



UNIVERSITÀ  
DI PAVIA

SCUOLA DI ALTA FORMAZIONE DOTTORALE  
MACRO-AREA SCIENZE E TECNOLOGIE

Dottorato di Ricerca in Scienze della Terra e dell'ambiente

Alice Chiodi

**Molecular markers and bioinformatics in species  
detection: two case studies in *Pelophylax* spp.  
(Amphibia, Ranidae) and in a novel bacterial  
endosymbiont, in ciliate protista**

Anno Accademico 2019-2020  
Ciclo XXXII

Coordinatore  
Prof. Roberto Sacchi

Tutor  
Prof. Roberto Sacchi

Co-tutors  
Prof. Claudio Bandi,  
Dott. Matteo Brilli

*To my family,  
To my love,  
To everyone that have made this possible*

# Table of contents

CHAPTER 1.....	6
General Introduction.....	6
CHAPTER 2.....	18
Molecular characterization of the species of the <i>Pelophylax esculentus</i> complex in Northern Italy .....	18
CHAPTER 3.....	30
SeqDex: a sequence deconvolution tool for genome separation of endosymbionts from mixed sequencing samples .....	30
CHAPTER 4.....	54
Three novel bacterial endosymbionts of <i>Spirostomum</i> spp. ....	54
CHAPTER 5.....	73
General discussion.....	73
REFERENCES .....	77
APPENDIX.....	92

# Abstract

Species detection is one of the older problems encountered by science. This detection implicitly needs the identification of several features that are peculiar of a group, or taxa, such that their cumulative presence or absence defines the category of the target. Initially, taxonomies were based on morphological characters, with some limitations as there are species lacking the needed informative features, as prokaryotic or cryptic species. From the discovery of genotype, sequences have proven as powerful features to be used to infer taxonomy and phylogenetic trees. A sample can be affiliated to a taxonomy by finding a homology to an already taxonomically classified sequence present in a public database. However, if the target sequence does not have a match, then a phylogenetic tree constructed together with the found similar sequences can be used to infer their relationships. This type of taxonomy annotation can be challenging as there is not a consensus about how much distant two taxa have to be to be considered of different species.

In this PhD thesis I consider two taxonomically challenging case studies, the first involving a complex of morphologically cryptic interbreeding Anuran species, the second a bacterial endosymbiont of a ciliate protozoa.

The first case is focused on *Pelophylax*, a genus of morphologically cryptic anuran species that show a form of sexual parasitism, called hybridogenesis, where inter-species mate produces viable offspring. Some species can be identified on a bio-acoustic basis, but the presence of hybrids makes the detection unaffordable. The identification is more precisely performed using mitochondrial DNA (mtDNA) and Short Tandem Repeats (STR) markers. I coupled these two techniques to classify animals sampled in the Po Valley, near Pavia, where two autochthonous taxa can be found together with allochthonous. A correct detection of these species is important to assess the conservation status of the local taxa and the impact of the allochthonous ones.

The second case is focused on the study of a prokaryotic endosymbiont of a ciliate Protista, where morphological traits are virtually useless to determinate the taxonomy. Due to differences between host and symbionts, it is possible to use molecular traits to discern the species, as using 16s rRNA for Bacteria and 18s rRNA for Eukaryotes. However, this approach may only provide species identification, not genomic information, which is needed to get a functional understanding of the symbiotic system, as also of the endosymbiotic species, which may be unknown. With this purpose, both organisms are sequenced together by using Whole Genome Sequencing (WGS) with Next Generation Sequencing (NGS) technologies. However, this procedure allows to obtain portion of the genomes fragmented into different contigs, which then have to be deconvolved to obtain separate genomes. We then decided to develop a fully automated tool, called SeqDex, able to deconvolve host-endosymbiont dataset by coupling partial taxonomic affiliations (homology derived) to composition analysis to predict the affiliations of all the sequences

using state of the art machine learning algorithms. The second case study is composed by three *Spirostomum* samples, which showed evidence of presence of a *Neisseriales* bacterium inside the ciliate cells. I have used SeqDex to deconvolve this dataset to reconstruct partially the endosymbionts genomes and perform functional analysis to infer their role and the nature of the relationship that bound the hosts and the bacteria.

---

# CHAPTER 1

General Introduction

# Thesis outline

The aim of my thesis is to define methodological approaches that can be used in species detection, primarily in cases where common widely used features fails.

To do so, in Chapter 2 I used molecular markers to discriminate organism belonging to cryptic taxa, in condition where most of these species coexist. The picture is complicated by the presence of a form of sexual parasitism that bond the species, as also allochthonous taxa introductions. I chose to use a double marker approach, by coupling mitochondrial DNA phylogenetic signal with nuclear DNA regions, able to give species specific discrimination.

Chapter 3 will describe the tool developed to deconvolve dataset of whole genome sequencing of both host and its endosymbionts by using state of the art machine learning classifiers and k-mer frequencies. I used this tool in Chapter 4 to deconvolve the datasets of three different *Spirostomum* samples sequenced with their endosymbionts. The tool developed in Chapter 3 allowed to obtain genomic sequences of the three bacteria, which have been used to perform functional analysis and infer the relationships that bond hosts and endosymbionts in Chapter 4.

Due to the high diversity between the two case studies, Chapter 1 is a general introduction to the problems that will be discussed in detail in the Chapters 2, 3 and 4, while Chapter 5 is a general discussion of the two case studies.

## The problem of species classification

The classification of organisms is one of the oldest problems faced by science. The capability of including an organism in a precise taxon requires the identification of several features that are peculiar of the known taxa such that their cumulative presence or absence defines the category of the target. This bears the need to have previously defined a taxon by means of its peculiar and conserved characteristic. Classification initially was mostly focused on the description of edible, beneficial or poisonous species in order to preserve and transmit this information. In ancient Egypt, these species were represented in wall paintings or transcribed into paper rolls, as the Ebers Papyrus (1550 BC), one of the oldest papyrus rolls containing the description of medical plants (Aboelsoud, 2010). Later, Aristotele (382-322 BC) expanded this description to all organism (Mayr, 1982). His purpose was not only to record which organisms are harmful or beneficial to humans, but to find the characteristics shared by similar organisms and assign names reflecting those peculiarities and similarities to others, giving birth to the first detailed classification of living things.

# Historical considerations from Linnaeus to Woese

The Aristotelian classification was maintained until the 18th century, when Carl Linnaeus used it as a basis for creating the binary nomenclature still in use. Linnaeus introduced species names coupled with the Genus, and also the ranked hierarchy where species with similarities have to be grouped together in higher order groups, the actual Genera, Families, Orders, Classes, Phyla and Kingdoms (Linnaeus, 1753; Linné, 1735). Key to the success of the Linnaean systematics was the identification of traits (or combinations thereof) with high discriminatory power. He took advantage of these features to write dichotomous keys allowing to define species affiliations. This is the birth of modern taxonomy. A species was defined as a group of organisms that share fixed properties (a type) and is nowadays referred to as typological species definition (Gould, 1979; Smith, 1989). This definition bears a lot of problems, as there are cases where variation at certain features is not sufficient for discrimination or might be absent, such as in microorganisms or cryptic species, which results in the inclusion of potentially unrelated organisms in a same taxon (Lewin, 1981; Ruse, 1969). The definition of species was reviewed in 1970 by Robert Sokal, who introduced the phenetic concept of species: a group of organisms constitute a species if they have a similar phenotype (term used to indicate all the observable characteristics) which is different from other organisms (Sokal and Crovello, 1970). It differs from the typological species concept as it involves distance/similarity matrix summarizing the comparison among organisms based on multiple characteristics to cluster the organisms, and it accepts some degree of variation. This concept is still used nowadays by taxonomists to recognize species in the field (Ghiselin, 1974). It is a useful operational definition which allows easy classification of organisms, but still bears the problems of the typological species concept: it is hard to apply to organisms that lacks clear informative characters.

The Linnaean nomenclature is still used nowadays, even if it changed meaning after the evolutionary theory was published by Charles Darwin in *On the origin of species* (1859). As more and more evidences in support to Darwin's theory were found, academics begun to interpret the taxonomy in an evolutionary perspective, leading to *evolutionary systematics*. Focusing on its persistence over time, the species was defined by George Gaylord Simpson (1951) as "an entity composed of organisms which maintain their identity from other such entities through time and over space, and which has its own independent evolutionary fate and historical tendencies" (Laporte, 1994; Mayden, 1997). Evolutionary systematics rose from studies that linked fossils to modern species and it is focused on reconstructing the evolutionary history of current and past organisms, together (Cavalier-Smith, 2010; Chambers, 2009; Huxley, 1882).

In the early 1900s, another approach entered the toolbox of taxonomists: *cladistics* or *phylogenetic systematics*, thanks to the birth of phylogenetic methods. In particular, phylogeny aims to reconstruct the evolutionary history and relationships among organisms by using similarities/dissimilarities between groups of taxa and then report the most probable tree. The first phylogenetic tree of life was drawn by Ernst Haeckel in 1866 (Figure 1), even though the term phylogenesis was introduced only in 1921, by Robert John Tillyard (Tillyard M. A., 1921), and the mathematical models still in use nowadays were developed in the second half of the 20th century, e.g. Maximum Likelihood (Cavalli-Sforza and Edwards, 1967); Parsimony (Camin and Sokal, 1965); Neighbour Joining (Saitou and Nei, 1987); Bayesian Inference using Markov Chain-Monte Carlo (Li et al., 2000; Mau et al., 1999; Rannala and Yang, 1996). The formalization of cladistics is attributed to Willi Hennig in his book *Phylogenetic systematics* (Hennig, 1966). A cladistics species is the smallest group of individuals that can be identified by a set of features. Even though cladistic and evolutionary systematics originated in the same period, the latter dominated until the past few decades.

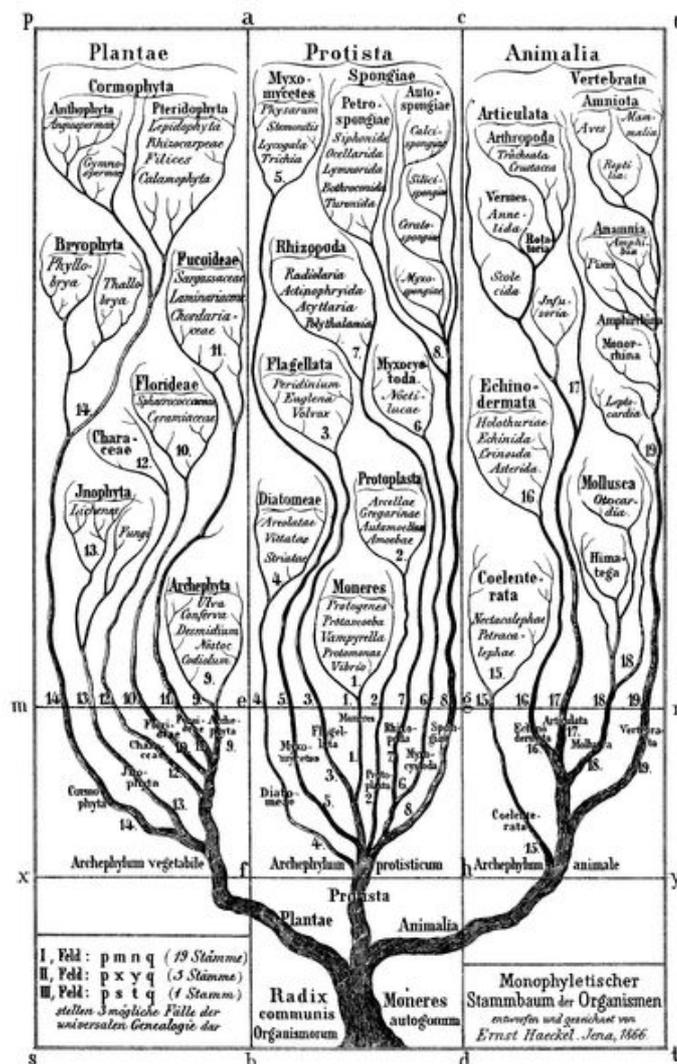


Figure 1 - The tree of life drawn by Ernst Haeckel in 1866; figure from: (Hossfeld and Levit, 2016). The root suggests the idea of a common primordial ancestor from which all other forms emerged. Haeckel drew this tree based on paleontological, embryological and systematic data.

All these classifications were based on phenotypic features, since molecular characters were not accessible until 20th century. The major limitation was that it was hard to take into account phenotypic plasticity, leading to possible misclassifications. For instance: it is impossible to correctly assign cryptic taxa, as they lack distinctive characters with respect to their taxon; moreover, some morphological characters can be typical of, and limited to, some life stages or gender, hindering the identification of the set of features with maximum informative power. For this reason, methods based on phenotypic traits are highly dependent on the ability and the knowledge of the target organism by specialists, and misclassifications are consequently common.

At the beginning of 1900 Hugo de Vries rediscovered Gregor Mendel's works, published in 1866, leading to the rise of genetics and an explosion of new disciplines. *Genetics* focused at the beginning on the identification of hereditary blocks, called genes, and the mechanisms by which these interact. The term genetics was first coined by William Bateson (Bateson, 1906). This discipline was not only influenced by Mendel's works, but also by the discoveries of nucleic acids by Friedrich Miescher (1869), of chromosomes and their behaviour during cell division (Walther Flemming, Eduard Strasburger and Edouard Van Beneden, 1880-1890), and by the theory of the equilibrium model of a gene in a population proposed by G. H. Hardy and Wilhelm Weinberg (Hardy, 1908). *Molecular biology* arose in 1930, even though the term was coined in 1938 by Warren Weaver, as a synthesis of genetics, biochemistry, microbiology and physics and studies life phenomena and organisms focusing on macromolecules, either nucleic acids and/or proteins (Weaver, 1970). *Population genetics* arose first with the work of Ronald Fisher in 1919 (Fisher, 1919, 1930), and then of J. B. S. Haldane, Sewall Wright, John Maynard Smith and Theodosius Dobzhansky. It applies genetic mechanisms to the study of population dynamics.

In the first half of the 20th century, Ernst Mayr (Mayr, 1943) defined the species as a group of populations potentially interbreeding and reproductively isolated by others. This definition, named 'the biological species concept', is one of the most widely used and encounters problems when dealing with: asexual taxa, such as prokaryotes, or parthenogenetic species (Gevers et al., 2005; Rosselló-Mora, 2001; Templeton, 1989); interbreeding taxa, like those that admit some degree of hybridization, or ring species (Figure 2), where the members of adjacent populations interbreed but distant ones cannot (Zachos, 2016); cases where it is impossible (or hardly possible) to verify reproductive isolation, like in extinct taxa (Teueman, 1924).

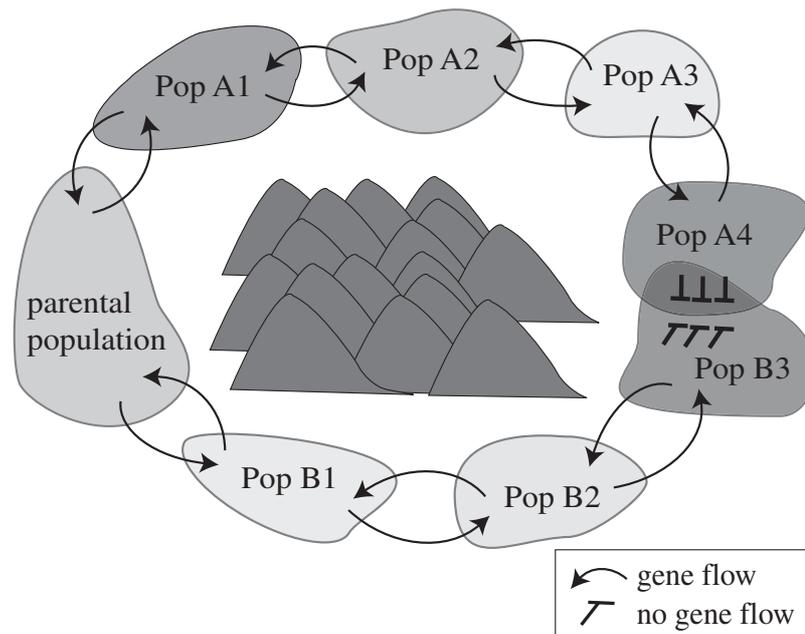


Figure 2 - The figure reassumes the problematics of ring species. A parental population diverges along two fronts of a barrier in A and B. The populations of each lineage can interbreed and allow some degree of gene flow, however the terminal populations that came in contact at the opposite fronts of the barrier cannot hybridize and there is not gene flow. Figure from: (Cacho and Baum, 2012)

All these disciplines needed to be supported by technical discoveries for data manipulation and for hypothesis testing. Among all, the description of the DNA double helix structure (Watson and Crick, 1953) and the experimental procedure to isolate nucleic acids (Avery et al., 1944) opened the door to the DNA era. Starting with this era, evolutionary systematics has been replaced by cladistics, which moreover shifted its focus from morphological characters to genes and genomes. The advancements in genetics influenced molecular biology, population genetics and phylogenetics studies. By applying the evolutionary theory to genetics, scientists observed that even distantly related organisms may share features, as the molecular apparatuses for DNA and protein synthesis. The comparisons between these highlighted that their sequences are highly similar but not identical and the variations may be the results of mutations that can be used as informative features on which one could implement phylogenetic methods to be used to define taxonomy. In 1977 two laboratories independently developed the first sequencing technologies and were able to sequence the first genomes (Maxam and Gilbert, 1977; Sanger et al., 1977). The results obtained during this DNA era incited some scientists to focus on genomes, leading to the rise of the genomic era and the development of more affordable, faster and cheaper technologies, nowadays called Next Generation Sequencing (NGS) methods. These technologies are based on another important methodological discovery of this period: the Polymerase Chain Reaction (PCR), by Kary Mullis (Mullis et al., 1987). This technique allows to duplicate exponentially a DNA molecule to obtain enough copies for sequence it. The availability of genes and proteins sequences allowed to pinpoint the presence of homologies, with a certain degree of variation, shared between even distant taxa. The

mathematical models developed for the morphology based phylogenesis are then used to infer the phylogenetic relationship on distance matrix calculated on the differences of aligned sequences.

In the early 1960s, Émile Zuckerkandl and Linus Pauling coupled speciation events dated with fossils information with differences among species sequences and proposed the molecular clock concept: by observing a fairly constant substitution rate among lineages, they proposed that this rate can be considered constant and therefore the distance based on the comparison of the sequences can be translated into time since evolutionary separation, with the possibility of dating speciation events (Zuckerkandl and Pauling, 1962). This “molecular clock” hypothesis was based only on empirical evidences, but the theoretical basis soon followed when Motoo Kimura completed the theoretical formulation of the neutral theory (Kimura, 1968). Unfortunately, the next years showed that the substitution rate of different proteins or within the same protein but at different sites is not always as constant as proposed, leading to possible problems in the translation of sequence differences into time units.

Initially, phylogenetic studies on sequences were focused only on eukaryotes. Microbiologists used phenetic traits, some showed in Figure 3, to differentiate among species. Carl Woese in the sixties was working on bacterial proteins, the protein making machinery, the ribosomes, and rRNA genes (Prakash et al., 2013). Based on its results and intuition, he supposed that ribosomes were the most conserved macromolecular complexes in prokaryotes and therefore they may represent the best targets to reconstruct their phylogenetic relationship. Then, he begun to catalogue the ribosomal genes sequences of known prokaryotes around 1966. In 1977 he proposed a method to use the SSU rRNA (Small Subunit ribosomal RNA) to reconstruct the relationships among taxa and published the first prokaryotic phylogenetic tree (Prakash et al., 2013; Woese and Fox, 1977). This approach proved to be so powerful to revolutionize prokaryotes studies, preparing the ground to the analysis of complex samples through metagenomic. In this latter field, relevants have been the works of Norman R. Pace, one of the collaborators of Woese, which implemented an experimental procedure to use a ribosomal based taxonomy even on not cultivable microorganisms. In detail, he cloned target sequences into cultivable organisms to then sequence them (Pace et al., 1986).

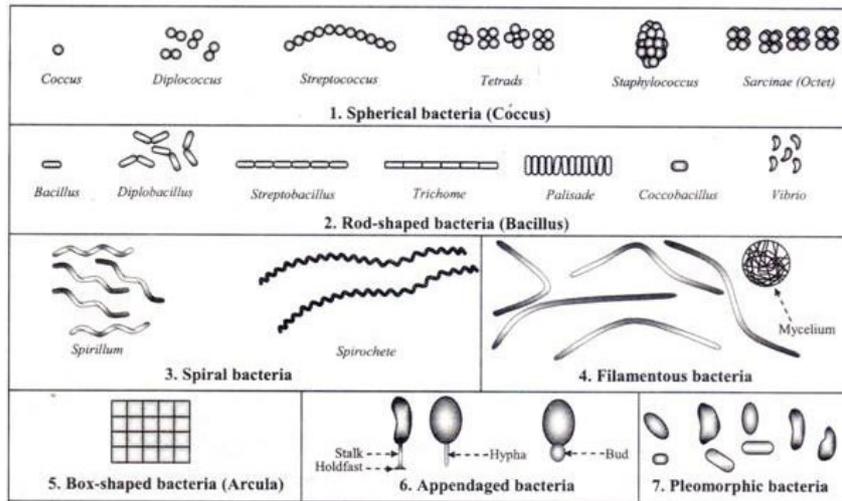


Figure 3 - The figure shows the different morphological features used to phenetically classify bacteria. (credits: microbiologyinfo.com)

Of all the SSU macromolecular components, the 16S sequence has been chosen as a marker as it was supposed that it was present in all bacteria (Woese et al., 1990). The comparison of sequences coming from different organisms later highlighted the presence of 9 hypervariable regions interleaved with highly conserved regions (Janda and Abbott, 2007), such that the variable ones provide resolution for distinguishing different species and the conserved allow to find anchoring points for quasi-universal PCR primers. Recent studies indicated in certain cases an insufficient resolution power at genus and species level for some lineages (Adeolu and Gupta, 2013; Drancourt et al., 2000; Mignard and Flandrois, 2006; Woo et al., 2003), probably due to the presence of species sharing highly similar or identical sequences, as also the presence of some erroneous taxonomic assignment of some 16S sequences deposited in databases (Janda and Abbott, 2007). These limitations of the applicability of 16S induced scientists to search for other markers. The first eukaryotic phylogenetic marker proposed was the 18S rRNA gene and it was used to draw the first phylogenetic tree of the animal kingdom (Field et al., 1988). At the beginning of 2000s, some studies reported the difficulty in interpreting the phylogenetic relationships at the species level for some mollusc taxa (Meyer et al., 2010). This is likely due to the presence of introns in the genes and of recombination events in the eukaryotes genomes, which could lead to insertion or deletion (indel) that may create problems in the comparison (Doyle and Gaut, 2000). Some others criticisms are bound to the polyploidy (Doyle and Gaut, 2000). In 2003, Paul Hebert proposed a new approach to use gene information to determine the taxonomy of eukaryotic species, called DNA barcoding. The concept at the base was using a gene, universally present in all organisms, as a barcode to recognize species. He proposed to use mitochondrial genes instead of 18S rRNA (Hebert et al., 2003) as they seldom have indels (that can lead to non-functional frameshifts) and introns, are haploid and rarely recombine. In detail, he proposed to use the cytochrome c oxidase I (COI) gene as it showed robust conserved flanking regions found previously

by Folmer and colleagues (Folmer et al., 1994) who exploited them to design primers able to recognize phylogenetically distant taxa too. The use of this region initially showed better phylogenetic signal than any other mitochondrial gene (Knowlton and Weigt, 1998), even though later studies reported some problem in the species discrimination capability of certain taxa, inducing the search for other suitable genes.

## Sequence-based modern-day classification

The first generation of sequencing technologies, in particular the Sanger sequencing, constituted an enormous advancement for time but they were characterized by a limited throughput, such that even sequencing a full bacterial genome required significant efforts in terms of cloning, sequencing and assembling. With the recent advancements in technology, that led to the NGS methods, the situation drastically changed, even if they required the development of additional tools and the concomitant increase of computational resources to manage the huge amounts of short-reads generated in each run. Basically, the huge advantage of those new approaches is that the cloning step can be avoided, and that the throughput is so high that even tiny quantities of DNA or RNA can be used proficiently. The result of the introduction of NGS technologies was an exponential increase in the number of full genomes and specific DNA regions in the public databases, such that modern-day classification started to heavily rely on the comparison of new sequences from a sample of interest with those available in specialized databases.

One of the most widely used algorithms to compare sequences to large databases is BLAST (Basic Local Alignment Tool, (Altschul, 1990)) that is based on the Smith-Waterman algorithm. BLAST was born in an era when biological sequence databases were much smaller than today and therefore speed was not the limiting issue, and none the less is still one of the most used software, even if much faster tools are today available.

Alignment algorithms can be classified in global or local: the first ones align entire sequences while the second looks for local alignments on sub-sequences such that they are more appropriate when two sequences share homologies that do not span their entire length. This latter is preferred as it allows to use raw sequences instead of recovering sequences that are homologous over their entire length. BLAST is based on a local alignment algorithm. Its functioning is reassumed in Figure 4. Searching for local alignments make the algorithm more affordable. However, this alignment algorithm becomes computationally intensive when comparing large numbers of sequences to big databases. At today several improvements exist of this algorithm that reduce execution time.

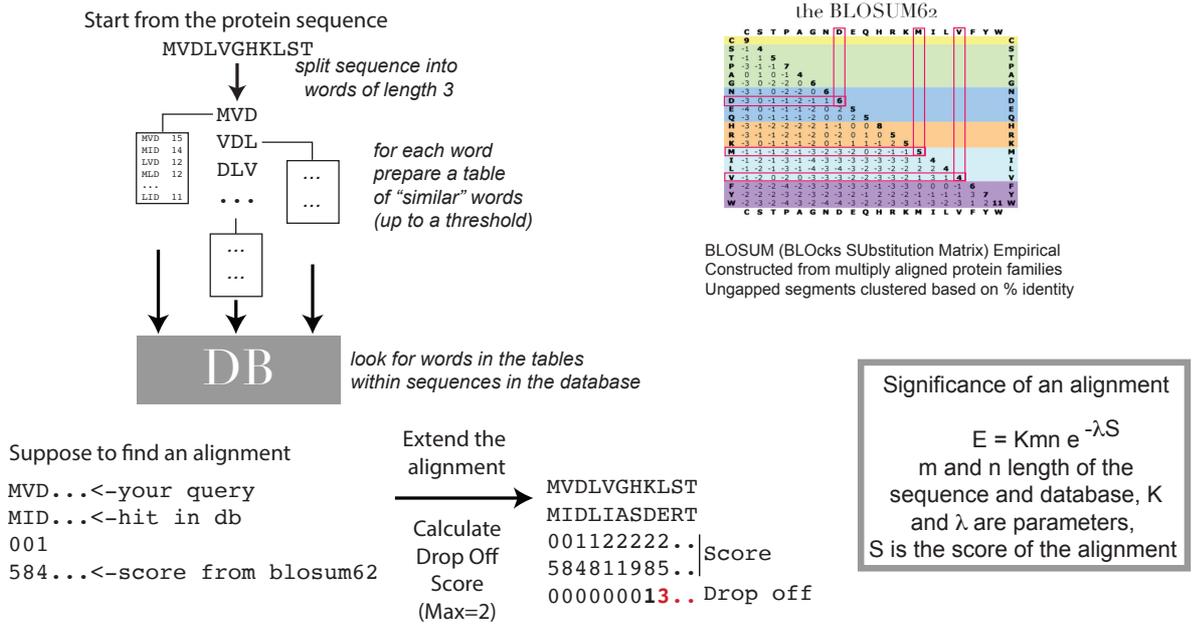


Figure 4 - The figure describes the functioning of the Smith-Waterman algorithm implemented in Blast. The algorithm starts by splitting the query sequence in all the overlapping words (for instance, w=3 for proteins in the default parameterization of Blastp). For each word  $w_i$ , the algorithm prepares a table composed by all words of length w and then exploits a substitution matrix (by default it is Blosum62), to quantify the degree of similarity with  $w_i$ . At this point, only words whose alignment's score is above a defined threshold are kept as possible hits to  $w_i$ . This procedure is repeated for all words contained in the query sequence. All the kept words are searched throughout the database, and when an identical match is found, the algorithm proceeds by trying to extend this seed alignment. Extension proceeds one amino acid (or nucleotide) at a time, and after each addition, the updated alignment score is calculated. Extension proceed up to the site that causes the score to decrease of a value larger than the drop off threshold. In this example, the similarity index increases from 5 to 29, and then further extension causes the score to drop below the threshold and therefore the algorithm stops and output a local alignment that is called HSP, high scoring pair. To guide the identification of significant alignments, Blast calculates the so-called E-value (E in the figure) that depends not only on the score S of the alignment, but also on the size of the query and the database.

To speed up the comparison, some authors developed tools that find similar sequence, with high accuracy, but avoiding the alignment step. One of the examples is the naïve Bayesian classifier implemented within the framework of the Ribosome Database Project (RDP, (Cole et al., 2014)). This algorithm takes all the 16S rDNA gene sequences in the database to calculate the probability of observing every possible word of length k (usually called k-mers, here the length is 8) (Wang et al., 2007). This can be achieved by exploiting the Jeffreys-Perks law of succession, i.e. that the probability of words is given by the number of their occurrences divided by the total number of words. Similarly, the probability of observing every word can be calculated on a subset of the entire sequence database, for instance by only considering sequences from a well-defined taxonomic group. While in the previous case the probabilities can be thought of as background probabilities in the entire database, when calculating them by only considering sequences from a certain taxonomic group, we are calculating the probabilities specific for that group. At the end of the procedure, each word has its probability calculated on the basis of the probability model developed for each taxonomic group present in the database. These models can

be used to calculate the probability that a certain sequence comes from one of the taxonomic groups present in the database by applying the Bayes theorem.

As there are no complicate calculations to be performed, alignment-free tools are very fast, even when working with large dataset. However, this kind of implementation strongly relies on the presence of a previously trained model for the taxonomic group from where the query sequence really comes from, otherwise the algorithm will output nonetheless the taxonomic group with the highest probability, only because the database misses the model for the right taxonomic rank.

The taxonomic identification of new samples based on the comparison to sequences deposited in specialized database is not free of problems. By comparing two sequences is possible to observe some degree of variation. The problem lays in the impossibility to define a consensus as different sequences may show different degrees of intra-species and inter-species variations (Janda and Abbott, 2007). This influence also the selection of the regions to be used for the comparison. In a good marker the intra-species variation must be less than the inter-species (Liu et al., 2017). Interbreeding, or horizontal gene transfer in prokaryotes, can confuse the phylogenetic signal if the gene under studies is involved in this interspecific exchange.

## Aims of the work

In this Thesis, I will consider two distinct case studies, the first concerning morphologically cryptic interbreeding species of frogs (Amphibia), and the second concerning endosymbiont bacteria of a ciliate (Protozoa). Cryptic species are challenging to recognize morphologically due to the absence of informative features. Also, as discussed above, the interbreeding capability complicate the possibility to recognize taxa by using molecular markers. Therefore, I used an approach where multiple markers are selected to differentially amplify the species and that can be combined to get the classification, even for hybrids.

Similarly, the morphology of bacteria is not taxonomically informative, such that the most common approach is today based on the analysis of DNA sequences. Even in this case, the identification/classification can be challenging, especially when dealing with species that are under-represented in the databases or strongly divergent in terms of sequences; this is particularly true having to do with endosymbionts, as their life-style most of the times implies non-cultivability and a peculiar evolutionary path, often proceeding in a completely independent way with respect to related, free-living species. As a consequence, the genomics of endosymbionts often requires to sequence the bacterium and its host together and then apply tools able to identify the sequences coming from different sources. To improve the latter step, that is the deconvolution of DNA molecules of different origin in a sequencing run containing several organisms, I implemented a tool exploiting state-of-the-

art machine learning techniques. This tool was then used to obtain the genomic sequences of the *Spirostomum* endosymbionts, and I conclude by performing computational functional analysis to better characterize the nature of this strict relationship.

---

## CHAPTER 2

Molecular characterization of the species of the *Pelophylax  
esculentus* complex in Northern Italy

# Introduction

*Pelophylax* is widespread in Eurasia and it comprises at least 12 morphologically cryptic species around the Mediterranean basin and Europe. It has been supposed that these species have originated in allopatry, as a result of the geographical isolation associated to glaciations (Canestrelli and Nascetti, 2008).

As cryptic species lack the morphological features that univocally mark the taxa, it is impossible to apply phenetic or typological species concepts. Some of these species can be discriminated by using bioacoustic records of male mating calls (Sinsch and Schneider, 2009) but even when exploitable for taxonomy assignment, this approach has huge limits: only reproductive males produce mating calls and therefore the technique is blind to females and non-reproductive males; moreover, it gives no information about population composition and dynamics.

Species discrimination based on molecular markers can also be difficult. Indeed, these taxa show an outstanding form of sexual parasitism, called hybridogenesis (Schultz, 1969) where two species mate producing a viable and fertile hybrid offspring that excludes one of the parental genomes in the germinal line cells (Figure 5). The hybrid lineage can then survive only by sharing its areal with the parental taxon whose genome has been excluded, whereas the other taxon needs to be absent.

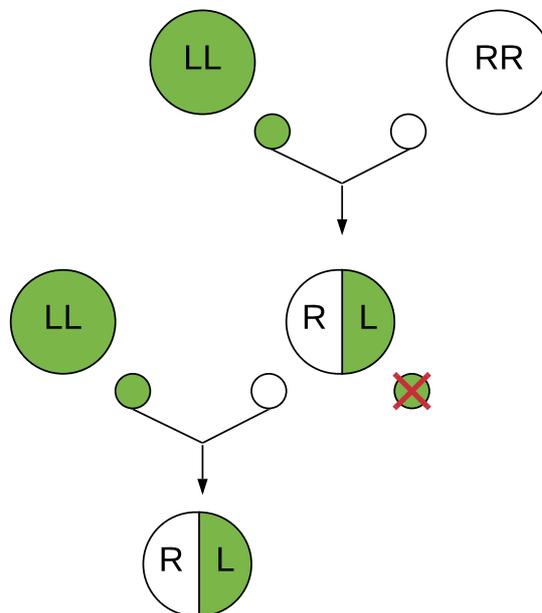


Figure 5 – The hybridogenetic system probably originated from an ancient interspecific mate between RR (*Ridibundus* lineage genome) and LL (*Lessonae* lineage genome). This produced viable offspring RL which exclude one of the parental genomes from the germinal cells, in this case L. This hybrid generation produces only gametes R and a backcross with the parental LL is needed to maintain the lineage; thus, they have to coexist. Backcross with the parental RR will produce only RR offspring, so this species needs to be absent to allow the survival of the hybrid lineage.

Various interspecific mating occurred in the history of the *Pelophylax* genus, as shown in Figure 6, leading to the formation of several hybridogenetic complexes: the *P. hispanicus* complex in the Italian peninsula (Uzzell and Hotz, 1979), the *P. grafi* complex in Spain (Graf et al., 1977), and the most studied, the *P. esculentus* complex, in Northern Italy and central Europe (Dubois, 1992; Frost et al., 2006). In this latter complex, ancient mating between *P. lessonae* and *P. ridibundus* produced the *P. esculentus* hybrid (Tunner and Heppich, 1981). This complex is one of the most common in Europe and produced different outcomes that have been studied in detail. The *P. esculentus* complex is divided into three systems: the L-E, the R-E and the E system (Christiansen and Reyer, 2009; Tunner and Heppich, 1981; Uzzell et al., 1976; Vinogradov et al., 2008). The first, and the most studied one, is diffused in Northern Italy, Swiss and France (Figure 6), and is composed by *P. lessonae* and *P. esculentus* which in its germinal line cells exclude the *P. lessonae* genome (hereafter L) and transmit the *P. ridibundus* genome (hereafter R) clonally (Pagano et al., 1997). Here, the two species need to co-occur to maintain the systems: *P. esculentus* produce only R gametes and needs L gametes to maintain the hybridogenetic lineage (Figure 5, Figure 6); instead the *P. ridibundus* parental species need to be completely absent, as crossing between them and the hybrid will produce pure RR offspring, leading to the extinction of *P. esculentus*. Mating between hybrids have been reported to produce not viable offspring in most cases (97%) as the clonally transmitted genome accumulate aberrant mutation that may be not compatible with life if in homozygosis (Holsbeek and Jooris, 2010). As male frogs usually choose a bigger female to reproduce with, it has been supposed that the first interspecific mating happened between males *P. lessonae* and females *P. ridibundus* (this species usually have body size bigger than the previous). The hybrids RL were supposed to have the mitochondrial DNA (mtDNA) of *P. ridibundus*, which is maternally inherited. However, RL hybrids show prevalence of mtDNA of *P. lessonae*, probably due to introgression of this parental character imputable to crossing with LL females, which was then maintained, as it may bring advantages in tadpole development in hypoxic water pond (Plenet et al., 2000a, 2000b; Spolsky and Uzzell, 1984).

