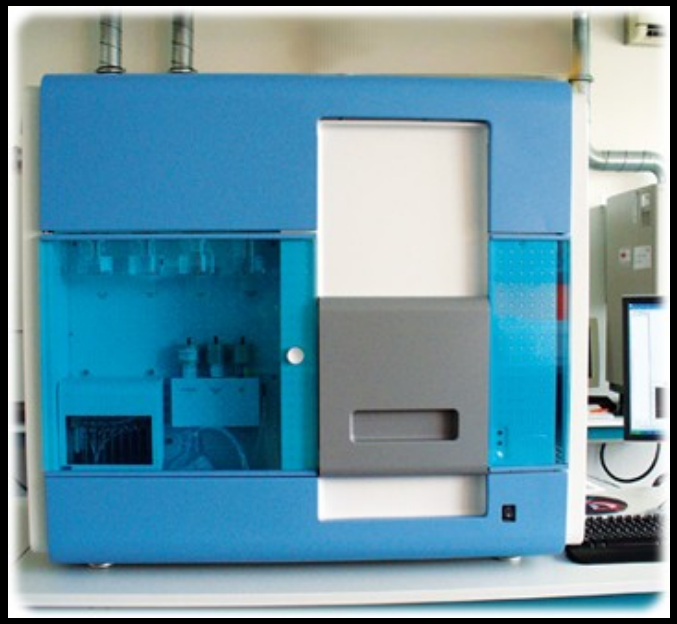
ALLPATHS: *de novo* assembly of whole genome micro-reads

by Butler *et al.*

Presented by Tim Smith
CSC2431 2008/03/12

NGS data presents new challenges and opportunities

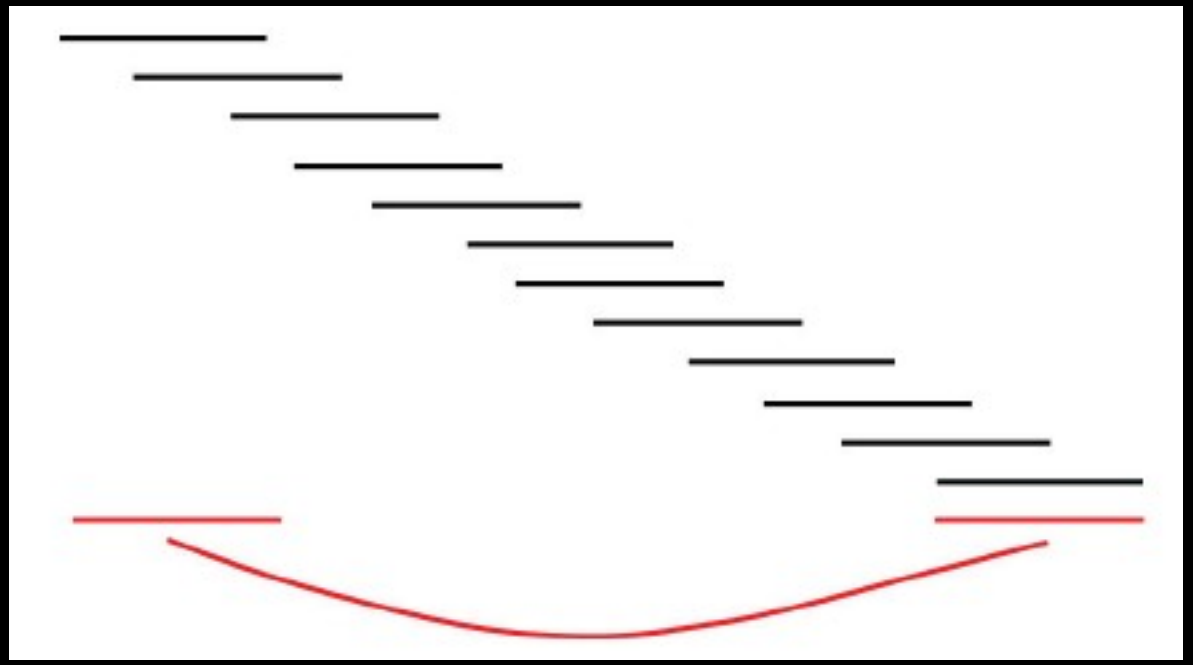


“Find all overlaps” is not adequate for NGS data

K	*E. coli*	*S. cerevisiae*	*A. thaliana*	*H. sapiens*
200	0.063	0.26	0.053	0.18
160	0.068	0.31	0.064	0.49
120	0.074	0.39	0.086	1.7
80	0.082	0.49	0.15	7.2
60	0.088	0.58	0.27	18
50	0.091	0.63	0.39	32
40	0.095	0.69	0.65	78
30	0.11	0.77	1.5	330
20	0.15	1.0	5.7	2100
10	18	63.8	880	40000

