UNIVERSITY OF PADOVA

Department of Information engineering

Master degree in Computer engineering

Master degree Thesis

A Study on the Effects of Using Sampling for Metagenomic Comparison

Supervisor
prof. Cinzia Pizzi

Candidate
Giorgio Gallina

21 April 2023 — a.y. 2022–2023

Abstract

Nowadays, ecological sciences depend heavily on genetic studies. Among these, analysis of environmental genetic material — i.e., metagenomics — is becoming increasingly popular for inferring essential information about microbial life and its interaction with ecosystems. An interesting application of metagenomics in this field is metagenomic comparison, that is the assessment of biotic dissimilarity between microbial environments. Current technologies allow us to produce Terabytes of metagenomic data with little effort. Consequently, the analysis of datasets of such size requires a large amount of computational resources. This led to the development and application of several strategies of dimensionality reduction, which are now being exploited for metagenomic comparison too.

In this thesis, we analyse three different methods of reducing dimensionality to see what an impact they have in relation to reference-based methods. Our results show that a sketching on distinct k-mers, as implemented in the tool SimkaMin, have almost no impact on both abundance-based and presence-absence-based comparison for a sketching size larger than 10^5 distinct k-mers. On smaller sketches, quality of results decreases. On Spriss' sampling scheme, in which reads are selected uniformly at random with replacement, abundance-based Bray-Curtis dissimilarity showed no significant variations on moderated sampling rates — e.g., above 2 % — and a marked quality decline on lower sampling rates. When the k-mers used are too short, 12 bp for instance, this sampling scheme seems to improve drastically dissimilarity measures. On the presence-absence Jaccard distance, instead, Spriss' subsampling scheme improves the correlation between reference-based and compositional-based methods at moderate sampling rates. Lastly, comparison of approximate sets of frequent k-mers, as outputted by Spriss, hold lower correlation with reference-based dissimilarities, except on very short k-mers.

Overall, our study suggests that rare k-mers are of both types: weakly informative and noise. Their impact is imperceivable on abundance-based dissimilarity, whereas the noisy part of them affect negatively the quality of the Jaccard index, which benefits from a moderate subsampling indeed.

Contents

1 Introduction 5

1.1 Background 5

1.2 Purpose of the Thesis 7

1.3 Organization of the Work 8

2 Biodiversity Measures 9

2.1 Biodiversity 9

2.2 The Notion of distance 12

2.3 Diversity Indices 14

2.3.1 Notation 15

2.3.2 Classification of Diversity Indices 16

2.3.3 Compatibility of Diversity Estimates 18

2.4 Alpha Diversities 18

2.4.1 Species Richness 18

2.4.2 Species Evenness 21

2.4.3 Taxonomic Indices 24

2.4.4 Taking Scale Into Account 25

2.5 Beta Diversities 26

2.5.1 Taking Scale Into Account 27

2.5.2 Incidence-Based Indices 28

2.5.3 Abundance-Based Indices 30

2.5.4 Taxonomy-Based Indices 31

2.6 Gamma Diversities 32

2.6.1 Observations 32

3 Tools and Experimental Settings 35

3.1 Running Environment 35

3.2 Procedure Outline 36

3.3 True Dissimilarities Computation 37

3.4 Reference-based Metagenomic Comparison 37

3.4.1 Kraken 2 38

3.4.2 Bracken 39

3.4.3 Reference Database 40
3.4.4 Technical Issues 41
3.5 Simka and SimkaMin for k-mer-based Dissimilarities 41
3.5.1 Simka 43
3.5.2 SimkaMin 44
3.6 Reads Sampling with Spriss 44
3.6.1 Frequent k-mers Estimates 46
3.6.2 Running Simka on Spriss' Samples 46
3.7 Examples of k-mers and Reads Sampling Schemes Effects 46
3.8 Statistical Analysis 48

4 Analysis of Results 49
4.1 Datasets Description 49
4.1.1 The HMP dataset 49
4.1.2 The GOS dataset 50
4.1.3 The CAMI II Toy Human Metagenome Project dataset . . . 50
4.2 SimkaMin Subsampling Effects 53
4.2.1 Experiments on Bray-Curtis Dissimilarity 53
4.2.2 Experiments on the Jaccard Distance 55
4.3 Spriss Subsampling Effects 63
4.3.1 BC Dissimilarity on Spriss Subsample 64
4.3.2 BC Dissimilarity on Spriss Frequent k-mers Estimates . . . 75
4.3.3 Experiments on the Jaccard Distance 76

5 Conclusion 83

A Datasets Details 85

B Spriss Samples Details 91

C Supplementary Plots 93

Glossary 99

Abbreviations 104

Bibliography 106

Chapter 1

Introduction

Due to a global concern about detrimental environmental changes, which, already predicted, are now being witnessed, *ecology* is becoming an increasingly popular science. Besides, thank to rapid technological development, which has made fast and cheap genomic and *metagenomic* sequencing feasible, it is now possible to study ecology by means of metagenomics, which is the focus of our thesis.

1.1 Background

Ecology knowledge has numerous applications and it is deeply influential for our society. Indeed, it is essential in evolutionary biology and conservation biology, it expands our understanding of geology, meteorology and medicine, and drives plenty of our thoughts and actions, thence any real philosophical investigation. Here, we are interested in *microbial comunities*, which are ubiquitous in our biosphere and play vital roles in every ecosystem. Microbes, indeed, can remediate toxins in the environment — including oil and chemical spills resulting from human activity, — can transform CO_2 into organic carbon, digest food that their hosts cannot digest and provide them with essential nutrients, suppress some pathogens and regulate immune response and epigenetic expression of their hosts [1, 9, 36, 45]. It is crystal clear, thus, how useful it is for us to understand, control and manage microbial communities. For instance, human microbiota composition may discriminate between healthy and ill individuals, hence providing reliable diagnosis methods and, maybe even more importantly, a new landscape in disease comprehension and treatment. We refer the reader to [7] for a wider view on current applications, interests and future perspectives.

As far as microbial comunities are concerned, metagenomics comes to our help. It is common practice, indeed, to infer their microbial compositions from the genomes collected from their environments, which constitute the metagenomes of those communities. Thank to the genome–individual bijection generalised into a function from genome to species, or other Operational Taxonimc Units[1] (OTUs), it is possible to determine the microbial composition of an environment via taxonomic binning, provided all the genomes present in it are both detected and known. Unfortunately, this is never the case, but relevant information can be extrapolated from available data nonetheless. For instance, it is still possible to estimate statistics like the number of different species in a sample (species richness) and the distribution of species abundances. Many of such statistics are commonly referred to as *a-diversities*, which are presented in Section 2.4.

Another bioinformatic technique to infer useful ecological information is to *compare distinct metagenomic samples* between each others or with some reference samples. In other words, we try to assess biological dissimilarity of distinct samples, commonly referred to as *β-diversity*. Several approaches are used to carry out such comparisons, some of which are described in Section 2.5. In fact, this is the technique specifically addressed by this thesis.

Compared samples may have been collected from separate environments, or from one environment in different moments. Noticeable is that a “community of transcripts” (RNA molecules), a “community of proteins” or an “arrangement of biological functions” are addressable as well by these methods.[2] Therefore, this technique applies to multiple purposes. Among them:

1. Diagnosis: compare patient’s microbiomes against references;

2. Monitoring microbial comunities over time (of great impact for agriculture, food industry and conservation biology, but also for personalised medical treatment);

3. Evaluation of the health of an environment (similar to medical diagnosis);

4. Assessment of compatibility of results of different Next Generation Sequencing (NGS) technologies, which are often biased [7];

5. Appraisal of human activity impact on environment, especially useful for assuring safety of new products or services.

6. It might even corroborate dating in archaeology and evolutionary biology.

Moreover, in the current global scenario, applications of metagenomic comparison might reveal to be even more useful than one may believe. In fact, if we think of farming or food industry, not only it would provide an alternative quality control, but even an additional sanity check. This is because increasingly often new pathogens are emerging and they might elude traditional inspections.

[1] Operational definition used to classify groups of closely related individuals.

[2] However, for the sake of simplicity, we only target metagenomics.

We could even force our creativity and speculate on purely imaginative applications. Observing animals' behaviour, we notice that some of them feel in advance geological events like earthquakes. Therefore, in spite of the difficulty in guessing any evolutionary advantages for microbial comunities to predict an earthquake and reorganise itself accordingly, what if few of them were nonetheless affected in advance by such phenomena, like animals are? Furthermore, it is reasonable to think that weather conditions both influence and are influenced by microbial communities and it may be scientifically interesting to understand if and how they change with weather. Even more so, life has proved to actively influence global climate [3, section 1.2]; what if peculiar communities could help weather forecasting?

Lastly, despite having adduced a consistent amount of reasons supporting interest on this field, as more and more evidence of essentiality and impact of microbiomes is collected, we shall not forget the risk of overestimating both their importance and our understanding of their dynamics. Moreover, we reckon that human activity, especially its intense chemical and electromagnetic pollution, has been confusing myriads of living beings and so we ought to consider instability of methods based on microbial life.

1.2 Purpose of the Thesis

The classification of diversity measures, and β-diversities in particular, into reference-based and reference-free diversities is of great interest in our scope. In the first case, the species (or whatever OTUs or functions) detected in the sample are firstly determined via taxonomic binning and then considered in subsequent computation in place of reads. In the other case, instead, reads in the sample are directly used in the computation.

Ideally, reference-based approaches should be the most meaningful and reliable since the real target of comparison are "effective" species rather than genetic sequences. However, for genomes of most species are still unknown, reliability is doubtable.

Furthermore, a big issue in metagenomics is data dimensionality and, hence, computational time and resources requested for metagenomic analysis. Indeed, Terabytes of data can be easily produced, whereas their analysis is a bottleneck. Reference-free methods are usually faster as they do not require the identification of the source of each read, yet they benefit from dimensionality reduction nonetheless. However, while time efficiency would naturally improve, its impact on results is still unexplored. For example, subsampling would likely impact differently rare and abundant species, hence possibly biasing results.

With this thesis we try to enlighten such an impact from an empirical point of view. To do so, we compare β-diversity estimates obtained using k-mer-based methods on subsampled metagenomic data with both those obtained without subsampling and those calculated with reference-based approaches. The latter is carried out by classifying metagenomes with the Kraken 2 software and by estimating species abundance through the tool Bracken. For k-mer-based dissimilarity computation we rely on Simka, which also provides the first subsampling technique we study. Other two subsampling approaches come from the tool Spriss.

1.3 Organization of the Work

In Chapter 2 we firstly introduce some relevant vocabulary of ecology and *biodiversity*. Afterwards, we consider the mathematical notions of *diversity* and *dissimilarity* and we present how these tools are applied to evaluate biodiversity. Hence, *alpha*, *beta* and *gamma diversities* are explored in details, with particular focus on beta diversities, which are the core topic of this thesis.

Subsequently, experimental settings and the implementation of the analysis is described, with essential details about the tools we exploited.

Afterwards, in Chapter 4 we present our analysis of the results we collected on each dissimilarity measure, on each dataset, for each subsampling scheme. Lastly, a brief recapitulation of the results we observed concludes the thesis.

Chapter 2

Biodiversity Measures

Whenever an artisan manufactures some implement, he/she must take into account the necessities of his/her customers. The latter, on the flip side, interface with the craftsman to know how to properly use their purchase. Similarly it goes for bioinformatics computer engineers and biologists.

Throughout this chapter we build such an interface. Therefore, we shall report clearly what we intend by "diversity" in a metagenomic context, how we measure such diversities, warn about possible conflicting concepts and methods, and explain how results are shown. Before diving into metagenomic diversities, however, a brief summary on biodiversity is deserved. Indeed, that is the field in which we move and we ought to agree on the vocabulary we use.

2.1 Biodiversity[1]

Metagenomic comparison, *any possible diversity/dissimilarity measures based on metagenomic samples*, appears as a means for measuring life diversity. Therefore, we likely think of it as a gauge of biological diversity, i.e., *biodiversity*. However, while being a powerful tool to asses the latter — genomic diversity has been regarded as "fundamental currency" of biodiversity, actually, — we should remark that biodiversity has a broader meaning among scientists. In fact, besides living organisms and their complex interactions, biodiversity is usually defined so to include their relationships with abiotic (non-living) aspects of their environments as well. A formal definition of **biodiversity** is: *the variety of life on Earth at all its levels, from genes to ecosystems, and the ecological and evolutionary processes that sustain it.*

On a large scale, the distinction just highlighted is definitely relevant. Indeed, two communities equally composed can act rather differently in diverse ecosystems: just consider the behavioural discrepancy of the same individual in summer and

[1] We refer to [3, 7] as main sources of information for the content of this Section 2.1

winter. Moreover, both an oak tree and an acorn belong to the same species, yet their impact on their ecosystem is irrefutably different. Notwithstanding, coming to microbial comunities, their richness in species composition and their high mutation rate may result in sharp discriminants between distinct ecosystems; therefore, such communities might carry information about their relation with abiotic environment too. Nonetheless, by the same argument and remembering redundancy of microbial individuals, though improbable, divergent species composition and distribution might be detected from ecosystems that we would characterise as (instantaneously) very similar.

In ecology, many concepts that we have already been using widely in the text have precise meaning, which we must report to avoid confusion. Before that, of paramount importance is to reason in a hierarchical way: different concepts are arranged in hierarchies, meanwhile one term can be used at different levels with the same meaning but unlike effect.

Biological **taxonomy** is the first hierarchy we find. While there are no exhaustive definitions of species and there are several possible categorisation approaches[2], there is no vocabulary confusion and, moreover, we will stick to species level only. Thus, we avoid expanding this subject. We shall nonetheless recall that species are not the only possible fundamental block on which a taxonomy is built; in fact, species constitute just some of several possible **Operational Taxonomic Units** (OTUs), which are operational definitions used to cluster together groups of closely related individuals. More relevant is, instead, the relation between ecosystems, communities, populations, environments and so on, which we now describe with the help of Fig. 2.1.

Groups of cohesive individuals of a same species which share genetic or demographic aspects more closely with each other than with other individuals of the same species are called **populations**.[3] It is possible to define a population diversity[4], but we do not bother about that since metagenomic data are intrinsically unsuitable for the purpose. In spite of that, microbial populations detection has lately been proposed on the base of amount of identity in whole genome alignment as an indicator of recent horizontal gene transfer [4].

Several populations of different species naturally interacting in some environment form a **community**. For instance, the collection of species associated with ripening figs in a tropical forest is a community. This concept, however, is probably even more cumbersome than that of "species". A small community can be part of a wider one, comprising several communities and some populations might be assigned to

[2] Cladistic, phenetic and evolutionary taxonomies, for example. See [28, chapter 6]

[3] Hence, populations are clusters of individuals of a same species. Unfortunately, the same word "population" in used in statistics with a different meaning; we will try not to use the statistical meaning in order to avoid confusion.

[4] Diversity or dissimilarity. We clarify the difference between the two notions in Section 2.2; in this section we just use the first of the two terms for simplicity.

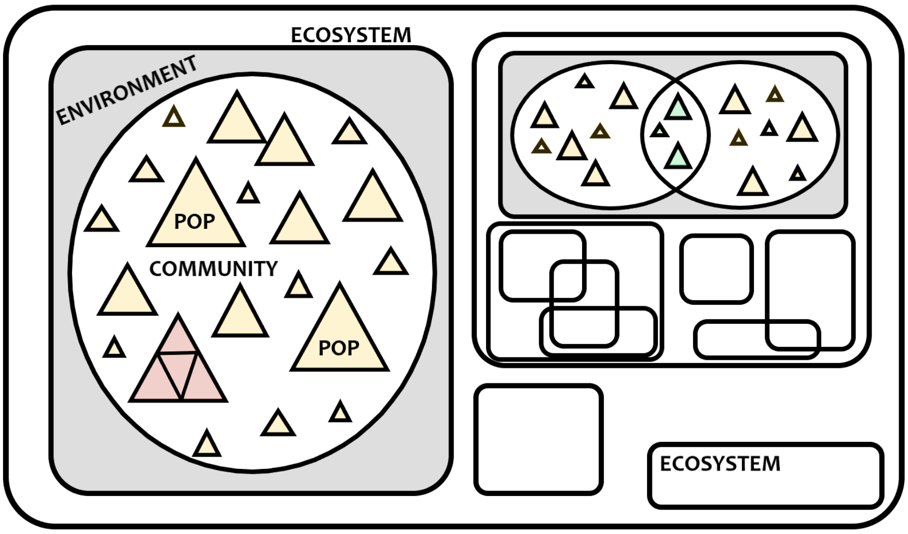


Figure 2.1: Representation of some relationships between populations (triangles), communities (circles), abiotic environment (grey shade) and ecosystems (rectangles).

different communities depending on the interest of the biologists who study them.

Taken *together, a community and its environment* form an **ecosystem**. Following the ambiguous definition of "community", we can find larger and larger ecosystems up to the entire biosphere. Ecosystem diversity takes into account species numerosity and distribution, their functions and relationships, but also physical characteristics of the environment.

As regards environment, **ecoregions** and **landscape** diversities are addressed too by biologists. These regions being dealing with large regions on earth, however, they are not of our interest.

Two key concepts in our scope are those of *microbiome* and *microbiota*. While biology being an intrinsically "imprecise" science in many of its definitions, experts from all over the globe have recently forgathered[5] to make order on the definition of these two terms. Formerly, they were used as synonyms by someone; others referred to microbiota as the effective microbial comunity whereas its collective genomes being a microbiome; and still other definitions have been used. Moreover, it was common to reserve the two buzzwords for those microbial comunities occurring in and on host multicellular organisms. As a result of the discussion, researchers have proposed to address *any community of living microorganisms present in a defined environment* (not necessarily a living host) as a **microbiota**. Viruses, phages, plasmids, prions and the similar are therefore not part of a microbiota. Instead, the

[5] March 2019, MicrobiomeSupport project workshop.

latter are part of a **microbiome**, which *comprises a microbiota and its biotic and abiotic "theatre of activity"*.

Consequently, **metagenomic diversity** is related to many of the diversities within and between the presented "entities", yet distinct from them all. As metagenomic samples contain large portions of *relic DNA* sequences, which are extracellular nucleic acids derived from dead cells, they usually embrace abiotic elements. On the other hand, they cannot contain every kind of entities present in an environment. Abiotic compounds aside, think that even DNA and RNA are sequenced separately, which implies bacteria and several viruses being sequenced using different technologies, which are not comparable in principle. Were these distinctive aspects of metagenomic diversity not enough, the availability of genetic material can vary enormously between individuals: therefore, genome abundance is not even proportional to species abundance [37].

For the sake of completeness, we shall mention that chemical methods for excluding extracellular DNA before sequencing have been proposed [27], hence allowing to estimate microbiota diversity. However, such methods are still not universally adopted. Moreover, despite relic DNA having shown to have little effect on estimates of taxonomic and phylogenetic diversity [20], in peculiar cases it could introduce "history" into measures of environmental diversities. If the latter consideration is definitely daring and rather useless on its own, special attention should be paid when exploiting metagenomic β-diversity for temporal studies and security assessment of novel technologies.

Having clarified relevant vocabulary and differences between metagenomic comparisons and other comparisons, we shall discuss briefly the notion of diversity itself.

2.2 The Notion of distance

Diversity is mathematically measured through the concept of **distance**, or **metrics**, which, on a set $\mathcal{M}$ is a function $d\colon \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ such that for any $x, y, z \in \mathcal{M}$ it holds:

$$d(x,y) = 0 \iff x = y \tag{2.1a}$$

$$d(x,y) = d(y,x) \tag{2.1b}$$

$$d(x,y) \leq d(x,z) + d(z,y) \tag{2.1c}$$

The couple $(\mathcal{M}, d)$ is called a **metric space**.[6].

A distance is, then, what allows us to measure diversity, provided we managed to define a sensible metric space. The key point is the codomain being $\mathbb{R}$, which is an

[6] Not to be confused with a measurable space. Notice that $d(x,y) \geq 0$ as a consequence of the definition.

ordered field, thus making *comparisons* feasible. Intuitively, the higher the distance $d(x, y)$ between two objects $x, y \in (\mathcal{M}, d)$ is, the more diverse or dissimilar they are.

Sometimes it is convenient to relax some of the constraints (2.1a) to (2.1c): a non negative function satisfying only Eqs. (2.1a) and (2.1b) is called a **semi-metrics**, whereas a function satisfying Eqs. (2.1b) and (2.1c) and such that $d(x, y) = 0 \Leftarrow x = y$ is a **pseudo-metrics**. The latter arises naturally when calculating the distance between elements of a set over which an *equivalence relation* maps those elements to their equivalence class in the *quotient set* over which a metric is defined. Notice that a pseudo-metrics captures well the meaning of geometrical similarity. As we will see, also semi-metrics are profitably used as diversity measures.

Some authors prefer to use the term *distance* only for strictly metric functions, and talk about **dissimilarity** as a general expression. While this distinction is quite reasonable, we find in literature unclear differentiations between *diversity* and *dissimilarity.* Often, the former is reserved to α-diversity, whereas the second one is used for β-diversity indices [2, 31], which, however, seems to us inconsistent with the name "β-diversity" itself. Elsewhere, we feel that diversity is used for unbounded functions while dissimilarity for bounded ones, where total dissimilarity is the higher bound and perfect similarity the lower one. Because of this confusion, we will just use these two terms interchangeably.

While similarity as just defined suite well *β-diversities*, evaluation of *a-diversities* requires different approaches. In that case, indeed, one's interest is to quantify the overall diversity present *within* one single sample. Therefore, **diversity indices** are used, which not always make use of distance functions. We discuss some of them in Section 2.4.

Consequence of the above definitions is that, given a very general universe $\mathcal{M}$, in order to measure difference between two objects in it, we would need to define the values of a distance d case by case. For this is impossible in an infinite universe, rather than assigning real numbers to each couple of objects, we define a *procedure* which enables us to determine such values in an automatised way. The physical measurement of distances by comparison with a unit of measure is probably the easiest of such procedures.

There exist, however, sets where such procedures are expressed by mathematical formulas. In the plane of real numbers $\mathbb{R}^2$ we *calculate* the Euclidean distance between two points (which are advantageous *representations* of two objects), provided their coordinates have been physically measured. Because we commonly work on metric spaces $(\mathcal{M}, d)$ where d is algorithmically defined, it is worth underlying that the difference between two objects is not only captured by some transformation of d but also by the *map* from "real objects" to $\mathcal{M}$.

As far as our work is concerned, parts of the map from environmental samples to some tractable space $\mathcal{M}$ are:

- Sequencing:

  - Everything that is not a (ribo-)nucleic acid is discarded;
  - Nucleic acids are sampled;
  - Their structure and epigenetic markers are ignored;
  - Integrity of the genomic sequences is lost;
  - Letters of the alphabet map signals detected from different nitrogenous bases;

- Quality filtering;
- Eventual further transformations of metagenomic data (e.g. k-mer decomposition);
- Eventual dimensionality reduction.

Thus, plenty of differences are disregarded from the beginning and would never be quantified by any measure we define. This shed some more light on the distinctions between comparisons remarked in the previous section. Finally, the last point in the list clarifies the goal of our thesis work: we empirically measure the impact of dimensionality reduction in the assessment of metagenomic diversity/dissimilarity.

2.3 Diversity Indices

Several *indices* of biological diversity are employed by the scientific community. While being primarily interested in those based on *metagenomic samples comparisons*, we shall provide an overview of the subject from a wider perspective. In this section we summarise the notation we use and give a general introduction, discussing briefly various proposed ways of categorising such indices. Next sections present some of them more pragmatically.

As stated formerly, the notion of *diversity* is often captured by that of distance. However, sometimes it is impossible to precisely measure some distance; even more, rigorous universal method for the purpose could be rather difficult to settle. In such cases, it is more appropriate to substitute the term “measure” with that of “index”, like someone pointing to some direction while still not touching the goal. The same concept of index also applies to that of *summary*, when many diversity values are considered at once.

There are numerous “entities” that biologists and ecologists need to differentiate quantitatively: individuals of the same species, OTUs, communities, and ecosystems, to mention only four of them. As far as we are concerned, the basic characteristics that distinguish environments from a biological viewpoint are:

Species richness: The tally of different species;

Species abundance: The numerosity, or proportion, of individuals for each species;

Taxonomic variety: The taxonomic, or phylogenetic, distance among species.

All of these three components of biological diversity can be easily extended to consider other *OTUs* in place of species.[7] Actually, the majority of the traditional literature refers only to the first two as such components. Indeed, it is usually assumed that **1.** individuals within the same species (OTU) are equivalent, **2.** all species are "equally different" from one another and receive equal weighting, and **3.** diversity is measured in appropriate units of observation [14, 39]. Notwithstanding, for the third component is equivalent to a relaxation of the second assumption and, in fact, is vastly used as well, we feel adequate to integrate it as a component of biodiversity.

2.3.1 Notation

A typical scenario is that of disposing of biological data collected from various environments. Let us collect here the symbols we will use throughout this report.

$N^\star$ True number of individuals in an environment.

$S^\star$ True number of species (or OTUs) in an environment, that is its *species (OTU) richness.*

N_i *True species absolute abundance*: the numerosity of individuals of the i-th species, $i \in \{\, 1 \dots S^\star \,\}$, in an environment: $N^\star = \sum_{i=1}^{S^\star} N_i$.

p_i True probability of randomly selecting an individual of species i — i.e., its true relative abundance — in an environment. Trivially, $p_i = {}^{N_i}\!/_{N^\star}$.

n Number of individuals collected in a sample.

s Count of observed species in a sample.

x_i Tally of individuals of the i-th species in a sample: $n = \sum_{i=1}^{S^\star} x_i$.

ψ_i Observed relative abundance of the i-th species in a sample, that is $\psi_i = {}^{x_i}\!/_{n}$.

f_k *Abundance frequency counts*: number of species in an assemblage[8] for each of which exactly k individuals were observed. Therefore, $f_0 = S^\star - s$ is the number of unobserved species. It holds: $\sum_{k=0}^{n} f_k = S^\star$ and $\sum_{k=1}^{n} f_k = s$.

An additional subscript is employed for referring to some specific sample. For instance, $p_{\mathbf{A},i}$ is the relative abundance of species i in the environment from which a sample $\mathbf{A}$ was drawn, thence $p_{\mathbf{A},i} = {}^{N_{\mathbf{A},i}}\!/_{N^\star_{\mathbf{A}}}$.

[7] In what follows, we will mainly refer to species, as these are the most targeted OTU. It is nonetheless clear that any other OTU could substitute species whenever sensible.

[8] In our scope, an assemblage is the set of individuals exposed to our sampling effort in a defined area or point [17]

2.3.2 Classification of Diversity Indices

The principal categorisation of diversity indices discerns OTU diversity *within* an assemblage and OTU diversity *between* assemblages. The former is commonly called **α-diversity**, the latter **β-diversity**. Depending on the amplitude of the ecosystem at hand, ecologists define **γ-diversity**, **δ-diversity** and **ε-diversity** too. Alpha, gamma and epsilon diversities are *inventory diversities* [18, 25, 39], targeting overall diversity within larger and larger samples. For instance, α-diversity could be employed for measuring community diversity, γ-diversity for regions and ε-diversity for landscapes. Beta and delta diversities, on the other hand, are indices of difference between samples: β-diversity between α-scale assemblages within a "γ-region", whereas δ-diversity measures difference between environments at γ scale within an "ε-region".

Unfortunately, even when discarding δ- and ε-diversities, which are rarely considered, there is no universal agreement on how proposed indices should be grouped into these categories. For example, according to *Jurasinski et al.* [18] there is no real distinction between inventory diversities, whereas *Tuomisto* [39] would rather distinguish α-diversity and γ-diversity by the weights given to each separate subplot.[9] Furthermore, the latter proposed entropy and probability ought to be referred to as such instead of being hidden "behind the vague umbrella term 'index of diversity'" [39]. Besides, β-diversity can be defined either as a link between alpha and gamma, or from scratch by only considering the elements of two samples. As a consequence, it assumes a variety of meanings and scopes.

In addition, even when adopting the same convention, ecologists may refer to alpha diversity of a microbial community as well as the alpha diversity of an island, which are definitely different in scale [44]. In order to avoid all this confusion, we would rather focus only on α- and β-diversity indices of generic assemblages, ignoring their spacial amplitude. The first might be named **sample intra-diversity** indices, or *inventory-diversity* indices, whereas the second could be called **sample inter-diversity** indices, or *sample-comparison* indices. Our main concern is, indeed, comparison between metagenomic samples, which is conceptually quite different from traditional biodiversity indices, even though closely related to them. Notwithstanding, although it may be useful to adopt diverse vocabulary to avoid confusion, we will stick to the common terminology for compatibility. We shall also report some of the traditional work done on the α-, β- and γ- triad in the following sections.

[9] However, we notice in her paper a possible minor inconsistency: by distinguishing between α- and γ-diversity indices, it might have been mathematically more coherent in formula (7) at p.857 to write ${}^{q}D_{\gamma,j}$ in place of ${}^{q}D_{\alpha,j}$. This observation highlights a natural difficulty posed by her method: the inventory diversity at low scale (alpha) is computed as a function of those indices used at a higher scale (gamma). To avoid confusion, for an assemblage $\textsc{j}$, we will use the notation ${}^{q}D_{\textsc{j}}$ which does not take scale into account, and thence ${}^{q}D_{\textsc{j}} = {}^{q}D_{\gamma,\textsc{j}}$.

A secondary classification is founded on the type of sampling data structure: **Individual-based (abundance) data** and **Sample-based (incidence) data** [10]. *We will only target the first class*, in which individuals are sampled and assigned to their OTUs, and we will consider sequences as individuals. In incidence data, instead, individuals are replaced by sampling units; see [14] for a review. Worth noting is that this categorisation is fundamental in estimating α-diversities, but we have not found it applied to β-diversities yet. However, the same terms “abundance” and “incidence” are used for distinguishing β-diversities based on species abundance and species richness, respectively: they should not be confused with the type of data sampling.

When it comes to *metagenomics*, another distinction is set by the *approach* adopted to detect and quantify species in a metagenomic sample. **Reference-based methods** make use of a reference database to classify the collected metagenomic sequences at the level of the desired OTU (e.g., species) and thus employ such a classification in subsequent computation. **Reference-free methods**, on the flip side, cluster together reads from the same OTU by means of sequence composition similarity [11]. Such methods are also known as *taxonomic binning/classification* and *genome binning/classification*, respectively. Ideally, the first category should provide better results by being more readily comparable with traditional measures. However, reference databases are extremely poor compared to the myriads of unknown and uncultivable microbes actually present in our biosphere. Consequently, genome binning is often more accurate [12].

Finally, a very relevant differentiation has been recently pointed out by Sun et al. [37]. *Taxonomic classification* comprises: **1.** alignment-based, **2.** marker-based, and **3.** composition-based methods [11]. In contrast, the authors of [37] recommend to unambiguously discriminate between **sequence abundance** and **taxonomic abundance** when dealing with reference-based methods. Taxonomic abundance is generally obtained by marker-based approaches only, whereas methods 1 and 3 supply sequence abundance. As they displayed, indices computed from these two kind of data are incompatible both mathematically and practically. The point is that length and ploidy of species' genomes are hugely variable and, hence, sequence abundance does not reflect species abundance, which is a taxonomic abundance.

Due to such an incompatibility, we remark that, despite considering *individual-based data* only, we had better not to confuse traditional biodiversity indices with those based on sequence abundance. It would be interesting, though, to study whether *incidence data* might overcome this latest distinction if properly managed.

Despite being interested in metagenomics comparison in the sense of sequence-based comparison, we report in the following the classic formulation of several dissimilarity indices in terms of species and individuals: the transposition to metagenomic data is immediate.

2.3.3 Compatibility of Diversity Estimates

Besides incommensurability of sequence-based indices with taxonomy-based indices, a number of other cautions are to be taken. In fact, neither true species richness nor true species frequencies can be known – a common difficulty in measurement theory. Therefore, only estimates of them are available. However, such estimates are likely to be biased by the methods adopted to calculate them. Discrepancies in sampling effort — e.g., number of individuals collected, area amplitude, quantity of traps employed etc. — demands particular care when benchmarking α-diversity indices of separate samples — i.e., *proportional diversity* in [18] terminology — and for measuring β-diversities.

One proposed solution for the problem is **rarefaction**, a subsampling technique for reducing all datasets down to the lowest available sampling effort. However, while sampling effort being often measured in number of collected individuals, it can also refer to sampled volume — and organisms density can very greatly, — or to time dedicated for capturing individuals, or still other characteristics. Moreover, rarefaction implies a loss of information and is said to be "neither justifiable nor necessary" [43] in the context of comparison of relative abundances.

When comparing two environmental samples, the matter gets even more delicate. Indeed, it is essential to measure diversity between assemblages regardless of efficiency of sampling methods. The latter, then, may need to be commensurate with the effective population's size of the habitats. This becomes of utmost relevance in benchmarking, as reference samples may be targeted by different studies.

However, we shall not expand this issue here.

2.4 Alpha Diversities

Alpha diversity indices try quantify how biologically differentiated is an environment at an elementary scale. For instance, some α-diversity of a metagenomic sample should be an index of variety of the microbial community sampled. In order to measure such a variety, each of the components of biological diversity listed above can be employed.

Hereafter we will use Table 2.1 as a simple instance to show how to compute diversity indices.

2.4.1 Species Richness

The easiest way of quantifying sample intra-diversity is to provide the tally of diverse species (or OTUs) present in the environment under study. Since the true species richness $S^\star$ is quite impossible to measure, several *non-parametric estimators* have been proposed, while other statistical strategies applied so far have proven to

Table 2.1: A toy example of species abundances observed in three samples. Symbols in parenthesis. Units are number of individuals.

Species (i)	Sample A ($x_{\mathbf{A},i}$)	Sample B ($x_{\mathbf{B},i}$)	Sample C ($x_{\mathbf{C},i}$)
1	10	0	1
2	15	0	0
3	8	9	10
4	13	9	15
5	6	9	14
6	1	0	2
7	0	9	3
8	0	9	1
9	0	0	1
10	0	9	1
Total (n)	53	54	48

be unsuccessful [14, 15]. Indices of species richness should be expressed in *actual number of species*, symbol " sp".

Trivially, a metric space $(\mathcal{S}, d')$ for measuring species richness is one where $\mathcal{S}$ is the set of every possible species and d' is the 1-complement of the Kronecker delta function:

$$d'(x, y) = 1 - \delta_{xy} = \begin{cases} 0 & \text{if individuals } x = y \\ 1 & \text{otherwise} \end{cases} \tag{2.2}$$

The set $\mathcal{S}$ can be easily viewed as a quotient set of the set of all living organisms with an equivalence relation yielded by species membership. Species richness is then the sum, over each detectable species, of their average distance from different species.

Alternatively — and rather profitably, I believe, — we could settle a metric space $(\mathcal{M}, d)$ on a set $\mathcal{M}$ of *leaf nodes* of a **star graph** $G(V, E)$, which is a tree with only leafs except the root $r \in V$, where each leaf represent a species and thus is an equivalence class on individuals. Therefore, $\mathcal{M} \subseteq V \setminus \{r\}$. We set $w_e = c$, $\forall e \in E$ for some constant value c, for which we suggest $c = 1/2$. The distance $d(u, v)$ between two nodes $u, v \in V \setminus \{r\}$ — i.e., two species — is then the length of the shortest path connecting the two nodes in G, divided by $2c$. The species richness of an environment represented by $\mathcal{M}$ is then the length of the shortest cycle traversing every node in the smallest connected subgraph $G_{\mathcal{M}}(V_{\mathcal{M}}, E_{\mathcal{M}})$ induced[10]

[10] Formally, $V_{\mathcal{M}} = \mathcal{M} \cup \{r\}$ and $E_{\mathcal{M}} = \{\, \mathcal{M} \times \{r\} \,\} \cup \{\, \{r\} \times \mathcal{M} \,\} = \{\, e \in E \colon e \cap \mathcal{M} \neq \emptyset \,\}$.

by $\mathcal{M}$, divided by $2c$.

Chao1 index

A simple, yet accurate, *lower bound* for species richness is *Chao1 estimator*, for which the variance is known [14]. The assumption is that rare species carry the majority of information about true species richness.

