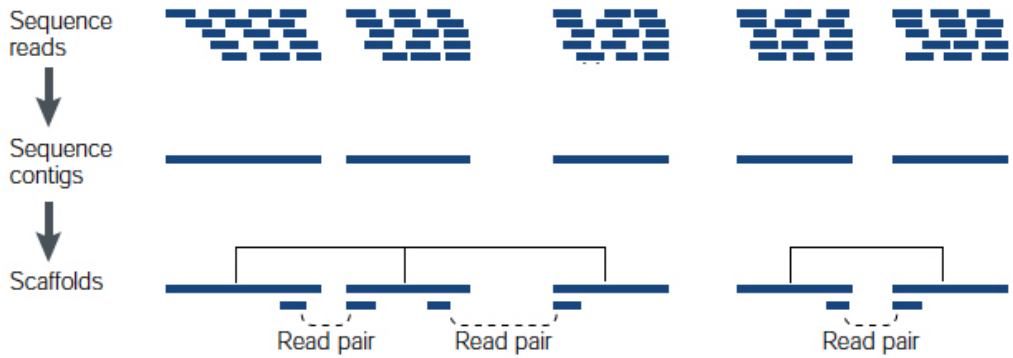
# A Roadmap to Genome De-Novo Assembly

Stefan Prost [1]

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

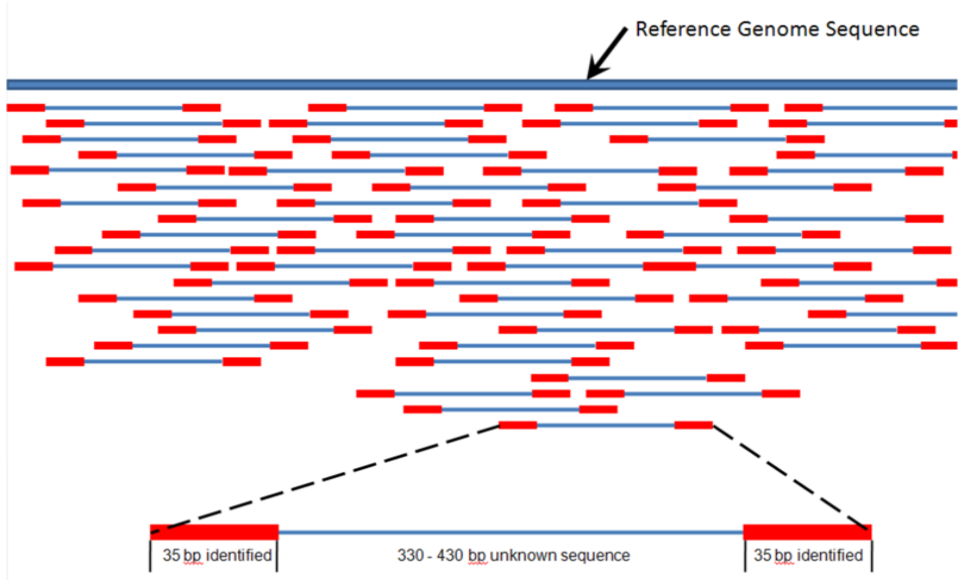
October 26-27th, 2015

## *De-novo* Assembly



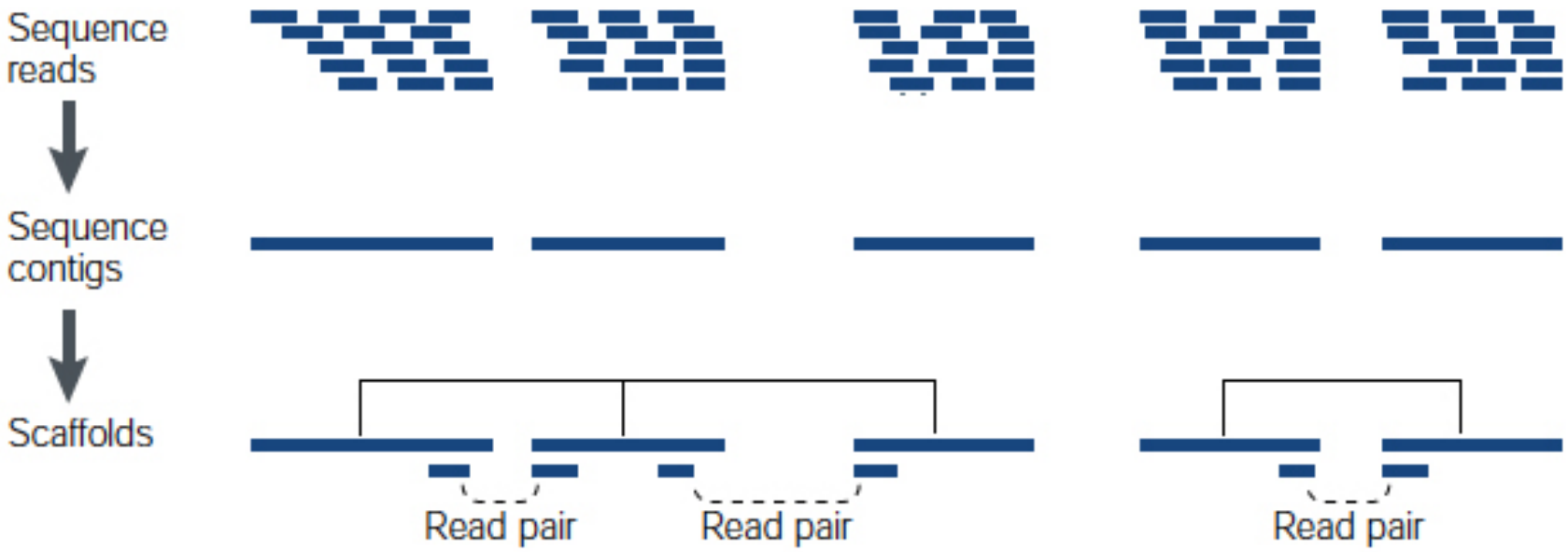

Changed after Green 2001

## Reference-based Mapping



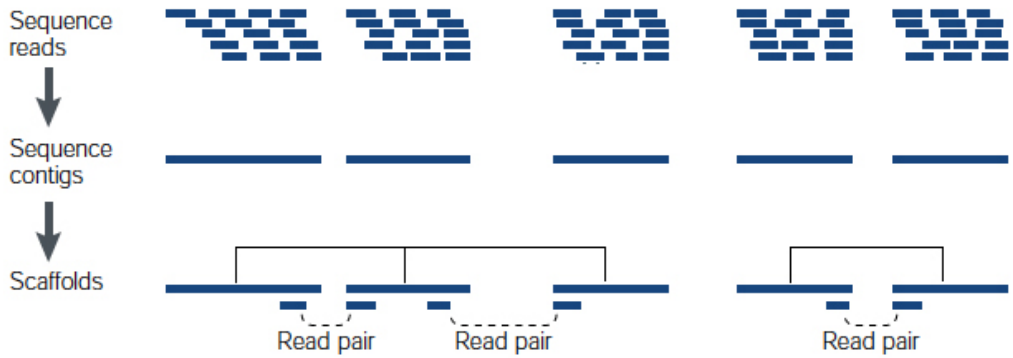

Wikipedia

# De-novo Assembly



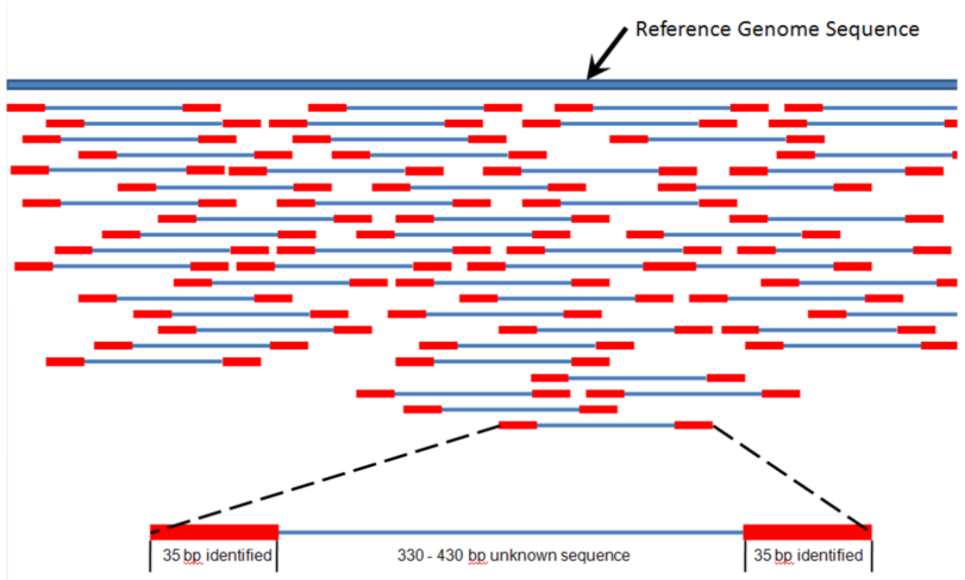

Changed after Green 2001

## *De-novo* Assembly

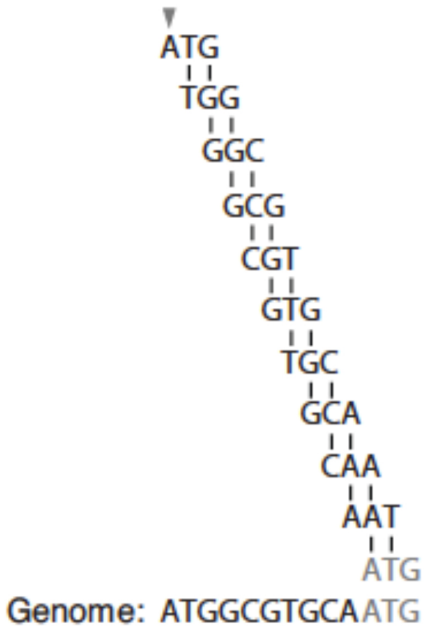



Changed after Green 2001

## Reference-based Mapping



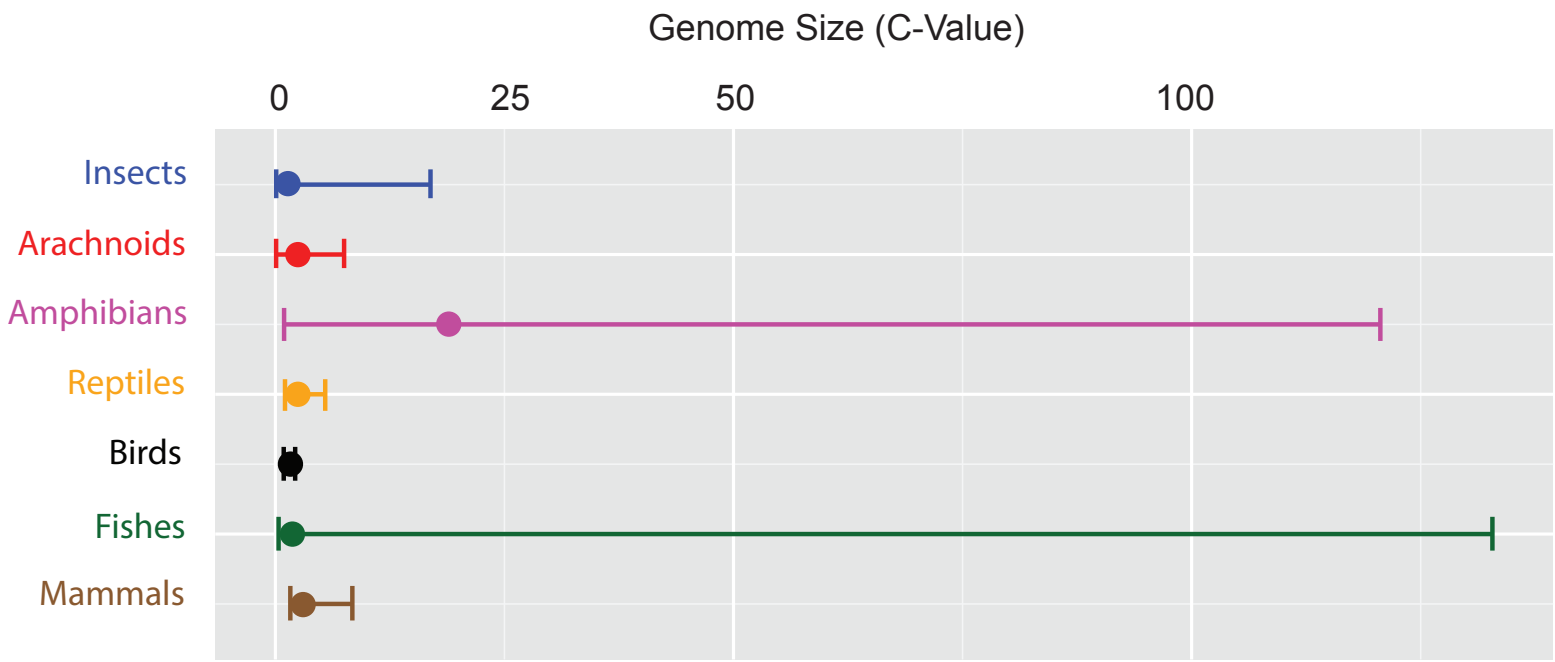

Wikipedia




## Definition: Kmer

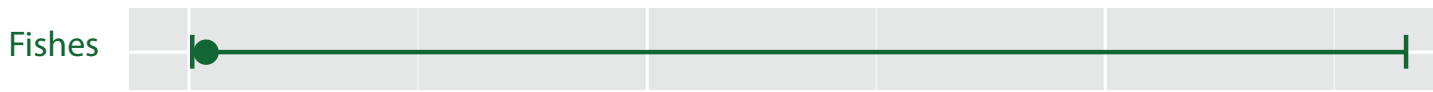
Short, unique element of DNA sequence of length n

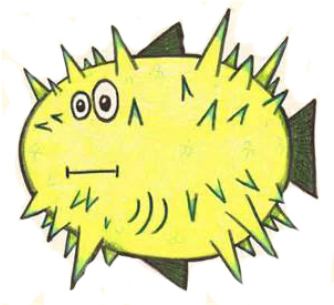
- Expected Genome Size
- Expected Repeat Content
- Expected Heterozygosity
- Diploid or Polyploid?

