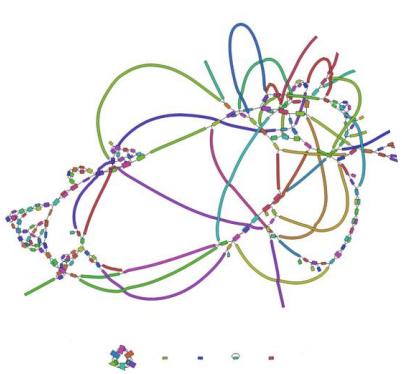
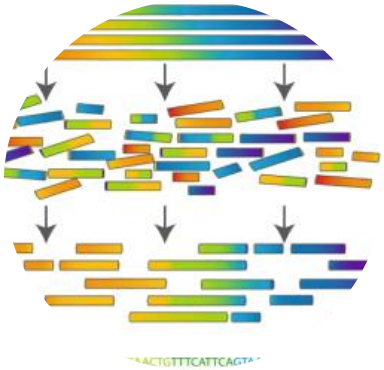
MASTER IN BIOPHARMACEUTICAL SCIENCES

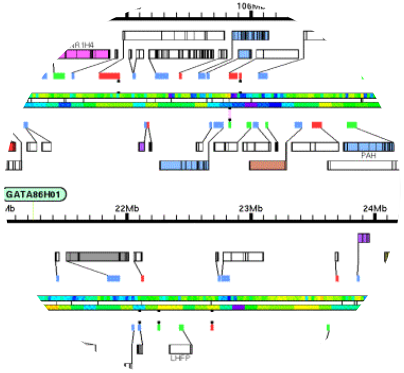
Bacterial Molecular Genetics

LISBON 2022

De novo Assembly







João Perdigão

imed Research Institute for Medicines

Going Genome-Wide!

Before … and after NGS and Whole-Genome Sequencing

illumina®

Oxford NANOPORE Technologies

PACBIO®

Roche 454 SEQUENCING

ion torrent by life technologies™

	Illumina	Ion Torrent	PacBio	Oxford Nanopore
Read Length (bp)	50-300	200-400	10k-40k	1Mb
Output (Gb)	6000	0.05-1	0.5-1	5-40
Cost/MB (USD)	0.05-0.15	1	0.13-0.60	0.13-0.60

Sanger Cost per MB: 2400USD

WGS and Microevolution towards Drug Resistance

Microevolutionary trajectory of the Lisboa3 clade shows multiple development of M/XDR-TB

Lisboa3

INH/STR
inhA C-15T/S94A
rpsL K43R

1973
RIF
rpoB S531L

FQ
gyrA D94G

1990
EMB/FQ/KAN
embB P397T
gyrA S91P
eis G-10A

CAP
tlyA Ins755GT

PZA
pncA
various mutations

RIF
rpoB S531L

RIF
rpoB S531L

EMB
embB M306I

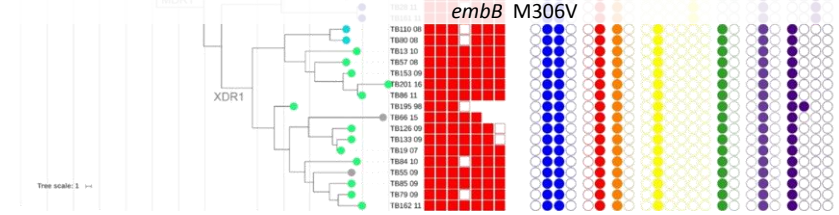
RIF 1981
rpoB D516V

RIF 1975
rpoB S531L

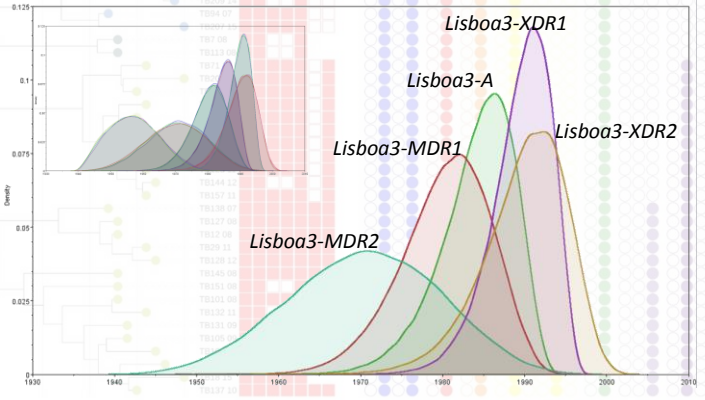
FQ
gyrA D94G

1989
EMB/PZA/KAN
embB M306V
pncA L120P
eis G-10A

EMB
embB M306V



Microevolutionary trajectory of the Q1 clade shows sequential development of M/XDR-TB



Bayesian Dating Estimates (BEAST)
Uncorrelated lognormal clock
Four 50 million MCMC Runs

Q1

1968
INH/STR
inhA C-15T/I194T
gidB A80P

1977
RIF/EMB/PZA
rpoB S531L
embB M306V
pncA V125G

FQ
gyrA D94A

1994
SLID
rrs A1401G

MDR-TB Beijing Strains in Portugal and Guinea-Bissau: missing links?

Emerging Microbes & Infections
2020, VOL. 9
https://doi.org/10.1080/22221751.2020.1774425

Emergence of multidrug-resistant *Mycobacterium tuberculosis* of the Beijing lineage in Portugal and Guinea-Bissau: a snapshot of moving clones by whole-genome sequencing

João Perdigão [a*], Carla Silva [a†], Fernando Maltez[b†], Diana Machado [c], Anabela Miranda [d], Isabel Couto [c], Paulo Rabna[e], Paola Florez de Sessions [f], Jody Phelan [g], Arnab Pain [h], Ruth McNerney [i], Martin L. Hibberd [g], Igor Mokrousov [j], Taane G. Clark [g], Miguel Viveiros [c] and Isabel Portugal [a*]

14 Isolates with available WGS

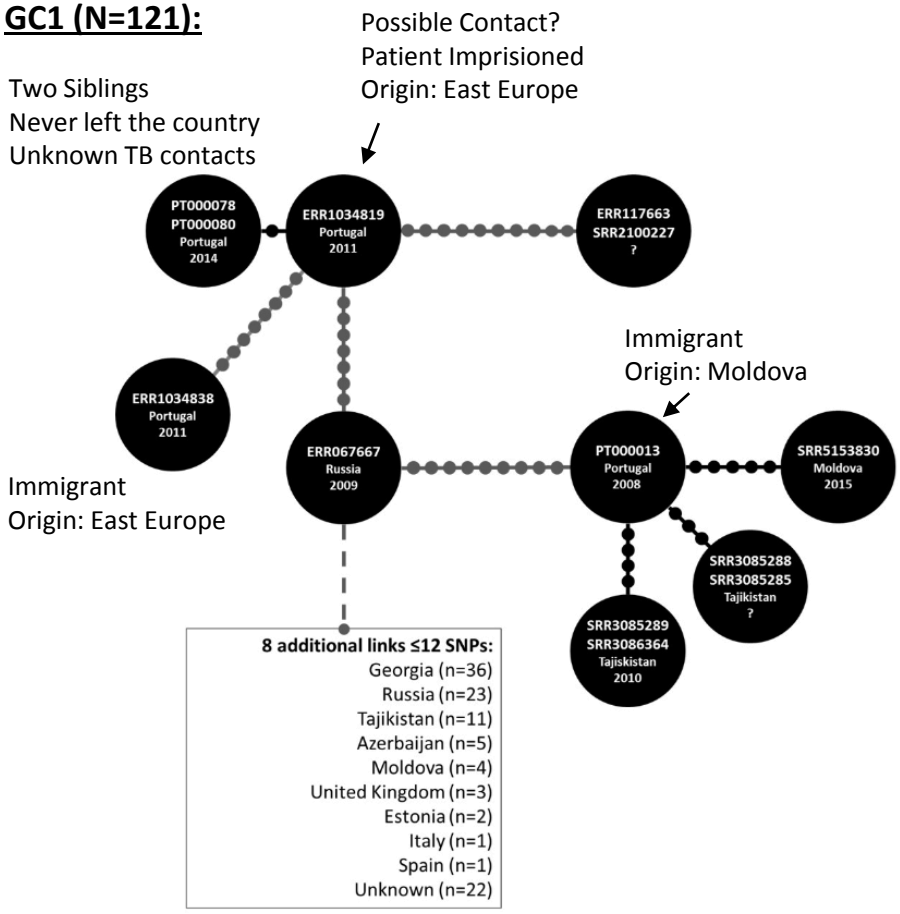
+

5 296 Lineage 2 strains (26 740 coreSNPS)

