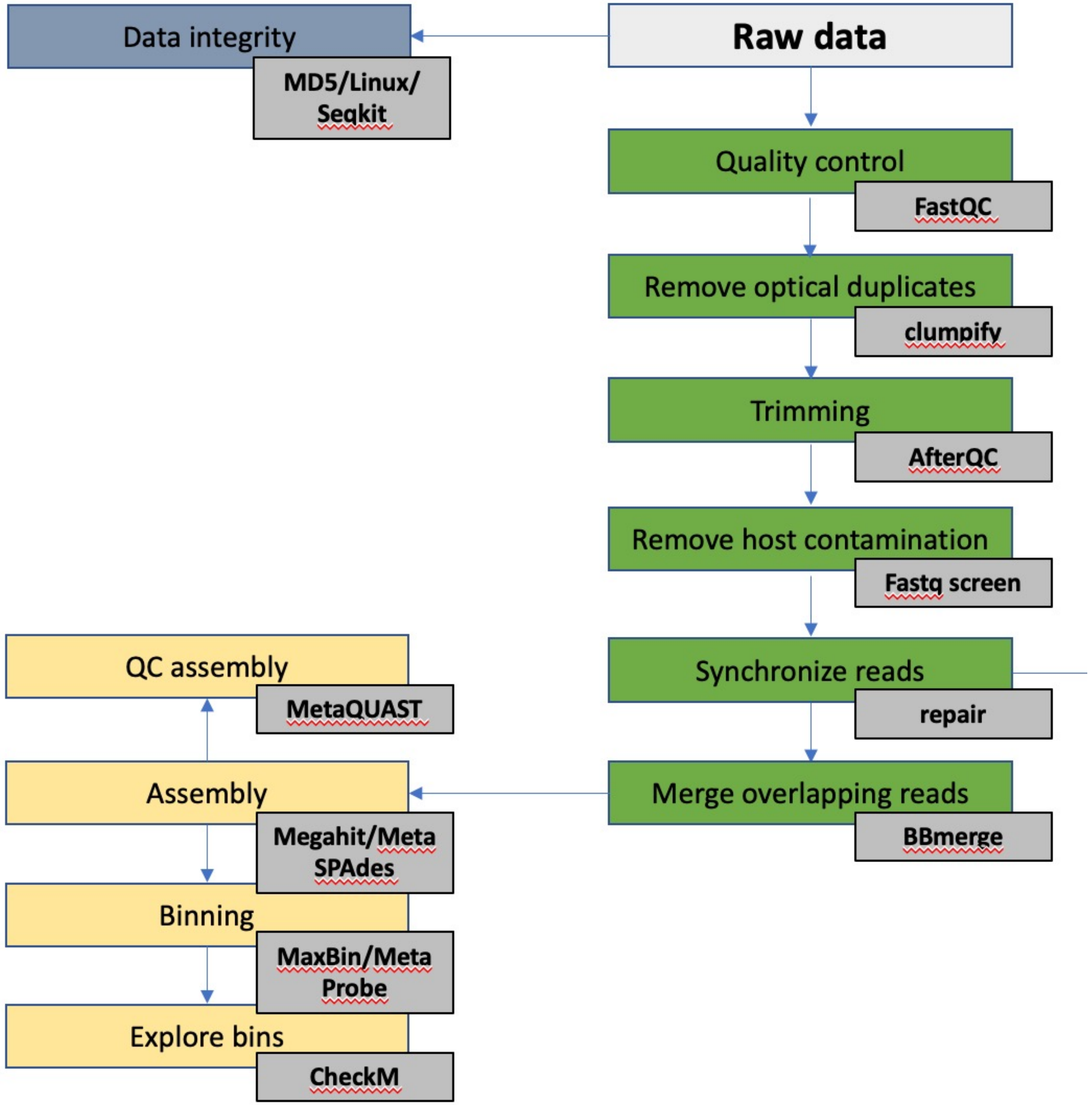
6. Assembly and validation

Overview

Obtaining a genome sequence

Metagenomic assembly

Evaluation of metagenomic assemblies



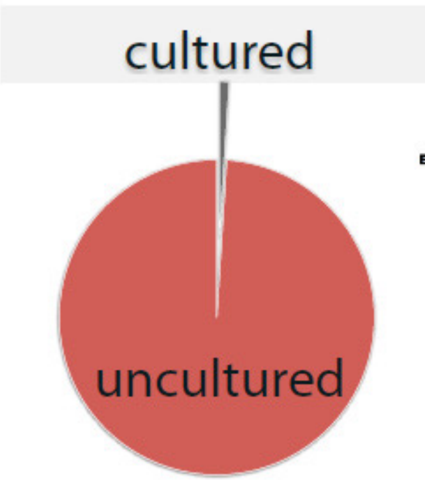
Assembly is the computational reconstruction of a longer sequence from smaller sequence reads

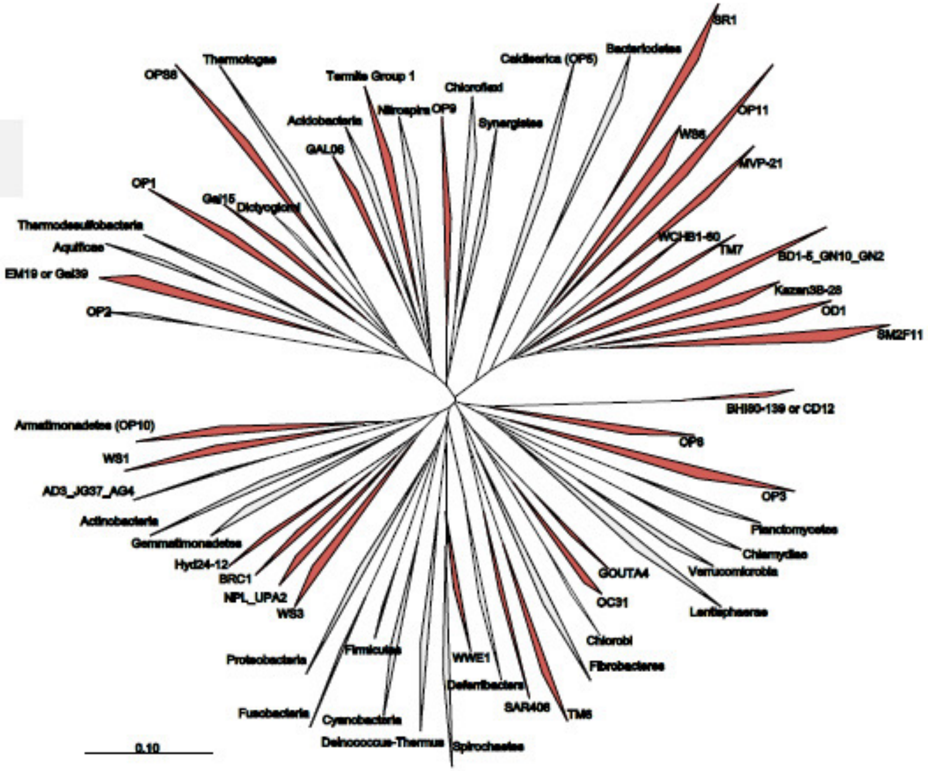
Which method should I choose that will produce the highest-quality assembly with the data that I have?

Why do we want to sequence metagenomes?

Uncultivable organisms – microbial dark matter

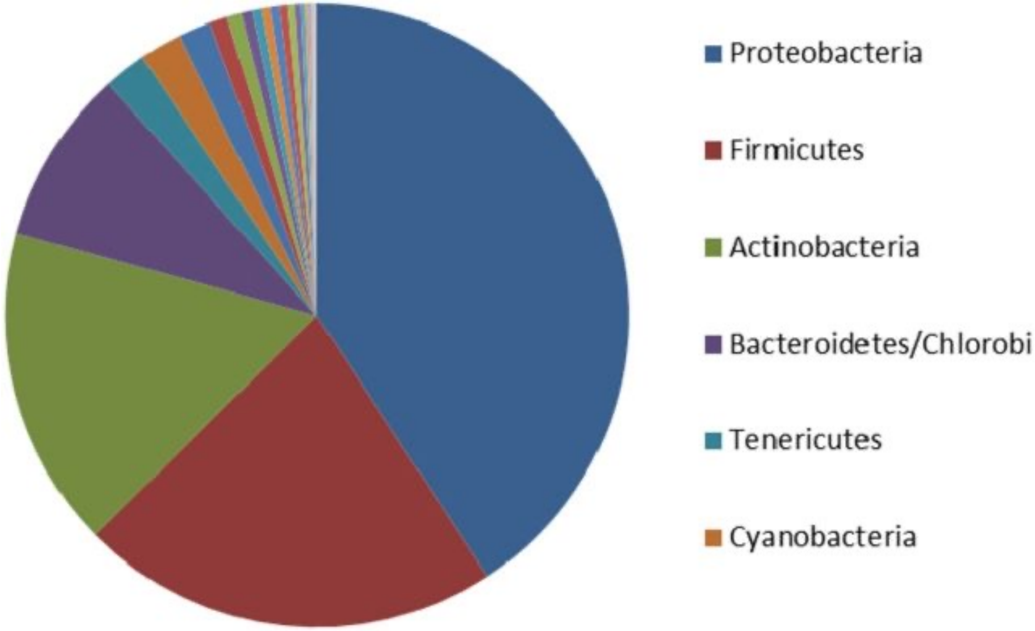
Many lineages known from 16 rRNA sequencing lacks a genomic representative





Christian Rinke / Tanja Woyke, DOE JGI

16S rRNA tree of known bacterial phyla

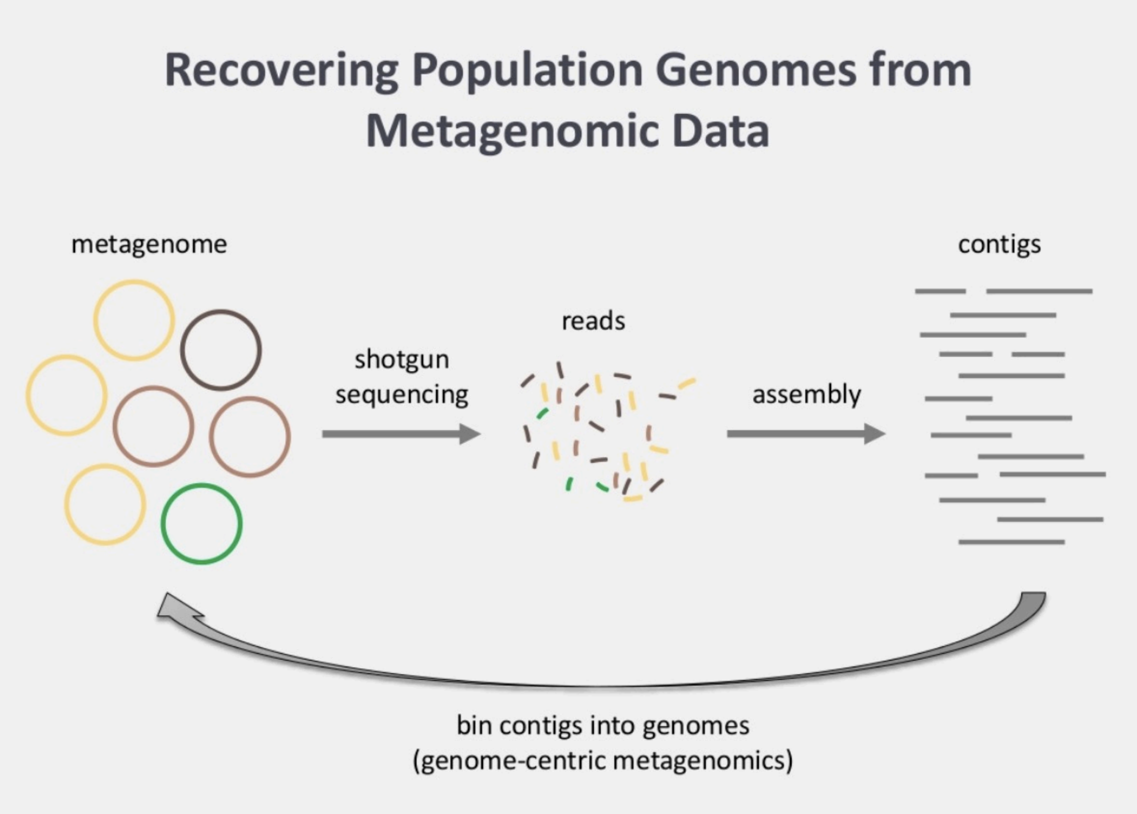


RefSeq

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


Recovering Population Genomes from Metagenomic Data
metagenome
shotgun sequencing
reads
assembly
contigs
bin contigs into genomes
(genome-centric metagenomics)

Donovan Parks, Australian school of ecogenomics

Obtaining a genome sequence

Genome sequencing => Requires culturing

Single cell sequencing => Incomplete genomes

Metagenomic sequencing => The solution?

