# Genome Sequencing & Assembly

Michael Schatz

May 2, 2013

Human Microbiome Consortium

# Outline

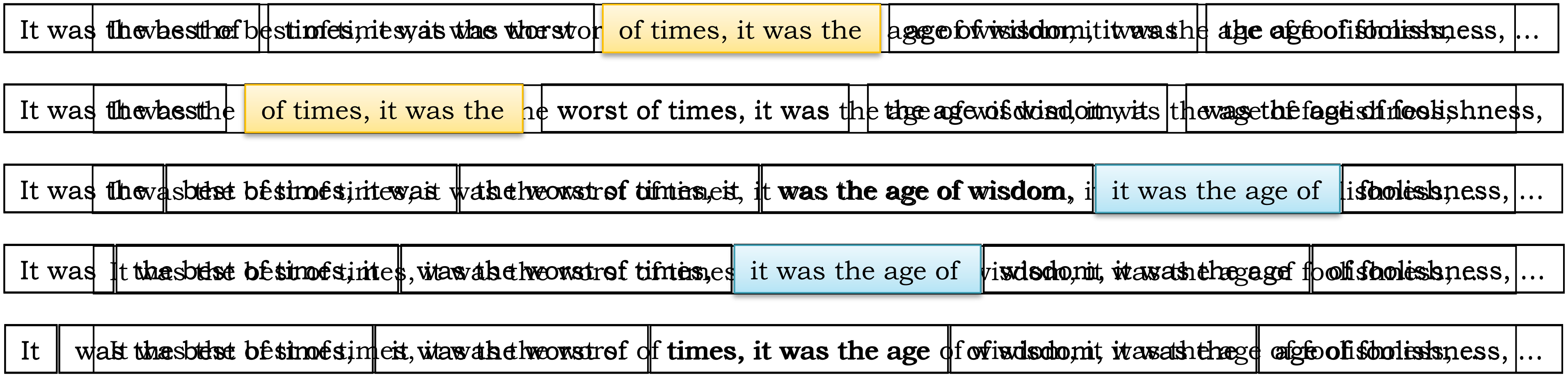
1. Assembly theory
   1. Assembly by analogy
   2. De Bruijn and Overlap graph
   3. Coverage, read length, errors, and repeats