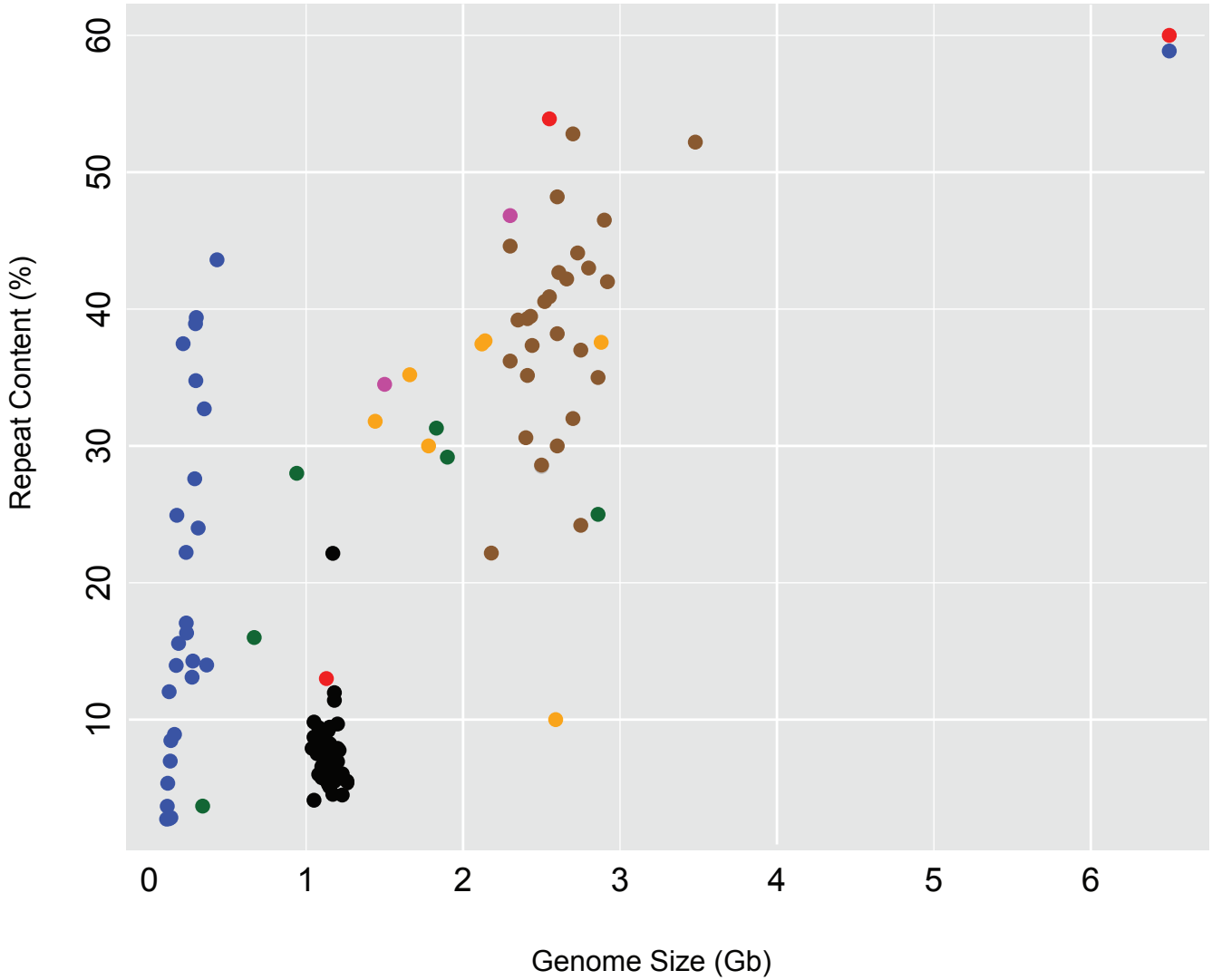


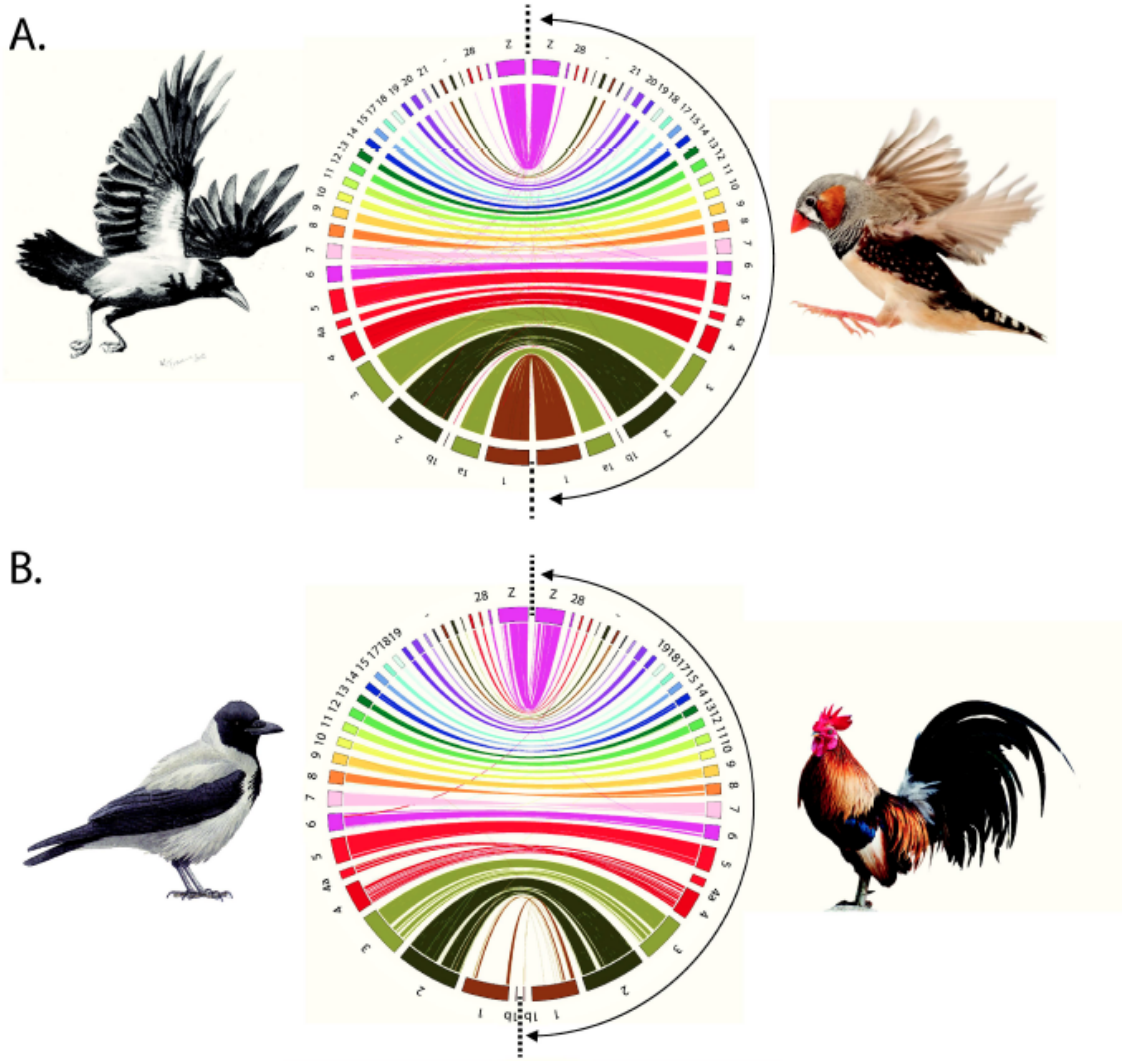

C-values obtained from: www.genomesize.com/

## Genome Size Range in Fishes




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




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?

**1st Generation**

ABI (Sanger)

**2nd Generation**

Roche 454

Illumina HiSeq

Life Technologies IonTorrent

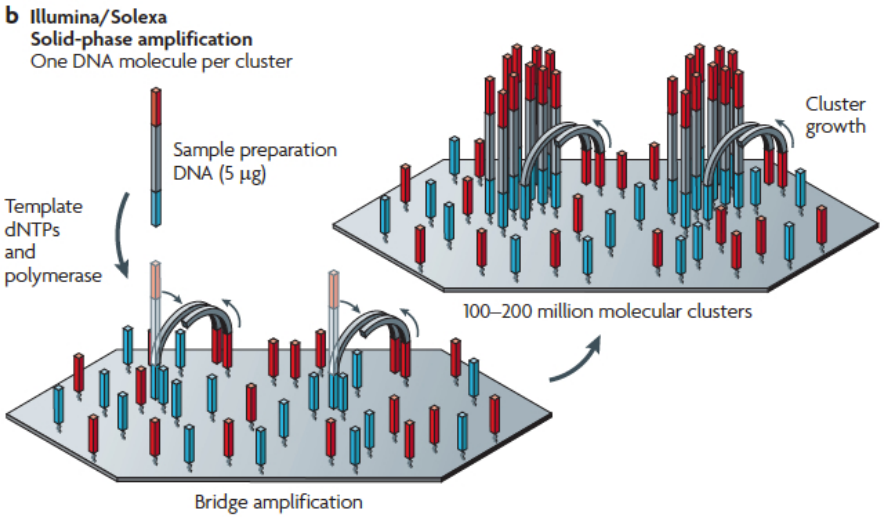
**3rd Generation**


Oxford Nanopore MinIon

Pacific Biosciences RSII

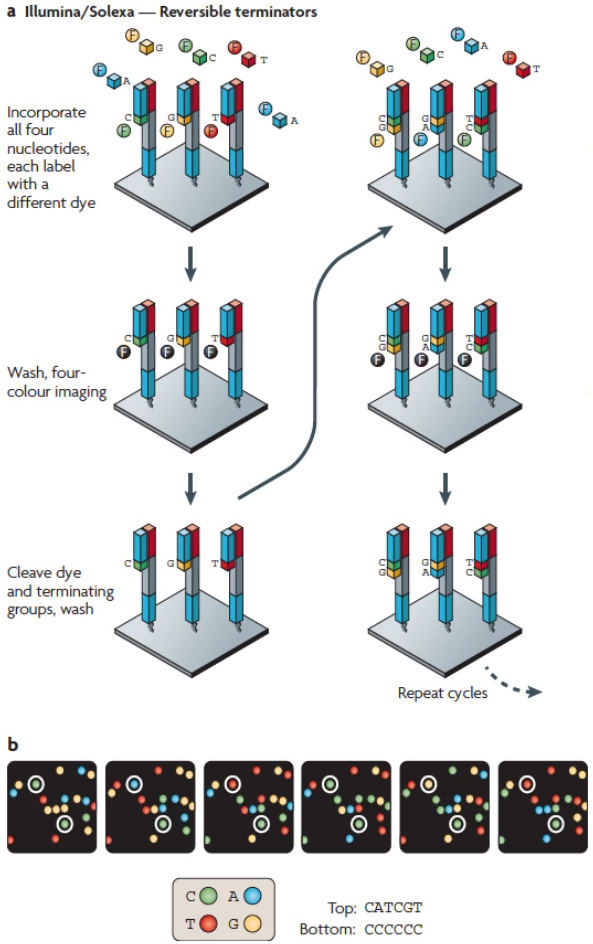
- First Generation Sequencing
  - Sanger Sequencing
- Second Generation Sequencing (*PCR Needed*)
  - Illumina: MiSeq & HiSeq
  - Roche: 454
  - Life Science: IONtorrent & IONproton
  - ABI: SOLiD
- Third Generation Sequencing (*Single Molecule Sequencing*)
  - Helicos Biosciences: Heliscope
  - Pacific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

# Illumina Sequencing



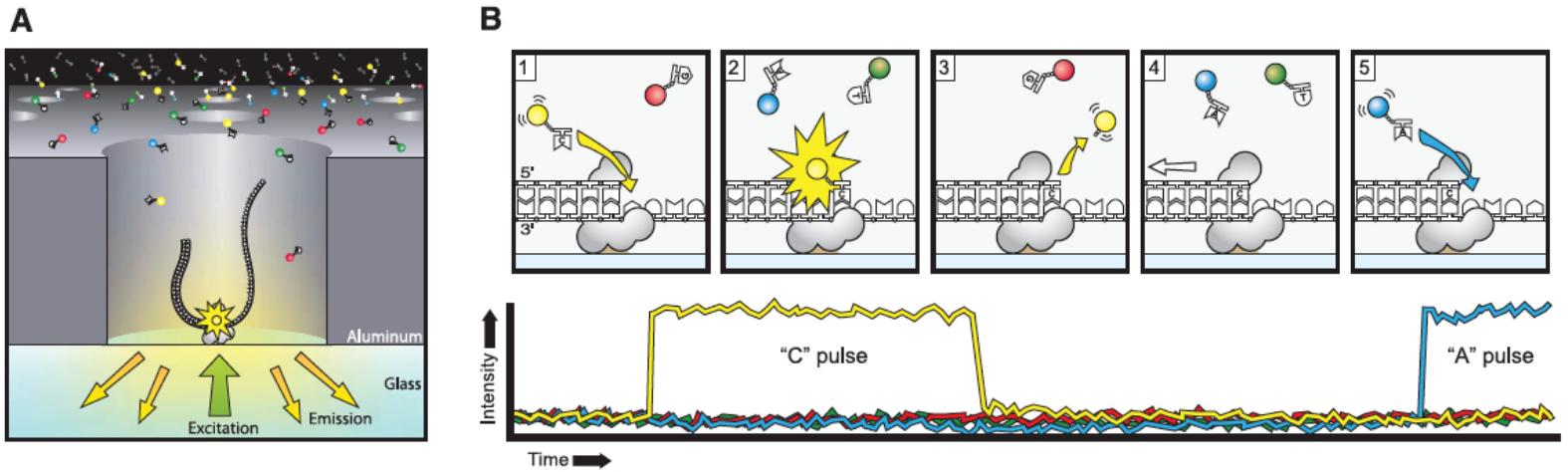

Metzker 2010

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




Metzker 2010

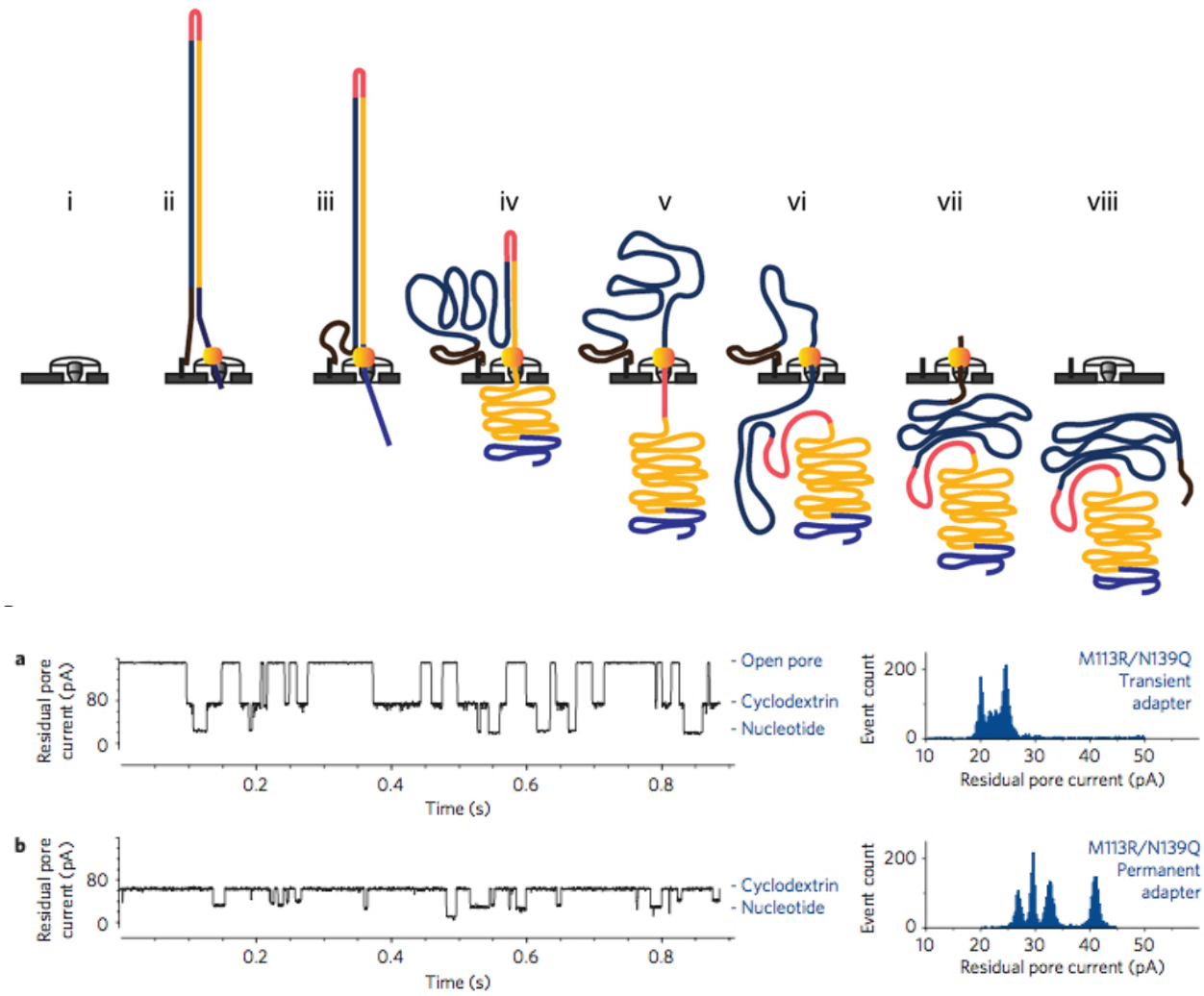
# Pacific Biosciences




Eid et al. 2009

## Oxford Nanopore

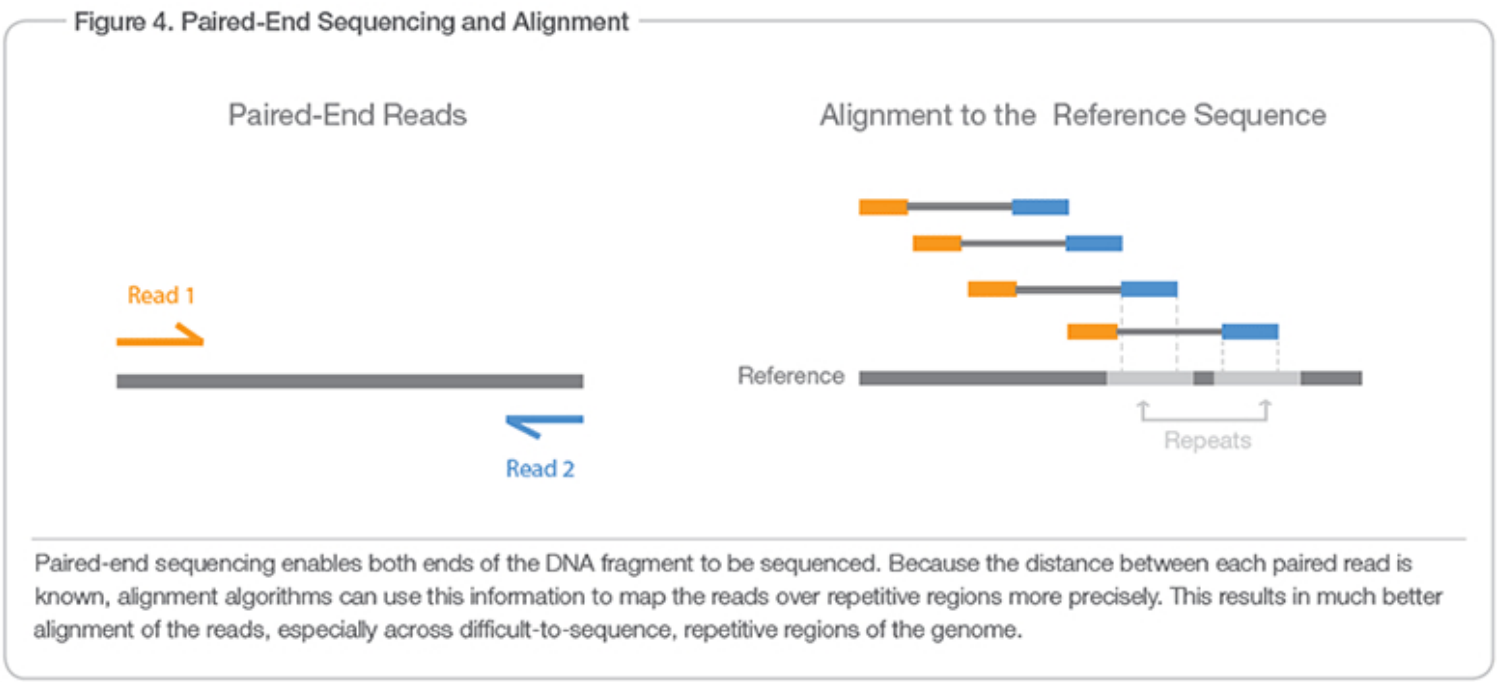
# Oxford Nanopore




Jain et al. 2014; Clarke et al. 2009

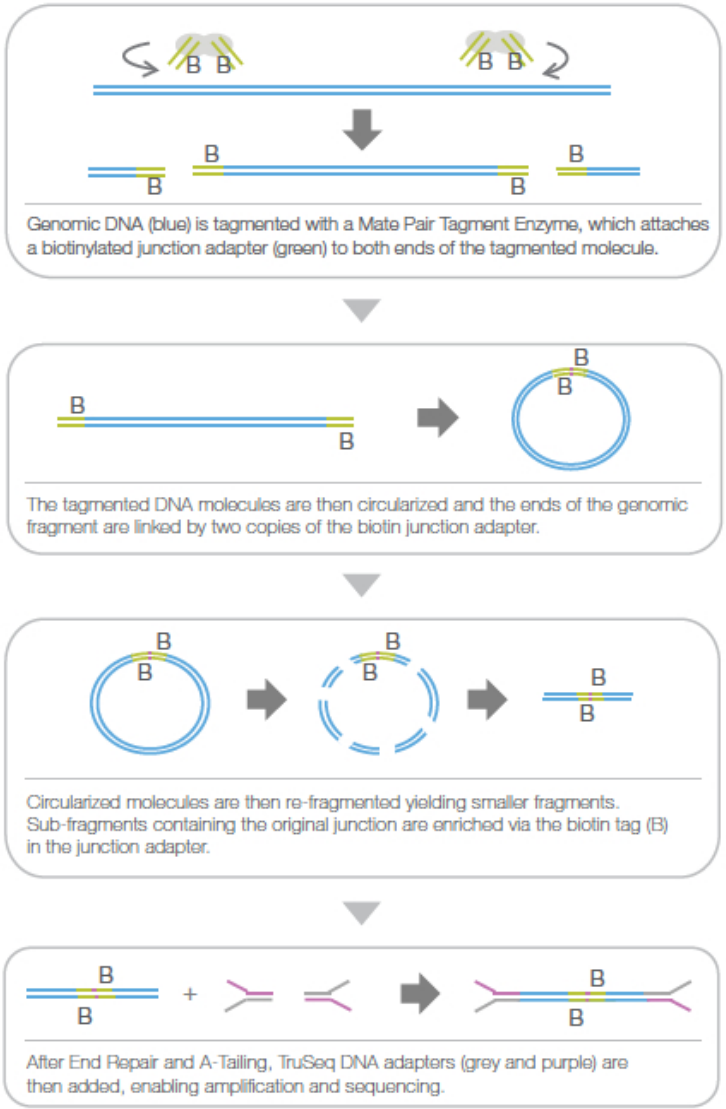
- First Generation Sequencing
  - Sanger Sequencing
- Second Generation Sequencing (*PCR Needed*)
  - Illumina: MiSeq & HiSeq
  - Roche: 454
  - Life Science: IONtorrent & IONproton
  - ABI: SOLiD
- Third Generation Sequencing (*Single Molecule Sequencing*)
  - Helicos Biosciences: Heliscope
  - Pacific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

