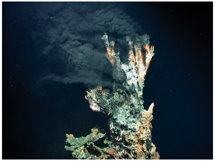
Current Topics in Bacterial Identification from Genomic Data

Dr. Gregoire Siekaniec & Prof. Dr. Sven Rahmann

Algorithmic Bioinformatics, Saarland University and Center for Bioinformatics,
Saarland Informatics Campus, Saarbrücken, Germany

Ubiquity of bacteria vs the difficulty to study them

- Unicellular, microscopic and prokaryotic living organisms.
- Present in diverse ecosystems (from deep-sea vents to clouds, the human gut, and dairy products).

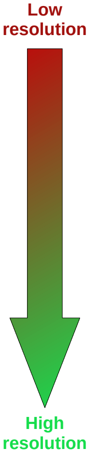


- Important role in health and ecology (e.g. carbon and nitrogen cycle).
- Many bacteria are unculturable ($\sim$ 98%*)
  → hard to study without sequencing their genomes
  → hard to classify taxonomically (lack of clear definition, HGT, ...)

* Wade. 2002. Unculturable bacteria—the uncharacterized organisms that cause oral infections. J. R. Soc. Med.

Identification methods

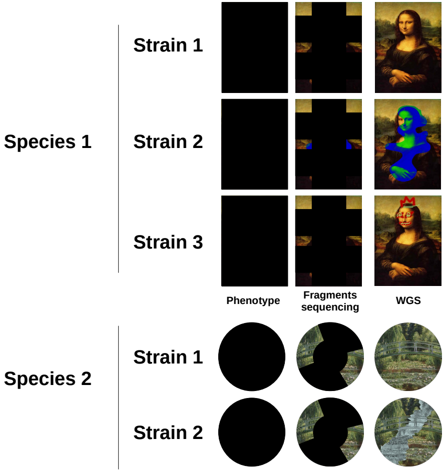
- Phenotypic information (Microscopy, Gram stain, Metabolism) → shape, motility, biofilm formation, *trophic type*.
- Genotypic informations
  - Biomolecular methods (Restriction enzyme digestion, DDH).
  - Methods based on genome sequencing:
    - Fragments sequencing:
      - 16S rRNA gene amplification (PCR) and sequencing
      - Multilocus sequence typing (MLST): analyses of few orthologous genes
    - **Whole genome/metagenome sequencing: Focus of the seminar**



Low resolution

High resolution

Resolution of the identification



Species 1

Strain 1

Strain 2

Strain 3

Phenotype

Fragments sequencing

WGS

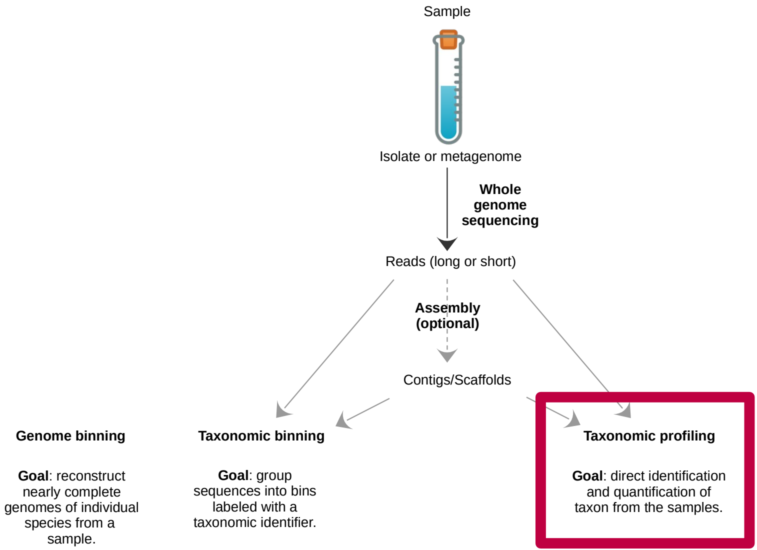
Species 2

Strain 1

Strain 2

Method used depends on the desired identification resolution.

