

# Metagenomics 101

## Session 5: Assembly

Anna Heintz-Buschart

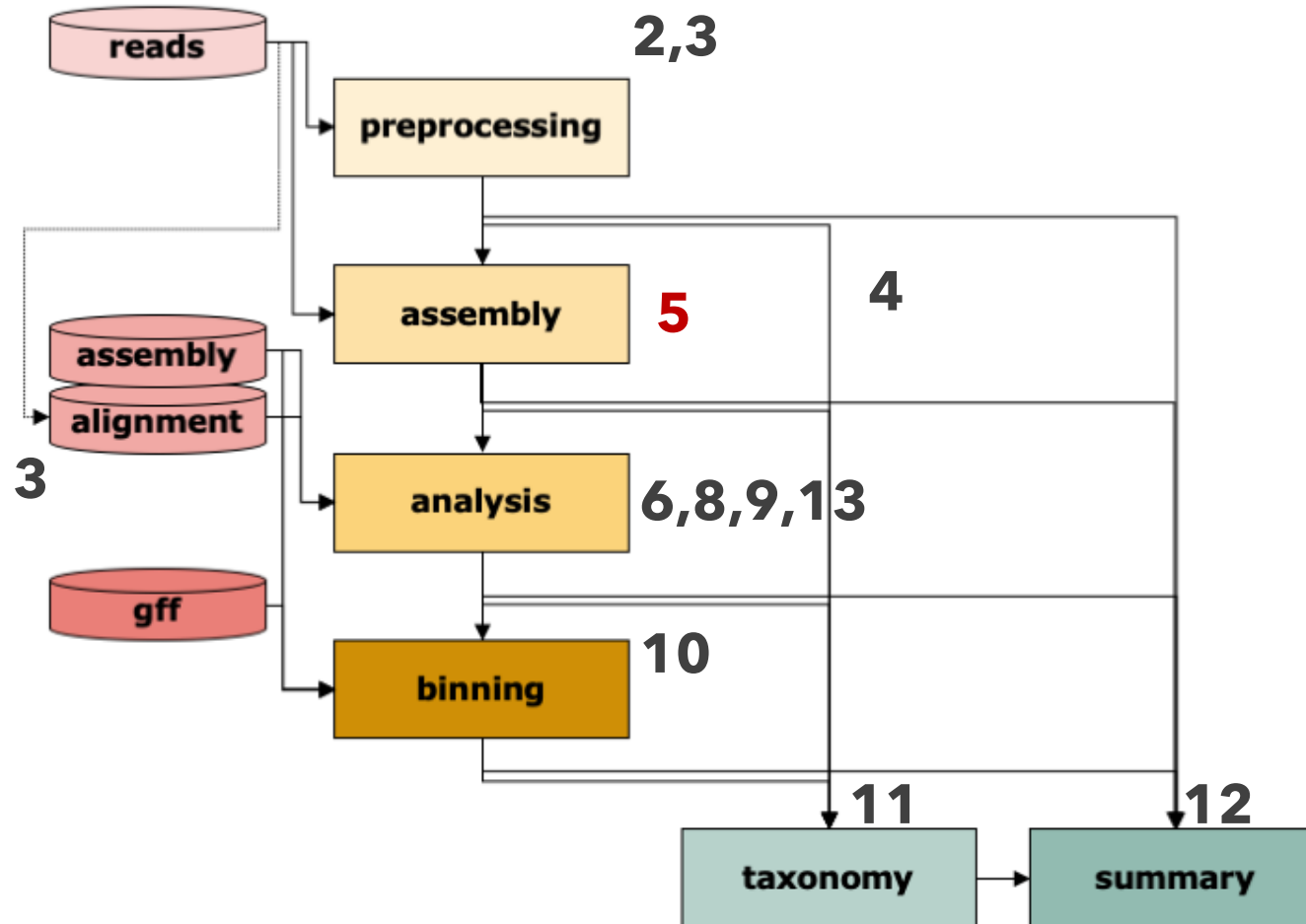
March 2022



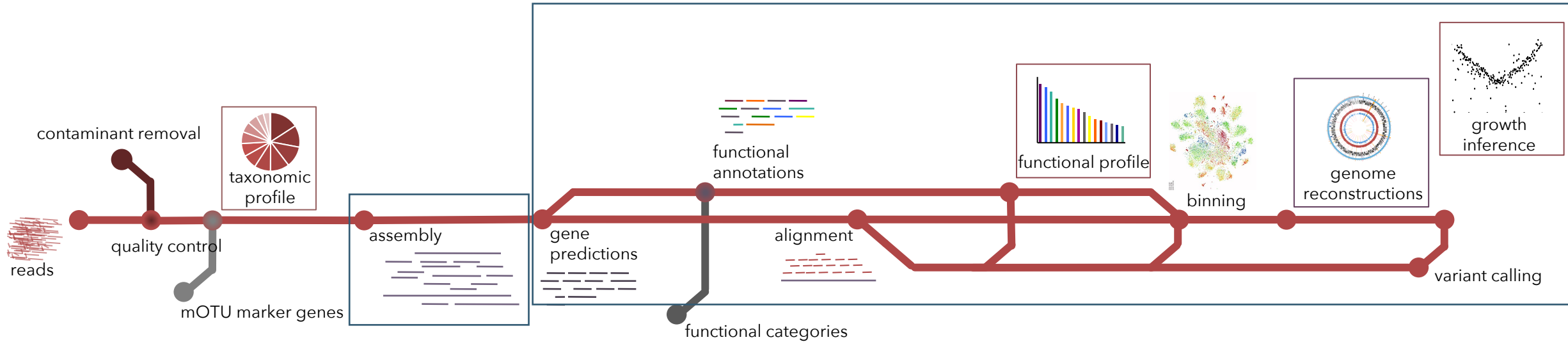
# Today

- idea of assembly
- alternatives
- how does assembly work?
- how to inspect assemblies
- what to assemble?