## Illumina Paired-end Sequencing Libraries




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** |

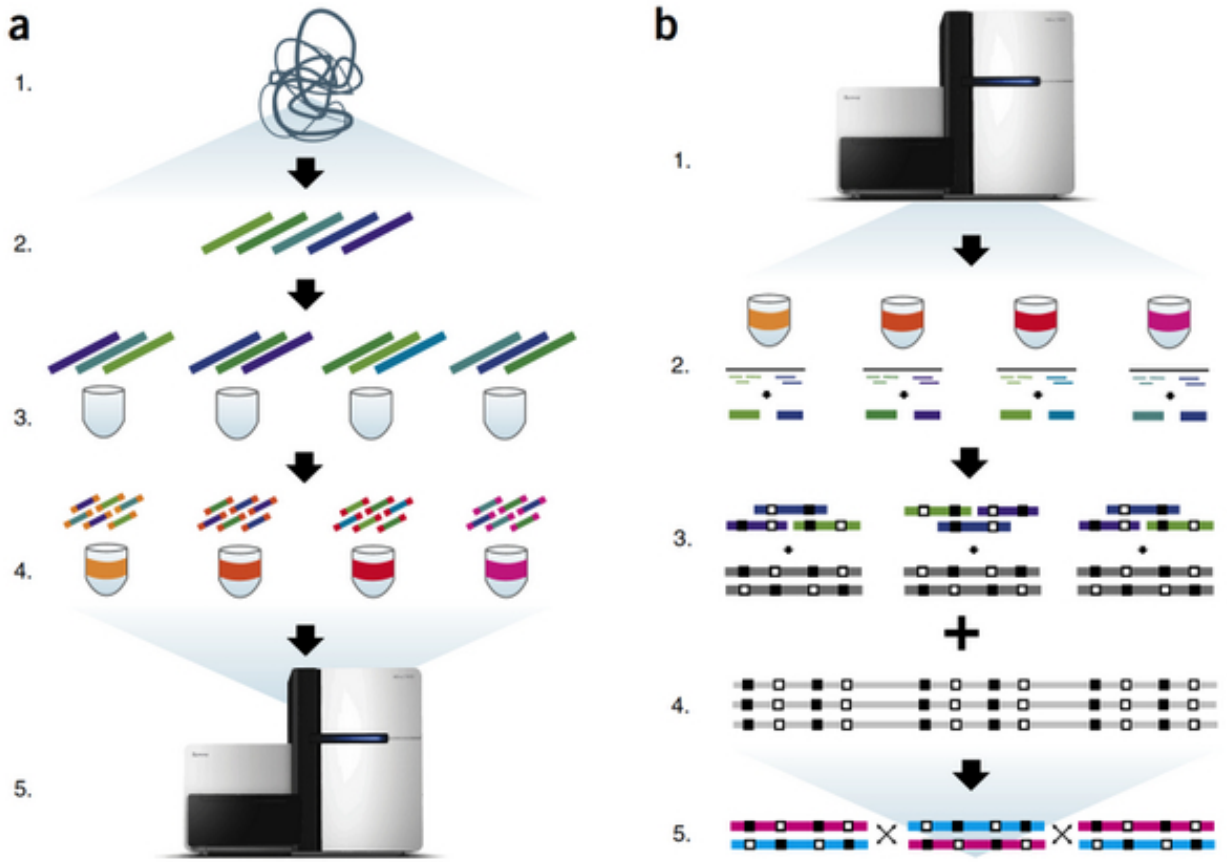
Custom Recipe for 1.3Gb Bird Genome (<15% Repeats)

- 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)

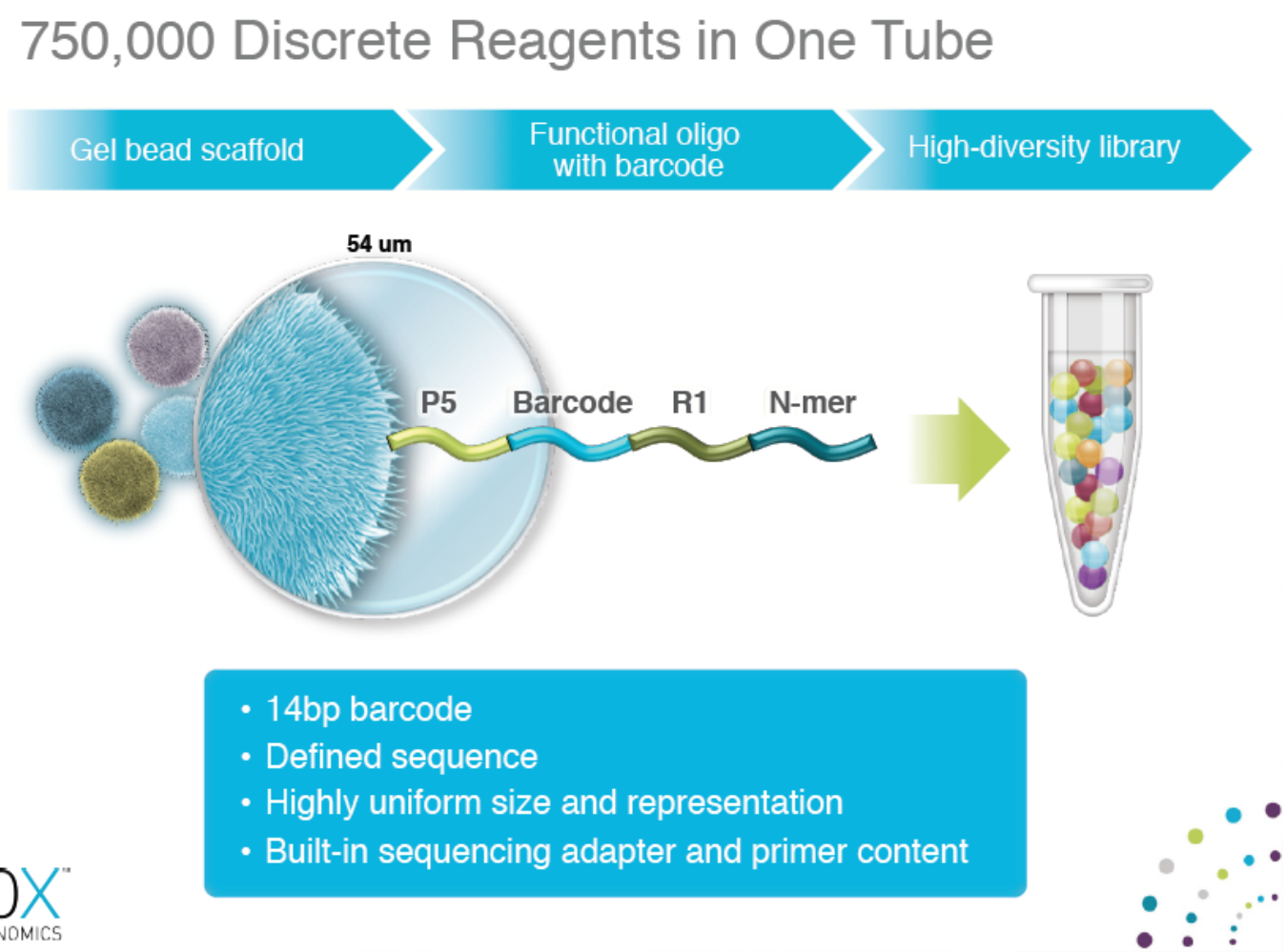
- 180bp PE - 4 lanes HiSeq (50x coverage)
- 3kb MP - 4 lanes HiSeq (40x coverage)
- 8kb/15kb MP - 1 lane (10x coverage)