Recalling notation from Section 2.3.1, s is the number of observed species, f_1 is the amount of species of which only one individual was observed, and f_2 is the tally of species of which exactly two individuals were observed.

$$\widehat{S}_{\text{Chao1}} = \begin{cases} s + {}^{f_1^2}\!/_{2f_2} & \text{if } f_2 > 0 \\ s + {}^{f_1(f_1-1)}\!/_{2} & \text{if } f_2 = 0 \end{cases} \tag{2.3}$$

Considering the example in Table 2.1, there are no species in $\mathbf{A}$ with exactly two individuals, therefore $f_{\mathbf{A},2} = 0$; moreover, species **6** is the only one detected exactly once in $\mathbf{A}$, so $f_{\mathbf{A},1} = 1$; and, trivially, there are $s_{\mathbf{A}} = 6$ observed species in $\mathbf{A}$. Similarly we have $f_{\mathbf{C},2} = 1$, $f_{\mathbf{C},1} = 4$, and $s_{\mathbf{C}} = 9$. Then, samples $\mathbf{A}$ and $\mathbf{C}$ hold, respectively:

$$\widehat{S}_{\text{Chao1},\mathbf{A}} = 6 + \frac{1 \cdot (1-1)}{2} = 6\,\text{sp}$$

$$\widehat{S}_{\text{Chao1},\mathbf{C}} = 9 + \frac{4^2}{2 \cdot 1} = 17\,\text{sp}$$

Abundance-based Coverage Estimator – ACE

A more sophisticate estimator is the *Abundance-based Coverage Estimator*, which uses information of rare species up to the frequency of κ individuals per species. Empirically, a cut-off of $\kappa = 10$ works well in practice [14]. Coverage ($\widehat{C}_{\text{rare}}$) is an estimated proportion of the n individuals recorded in the sample over the total true numerosity $N^\star$ of the population.

$$\widehat{S}_{\text{ACE}} = s_{\text{abun}} + \frac{s_{\text{rare}}}{\widehat{C}_{\text{rare}}} + \frac{f_1}{\widehat{C}_{\text{rare}}}\widehat{\gamma}^2_{\text{rare}} \tag{2.4}$$

With:

$$\begin{aligned}
s_{\text{rare}} &= \sum_{i=1}^{\kappa} f_i \\
s_{\text{abun}} &= \sum_{i=\kappa+1}^{n} f_i = s - s_{\text{rare}} \\
n_{\text{rare}} &= \sum_{i=1}^{\kappa} i f_i \\
\widehat{C}_{\text{rare}} &= 1 - {}^{f_1}\!/_{n_{\text{rare}}} \\
\widehat{\gamma}^2_{\text{rare}} &= \max\left\{ \frac{s_{\text{rare}}}{\widehat{C}_{\text{rare}}} \cdot \frac{\sum_{i=1}^{\kappa} i(i-1) f_i}{n_{\text{rare}}(n_{\text{rare}}-1)} - 1, \quad 0 \right\}
\end{aligned}$$

In sample **c** of Table 2.1, with $\kappa = 8$ we have:

$$\begin{aligned}
\widehat{S}_{\text{ACE},\mathbf{C}} &= 3 + \frac{6}{1 - {}^{4}\!/_{9}} + \frac{4}{1 - {}^{4}\!/_{9}} \cdot \max\left\{ \frac{6}{1 - {}^{4}\!/_{9}} \cdot \frac{(1 \cdot 0 \cdot 4) + (2 \cdot 1 \cdot 1) + (3 \cdot 2 \cdot 1)}{9 \cdot 8} - 1, \quad 0 \right\} \\
&= 3 + \frac{6 \cdot 9}{5} + \frac{4 \cdot 9}{5} \cdot \frac{1}{5} \\
&= 15\,\text{sp}
\end{aligned}$$

Other Indices

Thukral reported in [38] a mathematical characterisation of indices k of species richness as being such that for the estimated number of species richness it holds $\mathbb{E}[S] = f(k, n)$ for some function f. Accordingly, several such indices have been proposed, e.g. ratios between observed species richness s and some function of the number of detected individuals n. For such indices, collected in [38], adequate units of observation should be declared.

Worth noticing is that many of these indices do only account for species richness, whereas Chao1 and ACE indices provide estimated lower bounds of it by also considering species abundance distribution. Hence, the latter are preferable. Furthermore, as remarked in [43], estimators and indices for which their *variance* is unknown had better be discarded from scientific research.

2.4.2 Species Evenness

Consider samples **a** and **b** in Table 2.1: they equal in number of observed species, yet they appear rather dissimilar. Such a difference is caught by species relative abundance, on which various **indices of evenness** are built. Those indices convey information on how much species are equally distributed in an environment. Although alternative *densities* could be targeted, evenness in this context traditionally refers only to probabilities $p_i = {}^{N_i}\!/_{N^\star}$, or their estimates.

Shannon Entropy

Already a milestone in information theory, *Shannon Entropy* $\mathrm{H_{Sh}}$ has been profitably applied to ecology as well. Let $J\colon x \mapsto i$ be a random variable associating a randomly chosen individual x from some environment to its species i: $P[J = i] \doteq p(i) = p_i,\ \forall i \in \{\,1 \ldots S^\star\,\}$. Then, with the convention that $0\log(0) = 0$, the biological entropy of the environment is

$$\mathrm{H_{Sh}} = \mathbb{E}[-\log(p(J))] = -\sum_{i=1}^{S^\star} p_i \log p_i \tag{2.5}$$

If the natural logarithm is employed — which is advisable, — then entropy is measured in *nats*. If $\log_2$ is used, *bits* are the unit of measure; when base 10 is used, the units are *decits* or *hartleys* [38]. For instance, referring to Table 2.1 with the assumption of completeness of the samples — i.e., observed data equal true data, — we calculate:

$$\begin{aligned}
\mathrm{H_{Sh,\mathbf{A}}} &= \frac{10}{53}\ln\frac{53}{10} + \frac{15}{53}\ln\frac{53}{15} + \frac{8}{53}\ln\frac{53}{8} + \cdots \approx 2\,\mathrm{nat}\\
\mathrm{H_{Sh,\mathbf{B}}} &= 6 \times \frac{9}{54}\ln\frac{54}{9} \approx 2\,\mathrm{nat}\\
\mathrm{H_{Sh,\mathbf{C}}} &= 4 \times \frac{1}{48}\ln 48 + \frac{10}{48}\ln\frac{48}{10} + \cdots \approx 2\,\mathrm{nat}
\end{aligned}$$

Entropy, here, measures the expected amount of information carried by the identification of the species of a randomly chosen individual. Since common species convey little information in this sense, samples with highly uneven species abundances have low entropy. Uniform species distributions, then, yield maximum entropy.

Calling $\mathrm{H_{Sh}}$ a diversity index, however, has been discouraged since it more properly is an entropy [39]. The term “diversity” is preferably reserved to Hill numbers only (see below, Section 2.4.2). Nevertheless, entropy has been largely exploited, despite an estimator of its true value is “extremely sensitive to the singleton count, which is often difficult to determine in microbiome studies” [43].

As a result, the equivalent in **effective species richness** of Shannon entropy, $e^{\mathrm{H_{Sh}}}$, should be considered, provided natural logarithm is used.[11] Such an index can be interpreted as *the number of species yielding an identical entropy when evenly distributed*, number which is never higher than the **actual species richness**. Accordingly, we quantify $e^{\mathrm{H_{Sh}}}$ in terms of **effective number of species** ($\mathrm{sp_E}$), which however needs not be an integer. As an example, the former indices would be

[11] Otherwise, the transformation must be adjusted accordingly: $2^{\mathrm{H_{Sh}(\,nat)}}$ or $10^{\mathrm{H_{Sh}(\,decit)}}$ for instance.

transformed to:

$$
\begin{aligned}
&e^{\mathrm{H}_{\mathrm{Sh},\mathbf{A}}} \approx e^{1.624} \approx 5\,\mathrm{sp}_{\mathrm{E}}\\
&e^{\mathrm{H}_{\mathrm{Sh},\mathbf{B}}} = e^{\ln 6} = 6\,\mathrm{sp}_{\mathrm{E}}\\
&e^{\mathrm{H}_{\mathrm{Sh},\mathbf{C}}} \approx e^{1.678} \approx 5\,\mathrm{sp}_{\mathrm{E}}
\end{aligned}
$$

Gini-Simpson Index

The probability of randomly selecting (with replacement[12]) two individuals of distinct species holds another index for quantifying evenness, which is commonly known as the *Gini-Simpson index*. Indeed, when few highly abundant species are present, it would be more likely to pick two individuals of one of such species. Under equal species richness, the maximum probability of selecting two diverse individuals is obtained when species are evenly distributed.

$$P_{\mathrm{GS}} = 1 - \sum_{i=1}^{S^\star} p_i^2 \tag{2.6}$$

Being a probability, P_{GS} is a pure number. Again, for it is not a measure of species richness, its use as an α-diversity index is discouraged [43]. Its equivalent in species richness is the *inverse Simpson index* $D_{\mathrm{IS}} = (1 - P_{\mathrm{GS}})^{-1}$, which actually is the second-order Hill number, discussed below (Section 2.4.2). Following the previous example,

$$
\begin{aligned}
&P_{\mathrm{GS},\mathbf{A}} = 1 - \left(\frac{10}{53}\right)^2 - \left(\frac{15}{53}\right)^2 - \left(\frac{8}{53}\right)^2 - \cdots \approx 0.788\\
&P_{\mathrm{GS},\mathbf{B}} = 1 - 6 \times \frac{1}{6^2} = \frac{5}{6} = 0.8\overline{3}\\
&P_{\mathrm{GS},\mathbf{C}} = 1 - 4 \times \left(\frac{1}{48}\right)^2 - \left(\frac{10}{48}\right)^2 - \cdots \approx 0.766
\end{aligned}
$$

Hill Numbers

Aforementioned, a desirable property is for α-diversities indices to be measured in units of (actual/effective) species richness [43]. In addition, a **replication principle**, or **doubling property** should be met[13]: *if M equally diverse assemblages with no shared species are pooled in equal proportions, then the diversity of the pooled*

[12] In real environments, the immense amount of living organisms make selection with replacement a sensible model. Minimum Variance *Unbiased* Estimators, instead, use selection without replacement as model [14].

[13] Equivalent Shannon entropy $e^{\mathrm{H}_{\mathrm{Sh}}}$ already satisfies these properties.

assemblage should be M times the diversity of each single assemblage [14]. *Hill numbers* are a class of diversity indices for which such properties hold. They are therefore measured in *effective number of species* ($\mathrm{sp_E}$), suggesting that an even distribution of such "equivalent" number of species would yield the same value for the adopted index. These indices are

$$ {}^{q}D = \left(\sum_{i=1}^{S^\star} p_i^q \right)^{\frac{1}{1-q}} \tag{2.7} $$

The order q controls the sensitivity of the measure to species relative abundances. For low order values, $q < 1$, rare species are weighted more. At $q = 0$ we get species richness: ${}^0D = S^\star$. For $q > 1$, common species contribute increasingly more. As anticipated, at $q = 2$ the inverse Simpson index is obtained: ${}^2D = {}^1\!/_{1-P_{\mathrm{GS}}}$, which severely discounts the contribution of rare species. The order $q = 1$ Hill number is undefined, but [14, 43]

$$ {}^1D \doteq \lim_{q\to 1} {}^qD = e^{\mathrm{H_{Sh}}} \tag{2.8} $$

Remarkably, every Hill number is measured in effective species richness, fact which makes this a coherent *class of indices*.[14] Thank to such a coherence, *diversity profiles* of effective species richness versus q can be drawn. We refer to [14] for an explanation of them, as well as for estimators of Hill numbers.

2.4.3 Taxonomic Indices

Suppose two islands where observed, both with identical species richness and evenness. In the first one, only angiosperms are present; in the second one, a both angiosperms and gymnosperms grow. Then, the second island would be classified as *taxonomically* richer in biodiversity than the first one because of the presence of more distantly related species [3, 14].

To quantify *taxonomic or evolutionary relatedness*, it suffices a generalisation of the star graph introduced in Section 2.4.1. Notoriously, taxonomies and phylogenies are organised in **cladograms**, which are trees (rooted acyclic connected graphs). Consequently, the metric space for measuring taxonomic diversity is $(\mathcal{M}, d)$ as before, where only the star graph $G_{\text{star}}(V, E)$ is replaced with a taxonomic/phylogenetic tree $G_{\text{tax}}(W, E')$, with edge weights w_e needing not be constant. Species still correspond to leaf nodes: $\mathcal{M} \subset W$. Such trees can either be **ultrametric**, when all leaf nodes are at equal distance from the root, or **nonultrametric** otherwise [14].

Due to the variability of $w_e \in E'$, the distance $d(u, v)$ between $u, v \in \mathcal{M} \subset W$ must be adapted in the normalisation strategy — e.g., no normalisation, or division

[14] The same is not true for Good's series of indices [38], for which reason we have not reported it.

by $2\overline{w} = 2\mathbb{E}[w_e]$, — but the length of the *shortest path* linking the two nodes still applies. Alternative distances might be employed, provided they are metric functions. For instance, maximum depth from the lowest common ancestor is viable, but minimum such depth would not satisfy the triangular inequality (2.1c).

Among the proposed α-diversity indices accounting for taxonomic relatedness, **cladistic diversity** *is the tally of nodes in the minimal tree encompassing all the detected species.* Faith's **phylogenetic diversity** is, instead, the *sum of the branch lengths* (edge weights) *in a minimal phylogeny comprising all species in the target assemblage.*[15] These two measures do not make any use of species abundances, which are instead integrated in *Rao's quadratic entropy* and the *phylogenetic entropy.* For the taxonomically aware equivalent of Hill numbers we link to [14].

Rao's Quadratic Entropy

Let $L\colon \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ be a random variable denoting the pseudo-distance $\delta(x, y) = d(i, j)$ between two randomly selected individuals $x, y \in \mathcal{M}$, with $i = {}^{x}/_{\sim} = [x], \quad j = {}^{y}/_{\sim} = [y]$ being their equivalence classes in the space $\mathcal{S} = {}^{\mathcal{M}}/_{\sim}$ of species. Then its expected value, that is *Rao's quadratic entropy*, is an α-diversity index accounting both for taxonomic distance and species relative abundance. It can be viewed as a generalisation of Gini-Simpson index, which is actually obtained when $(\mathcal{M}, d)$ is defined on a star as in Section 2.4.1.

$$Q_{\text{Rao}} = \mathbb{E}[L] = \sum_{x,y\in\mathcal{M}} \frac{\delta(x, y)}{|\mathcal{M}|^2} = \sum_{i,j\in\mathcal{S}} d(i, j) p_i p_j \tag{2.9}$$

Phylogenetic Entropy

Let $\sigma\colon \mathcal{M} \to W$ link a species $i \in \mathcal{M}$ to its correspondent leaf node $\sigma(i) \in W$ of an evolutionary tree $G_{\text{phyl}}(W, E')$ rooted at node $r \in W$. Let $\lambda_{\sigma(i)} = d(r, \sigma(i))$. Then, the phylogenetic entropy of an assemblage is

$$\mathrm{H}_{\text{phyl}} = -\sum_{i\in\mathcal{M}} \lambda_{\sigma(i)} p_i \log p_i \tag{2.10}$$

See [14] for an equivalent characterisation.

2.4.4 Taking Scale Into Account

Every α-diversity index introduced so far can be used as a γ-diversity index as well. Let us now examine a collection of K_Γ subplots (*compositional units*) within a larger

[15] In an ultrametric tree, Faith's phylogenetic diversity would equal the number s of observed species multiplied by the distance of any species from their lowest common ancestor.

plot $\mathbf{\Gamma}$. Let $p_{\mathbf{z},i} = {}^{N_{\mathbf{z},i}}/_{N^\star_{\mathbf{z}}}$ be the relative abundance of species $i \in \{1 \dots S^\star\}$ in sample $\mathbf{z} \in \{1 \dots K_{\mathbf{\Gamma}}\}$.

We could devise an **α-diversity index of the region** as an *average biological diversity per subplot.* Such an index should meet Hill numbers' properties and be measured in **effective species richness per effective compositional unit** ($\mathrm{sp_E\,CU_E}^{-1}$) if adjusted for species relative abundance, or in **species richness per compositional unit** ($\mathrm{sp\,CU}^{-1}$) otherwise [39].

The class of the former indices is

$$^{q}D_{\alpha,\mathbf{\Gamma}} = \left[\sum_{\mathbf{z}=1}^{K_{\mathbf{\Gamma}}} \left(w_{\mathbf{z}} \sum_{i=1}^{S^\star} p_{\mathbf{z},i}^{q} \right)\right]^{\frac{1}{1-q}} = \left[\sum_{\mathbf{z}=1}^{K_{\mathbf{\Gamma}}} w_{\mathbf{z}} \left({}^{q}D_{\mathbf{\Gamma}} \right)^{1-q}\right]^{\frac{1}{1-q}} \tag{2.11}$$

where $w_{\mathbf{z}} = {}^{N^\star_{\mathbf{z}}}/_{N^\star_{\mathbf{\Gamma}}}$ is the portion of individuals of the region $\mathbf{\Gamma}$ contributed by subunit $\mathbf{z}$.

If we let $w_{\mathbf{z}} = {}^{1}/_{K_{\mathbf{\Gamma}}}$ instead, and $q = 0$ in Eq. (2.11), then we get the α-index of species richness of the region $\mathbf{\Gamma}$, which belongs to the second type of indices (measured in $\mathrm{sp\,CU}^{-1}$).

$$S_{\alpha,\mathbf{\Gamma}} = \frac{1}{K_{\mathbf{\Gamma}}} \sum_{\mathbf{z}=1}^{K_{\mathbf{\Gamma}}} S^\star_{\mathbf{z}} \tag{2.12}$$

2.5 Beta Diversities

As previously mentioned, **β-diversities** attempt to *quantify the diversity between assemblages.* Whereas all α-diversities are indices of how much is some environment diversified, although from different perspectives maybe, β-diversities assume several different meaning. Among them, we find species turnover[16], some relation between shared and unshared species, and number of equivalent units constrained to satisfy particular properties. Moreover, at least three radically different approaches exists for comparing biotic differentiation between distinct sites: **1.** comparison of their *a-diversities*, this has been called “proportional diversity” [18]; **2.** comparison of assemblages composition[17], *species by species* (“differentiation diversity” [18]), which is the only approach we report in what follows; and **3.** comparison of *shared-abundance* of the samples, that is the total number individuals belonging to any shared species [14, 17]. Such a variety of meanings should be considered seriously in order to avoid confusion and wrongful conclusions.

[16] Species turnover: Rate, or magnitude, of change in species composition *along predifined* spatial, environmental or time *gradients* [40].

[17] By composition of an assemblage we mean the list of species comprising it and their *relative* abundances, i.e., their distribution.

In order to guarantee validity of ecological inferences, every index of *compositional* differentiation between assemblages should satisfy at least the following three properties [17]:

Monotonicity it must monotonically increase when assemblages differentiation unambiguously decreases;

Density invariance it should only depend on *relative* abundances;

Replication invariance those indices ranging between 0 and 1 shall be invariant with respect to pooling of identical subsets (essential property in part-to-whole studies).

All the incidence- and abundance-based indices described below satisfy these three properties, except the Bray-Curtis index, which is not density invariant.

2.5.1 Taking Scale Into Account

Traditionally, β-diversity has been proposed as a *multiplicative link* between α- and γ-diversity [42] (see also below, Section 2.6). Given a region $\mathbf{\Gamma}$ and its inventory diversities ${}^{q}\mathcal{D}_{\alpha,\mathbf{\Gamma}}$ and ${}^{q}\mathcal{D}_{\gamma,\mathbf{\Gamma}} = {}^{q}\mathcal{D}_{\mathbf{\Gamma}}$, β-diversity indicates the number of *distinct virtual compositional units* — i.e., non overlapping assemblages with no species in common, — each of α-diversity ${}^{q}\mathcal{D}_{\alpha,\mathbf{\Gamma}}$, needed to make up a region of ${}^{q}\mathcal{D}_{\gamma,\mathbf{\Gamma}}$ γ-diversity [39]. Formally,

$$ {}^{q}\mathcal{D}_{\beta} = \frac{{}^{q}\mathcal{D}_{\gamma}}{{}^{q}\mathcal{D}_{\alpha}} \tag{2.13}$$

The units of observation of ${}^{q}\mathcal{D}_{\beta}$ are **effective compositional units** (CU$_{\text{E}}$), as just explained.

A simpler index of β-diversity carrying a similar meaning, but measured in **actual compositional units** (CU), is based on presence-absence data only, hence ignoring species abundances. This is the original **Whittaker index** S_{β}, i.e., the total count of species present in a region divided by the average number of species per subunit.

$$ S_{\beta} = \frac{S_{\gamma}}{S_{\alpha}} \tag{2.14}$$

More pragmatically, in a region $\mathbf{\Gamma}$ comprising $K_{\mathbf{\Gamma}}$ subunits,

$$ S_{\beta,\mathbf{\Gamma}} = \frac{S_{\gamma,\mathbf{\Gamma}}}{S_{\alpha,\mathbf{\Gamma}}} = \frac{S^{\star}_{\mathbf{\Gamma}}}{\frac{1}{K_{\mathbf{\Gamma}}}\sum_{\mathbf{z}=1}^{K_{\mathbf{\Gamma}}} S^{\star}_{\mathbf{z}}} \tag{2.15}$$

As a simple example, let us suppose the three samples $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ from Table 2.1 are the true species distribution of three subunits of which a larger region is composed.

Then,

$$
\begin{aligned}
S_\gamma &= 10\,\mathrm{sp} \\
S_\alpha &= \frac{6+6+9}{3} = 7\,\mathrm{sp}\,\mathrm{CU}^{-1} \\
S_\beta &= \frac{10}{7} \approx 1\,\mathrm{CU}
\end{aligned}
$$

Remarkably, such β-diversity indices should not be regarded as distances, nor as summaries of distances, but rather as the height of a parallelepiped of which the volume represents γ-diversity and the area of its base represents α-diversity [39]. Alternatively, they could also be visualised as some measure of a "β-volume" in which a "γ-mass" is dispersed with "α-density". Obviously, indeed, this is not a measure of difference between two samples, but somewhat an *index of how well species in a vast region are physically clustered into smaller groups.*

Other definitions of "scale-sensible" biological diversity, or "summary" diversity, have been advanced. The indices above relate to a **multiplicative** approach — i.e., $\gamma = \alpha\beta$, — to which an **additive** $\gamma = \alpha + \beta$ approach is opposed. In this latter case, all the diversity indices must be commensurable, thus measured by the same units of observation, which is quite counterintuitive. Its advantage over the multiplicative approach is that it needs not "α-samples" to be of equal sizes. However, empirical studies as well as mathematical considerations discourage its use [44]. Another proposed method, a hybrid of the formers, is a **proportional** approach $\beta = {}^{(\gamma-\alpha)}/_{\gamma}$, which again has no obvious physical interpretation.[18]

2.5.2 Incidence-Based Indices

Two distinct biotic assemblages can be compared with each others by means of *presence-absence* of species in them. This is the case of so-called **incidence-based indices** of β-diversity [14, 17], which are not to be confused with incidence-based sample collection as seen in Section 2.3.2.

Generally, such indices are real-valued functions taking two (or more) elements of a collection $\mathscr{C}$ of *sets of species*. As seen in Section 2.4.1, such sets of species can be modelled as subsets of a metric space: $\mathscr{C} \ni \mathcal{A} \subset (\mathcal{S}, d')$ or $\mathscr{C} \ni \mathcal{A} \subset (\mathcal{M}, d)$.

Whittaker's β-diversity index S_β is an example of incidence-based indices. However, while Whittaker's index provides a conceptual link between local and regional diversity [44], β-diversity indices are more commonly defined as some dissimilarity measure between two samples, with no need of geographical considerations.[19] Among

[18] In this case, such a β-diversity would be a pure number.

[19] We report that attempts to bring β-diversity indices interpretation back to Whittaker's definition have been made [31].

these indices, we report below the two most widely adopted ones: the Jaccard and the Sørensen indices, which actually differ in scale but are monotone one to the other. This means that if the Jaccard distance between a couple of sets is greater than that between another couple of sets, then so is the Sørensen dissimilarity index [19]. It is possible to generalise such coefficients to work on more than two assemblages [14, 17], but we will stick to the classical definitions for simplicity.

Jaccard Distance

The *Jaccard distance* compares the number of unshared species to the total number of species in the combined assemblages. Therefore, it provides a pure number in [0,1] which can be interpreted as the probability of randomly selecting an unshared species among all the species detected in the samples. In this sense, the Jaccard index takes a global view of the problem.[17] If we model two assemblages **A**, **B** with two sets $\mathcal{A}, \mathcal{B} \in \mathscr{C}$ respectively, then the Jaccard distance

$$\mathrm{B}_{\text{Jac}}(\mathbf{A}, \mathbf{B}) = 1 - \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|} \tag{2.16}$$

is a metrics on $\mathscr{C}$.

Back to example in Table 2.1, it holds:

$$\mathrm{B}_{\text{Jac}}(\mathbf{A}, \mathbf{B}) = \frac{3}{9} = 0.\bar{3} \qquad \mathrm{B}_{\text{Jac}}(\mathbf{A}, \mathbf{C}) = \frac{5}{10} = 0.5 \qquad \mathrm{B}_{\text{Jac}}(\mathbf{B}, \mathbf{C}) = \frac{6}{9} = 0.\bar{6}$$

Sørensen Index

The *Sørensen similarity index* between two sets $\mathcal{A}, \mathcal{B} \in \mathscr{C}$ — representing species in the two respective assemblages **A** and **B** — is the harmonic mean of the proportion of shared species within one assemblage and the proportion of shared species within the other assemblage [14]. *Sørensen dissimilarity index* is thus the one-complement of the former:

$$\mathrm{B}_{\text{Sør}}(\mathbf{A}, \mathbf{B}) = 1 - \frac{2|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A}| + |\mathcal{B}|} \tag{2.17}$$

The Sørensen dissimilarity coefficient can be interpreted as the probability of randomly selecting an unshared species among all the species present in one of the the two samples, randomly selected. In this sense, it takes a local view of the problem [17]. Noticeably, it is only a semi-metrics as it violates triangular inequality (2.1c) [19].

In the example from Table 2.1, we have:

$$\mathrm{B}_{\text{Sør}}(\mathbf{A}, \mathbf{B}) = \frac{6}{6+6} = 0.5 \qquad \mathrm{B}_{\text{Sør}}(\mathbf{A}, \mathbf{C}) = \frac{10}{6+9} = 0.\bar{6} \qquad \mathrm{B}_{\text{Sør}}(\mathbf{B}, \mathbf{C}) = \frac{12}{6+9} = 0.8$$

2.5.3 Abundance-Based Indices

In contrast with incidence-based indices, **abundance-based indices** are indicators of β-diversity taking into account the abundance of each species in assemblages. The easiest way of modelling the latter to compute such indices is by point representations in $\mathbb{R}^S$, where S is the number of known species, or that of species of interest. There is a bijection between dimensions of $\mathbb{R}^S_+$ and species, and the coordinate of a point in the i-th direction corresponds to the abundance of the i-th species in the respective assemblage.

It is worth noting that by means of the equivalence relation $\sim\colon \mathbb{R} \to \{0,1\}$ such that $\forall x \in \mathbb{R}\colon x \sim 0 \iff x = 0$, trivially expanded component-wise on $\mathbb{R}^S$ to $\{0,1\}^S$, we can pass from abundance-based to incidence-based indices. In this section, indeed, we present three indices which are commonly considered as quantitative extensions of the Sørensen similarity or dissimilarity index, which means that they are equal when calculated on assemblages with even species distribution.

Bray-Curtis Dissimilarity

One of the most frequently adopted indices of β-diversity is the *Bray-Curtis dissimilarity index*[20] *(BC)*. Like the Sørensen dissimilarity index is a semi-metric on $\mathscr{C}$, BC index too is a semi-metric on $\mathbb{N}^S \subset \mathbb{R}^S_+$. It measures, indeed, how many individuals from one of the two assemblages cannot be coupled with an individual of the same species from the other assemblage, and normalises such quantity by the mean number of individuals per plot. Resuming the notation given in Section 2.3.1, $S^\star$ being the true number of species in the pooled assemblage,

$$\mathrm{B}_{\mathrm{BC}}(\mathbf{A},\mathbf{B}) = 1 - 2\frac{\sum_{i=1}^{S^\star} \min(N_{\mathbf{A},i}, N_{\mathbf{B},i})}{N^\star_{\mathbf{A}} + N^\star_{\mathbf{B}}} = \frac{\sum_{i=1}^{S^\star} |N_{\mathbf{A},i} - N_{\mathbf{B},i}|}{N^\star_{\mathbf{A}} + N^\star_{\mathbf{B}}} \tag{2.18}$$

It is worth underlying that BC index is computed on *absolute species abundances* [17, 19, 30], hence accounting for differences in population sizes between compared assemblages, in addition to their composition. In practice, the unknown true abundances are replaced by observed (sample) abundances. This approach rises some issues: it confounds composition similarity with density and it becomes *statistically meaningless in case of unequal sampling fraction* (ratio between sampling and population sizes) [17].

Notwithstanding, it is not rare to find BC dissimilarity computed on data normalised by sample size, therefore on *relative* abundances, as reported by Calle [9] ans adopted by Dubinkina et al. [12] and Sentinella et al. [35], for instance.

[20] Curiously, such an index of dissimilarity is named after Bray and Curtis since it was asserted in the field after their publication on an ordination method [8], yet the index itself was previously introduced by the Polish mathematician H. Steinhaus [19].

Such method has empirically proven to be useful and informative, yet it is strongly discouraged [14]. As a matter of fact, though, by substituting absolute quantities with their respective relative quantities in Eq. (2.18), we get a *Manhattan distance*, which is a metric function on relative abundances. On absolute abundances we get a pseudo-metrics, instead. In other words, BC index becomes a distance when computed on spaces $\left\{ x \in \mathbb{R}^S \colon \|x\|_1 = N \right\}$, for every $N \in \mathbb{R}$.

As an example, following Table 2.1 we apply two different formulation of the BC index to get:

$$\begin{aligned}
\mathrm{B}_{\mathrm{BC}}(\mathbf{A},\mathbf{B}) &= \frac{10+15+1+4+3+1+9+9+0+9}{53+54} = \frac{61}{107} \approx 0.57\\
\mathrm{B}_{\mathrm{BC}}(\mathbf{B},\mathbf{C}) &= 1 - \frac{0+0+9+9+9+0+3+1+0+1}{\frac{54+48}{2}} = \frac{51-32}{51} = \frac{19}{51} \approx 0.37
\end{aligned}$$

Horn Overlap Measure

Another abundance formulation of the Sørensen dissimilarity index is the *Horn overlap measure*, which measures samples dissimilarity based on Shannon's entropy. This index is particularly useful when compositional differentiation is to be assessed and rare species are important, e.g. in conservation biology [14].

$$\mathrm{B}_{\mathrm{Horn}}(\mathbf{A},\mathbf{B}) = 1 - \frac{1}{\log(2)} \sum_{i=1}^{S^\star} \left[\frac{p_{\mathbf{A},i}}{2} \log\left(1 + \frac{p_{\mathbf{B},i}}{p_{\mathbf{A},i}}\right) + \frac{p_{\mathbf{B},i}}{2} \log\left(1 + \frac{p_{\mathbf{A},i}}{p_{\mathbf{B},i}}\right) \right] \tag{2.19}$$

Morisita-Horn Similarity Measure

When an index of compositional differentiation robust to undersampling is required, the one-complement of the *Morisita-Horn similarity measure* is among the most favourable ones. Such a similarity index consists of the probability of selecting two individuals of equal species, one from each assemblage, normalised by the mean probability of selecting two individuals of the same species within an assemblage [14].

$$\mathrm{MH}(\mathbf{A},\mathbf{B}) = \frac{2\sum_{i=1}^{S^\star} p_{\mathbf{A},i} p_{\mathbf{B},i}}{\sum_{i=1}^{S^\star} p_{\mathbf{A},i}^2 + \sum_{i=1}^{S^\star} p_{\mathbf{B},i}^2} \tag{2.20}$$

2.5.4 Taxonomy-Based Indices

The former indices of β-diversity ignore taxonomic/phylogenetic diversity between assemblages. When they are of interest, a first approach is to consider branch lengths in a taxonomic tree instead of species tally. In this way, the Jaccard and the Sørensen dissimilarity measures are adapted so that the cardinality $|\mathcal{A}|$ of a set of species is replaced by the total branch lengths in the minimal phylogeny comprising all species in $\mathcal{A}$ up to a desired time interval t [14]. Such quantity is just the *Faith's phylogenetic a-diversity* $L_{\mathbf{A},t}$ of the assemblage $\mathbf{A}$, mentioned in Section 2.4.3.

The phylogenetic version of the Jaccard distance is the unweighted **UniFrac distance** [21], which measures the fraction of branch lengths unique to only one of the two communities being compared on a phylogenetic tree embracing them both [41]. If we let t be the depth of such phylogeny, then the UniFrac metrics is

$$\mathrm{UF}(\mathbf{A},\mathbf{B}) = 1 - \frac{L_{\mathbf{A}\cap\mathbf{B},t}}{L_{\mathbf{A}\cup\mathbf{B},t}} \tag{2.21}$$

Such a taxonomic distance does only consider presence-absence of species in the assemblages. There also exists an abundance-based version of it, which is the **weighted UniFrac** measure [22]:

$$\mathrm{wUF}(\mathbf{A},\mathbf{B}) = \sum_{j=1}^{b} l_j |p_{\mathbf{A},j} - p_{\mathbf{B},j}| \tag{2.22}$$

where b is the number of branches in the tree, l_j is the length of the j-th branch, and $p_{\mathbf{X},j}$ is the fraction of elements in an assemblage $\mathbf{X}$ that are grouped under the taxon relative to branch j. In case the phylogenetic tree is *nonultrametric*, a normalisation to wUF is applied, for which we refer to the original paper introducing the measure [22].

When we come to *metagenomics*, though, the application of the UniFrac measures is to be done with caution: it was originally designed to analyse 16S sequences and only recently its application on Whole Genome Shotgun sequencing (WGS) data has been studied [41].

2.6 Gamma Diversities

When a vast assemblage comprises several smaller assemblages, we differentiate inventory diversity at different scales. **Gamma diversity** is then an *index of how much the wider environment is biologically differentiated.* As priorly introduced, γ-diversity indices ignore how organisms are clustered within the region of interest, but they merely account for species abundance. Thus, all the "scale-insensible" α-diversity indices reported in Section 2.4 are indeed γ-diversity indices when computed at higher scale.

2.6.1 Observations

We believe that the use of mathematically identical indices assuming different names in different context is just misleading. Furthermore, we have also displayed why renaming α-diversity indices to "γ-diversity indices", and retain α-diversity only for average species diversity per samples in a region would be terribly confusing. To make things easier, we have suggested adopting new terms distinguishing *intra-diversity* within samples from *inter-diversity* between samples. A similar division had previously been advanced by naming "diversity" the former, while "dissimilarity" the latter, which however contrasts with the term "β-diversity " itself and is already messy.

However, if we wished to maintain the current terminology, we would rather reserve:

1. α-diversity for plain inventory diversities ignoring any (spatial) assemblage structure. Whether functional diversity could be encompassed in this class or not is questionable.

2. β-diversity for diversities between assemblages. Beta diversity in Whittaker's definition might be renamed as "clustering-diversity" or "clustering-efficiency", instead.

3. γ-diversity for inventory diversities that, somehow, also account for individuals' organisation into cohesive subunits (clusters).

How to classify indices of average α-diversity per compositional unit, ${}^{q}\!D_\alpha$ and S_α, is debatable too.

Obviously, the only need for such a clarification is posed by the many divergent approaches being gathered under the same names α-, β-, and γ-diversity. *In this section we do not intend to suggest such a reorganisation, however: we are merely indicating an alternative construction in order to highlight the need of not mixing up these terms as it is being done.*[21]

A viable approach for measuring γ-diversity in these terms, then, might be the following. Let $G'(V', E')$ be a graph where each node $v \in V'$ represents a compositional unit, of which a quantity c_v is some index of its α-diversity. We advise letting G' be a complete graph, but it need not. For every edge $e = \{u, v\} \in E'$, let $d_e = d(u, v)$ be a β-diversity index — preferably, a metrics taking values in [0,1] — between the two corresponding subunits. Let $\delta'(v) = \{\, e \in E' \colon v \in e \,\}$ be the set of incident edges to v. Lastly, let the weight (summarising importance, contribution) of each compositional unit be its mean β-diversity from other such units:

$$w'_v = \frac{1}{|\delta'(v)|} \sum_{e \in \delta'(v)} d_e$$

[21] We would softly remind of the Genesis' tale of Babel tower: scientists are building a massive tower toward the sky, but once again the are being confused in their language. The problem stands not in diversified lexicons, but rather in divergent semantics within one single vocabulary. Such a divergence seems to be quite inevitable since the immensity of the tower makes tough for two distant builder to keep in touch. While being a brake in science, ambiguity is nevertheless a huge wealth for Human life.

Then we could define a γ-diversity as the sum of α-diversities weighted by their mean β-diversities

$$\Delta'_\gamma \doteq \sum_{v \in V'} \big(w'_v c_v \big) \tag{2.23}$$

Consider a plot **G** composed of two subplots **A** and **B**. If the two subplots are totally distinct, then $d(a, b) = 1$ and, reasonably, $\Delta'_\gamma = c_a + c_b$. On the contrary, suppose the two subplots are identical. Then $c_a = c_b$ and $d(a, b) = 0$, whence $\Delta'_\gamma = 0$, which is irrefutably undesirable.

We could overcome this issue by adding to G' a special node z, holding $c_z = 0$, and edges $\{z, v\}$, $\forall v \in V'$. We set $d_e = 1$, $\forall e \in \{\, \{z, u\} \colon u \in V' \,\}$. Let us call $G(V, E)$ the graph thus built.[22] Then, let

$$\begin{aligned} \delta(v) &= \{\, e \in E \colon v \in e \,\} \\ w_v &= \frac{1}{|\delta(v)|} \sum_{e \in \delta(v)} d_e \\ \Delta_\gamma &\doteq \sum_{v \in V} \big(w_v c_v \big) \end{aligned} \tag{2.24}$$

Consequently, we correctly get $\Delta_\gamma = c_a + c_b + c_z = c_a + c_b = \Delta'_\gamma$ in the first case, and $\Delta_\gamma = {}^{c_a}\!/_2 + {}^{c_b}\!/_2 + c_z = c_a$ in the second one.

Such a γ-diversity index Δ_γ would then fall into the class of multiplicative gamma indices. The difference with the previous definition is that γ-diversity would result from α- and β-diversity, rather than the latter deriving from α- and γ-diversity.