# Metagenomics (+ other omics) pipeline

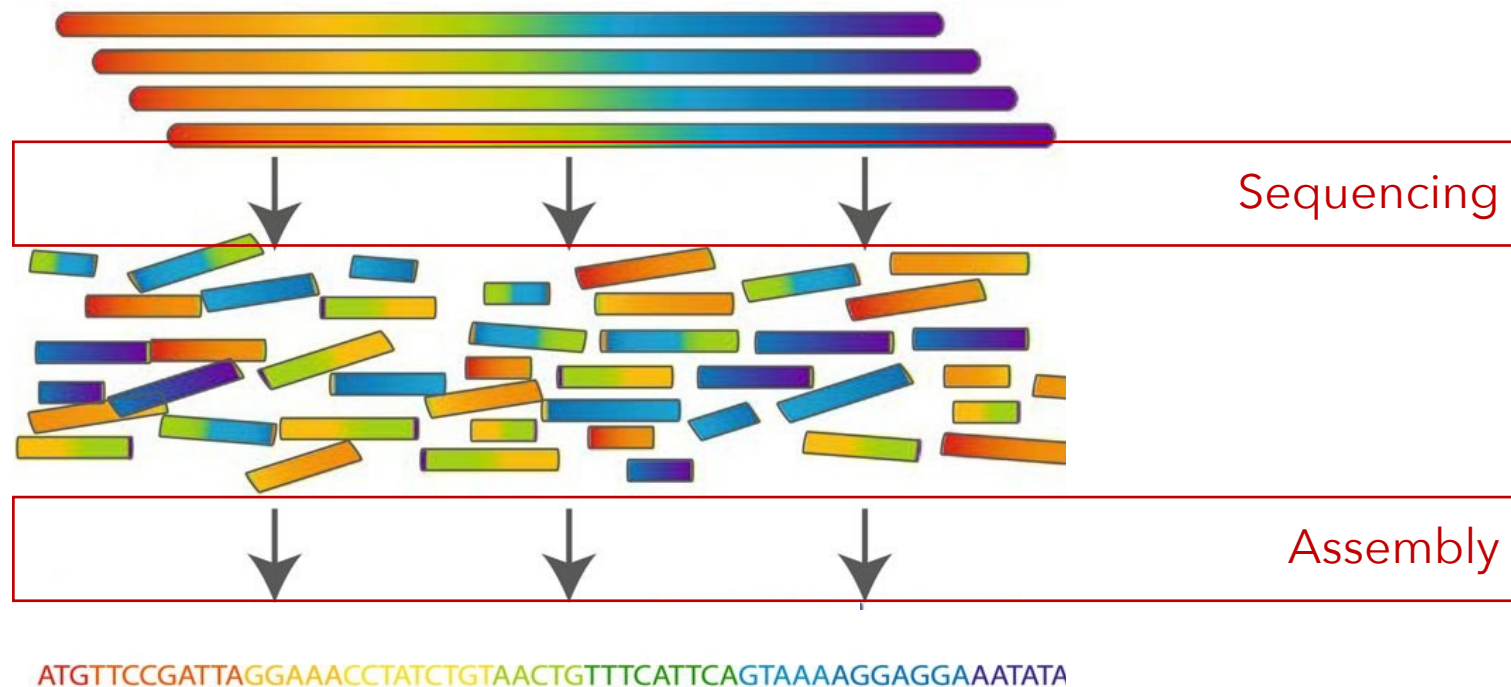


# Metagenomics (+ other omics) pipeline



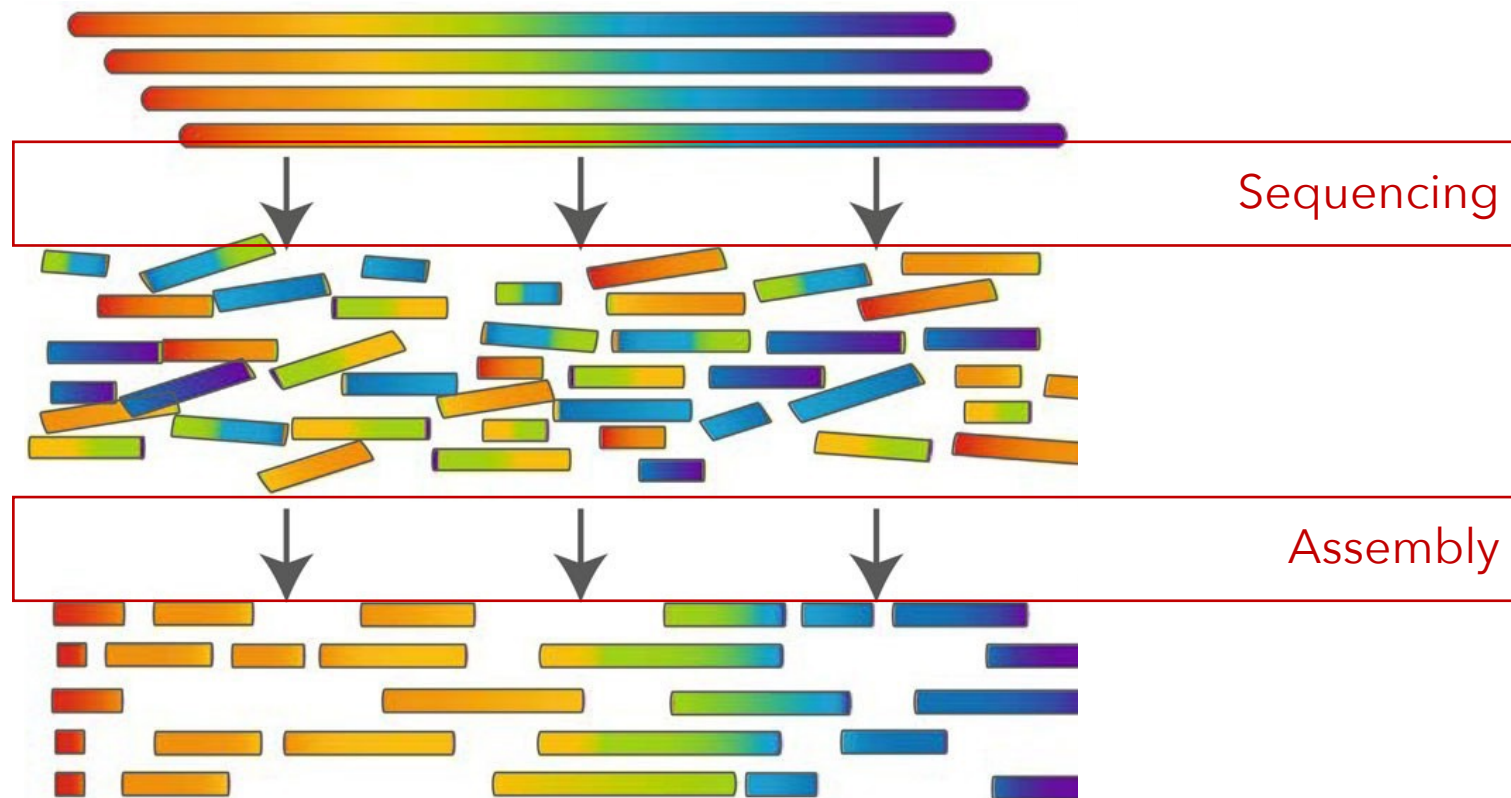
# What is "assembly"?

- puzzling sequencing reads back together



# What is "assembly"?

- puzzling sequencing reads back together

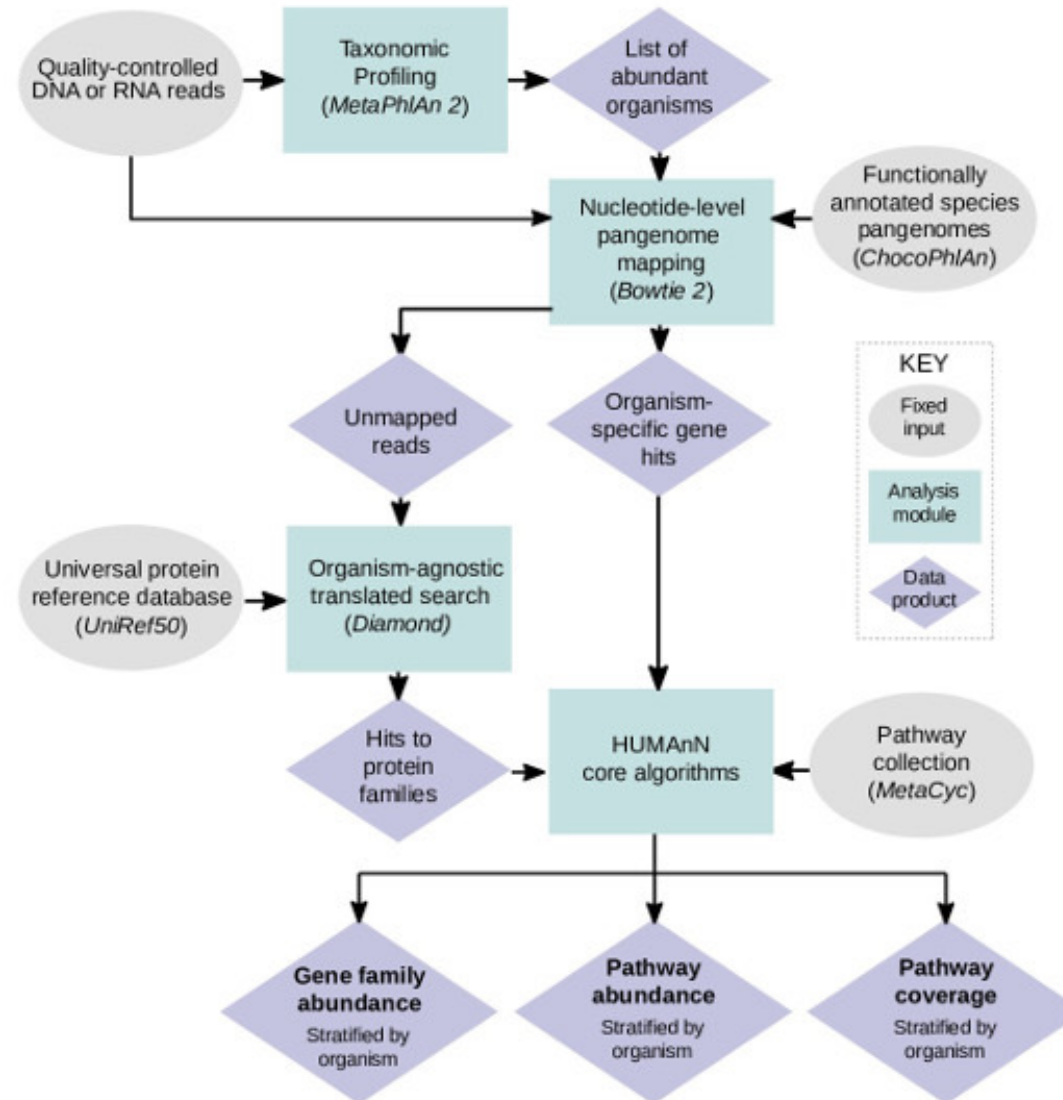


# Do we need assembly?

- assembly: “de novo” approaches
- also, usually: “genome-centric” approaches
- no assembly: “reference-based” approaches