## TruSeq Synthetic Long-Read Libraries (Moleculo)

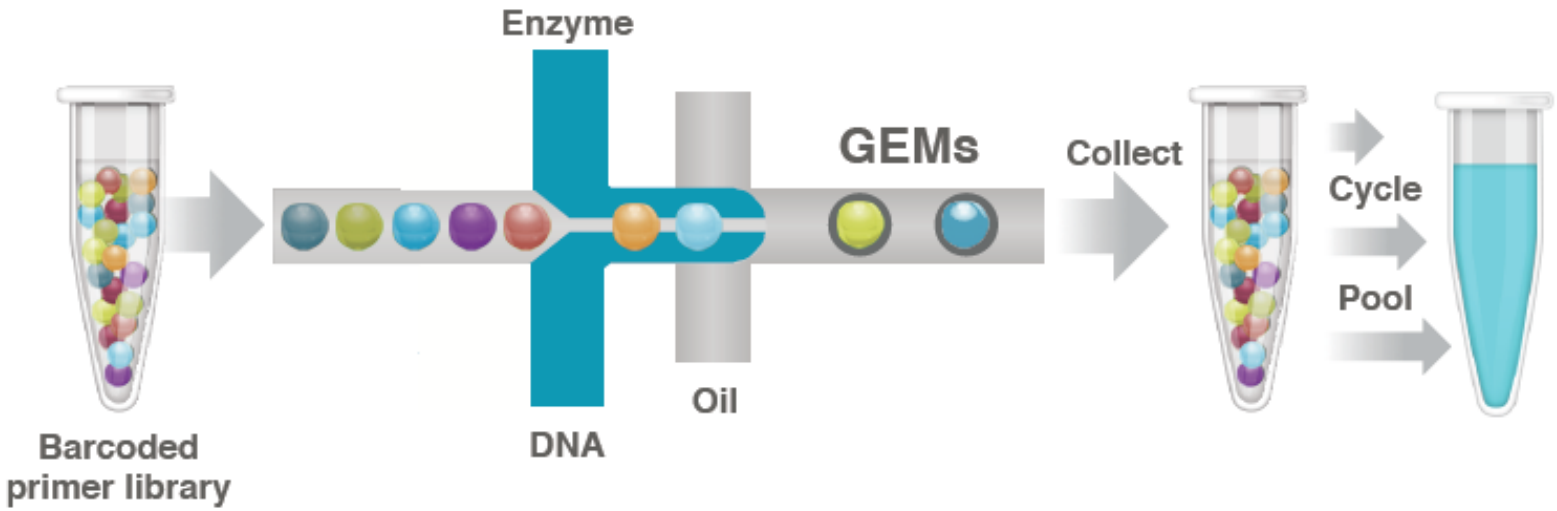



Nextera Kit Manual

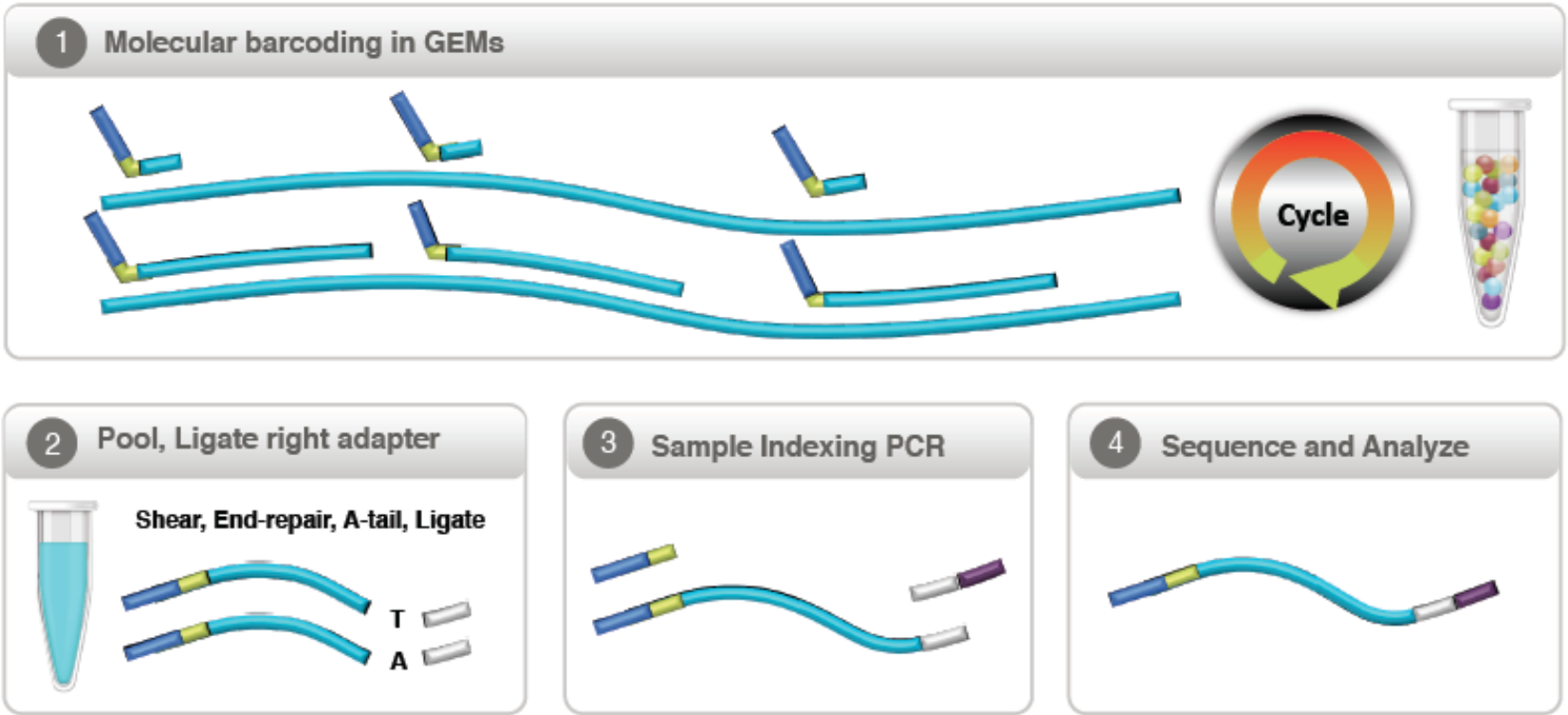
# 10X linked Long-Read Libraries



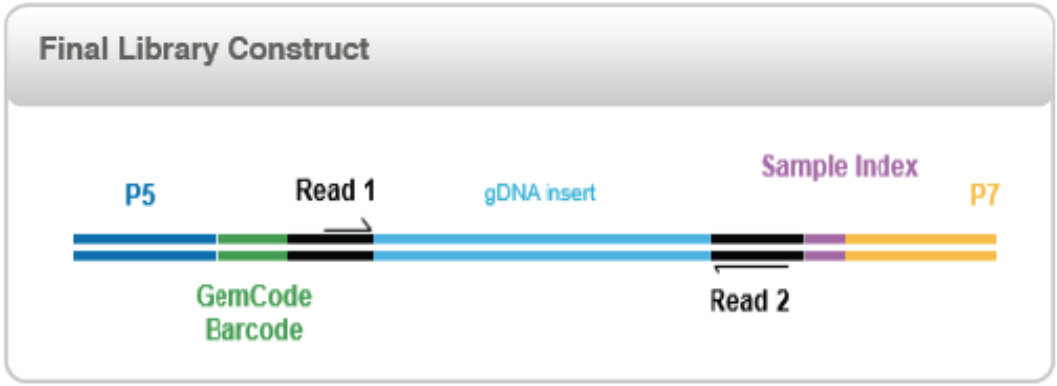
## 10X linked Long-Read Libraries

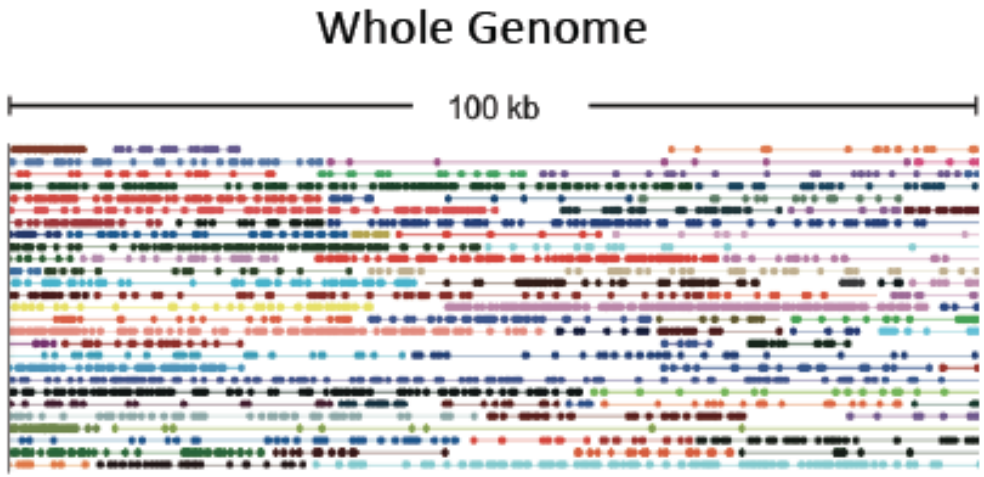


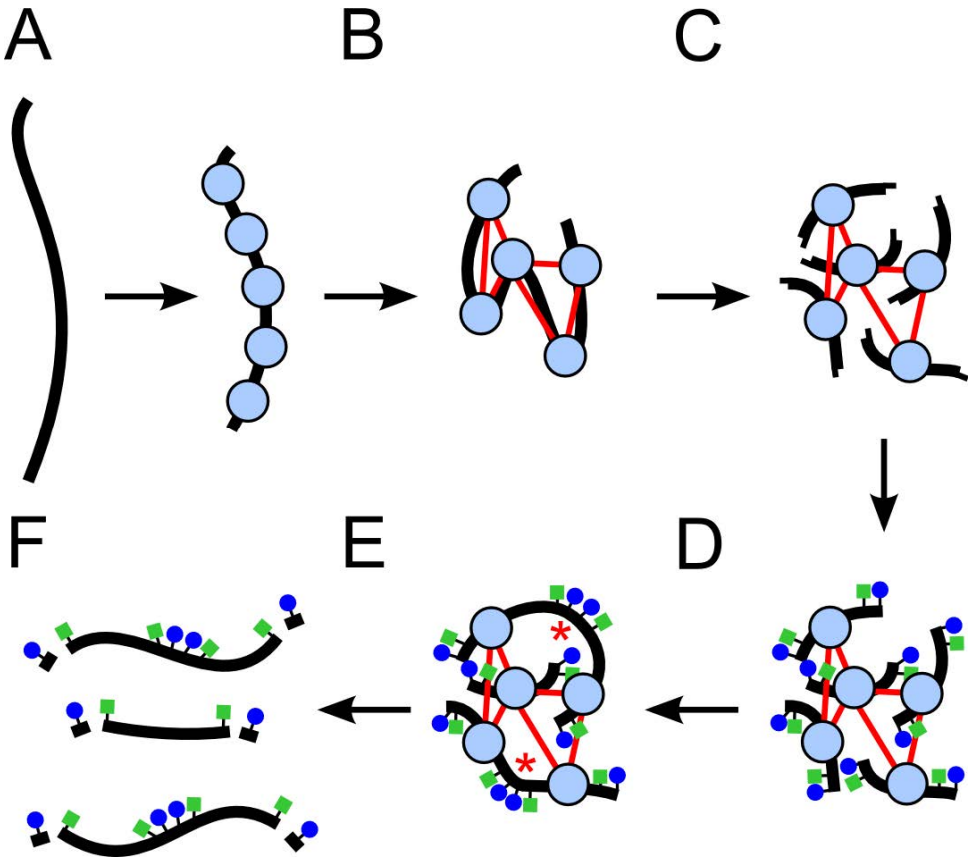

10X linked Long-Read Libraries



# 10X linked Long-Read Libraries






Putnam et al. 2015 (ArXiv)

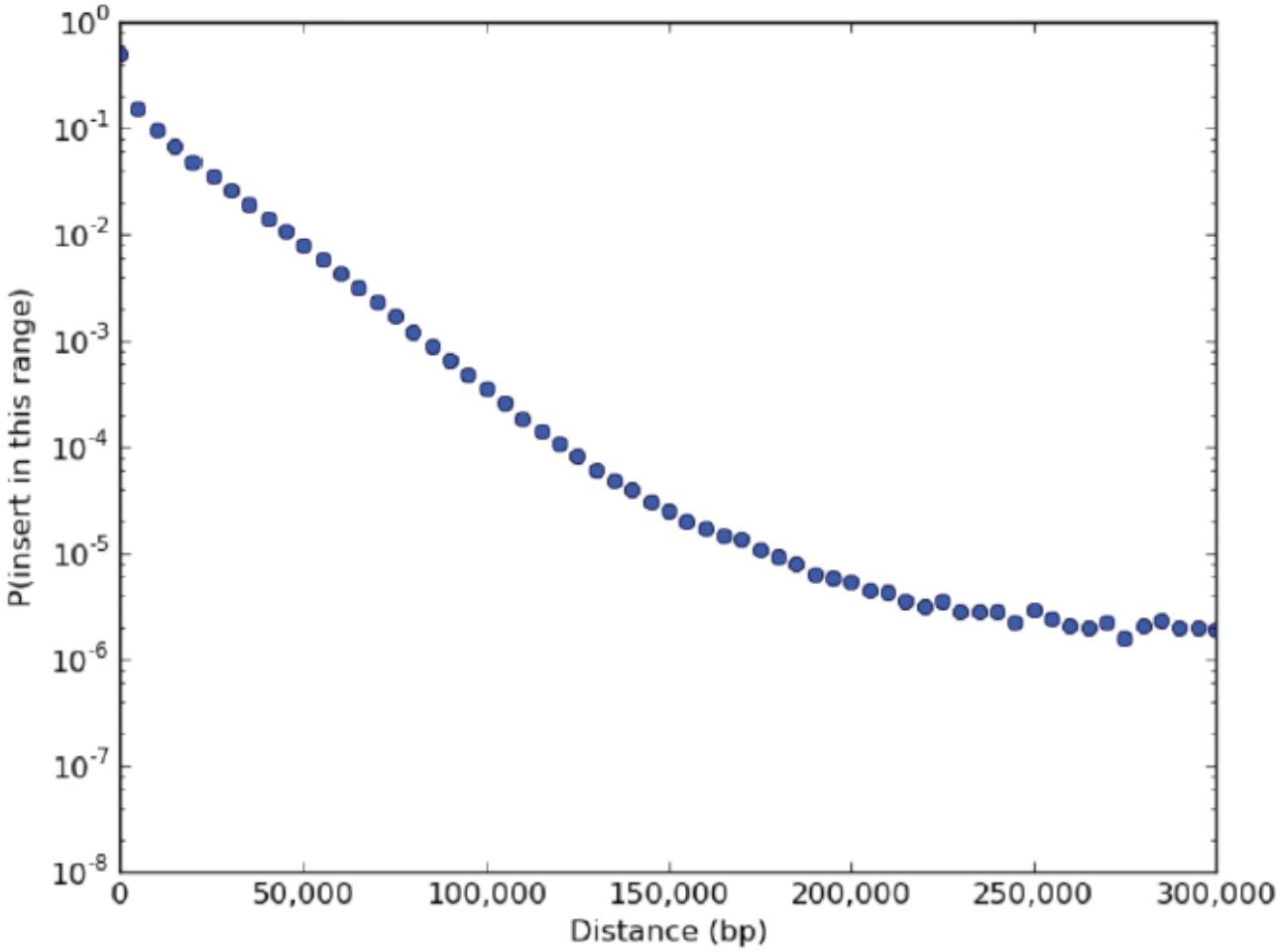
# Tuatara Assembly Statistics

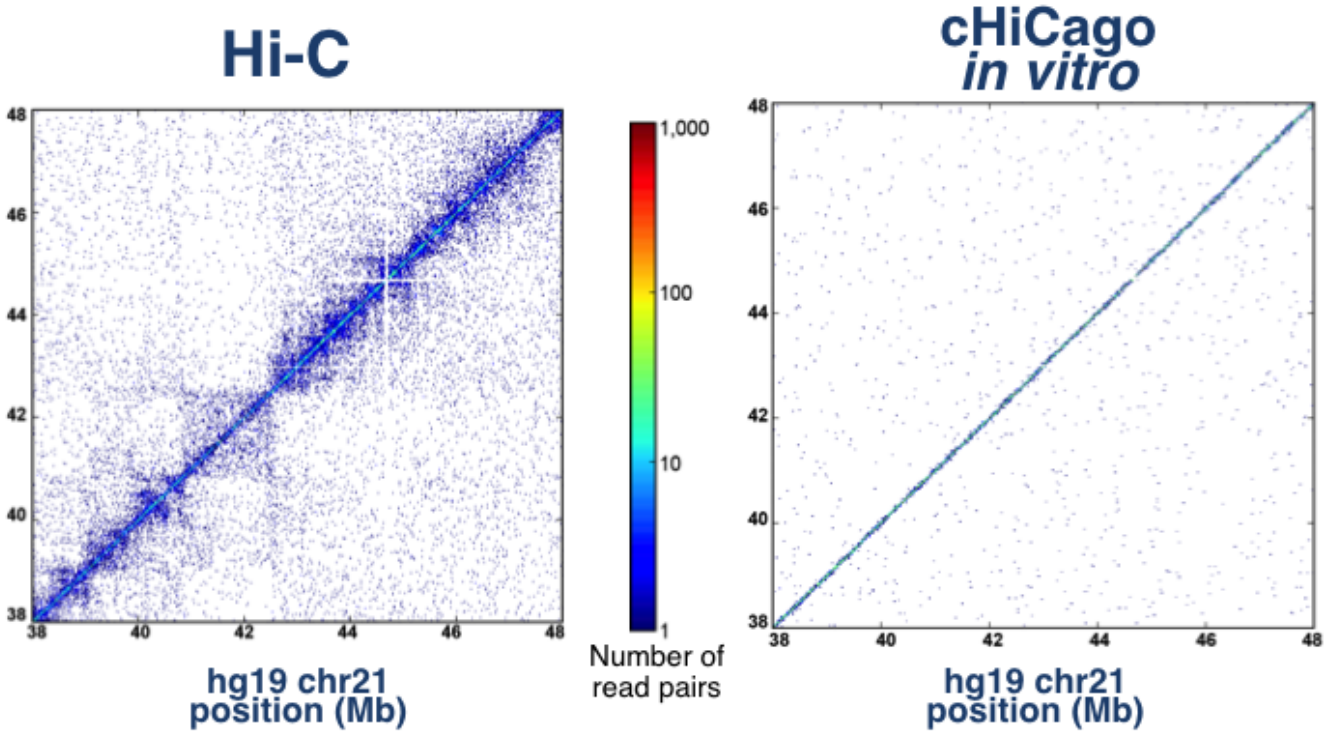
| | Starting Assembly | Final Assembly | Fold Increase |
|---|---|---|---|
| **Genome Size** | 4.2691 Gb | 4.2718 Gb | - |
| **N50** | 348 Kb | 2.23 Mb | 6.4X |
| **N90** | 69 Kb | 438 Kb | 6.3X |

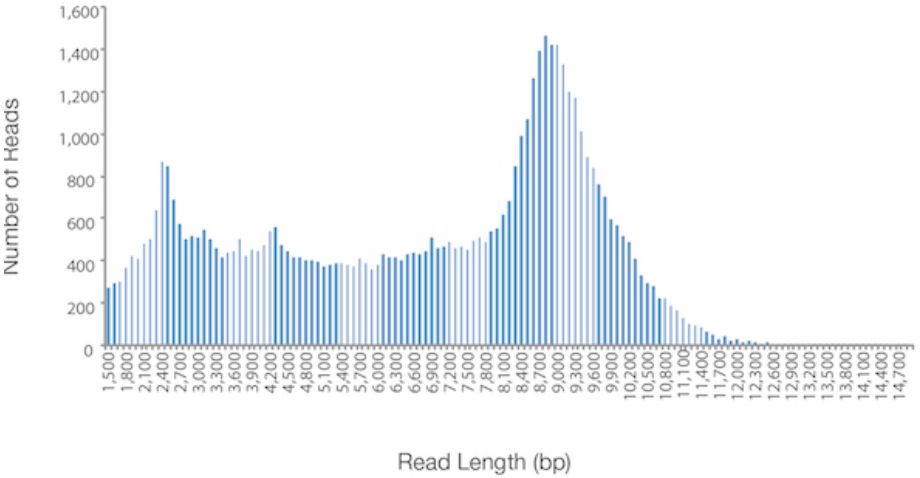
Estimated Dovetail coverage (1-50 Kb pairs): 43.5X

| | 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 |

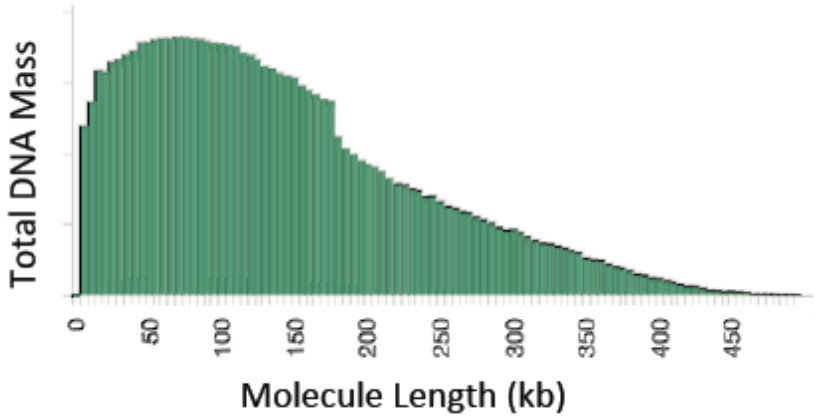
# Dovetail Library Insert Distributions



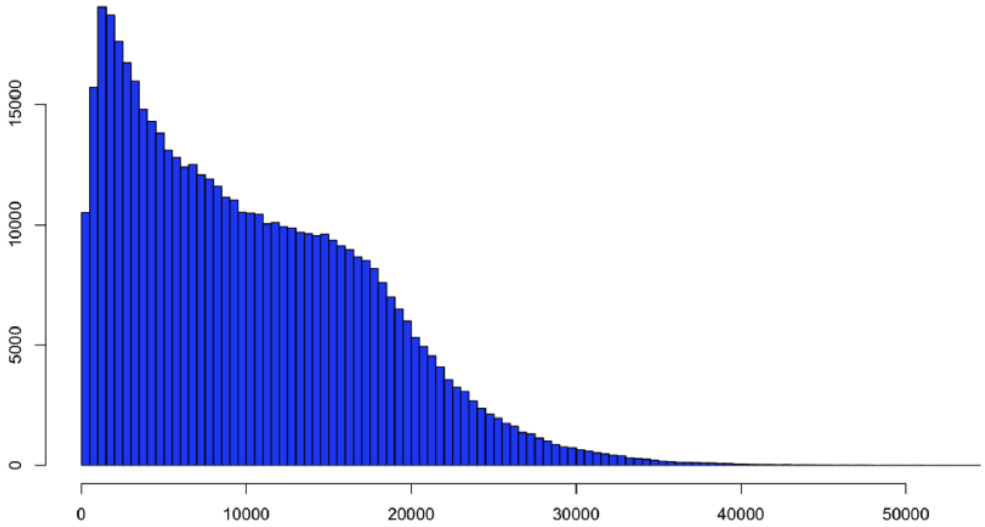

Hi-C
48
46
44
42
40
38
38
40
42
44
46
48
hg19 chr21
position (Mb)
1,000
100
10
1
Number of
read pairs
cHiCago
in vitro
48
46
44
42
40
38
38
40
42
44
46
48
hg19 chr21
position (Mb)