CCs: distance within ≤ 12 SNPs

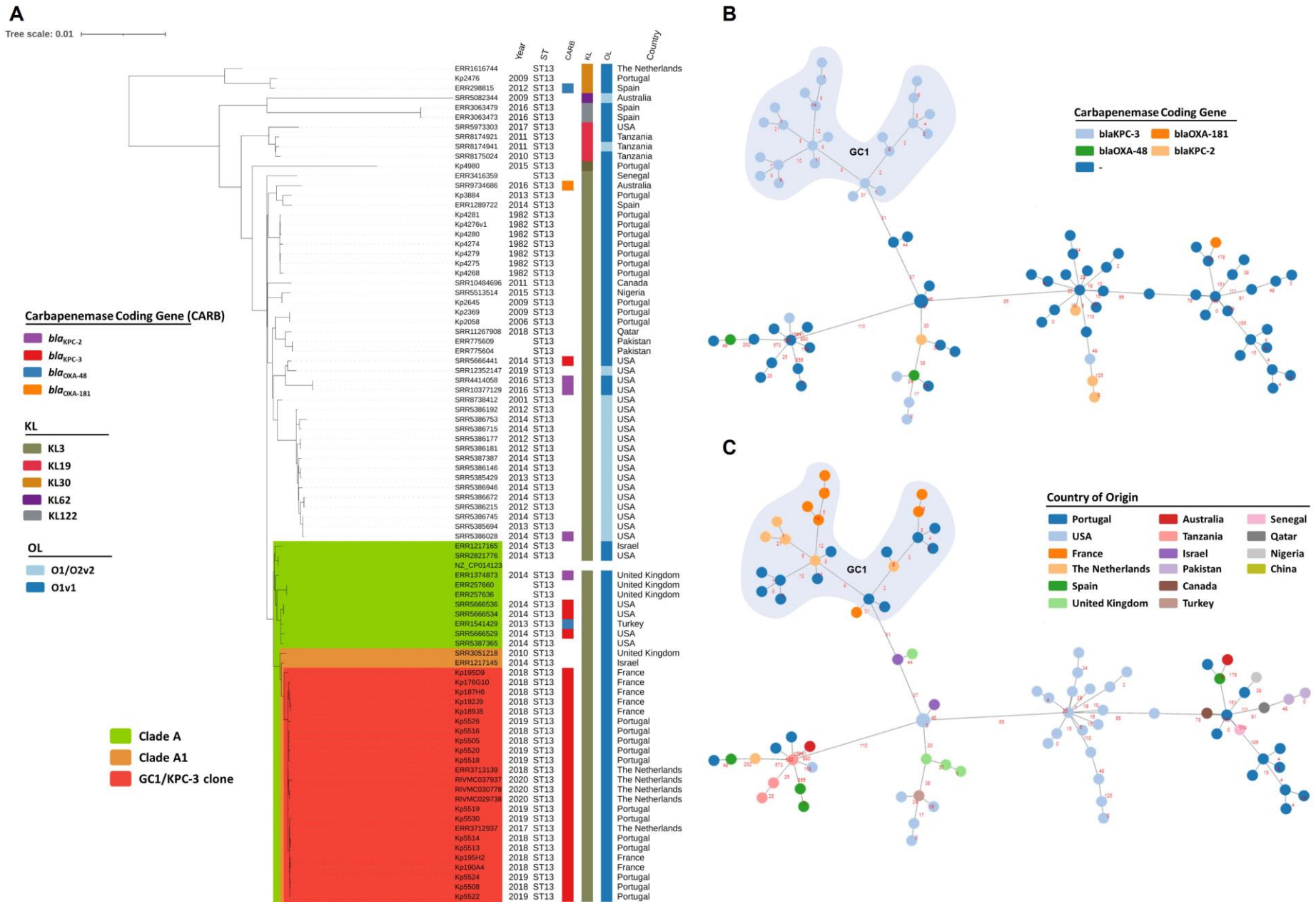
All strains in GC1 were positive for the kdpD binucleotide deletion marker for the Russian B0/W148, part of the MIRU-VNTR 100-32 supercluster.

GC1 (N=121):