Bacterial identification from sample



Sample

Isolate or metagenome

Whole genome sequencing

Reads (long or short)

Assembly (optional)

Contigs/Scaffolds

Genome binning

Goal: reconstruct nearly complete genomes of individual species from a sample.

Taxonomic binning

Goal: group sequences into bins labeled with a taxonomic identifier.

Taxonomic profiling

Goal: direct identification and quantification of taxon from the samples.

Focus of this seminar

Goal of this seminar

- Exploring the latest developments in bacterial identification software (taxonomic profiler).
- Topics will be selected from current research literature.
  - **focus on the algorithms and data structures** that underpin the tools.

Basic of (almost) all methods

- **Indexing part**: Compact storage of the reference sequences (typically assemblies).
- **Query part**: Querying the index with reads, which leads to the assignment of reads to reference sequences.
- **Identification and quantification part**: Identifying taxa from the read assignments and quantifying their abundance.

What is expected from you (in groups of 2 or 3)

1. **Summary Task**: Read, understand, and write a **3-4 pages summary** of a recent scientific paper, **focusing primarily on the methodology**.

   ⚠ This will require consulting related papers, software documentation, and supplementary materials to understand how the software operates.

   - **Draft Submission**: Submit a **1-page draft**, outlining the summary plan in bullet points.
   - **Summary Submission**: Complete and submit the final version of the summary.
   - **Revisions**: **Make modifications** to your summary based on feedback received.

   Progress Meeting: Attend a meeting to discuss progress (optional, but recommended).

2. **Oral Presentation**: Prepare and deliver a **35-minutes oral presentation** of your summary, followed by a **10-minutes** Q&A session.

   - **Question for Peers**: Each group must prepare and **ask at least one question** to the group that is presenting.

List of 13 proposed scientific papers (from the oldest to the newest)

- Centrifuge [Kim et al., 2016]
- MetaMLST [Zolfo et al., 2017]
- StrainSeeker [Roosaare et al., 2017]
- KrakenUniq [Breitwieser et al., 2018]
- Kraken2 [Wood et al., 2019]
- Metamaps [Dilthey et al., 2019]
- ganon [Piro et al., 2020]
- CCMetagen [Marcelino et al., 2020]
- ORI [Siekaniec et al., 2021]
- StrainFLAIR [Silva et al., 2021]
- StrainGE [van Dijk et al., 2022]
- MetaBIDx [Pham and Phan, 2024]
- Mibianto [Hirsch et al., 2024]

Most of the papers use:

- k-mer-based approach.
- at least one other tool/method.
- a database of known references such as Refseq (NCBI Reference Sequence Database).

1 - Centrifuge [Kim et al., 2016]

Developed to identify and categorize species from metagenomics data. For short sequencing reads but can handle long reads and is even use in What's In My Pot (real-time metagenomics analysis tool developed by Oxford Nanopore Technologies).

- Indexing part: data structure based on the Ferragina-Manzini (FM) index.
- Query part: Own specific read mapping.
- Identification/Quantification part: bottom-up algorithm on taxonomic tree and Expectation-Maximization (EM) algorithm.

Better to know

BWT, FM index, basic mapping algorithm, EM algorithm

2 - MetaMLST [Zolfo et al., 2017]

Designed for strain-level identification of bacteria in metagenomic samples using a method known as Multilocus Sequence Typing (MLST).

- Indexing part: no real indexation of the data. Use of a reference database of MLST profiles.
- Query part: Mapping of reads against MLST profils using Bowtie2 (FM Index) + reconstruction of the MLST loci.
- Identification part: ST calling.

⚠ Be careful not to lose yourself only in the Results part of the paper and focus on the Material and Methods part, typically here you will have to read the Bowtie2 paper to explain how the alignment work.