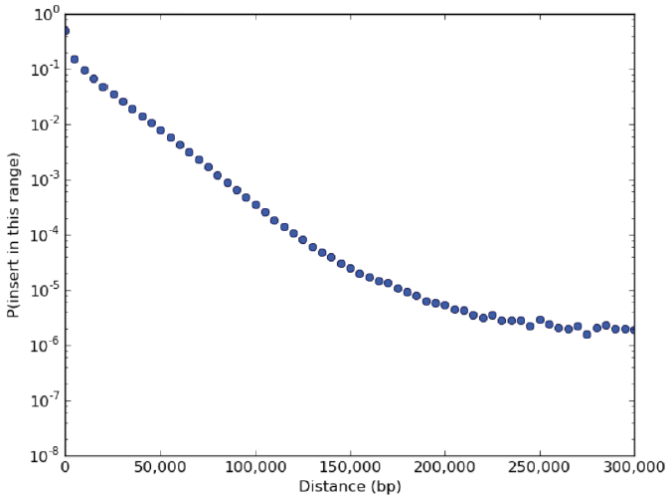

Moleculo




10X Genomics

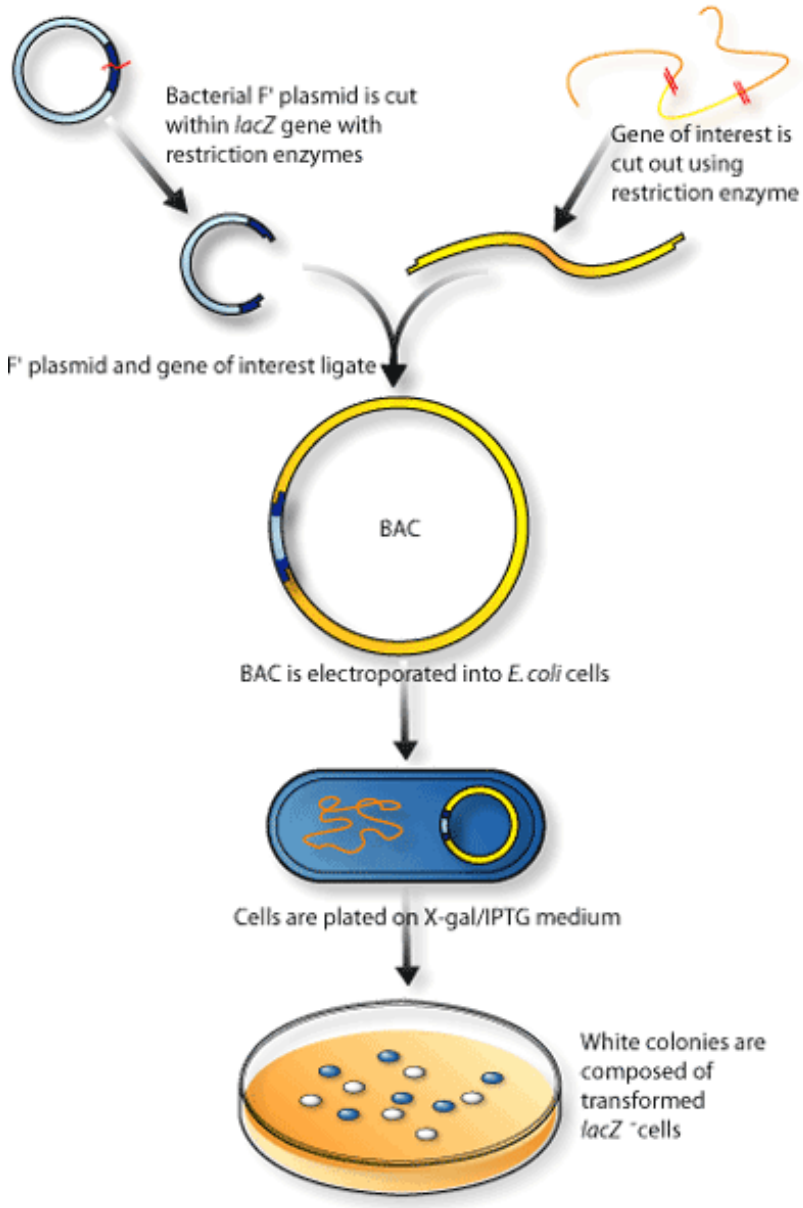


PacBio




Dovetail Genomics

- 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(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

## 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.........................
 ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./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
- Preqc

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

G = Genome Size

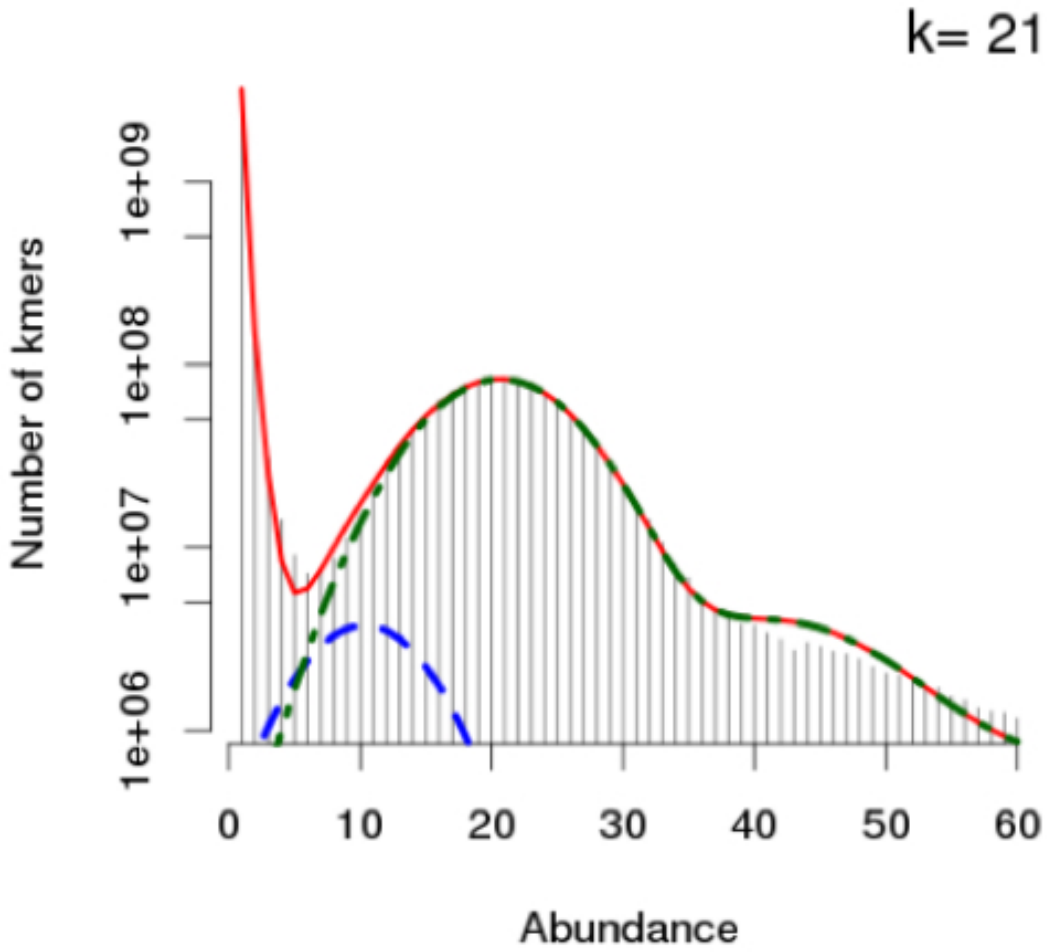
pn = proportion of correct reads

l = read length

k = kmer length

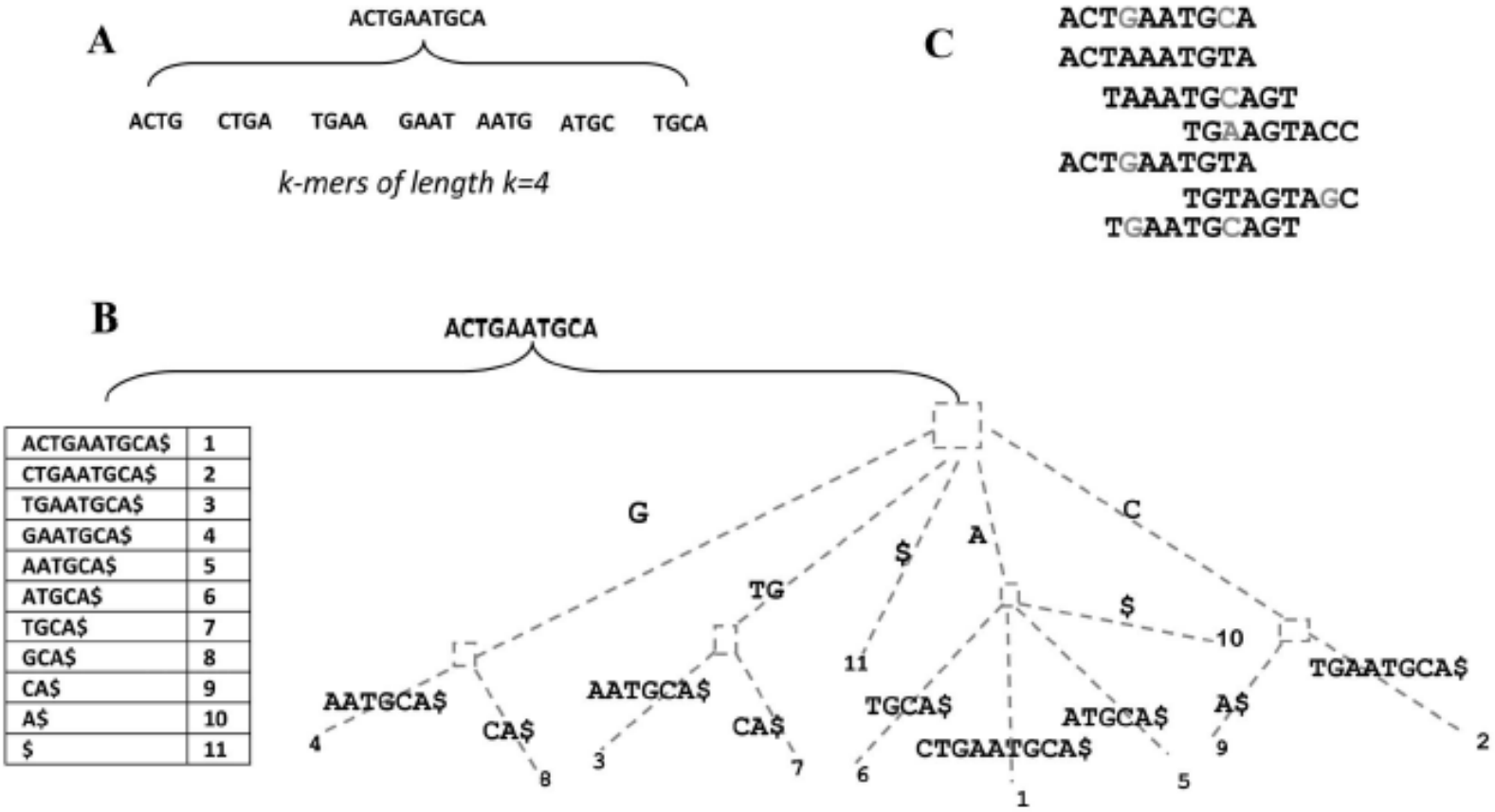
$\lambda_k$ = mode of the k-mer count histogram

Simpson 2013, arXiv


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

- 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

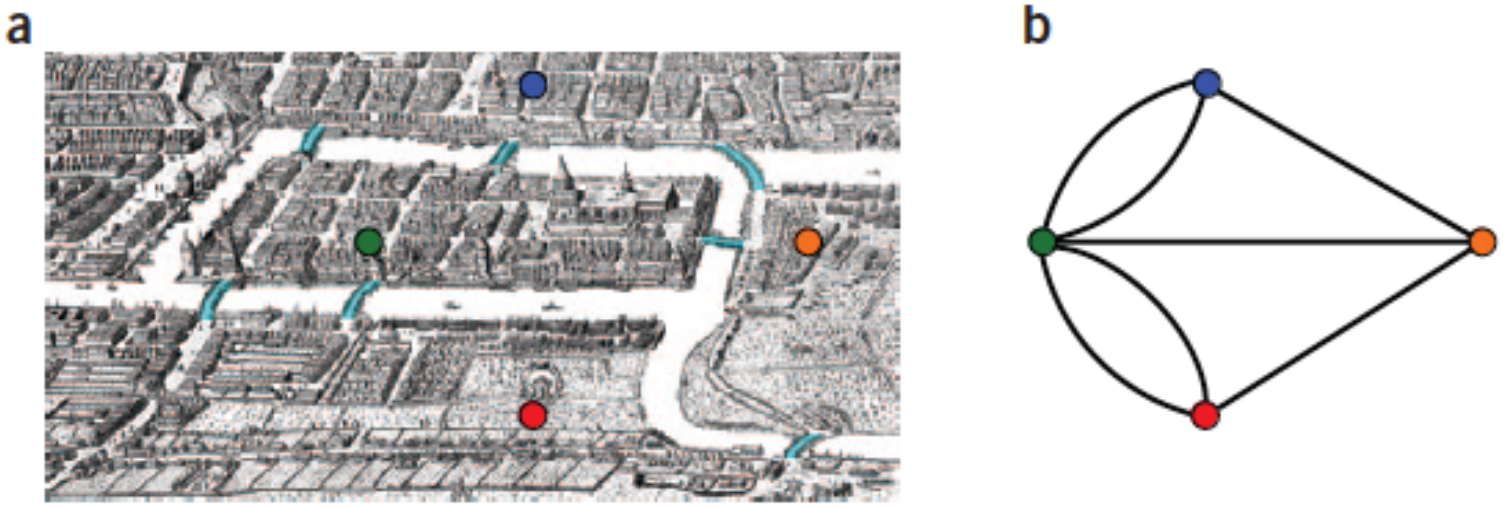
## Error Correction Strategies



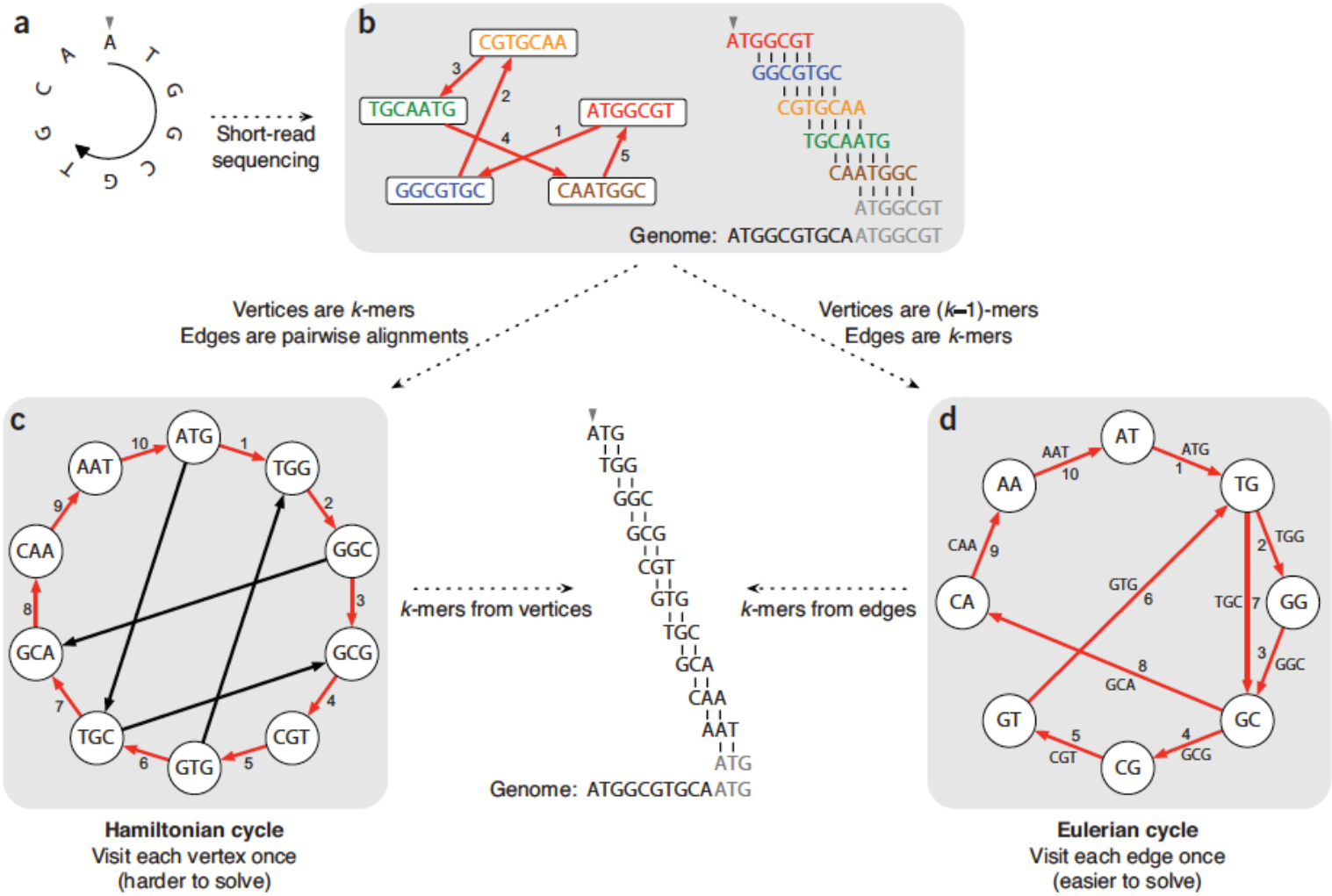

El-Metwally et al. 2013

- 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

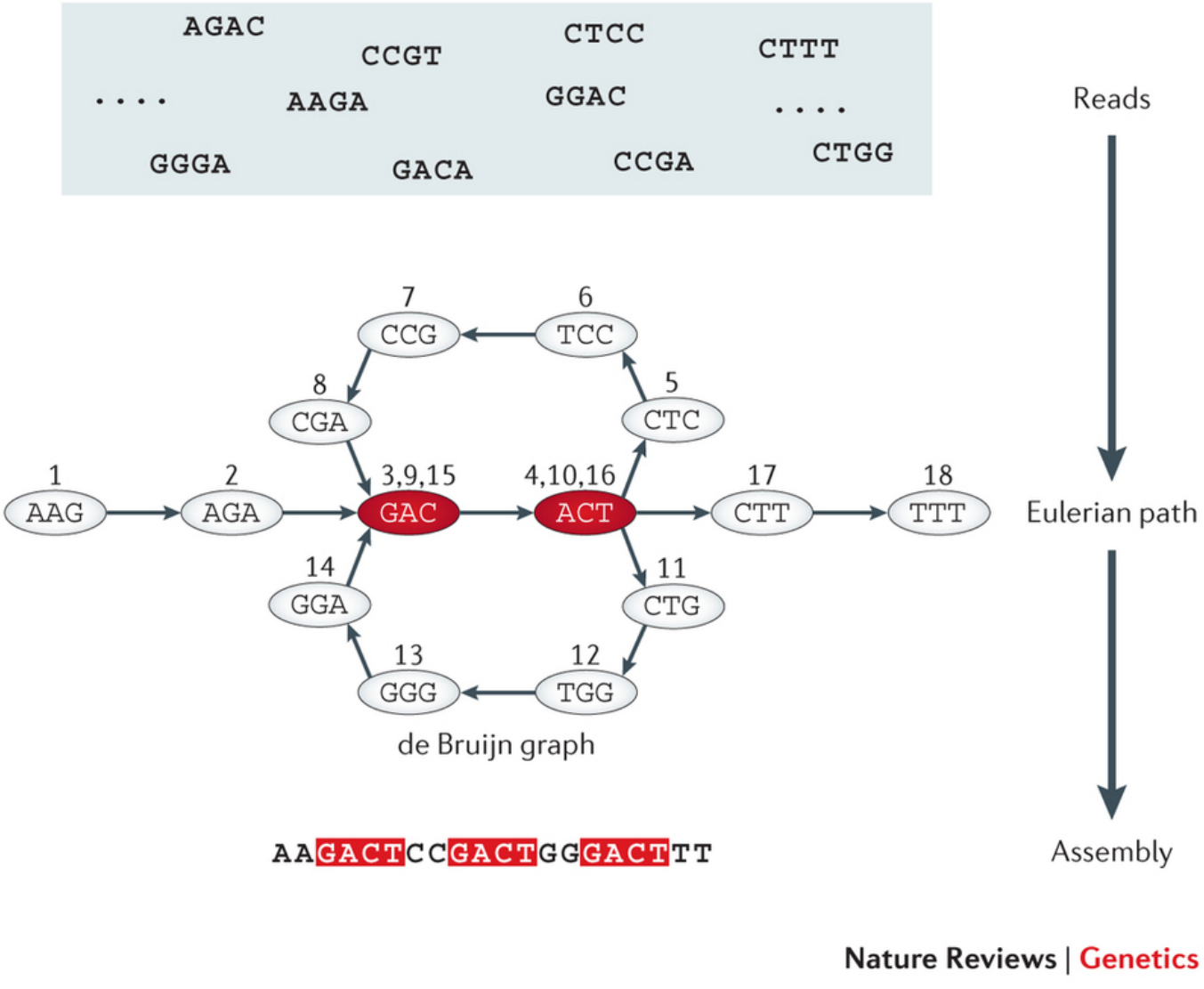
Bridges of Koenigsberg problem (Graph Theory by Euler)