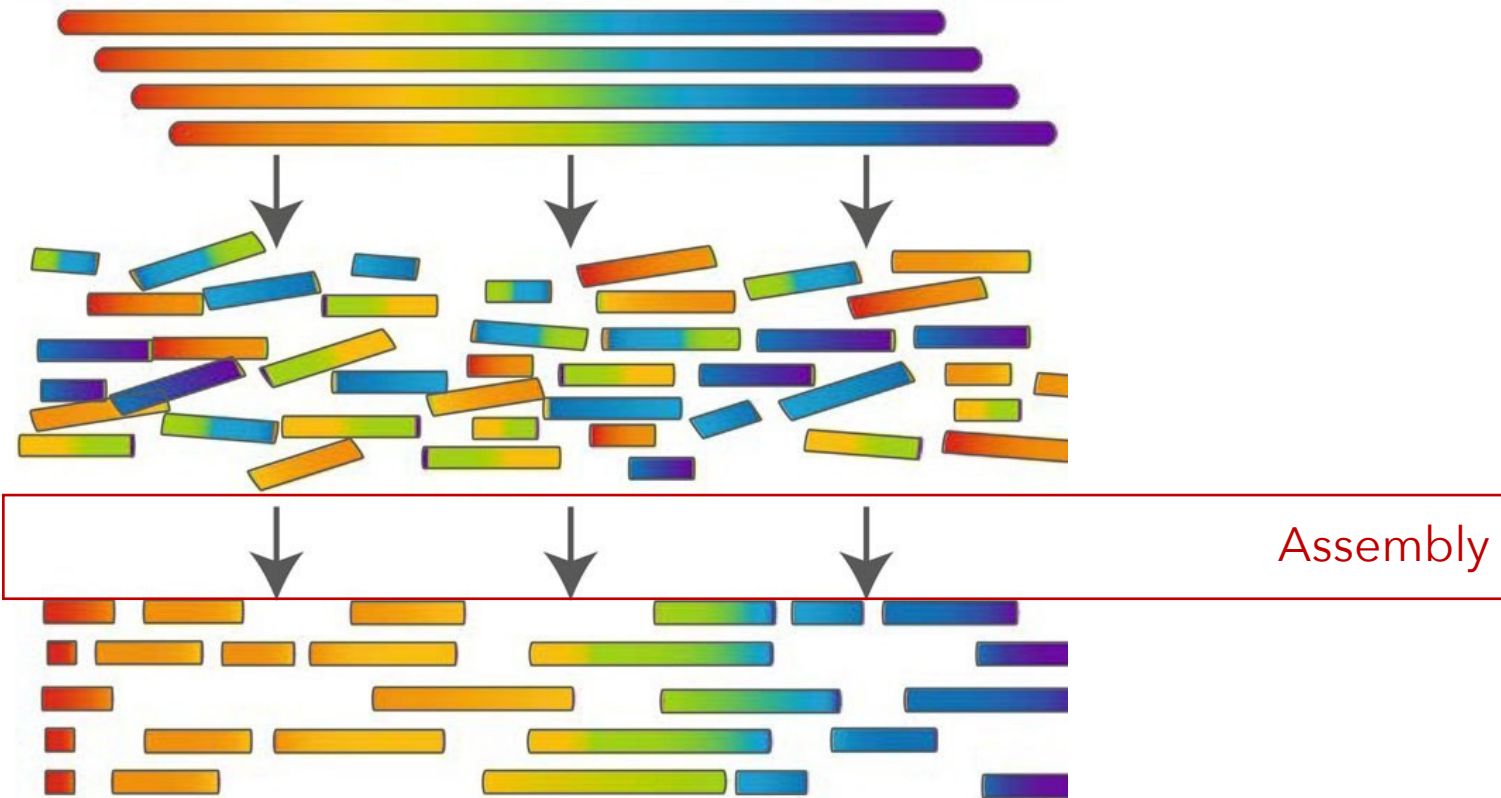
# Reference-based approaches

- HUMAnN 2

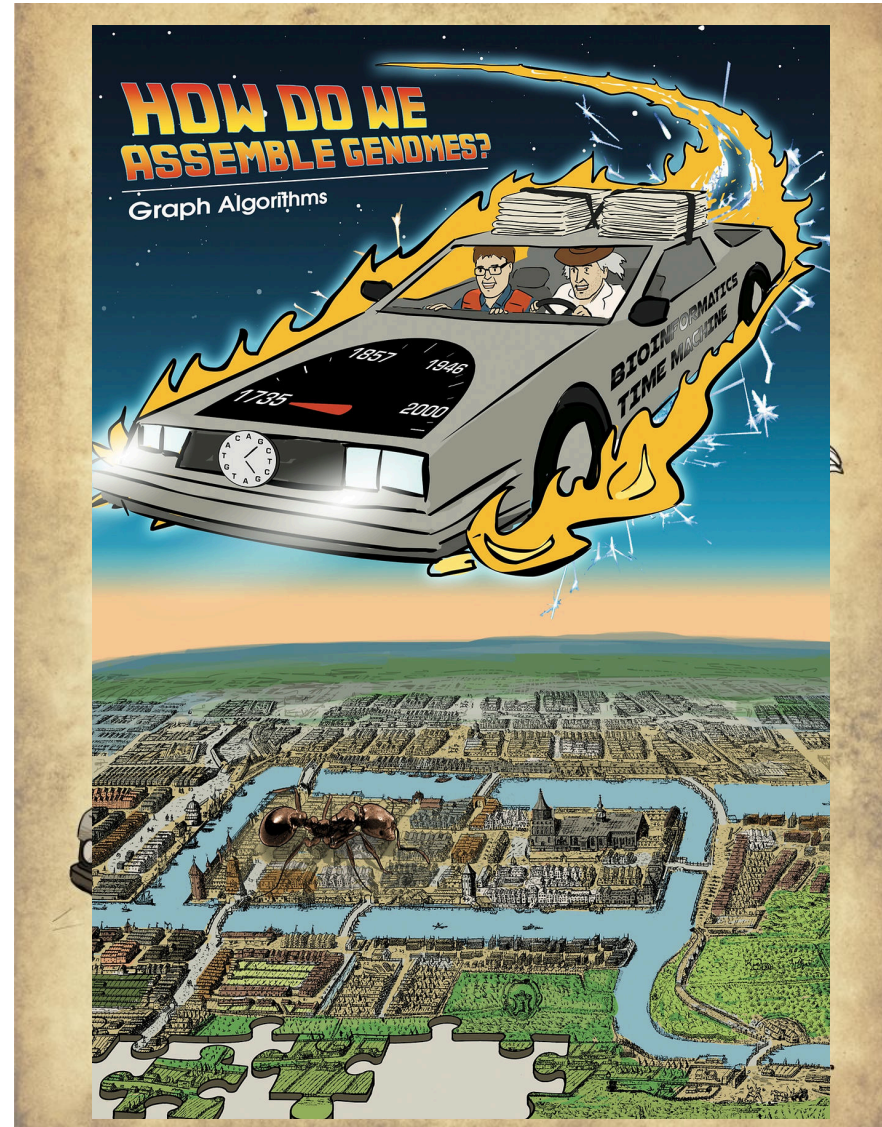




# How to solve the puzzle?



# How does assembly work?

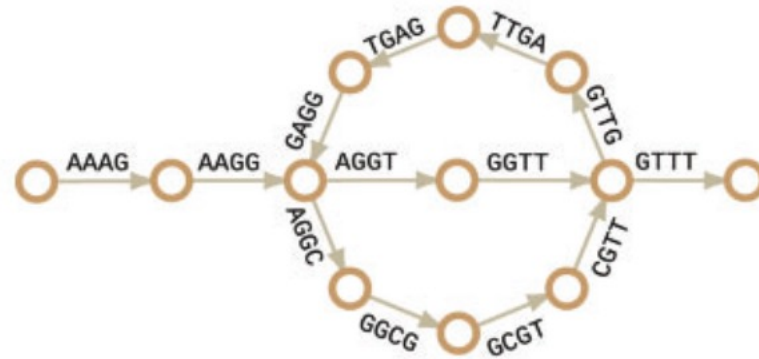


# How does assembly work?

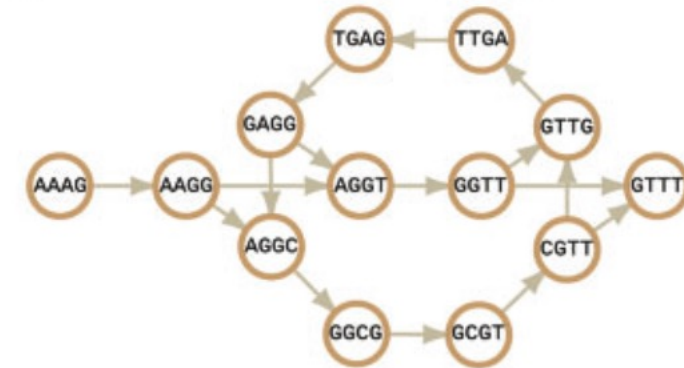
AAAGGCGTTGAGGTT

AAAG  
AAGG  
AGGC  
GGCG  
GCGT  
CGTT  
GTTG  
TTGA  
TGAG  
GAGG  
AGGT  
GGTT

**B** Eulerian de Bruijn graph

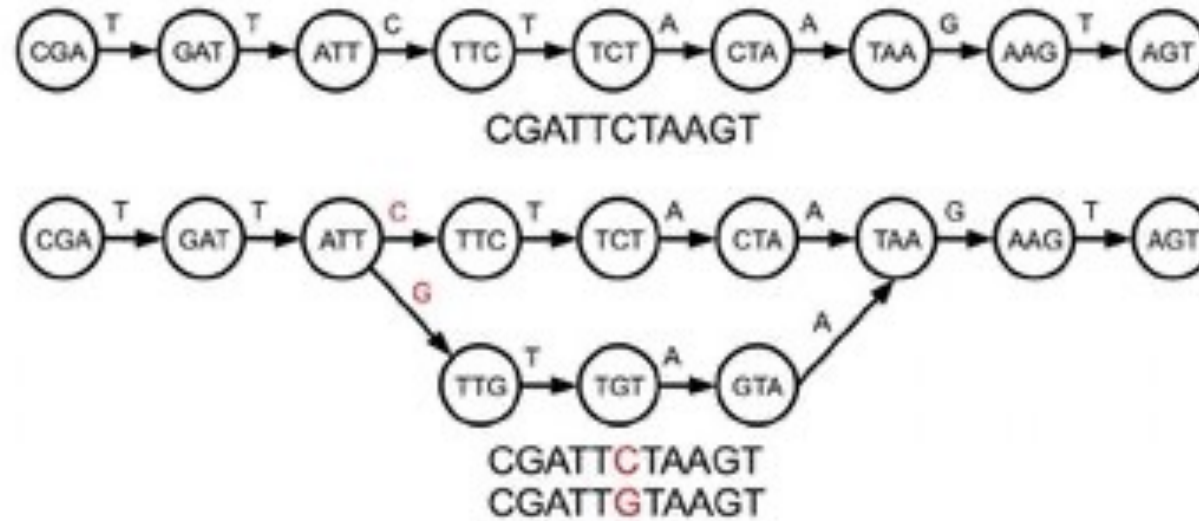


**C** Hamiltonian de Bruijn graph



# How does assembly work?

- effect of sequencing errors:

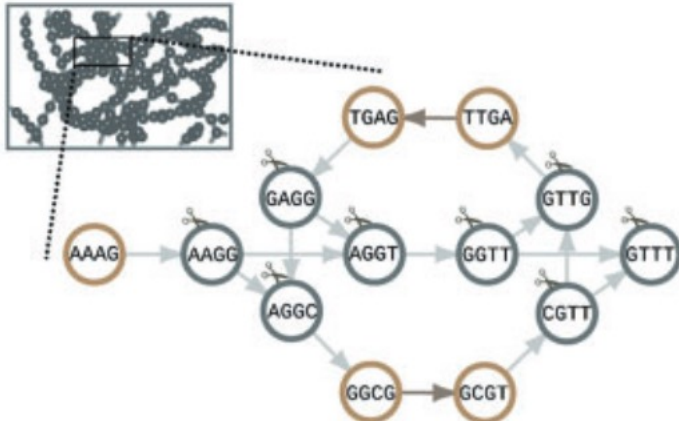


Leggett RM, Ramirez-Gonzalez RH, Verweij W, Kawashima CG, Iqbal Z, Jones JD, Caccamo M, Maclean D. Identifying and classifying trait linked polymorphisms in non-reference species by walking coloured de bruijn graphs. PLoS One. 2013;8(3):e60058

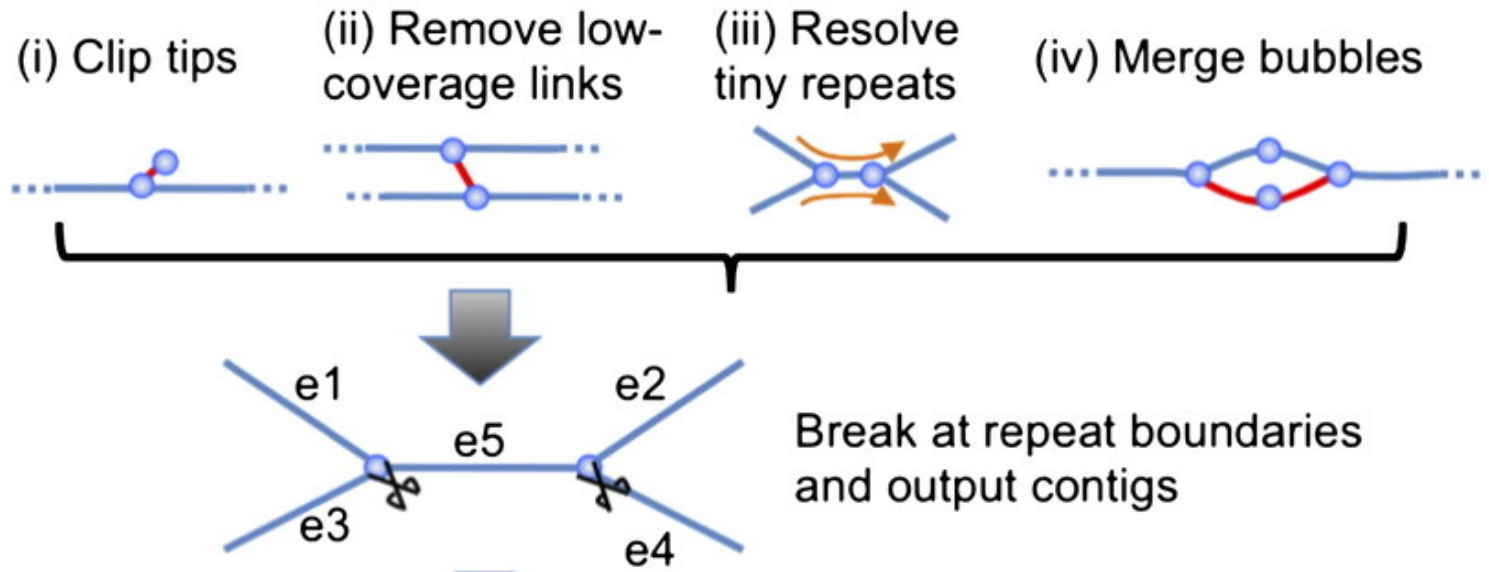
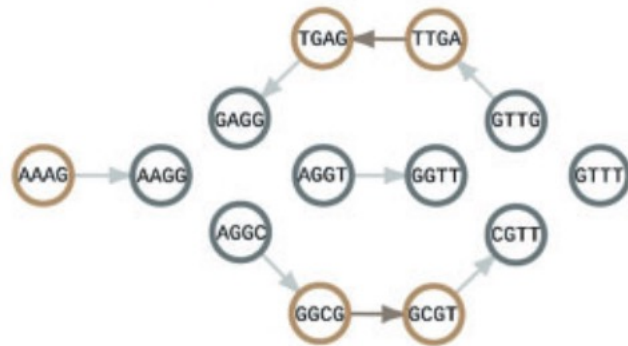
# How does assembly work?

## C Simplification of Hamiltonian graph

Complicated graph

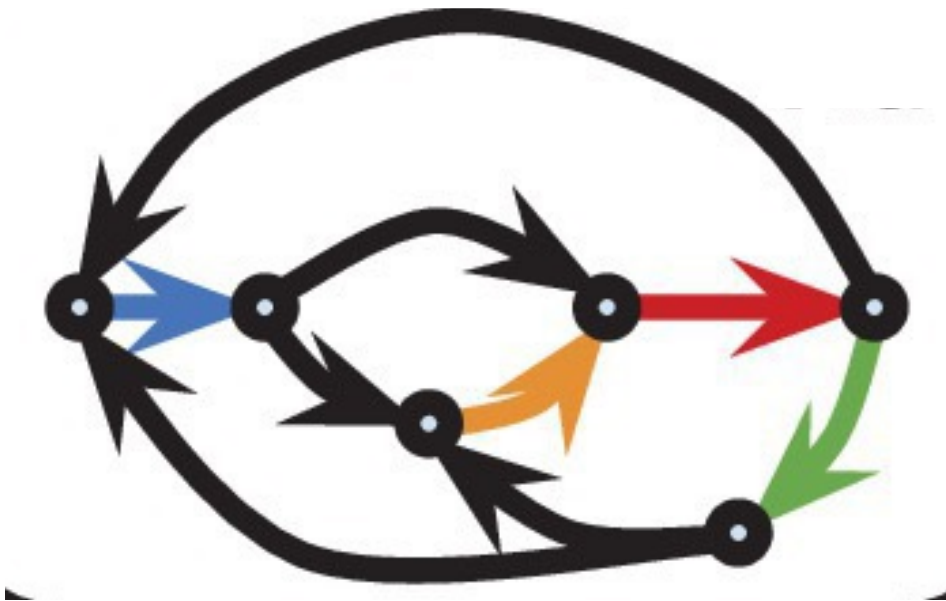


Simplification by removing branches



# Assembly outputs

- assembly graph (FASTG):



