

Whole Genome Assembly with iPlant

Michael Schatz & Shoshana Marcus

Dec 4, 2013

CSHL Plant Genomes and Biotechnology



iPlant Collaborative™ *Empowering A New Plant Biology*

Outline

1. Assembly theory
 1. Assembly by analogy
 2. De Bruijn and Overlap graph
 3. Coverage, read length, errors, and repeats
2. Genome assemblers
 1. Assemblathon
 2. ALLPATHS-LG
 3. Celera Assembler
3. Assembly Tutorial with iPlant



Shredded Book Reconstruction

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

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	best	of	times,	it	was	the	worst	of	times,	it	was	the	age	of	wisdom,	it	was	the	age	of	foolishness,	...	
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	best	of	times,	it	was	the	worst	of	times,	it	was	the	age	of	wisdom,	it	was	the	age	of	foolishness,	...	
It	was	the	best	of	times,	it	was	the	worst	of	times,	it	was	the	age	of	wisdom,	it	was	the	age	of	foolishness,	...

- 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

Greedy 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

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

de Bruijn Graph Construction

- $D_k = (V, E)$
 - $V =$ 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

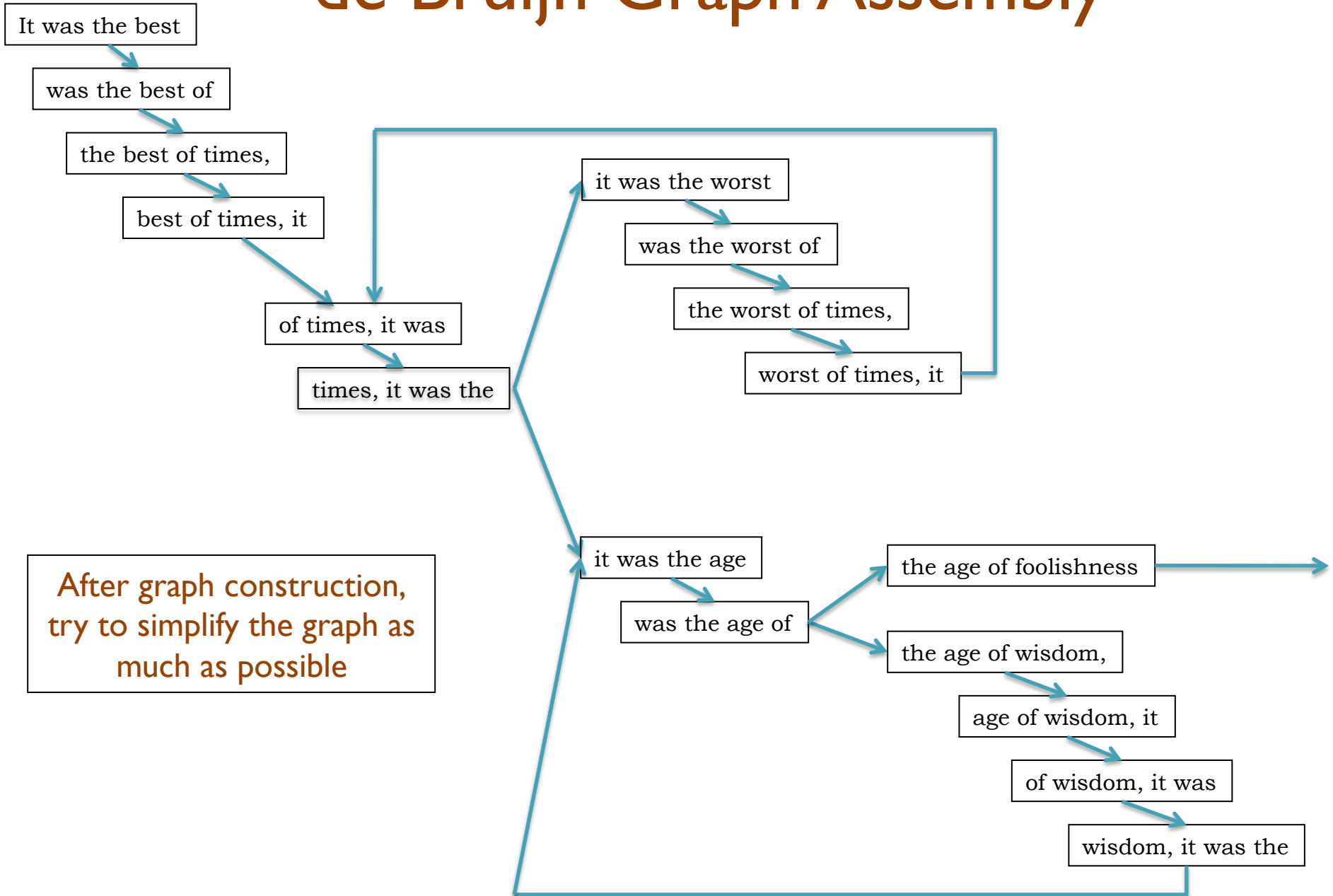
- Locally constructed graph reveals the global sequence structure
 - Overlaps between sequences implicitly computed

de Bruijn, 1946

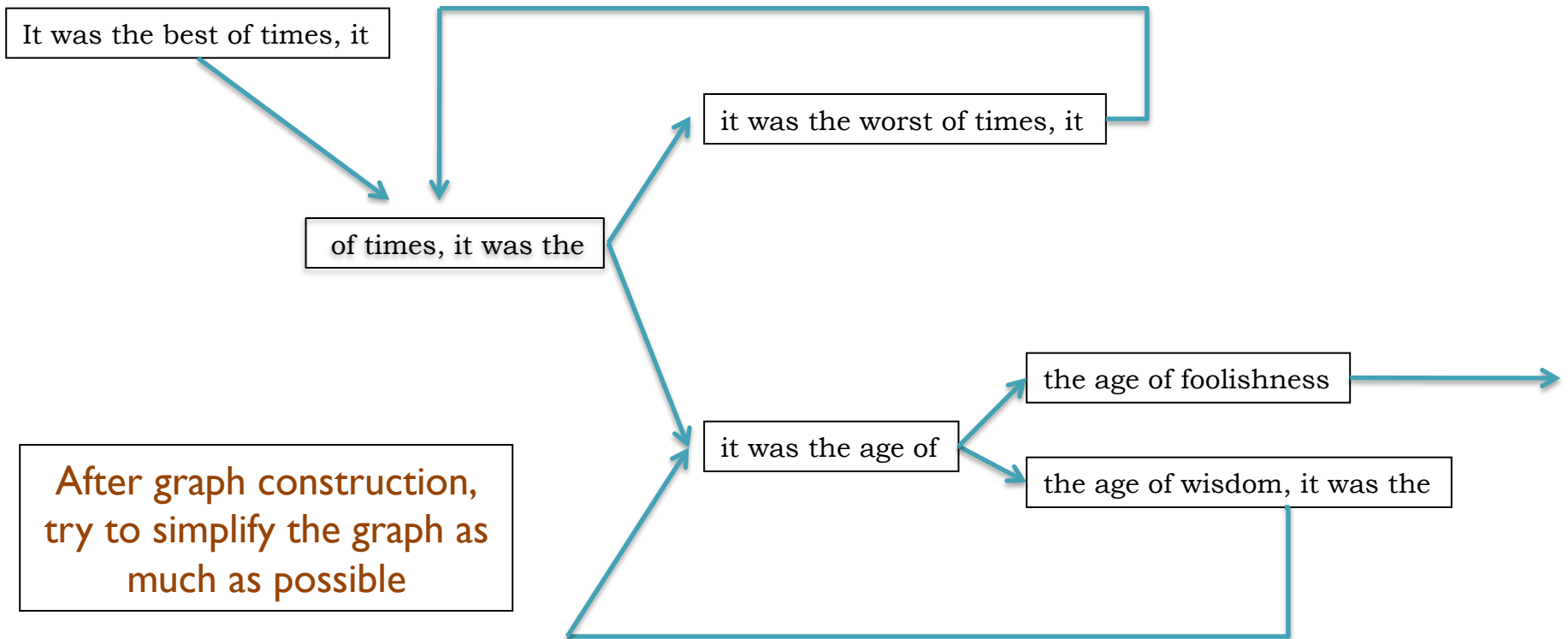
Idury and Waterman, 1995

Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly

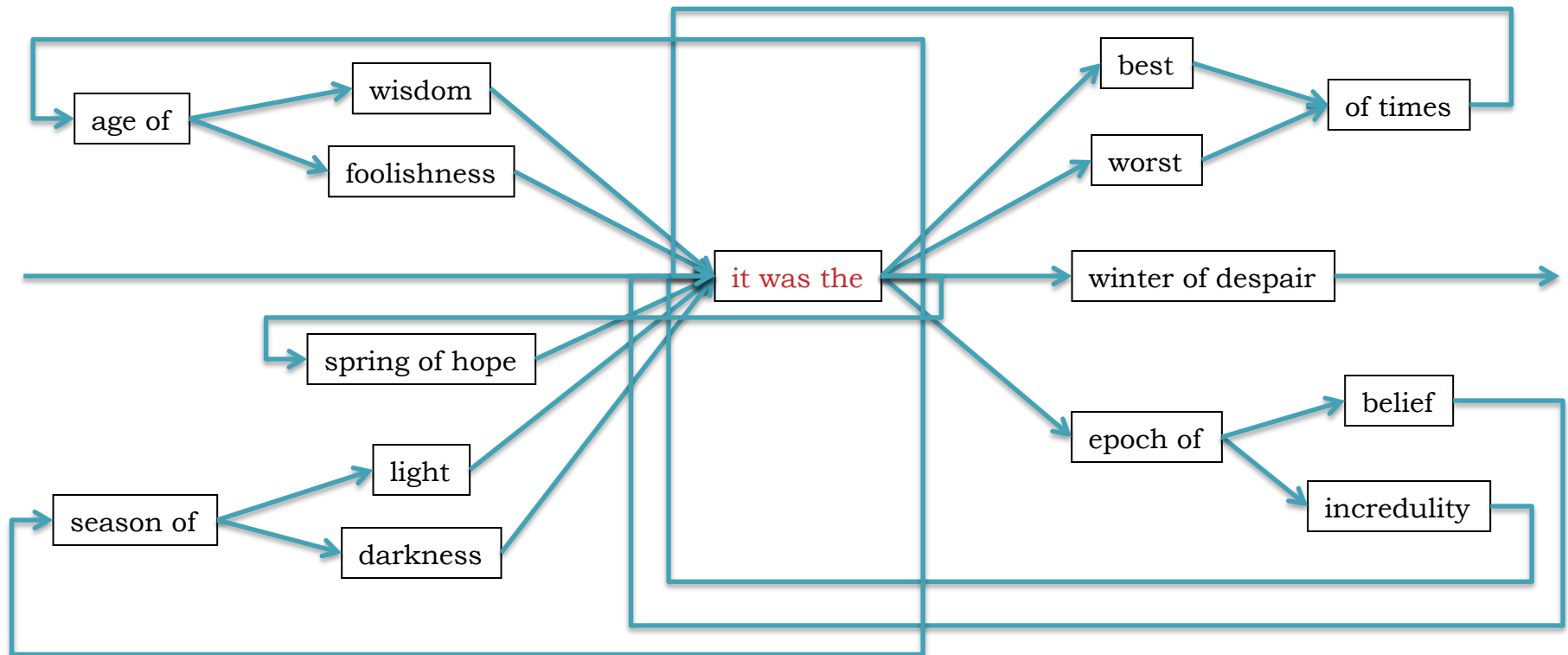


de Bruijn Graph Assembly



The full tale

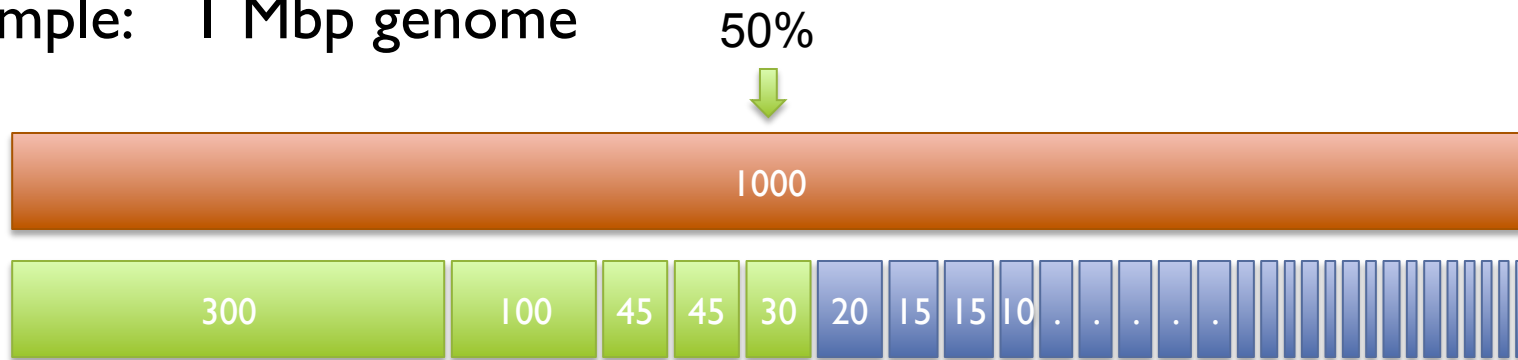
... it was the best of times it was the worst of times ...
... it was the age of wisdom it was the age of foolishness ...
... it was the epoch of belief it was the epoch of incredulity ...
... it was the season of light it was the season of darkness ...
... it was the spring of hope it was the winter of despair ...



N50 size

Def: 50% of the genome is in contigs as large as the N50 value

Example: 1 Mbp genome



N50 size = 30 kbp

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

Note:

N50 values are only meaningful to compare when base genome size is the same in all cases

Assembly Applications

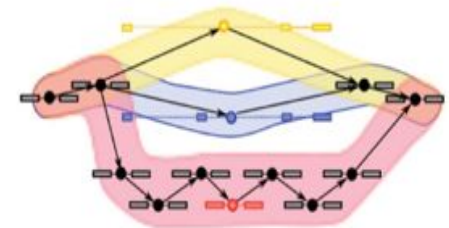
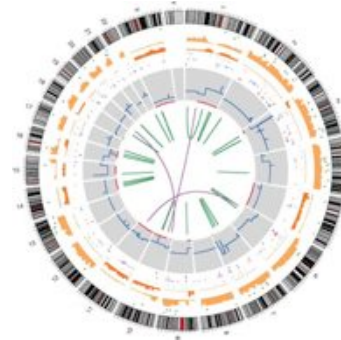
- Novel genomes



- Metagenomes

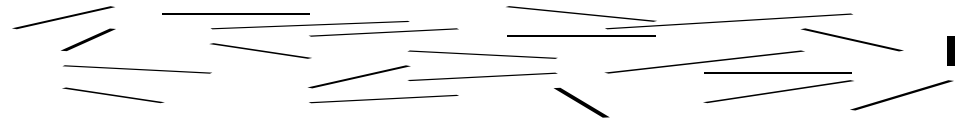


- Sequencing assays
 - Structural variations
 - Transcript assembly
 - ...



