Genome Sequencing & Assembly

Deb Triant
University of Florida
Museum of Natural History
University of Virginia
Dept. Biochemistry & Molecular Genetics

25 March 2017
Programming for Biology
Cold Spring Harbor, NY

Lecture outline

1. 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

*This evening - assembly workshop

Lecture outline

1. 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

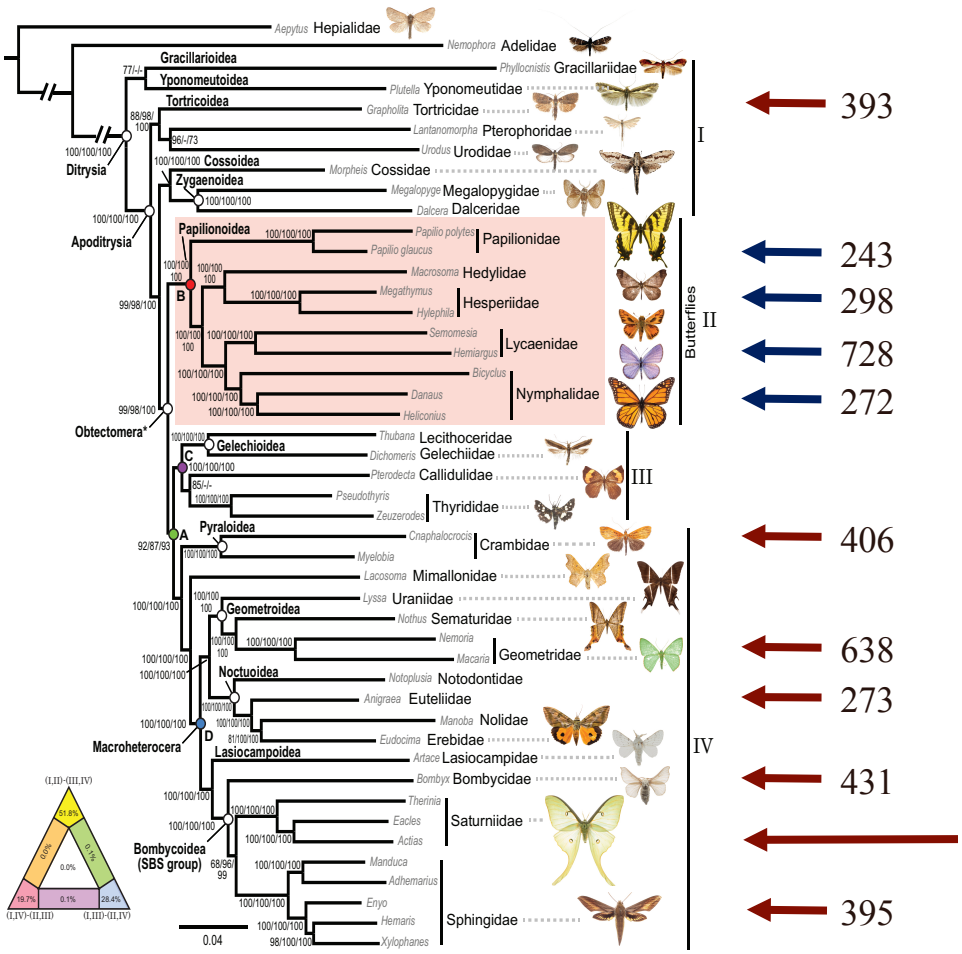
*Evening - assembly workshop

Lepidopteran genome projects

Automeris io

Actias luna

- Giant silk moths - Superfamiy Bombycoidea
- Over 2,000 described species
- Some species with well-developed eyespots
- Reduced proboscis or absent

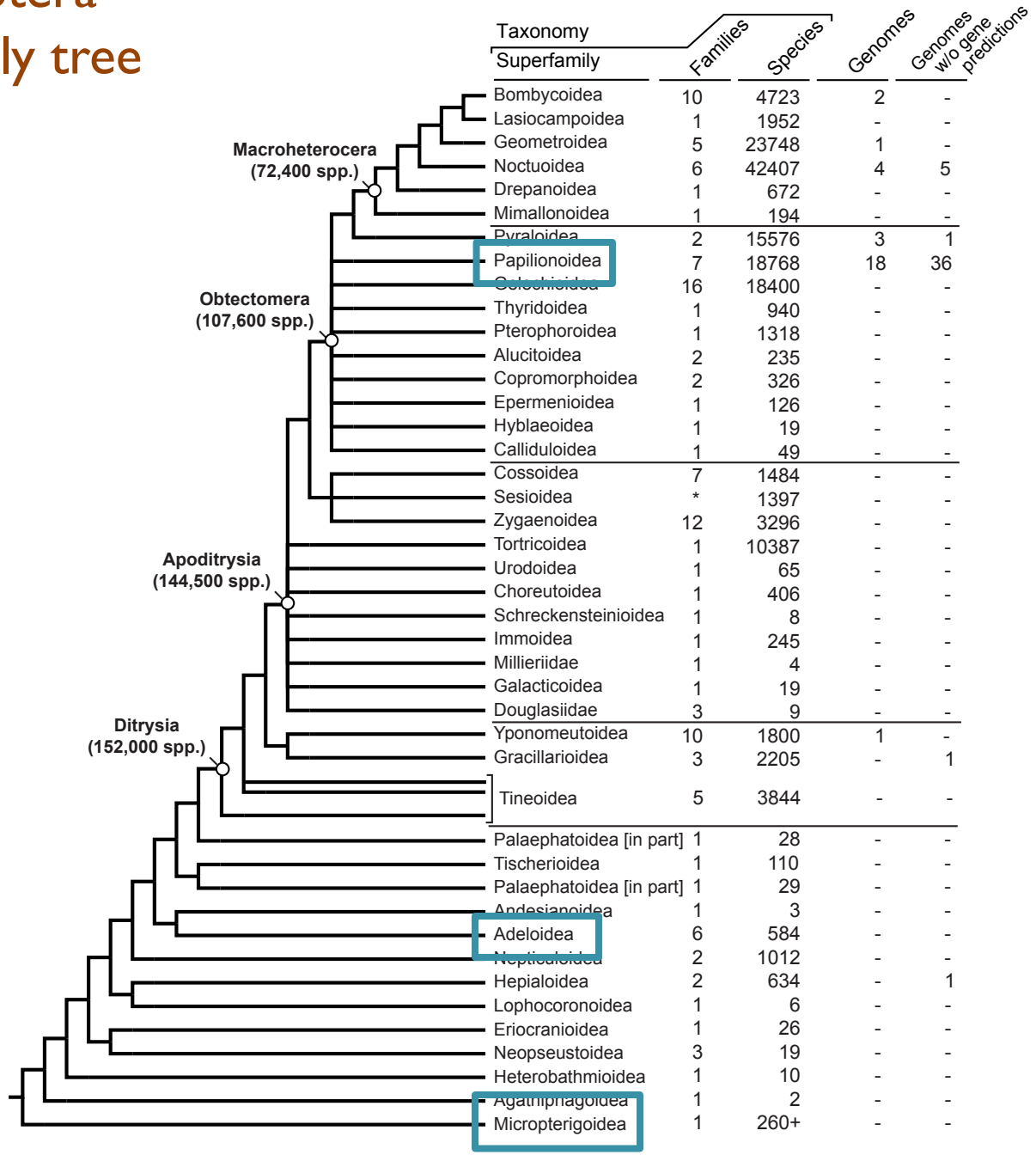


Kawahara and Breinholt (2014): *Proc. R. Soc. B*

Lepidoptera Genomes

- ~25 assemblies (LepBase, NCBI)
  - 12 families; 8 moth species
  - 275 - 725 Mb

Lepidoptera superfamily tree



Taxonomy Superfamily	Families	Species	Genomes	Genomes w/o gene predictions
Bombycoidea	10	4723	2	-
Lasiocampoidea	1	1952	-	-
Geometroidea	5	23748	1	-
Noctuoidea	6	42407	4	5
Drepanoidea	1	672	-	-
Mimallonoidea	1	194	-	-
Pyraloidea	2	15576	3	1
Papilionoidea	7	18768	18	36
Geloechioidea	16	18400	-	-
Thyridoidea	1	940	-	-
Pterophoroidea	1	1318	-	-
Alucitoidea	2	235	-	-
Copromorphoidea	2	326	-	-
Epermenioidea	1	126	-	-
Hyblaeoidea	1	19	-	-
Calliduloidea	1	49	-	-
Cossoidea	7	1484	-	-
Sesioidea	*	1397	-	-
Zygaenoidea	12	3296	-	-
Tortricoidea	1	10387	-	-
Urodoidea	1	65	-	-
Choreutoidea	1	406	-	-
Schreckensteinioidea	1	8	-	-
Immoidea	1	245	-	-
Millieriidae	1	4	-	-
Galacticoidea	1	19	-	-
Douglasiidae	3	9	-	-
Yponomeutoidea	10	1800	1	-
Gracillarioidea	3	2205	-	1
Tineoidea	5	3844	-	-
Palaephatoidea [in part]	1	28	-	-
Tischerioidea	1	110	-	-
Palaephatoidea [in part]	1	29	-	-
Andesianoidea	1	3	-	-
Adeloidea	6	584	-	-
Nepticuloidea	2	1012	-	-
Hepialoidea	2	634	-	1
Lophocoronoidea	1	6	-	-
Eriocranioidea	1	26	-	-
Neopseustoidea	3	19	-	-
Heterobathmioidea	1	10	-	-
Agathiphagoidea	1	2	-	-
Micropterigoidea	1	260+	-	-

Lepidopteran genome projects

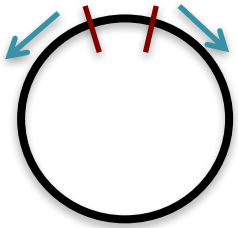
Cyclargus thomasi - Miami blue, Florida

Tuta absoluta – Tomato leafminer
Latin America: Europe N. Africa, Mid East

Sequencing Silene

1. 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

1. 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."

