

Genome Sequencing & Assembly

Deb Triant

University of Virginia, USA

Dept. Biochemistry & Molecular Genetics

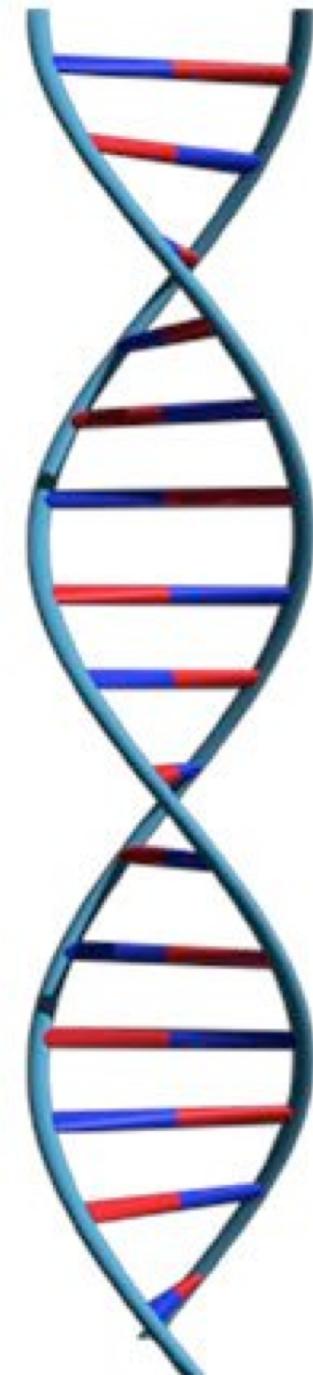
University of Missouri, USA

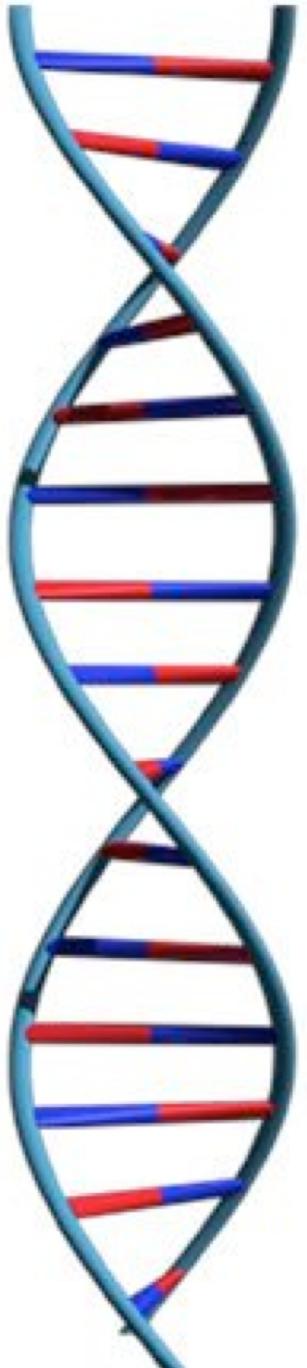
Dept. of Animal Sciences

21 October 2019

Programming for Biology

Cold Spring Harbor, NY





Lecture outline

1. General background of genome assembly & theory
2. Comparison of assembly methods
3. Recommendations for a good assembly project
4. Assembly workshop

It was the best of

age of wisdom, it was

best of times, it was

it was the age of

it was the age of

it was the worst of

of times, it was the

of times, it was the

of wisdom, it was the

the age of wisdom, it

the best of times, it

the worst of times, it

times, it was the age

times, it was the worst

was the age of wisdom,

was the age of foolishness,

was the best of times,

was the worst of times,

wisdom, it was the age

worst of times, it was

Book (Genome) Reconstruction

It was the best of

was the best of times,

the best of times, it

best of times, it was

of times, it was the

of times, it was the

times, it was the worst

times, it was the age

The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

Genome assembly!



BOVINEGENOME.ORG

THE BOVINE GENOME DATABASE



BovineMine BLAST JBrowse Apollo Lookup Tools Downloads Genome Consortia Assemblies

[UMD3.1 JBrowse](#)

[Btau_4.6.1 JBrowse](#)



A goal of the Bovine Genome Database project is to support the efforts of bovine genomics researchers by providing tools and annotation for the bovine reference genome based on the hereford cow, L1 Dominette 01449.

BDG provides tools for data mining ([BovineMine](#)), sequence database searching ([BLAST](#)), genome browsing ([JBrowse](#)). The bovine reference genome assembly has been [revised several times](#). BGD currently focuses on the [UMD3.1 assembly](#) and [Btau_4.6.1](#). We anticipate the release of an improved bovine reference genome assembly in early 2018, and plan to release it as soon as it is available.

The newest BGD tool is a data mining resource called [BovineMine](#), which integrates the UMD3.1 genome assembly and annotation with other genomic resources including genes, proteins, orthologs, pathways, gene ontology, gene expression, and more. BovineMine is designed to accelerate genomics analysis by enabling researchers without scripting skills to create and export customized annotation datasets merged with their own research data for use in downstream analyses. BovineMine allows researchers to leverage the curated biological data from the Bovine Genome Database, especially useful for GO and pathway analyses in conjunction with GWAS and QTL studies. BovineMine also facilitates cross-species data mining by tracking identifiers across gene sets and genome assemblies.

use of BGD, including BovineMine, JBrowse, Apollo, BLAST and other BGD tools:

MaizeMine v1.1 An integrated data warehouse for [MaizeGDB](#)

Home MyMine Templates Lists QueryBuilder Regions Data Sources API [release updates](#)

Contact Us | Log in Search: e.g. Zm00001d023208, GR | GO

Search
Search MaizeMine. Enter names, identifiers or keywords for genes, proteins, pathways, ontology terms, etc. (e.g. Zm00001d023210, GRMZM2G109674, Zm00001d023208_T002, GR2b, 100037783, NM_001111367.1).
Zm00001d023208
SEARCH

Analyze
Enter a list of identifiers.
Gene: Zm00001d023208, Zm00001d015434, Zm00001d008939, Zm00001d022478, Zm00001d015997, Zm00001d04217, Zm00001d013095
advanced **ANALYSE**

About v1.1 and Templates
MaizeMine 1.1 (release updates) integrates the Zea mays B73 genome assemblies (v3 and v4) with many biological data sets. Homologues are provided for several other monocots and *Arabidopsis thaliana*. The tabs below show categories for template queries, which provide simple search menus. "Alias and DBxref" template queries are for id conversion between the *Z. mays* AGPv3.21, AGPv4 and RefSeq gene sets.
TAKE A TOUR

GENES GENE EXPRESSION PROTEIN HOMOLOGY FUNCTION ENTIRE GENE SET ALIAS AND DBXREF



Hymenoptera Genome Database

Welcome to HGD

The Hymenoptera Genome Database (HGD) is a genome informatics resource that supports the research of insects of the order Hymenoptera (e.g. bees, wasps, ants). HGD is divided into three main divisions: [BeeBase](#), which hosts bee genomes and the [Bee Pests and Pathogens](#) resource; [NasoniaBase](#), which hosts the genome of the Jewel wasp (*Nasonia vitripennis*); and the [Ant Genomes Portal](#), which hosts the genomes of several ant species.

HGD provides tools for data mining ([HymenopteraMine](#)), sequence searching ([BLAST](#)), genome browsing ([JBrowse](#)), genome annotation (Apollo) and data download. HymenopteraMine and BLAST are shared across all three HGD divisions, while JBrowse, Apollo and data download pages are specific to each division.

[HymenopteraMine](#) is the newest HGD resource. [HymenopteraMine](#) is a data mining tool that integrates all of the genomes in HGD with biological data from a variety of sources. The goal of [HymenopteraMine](#) is to accelerate genomics analysis by enabling researchers without scripting skills to create and export customized annotation datasets merged with their own research data for use in downstream analyses. [HymenopteraMine](#) facilitates cross-species data mining based on orthology and supports meta-analyses by tracking identifiers across gene sets and genome assemblies.

How to Cite:

- Cite the use of HGD and HymenopteraMine with the most recent HGD publication:
Elsik CG, Tayal A, Diesh CM, Unni DR, Emery ML, Nguyen HN, Hagen DE. Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine. Nucleic Acids Research. 2016 Jan 4;44(D1):D793-800. doi: 10.1093/nar/gkv1208. Epub 2015 Nov 17. PubMed PMID: 26578564.
- Cite the appropriate [consortium publication](#) for the genome you used.



Nasonia vitripennis. ©Oliver Niehuis

Lepidopteran genome projects



Automeris io



Actias luna



Cyclargus thomasi



Papilio aristodemus

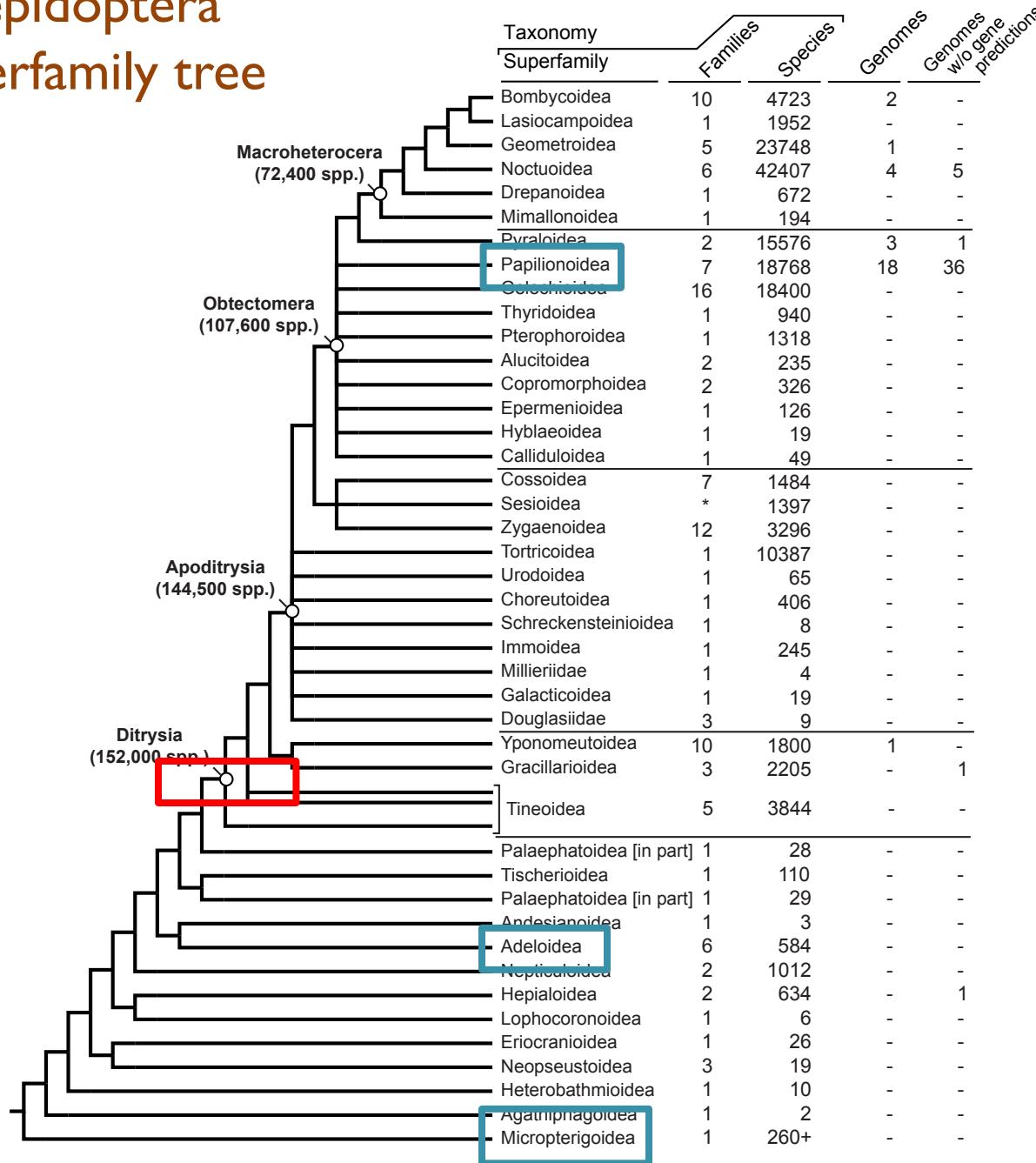


Eumeus atala

Deciding on assembly projects

- Financial limitations
- Specimen collection/access to fresh tissue
- Computational resources
- Inbred lines/rearing difficulties

Lepidoptera superfamily tree



Deciding on assembly projects

- Financial limitations
- Specimen collection/access to fresh tissue
- Computational resources
- Inbred lines/rearing difficulties

Silene

- Angiosperm - *Caryophyllaceae*
- Over 600 species
- *S. vulgaris*
 - organelle evolution
 - population genetics
 - biology of invasions
 - chemical pathways
 - niche modeling



A. Berardi

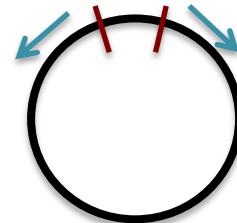
Wouldn't it be great if we had a reference genome?

Sequencing Silene



I. Greenhouse crosses

2. Library construction



3. Sequencing

4. Data analysis

GACCTACA
ACCTACAA
CCTACAAG
CTACAAGT
TACAAGTT
ACAAGTTA
CAAGTTAG



More than six months later.....

The Genome Plant



Library construction

I. Greenhouse crosses

2. Library construction

- two small-insert
- three long-insert
 - 2 kb, 5kb, 10kb outsourced

Months later.....

Library construction

“.....we completed your Illumina mate pair libraries and the QC looked great for sequencing”

“.... Unfortunately, the samples were inadvertently discarded by one of my staff”

“... I view this as the most terrible occurrence since I've opened the lab.”