Better to know

MLST, BWT, FM index

3 - StrainSeeker [Roosaare et al., 2017]

Designed for the identification of bacterial isolates (which also works very well on small mixes of strains in my experience).

- Indexing part: guide tree using k-mers specifiq to nodes/leafs.
- Query part: top-down search of k-mers reads in the guide tree.
- Identification/Quantification part: Tests on observed/expected k-mers ratio for each nodes.

⚠ You'll have to read GenomeTester4's paper to understand how StrainSeeker works (probably also the MEGA and MAFFT papers).

Better to know

Tree construction methods (neighbor-joining, UPGMA), multiples alignment technics, have good statistical knowledges

4 - KrakenUniq [Breitwieser et al., 2018]

Software based on Kraken but modify to allows for more accurate quantification of microbial species and strain-level identification.

- Indexing part: Same as Kraken (hash table).
- Query part: Same as Kraken + HyperLogLog algorithm for uniq k-mer counting.
- Identification/Quantification part: Same as Kraken (pruned classification tree and Root to Leaf (RTL) approach).

⚠ You'll have to read Kraken's paper to understand how Krakenuniq works.

Better to know

Hash table, Linear probing, Minimizer, LCA, HLL

5 - Kraken2 [Wood et al., 2019]

Kraken2 is one of the most widely used k-mers-based metagenomic classifier because of its speed and ease of use. However, it does not allow identification at low taxonomic levels such as strain.

- Indexing part: Compact hash table with use of minimizers and spaced seed.
- Query part: Same as Kraken.
- Identification/Quantification part: Similar to Kraken.

⚠ As for Krakenuniq, you'll have to read Kraken's paper to understand how Kraken2 works.

Better to know

Hash table, Linear probing, Minimizer, Spaced seeds, LCA

6 - Metamaps [Dilthey et al., 2019]

Designed for strain-level classification and read mapping in metagenomic datasets particularly optimized for long-reads sequencing data.

- Indexing part: Minimizer index construction from reference.
- Query part: Own fast approximate long-read mapping algorithm using MinHash sketch of the minimizer.
- Identification/Quantification part: EM algorithm using mapping qualities.

Better to know

MinHash technique, Minimizer, Bayesian inference, EM algorithm

7 - ganon [Piro et al., 2020]

Designed for taxonomic classification of metagenomics sequencing data.

- Indexing part: taxonomic clustering using TaxSBP then Interleaved Bloom Filters (IBF) creation.
- Query part: K-mer counting lemma (minimum number of matches between the read and the reference) + filters
- Identification/Quantification part: LCA

⚠ You'll have to read DREAM-Yara's papers and look how TaxSBP works to understand how ganon works.

Better to know

Bloom filter, LCA, have good statistical knowledges

8 - CCMetagen

Designed to produce accurate taxonomic classifications from metagenomic data.

- Indexing part: Hash table to store references (Part of KMA).
- Query part: use ConClave sorting scheme to perform alignment of reads on references (KMA).
- Identification/Quantification part: alignments filter using sequence similarity threshold to define the lowest taxonomic rank for each reads.

⚠ You'll have to read KMA's papers to understand how CCMetagen works.

Better to know

Hash table, basic alignment algorithms (NW, seed-and-extend...)

9 - ORI [Siekaniec et al., 2021]

Developed to enable strain-level bacterial identification from long nanopore sequencing data (also works with PacBio).

⚠ As the creator of ORI and having written the associated paper, I'm bound to be biased towards this one. It's a double-edged sword, because I'd be better able to help you if you didn't understand something, but I'd also inevitably see the slightest mistake.

- Indexing part: data structure based on a Bloom Filter (HowDeSBT).
- Query part: use of spaced k-mers.
- Identification/Quantification part: use of Answer Set Programming (ASP) a Declarative programming, Logical language, Implements SAT solving and constraint programming techniques.

⚠ You'll have to read HowDeSBT's paper to understand how ORI works.

