Reference-guided assembly

Acknowledgements

- Dr. Mihai Pop: Professor, Computer Science, University of Maryland College Park

- Dr. Victoria Cepeda: Formerly PhD student, Pop lab

What is reference-guided assembly?

- An assembler with a *strong assumption* that genomes in your metagenome look a lot like those that are in a reference database.

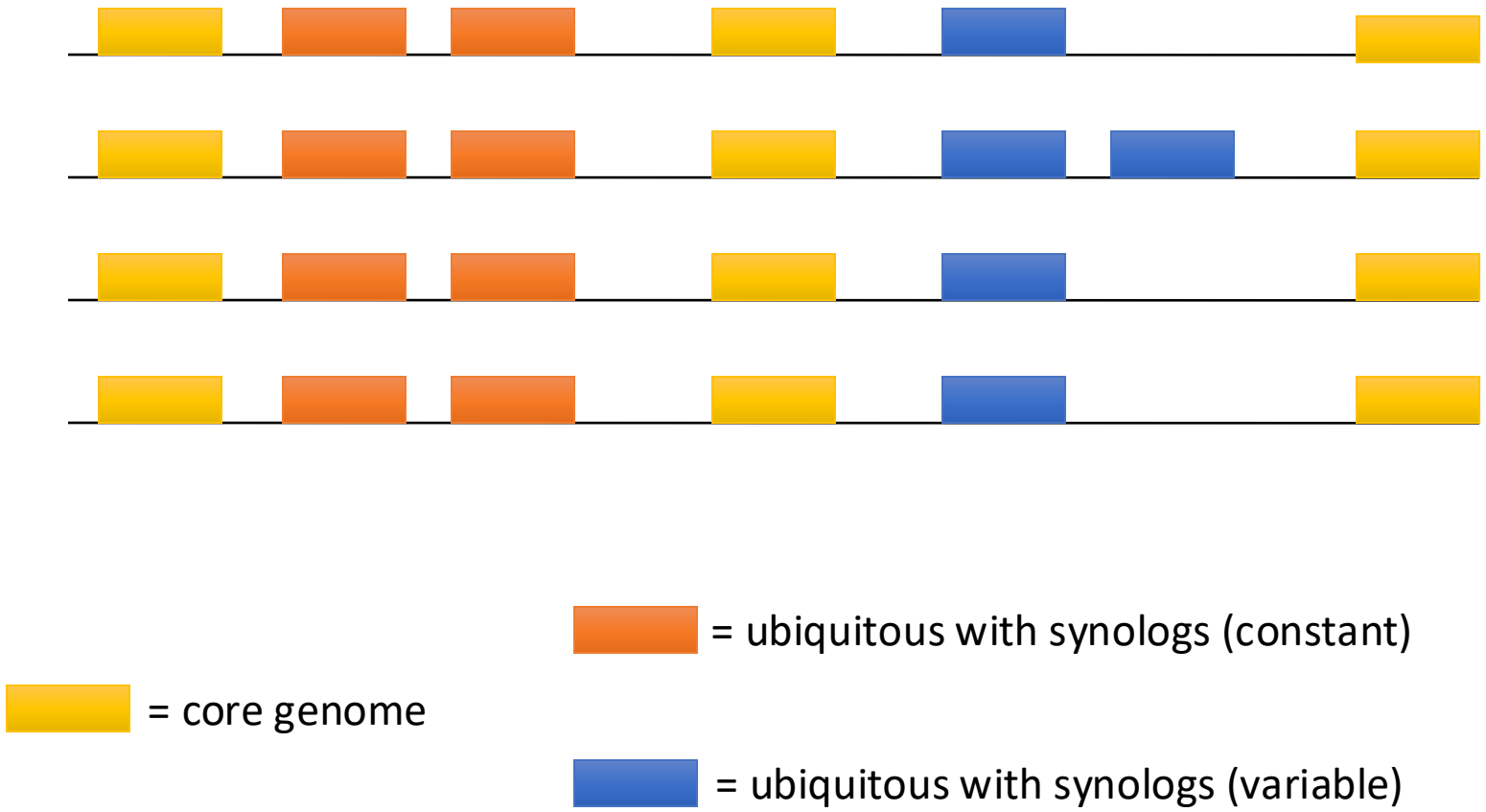
- If this is a reasonable assumption, **proceed with caution:**
  - Hint: rearrangements, horizonal gene transfer, and duplications are common!
- If this is not a reasonable assumption (viral genomes, soil samples), **think de novo assembly:**
  - Megahit
  - MetaSpades

Microbial genomes evolve over time

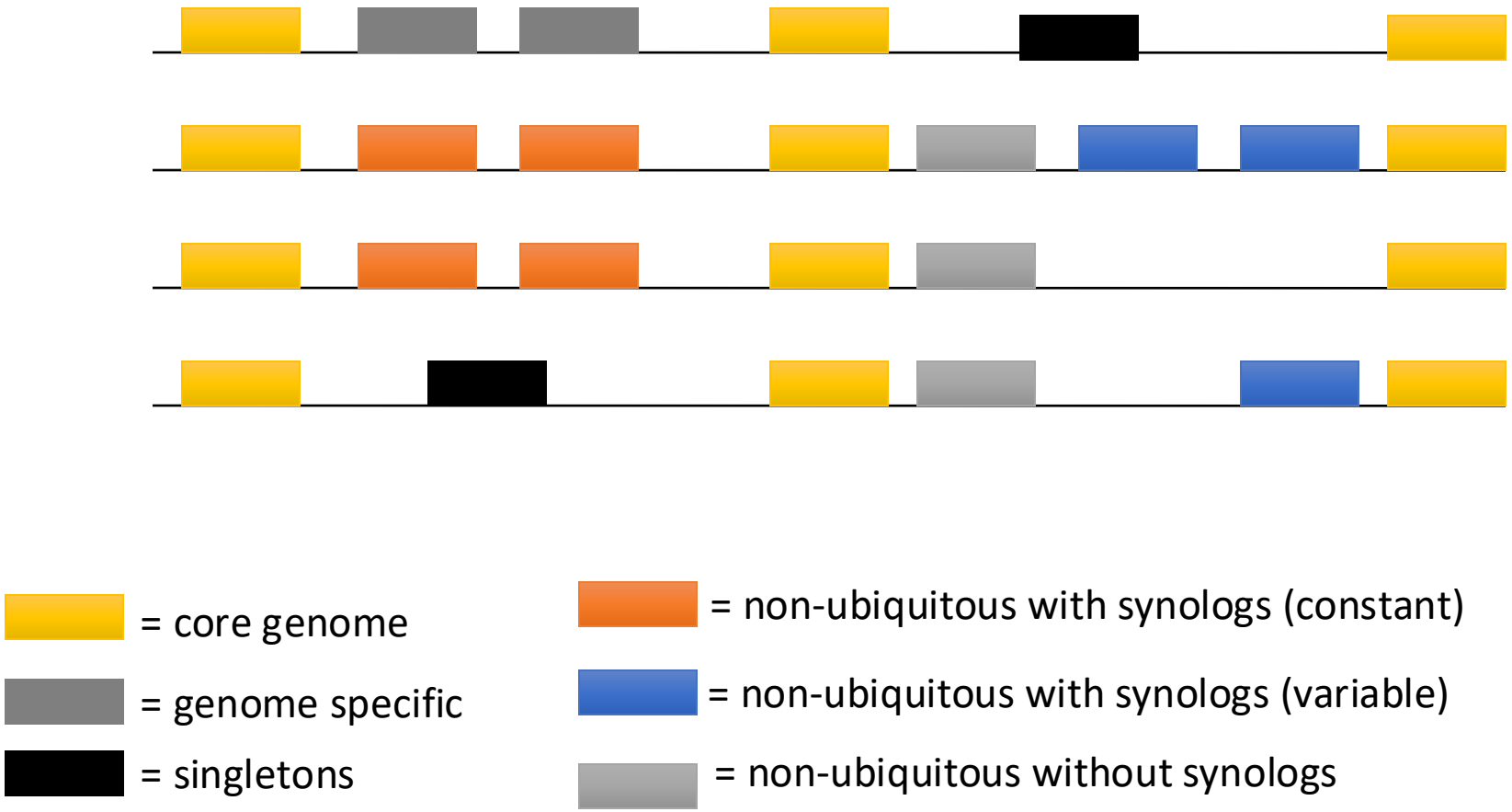
- *The presence of two or more homologous sequences within a single genome might reflect the acquisition of DNA sequence from a foreign source rather than the duplication of a resident gene.*