One first difficulty with this approach is probably the physical understanding of such diversities: which units of measure could be used, indeed? Beta-diversity may be a pure number, eventually interpretable as a probability, but alpha and gamma would hold different meanings, yet being commensurable. Notwithstanding, since proposing such measures is not our aim, we believe we can conclude here our introduction.

[22] A physical intuition behind the insertion of this dummy node z is the will to account for the observer viewpoint. It is somehow like the ground from which a stairway arises: alpha-plots are steps of the stairway, but there would be no first step unless a z ground floor existed.

Chapter 3

Tools and Experimental Settings

Our study focuses on the effects of sampling on reference-free metagenomic comparison techniques. The aim is to understand if, under certain conditions, subsampling actually improves results' quality by removing noise from the data, if relevant information is lost instead, or if results do not change consistently as in the case of redundancy removal.

In this work, the assessment of such effects is carried out experimentally by running sequence-composition-based tools for k-mer counting and metagenomic comparison on subsampled metagenomes. Two different subsampling strategies are studied: a k-mers-based sampling implemented in the tool SimkaMin [6] on the one hand, and a more sophisticate reads sampling operated by the tool Spriss [34] on the other hand. As reference, we adopted both a dataset of simulated metagenomes with known taxonomy classification as ground truth, and collections of real metagenomes classified by the reference-based classifier Kraken 2 [46].

A description of the computing environment and of the software exploited for the analysis is provided below in this chapter, whereas datasets and results are presented in Chapter 4. *All the scripts used for the analysis are publicly available* at:
gitlab.com/giorgio.gallina.1/dimReductionEffectsOnMetagenComparison.git

3.1 Running Environment

The experiments have been run on a shared computing cluster of the Information Engineering Department of the University of Padova. It is equipped with several processing nodes, up to 96 available CPUs for each node, and up to 3 TB of RAM available. At the moment of writing, more information about the cluster is available at the page: https://clusterdeiguide.readthedocs.io/en/latest/Overview.html

The software was installed and run inside a *Singularity container* built on February 2023 on the *docker*'s container `ubuntu:jammy`. Its Singularity definition file is retrievable from the main page of the aforementioned git repository.

3.2 Procedure Outline

Given a collection of metagenomic samples[1], and an essential description of it so to be able to asses the results, we proceed as follows:

1. For the collection of *simulated* metagenomes, we computed their metagenomic dissimilarity matrices by using true species abundances.
2. We classified each metagenome by means of KRAKEN 2, and then estimated species abundances using BRACKEN. We then used such abundance vectors to get dissimilarity matrices through the `distance` function from the library `ecodist` in the R language.
3. We ran SIMKAMIN on a grid of sketch sizes and k-mer lengths k to get k-mer-based dissimilarity matrices of the collection obtained by sampling uniformly at random distinct k-mers of the metagenomes.
4. We ran SIMKA to get the exact k-mer-based dissimilarity matrices computed over all available k-mers except singletons, for each of the former k-mer lengths k.
5. We executed SPRISS on a grid of *minimum frequency cut-off* ϑ and k-mer lengths k to get estimated abundances of frequent k-mers, over which we obtained dissimilarity matrices. By doing so we can observe the influence of frequent k-mers in dissimilarity indices more directly.
6. Since, for each metagenome, SPRISS also produces a subsample appropriately designed for its estimation of frequent k-mers, we also used such subsamples to compute dissimilarity matrices on their k-mer composition. We did so by inputting these subsamples to SIMKA. In this way, we observe the effects of reads sampling, besides k-mer sampling.

We developed some shell scripts for automatised execution of the programs with the desired parameter settings, which are available in the `BashScripts` directory within our GitLab repository. The analysis of results has been developed in R language, for which the R scripts are collected into the `Rscripts` folder. Other

[1] In order to avoid confusion, we will reserve the words *collection* and *dataset* for a set of metagenomic samples only, even though it would be reasonable to call "collection" or "dataset" a metagenome as well.

useful scripts for downloading the datasets and for the analyses are stored in the directories `Datasets/Datasets_preparation` and `src`, respectively.

3.3 True Dissimilarities Computation

Simulated metagenomic samples from the Critical Assessment of Metagenome Interpretation (CAMI) initiative [13, 26] do have *true taxonomic classifications* available for benchmarking. Among them, we tested the popular taxonomy from the National Center for Biotechnology Information (NCBI) taxonomy database.

In order to get species dissimilarity matrices between such samples, we used the files mapping reads of each metagenome to their true genome identifiers — i.e., file `$sample_dir/reads/reads_mapping.tsv` in the CAMI folder structure of the dataset — and the mapping from genome identifiers to their relative NCBI taxonomic identifiers — i.e., file `metadata.tsv`. The latter are all specified at *species level*, thus our analysis is carried out at that taxonomic rank. To this goal, we wrote the C++ program `betaDiversity_camitrue.cpp`, publicly available in the directory `src` in our git repository, that maps each read to its species identifier, computes abundance vectors accordingly, and calculates BC and Jaccard dissimilarities on them. It relies on the dataset's files naming and organisations provided by the automatised download gained through the shell script download_CAMI_HMP_bis.bash. Notice that *we only counted one of the two pair ended sequences* to get species abundances of these metagenomes, which actually does not change the results.

3.4 Reference-based Metagenomic Comparison

On real data, for which true species distributions are not available, we used a metagenomic classifier to estimate them. We shall mention, however, that these estimates are heavily *biased* by the availability of reference sequences. Furthermore, as the reference databases grow with more and more genomes added, in combination with the high sequence similarity between species, classification is becoming increasingly less accurate at species level, while improving at the genus level [24]. Nevertheless, up to our knowledge, reference-based techniques are still the only ones assuring results reasonably directly comparable with traditional definitions of β-diversities. Furthermore, since similar environments are better clustered at species level, we only computed reference-based dissimilarity indices at *species level*, except when explicitly stated.

Driven by difficulties in the reference database construction, which followed an heavy modification of the NCBI taxonomy database in 2019, constrained by limitation of resources available to us, and aware of some issues in estimating real abundances from a bare classification, we finally opted for using BRACKEN to solve the latter problem, which relies on KRAKEN 2 for the taxonomic binning. Once the abundances were obtained for each metagenome of a dataset, we run a simple R script, betaDiversity_bracken.R, which reads them, merges them taxon-wise and so exploits the R library `ecodist` for dissimilarity matrices computation.

3.4.1 Kraken 2

Kraken 2 is a *sequence-composition-based* metagenomic classifier, i.e., it uses sequences' k-mers rather than whole sequence alignment for classification. Its workflow is naturally divided into two major blocks: **1.** reference index building, which is done once and needs not be repeated, and **2.** sample classification.

We can sketch it as follows [46]:

1. **Building of the reference database index.** In this step, a hash table is produces. Its *key-value pair* is given by k-mers as keys, and taxonomic identifiers (id) as values.

   (a) A taxonomy (NCBI's by default) and the reference genomes are read.

   (b) An internal taxonomy tree is created starting from the one given at the previous step by: i. pruning nodes which do not correspond to any of the reference genomes and ii. renumbering nodes sequentially.

   (c) For each reference sequence, its k-mers are hashed in two steps:

      i. for each k-mer $\boldsymbol{w}$ of length k, the lexicographically smaller canonical l-mer $\boldsymbol{v}$ of $\boldsymbol{w}$ is extracted, with $l \leq k$; then a spaced seed[2] is applied, allowing s alternate mismatches in the trailing positions of the l-mer, with a match in the last one. For example, if $l = 10$ and $s = 3$, the spaced seed applied is 1111010101. Let us denote the final result by $\boldsymbol{z}$.

      ii. A proper *hash function* h is calculated on the trimmed l-mer $\boldsymbol{z}$.

      Let us define the overall hash function $\eta(\boldsymbol{w}) = h(\boldsymbol{z})$.

   (d) The thus calculated hash value $\eta(\boldsymbol{w})$ is used to index a **compact hash table** of $|T|$ 32 bit cells where to store a taxonomic identifier of the genome to which $\boldsymbol{w}$ belongs. The table size $|T|$ is fixed so to reach an estimated *load factor* of 70%. This process goes as follows:

      - Assume $q < 32$ bits are necessary to store the taxonomic (internal) id value, i.e., there are fewer than 2^q nodes in the internal taxonomy tree. The q least significant bits of a cell are then reserved to accommodate a taxonomic identifier.

[2] A **spaced seed** is a mask, usually represented as a binary string, which tells which positions of characters in a string are to be retained (matched) and which other ones shall be filtered out (mismatched).

- The hash value modulo the size of the table, $\eta(\boldsymbol{w}) \mod |T|$, locates a cell of the table.
- The $\bar{q} = 32 - q$ most significant bits of the hash complete the previous location by indexing $\boldsymbol{w}$ to the first subsequent cell storing the value $\lfloor \eta(\boldsymbol{w})/2^{\bar{q}} \rfloor$ in its $\bar{q}$ most significant bits.
- If, starting from the position $\eta(\boldsymbol{w}) \mod |T|$, an empty cell is found before than a cell storing $\lfloor \eta(\boldsymbol{w})/2^{\bar{q}} \rfloor$ in its indexing part, the lowest-level taxonomic identifier of the genome to which the k-mer $\boldsymbol{w}$ belongs and the most significant part of its hash value are stored accordingly in that cell.
- If a **collision** happens, that is, a cell holding the value $\lfloor \eta(\boldsymbol{w})/2^{\bar{q}} \rfloor$ in its $\bar{q}$ most significant bits is found, the Lowest Common Ancestor (LCA) between the hit taxon and the hitting taxon replaces the former.

2. **Sample classification.** For each read of a metagenomic sample:

   (a) it is decomposed into its k-mers, which are hashed as previously described to query the reference index;

   (b) for each of its k-mers, if a hit occurs, the retrieved taxon is counted in a read-specific pruned tree;

   (c) a k-mer $\boldsymbol{w}$ remains unclassified if, starting from position $\eta(\boldsymbol{w}) \mod |T|$ of the compact hash table and proceeding sequentially, an empty cell is found before than a cell storing the $\lfloor \eta(\boldsymbol{w})/2^{\bar{q}} \rfloor$ part of the hash;

   (d) when all the k-mers of a read are queried, the latter is classified with the leaf node of the maximally scoring root-to-leaf path in its specific tree.

3.4.2 Bracken

Once the reads in a metagenomic sample have been classified via Kraken 2, taxon abundances in it are estimated by Bracken (Bayesian Re-estimation of Abundance after Classification with KrakEN). A naive solution to the problem of estimating abundances at a desired taxonomic level, given a Kraken 2' classification, would retain only reads classified at the taxonomic rank of interest and reject the others; however, such an approach would both ignore important information and give poor estimates. Bracken, instead, undertakes a probabilistic approach to redistribute reads categorised at other taxonomic ranks to the one of interest [23]. In doing so, it takes into account both the Kraken 2' classification strategy and the database queried.

If we denote with $P[G_j(\boldsymbol{r})]$ the probability that a read $\boldsymbol{r}$ is *classified* by Kraken 2 at genus level G_j, and the probability $P[S_i(\boldsymbol{r})]$ that it actually *belongs to* genome S_i, then the probability that a read $\boldsymbol{r}$ classified by Kraken 2 at genus G_j belongs to the genome S_i is expressed by the Bayesian law:

$$P[S_i(\boldsymbol{r})|G_j(\boldsymbol{r})] = \frac{P[G_j(\boldsymbol{r})|S_i(\boldsymbol{r})]\,P[S_i(\boldsymbol{r})]}{P[G_j(\boldsymbol{r})]} \tag{3.1}$$

Given a read $\boldsymbol{r}$ classified at genus G_j in a metagenome $\mathcal{M}$ analysed by Kraken 2 against some defined database $\mathscr{D}$ containing genomes S_i, Bracken estimates the probabilities of Eq. (3.1) as follows [23]:

- $P[G_j(\boldsymbol{r})]$ is clearly unitary.
- To estimate $P[G_j(\boldsymbol{r})|S_i(\boldsymbol{r})]$, roughly writing, we can consider every genome in the database $\mathscr{D}$ being decomposed into its r-mers, with r being the *length of the reads* in the metagenome $\mathcal{M}$. All such r-mers are then pooled together to form an ideal new metagenome $\mathcal{D}$ that is thus classified by Kraken 2 against the index built on the same database $\mathscr{D}$. Finally, for each genome S_i, the probability $P[G_j(\boldsymbol{r})|S_i(\boldsymbol{r})]$ is estimated as the proportion of its reads that has been classifies at genus level G_j.
- To estimate $P[S_i(\boldsymbol{r})]$, Bracken begins with calculating the proportion U_{S_i} of reads of a genome S_i that are uniquely assigned by Kraken 2 to it in $\mathcal{D}$. Consequently, the number R_{S_i} of reads classified at level S_i in $\mathcal{M}$ are adjusted according to the unshared proportion of S_i: $\hat{R}_{S_i} = {}^{R_{S_i}}/_{U_{S_i}}$. Thence, the probability $P[S_i(\boldsymbol{r})]$ is estimated as $P[S_i(\boldsymbol{r})] = \frac{\hat{R}_{S_i}}{\sum_{S_g \in \mathcal{G}_j} \hat{R}_{S_g}}$, where $\mathcal{G}_j$ is the set of all the genomes S_g in the database $\mathscr{D}$ belonging to genus G_j.

Similarly, all the probabilities $P[S_i(\boldsymbol{r})|T_a(\boldsymbol{r})]$ of a read $\boldsymbol{r}$ classified at each taxon level T_a above genus level are computed. Finally, every read classified below species level is summed up to its species and all those reads classified above species level are redistributed to their species level according to the conditional probabilities just computed. Abundances at other taxonomic levels are calculated analogously.

Moreover, Bracken allows users to scale abundances based on the length of their reference sequences. However, for ploidy is still ignored and in line with Sun et al.'s advice [37], we have not taken into account this possibility.

3.4.3 Reference Database

For the analysis, we built up a reference database by means of our shell script `bracken_db_builder.bash` which performs as follows:

- Download taxonomy:
  ```
  kraken2-build --download-taxonomy --db $db_path
  ```
  where `$db_path` is the path to the directory where either *stdDB* or *microDB* is to be stored.

- Download reference genomes:
  `kraken2-build --download-library $lib_name --db $db_path`
  for each library `$lib_name` of interest and with option `--use-ftp` added if necessary. In particular, we set `$lib_name` to `archaea`, `bacteria`, `plasmid`, and `viral`.

- Build Kraken 2 reference index for classification:
  `kraken2-build --build --db $db_path`

In addition, *for each dataset's average read length* `$read_len`, a k-mers distribution index is built for Bracken to estimate abundances:

```
bracken-build -d $db_path -t $ncpus -l $read_len
```

where `$ncpus` is the number of processors available for the computation, which we fixed to 40. All the other settings are left as default, i.e.,

- k-mer length $k = 35$
- minimiser length $l = 31$
- spaced seed number of mismatches $s = 7$
- a low-complexity sequences masker (NCBI's `dustmasker`) is used.

3.4.4 Technical Issues

We installed from the respective git repositories:

- Bracken version 2.8, and
- Kraken 2 version 2.1.2.

However, due to a few bugs of Kraken 2, we had to fix it manually as suggested in the issues section of its GitHub repository. All the changes we made are detailed in the `README` file from the `src` directory of our git repository, where the modified code is collected as well. We also slightly modified the Bracken main script so to terminate rising an informative exit code on errors. An installation script install_krakenSuite_git.bash is available for convenience.

3.5 Simka and SimkaMin for k-mer-based Dissimilarities

In spite of the plain interpretation of reference-based approaches, the probably easiest way of comparing two metagenomic samples is by means of their mere k-mer decomposition. Such an operation is indeed very easy to implement: it just suffices to decompose each read into its k-mers and use their abundances in place of taxa abundances in the formulas. Notwithstanding, this strategy does not elude the requirement of massive computational resources for their computation on huge datasets. When dissimilarity indices are additive, though, efficient parallelised algorithms exist for the task. SIMKA implements one of such algorithms, as described in Section 3.5.1. However, despite allowing computation on a feasible amount of resources, SIMKA is still demanding. Consequently, their developers have provided a version of it which exploits k-mers subsampling in order to reduce dimensionality, that is SIMKAMIN. This is the first subsampling technique we analysed.

Dissimilarity matrices collection is carried out by running the `run_simkamin.bash` shell script. Since both SIMKA and SIMKAMIN must take the complete dataset as input to produce its dissimilarity matrices, an explanation of how to treat uneven sample sizes is deserved. SIMKAMIN does only compute Jaccard and BC dissimilarity matrices, which are also the two most popular indices for metagenomic comparison, hence we focused on these two indices only, and on the BC in particular. This is why *we retained the complete datasets for SIMKA computation*, even in case of uneven metagenome sizes. We have explained in Section 2.5.3, indeed, that the BC index does not make any correction for uneven assemblages sizes, as it makes statistical sense when sampling fraction, rather than sample sizes, is constant. The same holds for SIMKAMIN, even though *a constant number of distinct k-mers is retained*, regardless of the metagenome size. We therefore defined a grid of numbers of distinct k-mers to retain as SIMKAMIN's sampling sizes and thence collected the dissimilarity matrices thus obtained.

Technically, the programs were run with the following settings:

- **Sketch sizes**: 500, 750, 1000, 2500, 5000, 7500, 10000, 25000, 50000, 75000, 100000, 250000, 500000, 750000, 1000000 distinct k-mers. Notice that if in a collection of metagenomes there were fewer distinct k-mers than how many allowed in a SIMKAMIN's sketch[3], then no subsampling would be applied in practice. We have not checked how many distinct k-mers are there in our datasets, however it is possible to verify a posteriori that a subsampling has actually occurred whenever SIMKAMIN's results do not coincide with SIMKA's ones.

- **Singletons** — i.e., k-mers appearing only once in a metagenome — were filtered out both from SIMKA (`-abundance-min 2` option) and SIMKAMIN (`-filter` option). Notice that such a choice is particularly reasonable since: **1.** unique k-mers do often come from sequencing errors, **2.** in this context single k-mers are not even representative the read they come from, and **3.** if

[3] Technically, we correctly use the term *sketch* here, which is a more powerful data summary than a mere *subsample* [32]. Notwithstanding, because in our scope such a distinction is of little relevance and since a sketch is indeed a subsample — whereas the vice versa is false, — we will refer to SIMKAMIN's sketching as to subsampling as well.

the probability of randomly extracting a give singleton in the metagenome it comes from, we can confidently assume the probability of extracting an equal k-mer from other metagenomes to be even smaller.

- One computation node was used, with as many **CPUs** as double the number of metagenomes in the collection, plus one free. For instance, on a dataset comprising 12 metagenomes we required 25 processors and input to the programs the option `-nb-cores 24`.[4]
- Up to 2 TB of **RAM** were reserved for the execution, while a little less was explicitly allowed (`-max-memory 1800000`).

3.5.1 Simka

Simka exploits additivity of many dissimilarity indices to calculate efficiently *exact* dissimilarity matrices. Thanks to additivity, indeed, an abundance vector can be split into chunks, computation performed independently over each chunk simultaneously, and these partial results combined into the final result. Its algorithm is described as follows [5]:

1. By using a same distribution function on canonical representation[5] of each k-mer on each of the N samples independently, k-mers are separated into P partitions, each divided into N chunks, and stored on disk.
2. k-mers are then sorted and counted independently on each of the $P \times N$ sub-partitions.
3. Eventually, low-abundance k-mers, usually generated as sequencing errors, are filtered out.
4. For each of the P partitions, k-mer counts are parallelly merged into k-mer abundance vector. The efficiency of this operation is guaranteed by the previous sorting phase.
5. P independent processes are run in parallel to combine k-mer abundance vectors into P partial additive matrices of crossed terms used in dissimilarity

[4] We shall warn that, since CPU usage was inefficient, we tried reducing the number of processors dedicated to the program, but some issues arose. For example, with option `-nb-cores 13` on the 12-samples HMP dataset the programs failed due to `core dumped` error.

[5] The canonical representation of genetic sequence takes the former in lexicographic order between the sequence itself and its reverse complement, which is the sequence read right-to-left and with each nitrogenous base complemented.

calculation. An example of crossed term is the min calculated in Eq. (2.18) for the BC index.[6]

6. Finally, the P partial matrices are combined together to get the resulting dissimilarity matrix, for each dissimilarity function.

3.5.2 SimkaMin

SimkaMin runs essentially the Simka algorithm, except for adding a threshold on the number of distinct k-mers counted. It does so by implementing a generalised version of the k-minimum values (KMVs) MinHash sketching scheme, which was originally devised for Jaccard distance estimation [32]. More explicitly, in the process of k-mer counting, each k-mer is mapped through a unique hash function and only those holding one of the minimum w of such hash values are counted; for these, a vector of abundances instead of a vector of incidence is computed. Such an operation constitutes the first phase of the SimkaMin algorithm, whereas the second phase is equal to that of Simka [6].

It has been shown that the BC dissimilarity estimated by SimkaMin has both *bias* and *variance* which are linear in the inverse size w of the sketch, hence it is dominated by its *standard deviation* as w increases [6, supplementary data].

3.6 Reads Sampling with Spriss

While SimkaMin takes a *sketching* approach to improve time efficiency, Spriss (SamPling Reads algorIthm to eStimate frequent k-merS) is an efficient algorithm exploiting an intelligent *sampling scheme* to extract frequent k-mers and provide estimates of them with rigorous guarantees [34]. In fact, Spriss implements an approximate solution to the computational problem:

> Given a metagenomic sample $\mathcal{D}$, a positive integer k and a *minimum frequency threshold* $\vartheta \in (0,1]$, find the set $\mathrm{FK}(\mathcal{D}, k, \vartheta)$ of all the k-mers $K \in \Sigma^k$, with alphabet Σ, whose frequency in $\mathcal{D}$ is at least ϑ, and their frequencies.

In order to do so efficiently, Spriss firstly draws a subsample of a metagenome $\mathcal{D}$, comprising $m \times l$ *reads* extracted independently and uniformly at random, with replacement, form $\mathcal{D}$. These reads are conceptually organised into m bags of l *reads* each. The user can choose how many reads they want in each bag by inputting l to the algorithm; however, *in all our experiments we have used the default bag size*

[6] Notice, indeed, that, given the tally of individuals in each sample as constants, the BC similarity index is a linear function of such minima.

$l = \lfloor 0.9/\vartheta g_{\mathcal{D},k} \rfloor$, where, using the same notation as in [34], $g_{\mathcal{D},k}$ is the mean number of k-mers per read in a sample $\mathcal{D}$. On the other hand, m is internally determined as in [34, Proposition 1], except substituting $\ln\left(\frac{1}{\delta}\right)$ with $\ln\left(\frac{2}{\delta}\right)$ as Santoro et al. explain in the cited paper:

$$m = \left\lceil \frac{2}{\varepsilon^2} \left(\frac{1}{l g_{\mathcal{D},k}} \right)^2 \left(\left\lfloor \log_2 \min\left(2 l g_{\max.\mathcal{D},k},\ \sigma^k\right) \right\rfloor \right) + \ln\left(\frac{2}{\delta}\right) \right\rceil \tag{3.2}$$

where $g_{\max.\mathcal{D},k}$ is the maximum number of k-mers per read in $\mathcal{D}$, σ is the cardinality of the alphabet of $\mathcal{D}$, and δ and ε are parameters tuning the quality of the approximation.

Santoro et al. have shown that, by *estimating* adequate statistics on such m bags, it is possible to approximate the set $\mathrm{FK}(\mathcal{D}, k, \vartheta)$ of frequent k-mers with a set A of k-mers and their estimated frequency which, with probability at least $1 - 2\delta$ over the choice of the subsample, contains only k-mers K with frequency $f_{\mathcal{D}}(K) \geq \vartheta - \varepsilon$ and empirically holds a small false negative rate, for fixed frequency threshold ϑ, error parameter $\varepsilon \in (0, \vartheta)$, and confidence parameter $\delta \in (0,1)$. In practice, then, SPRISS outputs an overestimate of the mentioned estimates because it approximates the amount $R[K]$ of reads containing a k-mer K in a metagenome $\mathcal{D}$ with the tally $T[K]$ of occurrences of K in $\mathcal{D}$. Notwithstanding, the overestimation is considerable only for highly repetitive k-mers.

Once SPRISS creates and stores its subsample, it invokes the program `KMC 3` to count canonical k-mers and thus it can easily compute the mentioned statistics.

Technical Details

In all our experiments, we fixed the following parameters as suggested by Santoro et al.:

- $l = \lfloor \frac{0.9}{\vartheta g_{\mathcal{D},k}} \rfloor$ as anticipated;
- $\varepsilon = \vartheta - \frac{2}{t_{\mathcal{D},k}}$, where $t_{\mathcal{D},k}$ is the tally of k-mer in a metagenomic sample $\mathcal{D}$;
- $\delta = 0.1$, implying the result being almost an ε-approximation of $\mathrm{FK}(\mathcal{D}, k, \vartheta)$ with probability at least $1 - 2\delta = 80\%$ in the definitions given in [34].

Tractable values of frequency threshold ϑ are dataset-dependent: indeed, for high ϑ, the list of frequent k-mers could be empty; for low ϑ, on the other hand, the subsample size becomes grater than the original sample size, thus making the approximation senseless.

Furthermore, because the main interest of the authors was to present the algorithm and the sampling scheme, rather than providing a performant and user-friendly software implementing such an algorithm, we had to modify two of the SPRISS' scripts so to meet our needs like, for instance, working with metagenomes in `fasta` format and accommodating available resources. Our modifications are available in the directory `src` of the git repository. The analysis pipeline for some given dataset is available under the name `BashScripts/run_spriss.bash`.

3.6.1 Frequent k-mers Estimates

Together with canonical k-mers representation and their estimated frequency, Spriss outputs their estimated absolute abundances as well. We used such output[7] to compute BC and Jaccard dissimilarity indices in order to observe their behaviour along with the increase of the frequency threshold, i.e., decreasing subsample sizes.

Due to heavy inefficiency of the programming language R, this was totally unsuitable for computing such dissimilarities on large datasets. Therefore, we developed the ad hoc C++ program `betaDiversity_spriss.cpp`, exploiting parallelism similarly to how Simka works, retrievable from the `src` folder in the git repository.

3.6.2 Running Simka on Spriss' Samples

Thanks to Spriss storing subsamples of the metagenomes in the same `fasta/fastq` format of the latter, it was possible to compute the exact k-mer-based dissimilarity indices between such subsamples by means of Simka.

3.7 Examples of k-mers and Reads Sampling Schemes Effects

Let as recapitulate the subsampling techniques presented in this chapter by means of an example, so to better visualise them and interpret results more easily. Suppose we have two samples, **S** and **Q**, for which the distribution of k-mers is:

Sample	A	B	C	D	E	F	G	H	I	J	K	Tot.
S	15	18	11	3	0	1	2	0	1	1	2	54
Q	16	12	6	6	1	1	2	1	0	0	1	46

Due to indices availability through the software we used, we are only focusing on Jaccard and BC dissimilarities, which here hold:

$$\mathrm{B}_{\mathrm{BC}}(\mathbf{S},\mathbf{Q}) = \frac{20}{100} = 0.2, \qquad \mathrm{B}_{\mathrm{Jac}}(\mathbf{S},\mathbf{Q}) = \frac{7}{11} \approx 0.636$$

[7] Using abundances instead of frequencies is in line with the BC index definition, which is not density invariant. k-mers tally of each metagenome is considered informative for k-mer-based BC index computation, indeed.

The first sampling method we have introduced is that implemented in SimkaMin. The latter builds a generalised MinHash *sketch* by retaining a fixed number of *distinct k-mers*, chosen at random, for which the real abundances are annotated. For instance, if we fixed the sketch's size to six k-mers, a possible solution may be the following, where retained k-mers are marked in dark red.

Sample	A	B	C	D	E	F	G	H	I	J	K	Tot.
$\mathbf{S}_{\text{SM}}$	15	18	11	3	0	1	2	0	1	1	2	32
$\mathbf{Q}_{\text{SM}}$	16	12	6	6	1	1	2	1	0	0	1	31

$$\mathrm{B}_{\text{BC}}(\mathbf{S}_{\text{SM}}, \mathbf{Q}_{\text{SM}}) = \frac{11}{63} \approx 0.175, \qquad \mathrm{B}_{\text{Jac}}(\mathbf{S}_{\text{SM}}, \mathbf{Q}_{\text{SM}}) = \frac{4}{6} \approx 0.667$$

Notice that the set of k-mers kept in such a sketch have to be the same across all the samples.

Secondly, the sampling scheme adopted by Spriss extracts reads uniformly at random with replacement. Suppose, for the sake of simplicity, that: **1.** $\mathbf{Q}$ and $\mathbf{S}$ are sampled with equal sampling rate at 50%, and **2.** k-mers are uniformly distributed over reads; then the next table shows an instance of such a subsample.

Sample	A	B	C	D	E	F	G	H	I	J	K	Tot.
$\mathbf{S}_{\text{RS}}$	7	9	6	2	0	0	1	0	1	0	2	28
$\mathbf{Q}_{\text{RS}}$	9	6	3	3	0	0	1	1	0	0	2	25

$$\mathrm{B}_{\text{BC}}(\mathbf{S}_{\text{RS}}, \mathbf{Q}_{\text{RS}}) = \frac{11}{53} \approx 0.208, \qquad \mathrm{B}_{\text{Jac}}(\mathbf{S}_{\text{RS}}, \mathbf{Q}_{\text{RS}}) = \frac{6}{8} = 0.75$$

In practice, however, since compared metagenomes ought to be collected with equal technologies, with the chosen parameters reported in Section 3.6, and with subsample size equal to ml reads for each metagenome, with m given by Eq. (3.2), the *sampling rate* is not constant across the dataset, unless metagenomes are equal in sizes. This is because the only dependence on the sample size is laid by a small contribution of the error parameter ε; therefore, such subsamples happen to be comparable in size even if the original metagenomes were not, and so they hold different sampling rates.

Lastly, if we compare frequent k-mers's estimated abundances like those provided by Spriss, with a threshold $\vartheta = 0.1$ we may have a situation similar to the following:

Sample	A	B	C	D	E	F	G	H	I	J	K	Tot.
$\mathbf{S}_{\text{SF}}$	15	20	10	–	–	–	–	–	–	–	–	45
$\mathbf{Q}_{\text{SF}}$	17	13	6	6	–	–	–	–	–	–	–	42

$$\mathrm{B}_{\text{BC}}(\mathbf{S}_{\text{SF}}, \mathbf{Q}_{\text{SF}}) = \frac{19}{87} \approx 0.213, \qquad \mathrm{B}_{\text{Jac}}(\mathbf{S}_{\text{SF}}, \mathbf{Q}_{\text{SF}}) = \frac{3}{4} = 0.75$$

In the above table we have omitted rare k-mers' abundances because SPRISS actually *selects k-mers* to output by means of a biased frequency estimate, while the *reported frequencies and abundances* are unbiased estimates of such quantities.

3.8 Statistical Analysis

To compare two dissimilarity matrices, we used the `cor.test()` function of R to get their correlation. Since such matrices are symmetric with a 0-diagonal by definition, we only retained the dissimilarities in the upper triangle for the comparison. Let then `x` and `y` be two vectors storing such non-redundant dissimilarities of the first and second matrix, respectively, and let `METHOD` be either the string `‘‘spearman’’` or `‘‘pearson’’`; we obtained the correlation of the two matrices and its *p-value* by running `cor.test(x, y, alternative = ‘‘two.sided’’, method = METHOD)`.

We run a *two-sided* test because we are interested in testing comparability of two matrices rather than testing one be greater than the other. As far as the method is concerned, with the **Pearson correlation** we test linear correlation, whereas with the **Spearman correlation** we observe mutual monotonicity — i.e., linear correlation of the *ranks* of the two vectors.

In order to obtain basic statistics of the metagenomes, we have exploited the R library `ShortRead`, which allows to read `fasta` and `fastq` formatted file and already provides the most essential statistics. Referring to the git repository of the project, the two R scripts performing this job are available in the `Rscripts` folder, named `datasets_glob_statistics.R` and `datasets_kmer_statistics.R`. Both of them are automatically run within the shell script `parameters_setup.bash`, which can be found inside the directory `BashScripts`. For help on their usage, the user can read the top comment in them.

Chapter 4

Analysis of Results

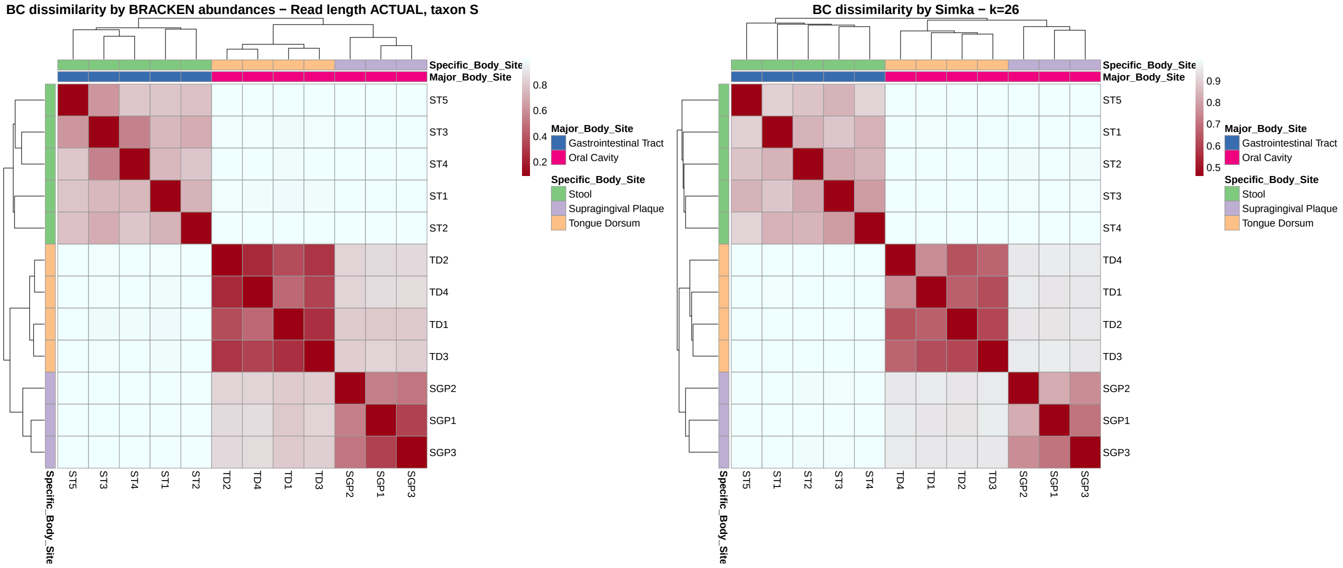
4.1 Datasets Description

In our experiments, we have used three different datasets, which we describe below. All of them were classified by means of the tool Bracken against a reference database comprising *bacterial, archaea, viral and plasmid* sequences, built as explained in Section 3.4.3. Heatmaps of BC dissimilarities on data collections are reported in this section to help visualising their variety; such figures provide a reference for subsequent analysis as well. For ease of reading, tables detailing datasets have been collected in Appendix A.

4.1.1 The Human Metagenome Project (HMP) dataset

We downloaded a collection of 12 large metagenomes from two different body sites of healthy humans: **1.** gastrointestinal tract (stools), and **2.** oral cavity (supra-gingival plaque, tongue dorsum) from the web page www.hmpdacc.org/hmp/HMASM/#data of the HMP. Their selection was carried out so to dispose of large metagenomes of comparable sizes around 10^8 reads (see Tables A.1 and B.2). These samples were collected as 101 bp paired-end reads on the Illumina GAIIx platform [16, Supplementary Information] and reads' alphabet comprises the canonical four nitrogenous bases A, C, T, and G, plus the N character for unknown bases. Human sequences and duplicated reads resulting as artefacts of the sequencing technology were removed and low quality sequences trimmed before their publication (see www.hmpdacc.org/hmp/doc/ReadProcessing_SOP.pdf).

Reads for which their paired-end mate had been excluded by quality filtering were discarded as well. Overall, we exactly mimic the library creation operated by Santoro et al. in [34] on a different set of metagenomes, described in Tables A.1 and A.2.



(A) Species-level Bracken-based BC dissimilarity.

(B) 26-mers-based BC dissimilarity.

Figure 4.1: BC dissimilarities on the **HMP collection**.