Better to know

Bloom filter, Graph theory, spaced seeds, ASP (solver, constraint programming)

10 - StrainFLAIR [Silva et al., 2021]

Developed for strain-level analysis in metagenomic datasets, focusing on the identification and tracking of bacterial strains in complex microbial communities.

- Indexing part: Gene prediction+clustering, variation graph (vg) construction using this genes.
- Query part: mapping reads over the vg.
- Identification/Quantification part: Colored-path attribution and use of genome specific gene abundances.

⚠ StrainFLAIR uses a variety of tools whose operation you'll need to understand, so you'll need to read other papers such as the Prodigal paper use to gene prediction.

Better to know

Graph theory, basic clustering technics, LASSO and linear model

11 - StrainGE [van Dijk et al., 2022]

Designed to identify strains in a sample even with a low sequences coverage. It is separate in two pipeline StrainGST (Strain Genome Search) and StrainGR (Strain Genome Recovery).

- Indexing part: database of k-mer profiles from reference genomes.
- Query part: k-mers composition comparison.
- Identification/Quantification part: combination of three metrics based on common k-mers between sequence and reference.

⚠ You'll have to read BWA's paper to understand how StrainGR works.

Better to know

minhash technique, basic clustering technics, BWT, FM index

12 - MetaBIDx [Pham and Phan, 2024]

Designed to accurately identify species in complex microbiomes.

- Indexing part: Modified Bloom filter allowing to differentiate unique k-mers specific to a species.
- Query part: Query reads against the index.
- Identification/Quantification part: Cluster reads into groups of similar coverages.

⚠ As the paper is relatively imprecise on the method to be used, it will be useful to go directly to the github directory.

Better to know

Bloom filter, basic clustering methods (K-means here)

13 - Mibianto [Hirsch et al., 2024]

Online whole metagenome sequencing data analysis web server centered around quick compositional analysis.

⚠ As Mibianto is a web-server, the methods used are separated between the user interface and the server side, complicating the taxonomic profiling part. In addition, Mibianto uses a large number of other software/methods, which makes the study of the method more complex than for other proposed papers.

- Indexing part: No true indexing part here. Usage of the Genome Taxonomy Database (GTDB).
- Query part: Reads FracMinHash using Sourmash.
- Identification/Quantification part: approximate taxonomic abundance using sourmashconsumr.

⚠ You'll have to read Sourmach's paper witch is the core of the taxonomic profiling part from Mibianto.

Better to know

Client and server communication, minHash technics

Preliminary schedule

	Event	Time	
1	Kick-off meeting	Today (29.10.2024)	mandatory
2	Draft Submission	22.11.2024	mandatory
3	Progress Meeting	November/December	optional
4	Summary Submission	03.01.2025	mandatory
5	Block presentations	End of January	mandatory

Find the papers and other information here: https://cms.sic.saarland/big

The seminar provides 7 ECTS

Additional information I

Summary

- Main structure: title page, main text (3 to 4 pages of text, excluding title, references, figures, tables etc...), references → for example: 3 pages + one big figure about the main part of the method.
- It is recommended to write the summary using LaTeX to train your scientific writing.
- use the figure/table to explain in details the parts of the methods that are more complex to understand in writing only.

Presentation

- 35 minutes presentation.
- 10 minutes of questions.
- Find question to ask for each presentation.
- Sven Rahmann will be with me to evaluate your presentation.

Additional information II

- Prioritize understanding the methods over admiring the results. I will be more tolerant with a 2-page text that explains in detail how the method works rather than with a 4-page text that focuses only on the results of the paper.
- For the summary, try to use a size 12 font and don't try to save too much space with the margins.
- If possible go further and give the advantages but also the disadvantages of the method used.
- Don't rely on the paper's visuals, draw your own figures.
- Be careful of your speaking time for the presentation.
- Don't forget to provide references.

