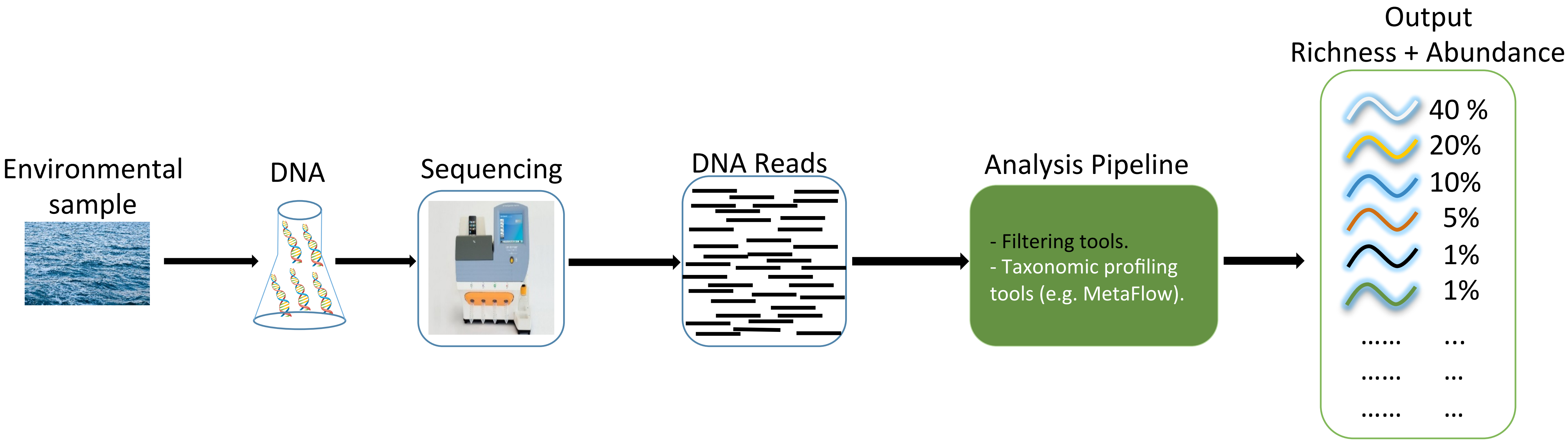
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

TIETOJENKÄSITTELYTIETEEN LAITOS
INSTITUTIONEN FÖR DATAVETENSKAP
DEPARTMENT OF COMPUTER SCIENCE

MetaFlow: Metagenomic profiling based on whole-genome coverage analysis with min-cost flows

Ahmed Sobih, Alexandru I. Tomescu, Veli Mäkinen

Metagenomic taxonomic profiling



Environmental sample

DNA

Sequencing

DNA Reads

Analysis Pipeline

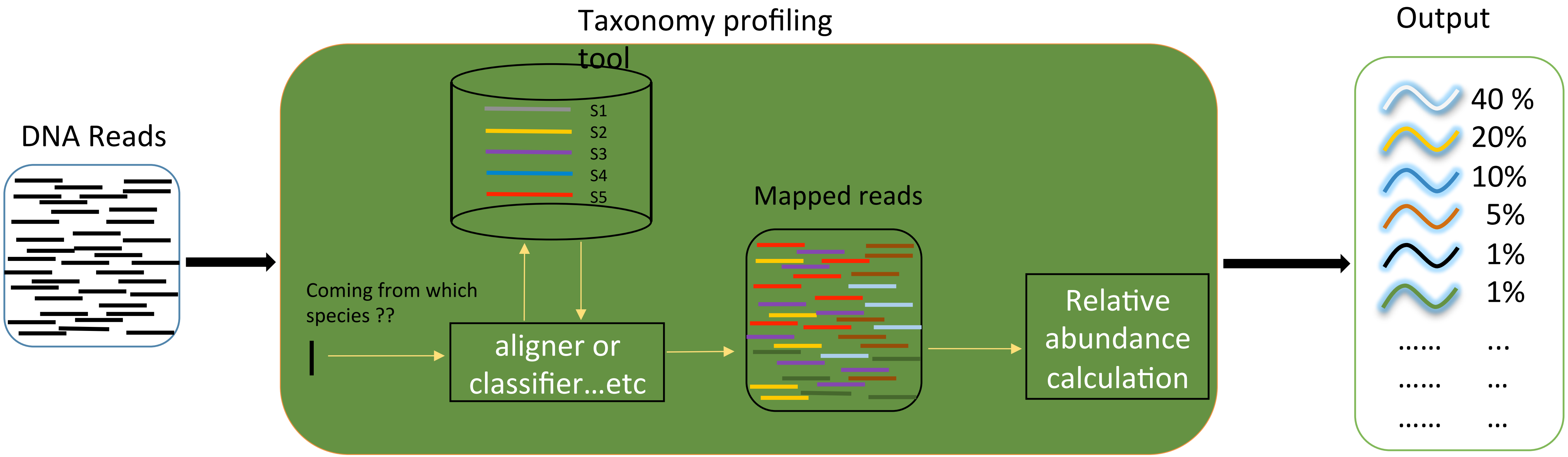
- Filtering tools.
- Taxonomic profiling tools (e.g. MetaFlow).