2. Genome assemblers
   1. Assemblathon 1 & 2
   2. Hybrid assembly with the Celera Assembler

3. Resources

# 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,

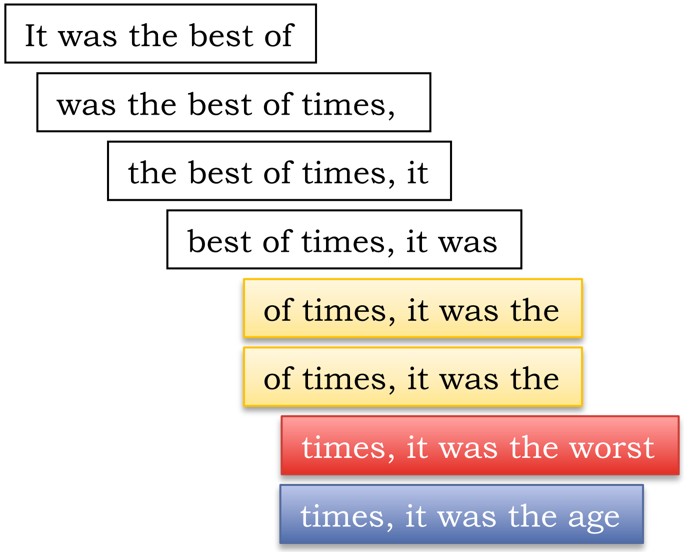
was the age of foolishness,

was the best of times,

was the worst of times,

wisdom, it was the age

worst of times, it was



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

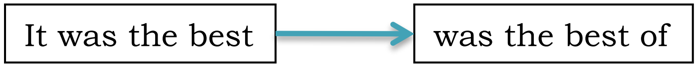
# de Bruijn Graph Construction

- $D_k = (V,E)$
  - 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

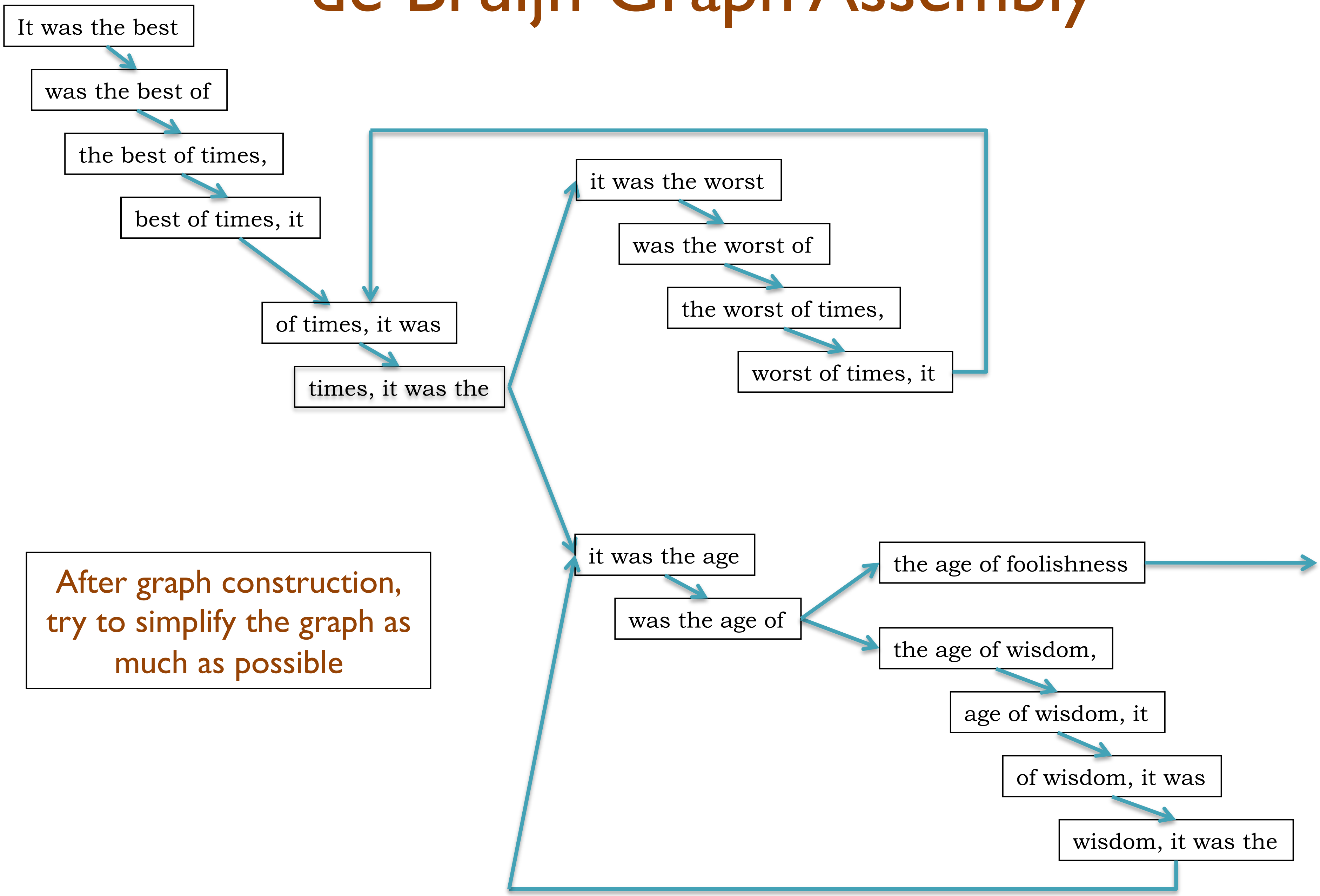
Directed Edge



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

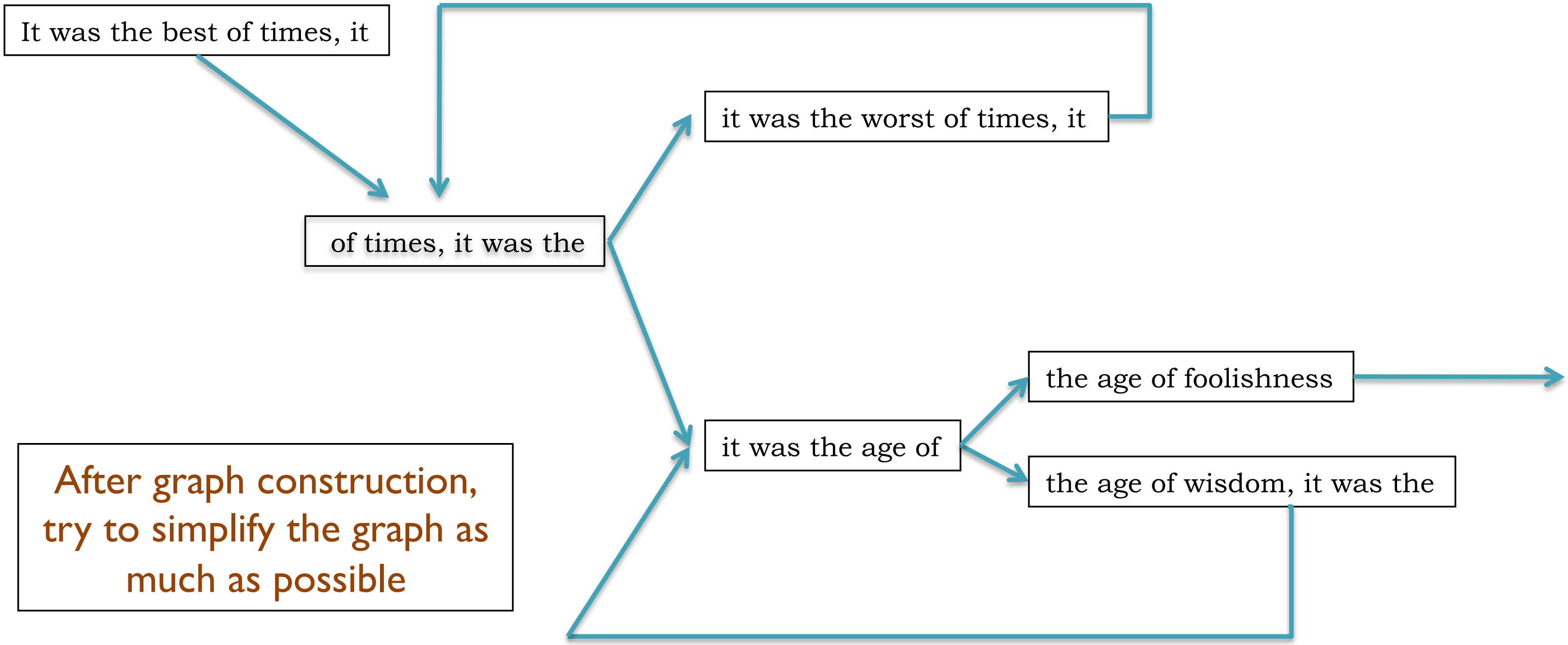
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

# de Bruijn Graph Assembly



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

# de Bruijn Graph Assembly



It was the best of times, it

of times, it was the

it was the worst of times, it

it was the age of

the age of foolishness

the age of wisdom, it was the

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

# The full tale

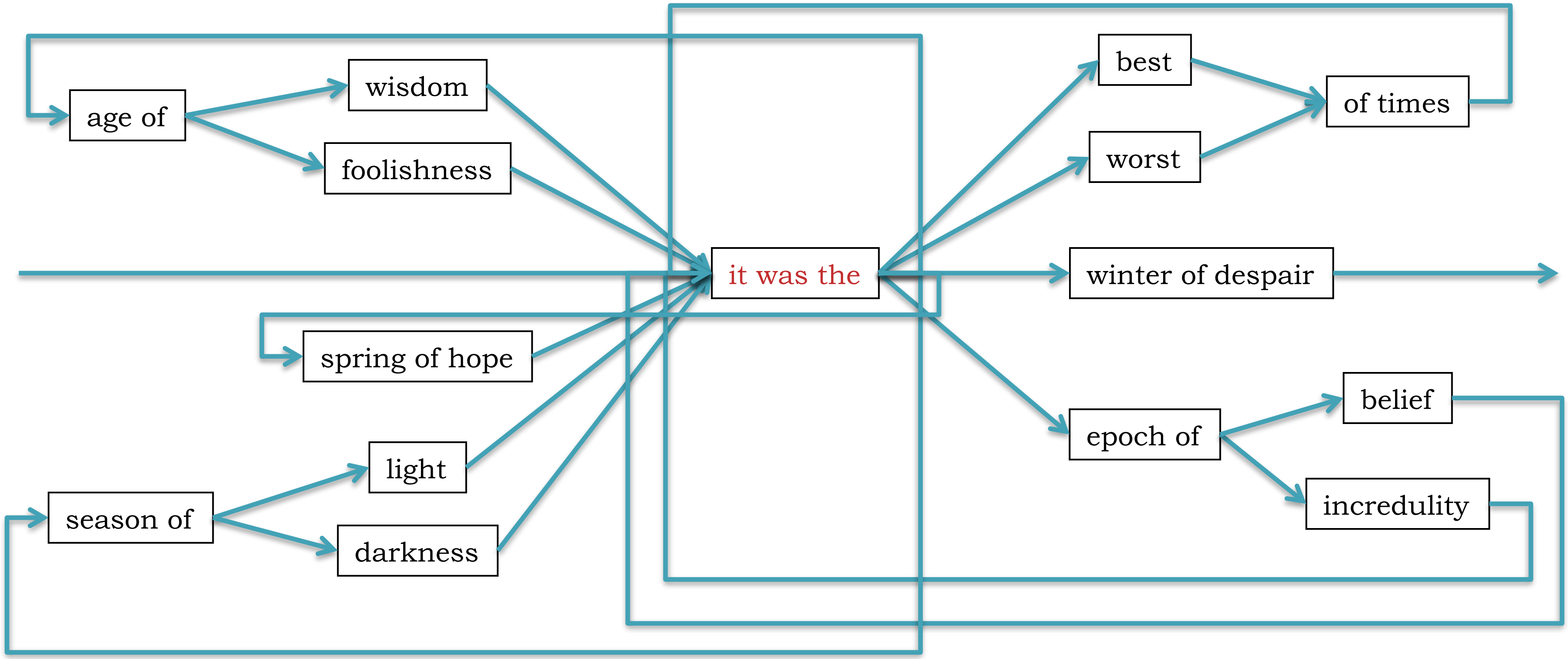
… it was the best of times it was the worst of times …

… 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 …

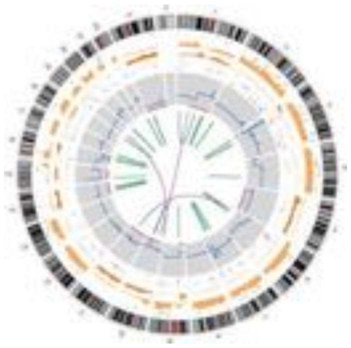


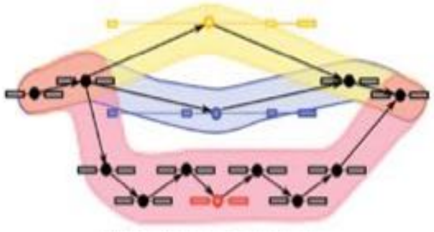
# Assembly Applications

- Novel genomes

- Metagenomes

- Sequencing assays
  - Structural variations
  - Transcript assembly

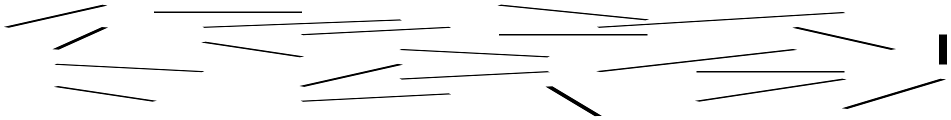




Like Dickens, have to reconstruct from short fragments

# Assembling a Genome

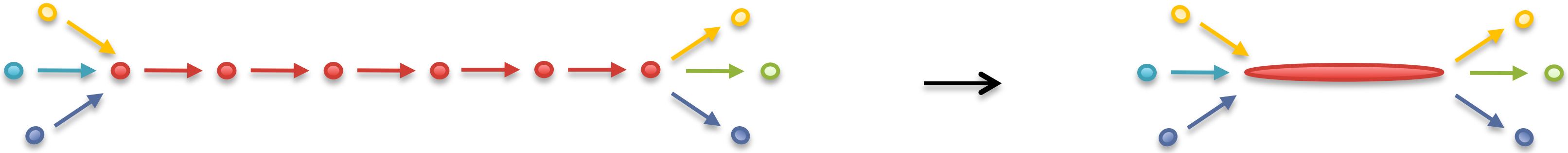
1. Shear & Sequence DNA



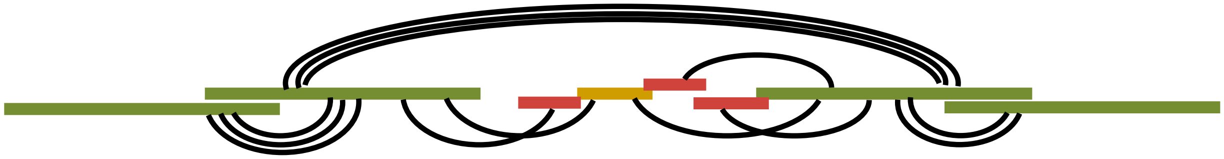
2. Construct assembly graph from overlapping reads

```
...AGCCTAGGGATGCGCGACACGT
          GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
                                          CAACCTCGGACGGACCTCAGCGAA...
```