- assembled contigs (FASTA):



hundreds of Mbp

ten-hundred thousands of contigs

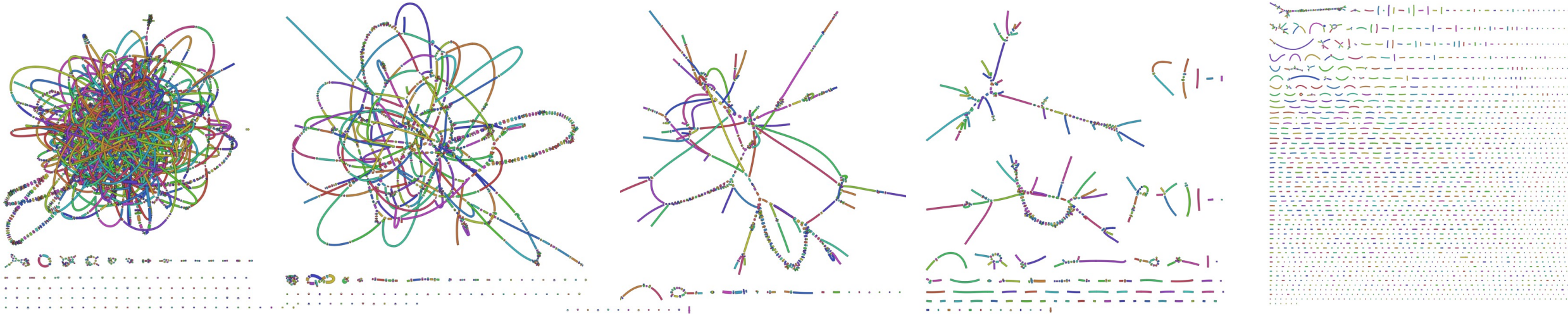
# *k*mers, again

- word sizes for alignment seeding: BLAST default 11, BWA default 19
- *k*mers for taxonomy: kraken1/2 default 31/33
- *k*mers for diversity: nonpareil 24
- *k*mers for assembly: metaSPAdes between 25 and 127

# Effect of *kmer* sizes

- small *kmers* work better on lower coverage
- larger *kmers* can resolve short repeats

increasing *kmer* size ->





# Metagenomics challenges

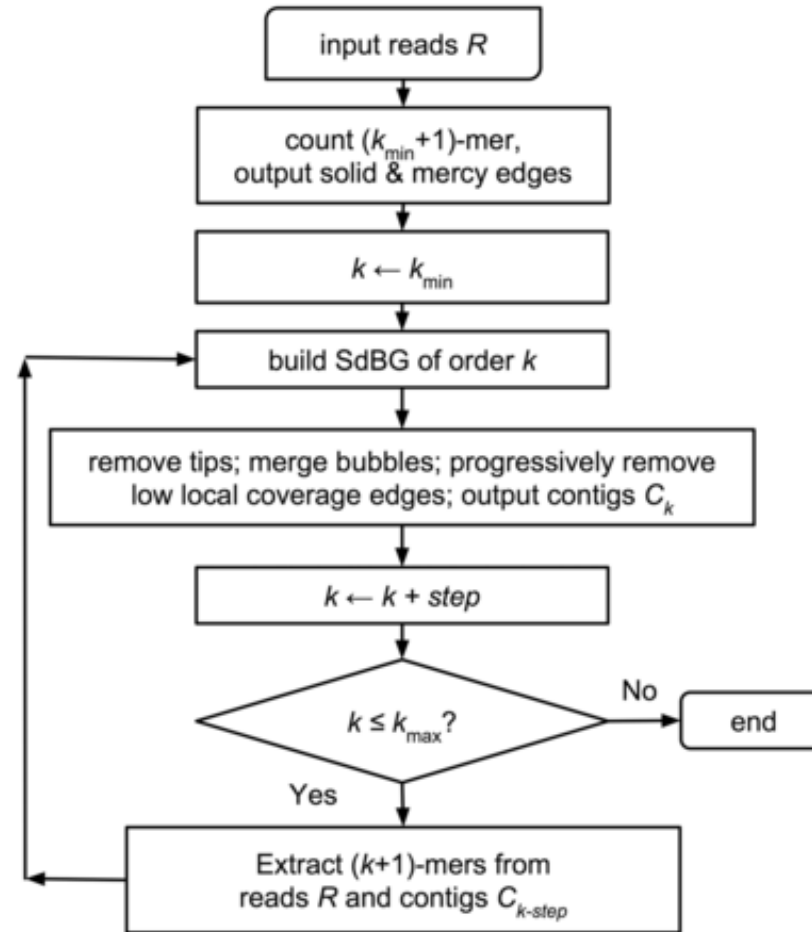
- diversity



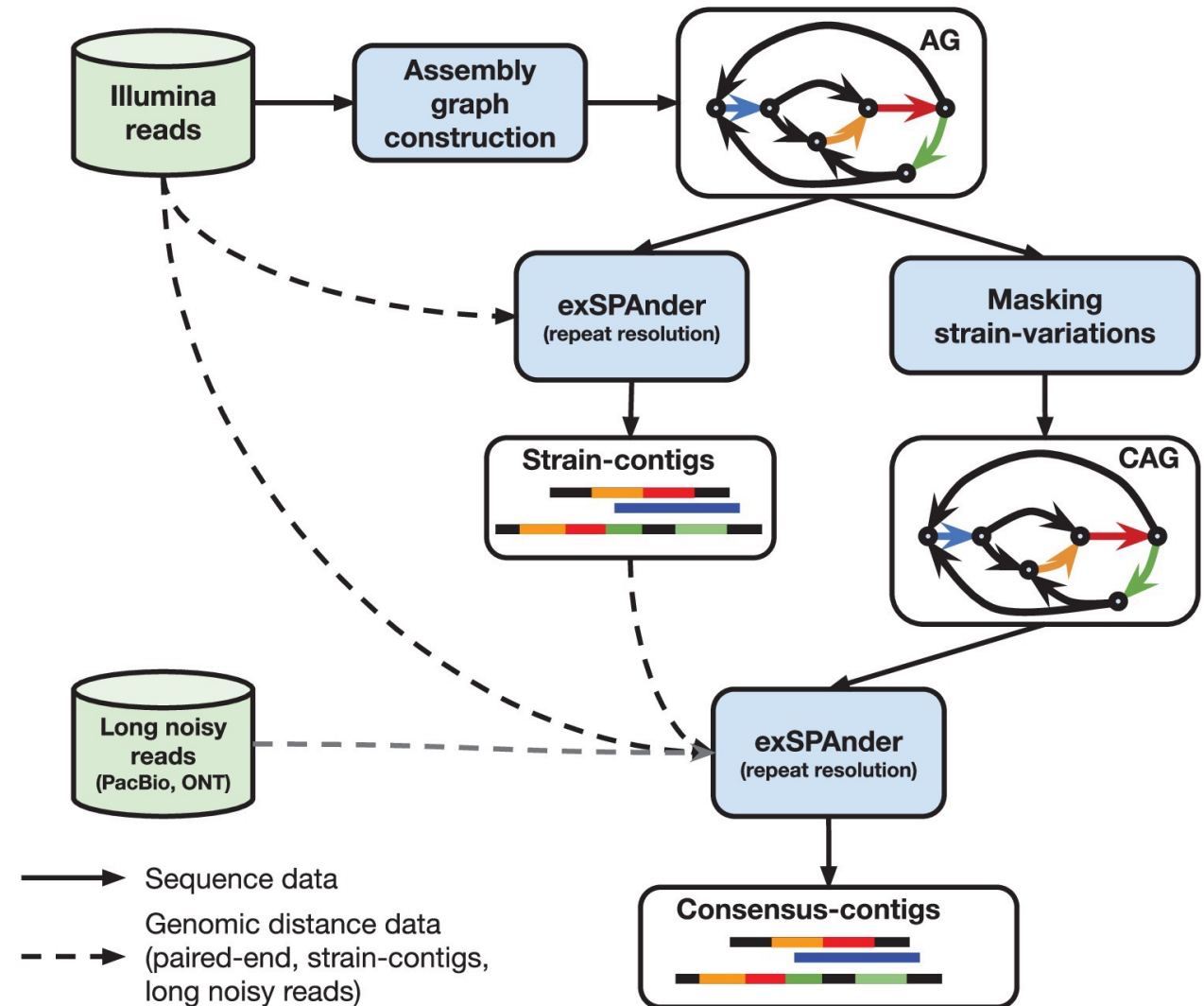
Green circles are scaffolds from low-abundance members of the community that are closely related to the abundant strain (blue)

