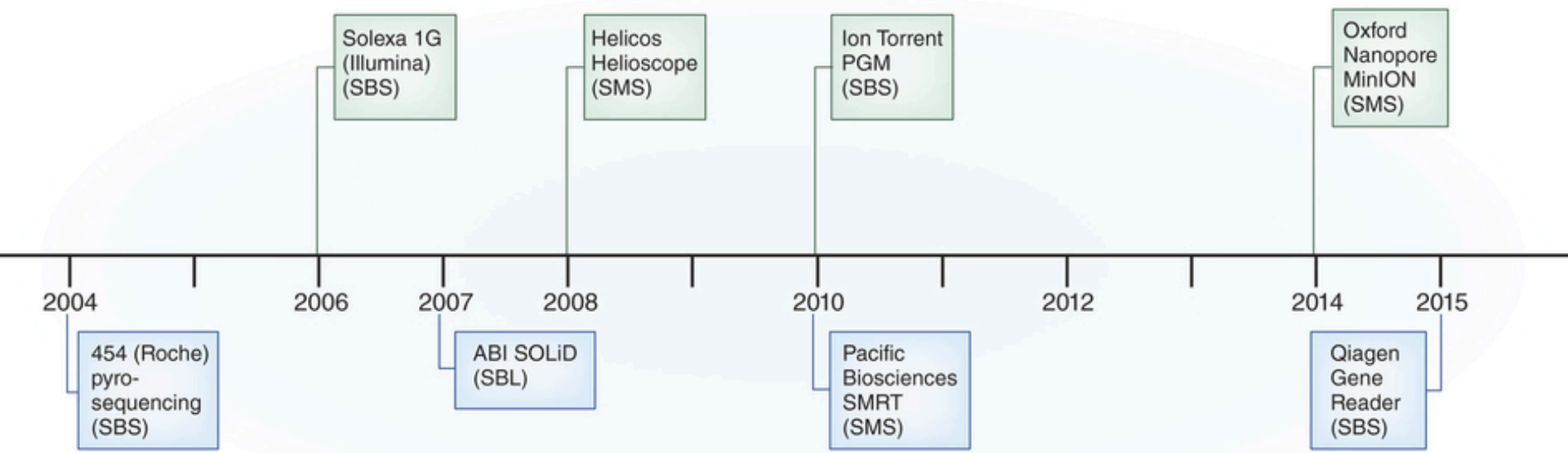
Sequence Assembly

Dr. Jung Soh

657.000 SS 2017

NGS platforms timeline



- Year of introduction of each of the NGS platforms that successfully achieved commercial introduction during the past decade.
- SBS: sequencing by synthesis, SMS: single-molecule sequencing, SBL: sequencing by ligation.

Mardis, E.R. (2017) DNA sequencing technologies: 2006–2016, Nature Protocols 12:213–218

Comparison of NGS platforms

Company	Read length	Applications
454/Roche	400 bp (single end)	Bacterial and viral genomes, multiplex-PCR products, validation of point mutations, targeted somatic-mutation detection
Illumina	150–300 bp (paired end)	Complex genomes (human, mouse and plants) and genome-wide NGS applications, RNA-seq, hybrid capture or multiplex-PCR products, somatic-mutation detection, forensics, noninvasive prenatal testing
ABI SOLiD	75 bp (single end) or 50 bp (paired end)	Complex genomes (human, mouse, plants) and genome-wide NGS applications, RNA-seq, hybrid capture or multiplex-PCR products, somatic-mutation detection
Pacific Biosciences	Up to 40 kb (single end or circular consensus)	Complex genomes (human, mouse and plants), microbiology and infectious-disease genomes, transcript-fusion detection, methylation detection
Ion Torrent	200–400 bp (single end)	Multiplex-PCR products, microbiology and infectious diseases, somatic-mutation detection, validation of point mutations
Oxford Nanopore	Variable: depends on library preparation (1D or 2D reads)	Pathogen surveillance, targeted mutation detection, metagenomics, bacterial and viral genomes
Qiagen GeneReader	107 bp (single end)	Targeted mutation detection, liquid biopsy in cancer

Mardis, E.R. (2017) DNA sequencing technologies: 2006–2016, Nature Protocols 12:213–218

NGS platform trade-off

- Number of reads
- Read lengths
- Cost
- Accuracy
- Application

http://www.allseq.com/knowledgebank/sequencing-platforms

Platform considerations

- "Second generation"
  - 454: longer read length
  - Illumina: shorter read length, low cost
  - SOLiD: even shorter read length, more reads

- "Third generation"
  - PacBio RS: much longer read length, high error rate
  - Ion Torrent PGM: low machine cost
  - 454 GS Junior, Illumina MiSeq

NGS reads

- Short subsequences of the genome
  - No idea on the original position in the genome
  - Orientation (strand) unknown
- Oversampled (high coverage)
  - Reads overlap: only clue for assembly
- Base errors
  - Indels, substitutions, characteristic biases
- Supposed to cover entire genome
  - Not always true
  - Coverage not uniform over genomic positions

Genome assembly metaphor

DNA clones

Reads

Reconstructed genome

The Call—Chronicle—Examiner

SAN FRANCISCO, THURSDAY, APRIL 19, 1906.

