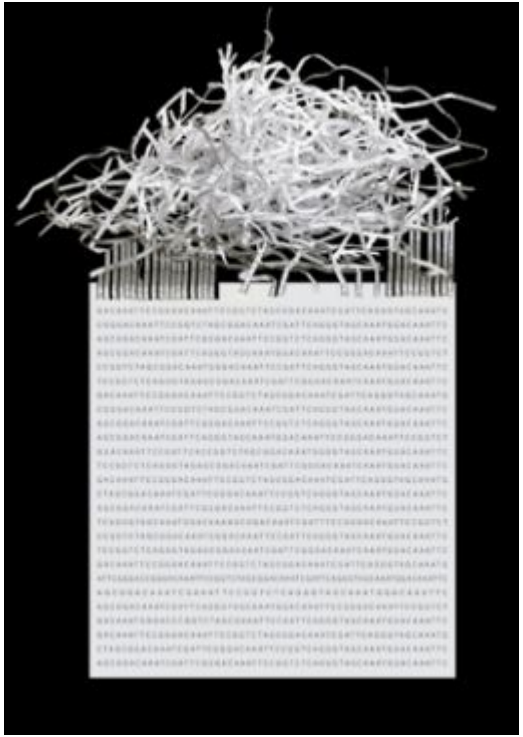
Week 3: Where do all these references come from, anyway? (de novo assembly)



Why to assemble?

Why to assemble?

- Sequencing data
  - Billions of short reads
  - Sequencing errors
  - Contaminants

Hard to perform analysis

- Assembly
  - ✓ Corrects sequencing errors
  - ✓ Much longer sequences
  - ✓ Each genomic region is presented only once
  - × May introduce errors

Why to assemble?

when the genome of an organism of interest is unknown (no reference)

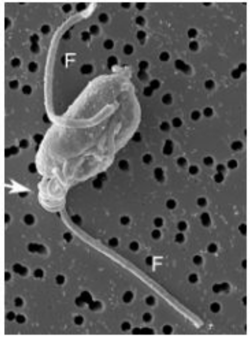
non-model species

conservation biology

very divergent species

Li, 2009

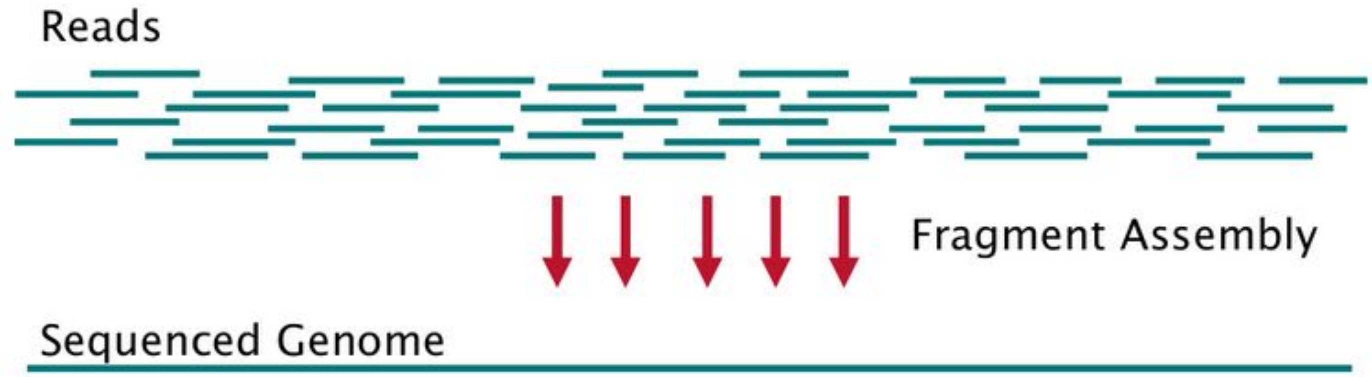
Cho, 2013



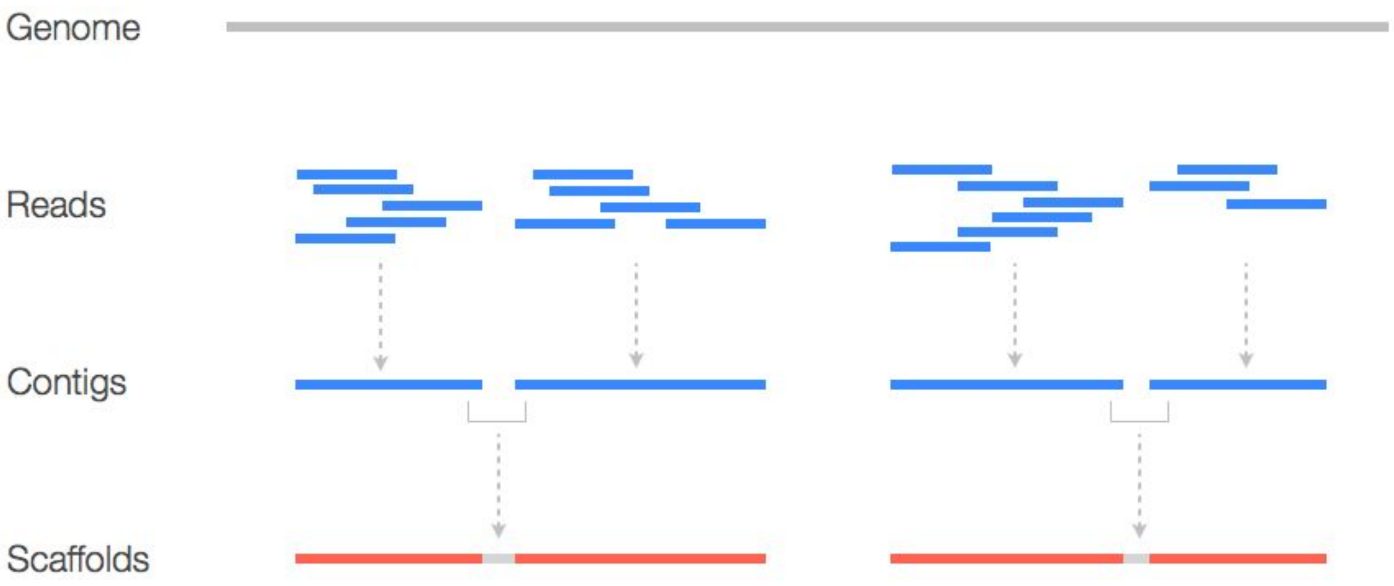
Hovde, 2015

Assembly basics

Assembly in a perfect world

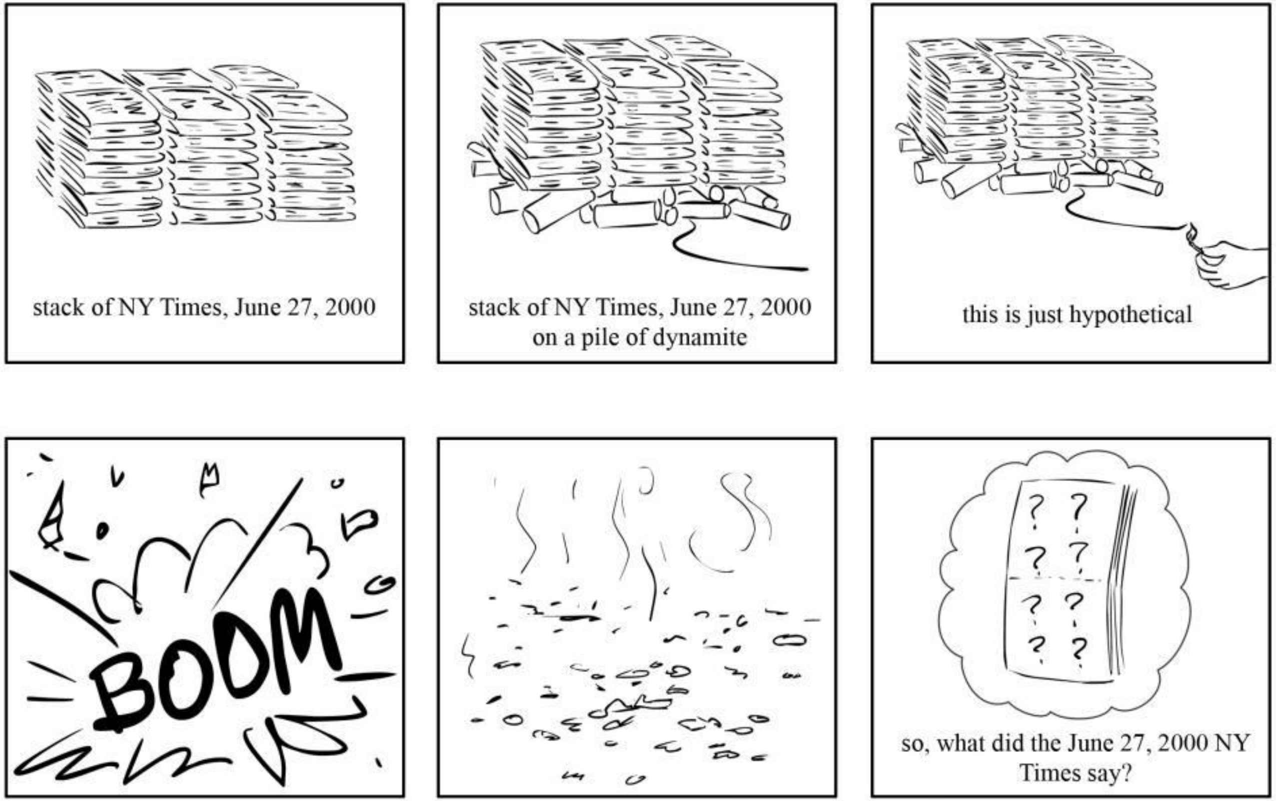


Assembly in real world



De novo whole genome assembly

De novo whole genome assembly


stack of NY Times, June 27, 2000
stack of NY Times, June 27, 2000
on a pile of dynamite
this is just hypothetical
BOOM
? ? ? ? ? ? ? ?
so, what did the June 27, 2000 NY
Times say?

Early days

- Sanger sequencing
  - Long reads
  - Low coverage

- Overlap-Layout-Consensus (OLC)
  - Find overlaps between all reads (BLAST)
  - Order reads according to the overlaps
  - Merge reads into consensus string

NGS and OLC

- Overlap-Layout-Consensus is not applicable
  - Hard to find overlaps between short reads
  - Impossible to scale to such amount of reads
- De Bruijn graph approach

  (Pevzner et al., 2001)
  (Zerbino et al., 2008)
- String Graph approach

  (Meyers, 2005)
  (Simpson, Durbin 2011)