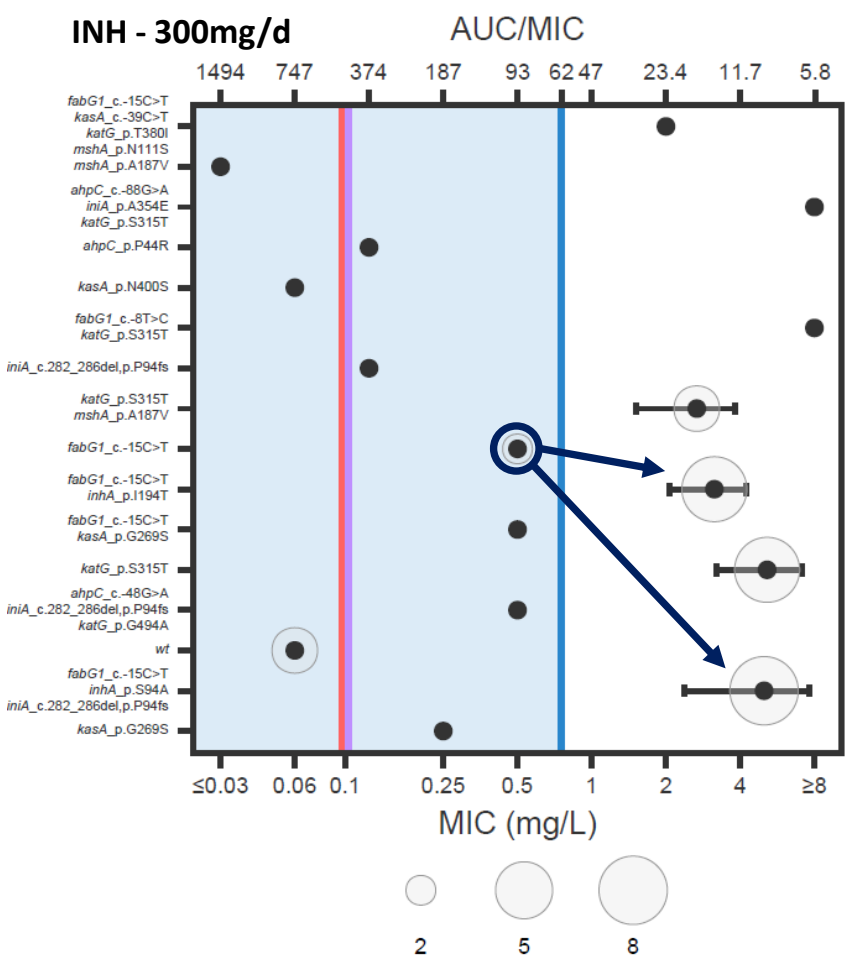
imed Research Institute for Medicines

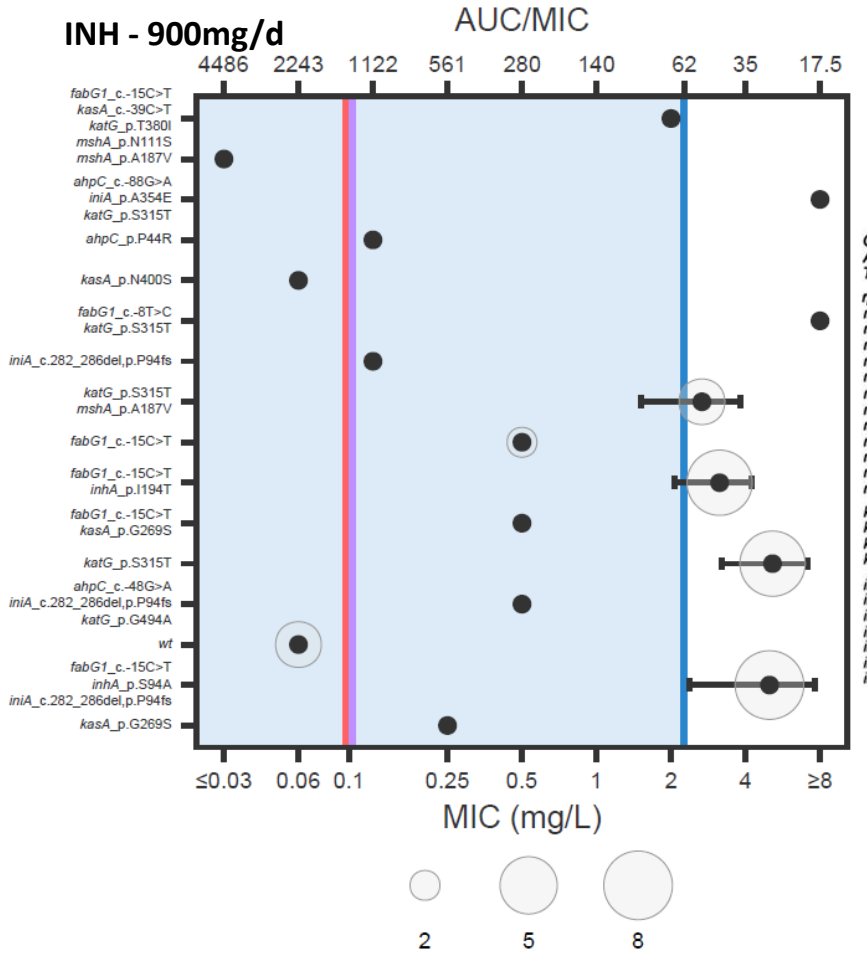
K. pneumoniae ST13/KPC-3 – going international from PT!



WGS and Microevolution towards Higher Drug Resistance Levels!!

Double inhA mutations associated with Lisboa3 and Q1 clades are associated with higher INH MICs