- distinguishing true diversity from sequencing errors

# Megahit



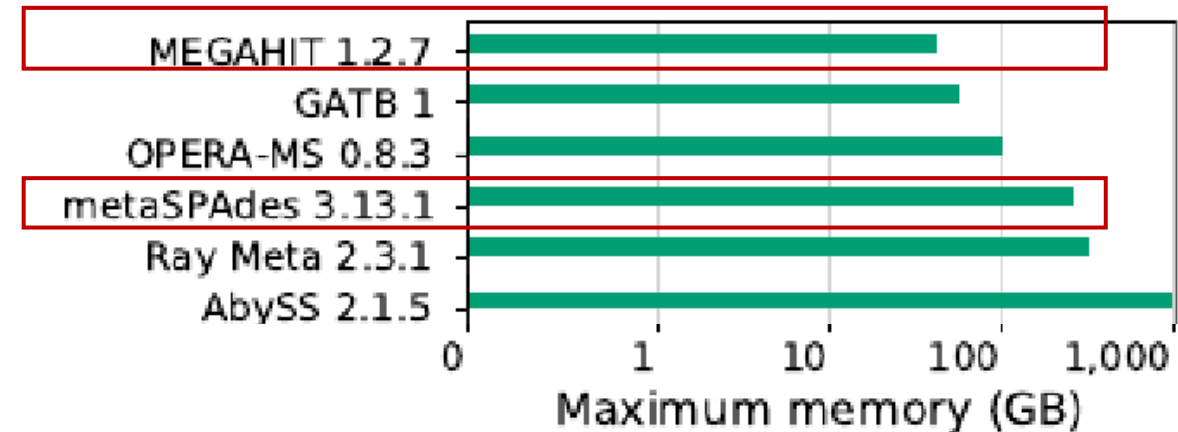
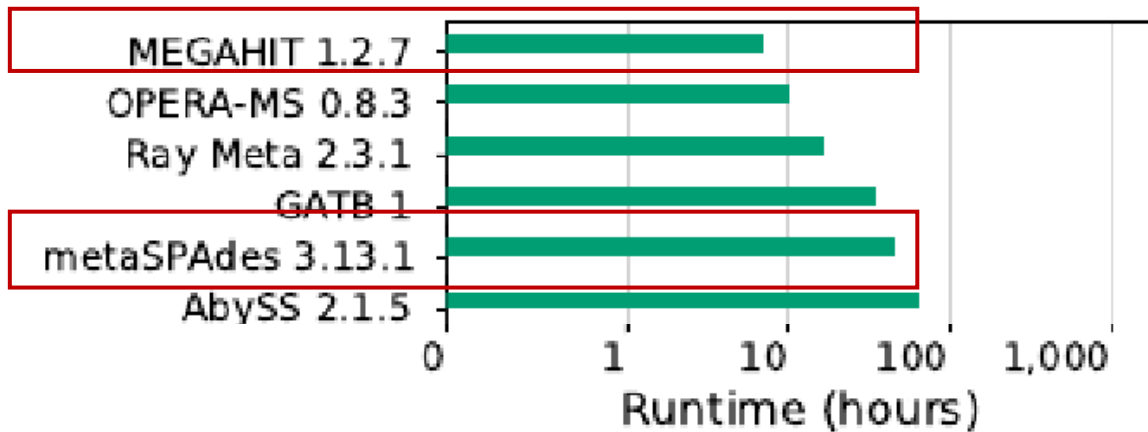
# MetaSpades



Lapidus AL, Korobeynikov AI. Metagenomic Data Assembly - The Way of Decoding Unknown Microorganisms. Front Microbiol. 2021 Mar 23;12:613791

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27(5):824-834

# Assembly has high computational demands

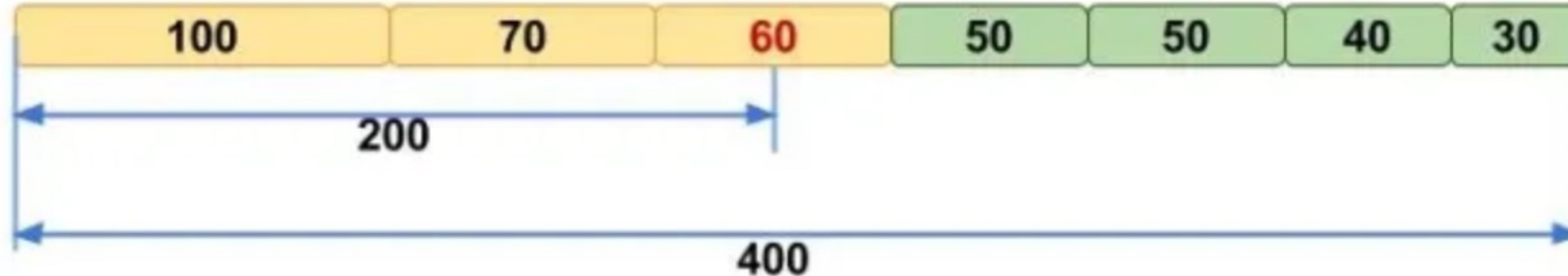


- run time depends on genome length (and algorithm)
- memory depends on the *kmers* (and algorithm)

# How to inspect assemblies?



1a. Contigs, sorted according to their lengths.

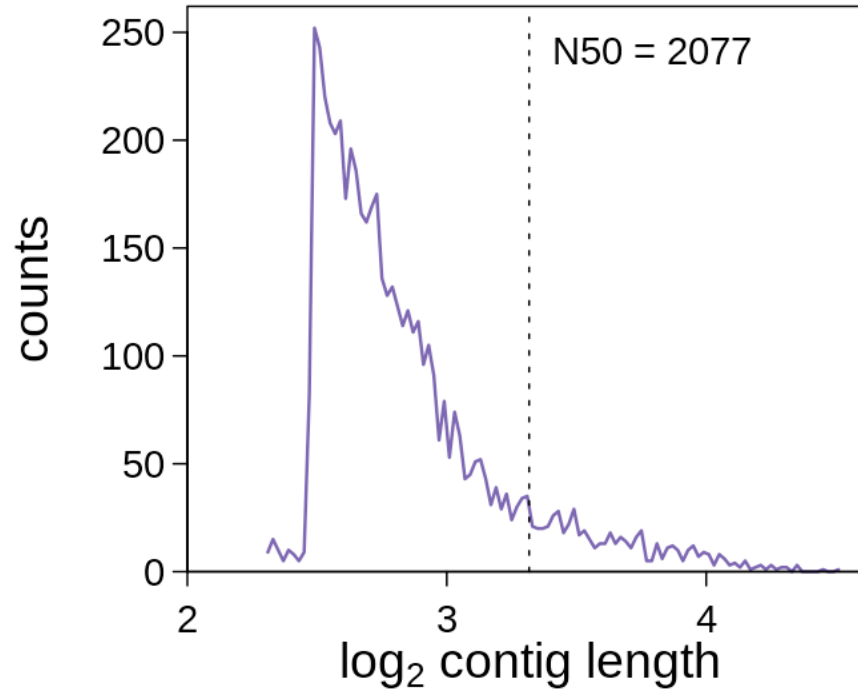


1b. Calculation of  $N_{50}$  using sorted contigs.

Fig. 1. Example of calculating  $N_{50}$  for a set of seven contigs.

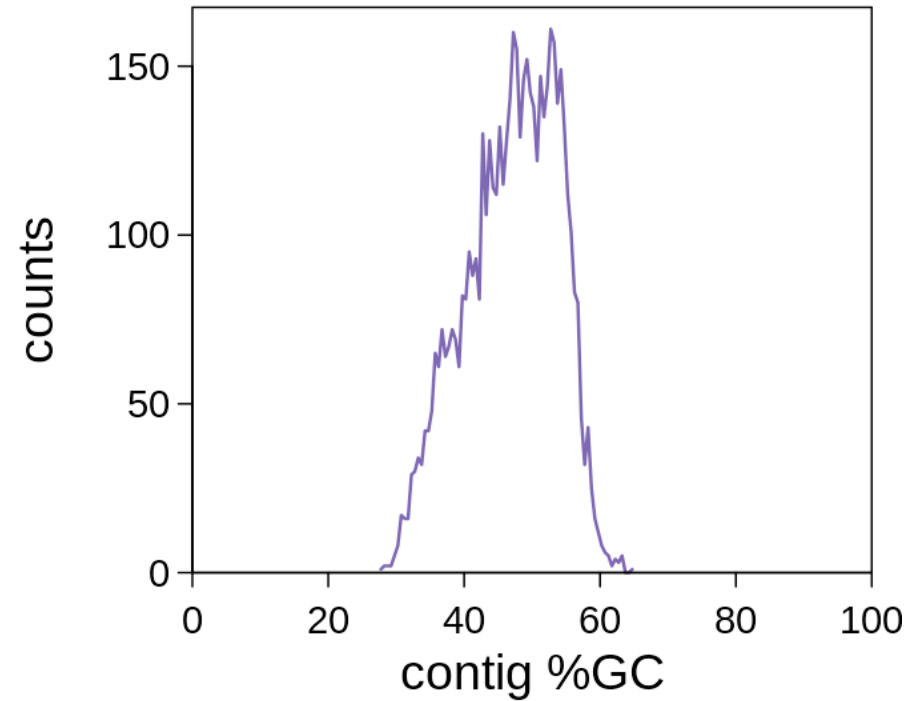
Here  $N_{50}$  equals 60 kbp.