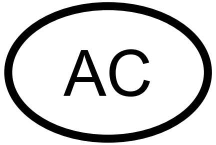
De Bruijn graph

ACGTCCGTAA

De Bruijn graph

ACGTCCGTAA

k=2



De Bruijn graph

ACGTCCGTAA

k=2

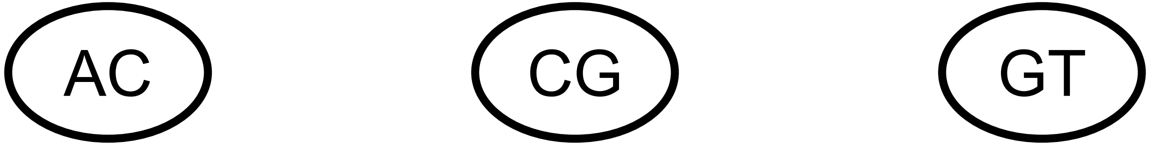
AC

CG

De Bruijn graph

ACGTCCGTAA

k=2



AC

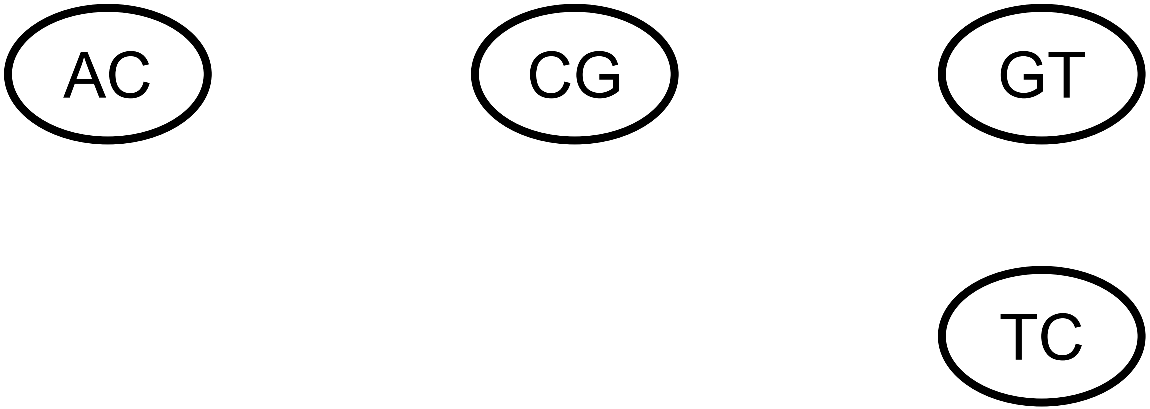
CG

GT

De Bruijn graph

ACGTCCCGTAA

k=2



AC

CG

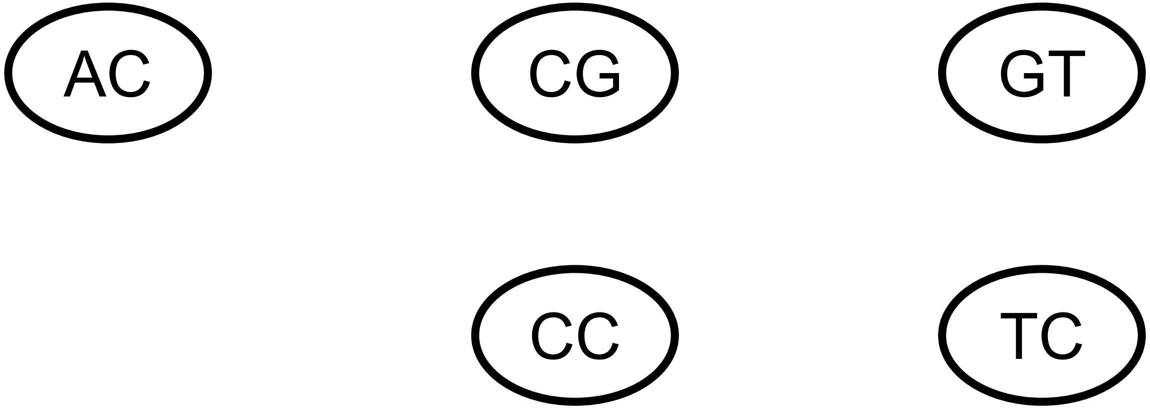
GT

TC

De Bruijn graph

ACGTCCGTAA

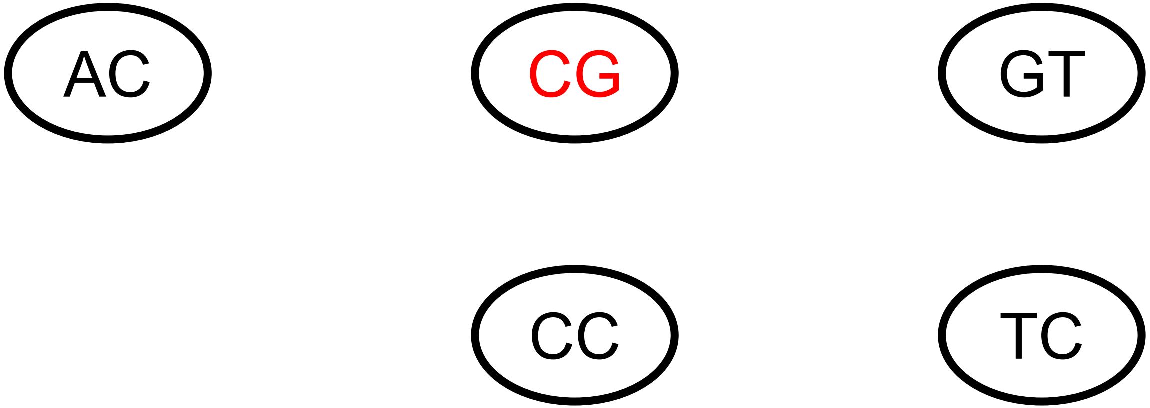
k=2



De Bruijn graph

ACGTCCGTAA

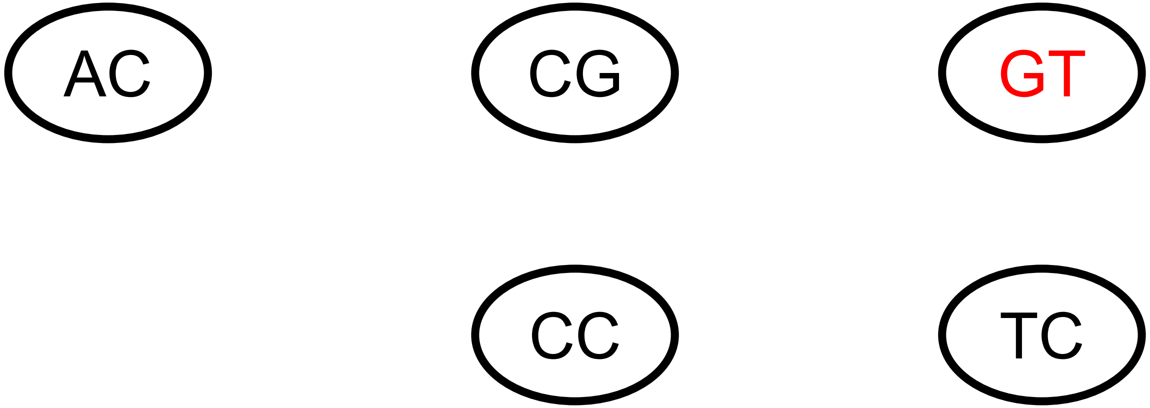
k=2



De Bruijn graph

ACGTCCGTAA

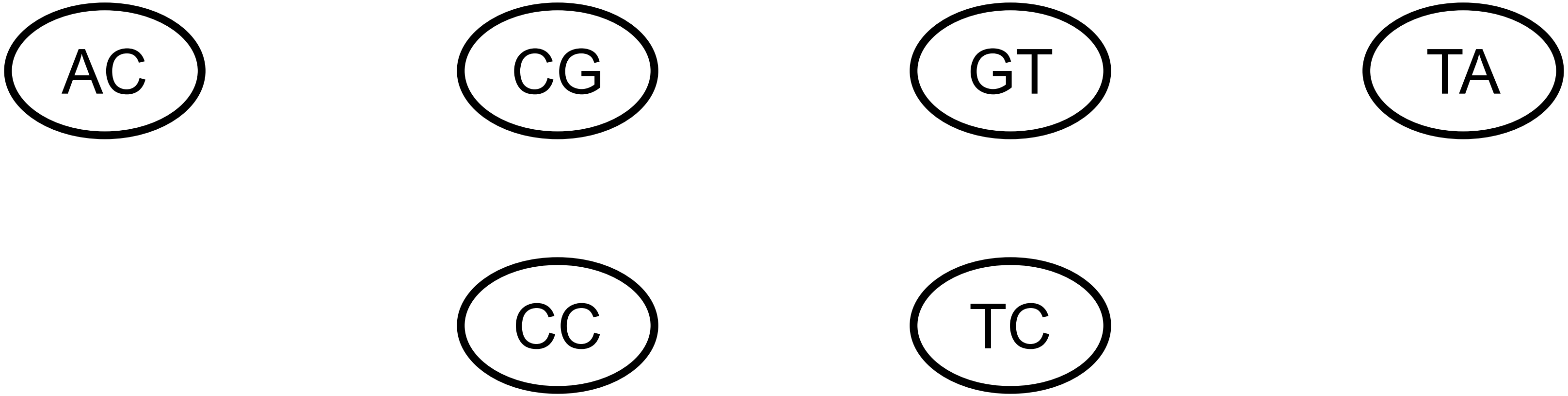
k=2



De Bruijn graph

ACGTCCGTAA

k=2



AC

CG

GT

TA

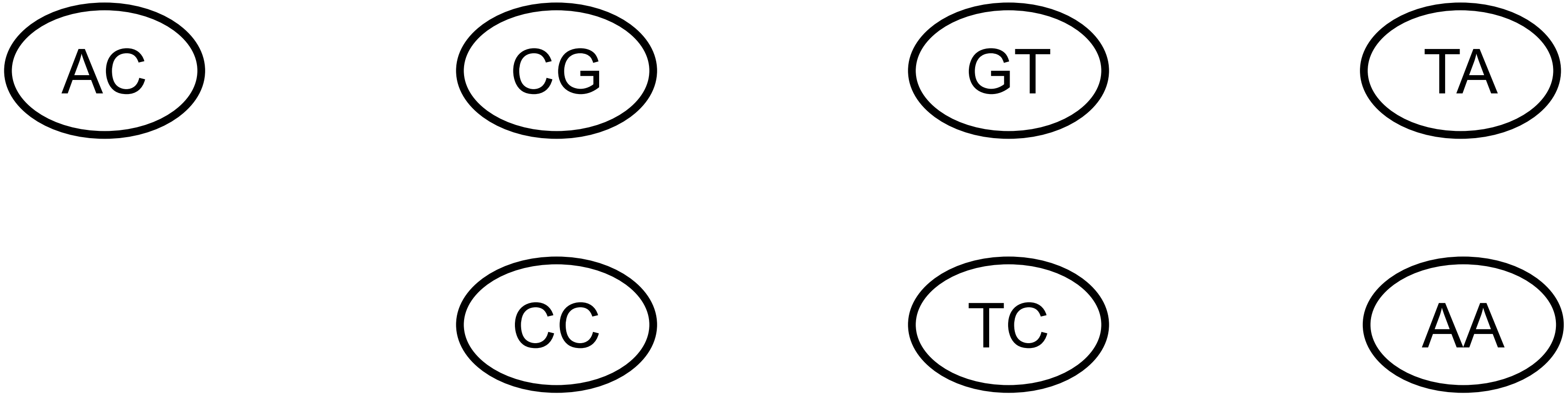
CC

TC

De Bruijn graph

ACGTCCGTAA

k=2



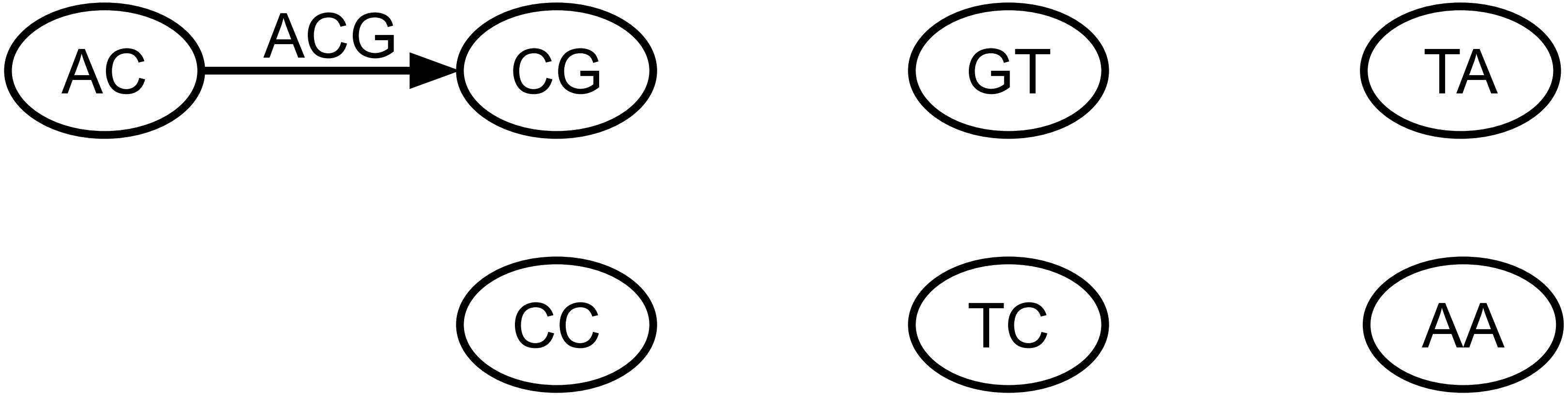
AC CG GT TA

CC TC AA

De Bruijn graph

ACGTCCGTAA

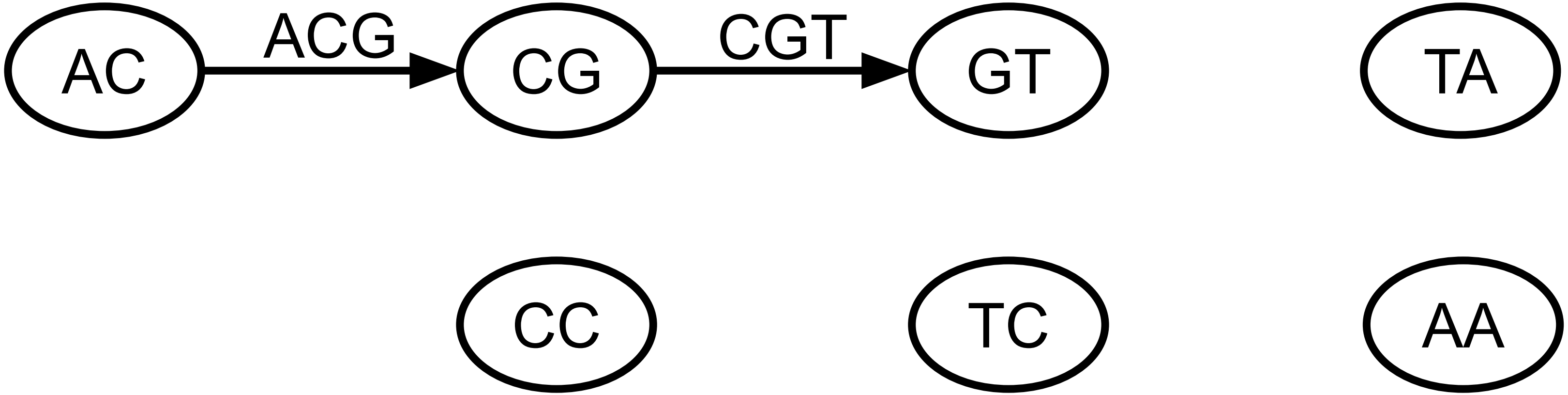
k=2



De Bruijn graph

ACGTCCGTAA

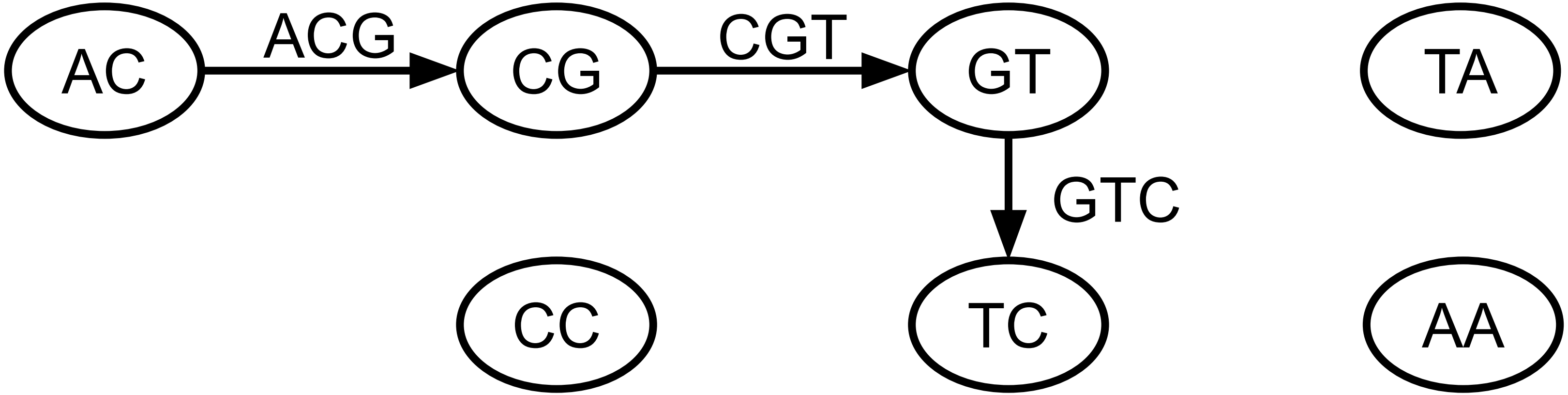
k=2



