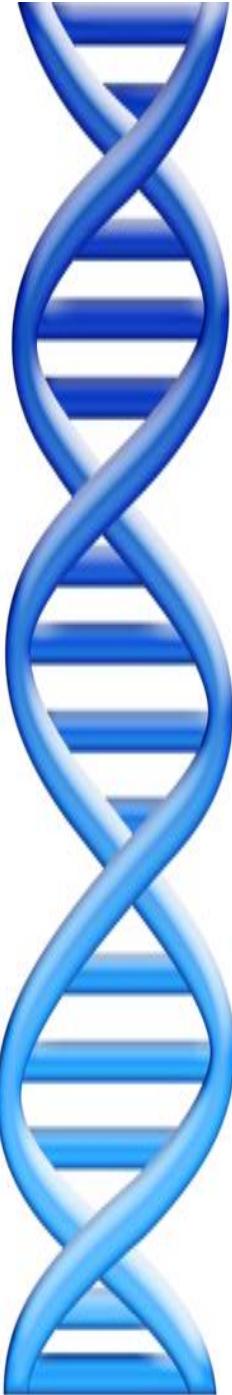


# Graphs and Genomes

Michael Schatz

July 27, 2012  
CSHL Undergraduate Research Program





# Outline

- I. Graph Searching
2. Assembly by analogy
3. Genome Assembly

# Biological Networks

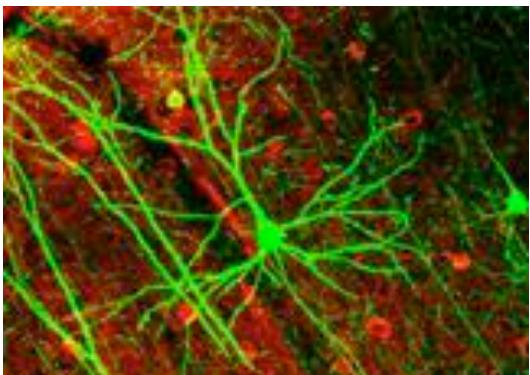
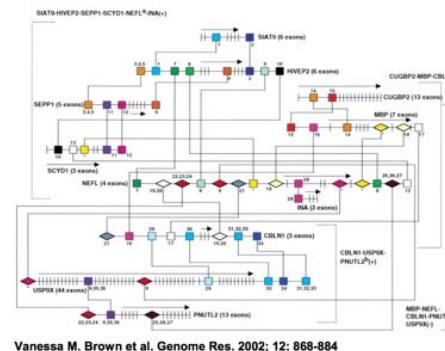
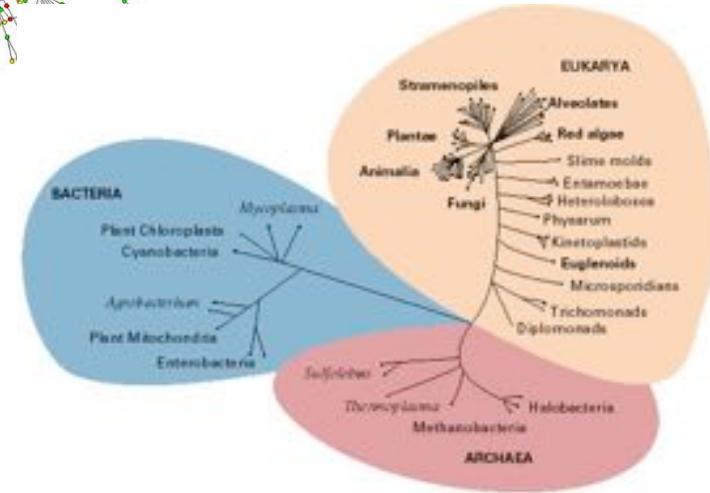
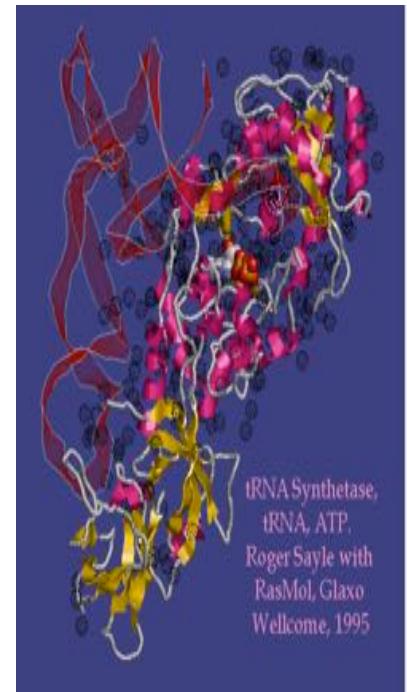
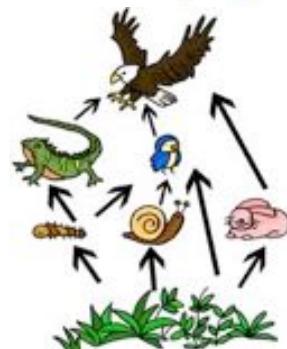
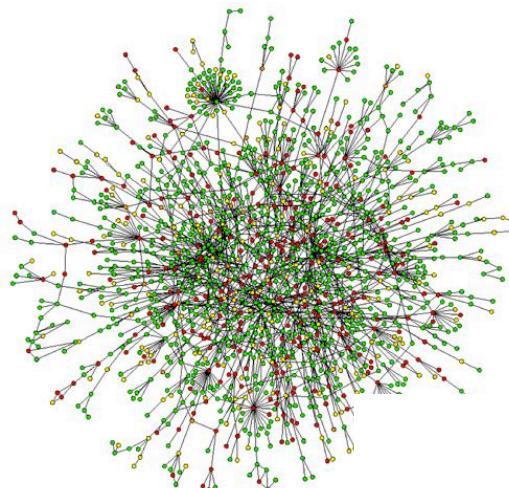
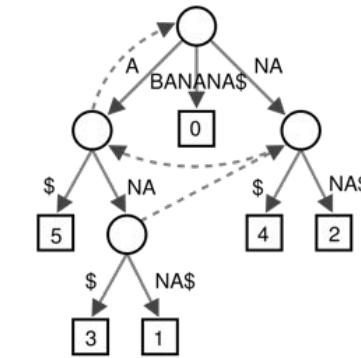
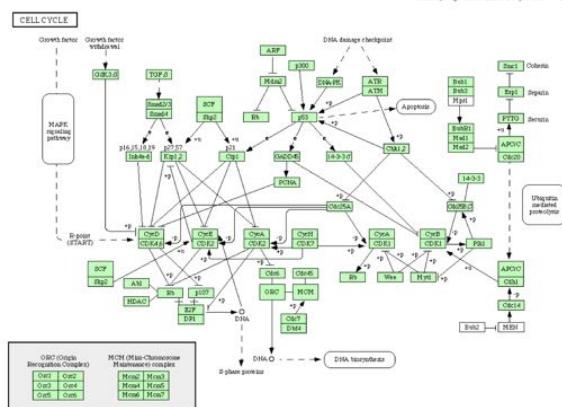


Figure 5 Putative regulatory elements shared between groups of correlated and anticorrelated genes

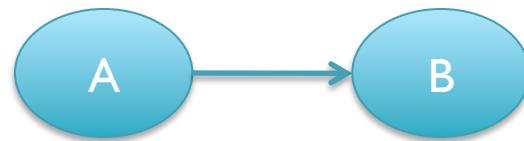


Vanessa M. Brown et al. Genome Res. 2002; 12: 868-884

Cold Spring Harbor Laboratory Press



# Graphs

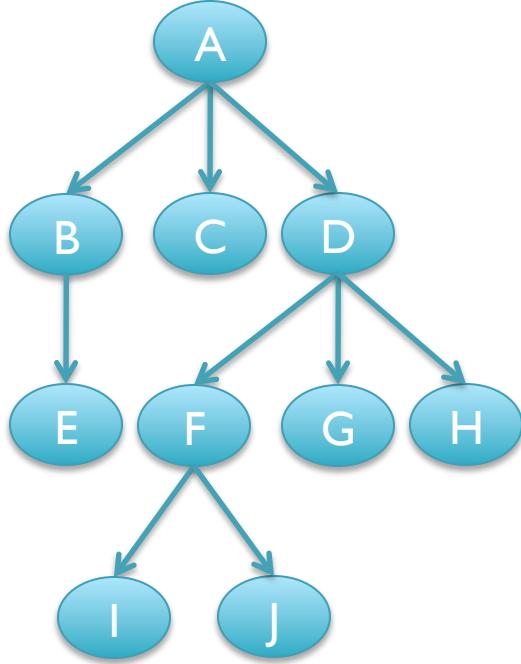


- **Nodes**
  - People, Proteins, Genes, Neurons, Sequences, Numbers, ...
- **Edges**
  - A is connected to B
  - A is related to B
  - A regulates B
  - A precedes B
  - A interacts with B
  - A activates B
  - ...