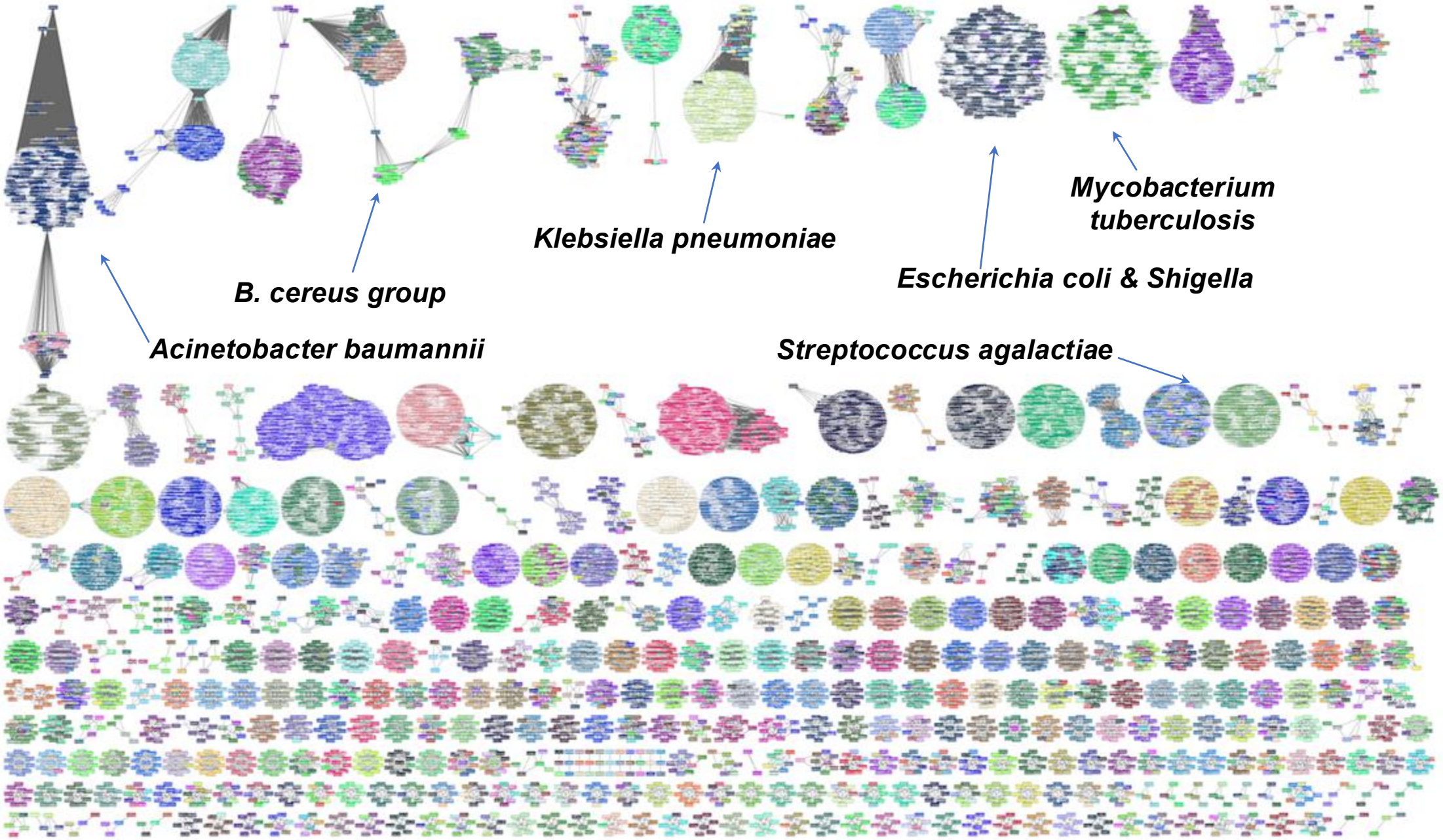
- Thus, since we do not know the origin a priori, we refer to these potential paralogs or xenologs as **synologs** (Lerat et al 2005).

Ubiquitous sequence fragments



Non-Ubiquitous sequence fragments



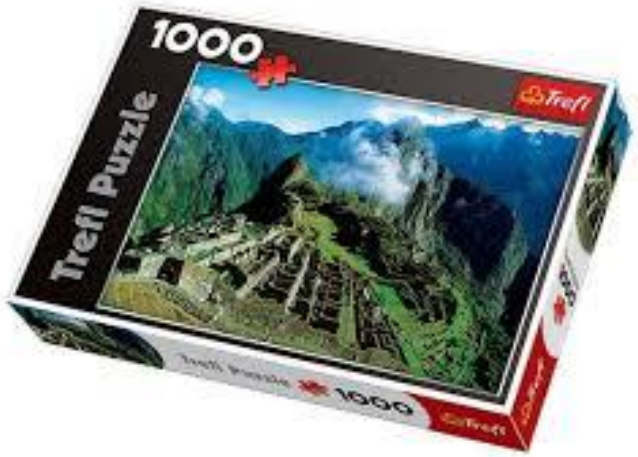

Mycobacterium tuberculosis
Klebsiella pneumoniae
Escherichia coli & Shigella
B. cereus group
Acinetobacter baumannii
Streptococcus agalactiae

Reference-guided metagenomic assembly

Ready, set, go!

Reference-guided genome assembly

- Reconstructing the original DNA sequence by aligning reads to a genome.
- Intuitively like a puzzle
- But we have the box!



