Department of Computer Science
Series of Publications A
Report A-2022-12

Probabilistic Methods for High-Resolution Metagenomics

Tommi Mäklin

Doctoral dissertation, to be presented for public examination with the permission of the Faculty of Science of the University of Helsinki in Auditorium CK112, Exactum building, Kumpula campus on the 28th of October 2022 at 12 o'clock.

University of Helsinki
Finland

Supervisors
Antti Honkela, University of Helsinki, Finland
Jukka Corander, University of Helsinki, Finland

Pre-examiners
Ashlee Earl, Broad Institute of MIT and Harvard, USA
Tommi Vatanen, University of Helsinki, Finland

Opponent
Leo Lahti, University of Turku, Finland

Custos
Antti Honkela, University of Helsinki, Finland

Contact information

Department of Computer Science
P.O. Box 68 (Pietari Kalmin katu 5)
FI-00014 University of Helsinki
Finland

Email address: info@cs.helsinki.fi
URL: http://cs.helsinki.fi/
Telephone: +358 2941 911

ISSN 1238-8645 (print)
ISSN 2814-4031 (online)
ISBN 978-951-51-8694-2 (paperback)
ISBN 978-951-51-8695-9 (PDF)
Helsinki 2022
Unigrafia

Probabilistic Methods for High-Resolution Metagenomics

Tommi Mäklin

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
tommi.maklin@helsinki.fi
https://maklin.fi/

PhD Thesis, Series of Publications A, Report A-2022-12
Helsinki, October 2022, 86+94 pages pages
ISSN 1238-8645 (print)
ISSN 2814-4031 (online)
ISBN 978-951-51-8694-2 (paperback)
ISBN 978-951-51-8695-9 (PDF)

Abstract

Metagenomics is the analysis of DNA sequencing data from samples obtained directly from the environment and containing several different organisms at once. Common tasks in metagenomics are taxonomic profiling, where the goal is to identify the organisms present in the sample and assign relative abundances to them, and taxonomic binning, where the sequencing data from the sample is divided into bins that correspond to some sensible taxonomic units. This thesis introduces methods for performing these two tasks at a high-resolution capable of distinguishing between lineages of bacterial species. The first of these methods is mSWEEP, which solves the profiling task by utilizing a collection of grouped bacterial reference sequences, pseudoalignment, and a probabilistic model. The second method, mGEMS, builds upon mSWEEP to solve the binning task using an assignment rule derived from the fundamentals of the probabilistic model used by mSWEEP. Both methods are accompanied by efficient implementations that utilize fast variational inference and pseudoalignment to fit the model in a reasonable time, rendering them applicable to large-scale datasets.

Both mSWEEP and mGEMS have been developed for application in either the traditional whole community metagenomics context, where the direct-from-environment samples are analysed, or in the plate sweep metagenomics context, where the sample has been plated once on a selective medium. While the latter is not metagenomics in the traditional sense, this thesis advocates for its use when high depth sequencing data is required from some species and the other organisms are not of interest. Regardless of the type of metagenomics data used, the ultimate goal of both mSWEEP and mGEMS is to enable performing standard genomic epidemiological analyses directly from data containing several strains of the same bacteria, skipping the typically used isolation steps required to separate them. Due to the implied cost-savings from reducing the number of cultures that need to be performed as well as the better capture of variation in the samples through using metagenomics data, mSWEEP and mGEMS enable performing entirely novel types of analyses in the field of genomic epidemiology.

Computing Reviews (2012) Categories and Subject Descriptors:

Mathematics of computing → Probability and statistics → Statistical computing
Computing applications → Biosciences

General Terms:
Algorithms, Experimentation

Additional Key Words and Phrases:
genomic epidemiology, plate sweeps, probabilistic modeling, pathogen surveillance, taxonomic profiling, taxonomic binning, metagenomics

Acknowledgements

First I would like to express my sincere thanks to my supervisors Antti Honkela and Jukka Corander. Without your guidance and occasional prodding none of the experiences I have been fortunate enough to collect during this journey would not have been possible. You have been extraordinarily helpful when needed and necessarily stern when required.

Like many great things in life, the research in this thesis is the result of collective work. Accordingly, I want to thank all of the incredible individuals who made the bigger picture manifest by collaborating and coauthoring with me. Without all of you, both the results and the road there would have been a much rougher ride.

In addition to my collaborators, I wish to recognize the impact of my colleagues in Antti's group who I have had the pleasure of working along for the past six years and who have provided excellent company for extensive lunch & coffee breaks and certain extracurricular activities. Although our academically measurable contributions were limited, the impact of a supportive environment is heartfelt and I owe you my gratitude. I also want to thank all the other brilliant people who I've met during my doctoral studies and other activities at the University of Helsinki, and who I've had the pleasure of chatting, advocating, and simply existing with. The University of Helsinki, and the Academy of Finland, deserve my additional gratitude for funding my studies through various projects.

Finally, I wish to thank all my friends who have been there for me during these past years, and my family for their unwavering support.

Helsinki,
October 2022
Tommi Mäklin

Contents

List of original publications

Publication I — High-resolution sweep metagenomics using fast probabilistic inference

By Tommi Mäklin, Teemu Kallonen, Sophia David, Christine J Boinett, Ben Pascoe, Guillaume Méric, David M Aanensen, Edward J Feil, Stephen Baker, Julian Parkhill, Samuel K Sheppard, Jukka Corander, and Antti Honkela. Published in *Wellcome Open Research* (2021), 5:14.
doi: 10.12688/wellcomeopenres.15639.2.

Publication II — Bacterial genomic epidemiology with mixed samples

By Tommi Mäklin, Teemu Kallonen, Jarno Alanko, Ørjan Samuelsen, Kristin Hegstad, Veli Mäkinen, Jukka Corander, Eva Heinz, and Antti Honkela. Published in *Microbial Genomics* (2021) 7:11.
doi: 10.1099/mgen.0.000691.

Publication III — Strong pathogen competition in neonatal gut colonization

By Tommi Mäklin, Harry A Thorpe, Anna K Pöntinen, Rebecca A Gladstone, Yan Shao, Maiju Pesonen, Alan McNally, Pål J Johnsen, Ørjan Samuelsen, Trevor D Lawley, Antti Honkela, and Jukka Corander. Submitted; preprint available from *bioRxiv* (2022).
doi: 10.1101/2022.06.19.496579.

Chapter 1

Introduction

Public health research focusing on bacterial pathogens has been transformed by analysis of the contents of bacterial genomes obtained by whole-genome sequencing (WGS) since 2010 [1]. In this time frame, the price of sequencing has decreased tremendously [2, 3], enabling adoption of sequencing as a standard tool in the infectious disease, evolutionary, and genomic epidemiology toolkits [4–6]. Many of the standard analyses in these fields require data from pure bacterial cultures, created by isolating a bacterium from an initial mixed culture, which often contains several distinct bacteria and even other micro-organisms. Isolating all of these presents substantial economical barriers to more widespread adoption of WGS as a routine tool since the cost and turnaround time of the library preparation and DNA extraction steps, performed once per each isolated organism, are approaching the price of sequencing itself [7].

Whole community metagenomics, where DNA is extracted and sequenced directly from the original environmental sample, presents a potential cost-effective alternative to the isolate sequencing approach. Contrary to isolate sequencing, whole community metagenomics requires only a single library preparation and DNA extraction step and no cultivation steps since the sample is sequenced directly. However, direct sequencing may require significantly higher sequencing depths due to presence of host DNA [8, 9] and contamination [9, 10]. In addition, low biomass samples are challenging for metagenomics to use reproducibly. Due to these factors, whole community metagenomics is difficult to apply when only a

subset of the diversity is of interest but the planned analyses require high sequencing depths, which is typical in genomic epidemiological studies.

Genomic epidemiology is generally speaking the study of the spread of bacterial pathogens generally speaking based on WGS data. Sequencing the genomes of bacterial pathogens during an outbreak allows for comparing accumulated mutations in their genomes [4], elucidating their short-term evolutionary history and enabling case linking when combined with appropriate metadata [5, 11]. Similarly, long-term routine surveillance aids in hastening the detection of outbreaks [12, 13], identifying potential high-risk clones [14], or reservoirs for antimicrobial resistance [15, 16]. Many of these analyses require assembling the genomes of the bacteria from the sequencing reads which has led to dominance of the isolate sequencing approach and a lack of studies attempting to solve the epidemiological problems with metagenomics.

When choosing between whole community metagenomics and isolate sequencing, a middle-ground can be found plate sweep metagenomics [17] — sometimes also called limited-diversity metagenomics [18]. In this approach the initial culture from a sample is swept and DNA extracted from it is sequenced en masse rather than preparing several isolates from it. Since selective culture media are available for most clinically relevant bacteria [19], plate sweep metagenomics simultaneously both reduces the number of library preparation and DNA extraction steps by using only a single culture, and solves the host DNA overabundance and sequencing depth issues in whole community metagenomics through selective media enrichment. Incorporating an enrichment step has also been found to increase the sensitivity to low-abundance organisms that might be missed in direct sequencing [20].

While both plate sweep metagenomics and whole community metagenomics have technically been possible for many years with the latter appearing around 2004 [21, 22], the development of computational methods has largely focused on analysing sequencing reads from a single organism at a time. Although many methods for analysis of metagenomic data at the level of identifying strains (in this thesis a strain is the biological organism corresponding to a single cell colony) or lineages (a collection of strains that descend from the same strain and have maintained similar genetic sequences) have been developed [23], these do not typically perform

well when applied to data containing multiple lineages of the same species at once [24], which will be referred to as lineage-level variation further in the thesis. Methods diverging from the traditional relative abundance estimation (taxonomic profiling) [25] or metagenomic sequence read demixing (taxonomic binning) [26] context do successfully tackle lineage-level variation but do not easily translate to replacing assembly-based analyses such as SNP calling or phylogenetic inference.

This thesis presents two computational methods that enable taxonomic profiling and taxonomic binning from either whole community metagenomic or plate sweep metagenomic short-read sequencing data. While the plate sweep metagenomics approach was the focus of both methods during their initial publication, further research has shown that they also perform reliably when applied to whole community metagenomics data. Using either of the two approaches to reduce the costs associated with data collection, the methods presented here enable performing routine genomic epidemiological analyses when significant lineage-level variation is present in the collected sequencing data.

The first of the two methods, called mSWEEP, consists of a probabilistic model for estimating the relative abundances of lineages of a bacterial species in a set of sequencing reads [17]. mSWEEP leverages pseudoalignment [27] of the reads against a set of reference sequences that have been grouped together into lineages and outputs estimates of the lineage-level abundances. The second method, mGEMS, processes the output from mSWEEP to construct an assignment rule for assigning each read to one or more bins corresponding to a reference lineage [28]. Both methods explicitly account for the fundamental characteristic of sequencing data containing multiple lineages of the same species where each read can, and often does belong to several lineages of the same species at the same time. The combination of mSWEEP and mGEMS enables effective computational quantification of metagenomic data at a high resolution within the species, and enables downstream processing of mixed samples with results often comparable to using isolate data.

Even though both mSWEEP and mGEMS were originally designed with applications in plate sweep metagenomics in mind, Publication III [29] demonstrates applicability of both methods to whole community metagenomics data. The data analysed in Publication III was collected

from a cohort of UK neonates [30] and the samples were submitted for whole community metagenomics sequencing. Results from this data show strong competition between bacterial species and strains during the initial colonization of the newborn gut microbiome. More importantly from a methods perspective, this analysis shows that mSWEEP and mGEMS provide (so-far) completely unprecedented levels of resolution in analysis of metagenomic sequencing data.

Together Publications I-III represent foundational methodological steps in both opening up high-resolution exploration of bacterial diversity as well as making such analyses more accessible to resource-constrained laboratories.

1.1 Three approaches to sequencing bacterial DNA

Preparing bacterial DNA for sequencing is often done after a culture step that enriches the number of bacterial cells from a target group of microorganisms. Culturing is performed by plating a sample and inoculating it for a period of time that allows the bacteria to multiply [31]. After inoculation, visible colonies may be isolated and further propagated on their own plates [31], or the entire plate may be prepared for DNA extraction to produce plate sweep metagenomic data. Alternatively, in whole community metagenomics, the whole culture procedure is skipped, and DNA is extracted directly from the sample with the extract procedure depending on the sample type [32]. When it comes to the end-result — the sequencing reads — all three approaches have their own characteristics that affect the available downstream analyses.

Whole community metagenomics, where all or most of the DNA in a sample is extracted (Figure 1.1a), has emerged as a tool for analysing the full breadth of variety in various microbiomes [20, 30, 33–35]. Exploring this diversity comes at a price, however, since the produced sequencing reads are split across the numerous organisms possibly present, resulting in a need to sequence the sample more deeply to capture the less abundant organisms [20, 36, 37]. Combined with other issues related to host DNA abundance [20, 38, 39], the shortcomings of whole community metagenomics have so-far hindered its adoption in genomic epidemiology.

Plate sweep metagenomics proposes a middle-ground between the direct sequencing of whole community metagenomics and isolate studies by incorporating a single culture step [17]. In this approach, the sample is cultured on an appropriate selective medium and the entire complexity of the plate is subjected to DNA extraction and sequenced after a suitable inoculation period (Figure 1.1b). The inclusion of a culturing step allows for generating large numbers of sequencing reads from the bacteria that thrive in the chosen medium, circumventing both the sequencing depth and host DNA issues in whole community metagenomics while improving the sensitivity to bacteria found in low abundance in the original sample [20, 40, 41]. Furthermore, focusing the sequencing efforts on the relevant bacteria enables application of bioinformatics tools that require a high sequencing depth provided that the reads from different organisms can be computationally separated. Developing a tool to solve the aforementioned deconvolution problem is one of the key contributions of this thesis.

In the third approach, whole-genome sequencing of isolates (Figure 1.1c), visible colonies from the initial culture are picked and transferred to new plates. After letting the transferred colonies grow, the resulting culture will consist only of the descendants of the original colony, allowing for massive numbers of sequencing reads to be generated from the isolated organisms. Since visible colonies on the initial culture are typically assumed to contain clones of the same organism, this approach effectively gets rid of most of the variation found in both the sample and the initial culture. While the whole-genome sequencing approach is excellent for generating high-coverage and high-quality data from a single bacterial strain, in practice the number of colonies that can be isolated is often constrained by laboratory resources and nearly always restricted to rapidly growing colony-forming phenotypes.

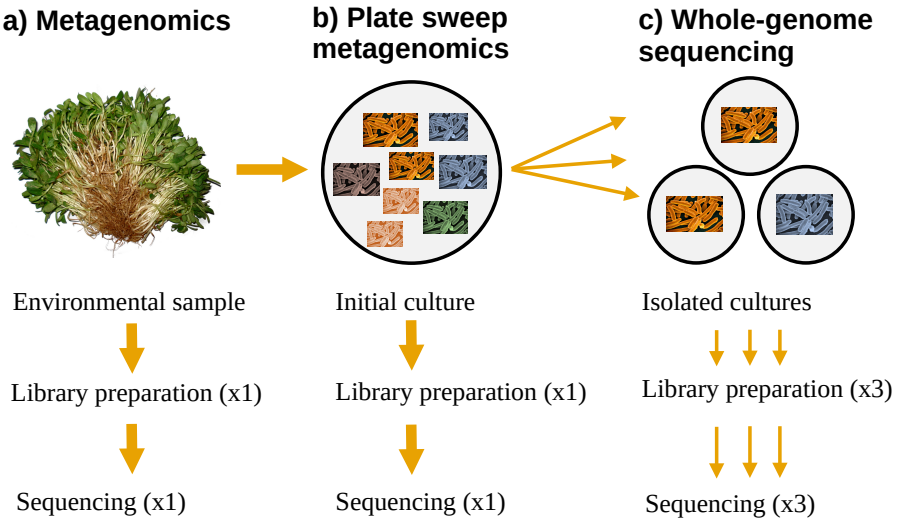

Figure 1.1: Different approaches to sequencing bacterial DNA. Panel **a)** depicts the whole community metagenomics approach, where sequence data is produced directly from the sample. Panel **b)** depicts the plate sweep metagenomics approach, where the sample is plated on a selective medium and DNA extracted and sequenced from the whole plate after an inoculation period. Panel **c)** depicts the whole-genome sequencing approach, where a subset of visible colonies on the inoculated culture is extracted, and DNA from them prepared for sequencing.

In genomic epidemiological analyses, the whole-genome sequencing approach has been dominant due to its strengths in producing highly accurate data capable of SNP calling and differentiating between organisms [44]. Producing the same results using whole community or plate sweep metagenomics has some obvious benefits in both increasing the number of samples that can be processed as well as in capturing more of the diversity in the samples, but the existing metagenome-analysis tools have not been able to reach the required level of resolution [24, 28, 45]. Here, the issue is tackled through methodological advances that open up more widespread use of metagenomic sequencing data in genomic epidemiology.

Figure 1.1 source: Adapted from [42] and [43].

1.2 Analysing metagenomic sequence data

Metagenomic sequencing data analysis presents several challenges to bioinformaticians). Firstly, the increased diversity of species requires much larger computational resources to analyse [46]. Secondly, the possible presence of lineage-level variation complicates analyses that attempt to separate the reads to distinct taxonomic units because the differences between strains within a species can be minimal [47]. Subsequently, the bulk of method development has focused on operating with the assumption that only a single strain from each species is present in the same sample [23]. This section will briefly cover some of the previous approaches and describe how mSWEEP and mGEMS fit into the metagenomics toolkit.

One of the more commonly used tools for analysing whole community metagenomics data are metagenome assemblers. Similarly to genome assemblers, metagenome assemblers aim to produce a set of contigs (sets of overlapping sequencing reads) that correspond overlapping sequencing reads. Since reads from metagenomic sequencing contain several organisms, metagenome assemblers are often paired with metagenome binners that attempt to assign the metagenome-assembled contigs to bins that correspond to a taxonomic unit. These units are typically assumed distinct enough that they are not mistaken for sequencing error or minor genetic variation. When the assumption is fulfilled, metagenome assemblers produce contigs that are adequately accurate for several types of analyses, and have subsequently been adopted among the standard tools in a variety of microbiome studies [34, 48, 49].

Metagenome binners are closely related to another type of analysis, where the aim is to assign (relative) abundances to the taxonomic units that were identified in the sequencing reads, called taxonomic profiling. These two approaches sometimes go hand-in-hand since the abundance of a taxonomic unit can naively be defined as the number of reads that align to contigs which have been assigned to the unit. When the units are closely related species or strains, more sophisticated methods are necessary since both the reads and short contigs may plausibly belong to several taxa, which has led to the development of dedicated taxonomic profilers that do not attempt to bin the contigs or the sequencing reads [17, 25, 26, 50].

Recently, a third category of methods for tracking the presence/absence of a specific strain across several samples has emerged [25, 26, 51]. These methods aim to infer similarity and shared strains and provide an attractive tool for transmission analysis when the genomes of the individual strains are not required or cannot be assembled due to low sequencing depth. Ideally, the tools for tracking strains would be combined with those for extracting the contig or read bins, together enabling both a wide analysis covering all samples and strains and a focused analysis of the strains that are abundant enough to assemble their genomes.

All of the above can be further divided into reference-based methods that leverage reference data in their analysis, and reference-free methods that perform the analysis solely based on the sequencing reads. While reference-free methods are able to handle data containing previously unknown bacteria that do not have any available genome assemblies, reference-based methods provide an easily interpretable context for the results and typically reach a higher resolution [52, 53]. If detailed quantification is only required for some subset of organisms in the sample, reference-based methods frequently also provide means to filter out the reads belonging to the uninteresting organisms.