3. Simplify assembly graph

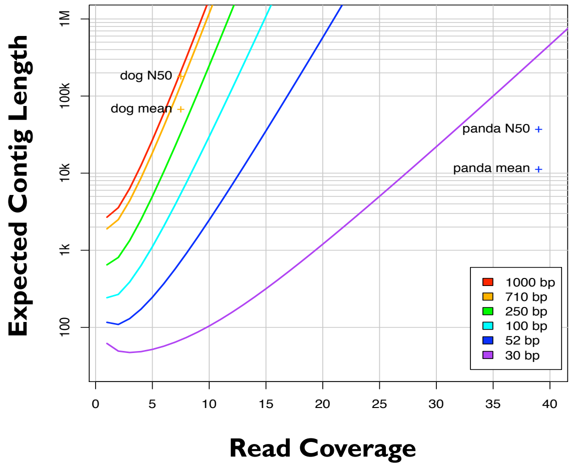


4. Detangle graph with long reads, mates, and other links



# Ingredients for a good assembly

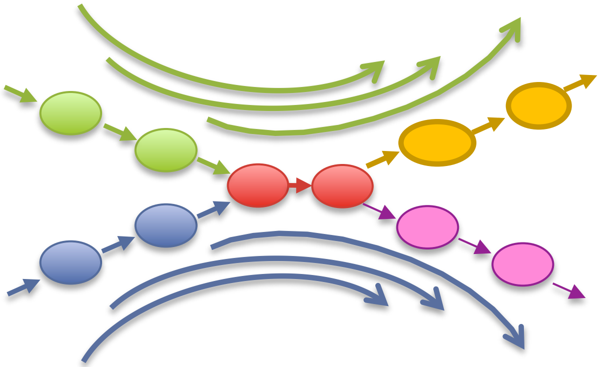
## Coverage



***High coverage is required***

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

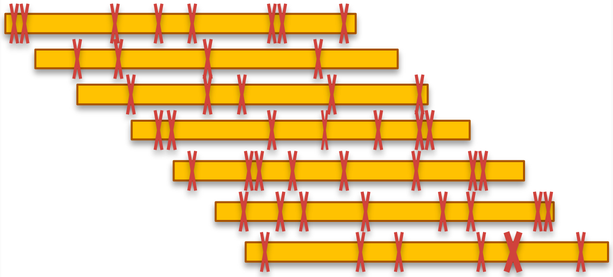
## Read Length



***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

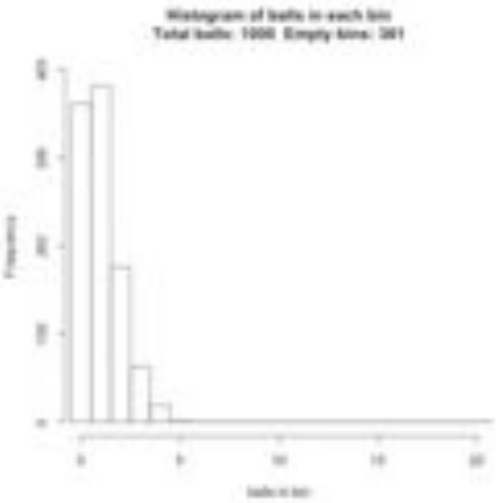


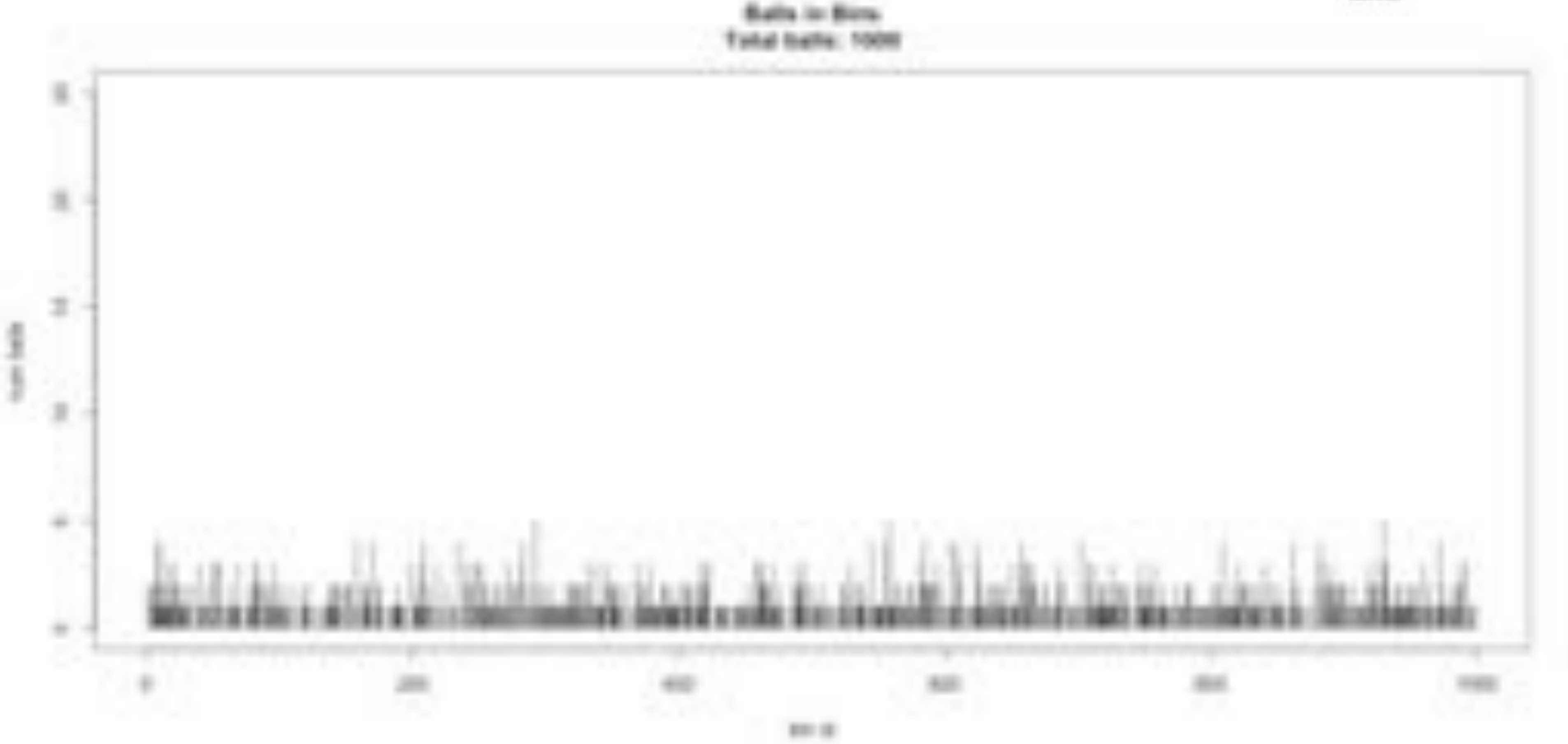
***Errors obscure overlaps***

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

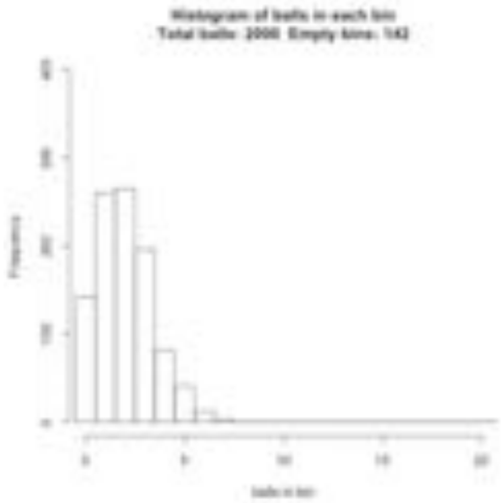
**Current challenges in *de novo* plant genome sequencing and assembly**
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