Assembling a Genome

1. Shear & Sequence DNA



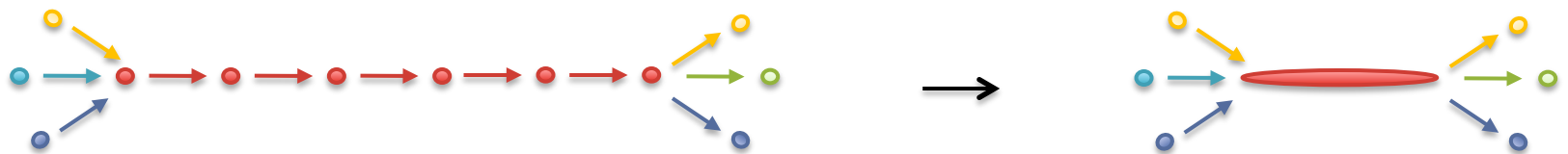
2. Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT

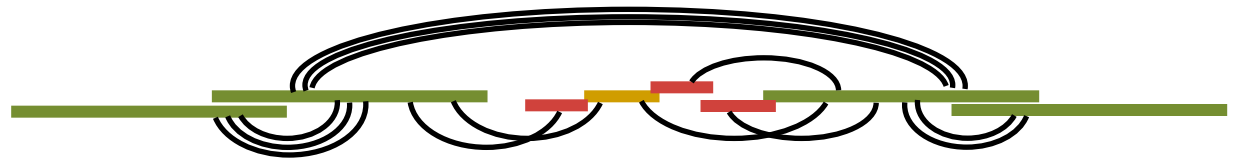
GGATGCGCGACACGT CGCATATCCGGTTTGGT CAACCTCGGACGGAC

CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph

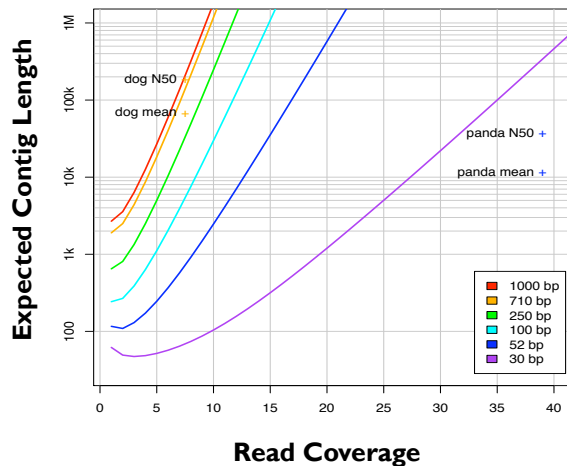


4. Detangle graph with long reads, mates, and other links



Ingredients for a good assembly

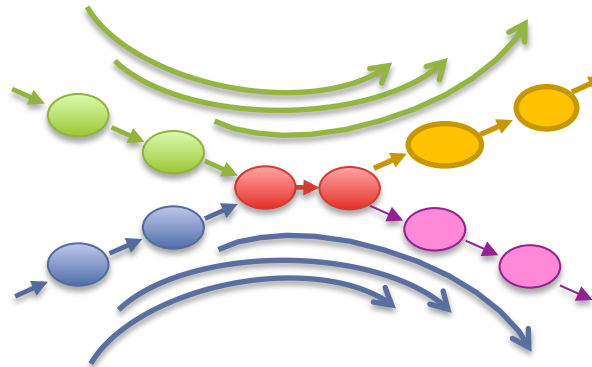
Coverage



High coverage is required

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

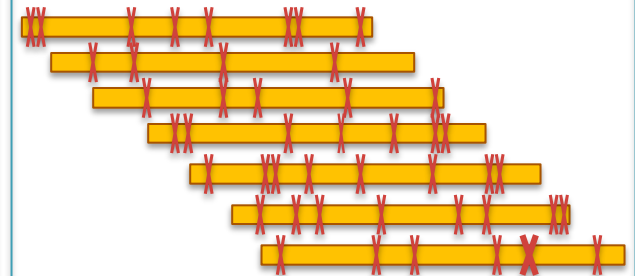
Read Length



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

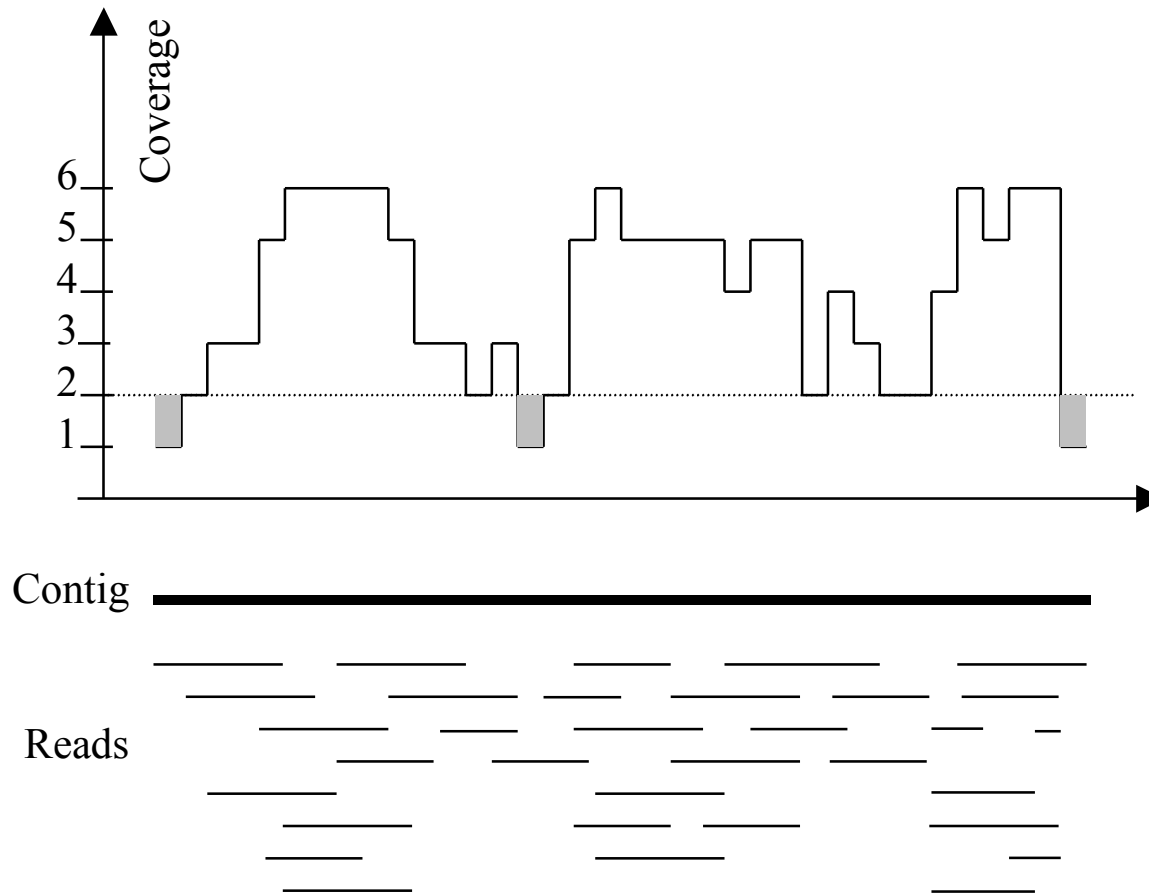
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in *de novo* plant genome sequencing and assembly

Schatz MC, Witkowski, McCombie, WVR (2012) *Genome Biology*. 12:243

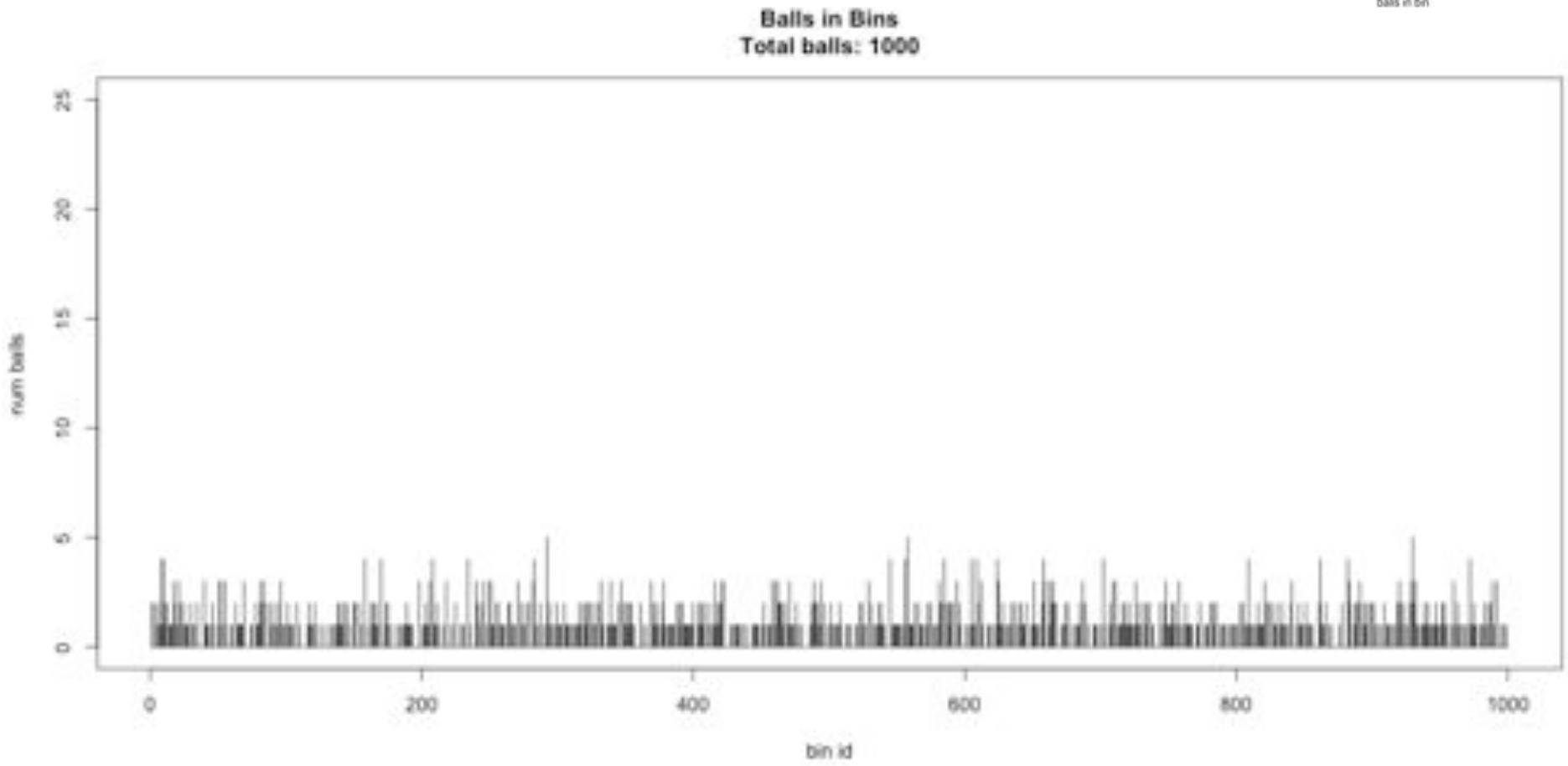
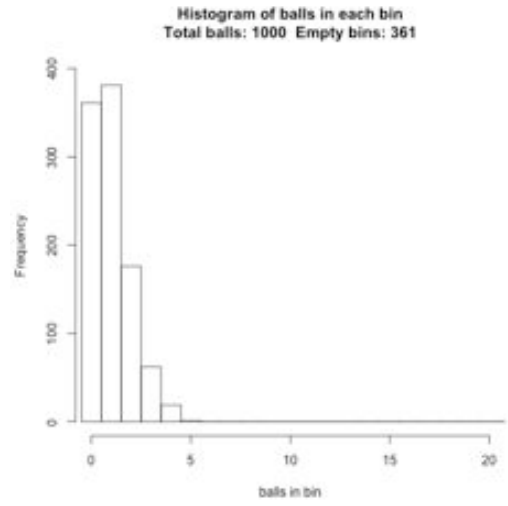
Coverage

Typical contig coverage

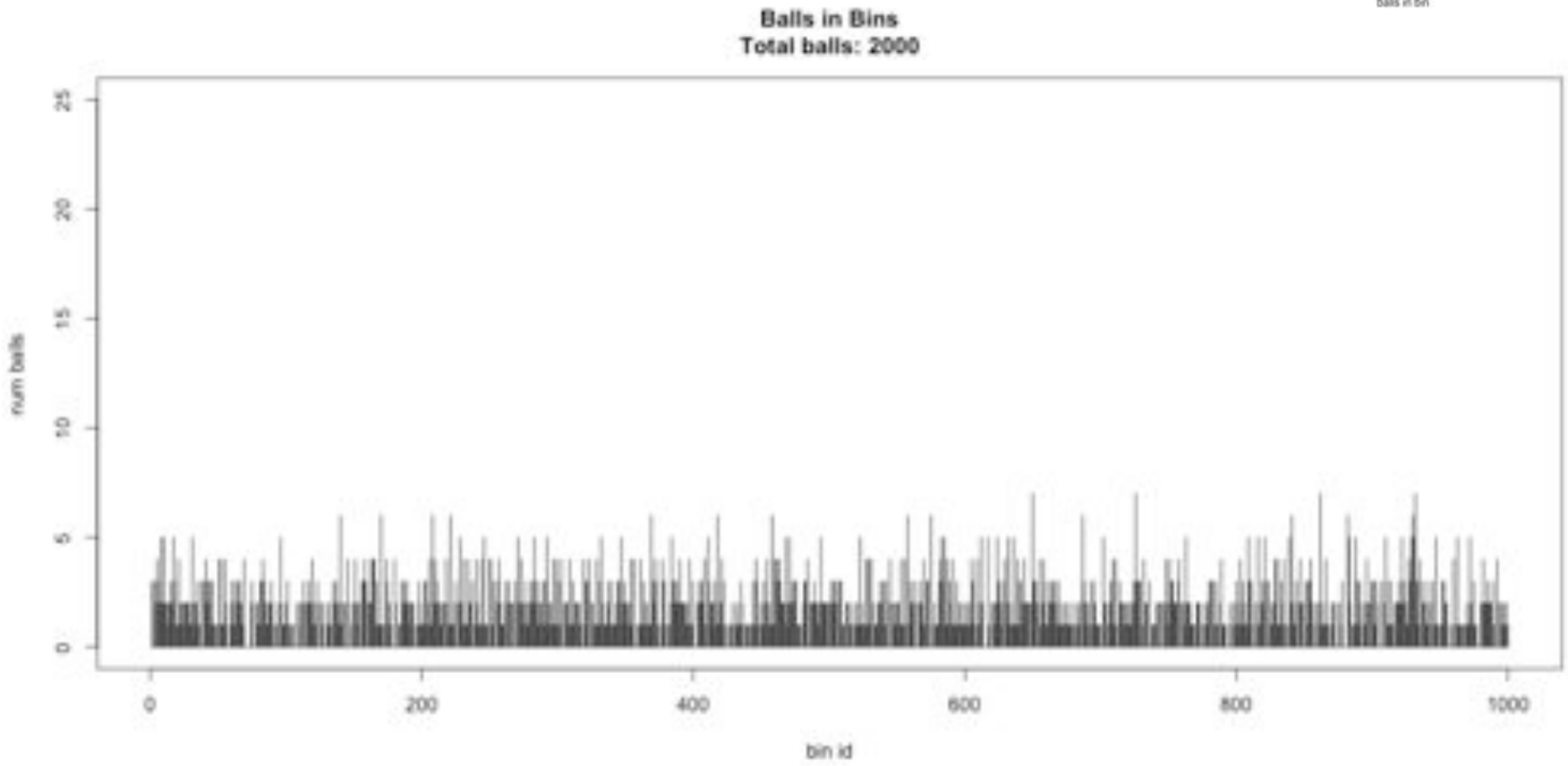
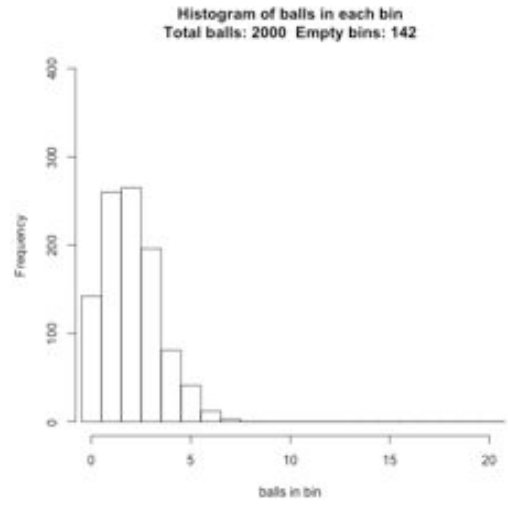