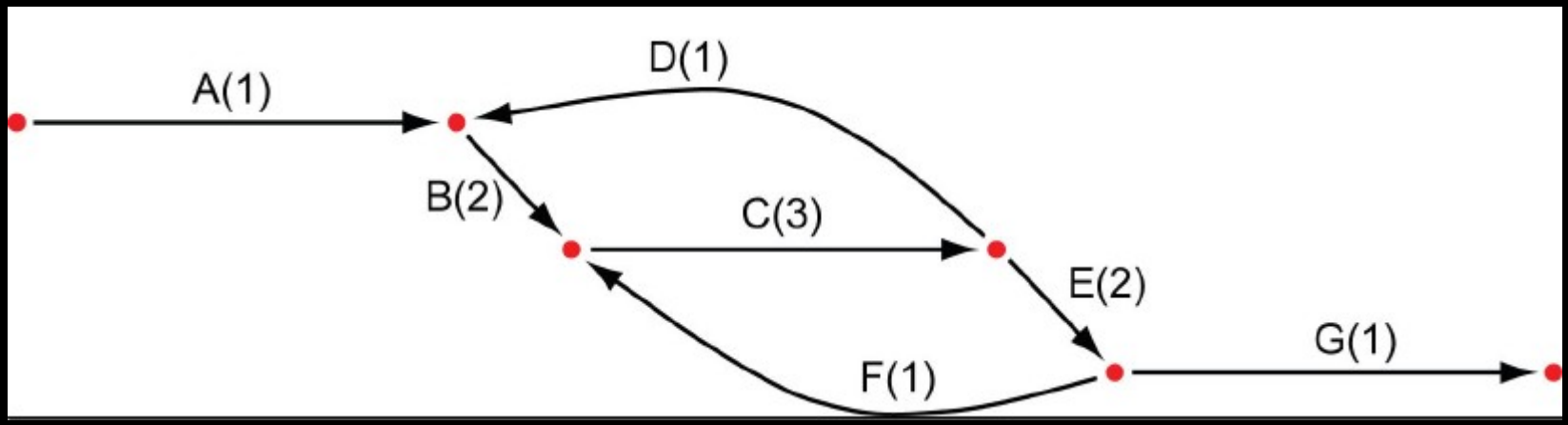
Mean number of false placements of K-mers

ALLPATHS finds all *paths* across read pairs



Gaps in read pairs are “walked” from one read to the other by filling in the gap with overlapping reads

