# Whole Genome Assembly with iPlant

Michael Schatz & Shoshana Marcus

Dec 4, 2013

CSHL Plant Genomes and Biotechnology

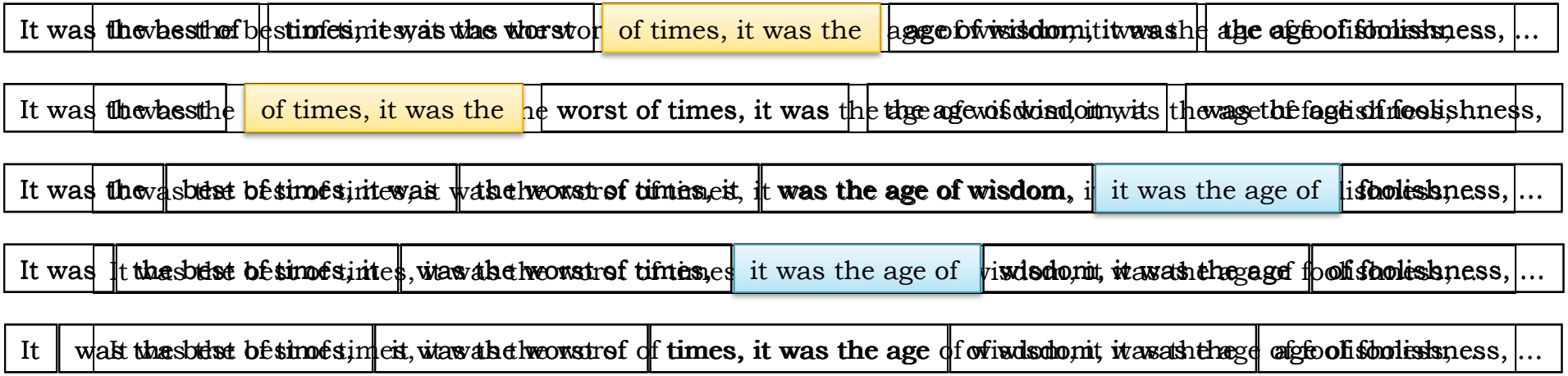
iPlant Collaborative™ *Empowering A New Plant Biology*

# 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

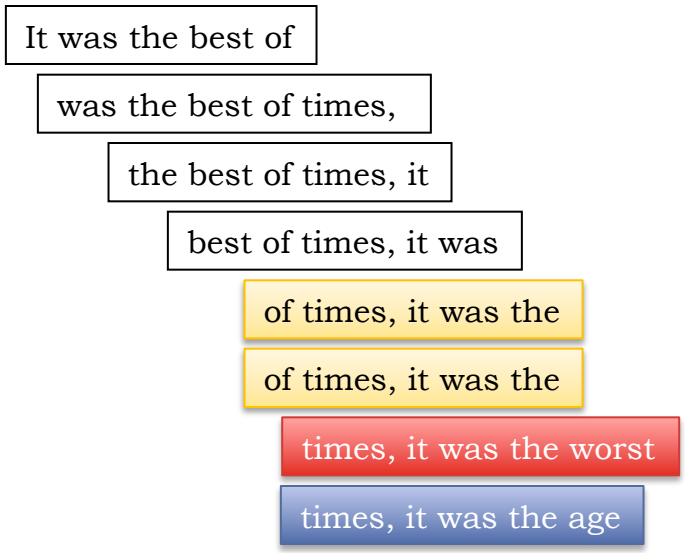
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools




- 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



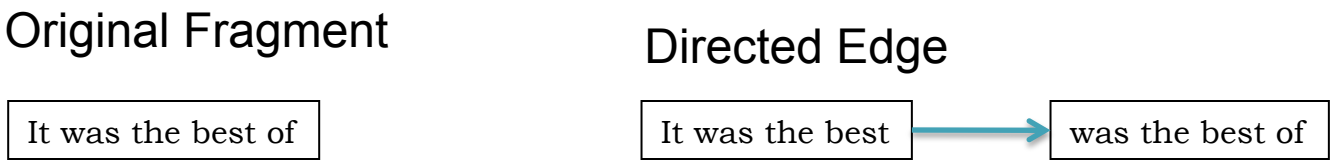

The repeated sequence make the correct reconstruction ambiguous

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

Model the assembly problem as a graph problem

# de Bruijn Graph Construction

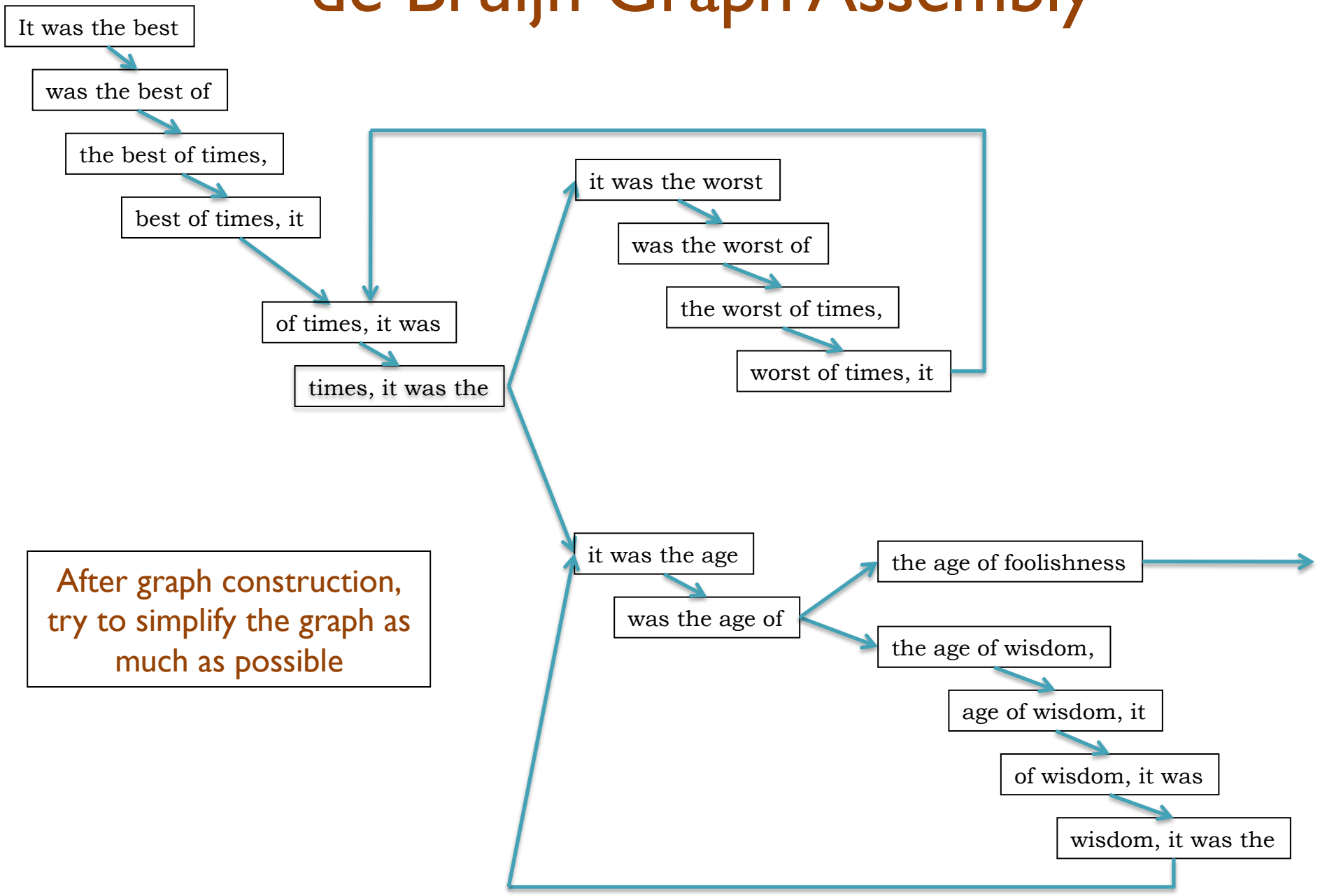
- $D_k = (V,E)$
  - V = All length-k subfragments (k < l)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words




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

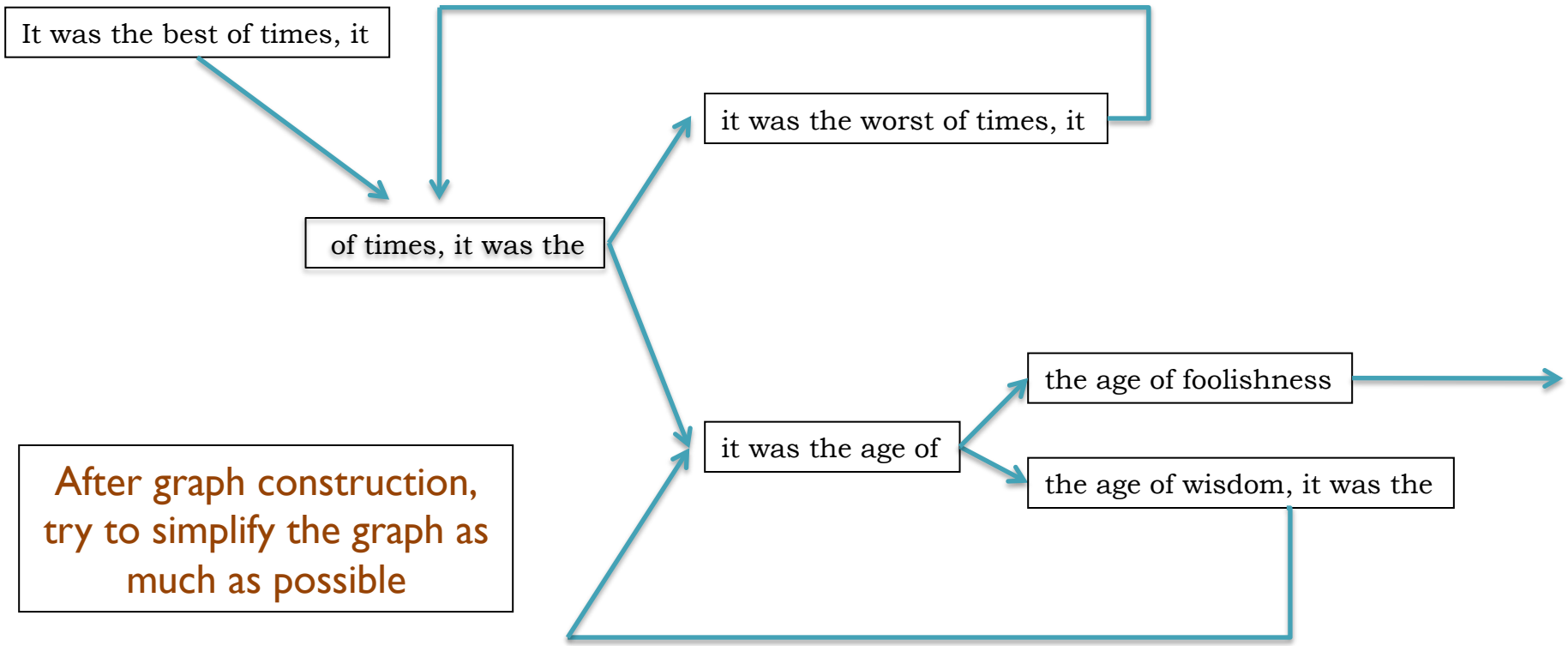
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

# de Bruijn Graph Assembly




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

# de Bruijn Graph Assembly




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

# The full tale

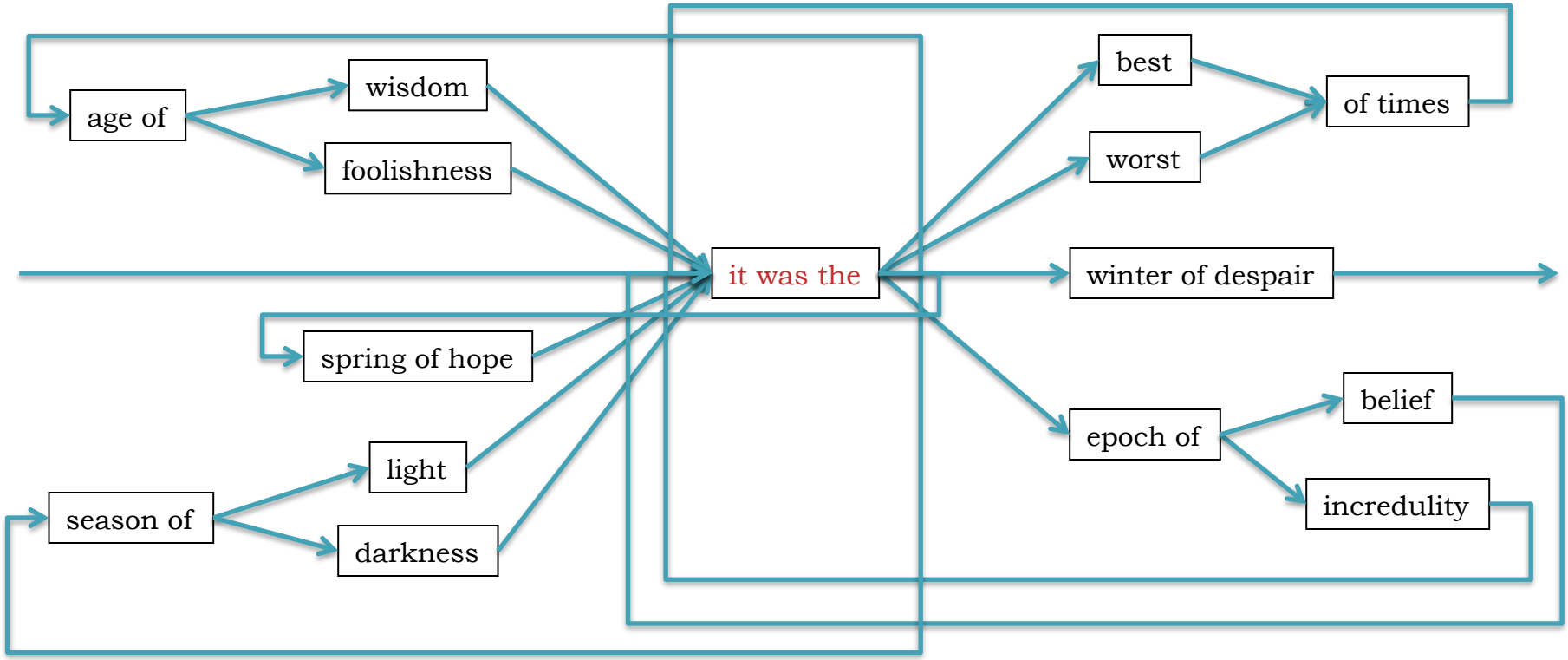
… it was the best of times it was the worst of times …

… 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 …

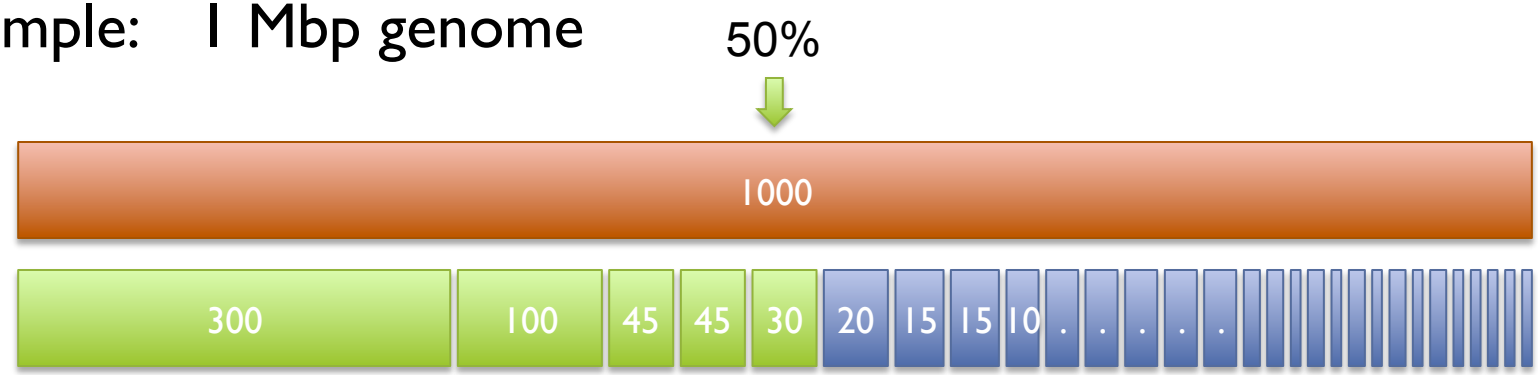
… it was the spring of hope it was the winder of despair …



# 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 >= 500kbp)

Note:

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

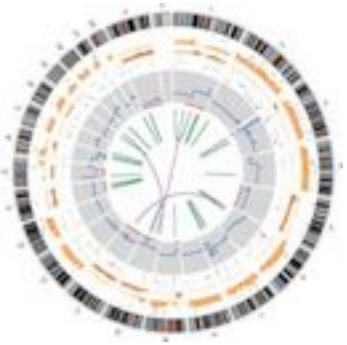
# Assembly Applications

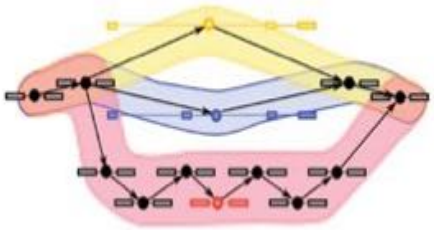
- Novel genomes



- Metagenomes

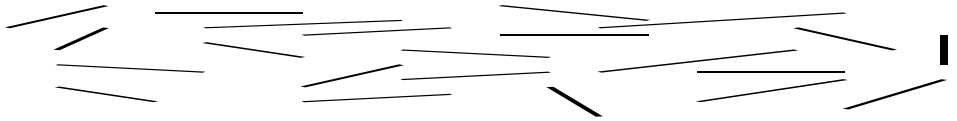

- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …





# Assembling a Genome

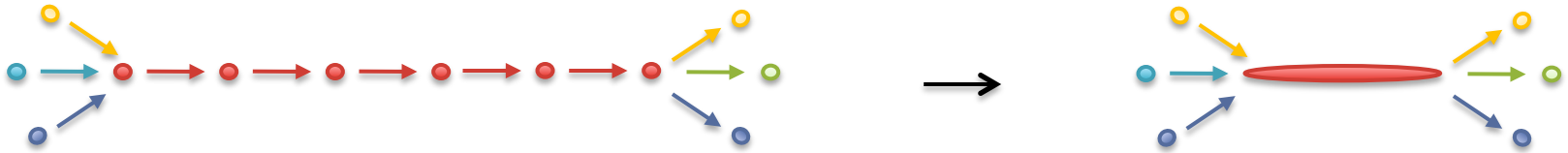
1. Shear & Sequence DNA



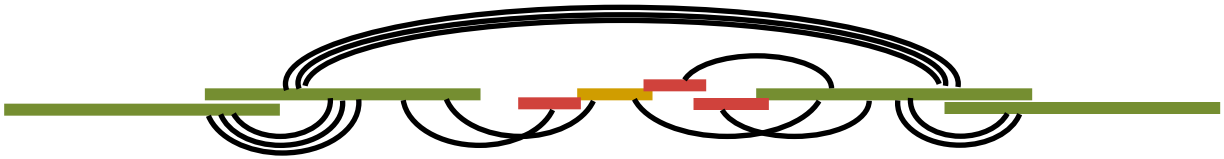
2. Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph

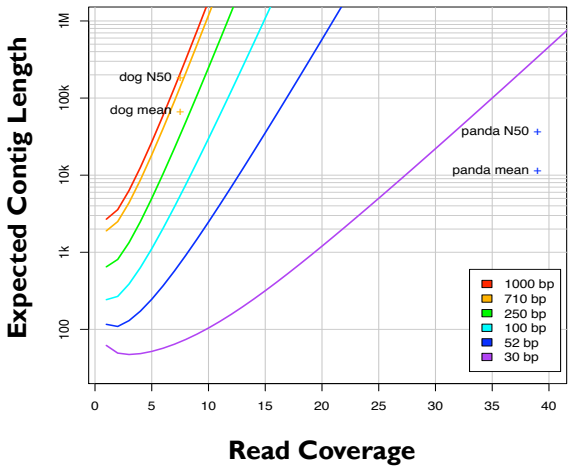


4. Detangle graph with long reads, mates, and other links



# Ingredients for a good assembly

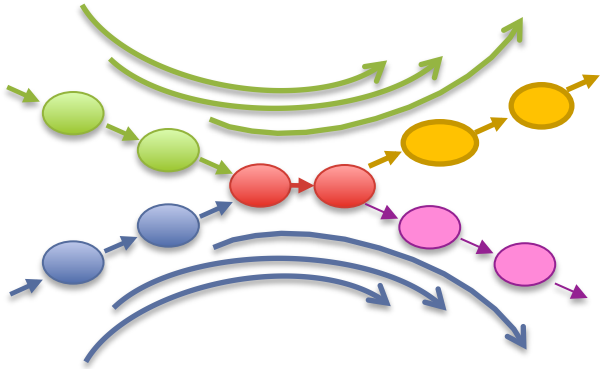
## Coverage




***High coverage is required***

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

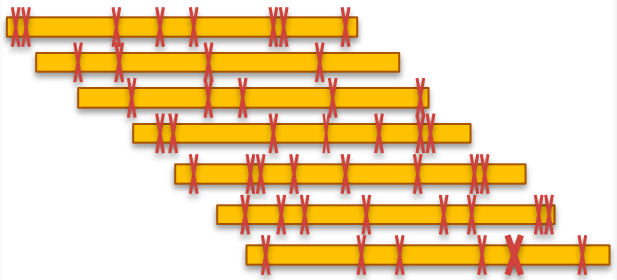
## Read Length



***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



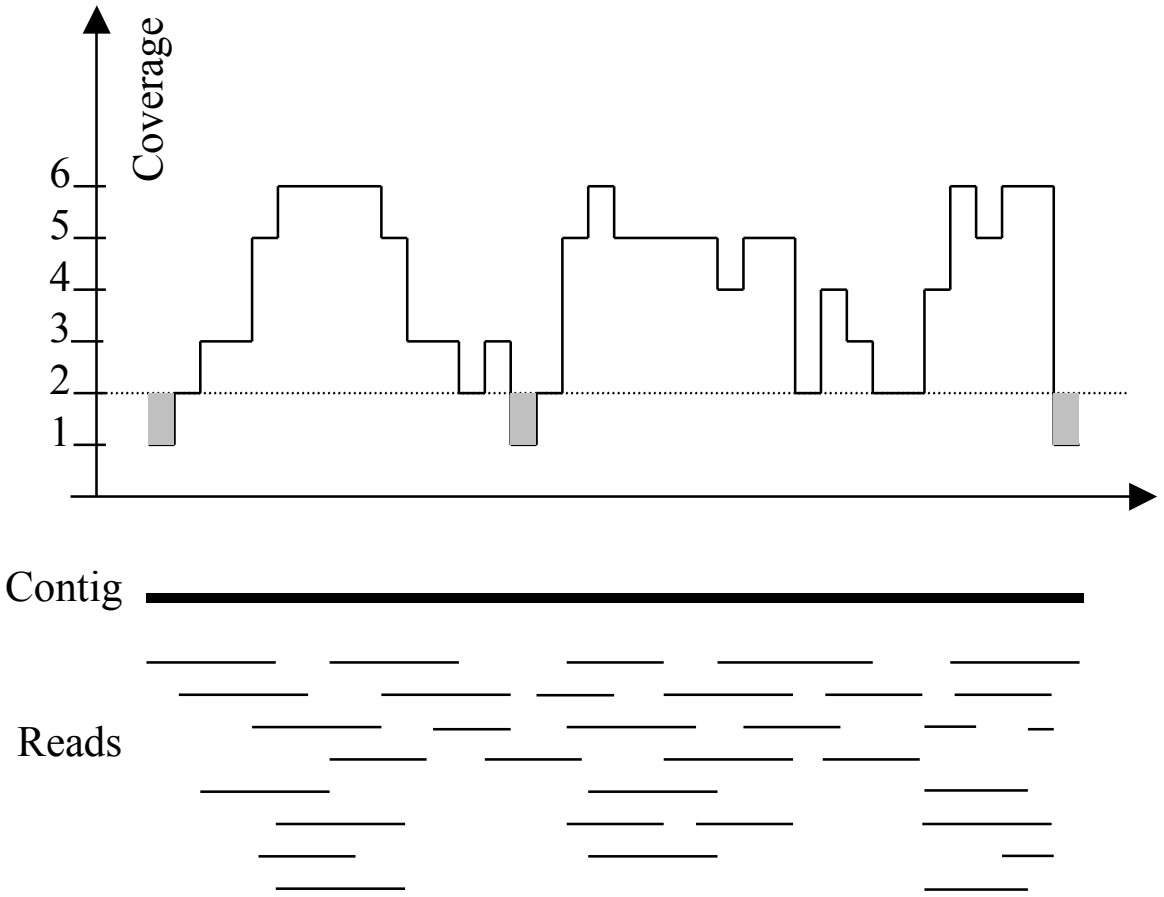
***Errors obscure overlaps***

- 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, WR (2012) *Genome Biology*. 12:243

