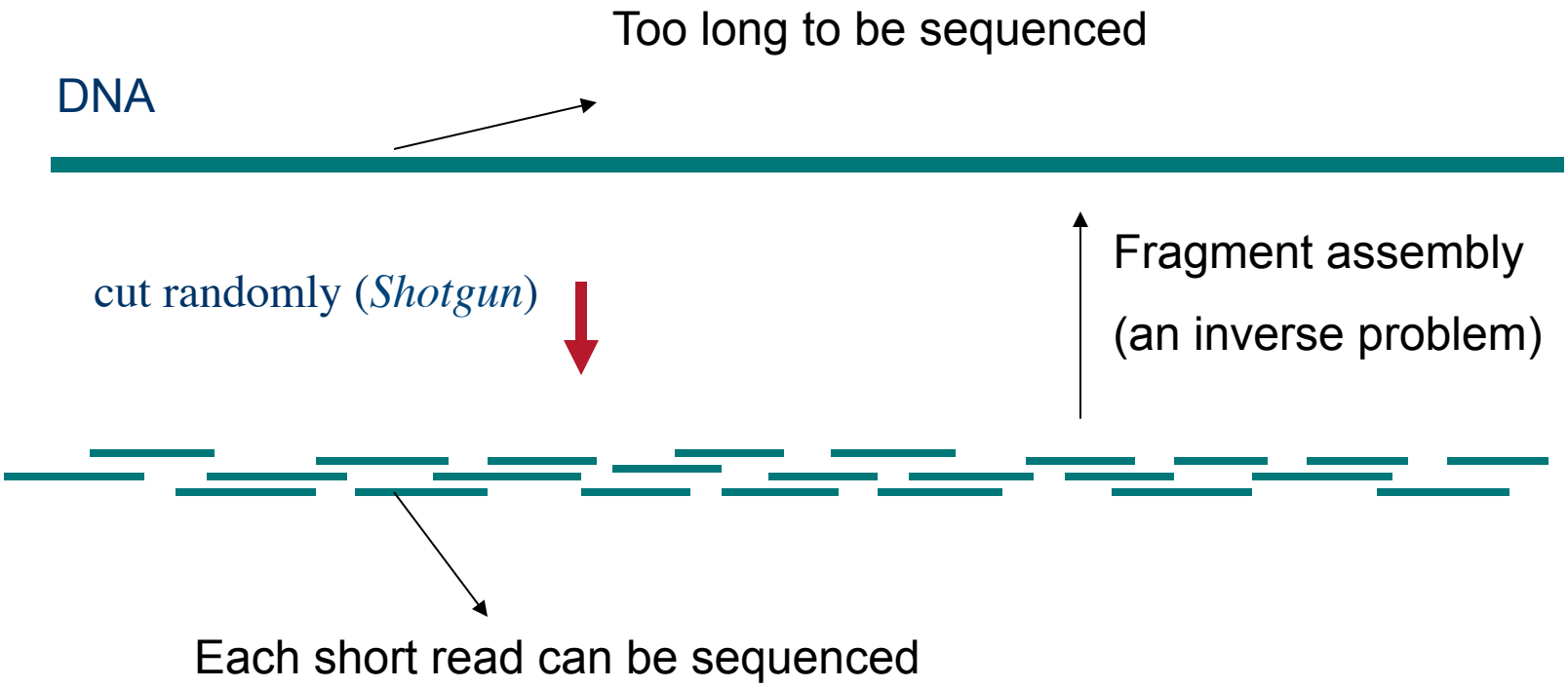
Genome assembly

Yuzhen Ye (yye@indiana.edu)
School of Informatics & Computing, IUB

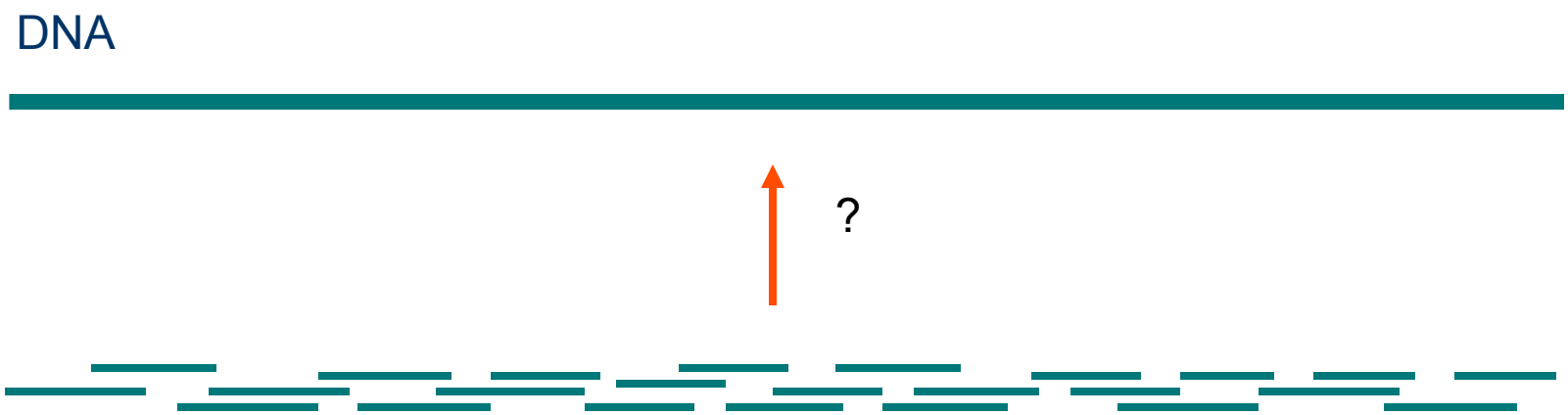
Contents

- Genome assembly problem
- Approaches
  - Comparative assembly
  - De novo assembly (OLC and de Bruijn Graph based approaches)
- Fundamentals
  - Read coverage
  - Sequencing errors
  - Assembly quality metric
  - Assembly evaluation
- Challenges
- Choose the right assembler

Shotgun sequencing



Fragment assembly (Genome assembly)



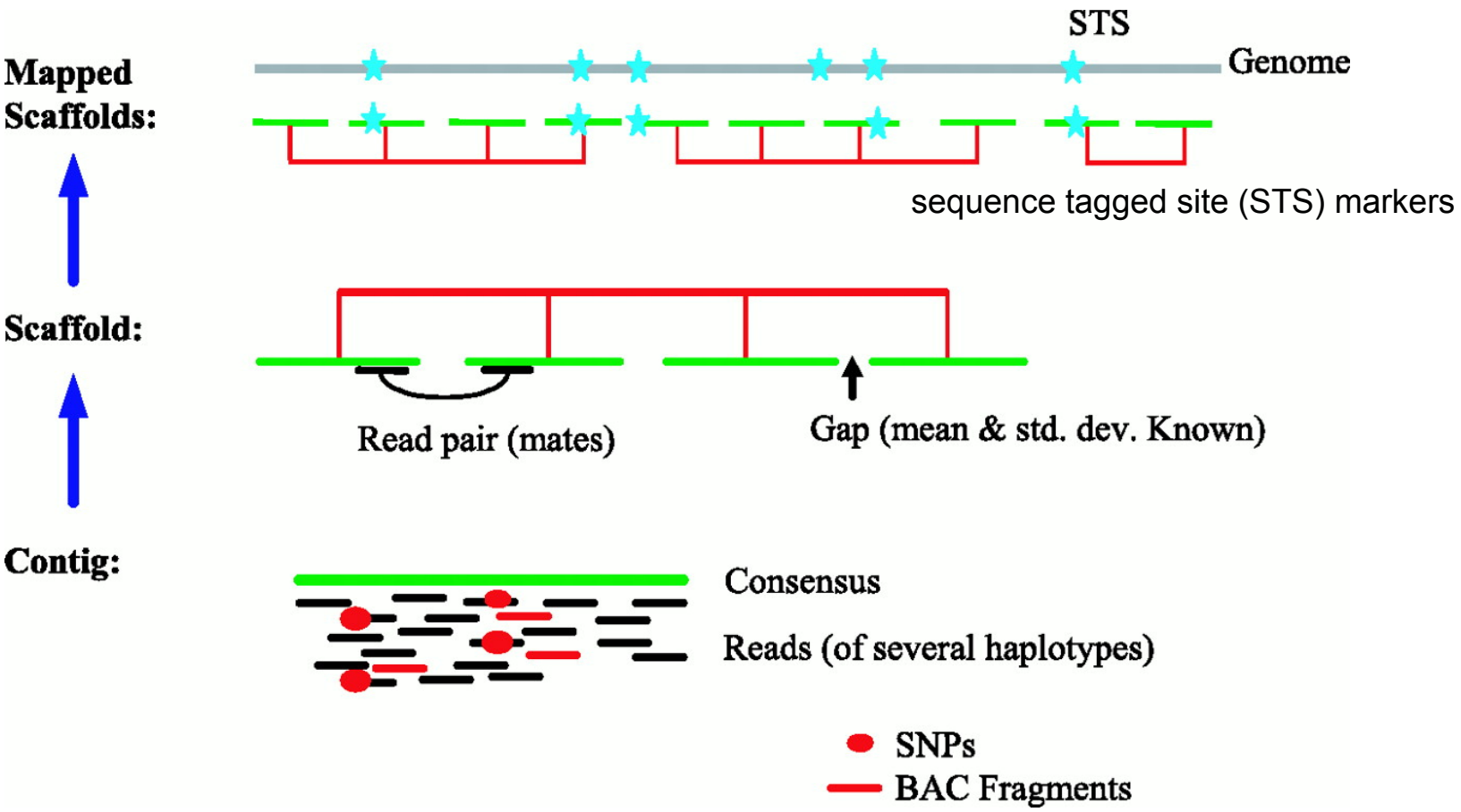
Shotgun sequencing: from small viral genomes to larger genomes