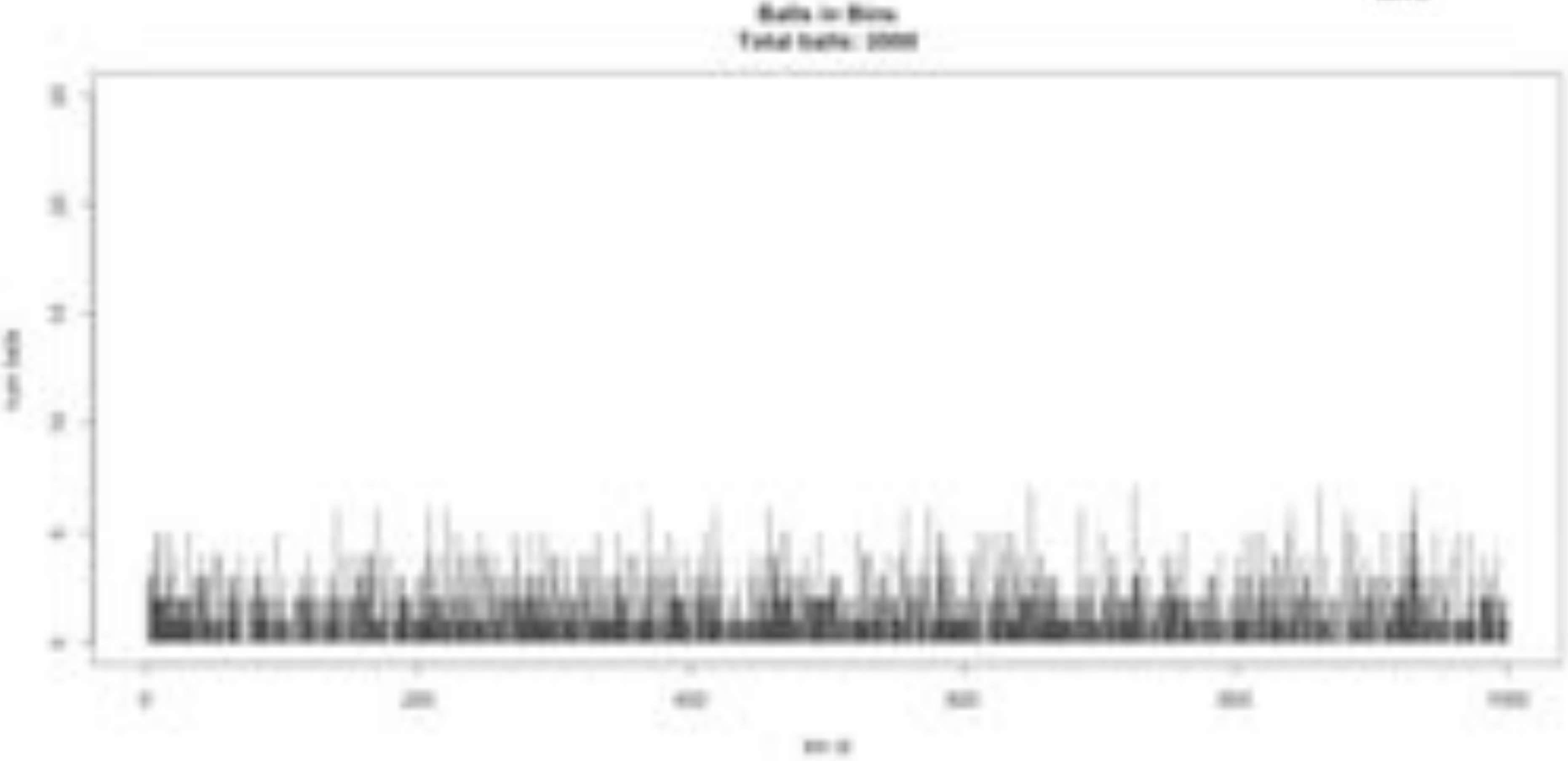
# Balls in Bins 1x



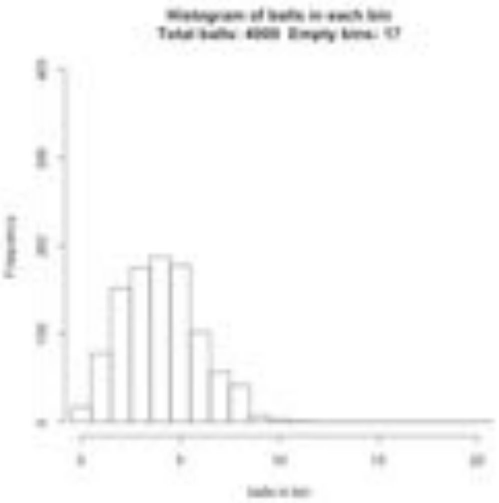


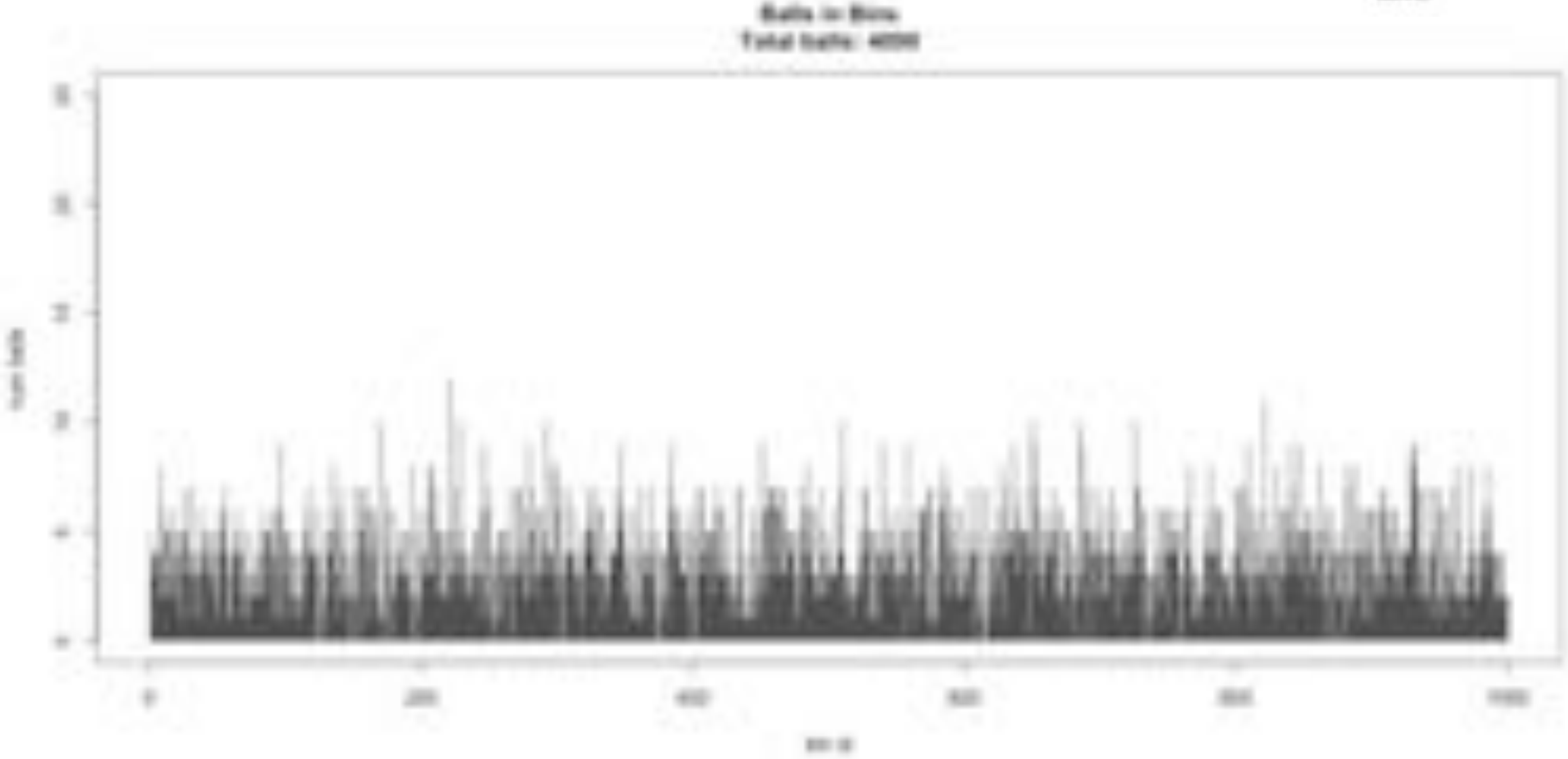
# Balls in Bins 2x



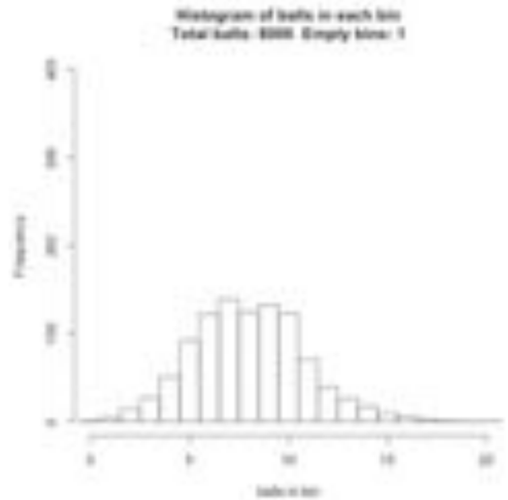


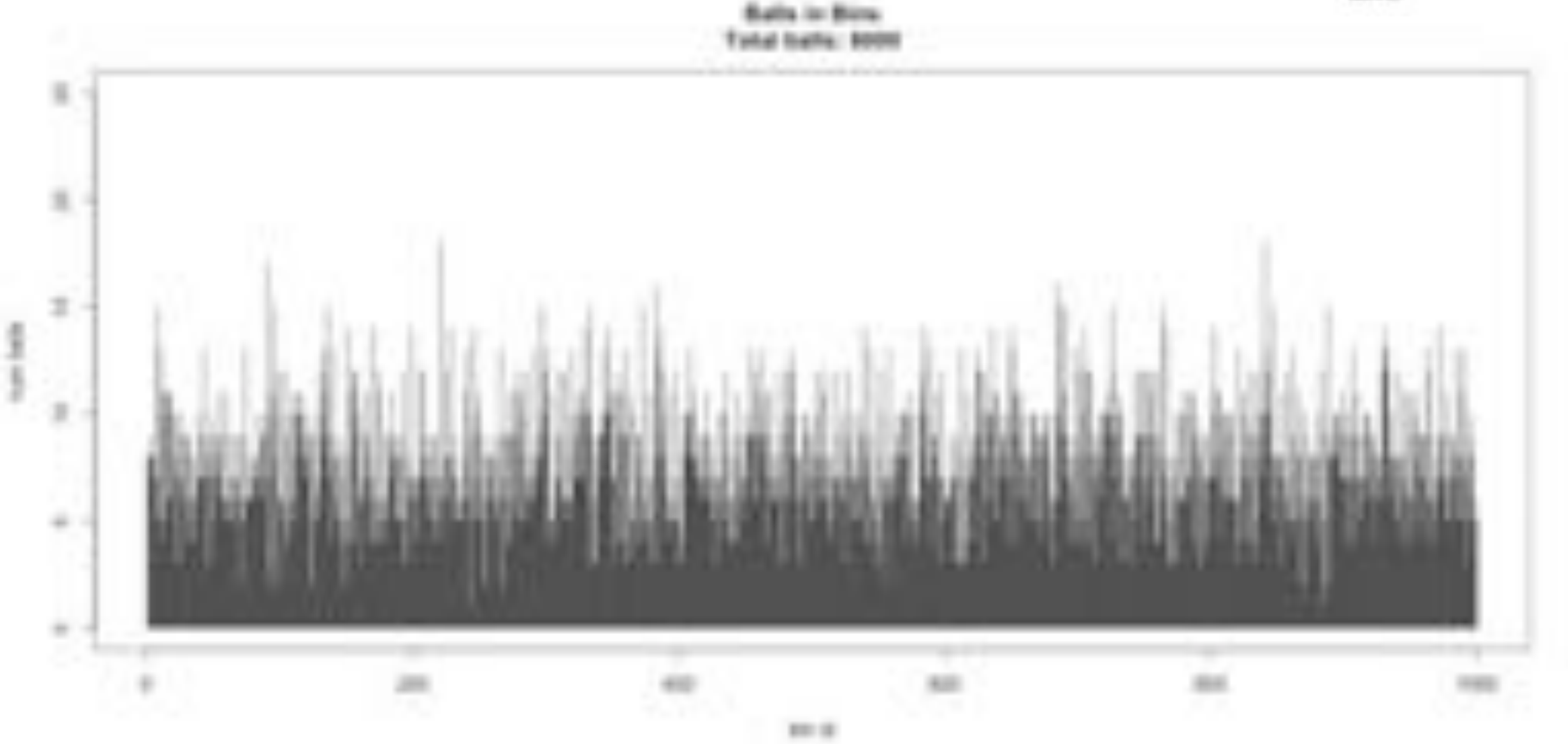
# Balls in Bins 4x





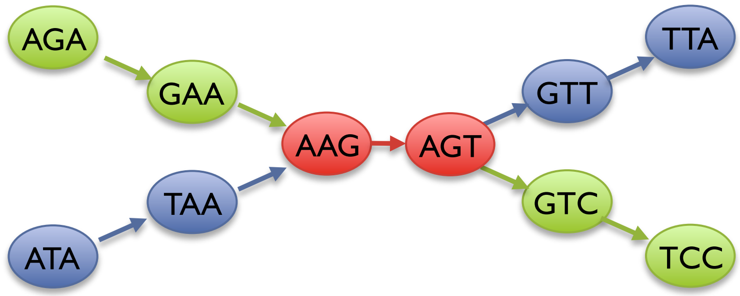
# Balls in Bins 8x





# Two Paradigms for Assembly

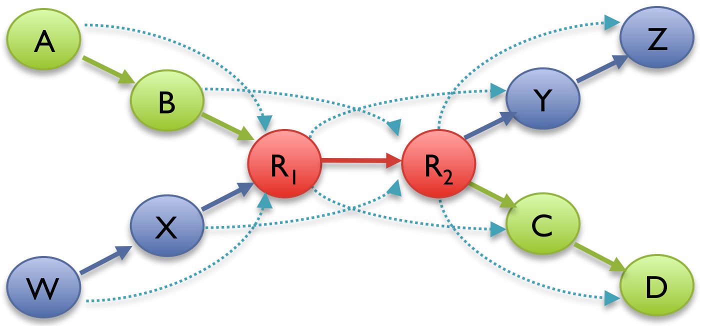