Output
Richness + Abundance

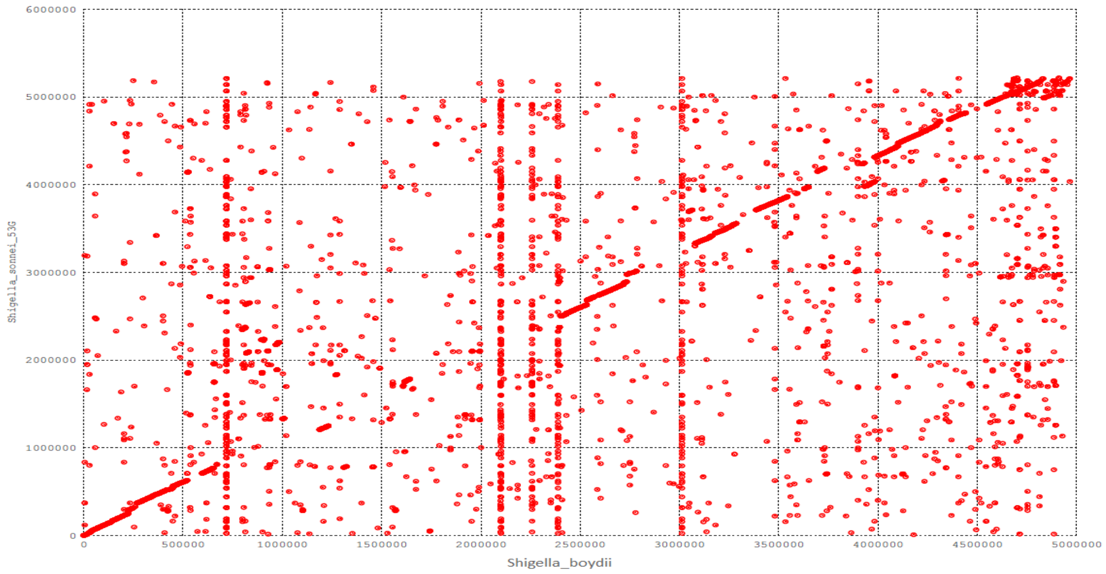
40 %
20%
10%
5%
1%
1%
....
....
....

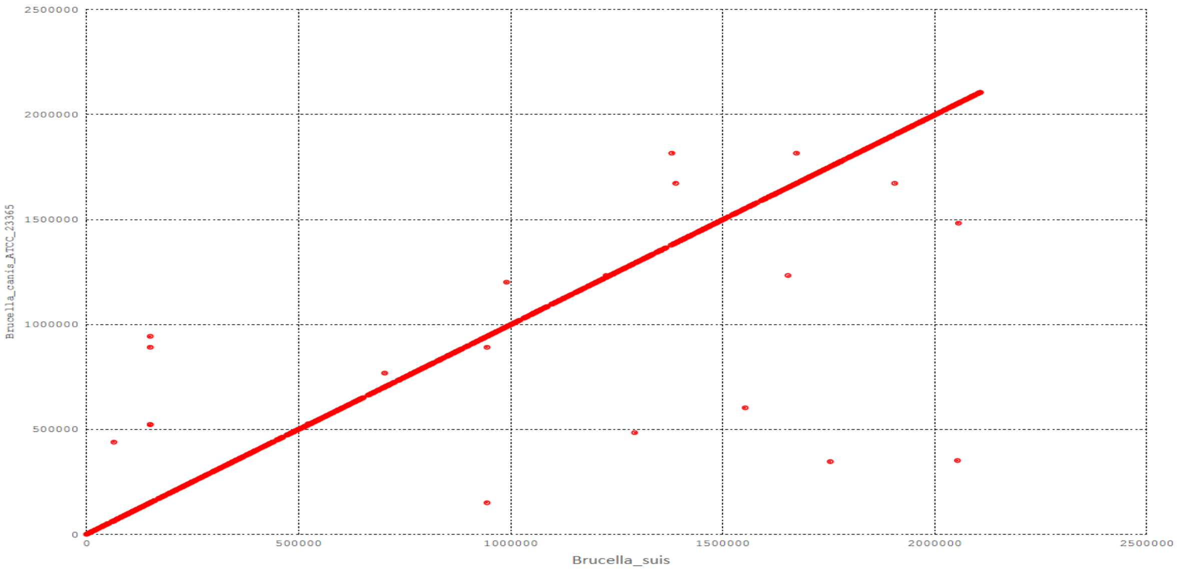
Shotgun taxonomy-dependent analysis

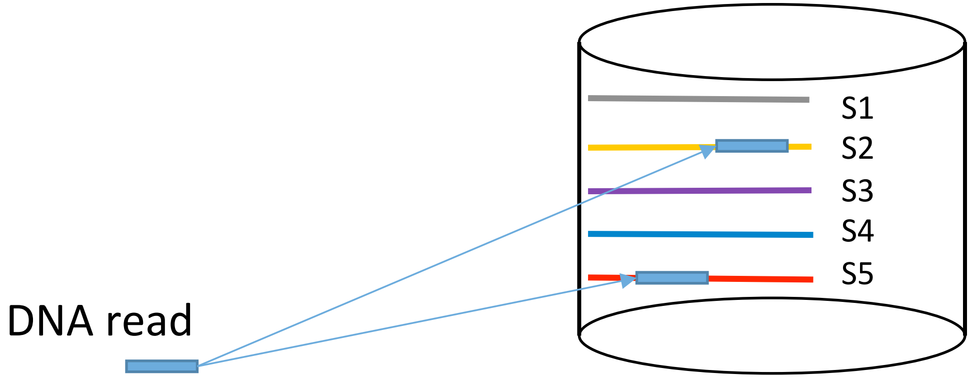
- Shotgun: Sequencing all of the DNA materials.
- Taxonomy-Dependent: Using a reference DB of genomes, HMMs, or marker genes…etc.