De Bruijn graph

ACGTCCGTAA

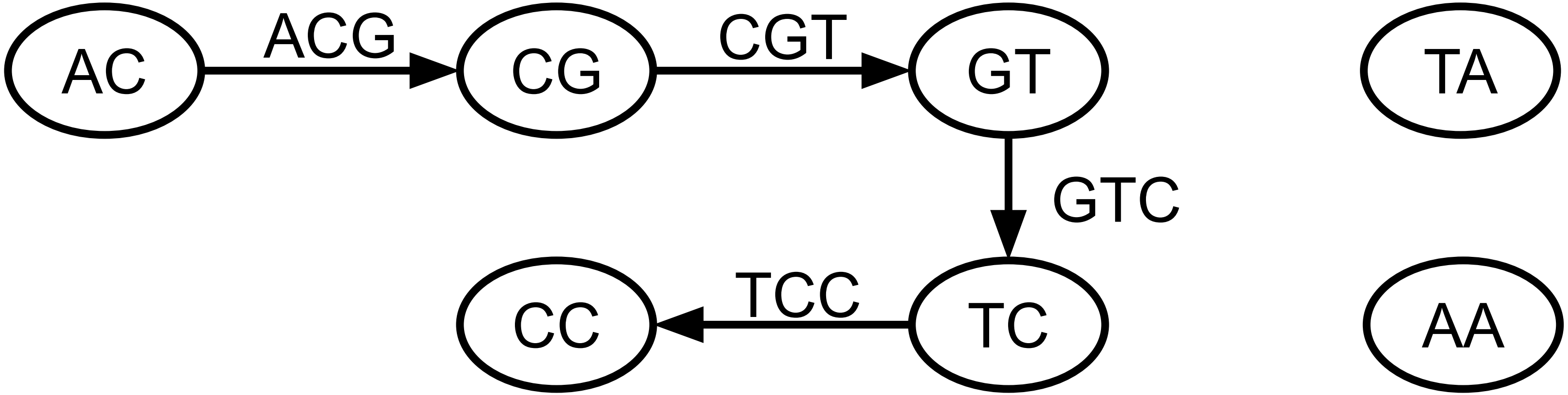
k=2



De Bruijn graph

ACGTCCGTAA

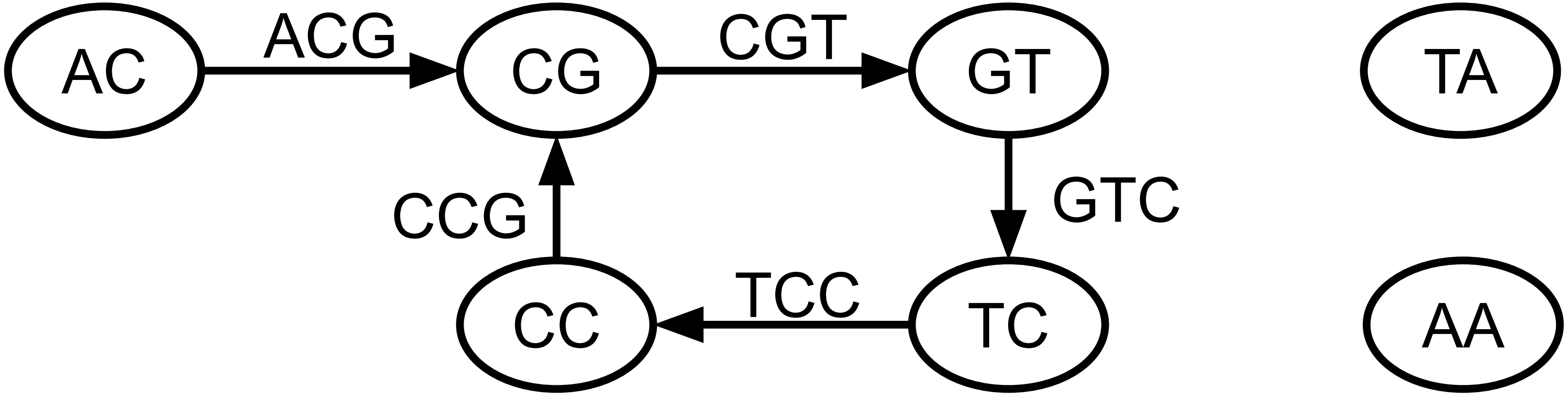
k=2



De Bruijn graph

ACGTCCGTAA

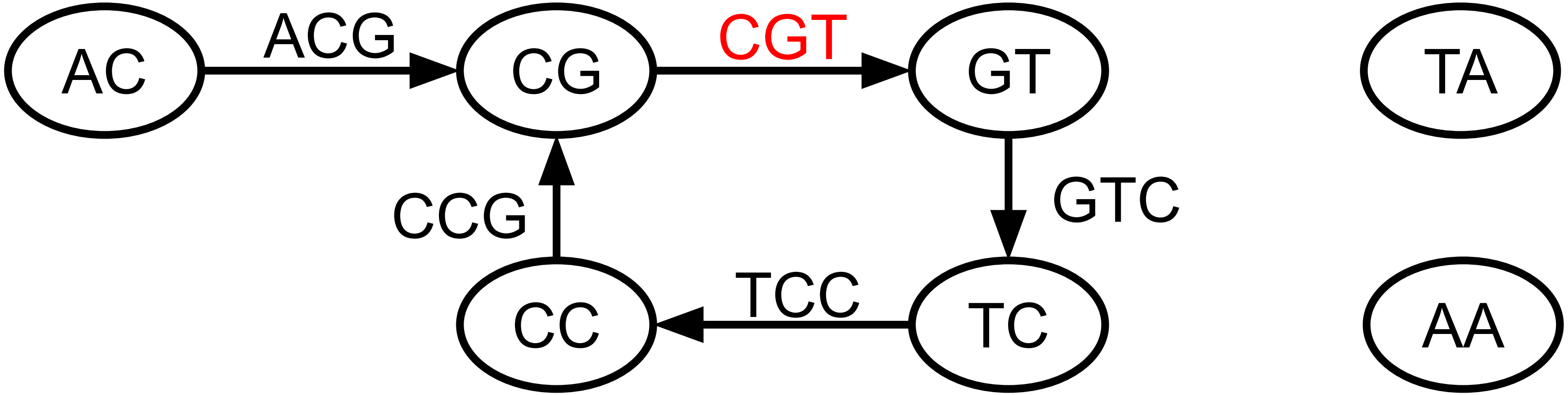
k=2



De Bruijn graph

ACGTCCGTAA

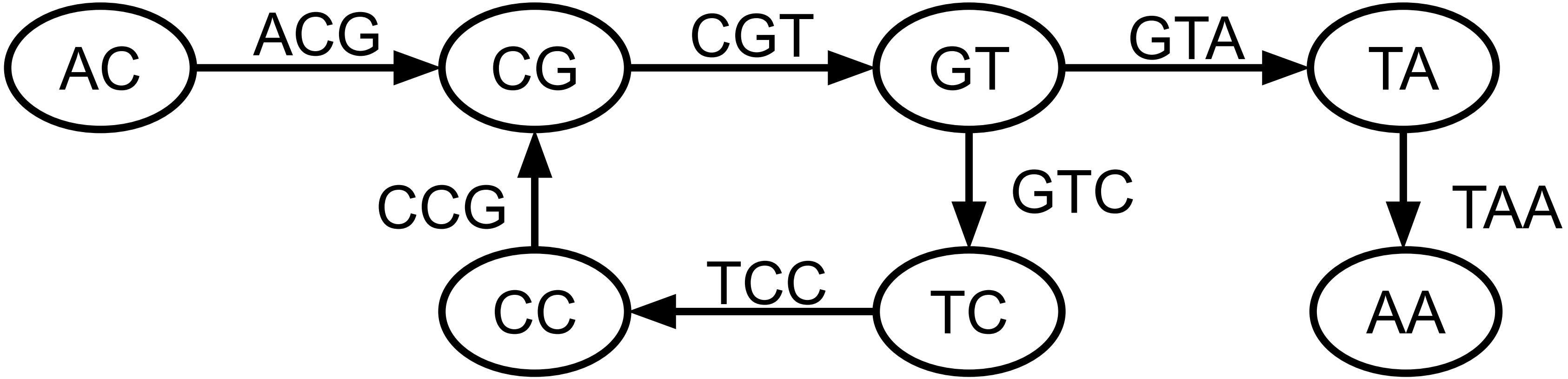
k=2



De Bruijn graph

ACGTCCGTAA

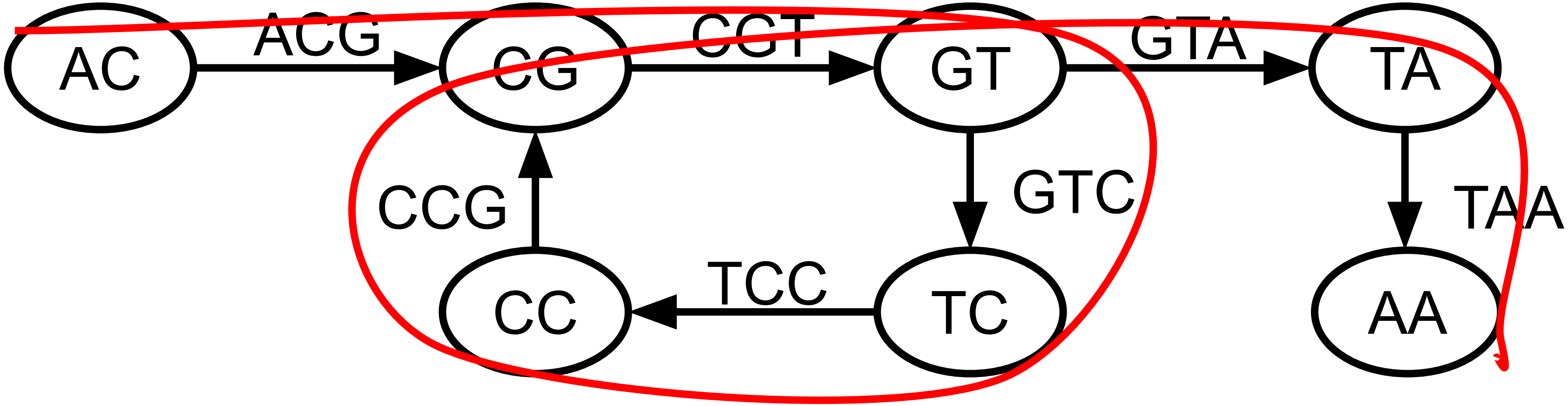
k=2



De Bruijn graph

ACGTCCGTAA

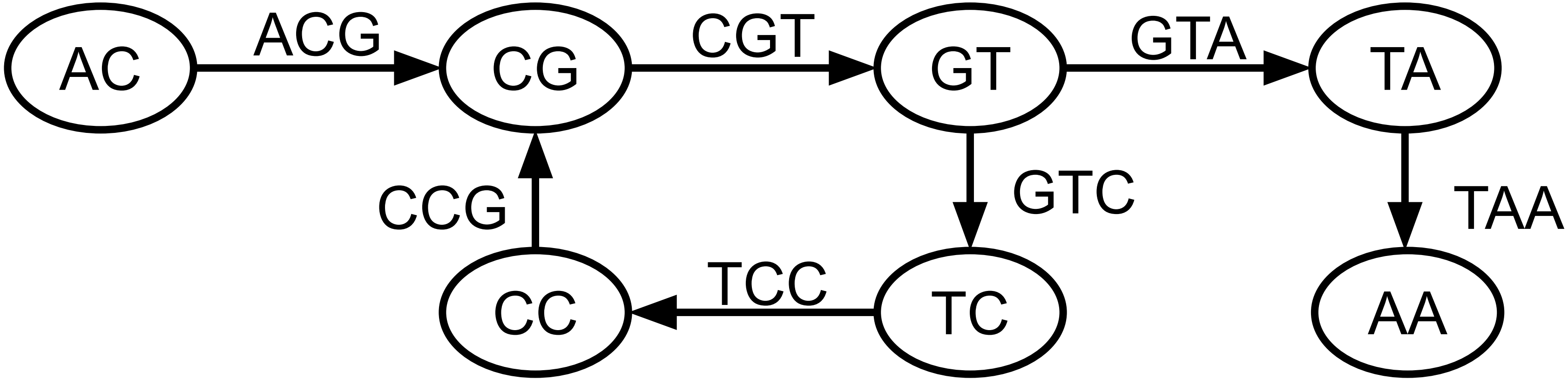
k=2



Condensed de Bruijn graph

ACGTCCGTAA

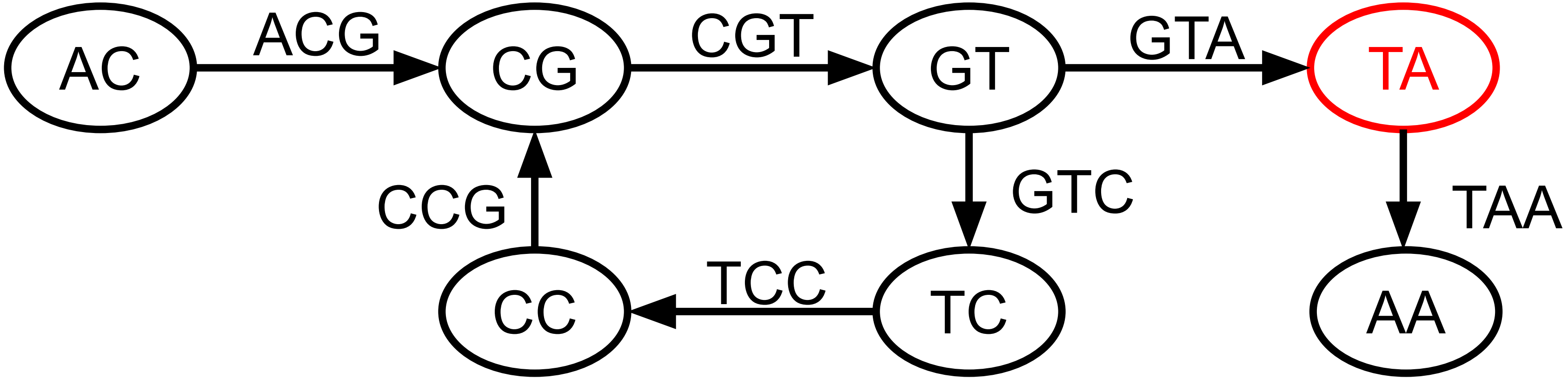
k=2



Condensed de Bruijn graph

ACGTCCGTAA

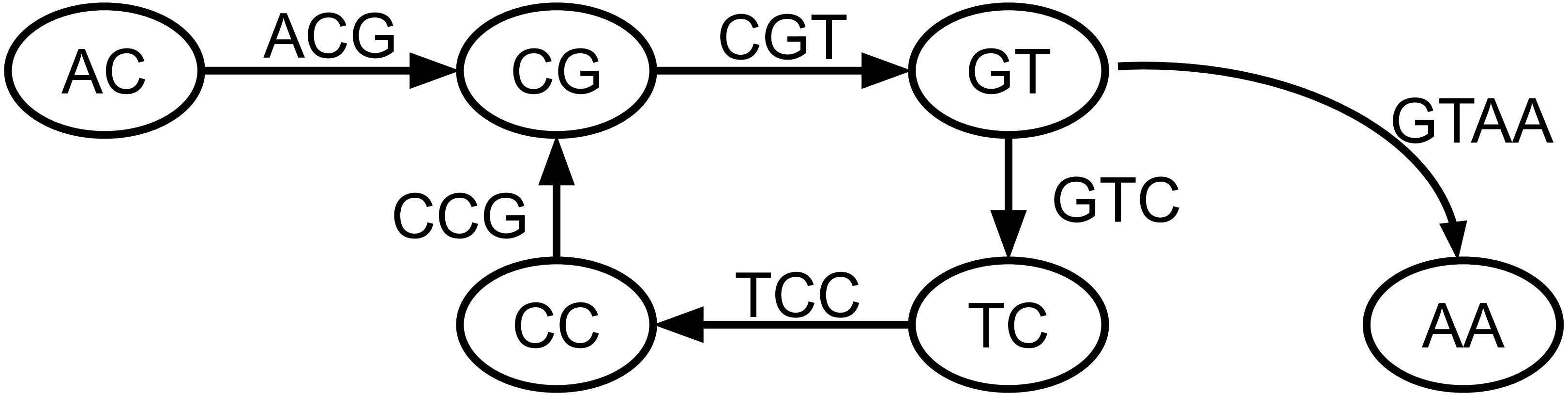
k=2



Condensed de Bruijn graph

ACGTCCGTAA

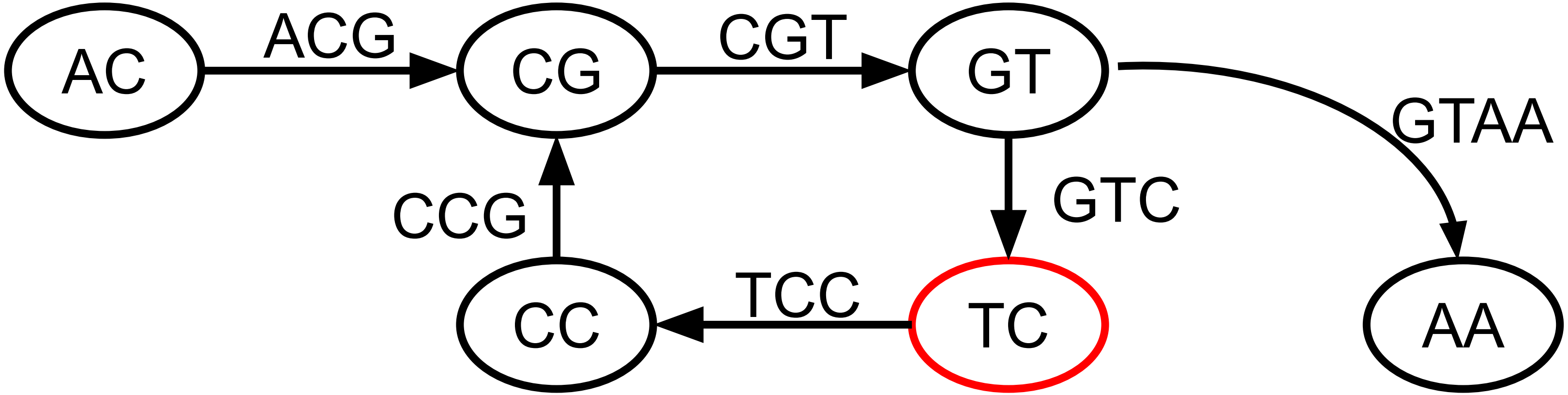
k=2



Condensed de Bruijn graph

ACGTCCGTAA

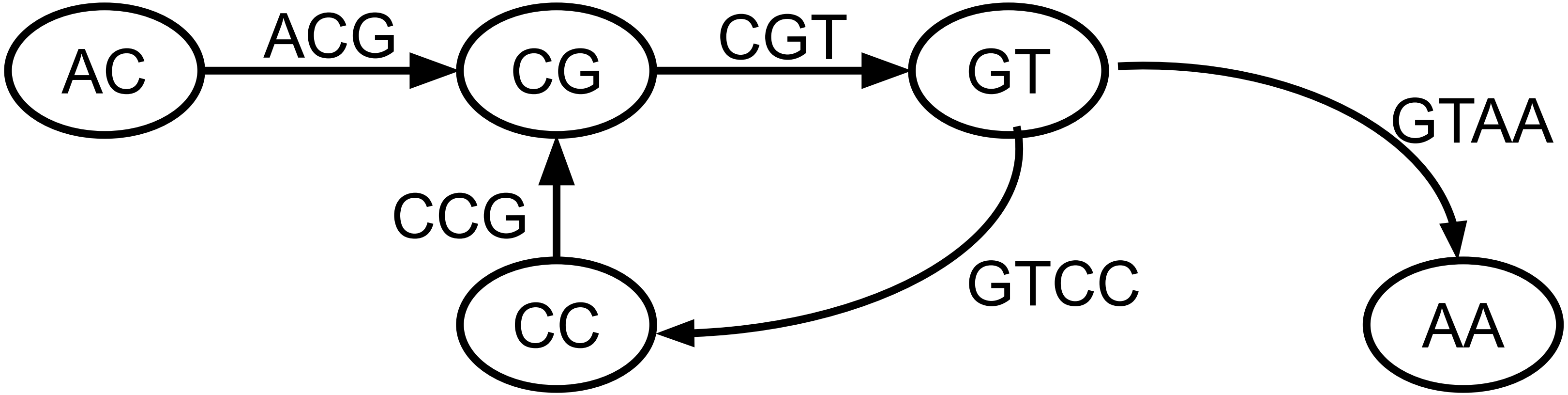
k=2



Condensed de Bruijn graph

ACGTCCGTAA

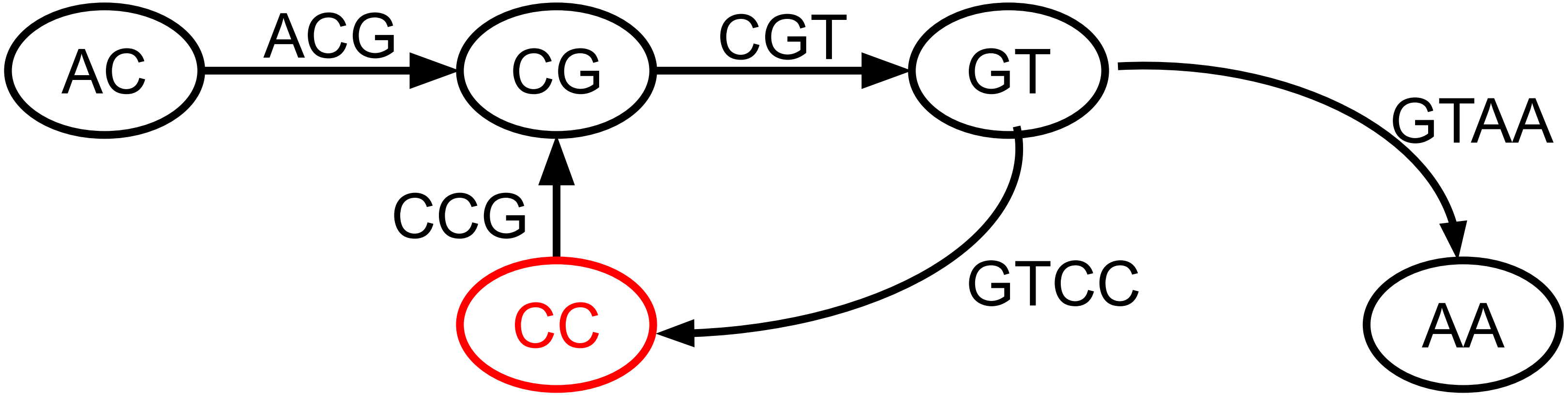
k=2



Condensed de Bruijn graph

ACGTCCGTAA