Sequencing Silene

I. Greenhouse crosses



2. Library construction

3. Sequencing at outside facility

- Illumina Hiseq



4. Data analysis

GACCTACA
ACCTACAA
CCTACAAG
CTACAAGT
TACAAGTT
ACAAGTTA
CAAGTTAG
TACAAGTC
ACAAGTCC
CAAGTCCG

Sequencing Silene

- Data analysis
 - read quality
 - assembly & annotation

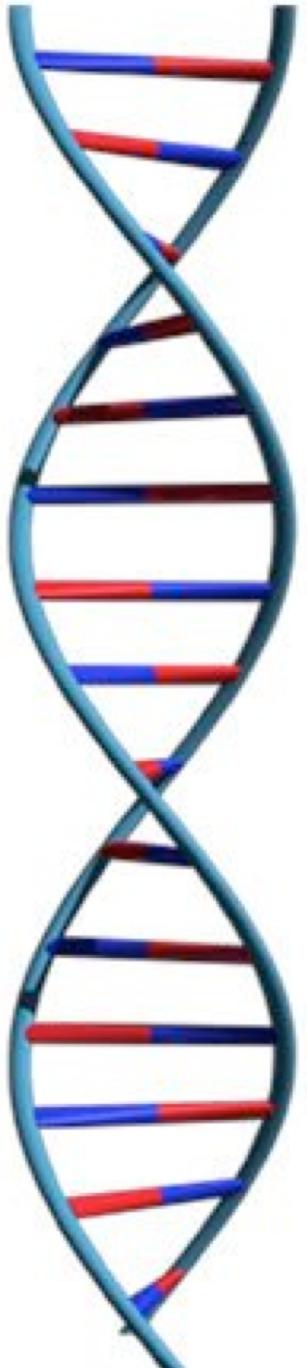
GACCTACA
ACCTACAA
CCTACAAG
CTACAAGT
TACAAGTT
ACAAGTTA
CAAGTTAG
TACAAGTC
ACAAGTCC
CAAGTCCG

“I think it makes sense to find someone else to do the ‘grunt work’....”

(‘grunt work’ = entire assembly and analysis!!!)

“.....maybe Deb can do it”

“Have a plan”!



Lecture outline

- I. My background & introduction to genome assembly
2. General background of genome assembly & theory
3. Comparison of assembly methods
4. Recommendations for a good assembly project

*Afternoon - assembly and alignment workshop

It was the best of

age of wisdom, it was

best of times, it was

it was the age of

it was the age of

it was the worst of

of times, it was the

of times, it was the

of wisdom, it was the

the age of wisdom, it

the best of times, it

the worst of times, it

times, it was the age

times, it was the worst

was the age of wisdom,

was the age of foolishness,

was the best of times,

was the worst of times,

wisdom, it was the age

worst of times, it was

Shredded Book Reconstruction

It was the best of

was the best of times,

the best of times, it

best of times, it was

of times, it was the

of times, it was the

times, it was the worst

times, it was the age

The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

Model the assembly problem as a graph problem

Graph Construction

- Graph representing overlaps between subfragments
- $D_k = (V, E)$
 - $V = \text{All length-}k \text{ subfragments } (k < l)$
 - $E = \text{Directed edges between consecutive subfragments}$
 - Nodes overlap by $k-1$ words

Original Fragment

It was the best of

Directed Edge

It **was the best**



was the best of

- Locally constructed graph reveals the global sequence structure
 - Overlaps between sequences implicitly computed

de Bruijn, 1946

Idury and Waterman, 1995

Pevzner, Tang, Waterman, 2001

Graph Assembly

It was the best

was the best of

the best of times,

best of times, it

of times, it was

times, it was the

it was the worst

was the worst of

the worst of times,

worst of times, it

it was the age

was the age of

the age of foolishness

the age of wisdom,

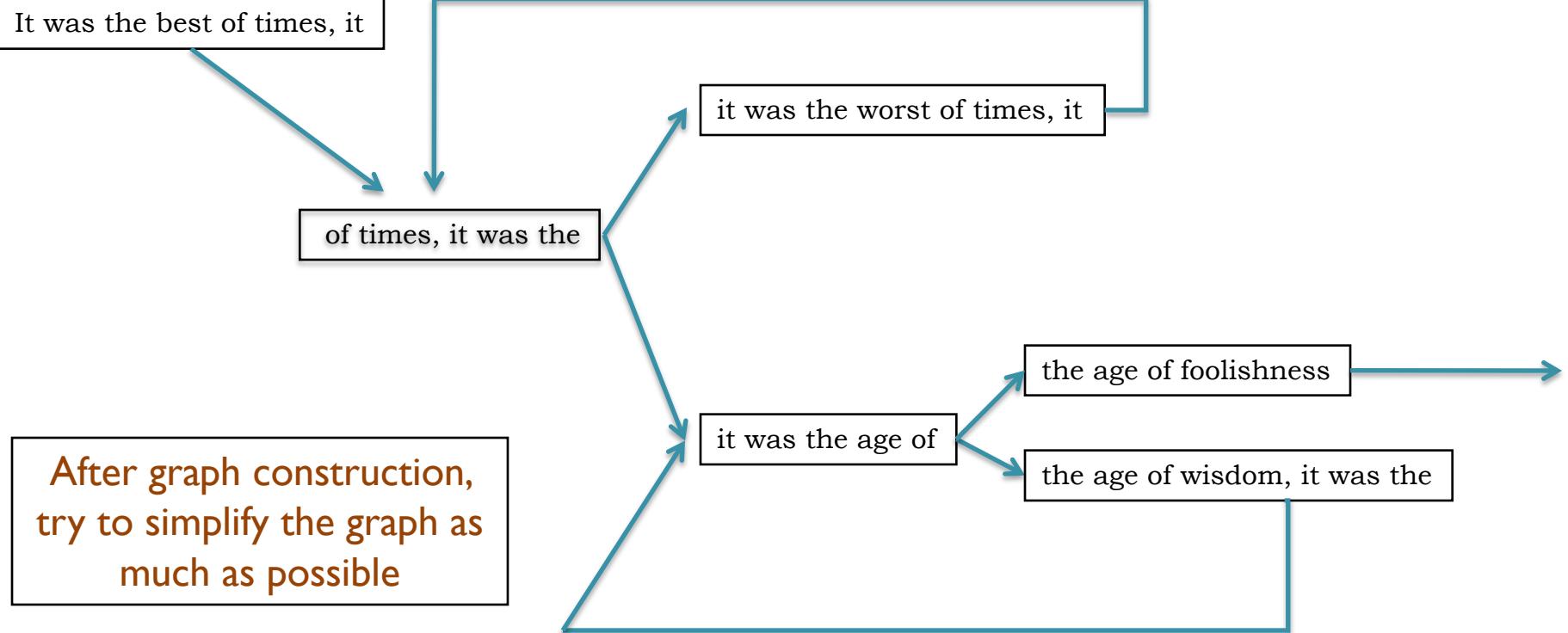
age of wisdom, it

of wisdom, it was

wisdom, it was the

After graph construction,
try to simplify the graph as
much as possible

Graph Assembly



Assembly outline

Overlap → Layout → Consensus

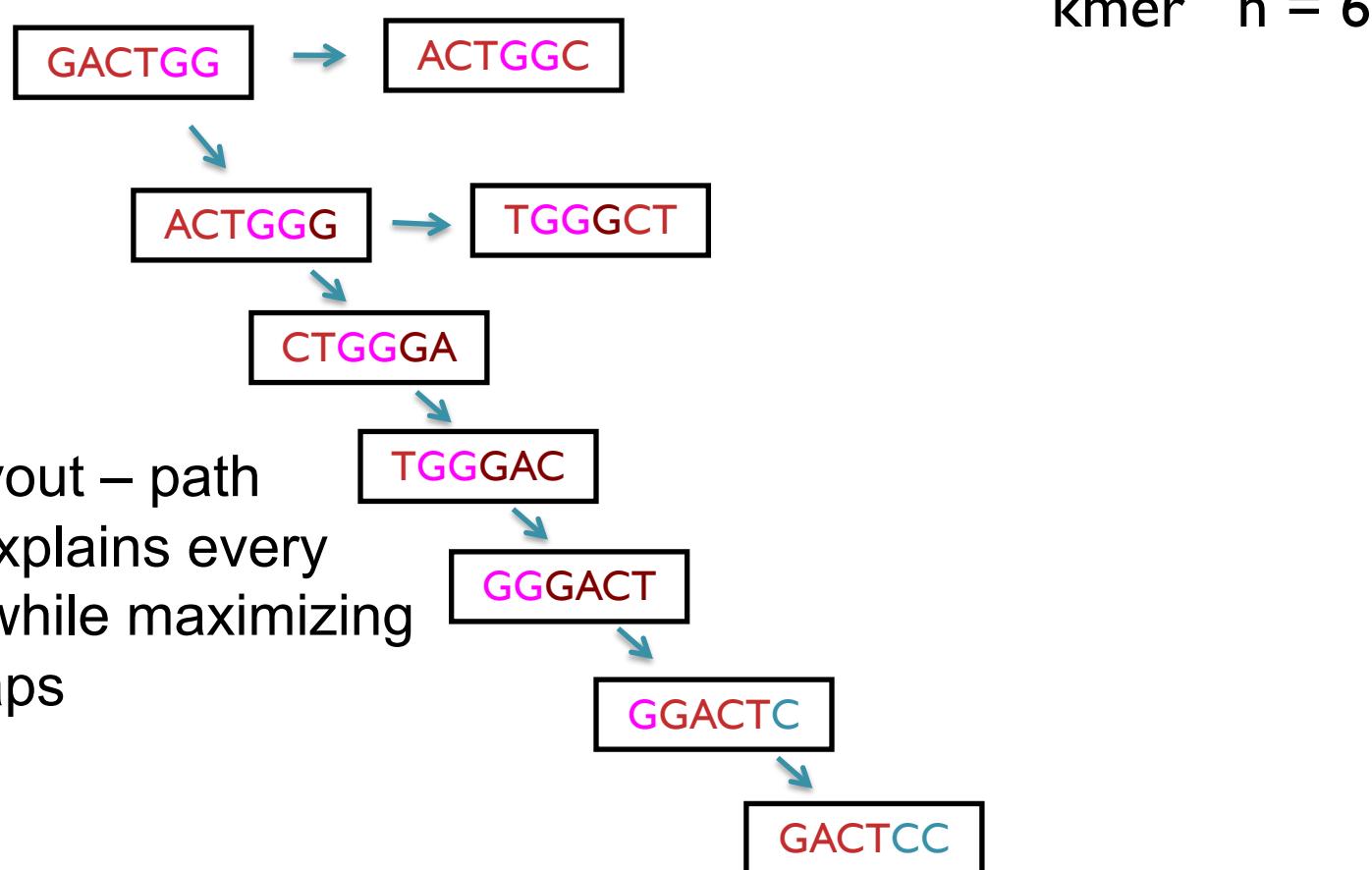
1. Overlap: identify overlapping reads
2. Layout: order reads
3. Consensus: merge reads into a sequence while correcting errors

Graph Assembly

I. Overlap

- Shredded words k-mer

GACTGGGAACTCC



2. Layout – path
that explains every
read while maximizing
overlaps

Graph Assembly

3. Consensus

- sequence alignments to represent the same position

G TGGGACTCCGGCATTAGC TA
GA~~T~~CAGACTCCGGCATTAGCCTA
GA~~T~~CTGGGACTCCGGCATTAGCCTA
GA~~T~~CTGGG TCCGGCATT~~C~~GCCTA
GA~~T~~CTGGGACTCCGGCATTAGCCTA
GA~~T~~CTGGGACTCCGGCATTA CCTA

↓ ↓ ↓ ↓ ↓

GA~~T~~CT~~GG~~ACTCCGGCATT~~A~~GCCTA

The full tale

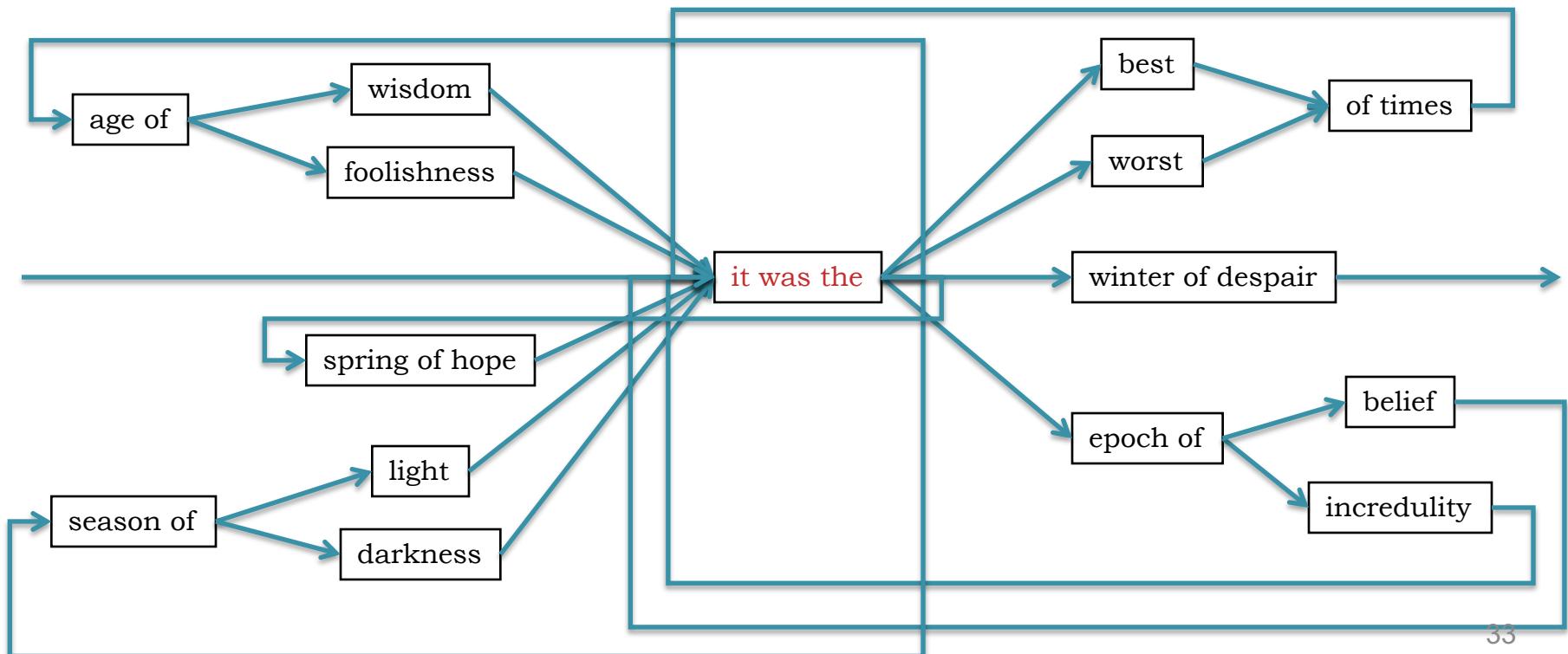
... it was the best of times it was the worst of times ...