- Early applications of shotgun approach
  - small ***viral genomes*** (e.g., lambda virus; 1982)
  - 30- to 40-kbp segments of larger genomes that could be manipulated and amplified in ***cosmids or other clones (physical mapping) -- hierarchical genome sequencing (divide-and-conquer sequencing)***
- 1994, *Haemophilus influenzae* -- ***whole-genome shotgun (WGS) sequencing***
  - Critical to this accomplishment: use of pairs of reads, called ***mates***, from the ends of 2-kbp and 16-kbp inserts randomly sampled from the genome (which used for ordering the contigs)
- 2001 whole-genome shotgun sequencing of Human genome

Human genome

- **2001 Two assemblies of initial human genome sequences published**
  - International Human Genome project (Hierachical sequencing; BACshotgun)
  - Celera Genomics: WGS approach;
- **Initial impact of the sequencing of the human genome (Nature 470:187–197, 2011)**

Assembly of human genome

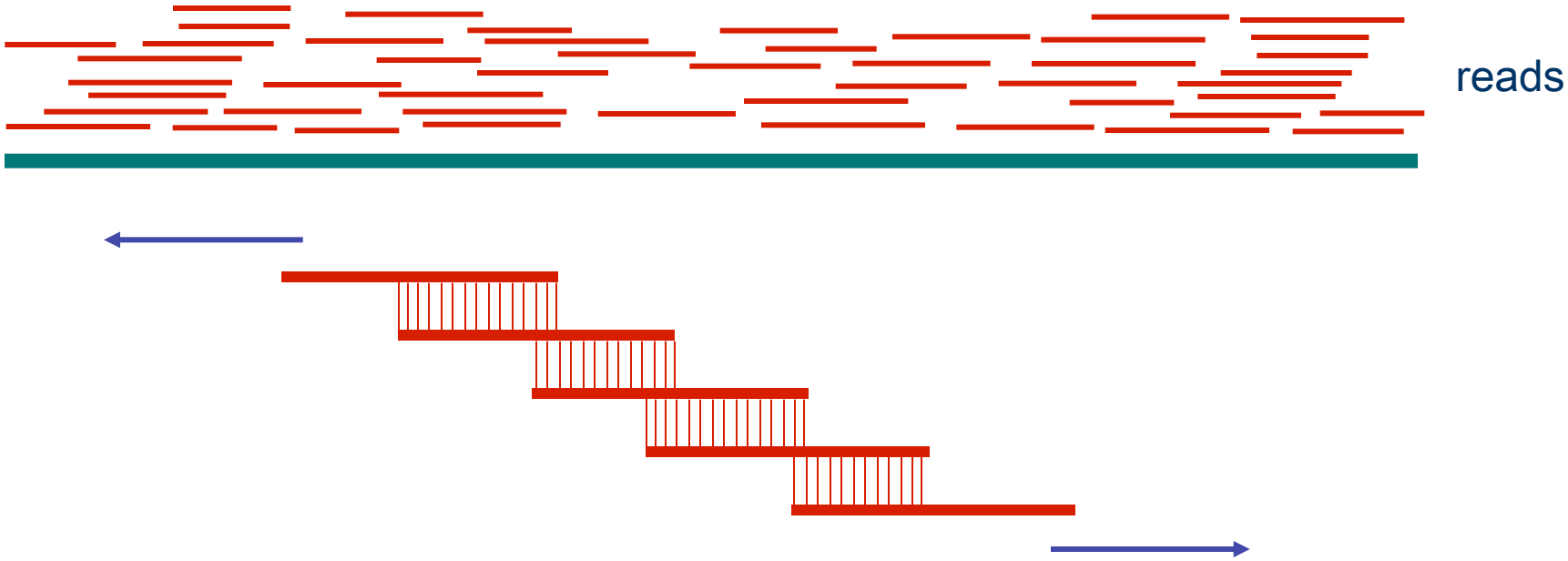


J. C. Venter et al., Science 291, 1304 -1351 (2001)

Assembly approaches

- Comparative assembly
  - Comparative (re-sequencing) approaches that use the sequence of a closely related organism as a guide during the assembly process.
- De novo assembly
  - reconstructing genomes that are not similar to any organisms previously sequenced
  - proven to be difficult, falling within a class of problems (NP-hard)
  - main strategies: greedy, overlap-layout-consensus, and Eulerian
- The two approaches are not exclusive
  - Even if a reference genome is available, regions of the sequenced genome that differ significantly from the reference (e.g. large insertions) can only be reconstructed through *de novo* assembly

Same principle: detecting and utilizing overlaps between reads



Overlap-layout-consensus approach

Assemblers: ARACHNE, PHRAP, CAP, TIGR, CELERA

Overlap: find potentially overlapping reads

Layout: merge reads into contigs

Consensus: derive the DNA sequence and correct read errors

..ACGATTACAATAGGTT..

Overlap computation

- Find the best match between the suffix of one read and the prefix of another
- Due to sequencing errors, need to use dynamic programming to find the optimal *overlap alignment*
- Apply a filtration method to filter out pairs of fragments that do not share a significantly long common substring

Overlap computation

- Sort all *k*-mers in reads (*k* ~ 24)
- Find pairs of reads sharing a k-mer
- Extend to full alignment – throw away if not >95% similar

Layout

Create local multiple alignments from the overlapping reads

Derive consensus sequence

Derive multiple alignment from pairwise read alignments

Derive each consensus base by weighted voting

Consensus

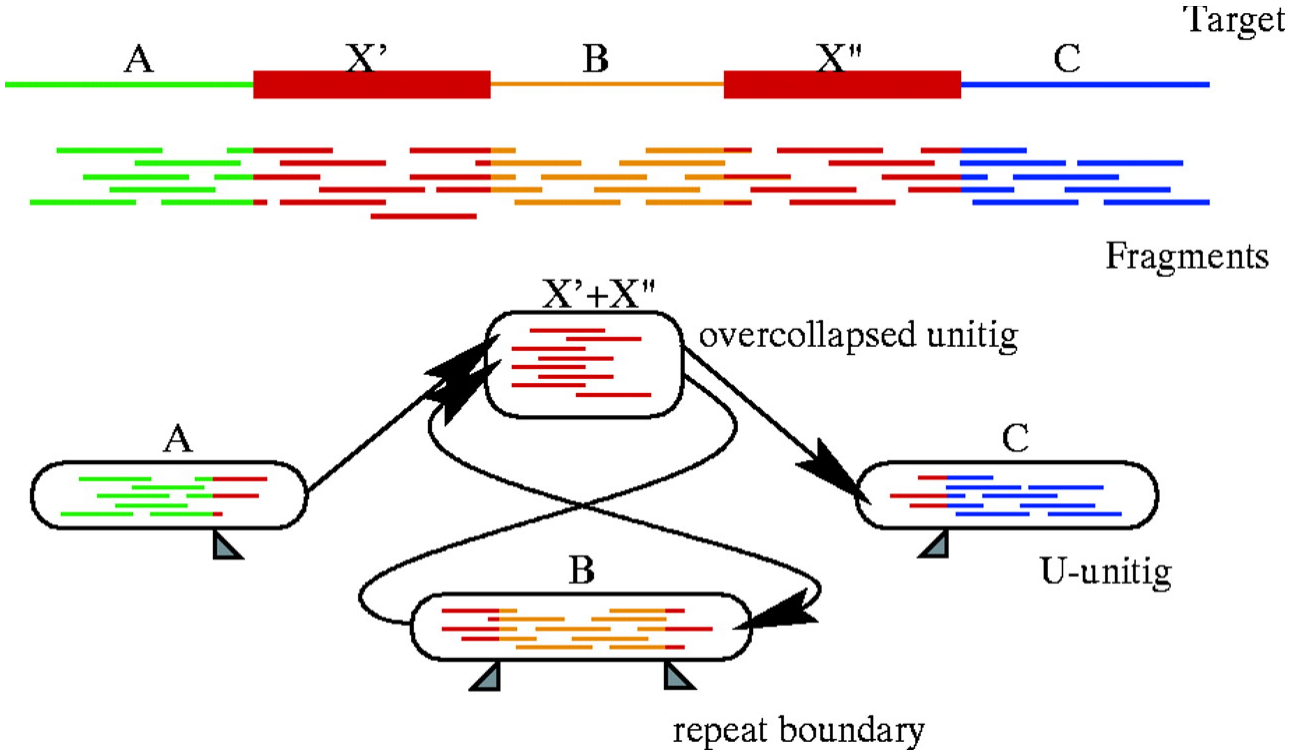
- A consensus sequence is derived from a profile of the assembled fragments
- A sufficient number of reads are required to ensure a statistically significant consensus.
- Reading errors are corrected