The lineage-level methods from this thesis take the reference-based approach and specifically leverage bespoke reference collections. Tailoring the collections to fit the presumed contents of the samples allows both mSWEEP and mGEMS to perform at a resolution that is mainly limited by the quality and variety of the available assemblies. Contrary to many existing reference-based methods, mSWEEP and mGEMS do not attempt identification of the individual reference sequences, but rather incorporate a clustering of the reference assemblies to biologically interpretable and phylogenetically reflected lineages. Since differences between lineages are generally more pronounced, the identification task becomes significantly easier [54] and extends to handling cases where the sequencing reads originate from a previously unknown sequence which nevertheless belongs to a known reference lineage. Trading the ability to imprecisely identify the exact sequence to precisely identifying the lineage is especially useful in clinical settings, where the diversity of potential pathogenic bacteria has been thoroughly studied using whole-genome sequencing and the clinically relevant lineages are often well-known.

1.3 Metagenomics data in genomic epidemiology

Aside metagenomics tools, the second significant aspect of this thesis has to do with their application in genomic epidemiology, where an important goal is to trace the transmission of pathogenic bacteria using genomic sequence data. Genome-informed analyses have in recent years greatly expanded the ability of researchers to investigate outbreaks, identify epidemiologically relevant genetic elements, and detect emerging public health concerns [4–6, 55]. Due to the high level of accuracy required, such analyses have been performed using isolate WGS data, which has a relatively high economical cost and slow turnaround time [7], rendering the research more reactive in nature. One of the goals of mSWEEP and mGEMS is to enable partially replacing the use of isolate data with metagenomics data, decreasing both the cost and turnaround time of the existing genomic epidemiology pipelines.

In addition to improving both cost- and time-effectiveness, incorporating some kind of metagenomics data into genomic epidemiology presents some obvious advantages in increasing the sensitivity to genetic diversity that might be obscured by the use of isolate data in routine surveillance. As an example, a recent study into within-host diversity of the common respiratory pathogen *Streptococcus pneumoniae* utilized an analogue of plate sweep metagenomics and found low-frequency co-colonization by lineages corresponding to epidemic serotypes alongside lineages of known carriage serotypes [40]. This finding helped explain the previously unknown source of the epidemic serotypes in outbreaks of disease, which could not be fully explained by isolate sequencing data. Since naturally occurring variation is common in many species of clinical interest [40, 56–59] similar findings in other fields are likely with more widespread use of whole community and plate sweep metagenomics data.

Another related aspect in favour of using more metagenomics-oriented approaches arises from simple practicality: sequencing several organisms at once is simply easier than performing the several steps required to isolate an organism for DNA extraction. Direct sequencing of the samples combined with nearly equally accurate analyses should, in principle, make implementing routine surveillance significantly more accessible to locations and laboratories lacking in funding and resources. This in turn

combined with data sharing practices across borders has the potential to vastly increase the capabilities of proactive surveillance. Furthermore, sequencing the whole sample and publicly archiving the reads has the benefit of preserving DNA from the full variety of organisms in the sample and making it available for future analyses with different goals from the original studies.

In conclusion, the field of genomic epidemiology that was established with the emergence of rapid and scalable WGS sequencing in the early 2010's can be seen as entering a transformative period with both data generation and more powerful computation tools becoming increasingly available and accessible. The development of methods such as mSWEEP and mGEMS will facilitate a further speed up of this transformation and enable entirely novel types of analyses and discoveries through the inclusion of metagenomic data.

1.4 Contributions

This thesis comprises three publications covering both mSWEEP [17] (Publication I), mGEMS [28] (Publication II), and a third article (Publication III) demonstrating their application to whole-genome shotgun metagenomic sequencing data [29]. Publications I-II are accompanied by software implementations [60, 61]. Publication III is more applied in nature, exploring in more detail the types of analyses enabled by Publications I-II.

Publication I — High-resolution sweep metagenomics using fast probabilistic inference

By Tommi Mäklin, Teemu Kallonen, Sophia David, Christine J Boinett, Ben Pascoe, Guillaume Méric, David M Aanensen, Edward J Feil, Stephen Baker, Julian Parkhill, Samuel K Sheppard, Jukka Corander, and Antti Honkela. Published in *Wellcome Open Research* (2021), 5:14, doi: 10.12688/wellcomeopenres.15639.2.

Publication I [17], presented and benchmarked the mSWEEP method for taxonomic profiling of sequencing data containing multiple strains from the same bacterial species. The author contributed to conceptualization

of the study, formal analysis and investigation of the data, developing the methodology and software implementations, validation and visualization of the results, and writing and editing both the original draft and the revised manuscript.

Software implementation of the ideas presented in Publication I is available from GitHub at https://github.com/PROBIC/mSWEEP (latest version). The latest version at the time of writing is archived and available in Zenodo [60].

Publication II — Bacterial genomic epidemiology with mixed samples

By Tommi Mäklin, Teemu Kallonen, Jarno Alanko, Ørjan Samuelsen, Kristin Hegstad, Veli Mäkinen, Jukka Corander, Eva Heinz, and Antti Honkela. Published in *Microbial Genomics* (2021) 7:11, doi: 10.1099/mgen.0.000691.

Publication II [28] continued to build upon mSWEEP by developing an algorithm for binning sequencing reads at the lineage-level of mSWEEP analyses. This approach, and the accompanying software implementation, are both called mGEMS. The author contributed to Publication II by taking part in conceiving the study, developing the mGEMS pipeline, in designing both the synthetic and the *in vitro* experiments, developing the mGEMS assignment algorithm, running the experiments, creating the visualizations, interpreting the results, and in writing and editing the main manuscript and the final published version

Software implementation of the ideas presented in Publication II is available from GitHub at https://github.com/PROBIC/mGEMS (latest version). The latest version at the time of writing is archived and available in Zenodo [61].

Publication III — Strong pathogen competition in neonatal gut colonization

By Tommi Mäklin, Harry A Thorpe, Anna K Pöntinen, Rebecca A Gladstone, Yan Shao, Maiju Pesonen, Alan McNally, Pål J Johnsen, Ørjan Samuelsen, Trevor D Lawley, Antti Honkela, and Jukka Corander. Submitted; preprint available from *bioRxiv* (2022), doi: 10.1101/2022.06.19.496579.

Publication III [29] provides an example of applying mSWEEP and mGEMS to whole community metagenomics sequencing data and explores the dynamics of pathogen competition and colonization in the gut microbiome of babies in their first three weeks of life. The author contributed to Publication III in running the mSWEEP/mGEMS pipeline on all data used in Publication III, updating the reference databases for the investigated species, performing the analysis of the mSWEEP/mGEMS results for the samples containing *E. coli*, and aiding the co-authors in analysing the other species. Additional contributions included creating the visualisations, interpreting the results, and naturally writing the publication.

1.5 Structure

The rest of the thesis is structured into three chapters that describe the contents of Publications I-III and how they contribute to the topics presented in the introduction chapter. The first of the three chapters (Chapter 2) describes the basic ideas behind the mSWEEP and mGEMS methods and provides historical context for the parts of the methods that have their origins within analysis of RNA sequencing data. The second chapter (Chapter 3) describes the experimental results from Publications I-III in more detail, focusing more on the applied part rather than the theoretical foundations. The third chapter (Chapter 4) is more speculative in nature, covering both the demonstrated applications from Publications I-III as well as exploring potential future avenues for use of the developed methods. The three chapters are followed by a concluding chapter (Chapter 5) which in the physical copy of the thesis is further followed by reprints of the three included original publications.

Chapter 2

Mixture modeling of sequence data

Mixture models are a family of probabilistic models which model sampling from an overall population as a mixture of sampling from several distinct subpopulations. Each subpopulation is typically assumed to have its own distribution, which can be from the same or a different distribution family, and mixing parameters that determine the percentages of data each subpopulation contributes to the overall population.

In sequencing data analysis, a key area of application for mixture models has been in RNA-Seq, where identifying the expression levels (relative contributions) of protein isoforms in some set of RNA sequencing reads is one of the main problems [62, 63]. Specifically, mixture models are useful in cases where the sequencing reads do not uniquely identify the isoform but could plausibly be the product of several genes. This is in contrast to the microarray technology that preceded RNA-Seq, where the technology itself allows for unique identification of the expressed isoform and applications of probabilistic models focused more on obtaining uncertainty estimates [64, 65].

The ability of mixture models to differentiate between expression of isoforms with similar nucleotide contents makes them ideal for analysis of short-read sequencing data from bacterial strains. Since the strains within a species typically share a large percentage of their genome (although the exact values vary greatly by species [66, 67]), sequencing reads from one will match with a large number of genomes from the same species. In-

deed, the models from RNA-Seq have been adapted almost directly to identify bacterial strains from sequencing data [54] with great effect but not without some caveats related to more general applicability across different bacterial species. The work in this thesis extends the previous work in the field [54] by introducing a more general formulation of the model that generalizes well to arbitrary bacterial species and allows for assigning sequencing reads to the bacterial strains in addition to identifying their relative contributions in a set of sequencing reads.

2.1 mSWEEP and mGEMS

The mSWEEP method is a tool for estimating the relative abundances of lineages of bacterial species in a set of sequencing reads. The method consists of two parts: preparing and clustering a reference genome assembly collection, and estimating the relative abundances using pseudoalignment [27] and probabilistic modelling. In the preparation part, a reference collection consisting of genome assemblies for some predefined set of bacterial species is constructed and prepared for analysis by clustering the assemblies into biologically sensible lineages. In the analysis part, short-read sequencing data are pseudoaligned against the reference collection and the alignments are used alongside the lineage clustering as the input to the mSWEEP probabilistic model. With results of the pseudoalignment mSWEEP estimates the relative abundances of each lineage in the reference collection using a mixture model and variational inference. The outputs from mSWEEP are the relative abundances of the lineages defined in the reference collection and a probability matrix describing the fit of each sequencing read to each reference lineage.

Accompanying mSWEEP is the mGEMS pipeline, which is a method for assigning each read in a sample to some (or none if the read does not pseudoalign against any reference sequences) of the reference lineages. mGEMS utilizes the relative abundance estimates and the probability matrix from mSWEEP to assign the lineage membership of each read. Importantly, mGEMS allows for multi-lineage membership, since many reads can plausibly originate from several strains within the same species.

Although both mSWEEP and mGEMS are novel methods that have been published in Publications I-II, the roots of mSWEEP especially lie

in the mixture modelling context from RNA-Seq and the subsequent BIB method [54] . These roots will be examined in more detail in the next section, which explains how they relate to the approach used in mSWEEP for bacterial data. The differences between reads from bacteria and RNA-Seq necessitate some changes to the probabilistic models employed in RNA-Seq, which eventually produces the model used in mSWEEP.

2.1.1 Relationship to RNA-Seq methods

In RNA-Seq, mixture models were proposed as a solution to the isoform expression level estimation problem around 2010 with several methods appearing around the same time [68–71]. In these methods, the model is defined through latent indicator variables that denote the source isoform for each sequencing read and the parameter of interest (the expression levels) are the proportions of reads assigned to each indicator variable. The proportions are inferred using either a likelihood function based on assessing the fit of the read to the reference isoforms based on sequence alignment [68, 69], or by assuming a Poisson distribution on the numbers of reads that are compatible with each reference [70, 71]. Estimating the parameters themselves was performed using a variety of algorithms ranging from Markov chain Monte Carlo (MCMC) sampling [68] and importance sampling [70] to maximum likelihood estimation [71] and the EM algorithm [69].

From the perspective of this thesis, a significant development of the methods appeared in 2012 with the introduction of BitSeq [72]. BitSeq extended the previous models by being the first of the methods to perform Bayesian inference on the relative isoform expression levels and derived update equations for a collapsed Gibbs sampler to implement MCMC sampling over the posterior distribution defined by the model. A further development of BitSeq appeared in 2015 with the introduction of BitSeqVB [73], where the sampling approach was supplemented by a collapsed variational Bayes approach that is significantly faster in fitting the model than the collapsed Gibbs sampler in BitSeq.

2.1.2 Applying RNA-Seq methods to data from bacteria

Solving the RNA-Seq isoform expression estimation problem had some unforeseen consequences in that the mixture models used can be almost directly applied to estimating the relative abundances of different strains of bacteria in a set of DNA sequencing reads. In 2016, the alikeness of the two problems was noted by the BIB method [54] which solved the analogous bacterial strain relative abundance problem leveraging the work from BitSeq and BitSeqVB. In BIB, the reference isoforms are simply replaced by the genomes of reference bacterial strains, turning the expression level estimates into relative abundances of these strains. However, due to the fact that the strains within a bacterial species are more alike than the isoforms BitSeq was developed to handle, BIB incorporated a step where the reference sequences were made more differentiable by clustering them into lineages. Each lineage was represented in the reference collection by a reference sequence randomly sampled from all those belonging to the lineage, and the representative sequences were further trimmed down to contain only the core genome of the species. The core genome refers to genomic sequences that are shared by all, or nearly all, members of a species. In some cases it may be preferable to define the core genome for subunits within the species, such as lineages, and particularly if the species definition is not based on or conforming to the genetic sequences.

The relative abundance estimation method from this thesis, mSWEEP, builds upon the work in BIB by using both the core genome and the accessory genome (accessory meaning the genome contents that are not contained in the core genome) of the reference sequence assemblies, and by removing the need to select a representative sequence from each lineage. Instead of selecting a representative sequence, mSWEEP uses all available assemblies from each lineage as the reference sequences, which gets around the problem of having to define an adequate sequence to represent the whole lineage. Furthermore, using all available sequences from each lineage provides better coverage of the variation in the now-included accessory genomes and allows applying the method to species that do not have as stable core genomes as *Staphylococcus aureus*, which was used as one of the example organisms in BIB [54].

In order to make the alignment against a much larger reference sequence collection feasible, mSWEEP additionally replaces the use of the

location-based alignment in BIB with pseudoalignment [27] which reports only a 0 or 1 depending on whether the read aligns somewhere (1) within a reference sequence or not at all (0), and massively speeds up the alignment part. This also allows for simplifying the likelihood function used in the mixture model by using just the pseudoalignment count within each lineage as the observations. In this sense, the mixture model used in mSWEEP can be seen as a descendant of both Poisson RNA-Seq models [70, 71], which used the alignment counts, and the models leveraging location-based information about the alignments [68, 69, 72] through the relation to BitSeq through BIB. Pseudoalignment-based identification of the relative abundances of some reference sequences is also implemented in the metakallisto [74] method but the inclusion of the probabilistic model from mSWEEP is necessary for high-resolution accuracy as demonstrated in Publication I.

2.1.3 Differences between RNA-Seq and bacterial data

Sequencing data and reference sequences from bacteria have some unique characteristics that distinguish them from data originating from humans or other more complex organisms. Mainly, the generation time for bacterial organisms is much shorter, measured in hours or even minutes depending on the environmental conditions (for example in the lab or in the wild) and species [75]. This has implications for analyses that incorporate the use of reference data from previously sequenced organisms. First, major changes in the genomic contents happen within human-observable timeframes and are reflected in sequence data obtained from what is assumed to be the same strain, although the accumulation rate is highly variable [76]. Secondly, bacterial genomes can undergo major horizontal gene transfer events even across large evolutionary distances, resulting in major genomic differences [77]. Together these factors imply that reference sequences for any set of bacteria are almost certainly at least somewhat different than what would be obtained from sequencing descendants of the organisms corresponding to the original reference sequence.

As noticed in the BIB method, the problems introduced by quicker evolution of the bacterial organisms can be solved by replacing the individual reference sequences with lineages within a bacterial species as the unit for relative abundance estimation [54]. When estimation is performed

for suitably clustered sequences, the problem becomes significantly easier since the short-term genetic variation, at least presumably, is more contained within the lineages, provided that they are biologically sensible. Since mSWEEP allows representing the lineages through several reference sequences, they become easier to distinguish because the differences between lineages are larger than the differences between strains within the same lineage which could potentially coexist in a sample. However, selecting the clustering algorithm for identifying the lineages needs to be performed carefully, since the estimation will be reliant on the signals that are contained within the lineages.

2.2 Significance of reference databases

In addition to the lineage definitions, the reference collection of some available genome assemblies for bacterial species of interest, or several, lies at the very core of mSWEEP. Since the method estimates the relative abundances of the lineages based on information provided by pseudoalignment of the reads against the reference sequences, the accuracy of the results is naturally constrained by the quality of the reference collection. The included sequences can be tailored to the problem at hand since mSWEEP does not place constraints on the kind of assemblies that are used. This bespoke approach to reference building is particularly useful when isolate sequencing data is available from the same or closely related organisms assumed to be present in the analysed sequencing reads.

Due to the disadvantages of requiring significant user effort in constructing the reference collection, many metagenomics methods rely on prebuilt references covering multiple species that have a high availability of sequences assemblies. For mSWEEP, supplying similar prebuilt references for a wide variety of species is not currently feasible due to the computational requirements of the pseudoalignment step, hence opting to use study or species-specific collections instead. Nevertheless, Publications I-III do include the databases that were used as parts of them, and allow for their reuse in future analyses. However, extending them with further isolate data is highly encouraged.

Another significant step in building the reference collection is deciding on the desired level of detail in the lineage definitions. While the fit of

the reference sequences to the sequencing reads ultimately determines the relative abundance estimates, tweaking the depth of the lineage definitions can enable identification in cases where the reference sequences are not exactly from a comparable source to the sample reads. Publications I-III employ several different approaches to the lineage definitions, making use of multilocus sequence typing (MLST) [78], and several clustering algorithms [79–81].

2.2.1 Clustering bacterial sequences

Various methods for clustering bacterial genomes have been developed. One of the most commonly used of these is MLST, where sequence clusters are defined based on variation observed at housekeeping loci [78], the combinations of which correspond to a unique sequence type. For many biological applications, the sequence types defined by MLST correspond to taxonomic units that have observable differences in phenotypes such as antimicrobial resistance [82, 83]. This has subsequently led to widespread adoption of the method among microbiologists. The downside of MLST is that it only offers a limited resolution by considering a small fraction of the variation present in a genome.

The PopPUNK method [79] provides an alternative to MLST that uses nucleotide distances and a Gaiussian mixture model or DBSCAN [84] to define the lineages. In practice, the lineages that PopPUNK identifies typically correspond to clonal complexes [79] which are sequence clusters containing the sequences assigned to a central multilocus sequence type (ST) and closely related single or double locus variants of the central ST. Main advantage of using the clonal complex analogues provided by PopPUNK is the ability to assign arbitrary reference sequences to lineages while mostly conforming to the MLST complexes. Additionally, using PopPUNK allows for including the accessory genome in defining the lineages if desired, making PopPUNK an ideal choice for defining the reference lineages for mSWEEP.

2.2.2 Sequence alignment

The reference collection is used in mSWEEP as the target for pseudoalignment. Contrary to the location-based alignment method employed in BIB

[54], pseudoalignment only reports a 0 when the read being pseudoaligned does not align anywhere within a reference sequence, and 1 in the when the read does align somewhere. In the original mSWEEP publication, the kallisto method [27], which introduced the pseudoalignment concept, was used to pseudoalign the reads, but Publication II introduced a more scalable method, called Themisto, that replaces kallisto in the pipeline. In addition to its scalability, Themisto also provides an exact version of the kallisto pseudoalignment algorithm [28].

Pseudoaligning the reads has the advantage of being much quicker to compute, enabling more extensive reference collections to be employed by mSWEEP. Although the disadvantage in information loss from binarizing the alignments does require some adjustments of the likelihood function in mSWEEP when compared to BIB, the added reference coverage more than makes up for any potential losses in accuracy. The next section will cover this mixture model formulation and the changes introduced by mSWEEP in more detail, as well as the theory behind the mGEMS algorithm for assigning the sequencing reads themselves to bins corresponding to the reference lineages.