Imagine raindrops on a sidewalk

Balls in Bins Ix

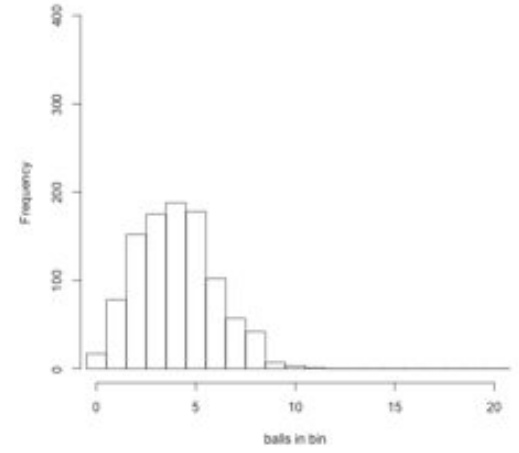


Balls in Bins 2x

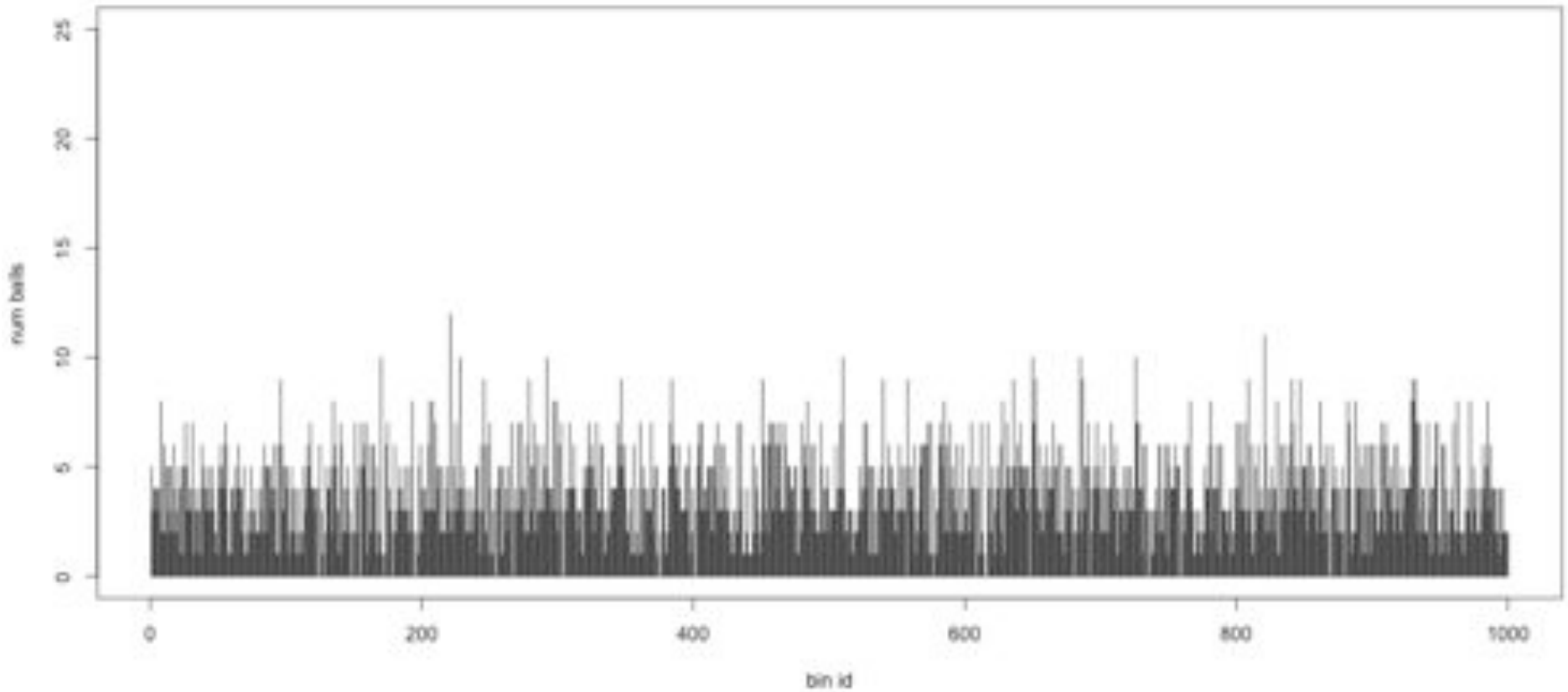


Balls in Bins 4x

Histogram of balls in each bin
Total balls: 4000 Empty bins: 17

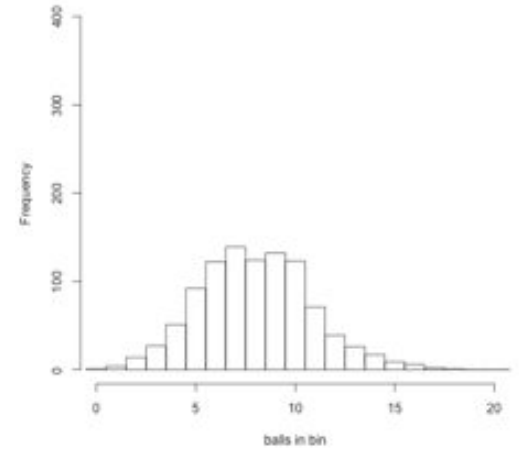


Balls in Bins
Total balls: 4000

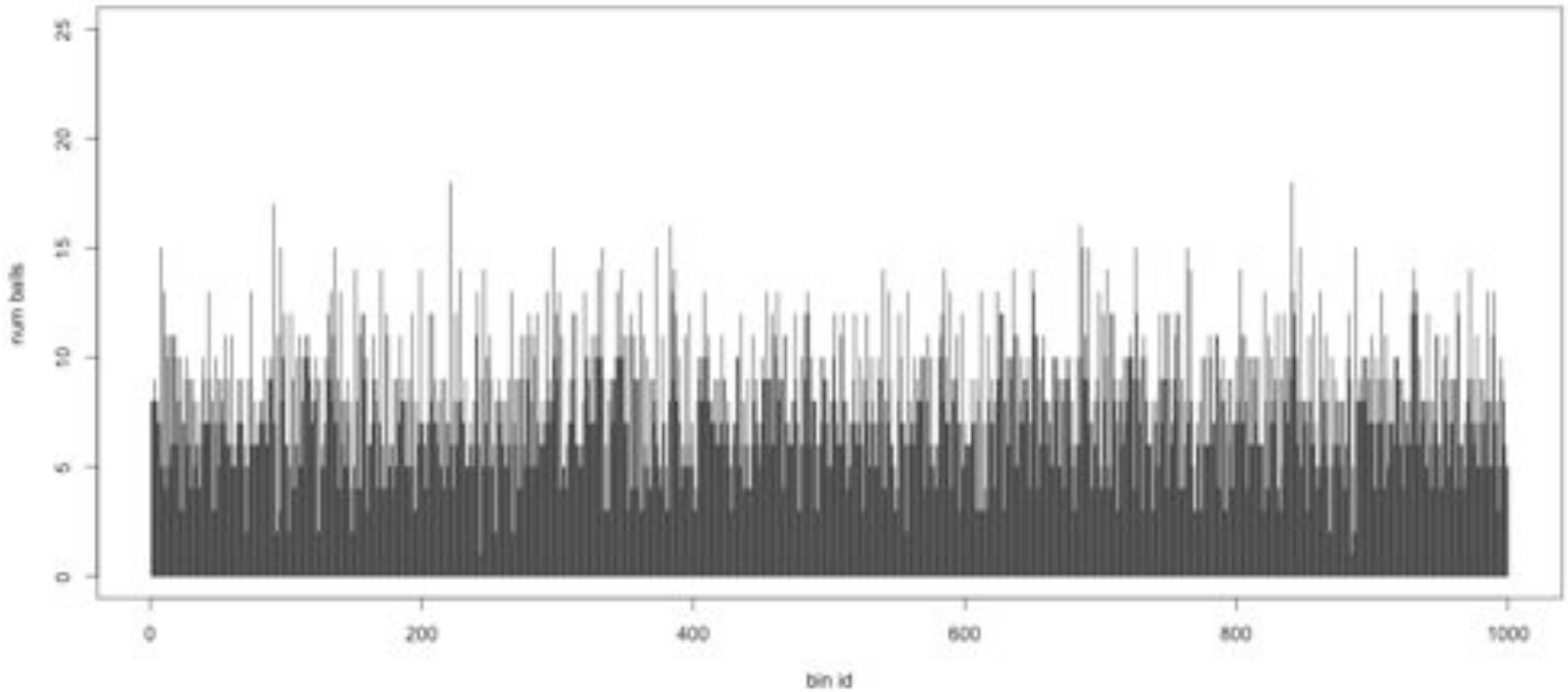


Balls in Bins 8x

Histogram of balls in each bin
Total balls: 8000 Empty bins: 1



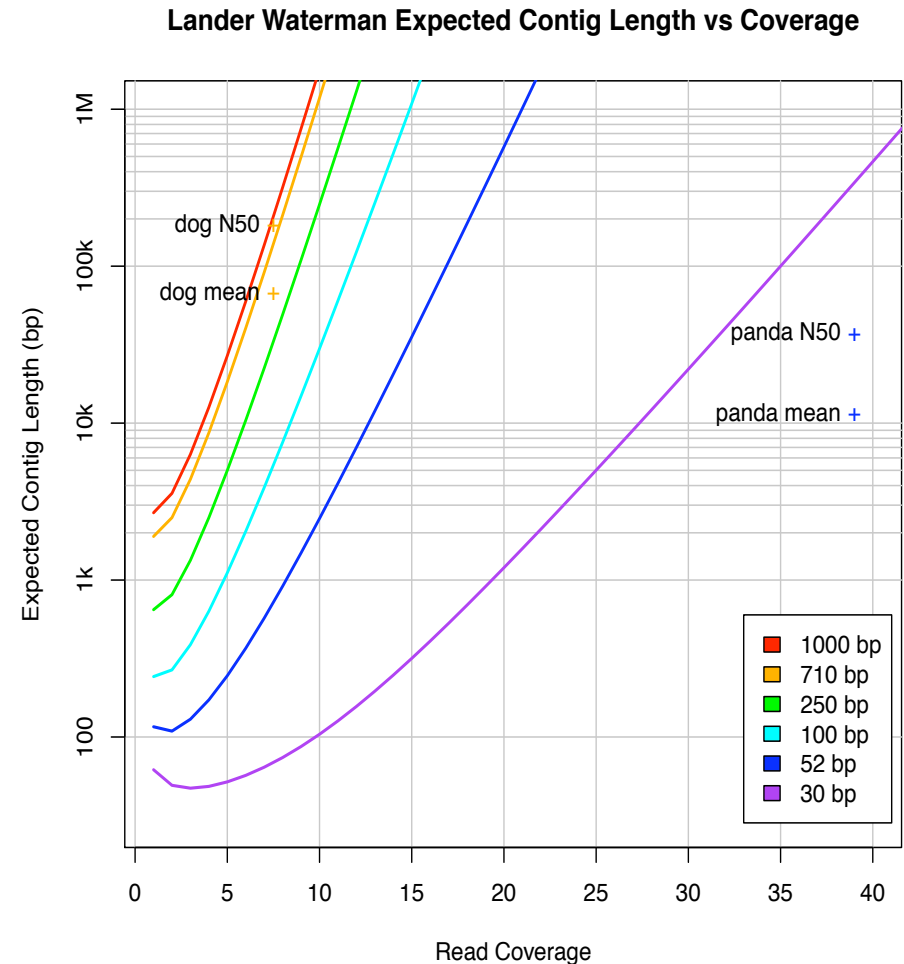
Balls in Bins
Total balls: 8000



Coverage and Read Length

Idealized Lander-Waterman model

- Reads start at perfectly random positions
- Contig length is a function of coverage and read length
 - Short reads require much higher coverage to reach same expected contig length
- Need even high coverage for higher ploidy, sequencing errors, sequencing biases
 - Recommend 100x coverage

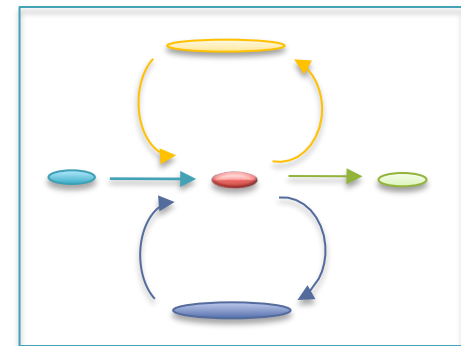
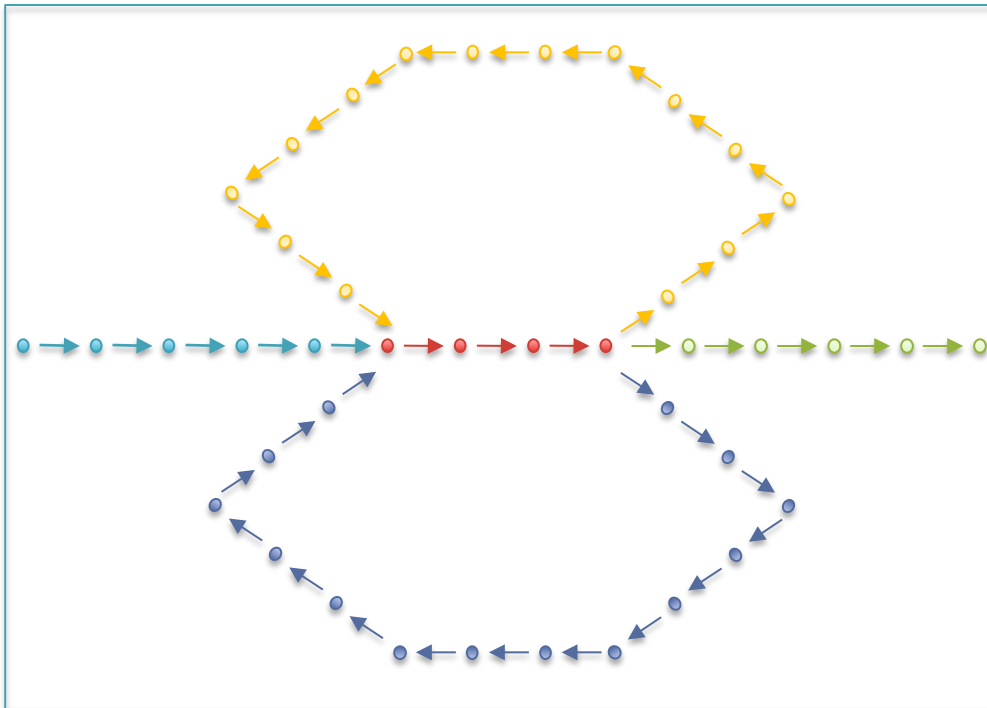


Assembly of Large Genomes using Second Generation Sequencing

Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research*. 20:1165-1173.

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, and (3) repeats



Errors in the graph



(Chaisson, 2009)

Clip Tips	Pop Bubbles
<p data-bbox="846 537 1247 597">was the worst of times,</p> <p data-bbox="846 651 1247 711">was the worst of tymes,</p> <p data-bbox="865 756 1228 816">the worst of times, it</p>	<p data-bbox="1486 518 1887 578">was the worst of times,</p> <p data-bbox="1486 607 1887 667">was the worst of tymes,</p> <p data-bbox="1505 698 1869 758">times, it was the age</p> <p data-bbox="1495 787 1879 847">tymes, it was the age</p>
<p data-bbox="926 1068 1264 1128">the worst of tymes,</p> <p data-bbox="846 1162 1142 1222">was the worst of</p> <p data-bbox="915 1256 1245 1317">the worst of times,</p> <p data-bbox="1016 1351 1316 1411">worst of times, it</p>	<p data-bbox="1619 1068 1766 1128">tymes,</p> <p data-bbox="1381 1162 1682 1222">was the worst of</p> <p data-bbox="1717 1162 1971 1222">it was the age</p> <p data-bbox="1612 1256 1749 1317">times,</p>

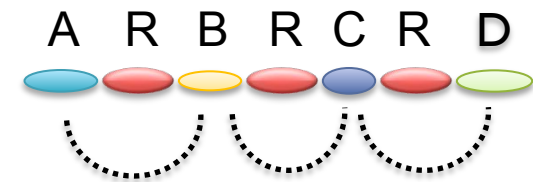
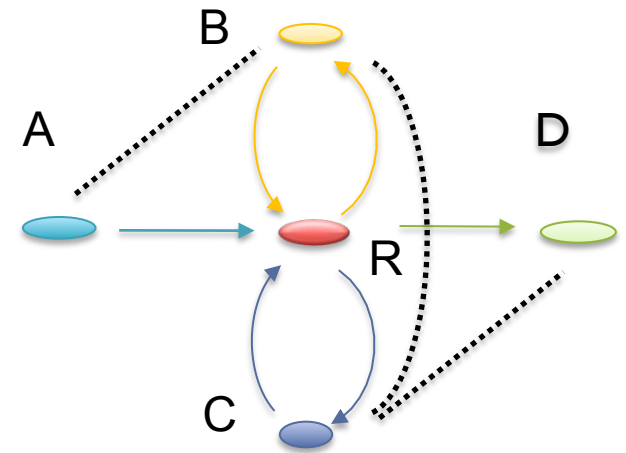
Repetitive regions

Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1b_2\dots b_k)^N$ where $1 \leq k \leq 6$ CACACACACACACACACA	2%
SINEs (Short Interspersed Nuclear Elements)	<i>Alu</i> 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
 - Wheat: 16 Gbp; Pine: 24 Gbp

