

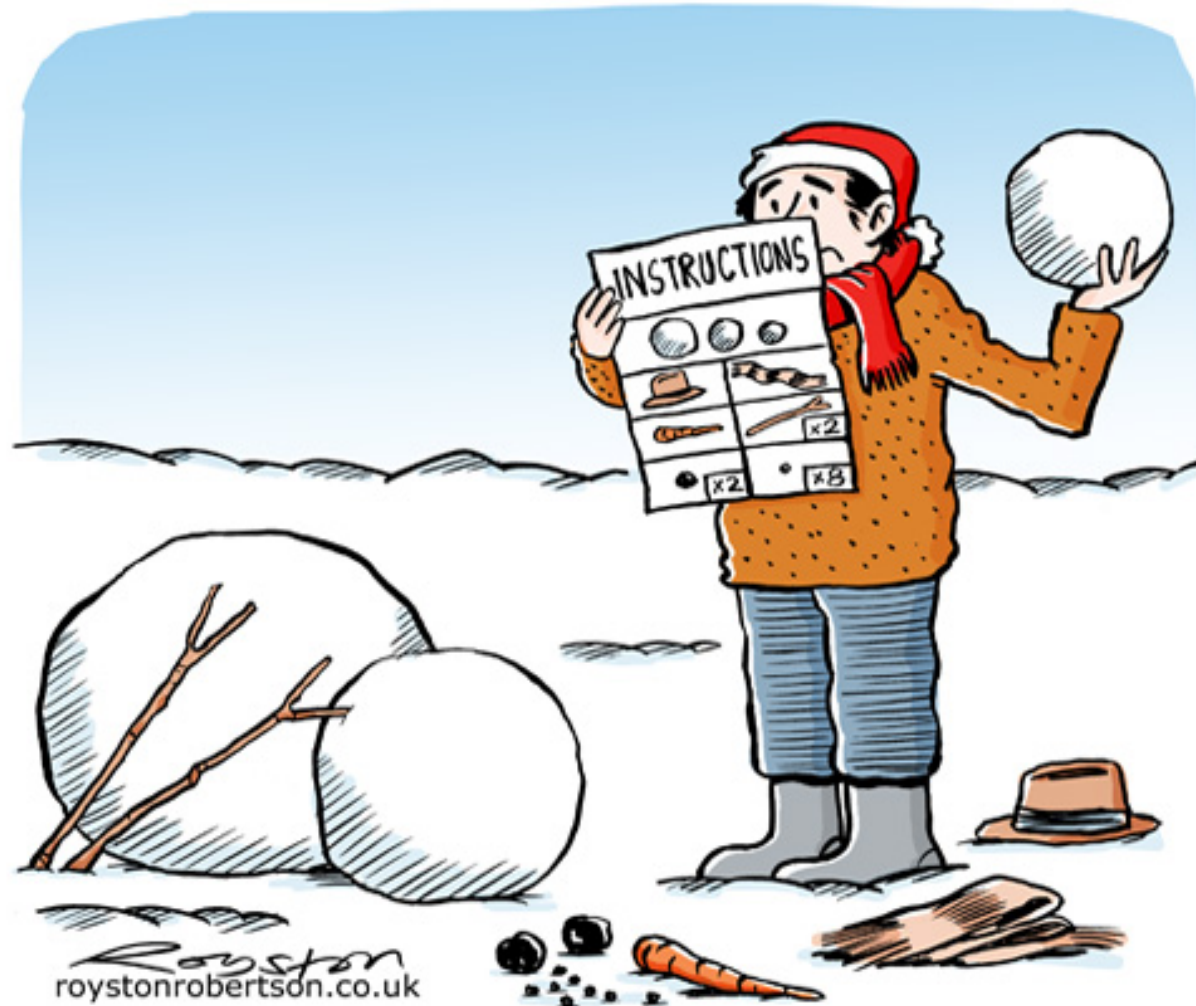
Assembly & validation

Workshop in marine metagenomics

Tromsø November 2018

Assembly is the computational reconstruction of a longer sequence from smaller sequence reads

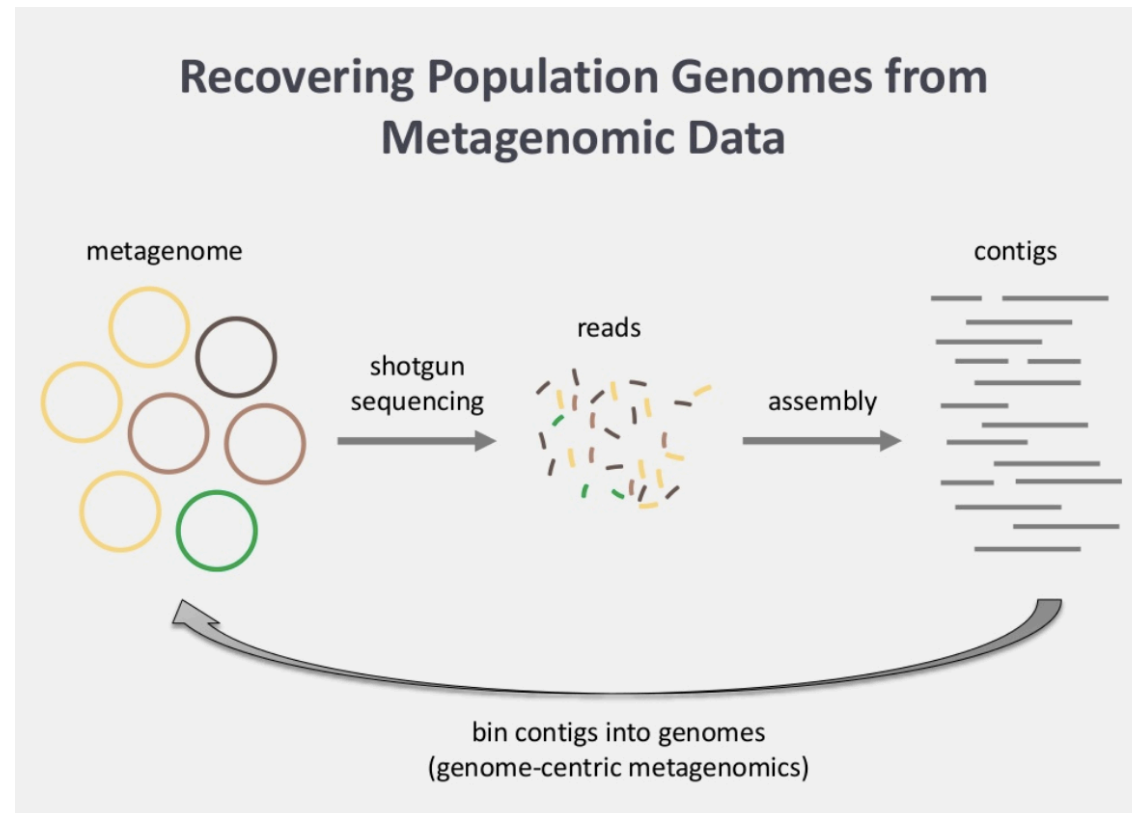
Which method should I choose that will produce the highest-quality assembly with the data that I have?



Why do we want to sequence metagenomes?

Important for understanding the biology and functional potential of hard-to-culture microorganisms

Metagenomic recovery of complete or draft microbial genomes is a starting point to analyze the “taxon-specific” potential of organisms within their community and ecosystem context



Donovan Parks, Australian school of ecogenomics

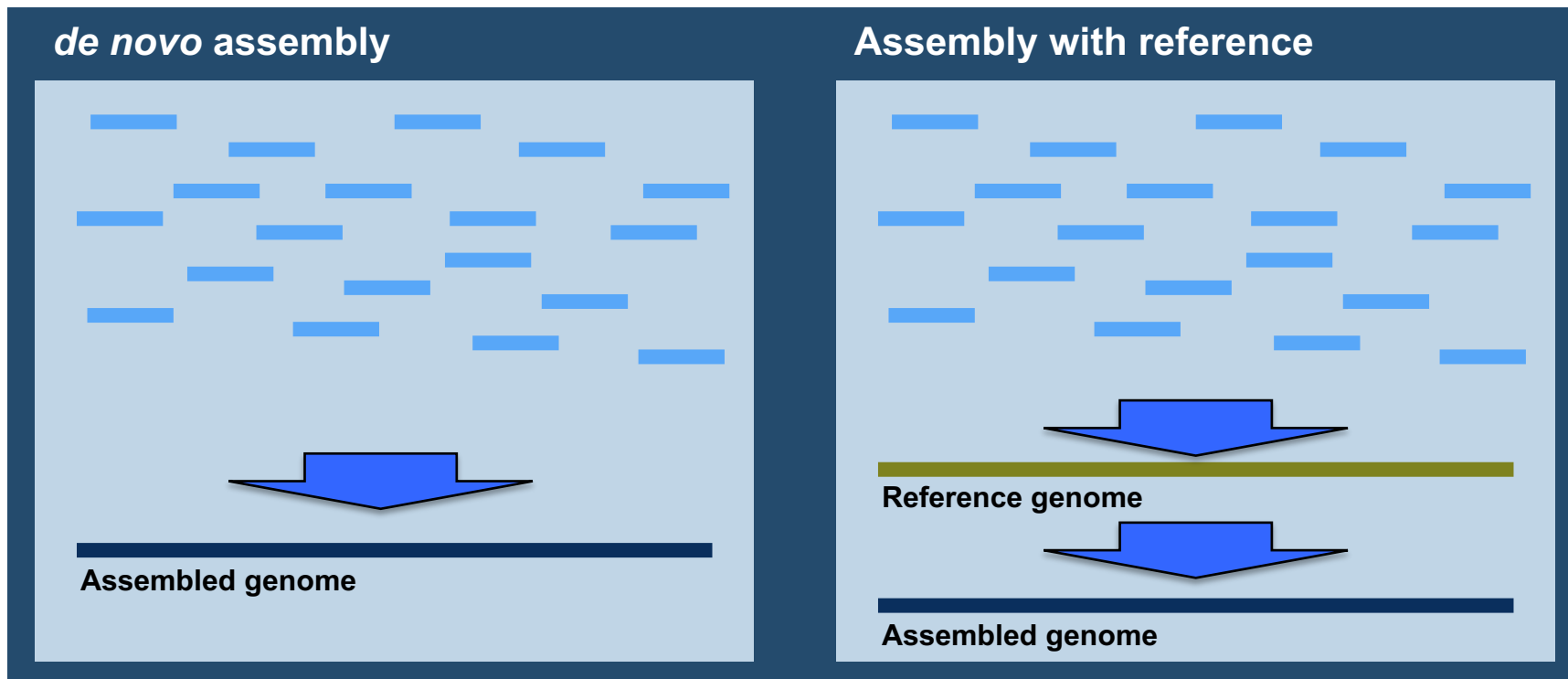
There are two approaches for sequence assembly

de novo assembly:

Reconstructing a DNA sequence with no prior knowledge of the sequence

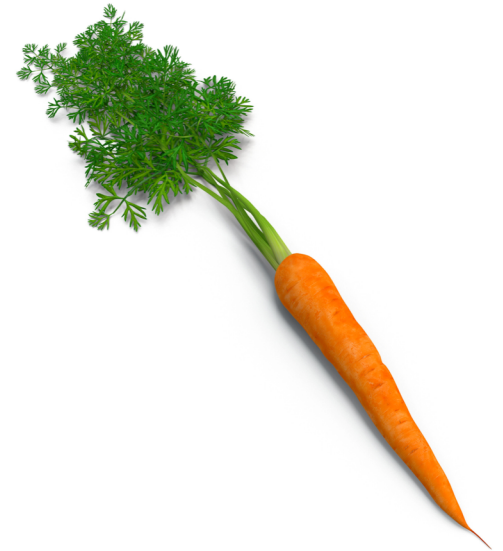
Assembly with reference sequences:

Mapping sequence reads using a reference sequence



How do we perform sequence assembly of single genomes?