2.3 A probabilistic model for sequences from mixed sources

The probabilistic model used by mSWEEP is an extension of the mixture model for grouped reference sequences used by BIB. Compared to BIB, which requires selecting a representative sequence for each reference lineage, the mSWEEP model allows including an arbitrary number of sequences to represent the variation in each lineage. In addition, to improve scalability aligning sequencing reads against the expanded reference collection, mSWEEP replaces the location-based alignment from BIB with the use of pseudoalignments. In practice, using pseudoalignments translates to observing only the number of reference sequences a read pseudoaligns against in each reference lineage. Combined, pseudoalignment and the use of many representative sequences for a lineage lead to a significantly improved accuracy when dealing with species that exhibit variability within the reference lineages, while also enabling the use of much larger reference sequence collections.

2.3.1 Mixture model formulation

Assume some set of sequencing reads $R = \{r_1, \ldots, r_N\}, N \in \mathbb{N}_+$ that are conditionally independent given the mixing proportions θ and identically distributed. While the assumption about conditional independence is useful in formulating the model, it does assume a certain structure in the reads that may not always hold depending on the sequencing technology used. However, the model in practice performs well with the assumption, justifying its use to speed up the analyses and simplify the model formulation.

The joint distribution for a generative mixture model that produced these reads can be written down by defining latent indicator variables $I = \{I_1, \ldots, I_N\}$ that follow some mixing proportions $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_S), S \in \mathbb{N}_+, \sum_{s=1}^{S} \theta_s = 1$. Because conditional independence between the reads R, $r_j \perp\!\!\!\perp r_i|\theta$ for all $i \neq j, 1 \leq i \leq N, 1 \leq j \leq N$ was assumed, the joint distribution for this generative model is

$$p(R, I, \boldsymbol{\theta}) = \prod_{n=1}^{N} p(r_n|I_n)\, p(I_n|\boldsymbol{\theta})\, p(\boldsymbol{\theta}). \tag{2.1}$$

Now, assume that for each read $r_n, 1 \leq n \leq N$, only alignments $r_{n,s}$ against the reference sequences $s = 1, \ldots, S$ are observed. The information contained in $r_{n,s}$ may be anything about the alignment such as its length, quality, or location. Furthermore, assume conditional independence between the alignments against different reference sequences $r_{n,i} \perp\!\!\!\perp r_{n,j}|\theta$ for all $i \neq j, 1 \leq i \leq S, 1 \leq j \leq S$. This leads to the joint distribution from Equation 2.1 factorizing into

$$p(R, I, \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{n=1}^{N} \prod_{s=1}^{S} p(r_{n,s}|I_n = s)\, p(I_n = s|\boldsymbol{\theta}). \tag{2.2}$$

The model in Equation 2.2 corresponds to the mixture model that has been historically used in the RNA-Seq context and the predecessor of mSWEEP. The difference between the various methods utilizing the model of Equation 2.2 is in the formulation of the likelihood term $p(r_{n,s}|I_n = s)$ and consideration of either reference sequences or reference lineages as the target of the latent indicator variables. Note that indexing with s in this

particular model denotes that the targets are reference sequences which represent a single taxonomic unit, and the inferred the inferred relative abundances $\boldsymbol{\theta}$ are the proportions of these reference sequences in the full set of reads R.

2.3.2 Incorporating grouped reference sequences

The model in Equation 2.2 performs admirably when the reference sequences s are sufficiently different from each other such as in the RNA-Seq context, however attempts to estimate the relative abundances of individual reference sequences s fail when the degree of relatedness is increased. Especially when applying the model to reference data containing sequences from strains of the same bacterial species, the abundances $\boldsymbol{\theta}$ tend to become scattered among the most closely related sequences — even if the correct sequence is contained in the reference.

The model used by BIB incorporates a clustering of the reference sequences into lineages to solve the problem presented by bacterial data. Including a clustering means that instead of estimating the relative abundances $\boldsymbol{\theta}$ for the individual reference sequences s, the abundance is estimated for some cluster of sequences $k, k = 1, \ldots, K, K \ll S$. Although this approach introduces an obvious loss of resolution when compared to the sequence-based approach, incorporating the use of a clustering provides advantages in accommodating for naturally occurring variation as well as improving the scalability of the inference part by reducing the number of reference units.

In terms of the model, the clustering is included in Equation 2.2 by replacing alignments against the reference sequences $r_{n,s}$ with alignments against the clusters $r_{n,k}, k = 1, \ldots, K$. With this replacement, the changes to the model are minimal: the s:s are simply replaced by k:s

$$p\left(R, I, \boldsymbol{\theta}\right) = p\left(\boldsymbol{\theta}\right) \prod_{n=1}^{N} \prod_{k=1}^{K} p\left(r_{n,k} | I_n = k\right) p\left(I_n = k | \boldsymbol{\theta}\right). \tag{2.3}$$

With an appropriate definition of the likelihood term $p\left(r_{n,k} | I_n = k\right)$ and consideration of the alignments $r_{n,k}$ as alignments against representative sequences from the cluster k, this model is the model used by BIB. The

BIB approach adequately solves the inference problem for several species of bacteria with well-defined clusterings within the species [54].

The model of Equation 2.3 has, however, several issues that render it difficult to apply in some scenarios. First, the model requires selecting a representative sequence for each cluster k. This selection is by no means an easy task and, secondly, using a representative sequence implies assumptions about the clustering: namely that there must be minimal variation within the clusters in terms of genomic content and that each cluster is clearly separated from the others; otherwise selecting a representative sequence is not feasible. In BIB, these requirements are somewhat alleviated by defining the core genome of the species and using only the parts of the representative sequence that belong to the core. Unfortunately, this introduces a third problem: increasing the number of sequences for any species of bacteria tends to shrink the core genome estimate, depending on the method used [85].

2.3.3 Modelling alignments against sequence groups

mSWEEP solves the issues present in the BIB model by replacing alignments against representative sequences with pseudoalignments (see Section 2.2.2 for more details) against all available reference sequences from each cluster. Although pseudoalignment reports less information about the relationship between the reads and the reference sequences than traditional alignment, including more reference sequences leads to excellent performance in cases where the BIB model fails and provides similar resolution in cases where the BIB model performs well [17].

With the changes in the mSWEEP model, the observations $r_{n,k}$ become the numbers of observed pseudoalignments $r_{n,k}, 0 \leq r_{n,k} \leq M_k$ against the M_k sequences belonging to a cluster k. If assumptions about conditional independence between the clusters are kept, the formulation for the model remains the same as the one presented in Equation 2.3 with the only changes being to the likelihood term $p\left(r_{n,k}|I_n = k\right)$.

2.3.4 Likelihood for a clustered reference

When dealing with pseudoalignments against clustered reference sequences, the likelihood term $p\left(r_{n,k}|I_n = k\right)$ in Equation 2.3 needs to be

carefully defined to account for several factors arising from the biology affecting the reference sequences. One, the clusters may vary greatly in size, with some of them having just one reference sequence and some hundreds or even thousands. Two, due to sequencing errors, reference errors (assembly errors or lack of closely related reference sequences from a cluster), and mutations, the read may not necessarily pseudoalign against any sequences in a cluster even though it belongs to the cluster. Three, since the sequences in a cluster share a significant degree of genetic material, a cluster with a higher fraction of sequences that the read aligned against should always be a better candidate for having produced the read. Four, the read can plausibly pseudoalign against several or even all of the clusters.

These four factors lead to considering a likelihood with the following properties: 1) within each cluster, and ignoring the case where no pseudoalignments are observed, the likelihood function must be increasing in the number of pseudoalignments (more alignments always means a better fit to the cluster); 2) the likelihoods from different clusters should be on the same scale regardless of the number of sequences in the cluster; and 3) the model should include zero inflation to account for nonalignment due to errors in the reads or the reference. This leads to defining the likelihood $p\left(r_{n,k}|I_n = k\right)$ in three parts

$$p\left(r_{n,k}|I_n = k\right) = \begin{cases} 0.01 \text{ if } r_{n,k} = 0, \\ 0.99 \text{ if } r_{n,k} = 1 \text{ and } M_k = 1, \\ 0.99 f\left(r_{n,k}, M_k\right) \text{ if } r_{n,k} \geq 1 \text{ and } M_k > 1, \end{cases} \tag{2.4}$$

where $f\left(r_{n,k}, M_k\right)$ is the main term defining the likelihood for clusters with more than one sequence, and is the term that should fulfill the requirements for the likelihood function.

In Equation 2.4, the first part provides a slight zero-inflation for the model, corresponding (roughly) to the error rate in Illumina sequencing data with a Phred quality score of Q20 [86, 87]. The second part handles the special case where the cluster k contains only one sequence ($M_k = 1$). For the final case, which represents the majority of the reference sequences in a setting where they can be plausibly assigned to clusters, the likelihood is defined by the term $f\left(r_{n,k}, M_k\right)$ which is a function of the pseudoalignment counts $r_{n,k}$ and the cluster size M_k.

Had the assumption about the comparability of fractions of alignments between clusters of different sizes not been made, a reasonable choice for $f\left(r_{n,k}, M_k\right)$ would be the beta-binomial distribution. This distribution is an extension of the binomial distribution that allows for modelling count data with over/under-dispersion through a 2-parameter formulation. With parameters $(n, \alpha, \beta), n \in \mathbb{N}, \alpha > 0, \beta > 0$, the beta-binomial distribution has the following probability mass function $p\left(k|n, \alpha, \beta\right)$ on the support $k \in \{0, \ldots, n\}$

$$p\left(k|n, \alpha, \beta\right) = \binom{n}{k} \frac{B\left(k + \alpha, n - k + \beta\right)}{B\left(\alpha, \beta\right)}. \tag{2.5}$$

In Equation 2.5, $B\left(\alpha, \beta\right)$ is the beta function

$$B\left(\alpha, \beta\right) = \int_0^1 t^{\alpha - 1} \left(1 - t\right)^{\beta - 1} dt, \tag{2.6}$$

and $\binom{n}{k}$ is the binomial coefficient.

However, since the assumptions made for the likelihood function require that a cluster with 100% of sequences compatible with the read is a better fit than another with only 99% regardless of their sizes M_k, the beta-binomial distribution of Equation 2.5 cannot be directly used. Nevertheless, a version of the likelihood function that is inspired by the beta-binomial distribution but fulfills the assumptions can be found. Namely, the function $p\left(k|n, \alpha, \beta\right)$ is modified by dividing each $p\left(k|n, \alpha, \beta\right)$ with their respective maximum values $p\left(n|n, \alpha, \beta\right)$, changing their range to $[0, 1]$ regardless of the parameter values (n, α, β). Note that achieving the maximum value at $k = n$ requires restricting the parameter values of the original beta-binomial distribution so that its probability mass function is increasing. This assumption is fulfilled when $\alpha \left(\alpha + \beta\right)^{-1} \in (0.5, 1)$ [88].

Performing the scaling of Equation 2.5 by the maximum value $p\left(n|n, \alpha, \beta\right)$ results in the following scaled likelihood function $p^\star\left(k|n, \alpha, \beta\right)$

$$\begin{aligned} p^{\star}\left(k|n,\alpha,\beta\right) &= \frac{p\left(k|n,\alpha,\beta\right)}{p\left(n|n,\alpha,\beta\right)} \\ &= \binom{n}{k}\frac{B\left(k+\alpha,n-k+\beta\right)}{B\left(\alpha,\beta\right)}\binom{n}{n}^{-1}\frac{B\left(\alpha,\beta\right)}{B\left(n+\alpha,\beta\right)} \\ &= \binom{n}{k}\frac{B\left(k+\alpha,n-k+\beta\right)}{B\left(n+\alpha,\beta\right)}. \end{aligned} \tag{2.7}$$

Since the scaling in Equation 2.7 is by a constant, $p\left(n|n,\alpha,\beta\right)$, the resulting function $p^{\star}\left(k|n,\alpha,\beta\right)$ remains an increasing function of k when the original function $p\left(k|n,\alpha,\beta\right)$ is increasing.

With the probability mass function of the distribution $p^{\star}\left(k|n,\alpha,\beta\right)$, the full definition for the third part of the likelihood in Equation 2.4 is

$$p\left(r_{n,k}|I_n=k\right) = 0.99\frac{p^{\star}\left(r_{n,k}|M_k,\alpha,\beta\right)}{Z\left(r_{n,k}\right)} \text{ if } r_{n,k}\geq 1 \text{ and } M_k > 1, \tag{2.8}$$

where $Z\left(r_{n,k}\right)$ is a normalizing constant. The scaling in Equation 2.7 fulfills the requirement that the likelihood of each cluster must be on the same scale despite different size. The next section will derive a closed form for the normalizing constant $Z\left(r_{n,k}\right)$.

2.3.5 Normalizing the likelihood

While the function $p^{\star}\left(k|n,\alpha,\beta\right)$ in Equation 2.7 closely resembles the probability mass function of a beta-binomial distribution (Equation 2.5), the function $p^{\star}\left(k|n,\alpha,\beta\right)$ by itself does not sum to 1 over its support $k\in\{1,\ldots,K\}$, which means that the function is not a proper probability mass function. To remedy this, the normalizing constant $Z\left(r_{n,k}\right)$ is needed.

In principle any distribution on a finite support can be normalized but in many cases the normalizing constant does not have a closed form. Fortunately, it turns out that — thanks to the properties of the beta function (see Equation 2.6 for the definition) — $Z\left(r_{n,k}\right)$ does have a closed form. Deriving this closed form requires using the following identity for the beta function which is derived in Theorem 2.1.

Theorem 2.1

$$B\left(a+1,b\right) = \frac{a}{a+b}B\left(a,b\right).$$

Proof. Follows from $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ [89] and $\Gamma(z+1) = z\Gamma(z)$, for all $z > 0$ [90], where $\Gamma(z)$ is the gamma function $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx, z > 0$. Using these two identities, $B(a+1,b)$ can be written as

$$\begin{aligned} B(a+1,b) &= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\ &= \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} \\ &= \frac{a}{a+b}B(a,b). \end{aligned}$$

□

Applying Theorem 2.1 leads to the closed form of the normalizing constant $Z(r_{n,k})$.

Theorem 2.2 *Let*

$$f(k,n) = \binom{n}{k}\frac{B(\alpha+k, n-k+\beta)}{B(\alpha+n,\ \beta)}, 0 \le k \le n,\ \alpha > 0,\ \beta > 0,$$

and

$$Z(n) = \prod_{j=1}^{n}\frac{\alpha+n+k-j}{\alpha+\beta+2n-j},$$

then

$$\sum_{k=0}^{n}\frac{f(k,n)}{Z(n)} = 1.$$

Proof. Consider a beta-binomial distribution with the parameters $(n, \alpha+n, \beta), n \in \mathbb{N}_+, \alpha > 0, \beta > 0$. This distribution has the probability mass function $g : 0, \ldots, n \to (0,1)$, where

$$g(k \mid n, \alpha+n, \beta) = \binom{n}{k}\frac{B(\alpha+n+k, n-k+\beta)}{B(\alpha+n,\beta)},\ 0 \le k \le n.$$

Using the identity $B(a+1,b) = B(a,b)\frac{a}{a+b}, a > 0, b > 0$ (Theorem 2.1) results in an alternative form for g:

$$
\begin{aligned}
g(k) &= \binom{n}{k} \frac{B(\alpha+n+k, n-k+\beta)}{B(\alpha+n, \beta)} \\
&= \binom{n}{k} \frac{B(\alpha+n+k-1, n-k+\beta)}{B(\alpha+n, \beta)} \frac{\alpha+n+k-1}{\alpha+n+k-1+n-k+\beta} \\
&= \binom{n}{k} \frac{B(\alpha+n+k-1, n-k+\beta)}{B(\alpha+n, \beta)} \frac{\alpha+n+k-1}{\alpha+\beta+2n-1} \\
&= \binom{n}{k} \frac{B(\alpha+n+k-2, n-k+\beta)}{B(\alpha+n, \beta)} \frac{\alpha+n+k-1}{\alpha+\beta+2n-1} \frac{\alpha+n+k-2}{\alpha+\beta+2n-2} \\
&= \binom{n}{k} \frac{B(\alpha+n+k-2, n-k+\beta)}{B(\alpha+n, \beta)} \prod_{j=1}^{2} \frac{\alpha+n+k-j}{\alpha+\beta+2n-j}.
\end{aligned}
$$

Above, Theorem 2.1 was applied twice. Repeatedly applying Theorem 2.1 n times yields the alternative form

$$
\begin{aligned}
g(k) &= \binom{n}{k} \frac{B(\alpha+n+k-n, n-k+\beta)}{B(\alpha+n, \beta)} \prod_{j=1}^{n} \frac{\alpha+n+k-j}{\alpha+\beta+2n-j} \\
&= \binom{n}{k} \frac{B(\alpha+k, n-k+\beta)}{B(\alpha+n, \beta)} \prod_{j=1}^{n} \frac{\alpha+n+k-j}{\alpha+\beta+2n-j} \\
&= f(k, n) \prod_{j=1}^{n} \frac{\alpha+n+k-j}{\alpha+\beta+2n-j}.
\end{aligned}
$$

Since $g(k)$ is a probability mass function, this implies that

$$
f(k, n) \left(\prod_{j=1}^{n} \frac{\alpha+n+k-j}{\alpha+\beta+2n-j} \right)^{-1}
$$

is also a probability mass function. Thus, setting

$$
Z(n) = \prod_{j=1}^{n} \frac{\alpha+n+k-j}{\alpha+\beta+2n-j}
$$

is sufficient to normalize $f(k, n)$ and prove Theorem 2.2. □

2.3.6 Likelihood hyperparameters

Instead of the traditional parametrization for the beta binomial distribution through $\alpha > 0, \beta > 0$, in Publication I the distribution is reparametrized to slightly change the interpretation of the parameters. The reparametrised forms for α and β are

$$\pi = \frac{\alpha}{\alpha + \beta}, \phi = \frac{1}{\alpha + \beta}, \tag{2.9}$$

where the first parameter π has the range $\pi \in (0, 1)$ unless constraints are placed on α and β and represents the mean success rate in repeated draws from the beta binomial distribution. The second parameter $\phi > 0$ measures the variation in the success rate for each draw [91]. In the formulation for the likelihood in Equation 2.8, each cluster k has its own parameters π_k, ϕ_k.

Although methods such as Bayesian optimization [92] could be employed to find optimal values for the parameters π_k, ϕ_k in Equation 2.9, their values are set based on a reasonable compromise that performed well in Publication I. The values of π_k, ϕ_k are set to

$$\begin{aligned} \pi_k &= 0.65, \text{ for all } k = 1, \ldots, K, \\ \phi_k &= 1 - \pi_k + 0.01 M_k^{-1}. \end{aligned} \tag{2.10}$$

2.3.7 Fitting the model using variational inference

With the likelihood defined in Equations 2.4 and 2.8, the remaining task is to come up with a suitable method to infer the relative abundances $\boldsymbol{\theta}$. Since the model is principally the same as the one used in BIB (Equation 2.3), just with a different formula for the likelihood term $p\left(r_{n,k} = k | I_n = k\right)$, the variational inference algorithm from BIB can be adjusted by simply changing the likelihood term to that of Equation 2.4 and the rest of the algorithm remains the same.

Variational inference by itself is an extremely broad topic that lies somewhat outside the scope of this dissertation. Thus, this section only covers the parts that are directly relevant to the contributions from this thesis — namely, how the probability matrix that mSWEEP generates and mGEMS leverages is obtained. For a more thorough coverage of variational inference in this context, the BitSeqVB publication [73] provides

an explanation for the case where the algorithm is derived for the same model, only with a different likelihood function.