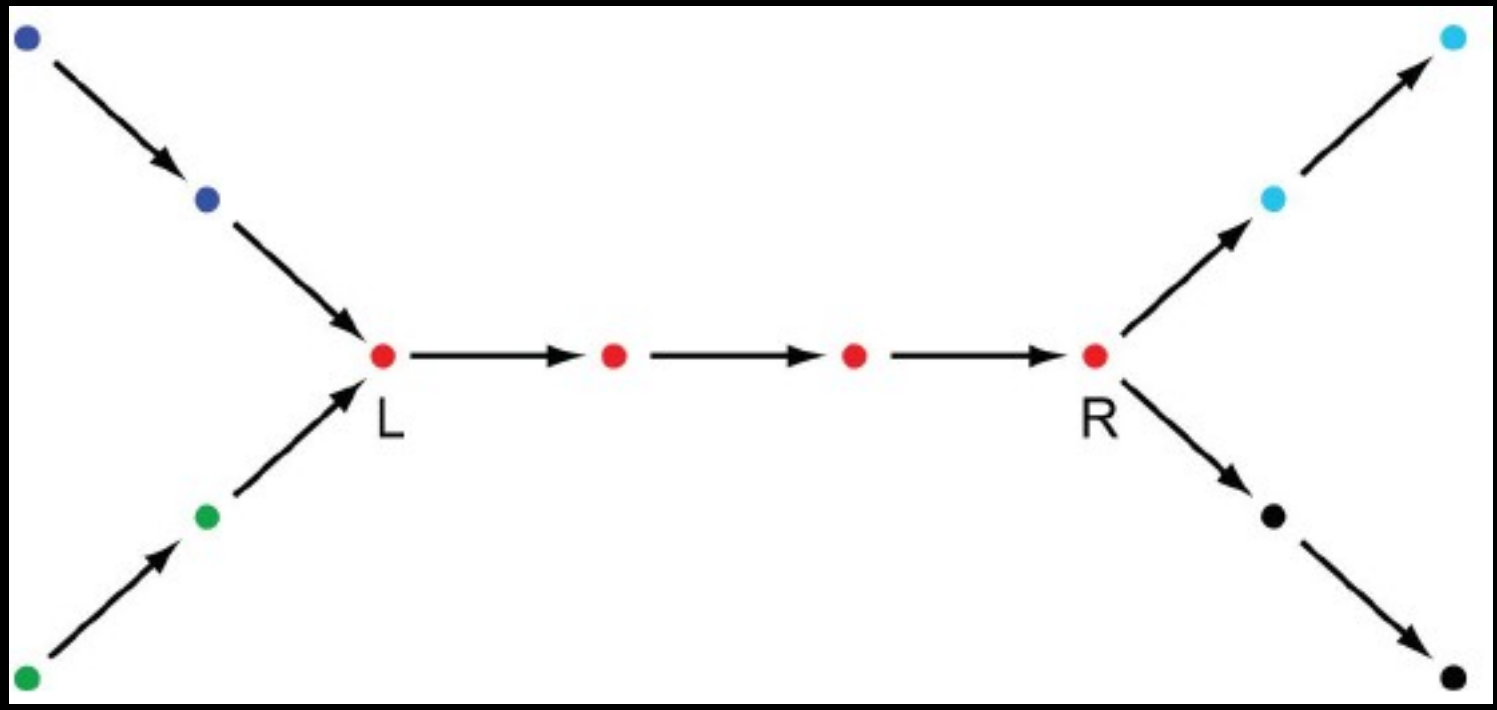
ALLPATHS introduces the concept of *unipath graphs*



Sequence graph of *C. jejuni* with K = 6000 bases

Two valid paths: ABCDBCEFCEG and ABCEFCDBCEG

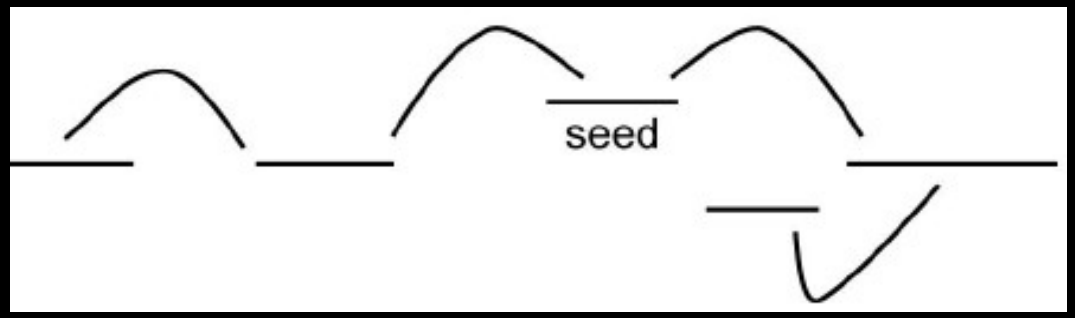
ALLPATHS finds approximate unipaths between read pairs



Unipaths with low copy number become seeds

- Ideally, seeds are long and unique
- Copy number is inferred from read coverage of unipath components
- Read pairing is used to optimize seed selection