Challenge if you don't know what the genome should look like



We have few ways to distinguish true insight from wrongly assembled genome sequence

What is real, what is missing, and what is experimental artifact?



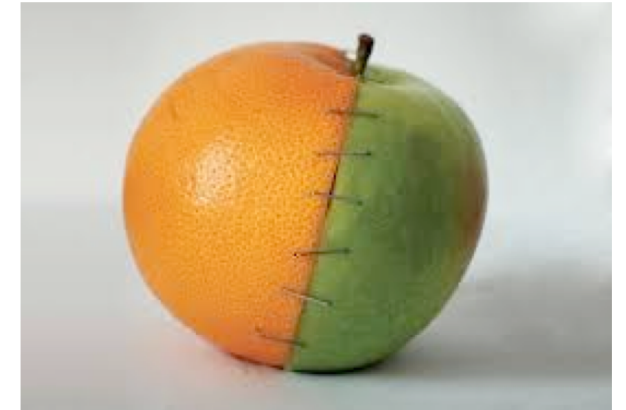
How do we perform sequence assembly of metagenomes?

Even more challenging for metagenomes



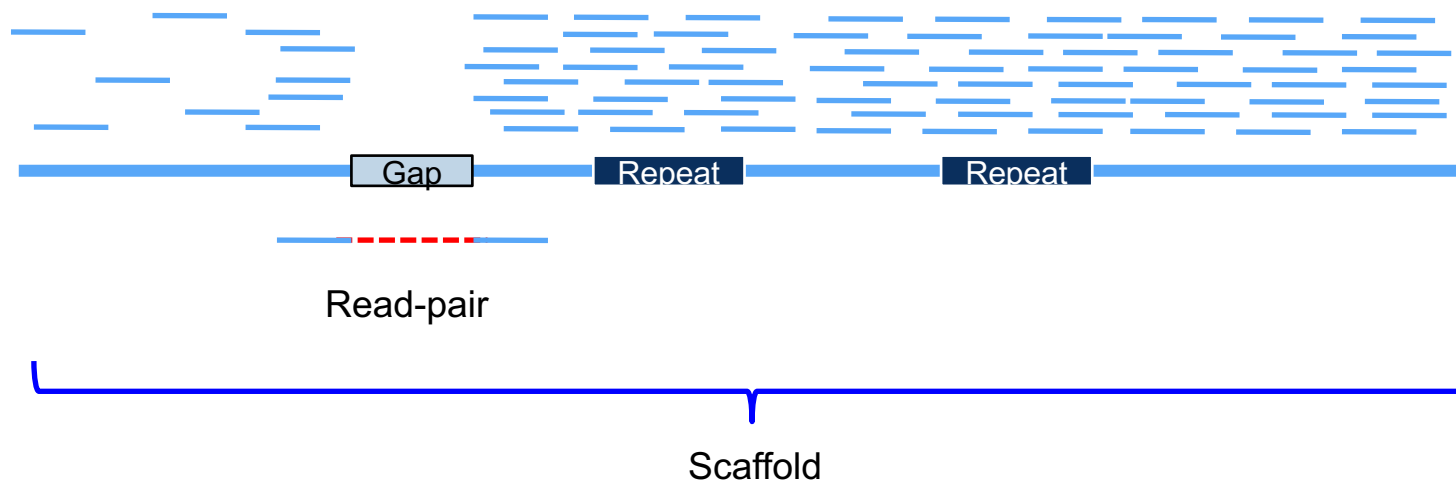
How do we perform sequence assembly of metagenomes?

Diverse samples – more challenging as it is not possible to sequence the complete DNA



Some definitions of terms

- Contig = Consensus sequence of overlapping sequence reads
- Scaffold = Contigs joined together using read-pair information
- Gap = Regions of the original DNA sequence that are not covered
- Repeats = Identical regions of DNA



Some definitions of terms

Contig = Consensus sequence of overlapping sequence reads

Scaffold = Contigs joined together using read-pair information

Gap = Regions of the original DNA sequence that are not covered

Repeats = Identical regions of DNA

Coverage = The average number of reads that cover each base



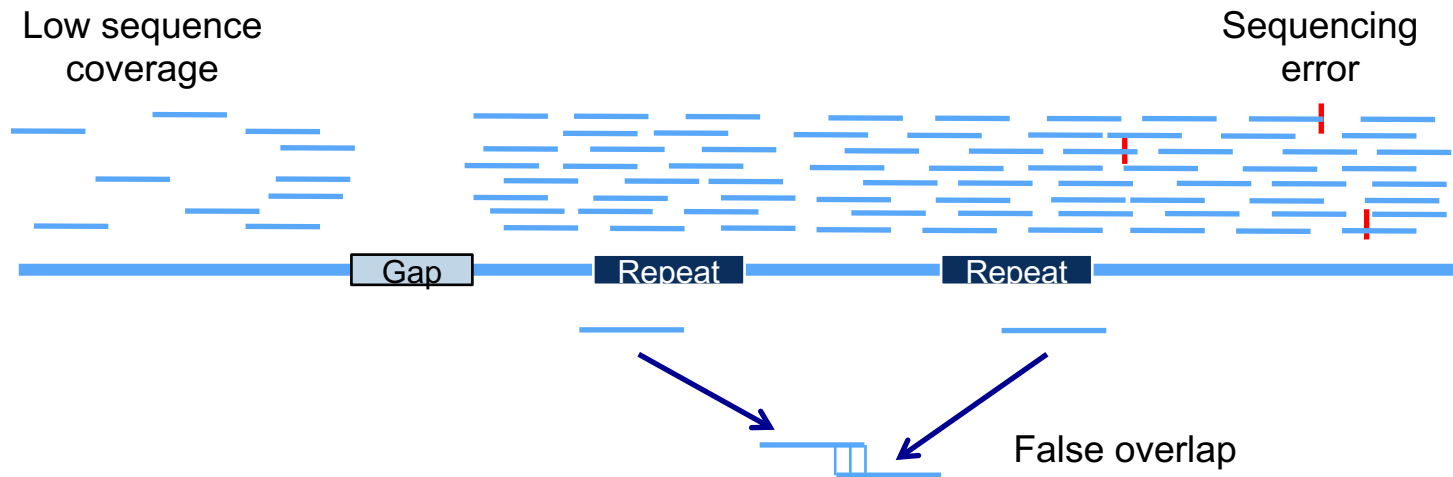
$$\frac{\text{Number of reads (n)} \times \text{Length of reads (l)}}{\text{Length of metagenome (L)}}$$

Some assembly challenges

Uncovered regions

Noise in the data (1-2% of the bases are wrong)

Sequence repeats (bacterial genomes ~5%, mammals ~50%)



Merge overlapping paired-end reads prior to assembly

Generate longer reads by overlapping and merging read pairs before assembling a sequence

S. aureus – PE illumina	Original assembly	FLASH
Total contig size (Mb)	2.91	2.94
Contig N50 size (kb)	1.45	8.40
Contig maximum (kb)	8.18	36.07
Scaffold N50 (kb)	2.07	8.80
Scaffold maximum (kb)	11.23	36.07

Magoč and Salzberg, Bioinformatics. 2011 Nov 1; 27(21): 2957–2963.






Short-read sequencing technologies have made the computational challenge harder

Highly memory-intensive task (TB) and storage demanding (TB)


45 GB of raw sequencing data for 32 × coverage of a human genome (three Illumina HiSeq2500 runs)



Ten steps to get started in Genome Assembly and Annotation [version 1; referees: awaiting peer review]

Victoria Dominguez Del Angel ¹, Erik Hjerde ², Lieven Sterck ^{3,4},
Salvadors Capella-Gutierrez^{5,6}, Cederic Notredame^{7,8}, Olga Vinnere Pettersson⁹,
Joelle Amsellem ¹⁰, Laurent Bouri ¹, Stephanie Bocs ¹¹⁻¹³, Christophe Klopp ¹⁴,
Jean-Francois Gibrat ^{1,15}, Anna Vlasova ⁸, Brane L. Leskosek¹⁶, Lucile Soler¹⁷, Mahesh Binzer-
Panchal ¹⁷,  Henrik Lantz ¹⁷

Lessons learned from implementing a national infrastructure in Sweden for storage and analysis of next-generation sequencing data

[Samuel Lampa](#), [Martin Dahlö](#), [Pall I Olason](#), [Jonas Hagberg](#) and [Ola Spjuth](#) 

GigaScience 2013 2:9 | DOI: 10.1186/2047-217X-2-9 | © Lampa et al.; licensee BioMed Central Ltd. 2013

Some computational considerations

Reference Genome	Size	Software	Input (space used on disk)	CPU/RAM Available	Real time	Max RAM Used
<i>Aliivibrio wodanis</i>	4 972 754 bp	SPAdes v3.10	200x Illumina reads (760 MB)	4 CPU/16GB RAM	2h17m3s	2,94GB
				12 CPU/256GB RAM	38m8s	9,37GB
<i>Caenorhabditis elegans</i>	100 272 607 bp	Smartdenovo	20x Pacbio P6C4 Corrected long reads (1,9 GB)	8 CPU/16GB RAM	24m47s	1,92GB
			80x Pacbio P6C4 Corrected long reads (7,6 GB)	8 CPU/16GB RAM	5h38m16s	7,29GB
		REPET v2.5	<i>C. Elegans</i> genome (100 MB) Repbse aa 20.05 (20 MB) Pfam 27 (GypsyDB) (1,2 GB) rRNA from eukaryota (2,6 MB)	8 CPU/16 GB RAM	1h53m11s + 19h9m40s	8,96GB
		Eugene v4.2a	<i>C. Elegans</i> genome (100 MB) Repbse aa 20.05 (20 MB) Proteins sequences (swissprot) (2,8 MB) ESTs sequences (29 MB)	8 CPU/32 GB RAM	5h2m30s	16,94GB
<i>Arabidopsis thaliana</i>	134 634 692 bp	Smartdenovo	20x Pacbio P5C3 corrected long reads (2,7 GB)	8 CPU/16GB RAM	1h16m20s	2,4GB
		REPET v2.5	<i>A. Thaliana</i> genome (130 MB) Repbse aa 20.05 (20 MB) Pfam 27 (GypsyDB) (1,2 GB) rRNA from eukaryota (2,6 MB)	8 CPU/16 GB RAM	5h6m23s + 33h10m34s	10,25GB
		Eugene v4.2a	<i>A. Thaliana</i> genome (130 MB) Repbse aa 20.05 (20 MB) Proteins sequences (swissprot) (9,2 MB) ESTs sequences (31 MB)	8 CPU/32 GB RAM	6h17m18s	17,25GB
<i>Theobroma cacao</i>	324 761 211 bp	Eugene v4.2a	<i>T. Cacao</i> genome (315 MB) Repbse aa 20.05 (20 MB) Proteins sequences (swissprot) (31 MB) ESTs sequences (100 MB)	8 CPU/188 GB RAM	41h27m13s	72,5GB

Some questions you should ask before you start genome sequencing

What is the purpose of sequencing the metagenome?