Celera assembler

- “The key to not being confused by repeats is the exploitation of ***mate pair*** information to circumnavigate and to fill them”
- A mate pair are two reads from the same clone -- we know the distance between the two reads

Myers et al. 2000 “A Whole-Genome Assembly of Drosophila”. *Science*, 287:2196 - 2204

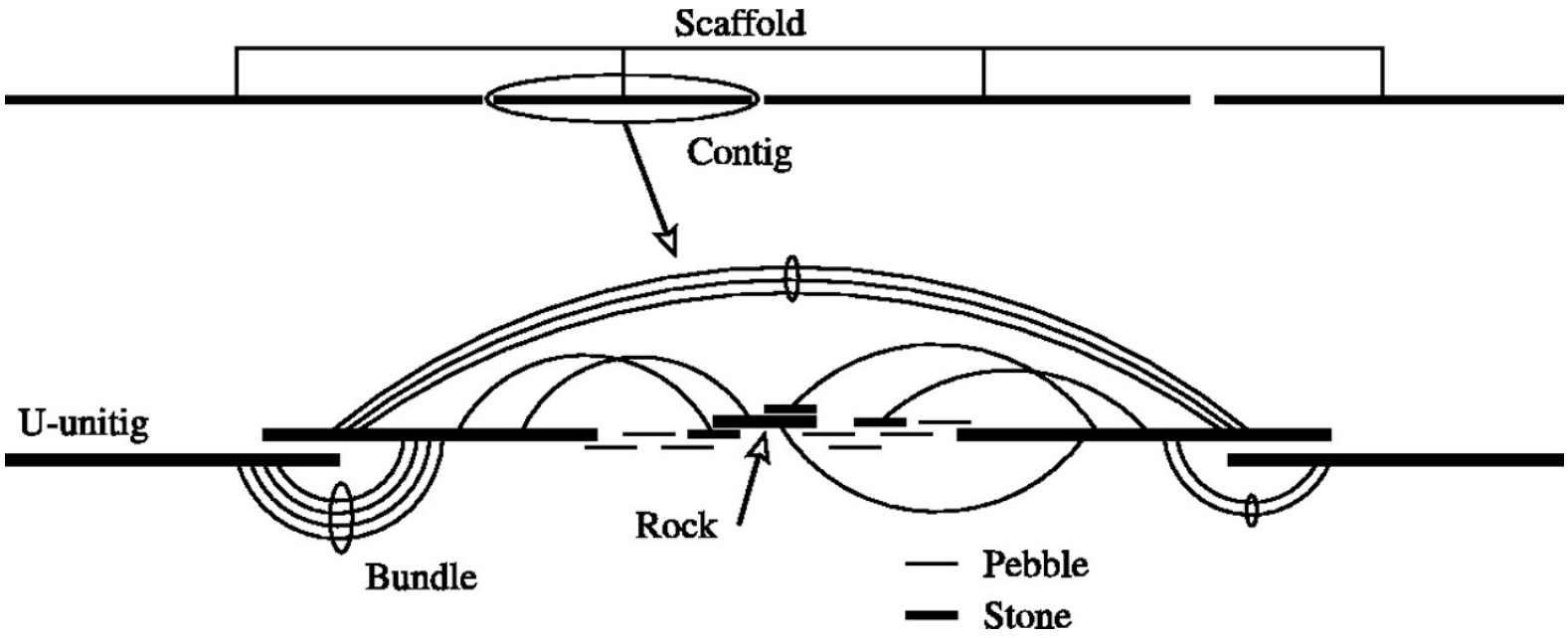
Celera assembler: unitig



Unitig: a maximal interval subgraph of the graph of all fragment overlaps for which there are no conflicting overlaps to an interior vertex

A-statistic: log-odds ratio of the probability that the distribution of fragment start points is representative of a "correct" unitig versus an overcollapsed unitig of two repeat copies.

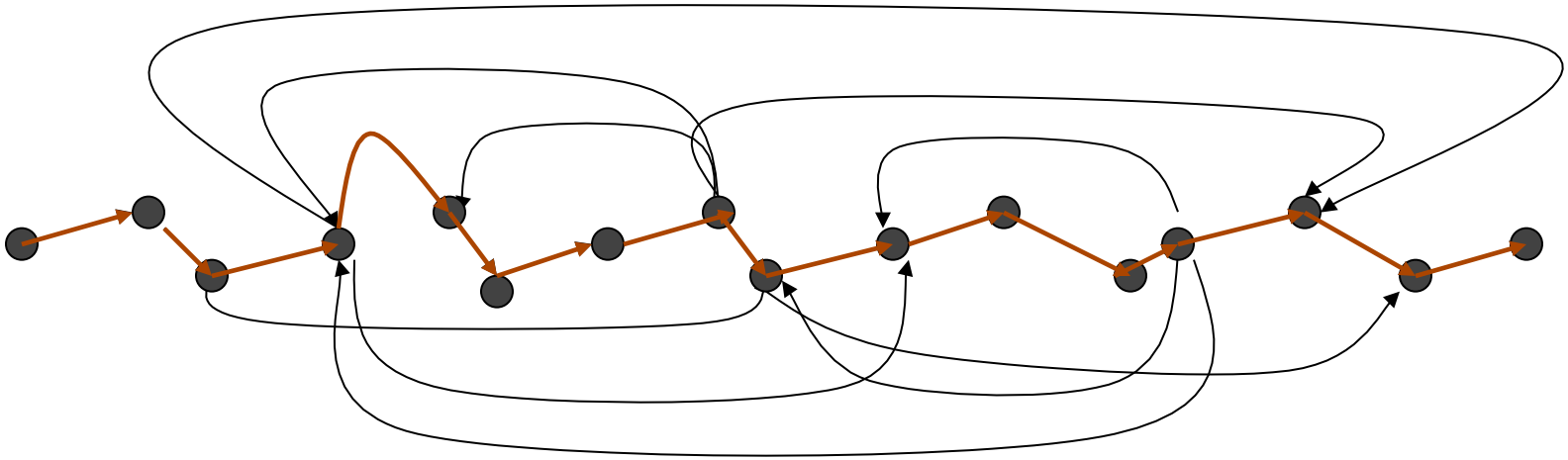
Celera assembler: scaffold



Contigs that are ordered and oriented into ***scaffolds*** with approximately known distances between them (using ***mate pairs*** or BAC ends)

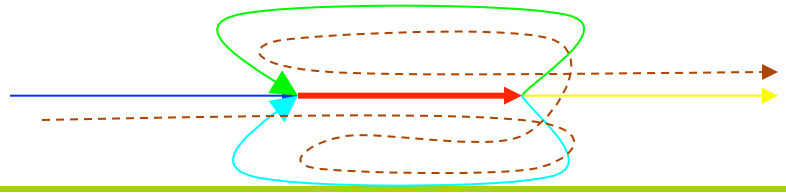
De novo assembly: two alternative choices