“Neighborhoods” are built around seeds



Unipaths assigned coordinates relative to the seed

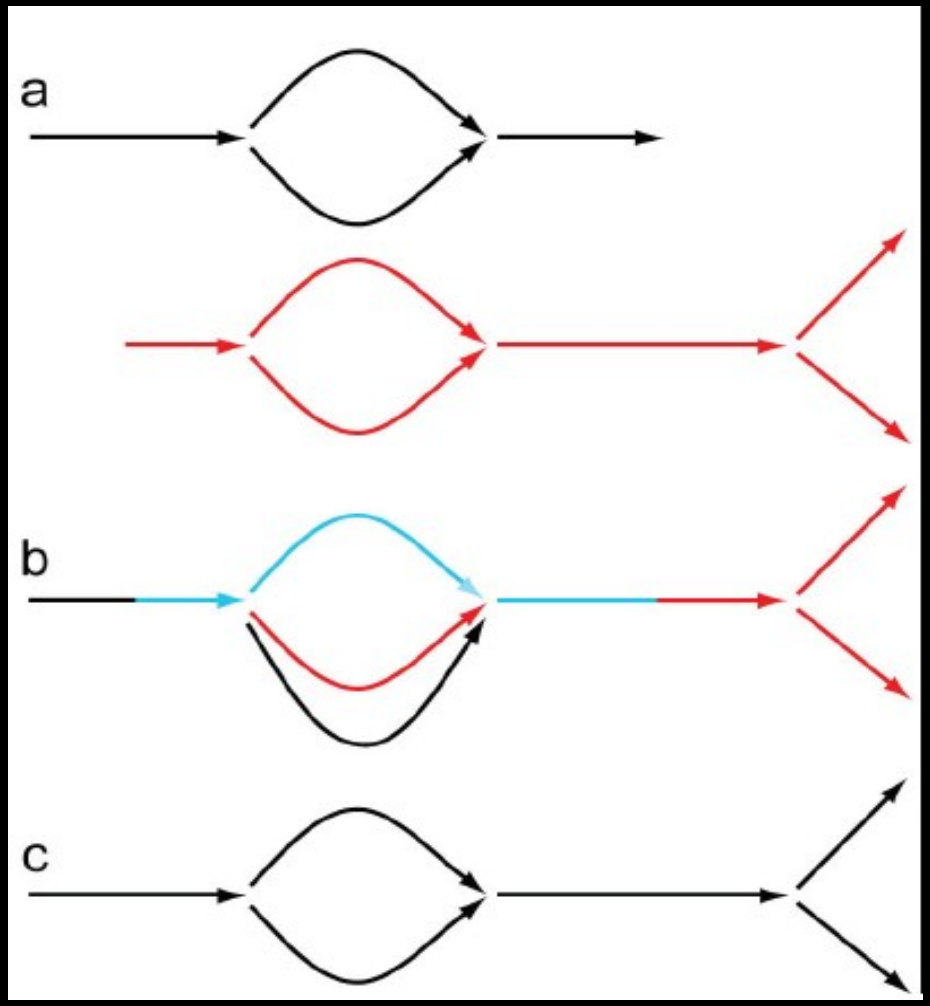
Read “partners” added to primary cloud

Repetitive read pairs are placed in the secondary cloud

All paths between merged short-fragment pairs are found

- Paths between merged short-fragment pairs are computed
- Resulting set of paths covers neighborhood
- Paths are then used as reads to walk mid-length (~5 kb) read pairs from the primary read cloud