The Genome Plant

R.I.P.

Sequencing Silene

1. 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

"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

1. My background & introduction to genome assembly

2. General background of genome assembly & theory

3. Comparison of assembly methods

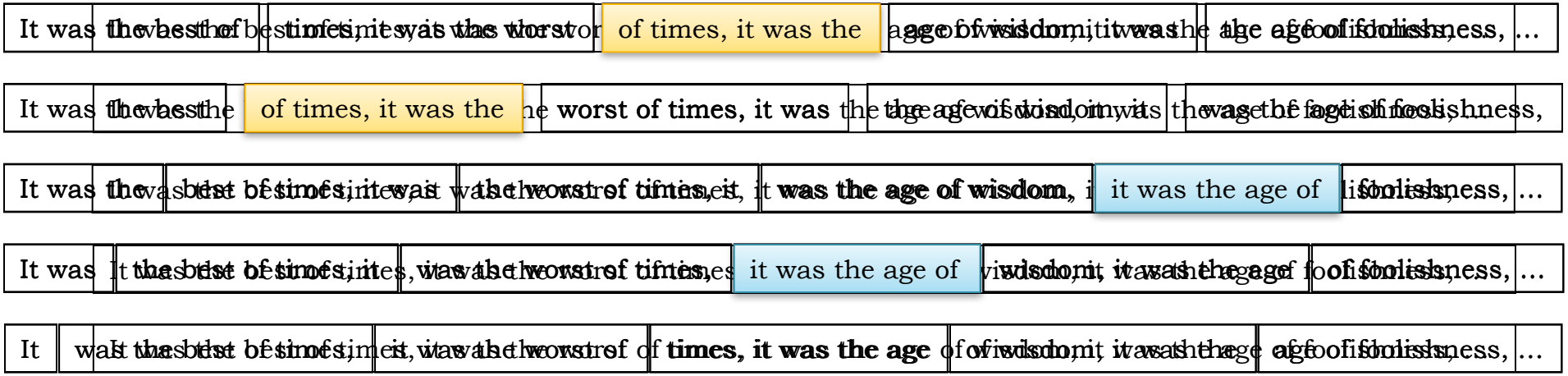
4. Recommendations for a good assembly project

*Afternoon - assembly workshop

Shredded Book Reconstruction

**Based on example by Michael Schatz

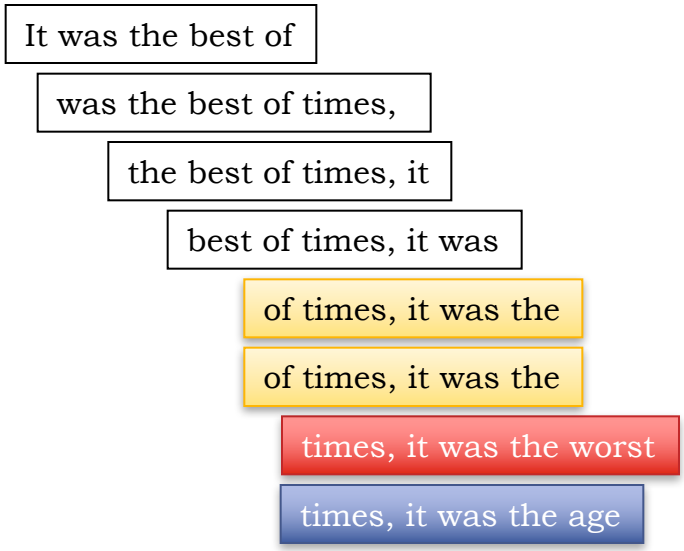
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



- How can he reconstruct the text?
  - 5 copies x 138, 656 words / 5 words per fragment = 138k fragments
  - The short fragments from every copy are mixed together
  - Some fragments are identical

Shredded Book Reconstruction

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



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 = All length-k subfragments (k < l)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words

Original Fragment

It was the best of

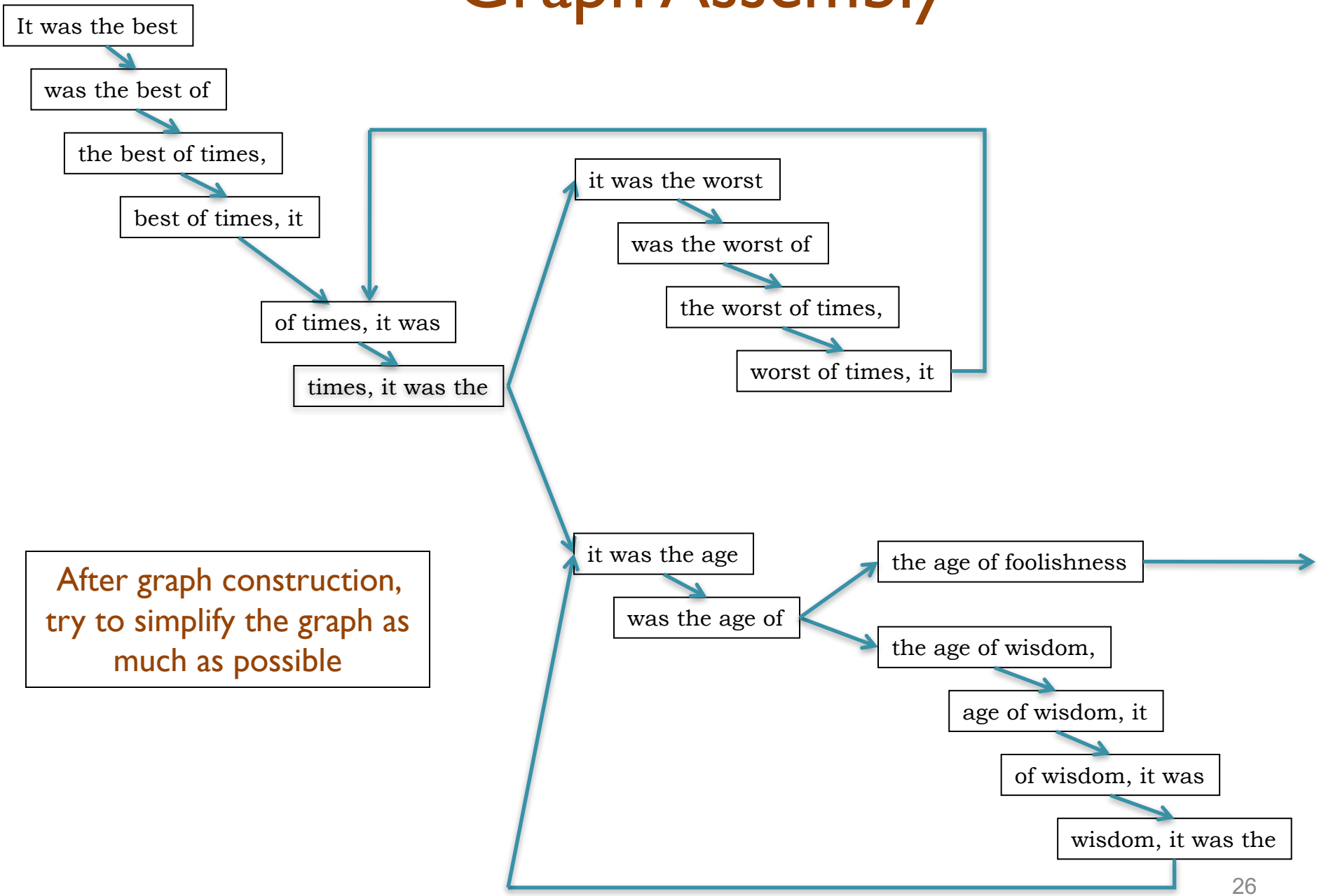
Directed Edge

It **was the best** → **was the best** of

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

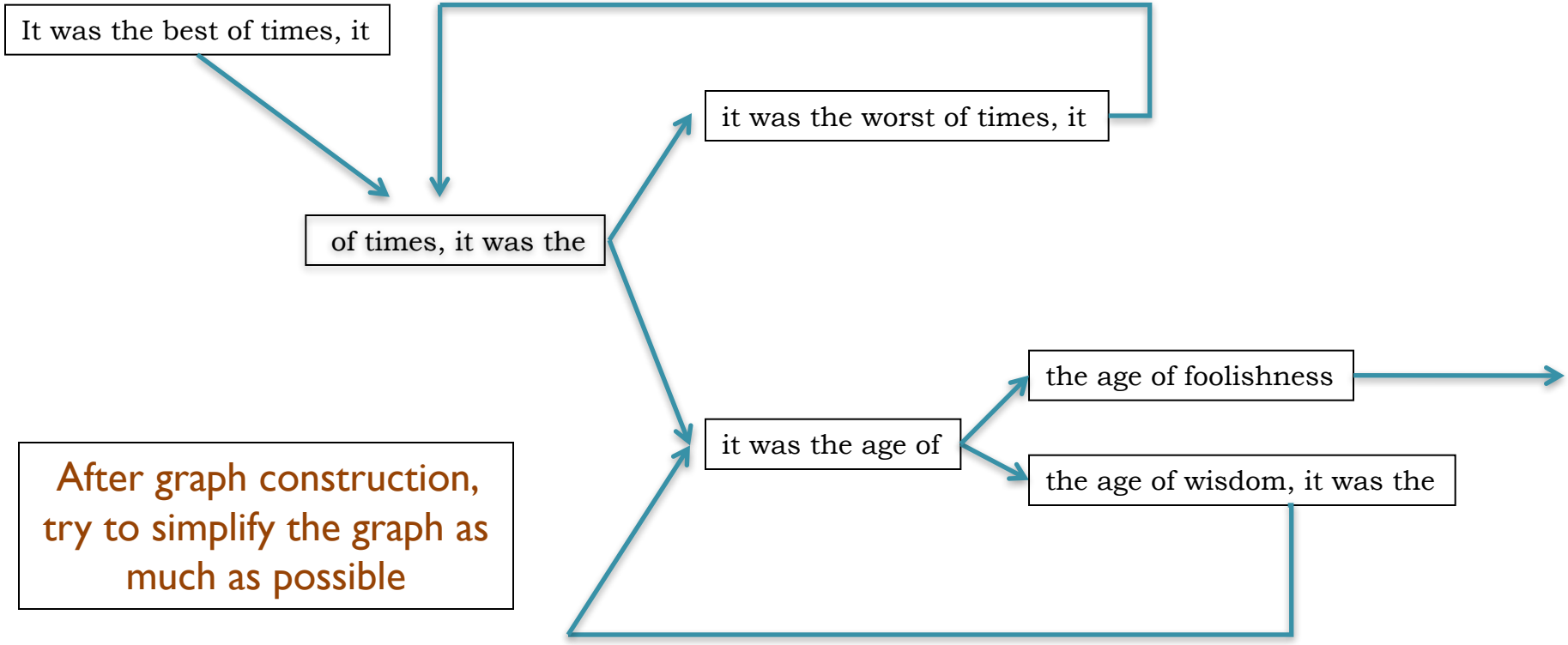
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