Challenges: Similarity between microbial species

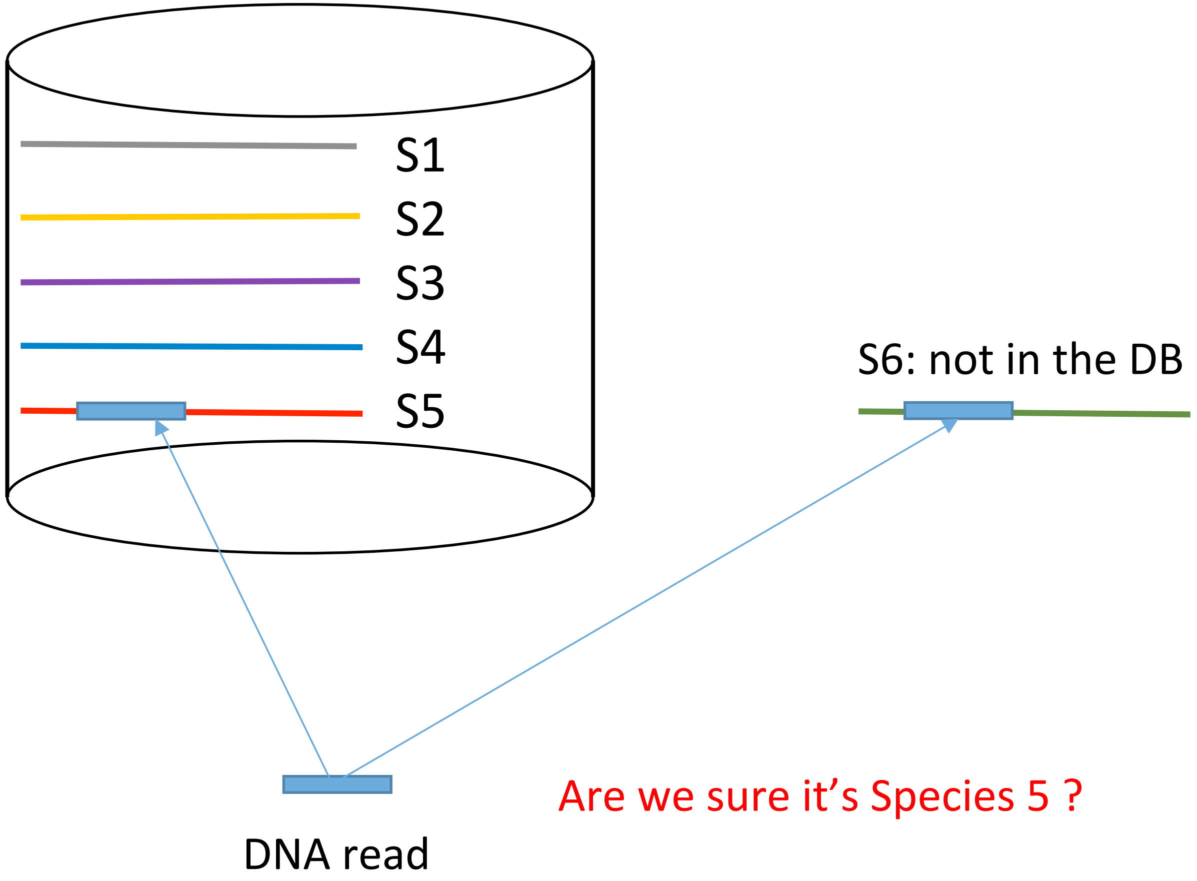






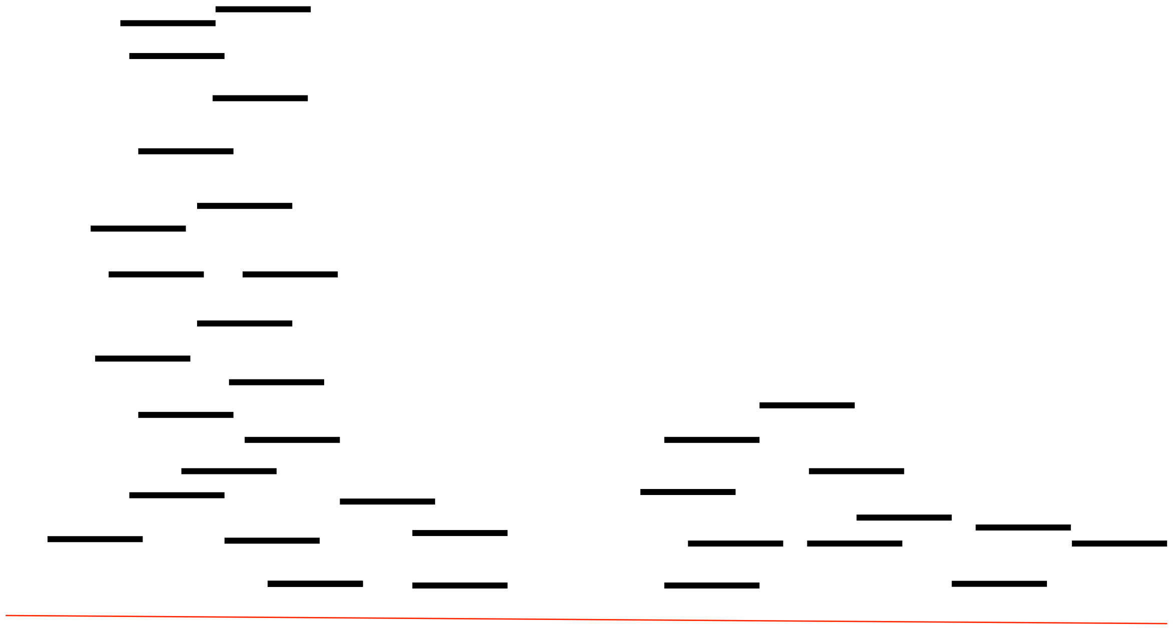
Species 2 or Species 5 ?

Challenges: Incomplete reference DB



Are we sure it’s Species 5 ?

