



# Assembly

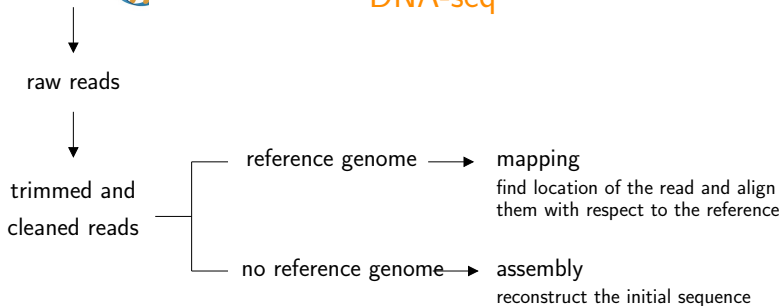
Hélène Touzet

`helene.touzet@univ-lille.fr`

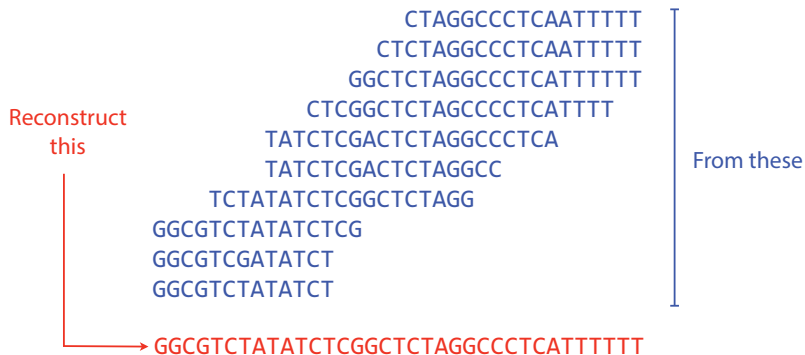
CNRS, Bonsai, CRISAL



## DNA-seq



# The assembly problem

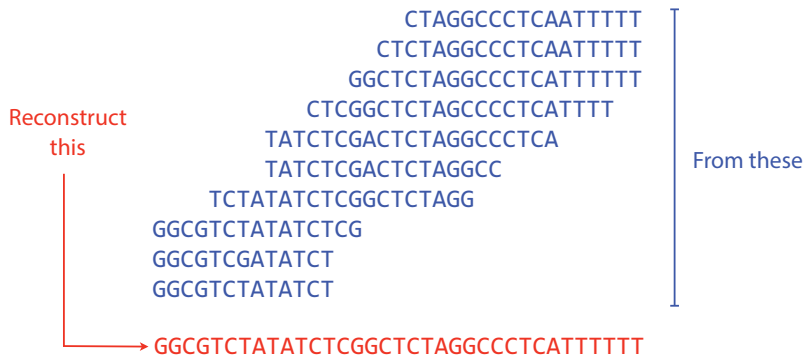


Courtesy of Ben Langmead (Johns Hopkins University)

# Why assembling reads ?

- annotation of genomes
  - discovery of new genes
  - gene order, structural variants
  - noncoding regions
  - evolutionnary genomics, phylogenomics
- transcriptome
  - reconstruction of transcripts
  - identification of alternative transcripts
- metagenomics
  - identification of species

# The assembly problem



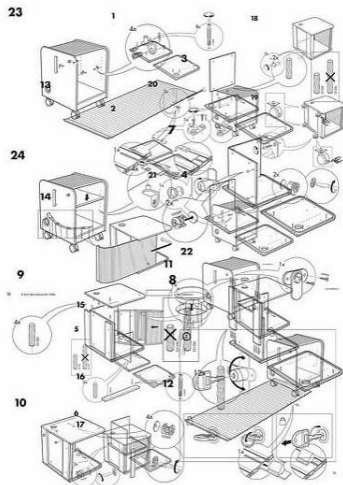
Reconstruct  
this

CTAGGCCCTCAATTTT  
GGCGTCTATATCT  
CTCTAGGCCCTCAATTTT  
TCTATATCTCGGCTCTAGG  
GGCTCTAGGCCCTCATTTTT  
CTCGGCTCTAGCCCCTCATTTT  
TATCTCGACTCTAGGCCCTCA  
GGCGTCGATATCT  
TATCTCGACTCTAGGCC  
GGCGTCTATATCTCG