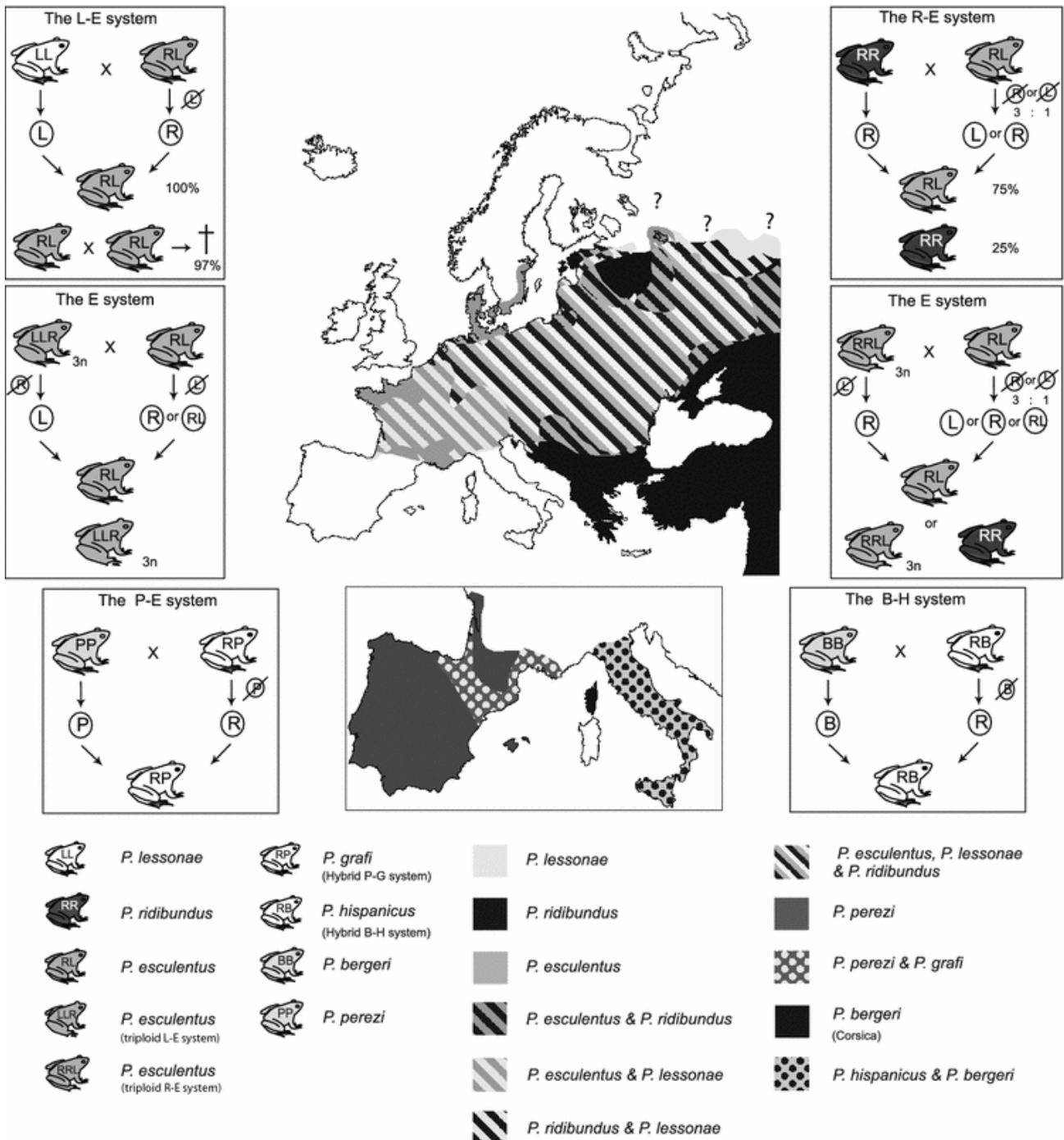


Figure 6 - Overview of the hybridogenetic systems of the *Pelophylax* complexes in Europe and the geographical distribution of the parental and hybridogenetic species. Figure from: (Holsbeek and Jooris, 2010)

The R-E system is diffused at the opposite limit of the *Pelophylax* range respect to the L-E system, in East Europe (Figure 6), and is composed by *P. ridibundus* and *P. esculentus* which exclude in the germinal line cells the R or L in a proportion of 3:1 (Uzzell et al., 1976; Vinogradov et al., 2008).

Between the L-E and R-E ranges, there are areas where all three species co-occur, or where only hybrids can be found, the E system. Here, among diploid RL hybrids, also polyploid hybrids can be found (usually triploid), which genotype is formed by one or

multiple copies of L or R, and that exclude one of the two parental genome to guarantee the hybrids lineages maintenance (Christiansen and Reyer, 2009; Uzzell et al., 1980).

In the past few decades the Italian range of the L-E system in the Po Valley has endured the introductions of alien *Pelophylax* taxa ascribable to *P. ridibundus* lineages coming from Southern, Eastern, Central and Western Europe, called *P. kurtmuelleri*, *P. bedriagae* and *P. ridibundus* (East and West Europe clade) (Lanza, 1962).

It has been supposed that these introductions may have different outcomes. The Southern, Central and Eastern taxa seem to be not capable to induce the sexual parasitism. For this reason, it has been hypothesized that the mate with the *P. lessonae* will produce sterile hybrids, which will also compete, and probably win due to higher body mass, with the autochthonous taxa (Holsbeek and Jooris, 2010; Hotz et al., 1985; Hotz and Uzzell, 1983). On the contrary, the Western Europe taxon is capable to induce hybridogenesis, and mating with *P. lessonae* will lead to establishment of new hybridogenetic lineages, that may substitute the autochthonous (Holsbeek and Jooris, 2010). Also, *P. lessonae* males should mate preferentially with *P. ridibundus* females due again to the higher body mass, leading to a reduction of density of the *P. lessonae* populations. For these reasons, methods are needed to correctly identify autochthonous and allochthonous taxa, as also to recognize the hybrids and their origins. Some approaches have been proposed, but none of these are fully comprehensive.

Some authors identified some morphological characters that might help scientists to discriminate among *P. lessonae* lineages, *P. ridibundus* lineages and their hybrid. However two problems comes with these measurements: these publications are in German, so not accessible to most, dissipating their usefulness; these features seem to be capable to discriminate lineages, but not the species (or clade), and therefore they are unable to discriminate between European *P. lessonae* and Italian *P. bergeri*, or between Center Europe *P. ridibundus* and *P. bedriagae* and *P. kurtmuelleri*. Plotner compared these morphological characters to other molecular methods to try to discern the geographical origins of individuals and so infer their species affiliations (Plotner et al., 2008). He concluded that using morphological features coupled with mtDNA ND3 gene allows to correctly assign the taxonomies. However, the mitochondrial DNA allows to reconstruct the phylogenesis of this maternally transmitted organelle, not to discriminate among hybrids, which do not have their own mtDNA haplotype, nor to reconstruct the history of both maternal and paternal genomic DNAs.

Another approach involves the use of particular genomic regions, called microsatellites or Short Tandem Repeats (STR), which are loci of repetitive DNA where a motif is repeated a certain number of times. The first microsatellite where discovered in 1984, even if the term was coined later, in 1989 (Richard et al., 2008). The use of STR regions spreads around the 1990s for paternity test. These regions show higher mutation rates compared to genes, due to the repetitions themselves. The high variability of a marker makes it suitable to obtain information on taxa at low taxonomic levels, even at the population or individual level. Some

authors found microsatellites specific for L, R or both lineages, where the two genomes are discriminated by a very different numbers of repetitions. Moreover, some of these STRs can be used to quantify the ploidy and allowing the study of polyploid hybrids (Arioli, 2007; Christiansen and Reyer, 2009; Garner et al., 2000; Zeisset et al., 2000). However, the information derived from these regions are hard to use as they were selected for discriminating L and R lineages, not the species.

We then decided to use a mixed approach, which involves the use of mtDNA to reconstruct phylogenetic history of this maternally inherited marker, coupled with the use of microsatellite loci to support mitochondrial data, to find hybrids and to infer their parental taxa. The main aim of this chapter is to use this molecular markers-based approach to identify *Pelophylax* species in the Po valley: the autochthonous and the allochthonous taxa, but also the hybrids, native or newly established.

## Material and Methods

### Field sampling

Before performing any collection of animal samples, we asked and obtained permission from the Italian Ministry of the Environment (Prot. 0003221/PNM of the 15/02/2017 DIV II). A total of 90 animals have been captured by hand or by net, mostly at night and using flashlights, from middle March until late July 2017. Sampling sites were chosen according to bibliography and to indications of presence of the taxa of interest (see Table 1). We selected sites to collect samples from populations composed by only one of the species of interest (hereafter referred as 'pure') or by multiple species mixed together (hereafter referred as 'mixed'). The allochthonous taxa are frequently present in the Po plain in mixed populations. Where possible, we chose to collect specimens of allochthonous taxa in the first site of introduction to be highly sure to sample pure populations. This was possible for *P. kurtmuelleri* and *P. ridibundus* Western Europe (Table 1).

Samples were collected by toe clipping, a technique to collect biological tissue that involves the asportation, with sterilized scissor, of the last part of a toe (except the third). Samples were preserved in ethanol 96% until subsequent analyses.

Table 1 – List of the sampling sites chosen according to (Lanza, 1962; Lanza et al., 2007)

BASIN	PROVINCE - LINEAGES (TAXA)
Ticino	Pavia - native (L-E system)
	Milano - native (L-E system)
Staffora	Pavia - alien ( <i>P. kurtmuelleri</i> , <i>P. bedriagae</i> , <i>P. ridibundus</i> East Europe)
Adige	Trento – alien ( <i>P. ridibundus</i> West Europe)
Impero	Imperia – alien ( <i>P. kurtmuelleri</i> )
Neva	Savona – alien ( <i>P. kurtmuelleri</i> )

## DNA analysis

Previous to extraction, tissue samples have been rehydrated in sterile water for 10 minutes at room temperature. The genome extractions have been performed using GenElute Mammalian Genomic DNA Miniprep Kit (Sigma-Aldrich, Saint Louis, USA), following the manufacturer instructions. For each sample, two markers were amplified: the mitochondrial gene ND3 (NADH dehydrogenase subunit 3) and 9 microsatellite loci, selected as they amplify only the L, only R, or both (see Table 2). All the amplifications were done using 0.05 U of the Hot Start Taq DNA polymerase (Biotechrabbit, Anaheim US) in a total reaction volume of 20  $\mu$ l for the ND3 and 10  $\mu$ l for the microsatellite. The thermal cycles used for each marker where the same reported in the publications.

The ND3 was amplified in each sample and then verified by gel electrophoresis; the fragments of the target length have been excised and purified using GenElute Gel Extraction Kit (Sigma-Aldrich, Saint Louis, USA), following the manufacturer's instructions. The purified PCR products have been sequenced using the forward primer (Eurofins Genomics, Ebersberg, Germany).

Considering the STR marker, the forward primer of each microsatellite was fluorescently labelled for detection on an ABI3130 capillary sequencer. Each STR locus was amplified individually but, previous to sequencing, we mixed them considering fluorescence dye and expected lengths. Sequencing was carried out at Eurofins Genomics (Ebersberg, Germany).

Table 2 – List of all the primer used to amplify the 9 microsatellite loci and the mtDNA ND3 marker

Primer	Sequence (5'-3')	Type	Target	Reference
ND3 L	AGTACACGTGACTTCCAATC	mtDNA	L, R	(Plotner et al., 2008)
ND3 H	TTGAGCCGAAATCAACTGTC	mtDNA	L, R	(Plotner et al., 2008)
RICA18 F	CTCTGCTCCCTCAGCTATGC	STR	L	(Garner et al., 2000)
RICA18 R	AAAAAGTGGTCCTTTCATTTTGAG	STR	L	(Garner et al., 2000)
RICA1a27 F	CAAATGGGTCATCCACACC	STR	L	(Christiansen and Reyer, 2009)
RICA1a27 R	GTTCAAGGGGGTCGAAATAC	STR	L	(Christiansen and Reyer, 2009)
RICA5 F	CTTCCACTTTGCCCATCAAG	STR	L	(Garner et al., 2000)
RICA5 R	ATGTGTCTGGCAGCTATGTTC	STR	L	(Garner et al., 2000)
Res22 F	ATACAGGGCTTAGTGAAATGAA	STR	R	(Zeisset et al., 2000)
Ress22 R	AAGGGGTAAAGGTGTGACTAT	STR	R	(Zeisset et al., 2000)
Re2CAGA3 F	ATGTCGTTAGAGTTCATAGG	STR	R	(Arioli, 2007)
Re2CAGA3 R	ATCTCAAGTAATCTGTCTGTC	STR	R	(Arioli, 2007)
Rrid169A F	CGGAACTCCGCTTTAATCAC	STR	R	(Christiansen and Reyer, 2009)
Rrid169A R	CCCATGTTGTCGTTGAGCTA	STR	R	(Christiansen and Reyer, 2009)
Ga1a19red F	GCACACTATTTCTGCTGTATTGC	STR	L, R	(Arioli, 2007; Christiansen and Reyer, 2009)
Ga1a19 R	CAGGGGATTTTCCCATCAG	STR	L, R	(Arioli, 2007; Christiansen and Reyer, 2009)
Ca1b6 F	AAACTCGCGGTTTCCCTTAG	STR	L, R	(Arioli, 2007)
Ca1b6 R	GAGCCAGGTTAAGATAACTGGAG	STR	L, R	(Arioli, 2007)
RICA1b5 F	CCCAGTGACAGTGAGTACCG	STR	L, R	(Garner et al., 2000)
RICA1b5 R	CCCAACTGGAGGACCAAAG	STR	L, R	(Garner et al., 2000)

## Data analysis

ND3 sequences electropherograms have been imported in Geneious (Drummond et al., 2011) and manually checked. The ND3 sequences of all 90 samples have been then aligned and compared to already published sequences using online BLAST algorithm (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) to identify species-specific lineages. Also, MEGA 6 (Tamura et al., 2013) was used to build a Neighbour-Joining (NJ) tree according to p-distance between pairs, using reference sequences both from native and alien lineages previously detected (data not shown). Node supports were estimated by bootstrap procedure (1000 replicates, (Felsenstein, 1981)). *P. nigromaculatus* ND3 sequence was used as outgroup. A haplotype network of the obtained ND3 haplotypes was built in TCS1.21 (Clement et al., 2000) according to the parsimony algorithm of (Templeton et al., 1992), which estimates the minimum number of connections required to join haplotypes assuming a statistical threshold of 95%.

STR alleles scoring and dimensioning were conducted in Geneious 11. Structure 3.2 (Pritchard and Wen, 2002) was used to perform a cluster analysis: each STR allele is assigned to a group, allowing to infer the belonging of each individual to one or more taxa. We performed the Structure analysis using admixture, with 100'000 runs, burn-in of 10'000, 6 replicates per run and testing for different number of clusters (from 1 to 8). The best number of clusters was inferred using Structure Harvester (Earl and VonHoldt, 2012). Prior to clustering, we checked for deviation from Hardy-Weinberg equilibrium (HWE) by using Genepop On The Web software (Raymond and Rousset, 1995; Rousset, 2008), and for presence of artefacts in the STR data (null alleles, large allele dropout, stuttering) by using Micro-Checker (Van Oosterhout et al., 2004).

## Results

Genomic DNA was successfully obtained from all toe-clipped samples. The extracts were enough concentrated and free of inhibitors to allow both mtDNA ND3 and microsatellite amplifications.

The NJ tree and the haplotype network analyses (Figure 7) allow us to discern among 5 different mtDNA lineages, corresponding to *P. ridibundus* (Western Europe lineage, RIDW), *P. ridibundus* (Eastern Europe lineage, RIDE), *P. kurtmuelleri* (KURT), *P. bedriagae* (BED) and a clade that merges *P. lessonae* (L) and native *P. esculentus* haplotypes (E). As expected, the ND3 was not able to identify the hybrids. As discussed before, the *P. esculentus* does not have its own mitochondrial haplotype but has the *P. lessonae* one. Thus, by using only this marker is impossible to identify the hybrids.

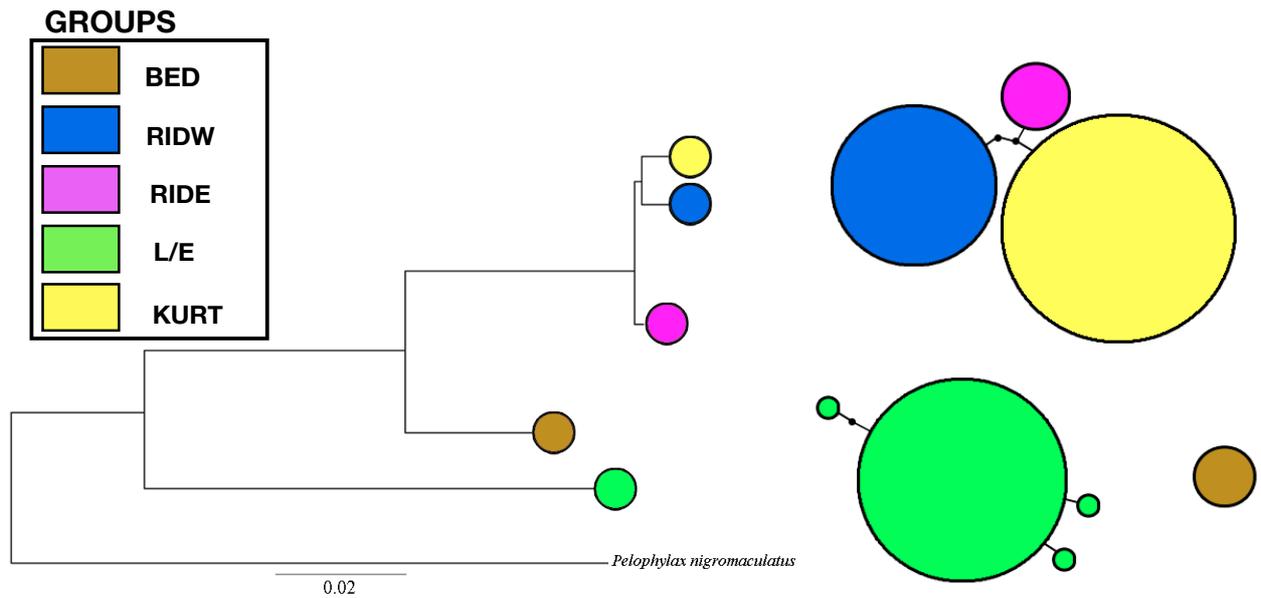


Figure 7 – On the left: the Neighbour Joining (NJ) tree constructed on the ND3 sequences of the 90 samples under analysis. *Pelophylax nigromaculatus* has been used as an outgroup. On the right: the haplotype network build using the ND3 haplotypes. Each circle is a haplotype associated to a species. In the Groups box each colour is associated to a species: BED stands for *P. bedriagae*; RIDW stands for *P. ridibundus* from Western Europe; RIDE stands for *P. ridibundus* Eastern Europe; L/E stands for *P. lessonae* and *P. esculentus*; KURT stands for *P. kurtmuelleri*.

To assess the absence of deviation from HWE and artefacts we used only the data that amplified only the L genome or the R. In detail, we obtained two sub-datasets by extracting the samples where only the STR loci L (R) specific were amplified, without the loci R (L) specific. Then, we checked for HWE and artefacts only in these two sub-datasets considering all the STR regions amplified (the three L (R) specific plus the three a-specific). No artefacts and no statistically significant HWE deviations were detected.

By using the Structure Harvester tool, we chose 4 as best number of clusters. We then used the taxonomy of the individuals of the pure populations, inferred by the sampling locality and confirmed by the ND3, to extend the information to each cluster. The Structure results are reassumed in Figure 8. The clusters correspond to (1) in red *P. ridibundus* East Europe and *P. bedriagae*; (2) in yellow *P. kurtmuelleri*; (3) in blue *P. ridibundus* West Europe; (4) in green *P. lessonae*.

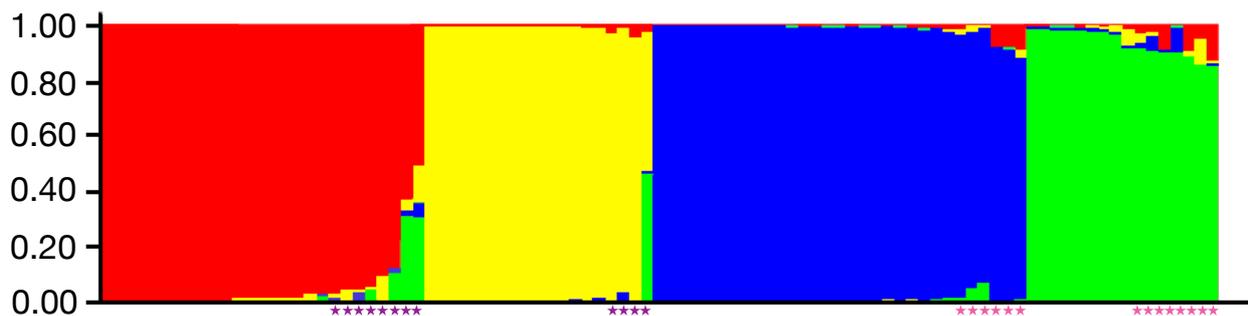


Figure 8 – The graph reassumes the results of the cluster analysis performed with Structure with  $k=4$ . On the x-axis there are the individuals and on the y-axis is reported the likelihood of the membership to each cluster. The red cluster is composed by RIDE and BED; the yellow is composed by KURT; the blue is composed by RIDW; the green is composed by L. As is possible to see, most of the samples have haplotypes completely belonging to a cluster, whereas others show a mixed signal. These are supposed to be hybrids, autochthonous (pink stars, E) and allochthonous (purple stars).

In the Figure 8, considering the likelihood of each sample, it is possible to observe the presence of individuals that fully or almost completely belongs to a unique cluster, whereas others show a mixed signal, suggesting that these can be the hybrid lineages. We hypothesized that the samples that in the figure are marked with the pink stars are the native *P. esculentus* hybrids, whereas the one marked with the purple stars are new hybrid lineages. We were able to discriminate between native and new hybrids making some considerations. The *Pelophylax* species, not considering the hybrids, can be reassumed as belonging to two lineages: the *ridibundus* and the *lessonae*. In the sampling sites we recovered only one *lessonae*-lineage species, the *P. lessonae*. However, all other species interested in this study belong to the *ridibundus*-lineage. Indeed, we are absolutely sure that the STR targeting the L genome amplify only the haplotypes present in the *P. lessonae*, but this does not hold true for the R. The Eastern and the Western Europe *P. ridibundus* are phylogenetically close, as also RIDE and *P. bedriagae*, where the latter is considered by some author a sub-species of the first. It is possible that Structure software fails to assign an organism to a cluster because the multi-locus genotypes might not be enough specific to discriminate such close *ridibundus*-lineage species. Thus, we coupled the output of the Structure with the mtDNA ND3 haplotype and the sampling sites to assess the nature of the hybrids: if the sample has a likelihood to belong to the L green cluster and it was sampled in pure *P. lessonae* populations, then we supposed it is a native *P. esculentus*; If the sample comes from mixed populations, then we supposed that it is a new hybrid.

# Discussion

The described molecular marker-based approach allows to assign a highly trustable taxonomy to *Pelophylax* tissue samples. The combined use of the two markers, the mtDNA and the STR, has proven to be a valuable approach to develop a molecular based taxonomy of these morphologically cryptic species, showing however some limitations.

The STR markers are not specifically developed to discriminate the *Pelophylax* species, thus leading to some uncertainty in the accuracy of the Structure output that we overcame by coupling it to the mtDNA and geographical data. This coupling was done by hand, which then resulted in a procedure highly dependent on the user. Using the geographical data may attempt the validity of the taxonomies inferred: we have considered the sampling sites as in bibliography allochthonous species where not recorded in certain localities, such as the valley northern to the Po river; however we cannot know if the aliens are really absent in these localities or if their presence are not already been detected.

The described molecular marker-based approach is a first attempt to use molecular data to identify morphologically cryptic species and has proven to be a valuable method. However, the limitations shown highlighted the need to search for STR markers highly species specific, not only lineage specific, as also to implement the coupling of the data to be completely user free to make this method widely applicable.

In conclusion, the results shown are promising but further advancements are needed to be able to apply this approach by routine and on huge dataset.

---

## CHAPTER 3

SeqDex: a sequence deconvolution tool for genome separation of endosymbionts from mixed sequencing samples

# Introduction

The first observations of the presence of bacteria living inside the cells of other organisms date at the beginning of the 20th century. In this period microorganisms could be studied only by using the microscope. This technique allowed Lynn Margulis to observe the features that lead her to hypothesize that the mitochondrion might have originated after an endosymbiotic event (Sagan, 1967), an idea that was accepted coldly by the scientific community. Working on the 16S rRNA gene, Woese developed the tools that allowed to obtain strong support to the endosymbiont hypothesis (Schwartz and Dayhoff, 1978), which provided renewed impetus to the theory. During the 20th century, revealed that endosymbiosis is widespread and typically involves a eukaryotic host and a bacterial symbiont.

Endosymbiotic systems were initially studied by both optical and electron microscopy techniques to observe the presence of bacteria within cells of a eukaryotic host, their structure and if they were actively replicating. These methods allow to visually localize the bacteria inside the cell of the host but gives no information about the nature of the endosymbiotic relationship. Endosymbionts are usually unculturable, and it is therefore challenging to retrieve enough DNA for sequencing in a pure form. They also often lack sequenced relatives and, when available, they are often quite divergent, limiting their utility as reference genomes. Also, the amplification of 16S can be difficult as the primer may amplify the mtDNA of the hosts as well, or other associated bacteria, confounding the signal. The works of Pace, who developed the technique to sequence 16S genes of mixed samples, make possible to obtain the 16S of the endosymbiont and surpass this limitation (Pace et al., 1986). However, to demonstrate that a bacterium is actually an endosymbiont it is not sufficient to show that the sequence belongs to an intracellular bacterium. The amplified 16S sequence can be used to develop probes for Fluorescence *In Situ* Hybridization (FISH). The probe will anneal only to the endosymbiont, marking it univocally. Using this technique is possible to verify the presence of the bacteria under analysis and their localization. However, this approach gives no information about the nature of the relationship among host and endosymbiont.

The turning point in endosymbiotic studies was the possibility to sequence the full genome of an organism from tiny DNA quantities. By analysing the complete genomic sequences of an endosymbiont, we can perform functional analysis to obtain more information about the nature of the symbiosis. To sequence the endosymbiont genome some authors adopt molecular procedures to physically separate it from the host (Ishida et al., 2014; Matsuo et al., 2010). However, this method is hardly widely applicable. A more general approach is performing the sequencing using NGS of the whole endosymbiotic system such that the output contains both genomic information of the host and of the symbiont. The sequences of the two genomes are separated using bioinformatics tools, to obtain the taxonomic affiliations and to perform functional and comparative analysis. The few available methods

performing this separation, also called deconvolution or binning, generally exploit a composition analysis of the sequences.

GC is the simplest compositional measure of a DNA sequence, and it is known that it has some phylogenetic inertia, i.e. it is maintained in evolutionary time such that related organisms tend to be similar in GC content. However, as it is a very simplistic measure, phylogenetically distant species can have similar GC content. Karlin and colleagues in the 1990s studied the composition of DNA sequences by considering the frequencies of short polynucleotide of variable length  $k$ , called  $k$ -mers. They observed that these bring a stronger phylogenetic signal than the GC content, that likely derives from different codon usage, differences in the specificity of the restriction endonucleases, and from slightly different mechanisms of DNA modification, replication and repair (Gentles, 2001; Karlin, 1998; Karlin et al., 1994, 1998; Karlin and Burge, 1995; Karlin and Ladunga, 1994; Teeling et al., 2004). The resolution power of this compositional analysis improves with the length of the  $k$ -mers, up to a limit that depends on the length of the sequences where the  $k$ -mers are counted. Intuitively, the longer the  $k$ -mer, the longer the DNA sequence has to be to minimize the noise in the counts.

Another useful variable to discern the symbiont's and the host's DNAs is the sequencing coverage, that is how many reads in total are mapping on a certain sequence normalized by its total length. This should be able to provide a partial separation of the DNA from the different sources because we can expect a different multiplicity of the two genomes. In general, the host has a larger genome than the endosymbiont, but at the same time the symbiont may have multiple genome copies per cell or multiple cells per host cell. The information of the coverage however may not be sufficient for host/endosymbiont DNA separation. It is moreover possible that non-target sequences may be present in the sample and thus sequenced. In example, protozoan usually fed on bacteria and, even if it is starved and washed prior to whole genome DNA extraction, the host may retain some DNA molecules of the eaten cells, which are then sequenced, 'wasting' reads and influencing coverage values of the whole sample. The more these signals are present, the stronger is the influence of this factor over the sequencing and over the coverage of the host and endosymbiont sequences. Also, endosymbiotic systems that involve more than a bacterial endosymbiont are not uncommon (Brown et al., 2016, 2018; Campbell et al., 2015; Gruber-Vodicka et al., 2019; Husnik and McCutcheon, 2016; Seah and Gruber-Vodicka, 2015). The number of cells of each bacterial taxon per host may be lower compared to a system where there is only one endosymbiont. In parallel, the coverage might result reduced as if there are more genomes the sequencing depth is divided onto these. Lastly, it has to be considered that NGS technologies are based on amplification, and different sequences may be amplified not with the same efficiency leading to unpredictable unequal coverage.

## Published deconvolution tools

The only tool developed for binning sequences coming from endosymbiotic Whole Genome Sequencing (WGS) data is Blobology (Kumar et al., 2013), which exploits the GC content and the coverage to separate contigs coming from the host from contigs of the symbiont. The method starts with the alignment of the contigs against an appropriate database to associate taxonomical categories to the contigs. It should be stressed that very often this step results in a very partial taxonomy coverage of the starting contigs, and this is the reason making tools like the Blobology so important in this context. We will call (taxonomy) labelled contigs those for which taxonomy information is available after this comparison step. The GC content and the coverage are used by Blobology as a bidimensional coordinate system for contig positioning, and the taxonomy from the previous step is used as a colour scheme. At this point the user leverages the positions of the labelled contigs to define a region that mostly contains sequences from the symbiont. As the partition is far from perfect, a post-processing of the results is necessary to reduce false positives and negatives, usually by performing additional comparisons with sequence databases, also at the protein level. This step is however time consuming and the whole procedure highly subjective, as it requires the user to take important decisions on the basis of relatively few labelled contigs. As a consequence, there are no easy ways to assess how changing the region for selection affects the performances of the classification. When Blobology's features do not allow to efficiently separate sequences from the host and the symbiont, most authors select contigs of interest based upon taxonomic affiliations obtained through comparison with public databases or databases specifically built with genomes of organisms related to those in the sample, or they exploit additional features, but this makes the procedure even more subjective (Brown et al., 2016, 2018; Kostygov et al., 2016; Small et al., 2016). Additionally, reads are sometimes mapped back to the assembled sequences and only those mapping on target contigs are re-assembled; this is repeated iteratively until no more contigs are added or no more sequences are elongated (Chung et al. 2017). In other works, Blobology is performed on contigs selected on the basis of the expected taxonomic affiliations (Wang and Chandler, 2016), or sequences are manually inspected to try to locate overlapping sequences and obtain a circular bacterial chromosome (Kostygov et al., 2016). As in classical genome sequencing efforts, the use of different sequencing technologies has been exploited to improve the genome reconstructions, but this clearly requires a larger budget, and DNA in higher amounts and with higher quality (Campbell et al., 2015; Floriano et al., 2018; Husnik and McCutcheon, 2016; Nikoh et al., 2018).

The problem of identifying DNA sequences of a symbiont in a sample that also contains host DNA bears strong similarities to the taxonomical binning used in shotgun metagenomics. However, the much higher complexity of DNA mixtures characteristic of metagenomic samples with respect to symbiotic systems makes the algorithmic requirements slightly different in the two cases. In the specific case of endosymbionts,

additional assumptions can be done to improve the performances of the classification. Therefore, while existing metagenomic solutions could be appropriate with endosymbiotic systems, specific tools could benefit by leveraging biological knowledge on the system. Metagenomic tools to address the binning of metagenomic sequences can be (i) reference based (Gregor et al., 2016; Seah and Gruber-Vodicka, 2015) or (ii) reference-free (Kang et al., 2019; Teeling et al., 2004; Wu et al., 2014, 2016). Tools belonging to the former category obtain taxonomic annotation of contigs through homology searches, and the performances are consequently strongly dependent on the presence of related genome sequences in the reference database. As it has been recently demonstrated, this is rarely the case in metagenomic samples, especially for less studied environments (Pasolli et al., 2019). Reference-free methods often exploits the differences in composition of the genomes in the community and are therefore less dependent on existing sequences in the databases. Briefly, DNA k-mers of a predefined length are counted in the contigs under analysis, and a clustering/classification algorithm is run to group together the compositionally similar sequences. The classification is in this case completely *unsupervised*, as it makes no use of available taxonomical information. This allows the identification of Operational Taxonomic Units (OTUs) with no counterpart in public repositories, and consequently the resulting OTUs can only partially be mapped to existing taxonomic groups, depending on the availability of similar DNA sequences.

Currently, a large number of reference free methods are available, as they are easily scalable to the size of metagenomic datasets; here we briefly describe some recent implementations that we used to put in scale the performance of our tool. MetaBAT2 (Kang et al., 2019) uses tetranucleotide frequencies and coverage to calculate a distance matrix among the contigs and to group them using a K-medoid clustering approach. The calculation of the distances can be unreliable with short sequences (the authors suggest to avoiding using contigs shorter than 2 kbp), which may result in a strong reduction of the number of contigs in highly fragmented assemblies. MaxBIN (Wu et al., 2016) implements an expectation maximization algorithm where tetramer frequencies and coverage are used separately to calculate the probability that two contigs come from the same genome; the probabilities are then combined, up to convergence of the parameters. The software provides additional information about the identified *bins*, like inferred genome size, GC content, completeness and coverage. BusyBee Web (Laczny et al., 2017) bins sequences in metagenomic samples by using a hybrid classification approach: calculation of the k-mer frequencies (either 4 or 5 bp long) is followed by unsupervised binning on a subset of the data using DBSCAN (Ester et al., 1996); at the end a Random Forest (Cutler et al., 2007) model trained on the labels from the unsupervised step is used to predict the unused part of the data. BusyBee can moreover integrate the identified bins with Prokaryotic taxonomic information or user-provided custom affiliations.

Here we present SeqDex, a tool written in R that combines partial taxonomic affiliations, obtained through combined homology searches from different databases, with

composition analysis to predict the taxonomic affiliations of all the contigs present in an assembly produced from the sequencing of mixed samples involving a host and its endosymbiont(s). SeqDex is innovative as it additionally implements a graph-based strategy to transfer taxonomical labels and because we provide a full characterization of the performances in a case-by-case way, helping the user to understand how effective the classification is. We provide both a comparison with similar methods, and several performance measures to summarize and rank the different tools.

# Methods

## The SeqDex workflow

SeqDex couples both Unix based programs and custom R script developed using the following R packages: Seqinr (Charif and Lobry, 2007), Taxonomizr ([cran.r-project.org/web/packages/taxonomizr/index.html](http://cran.r-project.org/web/packages/taxonomizr/index.html)), randomForest (Liaw and Wiener, 2002), e1071 ([cran.r-project.org/package=e1071](http://cran.r-project.org/package=e1071)), uwot ([cran.r-project.org/package=uwot](http://cran.r-project.org/package=uwot)), DBscan (Ester et al., 1996), igraph (Csardi and Nepusz, 2006). SeqDex bash script is available at Appendix 5.

All the step of the SeqDex workflow are described below and, extensively, in the User Manual in Appendix 4.

### *1. Coverage calculation*

SeqDex calculates sequencing depth using the BEDtools coverage (Quinlan and Hall, 2010); fragment counts per contig are then normalized by contig length. SeqDex considers the case of using FLASH or similar software to merge overlapping reads. Such software produces single-end reads when a pair was overlapped otherwise both reads in a pair are kept. When this happens, SeqDex considers the paired- and the single-end reads separately to provide a correct estimation of the coverage expressed in number of sequenced DNA fragments divided by the length of the contig. Moreover, when the mates map on different contigs, they contribute half a count.

### *2. k-mer frequency calculation*

GC content and k-mer frequencies are calculated by SeqDex with the Seqinr package on both strands. The counts for complementary k-mers are combined such that all k-mers, comprising the palindromes, get the same weight. As count matrices for long k-mers have a high dimensionality, the computational time needed for the analysis increases, and this step provides a reduction of the dimensions with no loss of information. SeqDex considers

contigs longer than a user-defined threshold as short sequences can diverge from the genome composition reducing the prediction capability of the model.

SeqDex by default calculates frequencies for 3-mers on contigs longer than 1 kbp.

### *3. Taxonomy affiliations*

SeqDex assigns taxonomy affiliations to contigs on the basis of homologies: contigs are compared to a nucleotide database by using BLAST+ (Camacho et al., 2009), after filtering at a defined percentage of identity and length of the HSP (High Scoring Pair). Thresholds can be changed by the user by editing the SeqDex bash script (see below for default values). The association of taxonomical codes to contigs is obtained by using Taxonomizr. As contigs potentially have multiple homologs in the database, SeqDex calculates the proportion (TaxonDensity) of alignments pointing towards a certain taxonomic category over the whole set of homologies for each contig. All contigs with a TaxonDensity value below the defined threshold, have their taxonomy label removed; in this way SeqDex reduces the risk of wrong or inconsistent taxonomical assignments. The database can be defined by the user, and it is used to assign a taxonomic origin to contigs in the input. The only limitation is that sequence titles must conform to NCBI format.

One of the problems working with symbiont sequences is the usually low identity level they share with sequences present in databases, therefore we endowed SeqDex with the ability of merging the taxonomic information retrieved for both nucleotide and proteins. Protein evolutionary rates are much slower than those characterizing nucleotide sequences, meaning that in some case it might be possible to find homologs by using the protein but not the corresponding gene. Protein coding genes are predicted with Prodigal (Hyatt et al., 2010) (default options, except for procedure set to 'meta'), and protein sequences are compared to a protein reference database by using Diamond (Buchfink et al., 2015) (default options). As before, the reference database may be the NCBI nr or a custom database with titles in NCBI format.

We stress the fact that considering the taxonomy coming from protein comparisons can be particularly advantageous when the symbiont is from taxa that are under-represented in the public repositories. However, adding the protein derived taxonomy affiliations may increase calculation time.

SeqDex exploits the presence of 16S genes within the assembly to identify the contigs of the target organisms in a final step of the workflow. rRNA genes in the assembly are identified by using Barrnap (<https://github.com/tseemann/barrnap>) then the contigs carrying the 16S genes are compared to RDP 16S database by using BLAST to add a taxonomic label to the 16S gene.

All homologies detected by BLAST and Diamond are taken into account, not only the best, allowing to compare multiple significant taxonomy affiliations for the same contig, which may highlight incongruencies.

If both nucleotide and protein homologies are used, SeqDex merges the affiliations and assigns a unique label to contigs.

By default, SeqDex considers contigs longer than 1kbp, nucleotide (protein) HSP length over 200 bp (70 aa), identity percentage over 70% (80%), and a final TaxonDensity larger than 0.75 (implying that 75% of the HSPs in the BLAST provided the same taxonomic information).

### *Extending taxonomy information*

Symbionts have often evolved for long time in an environment endowed with a very peculiar fitness landscape; this translates in their usually divergent genomic properties (extremely reduced gene content, AT richness...). Furthermore, we are only scratching at the surface of the diversity of existing symbionts and very often the study of novel symbionts leads to the discovery of novel genera or even families (Castelli et al., 2019). For the above reasons, the taxonomy assignment based on blast generally leads to relatively few contigs being labelled. This affects the parameterization of the machine learning models in a negative way, as it reduces the number of labelled cases on which the models are trained. One way to cope with this in SeqDex is by including protein comparisons. Nonetheless, the taxonomy coverage of contigs from symbiotic communities is often low (i.e. only a small fraction of contigs has significant similarities to sequences in the database such that we can assign a taxonomic label to the contig).