EARTHQUAKE AND FIRE: SAN FRANCISCO IN RUINS

NO HOPE LEFT FOR SAFETY OF ANY BUILDINGS

BLOW BUILDINGS UP TO CHECK FLAMES

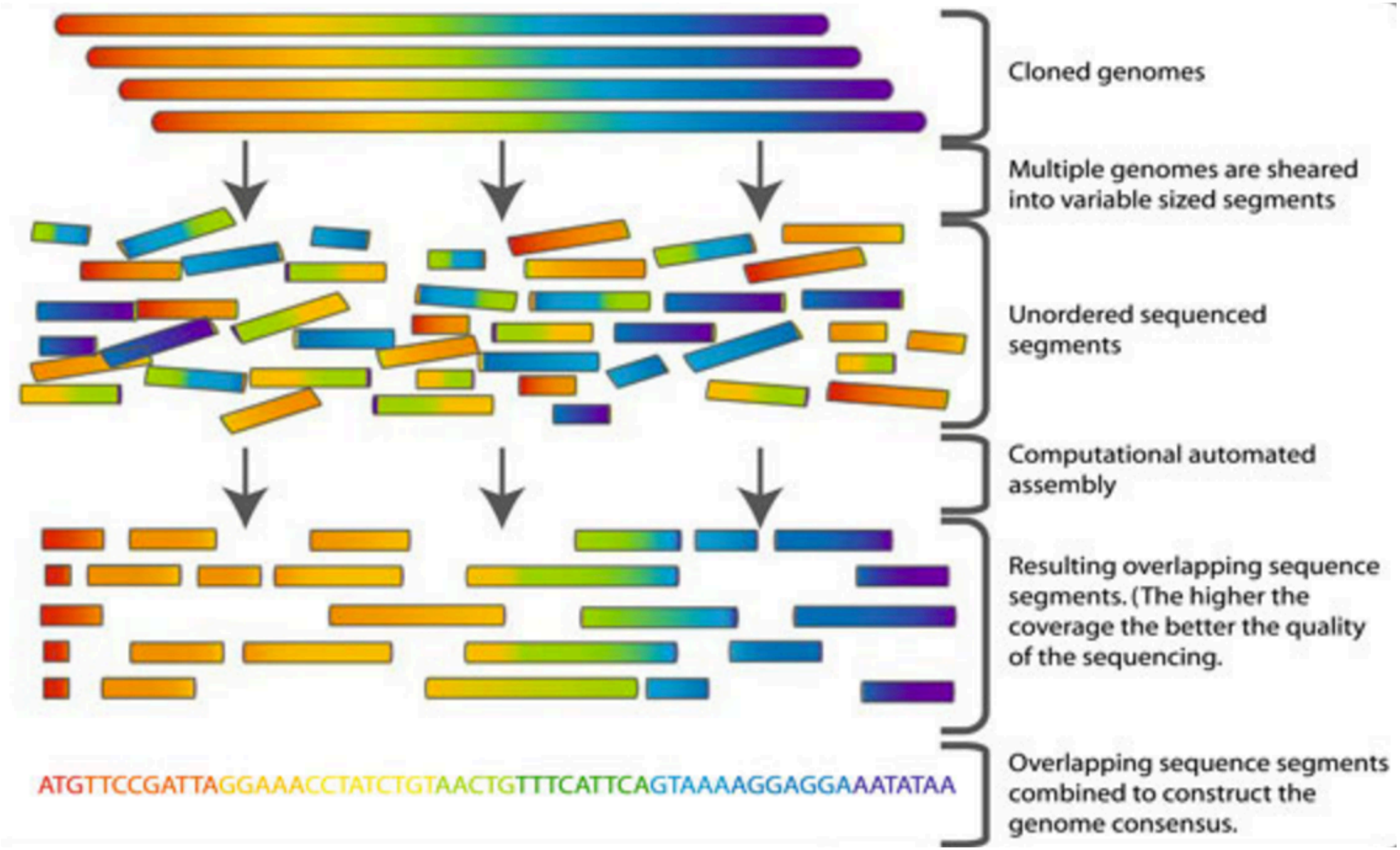
WHOLE CITY IS ABLAZE

CHURCH OF SAINT IGNATIUS IS DESTROYED