From these

GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT





*Some assembly is required*



# How to assemble reads ?

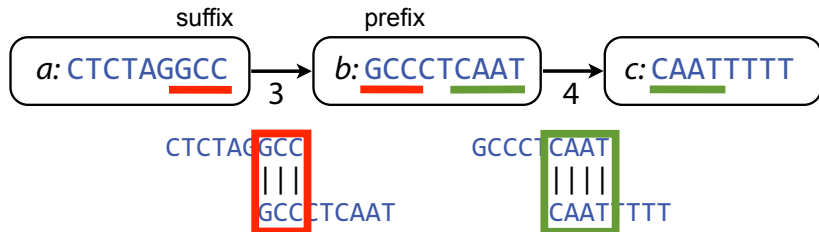
Historical perspective



Key features : overlaps between reads

# How to assemble reads ?

## Historical perspective



Key features : overlaps between reads

$R_1$ 

C	T	G	A	G	A	A	C	C	T	G	T
---	---	---	---	---	---	---	---	---	---	---	---

 $R_2$ 

C	C	T	G	T	A	A	G	A	T
---	---	---	---	---	---	---	---	---	---

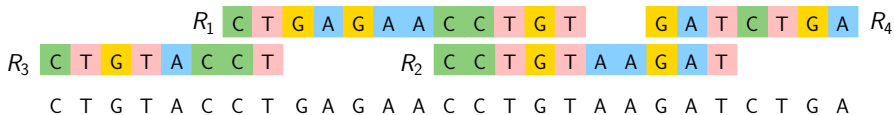
 $R_3$ 

C	T	G	T	A	C	C	T
---	---	---	---	---	---	---	---

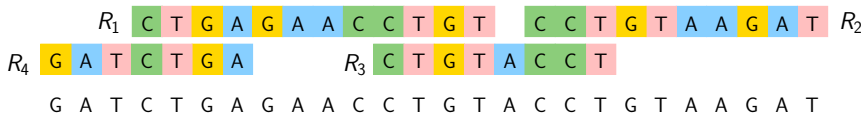
 $R_4$ 

G	A	T	C	T	G	A
---	---	---	---	---	---	---

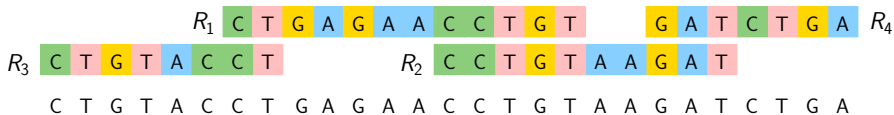




Length of the assembly : 27

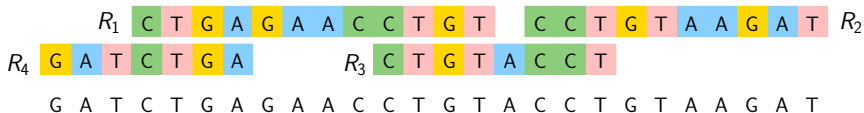


Length of the assembly : 26



Length of the assembly : 27

joining together the reads in decreasing order of the quality of their overlaps



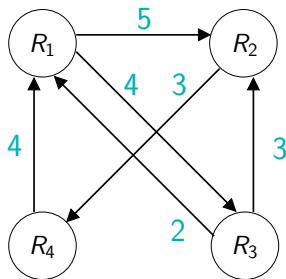
Length of the assembly : 26

trying to maximize the total length of read overlaps



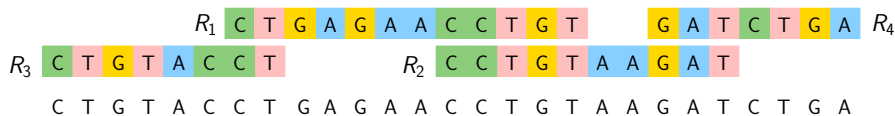