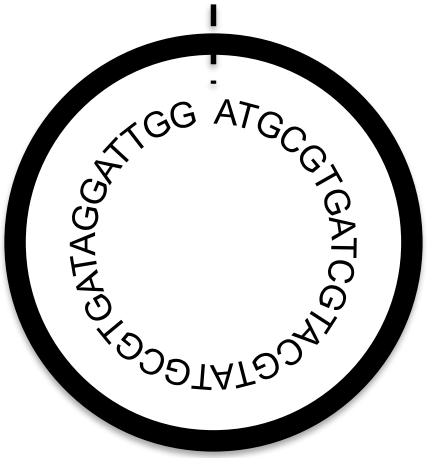




imed Research Institute for Medicines

Whole Genome Sequencing… the Illumina way

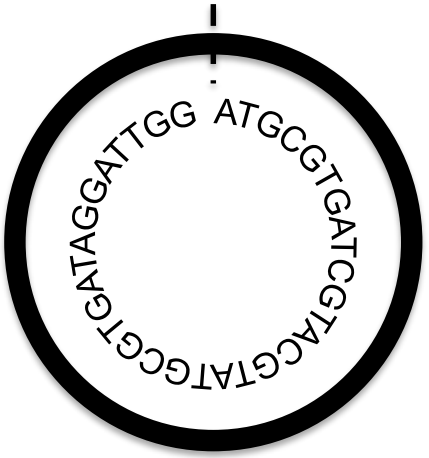
Ideally…



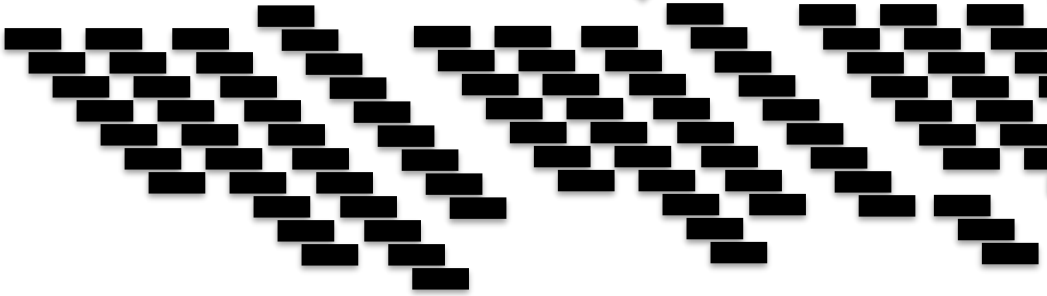
Sequencing

ATGCGTGATCGTACGTATGCGTGATAGGATTGG

The harsh reality…

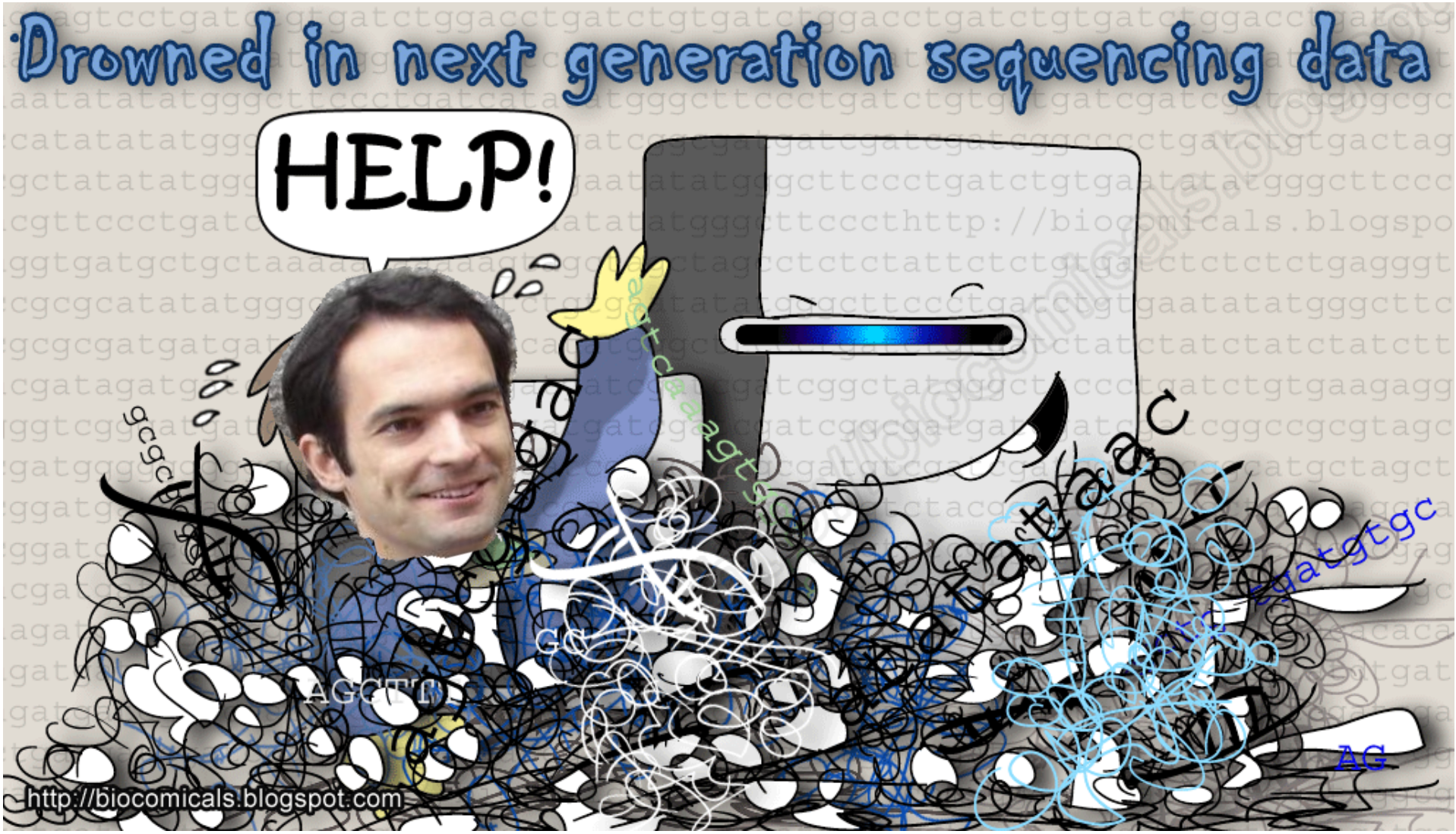


Sequencing



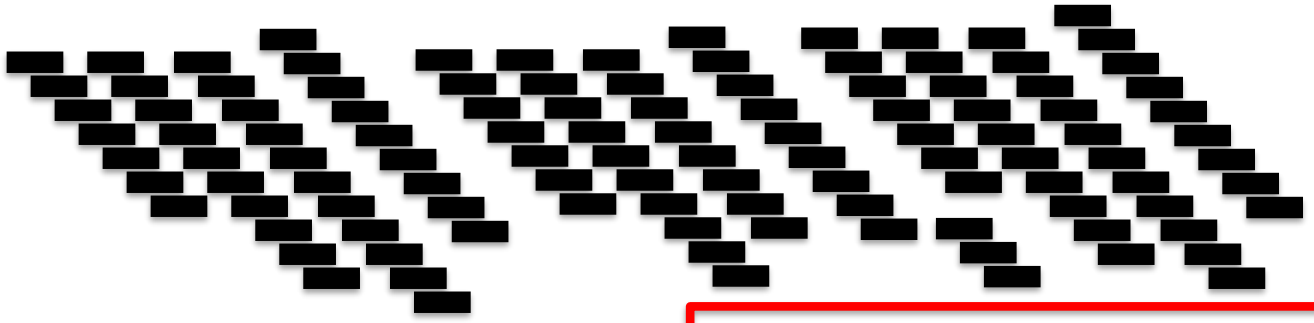
imed Research Institute for Medicines

Whole Genome Sequencing



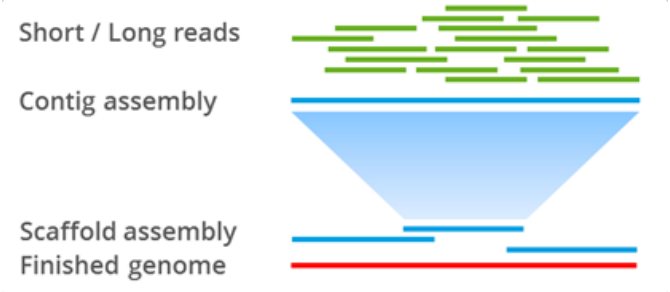
imed Research Institute for Medicines

Two main avenues for handling short-read data…



Mapping or Reference Assembly

De novo Assembly



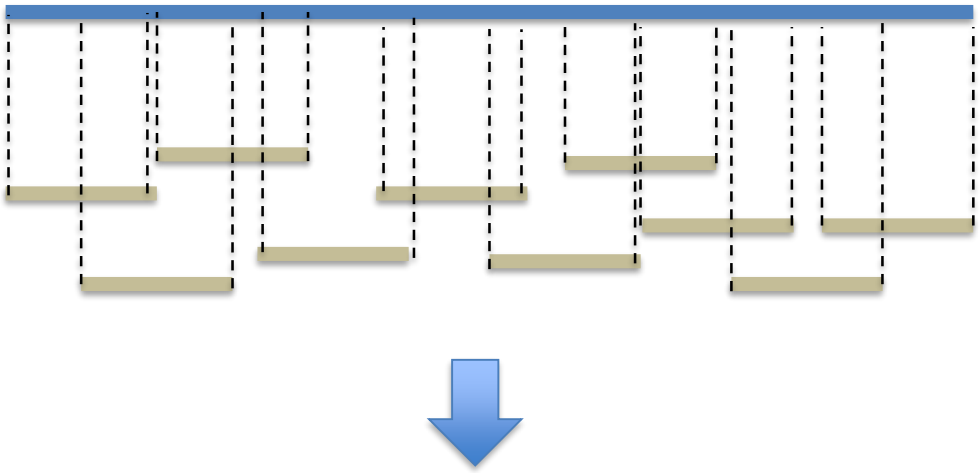
De novo Assembly

De novo assembly – genome assembly without a reference genome, i.e., starting *de novo* from sequence data

De novo assembly ≠ Reference assembly

Unknown Genome

Sequence Reads



What we think the Genome is!!

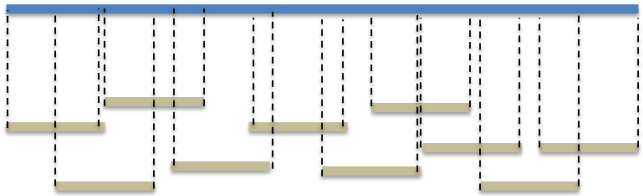
imed Research Institute for Medicines

De novo Assembly

*de novo a*ssembly attempts to reconstruct genomes by exploring read overlapping and contiguity

Problems and Challenges:

- Large volume of sequence reads
- Sequencing errors
- Genomic repeat patterns/regions and homopolymers
- Uneven coverage/sequencing



But, what is the definition of an assembly?

Best set of sequences that can approximate the sequenced genetic material

Implications?

Why to assemble?

