Graphs and Genomes

Michael Schatz

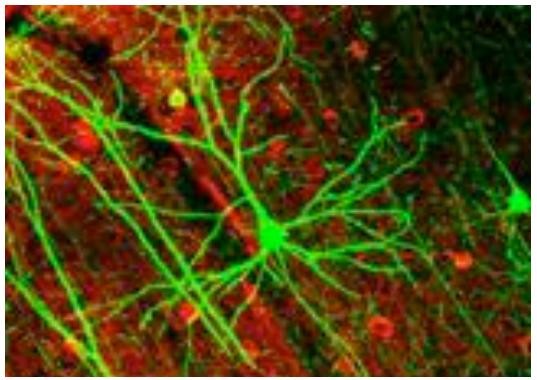
July 26, 2013

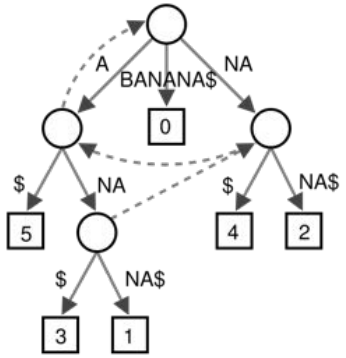
CSHL Undergraduate Research Program

Outline

1. Graph Searching
2. Assembly by analogy
3. Genome Assembly

Biological Networks





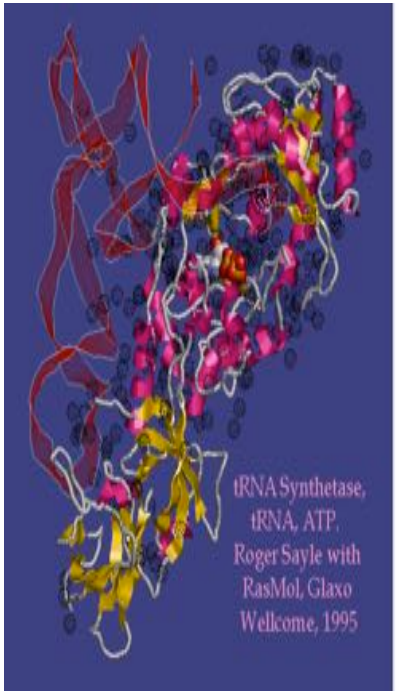
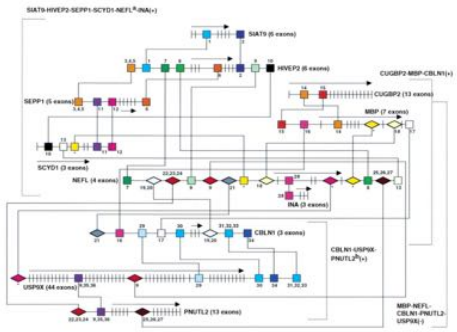
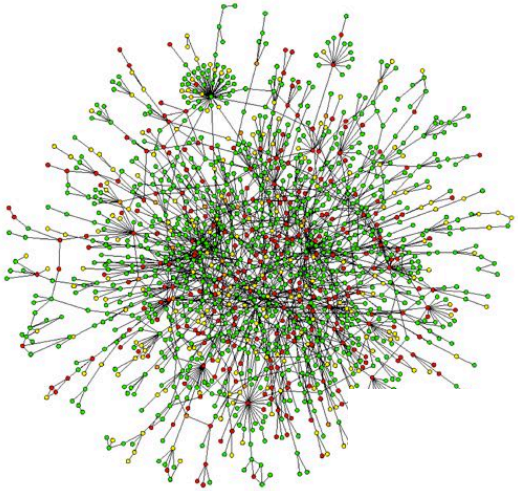


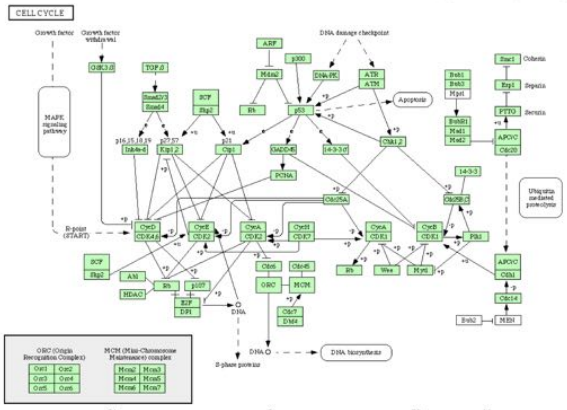
Figure 5 Putative regulatory elements shared between groups of correlated and anticorrelated genes

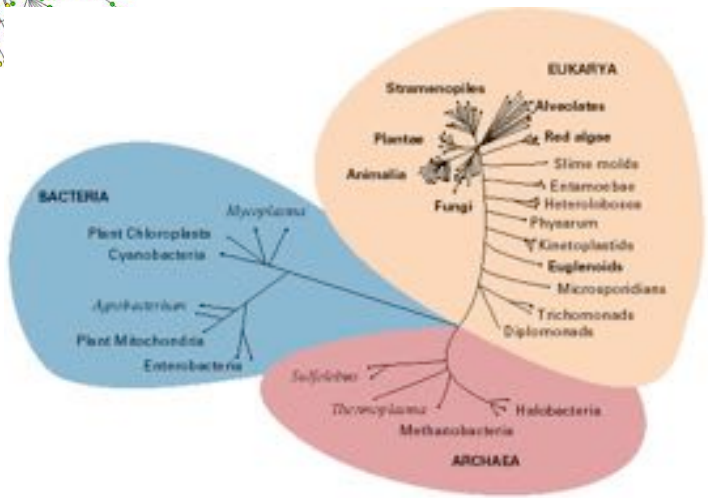


Vanessa M. Brown et al. Genome Res. 2002; 12: 868-884

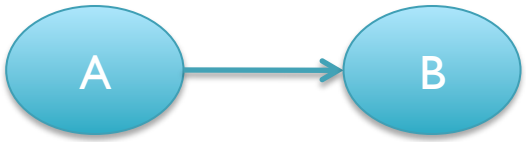
Cold Spring Harbor Laboratory Press





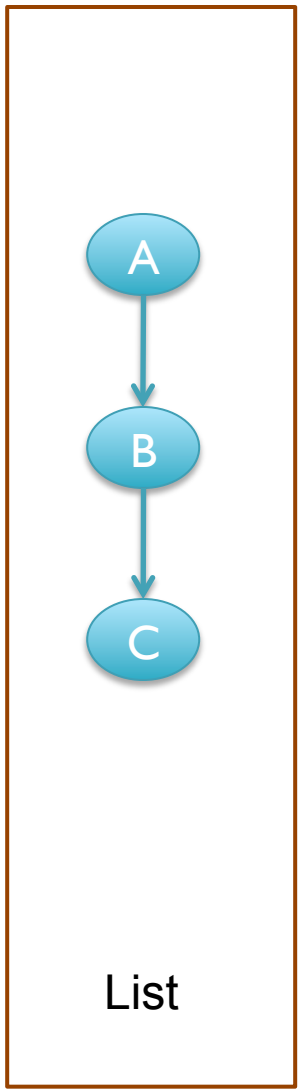


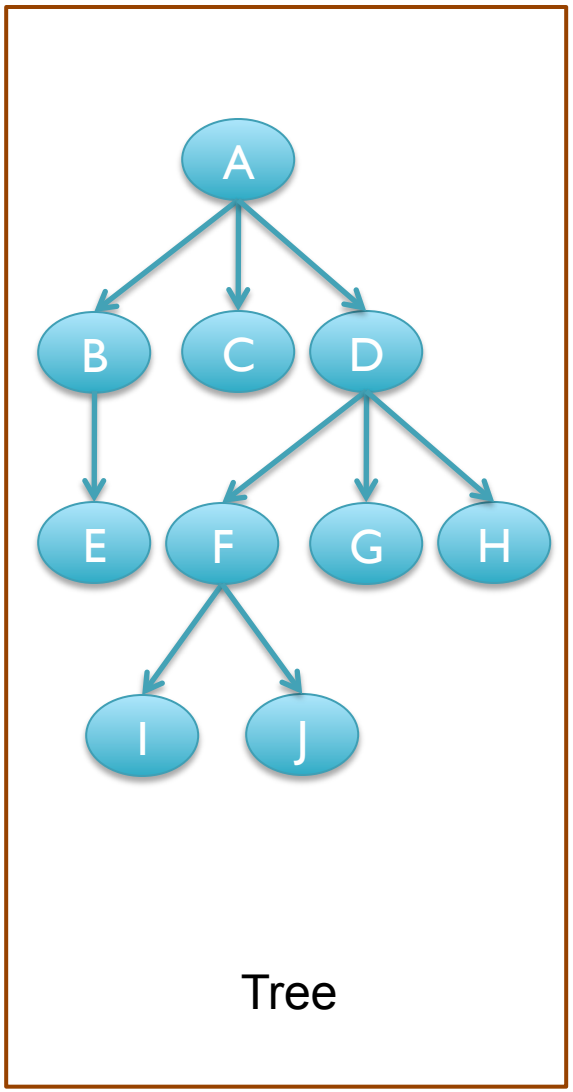
Graphs

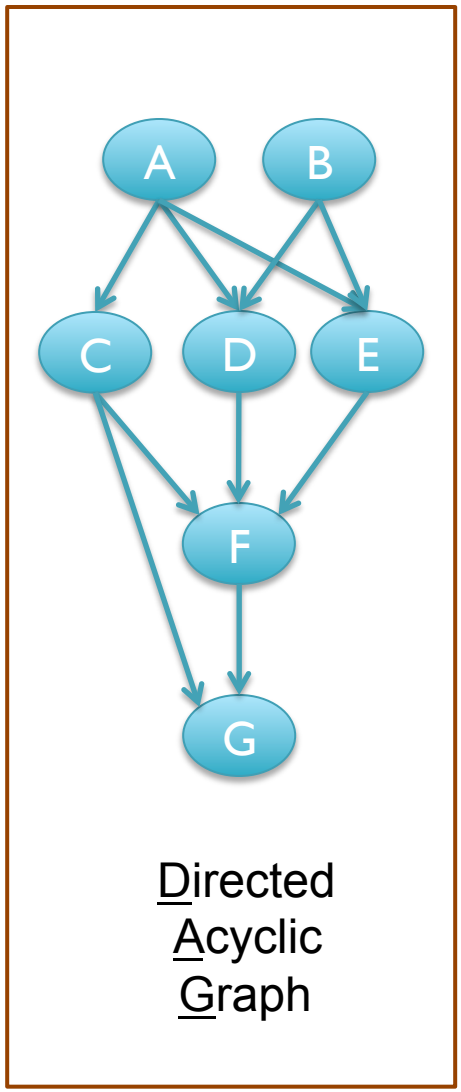


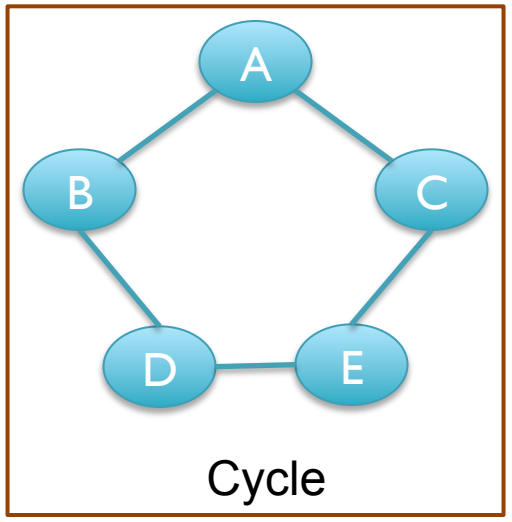
- Nodes
  - People, Proteins, Genes, Neurons, Sequences, Numbers, …
- Edges
  - A is connected to B
  - A is related to B
  - A regulates B
  - A precedes B
  - A interacts with B
  - A activates B
  - …

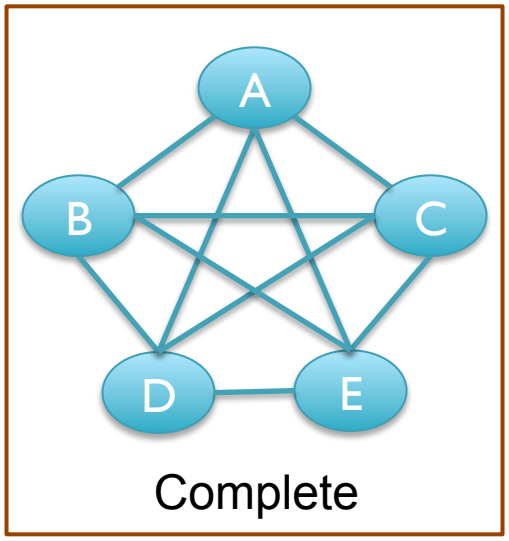
Graph Types


A
B
C
List

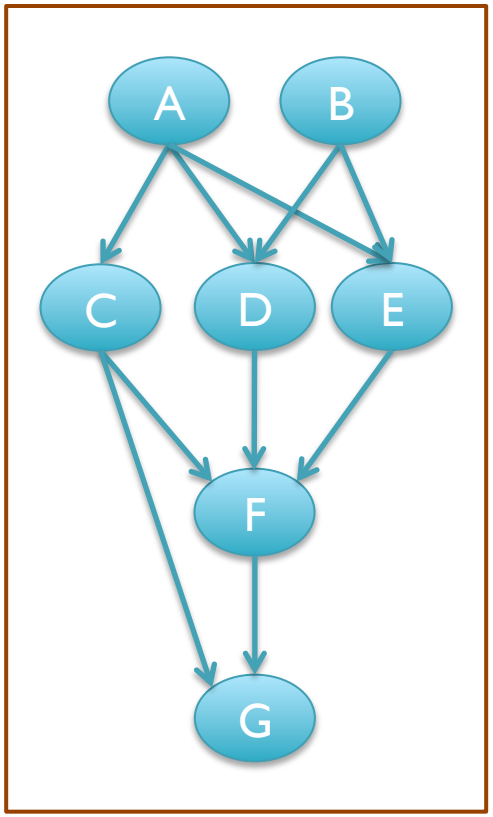

A
B
C
D
E
F
G
H
I
J
Tree


A
B
C
D
E
F
G
Directed
Acyclic
Graph


A
B
C
D
E
Cycle


A
B
C
D
E
Complete

Representing Graphs



Adjacency Matrix

Good for dense graphs
Fast, Fixed storage: N^2 bits

	A	B	C	D	E	F	G
A			1	1	1		
B				1	1		
C						1	1
D						1	
E						1	
F							1
G							

Adjacency List

Good for sparse graphs
Compact storage: 4 bytes/edge

A: C, D, E
B: D, E
C: F, G
D: F
E: F
G:

Edge List

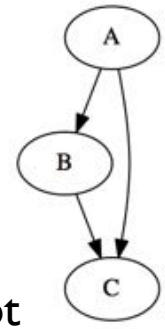
Easy, good if you (mostly) need to iterate through the edges
8 bytes / edge

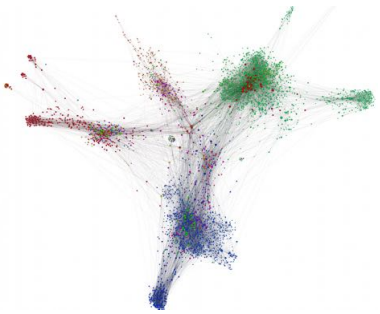
A,C A,D A,E
B,C B,D B,E
C,F C,G D,F
E,F F,G

Tools

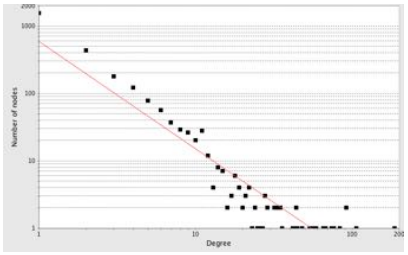
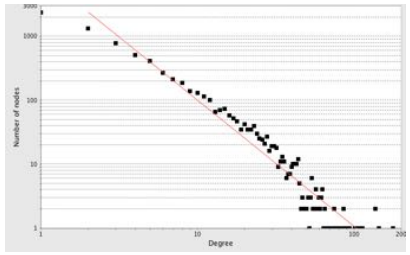
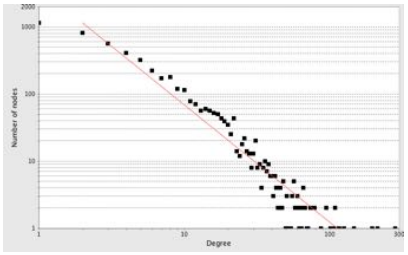
Matlab: http://www.mathworks.com/
Graphviz: http://www.graphviz.org/
Gephi: https://gephi.org/
Cytoscape: http://www.cytoscape.org/

```
digraph G {
    A->B
    B->C
    A->C
}
dot –Tpdf -og.pdf g.dot
```





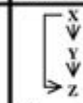
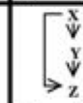
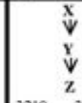
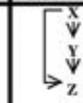
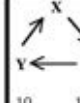
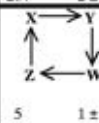
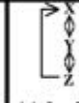
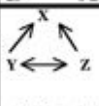
Network Characteristics

	C. elegans	*D. melanogaster*	*S. cerevisiae*
# Nodes	2646	7464	4965
# Edges	4037	22831	17536
Avg. / Max Degree	3.0 / 187	6.1 / 178	7.0 / 283
# Components	109	66	32
Largest Component	2386	7335	4906
Diameter	14	12	11
Avg. Shortest Path	4.8	4.4	4.1
Data Sources	2H	2x2H, TAP-MS	8x2H, 2xTAP, SUS
Degree Distributions			

Small World: Avg. Shortest Path between nodes is small
Scale Free: Power law distribution of degree – preferential attachment

Network Motifs

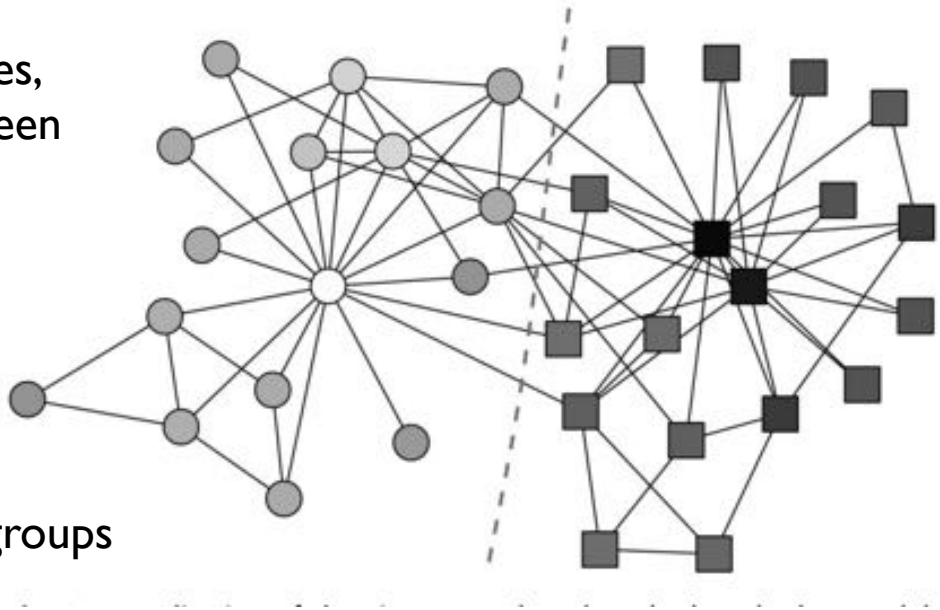
- Network Motif
  - Simple graph of connections
  - Exhaustively enumerate all possible 1, 2, 3, … k node motifs
- Statistical Significance
  - Compare frequency of a particular network motif in a real network as compared to a randomized network
- Certain motifs are "characteristic features" of the network

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)			 **Feed-forward loop**			 **Bi-fan**					
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons			 **Feed-forward loop**			 **Bi-fan**			 **Bi-parallel**		
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs			 **Three chain**			 **Bi-parallel**					
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			
Electronic circuits (forward logic chips)			 **Feed-forward loop**			 **Bi-fan**			 **Bi-parallel**		
s15850	10,383	14,240	424	2 ± 2	285	1040	1 ± 1	1200	480	2 ± 1	335
s38584	20,717	34,204	413	10 ± 3	120	1739	6 ± 2	800	711	9 ± 2	320
s38417	23,843	33,661	612	3 ± 2	400	2404	1 ± 1	2550	531	2 ± 2	340
s9234	5,844	8,197	211	2 ± 1	140	754	1 ± 1	1050	209	1 ± 1	200
s13207	8,651	11,831	403	2 ± 1	225	4445	1 ± 1	4950	264	2 ± 1	200
Electronic circuits (digital fractional multipliers)			 **Three-node feedback loop**			 **Bi-fan**			 **Four-node feedback loop**		
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web			 **Feedback with two mutual dyads**			 **Fully connected triad**			 **Uplinked mutual dyad**		
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4±4e2	15,000	1.2e6	1e4 ± 2e2	5000

Network Motifs: Simple Building Blocks of Complex Networks
Milo et al (2002) *Science. 298:824-827*

Modularity

- Community structure
  - Densely connected groups of vertices, with only sparser connections between groups
  - Reveals the structure of large-scale network data sets



- Modularity
  - The number of edges falling within groups minus the expected number in an equivalent network with edges placed at random
  - Larger positive values => Stronger community structure
  - Optimal assignment determined by computing the eigenvector of the modularity matrix

$$Q = \frac{1}{4m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) (s_i s_j + 1)$$

Normalization factor

Adjacency matrix

Random Prob. (product of degrees)

Indicates same group

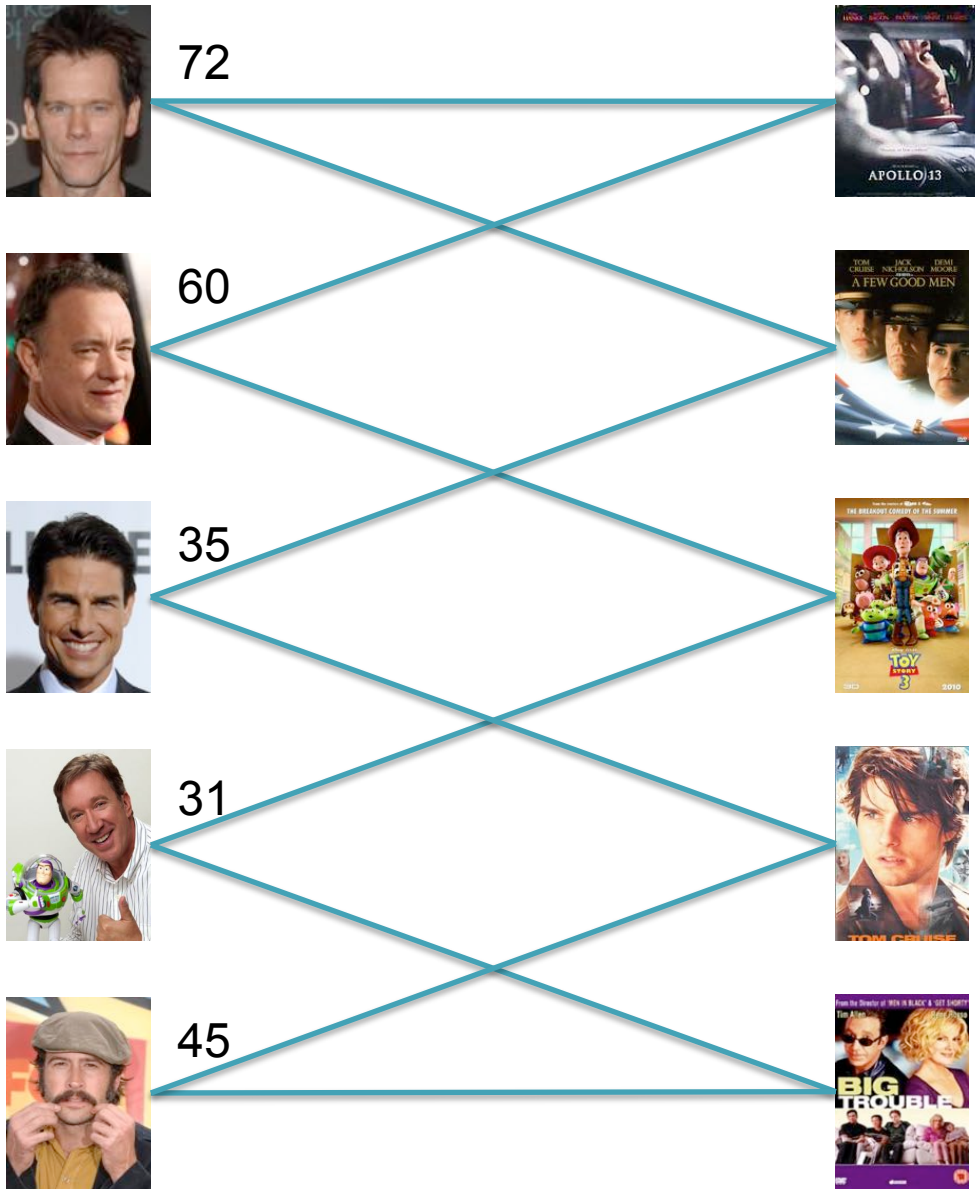
Modularity and community structure in networks.
Newman ME (2006) *PNAS. 103(23) 8577-8582*

Kevin Bacon and Bipartite Graphs

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

Breadth First Search: 4 hops

Bacon Distance: 2



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

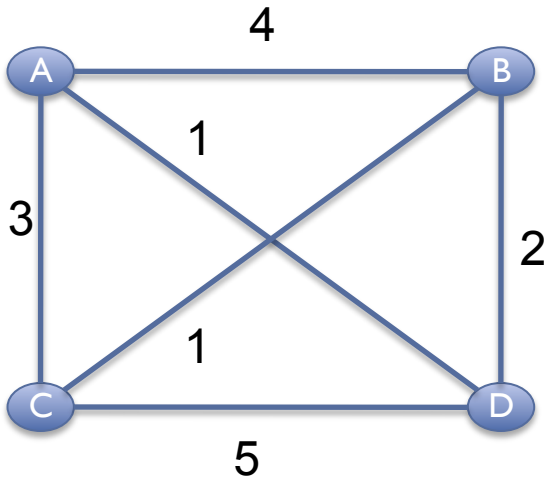
ACDBA: 3+5+2+4 = 14*

ABCDA: 4+1+5+1 = 11

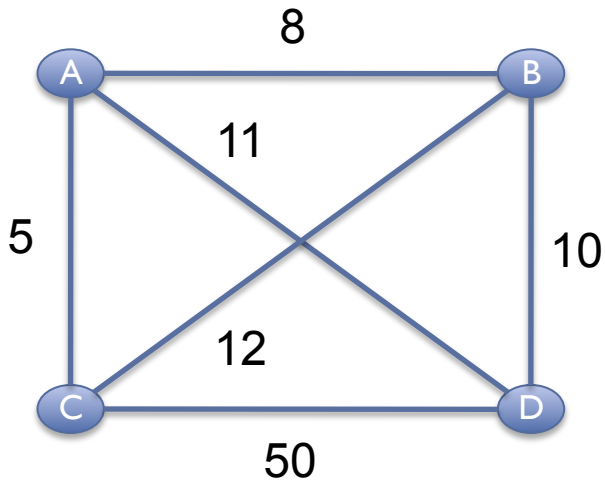
ADCBA: 1+5+1+4 = 11*

ACBDA: 3+1+2+1 = 7

ADBCA: 1+2+1+3= 7 *



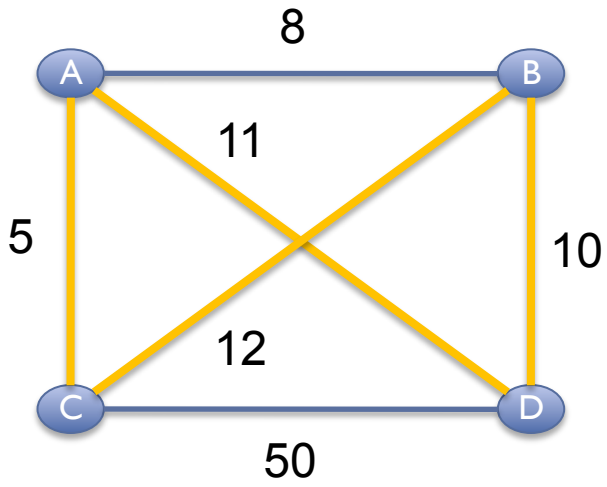
Greedy Search



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

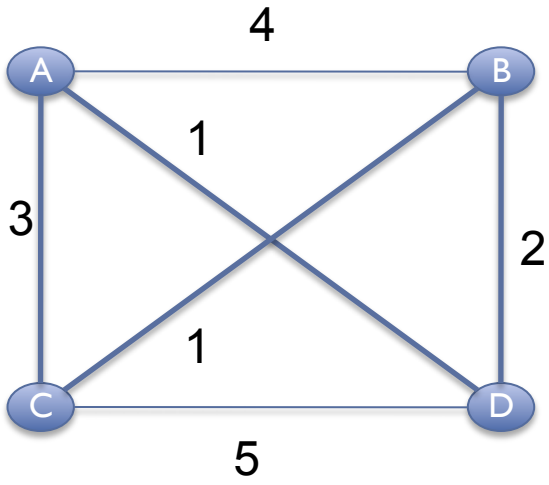
1. Greedy Choice: Local is correct without reconsideration
2. Optimal Substructure: Problem can be split into subproblems