# How to inspect assemblies?



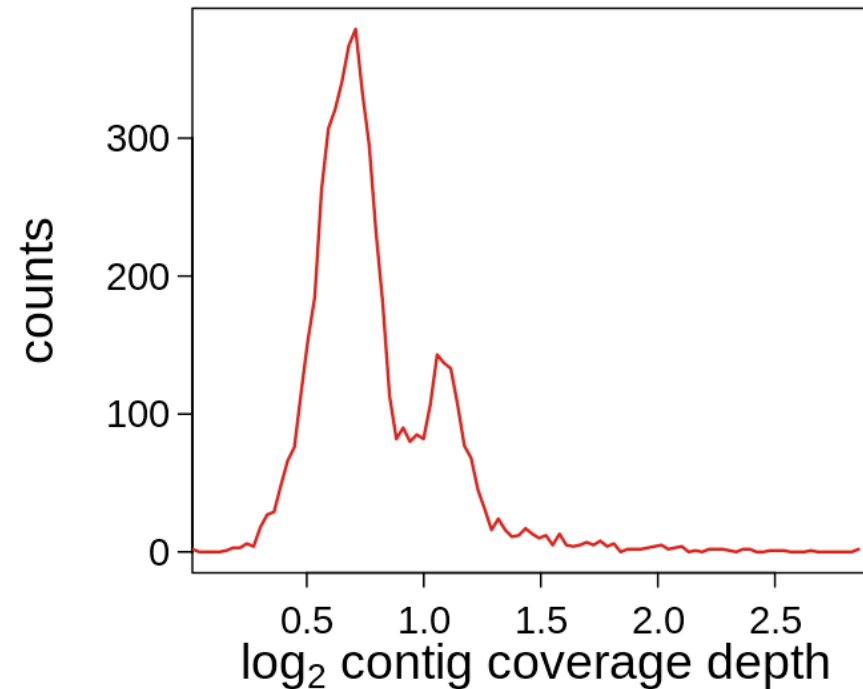
- contigs are often size filtered before analysis and further processing (e.g. min 500 or 1000 bp)

# How to inspect assemblies?



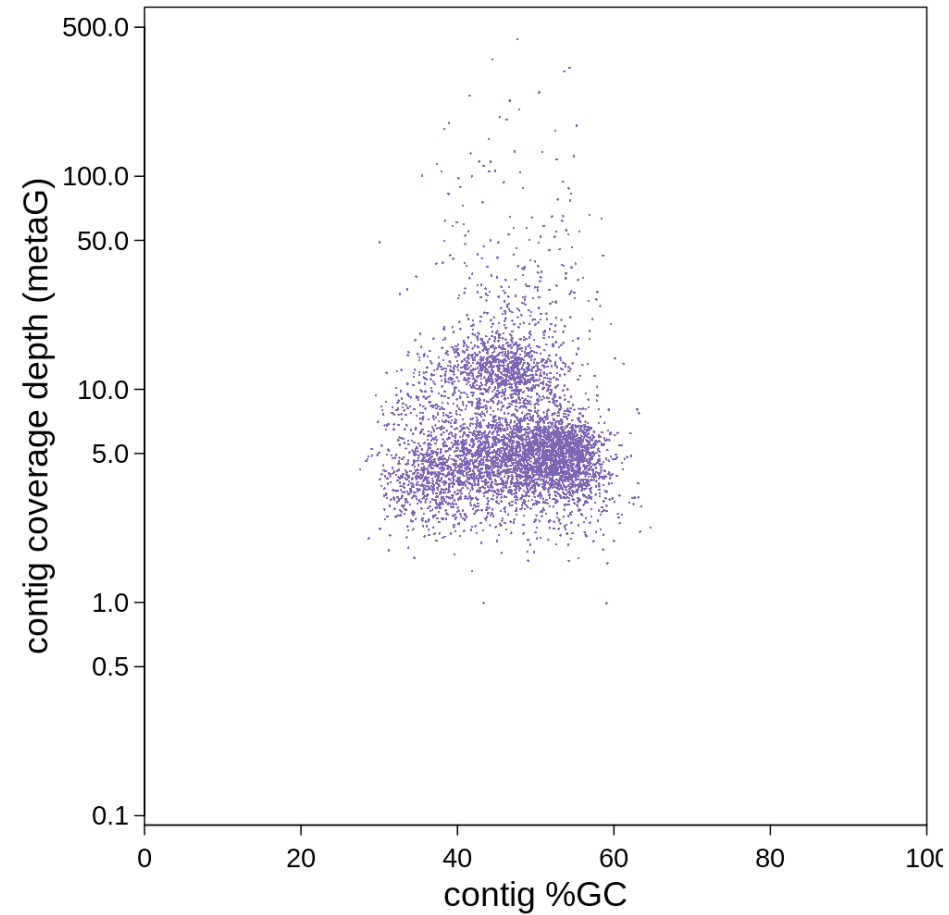
# How to inspect assemblies?

- mapping reads back on contigs



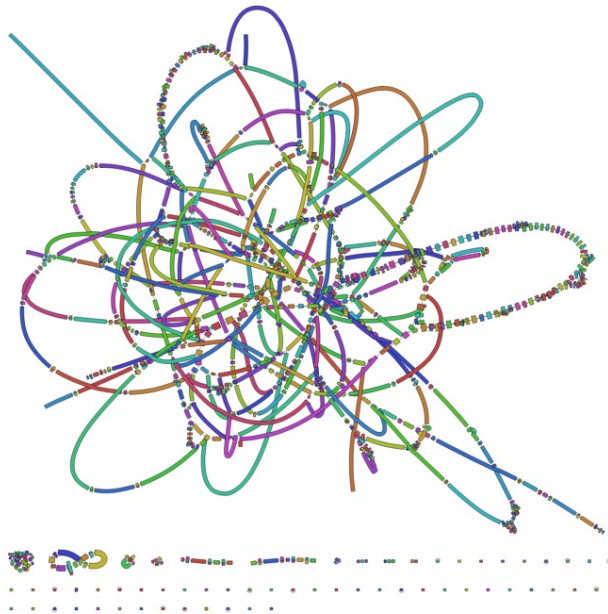


# How to inspect assemblies?



# How to inspect assemblies?

- Bandage:



Ryan R. Wick, Mark B. Schultz, Justin Zobel, Kathryn E. Holt (2015), Bandage: interactive visualization of de novo genome assemblies, *Bioinformatics* 31: 3350-3352, <https://doi.org/10.1093/bioinformatics/btv383>

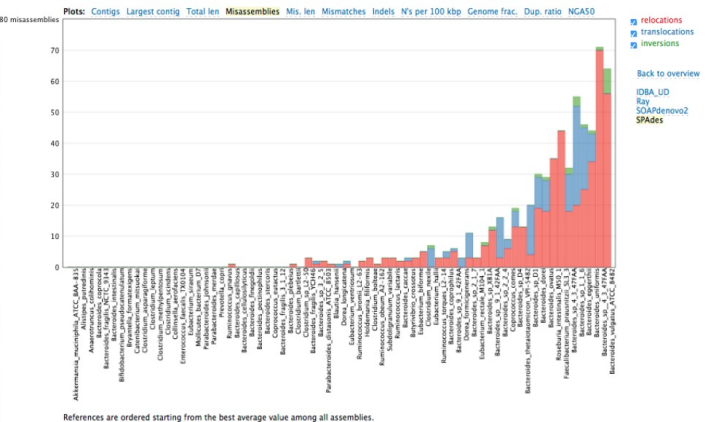
- Metaquast:

MetaQUAST report for assemblies of the MH0045 sample from MetaHit (Qin et al., 2010)

09 November 2015, Monday, 20:03:39  
 All statistics are based on contigs of size > 500 bp, unless otherwise noted (e.g., "k contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs.)  
 Rows show values for the whole assembly (column name) vs. combined reference (concatenation of input references).  
 Clicking on a row with **-** sign will expand values for contigs aligned to each of input references separately.  
 Note that some metrics (e.g. # contigs) may not sum up, because one contig may be aligned to several references and thus, counted several times.  
 All metrics that depend on the reference length (such as NGS, LGS), etc., plus the GC %, are not calculated for the combined reference.  
 The combined reference is just a concatenation of all available reference genomes of the species, presumably represented in the metagenomic dataset, but not necessarily the real content.  
 So it might miss many correctly assembled species, and therefore it doesn't make sense to apply the size and the GC content of the combined reference for assembly evaluation.

