MaCuMBA

Grupo de Genómica Evolutiva
Evolutionary Genomics Group

PROKARYOTIC GENOME ASSEMBLY AND ANNOTATION

Mario López-Pérez

μgen

Gobierno de España
Ministerio de Ciencia e Innovación

Consolider

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 311957.

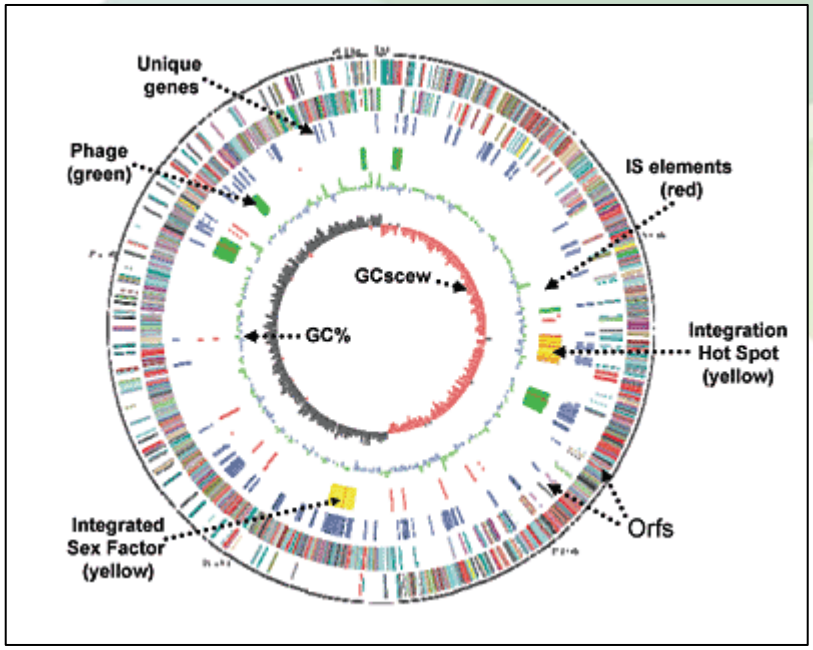
www.macumbaproject.eu

Bacterial genome organization

Chromosomes

- Most commonly single circular chromosome (always DNA)
  - BUT many species have linear chromosome(s) (e.g. *Borrelia, Streptomyces, Rhodoccus*)
  - BUT a few species with two chromosomes (e.g. *Vibrio cholerae*)
- Can be mix of circular and linear (e.g. *Agrobacterium tumefaciens, B. burgdoferi*)
- Operons with promoters just upstream

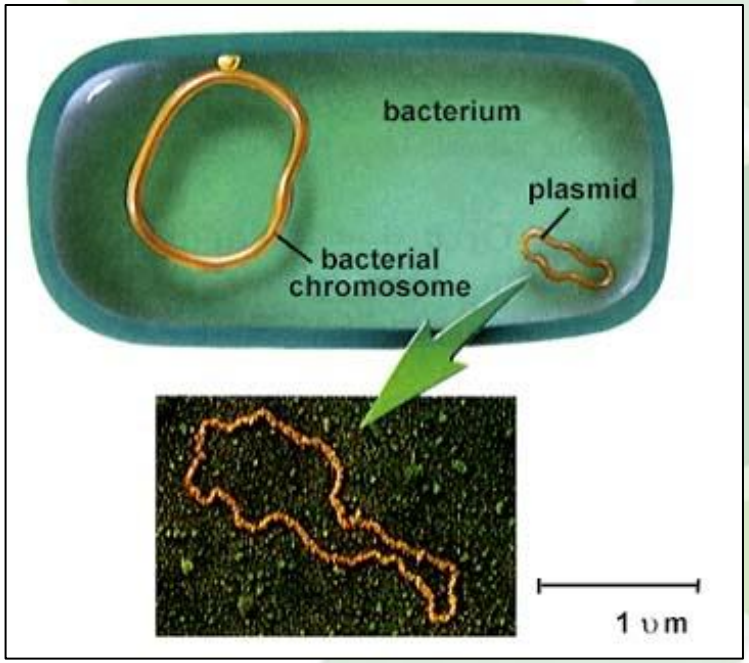
- Gene density high (>90%)
  - intergenic regions short
  - very little repetitive or non-coding DNA
  - Introns very rare
- Protein-coding genes (CDS) short (~1kbp)



Bacterial genome organization

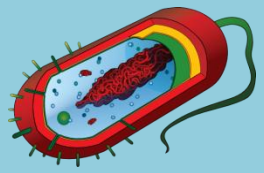
Plasmids

- Independent autonomous replicon, can be circular or linear
- may integrate into chromosome
- copy number varies 1 to 10s
- often carry non-essential genes that confer an adaptive advantage in certain conditions



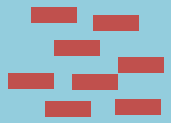
Previous Works

Wet lab work



Extract DNA

Shear DNA



Sequence

Sequencing

Draft assembly

Finishing

Annotation

Public release

Contents

Contents

What is genome assembly?

Genome assembly is the process of reconstructing the original DNA sequence(s) of an organism from the read sequences

Genome
(3.000.000 letters)

Sequencing →

Reads
(50-500 letters each)

Assembly →

Genome
(3.000.000 letters)

Ideal world

Reads unambiguous (long) and error-free
Simple deduction problem

Real world

Reads ambiguous (too short) and error-prone
Complicated inference problem

Fastq format

@Sequence ID

Sequence (A, C, G, T, N)

+Sequence ID

ASCII encoded quality score

s_?_1_sequence.txt

```
@HWI-ST741_0085:1:1101:1444:1939#0/1
ATAGTTACAATCGATCCATTTGCAGAGTACAGATACATGATAGGGGAATACCATTTAGTTCAAGGGAAAGCCAGCAAAGTTCGATCAAGGATAGACCGAGG
+HWI-ST741_0085:1:1101:1444:1939#0/1
ffffdfdfffffgggfafffcdfcfffbfdddeaegfgfgafaffW^a]a`b^d[dbdbbcccccb[`dW\a`a`^caa[ffcffafaccfcd_fcfc_cb
@HWI-ST741_0085:1:1101:1417:1939#0/1
GCTCAGTACTGGGACCCCAGCTATTTACAAAATATATTAATGATCTGGATGAGGGATTTGAAGGCAATATCTCCAAGTTTGCGGATGACATTAAGCTGGGG
+HWI-ST741_0085:1:1101:1417:1939#0/1
hhhhghhhhhhhhhhhhhhhhhhhhehhhhhhhghhhhhhhhhhhhhghhhghhghhhhahfghhhfhhhhhggaffffhhhdghg[hghhffhhhageh
...
```

s_?_2_sequence.txt

```
@HWI-ST741_0085:1:1101:1444:1939#0/2
CCCAGCTTATCCTTGCAACTCTTCTTAAATAGAGGCACAACATTAATCACCCTCCCTTCTATGGACTCCTTTTATACCTCACCCTTCCTCGGCAGGGCCAG
+HWI-ST741_0085:1:1101:1444:1939#0/2
edeaadffffcaffcdaeaeffdfdecfefaceccfdffdfddffffdffdfffdcaffffcffgaggggggg`ggggggfgggggggd]eedaa_^BB
@HWI-ST741_0085:1:1101:1417:1939#0/2
ACCATTCGCTTTCTTTACTGCCTGCTGCACCTGCATGCCTACCTTCAATGACTGGTGTACCATGACACCCAGGTCTCGCTGCATCTCCCCCTTTCCCAATC
+HWI-ST741_0085:1:1101:1417:1939#0/2
hhhhhghhhhhhhhhhhhhhhhghhhdhhghhhhhhhhhchhhhhhhhhfghhhhfhhdhfhhdhhahhhfefhhfdfffedhhhehhhhhehhcfffffd
...
```

Fastq format

```
      1          2  3    4    5  6 7
@HWI-ST741_0085:1:1101:1444:1939#0/1
ATAGTTACAATCGATCCATTTGCAGAGTACAGATACATGATACGGGAAT
+HWI-ST741_0085:1:1101:1444:1939#0/1
ffffdfdfffffgggfafffcdfcfffbfdddeaegfgfgafaffW^a]
```

```
      1          2  3    4    5  6 7
@HWI-ST741_0085:1:1101:1444:1939#0/2
CCCAGCTTATCCTTGCAACTCTTCTTAAATAGAGGCACAACATTAATCA
+HWI-ST741_0085:1:1101:1444:1939#0/2
Edeaadffffcaffcdaeaeffdfdecfefaceccfdffdfddffffd
```

1. the unique instrument name
2. flowcell lane
3. tile number within the flowcell lane
4. 'x'-coordinate of the cluster within the tile
5. 'y'-coordinate of the cluster within the tile
6. index number for a multiplexed sample (0 for no multiplexing)
7. the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

Trim sequences

- Quality trimming
  - Based on quality scores
- Ambiguity trimming
  - Remove stretches of Ns
- Adapter sequence trimming
  - Remove sequence adapters
- Base trim
  - Remove a specified number of bases at either 3' or 5' end of the reads
- Length trimming
  - Remove reads shorter or longer than a specified threshold

Assembly approaches

Reference assembly

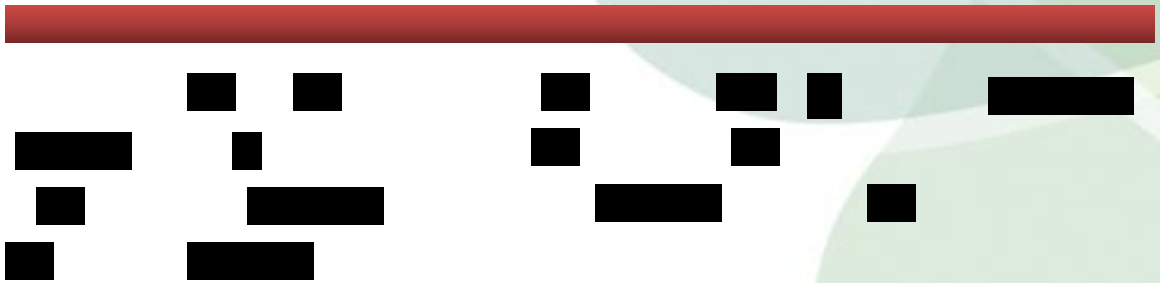
-We have sequence of similar genome
-Reads are aligned to the reference
-Can guide, but also mislead

***de novo* assembly**

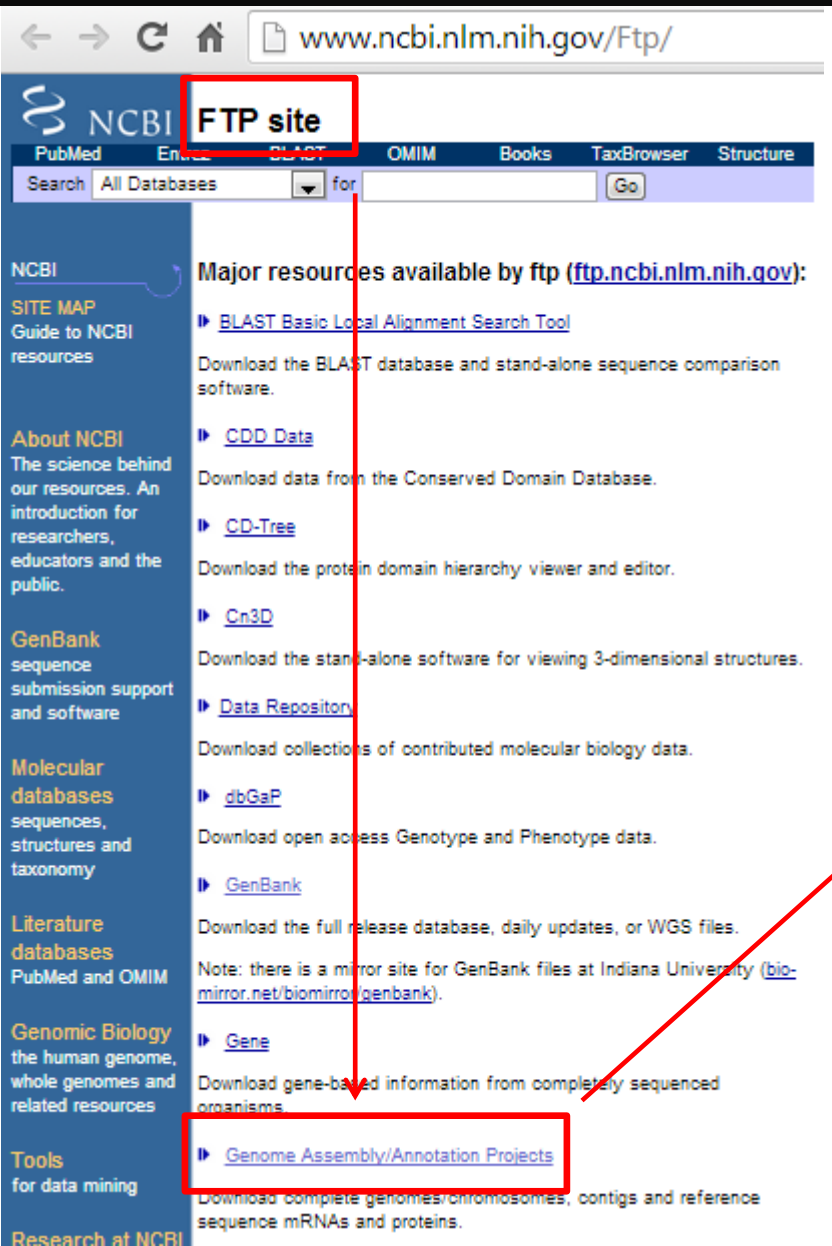
-No prior information about the genome
-Only supplied with read sequences
-Necessary for novel genomes