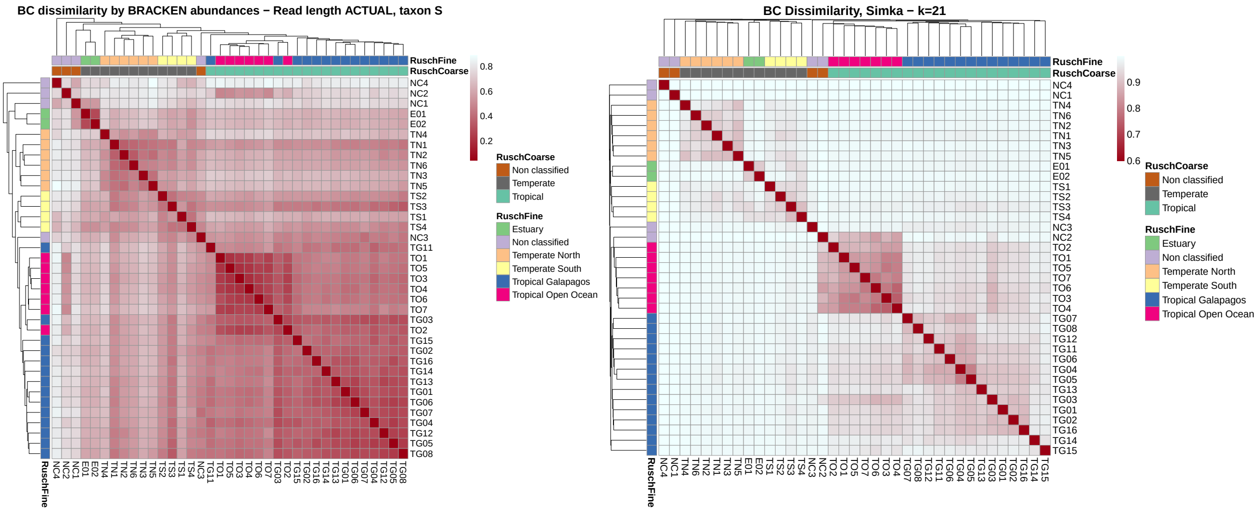
4.1.2 The Global Ocean Sampling expedition (GOS) dataset

A popular metagenomic dataset comes from the Global Ocean Sampling Expedition. This is a collection of small oceanic metagenomes taken between 2004 and 2006 and sequenced by means of gel electrophoresis [33]. Of these, excluding those from the Sargasso Sea, we considered the 37 samples analysed by Rusch et al., and reported in [34] as well. These samples are rather heterogeneous in size, with up to one order of magnitude of difference, and are composed of sequences of average length ≈ 1070 bp. See Table A.4 for further details.

Due to its scarcity and heterogeneity, this is a challenging dataset to analyse. Moreover, our reference database is probably insufficient for its classification: eukaryotic sequence, indeed, are likely to comprise part of its metagenomes because of the presence of algae and planktons. Unfortunately, we disposed of insufficient resources for using a larger reference database, yet the results we got are quite consistent.

4.1.3 The CAMI II Toy Human Metagenome Project dataset

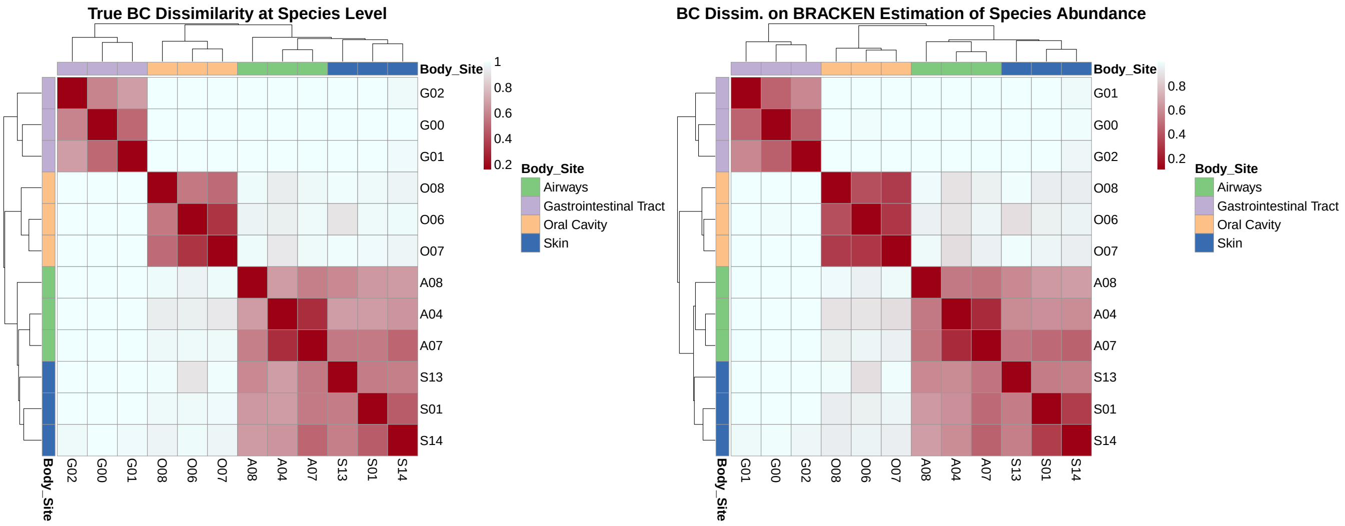
Common practice in empirical studies is to compare results with known *ground truth* to establish their trustfulness. For the purpose, we applied our analysis pipeline to a dataset of 12 synthetic WGS short-read metagenomes provided by the CAMI initiative [13], simulating HMP samples. These metagenomes have approximately equal size and all their paired-end reads are 150 bp long; however, their alphabet is not limited to the four canonical nucleotides plus the **N** (see Table A.7). Further properties of this dataset are listed in Table A.6.



(A) Species-level BRACKEN-based BC dissimilarity.

(B) 21-mers-based BC dissimilarity.

FIGURE 4.2: BC dissimilarities on the **GOS collection**.



(A) Species-leve true BC dissimilarity

(B) Species-level BRACKEN-based BC dissimilarity.

FIGURE 4.3: BC dissimilarities at species level on the **CAMI collection**.

Since a ground truth is known only on the CAMI dataset, this is our main collection of reference. Nevertheless, because effects of subsampling depend on the distribution of the population[1] being subsampled, we have also analysed the two real datasets just introduced. Indeed, even the real HMP collection we used is rather different from the simulated human metagenomes. For confirmation, we compared them by means of their k-mer-based BC index computed by SIMKA on the two datasets pooled together (Fig. 4.4).

Ideally, this would provide an unbiased vision of the diversity between the collections since their *sequences composition* is the only information used. In reality, however, the two datasets are not genuinely comparable due to their different metagenomes' sizes and reads length. To mitigate the issue, we exploited `-max-reads 0` option of SIMKA for reading an equal number of reads from each metagenome and we scaled crossed-dissimilarities so to range from 0 to 1: $\mathrm{B}_{\mathrm{BC}}(\mathbf{R},\mathbf{S})_{\mathrm{norm}} = 1 - \frac{5}{4}(1 - \mathrm{B}_{\mathrm{BC}}(\mathbf{R},\mathbf{S}))$ for each real metagenome $\mathbf{R}$ versus each simulated metagenome $\mathbf{S}$.[2]

In spite of these comparability issues, the results show with little doubt that the two human datasets have divergent k-mer distributions.

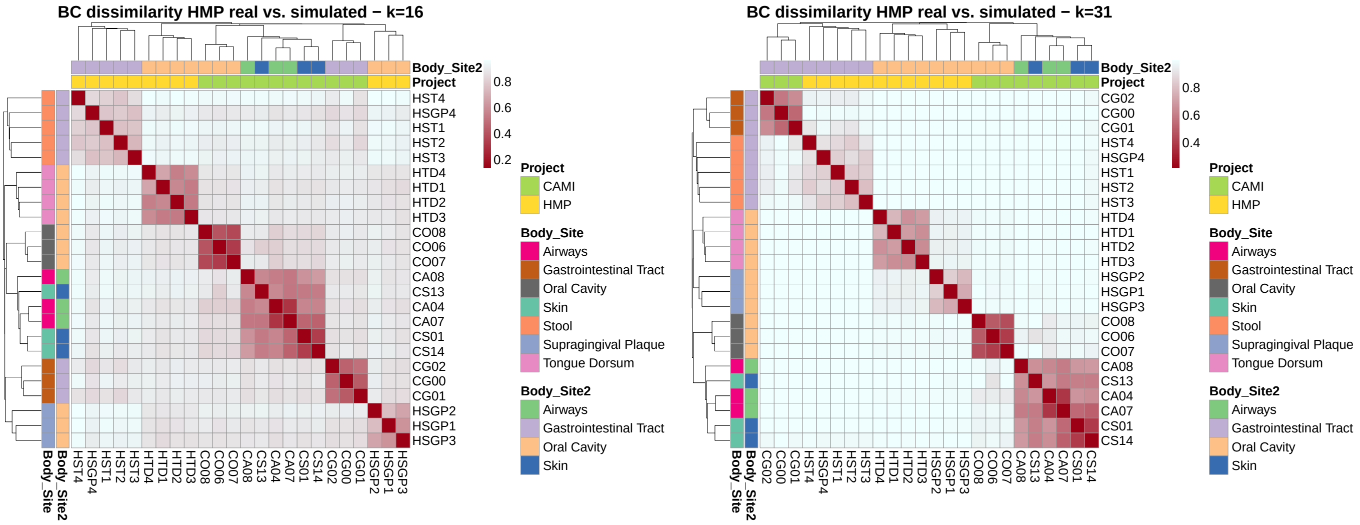


FIGURE 4.4: Comparison of real (HMP) versus simulated (CAMI) metagenomic samples via BC dissimilarity indices. Indices have been obtained by running SIMKA on the pooled dataset with automatic limit on the number of reads (option `-max-reads 0`) and with post-processing normalisation of crossed-dissimilarities as explained in the text.

[1] We adopt here the statistical meaning of "population".

[2] The motivation of such a form of normalisation is that with equal number of reads retained and with read length being 150 bp for CAMI dataset whereas ≈ 100 bp for the HMP, in case of identical k-mers distribution among two samples, their BC *similarity* would be equal to 4/5.

4.2 SimkaMin Subsampling Effects

In this and the following studies, we began by observing how well k-mer-based β-diversity indices computed on subsampled metagenomes correlate with their respective reference-based indices obtained from the whole metagenome. In order to see if subsampling brings either an improvement or a worsening on results, indeed, it just suffices to observe if such correlations decrease or increase, respectively, with the size of the subsample. We ran these kind of analysis for different k-mer lengths.

As mentioned in Section 3.8, we compared both Pearson and Spearman correlations, since both of them carry relevant information. A high Pearson correlation indicates a strong linear relation between two variables. Consequently, it is reasonable to expect that two dissimilarity matrices highly correlated in Pearson's sense would hold similar clustering at higher levels of the clustering tree. A high Spearman correlation, on the other hand, being a *rank* correlation and thus indicating mutual monotonicity of two variables, is more likely to produce similar clustering at low level of the tree. The main reason for considering Spearman correlation, though, is that k-mer-based and taxonomy-based dissimilarities are not linearly related in principle, yet we are interested in their discriminative power, which is captured by ranking dissimilarities.

We believe that, to our knowledge, this is the most proper way of assessing subsampling effects in our scenario because, in light of a recent study [37], watching at bare dissimilarities might be doubly misleading: **1.** it could lead to failures in interpretation of results, and **2.** it may give the impression that k-mer-based dissimilarities are *good estimates* of reference-based ones, whereas they are only *good substitutes*.

Baring these warnings in mind, it is nonetheless convenient to actually see how dissimilarity measures behave directly to better understand the former results. We therefore analyse their trend in absolute value by means of scatter plots (sampled k-mer-based versus complete reference-based dissimilarities) and line plots of the dissimilarity against sampling size. Due to the considerable amount of data we produced, we have selected a small representative subset of them. The complete set of results is retrievable through GitLab at the `Results` directory.

4.2.1 Experiments on Bray-Curtis Dissimilarity

We firstly analyse the behaviour of SimkaMin's sketched Bray-Curtis dissimilarity compared to ground truth in the **simulated human dataset** (Figs. 4.5a and 4.5b). Both their Pearson and Spearman correlations reveal that, no matter which k-mer length $k \in \{ 21, 26, 31 \}$, such dissimilarity indices computed on sketches comprising more than 10^5 distinct k-mers chosen at random according to the MinHash sketching technique are barely distinguishable from one another. All of them, indeed, appear to be constant along with increasing sketching size up to non-subsampled metagenomes.

Moreover, they are also extremely high — i.e., neatly above 97%, — vouching validity of sequence-composition reference-free methods for measuring metagenomic diversity (Table 4.1). Tighter subsampling generally worsen results, thus leaving out relevant information. Overall, Fig. 4.8 confirms these considerations.

When 16-mers were used, however, a clear decay is manifest. Such a phenomenon deserves further investigation for an explanation, yet we suppose a combination of metagenomes size and the true distribution and phylogenetic diversification of species to be its leading cause. Such hypothesis is corroborated by Fig. 4.8a, where we can see the drop in correlation being brought by pairs of highly dissimilar metagenomes that hold milder 16-mers-based indices. This suggests, indeed, that 16-mers are too short for metagenomic comparison at species level on this collection. Hence, It would be interesting to study how these k-mer-based dissimilarities correlate with truth at genus level, instead.

Bracken-based BC dissimilarity also correlates very well with truth on the CAMI collection, although it generally underestimates BC dissimilarity proportionally to true BC similarity. This allows us to rely on the former for assessment of real datasets. Of these, the collection from the **GOS** expedition shows a similar behaviour to the simulated dataset in spite of the huge diversity of the two datasets. In these metagenomes, however, the highest linear correlation between reference-free and reference-based indices is reached using 16-mers and it is markedly lower than the previously analysed ones, although being still high (Fig. 4.7 and Table 4.3). Spearman correlation peaks on 21-mers, which hold the second-best Pearson correlation. Interestingly, these also provide the best clustering of metagenomes according to the classification proposed in [33], which is even better than the reference-based one, as it can be visualised in Fig. 4.2. Therefore, we shall avoid lucubrating on lower or higher values of correlation between k-mer-based and reference-based metagenomic comparison on this dataset. Nevertheless, it is evident that moderate sketching effort have practically no impact.

More peculiar is the **HMP dataset**. Here Pearson correlations sticks to our previous observations but for a slightly clearer differentiation between indices based on different k-mer lengths, which was not perceivable in the CAMI dataset. 21-mers provided the highest linear correlation with Bracken's results at species level. Outstanding is the appearance of Spearman correlations, instead. These are markedly lower than their respective Pearson correlations, they are rather fluctuating in function of sketch size, and 21-mers hold the worst results. Figure 4.9 helps understanding this plot. We see in Fig. 4.9c, indeed, a high density of maximum dissimilarity pairs, which is perfectly desirable, yet the ranking strategy underlying Spearman correlation fails to stand up to this scenario.

To sum up, all the three datasets analysed showed a similar behaviour on the SimkaMin sketching technique: as expected [6, Supplementary file], a reasonably small amount of distinct k-mers was just enough to estimate with very high precision k-mer-based Bray-Curtis dissimilarity index. In light of the example given in

Section 3.7, we can intuitively explain this result as follows. Suppose we ordered k-mers abundances and showed their distribution in a bar-plot, and then we randomly removed some of them leaving the others unchanged, then the shape drawn by the second bars would be similar to the original one, except only for steepness. However, it is impossible to assess neither an improvement nor a worsening along with light subsampling: just redundant information is cut off. With heavier subsampling, instead, an obvious worsening of results becomes manifest.

4.2.2 Experiments on the Jaccard Distance

We repeated the former analysis to study Jaccard distance's behaviour on increasingly intensive subsampling effort (Fig. 4.11a).However, from the very beginning, this index appears to be much less linearly correlated in its k-mer form compared to the true taxonomic one at species level. This becomes incontrovertible in Fig. 4.11c, where plots' line do not respect colour stratification at all on weakly dissimilar metagenomes pairs. Notwithstanding, we may notice that, as shorter k-mers provide Jaccard distances more linearly correlated with the true ones, such correlation slightly increases with heavy sketching effort on longer k-mers. This suggests that a k-mer-based version of the Jaccard distance should not rely on too much discriminative information.

Spearman correlation between true and SimkaMin's Jaccard distances, on the contrary, is high and even higher than the that with reference-based Bracken Jaccard index. Nevertheless, on the basis of previous considerations and the wealth of plots we have drawn, we reckon it cannot be said whether SimkaMin's sketching scheme removes either noise or information in this scenario. As before, however, a moderate subsampling have nearly no impact on results, although it cannot be said that only redundancy is left out.

Table 4.1: Correlations between BC dissimilarities computed on SimkaMin sketches and the true ones on **CAMI simulated dataset**. Sketches' correlations are reported as average of those with size indicated in column 2 with standard deviation in error notation.

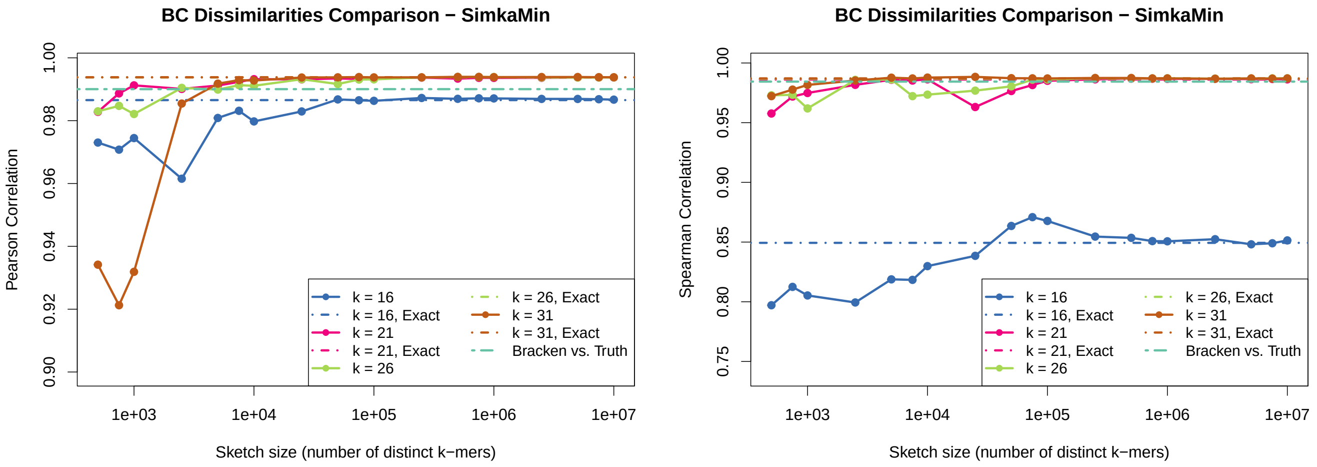
k	Sketch size	Pearson cor.	Pearson p-value	Spearman cor.	Spearman p-value
16	>=1e+05	0.9869(3)	$2(2) \times 10^{-52}$	0.853(6)	$1.3(8) \times 10^{-19}$
16	unsampled	0.987	4.66×10^{-52}	0.849	2.02×10^{-19}
21	>=1e+05	0.9936(1)	$3(2) \times 10^{-62}$	0.9864(6)	$2(4) \times 10^{-51}$
21	unsampled	0.994	1.06×10^{-62}	0.986	1.26×10^{-51}
26	>=1e+05	0.9937(2)	$3(7) \times 10^{-62}$	0.9870(3)	$3(3) \times 10^{-52}$
26	unsampled	0.994	9.82×10^{-63}	0.987	2.87×10^{-52}
31	>=1e+05	0.993 86(6)	$7(2) \times 10^{-63}$	0.9871(2)	$1.4(7) \times 10^{-52}$
31	unsampled	0.994	7.96×10^{-63}	0.987	1.24×10^{-52}
Bracken		0.9900	3.32×10^{-56}	0.9844	5.16×10^{-50}

Table 4.2: Correlations between BC dissimilarities computed on SimkaMin sketches and those Bracken-based in **HMP dataset**. See Table 4.1 for table structure.

k	Sketch size	Pearson cor.	Pearson p-value	Spearman cor.	Spearman p-value
16	>=1e+05	0.9822(6)	$5(5) \times 10^{-48}$	0.960(4)	0
16	unsampled	0.982	2.18×10^{-48}	0.96	0
21	>=1e+05	0.9853(4)	$1(1) \times 10^{-50}$	0.90(2)	$4(10) \times 10^{-27}$
21	unsampled	0.985	7.86×10^{-51}	0.902	0
26	>=1e+05	0.9826(8)	$7(20) \times 10^{-48}$	0.909(7)	0
26	unsampled	0.983	6.3×10^{-49}	0.913	1.5×10^{-26}
31	>=1e+05	0.9810(3)	$3(1) \times 10^{-47}$	0.913(7)	$7(20) \times 10^{-27}$
31	unsampled	0.981	4.92×10^{-47}	0.915	0

Table 4.3: Correlations between BC dissimilarities computed on SimkaMin sketches and those Bracken-based in **GOS dataset**. See Table 4.1 for table structure.

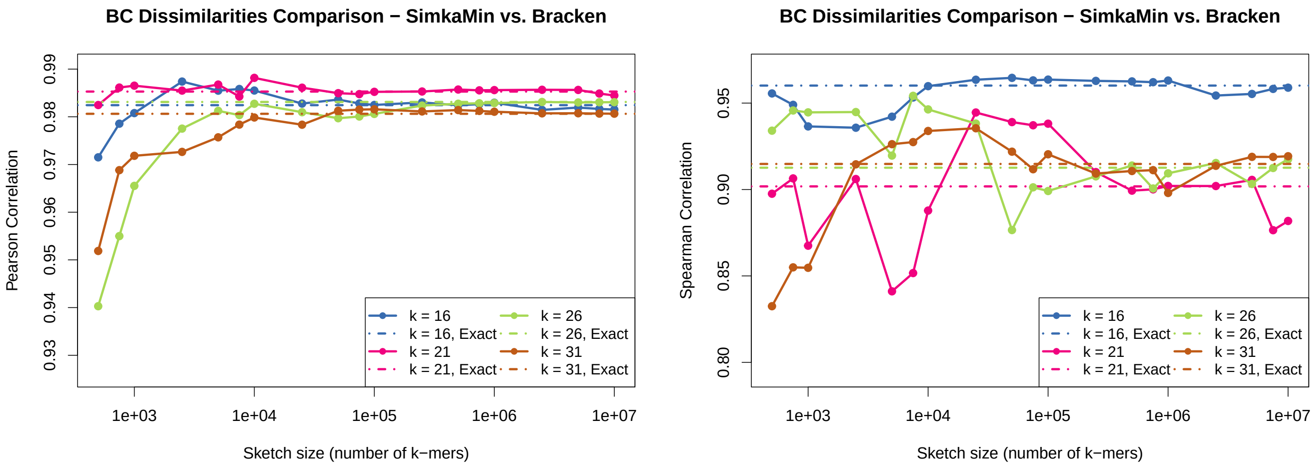
k	Sketch size	Pearson cor.	Pearson p-value	Spearman cor.	Spearman p-value
12	>=1e+05	0.446(1)	$7(3) \times 10^{-34}$	0.4268(7)	$5(4) \times 10^{-31}$
12	unsampled	0.447	5.38×10^{-34}	0.427	0
14	>=1e+05	0.692(1)	$6(6) \times 10^{-96}$	0.665(2)	$2(3) \times 10^{-86}$
14	unsampled	0.694	1.16×10^{-96}	0.666	0
16	>=1e+05	0.8389(3)	1.66×10^{-177}	0.8540(5)	1.77×10^{-190}
16	unsampled	0.839	6.16×10^{-178}	0.855	0
21	>=1e+05	0.7490(3)	$9(3) \times 10^{-121}$	0.9014(9)	7×10^{-242}
21	unsampled	0.749	9.83×10^{-121}	0.902	9.91×10^{-245}
26	>=1e+05	0.7103(7)	$3(2) \times 10^{-103}$	0.8898(9)	1.77×10^{-227}
26	unsampled	0.711	1.38×10^{-103}	0.89	1.52×10^{-228}
31	>=1e+05	0.687(1)	$6(10) \times 10^{-94}$	0.878(2)	1.15×10^{-209}
31	unsampled	0.687	3.79×10^{-94}	0.879	9.07×10^{-216}



(a) Pearson correlation, CAMI dataset

(b) Spearman correlation, CAMI dataset

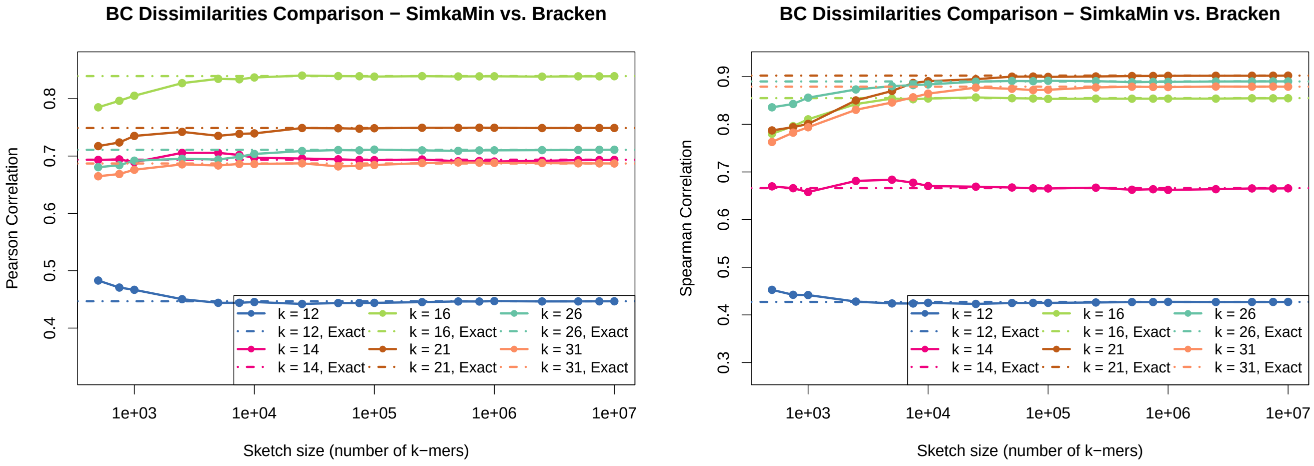
Figure 4.5: Correlation between BC indices computed by SimkaMin and the true species-level BC dissimilarity on the **CAMI dataset**, varying sampling effort. Correlation between the true dissimilarities and that calculated on Bracken results is reported too.



(A) Pearson correlation, HMP dataset

(B) Spearman correlation, HMP dataset

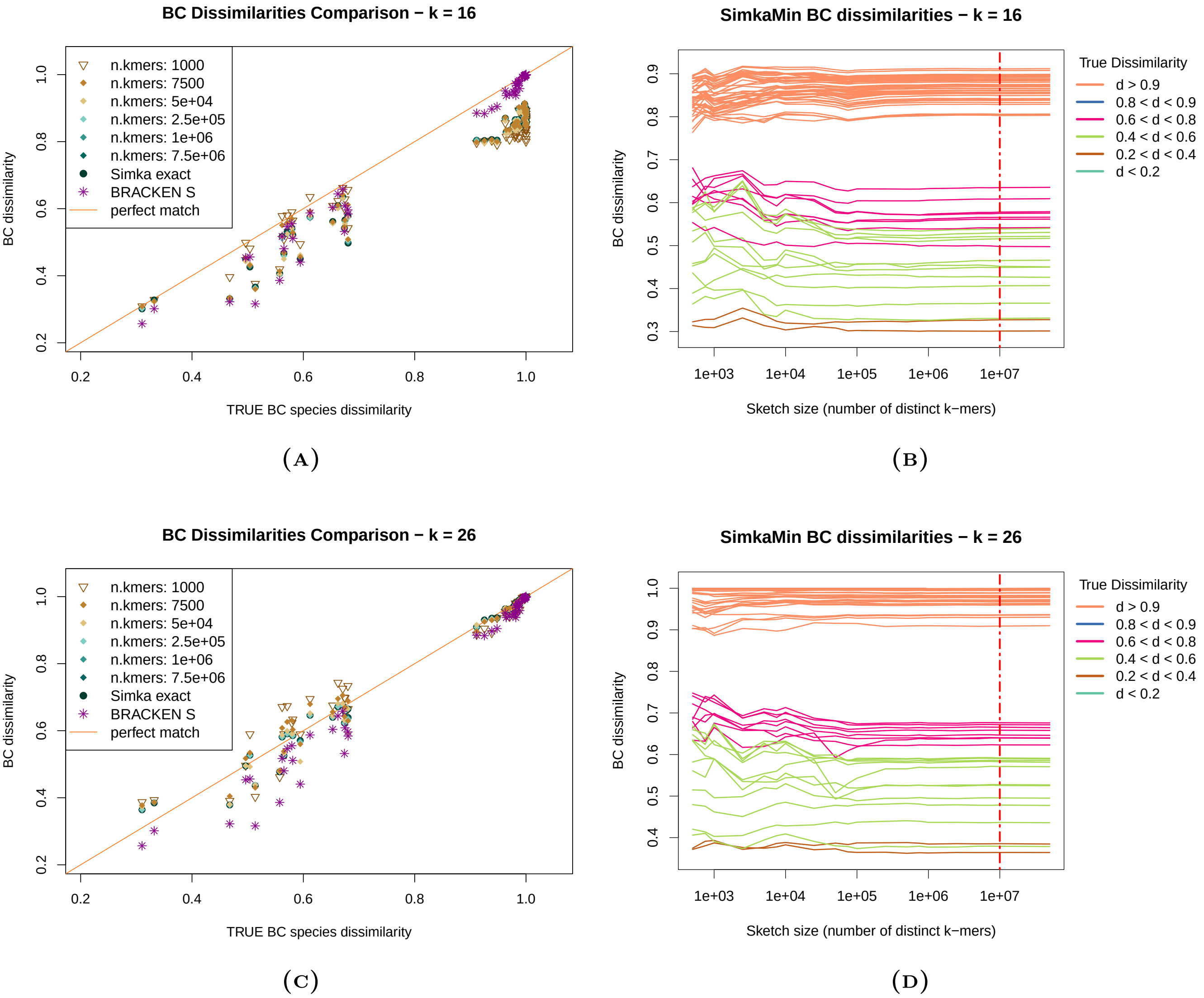
FIGURE 4.6: Correlation between BC indices computed by SIMKAMIN and the species-level BRACKEN-based BC dissimilarity on the **HMP dataset**, varying sampling effort.



(A) Pearson correlation, GOS dataset

(B) Spearman correlation, GOS dataset

FIGURE 4.7: Correlation between BC indices computed by SIMKAMIN and the species-level BRACKEN-based BC dissimilarity on the **GOS dataset**, varying sampling effort.



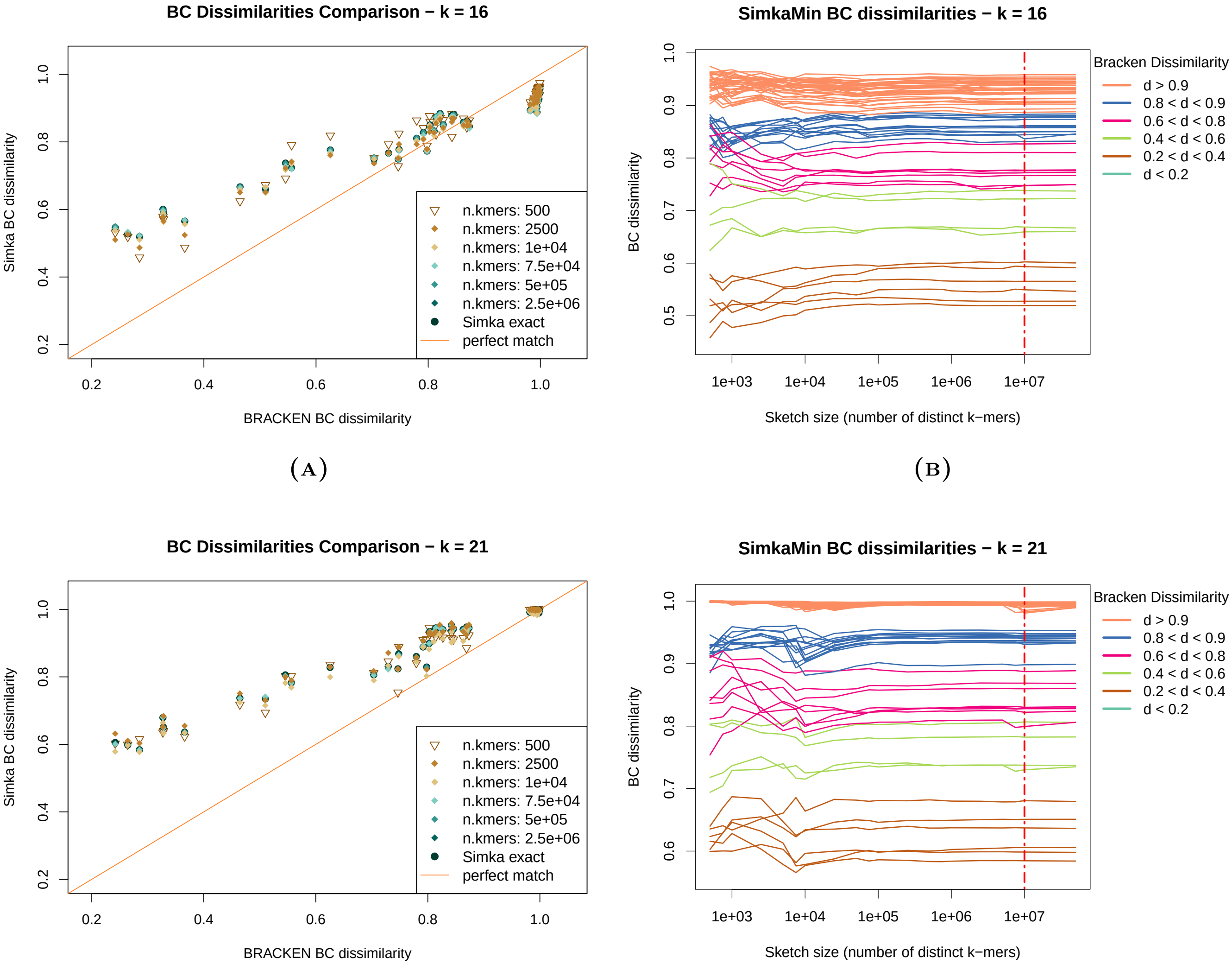
(a)

(b)

(c)

(d)

Figure 4.8: Pairwise SimkaMin and true BC dissimilarities compared on the simulated **CAMI dataset**. On the right, dissimilarity values of each pair of metagenomes are reported varying sketch's size. Each line correspond to one such pair, and colouring is based on reference dissimilarity as shown in legend, i.e., k-mer-based dissimilarities should ideally respect the colour legend at least in stratification. The highest sketch's size is met in correspondence with the vertical dashed red line, after which dissimilarity without subsampling is reported at the extreme right.



(A)

(B)

(C)

(D)

Figure 4.9: Pairwise SimkaMin and Bracken BC dissimilarities compared on the **HMP dataset**. On the right, the highest sketch's size is met in correspondence with the vertical dashed red line, after which dissimilarity without subsampling is reported at the extreme right.