Reference size: 306 971 432 bp

Reference	Size, bp	GC, %
Akkermansia_muciniphila_ATCC_BAA-835	2 864 102	52.76
Alisipites_putredinis	2 550 878	53.27
Anaerotruncus_culturomis	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_dornii	6 060 928	42.2
Bacteroides_eggerthii	4 611 535	44.71
Bacteroides_finegoldii	5 124 100	42.6
Bacteroides_fragilis_3_1_12	5 530 115	43.62
Bacteroides_fragilis_NCTC_9343	5 205 140	43.19
Bacteroides_fragilis_YCH46	5 272 274	43.27
Bacteroides_intestinalis	4 605 106	43.54
Bacteroides_ovatus	7 010 996	42.3
Bacteroides_pectinophilus	29 332	36.86
Bacteroides_plebeius	4 421 924	44.31
Bacteroides_sp_1_1_6	6 760 735	43.02
Bacteroides_sp_1_1_7	3 180 144	45.08
Bacteroides_sp_2_2_4	7 101 224	42.13
Bacteroides_sp_2_2_5	3 116 282	43.17
Bacteroides_sp_4_3_47FAA	5 442 925	42.7
Bacteroides_sp_4_3_47FAA	5 622 644	42.33
Bacteroides_sp_D1	5 974 559	41.88
Bacteroides_sp_D4	5 538 248	41.75
Bacteroides_sp_D4	5 976 145	41.89
Bacteroides_sp_4_3_47FAA	4 440 920	42.7
Bacteroides_sp_9_1_42FAA	4 884 745	42.2
Bacteroides_stercoris	4 102 660	31.33
Bacteroides_thetaosiamocrocei_VPI-5482	6 260 361	42.84
Bacteroides_uniformis	4 835 507	46.49
Bacteroides_vulgatus_ATCC_8482	5 163 189	42.1
Bifidobacterium_pseudocatenulatum	2 313 572	56.38
Blautia_hansenii	3 058 721	38.99
Byastinia_fornaxiensis	4 544 860	49.55
Butyrivibrio_croosotus	2 496 039	37.75
Catenibacterium_mitsuokai	2 671 313	36.62
Clostridium_anaeraglyforme	6 413 332	55.6
Clostridium_bartlettii	2 972 256	28.84
Clostridium_boleae	6 538 460	49.19
Clostridium_leptum	3 270 209	50.19
Clostridium_methylpentosum	2 478 423	51.82
Clostridium_mexiae	1 993 628	40.090
Clostridium_scindens	1 631 609	46.03
Clostridium_sp_12_50	2 954 816	41.37
Collinsella_aerofaciens	4 439 869	60.55
Coproccoccus_comes	3 242 215	42.49
Coproccoccus_nictatus	3 102 987	43.09
Dorea_fornicigenans	3 843 583	40.340
Dorea_longicatena	2 915 433	41.44
Enterococcus_faecalis_T30104	3 155 474	37.270
Lubacterium_biforme	2 517 763	33.79
Lubacterium_halli	3 290 996	38.19
Lubacterium_rectak_M104_1	3 698 419	40.550
Lubacterium_siraeum	2 664 035	44.97
Lubacterium_siraeum	2 879 795	34.92
Facaliclostridium_praeurtizii_SL3_3	3 214 418	55.65
Holdemania_bififormis	3 932 923	50.18
Mollicutes_bacterium_D7	3 561 737	31.37
Parabacteroides_distansum_ATCC_8503	4 811 379	45.06
Parabacteroides_johnsonii	4 629 061	45.13
Parabacteroides_merdae	4 453 741	45.25
Prevotella_copri	3 512 473	44.85
Roseburia_intestinalis_MS0_1	4 143 550	42.41
Ruminococcus_bromii_L2-63	2 240 085	41.39






Alla Mikheenko, Vladislav Saveliev, Alexey Gurevich, *MetaQUAST: evaluation of metagenome assemblies*, *Bioinformatics* (2016) 32 (7): 1088-1090. doi: 10.1093/bioinformatics/btv697

# How to choose a assembler? Benchmarks

ANALYSIS

OPEN

## Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software

Alexander Sczyrba<sup>1,2,48</sup>, Peter Hofmann<sup>3-5,48</sup>, Peter Belmann<sup>1,2,4,5,48</sup>, David Koslicki<sup>6</sup>, Stefan Janssen<sup>4,7,8</sup>, Johannes Dröge<sup>3-5</sup>, Ivan Gregor<sup>3-5</sup>, Stephan Majda<sup>3,47</sup>, Jessika Fiedler<sup>3,4</sup>, Eik Dahms<sup>3-5</sup>, Andreas Bremges<sup>1,2,4,5,9</sup>, Adrian Fritz<sup>4,5</sup>, Ruben Garrido-Oter<sup>3-5,10,11</sup>, Tue Sparholt Jørgensen<sup>12-14</sup>, Nicole Shapiro<sup>15</sup>, Philip D Blood<sup>16</sup>, Alexey Gurevich<sup>17</sup>, Yang Bai<sup>10,47</sup>, Dmitriy Turaev<sup>18</sup>, Matthew Z DeMaere<sup>19</sup>, Rayan Chikhi<sup>20,21</sup>, Niranjana Nagarajan<sup>22</sup>, Christopher Quince<sup>23</sup>, Fernando Meyer<sup>4,5</sup>, Monika Balvočiūtė<sup>24</sup>, Lars Hestbjerg Hansen<sup>12</sup>, Søren J Sørensen<sup>13</sup>, Burton K H Chia<sup>22</sup>, Bertrand Denis<sup>22</sup>, Jeff L Froula<sup>15</sup>, Zhong Wang<sup>15</sup>, Robert Egan<sup>15</sup>, Dongwan Don Kang<sup>15</sup>, Jeffrey J Cook<sup>25</sup>, Charles Deltel<sup>26,27</sup>, Michael Beckstette<sup>28</sup>, Claire Lemaitre<sup>26,27</sup>, Pierre Peterlongo<sup>26,27</sup>, Guillaume Rizk<sup>27,29</sup>, Dominique Lavenier<sup>21,27</sup>, Yu-Wei Wu<sup>30,31</sup>, Steven W Singer<sup>30,32</sup>, Chirag Jain<sup>33</sup>, Marc Strous<sup>34</sup>, Heiner Klingenberg<sup>35</sup>, Peter Meinicke<sup>35</sup>, Michael D Barton<sup>15</sup>, Thomas Lingner<sup>36</sup>, Hsin-Hung Lin<sup>37</sup>, Yu-Chieh Liao<sup>37</sup>, Genivaldo Gueiros Z Silva<sup>38</sup>, Daniel A Cuevas<sup>38</sup>, Robert A Edwards<sup>38</sup>, Surya Saha<sup>39</sup>, Vitor C Piro<sup>40,41</sup>, Bernhard Y Renard<sup>40</sup>, Mihai Pop<sup>42,43</sup>, Hans-Peter Klenk<sup>44</sup>, Markus Göker<sup>45</sup>, Nikos C Kyrpides<sup>15</sup>, Tanja Woyke<sup>15</sup>, Julia A Vorholt<sup>46</sup>, Paul Schulze-Lefert<sup>10,11</sup>, Edward M Rubin<sup>15</sup>, Aaron E Darling<sup>19</sup> , Thomas Rattei<sup>18</sup>  & Alice C McHardy<sup>3-5,11</sup> 



# What to assemble?

- single-sample      depth can be limiting
- multi-sample      diversity can cause troubles
  
- short reads
- long reads      different approaches, errors & memory can be an issue
- short and long reads  
    long reads for scaffolding or co-assembly
  
- metagenomics, metatranscriptomics or both  
    introns or not? can increase depth, different coverage can be problematic



# Thanks for your attention!



[a.u.s.heintzbuschart@uva.nl](mailto:a.u.s.heintzbuschart@uva.nl)

SP C2.205



[github.com/a-h-b](https://github.com/a-h-b)



[twitter.com/\\_a\\_h\\_b\\_](https://twitter.com/_a_h_b_)