... it was the age of wisdom it was the age of foolishness ...

... it was the epoch of belief it was the epoch of incredulity ...

... it was the season of light it was the season of darkness ...

... it was the spring of hope it was the winter of despair ...



History of Genome Assembly

1977. Sanger et al. 1st Complete Organism bacteriophage 5375 bp

1995. Fleischmann et al. 1st Free Living bacteria; *Haemophilus influenzae*; TIGR Assembler. 1.8Mb

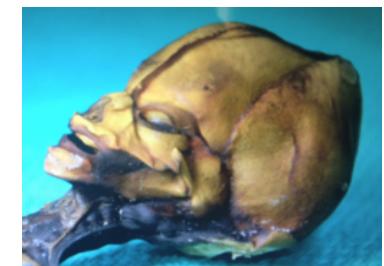
1998. *C.elegans* SC 1st Multicellular Organism BAC-by-BAC Phrap. 97Mbp

2000. *Drosophila* genome; Myers et al. 1st Large WGS Assembly Celera Assembler. 116 Mbp



Human Genome

Public: 13-year project began 1990, Dept Energy & NIH,
\$3 billion; millions of small fragments
2003 – announced as complete



Bhattacharya et al. 2018.
Genome Research



Private: Craig Venter, Celera Genomics; 1998, \$300 million
Could not be patented.

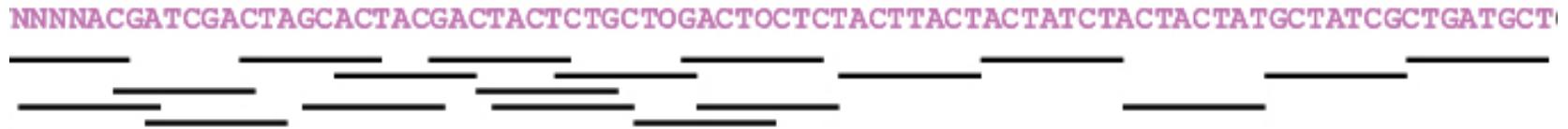


Why are genomes so difficult to assemble?

- Biological
 - Heterozygosity, repetitive regions, ploidy
- Sequencing
 - Genome size, sequencing errors, inconsistencies
- Computational
 - Million or billions of reads, complexity
- Accuracy
 - Difficult to assess accuracy - assemblers

Considerations for assembly projects

Coverage



- How many times has genome been sequenced?
reads, read length, genome size
- Too much? Too little? Aim for oversampling ~40-60x. 100x
- Raw read depth vs Mapped read depth
 - driven by efficiency of alignment process

Considerations for assembly projects

Coverage



Much is driven by funding and application

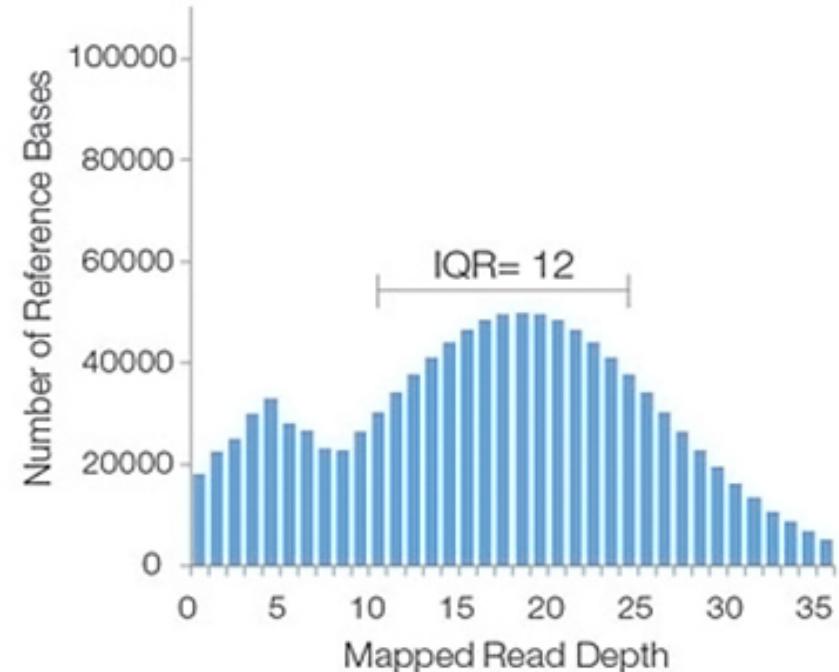
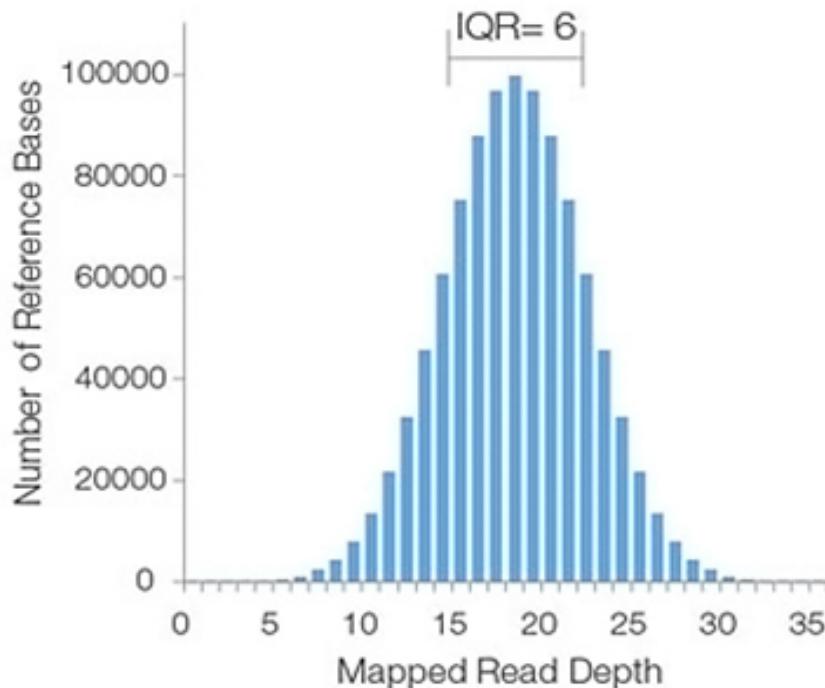
- SNPs and genome rearrangements
- particular coding regions
- ‘complete’ genome

Depth vs Coverage

Coverage - how much of genome is sequenced; how many times nucleotide is represented in a read

Depth - number of reads aligned to that position

Coverage histograms



Assumes reads randomly distributed across the genome

<https://www.illumina.com/science/education/sequencing-coverage.html>

Calculating coverage

- $C = LN/G$
- $G = \text{Genome size: } 2 * 10^9$
- $N = \text{Number reads: } 150 * 10^6$ (HiSeq lane)
- $L = \text{Length of reads: } 100 \text{ nucl} * 2$ (Paired-End)
 - $3 * 10^{10}$ (15X coverage) Goal: 80X = ~5 - 6 lanes

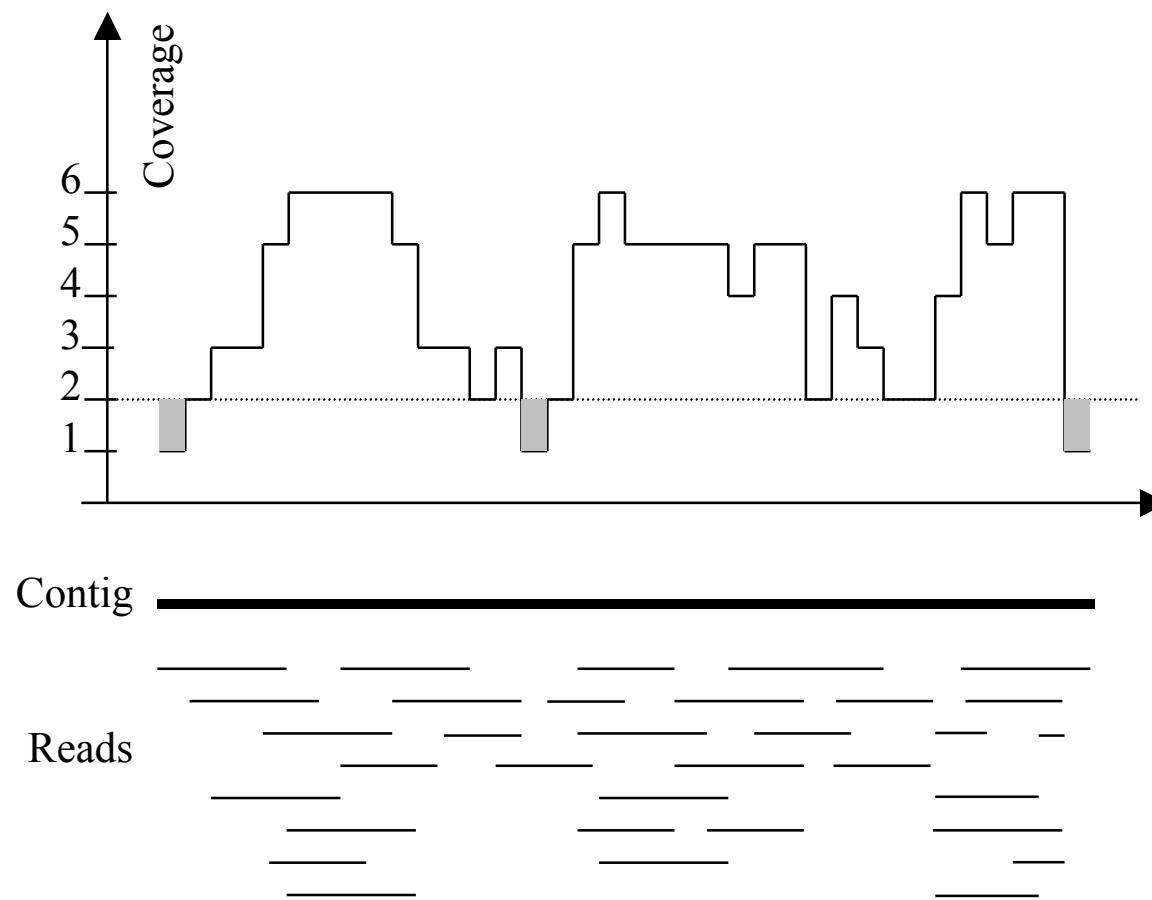
Many coverage calculators available online:

- genome size, technology

https://www.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf

<http://www.slimsuite.unsw.edu.au/servers/pacbio.php>

Typical sequencing coverage



Considerations for assembly projects

Read Length

- Read lengths *must* be longer than repetitive regions
- Short vs long read technologies