Optimal Greedy: Making change with the fewest number of coins

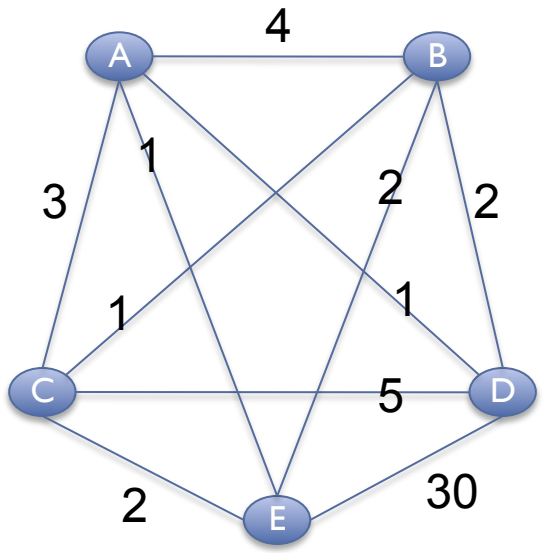
TSP Complexity

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

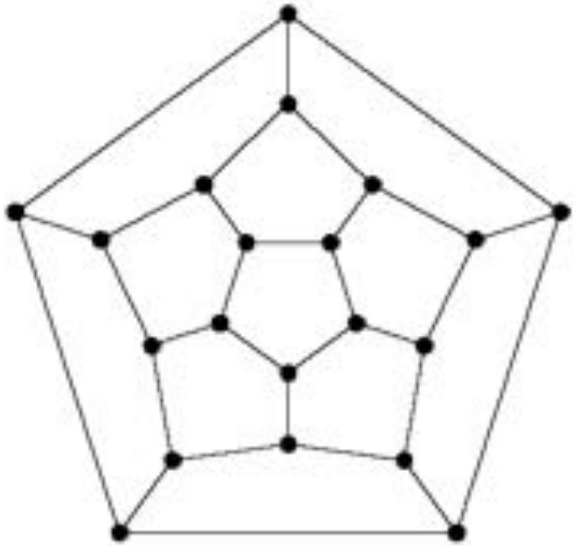
[How many possible tours for n cities?]



- 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 often 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

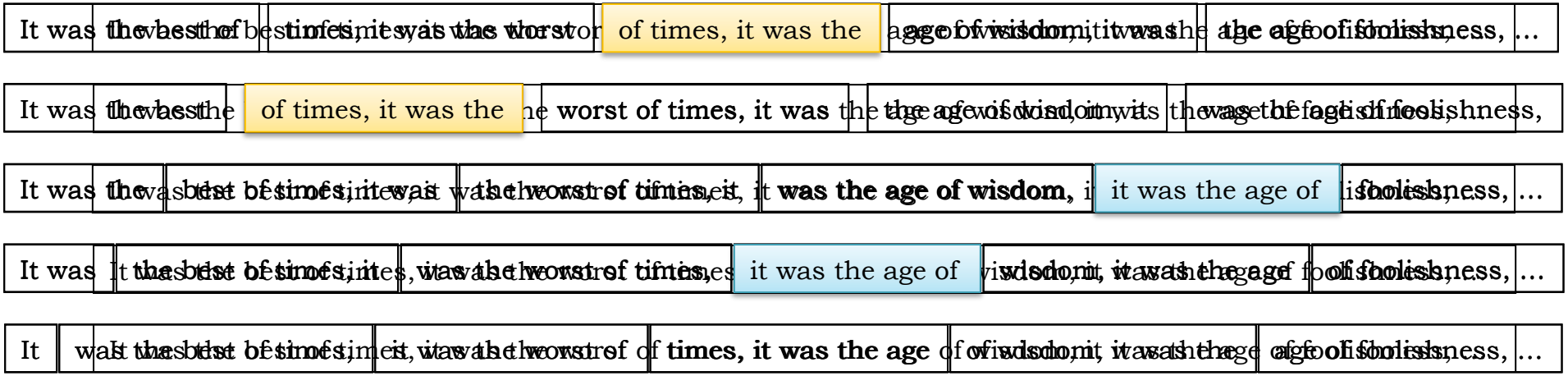
- WARNING: Many biological problems are in this class
  - Find a tour the visits every node once (Genome Assembly)
  - Find the smallest set of vertices covering the edges (Essential Genes)
  - Find the largest clique in the graph (Protein Complexes)
  - Find the highest mutual information encoding scheme (Neurobiology)
  - Find the best set of moves in tetris
  - …
  - http://en.wikipedia.org/wiki/List_of_NP-complete_problems

Outline

1. Graph Searching
2. Assembly by analogy
3. Genome Assembly

Shredded Book Reconstruction

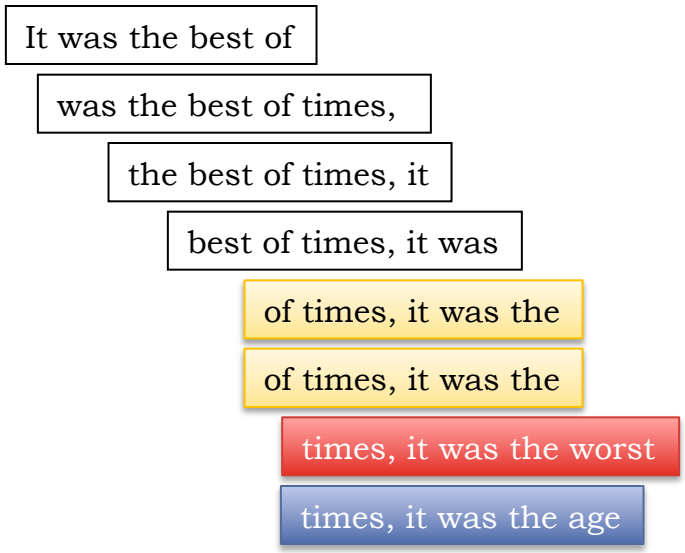
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



- 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



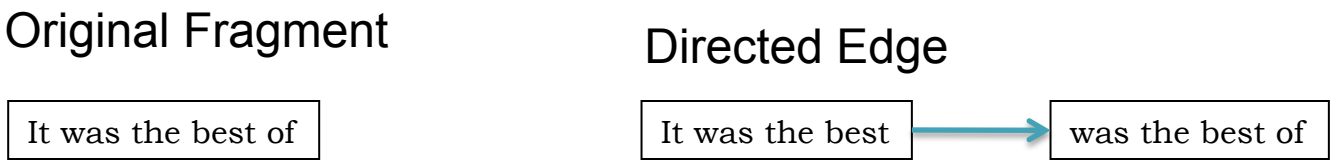
The repeated sequence make the correct reconstruction ambiguous

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

Model sequence reconstruction as a graph problem.

de Bruijn Graph Construction

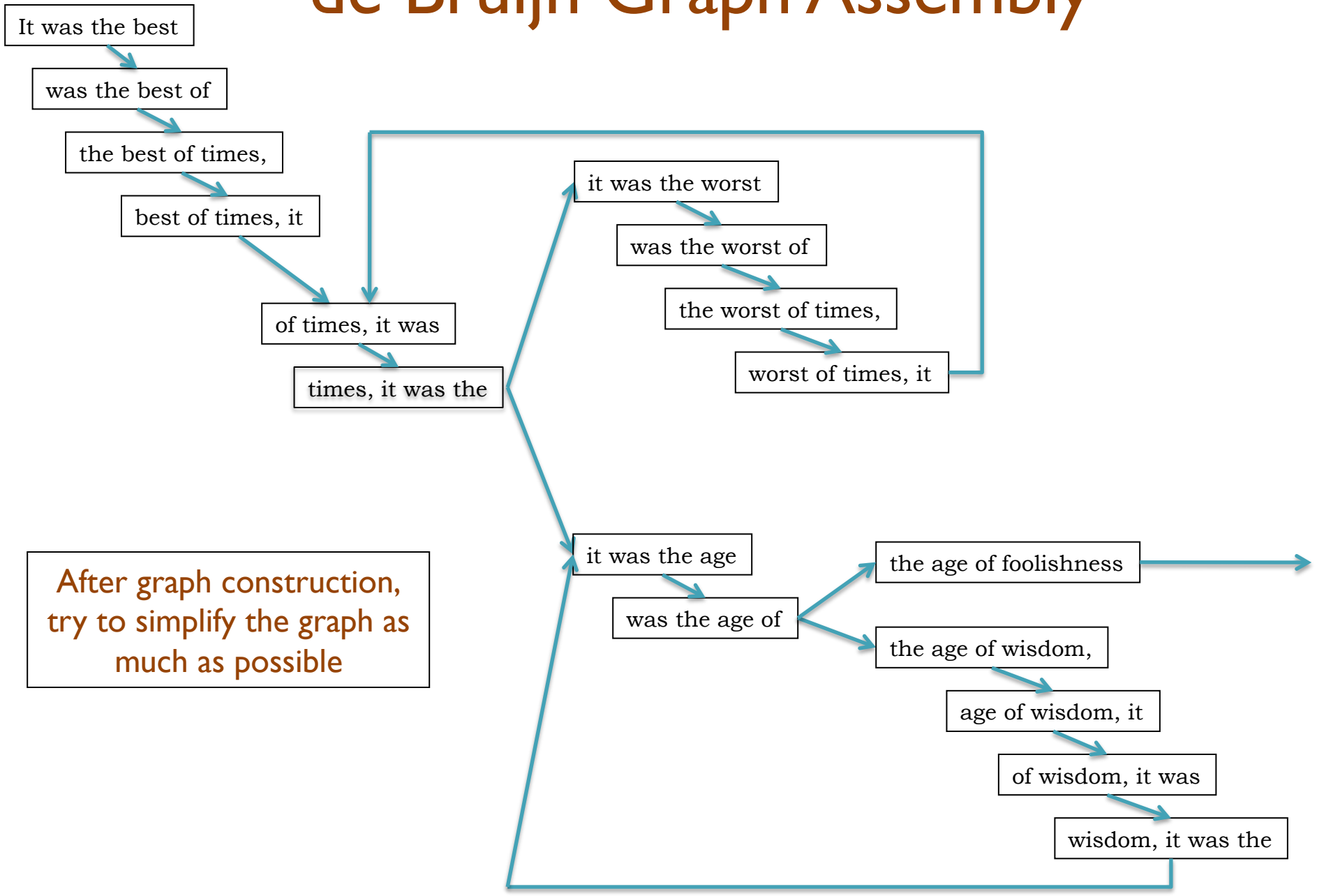
- $D_k = (V,E)$
  - V = All length-k subfragments (k < l)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words



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

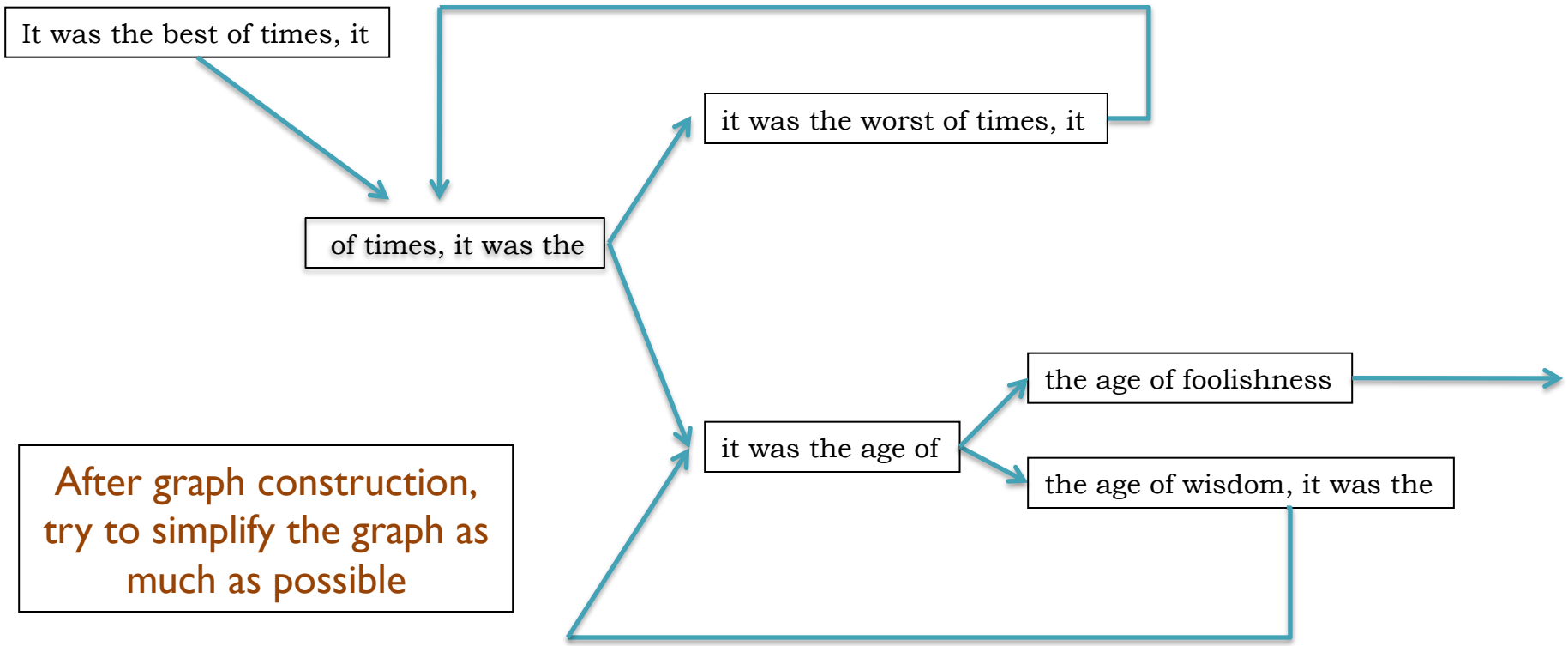
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly



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

de Bruijn Graph Assembly

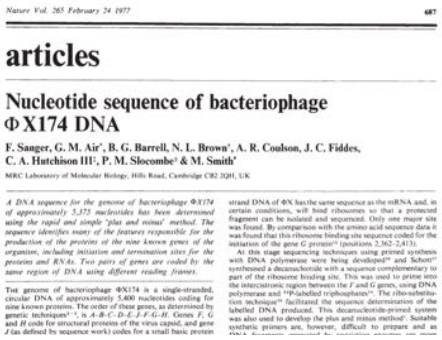


After graph construction, try to simplify the graph as much as possible

Outline

1. Genome Assembly by Analogy
2. Graph Searching
3. Genome Assembly

Milestones in Genome Assembly


Nature Vol. 265 February 24 1977

articles

Nucleotide sequence of bacteriophage ΦX174 DNA

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

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

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

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

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

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

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

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

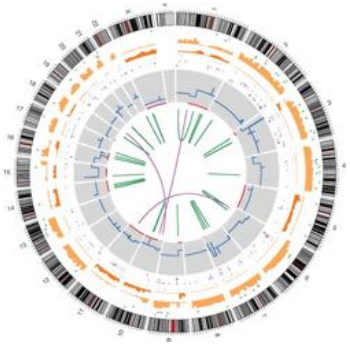
Like Dickens, we must computationally reconstruct a genome from short fragments

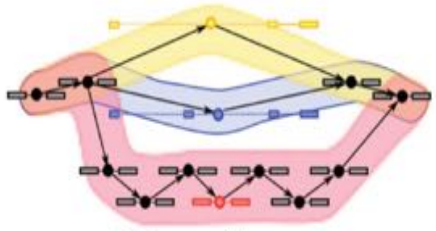
Current Applications

- Novel genomes

- Metagenomes

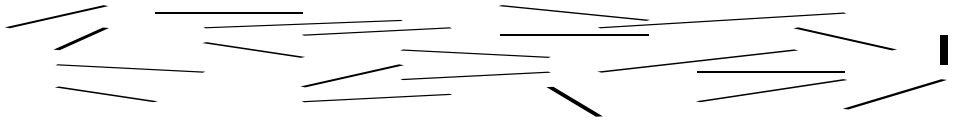
- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …





Assembling a Genome

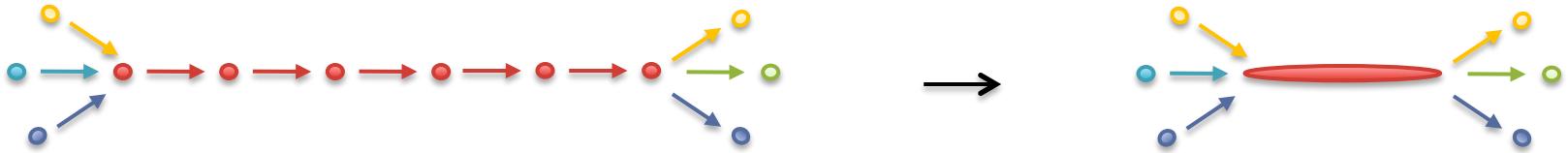
1. Shear & Sequence DNA



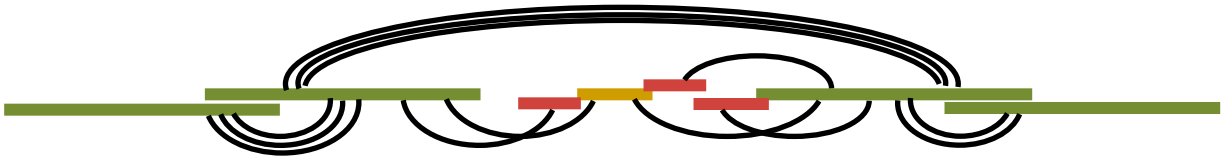
2. Construct assembly graph from overlapping reads

...AGCCTAGACCTACAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT...

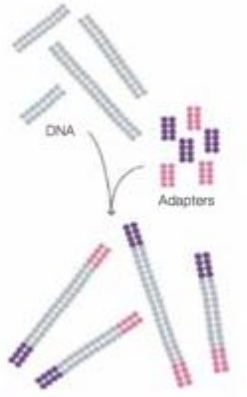
3. Simplify assembly graph



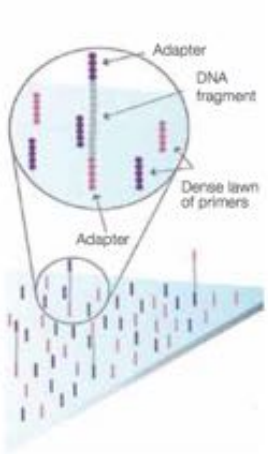
4. Detangle graph with long reads, mates, and other links



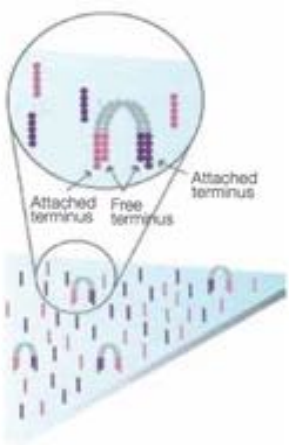
Illumina Sequencing by Synthesis



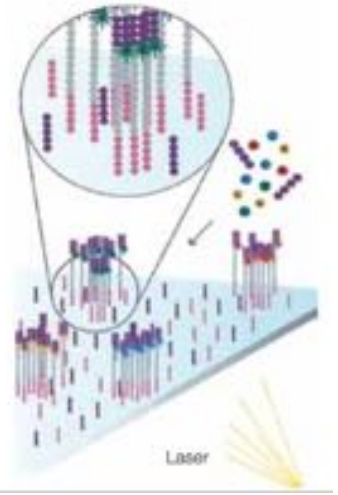
1. Prepare



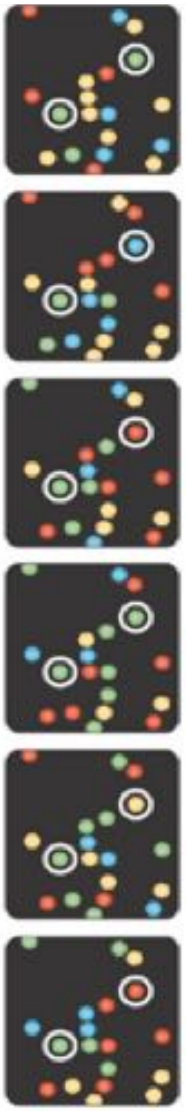
2. Attach



3. Amplify



4. Image

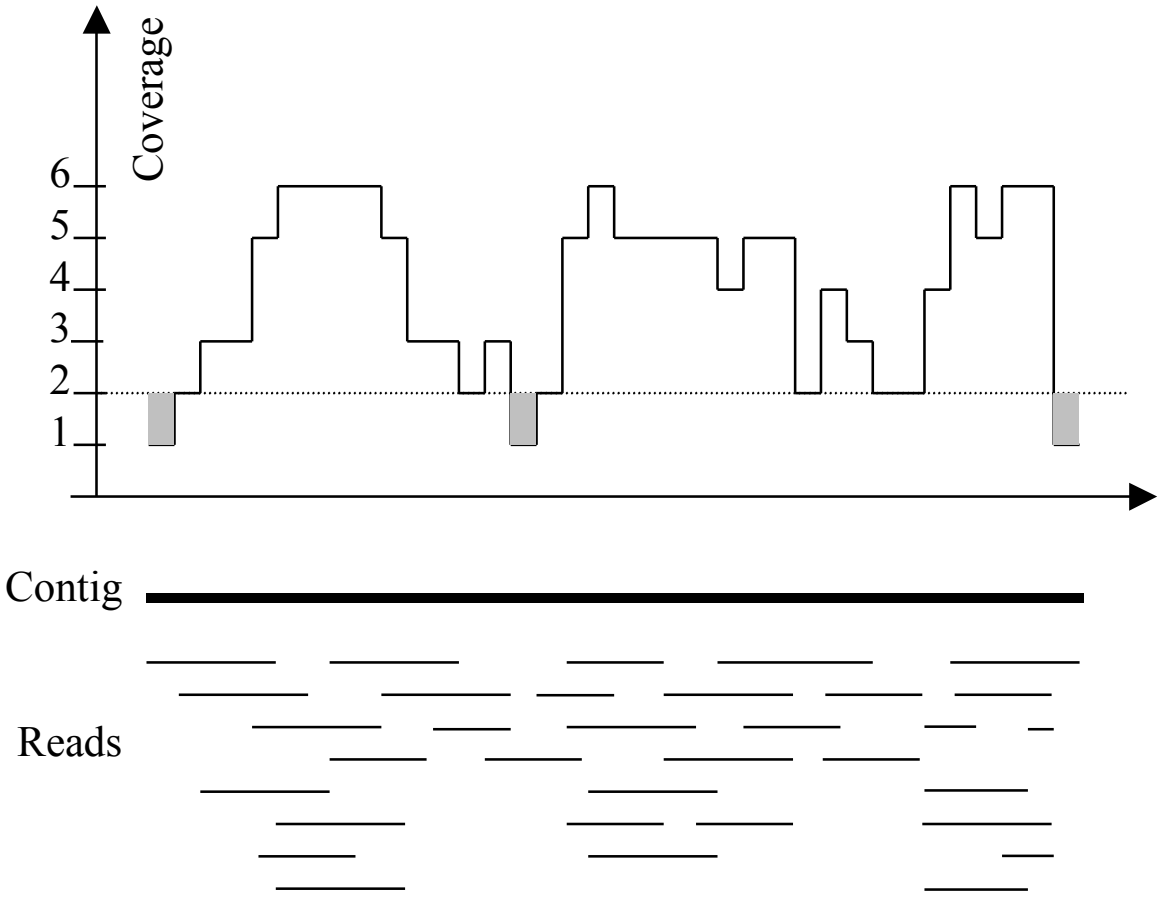


5. Basecall

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

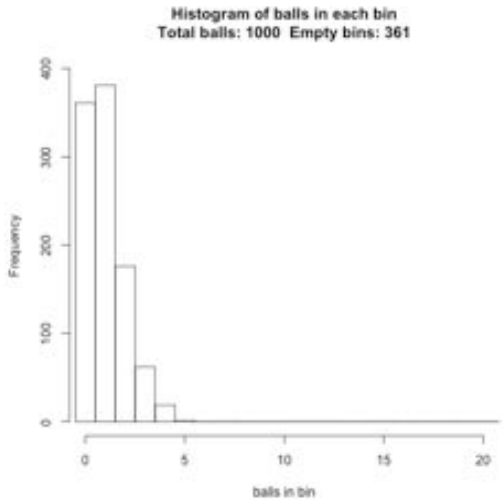
http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