MAYOR CONFERS WITH MILITARY AND CITIZENS

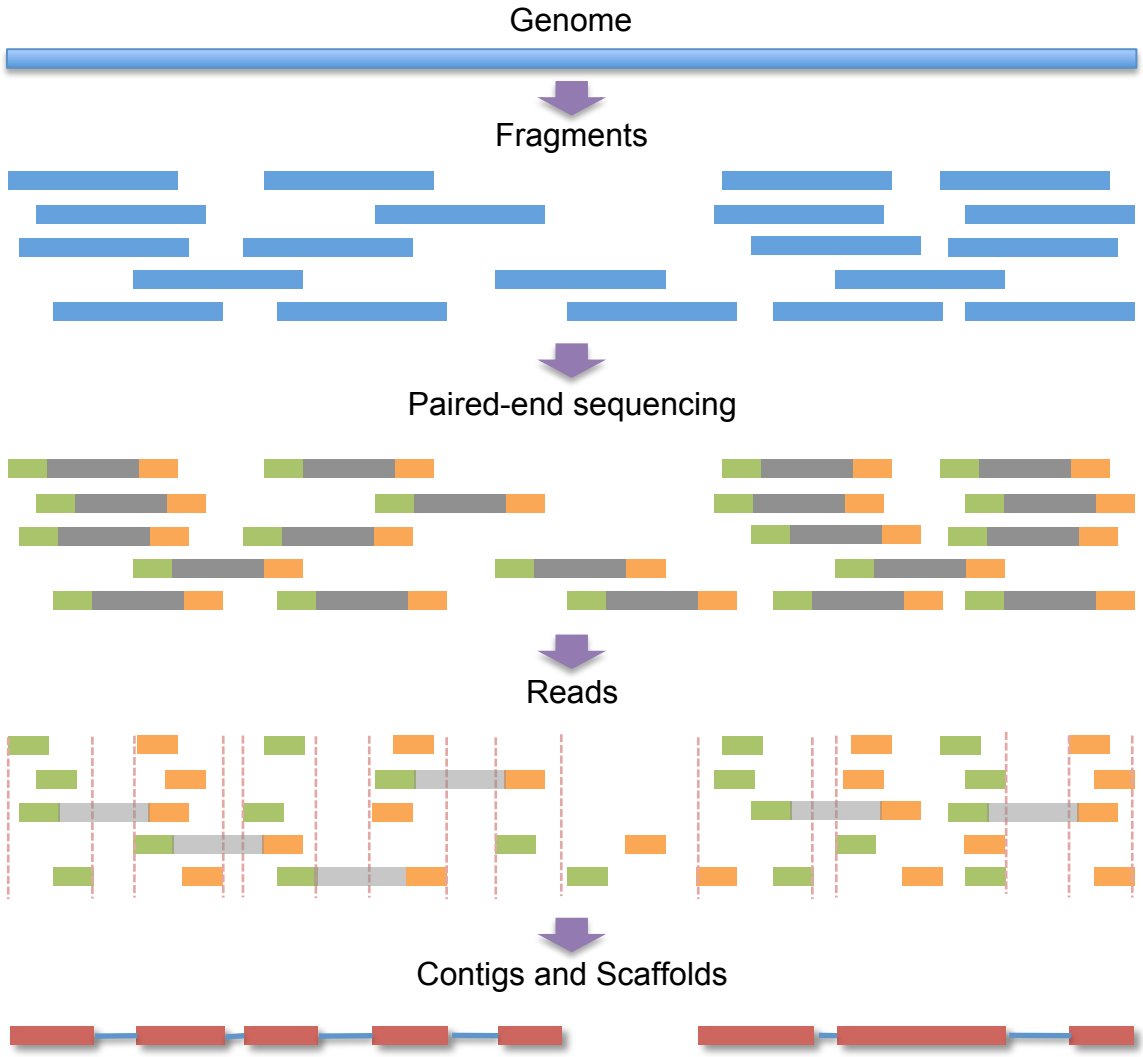
http://www.vicbioinformatics.com/documents/Genome%20Assembly%20Strategies%20-%20Torsten%20Seemann%20-%20IMB%20-%205%20Jul%202010.pdf