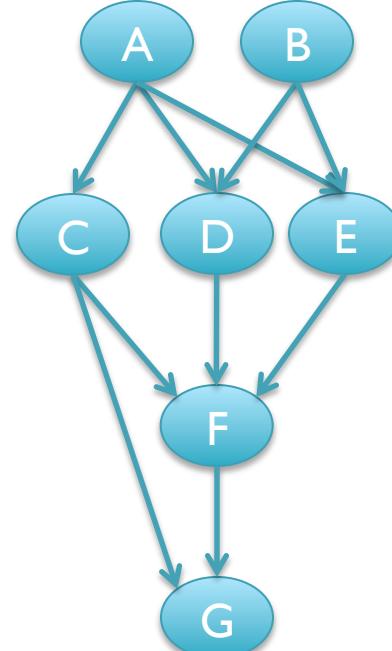
# Graph Types



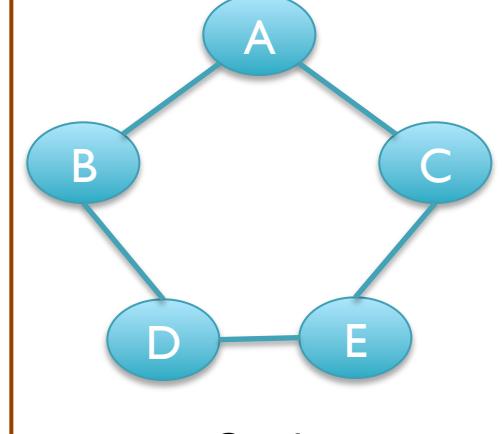
List



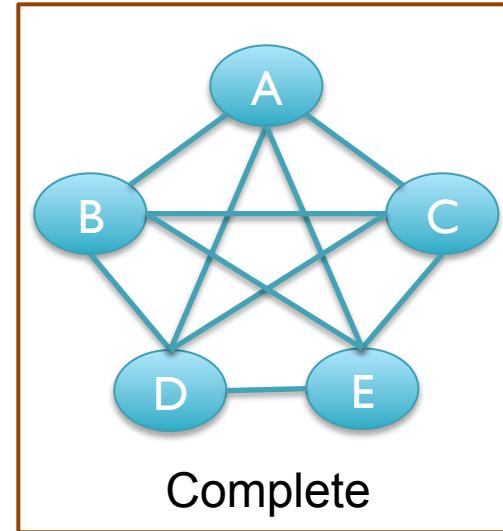
Tree



Directed  
Acyclic  
Graph



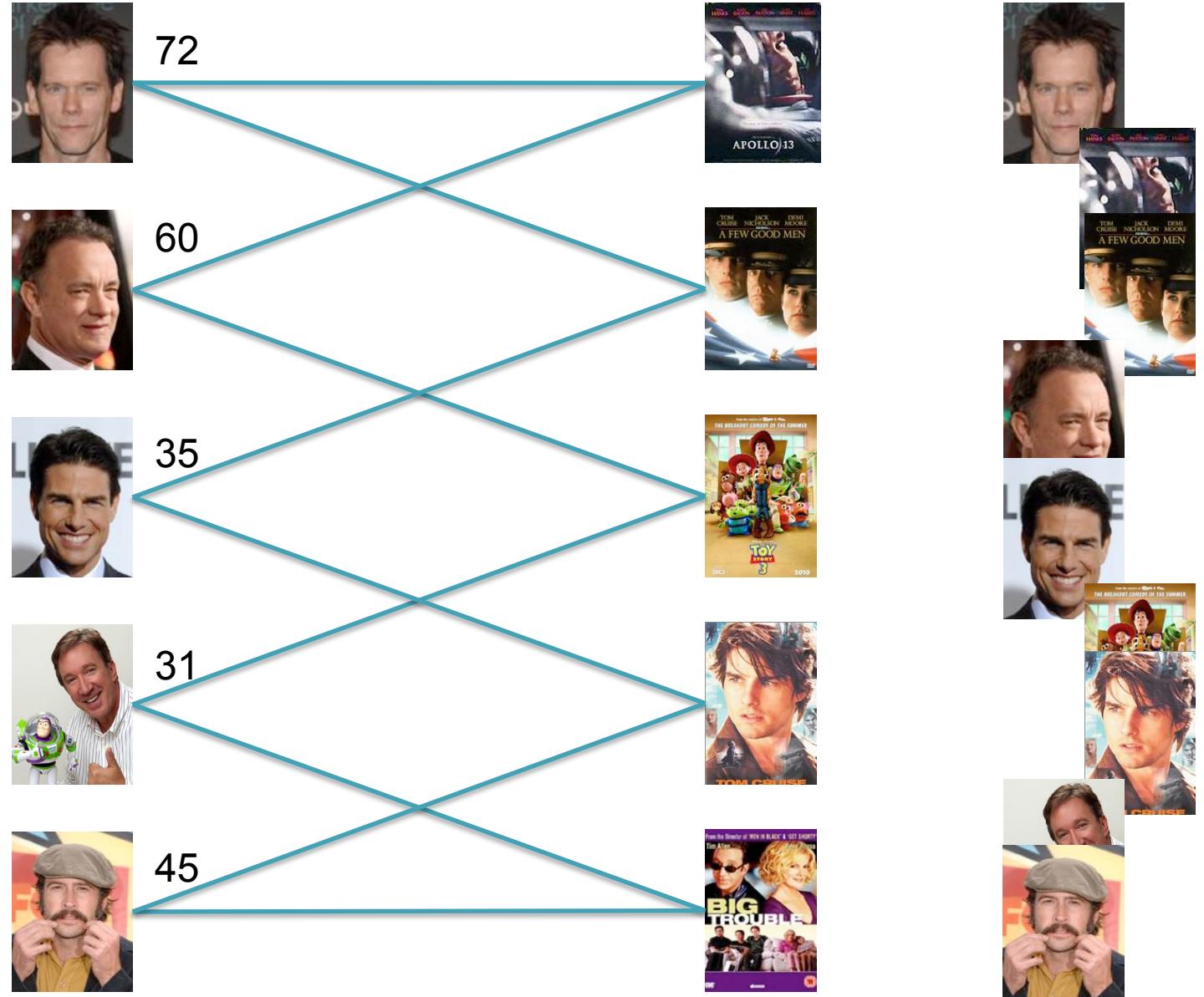
Cycle



Complete

# Kevin Bacon and Bipartite Graphs

Find the **shortest** path from Kevin Bacon to Jason Lee



# BFS and TSP

- BFS computes the shortest path between a pair of nodes in  $O(|E|) = O(|N|^2)$
- What if we wanted to compute the shortest path visiting every node once?
  - Traveling Salesman Problem

ABDCA:  $4+2+5+3 = 14$

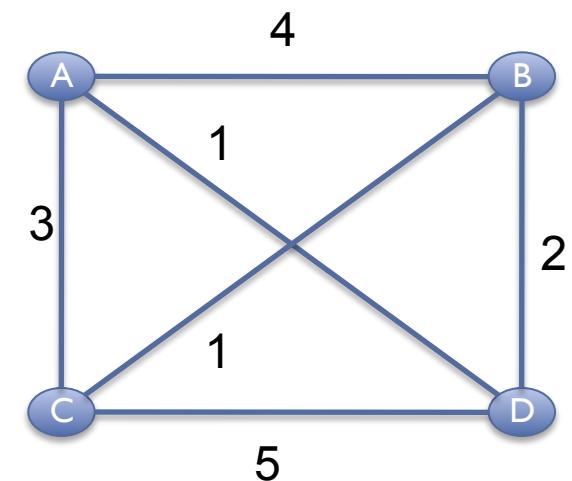
ACDBA:  $3+5+2+4 = 14^*$

ABCDA:  $4+1+5+1 = 11$

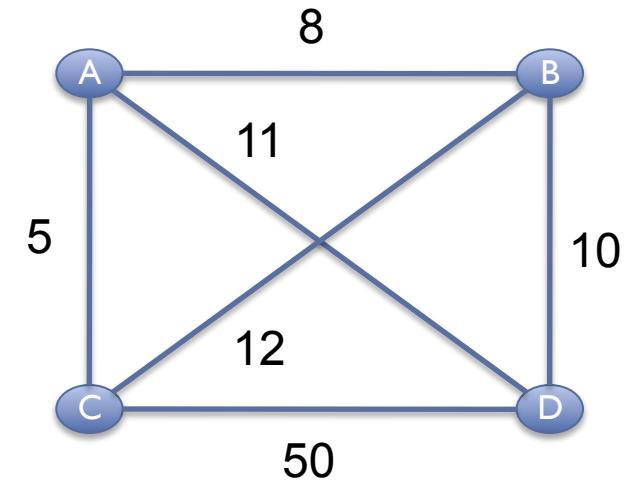
ADCBA:  $1+5+1+4 = 11^*$

ACBDA:  $3+1+2+1 = 7$

ADBKA:  $1+2+1+3= 7 *$



# Greedy Search



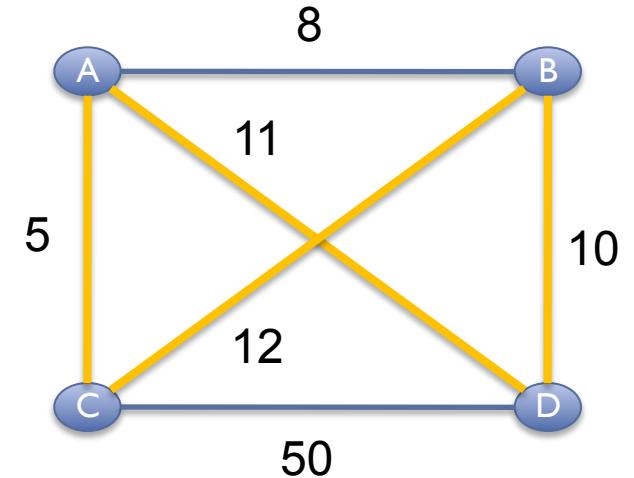
# Greedy Search

## **Greedy Search**

```
cur=graph.randNode()
```

```
while (!done)
```

```
    next=cur.getNextClosest()
```



Greedy:  $ABDCA = 5+8+10+50= 73$

Optimal:  $ACBDA = 5+11+10+12 = 38$

Greedy finds the global optimum only when

1. Greedy Choice: Local is correct without reconsideration
2. Optimal Substructure: Problem can be split into subproblems

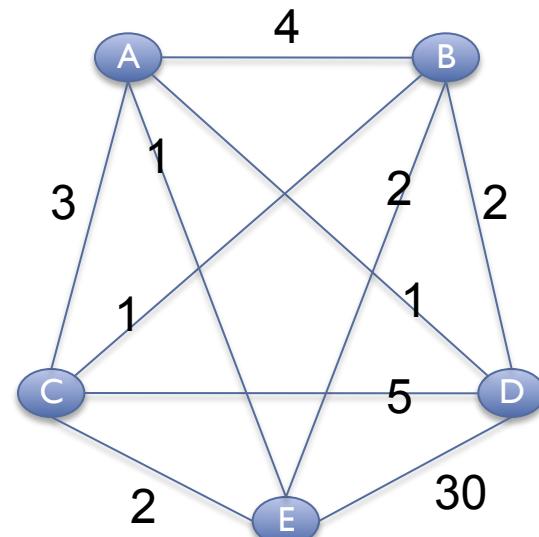
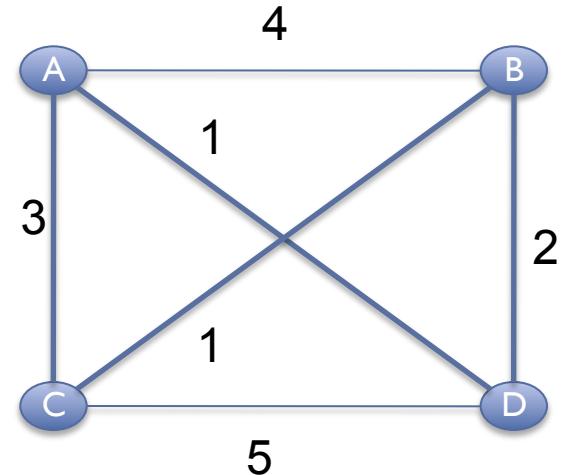
Optimal Greedy: Making change with the fewest number of coins

# TSP Complexity

- No fast solution
  - Knowing optimal tour through  $n$  cities doesn't seem to help much for  $n+1$  cities

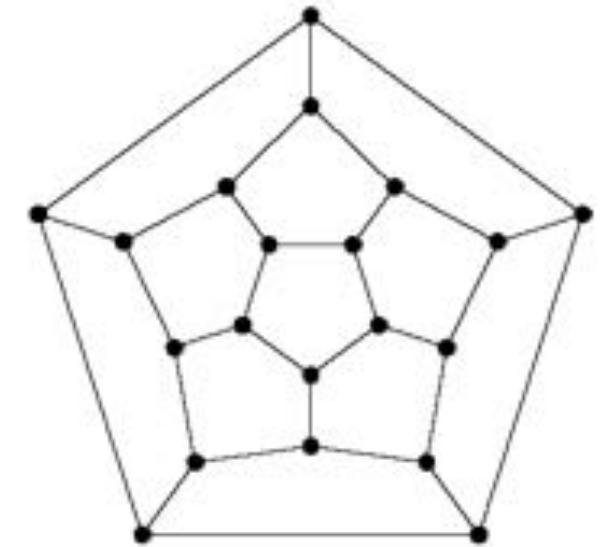
[How many possible tours for  $n$  cities?]

- Extensive searching is the only provably correct algorithm
  - Brute Force:  $O(n!)$ 
    - ~20 cities max
    - $20! = 2.4 \times 10^{18}$
  - Branch-and-Bound can sometimes help



# TSP and NP-complete

- TSP is one of many extremely hard problems of the class NP-complete
  - Extensive searching is the only way to find an exact solution
  - Often have to settle for approx. solution
- **WARNING:** Many biological problems are in this class
  - Find a tour the visits every node once (Genome Assembly)
  - Find the smallest set of vertices covering the edges (Essential Genes)
  - Find the largest clique in the graph (Protein Complexes)
  - Find the highest mutual information encoding scheme (Neurobiology)
  - Find the best set of moves in tetris
  - ...
  - [http://en.wikipedia.org/wiki/List\\_of\\_NP-complete\\_problems](http://en.wikipedia.org/wiki/List_of_NP-complete_problems)





# Outline

- I. Graph Searching
2. Assembly by analogy
3. Genome Assembly

# 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?
  - $5 \text{ copies} \times 138,656 \text{ words} / 5 \text{ words per fragment} = 138k \text{ fragments}$
  - The short fragments from every copy are mixed together
  - Some fragments are identical

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

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 sequence reconstruction 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

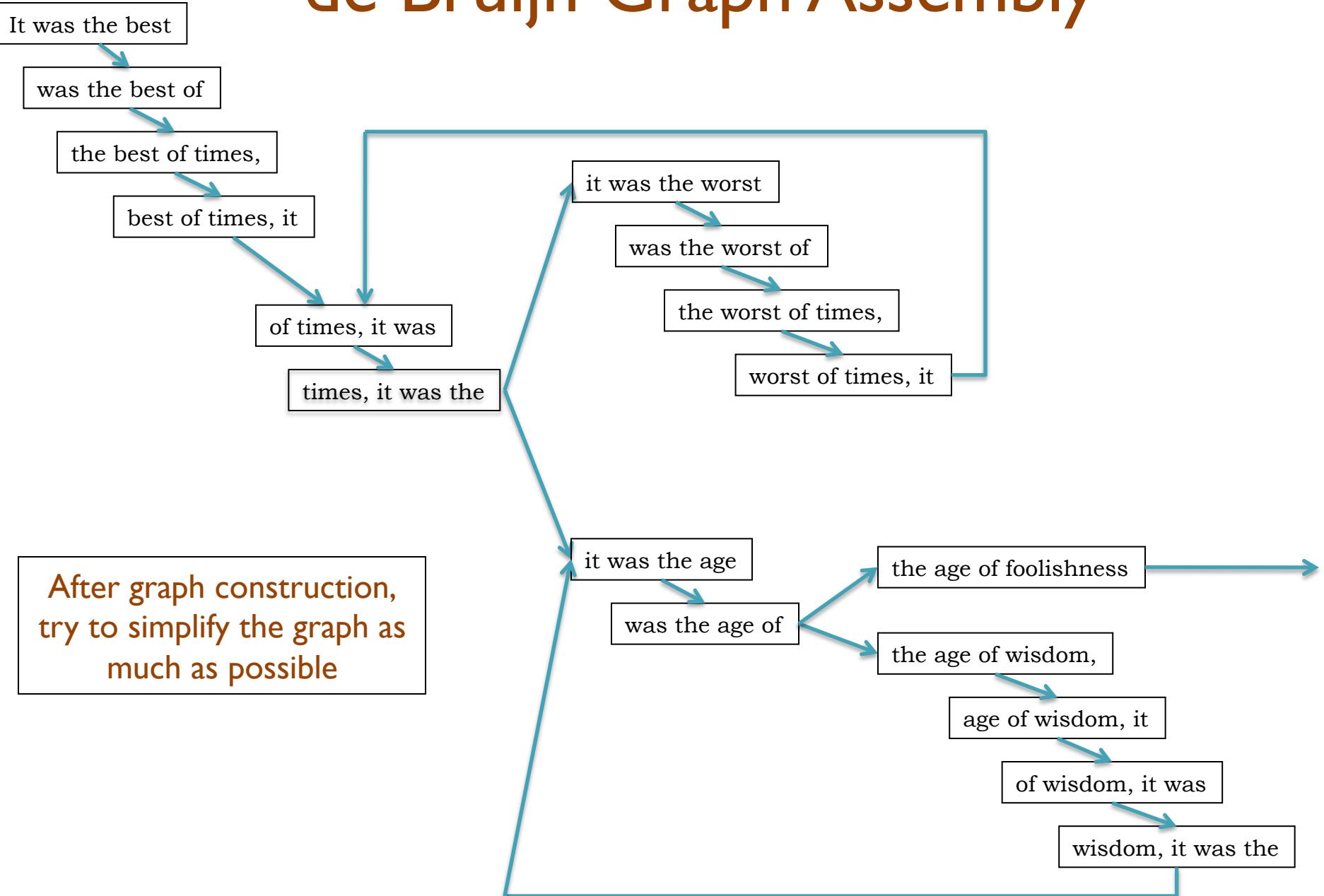
- Locally constructed graph reveals the global sequence structure
  - Overlaps between sequences implicitly computed