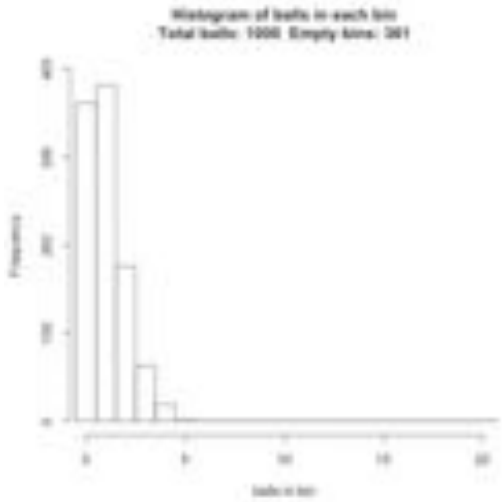
Coverage

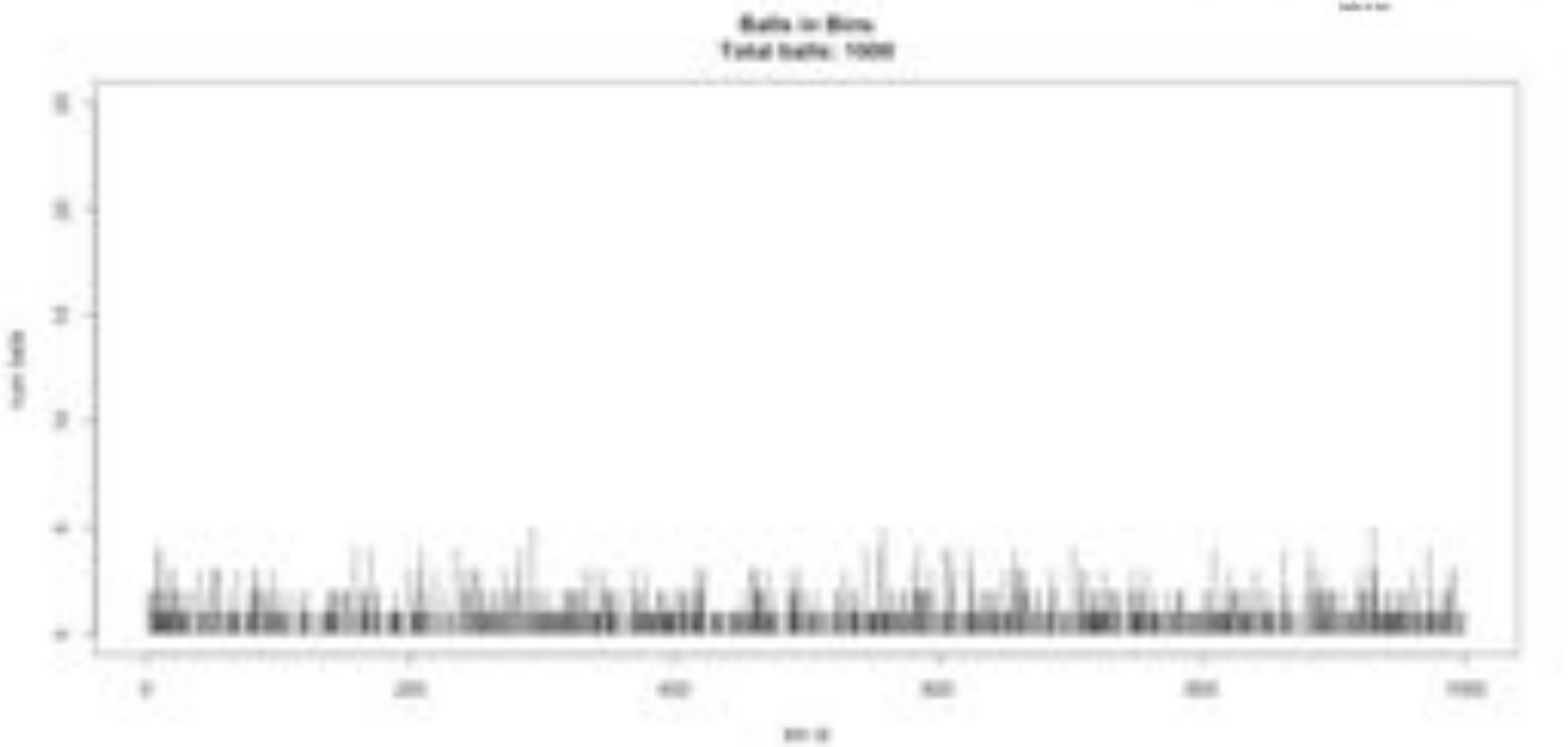
# Typical contig coverage



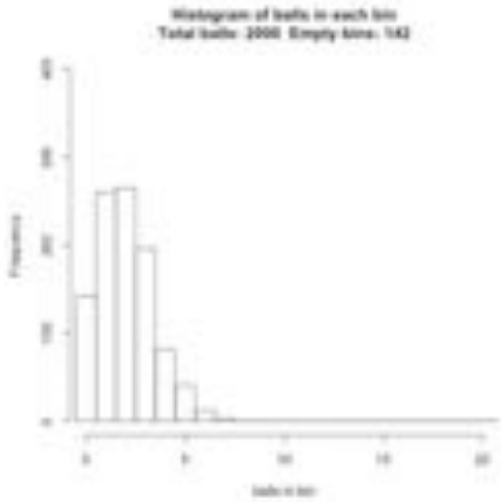

Imagine raindrops on a sidewalk

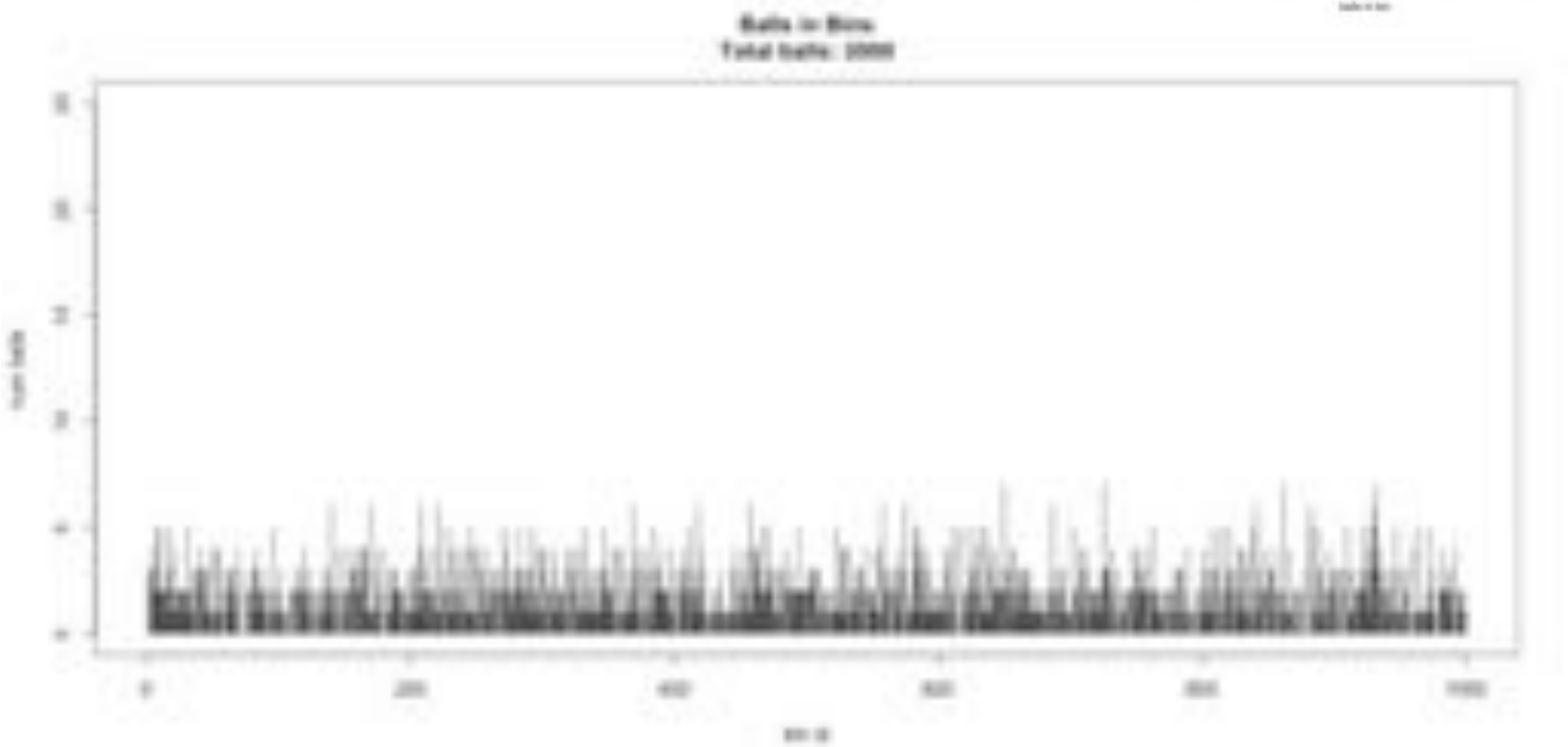
# Balls in Bins 1x



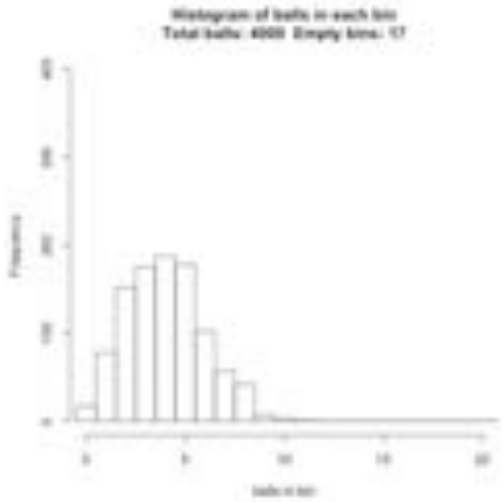


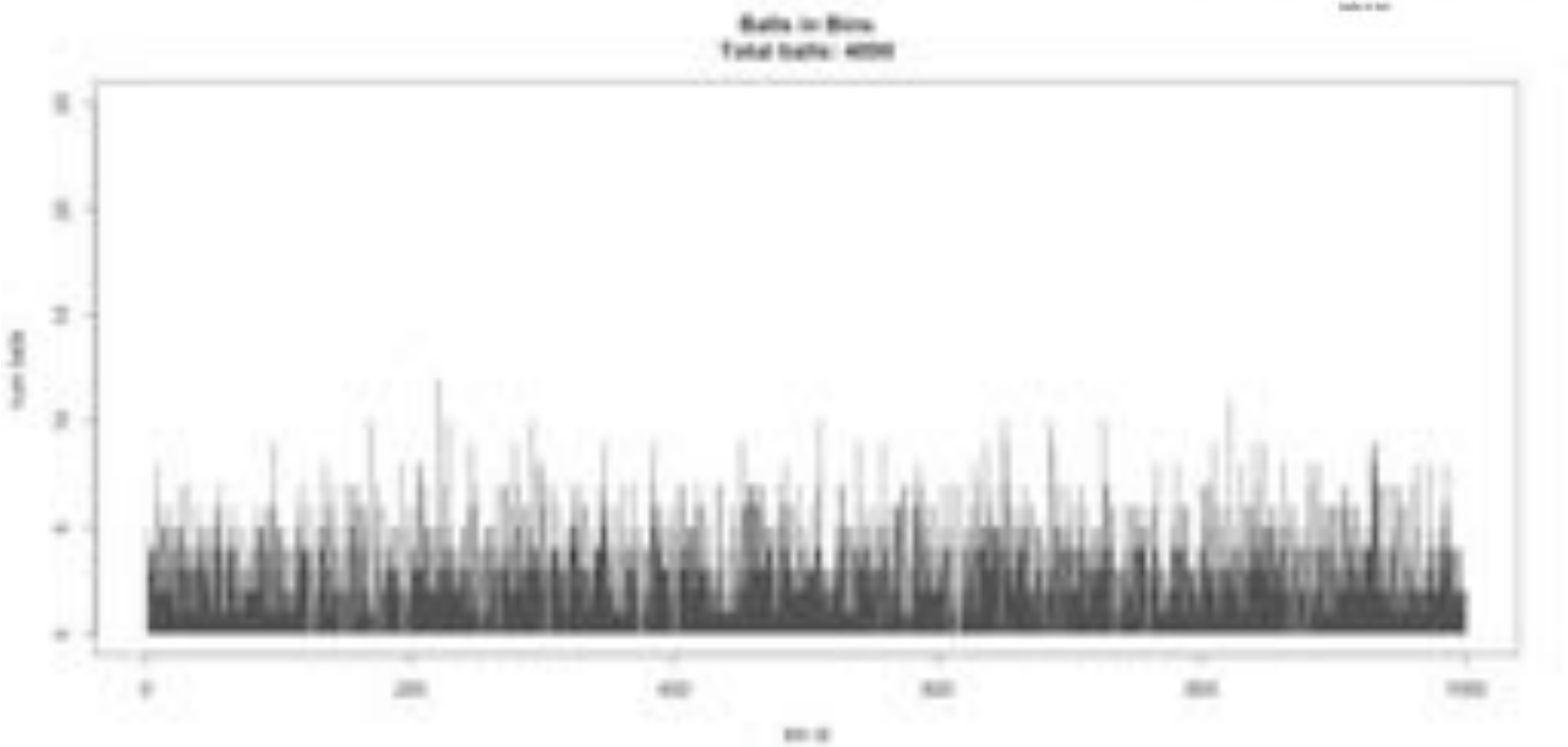
# Balls in Bins 2x



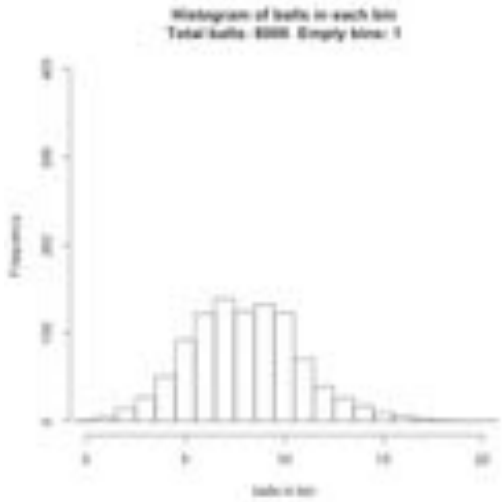


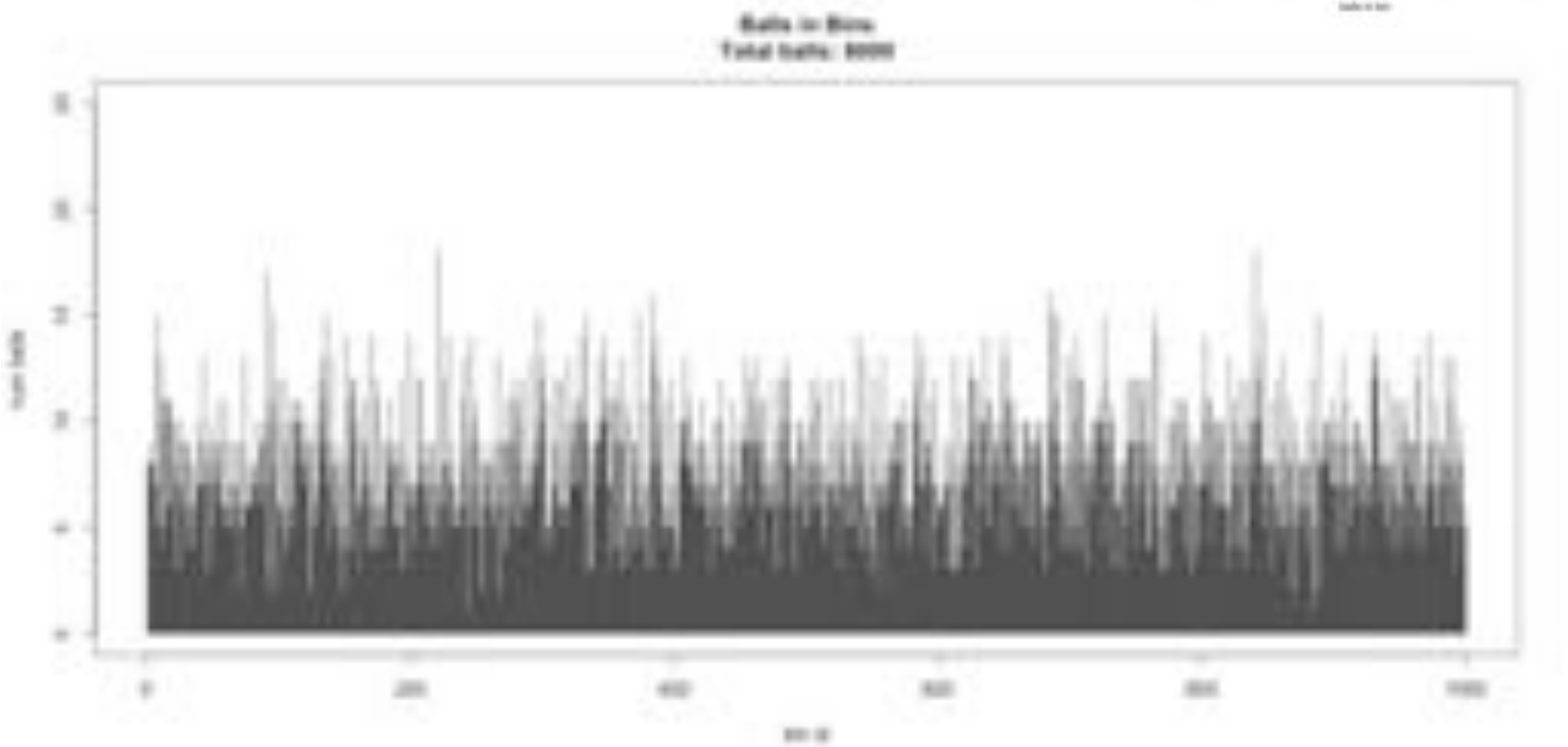
# Balls in Bins 4x





# Balls in Bins 8x

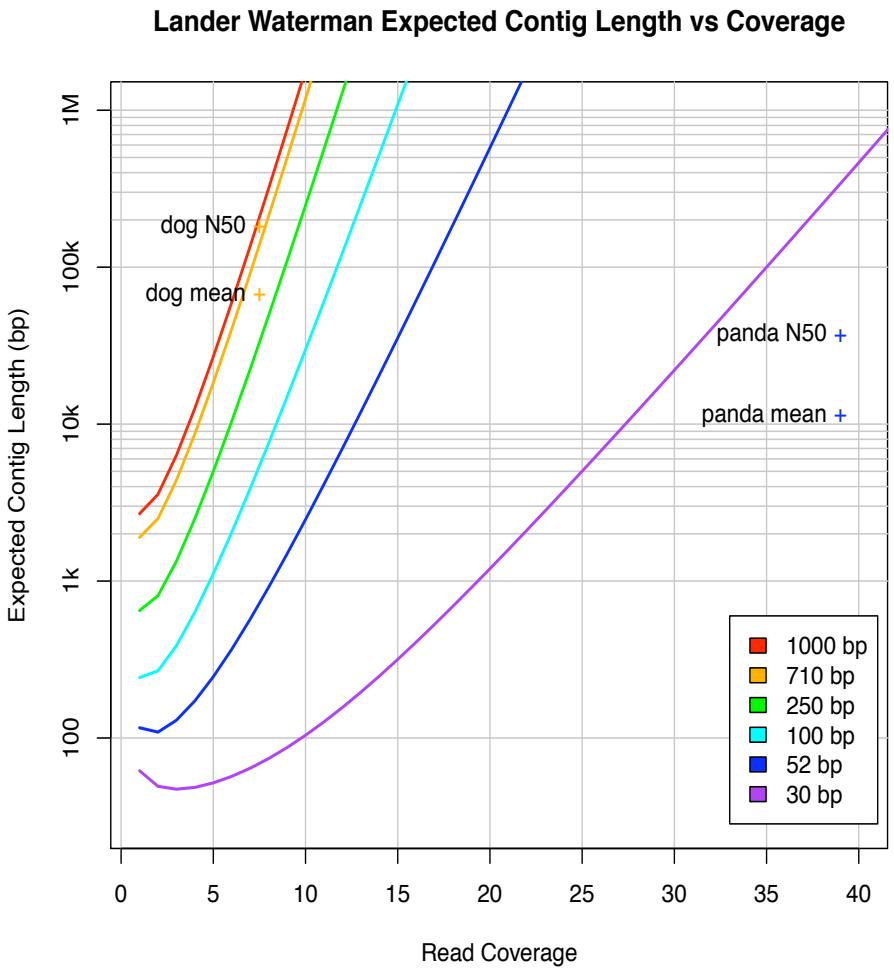




# Coverage and Read Length

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

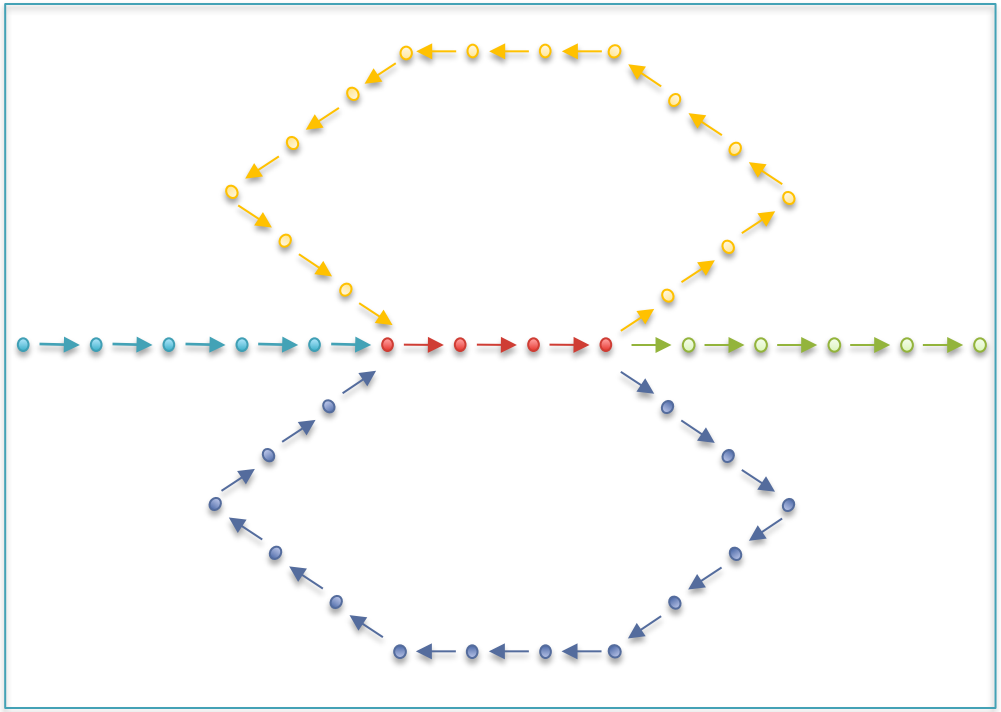


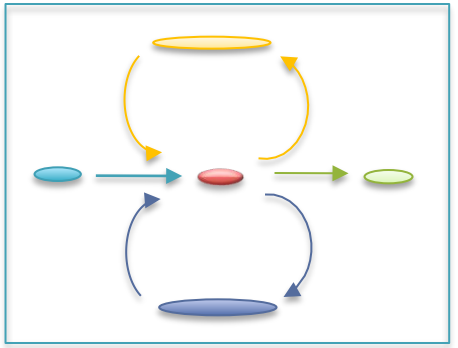

**Assembly of Large Genomes using Second Generation Sequencing**
Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research.* 20:1165-1173.

Errors

# 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 |
| --- | --- |
| was the worst of times,<br>was the worst of t**y**mes,<br>the worst of times, it | was the worst of times,<br>was the worst of t**y**mes,<br>times, it was the age<br>t**y**mes, it was the age |
| the worst of t**y**mes,<br>was the worst of<br>the worst of times,<br>worst of times, it | t**y**mes,<br>was the worst of<br>it was the age<br>times, |

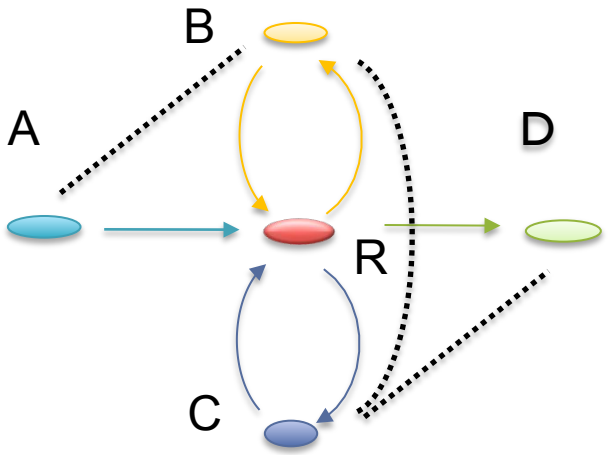
# Repetitive regions

| Repeat Type | Definition / Example | Prevalence |
|---|---|---|
| Low-complexity DNA / Microsatellites | $(b_1 b_2 \ldots b_k)^N$ where $1 \leq k \leq 6$<br>CACACACACACACACACACA | 2% |
| SINEs (Short Interspersed Nuclear Elements) | *Alu* sequence (~280 bp)<br>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% |

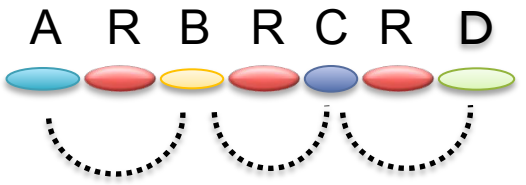
- Over 50% of mammalian genomes are repetitive
  - Large plant genomes tend to be even worse
  - Wheat: 16 Gbp; Pine: 24 Gbp

# 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

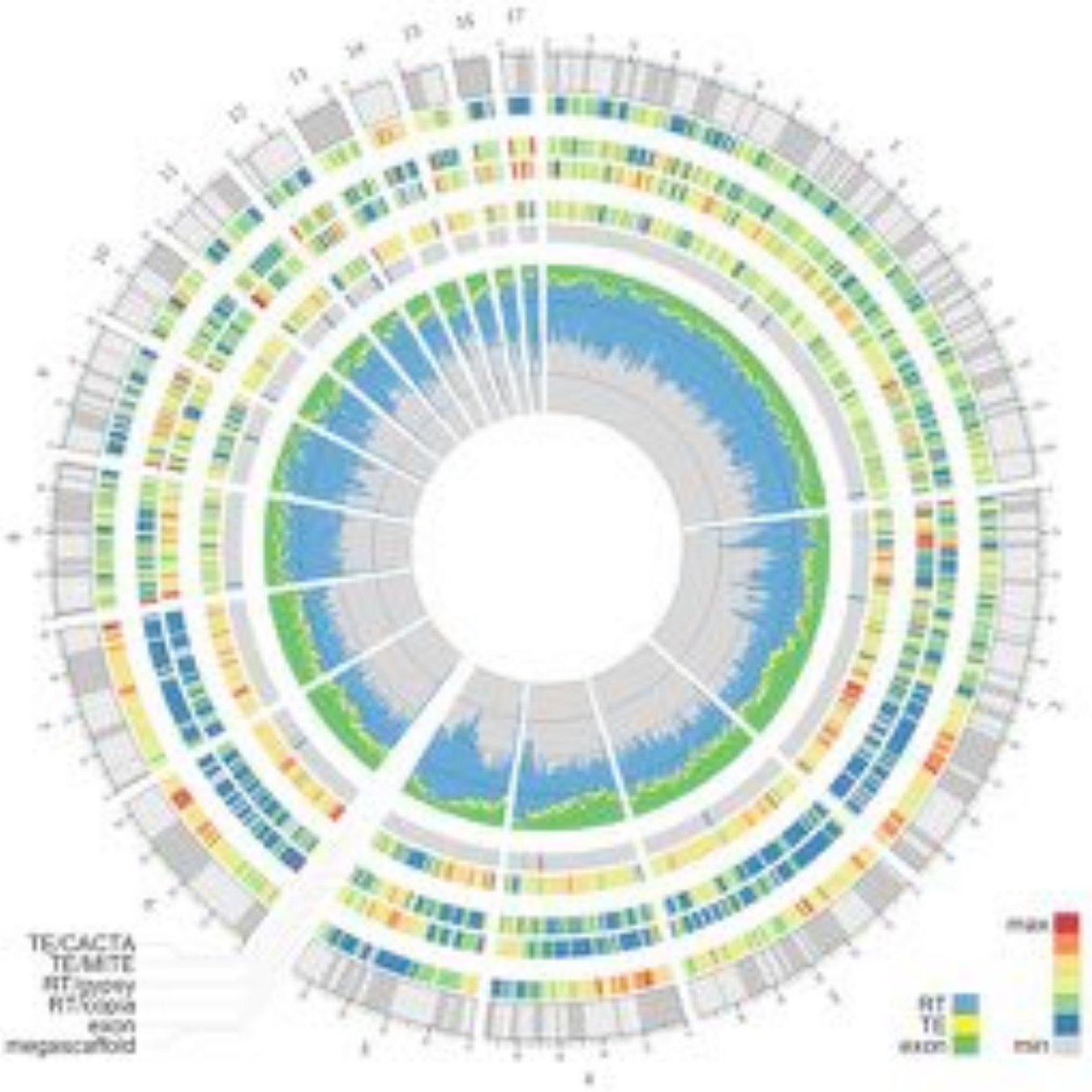



- Final scaffold may have internal gaps called sequencing gaps
  - We know the order, orientation, and spacing, but just not the bases. Fill with Ns instead