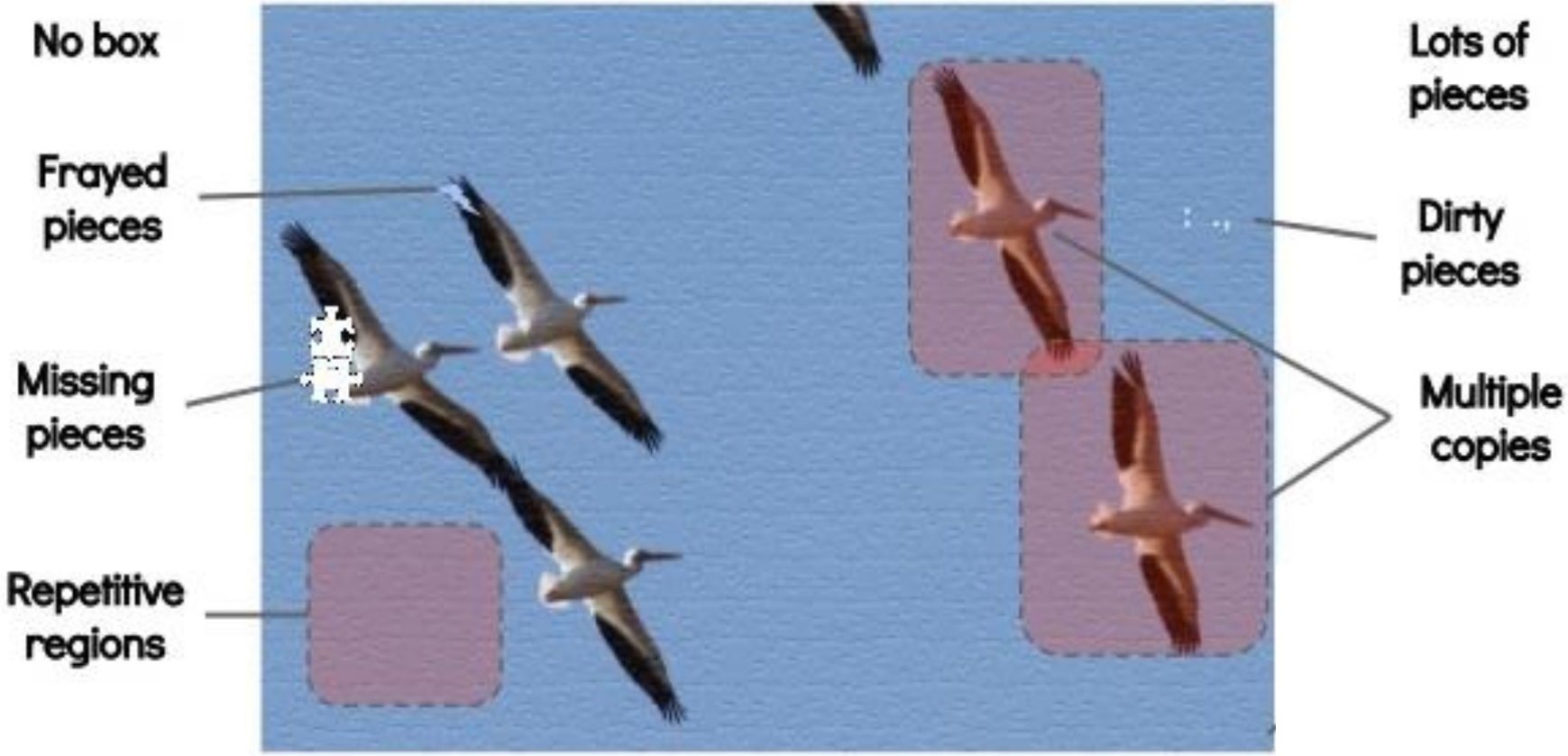
Reference-guided metagenome assembly

- Reconstructing original DNA sequences aligning reads to a set of genomes.
- Intuitively like multiple puzzles
- But we need to find the boxes!

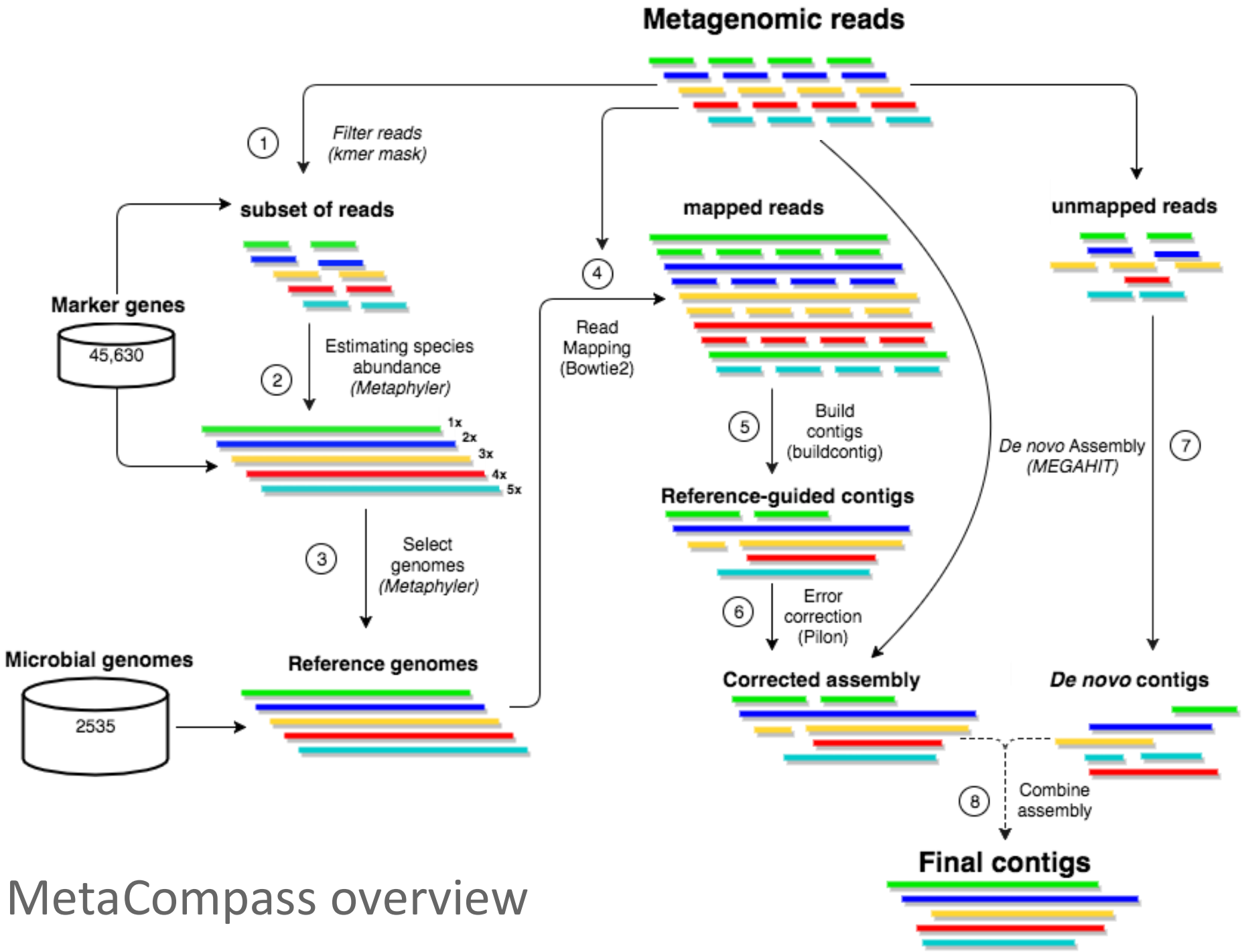
Reference-guided metagenome assembly

- Step 1: Find the puzzle boxes (reference selection)
- Step 2: Bin pieces into the right boxes (read mapping)
- Step 3: Solve each puzzle (assembly)