de Bruijn, 1946

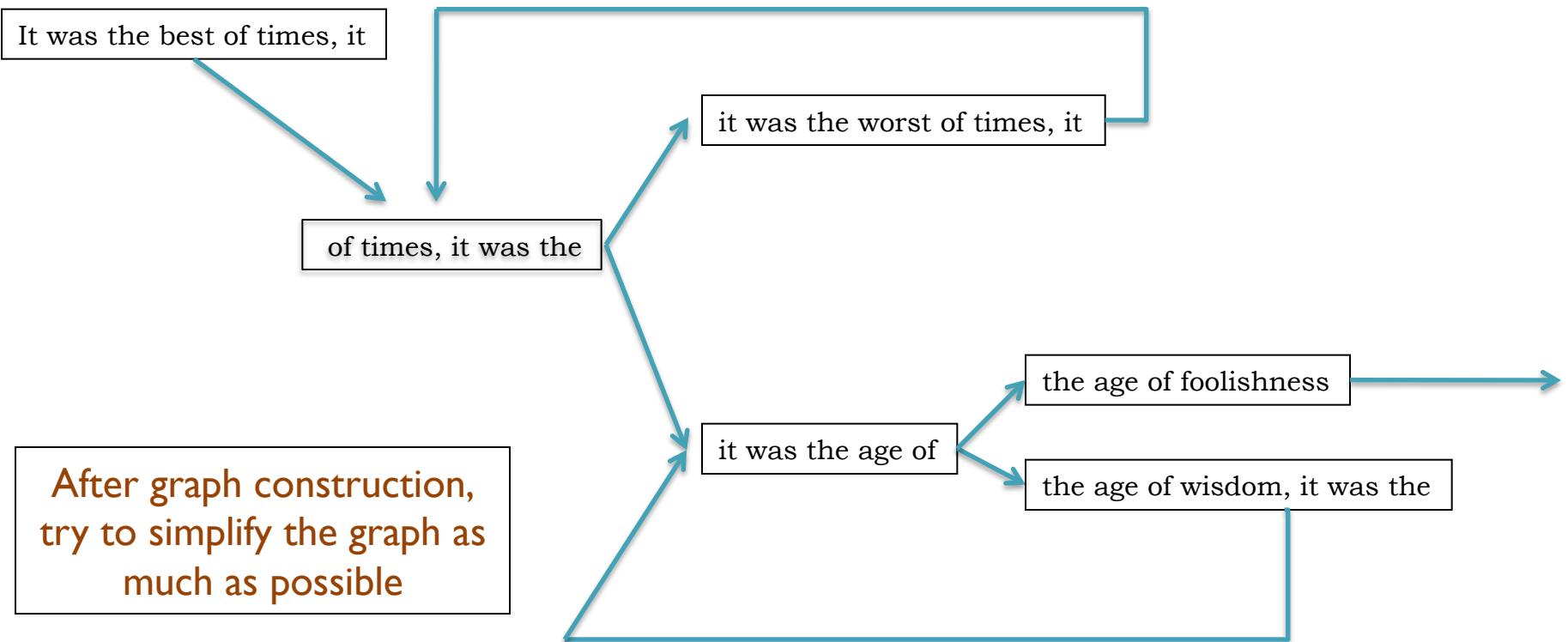
Idury and Waterman, 1995

Pevzner, Tang, Waterman, 2001

# de Bruijn Graph Assembly



# de Bruijn Graph Assembly





# Outline

- I. Genome Assembly by Analogy
2. Graph Searching
3. Genome Assembly

# Milestones in Genome Assembly

Nature Vol. 265 February 24 1977 487

## articles

### Nucleotide sequence of bacteriophage $\Phi$ X174 DNA

F. Sanger, G. M. Air<sup>1</sup>, B. G. Barrell, N. L. Brown<sup>1</sup>, A. R. Coulson, J. C. Fiddes,  
C. A. Hutchison III<sup>1</sup>, P. M. Slocombe<sup>1</sup> & M. Smith<sup>2</sup>

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

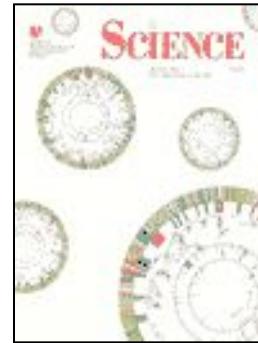
A DNA sequence for the genome of bacteriophage  $\Phi$ X174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the various proteins known to be produced by the organism, including initiation and termination sites for the proteins and RNAs. Two pairs of genes are coded by the same region of DNA using different reading frames.

The genome of bacteriophage  $\Phi$ X174 is a single-stranded, circular molecule containing approximately 5,375 nucleotides and nine known proteins. The order of these genes, as determined by genetic techniques<sup>1-3</sup>, is A-B-C-D-E-F-F-G-H. Genes F, G and H code for structural proteins, while genes A, B, C and D code for enzymes. Two pairs of genes are coded by the same region of DNA using different reading frames.

1977. Sanger et al.

1<sup>st</sup> Complete Organism

5375 bp



2000. Myers et al.  
1<sup>st</sup> Large WGS Assembly.  
Celera Assembler. 116 Mbp

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

2010. Li et al.  
1<sup>st</sup> Large SGS Assembly.  
SOAPdenovo 2.2 Gbp

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

# Current Applications

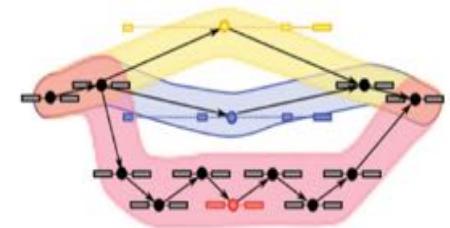
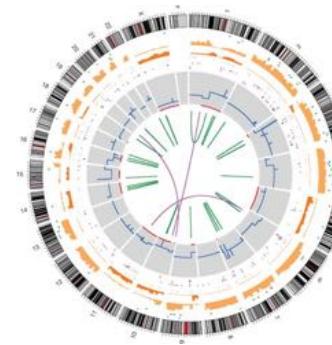
- Novel genomes



- Metagenomes

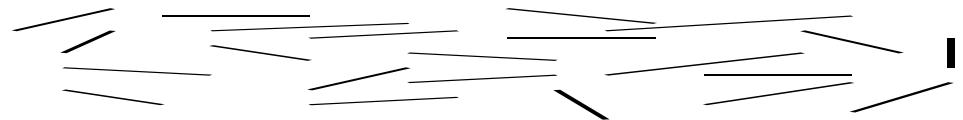


- Sequencing assays
  - Structural variations
  - Transcript assembly
  - ...



# Assembling a Genome

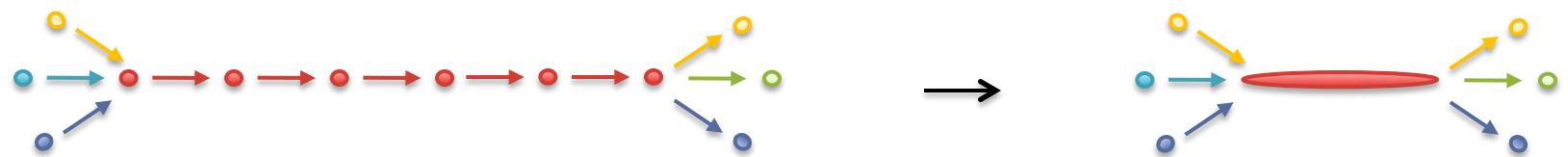
## 1. Shear & Sequence DNA



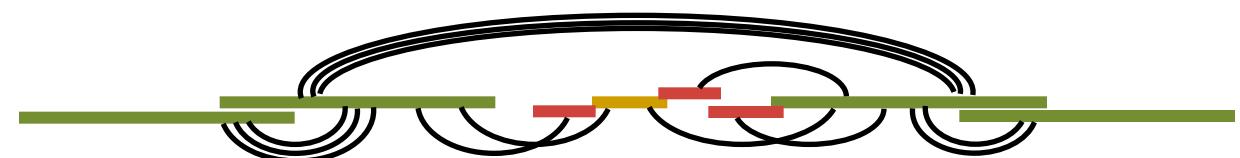
## 2. Construct assembly graph from overlapping reads

...AGCCTAGACCTACA**GGATGCGCGACACGT**  
**GGATGCGCGACACGT**CGCATATCCGGT...

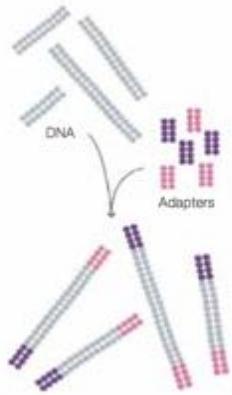
## 3. Simplify assembly graph



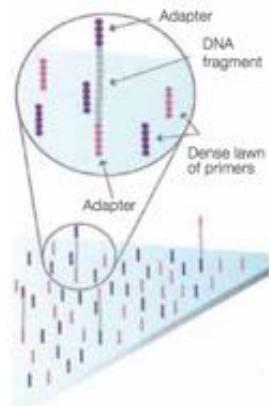
## 4. Detangle graph with long reads, mates, and other links



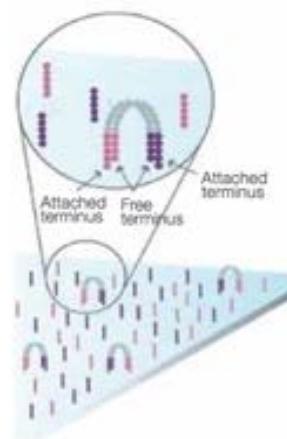
# Illumina Sequencing by Synthesis



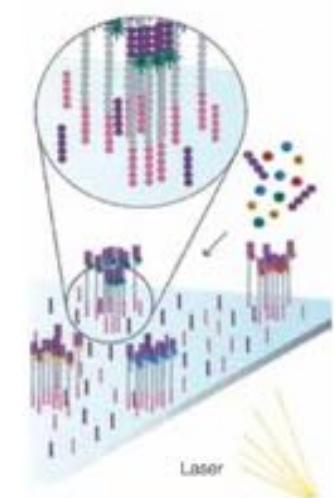
1. Prepare



2. Attach



3. Amplify



4. Image



5. Basecall

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

[http://www.illumina.com/documents/products/techspotlights/techspotlight\\_sequencing.pdf](http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf)