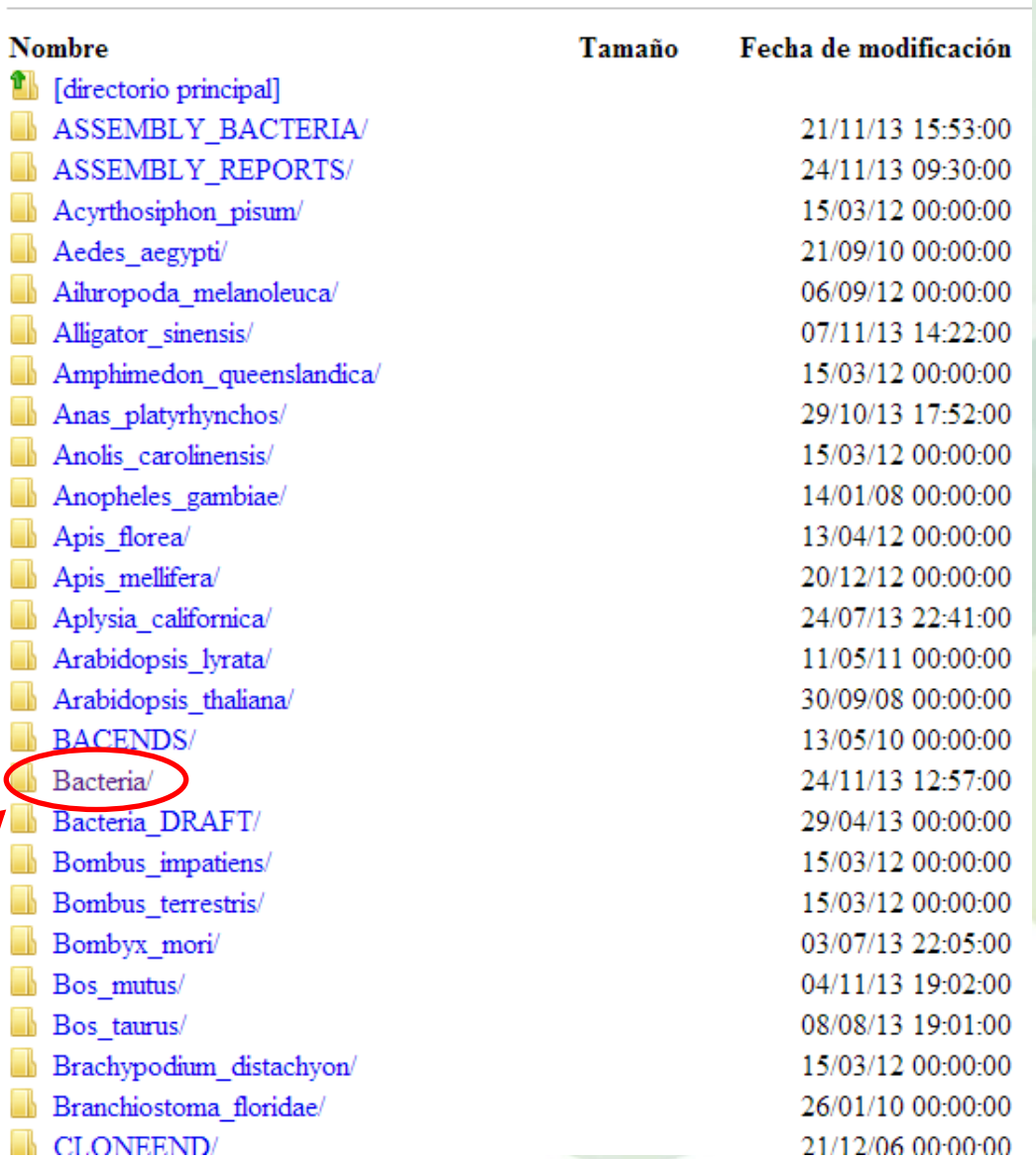
Reference genome

Contigs



Reference Genome





Nombre	Tamaño	Fecha de modificación
[directorio principal]		
ASSEMBLY_BACTERIA/		21/11/13 15:53:00
ASSEMBLY_REPORTS/		24/11/13 09:30:00
Acyrthosiphon_pisum/		15/03/12 00:00:00
Aedes_aegypti/		21/09/10 00:00:00
Ailuropoda_melanoleuca/		06/09/12 00:00:00
Alligator_sinensis/		07/11/13 14:22:00
Amphimedon_queenslandica/		15/03/12 00:00:00
Anas_platyrhynchos/		29/10/13 17:52:00
Anolis_carolinensis/		15/03/12 00:00:00
Anopheles_gambiae/		14/01/08 00:00:00
Apis_florea/		13/04/12 00:00:00
Apis_mellifera/		20/12/12 00:00:00
Aplysia_californica/		24/07/13 22:41:00
Arabidopsis_lyrata/		11/05/11 00:00:00
Arabidopsis_thaliana/		30/09/08 00:00:00
BACENDS/		13/05/10 00:00:00
Bacteria/		24/11/13 12:57:00
Bacteria_DRAFT/		29/04/13 00:00:00
Bombus_impatiens/		15/03/12 00:00:00
Bombus_terrestris/		15/03/12 00:00:00
Bombyx_mori/		03/07/13 22:05:00
Bos_mutus/		04/11/13 19:02:00
Bos_taurus/		08/08/13 19:01:00
Brachypodium_distachyon/		15/03/12 00:00:00
Branchiostoma_floridae/		26/01/10 00:00:00
CLONEEND/		21/12/06 00:00:00

Genome assembly

Bioinformatics

50-500 bp — **Read** → A sequenced piece of DNA

Contig → A set of overlapping DNA segments that represents a consensus region of DNA

N — **Scaffold** → Contigs separated by gaps of known length

4x — **Coverage** → The number of times a specific position in the genome is covered by reads

Challenges with Fragment Assembly

The main problems during the assembly process are:

-Sequencing errors

substitutions, insertions, deletions, and others.

TTTTTATA*GA* (substitution), CCTT—TAAACG (deletion and insertion)

-Repeats

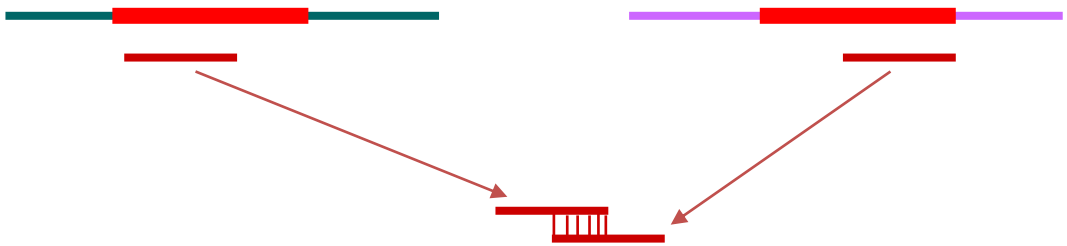
A segment of DNA that occurs more than once in the genome

(Transposons, IS elements**,** Gene duplication)

-Lack of coverage

Size of the data

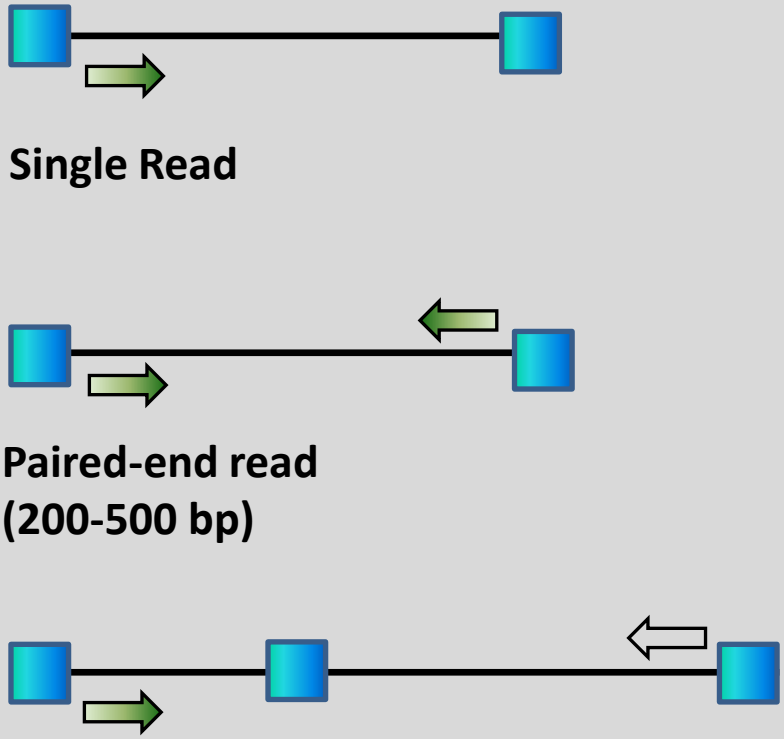
-Unknown orientation



false overlap due to repeat

Resolving Repeats

Sequencing Types



-Sequence two ends of a fragment of known size.

-Currently fragment length (insert size) can range from 200 bps – 25,000 bps

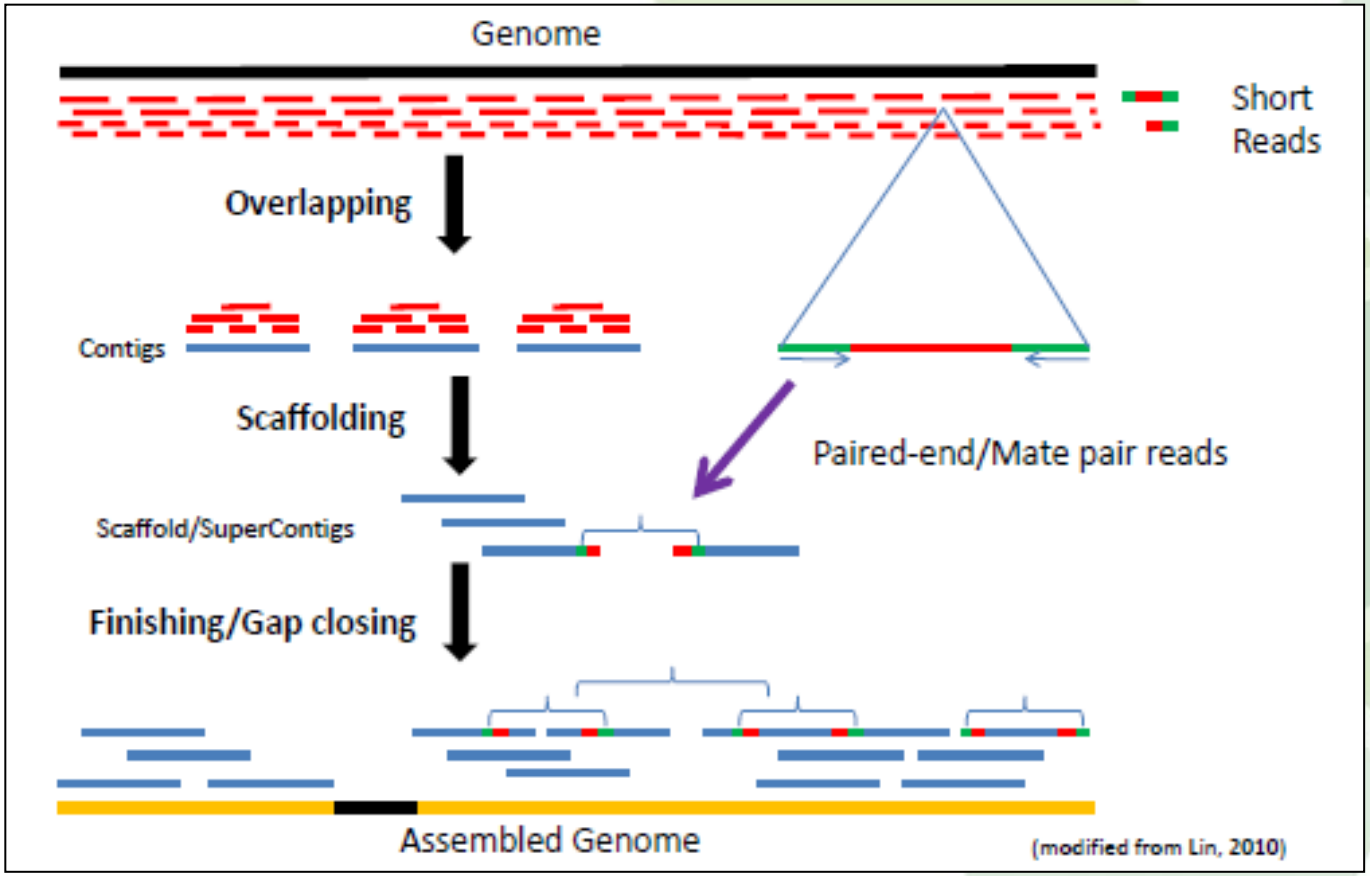
-Small scale : paired-end reads (<1Kbp)

-Medium scale: mate-pair ends, 2-25Kbp

-Large scale: fosmids/BAC libraries 40-200Kbp

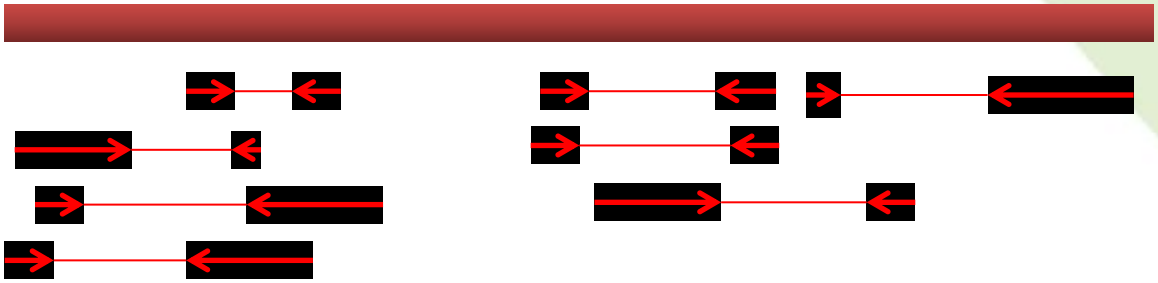
-Huge scale: Optical maps, 10-100 Mbp

Paired-end and Matepair Reads

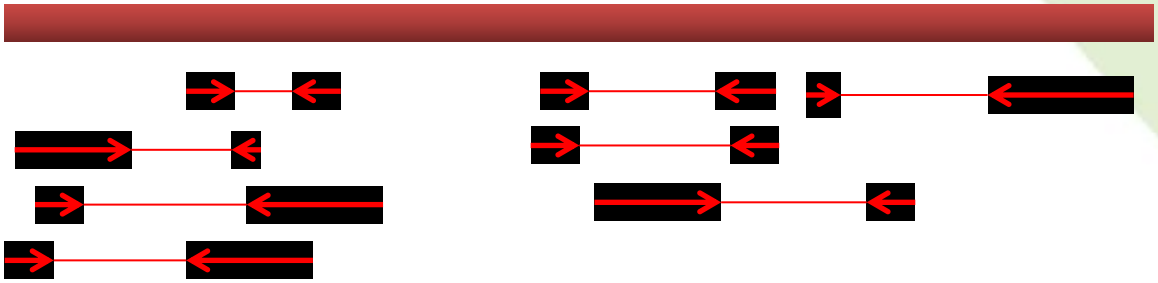


Why are Pairs so Useful?

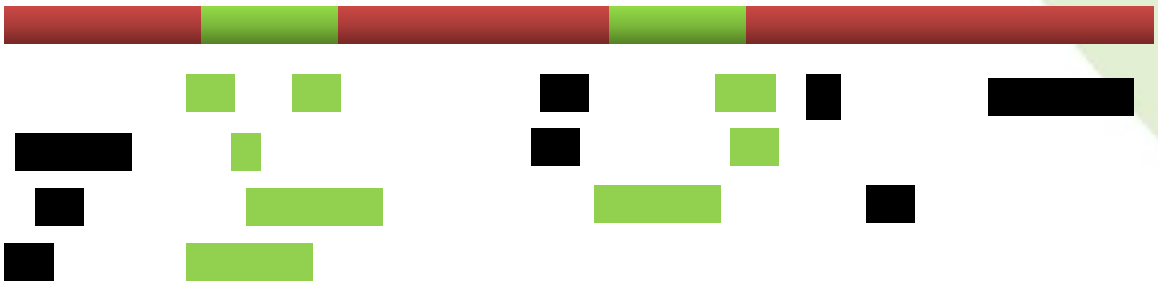
Why are Pairs so Useful?