Objective/Purpose of the Assembly:

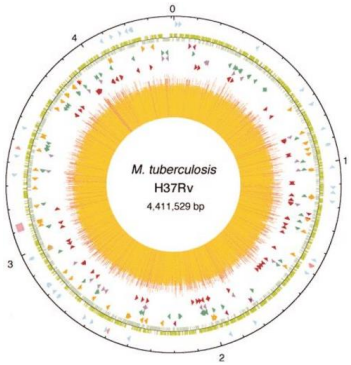
Obtain a reference genome

↓

PacBio

↓

Finished Genome Assembly



Manual closure
Mate-pair sequencing

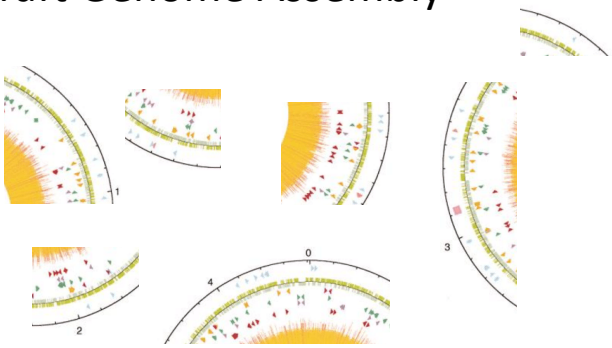
Gene content, insertions, deletions

↓

Illumina

↓

Draft Genome Assembly



imed Research Institute for Medicines

Assembly Methods

Three major methods for assembly:

i) overlap-and-extend

Finds read overlap and extends – suffix of a read is equal to the prefix of another read with a length that meets a defined threshold.

Software: SSAKE, VCAKE and SHARCGS

ii) string graph

Construction of string graph where each read is a vertex with edges connecting overlapping nodes.

Software: Edena and BOA

Problems: high memory comsumption and sequencing errors

Assembly Methods

Three major methods for assembly:

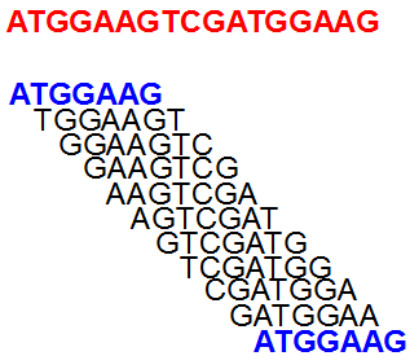
iii) de Bruijn Graph

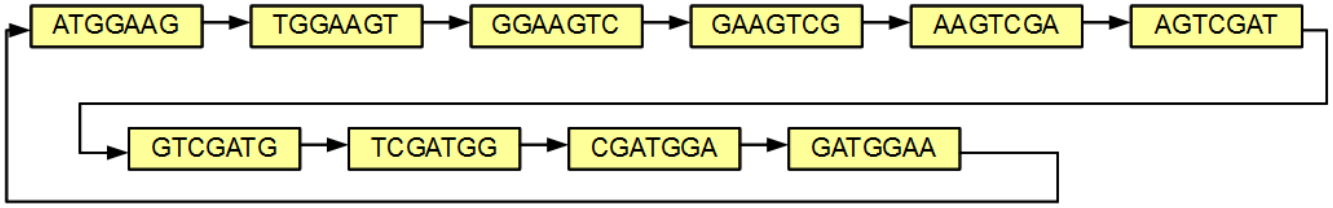
Each vertex represents a substring of length *k* (*k*-mer) in a read. Edges connect vertexes if these are consecutive vertexes, i.e., the last *k*-1 nucleotides in k-mer *u* are the same as the as the first *k*-1 nucleotides of k-mer *v.*

Software: Velvet, SOAPdenovo, SPAdes

Most widely used approaches.

Objective: represent every possible *k*-mer present in the genome!





Assembly Methods

Three major methods for assembly:

iii) de Bruijn Graph

Connecting the nodes:

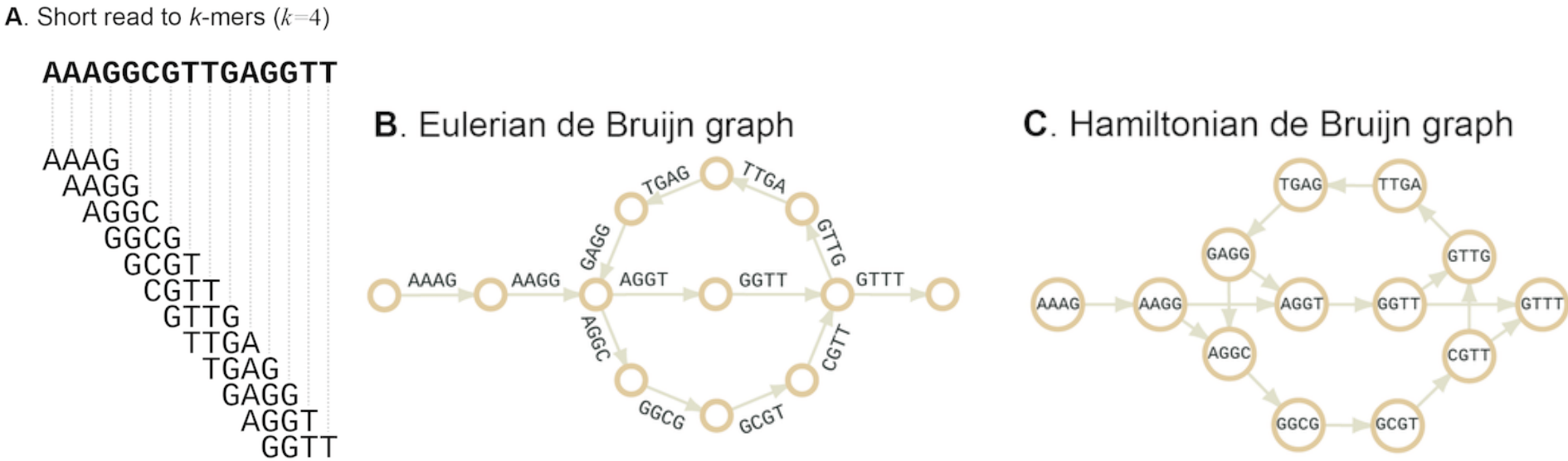
•*Hamiltonian path*: visits **every vertex in the graph** (exactly once, because it is a path)

•*Eulerian trail*: visits **every edge in the graph** exactly once (because it is a trail, vertices may well be crossed more than once.)