Graph Assembly



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

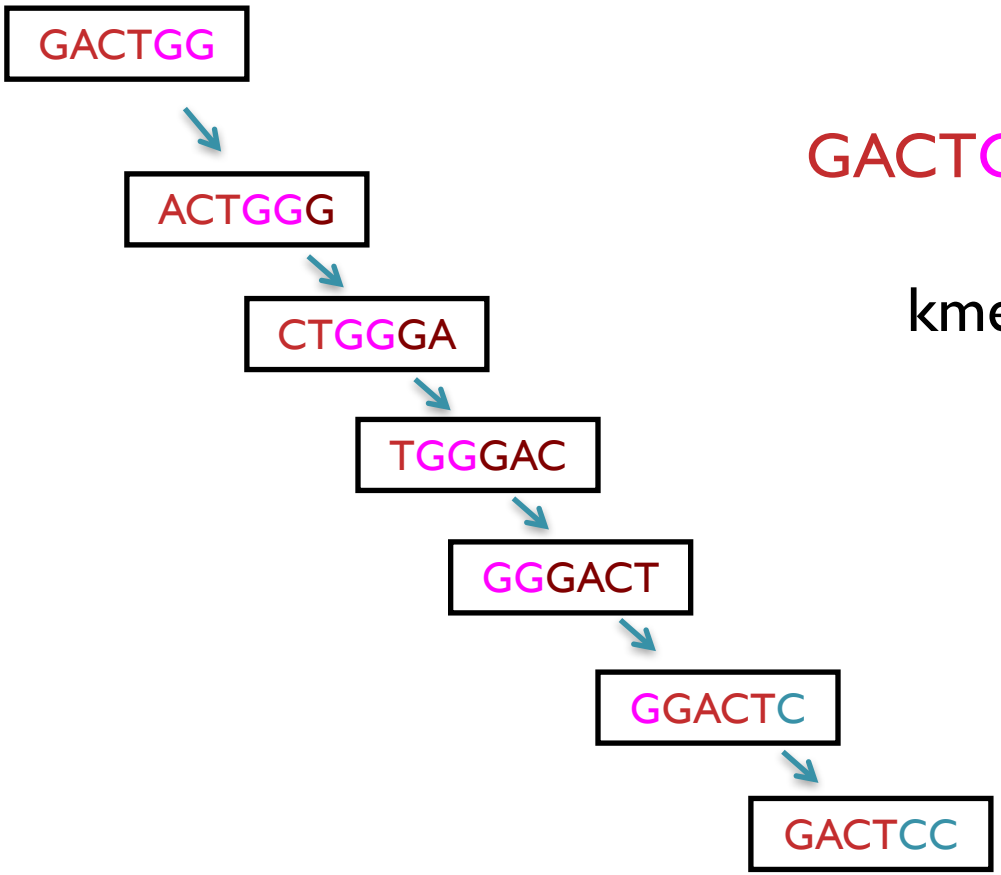
Graph Assembly



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

Graph Assembly

- Shredded words → k-mer



GACTGGGACTCC

kmer n = 6

The full tale

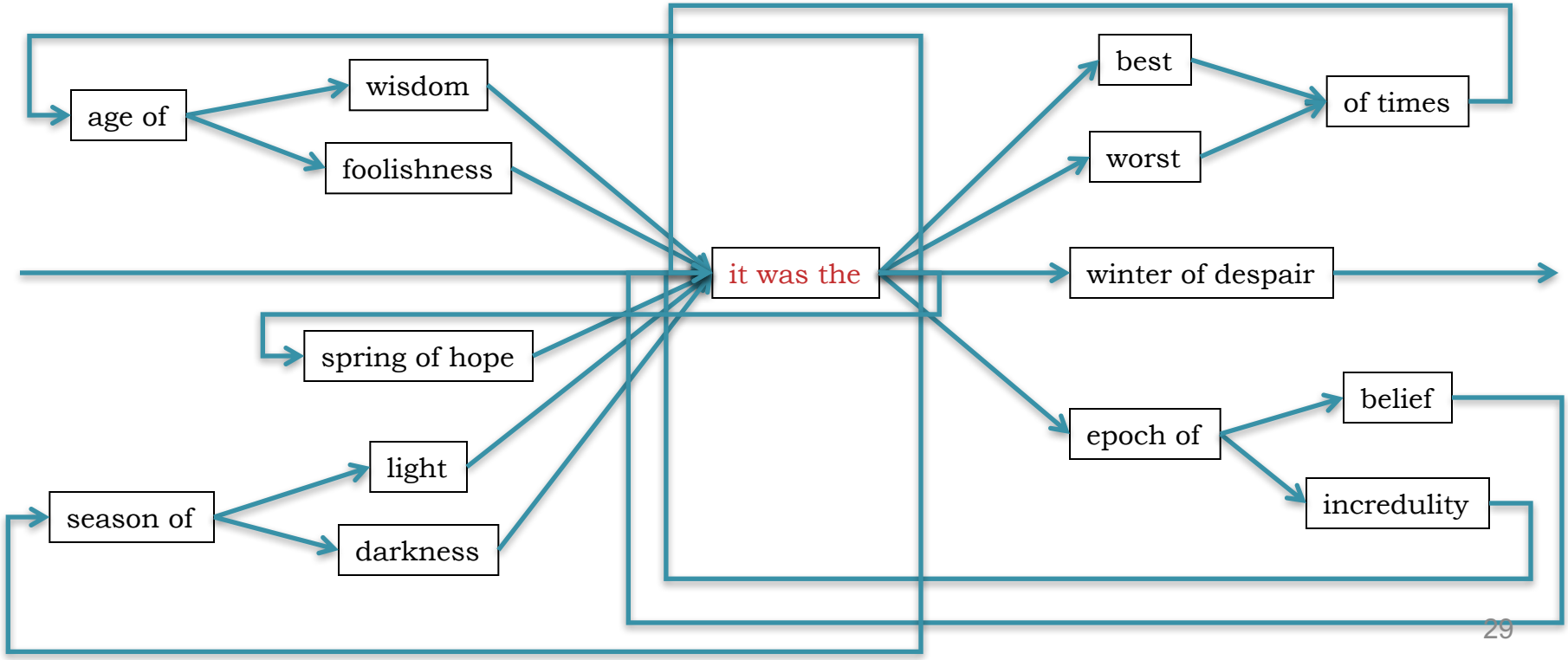
… it was the best of times it was the worst of times …

… 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. *Drosopjhila 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

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

GENOME 10K©

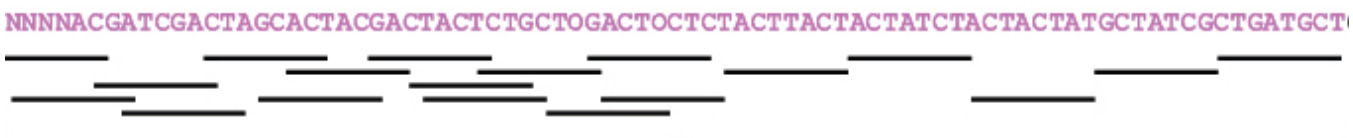
i5K

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

Recipe for a good assembly

- Coverage
  - How many times has genome been sequenced?
  - Too much? Too little? Aim for oversampling ~40-60x. 100x
  - Much is driven by funding – (# reads, read length, genome size)



- Read Length
  - Read lengths must be longer than repetitive regions

- Quality
  - reads assembled by shared regions
  - Find kmers shared in pair of reads to assemble