Compeau et al. 2011

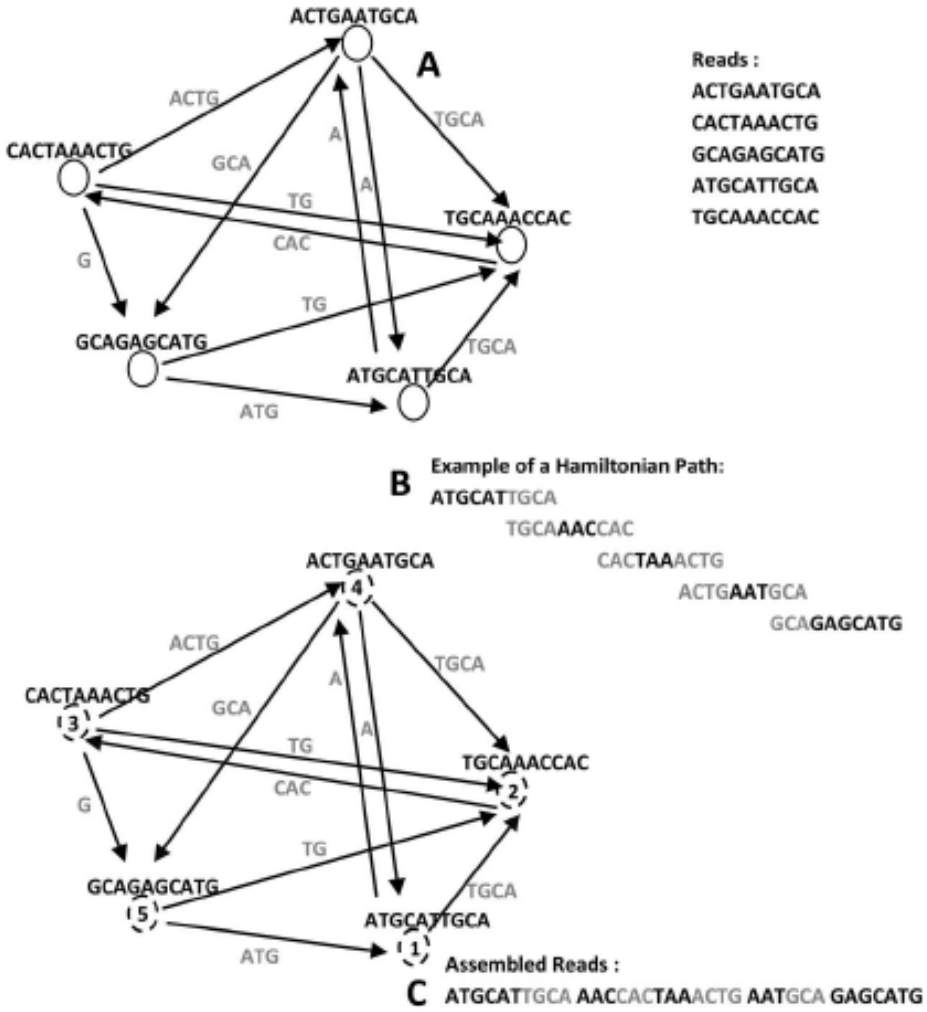



Compeau et al. 2011

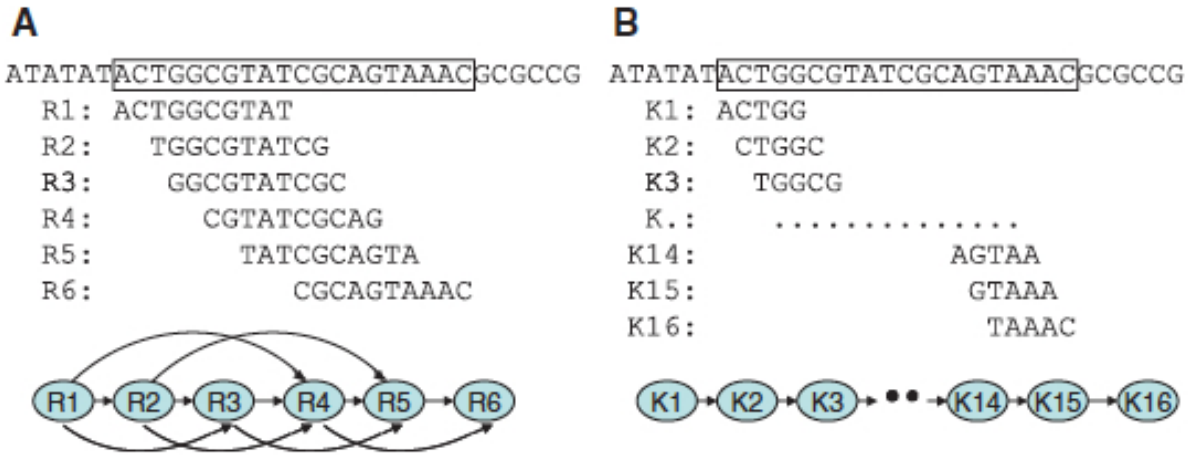



Berger et al. 2013

## Overlap-layout-consensus Graph



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




Li et al. 2011, Briefings in Functional Genomcis

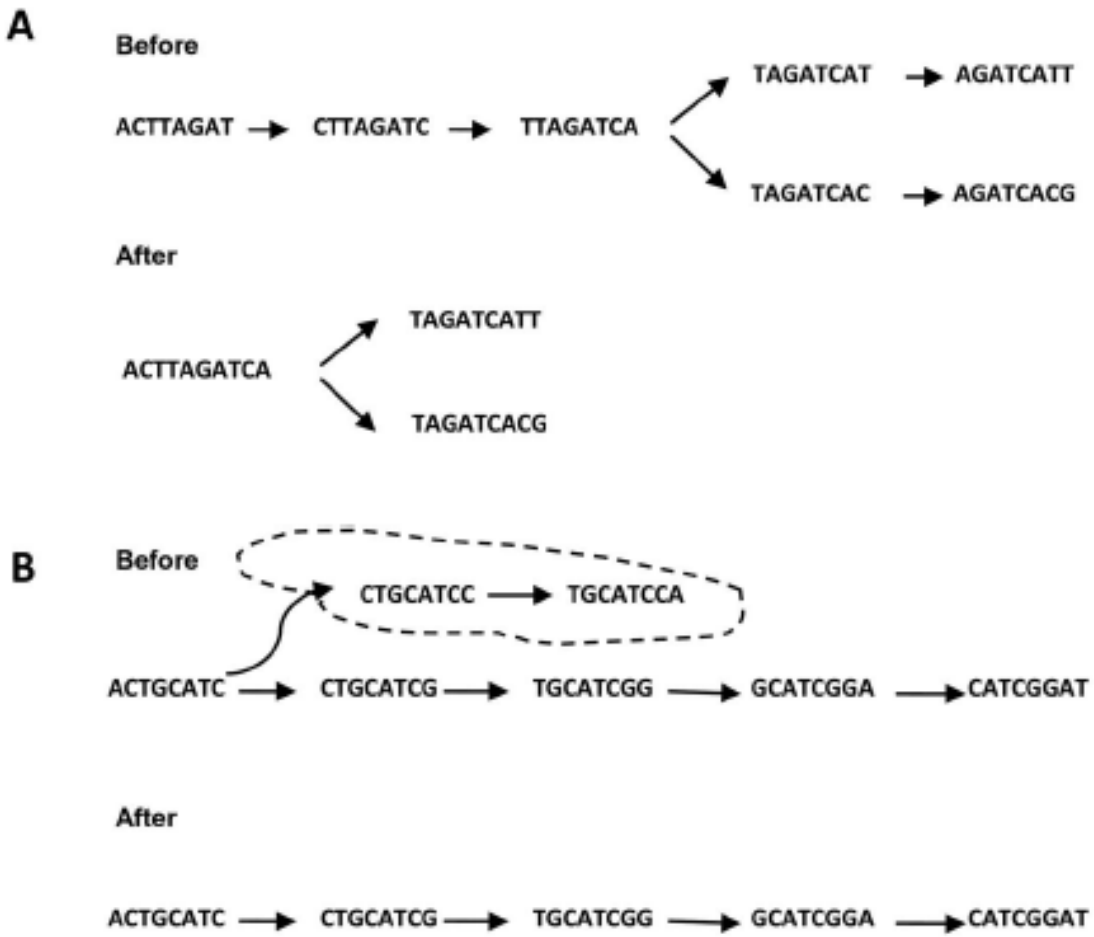
## Overlap-layout-consensus based Tools:

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

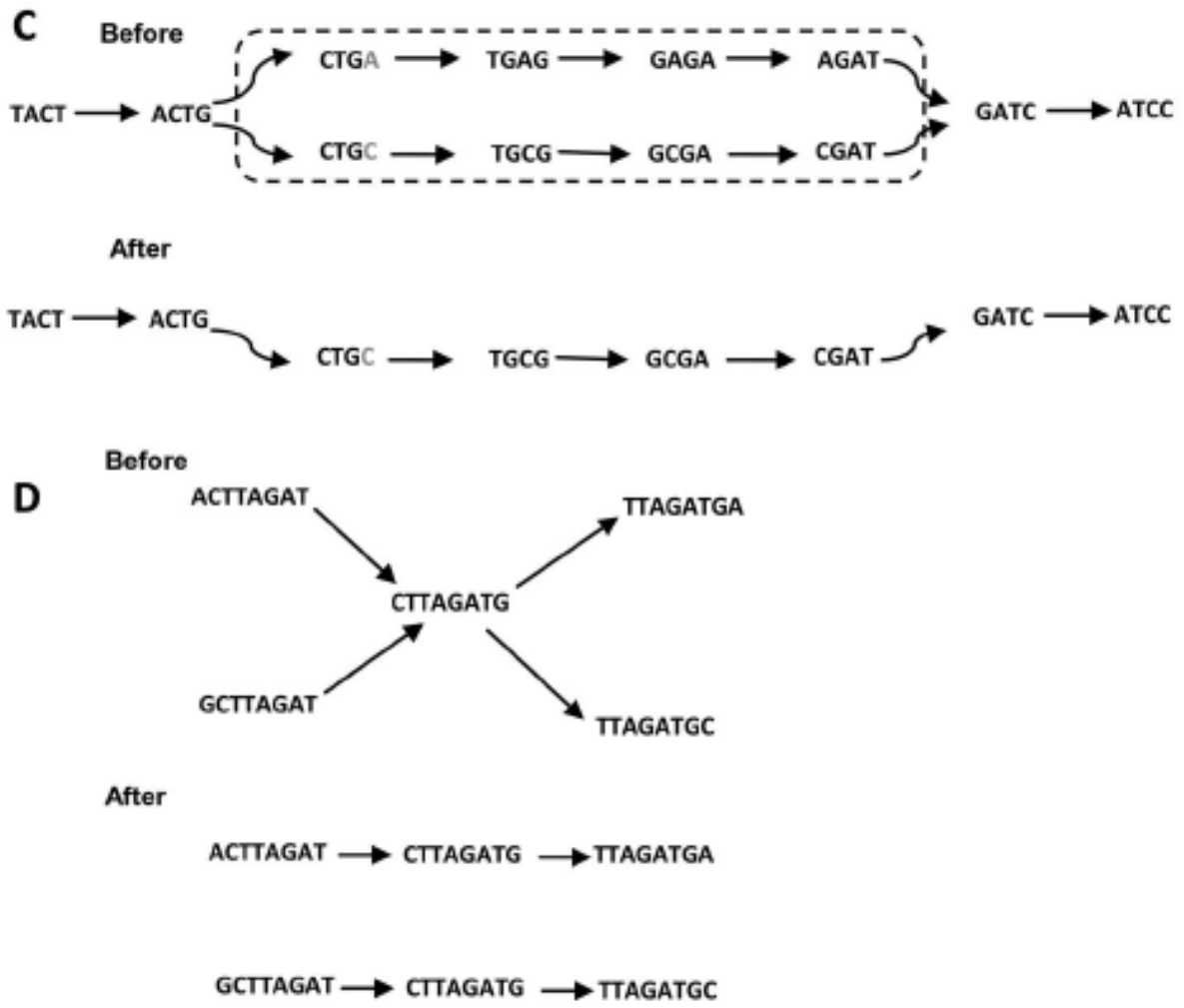
## *de Bruijn* graph based Tools:

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

## Graph Simplification



## Graph Simplification



**Short Read Assembler**

- *De Bruijn* Graph Assembler
  - Allpaths-LG (Gnerre et al. 2011)
  - Abyss (Simpson et al. 2009)
  - 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)

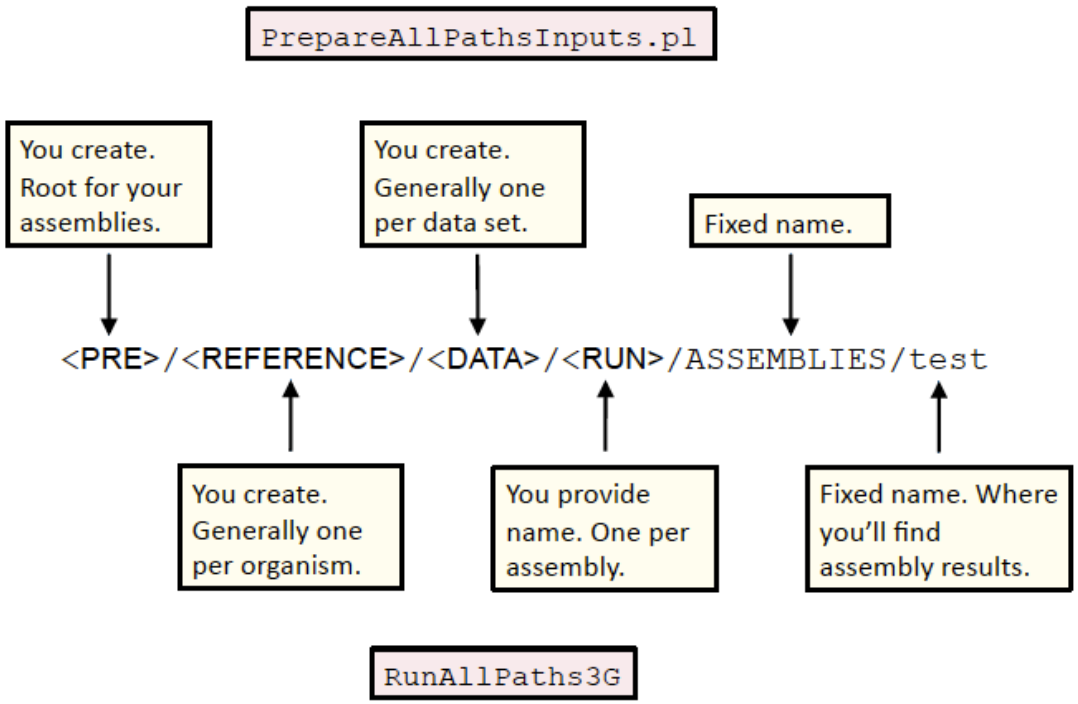
**Long Read Assembler**

- 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)

- Input Data
  - BAM, FASTQ, FASTA/QUALA
- Input Files
  - in_libs.csv
  - in_groups.csv
  - ploidy.txt




Schatz Presentation

```
PrepareAllPathsInputs.pl
    IN_GROUPS_CSV=<in groups file>
    IN_LIBS_CSV=<in libs file>
    DATA_DIR=<full path of data directory>
    PLOIDY=<ploidy, either 1 or 2>
    PICARD_TOOLS_DIR=<picard tools directory>
    HOSTS=<list of hosts to be used in parallel>
```

```
RunAllPathsLG                    \
   PRE=<prefix path>               \
   REFERENCE_NAME=<reference dir>   \
   DATA_SUBDIR=<data dir>            \
   RUN=<run dir>
```

Schatz Presentation

- Output Files
  - final.assembly.fasta
  - final.assembly.efasta
  - final.contigs.fasta
  - final.contigs.efasta
  - final.summary
  - final.rings
  - final.superb
  - assembly_stats.report
  - library_coverage.report
  - assembly.report