Finding a path visiting every VERTEX exactly once in the OVERLAP graph:
Hamiltonian path problem



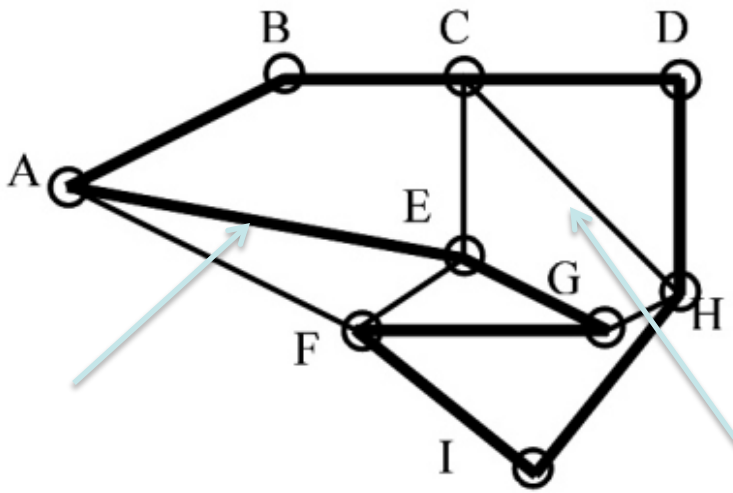
NP-complete problem: algorithms unknown

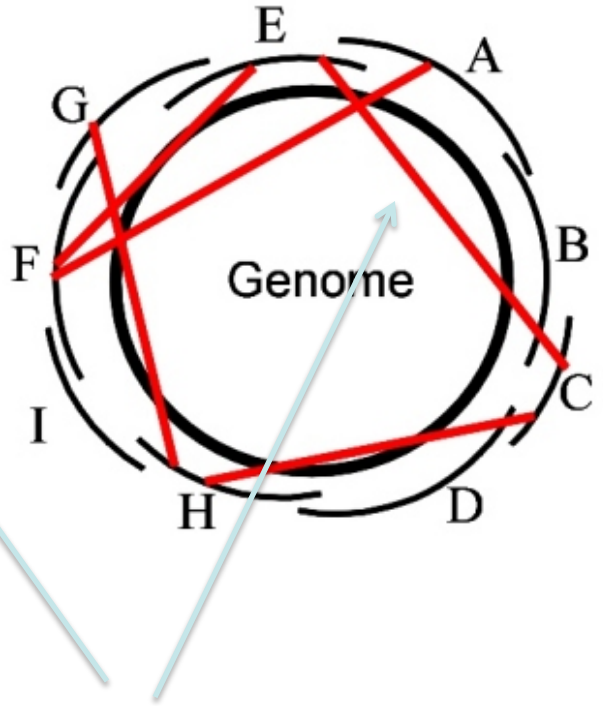
Find a path visiting every EDGE exactly once in the REPEAT graph:
Eulerian path problem



Linear time algorithms are known

Overlap graph

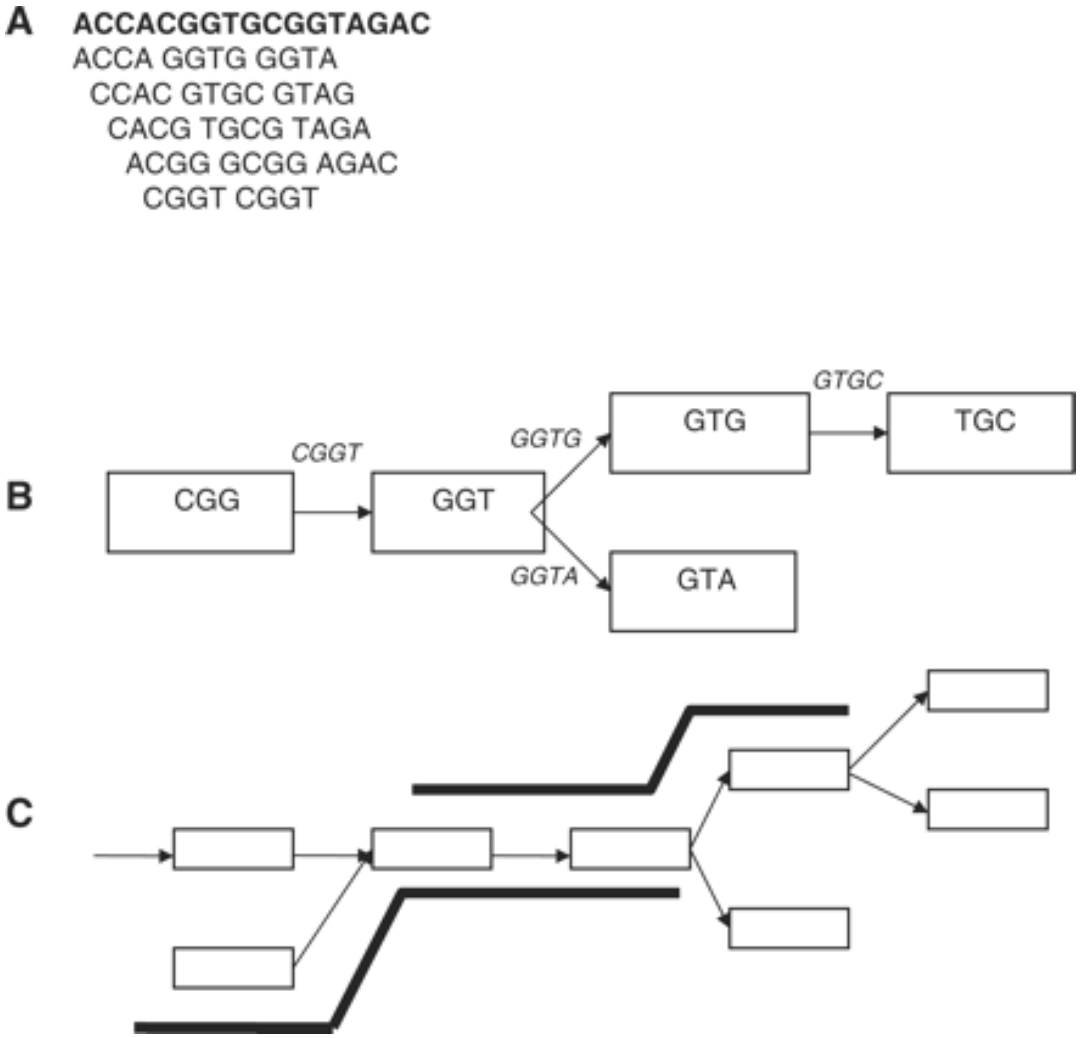




thick edges (a Hamiltonian cycle) correspond to the correct layout of the reads along the genome

False overlaps induced by repeats

Eulerian path approach



Pairwise overlaps between reads are never explicitly computed, hence no expensive overlap step is necessary

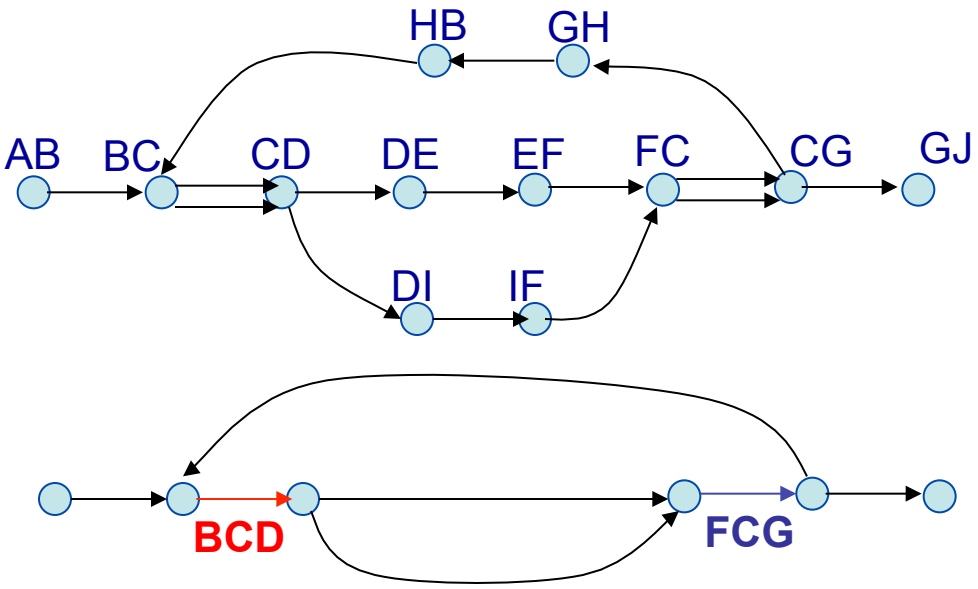
Overlap between two reads (bold) that can be inferred from the corresponding paths through the deBruijn graph