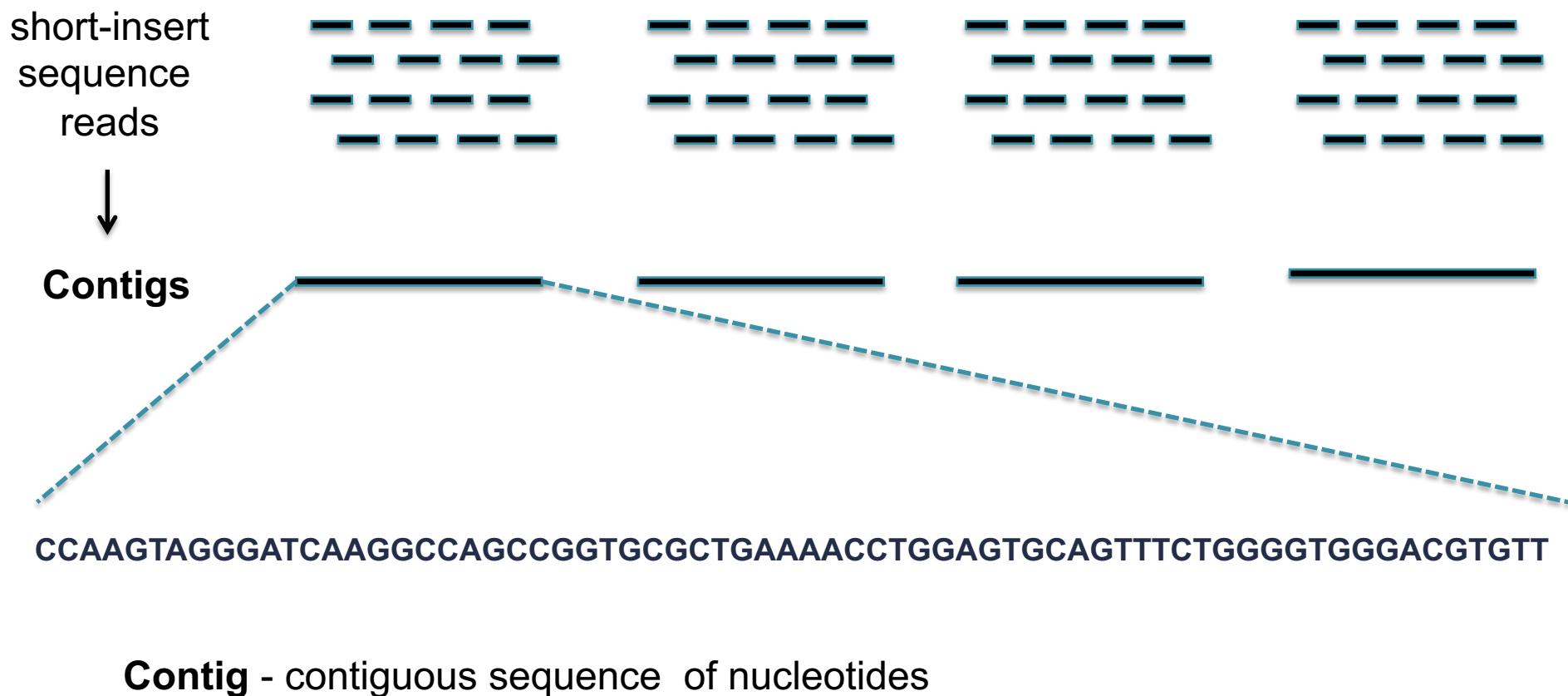
Quality

- Reads and template material

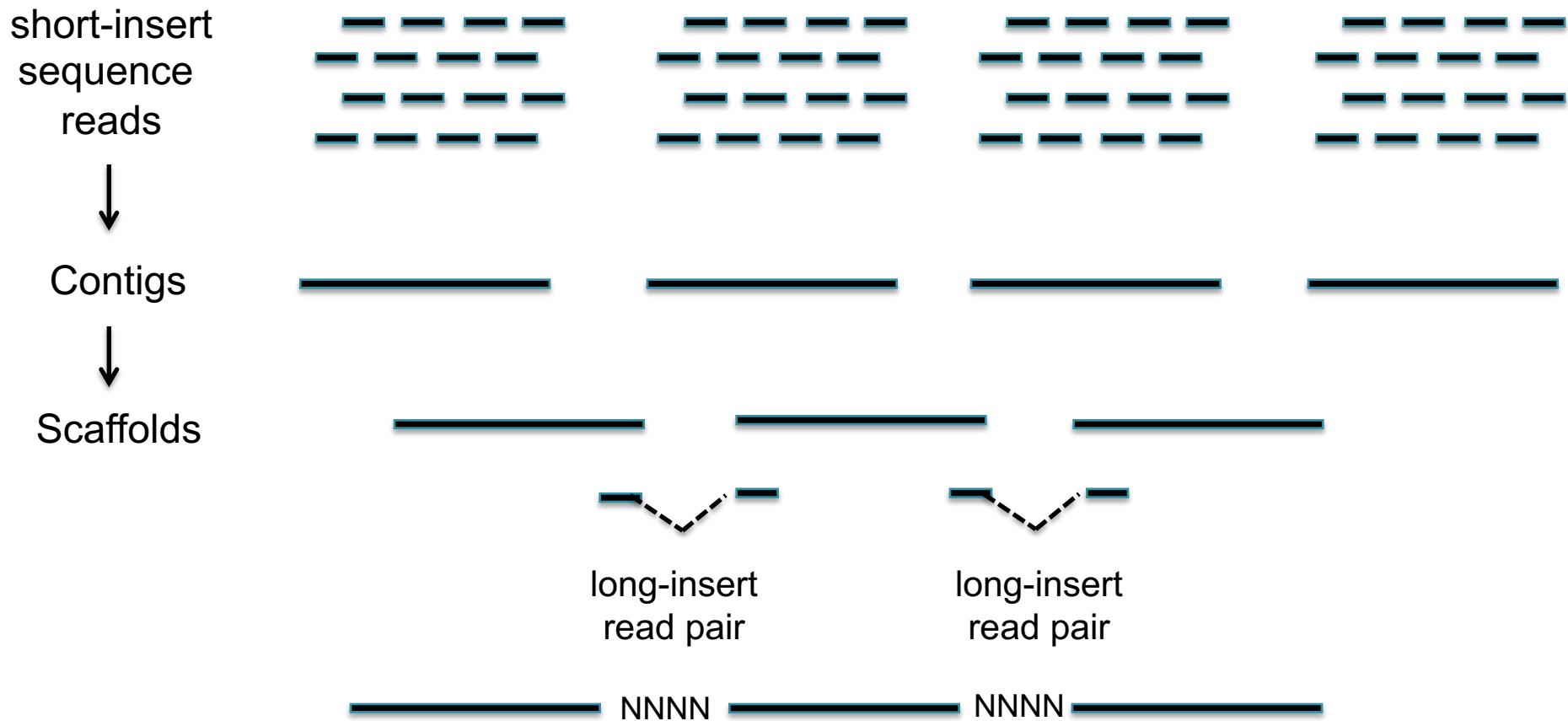
DNA requirement projections

- High quality DNA! **often our biggest limitation
- Number and type of libraries required
- Potential projects resulting from assembly

Assembly construction – Hierarchical process



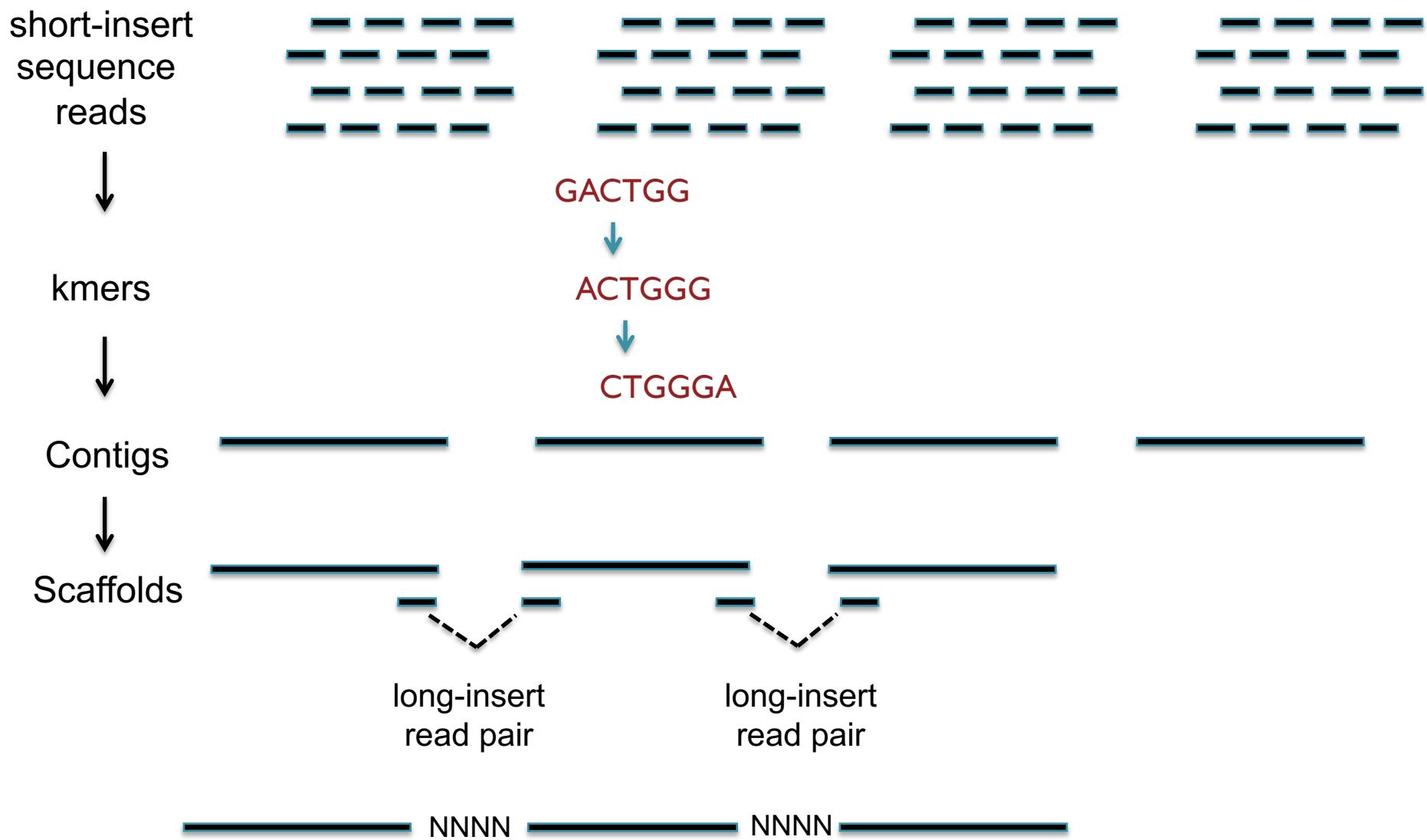
Assembly construction



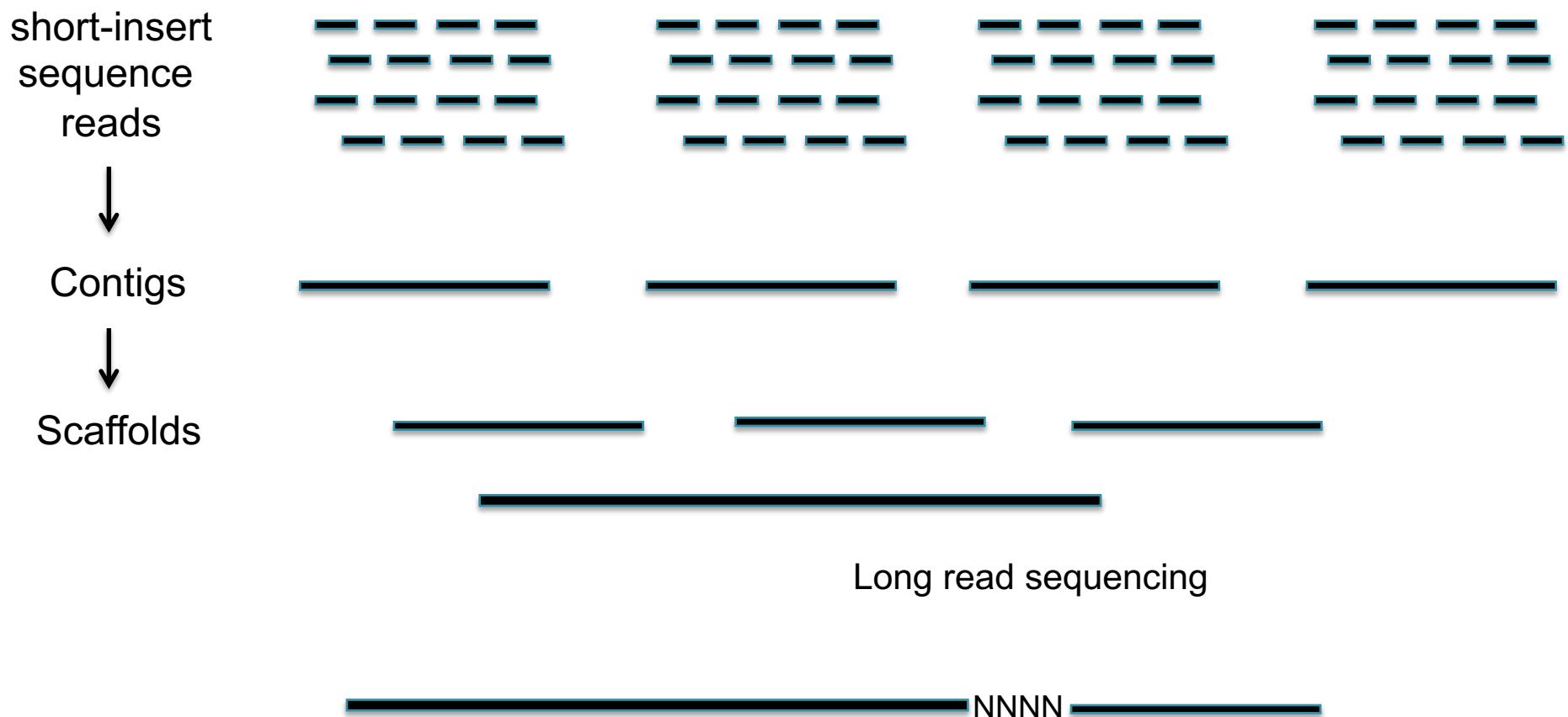
Scaffold - sequence of contigs, separated by gaps - Ns are predicted gap size

```
>GTAGTATTCTAGAAAATGTTAACATAGATAGTTGTTANATCTGTTAGTGTCAGATGCTACTGAATAGTTGGAANNNNNNNNNNNNNNNNNN  
NNNNNNNTGTGAGGTTTAGCTCATGAAAGTTATGATTATTGCACCCCTACTCACAAACGAATCCCTATTCTTATCTTTNNNNNNNNNNNNNNN  
CATGTCAGTTTATTATTGGCTGCAGAAGTCCTTGTGCTGTTAATTTGGAGTTCTCCTGTCGTATATAAGCTTCTTCTTCAGTT  
TAAATTATTGAACCTTACTATCTTCTAACAAATAATTGGAATTATCAACGAAAACATAGGNNNNNNNNNNNGTCCTTATACGAAAGCTATATAG  
TGTAGGCTTCTTTNNNNNNNNGGTGATGTTGTTAATGGTGCCTTCTGGAATCTTACTAAATCAGTTGCTGTTACTGTATAGTTG
```