In brief, variational inference for the mSWEEP model [93] consists of finding a distribution $q(\boldsymbol{\theta}, I)$ that minimizes the Kullback-Leibler divergence to the true posterior $p(\boldsymbol{\theta}, I|R) \approx q(\boldsymbol{\theta}, I)$. For simplicity, assume that the approximation $q(\boldsymbol{\theta}, I)$ factorizes into $q(\boldsymbol{\theta}, I) = q(\boldsymbol{\theta})\, q(I)$. Because each I_n has a categorical distribution $Cat(\boldsymbol{\theta})$ and they are furthermore assumed independent of each other given the mixing proportions, $I_n \perp\!\!\!\perp I_m, n \neq m|\theta$, the second term $q(I)$ simplifies to

$$q(I) = \prod_{n=1}^{N} \prod_{k=1}^{K} \gamma_{n,k}^{I_{n,k}}. \tag{2.11}$$

In practice, the best approximation $q(\boldsymbol{\theta}, I)$ is found when optimal values for the parameters $\gamma_{n,k}$ in Equation 2.11 are found [73]. The Riemannian conjugate gradient method and variational Bayesian expectation maximization steps are used to find the optimal values for $\gamma_{n,k}$ [73, 93, 94]. A generic, parallel and distributable implementation for arbitrary likelihood with this mixture model structure is available from `https://github.com/tmaklin/rcgpar`.

2.3.8 Alternative fit using MCMC sampling

Alternatively, the model could be fitted using Markov chain Monte Carlo (MCMC) sampling methods [72]. Instead of the variational inference approach of finding an approximating distribution $q(\boldsymbol{\theta}, I)$, MCMC sampling attempts to produce a set of samples from the true posterior $p(\boldsymbol{\theta}, I|R)$. Averaging over the values produced via MCMC sampling $\hat{\boldsymbol{\theta}}$ (asymptotically) produces the true parameters $\boldsymbol{\theta}$ as the number of samples increases.

Even though producing the true values is tempting, MCMC has several problems when applied in practice [95]. First, the true values are only found asymptotically, meaning that it is difficult to determine when the MCMC sampler has converged to the true posterior. Secondly, due to the first point, the number of samples that need to be drawn may be excessively high, resulting in long run times when compared to the significantly faster variational inference [95].

Compared to variational inference, MCMC does have the advantage that, provided sufficient runtime, the samples will be from the true posterior. In practice, this leads to better estimates of the covariance between the sampled parameters when the model is of the form in Equation 2.1 [73]. However, replacing the variational inference algorithm in the mSWEEP model with the Gibbs sampler from the original BitSeq [72] does not seem to produce any significant differences between the parameter estimates (Figures 2.1 and 2.2).

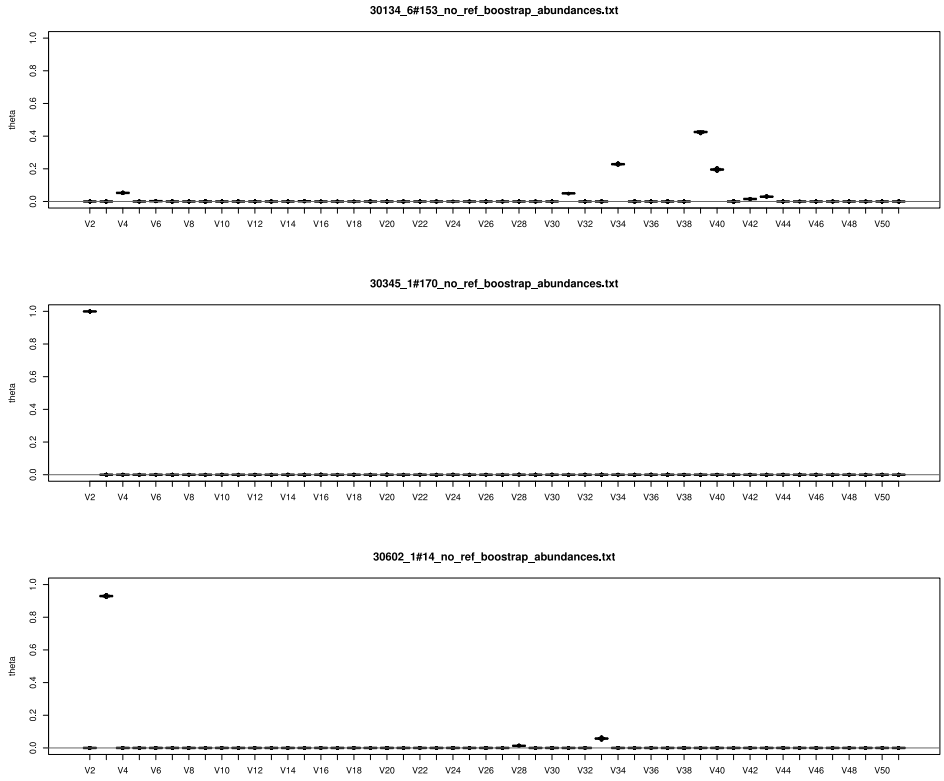

Figure 2.1: Parameter estimates for the mSWEEP model inferred using the variational inference implemented in BitSeqVB [73] on the *in vitro* samples from Publication II [28]. The true value for each subplot is 1.0 at the column corresponding to the highest estimate. Each boxplot contains parameter estimates from bootstrapping the pseudoalignments that are used as input to the variational inference algorithm.

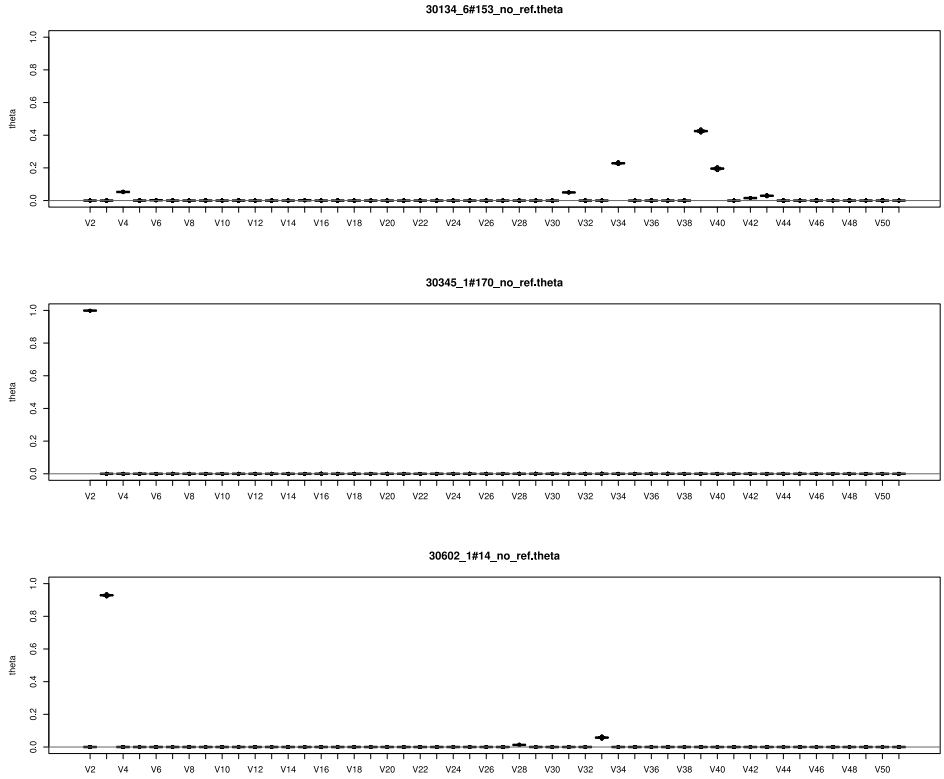

Figure 2.2: Parameter estimates for the mSWEEP model inferred using the collapsed Gibbs sampler implemented in BitSeq [72] on the *in vitro* samples from Publication II [28]. The true value for each subplot is 1.0 at the column corresponding to the highest estimate. Each boxplot contains the samples from the posterior obtained using the collapsed Gibbs sampler.

2.4 From profiling to binning

This section covers using the model from mSWEEP to derive an algorithm for assigning the sequencing reads $r_{n,k}$ to the reference lineages k, also known as binning. Binning differs from relative abundance estimation in that the goal is to produce some assignment of the reads to reference units. Typically the reference units and the created bins correspond to some species or even genera. In this thesis, the bins will be created on the level of the reference lineages/clusters that mSWEEP reports the relative abundances for. Compared to estimating only the abundances, the addition of binning provides much extra detail about the contents of a sample, since the creation of lineage-level sequencing read bins allows performing many downstream analyses that require sequencing reads or even assemblies. The binning algorithm presented in this section is called the mGEMS binning algorithm, which in itself is a part of the mGEMS pipeline for binning sequencing reads. The work in this section is based on the results from Publication II.

2.4.1 The mGEMS binning algorithm

A crucial feature for the binning algorithm to handle lineage-level differences between bacteria is that the algorithm must allow for assignment of a single read to multiple bins at the same time. Because of the relatively small differences between different lineages of a bacterial species, the read could easily have been generated from several of them. This differs from most work on binning, where the reads are typically only allowed an assignment in a single bin at a time because the variation between the organisms belonging to different bins is assumed large enough that multi-bin assignment may not be necessary. For lineage-level binning, this assumption obviously does not hold because of the shared genomic contents when the species are defined in a manner that reflects their phylogenetic characteristics. This means that the mGEMS binning algorithm must be explicitly defined in a way that allows for assignment to multiple bins.

The mGEMS binning algorithm consists of a rule for assigning the reads to the bins. This rule is derived by leveraging the assignment probabilities for each read $\gamma_{n,k} \in (0,1)$, $n = 1,\dots,N$, $k = 1,\dots,K$, $\sum_{k=1}^{K} \gamma_{n,k} = 1$ produced by the variational approximation used in fitting

the mSWEEP model (Equation 2.11). To derive the assignment rule, some further assumptions regarding the reads and the reference sequences are required. Firstly, the sequencing reads are assumed to be generated from only one strain belonging to the same lineage — similarly to the typical metagenomic binners assuming only variation at the level they operate on. Secondly, should the true reference sequence that generated the reads be missing from the reference, the set of reference sequences in the lineage that generated the reads is assumed to adequately cover the variation in the missing sequence. The second assumption is necessary since reads that do not pseudoalign to any reference sequence in the collection are discarded by mSWEEP.

To fulfill our requirement that a read can be assigned to several bins at the same time, the bins G_k for each cluster k are defined as a subset of sequencing reads r_n such that

$$G_k = \{r_n : \gamma_{n,k} \geq q_k\} \tag{2.12}$$

holds for some threshold $q_k \in [0, 1]$. Note that the threshold may be different for each cluster k. Because of the way the bins G_k are defined in Equation 2.12, this definition obviously allows for a read to belong to several bins (for a trivial example, consider the case where $q_k = 0$ for all k).

2.4.2 Assignment rule for multi-cluster membership

Next, the thresholds q_k should be assigned some sensible value that maximizes the probability $A_{n,k}$ of assigning the read r_n to the bin G_k if the cluster k (could have) generated the read r_n. Ideally, the probabilities $A_{n,k}$ could be defined through other probabilities $B_{n,k}$ with the meaning: the cluster k contains a sequence that contains the true (error-free) nucleotide sequence of the read r_n. However, the probabilities $B_{n,k}$ are quite difficult to estimate directly since 1) the reads cannot be error-corrected with full accuracy, and 2) the reference collection is nearly always incomplete.

These two problems can be remedied by assuming that the sample is mostly composed of closely related organisms, which implies that when $P[A_{n,k} = 1] \geq \theta_k$, then $P[B_{n,k} = 1]$ must be "large" because the cluster must contain a sequence to generate it. A more detailed derivation for this

statement about the magnitude of $B_{n,k}$ and is provided in the methods section of Publication II, supplied in the appendix for this thesis and omitted from here.

The implied statement about the magnitude of $B_{n,k}$ when $P[A_{n,k} = 1] \geq \theta_k$ means that there is a high chance that $B_{n,k} \to 1$ when the former holds. Because of this relationship, an assignment rule can be derived by using the estimates $\gamma_{n,k}$ (Equation 2.11) for the probability $P[A_{n,k} = 1]$

$$\text{if } \gamma_{n,k} \geq \theta_k, \text{ assign the read } r_n \text{ to } G_k. \tag{2.13}$$

Equation 2.13 provides an inequality whose validity can be checked to assess the probability of the event $B_{n,k} = 1$ which could not be estimated directly.

2.4.3 Practical considerations

While the assignment rule in Equation 2.13 provides a theoretically sound tool to assign reads r_n to the bins G_k, applying it in practice requires a slight adjustment due to computational accuracy. Namely, when estimating the relative abundances θ_k of N reads, any estimate that falls below $\frac{1}{N}$ means that zero reads originated from the cluster k. Because of this, values $\theta_k < \frac{1}{N}$ are in some sense meaningless, and all represent the same case of 0 reads from the clusters where the inequality is true. Due to the constraint that θ_k must sum up to 1 over k, these essentially-zero values do, however, contribute a small amount of noise to the other estimates that exceed $\frac{1}{N}$. Since there are K clusters, the fraction of noise d is (in the worst-case scenario) at most

$$d = (K - 1)\frac{1}{N}. \tag{2.14}$$

The noise-level in the worst-case scenario of Equation 2.14 means that when evaluating the validity of the inequality in Equation 2.13, the thresholds θ_k should be adjusted with $1 - d$. This adjustment in turn means that (in the worst-case scenario) only the fraction of relative abundance that is assigned to nonzero estimates is considered. Adjusting Equation 2.13 with d produces the final assignment rule that is used in mGEMS:

$$\text{if } \gamma_{n,k} \geq (1 - d)\theta_k, \text{ assign the read } r_n \text{ to } G_k. \tag{2.15}$$

Because the variational approximation used to fit the mSWEEP model already provides both the estimates $\gamma_{n,k}$ and the relative abundances θ_k, the assignment rule in Equation 2.15 is in practice inexpensive to evaluate after the model has been fitted.

2.4.4 General applicability of the assignment rule

Since the relative abundances θ_k are derived from the values $\gamma_{n,k}$ by averaging over $n = 1, \ldots, N$, the assignment rule in Equation 2.15 can be seen as a way to cluster the rows (or columns) of a generic probability matrix, whose rows (or columns) sum up to 1. This rule in particular allows assigning each row (or column) to several clusters at the same time. However, more general applicability of the rule to probability matrices is not explored further in this thesis beyond this acknowledgement that the rule could be applied to more general scenarios where a probability matrix needs to be clustered.

The next chapter will cover the application of mSWEEP and mGEMS to different kinds of sequencing data and explore how the methods enable new directions in analysis of sequencing data. The chapter also includes a coverage of the benchmarks and experiments presented in Publication I and Publication II.

Chapter 3

High-resolution metagenomics

High-resolution metagenomics (in the context of this thesis) refers to applying some method capable of recovering variation at the lineage-level to some kind of metagenomics data as defined in Section 1.1. In this chapter, the methods of interest are mSWEEP and mGEMS. The chapter will deal with the benchmarking and experimental results from Publication I and Publication II, which focus on analysing plate sweep metagenomics data. At the end of the chapter some technicalities arising from the model formulation presented in Chapter 2 will also be covered. These mostly cover questions about reliability of the results when applied to realistic use-cases.

3.1 Plate sweep and whole community metagenomics

The primary methods for obtaining metagenomics sequencing data considered are plate sweep metagenomics and whole community metagenomics (see Section 1.1 for more details). Of these two, plate sweep metagenomics has the advantage of being able to focus sequencing efforts to species that are known to grow on specific culture media, while whole community metagenomics provides data of *all* organisms on some sample (including for example host DNA, fungi, commensal species, and so on). When the goal is to investigate lineage-level variation, both approaches have their uses in either producing more data from the species of interest or provid-

ing a less biased view of the sample contents at the expense of sequencing depth. Publications I-II were written with only plate sweep metagenomics in mind but Publication III demonstrated that mSWEEP and mGEMS are applicable in the whole community metagenomics context. Thus, this chapter will not discriminate between the two.

3.1.1 Benefits of metagenomics over culturing

When comparing metagenomics approaches to culturing isolates, the major difference between the approaches is that metagenomics will provide a better overview of the microbiome in the sample. Although isolate data has mostly been used in the past in epidemiological studies, some demonstrations of the benefits of applying a metagenomics approach have emerged. In particular, a recent study utilizing mSWEEP demonstrated that using isolate data alone does not allow identifying the presence of non-dominant variants of *Streptococcus pneumoniae* [40]. Currently it is a somewhat of an open question whether similar naturally occurring variation is found ubiquituously across different microbiomes

mSWEEP and mGEMS enable answering the question regarding lineage-level variation by providing methods that can be targeted to capture the variation present in samples containing members of some bacterial species. Since many bacteria of clinical importance have been studied for several decades with significant sequencing efforts aimed at them to obtain high-quality genome assemblies, the reference-based approach employed by mSWEEP and mGEMS is ideal for disentangling variation in the same clinical setting. When combined with tools that take different approaches to analysing metagenomic data, such as the StrainPhlAn [25] and StrainGE [26] methods that track bacterial strains across samples, the combination has the potential for providing unprecedented level of detail in future epidemiological analyses and routine surveillance.

While the previous chapter covered the theoretical foundations of mSWEEP and mGEMS, this chapter will focus on practical considerations regarding the two methods. Namely, the chapter briefly overviews their performance in various settings, and covers questions related to reliability of the approaches and sensitivity to the reference sequence collection. The last part of the chapter covers additional results from Publications I-III, exploring what kind of information metagenomics-derived results provide.

3.1.2 The mSWEEP/mGEMS pipeline

Combining the results from Sections 2.3 and 2.4 produces the complete mSWEEP/mGEMS workflow for analysing sequencing data from mixed sources. This workflow is originally presented in Publication II, where it is referred to as the mGEMS pipeline. Figure 3.1 provides an overarching diagram representing the steps described in this section.

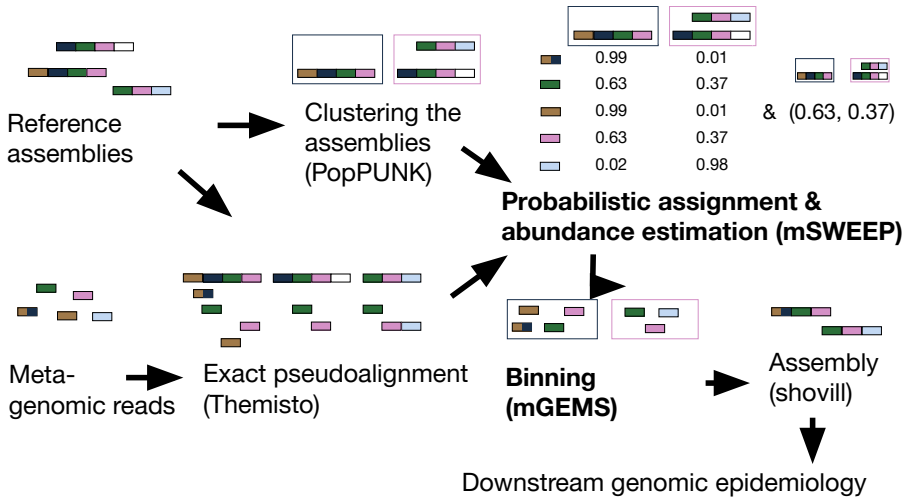

Figure 3.1: Flowchart describing a genomic epidemiology workflow with the mGEMS pipeline. The figure shows the various steps of the pipeline. Steps with programme names in brackets constitute the parts of the mGEMS pipeline. Presented values from mSWEEP and mGEMS binner are the actual results of running the pipeline with the described input. The steps that are perfomed by the methods described in this thesis are bolded (mSWEEP and mGEMS).

Figure 3.1 and legend source: adapted from Publication II, Figure 1 [28].

Constructing the reference database

The pipeline begins with constructing a set of reference sequences that represent the variation in the target species of interest. In an ideal scenario, the reference should consist of high-quality assemblies from each lineage that is expected to be found in the sequencing reads. Since this is a rather strong assumption, typical use cases exploit published datasets and possibly combine them with bespoke isolate sequencing data. Either previously published assemblies, or even curated genomes from databases such as RefSeq [96], may be used. The reference may also include newly assembled sequences or otherwise be tailored to the problem at hand. In both cases, a cautious approach regarding the inclusion of potentially low-quality assemblies in the reference is recommended, as the quality of the reference sequence collection is the most important factor in obtaining trustworthy results from the pipeline.