# Paired-end and Mate-pairs

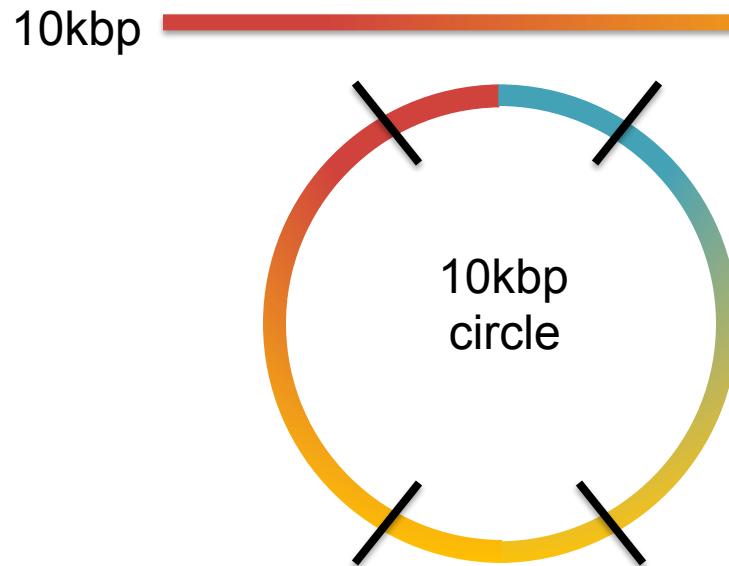
## Paired-end sequencing

- 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



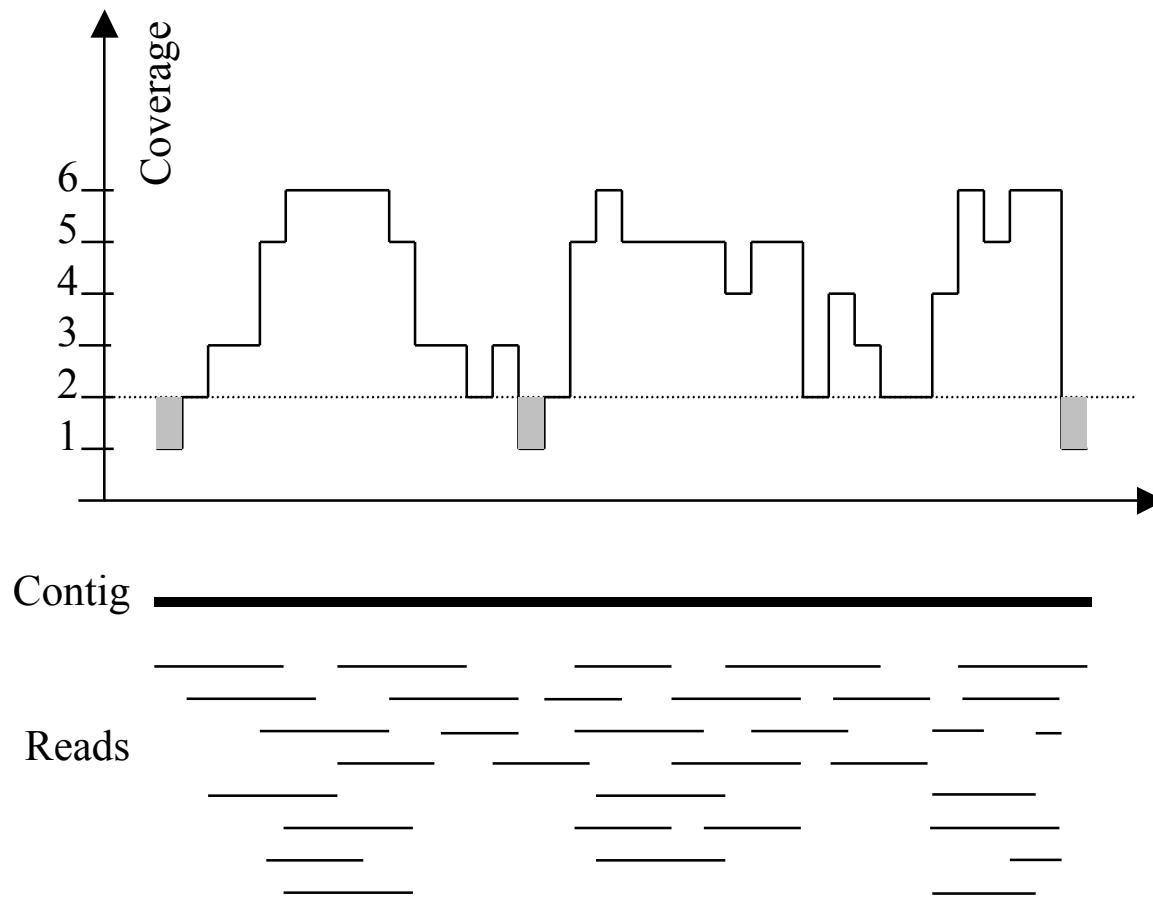
2x100 @ ~10kbp (outies)



2x100 @ 300bp (innies)

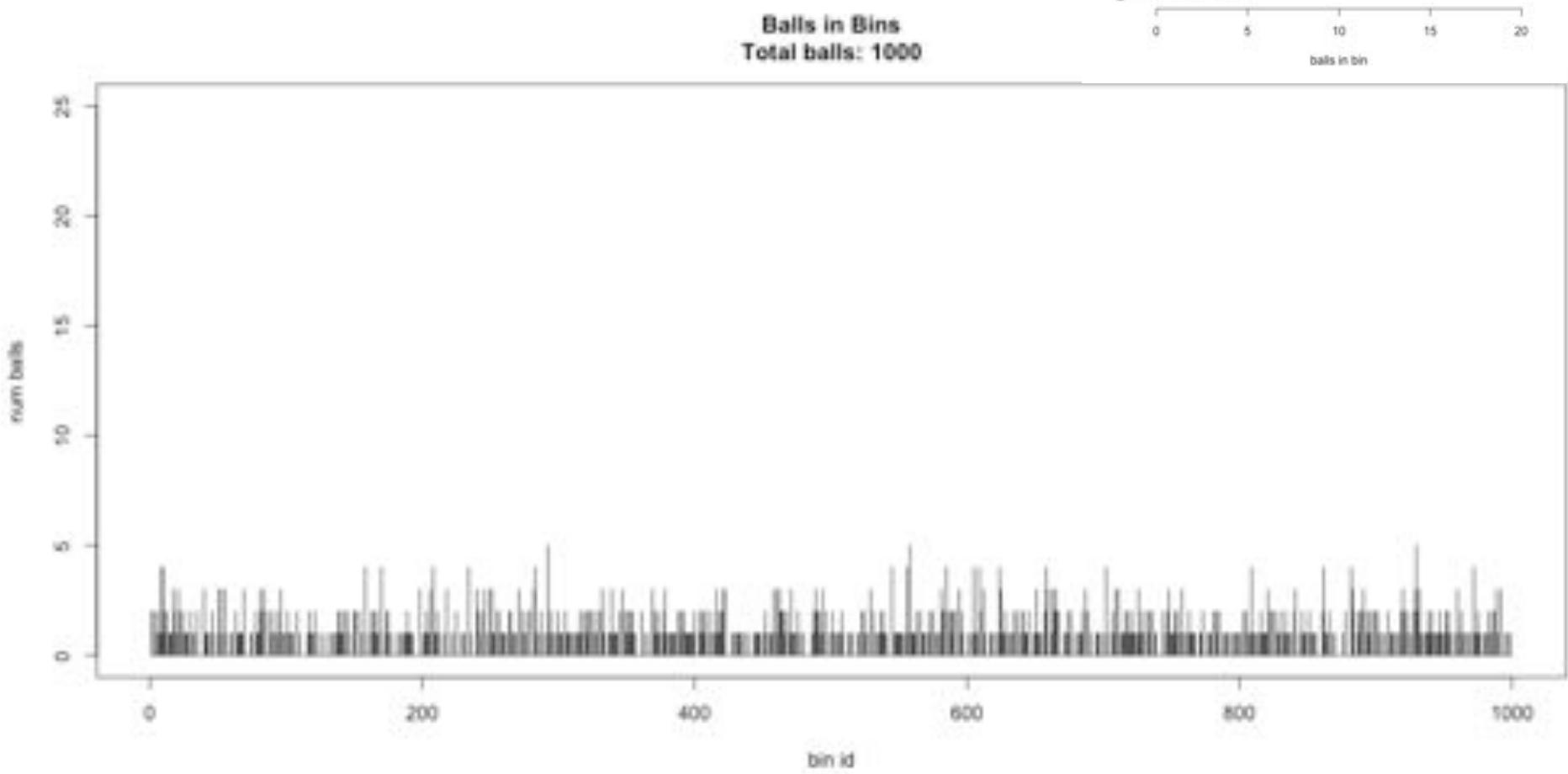


# Typical contig coverage



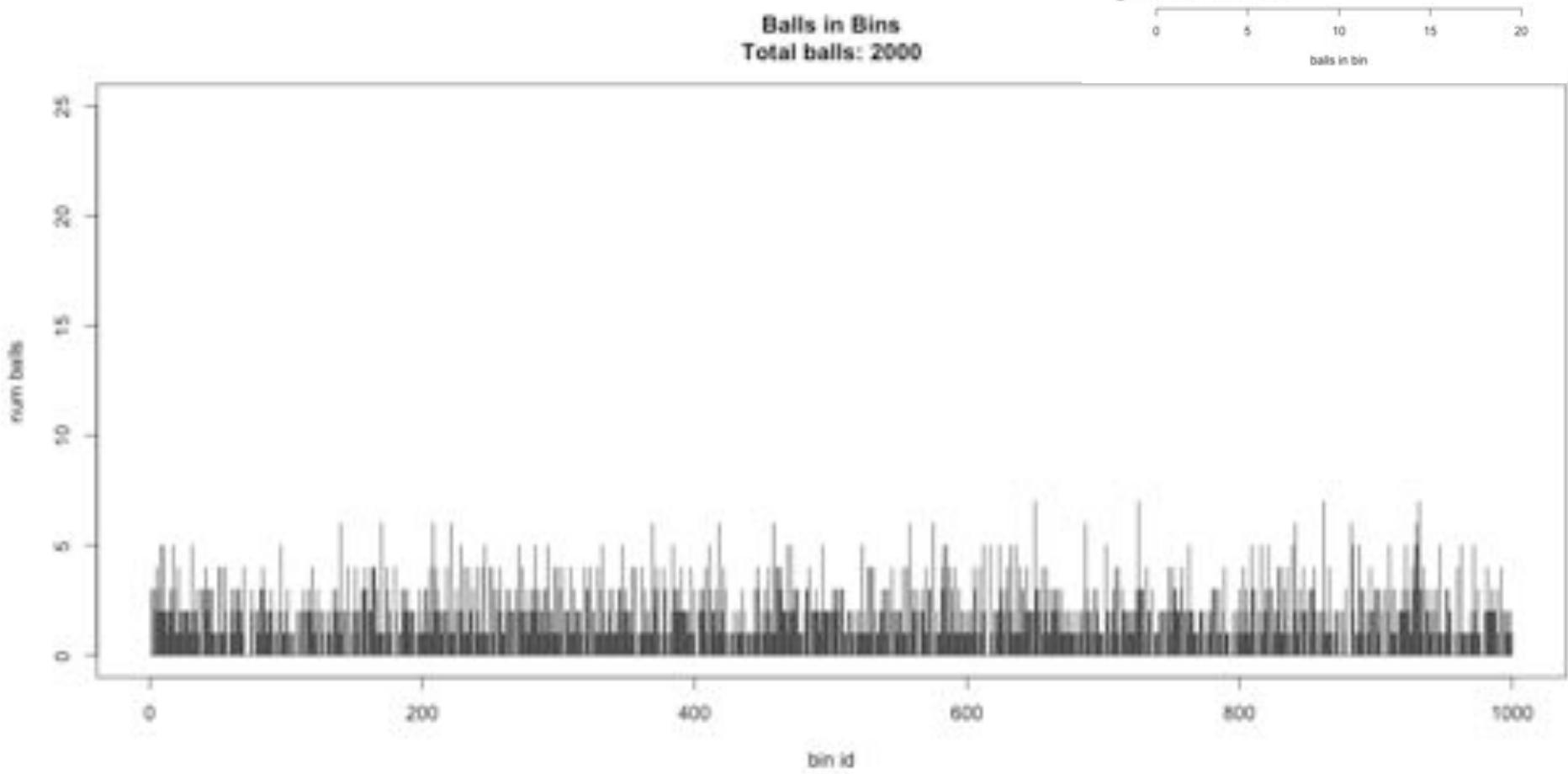
Imagine raindrops on a sidewalk

# Balls in Bins IX



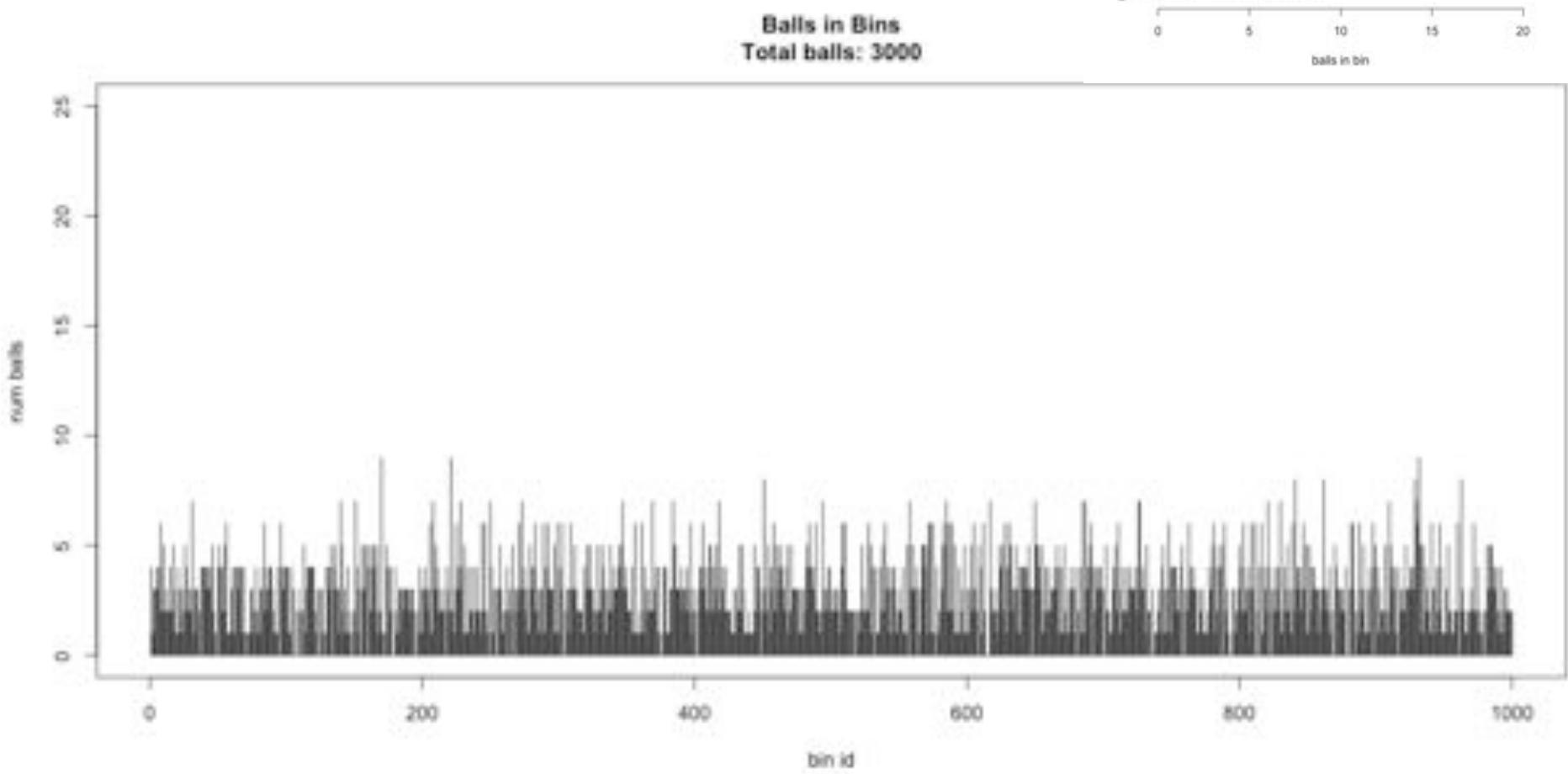
Histogram of balls in each bin  
Total balls: 1000

