CSC 334: TOPICS IN COMPUTATIONAL BIOLOGY

"Algorithms for Genomic Data"

Fall 2015

Smith College

Instructor: Prof. Sara Sheehan

Outline: 9/18

- Velvet paper

Assembly vocabulary

- **Long read**: a fragment that has been "read" from a genomic sequence (DNA for us), usually > 1000 bp
- **Short read**: same as a long read but usually < 1000 bp
- **Paired-end read**: both ends of a fragment are "read", but the portion between them is unknown
- **bp**: base pair
- **kb, Mb, Gb**: kilo bases 10^3, mega bases 10^6, giga bases 10^9

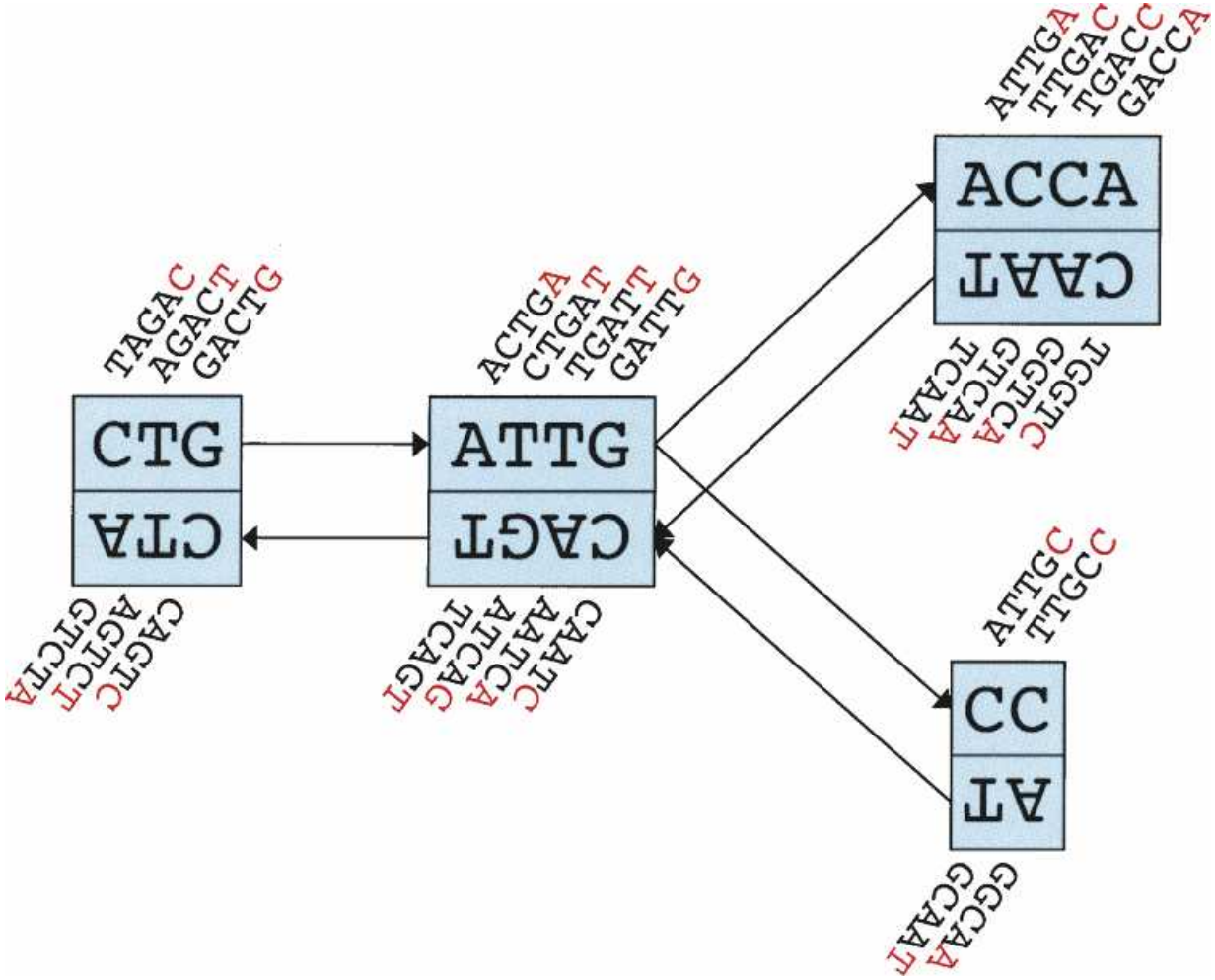
Assembly vocabulary (cont)

- **Coverage**: total number of bases in all reads, divided by the length of the genome (short reads: need higher coverage)