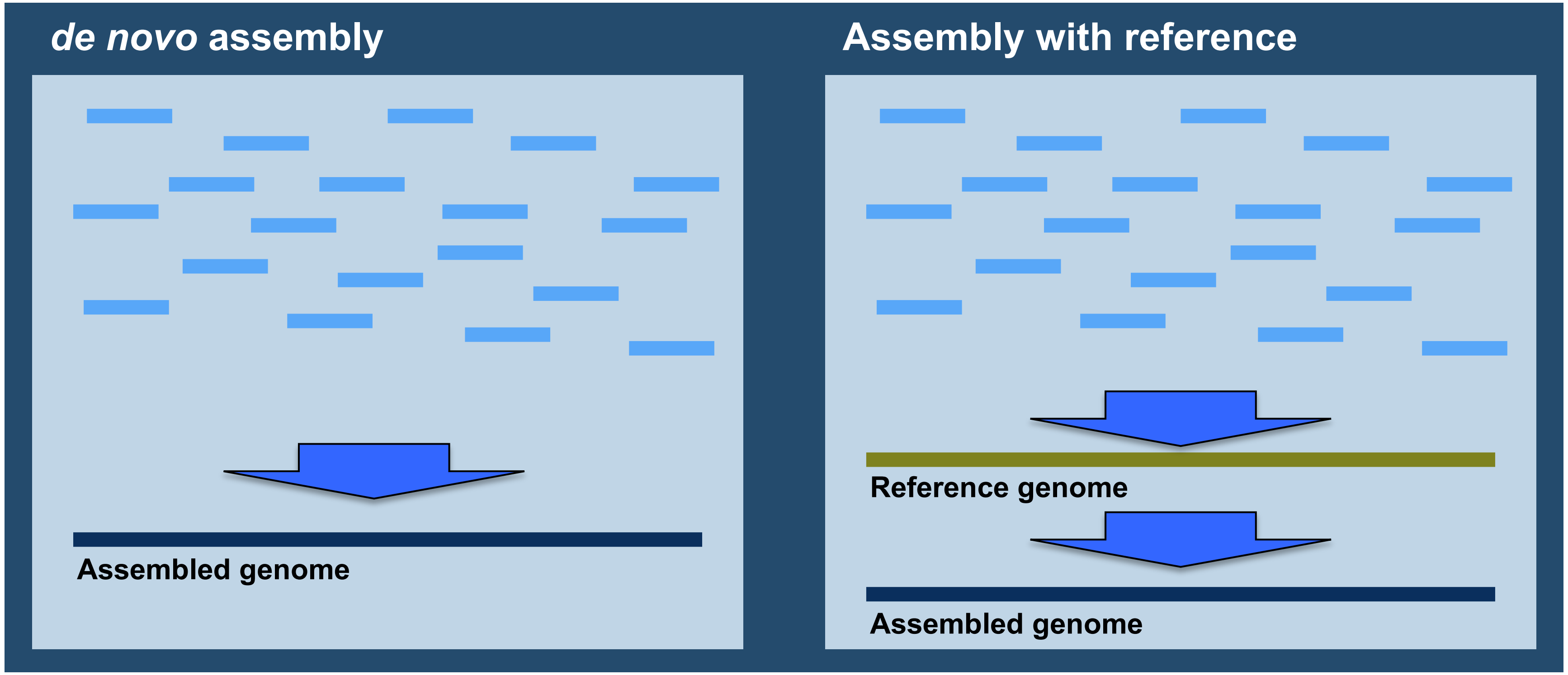
There are two approaches for sequence assembly

de novo assembly:

Reconstructing a DNA sequence with no prior knowledge of the sequence

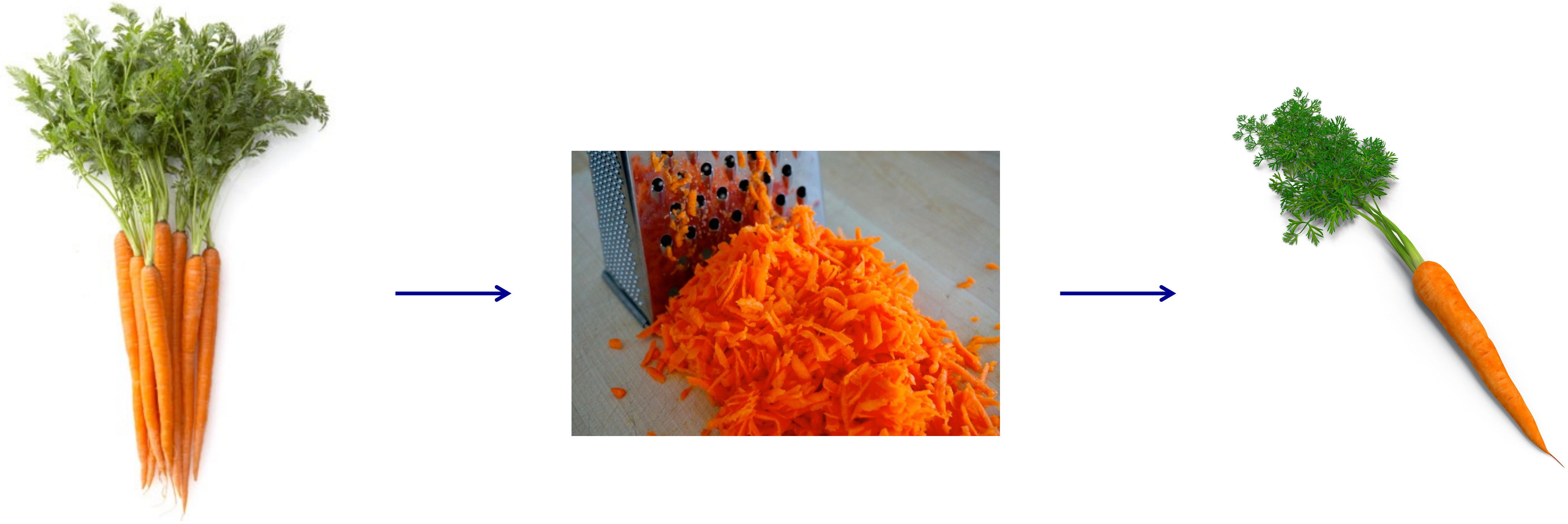
Assembly with reference sequences:

Mapping sequence reads using a reference sequence



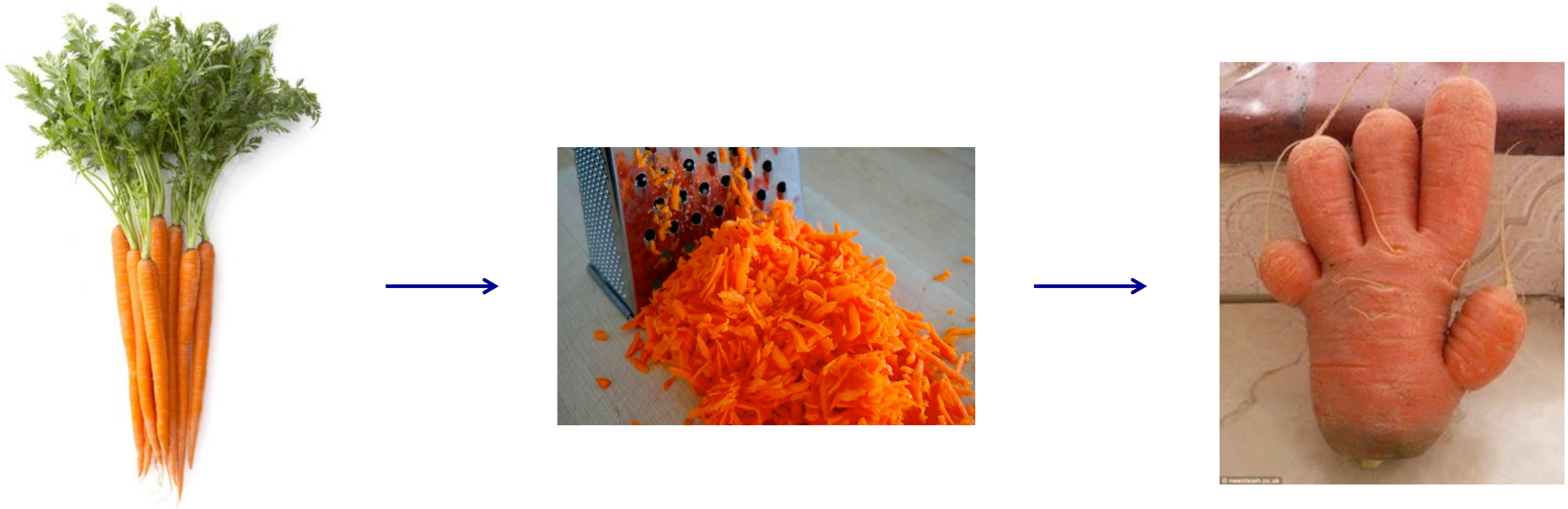
How do we perform sequence assembly of single genomes?

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?

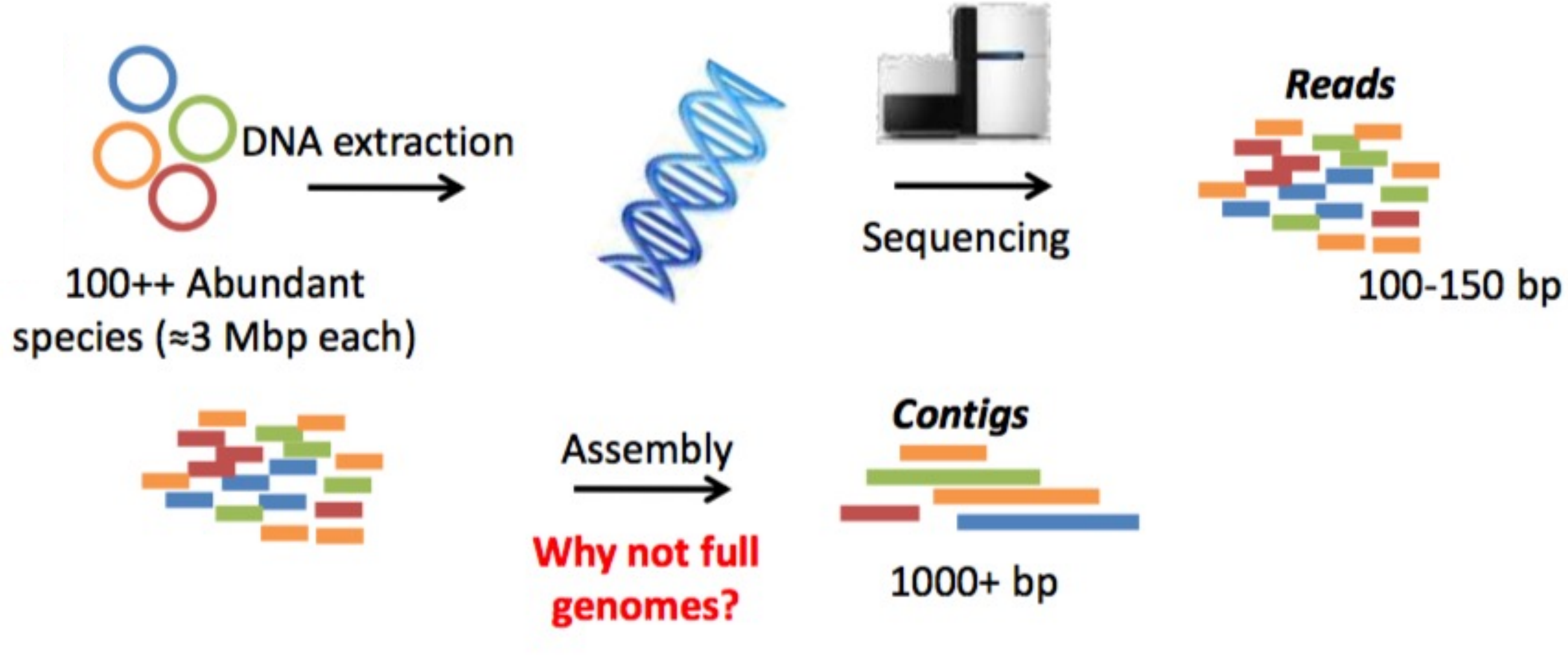
Diverse samples – more challenging as it is not possible to sequence the complete DNA

Obtaining a genome sequence from a metagenomic sample

Metagenomic Assembled Genomes (MAGs)

Similar as to genome sequencing

Trying to reconstruct the individual genomes of a mixture of DNA from an entire population



DNA extraction

100++ Abundant species (≈3 Mbp each)

Sequencing

Reads

100-150 bp

Assembly

Why not full genomes?