Why are Pairs so Useful?

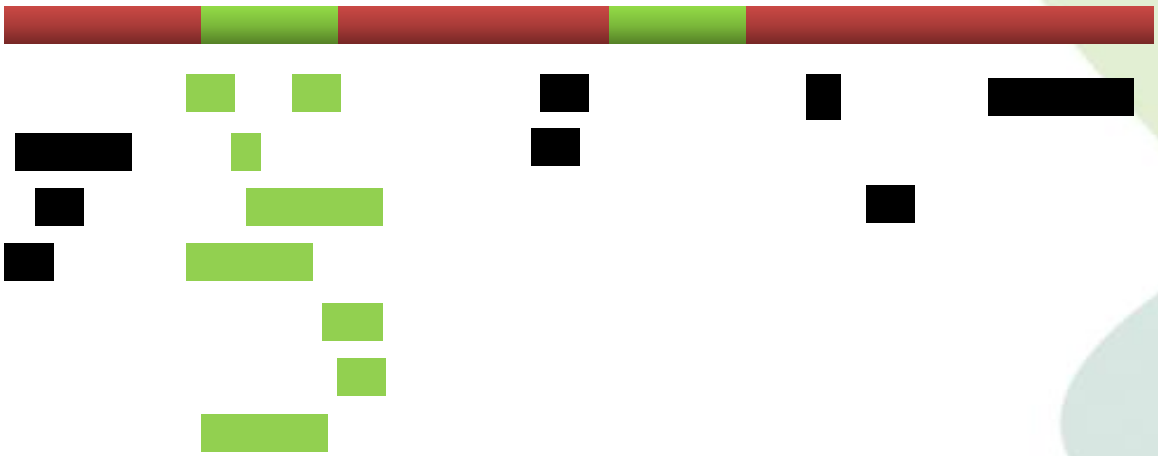


Why are Pairs so Useful?



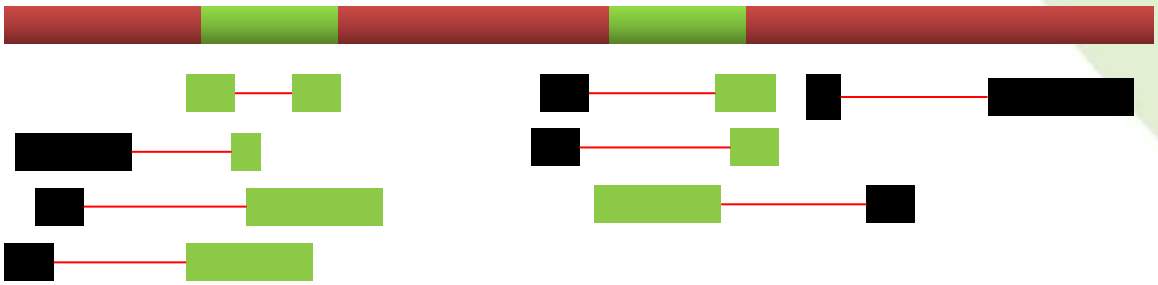
Pairs are Useful – Orientation and Separation

Why are Pairs so Useful?

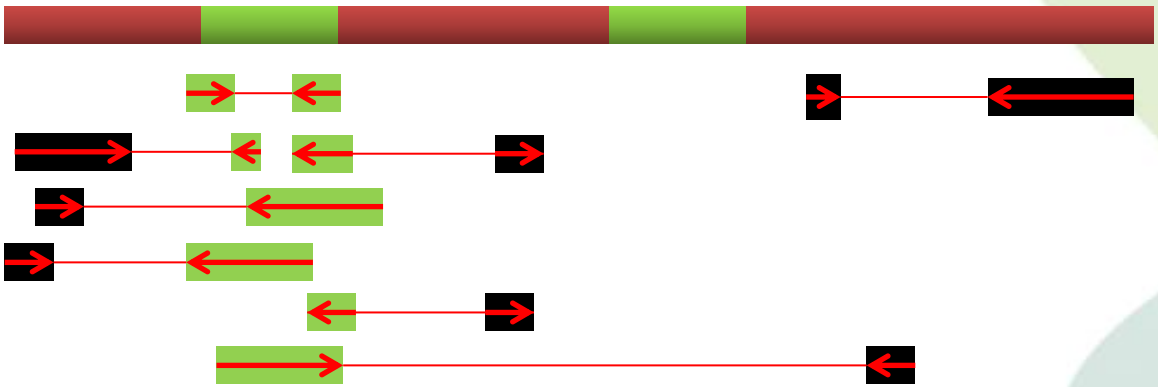


Pairs are Useful – Orientation and Separation

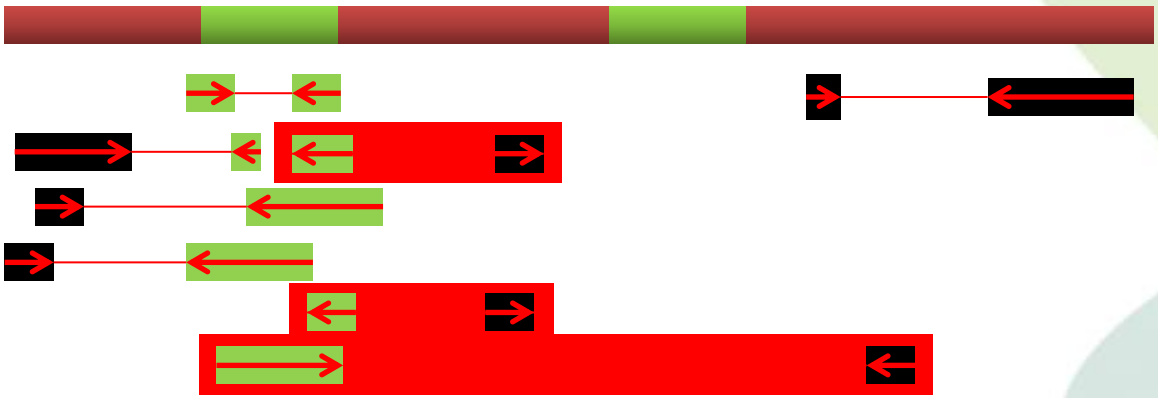
Why are Pairs so Useful?



Why are Pairs so Useful?



Why are Pairs so Useful?



Incorrect orientation

Incorrect distance

Optical mapping

Optical mapping is a technique for constructing **ordered**, genome-wide, high-resolution restriction maps from single, stained molecules of DNA

1. Cells are lysed to retrieve genomic DNA

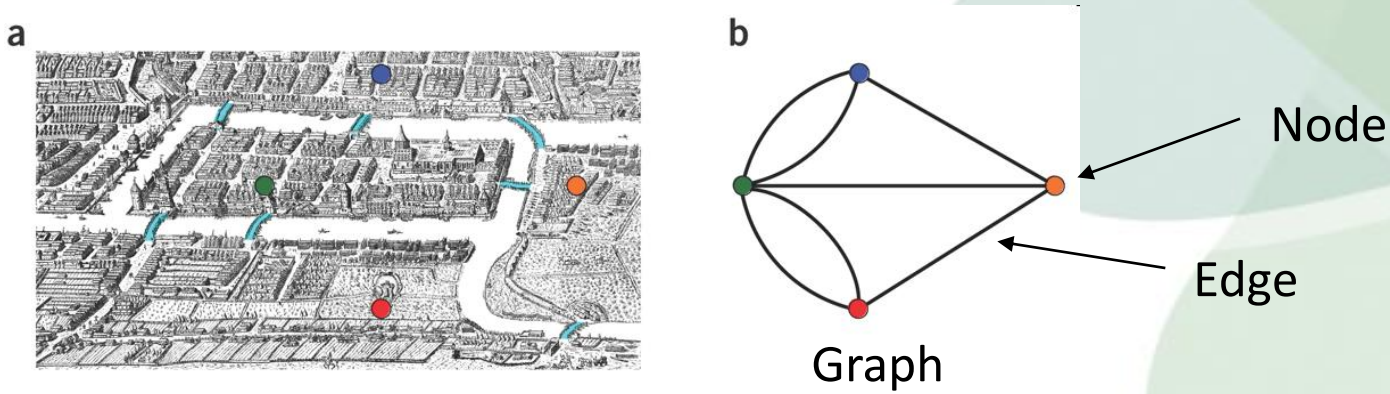
Assembly methods

Vocabulary

Node: A point, read or kmer

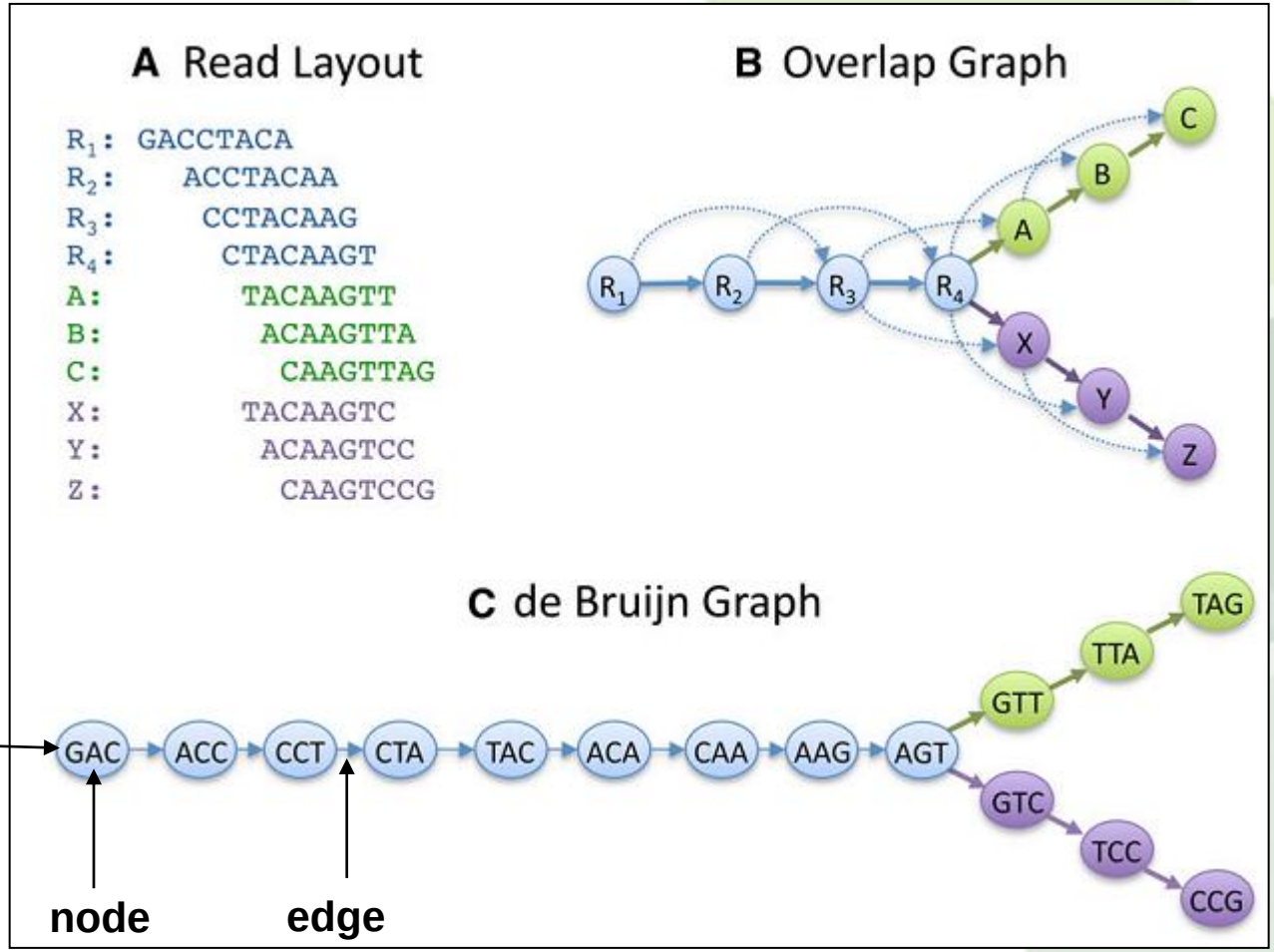
Edge: a line connecting two nodes

Graph: a network of nodes connected by edges



Compeau et al., Nature Biotechnology (2011)

Assembly algorithms



Kmer

(k=3)

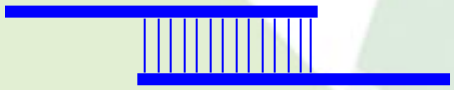
node **edge**

Schatz et al., Genome Res. (2010)

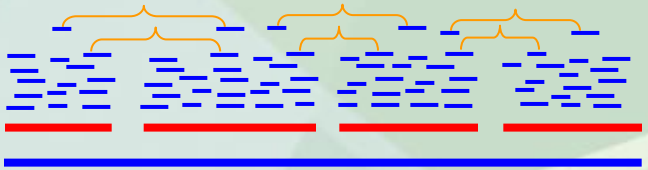
Overlap/Layout/Consensus graphs

All against all pairwise comparison

Overlap: find potentially overlapping reads



Layout: merge reads into contigs and contigs into supercontigs



Consensus: derive the DNA sequence and correct read errors

..ACGATTACAATAGGTT..