William Hamilton

Leonhard Euler

Nicolaas de Bruijn



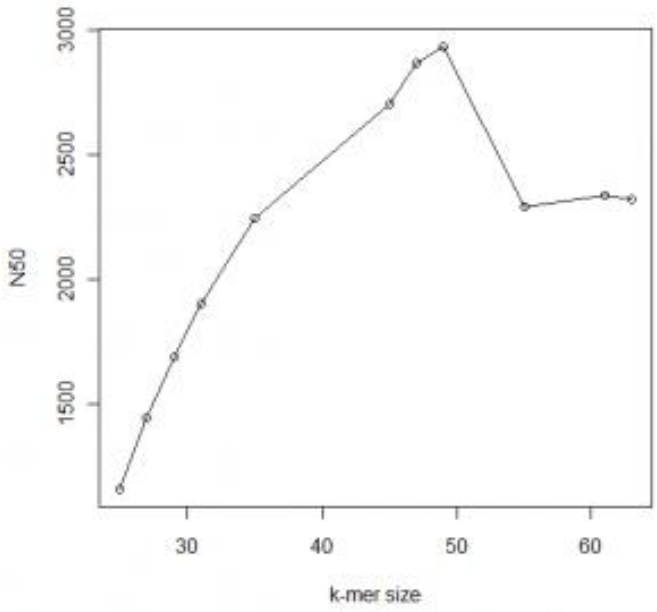
Assembly Methods

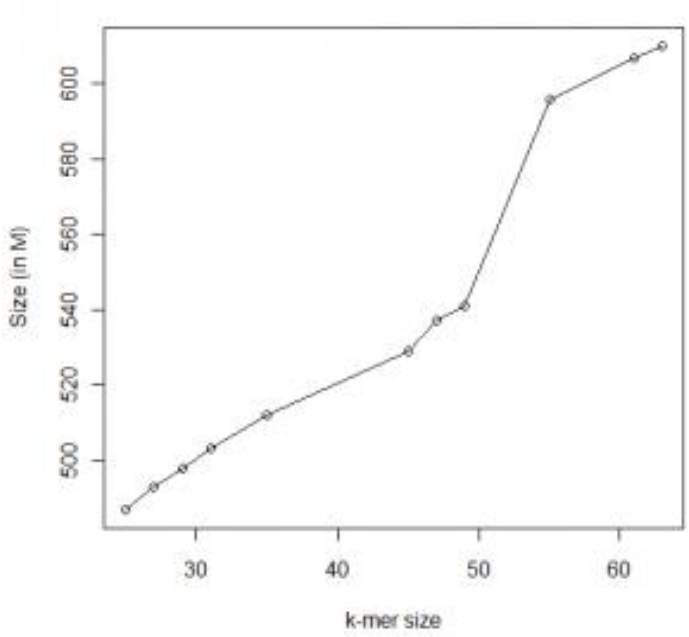
Three major methods for assembly:

iii) de Bruijn Graph

Choice of the k-mer length: always below the read-length of your data!

The k-mer length should be an odd number: avoid palindromes – an odd-sized k-mer cannot form palindromes when reverse-complemented!



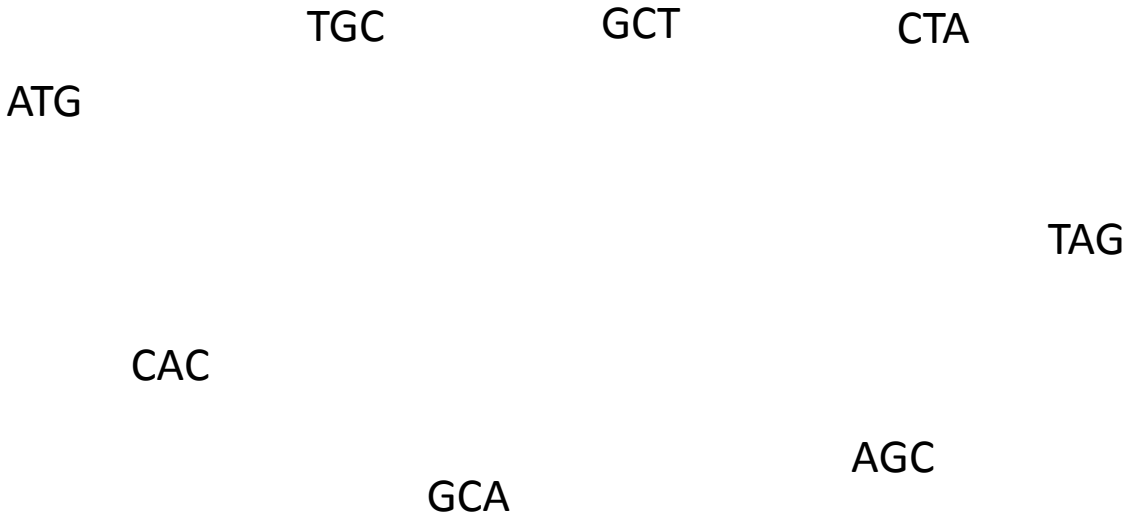


https://homolog.us/blogs/genome/2012/10/10/multi-kmer-de-bruijn-graphs/

De Bruijn Graphs - Exercise

Let's consider the following reads:

ATGCTA
GCTAGC
TAGCAC
CTAGCA

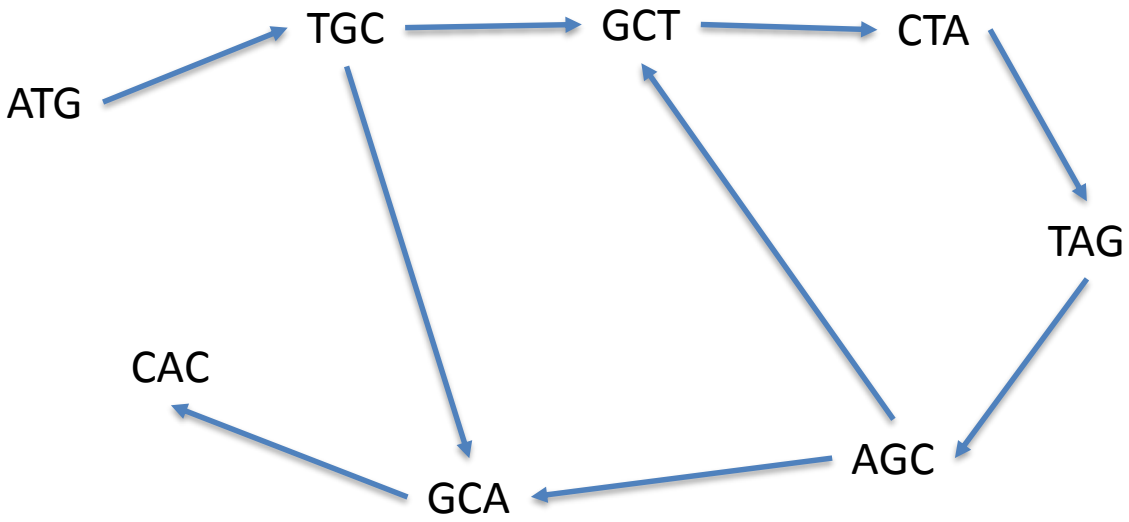


1 List all 3bp k-mers
2 Establish links (edges) between k-mers differing by k-1 nucleotides
3 Visit all nodes and use the minimal path length

De Bruijn Graphs - Exercise

Let's consider the following reads:

ATGCTA
GCTAGC
TAGCAC
CTAGCA

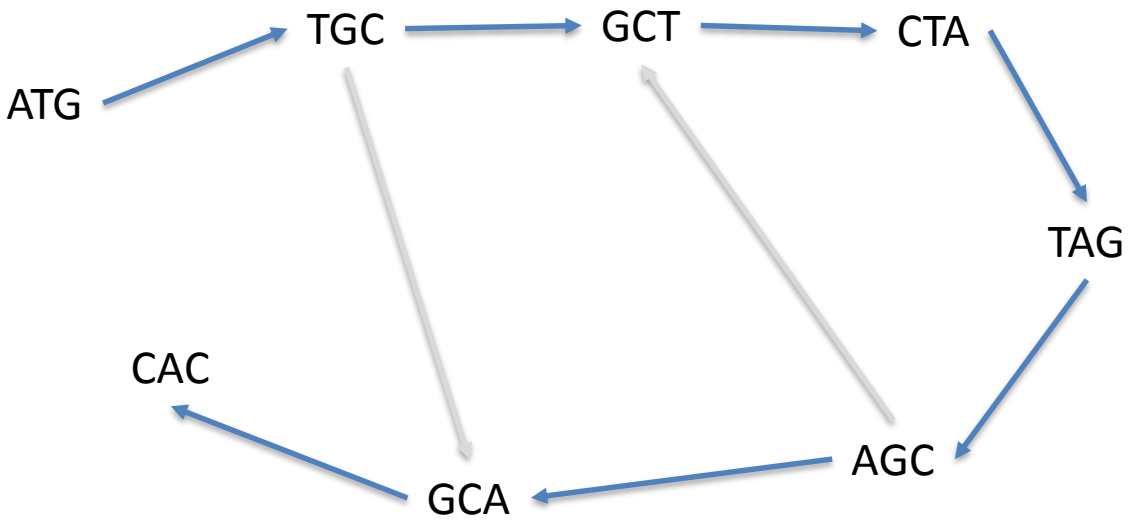


1 List all 3bp k-mers
2 Establish links (edges) between k-mers differing by k-1 nucleotides
3 Visit all nodes only once

De Bruijn Graphs - Exercise

Let's consider the following reads:

ATGCTA
GCTAGC
TAGCAC
CTAGCA



ATGCTAGCAC

1 List all 3bp k-mers
2 Establish links (edges) between k-mers differing by k-1 nucleotides
3 Visit all nodes only once

De Bruijn Graphs - Exercise

Let's consider the following reads:

ATGCTA
GCTAGC
TAGCAC
CTAGCA

The Eulerian Path

1 List all 3bp k-mers
2 Define nodes between edges
3 Visit all edges only once

De Bruijn Graphs - Exercise

Let's consider the following reads:

ATGCTA
GCTAGC
TAGCAC
CTAGCA

The Eulerian Path

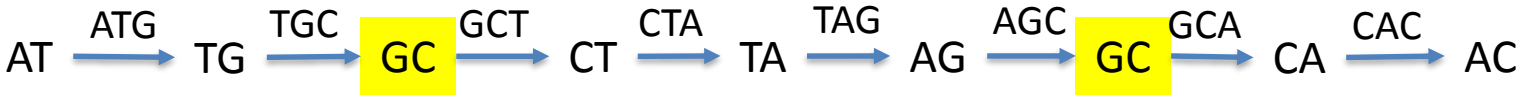
1 List all 3bp k-mers
2 Define nodes between edges
3 Visit all edges only once

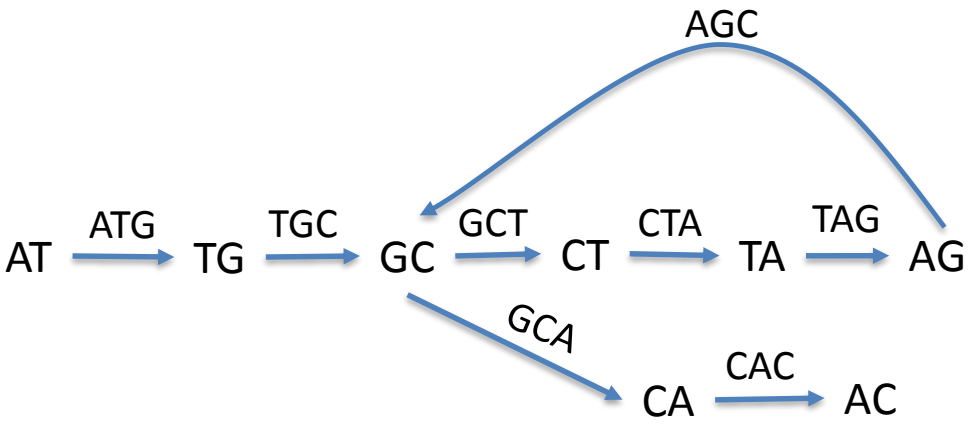
De Bruijn Graphs - Exercise

Let's consider the following reads:

ATGCTA
GCTAGC
TAGCAC
CTAGCA

The Eulerian Path





ATGCTAGCAC

1 List all 3bp k-mers
2 Define nodes between edges
3 Visit all edges only once

Evaluating and Comparing Assemblies

Metrics to Evaluate and Compare Assemblies:

Evaluating an assembly can be reference-free or comparing to a reference genome**!**

Purpose:

1. Assess the individual quality of an assembly;
2. Compare assemblers.

Metrics commonly used:

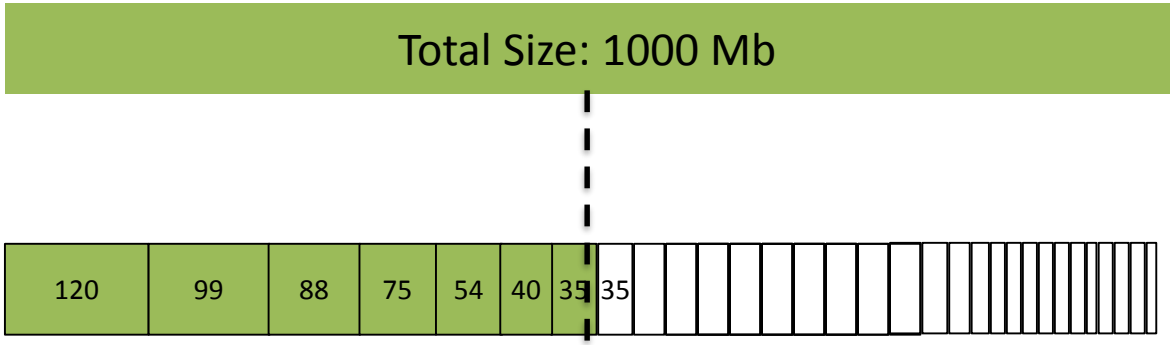
- Number of contigs/scaffolds
- Total length of the assembly
- Length of the largest contig/scaffold
- Percentage of gaps in scaffolds ('N')
- N50/NG50 of contigs/scaffolds
- Number of predicted genes
- Number of core single-copy genes

Software:

QUAST – allows comparison against reference

BUSCO – evaluates the presence of core single-copy orthologous genes

N50 and L50 and other metrics



N50 – shortest contig length spanning the midpoint of the assembly length (after sorting from largest to smallest contig);

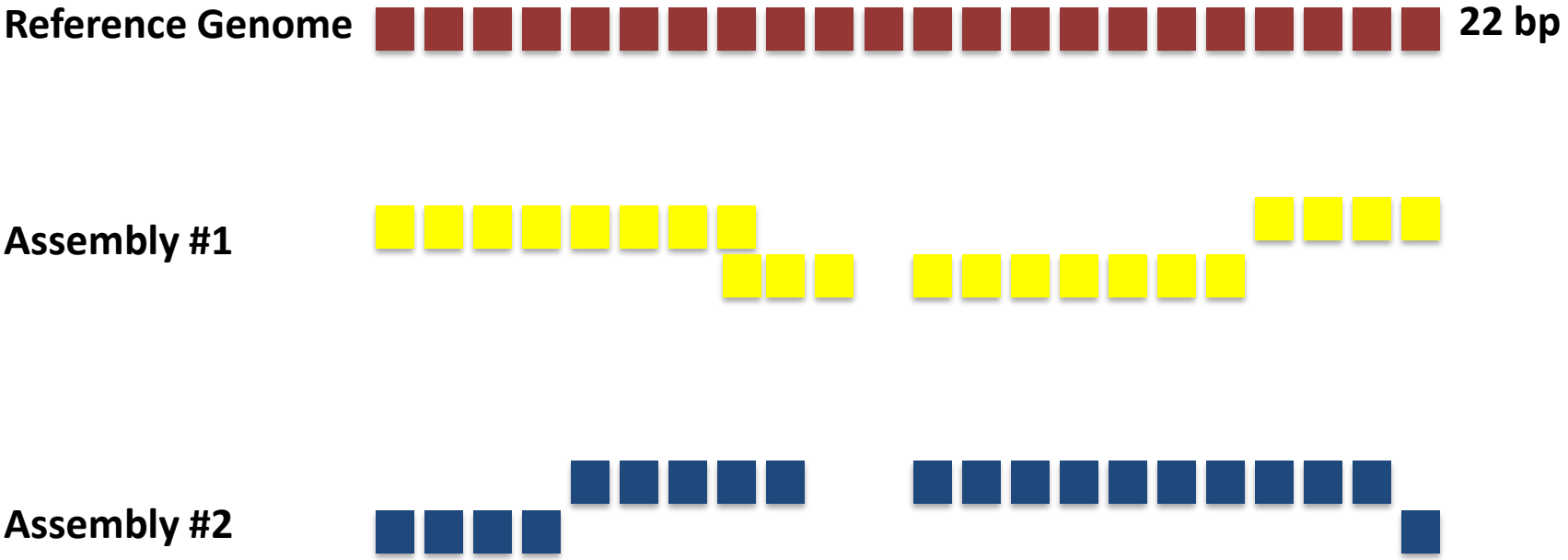
NG50 – shortest contig length spanning the midpoint of the estimated genome size (after sorting from largest to smallest contig);

L50 – number of contigs necessary to span the midpoint of the assembly length (after sorting from largest to smallest contig)

LG50 – number of contigd necessary to span the midpoint of the estimated genome size (after sorting from largest to smallest contig)

N50 and L50 in this exemple?