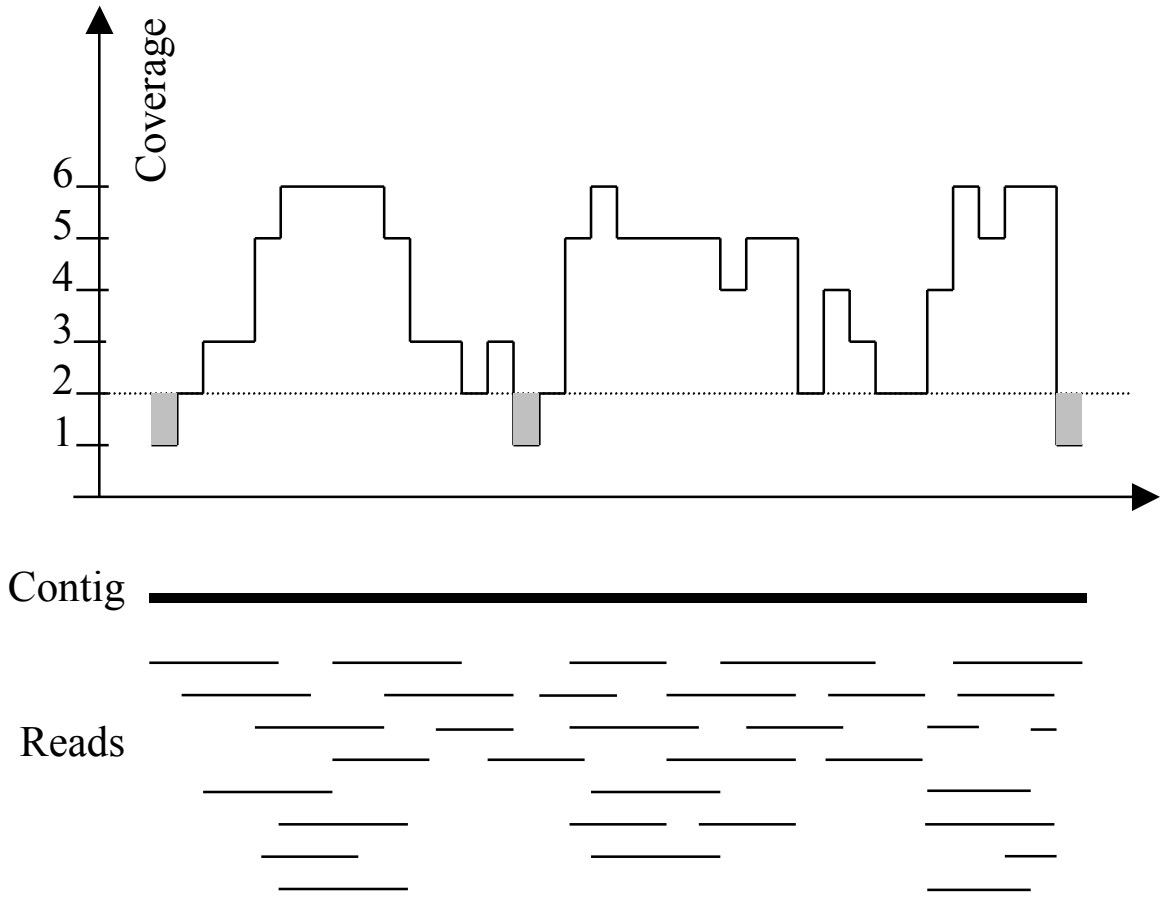
Calculating coverage

- *Genome size*: $2 * 10^9$

- Max read length: $150 * 10^6$ (HiSeq lane)
- Read length: 100 nucl * 2 (Paired-End)
  - $3 * 10^{10}$ (15X coverage) Goal: 80X = ~5 - 6 lanes

- DNA requirement projections
  - High quality DNA!!! **often our biggest limitation
  - number and type of libraries required
  - potential projects resulting from assembly