Scaffolding

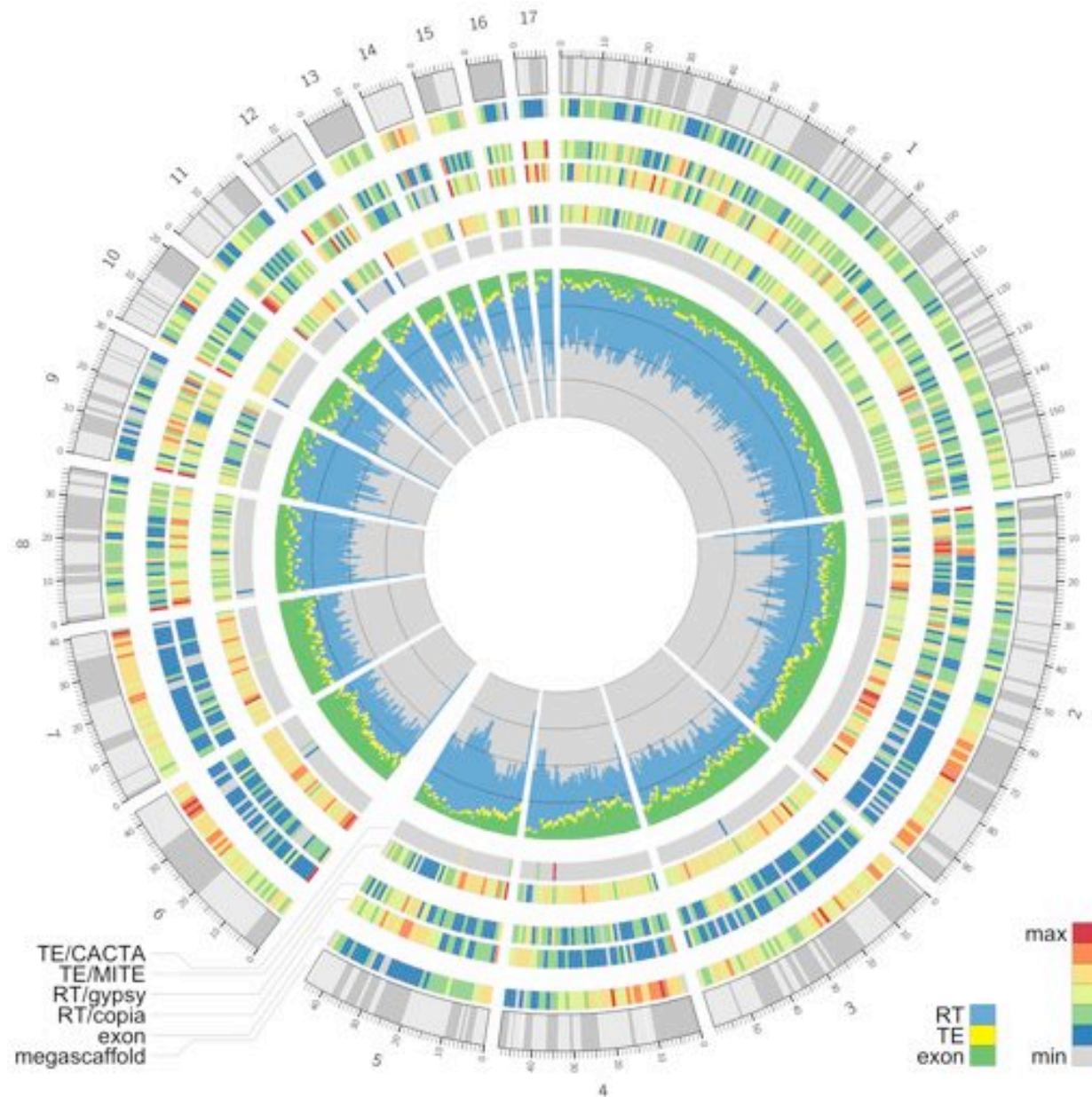
- Initial contigs (*aka* unipaths, unitigs) terminate at
 - Coverage gaps: especially extreme GC
 - Conflicts: errors, repeat boundaries
- Use mate-pairs to resolve correct order through assembly graph
 - Place sequence to satisfy the mate constraints
 - Mates through repeat nodes are tangled
- Final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



Post-assembly Analysis

After assembly:

- Validation
- CEGMA
- BLAST
- Gene Finding
- Repeat mask
- RNA-seq
- *-seq
- ...
- Publish! 😊



Outline

1. Assembly theory
 1. Assembly by analogy
 2. De Bruijn and Overlap graph
 3. Coverage, read length, errors, and repeats
2. **Genome assemblers**
 1. Assemblathon
 2. ALLPATHS-LG
 3. Celera Assembler
3. Assembly Tutorial with iPlant



THE ASSEMBLATHON

- Attempt to answer the question:
 “What makes a good assembly?”
- Organizers provided sequence data to assembly experts around the world
 - Assemblathon 1: ~100Mbp simulated genome
 - Assemblathon 2: 3 vertebrate genomes each ~1GB
- Results demonstrate trade-offs assemblers must make

Assemblathon 1: A competitive assessment of de novo short read assembly methods.

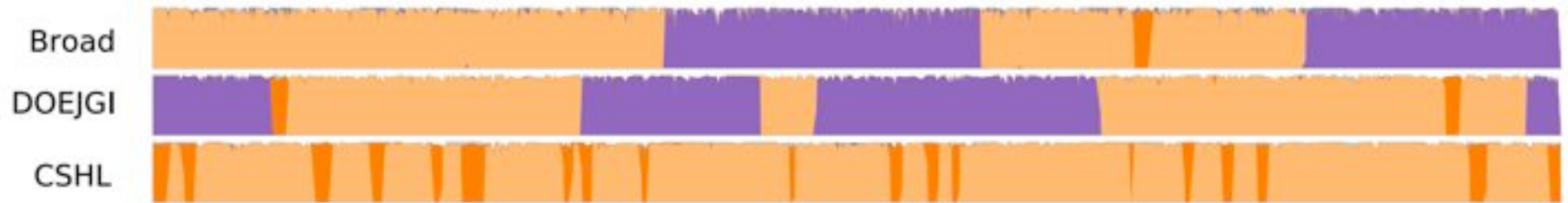
Earl, DA, et al. (2011) Genome Research. doi: 10.1101/gr.126599.111

Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species

Bradnam, KR. et al (2013) GigaScience 2:10 doi:10.1186/2047-217X-2-10

Assembly Results

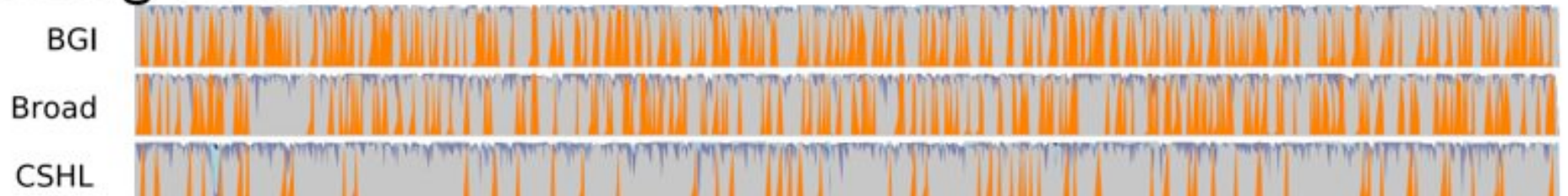
Scaffolds



Scaffold Paths



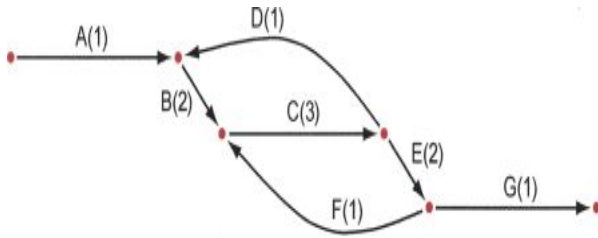
Contig Paths



Final Rankings

ID	Overall	CPNG50	SPNG50	Struct.	CC50	Subs.	Copy. Num.	Cov. Tot.	Cov. CDS
BGI	36	★					★	★	★
Broad	37	★	★	★	★				
WTSI-S	46		★	★	★	★			
CSHL	52	★							★
BCCGSC	53							★	★
DOEJGI	56		★	★	★	★			
RHUL	58								
WTSI-P	64							★	
EBI	64						★		
CRACS	64					★			

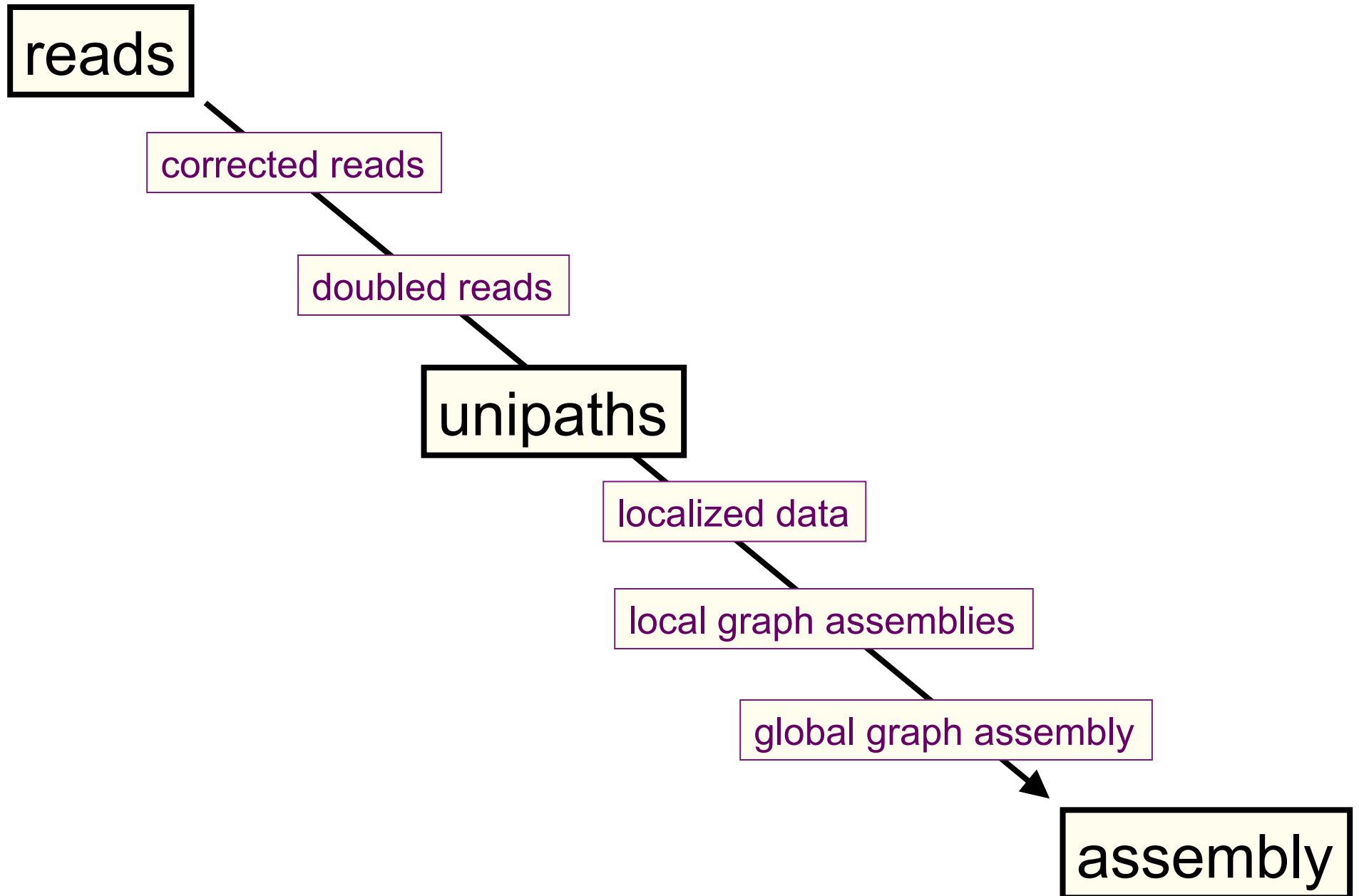
- ALLPATHS and SOAPdenovo came out neck-and-neck followed closely behind by Celera Assembler, SGA, and ABySS
- My recommendation for “typical” short read assembly is to use ALLPATHS
- Single molecule sequencing becoming extremely attractive if you have access



Genome assembly with ALLPATHS-LG

Iain MacCallum

How ALLPATHS-LG works



ALLPATHS-LG sequencing model

Libraries (insert types)	Fragment size (bp)	Read length (bases)	Sequence coverage (x)	Required
Fragment	180*	≥ 100	45	yes
Short jump	3,000	≥ 100 preferable	45	yes
Long jump	6,000	≥ 100 preferable	5	no**
Fosmid jump	40,000	≥ 26	1	no**

*See next slide.

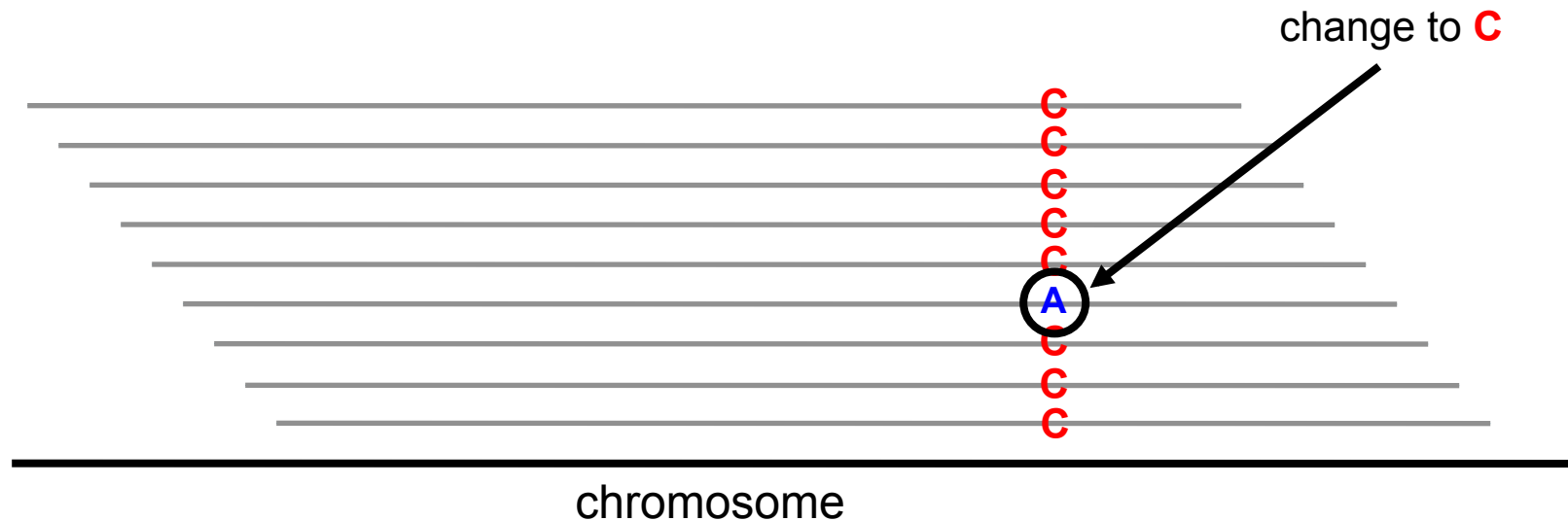
**For best results. Normally not used for small genomes.
However essential to assemble long repeats or duplications.

Cutting coverage in half still works, with some reduction in quality of results.

All: protocols are either available, or in progress.