Complete sequence (Base-perfect sequencing)

Draft sequence

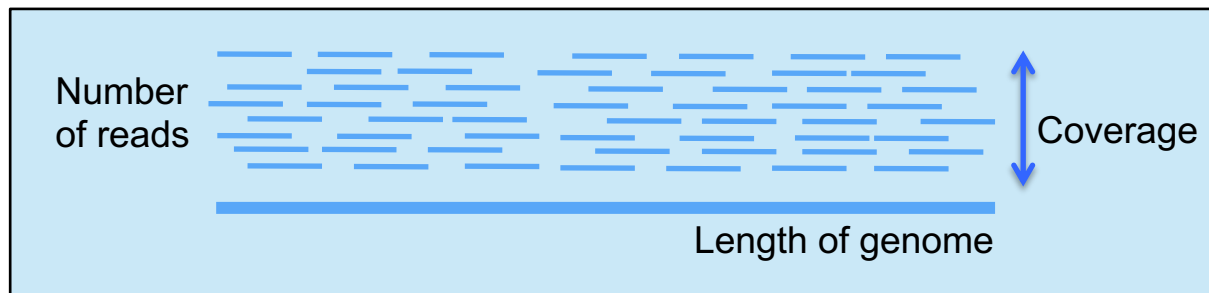
How much data (and what technology) do you need?

Access to computational resources?

Plan for analyses?



<http://www.sullivan-financial.com/p/planning-your-financial-future>



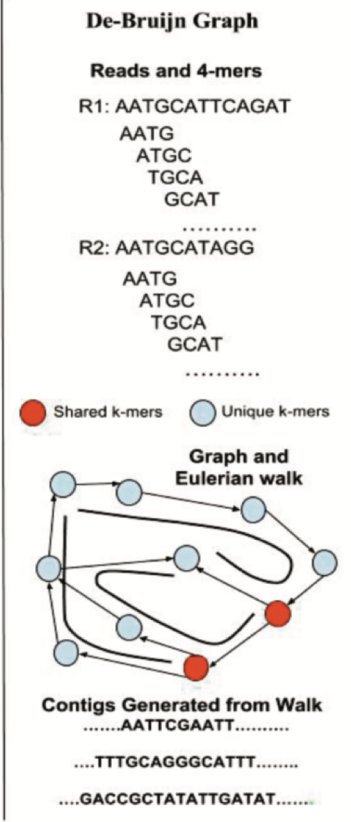
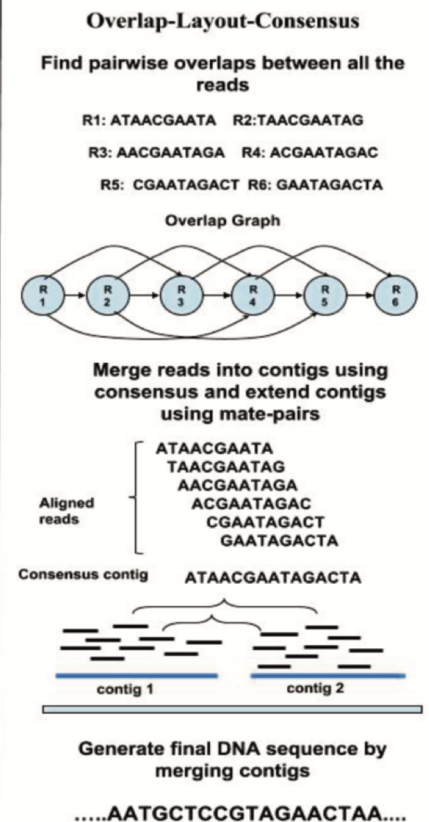
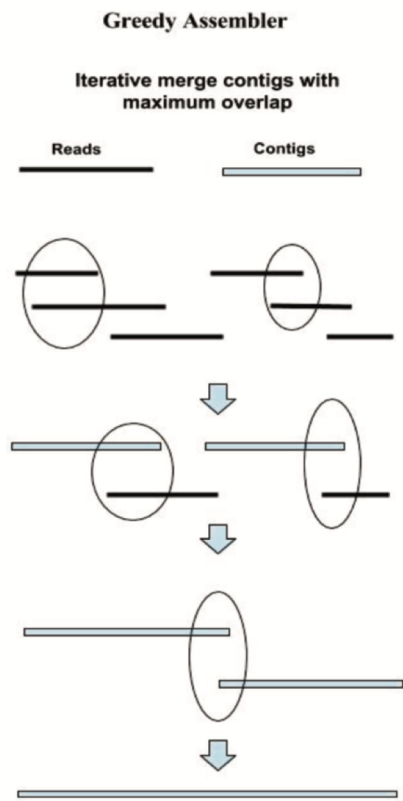
$$\text{Coverage} = \frac{\text{Number of reads} \times \text{Length of read}}{\text{Length of genome}}$$

Graph-based assembly methods

Greedy graph assembly (greedy extension, or extension-based)

Overlap-Layout-Consensus assembly (OLC)

De Bruijn graph assembly (DBG)



'Bridges of Königsberg problem' - Leonhard Euler in 1735

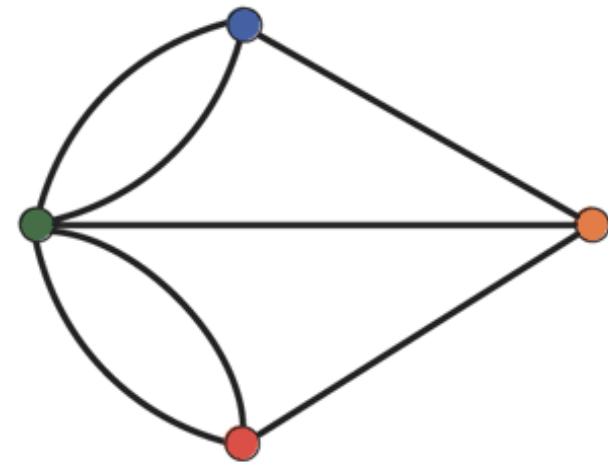
Seven bridges joined the four parts of the city located on opposing banks of the Pregel River and two river islands.

Could every part of the city could be visited by walking across each of the seven bridges exactly once and returning to one's starting location?

a



b



'Bridges of Königsberg problem' - Leonhard Euler in 1735

Euler represented each landmass as a point (called a node) and each bridge as a line segment (called an edge) connecting two points.

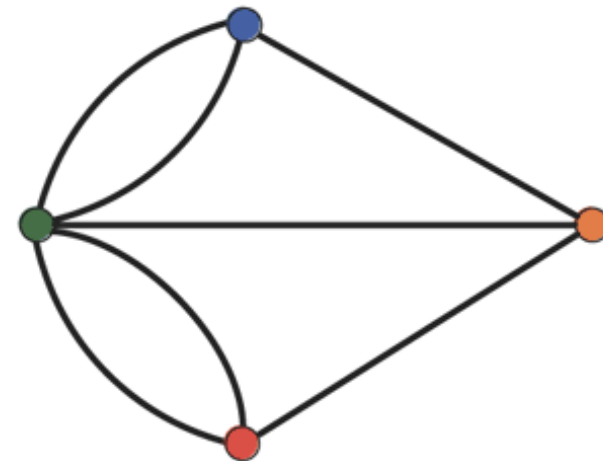
This creates a graph—a network of nodes connected by edges

Algorithm determining whether an arbitrary graph contains a path that visits every edge exactly once and returns to where it started

a



b

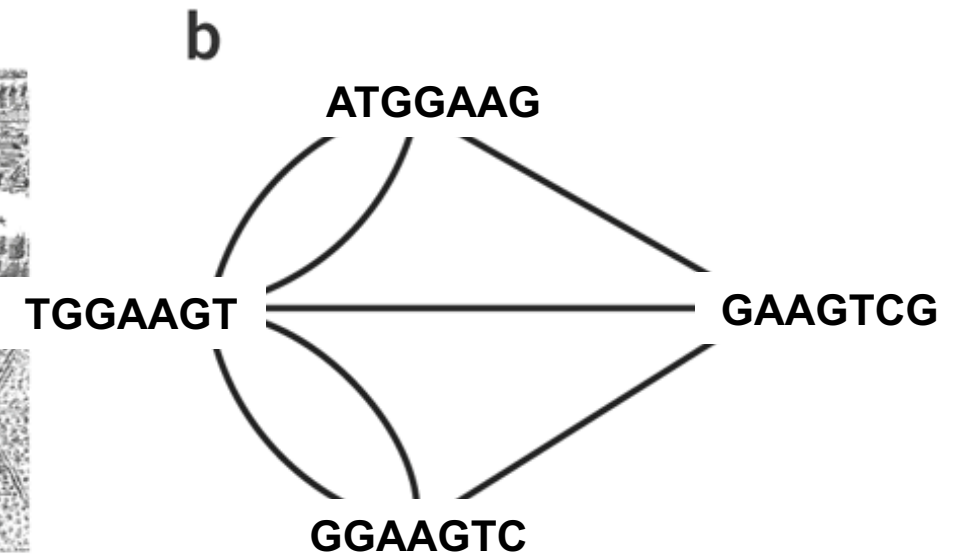


'Bridges of Königsberg problem' - Leonhard Euler in 1735

Euler represented each landmass as a point (called a node) and each bridge as a line segment (called an edge) connecting two points.