After the appropriate reference sequences have been collected, they should be clustered in some meaningful way to obtain the lineage grouping. For some species, multilocus sequence typing is sufficient, but for others with more variable genome contents, algorithms that attempt to identify clonal complex analogues (central multilocus sequence type and its 1 or 2 locus variants) may be useful. One such algorithm is PopPUNK [79], which is demonstrated to perform well in Publication III. PopPUNK clusters the reference sequences based on accessory and core genome distances with an option to perform the clustering only based on the core-genome. The resulting clustering from PoPPUNK often corresponds to clonal complexes. Using a computational approach like PopPUNK instead of a curated database approach like the sequence types and clonal complexes has the advantage of providing means to assign sequences that have not yet been included in the curated databases, or work with species for which such databases do not exist.

After clustering the reference sequences, the next step is to build an index for pseudoalignment, and pseudoalign the reads from the samples against the index. In the mGEMS pipeline, the Themisto method is used [28] to perform both the index construction and the pseudoalignment. The pseudoalignment step produces binary pseudoalignment vectors for each read against every reference sequence, which are used as the input to mSWEEP.

Estimating relative abundances and binning the reads

The next step in the pipeline is to use mSWEEP to estimate the relative abundances of the reference groups based on the pseudoalignments. This is performed directly on the output from Themisto, with no intermediate steps required. After the relative abundances have been estimated, the results are fed to mGEMS which produces the read bins and optionally also extracts the reads corresponding to each bin from the original set.

Assembling the read bins

In Publication II, the mGEMS pipeline also contains an optional step to assemble the reads placed in each bin. The suggested assembler is shovill[1], which is an assembly pipeline built around the SPAdes assembler [97] but incorporates some pre- and post-processing steps. Naturally, other assemblers may also be used, or the assembly step skipped entirely and the analysis instead focused on the reads. In Publication II, the analyses mostly focused on using the assemblies, as including an assembly step is equivalent to adding a post-processing step that aids in filtering out reads that may mistakenly have been assigned to the wrong bin.

Since mGEMS allows for assigning a sequencing read to several bins/lineages at once, the produced bins may result in very high coverages for genomic sequences that are shared by multiple organisms belonging to different bins. Consequently, Publication II investigated the effect of replacing the isolate-data optimised shovill with metagenomic assemblers [98–100], which presumably implement better handling of variable coverage in the produced genomes. Although this resulted in some differences in the resulting assemblies (Figure 3.2), particularly when mGEMS was paired with IDBA-UD which resulted in highly fragmented assemblies or metaSPAdes which completely failed to assemble some sequences, there was no conclusive evidence in favour of using either of the best performing approaches (mGEMS/shovill or mGEMS/MEGAHIT). Regardless of the assembler choice, using the assemblies from the mGEMS pipeline in downstream analyses performed similarly to using those created from corresponding isolate sequencing data.

[1] `https://github.com/tseemann/shovill`

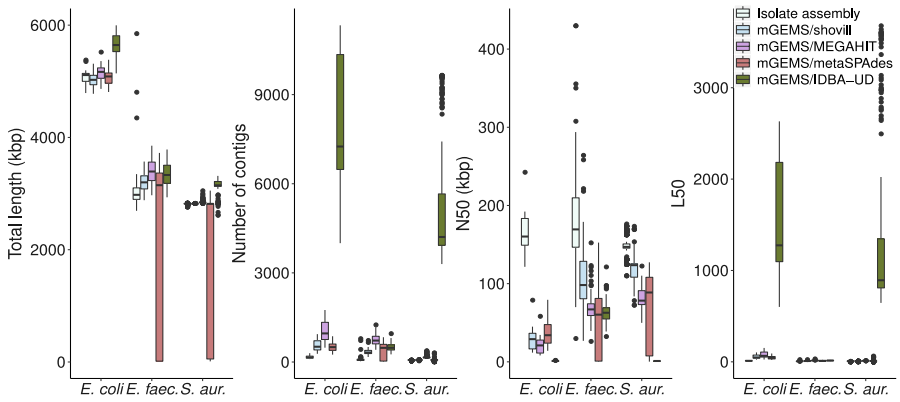

Figure 3.2: Comparing mGEMS-derived assemblies with different assembler choices (shovill, MEGAHIT [99], metaSPAdes [100], and IDBA-UD [98]). The boxes are colored according to the assembler used. Presented statistics are the summed lengths of all contigs (total length), the number of contigs, the sequence length of the shortest contig at 50% genome length (N50), and the smallest number of contigs whose sum of lengths is at least 50% of the genome length (L50).

Quality control

The mGEMS pipeline as described above is the method that has been applied in Publication III with the additional inclusion of a quality control (QC) step attempting to identify whether the reference sequences suitably cover the variation in the sequencing reads. This QC step, called demix_check[2], performs several checks on the results from mSWEEP and mGEM, to determine whether the created read bins correspond to some reference cluster. Although the demix_check step was not used in Publications I-II that introduced mSWEEP and mGEMS, its inclusion addresses an important question regarding the applicability of the results from mSWEEP/mGEMS. Therefore, including demix_check — or other

Figure 3.2 and legend source: adapted from Publication II, Figure 3 [28].

[2] `https://github.com/harry-thorpe/demix_check`

similar approach — as part of the mGEMS pipeline between the mGEMS and the assembly steps is recommended for a rigorous approach. This and other questions related to quality control and reliability of the results are explored further down in this chapter.

3.1.3 Other approaches for metagenomic analyses

While this thesis deals with the development and usage of mSWEEP and mGEMS, one has to acknowledge that the use of metagenomic sequencing data is by no means an understudied field. In fact, many methods exist that aim to perform similar tasks ranging from genome assembly from metagenomic sequencing data (metagenome assemblers) [98–100] to taxonomic binning (metagenomic binners) [101–103] and profiling [26, 50], and strain tracking (StrainGE and Strainphlan) [25, 26]. Compared to mSWEEP and mGEMS, these methods typically assume that the samples only contain a single strain from each species, with the exception of StrainGE, which explicitly addresses the presence of several strains, which enables them to solve the task when the assumption holds, but does not allow for extraction of the reads like mGEMS.

3.2 Benchmarking mSWEEP and mGEMS

This section will briefly cover the results related to benchmarking the performance of mSWEEP and mGEMS in Publication I and Publication II. A majority of these benchmarks were performed on synthetic mixtures of real sequencing reads that were obtained from isolate cultures. Publication II additionally provides a benchmark that used *in vitro* mixtures of DNA from isolate cultures. Figures from Publications I-II have been reproduced when necessary and are marked appropriately.

3.2.1 mSWEEP

Publication I presented a comparison of mSWEEP with the Bayesian Identification of Bacteria (BIB) [54] and the pseudoalignment-based metakallisto methods [74]. The vast majority of other metagenomics tools published at the time either did not attempt lineage-level profiling or had been developed for cases with only one strain present from

each species. Because of these limitations, Publication I only makes the comparisons with BIB and metakallisto, which have been developed for settings with several strains present. Additionally, since the preceding metagenomics tools have typically not been developed with the strain-complexity in mind, a performance comparison between them and mSWEEP (or mGEMS) is not meaningful nor fair.

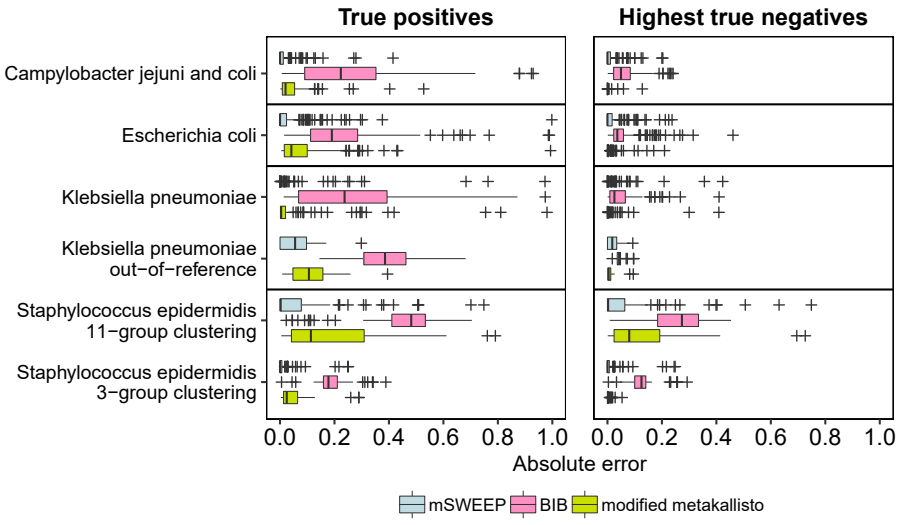

Figure 3.3: Comparison between mSWEEP, BIB, and modified metakallisto. This figure shows the differences in accuracy for the abundance estimates from mSWEEP, BIB, and a modified version of metakallisto. Modified metakallisto sums up the abundances within the lineages rather than simply reporting the abundances for the individual reference sequences. True positives refer to the relative abundance estimates in the true lineage. Highest true negatives refer to the highest estimate in the incorrect lineages. The absolute error is the difference from an abundance of one (True positives) or from zero (Highest true negatives).

Figure 3.3 and legend source: adapted from Publication I, Figure 2 and Extended Data figure S1 [17].

The approach used by BIB is similar to the one in mSWEEP in that the reference sequences are grouped together into lineages and estimation is performed on the level of these lineages. Metakallisto attempts the much more ambitious task of estimating the relative abundances for the individual sequences. Hence, a direct comparison between the three is not possible because metakallisto reports the abundance estimates for the sequences. This was addressed in Publication I by modifying the output from metakallisto to include a step where the abundance estimates within the same lineage are summed up. Even though this step was not included in the original metakallisto publication, its addition helps to compare the estimates from mSWEEP, BIB, and metakallisto.

One of the main results of Publication I is that mSWEEP outperforms both BIB and metakallisto (Figure 3.3), and that incorporating the probabilistic model from mSWEEP proves to be a necessary step in obtaining accurate information. Although these performance benchmarks were only performed on data containing a single strain in each sample, Publication I demonstrated through stochastic dominance [104, 105] that the methods which do not succeed with single-strain estimates are unlikely to provide accurate results from multi-strain samples.

Another result from Publication I concerns benchamrking the performance in presence of several strains from the same species. In this benchmark, sequence data from three strains, obtained via isolate sequencing, were mixed together in single sample at known proportions. Then, mSWEEP was applied to estimate the proportions when the real reference sequences were removed from the reference collection but at least one close representative from the same lineage was still available. In this setting, mSWEEP demonstrated very accurate performance when measured on both true positive and true negative estimates (Figure 3.4). The *K. pneumoniae* benchmark proved somewhat more challenging than the others, possibly due to the excessive genomic variation present in the species [106], but the errors in the results were nevertheless within acceptable limits when measured by the relative abundances for the correct lineage and the incorrect lineages.

Figure 3.4 and legend source: adapted from Publication I, Figure 4 [17].

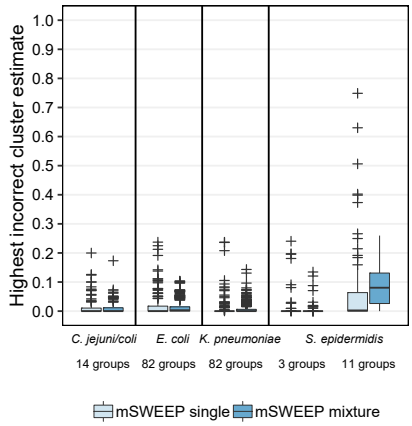

Figure 3.4: Abundance estimates from synthetic mixtures of three lineages do not result in higher number of false positive estimates when compared to estimates from the single-colony samples, as measured by the largest estimate for a lineage that does not contribute any sequencing reads. The only exception is the S. epidermidis 11-cluster case which is not accurately identified in neither the synthetic mixtures nor the single-colony samples.

3.2.2 mGEMS

The taxonomic binner mGEMS was, in turn, benchmarked in Publication II using again both synthetic mixtures of reads from isolate sequencing, and on an *in vitro* that contained measured amounts of DNA from known sources. For mGEMS, the chief measures of accuracy used were related to those that are the main objects of interest in genomic epidemiological analyses: SNPs and phylogenies estimated from the SNP data. In the synthetic mixtures, the performance of mGEMS was evaluated at the level of distinguishing between different clades within a specific sequence type using data from *E. coli* [107], at the level of distinguishing between different sequence types using data from *E. faecalis* [108], and at an extreme level with only dozens of SNPs separating the different reference clusters using data from *S. aureus* [56].

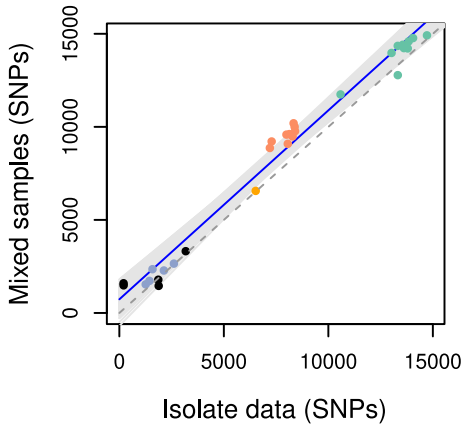

Figure 3.5: SNP calling from mGEMS-derived assemblies versus isolate assemblies for *E. coli* ST131. SNPs were called from contigs after assembling the reads. Points are colored according to ST131 sublineages. The dashed gray line represents a perfect match. The blue line is the posterior mean and the shaded area the 95% posterior credible region calculated from 10000 posterior samples using a Bayesian linear regression model.

Synthetic mixture benchmarks

The *E. coli* benchmark investigated how mGEMS-derived assemblies performed for maximum likelihood phylogeny estimation (using RAxML-NG, Gamma+GTR4M model) [109] when compared to using isolate sequencing data in the same pipeline. The core genome alignment required by RAxML-NG was estimated using snippy[3] with the same reference genome for both mGEMS-derived and isolate data assemblies. Calling SNPs from the reuslting assemblies shows that mGEMS tends to slightly overestimate the number of SNPs in these assemblies (Figure 3.5) but the phylogenetic relationships (Figure 3.6) are recovered well. This benchmark was performed at the level of variation within a sequence type (*E. coli* ST131) using sublineages defined in a previous study [82].

[3] `https://github.com/tseemann/snippy`

Figure 3.5 and legend source: adapted from Publication II, Figure 3 [28].
Figure 3.6 and legend source: adapted from Publication II, Figure 4 [28].

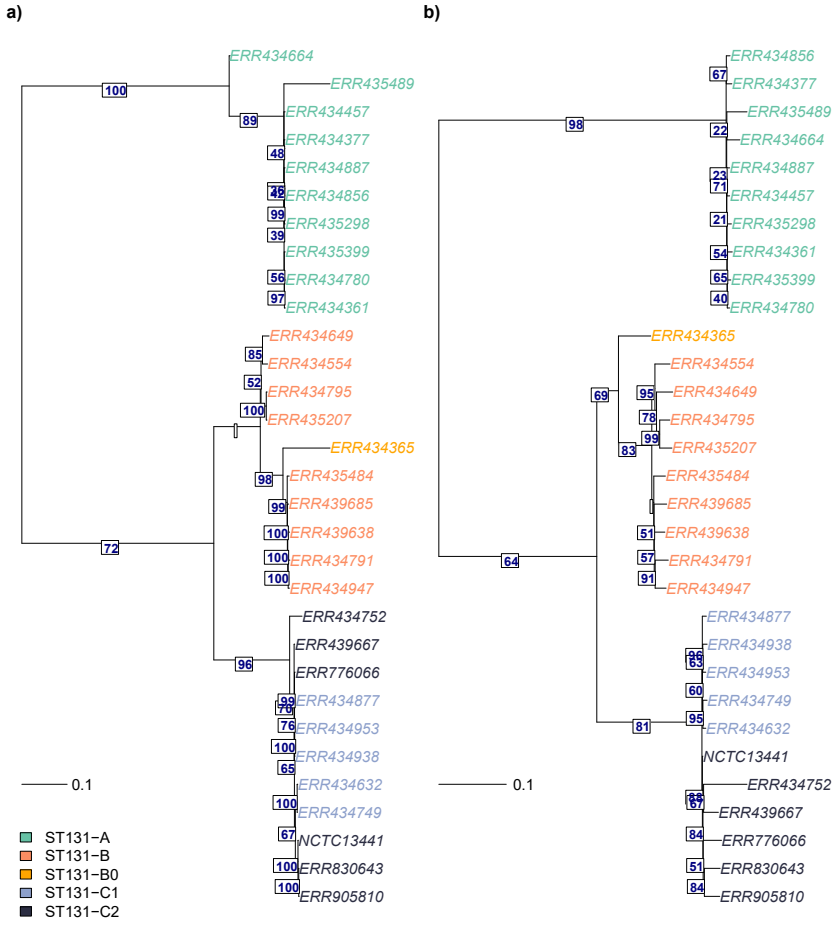

Figure 3.6: Midpoint-rooted maximum likelihood trees from core SNP alignment of *E. coli* ST131 strains. The phylogeny in panel **a)** was constructed from isolate sequencing data from 30 E. coli ST131 strains, and the phylogeny in panel **b)** with the mGEMS pipeline from ten synthetic plate sweep samples, each mixing three isolate samples from the ST131 sublineages (A, B, B0, C1, or C2). Boxed numbers below the edges indicate bootstrap support values from RAxML-NG for the next branch towards the leaves of the tree.

In the *E. faecalis* benchmark, the assessment was performed similarly to *E. coli* but with the change to investigating performance with between-ST variation. Additionally, *E. faecalis* is known to have a relatively high rate of recombination within the species across sequence types, particularly in nosocomially adapted lineages [110], which adds additional difficulty to the problem. Nevertheless, the mGEMS-derived assemblies do recover the overall structure of the phylogeny well and place sequences from the same ST to the same clade (Figure 3.7). Although the global structure is somewhat different from the isolate assembly phylogeny, even in phylogenies estimated from isolate sequencing data global differences are often explained by uncertainty arising from recombination affecting the placement of the STs within the overall phylogeny. This phenomenon is also apparent in the bootstrap support values for both the isolate and mGEMS-derived phylogenies, partially explaining the differences between the two.

The final synthetic benchmark in Publication II investigated phylogeny recovery within *S. aureus* ST22 with sublineages that are separated by a few dozen of SNPs [56], source study). The performance in this benchmark was not quite as good as in the *E. coli* and *E. faecalis* benchmarks but the mGEMS-derived results do still manage to replicate the important parts of the results transmission analysis wise. Namely, the samples that were determined as the likely source of the pathogenic strain in the sequenced patients (Figure 3.8) were placed at the root of the phylogeny when using both mGEMS and isolate data. Further down in the tree there is some lack of detail which is likely a result of the small degree of separation between the sublineages, or possibly of the filtering of the assemblies performed in the original study that was not replicated for the mGEMS-derived assemblies.

Figure 3.7 and legend source: adapted from Publication II, Figure 5 [28].