Error correction

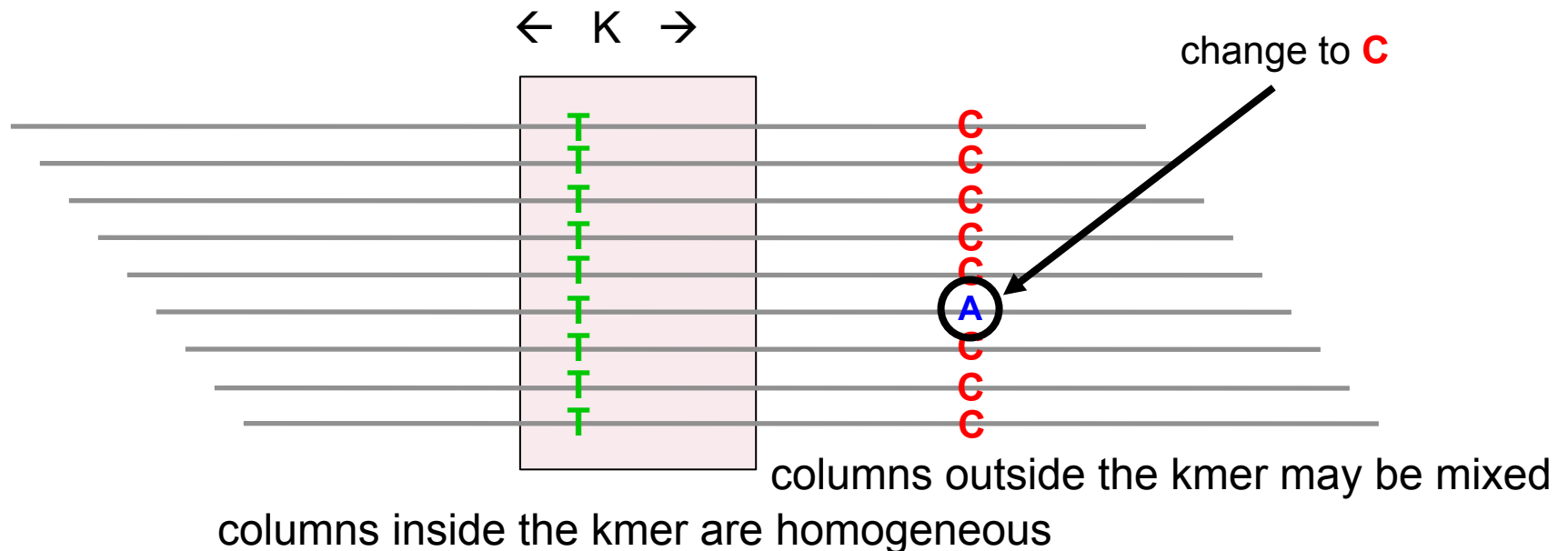
Given a crystal ball, we could stack reads on the chromosomes they came from (with homologous chromosomes separate), then let each column 'vote':



But we don't have a crystal ball....

Error correction

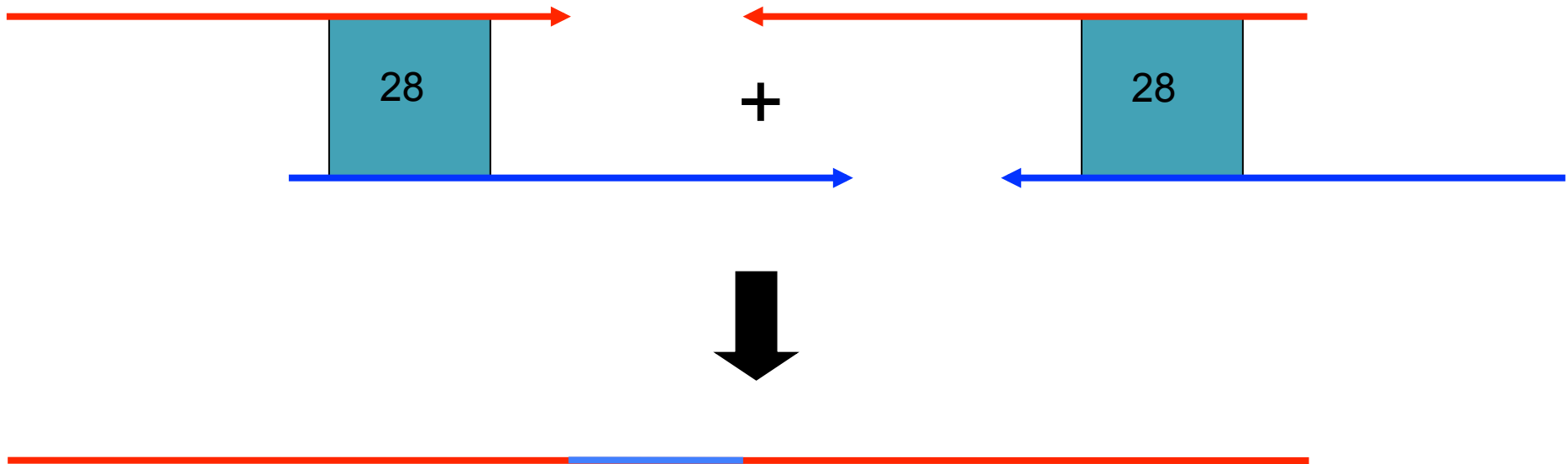
ALLPATHS-LG. For every K-mer, examine the stack of all reads containing the K-mer. Individual reads may be edited if they differ from the overwhelming consensus of the stack. If a given base on a read receives conflicting votes (arising from membership of the read in multiple stacks), it is not changed. (K=24)



Two calls at Q20 or better are enough to protect a base

Read doubling

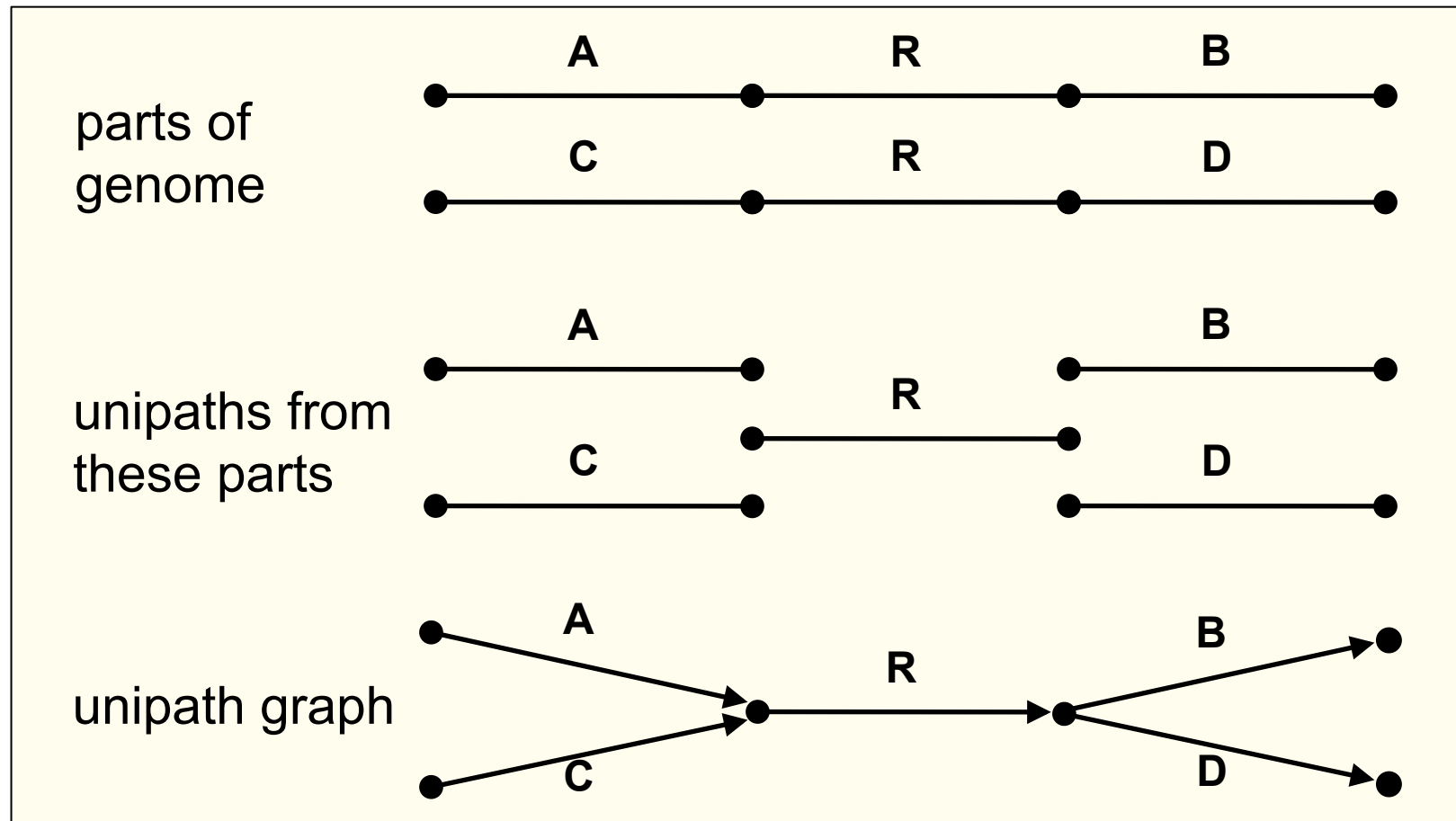
To close a read pair (red), we require the existence of another read pair (blue), overlapping perfectly like this:



More than one closure allowed (but rare).

Unipaths

Unipath: unbranched part of genome – squeeze together perfect repeats of size $\geq K$



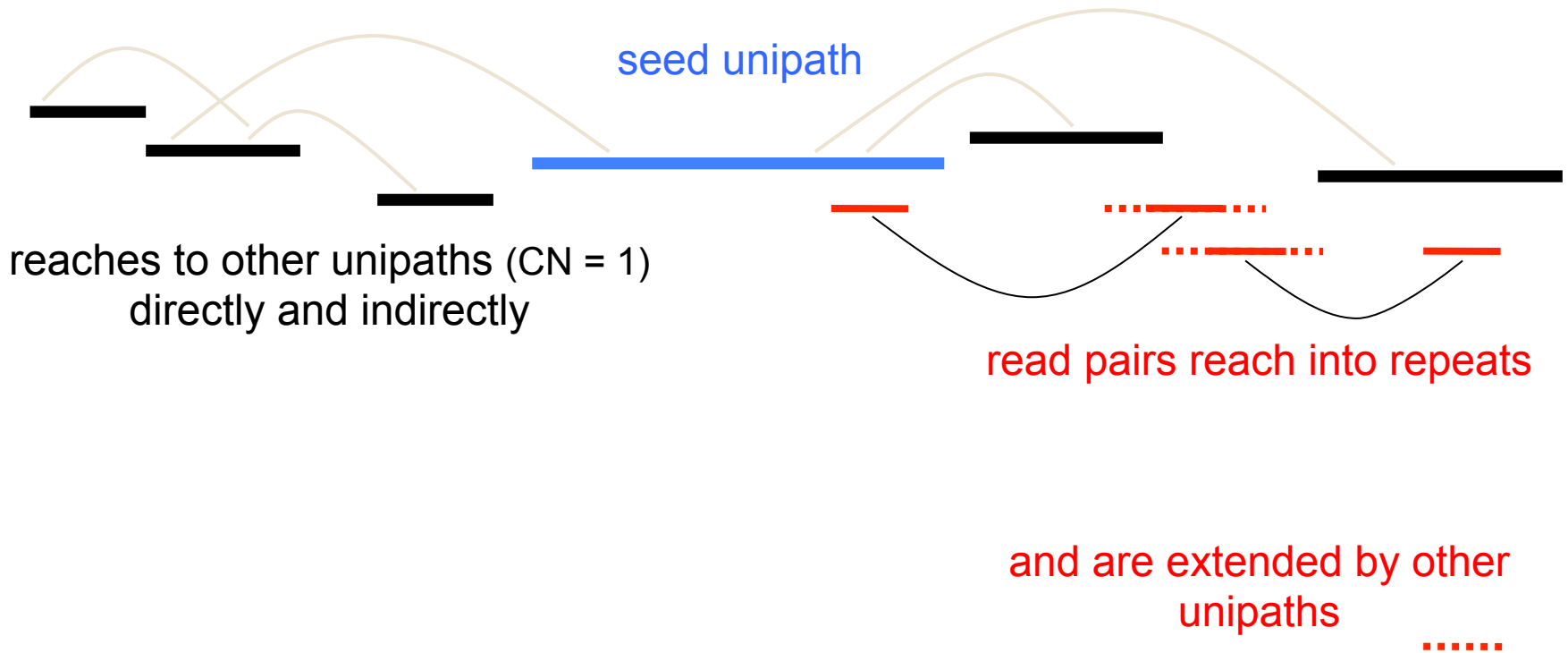
Adjacent unipaths overlap by $K-1$ bases

Localization

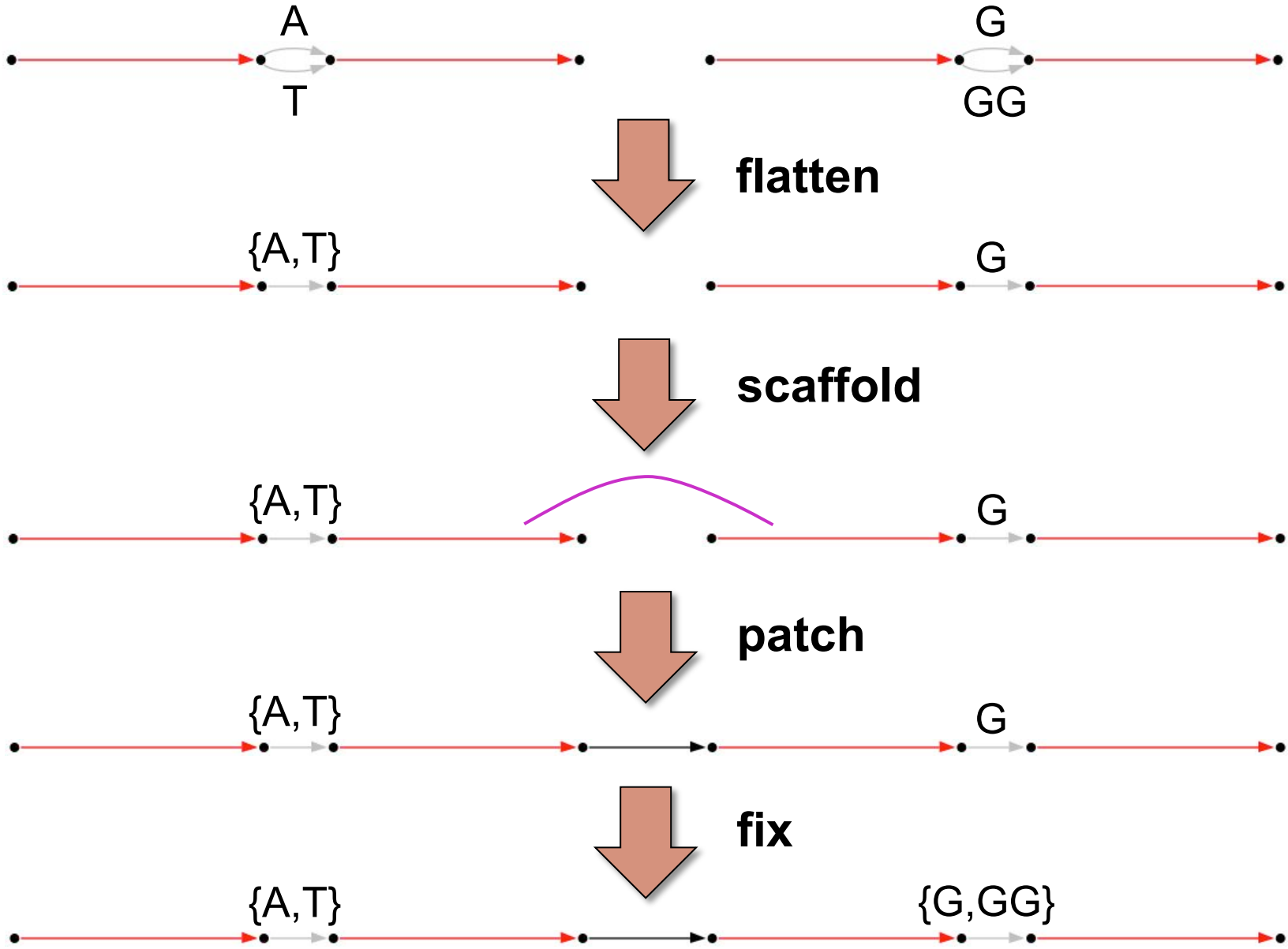
I. Find 'seed' unipaths, evenly spaced across genome
(ideally long, of copy number $CN = 1$)