This creates a graph—a network of nodes connected by edges

Algorithm determining whether an arbitrary graph contains a path that visits every edge exactly once and returns to where it started



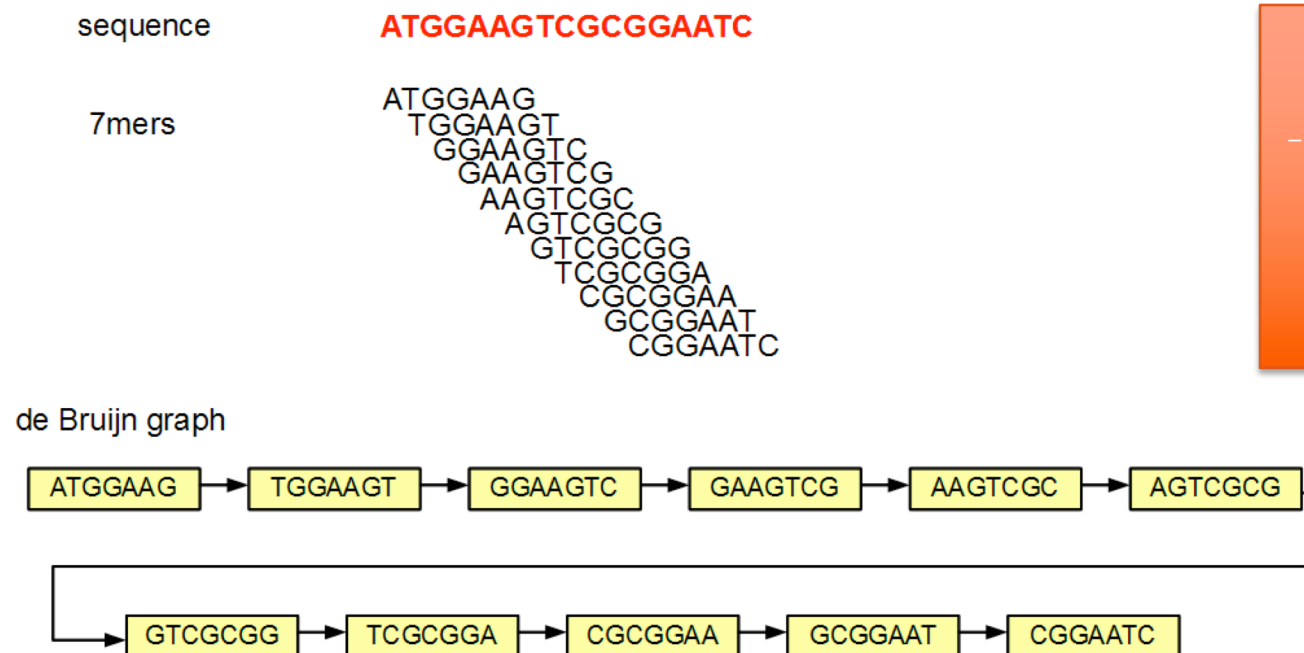
de bruijn graph are used by most modern de novo assemblers

Creates a sorted table of all sub-sequences (words) found in the reads

The words are relatively short, e.g. about 20 (20 mers)

Given any word in the table, it will look up potential neighbouring words

The algorithm tries to make a graph (Eulerian path) connecting all words



Construct a de Bruijn graph (DBG)

- Nodes = one for each unique k-mer
- Edges = k-1 exact overlap between two nodes

Graph simplification

- Merge chains, remove bubbles and Rps

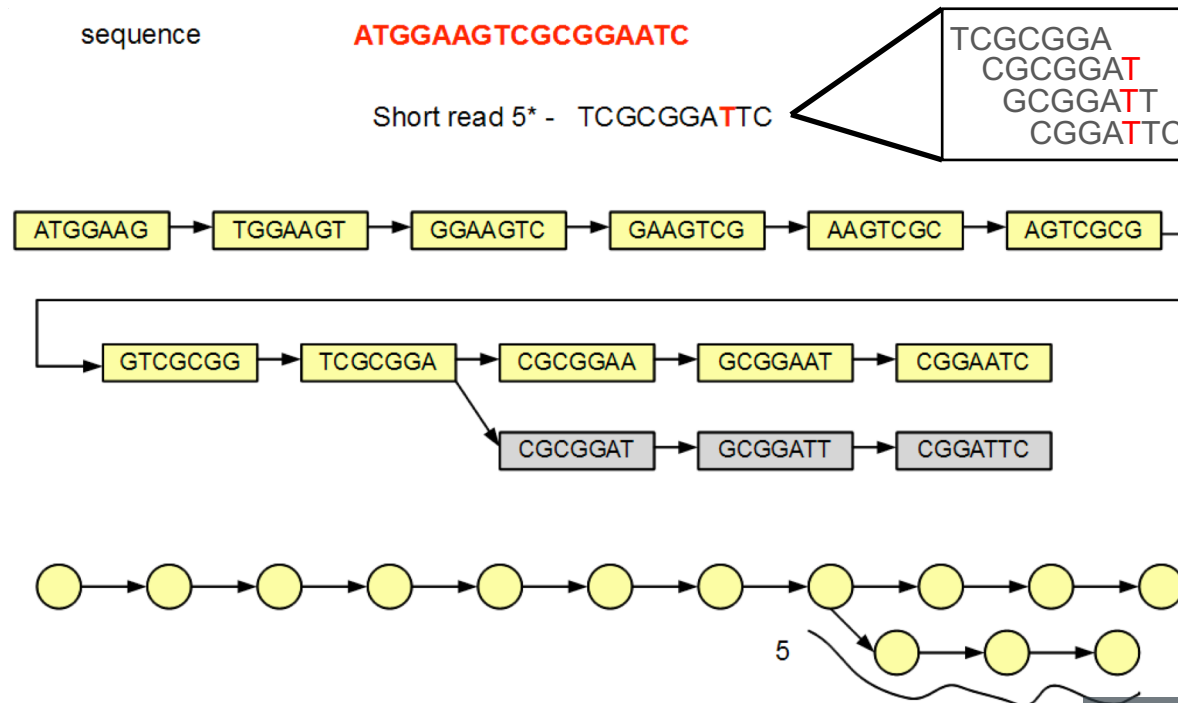
Find a Eulerian path through the graph

de bruijn graph are used by most modern de novo assemblers

SNPs or a sequencing errors will create so-called bubbles

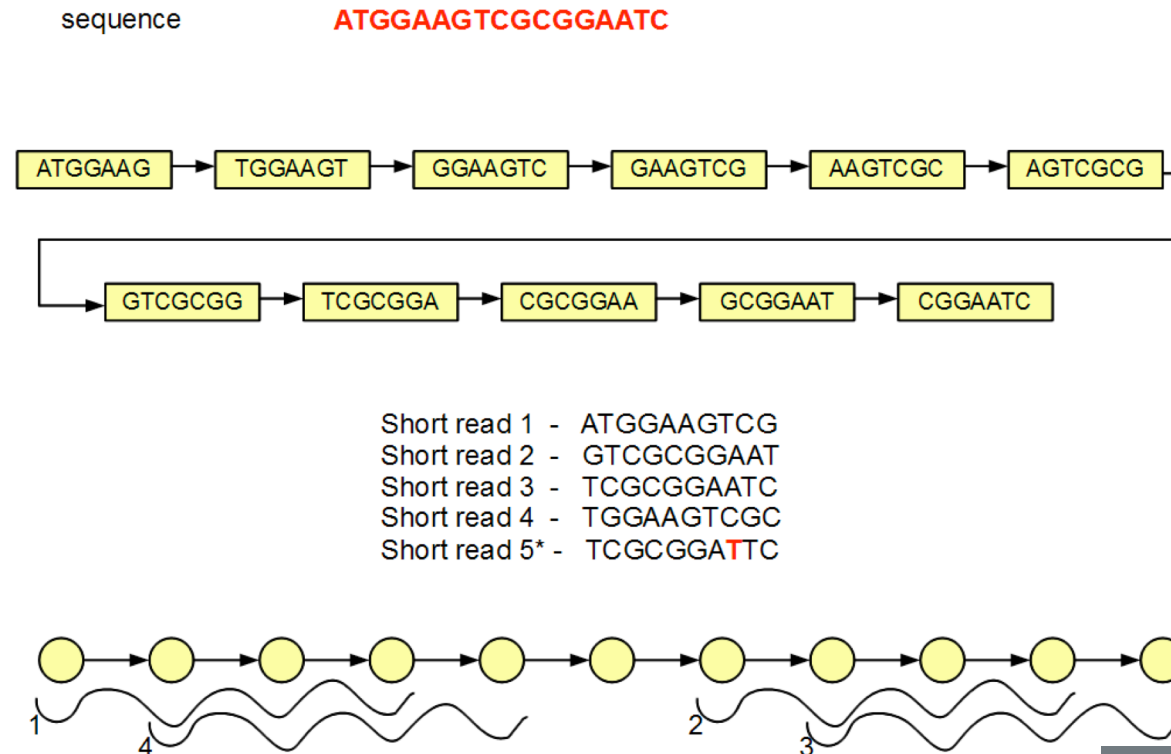
For sequencing errors the deviating word occurs only once

For heterozygous SNPs both paths represented more or less equally



de bruijn graph are used by most modern de novo assemblers

It will continue to add words – build coverage of the assembly

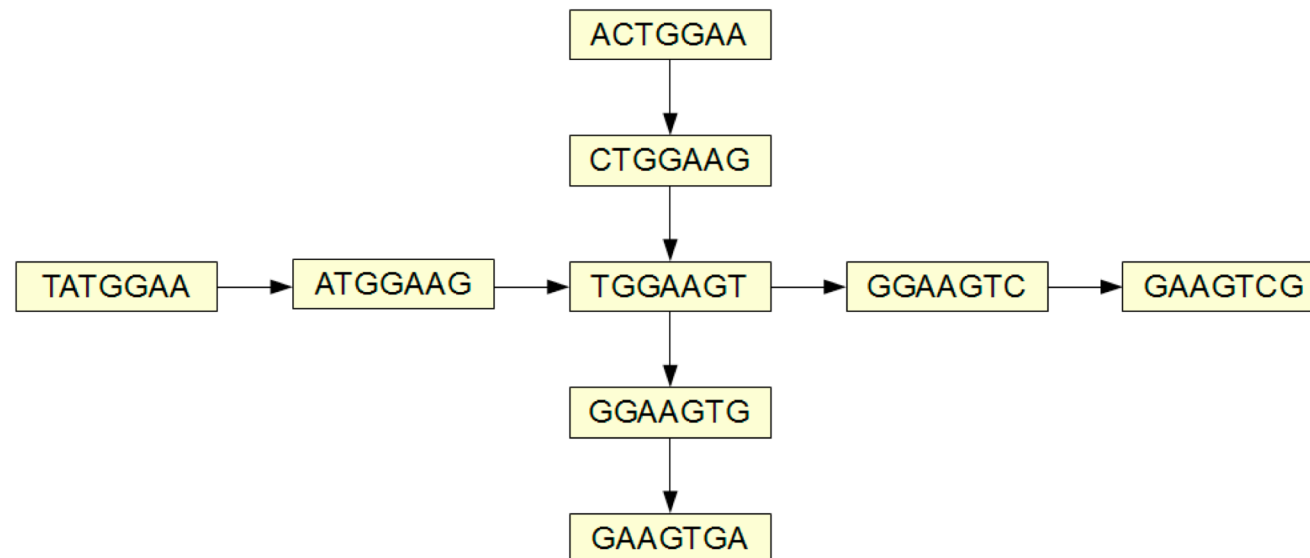


de bruijn graph are used by most modern de novo assemblers

Repeats are the most difficult problem for the de novo assembly

Impossible to resolve if the repeat is longer than the paired distance of read pairs

Such repeats will cause the assembler to spit the graph – make contigs

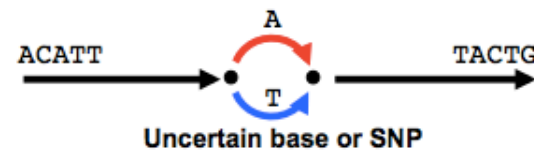
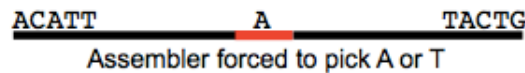


Many assemblers produce an assembly graph in FASTG format (G=graph)

Unlike FASTA (linear representation), FASTG can express branching arising from eg. ambiguities and repetitive segments