# Balls in Bins 2x

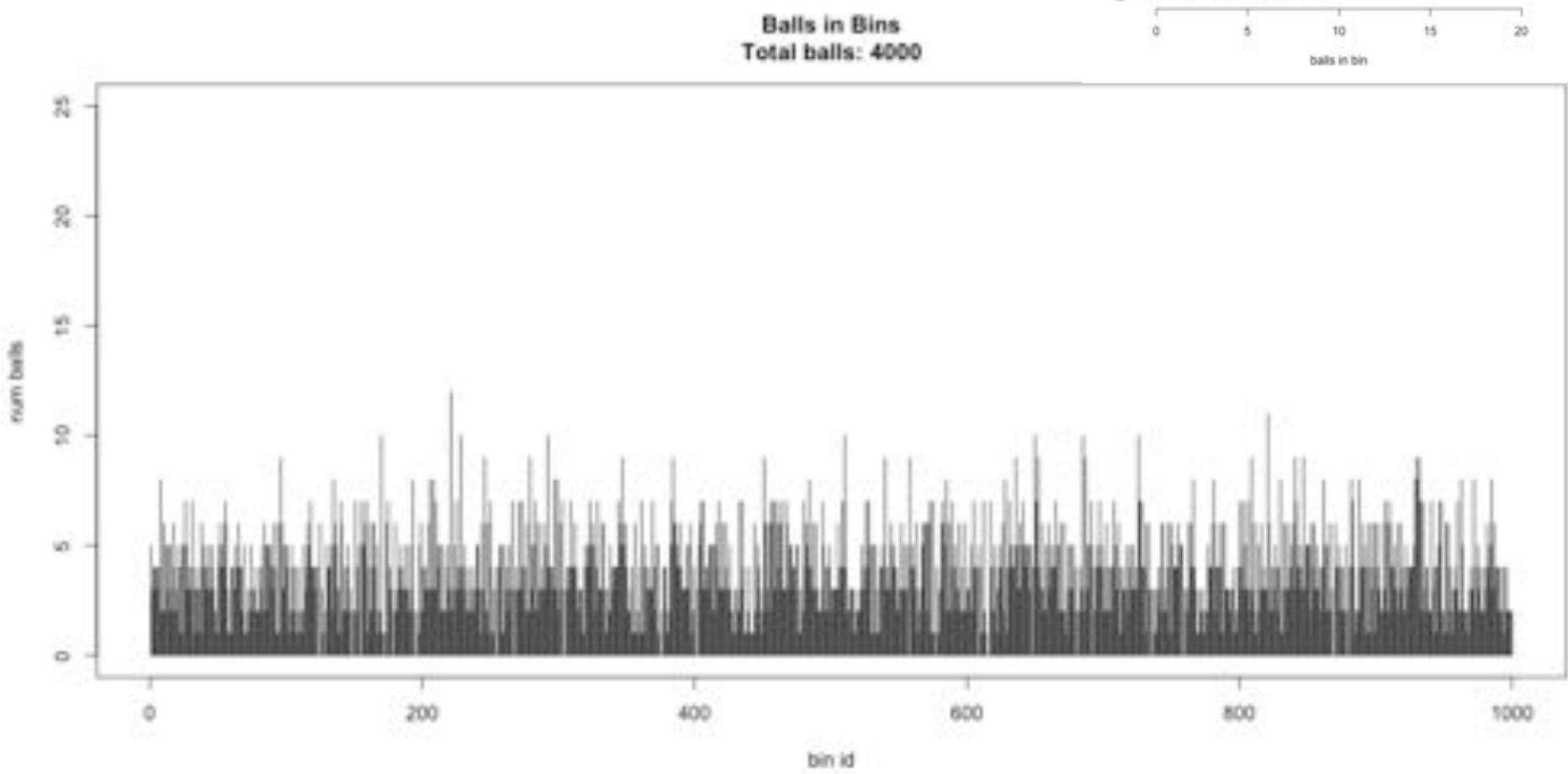


Histogram of balls in each bin  
Total balls: 2000

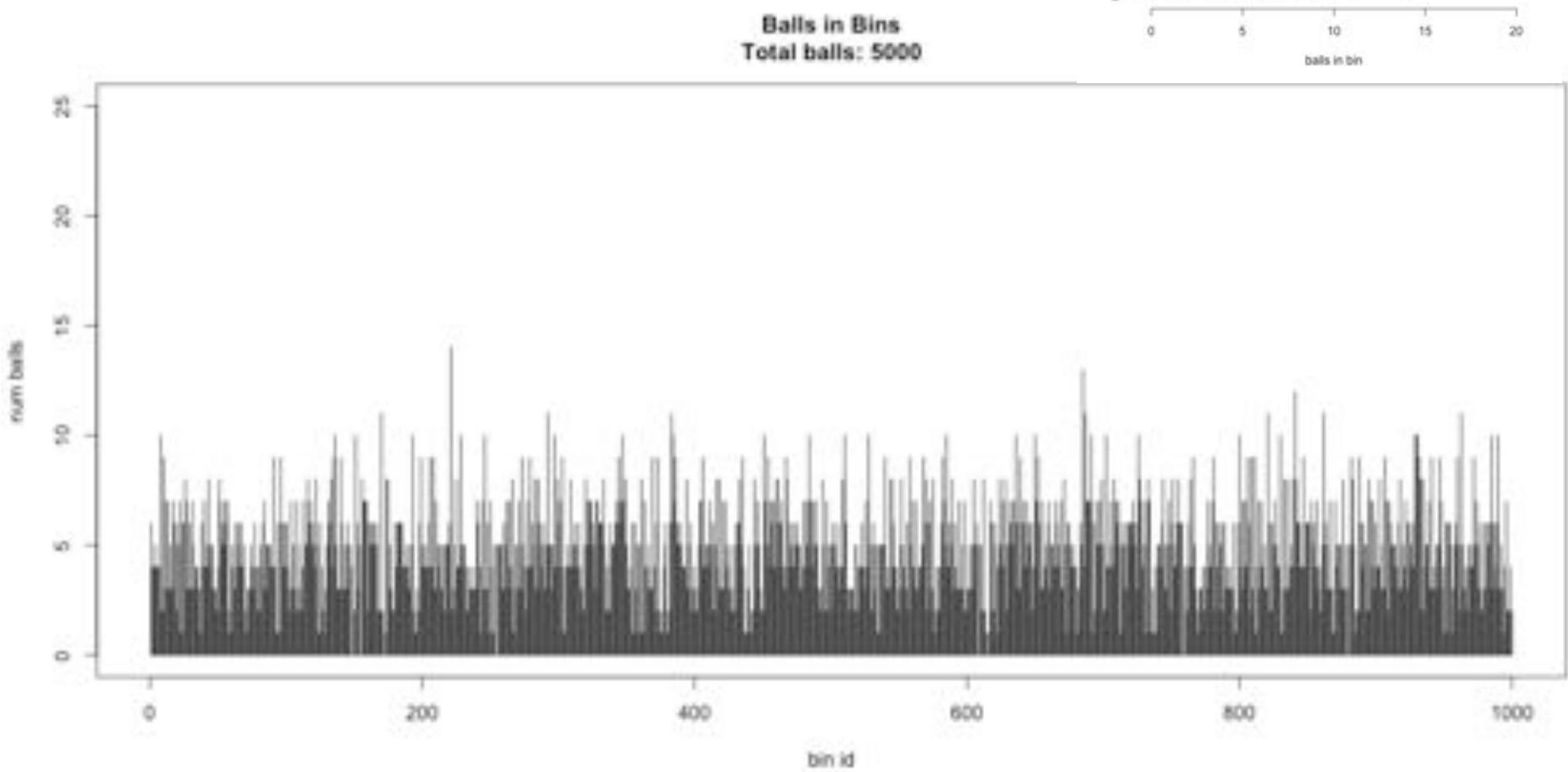
# Balls in Bins 3x



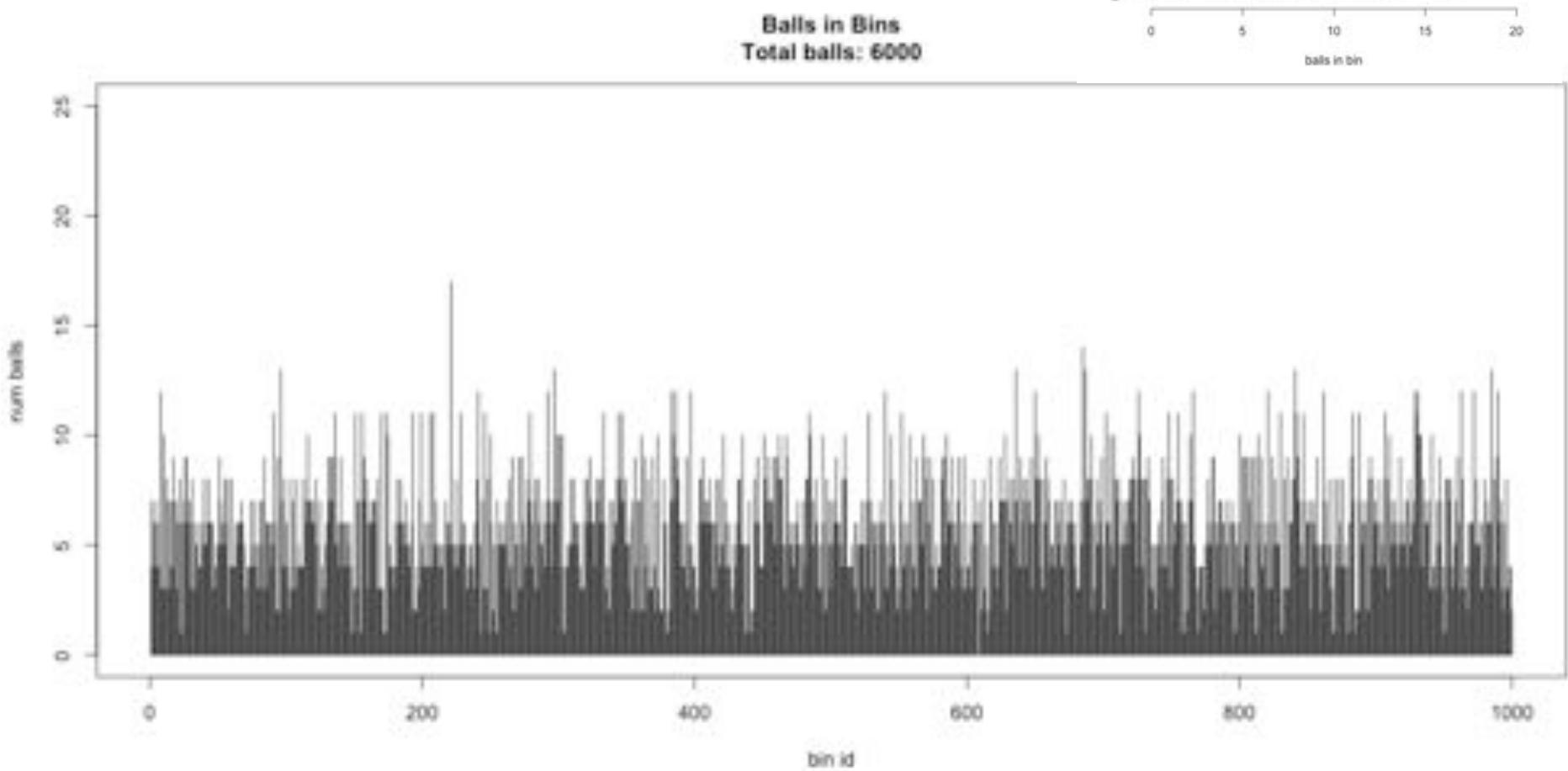
# Balls in Bins 4x



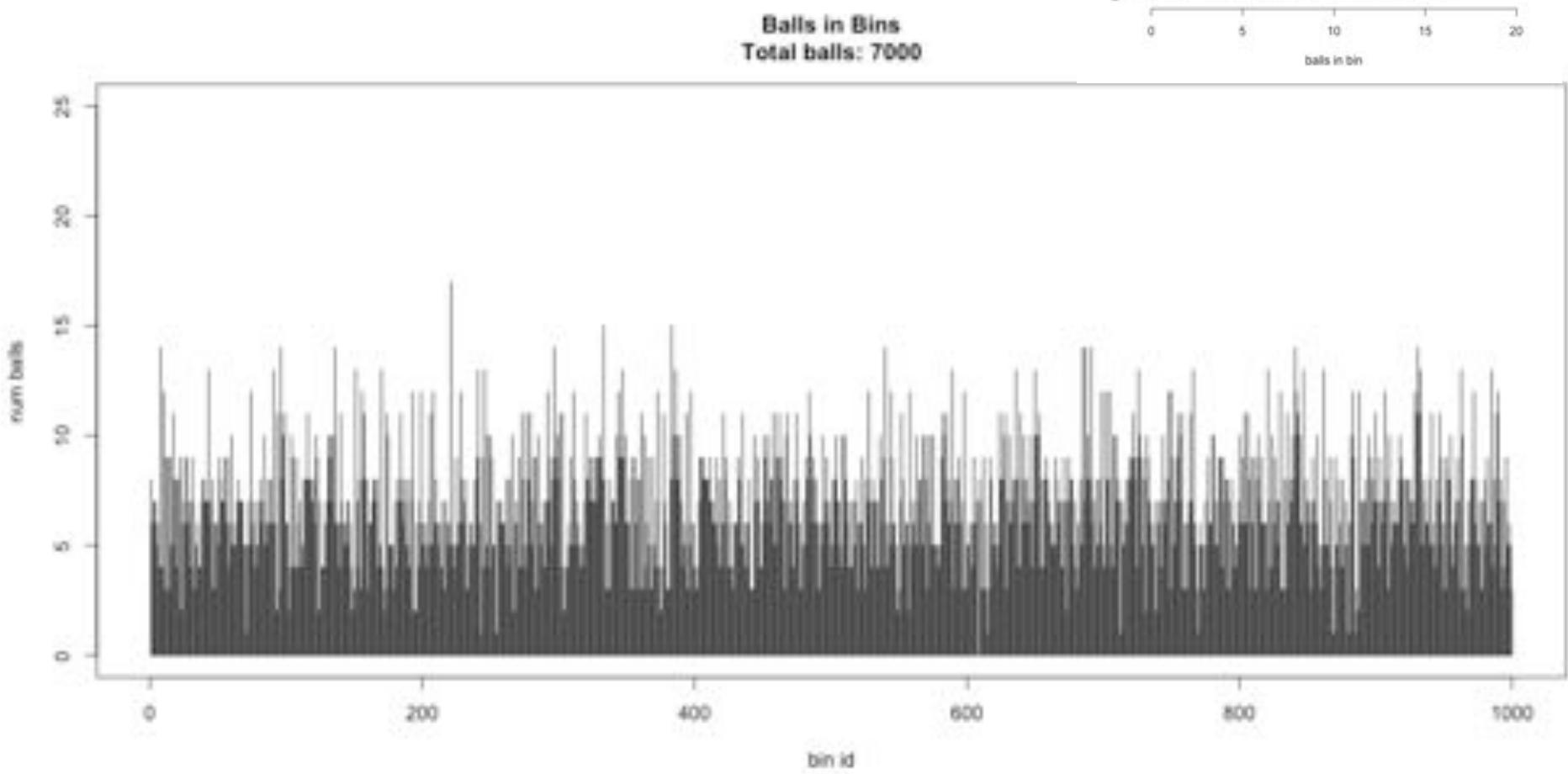
# Balls in Bins 5x



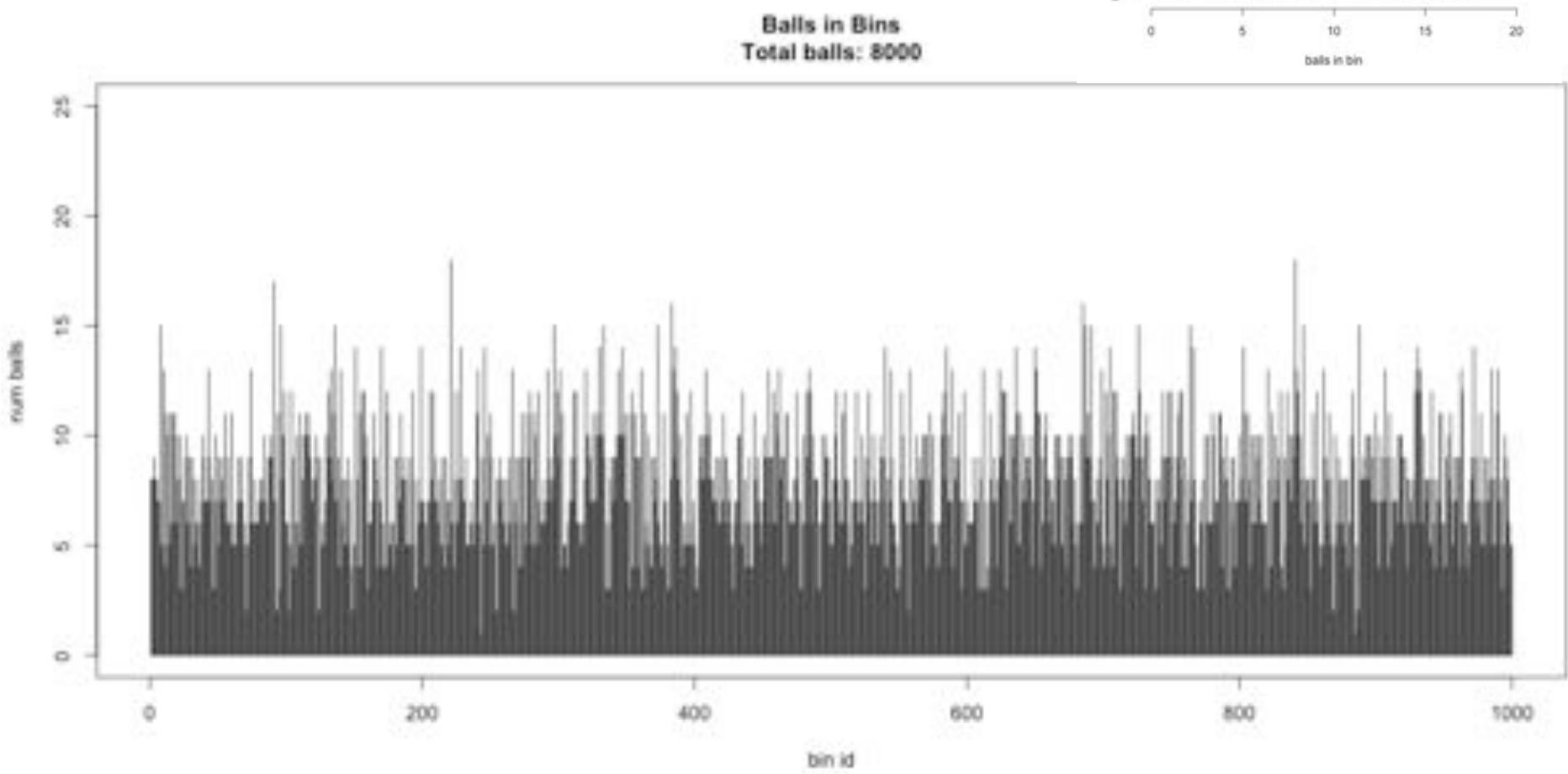
# Balls in Bins 6x



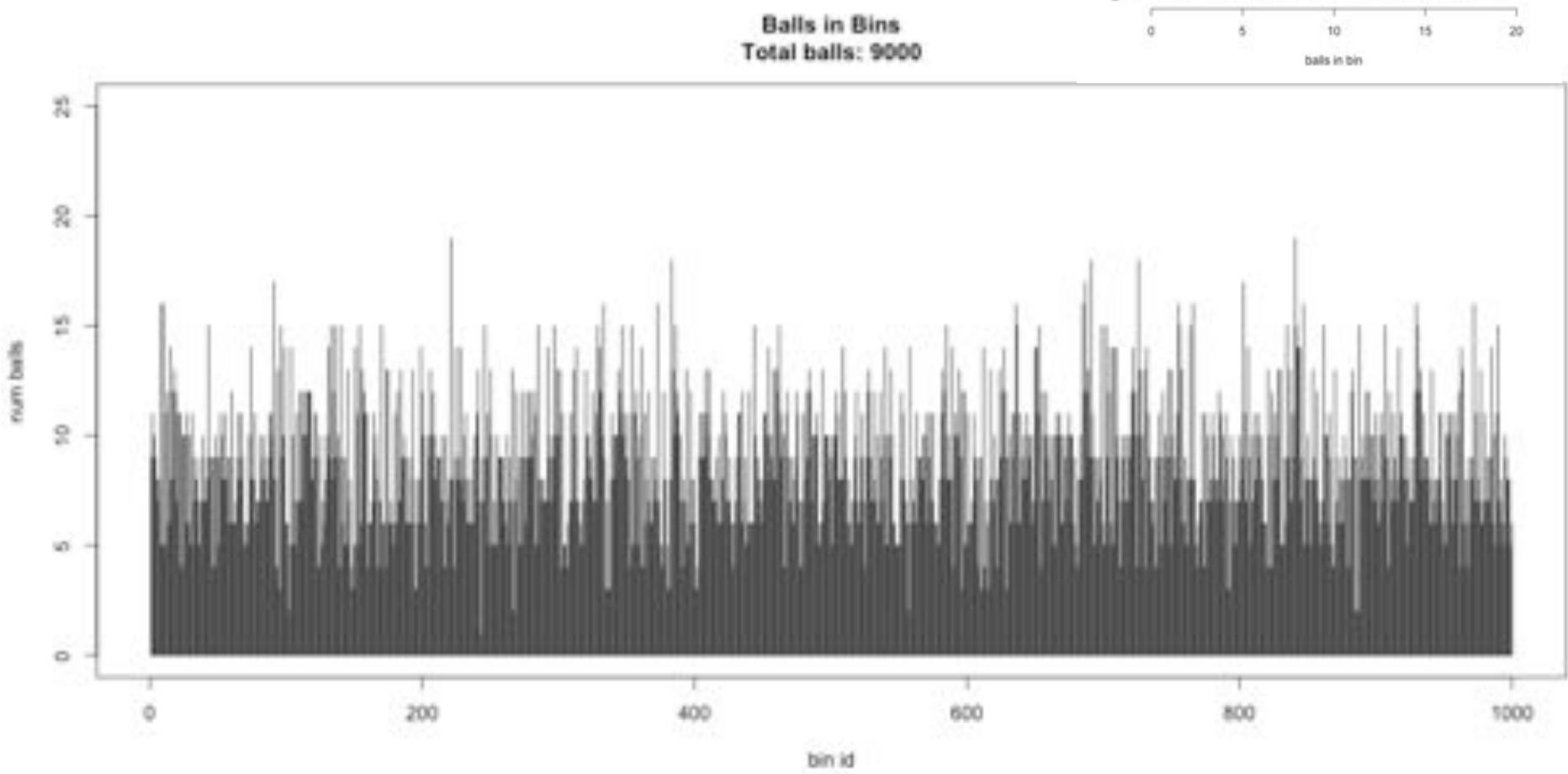
# Balls in Bins 7x



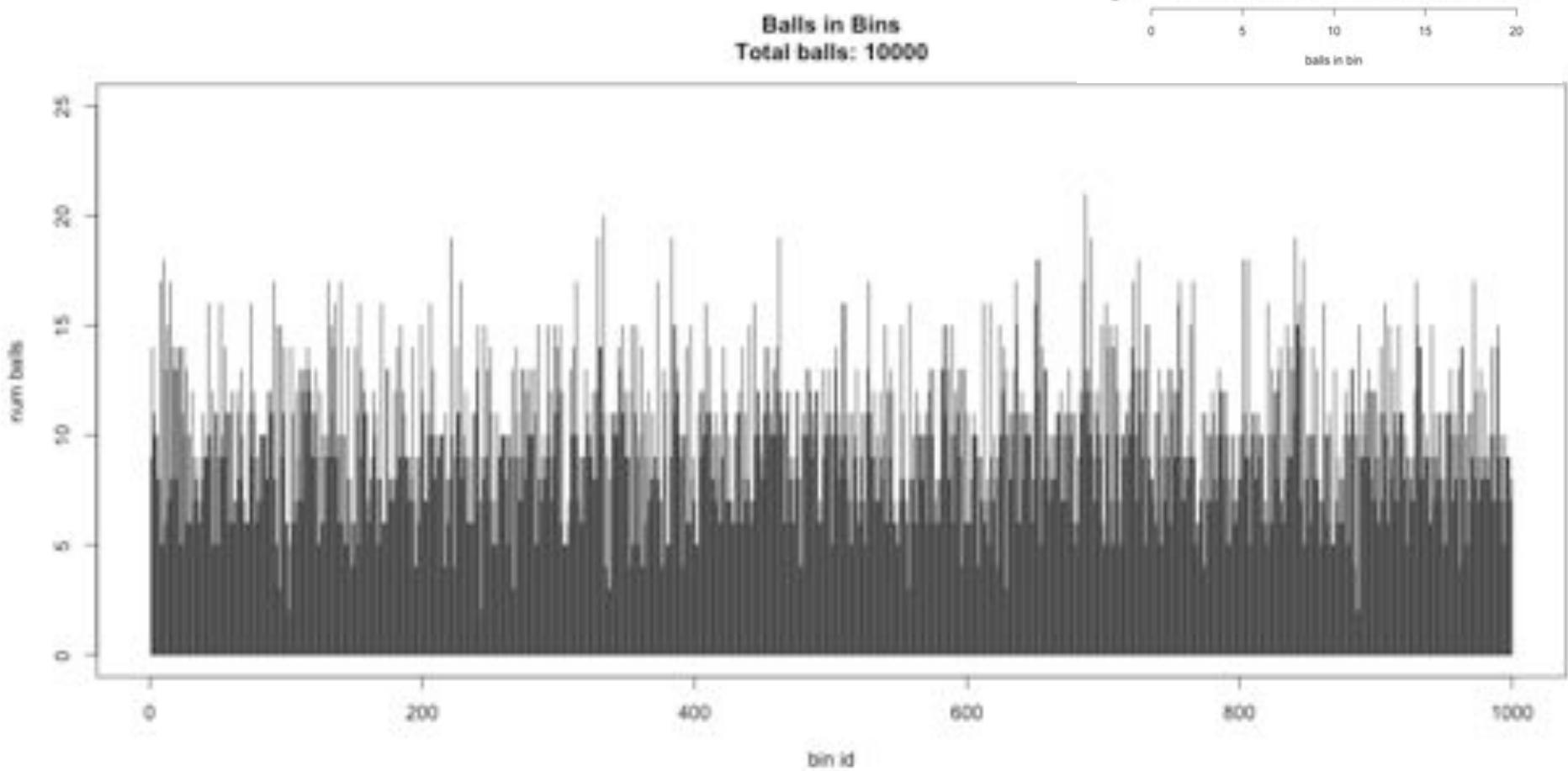