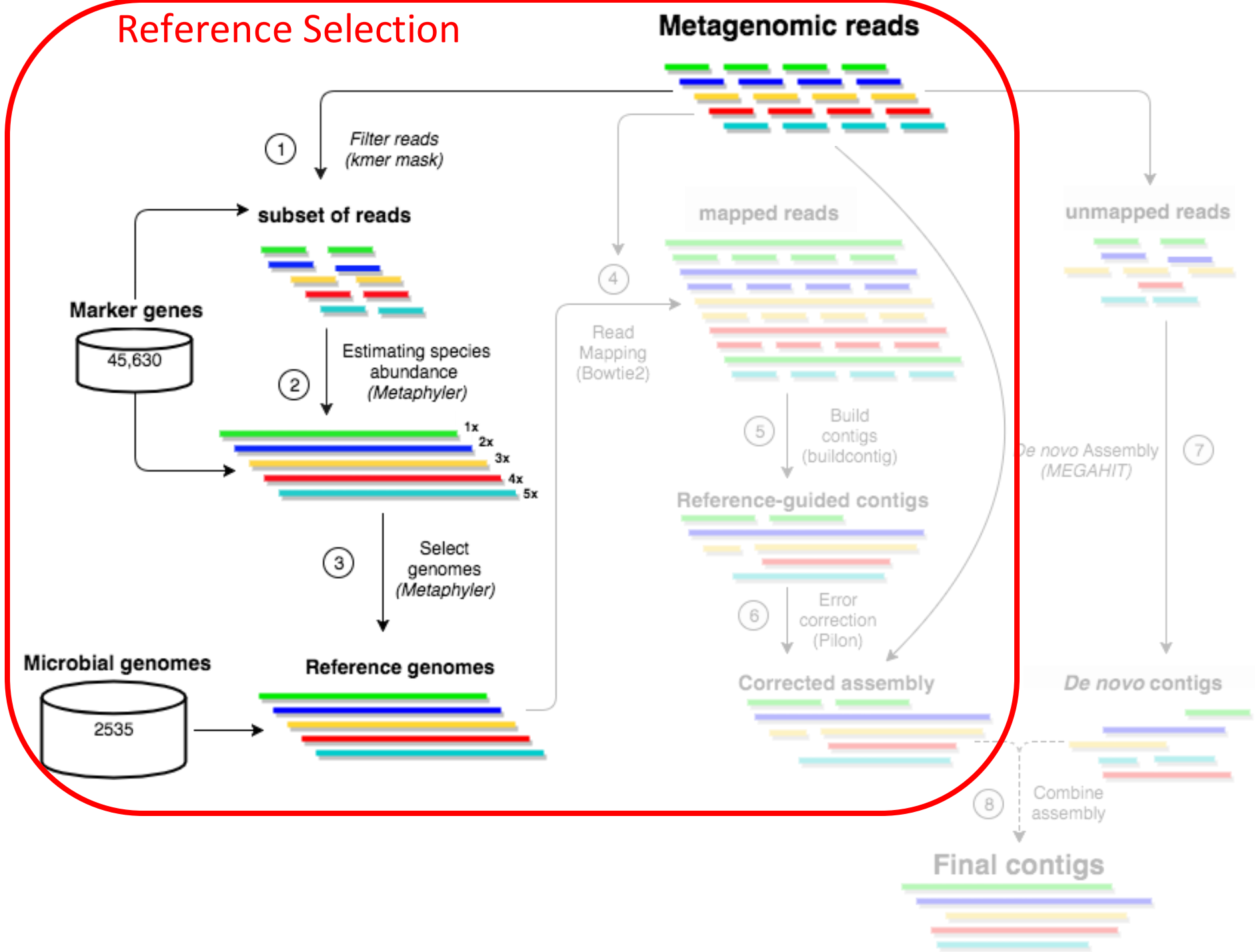
What makes a puzzle hard?

