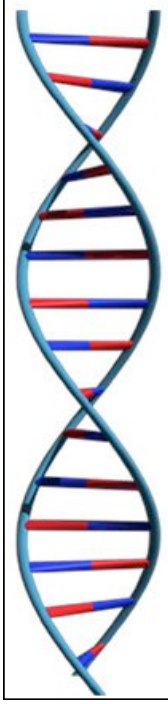
Genome Sequencing & Assembly

Deb Triant
Based on material by Michael Schatz

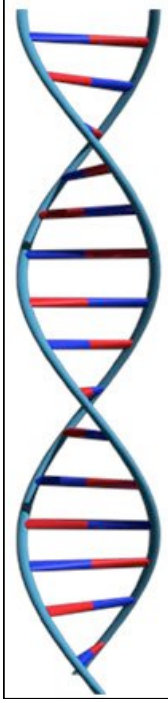
18 October 2015
Programming for Biology



Outline

1. Assembly theory
   1. Assembly by analogy
   2. Overlap graph
   3. Coverage, read length, errors, and repeats
2. Whole Genome Alignment
   1. Aligning & visualizing with MUMmer
3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Long-read assemblies
4. Summary & Recommendations

2



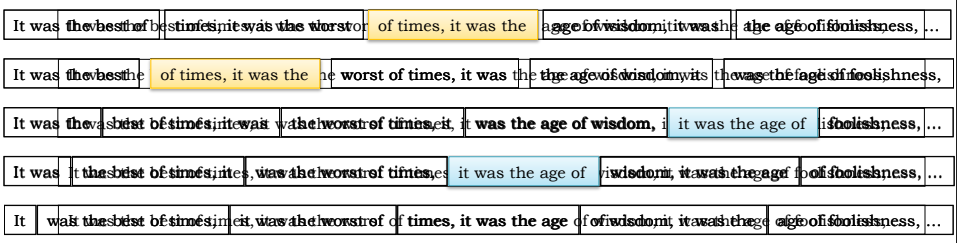
Outline

1. Assembly theory
   1. Assembly by analogy
   2. Overlap graph
   3. Coverage, read length, errors, and repeats
2. Whole Genome Alignment
   1. Aligning & visualizing with MUMmer
3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Long-read assemblies
4. Summary & Recommendations

3

Shredded Book Reconstruction

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

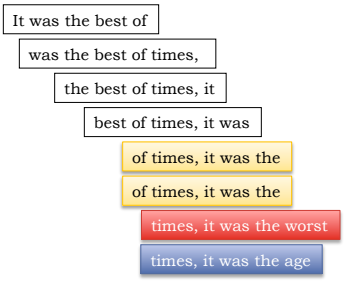


- How can he reconstruct the text?
  - 5 copies x 138, 656 words / 5 words per fragment = 138k fragments
  - The short fragments from every copy are mixed together
  - Some fragments are identical

4

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

Greedy Reconstruction



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

5

de Bruijn 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

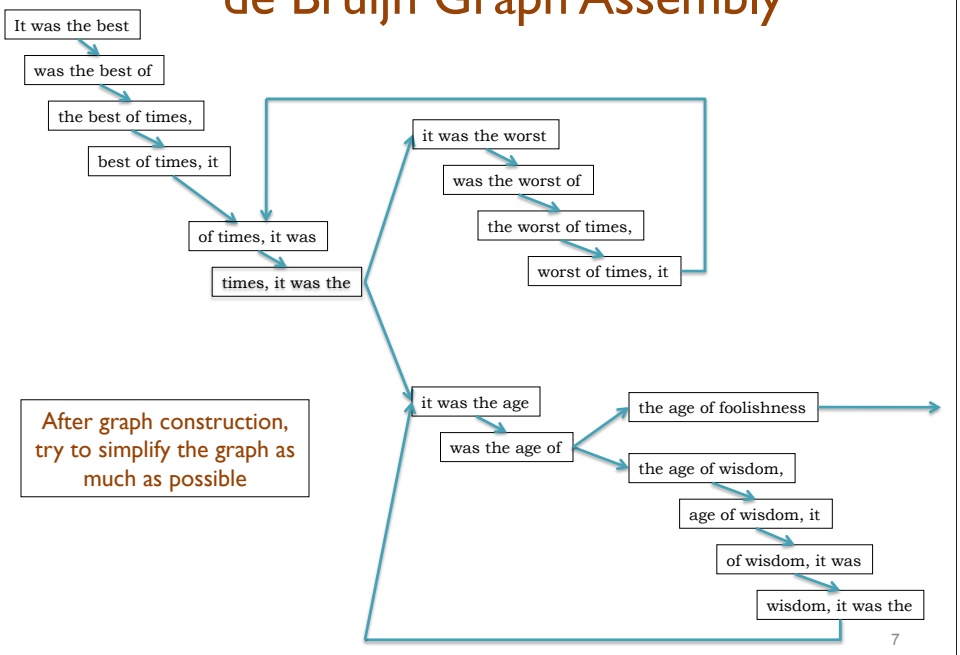
It **was the best** → **was the best** of