Contigs

1000+ bp

www.slideshare.net Mads Albertsen, University of Vienna

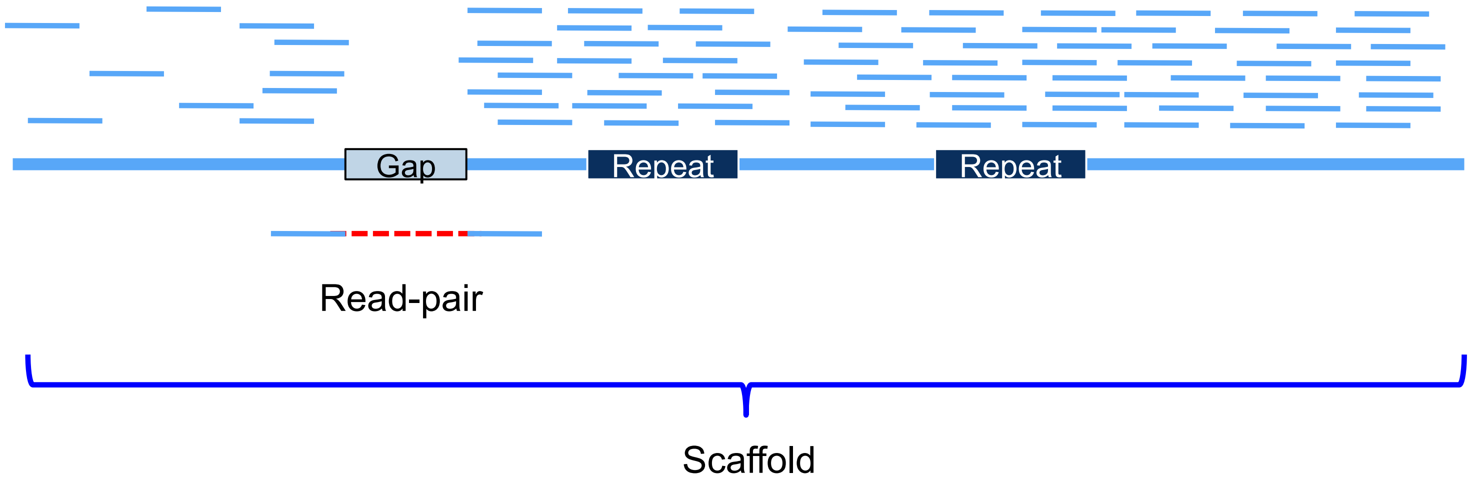
Some definitions of terms

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

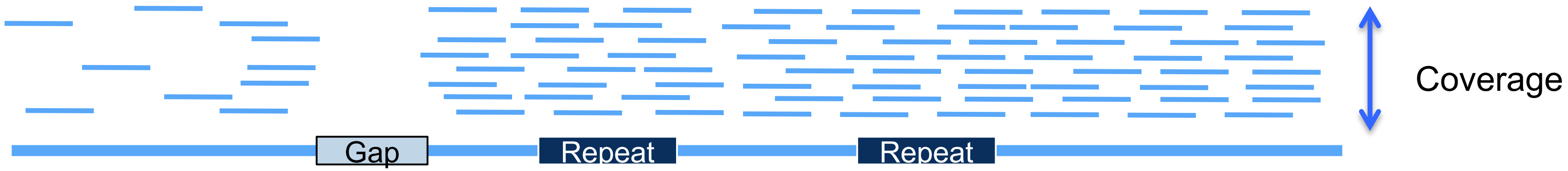
Contig = Consensus sequence of overlapping sequence reads

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



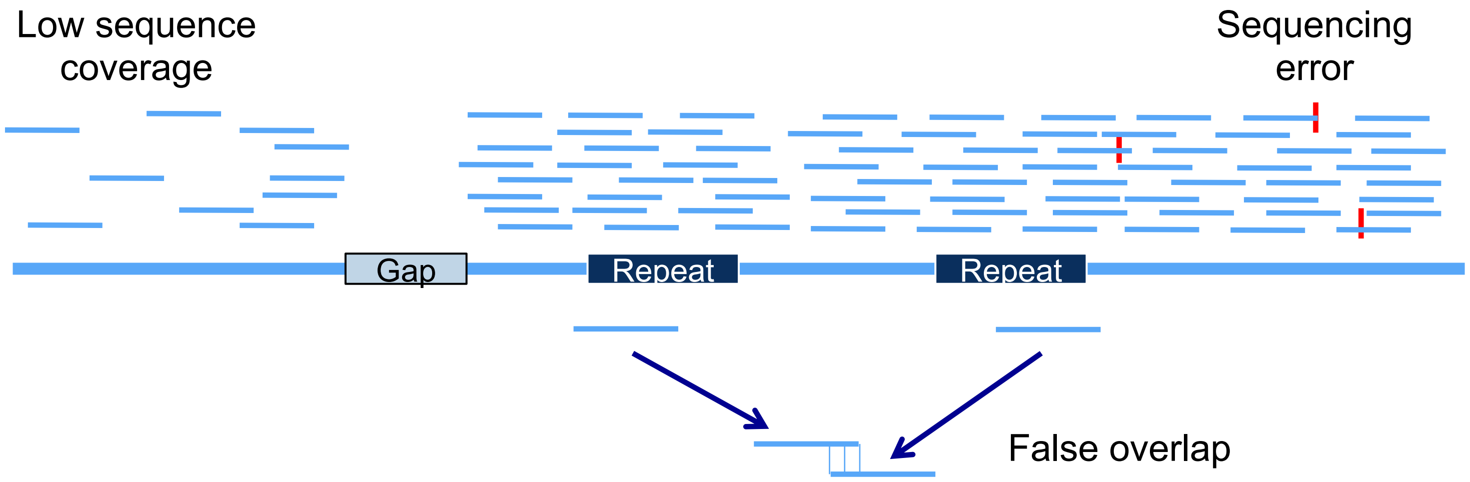
$$\frac{\text{Number of reads (n) x Length of reads (l)}}{\text{Length of metagenome (L)}}$$

Some assembly challenges

Uncovered regions

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

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

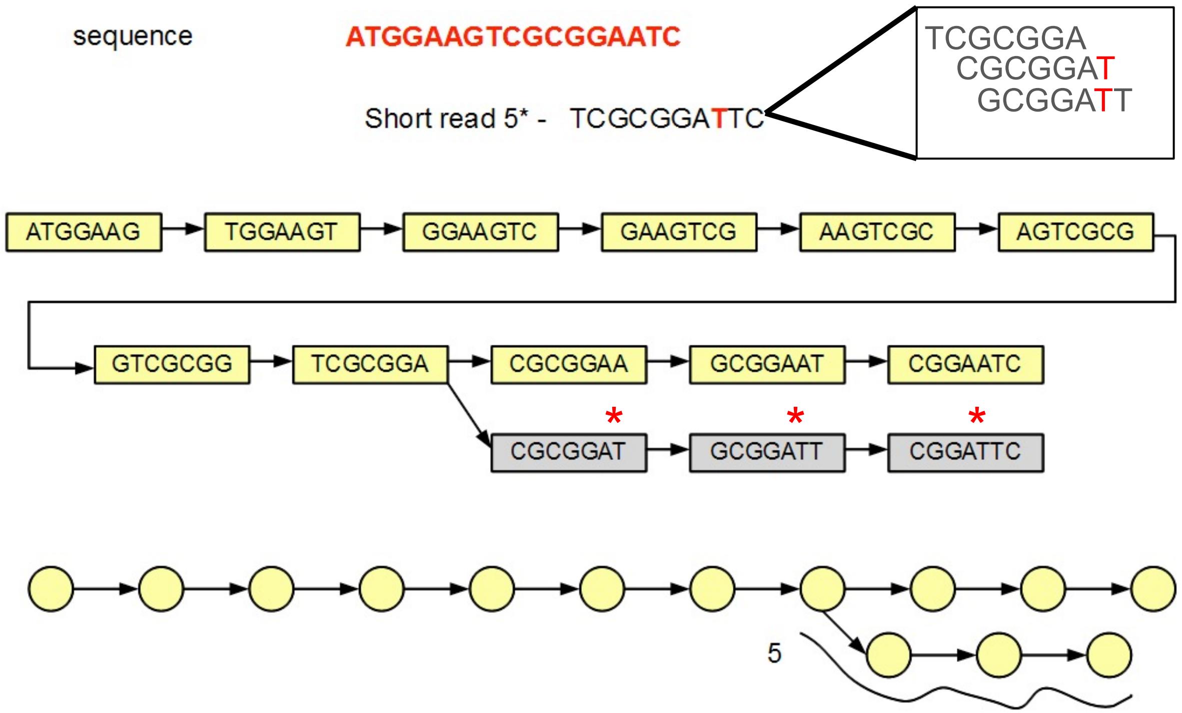


Challenges in assembly of genome sequences

Identify overlapping sequence reads or K-mers and create a graph

Challenges when there are variations or repeats

- Creates bubbles in the graph
- Split into contigs



homes.cs.washington.edu

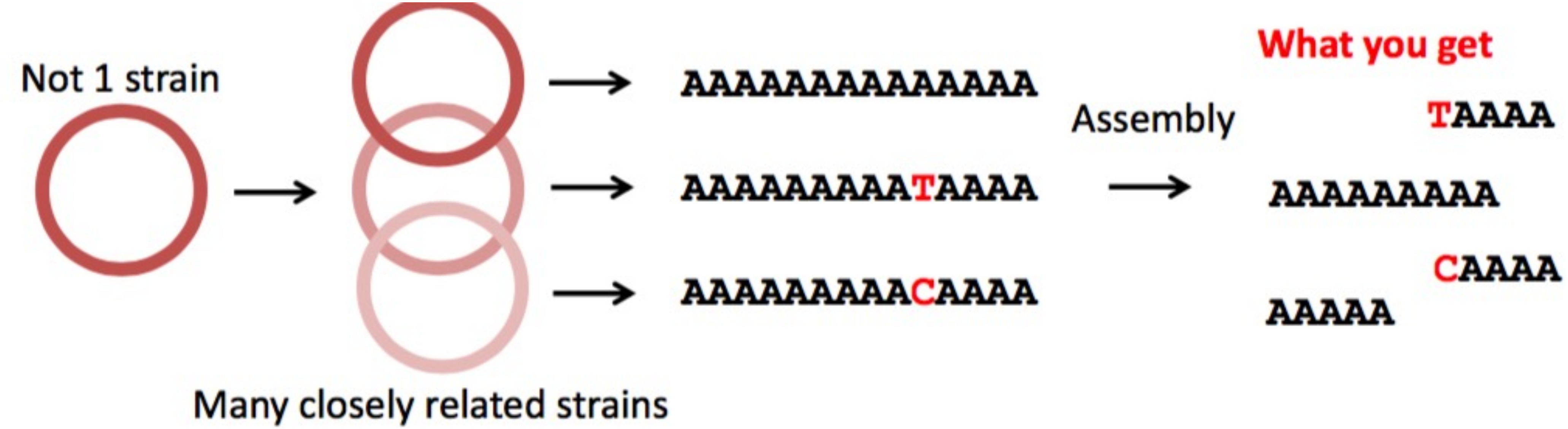
Challenges obtaining a MAGs

Very fragmented and rarely complete genomes in the sample

Highly diverse DNA (extremely many K-mers?)