# 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

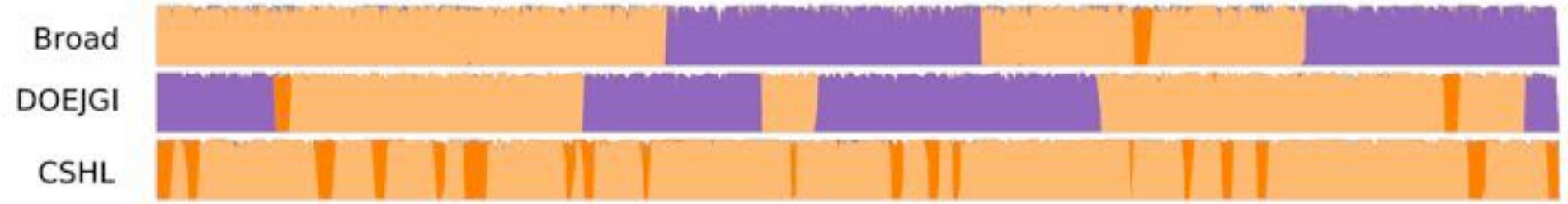
- Results demonstrate trade-offs assemblers must make

**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

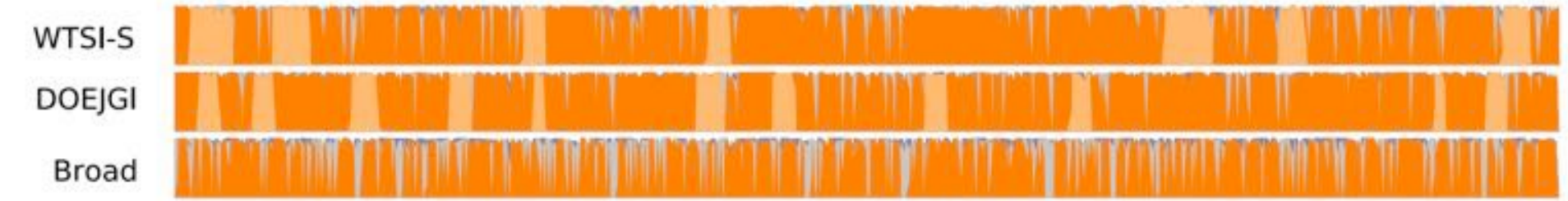
**Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species**
Bradnam, KR. et al (2013) GigaScience 2:10 doi:10.1186/2047-217X-2-10

# Assembly Results

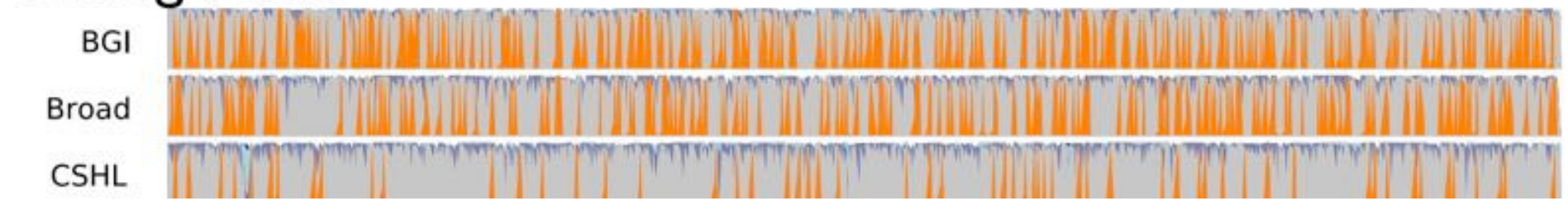
## Scaffolds




## Scaffold Paths




## Contig Paths

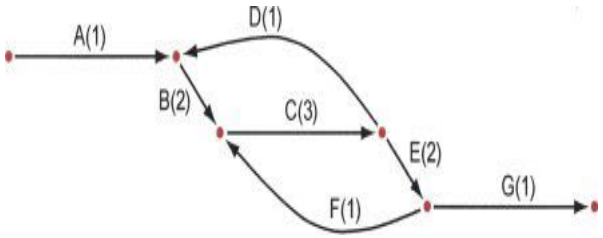



# Final Rankings

| 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

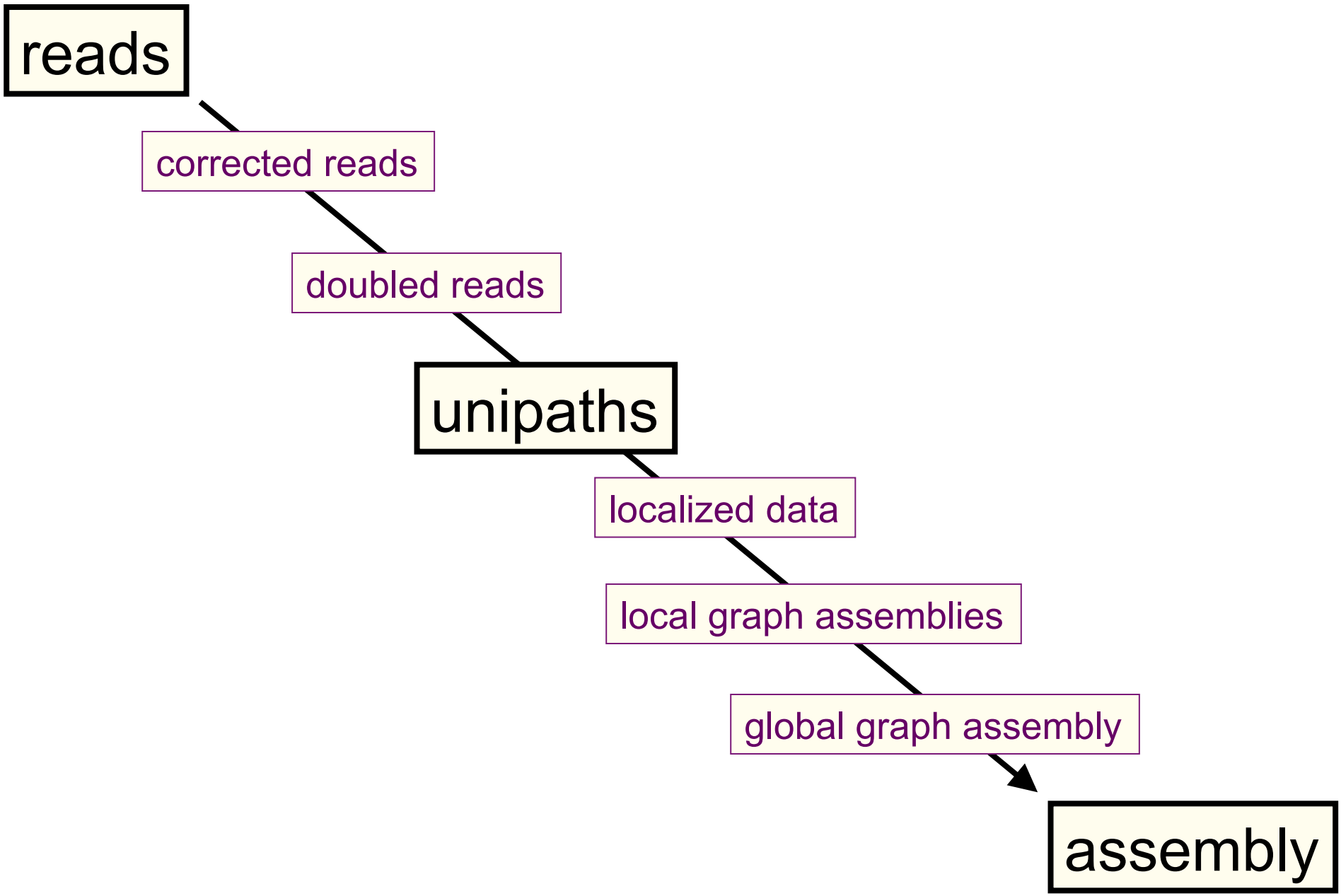



# Genome assembly with ALLPATHS-LG

Iain MacCallum

BROAD INSTITUTE

## 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.

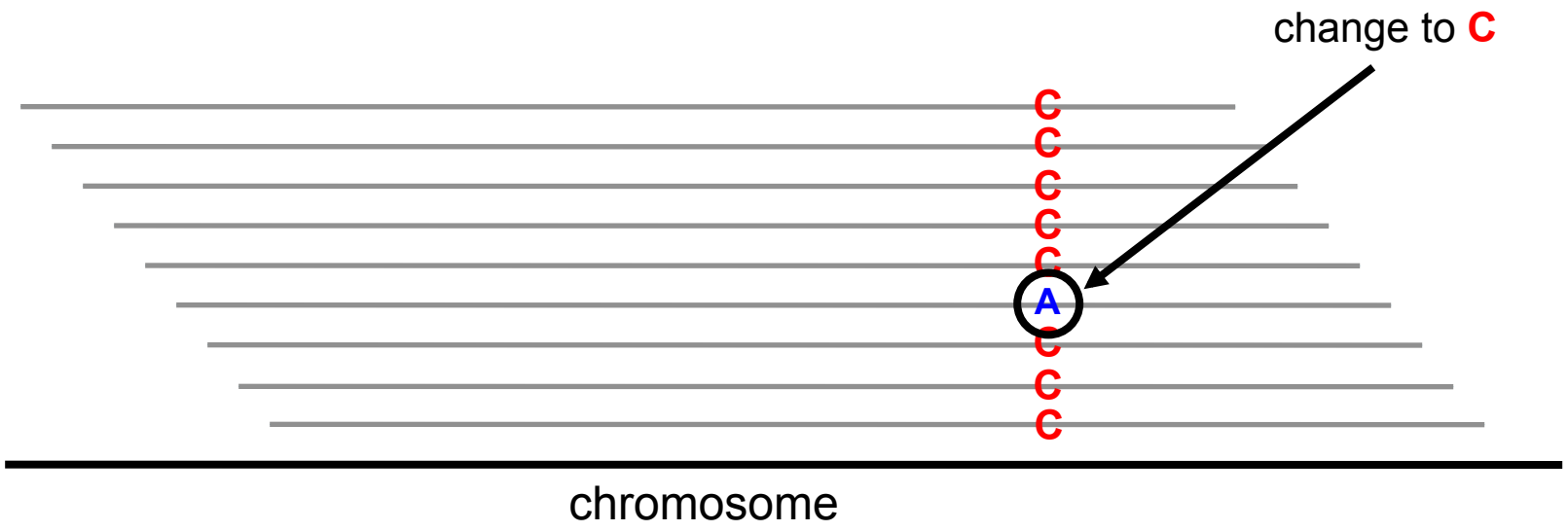
**For best results. Normally not used for small genomes.
However essential to assemble long repeats or duplications.

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

All: protocols are either available, or in progress.

## Error correction

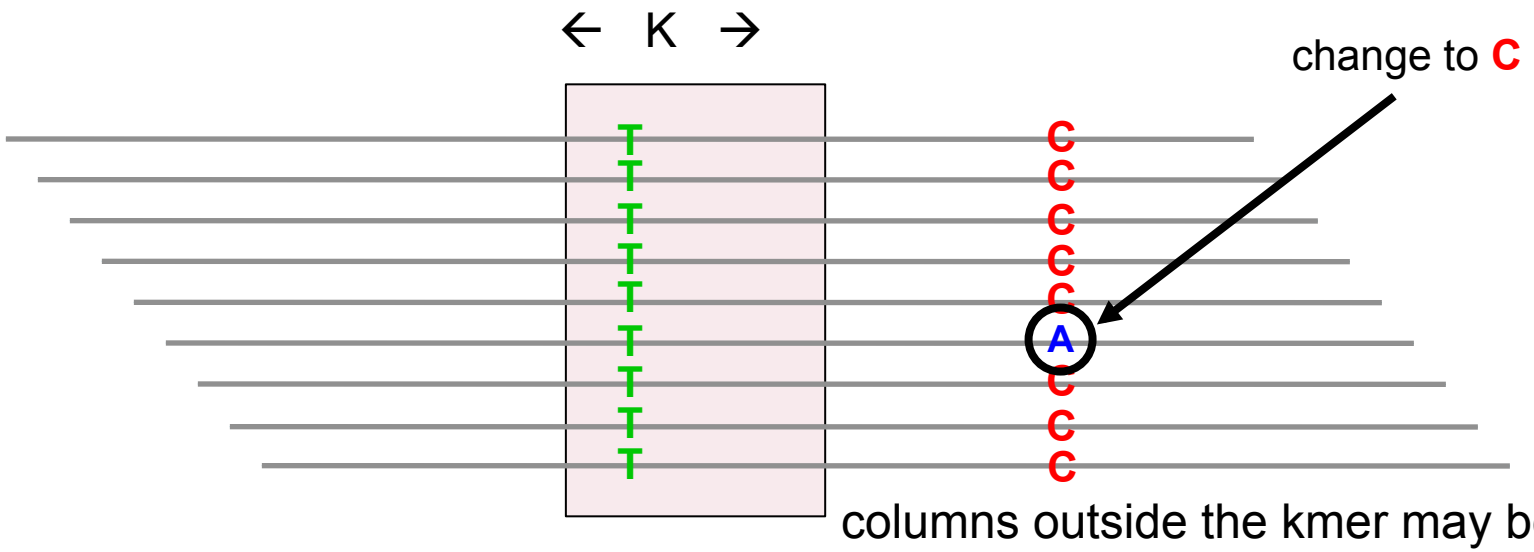
Given a crystal ball, we could stack reads on the chromosomes they came from (with homologous chromosomes separate), then let each column ‘vote’:




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)