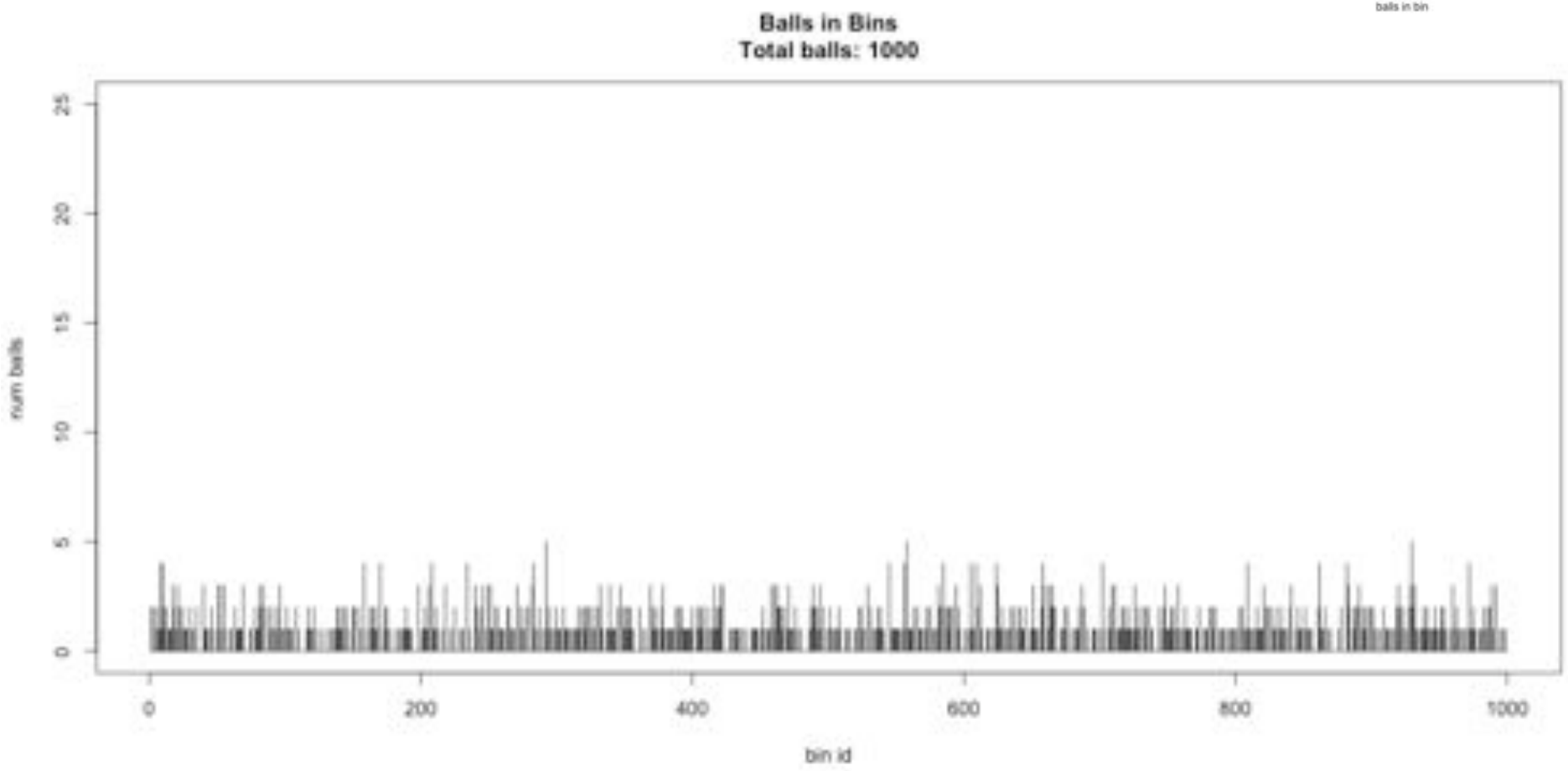
Typical contig coverage



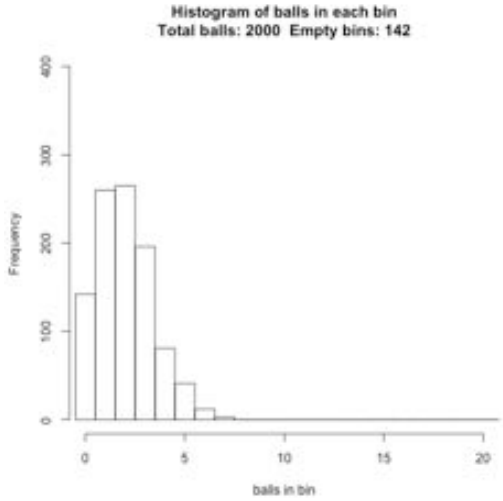
Imagine raindrops on a sidewalk

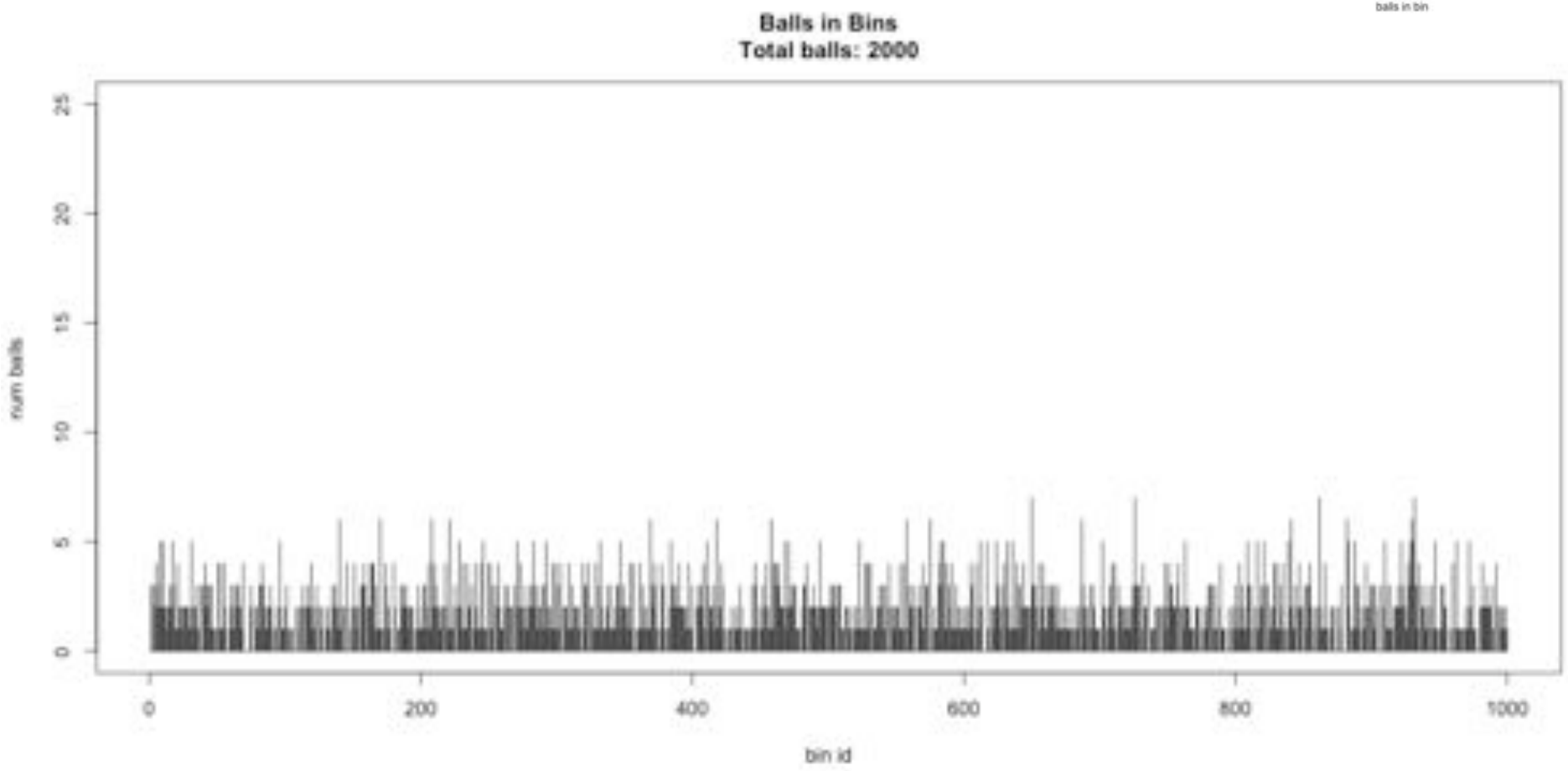
Balls in Bins 1x



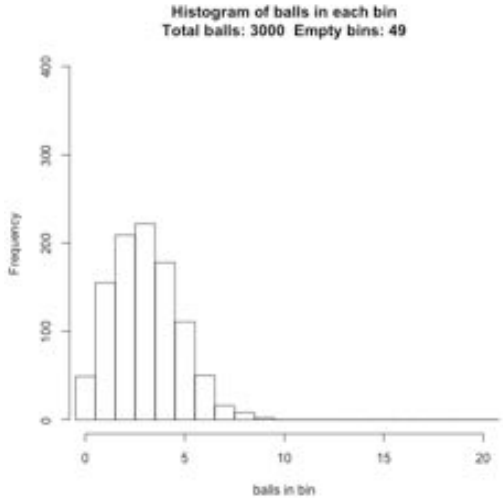


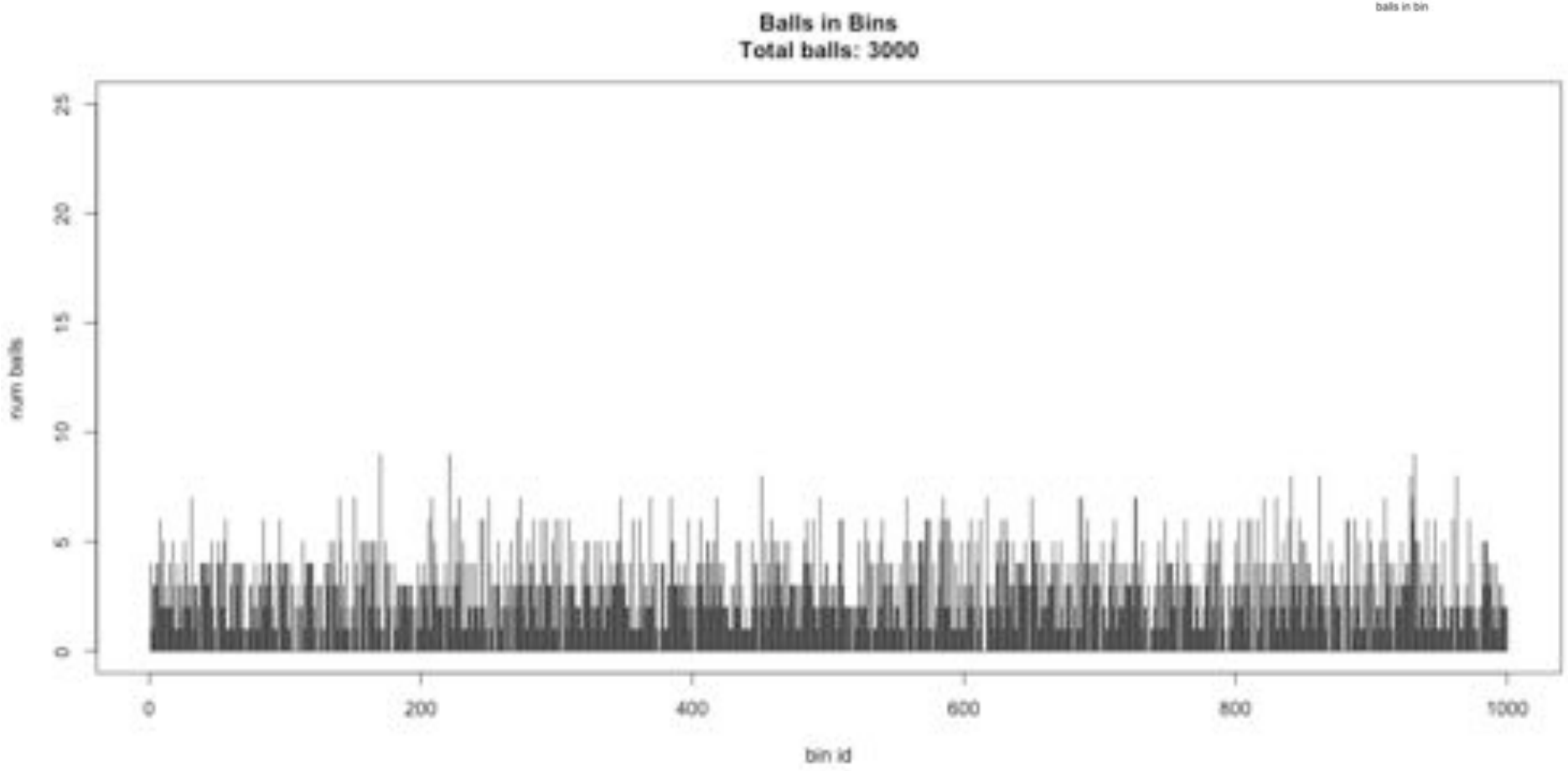
Balls in Bins 2x



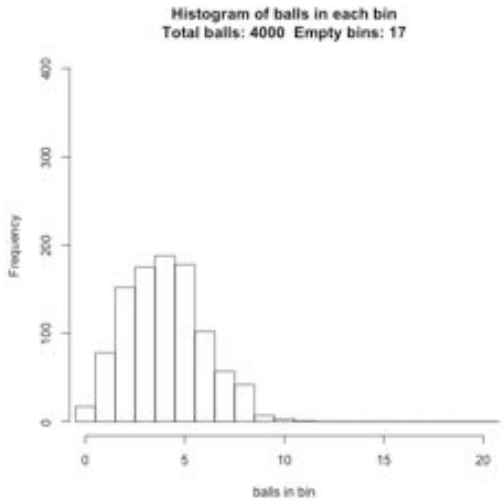


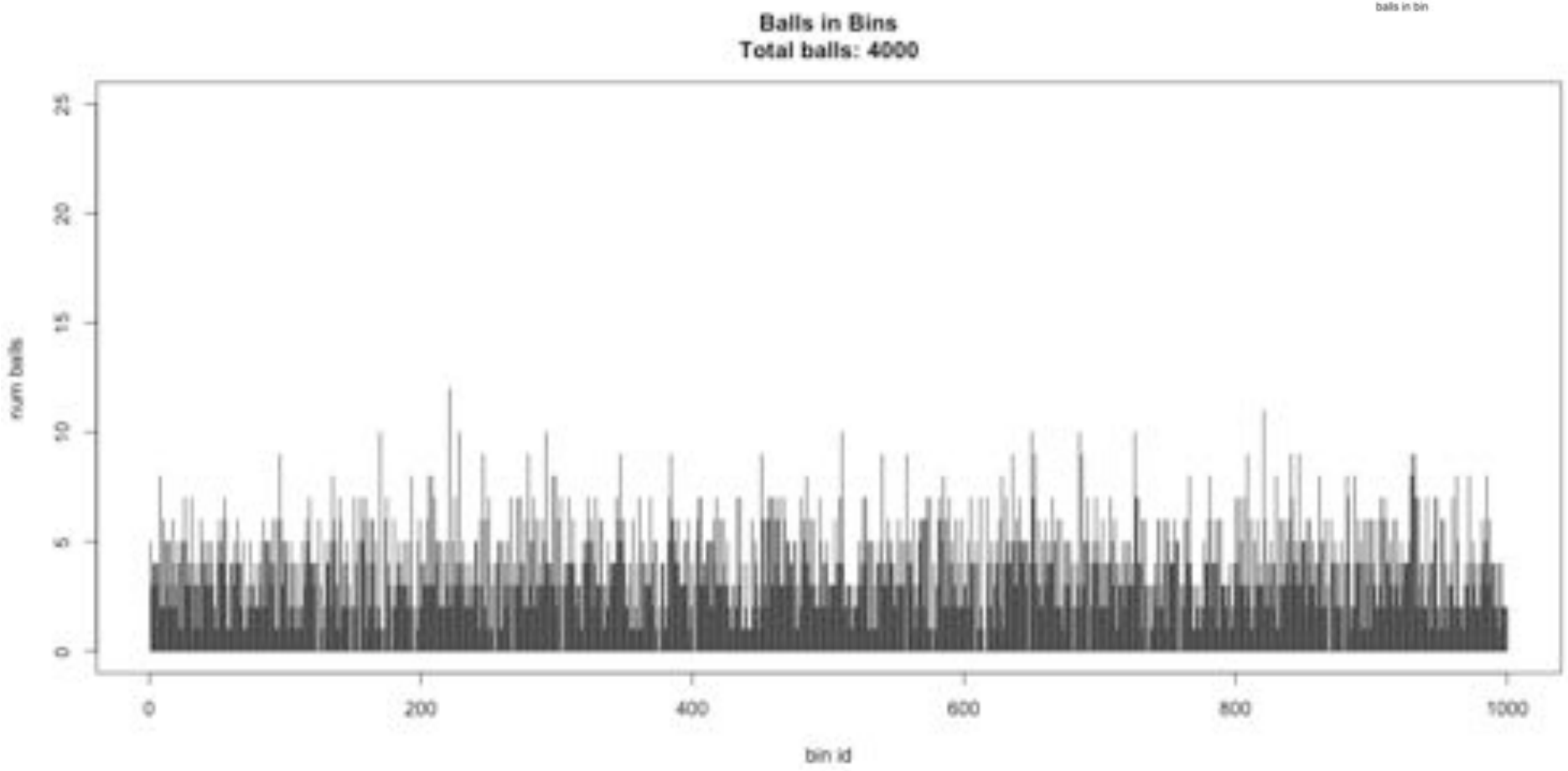