- Diverse level of abundance
- Different relatedness to each other (same specie but different strains?)

Computational challenging



www.slideshare.net Mads Albertsen, University of Vienna

Challenges obtaining a MAGs

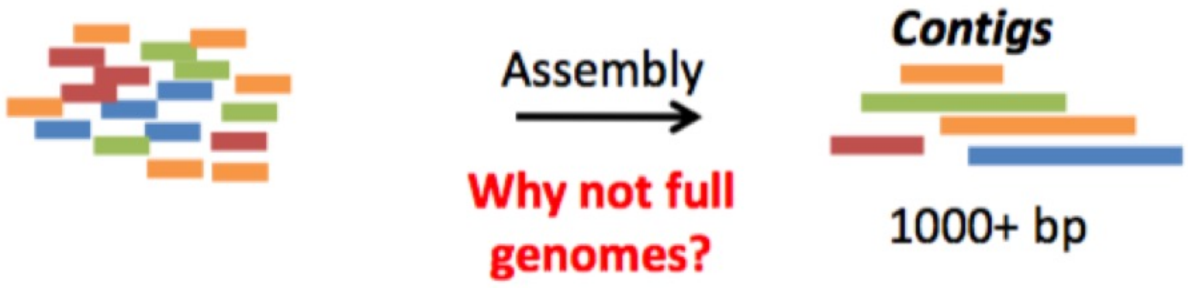
Most metagenomics assemblers use de Bruijn graphs

Algorithms for single genome assemblies cannot be used directly

Digital normalization aims to eliminate redundant reads

Partition the de Bruijun graph prior to assembly – lower memory costs

'Bubble popping' procedure. Parallel paths in the graph that differ by only a small amount, these paths are collapsed into one



- Metagenomic assemblies will still be highly fragmented - Binning

Merge overlapping paired-end reads prior to assembly

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

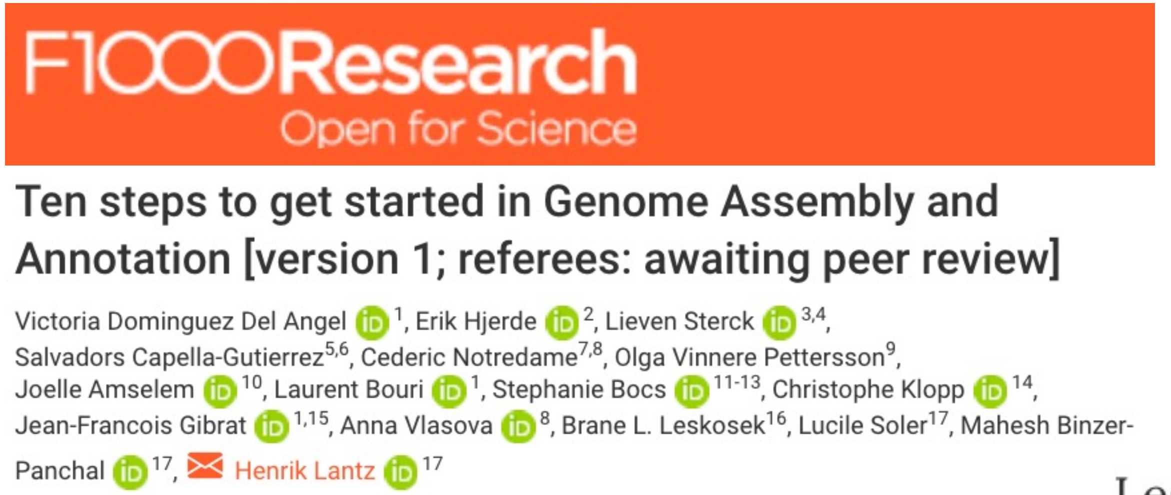
S. aureus – PE illumina	Original assembly	FLASH
Total contig size (Mb)	2.91	2.94
Contig N50 size (kb)	1.45	8.40
Contig maximum (kb)	8.18	36.07
Scaffold N50 (kb)	2.07	8.80
Scaffold maximum (kb)	11.23	36.07

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)

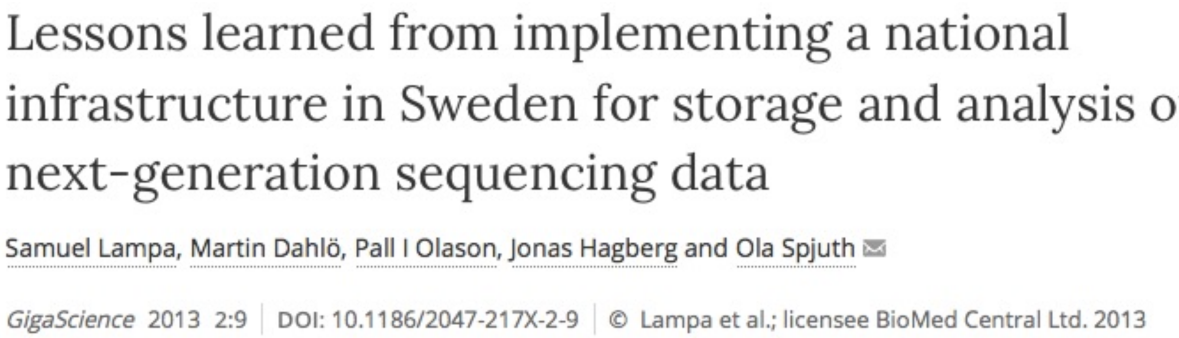
45 GB of raw sequencing data for 32 × coverage of a human genome (three Illumina HiSeq2500 runs)



F1000Research
Open for Science

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 Pettersson [9], Joelle Amselem [10], Laurent Bouri [1], Stephanie Bocs [11-13], Christophe Klopp [14], Jean-Francois Gibrat [1,15], Anna Vlasova [8], Brane L. Leskosek [16], Lucile Soler [17], 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 questions you should ask before you start sequencing

What is the purpose of sequencing the metagenome?

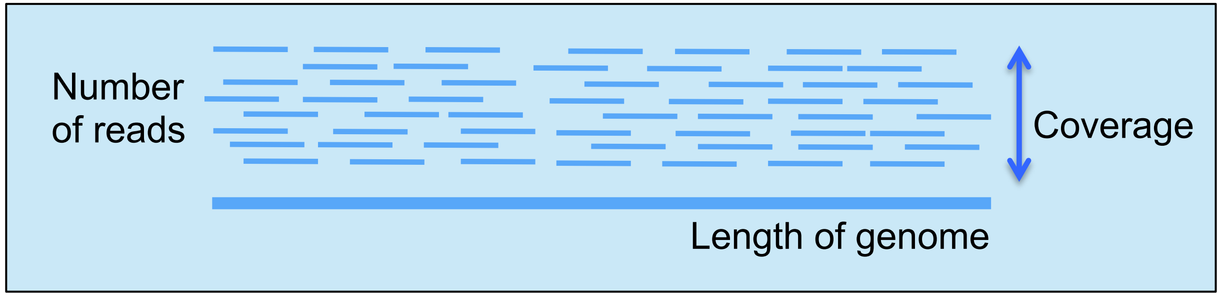
- Complete sequence (Base-perfect sequencing)
- Draft sequence

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



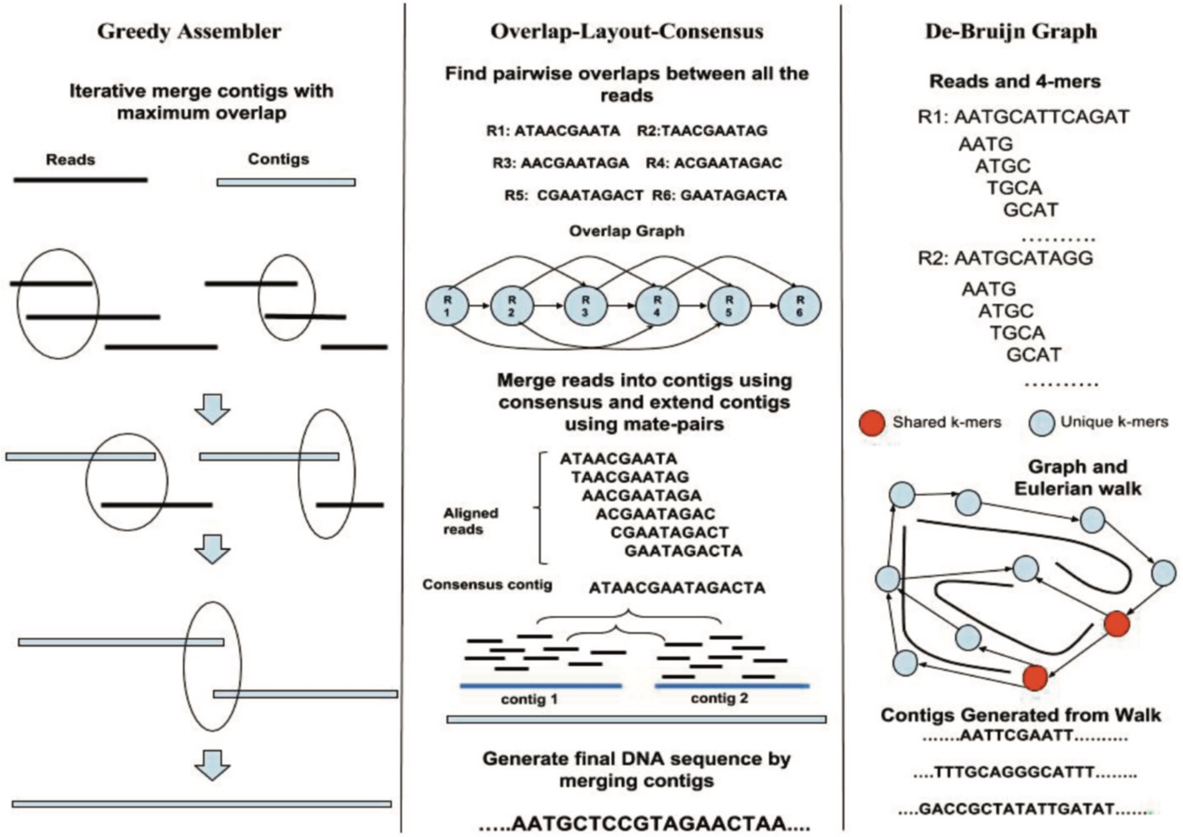
$$\text{Coverage} = \frac{\text{Number of reads x Length of read}}{\text{Length of genome}}$$

Graph-based assembly methods

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

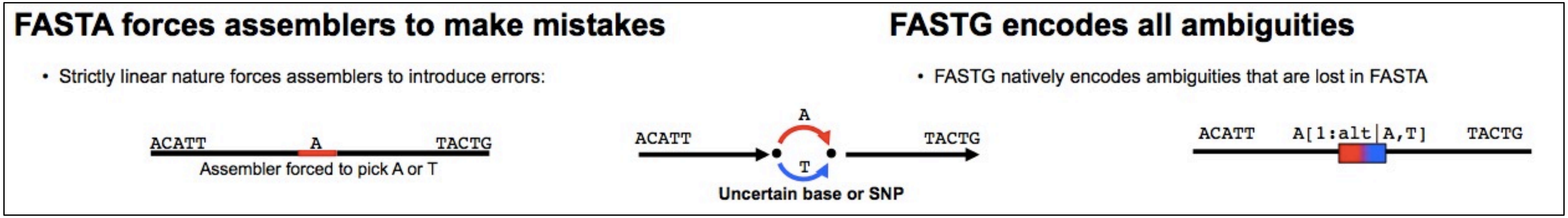
Overlap-Layout-Consensus assembly (OLC)

De Bruijn graph assembly (DBG)