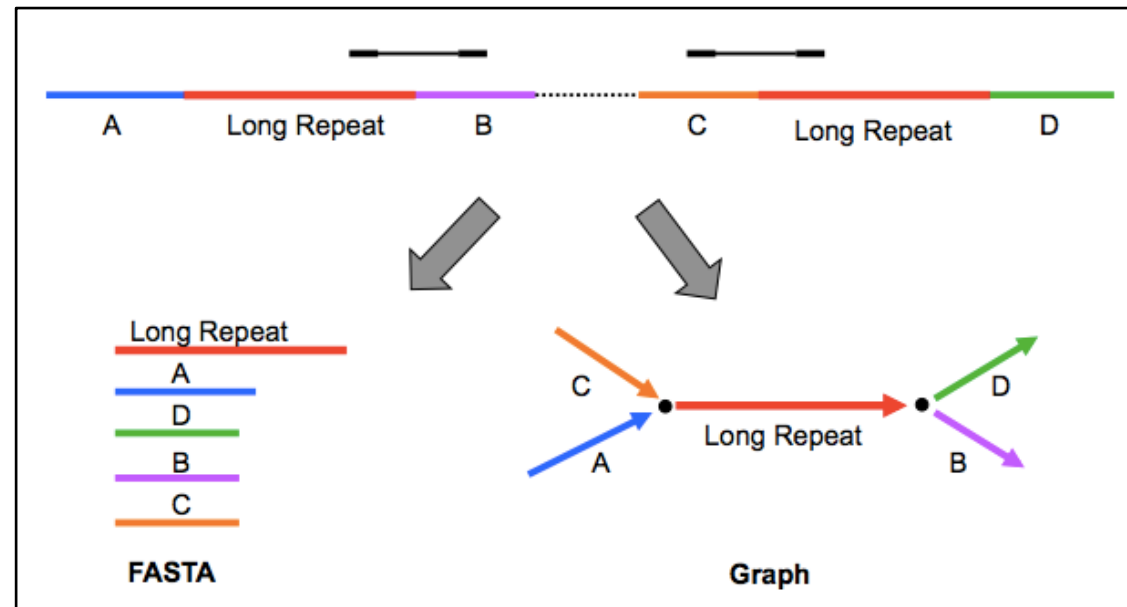
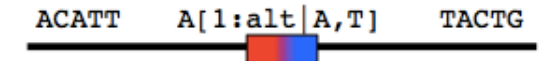
FASTA forces assemblers to make mistakes

- Strictly linear nature forces assemblers to introduce errors:



FASTG encodes all ambiguities

- FASTG natively encodes ambiguities that are lost in FASTA



FASTG can easily be converted to FASTA

FASTG and derived FASTA files share the same base co-ordinate system

FASTA + Markup will produce the original FASTG

FASTG

```
>contig1;  
TACCGCNNNN[4:gap:size=(4,3..5)]AGCCTGCC  
GTTATAC[1:alt:allele|C,G]TCCCTGGATACGTT  
TAGGATATAT[6:tandem:size=(3,2..5)|AT]CC
```



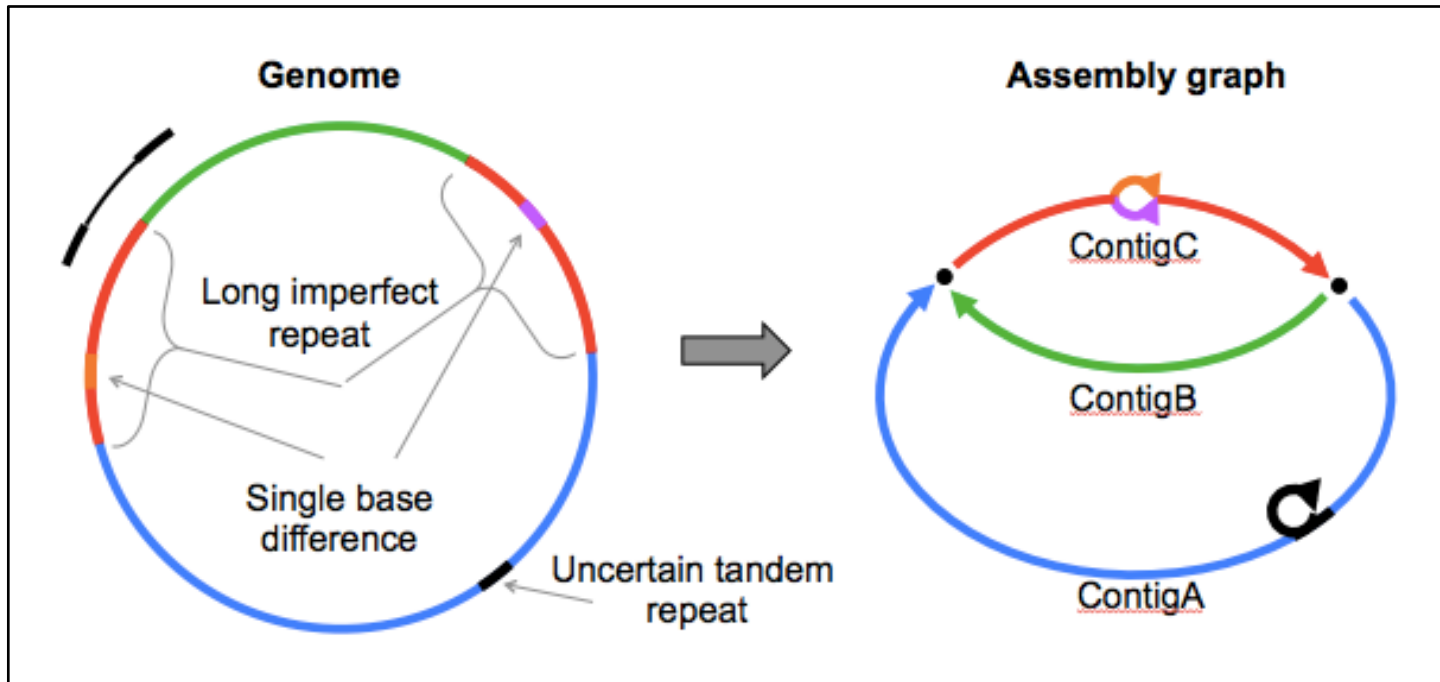
FASTA

```
>contig1  
TACCGCNNNNAGCCTGCC  
GTTATACCTCCCTGGATA  
CGTTTAGGATATATCC
```

+

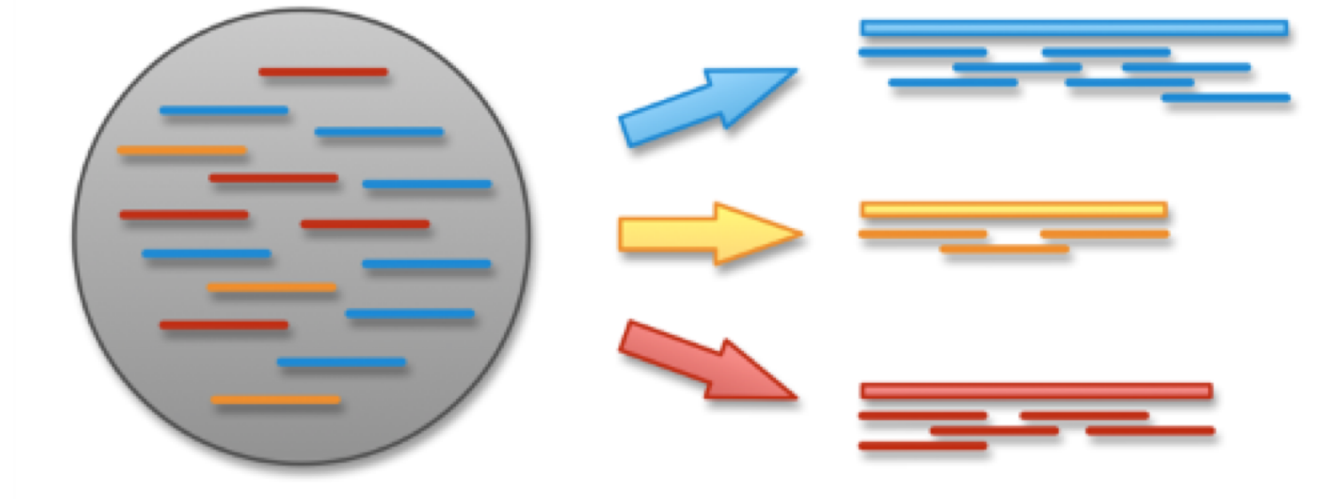
Markup

```
>contig1;  
6 [4:gap:size=(4,3..5)]  
26 [1:alt:allele|C,G]  
52 [6:tandem:size=(3,2..5)|AT]
```



Metagenome assembly tools

Megahit
MetaSPAdes
Snowball
MetaVelvet
Ray Meta
MetAMOS



Andreas Bremges

CAMI - challenge the developers to benchmark their programs

- Highly complex and realistic data sets
- ~700 newly sequenced microorganisms
- ~600 novel viruses and plasmids
- Assembly and genome binning
- Taxonomic profiling and binning

nature|methods

Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software

Alexander Sczyrba , Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, Stephan Majda, Jessika Fiedler, Eik Dahms, Andreas Bremges, Adrian Fritz, Ruben Garrido-Oter, Tue Sparholt Jørgensen, Nicole Shapiro, Philip D Blood, Alexey Gurevich, Yang Bai, Dmitrij Turaev, Matthew Z DeMaere, Rayan Chikhi, Niranjana Nagarajan, Christopher Quince, Fernando Meyer, Monika Balvočiūtė, Lars Hestbjerg Hansen, Søren J Sørensen, Burton K H Chia, Bertrand Denis, Jeff L Froula, Zhong Wang, Robert Egan, Dongwan Don Kang, Jeffrey J Cook, Charles Deltel, Michael Beckstette, Claire Lemaitre, Pierre Peterlongo, Guillaume Rizk, Dominique Lavenier, Yu-Wei Wu, Steven W Singer, Chirag Jain, Marc Strous, Heiner Klingenberg, Peter Meinicke, Michael D Barton, Thomas Lingner, Hsin-Hung Lin, Yu-Chieh Liao, Genivaldo Gueiros Z Silva, Daniel A Cuevas, Robert A Edwards, Surya Saha, Vitor C Piro, Bernhard Y Renard, Mihai Pop, Hans-Peter Klenk, Markus Göker, Nikos C Kyrpides, Tanja Woyke, Julia A Vorholt, Paul Schulze-Lefert, Edward M Rubin, Aaron E Darling, Thomas Rattei & Alice C McHardy  - Show fewer authors