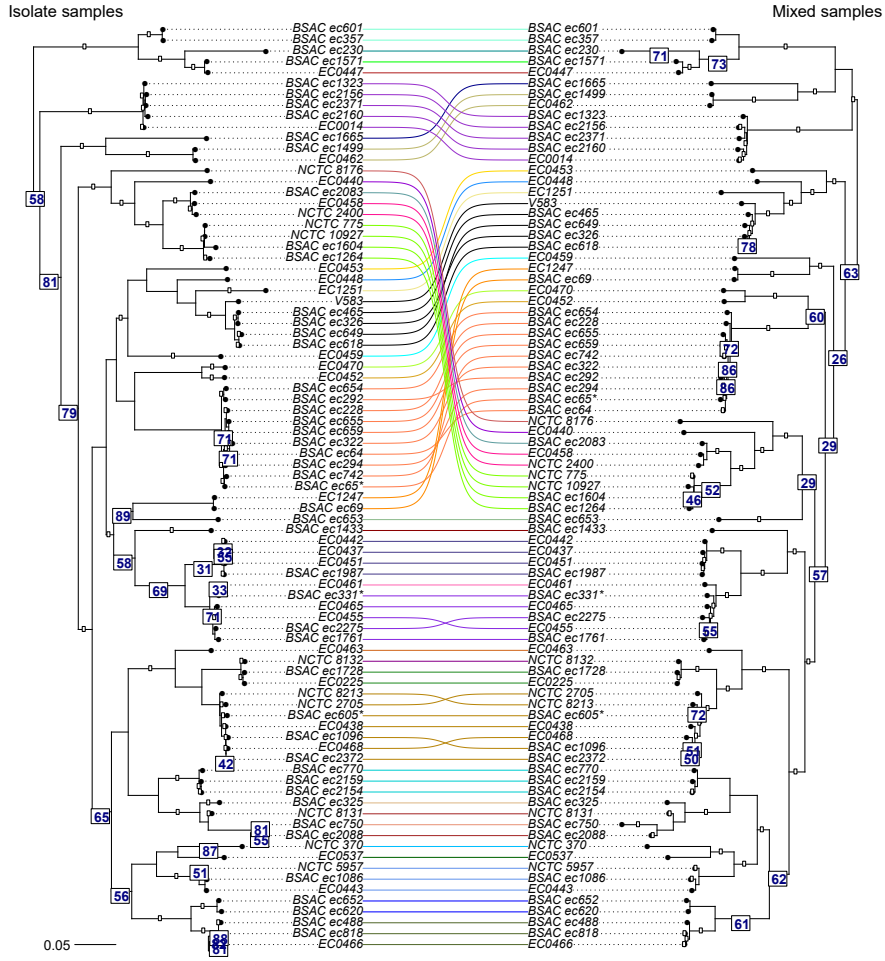

Figure 3.7: Tanglegram of two midpoint-rooted maximum likelihood trees from core SNP alignment of *E. faecalis* strains. The phylogenies were inferred with RAxML-NG. Numbers below the edges indicate bootstrap support values from RAxML-NG for the next branch towards the leaves of the tree. Only values under 90 are shown. Branches are coloured according to the *E. faecalis* STs.

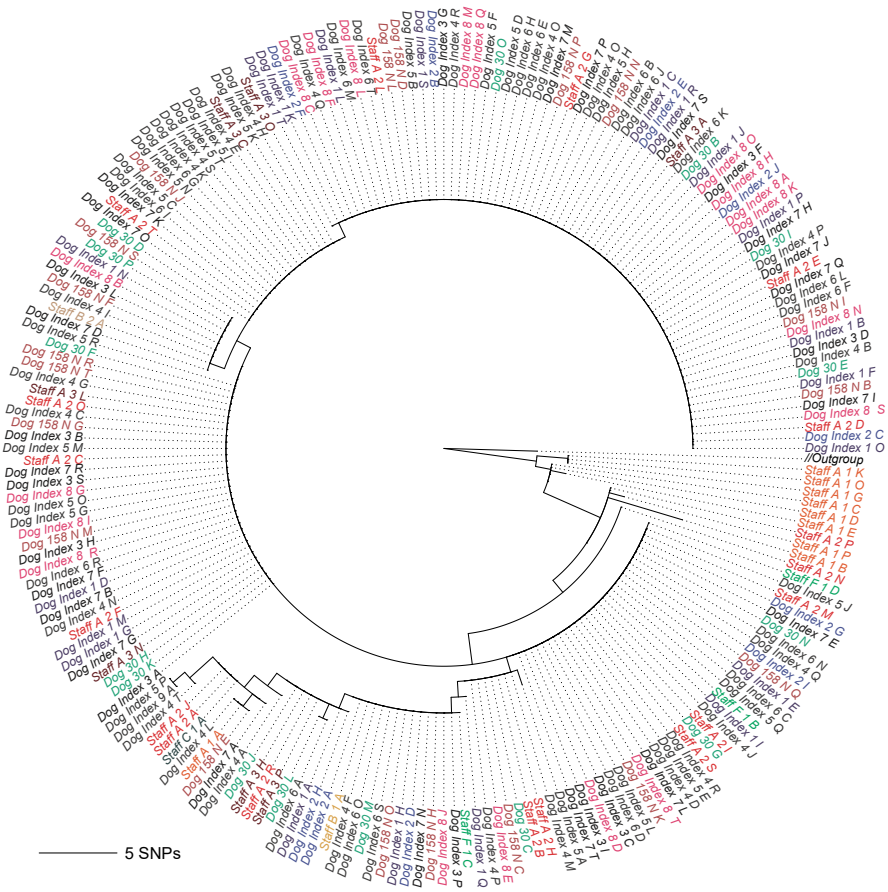

Figure 3.8: Midpoint-rooted maximum likelihood tree from core SNP alignment of *S. aureus* ST22 sublineage (clade 1). The phylogeny was inferred from a combined set of assemblies from 60 isolate sequencing samples (leaves labelled Staff A-G 1 A-T, corresponding to the temporally first samples from each staff member) and 312 mGEMS-derived assemblies from synthetic mixed samples containing sequencing data from each of the three different *S. aureus* ST22 sublineages (clades 1, 2, and 3).

Figure 3.8 and legend source: adapted from Publication II, Figure 6 [28].

In vitro benchmark

Publication II also included an *in vitro* benchmark data already mentioned earlier, where known amounts of DNA from three different strains were mixed together using Qbit [28], and both the mixed sample and the corresponding isolate cultures were sequenced. This data was used to re-test the performance of both mSWEEP and mGEMS.

The mGEMS part of the test examined the recovery of SNPs from either the isolate sequencing data or the Qbit-mixed sequencing data. In both the *E. coli* and *E. faecalis* benchmark samples the SNPs recovered from the mGEMS data reflect the values from the isolate data quite closely (Figure 3.9). There is some difficulty in separating the ST131-C2 sublineages 4 and 6, however. Similar results are obtained when mSWEEP is applied, with the *E. coli* benchmark being more challenging than the *E. faecalis* benchmark and the main difficulty being in distinguishing between the *E. coli* ST131-C2-4 and ST131-C2-6 sublineages. Nevertheless, mSWEEP manages to identify the presence of both clades quite well.

The accuracy hit in separating the ST131 sublineages is likely explained by their construction. While the primary sublineages (A, B, B0, C1, or C2, established in [82]) are defined using the core genome, the further sublineages within sublineages (ST131-C2-2, ST131-C2-6 and so on) incorporate information from the accessory genome, which is by definition much more variable than the stable core genome. This leads to a situation where, while incorporating the accessory information does further separate the established ST131-C2 sublineage, the resulting split is possibly only valid for a certain collection of sequences with the same accessory contents, and does not necessarily extend to other sequences collected from a different environment or at a different time. However, since the sequence types, or sometimes their well-established sublineages such as in the case of *E. coli* ST131, are typically the taxonomic unit of interest in practical analyses, the observed difficulties in separating between the accessory genome based specific sublineages are likely not relevant in applications of the method.

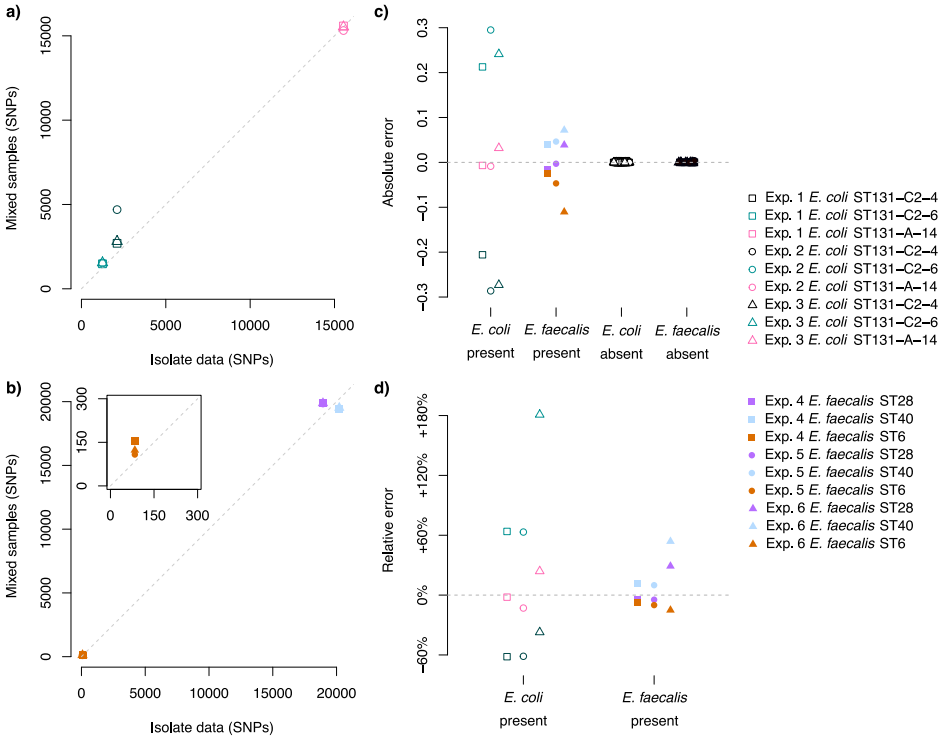

Figure 3.9: Evaluating mGEMS and mSWEEP on the in vitro benchmark data. Panels **a)** E. coli and **b)** E. faecalis compare the results of SNP calling from the isolate sequencing data (horizontal axis) against the results of SNP calling from the mixed samples with the mGEMS pipeline (vertical axis). The subplot in panel **b)** contains a zoomed-in view of the points around the origin. Panels **c)** and **d)** compare the abundance estimates from mSWEEP to the ground truth relative abundances. Panel **c)** shows the absolute difference between the estimates from mSWEEP and the true abundance. The values shown are split into E. coli and E. faecalis lineages truly present in the samples, and lineages truly absent. Panel **d)** shows the relative error in the truly present lineages.

Figure 3.9 and legend source: adapted from Publication II, Figure 2 [28].

Together, these benchmarks show that mSWEEP and mGEMS method can be reliably used to disentangle metagenomic sequencing data and create lineage-specific read bins. These bins can in turn be directly used in standard epidemiological analyses in place of isolate sequencing data and produce similar results. While mGEMS does not completely replace the use of isolate sequencing data in epidemiology — as the availability and continued production of high-quality reference sequences from isolate sequencing remains a critical part of the pipeline — the method shows promise in reducing the number of isolate cultures that need to be created when the isolates circulating in the samples are known. Additionally, applying mGEMS to whole community metagenomics sequencing data can yield results that previously available tools have not been able to produce as will be shown in more detail in Chapter 4.

3.2.3 Sequencing depth requirements

In addition to the benchmarks that assessed performance of mSWEEP and mGEMS in presence of complex strain variation, Publication I also investigated the impact of varying the sequencing depth (number of base pairs from a lineage in a sample divided by the average length of a genome from the same lineage) in the reads. In this benchmark, reads from 10 *E. coli* lineages were mixed synthetically at depths varying from 0.10x to 50x, and mSWEEP was applied to retrieve the relative abundances.

The results show that mSWEEP recovers the real values with admirable accuracy when the lineage in question was sequenced at high depths between 50x and 12.50x (Figure 3.10). Lineages mixed at intermediate depths between 6.25x and 1.56x were also recovered, but reducing the depths below 1x to values ranging between 0.78x and 0.10x ultimately overcame the detection accuracy of the method with none to very few of the lineages mixed at these depths identified at all. This limitation is, however, expected since the differences between the lineages are not large enough to accurately distinguish between them without sequencing reads from across the whole genome. These results imply that mSWEEP performs well up to low coverages between 1x and 2x, which translates to accurately recovering lineages with a relative abundance between 0.01 to 0.02 if the whole sample is sequenced at a depth that would correspond to 100x coverage in reads originating from a single source.

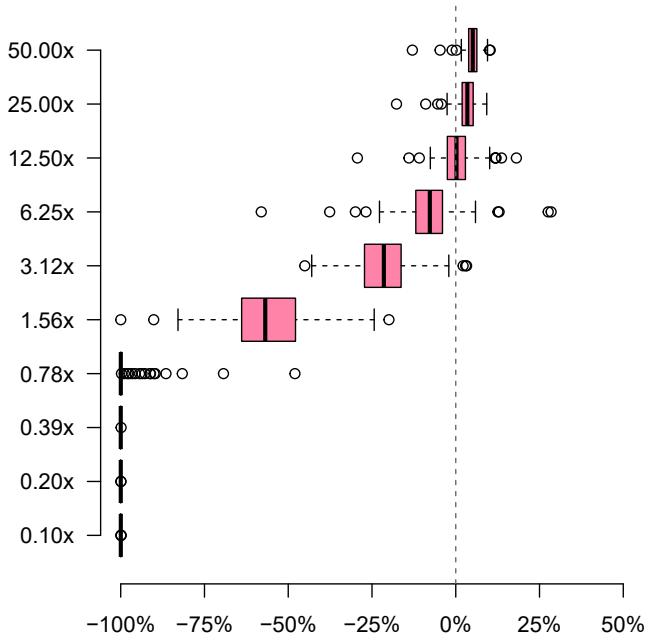

Figure 3.10: Relative error in 87 complex synthetic mixtures comprising 10 E. coli lineages at varying sequencing depths. The boxplot displays the relative error in the relative abundance estimates from mSWEEP compared against the true values. Error greater than 0% (horizontal axis) denotes estimates from mSWEEP exceeding the true value, while error less than 0% denotes estimates lower than the true value. The dashed gray line corresponds to 0% error. The rows (vertical axis) separate the estimates by their approximate sequencing depth, with each sample contributing one value (estimate for one lineage) to each row.

Figure 3.10 and legend source: adapted from Publication I, Figure 5 [17].

3.3 Assessing the detection accuracy

An important question when applying mSWEEP/mGEMS in practice is when the results from the pipeline are accurate enough. Inaccuracies may arise from for example a lack of reference sequences for a lineage that is present in the samples, which results in false positive estimates. Additionally, since the methods use a Bayesian approach to estimating the relative abundances of the reference lineages, none of the abundances will ever be exactly zero even though they contributed no sequencing reads in reality. These false positives (relative abundance estimates that differ from zero in some way that is deemed significant) may arise when the sequencing data contains DNA from a lineage that is not covered by the reference sequences. In this case, reads from the missing lineage are assigned to closely related covered lineages instead because the close relatives are the best plausible explanation for the presence of the reads in the sample. Although assignments produced in this scenario may still be useful for further analyses, this would require manual inspection of the results in order to determine which of the several lineages the reads were split between should be merged to obtain a set of reads corresponding to the missing lineage. In the scope of this thesis the falsely assigned reads are not considered for further analysis but their better handling is a potential avenue for future research on this topic.

These considerations naturally beg the question: "When is a lineage truly present in the sample?" when mSWEEP is applied in practice. Furthermore since the abundance estimates are the basis for the mGEMS pipeline, it is important to know whether the results for some lineage(s) with a high, nonzero relative abundance are truly correct or not. not. Some answers and means to investigate these questions will be provided in this section.

3.3.1 Detection thresholds for lineages

Answering the question about the presence/absence of a lineage in a sample requires defining some sort of threshold on the abundance estimates. Any estimate that falls below the threshold is then considered unreliable, and those that exceed it can be used in downstream analyses. One answer, called *detection thresholds*, is provided in Publication I. The approach pro-

vides a minimum abundance that the estimates must exceed and gives accompanying p-value-analogues that measure the degree of trust in values exceeding the threshold. Including detection thresholds in an analysis provides means for more theoretically sound consideration of the relative abundance results than using simple heuristics such as filtering by a minimum abundance.

The detection thresholds are created by taking a collection of reference sequences with corresponding short-read sequencing data available, and using a bootstrapping approach to determine the thresholds for each lineage. Within each lineage, a single reference sequence is removed from the reference, and a new set of sequencing reads is bootstrapped from the reads that were used to assemble the reference sequence. The bootstrapped reads are then put through mSWEEP to obtain a bootstrapped abundance estimate. Repeating the bootstrapping for several reference sequences, each removed in turn, within the same lineage yields an empirical distribution for estimates from the lineage when the true sequence is missing. The process should then be repeated for all lineages which have more than one reference sequence, and the values combined to obtain the empirical distributions for all of them.

The empirical distribution can be used to define the thresholds by taking an upper quantile of the distribution as a cutoff point. The cutoff defines the threshold, and any estimates that exceed it are considered reliable. These estimates come with an accompanying p-value, which depends on the number of estimates that exceed the chosen cutoff point. For lineages which have only one reference sequence, and naturally cannot be bootstrapped from in this manner, the cutoff point is defined as the maximum over all cutoffs that could be bootstrapped. The approach is described in more detail in Publication I.

Using bootstrapping to estimate the magnitude of the estimates in the absent lineages provides theoretically sound means to define the thresholds, and the accompanying p-values allow for some calibration in how trustworthy the results should be. The approach does, however have some computational issues, which limit the scaling of the method when dealing with large reference collections. Namely, the need to remove a single reference sequence at a time implies a need to reconstruct the pseudoalignment index every time the removal is performed. Dynamic indexing, which

refers to removing/adding sequences to the index without reconstructing the whole index, would solve this issue but remains an open question at the time of writing. However, the computational demands can be somewhat alleviated by employing a block-design based strategy to remove several sequences at the same time, which would lead to roughly similar results.

Unfortunately, the detection threshold approach has not found much use in the so-far published studies that use mSWEEP. The issue of defining trustworthy estimates has mostly been tackled by utilizing minimum abundance based thresholds, or by other means such as pseudocoverage, which will be defined in the next section. However, the detection thresholds could still be adopted more widely by including the thresholds as parts of prebuilt reference collections, or by providing more computationally scalable means to construct the thresholds, which earns them a mention here.

3.3.2 Pseudocoverage as a threshold

An alternative to constructing the detection thresholds can be found in using the number of aligned reads and the abundance estimates together to estimate the *pseudocoverage* of the reference lineage in the analysed reads. The pseudocoverage $c_k^\star$ for reference lineage k is defined as the product of the abundance estimate θ_k for lineage k and the total number of bases b in the reads that pseudoaligned against any reference sequence, divided by the average genome length l_k in lineage k

$$c_k^\star = \frac{\theta_k b}{l_k}. \tag{3.1}$$

Although the definition in Equation 3.1 is similar to the basic definition of average coverage (number of bases in aligned reads divided by genome length), these two definitions are not necessarily the same.

The difference between coverage and pseudocoverage can be elucidated by considering a set of sequencing reads originating from two bacterial strains, both with 100 base pair long genomes. Now, assume that these strains differ by a single base pair and are present in equal proportions in the sample providing the sequencing reads. Assuming 100 one base pair long reads from each position in both genomes are available, the average coverage of both strains would be 199 since $9900 \cdot 2$ reads will align to

both genomes and only $100 \cdot 2$ reads will align to just one. However, when considering the relative abundance estimates for these two strains, their true values are 0.50 and 0.50 — contrary to the fractions of reads that fit with each strain, 0.99 and 0.99 — since both strains were assumed present in equal proportions. In this setup, both strains would have a pseudocoverage of 99.5 which is a conservative value compared to the traditional definition of coverage.

Pseudocoverage can be used to define a threshold on the relative abundance estimates by finding the value of θ_k that provides a pseudocoverage of *at least* 1x, or another higher/lower value. Since the pseudocoverage is a conservative estimate of the true coverage (the exact, typically more complex than in the thought experiment above, relationship depending on the relatedness of the lineages k), this minimum value can be taken as a threshold on the relative abundances. If combined with means to investigate whether the lineages that remain after filtering by this approach are a good fit to the reference lineages, pseudocoverage provides an adjustable rule that is more easy to construct than the detection thresholds earlier.