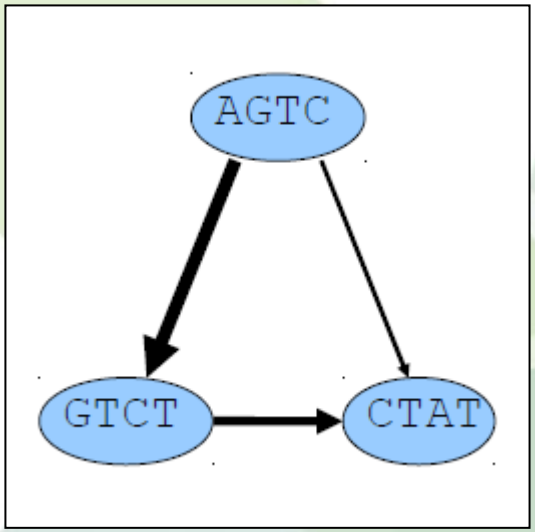
Assemblers: ARACHNE, PHRAP, CAP, TIGR, CELERA

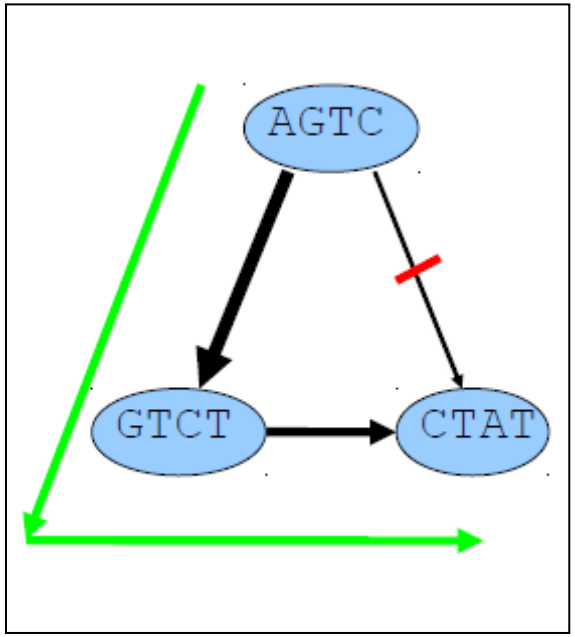
Overlap/Layout/Consensus graphs

Nodes are the 3 **read** sequences

Edges are the **overlap** alignment with orientation

Edge thickness represents score of overlap





a**GTCTCT**at

Optimal path shown in green

Un-traversed weak overlap in red

Consensus is read by outputting the overlapped nodes along the path

de Bruijn graphs

No need for all against all overlap discovery

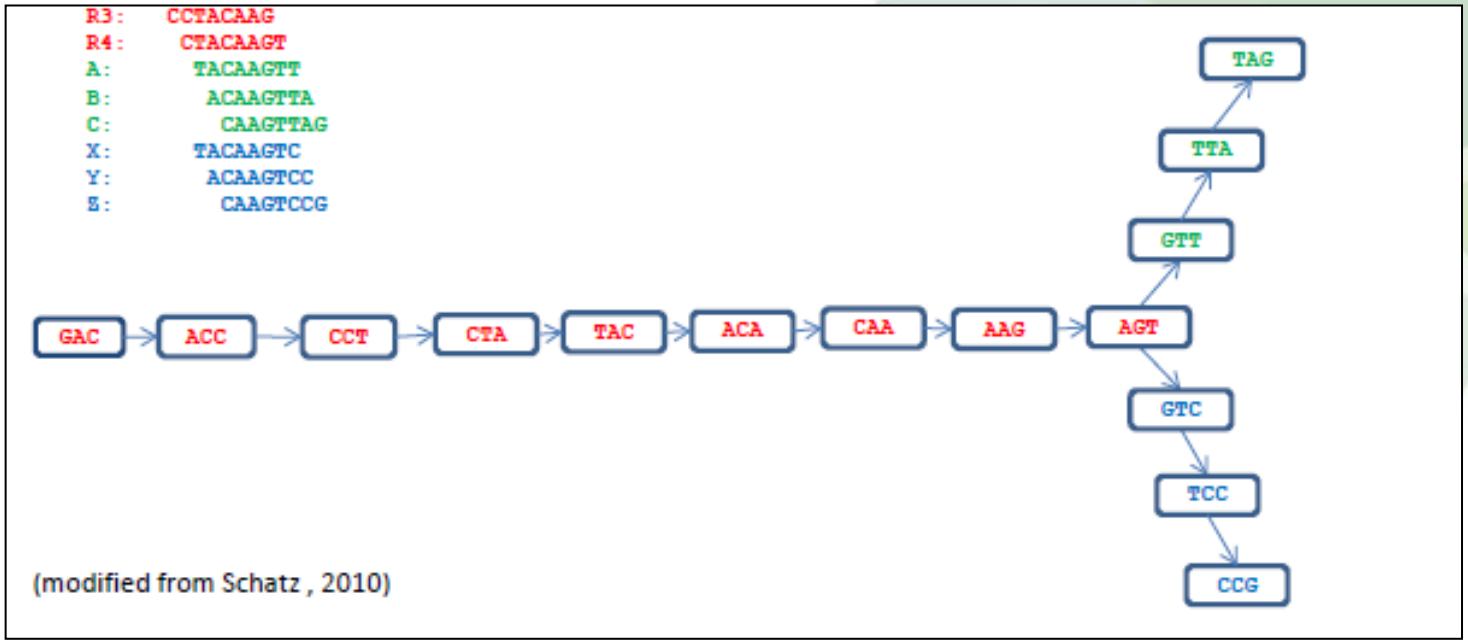
Break reads into smaller sequences of DNA (K-mers, K denotes the length in bases of these sequences).

Captures overlaps of length K-1 between these K-mers

More sensitive to repeats and sequencing errors

By construction, the graph contains a path corresponding to the original sequence.

Example assemblers: Euler, **Velvet**, ABySS, AllPaths, SOAPdenovo, CLC Bio



(modified from Schatz , 2010)

de Bruijn graphs

Step 1:
Convert reads into “K-mers”

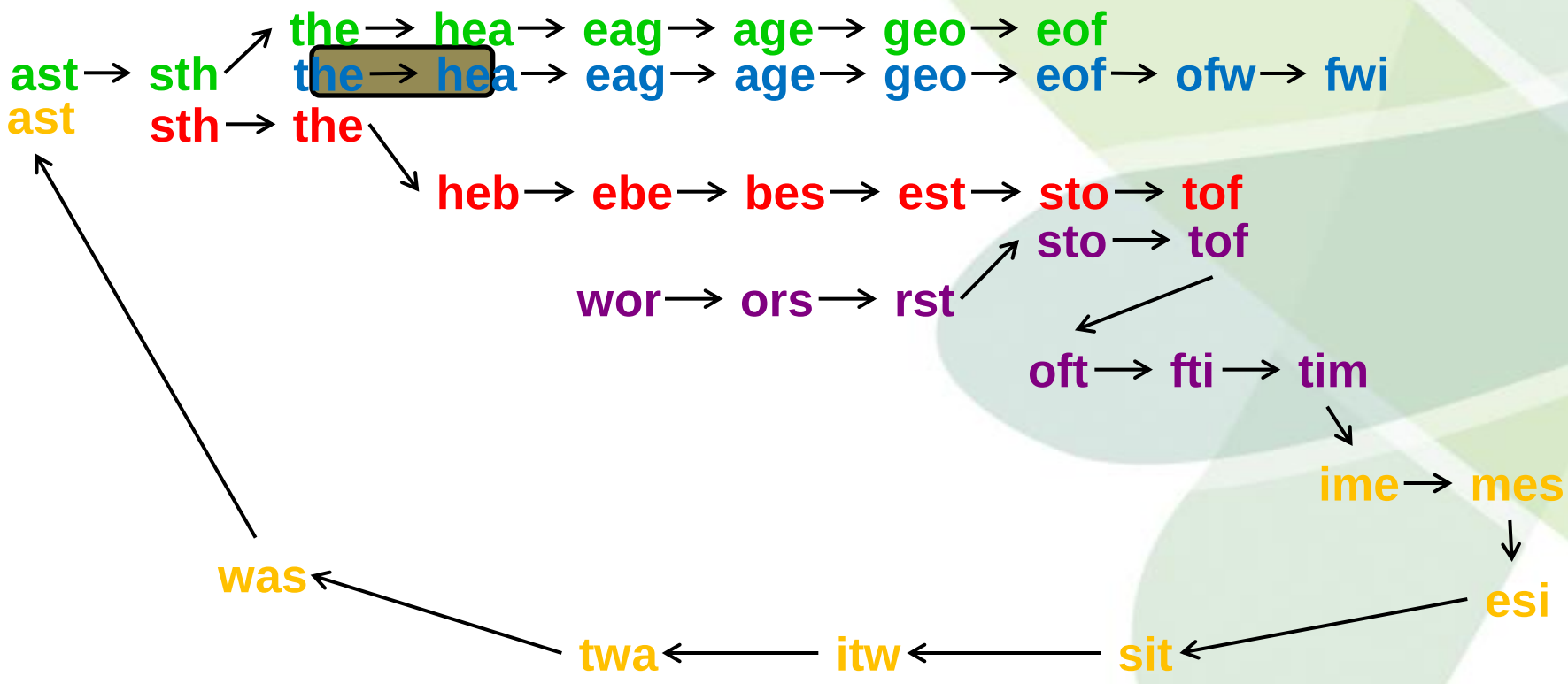
K-mer: a substring of defined length

Reads:	theageofwi	sthebestof	astheageof	worstoftim	imesitwast
K-mers : (k=3)	the	sth	ast	wor	ime
	hea	the	sth	ors	mes
	eag	heb	the	rst	esi
	age	ebe	hea	sto	sit
	geo	bes	eag	tof	itw
	eof	est	age	oft	twa
	ofw	sto	geo	fti	was
	fwi	tof	eof	tim	ast

…..etc for all reads in the dataset

de Bruijn graphs

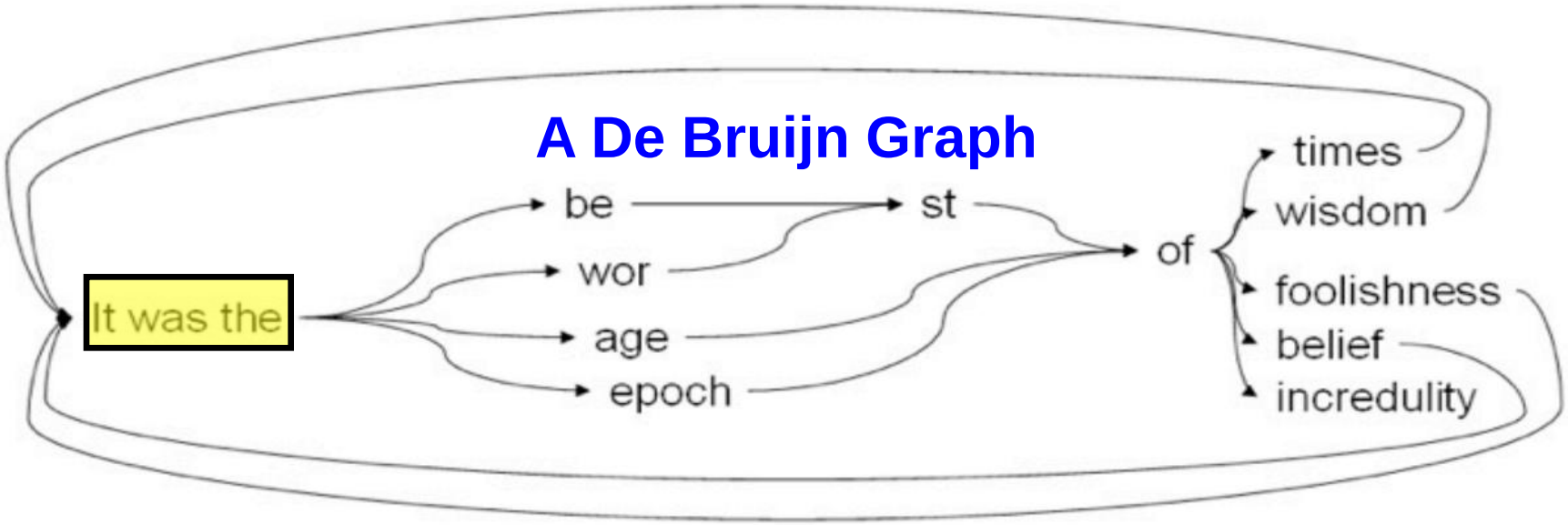
Step 2:
Build a De-Bruijn graph from the kmers



…..etc for all 'kmers' in the dataset

de Bruijn graphs

Step 3:
Simplify the graph as much as possible:

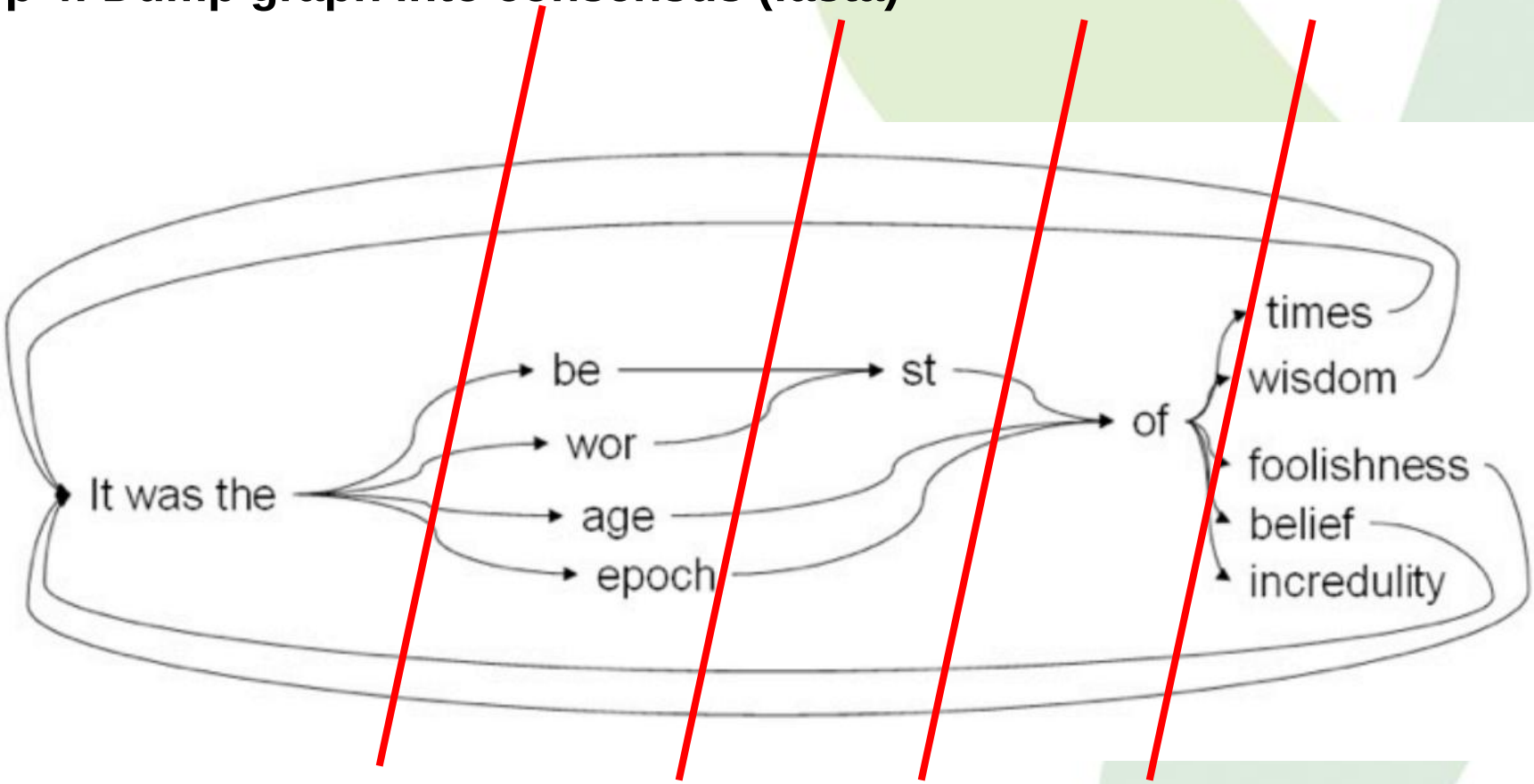


De Bruijn assemblies 'broken' by repeats longer than kmer

'It was the best of times, It was the worst of times, It was the age of wisdom, It was the age of foolishness, It was the epoch of belief, It was the epoch of incredulity,.... "

de Bruijn graphs

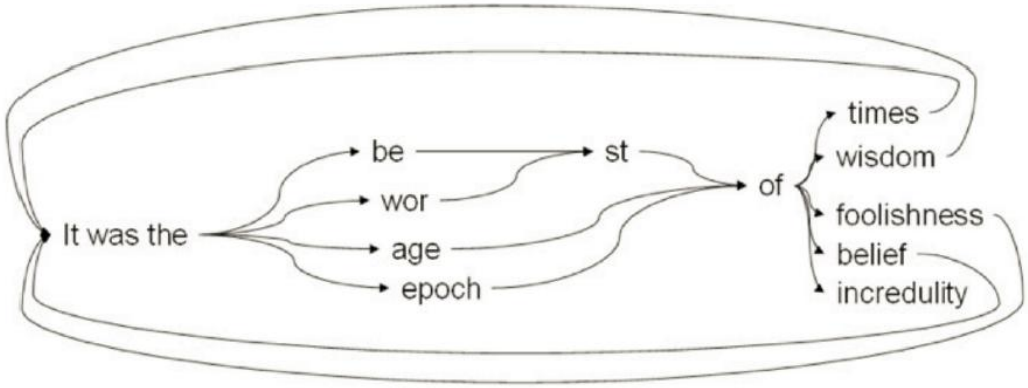
Step 4: Dump graph into consensus (fasta)



No single solution!