To further improve the taxonomy coverage of a sample, SeqDex exploits the information related to the paired end reads mapped back on the assembly. Basically, SeqDex builds a graph where two vertices (contigs) “a” and “b” are connected if there is at least one pair of reads for which one mate maps on contig “a” and the other on contig “b”. This graph is related to the graph used by assemblers for scaffolding. Edges are weighted by the number of read pairs in support, therefore they can be filtered to only keep the highly supported ones (option EDGES, default = 10). If we assume that the genomes present in the sequenced pool are different enough, as we expect in host-symbiont cases, then the connected components (CC) in this graph mostly comprise vertices corresponding to non-overlapping regions from the same genome. Therefore, the taxonomy label associated to one vertex can in principle be transferred to vertices of the same CC. Reads can however randomly map on genomes from phylogenetically distant genomes and filtering edges on their weight provides a way to remove most of the chimeric associations. Additionally, the user can control the maximum degree of vertices, as highly connected ones are more prone to be responsible for the connections involving contigs from different sources (VERTICES, default =5). As a partial error control strategy, SeqDex checks for discordant taxonomical signals within each CC on the basis of the homology-defined taxonomy, and it only applies the transfer when most labelled contigs in the CC provide the same information (can be controlled by setting MIXEDCOMP, default = 0.2). Alternatively, the user can choose to transfer the labels up to a certain distance from labelled vertices

(VERTEXDIST, default=all), which may be a compromise between the risk of wrongly propagating taxonomic information (likely reduction in precision when extending taxonomic information far away from labelled vertices), and the opposite risk of strongly reducing the recovered information (likely reduction in recall when not transferring labels to short contigs, especially in the case of highly fragmented assemblies). In this case, transfer proceeds up to the defined distance from the labelled node but if node n has a different taxonomy label, the transfer is performed up to node n-2. The increment of performances related to this approach is shown in Appendix Table 7.

#### *4. Predicting taxonomic affiliations*

The above steps prepare the input for the classification tasks performed by SeqDex and is represented by the matrix containing k-mer frequencies for each contig and the corresponding taxonomic affiliations. In a standard run, SeqDex performs a classification of the contigs at the level of Superkingdom. In this way, host and symbiont/contaminants contigs are separated. Model training is performed with RandomForest and Support Vector Machine (for further detail see Appendix 2). In both cases, SeqDex performs model training 100 times on 66% of the contigs fulfilling all the thresholds and calculates the performances of the classification on the remaining 33%. All models are kept in memory and are used for performing the classification of the contigs without a taxonomy label. Different models can assign different labels to the same contig, therefore after 100 predictions, SeqDex returns the percentage of times each contig was included in a certain taxonomy category, and the final label corresponds to the category with the highest percentage.

#### *Extending the predictions*

In the next step, the graph obtained by exploiting the pairing information is used again, to improve the predictions through a transfer strategy and consistence check similar to the one used for extending taxonomy labels. The feasibility of such a transfer is decided on a case-by-case basis by following the same rules defined in section “Extending taxonomy information” with the difference that, when it is not possible to extend the predictions, the taxonomy labels predicted for contigs are discarded (and marked as ‘misclassified’) instead of being kept.

#### *5. Unsupervised clustering*

In a hypothetical condition, only host and endosymbiont genome sequences will be present in the dataset, so the classification step at Superkingdom level will allow to retrieve the Bacterial contigs. However, this is rarely the case. Usually, contaminants are also present, but we noticed that SVM and RF are not able to provide satisfactory performances in these cases (data not shown). For this reason, SeqDex performs a final step to cluster sequences

in groups of similar composition and thanks to the identification of the cluster containing the target 16S gene, is able to recover contigs deriving from the target genome. By enabling this optional analysis (CLUSTERING, default = yes), SeqDex will (i) take the output of the classification step at the lowest selected taxonomic rank, (ii) apply a UMAP transformation to the data (R package uwot, NCOMP, default = 2) (iii) cluster the new variables with DBscan (Ester et al., 1996) and (iv) identify the cluster comprising the contig carrying the target 16S gene, (v) flag all the contigs in the same cluster as belonging to the target genome (for further information, see Appendix 2 and 4). This is the final taxonomy prediction made by SeqDex and the results are based on this set of contigs.

If several 16S genes with the taxonomic affiliation of interest are present, SeqDex will use the one with the highest coverage.

As discussed before, coverage and/or GC content may be also highly informative, depending on the specific symbiotic system. In these cases, the user can decide to perform the clustering by adding the coverage and/or the GC content to the data matrix storing the UMAP coordinates (TYPE, can be k-mers, gc, cov and combinations thereof e.g. TYPE=k-mers, gc adds the GC content as a variable in the clustering together with k-mers; default = k-mers).

### *Extending the final taxonomy prediction*

As done at the end of the machine learning classification, the clustering can be improved by using the same transfer strategy based on the read pairs-based graph. This is because methods based on k-mers are meaningful only when performed on contigs above a certain length (which depends on the selected k). Sometimes a consistent proportion of the assembly is excluded for this reason, with information loss. However, since short contigs are present in the graph built using the pairing information, SeqDex transfers the clustering belonging within a CC as done in “Extending the prediction”.

## *6. Standard SeqDex Output*

In standard usage, SeqDex produces the following output files:

Taxonomy folder: several files for the homology searches and taxonomy affiliations.

Coverage folder: k-mer frequencies, GC content, coverage;

SVMoutput and RFoutput folders: input and output files for the machine learning step.

ClusteringOutputSVM and ClusteringOutputRF folders: several files related to the DBscan clustering output. More specifically, this folder also contains the file with the name of the contigs in the target clusters and the fasta file with the sequences of the target contigs.

See Appendix 4 for further details.

## Data

To develop and test SeqDex we used three datasets: (i) a simulated dataset composed by *Saccharomyces cerevisiae* and *Neisseria gonorrhoeae*; (ii) a published dataset of an endosymbiont sequenced together with the host (*Ca. Fokinia solitaria*) (Floriano et al., 2018) that was extensively curated and re-sequenced to complete and close the genome; (iii) a dataset of a nematode (*Pratylenchus penetrans*) sequenced with his two endosymbionts (*Wolbachia pipientis* and *Ca. Cardinium hertigii*) (Brown et al., 2016, 2018); in this case the genomes of the two symbionts were only partially assembled by the authors.

### 1. Simulated dataset

We simulated paired-end reads from the *Saccharomyces cerevisiae* and *Neisseria gonorrhoeae* genomes by using the wgsim package (<https://github.com/lh3/wgsim>) changing the following parameters: number of read pairs set to one million; read length: 100 bp; fraction of indels set to 0.01; probability that an indel is extended equal to 0.05; outer distance between the two ends of 2000 bp. Paired reads for both genomes were randomly sampled and merged to obtain a dataset composed by *Saccharomyces cerevisiae* and *Neisseria gonorrhoeae* in a 9:1 proportion. We assembled the reads using SPADes (Bankevich et al., 2012), with k-mer length ranging from 31 to 91. We selected the best assembly using QUAST (Gurevich et al., 2013) based upon the N50 statistic. SeqDex were run using only nucleotide comparisons against a custom database composed by the two genomes present in the dataset, classification at Superkingdom level with k-mers of length 3 and both machine learning algorithms, and the clustering was disabled.

### 2. Real world dataset – *Ca. Fokinia solitaria*, endosymbiont of a ciliate

*Ca. Fokinia solitaria* and its host were sequenced using Illumina HiSeq 2500 to generate 14'783'394 150 bp paired-end reads, as reported by the authors (Floriano et al., 2018). We assembled the reads obtained by the authors using SPADes, with k-mer length ranging from 31 to 91, and then chose the best assembly using QUAST, based on the N50 statistics. In SeqDex taxonomies were assigned using BLAST against the NCBI nt database (downloaded in October 2018), with default options but excluding *Ca. Fokinia solitaria* genome; the 16S rRNA genes were compared to RDP 16S database downloaded in October 2018. SeqDex was run with default parameters using Alphaproteobacteria as target class for the final clustering.

### 3. Real world dataset – *Wolbachia-Cardinium* dual endosymbionts of a nematode

*P. penetrans* and its endosymbionts were sequenced using Illumina MiSeq to generate 301 bp long paired end reads (accession: SRR3097580) for a total of 10'563'810 pairs (Brown et al., 2016). Reads were quality checked with fastQC (Andrews, 2010), adapters were removed using Trimmomatic (Bolger et al., 2014) and overlapping pairs were merged using FLASH. Assembly was performed with SPADes, using default parameters and k-mer length of 21, 33, 55, 77, 99. The best assembly was chosen using QUASt. SeqDex was run on both endosymbionts using both nt and nr NCBI databases, the same RDP 16S databases of *Ca. Fokinia solitaria*, two classification iterations on Superkingdom and Class taxonomic levels and the final clustering searching as target “unclassified\_Bacteroidetes” for *Ca. Cardinium hertigii* and “Alphaproteobacteria” for *Wolbachia*.

## Performance calculation

Regarding the study cases shown in the paper, we calculated additional statistics to highlight the behaviour of SeqDex that exploit the availability of genome sequences of the symbionts as from previous publications (Brown et al., 2016, 2018; Floriano et al., 2018). In these cases, we perform comparisons of the performances of the different methods based on counting True Positives and Negatives (TP, TN respectively), False Positives and Negatives (FP, FN respectively), that are used to calculate sensitivity, accuracy, precision and F1 scores (Appendix Table 1, for further information see Appendix 1). We stress that these statistics can be calculated here because the true labels can be derived for all contigs thanks to the availability of the symbiont genomes. The performances of tools performing the taxonomical classification of contigs are generally based on numbers of correct/wrong classifications; however, the many contigs obtained from short reads have very heterogeneous lengths such that weighting the error made in the classification with respect to the length of the sequences should provide a much better characterization of the true capability of a tool. For instance, a tool mis-classifying a very short contig performs better than one mis-classifying a very long one and yet both have the same performances if we refer to raw contig counts. For this reason, we also compare the contigs assigned to each taxonomical category with the source genome, and we calculate performances based on the total number of nucleotides that were correctly assigned, with respect to the genome length. To evaluate this, we used QUASt which provides a comparison among an assembly and a reference genome.

## Details about third party tools parameters

We compared SeqDex with the methods Blobology (Kumar et al., 2013), MetaBAT2 (Kang et al., 2019), BusyBee (Laczny et al., 2017) and MaxBin (Wu et al., 2014). If not otherwise specified, only contigs longer than 1000 bp were considered, with GC content and coverage values calculated as described before.

MetaBAT was run on contigs by changing minimum contigs length (-m, set to >1500 bp), percentage of 'good' contigs (--maxP, set to 90), minimum score of an edge for binning (--minS, set to 80) and minimum size of a bin as the output (-s, set to 150000). MaxBIN was performed using default parameters, and only k-mer length was changed in BusyBee Web (k=4).

# Results

## SeqDex pipeline

SeqDex is written mainly in R but can be run from a bash script where the user can change most parameters (Appendix Table 2 lists all scripts that are part of SeqDex, and that are available for download at [github.com/ComparativeSystemsBiologyGroup/SeqDex](https://github.com/ComparativeSystemsBiologyGroup/SeqDex); see Appendix 4 for the manual and Appendix 5 for the SeqDex bash script).

The workflow is shown in Figure 9 and each step is described in Methods:

- (1) Coverage calculation. We indicate the whole set of contigs in the assembly as A;
- (2) Identification of 16S rRNA genes to identify target bacteria (the identity of the symbiont in these situations is often achieved through PCR amplification and sequencing);
- (3) Comparison with sequence databases to associate contigs to taxonomic affiliations for a subset T (with  $T \subseteq A$ ) of the contigs; this can additionally be performed at the protein level;
- (4) Taxonomy extension using the paired read graph of the assembly;
- (5) k-mer frequencies are calculated for all contigs in C ( $C \subseteq A$  such that contigs in C are longer than a defined threshold).
- (6) Random Forest (RF) (Cutler et al., 2007) and Support Vector Machine (SVM) (Cortes and Vapnik, 1995) models are trained on data for contigs in  $T \cap C$  by exploiting k-mer frequencies and the partial taxonomical affiliations obtained in 3 and 4. Then, the trained models can be used to predict taxonomical affiliations of the contigs with no taxonomical label. At this point, all contigs in C have a taxonomical affiliation, coming from step 3 and 4 or predicted here. The problem is split into nested classifications by considering different taxonomical depths: a first classification separates Prokaryotic from Eukaryotic sequences; contigs included in the former can then be used for more stringent classifications by applying the model at a stricter level of taxonomical

categories. At the end of this step SeqDex provides a taxonomy for the contigs. This can be used as is or it can be processed in the following optional step.

- (7) Useful when the user knows there may be more species in the assembly (e.g. bacterial contaminants, in addition to the target organism(s)). The machine learning approach has unsatisfying performances at this level (data not shown) and therefore was replaced by the following strategy. First, the k-mer matrix undergoes dimension reduction using the UMAP transformation (McInnes et al., 2018); then a DBscan clustering (Ester et al., 1996) provides a partition of the contigs into clusters. The cluster containing the 16S rDNA gene with the right taxonomical affiliations is defined as the target cluster. Then, paired end reads mate graph is used to control the clustering and also to extend it to contigs shorter than the defined threshold (A – C), so that the final target cluster contain also contigs that were excluded in (5). The contigs falling in the target cluster are now retrieved.

The entire SeqDex procedure can be run using default parameters but the scripts are customizable, as the user is able to change several key parameters.

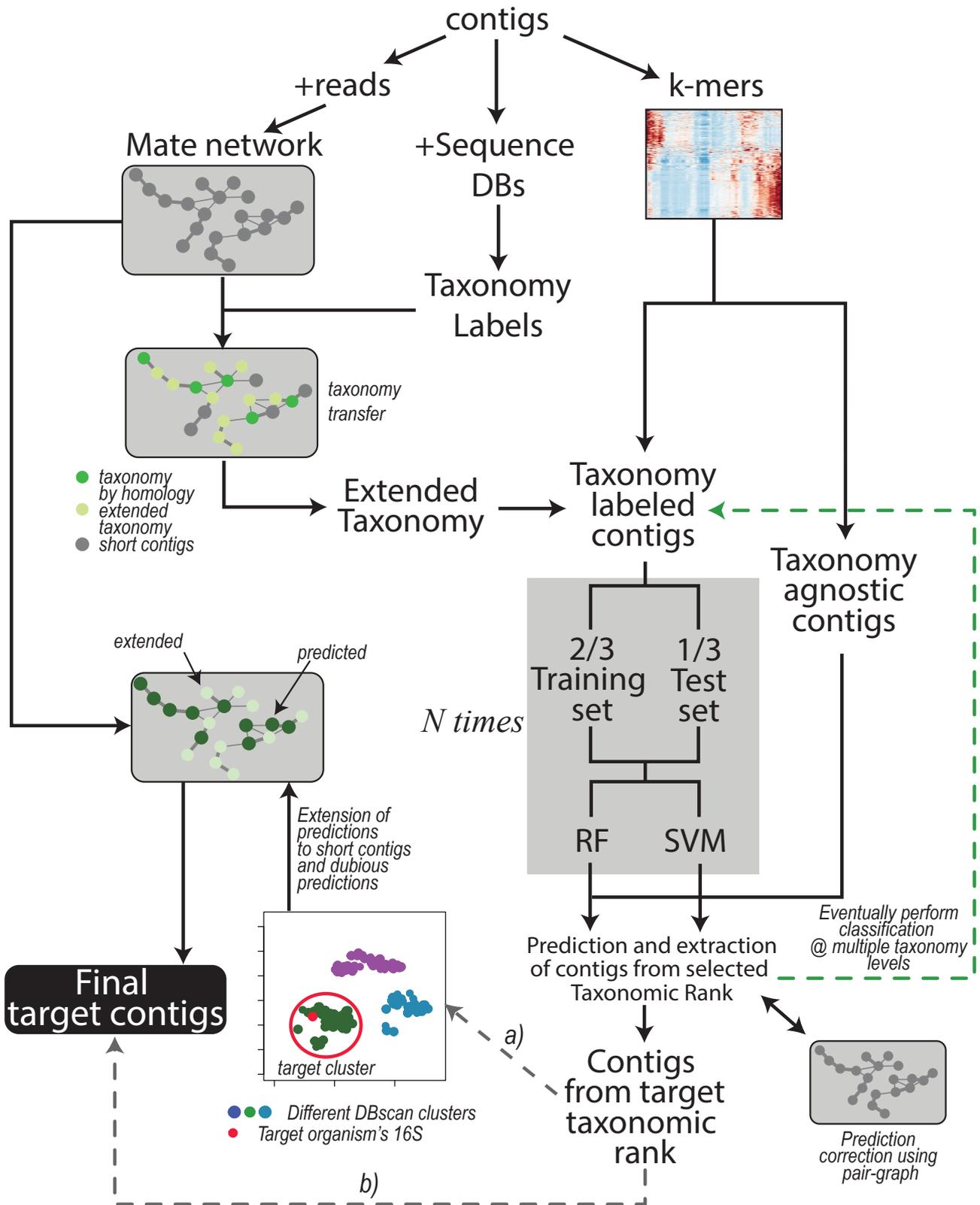


Figure 9 - Contigs are used to obtain the read-pair graph by exploiting the paired sequencing (left branch). The network is used in several steps of the procedure, for instance, to extend the taxonomy information obtained through sequence comparison (middle branch). The k-mer frequencies are also calculated (right branch) and combined with the (extended) taxonomy. The contig dataset is then split in two depending on the presence of taxonomy labels; the labelled contigs are used to train the machine learning models (grey box) after partitioning the contigs again into a training and a test set. Training of the models is repeated N times to provide error estimations that are independent of the actual contigs in the train and test sets. As a default, classification is performed at the only Superkingdom level; if the user wants to proceed down in the

taxonomy hierarchy, additional iterations, each time focusing on a different taxonomic rank (green branch) can be performed. After that, SeqDex uses the trained models to predict the taxonomic affiliations of unlabelled contigs. Again, the read-pair network can be used to correct the predictions made by the machine learning models. At this point, contigs can be recovered, and two possible alternatives exist: a) when there is more than one bacterium in the sequencing, the user can proceed by directing SeqDex on the flow indicated with (a): (i) run UMAP, (ii) DBscan on the UMAP transformed k-mer frequencies, (iii) identify the cluster containing the target 16S gene, (iv) extend predicted taxonomy information using the read-pair graph and (v) extract the contigs identified as coming from the target organism. Alternatively, (b) SeqDex can directly extract the contigs classified as coming from the target organism after the machine learning step.

## Case studies

### *1. Simulated Saccharomyces cerevisiae-Neisseria gonorrhoeae dataset*

We use this simulated dataset as an example of a very simple case with a eukaryote (as the host) and a bacterium (playing the role of the symbiont). Such simple situations are very rare in real-world cases, where usually more prokaryotes can be found, most of which are usually not the symbionts. Therefore, in this case we proceeded by only classifying at the Superkingdom level, which could be done when preliminary analyses (e.g. PCR amplification) show the presence of only one bacterial 16S rDNA.

To assess the performance of the Blobology approach, we selected the contigs included in the region defined by GC content  $\geq 0.3$  and coverage  $\leq 0.05$  fragments/nt on the basis of an enrichment of contigs with the target taxonomic affiliation in that region.

BusyBee crashed reporting an error after identifying one only cluster in the dataset. MetaBAT completed the analysis but still found only one bin. The results of these two tools are therefore not shown for this study case.

MaxBIN correctly identified two bins, one mainly composed by contigs with taxonomic affiliation Bacteria, which was selected as target.

Finally, we performed SeqDex with RF and SVM focusing on the Superkingdom level. Taxonomic affiliations were obtained by comparison to a database composed only by the genomes of the two organisms used for this simulation, thus this homology search was enough to classify all contigs. To use SeqDex we randomly discarded 33% of these affiliations.

The contigs retrieved after each method were used to calculate the fraction of the genome of *Neisseria* and to calculate sensitivity, precision, accuracy and F1 scores as described before (Figure 10a and Appendix Figure 4, Table 3 and Appendix Table 3).

SeqDex outperforms Blobology in F1 score and sensitivity, both considering the total length and the number of contigs, but the latter has higher precision. This could be explained considering that these two organisms have different GC content and that the simulated sequencing produced widely different average coverages for the two genomes. Considering MaxBIN, SeqDex give similar sensitivity and accuracy but higher precision and F1 score.

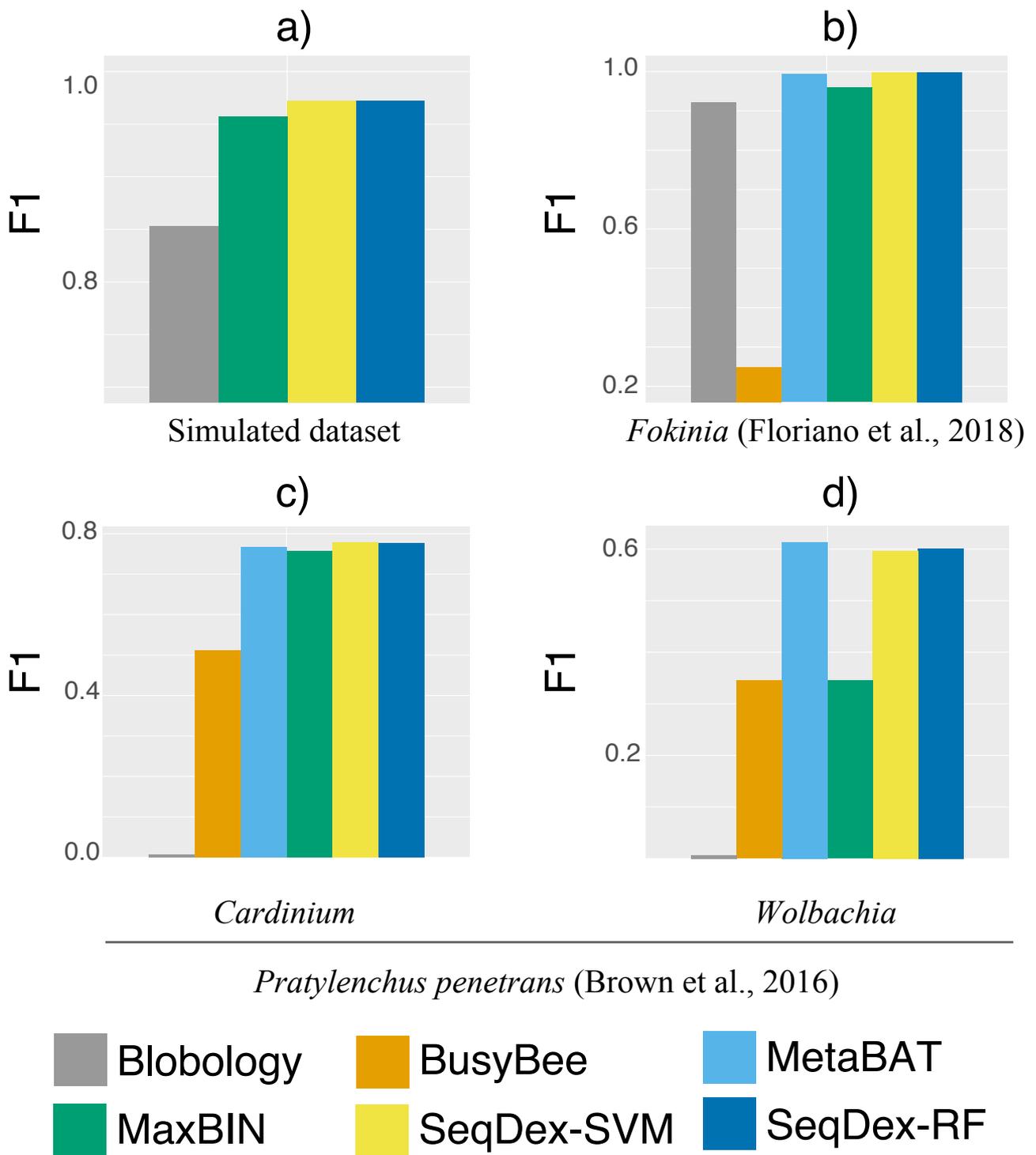


Figure 10 - Genome-based F1 scores. For all datasets and targets considered in the work. a) Simulated dataset; b) *Ca. Fokinia solitaria* dataset; c) and d) *Pratylenchus penetrans* dataset: c) *Ca. Cardinium hertigii*; c) *Wolbachia pipientis*.

Table 3 - Performances calculated with respect to the whole genome sequence of *Neisseria gonorrhoeae*.

	Blobology	MaxBin	SeqDex-SVM	SeqDex-RF
Sensitivity	0.7442	0.9634	0.9586	0.9586
Precision	1.0000	0.9518	0.9862	0.9862
Accuracy	0.9909	0.9969	0.9980	0.9980
F1 score	0.8533	0.9576	0.9722	0.9722

## 2. Real world *Ca. Fokinia solitaria* dataset

We applied the Blobology approach by using GC and coverage boundaries comparable to those used by the authors in the original publication (>30% GC, coverage between 0.3 and 8 fragments per nucleotide).

MetaBAT identified 10 bins, MaxBIN 5 bins and BusyBee Web 8 bins.

Our model exploited the presence of five complete 16S genes, belonging to Class *Gammaproteobacteria* (RDP code: S000653219), *Alphaproteobacteria* (two different 16S genes with RDP codes S000607898 and S004400661, the latter of which 100% identical to *Ca. Fokinia solitaria*), *Epsilonproteobacteria* (RDP code: S003597162), *Bacteroidia* (RDP code: S001493056). Of these, the contig containing the 16S gene from *Ca. Fokinia* have a coverage of 2.78 fragments per nucleotide, while the others range from 0.02 to 0.27, reflecting the presence of a much higher copy number for the endosymbiont with respect to the other bacterial species in the sample.

For this dataset, we run the entire classification pipeline implemented in SeqDex: SVM and RF are used to define the contigs coming from Eukaryotes and Bacteria; then, the k-mer frequencies of the contigs with Bacteria affiliation (predicted or deriving from the blast), undergo the UMAP transformation that produces two new variables that DBscan uses to define clusters. The whole procedure resulted in 16 and 8 clusters for SVM and RF, respectively. In Figure 10b and Table 4 we report the statistics relative to the cumulative length correctly classified by each approach, while performance statistics based on contig counts are shown in Appendix Table 4 and Appendix Figure 5. Considering cumulative length, BusyBee web performed poorly: even if it has sensitivity values that are comparable to the other methods, its precision, accuracy and F1 score are extremely low. Blobology shows performance statistics comparable to MetaBAT, MaxBIN and SeqDex. It has to be considered that this represents an uncommon situation: host and endosymbiont have different GC content and the symbiont is very abundant, at least compared to other bacteria present, as the sample to be sequenced was carefully selected in lab on the basis of the strength of the 16S signal by *Fokinia*. Among the remaining tools, they all performed good, with MaxBIN having lower precision and F1 scores and MetaBAT shows lower sensitivity

and F1 score than SeqDex. Among all, SeqDex with both machine learning algorithm shows higher sensitivity, precision, accuracy and F1 scores.

Table 4 - Performances of the classifications with respect to the whole *Ca. Fokinia solitaria* genome

	Blobology	BusyBee	MetaBAT	MaxBIN	SeqDex - SVM	SeqDex - RF
Sensitivity	0.8684	0.9966	0.9864	0.9966	0.9966	0.9966
Precision	0.9843	0.1422	1.0000	0.9246	1.0000	1.0000
Accuracy	0.9973	0.8881	0.9997	0.9984	0.9999	0.9999
F1 score	0.9227	0.2489	0.9932	0.9592	0.9983	0.9983

### 3. Real world *Wolbachia-Cardinium* dataset

This study case has additional levels of complexity because the host is multicellular, it contains at least two endosymbionts whose genomes are still incomplete. Moreover, as often for endosymbionts, the most closely related genomes from databases are not highly similar.

When plotting the contigs in Blobology space (GC Vs coverage), the sequences from the two symbionts did not form discernible clusters and also overlap with host's contigs (Figure 11); it is therefore difficult to define regions enriched in sequences coming from one or the other symbiont and with the exclusion of host's contigs. For the Blobology strategy, we tentatively defined the *Cardinium* region as defined by a GC content below 50% and by a coverage in between 0.001 and 0.3 fragments per nucleotide; the *Wolbachia* region was defined by a GC content below 40% and a coverage in between 0.01 and 0.1 fragments per nucleotide. We defined these thresholds based on the shape of the Blobology plot, by observing the location of contigs containing the symbionts 16S genes and exploiting the position of contigs aligning to the draft genomes available for the targets.

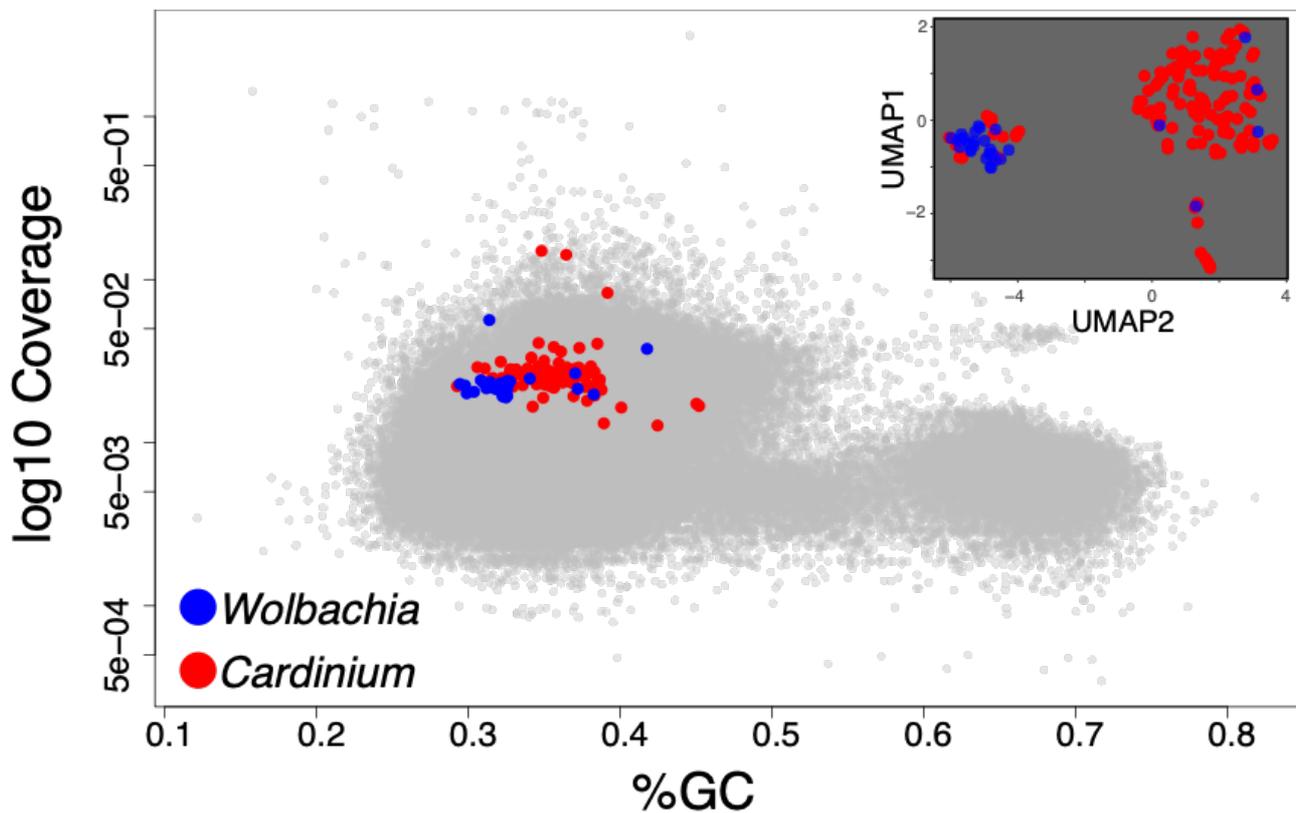


Figure 11 - In the main panel we show the Blobology plot obtained for the *Pratylenchus* dataset as an example of cases when host and symbiont(s) are not clearly discernible in the GC and coverage dimensions. In the inset we represent the contigs from the symbionts (as identified through homology) in the UMAP space used by SeqDex to partition the contigs from the symbionts.

BusyBee identified 29 bins, but the two targets belong to the same bin. MetaBAT discovers 11 bins and MaxBIN 24 and both manage to assign one bin per symbiont. Our analysis exploited the presence of four complete and three partial 16S genes, belonging to Class *Alphaproteobacteria* (3 complete genes with RDP codes: S003299234, S000830683, S001548999; the latter is identical to reference *Wolbachia pipientis* at 99.98%), Gammaproteobacteria (a complete gene with RDP code S000711119), *Betaproteobacteria* (a partial gene with RDP code S000691097), *Cytophagia* (a complete gene with RDP code S004482339 and a partial gene with RDP code S004414660, identical to *Ca. Cardinium hertigii* at 100%).

In this dataset, we run SeqDex with two nested iterations. In detail: (1) taxonomic affiliations were used to predict the Superkingdom of contigs having no homologs in the database and then we only select contigs predicted as Bacteria (predicted and derived from homology); (2) the second iteration works on these contigs to predict the Class; (3) Contigs with *Cytophagia* affiliation were selected as potentially containing *Cardinium* contigs, and the additional clustering step was performed; (4) Similarly, the contigs predicted as coming from the *Alphaproteobacteria* Class were used for the clustering step to identify the *Wolbachia* contigs. SeqDex with SVM (RF) identified 8 (5) clusters within the *Cytophagia* dataset and 17 (6) within the *Alphaproteobacteria* one. As for the *Ca. Fokinia*, we then

identified the cluster of interest by looking for the *Cardinium* and *Wolbachia* 16S genes. The performance statistics are summarized in Figure 10c, Table 5, Appendix Figure 6 and Appendix Table 5 for the *Cardinium*; Figure 10d, Table 6, Appendix Figure 7, Appendix Table 6 for the *Wolbachia*.

We concluded that the Blobology approach shows high sensitivity, comparable to other methods, but low precision, accuracy and F1 scores for both organisms. BusyBee performed similar to Blobology, except for accuracy, which is higher. Both methods basically failed to classify the two endosymbionts. Considering the total length of contigs correctly deconvolved for *Cardinium*, SeqDex, MaxBIN and MetaBAT showed comparable performance statistics, even though MetaBAT performed worse than the other two in accuracy, and SeqDex slightly better in F1 score. Instead, considering total length of *Wolbachia*, MaxBIN performed worse than others in precision and F1 scores.

Table 5 - Performances of the classifications with respect to the whole *Ca. Cardinium hertigii* genome retrieved from the *Pratylenchus penetrans* dataset.

	Blobology	BusyBee	MetaBAT	MaxBIN	SeqDex – SVM	SeqDex - RF
<b>Sensitivity</b>	1.0000	0.9781	0.8793	0.9289	0.9388	0.929
<b>Precision</b>	0.0033	0.3466	0.6789	0.6394	0.6649	0.6682
<b>Accuracy</b>	0.1423	0.9946	0.9981	0.9983	0.9985	0.9985
<b>F1 score</b>	0.0067	0.5118	0.7662	0.7575	0.7785	0.7773

Table 6 - Performances of the classifications with respect to the whole *Wolbachia pipientis* genome retrieved from the *Pratylenchus penetrans* dataset.

	Blobology	BusyBee	MetaBAT	MaxBIN	SeqDex – SVM	SeqDex - RF
<b>Sensitivity</b>	0.9974	0.9922	0.9800	0.9810	0.9868	0.9868
<b>Precision</b>	0.0034	0.2091	0.4469	0.2096	0.4273	0.4327
<b>Accuracy</b>	0.5042	0.9936	0.9974	0.9936	0.9977	0.9978
<b>F1 score</b>	0.0068	0.3454	0.6139	0.3453	0.5963	0.6016

# Discussion

The comparison of our model with Blobology, BusyBee Web, MaxBIN and MetaBAT points out the generally superior performances of our method for all tested datasets.

In the simulated dataset Blobology, MetaBAT and BusyBee Web failed to separate *Neisseria* from *Saccharomyces* while SeqDex and MaxBIN showed similar good performances. In the *Ca. Fokinia* dataset, SeqDex and MetaBAT showed similarly good performance, while Blobology and BusyBee performed worse. In the *Pratylenchus penetrans* dataset, SeqDex with both SVM and RF performed slightly better than MetaBAT and MaxBIN concerning *Cardinium*, and it outperforms MaxBIN on *Wolbachia*. By comparing the whole-genome based performances for *Ca. Fokinia solitaria* (Figure 10b, Table 4, Appendix Figure 5, Appendix Table 4) with those calculated for *Cardinium* (Figure 10c, Table 5, Appendix Figure 6, Appendix Table 5) and for *Wolbachia* (Figure 10d, Table 6, Appendix Figure 7, Appendix Table 6), we see they are lower in the latter. The published target genomes are however incomplete, and this might explain this difference, as the presence of correctly assigned contigs that are missing from the assembly would artifactually degrade the performances. Indeed, this dataset illustrates that SeqDex can also be helpful with complex datasets.

We tested our model in a variety of condition: an unrealistic simulated dataset composed by only two organisms; a real dataset sequenced with high coverage, with a strong signal from the endosymbiont, and lower for non-target bacteria; a final real dataset containing two different endosymbionts and contaminant sequences. The performance analysis pointed out that SeqDex has comparable and sometimes superior performance to the other tools, which likely reflect the slightly different purpose for which most of the other tested tools were designed. The use of the paired-reads derived graph provides a boost to the performances when taxonomy labels derived from homology searches are particularly deficient. For instance, in all tested cases the use of the graph to refine and extend taxonomies and predictions only provided a marginal improvement, except in the case of *Wolbachia* from the *Pratylenchus* dataset, for which the precision increases 10-100 times depending on the algorithm used for the classification. This shows that our procedure can be extremely helpful in particular cases.

In conclusion, SeqDex showed high reliability on all datasets, with high precision, accuracy and F1 score. Differently from other tools, it provides and returns error estimation of the classification, such that the user understands if additional refinements are necessary and more importantly if the method can be applied.

We stress that in many situations it should be better to combine the output of different tools to achieve optimal results. This can be done in a conservative way, for instance retaining only the contigs predicted as coming from the target by all applied tools, or using some sort of majority rule.

Future developments will focus on a modification of the machine learning algorithms to include sequence length-dependent weights for contigs (Freitas et al., 2007) as the machine learning algorithms that are commonly employed in these situations seek an optimization of the classification based on contig counts only (e.g. same weight is given to the wrong/right classification of a contig of 100'000 nucleotides and a contig of 2'000). To conclude, another way to improve these approaches is the integration of clade-specific gene syntenies to further refine the composition-based classification.