3.3.3 Compatibility of the clustering and the reads

The remaining question regarding the reliability of the relative abundance estimates concerns the fit of the reference lineages with the estimated contents of the bins produced by mGEMS. This issue was not covered in neither Publication I nor Publication II but since the publication of the two, an external method has been developed to address the question. This method, demix_check[4], uses mash [111] to calculate distances between the read bins from mGEMS and the corresponding reference sequences in each lineage. These distances are used to evaluate the fit between the read bin and the lineage by comparing their distributions. In practice, demix_check has proved an integral part of the mSWEEP/mGEMS pipeline in evaluating the reliability of the results from mGEMS, and is also an integral part of the processing pipeline that was used in Publication III. Since this method was not developed by the author of this thesis, it won't be covered in more detail although its inclusion in mSWEEP/mGEMS analyses is strongly recommended.

[4] `https://github.com/harry-thorpe/demix_check`

Chapter 4

Metagenomic epidemiology

Genomic epidemiology refers to the use of sequencing data to identify pathogen transmissions chains and analyse the spread and diversity of the pathogen population [4–6]. These analyses are typically performed using isolate sequencing data, which is obtained by cultivating bacteria of interest on some selective media and isolating them for sequencing. The main reason for using isolate sequencing data is to produce reads with sufficiently deep sequencing depth for SNP calling, assembly, and other genomic analyses. Conversely in the spirit of genomic epidemiology, metagenomic epidemiology refers to performing the genomic epidemiology tasks but forgoing the culture step and using metagenomics data (whole community, plate sweep or other metagenomics approaches) to identify and analyse the pathogens with similar methods with particular attention paid to interactions within the microbiome [112, 113]. In the previous literature that could be called metagenomic epidemiology has typically been performed using genus- or species-level resolution tools like 16S rRNA sequencing. In this chapter, the term also encompasses the use of mSWEEP/mGEMS to perform the analyses at the lineage-level.

4.1 From genomic to metagenomic epidemiology

The previous chapter described the enabling effect of mSWEEP and mGEMS on high-resolution analysis of metagenomic sequencing data by extending the application of methods designed for isolate data to metage-

nomic sequencing reads. Especially when it comes to epidemiologically relevant analyses, such as SNP calling, assembly, and phylogenetic inference, the mGEMS-derived read bins perform nearly identically to isolate sequencing data. This means that most standard epidemiological analyses can be performed with metagenomic sequencing data by applying the mGEMS pipeline — provided that a sufficiently accurate reference database exists for the species of interest. Using metagenomic sequencing data in place of isolate sequencing data comes with the previously covered benefits related to cost-efficiency, vast expansion of throughput, and a more thorough coverage of the variation in a sample. Metagenomic epidemiology also enables performing various novel analyses by allowing lineage-level analysis of either the less biased data produced by whole community metagenomics, or detailed exploration of the diversity within some restricted set of taxons that can be enriched via the plate sweep metagenomics approach.

This chapter will briefly cover some of the metagenomic epidemiology results from Publications I-III as well as the advantages and disadvantages of incorporating metagenomic sequencing data into the analyses. Out of the three included, Publication I presented an experiment with real-world plate sweep metagenomics data, and Publication III provides an example of applying the mGEMS pipeline to whole community metagenomics data. While Publication II did not include a real-world application, the *in vitro* mixture samples analysed in Publication II do highlight some challenges for the methods that are relevant to this chapter. The chapter concludes with some speculation about the future applicability of the methods and the types of analyses that are possible with the introduction of mSWEEP and mGEMS.

4.1.1 Metagenomics-derived results

Epidemiological analyses performed on metagenomic sequencing data have the major advantage of covering the full spectrum of bacteria in a sample without the bias introduced by using cultivation steps. Because of this, results derived from metagenomic data should, in theory and with sufficiently deep sequencing, be capable of providing roughly the same results as non-metagenomics based approaches, although sequencing a pool of strains cannot resolve the lineage of the particular variants unless

long reads are used. Additionally, metagenomic approaches vastly extend the possibilities with regards to interactions between non-pathogenic and pathogenic species, and in tracking non-dominant strains of pathogenic species. When it comes to applying these methods in practice and interpreting the results, there are some obstacles standing in the way of rendering isolate sequencing studies completely redundant.

One of the immediate questions in analysing metagenomics-derived results in practice is what to do with the diversity that can be found in most samples. For practical use, most of the species (in clinical settings especially the commensal ones) are, firstly, not of any interest from the epidemiological point of view. Secondly, when dealing with microbiomes that are less extensively studied than the clinically relevant ones, many of the sequenced bacteria will not correspond to any previously sequenced lineages, species, or even genus. For reference-based approaches like mSWEEP and mGEMS this presents significant problems as the reads may only be analysed at the level where related reference sequences are available. Even for reference-free approaches, it can be difficult to place the results in a meaningful context if the samples contain significant amounts of unknown diversity. These factors imply that when talking about metagenomic epidemiology, the interpretation and analyses in practice might only focus on species that have already been studied using isolate sequencing.

4.1.2 Challenges

The major challenges in using metagenomic sequencing data again relate to the diversity found in the samples. Sometimes the (pathogenically) interesting species is encountered at very low abundances, resulting in a need to sequence the sample at much higher depth than what would be enough in an isolate study. This low abundance can also cause problems in identifying the presence of the species in the first place, as the number of reads that are *unique* to the species is even lower. Many of the other sequencing reads will be generated from commensal or even contaminant species which usually have no use in the downstream analyses. Additionally, whole community metagenomics sometimes results in an overabundance of host DNA in the sample [8, 9], which then dominates the contents of the reads from a sequencing run without host DNA depletion methods.

A second problem with metagenomics analyses relates to the lack of reference data from many domains of life that might be found in direct sequencing a sample. Even disregarding the presence of non-bacterial domains such as fungi, yeasts, and bacteriophages that are commonly found alongside the interesting bacteria, restricting the analyses to the bacterial domain still leaves a massive number of bacteria that have not been sequenced or studied before. Although the amount of "microbial dark matter" [114] is difficult to estimate, the discovery of completely new species [115] or even genera is not completely unheard of [116, 117]. This presents significant challenges for reference-based methods if the goal is to analyse the full diversity of the sample. Focusing on the more well-known bacteria does help in resolving the issue but when going further down to the lineage-level, it is almost a certainty to find new lineages of the species when the sequencing effort is sufficiently large simply due to the short timescale that bacterial evolution happens on.

The third issue is related to the lack of maturity in methods for analysing metagenomic data at the lineage-level, perhaps connected to the difficulties in solving the previous issues. Although mSWEEP and mGEMS provide tools for solving the problem, they are by no means alone capable of performing all analyses that might be of interest. In practice, this sometimes means that a human intervention in analysis pipelines might be required to identify cases that are problematic and to remove them from further consideration. This can be a surprisingly daunting, difficult, and especially time-consuming task.

4.1.3 Advantages

Although using whole community or plate sweep metagenomics in practice has its disadvantages, some major advantages in favour of the approaches do exist. The foremost of these is the ability to analyse the complete diversity with a single sequencing run since metagenomic sequencing produces vastly more data about the contents and composition of the sample than what would be obtained even with several different culture media and subsequent isolate sequencing runs. This information in turn enables making inferences about the coexistence and competition dynamics between different taxonomic units or possibly even co-transmission when considering epidemiological applications.

Another advantage of metagenomics-based analyses is their capability to increase the number of samples that can be processed since the plating steps may be entirely skipped. If obtaining high sequencing depths is not a priority, then the samples may simply be processed through the standard whole community metagenomics pipeline and disentangled computationally. For a higher depth, the plate sweep approach may be used to generate more reads from some interesting species. Regardless, both approaches significantly reduce the amount of laboratory work that is needed and allow processing of samples even in less-well resourced facilities.

4.2 Metagenomic epidemiology in practice

While the previous section covered the factors related to using metagenomics-derived results in practice in a more theoretical context, this section will focus on briefly summarizing the practical application of mSWEEP and mGEMS to both plate sweep and whole community metagenomics data in Publications I-III. The first subsection will cover results from the plate sweep approach that was initially employed, and required, by both mSWEEP and mGEMS. The second subsection shows results from applying the two methods to whole community metagenomics data, showing that the methods are not restricted to the plate sweep approach. The results are presented in more detail in their respective publications but a summary is provided here to elaborate on the potential applications of mSWEEP and mGEMS.

4.2.1 Plate sweep metagenomics

In Publication I, the method was applied to a set of *in vitro* plate sweep samples from children sequenced at a Vietnamese hospital. These samples were paired, with the first being taken before and the second after exposure to antibiotic treatment for diarrhea. The samples were plated on a media selecting for *E. coli* growth, and the whole plate was sweeped and the DNA sequenced in accordance to the plate sweep protocol presented in Publication I. The samples were analysed with mSWEEP, and the results used to investigate differences in *E. coli* sequence type contents and their relative abundances pre and post-treatment.

The results indicated a significant difference in the lineage composition between the paired samples, with more commensal lineages such as ST10 [29] being much more common in the pre-treatment samples. In the post-treatment samples the more invasive ST131 [29] had taken over and became the dominant lineages. No significant difference was detected in the composition of the samples (magnitudes of the relative abundance estimates), meaning that there was no significant difference in the number of samples that contained coexisting lineages pre- or post-treatment.

This analysis demonstrates simple means to analyse the lineage-contents of some sets of samples using only the relative abundance estimates. While it would be optimal to include the mGEMS pipeline steps (unavailable at the time Publication I was written), the results obtained using only abundance estimates are in line with the hypothesis/knowledge that commensal lineages may be replaced by antibiotic-resistance harbouring lineages when they are exposed to treatment. With more thorough follow-up sampling, the abundance estimates alone would be enough to identify when, or if ever, the lineage composition shifts back to the previous presumably stable composition in the pre-treatment cohort.

Focusing the analysis on the *E. coli* diversity only using the plate sweep approach provided high enough detail in the sequencing reads that exploration could even be performed at the level of identifying lineages within a sequence type. Contrary to performing whole community metagenomics on the same samples, the plate sweeps added to the sequencing depth in the reads and allowed for accurate identification. With the development of mGEMS, the analysis would become even more powerful with the possibility to separate the different strains (in cases that exhibited coexistence) and allow subsequent use of the strain-specific bins in downstream antibiotic resistance gene finders or phylogenetic analyses.

4.2.2 Whole community metagenomics

Publication III shows a set of results from applying mSWEEP and mGEMS to whole community metagenomics sequencing data. This data was obtained from another study [30] that investigated the differences in colonization of the newborn human gut using whole community metagenomics data from the first 21 days of life. In the original study, the analyses were performed only on the species-level. Using the same dataset demon-

strated that mSWEEP and mGEMS can be used to provide additional insights into the lineage-level dynamics present in the samples by focusing on the species that are known to be pathogenic and have extensive available reference collections.

The results from Publication III demonstrate one of the first insights into colonisation dynamics at the lineage-level in a virgin microbiome. The time-series data from the first 21 days showed a strong competition in the gut, with the first strain to colonize the gut often becoming the dominant strain and preventing others from displacing it. Additionally, in very few cases the samples contained several strains of the same species at the same time, providing more evidence in favour of the previous finding. Based on the results from mSWEEP/mGEMS, newborn babies are initially colonized by a single strain that typically persists or disappears in the gut for at least the first 21 days. Switches to another strain occurred very rarely within this time period but were more commonly observed in a single follow-up sampling somewhere between 4-12 months of age.

Coexistence at the lineage-level was rarely observed across all species analysed in the study (*Klebsiella* genus, *E. coli*, *E. faecalis*). Within the *Klebsiella* genus, which is composed of several related species, the results showed some synergistic relationships between the various *Klebsiella* species that were identified by mSWEEP and mGEMS. As a final analysis, the mGEMS-derived assemblies for the *Klebsiella* species were put through the Kleborate [118] pipeline to perform analyses of the resistance and virulence factors in them. Similar analysis was performed for the *E. faecalis* lineages using AMRFinderPlus [119]. These analyses and the distribution of the *E. coli* lineages somewhat surprisingly show no significant differences between the vaginally delivered and caesarean section delivered babies when it comes to the AMR gene contents.

All of these analyses required the use of mSWEEP and mGEMS, as no other methods exist for assembly-based high-resolution analyses of the strain content exist. The results together with the plate sweep results demonstrate the usefulness of having a method capable of targeted, high-resolution analysis of some parts of the microbiome. Additionally, since the results from the third study were derived from whole community metagenomics data, the study demosntrated the removal of a significant barrier in requiring performing plate sweep metagenomics, which has pre-

vented more widespread application of mSWEEP and mGEMS in the past. These results show that mSWEEP and mGEMS can be used alongside established tools for metagenome analyses when more information is desired about some particular subset of organisms within the samples.

Chapter 5

Conclusions and future directions

This thesis summarized the development and introduction of the mSWEEP and mGEMS methods for untangling lineage-level variation in metagenomic sequencing reads. Incorporating metagenomics data in genomic epidemiological studies was shown to enable novel insights into colonization dynamics and co-carriage of various species of bacteria that are capable of causing disease under the right conditions. These types of results that rely on sampling the full breadth of the bacterial species present in a host would not be possible with isolate sequencing data alone, demonstrating the need for methods like mSWEEP and mGEMS. Together, these two tools have the potential to enable entirely new types of analyses and broader exploration of the bacterial diversity by using metagenomic sequencing.

All experiments presented were performed using high-throughput short-read sequencing data, ignoring the more recent Oxford Nanopore and PacBio long-read sequencing technologies. Long-read sequencing can provide unprecedented detail into analysis of mobile genetic elements and difficult-to assemble regions of the genome, which cannot be accurately quantified using short reads. Long-read sequencing has been succesfully applied in metagenomics studies [48, 49], and the extension of mSWEEP and mGEMS to work on these technologies would be an important direction for future research.

Another promising area for future development is the use of either short or long-read sequencing data in combination with mSWEEP/mGEMS to

investigate plasmids. Plasmids are mobile genetic elements carried by bacteria, and they sometimes carry virulence factors or resistance genes [106, 120, 121], or otherwise help opportunistic pathogen species adapt to hospital environments [122]. Especially in short-read sequencing, current methods are often not able to differentiate between plasmid and chromosome derived reads. By constructing adequate reference databases for plasmids, mSWEEP and mGEMS could be applied to extract plasmid-derived reads from a set of reads by treating it as a metagenomic sample composed of the chromosomal and plasmid-derived parts.

The third point relates to combining plate sweeps and whole community metagenomics. Using whole community metagenomics alone has some issues in detecting low abundance organisms of clinical importance, such as *K. pneumoniae* [123, 124], but its inclusion can provide useful information about other species present in the samples. When analysing whole community metagenomics data, plate sweeps could be incorporated into the pipelines to enrich for species of clinical interest that cannot be adequately identified without extremely deep direct sequencing of the sample. Since mSWEEP and mGEMS can analyse both types of data, they could be used to screen metagenomics datasets for samples that should be enriched for interesting species.

The introduction of mSWEEP and mGEMS facilitates these types of analyses and other research directions. Consideration of lineage-level variation in epidemiological studies is likely to advance the field of genomic epidemiology beyond the isolate era. With the rapid production of high quality reference genomes from long-read sequencing studies, the applications of the methods can likely be extended beyond the species that were analysed in this thesis, and insights into the already possible species expanded through the inclusion of high-resolution metagenomics.

References

[1] G. L. Armstrong *et al.*, "Pathogen genomics in public health," *New England Journal of Medicine*, vol. 381, no. 26, pp. 2569–2580, 2019.

[2] K. A. Wetterstrand. "DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP)." (2021), [Online]. Available: `https://www.genome.gov/sequencingcostsdata` (visited on 05/21/2022).

[3] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: Ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, no. 6, pp. 333–351, 2016.

[4] P. Tang, M. A. Croxen, M. R. Hasan, W. W. Hsiao, and L. M. Hoang, "Infection control in the new age of genomic epidemiology," *American Journal of Infection Control*, vol. 45, no. 2, pp. 170–179, 2017.

[5] Y. H. Grad and M. Lipsitch, "Epidemiologic data and pathogen genome sequences: A powerful synergy for public health," *Genome Biology*, vol. 15, no. 11, pp. 1–14, 2014.

[6] J. C. Kwong, N. McCallum, V. Sintchenko, and B. P. Howden, "Whole genome sequencing in clinical and public health microbiology," *Pathology*, vol. 47, no. 3, pp. 199–210, 2015.

[7] J. W. Rossen, A. W. Friedrich, J. Moran-Gilad, *et al.*, "Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology," *Clinical Microbiology and Infection*, vol. 24, no. 4, pp. 355–360, 2018.

[8] J. Pereira-Marques *et al.*, "Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis," *Frontiers in Microbiology*, p. 1277, 2019.

[9] A. J. McArdle and M. Kaforou, "Sensitivity of shotgun metagenomics to host DNA: Abundance estimates depend on bioinformatic tools and contamination is the main issue," *Access Microbiology*, vol. 2, no. 4, 2020.

[10] S. J. Salter *et al.*, "Reagent and laboratory contamination can critically impact sequence-based microbiome analyses," *BMC Biology*, vol. 12, no. 1, pp. 1–12, 2014.

[11] V. Hill, C. Ruis, S. Bajaj, O. G. Pybus, and M. U. Kraemer, "Progress and challenges in virus genomic epidemiology," *Trends in Parasitology*, vol. 37, no. 12, pp. 1038–1049, 2021.

[12] D. W. Eyre *et al.*, "A pilot study of rapid benchtop sequencing of staphylococcus aureus and *Clostridium difficile* for outbreak detection and surveillance," *BMJ Open*, vol. 2, no. 3, e001124, 2012.

[13] J. L. Gardy and N. J. Loman, "Towards a genomics-informed, real-time, global pathogen surveillance system," *Nature Reviews Genetics*, vol. 19, no. 1, pp. 9–20, 2018.

[14] D. M. Aanensen *et al.*, "Whole-genome sequencing for routine pathogen surveillance in public health: A population snapshot of invasive staphylococcus aureus in Europe," *mBio*, vol. 7, no. 3, e00444–16, 2016.

[15] R. A. Weingarten *et al.*, "Genomic analysis of hospital plumbing reveals diverse reservoir of bacterial plasmids conferring carbapenem resistance," *mBio*, vol. 9, no. 1, e02011–17, 2018.

[16] C. E. Coipan *et al.*, "Genomic epidemiology of emerging ESBL-producing *Salmonella kentucky* bla CTX-M-14b in Europe," *Emerging Microbes & Infections*, vol. 9, no. 1, pp. 2124–2135, 2020.

[17] T. Mäklin *et al.*, "High-resolution sweep metagenomics using fast probabilistic inference," *Wellcome Open Research*, vol. 5, p. 14, Oct. 8, 2021, ISSN: 2398-502X.

[18] D. Cocker *et al.*, "Drivers of resistance in Uganda and Malawi (DRUM): A protocol for the evaluation of one-health drivers of extended spectrum beta lactamase (ESBL) resistance in low-middle income countries (LMICs)," *Wellcome Open Research*, vol. 7, p. 55, Feb. 15, 2022, ISSN: 2398-502X.

[19] J.-C. Lagier, S. Edouard, I. Pagnier, O. Mediannikov, M. Drancourt, and D. Raoult, "Current and past strategies for bacterial culture in clinical microbiology," *Clinical Microbiology Reviews*, vol. 28, no. 1, pp. 208–236, 2015.

[20] F. J. Whelan *et al.*, "Culture-enriched metagenomic sequencing enables in-depth profiling of the cystic fibrosis lung microbiota," *Nature Microbiology*, vol. 5, no. 2, pp. 379–390, 2020.