Assembly construction



Assembly construction



Scaffold - sequence of contigs, separated by gaps - Ns are predicted gap size

```
>GTAGTATTCTAGAAAATGTTAACATAGATAGTTGTTANATCTGTTAGTGTCAGATGCTACTGAATAGTTGGAANNNNNNNNNNNNNNNNNN  
NNNNNNNTGTGAGGTTTAGCTCATGAAAGTTATGATTATTGCACCCCTACTCACAAACGAATCCCTATTCTTATCTTTNNNNNNNNNNNNNNN  
CATGTCACGGTTATTTATTTGTGGCTGCAGAAGTCCTTGTGCTGTTAATTTGGAGTTCTCCTGTCGTATATTAAAGCTTCTTCTTCAGTT  
TAAATTATTAACCTTACTATCTTCTAACATAAAATTGGAATTATCAACGAAAACATAGGNNNNNNNNNNNGTCCTTATACGAAAGCTATATAG  
TGTAGGCTTCTTTNNNNNNNNGGTGATGTTGTTAATGGTGCCTTCTGGAATCTTACTAAATCAGTTGCTGTTACTGTATAGTTG
```

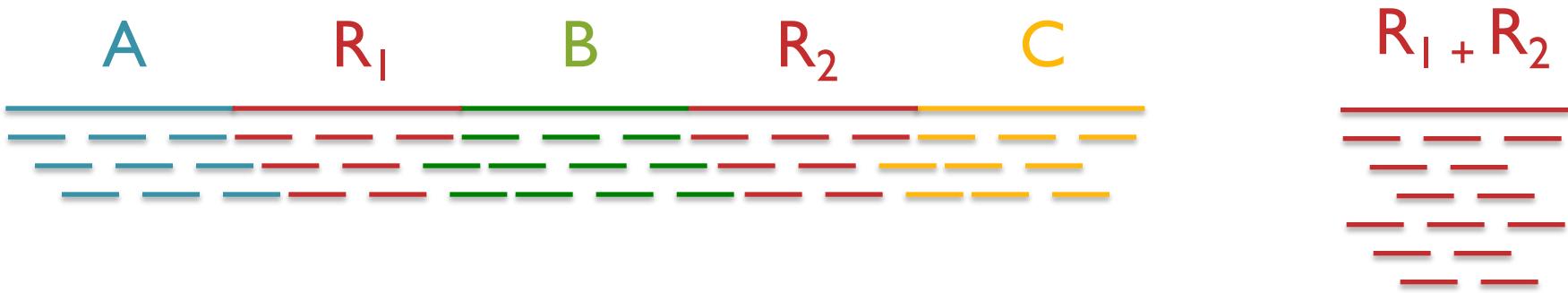
Assembly construction



Repetitive regions

- Over 50% of mammalian genomes are repetitive
 - Large plant genomes tend to be even worse
 - SINEs - Short Interspersed Nuclear Elements
 - LINES - Long Interspersed Nuclear Elements
 - LTR - Long Terminal Repeats, retrotransposons
 - Segmental duplications
 - Low-complexity - Microsatellites or homopolymers

Repeats and Coverage Statistics



If reads are a uniform random sample of the genome, we would expect relatively uniform distribution. If we see more reads than expected, likely a collapsed repeat.

Sequencing challenges

- Resolving repeats
 - Repeats longer than reads *cannot* be resolved
 - Merge reads up to repeat boundaries
 - Tandem repeats often not resolved
 - RepeatMasker - <http://repeatmasker.org>
screens DNA sequences for interspersed repeats and low complexity DNA sequences

- Error correction
 - incorporation of long reads
 - correction algorithms - Quake

Correction by coverage.

Figure S16, evaluates the ability of high-coverage to correct for sequencing errors. With sufficient coverage, even high error rates can be compensated for.

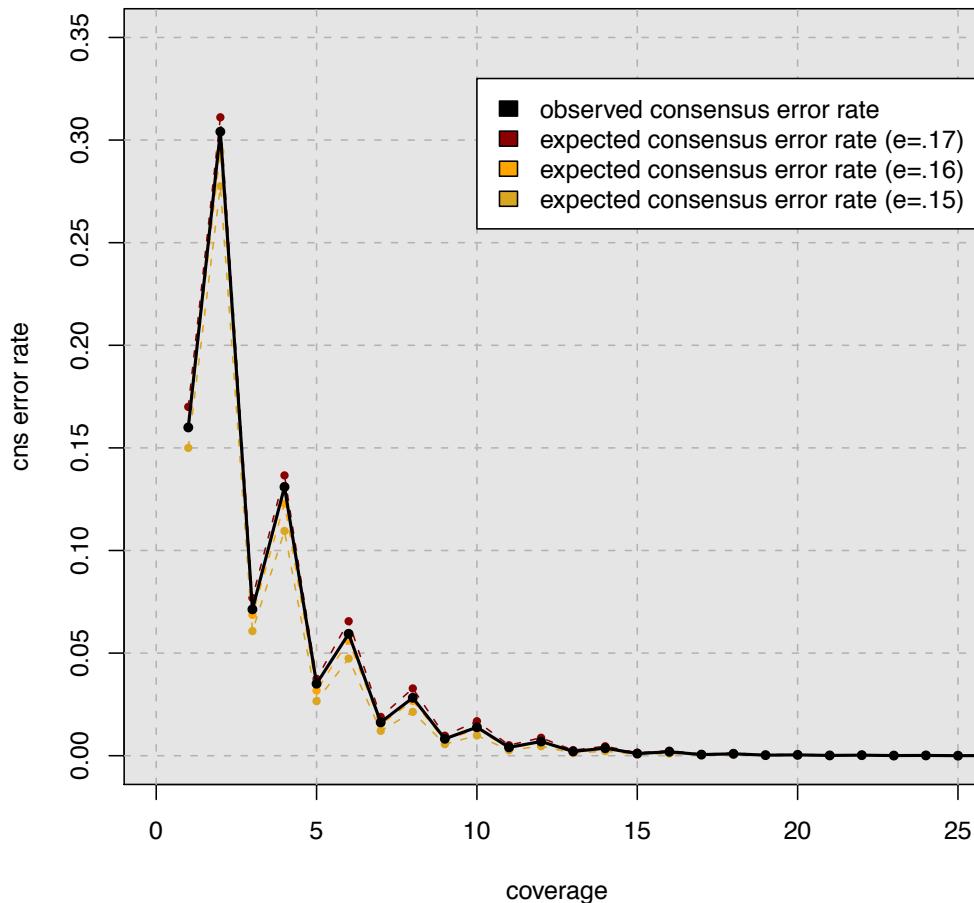
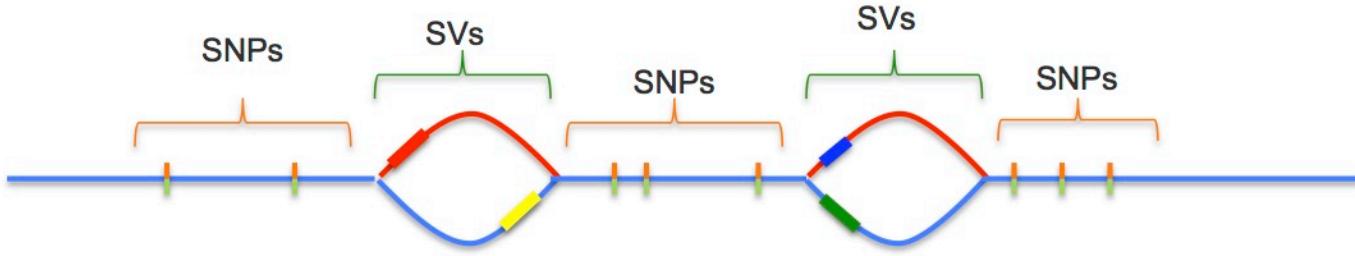


Figure S16. Coverage can overcome most random errors. 1,000 bp reads for *E. coli* K12 were simulated with random errors and the resulting consensus accuracy was measured. Even with high errors, coverage over 10X is sufficient to generate an accurate consensus. The periodic fluctuation in consensus error rate is an artifact of the tie-breaking scheme used in the consensus simulation (even numbers of reads can have ties and odd cannot).

Hybrid error correction and de novo assembly of single-molecule sequencing reads.

Koren et al (2012) *Nature Biotechnology*. doi:10.1038/nbt.2280

Phased genome assemblies



“Pseudo-haplotypes” - mixture of both types

How do we get phased genomes?

Higher heterozygosity - easier to phase.

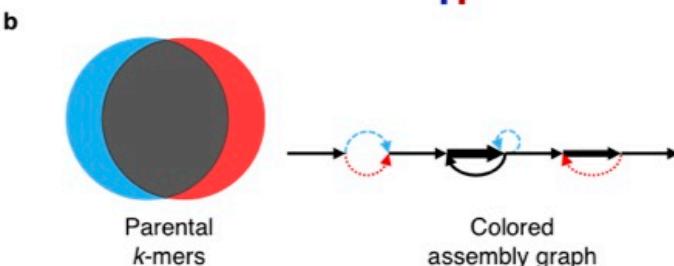
Chin et al. *Nat Meth* 2016

Assembly of heterozygous genomes

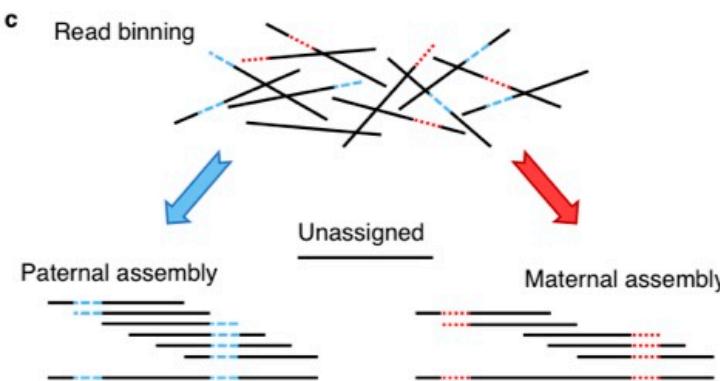
Trio-binning



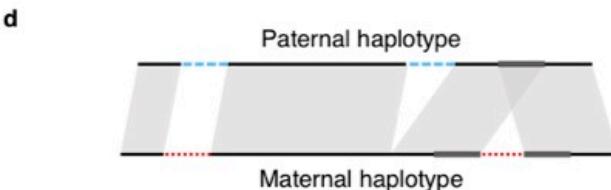
Short reads from two parental genomes to identify unique k-mers - breed specific



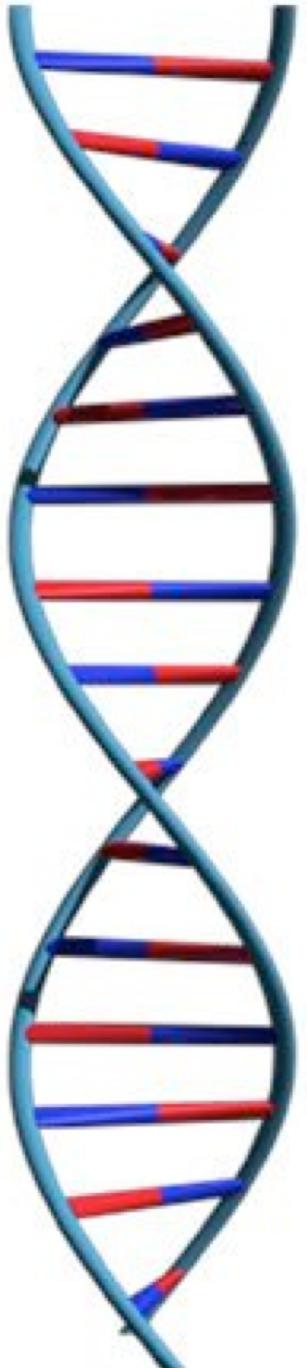
Long reads from offspring partitioned into into haplotype-specific sets prior to assembly