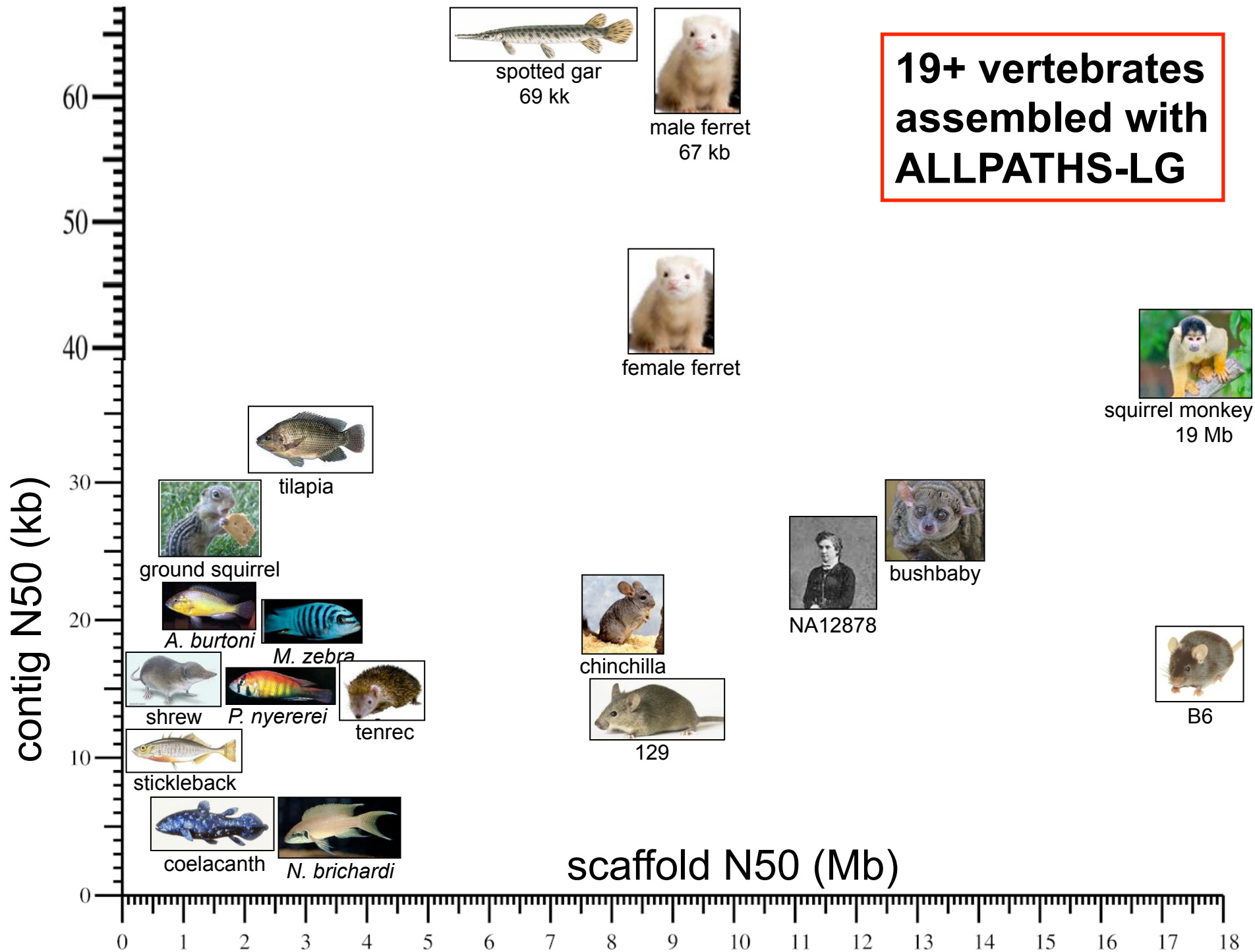
II. Form neighborhood around each seed

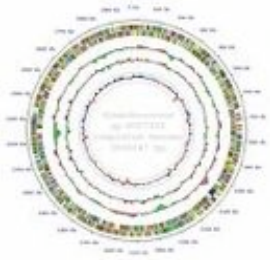


Create assembly from global assembly graph



**19+ vertebrates
assembled with
ALLPATHS-LG**



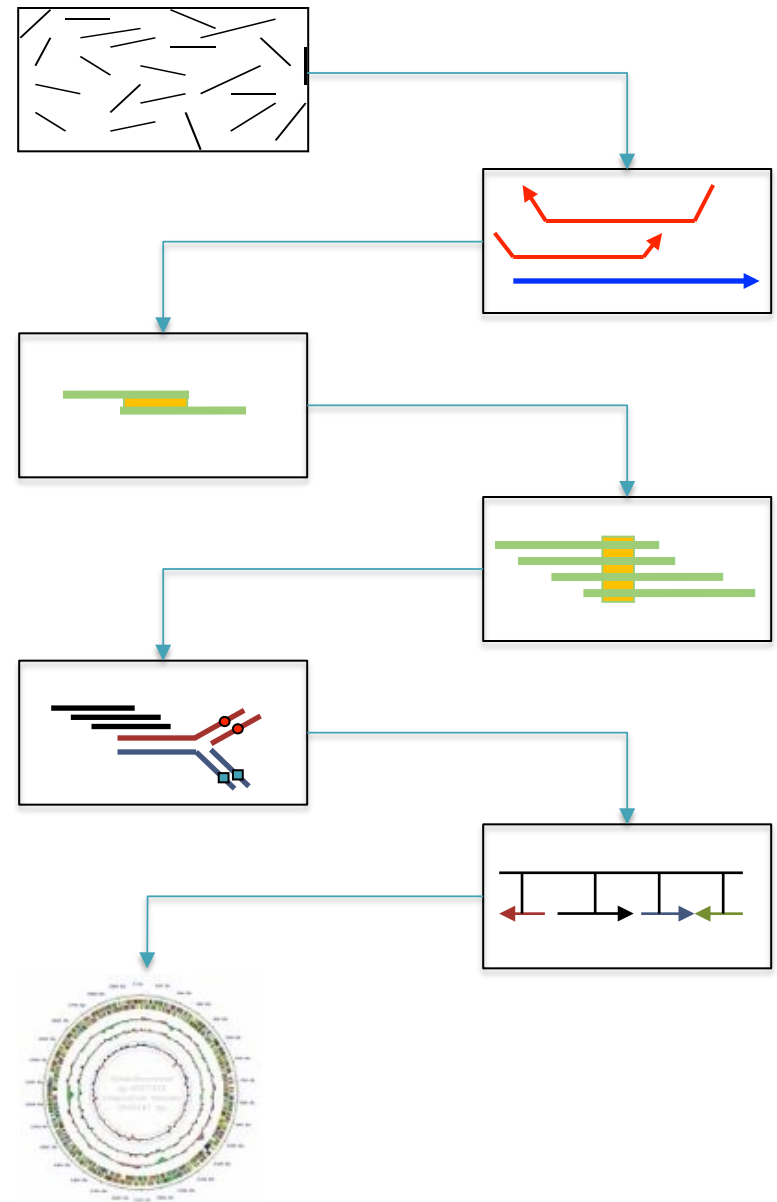


Genome assembly with the Celera Assembler

Celera Assembler

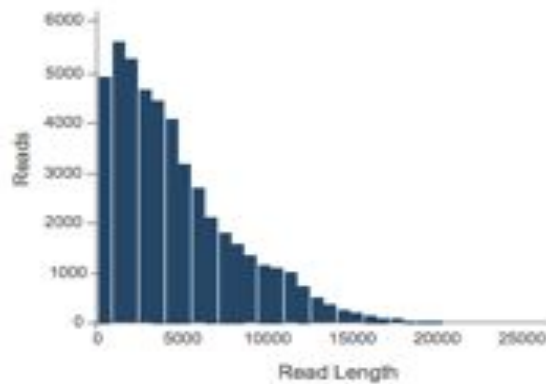
<http://wgs-assembler.sf.net>

1. Pre-overlap
 - Consistency checks
2. Trimming
 - Quality trimming & partial overlaps
3. Compute Overlaps
 - Find high quality overlaps
4. Error Correction
 - Evaluate difference in context of overlapping reads
5. Unitigging
 - Merge consistent reads
6. Scaffolding
 - Bundle mates, Order & Orient
7. Finalize Data
 - Build final consensus sequences

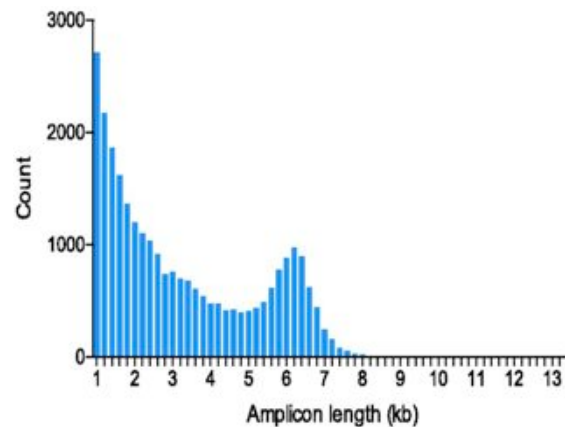


Single Molecule Sequencing Technology

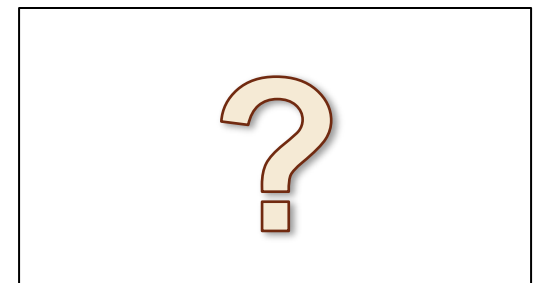
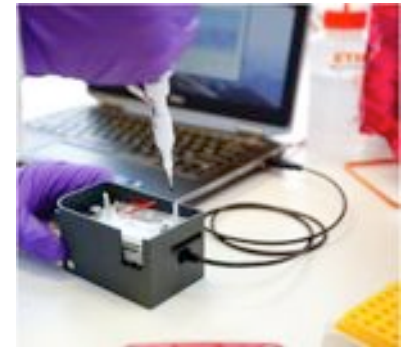
PacBio RS II



Moleculo



Oxford Nanopore



Hybrid Sequencing



Illumina

Sequencing by Synthesis

High throughput (60Gbp/day)

High accuracy (~99%)

Short reads (~100bp)



Pacific Biosciences

SMRT Sequencing

Lower throughput (1Gbp/day)

Lower accuracy (~85%)

Long reads (5kbp+)

Hybrid Error Correction: PacBioToCA

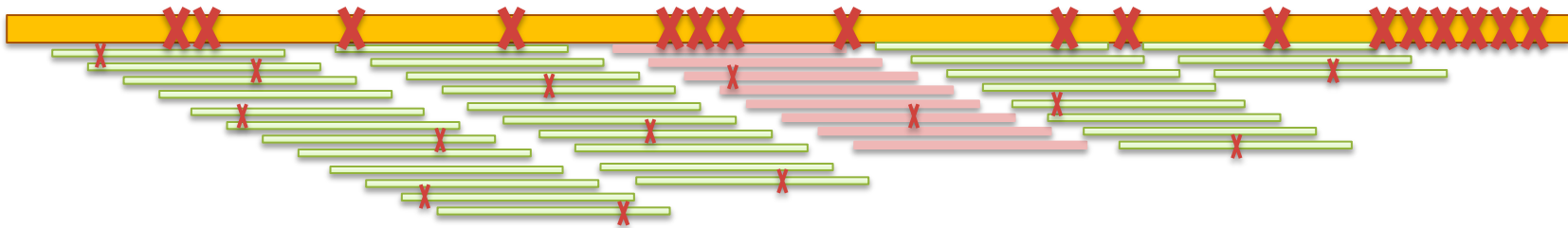
<http://wgs-assembler.sf.net>

I. Correction Pipeline

1. Map short reads to long reads
2. Trim long reads at coverage gaps
3. Compute consensus for each long read



2. Error corrected reads can be easily assembled, aligned

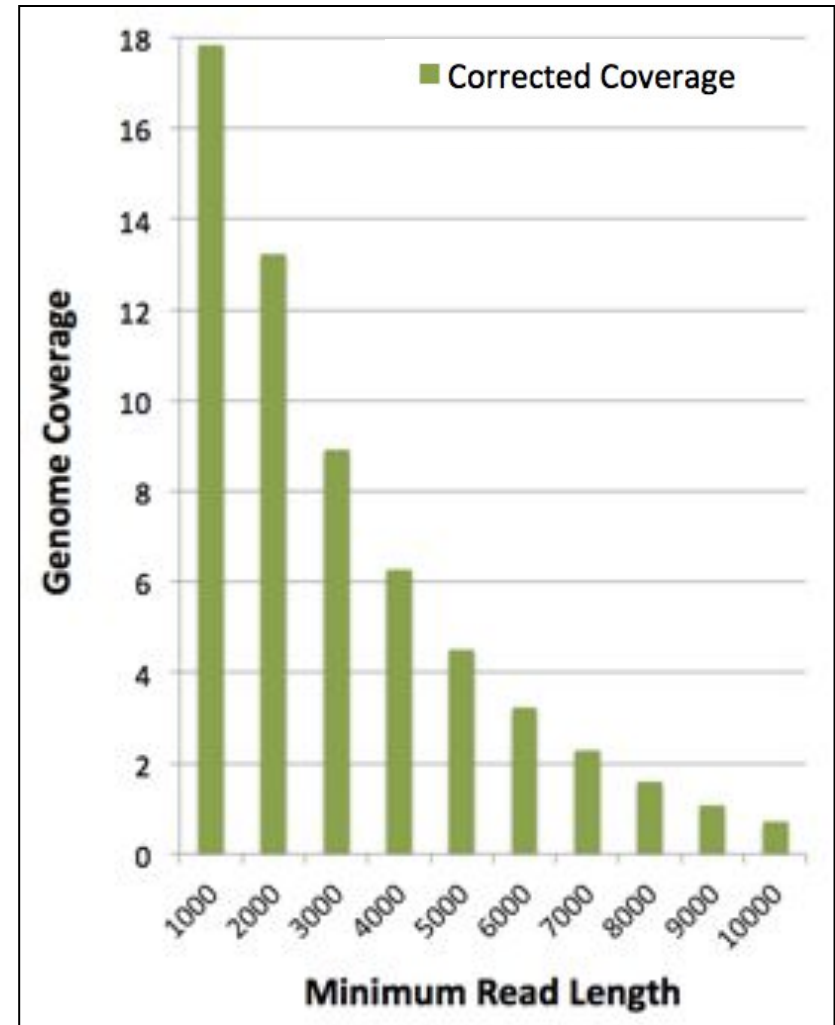


Hybrid error correction and de novo assembly of single-molecule sequencing reads.