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

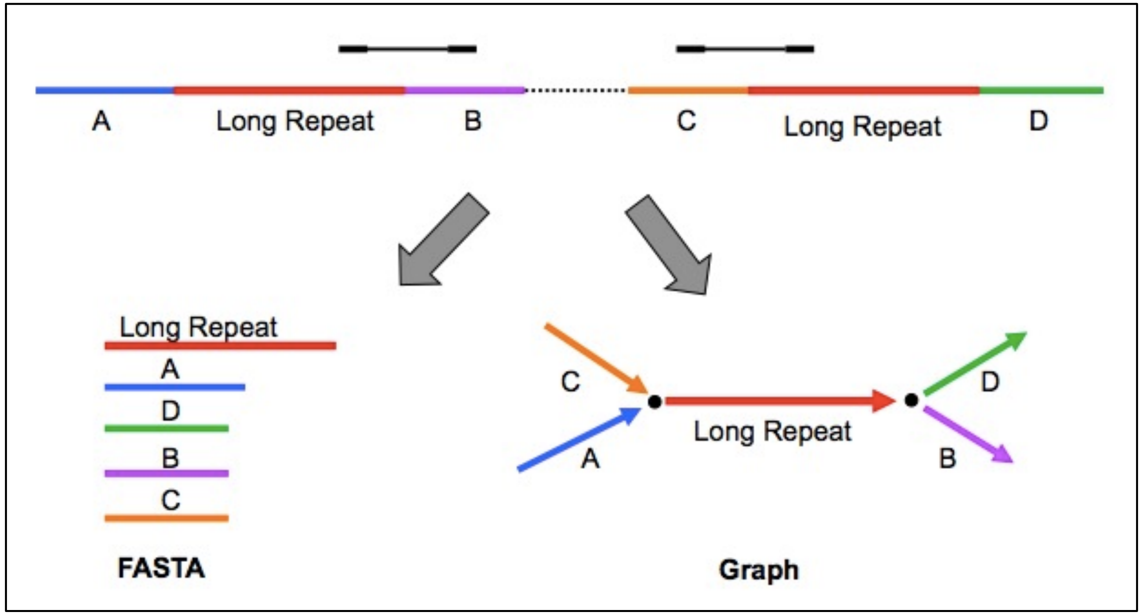


FASTA forces assemblers to make mistakes

- Strictly linear nature forces assemblers to introduce errors:

FASTG encodes all ambiguities

- FASTG natively encodes ambiguities that are lost in FASTA

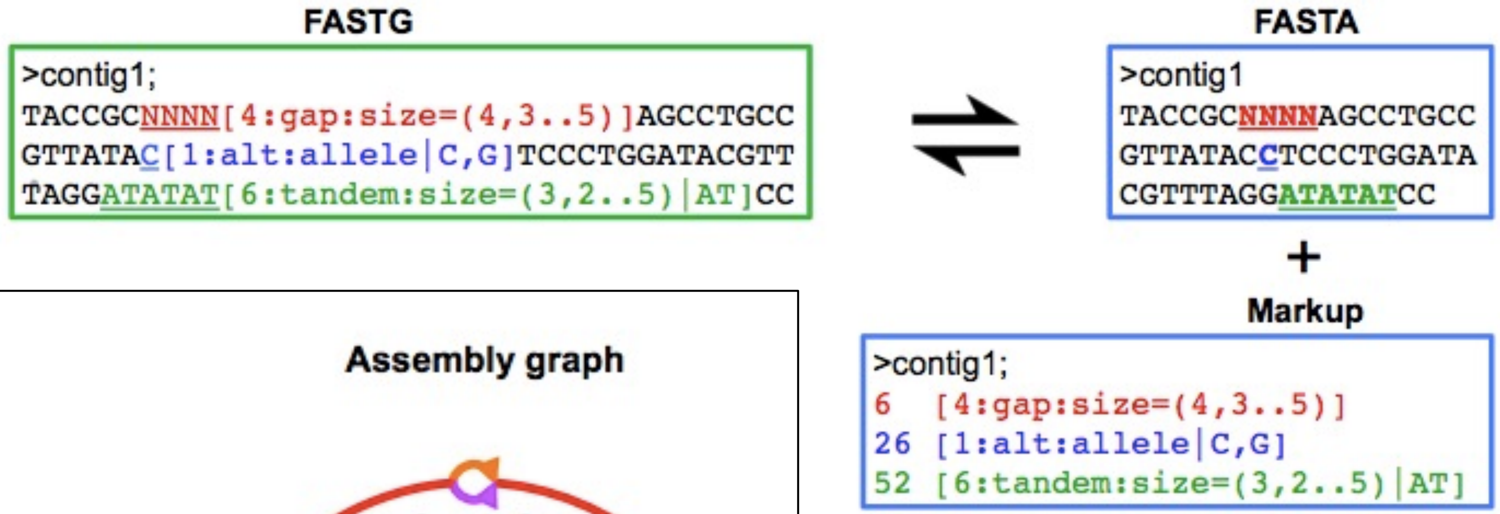


Iain MacCallum, David B. Jaffe

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
```

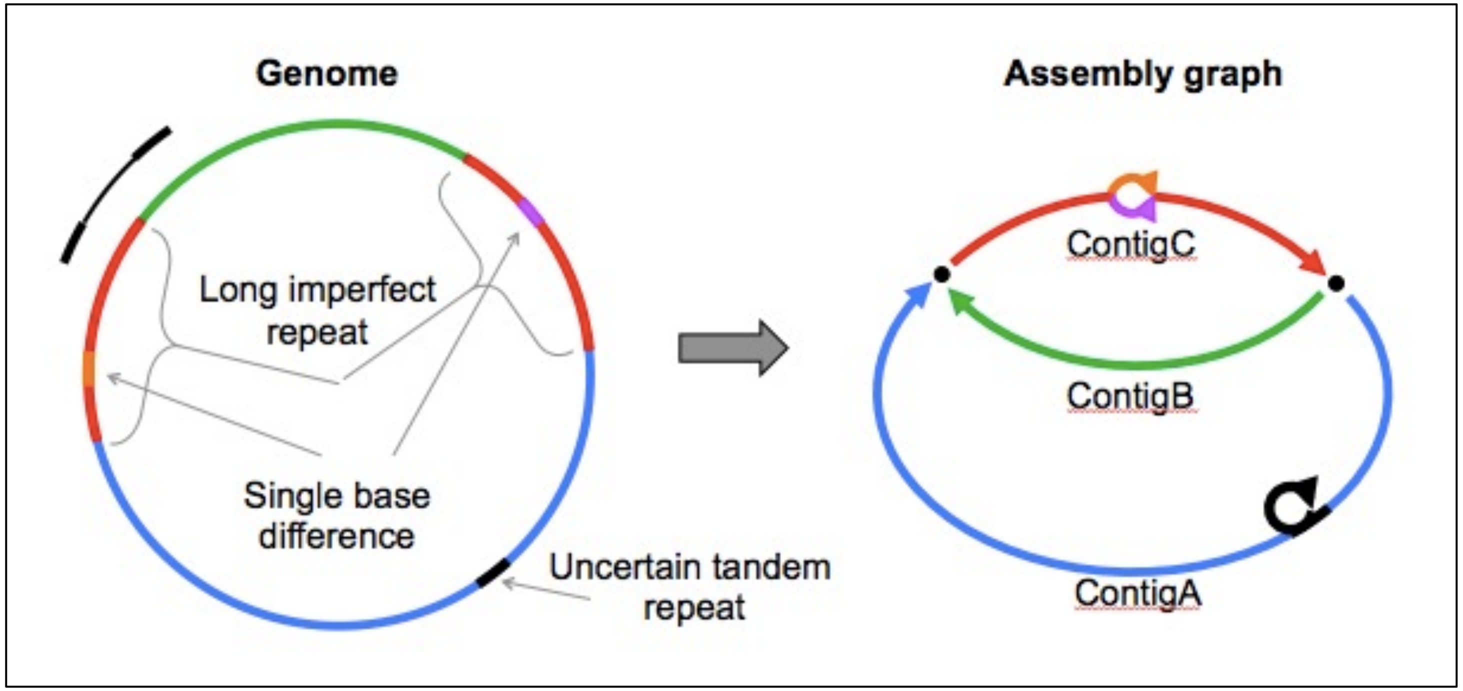
FASTA