Balls in Bins 3x



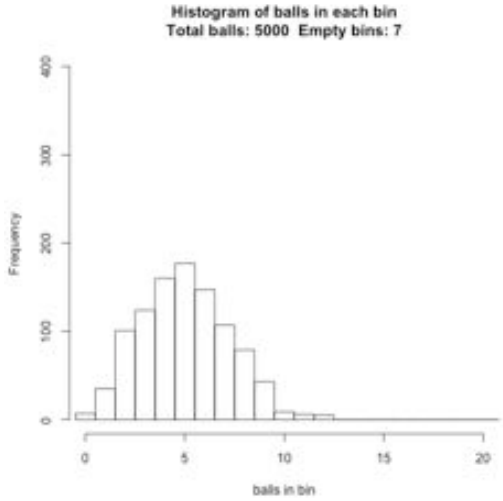


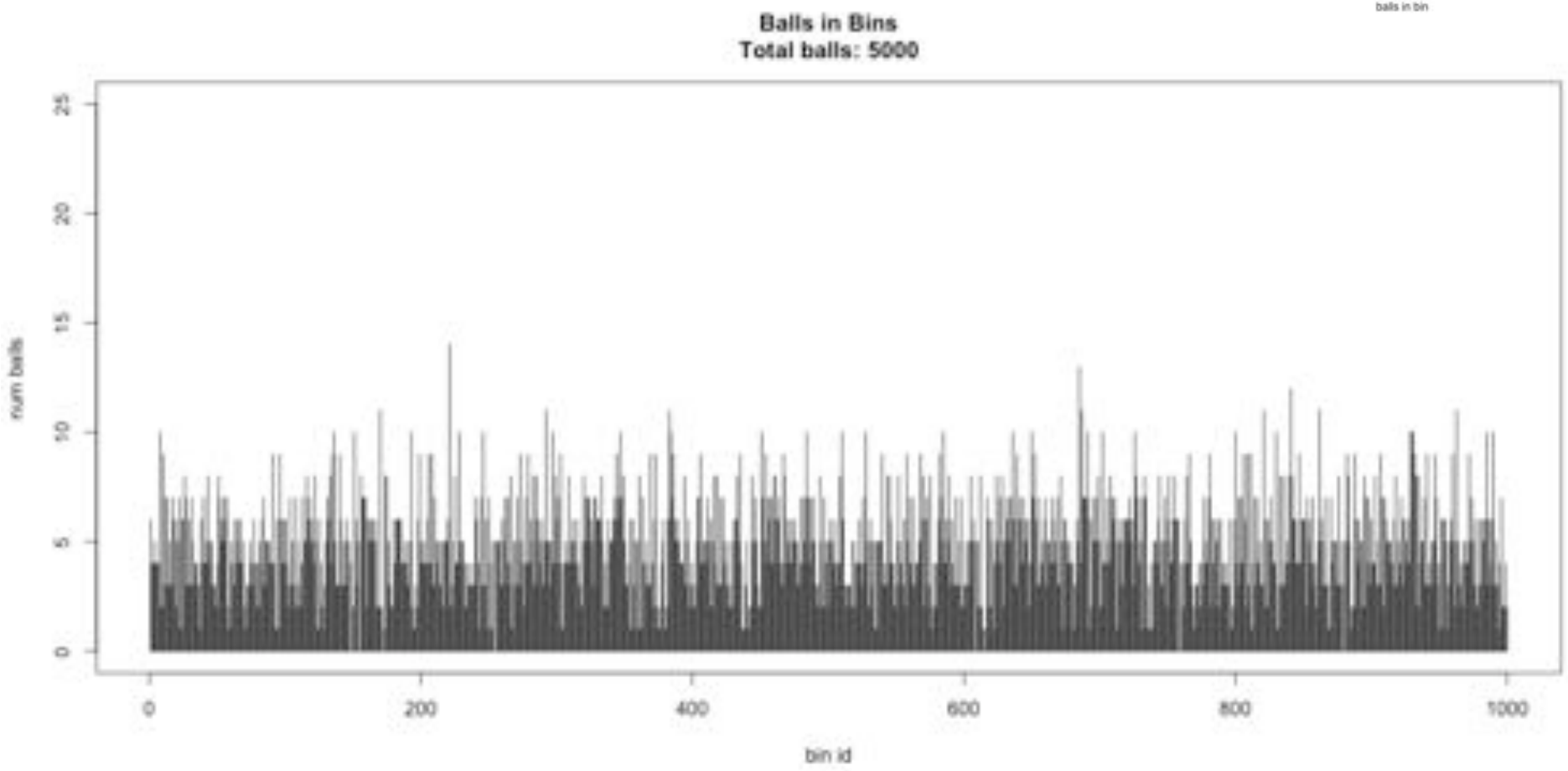
Balls in Bins 4x



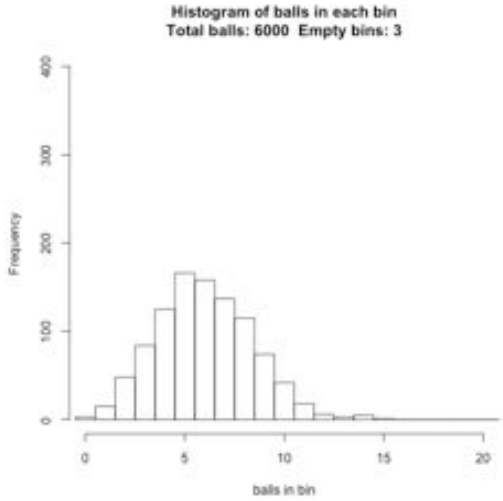


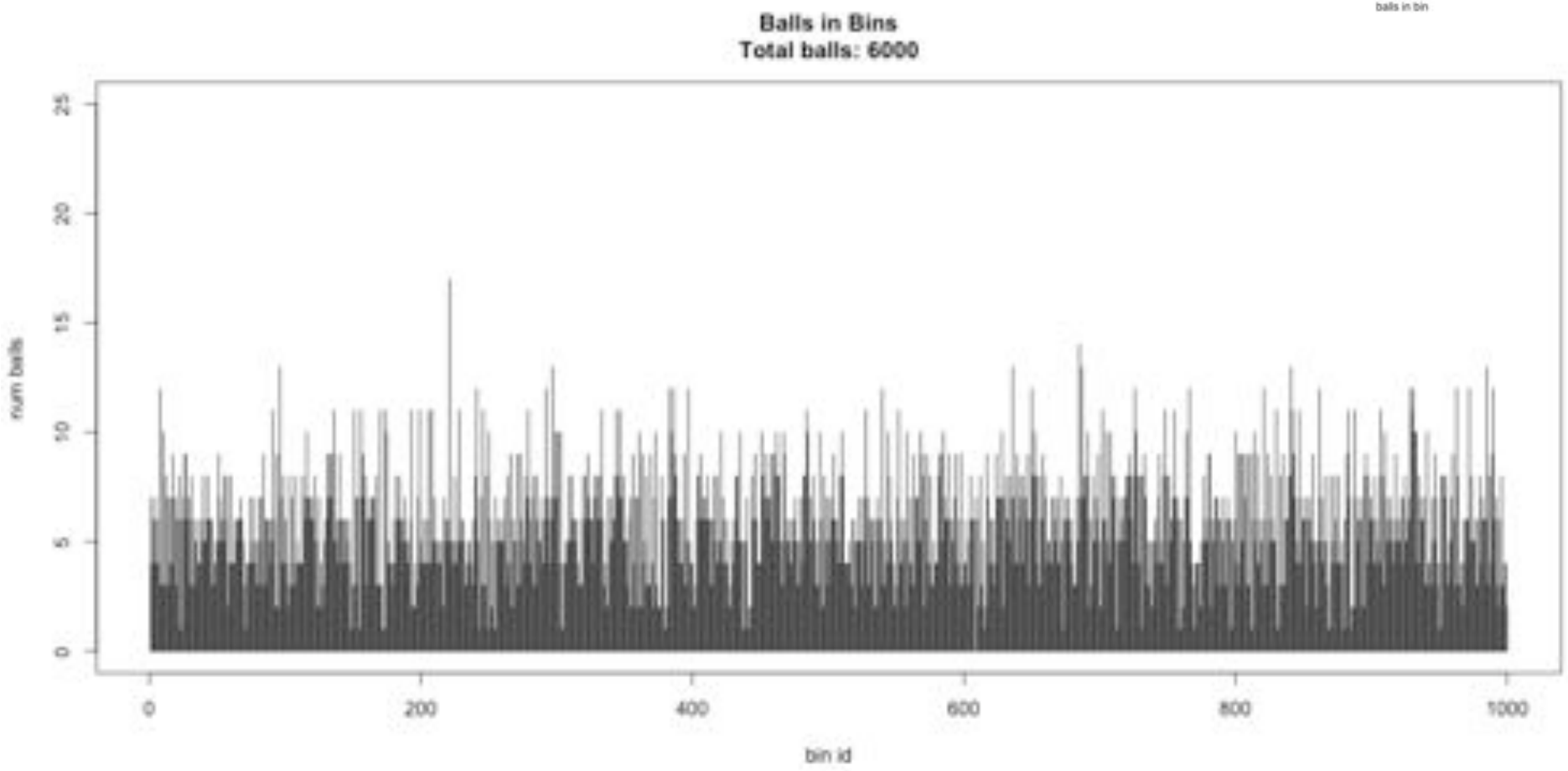
Balls in Bins 5x



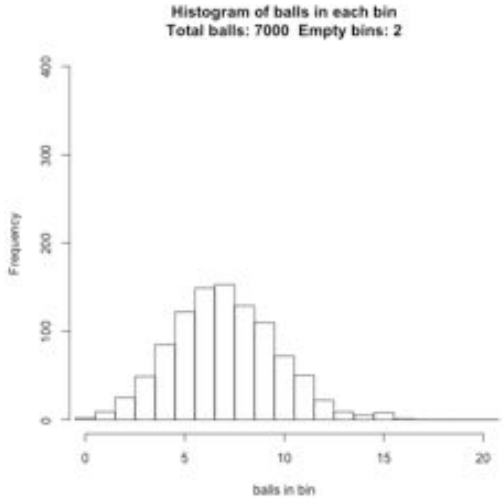


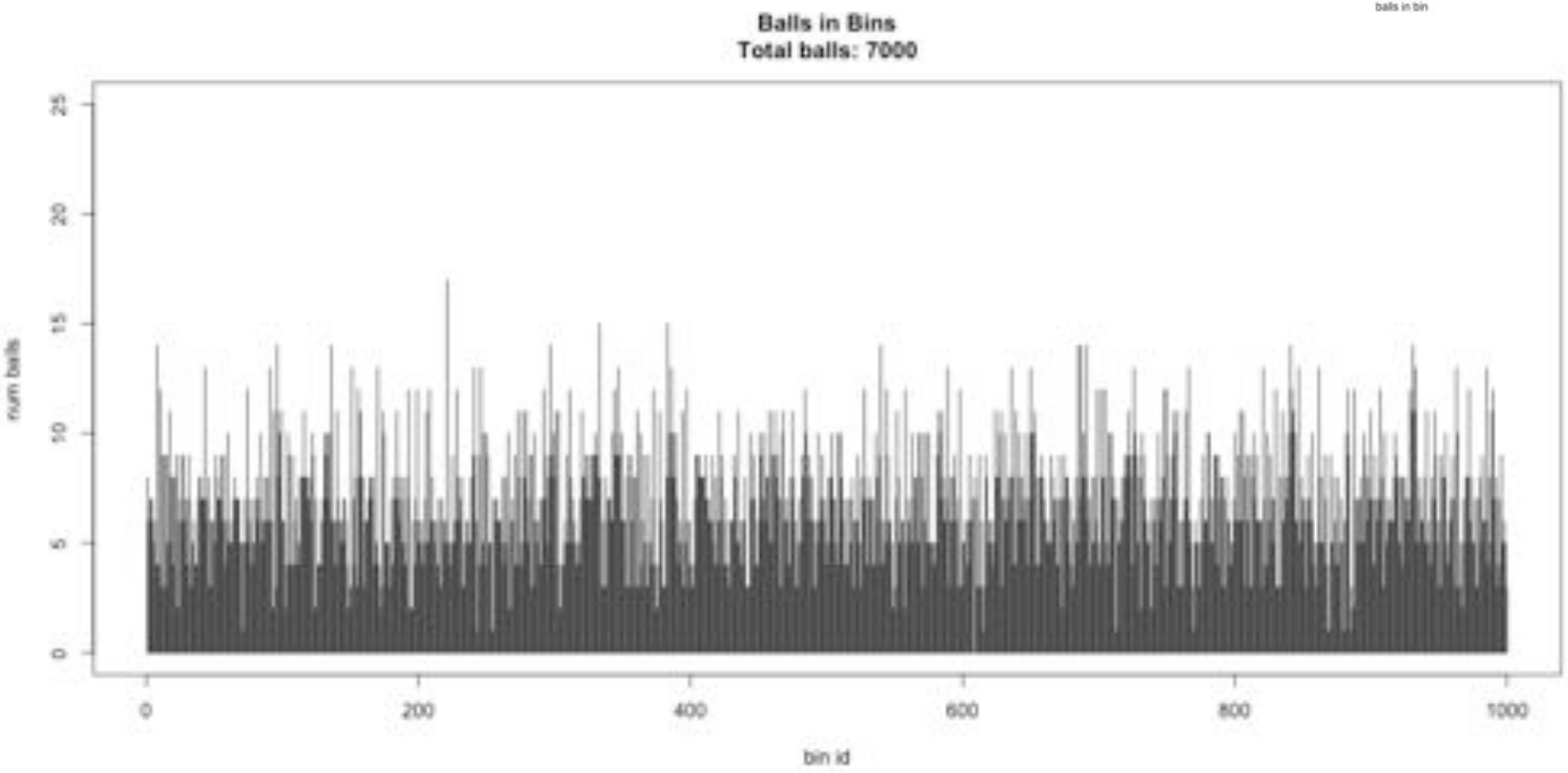
Balls in Bins 6x



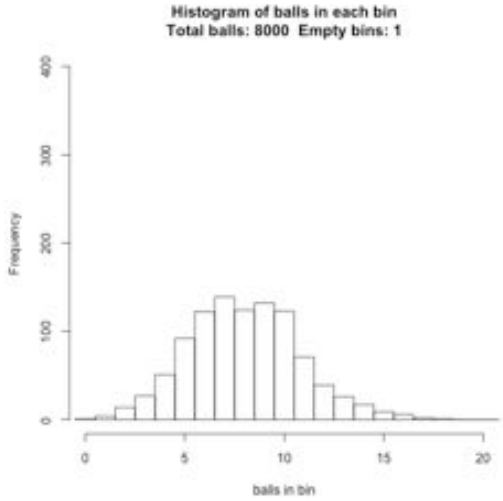