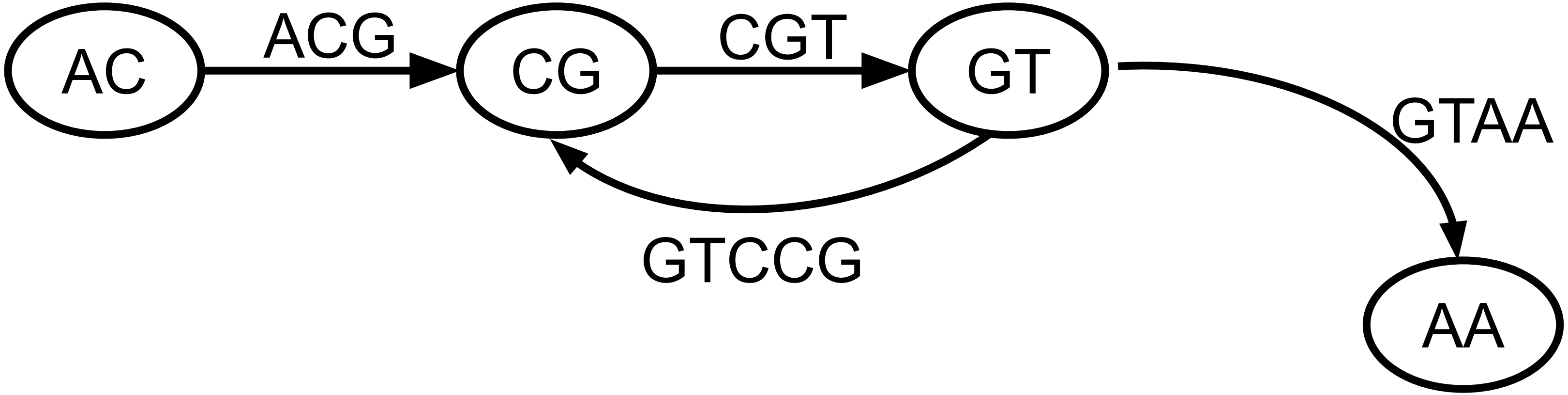
k=2



Condensed de Bruijn graph

ACGTCCGTAA

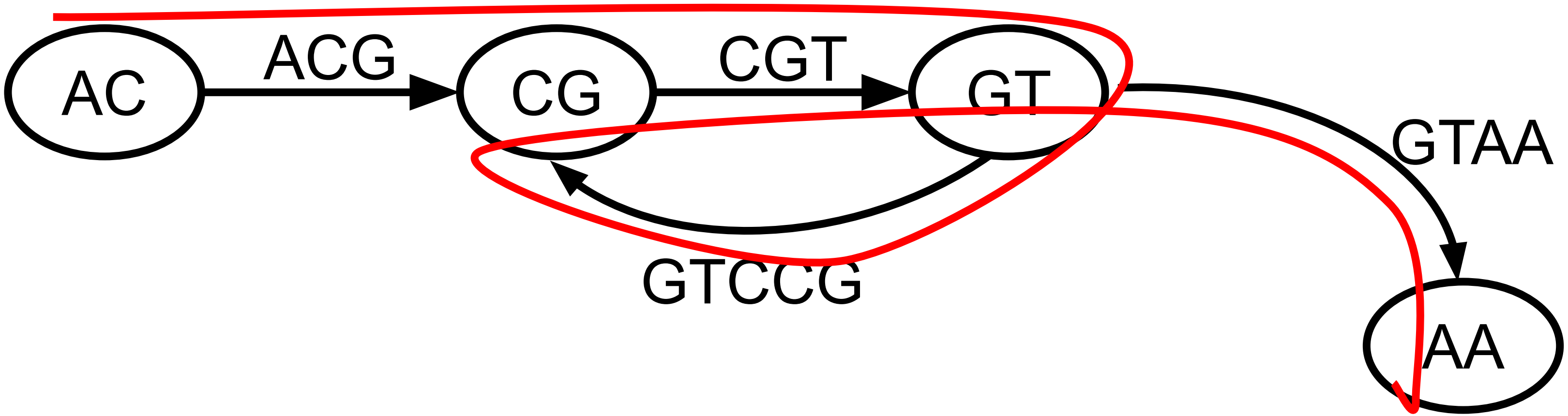
k=2



Condensed de Bruijn graph

ACGTCCGTAA

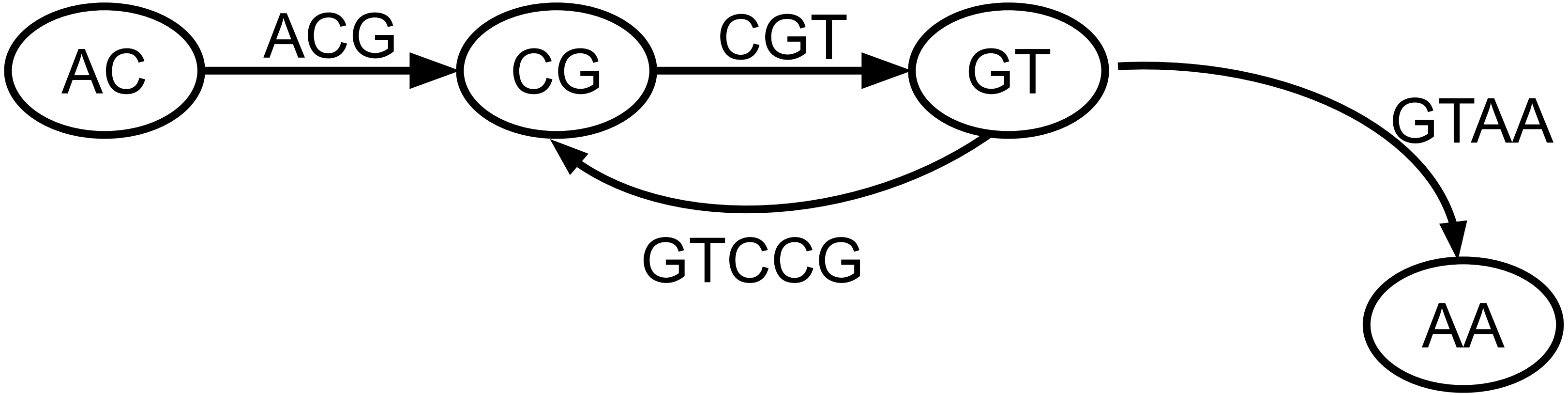
k=2



Repeats in de Bruijn graph

ACGTCCGTAA

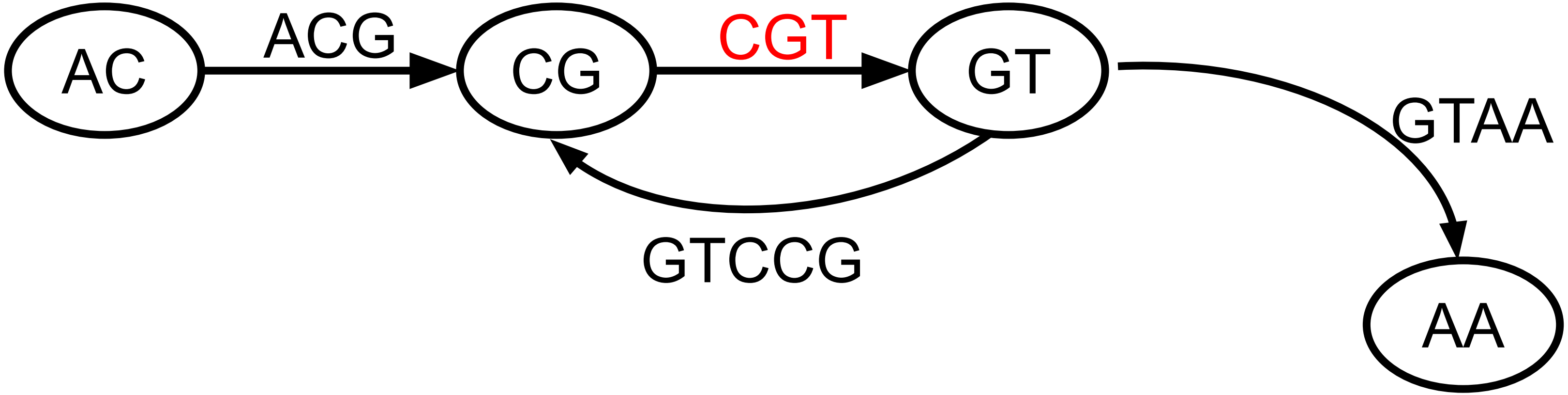
k=2



Repeats in de Bruijn graph

ACGTCCGTAA

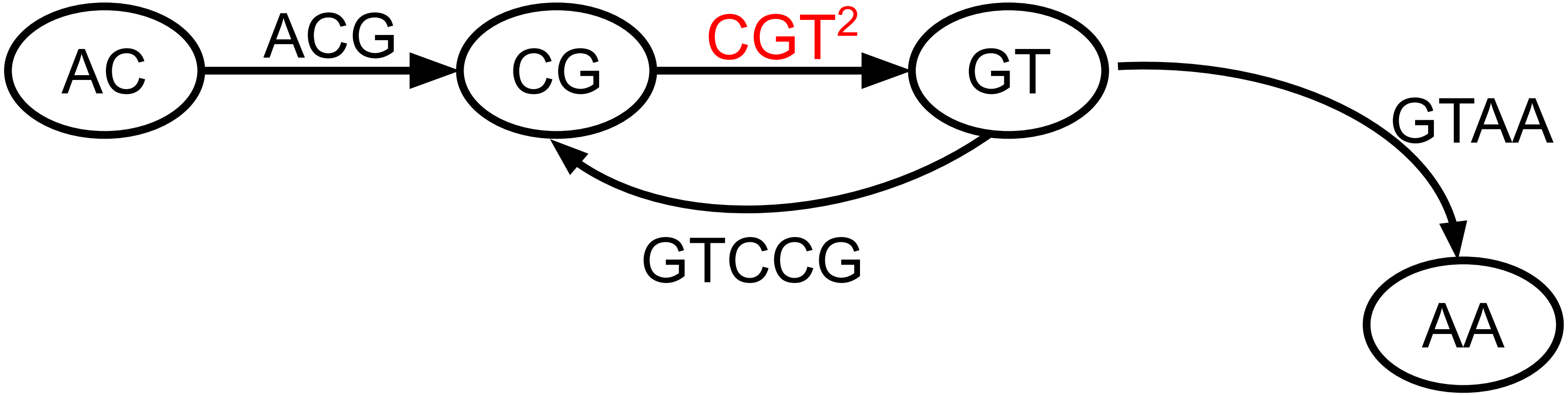
k=2



Repeats in de Bruijn graph

ACGTCCGTAA

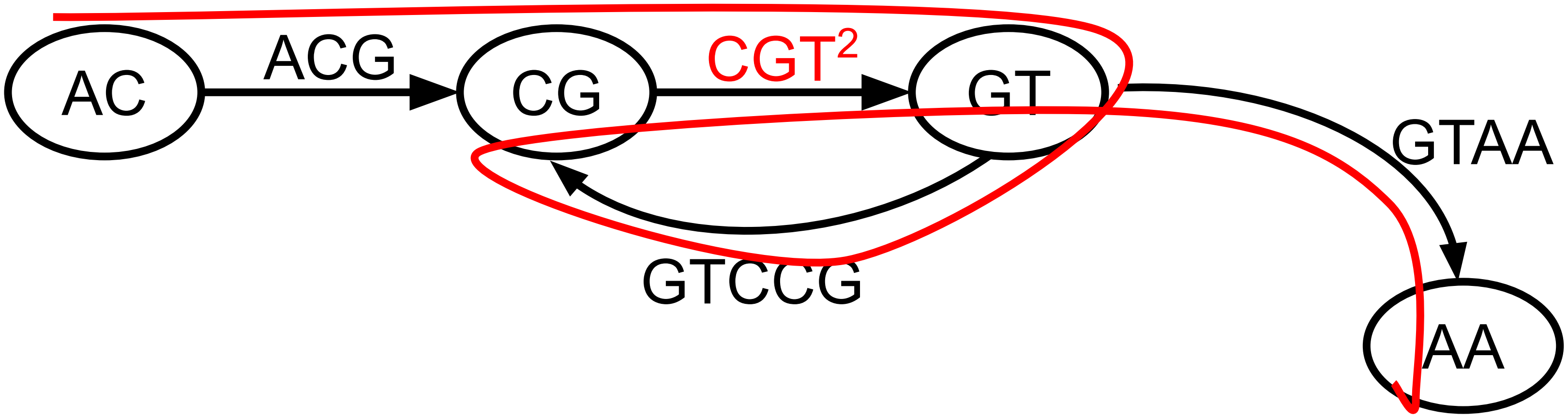
k=2



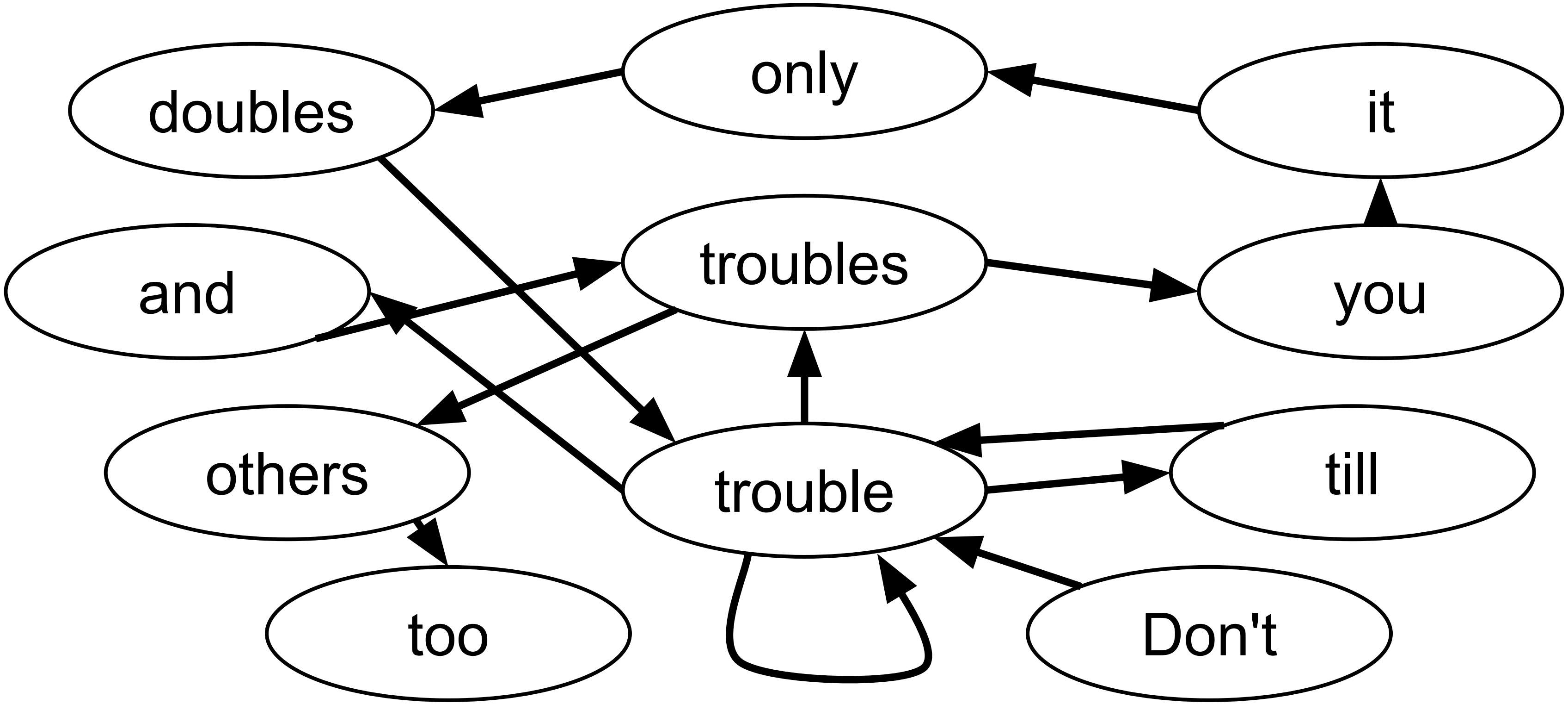
Eulerian path with multiplicities

ACGTCCGTAA

k=2



De Bruijn graph in a nutshell



Oh, repeats...

- Ribosomal operons (5-8 kbp)
- ALU, SINEs
    - < 1 kbp, extremely high multiplicity
- LINEs
    - ~ 6-7 kbp, high multiplicity
- Tandem repeats

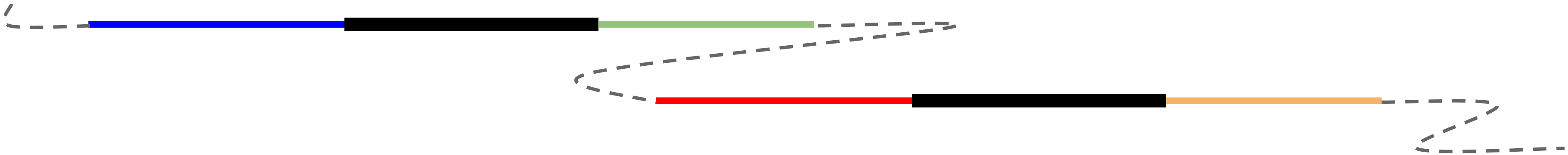
Oh, repeats...

NCBI contains assemblies with 100K+ scaffolds!

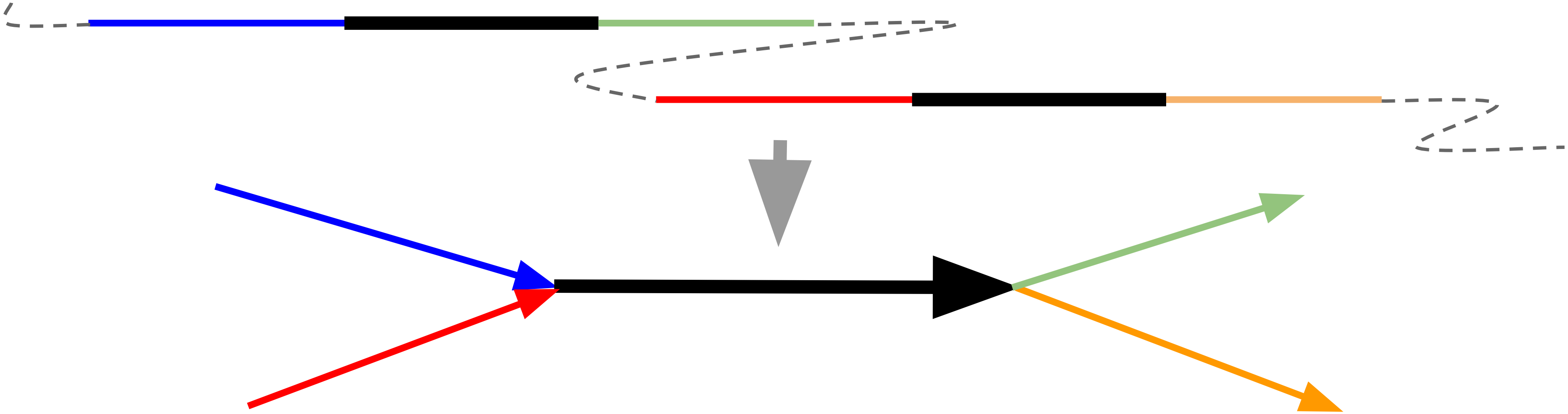
“These are not the genomes I wanted you to assemble”