```
>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]
```



Iain MacCallum, David B. Jaffe

Metagenome assembly tools

Megahit

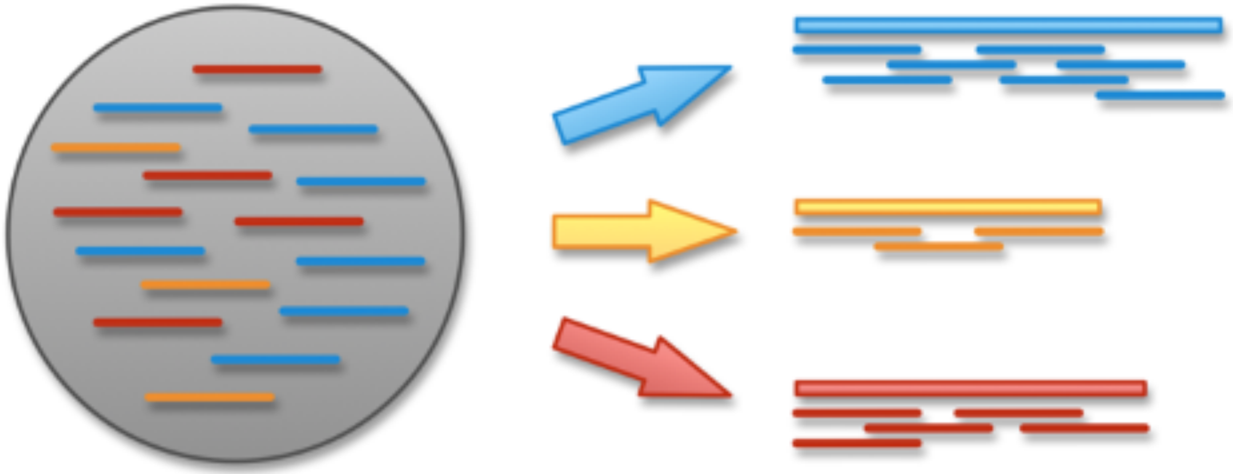
MetaSPAdes

Snowball

MetaVelvet

Ray Meta

MetAMOS

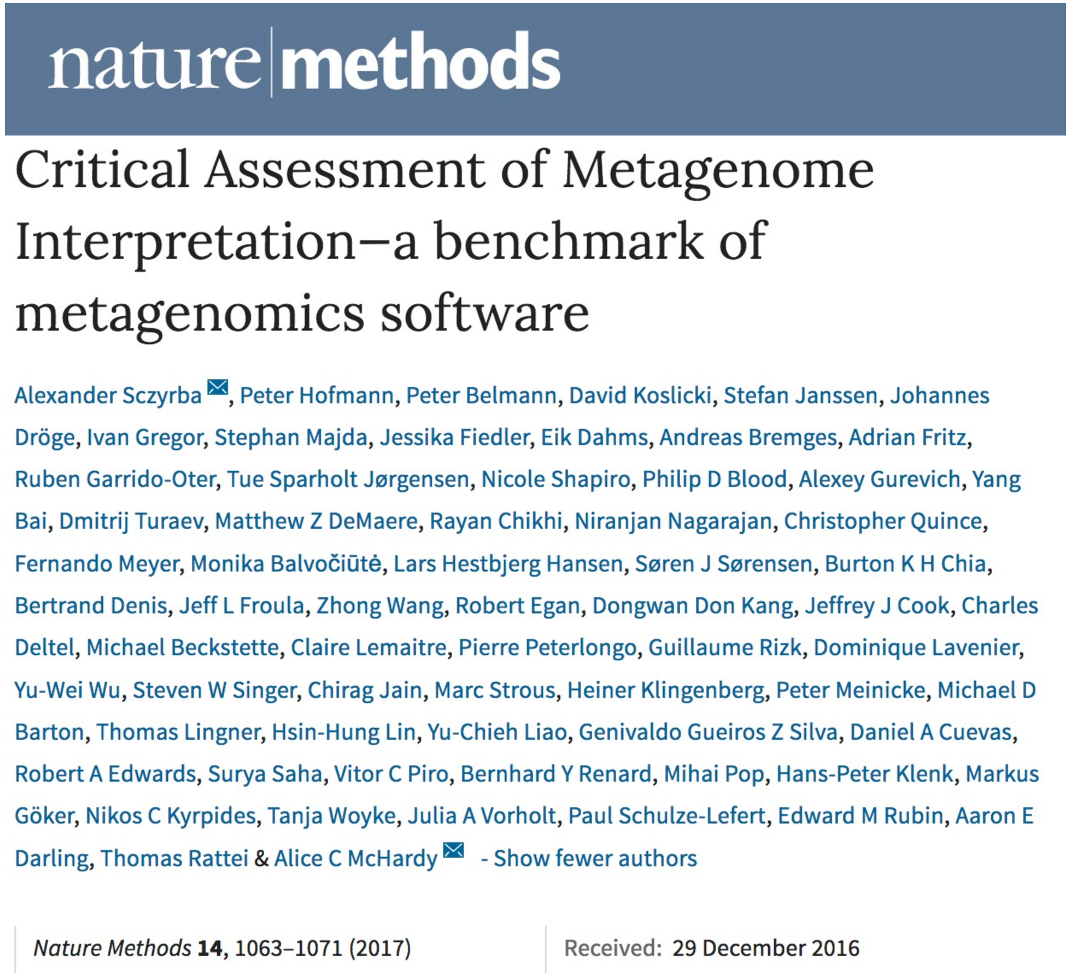


Andreas Bremges

Metagenome assembly tools - performance

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, Niranjan 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

Nature Methods **14**, 1063–1071 (2017) | Received: 29 December 2016

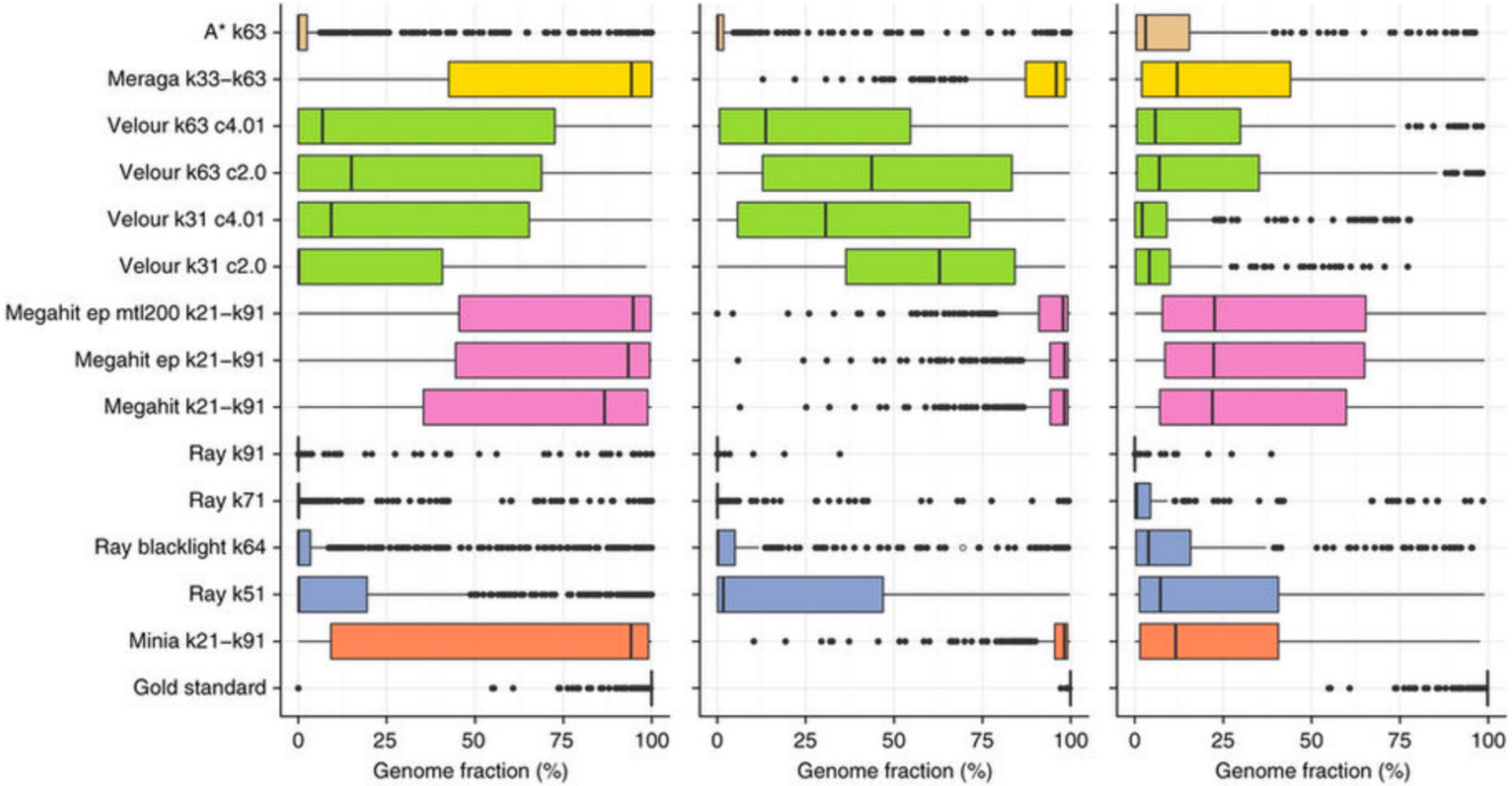
Metagenome assembly tools - performance

Main conclusion:

Assembly is substantially affected by the presence of related strains

Parameter settings markedly affected performance

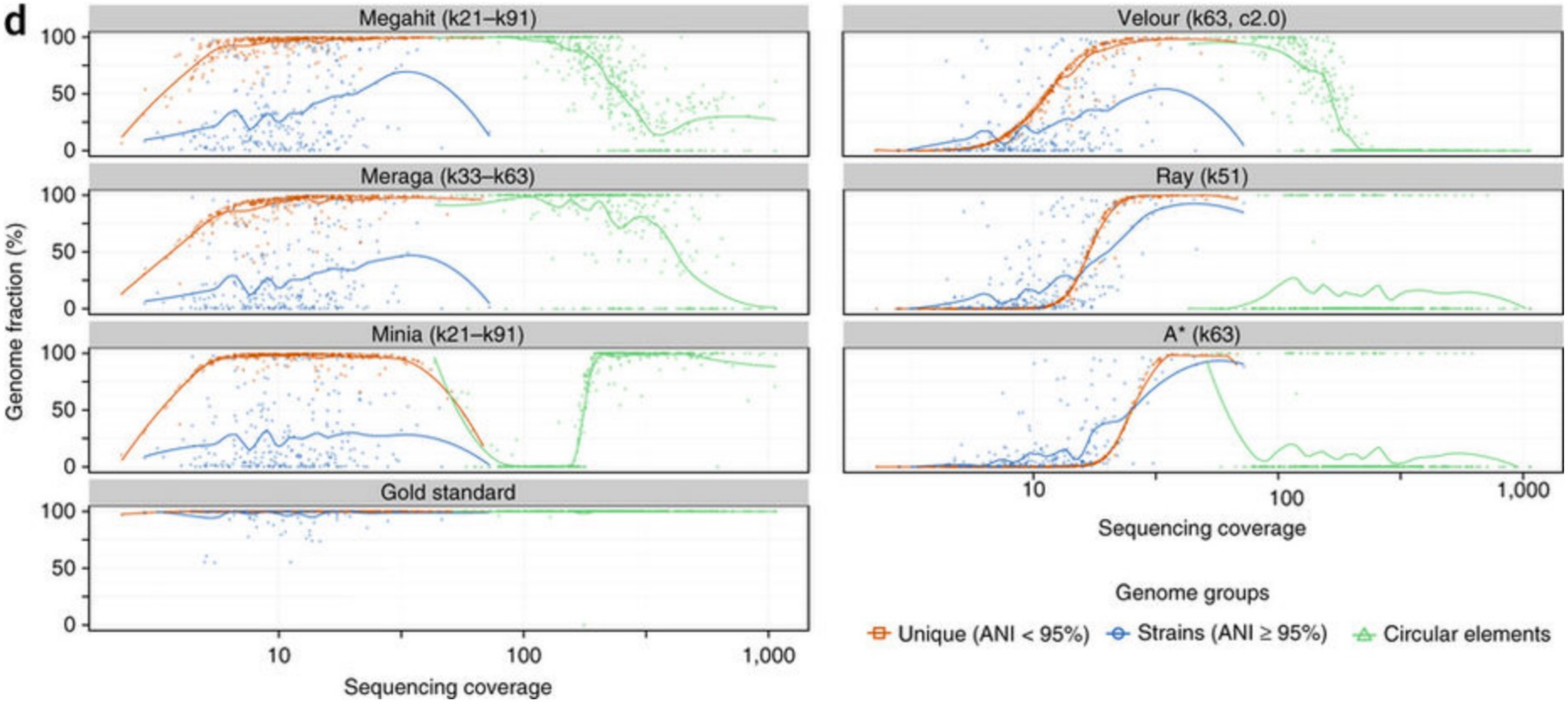
Assemblers using multiple k-mers (Minia, MEGAHIT and Meraga) substantially outperformed single k-mer assemblers



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

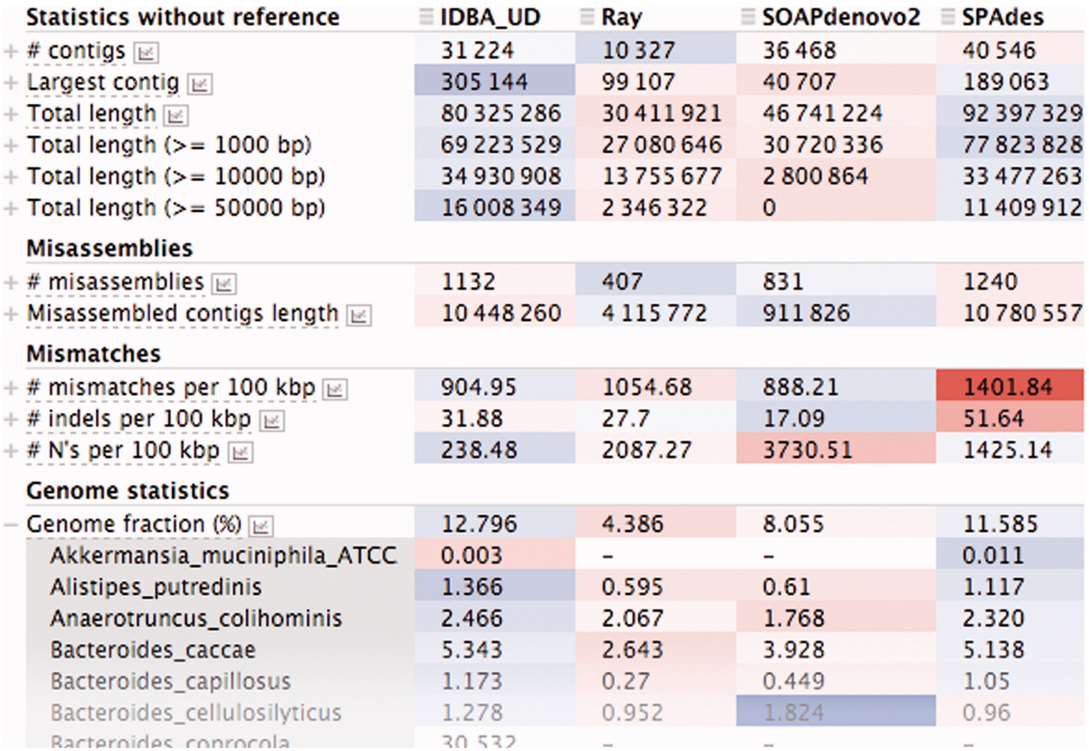
Assembly accuracy is difficult to measure!!!!

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
- Misassembly where two parts of the same contig align to distinct references
- Contigs that include both large aligned and unaligned fragments



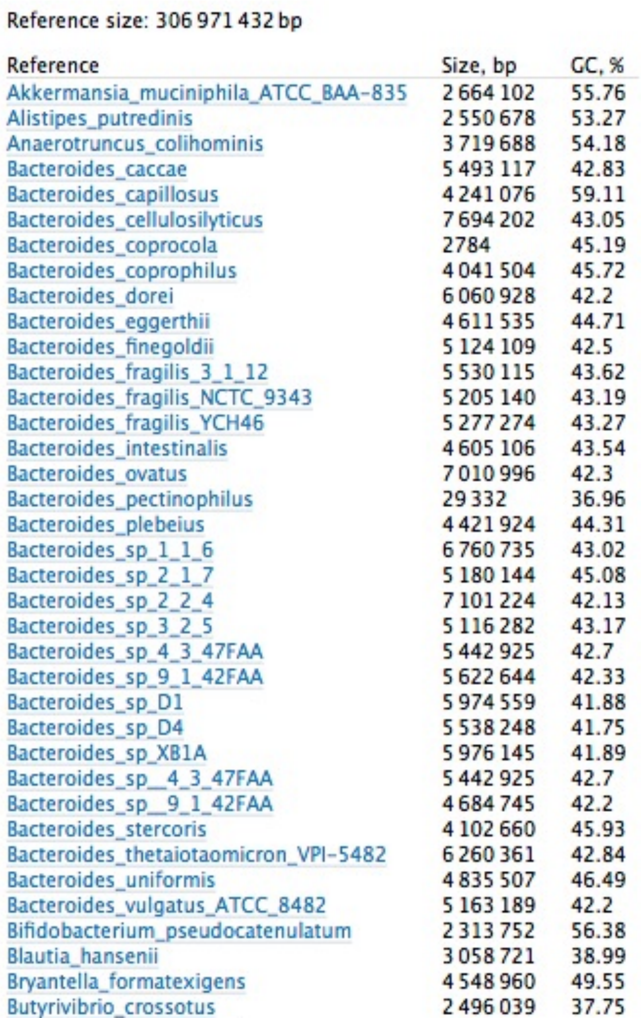
Statistics without reference	IDBA_UD	Ray	SOAPdenovo2	SPAdes