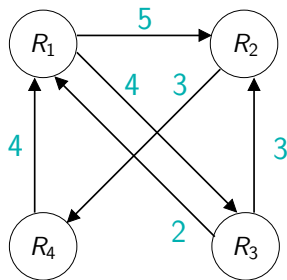
$\nearrow$	$R_1$	$R_2$	$R_3$	$R_4$
$R_1$		5	4	0
$R_2$	0		0	3
$R_3$	2	3		0
$R_4$	4	0	0	

Length of the longest suffix of  $R_i$   
which is also a prefix of  $R_j$

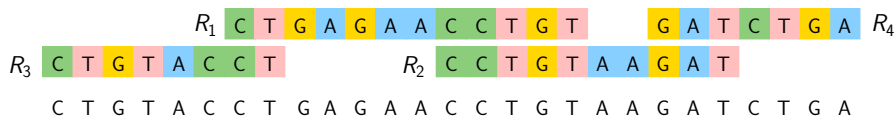
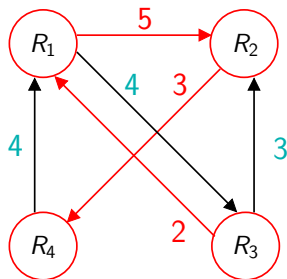


Overlap graph

# Paths in the graph

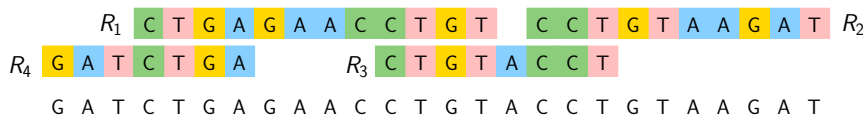
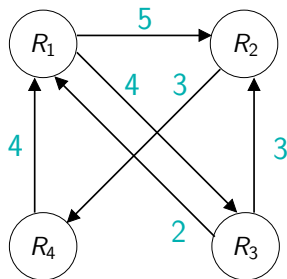


# Paths in the graph

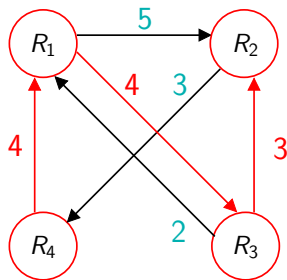




# Paths in the graph



# Paths in the graph



several paths = several assemblies

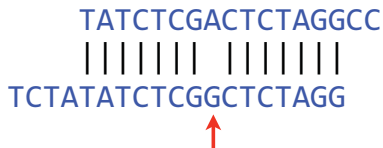
# Overlap assemblies in real life

- risk of contamination
- existence of sequencing errors
- existence of repeats
- diploid and polyploid genomes
- low coverage or uneven coverage

+ unable to handle the large number of NGS sequencing reads

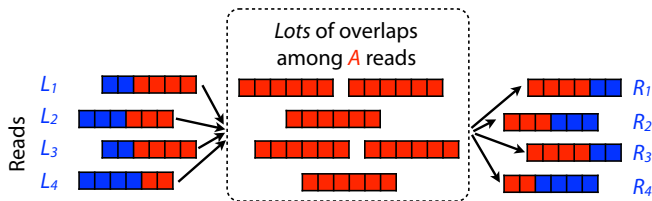
# Sequencing errors

TATCTCGACTCTAGGCC  
||||| |||||  
TCTATATCTCGGCTCTAGG



- Approximate overlaps : Construction of alignments between reads
- Assembly : consensus sequence