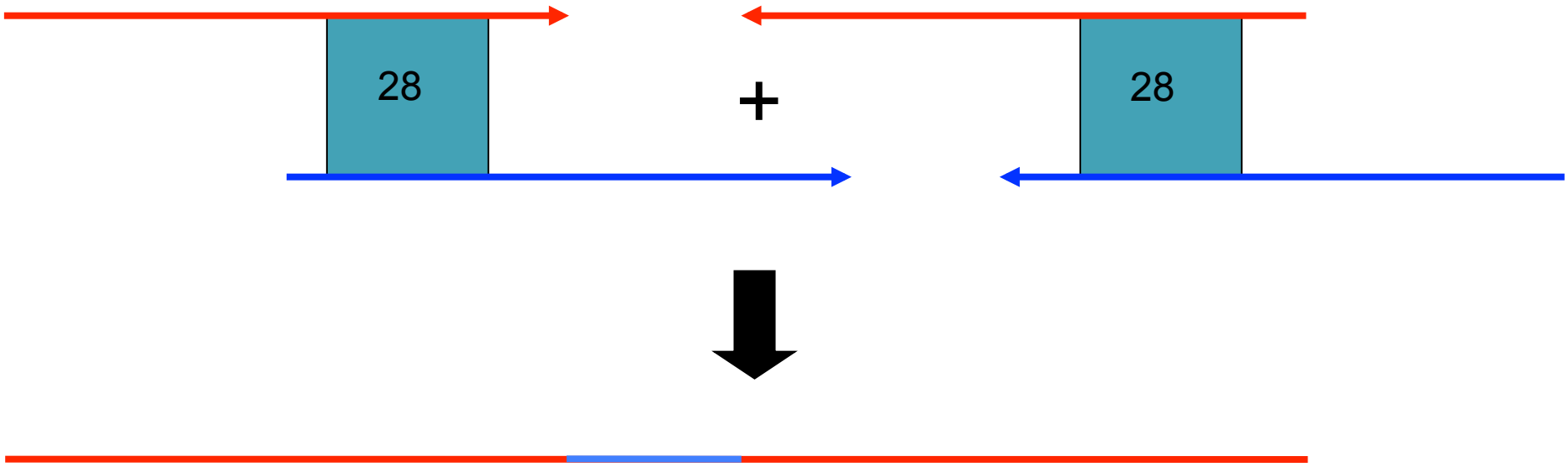

columns outside the kmer may be mixed

columns inside the kmer are homogeneous

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

## Read doubling

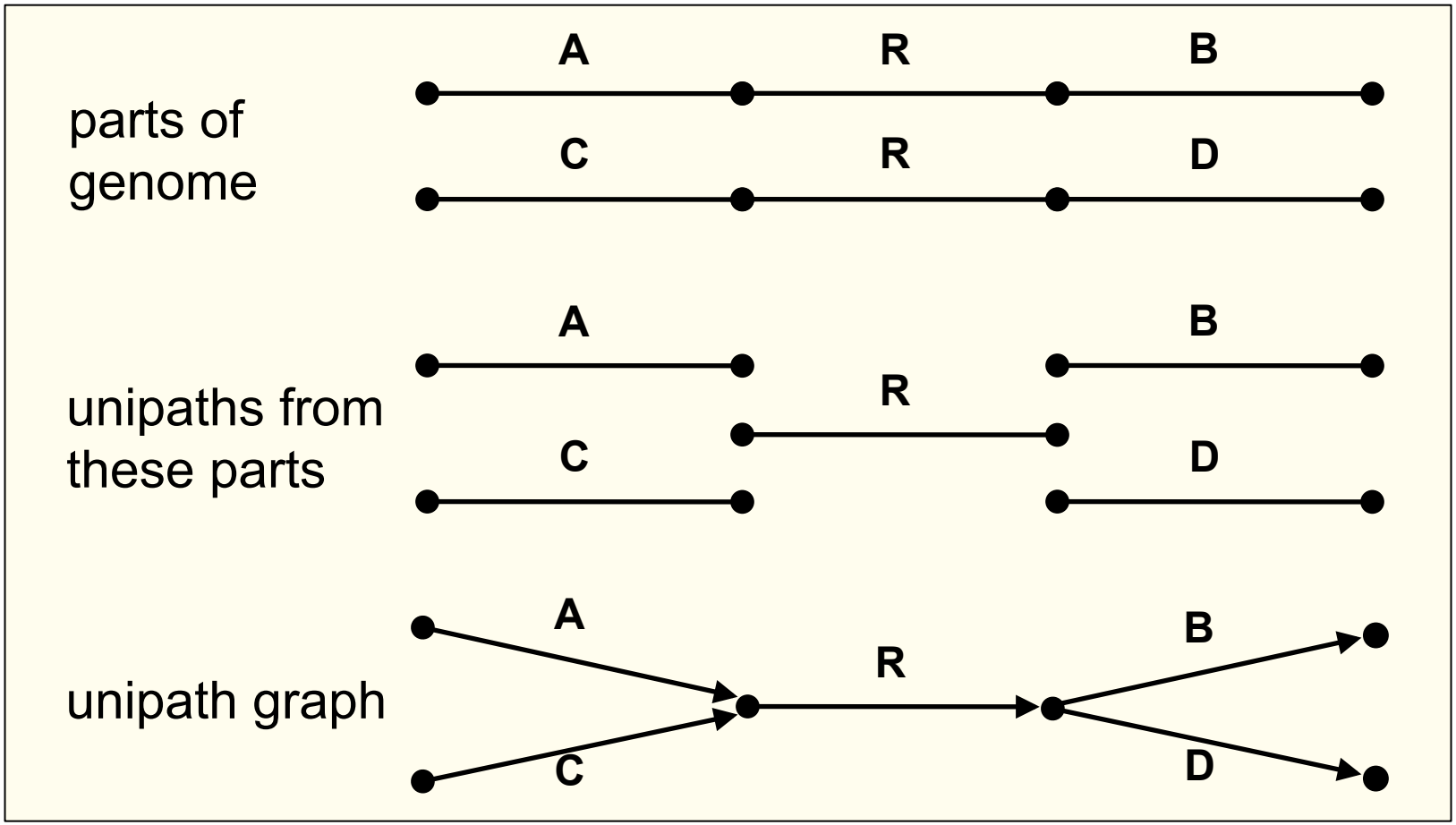
To close a read pair (red), we require the existence of another read pair (blue), overlapping perfectly like this:




More than one closure allowed (but rare).

## Unipaths

*Unipath*: unbranched part of genome – squeeze together perfect repeats of size ≥ K

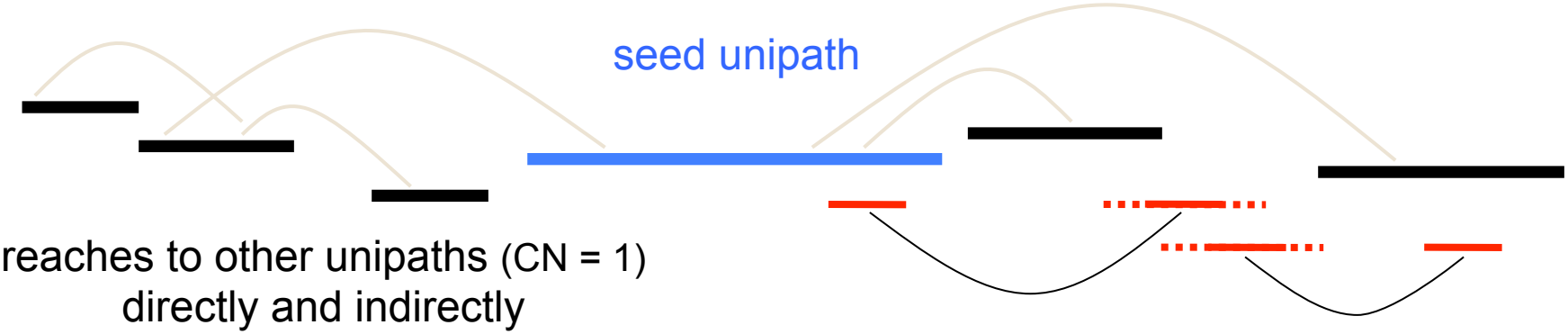



Adjacent unipaths overlap by K-1 bases

# Localization

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

**II. Form neighborhood around each seed**



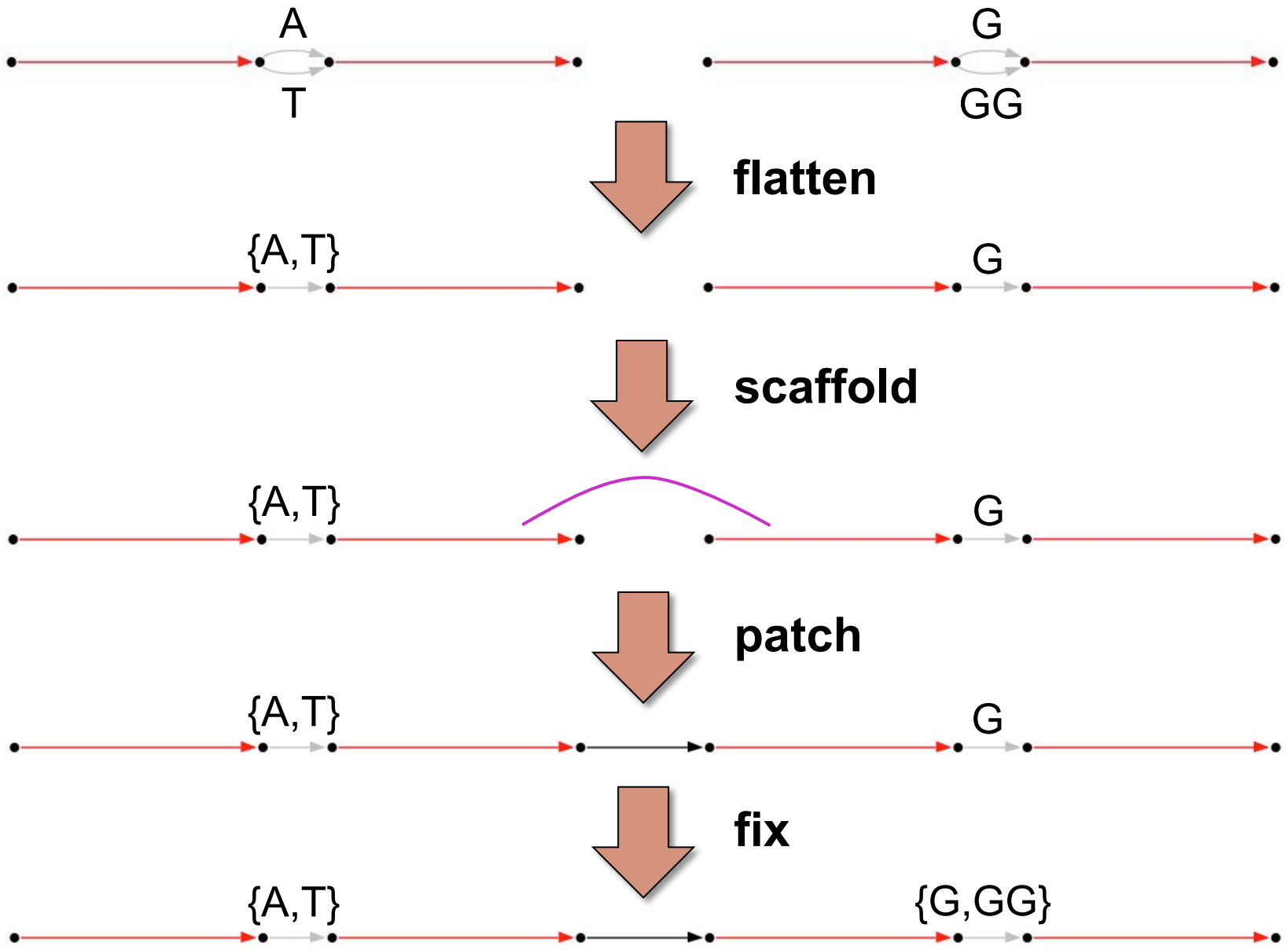

reaches to other unipaths (CN = 1)
directly and indirectly

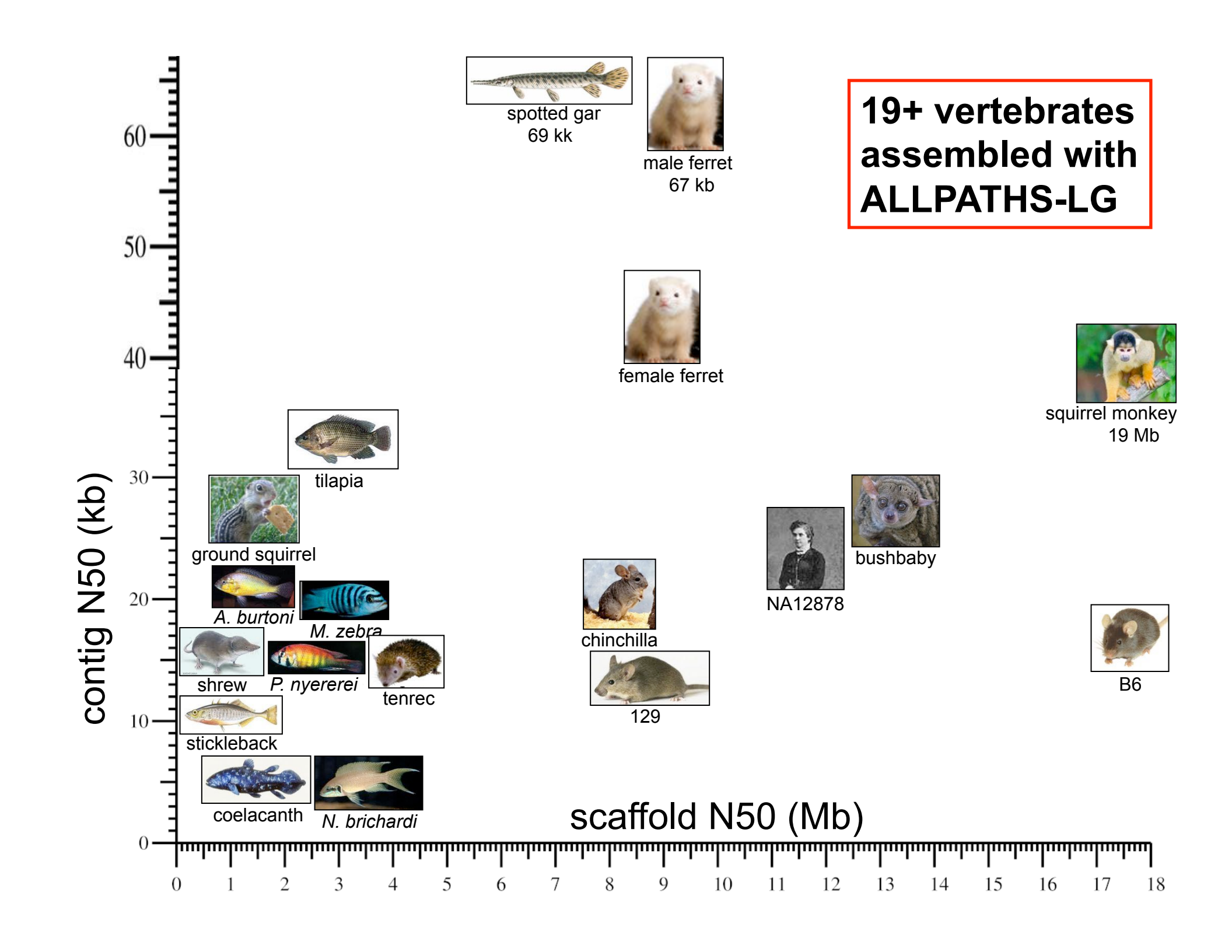
read pairs reach into repeats

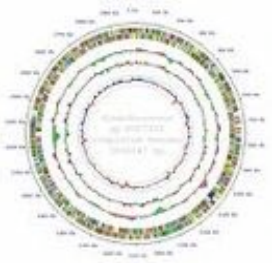
and are extended by other unipaths

.....

## Create assembly from global assembly graph




19+ vertebrates assembled with ALLPATHS-LG
spotted gar
69 kk
male ferret
67 kb
female ferret
squirrel monkey
19 Mb
tilapia
ground squirrel
A. burtoni
M. zebra
shrew
P. nyererei
tenrec
stickleback
coelacanth
N. brichardi
chinchilla
129
NA12878
bushbaby
B6
contig N50 (kb)
scaffold N50 (Mb)
0
10
20
30
40
50
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

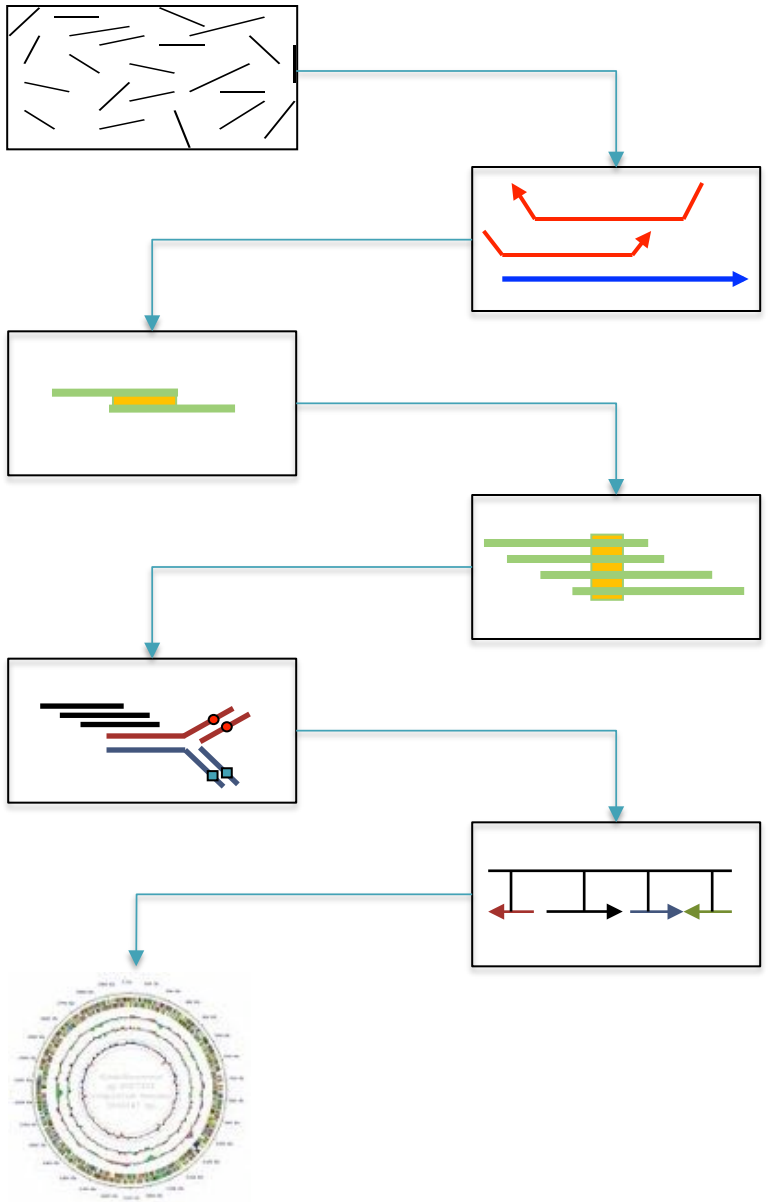


# Genome assembly with the Celera Assembler

# Celera Assembler

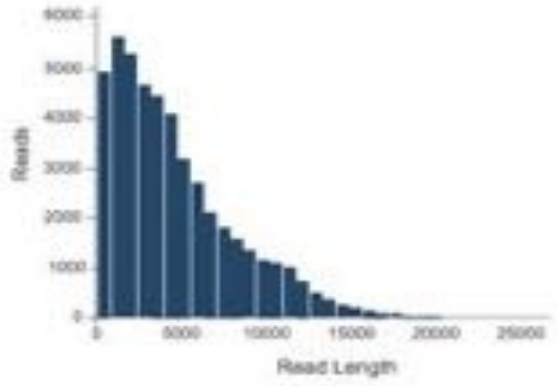
*http://wgs-assembler.sf.net*

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



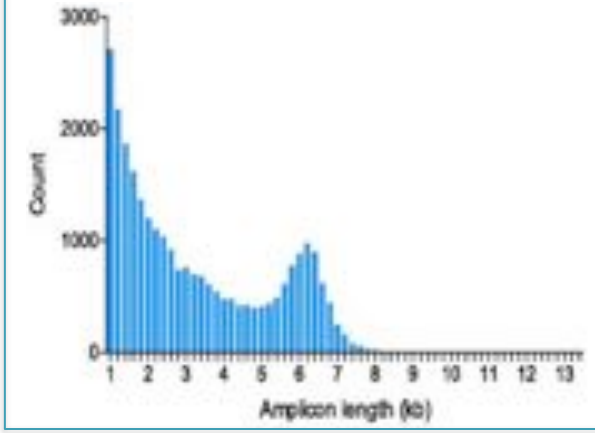
# Single Molecule Sequencing Technology

## PacBio RS II




## Moleculo

illumina
moleculo




## Oxford Nanopore

?