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

Config file:

```
max_rd_len=100
[LIB]
avg_ins=450
reverse_seq=0
asm_flags=3
rd_len_cutoff=100
rank=1
pair_num_cutoff=3
map_len=32
q1=/scratch/stefan/bop/lyco_soap/1_140430_BC4682ANXX_P974_101_1.fastq.gz
q2=/scratch/stefan/bop/lyco_soap/1_140430_BC4682ANXX_P974_101_2.fastq.gz
[LIB]
avg_ins=5000
reverse_seq=1
asm_flags=3
rd_len_cutoff=100
rank=2
pair_num_cutoff=5
map_len=35
q1=/scratch/stefan/bop/lyco_soap/3_140723_AC49LVACXX_P1326_101_1.fastq.gz
q2=/scratch/stefan/bop/lyco_soap/3_140723_AC49LVACXX_P1326_101_2.fastq.gz
```

Output Files

- lycocorax_soap.scafSeq
- lycocorax_soap.contig
- lycocorax_soap.err
- lycocorax_soap.scafStatistics

## 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'
```

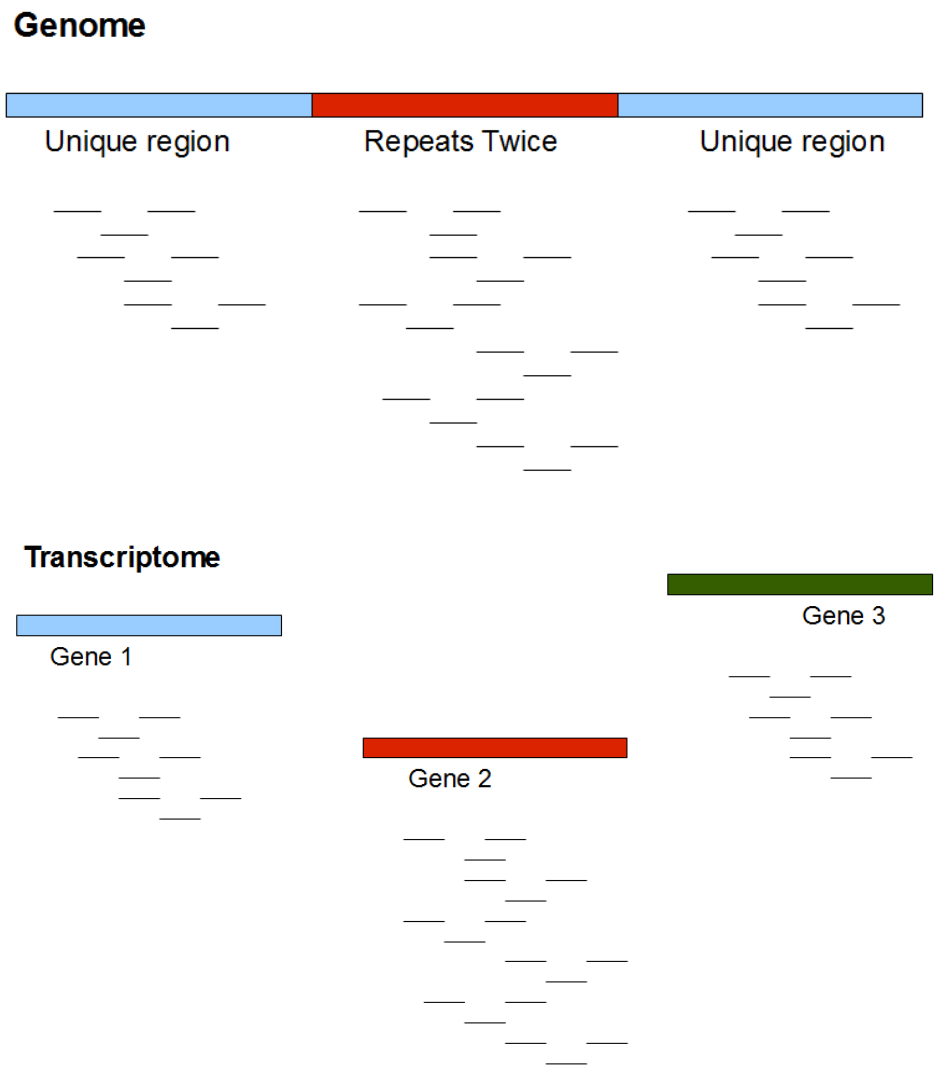
## Output Files

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

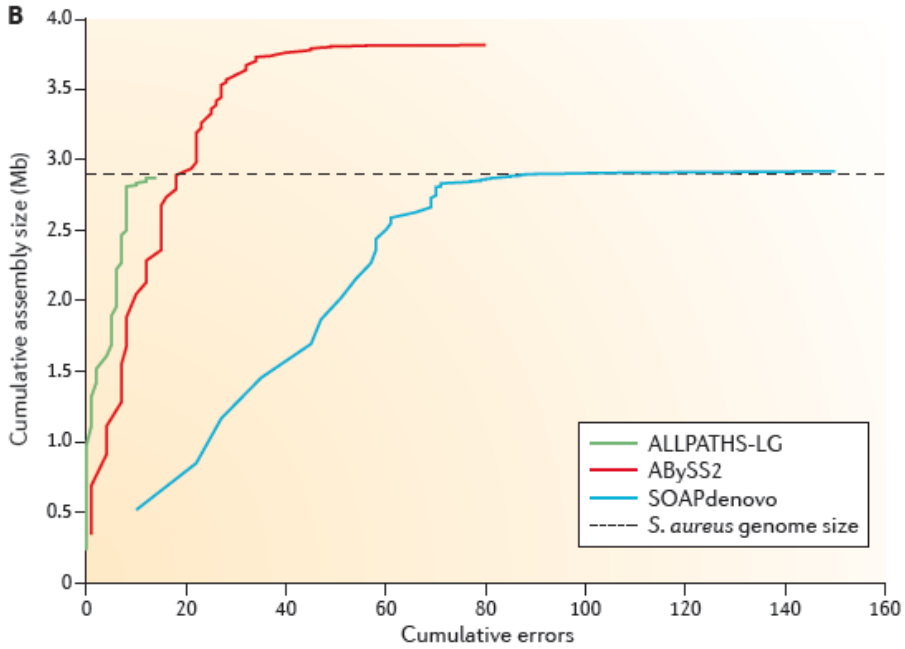
### Get Assembly Stats:

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 |

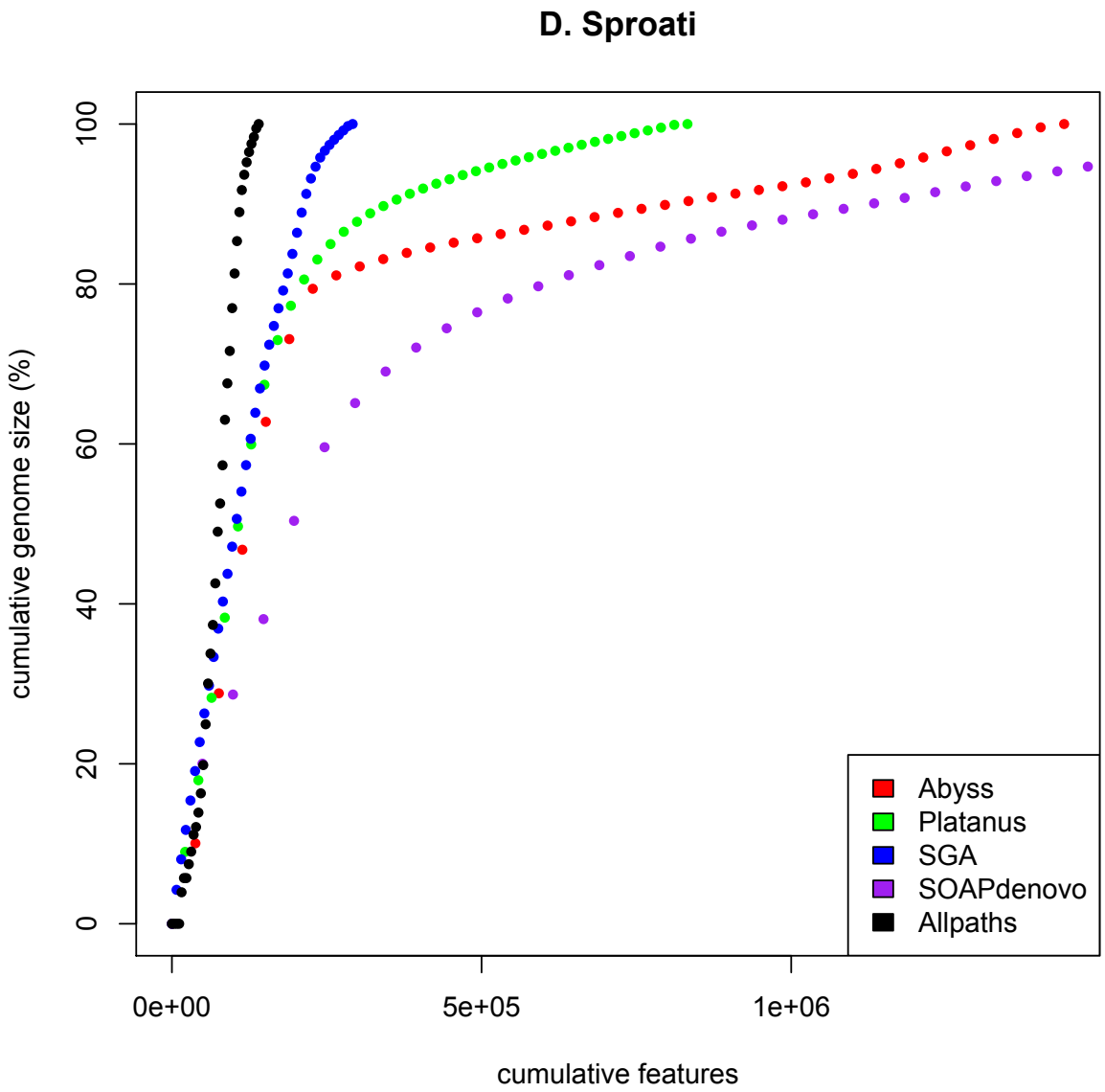

Genome
Unique region
Repeats Twice
Unique region
Transcriptome
Gene 1
Gene 2
Gene 3

- 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
Platanus
SGA
SOAPdenovo
Allpaths

- 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

## 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)

## 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 |

**Gap Filling**

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

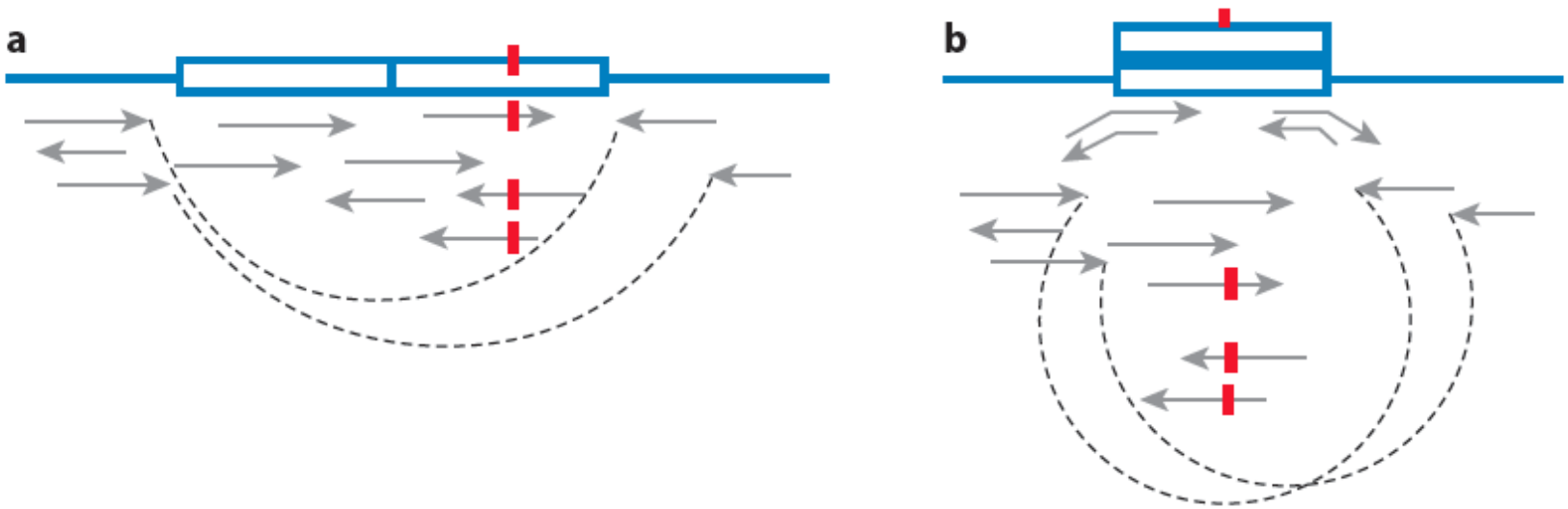
**Resolving Misassemblies**

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

**Genome Merging**

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

# Finding Misassemblies



**Gap Filling**

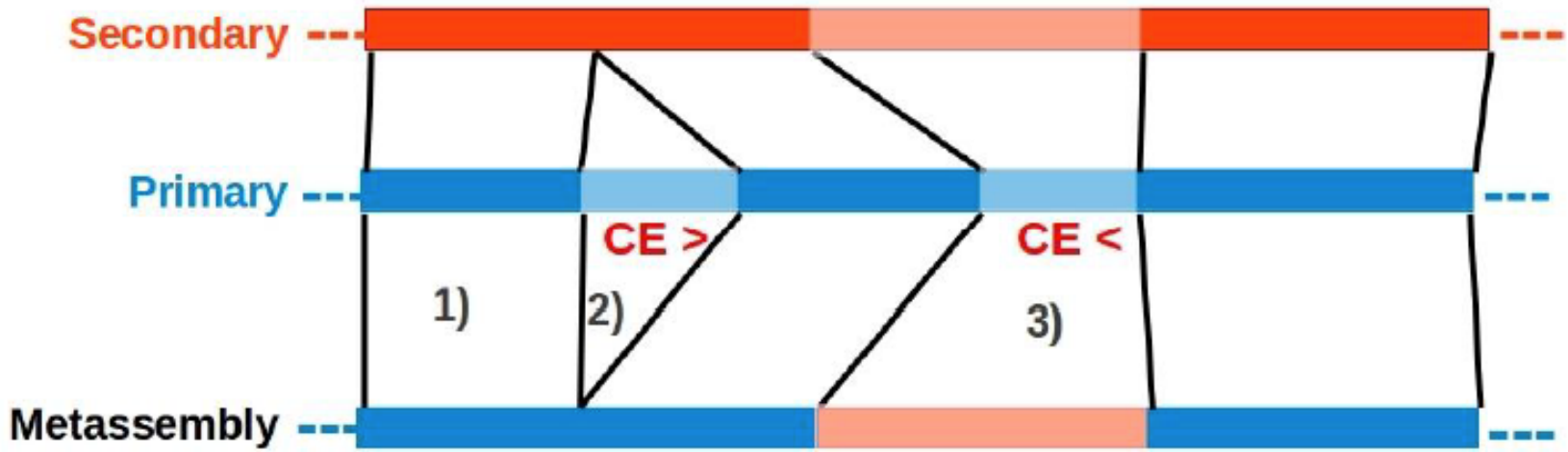
- Sealer (Paulino et al. 2015)