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

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

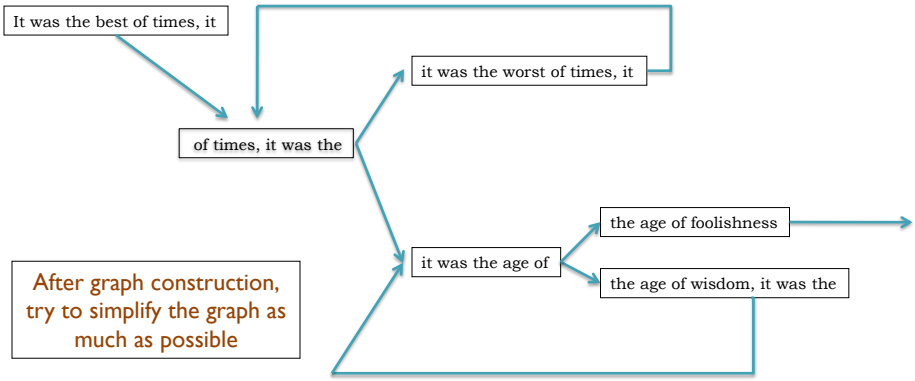
6

de Bruijn Graph Assembly



7

de Bruijn Graph Assembly



8

de Bruijn Genome Assembly

- Shredded words → k-mer

GACTGG
↓
ACTGGG
↓
CTGGGA
↓
TGGGAC
↓
GGGACT
↓
GGACTC
↓
GACTCC

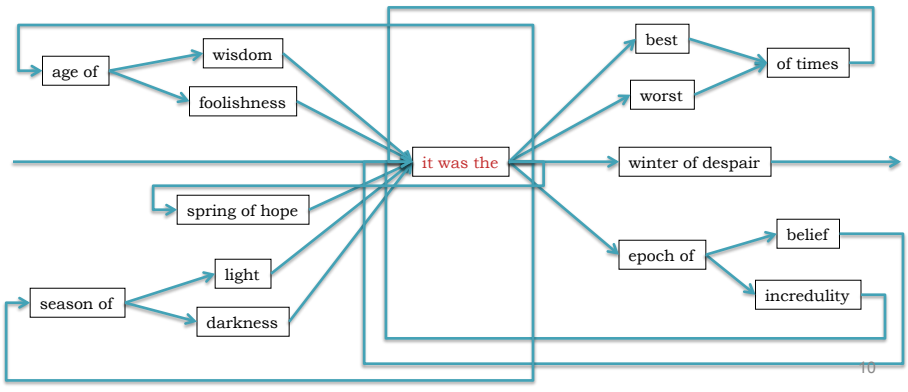
GACTGGGACTCC

kmer n = 6

9

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 …



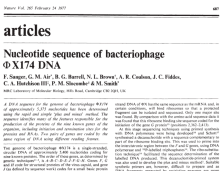
10

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

11

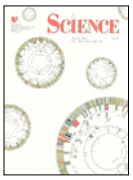
Milestones in Genome Assembly


articles

Nucleotide sequence of bacteriophage Φ X174 DNA

F. Sanger, G. M. Air*, B. G. Barrell, N. L. Brown*, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III*, P. M. Slocombe* & M. Smith*

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



1995. Fleischmann *et al.*
1st Free Living Organism
TIGR Assembler. 1.8Mbp

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

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

2001. Venter *et al.*, IHGSC
Human Genome
Celera Assembler/GigaAssembler. 2.9 Gbp

2010. Li *et al.*
1st Large SGS Assembly.
SOAPdenovo 2.2 Gbp

Like Dickens, we must computationally reconstruct a genome from short fragments

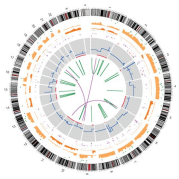
18

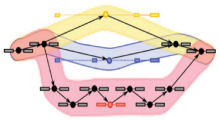
Assembly Applications

- Novel genomes

- Metagenomes

- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …

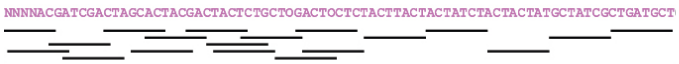




13

Recipe for a good assembly

- Coverage
  - How many times has genome been sequenced?
  - Too much? Too little?

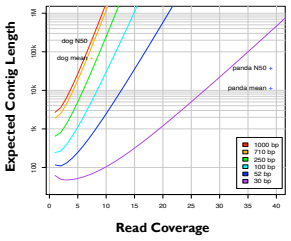


- Read Length
  - Read lengths must be longer than repetitive regions
- Quality
  - reads assembled by shared regions

14

Ingredients for a good assembly

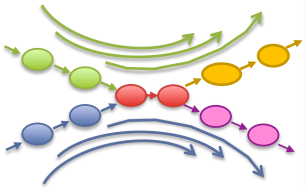
Coverage



High coverage is required

- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

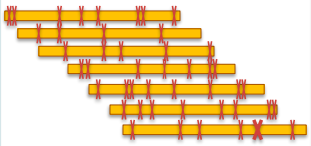
Read Length



Reads & mates must be longer than the repeats

- Short reads will have ***false overlaps*** forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality



Errors obscure overlaps

- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds (unipaths), increasing complexity and forming assembly hairballs

Current challenges in *de novo* plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

15

Illumina sequencing

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

16

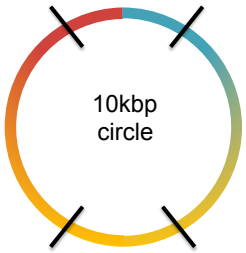
Paired-end and Mate-pairs

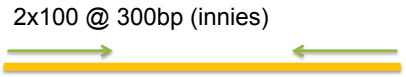
Paired-end sequencing

- 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

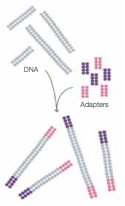
- Circularize long molecules (1-10kbp), shear into fragments, & sequence
- Mate failures create short paired-end reads



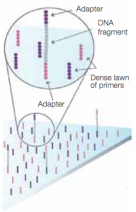


17

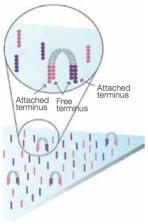
Illumina Sequencing by Synthesis



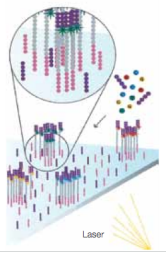
1. Prepare



2. Attach



3. Amplify



4. Image

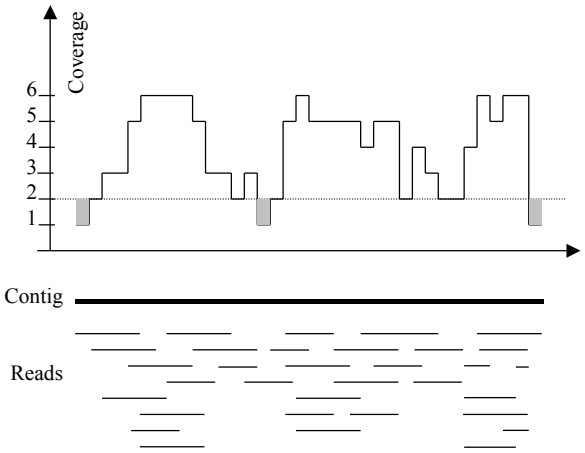
5. Basecall

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

http://www.youtube.com/watch?v=I99aKKHcxC4

18

Typical sequencing coverage



Imagine raindrops on a sidewalk

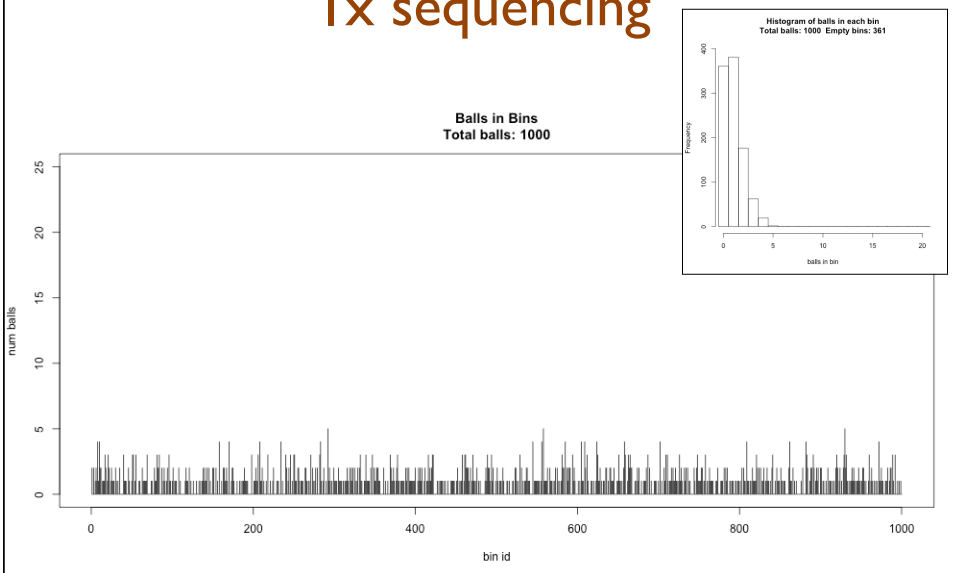
We want to cover the entire sidewalk but each drop costs $$$$!