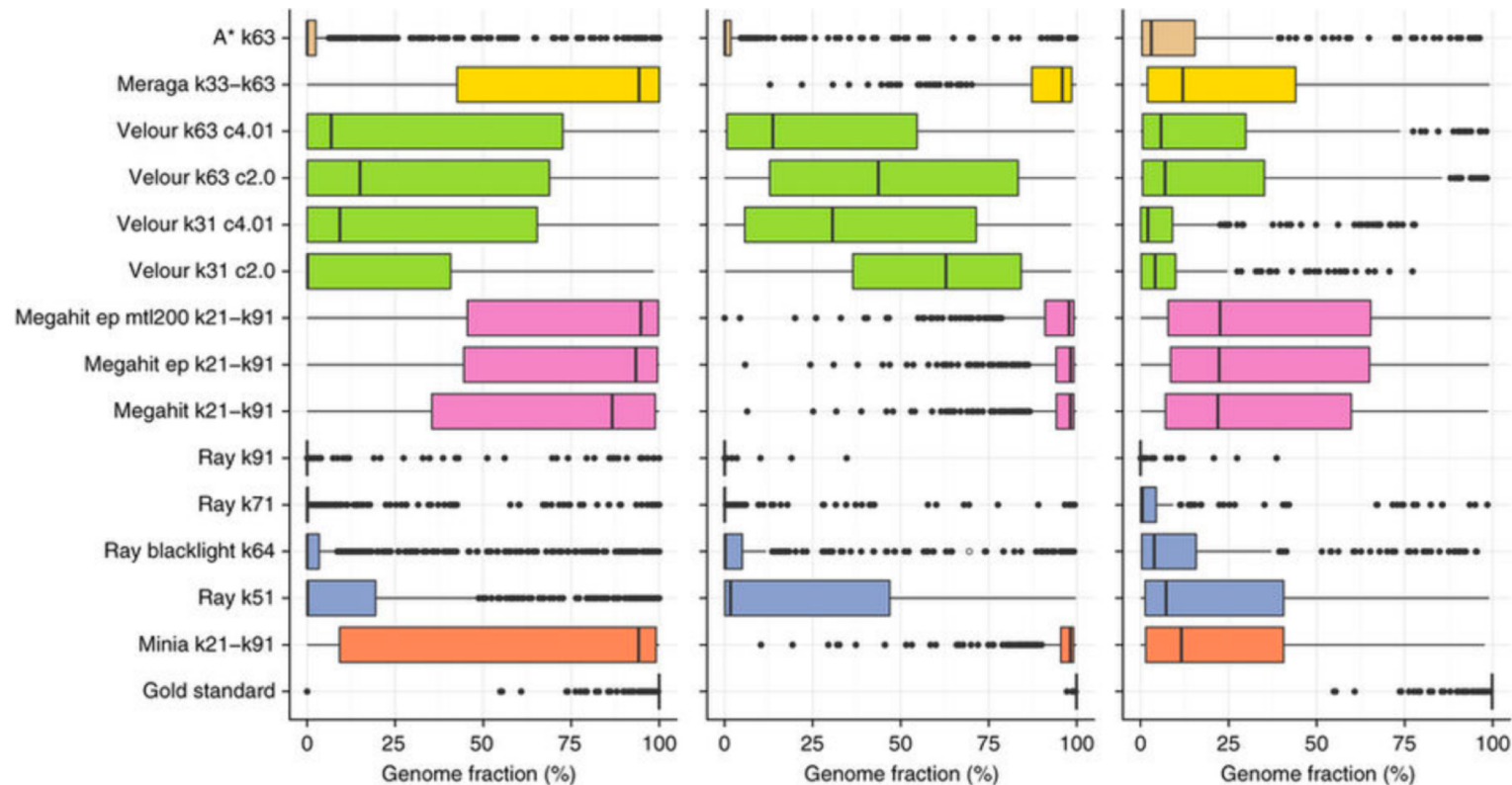
Metagenome assembly tools - performance

Main conclusion:

Assembly is substantially affected by the presence of related strains

Parameter settings markedly affected performance

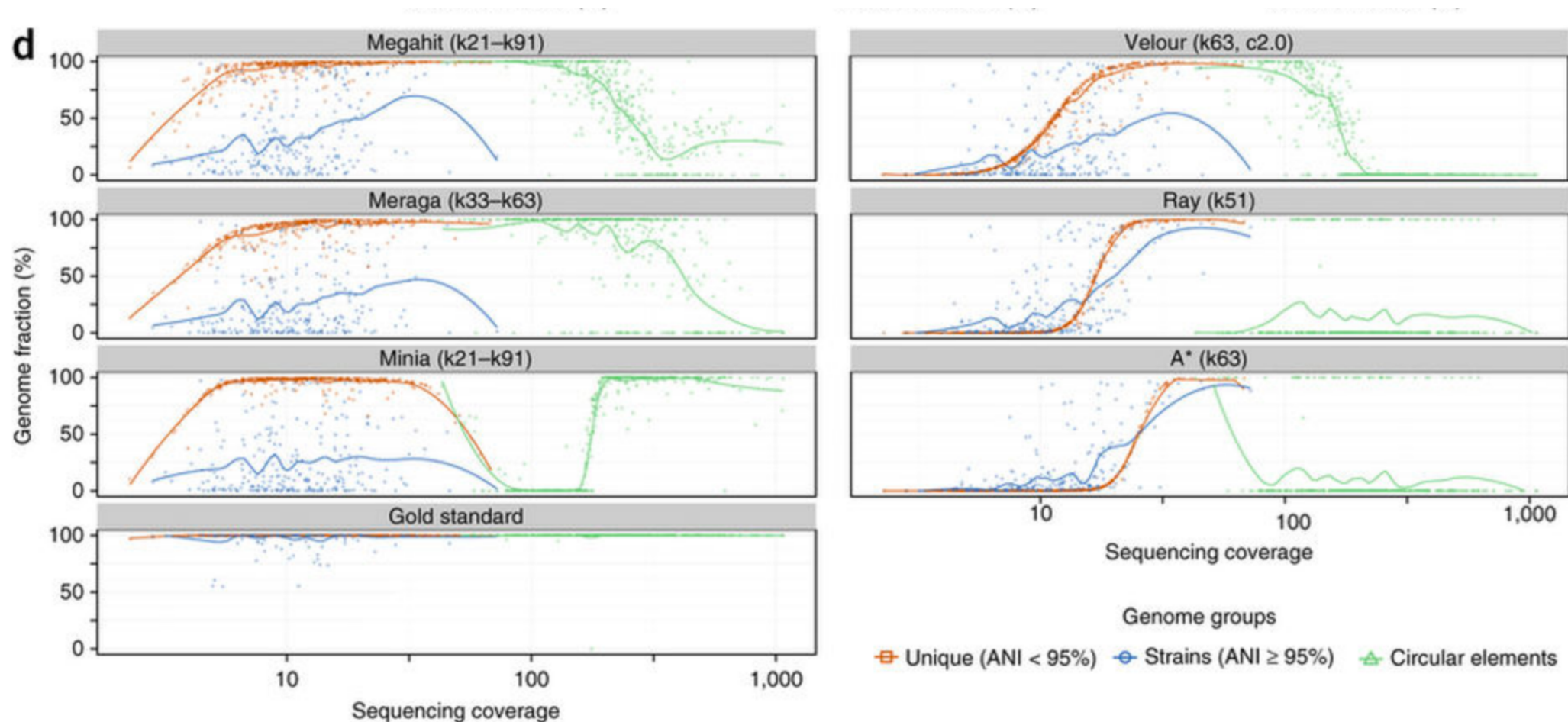
Assemblers using multiple k-mers (Minia, MEGAHIT and Meraga) substantially outperformed single k-mer assemblers



Metagenome assembly tools - performance

Main conclusion:

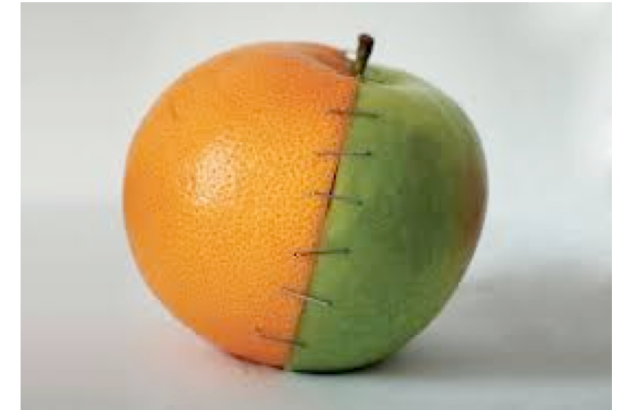
Most assemblers except for Meraga and Minia did not recover very-high-copy circular elements



Evaluation of metagenome assemblies

Assembly accuracy is difficult to measure!!!!

Few ways to distinguish true insight from wrongly assembled metagenome sequences



Contiguity-based evaluation of sequence assemblies

MetaQUAST evaluates and compares metagenome assemblies based on alignments to close references

N₅₀ = the smallest of the largest contigs covering 50% of the total size of all contigs

Misassemblies where two parts of the same contig align to distinct references

Contigs that include both large aligned and unaligned fragments

Statistics without reference	IDBA_UD	Ray	SOAPdenovo2	SPAdes
+ # contigs	31 224	10 327	36 468	40 546
+ Largest contig	305 144	99 107	40 707	189 063
+ Total length	80 325 286	30 411 921	46 741 224	92 397 329
+ Total length (>= 1000 bp)	69 223 529	27 080 646	30 720 336	77 823 828
+ Total length (>= 10000 bp)	34 930 908	13 755 677	2 800 864	33 477 263
+ Total length (>= 50000 bp)	16 008 349	2 346 322	0	11 409 912
Misassemblies				
+ # misassemblies	1132	407	831	1240
+ Misassembled contigs length	10 448 260	4 115 772	911 826	10 780 557
Mismatches				
+ # mismatches per 100 kbp	904.95	1054.68	888.21	1401.84
+ # indels per 100 kbp	31.88	27.7	17.09	51.64
+ # N's per 100 kbp	238.48	2087.27	3730.51	1425.14
Genome statistics				
- Genome fraction (%)	12.796	4.386	8.055	11.585
Akkermansia_muciniphila_ATCC	0.003	-	-	0.011
Alistipes_putredinis	1.366	0.595	0.61	1.117
Anaerotruncus_colihominis	2.466	2.067	1.768	2.320
Bacteroides_caccae	5.343	2.643	3.928	5.138
Bacteroides_capillosus	1.173	0.27	0.449	1.05
Bacteroides_cellulosilyticus	1.278	0.952	1.824	0.96
Bacteroides_coprocola	30.532	-	-	-



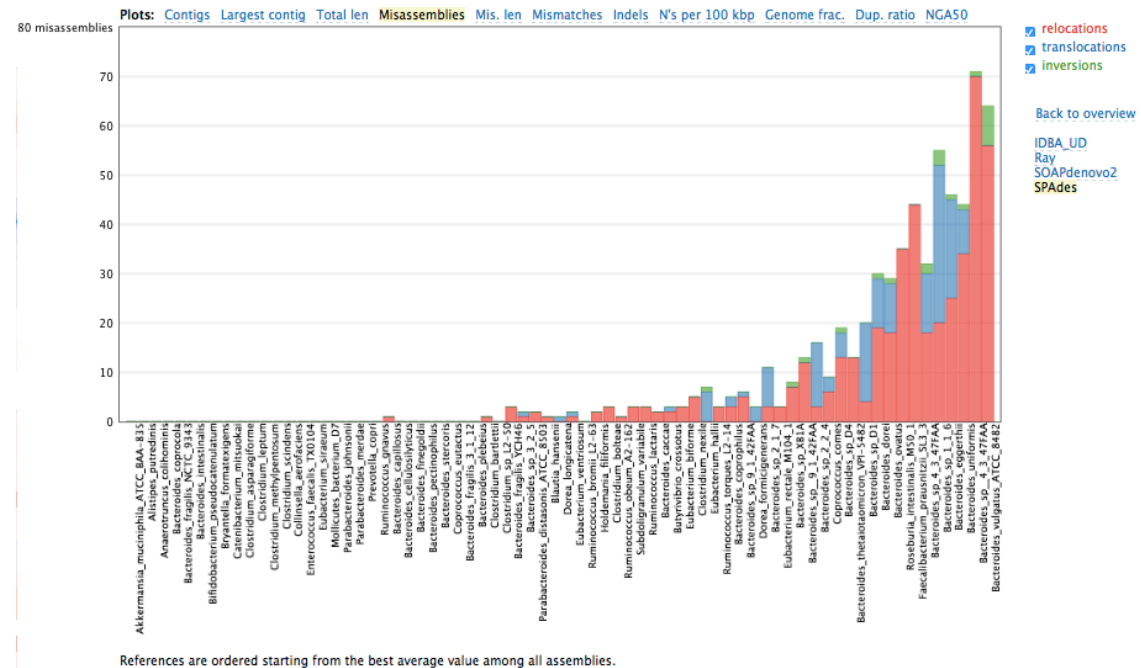
MetaQUAST: evaluation of metagenome assemblies
 Bioinformatics. 2015;32(7):1088-1090.
 doi:10.1093/bioinformatics/btv697