# Balls in Bins 8x



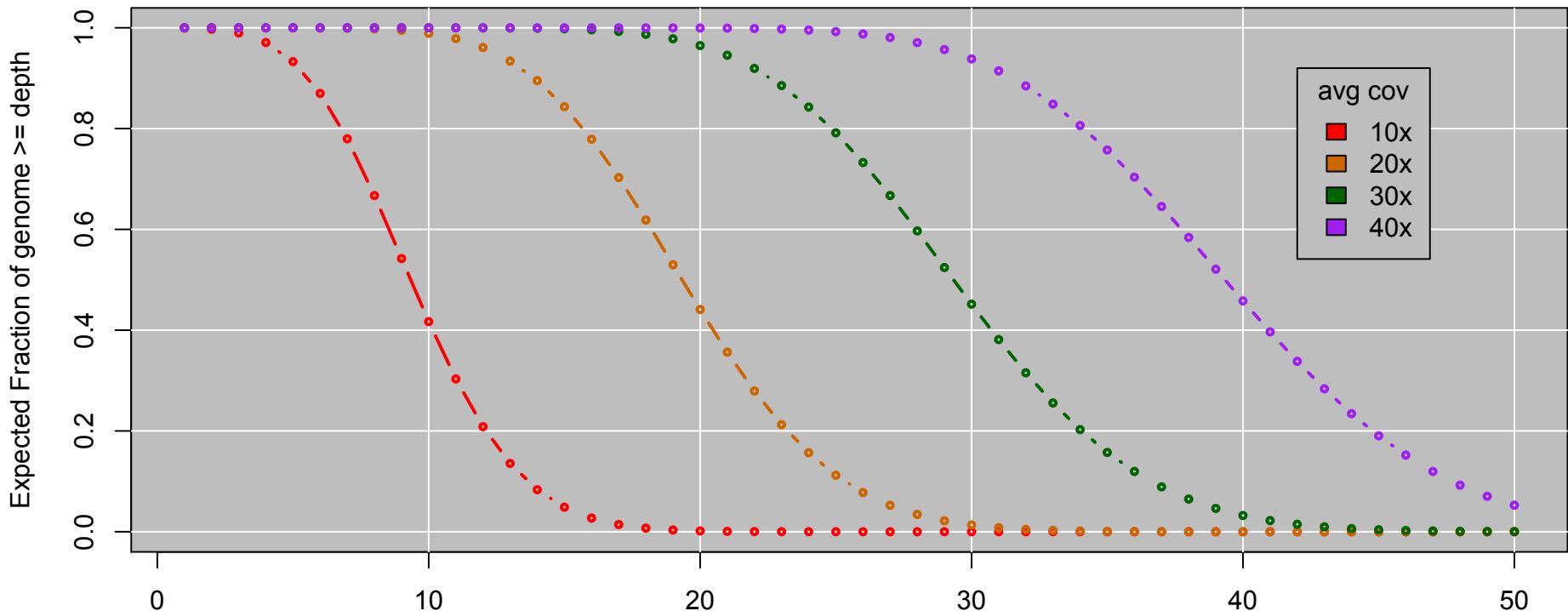
# Balls in Bins 9x



# Balls in Bins 10x



# Genome Coverage Distribution

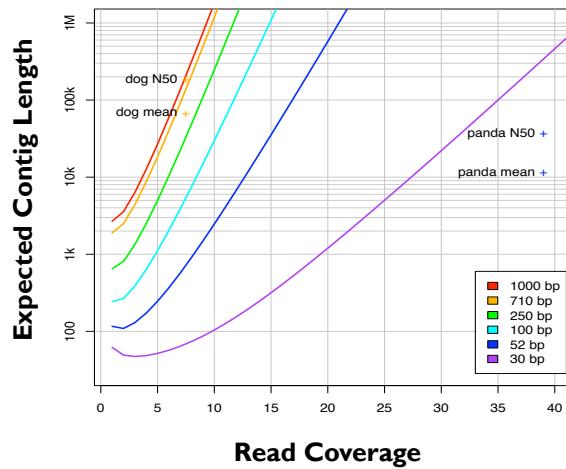


Expect Poisson distribution on depth  
Standard Deviation =  $\sqrt{\text{cov}}$

This is the mathematically model  $\Rightarrow$  reality may be much worse  
Double your coverage for diploid genomes

# Ingredients for a good assembly

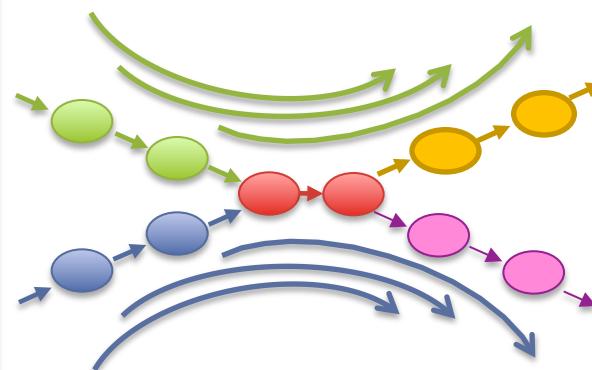
## Coverage



### High coverage is required

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