+ # contigs	31 224	10 327	36 468	40 546
+ Largest contig	305 144	99 107	40 707	189 063
+ Total length	80 325 286	30 411 921	46 741 224	92 397 329
+ Total length (>= 1000 bp)	69 223 529	27 080 646	30 720 336	77 823 828
+ Total length (>= 10000 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	–	–	–

Worst
Median
Best

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

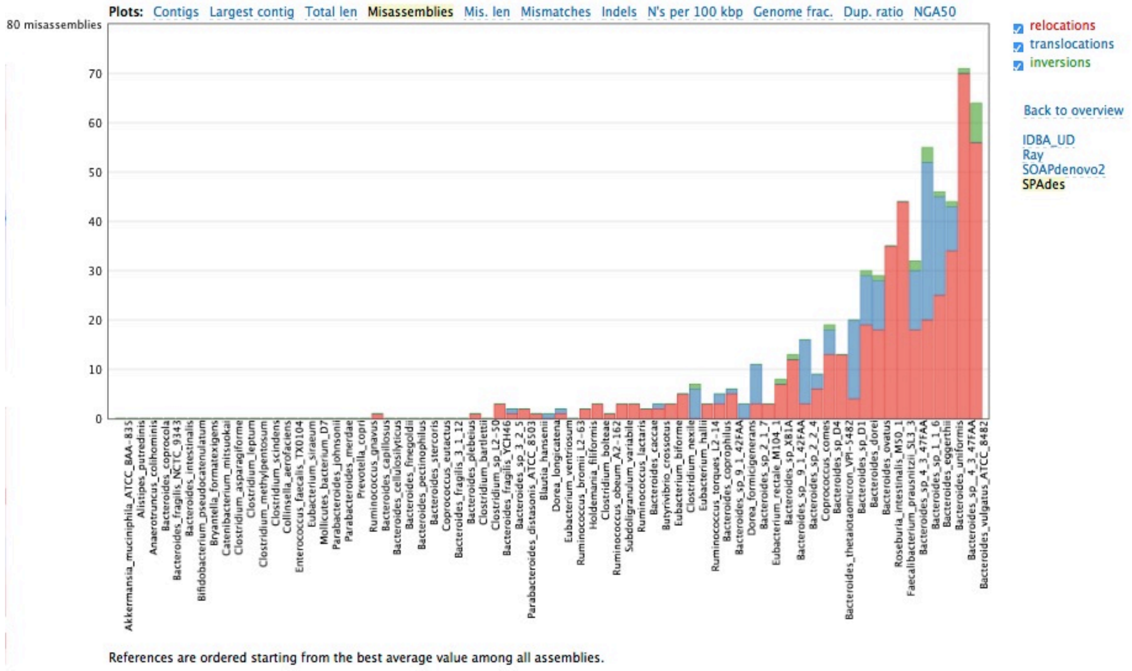
Compare the assembly from different assemblers

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	2784	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



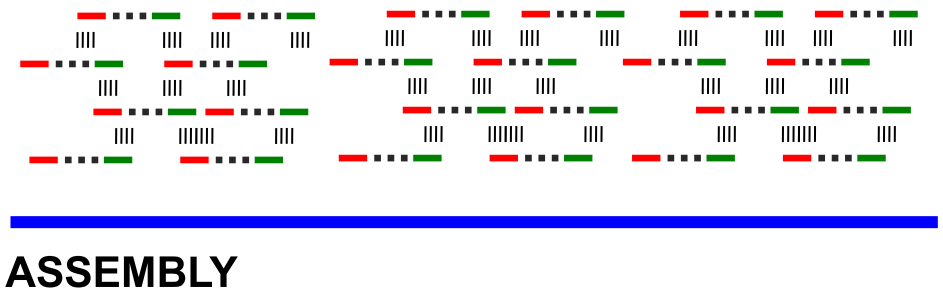
References are ordered starting from the best average value among all assemblies.

Consistency-based evaluation of sequence assemblies – In Exercise 7

Align reads against assembly of itself (not against reference)

Erroneous placement of reads within the assembly

These signatures that can be detected computationally

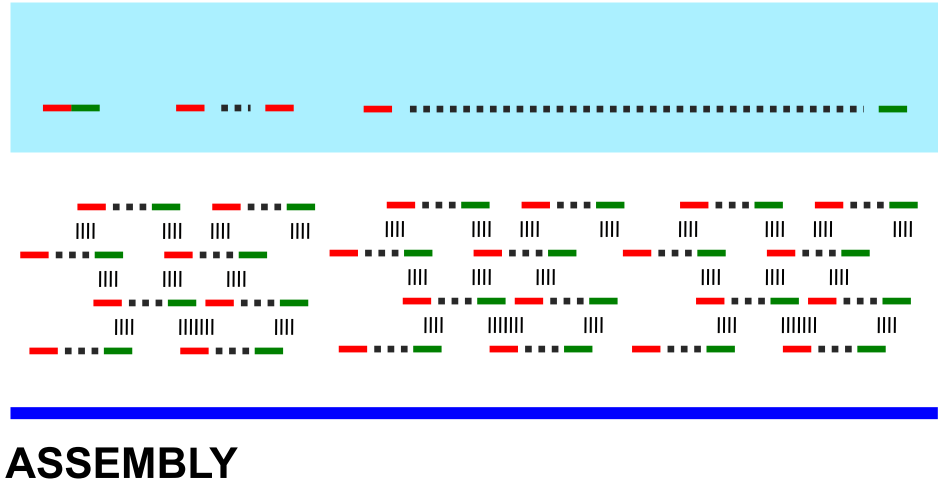


Consistency-based evaluation of sequence assemblies

Align reads against assembly of itself (not against reference)

Erroneous placement of reads within the assembly

These signatures that can be detected computationally

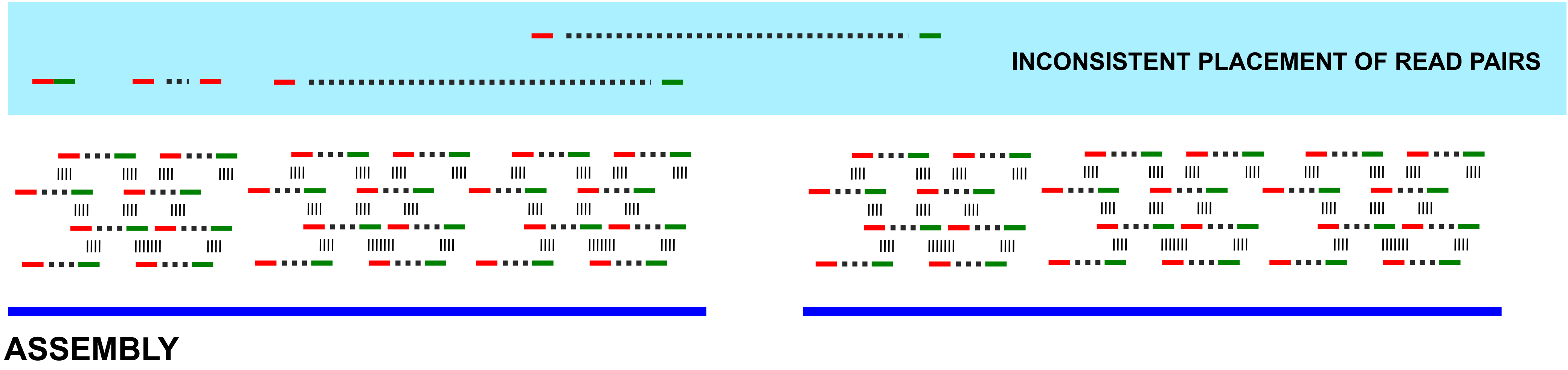


Consistency-based evaluation of sequence assemblies

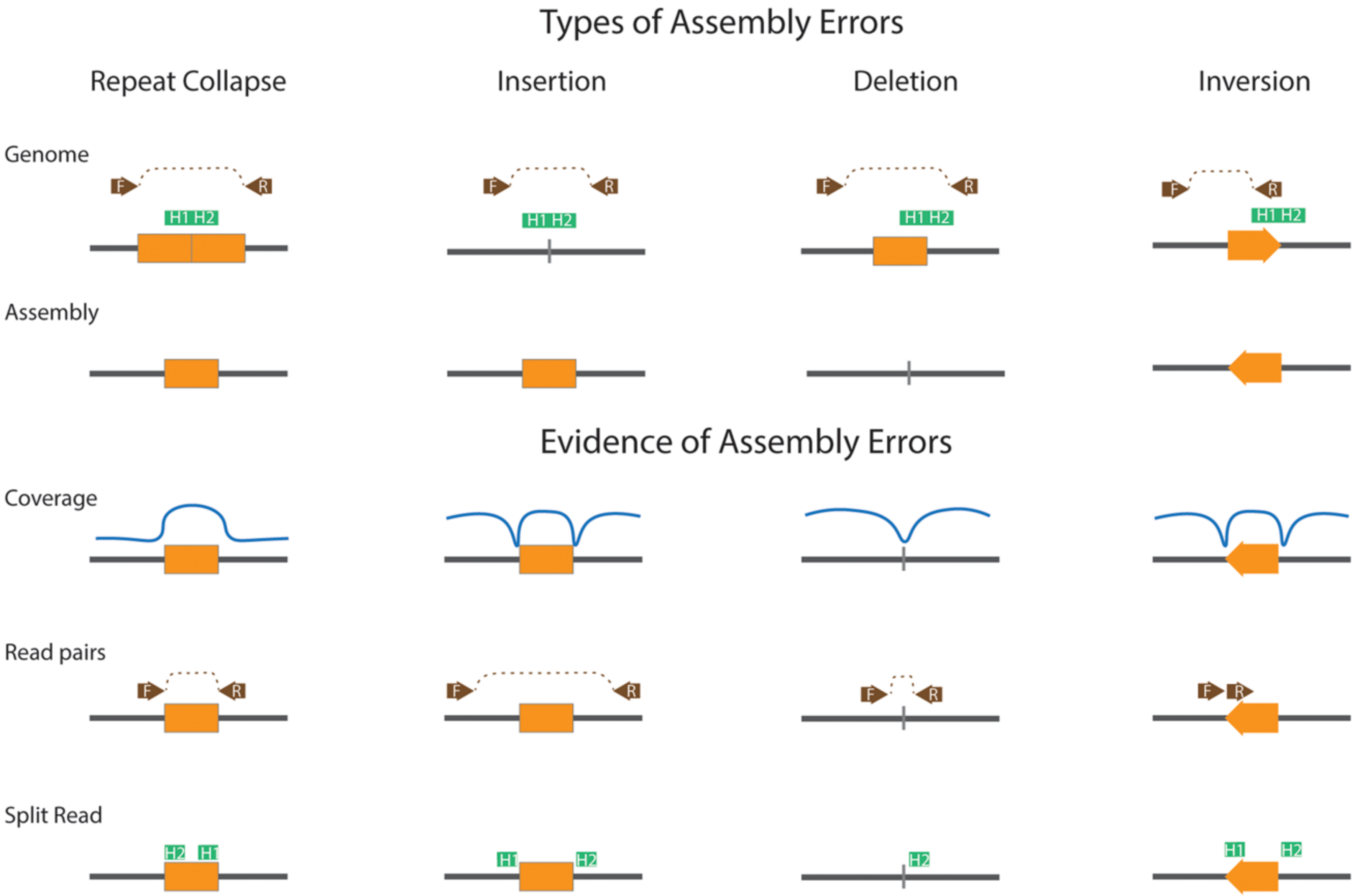