## de Bruijn Graph



Short read assemblers

- Repeats depends on word length
- Read coherency, placements lost
- Robust to high coverage

## Overlap Graph



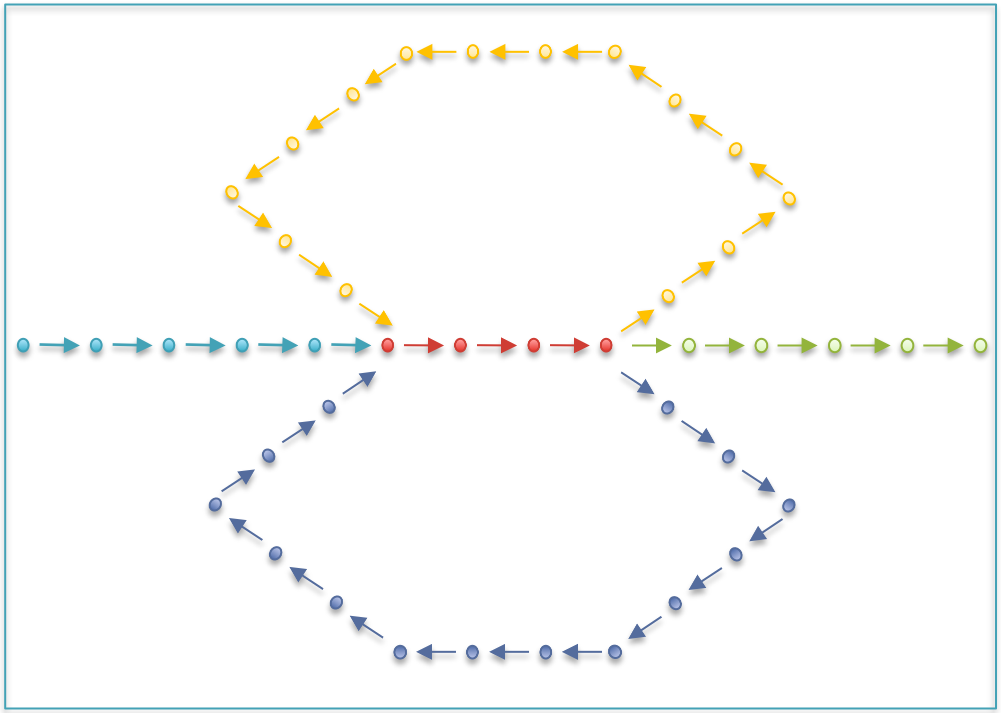
Long read assemblers

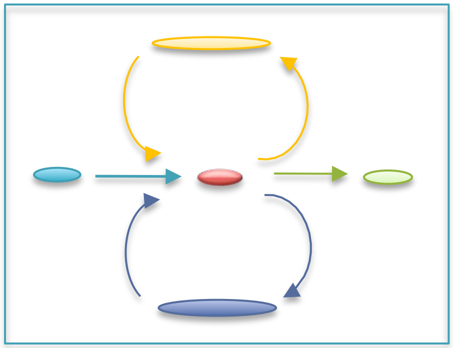
- Repeats depends on read length
- Read coherency, placements kept
- Tangled by high coverage