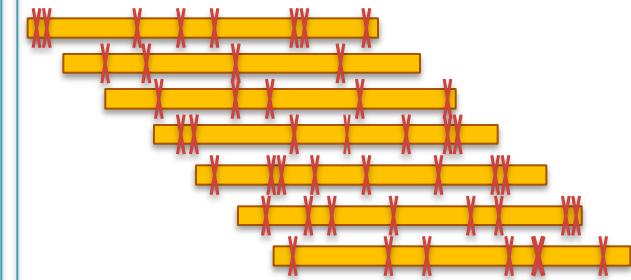
## Read Length



### 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



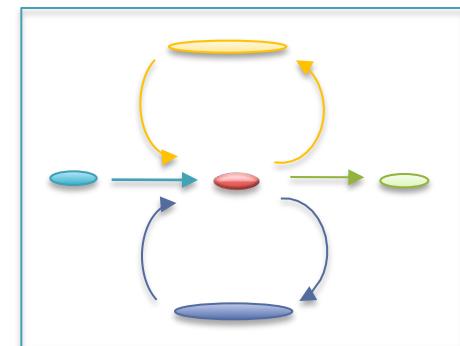
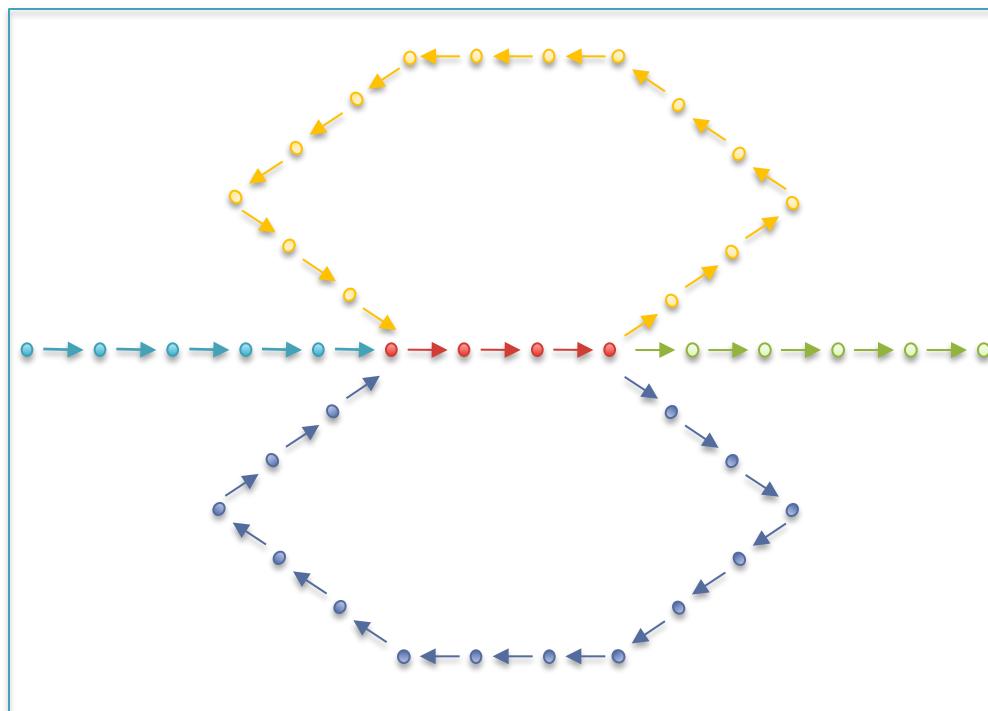
### Errors obscure overlaps

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

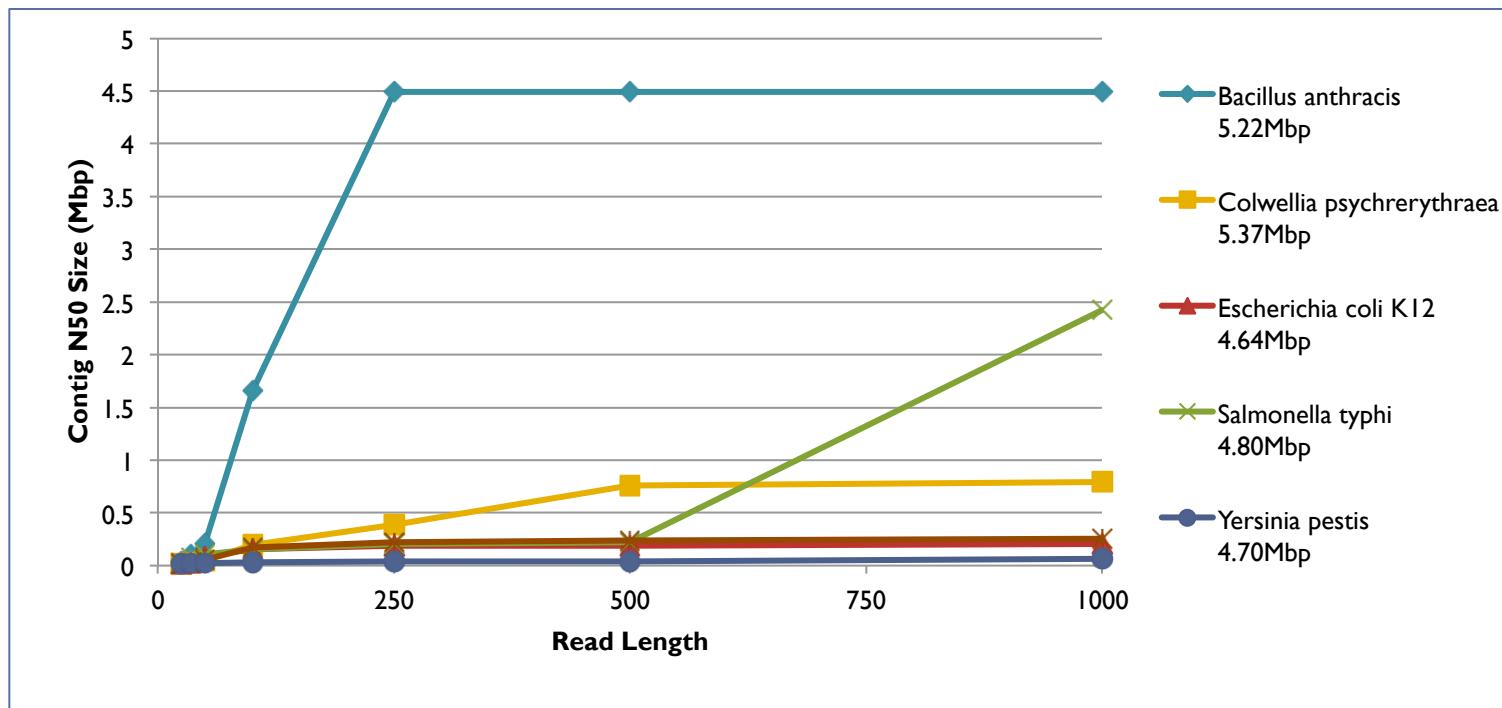
**Current challenges in *de novo* plant genome sequencing and assembly**  
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*.