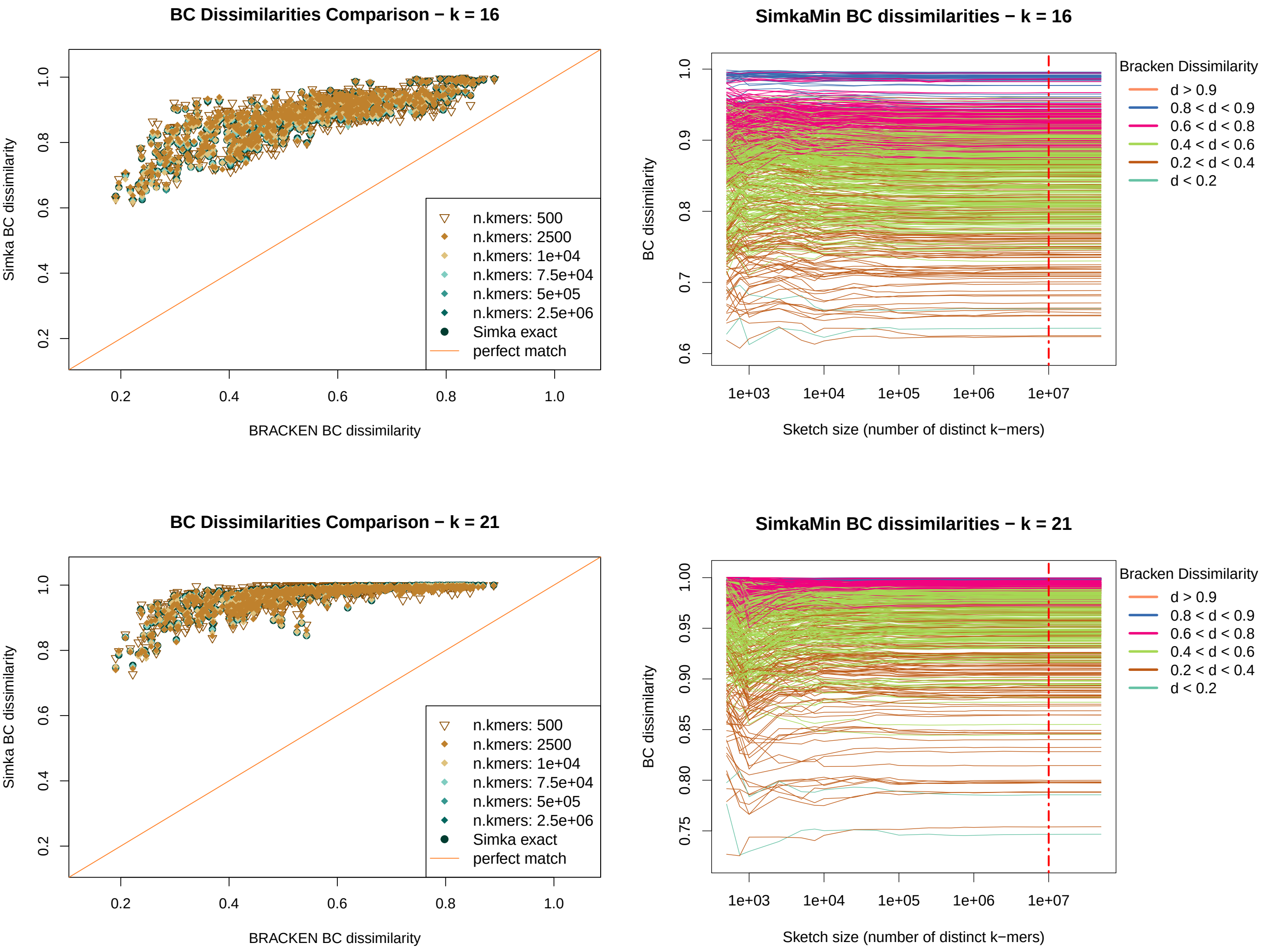
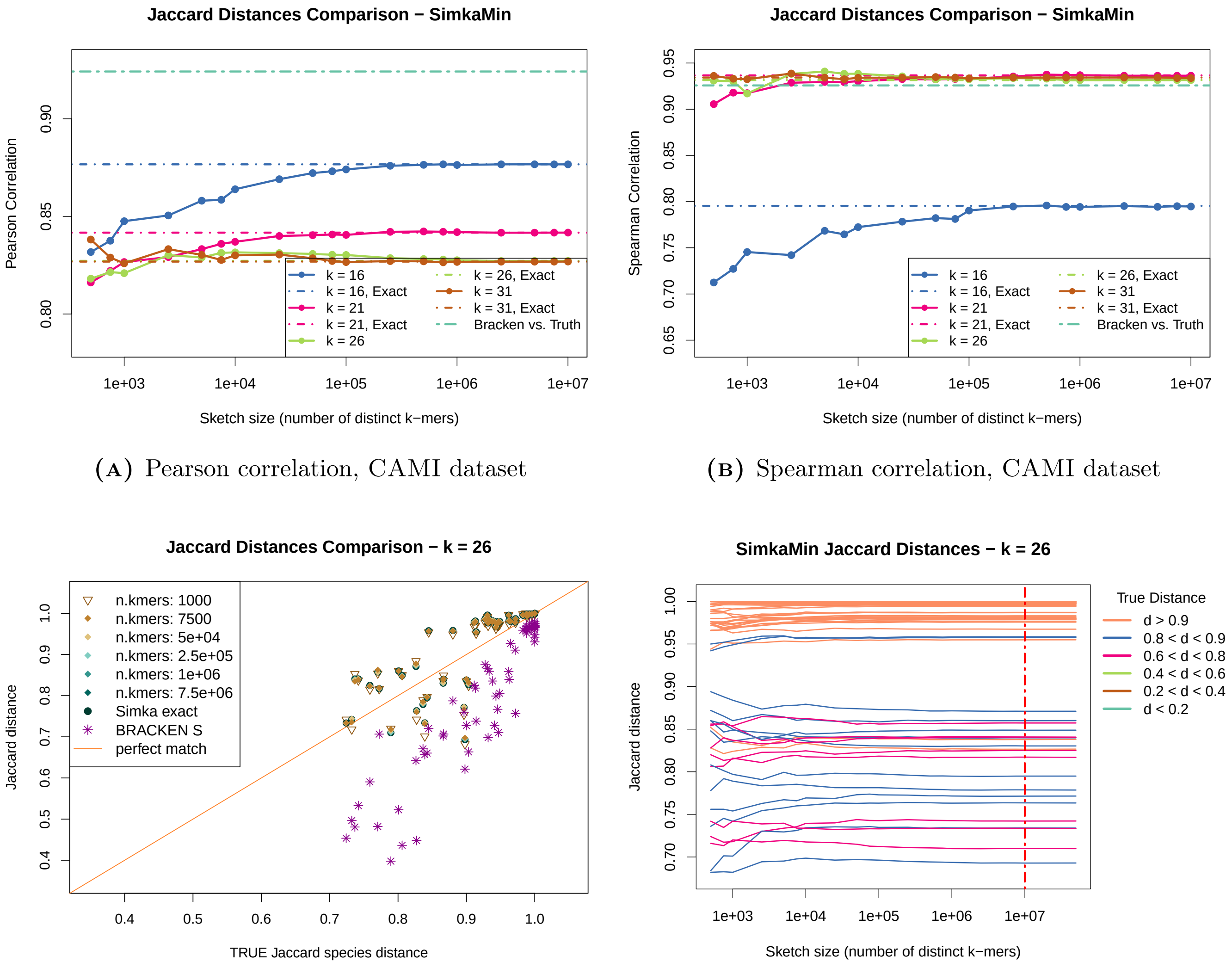


Figure 4.10: Pairwise SimkaMin and Bracken BC dissimilarities compared on the **GOS dataset**, with k-mer lengths which provide the best correlation between the two types of dissimilarities. On the right, the highest sketch's size is met in correspondence with the vertical dashed red line, after which dissimilarity without subsampling is reported at the extreme right.



(A) Pearson correlation, CAMI dataset

(B) Spearman correlation, CAMI dataset

(C) SimkaMin versus truth with $k = 26$

(D) SimkaMin on increasing subsample size, $k = 26$

Figure 4.11: SimkaMin estimates of Jaccard distance compared with true species Jaccard index on the **CAMI dataset**. On the bottom-right, the highest sketch's size is met in correspondence with the vertical dashed red line, after which dissimilarity without subsampling is reported at the extreme right. Only plots on 26-mers are reported for ease of reading since other k-mer lengths hold very similar results. However, all plots and results can be downloaded from the aforementioned git repository if of interest for the reader.

4.3 Spriss Subsampling Effects

Instead of sampling randomly detected distinct k-mers, Spriss *samples reads uniformly at random with replacement.* We can thus expect different results because the probability mass would probably shift from rare k-mers, which tend to disappear, to frequent ones.

By fixing error and confidence parameters, subsample sizes are controlled by the *minimum frequency threshold* ϑ: higher such thresholds hold smaller subsamples. To better understand the actual size of subsamples, then, we report in Tables 4.4 to 4.6 and Tables B.1 to B.3 in Appendix B a summary of average actual subsample sizes per dataset. Importantly, notice that replacement in the sampling scheme implies that a sampling rate of 100% does not mean all available data is retained.

The choice of reasonable values of ϑ depends on the collections through their original size [29]. At too high thresholds, indeed, subsamples would always be produced, yet the set of frequent k-mers may be empty. Too low thresholds, on the flip side, require subsamples of larger size than the originally available data; these are therefore not produced. These phenomena have great impact on small heterogeneous collections, like the GOS dataset, which led us to reduce it to its 26 largest metagenomes, listed in Table A.8 in Appendix A.

Table 4.4: Spriss' sampling rates on **CAMI dataset**. Rows are minimum frequency thresholds inputted to Spriss, columns refer to different k-mer lengths. Sampling rates are expressed as number of k-mers in the subsample divided by the number of k-mers available in the complete metagenome, which equals the rate of sampled reads since all reads have equal length. The values displayed are taken as arithmetic mean of all sampling rates on the dataset, with their standard deviation expressed as an error in parenthesis. See also Table B.1 in Appendix B for more details on sampling sizes when 26-mers are used.

ϑ	$k = 16$	$k = 21$	$k = 26$	$k = 31$
$\mathbf{1.00 \times 10^{-6}}$	0.011 401(2)	0.011 840(2)	0.012 314(2)	0.012 827(2)
$\mathbf{5.00 \times 10^{-7}}$	0.024 003(5)	0.024 927(5)	0.025 924(5)	0.027 004(5)
$\mathbf{3.00 \times 10^{-7}}$	0.041 339(8)	0.042 928(8)	0.044 647(9)	0.046 507(9)
$\mathbf{2.00 \times 10^{-7}}$	0.065 01(1)	0.067 51(1)	0.070 21(1)	0.073 14(1)
$\mathbf{1.50 \times 10^{-7}}$	0.086 68(2)	0.090 01(2)	0.093 61(2)	0.097 51(2)
$\mathbf{1.25 \times 10^{-7}}$	0.104 02(2)	0.108 02(2)	0.112 34(2)	0.117 02(2)
$\mathbf{1.00 \times 10^{-7}}$	0.136 02(3)	0.141 25(3)	0.146 90(3)	0.153 02(3)
$\mathbf{7.50 \times 10^{-8}}$	0.181 36(3)	0.188 34(4)	0.195 87(4)	0.204 03(4)
$\mathbf{5.00 \times 10^{-8}}$	0.284 04(5)	0.294 97(6)	0.306 77(6)	0.319 55(6)
$\mathbf{3.50 \times 10^{-8}}$	0.405 78(8)	0.421 38(8)	0.444 41(8)	0.462 93(9)
$\mathbf{3.00 \times 10^{-8}}$	0.480 07(9)	0.4985(1)	0.5185(1)	0.5401(1)
$\mathbf{2.50 \times 10^{-8}}$	0.6001(1)	0.6232(1)	0.6481(1)	0.6751(1)
$\mathbf{2.00 \times 10^{-8}}$	0.7501(1)	0.7893(2)	0.8209(2)	0.8551(2)

Table 4.5: Spriss' sampling rates on **HMP dataset**. Rows are minimum frequency thresholds inputted to Spriss, columns refer to different k-mer lengths. Sampling rates are expressed as number of k-mers in the subsample divided by the number of k-mers available in the complete metagenome. The values displayed are taken as arithmetic mean of all sampling rates on the dataset, with their standard deviation expressed as an error in parenthesis. See also Table B.2 in Appendix B for more details on sampling sizes when 26-mers are used.

ϑ	$k = 16$	$k = 21$	$k = 26$	$k = 31$
$\mathbf{1.00 \times 10^{-6}}$	0.0060(6)	0.0064(7)	0.0068(7)	0.0073(8)
$\mathbf{5.00 \times 10^{-7}}$	0.013(1)	0.013(1)	0.014(2)	0.015(2)
$\mathbf{3.00 \times 10^{-7}}$	0.022(2)	0.023(2)	0.025(3)	0.027(3)
$\mathbf{2.00 \times 10^{-7}}$	0.034(4)	0.036(4)	0.039(4)	0.042(5)
$\mathbf{1.50 \times 10^{-7}}$	0.045(5)	0.048(5)	0.052(6)	0.056(6)
$\mathbf{1.25 \times 10^{-7}}$	0.054(6)	0.058(6)	0.062(7)	0.067(7)
$\mathbf{1.00 \times 10^{-7}}$	0.070(8)	0.075(9)	0.081(9)	0.09(1)
$\mathbf{7.50 \times 10^{-8}}$	0.09(1)	0.10(1)	0.11(1)	0.12(1)
$\mathbf{5.00 \times 10^{-8}}$	0.15(2)	0.16(2)	0.17(2)	0.18(2)
$\mathbf{3.50 \times 10^{-8}}$	0.21(2)	0.23(3)	0.24(3)	0.26(3)
$\mathbf{3.00 \times 10^{-8}}$	0.25(3)	0.26(3)	0.28(3)	0.31(4)
$\mathbf{2.50 \times 10^{-8}}$	0.31(3)	0.33(4)	0.35(4)	0.38(4)
$\mathbf{2.00 \times 10^{-8}}$	0.39(4)	0.41(4)	0.44(5)	0.48(5)
$\mathbf{1.50 \times 10^{-8}}$	0.52(6)	0.56(6)	0.59(7)	0.65(8)
$\mathbf{1.20 \times 10^{-8}}$	0.68(8)	0.72(8)	0.77(9)	0.8(1)

4.3.1 BC Dissimilarity on Spriss Subsamples

In this set of experiments, we sample reads with Spriss and then compute the k-mer frequency in the sampled set with Simka. When directly sampling k-mers (see Section 4.2), 16-mers proved to be inadequate for comparing *large human datasets*; here they do perform better, yet worse than longer k-mers (see Fig. 4.13). Hence we focus on the latter, which behave similarly to one another. For convenience, results obtained on 26-mers are displayed here, while figures of 21-mers and 31-mers are reported in Appendix C.

On the **CAMI dataset**, linear correlation between k-mer-based and true BC dissimilarities approaches unity on high subsampling rates (Figs. 4.12 and 4.13). In fact, high values of such a correlation have been reported in all our results, as far as proper k-mer lengths were used. We shall therefore notice that this is not guaranteed a priori, in spite of observations. Were we to explain such a behaviour of the two different BC indices, by considering that they can be expressed as additive functions of abundances [6], we would suppose the existence of a relation from species to

Table 4.6: Spriss' sampling rates on the **reduced GOS dataset** (26 metagenomes). Rows are minimum frequency thresholds inputted to Spriss, columns refer to different k-mer lengths. Sampling rates are expressed as number of k-mers in the subsample divided by the number of k-mers available in the complete metagenome. The values displayed are taken as arithmetic mean of all sampling rates on the dataset, with their standard deviation expressed as an error in parenthesis. See also Table B.3 in Appendix B for more details on sampling sizes when 16-mers are used.

ϑ	$k = 12$	$k = 14$	$k = 16$	$k = 21$	$k = 26$	$k = 31$
$\mathbf{8.00 \times 10^{-6}}$	0.04(1)	0.04(1)	0.04(1)	0.04(1)	0.04(1)	0.04(1)
$\mathbf{7.00 \times 10^{-6}}$	0.04(1)	0.04(1)	0.04(1)	0.04(1)	0.04(1)	0.04(1)
$\mathbf{6.00 \times 10^{-6}}$	0.05(2)	0.05(2)	0.05(2)	0.05(2)	0.05(2)	0.05(2)
$\mathbf{5.50 \times 10^{-6}}$	0.05(2)	0.05(2)	0.05(2)	0.05(2)	0.05(2)	0.05(2)
$\mathbf{5.00 \times 10^{-6}}$	0.06(2)	0.06(2)	0.06(2)	0.06(2)	0.06(2)	0.06(2)
$\mathbf{4.00 \times 10^{-6}}$	0.08(2)	0.08(2)	0.08(2)	0.08(2)	0.08(2)	0.08(2)
$\mathbf{3.50 \times 10^{-6}}$	0.09(3)	0.09(3)	0.09(3)	0.09(3)	0.09(3)	0.09(3)
$\mathbf{3.00 \times 10^{-6}}$	0.10(3)	0.10(3)	0.10(3)	0.10(3)	0.10(3)	0.10(3)
$\mathbf{2.50 \times 10^{-6}}$	0.12(4)	0.12(4)	0.12(4)	0.12(4)	0.12(4)	0.12(4)
$\mathbf{2.00 \times 10^{-6}}$	0.16(5)	0.16(5)	0.16(5)	0.16(5)	0.16(5)	0.16(5)
$\mathbf{1.50 \times 10^{-6}}$	0.21(7)	0.21(7)	0.21(7)	0.22(7)	0.22(7)	0.22(7)
$\mathbf{1.00 \times 10^{-6}}$	0.3(1)	0.3(1)	0.3(1)	0.3(1)	0.3(1)	0.3(1)
$\mathbf{8.00 \times 10^{-7}}$	0.4(1)	0.4(1)	0.4(1)	0.4(1)	0.4(1)	0.4(1)
$\mathbf{5.00 \times 10^{-7}}$	0.7(2)	0.7(2)	0.7(2)	0.7(2)	0.7(2)	0.7(2)

k-mers such that an approximately constant number of k-mers, likely dependent on the environment studied, distinguishes each species. Non-discriminative k-mers, shared among many species, would then carry information about distribution of higher-ranked taxa. The variance brought by the latter, however, might reasonably be negligible compared to species variance when similar environments are compared — i.e., we expect linearity would drop down if, for instance, human and oceanic metagenomes were pooled together and compared. Diversity between samples at higher taxonomic ranks would then affect k-mer-based dissimilarity indices by some constants rather than biasing their type of relation with their respective species-level reference-based indices. We pose it here as an hypothesis, which has not been verified yet.

Figure 4.13 shows an odd behaviour of the BC index on Spriss' subsamples on the **CAMI collection**: a very mild subsampling in terms of sample size results in a sudden increase of all pairwise dissimilarities, which then decrease tidily towards their value on original samples, and finally increase again more messily. Such an early peak is likely due to deterministic removal of unique k-mers actuated by Simka, the impact of which is milder on Spriss subsamples at high sampling rates because of replacement in Spriss' reads-sampling scheme, which inevitably duplicate many originally unique k-mers. Coherently, at lower sampling rates, singleton removal is partially reached. Figure 4.14 sustains our hypothesis.

Nevertheless, although being barely perceivable in Fig. 4.12c,[3] Pearson correlation between true BC indices and those based on k-mers on Spriss' subsamples with minimum frequency thresholds from 2.5×10^{-8} to 2.0×10^{-7} is the highest. For instance, with k-mer length $k = 26$, at a minimum frequency threshold $\vartheta = 1 \times 10^{-7}$, which corresponds to a sampling rate of 14.7 %, k-mer-based BC dissimilarity on such a subsample has a Pearson correlation of 0.9949 with the true BC, whereas Simka results on the complete dataset with unique k-mers removal holds a linear correlation of 0.9938 with truth, and Bracken-based correlation is 0.9900. At heavier subsampling, though — i.e., sampling rates[4] below 2.6 %, — the quality of k-mer based BC dissimilarity quickly drops, thus suggesting that important information is lost.

Spearman correlation with true BC dissimilarity, on the other hand, holds highest values on Spriss' subsamples at high sampling rates. Such correlations decline as soon as a $\vartheta = 5.0 \times 10^{-8}$ threshold is applied, in correspondence with a 30.7 % sampling rate. It is therefore very coherent with what can be observed in Fig. 4.13d: Spearman correlation is fairly constant as long as pairwise dissimilarities maintain their mutual ordering (rank), but decline as soon as such an order begins to be lost. As we mentioned earlier (see Section 4.2), Spearman correlation is more informative than Pearson correlation on low levels of a clustering tree built on these dissimilarities. However, due to marked linear correlation between k-mer-based and true BC indices on the CAMI dataset, we would privilege Pearson correlation for comparison.

The collection of real metagenomes from the **HMP project** holds similar results to the previously analysed ones (Fig. 4.15), except that: **1.** Pearson correlation between bracken-based BC index and its k-mer version on Spriss' subsamples is never higher than that between the former and its k-mer version computed on the complete dataset, and **2.** dissimilarity measures monotonically increase along with diminishing subsampling rates.

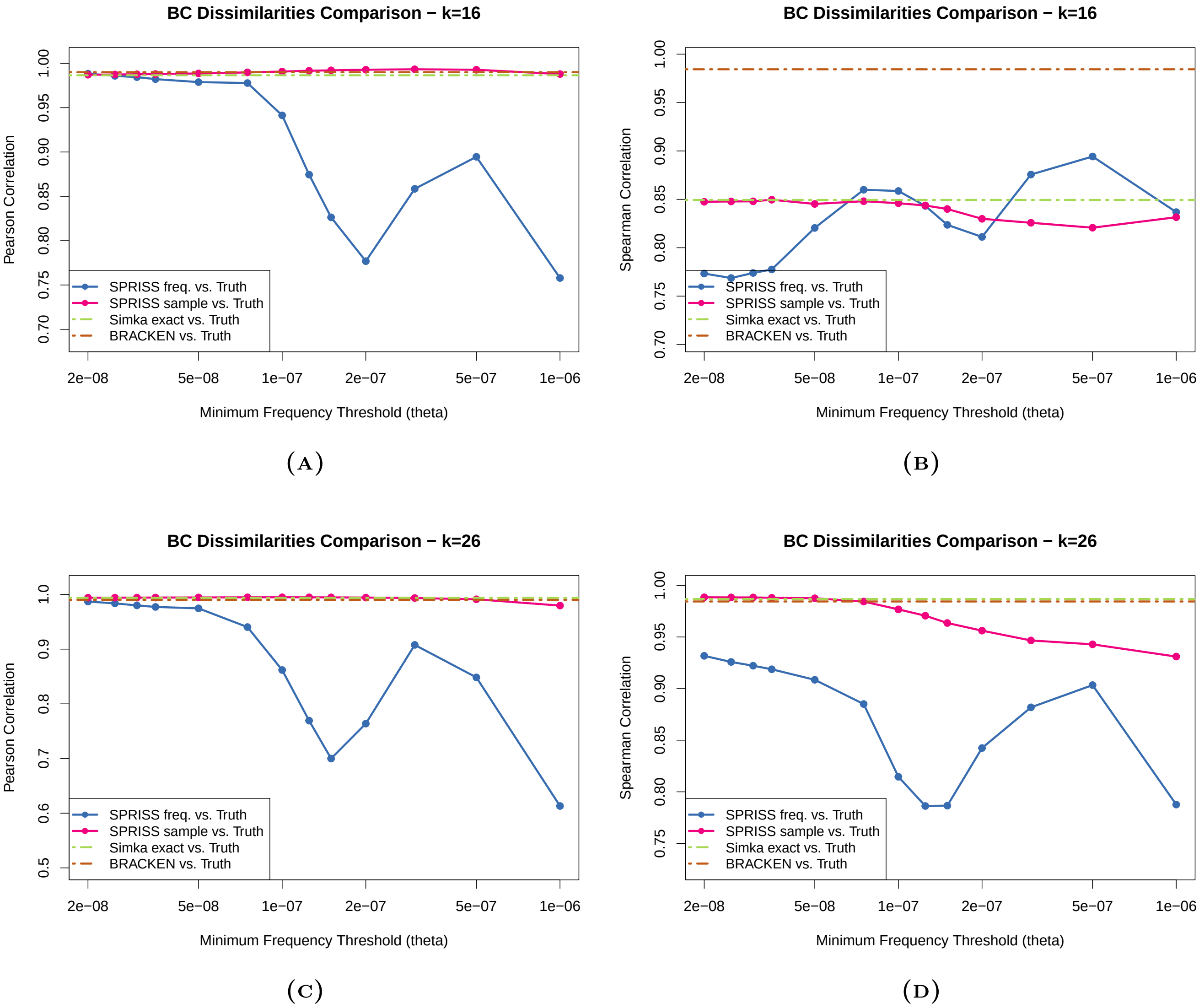
Therefore, both real and simulated collections of large human metagenomes constituted of short reads provided no evidence of Spriss subsampling removing noise nor important information as far as the sampling rate is not excessively low for the dataset.

Lastly, as seen in Section 4.2.1, the best correlation between species-level Bracken-based BC dissimilarity and its k-mer-based version on the **reduced**

[3] The interested reader can download all the results from the aforementioned git repository, where correlation values on Spriss experiments are stored in `csv` files inside directories named `Spriss_MTDabs__vs__Bracken_BacArcVirPla_lvlS`.

[4] See Table 4.5

GOS dataset is reached on 16-mers. On such settings, results are shown in Fig. 4.16, whereas Fig. 4.17 displays analogous results on 21-mers. In this scenario, Spriss subsampling improves correlation between reference-based and k-mer-based BC indices, which is rather surprising if considering the scarcity of the dataset in comparison with the richness of oceanic environments. As observable in Figs. 4.2, 4.16d and 4.17d, indeed, subsampling here results in a clear rise of the majority of the pairwise dissimilarities, yet these provide a better means of metagenomic comparison. Interestingly, Spriss' subsampling improves drastically metagenomic comparison on too short k-mer lengths (Fig. 4.18), suggesting that frequent relatively short k-mers might be as informative as complete sets of longer k-mers.



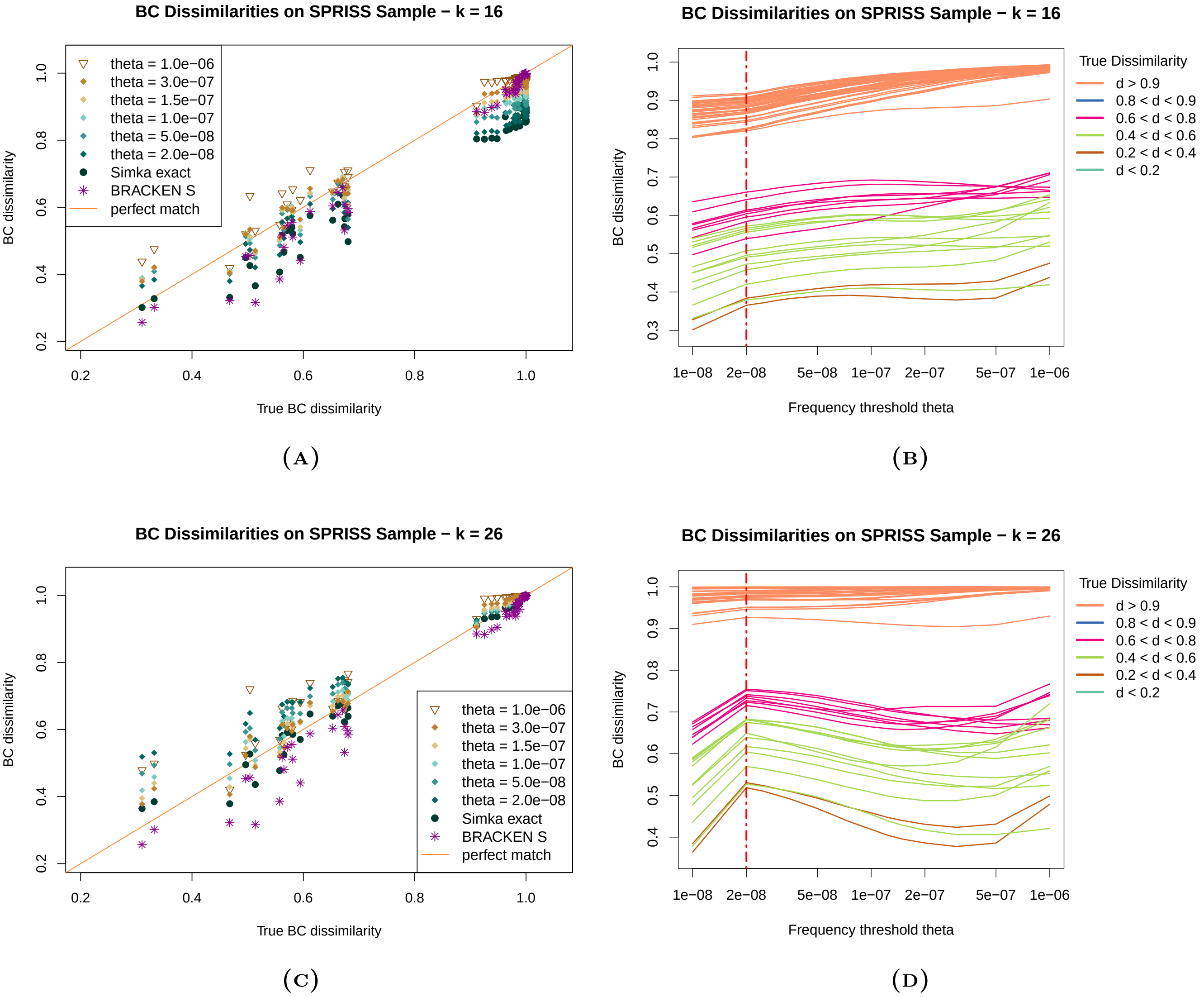
(A)

(B)

(C)

(D)

FIGURE 4.12: Correlation between SPRISS-based and true BC indices on the **CAMI dataset**. Also correlation between truth and BRACKEN-based BC dissimilarity at species level is plotted. In legend, "SPRISS freq." stands for dissimilarity computed on frequent k-mers' estimates of abundance; "SPRISS sample" denotes the exact index computed by SIMKA, with unique k-mers removal, on metagenomes subsampled by SPRISS.



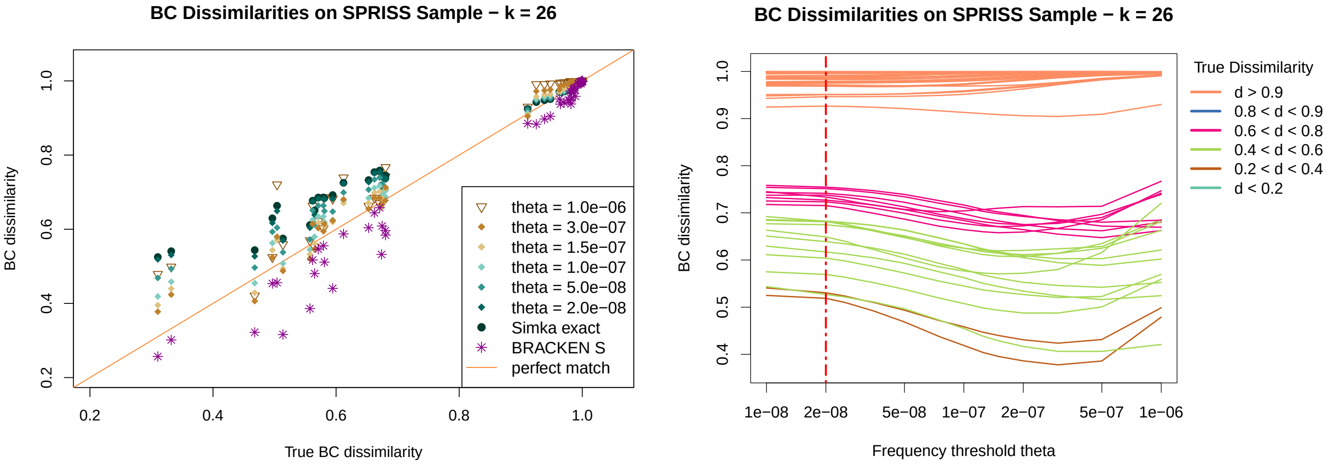
(a)

(b)

(c)

(d)

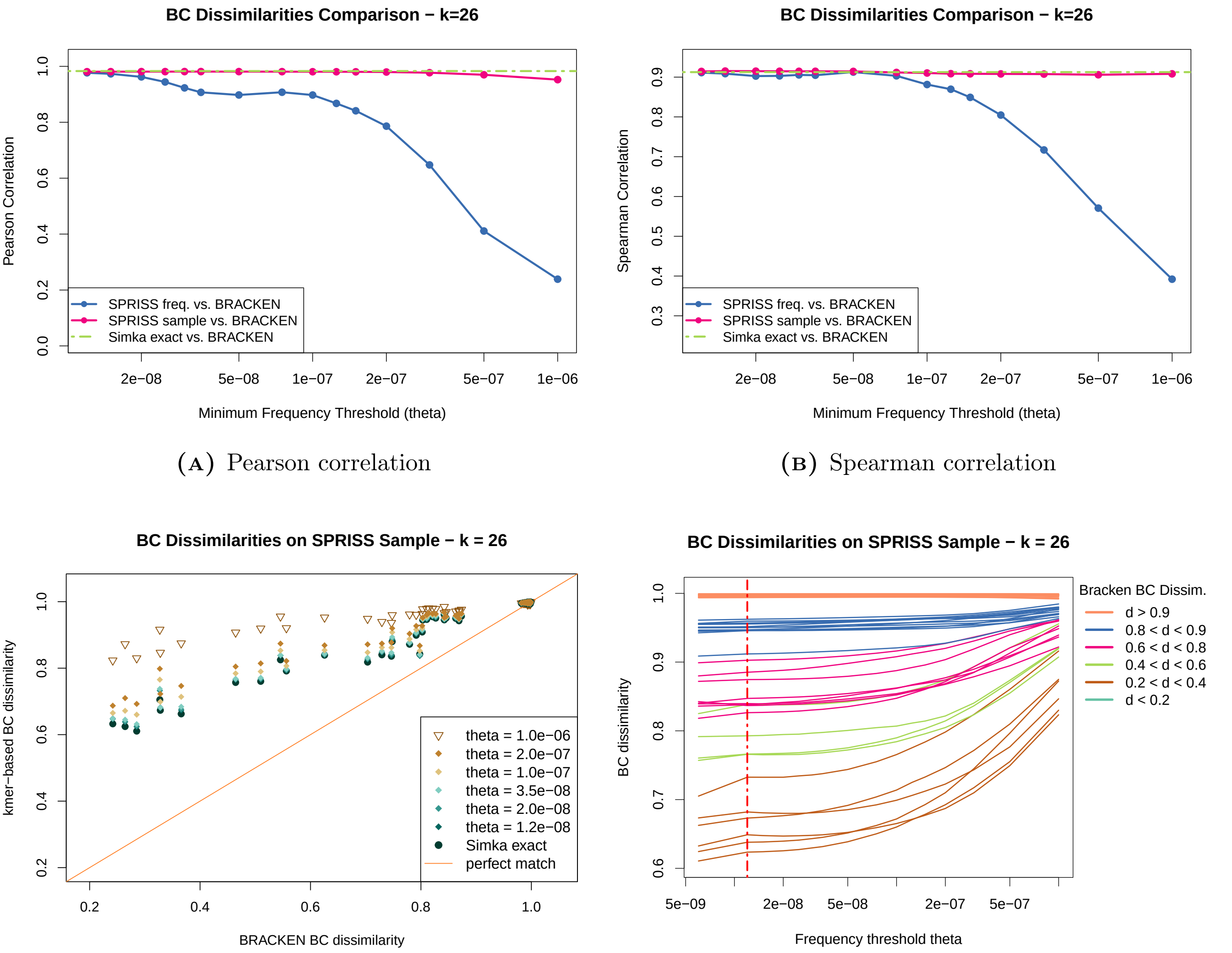
Figure 4.13: Spriss-sample-based and Bracken-based BC indices referred to truth, at species level, on the **CAMI dataset**. On the right, a vertical dashed red line indicates the lowest frequency threshold applied: at its left, the dissimilarity values on complete datasets, with unique k-mers excluded, as computed by Simka are reported.



(a)

(b)

Figure 4.14: Spriss-sample-based and Bracken-based BC indices referred to truth, at species level, on the **CAMI dataset** as in Figs. 4.13c and 4.13d, but exact k-mer-based BC dissimilarities on the non-sampled dataset are computed by Simka without filtering out singletons (dark bullets in the plot on left, leftmost end on the plot in the right).



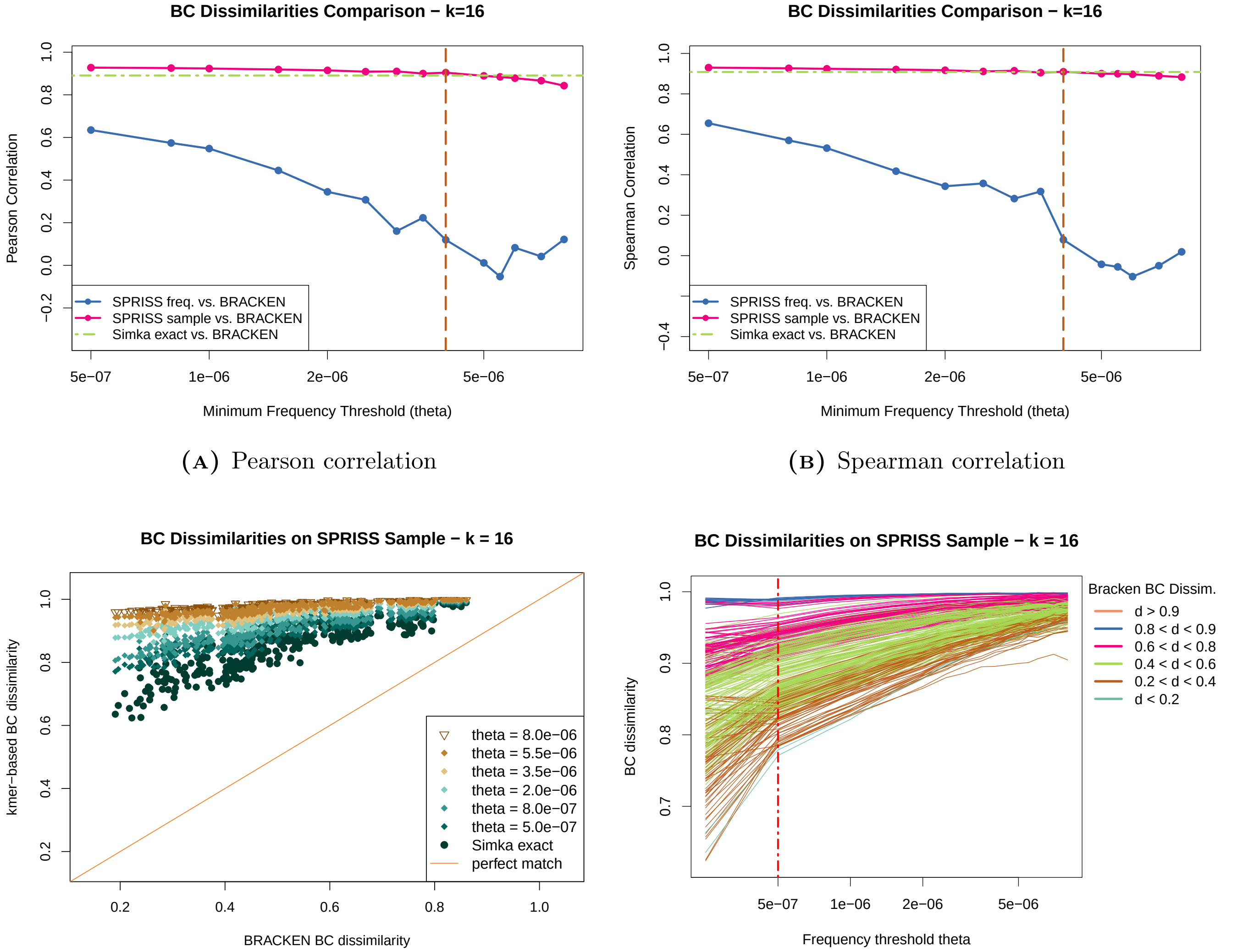
(A) Pearson correlation

(B) Spearman correlation

(C) k-mer-based versus reference-based dissimilarities

(D) k-mer-based dissimilarities along with decreasing sampling rate

Figure 4.15: On the top, correlation between Spriss-sample-based and Bracken-based BC indices on the **HMP dataset**; "Spriss freq." in the legend stands for dissimilarity computed on frequent k-mers' estimates of abundance, "Spriss sample" denotes the exact index computed by Simka, with unique k-mers removal, on metagenomes subsampled by Spriss. On the bottom, Spriss-based compared to species-level Bracken-based BC indices. A vertical dashed red line in Fig. 4.15d indicates the lowest frequency threshold applied: at its left, dissimilarity values on complete datasets with unique k-mers excluded as computed by Simka are reported.