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

# Unitigging / Unipathing

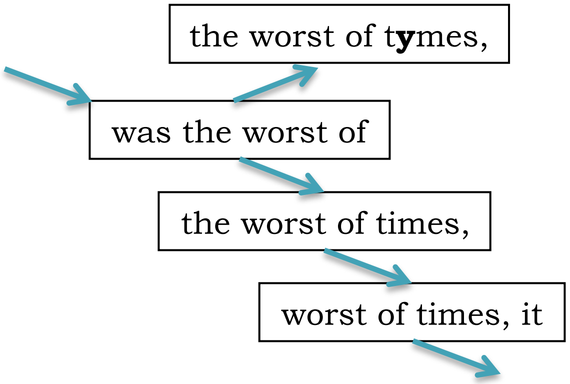
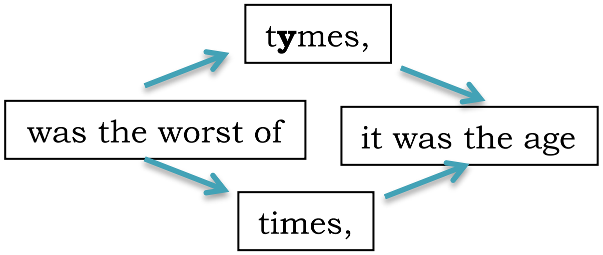
- After simplification and correction, compress graph down to its non-branching initial contigs
  - Aka "unitigs", "unipaths"
  - Unitigs end because of (1) lack of coverage, (2) errors, and (3) repeats



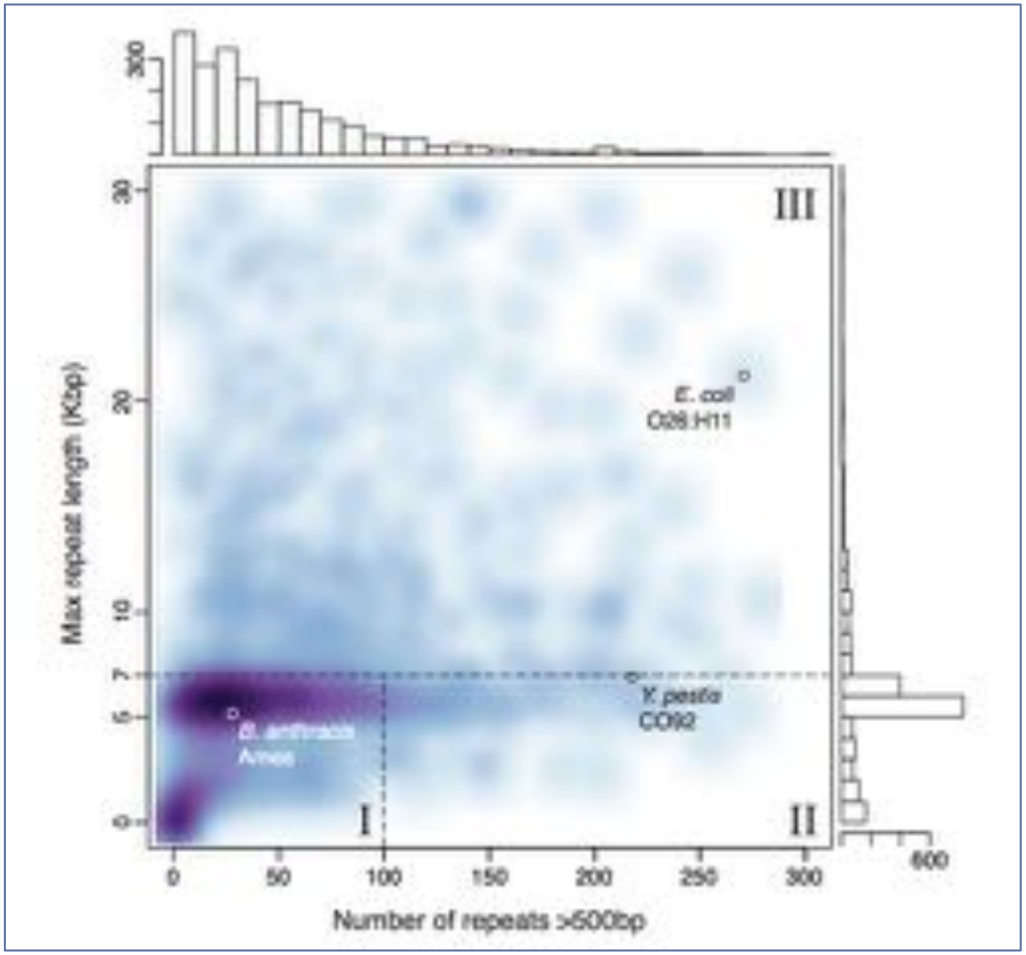


# Errors in the graph

(Chaisson, 2009)

| Clip Tips | Pop Bubbles |
|---|---|
| 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, |

# Repeats and Read Length

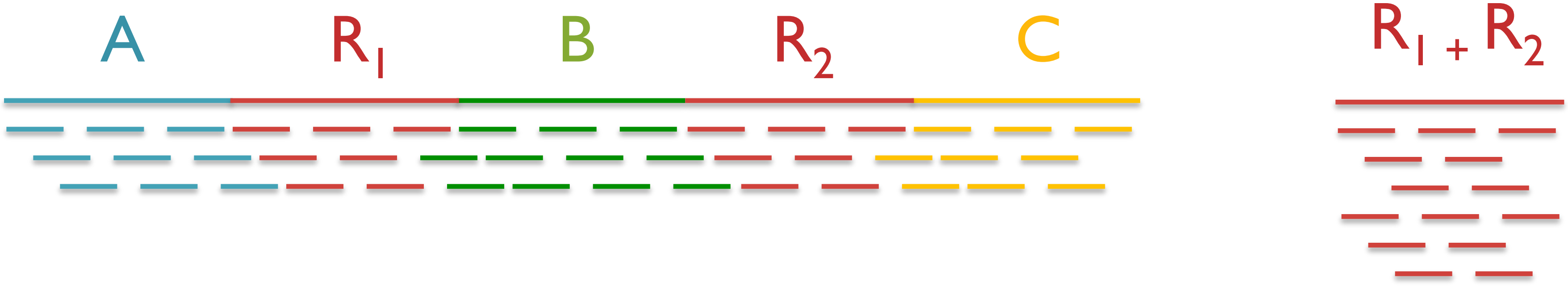


- All microbes have repeats
  - Analyzed all 2,267 available microbial genomes
  - Most are < 7kbp in length and occur in < 100 copies
  - Most repeats are rRNA operons or IS elements

- With enough coverage, contig sizes will be determined by the repeats
  - 5-50kbp contig N50 sizes are common

**Reducing assembly complexity of microbial genomes with single-molecule sequencing**
Koren S. *et al.* (2013) *Under Review.* http://arxiv.org/abs/1304.3752

# Repeats and Coverage Statistics



- If *n* reads are a uniform random sample of the genome of length *G*, we expect $k=n\Delta/G$ reads to start in a region of length $\Delta$.
  - If we see many more reads than k (if the arrival rate is > A) , it is likely to be a collapsed repeat
  - Requires an accurate genome size estimate