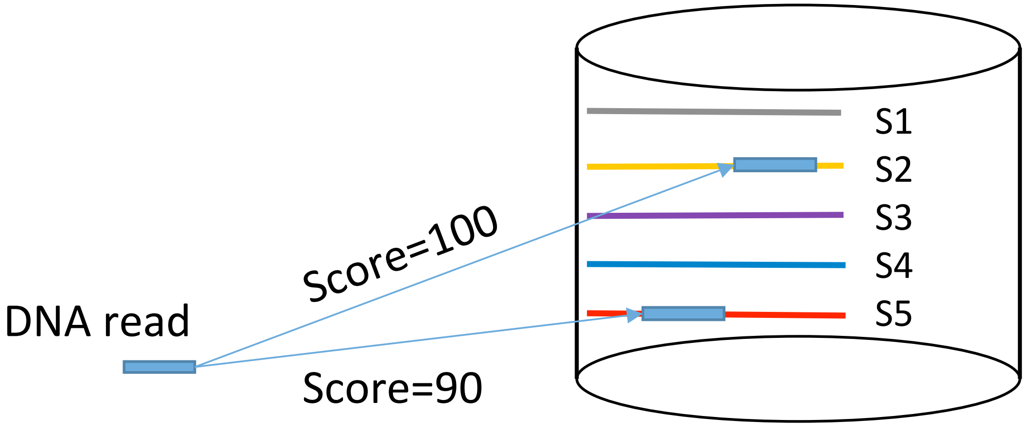
Challenges: Sequencing biases



Species 1

What is the abundance ?

Challenges: Sequencing errors



Are we sure it’s Species 2 ?

How to:

- Break ties between equally good alignments.
- Minimize (or eliminate) false positives.
- Minimize (or eliminate) false negatives.
- Calculate abundances accurately.
- Estimate the abundances of unknown species.

Metagenomics taxonomic profiling tools

- MEGAN (Huson, D et al., Genome research 2007).
- PhymmBL (Brady, A and Salzberg, S, Nat. methods 2009).
- NBC (Gail L. Rosen et al., Bioinformatics 2010).
- MetaPhlAn (Segata, N et al., Nat. methods 2012).
- mOTU (Sunagawa, S et al., Nat. methods 2013).
- GSMer (Qichao Tu et al., Nucleic Acids Res 2014).

MetaFLow: Coverage sensitive metagenomic mapping

- Input:

  - A set of BLAST hits of the metagenomics reads inside a collection of reference genomes.

- Output:

  - The richness of the sample and the relative abundance of each known species.

- Objective:

  - Select a subset of read mapping where every read will have exactly one hit in a reference genome, or classifying it as originating from a species not in the reference database, such that:

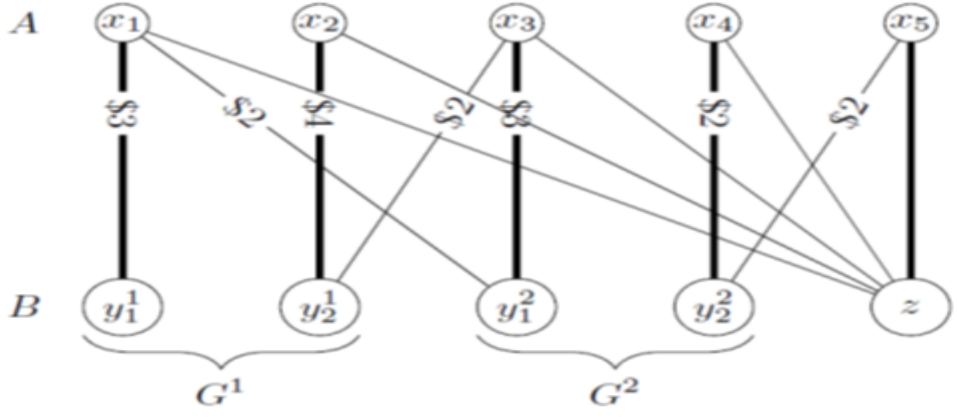
    a- Most regions of each reported genome are covered.

    b- Read coverage in each genome is close to uniform.

    c- Take into account the BLAST scores of the mappings.

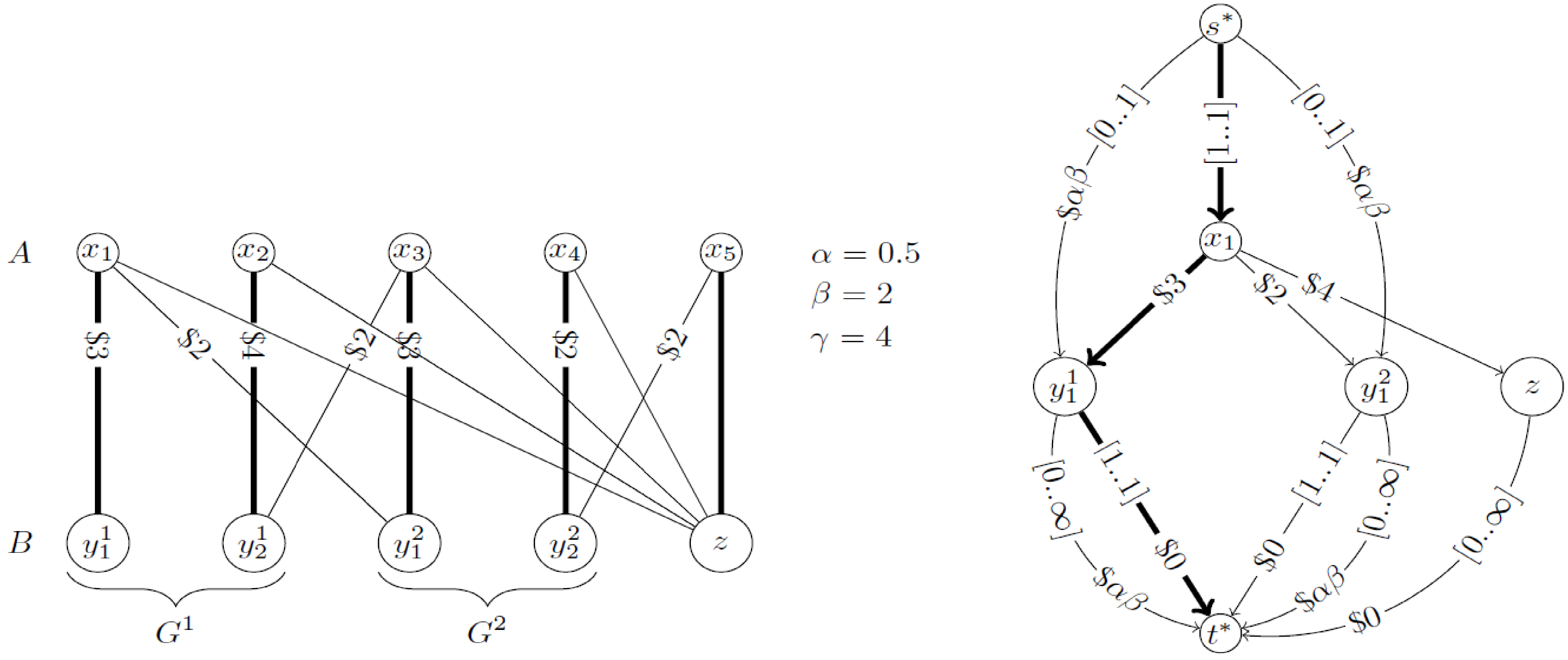
MetaFLow: Coverage sensitive metagenomic mapping