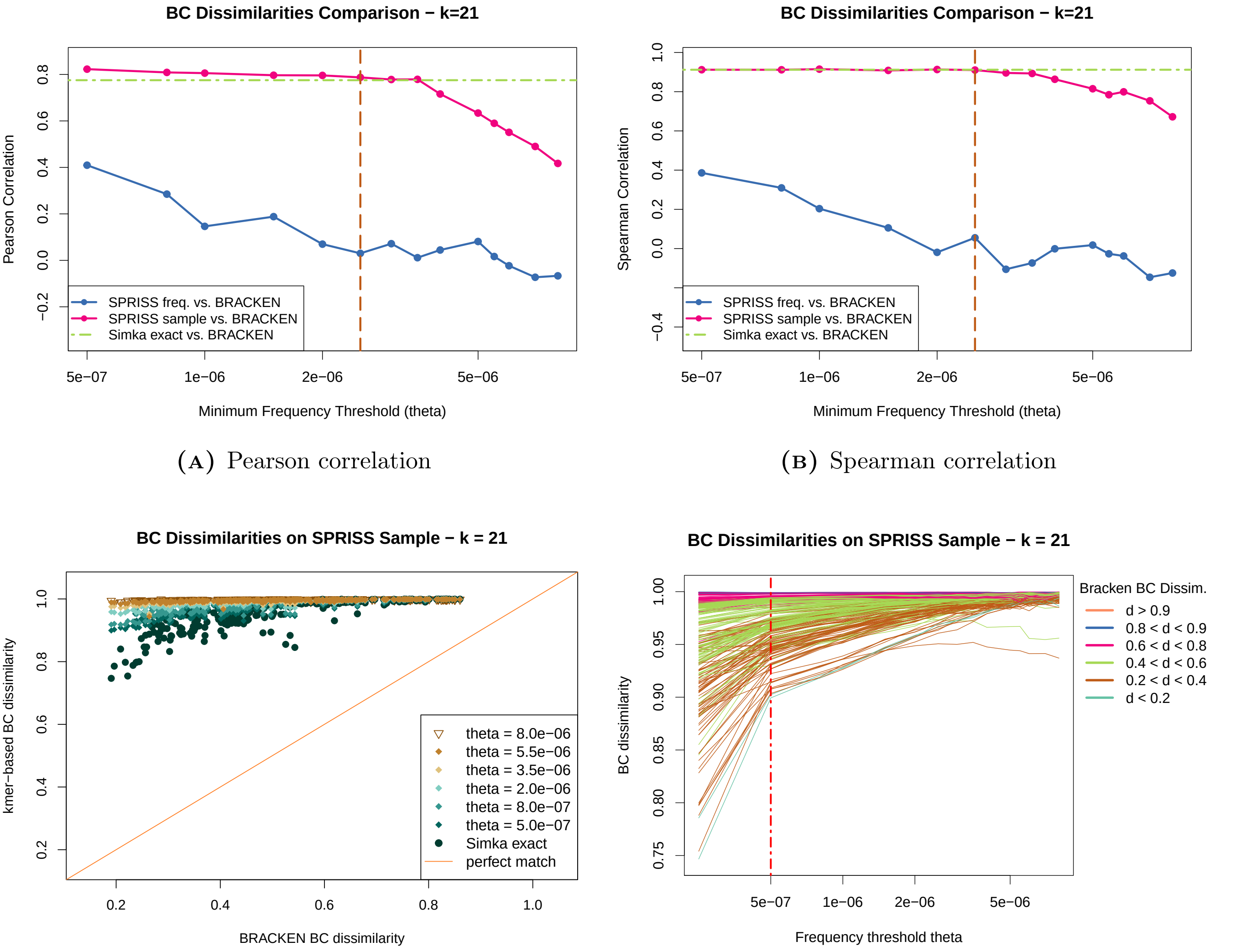
(A) Pearson correlation

(B) Spearman correlation

(C) k-mer-based versus reference-based dissimilarities

(D) k-mer-based dissimilarities along with decreasing sampling rate

FIGURE 4.16: Comparison of SPRISS-sample-based and BRACKEN-based BC indices on the **reduced GOS dataset** when 16-mers are used by SPRISS. A vertical dashed brown line in the top Figs. 4.16a and 4.16b signals the lowest frequency threshold for which at least one metagenome holds an empty frequent k-mers set, which is considered to be at unitary dissimilarity from every other frequent k-mers set. On the top, "SPRISS freq." in the legends stands for dissimilarity computed on frequent k-mers' estimates of abundance, "SPRISS sample" denotes the exact index computed by SIMKA, with unique k-mers removal, on metagenomes subsampled by SPRISS. In Fig. 4.16d, a vertical dashed red line indicates the lowest frequency threshold applied: at its left, dissimilarity values on complete datasets with unique k-mers excluded, computed by SIMKA, are reported.



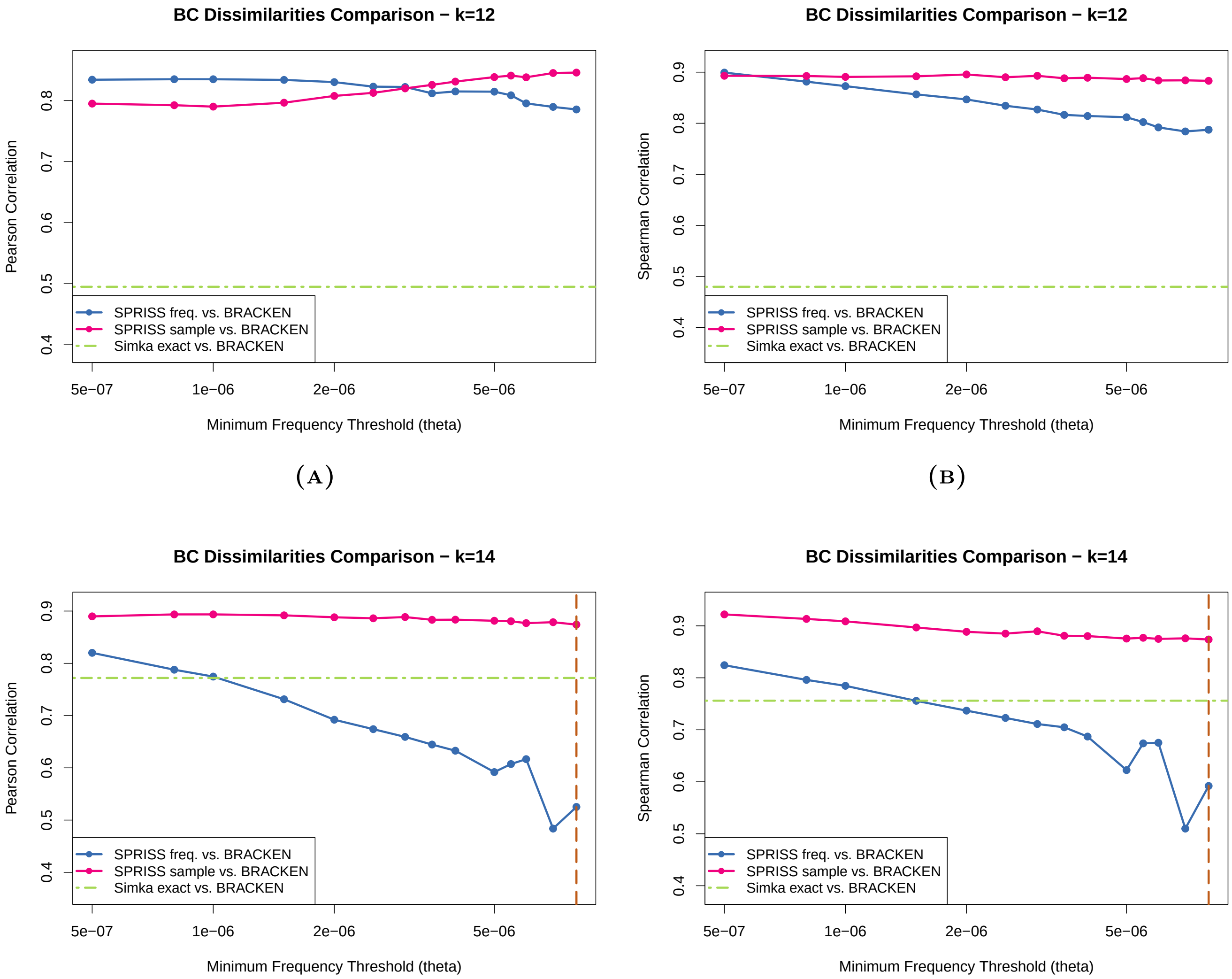
(A) Pearson correlation

(B) Spearman correlation

(C) k-mer-based versus reference-based dissimilarities

(D) k-mer-based dissimilarities along with decreasing sampling rate

Figure 4.17: Comparison of Spriss-sample-based and Bracken-based BC indices on the **reduced GOS dataset** when 21-mers are used by Spriss. A vertical dashed brown line in the top Figs. 4.16a and 4.16b signals the lowest frequency threshold for which at least one metagenome holds an empty frequent k-mers set, which is considered to be at unitary dissimilarity from every other frequent k-mers set. On the top, “Spriss freq.” in the legends stands for dissimilarity computed on frequent k-mers’ estimates of abundance, “Spriss sample” denotes the exact index computed by Simka, with unique k-mers removal, on metagenomes subsampled by Spriss. In Fig. 4.17d, a vertical dashed red line indicates the lowest frequency threshold applied: at its left, dissimilarity values on complete datasets with unique k-mers excluded, computed by Simka, are reported.



(A)

(B)

(C)

(D)

Figure 4.18: Correlation between Spriss-based and Bracken-based BC indices on the **reduced GOS dataset** on short k-mers. In legend, "Spriss freq." stands for dissimilarity computed on frequent k-mers' estimates of abundance; "Spriss sample" denotes the exact index computed by Simka, with unique k-mers removal, on metagenomes subsampled by Spriss. A vertical dashed brown line in the top Figs. 4.16a and 4.16b signals the lowest frequency threshold for which at least one metagenome holds an empty frequent k-mers set, which is considered to be at unitary dissimilarity from every other frequent k-mers set.

4.3.2 BC Dissimilarity on SPRISS Frequent k-mers Estimates

Since SPRISS is a tool for estimating frequent k-mers' abundances, we compared the BC index computed on the latter with its reference-based version to asses this third subsampling technique. With few exceptions, BC dissimilarity between frequent k-mers estimates correlates generally worse than all the other computations of BC dissimilarity (blue lines in Figs. 4.12, 4.15a, 4.15b, 4.16a, 4.17b and 4.18, and Figs. C.2, C.4 and C.5 in Appendix C).

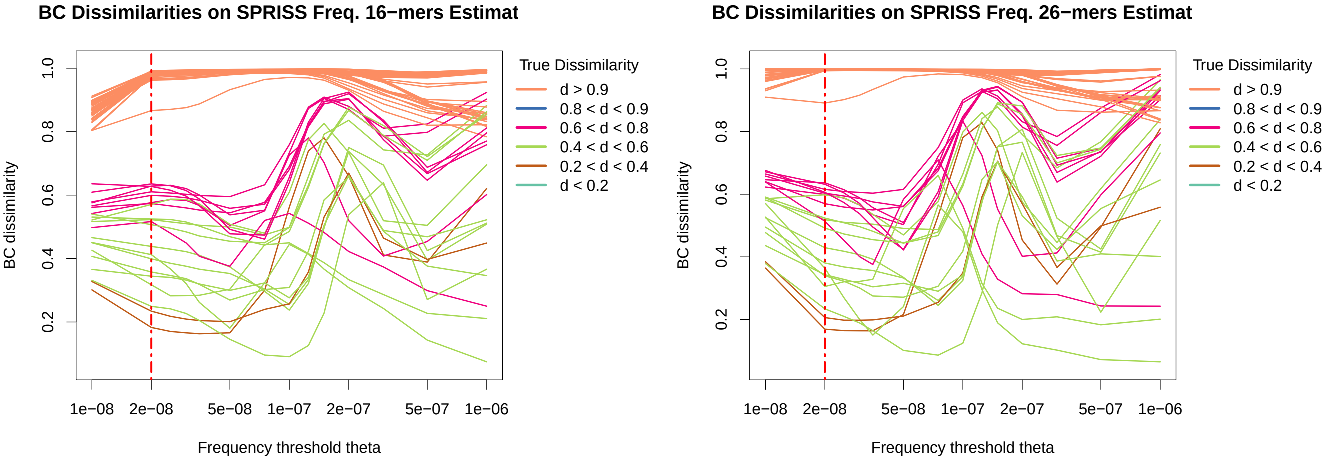


FIGURE 4.19: BC dissimilarity indices computed on SPRISS' *frequent k-mers* abundances estimation between pairs of metagenomes on the CAMI collection, varying sampling effort. See Fig. 4.13 for further details on the plot structure. A vertical dashed red line indicates the lowest frequency threshold applied: at its left, the dissimilarity values on complete datasets, with unique k-mers excluded, as computed by SIMKA are reported.

However, on simulated large human metagenomes composed of short-reads — i.e., the **CAMI dataset**, — when *minimum frequency thresholds* are $\vartheta \leq 5 \times 10^{-8}$, Pearson correlation to true BC index is comparable with those held by BRACKEN and by SIMKA, whereas Spearman correlation is markedly worse (Fig. 4.12 and Fig. C.2 in Appendix C). Both correlations have a steep decline peaking to a minimum around the threshold $\vartheta = 1.5 \times 10^{-7}$, then increasing again. An explanation of such an odd phenomenon might be that mildly frequent k-mers are the most unshared and, hence, they enhance dissimilarity; consequently, when they are excluded by higher minimum frequency thresholds the BC index improves again before declining definitely. Such an hypothesis is weakly sustained by Fig. 4.19, however it deserves further investigation. More likely, the phenomenon could be explained by considering the interaction of the two different estimates used: one for frequent k-mer selection, and one for frequency estimation. Indeed, with the settings we applied — i.e., $\varepsilon = \vartheta - {}^{2}/_{t_{\mathcal{D},k}}$ as suggested by Santoro et al., — the frequent k-mers set output by

SPRISS does only guarantee not to report any k-mer appearing fewer than three times in the original metagenome [34, Def. 1, point 2]; therefore, the observed hollow may depend on the distribution of false negative and false positive k-mers in the frequent k-mers set.

On the **HMP dataset**, Spearman correlation between BC dissimilarities based on frequent k-mers' abundances estimates and the BRACKEN-based ones are more comparable with those between the latter and the SIMKA-based BC (Fig. 4.15d). Nonetheless, low minimum frequency thresholds provide high Pearson correlation as well, while the behaviour noticed on the CAMI dataset is only weakly perceivable here (Figs. 4.15d and 4.20a).

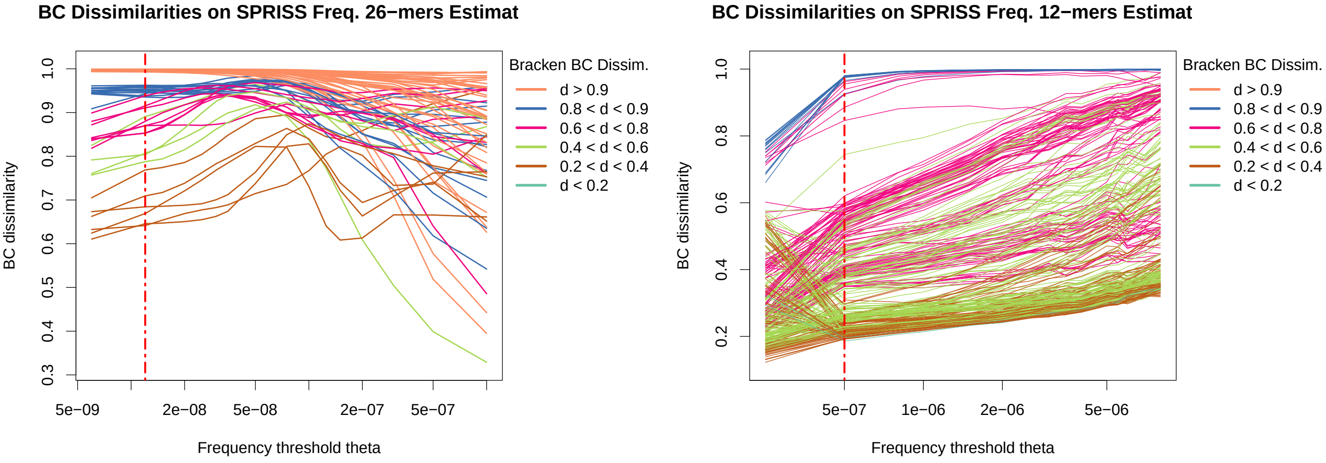
Exceptionally, as far as the **reduced GOS dataset** is concerned, although non-optimal, the linear correlation of BC indices on estimated abundances of frequent 12-mers with those based on BRACKEN's abundances is better than that computed on all k-mers, as displayed in Fig. 4.18. The respective plot of Pearson correlation in this figure (top-left) clearly demonstrate that infrequent 12-mers can be considered as a noise when BC dissimilarity is to be computed on their abundances: not only the SPRISS' approximate set of frequent 12-mers holds the highest linear correlation with true BC dissimilarity on all minimum frequency thresholds lower than $\vartheta = 3 \times 10^{-6}$, but also SPRISS' subsamples provide better dissimilarity estimates on subsamples with lower sampling rates. Notwithstanding, the dependence on k-mer length $k = 12$, the limitations of the GOS dataset, and the visualisation of such BC dissimilarities in Fig. 4.20b do not allow us to generalise from this single result.

Overall, limiting metagenomic comparison to estimated abundances of frequent k-mers is never improving k-mer-based BC index on taxonomic interpretation of it at species level on higher k-mer lengths. As a consequence, our results strongly suggest that, despite frequent k-mers carrying most of the relevant taxonomic information, part of rarer k-mers are informative too when comparing environment by means of the BC dissimilarity index on their sequenced metagenomes.

4.3.3 Experiments on the Jaccard Distance

Lastly, we observed the effects of subsampling on the k-mer-based computation of the Jaccard distance between two metagenomes. Once again, k-mer lengths $k = 21$, 26 and 31 provide very similar results, so we are only reporting those on 26-mers as representative.

On the **CAMI dataset**, we observe a sampling window on minimum frequency thresholds $\vartheta = 5.00 \times 10^{-8}$ to 1.50×10^{-8} where k-mer-based Jaccard distance computed by SIMKA with singletons removal on the SPRISS subsamples is unequivocally better correlated to the true Jaccard distance than any other is, both in Pearson and Spearman methods (Fig. 4.22). Figure 4.23b shows that pairs of metagenomes holding the lowest Jaccard distance get better clustered together toward low distance values when such thresholds are applied. We therefore believe



(A) The HMP dataset

(B) The GOS dataset

FIGURE 4.20: BC dissimilarity indices computed on SPRISS' *frequent k-mers* abundances estimation between pairs of metagenomes, varying sampling effort. A vertical dashed red line indicates the lowest frequency threshold applied: at its left, the dissimilarity values on complete datasets, with unique k-mers excluded, as computed by SIMKA are reported.

that rare k-mers are inadequate when inferring information about presence-absence dissimilarities. On such dissimilarities, indeed, frequent and rare k-mers are equally weighted. Therefore, were we to explain why infrequent k-mers are informative, although weakly, for computing metagenomic BC dissimilarity while being rather noisy on the Jaccard distance, we would suggest that there are both informative and noisy infrequent k-mers, and that the way through which they are weighted by the respective indices makes the one or the other part of them emerge.

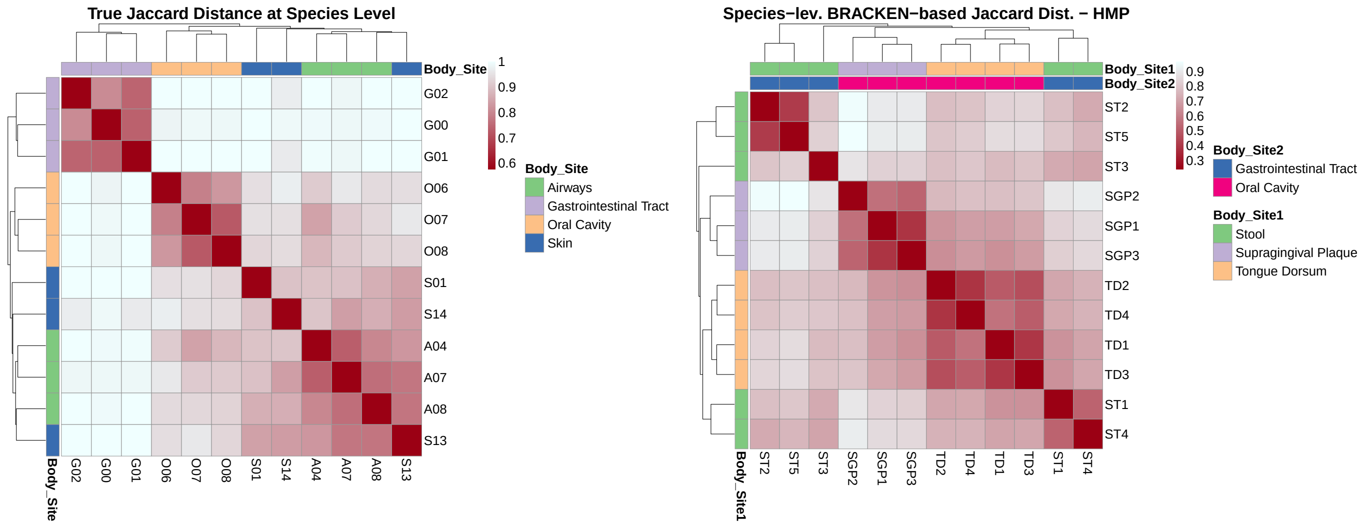
Interestingly, the high peak in correlation just described happens near to the lowest correlation between the Jaccard distance computed on SPRISS' approximate frequent k-mers set and the true taxonomic Jaccard distance. This might suggest that infrequent k-mers on the subsamples just indicated are the most informative for presence-absence β-diversity. However, we think the second hypothesis proposed in Section 4.3.2 applies here as well and could be the right explanation. Consequently, a specific study on the distribution of false positives and false negatives on the respective SPRISS' approximate frequent k-mers sets and their impact on metagenomic comparison should be carried out to verify our hypothesis.

The **HMP dataset** provides similar results (Figs. 4.24 and 4.25). Notwithstanding, we see a lower correlation, which is maximum on 16-mers, and is less variant on subsampling rates higher than 1 %. The main reason for a worse correlation may be the lower number of pairs of metagenomes at nearly unitary distance in this collection, as clear in Fig. 4.21. Indeed, our results suggest that k-mer-based

Jaccard distance does not correlate well with its reference-based version on pairs of similar samples in general, as visible in both CAMI and HMP results displayed in Figs. 4.23b and 4.25, respectively.

K-mer based Jaccard distance on the **reduced GOS dataset** correlates quite badly with their respective reference-based Jaccard distance, holding a coefficient always well below 0.75. However, because clustering of this dataset is sharply better[5] on k-mer-based Jaccard distance than on its reference-based version (Fig. C.1 in Appendix C), our analysis is not applicable on this dataset.

Therefore, a sampling scheme selecting reads uniformly at random with replacement, as that applied by the tool SPRISS, seems to reduce the noise which affects Jaccard distance computation for metagenomic comparison, at proper sampling rates. However, despite being reasonably well correlated with reference-based Jaccard distance, k-mer-based Jaccard distance is not a really good substitute of the former, and subsampling, even where it improves such a correlation, is still not powerful enough for that aim.



(A) True Jaccard distance at species level on the **CAMI dataset**

(B) BRACKEN-based Jaccard distance at species level on the **HMP dataset**

FIGURE 4.21: Comparison of Jaccard distance on human datasets.

[5] Clustering is better with that provided by Rusch et al. as a reference.

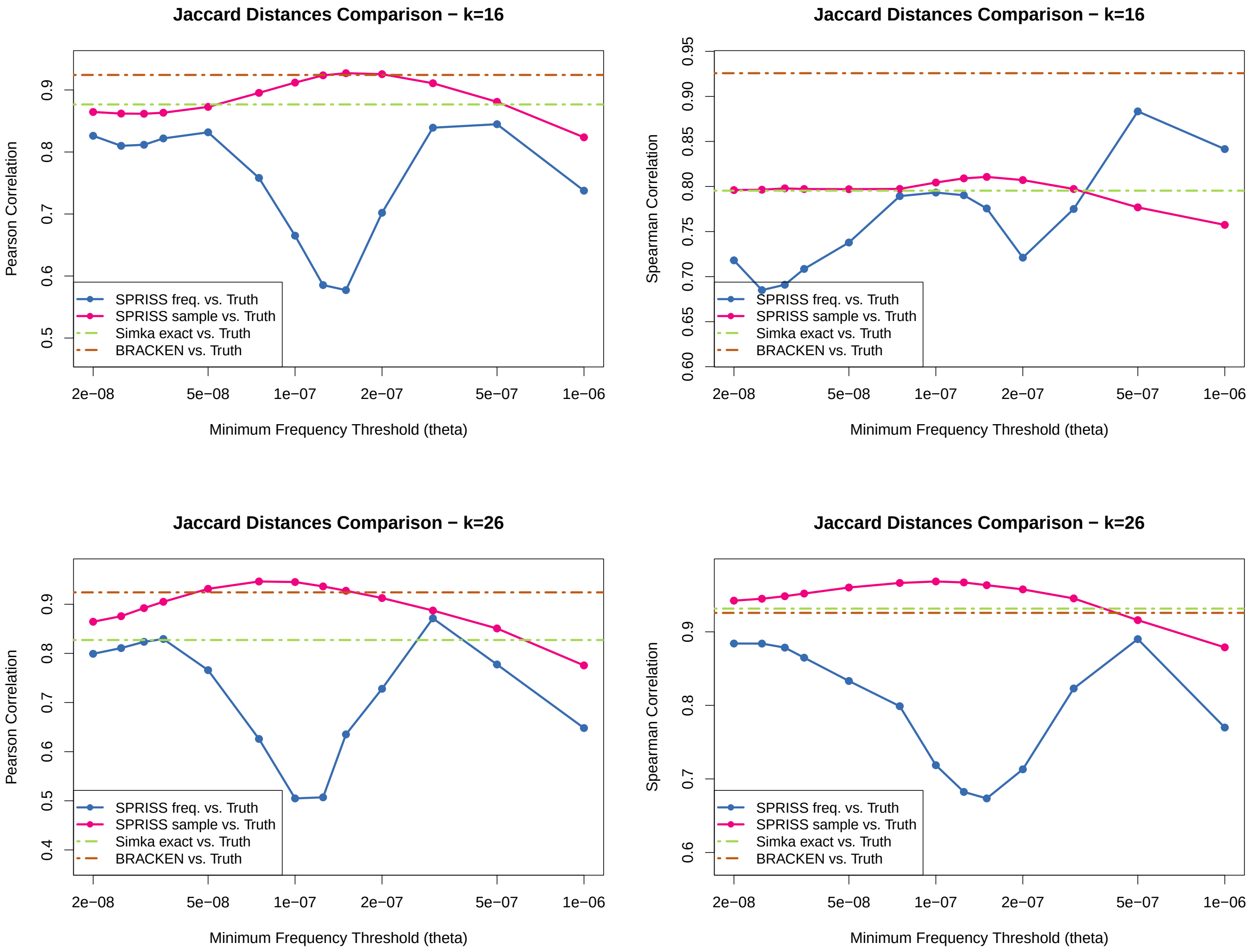
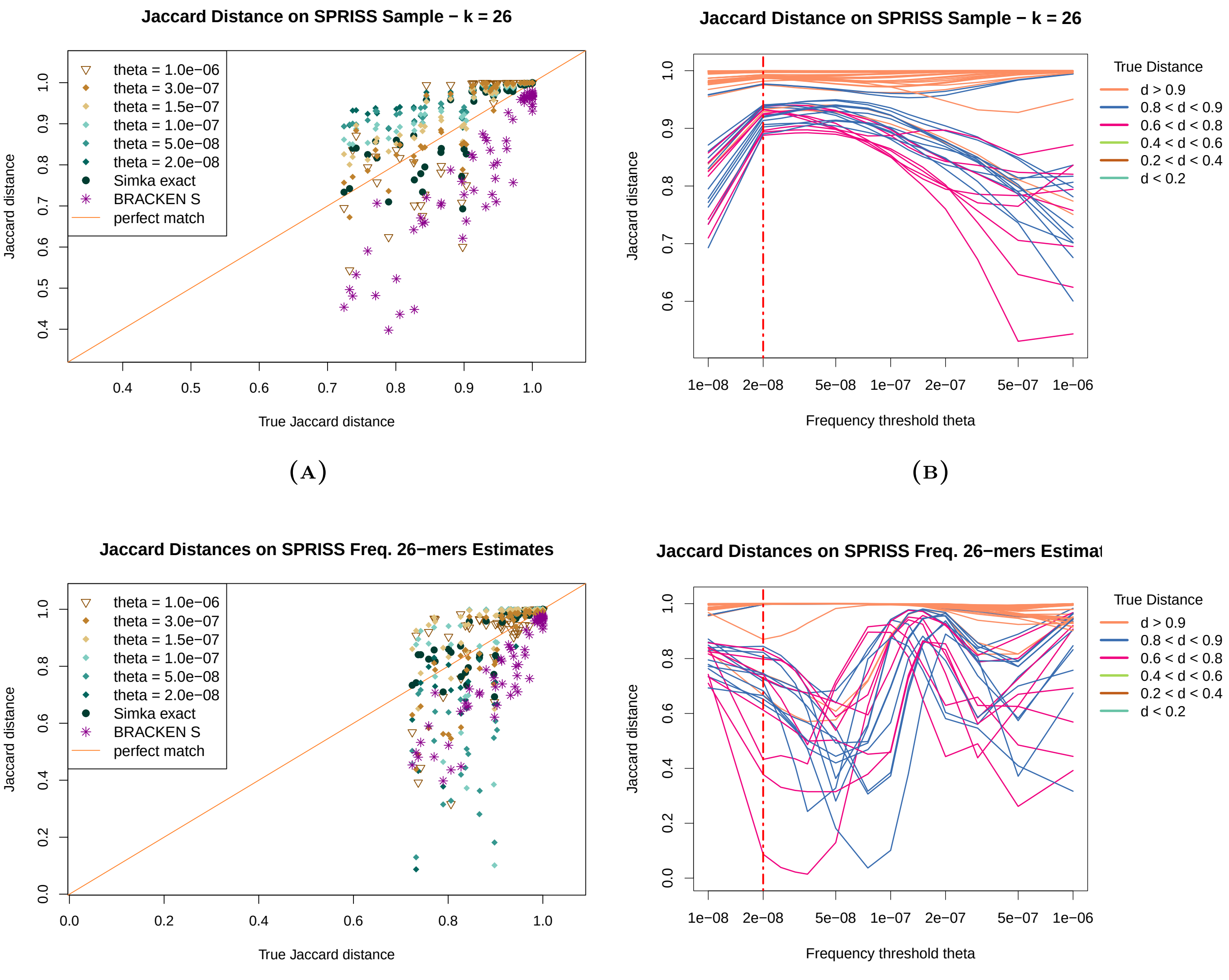


Figure 4.22: Correlation between Spriss-sample-based Jaccard distances and true Jaccard distances on the **CAMI dataset**. Also correlation between truth and Bracken-based Jaccard distance at species level is plotted. In legend, “Spriss freq.” stands for dissimilarity computed on frequent k-mers’ estimates of abundance; “Spriss sample” denotes the exact index computed by Simka, with unique k-mers removal, on metagenomes subsampled by Spriss.



(A)

(B)

(C)

(D)

Figure 4.23: Spriss-based and Bracken-based Jaccard distances referred to true Jaccard distance on the **CAMI dataset**. On the right, a vertical dashed red line indicates the lowest frequency threshold applied: at its left, the distance values on complete datasets, with unique k-mers excluded, as computed by Simka are reported.

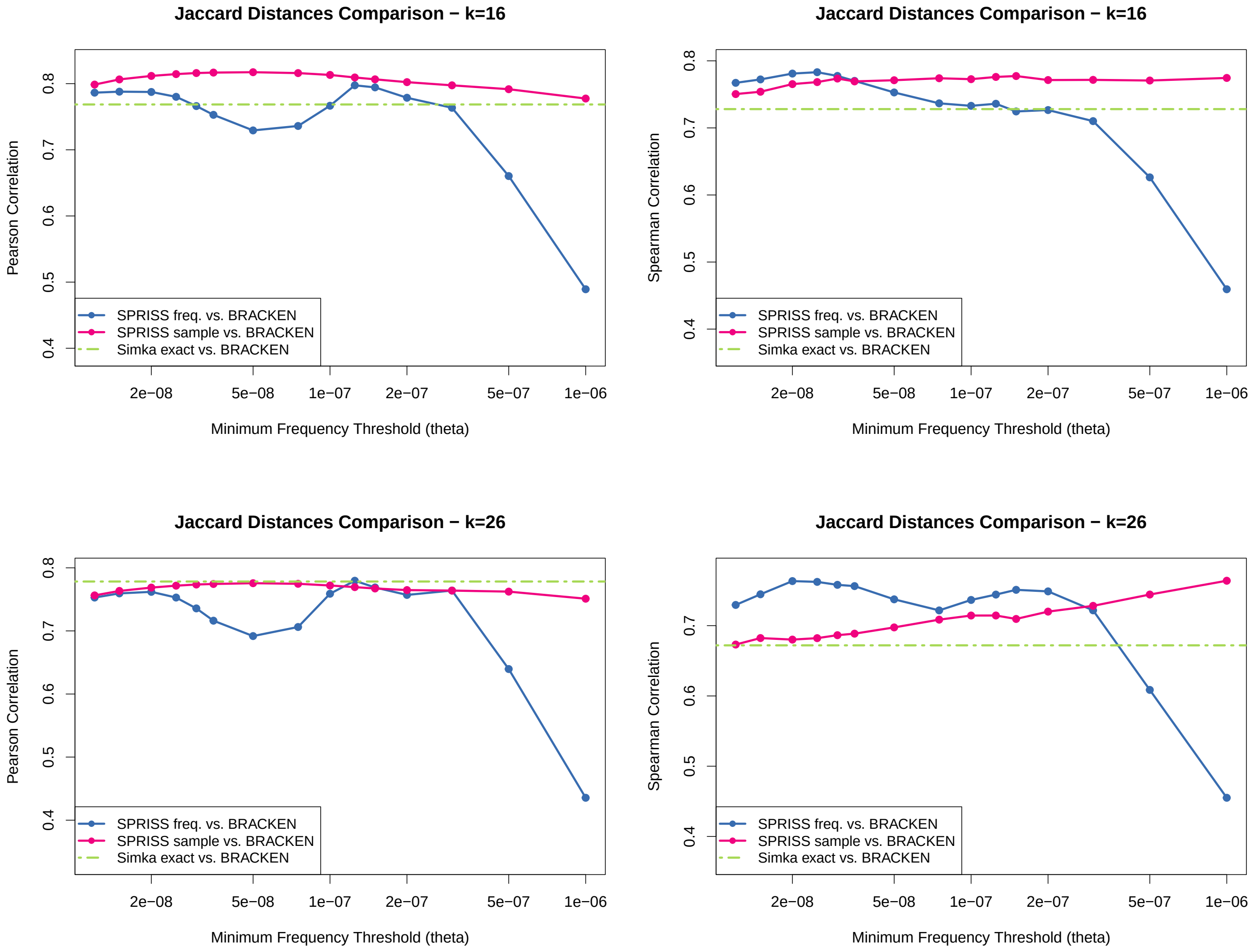
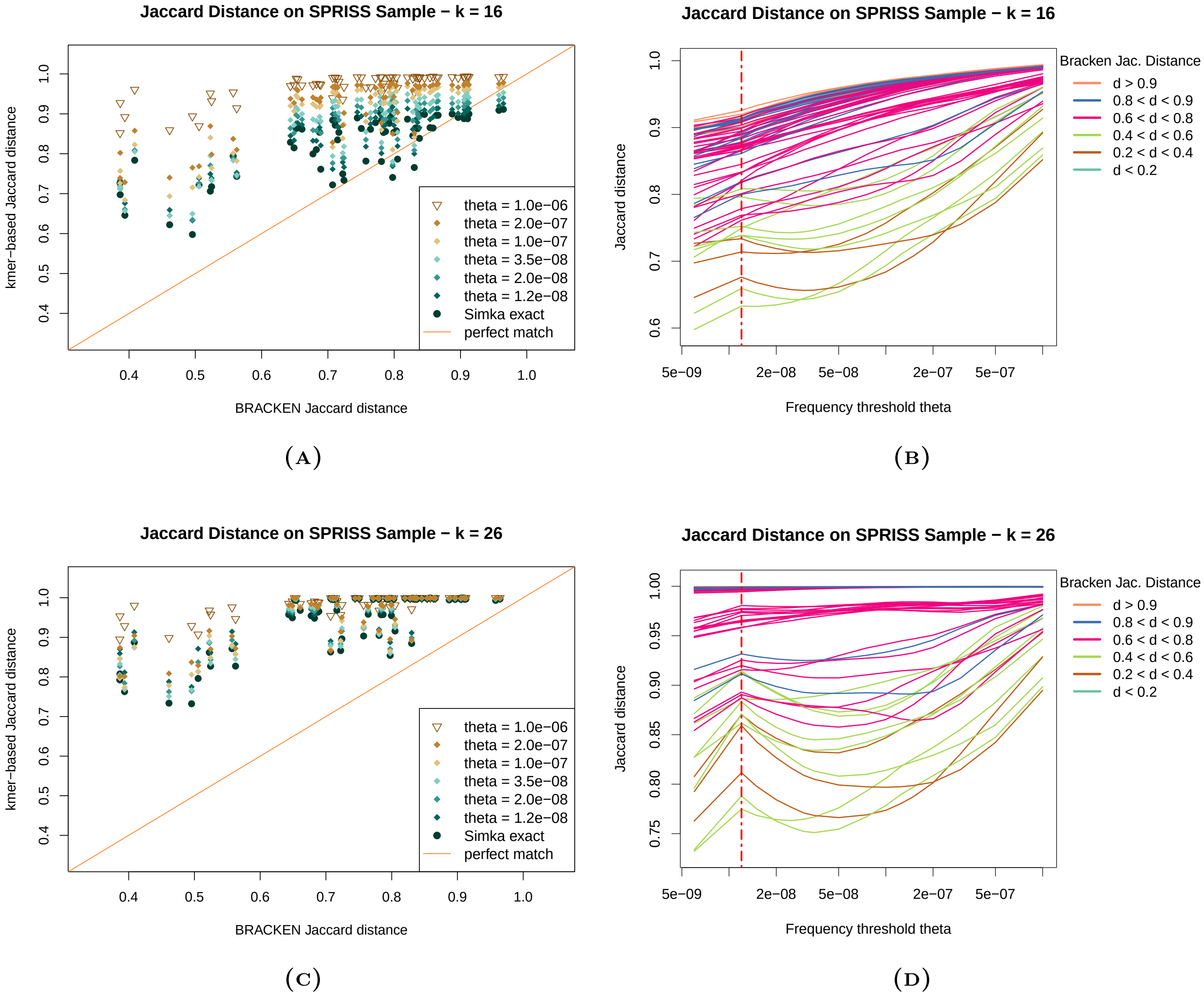


Figure 4.24: Correlation between Spriss-based and Bracken-based Jaccard distances on the **HMP dataset**. In legend, "Spriss freq." stands for dissimilarity computed on frequent k-mers' estimates of abundance; "Spriss sample" denotes the exact index computed by Simka, with unique k-mers removal, on metagenomes subsampled by Spriss.