[21] G. W. Tyson *et al.*, "Community structure and metabolism through reconstruction of microbial genomes from the environment," *Nature*, vol. 428, no. 6978, pp. 37–43, 2004.

[22] J. C. Venter *et al.*, "Environmental genome shotgun sequencing of the Sargasso Sea," *Science*, vol. 304, no. 5667, pp. 66–74, 2004.

[23] F. P. Breitwieser, J. Lu, and S. L. Salzberg, "A review of methods and databases for metagenomic classification and assembly," *Briefings in Bioinformatics*, vol. 20, no. 4, pp. 1125–1136, 2019.

[24] A. Sczyrba *et al.*, "Critical assessment of metagenome interpretation—a benchmark of metagenomics software," *Nature Methods*, vol. 14, no. 11, pp. 1063–1071, 2017.

[25] D. T. Truong, A. Tett, E. Pasolli, C. Huttenhower, and N. Segata, "Microbial strain-level population structure and genetic diversity from metagenomes," *Genome Research*, vol. 27, no. 4, pp. 626–638, 2017.

[26] L. R. van Dijk *et al.*, "StrainGE: A toolkit to track and characterize low-abundance strains in complex microbial communities," *Genome Biology*, vol. 23, no. 1, pp. 1–27, 2022.

[27] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-Seq quantification," *Nature Biotechnology*, vol. 34, no. 5, pp. 525–527, 2016.

[28] T. Mäklin *et al.*, "Bacterial genomic epidemiology with mixed samples," *Microbial Genomics*, vol. 7, no. 11, Nov. 16, 2021, ISSN: 2057-5858.

[29] T. Mäklin *et al.*, "Strong pathogen competition in neonatal gut colonisation," bioRxiv, preprint, Oct. 7, 2022.

[30] Y. Shao *et al.*, "Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth," *Nature*, vol. 574, no. 7776, pp. 117–121, 2019.

[31] E. R. Sanders, "Aseptic laboratory techniques: Plating methods," *Journal of Visualized Experiments*, no. 63, e3064, 2012.

[32] N. L. Bachmann, R. J. Rockett, V. J. Timms, and V. Sintchenko, "Advances in clinical sample preparation for identification and characterization of bacterial pathogens using metagenomics," *Frontiers in Public Health*, vol. 6, pp. 363–363, 2018.

[33] P. Ghensi *et al.*, "Strong oral plaque microbiome signatures for dental implant diseases identified by strain-resolution metagenomics," *npj Biofilms and Microbiomes*, vol. 6, no. 1, pp. 1–12, 2020.

[34] D. Bertrand *et al.*, "Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes," *Nature Biotechnology*, vol. 37, no. 8, pp. 937–944, 2019.

[35] D. Danko *et al.*, "A global metagenomic map of urban microbiomes and antimicrobial resistance," *Cell*, vol. 184, no. 13, pp. 3376–3393, 2021.

[36] J. Vollmers, S. Wiegand, and A.-K. Kaster, "Comparing and evaluating metagenome assembly tools from a microbiologist's perspective-not only size matters!" *PLOS One*, vol. 12, no. 1, 2017.

[37] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata, "Shotgun metagenomics, from sampling to analysis," *Nature Biotechnology*, vol. 35, no. 9, pp. 833–844, 2017.

[38] M. I. Ivy *et al.*, "Direct detection and identification of prosthetic joint infection pathogens in synovial fluid by metagenomic shotgun sequencing," *Journal of Clinical Microbiology*, vol. 56, no. 9, e00402–18, 2018.

[39] W. Gu, S. Miller, and C. Y. Chiu, "Clinical metagenomic next-generation sequencing for pathogen detection," *Annual Review of Pathology*, vol. 14, p. 319, 2019.

[40] G. Tonkin-Hill *et al.*, "Pneumococcal within-host diversity during colonisation, transmission and treatment," bioRxiv, preprint, Feb. 21, 2022.

[41] Z. Zhang, G. Zhang, and F. Ju, "Using culture-enriched phenotypic metagenomics for targeted high-throughput monitoring of clinically-important fraction of beta-lactam resistome," bioRxiv, preprint, Jun. 4, 2022.

[42] Miansari66. "Freshly sprouted Qasuri Methi, CC BY 3.0 https://creativecommons.org/licenses/by/3.0." (Feb. 23, 2011), [Online]. Available: `https://commons.wikimedia.org/wiki/File:Fresly_sprouted_Qasuri_Methi.JPG` (visited on 06/21/2022).

[43] National Institute of Allergy and Infectious Diseases, National Institutes of Health. "*E. coli* bacteria, CC BY-NC 2.0 https://creativecommons.org/licenses/by-nc/2.0/." (Feb. 2, 2016), [Online]. Available: `https://www.flickr.com/photos/nihgov/24661308922` (visited on 04/04/2022).

[44] E. P. on Biological Hazards (EFSA BIOHAZ Panel) *et al.*, "Whole genome sequencing and metagenomics for outbreak investigation, source attribution and risk assessment of food-borne microorganisms," *EFSA Journal*, vol. 17, no. 12, e05898, 2019.

[45] A. B. McIntyre *et al.*, "Comprehensive benchmarking and ensemble approaches for metagenomic classifiers," *Genome Biology*, vol. 18, no. 1, pp. 1–19, 2017.

[46] C. Yang *et al.*, "A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 6301–6314, 2021.

[47] F. Meyer *et al.*, "Critical assessment of metagenome interpretation: The second round of challenges," *Nature Methods*, vol. 19, no. 4, pp. 429–440, 2022.

[48] V. Somerville *et al.*, "Long-read based *de novo* assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system," *BMC Microbiology*, vol. 19, no. 1, pp. 1–18, 2019.

[49] R. D. Stewart, M. D. Auffret, A. Warr, A. W. Walker, R. Roehe, and M. Watson, "Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery," *Nature Biotechnology*, vol. 37, no. 8, pp. 953–961, 2019.

[50] F. Beghini *et al.*, "Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3," *eLife*, vol. 10, e65088, 2021.

[51] S. Nayfach, B. Rodriguez-Mueller, N. Garud, and K. S. Pollard, "An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography," *Genome Research*, vol. 26, no. 11, pp. 1612–1625, 2016.

[52] S. Hiraoka, C. Yang, and W. Iwasaki, "Metagenomics and bioinformatics in microbial ecology: Current status and beyond," *Microbes and Environments*, ME16024, 2016.

[53] T. Thomas, J. Gilbert, and F. Meyer, "Metagenomics-a guide from sampling to data analysis," *Microbial Informatics and Experimentation*, vol. 2, no. 1, pp. 1–12, 2012.

[54] A. Sankar *et al.*, "Bayesian identification of bacterial strains from sequencing data," *Microbial Genomics*, vol. 2, no. 8, 2016.

[55] N. Van Goethem *et al.*, "Status and potential of bacterial genomics for public health practice: A scoping review," *Implementation Science*, vol. 14, no. 1, pp. 1–16, 2019.

[56] G. K. Paterson *et al.*, "Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission," *Nature Communications*, vol. 6, no. 1, pp. 1–10, 2015.

[57] S. Zlitni *et al.*, "Strain-resolved microbiome sequencing reveals mobile elements that drive bacterial competition on a clinical timescale," *Genome Medicine*, vol. 12, no. 1, pp. 1–17, 2020.

[58] O. V. Dixit, C. L. O'Brien, P. Pavli, and D. M. Gordon, "Within-host evolution versus immigration as a determinant of *Escherichia coli* diversity in the human gastrointestinal tract," *Environmental Microbiology*, vol. 20, no. 3, pp. 993–1001, 2018.

[59] M. Mosavie *et al.*, "Sampling and diversity of *Escherichia coli* from the enteric microbiota in patients with *Escherichia coli* bacteraemia," *BMC Research Notes*, vol. 12, no. 1, pp. 1–5, 2019.

[60] [SW] T. Mäklin and A. Honkela, *mSWEEP* version v1.5.2, Nov. 20, 2021, Zenodo, DOI: `10.5281/zenodo.5715877`,

[61] [SW] T. Mäklin, *mGEMS* version v1.2.0, Nov. 20, 2021, Zenodo, DOI: `10.5281/zenodo.5715888`,

[62] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-Seq," *Nature Methods*, vol. 8, no. 6, pp. 469–477, 2011.

[63] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: A revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[64] M. Rattray, X. Liu, G. Sanguinetti, M. Milo, and N. D. Lawrence, "Propagating uncertainty in microarray data analysis," *Briefings in Bioinformatics*, vol. 7, no. 1, pp. 37–47, 2006.

[65] X. Liu, K. K. Lin, B. Andersen, and M. Rattray, "Including probe-level uncertainty in model-based gene expression clustering," *BMC Bioinformatics*, vol. 8, no. 1, pp. 1–19, 2007.

[66] W. F. Doolittle and R. T. Papke, "Genomics and the bacterial species problem," *Genome Biology*, vol. 7, no. 9, pp. 1–7, 2006.

[67] T. Van Rossum, P. Ferretti, O. M. Maistrenko, and P. Bork, "Diversity within species: Interpreting strains in microbiomes," *Nature Reviews Microbiology*, vol. 18, no. 9, pp. 491–506, 2020.

[68] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge, "Analysis and design of RNA sequencing experiments for identifying isoform regulation," *Nature Methods*, vol. 7, no. 12, pp. 1009–1015, 2010.

[69] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey, "RNA-Seq gene expression estimation with read mapping uncertainty," *Bioinformatics*, vol. 26, no. 4, pp. 493–500, 2010.

[70] H. Jiang and W. H. Wong, "Statistical inferences for isoform expression in RNA-Seq," *Bioinformatics*, vol. 25, no. 8, pp. 1026–1032, 2009.

[71] X. Wang, Z. Wu, and X. Zhang, "Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-Seq," *Journal of Bioinformatics and Computational Biology*, vol. 8, no. supp01, pp. 177–192, 2010.

[72] P. Glaus, A. Honkela, and M. Rattray, "Identifying differentially expressed transcripts from RNA-Seq data with biological variation," *Bioinformatics*, vol. 28, no. 13, pp. 1721–1728, 2012.

[73] J. Hensman, P. Papastamoulis, P. Glaus, A. Honkela, and M. Rattray, "Fast and accurate approximate inference of transcript expression from RNA-Seq data," *Bioinformatics*, vol. 31, no. 24, pp. 3881–3889, 2015.

[74] L. Schaeffer, H. Pimentel, N. Bray, P. Melsted, and L. Pachter, "Pseudoalignment for metagenomic read assignment," *Bioinformatics*, vol. 33, no. 14, pp. 2082–2088, 2017.

[75] B. Gibson, D. J. Wilson, E. Feil, and A. Eyre-Walker, "The distribution of bacterial doubling times in the wild," *Proceedings of the Royal Society B*, vol. 285, no. 1880, p. 20 180 789, 2018.

[76] B. Gibson and A. Eyre-Walker, "Investigating evolutionary rate variation in bacteria," *Journal of Molecular Evolution*, vol. 87, no. 9, pp. 317–326, 2019.

[77] B. J. Arnold, I. Huang, W. P. Hanage, *et al.*, "Horizontal gene transfer and adaptive evolution in bacteria," *Nature Reviews Microbiology*, vol. 20, no. 4, pp. 206–218, 2022.

[78] M. C. Enright and B. G. Spratt, "Multilocus sequence typing," *Trends in Microbiology*, vol. 7, no. 12, pp. 482–487, 1999.

[79] J. A. Lees *et al.*, "Fast and flexible bacterial genomic epidemiology with PopPUNK," *Genome Research*, vol. 29, no. 2, pp. 304–316, 2019.

[80] L. Cheng, T. R. Connor, J. Sirén, D. M. Aanensen, and J. Corander, "Hierarchical and spatially explicit clustering of dna sequences with BAPS software," *Molecular Biology and Evolution*, vol. 30, no. 5, pp. 1224–1228, 2013.

[81] J. Corander and P. Marttinen, "Bayesian identification of admixture events using multilocus molecular markers," *Molecular Ecology*, vol. 15, no. 10, pp. 2833–2843, 2006.

[82] T. Kallonen *et al.*, "Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131," *Genome Research*, vol. 27, no. 8, pp. 1437–1449, 2017.

[83] S. Shaik *et al.*, "Comparative genomic analysis of globally dominant ST131 clone with other epidemiologically successful extraintestinal pathogenic *Escherichia coli* (ExPEC) lineages," *mBio*, vol. 8, no. 5, e01596–17, 2017.

[84] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, vol. 96, 1996, pp. 226–231.

[85] G. Tonkin-Hill *et al.*, "Producing polished prokaryotic pangenomes with the Panaroo pipeline," *Genome Biology*, vol. 21, no. 1, pp. 1–21, 2020.

[86] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, "Base-calling of automated sequencer traces using phred. I. Accuracy assessment," *Genome Research*, vol. 8, no. 3, pp. 175–185, 1998.

[87] B. Ewing and P. Green, "Base-calling of automated sequencer traces using phred. II. Error probabilities," *Genome Research*, vol. 8, no. 3, pp. 186–194, 1998.

[88] S. Berg, "Condorcet's jury theorem, dependency among jurors," *Social Choice and Welfare*, vol. 10, no. 1, pp. 87–95, 1993.

[89] E. Artin, *The Gamma Function*, trans. by M. Butler. Mineola, New York: Dover Publications, Inc., 1964.

[90] P. J. Davis, "Leonhard Euler's integral: A historical profile of the gamma function," *The American Mathematical Monthly*, vol. 66, no. 10, pp. 849–869, 1959.

[91] D. Griffiths, "Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease," *Biometrics*, vol. 29, no. 4, pp. 637–648, 1973.

[92] J. Močkus, "On Bayesian methods for seeking the extremum," in *Optimization Techniques IFIP Technical Conference: Novosibirsk, July 1–7, 1974*, G. I. Marchuk, Ed. Springer Berlin Heidelberg, 1975, pp. 400–404, ISBN: 978-3-662-38527-2.

[93] J. Hensman, M. Rattray, and N. Lawrence, "Fast variational inference in the conjugate exponential family," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[94] A. Honkela, T. Raiko, M. Kuusela, M. Tornio, and J. Karhunen, "Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes," *The Journal of Machine Learning Research*, vol. 11, pp. 3235–3268, 2010.

[95] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[96] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Research*, vol. 35, no. suppl_1, pp. D61–D65, 2007.

[97] A. Prjibelski, D. Antipov, D. Meleshko, A. Lapidus, and A. Korobeynikov, "Using SPAdes de novo assembler," *Current Protocols in Bioinformatics*, vol. 70, no. 1, e102, 2020.

[98] Y. Peng, H. C. Leung, S.-M. Yiu, and F. Y. Chin, "IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth," *Bioinformatics*, vol. 28, no. 11, pp. 1420–1428, 2012.

[99] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, "MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph," *Bioinformatics*, vol. 31, no. 10, pp. 1674–1676, 2015.

[100] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, "metaSPAdes: A new versatile metagenomic assembler," *Genome Research*, vol. 27, no. 5, pp. 824–834, 2017.

[101] D. D. Kang *et al.*, "MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies," *PeerJ*, vol. 7, e7359, 2019.

[102] Y.-W. Wu, B. A. Simmons, and S. W. Singer, "MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets," *Bioinformatics*, vol. 32, no. 4, pp. 605–607, 2016.

[103] C. M. Sieber *et al.*, "Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy," *Nature Microbiology*, vol. 3, no. 7, pp. 836–843, 2018.

[104] J. Hadar and W. R. Russell, "Rules for ordering uncertain prospects," *The American Economic Review*, vol. 59, no. 1, pp. 25–34, 1969.

[105] V. S. Bawa, "Optimal rules for ordering uncertain prospects," *Journal of Financial Economics*, vol. 2, no. 1, pp. 95–121, 1975.

[106] K. L. Wyres and K. E. Holt, "*Klebsiella pneumoniae* population genomics and antimicrobial-resistant clones," *Trends in Microbiology*, vol. 24, no. 12, pp. 944–956, 2016.

[107] H. J. Brodrick *et al.*, "Longitudinal genomic surveillance of multidrug-resistant *Escherichia coli* carriage in a long-term care facility in the United Kingdom," *Genome Medicine*, vol. 9, no. 1, pp. 1–11, 2017.

[108] K. E. Raven *et al.*, "Genome-based characterization of hospital-adapted *Enterococcus faecalis* lineages," *Nature Microbiology*, vol. 1, no. 3, pp. 1–7, 2016.

[109] A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, and A. Stamatakis, "RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference," *Bioinformatics*, vol. 35, no. 21, pp. 4453–4455, 2019.

[110] A. K. Pöntinen *et al.*, "Apparent nosocomial adaptation of *Enterococcus faecalis* predates the modern hospital era," *Nature Communications*, vol. 12, no. 1, pp. 1–13, 2021.

[111] B. D. Ondov *et al.*, "Mash: Fast genome and metagenome distance estimation using MinHash," *Genome Biology*, vol. 17, no. 1, pp. 1–14, 2016.

[112] S. S. Francis and L. W. Riley, "Metagenomic epidemiology: A new frontier," *Journal of Epidemiology & Community Health*, vol. 69, no. 4, pp. 306–308, 2015.

[113] F. Baquero, "Metagenomic epidemiology: A public health need for the control of antimicrobial resistance," *Clinical Microbiology and Infection*, vol. 18, pp. 67–73, 2012.

[114] C. Rinke *et al.*, "Insights into the phylogeny and coding potential of microbial dark matter," *Nature*, vol. 499, no. 7459, pp. 431–437, 2013.

[115] H. A. Thorpe *et al.*, "One health or three? transmission modelling of *Klebsiella* isolates reveals ecological barriers to transmission between humans, animals and the environment," bioRxiv, preprint, Aug. 11, 2021.

[116] O. V. Conle, F. H. Hennemann, Y. Bellanger, P. Lelong, T. Jourdan, and P. Valero, "Studies on neotropical Phasmatodea XX: A new genus and 16 new species from French Guiana," *Zootaxa*, vol. 4814, no. 1, pp. 1–136, 2020.

[117] A. Pitt, J. Schmidt, U. Koll, and M. W. Hahn, "*Aquirufa antheringensis* gen. nov., sp. nov. and *aquirufa nivalisilvae* sp. nov., representing a new genus of widespread freshwater bacteria," *International Journal of Systematic and Evolutionary Microbiology*, vol. 69, no. 9, pp. 2739–2749, 2019.

[118] M. Lam, R. R. Wick, S. C. Watts, L. T. Cerdeira, K. L. Wyres, and K. E. Holt, "A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex," *Nature Communications*, vol. 12, no. 1, pp. 1–16, 2021.

[119] M. Feldgarden *et al.*, "AMRFinderPlus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence," *Scientific Reports*, vol. 11, no. 1, pp. 1–9, 2021.

[120] E. Denamur, O. Clermont, S. Bonacorsi, and D. Gordon, "The population genetics of pathogenic *Escherichia coli*," *Nature Reviews Microbiology*, vol. 19, no. 1, pp. 37–54, 2021.

[121] K. L. Palmer, V. N. Kos, and M. S. Gilmore, "Horizontal gene transfer and the genomics of enterococcal antibiotic resistance," *Current Opinion in Microbiology*, vol. 13, no. 5, pp. 632–639, 2010.

[122] S. Arredondo-Alonso *et al.*, "Plasmids shaped the recent emergence of the major nosocomial pathogen *Enterococcus faecium*," *mBio*, vol. 11, no. 1, e03284–19, 2020.

[123] C. L. Gorrie *et al.*, "Gastrointestinal carriage is a major reservoir of *Klebsiella pneumoniae* infection in intensive care patients," *Clinical Infectious Diseases*, vol. 65, no. 2, pp. 208–215, 2017.

[124] R. M. Martin *et al.*, "Molecular epidemiology of colonizing and infecting isolates of *Klebsiella pneumoniae*," *mSphere*, vol. 1, no. 5, e00261–16, 2016.