# Hybrid Sequencing

**Illumina**

*Sequencing by Synthesis*

High throughput (60Gbp/day)
High accuracy (~99%)
Short reads (~100bp)

**Pacific Biosciences**

*SMRT Sequencing*

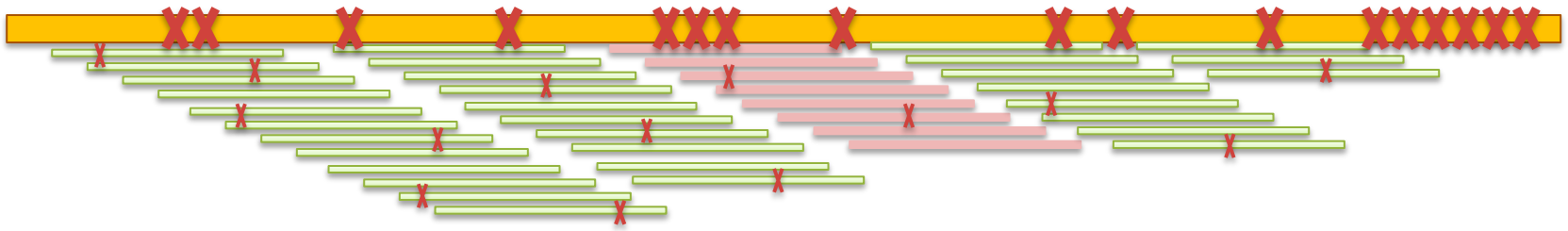
Lower throughput (1Gbp/day)
Lower accuracy (~85%)
Long reads (5kbp+)

# Hybrid Error Correction: PacBioToCA

http://wgs-assembler.sf.net

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

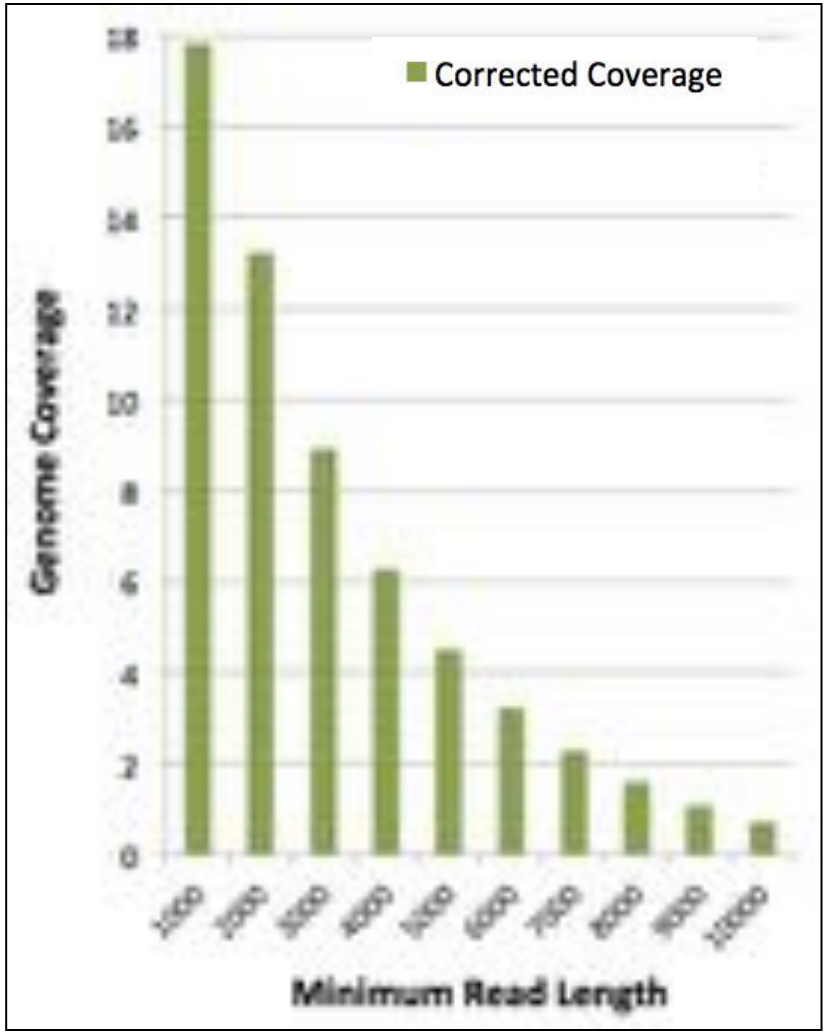
2. Error corrected reads can be easily assembled, aligned



**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<br>50x 2x100bp @ 180 | 3,925 |
| MiSeq Fragments<br>23x 459bp<br>8x 2x251bp @ 450 | 6,332 |
| "ALLPATHS-recipe"<br>50x 2x100bp @ 180<br>36x 2x50bp @ 2100<br>51x 2x50bp @ 4800 | 18,248 |




In collaboration with McCombie & Ware labs @ CSHL

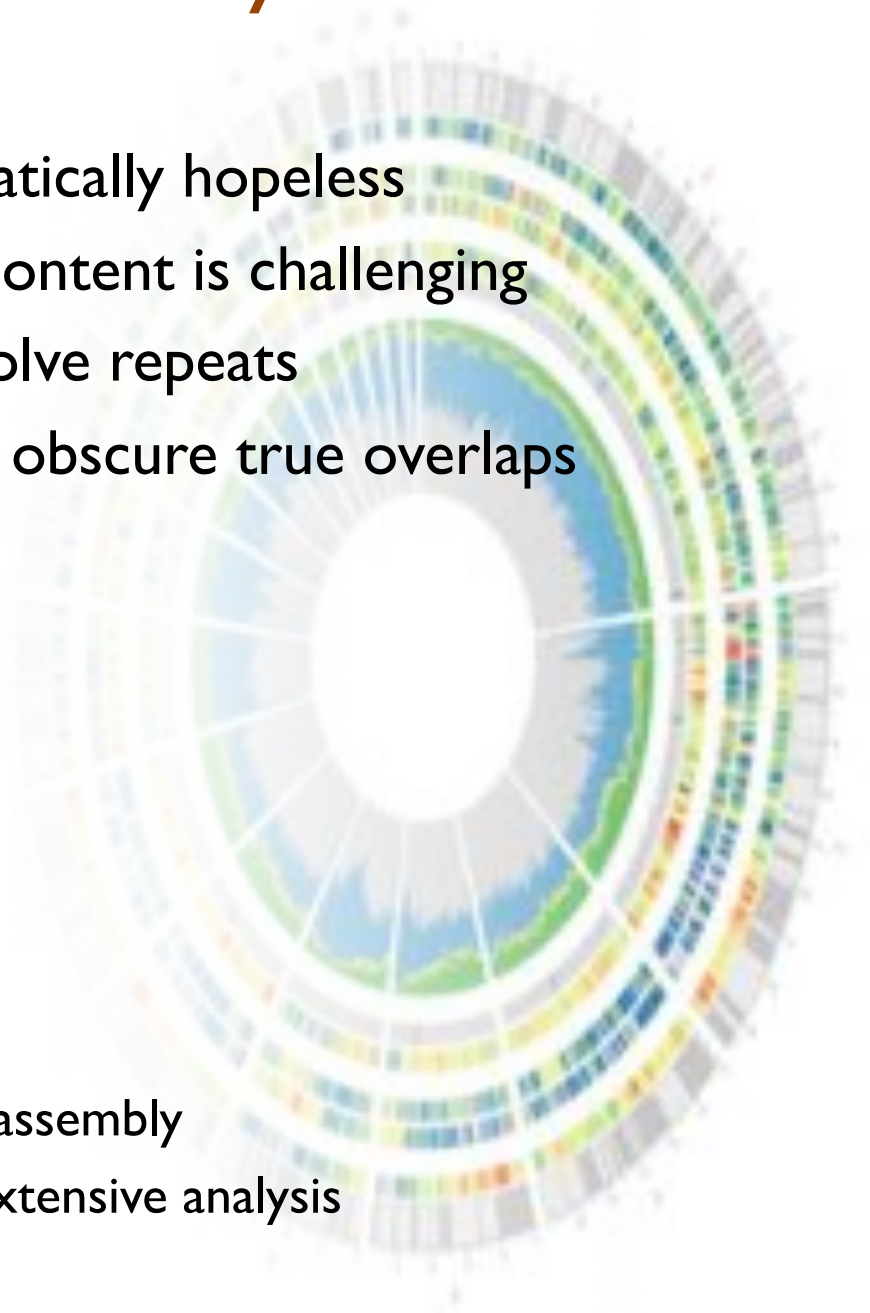
# Assembly Summary

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.br adinstitute.org/pub/crd/ LPATHS/Release-LG/
% configure && ma  make install
```

1. Collect the BA that you w Create a in_libs.c ribe roups.csv metadata file t

2. Prepare input files

```
% cd /tmp/cshl/asm
% PrepareAllPathsInpu
    DATA_DIR=`pwd`
```

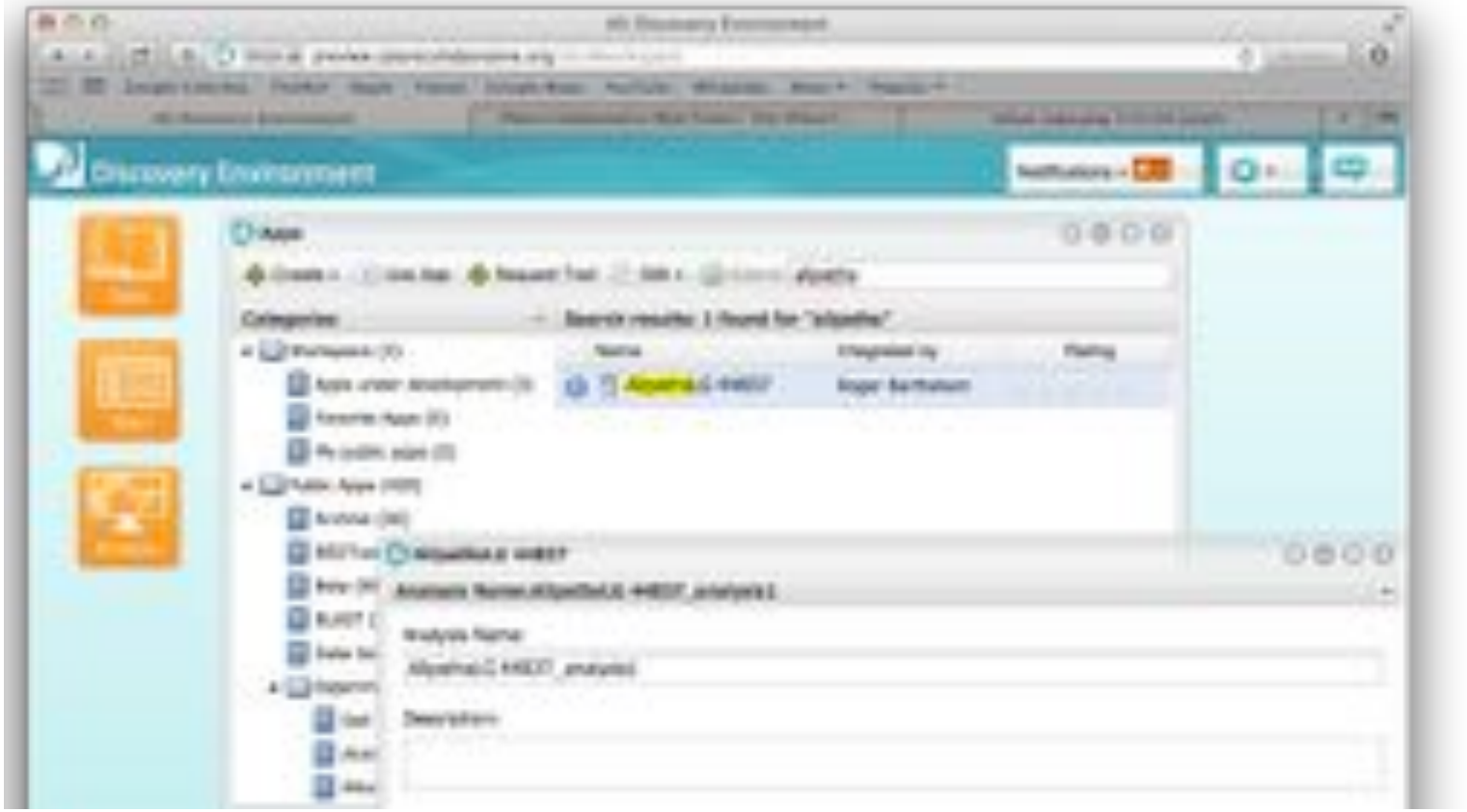
3. Assemble.