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!!!
  - number and type of libraries required
  - potential projects resulting from assembly

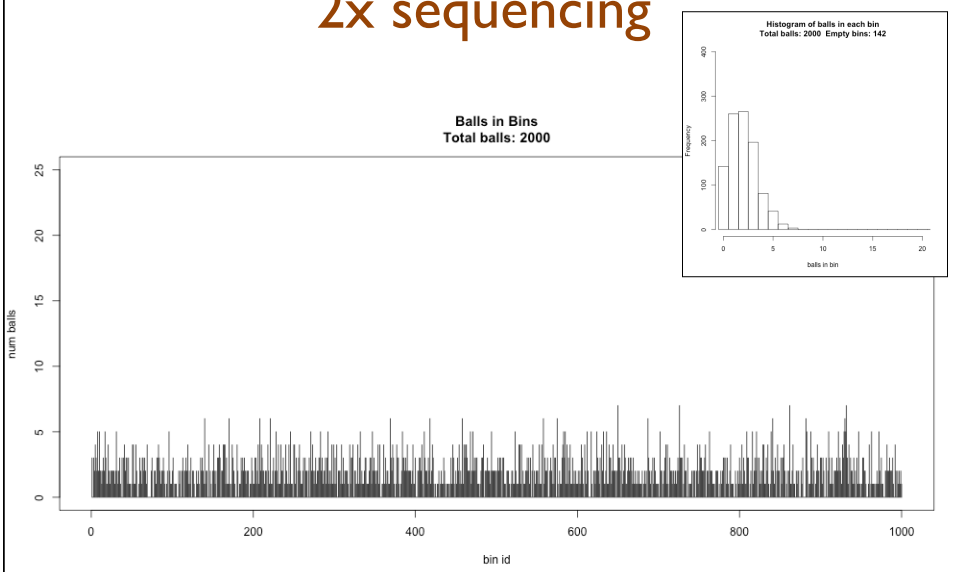
20

1x sequencing



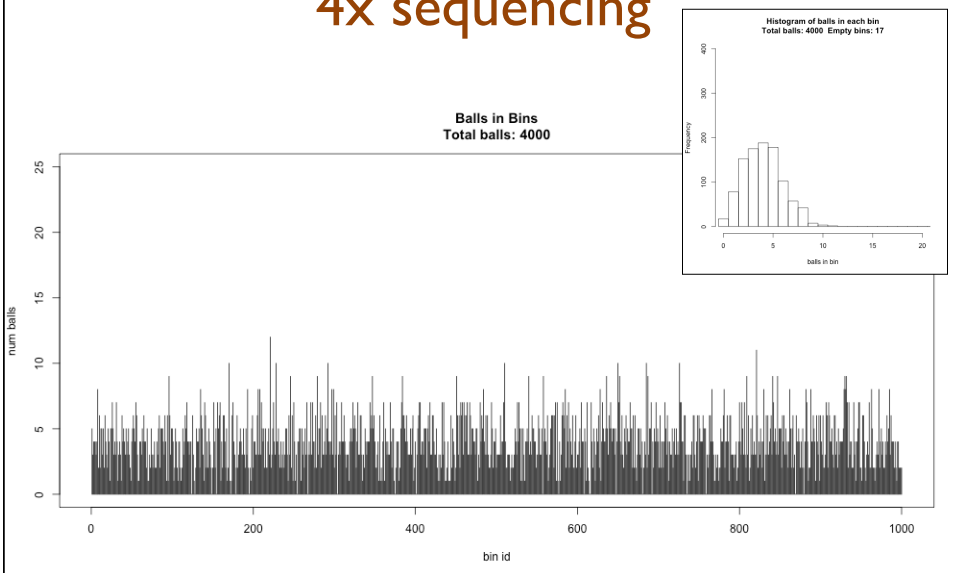
21

2x sequencing



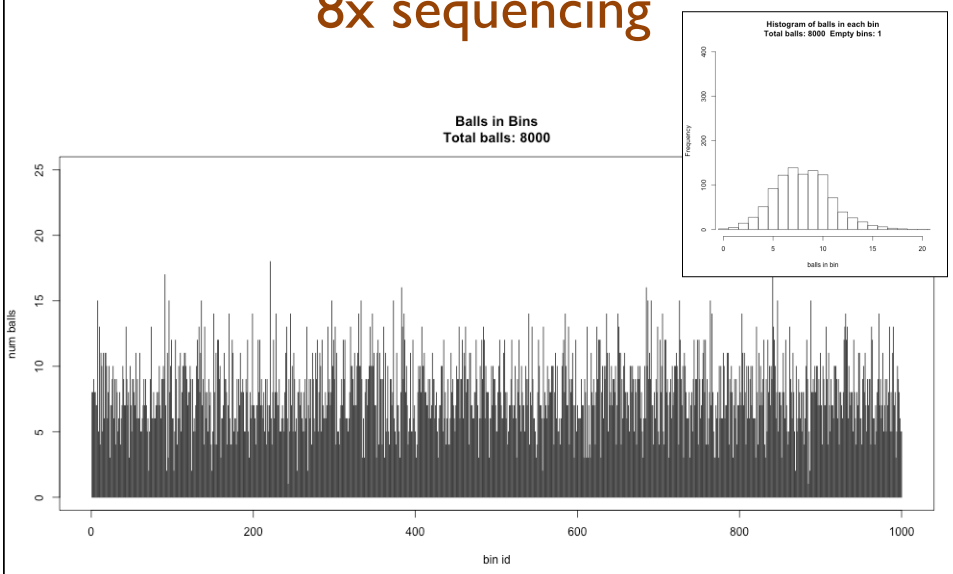
22

4x sequencing



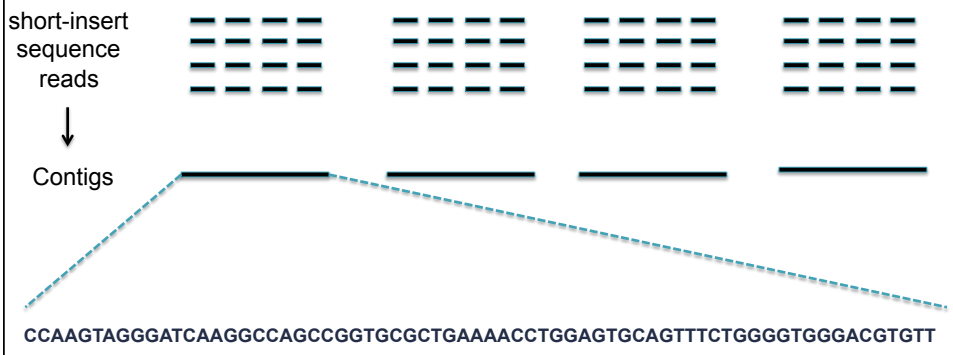
23

8x sequencing



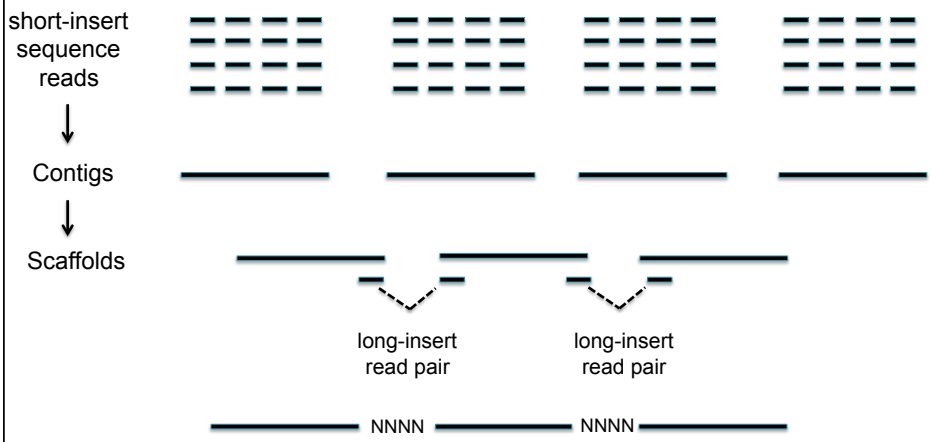
24

Assembly construction



Contig - contiguous sequence of nucleotides

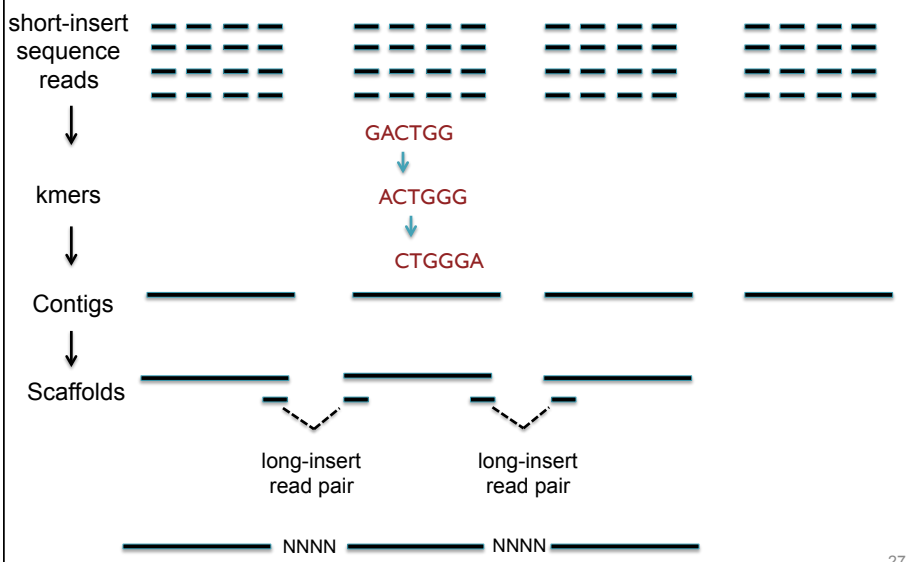
Assembly construction



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

>GTAGTATTCTAGAAAATGTTTTAAATATAGATAGTTGTTTTANATCTGTTAGTTGTCAGATGCTTACTGAATAGTTGGAANNNNNNNNNNNNNNNNNNNNNNNNTGTGAGGTTTTTAGCTCATGAAAGTTTATGATTATTGCACCCCTACTCACAACGAATCCCTATTCTTATCTTTTNNNNNNNNNNNNNNCATGTCACTGGTTTATTTATTTTTTGTGGCTGCAGAAGTCCTTTGTCGCTGTTAATTTTTTGGGAGTTCTCCTGTCGTATATTAAAGCTTTCTTTCTTTTCAGTTTTAAATTTATTTGAACCTTACTATCTTTCTAACAATAAATTGTGGAATTATCAACGAAAACATAGGNNNNNNNNNNGTCCTTTATACGAAAGCTATATAGTGTTAGGCTTTTCTTTTTTTTTNNNNNNNNGGTGATGTTGTTAATGGTGCCCTTTTCTGGTAATCTTACTAAATCAGTTTGCTTGTTACTTGTATAGTTGTA

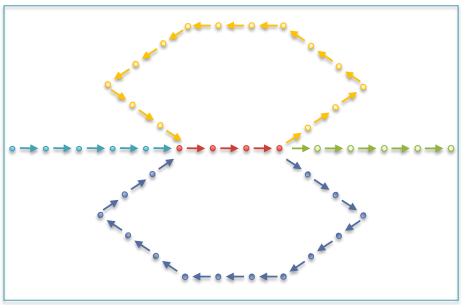
Assembly construction

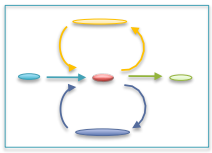


27

Unitigging / Unipathing

- After simplification and correction, compress graph down to its non-branching initial contigs