Local assemblies are glued together

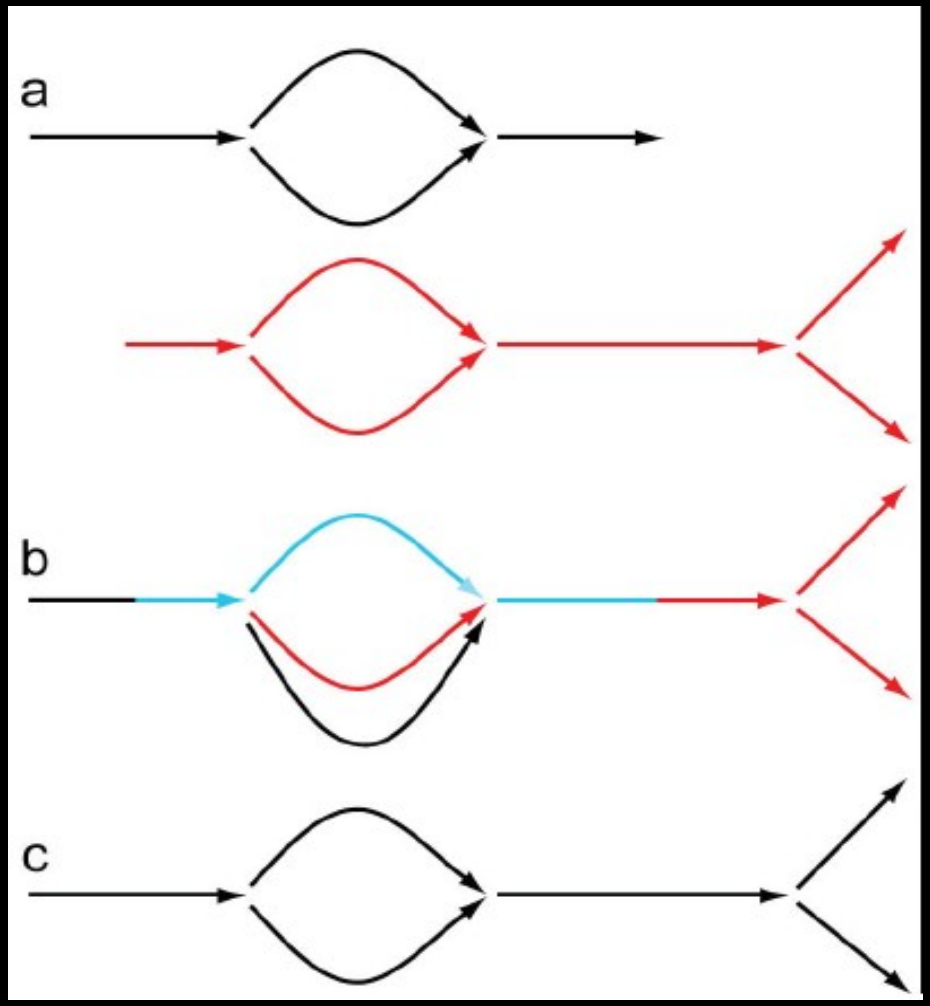


(a) Sequences around bubble match

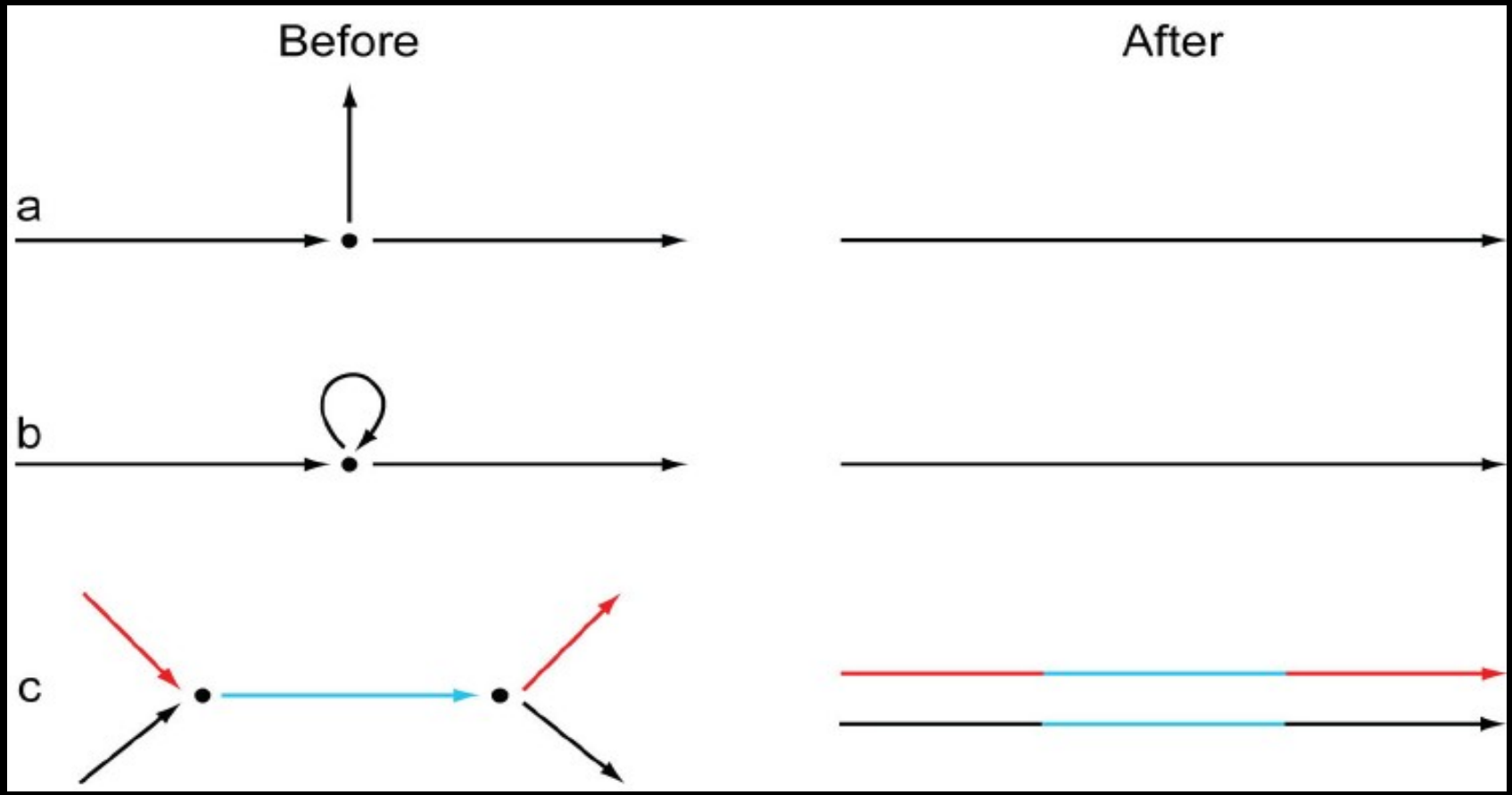
(b) Common path identified

(c) Edges “zipped up”

The global assembly is glued together



The global assembly is edited



Evaluation was performed using “simulated short reads”

- Ten reference genomes (2-39 Mb)
- 10Mb segment of reference human genome
- Segmented into 30 base “reads”
  - 1X coverage from long fragments (~50 kb)
  - 39.5X from medium fragments (~6 kb)
  - 39.5X from short fragments (~500 bases)
  - Total of 80X coverage

The results were promising

INPUTS				OUTPUTS				
Species	**Ploidy**	**Genome size (kb)**	**Reference N50 (kb)**	**Component N50 (kb)**	**Edge N50 (kb)**	**Ambiguities per Mb**	**Coverage (%)**	**Coverage by perfect edges ≥ 10 kb (%)**
C. jejuni	1	1,800	1,800	1,800	1,800	0.0	100.0	100.0
E. coli	1	4,600	4,600	4,600	4,600	0.0	100.0	100.0
B. thailandensis	1	6,700	3,800	1,800	890	2.7	99.8	99.5
E. gossypii	1	8,700	1,500	1,500	890	2.6	100.0	99.9
S. cerevisiae	1	12,000	920	810	290	28.7	98.7	94.9
S. pombe	1	13,000	4,500	1,400	500	19.1	98.8	97.5
P. stipitis	1	15,000	1,800	900	700	8.6	97.9	96.3
C. neoformans	1	19,000	1,400	810	770	4.5	96.4	93.4
Y. lipolytica	1	21,000	3,600	2,200	290	6.2	99.1	98.6
N. crassa	1	39,000	660	640	90	17.4	97.0	92.5
H. sapiens region	2	10,000	10,000	490	2	68.2	97.3	0.2

ALLPATHS accuracy is still unknown

- Comparisons were against “reference” genomes
- No “coverage bias” in simulated reads
- Is ALLPATHS actually accurate, or just biased in the same way as Sanger?