# SeqDex: A Sequence Deconvolution Tool for Genome Separation of Endosymbionts From Mixed Sequencing Samples

Alice Chiodi<sup>1,2\*</sup>, Francesco Comandatore<sup>3,6</sup>, Davide Sassera<sup>4</sup>, Giulio Petroni<sup>5</sup>, Claudio Bandi<sup>2,3</sup> and Matteo Brilli<sup>2,3\*</sup>

<sup>1</sup> Department of Earth and Environmental Sciences, University of Pavia, Pavia, Italy, <sup>2</sup> Department of Biosciences, University of Milan, Milan, Italy, <sup>3</sup> Pediatric Clinical Research Center "Romeo ed Enrica Invernizzi", University of Milan, Milan, Italy, <sup>4</sup> Department of Biology and Biotechnology, University of Pavia, Pavia, Italy, <sup>5</sup> Department of Biology, University of Pisa, Pisa, Italy, <sup>6</sup> Department of Biomedical and Clinical Sciences "L. Sacco", University of Milan, Milan, Italy

## OPEN ACCESS

### Edited by:

Alessio Mengoni,  
University of Florence, Italy

### Reviewed by:

Tarini Shankar Ghosh,  
University College Cork,  
Ireland

Irene Stefanini,  
University of Warwick,  
United Kingdom

Florent Lassalle,  
Imperial College London,  
United Kingdom

### \*Correspondence:

Alice Chiodi  
alice.chiodi01@universitadipavia.it  
Matteo Brilli  
matteo.brilli@unimi.it

### Specialty section:

This article was submitted to  
Bioinformatics and  
Computational Biology,  
a section of the journal  
Frontiers in Genetics

Received: 11 June 2019

Accepted: 15 August 2019

Published: 19 September 2019

### Citation:

Chiodi A, Comandatore F,  
Sassera D, Petroni G, Bandi C and  
Brilli M (2019) SeqDex: A Sequence  
Deconvolution Tool for Genome  
Separation of Endosymbionts From  
Mixed Sequencing Samples.  
*Front. Genet.* 10:853.  
doi: 10.3389/fgene.2019.00853

In recent years, the advent of NGS technology has made genome sequencing much cheaper than in the past; the high parallelization capability and the possibility to sequence more than one organism at once have opened the door to processing whole symbiotic consortia. However, this approach needs the development of specific bioinformatics tools able to analyze these data. In this work, we describe SeqDex, a tool that starts from a preliminary assembly obtained from sequencing a mixture of DNA from different organisms, to identify the contigs coming from one organism of interest. SeqDex is a fully automated machine learning-based tool exploiting partial taxonomic affiliations and compositional analysis to predict the taxonomic affiliations of contigs in an assembly. In literature, there are few methods able to deconvolve host-symbiont datasets, and most of them heavily rely on user curation and are therefore time consuming. The problem has strong similarities with metagenomic studies, where mixed samples are sequenced and the bioinformatics challenge is trying to separate contigs on the basis of their source organism; however, in symbiotic systems, additional information can be exploited to improve the output. To assess the ability of SeqDex to deconvolve host-symbiont datasets, we compared it to state-of-the-art methods for metagenomic binning and for host-symbiont deconvolution on three study cases. The results point out the good performances of the presented tool that, in addition to the ease of use and customization potential, make SeqDex a useful tool for rapid identification of endosymbiont sequences.

**Keywords:** symbiont, deconvolution, machine learning, binning, NGS

## INTRODUCTION

In recent years, we experienced a huge improvement in sequencing technologies. In particular, NGS machines have reached throughput levels and costs that make whole genome and metagenome sequencing technically easy and cheap.

In this article, we deal with a specific problem that arises when the sequencing is performed on heterogeneous DNA mixtures containing the DNA of a host and of its symbiont(s). Such "mixed samples" sequencing approach is widely used in the study of symbionts (Brown et al., 2016; Brown

---

## CHAPTER 4

Three novel bacterial endosymbionts of *Spirostomum* spp.

# Introduction

Symbiosis is a term coined in 1877 by Albert Bernhard Frank to describe the mutualistic interaction between algae and fungi in lichens. The current definition was instead formulated in 1879 by Heinrich Anton de Bary as *the living together of unlike organisms* (Douglas, 1994). In 1949, Edward Haskell proposed to classify symbioses based on the effect each organism involved has on the other(s). Basically, the interaction can be positive (+), negative (-) or neutral (0); their combination for a couple of organisms defines 6 type of interactions (Table 7): mutualism (+,+), commensalism (0,+ or +,0), predation/herbivory/parasitism (-,+ or +,-), amensalism (0,- or -,0), competition (-,-), neutralism (0,0) (Pringle, 2016). These interactions can be facultative or obligate: in the first case the two organisms can live separately; in the second case one or both of them depend on the other for survival.

Table 7 - The interaction between two organisms can be described and defined by the type of action each has on the other. The table reassumes all type of interaction and provide their current definition. + means that the action of an organism over the other is positive, 0 that is neutral and - that is negative.

		Org 2		
		+	-	0
Org 1	+	Mutualism	Predation/herbivory/parasitism	Commensalism
	-	Predation/Herbivory/Parasitism	Competition	Amensalism
	0	Commensalism	Amensalism	Neutralism

In 1967 Lynn Margulis proposed a new type of symbiotic interaction, the endosymbiosis, where an organism, usually a bacterium, lives inside the cell of a host (Sagan, 1967). In detail, she proposed that the mitochondrion originated from an endosymbiosis. This theory received little support in the beginning, but when Robert Schwartz and Margaret Dayhoff published new evidences supporting the bacterial ancestry of both mitochondria and chloroplasts (Schwartz and Dayhoff, 1978), it became clear to the scientific community that this type of symbiotic interaction had a fundamental importance in the evolution of eukaryotes, raising the interest in characterizing such systems. Studies dealing with different aspects of endosymbiosis increased, also as a consequence of the methodological improvements taking place in those years, resulting in a wide recognition of the evolutionary importance of symbioses in general and endosymbiosis in particular. Endosymbiotic relationships can be observed throughout the whole tree of life and are frequent in *Protozoans*, as in *Trypanosomatide*, *Ciliates*, and *Amoebidae*, but also in *Hexapoda*, as aphids, ants, psyllids, and cycads.

One of the first observed endosymbionts is *Wolbachia pipientis* (*Alphaproteobacteria*, order *Rickettsiales*), which was found inside the cells of the mosquito *Culex pipiens* (Hertig, 1936; Hertig and Wolbach, 1924). In the last few decades various studies have found *Wolbachia* species in numerous arthropods and nematodes (Bandi et al., 1998, 2001;

Jeyaprakash and Hoy, 2000; Taylor and Hoerauf, 1999; Wenseleers et al., 1998; Werren et al., 1995; Werren and Windsor, 2000). *Wolbachia* is usually located in the gonads of the host, behaving as a sexual parasite as it provokes male killing, feminization, parthenogenesis or, more importantly, cytoplasmic incompatibility in gametes, a situation where infected males can produce offspring only if mating with infected females (Figure 12). Although behaving as a parasite, organisms infected by *Wolbachia* usually have fitness advantages over the others, up to the extreme that some hosts are not able to reproduce without the symbiont, as it is the case with some nematode (*Brugia Malayi* and *Wuchereria bancrofti*); other hosts may benefit from the symbiosis as they develop resistance to viruses, insecticides or show higher percentage of successful emergence of adults from larvae (Berticat et al., 2002; Foster et al., 2005; Kaiser et al., 2010; Teixeira et al., 2008).

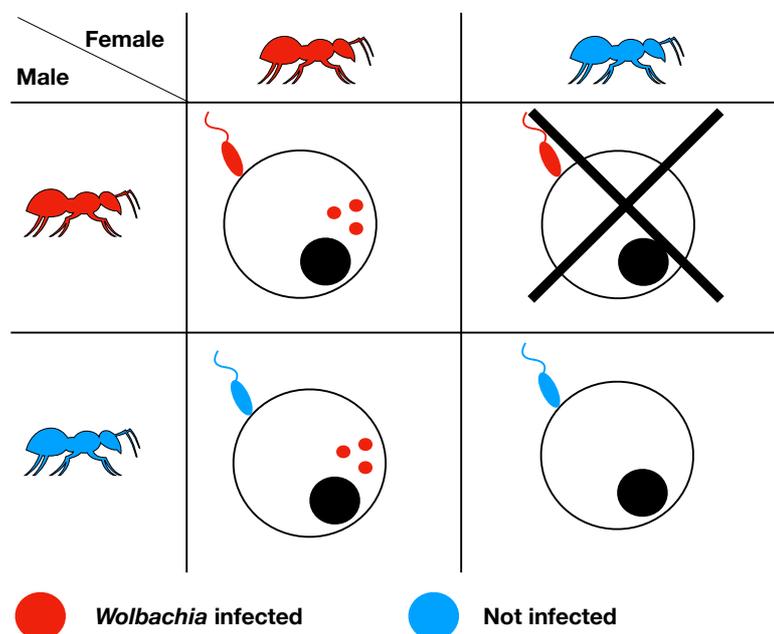


Figure 12 – The image explains cytoplasmic incompatibility phenomenon. Infected females can proficiently mate with both infected and non-infected males; a non-infected female mating with a non-infected male also produces viable progeny. On the contrary, the mating between a non-infected female and an infected male does not produce offspring. This results in the infected female being compatible with both infected and non-infected male, whereas the non-infected female can produce offspring only when mating with the non-infected male. This leads to a fitness advantage of the infected female as it ensures the transmission of *Wolbachia*, which can be passed to the offspring only by the mother.

Another well studied case is the aphid endosymbiont, *Buchnera aphidicola*. It has one of the smallest genomes known, as the long term and stable relationship with the aphid resulted in the loss of many genes (van Ham et al., 2003). The bacterium cannot live outside the host, which gain some benefits from this association: aphids fed on plant sap that has serious nutritional deficiencies, especially concerning amino acids, that are however produced and released to the host by *Buchnera* (Douglas, 1998).

The studies on insect endosymbionts revealed that some hexapods harbour two bacteria: a primary endosymbiont, which derives from an ancient association, and a secondary

endosymbiont, whose relationship to the host is more recent. A well-studied example is *Homalodisca vitripennis*, a species of cicadas that harbour a primary endosymbiont - *Sulcia muelleri* - and a secondary endosymbiont - *Baumannia cicadellinicola*. These bacteria are both obligate and are present in the cells of a specific organ called *bacteriome*. The relationship that keeps together the host and the endosymbionts is mutualistic: *Sulcia* has one of the smallest genomes known and is unable to live outside the host, but it produces and shares essential amino acids; *Baumannia* seems to be able to produce its own membrane and plasma, but misses important functions, such as the production of LPS, and thus is not able to live outside the host. Its role is to provide vitamins and cofactors (Wu et al., 2006) to the host. *S. muelleri* has been found in other sap feeding insects inhabiting the same specialized tissue.

In Protista lots of endosymbiosis are found, mainly in ciliates, probably due to their diet, as they fed on bacteria. One example is the endosymbiosis between host of the genus *Paramecium* and *Holospira* bacterial species. First observations date back to 1969 (Preer, 1969). *Holospira* species inhabit the nucleus of the host and can be found in two forms: a reproductive short form and an infective elongated form, a dimorphism that is quite common in the *Alphaproteobacteria* group. The reproductive form divides in the host nucleus during growth. Then, if the *Paramecium* starves or if protein synthesis is inhibited, the reproductive form differentiates into the infective, which divides into the nucleus and escapes the host to infect another cell. An excess of the infective *Holospira* form into the nucleus can inhibit the *Paramecium* growth and eventually kill it. Due to this behaviour, it was initially supposed that *Holospira* was a parasite; however, later studies demonstrated that infected host cells acquire the capability to resist to heat and osmotic stress after colonization. *Holospira* species are unculturable outside the host indicating that this relationship is likely obligate for the bacterium. However, the symbiont is not essential for the host (Fujishima and Kodama, 2012).

All these endosymbiotic systems examples reveal quite a complicated picture: the organisms may establish positive, negative or neutral relationships, as in symbiosis, that can moreover be facultative or obligatory. It can be hard to define the type of relationship as suggested by the *Wolbachia* and the *Holospira* examples: the symbionts may act as parasites in some conditions, but they might provide some advantage to the hosts. This highlights the difficulties to restrain such multiple faceting relationships to a unique type. The studies conducted on the genome of the endosymbionts have pointed out the unicity of obligate endosymbiotic systems. As obligatory, the relationship is vital for the bacterium, which means that the prokaryote can hardly be cultivated because is adapted to the stable and rich condition of the cytoplasm of the host cell. The endosymbiosis can be essential also for the host such that the bacterium is needed for survival and/or reproduction.

In the past few decades Siv Andersson and colleagues compared the available sequences of the endosymbionts genomes and observed that these bacteria showed similar characteristics: reduced genome size, high level of pseudogenization, lack of genes that

are essential in free-living species, i.e. the genes involved in growth, replication and survival. They supposed that the isolation in which the endosymbionts incur in hosts cells facilitated the accumulation of degenerative mutations on genes which are no more needed, as those involved in sensing the environment or in the production of molecules that are usually essentials but that are available in the intracellular milieu (Andersson and Kurland, 1998; Gil et al., 2002; McCutcheon and Moran, 2012; Moran and Wernegreen, 2000). As a result, the adaptations of the endosymbionts often take evolutionary paths that are highly divergent with the respect to their free-living counterparts. This, together with a true genetic isolation of the endosymbiont, makes so that its sequences evolve independently from its relatives and, as a consequence, taxonomic assignment by single gene comparisons (e.g. even the 16S rDNA) can become problematic.

Besides this, reaching a trustable taxonomic assignment nonetheless gives no information about the nature of the relationship that keeps the organisms together. One may for instance ask if there is some sort of metabolic complementation or dependency, because this may indicate a synergy of the system over the single entities and therefore a relationship which is beneficial to one or all the components of the system.

When studying a newly discovered endosymbiosis, analyses going beyond the simple taxonomical assignment are needed.

Given the difficulty in obtaining a reliable taxonomic assignment based on a single gene, and the interest in understanding the nature of the symbiotic relationship, sequencing the genome of the endosymbiont represents today one of the most proficient approaches. It requires the post-processing phase described above to separate sequences coming from the different genomes in the system, which may be tricky, but it can allow to perform taxonomic assignment by using many genes e.g. it enables a phylogenomic approach that is certainly superior with respect to the single gene approach (Ciccarelli, 2006; Gao and Gupta, 2012; Rokas et al., 2003). In addition, by analysing the bacterial genomes and their functional content we can get information to derive a more complete picture of the system under study.

## Endosymbiosis in the genus *Spirostomum*

The ciliate protozoa of the second case study of this thesis belong to the genus *Spirostomum*. This taxon is characterized by unicellular organisms with an elongated, flexible and highly contractile cell that can reach 4 mm in length. The species of this genus can inhabit fresh or salt water (Lynn, 2010). *Spirostomum* species are mainly described morphologically, even if diagnostic traits are often difficult to interpret. Indeed, molecular markers, such as 18S, have been used to revise and provide a more rigorous taxonomy (Schmidt et al., 2007).

*Spirostomum* is not a widely studied ciliate genus. However, some studies showed that the species of this genus can be associated with a variety of epibiont bacteria that are localized

at the outer membrane and endosymbiont bacteria living within cells or even associated with the mitochondrion (Fokin et al., 2005; Harrison et al., 1976). In (Fokin et al., 2005) bacterial taxonomy of these *Spirostomum*-associated bacteria was inferred by fluorescence *in situ* hybridization using probes specific for *Alphaproteobacteria*, *Betaproteobacteria*, *Gammaproteobacteria*, *Deltaproteobacteria* and gram-positive bacteria. The authors mostly detected *Alphaproteobacteria* and *Gammaproteobacteria*, but much more can be discovered.

## Material and Methods

### DNA analysis

We collected a total of three samples: two from Kolleru lake (India, named S8 and S9) and one from Bosco Bazzoni (Trieste, Italy; S0). The samples were collected and maintained as in (Boscaro et al., 2014). Through 18S and 16S gene sequencing we assigned a taxonomy to the hosts and the endosymbionts.

The host and the endosymbiont genomes were sequenced together, by performing a Whole Genome Amplification (WGA) on live cells using the REPLI-g Single Cell kit (Qiagen). Libraries for sequencing were prepared using the Nextera XT library Prep Kit (Illumina) and sequenced on an HiSeq X in an external facility (Admera Health LLC, NJ, USA) in paired end, with 150 bp read long.

### Data analysis

The quality of the reads was assessed with fastQC; Trimmomatic was used to cut the remaining adaptors and the low qualities regions. Read pre-processing ended by performing FLASH to merge the overlapping read pairs (Andrews et al., 2012; Bolger et al., 2014; Magoc and Salzberg, 2011). The resulting reads were assembled using SPADes with k-mers length ranging from 31 to 101 (Bankevich et al., 2012). The best k-mer length for each sample was identified using Quast (Gurevich et al., 2013) that calculates several statistics on the assembly, such as the distribution of contig lengths, the number of contigs and the N50.

We ran SeqDex (see Chapter 3, (Chiodi et al., 2019)) on each sample to retrieve the sequences coming from the endosymbiont by exploiting both nucleotide (nt, downloaded by NCBI in October 2018) and protein (nr, downloaded in November 2018) homology searches for assigning preliminary taxonomy affiliations, at a minimum percentage of identity of 80% and 85%, respectively. We also enabled the use of GC content and coverage together with k-mers in the final clustering step (for further information, see the user manual in Appendix 4). The mapping files of each of the three samples used by SeqDex were obtained with BEDtools (Quinlan and Hall, 2010).

Reads mapped on the contigs identified by SeqDex as of endosymbiont origin were re-assembled using SPADes, as above. The re-assembled contigs were analysed with BUSCO, a software that exploits a database of universally conserved genes to predict the percentage of completeness of an assembly. As the endosymbionts come from the *Betaproteobacteria*, we ran BUSCO using the corresponding database (downloaded from busco.ezlab.org on May 2019) (Simão et al., 2015). Prokka (Seemann, 2014) was used to predict coding sequences (CDS), and to annotate the corresponding proteins. The predicted proteomes of the three endosymbionts were used as input for OrthoFinder (Emms and Kelly, 2015, 2018), together with the proteomes of 168 additional genomes downloaded from the NCBI repository, chosen among the *Neisseriales* in NCBI and those classified as *Neisseriaceae* or *Chromobacteriaceae* in GTDB (gtdb.ecogenomic.org), and 9 genomes of *Gammaproteobacteria* or non-*Neisseriales Betaproteobacteria* as outgroups; they mostly come from complete genomes but also comprise incomplete assemblies coming for instance from metagenomic samples (Appendix 6).

After running OrthoFinder, we selected all the sequences of single copy core orthologs from each organism for alignment and phylogenetic analysis. The sequences were concatenated together and aligned with Muscle (Edgar, 2004); Gblocks was used to select conserved regions (default parameters) (Castresana, 2000); RAxML was used to reconstruct the phylogenetic relationships among the species using an empirically estimated gamma distribution for taking into account evolutionary rate variation, the WAG evolutionary model and empirically estimated amino acid equilibrium frequencies (GAMMA+WAG+F), with 100 bootstrap replicates (Stamatakis, 2014).

To obtain biologically meaningful information on these symbionts, we performed preliminary comparative genomics analysis to assess the ability of the symbionts to interact with the environment; we reasoned that it surely requires the ability to sense the conditions of the surrounding space, a task that in bacteria is often associated to two-component systems. We hypothesize that the number of proteins involved in two component systems might change depending on the necessity for interaction with the outside environment and therefore we counted proteins with similarity to Pfam models (El-Gebali et al., 2019) built on Histidine Kinases, Methyl-accepting chemotaxis proteins and Response Regulator domains. Scanning on proteins was made by using the software HMMer3.1b2 (hmmer.org). Additionally, a bacterium interacts with its environment by importing/exporting chemicals through specific transporters and also in this case one can expect that a free-living species should require a vaster array of transporters. To explore if this evolutionary pattern is indeed significant for our species, we counted the number of ABC transporters in these genomes. Finally, the *secretome* of the bacteria - the ensemble of proteins that get secreted outside the cell - might also show interesting evolutionary patterns when contrasting free-living and endosymbiont species, as the latter lives in a very constant and protected environment, and are therefore expected to show a reduction in the number of secreted proteins. To have an approximation of the size of the secretome, we used SignalP (Almagro Armenteros

et al., 2019) with the gram negative model for predicting the presence of secretion signal peptides at the N-terminal of the proteins. Another ability that might not be necessary during endosymbiosis is motility, and we checked this by using a Pfam model of the PilC protein, that is involved in the pilus mediated adherence to surfaces or the host cells in commensals and pathogens. The Pfam models used are listed in Table 8.

Table 8 – List of the Pfam models used to predict with HMMer the presence of protein domains involved in the interaction with the environment as stimuli sensing (HisKA, MCP), response (RR), transmembrane transport (ABC) and movement (PilC)

Target	Model	Pfam code
Histidine kinase A	HisKA	PF00512
Methyl-accepting chemotaxis	MCPsignal	PF00015
Response Regulator receiver	RR	PF00072
Neisseria PilC propeller	Neisseria_PilC	PF05567
ABC transporter	ABC_tran	PF00005

## Results

The comparison of the 18S genes sequenced from our three *Spirostomum* species revealed that S8 corresponds to *S. teres* while both S0 and S9 to *S. minus*. On the converse, the 16S sequences of the endosymbionts were not sufficient to identify the endosymbiont's species by similarity only. A preliminary phylogenetic tree was built using all the 16S gene sequences of the *Neisseriales* available in NCBI (data not shown) and showed that the *Betaproteobacteria* found in our samples belong to the *Neisseriaceae* family.

The total DNA extractions allowed to obtain enough material to perform the WGS of the three samples, resulting in 43'766'268 (S0), 41'701'334 (S8) and 41'525'275 (S9) reads.

With SeqDex we were able to obtain genomic contigs of the endosymbionts in samples S0 and S9, but processing of the S8 sample did not worked as expected. By comparing the coverage of the 16S genes of the three samples, we observed that the coverage in the S8 sample is almost 10 times lower than the coverage in the S0 and S9 samples. We supposed that this influenced the quality of the assembly and the performance of SeqDex on this sample (Table 9).

Table 9 – The table shows the coverage values of the contigs containing the 16S target gene in the three samples, the total number of fragments used in the assemblies and the coverage normalized by number of fragments and multiplied by 10<sup>6</sup>

Sample	Coverage	Number of fragments	Coverage/reads
S0	5.60	36851159	0.15
S9	5.33	41525275	0.14
S8	0.78	41701334	0.02

We decided to compare the genomic contigs of the three endosymbionts obtained with SeqDex by using the Average Nucleotide Identity (ANI) calculator of EZBioCloud (Yoon et al., 2017) to assess if we could use S0 and/or S9 as a reference to improve the processing of sample S8. The results are shown in Table 10: S0 and S8 are highly similar and thus can be considered as part of the same species.

Table 10 - The table shows the Average Nucleotide Identity (ANI) calculated using EZBioCloud web tool.

	Total length (bp)	ANI value vs S9	ANI value vs S8
S0	999600	70.71 %	96.76 %
S9	1052640	-	-
S8	358020	70.94 %	-

We used the reassembled S0 contigs as a reference to obtain the contigs of the endosymbiont in S8 by using Quast. The reads mapping only to the S8 endosymbiont contigs were retrieved and assembled using SPADes as done for the other two samples; the best assembly was chosen based on the Quast output.

In Table 11 we show the total length and the number of reassembled contigs for each sample together with the number of CDS found by Prokka and the percentage of completeness calculated with BUSCO.

Table 11 – The table shows various characteristic relative to the reassembled genomic contigs of the three endosymbionts

	Number of Contigs	Total length (Bp)	CDS (Prokka)	Completeness in % (BUSCO)
S0	34	1014825	948	63.4
S9	55	1076025	962	61.9
S8	26	894166	823	57.2

With OrthoFinder we found 37 single copy core orthologs. After the elaboration with Muscle and Gblocks we ended with an alignment of 6250 positions that was used to reconstruct the phylogenetic relationships among these genomes with RAxML (Figure 13). In the tree, four main named groups can be observed: *Neisseriaceae* is composed only by genomes belonging to this family (hereafter: clade *Neisseriaceae*); *Chromobacteriaceae*, is composed by genomes of the family and a *Xenophilus* genome (hereafter: clade *Chromobacteriaceae*); *Chromobacteriaceae + Neisseriaceae* is instead composed by *Chromobacteriaceae* plus a *Neisseriaceae* genome (hereafter: Clade A); *Chromobacteriaceae + others* is composed mainly by *Chromobacteriaceae* plus some *Burkholderiales* and a *Neisseriaceae* (hereafter Clade B; for further detail, see Figure 13). We compared our tree to the last published phylogenomic tree of the *Neisseriales* family (Adeolu and Gupta, 2013). In this article the authors observed that the order is composed by two families: the *Neisseriaceae* and the *Chromobacteriaceae*. These two families are congruent to our *Neisseriaceae* and *Chromobacteriaceae* clades. However, we included more genomes than Adeolu and Gupta in our analysis, and not only complete assemblies but partial ones that often come from metagenomic studies, whom assigned taxonomy might not be sure.

The three endosymbionts under analysis are basal to the *Neisseriaceae* clade.

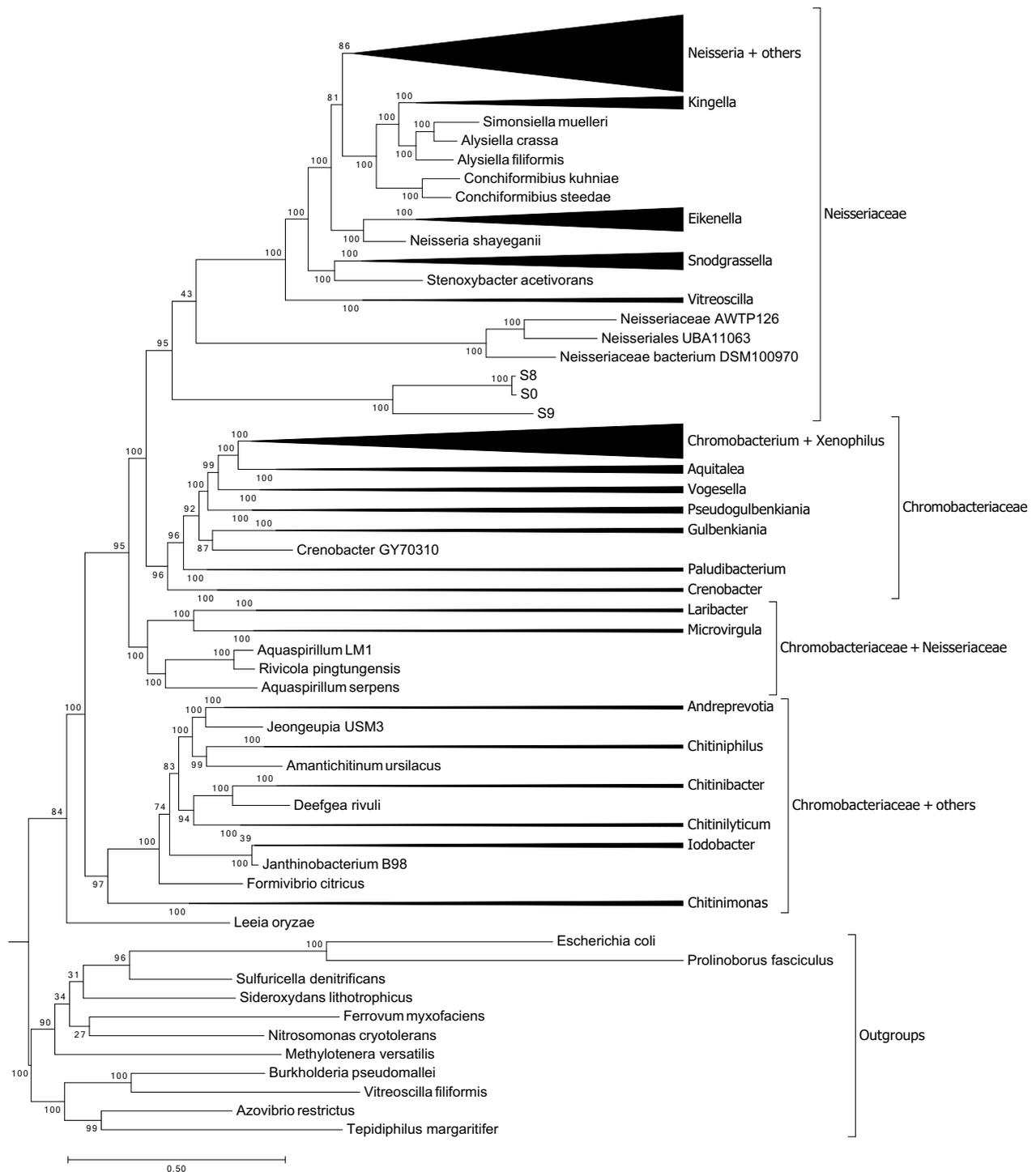


Figure 13 – Phylogenetic tree of the *Neisseriales* obtained with RAxML. The family is composed by two order: *Chromobacteriaceae* and *Neisseriaceae*, which in the tree are well differentiated and form two separated sister clusters. Inside *Chromobacteriaceae* order, *Chromobacterium* genus contains *Xenophilus* AP218F, which is classified as a *Burkholderiales*. Inside the *Neisseria* genus there are also two *Morococcus cerebrosus* genomes, a *Bergella denitrificans*, a *Uruburella suis* and a *Kingella potus*, all species of the genus *Neisseriaceae*. In the tree there are also other two clusters, one composed mainly by *Chromobacteriaceae* plus *Rivicola pingtungensis*, which is classified as a *Neisseriaceae* (clade A), the other composed mainly by *Chromobacteriaceae* plus *Burkholderiales* (genus *Chitinimonas*, 3 genomes, and *Janthinobacterium*, 1 genome) and *Neisseriaceae* (*Amantichitinium usialicus*) (clade B).

Figure 14 shows the tree obtained with RAxML integrated with the information about the number of predicted domains for each Pfam model used, the number of proteins with a signal peptide found by signalP, the number of xenologs genes as defined by OrthoFinder - genes that likely reached a genome from another by means of Horizontal Gene Transfer (HGT) - and the number of duplicated genes also found by OrthoFinder. Considering the tree, we highlighted 5 major clades: the blue clade is the *Neisseriaceae* clade of the tree in Figure 13; the red clade is inside the blue one and is composed by the three endosymbionts under analysis (hereafter: clade E); the green clade is the *Chromobacteriaceae* clade; the orange is clade B; the purple is composed by outgroup genomes. We decided to not consider the clade A from Figure 13 for further analysis as it is composed only by few genomes. In the heatmap the counts of each kind of protein were normalized with respect to the total number of proteins in each genome and is presented for clarity in a log10 scale. As a consequence of the partiality of our genomes, some genes may be missing because they are not in the genome or because they have not been sequenced or retrieved, but without a complete genome to be used as a reference, it is impossible to tell which one explanation is true for all genes. However, we can make the hypothesis that the loss of genes due to the methodology should be homogeneously distributed among all protein families, which should not be true for proteins or protein domains that were purged from the genome in evolutionary time as a consequence of the life style of the bacterium. Therefore, with this analysis we hoped to detect some pattern in the evolution of the size of the different protein families and to be able to connect them to the endosymbiotic lifestyle. All the clades are well differentiated, indicating that within a clade there is a tendency to conserve a certain family size, but this is not true when evolutionary time increases, and species from different clades are more different. The *Chromobacteriaceae* and A clades seem concordant even if the first is pure and the second is mixed. The clade E, basal to the *Neisseriaceae*'s one, seem to diverge from it when considering the proportion of ABC transporter domains and signal peptides.

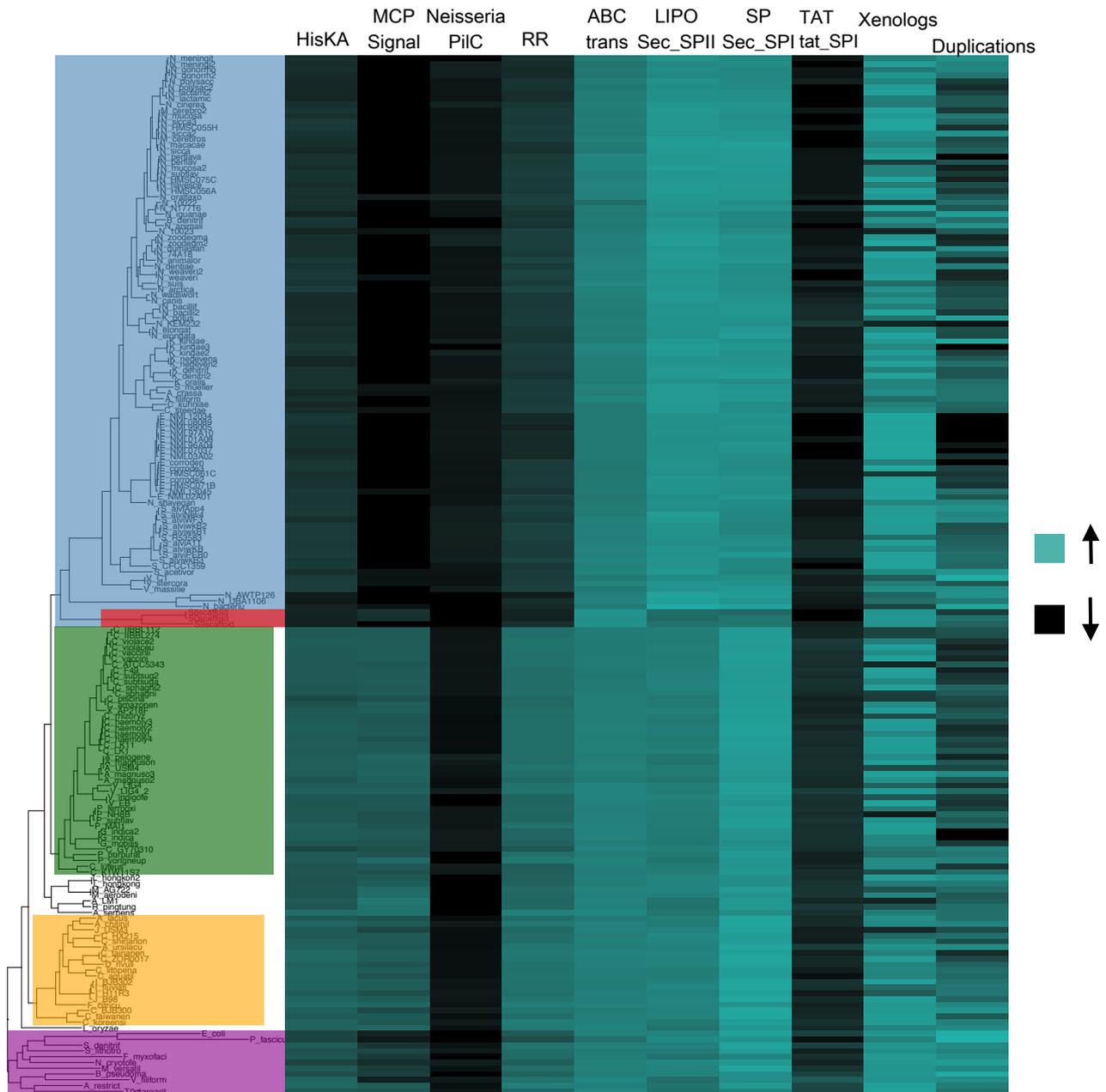


Figure 14 - The figure shows the phylogenetic tree obtained with RAxML together with a heatmap reassuming for each genome the number of predicted Pfam domains, signal peptides, xenologs and duplicated genes, normalized by total number of proteins and reported in a log 10 scale. In the tree we highlighted 5 major clades: the blue is composed only by *Neisseriaceae*; the green that is composed only by *Chromobacteriaceae*; the orange that is composed mainly by *Chromobacteriaceae*; the purple that is composed only by the outgroups; the red clade is inside the blue clade and is composed by the three endosymbionts.

Figure 15 highlights the differences among the 5 clades. Each chart shows for each marker the proportion reported in the heatmap of Figure 14. Clade E is similar to the *Neisseriaceae*'s one when considering HisKA, MCPsignal, RR, TAT signal peptides, proportion of xenologs and duplicated genes. However, it shows a higher proportion of ABC\_transporters predicted domains and lower LIPO and SP signal peptides compared to the other clades.

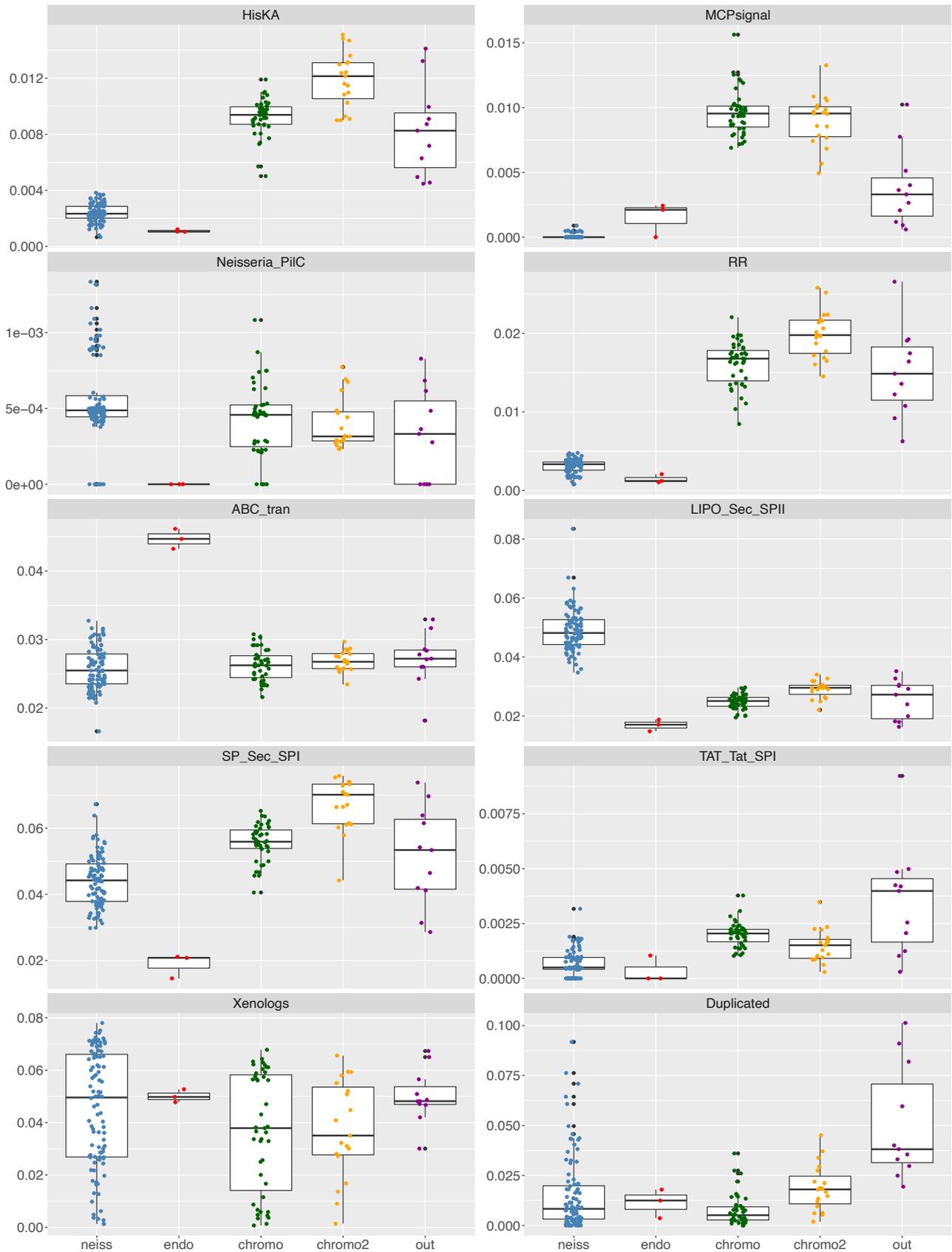


Figure 15 - Each chart refers to a different proportion of the marker considered (proteins domains predicted with HMMer using Pfam model, signal peptides predicted with signalP, xenologs and duplicated genes calculated by OrthoFinder). In the graphs, the points are the numbers of each marker normalized by total number of proteins in the, whereas the boxplot reassumes their distribution for each clade. The colours of the points are concordant to the clades described previously. Blue= *Neisseriaceae* clade (neiss); red=

endosymbionts under analysis (endo); Green= *Chromobacteriaceae* clade (chromo); Orange= clade B, composed mainly *Chromobacteriaceae* (chromo2); Purple= outgroup clade(out).