Gene Meyers

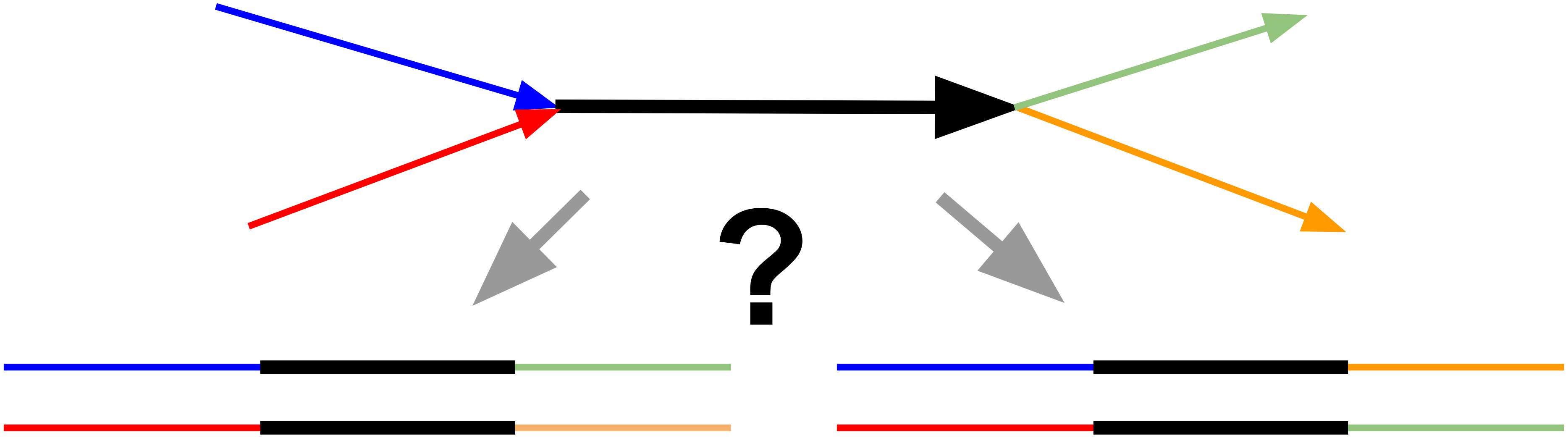
Resolving repeats



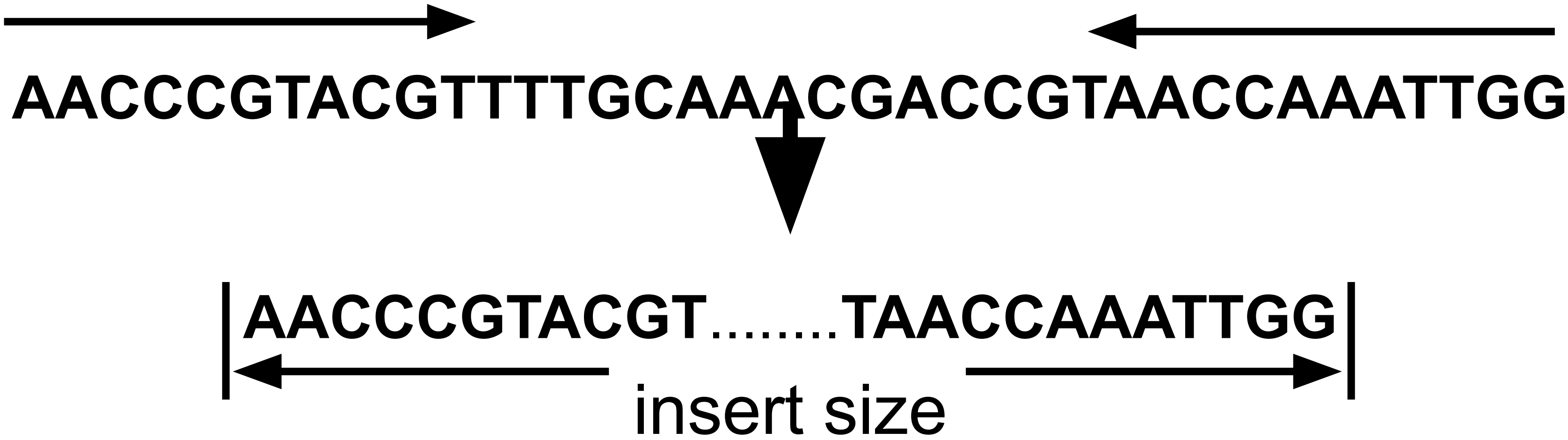
Resolving repeats



Resolving repeats



Paired reads

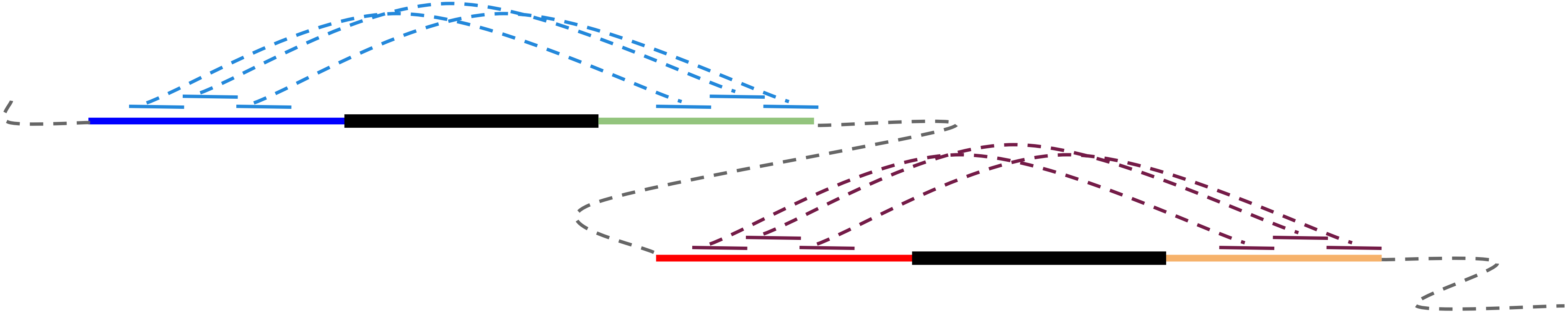


AACCCGTACGTTTTGCAAACGACCGTAACCAAATTGG

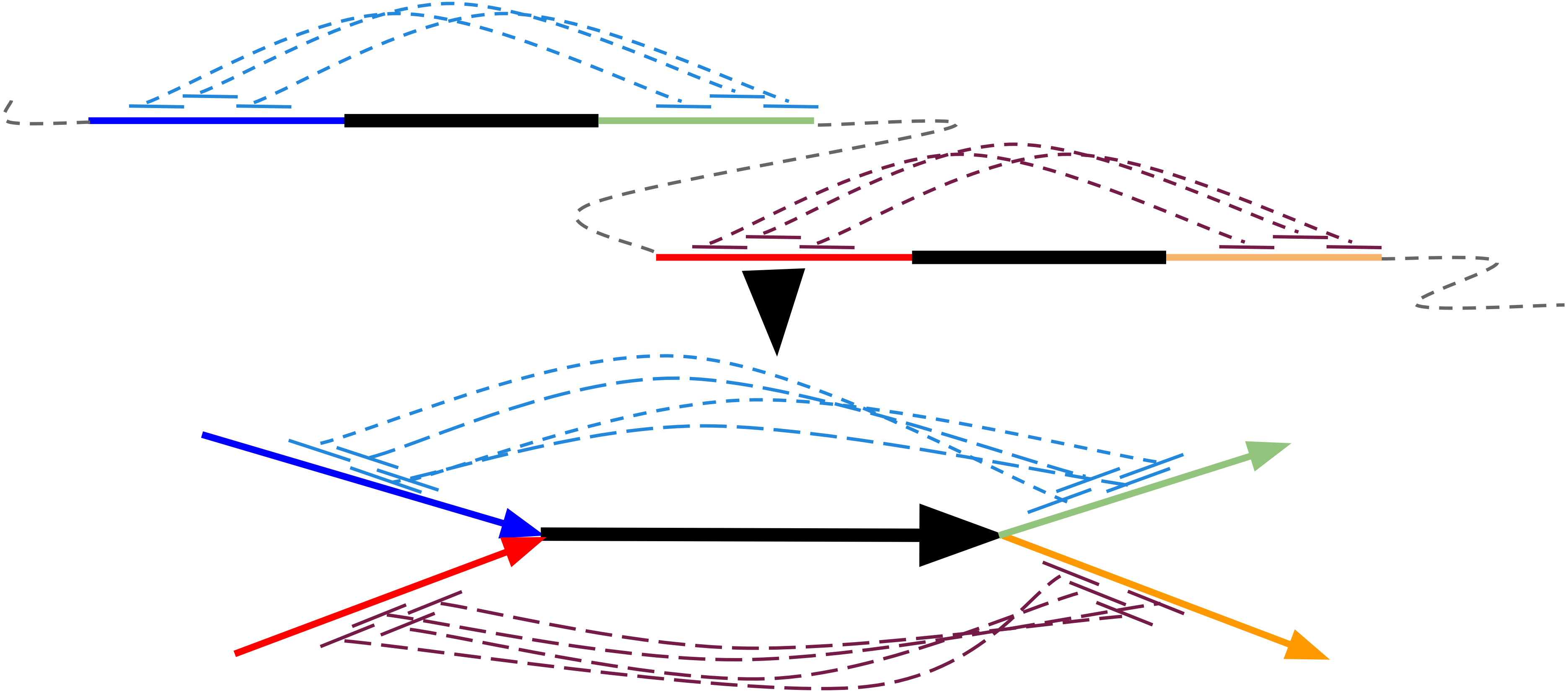
AACCCGTACGT........TAACCAAATTGG

insert size

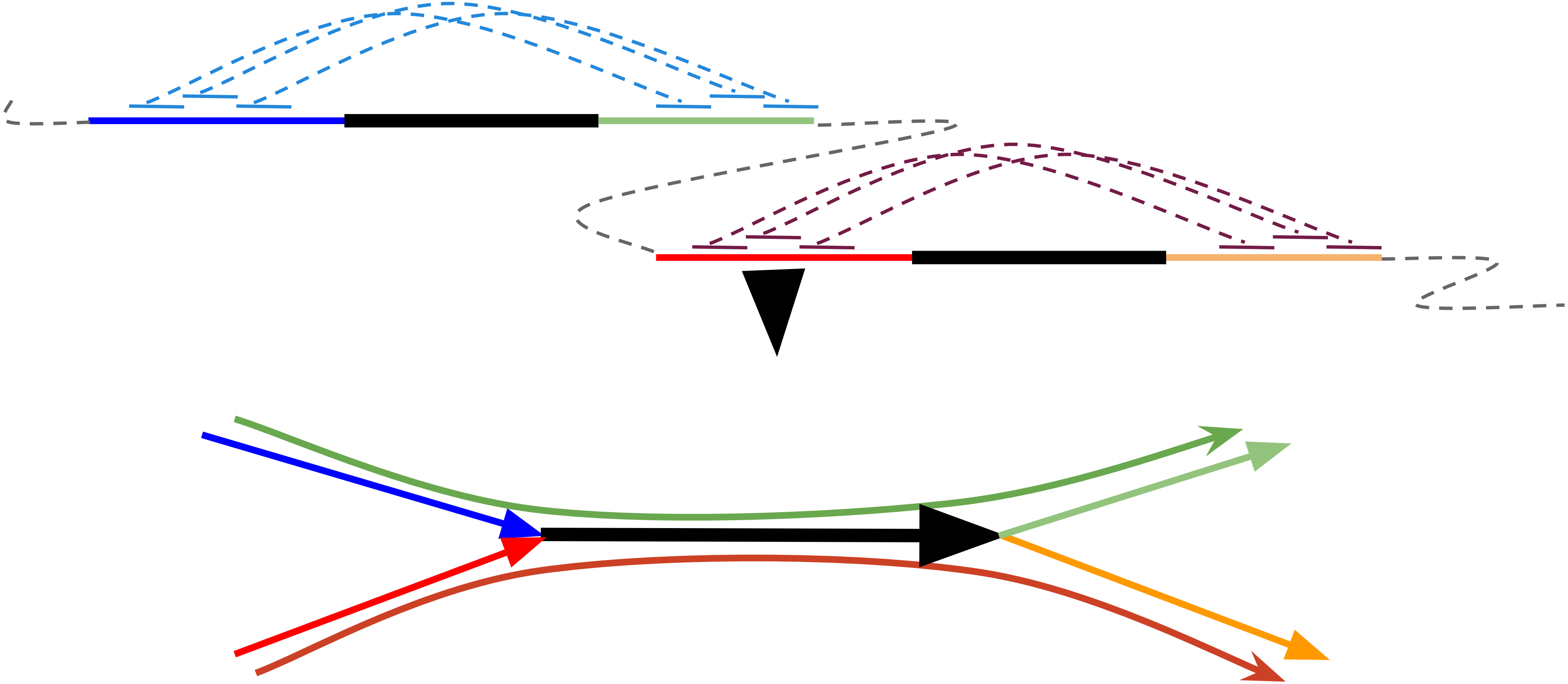
Resolving repeats



Resolving repeats

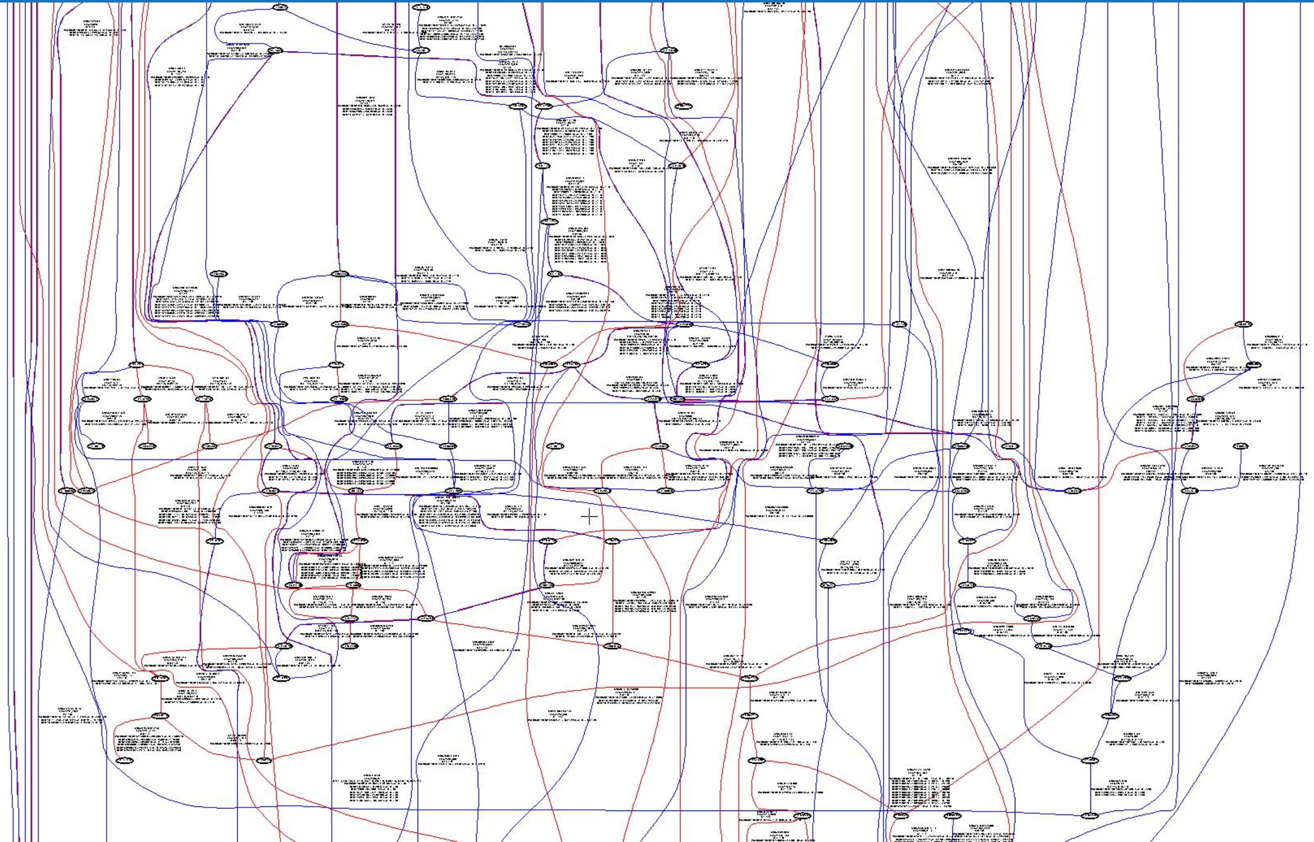


Resolving repeats



Real life

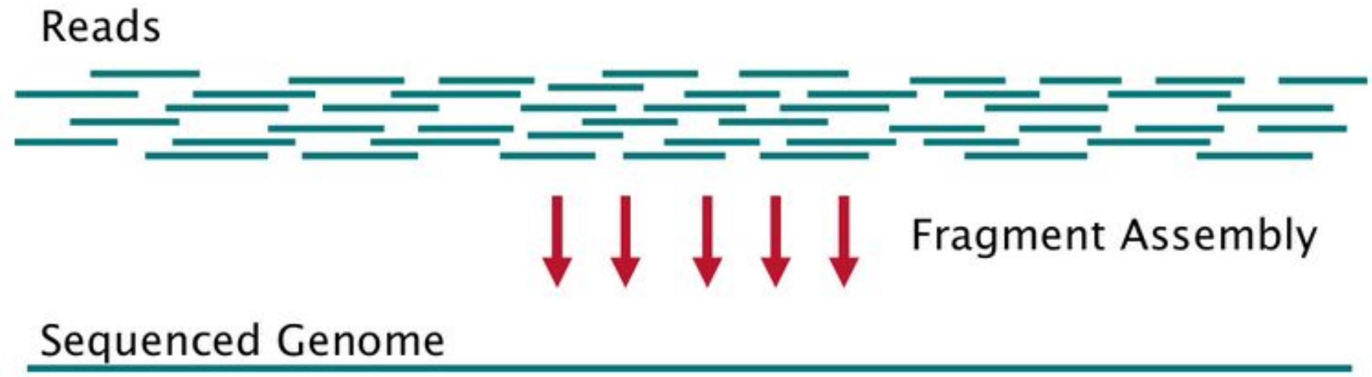
Part of *E.coli* genome, K = 99



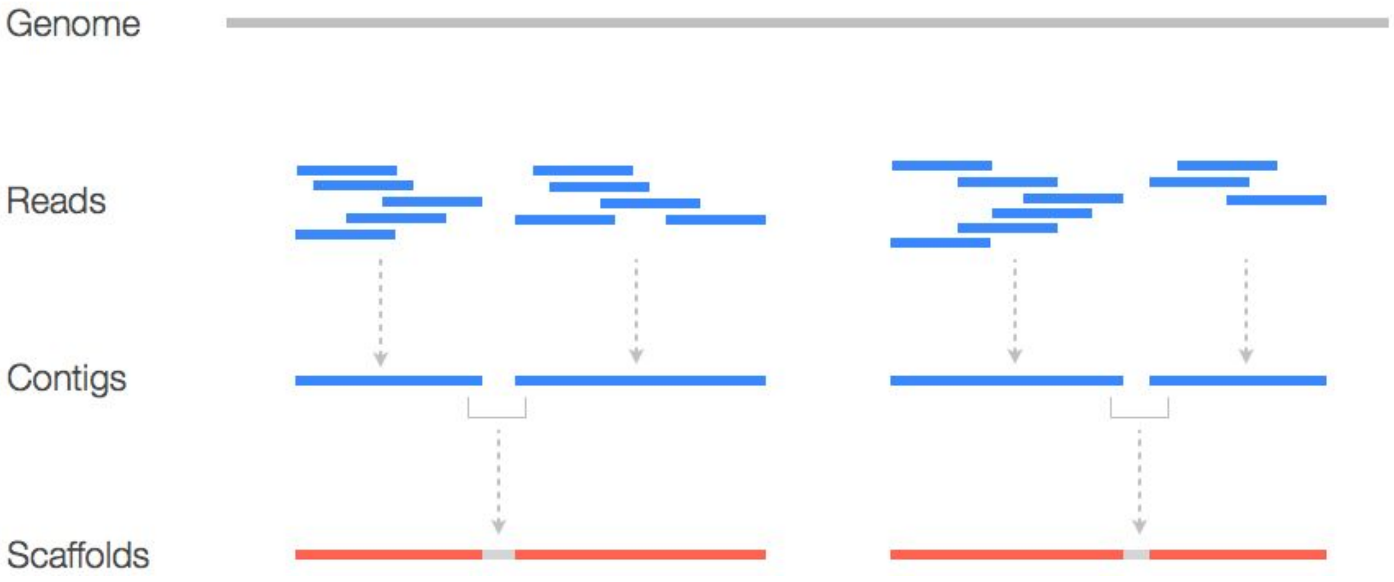
Genome assembly evaluation with QUAST

Center for Algorithmic Biotechnology
SPbU

Assembly in a perfect world



Assembly in real world



Which assembler to use?

- ABySS
- ALLPATHS-LG
- CLC
- IDBA-UD
- MaSuRCA
- MIRA
- Ray
- SOAPdenovo
- SPAdes
- Velvet
- and many more...

Which assembler to use?

- Different technologies (Illumina, 454, IonTorrent, ...)
- Genome type and size (bacteria, insects, mammals, plants, ...)
- Type of prepared libraries (single reads, paired-end, mate-pairs, combinations)
- Type of data (multicell, metagenomic, single-cell)

There is no best assembler

Which assembler to use?

- Assemblathon 1 & 2
  - Simulated and real datasets
  - More than 30 teams competing
- Independent studies
  - Papers (GAGE, GAGE-B, GABenchToB)
  - Web-sites (nucleotid.es, …)
  - Surveys
- Genome assembly evaluation tools
  - QUAST
  - GAGE

Genome Assembly Gold-Standard Evaluations

Assembly evaluation

- Basic evaluation
  - No extra input
  - Very quick
- Reference-based evaluation
  - A lot of metrics
  - Very accurate
- *De novo* evaluation
  - Advanced analysis of *de novo* assemblies

Basic statistics

- Only assemblies are needed (no additional input)
- Very fast to compute

Contig sizes

- Number of contigs

Contig sizes

- Number of contigs
- Number of large contigs (i.e. > 1000 bp)

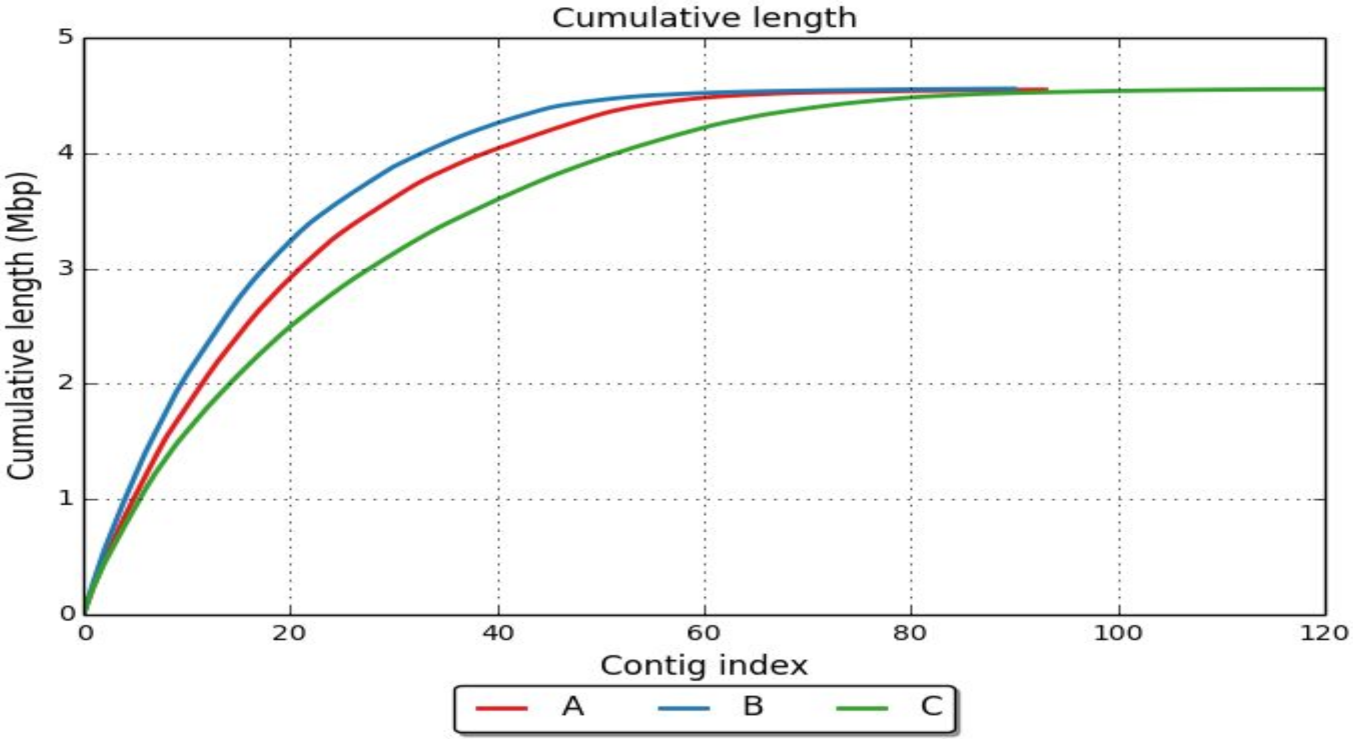
Contig sizes

- Number of contigs
- Number of large contigs (i.e. > 1000 bp)
- Largest contig length