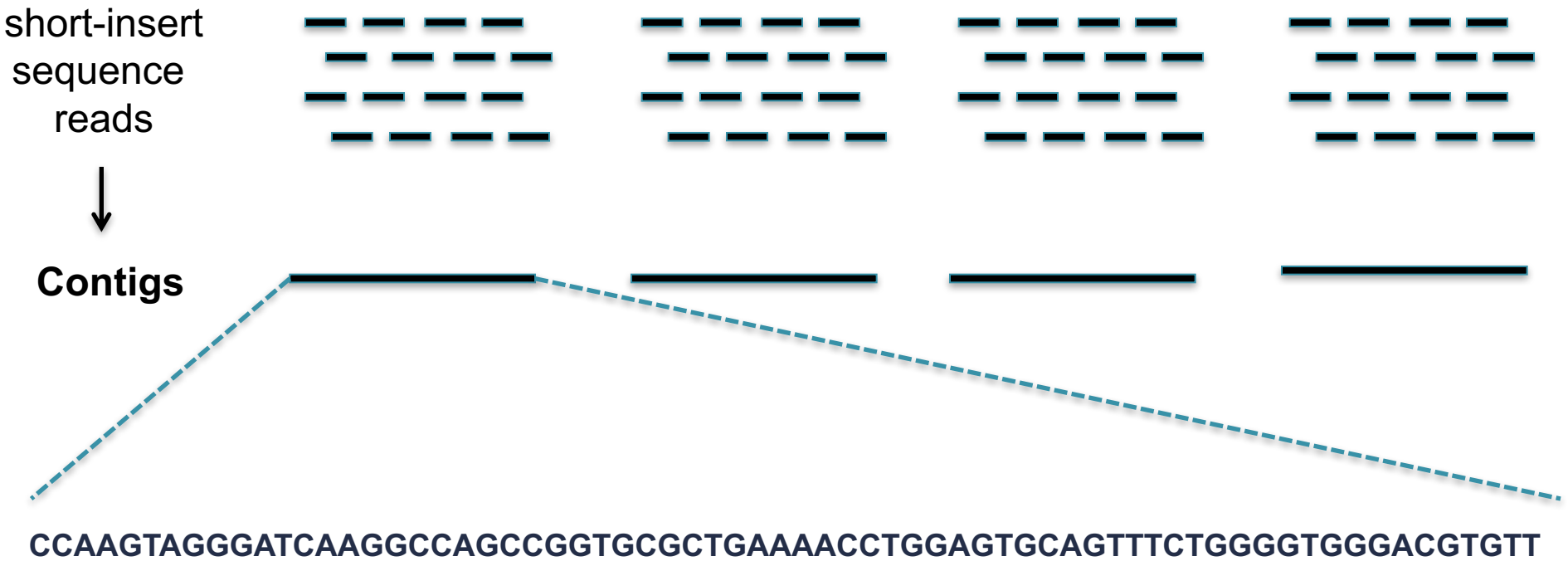
Typical sequencing coverage



Imagine raindrops on a sidewalk

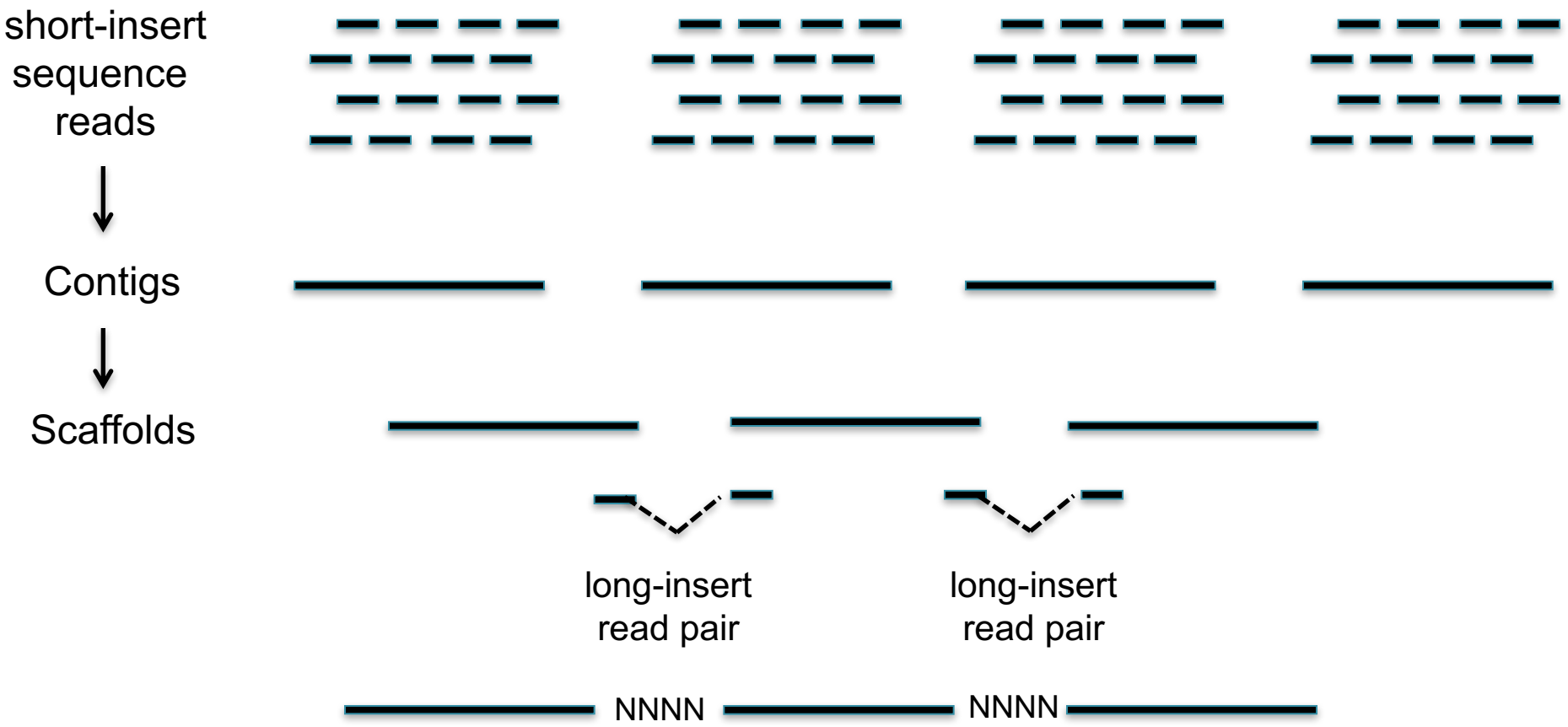
*Slide from Michael Schatz

Assembly construction – Hierarchical process



Contig - contiguous sequence of nucleotides

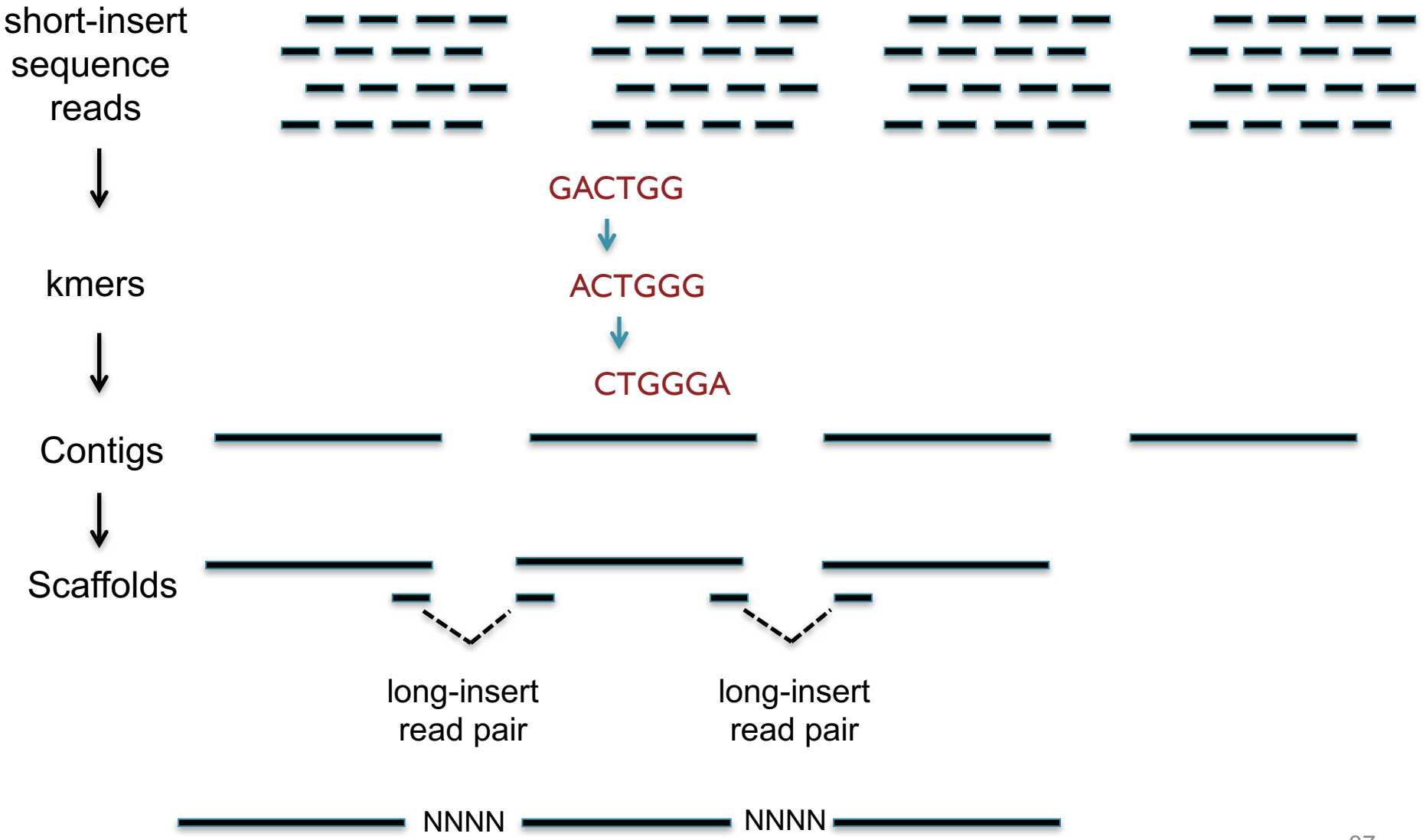
Assembly construction



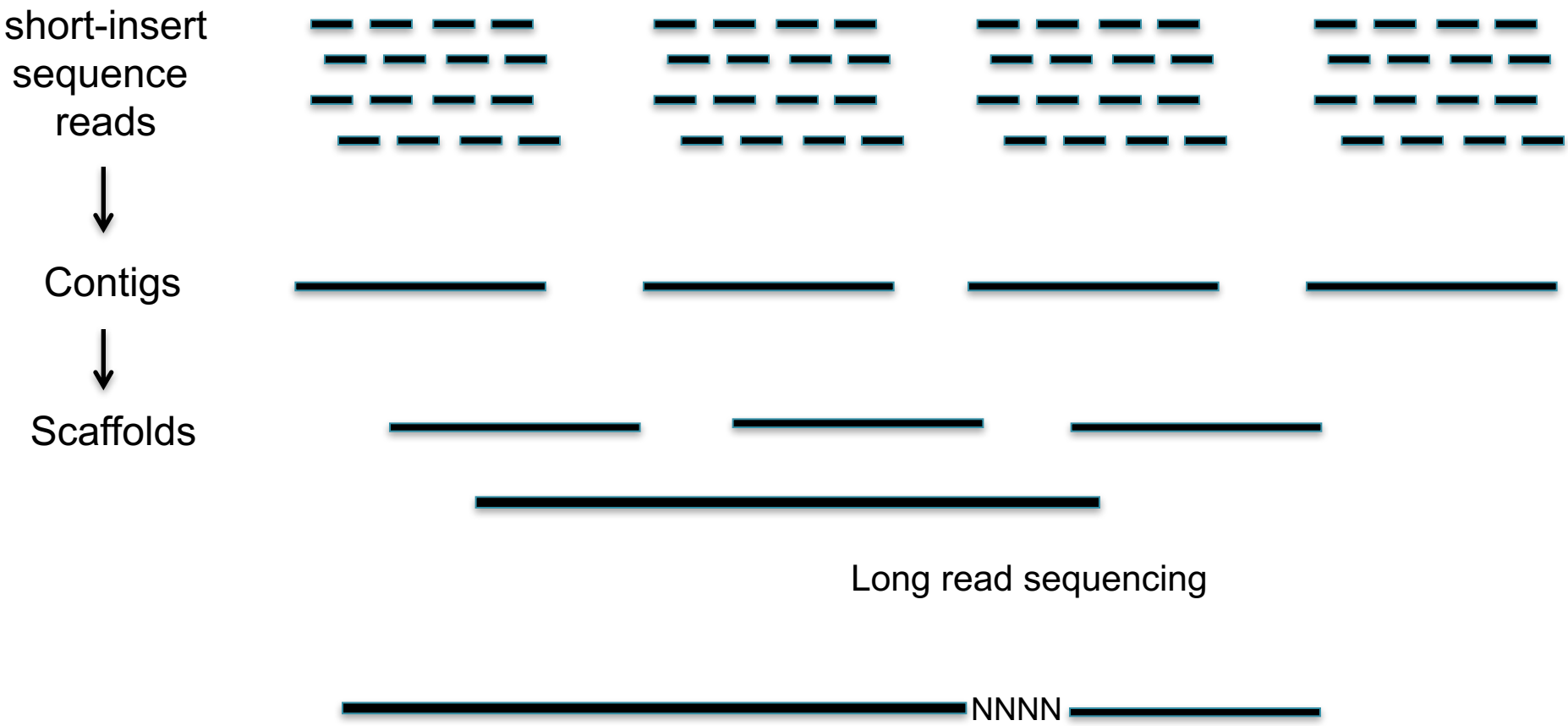
Scaffold - sequence of contigs, separated by gaps - Ns are predicted gap size

>GTAGTATTCTAGAAAATGTTTTAAATATAGATAGTTGTTTTANATCTGTTAGTTGTCAGATGCTTACTGAATAGTTGGAANNNNNNNNNNNNNNNNNN
NNNNNNTGTGAGGTTTTTAGCTCATGAAAGTTTATGATTATTGCACCCCTACTCACAACGAATCCCTATTCTTATCTTTTNNNNNNNNNNNNNNNCATGT
CACTGGTTTATTTATTTTTTGTGGCTGCAGAAGTCCTTTGTCGCTGTTAATTTTTTGGGAGTTCTCCTGTCGTATATTAAAGCTTTCTTTCTTTTCAGTTT
TAAATTTATTTGAACCTTACTATCTTTCTAACAATAAATTGTGGAATTATCAACGAAAACATAGGNNNNNNNNNNNGTCCTTTATACGAAAGCTATATAG
TGTTAGGCTTTTCTTTTTTTTTNNNNNNNNGGTGATGTTGTTAATGGTGCCCTTTTCTGGTAATCTTACTAAATCAGTTTGCTTGTTACTTGTATAGTTGT

Assembly construction



Assembly construction



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

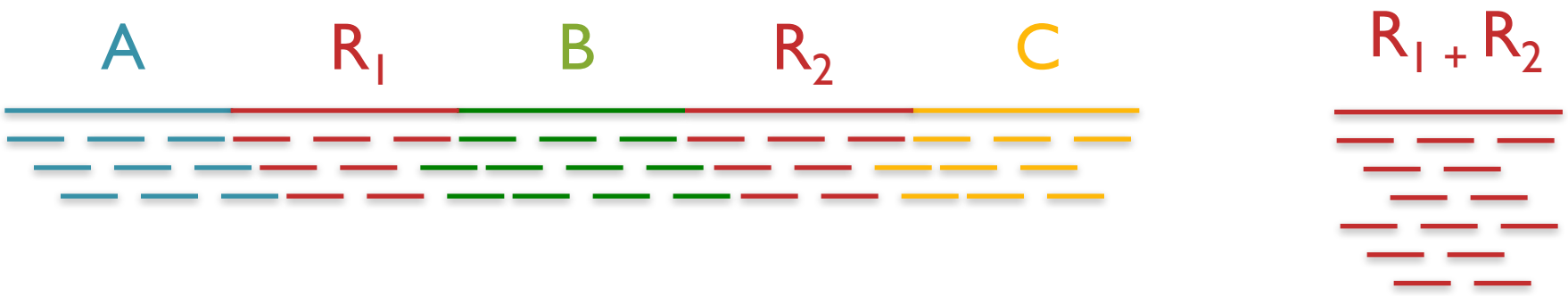
>GTAGTATTCTAGAAAATGTTTTAAATATAGATAGTTGTTTTANATCTGTTAGTTGTCAGATGCTTACTGAATAGTTGGAANNNNNNNNNNNNNNNNNN
NNNNNNTGTGAGGTTTTTAGCTCATGAAAGTTTATGATTATTGCACCCCTACTCACAACGAATCCCTATTCTTATCTTTTNNNNNNNNNNNNNNNCATGT
CACTGGTTTATTTATTTTTTGTGGCTGCAGAAGTCCTTTGTCGCTGTTAATTTTTTGGGAGTTCTCCTGTCGTATATTAAAGCTTTCTTTCTTTTCAGTTT
TAAATTTATTTGAACCTTACTATCTTTCTAACAATAAATTGTGGAATTATCAACGAAAACATAGGNNNNNNNNNNNGTCCTTTATACGAAAGCTATATAG
TGTTAGGCTTTTCTTTTTTTTTNNNNNNNNGGTGATGTTGTTAATGGTGCCCTTTTCTGGTAATCTTACTAAATCAGTTTGCTTGTTACTTGTATAGTTGT

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.