Evaluating and Comparing Assemblies

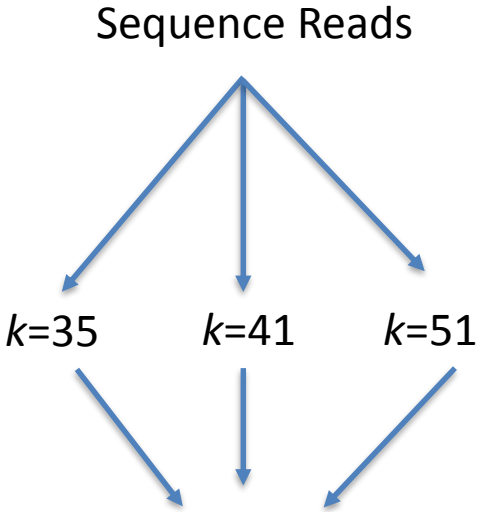


Calculate:	Assembly #1	Assembly #2
N50	**7**	**10**
NG50	7	**5**
Coverage	**95.4%**	**90.9%**
Total Length of Assembly	**22**	**20**

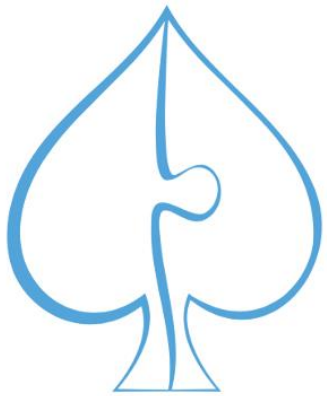
Multi *k-mer* assembly – the way forward

Multi k-mer asssembly:

Multi-k-mer stategies always provide better results than single k-mer



Merge Assemblies -> Velvet+cd-hit+minimus2
Iterative Assembly removing assembled reads – IDBA
Multi-kmer de Bruijn graph - **SPAdes**



SPAdes 3.13.0

imed Research Institute for Medicines

Assembly pipelines

Unicycler

As input, Unicycler takes one of the following:
•Illumina reads from a bacterial isolate (ideally paired-end, but unpaired works too)
•A set of long reads (either PacBio or Nanopore) from a bacterial isolate (uncorrected long reads are fine, though corrected long reads should work too)
•Illumina reads and long reads from the same isolate (best case)

Unicycler does:
- ***Short-read assembly***
- ***Long-read assembly***
- ***Hybrid assembly***

Reasons to use Unicycler:
•It circularises replicons without the need for a separate tool like Circlator.
•It handles plasmid-rich genomes.
•It can use long reads of any depth and quality in hybrid assembly. 10x or more may be required to complete a genome, but Unicycler can make nearly-complete genomes with far fewer long reads.
•It produces an assembly *graph* in addition to a contigs FASTA file, viewable in Bandage.
•It has very low misassembly rates.
•It can cope with very repetitive genomes, such as *Shigella*.
•It's easy to use: runs with just one command and usually doesn't require tinkering with parameters.

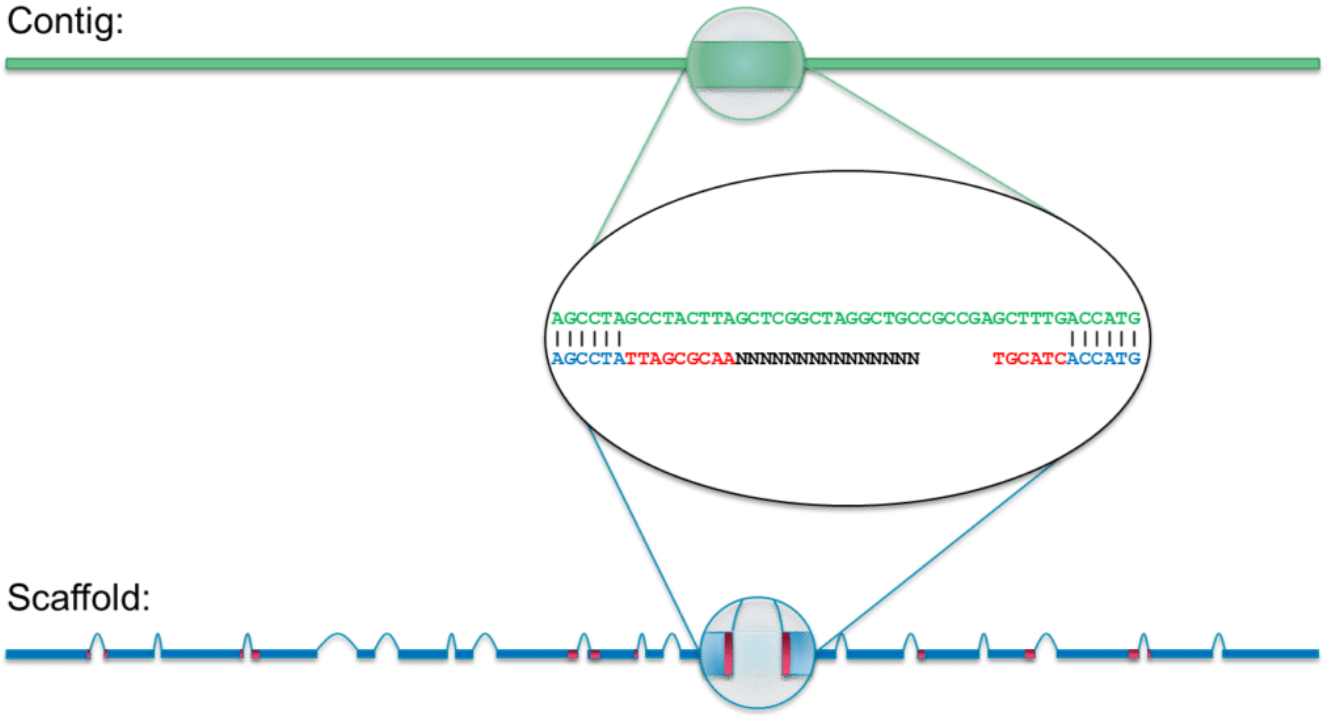
Reasons to not use Unicycler:
•You're assembling a eukaryotic genome or a metagenome (Unicycler is designed exclusively for bacterial isolates).
•Your Illumina reads and long reads are from different isolates (Unicycler struggles with sample heterogeneity).
•You're impatient (Unicycler is thorough but not especially fast).

https://github.com/rrwick/Unicycler

Scaffolding

Contigs are continuous stretches of sequence containing only A, C, G, or T bases without gaps.

*Scaffolds are created by chaining contigs together **using additional information** about the relative position and orientation of the contigs in the genome*



ABACAS

MAUVE

ARCS