Contig sizes

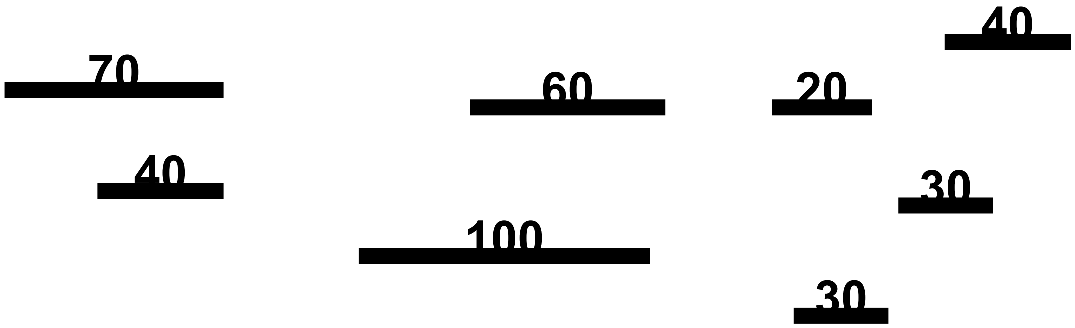
- Number of contigs
- Number of large contigs (i.e. > 1000 bp)
- Largest contig length
- Total assembly length

Cumulative length plot



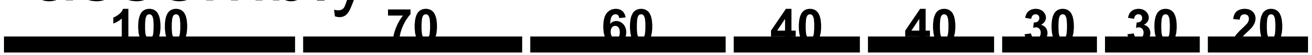
N50

The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50%** of the assembly



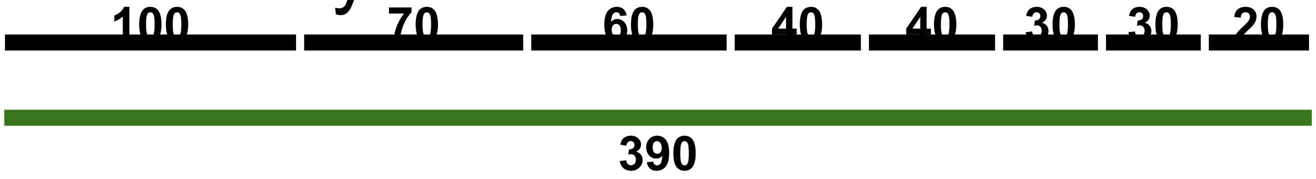
N50

The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50%** of the assembly



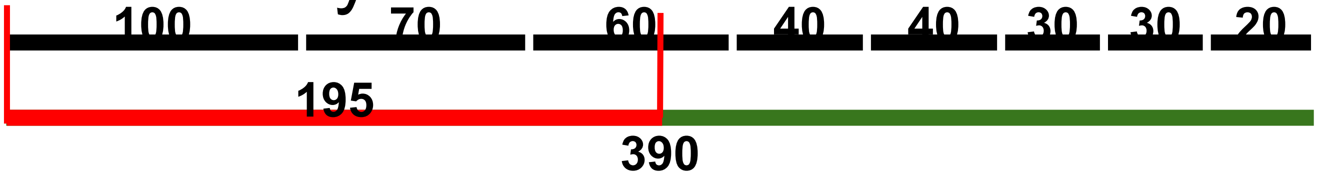
N50

The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50%** of the assembly



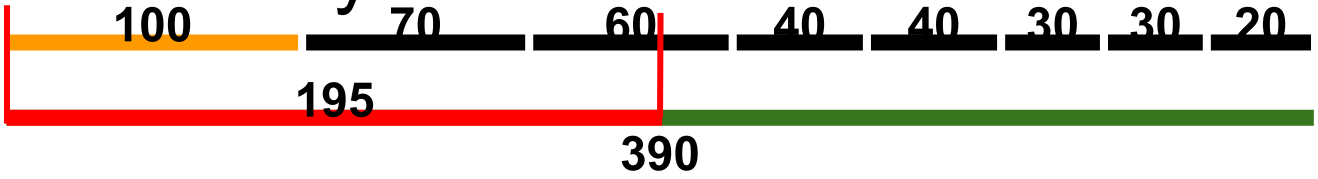
N50

The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50%** of the assembly



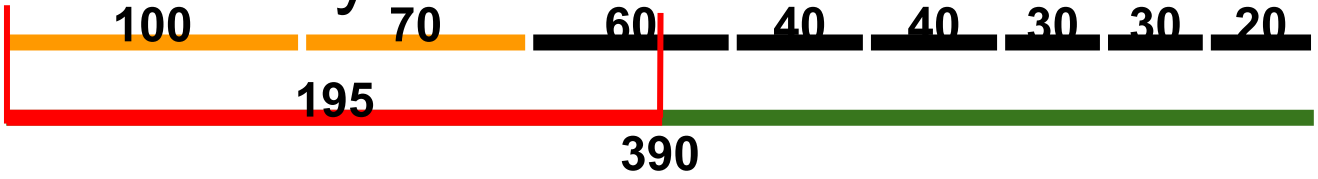
N50

The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50%** of the assembly



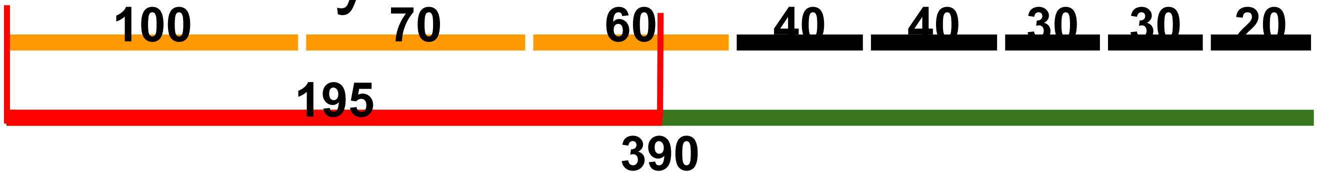
N50

The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50%** of the assembly



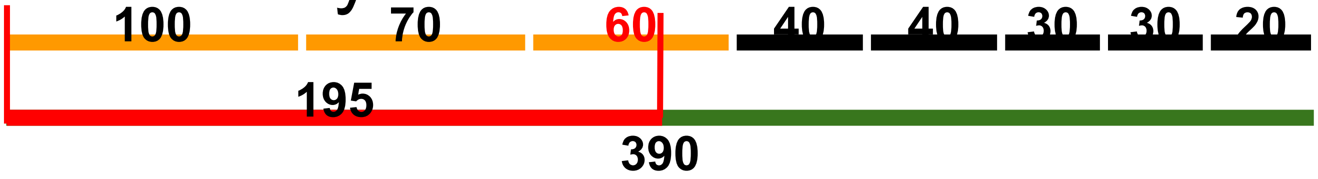
N50

The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50%** of the assembly



N50

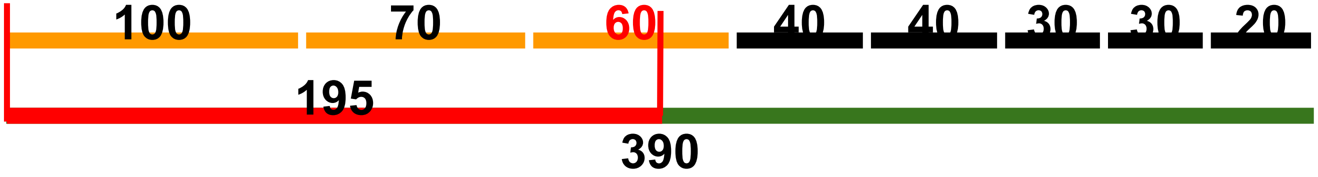
The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50%** of the assembly



N50 = 60

L50

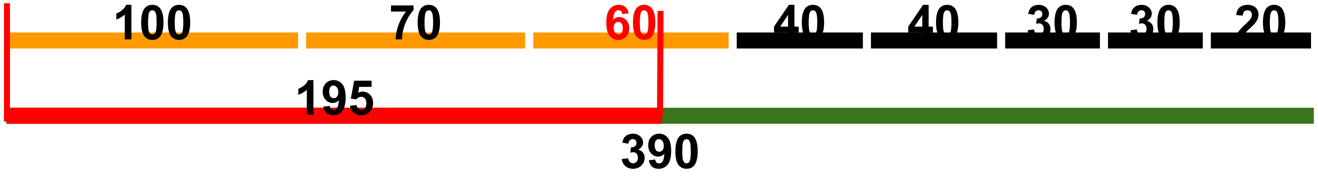
The minimum number **X** such that **X** longest contigs cover at least **50%** of the assembly



L50 =

L50

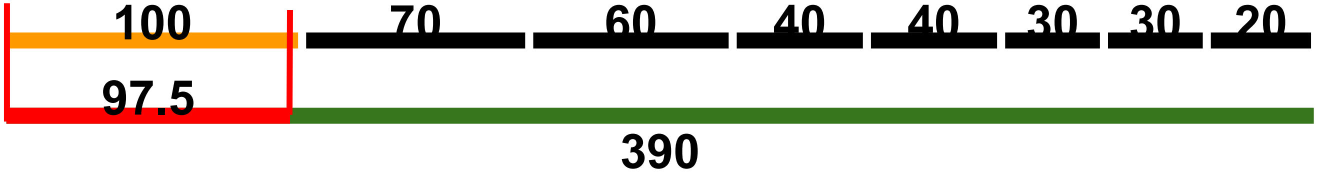
The minimum number **X** such that **X** longest contigs cover at least **50**% of the assembly



L50 = 3

N50-variations

- N25, N75
- L25, L75

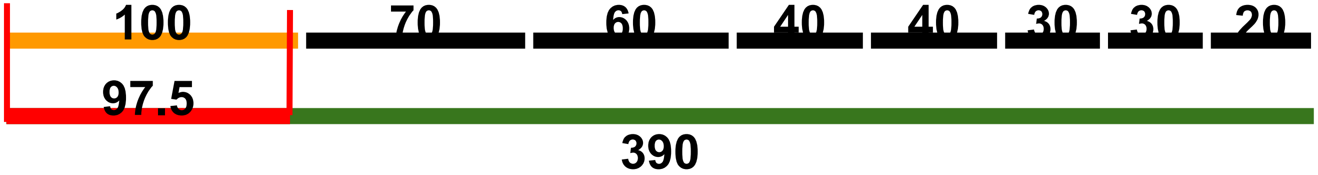


N25 = , N75 =

L25 = , L75 =

N50-variations

- N25, N75
- L25, L75



N25 = 100, N75 = 40

L25 = 1, L75 = 5

N50-variations

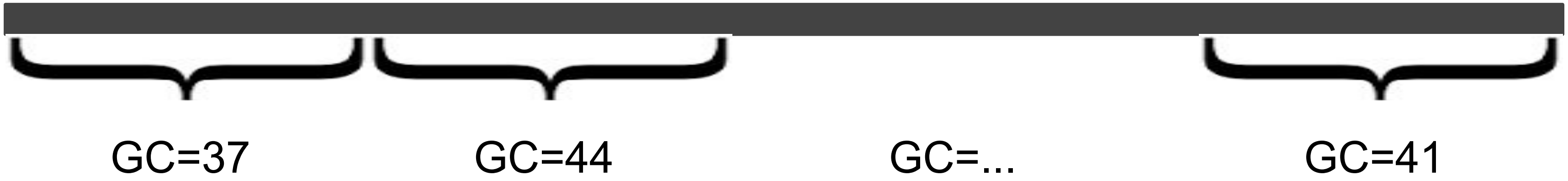
- N25, N75
- L25, L50, L75

N50-variations

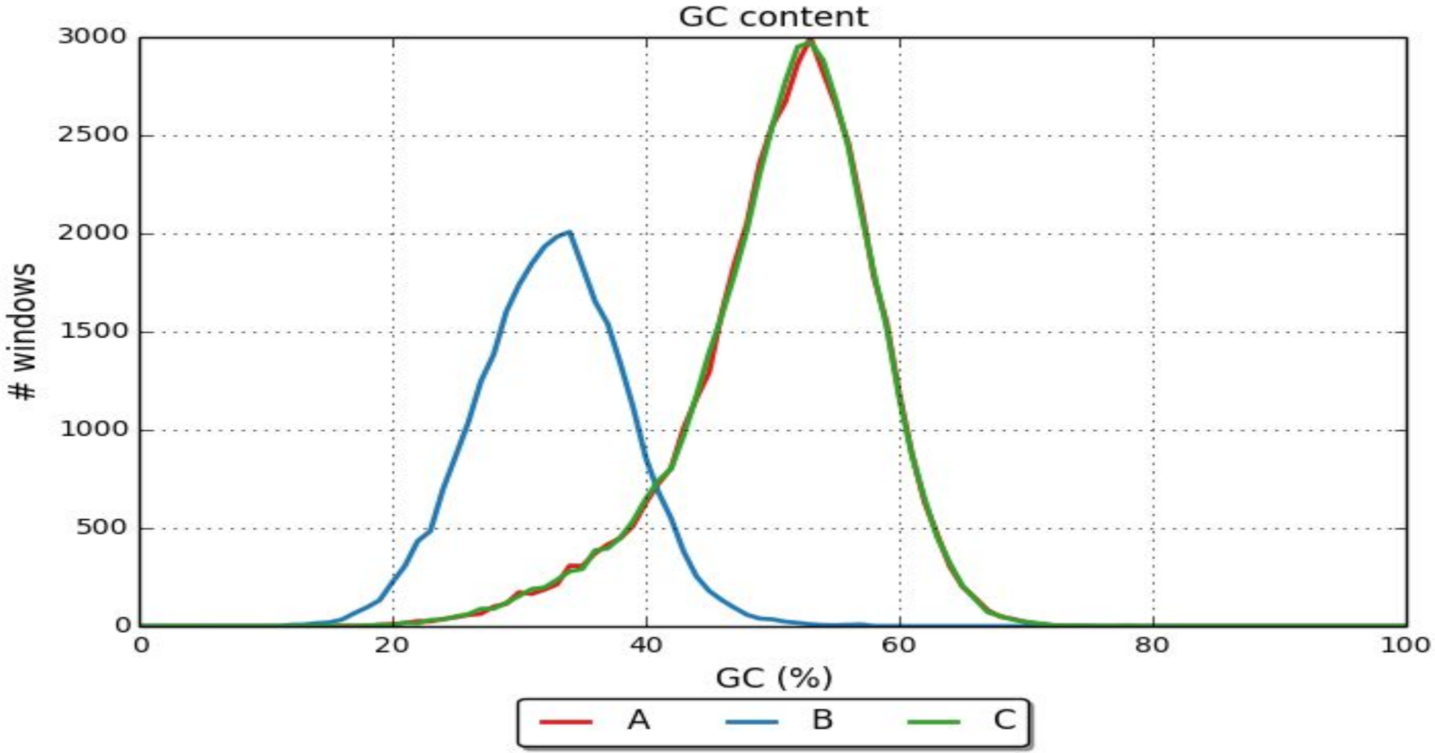
- N25, N75
- L25, L50, L75
- Nx, Lx

Other

- Number of N’s per 100 kbp
- GC %
- Distributions of GC % in small windows:



Other


GC content
3000
2500
2000
1500
1000
500
0
# windows
0
20
40
60
80
100
GC (%)
A
B
C