De Bruijn graph → repeat graph (no sequencing errors)

ABCDEFCGHBCDIFCGJ

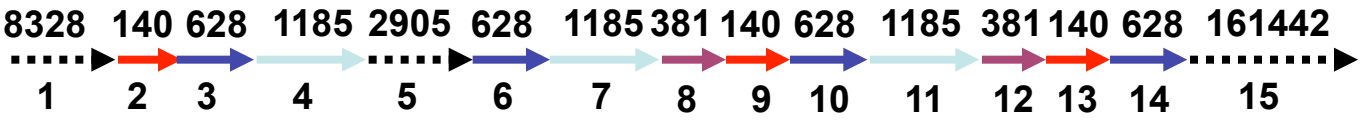
Vertices: (*k-1)*-mers from the sequence

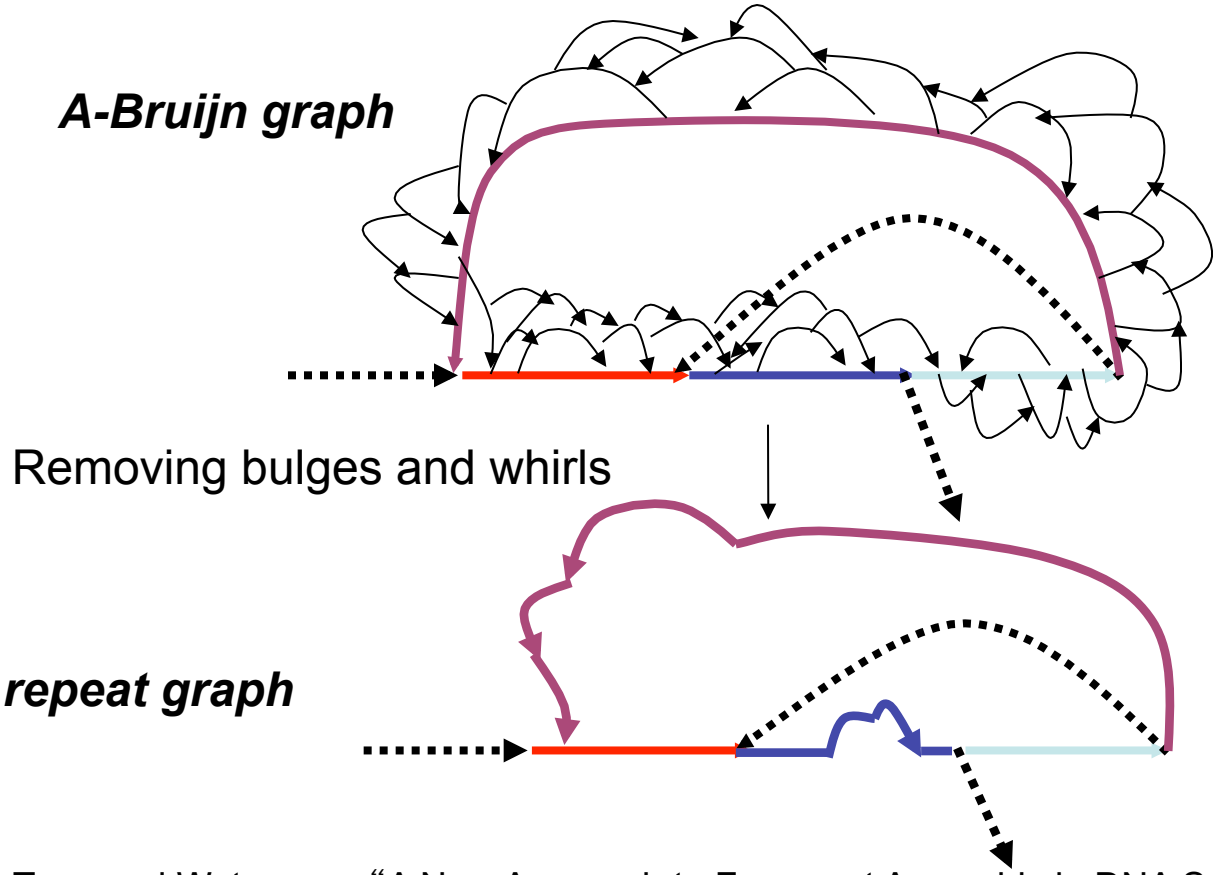
Edges: *k*-mers from the sequence



Every sub-repeat is represented as a repeat edge in the graph.

Repeat graph

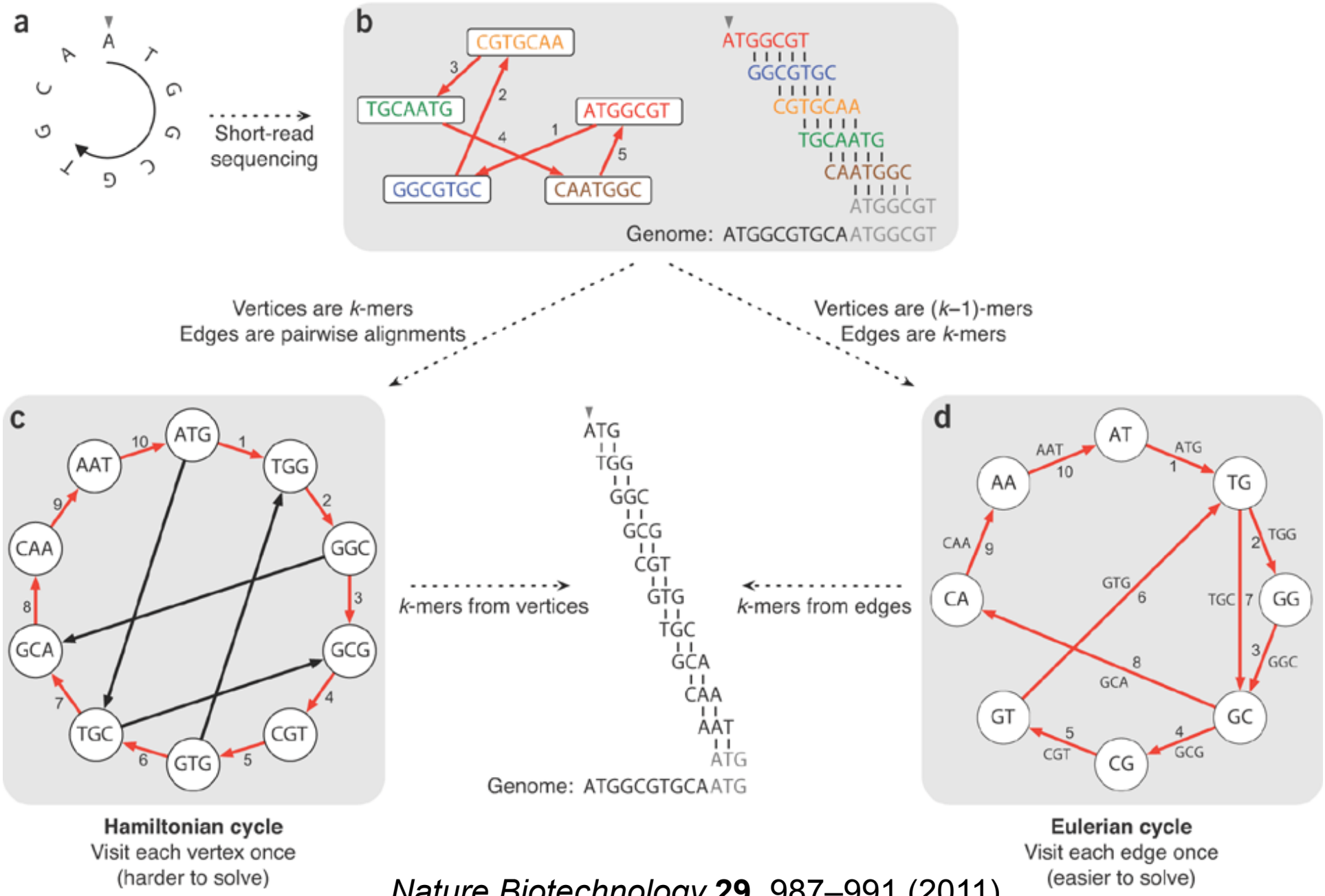




Pevzner, Tang and Waterman. “A New Approach to Fragment Assembly in DNA Sequencing”. RECOMB01

How to apply de Bruijn graphs to genome assembly

Nature Biotechnology 29, 987–991 (2011)



Nature Biotechnology **29**, 987–991 (2011)

Fundamentals #1: Read coverage

Assuming uniform distribution of reads:
Length of genomic segment: ***L***
Number of reads: ***n***
Length of each read: ***l***

Coverage $\lambda = n\,l / L$

How much coverage is enough (or what is sufficient oversampling)?