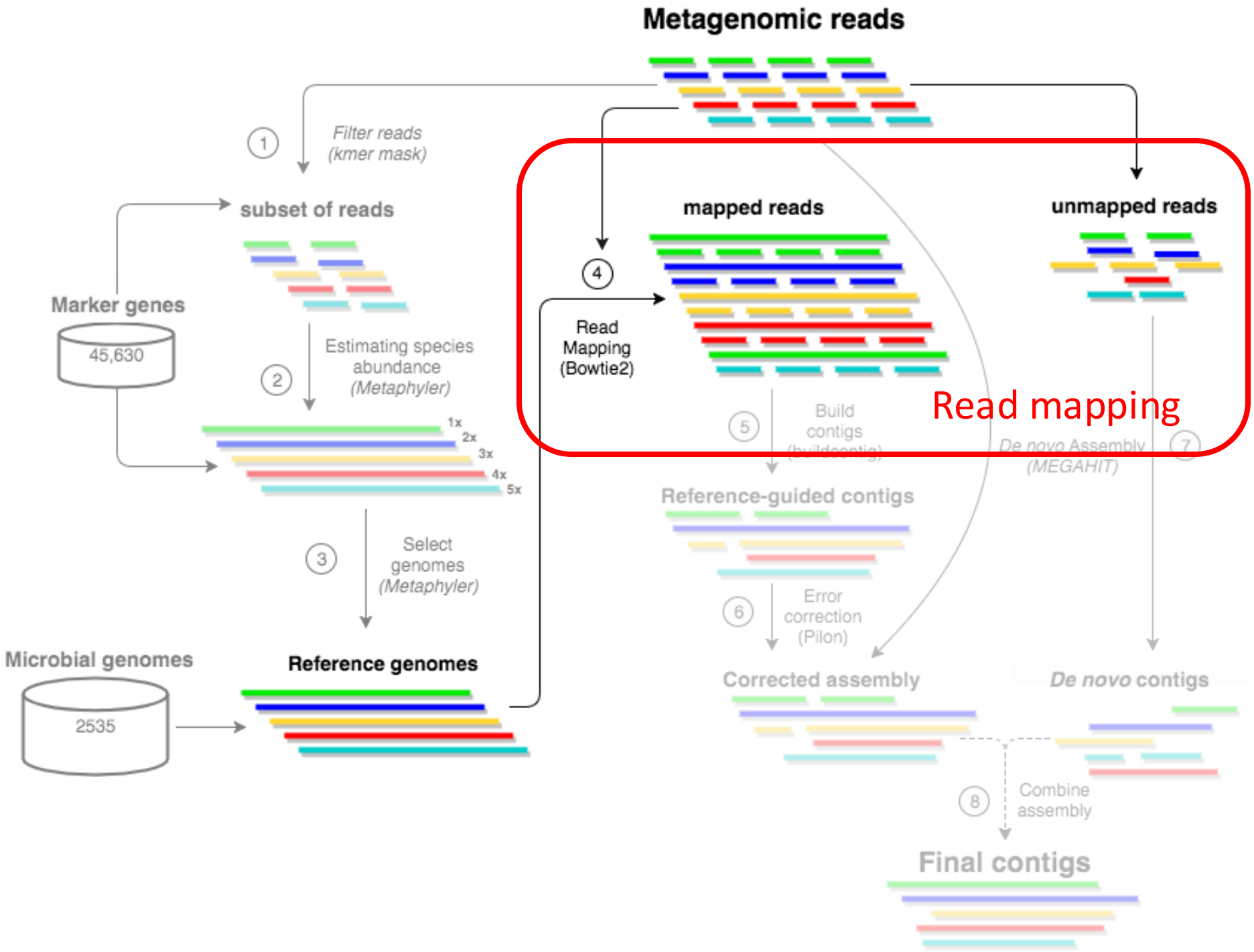


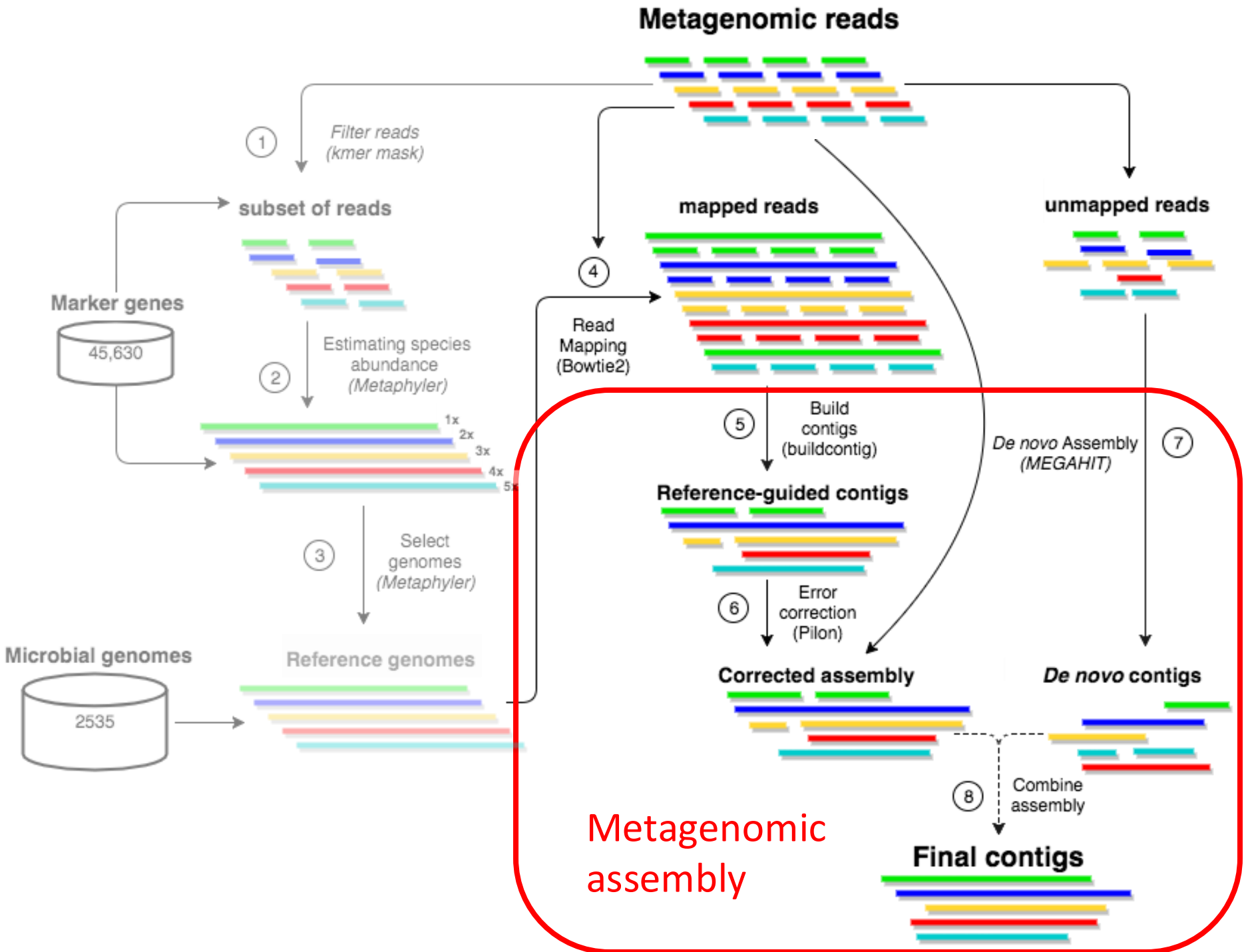
Slide credit Torsten Seemann



MetaCompass overview


Reference Selection
Metagenomic reads
1
Filter reads
(kmer mask)
subset of reads
Marker genes
45,630
2
Estimating species
abundance
(Metaphyler)
1x
2x
3x
4x
5x
3
Select
genomes
(Metaphyler)
Microbial genomes
2535
Reference genomes
4
Read
Mapping
(Bowtie2)
mapped reads
5
Build
contigs
(buildcontig)
Reference-guided contigs
6
Error
correction
(Pilon)
Corrected assembly
unmapped reads
De novo Assembly
(MEGAHIT)
7
De novo contigs
8
Combine
assembly
Final contigs


Metagenomic reads
1
Filter reads
(kmer mask)
subset of reads
Marker genes
45,630
2
Estimating species
abundance
(Metaphyler)
1x
2x
3x
4x
5x
3
Select
genomes
(Metaphyler)
Microbial genomes
2535
Reference genomes
4
Read
Mapping
(Bowtie2)
mapped reads
unmapped reads
Read mapping
5
Build
contigs
(buildcontig)
Reference-guided contigs
6
Error
correction
(Pilon)
Corrected assembly
De novo Assembly
(MEGAHIT)
7
De novo contigs
8
Combine
assembly
Final contigs

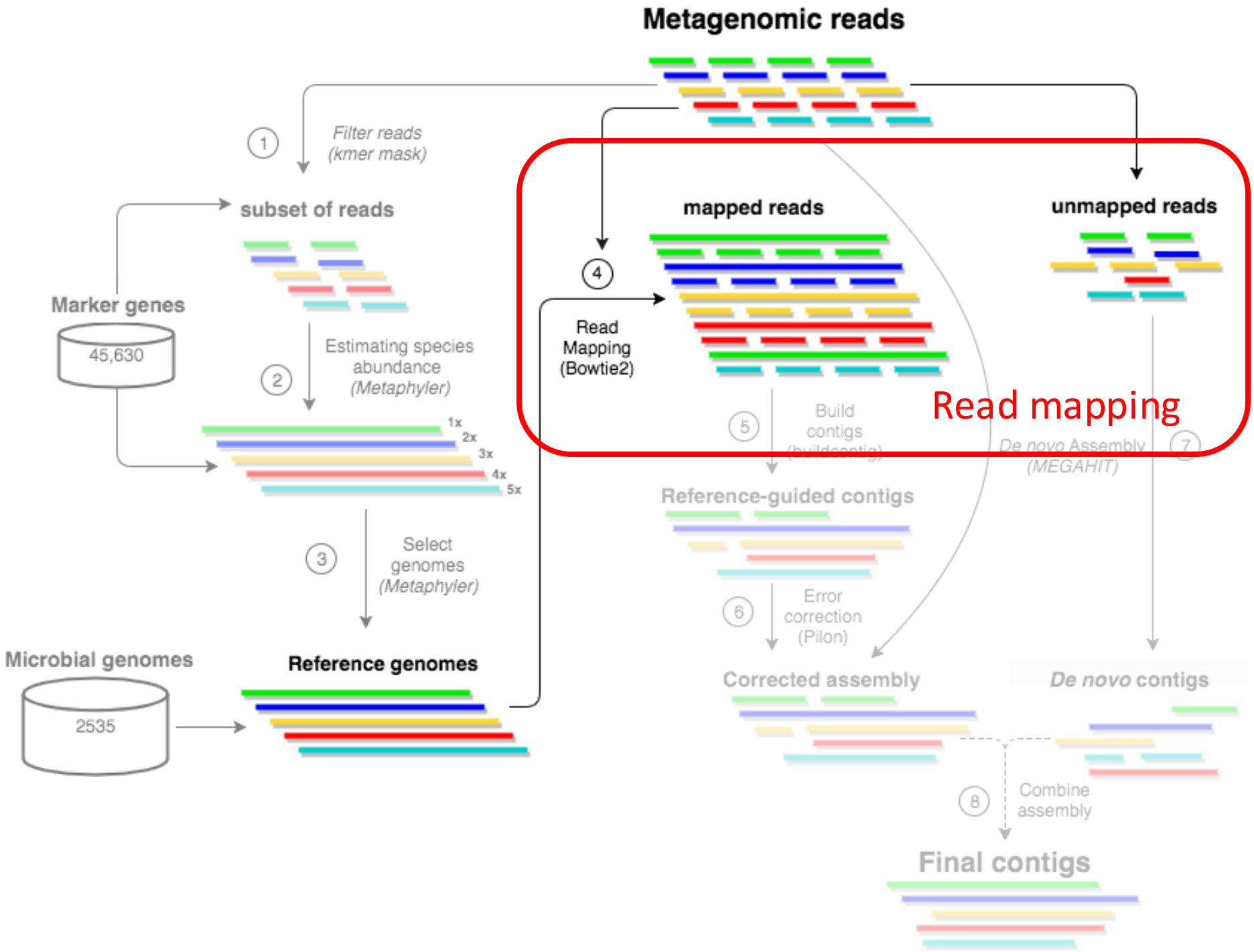

Metagenomic reads
1
Filter reads
(kmer mask)
subset of reads
Marker genes
45,630
2
Estimating species
abundance
(Metaphyler)
1x
2x
3x
4x
5x
3
Select
genomes
(Metaphyler)
Microbial genomes
2535
Reference genomes
4
Read
Mapping
(Bowtie2)
mapped reads
unmapped reads
5
Build
contigs
(buildcontig)
Reference-guided contigs
6
Error
correction
(Pilon)
Corrected assembly
De novo Assembly
(MEGAHIT)
7
De novo contigs
8
Combine
assembly
Final contigs
Metagenomic
assembly

Choose your own adventure: how shall we identify the reference genomes in our microbiomes?