Break graph to produce final assembly

de Bruijn graphs



=

Contigs

It was the

be

wor

age

epoch

st

times

wisdom

foolishness

of

incredulity

belief

de Bruijn graphs

k-mer size

- Need to choose a “k” the k-mer size
  - Must be odd (avoids palindrome issues)
  - Must be less than or equal to read length
- Small “k”
  - Increase connectivity
  - More ambiguous repeats
- Large “k”
  - Increase specificity
  - Decrease connectivity

Determine k considering “**sensitivity**” and “**specificity**”

OLC vs DBG

DBG

- More sensitive to repeats and read errors
- Graph converges at repeats of length k
- One read error introduces k false nodes
- Parameters: kmer_size cov_cutoff ...

OLC

- Less sensitive to repeats and read errors
- Graph construction more demanding
- Doesn't scale to voluminous short reads
- Parameters: minOverlapLen %id ...

Overview of Tested Assemblers

Algorithm	Description	Strength	Genomes Assembled
Velvet	De Bruijn graph based Error corrections after graph is built	Fast (~30 mins) Easy to use Larger supercontig N50	Bacterial (Ref. 1; this technical note)
SOAPdenovo	De Bruijn graph based Error correction before graph is built	Easy to use Multi-threaded mode	Panda, Bacterial (Ref. 11; this technical note)
ABySS	De Bruijn graph based Can be run in parallel Distributed memory model (efficient)	Easy to use Largest contigs/scaffolds Best suited for large genomes	Human (Ref. 3; this technical note)
Forge	Overlap-layout-consensus method Modifications to accommodate Illumina reads	Largest contigs/supercontigs Good "long read" assembler	Bacterial (this technical note)

Assembly Parameters

- Number of contigs/scaffolds
  - Fewer is better, one is ideal

- Contig sizes
  - Maximum, average, median, “N50”

- Total size
  - Should be close to expected genome size
  - Repeats may only be counted once

- Number of “N”s
  - N is the ambiguous base, fewer is better

Assembly Parameters

The "N50" metric

The N50 is the most widely reported metric for *de novo* assemblies

The N50 of a set of contigs is the size of the largest contig for which half the total size is contained in that contigs and those larger.

Example:

7 contigs totalling 20 units: 7, 4, 3, 2, 2, 1, 1

N50 is 4, as 7+4=11, which is > 50% of 20

It can be calculated from the length of the original genome (*G*), the number of reads(*N*), and the average read length(*L*) as

Assembly Parameters

Coverage The percentage of bases in the reference covered by the assembled contigs

It can be calculated from the length of the original genome (*G*), the number of reads(*N*), and the average read length(*L*) as N x L/G

Example: hypothetical genome with 2,000 base pairs reconstructed from 8 reads with an average length of 500 nucleotides will have 2x coverage

Effect of coverage on assembly quality

Coverage	N50 contig size	Largest contig	Genome coverage
320×	95,313 bp	215,645 bp	99.47%
160×	95,368 bp	209,234 bp	99.72%
50×	97,333 bp	223,793 bp	99.72%
21×	35,828 bp	119,071 bp	99.38%

Contig sizes remain stable at higher coverage

It can be calculated from the length of the original genome (G), the number of reads(N), and the average read length(L) as

Assembly Parameters

Read Length

Contig sizes for each assembly (Brujin graph – Velvet – kmers 31)

Sample	N50 contig size	Largest contig	Genome coverage
E. coli, 100 bp pe	132,786 bp	326,886 bp	99.87 %
E. coli, 400 bp sr	22,902 bp	127,976 bp	99.87 %
Chr. 20, 100 bp pe	70,744 bp	484,312 bp	92.69 %
Chr. 20, 400 bp sr	2,319 bp	22,823 bp	92.65 %

Paired-end reads (100 bp)

Sample (100 bp reads)	N50 contig size	Largest contig	Genome coverage
E. coli, paired-end	132,786 bp	326,886 bp	99.87 %
E. coli, single read	23,326 bp	127,976 bp	99.87 %
Chr. 20, paired-end	70,744 bp	484,312 bp	92.69 %
Chr. 20, single read	2,320 bp	22,823 bp	92.43 %

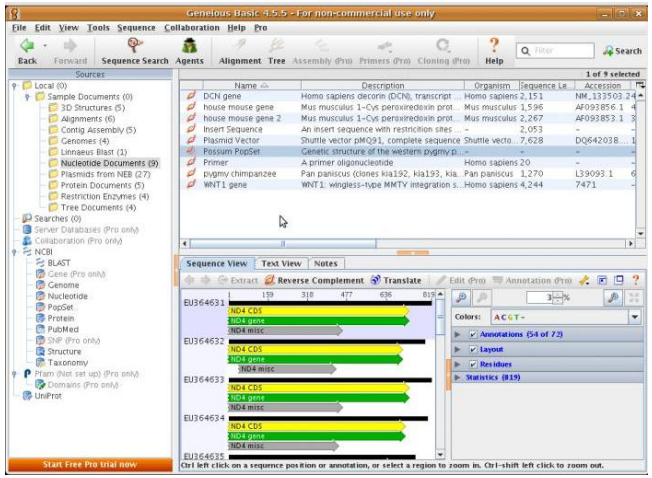
Assembly quality decreases strongly when not using paired-ends

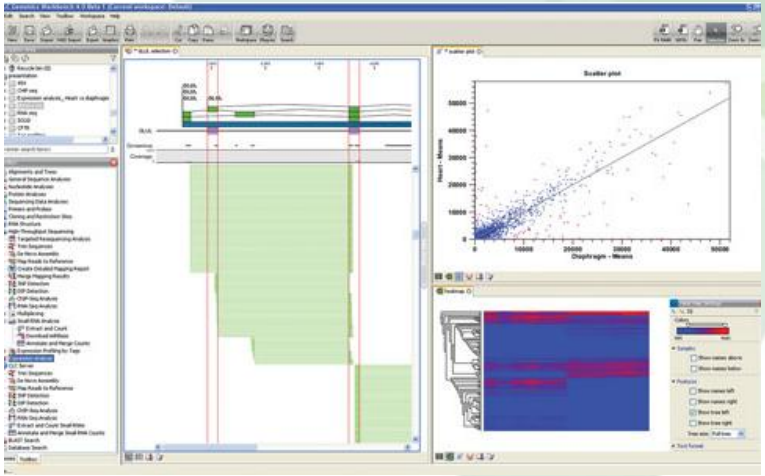
It can be calculated from the length of the original genome (G), the number of reads(N), and the average read lenght(L) as

Genome finishing

After Assembly

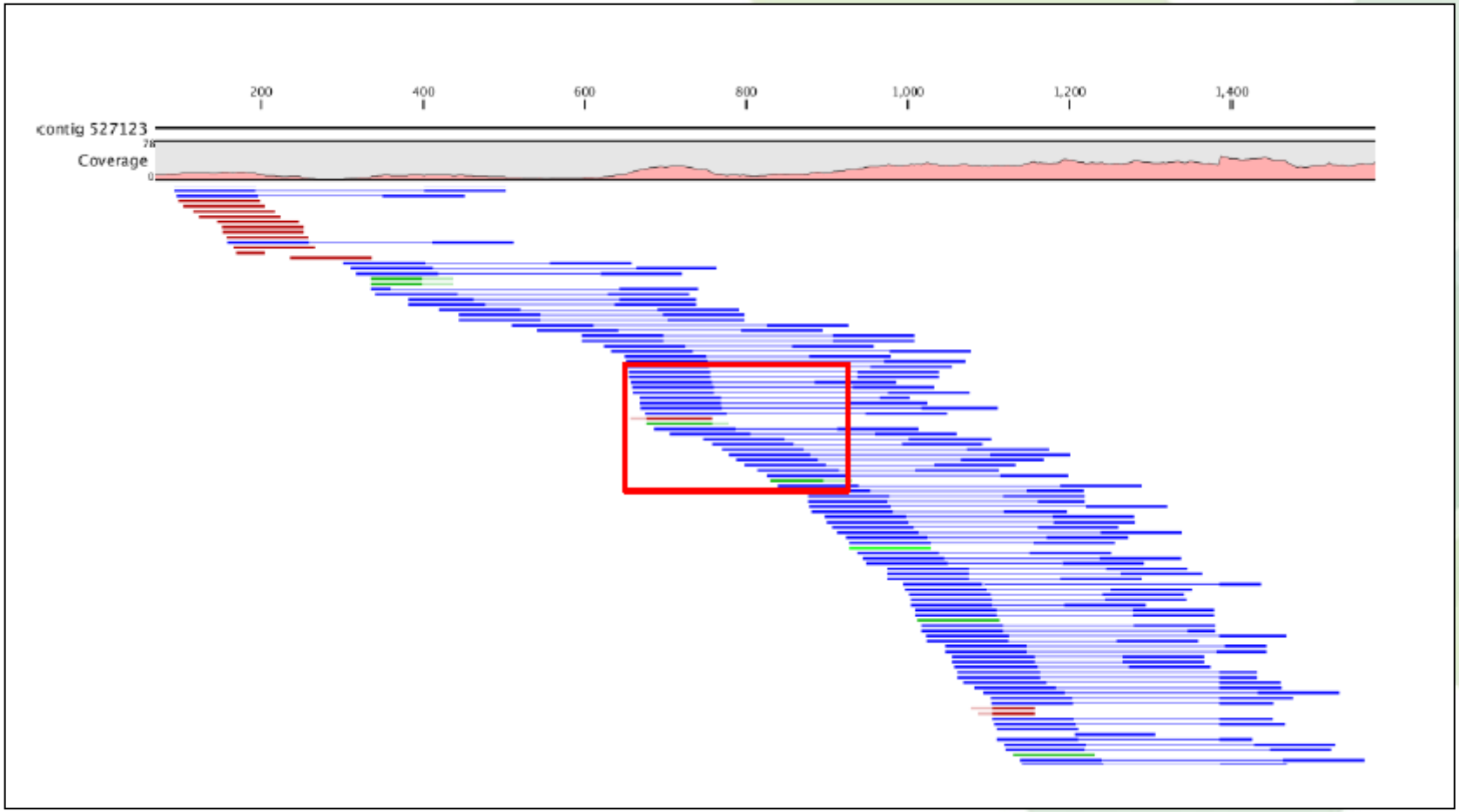
User-friendly software for analyze, compare and visualize NGS data





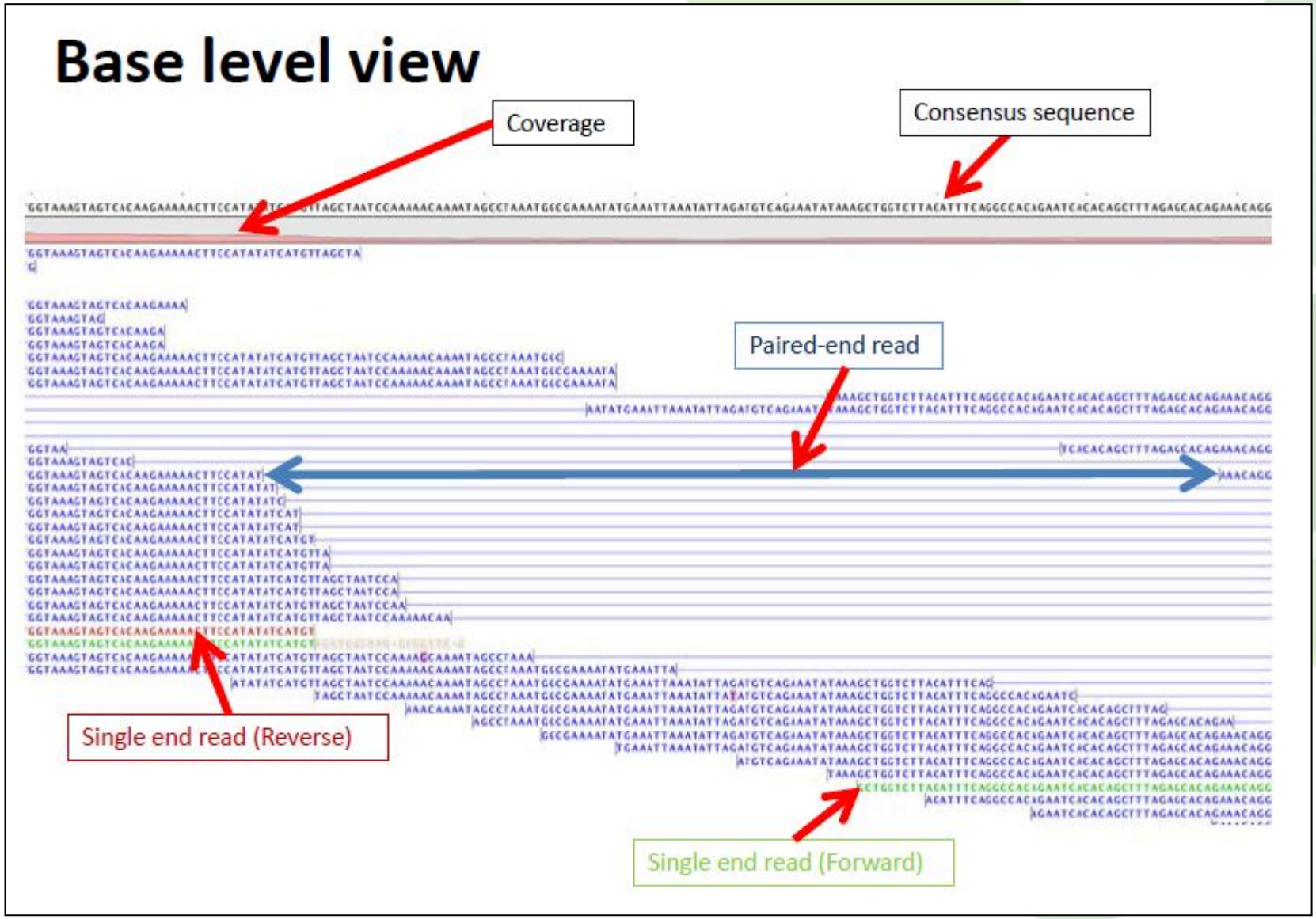
It can be calculated from the length of the original genome (G), the number of reads(N), and the average read lenght(L) as

Genome finishing



It can be calculated from the length of the original genome (G), the number of reads(N), and the average read length(L) as

Genome finishing



It can be calculated from the length of the original genome (G), the number of reads(N), and the average read length(L) as

Genome finishing

Genome finishing : aims

Produce a single "closed" DNA sequence
– No gaps or ambiguous bases (only A,G,T,C)
– No true contigs excluded

Possible?
– Yes, for bacteria and virus
– Troublesome, for larger genomes

Necessary?
– Unfinished draft genomes still very useful
– Advantage is simpler analysis, global structure

Genome finishing: methods

Close gaps (runs of Ns)
– Design custom oligos each side of Ns
– Get PCR product (hopefully only one band)
– Sanger sequence the product

Join contigs/scaffolds
– Primer walking to span long repeats
– Try out oligo pair combinations

It can be calculated from the length of the original genome (G), the number of reads(N), and the average read length(L) as .