## Discussion

The molecular and bioinformatic analyses allowed to obtain fairly complete genomes of two samples out of three. The S8 sample has a lower 16S gene coverage compared to S0 and S9. This could be due to a lower abundance of the endosymbiont in the sample or a larger contaminant level. Even if ciliate cells are starved and washed prior to amplification and sequencing, they might retain DNA sequences coming from the bacteria they have fed on, and these can end being sequenced. By using S0 as a reference, we obtained the genomic sequences in the S8 samples, even if the total length of the reassembly is shorter than the other two samples (Table 11). This suggest that maybe the low performance of SeqDex can be imputable to both the highly presence of contaminating sequences and a lower abundance of the bacteria inside the host cell.

The percentages of completeness calculated by BUSCO and reported in Table 11 suggest that the obtained genomes of the three samples are incomplete. To calculate these percentages, BUSCO compares the input proteome to a database composed by the core genes of a certain taxonomic group. We used the *betaproteobacteria* database, that is composed mainly by free living and pathogen species. As endosymbionts are missing from the set of genomes used to define the *Betaproteobacteria* core, we explain the low percentage of completeness reported by BUSCO as derived from two contributions: the first is the decrease due to the endosymbiotic life style of the bacterium, and this is a true genome reduction; the second is the fact that our sequencing was not able to catch the entire genome of the endosymbiont. For these reasons, we think that the BUSCO percentage represents a lower bound for the true completeness of our assemblies.

The phylogenetic tree in Figure 13 shows high bootstrap values, so it is highly affordable, even if presents some mixed groups composed by organism of different classification. We compared our phylogenesis to the one recently published by (Adeolu and Gupta, 2013). The authors used a phylogenomic approach similar to ours: they build a long protein concatemer to then reconstruct the phylogenetic tree based on the maximum-likelihood. The differences of our tree with the published one could be mainly due to wrong taxonomic assignment of metagenomic derived samples, but we can't exclude some phylogenetic misplacement that however, seeing the exceptionally high bootstrap values, should be only a few. Moreover, our tree includes more species than the tree published by Gupta et al., and the phylogenesis might have slightly changed.

The endosymbionts in our samples are basal to the *Neisseriaceae*, with S0 and S8 most close, probably belonging to the same species, and S9 just slightly farther. S0 and S8 come from different host species, *S. minus* and *S. teres*, respectively while the two *Spirostomum* of the same species seem to have different endosymbionts (S0 and S9, found both in *S.*

*minus*). Thus, we may infer that there is no co-cladogenesis nor co-phylogenesis among host and endosymbionts, a condition that is instead frequently found in most obligate endosymbiosis indicating that a species is strictly associated with a certain host species since long time. At the same time, the fact that we find the same endosymbiont in different host species might indicate that this endosymbiosis might be quite widespread among ciliates.

In stable endosymbiosis, bacteria usually face genomic rearrangements with many genes loss as they adapt to the stable and rich conditions of the intracellular milieu. The genes involved in the interaction with the external environment may be the ones more drastically affected by these rearrangements. We chose to perform a functional analysis on the proteins of all the organisms used in the phylogenomic analysis to be able to compare genomic properties of our endosymbionts to their relatives and highlight the differences in an evolutionary perspective. We decided to predict the presence of protein domains involved in environmental signal perception and processing (HisKA, MCPsignal, RR), motility (*Neisseria*\_PilC) and the exchange of substances (ABC\_tran), as also to predict the presence of standard (SP\_Sec\_SPI), lipoprotein (LIPO\_sec\_SPII) or Tat (TAT\_tat\_SPI) signal peptides.

The closest relatives of the endosymbionts under analysis belong to the *Neisseriaceae* clade, which is a family mainly comprising obligate pathogens or commensals (Adeolu and Gupta, 2013). This clade shows a generally low proportion of domains for detecting and responding to external stimuli, as some HisKA and RR are present, but not MCP was observed. The endosymbionts even show a reduced proportion of domains involved in the detection and response to stimuli: HisKA and RR domains have low abundance in these genomes, whereas some MCP signal domains are present, in opposition to the *Neisseriaceae* clade. The differences between the two clades considering these three domains are, although, not so pronounced in agreement with their phylogenetic placement (Figure 13). The proportion of domains involved in the detection and response to external stimuli highly reflects the lifestyle of the organisms. *Chromobacteriaceae* is a family composed mainly by free-living species, with some facultative pathogens, and the *Chromobacteriaceae* and B clades, which is mainly composed by *Chromobacteriaceae*, show the higher proportion of these environment interaction domains. The obligate host-associated lifestyle of the *Neisseriaceae* reflects in their proportion of proteins containing HisKA, RR and MCP signal domains, that are lower than in the *Chromobacteriaceae* clade. The *Neisseriaceae* are adapted to a life in association with an host, thus they need to respond to external stimuli, as these capabilities may be involved in the establishment of a commensal relationship or in the development of the pathogenicity, however they live in a more stable condition compared to the free-living. The endosymbionts show the lower proportion of these domains among all the clades analysed. Thus, in accordance to the observed trend, they might be adapted to a more stable condition compared to the *Neisseriaceae*, as the intracellular milieu of a host.

The clade E differentiates from the *Neisseriaceae* one as it shows an increased proportion of ABC transporter domains, more than all other clusters, and a lower proportion of standard and lipoproteins signal. The presence of a higher proportion of ABC transporter domains among all *Neisseriales* highlights the importance of the exchange of products with the external environment: the endosymbionts may not have more ABC transporter domains than the other *Neisseriales* but the fact that their proportions are larger, might suggest that they are important for the survival, as while likely undergoing a genome reduction, these species kept ABC transporters in higher proportion.

The signal peptides are regions that flags the proteins that will be targeted by the secretion system and are involved in the interaction with the environment, as in virulence and pathogenesis. Bacteria secrete proteins to identify, bind, degrade complex metabolites and transport them into the cell, as also to adapt and interact with the environment, and to communicate (Gagic et al., 2016). Species of the *Neisseria* genus secrete proteins and lipoproteins to actively uptake iron when its concentration is low, as when growing in the blood of mammals where free iron is extremely low; to prevent the activation of the immune system of the host; to produce biofilms in commensals lifestyles, or to induce the aggregation of multiple bacterial cells as a defence against phagocytosis during infections; to adhere to the host cells and invade them (Tomassen and Arenas, 2017). The gram negative organisms, as the *Neisseriales*, usually have a secretion system of type Sec, which transports standard signal peptides (SP) and lipoproteins (LIPO), but not TAT peptides (these are transported by the Tat secretion systems); this is confirmed by the charts in Figure 15 that show the fairly absence of TAT signal peptides in the proteins of the considered clades. We show that our endosymbionts have a reduced proportion of proteins carrying secretion signal peptide. This may be compatible to a symbiotic lifestyle. If these bacteria have adapted to the stable and rich environment present inside the host cell, it is possible that genes encoding proteins needed for the interaction with the external milieu, communication and also, maybe, pathogenicity are no more needed. Thus, these genes may have faced degenerative mutation, deletions, up to removal.

An evidence that might suggest that these *Spirostomum* and endosymbiont species are bound by a stable symbiotic relationship is the absence of any gene related to the production, secretion and maintenance of the LPS barrier. During a manual exploration of the proteins obtained with Prokka by using the KEGG database we observed the complete absence of the lipopolysaccharide biosynthetic pathway in all three endosymbionts (Kanehisa et al., 2016). Besides knowing that our genomes are incomplete, the total absence of an entire pathway of almost 20 genes in the three samples is a strong indication that the process is indeed missing in the genome. The LPS barrier is important for the survival of the gram negatives but, as suggested by the *Baumannia cicadellinicola* example discussed previously, being mostly important for protection, it may be one essential function becoming accessory once the environment becomes as stable as the intracellular milieu.

Obligate endosymbionts, such *Buchnera aphidicola*, are considered to have the more stable genomes among bacteria as they show no rearrangements, no HGT and low mutation rate (Mira et al., 2002). Horizontal transfer of genes is an important source of variation, introduces novelties and allows to replace degenerated genes. However, the frequency of such events is reduced by an intracellular lifestyle like in endosymbionts, particularly in the obligated ones, or obligate parasites such as *Chlamydia* and *Rickettsia* species (Wernegreen, 2015). It is possible that the regions involved in the integration of horizontally transmitted genes may be involved in the loss of portion of genomic regions that results from the high mutation rate, duplications and pseudogenization that bacteria may face during the establishment of an obligate endosymbiosis (Andersson and Kurland, 1998; Darby et al., 2007; Gil et al., 2002; McCutcheon and Moran, 2012; Mira et al., 2002; Moran and Wernegreen, 2000). In our endosymbionts, it is possible that the sequences that OrthoFinder catalogues as xenologs are really genes present in their genome due to HGT, but they could also represent genes that were erroneously included in the assembly. Also, the fact that the proportion of xenologs genes shown in Figure 14 and Figure 15 are completely in line with the other clades suggests that these genomes have faced horizontal gene transfers, but it is not known if they are a residual trace of HGT not already removed from the genome, or if the bacteria is still currently able to exchange genes.

Another important source of variation is the duplication of genes or portion of the genomes. Bacteria, both free-livings and commensals, show a high level of duplication events in their genomes, as these allow to relax selective constrains on genomic regions and thus favour the introduction of novelties. Comparative studies on genomic rearrangements show rare rearrangements in obligate endosymbionts (Wernegreen, 2015). However, studies conducted on facultative and young obligate endosymbionts showed a different picture: in the very first phase of the establishment of the relationships the bacterial genome shows high pseudogenization, explosion of insertion sequences and other mobile elements, as also inversions and duplications (Wernegreen, 2015). We are not able to compare the duplication rate of endosymbionts during the establishment of the symbiotic relationship to other bacteria as there are not published studies about it. However, we consider the fact that our bacteria show a proportion of duplicated genes consistent with the non-endosymbionts, as the *Neisseriales* (Figure 14 and Figure 15), as indicating the presence of genomic rearrangements compatible with the establishment of an endosymbiosis. Also, in parallel to the considerations done for the xenolog proportions, we are not able to say if our endosymbionts are currently facing these rearrangements or if these are the residuals of events already happened but not still purged.

In conclusion we can say that besides the absence of co-cladogenesis or co-phylogenesis, the presence of closely related endosymbionts in *Spirostomum* species coming from distant sampling sites, together with all the considerations done on the comparative functional analysis performed, suggests that this relationship could be stable even if likely

not obligate, and widespread across the host genus. However, more effort to complete the three genomes is needed to completely characterize the system.

This work represents one of the few examples of *Betaproteobacteria* endosymbionts, and as far as known, one of the firsts *Neisseriales* endosymbionts description, highlighting the fact that the order may be composed also by not pathogen or commensal species that, at today, are completely underrepresented.

---

## CHAPTER 5

General discussion

This thesis considers two case studies where the currently used techniques for species identification do not provide an affordable taxonomic affiliation. The first one concerns the identification of morphologically cryptic anuran taxa which is complicated by the presence of interspecific breeding with production of viable and fertile offspring due to a form of sexual parasitism. The second case study is a putative endosymbiosis involving a bacterium of the *Neisseriales* order and a host belonging to the genus *Spirostomum*. Here, the taxonomy determination is complicated by the fact that the bacteria belong to new and not already described species with no closely related sequences in the public databases. Cryptic species are morphologically similar groups of taxa, such that some may completely lack phenotypic informative characters that may be used to determine the taxonomic affiliations. It is therefore important to find features that can be used for a correct taxonomic affiliation. The characters investigated may be the life history, the behaviour, the vocalisations, the physiology, as also the genotypic data. Lots of examples can be found in the literature. One is *Bemisia tabaci*, a worldwide vegetable, ornamental and field crop pest, that was thought to be cosmopolitan but is in reality a complex of morphologically indistinguishable cryptic species. In this case the species can be determined by using a singular informative mtDNA marker, but a mating test is needed to confirm the reproductive isolation and thus the existence of different species according to the biological species concept (Xu et al., 2010). However, the breeding tests are not error-free, as the absence of offspring may be due to reproductive isolation or induced by the absence of some environmental condition needed for successful reproduction, as observed for wild animals kept in captivity. Another example is provided by the lizards of the genus *Diporiphora*, which exhibit little morphological informative characters. The authors in (Smith et al., 2011) used two markers, one nuclear and the other mitochondrial, to be able to differentiate the two species under study. As said before, the mtDNA genes have demonstrated to have good resolution power and good stability to be used as universal marker for species determination. These, moreover, have been preferred to nuclear markers which have introns, can undergo recombination events and diverge through the acquisition of insertions and/or deletion that complicate sequence alignment and taxonomic reconstruction. Thus, exploiting a single marker may be error prone, especially when rearrangements are common in one or more of the taxa under analysis.

The approach used in Chapter 2 allows to take advantage of the resolution power of the mtDNA marker and to overcome its limitation in the detection of hybrids by using STR nuclear markers, avoiding all the problems linked to nuclear genes sequences. Our study is not the first one in which mtDNA data are coupled to STRs to identify cryptic species and putative hybrids (Fekete et al., 2012; Trigo et al., 2013); however, it represents the first attempt to use these data to identify pure species and hybrids in the *Pelophylax* complex. Indeed, this approach needs more implementations, as species specific microsatellite are needed to make it more affordable.

The second study case of this thesis focused on three samples of ciliates of the genus *Spirostomum* that harbour endosymbionts. This case study is complicated by the difficulty to obtain the data to perform the taxonomic affiliation of the bacteria, as morphology is little informative, the endosymbiont sequences are hard to isolate, and the databases miss sequences that are closely related to those from the endosymbiont. Moreover, these three bacteria likely belong to new species. In this case, it is also important to perform a functional analysis to get an idea of the possible role of the endosymbiont within the host; to get this information, we worked to extract the genomic sequences of the endosymbiont. To achieve this task, we developed SeqDex (Chapter 3), a bioinformatic tool designed to partition the different genome sequences present in a sample. In our case, SeqDex helped in isolating the endosymbiont genomic sequences from all the rest (host + contaminants).

SeqDex is based on state-of-the-art machine learning models that use compositional properties to provide the separation of genomic sequences from different sources in a mixed sample. Additionally, I worked carefully to implement the code required to calculate performance measure of the classification made by SeqDex, to inform the user about the goodness of the solution found.

SeqDex was tested on several test-cases before running it on the three samples of my thesis, on which it performed quite well, allowing to retrieve a good proportion of the genome of each symbiont.

As the preliminary 16S gene sequence analysis suggested that the three endosymbionts belong to the *Neisseriales* family, we used the available genomes to place our samples in a phylogenomic tree of the family. This analysis confirmed the absence of very close relatives to the endosymbionts in the family. The complete absence of known *Neisseriales* endosymbionts in the literature makes hard to analyse and interpret the genomic data: not knowing phylogenetically close examples means that we have no idea of the molecular functions or pathways that might be important for these bacteria. Many endosymbiotic systems have been previously studied, but often show peculiarities depending on the specific requirements of the host. The only shared characteristics are the rearrangements and gene loss faced by the obligate endosymbionts, but what genes and pathways are lost, it depends on the specific situation. For instance, some symbiont produces essential amino acids, other vitamins and so on, depending on the metabolic abilities of the host and the environmental requirements. One may be tempted to think that symbionts from different clades that are in a relationship with host's also from different clades should nevertheless show some common evolutionary trend and therefore should have faced similar genomic rearrangements in addition to specific ones, but the truth is that we do not have enough information to confirm this. One property that should be general enough is the ability of interacting with the environment. Free living bacteria must be able to detect population density, the presence of nutrients in the environment and to transduce this information in the cell with appropriate signalling pathways. For this reason, we focused our comparative

genomics analyses on the interaction with the environment (detection of stimuli, transport, signalling), in addition to HGT and gene duplications.

The picture that emerged from this analysis suggests that these bacteria might be facultative symbionts still undergoing rearrangements and gene losses probably towards an obligate symbiotic behaviour. The high proportion of transporter domains compared to the other clades suggest that these genes are conserved, maybe because are involved in the interaction with the host.

The nature of the relationship, thus the action each organism takes on the other in the endosymbiotic system, remains unclear. The examples described in the introduction of Chapter 4 highlighted the difficulty to restrain these actions to a unique type. It is possible that our bacteria are endosymbionts, facultative or nearly obligated, that interact positively or in a neutral manner with the host; it is also possible that they are pathogens, as most of the *Neisseriaceae* are, or that they have been pathogens which then lose their pathogenicity and established a mutualistic and, maybe, obligate relationship. To be able to better define the endosymbiotic system more analyses, and maybe more samples, are needed.

The analyses conducted in Chapter 4 represent the first attempt to study completely *ex novo* an endosymbiotic relationship involving a *Neisseriaceae* bacterium. Also, the analyses allowed to obtain some information about the bacteria and the endosymbiotic relationship, even if these are not exhaustive.

---

## REFERENCES

- Aboelsoud, N. H. (2010). Herbal medicine in ancient Egypt. *J. Med. Plants Res.* 4, 082–086. doi:10.5897/JMPR09.013.
- Adeolu, M., and Gupta, R. S. (2013). Phylogenomics and molecular signatures for the order *Neisseriales*: proposal for division of the order *Neisseriales* into the emended family *Neisseriaceae* and *Chromobacteriaceae* fam. nov. *Antonie Van Leeuwenhoek* 104, 1–24. doi:10.1007/s10482-013-9920-6.
- Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423. doi:10.1038/s41587-019-0036-z.
- Altschul, S. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* 215, 403–410. doi:10.1006/jmbi.1990.9999.
- Andersson, S. G. e., and Kurland, C. G. (1998). Reductive evolution of resident genomes. *Trends Microbiol.* 6, 263–268. doi:10.1016/S0966-842X(98)01312-2.
- Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., and Wingett, S. (2012). FastQC.
- Arioli, M. (2007). Reproductive patterns and population genetics in pure hybridogenetic water frog populations of *Rana esculenta*. doi:10.5167/uzh-163641.
- Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.* 79, 137–158. doi:10.1084/jem.79.2.137.
- Bandi, C., Anderson, T. J. C., Genchi, C., and Blaxter, M. L. (1998). Phylogeny of *Wolbachia* in filarial nematodes. *Proc. R. Soc. London. Ser. B Biol. Sci.* 265, 2407–2413. doi:10.1098/rspb.1998.0591.
- Bandi, C., Trees, A. J., and Brattig, N. W. (2001). *Wolbachia* in filarial nematodes: evolutionary aspects and implications for the pathogenesis and treatment of filarial diseases. *Vet. Parasitol.* 98, 215–238. doi:10.1016/S0304-4017(01)00432-0.
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* 19, 455–477. doi:10.1089/cmb.2012.0021.
- Bateson, W. (1906). The progress of genetic research. in *Report of the Third International Conference 1906 on Genetics*, 91.
- Berticat, C., Rousset, F., Raymond, M., Berthomieu, A., and Weill, M. (2002). High *Wolbachia* density in insecticide-resistant mosquitoes. *Proc. R. Soc. London. Ser. B Biol. Sci.* 269, 1413–1416. doi:10.1098/rspb.2002.2022.
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170.
- Boscaro, V., Carducci, D., Barbieri, G., Senra, M. V. X., Andreoli, I., Erra, F., et al. (2014). Focusing on Genera to Improve Species Identification: Revised Systematics of the Ciliate *Spirostomum*. *Protist* 165, 527–541. doi:10.1016/j.protis.2014.05.004.

- Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi:10.1023/A:1010933404324.
- Brown, A. M. V., Wasala, S. K., Howe, D. K., Peetz, A. B., Zasada, I. A., and Denver, D. R. (2016). Genomic evidence for plant-parasitic nematodes as the earliest *Wolbachia* hosts. *Sci. Rep.* 6, 34955. doi:10.1038/srep34955.
- Brown, A. M. V., Wasala, S. K., Howe, D. K., Peetz, A. B., Zasada, I. A., and Denver, D. R. (2018). Comparative Genomics of *Wolbachia*–*Cardinium* Dual Endosymbiosis in a Plant-Parasitic Nematode. *Front. Microbiol.* 9, 2482. doi:10.3389/fmicb.2018.02482.
- Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi:10.1038/nmeth.3176.
- Cacho, N. I., and Baum, D. A. (2012). The Caribbean slipper spurge *Euphorbia tithymaloides*: the first example of a ring species in plants. *Proc. R. Soc. B Biol. Sci.* 279, 3377–3383. doi:10.1098/rspb.2012.0498.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST plus: architecture and applications. *BMC Bioinformatics* 10, 1. doi:Artn 421\nDoi 10.1186/1471-2105-10-421.
- Camin, J. H., and Sokal, R. R. (1965). A Method for Deducing Branching Sequences in Phylogeny. *Evolution (N. Y.)* 19, 311. doi:10.2307/2406441.
- Campbell, M. A., Van Leuven, J. T., Meister, R. C., Carey, K. M., Simon, C., and McCutcheon, J. P. (2015). Genome expansion via lineage splitting and genome reduction in the cicada endosymbiont *Hodgkinia*. *Proc. Natl. Acad. Sci.* 112, 10192–10199. doi:10.1073/pnas.1421386112.
- Canestrelli, D., and Nascetti, G. (2008). Phylogeography of the pool frog *Rana (Pelophylax) lessonae* in the Italian peninsula and Sicily: multiple refugia, glacial expansions and nuclear-mitochondrial discordance. *J. Biogeogr.* 35, 1923–1936. doi:10.1111/j.1365-2699.2008.01946.x.
- Castelli, M., Sabaneyeva, E., Lanzoni, O., Lebedeva, N., Floriano, A. M., Gaiarsa, S., et al. (2019). *Deianiraea*, an extracellular bacterium associated with the ciliate *Paramecium*, suggests an alternative scenario for the evolution of *Rickettsiales*. *ISME J.* 13, 2280–2294. doi:10.1038/s41396-019-0433-9.
- Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol. Biol. Evol.* 17, 540–552. doi:10.1093/oxfordjournals.molbev.a026334.
- Cavalier-Smith, T. (2010). Deep phylogeny, ancestral groups and the four ages of life. *Philos. Trans. R. Soc. B Biol. Sci.* 365, 111–132. doi:10.1098/rstb.2009.0161.
- Cavalli-Sforza, L. L., and Edwards, A. W. F. (1967). Phylogenetic analysis: models and estimation procedures. *Evolution (N. Y.)* 21, 550–570. doi:10.1111/j.1558-5646.1967.tb03411.x.
- Chambers, R. (2009). *Vestiges of the Natural History of Creation*. Cambridge: Cambridge

University Press doi:10.1017/CBO9780511693168.

- Charif, D., and Lobry, J. R. (2007). "SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis," in *Structural approaches to sequence evolution*, 207–232. doi:10.1007/978-3-540-35306-5\_10.
- Chiodi, A., Comandatore, F., Sasseria, D., Petroni, G., Bandi, C., and Brilli, M. (2019). SeqDex: A Sequence Deconvolution Tool for Genome Separation of Endosymbionts From Mixed Sequencing Samples. *Front. Genet.* 10, 853. doi:10.3389/fgene.2019.00853.
- Christiansen, D. G., and Reyer, H.-U. (2009). From clonal to sexual hybrids: genetic recombination via triploids in all-hybrid populations of water frogs. *Evolution (N. Y.)* 63, 1754–1768. doi:10.1111/j.1558-5646.2009.00673.x.
- Chung, M., Small, S. T., Serre, D., Zimmerman, P. A., and Dunning Hotopp, J. C. (2017). Draft genome sequence of the *Wolbachia* endosymbiont of *Wuchereria bancrofti* Wb. *Pathog. Dis.* 75, ftx115. doi:10.1093/femspd/ftx115.
- Ciccarelli, F. D. (2006). Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science (80- )*. 311, 1283–1287. doi:10.1126/science.1123061.
- Clement, M., Posada, D., and Crandall, K. A. (2000). TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* 9, 1657–1659. doi:10.1046/j.1365-294x.2000.01020.x.
- Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., et al. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, D633–D642. doi:10.1093/nar/gkt1244.
- Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi:10.1007/BF00994018.
- Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal* 1695, 1–9.
- Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., et al. (2007). Random forests for classification in ecology. *Ecology* 88, 2783–2792. doi:10.1890/07-0539.1.
- Darby, A. C., Cho, N.-H., Fuxelius, H.-H., Westberg, J., and Andersson, S. G. E. (2007). Intracellular pathogens go extreme: genome evolution in the *Rickettsiales*. *Trends Genet.* 23, 511–520. doi:10.1016/j.tig.2007.08.002.
- Denisko, D., and Hoffman, M. M. (2018). Classification and interaction in random forests. *Proc. Natl. Acad. Sci.* 115, 1690–1692. doi:10.1073/pnas.1800256115.
- Douglas, A. (1994). Symbiotic interactions. *J. Mar. Biol. Assoc. United Kingdom* 74, 743–743. doi:10.1017/S0025315400047810.
- Douglas, A. E. (1998). Nutritional Interactions in Insect-Microbial Symbioses: Aphids and Their Symbiotic Bacteria *Buchnera*. *Annu. Rev. Entomol.* 43, 17–37. doi:10.1146/annurev.ento.43.1.17.
- Doyle, J. J., and Gaut, B. S. (2000). Evolution of genes and taxa: A primer. *Plant Mol. Biol.*

42, 1–23. doi:10.1023/A:1006349518932.

- Drancourt, M., Bollet, C., Carlioz, A., Martelin, R., Gayral, J. P., and Raoult, D. (2000). 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. *J. Clin. Microbiol.* 38, 3623–3630.
- Drummond, A. J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., et al. (2011). Geneious v5.4. doi:http://www.geneious.com/.
- Dubois, A. (1992). Notes sur la classification des Ranidae (Amphibiens, Anoures). *Bull. Mens. la Société linnéenne Lyon* 61, 305–352. doi:10.3406/linly.1992.11011.
- Earl, D. A., and VonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi:10.1007/s12686-011-9548-7.
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi:10.1093/nar/gkh340.
- El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432. doi:10.1093/nar/gky995.
- Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16, 157. doi:10.1186/s13059-015-0721-2.
- Emms, D. M., and Kelly, S. (2018). OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences. *bioRxiv*, 466201. doi:10.1101/466201.
- Ester, M., Hans-Peter, K., Jorg, S., Xiaowei, X., and Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Clustering Algorithms for Discovering Clusters in large spatial databases with noise. *AAAI KDD-96 Proc.* 96, 635–654. doi:10.1.1.71.1980.
- Fekete, É., Fekete, E., Irinyi, L., Karaffa, L., Árnayasi, M., Asadollahi, M., et al. (2012). Genetic diversity of a *Botrytis cinerea* cryptic species complex in Hungary. *Microbiol. Res.* 167, 283–291. doi:10.1016/j.micres.2011.10.006.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17, 368–376. doi:10.1007/BF01734359.
- Field, K., Olsen, G., Lane, D., Giovannoni, S., Ghiselin, M., Raff, E., et al. (1988). Molecular phylogeny of the animal kingdom. *Science (80- )*. 239, 748–753. doi:10.1126/science.3277277.
- Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Trans. R. Soc. Edinburgh* 52, 399–433. doi:10.1017/S0080456800012163.
- Fisher, R. A. (1930). *The genetical theory of natural selection*. Available at: <https://academic.oup.com/sf/article-lookup/doi/10.2307/2570330>.
- Floriano, A. M., Castelli, M., Krenek, S., Berendonk, T. U., Bazzocchi, C., Petroni, G., et al. (2018). The Genome Sequence of “*Candidatus Fokinia solitaria*”: Insights on Reductive Evolution in *Rickettsiales*. *Genome Biol. Evol.* 10, 1120–1126. doi:10.1093/gbe/evy072.

- Fokin, S. I., Schweikert, M., Brümmer, F., and Görtz, H.-D. (2005). *Spirostomum* spp. (Ciliophora, Protista), a suitable system for endocytobiosis research. *Protoplasma* 225, 93–102. doi:10.1007/s00709-004-0078-y.
- Folmer, O., Black, M., Hoeh, W., Lutz, R., and Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* 3, 294–9.
- Foster, J., Ganatra, M., Kamal, I., Ware, J., Makarova, K., Ivanova, N., et al. (2005). The *Wolbachia* Genome of *Brugia malayi*: Endosymbiont Evolution within a Human Pathogenic Nematode. *PLoS Biol.* 3, e121. doi:10.1371/journal.pbio.0030121.
- Freitas, A., Pereira, A. da C., Brazdil, P., Costa-Pereira, A., and Brazdil, P. (2007). “Cost-Sensitive Decision Trees Applied to Medical Data,” in *Data Warehousing and Knowledge Discovery Lecture Notes in Computer Science.*, eds. I.-Y. Song, J. Eder, and T. M. Nguyen (Berlin, Heidelberg, Heidelberg: Springer Berlin Heidelberg), 303–312. doi:10.1007/978-3-540-74553-2\_28.
- Frost, D. R., Grant, T., Faivovich, J., Bain, R. H., Haas, A., Haddad, C. F. B., et al. (2006). The Amphibian Tree Of Life. *Bull. Am. Museum Nat. Hist.* 2006, 1–291. doi:10.1206/0003-0090(2006)297[0001:tatol]2.0.co;2.
- Fujishima, M., and Kodama, Y. (2012). Endosymbionts in *Paramecium*. *Eur. J. Protistol.* 48, 124–137. doi:10.1016/j.ejop.2011.10.002.
- Gagic, D., Ciric, M., Wen, W. X., Ng, F., and Rakonjac, J. (2016). Exploring the Secretomes of Microbes and Microbial Communities Using Filamentous Phage Display. *Front. Microbiol.* 7. doi:10.3389/fmicb.2016.00429.
- Gao, B., and Gupta, R. S. (2012). Microbial systematics in the post-genomics era. *Antonie Van Leeuwenhoek* 101, 45–54. doi:10.1007/s10482-011-9663-1.
- Garner, T. W. J., Gautschi, B., Rothlisberger, S., and Reyer, H.-U. (2000). A set of CA repeat microsatellite markers derived from the pool frog, *Rana lessonae*. *Mol. Ecol.* 9, 2173–2175. doi:10.1046/j.1365-294X.2000.105311.x.
- Gentles, A. J. (2001). Genome-Scale Compositional Comparisons in Eukaryotes. *Genome Res.* 11, 540–546. doi:10.1101/gr.163101.
- Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., et al. (2005). Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* 3, 733–739. doi:10.1038/nrmicro1236.
- Ghiselin, M. T. (1974). A Radical Solution to the Species Problem. *Syst. Biol.* 23, 536–544. doi:10.1093/sysbio/23.4.536.
- Gil, R., Sabater-Munoz, B., Latorre, A., Silva, F. J., and Moya, A. (2002). Extreme genome reduction in *Buchnera* spp.: Toward the minimal genome needed for symbiotic life. *Proc. Natl. Acad. Sci.* 99, 4454–4458. doi:10.1073/pnas.062067299.
- Gould, S. J. (1979). Quahog is a quahog. *Nat. Hist.* 88, 18.
- Graf, J. D., Karch, F., and Moreillon, M. C. (1977). Biochemical variation in the *Rana esculenta* complex: A new hybrid form related to *Rana perezi* and *Rana ridibunda*.

*Experientia* 33, 1582–1584. doi:10.1007/BF01934010.

- Gregor, I., Dröge, J., Schirmer, M., Quince, C., and McHardy, A. C. (2016). PhyloPythiaS+ : a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ* 4, e1603. doi:10.7717/peerj.1603.
- Gruber-Vodicka, H. R., Leisch, N., Kleiner, M., Hinzke, T., Liebeke, M., McFall-Ngai, M., et al. (2019). Two intracellular and cell type-specific bacterial symbionts in the placozoan *Trichoplax*H2. *Nat. Microbiol.* 4, 1465–1474. doi:10.1038/s41564-019-0475-9.
- Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi:10.1093/bioinformatics/btt086.
- Hahsler, M., Piekenbrock, M., and Doran, D. (2019). dbscan: Fast Density-based Clustering with R. *J. Stat. Softw.* 25, 409–416. Available at: <https://cran.r-project.org/web/packages/dbscan/vignettes/dbscan.pdf>.
- Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science (80- )*. 28, 49–50. doi:10.1126/science.28.706.49.
- Harrison, D. N., Dorsey, C. H., and Finley, H. E. (1976). Studies on a Macronuclear Endosymbiont of *Spirostomum ambiguum*. I. Isolation of the Microorganism from the Macronucleus. *Trans. Am. Microsc. Soc.* 95, 560. doi:10.2307/3225378.
- Hebert, P. D. N., Cywinska, A., Ball, S. L., and DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proc. R. Soc. London. Ser. B Biol. Sci.* 270, 313–321. doi:10.1098/rspb.2002.2218.
- Hennig, W. (1966). *Phylogenetic Systematics*. University of Illinois Press Available at: <http://www.annualreviews.org/doi/10.1146/annurev.en.10.010165.000525>.
- Hertig, M. (1936). The *Rickettsia*, *Wolbachia pipientis* (gen. et sp.n.) and Associated Inclusions of the Mosquito, *Culex pipiens*. *Parasitology* 28, 453–486. doi:10.1017/S0031182000022666.
- Hertig, M., and Wolbach, S. B. (1924). Studies on *Rickettsia*-Like Micro-Organisms in Insects. *J. Med. Res.* 44, 329-374.7. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19972605>.
- Holsbeek, G., and Jooris, R. (2010). Potential impact of genome exclusion by alien species in the hybridogenetic water frogs (*Pelophylax esculentus* complex). *Biol. Invasions* 12, 1–13. doi:10.1007/s10530-009-9427-2.
- Hossfeld, U., and Levit, G. S. (2016). “Tree of life” took root 150 years ago. *Nature* 540, 38–38. doi:10.1038/540038a.
- Hotz, H., Mancino, G., Bucciinnocenti, S., Ragghianti, M., Berger, L., and Uzzell, T. (1985). *Rana ridibunda* varies geographically in inducing clonal gametogenesis in interspecies hybrids. *J. Exp. Zool.* 236, 199–210. doi:10.1002/jez.1402360210.
- Hotz, H., and Uzzell, T. (1983). Interspecific hybrids of *Rana ridibunda* without germ line exclusion of a parental genome. *Experientia* 39, 538–540. doi:10.1007/BF01965196.