(A)

(B)

(C)

(D)

Figure 4.25: Spriss-based referred to Bracken-based Jaccard distances on the **HMP dataset**. On the right, a vertical dashed red line indicates the lowest frequency threshold applied: at its left, the dissimilarity values on complete datasets, with unique k-mers excluded, as computed by Simka are reported.

Chapter 5

Conclusion

In this thesis we have explored the effects of three different subsampling approaches on k-mer-based metagenomic comparison in relation to reference-based metagenomic comparison.

Our results show that SimkaMin's sketching approach — i.e., randomly selecting a small set of distinct k-mers to be counted in the metagenomes under comparison — does not improve nor worsen the correlation between k-mer-based and reference-based BC dissimilarity and Jaccard distance, unless extremely small sketches — i.e., comprising fewer than 10^5 distinct k-mers — are applied. In the latter case, the quality of results naturally drops. Our observation is, therefore, in line with SimkaMin's authors claim that bias and variance of dissimilarity measures computed on such sketches are linearly proportional to the inverse of the number of distinct k-mers in the sketch [6, Suppl.].

Similarly, Spriss sampling scheme — i.e., selecting reads uniformly at random with replacement — appear to have very little impact on k-mer-based BC dissimilarity on large human metagenomes composed of short reads, on sampling rates higher than about 1 %. Instead, on metagenomes from the GOS experiment, which are small metagenomes with average sequence length $\approx$ 1070 bp, such a sampling scheme has provided better correlation with reference-based BC dissimilarity on sampling rates higher than 10 %. In such settings, Spriss sampling scheme drastically improve BC dissimilarity computed on very short k-mers, e.g., $k = 12$ and 14.

Coherently, a third sampling scheme providing an approximate set of frequent k-mers with their estimated abundances, as the output of the Spriss algorithm, shows a huge improvement of short-k-mers-based BC dissimilarity. On longer k-mers, however, this third method clearly worsen metagenomic comparison. Although this allows us to affirm that relevant information is thus lost, we ought to be cautious before stating that informative k-mers are therefore removed. Indeed, information loss may be due to statistical effects, like distribution of false positives and negatives in the estimated set. Effects of such a distribution are probably heavier when comparing k-mer-based Jaccard distance with its reference-based version.

Accordingly, our results show that SPRISS' approximate frequent k-mers set is non-optimal for computing a reference-free Jaccard distance. Instead, the best way of computing such indices, as emerges from our results, is by means of a moderate subsampling in the SPRISS' scheme. On the latter, indeed, reference-based Jaccard distance correlates with the k-mer-based Jaccard distance even better than it does on non-subsampled metagenomes.

Our study has also highlighted few interesting problems to be solved. Firstly, there is a need of better relating optimal k-mer length with relevant characteristics of the metagenomes and the task to be accomplished; here we focused on metagenomic comparison at species level. Indeed, the research carried out by Dubinkina et al. [12] should be repeated with longer k-mers and variable datasets. In particular, our results suggests k-mer length should be proportional to metagenome size, but nothing about this problem can be definitely inferred form our study.

Secondly, the strong linear relation between k-mer-based and reference-based dissimilarity measures observable in our results deserves an explanation, as we have commented in Section 4.3.1.

To sum up, our study has shown a very weak impact of subsampling on the computation of reference-free BC dissimilarities between metagenomes, unless a consistent amount of data is excluded. Therefore, just uninformative data seem to be cut off by subsampling. Among such data, however, rare k-mers are likely to be more noisy on the Jaccard distance, therefore a subsampling scheme reducing their quantity improves sensibly the quality of such a distance.

Appendix A

Datasets Details

We gather here for convenience tables which enrich the datasets' description given in Section 4.1.

Table A.1: Sequencing properties of the **HMP dataset**.

Label	**Sample name**	**Reads tally**	**Average read length**	**Symbols tally**
SGP1	SRS053917	1.158×10^8	100.00	5
SGP2	SRS075410	9.949×10^7	95.06	5
SGP3	SRS011126	9.627×10^7	93.48	5
ST1	SRS012273	1.075×10^8	94.22	5
ST2	SRS045713	1.057×10^8	93.01	5
ST3	SRS016095	1.069×10^8	95.61	5
ST4	SRS011239	1.238×10^8	95.70	5
ST5	SRS024388	1.197×10^8	96.21	5
TD1	SRS012279	1.050×10^8	95.51	5
TD2	SRS011306	1.004×10^8	94.72	5
TD3	SRS023617	9.535×10^7	93.42	5
TD4	SRS062761	1.181×10^8	100.00	5

Table A.2: Sequencing sites of the **HMP dataset**.

Label	Name	Body Site	Sampling site
SGP1	SRS053917	Oral Cavity	Supragingival Plaque
SGP2	SRS075410	Oral Cavity	Supragingival Plaque
SGP3	SRS011126	Oral Cavity	Supragingival Plaque
ST1	SRS012273	Gastrointestinal Tract	Stool
ST2	SRS045713	Gastrointestinal Tract	Stool
ST3	SRS016095	Gastrointestinal Tract	Stool
ST4	SRS011239	Gastrointestinal Tract	Stool
ST5	SRS024388	Gastrointestinal Tract	Stool
TD1	SRS012279	Oral Cavity	Tongue Dorsum
TD2	SRS011306	Oral Cavity	Tongue Dorsum
TD3	SRS023617	Oral Cavity	Tongue Dorsum
TD4	SRS062761	Oral Cavity	Tongue Dorsum

Table A.3: Alphabet of the **HMP dataset**.

Label	A	C	G	T	N
SGP1	2.9×10^9	3.0×10^9	2.9×10^9	2.8×10^9	2.4×10^6
SGP2	2.5×10^9	2.3×10^9	2.3×10^9	2.4×10^9	1.5×10^6
SGP3	2.4×10^9	2.1×10^9	2.1×10^9	2.3×10^9	2.2×10^6
ST1	2.8×10^9	2.2×10^9	2.2×10^9	2.8×10^9	2.5×10^6
ST2	2.8×10^9	2.1×10^9	2.1×10^9	2.8×10^9	4.3×10^5
ST3	2.9×10^9	2.3×10^9	2.3×10^9	2.8×10^9	1.7×10^6
ST4	3.3×10^9	2.7×10^9	2.7×10^9	3.2×10^9	8.3×10^5
ST5	3.3×10^9	2.5×10^9	2.5×10^9	3.3×10^9	6.2×10^6
TD1	3.0×10^9	2.0×10^9	2.1×10^9	2.9×10^9	3.3×10^6
TD2	2.8×10^9	1.9×10^9	1.9×10^9	2.8×10^9	1.1×10^6
TD3	2.7×10^9	1.8×10^9	1.8×10^9	2.6×10^9	2.8×10^6
TD4	3.4×10^9	2.5×10^9	2.5×10^9	3.4×10^9	7.9×10^6

Table A.4: Properties of the **GOS dataset**.

Label	**Sample name**	**Reads tally**	**Average read length**	**Symbols tally**
E01	GOS011	1.244×10^5	1070.85	4
E02	GOS012	1.262×10^5	1078.62	4
NC1	GOS020	2.964×10^5	1063.42	4
NC2	GOS025	1.207×10^5	1075.50	4
NC3	GOS032	1.480×10^5	1035.97	4
NC4	GOS033	6.923×10^5	1054.10	4
TG01	GOS014	1.289×10^5	1085.58	4
TG02	GOS021	1.318×10^5	1088.44	4
TG03	GOS022	1.217×10^5	1077.41	4
TG04	GOS027	2.221×10^5	1068.65	4
TG05	GOS028	1.891×10^5	1084.40	4
TG06	GOS029	1.315×10^5	1093.47	4
TG07	GOS030	3.592×10^5	1090.61	4
TG08	GOS031	4.364×10^5	1057.91	4
TG11	GOS034	1.343×10^5	1058.45	4
TG12	GOS035	1.408×10^5	1078.30	4
TG13	GOS036	7.754×10^4	1106.01	4
TG14	GOS037	6.567×10^4	1045.40	4
TG15	GOS047	6.602×10^4	1035.10	4
TG16	GOS051	1.290×10^5	1089.28	4
TN1	GOS002	1.216×10^5	1058.98	4
TN2	GOS003	6.161×10^4	1086.07	4
TN3	GOS004	5.296×10^4	1074.83	4
TN4	GOS005	6.113×10^4	1079.37	4
TN5	GOS006	5.968×10^4	1082.72	4
TN6	GOS007	5.098×10^4	1087.31	4
TO1	GOS015	1.274×10^5	1083.79	4
TO2	GOS016	1.271×10^5	1081.48	4
TO3	GOS017	2.576×10^5	1091.93	4
TO4	GOS018	1.427×10^5	1096.20	4
TO5	GOS019	1.353×10^5	1081.94	4
TO6	GOS023	1.331×10^5	1079.49	4
TO7	GOS026	1.027×10^5	1061.74	4
TS1	GOS008	1.297×10^5	1062.25	4
TS2	GOS009	7.930×10^4	1063.36	4
TS3	GOS010	7.830×10^4	1052.62	4
TS4	GOS013	1.380×10^5	1079.51	4

Table A.5: Environmental characteristics of the **GOS dataset**. We report the categorisation made by Rusch et al. in [33] despite ignoring, for instance, how sample GOS014 could be labelled within the Galapagos group. NAEC stands for North America East Coast.

Label	**Name**	**Rusch et al. finer category**	**Rusch et al. large cluster**	**Geographic provenience**
E01	GOS011	Estuary	Temperate	Estuary NAEC
E02	GOS012	Estuary	Temperate	Estuary NAEC
NC1	GOS020	Non classified	Non classified	Freshwater Panama
NC2	GOS025	Non classified	Non classified	Trop. Eastern Pacific
NC3	GOS032	Non classified	Non classified	Galapagos
NC4	GOS033	Non classified	Non classified	Galapagos
TG01	GOS014	Galapagos	Tropical	Gulf Stream
TG02	GOS021	Galapagos	Tropical	Trop. Eastern Pacific
TG03	GOS022	Galapagos	Tropical	Trop. Eastern Pacific
TG04	GOS027	Galapagos	Tropical	Galapagos
TG05	GOS028	Galapagos	Tropical	Galapagos
TG06	GOS029	Galapagos	Tropical	Galapagos
TG07	GOS030	Galapagos	Tropical	Galapagos
TG08	GOS031	Galapagos	Tropical	Galapagos
TG11	GOS034	Galapagos	Tropical	Galapagos
TG12	GOS035	Galapagos	Tropical	Galapagos
TG13	GOS036	Galapagos	Tropical	Galapagos
TG14	GOS037	Galapagos	Tropical	Trop. Eastern Pacific
TG15	GOS047	Galapagos	Tropical	Trop. South Pacific
TG16	GOS051	Galapagos	Tropical	Polynesia Archipelagos
TN1	GOS002	Temp. North	Temperate	Coast NAEC
TN2	GOS003	Temp. North	Temperate	Coast NAEC
TN3	GOS004	Temp. North	Temperate	Coast NAEC
TN4	GOS005	Temp. North	Temperate	Embayment NAEC
TN5	GOS006	Temp. North	Temperate	Estuary NAEC
TN6	GOS007	Temp. North	Temperate	Coast NAEC
TO1	GOS015	Open Ocean	Tropical	Gulf Stream
TO2	GOS016	Open Ocean	Tropical	Gulf Stream
TO3	GOS017	Open Ocean	Tropical	Gulf Stream
TO4	GOS018	Open Ocean	Tropical	Gulf Stream
TO5	GOS019	Open Ocean	Tropical	Caribbean Sea
TO6	GOS023	Open Ocean	Tropical	Trop. Eastern Pacific
TO7	GOS026	Open Ocean	Tropical	Galapagos
TS1	GOS008	Temp. South	Temperate	Coast NAEC
TS2	GOS009	Temp. South	Temperate	Coast NAEC
TS3	GOS010	Temp. South	Temperate	Coast NAEC
TS4	GOS013	Temp. South	Temperate	Coast NAEC

Table A.6: The **CAMI simulated dataset**. In samples' names the original CAMI's sample numbering is kept.

Label	Sample name	Number of reads	Average read length	Number of symbols	Body Site
A04	CAMI_airways4	3.333×10^7	150	12	Airways
A07	CAMI_airways7	3.333×10^7	150	12	Airways
A08	CAMI_airways8	3.333×10^7	150	11	Airways
G00	CAMI_gastro0	3.333×10^7	150	7	Gastrointestinal Tract
G01	CAMI_gastro1	3.333×10^7	150	11	Gastrointestinal Tract
G02	CAMI_gastro2	3.333×10^7	150	4	Gastrointestinal Tract
O06	CAMI_oral6	3.332×10^7	150	9	Oral Cavity
O07	CAMI_oral7	3.332×10^7	150	12	Oral Cavity
O08	CAMI_oral8	3.333×10^7	150	12	Oral Cavity
S01	CAMI_skin1	3.333×10^7	150	11	Skin
S13	CAMI_skin13	3.333×10^7	150	10	Skin
S14	CAMI_skin14	3.333×10^7	150	11	Skin

Table A.7: Alphabet of the **CAMI dataset**.

Label	A	C	G	T	M	R	W	S	V	H	D	B	N
A04	1.4×10^9	1.1×10^9	1.1×10^9	1.4×10^9	219	443	468	176	365	315	0	14	2256
A07	1.4×10^9	1.1×10^9	1.1×10^9	1.4×10^9	105	178	121	51	217	87	3	0	817
A08	1.2×10^9	1.3×10^9	1.3×10^9	1.2×10^9	89	330	181	102	328	61	0	0	1180
G00	1.1×10^9	1.4×10^9	1.4×10^9	1.1×10^9	8	0	0	0	3	0	0	0	14
G01	1.2×10^9	1.3×10^9	1.3×10^9	1.2×10^9	4	17	7	2	17	14	0	0	77
G02	1.2×10^9	1.3×10^9	1.3×10^9	1.2×10^9	0	0	0	0	0	0	0	0	0
O06	1.5×10^9	1.0×10^9	1.0×10^9	1.5×10^9	12	3	5	0	4	0	0	0	18
O07	1.5×10^9	1.0×10^9	1.0×10^9	1.5×10^9	20	36	29	29	28	26	0	1	152
O08	1.5×10^9	1.0×10^9	1.0×10^9	1.5×10^9	276	449	312	189	330	229	0	3	1956
S01	1.5×10^9	1.0×10^9	1.0×10^9	1.5×10^9	51	96	134	27	162	68	0	0	453
S13	1.4×10^9	1.1×10^9	1.1×10^9	1.4×10^9	1	20	20	3	48	0	0	0	87
S14	1.6×10^9	9.2×10^8	9.2×10^8	1.6×10^9	162	529	393	137	786	93	0	0	2141

Table A.8: Properties of the GOS dataset reduced to 26 metagenomes.

Label	Sample name	Reads tally	Average read length	Symbols tally
E01	GOS011	1.244×10^5	1070.85	4
E02	GOS012	1.262×10^5	1078.62	4
NC1	GOS020	2.964×10^5	1063.42	4
NC2	GOS025	1.207×10^5	1075.50	4
NC3	GOS032	1.480×10^5	1035.97	4
NC4	GOS033	6.923×10^5	1054.10	4
TG01	GOS014	1.289×10^5	1085.58	4
TG02	GOS021	1.318×10^5	1088.44	4
TG03	GOS022	1.217×10^5	1077.41	4
TG04	GOS027	2.221×10^5	1068.65	4
TG05	GOS028	1.891×10^5	1084.40	4
TG06	GOS029	1.315×10^5	1093.47	4
TG07	GOS030	3.592×10^5	1090.61	4
TG08	GOS031	4.364×10^5	1057.91	4
TG11	GOS034	1.343×10^5	1058.45	4
TG12	GOS035	1.408×10^5	1078.30	4
TG16	GOS051	1.290×10^5	1089.28	4
TN1	GOS002	1.216×10^5	1058.98	4
TO1	GOS015	1.274×10^5	1083.79	4
TO2	GOS016	1.271×10^5	1081.48	4
TO3	GOS017	2.576×10^5	1091.93	4
TO4	GOS018	1.427×10^5	1096.20	4
TO5	GOS019	1.353×10^5	1081.94	4
TO6	GOS023	1.331×10^5	1079.49	4
TS1	GOS008	1.297×10^5	1062.25	4
TS4	GOS013	1.380×10^5	1079.51	4

Appendix B

Spriss Samples Details

Tables below show the actual Spriss' subsamples sizes for each dataset. For simplicity, subsamples on only one k-mer length is being shown for each dataset.

Table B.1: Spriss' subsamples sizes summary on CAMI dataset when 26-mers are used. Values displayed are taken as arithmetic mean of all sampling rates on the dataset, with their standard deviation expressed as an error in parenthesis if not null.

ϑ	**#reads original**	**#reads sampled**	**#26-mers original**	**#26-mers sampled**	**Sampling rate**
$\mathbf{1.00 \times 10^{-6}}$	$3.3328(6) \times 10^{7}$	4.1×10^{5}	$4.1660(8) \times 10^{9}$	5.13×10^{7}	$1.2314(2) \times 10^{-2}$
$\mathbf{5.00 \times 10^{-7}}$	$3.3328(6) \times 10^{7}$	8.64×10^{5}	$4.1660(8) \times 10^{9}$	1.08×10^{8}	$2.5924(5) \times 10^{-2}$
$\mathbf{3.00 \times 10^{-7}}$	$3.3328(6) \times 10^{7}$	1.49×10^{6}	$4.1660(8) \times 10^{9}$	1.86×10^{8}	$4.4647(9) \times 10^{-2}$
$\mathbf{2.00 \times 10^{-7}}$	$3.3328(6) \times 10^{7}$	2.34×10^{6}	$4.1660(8) \times 10^{9}$	2.93×10^{8}	$7.021(1) \times 10^{-2}$
$\mathbf{1.50 \times 10^{-7}}$	$3.3328(6) \times 10^{7}$	3.12×10^{6}	$4.1660(8) \times 10^{9}$	3.9×10^{8}	$9.361(2) \times 10^{-2}$
$\mathbf{1.25 \times 10^{-7}}$	$3.3328(6) \times 10^{7}$	3.74×10^{6}	$4.1660(8) \times 10^{9}$	4.68×10^{8}	$1.1234(2) \times 10^{-1}$
$\mathbf{1.00 \times 10^{-7}}$	$3.3328(6) \times 10^{7}$	4.9×10^{6}	$4.1660(8) \times 10^{9}$	6.12×10^{8}	$1.4690(3) \times 10^{-1}$
$\mathbf{7.50 \times 10^{-8}}$	$3.3328(6) \times 10^{7}$	6.53×10^{6}	$4.1660(8) \times 10^{9}$	8.16×10^{8}	$1.9587(4) \times 10^{-1}$
$\mathbf{5.00 \times 10^{-8}}$	$3.3328(6) \times 10^{7}$	1.02×10^{7}	$4.1660(8) \times 10^{9}$	1.28×10^{9}	$3.0677(6) \times 10^{-1}$
$\mathbf{3.50 \times 10^{-8}}$	$3.3328(6) \times 10^{7}$	1.48×10^{7}	$4.1660(8) \times 10^{9}$	1.85×10^{9}	$4.4441(8) \times 10^{-1}$
$\mathbf{3.00 \times 10^{-8}}$	$3.3328(6) \times 10^{7}$	1.73×10^{7}	$4.1660(8) \times 10^{9}$	2.16×10^{9}	$5.185(1) \times 10^{-1}$
$\mathbf{2.50 \times 10^{-8}}$	$3.3328(6) \times 10^{7}$	2.16×10^{7}	$4.1660(8) \times 10^{9}$	2.7×10^{9}	$6.481(1) \times 10^{-1}$
$\mathbf{2.00 \times 10^{-8}}$	$3.3328(6) \times 10^{7}$	2.74×10^{7}	$4.1660(8) \times 10^{9}$	3.42×10^{9}	$8.209(2) \times 10^{-1}$

Table B.2: Spriss' subsamples sizes summary on HMP dataset when 26-mers are used.

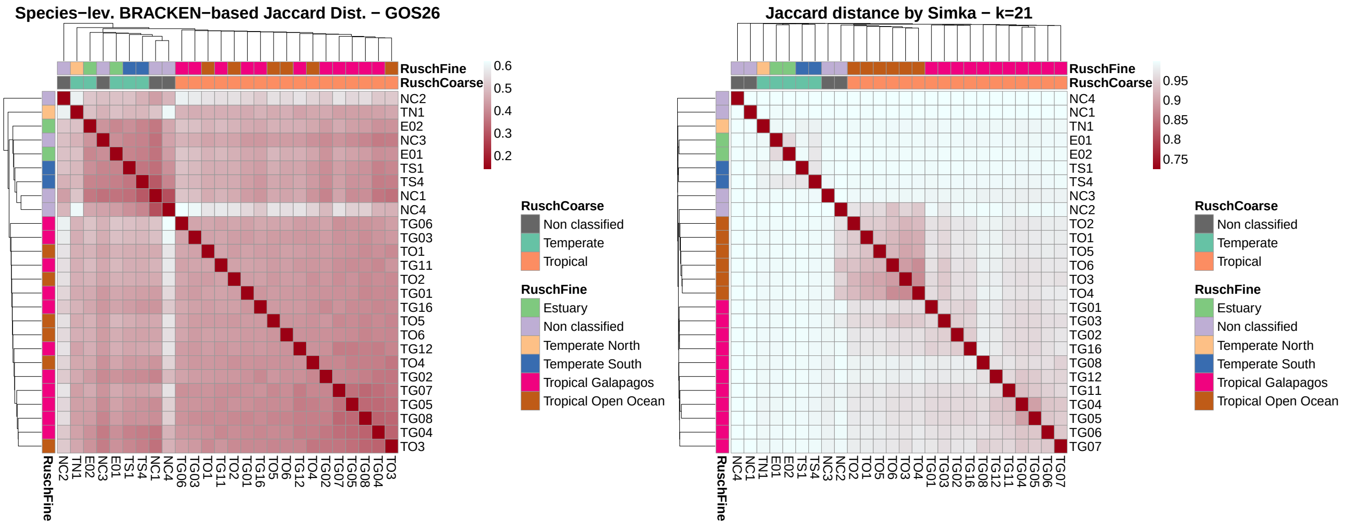
ϑ	**#reads original**	**#reads sampled**	**#26-mers original**	**#26-mers sampled**	**Sampling rate**
$\mathbf{1.00 \times 10^{-6}}$	$1.1(1) \times 10^{8}$	$7.3(2) \times 10^{5}$	$7.6(9) \times 10^{9}$	$5.1299(1) \times 10^{7}$	0.0068(7)
$\mathbf{5.00 \times 10^{-7}}$	$1.1(1) \times 10^{8}$	$1.53(5) \times 10^{6}$	$7.6(9) \times 10^{9}$	$1.079\,99(1) \times 10^{8}$	0.014(2)
$\mathbf{3.00 \times 10^{-7}}$	$1.1(1) \times 10^{8}$	$2.64(8) \times 10^{6}$	$7.6(9) \times 10^{9}$	$1.859\,98(1) \times 10^{8}$	0.025(3)
$\mathbf{2.00 \times 10^{-7}}$	$1.1(1) \times 10^{8}$	$4.1(1) \times 10^{6}$	$7.6(9) \times 10^{9}$	$2.924\,99(1) \times 10^{8}$	0.039(4)
$\mathbf{1.50 \times 10^{-7}}$	$1.1(1) \times 10^{8}$	$5.5(2) \times 10^{6}$	$7.6(9) \times 10^{9}$	$3.899\,98(1) \times 10^{8}$	0.052(6)
$\mathbf{1.25 \times 10^{-7}}$	$1.1(1) \times 10^{8}$	$6.6(2) \times 10^{6}$	$7.6(9) \times 10^{9}$	$4.679\,98(1) \times 10^{8}$	0.062(7)
$\mathbf{1.00 \times 10^{-7}}$	$1.1(1) \times 10^{8}$	$8.6(3) \times 10^{6}$	$7.6(9) \times 10^{9}$	$6.09(4) \times 10^{8}$	0.081(9)
$\mathbf{7.50 \times 10^{-8}}$	$1.1(1) \times 10^{8}$	$1.16(4) \times 10^{7}$	$7.6(9) \times 10^{9}$	$8.159\,98(2) \times 10^{8}$	0.11(1)
$\mathbf{5.00 \times 10^{-8}}$	$1.1(1) \times 10^{8}$	$1.79(6) \times 10^{7}$	$7.6(9) \times 10^{9}$	$1.259\,998(2) \times 10^{9}$	0.17(2)
$\mathbf{3.50 \times 10^{-8}}$	$1.1(1) \times 10^{8}$	$2.59(8) \times 10^{7}$	$7.6(9) \times 10^{9}$	$1.825\,712(2) \times 10^{9}$	0.24(3)
$\mathbf{3.00 \times 10^{-8}}$	$1.1(1) \times 10^{8}$	$3.0(1) \times 10^{7}$	$7.6(9) \times 10^{9}$	$2.129\,998(1) \times 10^{9}$	0.28(3)
$\mathbf{2.50 \times 10^{-8}}$	$1.1(1) \times 10^{8}$	$3.8(1) \times 10^{7}$	$7.6(9) \times 10^{9}$	$2.65(2) \times 10^{9}$	0.35(4)
$\mathbf{2.00 \times 10^{-8}}$	$1.1(1) \times 10^{8}$	$4.7(1) \times 10^{7}$	$7.6(9) \times 10^{9}$	$3.329\,998(2) \times 10^{9}$	0.44(5)
$\mathbf{1.50 \times 10^{-8}}$	$1.1(1) \times 10^{8}$	$6.4(2) \times 10^{7}$	$7.6(9) \times 10^{9}$	$4.48(3) \times 10^{9}$	0.59(7)
$\mathbf{1.20 \times 10^{-8}}$	$1.1(1) \times 10^{8}$	$8.3(3) \times 10^{7}$	$7.6(9) \times 10^{9}$	$5.83(3) \times 10^{9}$	0.77(9)

Table B.3: Spriss' subsamples sizes summary on the *reduced* GOS dataset when 16-mers are used.

ϑ	**#reads original**	**#reads sampled**	**#16-mers original**	**#16-mers sampled**	**Sampling rate**
$\mathbf{8.0 \times 10^{-6}}$	$2(1) \times 10^{5}$	$5.60(8) \times 10^{3}$	$2(1) \times 10^{8}$	$5.93(2) \times 10^{6}$	0.04(1)
$\mathbf{7.0 \times 10^{-6}}$	$2(1) \times 10^{5}$	$6.39(9) \times 10^{3}$	$2(1) \times 10^{8}$	$6.78(3) \times 10^{6}$	0.04(1)
$\mathbf{6.0 \times 10^{-6}}$	$2(1) \times 10^{5}$	$7.5(1) \times 10^{3}$	$2(1) \times 10^{8}$	$7.92(2) \times 10^{6}$	0.05(2)
$\mathbf{5.5 \times 10^{-6}}$	$2(1) \times 10^{5}$	$8.2(1) \times 10^{3}$	$2(1) \times 10^{8}$	$8.64(2) \times 10^{6}$	0.05(2)
$\mathbf{5.0 \times 10^{-6}}$	$2(1) \times 10^{5}$	$9.0(1) \times 10^{3}$	$2(1) \times 10^{8}$	$9.51(1) \times 10^{6}$	0.06(2)
$\mathbf{4.0 \times 10^{-6}}$	$2(1) \times 10^{5}$	$1.17(2) \times 10^{4}$	$2(1) \times 10^{8}$	$1.24(1) \times 10^{7}$	0.08(2)
$\mathbf{3.5 \times 10^{-6}}$	$2(1) \times 10^{5}$	$1.34(2) \times 10^{4}$	$2(1) \times 10^{8}$	$1.42(1) \times 10^{7}$	0.09(3)
$\mathbf{3.0 \times 10^{-6}}$	$2(1) \times 10^{5}$	$1.56(3) \times 10^{4}$	$2(1) \times 10^{8}$	$1.66(1) \times 10^{7}$	0.10(3)
$\mathbf{2.5 \times 10^{-6}}$	$2(1) \times 10^{5}$	$1.88(3) \times 10^{4}$	$2(1) \times 10^{8}$	$2.00(2) \times 10^{7}$	0.12(4)
$\mathbf{2.0 \times 10^{-6}}$	$2(1) \times 10^{5}$	$2.46(3) \times 10^{4}$	$2(1) \times 10^{8}$	$2.606(9) \times 10^{7}$	0.16(5)
$\mathbf{1.5 \times 10^{-6}}$	$2(1) \times 10^{5}$	$3.29(5) \times 10^{4}$	$2(1) \times 10^{8}$	$3.49(3) \times 10^{7}$	0.21(7)
$\mathbf{1.0 \times 10^{-6}}$	$2(1) \times 10^{5}$	$5.21(8) \times 10^{4}$	$2(1) \times 10^{8}$	$5.53(6) \times 10^{7}$	0.3(1)
$\mathbf{8.0 \times 10^{-7}}$	$2(1) \times 10^{5}$	$6.5(1) \times 10^{4}$	$2(1) \times 10^{8}$	$6.93(7) \times 10^{7}$	0.4(1)
$\mathbf{5.0 \times 10^{-7}}$	$2(1) \times 10^{5}$	$1.11(2) \times 10^{5}$	$2(1) \times 10^{8}$	$1.18(2) \times 10^{8}$	0.7(2)

Appendix C

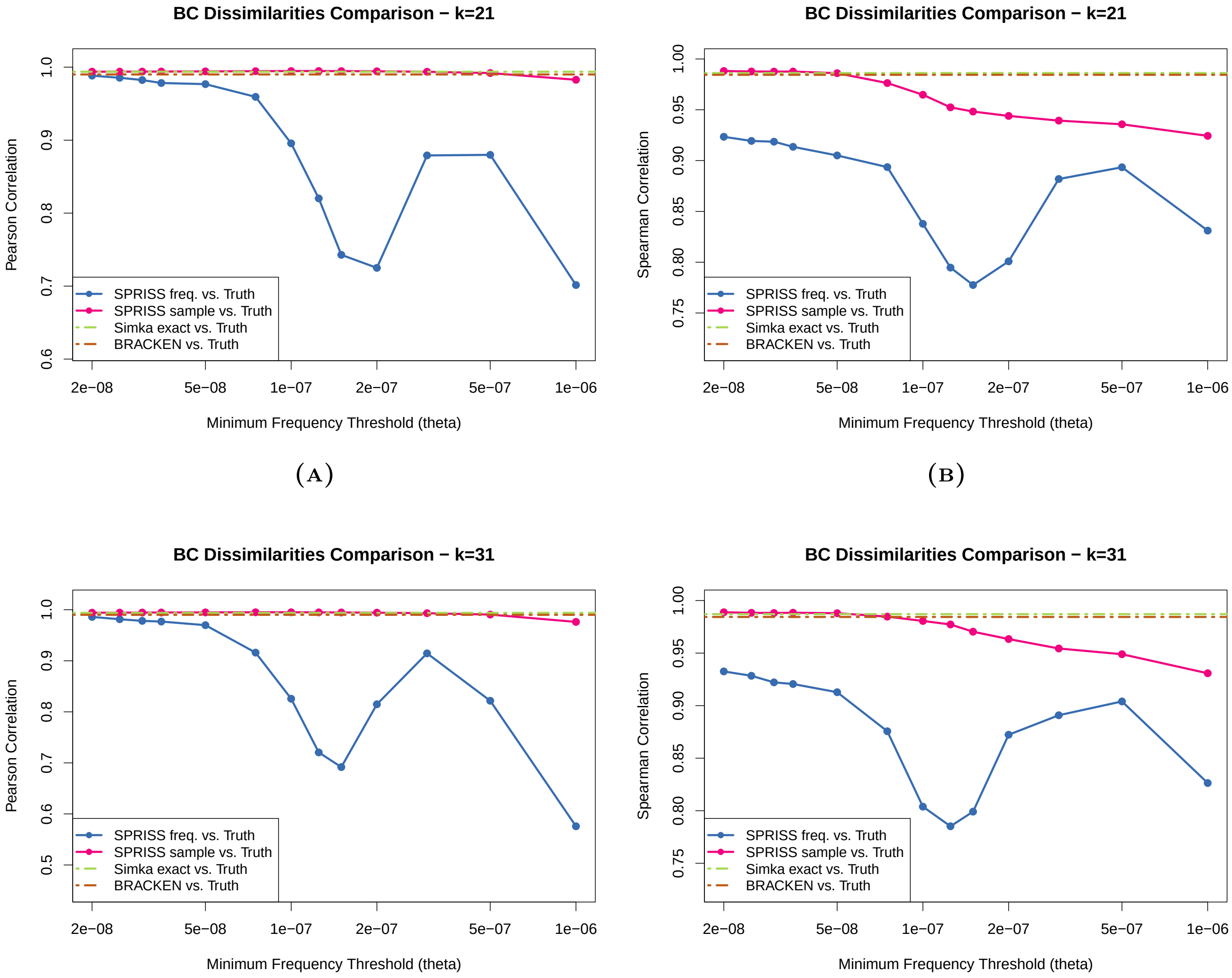
Supplementary Plots



(A) BRACKEN-based Jaccard Distance

(B) SIMKA-based Jaccard Distance on 21-mers

FIGURE C.1: Comparison of Jaccard-distance-based clusterings on the **reduced GOS collection**.