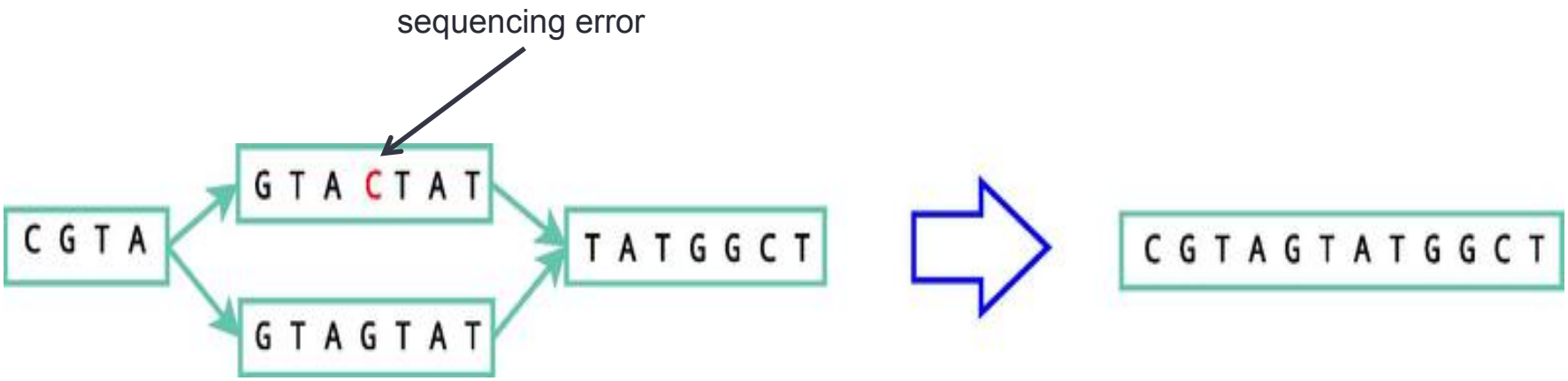
- **k-mer**: a genomic "word" of length *k*

- **N50**: for a set of contigs, N50 is the length such that at least half the bases of the assembly are in a contig with length N50 or longer, and at least half the bases are in a contig with length N50 or shorter