The fragment assembly string graph
Myers, EW (2005) Bioinformatics. 21(suppl 2): ii79-85.

Lecture outline

1. 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 workshop

Sequencing Technologies

- Sanger
  - 800 bp reads with low error rate, costly
- Illumina
  - 800 bp reads with low error rate, costly
  - Paired-End, Long-insert libraries
  - HiSeq (250pb); MiSeq (450bp)
  - Multiplexing; inexpensive, low error rates

Goodwin et al. 2006 *Nature Reviews*, 17:333-351.

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
  - Moleculo

  Illumina synthetic long-reads - 100kb?

https://dovetailgenomics.com/support/faqs/

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: **doi**: https://doi.org/10.1101/149997

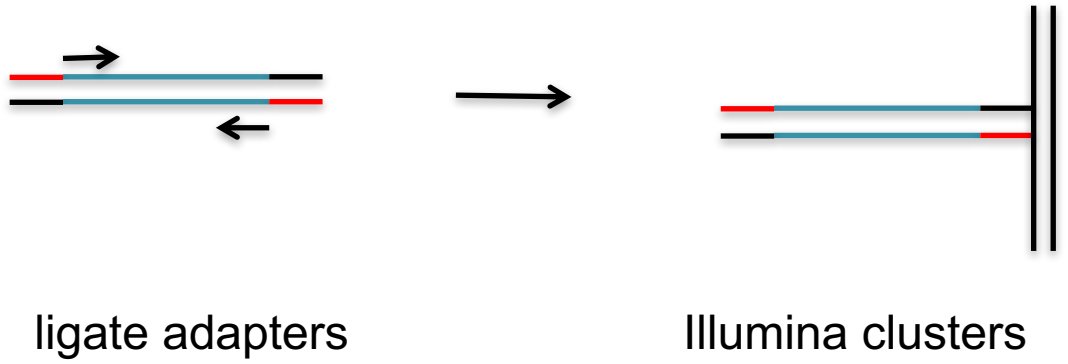
- Linear Assembly of a Human Y Centromere using Nanopore Long Reads. bioRxiv: https://doi.org/10.1101/170373

https://dovetailgenomics.com/support/faqs/

Short read sequencing - Illumina

- Three steps:
  1. Library Construction
  2. Cluster generation – Bridge PCR
  3. Sequencing

Library Construction



Paired-end, small-insert
~500 bp insert max
100 bp fragments

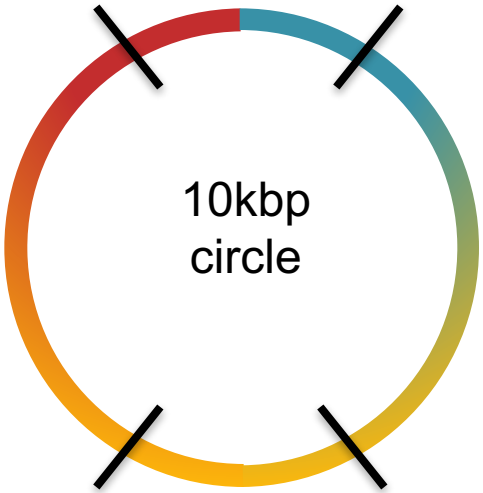
Paired-end and Mate-pairs

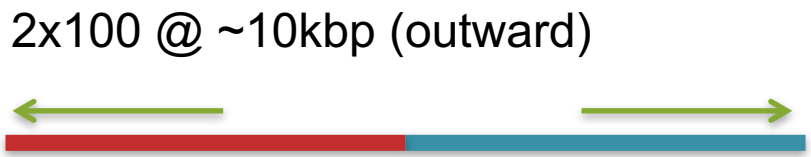
Paired-end sequencing (single-end, just one)

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

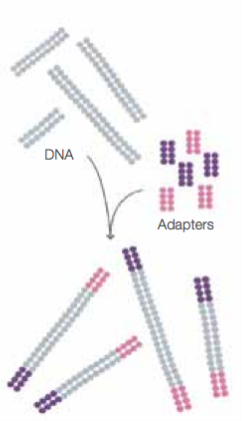
Mate-pair sequencing

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

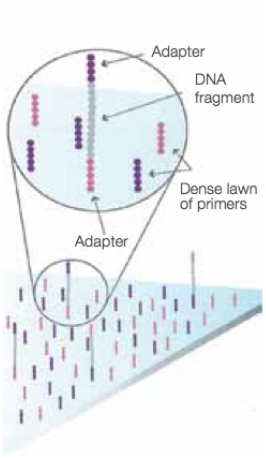




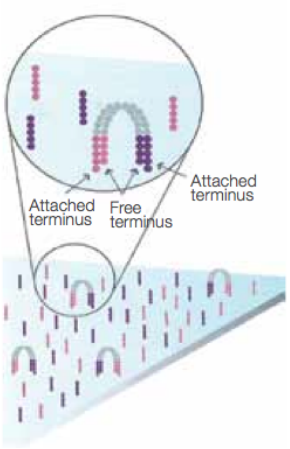
Illumina Sequencing by Synthesis



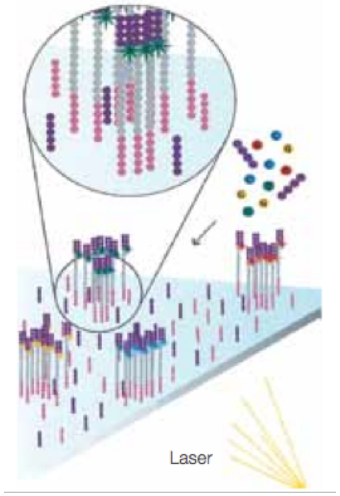
1. Prepare



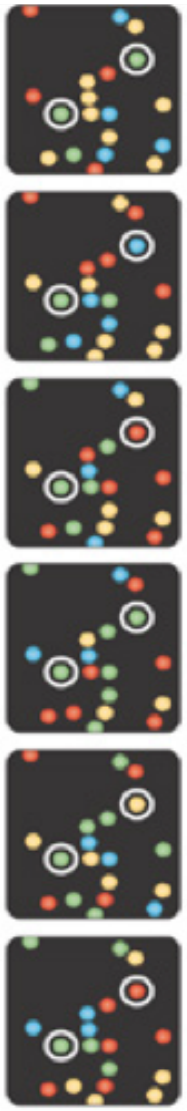
2. Attach



3. Amplify



4. Image



5. Basecall

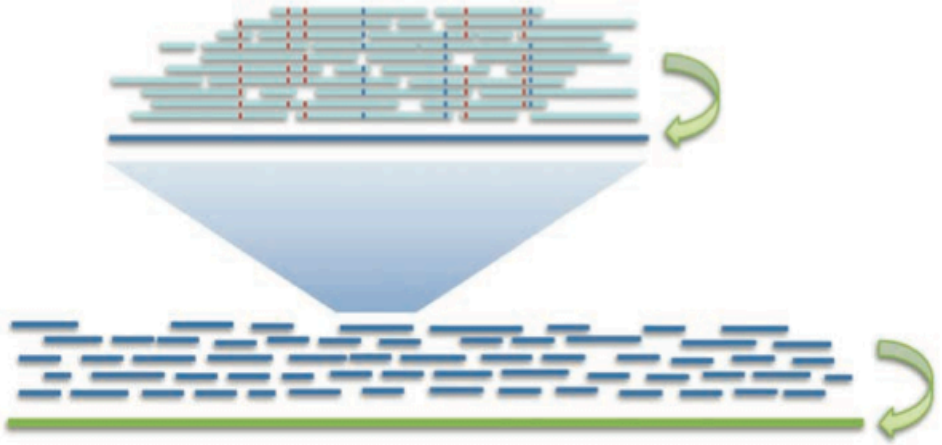
Metzker (2010) Nature Reviews Genetics 11:31-46