- GapCloser (Luo et al. 2012)
- GapFiller (Nadalin et al. 2011)

**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**

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

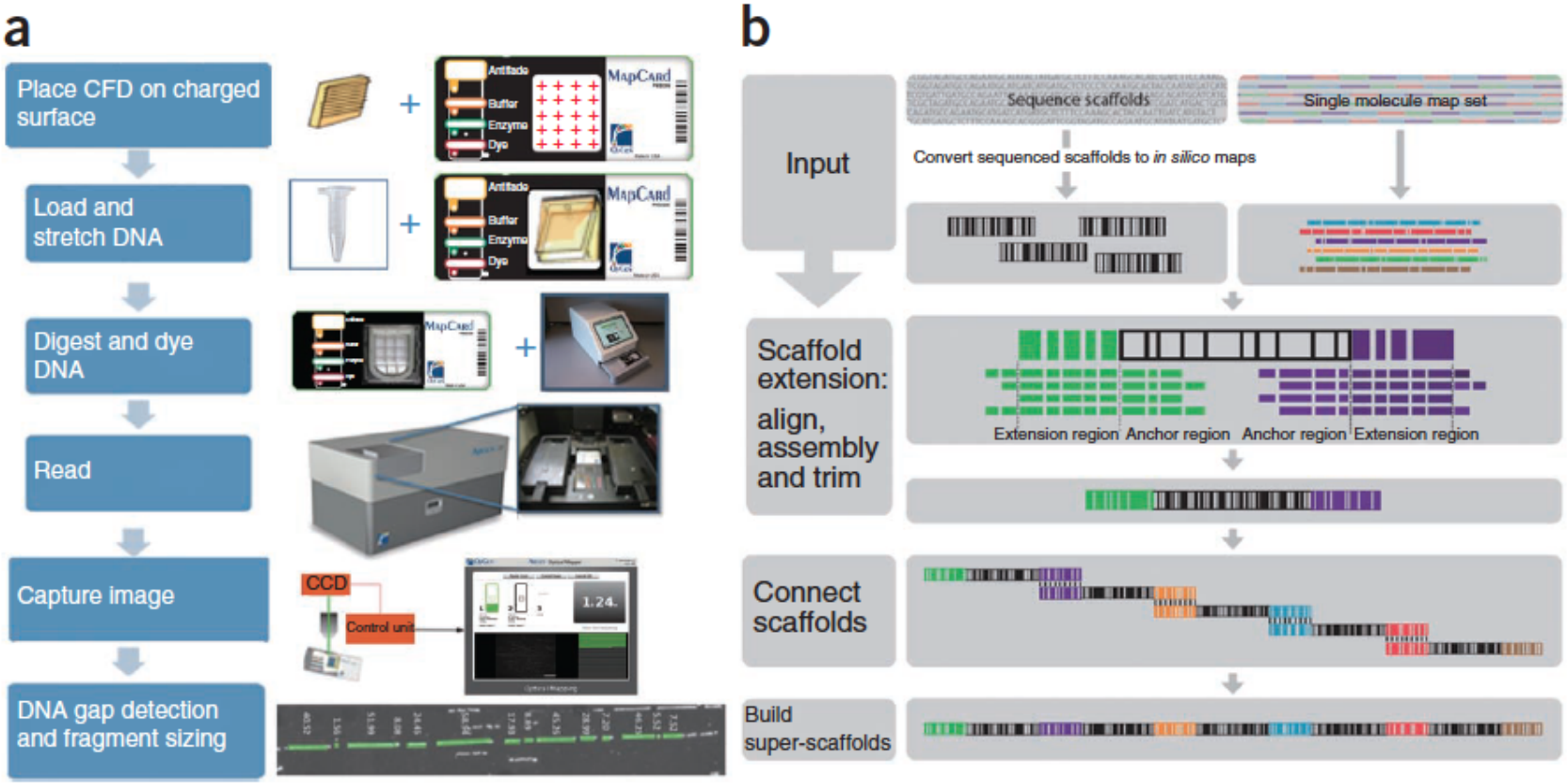
**Resolving Mis-Assemblies**

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

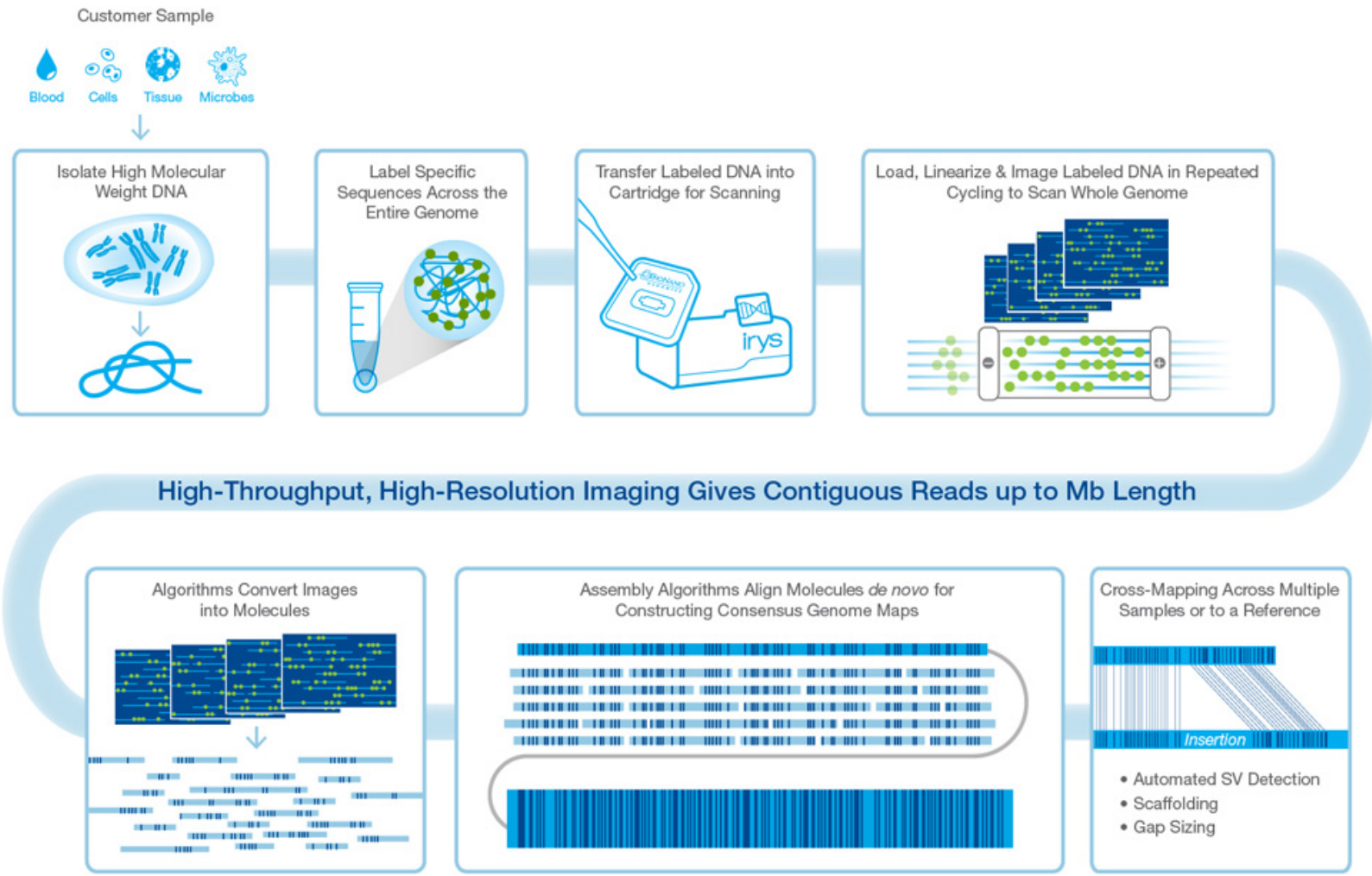
**Genome Merging**

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

# Optical Genome Mapping (OpGen)



# Nanochannel-based Genome Mapping (Bionano Irys)

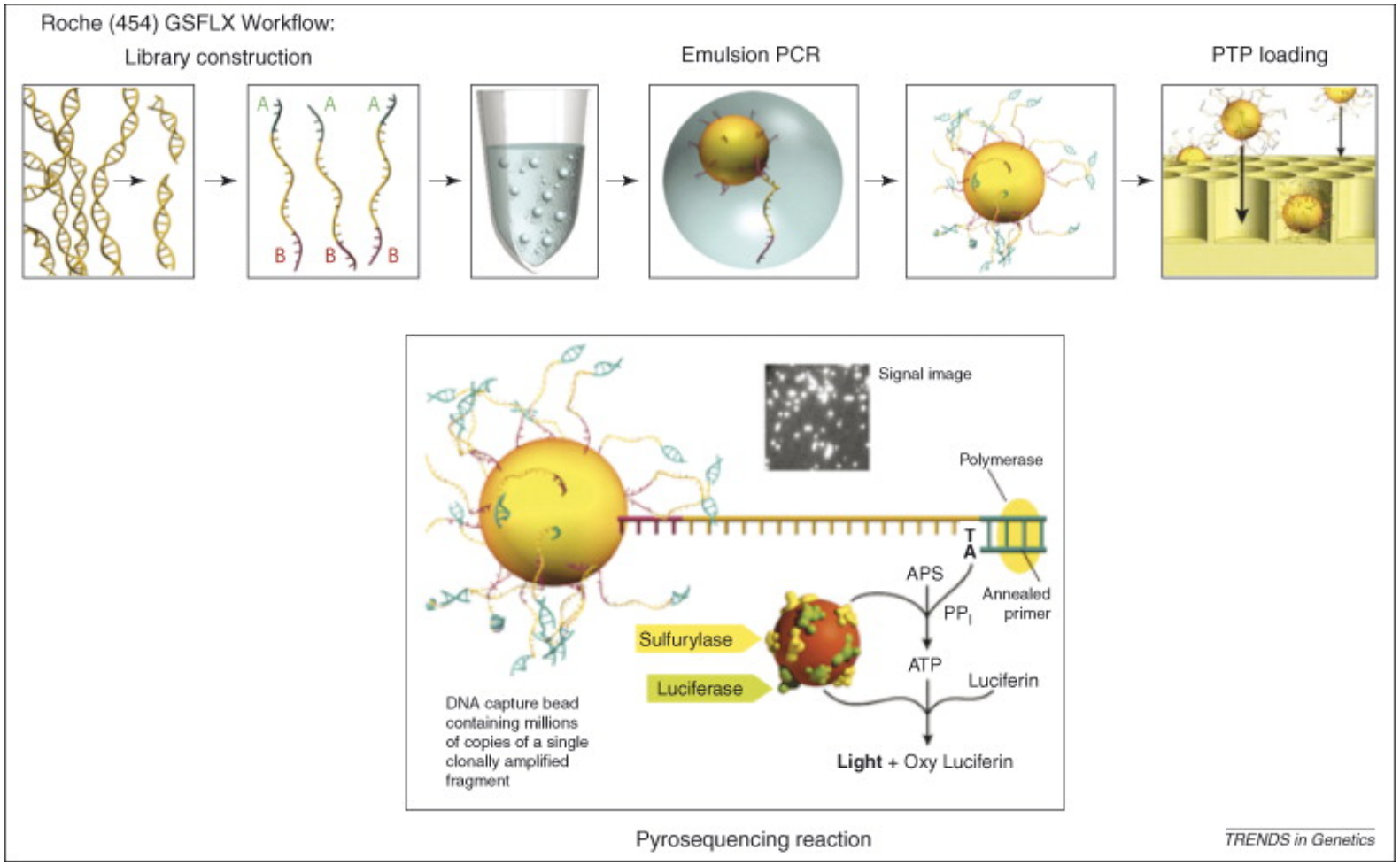


What is a good assembly and when is it finished?

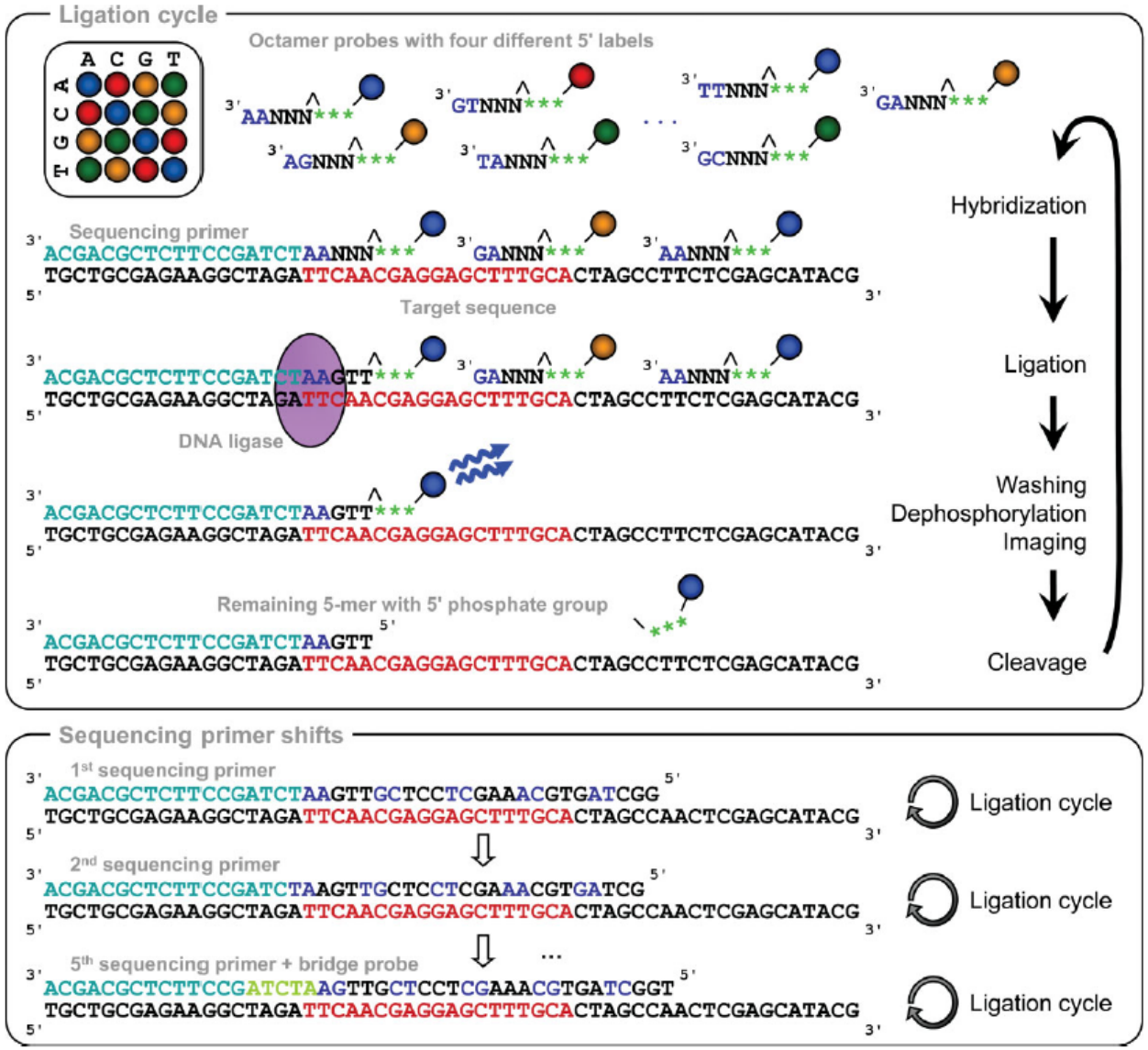
- No Finished Assemblies
  - >100 Euchromatic Gaps in Human Genome
  - *Drosophila melongaster* release 6.1 (More Centric Heterochromatic Sequences Added)
- Scaffold N50 Close to or Bigger than Average Gene Size
- Number of Scaffolds Should be Close to Number of Chromosomes

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

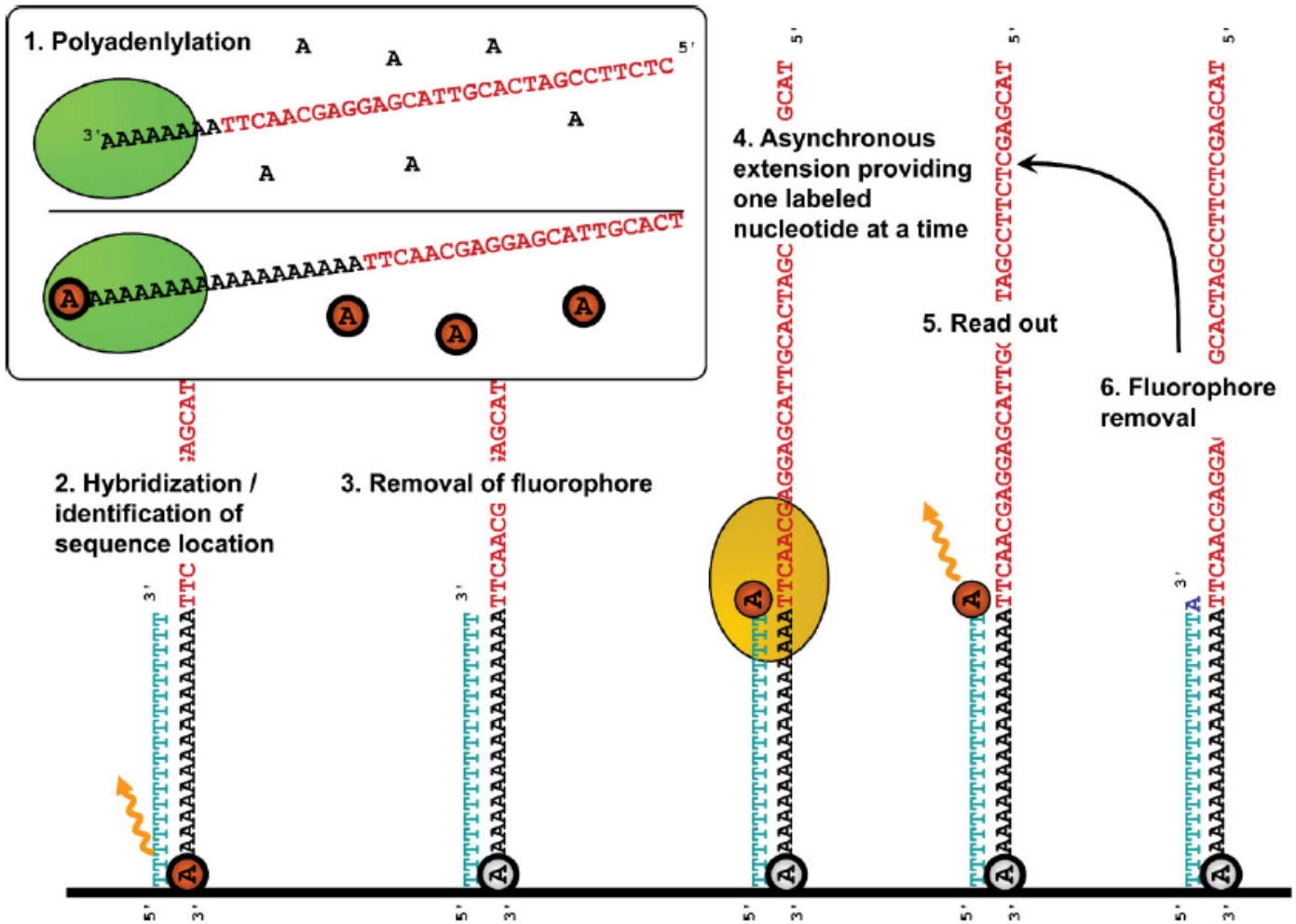
# Roche 454


Roche (454) GSFLX Workflow:
Library construction
Emulsion PCR
PTP loading
A
B
Signal image
Polymerase
T
A
APS
PPi
Annealed primer
Sulfurylase
Luciferase
ATP
Luciferin
DNA capture bead containing millions of copies of a single clonally amplified fragment
Light + Oxy Luciferin
Pyrosequencing reaction
TRENDS in Genetics

# ABI SOLiD


Ligation cycle
Octamer probes with four different 5' labels
A C G T
Hybridization
Sequencing primer
Target sequence
Ligation
DNA ligase
Washing
Dephosphorylation
Imaging
Remaining 5-mer with 5' phosphate group
Cleavage
Sequencing primer shifts
1st sequencing primer
2nd sequencing primer
5th sequencing primer + bridge probe
Ligation cycle
Ligation cycle
Ligation cycle

# Helicos




Kircher and Kelso 2010