```
% RunAllPa
    PRE=/t -csh
    DATA_SUB efault >&
```

4. Get the results (fou

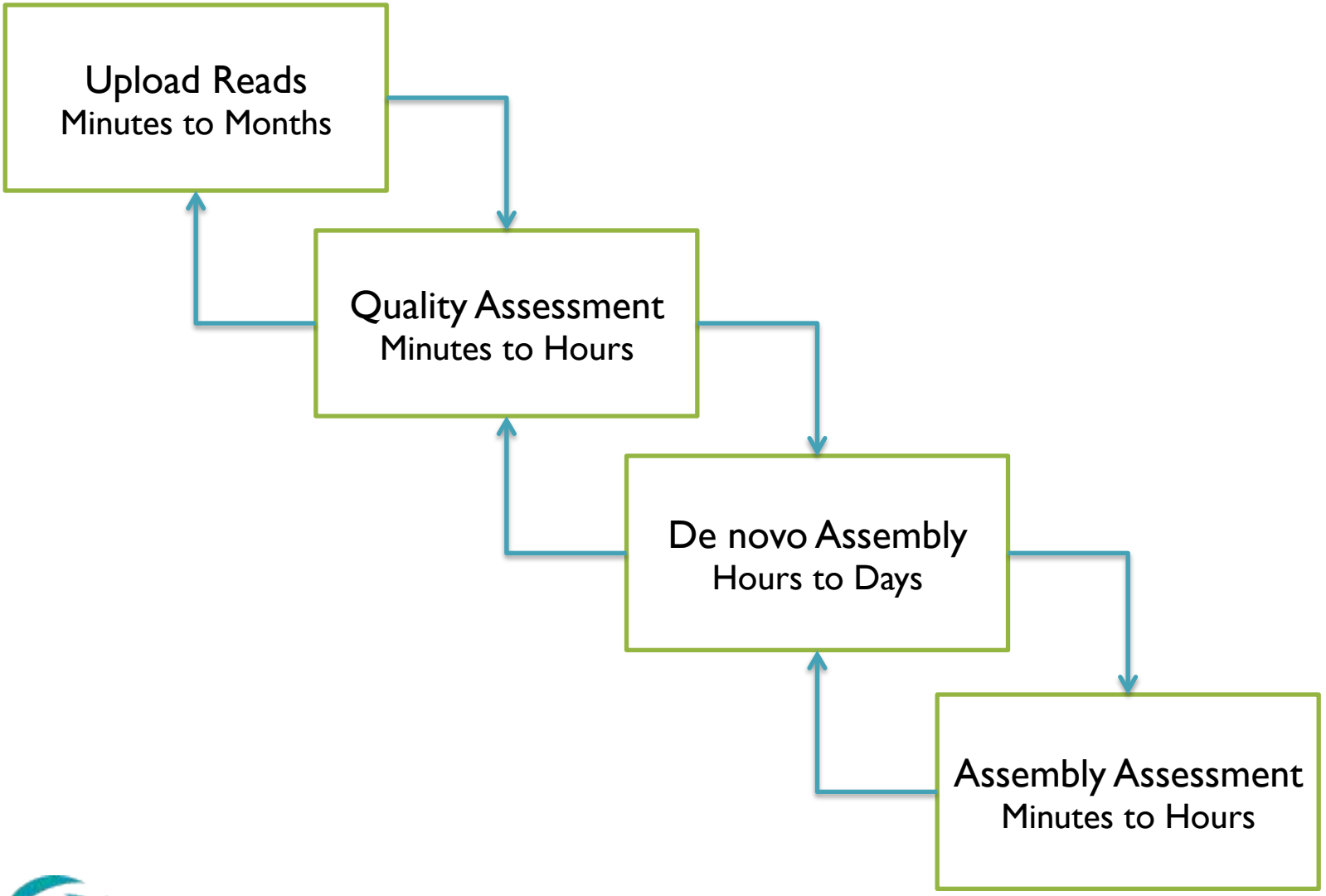
```
% cd /tmp/cshl/as /default/ASSEMBLIES/te t/
% less final.{assembly,contigs}.{fasta,efasta}
```

# Assembly with iPlant



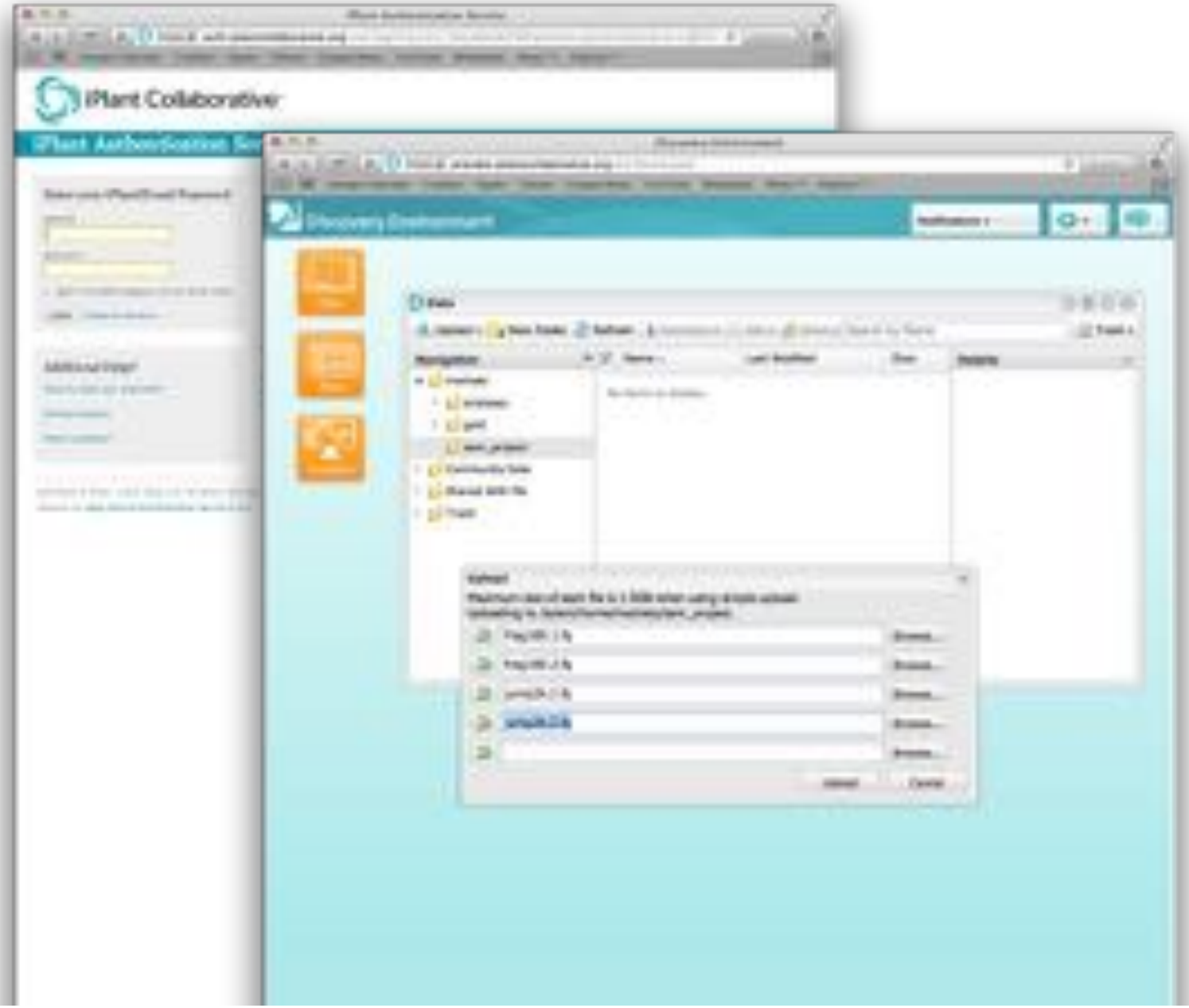

iPlant Collaborative™ *Empowering A New Plant Biology*

# Assembly Workflow

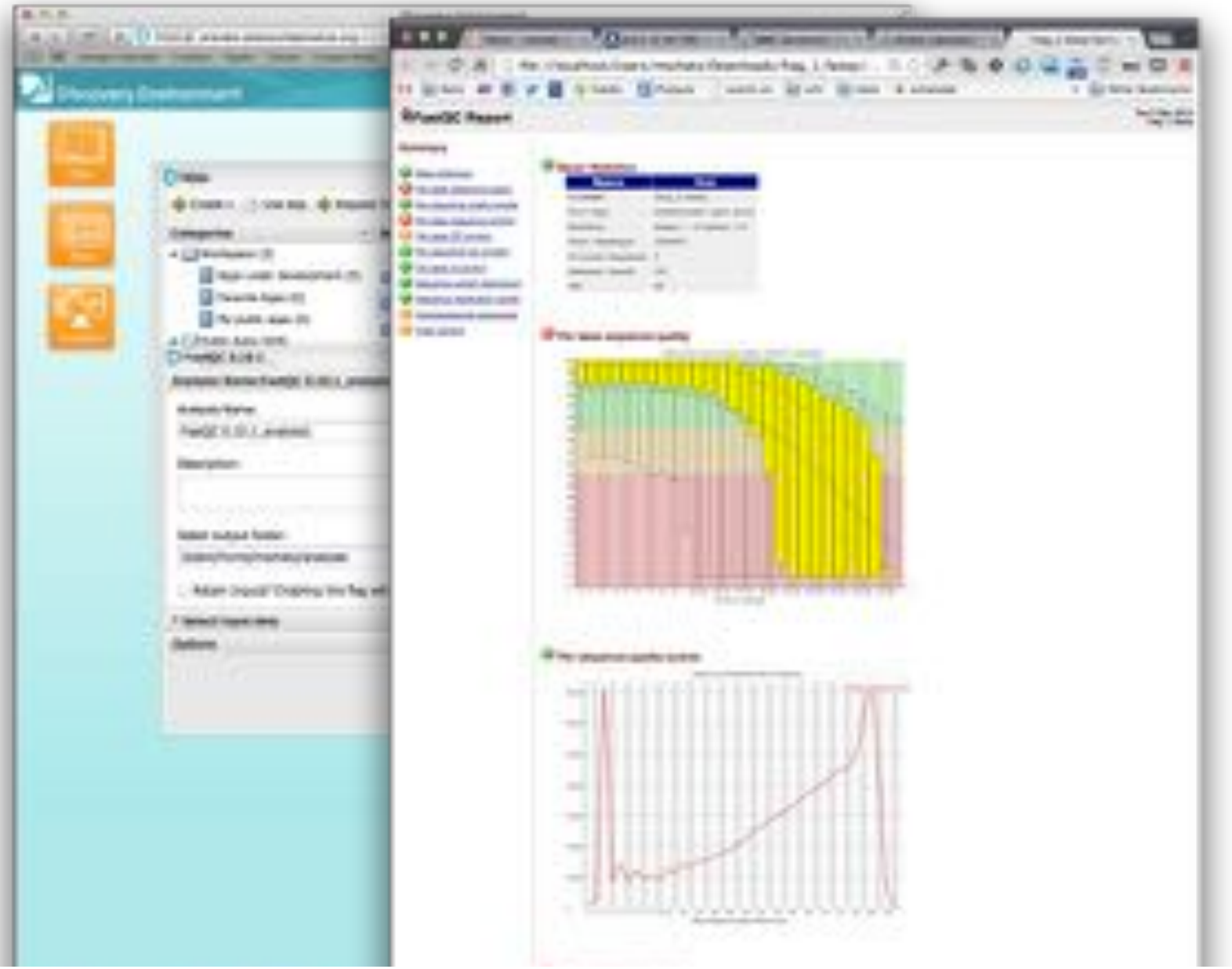



iPlant Collaborative™ *Empowering A New Plant Biology*

# Upload Reads



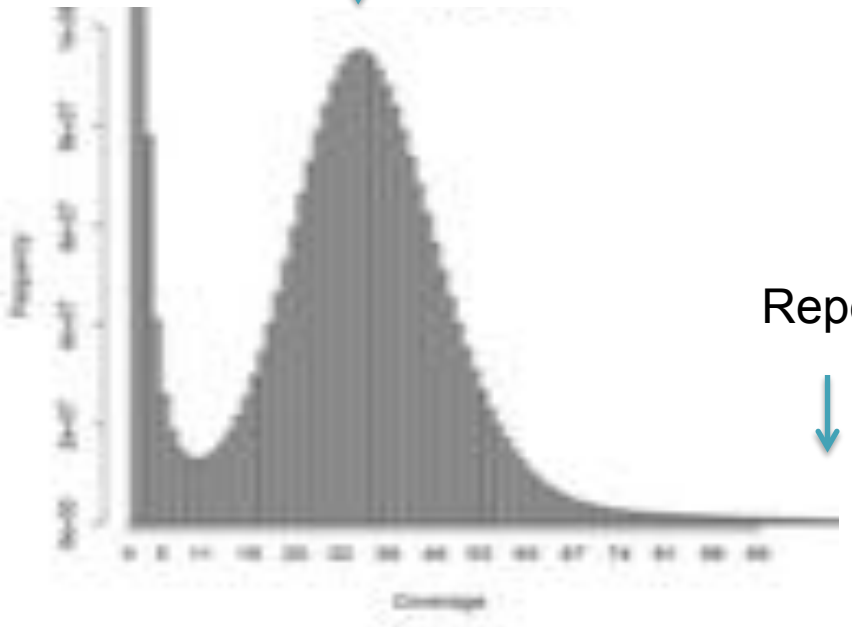
# QC: FastQC



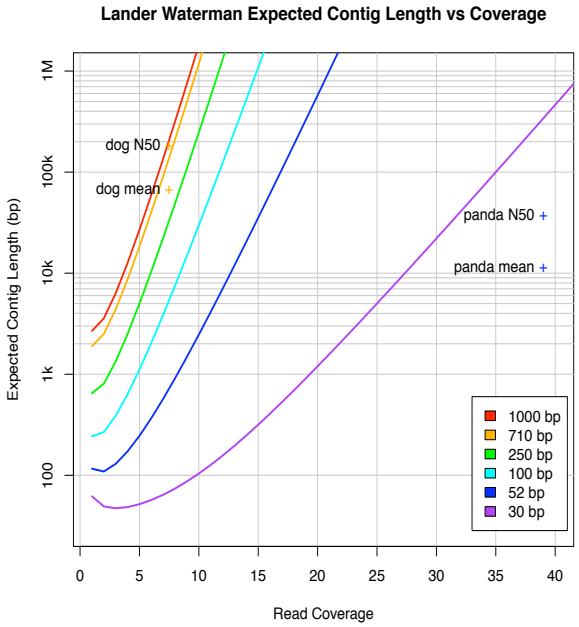
# QC: Read Coverage


Errors

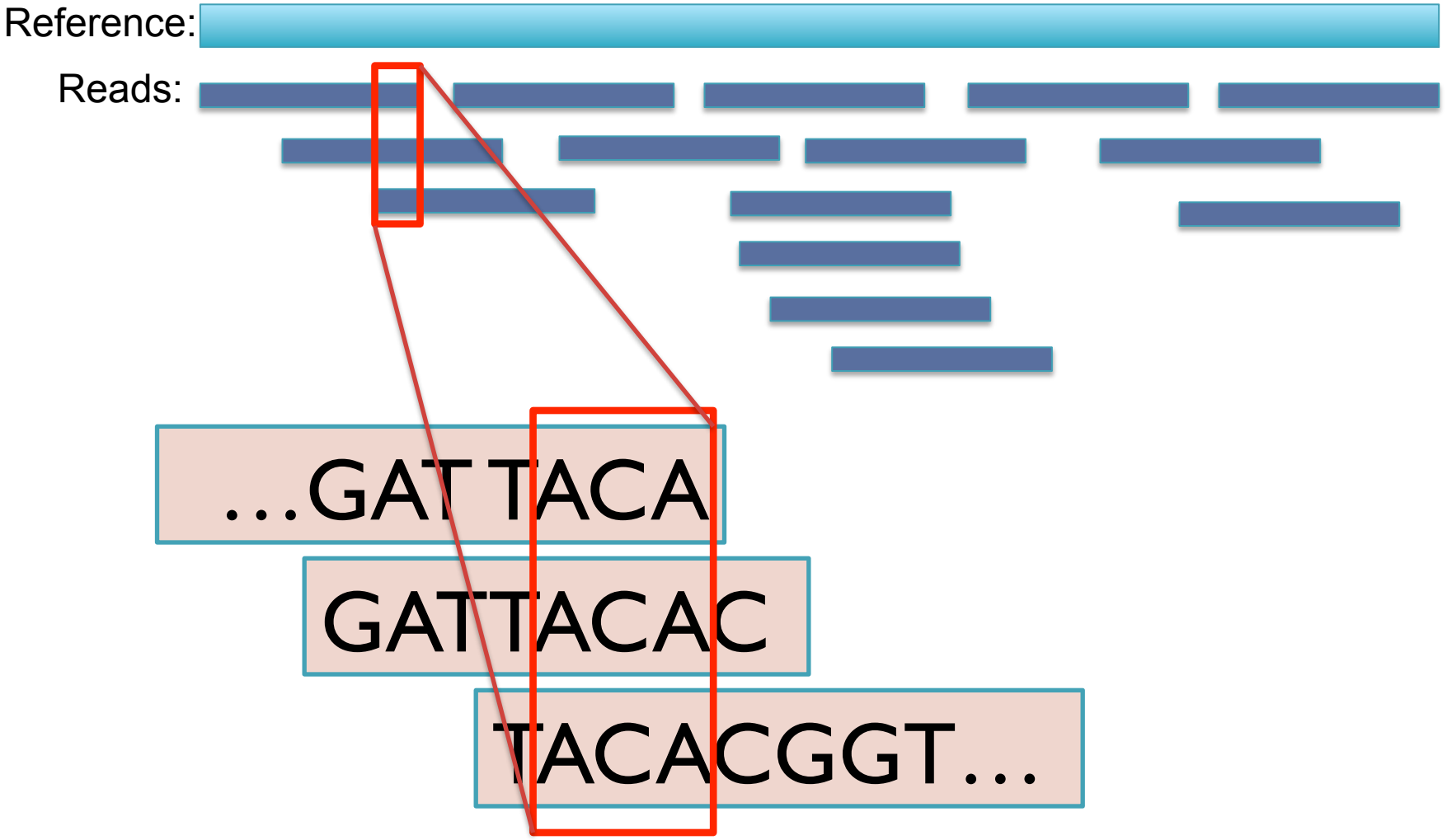
Coverage



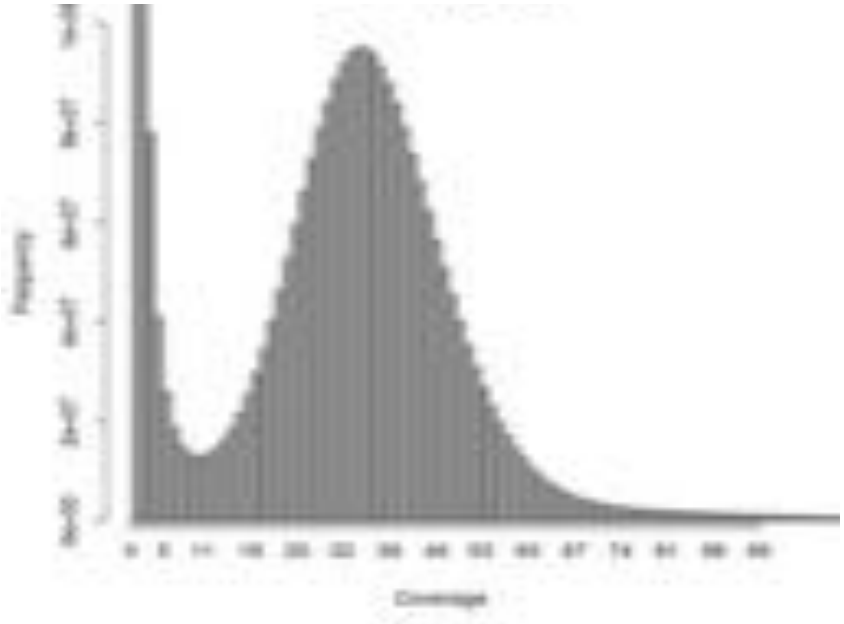
Repeats

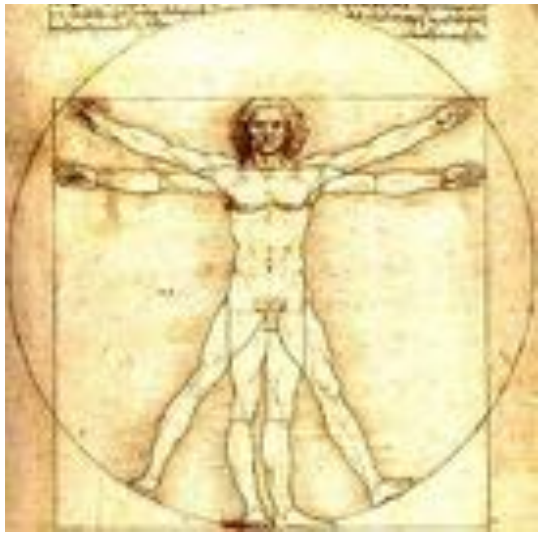


# Estimating coverage with Kmers



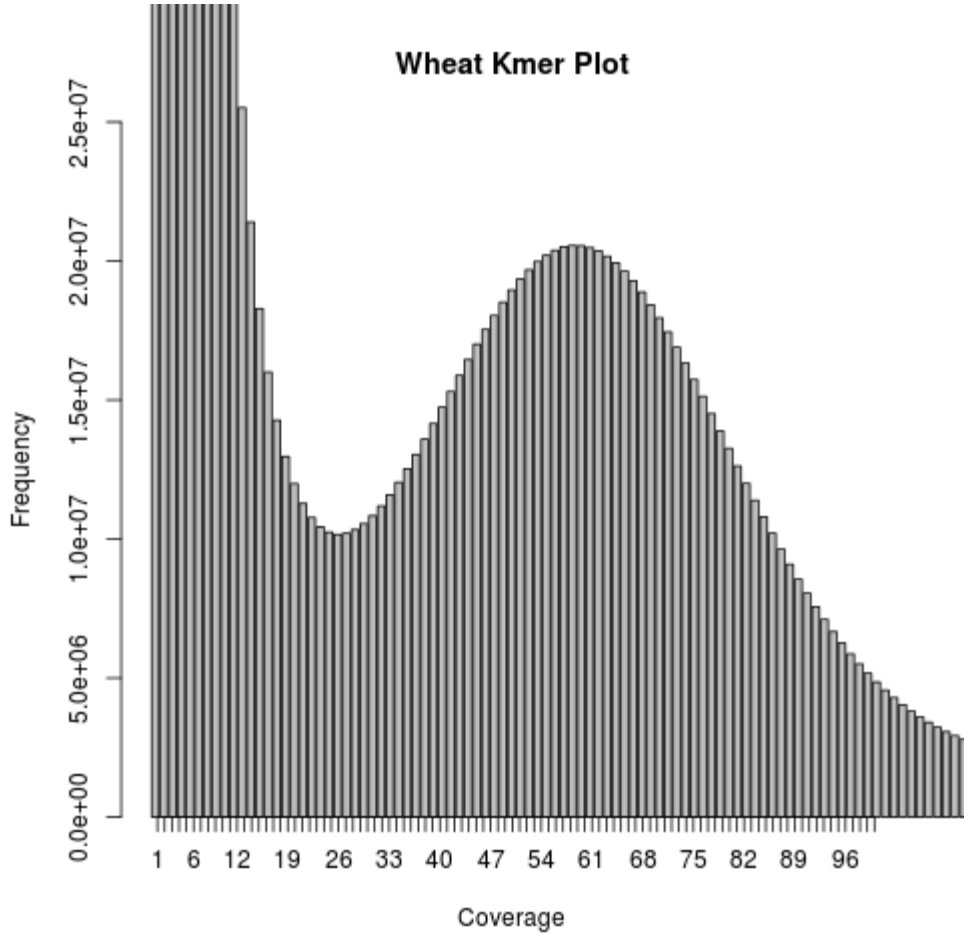
# Estimating coverage with Kmers

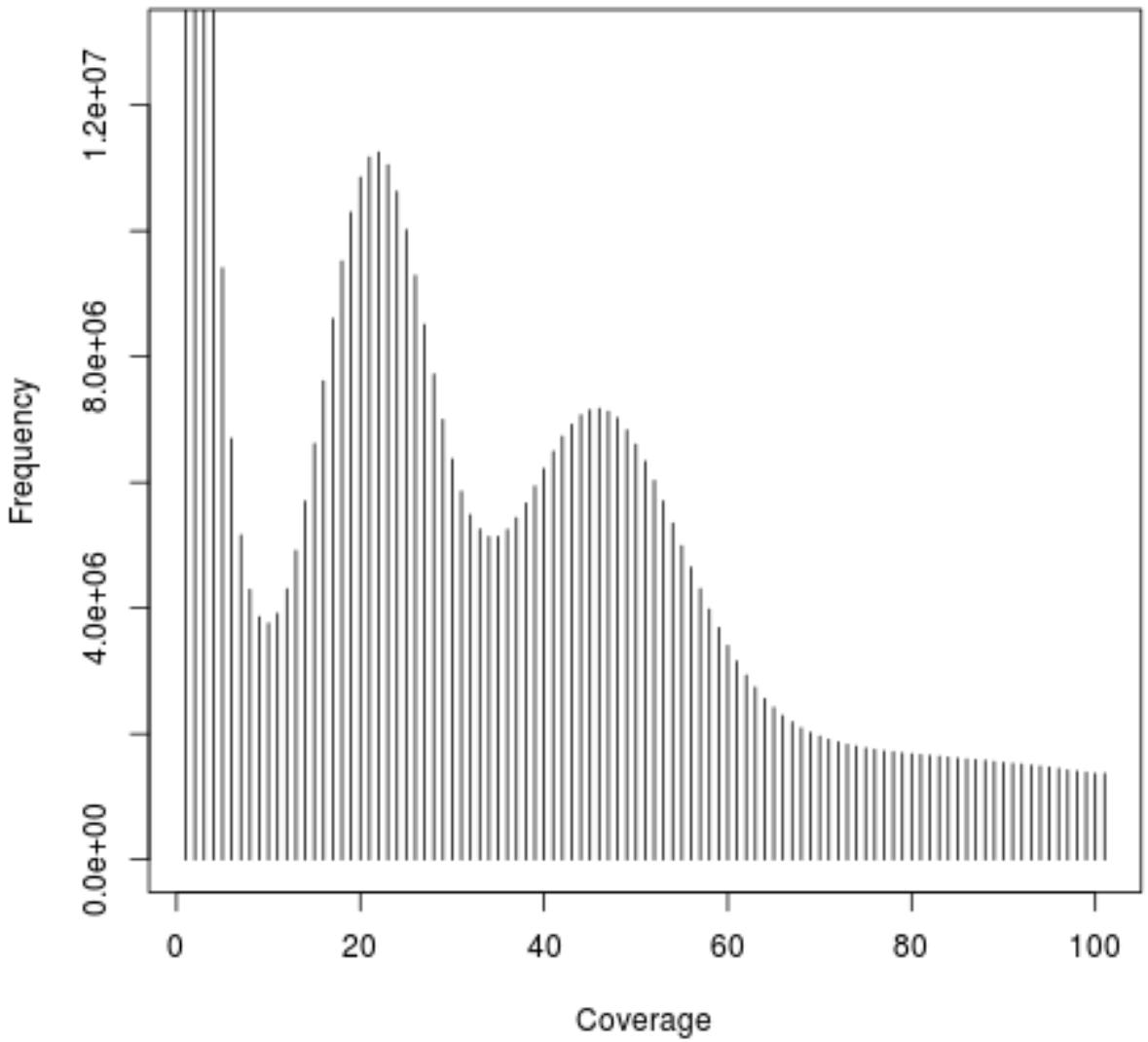


NA12878

# Wheat Genome
## (A. tauschi / CSHL)

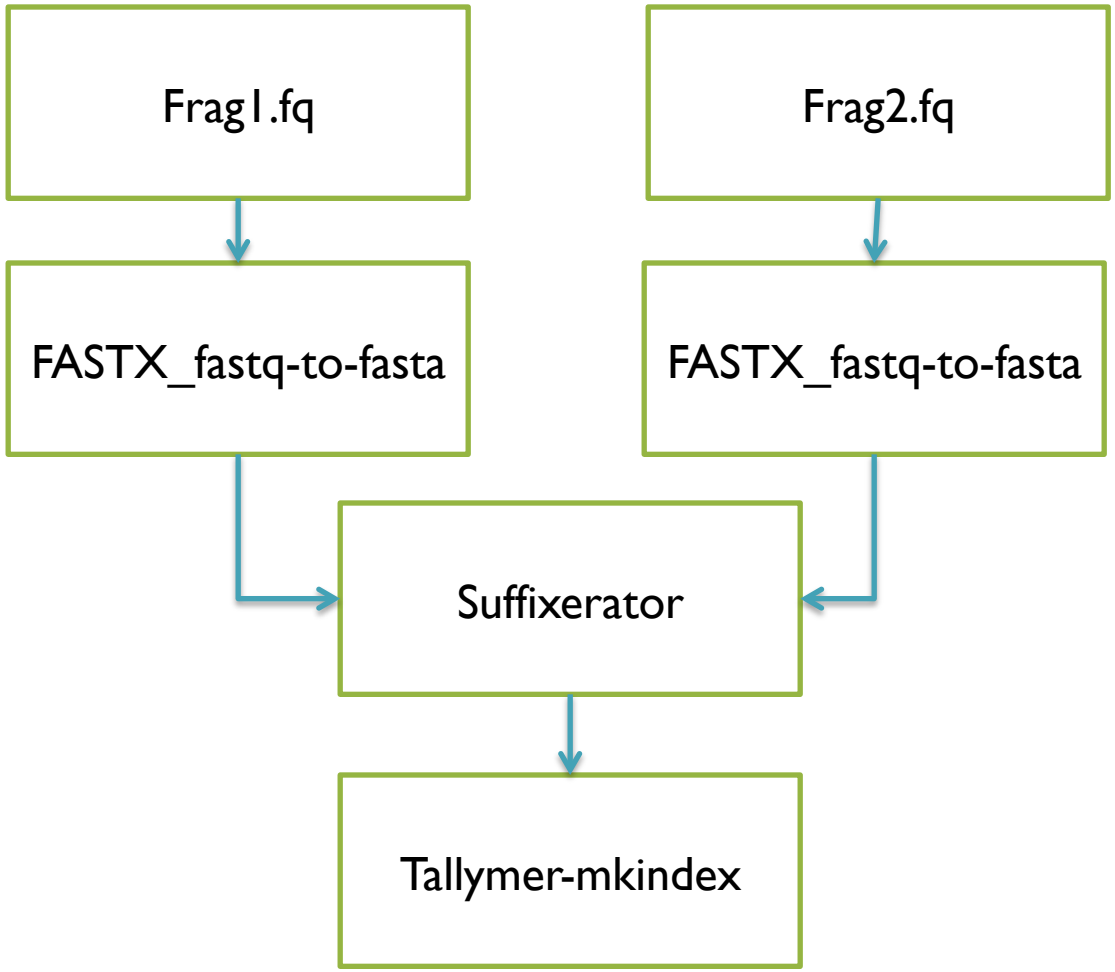


# Heterozygous Genome



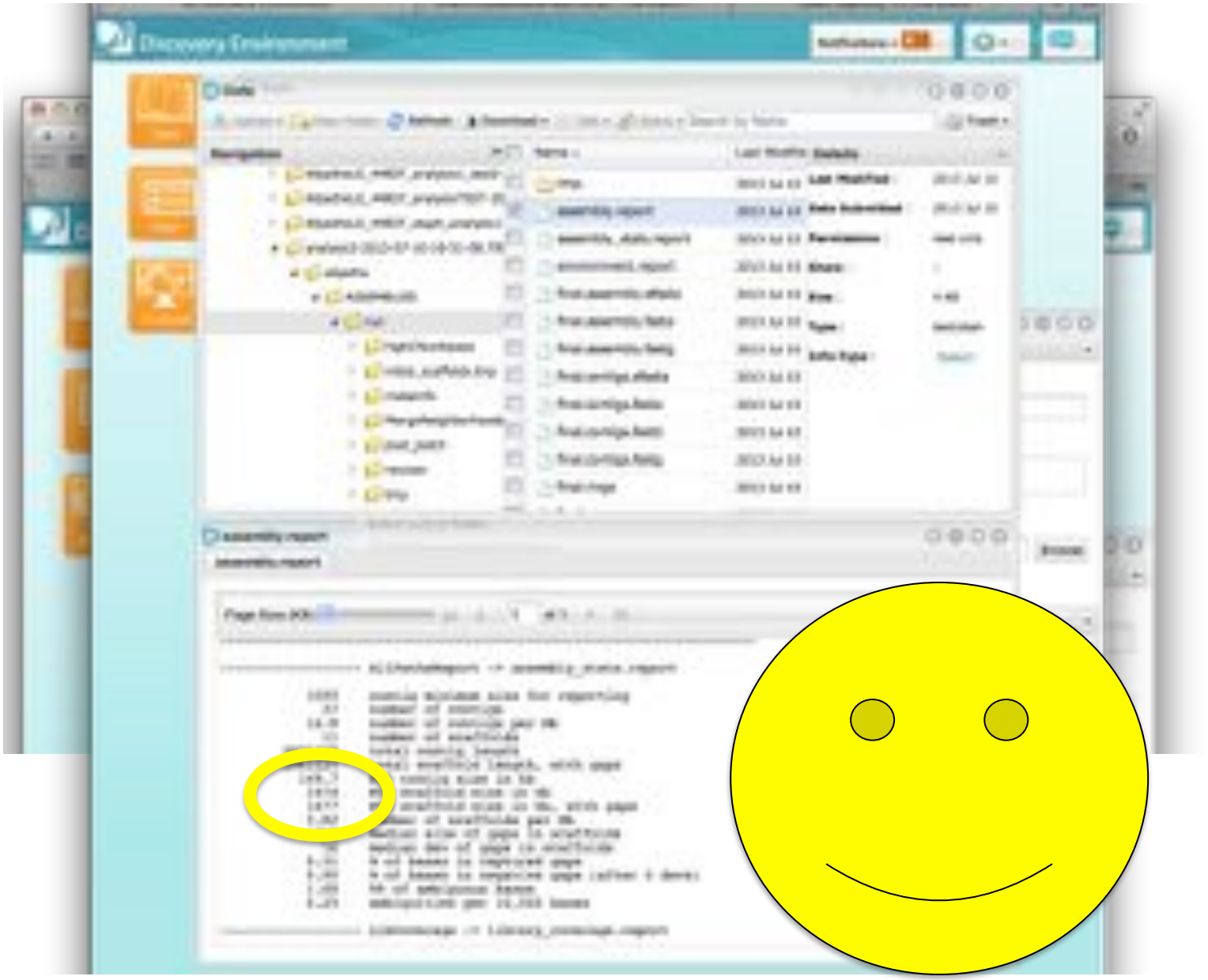

Contact: @mike_schatz

# 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

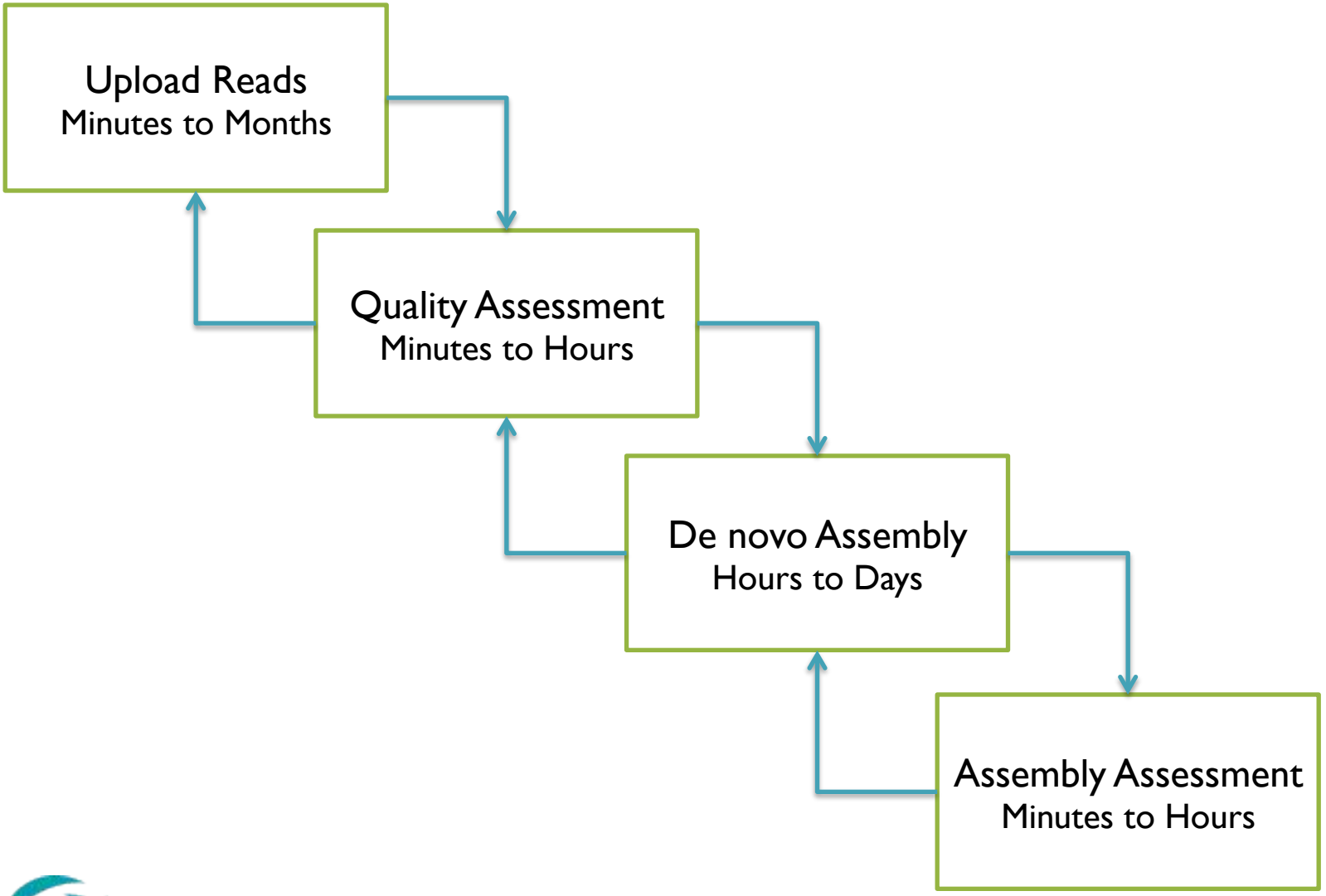