  - Aka "unitigs", "unipaths"
  - Unitigs end because of (1) lack of coverage, (2) errors, (3) heterozygosity, and (4) repeats
  - Reads offer two possible alternatives for a base.

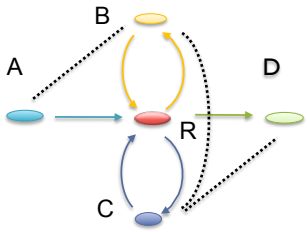




28

Scaffolding

- Initial contigs (*aka* unipaths, unitigs) terminate at
  - *Coverage gaps*: especially extreme GC
  - *Conflicts*: errors, repeat boundaries
- Use mate-pairs to resolve correct order through assembly graph
  - Place sequence to satisfy the mate constraints
  - Mates through repeat nodes are tangled
- Final scaffold may have internal gaps called sequencing gaps
  - We know the order, orientation, and spacing, but just not the bases. Fill with Ns instead



29

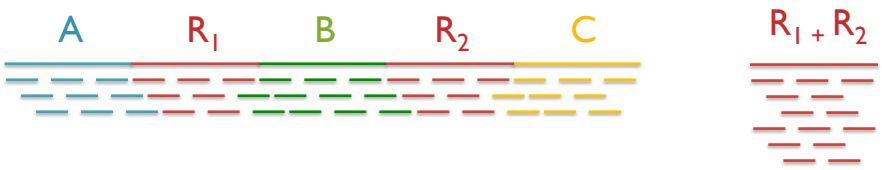
Repetitive regions

Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1b_2\ldots b_k)^N$ where $1 \leq k \leq 6$ CACACACACACACACACACA	2%
SINEs (Short Interspersed Nuclear Elements)	*Alu* sequence (~280 bp) Mariner elements (~80 bp)	13%
LINEs (Long Interspersed Nuclear Elements)	~500 – 5,000 bp	21%
LTR (long terminal repeat) retrotransposons	Ty1-copia, Ty3-gypsy, Pao-BEL (~100 – 5,000 bp)	8%
Other DNA transposons		3%
Gene families & segmental duplications		4%

- Over 50% of mammalian genomes are repetitive
  - Large plant genomes tend to be even worse

30

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.