# Repeats



The region  $A$  is longer than reads

# Diploidy and polyploidy

Read from Mother: TATCTCGACTCTAGGCC

|||||

Read from Father: TCTATATCTCGGCTCTAGG

Sequence from Mother: TCTATATCTCGACTCTAGGCC

Sequence from Father: TCTATATCTCGGCTCTAGGCC

# Coverage

CTAGGCCCTCAATTTTT  
CTCTAGGCCCTCAATTTTT  
GGCTCTAGGCCCTCATTTTTT  
CTCGGCTCTAGCCCCTCATTTT  
TATCTCGACTCTAGGCCCTCA  
TATCTCGACTCTAGGCC  
TCTATATCTCGGCTCTAGG  
GGCGTCTATATCTCG  
GGCGTCGATATCT  
GGCGTCTATATCT  
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

Coverage = 5

CTAGGCCCTCAATTTT  
CTCTAGGCCCTCAATTTT  
GGCTCTAGGCCCTCATTTTT  
CTCGGCTCTAGCCCCTCATTTT  
TATCTCGACTCTAGGCCCTCA  
TATCTCGACTCTAGGCC  
TCTATATCTCGGCTCTAGG  
GGCGTCTATATCTCG  
GGCGTCGATATCT  
GGCGTCTATATCT  
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

177 bases

35 bases

Average coverage =  $177 / 35 \approx 5$ -fold



# Overlaps - Historical perspectives

- Sanger sequencing
  - Celera (Myers, 2000)  
originally developed for the assembly of the human genome
  - SGA (Simpson, Durbin, 2012)
- not suitable for NGS short reads (Illumina)  
computationally expensive : construction of the graph, size of the graph, path discovery
- comeback with long reads (Nanopore, Pacbio)

# De Bruijn graphs

- introduced in bioinformatics to deal with NGS data
- used by almost all modern short-reads assembly tools  
seminal : Velvet (2008), Abyss (2009), SOAPdenovo2 (2012)  
state-of-the art : SPAdes (2012), MaSuRCA (2013), Megahit (2015)...

Genome assembly reborn : recent computational challenges. M. Pop, Briefings in Bioinformatics 2009 <https://doi.org/10.1093/bib/bbp026>

How to apply de Bruijn graphs to genome assembly. P.E.C. Compeau, P.A. Pevzner, G. Tesler, Nature Biotechnology 2011 [doi:10.1038/nbt.2023](https://doi.org/10.1038/nbt.2023)

# Rationale

- The genome can be reconstructed from the  $k$ -mers it contains
- Reads are decomposed into  $k$ -mers