Consistent path through each haplotype
Homozygotes rep. twice

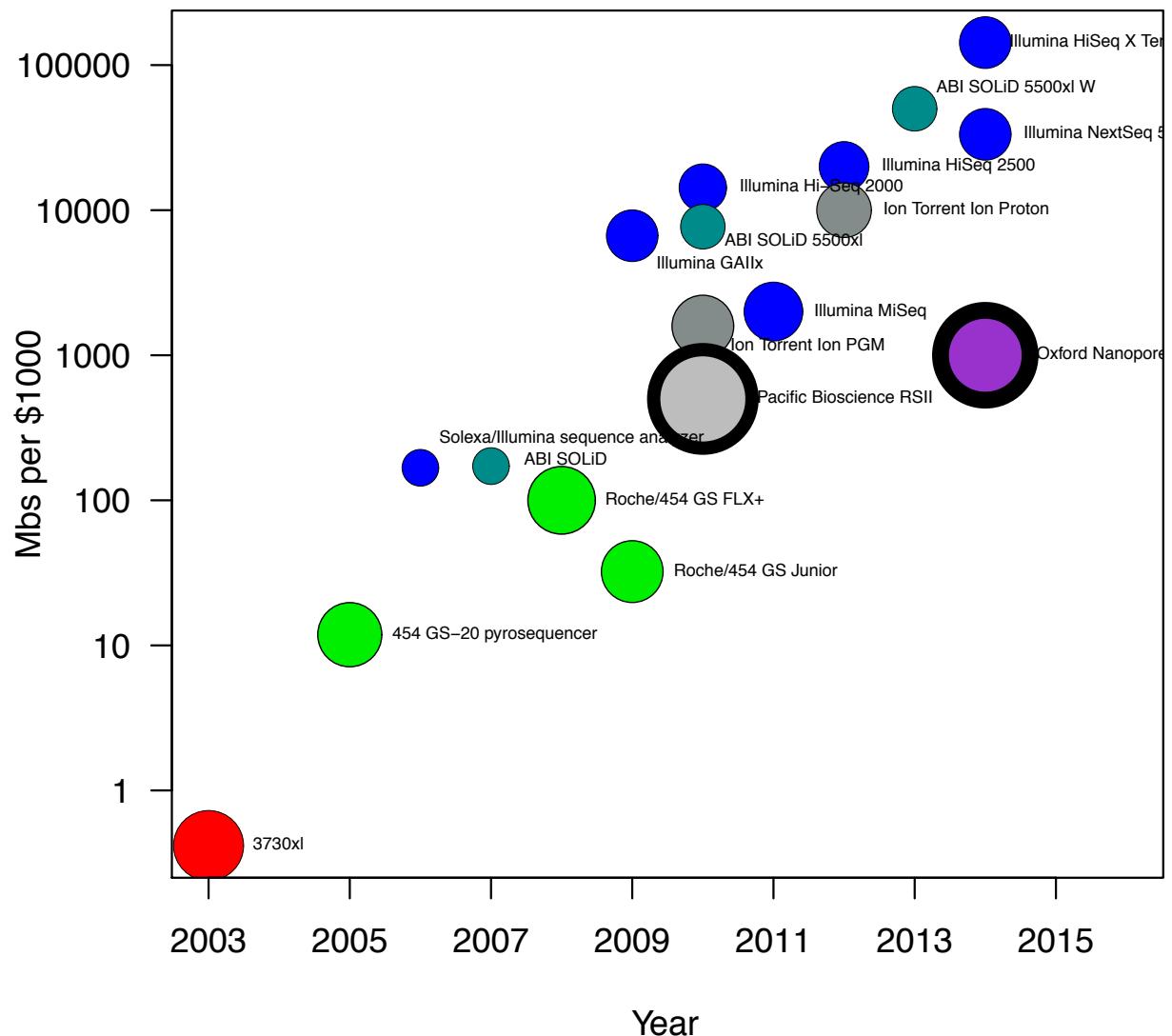


Each haplotype assembled separately
to represent diploid assembly



Lecture outline

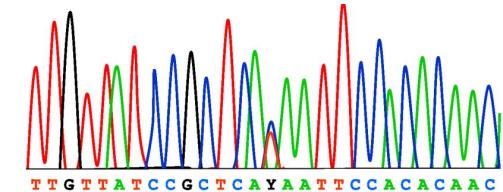
- I. General background of genome assembly & theory
2. Comparison of assembly methods
3. Recommendations for a good assembly project
4. Assembly workshop



(<http://www.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-186.pdf>)

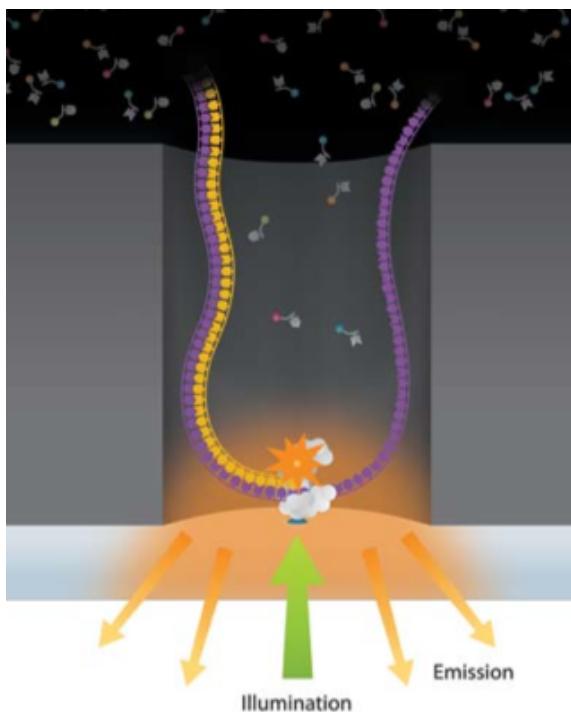
Sequencing Technologies

- Sanger
 - 800 bp reads with low error rate, costly
- 454 sequencing
- Reads up to 1 kb; poor with homopolymers, closed 2013
- Illumina
 - Paired-End, Long-insert libraries
 - HiSeq (300pb); MiSeq (500bp)
 - Multiplexing; inexpensive, low error rates
 - Miseq, Hiseq (2000, 3000, xTen), NexSeq

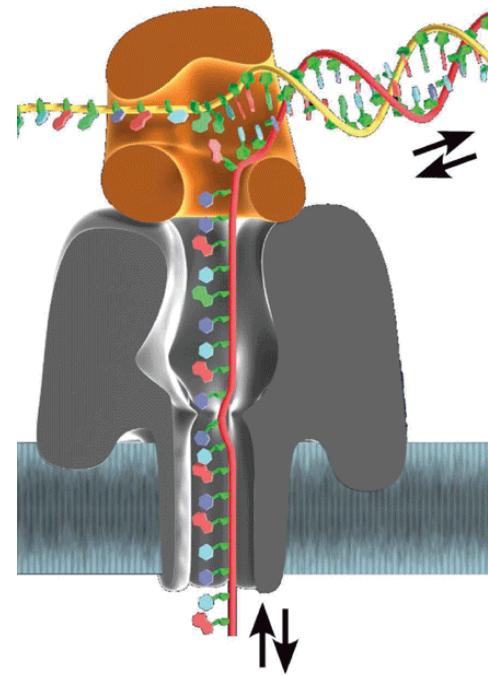


“3rd Generation” Sequencing

Pacific Biosciences



Nanopore



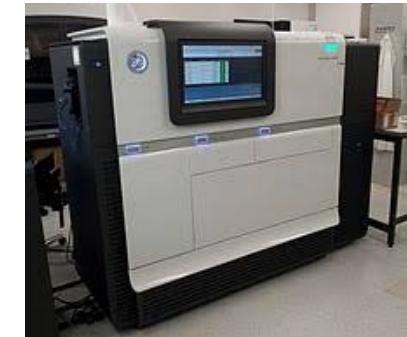
Images from Schadt *et al.* Human Molecular Genetics, 2010 and Schneider *et al.* Nature Biotechnology, 2012

Sequencing Technologies

- Single Molecule Real Time Sequencing

- Pac Bio

- Long read sequencing; 10 - 60kb reads; expensive
 - Relatively high error rate (~15-20%) but can error correct
 - Sequenced only on PacBio machines



- Oxford Nanopore - MiniON

- portable for sequencing on laptop in the field
 - > 100 kb
 - DNA sequenced by threading through microscopic pores with no amplification or chemical labeling of samples
 - Flash-drive sized sequencer run out of USB port



Sequencing Technologies

- Synthetic long reads
 - Dovetail Genomics (Chicago library method; Hi-C)
Reads up to 100kb.
Sequenced on Illumina platforms
Scaffolding platform
 - 10X Genomics
100kb synthetic length
Sequenced on Illumina platforms

Sequencing Technologies

- Oxford Nanopore
 - portable for sequencing in the field
 - >100 kb reads
 - DNA sequenced by threading through microscopic pores with no amplification or chemical labeling of samples
- High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. bioRxiv:
<https://www.nature.com/articles/s41467-018-03016-2>
- Linear Assembly of a Human Y Centromere using Nanopore Long Reads
<https://www.nature.com/articles/nbt.4109>

Paired-end and Mate-pairs

Paired-end sequencing (*single-end, just one*)

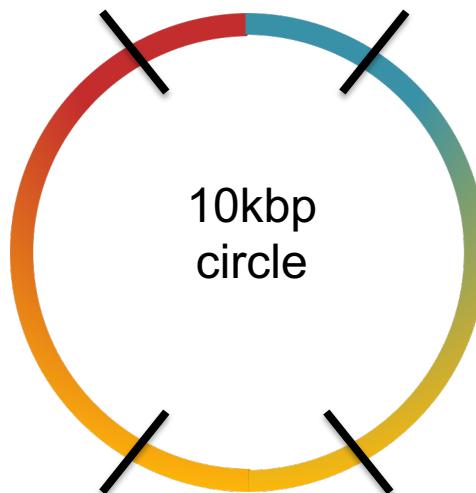
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation



Mate-pair sequencing

- Circularize long molecules (1-10kbp), shear into fragments, & sequence

10kbp



2x100 @ ~10kbp (outward)



PacBio library construction

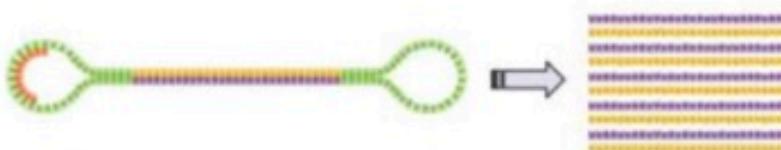
SMRT Sequencing Options

Standard



- Generates one pass for each molecule sequenced
- Large insert sizes, average reads over 5 kb (10 kb soon)
- Reads to over 20 kb
- ~13% error rate, random, mostly indels

Circular Consensus



- Small insert sizes
- Generates multiple sense and antisense reads of each molecule sequenced to generate high quality score



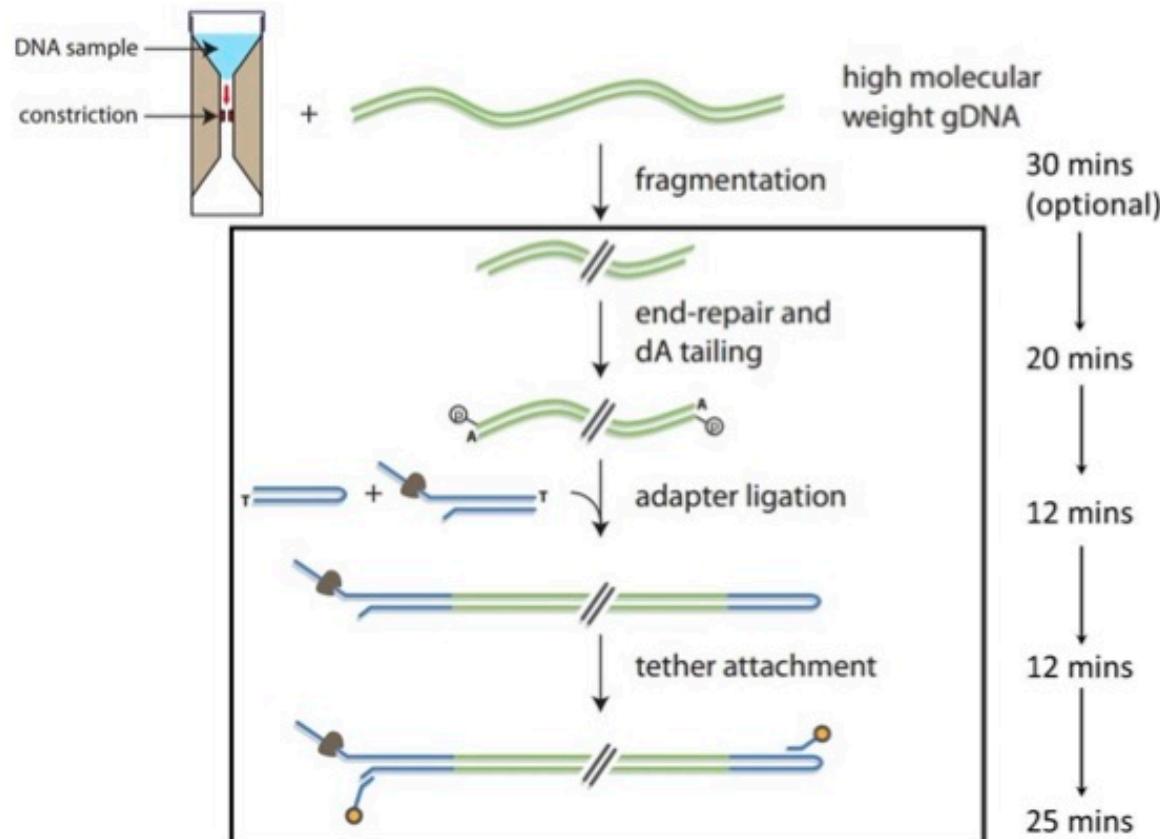
PACIFIC
BIOSCIENCES™

Nanopore library construction

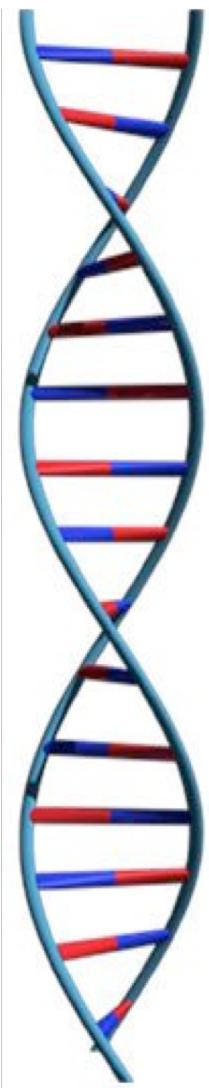
Library making - “quick & easy”

New TE-based method - 10 min!