31

Genome Assembly

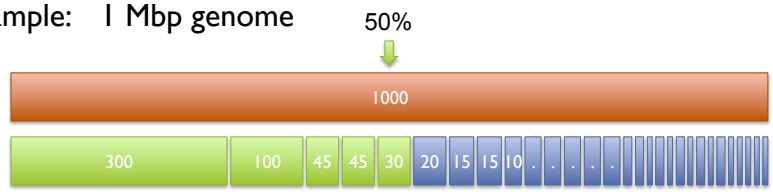
- Recommended to use multiple assemblers with different parameters to assess results
- How to assess our results?
  - Number of contigs/scaffolds
  - Longest contig/scaffold
  - L50 - 50% of the genome is contained in contigs longer than value

32

L50 size

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

Example: 1 Mbp genome



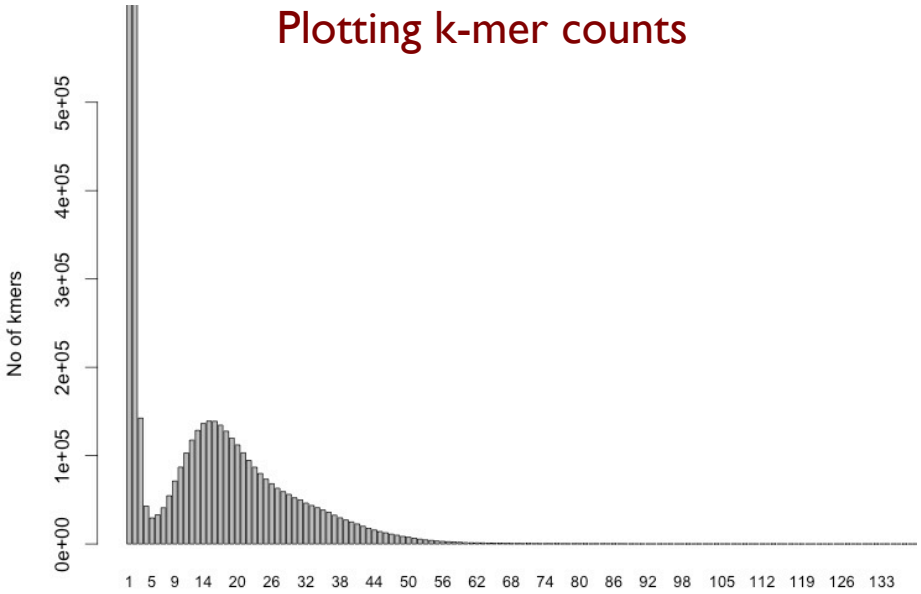
L50 size = 30 kbp
(300k+100k+45k+45k+30k = 520k >= 500kbp)

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

33

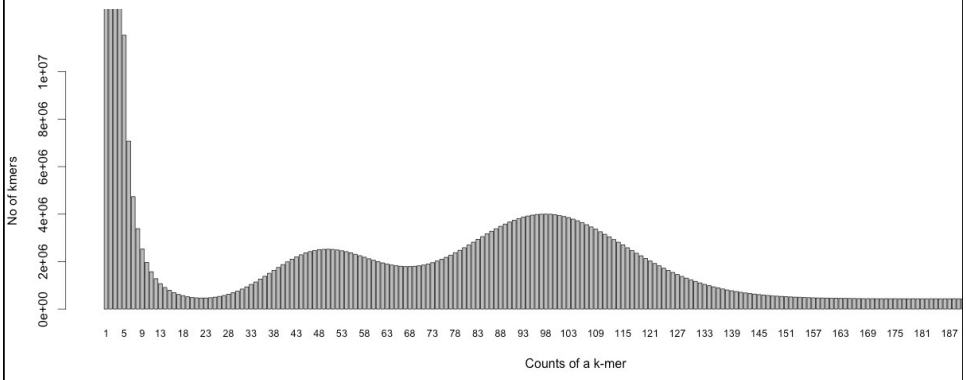
Plotting k-mer counts



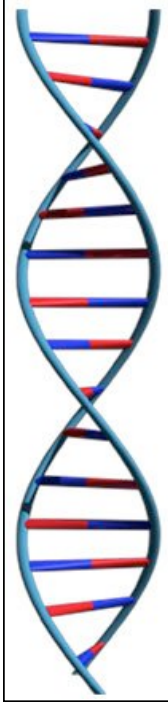
Quake. Kelley et al. Genome Biology 2010

34

Plotting k-mer counts



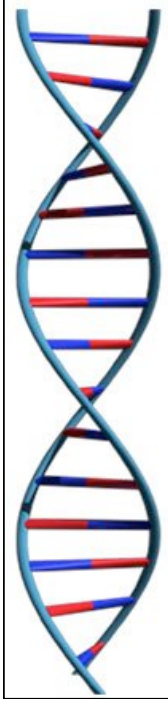
35



Outline

1. Assembly theory
   1. Assembly by analogy
   2. Overlap graph
   3. Coverage, read length, errors, and repeats
2. Whole Genome Alignment
   1. Aligning & visualizing with MUMmer
3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Long-read assemblies
4. Summary and Recommendations

36

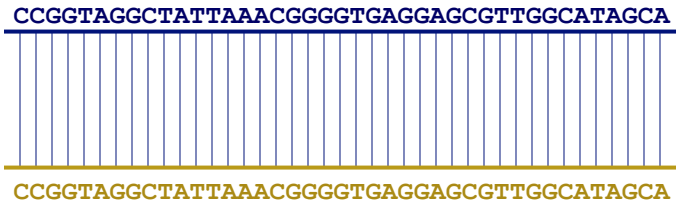


Whole Genome Alignment with MUMmer

Slides Courtesy of Adam M. Phillippy
University of Maryland

Goal of WGA

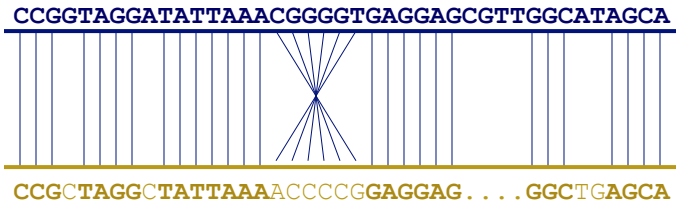
- For two genomes, *A* and *B*, find a mapping from each position in *A* to its corresponding position in *B*



38

Not so fast...