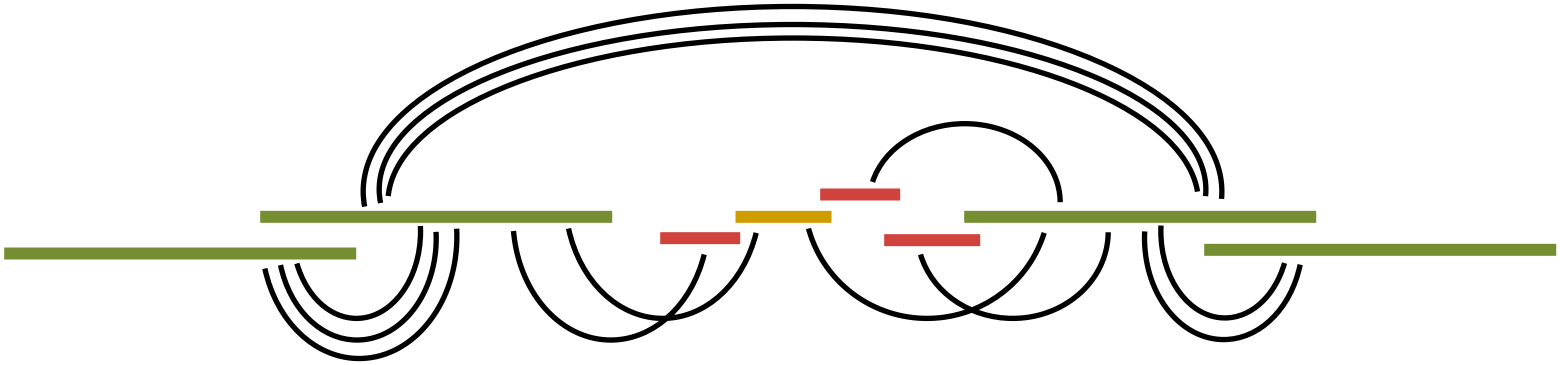
$$\Pr(X-copy)=\binom{n}{k}\left(\frac{X\Delta}{G}\right)^{k}\left(\frac{G-X\Delta}{G}\right)^{n-k}$$

$$A(\Delta,k)=\ln\left(\frac{\Pr(1-copy)}{\Pr(2-copy)}\right)=\ln\left(\frac{\frac{(\Delta n/G)^{k}}{k!}e^{\frac{-\Delta n}{G}}}{\frac{(2\Delta n/G)^{k}}{k!}e^{\frac{-2\Delta n}{G}}}\right)=\frac{n\Delta}{G}-k\ln 2$$

# Scaffolding

- Initial contigs (*aka* unipaths, unitigs) terminate at
  - *Coverage gaps*: especially extreme GC regions
  - *Conflicts*: sequencing errors, repeat boundaries

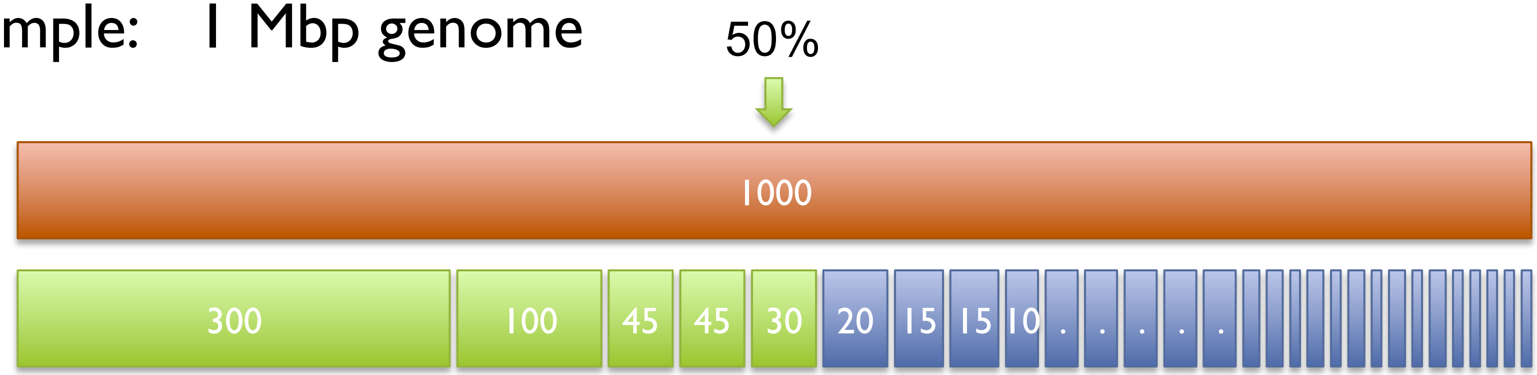
- Iteratively resolve longest, 'most unique' contigs
  - Both overlap graph and de Bruijn assemblers initially collapse repeats into single copies
  - Uniqueness measured by a statistical test on coverage



# N50 size

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

Example: 1 Mbp genome



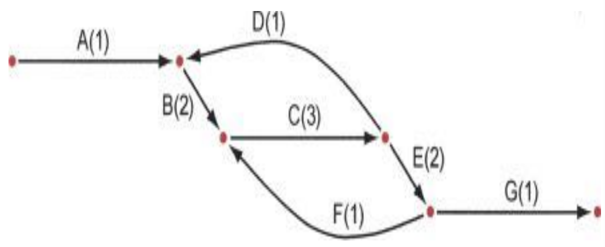
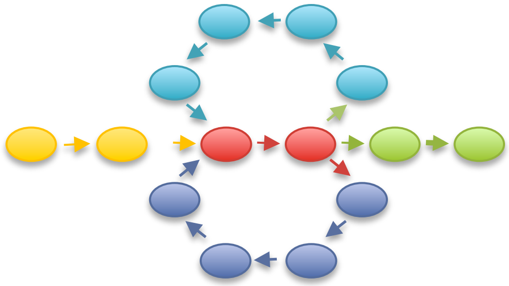
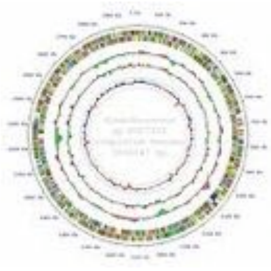
N50 size = 30 kbp

(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 Algorithms

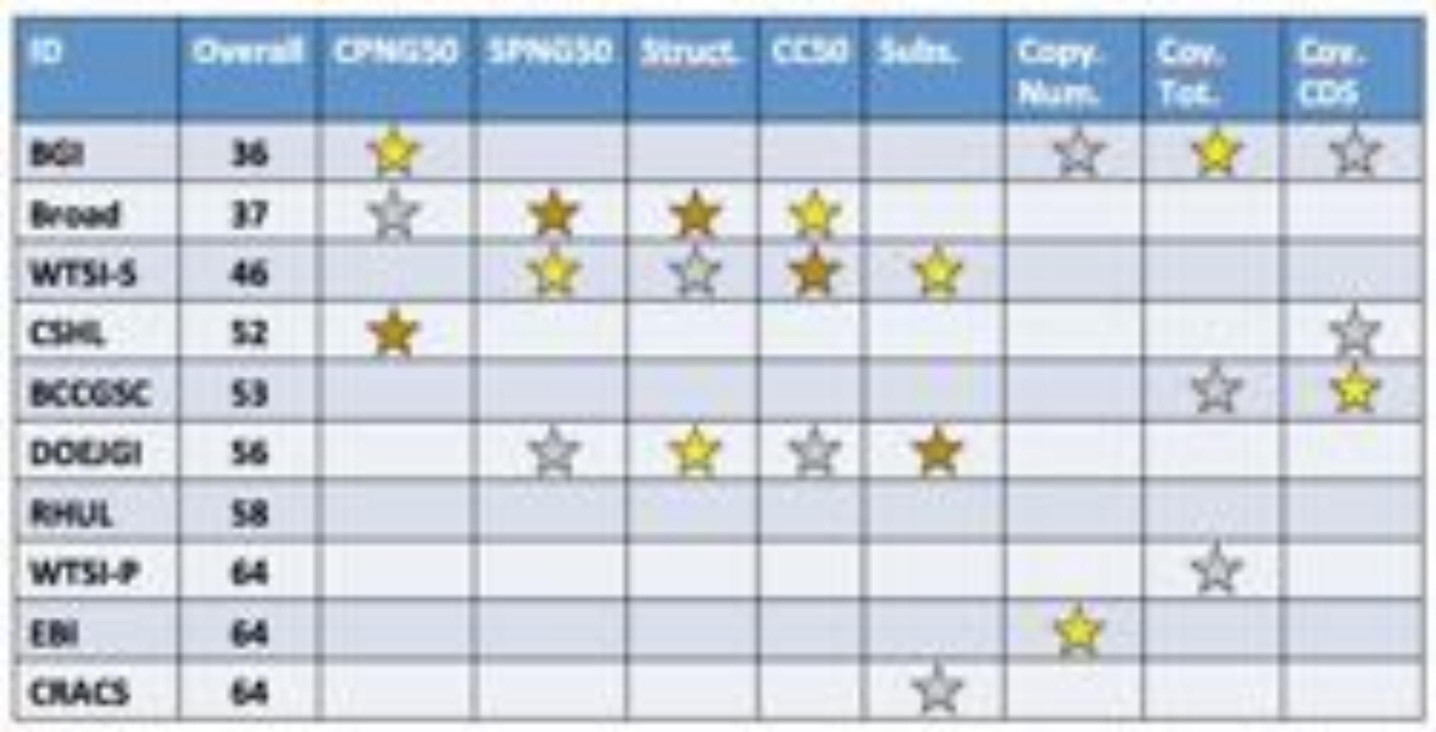
| ALLPATHS-LG | SOAPdenovo | Celera Assembler |
|---|---|---|
|  |  |  |
| Broad's assembler<br>(Gnerre et al. 2011) | BGI's assembler<br>(Li et al. 2010) | JCVI's assembler<br>(Miller et al. 2008) |
| De bruijn graph<br>Short + PacBio (patching) | De bruijn graph<br>Short reads | Overlap graph<br>Medium + Long reads |
| Easy to run if you have compatible libraries | Most flexible, but requires a lot of tuning | Supports Illumina/454/PacBio Hybrid assemblies |
| http://www.broadinstitute.org/software/allpaths-lg/blog/ | http://soap.genomics.org.cn/soapdenovo.html | http://wgs-assembler.sf.net |