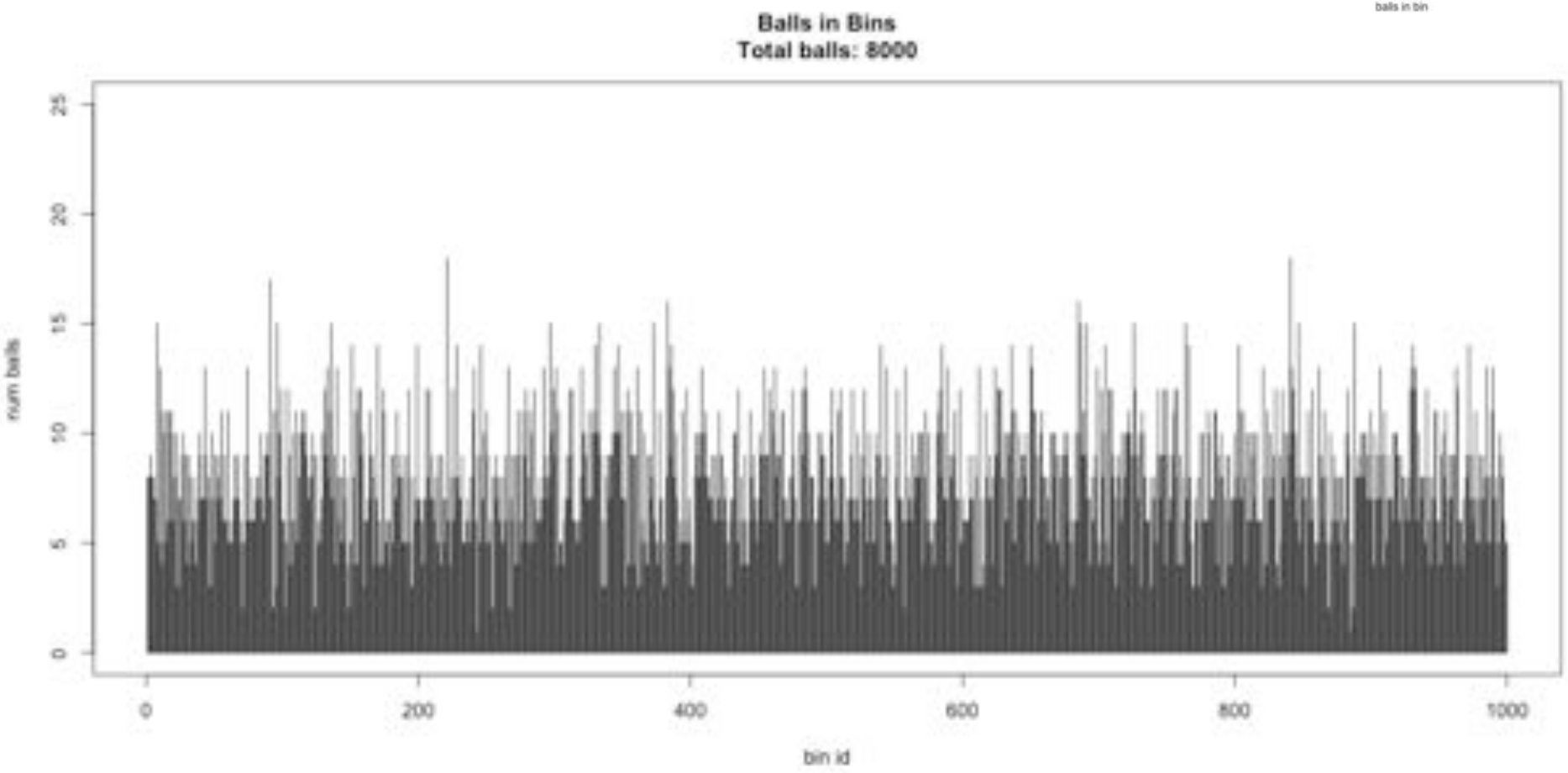
Balls in Bins 7x



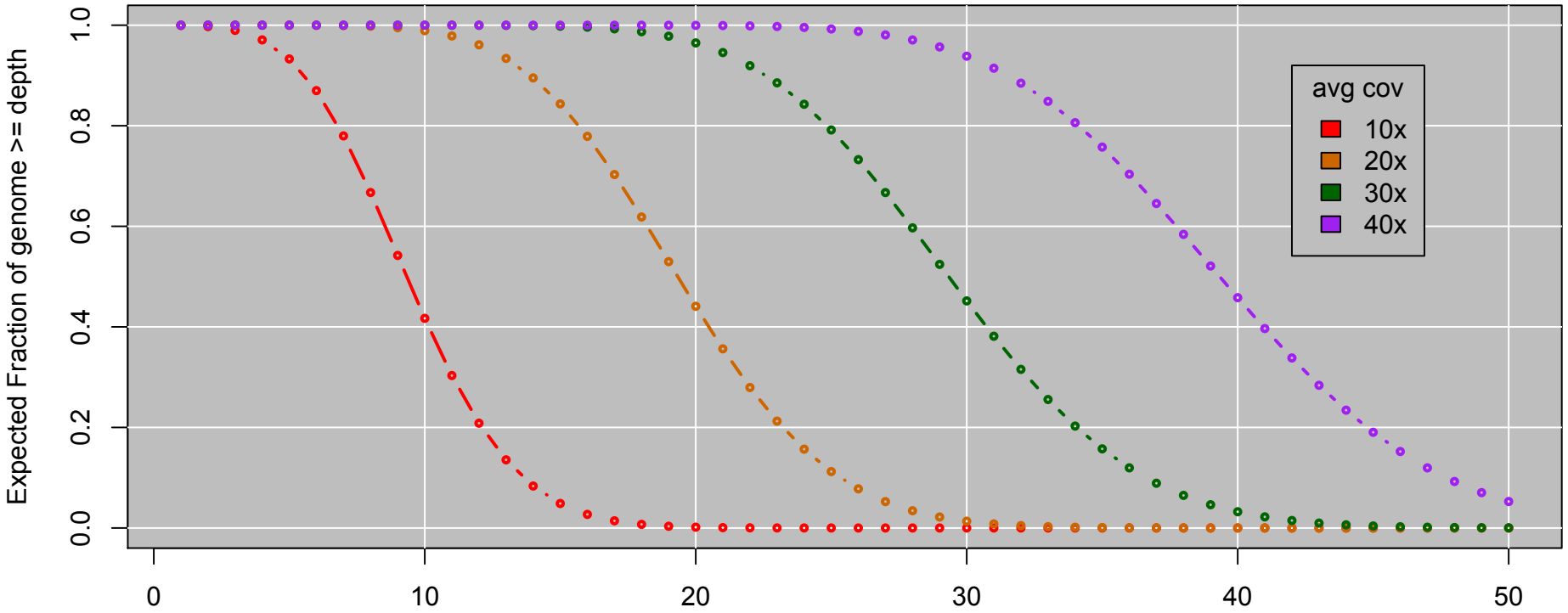


Balls in Bins 8x





Genome Coverage Distribution



Expect Poisson distribution on depth

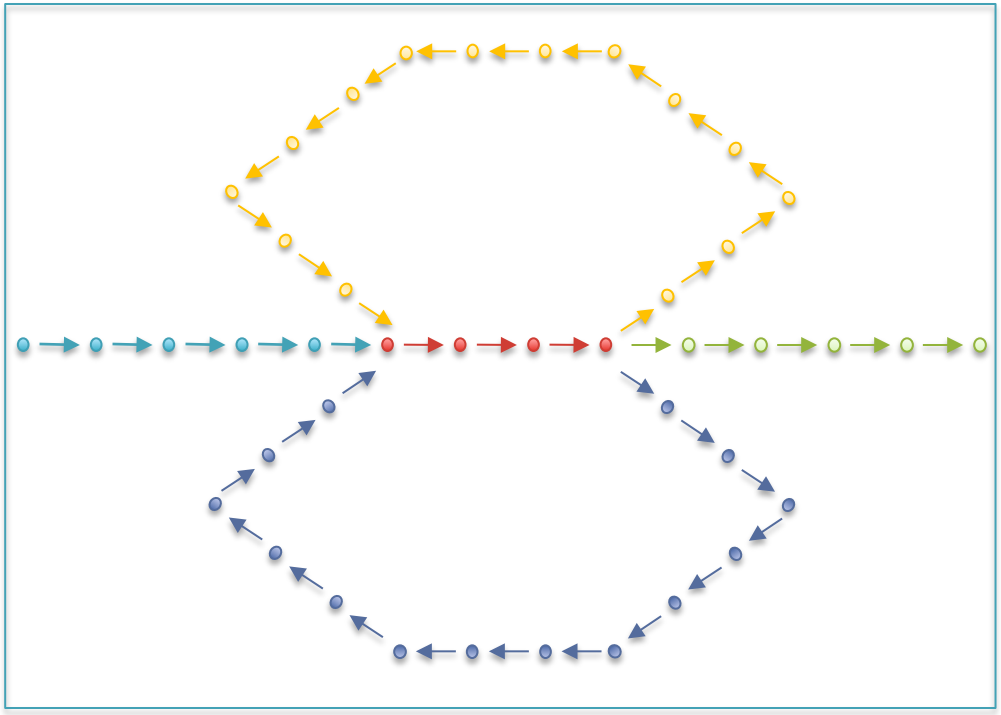
Standard Deviation = sqrt(cov)

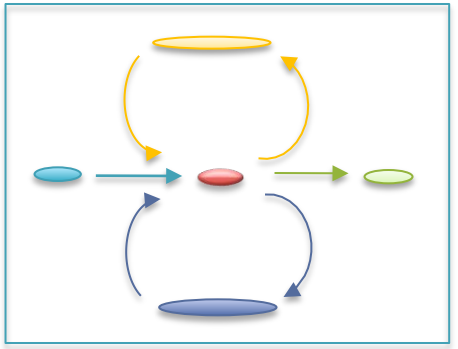
This is the mathematically model => reality may be much worse