Evaluation was also performed with “artificially paired” Solexa reads”

- 36 base *E. coli* Solexa reads mapped to reference genome
- Reads paired in same 80X coverage distribution as above
- Simulated error as a result in error in fragment length

Performance with real data was slightly worse

- ALLPATHS produced assembly of 58 components, with 99.1% coverage
- Components were ordered and oriented using read pair information to produce a single contiguous sequence
- Assembled sequence matches reference except in 12 locations

The performance on real paired read data is unknown

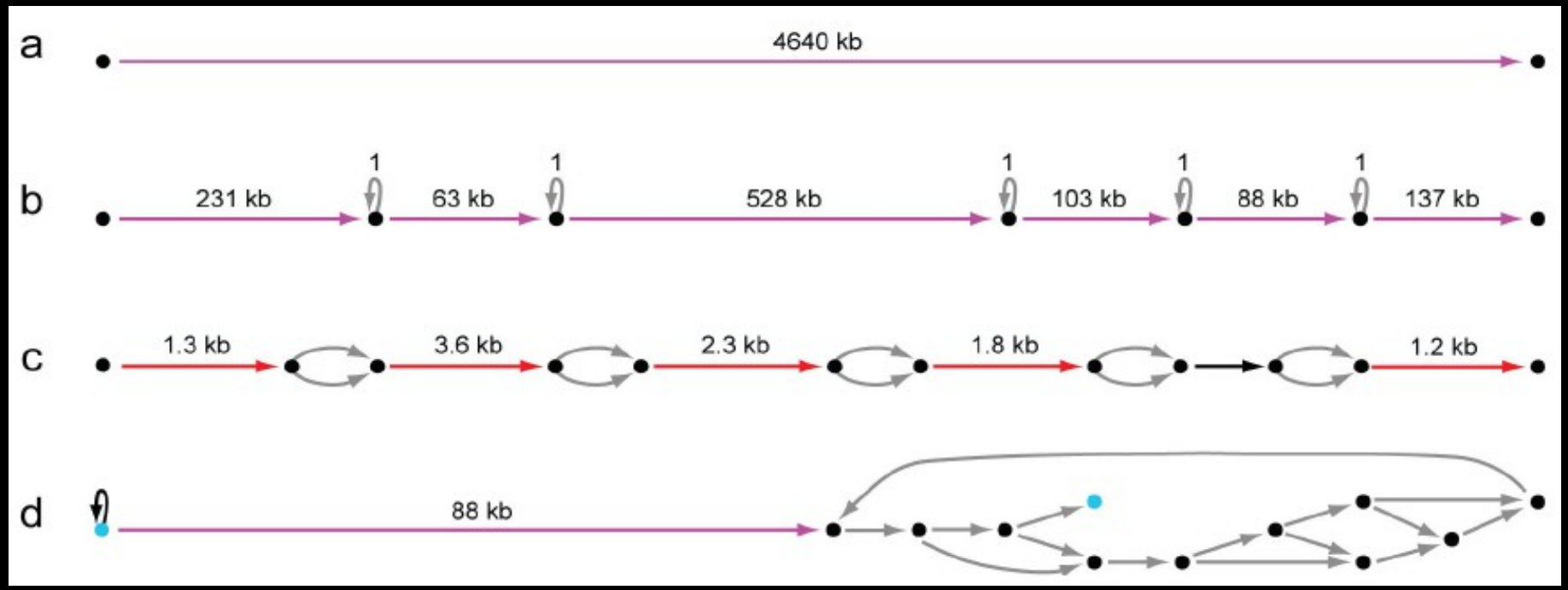
- Same problems with “simulated data” evaluation
- Bias in fragment size “error”?
- Lack of read error information

Variance in fragment size can cause “closure explosion”

	walk using entire genome		walk within 20 kb region	
	500 ± 1 %	500 ± 10 %	500 ± 1 %	500 ± 10 %
closures found	% of pairs	% of pairs	% of pairs	% of pairs
0	0.19	0.29	0.20	0.22
1	94.3	93.3	98.7	98.3
2	1.17	1.07	0.30	0.29
3 to 5	1.21	1.06	0.41	0.33
6 to 9	0.91	0.74	0.14	0.17
10^1 -	1.32	1.28	0.22	0.51
10^2 -	0.58	0.36	0.03	0.15
10^3 -	0.12	0.62	0	0.05
10^4 -	0.12	0.58	0	0
10^5 -	0.06	0.43	0	0
10^6 -	0.04	0.19	0	0
10^7 -	0.003	0.07	0	0

Number of read pair closures in E. coli using 30-base reads and K = 20

Unipath graphs offer a compact and informative representation of sequence components



Questions?