# Running ALLPATHS-LG



# Post-QC: CEGMA

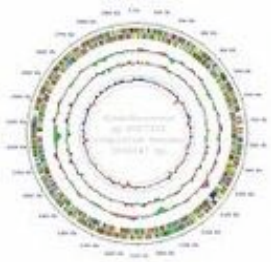
**CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes**
Parra G, Bradnam K, Korf I. (2007) Bioinformatics. 23 (9): 1061-1067.

# Assembly Workflow




iPlant Collaborative™ *Empowering A New Plant Biology*

# Resources



- iPlant
  - http://www.iplantcollaborative.org/
- Assembly Competitions
  - Assemblathon: http://assemblathon.org/
  - GAGE: http://gage.cbcb.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 |
|---|---|---|
| [illegible] | PI and Project Director, iPlant Collaborative | University of Arizona |
| [illegible] | Co-PI, iPlant Collaborative | University of Texas, Austin |

Staff

| Name | Title | Institution |
|---|---|---|
| [illegible] | Research Scientist | University of Texas, Austin |
| [illegible] | Research Associate | University of Texas, Austin |
| [illegible] | Bioinformatics Analyst | University of Arizona |
| [illegible] | QA Test Engineer | University of Arizona |

iPlant Collaborative™ *Empowering A New Plant Biology*

# Questions?

http://schatzlab.cshl.edu/

@mike_schatz