Align reads against assembly of itself (not against reference)

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

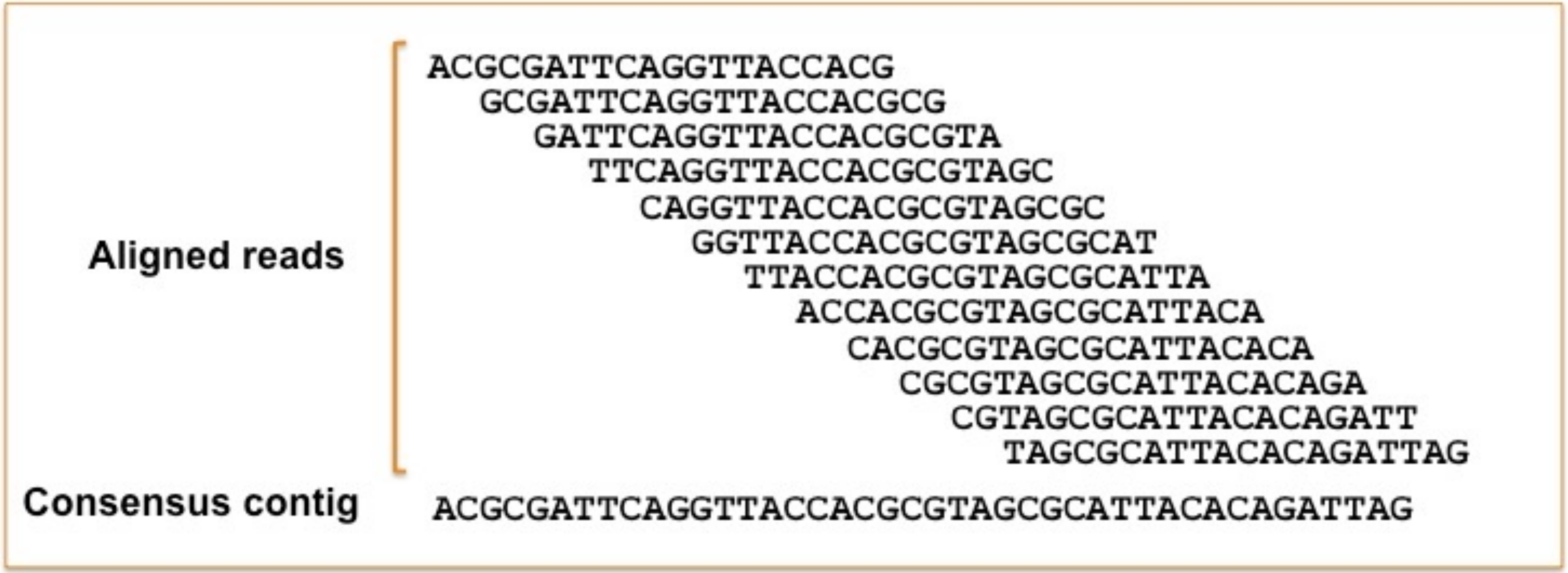


Brief Bioinform. Published online August 07, 2017. doi:10.1093/bib/bbx098

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



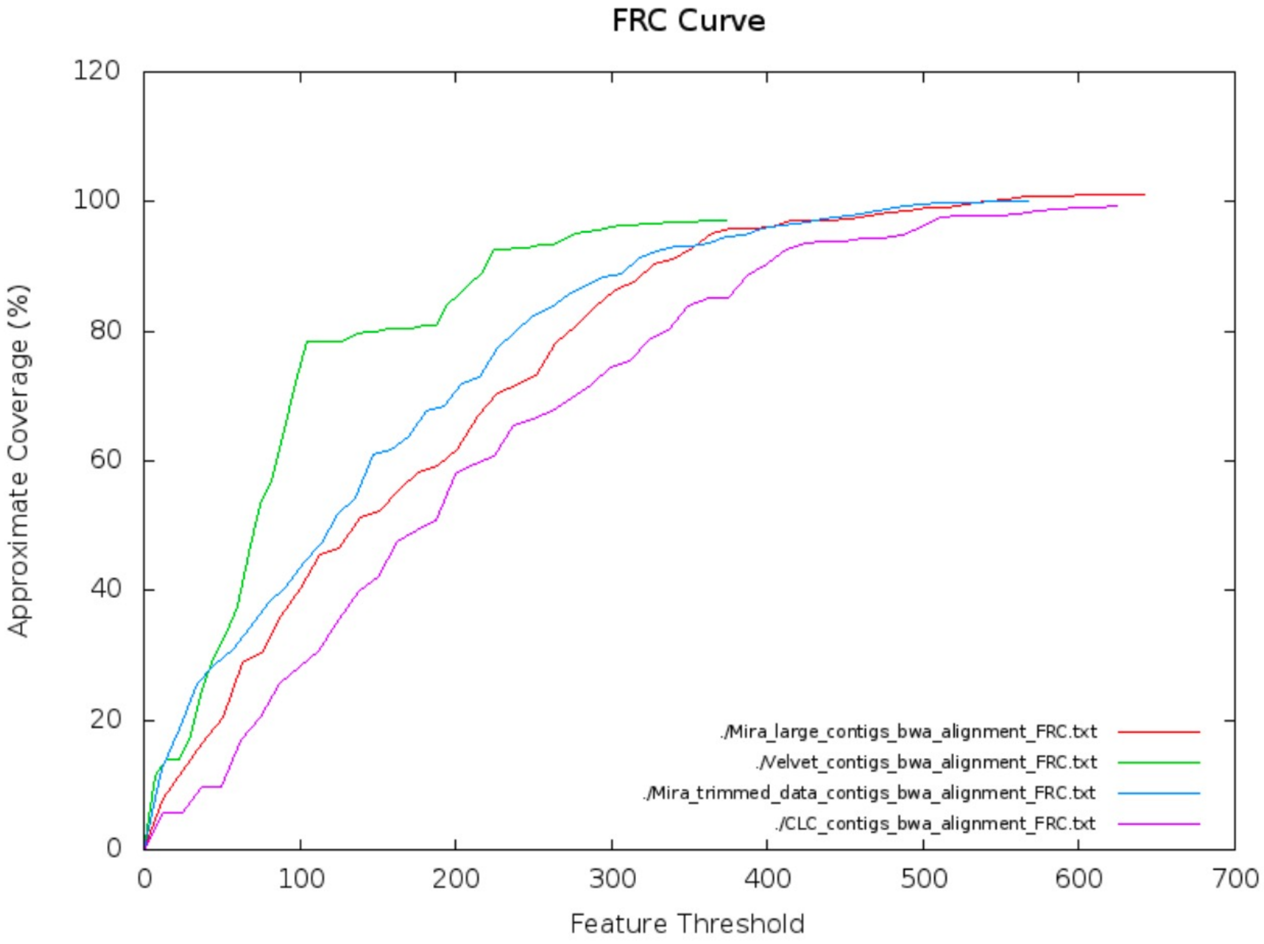
```
Aligned reads     ACGCGATTCAGGTTACCACG
                    GCGATTCAGGTTACCACGCG
                      GATTCAGGTTACCACGCGTA
                        TTCAGGTTACCACGCGTAGC
                          CAGGTTACCACGCGTAGCGC
                            GGTTACCACGCGTAGCGCAT
                              TTACCACGCGTAGCGCATTA
                                ACCACGCGTAGCGCATTACA
                                  CACGCGTAGCGCATTACACA
                                    CGCGTAGCGCATTACACAGA
                                      CGTAGCGCATTACACAGATT
                                        TAGCGCATTACACAGATTAG
Consensus contig  ACGCGATTCAGGTTACCACGCGTAGCGCATTACACAGATTAG
```

https://contig.wordpress.com/tag/alignment/

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)

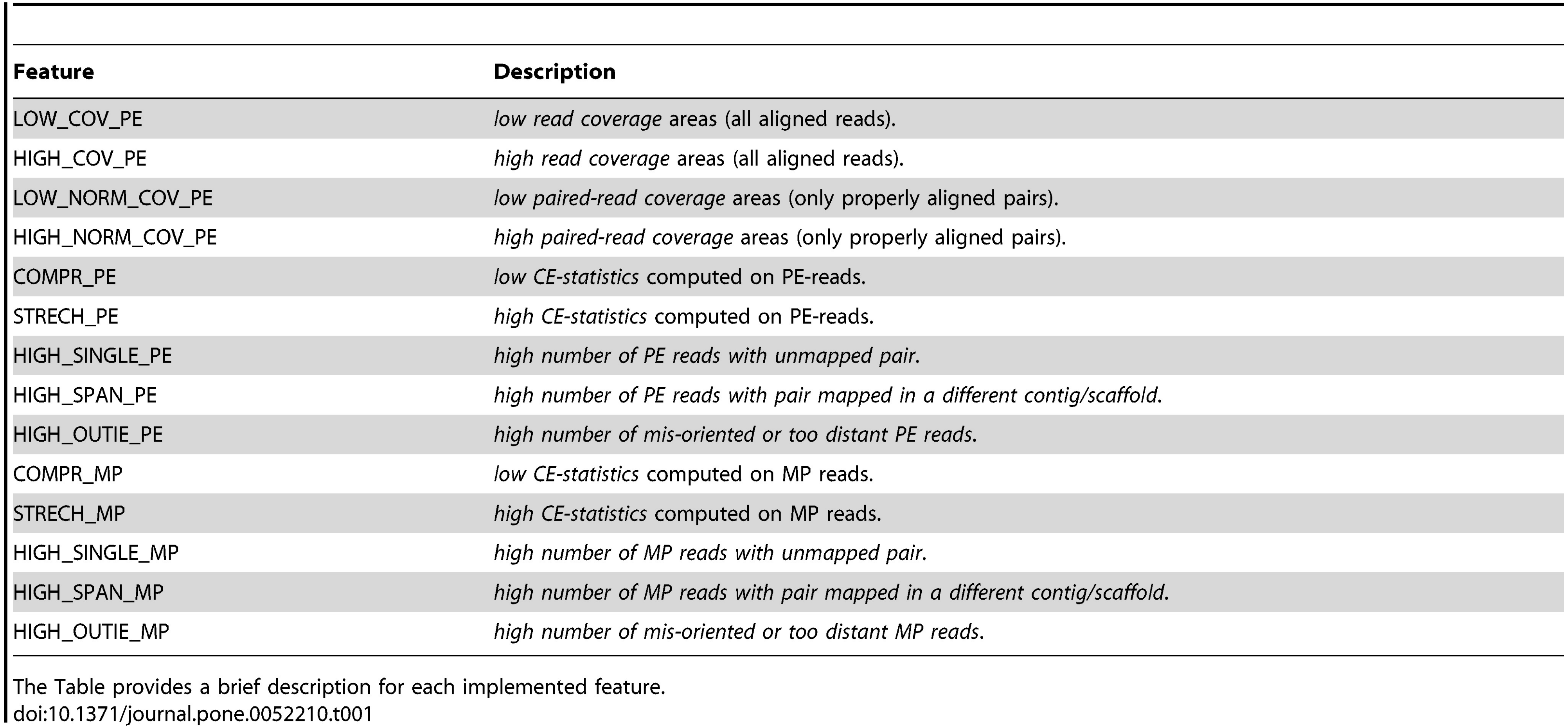
For example regions with many PE reads with pair mapped in different contigs



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

The Table provides a brief description for each implemented feature.
doi:10.1371/journal.pone.0052210.t001