- Breaking each reference genome into substrings of equal length (chunks).
- Introducing the unknown node (Z) to which all reads can be mapped to.
- Matching problem in a bipartite graph inspired by Lo et al. (2013).
- NP-hard.



(a)

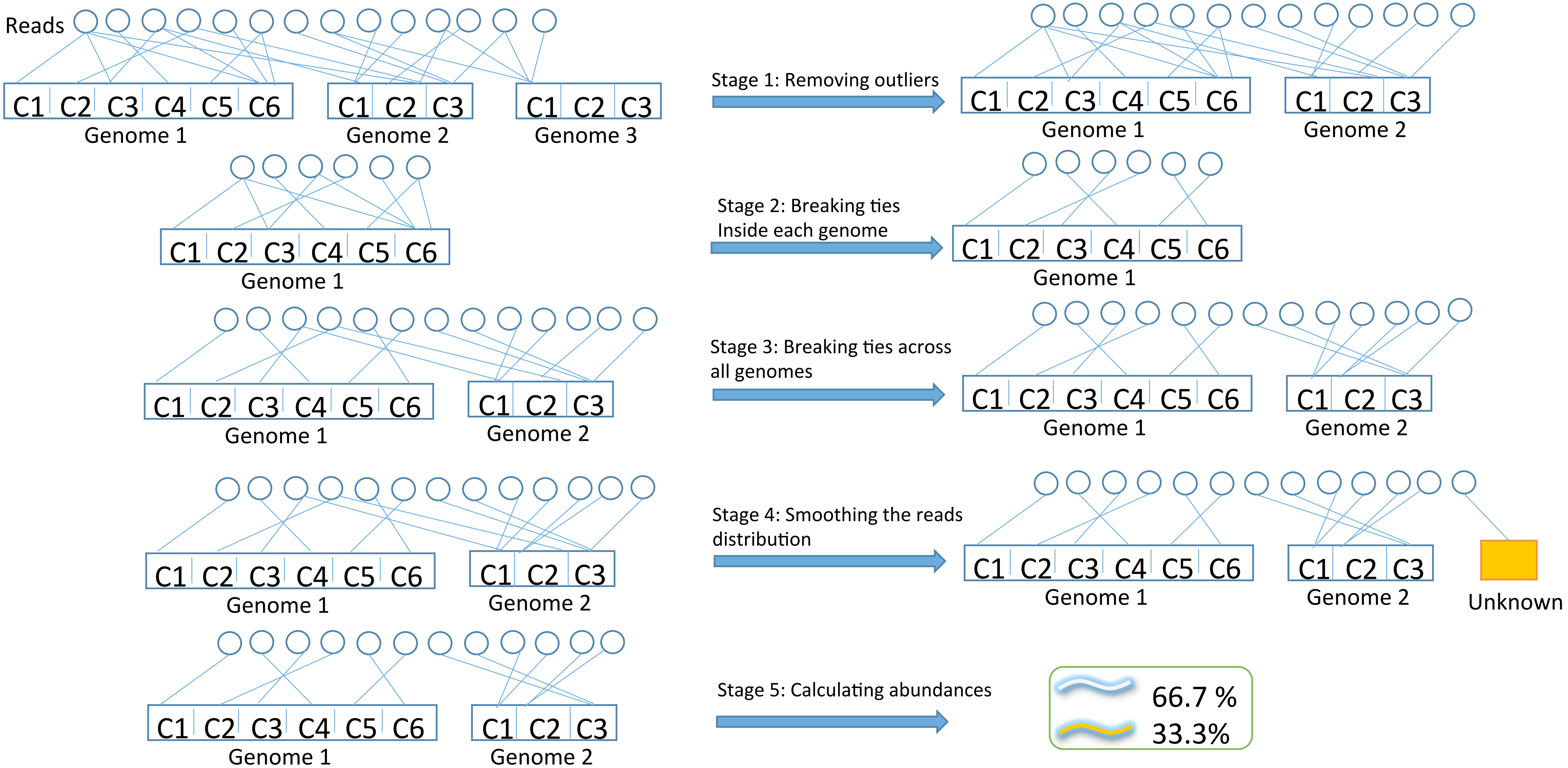
MetaFlow: Reduction to min-cost flows problem



(a)

(b)