How many distinct 3-mers are they in

$R_1$       C T G A G A A C C T G T

$R_2$       C C T G T A A G A T

$R_3$       C T G T A C C T

$R_4$       G A T C T G A

A A C

A T C

G A G

T A C

A A G

C C T

G A T

T C T

A C C

C T G

G T A

T G A

A G A

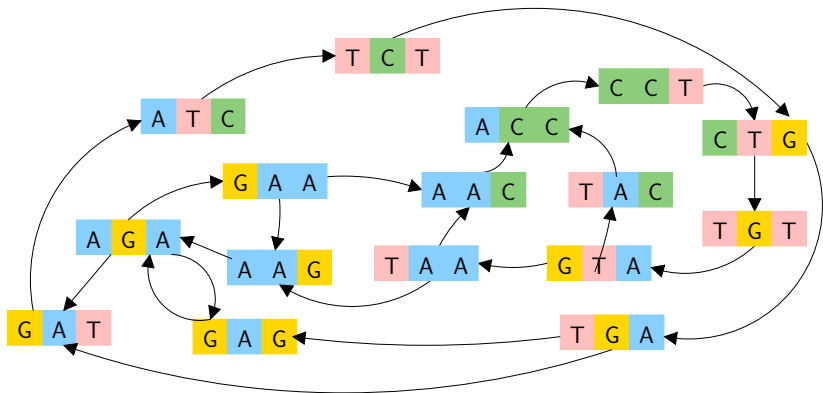
G A A

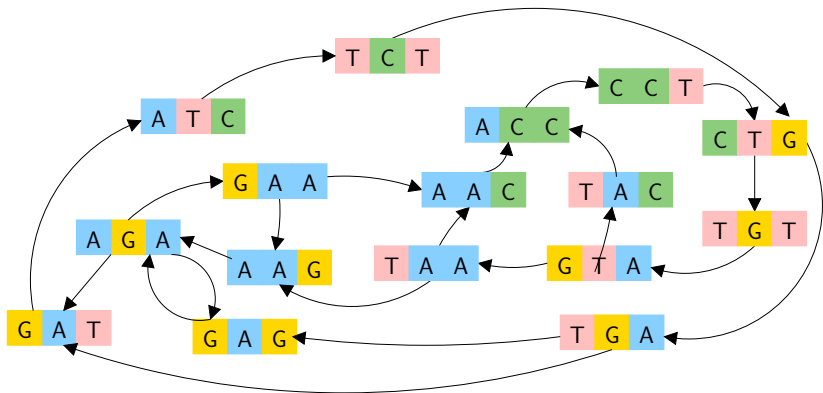
T A A

T G T

# De Bruijn Graph

- Nodes :  $k$ -mers present in the reads
- Arcs : overlaps of length  $k - 1$  between  $k$ -mers  
Do not depend on the set of reads
- Easy to construct, low memory footprint  
Great advantage over overlap graphs

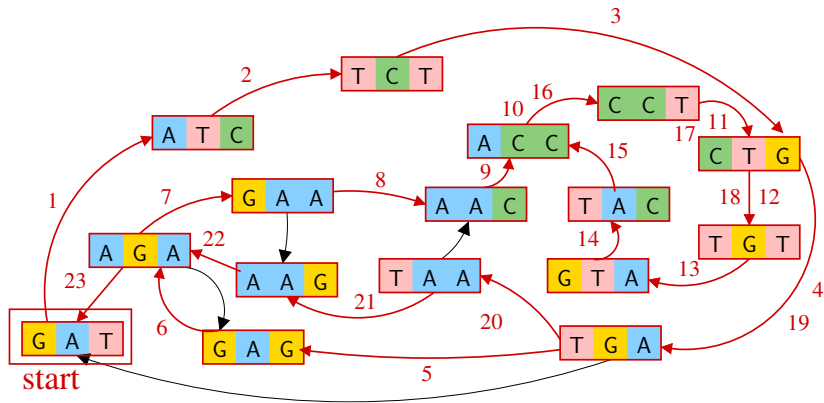




Assembly = path in the graph

Several paths = several assemblies





$R_1$  C T G A G A A C C T G T C C T G T A A G A T  $R_2$   
 $R_4$  G A T C T G A  $R_3$  C T G T A C C T  
 G A T C T G A G A A C C T G T A C C T G T A A G A T

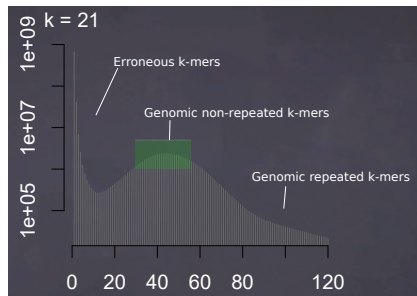
# De Bruijn Graphs in practice - choice of $k$

## Length of $k$ -mers

- small  $k$  :
  - pro : more non-erroneous  $k$ -mers
  - cons : less signal, more random overlaps, repeat collapsing
- large  $k$  :
  - pro : higher signal, less random overlaps, less repeat collapsing
  - cons : more erroneous  $k$ -mers
- generally  $k \geq 20$  (may be longer for large genomes)
- higher sequencing coverage means larger  $k$  values can be used
- multi- $k$  assembly ( $k = 21 \rightarrow k = 55 \rightarrow k = 72$ )  
IDBA, SPAdes, Megahit