Genome assembly: another view



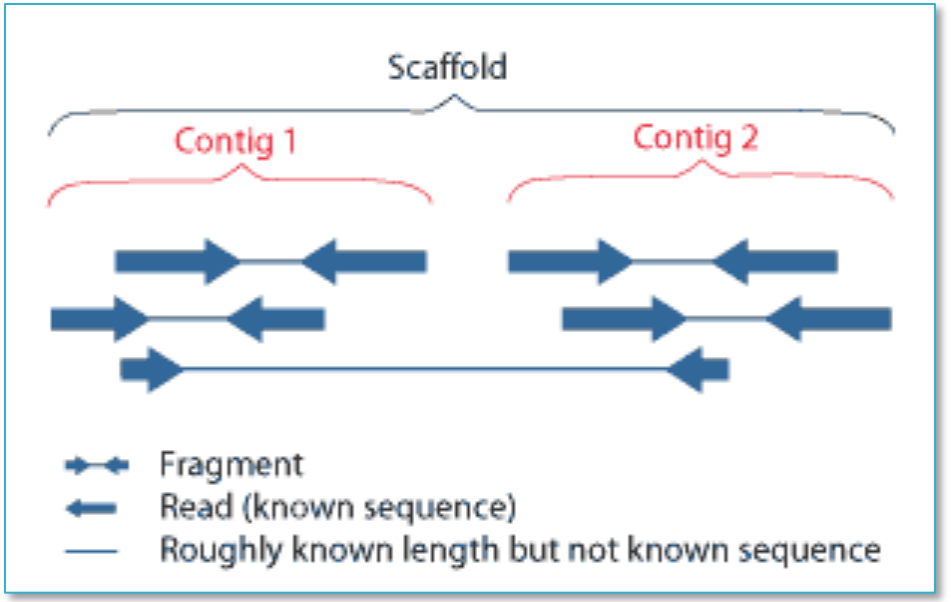
Commins J, Toft C, Fares MA. (2009) Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. Biol Proced Online. 11:52-78.

Random shotgun paired-end sequencing and assembly



Scaffolding

- Joining a set of contigs
  - Scaffolds have gaps (N's)
- Paired-end reads
  - Known sequences at either end of a fragment
  - Distance known (roughly), sequence unknown between ends
- Ends can belong to
  - 1 contig: not useful for scaffolding (most pairs)
  - 2 contigs: link the contigs



http://en.wikipedia.org/wiki/File:PET_contig_scaffold.png

Assembly output: contigs

- Contigs
  - Ideally one contig (genome)
  - Number of contigs: smaller is better?
- Contig size metrics
  - Larger is better?
  - Min, max, median, N50, N90, …
- Assembly size
  - Sum of all contig sizes (total number of bases in assembly)
  - Ideally genome size
- Number of ambiguous bases (N's)
  - Smaller is better

N50 (and L50)

- Most commonly used assembly metric
  - Others: N75, N90, N95
- Defined over a set of contigs as:
  - The size of the largest contig such that at least half the total size is contained in the contigs of that size or larger
- Example: contigs of sizes 10, 9, 8, 7, 6, 5, 4, 3, 2
  - Total size = 54, half size = 27
  - 10 + 9 + 8 = 27
  - N50 = 8, because 27 $\geq$ 27
  - L50 =3, because 3 contigs used to get N50
- Not always reliable
  - Can be made misleadingly large by erroneously long contigs

Test your understanding of Nx

- Given these contigs:
  - **GATTACA** (length 7)
  - **TACTACTAC** (length 8)
  - **ATTGAT** (length 6)
  - **GAAGA** (length 5)
- What are N50 and L50?
  - Total length = 26, half = 13, so…
- What are N75 and L75?
  - 75% =19.5, so …