MetaFlow: Algorithm



Stage 1: Removing outliers

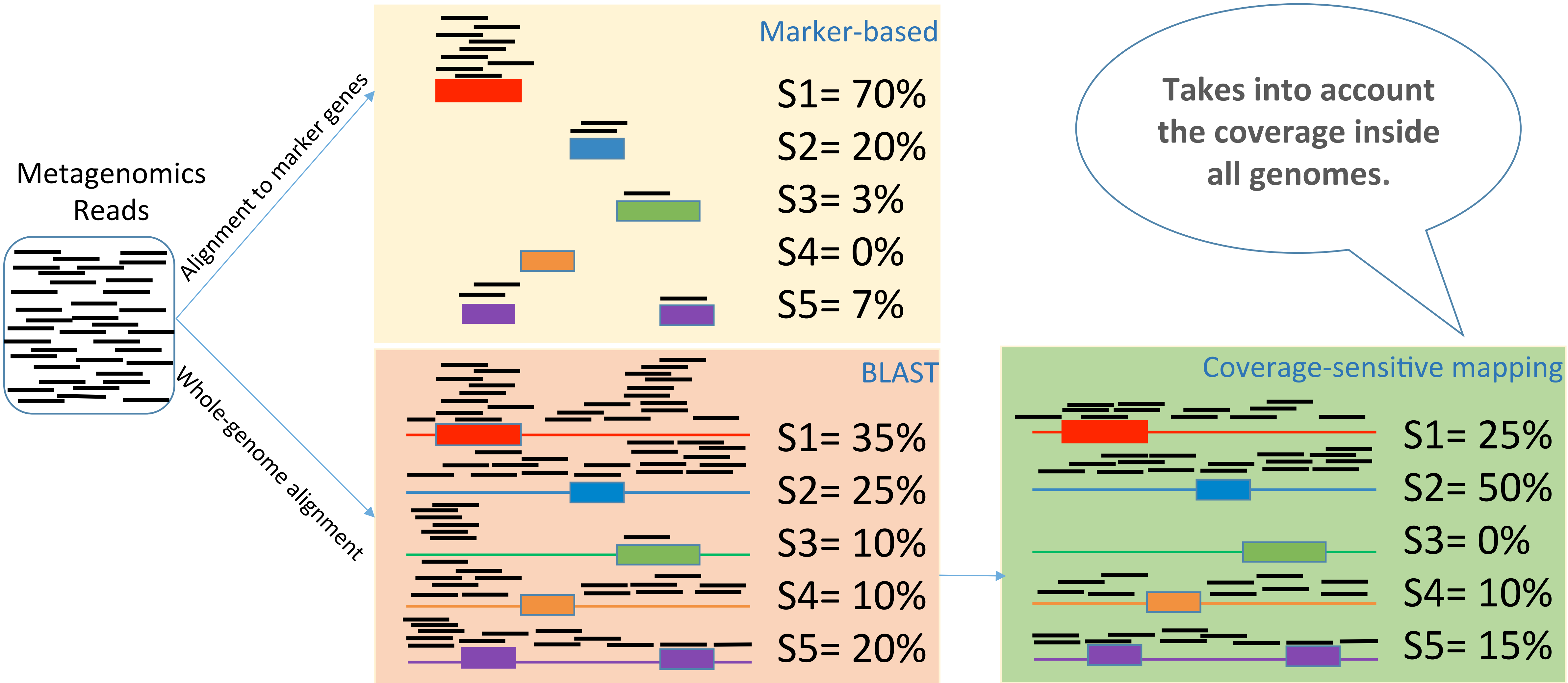
Stage 2: Breaking ties
Inside each genome

Stage 3: Breaking ties across
all genomes

Stage 4: Smoothing the reads
distribution

Stage 5: Calculating abundances

MetaFLow: Coverage sensitive metagenomic mapping



Metagenomics Reads

Alignment to marker genes

Whole-genome alignment

Marker-based

S1= 70%
S2= 20%
S3= 3%
S4= 0%
S5= 7%

BLAST

S1= 35%
S2= 25%
S3= 10%
S4= 10%
S5= 20%

Coverage-sensitive mapping

S1= 25%
S2= 50%
S3= 0%
S4= 10%
S5= 15%

Takes into account the coverage inside all genomes.

Experiments: Simulated data (46 datasets)

	# of datasets	# of species per dataset	# of reads per sample	Unknown species %	Species Selection method
LC-Known	15	15	4 M	0%	Based on similarity
LC-Unknown	15	15	4 M	20%	Based on similarity
HC-Known	8	100	40 M	0%	Random
HC-Unknown	8	100	40 M	15%	Random

Evaluation Criteria:

- Accuracy of the richness estimations:

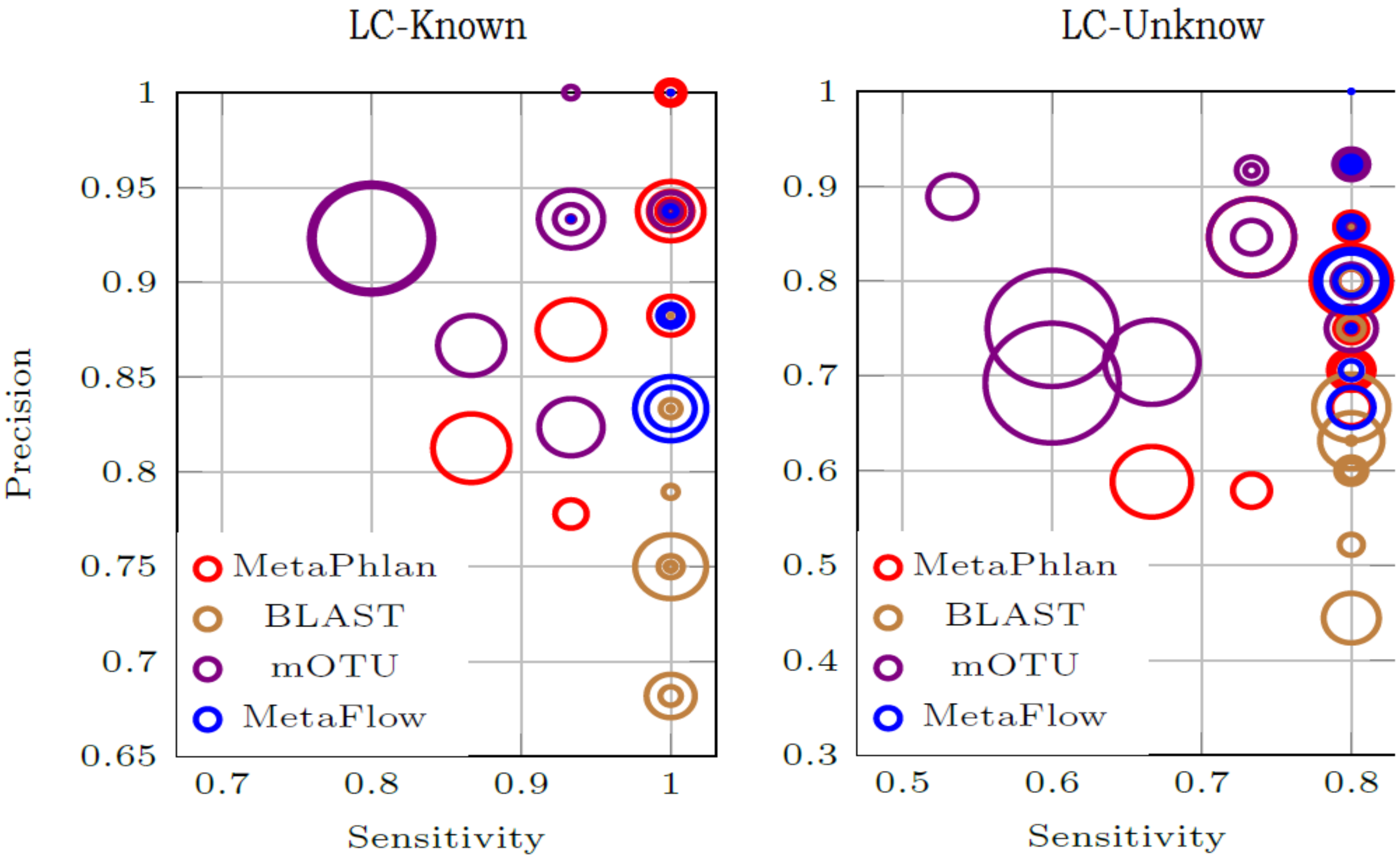
$$\mathit{Sensitivity} = \mathit{number\ of\ true\ postives} / \mathit{actual\ number\ of\ species\ in\ the\ sample}$$

$$\mathit{Precision} \quad = \mathit{number\ of\ true\ postives} / \mathit{number\ of\ predicted\ species}$$

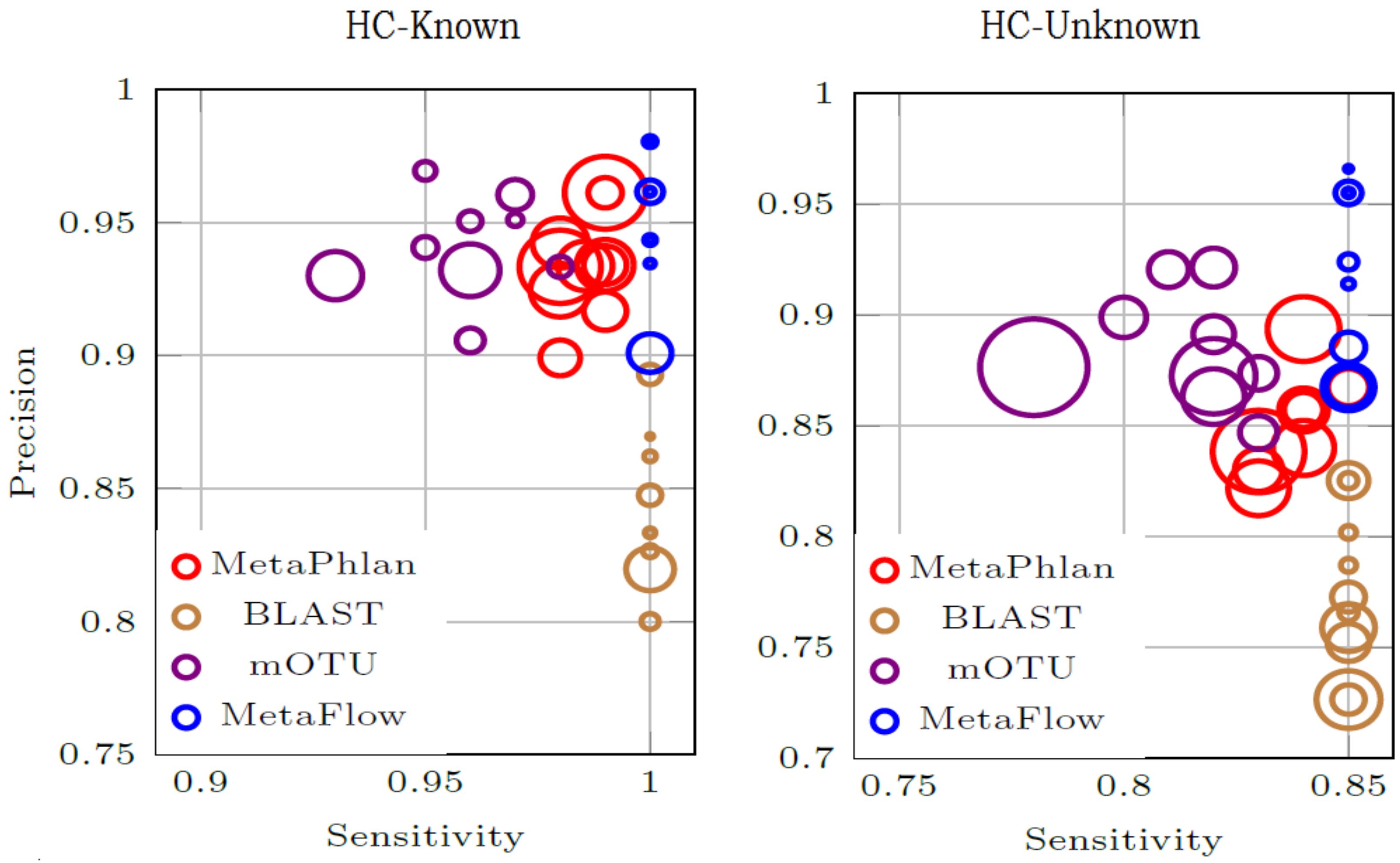
- Relative abundance predictions:

$$l_1\ \mathit{norm} \qquad = \sum_{k=1}^{n} |\mathit{actual\ abundance}_k - \mathit{predicted\ abundance}_k\,|$$

Results: Simulated data (LC)

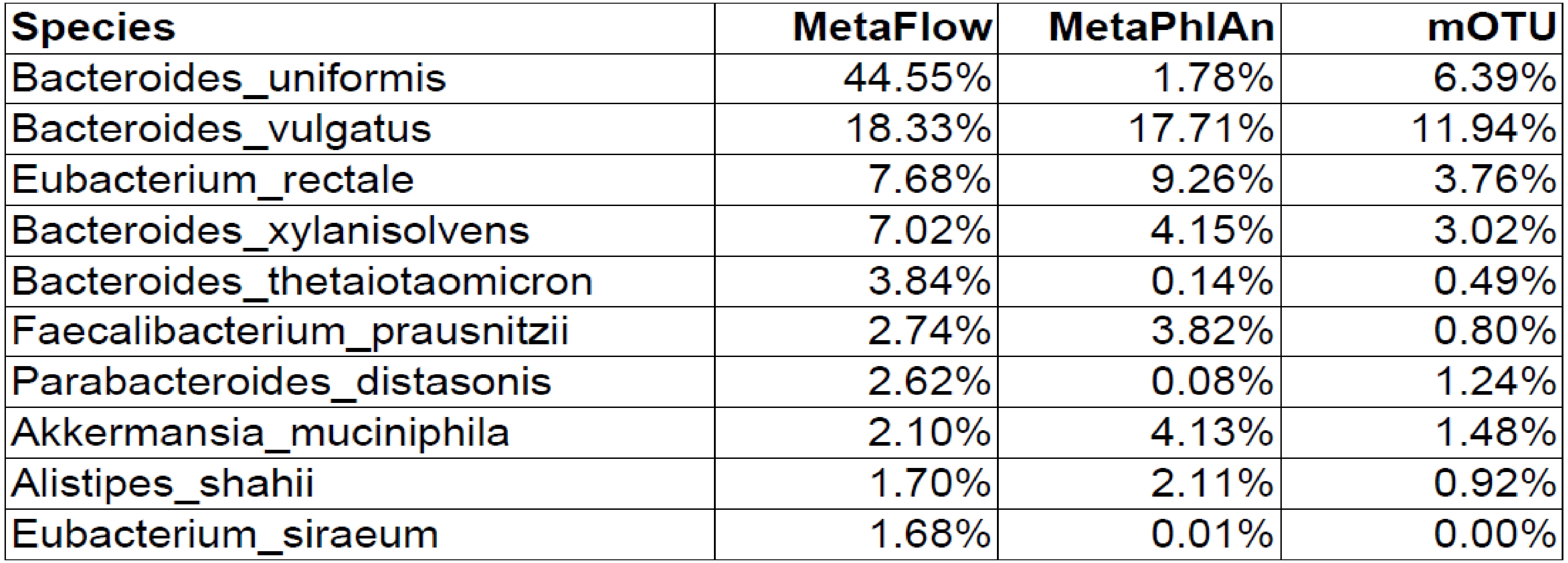


Results: Simulated data (HC)



Real metagenomics sample

- Merged 6 G_DNA_Stool samples of a female from the Human Microbiome Project.
- 287,565,377, out of which 82,486,518 BLAST mapped to one or more species.



Species	MetaFlow	MetaPhlAn	mOTU
Bacteroides_uniformis	44.55%	1.78%	6.39%
Bacteroides_vulgatus	18.33%	17.71%	11.94%
Eubacterium_rectale	7.68%	9.26%	3.76%
Bacteroides_xylanisolvens	7.02%	4.15%	3.02%
Bacteroides_thetaiotaomicron	3.84%	0.14%	0.49%
Faecalibacterium_prausnitzii	2.74%	3.82%	0.80%
Parabacteroides_distasonis	2.62%	0.08%	1.24%
Akkermansia_muciniphila	2.10%	4.13%	1.48%
Alistipes_shahii	1.70%	2.11%	0.92%
Eubacterium_siraeum	1.68%	0.01%	0.00%

Average running time (minutes)

	4M reads	40M reads	280M reads
MetaPhlAn	14	132	387
mOTU	9	84	380
GSMer	42	364	NA
BLAST	243	1572	3696
MetaFlow	28	459	2025

Summary

- Taking into account the coverage across the whole genome can improve the richness and abundance estimation.
- Coverage sensitive metagenomic mapping is NP-hard, and can be modeled using minimum cost flow.
- Perhaps a mixture between markers-based methods and whole genome coverage could give better estimates in a reasonable time.
- Estimating the abundance for ”unknown” species remains a challenging problem.

Thank you!