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

Preliminary Rice Assemblies

Assembly	Contig NG50
HiSeq Fragments 50x 2x100bp @ 180	3,925
MiSeq Fragments 23x 459bp 8x 2x251bp @ 450	6,332
“ALLPATHS-recipe” 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18,248



In collaboration with McCombie & Ware labs @ CSHL

Assembly Summary

Assembly quality depends on

1. **Coverage**: low coverage is mathematically hopeless
 2. **Repeat composition**: high repeat content is challenging
 3. **Read length**: longer reads help resolve repeats
 4. **Error rate**: errors reduce coverage, obscure true overlaps
- Assembly is a hierarchical
 - Reads -> unitigs -> mates -> scaffolds
 - > optical / physical / genetic maps
 - > chromosomes
 - Recommendations:
 - ALLPATH-LG for Illumina-only
 - HGAP for PacBio-only, CA for Hybrid assembly
 - See Assemblathon papers for a more extensive analysis



Outline

1. Assembly theory
 1. Assembly by analogy
 2. De Bruijn and Overlap graph
 3. Coverage, read length, errors, and repeats
2. Genome assemblers
 1. Assemblathon
 2. ALLPATHS-LG
 3. Celera Assembler
3. **Assembly Tutorial with iPlant**



Assembly with ALLPATHS-LG

0. Download and install ALLPATHS-LG source code

```
% wget ftp://ftp.broadinstitute.org/pub/crd/7/ALLPATHS/Release-LG/  
% configure && make install
```

1. Collect the BAM files that you want to assemble. Create a

```
in_libs.csv file to describe your libraries. Create a groups.csv  
metadata file to describe your groups.
```

2. Prepare input files

```
% cd /tmp/csh1/asm  
% PrepareAllPathsInput  
DATA_DIR=`pwd`
```

3. Assemble.

```
% RunAllPaths  
PRE=/tmp/csh1/asm  
DATA_SUBDIR=default >>
```

4. Get the results (four

```
% cd /tmp/csh1/asm/default/ASSEMBLIES/test/  
% less final.{assembly,contigs}.{fasta,efasta}
```

Assembly with iPlant

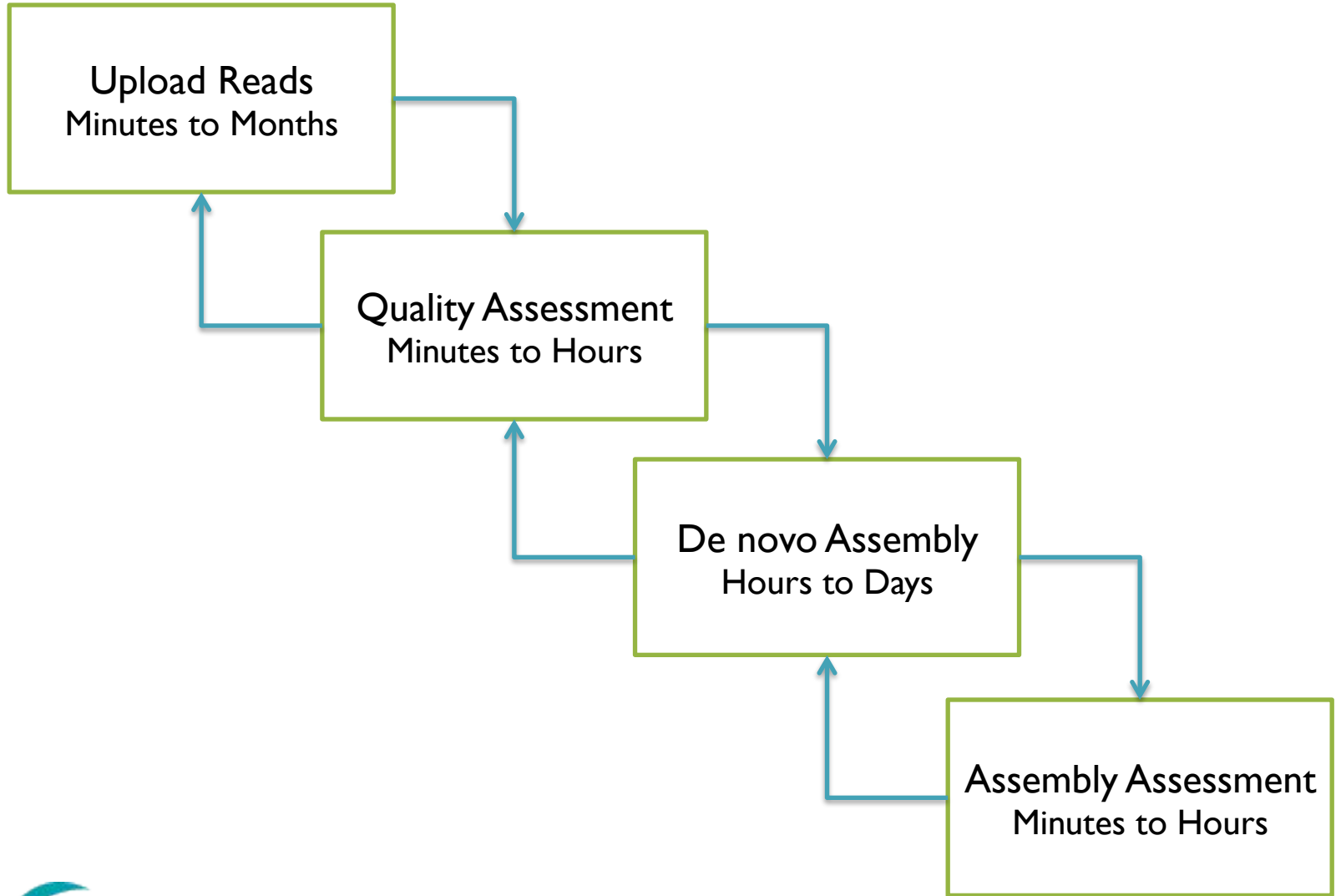


The screenshot shows the iPlant Collaborative Discovery Environment web interface. The browser address bar displays `https://preview.iplantcollaborative.org/de/#workspace`. The page title is "(4) Discovery Environment". The interface includes a navigation sidebar with icons for Data, Apps, and Analysis. The main content area shows a search for "allpaths" with one result: "AllpathsLG 44837" integrated by Roger Barthelson. A modal window is open for "AllpathsLG 44837", showing an analysis name "AllpathsLG 44837_analysis1" and a description field.



iPlant Collaborative™ *Empowering A New Plant Biology*

Assembly Workflow



Upload Reads



The screenshot displays the iPlant Collaborative Discovery Environment interface. The main window shows a file browser with a navigation pane on the left containing folders like 'machat', 'analyses', 'gold', 'asm_project', 'Community Data', 'Shared With Me', and 'Trash'. The main pane shows 'No items to display'. An 'Upload' dialog box is open in the foreground, displaying the following information:

Upload
Maximum size of each file is 1.9GB when using simple upload.
Uploading to /iplant/home/machat/asm_project.

File Name	Action
frag180.1.fq	Browse...
frag180.2.fq	Browse...
jump2k.1.fq	Browse...
jump2k.2.fq	Browse...
	Browse...

Buttons: Upload, Cancel



QC: FastQC

Discovery Environment

file://localhost/Users/mschatz/Downloads/frag_1_fastqc/...

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Contamination

Metric	Value
Filename	frag_1_fastqc
File type	FASTQ (gzipped)
Encoding	ASCII (12 bit)
Technical requirements	ok
Processed sequences	4
Sequence length	100
MD5	AA

Basic Statistics

Quality score distribution of each cycle (cycles 1-100)

Per base sequence quality

Quality score distribution for all sequences

Per sequence quality scores

Quality score distribution for all sequences

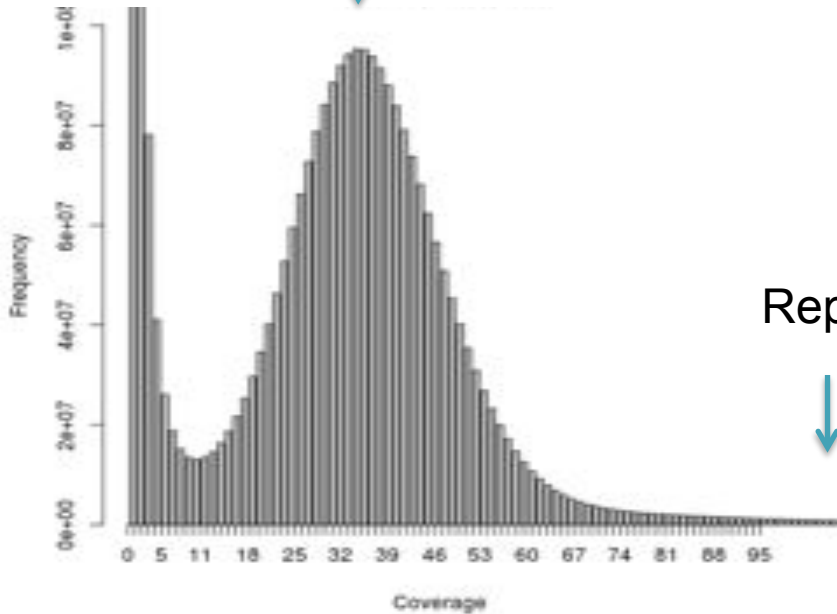
QC: Read Coverage

Reference: 

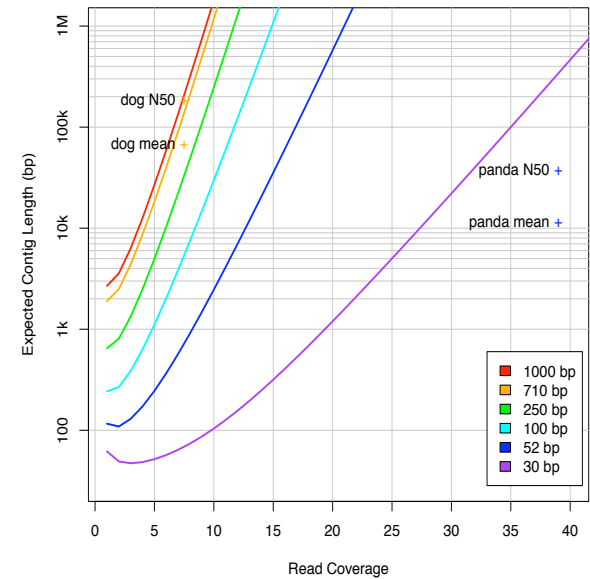


Errors

Coverage




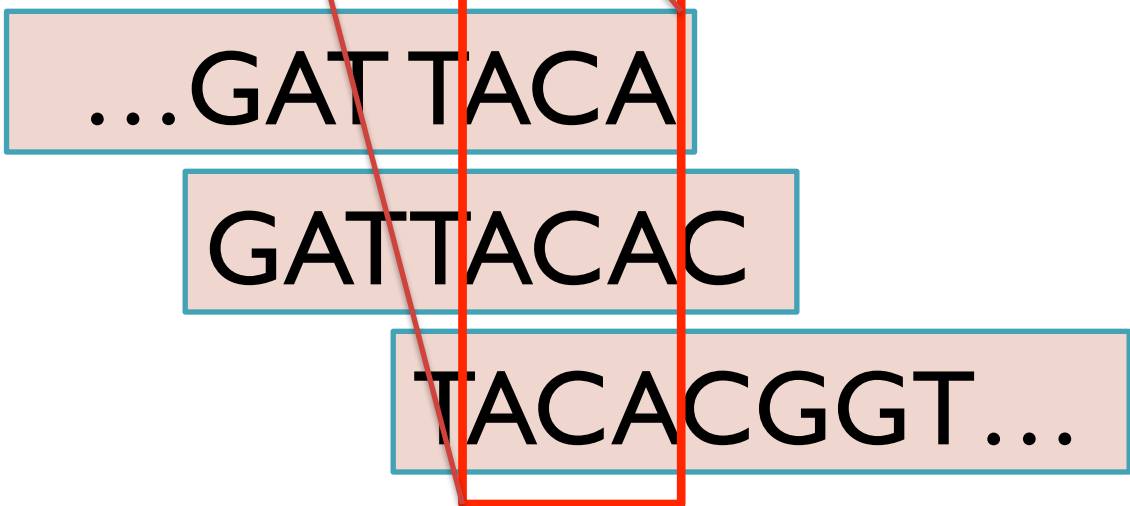
Lander Waterman Expected Contig Length vs Coverage



Estimating coverage with Kmers

Reference: 

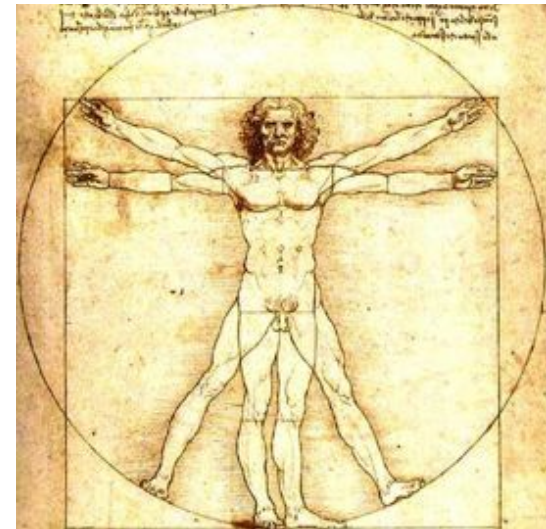
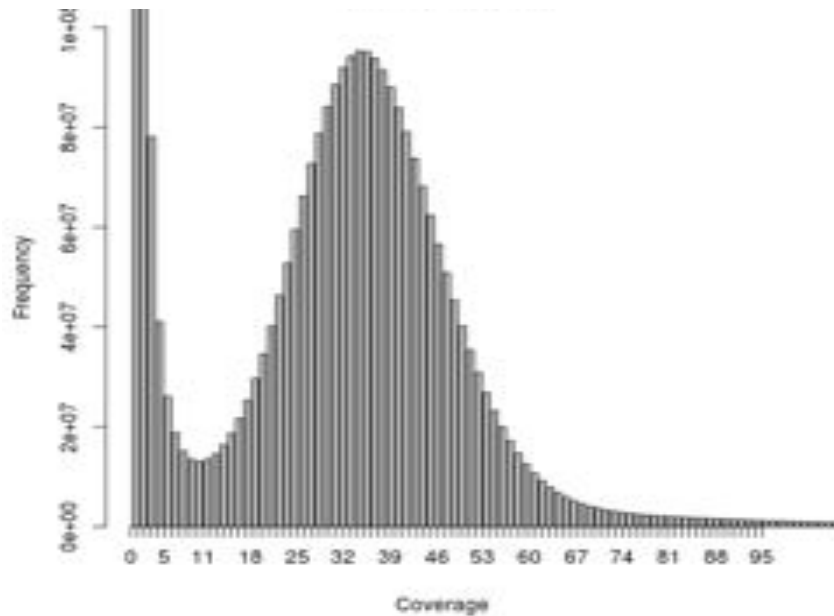
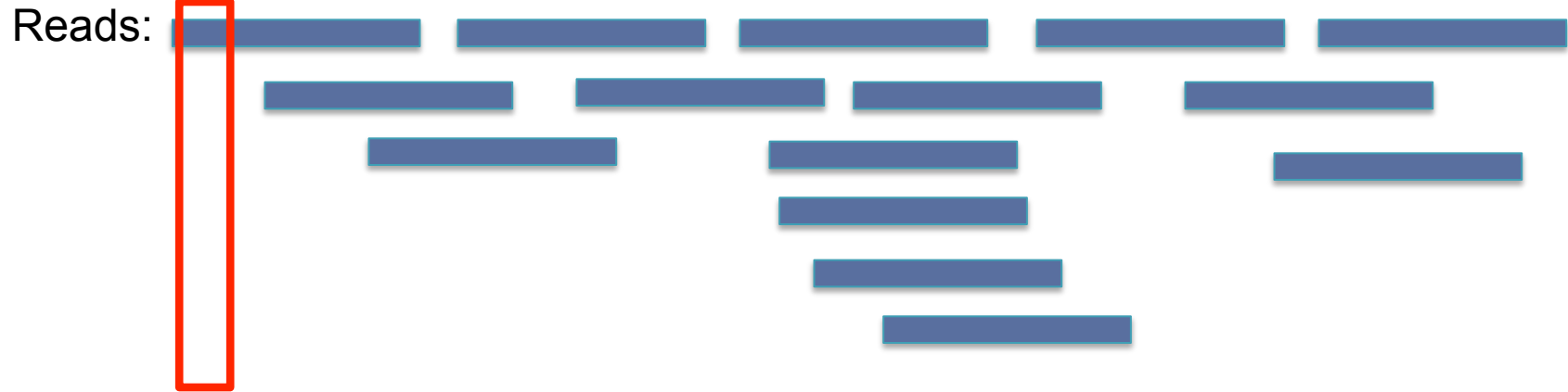
Reads: 



...GATTACA
GATTACAC
TACACGGT...