Velvet de Bruijn graph

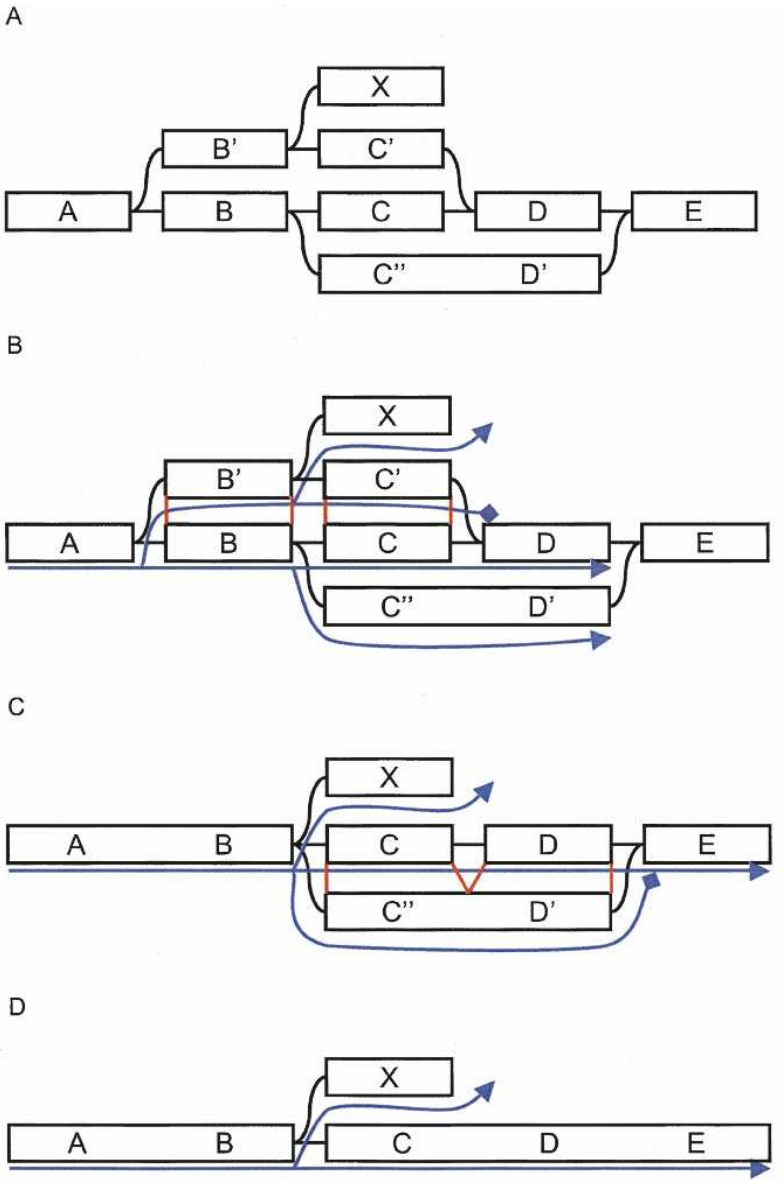


Bubbles



Velvet Assembler, Wikipedia

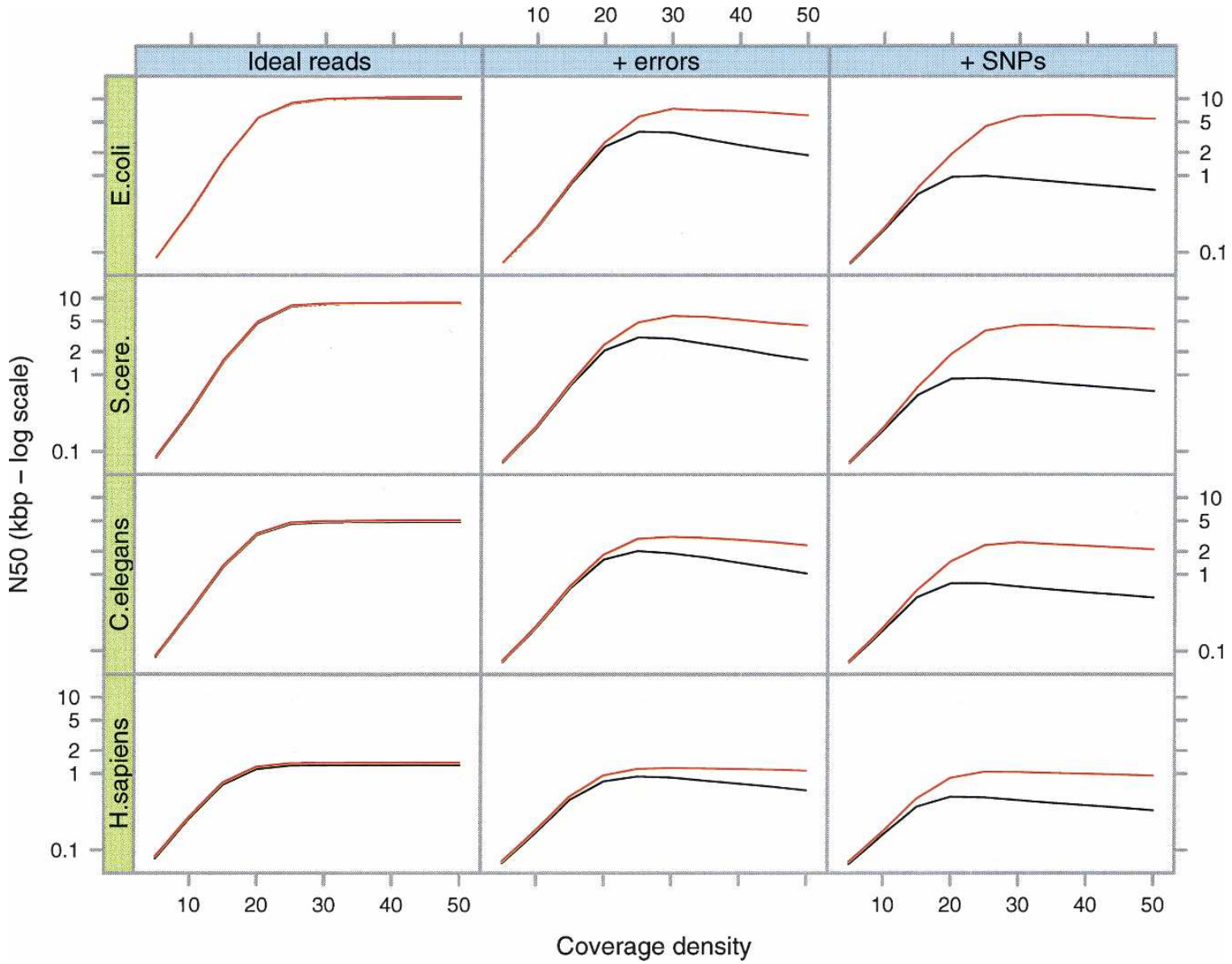
Velvet: Tour Bus algorithm



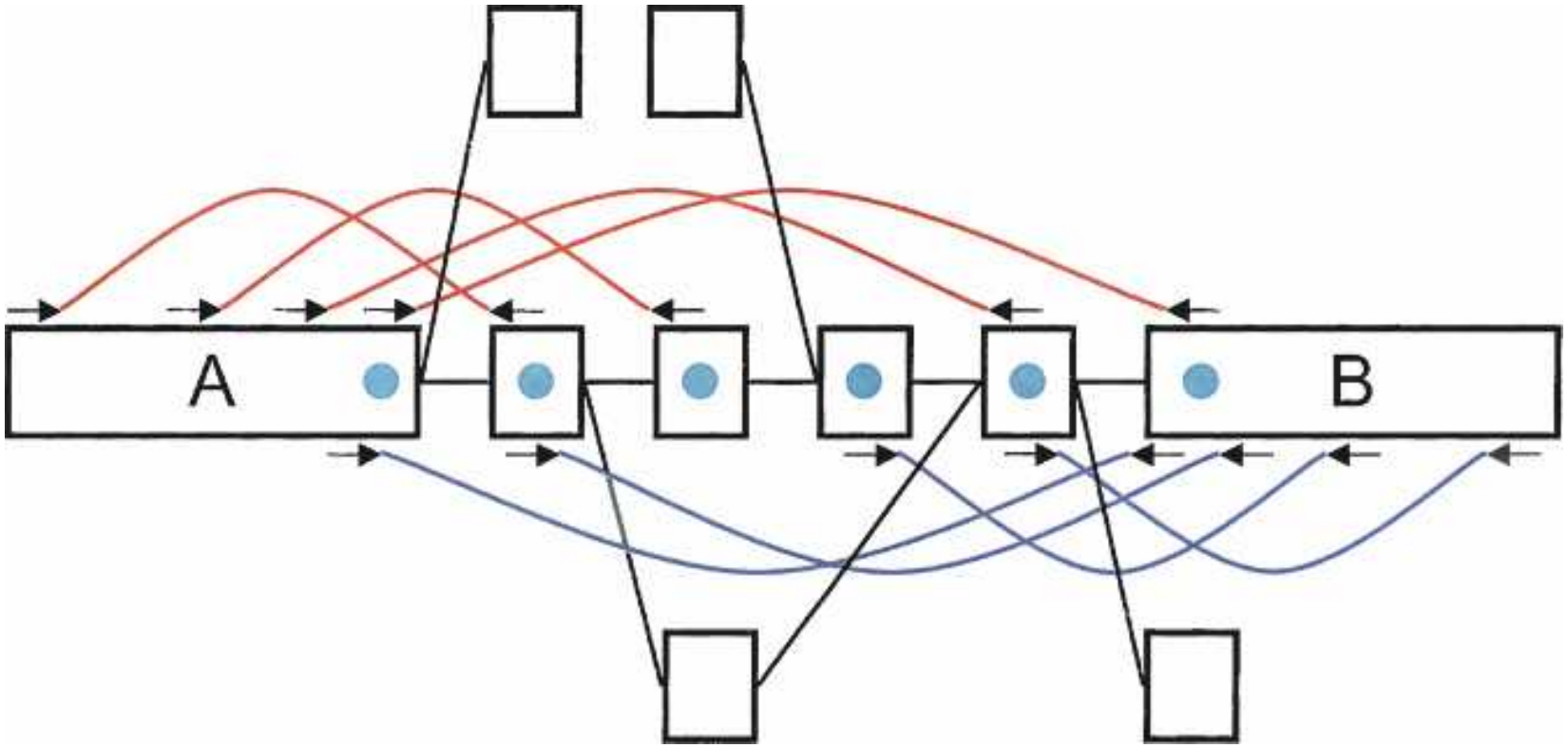
Species tested: model organisms

- *Escherichia coli:* E. coli (bacteria)
- *Saccharomyces cerevisiae:* yeast
- *Caenorhabditis elegans:* C. elegans (worm)
- *Homo sapiens:* Us :)

Velvet: N50 vs. coverage



Velvet: Breadcrumb algorithm



Velvet: comparison with other assemblers

Table 3. Comparison of short read assemblers on experimental *Streptococcus suis* Solexa reads

Assembler	No. of contigs	N50	Average error rate	Memory	Time	Seq. Cov.
Velvet 0.3	470	8661 bp	0.02%	2.0G	2 min 57 sec	97%
SSAKE 2.0	265	1727 bp	0.20%	1.7G	1 h 47 min	16%
VCAKE 1.0	7675	1137 bp	0.64%	1.8G	4 h 25 min	134%