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



Preassembled reads used for genome assembly

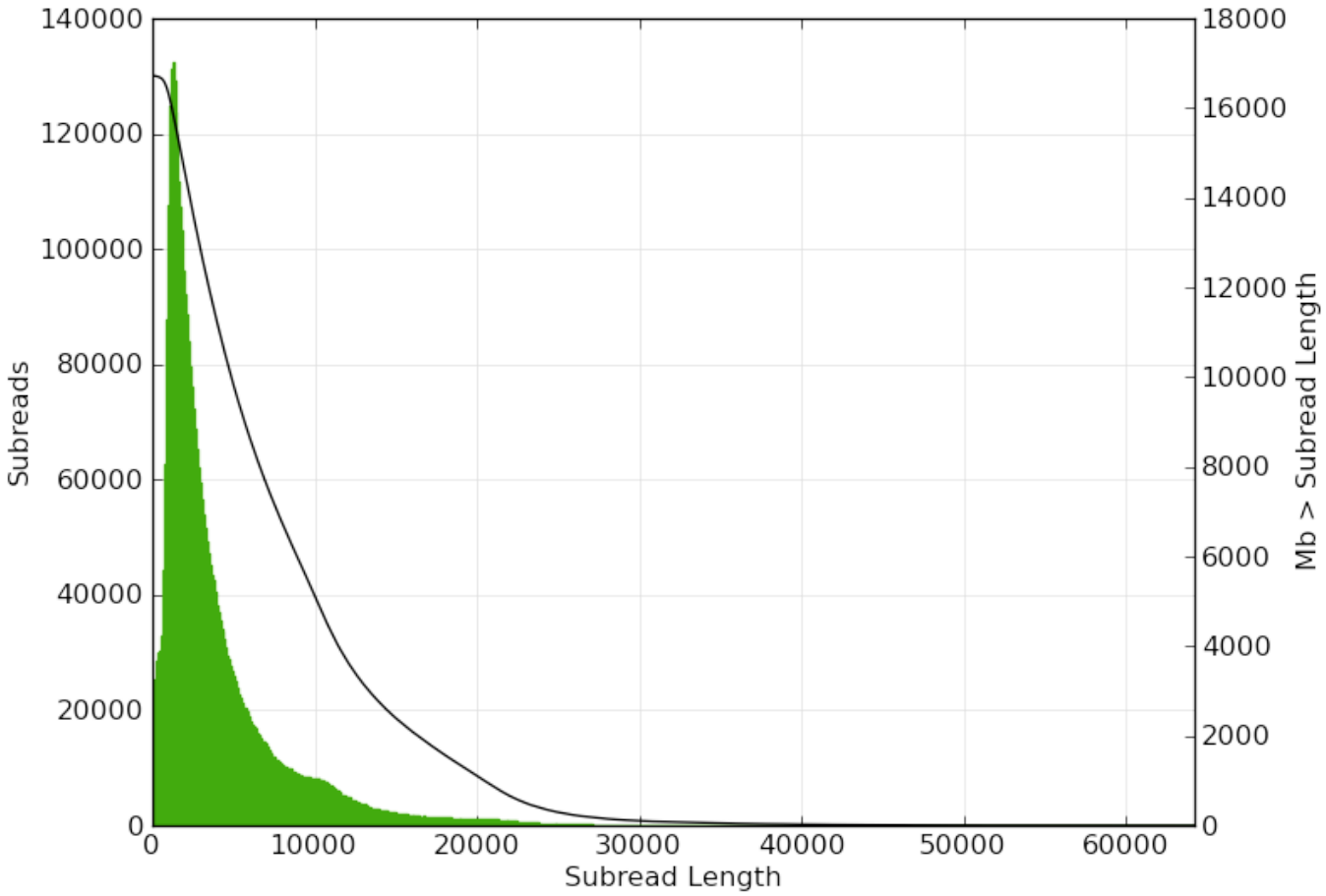
Error correction: mapping high-quality short PacBio reads to long reads

Assembly polishing with raw reads

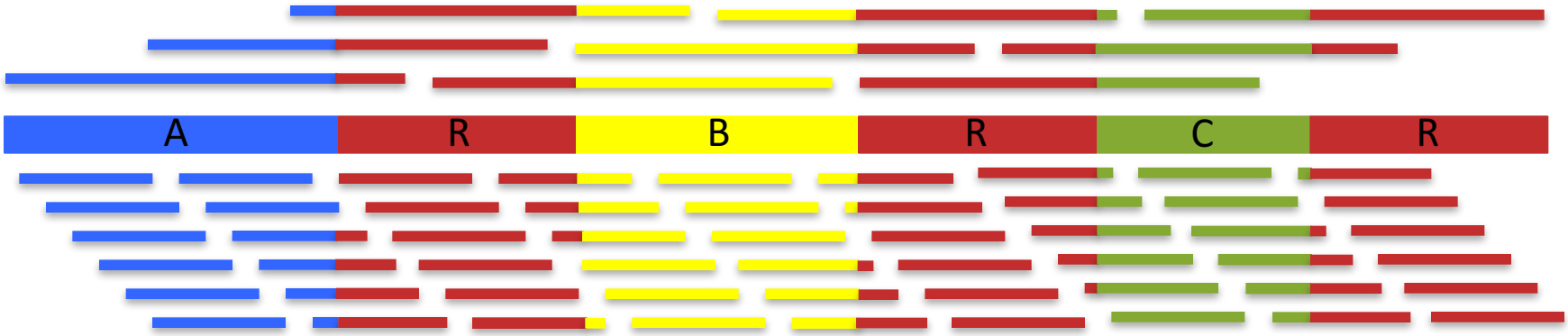
PB-only Correction & Polishing
Chin *et al* (2013) *Nature Methods.* 10:563–569

Long read sequencing

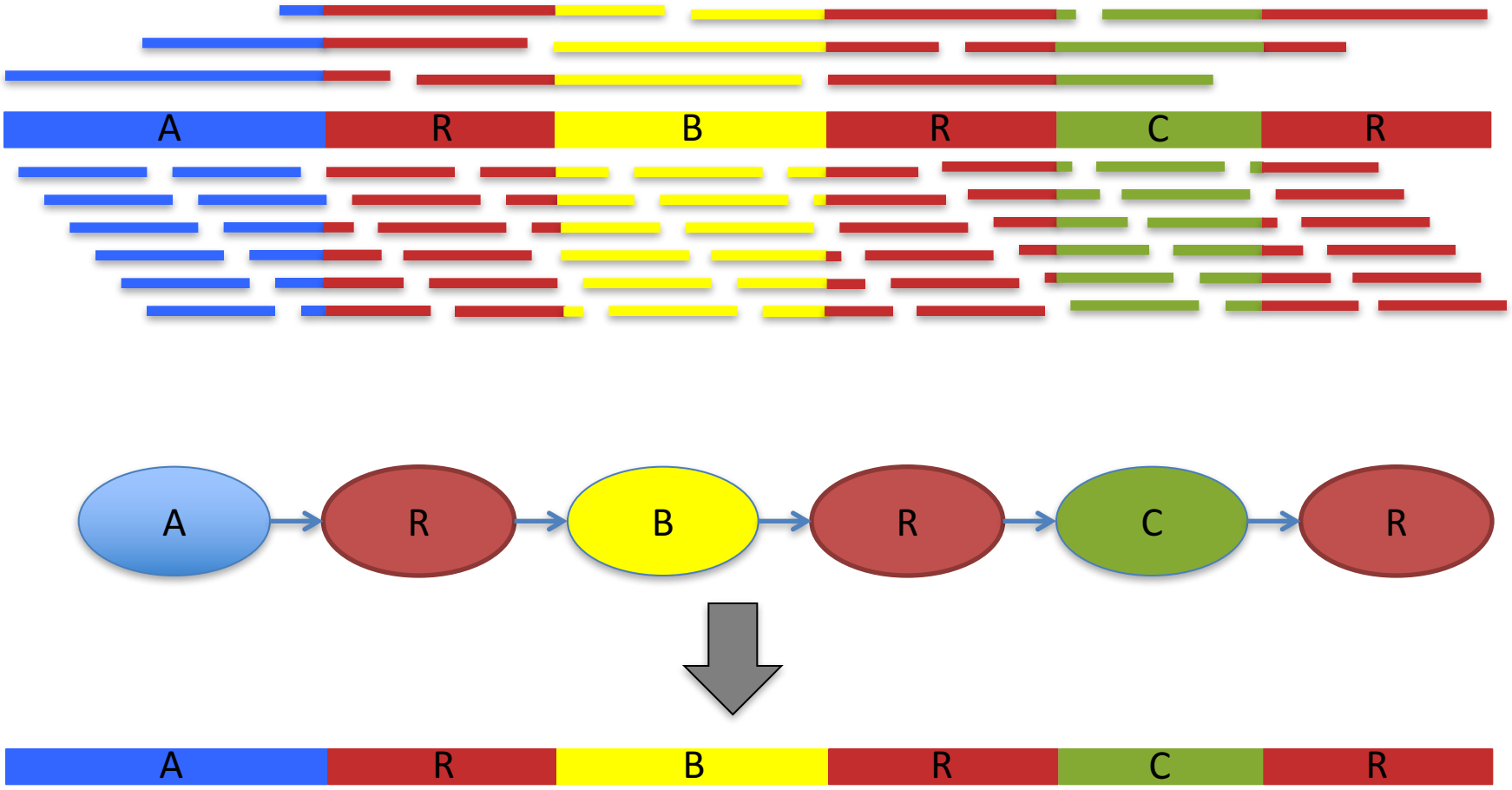
PacBio filtered subreads



Long Read Assembly - Complexity



Assembly Complexity



The advantages of SMRT (Single Molecule Real Time) sequencing
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology*. 14:405

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?
- How easy is it.....really?

File formatting

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

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

FASTA

```
>M63509.1 Human glutathione transferase M2 (GSTM2)
CGCAGCAACCAGCACCATGCCCATGACACTGGGGTACTGGAACATCCGCGGGCTGGCCCATTCCA
TCCGCCTGCTCCTGGAATACACAGACTCAAGCTACGAGGAAAAGAAGTACACGATGGGGGACGCT
CCTGATTATGACAGAAGCCAGTGGCTGAATGAAAAATTCAAGCTGGGCCTGGACTTTCCCAATCT
GCCCTACTTGATTGATGGGACTCACAAGATCACCCAGAGCAATGCCATCCTGCGGTACATTGCCC
GCAAGCACAACCTGTGCGGGGAATCAGAAAAGGAGCAGATTCGCGAAGACATTTTGGAGAACCAG
TTTATGGACAGCCGTATGCAGCTGGCCAAACTCTGCTATGACCCAGATTTTGAGAAACTGAAACC
AGAATACCTGCAGGCACTCCCTGAAATGCTGAAGCTCTACTCACAGTTTCTGGGGAAGCAGCCAT
GGTTTCTTGGGGACAAGATCACCTTTGTGGATTTCATCGCTTATGATGTCCTTGAGAGAAACCAA
GTATTTGAGCCCAGCTGCCTGGATGCCTTCCCAAACCTGAAGGACTTCATCTCCCGATTTGAGGG
CTTGGAGAAGATCTCTGCCTACATGAAGTCCAGCCGCTTCCTCCCAAGACCTGTGTTCACAAAGA
TGGCTGTCTGGGGCAACAAGTAGGGCCTTGAAGGCAGGAGGTGGGAGTGAGGAGCCCATACTCAG
CCTGCTGCCCAGGCTGTGCAGCGCAGCTGGACTCTGCATCCCAGCACCTGCCTCCTCGTTCCTTT
CTCCTGTTTATTCCCATCTTTACTCCCAAGACTTCATTGTCCCTCTTCACTCCCCCTAAACCCCT
GTCCCATGCAGGCCCTTTGAAGCCTCAGCTACCCACTATCCTTCGTGAACATCCCCTCCCATCAT
TACCCTTCCCTGCACTAAAGCCAGCCTGACCTTCCTTCCTGTTAGTGGTTGTGTCTGCTTTAAAG
GGCCTGCCTGGCCCTCGCCTGTGGAGCTCAGCCCCGAGCTGTCCCCGTGTTGCATGAAGGAGCAG
CATTGACTGGTTTACAGGCCCTGCTCCTGCAGCATGGTCCCTGCCTTAGGCCTACCTGATGGAAG
TAAAGCCTCAACCACAAAAAAAAAAA
```