- Husnik, F., and McCutcheon, J. P. (2016). Repeated replacement of an intrabacterial symbiont in the tripartite nested mealybug symbiosis. *Proc. Natl. Acad. Sci.* 113, E5416–E5424. doi:10.1073/pnas.1603910113.
- Huxley, T. H. (1882). *Lectures on evolution*. Cambridge: Cambridge University Press.
- Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11, 119. doi:10.1186/1471-2105-11-119.
- Ishida, K., Sekizuka, T., Hayashida, K., Matsuo, J., Takeuchi, F., Kuroda, M., et al. (2014). Amoebal Endosymbiont *Neochlamydia* Genome Sequence Illuminates the Bacterial Role in the Defense of the Host Amoebae against *Legionella pneumophila*. *PLoS One* 9, e95166. doi:10.1371/journal.pone.0095166.
- Janda, J. M., and Abbott, S. L. (2007). 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. *J. Clin. Microbiol.* 45, 2761–2764. doi:10.1128/JCM.01228-07.
- Jeyaprakash, A., and Hoy, M. A. (2000). Long PCR improves *Wolbachia* DNA amplification: wsp sequences found in 76% of sixty-three arthropod species. *Insect Mol. Biol.* 9, 393–405. doi:10.1046/j.1365-2583.2000.00203.x.
- Kaiser, W., Huguet, E., Casas, J., Commin, C., and Giron, D. (2010). Plant green-island phenotype induced by leaf-miners is mediated by bacterial symbionts. *Proc. R. Soc. B Biol. Sci.* 277, 2311–2319. doi:10.1098/rspb.2010.0214.
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462. doi:10.1093/nar/gkv1070.
- Kang, D., Li, F., Kirton, E. S., Thomas, A., Egan, R. S., An, H., et al. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ Prepr.* 7, e27522v1.
- Karlin, S. (1998). Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr. Opin. Microbiol.* 1, 598–610. doi:10.1016/S1369-5274(98)80095-7.
- Karlin, S., and Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* 11, 283–290. doi:10.1016/S0168-9525(00)89076-9.
- Karlin, S., Campbell, A. M., and Mrázek, J. (1998). Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* 32, 185–225. doi:10.1146/annurev.genet.32.1.185.
- Karlin, S., and Ladunga, I. (1994). Comparisons of eukaryotic genomic sequences. *Proc. Natl. Acad. Sci.* 91, 12832–12836. doi:10.1073/pnas.91.26.12832.
- Karlin, S., Ladunga, I., and Blaisdell, B. E. (1994). Heterogeneity of genomes: measures and values. *Proc. Natl. Acad. Sci.* 91, 12837–12841. doi:10.1073/pnas.91.26.12837.
- Kimura, M. (1968). Evolutionary Rate at the Molecular Level. *Nature* 217, 624–626. doi:10.1038/217624a0.
- Knowlton, N., and Weigt, L. A. (1998). New dates and new rates for divergence across the

- Isthmus of Panama. *Proc. R. Soc. London. Ser. B Biol. Sci.* 265, 2257–2263. doi:10.1098/rspb.1998.0568.
- Kostygov, A. Y., Dobáková, E., Grybchuk-Ieremenko, A., Váhala, D., Maslov, D. A., Votýpka, J., et al. (2016). Novel *Trypanosomatid*-Bacterium Association: Evolution of Endosymbiosis in Action. *MBio* 7, e01985-15. doi:10.1128/mBio.01985-15.
- Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M., and Blaxter, M. (2013). Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front. Genet.* 4, 1–12. doi:10.3389/fgene.2013.00237.
- Laczny, C. C., Kiefer, C., Galata, V., Fehlmann, T., Backes, C., and Keller, A. (2017). BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation. *Nucleic Acids Res.* 45, W171–W179. doi:10.1093/nar/gkx348.
- Lanza, B. (1962). On the Introduction of *Rana ridibunda* Pallas and *Rana catesbeiana* Shaw in Italy. *Copeia* 1962, 642. doi:10.2307/1441194.
- Lanza, B., Andreone, F., Bologna, M. A., Corti, C., and Razzetti, E. (2007). *Fauna d'Italia, vol. XLII, Amphibia*. Calderini, Bologna, XI.
- Laporte, L. F. (1994). Simpson on species. *J. Hist. Biol.* 27, 141–159. doi:10.1007/BF01058629.
- Lewin, R. A. (1981). Three species concepts. *Taxon* 30, 609–613. doi:10.2307/1219942.
- Li, S., Pearl, D. K., and Doss, H. (2000). Phylogenetic Tree Construction Using Markov Chain Monte Carlo. *J. Am. Stat. Assoc.* 95, 493–508. doi:10.1080/01621459.2000.10474227.
- Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. *R News* 2, 18–22. Available at: <https://cran.r-project.org/doc/Rnews/>.
- Linnaeus, C. von (1753). *Species plantarum, Vol II*.
- Linneé, C. von (1735). *Caroli Linnaei, Sveci, Doctoris Medicinae systema naturae, sive, Regna tria naturae systematice proposita per classes, ordines, genera, & species*. Lugduni Batavorum [Leiden, the Netherlands]: Apud Theodorum Haak: Ex Typographia Joannis Wilhelmi de Groot, doi:10.5962/bhl.title.877.
- Liu, J., Jiang, J., Song, S., Tornabene, L., Chabarria, R., Naylor, G. J. P., et al. (2017). Multilocus DNA barcoding – Species Identification with Multilocus Data. *Sci. Rep.* 7, 16601. doi:10.1038/s41598-017-16920-2.
- Lynn, D. (2010). “Subphylum 1. POSTCILIODESMATOPHORA: Class 2. HETEROTRICHEA – Once Close to the Top,” in *The Ciliated Protozoa* (Dordrecht: Springer Netherlands), 129–139. doi:10.1007/978-1-4020-8239-9\_6.
- Magoc, T., and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957–2963. doi:10.1093/bioinformatics/btr507.
- Matsuo, J., Oguri, S., Nakamura, S., Hanawa, T., Fukumoto, T., Hayashi, Y., et al. (2010). Ciliates rapidly enhance the frequency of conjugation between *Escherichia coli* strains

- through bacterial accumulation in vesicles. *Res. Microbiol.* doi:10.1016/j.resmic.2010.07.004.
- Mau, B., Newton, M. A., and Larget, B. (1999). Bayesian Phylogenetic Inference via Markov Chain Monte Carlo Methods. *Biometrics* 55, 1–12. doi:10.1111/j.0006-341X.1999.00001.x.
- Maxam, A. M., and Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* 74, 560–4. doi:10.1073/pnas.74.2.560.
- Mayden, Ri. (1997). “A Hierarchy of Species Concepts : the Denouement in the Saga of the Species Problem,” in *Species: The Units of Biodiversity*, 381–423.
- Mayr, E. (1943). Systematics and the Origin of Species. *Bird-Banding* 14, 89. doi:10.2307/4509787.
- Mayr, E. (1982). *The growth of biological thought: Diversity, evolution, and inheritance*. Cambridge: Harvard University Press.
- McCutcheon, J. P., and Moran, N. A. (2012). Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* 10, 13–26. doi:10.1038/nrmicro2670.
- McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* 3, 861. doi:10.21105/joss.00861.
- Meyer, A., Todt, C., Mikkelsen, N. T., and Lieb, B. (2010). Fast evolving 18S rRNA sequences from *Solenogastres* (Mollusca) resist standard PCR amplification and give new insights into mollusk substitution rate heterogeneity. *BMC Evol. Biol.* 10, 70. doi:10.1186/1471-2148-10-70.
- Meyer, D. (2014). Support vector machines, The interface to libsvm. *Support Vector Mach. Interface to libsvm Packag.* e1071.
- Mignard, S., and Flandrois, J. P. (2006). 16S rRNA sequencing in routine bacterial identification: A 30-month experiment. *J. Microbiol. Methods* 67, 574–581. doi:10.1016/j.mimet.2006.05.009.
- Mira, A., Klasson, L., and Andersson, S. G. E. (2002). Microbial genome evolution: Sources of variability. *Curr. Opin. Microbiol.* doi:10.1016/S1369-5274(02)00358-2.
- Moran, N. A., and Wernegreen, J. J. (2000). Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends Ecol. Evol.* 15, 321–326. doi:10.1016/S0169-5347(00)01902-9.
- Mullis, K. B., Erlich, H. A., Arnheim, N., Horn, G. T., Saiki, R. K., and Scharf, S. J. (1987). Process for amplifying, detecting, and/or-cloning nucleic acid sequences. *Biotechnol. Adv.* 5, 313. doi:10.1016/0734-9750(87)90384-3.
- Nikoh, N., Tsuchida, T., Maeda, T., Yamaguchi, K., Shigenobu, S., Koga, R., et al. (2018). Genomic Insight into Symbiosis-Induced Insect Color Change by a Facultative Bacterial Endosymbiont, “*Candidatus* Rickettsiella viridis.” *MBio* 9, e00890-18. doi:10.1128/mBio.00890-18.
- Pace, N. R., Stahl, D. A., Lane, D. J., and Olsen, G. J. (1986). “The Analysis of Natural

- Microbial Populations by Ribosomal RNA Sequences,” in *Advances in microbial ecology*, 1–55. doi:10.1007/978-1-4757-0611-6\_1.
- Pagano, A., Joly, P., and Hotz, H. (1997). Taxon composition and genetic variation of water frogs in the Mid-Rhone floodplain. *Comptes Rendus l'Académie des Sci. - Ser. III - Sci. la Vie* 320, 759–766. doi:10.1016/S0764-4469(97)84825-1.
- Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* 176, 649–662.e20. doi:10.1016/j.cell.2019.01.001.
- Plenet, Pagano, Joly, and Fouillet (2000a). Variation of plastic responses to oxygen availability within the hybridogenetic *Rana esculenta* complex. *J. Evol. Biol.* 13, 20–28. doi:10.1046/j.1420-9101.2000.00141.x.
- Plenet, S., Hervant, F., and Joly, P. (2000b). Ecology of the Hybridogenetic *Rana esculenta* Complex: Differential Oxygen Requirements of Tadpoles. *Evol. Ecol.* 14, 13–23. doi:10.1023/A:1011056703016.
- Plotner, J., Uzzell, T. M., Beerli, P., Spolsky, C., Ohst, T., Litvinchuk, S. N., et al. (2008). Widespread unidirectional transfer of mitochondrial DNA: a case in western Palaeartic water frogs. *J. Evol. Biol.* 21, 668–681. doi:10.1111/j.1420-9101.2008.01527.x.
- Prakash, O., Jangid, K., and Shouche, Y. S. (2013). Carl Woese: from Biophysics to Evolutionary Microbiology. *Indian J. Microbiol.* 53, 247–252. doi:10.1007/s12088-013-0401-4.
- Preer, L. B. (1969). Alpha, an Infectious Macronuclear Symbiont of *Paramecium aurelia*. *J. Protozool.* 16, 570–578. doi:10.1111/j.1550-7408.1969.tb02315.x.
- Pringle, E. G. (2016). Orienting the Interaction Compass: Resource Availability as a Major Driver of Context Dependence. *PLOS Biol.* 14, e2000891. doi:10.1371/journal.pbio.2000891.
- Pritchard, J. K., and Wen, W. (2002). Documentation for structure software: Version 2. *In Pract.*
- Quinlan, A. R., and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi:10.1093/bioinformatics/btq033.
- Rannala, B., and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J. Mol. Evol.* 43, 304–311. doi:10.1007/BF02338839.
- Raymond, M., and Rousset, F. (1995). GENEPOP (Version 1.2): Population Genetics Software for Exact Tests and Ecumenicism. *J. Hered.* 86, 248–249. doi:10.1093/oxfordjournals.jhered.a111573.
- Richard, G.-F., Kerrest, A., and Dujon, B. (2008). Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. *Microbiol. Mol. Biol. Rev.* 72, 686–727. doi:10.1128/MMBR.00011-08.

- Rokas, A., Williams, B. L., King, N., and Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798–804. doi:10.1038/nature02053.
- Rosselló-Mora, R. (2001). The species concept for prokaryotes. *FEMS Microbiol. Rev.* 25, 39–67. doi:10.1016/S0168-6445(00)00040-1.
- Rousset, F. (2008). genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol. Ecol. Resour.* 8, 103–106. doi:10.1111/j.1471-8286.2007.01931.x.
- Ruse, M. (1969). Definitions of Species in Biology. *Br. J. Philos. Sci.* 20, 97–119. doi:10.1093/bjps/20.2.97.
- Sagan, L. (1967). On the origin of mitosing cells. *J. Theor. Biol.* 14, 225-IN6. doi:10.1016/0022-5193(67)90079-3.
- Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* doi:10.1093/oxfordjournals.molbev.a040454.
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–7. doi:10.1073/pnas.74.12.5463.
- Schmidt, S. L., Foissner, W., Schlegel, M., and Bernhard, D. (2007). Molecular Phylogeny of the *Heterotrichea* (Ciliophora, Postciliodesmatophora) Based on Small Subunit rRNA Gene Sequences. *J. Eukaryot. Microbiol.* 54, 358–363. doi:10.1111/j.1550-7408.2007.00269.x.
- Schultz, R. J. (1969). Hybridization, Unisexuality, and Polyploidy in the Teleost *Poeciliopsis* (Poeciliidae) and Other Vertebrates. *Am. Nat.* 103, 605–619. doi:10.1086/282629.
- Schwartz, R., and Dayhoff, M. (1978). Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science (80- )*. 199, 395–403. doi:10.1126/science.202030.
- Seah, B. K. B., and Gruber-Vodicka, H. R. (2015). Gbtools: Interactive Visualization of Metagenome Bins in R. *Front. Microbiol.* 6, 1451. doi:10.3389/fmicb.2015.01451.
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi:10.1093/bioinformatics/btu153.
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi:10.1093/bioinformatics/btv351.
- Sinsch, U., and Schneider, H. (2009). Bioacoustic assessment of the taxonomic status of pool frog populations (*Rana lessonae*) with reference to a topotypical population. *J. Zool. Syst. Evol. Res.* 34, 63–73. doi:10.1111/j.1439-0469.1996.tb00811.x.
- Small, S. T., Reimer, L. J., Tisch, D. J., King, C. L., Christensen, B. M., Siba, P. M., et al. (2016). Population genomics of the filarial nematode parasite *Wuchereria bancrofti* from mosquitoes. *Mol. Ecol.* 25, 1465–1477. doi:10.1111/mec.13574.
- Smith, J. M. (1989). *Evolutionary genetics*. Oxford University Press.

- Smith, K. L., Harmon, L. J., Shoo, L. P., and Melville, J. (2011). Evidence of constrained phenotypic evolution in a cryptic species complex of agamid lizards. *Evolution (N. Y.)* 65, 976–992. doi:10.1111/j.1558-5646.2010.01211.x.
- Sokal, R. R., and Crovello, T. J. (1970). The Biological Species Concept: A Critical Evaluation. *Am. Nat.* 104, 127–153. doi:10.1086/282646.
- Spolsky, C., and Uzzell, T. (1984). Natural interspecies transfer of mitochondrial DNA in amphibians. *Proc. Natl. Acad. Sci.* 81, 5802–5805. doi:10.1073/pnas.81.18.5802.
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033.
- Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi:10.1093/molbev/mst197.
- Taylor, M. ., and Hoerauf, A. (1999). *Wolbachia* Bacteria of Filarial Nematodes. *Parasitol. Today* 15, 437–442. doi:10.1016/S0169-4758(99)01533-1.
- Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glöckner, F. O. (2004). TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* 5, 163. doi:10.1186/1471-2105-5-163.
- Teixeira, L., Ferreira, Á., and Ashburner, M. (2008). The Bacterial Symbiont *Wolbachia* Induces Resistance to RNA Viral Infections in *Drosophila melanogaster*. *PLoS Biol.* 6, e1000002. doi:10.1371/journal.pbio.1000002.
- Templeton, A. R. (1989). “The meaning of species and speciation: a genetic perspective,” in *Speciation and its Consequences*, 159–183.
- Templeton, A. R., Crandall, K. A., and Sing, C. F. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* 132, 619–633.
- Teueman, A. E. (1924). The Species-Concept in Palaeontology. *Geol. Mag.* 61, 355–360. doi:10.1017/S001675680008660X.
- Tillyard M. A., R. J. (1921). A new classification of the order perlaria. *Can. Entomol.* 53, 35–43. doi:10.4039/Ent5335-2.
- Tommassen, J., and Arenas, J. (2017). Biological Functions of the Secretome of *Neisseria meningitidis*. *Front. Cell. Infect. Microbiol.* 7. doi:10.3389/fcimb.2017.00256.
- Trigo, T. C., Schneider, A., de Oliveira, T. G., Lehugeur, L. M., Silveira, L., Freitas, T. R. O., et al. (2013). Molecular Data Reveal Complex Hybridization and a Cryptic Species of Neotropical Wild Cat. *Curr. Biol.* 23, 2528–2533. doi:10.1016/j.cub.2013.10.046.
- Tunner, H. G., and Heppich, S. (1981). Premeiotic genome exclusion during oogenesis in the common edible frog, *Rana esculenta*. *Naturwissenschaften* 68, 207–208. doi:10.1007/BF01047207.
- Uzzell, T., Günther, R., and Berger, L. (1976). *Rana ridibunda* and *Rana esculenta*: A Leaky

- Hybridogenetic System (Amphibia Salientia). *Proc. Acad. Nat. Sci. Philadelphia* 128, 147–171. Available at: <http://www.jstor.org/stable/4064723>.
- Uzzell, T., and Hotz, H. (1979). Electrophoretic and Morphological Evidence for Two Forms of Green Frogs (*Rana esculenta* Complex) in Peninsular Italy (Amphibia, Salientia). *Mitteilungen aus dem Museum für Naturkd. Berlin. Zool. Museum und Inst. für Spez. Zool.* 55, 13–27. doi:10.1002/mmnz.4830550105.
- Uzzell, T., Hotz, H., and Berger, L. (1980). Genome exclusion in gametogenesis by an interspecific *Rana* hybrid: Evidence from electrophoresis of individual oocytes. *J. Exp. Zool.* 214, 251–259. doi:10.1002/jez.1402140303.
- van Ham, R. C. H. J., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U., et al. (2003). Reductive genome evolution in *Buchnera aphidicola*. *Proc. Natl. Acad. Sci.* 100, 581–586. doi:10.1073/pnas.0235981100.
- Van Oosterhout, C., Hutchinson, W. F., Willis, D. P. M., and Shipley, P. (2004). micro-checker: software for identifying and correcting genotyping errors in microsatellite data. *Mol. Ecol. Notes* 4, 535–538. doi:10.1111/j.1471-8286.2004.00684.x.
- Vinogradov, A. E., Borkin, L. J., Gunther, R., and Rosanov, J. M. (2008). Two germ cell lineages with genomes of different species in one and the same animal. *Hereditas* 114, 245–251. doi:10.1111/j.1601-5223.1991.tb00331.x.
- Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi:10.1128/AEM.00062-07.
- Wang, Y., and Chandler, C. (2016). Candidate pathogenicity islands in the genome of ‘*Candidatus* Rickettsiella isopodorum’, an intracellular bacterium infecting terrestrial isopod crustaceans. *PeerJ* 4, e2806. doi:10.7717/peerj.2806.
- Watson, J. D., and Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* 171, 737–738. doi:10.1038/171737a0.
- Weaver, W. (1970). Molecular Biology: Origin of the Term. *Science (80- )*. 170, 581–582. doi:10.1126/science.170.3958.581-a.
- Wenseleers, T., Ito, F., Van Borm, S., Huybrechts, R., Volckaert, F., and Billen, J. (1998). Widespread occurrence of the microorganism *Wolbachia* in ants. *Proc. R. Soc. London. Ser. B Biol. Sci.* 265, 1447–1452. doi:10.1098/rspb.1998.0456.
- Wernegreen, J. J. (2015). Endosymbiont evolution: predictions from theory and surprises from genomes. *Ann. N. Y. Acad. Sci.* 1360, 16–35. doi:10.1111/nyas.12740.
- Werren, J. H., and Windsor, D. M. (2000). *Wolbachia* infection frequencies in insects: evidence of a global equilibrium? *Proc. R. Soc. London. Ser. B Biol. Sci.* 267, 1277–1285. doi:10.1098/rspb.2000.1139.
- Werren, J. H., Zhang, W., and Guo, L. R. (1995). Evolution and phylogeny of *Wolbachia*: Reproductive parasites of arthropods. *Proc. R. Soc. B Biol. Sci.* doi:10.1098/rspb.1995.0117.
- Woese, C. R., and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The

- primary kingdoms. *Proc. Natl. Acad. Sci.* 74, 5088–5090. doi:10.1073/pnas.74.11.5088.
- Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* 87, 4576–9. doi:10.1073/pnas.87.12.4576.
- Woo, P. C. Y., Ng, K. H. L., Lau, S. K. P., Yip, K. -t., Fung, A. M. Y., Leung, K. -w., et al. (2003). Usefulness of the MicroSeq 500 16S Ribosomal DNA-Based Bacterial Identification System for Identification of Clinically Significant Bacterial Isolates with Ambiguous Biochemical Profiles. *J. Clin. Microbiol.* 41, 1996–2001. doi:10.1128/JCM.41.5.1996-2001.2003.
- Wu, D., Daugherty, S. C., Van Aken, S. E., Pai, G. H., Watkins, K. L., Khouri, H., et al. (2006). Metabolic Complementarity and Genomics of the Dual Bacterial Symbiosis of Sharpshooters. *PLoS Biol.* 4, e188. doi:10.1371/journal.pbio.0040188.
- Wu, Y.-W., Simmons, B. A., and Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607. doi:10.1093/bioinformatics/btv638.
- Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A., and Singer, S. W. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2, 26. doi:10.1186/2049-2618-2-26.
- Xu, J., De Barro, P. J., and Liu, S. S. (2010). Reproductive incompatibility among genetic groups of *Bemisia tabaci* supports the proposition that the whitefly is a cryptic species complex. *Bull. Entomol. Res.* 100, 359–366. doi:10.1017/S0007485310000015.
- Yoon, S.-H., Ha, S., Lim, J., Kwon, S., and Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie Van Leeuwenhoek* 110, 1281–1286. doi:10.1007/s10482-017-0844-4.
- Zachos, F. E. (2016). *Species Concepts in Biology*. Cham: Springer International Publishing doi:10.1007/978-3-319-44966-1.
- Zeisset, I., Rowe, G., and Beebee, T. J. C. (2000). Polymerase chain reaction primers for microsatellite loci in the north European water frogs *Rana ridibunda* and *R. lessonae*. *Mol. Ecol.* 9, 1173–1174. doi:10.1046/j.1365-294x.2000.00954-2.x.
- Zuckerlandl, E., and Pauling, L. (1962). “Molecular disease, evolution, and genetic heterogeneity,” in *Horizons in biochemistry* (Academic Press), 189–225. doi:10.1021/ed053pA326.2.

---

## APPENDIX

# Appendix 1 – Performance calculation

During the development of a classification method it is important to calculate and evaluate performances. It is possible to achieve this task when the developed method is a classifier that requires testing on a set for which the true categories are known. Also, as today there are many bioinformatic implementations for performing widely different analyses, when presenting a new tool, it is important to provide a comparison with other available methods to understand how the new and the old tools perform on the same datasets.

Consider the case of a dataset  $D$  composed by elements associated to a set of numerical or categorical features and classified in two classes,  $a$  and  $b$ . The task is to measure how good a general classifier is in assigning objects to the  $a$  and  $b$  classes by exploiting the features associated to data points. We thus define True Positives (TP) the elements that were in class  $a$  ( $b$ ) and that the classifier assigns to class  $a$  ( $b$ ); True Negatives (TN), as the elements that the classifier tells do not belong to class  $a$  ( $b$ ) and indeed they don't; False Positives (FP) as the elements that it assigns to class  $a$  ( $b$ ) but they don't belong to it; and False Negatives (FN) the elements that belong to class  $a$  ( $b$ ) but the classifier assigned outside of  $a$  ( $b$ ).

To compare the performance of several classifiers, the above quantities can be used to summarize the different kinds of errors that can be done and therefore provide a compact representation of a classifier's behaviour. Additionally, the above quantities have been combined to obtain more meaningful measures.

Sensitivity, also called recall, is the proportion of TP over the total number of positives:

$$\frac{TP}{(TP + FN)}$$

Sensitivity quantifies the avoidance of FN. If the sensitivity is low, it means that the FN outperforms the TP and so the classification method used determine the wrong classification most of the times.

The accuracy is a measure of how well the method classify the elements of the dataset and is calculated as:

$$\frac{(TP + TN)}{(TP + FP + FN + TN)}$$

It is a measure of how much the classification is close to the reality.

Precision is a measure associated to the accuracy of the predictions. It is calculated as:

$$\frac{TP}{(TP + FP)}$$

The F1 score is a way to combine in an even more compact form two of the quantities introduced above. It is calculated as the harmonic mean between sensitivity and precision:

$$2 \cdot \frac{(\textit{Precision} \cdot \textit{Sensitivity})}{(\textit{Precision} + \textit{Sensitivity})}$$

# Appendix 2 – Machine Learning

## Classification Models

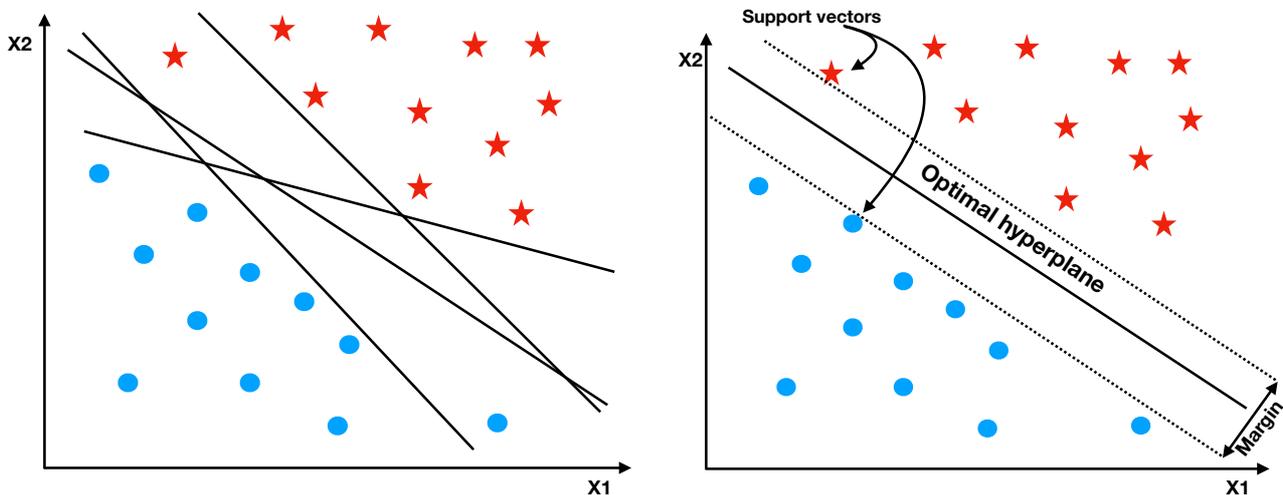
Machine learning (ML) is a family of algorithms that learns how to perform a defined task without having explicitly provided instructions. In particular, ML classification algorithms are used to make predictions on the data and thus classify them. There are two types of ML algorithms: supervised or unsupervised. Supervised ML uses data for which the classes to be predicted are known such that it can be trained on these. The model is then constructed to be able to correctly predict the already known classes. The so-obtained model can then be used to classify unlabelled objects, a situation that is common when the classification of the objects can be done precisely by using expensive (or time consuming) technology and one wants to check if the classification is feasible with a different kind of data which is cheaper (faster) to obtain. It is for instance the case of the prediction of antibiotic resistances in pathogens. The whole process takes weeks in *Mycobacterium tuberculosis* but by exploiting genome sequences it is possible to provide the same result in a matter of days; highly appealing.

Unsupervised algorithms are instead able to find clusters and patterns in the data, and they can use them as the labels used in supervised algorithms to predict new observations. ML classification algorithms are widely used to classify DNA or protein sequences: they can be used to deconvolve data obtained by the sequencing of mixed samples, as in metagenomic studies; or during comparison to a database to reconstruct the taxonomic affiliation of an organism.

Supervised ML algorithms frequently used to classify sequences are Support Vector Machine (SVM) and Random Forest (RF). As supervised algorithms, they need to have already classified objects in the dataset to train the model. These objects need to be representative of the variability of the dataset and their labels have to be highly reliable to construct a model with good classification performances. This labelled dataset is randomly divided into a training set and a test set; the former is used for obtaining the values of the parameters of the model, the latter to check how good a certain parameterization is, in a fairly standard cross validation framework. Most of the labelled objects go in the training set, but there is not a consensus (e.g. can be 66% train and 33% test, or 90% and 10%, respectively).

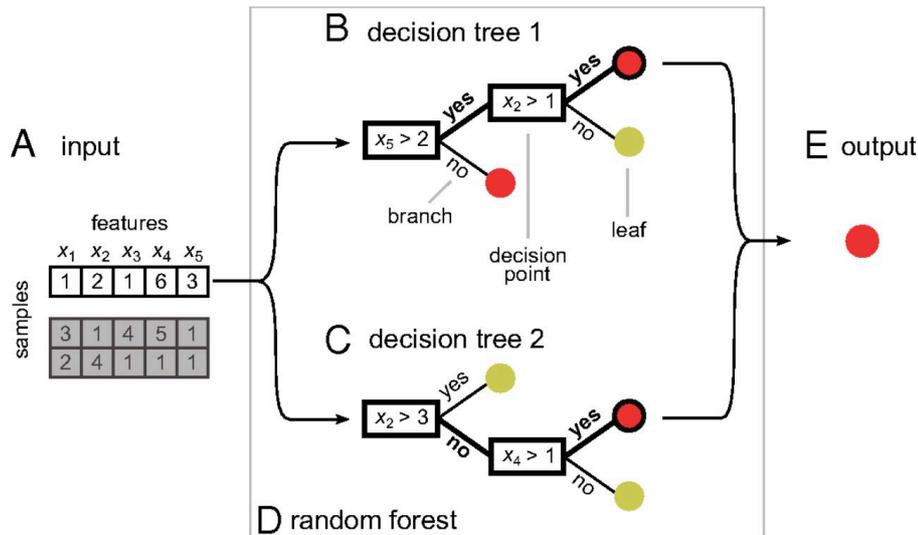
Support Vector Machine (SVM) is a machine learning algorithm used in classification and regression problems (Cortes and Vapnik, 1995; Meyer, 2014). Given a set of labelled objects described by  $n$  variables, SVM searches for the optimal hyperplane that separates the classes (Appendix Figure 1). In addition, among all the possible hyperplanes found, SVM chooses the one with maximal margin between the closest points to the plane, called

Support Vectors. SVM perform a linear separation of the space for the classification: each area delimited by the intersection of the planes is assigned to a label. However, if it is not possible to separate labels by using linear planes, the algorithm transforms the space using a kernel to reconduct it to a problem with a linear solution.



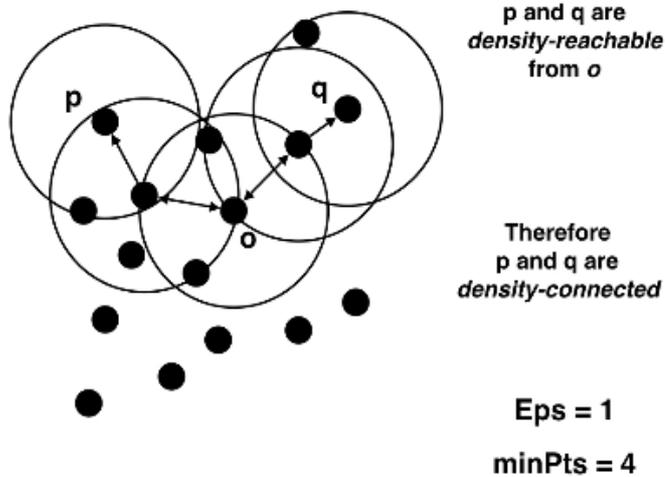
Appendix Figure 1 – The charts reassume how SVM works. In this example, we have to separate the objects of a dataset composed by blue points and red stars. Both are described by two variables:  $X_1$  and  $X_2$ . The algorithm searches for all possible planes to separate the two classes (left chart). Some degree of mistakes can also be allowed. Among all the found planes, SVM chooses the one with the maximum margin (right chart). The margin is defined as the distance between the closest point to the hyperplane among all in the dataset (called support vectors). As is possible to see, there are various plane found in the left chart; however, the algorithm chooses the one in the right chart as it makes no mistakes and have the higher marginal distance between the support vectors.

Random Forest (RF) is a ML algorithm based on the classification tree (Breiman, 2001; Cutler et al., 2007). A classification tree is formed by decision points that fork in branches. The decision points are associated to rules that are used to decide if a sample has to be assigned to a branch or another. At the end, branches end in leaves, each assigned to a class present in the dataset used for the construction of the tree. Random Forest grows a user-defined number of trees. During the construction of the model, for each tree the dataset is randomly divided into a training and a test set. At each split of each tree, the algorithm randomly samples a subset of the features and then test them to find a rule for the decision point. These trees are a modification of the classical classification ones and are called decision tree. At the end, the model is composed by all the tree constructed as described (see Appendix Figure 2). The prediction of the classification of the data is done by considering all the votes of all the trees in the forest.



Appendix Figure 2 – Random Forest classification algorithm from (Denisko and Hoffman, 2018). In the figure: A is the input dataset, where each object is characterized by 5 features; B and C are the two trees in the random forest D and constructed using at each split a rule build on one of the features randomly sampled; E is the output. A decision tree is composed by decision point that fork in branches. The decision points have rules on which decide to assign an object to one branch or the other. Branches terminates in the final leaves. The decision to which class belong each object of the dataset is done by combining all the votes of the trees of the forest.

The most frequently used unsupervised ML algorithm is the clustering analysis. This class of methods analyse the data to find similarities and dissimilarities. There exists various different type of clustering algorithms, and, due to the extent of the argument, they will not be discussed in detail. Briefly, the most widely used methods are k-means and hierarchical clustering. However, these algorithms work well in dataset that form circular (or spherical, depending on the number of dimensions) clusters, but the sequence analysis data might form groups of irregular shapes. In this condition, it is preferable to use an algorithm based on the proximity of the data. One affordable method is DBscan, which is a density-based clustering algorithm (Ester et al., 1996). As described in Appendix Figure 3, it identifies cores point as the ones that have at least a *minPts* points within a distance *Eps*; reachable points as points that are within *Eps* from a core points; outliers that within *Eps* do not have a core nor a reachable point. After defining all the points in the dataset, DBscan clusters together all the core and reachable points connected. This algorithm is based on spatial proximity and density, thus is able to identify clusters of irregular shape. The parameters *minPts* and *Eps* can simply be tuned on the dataset, thus allowing a simple customization and automatization of the method. Also, it allows easily to identify outliers that can be thus excluded from further analysis.



Appendix Figure 3 – The figure reassumes how DBscan classify the data. Eps is the maximum spatial distance considered. The circle around the data in the figure are designed on this value. minPts is the minimum points within Eps that have to be present to consider the point a core point. In the figure, as minPts is 4, O is defined as a core point. P is defined as a density reachable point as it near a core point within Eps; q is defined also as density reachable point as it is near within Eps to another density reachable point, that is near O. DBscan analyses singularly each point multiple times to revise their classification considering their neighbourhood, and give as an output the list of the point and their cluster belonging. All the outgroup points are classified as belonging to a fake cluster label (usually named 0). (from (Hahsler et al., 2019))

# Appendix 3 – Chapter 3 Supplementary material

Appendix Table 1 - Performance statistics, formula and meaning. TP = True Positive; FP = False Positive; TN = True Negative; FN = False Negative. The performances are calculated for both numbers of contigs and with respect to the known symbiont genomes. In the latter case, the positives are all the nucleotides from the target genome, and we consider what fraction of the genome has been correctly attributed to the target bin by the different tools.

Performance statistic	Formula	Meaning
Sensitivity/Recall	$\frac{TP}{(TP + FN)}$	The fraction of relevant instances that have been detected over the total amount of relevant instances.
Precision	$\frac{TP}{(TP + FP)}$	Or positive predictive value, is the fraction of relevant instances among the retrieved instances.
Accuracy	$\frac{(TP + TN)}{(TP + FP + FN + TN)}$	Proportion of correct results over all cases obtained
F1 score	$2 \cdot \frac{TP}{(2 \cdot TP + FP + FN)}$	Measure of test accuracy, integrating precision and sensitivity

Appendix Table 2 - List of SeqDex scripts

Name	Description
SeqDex.sh	SeqDex main executable file
rRNA16S.R	Identification of 16S genes in the sample
Taxonomy.R	Find taxonomic affiliation of the contigs and calculate the TaxonDensity
GCKmersCov.R	k-mers frequencies, coverage and GC calculations
SVM.R	Prediction of taxonomy using Support Vector Machine (SVM)
RF.R	Prediction of taxonomy using Random Forest (RF)
Clustering.R	DBScan of the output of SVM.R and/or RF.R
Func.R	Function used by the scripts

Appendix Table 3 - The table shows sensitivity, accuracy, precision and F1 scores calculated over the number of target organism's contigs correctly deconvolved by running Blobology, MaxBIN and SeqDex with both algorithm on the simulated dataset

Statistics	Blobology	MaxBIN	SeqDex-SVM	SeqDex-RF
Sensitivity	0.7916	0.9583	0.9560	0.9560
Precision	1.0000	0.9583	0.9886	0.9886
Accuracy	0.9743	0.9897	0.9935	0.9935
F1 score	0.8837	0.9583	0.9720	0.9720

Appendix Table 4 - The table shows sensitivity, accuracy, precision and F1 scores calculated over the number of target organism's contigs correctly deconvolved by running Blobology, BusyBee, MetaBAT, MaxBIN and SeqDex with both algorithm on the *Ca. Fokinia solitaria* dataset.

Statistics	Blobology	BusyBee	MetaBAT	MaxBin	SeqDex-SVM	SeqDex-RF
Sensitivity	0.8333	0.9444	0.7778	0.9444	0.9444	0.9444
Precision	0.6818	0.0265	1.0000	0.3696	1.0000	1.0000
Accuracy	0.9982	0.8885	0.9988	0.9946	0.9998	0.9998
F1 score	0.7500	0.0516	0.8750	0.5313	0.9714	0.9714

Appendix Table 5 - The table shows sensitivity, accuracy, precision and F1 scores calculated over the total number of *Cardinium* contig correctly deconvolved by running Blobology, BusyBee, MetaBAT, MaxBIN and SeqDex with both algorithm on the partially deconvolved real dataset

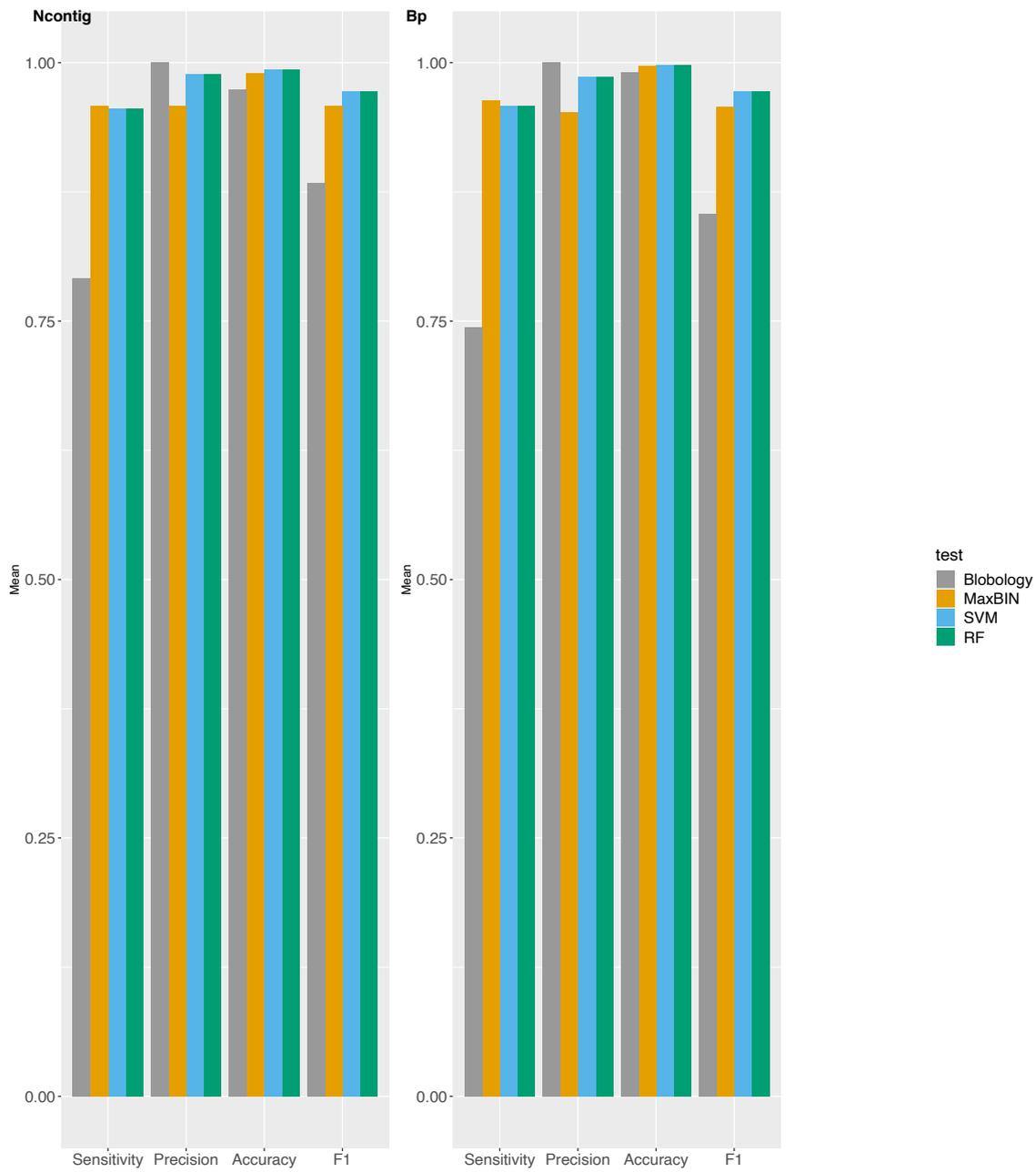
Statistics	Blobology	BusyBee	MetaBAT	MaxBin	SeqDex-SVM	SeqDex-RF
Sensitivity	1.0000	0.9292	0.7100	0.8142	0.8584	0.8319
Precision	0.0012	0.4234	0.7474	0.5786	0.6467	0.6667
Accuracy	0.1236	0.9986	0.9992	0.9992	0.9994	0.9994
F1 score	0.0024	0.5817	0.7282	0.6765	0.7376	0.7402

Appendix Table 6 - The table shows sensitivity, accuracy, precision and F1 scores calculated over the total number of *Wolbachia* contig correctly deconvolved by running Blobology, BusyBee, MetaBAT, MaxBIN and SeqDex with both algorithm on the partially deconvolved real dataset. Here, the taxonomic affiliations are obtained through combination of the information obtained by comparing contigs to nt and nr database.