# Initial Contigs

- After simplification and correction, compress graph down to its non-branching initial contigs
  - Aka “unitigs”, “unipaths”



# Repeats and Read Length



- Explore the relationship between read length and contig N50 size
  - Idealized assembly of read lengths: 25, 35, 50, 100, 250, 500, 1000
  - Contig/Read length relationship depends on specific repeat composition

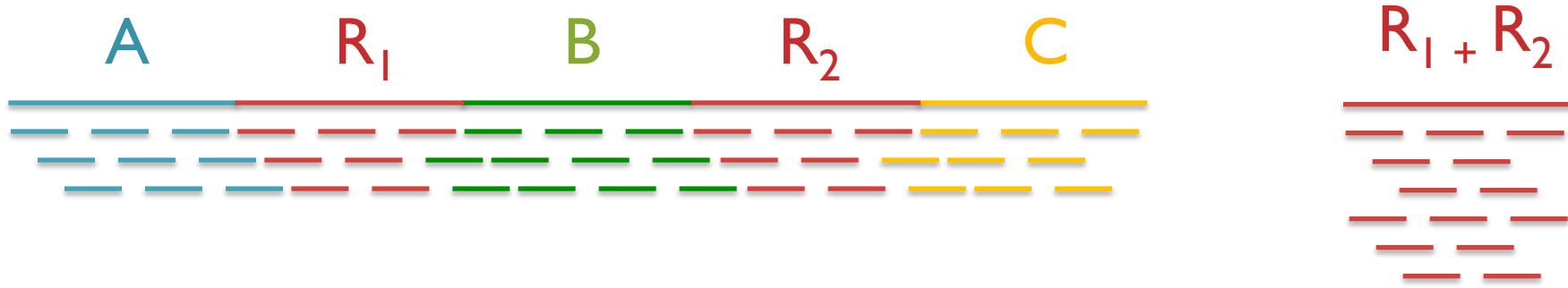
**Assembly Complexity of Prokaryotic Genomes using Short Reads.**  
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics*. 11:21.

# Repetitive regions

- Over 50% of the human genome is repetitive

Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1 b_2 \dots b_k)^N$ where $1 \leq k \leq 6$ CACACACACACACACACACA	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%

# Repeats and Coverage Statistics



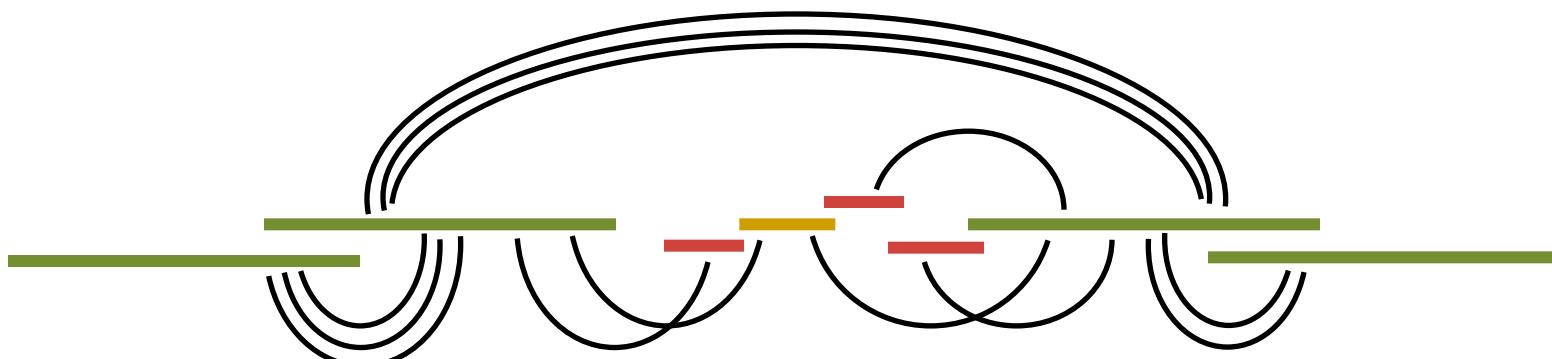
- If  $n$  reads are a uniform random sample of the genome of length  $G$ , we expect  $k = n \Delta/G$  reads to start in a region of length  $\Delta$ .
  - If we see many more reads than  $k$  (if the arrival rate is  $> A$ ) , it is likely to be a collapsed repeat
  - Requires an accurate genome size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

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

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

# 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 larger than N50

Example: 1 Mbp genome



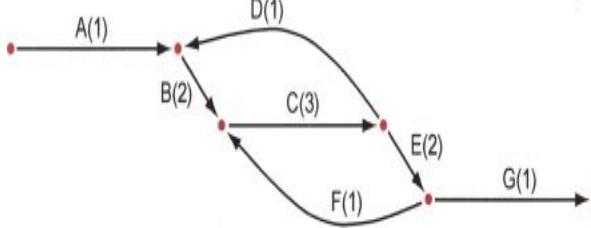
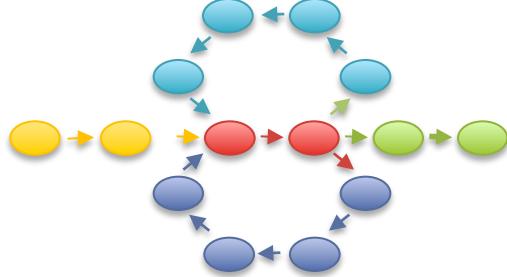
N50 size = 30 kbp

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

Note:

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

# Assembly Algorithms

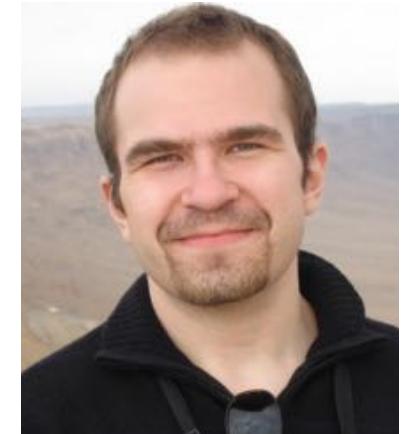
ALLPATHS-LG	SOAPdenovo	Celera Assembler
		
Broad's assembler (Gnerre et al. 2011)	BGI's assembler (Li et al. 2010)	JCVI's assembler (Miller et al. 2008)
De bruijn graph Short + PacBio (patching)	De bruijn graph Short reads	Overlap graph Medium + Long reads
Easy to run if you have compatible libraries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

# PacBio Error Correction & Assembly

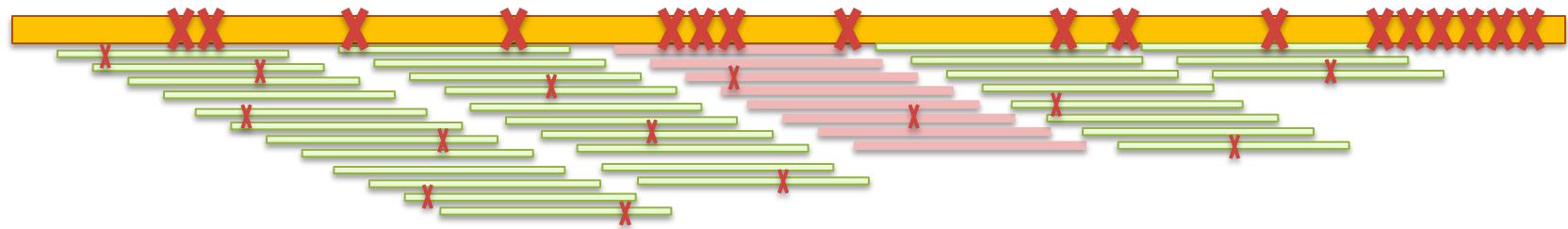
<http://wgs-assembler.sf.net>

## I. Correction Pipeline

1. Map short reads (SR) to long reads (LR)
2. Trim LRs at coverage gaps
3. Compute consensus for each LR



## 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, Walenz, BP, Martin, J, Howard, J, Ganapathy, G, Wang, Z, Rasko, DA, McCombie, WR, Jarvis, ED, Phillippy, AM. (2012) *Nature Biotechnology*. doi:10.1038/nbt.2280

# Assembly of Heterozygous Genomes

E. Biggers, M. Schatz

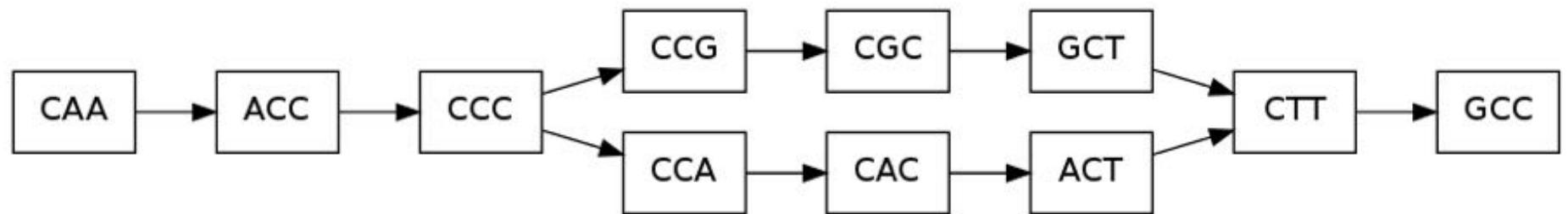


Genome assemblers developed to assembly genomes with low rates of heterozygosity

- 0-.1% (similar to human)

Assembly becomes more complicated with higher rates

Preprocess the reads to “smooth” the heterozygosity, assemble, and then restore variants



# Scalpel: Haplotype Microassembly

G. Narzisi, D. Levy, I. Iossifov, J. Kendall, M. Wigler, M. Schatz



- Use assembly techniques to identify complex variations from short reads
  - Improved power to find indels
  - Trace candidate haplotypes sequences as paths through assembly graphs



Ref: . . . TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTGCCCGGA . . .

Father: . . . TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTGCCCGGA . . .

Mother: . . . TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTGCCCGGA . . .

Sib: . . . TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTGCCCGGA . . .

Aut(1): . . . TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTGCCCGGA . . .

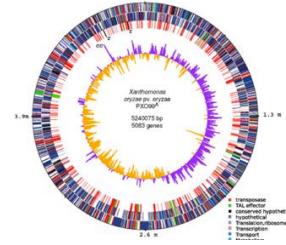
Aut(2): . . . TCAGAACAGCTGGATGAGATCTTACC-----CC**G**GGAGATTGTCTTGCCCGGA . . .

6bp heterozygous indel at chr13:25280526 ATP12A

# Assembly Summary

Graphs are ubiquitous in the world

- Pairwise searching is easy, finding features is hard



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

# Genomics Challenges



**The foundations of genomics will continue to be *observation, experimentation, and interpretation***

- Technology will continue to push the frontier
- Measurements will be made *digitally* over large populations, at extremely high resolution, and for diverse applications

**Rise in Quantitative and Computational Demands**

1. *Experimental design*: selection, collection & metadata
2. *Observation*: measurement, storage, transfer, computation
3. *Integration*: multiple samples, assays, analyses
4. *Discovery*: visualizing, interpreting, modeling

***Ultimately limited by the human capacity to execute extremely complex experiments and interpret results***

# Acknowledgements

## Schatzlab

Eric Biggers

Hayan Lee

Mitch Bekritsky

James Gurtowski

Rushil Gupta

Giuseppe Narzisi

Rob Aboukhalil

## CSHL

Hannon Lab

Iossifov Lab

Levy Lab

Lippman Lab

Martienssen Lab

McCombie Lab

Ware Lab

Wigler Lab

## NBACC

Adam Phillippy

Sergey Koren

## UMD

Steven Salzberg

Mihai Pop

Ben Langmead

Cole Trapnell



# Thank You



<http://schatzlab.cshl.edu/teaching/>  
@mike\_schatz