Compare the assembly from different assemblers

Or with raw data or trimmed/filtered data

Reference size: 306 971 432 bp

Reference	Size, bp	GC, %
Akkermansia_muciniphila_ATCC_BAA-835	2 664 102	55.76
Alistipes_putredinis	2 550 678	53.27
Anaerotruncus_colihominis	3 719 688	54.18
Bacteroides_caccae	5 493 117	42.83
Bacteroides_capillosus	4 241 076	59.11
Bacteroides_cellulosilyticus	7 694 202	43.05
Bacteroides_coprocola	2 784	45.19
Bacteroides_coprophilus	4 041 504	45.72
Bacteroides_dorei	6 060 928	42.2
Bacteroides_eggerthii	4 611 535	44.71
Bacteroides_finegoldii	5 124 109	42.5
Bacteroides_fragilis_3_1_12	5 530 115	43.62
Bacteroides_fragilis_NCTC_9343	5 205 140	43.19
Bacteroides_fragilis_YCH46	5 277 274	43.27
Bacteroides_intestinalis	4 605 106	43.54
Bacteroides_ovatus	7 010 996	42.3
Bacteroides_pectinophilus	29 332	36.96
Bacteroides_plebeius	4 421 924	44.31
Bacteroides_sp_1_1_6	6 760 735	43.02
Bacteroides_sp_2_1_7	5 180 144	45.08
Bacteroides_sp_2_2_4	7 101 224	42.13
Bacteroides_sp_3_2_5	5 116 282	43.17
Bacteroides_sp_4_3_47FAA	5 442 925	42.7
Bacteroides_sp_9_1_42FAA	5 622 644	42.33
Bacteroides_sp_D1	5 974 559	41.88
Bacteroides_sp_D4	5 538 248	41.75
Bacteroides_sp_XB1A	5 976 145	41.89
Bacteroides_sp_4_3_47FAA	5 442 925	42.7
Bacteroides_sp_9_1_42FAA	4 684 745	42.2
Bacteroides_stercoris	4 102 660	45.93
Bacteroides_thetaiotaomicron_VPI-5482	6 260 361	42.84
Bacteroides_uniformis	4 835 507	46.49
Bacteroides_vulgatus_ATCC_8482	5 163 189	42.2
Bifidobacterium_pseudocatenulatum	2 313 752	56.38
Blautia_hansenii	3 058 721	38.99
Bryantella_formatexigens	4 548 960	49.55
Butyrivibrio_crossotus	2 496 039	37.75

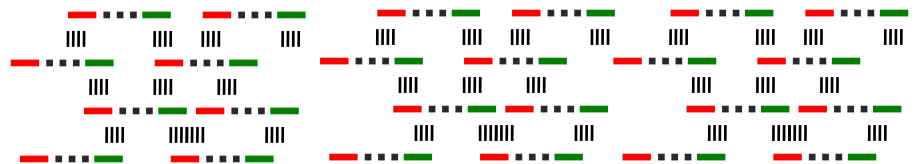


Consistency-based evaluation of sequence assemblies

Align reads against assembly of itself (not against reference)

Erroneous placement of reads within the assembly

These signatures that can be detected computationally



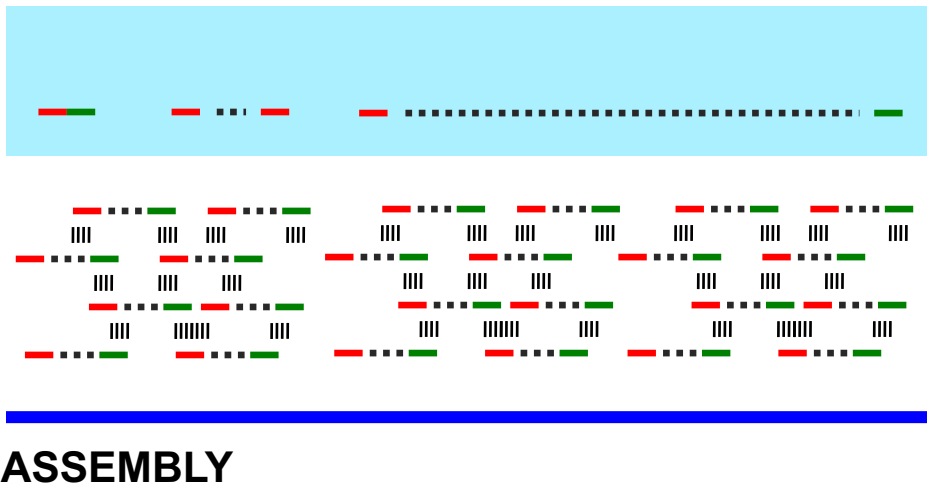
ASSEMBLY

Consistency-based evaluation of sequence assemblies

Align reads against assembly of itself (not against reference)

Erroneous placement of reads within the assembly

These signatures that can be detected computationally

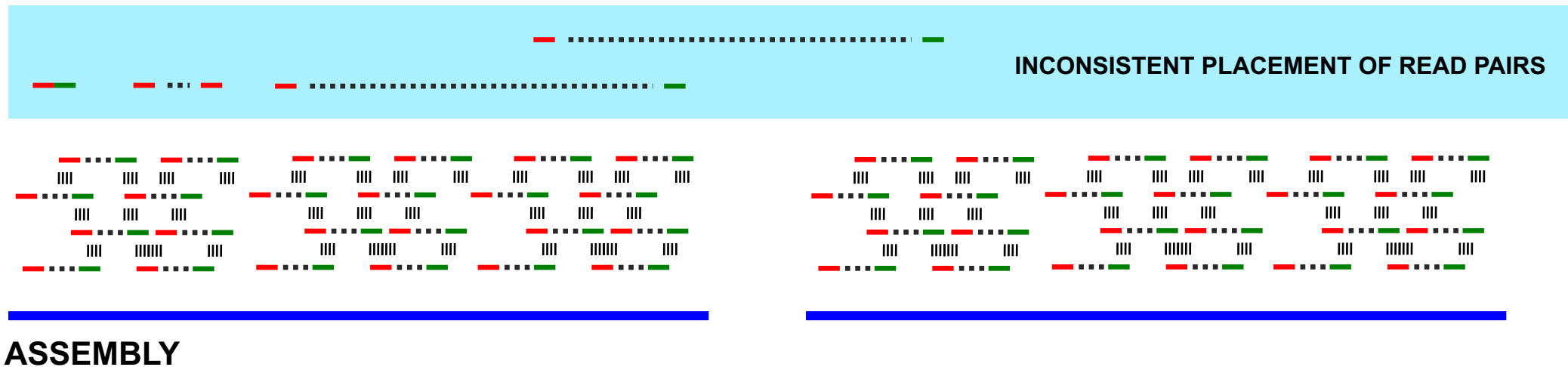


Consistency-based evaluation of sequence assemblies

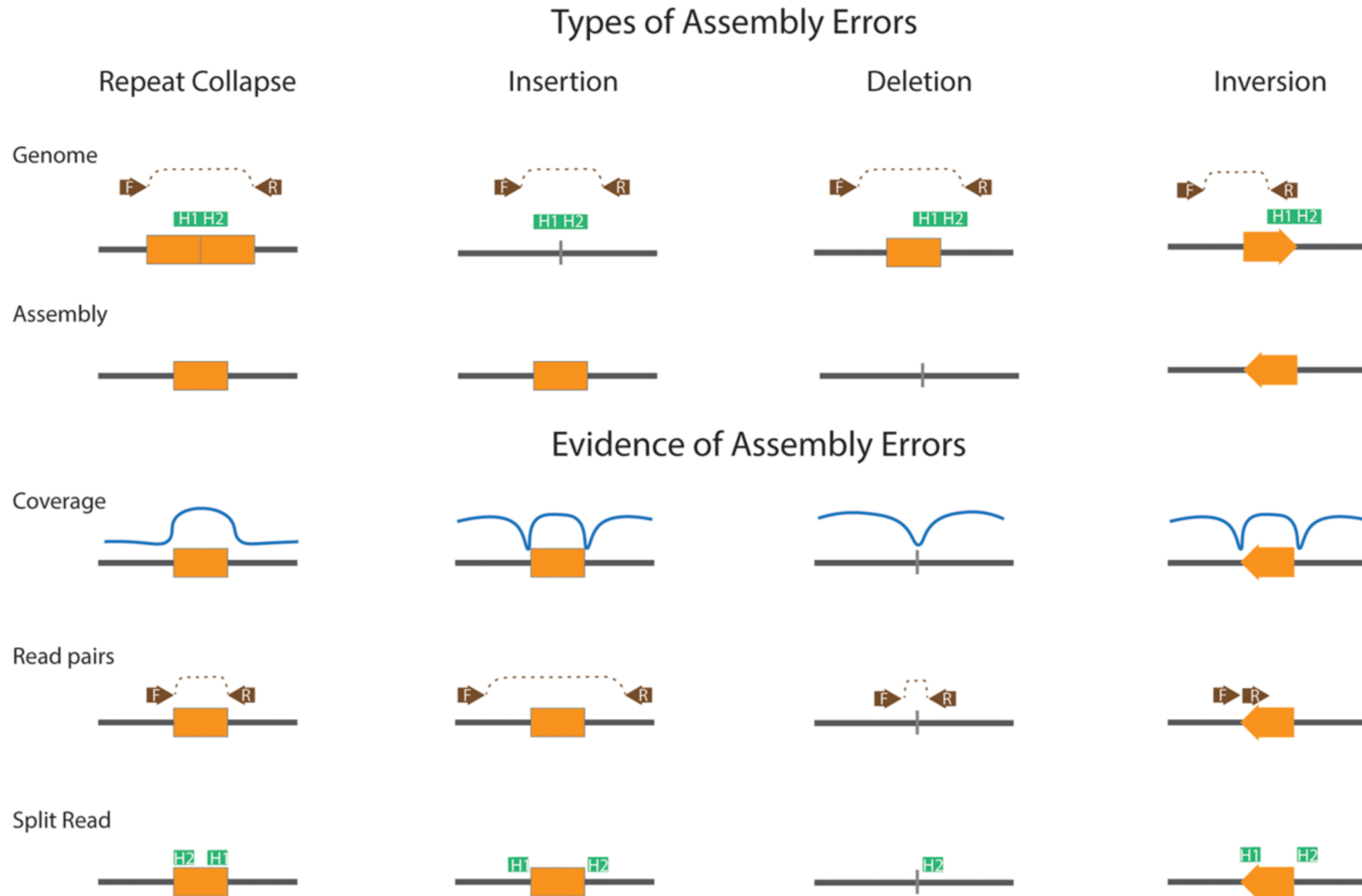
Align reads against assembly of itself (not against reference)