In this lecture : SPAdes

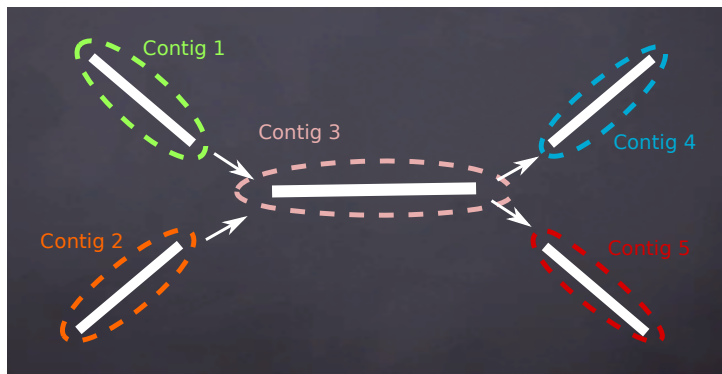
# De Bruijn Graphs in practice - cleaning $k$ -mers



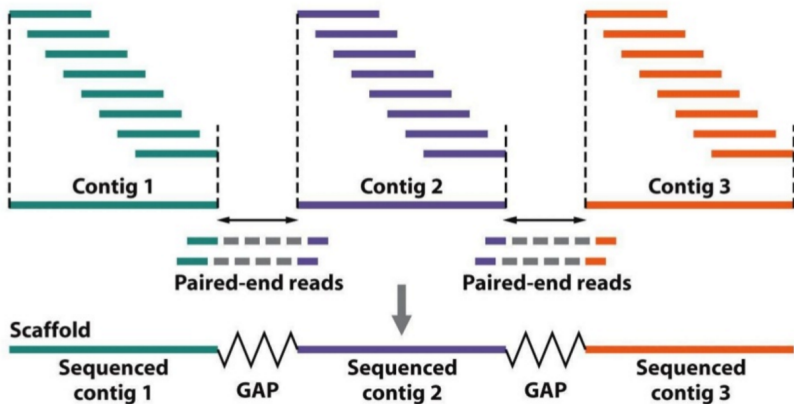
Courtesy of Rayan Chikhi (Institut Pasteur)  
horizontal axis : number of occurrences  
vertical axis : number of  $k$ -mers

- $k$ -mers with low frequency are likely to contain sequencing errors
- they can be removed before the construction of the graph

# De Bruijn Graphs in practice - contigs and scaffolds



Contigs = *simple* paths in the graph



Scaffold = link between contigs using paired-end reads  
Error-prone

# Short read assembly is still difficult

even with De Bruijn graphs

- risk of contamination
- existence of sequencing errors **solved**
- existence of repeats
- diploid and polyploid genomes
- low coverage or uneven coverage

+ unable to handle the large number of NGS sequencing reads  
**solved**

# Short read assembly is still difficult

- library design
  - longest read lengths
  - coverage  $\geq 50x$ ,  $\times$  ploidy number
  - for 1 bacterial genome, no point going above 200x
  - BROAD recipe : several mate pairs libraries of increasing size
- assembler
  - SPAdes for small genomes
  - unclear for large genomes
  - try at least two assemblers, try different parameters
  - high computational requirements overall
- an assembly is not the absolute truth, it is a mostly complete, generally fragmented and mostly accurate hypothesis

# How to compare/analyse assemblies?

- no trivial ranking between assemblies
- no simple criteria
- assembly with high coverage and short contigs / assembly with low coverage and long contigs



# Quast

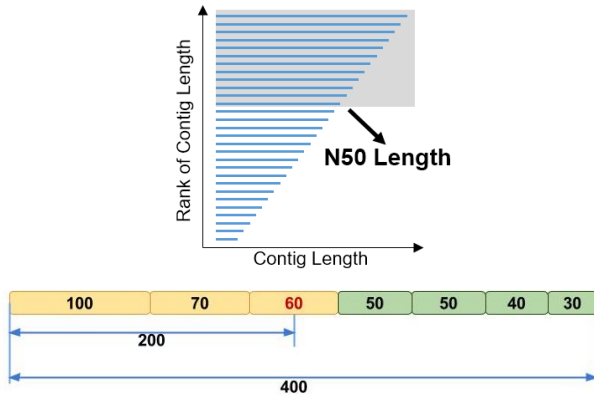
## Quality Assessment Tool for Genome Assemblies