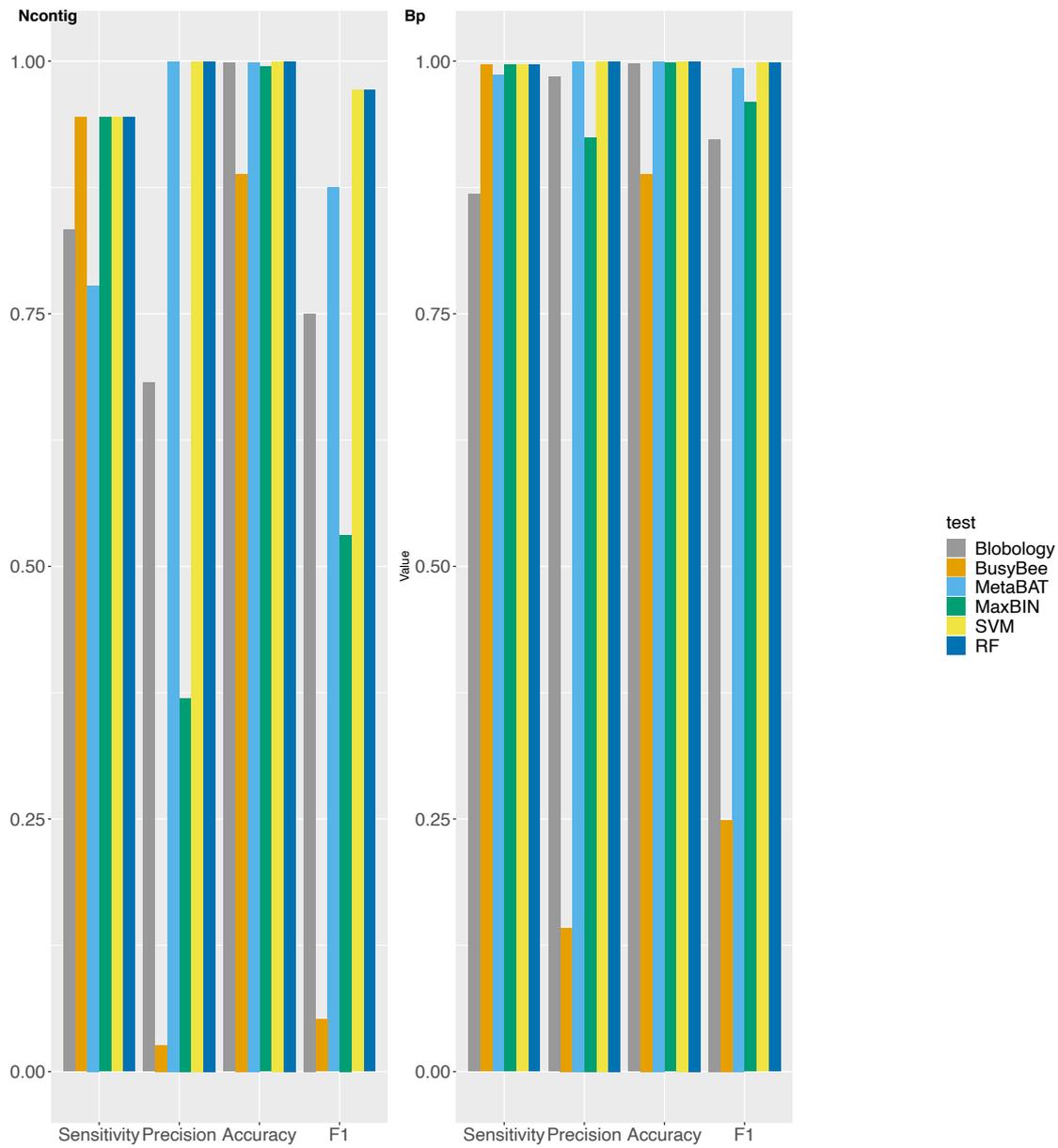
Statistics	Blobology	BusyBee	MetaBAT	MaxBin	SeqDex-SVM	SeqDex-RF
Sensitivity	0.9643	0.9286	0.8846	0.8574	0.9286	0.9286
Precision	0.0006	0.1048	0.5349	0.0390	0.3170	0.3333
Accuracy	0.5974	0.9979	0.9996	0.9945	0.9995	0.9995
F1 score	0.0012	0.1884	0.6667	0.0745	0.4727	0.4906

Appendix Table 7 - Performances of the classification made by SeqDex with or without the graph built by exploiting pairing information, and the ratio among the two.

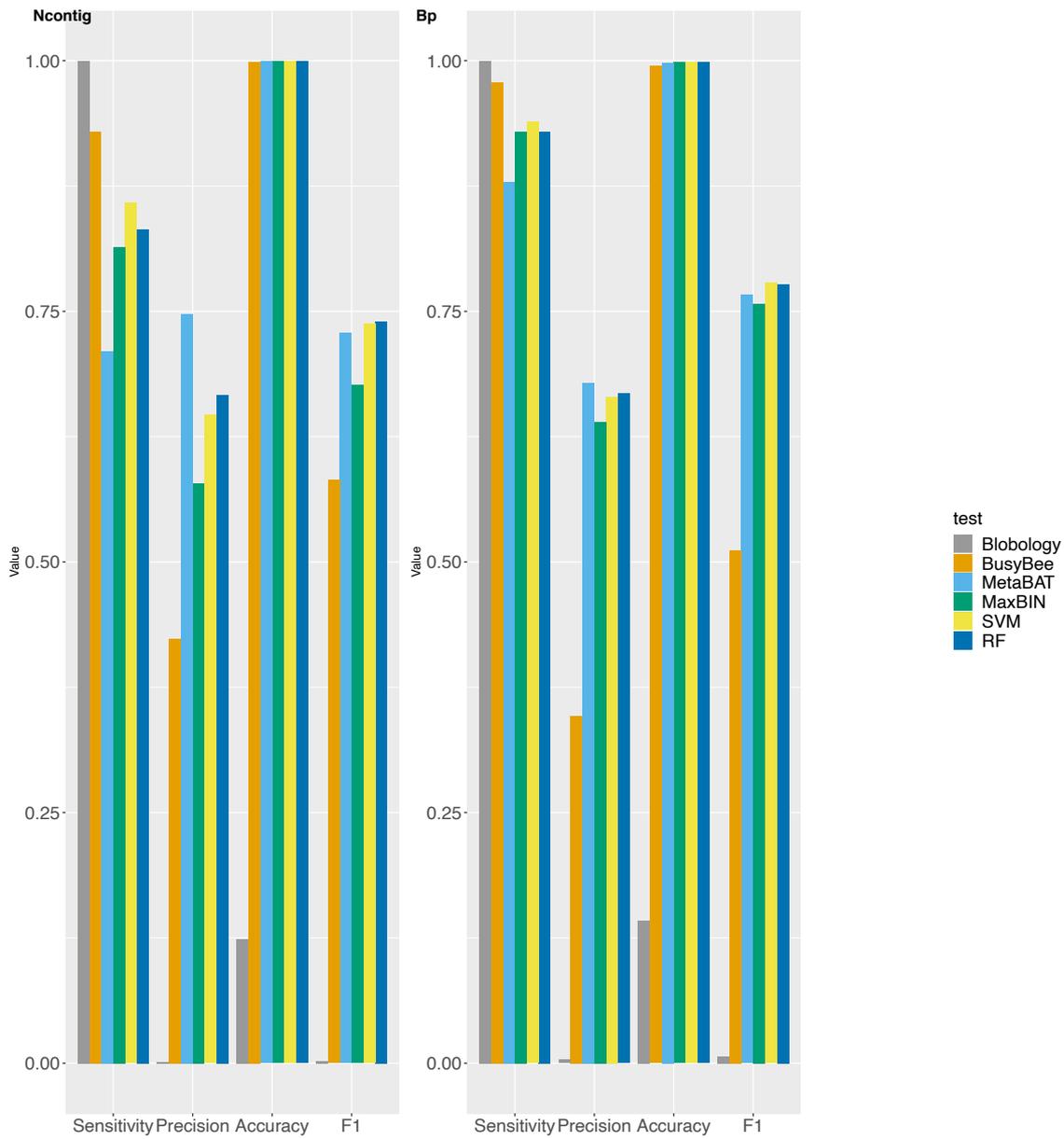
FOKINIA							
		Based on contigs number			Based on length		
Performance	ML	With graph	Without graph	Ratio w/o	With graph	Without graph	Ratio w/o
Sensitivity	SVM	0.971428571	0.833333333	1.165714286	0.996623147	0.973410781	1.023846424
precision	SVM	1	1	1	1	1	1
Accuracy	SVM	0.999836681	0.999464573	1.000372308	0.999937259	0.999505694	1.000431779
F1	SVM	0.985507246	0.909090909	1.084057971	0.998308718	0.986526262	1.011943377
Sensitivity	RF	0.971428571	0.833333333	1.165714286	0.996623147	0.973410781	1.023846424
precision	RF	1	1	1	1	1	1
Accuracy	RF	0.999836681	0.999464573	1.000372308	0.999937225	0.999505694	1.000431745
F1	RF	0.985507246	0.909090909	1.084057971	0.998308718	0.986526262	1.011943377
CARDINIUM							
		Based on contigs number			Based on length		
Performance	ML	With graph	Without graph	Ratio w/o	With graph	Without graph	Ratio w/o
Sensitivity	SVM	0.859649123	0.840707965	1.022530009	0.938849397	0.932346207	1.006975081
precision	SVM	0.649006623	0.655172414	0.990589055	0.665173353	0.663342462	1.002760099
Accuracy	SVM	0.999369281	0.999372763	0.999996516	0.998466562	0.998443755	1.000022843
F1	SVM	0.739622642	0.736434109	1.004329692	0.778664269	0.775169794	1.004508013
Sensitivity	RF	0.833333333	0.814159292	1.023550725	0.929061562	0.92437096	1.005074372
precision	RF	0.669014085	0.643356643	1.039880588	0.668413434	0.662456787	1.008991752
Accuracy	RF	0.999396704	0.999335867	1.000060877	0.998472048	0.998427092	1.000045027
F1	RF	0.7421875	0.71875	1.032608696	0.777473488	0.771798725	1.007352647
WOLBACHIA							
		Based on contigs number			Based on length		
Performance	ML	With graph	Without graph	Ratio w/o	With graph	Without graph	Ratio w/o
Sensitivity	SVM	0.928571429	0.964285714	0.962962963	0.986758222	0.991960115	0.994755946
precision	SVM	0.313253012	0.002291242	136.7175368	0.426951454	0.019491726	21.90424071
Accuracy	SVM	0.99946069	0.891543372	1.121045506	0.997716523	0.914579092	1.090902396
F1	SVM	0.468468468	0.004571622	102.4731398	0.596017506	0.0382322	15.58941167
Sensitivity	RF	0.928571429	0.892857143	1.04	0.986758222	0.984143075	1.002657284
precision	RF	0.329113924	0.036603221	8.991392405	0.432357843	0.187743061	2.302923154
Accuracy	RF	0.999497253	0.993902889	1.005628683	0.997765858	0.992685221	1.005118074
F1	RF	0.485981308	0.070323488	6.910654206	0.601265347	0.315331034	1.906775044



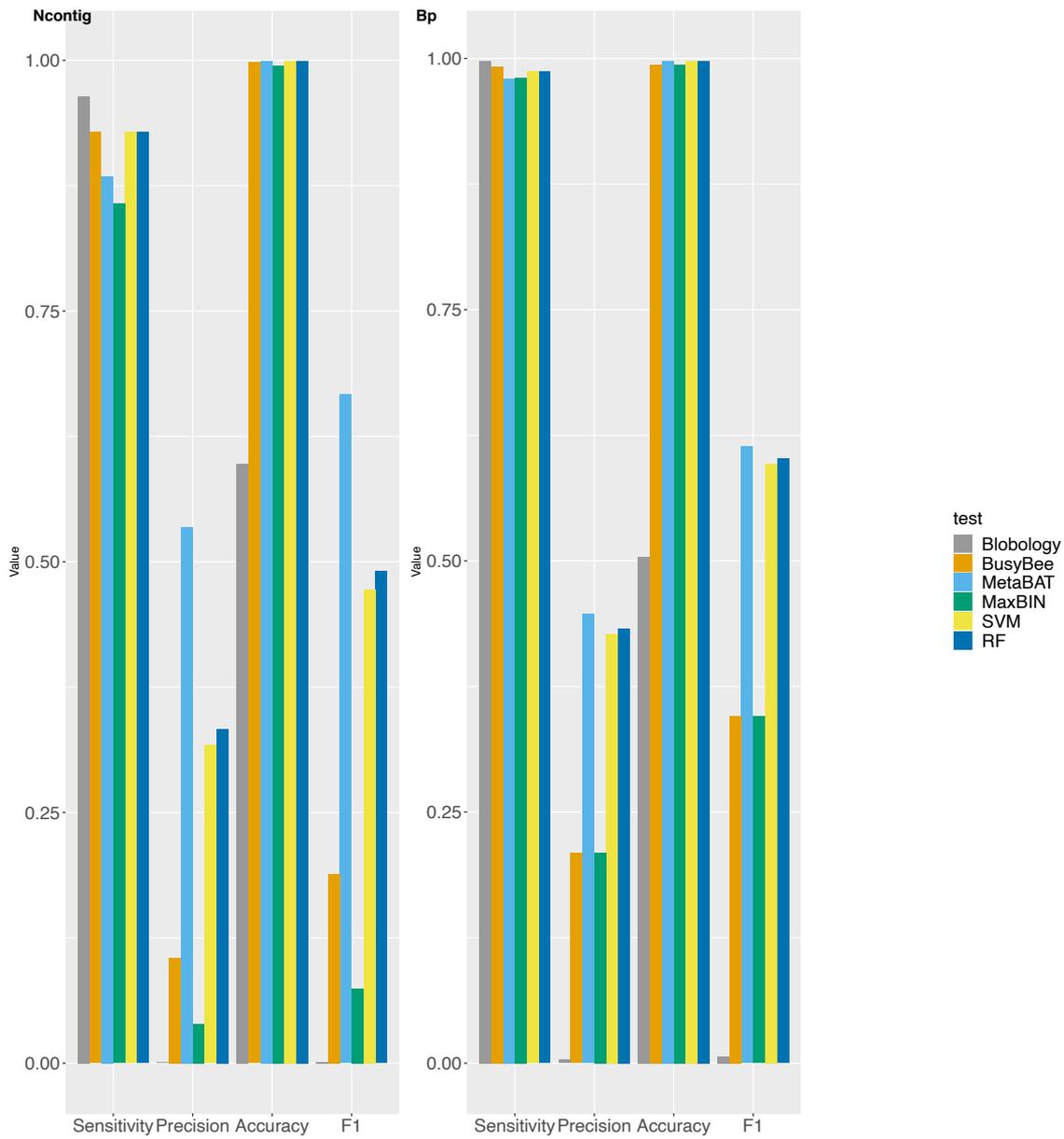
Appendix Figure 4 - The barplots show sensitivity, accuracy, precision and F1 score calculated for number of contigs (Ncontig) and amount of bp (Bp) of target organisms obtained by using Blobology, MaxBIN and SeqDex, with both algorithms, to deconvolve the simulated dataset.



Appendix Figure 5 - The barplots show sensitivity, accuracy, precision and F1 score of number of contigs (Ncontig) and amount of bp (Bp) of the *Ca. Fokinia solitaria* identified by using Blobology, BusyBee Web, MetaBAT, MaxBIN and SeqDex, with both machine learning algorithms, to deconvolve the dataset.



Appendix Figure 6 - The barplots show sensitivity, accuracy, precision and F1 score calculated for number of contigs (Ncontig) and amount of bp (Bp) of *Cardinium* correctly deconvolved by using Blobology, BusyBee Web, MetaBAT, MaxBIN and SeqDex, with both machine learning algorithms, to deconvolve the dataset.



Appendix Figure 7 - The barplots show sensitivity, accuracy, precision and F1 score calculated for number of contigs (Ncontig) and amount of bp (Bp) of *Wolbachia* correctly deconvolved by using Blobology, BusyBee Web, MetaBAT, MaxBIN and SeqDex, with both machine learning algorithms, to deconvolve the dataset.

# Appendix 4 - SeqDex Manual

## Introduction

SeqDex is an automated approach for the extraction of contigs from target species in whole genome sequencing of symbiont(s) together with the host. It uses taxonomic affiliations obtained through comparison to reference databases, coupled with a composition-based strategy to identify genomic contigs of the target organism(s). The workflow has four main phases: (i) data are prepared for subsequent analysis; (ii) taxonomic affiliation are associated to contigs and used to train supervised machine learning models (SVM, RF) on k-mer frequencies; (iii) the models are used to predict the taxonomy of unlabelled contigs; (iv - optional) unsupervised clustering (DBscan) can be used to refine the classification at lower taxonomic ranks.

## Dependencies

### *1. Packages*

The dependencies that can be installed via terminal are:

Samtools

Bedtools

NCBI-BLAST+

Barrnap

Prodigal

Diamond

Seqtk

Of these, only Prodigal and Diamond are optional (see above).

Most of SeqDex has been developed in R, so it requires the following R packages:

Taxonomizr

Seqinr

randomForest

e1071

Uwot

Dbscan

Parallel

doParallel

Foreach  
Optparse  
Ggplot2  
igraph

Of these, only Taxonomizr needs some manual curation. As described elsewhere (<https://cran.r-project.org/web/packages/taxonomizr/vignettes/usage.html>), Taxonomizr is used to convert NCBI accession number in taxonomic information. To do so, it uses a sql file that stores the information for the conversion. This file (`accessionTaxa.sql`) have to be downloaded and compiled. We make the user note that following the instruction for 'preparation' and use `prepareDatabase('accessionTaxa.sql')` command, a file containing only nucleotide NCBI accession number will be generated; for using SeqDex with proteins, then the 'Manual preparation database' instruction must be followed. Moreover, while this step needs to be done only once it has to be rerun if the database change.

SeqDex uses the standard filename `'accessionTaxa.sql'` so use this default name and do not create multiple versions, or it will return an internal error.

## *2. Databases*

SeqDex models combine supervised and unsupervised steps. In the supervised section, SeqDex needs BLAST databases for defining taxonomy labels of the contigs:

The minimal requirement databases are (i) a nucleotide database to which comparing the contigs, and (ii) the RDP rRNA database to which comparing the 16S genes found by `Barrnap` (download it form <https://rdp.cme.msu.edu/misc/resources.jsp>, in the unaligned format and then use `makeblastdb` from the BLAST plus suite to obtain the BLAST database. Do not remove the fasta file, as it is used by SeqDex to retrieve taxonomic information).

When the target endosymbionts are particularly distant from relative genome published online, or when the sample under analysis is particularly complex taxonomic affiliation from a nucleotide-nucleotide comparison can be insufficient to train the machine learning models. In these cases, coupling nucleotide and protein derived taxonomy may be useful. To do so, the user has to install two optional dependencies (`prodigal` and `diamond`) as well as download a protein database to be formatted using `diamond` (see instructions for building the database).

Please note that to use protein-derived taxonomy affiliations you must rerun the Taxonomizr preparation step to include proteins in the database (see [above](#)).

## SeqDex workflow

SeqDex can be divided into 3 main phases, discussed in detail below. These comprehend (i) an initial manipulation of the data to obtain the information needed in the subsequent phases; (ii) a classification/prediction through machine learning, and (iii) a final clustering and prediction step.

### *Input files*

The inputs to SeqDex.sh are: (1) the basename of the contig/scaffold assembly file in fasta format, and (2) the basename of the sam file obtained when mapping reads to the assembly. The quality of the assembly can influence the capability of the model to discriminate organisms, so please take care in this step. We suggest to evaluate sequencing quality by using FastQC and to remove adapters and low quality bases with Trimmomatic or similar tools. If the sequencing has been performed in paired end mode, it may be important to check the overlapping and merge the overlapping mates by using FLASH. At this point assemble the reads with SPADes or an equivalent assembler at several k-mer lengths to identify the best assembly using Quast. At this point, reads can be mapped back on the assembly. You can choose to use whenever tool you prefer, as long as the assembly file is in fasta format and the mapping file in sam format.

We considered paired end sequencing as most whole genome sequencing are performed by using this modality. However, SeqDex can be used also on single end sequencing data without modifying the code.

### *Phase 1 - data preparation*

#### *Coverage calculation*

FLASH merges overlapping paired reads, when found. Once performed, you will obtain as output the paired non-overlapping reads and the extended reads resulting from merging the overlapping mates. As Bedtools coverage do not differentiate among single reads and “mates” (here used in the general sense of the two reads composing a pair), we calculate the coverage for single and paired reads separately and then sum them. The outputs are saved in the Coverage folder.

### *Taxonomic affiliations*

SeqDex needs taxonomic affiliations to train the machine learning models. `SeqDex.sh` by default uses only nucleotide-derived taxonomic affiliations (`TAX=NT`). Insert the name of the nucleotide database (`NTI`) and the path to this file (`NT`) by editing the `SeqDex.sh` file. SeqDex will run `blastn` by using the contigs as queries. `SeqDex Taxonomy.R` script then reads the blast output to : (1) convert the NCBI codes into taxonomy information; (2) filter hits by the alignment length (`--aliLength`, default= 200 bp) and the percentage of identity (`--xid`, default= 70%); (3) for each taxonomic level (`--taxaLevels`, default= 1) the taxonomy Density (`--taxonDensity`, default= 0.75) is calculated and filtered at a defined threshold; (4) save the output in Taxonomy folder for subsequent analysis.

If you wish to use another tool similar to `blastn` to find the taxonomy affiliations derived from alignment to a nucleotide database, you can add your code at line 218 of the `SeqDex.sh` file. In this case, the output name and structure need to be equivalent to the one produced by our `blastn` code.

If you want to use protein-derived affiliations too, you must change in `SeqDex.sh` `TAX` to `NTNR`, provide the name of the database (`NRI`) and its path (`NTR`). Nucleotide workflow proceeds as explained above. The protein workflow implements `Prodigal` to obtain gene predictions and corresponding protein sequences, and then compare them to the NR database by using `Diamond`. Outputs are taken by `SeqDex Taxonomy.R` script to: (1) convert the NCBI code into taxonomic information; (2) filter hits by the alignment length (`--aliLength`; `--aliLengthNR`, default= 70 aa) and the percentage of identity (`--xid`; `--xidNr`, default 90 %); (3) for each taxonomic level (`--taxaLevel`) the taxonomy Density (`--taxonDensity`) is calculated and filtered at a defined threshold; (4) the shared affiliations combined with the unique affiliations, relative to both databases, are merged and the output is saved in Taxonomy folder.

You may wish to repeat the machine learning classification step on more than one taxonomy level, i.e.: on the Superkingdom level to select only bacterial contigs, and then at the level of bacterial classes, to select the class of interest (e.g. *Alphaproteobacteria*, see [below](#)). To do this, you need to tell to the model to prepare the taxonomic information for each level considered (Superkingdom and Class) by providing a comma separated list taxonomy levels desired, from superkingdom to species (run `Rscript Taxonomy.R --help` to see a full list of options; `--taxaLevels`, default=superkingdom). We suggest not to go too deep in the taxonomy levels: most of the times, when the sample under analysis shows low complexity (only one symbionts with high coverage), the classification at Superkingdom level is enough to then clustering and obtain the contigs relative to the target organism (this can be guided by preliminary rtPCR on 16S, for instance).

As we build SeqDex to be highly flexible, you can choose to prepare all the file at first entrance, run SeqDex as desired, evaluate the final output and then consider rerunning only certain parts by manually copying commands from `SeqDex.sh` to the terminal.

### ***Mate network construction***

Nucleotide and/or protein derived taxonomic affiliation can be too scarce information to build satisfactory predictive models. This is imputable to the peculiarities of the organism under analysis: stable obligate endosymbionts often show reduced genomes, with gene losses and high AT content. In these conditions, even if a related genome and its proteins are present in public databases (or, at least, in the databases used for inferring taxonomic affiliations), they could be too divergent to obtain good and valuable matches.

This will reduce the number of labelled contigs, affecting negatively the parametrization and the performance of the machine learning models. Including protein-derived taxonomical affiliations could improve the number of labels obtained. However, there is the risk to inflate the computational time and resources for little improvement. Also, due to possible erroneous taxonomic affiliation of some deposited NCBI sequences, there is the possibility to obtain different affiliations for the same contigs coming from nt and nr databases. In this case, there will be no shared labels between the two databases, and these will be discarded, reducing the number of labels retrieved.

To help improve the taxonomy coverage of the sample, we decided to use the information of paired end reads mapping in the assembly. In detail, SeqDex builds a graph based on mates mapping on different contigs. Here, if the edges (corresponding to the pairing information) that connect vertices (corresponding to contigs) are confirmed by `--Edges`, pairs (default= 10 pairs), we assume that the connected vertices (contigs) were parts of the same genome that have been put into different contigs due to lack of overlap. The complexity of such a graph can be high and the user can control it by setting the maximum degree of contigs (`--VerticesDegree`, default= 5). Highly connected contigs are mostly repeated regions, and their presence is basically at the basis of our difficulties in assembling complete genomes from short reads only. However, in this case the presence of intricate connected components (CC) provides a way to increase the taxonomy coverage of the dataset. Therefore, we set rules to exploit this graph to transfer taxonomy labels to contigs belonging to connected components where at least some of the contigs provide congruent taxonomical information derived from the homology search. The user can set `--componentSize`, which is the minimum size for a CC to be considered for trying label transfer (default=2).

### **Use with care!**

These graph, with the parameters and the rules described below, will be applied: (i) when parsing the taxonomy derived from nt and/or nt to transfer the affiliations to connected unlabelled contigs; (ii) after each machine learning prediction step, to correct errors or give

additional support; (iii) after the clustering, to collect connected contigs which may be erroneously assigned to another group, but also to retrieve the contigs shorter than the threshold for length (see [below](#)).

Rules to transfer taxonomy information:

- (1) if the CC has size 2 and only one vertex is labelled, then the label is always transferred;
- (2) if the CC has size 2 and the vertices have incongruent labels then: during (i) the two are maintained as they are while in (ii) and (iii) all the contigs in the CC will be labelled as 'misclassified'.
- (3) if the CC contains more than two vertices and there is only one type of label, then this is always extended;
- (4) if the CC contains more than two vertices, there are two labels and the underrepresented have a frequency less than `--mixedComponents`: during (i) the unlabelled contigs receive the most frequent label, whereas the second label is nonetheless maintained; however, in (ii) and (iii) the incongruent labels **are considered as 'erroneous' and corrected**;
- (5) if there are more than two labels: in (i) all the labels obtained through the BLAST search are kept, but no transfer occurs; in (ii) and (iii) all the contigs in the CC are labelled as 'misclassified'.

In (i) the user can choose to use all contigs in each CC to transfer the taxonomy affiliations with the rules described above, or to limit the transferring to a certain distance from the taxonomy labelled contigs (`--taxTransfer`, default = all).

### *rDNA 16S*

The 16S genes in the sample have to be located and their likely origin has to be identified at the end of the clustering step, when SeqDex looks for the cluster with the rRNA gene with the target taxonomy affiliation and the highest coverage among all the 16S rDNA genes present (if any other is present). `Barrnap` identifies putative 16S rRNA genes, that are then compared to the RDP database. The name of the fasta RDP file (`RDPF`), the name of the database file (`RDP1`) and the path to the folder where these are located (`RDP`) has to be specified in `SeqDex.sh`.

The output of `blastn` is analysed by `rRNA16S.R`, with the selection of only one among all the taxonomic affiliations obtained for each 16S contigs. Only affiliations of contigs longer than

`--minContigLen` (default= 1000) and with alignment length longer than 500 bp are saved in the output, in the Taxonomy folder. 16S genes can be fragmented in the assemblies, but we chose to consider only putative genes of at least 500 bp (change is possible, editing the `rRNA16S.R` script).

### *K-mer frequencies*

K-mer frequencies are used both to perform the machine learning classification and the clustering.

In the `GCKmersCov.R` script, k-mer frequencies are calculated by accounting for reverse, complement and palindromes. Frequencies are calculated by normalizing the counts by the  $1000/(\text{length of the sequence})$ .

SeqDex calculates 3-mers frequencies by default (`--kmers`). We choose this length because the main purpose of this model is to discriminate endosymbionts from their hosts, which are mostly eukaryotic, and so the trimers could implicitly summarize the information on codon usage and on coding density, which is very different in Bacteria and Eukaryotes. Moreover, the phylogenetic signal of k-mers increases with the length, but this introduces the necessity for setting a threshold on the length of contigs that becomes larger for longer k-mers. Using trimers reduces this problem, as there are only 64 combinations that reduces to XX variables thanks to reverse complementarity and palindrome compression.

Nonetheless, the user can choose whichever k-mer length but going above 5-6 can cause memory and computational time to increase a lot.

In `GCKmersCov.R` also calculates the total coverage per nucleotide by first summing for each contigs the single end reads coverage with half of the paired end reads coverage, to then divide it by the size.

The output of `GCKmersCov.R` is saved in the Coverage folder and it contains the k-mers frequencies, the GC content, the nucleotide coverage and the contigs length.

### *Phase 2 - machine learning classification*

Once completed, the output files of the phase 1 became the input file for phase 2, which couples K-mer frequencies with taxonomic affiliations to train machine learning models that are at the basis of the prediction of taxonomy for all contigs. The `SVM.R` and `RF.R` scripts both take as input the table with K-mer frequencies (`--gcCovKmersTable`), 16S rRNA gene table (`--rRNA16S`), the taxonomy table (`--taxonomy`) and the mate CC network (`--network`). Together with the network, arguments `--Edges`, `--VerticesDegree`, `--componentSize` and `--mixedComponents` are needed, as described [before](#).

Only contigs longer than `--minContigLength` and having a label assigned in the taxonomy table will be used to build the machine learning models (labelled contigs viz unlabelled). The labelled dataset is randomly split into training and test set, where the former represents two thirds of the total and is used to train the models while the latter is used to test the models and to calculate error rates. **This procedure is repeated, with a novel split of the dataset into train and test each time.** The number of model build is user defined by setting `-nmodel` (default = 100). The percentage of erroneous contigs/length is

reported as a cumulative measure over all 100 times permuted models and therefore represents a measure that is largely independent on the identity of the contigs in the train/test datasets. The scripts calculate sensitivity, precision, accuracy and F1 score based on both the number of contigs and the cumulative length on the comparison between empirically inferred taxonomy and the predicted.

After this procedure, the unlabelled contigs are predicted by using all the 100 models and then the percentage of inclusion within a taxonomical category is calculated. As output, only the most represented taxonomical category is reported. However, if a contig shows more affiliations and the 100 predictions are distributed among them such that are equally divided and the so percentages of belonging to each are equal (i.e.: two categories, 50% of presence each), the model keep them all.

The taxonomic affiliation, predicted or obtained through blast, is integrated and eventually corrected using the mate CC network discussed before and following the same rules described above. During this step uncertain predictions can be corrected. If there are more taxonomical affiliations than allowed by the mate CC network parameters, then the contigs involved are considered 'misclassified'. These contigs, as there is uncertainty about their origin, are included into the next iteration or in the output to avoid thrashing sequences which may turn out to be informative. Also, these scripts calculate a homogeneity (H) index, defined as the number of homogeneous CC divided by the total number of CC (both homogenous and heterogenous in terms of taxonomic predictions). It goes from 0 (all heterogeneous) to 1 (all homogeneous) to give an idea of how much the CCs are likely representing contigs from the same genome. When this value is low, we suggest not using the network at all.

The user can choose to print on screen these stats or only on the output stats file (`--verbose`, default `TRUE`).

The machine learning algorithms have been developed to be highly parallelizable, so both SVM and RF support `--threads` argument (default 8). Nonetheless, the entire strategy, mainly for the 100 iterations, requires several hours. To reduce execution time, the user can control the number of iterations.

We developed our SVM and RF scripts to allow easy customization of critical parameters for both.

Support vector is sensible to `--cost`, `--gamma` and `--cross` values. Our implementation automatically selects best cost and gamma from a provided interval of values. The user can choose to use default intervals (`--cost: 1e-1,1e3`; `--gamma: 1e-5,1e-1`; `--cross: 5`) or provide its own. Also, the SVM implementation in e1071 automatically selects the best kernel type for the input data. For further information see <https://cran.r-project.org/web/packages/e1071/index.html>.

The `randomForest` package used by `SeqDex` (in `RF.R` script) automatically chooses one of the major influential parameters of the random forest algorithm (the number of variables tried at each split, see <https://CRAN.R-project.org/package=randomFores>). However, our implementation allows the user to choose the number of trees in the forest and if the sampling have to be done with or without replacement (default: `--ntree 550`, `--replace TRUE`).

You can decide whether to pass the output of the machine learning to the clustering (Phase 3) by setting `CLUSTER=yes` (default). When the user avoids running the clustering, the model gives as final output the fasta file of the contigs listed in the machine learning output.

### *Phase 3 - clustering*

The clustering takes as input the taxonomy (`--taxonomy`), k-mer frequencies (`--gcCovKmersTable`), 16S genes (`--rRNA16S`), mate network (`--network`), and the Phase 2 output files (`--modelOutput`). If the machine learning prediction is done at several taxonomy ranks, only one of the taxonomy files need to be provided, as also the last prediction output.

The `cluster.R` script reduces the k-mer frequencies to a user defined number of components (`--ncomponents`, default= 2). As the `uwot` packages performs a parallelized form of UMAP, then the `THREADS` value set in the `sh` file will be used here too (`--threads`, default= 8) to reduce execution time.

The user can choose to do the clustering only on these new components or also on GC content and/or coverage value (`input`, default= `Kmers`). In this way the user can combine the *Blobology* and the higher order composition analysis spirits in the clustering step. Combining these variables at this step allows to give some more weight to k-mers (2 variables) than to coverage or GC (1 variable each).

After obtaining these new dimensions, `DBscan` clusters the data. The script calculates a `minPts` value, as equal to the logarithm of the number of contigs used in clustering, and on this calculates the best `Eps` value. Usually, `Eps` is chosen by plotting the k Nearest Neighbors distances (kNN) and selecting the distance values where the curve changes slope. We avoid this user-dependent part by rounding the kNN distances to the first digit, calculating the difference between consecutive and selecting the kNN value which has the greatest difference from its subsequent value.

After `DBscan`, each contig is assigned to a cluster and support to the assignments is evaluated by using the mate network, as previously done for the taxonomy and the machine learning prediction steps. However, here also contigs shorter than the provided length

(minContigLen, default= 1000) will be included in the output by extending the cluster belonging to all members of the CCs.

At this point SeqDex searches for the cluster containing the target 16S gene that has also the higher coverage among all genes with the same label.

Finally, the fasta file of the contigs of interest is returned as output.

### *Output files*

SeqDex produces various folders with output files, most of which are used as checkpoints to be able to rerun the analysis, if needed.

Coverage: coverage files (paired and single: sambasenamefile\_PAIRED\_end.bed and sambasenamefile\_SINGLE\_end.bed), the file with k-mer frequencies, GC content and total coverage informations (gckCovTable.txt), and the mate network file (contigNet.txt).

Taxonomy: (i) blastn output (ContigsvsNt.txt); (ii) if performed, prodigal predicted proteins (prodigalContigs.faa and gff format ProdContigs) and diamond output (ContigsvsNr.txt); (iii) the elaboration made by Taxonomy.R, which produces one file per taxonomy rank (superkingdomTaxonomyIteration.txt); (iv) 16S genes predicted by barrnap (barrnap16s\_contigs.gff, 16sContigs.fasta), the alignment on RDP (16sContigsvsRDP.txt), the RDP based taxonomy file (RDP16s\_taxa\_mod.txt); (v) the final elaboration performed by rRNA16S.R (rRNA16STaxonomy2.txt).

SVMoutput and RFoutput: for each iteration performed, the SVM/RF script produces a file containing statistics (superkingdomOutput\_statsSVM.txt, superkingdomOutput\_statsRF.txt; the number indicate the iteration therefore the taxonomy rank), a file with k-mer frequencies of the target contigs selected at the end of the iteration (superkingdomOutputSVM.txt, superkingdomOutputRF.txt) and the environment (superkingdomSVMModel.RData, superkingdomRFModel.RData; so that an expert user can use the SVM/RF models outside SeqDex). In the stats file is reported performance statistics of the models constructed in each machine learning R script. In detail, for each model, the prediction performed on the test set is compared to the relative empirically inferred taxonomy; the cumulative comparison of all test set are used to calculate cumulative percentage of error, sensitivity, precision, accuracy and F1 score by

considering both the number of contigs and the total length. At the end of this file is reported the H-index.

When the clustering step is not performed, these folders also contain the fasta files of the contigs selected as belonging to the target organism by the machine learning algorithms (`superkingdomOutputSVM_contigs.fasta`, `superkingdomOutputRF_contigs.fasta`).

The output file with the target contigs contains:

- name of the contigs;
- GC content;
- k-mer frequencies;
- Coverage;
- TaxonDensity value: if it is 'NoBlastHit' means that the contig affiliations is not empirically inferred but has been predicted in this phase; if it is '-1' means that the label is derived from the mate CC network-based extension;
- taxonomy labels converted into numbers (the order reported in stats file);
- percentage of times a contig has been classified in a certain taxonomical category, calculated over the predictions done over the 100 models performed. '-1' Could mean either that the contigs label is empirically inferred or that it comes from network-based extension.