Erroneous placement of reads within the assembly

These signatures that can be detected computationally

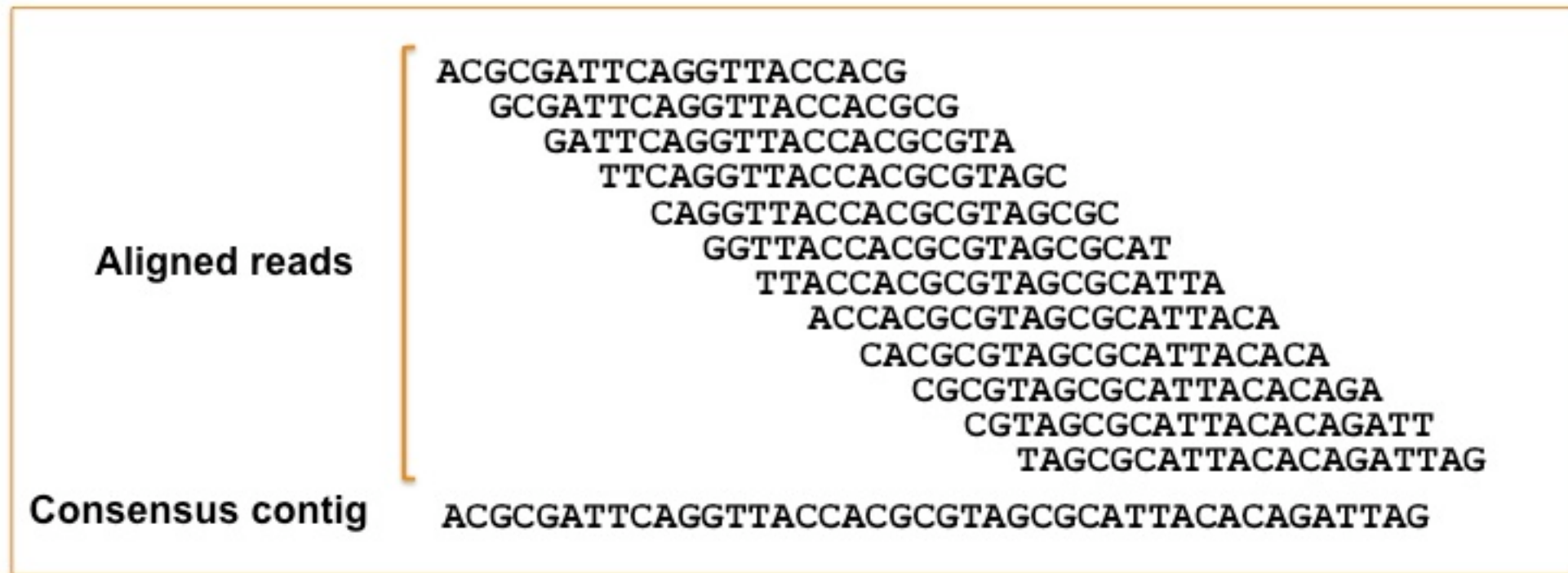


Four primary types of assembly errors that can be identified by mapping reads to the assembly



Use read alignment statistics to see how well do the reads align back to the draft assemblies

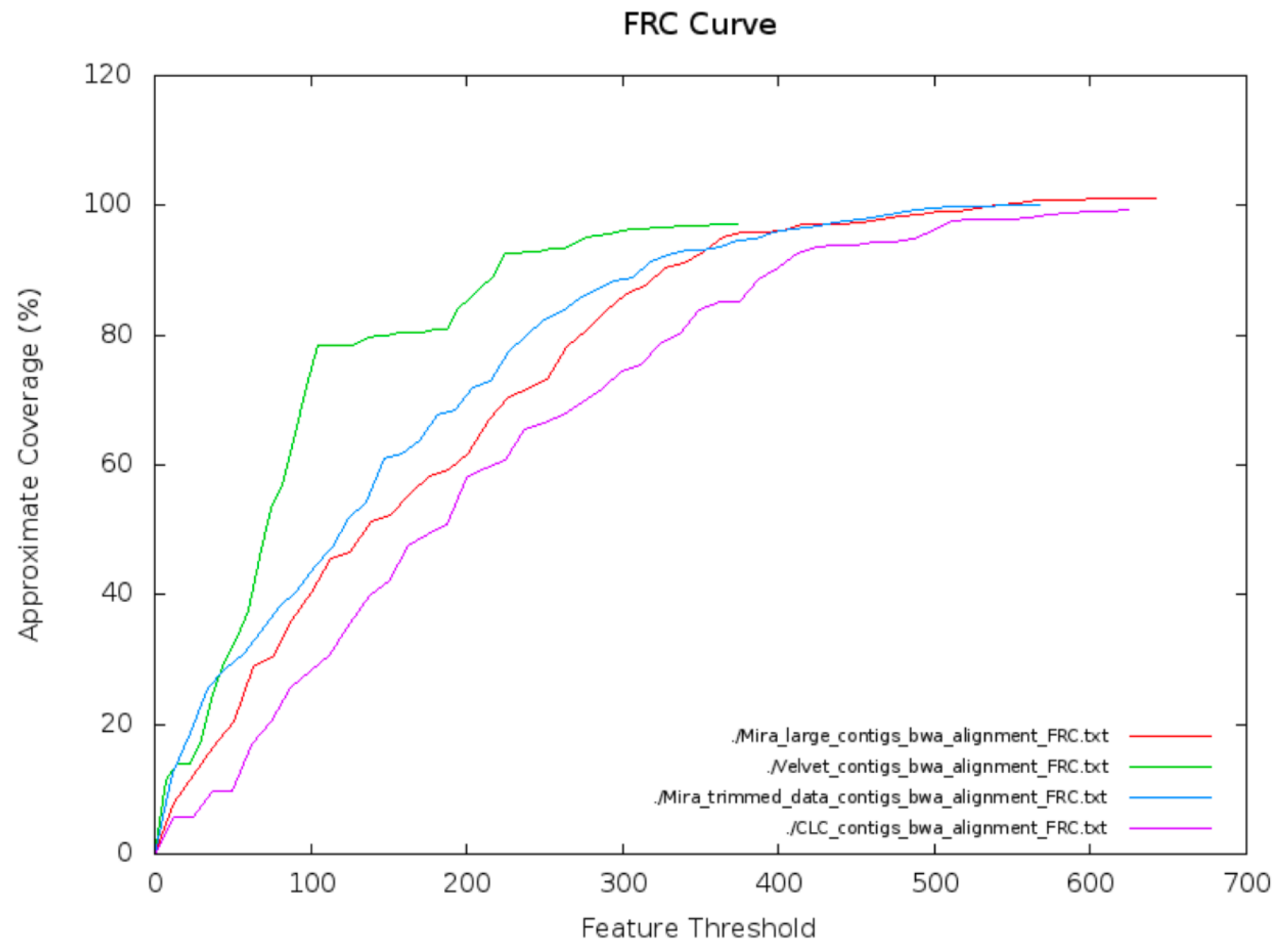
Read congruency is an important measure in determining assembly accuracy
Clusters of read pairs that align incorrectly are strong indicators of mis-assembly



FRCbam uses the alignment of reads to find regions of assembled sequence that appear to be inconsistent with the read data

Reports features (possible inconsistencies) in FRCs (Feature Response Curves)

For example regions with many PE reads with pair mapped in different contigs



For example regions with low coverage

FRCbam uses the alignment of reads to find regions of assembled sequence that appear to be inconsistent with the read data

Reports features (possible inconsistencies) in FRCs (Feature Response Curves)

Feature	Description
LOW_COV_PE	<i>low read coverage areas (all aligned reads).</i>
HIGH_COV_PE	<i>high read coverage areas (all aligned reads).</i>
LOW_NORM_COV_PE	<i>low paired-read coverage areas (only properly aligned pairs).</i>
HIGH_NORM_COV_PE	<i>high paired-read coverage areas (only properly aligned pairs).</i>
COMPR_PE	<i>low CE-statistics computed on PE-reads.</i>
STRECH_PE	<i>high CE-statistics computed on PE-reads.</i>
HIGH_SINGLE_PE	<i>high number of PE reads with unmapped pair.</i>
HIGH_SPAN_PE	<i>high number of PE reads with pair mapped in a different contig/scaffold.</i>
HIGH_OUTIE_PE	<i>high number of mis-oriented or too distant PE reads.</i>
COMPR_MP	<i>low CE-statistics computed on MP reads.</i>
STRECH_MP	<i>high CE-statistics computed on MP reads.</i>
HIGH_SINGLE_MP	<i>high number of MP reads with unmapped pair.</i>
HIGH_SPAN_MP	<i>high number of MP reads with pair mapped in a different contig/scaffold.</i>
HIGH_OUTIE_MP	<i>high number of mis-oriented or too distant MP reads.</i>

The Table provides a brief description for each implemented feature.

doi:10.1371/journal.pone.0052210.t001

Generate report and show to your boss 😊

MultiQC is a reporting tool that parses summary statistics from results and log files generated by other bioinformatics tools

Parses relevant information from log files to a HTML report file

MultiQC v1.3

General Stats

QUAST

Assembly Statistics

Number of Contigs

FastQC

Sequence Quality Histograms

Per Sequence Quality Scores

Per Base Sequence Content

Per Sequence GC Content

Per Base N Content

Sequence Length Distribution

Sequence Duplication Levels

Overrepresented sequences

Adapter Content

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2017-12-24, 14:22 based on data in: /Users/service/Box Sync/ELIXIR/Excellerate/MultiQC/reduced_data/multiqc

Welcome! Not sure where to start? [Watch a tutorial video](#) (6:06)

General Statistics

Copy table | Configure Columns | Plot | Showing 8/8 rows and 5/7 columns.

Sample Name	N50 (Kbp)	Length (Mbp)	% Dups
clean_megahit	3.4bp	29.6bp	
clean_metaspades	9.4bp	30.2bp	
sample_R1			0.0%
sample_R2			0.0%
sample_megahit	3.4bp	29.7bp	
sample_metaspades	4.0bp	29.7bp	
sample_trim_megahit	3.7bp	29.5bp	
sample_trim_metaspades	4.0bp	29.4bp	

MultiQC Toolbox

Rename Samples

From To +

Click here for bulk input.

Regex mode help

Now it is your turn to try!!!!

Metagenomic whole genome shotgun dataset from artificial marine mock sample

- Get to know the FASTQ file format – simple conversions

- Perform quality control of the sequence reads

- Merge overlapping read pairs

- Trim poor quality data

- Assemble the metagenome

- Validate the assembly

- Create a report

Practical – Day 2 - Summary