FASTQ

```
@M00747:32:000000000-A16RG:1:1112:15153:29246 1:N:0:1
TCGATCGAGTAACTCGCTGCTGTCAGACTGGTTTTTGGTCGATCGACTATTGTTTCAGTCGCAAGAAT
ATTGTGTCCAGTCGATCGACTGAATTCTGCTGTACGGCCACGGCGGATGCACGGTACAGCAGGCTCAG
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)OO=9)))2@624(/(/2/-
(.(6;9(((((.(.'((6-66<6(///
@M00747:32:000000000-A16RG:1:1112:15536:29246 1:N:0:1
GTAAAATTGAGGTAAATTGTGCGGAATTTAGCAATACCGTTTTTTTTATTATCACCGGATATCTATTC
TGCTGTACGGCCAAGGAGGATGTACGGTACAGCAGGTGCGAACTCACTCCGACGCTCAAGTCAGTGAC
TTAATGATAAGCGTG
+
?????<BBBBBB5<?BFFFFFFECHEFFECCFF?9AAC>7@FHHHHHHFG?EAFGF@EEDEHHDGHHC
BDFFGDFHF)<CCD@F,+3=CFBDFHBD++??DBDEEEDE:):CBEEEBCE68>?))5?**0?:AE*A
*0//:/*:*:**.0)
@M00747:32:000000000-A16RG:1:1112:15513:29246 1:N:0:1
GCTAGTCTTGTGTTTAGTTTTATGTTTTGCATGTTGTAACGGATTCATAAACATAGGTGTTTGTTTCT
TTTTATGGTTGTACAATTTGGCCCTAAGGCCCTACACTTACTTGTTTGTTTCTTTTATGGTACGACAT
TTGAGTGGTGGTTGA
+
```

SAM/BAM Sequence Alignment Map

```
D4ZHLFP1:53:D2386ACXX:6:2115:17945:68812     0    Mle_000001     18    42    108M *    0    0
     TCCCCCTGCATGTGTCCGTCTGGCTGGATGCCATGCTCCATGCAGTATAGCTCCCAGCATGAGTTACCGATCTGGACACCTGCTTG
GCCAAGATGTACTGAGATGCAT
C@CFDFFFHHGHHHFGGBFEGGDGGGGEHGIGGGJJJJIIIGIIB9BFBBFHGGHICEAGHGEGEDHIGEEDBECCACBDDC@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
TCCCCCTGCATGTGTCCGTCTGGCTGGATGCCATGCTCCATGCAGTATAGCTCCCAGCATGAGTTACCGATCTGGACACCTGCTTGGCCAA
GATGTACTGAGATGCAT
CCCFFFFFHHHHHHGHGGEGIJIIIGJFHJJJJJIIJJIIGIJIJFHJJIIIJJIFIIIIIIJIIJJJHHFFFCEEEEDDDDDDDDDDDDD
BDCDDEEEDDDDDDDDD    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
TCCCCCTGCATGTGTCCGTCTGGCTGGATGCCATGCTCCATGCAGTATAGCTCCCAGCATGAGTTACCGATCTGGACACCTGCTTGGCCAA
GATGTACTGAGATGCAT
CCCFFFFFHHHHHJJJJIHJJIIIIJJIJJJJJJJJJIJJJJJJJIIJJJJJJJJJJJJJJJJJIJJJGEFFFEEEEEDDDDDDDDDDDDD
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

Illumina data:

DISCOVAR denovo

https://software.broadinstitute.org/software/discovar/blog/?page_id=98

large genome assembler

ALLPATHS-lg

http://software.broadinstitute.org/allpaths-lg/blog/?page_id=12

- large genome assembler
- needs at least one mate-pair library and requires specific insert size for paired-end library

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

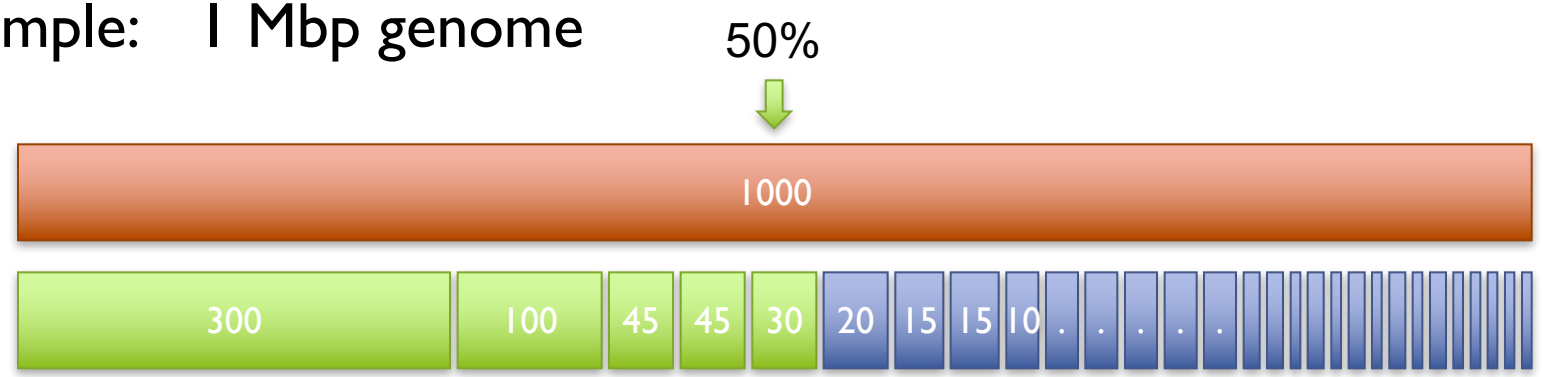
- Recommended to use multiple assemblers with different parameters to assess results
- How to assess our results?
  - Number of contigs/scaffolds
  - Longest contig/scaffold
  - L50 - 50% of the genome is contained in contigs longer than value
  - Busco - Simao et al. 2015. Bioinformatics

Annotation, alignments, assessing repetitive regions

L50 size

Def: 50% of the genome is in contigs as long as the L50 value

Example: 1 Mbp genome



L50 size = 30 kbp

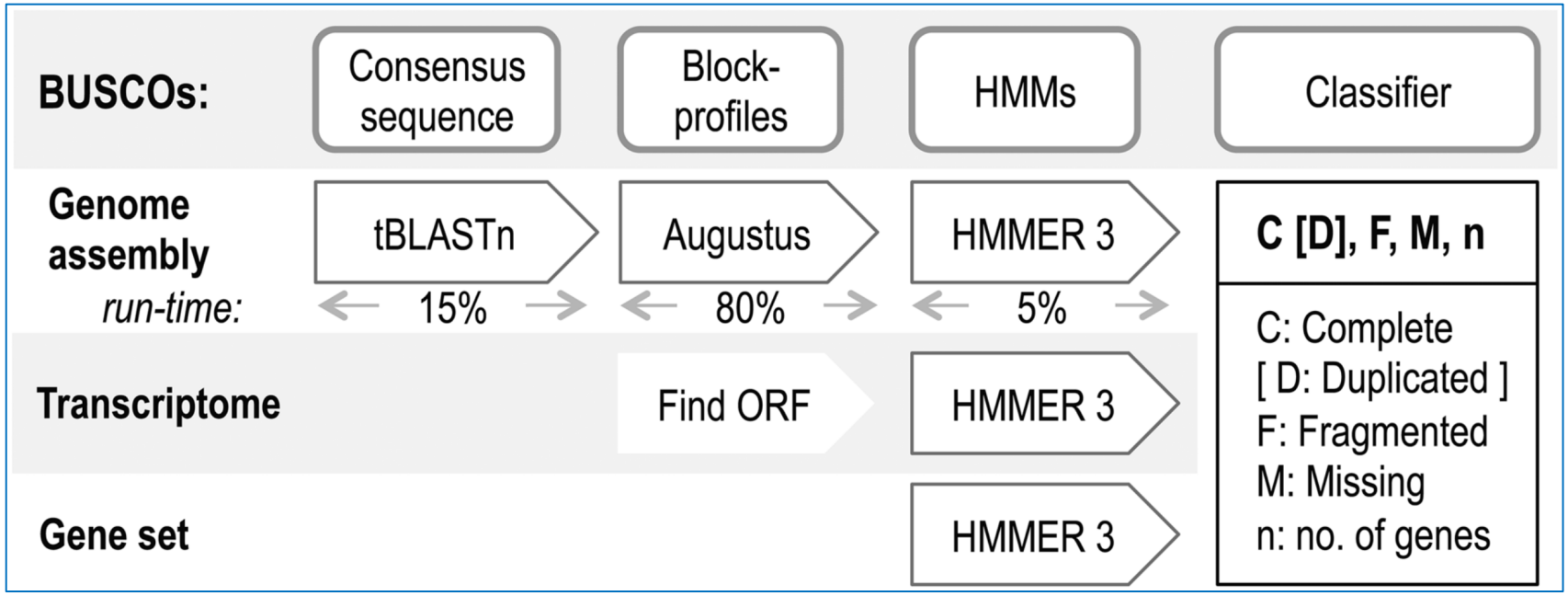
(300k+100k+45k+45k+30k = 520k >= 500kbp)

A greater L50 is indicative of improvement in every dimension:

- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization
- Better sequence for all downstream analysis

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

Lecture outline

1. 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 workshop

Assembly Summary

Assembly quality depends on

1. **Experimental design**: clear and organized with high-quality DNA
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 build scaffolds
  - 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