Reference-based metrics

- A lot of metrics
- Accurate assessment

Basic reference statistics

- Reference length
- Reference GC %
- Number of chromosomes

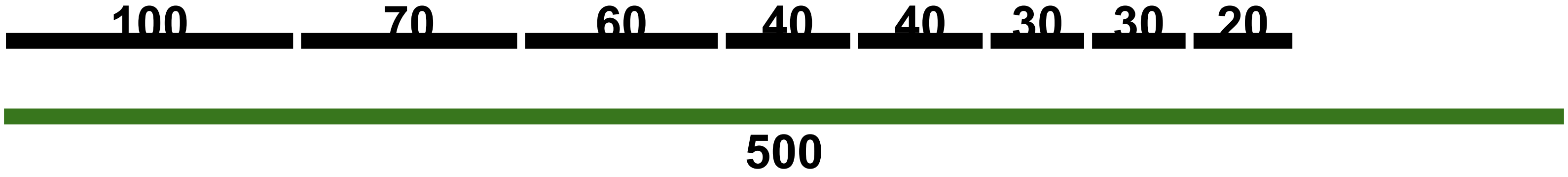
Basic reference statistics

- Reference length
- Reference GC %
- Number of chromosomes

- NGx, LGx

Basic reference statistics

- NGx, LGx

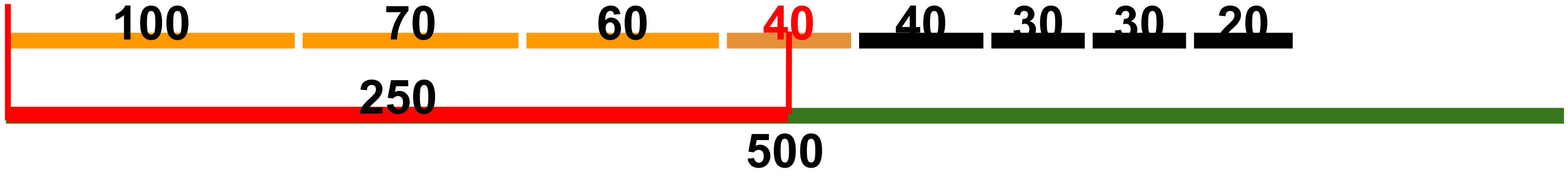


NG50 =

LG50 =

Basic reference statistics

- NGx, LGx

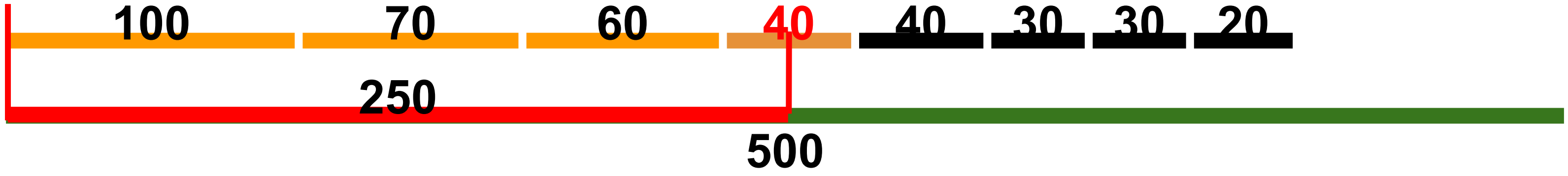


NG50 =

LG50 =

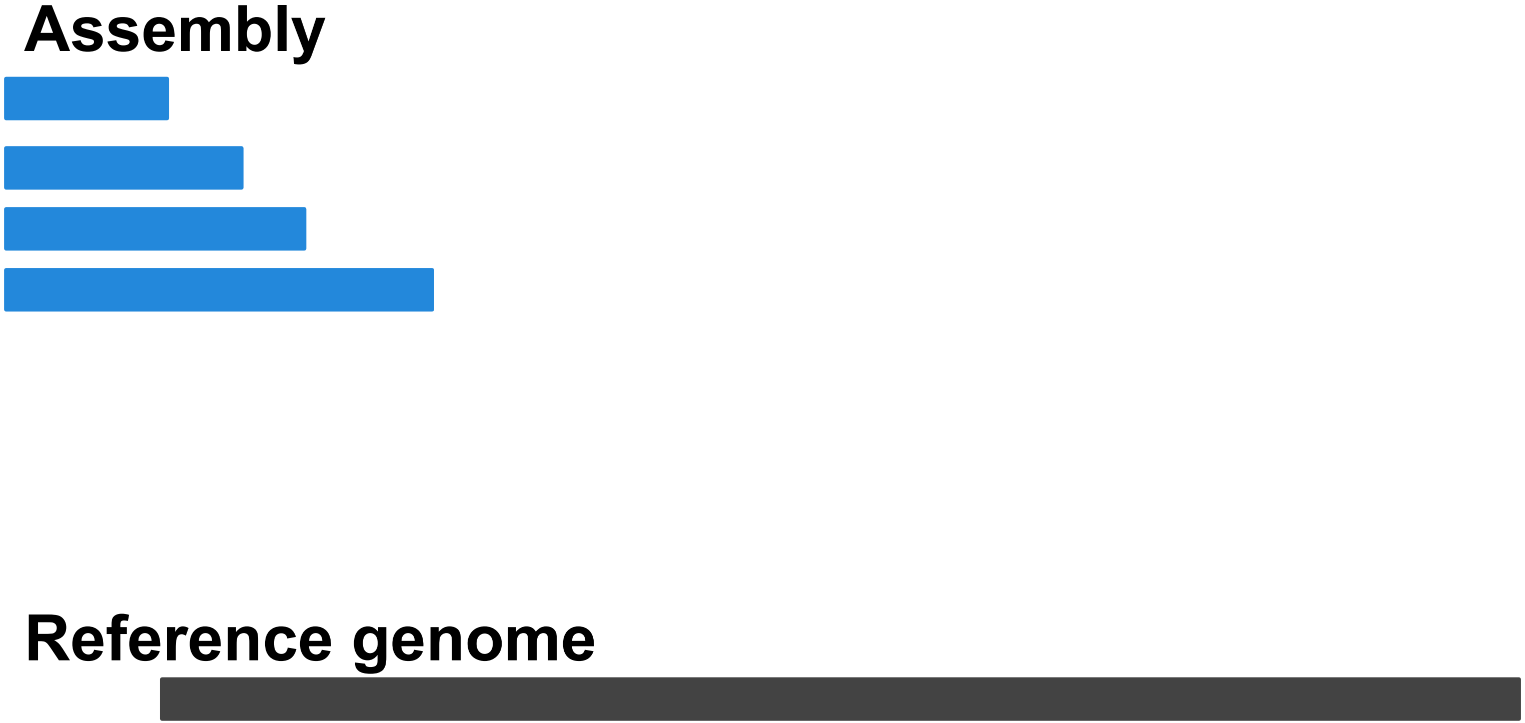
Basic reference statistics

- NGx, LGx



NG50 = 40
LG50 = 4

Alignment statistics

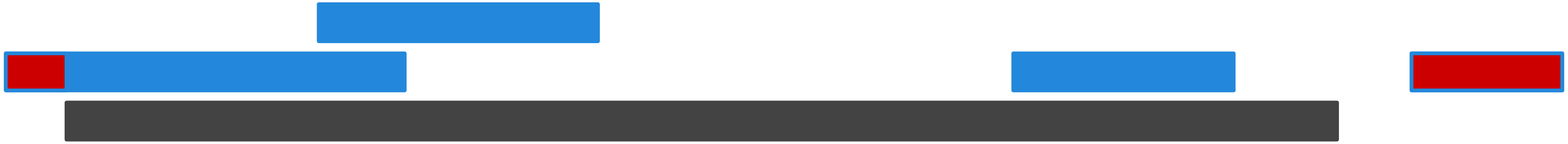


Assembly

Reference genome

Alignment statistics

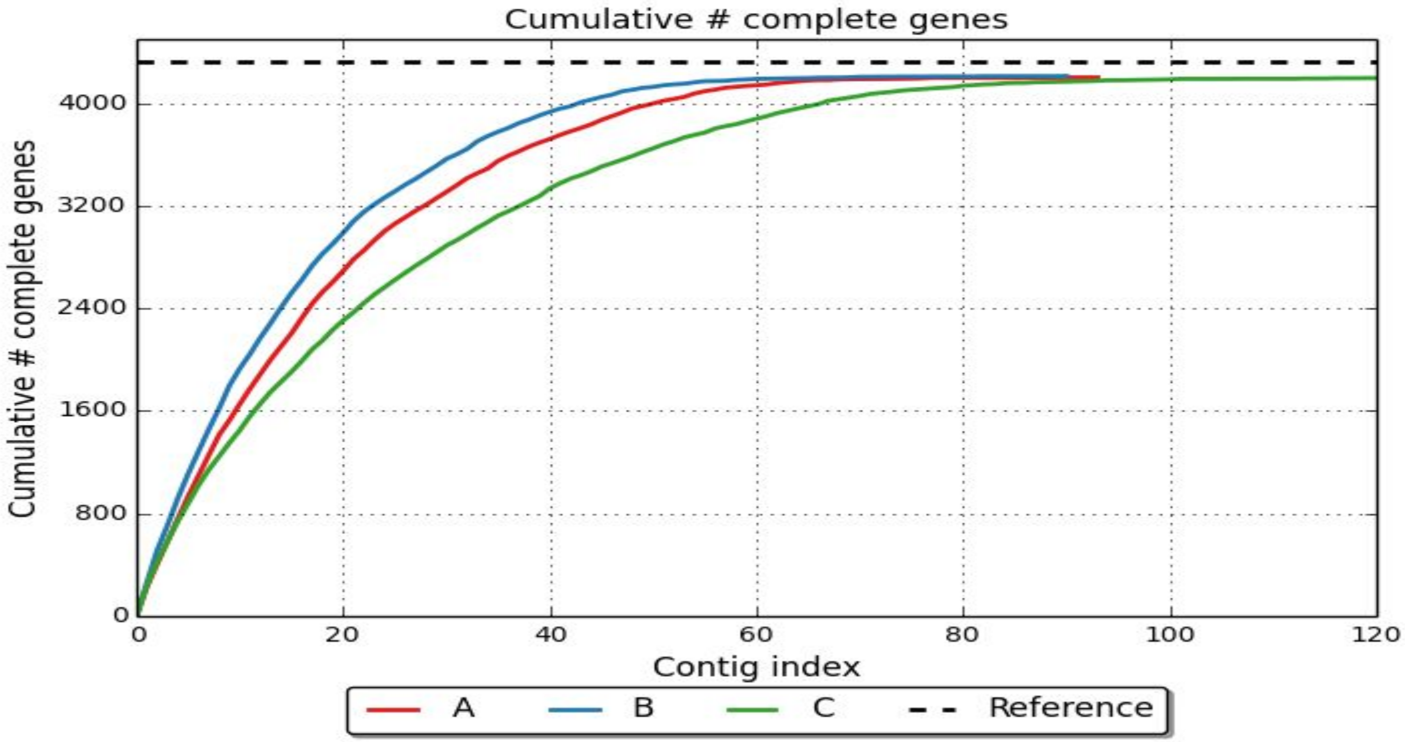
- Genome fraction %
- Duplication ratio
- Number of gaps
- Largest alignment length
- Number of unaligned contigs (full & partial)



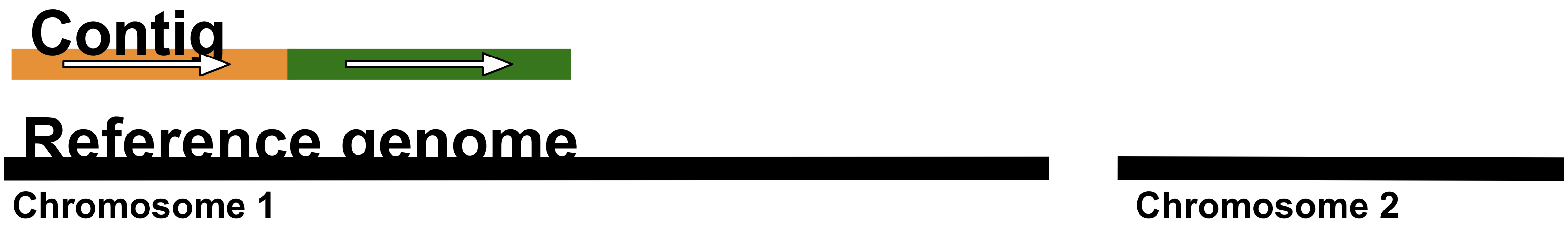
Alignment statistics

- Genome fraction %
- Duplication ratio
- Number of gaps
- Largest alignment length
- Number of unaligned contigs (full & partial)
- Number of mismatches/indels per 100 kbp
- Number of genes/operons (full & partial)

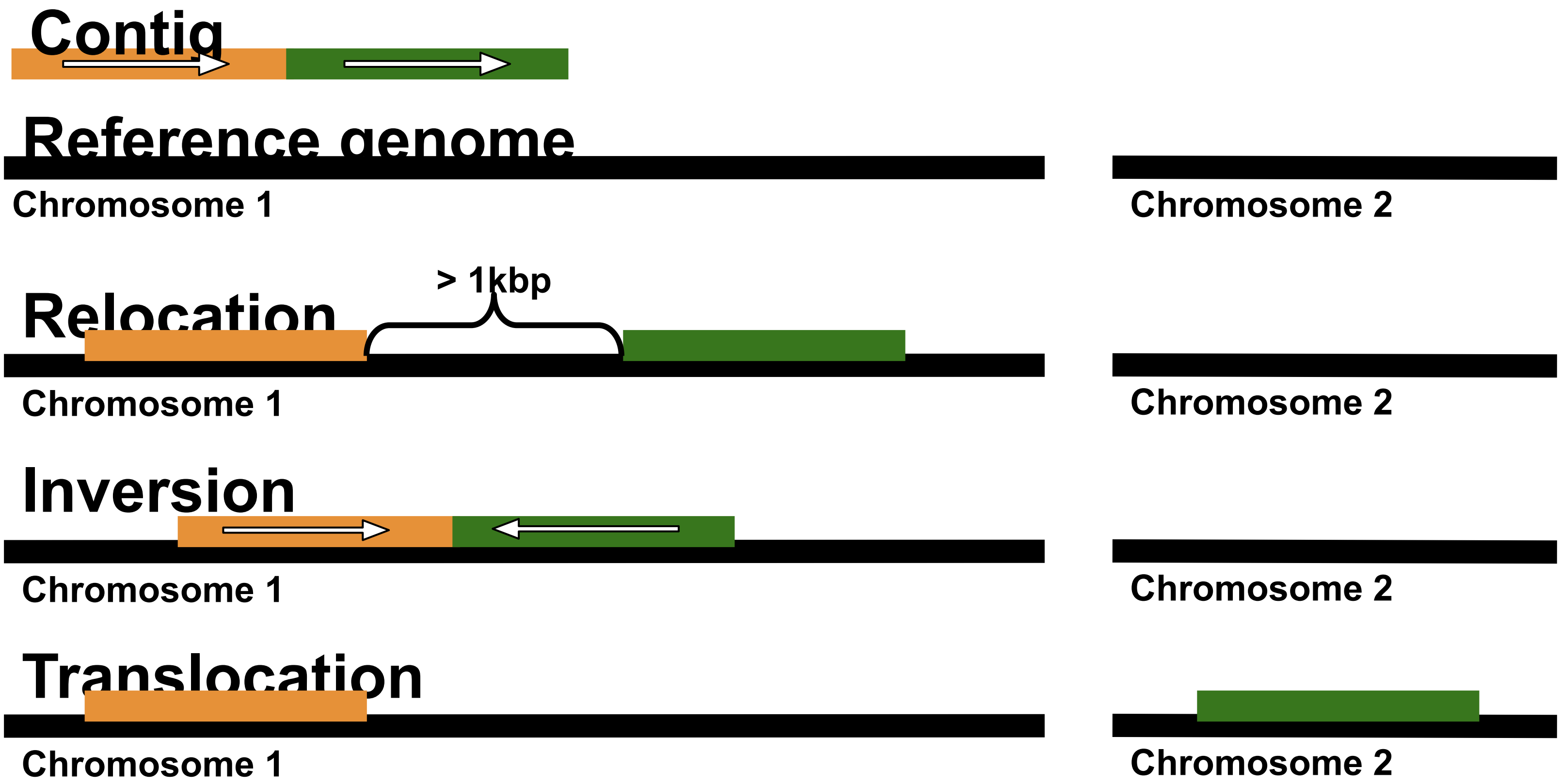
Alignment statistics



Misassemblies



Misassemblies



Contig

Reference genome

Chromosome 1

Chromosome 2

Relocation

> 1kbp

Chromosome 1

Chromosome 2

Inversion

Chromosome 1

Chromosome 2

Translocation

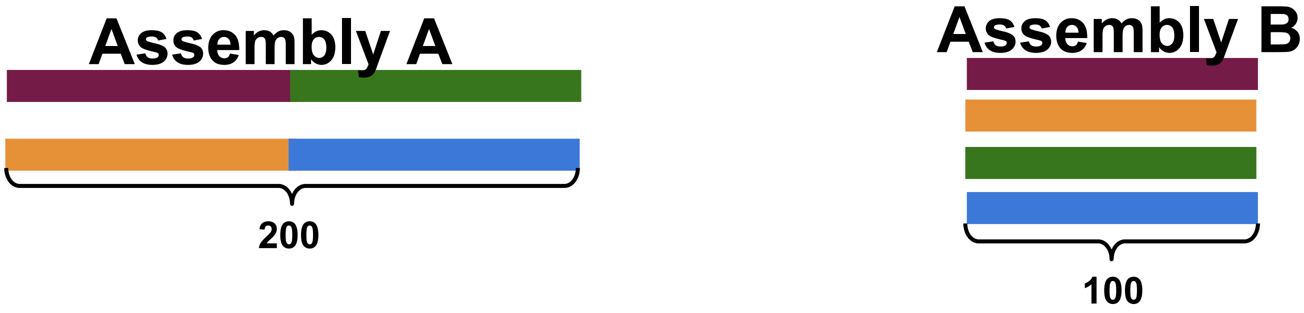
Chromosome 1

Chromosome 2

NB!