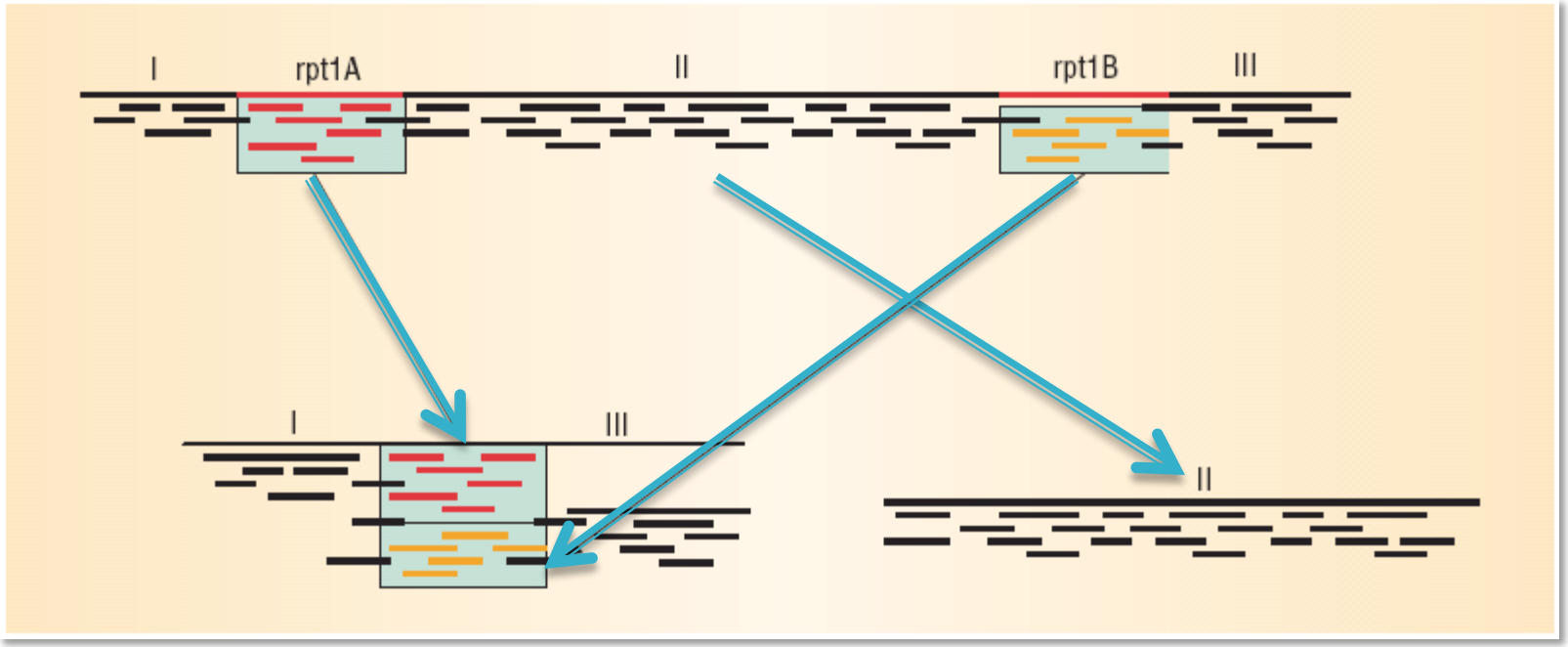
http://rosalind.info/problems/long

Levels of assembly (NCBI)

- Contig level
  - Only contigs, no ordering or further grouping
- Scaffold level
  - (Some) contigs ordered, oriented, assembled into a scaffold (supercontig), gaps filled with N's
- Chromosome level
  - Scaffolds ordered, oriented, assembled into a chromosome, gaps filled with N's
- Complete assembly
  - No sequencing gaps in assembled chromosomes
  - Currently unattainable for higher eukaryotes (e.g. human)

Assembly challenge: repeats

Correct layout of three DNA sequences.



Two copies of a repeat collapsed in a misassembled contig

Orphan contig between two repeat copies

Pop M, Salzberg, SL, Shumway M (2002) Genome sequence assembly: Algorithms and issues. Computer *35*(7), 47-54.

Two types of assembly

- *De novo* assembly
  - No reference genome, only reads
  - Needed for novel genomes
  - Two big groups of algorithms
    - Overlap-Layout-Consensus
    - de Bruijn Graph-based

- Reference alignment/mapping
  - Align reads to a reference genome
  - Sequencing errors not major concern
  - Guide reassembly or variation detection

Shortest superstring problem

- Superstring
  - For a collection of strings, a larger string containing every one of the smaller strings as a substring
- Genome assembly
  - Given DNA strings (reads), find a shortest superstring (contig)

```
>Rosalind_56
ATTAGACCTG
>Rosalind_57
CCTGCCGGAA
>Rosalind_58
AGACCTGCCG
>Rosalind_59
GCCGGAATAC
```

```
ATTAGACCTG
      CCTGCCGGAA
   AGACCTGCCG
         GCCGGAATAC
```