1. Universal marker gene- based approaches (MetaPhlan, etc)

2. MinHash based approaches (SourMash, etc)

3. Kmer + LCA based approaches (Kraken2, etc)


Metagenomic reads
1
Filter reads
(kmer mask)
subset of reads
Marker genes
45,630
2
Estimating species
abundance
(Metaphyler)
1x
2x
3x
4x
5x
3
Select
genomes
(Metaphyler)
Microbial genomes
2535
Reference genomes
4
Read
Mapping
(Bowtie2)
mapped reads
unmapped reads
Read mapping
5
Build
contigs
(buildcontig)
Reference-guided contigs
6
Error
correction
(Pilon)
Corrected assembly
De novo Assembly
(MEGAHIT)
7
De novo contigs
8
Combine
assembly
Final contigs

STEP 2: Read mapping

Software | Open Access

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

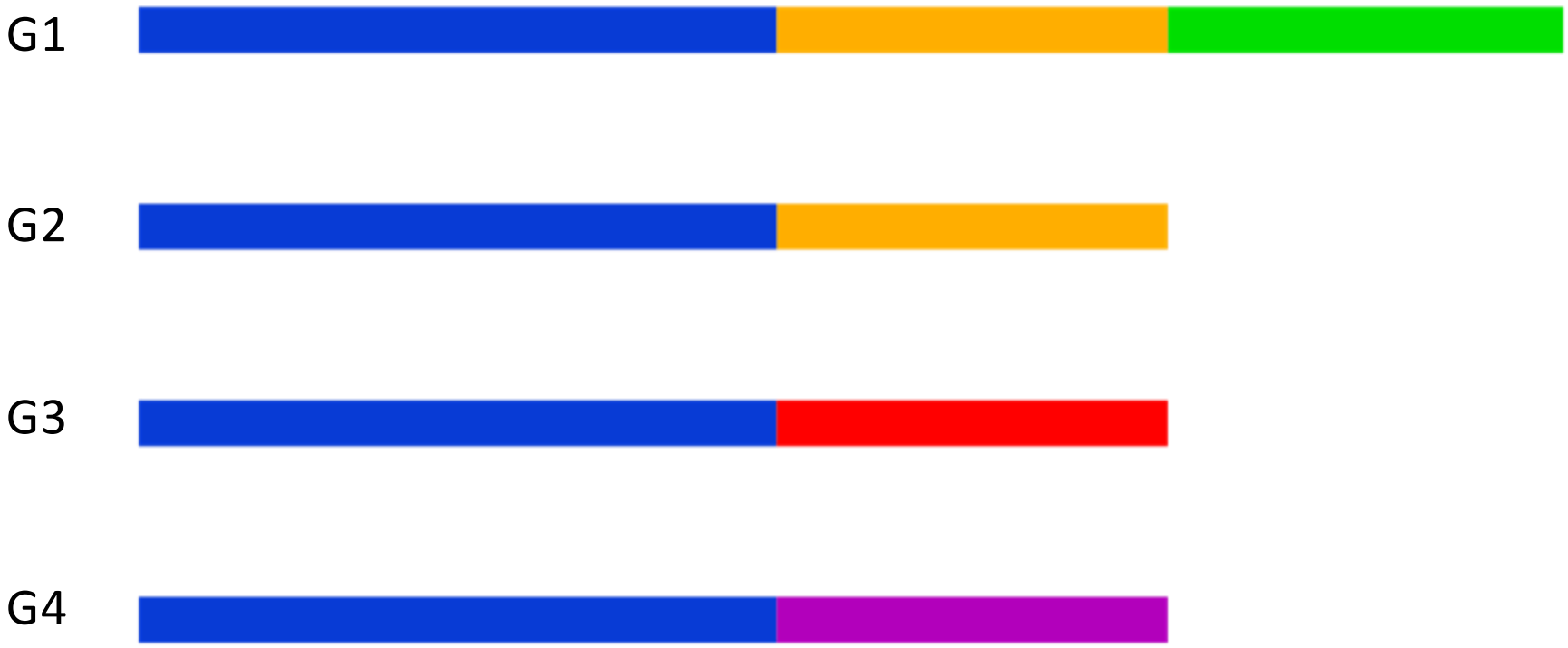
Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg

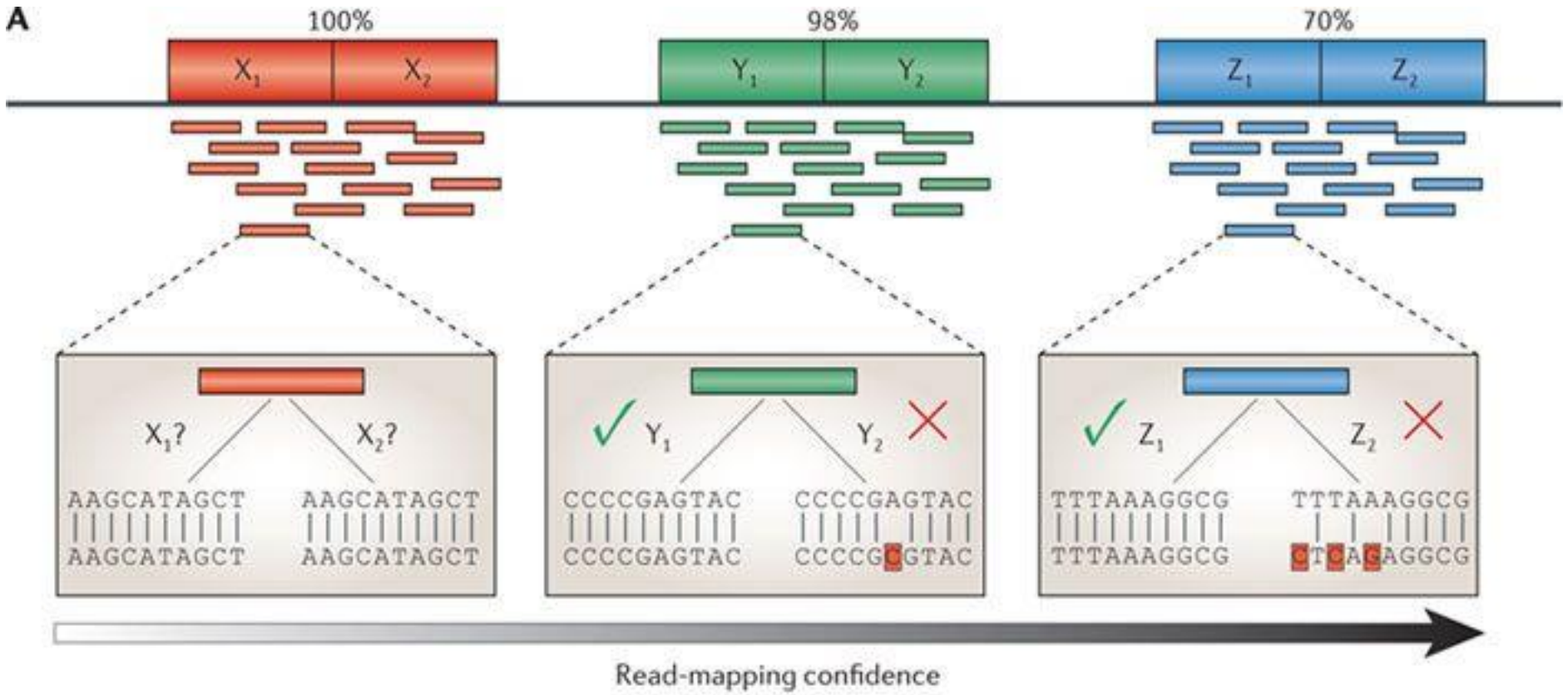
Genome Biology 2009 **10**:R25

https://doi.org/10.1186/gb-2009-10-3-r25

Received: 21 October 2008 | **Accepted:** 4 March 2009 | **Published:** 4 March 2009

STEP 2: Read mapping





Nature Reviews | Genetics

Choose your own adventure: how shall we map reads to the recruited genomes?