There is **no** best metric

NA50



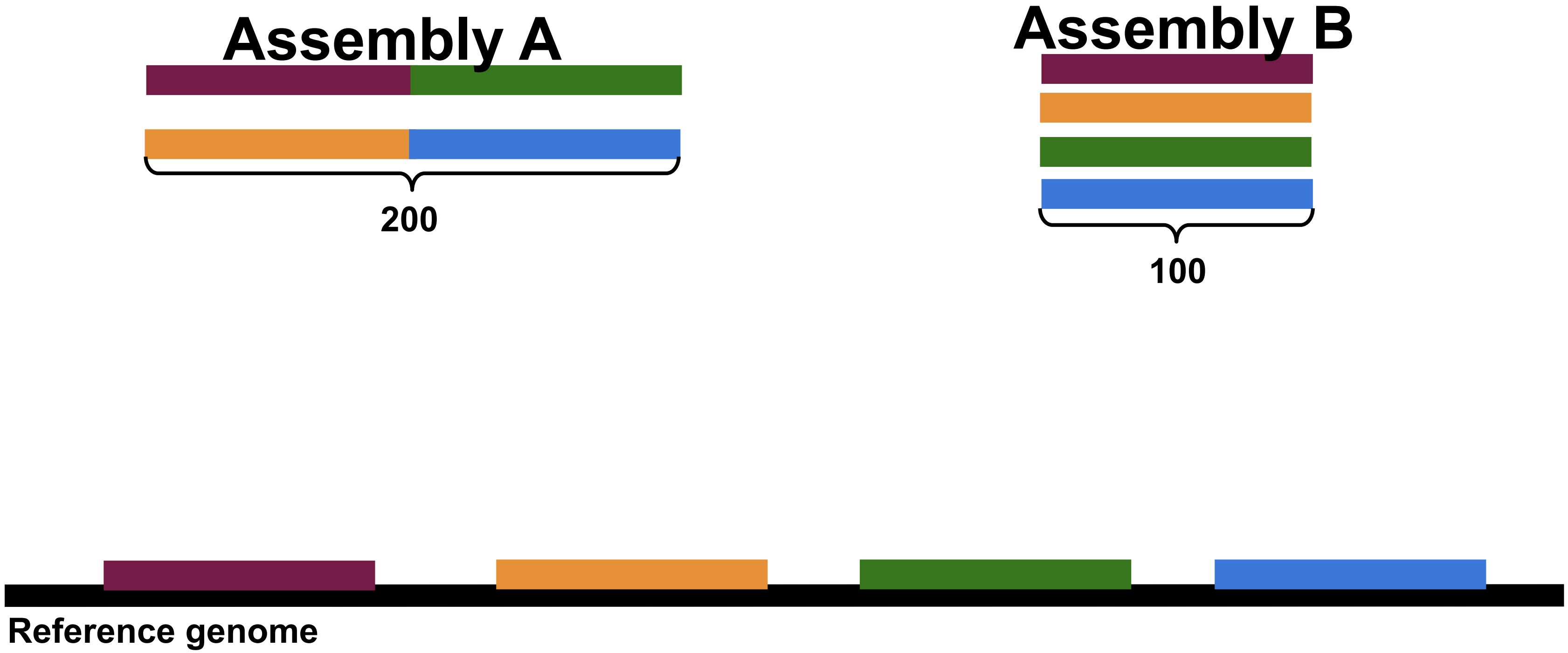
Assembly A

200

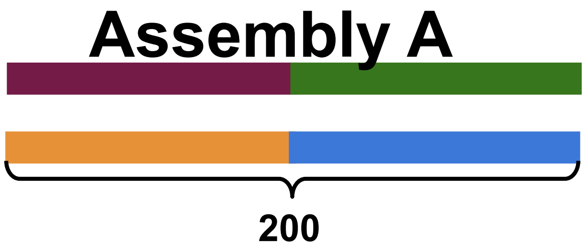
Assembly B

100

NA50



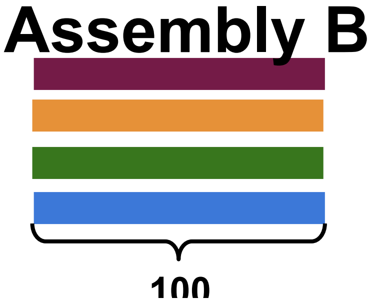
NA50



Assembly A

200

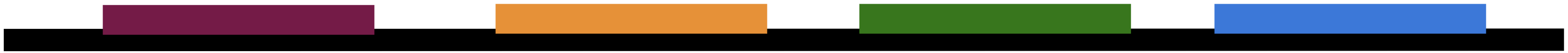
N50 = 200
# misassemblies = 2



Assembly B

100

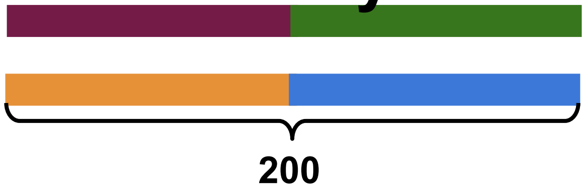
N50 = 100
# misassemblies = 0



Reference genome

NA50

Assembly A

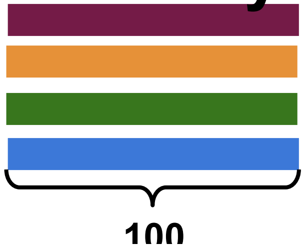


N50 = 200

# misassemblies = 2

NA50 = 100

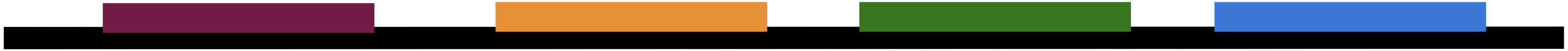
Assembly B



N50 = 100

# misassemblies = 0

NA50 = 100



Reference genome

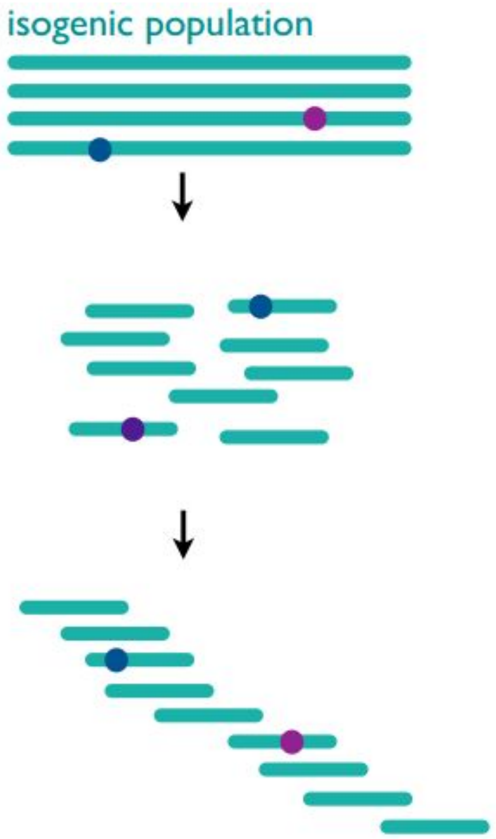
Error Correction

To correct the sequencing errors, we can

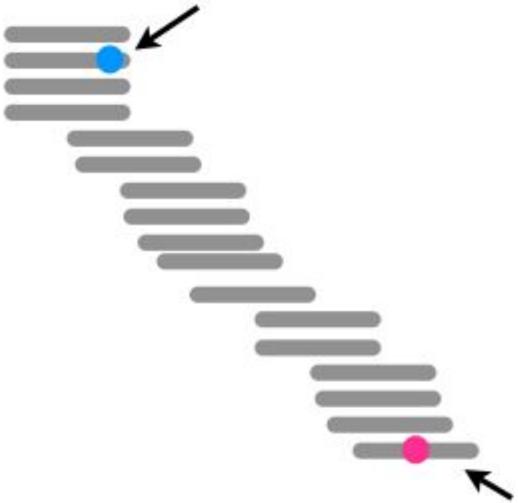
1) analyze k-mer distribution in sequencing data

2) use high abundance k-mers to correct data

/informatics - de novo assembly

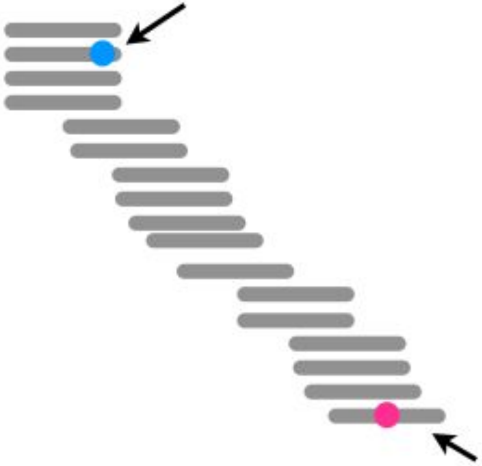


/informatics k-mer error correction



1) use k-mer distribution to identify suspicious (unique) k-mers

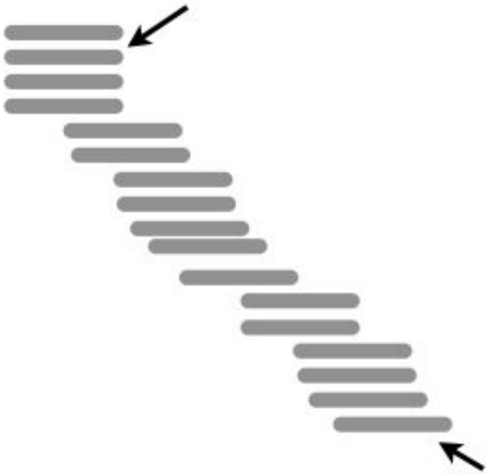
/informatics k-mer error correction



1) use k-mer distribution to identify suspicious (unique) k-mers

2) use good (high abundance) k-mers that are only 1 or 2 mutations away, and rewrite the suspicious k-mer

/informatics k-mer error correction



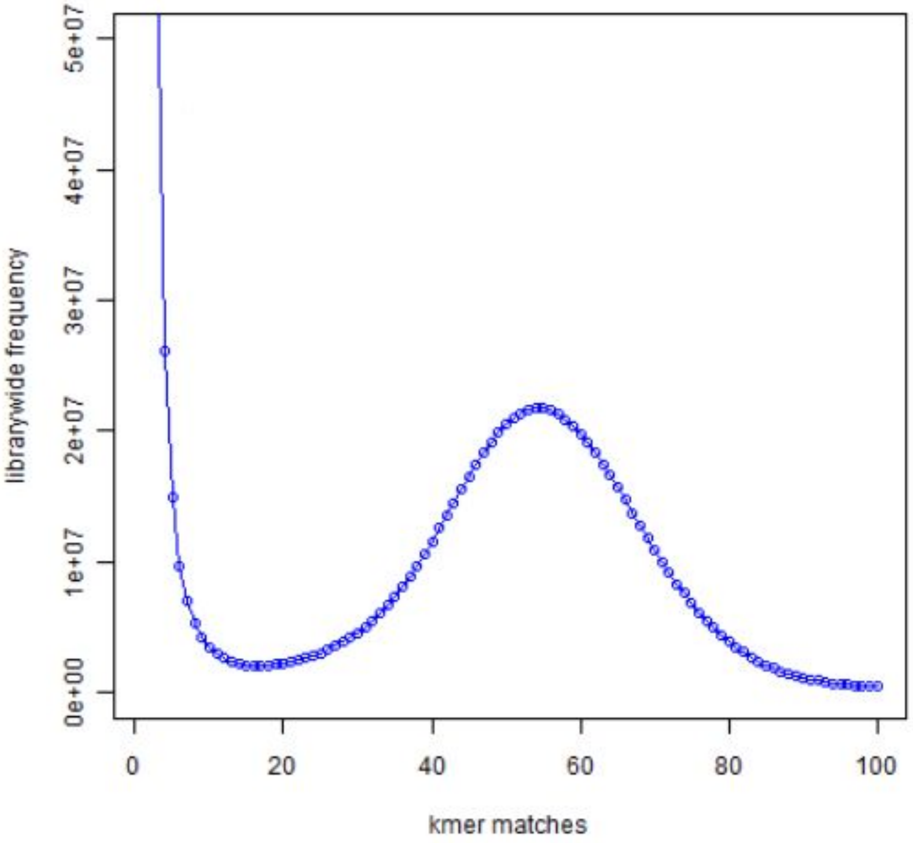
1) use k-mer distribution to identify suspicious (unique) k-mers

2) use good (high abundance) k-mers that are only 1 or 2 mutations away, and rewrite the suspicious k-mer

/informatics k-mer distribution

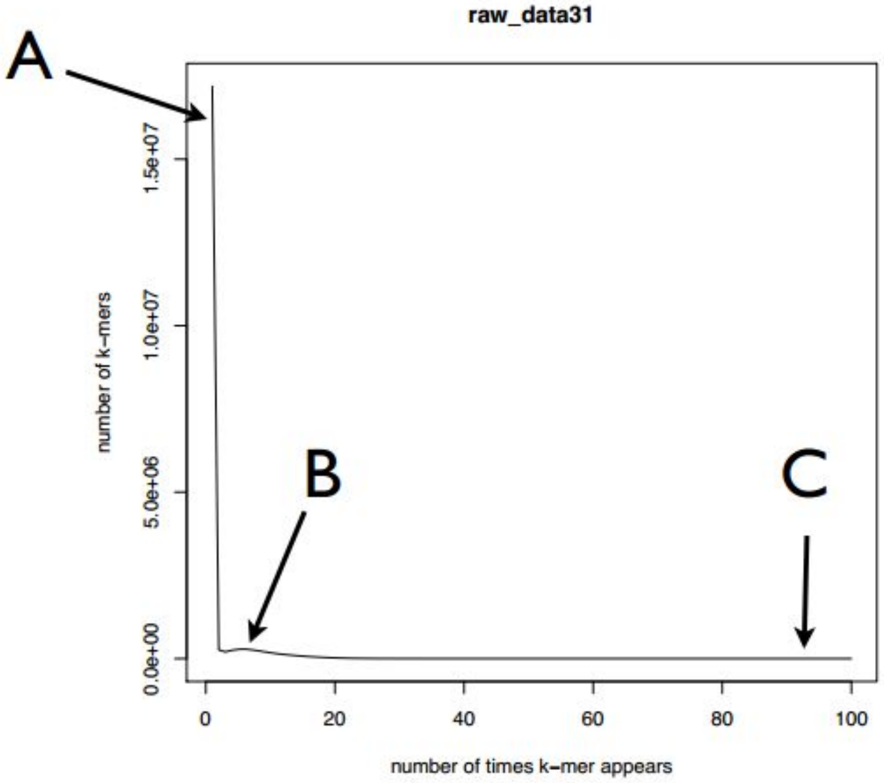
k-mer distribution for a given length k-mer in actual sequencing data

ie: k-mer = 30



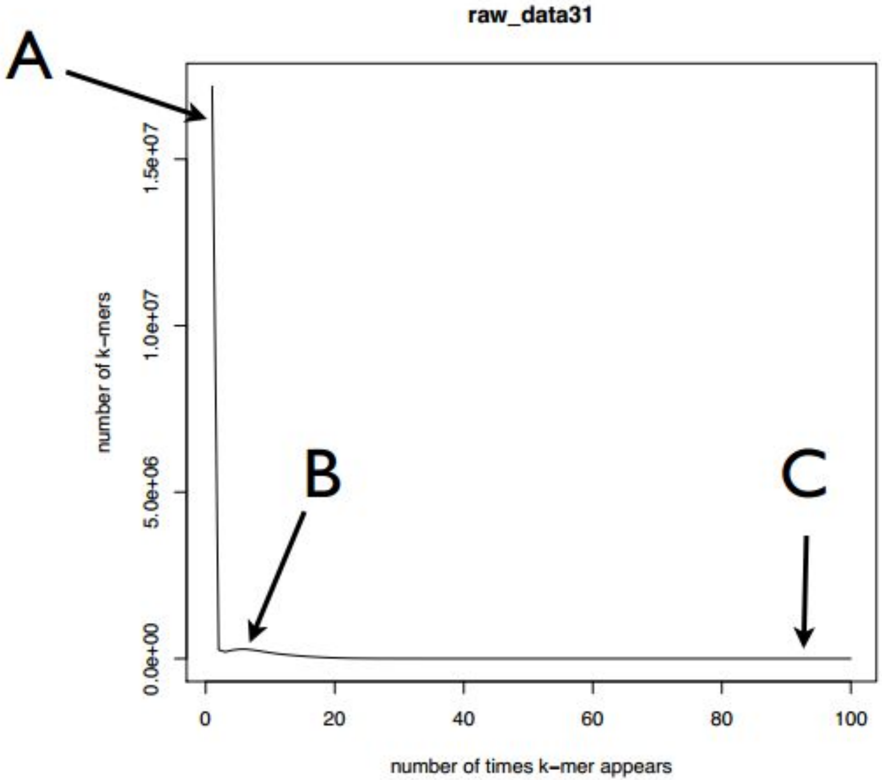
/informatics k-mer distribution

Which arrow is pointing to the sequencing errors?

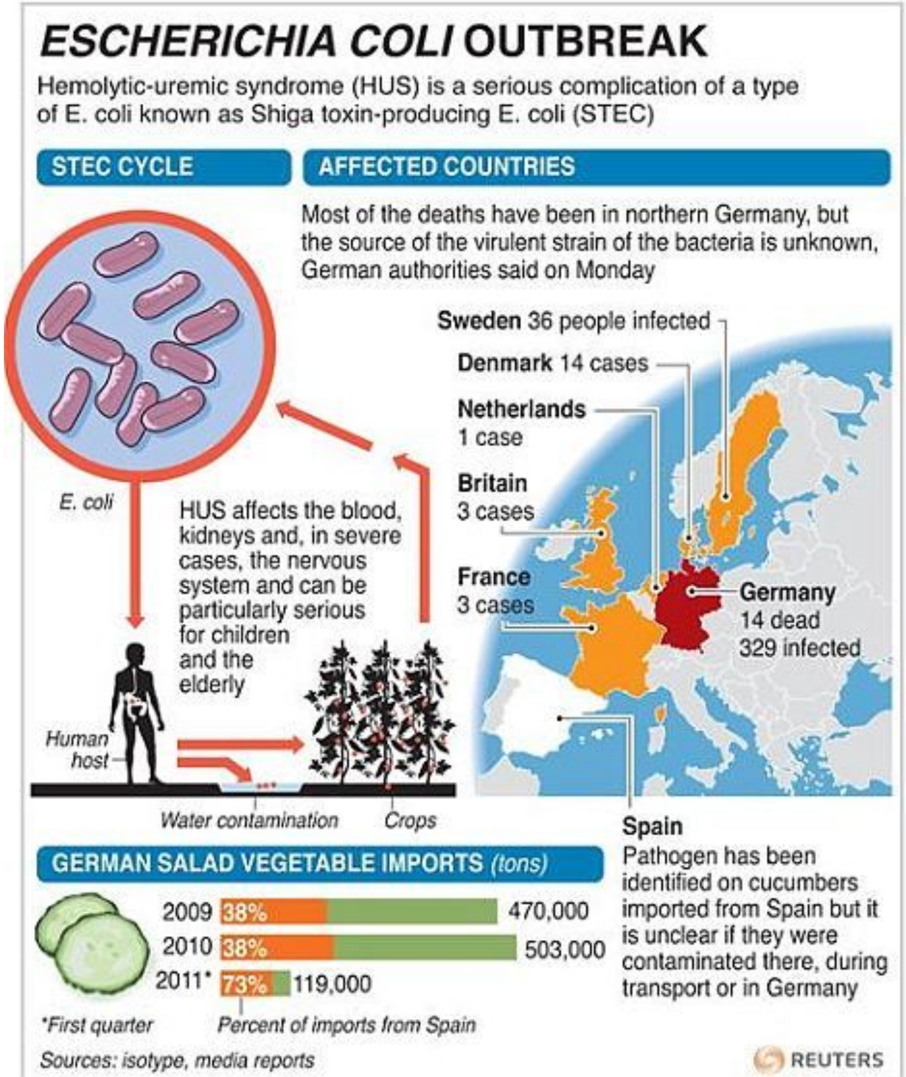


/informatics k-mer distribution

Which arrow is pointing to the kmers from repeated regions?



Case study - *E.coli* outbreak



ESCHERICHIA COLI OUTBREAK

Hemolytic-uremic syndrome (HUS) is a serious complication of a type of E. coli known as Shiga toxin-producing E. coli (STEC)

HUS affects the blood, kidneys and, in severe cases, the nervous system and can be particularly serious for children and the elderly

Most of the deaths have been in northern Germany, but the source of the virulent strain of the bacteria is unknown, German authorities said on Monday

Pathogen has been identified on cucumbers imported from Spain but it is unclear if they were contaminated there, during transport or in Germany

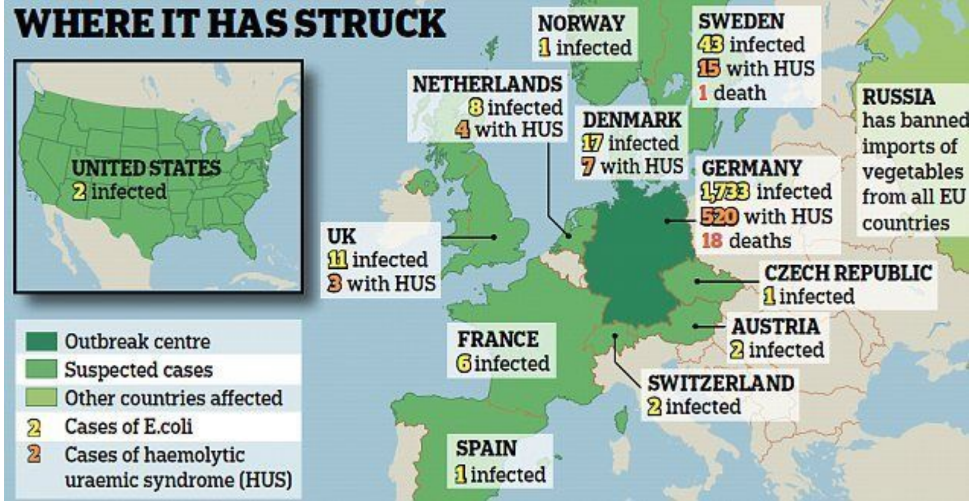
GERMAN SALAD VEGETABLE IMPORTS *(tons)*

2009	38%	470,000
2010	38%	503,000
2011*	73%	119,000

*First quarter

Percent of imports from Spain

Sources: isotype, media reports



WHERE IT HAS STRUCK

UNITED STATES 2 infected
NORWAY 1 infected
SWEDEN 43 infected 15 with HUS 1 death
NETHERLANDS 8 infected 4 with HUS
DENMARK 17 infected 7 with HUS
GERMANY 1,733 infected 520 with HUS 18 deaths
UK 11 infected 3 with HUS
CZECH REPUBLIC 1 infected
AUSTRIA 2 infected
FRANCE 6 infected
SWITZERLAND 2 infected
SPAIN 1 infected

RUSSIA has banned imports of vegetables from all EU countries

- Outbreak centre
- Suspected cases
- Other countries affected
- 2 Cases of E.coli
- 2 Cases of haemolytic uraemic syndrome (HUS)

- What is the genome sequence of *E.coli* X?
- What strain of *E.coli* is *E.coli* X most similar to? (Where did it come from?)
- What are the genes that *E.coli* X contains?
- Which of these genes make *E.coli* X distinct?
- How did *E.coli* X evolve to obtain these genes?
- How did *E.coli* X become pathogenic?