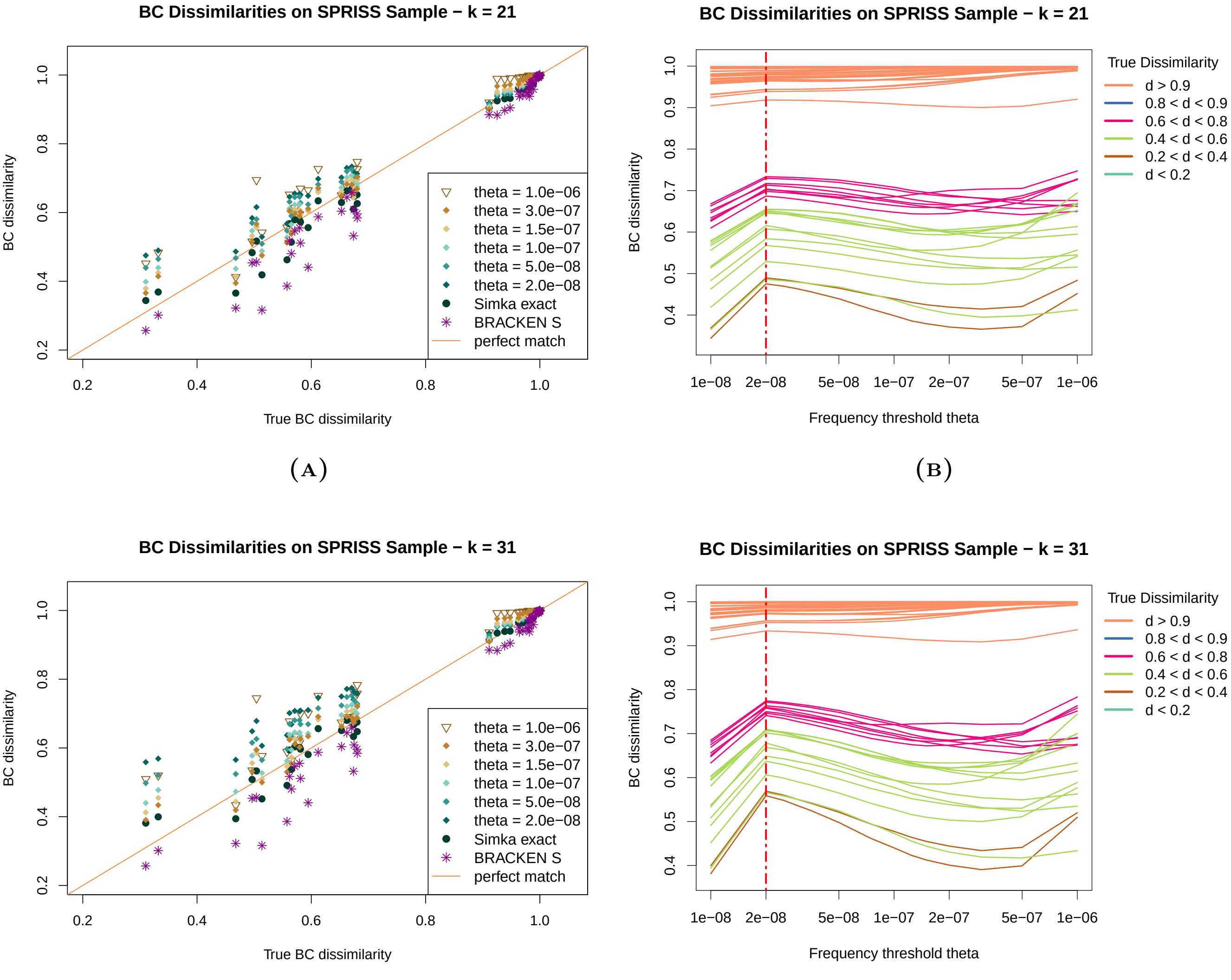
(a)

(b)

(c)

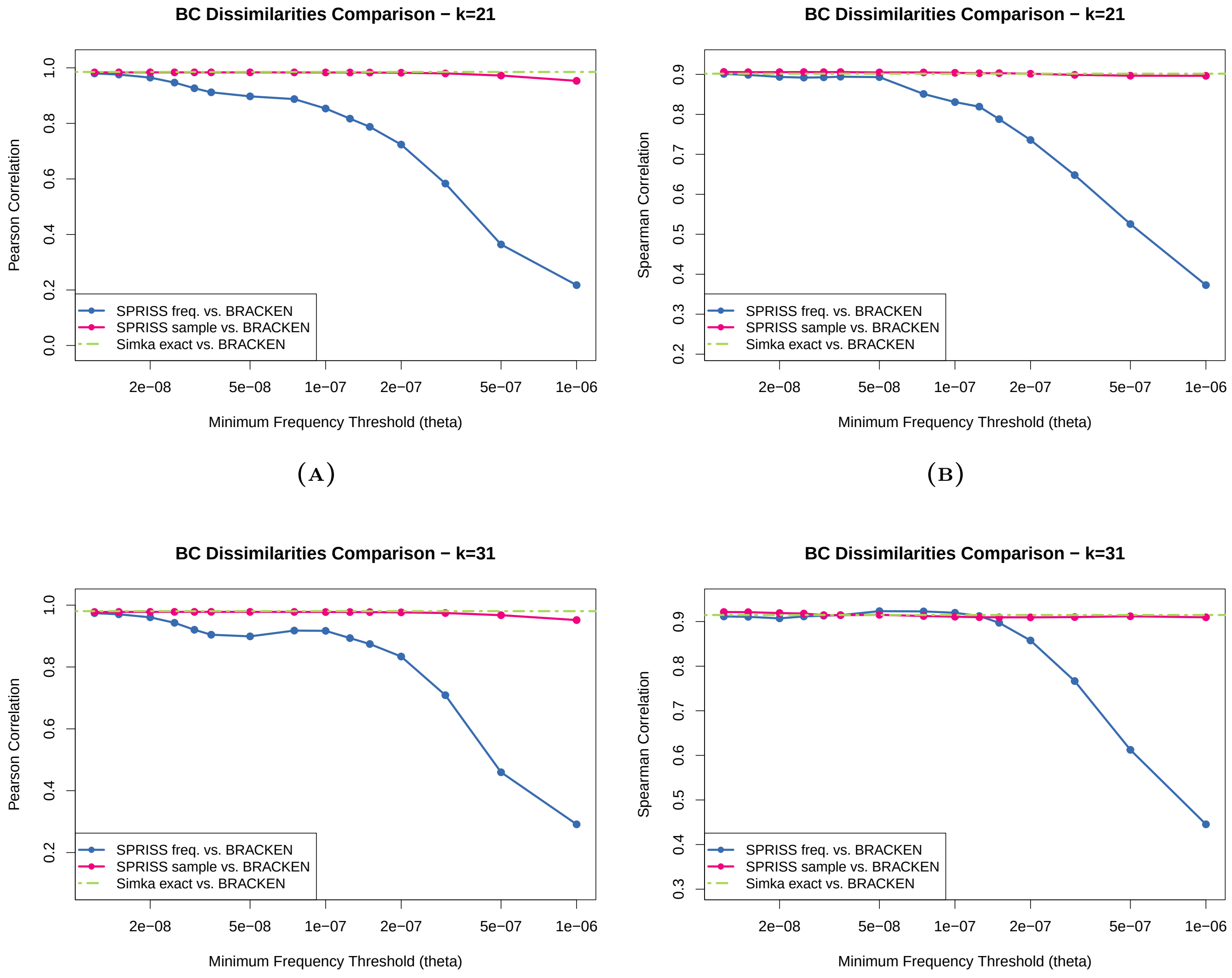
(d)

Figure C.2: Correlation between Spriss-sample-based and true BC indices on the **CAMI dataset**. Also correlation between truth and Bracken-based BC dissimilarity at species level is plotted. In legend, "Spriss freq." stands for dissimilarity computed on frequent k-mers' estimates of abundance; "Spriss sample" denotes the exact index computed by Simka, with unique k-mers removal, on metagenomes subsampled by Spriss.



(A) (B) (C) (D)

Figure C.3: Spriss-sample-based and Bracken-based BC indices referred to truth, at species level, on the **CAMI dataset**. On the right, a vertical dashed red line indicates the lowest frequency threshold applied: at its left, the dissimilarity values on complete datasets, with unique k-mers excluded, as computed by Simka are reported.



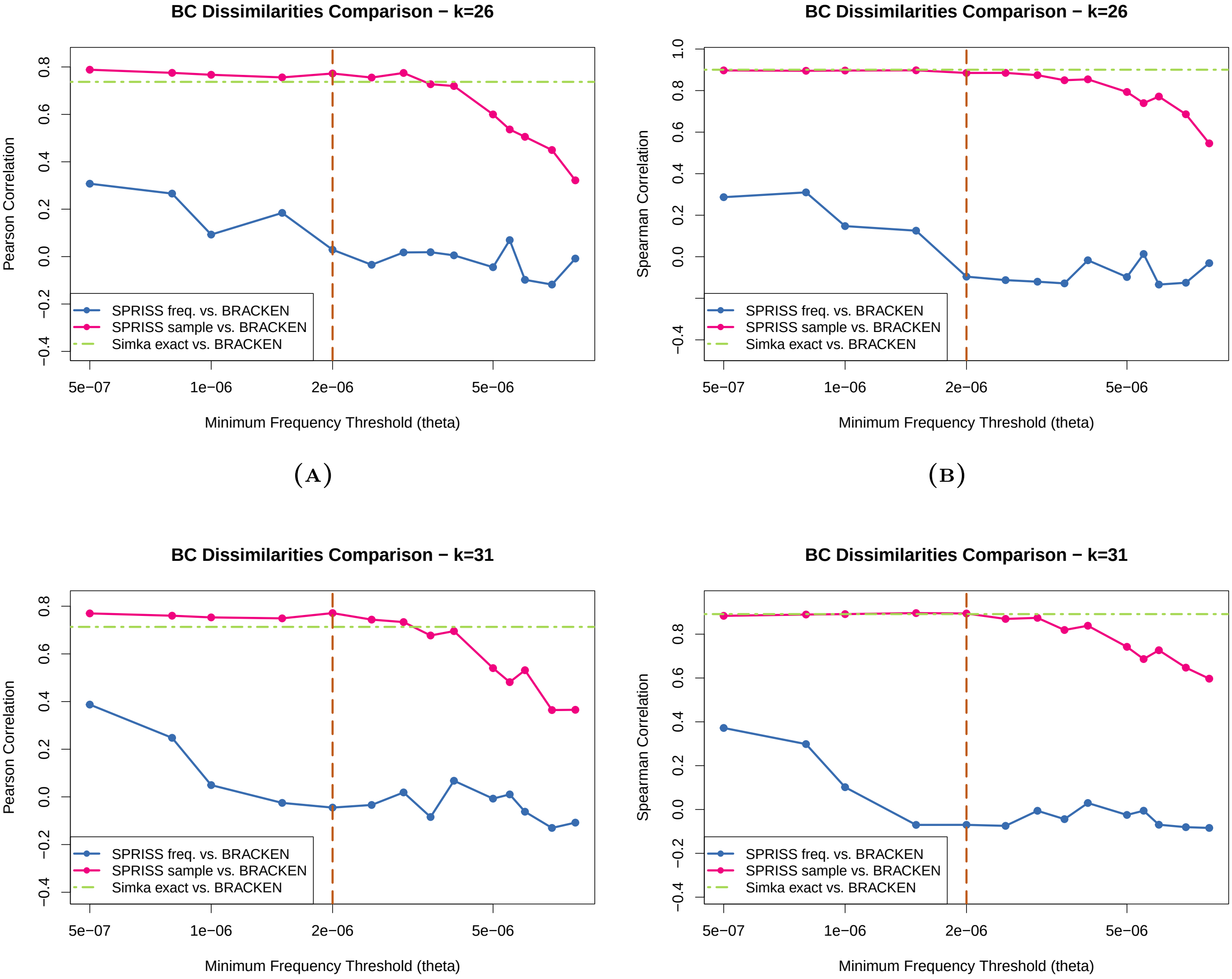
(A)

(B)

(C)

(D)

Figure C.4: Correlation between Spriss-sample-based and Bracken-based BC indices on the **HMP dataset**; In legend, "Spriss freq." stands for dissimilarity computed on frequent k-mers' estimates of abundance, "Spriss sample" denotes the exact index computed by Simka, with unique k-mers removal, on metagenomes subsampled by Spriss.



(a)

(b)

(c)

(d)

Figure C.5: Correlation between Spriss-based and Bracken-based BC indices on the **reduced GOS dataset** on long k-mers. In legend, "Spriss freq." stands for dissimilarity computed on frequent k-mers' estimates of abundance, "Spriss sample" denotes the exact index computed by Simka, with unique k-mers removal, on metagenomes subsampled by Spriss.

Glossary

Abundance vector Given N assemblages $\{\, \mathcal{A}_1 \ldots \mathcal{A}_N \,\}$ of individuals and a set $\mathcal{T}$ of taxa with which each individual is categorised, the abundance vector of a taxon $\tau \in \mathcal{T}$ is the vector $\begin{bmatrix} x_{\mathcal{A}_1,\tau} & \ldots & x_{\mathcal{A}_N,\tau} \end{bmatrix}$ of its abundances over the N assemblages. 36, 43

Assemblage In our scope, an assemblage is the set of individuals exposed to our sampling effort in a defined area or point [17]. 15, 16, 23, 25, 26

Binning The process categorising elements on the basis of a set of "known" categories.

Metagenomic binning The process of grouping metagenomic reads or contigs by their organism of origin. 99

Taxonomic binning reference-based metagenomic binning, which can be broadly devided into three categories: **1.** alignment-based, **2.** marker-based, or **3.** sequence-composition-based [11]. 6, 7

Biodiversity [*contraction of "biological diversity", from βίος (bíos) "life" and -λογία (-logía) "study of"*] The variety of life on Earth at all its levels, from genes to ecosystem, and ecological and evolutionary processes that sustain it. See [3]. 9, 99

Biosphere Our global ecosystem, that is all the "life-supporting stratum of Earth" [*Encyclopedia Britannica* (Oct 2022)]. 5, 11, 17

Cluster Given a set $\{\, S \,\}$ of elements, a cluster is a subset $\{\, C \,\} \in \{\, S \,\}$ such that its elements are similar to each others whereas they are different from elements outside $\{\, C \,\}$. Usually, the notion of dissimilarity is mathematically captured by that of distance over the metric space to which the set $\{\, S \,\}$ should belong. **Clustering**, then, is either the process or the result of partitioning a set into clusters. 10, 17, 28, 33, 100

Community As concerns biology, the populations of different species that naturally occur and interact in a particular environment. See [3, Chapter 14]. 5, 9–11, 14, 100

Microbial comunity A community of microbes [*μικρός (mikrós) "small" and βίος (bíos) "life"*], which are microscopic living organisms. 5–7, 10, 11

Conservation biology A multidisciplinary science addressed to investigate ecosystem perturbations in order to protect and maintain biodiversity.

See, among others, Gerger L. (2010), *Conservation biology* and Soulé M.E. (1985), *What is conservation biology?*. 5, 6, 31

Contig [*from contiguous*] It is a set of overlapping DNA segments that together represent a consensus region of DNA. 99

Dimensionality reduction A transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data. A variety of strategies are available: clustering, subsampling a sketching are three examples of them. 7

Distance *see* metric

Diversity measure A measure of how much two entities are incorrelated. 7

α-diversity Some measure of diversity between species within an environment or a sample. 6, 13, 16–18, 23, 25–28, 32–34

β-diversity Some measure of diversity between distinct communities, ecosystems or samples. 6, 7, 12, 13, 16–18, 26–28, 30, 31, 33, 34, 37, 53, 77

γ-diversity A measure of the overall diversity within a large region (composed of many communities). 16, 25, 27, 28, 32–34

Ecology The study of relationships between organisms and their environment.

The term was coined in 1866 by the zoologist Ernst Haeckel from Ancient Greek οἶκος *(oîkos)* "house", and -λογία *(-logía)* "study of". 5, 10

Ecosystem or **ecological system**, is "A community plus the physical environment that it occupies at a given time" [3]. In other words, it consists of all biotic and abiotic components which, in some environment at some time, are linked together through nutrient cycles and energy flows. 5, 9, 11, 14, 99, 100, *see also* ecology

Evolutionary biology The subfield of biology that studies the evolutionary processes that produced the diversity of life on Earth. 5, 6

Kmer Substring of k countiguous nucelotides (bases) of a genetic sequence. 1, 3, 7, 8, 14, 35, 36, 38, 39, 41–48, 52–55, 59, 61–84, 91, 94–97

Metagenome "The collective genomes and genes from the member of a microbiota." [7]

The word was derived from *genome*, which is the whole genetic information of an organism, with the prefix *meta*, from Ancient Greek μετά, which in this context means "beyond". We can therefore state that a metagenome is the (necessarily incomplete) genetic information of an environment rather than that of an individual organism. Hence, it is worth noticing that besides genetic sequences from several organisms, metagenomic data should come with many relevant "metadata" describing the environment from which the sample was taken and its physical state.

From a practical point of view, we call a metagenome also the collection of reads sequenced from an environment – i.e., a **metagenomic sample**. 5, 6, 35, 101

Metagenomics The study of metagenomes. It is “Both a set of research techniques, comprising many related approaches and methods, and a research field” [1, p. 13]. Metagenomics aims at “understand biology at the aggregate level, transcending the individual organism to focus on the genes in the community and how genes might influence each other’s activities in serving collective functions”. 5, 7, 9, 17

Metric A *communtative function* $d\colon \{ M \} \times \{ M \} \to \mathbb{R}$ on a set $\{ M \}$ such that $d(x,y) = 0 \iff x = y$ and the *triangular inequality* holds. 12, 13, 101

Metric space A couple $(\{ M \}, d)$, where $\{ M \}$ is a set and d a metric on $\{ M \}$. 12, 13, 19, 24

Microbiome A “characteristic microbial community occupying a reasonable well-defined habitat which has distinct physio-chemical properties. It not only refers to the microorganisms involved but also encompass their theatre of activity, which results in the formation of specific ecological niches. The microbiome, which forms a dynamic and interactive micro-ecosystem prone to change in time and scale, is integrated in macro-ecosystems including eukaryotic hosts, and here crucial for their functioning and health” [7]. See the cited article for the etymology of the word. 6, 7, 11, 12

Microbiota “The assemblage of living microorganisms (bacteria, archaea, fungi, protists and algae) present in a defined environment. As phages, viruses, plasmids, prions, viroids, and free DNA are usually not considered as living microorganisms, they do not belong to the microbiota” [7]. See the cited article for the etymology of the word. 5, 11, 100

Phylogenetics [*from Ancient Greek φυλή/φῦλον (phylé/phylon) “tribe, clan, race”, and γενετικός (genetikós) “origin, source, birth”*] The study of the evolutionary relatedness among groups of organisms.

Phylogeny Also known as **phylogenetic/evolutionary tree**, is a branching diagram or a tree showing the evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical or genetic characteristics. Not to be confused with taxonomy. 24, 25, 31, *see also* phylogenetics

Population Group of individuals of the same species that share aspects of their genetics or demography more closely with each other than with other groups of individuals of that species (where demography is the statistical characteristic of the population such as size, density, birth and death rates, distribution, and movement of migration), [3]. 10, 99

Read In bioinformatics, a genomic sequence output by a sequencing machine. NGS reads are usually 80 bp to 500 bp long. 7, 17, 35–37, 39, 83, 99, 100

Reference-based In bioinformatics, any reference-based approach is one that requires to query a database of references. For instance, reference-based metagenomics comparison needs to “map the reads obtained from sequencing the microbiome against a database of reference genomes” [11], taxa or functions. 7, 17, 35, 83, 99

Reference-free In bioinformatics, any reference-free approach is one that does not rely on external sets of references [11]. 7, 17, 35, 84

Relic DNA extracellular DNA derived from dead cells. 12

Sketch In a streaming model, a space-effcient data structure that can be used to provide estimates of (statistical) characteristics of a data stream. Therefore, **sketching** is the "process of generating — such — an approximate, compact summary of data" [32]. 36, 42, 44, 47, 100

Species This is a notoriously ambiguous concept in biology. Its definition we like the most is "The basic unit of classification and a taxonomic rank of an organism, as well as a unit of biodiversity" [*Wikipedia* (Oct. 2022)].

Alternatively, species are commonly defined as "groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups" [Ernst Mayr (1942)] and generating a fertile offspring. However, this definition is problematic as it only considers sexual organisms and, even within these, there exist pathological cases such as that of *Larus gulls.* Several other characterisations exist, — morphological, phylogenetic, ecological, etc. — none of which is satisfactory.

In any case, species should be regarded as conventional groups, not unequivocally defined and not even sharply delimited, rather than "real entities". 6, 7, 10, 99–101, *see also* taxon

Species abundance The number (cardinality) of individuals of a species in an environment. It is sometimes a synonymous of *species frequency*, which is the relative abundance of a species in an environment – i.e., species abundance divided by the total number of individuals in the environment. 6, 14

Species richness "The number of different species in a particular area" [3]. 6, 14

Species turnover Rate, or magnitude, of change in species composition *along predifined* spatial, environmental or time *gradients* [40]. 26

Streaming model A computational model consisting of a sequential machine with limited ammount of working memory and a continuous (one-way) stream in input. 102

Subsample A small portion of a larger sample. 7, 100

Taxon plural *taxa*, is "any unit used in the science of biological classification, or taxonomy. Taxa are arranged in a hierarchy from kingdom to subspecies, a given taxon ordinarily including several taxa of lower rank. In the classification of protists, plants, and animals, certain taxonomic categories are universally recognised; in descending order, these are kingdom, phylum (in plants, division), class, order, family, genus, species, and subspecies, or race" [*Encyclopedia Britannica* (Oct 2022)]. 39

Taxonomy "In a broad sense the science of classification, but more strictly the classification of living and extinct organisms – i.e., biological classification"

[*Encyclopedia Britannica* (Oct 2022)].
The term is derived from the Ancient Greek τάξις *(táxis)* “arrangement, ordering” and νόμος *(nómos)* “law, custom”. “Taxonomy is, therefore, the methodology and principles of systematic botany and zoology and sets up arrangements of the kinds of plants and animals in hierarchies of superior and subordinate groups”. 10, 101

Abbreviations

Alignment *Sequence alignment*, or, more specifically, *genome alignment*.
It is a bioinformatic techinque that aims at finding the best matching between two sequences of characters. We use this term also to concisely indicate the process of aligning (matching) a sequence against each sequence in a reference set and electing best alignment obtained. 99

BC Bray-Curtis dissimilarity index. 30, 31, 37, 42, 44, 46, 49–52, 54, 56–61, 64–77, 83, 84, 94–97

CAMI **C**ritical **A**ssessment of **M**etagenome **I**nterpretation.
A Bioinformatics challenge aimed at assessing realiability of metagenomic analysis by exploiting simulated metagenomes. See [26]. 37, 50–52, 54, 56, 57, 59, 62–66, 68–70, 75, 76, 78–80, 89, 91, 94, 95

GOS **G**lobal **O**cean **S**ampling expedition.
An expedition aimed at collecting oceanic metagenomes for further analysis. See [33] or www.jcvi.org/research/gos for more details.. 50, 51, 54, 57, 58, 61, 63, 65, 67, 72–74, 76–78, 83, 87, 88, 90, 92, 93, 97

HMP **H**uman **M**etagenome **P**roject.
A project aimed at generating resources to facilitate characterization of the human microbiota. Visit the NIH HMP site at hmpdacc.org/hmp/. 49, 50, 52, 54, 56, 58, 60, 64, 66, 71, 76–78, 81, 82, 85, 86, 92, 96

LCA **L**owest **C**ommon **A**ncestor.
In a phylogenetic tree, the ancestral taxon most near to all of a set of given taxons. In a taxonomic tree, the most specific taxonomy rank to which all taxons in a given set belong.. 39

NCBI **N**ational **C**enter for **B**iotechnology **I**nformation.
For more information, visit the site www.ncbi.nlm.nih.gov/. 37, 38, 41

NGS **N**ext **G**eneration **S**equencing.
It is any of several high-throughput approaches to DNA sequencing using the concept of *massively parallel processing*. Among them, *Illumina* and *Ion-Torrent* technologies stand out. 6, 101

OTU **O**perational **T**axonimc **U**nit.
Operational definition used to classify groups of closely related individuals. 6, 7, 10, 14–18

RNA RiboNucleic Acid. 6
WGS Whole **G**enome **S**hotgun sequencing. 32, 50

Bibliography

[1] National Research Council (US). *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet.* Washington, DC: The National Academies Press (US), 2007. DOI: 10.17226/11902. URL: https://www.ncbi.nlm.nih.gov/books/NBK54006/?report=classic (cit. on pp. 5, 101).

[2] S. R. Adke and M. V. Ratnaparkhi. "Measurement of Diversity and Dissimilarity for Stochastic Populations". In: *Biometrical Journal* 39.1 (1997), pp. 69–84. DOI: 10.1002/bimj.4710390108. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.4710390108. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.4710390108 (cit. on p. 13).

[3] Nora Bynum, ed. *What Is Biodiversity.* Feb. 5, 2009. URL: http://cnx.org/content/col10639/1.1/ (cit. on pp. 7, 9, 24, 99–102).

[4] Philip Arevalo et al. "A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations". In: *Cell* 178.4 (Aug. 2019), 820–834.e14. ISSN: 0092-8674. DOI: 10.1016/j.cell.2019.06.033. URL: https://www.sciencedirect.com/science/article/pii/S0092867419307366 (cit. on p. 10).

[5] Gaëtan Benoit et al. "Multiple Comparative Metagenomics Using Multiset k-mer Counting". In: *PeerJ Computer Science* 2 (Nov. 2016), e94. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.94 (cit. on p. 43).

[6] Gaëtan Benoit et al. "SimkaMin: Fast and Resource Frugal De Novo Comparative Metagenomics". In: *Bioinformatics* 36.4 (Sept. 2019), pp. 1275–1276. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz685. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/4/1275/48983908/bioinformatics_36_4_1275_s2.pdf. URL: https://doi.org/10.1093/bioinformatics/btz685 (cit. on pp. 35, 44, 54, 64, 83).

[7] Gabriele Berg et al. "Microbiome Definition Re-Visited: Old Concepts and New Challenges". In: *Microbiome* 8 (1 June 30, 2020). ISSN: 2049-2618. DOI: 10.1186/s40168-020-00875-0 (cit. on pp. 5, 6, 9, 100, 101).

[8] John Roger Bray and John Thomas Curtis. "An Ordination of the Upland Forest Communities of Southern Wisconsin". In: *Ecological Monographs* 27.4 (1957), pp. 326–349. ISSN: 00129615. URL: http://www.jstor.org/stable/1942268 (visited on 03/10/2023) (cit. on p. 30).

[9] M. Luz Calle. "Statistical Analysis of Metagenomics Data". In: *Genomics & Informatics* 17.1 (2019), e6–. DOI: 10.5808/GI.2019.17.1.e6. eprint: http://genominfo.org/journal/view.php?number=549. URL: http://genominfo.org/journal/view.php?number=549 (cit. on pp. 5, 30).

[10] Anne Chao and Chun-Huo Chiu. "Species Richness: Estimation and Comparison". In: Aug. 2016, pp. 1–26. ISBN: 9781118445112. DOI: 10.1002/9781118445112.stat03432.pub2 (cit. on p. 17).

[11] Matteo Comin et al. "Comparison of Microbiome Samples: Methods and Computational Challenges". In: *Briefings in Bioinformatics* 22.1 (June 2020), pp. 88–95. ISSN: 1477-4054. DOI: 10.1093/bib/bbaa121. eprint: https://academic.oup.com/bib/article-pdf/22/1/88/35934578/bbaa121.pdf (cit. on pp. 17, 99, 101, 102).

[12] Veronika B. Dubinkina et al. "Assessment of K-Mer Spectrum Applicability for Metagenomic Dissimilarity Analysis". In: *BMC Bioinformatics* 17.38 (Jan. 16, 2016). ISSN: 1471-2105. DOI: 10.1186/s12859-015-0875-7 (cit. on pp. 17, 30, 84).

[13] Adrian Fritz et al. "CAMISIM: Simulating Metagenomes and Microbial Communities". In: *Microbiome* 7.1 (Feb. 8, 2019). ISSN: 2049-2618. DOI: 10.1186/s40168-019-0633-6 (cit. on pp. 37, 50).

[14] Nicholas J. Gotelli and Anne Chao. "Measuring and Estimating Species Richness, Species Diversity, and Biotic Similarity from Sampling Data". In: *Encyclopedia of Biodiversity (Second Edition)*. Ed. by Simon A. Levin. Second Edition. Waltham: Academic Press, 2013, pp. 195–211. ISBN: 978-0-12-384720-1. DOI: 10.1016/B978-0-12-384719-5.00424-X. URL: https://www.sciencedirect.com/science/article/pii/B978012384719500424X (cit. on pp. 15, 17, 19, 20, 23–26, 28, 29, 31, 32).

[15] Jennifer B. Hughes et al. "Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity". In: *Applied and Environmental Microbiology* 67.10 (2001), pp. 4399–4406. DOI: 10.1128/AEM.67.10.4399-4406.2001. eprint: https://journals.asm.org/doi/pdf/10.1128/AEM.67.10.4399-4406.2001. URL: https://journals.asm.org/doi/abs/10.1128/AEM.67.10.4399-4406.2001 (cit. on p. 19).

[16] Curtis Huttenhower et al. "Structure, Function and Diversity of the Healthy Human Microbiome". In: *Nature* 486.7402 (June 1, 2012), pp. 207–214. ISSN: 1474-4687. DOI: 10.1038/nature11234 (cit. on p. 49).

[17] Lou Jost, Anne Chao, and Robin Chazdon. "Compositional Similarity and Beta Diversity". In: *Biological Diversity: Frontiers in Measurement and Assessment* (Jan. 2011), pp. 66–84 (cit. on pp. 15, 26–30, 99).

[18] Gerald Jurasinski, Vroni Retzer, and Carl Beierkuhnlein. "Inventory, Differentiation, and Proportional Diversity: A Consistent Terminology for Quantifying Species Diversity". In: *Oecologia* 159.1 (Feb. 1, 2009), pp. 15–26. ISSN: 1432-1939. DOI: `10.1007/s00442-008-1190-z` (cit. on pp. 16, 18, 26).

[19] Pierre Legendre and Louis Legendre. *Numerical Ecology.* Developments in Environmental Modelling. Elsevier Science, July 3, 2012. ISBN: 9780444538697 (cit. on pp. 29, 30).

[20] J. T. Lennon et al. "How, When, and Where Relic DNA Affects Microbial Diversity". In: *mBio* 9.3 (2018), e00637–18. DOI: `10.1128/mBio.00637-18`. eprint: `https://journals.asm.org/doi/pdf/10.1128/mBio.00637-18`. URL: `https://journals.asm.org/doi/abs/10.1128/mBio.00637-18` (cit. on p. 12).

[21] Catherine Lozupone and Rob Knight. "UniFrac: a New Phylogenetic Method for Comparing Microbial Communities". In: *Applied and Environmental Microbiology* 71.12 (2005), pp. 8228–8235. DOI: `10.1128/AEM.71.12.8228-8235.2005`. URL: `https://journals.asm.org/doi/abs/10.1128/AEM.71.12.8228-8235.2005` (cit. on p. 32).

[22] Catherine A. Lozupone et al. "Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities". In: *Applied and Environmental Microbiology* 73.5 (2007), pp. 1576–1585. DOI: `10.1128/AEM.01996-06`. URL: `https://journals.asm.org/doi/abs/10.1128/AEM.01996-06` (cit. on p. 32).

[23] Jennifer Lu et al. "Bracken: Estimating Species Abundance in Metagenomics Data". In: *PeerJ Computer Science* 3 (Jan. 2017), e104. ISSN: 2376-5992. DOI: `10.7717/peerj-cs.104` (cit. on pp. 39, 40).

[24] Jennifer Lu et al. "Metagenome Analysis Using the Kraken Software Suite". In: *Nature Protocols* 17.12 (Nov. 28, 2019), pp. 2815–1839. ISSN: 1750-2799. DOI: `10.1038/s41596-022-00738-y` (cit. on p. 37).

[25] Anne E. Magurran. *Measuring Biological Diversity.* Wiley, 2004. ISBN: 9780632056330 (cit. on p. 16).

[26] Fernando Meyer et al. "Critical Assessment of Metagenome Interpretation: the second round of challenges". In: *Nature Methods* 19.4 (Apr. 1, 2022), pp. 429–440. ISSN: 1548-7105. DOI: `10.1038/s41592-022-01431-4` (cit. on pp. 37, 104).

[27] Andreas Nocker, Ching-Ying Cheung, and Anne K. Camper. "Comparison of Propidium Monoazide With Ethidium Monoazide for Differentiation of Live vs. Dead Bacteria by Selective Removal of DNA From Dead Cells". In: *Journal of Microbiological Methods* 67.2 (2006), pp. 310–320. ISSN: 0167-7012. DOI: https://doi.org/10.1016/j.mimet.2006.04.015. URL: https://www.sciencedirect.com/science/article/pii/S0167701206001254 (cit. on p. 12).

[28] Samir Okasha. *Philosophy of Science: A Very Short Introduction.* Oxford University Press, May 2002. ISBN: 9780192802835. DOI: 10.1093/actrade/9780192802835.001.0001 (cit. on p. 10).

[29] Leonardo Pellegrina, Cinzia Pizzi, and Fabio Vandin. "Fast Approximation of Frequent k -Mers and Applications to Metagenomics". In: *Journal of Computational Biology* 27 (Dec. 2019). DOI: 10.1089/cmb.2019.0314 (cit. on p. 63).

[30] Carlo Ricotta and Janos Podani. "On Some Properties of the Bray-Curtis Dissimilarity and Their Ecological Meaning". In: *Ecological Complexity* 31 (July 27, 2017), pp. 201–205. ISSN: 1476-945X. DOI: 10.1016/j.ecocom.2017.07.003. URL: https://www.sciencedirect.com/science/article/pii/S1476945X17300582 (cit. on p. 30).

[31] Carlo Ricotta, László Szeidl, and Sandrine Pavoine. "Towards a Unifying Framework for Diversity and Dissimilarity Coefficients". In: *Ecological Indicators* 129 (2021), p. 107971. ISSN: 1470-160X. DOI: 10.1016/j.ecolind.2021.107971. URL: https://www.sciencedirect.com/science/article/pii/S1470160X21006361 (cit. on pp. 13, 28).

[32] Will P. M. Rowe. "When the Levee Breaks: A Practical Guide to Sketching Algorithms for Processing the Flood of Genomic Data". In: *Genome Biology* 20.199 (Sept. 2019). ISSN: 1474-760X. DOI: 10.1186/s13059-019-1809-x (cit. on pp. 42, 44, 102).

[33] Douglas B Rusch et al. "The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific". In: *PLOS Biology* 5.3 (Mar. 2007), pp. 1–34. DOI: 10.1371/journal.pbio.0050077 (cit. on pp. 50, 54, 78, 88, 104).

[34] Diego Santoro et al. "SPRISS: Approximating Frequent K-Mers by Sampling Reads, and Applications". In: *Bioinformatics* 38.13 (May 2022), pp. 3343–3350. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac180. eprint: https://academic.oup.com/bioinformatics/article-pdf/38/13/3343/44268791/btac180.pdf. URL: https://doi.org/10.1093/bioinformatics/btac180 (cit. on pp. 35, 44, 45, 49, 50, 75, 76).

[35] Alexander Sentinella et al. “Detecting Steps in Spatial Genetic Data: Which Diversity Measures Are Best?” In: *PloS one* 17 (Mar. 2022), e0265110. DOI: 10.1371/journal.pone.0265110 (cit. on p. 30).

[36] Harpreet Singh et al. “Computational Metagenomics: State-of-the-Art, Facts and Artifacts”. In: *Metagenomics: Techniques, Applications, Challenges and Opportunities.* Ed. by Reena Singh Chopra, Chirag Chopra, and Neeta Raj Sharma. Singapore: Springer Singapore, 2020, pp. 199–227. ISBN: 978-981-15-6529-8. DOI: 10.1007/978-981-15-6529-8_13. URL: https://doi.org/10.1007/978-981-15-6529-8_13 (cit. on p. 5).

[37] Zheng Sun et al. “Challenges in Benchmarking Metagenomic Profilers”. In: *Nature Methods* 18.6 (June 1, 2021), pp. 618–626. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01141-3 (cit. on pp. 12, 17, 40, 53).

[38] Ashwani Thukral. “A Review on Measurement of Alpha Diversity in Biology”. In: *Agricultural Research Journal* 54 (Jan. 2017), p. 1. DOI: 10.5958/2395-146X.2017.00001.1 (cit. on pp. 21, 22, 24).

[39] Hanna Tuomisto. “A Consistent Terminology for Quantifying Species Diversity? Yes, It Does Exist”. In: *Oecologia* 164.4 (Feb. 1, 2010), pp. 853–860. ISSN: 1432-1939. DOI: 10.1007/s00442-010-1812-0 (cit. on pp. 15, 16, 22, 26–28).

[40] Mark Vellend. “Do Commonly Used Indices of β-Diversity Measure Species Turnover?” In: *Journal of Vegetation Science* 12.4 (Sept. 2001), pp. 545–552. ISSN: 11009233, 16541103. URL: http://www.jstor.org/stable/3237006 (visited on 10/27/2022) (cit. on pp. 26, 102).

[41] Wei Wei and David Koslicki. “Using the UniFrac metric on Whole Genome Shotgun data”. In: *bioRxiv* (2022). DOI: 10.1101/2022.01.17.476629. eprint: https://www.biorxiv.org/content/early/2022/01/28/2022.01.17.476629.full.pdf. URL: https://www.biorxiv.org/content/early/2022/01/28/2022.01.17.476629 (cit. on p. 32).

[42] R. H. Whittaker. “Evolution and Measurement of Species Diversity”. In: *Taxon* 21.2/3 (May 1972), pp. 213–251. ISSN: 00400262. URL: http://www.jstor.org/stable/1218190 (visited on 10/27/2022) (cit. on p. 27).

[43] Amy D. Willis. “Rarefaction, Alpha Diversity, and Statistics”. In: *Frontiers in Microbiology* 10 (2019). ISSN: 1664-302x. DOI: 10.3389/fmicb.2019.02407. URL: https://www.frontiersin.org/articles/10.3389/fmicb.2019.02407 (cit. on pp. 18, 21–24).

[44] Brian J. Wilsey. “An Empirical Comparison of Beta Diversity Indices in Establishing Prairies”. In: *Ecology* 91.7 (2010), pp. 1984–1988. ISSN: 00129658, 19399170. URL: http://www.jstor.org/stable/25680451 (visited on 10/25/2022) (cit. on pp. 16, 28).

[45] Vivienne Woo and Theresa Alenghat. “Epigenetic Regulation by Gut Microbiota”. In: *Gut Microbes* 14.1 (2022). DOI: 10.1080/19490976.2021.2022407 (cit. on p. 5).

[46] Derrick E. Wood, Jennifer Lu, and Ben Langmead. “Improved Metagenomic Analysis With Kraken 2”. In: *Genome Biology* 20.1 (Nov. 28, 2019), p. 257. ISSN: 1474-760x. DOI: 10.1186/s13059-019-1891-0 (cit. on pp. 35, 38).