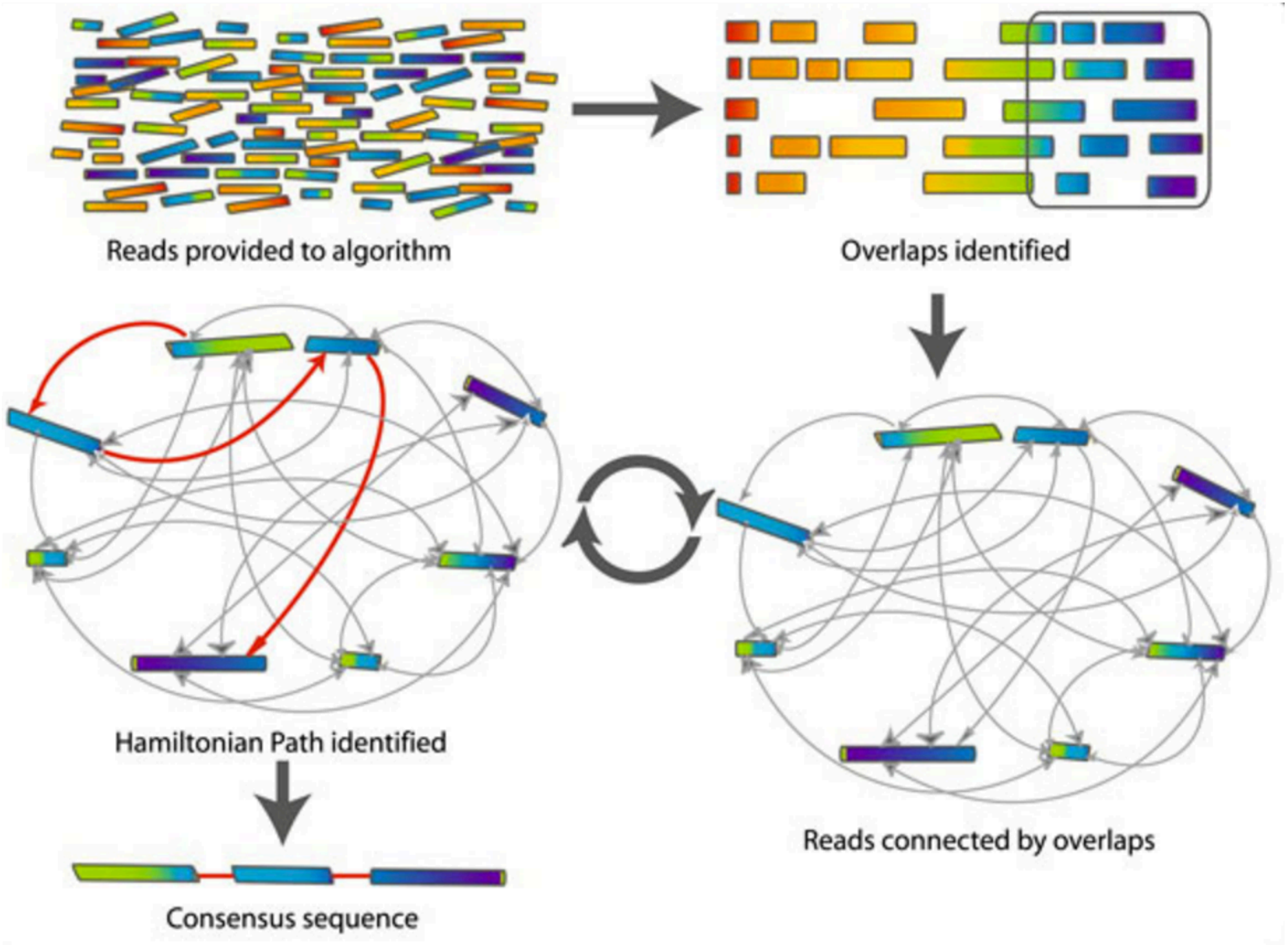
```
>contig_1
ATTAGACCTGCCGGAATAC
```

http://rosalind.info/problems/long

Overlap-layout-consensus (OLC)

- Three steps:
  - Overlaps among all reads are calculated
  - Layout the overlaps onto a graph
    - Overlapping reads (nodes) are connected (edge)
  - Consensus sequence is generated from the graph
- Huge computing resources needed for large data
  - Each read compared against every other read (all-to-all)
- Suitable for
  - Low-coverage long reads
- Examples:
  - Newbler, MIRA, Celera Assembler, CABOG, ARACHNE, Edena

OLC steps



Commins J, Toft C, Fares MA. (2009) Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. Biol Proced Online. 11:52-78.