How to close *de novo* a genome

454 mate-pair (. plate, 3kbp insert)
Good number of scaffolds and orphan contigs

Illumina paired-end (. lane, 200bp insert)
Correct homopolymer errors in 454 contigs
Extra sequence missed by 454

Optical map
Order and orient scaffolds

Finishing PCRs
Fill gaps, join contigs

Contents

Genome annotation

- **Open Reading Frame (ORF)**

  A stretch of DNA sequence with no stop codon

- **Coding Sequences (CDS)**

  DNA sequence with initiation codon and stop codon in the same frame

- **Protein encoding gene (PEG)**

  An ORF that could encode a protein

- **Hypothetical protein = putative protein**

  Something that has not been experimentally shown

Genome annotation

- Annotation is the addition of information about the predicted sequence features to the flat file of DNA code
- Identification of potential coding sequences – CDS
  -Gene prediction software **GLIMMER** (no rRNA and tRNA)
- Homology searches to predict function
- Other features can be annotated as well
  - rRNAs
  - Potential promoters
  - tRNAs
  - Small non-coding RNAs
  - Repeat sequences
  - Insertion sequences (ISs), transposons, gene fragments
- Location of the origin of replication
- Determination of the number of bases, genes, and G+C%.

How to go from this….?

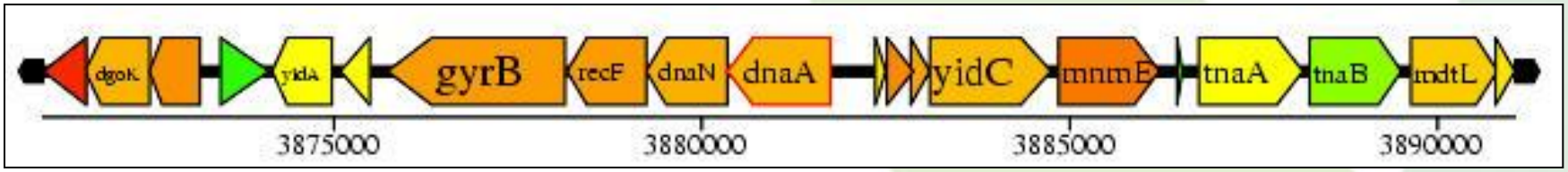
>Escherichia coli K-12 MG1655 3870656-3890655

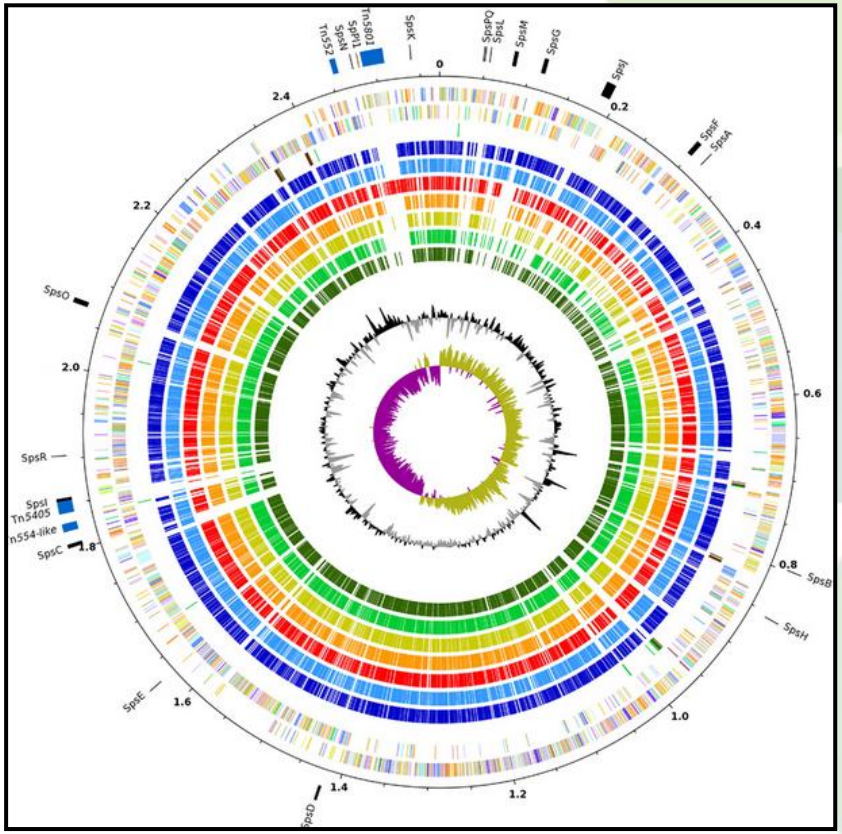
TGCTGCTGCCTGCTGCGCGGTGCGCTCTACGGATTGCCCGGCGCGATAGAGATCGCTGCCTAAGCCCGCCCCTGCACAACCTGCGTCTATCCACTGCGCCAGGTTTTCTGCGTCACGCCGCAACGGC
AAAGACTGCGATGTCCGATGGCAATACCGCTTTTAACGCTTTGATGTATTGCGGACCAAAAGCCGATGACGGAAATATTTTCAGCGCCTGCGGCGCCCGCTTCGAGCGCGGTAAAGGCTTCGGTCGC
CGTCGCGCAGCCGGGGCAGACGTCATGCCGTAGCCCACCGCACGGCGGATCACTTCACTATGGATATTGGGCGTAACGATGAGCTGACAGCCCATCCTGGCGAGCGCATCGACCTGTTCAGGTTTCA
GTACCGTACCTGCGCCAATCAACGCCTTGTCGCCGTACGCATCAACGATGCGGGAATGCTTTGCTCCCATTGTGGGGAATTCAGCGGGATTTCAACCGCGTCGAACCCGGCGTCAATCACCGCGCCA
ACATGCGCCAGCGCCTCGTCGGGCGTAATACCCGCGCAAAATGGCGATCAGCGGGAGTTTAGTTTGCCACTGCATGAGGATGCTCCTTATACCAGCCTGAAATGCCGTGTCGCCCGCCACCGCCGTCA
CGTCGCAACCCATCGCCTGAAAGGCTTGCTGGTAGCGCGCGGTCAGCGATGTTCCGGCGACAAGGGTGATGGCGTGTTGATGGGCCACATAGTCGCGCATACTGGGACCTCTGCGCCAATCAACAAA
CCAGAGAGAAATTCGCTGACCTGTTCGCGGGGAAGTGTTCCCAGCACATGCGAGGCGCGAACTTCAAAAAGCTGCGGCAATATGGCGGGCGTATTAAGACCACGCTCAAGGCCAGCTGTGAAGGCAT
CGGCAGGTTTTCCTGCGGCGGCAAACCTGCGCCAATCAATGAGTGATTTAACAGTAAATGATGTAATTCACCGGTCATCACGGTGCGAAAATCGTTGATTTGCTGGCTATCGGCCTGCACCCATTTG
CAATGGGTTCCGGGCATGACATAAAGAGAGGAAGAGCCAGAGCTCGCGCGCCGATCAATTGTGTTTCTTCGCCGCGCATCACATTGTGGTTATCGTCATGAGAGACACATAATCCGGGAATAATCCA
GATATTGTCGCCAACTGACGTTAATTGTTCGCCAATAGACGAAAAACAGGCAGGAACAGATAATACGGTGCAACTTTCCAGCCGACGTTGCTGCCAACCATTCCTGCCATTACCACTGGCGTTTTCT
CTTCACGCCAGTCGGTCGTGACTTCTGCTAACACCGCAGCCGGAGATTTTCCGTTCAGGCGCGTGACGCCTGCTTCTGATTGCCTGCTCTCAGGCAGTGGTCGCCCTGATAAAGCCAGGCGCGCAGA
TTGGTCGATCCCCAGTCAATTGCGATGTAGCGAGCTGTCATGTGATTTCCTTTAACCTTCGTGTCGAGCTGGCGATCATGGTAAGCGCCGCCTGCTCTGCCGCATCGCCGTCCTGATGCGTATCGCA
TCGAACAGCGCCTTATGTTCCTGGAGCGTTTGCGGCATGTTGGCCTCATCGCCCATCCAGGTTCGTTCAAAAACCGCCCGCTGCAGCGAACTGATCGCAATGCTAAGTTGCTGTAACACCGGGTTAT
GCACCGACTGCAGCACCGCTCGTGGTAGCGAATATCCGCTTCGTTAAACGCTTCGCGGTCCTGATTGTTGGCAATCATCTCGTTCAGCGCCGATTCAATCTGCGCCAGATCGCTGGAAGTCGCGCGC
TCTGCTCCCAACGGGCAATCGCCGGTTCCACCAGATTTCGCACTTCGTCATGGCACTGATAAGCCGTGGGTCGTAGTCATTTTCCAGCACCCATTGCAGTACGTCAGTGTCGAGGTAATTCCACTGG
TTACGCGGTGCCACAAACGCCCCGCGATAACGTTTCATTTCAATCAGCCGCTTCGCCATCAGCGAACGGAACACCCACGGATGATGTTGCGCGAGGTTGCAAACTCCTCACAGAGTTCCGCCTCAGC
CGGAAGCGGCGAGCCTGGCACGTATTTGCCGTGAACGATCTGTTTACCCAGCGTAATGACAATGCGATCGGTTTTTATTGAGAGTCATGGAGAGTCCTTGTGCTTGTATGTTCTTCTCTACTTTACCC
CGATCGATGCATAACGCGGCAACTTTGTAGTACCAGCGTGATGACGTTCGCGTTTGCCGTGCGTGTAATGTAGTACAAACTTATATTGTTGTACTACAATTTAGATCACAAAAAGAACAATGCATAA
AAAATGACATGCGTCGGGCAGAAATCTGAAAAGGGATATCAGGCGCTAAACAGGAGGGAAAGAAGAGATGCTTTCAACGGCTTAGCTACTCGTTTAAAGGATTAATCATGAAGTTGAATTTTAAGG
GATTTTTTAAGGCTGCCGGTTTATTCCCACTGCGCTGATGCTTTCAGGCTGTATCTCGTATGCTCTGGTTTCCCATACCGCAAAGGGTAGTTCAGGAAAGTATCAATCGCAGTCAGACACCATCACT
GGGCTATCGCAGGCAAAAGATAGTAATGGAACAAAAGGCTATGTTTTTGTAGGGGAATCGTGGATTACCTTATCACTGATGGTGCCGATGACATCGTTAAGATGCTCAATGATCCAGCACTTAACCG
GCACAATATTCAGGTTGCCGATGACGCAAGATTTGTTTTAAATGCGGGGAAAAAGAAATTTACCGGCACAATATCGCTTTACTACTACGGAATAACGAAGAAGAAAAGGCACTGGCAACGCATTATG
GTTTTGCCTGTGGTGTTCAACACTGTACCAGGTCACTGGAAAACCTAAAAGGCACAATCCATGAGAAAAATAAAAACATGGATTACTCAAAGGTGATGGCGTTCTACCATCCATTTAAGTGCGATTT
TATGAATACTATTCACCCAGAGGCATTCCGGGATGGTGTTTCCGCAGCATTACTGCCAGTGACTGTTACGCTGGACATCATTACTGCACCGCTGCAATTTCTGGTTGTATATGCAGTAAACCAATAA
TCAGTAAGCGGGCAAACCGTTTATGCTGTTTGCCCGCCCACAGATTAATTCAGCACATACTTCTCAATAGCAAACGCCACGCCATCTTCAAGGTTAGATTTGGTGACAAAGTTCGCCACTTCTTTCA
CTGAAGGAATAGCGTTATCCATCGCCACACCGACGCCTGCATATTAATCATTGCGATATCGTTTTCCTGATCGCCAATCGCCATGATTTCTTCCGGTTTAATACCTAACACGTCGGCCAGTGATTTC
ACCCCCGTACCTTTGTTAACGCGTTTATCGAGGATTTCGAGGAAGTACGGCGCACTTTTCAGCACGGTATATTCTCTTTCACTTCCTGCGGAATACGCGCGATAGCCTGGTCGAGGATGGCGGGTTC
ATCAATCATCATCACTTTCAGGAACTGGGTATTGGGGTCCATTTTCTCCGCTTCGCAGAACACCAGCGGAATGGTGGCAACGAAGGATTCATGCACCGTGTGTAGCTGATATCACGGTTGGCGGTGT
ACAGCGTGGTGCGGTCCAGGGCGTGGAAATGAGAACCGACTTCGCGAGAGAGTTTTTCCAGGAAACGATAGTCGTCATAGCTGAGAGCAGTTTGCGCCACGGTGCTACCATCAGCGGCCTTCTGTAC
CACGCGCCGTTATAAGTAATGCAGTAGTCGCCCGGCTGTTCCATATGCAGCTCTTTCAGGTAGTTGTGCACACCTGCATACGGGCGACCCGTCGTTAGCACGACATTCACGCCACGGGCGCGAGCTG
CGGCAATCGCATTTTTAACGGCGGGTGAAAGGTGTGATCGGGCAGCAGAAGGGTGCCATCCATATCGATAGCAATGAGTTTAATAGCCATGAGTTCCCCAGGTAGATTGGTTCCTGACCCATGCTAA
CGCGATTCCGCTCAAAAATCAGTACAACACCCGAGGGAAAAGGGGATGCAACGCGCGTGCGTGCTCCCTTTTTGCTTAGCGGAAGAGTTTCCCATGCCTGCGGAAAGCAGATC
GTTATTGGCTTGTGGTGACACTTCACCTTGCGGTGAGAGCGCATCAATAATCTTCGGCAATTGTTCTGCCAGTAAACTGGAAGCTGACTGGTATCCACGCCAAGTTTTTGCCCGAGATCGGACACCG
CATTTGTGCCGAGCGCCGATTCCAGTTGCTCGCCACTAACCGATTGATTGCCCTGTTGATTACTCAGCCAGGTTGAGAGAATGGCCCCTAAGCCGCCACTTTGCAGTTTTTCCACAGCACCTGAATG
CCGCCCTGCTCCTCAACCCAACTTAAAATAGCCTGATATTTCCCCGCATCGCCTTTCAGAAAGGCACCGACAACTTCATCAAAAAGCCCCATGATAATCACCTGTAAAGCGTTACGTGTTGACCCAA
AAAGTATAGATTTGCGGATGATAATTGCGGATTGCAGAAATAAAAAGGGCGGAGATGATCTCCGCCCTTTTCTTATAGCTTCTTGCCGGATGCGGCGTGAACGCCTTATCCGGCCTACAAAATCATG
AAAATTCAATACATTGCAAGATTTTCGTAGGCCTGATAAGCGTGCGCATCAGGCACGCTCGCATGGTTAGCGCCATTAAATATCGATATTCGCCGCTTTCAGGGCGTTCTCTTCAATAAACGCACGG
CGCGGTTCAACGGCGTCGCCCATCAGCGTGGTGAACAACTGGTCGGCAGCAATCGCATCTTTAACGGTAACCGCAGCATACGACGACTTTCCGGGTCCATAGTGGTTTCCCACAGCTGTTCCGGGTT
CATCTCGCCCAGACCTTTATAACGCTGGATGGAGAGGCCGCGACGGGACTCTTTCACCAGCCAGTCCAGCGCCTGCTCGAAGCTGGCTACCGGCTGACGCGCTCGCCACGTTCGATAAACGCATCTT
CTTCCAGCAAGCCACGCAGTTTCTCACCCAGCGTGCAGATACGACGATATTCGCCACCGGTGATAAACTCGTGATCCAGCGGATAGTCAGTATCCACACCGTGGGTACGCACGCGAACAATCGGCTC
AACAGGTTTTGCTCAGCATTGGTGTGAACATCAAACTTCCACTGGCTGCCGTGCTGTTCTTTGTCGTTCAGTTCGCTGACCAGCGCGTTCACCCAGCGGGTAACGGTCTGCTCATCAGAAAGGTCAG
CTTCCGTCAACGTCGGCTGATAGATAAGTCTTTCAGCATTGCTTTCGGATAACGACGCTCCATACGATTGATCATTTTCTGCGTCGCGTTGTACTCAGATACCAGTTTCTCTAACGCTTCGCCAGCC
AATGCCGGTGCACTGGCGTTGGTGTGCAGCGTTGCGCCGTCCAGCGCGATAGAGATTGGTACTGATCCATCGCTTCGTCGTCTTTAATGTACTGTTCCTGCTTGCCTTTCTTCACTTTGTACAGCGG
CGGCTGAGCGATGTAGACGTGACCGCGTTCAACGATTTCCGGCATCTGACGATAGAAGAAGGTCAACAGCAGCGTACGAATGTGGAGCCGTCGACGTCCGCATCGGTCATGATGATGATGCTGTGAT
AACGCAGTTTGTCCGGGTTGTACTCGTCACGACCGATACCACAGCCAAGCGCGGTGATAAGCGTCGCCACTTCCTGAGAAGAGAGCATCTTATCGAAGCGCGCTTTCTCGACTTGAGGATTTTACCC
TTCAGCGGCAGAATCGCCTGGTTCTTGCGGTTACGCCCCTGCTTCGCAGAGCCGCCCGCGGAGTCCCCTTCCACCAGGTACAGTTCGGAAAGCGCCGGATCGCGTTCCTGGCAGTCTGCCAGTTTGC
CCGGCAGGCCCGCAAGTCGAGCGCACCTTTACGGCGGGTCATTTCACGCGCGCGACGCGGCGCTTCACGGGCACGGGCAGCATCGATAATTTTGCCAACCACGATTTTCGCGTCGGTTGGGTTTTCC
AGCAGGTATTCTGCCAGCAGTTCGTTCATCTGCTGTTCAACGCCGATTTCACCTCAGAAGAAACCAGTTTGTCTTTGGTCTGGGAGGAGAATTTCGGGTCCGGCACTTTCACGGAAACGACCGCAAT
CAGGCCTTCACGCGCATCGTCACCGGTGGCGCTGACTTTGGCTTTTTTGCTGTAGCCTTCTTTGTCCATTAGGCGTTCAGGGTACGGGTCATCGCCGCACGGAAGCCTGCCAGGTGAGTACCGCCGT
CACGCTGCGGAATGTTGTTGGTAAAGCAGTAGATGTTTTCCTGGAAGCCATCGTCTCCACTGCAACGCCACTTCGACGCCAATACCGTCTTTTTCAGTGAGAAGTAGAAGATATTCGGGTGGATCGGC
GTTTTGTTCTTGTTCAGATATTCAACGAACGCCTTGATGCCGCCTTCATAGTGGAAGTGGTCTTCTTTGCCGTCGCGCTTGTCGCGCAGACGAATGGAAACGCCGGAGTTGAGGAACGACAACTCCG
CAGACGTTTCGCCAGAATTTCATATTCGAACTCGGTCACATTGGTGAAGGTTTCGAGGCTGGGCCGAGAACGCACCATGGTGCCGGTTTTTTCAGTCTCGCCGGTAACCGCCAGCGGGGCCTGCGGT
ACACCGTGTTCGTAGATCTGACGGTGATTTTACCCTCGCGCTGGATAACCAGCTCCAGTTTTTGCGACAGGGCGTTTACTACCGAAACACCAACGCCGTGCAGACCGCCGGACACTTTATAGGAGTT
ATCGTCAAATTTACCGCCTGCGTGCAGAACGGTCATGATCACTTCCGCCGCCGA