Lander-Waterman model: $P(x) = (\lambda^x * e^{-\lambda}) / x!$

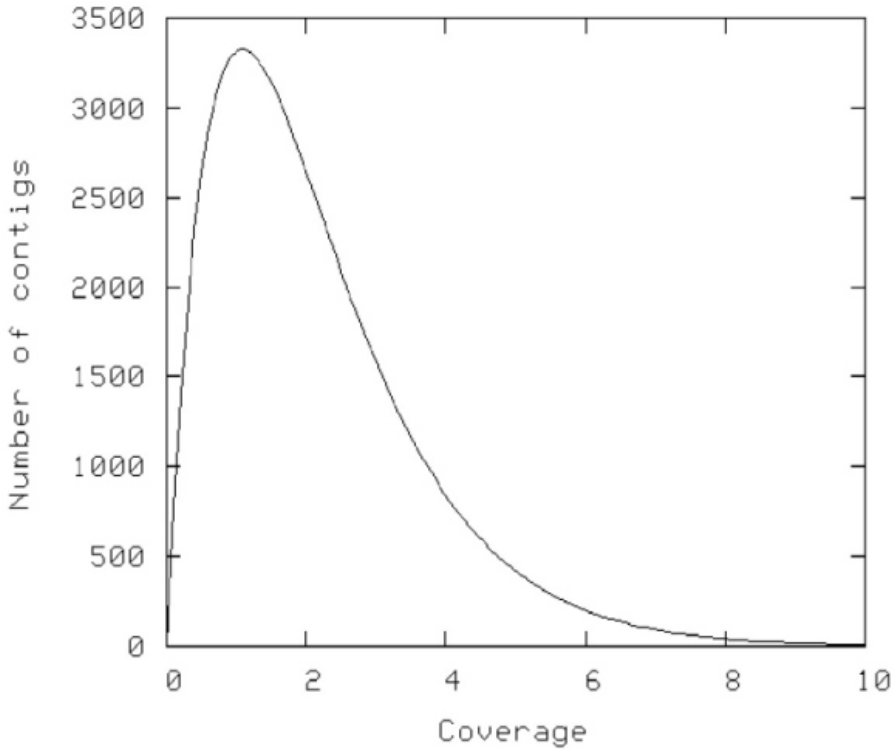
$P(x=0) = e^{-\lambda}$

where λ is coverage

Poisson distribution

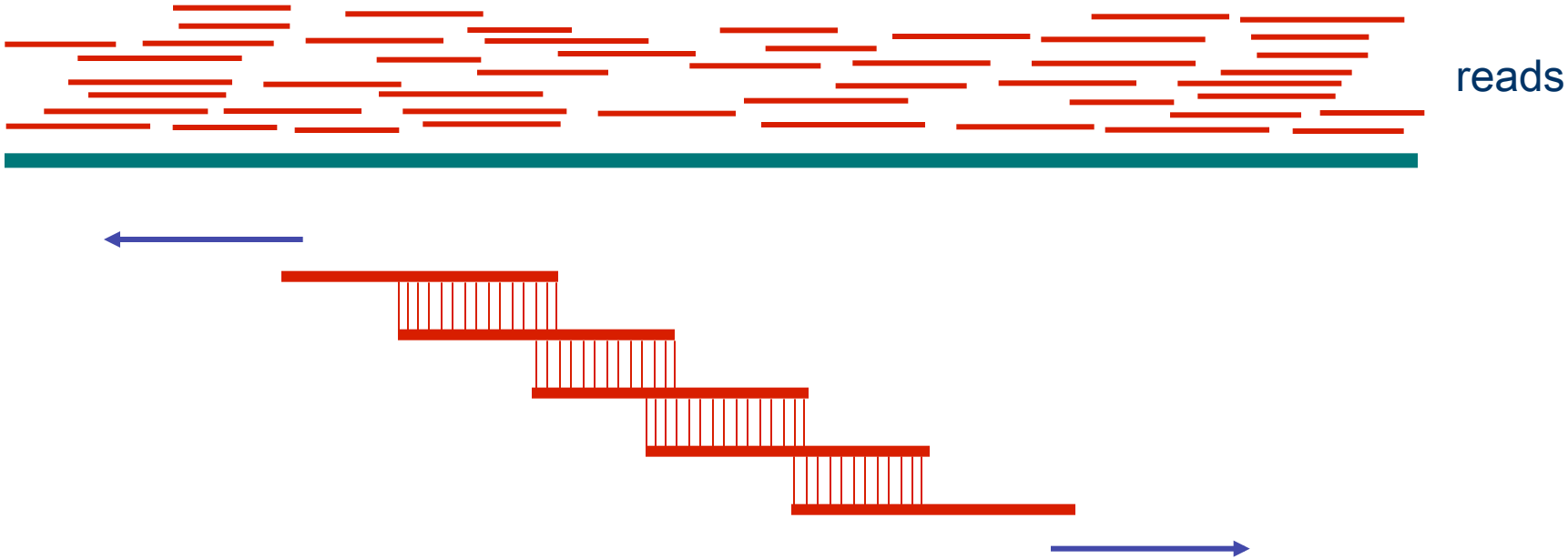
c	$P_0 = e^{-c}$	% not sequenced	% sequenced (1- P_0)
1	0.37	37%	63%
2	0.135	13.5%	87.5%
3	0.05	5%	95%
4	0.018	1.8%	98.2%
5	0.0067	0.6%	99.4%
6	0.0025	0.25%	99.75%
7	0.0009	0.09%	99.91%
8	0.0003	0.03%	99.97
9	0.0001	0.01%	99.99%
10	0.000045	0.005%	99.995%