List of 13 proposed scientific papers (from the oldest to the newest)

- Centrifuge [Kim et al., 2016]
- MetaMLST [Zolfo et al., 2017]
- StrainSeeker [Roosaare et al., 2017]
- KrakenUniq [Breitwieser et al., 2018]
- Kraken2 [Wood et al., 2019]
- Metamaps [Dilthey et al., 2019]
- ganon [Piro et al., 2020]
- CCMetagen [Marcelino et al., 2020]
- ORI [Siekaniec et al., 2021]
- StrainFLAIR [Silva et al., 2021]
- StrainGE [van Dijk et al., 2022]
- MetaBIDx [Pham and Phan, 2024]
- Mibianto [Hirsch et al., 2024]

References I

Breitwieser, F. P., Baker, D. N., and Salzberg, S. L. (2018).
KrakenUniq: confident and fast metagenomics classification using unique k-mer counts.
Genome Biology, 19(1):198.

Dilthey, A. T., Jain, C., Koren, S., and Phillippy, A. M. (2019).
Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps.
Nature Communications, 10(1):3066.
Publisher: Nature Publishing Group.

Hirsch, P., Molano, L.-A., Engel, A., Zentgraf, J., Rahmann, S., Hannig, M., Müller, R., Kern, F., Keller, A., and Schmartz, G. (2024).
Mibianto: ultra-efficient online microbiome analysis through k-mer based metagenomics.
Nucleic Acids Research, 52(W1):W407–W414.

References II

Kim, D., Song, L., Breitwieser, F. P., and Salzberg, S. L. (2016).
Centrifuge: rapid and sensitive classification of metagenomic sequences.
Genome Research, 26(12):1721–1729.
Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Marcelino, V. R., Clausen, P. T. L. C., Buchmann, J. P., Wille, M., Iredell, J. R., Meyer, W., Lund, O., Sorrell, T. C., and Holmes, E. C. (2020).
CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data.
Genome Biology, 21(1):103.

References III

Pham, D.-T. and Phan, V. (2024).
MetaBIDx: a new computational approach to bacteria identification in microbiomes.
Microbiome Research Reports, 3(2):N/A–N/A.
Publisher: OAE Publishing Inc.

Piro, V. C., Dadi, T. H., Seiler, E., Reinert, K., and Renard, B. Y. (2020).
ganon: precise metagenomics classification against large and up-to-date sets of reference sequences.
Bioinformatics, 36(Supplement_1):i12–i20.

References IV

Roosaare, M., Vaher, M., Kaplinski, L., Möls, M., Andreson, R., Lepamets, M., Kõressaar, T., Naaber, P., Kõljalg, S., and Remm, M. (2017).
StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees.
PeerJ, 5:e3353.
Publisher: PeerJ Inc.

Siekaniec, G., Roux, E., Lemane, T., Guédon, E., and Nicolas, J. (2021).
Identification of isolated or mixed strains from long reads: a challenge met on Streptococcus thermophilus using a MinION sequencer.
Microbial Genomics, 7(11):000654.
Publisher: Microbiology Society,.

References V

Silva, K. D., Pons, N., Berland, M., Oñate, F. P., Almeida, M., and Peterlongo, P. (2021).
StrainFLAIR: strain-level profiling of metagenomic samples using variation graphs.
PeerJ, 9:e11884.
Publisher: PeerJ Inc.

van Dijk, L. R., Walker, B. J., Straub, T. J., Worby, C. J., Grote, A., Schreiber, H. L., Anyansi, C., Pickering, A. J., Hultgren, S. J., Manson, A. L., Abeel, T., and Earl, A. M. (2022).
StrainGE: a toolkit to track and characterize low-abundance strains in complex microbial communities.
Genome Biology, 23(1):74.

References VI

Wood, D. E., Lu, J., and Langmead, B. (2019).
Improved metagenomic analysis with Kraken 2.
Genome Biology, 20(1):257.

Zolfo, M., Tett, A., Jousson, O., Donati, C., and Segata, N. (2017).
MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples.
Nucleic Acids Research, 45(2):e7.