…to this?

```
FT gene complement(9299..10702)
FT /db_xref="GenBank:2367266”
FT /gene="dnaA”
FT /note="b3702”
FT CDS complement(9299..10702)
FT /db_xref="GI:2367267”
FT /db_xref="PID:g2367267”
FT /function="putative regulator; DNA - replication, repair,
FT restriction/modification”
FT /codon_start=1
FT /protein_id="AAC76725.1”
FT /gene="dnaA”
FT /translation="MSLSLWQQCLARLQDELPATEFSMWIRPLQAELSDNTLALYAPNR
FT FVLDWVRDKYLNNINGLLTSFCGADAPQLRFEVGTKPVTQTPQAAVTSNVAAPAQVAQT
FT QPQRAAPSTRSGWDNVPAPAEPTYRSNVNVKHTFDNFVEGKSNQLARAAARQVADNPGG
FT AYNPLFLYGGTGLGKTHLLHAVGNGIMARKPNAKVVYMHSERFVQDMVKALQNNAIEEF
FT KRYYRSVDALLIDDIQFFANKERSQEEFFHTFNALLEGNQQIILTSDRYPKEINGVEDR
FT LKSRFGWGLTVAIEPPELETRVAILMKKADENDIRLPGEVAFFIAKRLRSNVRELEGAL
FT NRVIANANFTGRAITIDFVREALRDLLALQEKLVTIDNIQKTVAEYYKIKVADLLSKRR
FT SRSVARPRQMAMALAKELTNHSLPEIGDAFGGRDHTTVLHACRKIEQLREESHDIKEDF
FT SNLIRTLSS”
FT /product="DNA biosynthesis; initiation of chromosome
FT replication; can be transcription regulator”
FT /transl_table=11
FT /note="f467; 100 pct identical to DNAA_ECOLI SW: P03004;
FT CG Site No. 851”
```

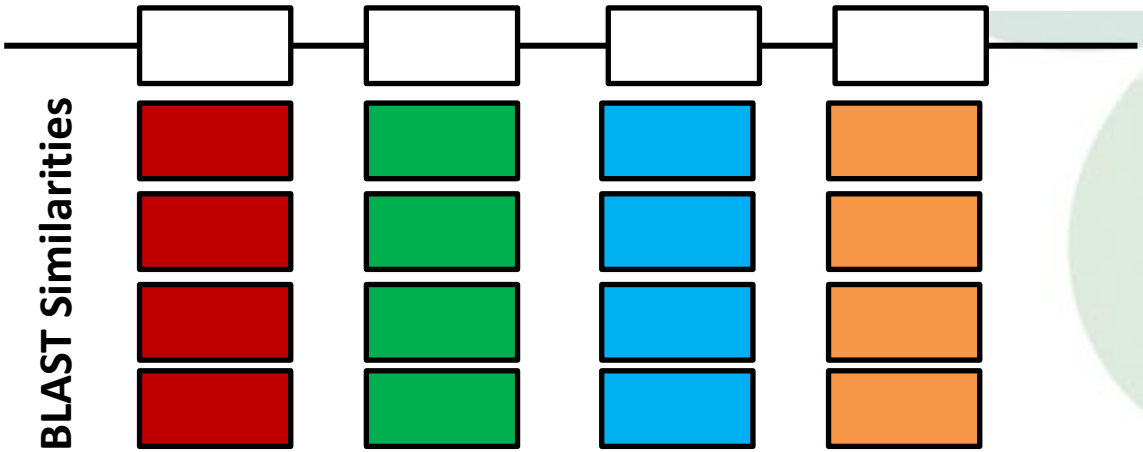
Or this?


dgoK
ykfA
gyrB
recF
dnaN
dnaA
yidC
mnmE
tnaA
tnaB
mdtL
3875000
3880000
3885000
3890000


Tn552
SpsN
SpsP1
Tn5801
SpsK
SpsQ
SpsL
SpsM
SpsG
SpsJ
SpsF
SpsA
SpsO
SpsR
SpsI
Tn5405
Tn554-like
SpsC
SpsB
SpsH
SpsE
SpsD
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4

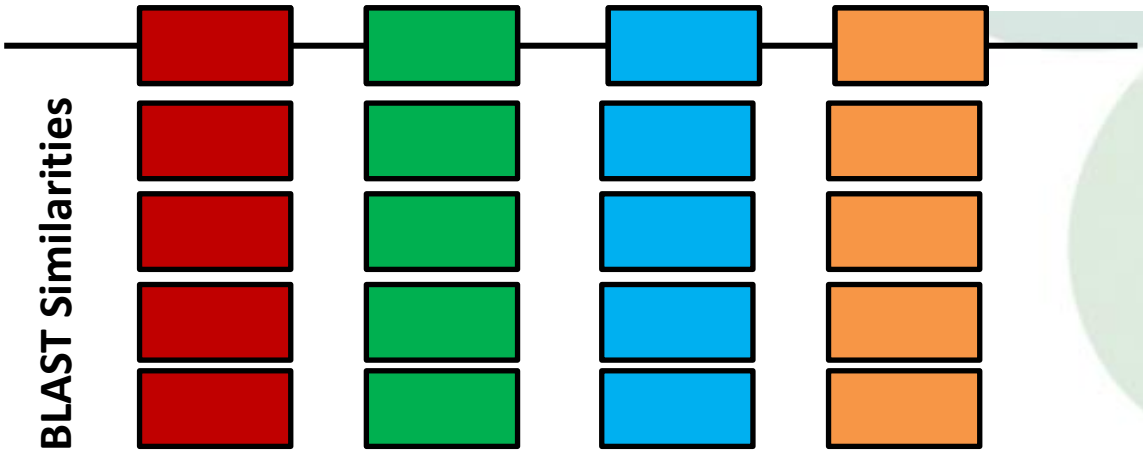
What is BLAST?

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
  - Developed in 1990, refined in 1997 (Stephen Altschul)
- **A method of searching sequence databases to find sequences similar to the input sequence**
  - Scans a database for alignments to a query sequence
- **Fastest and most frequently used sequence alignment tool**
  - the industry standard
- **Can be extremely informative, giving clues to**
  - functionality, evolutionary history, important residues
- **Basis for many forms of bioinformatic analysis**



What is BLAST?

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
  - Developed in 1990, refined in 1997 (Stephen Altschul)
- **A method of searching sequence databases to find sequences similar to the input sequence**
  - Scans a database for alignments to a query sequence
- **Fastest and most frequently used sequence alignment tool**
  - the industry standard
- **Can be extremely informative, giving clues to**
  - functionality, evolutionary history, important residues
- **Basis for many forms of bioinformatic analysis**



The several flavours of BLAST

- **BLASTP**
  - protein query versus protein sequence database.
- **BLASTN**
  - nucleotide query versus nucleotide sequence database.
- **BLASTX**
  - translated nucleotide query versus protein sequence database
- **TBLASTN**
  - protein query versus translated nucleotide sequence database
- **TBLASTX**
  - translated nucleotide query versus translated nucleotide sequence database.

Genome annotation

Once the ordered set of contigs has been obtained

Annotated the draft genome

Automated web-based tool
Manual curation of the results

Genome Annotation: A multistep process

3 general levels of annotation:

-1 Nucleotide-level (where)

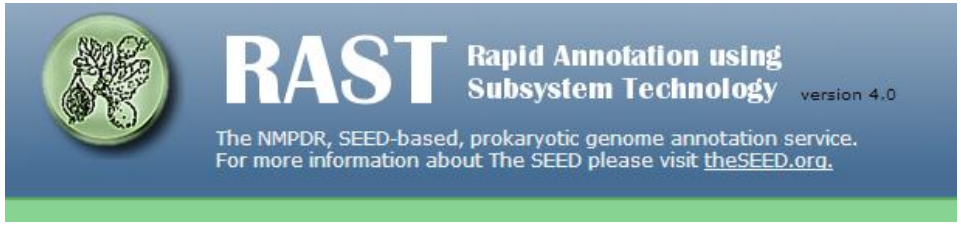
-2 Protein-level (what)

-3 Process-level (how)

Genome annotation

BASys: Bacterial Annotation System

https://www.basys.ca/



RAST: Rapid Annotation using Subsystems Technology

http://rast.nmpdr.org/

Prokka http://www.vicbioinformatics.com/software.prokka.shtml

BG7 http://bg7.ohnosequences.com/

NCBI https://www.ncbi.nlm.nih.gov/genome/annotation_prok/

Nucleotide level

- *Ab initio* gene prediction
  - By opening reading frame
    - Find ORFs
    - Find credible CDSs within ORFs
    - Resolve conflicting ORFs
  - By codon usage
  - By Markov models (Glimmer)
- By homology
  - Similarity Searches via protein or translated BLAST
  - Comparative genomics

Nucleotide level

- Search for Sequence Features
  - Promoters, Ribosome-binding Sites
  - Repeats, Inverted Repeats
  - Consensus Sequences for regulator binding site
  - Often rely on sequence motifs

- Other features

  -tRNA (tRNA scan, http://lowelab.ucsc.edu/tRNAscan-SE/)

  -rRNA (RNAmmer, http://www.cbs.dtu.dk/services/RNAmmer/)

  -ncRNA (Rfam, http://rfam.sanger.ac.uk/)

Protein level

-**Assign putative functions to proteins** of an organism

-**Classify proteins** into families:

-using similarities to better-characterized proteins of other species (**BLASTP**)

-on the basis of functional domains, motifs, and folds

-Search against protein databases of **functional domains**

-Pfam (http://pfam.sanger.ac.uk/)

-ProSite (http://prosite.expasy.org/)

-**InterProScan: integration of several protein databases**

-makes things much easier!