Contig numbers vs read coverage



Using a genome of 1Mbp

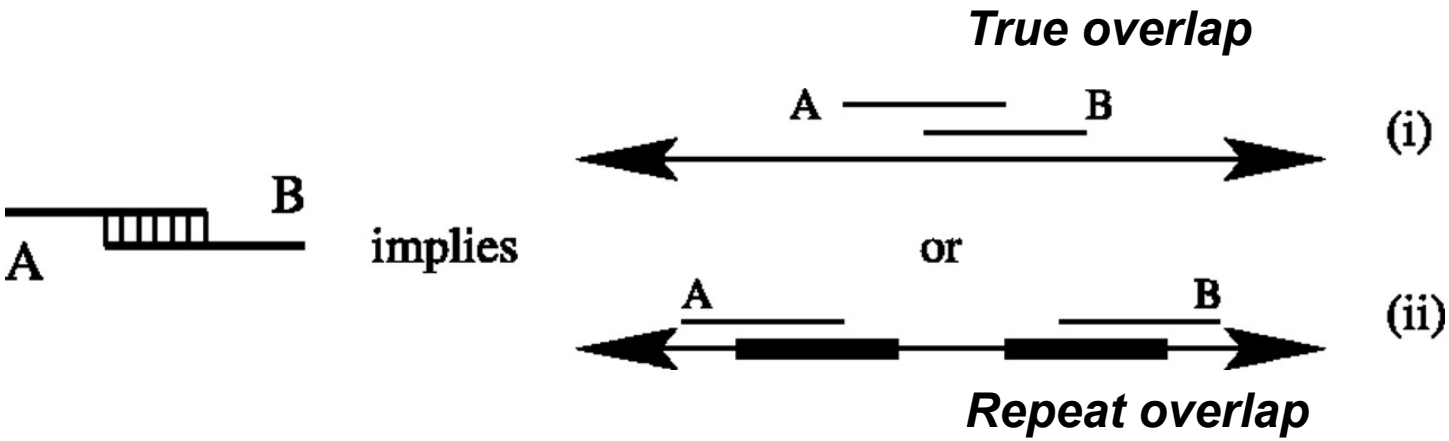
How much coverage is needed



Cover region with >7-fold redundancy

Overlap reads and extend to reconstruct the original DNA sequence

Repeats complicate fragment assembly



Fundamentals #2: Sequencing reads contain errors

$$Q_{\mathrm{PHRED}} = -10 \times \log_{10}(P_e)$$ Phrep quality score

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1/10	90%
20	1/100	99%

$$Q_{\mathrm{Solexa}} = -10 \times \log_{10}\left(\frac{P_e}{1 - P_e}\right)$$

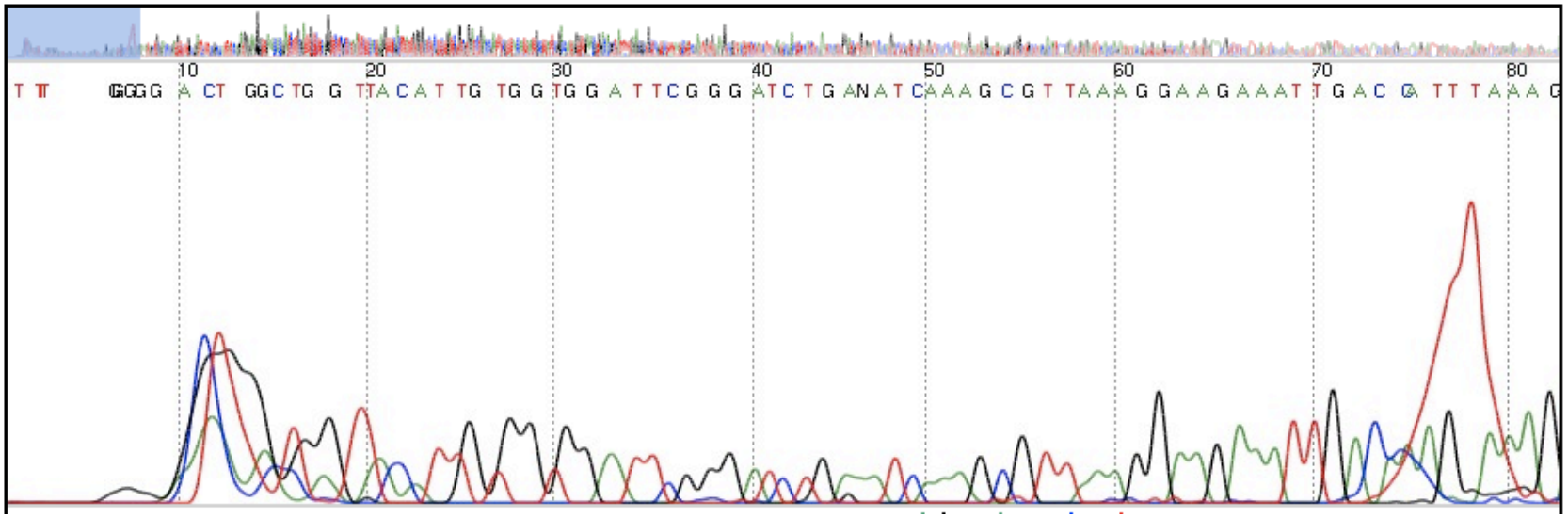
$$Q_{\mathrm{PHRED}} = 10 \times \log_{10}\left(10^{Q_{\mathrm{Solexa}}/10} + 1\right)$$

$$Q_{\mathrm{Solexa}} = 10 \times \log_{10}\left(10^{Q_{\mathrm{PHRED}}/10} - 1\right)$$

(for high values the two scores are asymptotically equal)

Base calling

- Determine the sequence of nucleotides from chromatograms or flowgram (trace files often in SCF format)
- Peak detection
- Phrep quality score



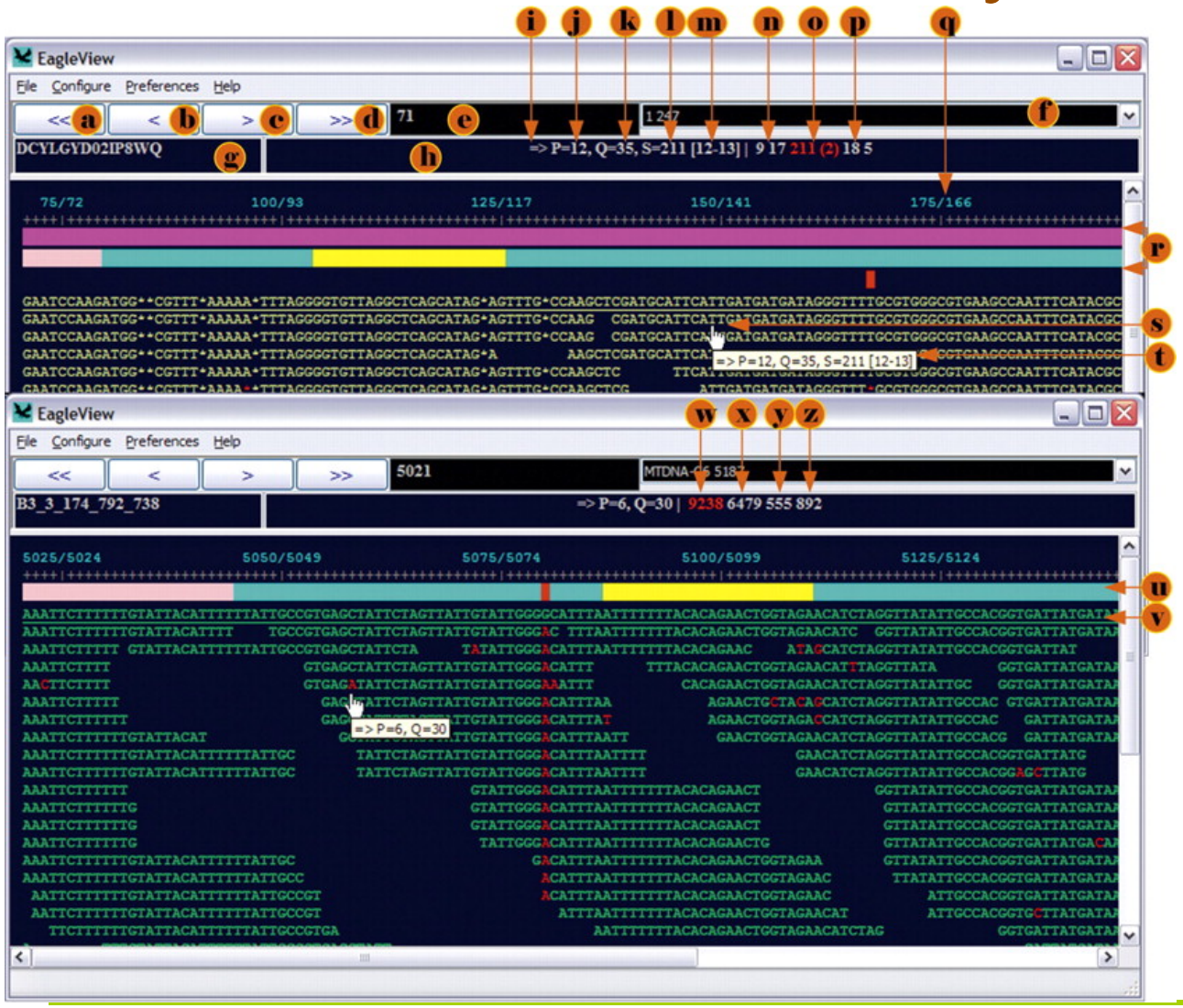
Fundamentals #3: Assembly quality metrics

- Number of contigs, the longest contig
- N50, defined as the contig length such that using equal or longer contigs produces *half* the bases of the genome (or all the contigs).
  - sorting all contigs from largest to smallest
  - contig sizes: 2M, 1M, 0.5M, 0.3M, 0.2M, … 500bp with total bases = 8M, then N50 = 0.2M

Fundamentals #4: Assembly validation

- No assembler is perfect
- Post-assembly validation is important
  - Detection of collapsed repeats: regions within the assembly that have unusually deep coverage. Such regions can be identified through statistical approaches.
  - The C/E statistic of Zimin estimates the likelihood that a cluster of mate-pairs indicates an insertion or deletion within the assembly.
  - Visual tools