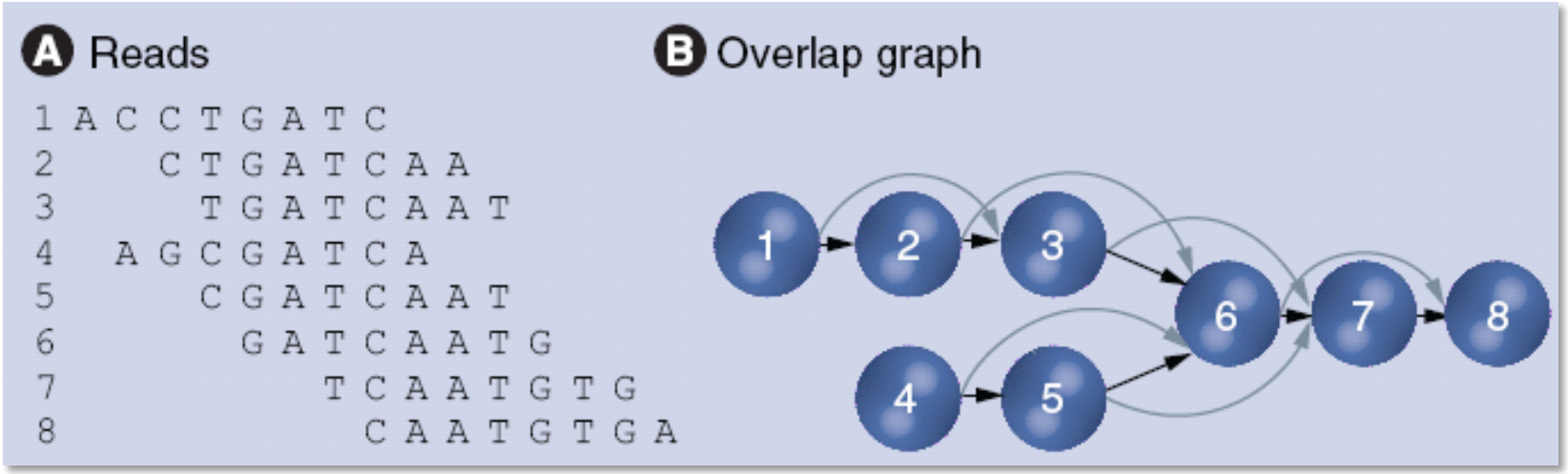
k-mer

- A substring of length *k*
- A string of length *L* has (L–*k*+1) *k*-mers
- Example:
  - If we have a read (string) "**ATGCTCGA**" (length *L*=8) and are looking for 5-mers (*k*=5), there are 4 (L–*k*+1 = 8–5+1) 5-mers :
    - **ATGCT**
    - **TGCTC**
    - **GCTCG**
    - **CTCGA**
- Basic unit of data in de Bruijn graph-based assembly algorithms

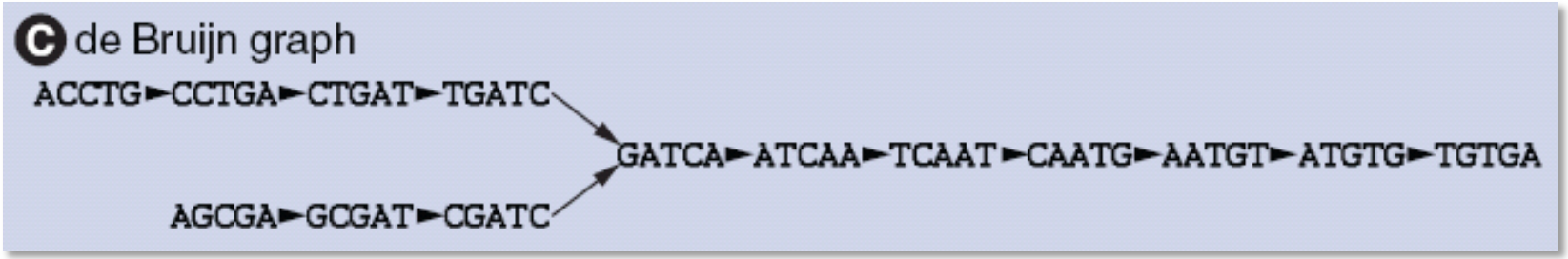
de Bruijn graph (DBG)-based

- Three steps:
  - Chop reads into *k*-mers
  - Form DBG using all *k*-mers
    - Create a node for each unique *k*-mer
    - Connect two nodes if two *k*-mers are adjacent in a read with *k*−1 base overlap
  - Infer genome sequence from the DBG
- Suitable for
  - Deep-sequenced short reads
  - Large genome assembly
- Examples:
  - ULER-USR, Velvet, ABySS, ALLPATHS-LG, SOAPdenovo, SPAdes

OLC vs. DBG



Nodes: reads, edges: overlaps



Nodes: *k*-mers, edges: adjacent *k*-1 base overlap (*k* = 5)