1. All: map all reads to every equally good mapping location

1. Random: randomly assign reads amongst equally good mapping location

1. Depth: genome with highest depth of coverage takes all of the reads that map to it

1. Breadth: genome with highest breadth of coverage takes all the reads that map to it

READ MAPPING: How to place reads?

Random assignment "coin flip"

Rank by depth of coverage

Rank by minimum set cover

Minimum set cover?

- Given a set of elements U { 1 , 2 , . . . , n } and a collection S of m sets whose union equals the universe, the set cover problem is to identify the smallest sub-collection of S whose union equals the universe.

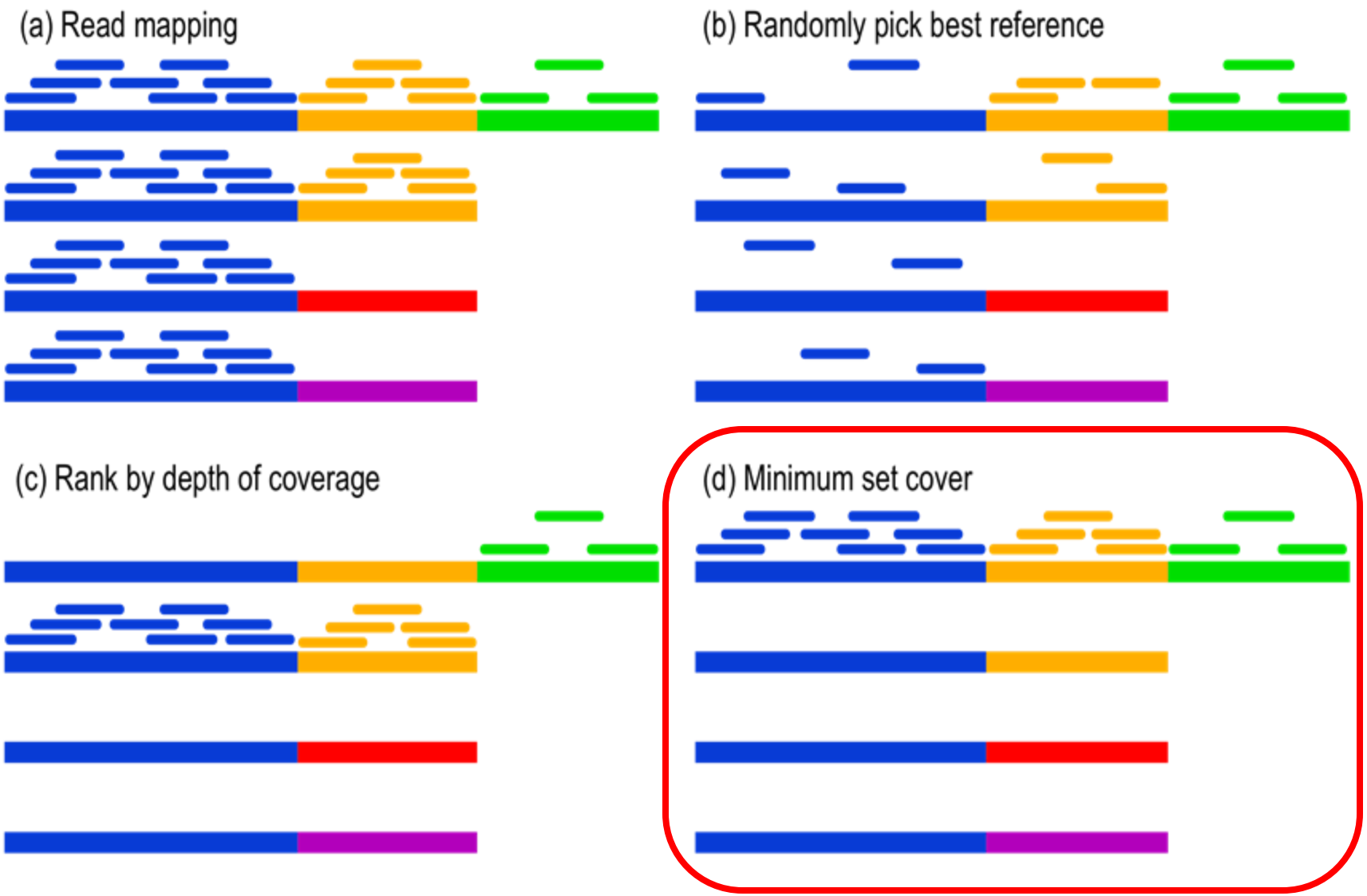
- For example, consider the universe U = { 1 , 2 , 3 , 4 , 5 } and the collection of sets S = { { 1 , 2 , 3 } , { 2 , 4 } , { 3 , 4 } , { 4 , 5 } } the union of S is U.

- The minimum set cover is the smallest number of m sets that cover U:

{ { 1 , 2 , 3 } , { 4 , 5 } }

- **For reference selection, it's the smallest number of genomes that cover all of the input reads.**

Read mapping selection: Minimum set cover



STEP 3: Building the contigs

```
                                   C
Reads      TGCACGGATG   TGCATGCACG
     TTAATGCACG   TG-ATGCATG
TGGATTAATG   TGGATG-ATG   C
TGGATTCATGCATGGATGCATGCATGCACG   Reference
TGGATTAATGCACGGATG-ATGCACTGCACG  Contig
```