Estimating coverage with Kmers

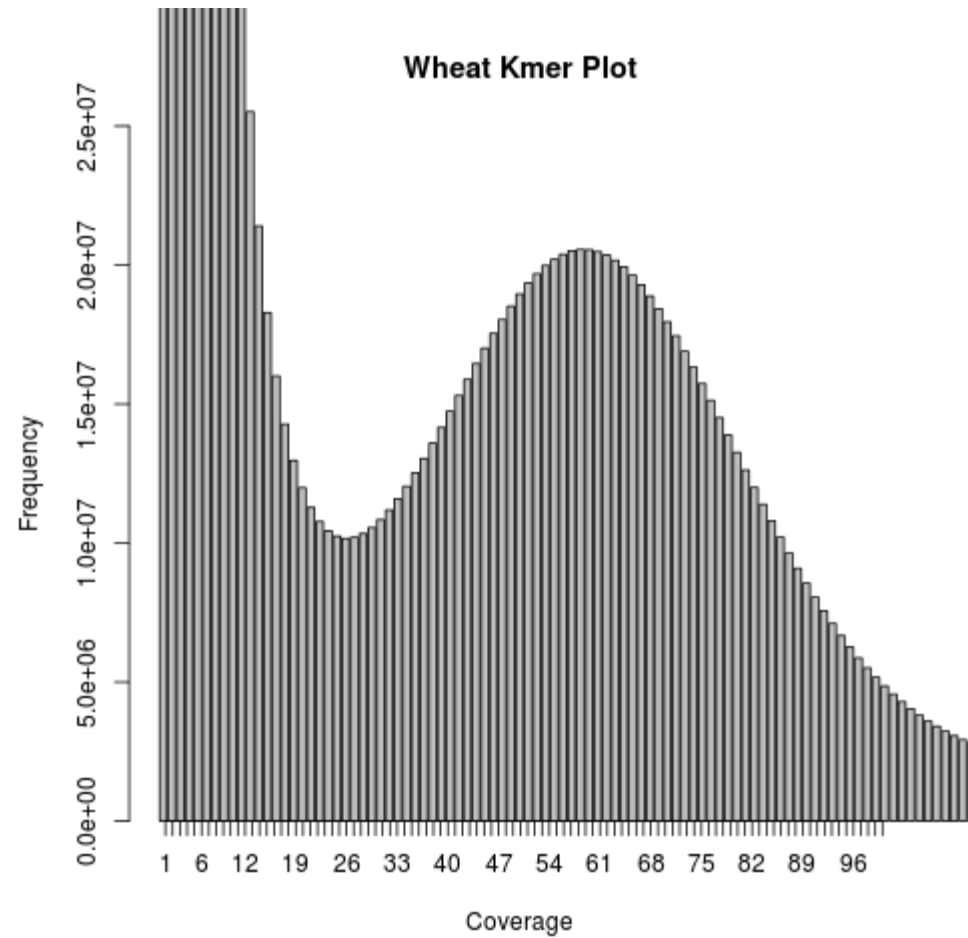
Reference: 



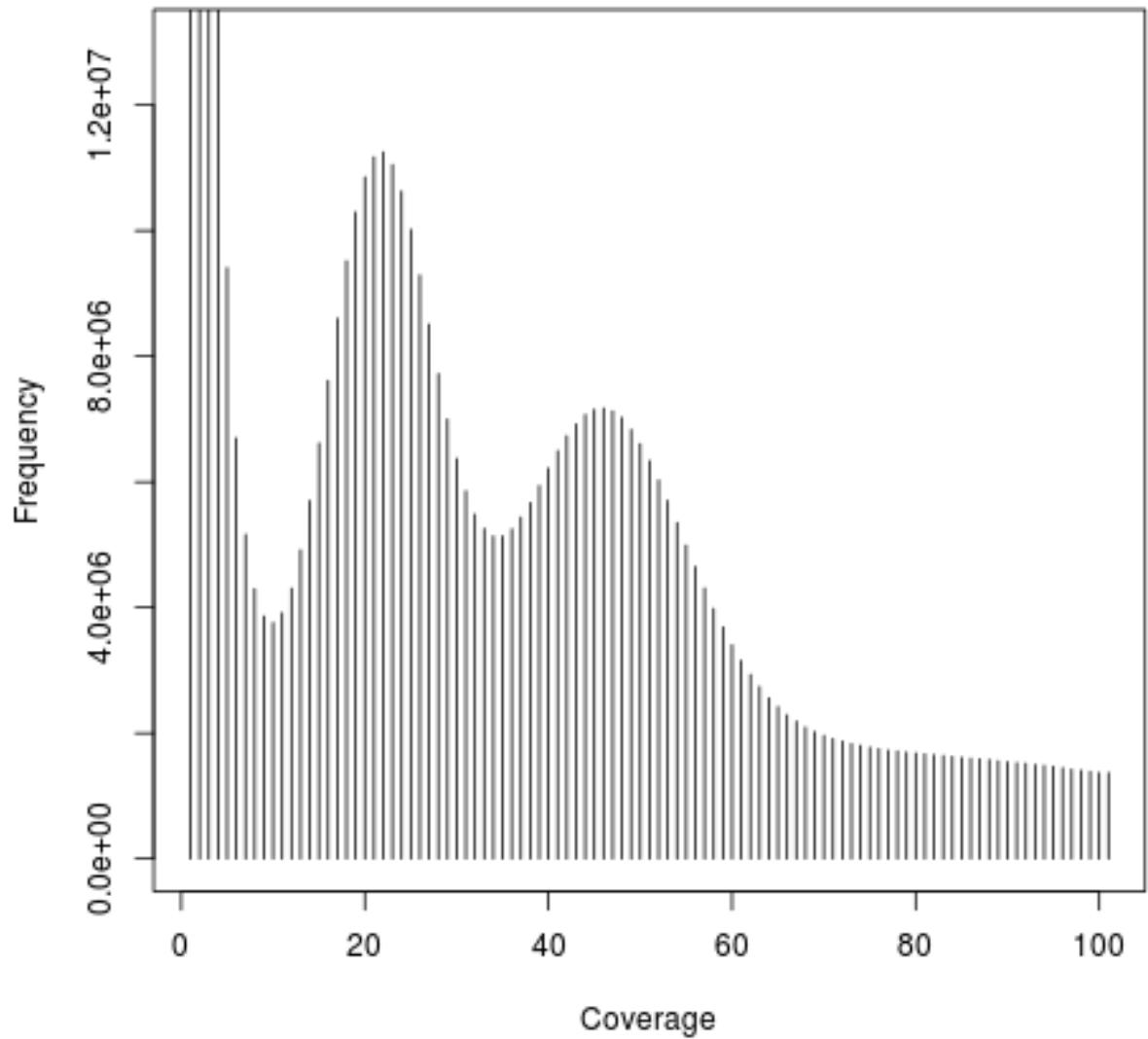
NA12878

Wheat Genome

(*A. tauschii* / CSHL)



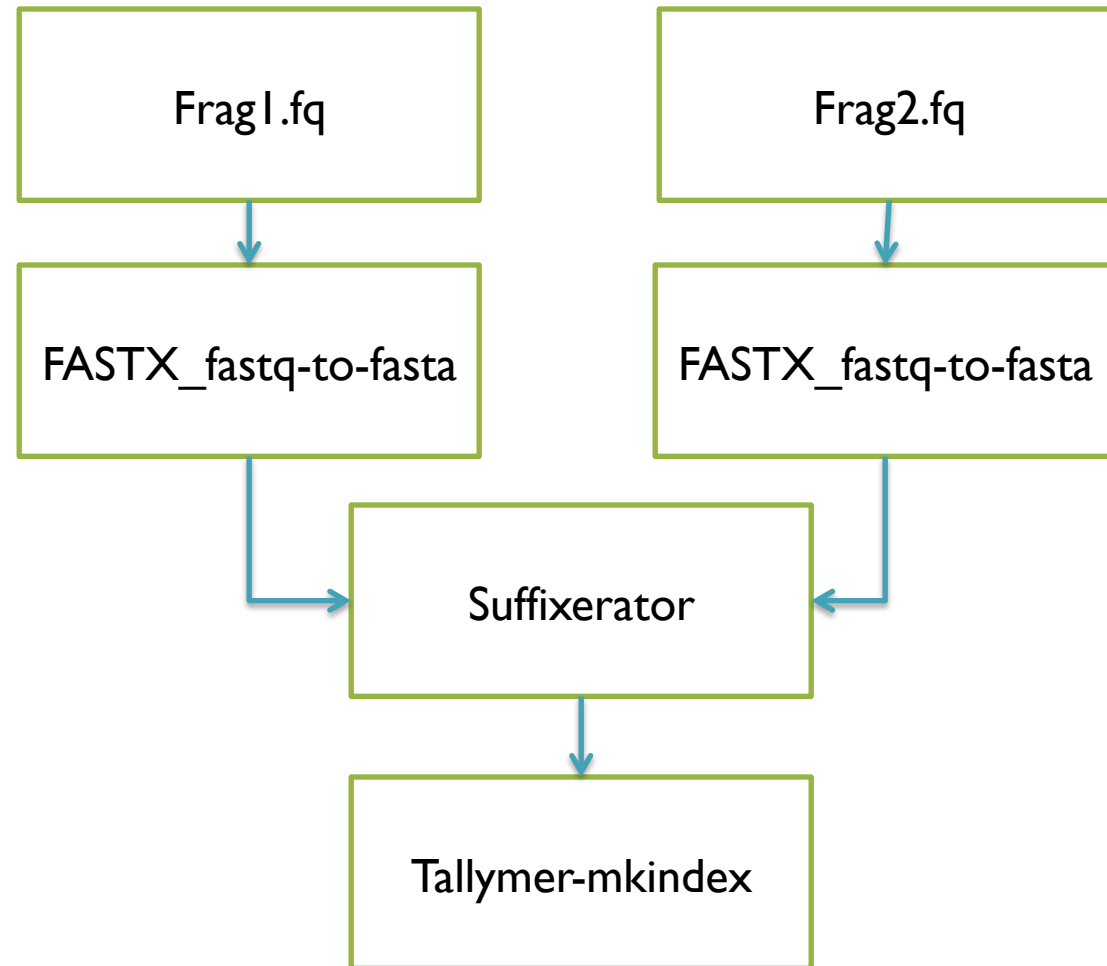
Heterozygous Genome



Contact: @mike_schatz



QC: Mer counts



A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes

Kurtz S, Narechania A, Stein JC, Ware D. (2008) BMC Genomics. 9:517

Running ALLPATHS-LG



The screenshot shows the Discovery Environment interface. The file explorer displays a directory structure under 'allpaths' with subfolders like 'ASSEMBLIES' and 'run'. The 'assembly.report' file is selected, and its contents are shown in a text viewer. The text includes 'AllPathsReport' and 'LibCoverage' sections. A yellow circle highlights the 'total scaffold length, with gaps' value of 283222. A yellow smiley face is overlaid on the bottom right of the screenshot.

Property	Value
costig minimum size for reporting	1000
number of contigs	37
number of contigs per Mb	12.8
number of scaffolds	11
total contig length	283222
total scaffold length, with gaps	283222
contig size in kb	149.7
scaffold size in kb	1474
scaffold size in kb, with gaps	1477
number of scaffolds per Mb	3.82
median size of gaps in scaffolds	16
median dev of gaps in scaffolds	0.31
% of bases in captured gaps	0.31
% of bases in negative gaps (after 5 devs)	0.00
% of ambiguous bases	1.68
ambiguities per 10,000 bases	0.25



Post-QC: CEGMA

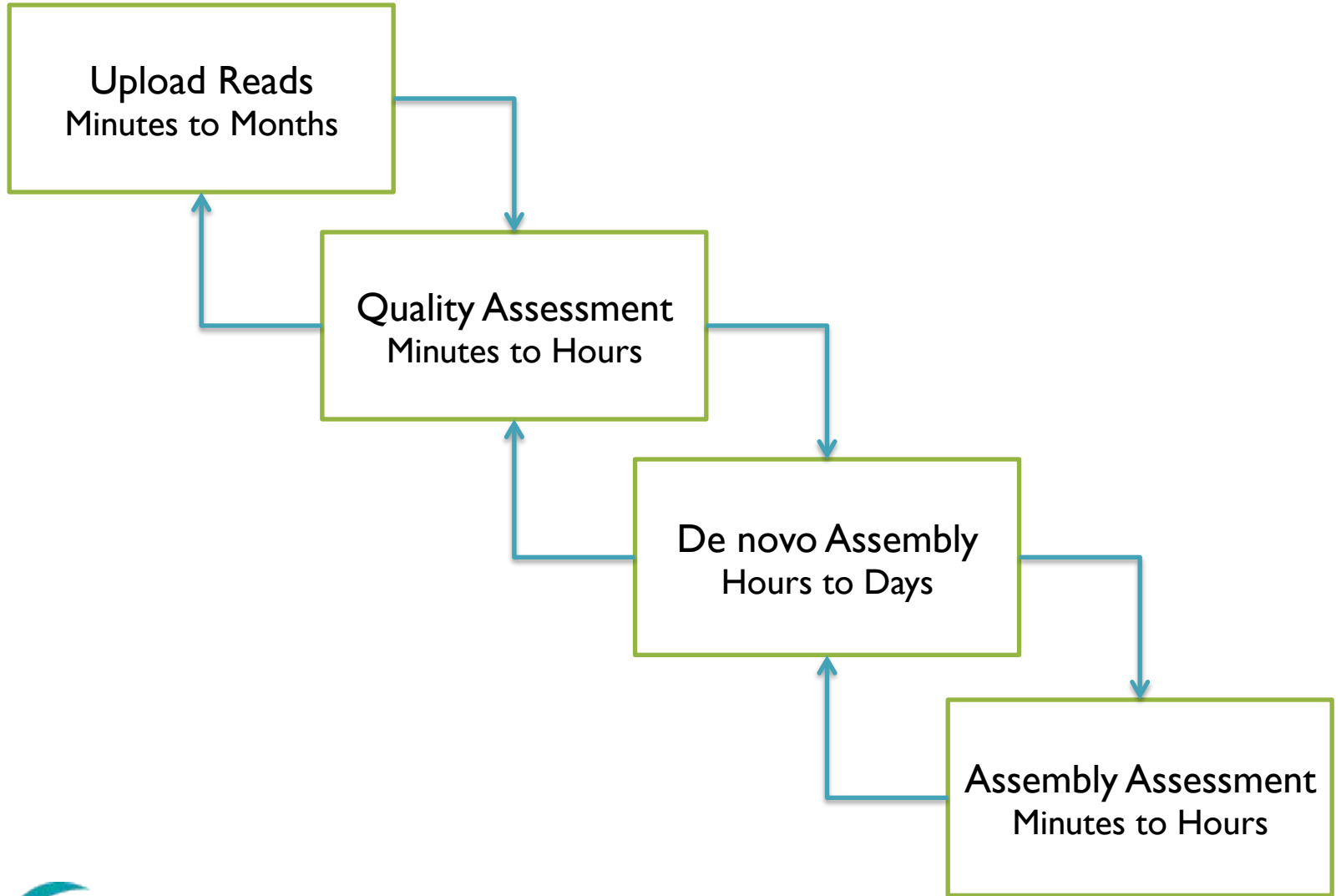
The screenshot displays the Discovery Environment interface. The 'Apps' window shows a search for 'cegma' with one result, 'CEGMA', integrated by Michael Crusoe. The 'CEGMA' window shows an analysis named 'CEGMA_analysis1'. The 'staph_cegma.completeness_report' window displays the following table:

	#Prots	%Completeness	-	#Total	Average	%Ortho
Complete	29	11.69	-	33	1.14	13.79
Group 1	5	7.58	-	5	1.00	0.00
Group 2	8	14.29	-	9	1.12	12.50
Group 3	7	11.48	-	8	1.14	14.29
Group 4	9	13.85	-	11	1.22	22.22
Partial	31	12.50	-	36	1.16	16.13
Group 1	5	7.58	-	6	1.20	20.00
Group 2	9	16.07	-	10	1.11	11.11
Group 3	7	11.48	-	8	1.14	14.29
Group 4	10	15.38	-	12	1.20	20.00

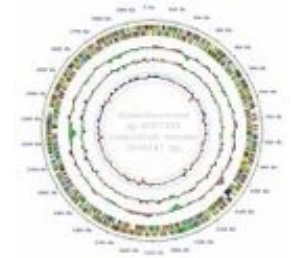
These results are based on the set of genes selected by Genis Parra ##
Key: ##
Prots = number of 248 ultra-conserved CEGs present in genome ##
%Completeness = percentage of 248 ultra-conserved CEGs present ##
Total = total number of CEGs present including putative orthologs ##
Average = average number of orthologs per CEG ##
%Ortho = percentage of detected CEGs that have more than 1 ortholog ##
Missing proteins ##
Complete

CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes
Parra G, Bradnam K, Korf I. (2007) *Bioinformatics*. 23 (9): 1061-1067.

Assembly Workflow



Resources



- iPlant
 - <http://www.iplantcollaborative.org/>
- Assembly Competitions
 - Assemblathon: <http://assemblathon.org/>
 - GAGE: <http://gage.cbc.umd.edu/>
- Assembler Websites:
 - ALLPATHS-LG: <http://www.broadinstitute.org/software/allpaths-lg/blog/>
 - SOAPdenovo: <http://soap.genomics.org.cn/soapdenovo.html>
 - Celera Assembler: <http://wgs-assembler.sf.net>
- Tools:
 - FastQC: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
 - Tallymer: <http://www.zbh.uni-hamburg.de/?id=211>
 - CEGMA: <http://korflab.ucdavis.edu/datasets/cegma/>

Acknowledgements

Special Thanks

Shoshana Marcus
James Gurtowski

Roger Barthelson
Stephen Goff
Nicole Hopkins
Dan Stanzione
Joshua Stein
Matthew Vaughn
Doreen Ware
Jason Williams

Executive Team

Name	Title	Institution
Stephen Goff	PI and Project Director, iPlant Collaborative	University of Arizona
Dan Stanzione	Co-PI, iPlant Collaborative	University of Texas, Austin

Staff

Name	Title	Institution
Greg Wilson	Research Scientist	University of Texas, Austin
Ellie Arora	Research Associate	University of Texas, Austin
Roger Barthelson	Bioinformatics Analyst	University of Arizona
Rob Smith	QA Test Engineer	University of Arizona



iPlant Collaborative™ *Empowering A New Plant Biology*

Questions?

<http://schatzlab.cshl.edu/>
@mike_schatz