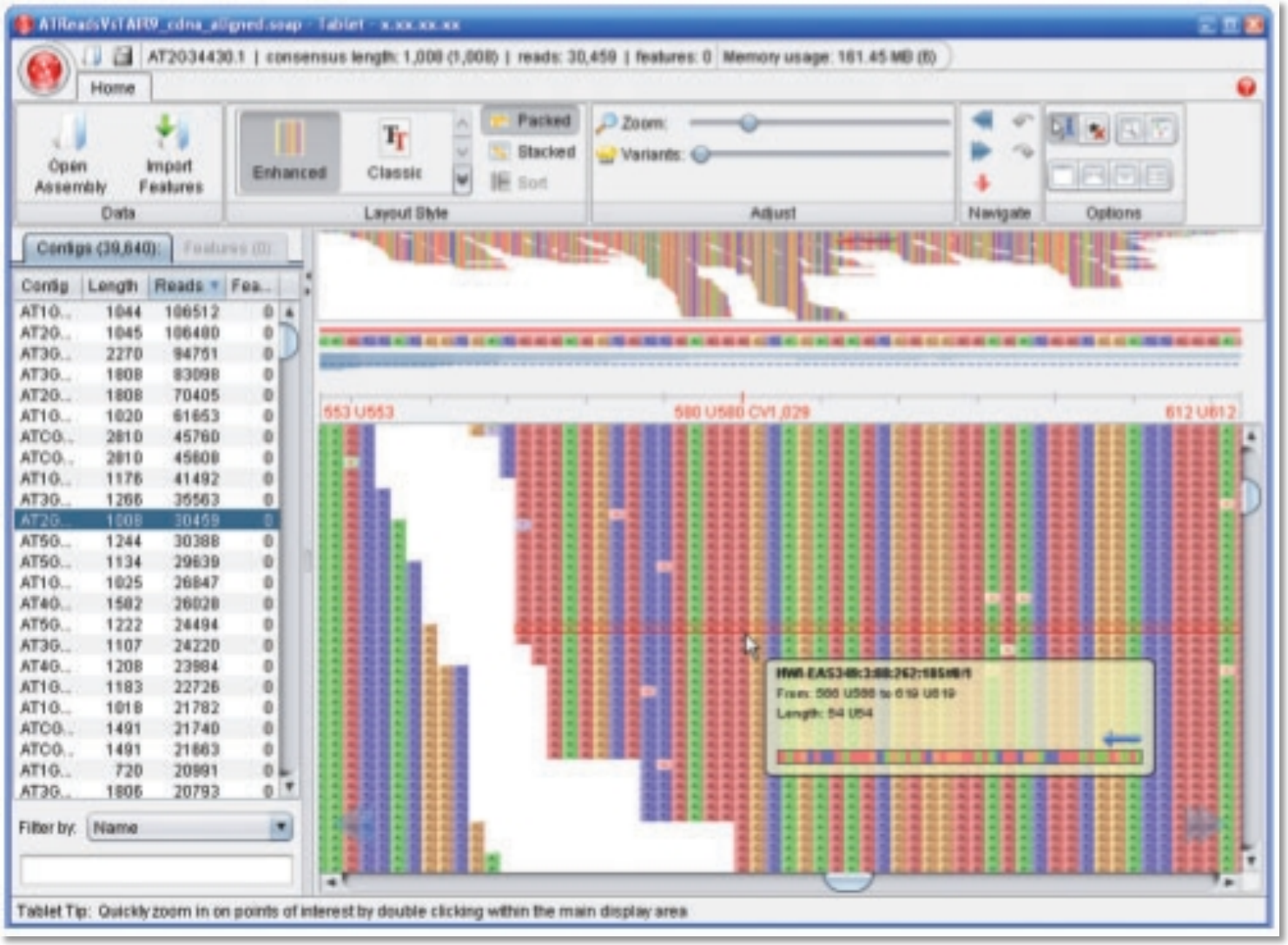
Henson J, Tischler G, Ning Z (2012) Next-generation sequencing and large genome assemblies. Pharmacogenomics 13(8), 901-915

Genome assembly visualization

- Visualization benefits
  - View which reads map to which contigs
  - Find/view paired-end reads
  - View coverage depth profiles
  - Coloration to highlight different aspects of assembly
  - Often detailed assembly metrics provided

- Assembly visualization tools
  - Tablet, EagleView, Icarus, Bandage

Tablet visualization example



Milne I, Bayer M, Cardle L, et al. Tablet—next generation sequence assembly visualization. Bioinformatics. 2010;26(3):401-402.

Icarus visualization example