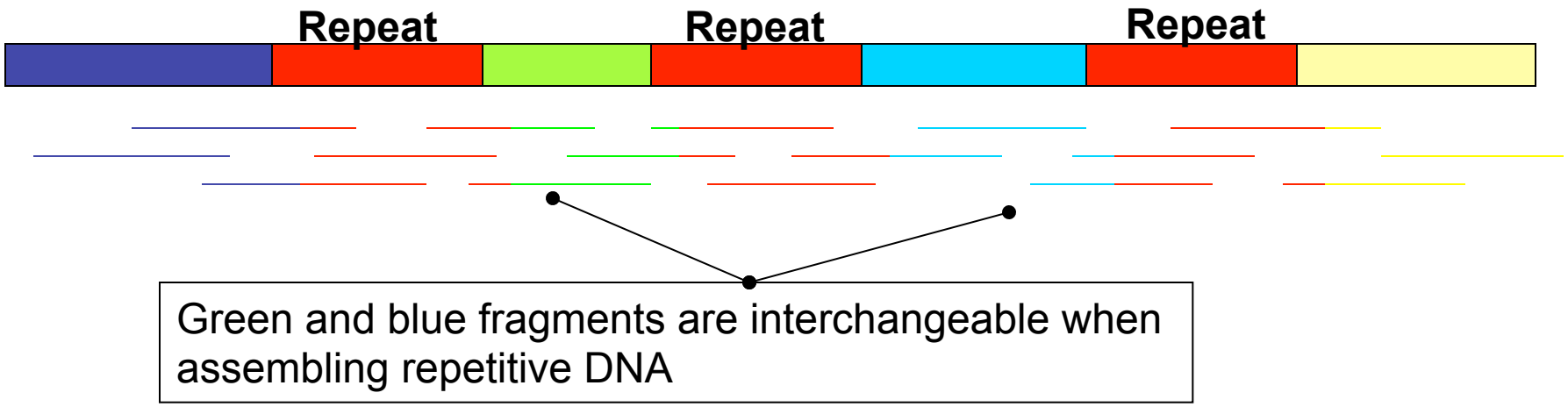
Genome assembly viewer



EagleView

Challenges in fragment assembly

- Repeats: A **major** problem for fragment assembly
- > 50% of human genome are repeats:
  - over 1 million *Alu* repeats (about 300 bp)
  - about 200,000 LINE repeats (1000 bp and longer)



Repeat types

- **Low-Complexity DNA** (e.g. ATATATATACATA…)
- **Microsatellite repeats** $(a_1 \ldots a_k)^N$ where k ~ 3-6
  (e.g. CAGCAGTAGCAGCACCAG)
- **Transposons/retrotransposons**
  - **SINE** Short Interspersed Nuclear Elements
    (e.g., *Alu*: ~300 bp long, 10^6 copies)
  - **LINE** Long Interspersed Nuclear Elements
    ~500 - 5,000 bp long, 200,000 copies
  - **LTR retroposons** Long Terminal Repeats (~700 bp) at each end
- **Gene Families** genes duplicate & then diverge
- **Segmental duplications** ~very long, very similar copies

More challenges

- Assemble large genomes using short reads
- Assemble multiple genomes from a sequencing data of a mixture (metagenomes)

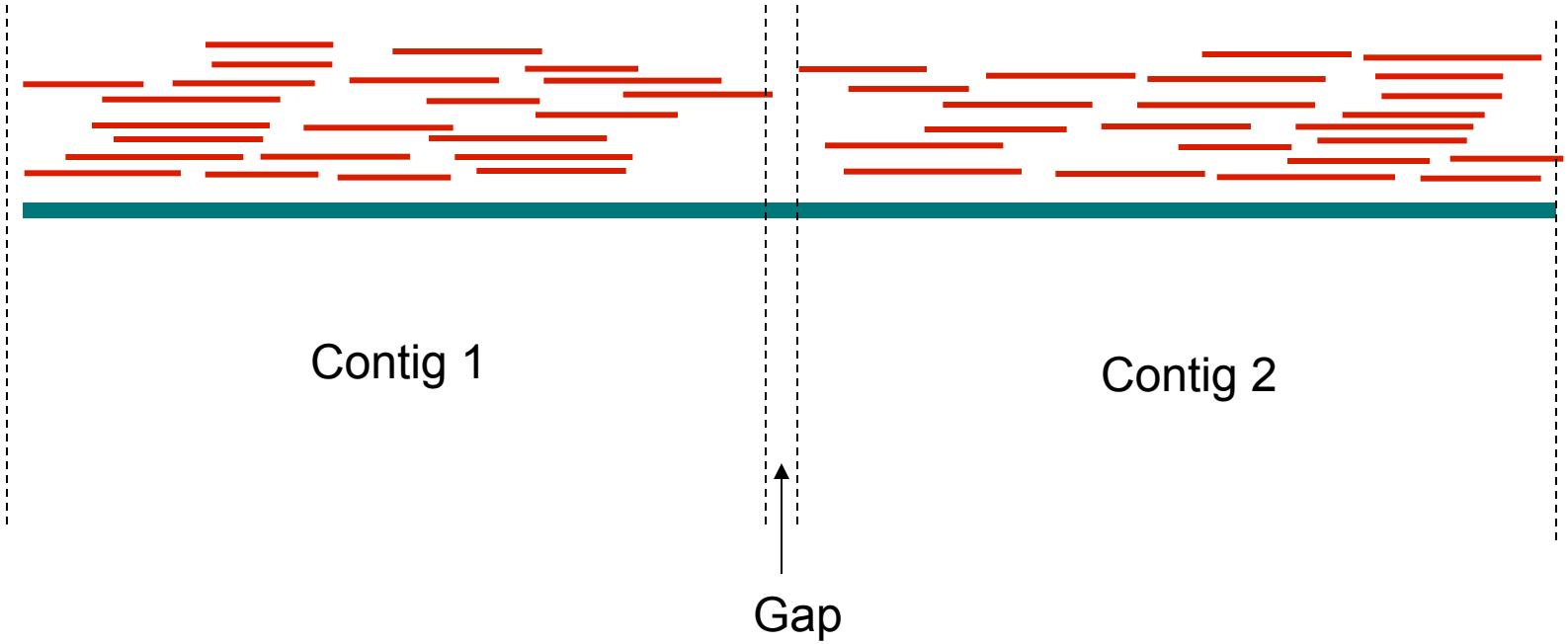
Choose the right assembler

- There is a good collection of assemblers out there
- Some were designed for specific sequencing platforms
- Genome assembly gold-standard evaluations (GAGE)
  - http://gage.cbcb.umd.edu/assemblers/index.html

Assemblers evaluated at GAGE

- ABySS (Assembly By Short Sequencing) (Birol et al): A denovo assembler for short read sequence data which uses a distributed representation of a de Bruijn graph
- ALLPATHS-LG (Gnerre et al): a de Bruijn graph-based *de novo* assembler for large (and small) genomes
- Bambus2: The second generation Bambus scaffolder relies on a combination of a novel method for detecting genomic repeats and algorithms that analyze assembly graphs to identify biologically meaningful genomic variants.
- Celera Assembler: an Overlap-Layout-Consenus based de novo whole-genome shotgun (WGS) DNA sequence assembler
- MSR-CA (pronounced "MizerKa") is a new technique that pre-processes the short read data and then performs the final assembly using a modified version of Celera Assembler
- SGA (Simpson et al): stands for String Graph Assembler. Experimental de novo assembler based on string graphs.
- SOAPdenovo (Li et al): is the short-read assembler that was used for the panda genome, the first mammalian genome assembled entirely from Illumina reads, and for several human genomes and other genomes subsequently (SOAPdenovo2)
- Velvet (Zerbino et al): Velvet is a *de novo* genome assembler specially designed for short read sequencing technologies, particularly Illumina reads, and was one of the first short-read assemblers to be published.

Gaps and contigs



Filling gap -- up the gaps by further experiments
Mates for ordering the contigs

How to best utilize the draft genomes?

de novo assemblies constructed from short-read data are highly fragmented

● Chromosomes [1] ◑ Scaffolds or contigs [16] ◔ SRA or Traces [0]

Organism	BioProject	Assembly	Status	Chrs	Plasmids	Size (Mb)	GC%	Gene	Protein
Treponema denticola H1-T	PRJNA189164, PRJNA64913	Trep_dent_H1-T_V1	◑	1	-	2.93	37.9	2,754	2,705
Treponema denticola ATCC 35405	PRJNA57583, PRJNA4	ASM818v1	●	1	-	2.84	37.9	2,838	2,767
Treponema denticola AL-2	PRJNA189158, PRJNA64903	Trep_dent_AL-2_V1	◑	1	-	2.84	38	2,681	2,632
Treponema denticola ATCC 35404	PRJNA189162, PRJNA64911	Trep_dent_ATCC_35404_V1	◑	1	-	2.77	38	2,583	2,533
Treponema denticola H-22	PRJNA189163, PRJNA64915	Trep_dent_H-22_V1	◑	1	-	2.76	37.9	2,569	2,520