- Genome *A* may have insertions, deletions, translocations, inversions, duplications or SNPs with respect to *B* (sometimes all of the above)

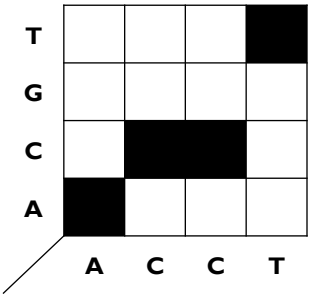


39

WGA visualization

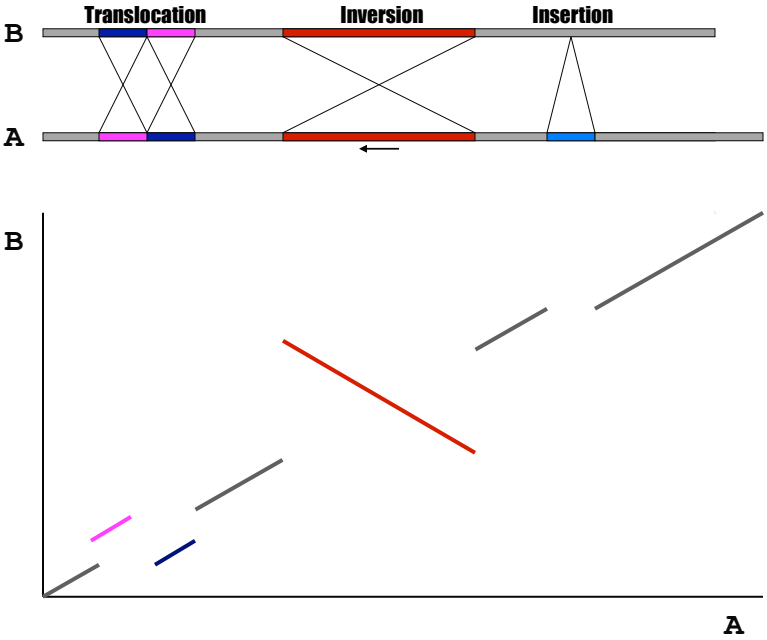
- How can we visualize *whole* genome alignments?

- With an alignment dot plot
  - *N* x *M* matrix
    - Let *i* = position in genome *A*
    - Let *j* = position in genome *B*
    - Fill cell *(i,j)* if A_i shows similarity to B_j



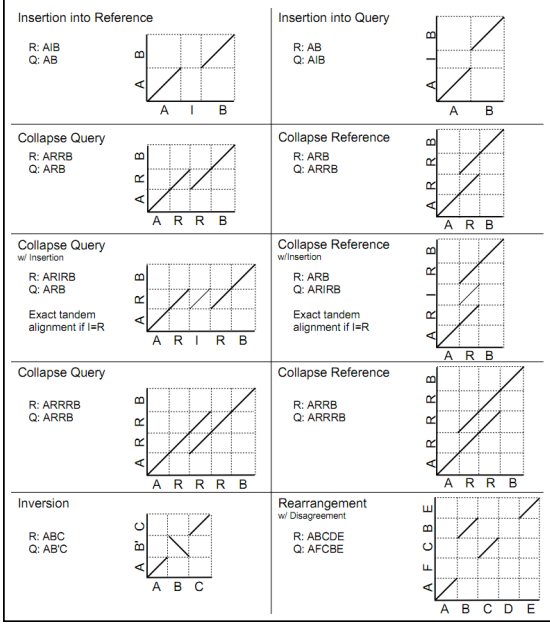
  - A perfect alignment between *A* and *B* would completely fill the positive diagonal

40



41

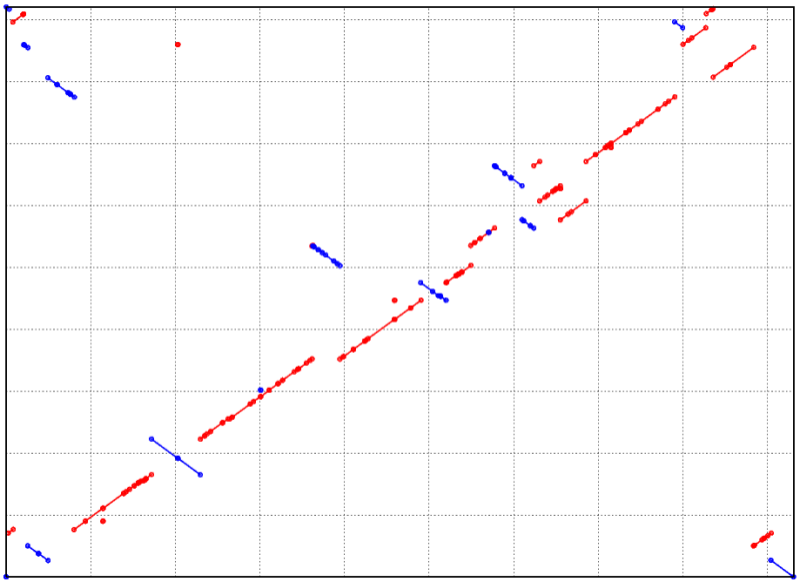
SV Types



- Different structural variation types / misassemblies will be apparent by their pattern of breakpoints
- Most breakpoints will be at or near repeats
- Things quickly get complicated in real genomes

http://mummer.sf.net/manual/AlignmentTypes.pdf

42



Alignment of 2 strains of *Y. pestis*
http://mummer.sourceforge.net/manual/

43

Outline

1. Assembly theory
   1. Assembly by analogy
   2. Overlap graph
   3. Coverage, read length, errors, and repeats
2. Whole Genome Alignment
   1. Aligning & visualizing with MUMmer
3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Long-read assemblies
4. Summary and Recommendations