Limited by DNA isolation



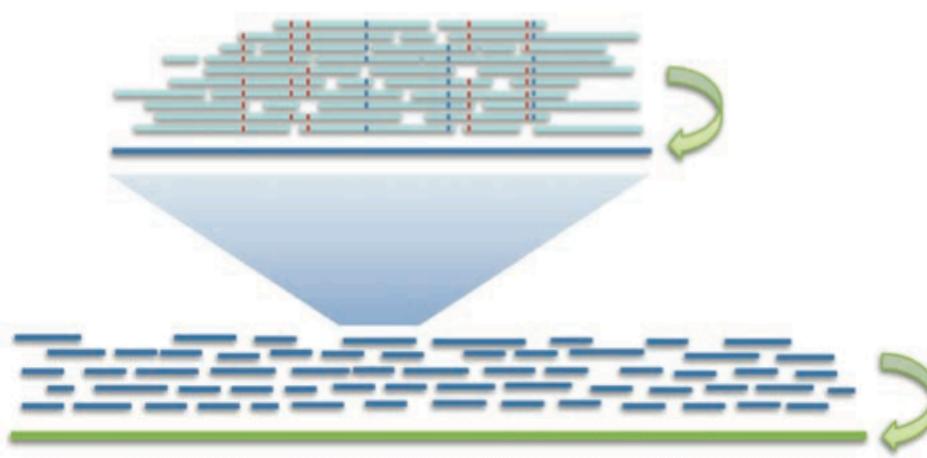
Long read sequencing



Longest sequencing
reads as seed data set

Use seed dataset
to map shorter
reads

Perform preassembly:
construct preassembled
reads from seed reads
through consensus
procedure

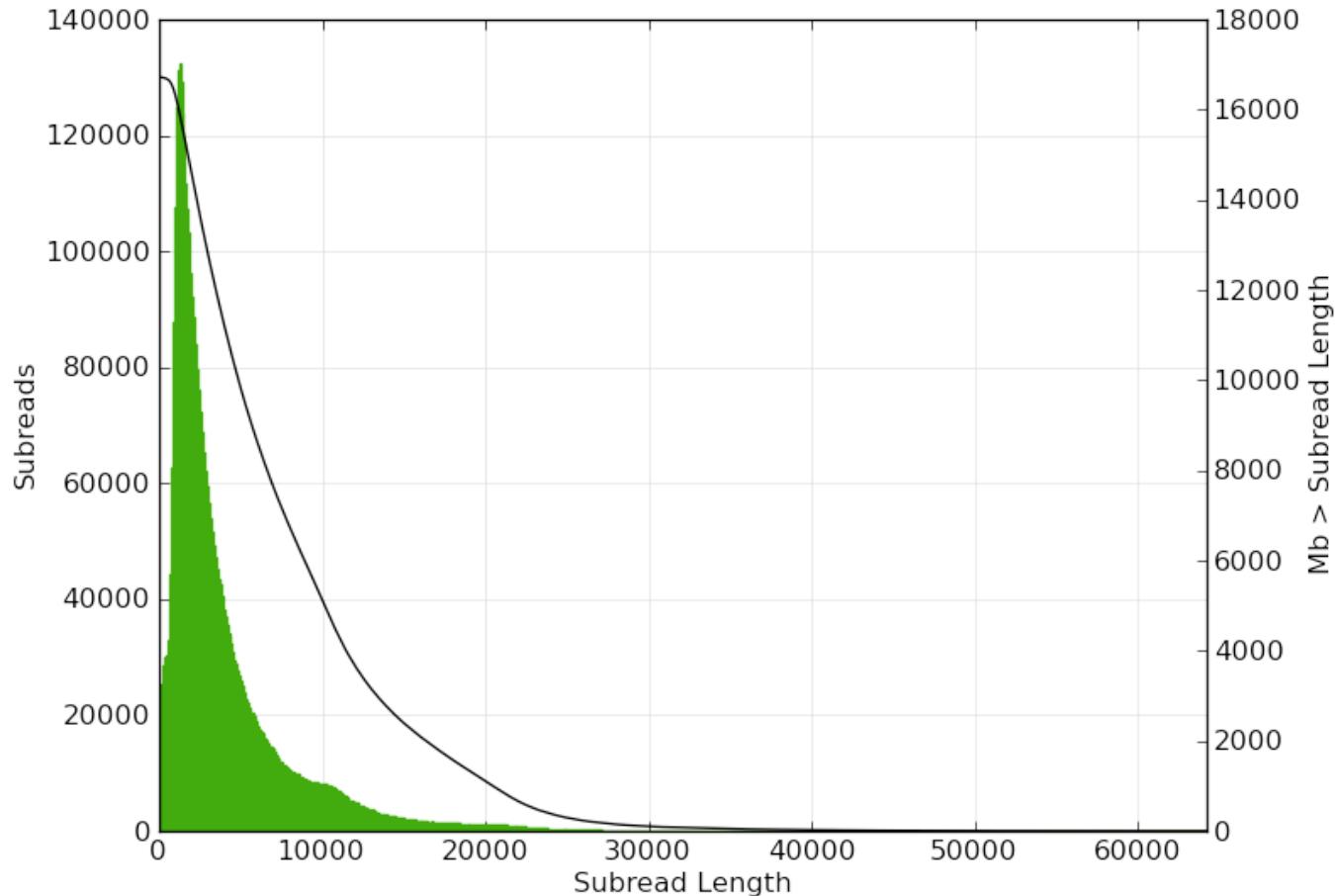


Error correction: mapping high-quality
short PacBio reads to long reads

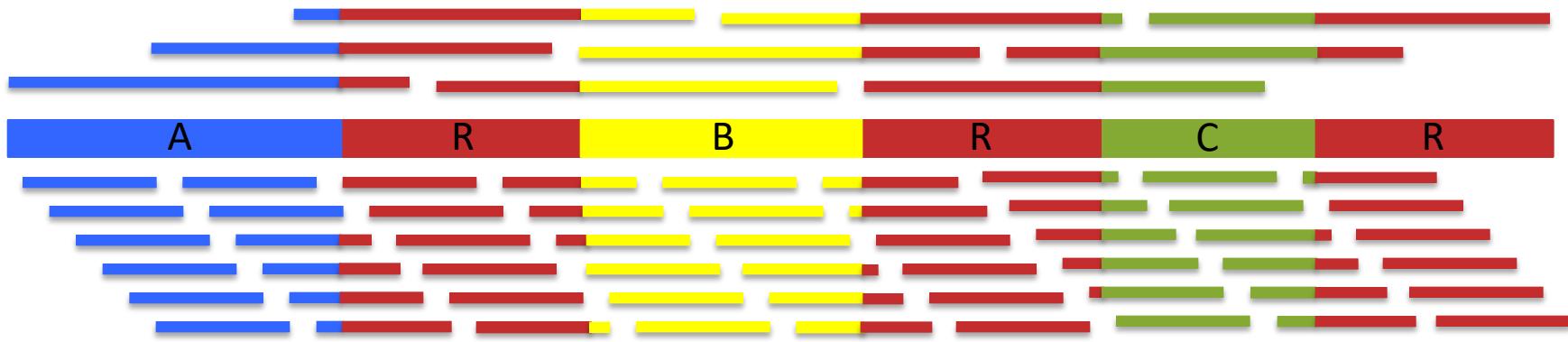
Assembly polishing with raw reads

Long read sequencing

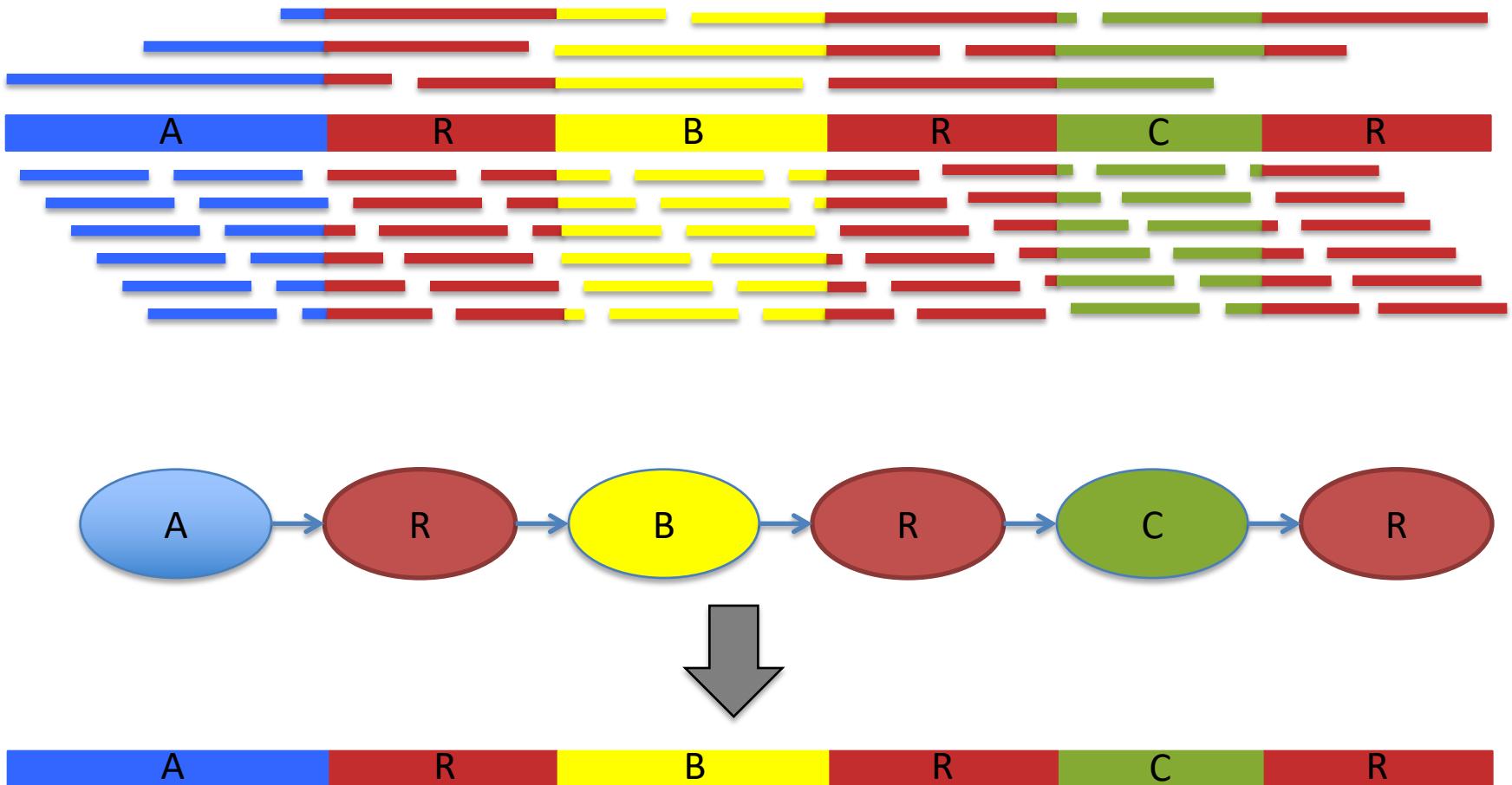
PacBio filtered subreads



Long Read Assembly - Complexity



Assembly Complexity



The advantages of SMRT (Single Molecule Real Time) sequencing

Sequencing Technologies

- Combine technologies to improve assemblies.
Long reads with short reads
- Hybrid assemblies
 - 80x Illumina 2x150bp PE - contig building
 - Illumina long insert 500bp library - scaffolding
 - Dovetail Genomics - contig ordering & orientation
 - 20x Pac Bio - gap-filling, fill in N's between scaffolds
- When are you finished? Gap-filling?
- How easy is it....really?

File formatting

- FASTA
- FASTQ
 - quality scores
- SAM/BAM
 - Sequence Alignment Map
 - Stores alignment information

<https://genome.ucsc.edu/FAQ/FAQformat.html>

FASTA

>M63509.1 Human glutathione transferase M2 (GSTM2)

CGCAGCAACCAGCACCATGCCATGACACTGGGTACTGGAACATCCGGCTGGCCATTCCA
TCCGCCTGCTCCTGGAATACACAGACTCAAGCTACGAGGAAAAGAAGTACACGATGGGGACGCT
CCTGATTATGACAGAACAGCAGTGGCTGAATGAAAAATTCAAGCTGGCCTGGACTTTCCAATCT
GCCCTACTTGATTGATGGGACTCACAAAGATCACCCAGAGCAATGCCATCCTGCGGTACATTGCC
GCAAGCACAACCTGTGCGGGAAATCAGAAAAGGAGCAGATTCGGAAGACATTGGAGAACCAAG
TTTATGGACAGCCGTATGCAGCTGCCAAACTCTGCTATGACCCAGATTTGAGAAACTGAAACC
AGAATACCTGCAGGCACTCCCTGAAATGCTGAAGCTCTACTCACAGTTCTGGGAAGCAGCCAT
GGTTTCTGGGGACAAGATCACCTTGTGGATTCATCGCTTATGATGTCCTTGAGAGAAACCAA
GTATTGAGCCCAGCTGCCCTGGATGCCCTCCAAACCTGAAGGACTTCATCTCCGATTTGAGGG
CTTGGAGAAGATCTCTGCCATCATGAAGTCCAGCCGCTCCTCCAAGACCTGTGTTACAAAGA
TGGCTGTCTGGGCAACAAGTAGGGCCTTGAAGGCAGGAGGTGGAGTGAGGAGCCATACTCAG
CCTGCTGCCAGGCTGTGCAGCGCAGCTGGACTCTGCATCCCAGCACCTGCCCTCGTTCTT
CTCCTGTTATTCCATCTTACTCCAAGACTTCATTGTCCTCTTCACTCCCCCTAAACCCCT
GTCCCATGCAGGCCCTTGAAGCCTCAGCTACCCACTATCCTCGTAACATCCCCTCCATCAT
TACCCCTCCCTGCACTAAAGCCAGCCTGACCTCCCTGTTAGTGGTTGTCTGCTTAAAG
GGCCTGCCTGGCCCTGCCCTGGAGCTCAGCCCCGAGCTGTCCCCGTGTTGCATGAAGGAGCAG
CATTGACTGGTTACAGGCCCTGCTCCTGCAGCATGGTCCCTGCCTAGGCCTACCTGATGGAAG
TAAAGCCTCAACCACAAAAAAAAAA