- provides a large number of statistics and metrics : contigs, missamblies, functional elements
- works both with and without a reference genome
- accepts multiple assemblies, thus is suitable for comparison

QUAST : quality assessment tool for genome assemblies. Bioinformatics 2013 <https://doi.org/10.1093/bioinformatics/btt086>

## Contigs

- number of contigs
- length of the largest contig
- total number of bases in the assembly (sum of contig lengths)



- *N50* : contig length  $N$  for which 50% of all bases in the sequences are in a contig of length  $L \geq N$
- *NG50* : contig length such that using equal or longer length contigs produces 50% of the expected length of the reference genome

## Misassemblies (requires a reference genome)

- **missassembly breakpoints** : position in the contig where the left flanking sequence aligns over 1 kb away from the right flanking sequence on the reference, or they overlap  $>1$  kb, or align on opposite strands or different chromosomes :
- **metrics** : total number of missambly breakpoints, number of contigs that contain misassembly breakpoints, number of bases contained in all contigs that have one or more misassemblies (Mummer)
- **number of unaligned contigs** : contigs that have no alignment to the reference sequence
- **number of ambiguously mapped contigs** : contigs that have multiple alignments to the reference genome

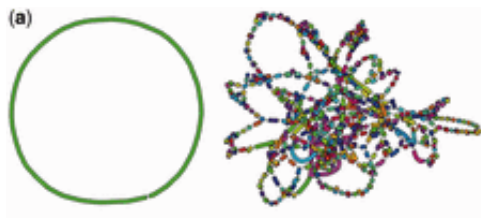
## Functional elements

- genome fraction (%) : number of aligned bases in the reference, divided by the genome size.
- duplication ratio : number of aligned bases in the assembly divided by the number of aligned bases in the reference
- number of mismatches and number of indels per 100 kb
- number of genes based on a user-provided annotated list of gene positions in the reference genome
- number of predicted genes in the assembly (GeneMark.hmm for prokaryotes and GlimmerHMM for eukaryotes)

# Bandage

## Bioinformatics Application for Navigating De novo Assembly Graphs Easily

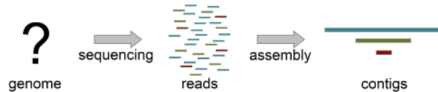
- interactive visualization of the assembly graph (such as de Bruijn graph)
- <https://github.com/rrwick/Bandage/wiki/>



Left : ideal bacterial assembly with one single contig  
Right : poor assembly with many short contigs



A researcher (who does not yet know the structure of the genome) sequences it, and the resulting 100 bp reads are assembled with a *de novo* assembler:



Because the repeated element is longer than the sequencing reads, the assembler was not able to reproduce the original genome as a single contig. Rather, three contigs are produced: one for the repeated sequence (even though it occurs twice), and one for each sequence between the repeated elements.

Given only the contigs, the relationship between these sequences is not clear. However, the assembly graph contains additional information which is made apparent in Bandage:



There are two principal underlying sequences compatible with this graph: two separate circular sequences that share a region in common, or a single larger circular sequence with an element that occurs twice:



## Exercise

Bacterial training dataset for Galaxy training network tutorials on Genome assembly : imaginary *Staphylococcus aureus* bacterium with a miniature genome.

<https://zenodo.org/record/582600>

- Download reads : mutant\_R1.fastq and mutant\_R1.fastq  
Paired-end, 150 bases long, read coverage 19x
- Galaxy : Create a collection of list of pairs
- Assembly with SPAdes, mode only assembler





Regular

Composite

Collection

Rule-based

 Drop files here

List

Pair

List of Pairs

Collection:

List of Pairs

Upload from Disk or Web


all:


fastqsanger



Reference (set all):

unspecified (?)

 Choose local file

 Choose remote files

 Paste/Fetch data

Start

Build

Pause

Reset

Close

- Quality metrics with Quast
  - Assembly file : scaffolds
  - Reference genome :  
*ASM1342v1 Staphylococcus aureus subsp. aureus NCTC 8325*  
→ NCBI
  - Report : PDF