44

Genome assembly with ALLPATHS-LG
Iain MacCallum

BROAD INSTITUTE

How ALLPATHS-LG works

reads

corrected reads

doubled reads

unipaths

localized data

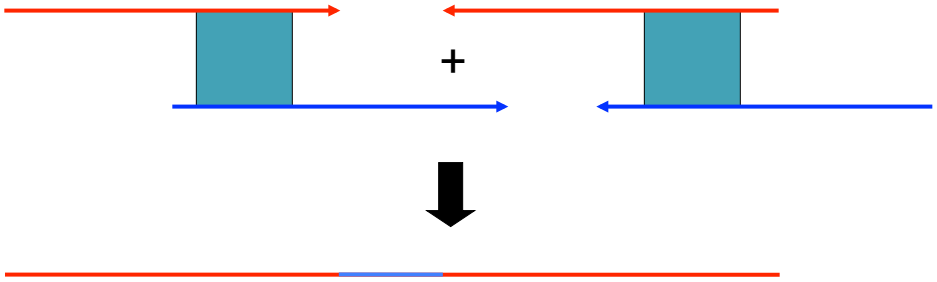
local graph assemblies

global graph assembly

assembly

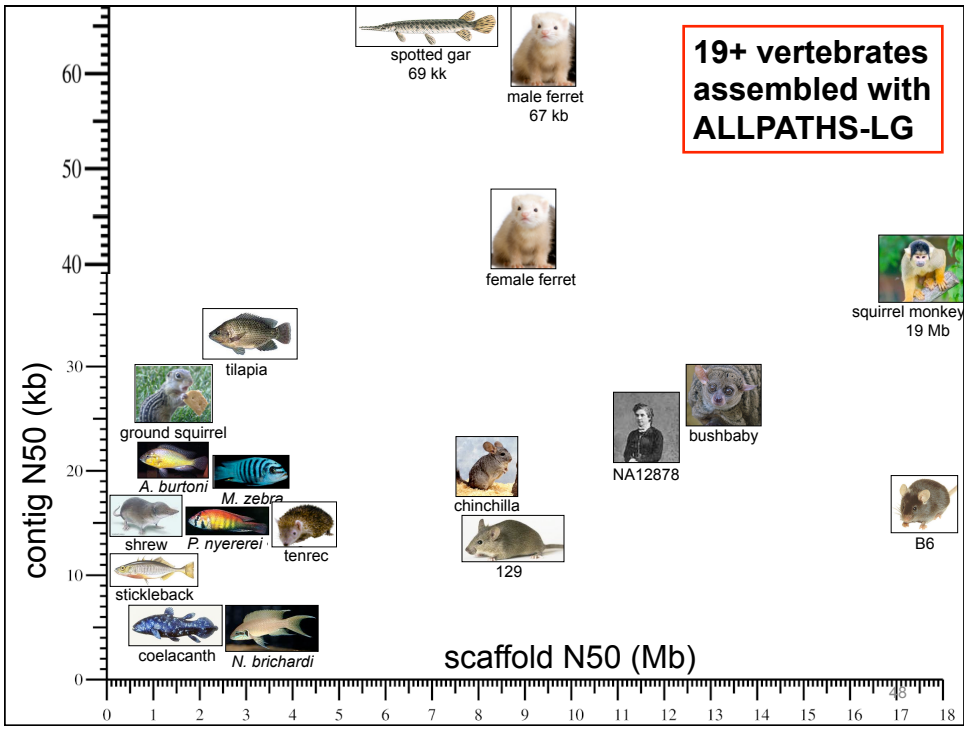
Read doubling

To close a read pair (red), we require the existence of another read pair (blue), overlapping perfectly like this:



More than one closure allowed (but rare).

47

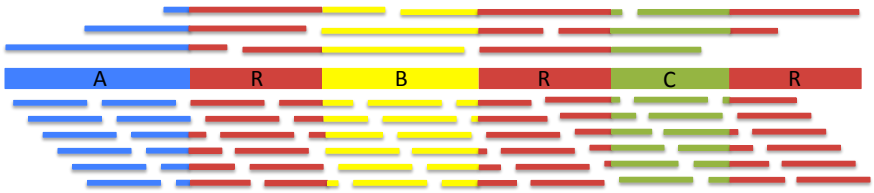


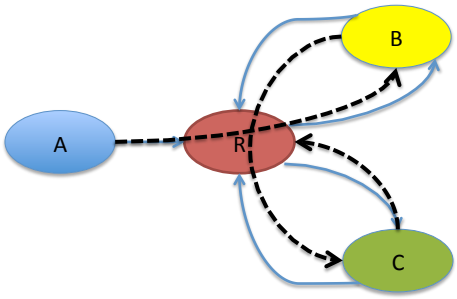
48

Genome assembly with long reads

49

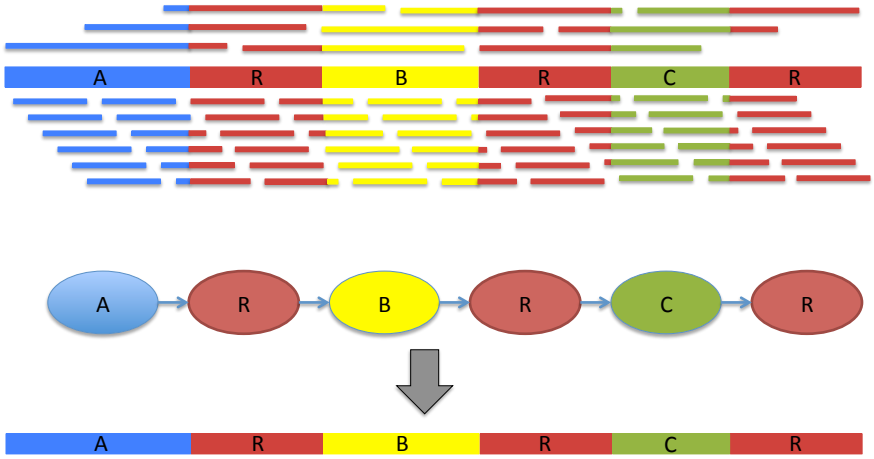
Assembly Complexity





50

Assembly Complexity

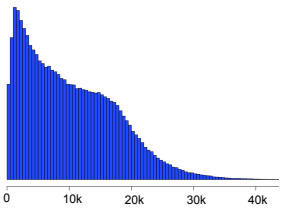


The advantages of SMRT (Single Molecule Real Time) sequencing
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology.* 14:405

51

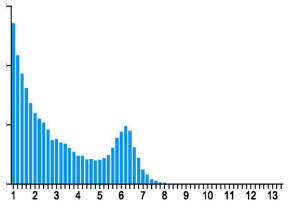
Long Read Sequencing Technology

PacBio
Pacific Biosciences



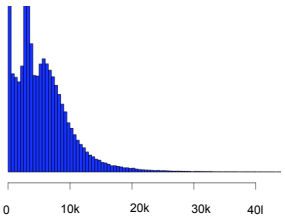
CSHL/PacBio

Moleculo



(Voskoboynik et al. 2013)

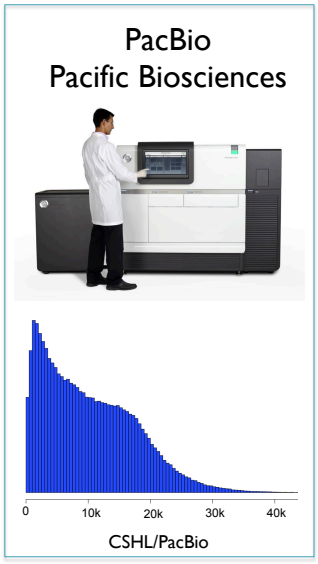
Oxford Nanopore



CSHL/ONT

52

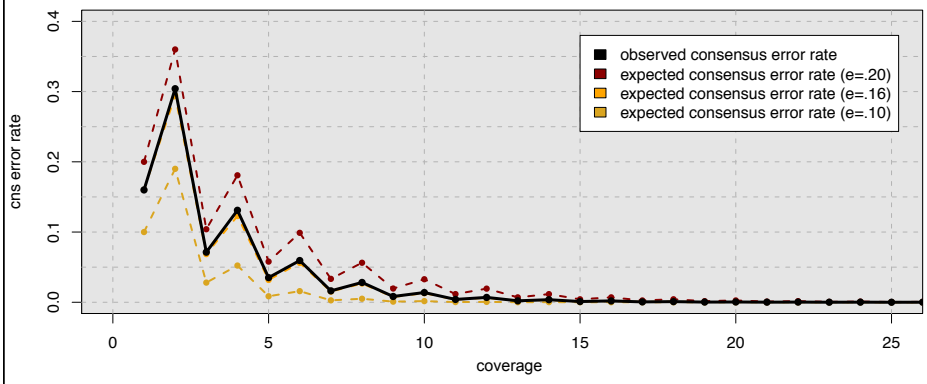
Long Read Sequencing Technology



- SMRT Sequencing - Single Molecule Real Time
- High error rate - 15-20% but error correction methods available
- Read lengths up to 20kb or higher
- Hybrid genomes: scaffolding with Illumina sequencing but now prices decreased and longer reads