Min. depth of coverage:2
Min. length:10

STEP 4: de novo assembly

Assembly unmapped reads to reference

De novo assembly using MEGAHIT

Evaluation Datasets

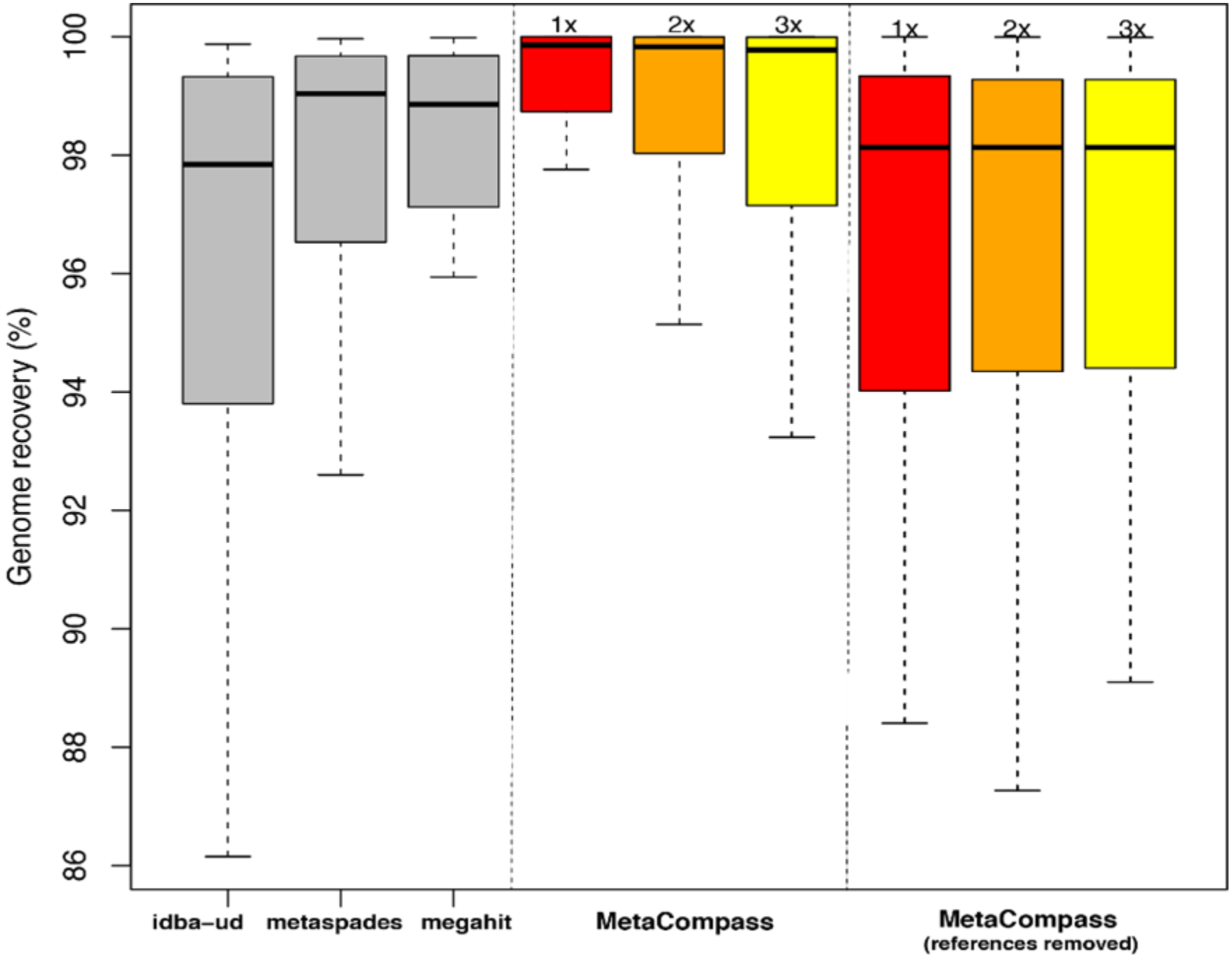
- Dataset 1: Synthetic dataset, Shakya et. al.
- Dataset 2: Down-sampled Dataset 1(low coverage)
- Dataset 3: 2,077 samples from HMP2

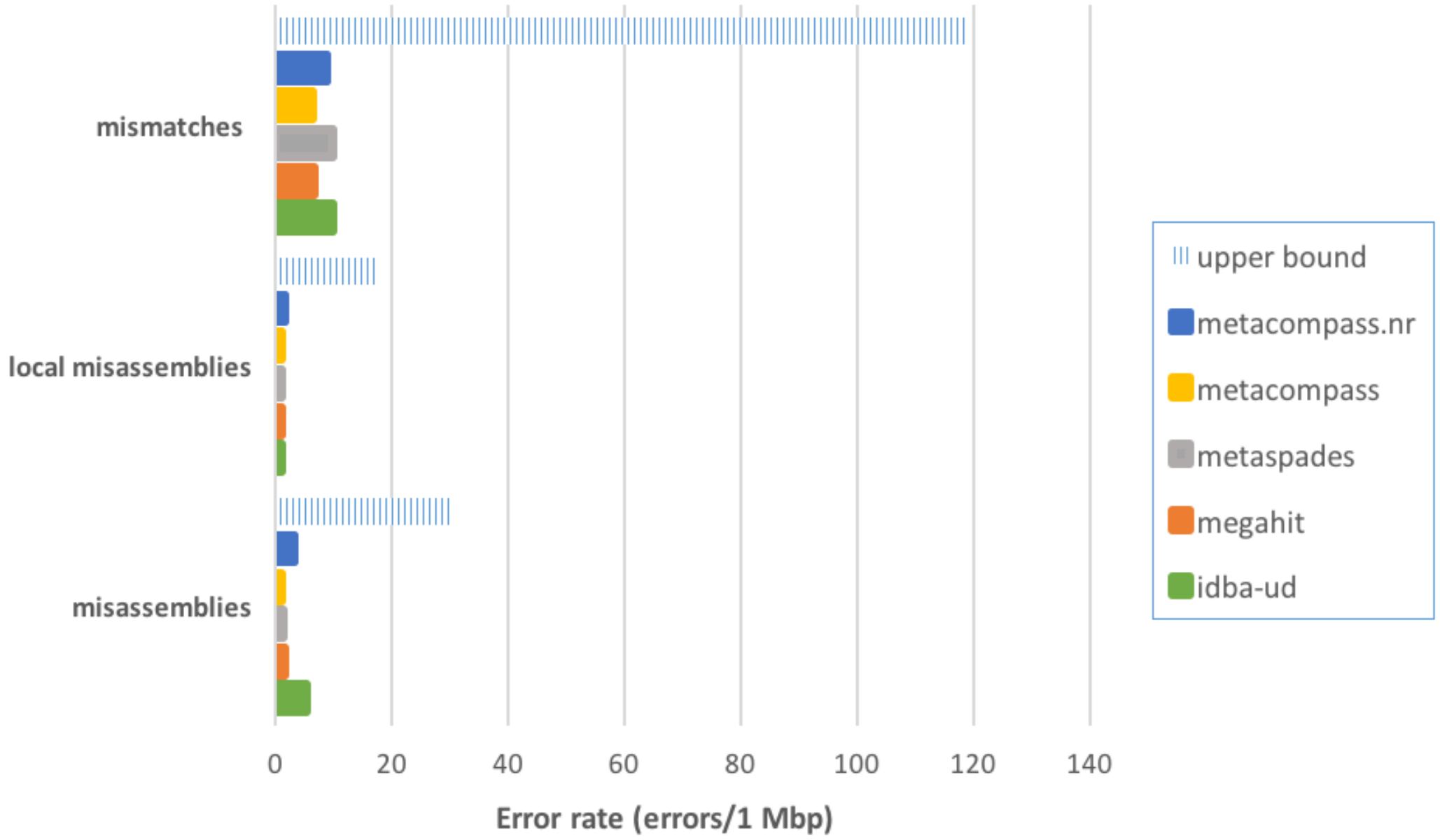
Results - Dataset 1

- Mixture of 64 bacterial and archaeal species (Shakya et al., 2013)
- 109 million reads with mean insert size 206 bp and 100 bp read length
- Easier to evaluate assembly since the truth is known

Results – Assembly Statistics

Method	No. Contigs	Longest Contig (bp)	Median genome recovery	Mismatches (Per 1Mbp)	Misassemblies (Per 1Mbp)
MetaCompass	**18,766**	**7,057,109**	**100%**	**61.9**	**1.9**
IDBA-UD	22,355	991,792	98%	98.6	6.3
MEGAHIT	35,351	1,151,857	99%	66.5	2.5
metaSPAdes	21,424	1,438,235	99%	97.1	2.3


Genome recovery (%)
100
98
96
94
92
90
88
86
1x
2x
3x
1x
2x
3x
idba–ud
metaspades
megahit
MetaCompass
MetaCompass
(references removed)


mismatches
local misassemblies
misassemblies
0
20
40
60
80
100
120
140
Error rate (errors/1 Mbp)
upper bound
metacompass.nr
metacompass
metaspades
megahit
idba-ud