# THE ASSEMBLATHON

- Attempt to answer the question:
  **“What makes a good assembly?”**
- Organizers provided simulated sequence data
  - Simulated 100 base pair Illumina reads from simulated diploid organism
  - 41 submissions from 17 groups

**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

# 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 SGA, Celera Assembler, ABySS
  - My recommendation for "typical" short read assembly is to use ALLPATHS
  - See Assemblathon 2 paper for more discussion

**Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species**
Bradman, KR. (2013) *Under Review.* http://arxiv.org/abs/1301.5406

# Hybrid Sequencing

**Illumina**
*Sequencing by Synthesis*

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

**Pacific Biosciences**
*SMRT Sequencing*

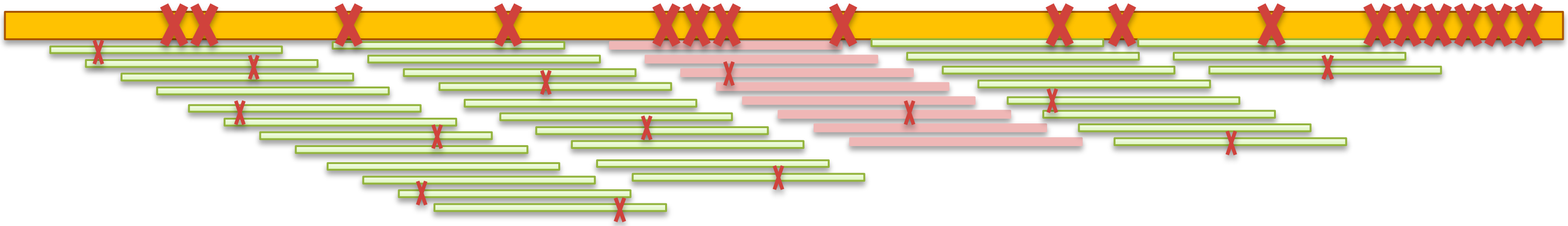
Lower throughput (600Mbp/day)
Lower accuracy (~85%)
Long reads (2-25kbp)

# PacBio Error Correction

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

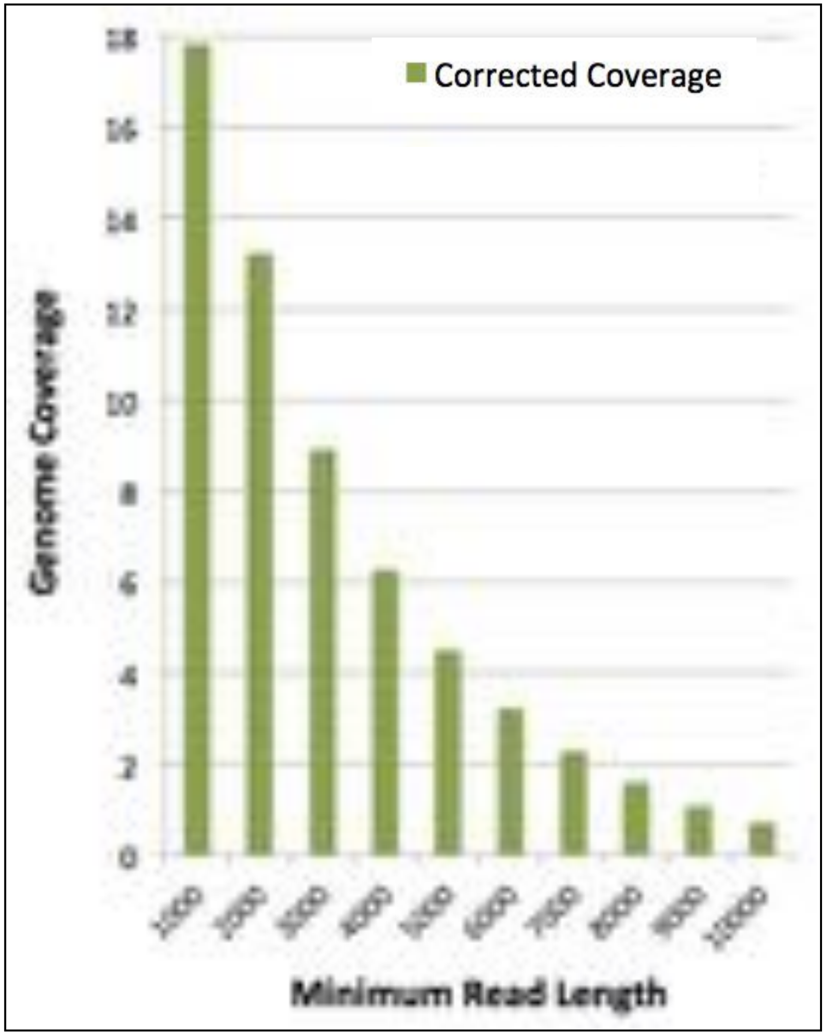
2. Error corrected reads can be easily assembled, aligned



**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 |
| PBeCR Reads<br>7x @ 3500 ** MiSeq for correction | 50,995 |
| PBeCR + Illumina Shred<br>7x @ 3500 ** MiSeq for correction<br>5x @ 3000bp shred | 59,695 |

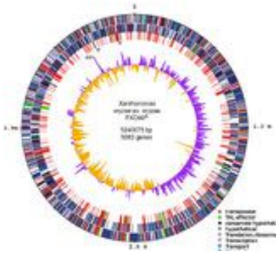


In collaboration with McCombie & Ware labs @ CSHL

# Other Resources

| Resource | URL | Description |
|---|---|---|
| Google | http://www.google.com | Internet Search |
| Google Scholar | http://scholar.google.com/ | Literature Searches |
| SeqAnswers | http://seqanswers.com/ | Bioinformatics Forum |
| Wikipedia | http://www.wikipedia.org/ | Overview on anything |
| | | |
| Clovr | http://clovr.org/ | Automated Sequence Analysis |
| Circos | http://circos.ca/ | Circular Genome Plots |
| Galaxy | http://usegalaxy.org | Sequence Analysis in the clouds |
| GraphViz | http://www.graphviz.org/ | Graph Visualization |
| IGV | http://www.broadinstitute.org/igv/ | Read Mapping Viz |
| R | http://www.r-project.org/ | Stats & Visualizations |
| | | |
| Schatz Lab | http://schatzlab.cshl.edu/teaching/ | Exercises and Lectures |

# 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, 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

# Acknowledgements

Schatz Lab
Giuseppe Narzisi
Shoshana Marcus
James Gurtowski
Alejandro Wences
Hayan Lee
Rob Aboukhalil
Mitch Bekritsky
Charles Underwood
Rushil Gupta
Avijit Gupta
Shishir Horane
Deepak Nettem
Varrun Ramani
Kelly Moffat
Eric Biggers
Aspyn Palatnick

CSHL
Hannon Lab
Gingeras Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Ware Lab
Wigler Lab

IT Department

NBACC
Adam Phillippy
Sergey Koren

SFARI
SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE

National Human
Genome Research
Institute

U.S. DEPARTMENT OF
ENERGY

NSF

# Thank You

http://schatzlab.cshl.edu

@mike_schatz