53

Consensus Accuracy and Coverage



Coverage can overcome random errors

- Dashed: error model from binomial sampling
- Solid: observed accuracy

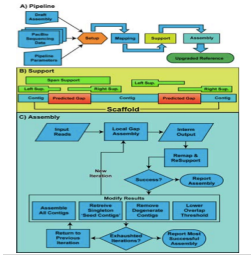
PacBio: "High error-rate"

Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

54

PacBio Assembly Algorithms

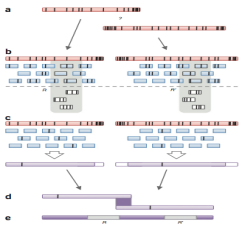
PBJelly



Gap Filling and Assembly Upgrade

English *et al* (2012)
PLOS One. 7(11): e47768

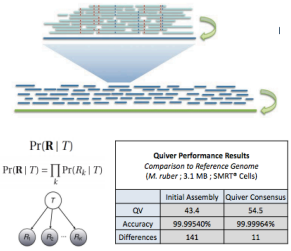
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

HGAP & Quiver Falcon; MHAP



$$\Pr(\mathbf{R} \mid T) = \prod_k \Pr(R_k \mid T)$$

Quiver Performance Results
Comparison to Reference Genome
(M. ruber ; 3.1 MB ; SMRT® Cells)

	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

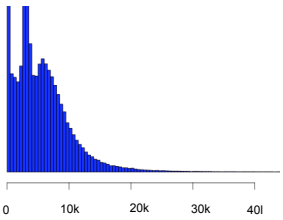
PB-only Correction & Polishing

Chin *et al* (2013)
Nature Methods. 10:563–569

< 5x PacBio Coverage > 50x

Long Read Sequencing Technology

Oxford Nanopore

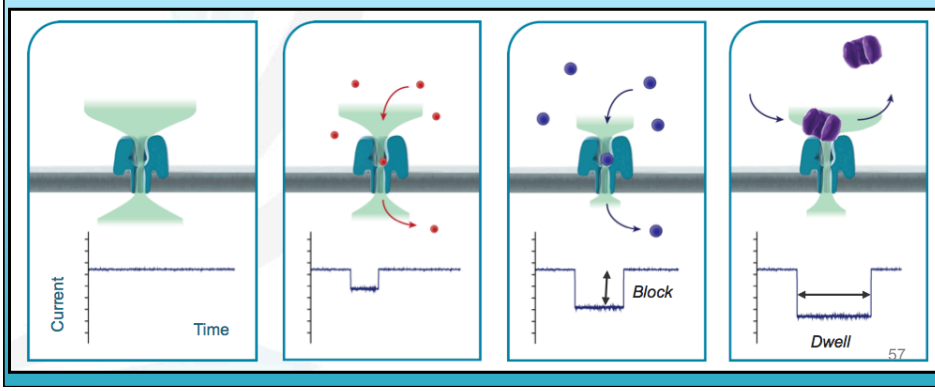


CSHL/ONT

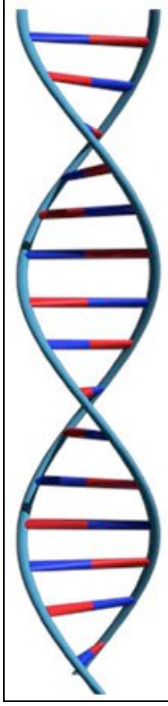
- Pocket sized for sequencing on laptop or in the field **isolation?
- DNA sequenced by threading through microscopic pores
- Read lengths 5 - 10kb
- No amplification or chemical labeling of samples

Oxford Nanopore MinION

- Thumb drive sized sequencer powered over USB
- Capacity for 512 reads at once
- Senses DNA by measuring changes to ion flow



57



Outline

1. Assembly theory
   1. Assembly by analogy
   2. Overlap graph
   3. Coverage, read length, errors, and repeats
2. Whole Genome Alignment
   1. Aligning & visualizing with MUMmer
3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Long-read assemblies
4. Summary and Recommendations

58

Assembly Summary

Assembly quality depends on

1. ***Experimental design***: clear and organized with high-quality DNA
2. ***Coverage***: low coverage is mathematically hopeless
3. ***Repeat composition***: high repeat content is challenging
4. ***Read length***: longer reads help resolve repeats
5. ***Error rate***: errors reduce coverage, obscure true overlaps

- Assembly is a hierarchical, starting from individual reads, build high confidence contigs/unitigs, incorporate the mates to build scaffolds
  - Extensive error correction is the key to getting the best assembly possible from a given data set
- Watch out for collapsed repeats & other misassemblies
  - Globally/Locally reassemble data from scratch with better parameters & stitch the 2 assemblies together

59

What should we expect from an assembly?

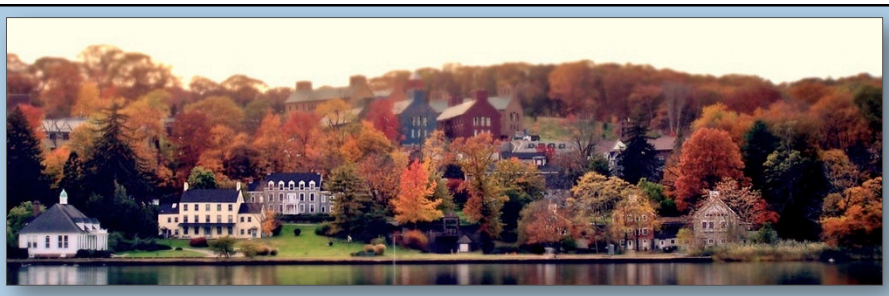
Analysis of dozens of genomes from across the tree of life with real and simulated data

Summary & Recommendations

< 100 Mbp:	HGAP/PacBio @ 100x expect near perfect chromosome arms
< 1GB:	HGAP/PacBio @ 100 high quality assembly: contig N50 over 1Mbp
> 1GB:	hybrid/gap filling expect contig N50 to be 100kbp – 1Mbp
> 5GB:	$$$$

Error correction and assembly complexity of single molecule sequencing reads.
Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz, MC
http://www.biorxiv.org/content/early/2014/06/18/006395

60



Thank you to Michael Schatz for sharing his lecture material and for his helpful insight.

http://schatzlab.cshl.edu
@mike_schatz

61