Double your coverage for diploid genomes

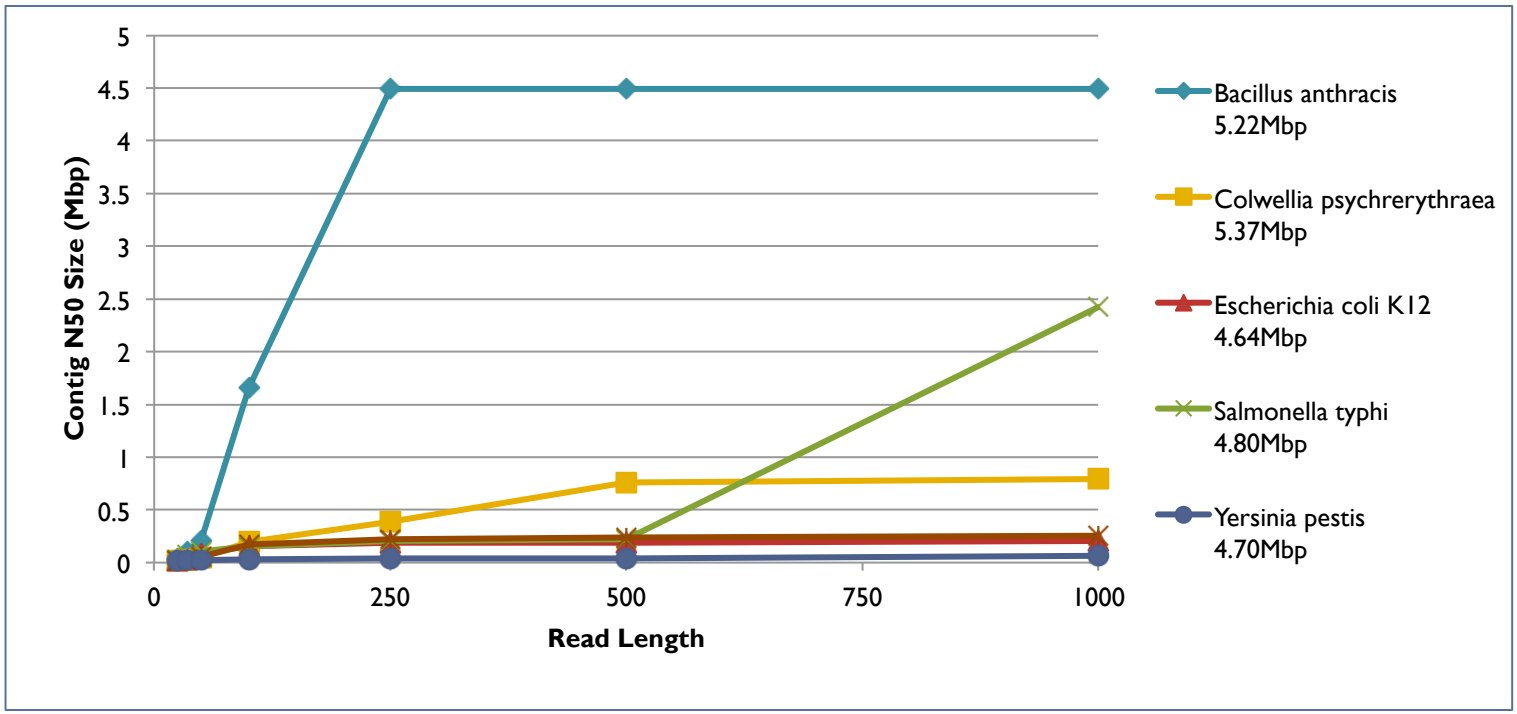
Initial Contigs

- 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

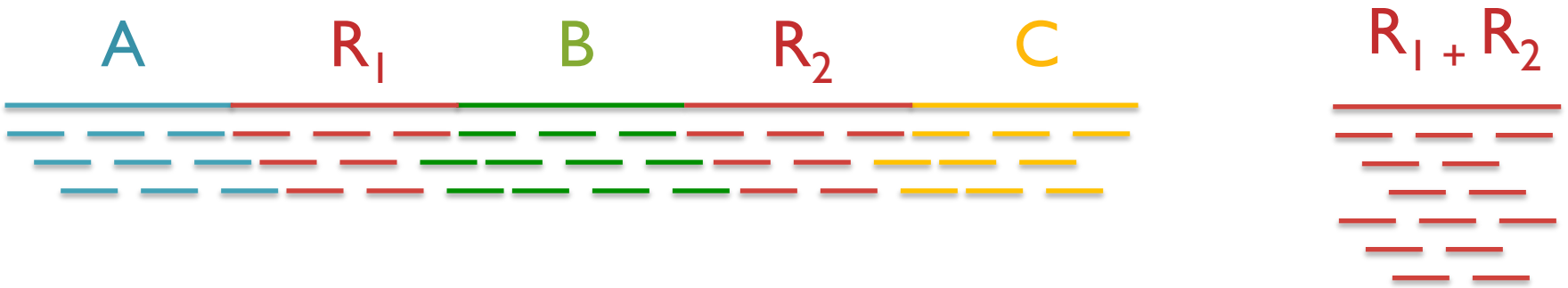
Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics*. 11:21.

Repetitive regions

- Over 50% of the human genome is repetitive

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

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 Δ.
  - 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

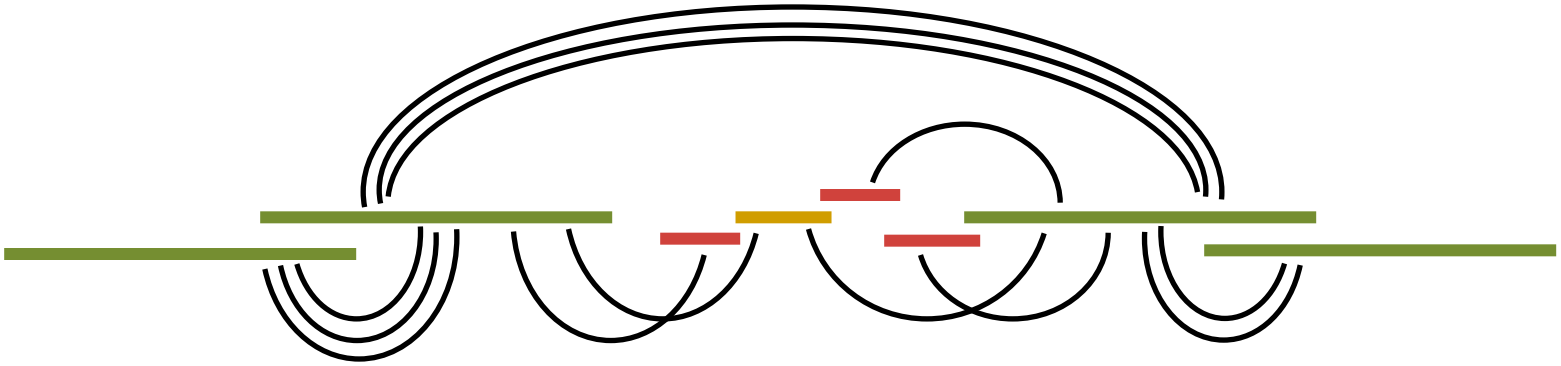
$$\Pr(X-copy)=\binom{n}{k}\left(\frac{X\Delta}{G}\right)^{k}\left(\frac{G-X\Delta}{G}\right)^{n-k}$$

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

Scaffolding

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

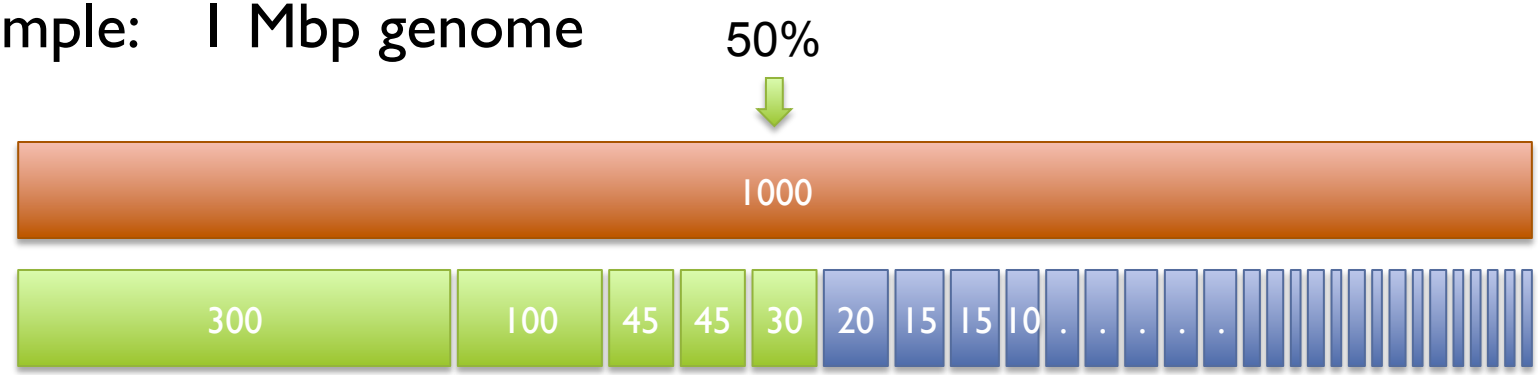
- Iteratively resolve longest, 'most unique' contigs
  - Both overlap graph and de Bruijn assemblers initially collapse repeats into single copies
  - Uniqueness measured by a statistical test on coverage



N50 size

Def: 50% of the genome is in contigs larger than N50

Example: 1 Mbp genome

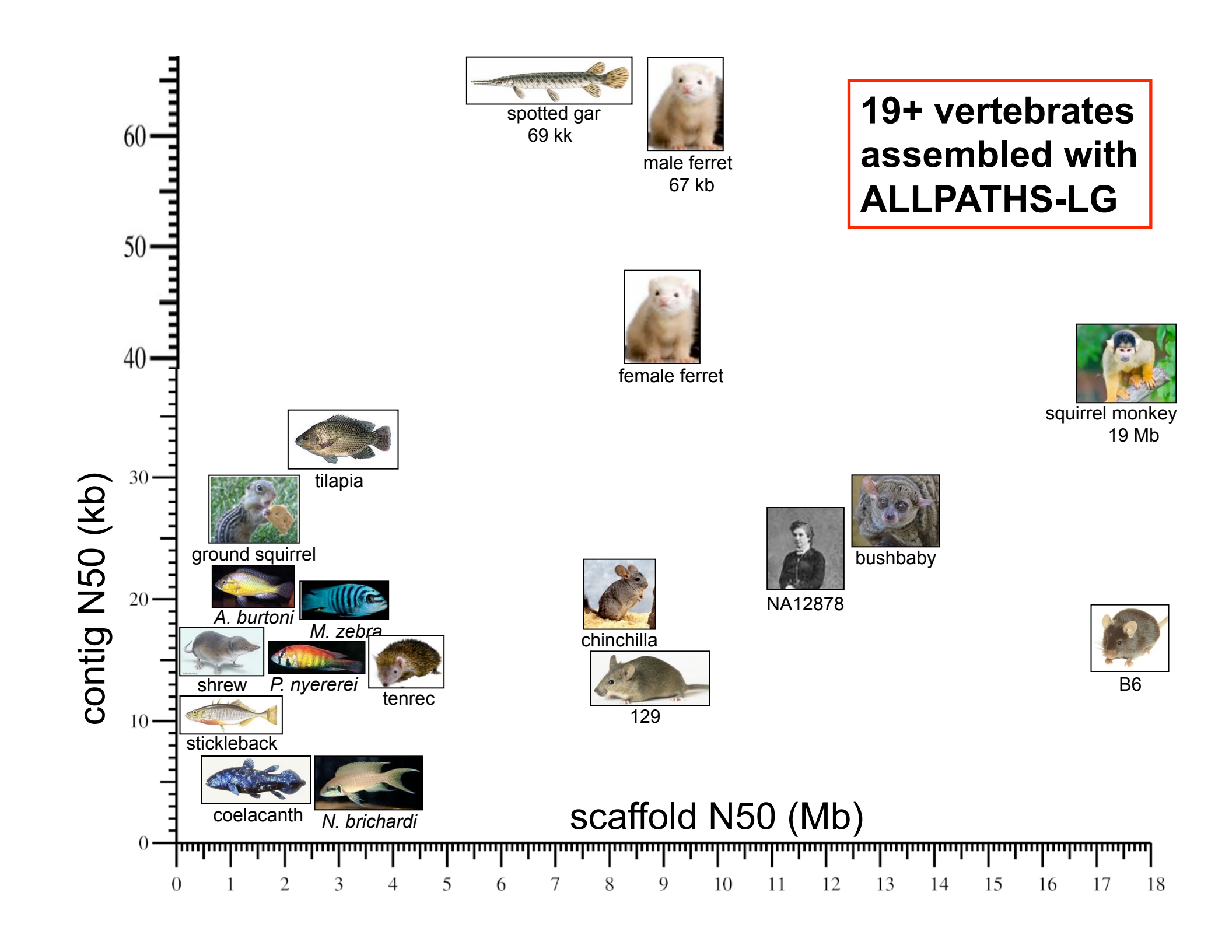


N50 size = 30 kbp

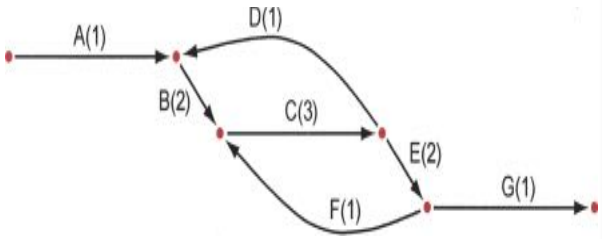
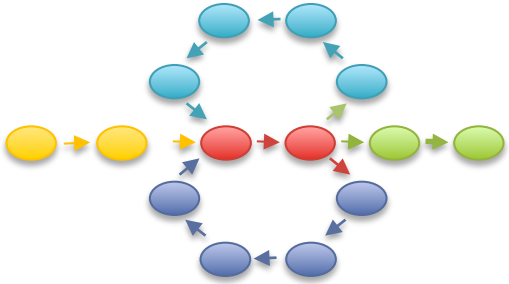
(300k+100k+45k+45k+30k = 520k >= 500kbp)

Note:

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


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

Assembly Algorithms

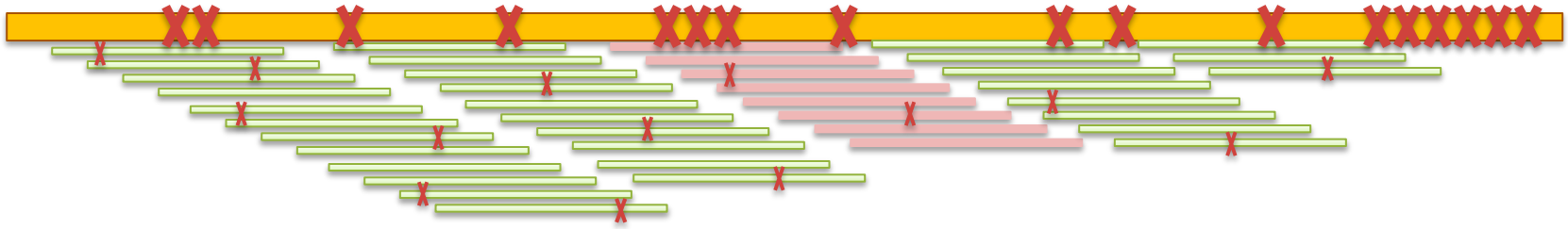
ALLPATHS-LG	SOAPdenovo	Celera Assembler
		
Broad's assembler (Gnerre et al. 2011)	BGI's assembler (Li et al. 2010)	JCVI's assembler (Miller et al. 2008)
De bruijn graph Short + PacBio (patching)	De bruijn graph Short reads	Overlap graph Medium + Long reads
Easy to run if you have compatible libraries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

PacBio Error Correction & Assembly

http://wgs-assembler.sf.net

1. 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, et al (2012) *Nature Biotechnology*. doi:10.1038/nbt.2280

Scalpel: Haplotype Microassembly

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

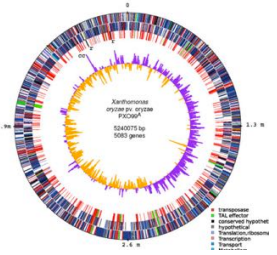
- Use assembly techniques to identify complex variations from short reads
  - Improved power to find indels
  - Trace candidate haplotypes sequences as paths through assembly graphs

```
Ref:     ...TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTTGCCCGGA...

Father:  ...TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTTGCCCGGA...
Mother:  ...TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTTGCCCGGA...
Sib:     ...TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTTGCCCGGA...
Aut(1):  ...TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTTGCCCGGA...
Aut(2):  ...TCAGAACAGCTGGATGAGATCTTACC------CCGGAGATTGTCTTTGCCCGGA...
```

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

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
Piyush Kansal
Greg Vurture
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/teaching/
@mike_schatz