Process-level Annotation

-**Linking the genome to biological processes**

-**Bench work required** (e.g. microarrays, RNAi, etc.)

-Classification scheme required**: Gene Ontology (GO)**

Gene Ontology

Gene Ontology (GO) is a structured vocabulary of terms describing gene products according to

- molecular function
- biological process
- cellular component

Molecular function:
the tasks performed by individual gene products
(examples are carbohydrate binding and ATPase activity)

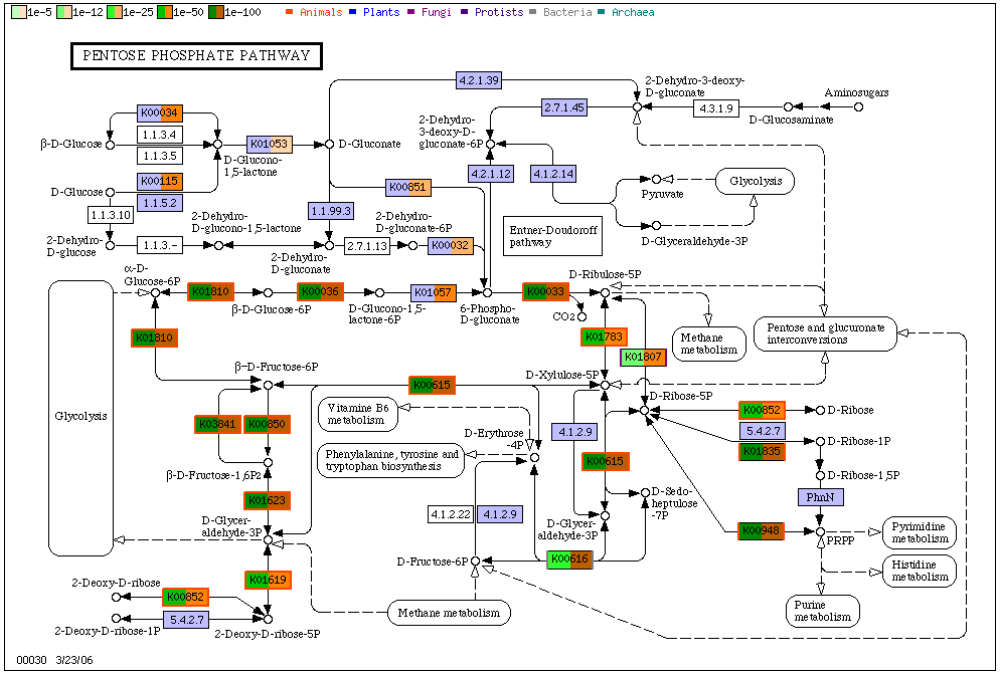
Biological Process:
biological goal or objective– broad biological goals
(such as mitosis or immune response)

Cellular Component:
Subcellular structures, locations, and macromolecular complexes
(examples include nucleus, telomere and RNA ploymerase II holoenzyme)

KEGG pathway database

KEGG:

Kyoto Encyclopedia of Genes and Genomes



The KEGG pathway maps are graphical diagrams representing knowledge on molecular interaction and reaction networks for metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases and drug development.

Manually entered from published materials

Deposit into GenBank

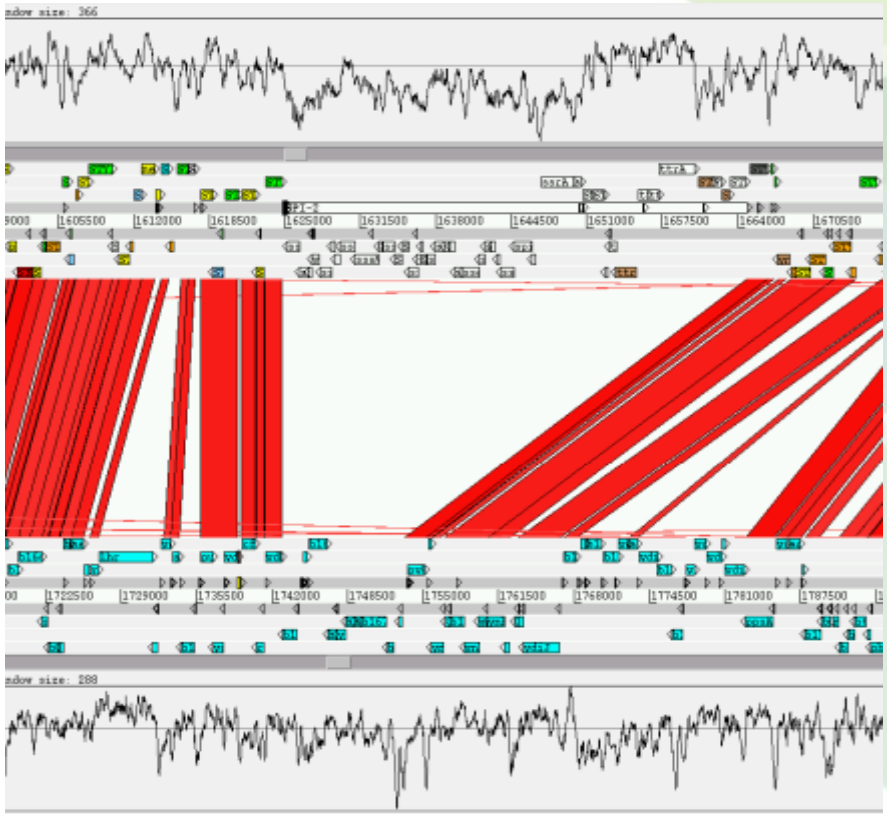
- Once analyses are complete and being written up they must be shared
- Use SequIn at NCBI to upload sequences and annotations
- All sequences must be deposited into GenBank for publication
- Annotation *not required* but essential for interpretations
- New levels defined
  - Standard draft
  - High quality draft
  - Annotation Grade
  - Non-contiguous finished
  - Finished

Contents

Artemis Comparison Tool

- **ACT** is a DNA sequence comparison viewer based on Artemis
- Can read complete EMBL and GenBank entries or sequence in FASTA or raw format
- Blast is used to compare the sequences
- ACT is a Java-based tool for visualizing pairwise comparison sequences, free software.

tRNA
phage/IS genes
Pseudogenes

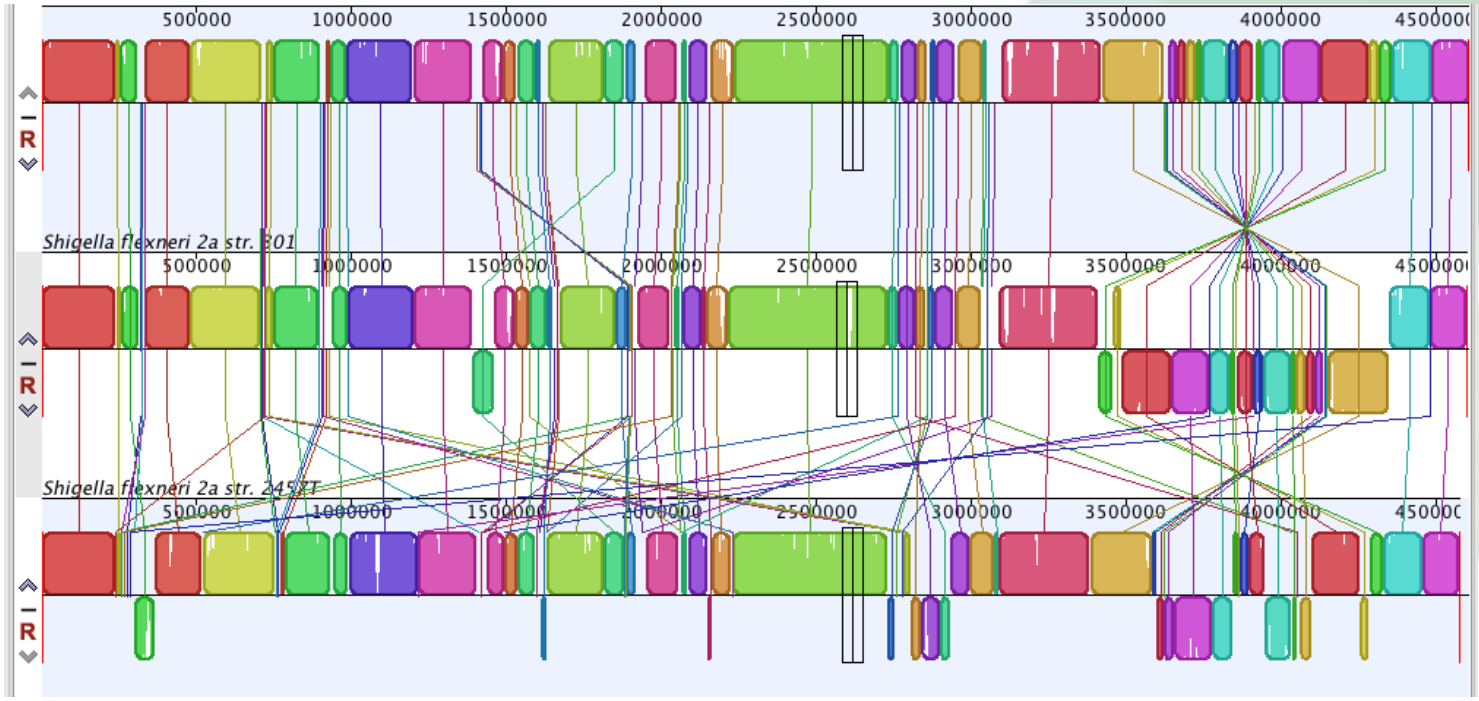


G+C

Blast hits

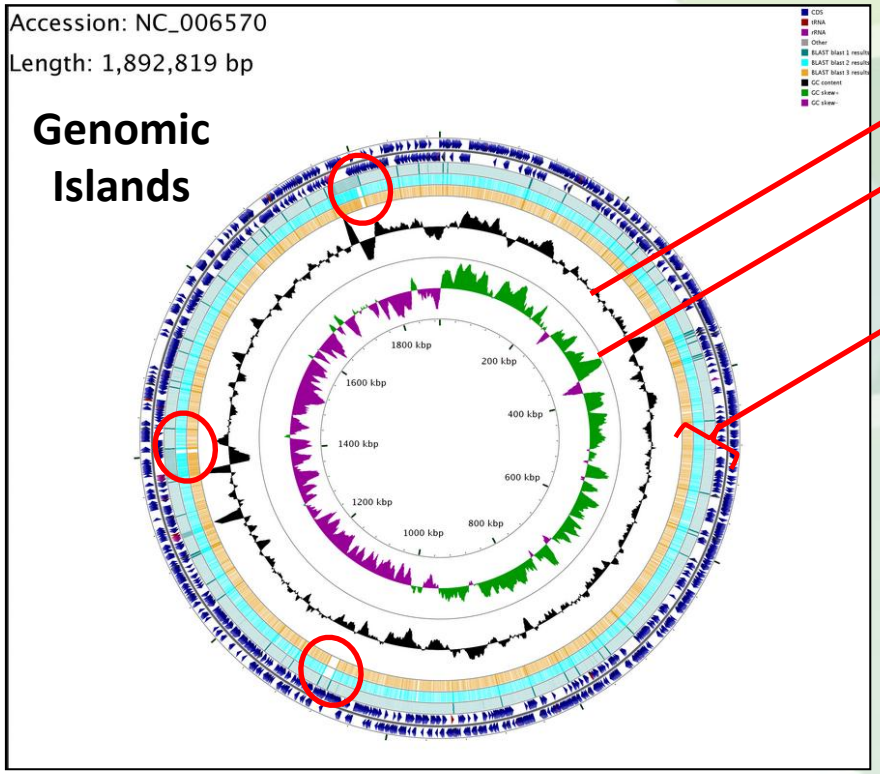
Mauve

-Java-based tool for multiple alignment of whole genomes
-Generates a multiple whole-genome alignment
-Identifies blocks of sequence homology
-Easy identify a) regions that are conserved among the whole genomes
b) regions that are uniques to subsets of genomes (islands)
c) single nucleotide polymorphism (SNPs)
d) arrangements and inversions



GC View

-Is a comparative genomics tool for circular genomes that allows sequence feature information to be visualized in the context of sequence analysis results
-Results plotted as a series of rings
-Identification **Genomic Islands**



GC content

GC skew

BLAST results

Genomic Islands

- Large chromosomal regions, part of the flexible gene pool
- Previously transferred by mobile genetic elements
- Present in some bacteria but absent in close relatives
- Carry multiple genes that increase phenotypic versatility
- Contribute to dynamic character of bacterial chromosomes and can be excised from the chromosome and transferred to other recipients

Bacterial mobile genetic elements

- **Transposons**
  - pieces of DNA that act as ‘jumping genes’ that change location on chromosome or plasmid chromosomal localization.
  - encode transposase that catalyses the transposition event
  - can carry resistance or virulence genes
- **Insertion sequences (IS elements)**
  - transposable elements that encode only the transposase
  - multiple copies of same IS within genome provide targets for homologous recombination, rearrangements and replicon fusions
- **Conjugative transposons**
  - normally integrated into the chromosome
  - excise then transferred to recipient cells by conjugation

Bacterial mobile genetic elements

- **Plasmids**
  - self-replicating extrachromosomal replicons
  - usually circular but can be linear
  - Can carry resistance or virulence genes
- **Bacteriophages**
  - bacterial viruses can carry virulence genes
  - can insert into bacterial chromosome as prophages (lysogeny)
- **Integrons**
  - complex natural cloning and gene expression systems able to capture promoter less gene cassettes by site-specific recombination
  - allow formation of large arrays of gene cassettes transferred as a whole between different replicons.


THANK YOU

Practical Groups

NAME	SURNAME	GROUP
Mariane	Schmidt	1
Blanca	Vera Gargallo	1
Steffen	Lott	1
Ricardo	Delgado Santander	1

NAME	SURNAME	GROUP
Catarina	Cúcio	5
Eduardo	Tosado Rodríguez	5
Rafael	Montalvo	5
Catalina María	Alejandro Marín	5

NAME	SURNAME	GROUP
Coralis del M.	Rodríguez García	2
Santiago	Català García	2
Javier	Miralles Lorenzo	2
Lejla	Pasic	2

NAME	SURNAME	GROUP
Mike	Winson	6
Saghaï	Aurélien	6
Miguel	Carda	6
Rafael Antonio	Rojas Herrera	6

NAME	SURNAME	GROUP
Mara F.	Cuebas Irzamy	3
Rakel	Arrazuria	3
Laura	Sanguino Casado	3
Laura	Leite	3

NAME	SURNAME	GROUP
Lucas	Stahl	7
Allali	Imane	7
Antonio	Picazo	7

NAME	SURNAME	GROUP
Gerard	Muyzer	4
Ylenia	Arizaga	4
Julliane	Medeiros	4
Elena	Gómez Sanz	4

NAME	SURNAME	GROUP
Clara	Cardoso	8
José	Moya Cuevas	8
Uljana	Hesse	8