FASTQ

```
@M00747:32:00000000-A16RG:1:1112:15153:29246 1:N:0:1
TCGATCGAGTAACTCGCTGCTGTCAGACTGGTTTGGTCGACTATTGTTCAGTCGCAAGAAT
ATTGTGTCCAGTCGACTGAATTCTGCTGTACGGCCACGGCGGATGCACGGTACAGCAGGCTCAG
ACGGATTAAACTGTT
+
5=9=9<=9,-5@<<55>,6+8AC>EE.88AE9CDD7>+7.CC9CD+++5@=-FCCA@EF@+**+--
55--AA---AA-5A<9C+3+<9)4++=E====<D94)00=9)))2@624(/(/2/-
(.(6;9((((.(.'((6-66<6///
@M00747:32:00000000-A16RG:1:1112:15536:29246 1:N:0:1
GTAAAATTGAGGTAAATTGTGCGGAATTAGCAATACCGTTTTTATTATCACCGGATATCTATT
TGCTGTACGGCCAAGGAGGATGTACGGTACAGCAGGTGCGAACTCACTCCGACGCTCAAGTCAGTGAC
TTAATGATAAGCGTG
+
?????<BBBBBBB5<?BFFFFFFECHEFFFECCFF?9AAC>7@FHHHHHHFG?EAFFG@EEDEHHDGHHC
BDFFGDFHF)<CCD@F,+3=CFBDFHBD++??DBDEEEDE:):CBEEEBCE68>?) )5?**0?:AE*A
*0//:/*:*:**:*.0)
@M00747:32:00000000-A16RG:1:1112:15513:29246 1:N:0:1
GCTAGTCTGTGTTAGTTATGTTTGCATGTTGTAACGGATTCAAACATAGGTGTTGTTCT
TTTATGGTTGTACAATTGGCCCTAACGCCCTACACTTACTGTTGTTCTTATGGTACGACAT
TTGAGTGGTGGTTGA
+
```

SAM/BAM Sequence Alignment Map

```
D4ZHLFP1:53:D2386ACXX:6:2115:17945:68812      0      Mle_000001      18      42      108M *      0      0
       TCCCCCTGCATGTGTCGTCTGGCTGGATGCCATGCTCCATGCAGTATAGCTCCAGCATGAGTTACCGATCTGGACACCTGCTTG
GCCAAGATGTACTGAGATGCAT
C@CFDFFFHHGHHFGBFEggDGggGEHGiGGGJJJIIIGIIB9BFBBFHGGHICEAGHGEGEDHIGEEDBECCACBDDC@CCDBCDD<
?2+4>@4>>CCCCAA@@      AS:i:-5      XN:i:0      XM:i:1      XO:i:0      XG:i:0      NM:i:1      MD:Z:0A107
       YT:Z:UU

D4ZHLFP1:53:D2386ACXX:7:2110:5214:83081 0      Mle_000001      18      42      108M *      0      0
       TCCCCCTGCATGTGTCGTCTGGCTGGATGCCATGCTCCATGCAGTATAGCTCCAGCATGAGTTACCGATCTGGACACCTGCTGGCAA
GATGTACTGAGATGCAT
CCCFFFFFHHHHHGGEGIJIIIGJFHJJJJIIIGIJIJFHJJIIJJFIIIIIIJIIJJHHFFFCEEEEDDDDDDDDDDDDDDDDD
BCDCDEEEEDDDDDDDDD      AS:i:-5      XN:i:0      XM:i:1      XO:i:0      XG:i:0      NM:i:1      MD:Z:0A107
       YT:Z:UU

D4ZHLFP1:53:D2386ACXX:7:2206:9985:31556 0      Mle_000001      18      42      108M *      0      0
       TCCCCCTGCATGTGTCGTCTGGCTGGATGCCATGCTCCATGCAGTATAGCTCCAGCATGAGTTACCGATCTGGACACCTGCTGGCAA
GATGTACTGAGATGCAT
CCCFFFFFHHHHJJJIHJJIIIIJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
DDCD@@CDCCDDCDCDC      AS:i:-5      XN:i:0      XM:i:1      XO:i:0      XG:i:0      NM:i:1      MD:Z:0A107
       YT:Z:UU
```

Genome assemblers

- Some useful assemblers:

Illumina data:

w2rap-contigger

<https://github.com/bioinfologics/w2rap-contigger>

- can take lots of computer power to run
- works with single paired-end library

soapdenovo2

<https://sourceforge.net/projects/soapdenovo2>

- relatively easy to install and run
- works with large genomes

Genome assemblers

PacBio data:

falcon & falcon-unzip

<https://github.com/PacificBiosciences/FALCON>

<http://profs.scienze.univr.it/delledonne/Papers/2016%20Chin%20NMethods.pdf>

Very powerful assemblers but not easy to install or use.

Quiver/Arrow/pbalign - genome alignment and polishing. Part of PacBio Genomic Consensus package.

<https://github.com/PacificBiosciences/GenomicConsensus>

Again, powerful but not the easiest to use.

Canu - evening workshop

<http://genome.cshlp.org/content/27/5/722>

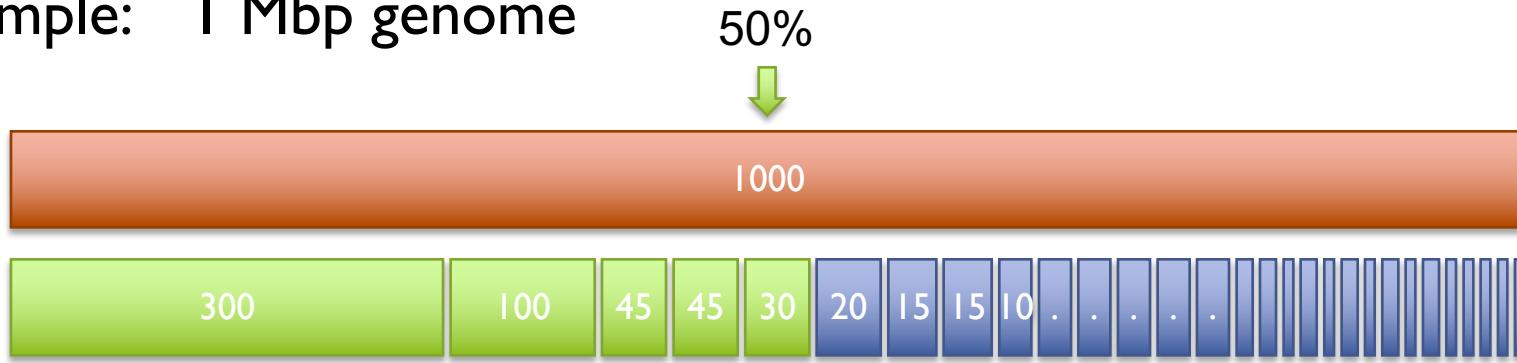
Genome Assembly

- Recommended to use multiple assemblers with different parameters to assess results
- How to assess our results?
 - Number of contigs/scaffolds
 - Longest contig/scaffold
 - N50 - percentage of the genome in contigs
 - L50 - number of contigs that are as long or longer than the N50 value

N50 size

50% of the genome in contigs as long as or larger than N50 value

Example: 1 Mbp genome



N50 size = 30 kbp

$$(300k + 100k + 45k + 45k + 30k = 520k \geq 500\text{kbp})$$

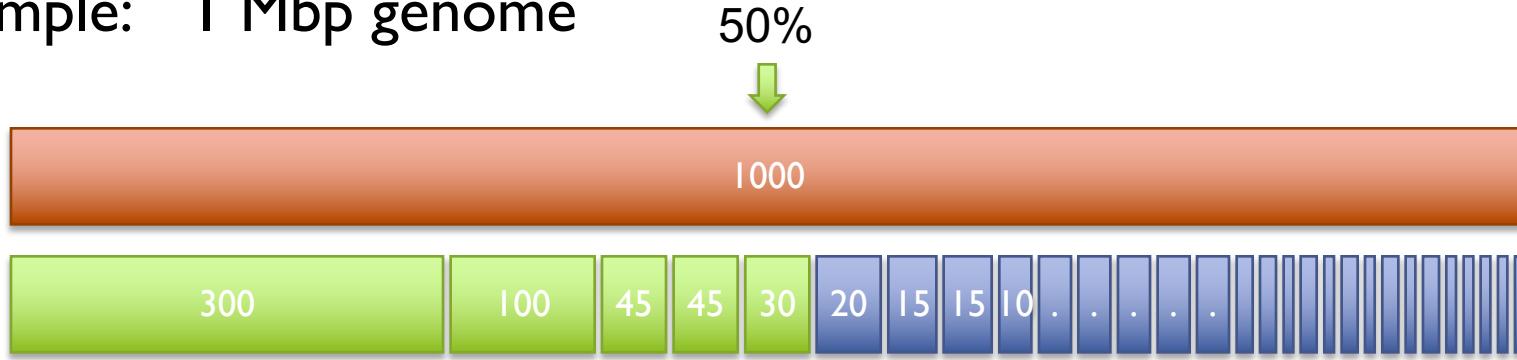
A greater N50 is usually a sign of assembly improvement

- Comparable with genomes of similar size
- Genome composition can bias comparisons
- Low L50 vs High N50

L50 size

Number of contigs that are as long or longer than the N50 value

Example: 1 Mbp genome



N50 size = 30 kbp

$$(300k + 100k + 45k + 45k + 30k = 520k \geq 500\text{ kbp})$$

L50 - number of contigs that sum to N50 length

L50 = how many?

- Low L50 vs High N50
 - longer sequences and fewer of them...in theory
 - lower stringency can inflate N50

Genome assembly quality

- How to assess our results?

Alignments:

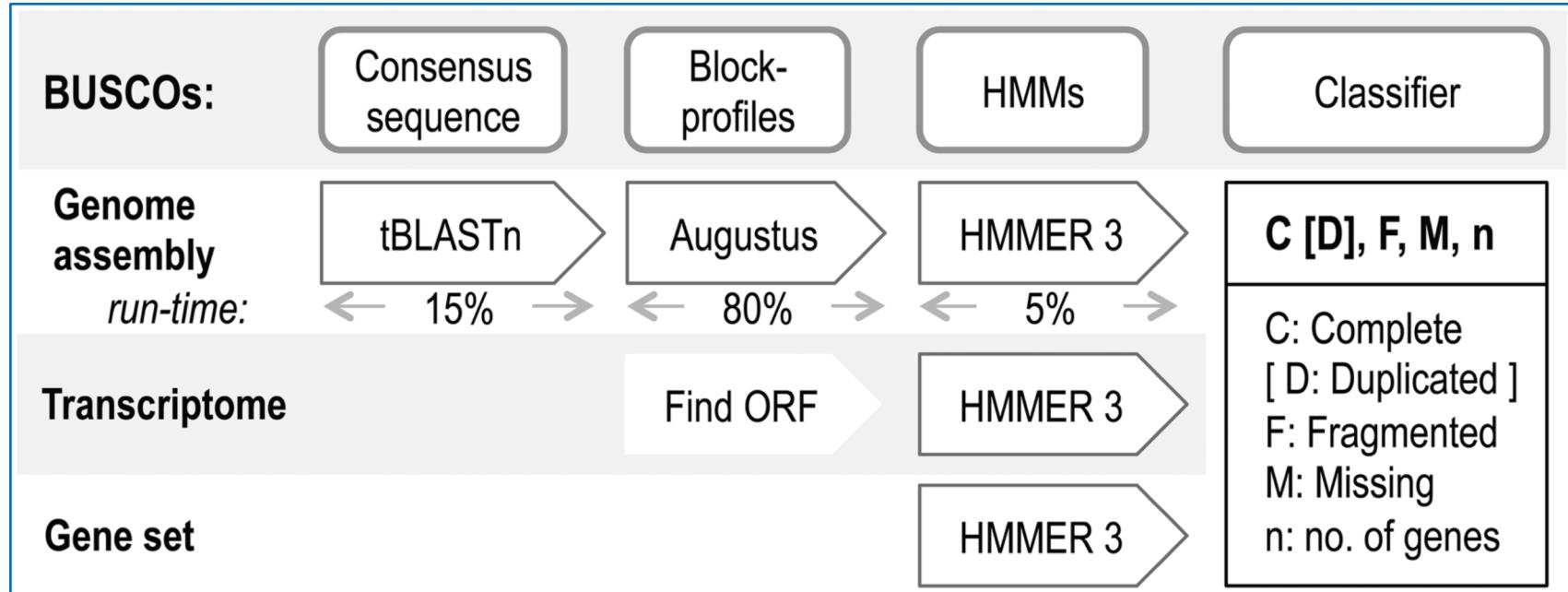
- compare to a reference genome
- align reads to your assembled genome
- assess repetitive regions
- Call SNPs

-Check for completeness:

- annotation, blast against reference gene set
- Busco - Simao et al. 2015. Bioinformatics

BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs

Bioinformatics. 2015;31(19):3210-3212. doi:10.1093/bioinformatics/btv351



BUSCO assessment workflow and relative run-times

Quality of genome vs. completeness

Genome assembly quality

- Do not take first version. Try correcting/polishing your genome.
 - quiver/arrow
 - pilon
- Go back and reassess your genome
 - Number of contigs/scaffolds/unitigs change?
 - L50 go up or down?
 - How much is fragmented?
 - Always new technologies and improvements

Assembly Summary

Assembly quality depends on

1. **Experimental design:** clear and organized with high-quality DNA, if possible
2. **Coverage:** Aim for high coverage
3. **Repeat composition:** high repeat content can be a challenge
4. **Read length:** incorporate some longer reads for scaffolding and help resolve repeats
5. **Error rate:** errors reduce coverage, obscure true overlaps. Error correction for long read assemblies
 - Assembly is a hierarchical, starting from individual reads, build contigs, incorporate long reads to join gaps, decrease fragmentation
 - Extensive error correction is the key to getting the best assembly possible from a given data set

What should we expect from an assembly?

- Annotation of assembly
- Comparison to closely related genomes
- Gene content
- Percent repetitive
- Another estimate
 - Flow cytometry

Genome assembly workshop

- Genome assembly with PacBio data using Canu assembler
- Python programming exercise