ClusteringOutput: if SeqDex is run enabling both SVM and RF, it will produce two clustering folders (`ClusteringOutputSVM` and `ClusteringOutputRF`) but only a folder named `ClusteringOutput` if only one machine learning algorithm was performed. The `Clustering.R` script produces: (i) the list of contigs in the target cluster, extended by using the mate network to (`OutputClustering.txt`); (ii) their sequences (`OutputClusteringSVM.fasta`, `OutputClusteringRF.fasta`); (iii) the clustering stats, reassuming the number of clusters obtained, the amount of contigs in each cluster, and the homogeneity index (`output_statsClustering.txt`); (iv) the final clustering table validated with the mate network (`extendedClusteringCC.txt`); (v) a file with the scatterplot of the contigs plotted by using two umap dimensions, and the scatterplot of the contigs colored by clusters (`Rplots.pdf`; cluster 0 is used to indicate outliers contigs, so the belonging to this cluster is not considered).

The `extendedClusteringCC.txt` contains:

- the names of the contigs;
- the new umap dimensions;
- the percent of assignment of a contigs to a label (calculated during the SVM/RF prediction phase);

- the identifier of the CC where each contig is found: if it is '-1' means that the contigs does not belong to a CC (it is an isolated vertex);
- The identifier of the cluster/group to which each contig belongs.

## Running SeqDex

SeqDex can be run by using default parameters or by customizing one or more parameters. In both cases, there are options that must be specified by the user. To do this, open the SeqDex.sh file as text and at least provide mandatory variables: the `THREADS` (the number of threads to be used), `NT` (the path to the nucleotide databases file), `NTI` (the name of the nucleotide database – without extension), `RDP` (the path to the RDP files – database and fasta file must be in the same folder), `RDPF` (the name of the RDP fasta file), `RDPI` (the name of the RDP database – without extension), `SCRIPT` (the path to the folder containing the SeqDex folder), `TAX` (specify `NT` or `NTNR` to consider `nt` or both `nt` and `nr`-derived taxonomy labels), `MLALG` (`SVM`, `RF`, `BOTH`: self-explanatory), `TRG` (the target category at Class level – required only when `CLUSTERING=YES`), `CLUSTER` (`YES/NO`, whether to perform the DBscan clustering), `ITER` (taxonomy level on which perform SeqDex) and `ITERTRG` (target taxonomy category for `ITER`). If also protein-derived taxonomy affiliations are needed then set also the optional variables `NR` (the path to the protein database file) and `NRI` (the protein Diamond database),

### *Default*

To run SeqDex by using default parameters, you only need to specify the above options. Then change permissions of the file, if needed, and run on the terminal

```
./SeqDex.sh basename_mapping_file basename_contigs_fastafile
```

Doing this, SeqDex.sh produces a taxonomy file at only the Superkingdom level, it predicts affiliation with both SVM and RF selecting only contigs from Bacteria (both obtained through BLAST or predicted by SeqDex) together with the misclassified ones, and then it performs the clustering step to retrieve the cluster with *TRG* 16S gene with higher coverage. The list of the default parameters is in tables 1-6.

## *Taxonomy parameters*

To change the target, the taxonomy level and the category on which performing the prediction and selection of the contigs:

- (1) change TRG argument in `SeqDex.sh`, to your taxonomy category at Class level;
- (2) change ITER (taxonomy level) and ITERTRG (taxonomy category for ITER level) variables to obtain taxonomic affiliation and prediction on a different level than the default.

## *Iteration parameters*

To perform the iterations (that is, to perform the training of the machine learning methods on multiple taxonomical categories, in series):

(1) change the default parameter in `SeqDex.sh` and provide a comma separated list of taxonomy levels (ITER), taxonomy target category for each level (ITERTRG) and the final target category at class level (TRG). By doing this, `SeqDex` will produce one taxonomy file for each level of interest (see [above](#)). These files will be named by the code corresponding to the taxonomy level pasted in front of the basename 'taxonomyIteration.txt' (i.e.: ITER=superkingdom,class, superkingdomTaxonomyIteration.txt for the first iteration; classTaxonomyIteration.txt for second iteration). These variables will be used also by both SVM.R and RF.R.

(2) provide to SVM.R and/or RF.R scripts a comma separated list, without spaces, of the taxonomy files produced by Taxonomy.R. By default, `SeqDex` will use ITER to reconstruct the name of these files, but if different filenames are needed, change the TAXFILE variable providing a comma separated list (without spaces) of the files (with also their path). Here, the scripts will (i) use the labels of the first iteration (superkingdom or superkingdomTaxonomyIteration.txt) to predict the unlabelled contigs; (ii) integrate the affiliations with the mate network to validate/correct labels; (iii) select contigs with the target label (Bacteria) and then (iv) use only these on the subsequent iteration. Step (i)-(iv) will be repeated for all taxonomic levels and targets provided;

(3) If the final clustering phase is not performed (CLUSTER=not), `SeqDex` will automatically recognize the last prediction file and return the FASTA file of the contigs in it. Else, if the final clustering step (CLUSTER=yes) is enabled, then the R script will use ITER together with MLALG to retrieve the last output prediction file and perform the clustering on it. However, if a different file has to be used, than simply change MODELOUT variable providing the file together with its path.

Adding iterations may be meaningless if the percentages of error returned by the model are high. Most of the times, in our experience, you may need to adjust Taxonomy.R values

to use only highly sure taxonomic affiliations and predict only at Superkingdom level. We suggest not to go too deep in taxonomic levels: SeqDex at each iteration uses the contigs with the taxonomic affiliations obtained with `Taxonomy.R` which have been selected in the previous iteration; like this the number of contigs on which the model is applied will decrease at each iteration, and the error committed in each iteration will affect the next ones, making nearly impossible to go from Superkingdom to *Species* with affordable predictions. This is indeed also true for metagenomics tools, that not necessarily are able to provide species assignments with low error rate. Consider also that endosymbionts, especially the obligate ones, may not show high percentage of identity at species level even if close relative genomes are available. In our experience, we used at maximum only two iteration (Superkingdom and Class).

### *Multiple targets dataset*

To deconvolve dataset with more than a target symbiont, the user can decide to:

- (1) run a SeqDex run specifying a single target organism: just run SeqDex for the first time on one of the target organisms, and then take advantage of coverage, taxonomy, rRNA16S, k-mers frequencies files already prepared to speed up subsequent run on a different target organism. `SeqDex.sh` will detect them automatically and skip the command lines needed to produce them, which are highly time consuming. Take care to rename the files of the prediction/clustering of the first organisms, or move them, to avoid overwriting;
- (2) modify the `SeqDex.sh` file to make SeqDex able to run all the targets deconvolution in sequences: (i) copy and paste all variable fields (line 7-143) and change them, in particular take care to change TRG, ITERTRG and the path of the input variable files to make sure it will find the file written in the first run and used by the subsequent other run (coverage, taxonomy, rRNA16S, k-mers frequencies); (ii) add command `mkdir name_new_folder; cd name_new_folder`; (iii) copy and paste the prediction and clustering, if performed, part (lines 301 to 517 of the file); (iv) add command `cd ..`; (v) repeat point (i)-(iv) for each subsequent target. Take care to change the path of each file that have to be provided to each new prediction-clustering R scripts. See example file.

Deconvolving multiple targets dataset can be tricky. Just take care of: percentage of error committed by the machine learning models, as low error may involve the possibility to go deeper in the taxonomy levels considered; use the `rRNA16sTaxonomy2.txt` to find on which contigs the targets 16S genes are, and then check if they are correctly deconvolved in different groups in the final clustering step (if performed). In the latter case, if in the cluster of the searched target there is the 16S contigs of another target, then this means that the

two have not been correctly deconvolved and so you may wish to add an iteration to go deeper in the taxonomy and arrive to a more correct separation.

### *Rerunning*

Similarly, once completed the analysis, you may want to rerun the entire script or part of it to see if the performance or the outputs will change significantly by changing parameters. The user can then choose to:

- (1) re-run the entire `SeqDex.sh` by changing parameters of interest and taking care to move or rename old files **or they will be overwritten**;
- (2) run only steps of interest, as each of our R scripts can be run independently by typing `Rscript namescript.R` in the terminal. Similarly, to see all the options available for the script, type `Rscript namescript.R --help` or `Rscript namescript.R -h`. As before, old files must be renamed or moved, **or they will be overwritten**.

When machine learning predictions of SeqDex return high percentage of errors, we suggest to try to change `aliLength` and `Xid` of `Taxonomy.R` (and the corresponding parameters for the protein searches, if used). Samples with low complexity may not be highly influenced by these thresholds, while samples with high complexity (multiple targets, contaminating sequences of other bacteria, etc.) might need more affordable taxonomic affiliations; using stricter threshold for these values may be the right choice. To test the influence of these on your dataset, move or rename files to avoid overwriting (point (1)).

## Default values of R scripts

Appendix Table 8 - Taxonomy.R default parameters with indication of the SeqDex.sh lines in which are used.

SeqDex.sh variables	Taxonomy.R	Default value	Meaning
BLAST	--blast	ContigsvsNt.txt	Contigs vs nucleotide database file. The path is needed if the R script is run in a different folder than this file (default: same folder)
ITER	--taxaLevels	1	Taxonomy level to consider. 1 means Superkingdom
ALILENGTH	--aliLength	200	Minimum alignment length to nucleotide hit
XID	--Xid	70	Minimum percentage of identity between contigs and nucleotide hit
TAXONDENSITY	--TaxonDensity	0.75	Minimum considerable percentage (in 0-1 scale) of belonging to a taxonomic affiliation of a contig
DIAMONDD	--diamond	ContigsvsNr_mod.txt	Contigs vs protein database file. The path is needed if the R script is run in a different folder than this file (default: same folder)
ALILENGTHNR	--aliLengthNr	70	Minimum alignment length to protein hit
XIDNR	--XidNr	80	Minimum percentage of identity between contigs and protein hit
NETWORK	--network	../Coverage/contigNet.txt	Mate CC network file with path
EDGES	--Edges	10	Minimum value for an edge to be considered
VERTICES	--VerticesDegree	5	Maximum vertex degree allowed
COMPONENTSIZE	--componentSize	2	Minimum size of the component CC
VERTEXDIST	--taxTransfer	all	If 'all' then all contigs in each CC is considered to transfer the taxonomy affiliations, whereas only the contigs distant no more than the provided value will be considered

Appendix Table 9 - rRNA16S.R default parameters with indication of the SeqDex.sh lines in which are used

SeqDex.sh variables	rRNA16S.R	Default	Meaning
BLASTRDP	--blastRDP	16sContigsvsRDP.txt	Output file of 16S genes vs RDP. The path is needed if the R script is run in a different folder than this file (default: same folder)
RDPTAXA	--taxaRDP	RDP16s_taxa_mod.txt	File with RDP taxonomy
MINLENGTH	--minContigLen	1000	Contigs length threshold: only contigs longer than this value will be considered

Appendix Table 10 - GCKmersCov.R default parameters with indication of the SeqDex.sh lines in which are used.

SeqDex.sh variables	GCKmersCov.R	Default	Meaning
(input)	--contigs	../"\${2}" .fasta	Contigs fasta file with its path. SeqDex uses the second argument to automatically find it
KMERS	--Kmers	3	K-mers length to be used
SINGLE	--covSingle	onlymapping_sorted_SINGLE.bed	Extended reads coverage file. If the R script is run in a different folder than this file, the path is also needed
PAIRED	--covPaired	onlymapping_sorted_PAISED.bed	Paired end reads coverage file. If the R script is run in a different folder than this file, the path is also needed
THREADS	--threads	"\$THREADS"	Number of threads to be used. SeqDex.sh uses the variable THREADS defined in line 7

Appendix Table 11 - SVM.R default parameters with indication of the SeqDex.sh lines in which are used.

SeqDex.sh variables	SVM.R	Default	Meaning
KMERSFREQ	--gcCovKmersTable	../Coverage/gckCovTable.txt	Output of GCKmersCov.R (with path)
RNA	--rRNA16S	../Taxonomy/rRNA16sTaxonomy2.txt	Output of rRNA16S.R (with path)
TAXFILE	--taxonomy	"\${ITER}"	Output of Taxonomy.R (with path)
NETWORK	--network	../Coverage/contigNetwork.txt	Mate CC network file (with path)
CROSS	--cross	5	Maximum number of allowed erroneous contigs in SVM model construction
COST	--cost	1e-1,1e3	Range of cost to be tested to tune SVM model
GAMMA	--gamma	1e-5,1e-1	Range of gamma values to be tested to tune SVM model
SCALE	--scale	F	Boolean value meaning whether to scale the data or not (usually not needed)
MINLENGTH	--minContigLen	1000	Contigs length threshold: only contigs longer than this value will be considered
NMODEL	--nmodel	100	Number of models constructed
ITERTRG	--targetName	Bacteria	Name of the taxonomical label to be selected after prediction
ITER	--TaxaName	Superkingdom	Taxonomical level on which perform the prediction
THREADS	--threads	"\$THREADS"	Number of threads to be used. SeqDex.sh uses the variable THREADS defined in line 7
EDGES	--Edges	10	Minimum value for an edge to be considered
VERTICES	--VerticesDegree	5	Maximum vertex degree allowed
COMPONENTSIZE	--componentSize	2	Minimum size of the component CC
MIXEDCOMP	--mixedComponents	0.2	Maximum proportion of alternative label in the component to consider it as erroneous and correct
VERBOSE	--verbose	T	Boolean value meaning if to print output stats of the model in screen or not (won't disable stats file writing)

Appendix Table 12 - RF.R default parameters with indication of the SeqDex.sh lines in which are used

SeqDex.sh variables	RF.R	Default	Meaning
KMERSFREQ	--gcCovKmersTable	../Coverage/gckCovTable.txt	Output of GCKmersCov.R (with path)
RNA	--rRNA16S	../Taxonomy/rRNA16sTaxonomy2.txt	Output of rRNA16S.R (with path)
TAXFILE	--taxonomy	"\${ITER}"	Output of Taxonomy.R (with path)
NETWORK	--network	../Coverage/contigNetwork.txt	Mate CC network file with path
MINLENGTH	--minContigLen	1000	Contigs length threshold: only contigs longer than this value will be considered
ITERTRG	--targetName	Bacteria	Name of the taxonomical label to be selected after prediction
NMODEL	--nmodel	100	Number of models constructed
ITER	--TaxaName	Superkingdom	Taxonomical level on which perform the prediction
REPLACE	--replace	T	Boolean value meaning whether the RF sampling have to be done with or without replacement
NTREE	--ntree	500	Number of tree to be constructed by RF
THREADS	--threads	"\$THREADS"	Number of threads to be used. SeqDex.sh uses the variable THREADS defined in line 7
EDGES	--Edges	10	Minimum value for an edge to be considered
VERTICES	--VerticesDegree	5	Maximum vertex degree allowed
COMPONENTSIZE	--componentSize	2	Minimum size of the component CC
MIXEDCOMP	--mixedComponents	0.2	Maximum proportion of alternative label in the component to consider it as erroneous and correct
VERBOSE	--verbose	T	Boolean value meaning if to print output stats of the model in screen or not (won't disable stats file writing)

Appendix Table 13 - Clustering.R default parameters with indication of the SeqDex.sh lines in which are used

SeqDex.sh variables	Clustering.R	Default	Meaning
MODELOUTSVM	--modelOutput	"\${ITER}"	Output file of SVM.R to be used for clustering
MODELOUTRF	--modelOutput	"\${ITER}"	Output file of RF.R to be used for clustering
KMERSFREQ	--gcCovKmersTable	../Coverage/gckCovTable.txt	Output of GCKmersCov.R (with path)
RNA	--rRNA16S	../Taxonomy/rRNA16sTaxonomy2.txt	Output of rRNA16S.R (with path)
TAXFILE	--taxonomy	"\${ITER}"	Output of Taxonomy.R (with path)
NETWORK	--network	../Coverage/contigNetwork.txt	Mate CC network file with path
TYPE	--input	Kmers	Which type of input variable to be used for clustering (Kmers=only k-mers frequencies reduced by umap)
MINLENGTH	--minContigLen	1000	Contigs length threshold: only contigs longer than this value will be considered
TRG	--targetName	"\$TRG"	Name of the target Class searched. SeqDex.sh uses TRG variable assigned in line 32
THREADS	--threads	"\$THREADS"	Number of threads to be used. SeqDex.sh uses the variable THREADS defined in line 7
EDGES	--Edges	10	Minimum value for an edge to be considered
VERTICES	--VerticesDegree	5	Maximum vertices degree to be considered
COMPONENTSIZE	--componentSize	2	Minimum size of the component to be considered
MIXEDCOMP	--mixedComponents	0.2	Maximum proportion of alternative label in the component to consider it as erroneous and correct
MDL	--modelType	"\$MDL"	Variable used to reconstruct input filename in case modelOutput=ITER

## Example files

Download the Example folder from [figshare.com/articles/SeqDex\\_example\\_data\\_zip/8845841](https://figshare.com/articles/SeqDex_example_data_zip/8845841).

This folder contains files to run the analysis on the simulated dataset. By running SeqDex on this example dataset you will be able to check if R dependencies have been installed successfully. To do so, enter in this folder in the terminal, change permissions to `SeqDex_example.sh` file (if needed), and then run `./SeqDex_example.sh remap contigs`

This command will automatically detect the files in `Taxonomy` and `Coverage` folders to pass them to `SVM.R` and `RF.R`, without final clustering step. We decided to provide these folders to allow an easy and fast check of the mandatory dependencies without the need to have also reference databases for nucleotide and 16S genes. Also, we compared the contigs against a database composed by only the two genomes used to create this simulated dataset which indeed provided an affordable taxonomy affiliation for all the contigs, making impossible to complete the predictive step. We then selected the `blastn` hit for 2/3 of the contigs (randomly sampled) to then perform the prediction on the other 1/3.

In addition, also the file used to perform SeqDex on the *P. penetrans* dataset (`seqDex_Ppenetrans.sh`) can be found, which contain at least two endosymbionts. We added it as an example of how the sh file can be modified to perform our model on more than a symbiont dataset.

# Appendix 5 - SeqDex bash script

```
#!/bin/bash
#$1 basename alignment file
#$2 basename contigs file
echo $1
echo $2
#####
#####
#MANDATORY variables: these have to be assigned to run SeqDex
THREADS=10

#path to folder containing blast database files as downloaded from
#ftp://ftp.ncbi.nlm.nih.gov/blast/db/
#or any other custom database with sequence titles fulfilling NCBI
sequence #titles formatting rules
NT=~/database/nt

#file name of the blast database in $NT
NTI=nt

#path to blast database built using $RDPF downloaded from
#https://rdp.cme.msu.edu/misc/resources.jsp
#or any other custom database with sequence titles fulfilling RDP
sequence #titles formatting rules
RDP=~/database/rdp16s

#fasta file used to build $RDPI
RDPF=current_Bacteria_unaligned.fa

#file name of the blast database in $RDP
RDPI=rdp16S

#home folder of SeqDex
```

```

SCRIPT=~/Downloads/

#taxonomy information to be used. Can be NT of NTNR
TAX=NT

#machine learning algorithm to be used. Can be RF, SVM, BOTH
MLALG=BOTH

#target taxonomy name to be used to identify the target cluster
TRG=Alphaproteobacteria

#if it is equal to 'yes', then SeqDex perform the final clustering
step
CLUSTER=yes

#Taxonomy level/s to be used in SeqDex. One level for only one
iteration; #a comma separated list (without spaces) for more than
an iteration
ITER=superkingdom

#Taxonomy category target for each level defined in $ITER
ITERTRG=bacteria
#####
#####
#OPTIONAL variables: these variables have to be assigned only if
$TAX=NTNR
#path to folder containing nr database file in diamond format
#or any other custom database with sequence titles fulfilling NCBI
sequence #titles formatting rules
NR=~/database/

#file name of the nr diamond database in $NR
NRI=nrTaxonomy.dmnd
#####
#####

```

*#input FILENAME used by the R scripts (default, usually no need to be #changed)*

*#length of the kmers frequency calculated in GCKmersCov.R*  
KMERS=3

*#kmers frequency table (output of gcKmersCov.R)*  
KMERSFREQ=../Coverage/gckCovTable.txt

*#16S rRNA file produced by rRNA16S.R*  
RNA=../Taxonomy/rRNA16sTaxonomy2.txt

*#Mate CC network file produced after coverage calculation*  
NETWORK=../Coverage/contigNet.txt

*#taxonomy file that will be used by SVM.R, RF.R and Clustering.R R scripts. #By default they reconstruct these files name by using \$ITER*  
TAXFILE=\$ITER

*#coverage files*  
SINGLE=onlymapping\_sorted\_SINGLE.bed  
PAIRED=onlymapping\_sorted\_PAIRED.bed

*#print at screen output or run in silent mode*  
VERBOSE=T

#####  
#####

*#Shared parameters between Rscript*

*#consider only contigs longer than \$MINLENGTH*  
MINLENGTH=1000

*#Number of SVM/RF model constructed in model training*  
NMODEL=100

```

#####
#####
#Taxonomy.R parameters ONLY
#blast output of the blast of the contigs vs $NTI
BLAST=ContigsvsNt.txt

#aliLength minimum length of an HSP to be considered for taxonomy
in blastn #output
ALILENGTH=200

#Xid minimum percentage of identity of an HSP to be considered for
taxonomy #in blastn output
XID=70
if [ $TAX == "NTNR" ]; then
    #diamond output of the diamond of the prodigal predicted protein
vs    #$NRI (only if $TAX=NTNR)
    DIAMOND=ContigsvsNr_mod.txt
    #aliLengthNR minimum length (in aa) of an HSP to be considered
for    #taxonomy in Diamond output
    ALILENGTHNR=70
    #XidNr minimum percentage of identity of an HSP to be considered
for    #taxonomy in Diamond output
    XIDNR=80
Fi

#TaxonDensity congruence score for taxonomy from blast
TAXONDENSITY=0.75
#####
#####
#rRNA16S.R parameters ONLY - file names are the default, change only
if #needed
#output of Barrnap 16S contigs vs RDP database (blastn)
BLASTRDP=16sContigvsRDP.txt

#RDP taxonomies of the hits found by blast

```

RDPTAXA=RDP16s\_taxa\_mod.txt

*#minimum alignment length between the 16S gene found by Barrnap and the #blast hit to consider it in the subsequent analysis*

MIN16SLEN=500

#####  
#####

*#CC mate network hyperparameters--use with care!*

*#minimum edge value considered*

EDGES=10

*#maximum vertices degree considered*

VERTICES=5

*#minimum component size*

COMPONENTSIZE=2

*#maximum proportion of alternative taxonomy over the total to consider the #alternative as erroneous and correct/uniform the CC taxonomy*

MIXEDCOMP=0.2

*#variable to control taxonomy transfer/correction: if all, then all vertex #in the component will be used;*

*#otherwise insert the maximum distance from taxonomy labelled vertex to #limit the taxonomy transfer/correction*

VERTEXDIST=all

#####  
#####

*#SVM hyperparameters--use with care! see e1071 package vignette for #explanation*

CROSS=5

COST=1e-1,1e3

GAMMA=1e-5,1e-1

SCALE=F

```

#####
#####
#RF hyperparameters--use with care! se randomForest package vignette
for #explanation
REPLACE=T
NTREE=500
#####
#####
#Clustering parameters
#Output file produced by by SVM.R and RF.R R scripts that have to
be used #by Clustering.R. By default they reconstruct these files
name by using # $ITER; if more than an iteration have been performed,
it automatically #uses the last $ITER taxonomy level.
MODELOUTSVM=$ITER
MODELOUTRF=$ITER

#type of dimensions used for clustering. kmers=only umap reduced
kmers #frequencies dimensions; to add coverage and/or GC content
write kmers,cov #or kmers,GC or kmers,cov,GC
TYPE=Kmers

#number of umap component to be produced to reassume Kmers
frequencies
NCOMP=2

#suffixes needed for reconstruct machine learning output file name
from # $ITER (do not have to be changed)
if [ $MLALG == "SVM" ]; then
    MDL=SVM
elif [ $MLALG == "RF" ]; then
    MLD=RF
elif [ $MLALG == "BOTH" ]; then
    MLD1=SVM
    MLD2=RF
fi

```

```

##coverage calculation
echo "coverage calculation"

if [ -d Coverage ];then
    echo "Using existing Coverage directory for the following
steps. If this is not ok, please rename the directory.";

else
    echo "The Coverage directory does not exist. Creating..."
    mkdir Coverage
fi

cd Coverage

if [ -f onlymapping_sorted_PAired.bed ];then
    echo "Using existing coverage outputs for the following steps.
If this is not ok, please remove the file.";

else
    echo "The coverage outputs do not exist. Running...This will
take a while"

    samtools view --threads "$THREADS" -b ../"${1}".sam -o
"${1}".bam

    samtools view -F4 --threads "$THREADS" "${1}".bam -o
"${1}"_onlymapping.bam

    samtools view -H ../"${1}".sam > header.sam

    samtools reheader header.sam "${1}"_onlymapping.bam >
"${1}"_onlymapping_h.bam

```

```
samtools sort --threads "$THREADS" "${1}"_onlymapping_h.bam >
"${1}"_onlymapping_sorted.bam
```

```
grep "@SQ" header.sam | cut -f2,3 -d ":" --output-delimiter "
" | sed -e 's/LN/1/g' > Scaffold_size.bed
```

```
samtools view -F1 --threads "$THREADS" -O BAM
"${1}"_onlymapping_sorted.bam | bedtools bamtobed >
"${1}"_SINGLE_end.bed
```

```
samtools view -f1 --threads "$THREADS" -O BAM
"${1}"_onlymapping_sorted.bam | bedtools bamtobed >
"${1}"_PAIRED_end.bed
```

```
coverageBed -a Scaffold_size.bed -bed -b "${1}"_SINGLE_end.bed
> onlymapping_sorted_SINGLE.bed
```

```
coverageBed -a Scaffold_size.bed -bed -b "${1}"_PAIRED_end.bed
> onlymapping_sorted_PAIRED.bed
```

```
samtools sort -n ../"${1}".sam -o ReadsRemap.bam -@ "$THREADS"
```

```
samtools view ReadsRemap.bam > ReadsRemap.sam
```

```
cut -f3,7 ReadsRemap.sam | grep -v "=" | grep -vi "*" | sort |
uniq -c > contigNet.txt
```

```
fi
```

```
cd ..
```

```
##taxonomic affiliations & 16S
```

```
echo "taxonomy affiliation and 16S"
```

```
if [ -d Taxonomy ];then
```

```
echo "Using existing Taxonomy directory for the following
steps. If this is not ok, please rename the directory.";
```

```

else
    echo "The Taxonomy directory does not exist. Creating..."
    mkdir Taxonomy;
fi

cd Taxonomy

if [ -f ContigsvsNt.txt ];then
    echo "Using existing blastn output for the following steps. If
this is not ok, please remove the file.";

else
    echo "The blastn output does not exist. Running blastn...This
will take a while"
    #blastn of the contigs file vs the database $NTI
    blastn -query ../"${2}".fasta -out ContigsvsNt.txt -db
"$NT"/"${NTI}" -outfmt "6 std qlen slen qcovs gaps qcovhsp" -
num_threads "$THREADS";
fi

if [ $TAX == "NTNR" ]; then
    if [ -f ContigsvsNr_mod.txt ];then
        echo "Using existing diamond output for the following
steps. If this is not ok, please remove the file.";

        else
            echo "The diamond output does not exist. Running prodigal
and diamond..."
            #blastn of the contigs file vs the database $NTI
            prodigal -i ../"${2}".fasta -a prodigalContigs.faa -f gff
-p meta -q -o ProdContigs
            diamond blastp --db "$NR"/"${NRI}" --query prodigalContigs.faa --
out ContigsvsNr.txt --outfmt 6 qseqid qlen sseqid slen qstart qend

```

```

sstart send evaluate bitscore length pident gaps staxids stitle qcovhsp
-p "$THREADS" --quiet
        cut -f1,2,3,4,5,6,7,8,9,10,11,12,13,14,16 ContigsvsNr.txt
> ContigsvsNr_mod.txt
        fi
fi

echo "16S rRNA genes in contigs"
if [ -f RDP16s_taxa_mod.txt ];then
        echo "Using existing 16S genes for the following steps. If this
is not ok, please remove the file.";

else
        echo "The 16S gene file does not exist. Running...This will
take a while"

#find rRNAs in contigs
        barrnap -kingdom bac --threads "$THREADS" ../"${2}".fasta --
quiet ON > barrnap16s_contigs.gff
#extract contigs with matches to 16S_rRNA
        grep -w 16S_rRNA barrnap16s_contigs.gff | cut -f1 >
16scontigsName.txt ; seqtk subseq ../"${2}".fasta 16scontigsName.txt
> 16sContigs.fasta
#blastn contigs containing 16S rRNA against $RDPI
        blastn -query 16sContigs.fasta -out 16sContigvsRDP.txt -db
"$RDP"/"${RDPI}" -outfmt "6 std qlen slen qcovs gaps qcovhsp" -
num_threads "$THREADS"
#retrieving RDP taxonomy
        cut -f2 16sContigvsRDP.txt > RDP16s.txt; grep "$RDP"/"${RDPF}"
-wf RDP16s.txt > RDP16s_taxa.txt
#adapting output
        sed "s/ /_/g" RDP16s_taxa.txt | sed "s:/_/g" | sed "s/#/_/g"
| sed "s/'/_/g" | sed "s/;/\t/g" | sed "s/>/_/g" >
RDP16s_taxa_mod.txt
        fi

```

```

cd ..
##Rscript

echo "Rscripts for kmers frequencies, SVM and final clustering step"
echo "Taxonomy"

cd Taxonomy
#create taxonomy tables
#build Taxonomy table with the following options (for full list run
Rscript #TaxonomyDef.R -h at the terminal)
#--taxaLevels 1 meaning contigs will be classified at superkingdom
level
#--blast output of the blast of the contigs vs $NTI
#--aliLength minimum length of an HSP to be considered for taxonomy
in #blastn output
#--Xid minimum percentage of identity of an HSP to be considered for
#taxonomy in blastn output
#--TaxonDensity congruence score for taxonomy from blast
#--diamond output of the diamond of the prodigal predicted protein
vs $NRI
#--aliLengthNR minimum length (in aa) of an HSP to be considered for
#taxonomy in Diamond output
#--XidNr minimum percentage of identity of an HSP to be considered
for #taxonomy in Diamond output
#--network network file
#--Edges minimum edge value to be considered
#--VerticesDegree maximum vertices degree to be considered
#--componentSize minimum component size
#--mixedComponents maximum proportion of alternative taxonomy over
the #total to consider the alternative as erroneous and
correct/uniform the CC #taxonomy
if [ $TAX == "NT" ]; then
    Rscript "$SCRIPT"/SeqDex/Taxonomy.R --blast "${BLAST}" --
taxaLevels "$ITER" --aliLength "$ALILENGTH" --Xid "$XID" --
TaxonDensity "$TAXONDENSITY" --network "${NETWORK}" --Edges "$EDGES"

```

```

--VerticesDegree "$VERTICES" --componentSize "$COMPONENTSIZE" --
mixedComponents "$MIXEDCOMP" --taxTransfer "$VERTEXDIST"
elif [ $TAX == "NTNR" ]; then
    Rscript "$SCRIPT"/SeqDex/Taxonomy.R --blast "${BLAST}" --
taxaLevels "$ITER" --aliLength "$ALILENGTH" --Xid "$XID" --
TaxonDensity "$TAXONDENSITY" --diamond "${DIAMOND}" --aliLengthNr
"$ALILENGTHNR" --XidNr "$XIDNR" --network "${NETWORK}" --Edges
"$EDGES" --VerticesDegree "$VERTICES" --componentSize
"$COMPONENTSIZE" --mixedComponents "$MIXEDCOMP" --taxTransfer
"$VERTEXDIST"
fi

#create tables with taxonomy assignments for 16S rRNA contigs (for
full #list run Rscript rRNA16S.R -h at the terminal)
#--blastRDP output of the blast of the 16S contigs vs RDP
#--taxaRDP file with taxonomy from $RDPF
#--minContigLeng minimum length of contigs considered by the model
(should ##be the same of SVM/RF)
Rscript "$SCRIPT"/SeqDex/rRNA16S.R --blastRDP "${BLASTRDP}" --
taxaRDP "${RDPTAXA}" --minContigLen "$MINLENGTH" --min16SLen
"$MIN16SLEN"
cd ..
echo "gck"
cd Coverage
if [ -f gckCovTable.txt ];then
    echo "Using existing K-mers frequency table. If this is not ok,
please remove the file.";
else
    echo "The k-mers frequency table does not exists, running
Rscript"
#create tables with kmers frequencies, GC content and contigs
coverage
#--contigs contigs fasta file
#--kmers length of the kmer to use for machine learning

```

```

#--covSingle --covPaired files created during coverage calculation
(see above)
#--threads number of threads to be used
    Rscript          "$SCRIPT"/SeqDex/GCKmersCov.R          --contigs
../"${2}".fasta --Kmers "$KMERS" --covSingle "${SINGLE}" --covPaired
"${PAIRED}" --threads "$THREADS"
fi
cd ..
if [ $MLALG == "SVM" ]; then
    echo "SVM"
    mkdir SVMoutput
    cd SVMoutput
#predicting taxonomic affiliation using SVM based upon output of
#TaxonomyDef.R (for full list run Rscript SVMDef.R -h at the
terminal)
#--gckCovTable path and filename of the output of GCKmersCovDef.R
files
#--rRNA16S path and filename to output of rRNA16S.R
#--taxonomy path and filename to output of Taxonomy.R
#--network network file
#--cross --cost --gamma --scale parameters of SVM algorithm
#--minContigLeng minimum length of contigs to be considered by SVM
#--targetName name of the target taxonomy of interest
#--TaxaName taxonomy level of interest
#--threads number of threads to be used
#--Edges minimum edge value to be considered
#--VerticesDegree maximum vertices degree to be considered
#--componentSize minimum component size
#--mixedComponents maximum proportion of alternative taxonomy over
the #total to consider the alternative as erroneous and
correct/uniform the CC #taxonomy
#--verbose set to true to print on the terminal performance
statistics and #percentage of error committed by SVM model
    Rscript          "$SCRIPT"/SeqDex/SVM.R          --gcCovKmersTable
"${KMERSFREQ}" --rRNA16S "${RNA}" --taxonomy "${TAXFILE}" --network

```

```

"${NETWORK}" --cross "$CROSS" --cost "$COST" --gamma "$GAMMA" --
scale "$SCALE" --minContigLen "$MINLENGTH" --targetName "$ITERTRG"
--TaxaName "$ITER" --threads "$THREADS" --Edges "$EDGES" --
VerticesDegree "$VERTICES" --componentSize "$COMPONENTSIZE" --
mixedComponents "$MIXEDCOMP" --verbose "$VERBOSE"
    cd ..
    if [ $CLUSTER == "yes" ]; then
        echo "umap e dbscan"
        mkdir ClusteringOutput
        cd ClusteringOutput
#clustering contigs of the target taxonomy found with SVMDef.R. The
#cluster of interest is identified by using 16S rRNA of target
taxonomy #with higher coverage
##--gckCovTable path and filename of the output of GCKmersCovDef.R
files
##--rRNA16S path and filename to output of rRNA16S.R
##--taxonomy path and filename to output of Taxonomy.R
##--network network file
##--input variables to be used in clustering algorithm
##--minContigLeng minimum length of contigs to be considered by SVM
##--targetName name of the target taxonomy of interest
##--threads number of threads to be used
##--ncomponents number of components to be extract by umap algorithm
##--Edges minimum edge value to be considered
##--VerticesDegree maximum vertices degree to be considered
##--componentSize minimum component size
##--mixedComponents maximum proportion of alternative taxonomy over
the #total to consider the alternative as erroneous and
correct/uniform the CC #taxonomy

        Rscript "$SCRIPT"/SeqDex/Clustering.R --modelOutput
"${MODELOUTSVM}" --gcCovKmersTable "${KMERSFREQ}" --rRNA16S
"${RNA}" --taxonomy "${TAXFILE}" --network "${NETWORK}" --input
"$TYPE" --minContigLen "$MINLENGTH" --targetName "$TRG" --threads
"$THREADS" --ncomponents "$NCOMP" --Edges "$EDGES" --VerticesDegree

```

```

"$VERTICES" --componentSize "$COMPONENTSIZE" --mixedComponents
"$MIXEDCOMP" --modelType "$MDL"
    seqtk subseq ../"${2}".fasta OutputClustering.txt >
OutputClustering.fasta
    cd ..
else
    cd SVMoutput
    A=$(echo "$ITER" | rev | cut -f1 -d "," | rev)
    A+="OutputSVM.txt"
    cut -f1 "${A}" > outputSVM_names.txt
    seqtk subseq ../"${2}".fasta outputSVM_names.txt >
outputSVM_contigs.fasta
    cd ..
fi
elif [ $MLALG == "RF" ]; then
    echo "RF"
    mkdir RFoutput
    cd RFoutput
#predicting taxonomic affiliation using SVM based upon output of
#TaxonomyDef.R (for full list run Rscript RFDef.R -h at the terminal)
##--gckCovTable path and filename of the output of GCKmersCovDef.R
files
##--rRNA16S path and filename to output of rRNA16S.R
##--taxonomy path and filename to output of Taxonomy.R
##--network network file
##--minContigLeng minimum length of contigs to be considered by RF
##--targetName name of the target taxonomy of interest
##--TaxaName taxonomy level of interest
##--replace if sampling cases of RandomForest should be done with or
without #replacement
##--ntree number of tree to be grown in the forest
##--threads number of threads to be used
##--Edges minimum edge value to be considered
##--VerticesDegree maximum vertices degree to be considered
##--componentSize minimum component size

```

```

#--mixedComponents maximum proportion of alternative taxonomy over
the #total to consider the alternative as erroneous and
correct/uniform the CC #taxonomy
#--verbose set to true to print on the terminal performance
statistics and #percentage of error committed by RF model
    Rscript "$SCRIPT"/SeqDex/RF.R --gcCovKmersTable "${KMERSFREQ}"
--rRNA16S "${RNA}" --taxonomy "${TAXFILE}" --network "${NETWORK}" -
--minContigLen "$MINLENGTH" --targetName "$ITERTRG" --TaxaName
"$ITER" --replace "$REPLACE" --ntree "$NTREE" --threads "$THREADS"
--Edges "$EDGES" --VerticesDegree "$VERTICES" --componentSize
"$COMPONENTSIZE" --mixedComponents "$MIXEDCOMP" --verbose
"$VERBOSE"
    cd ..
    if [ $CLUSTER == "yes" ]; then
        echo "umap e dbscan"
        mkdir ClusteringOutput
        cd ClusteringOutput
#clustering contigs of the target taxonomy found with RFDef.R. The
cluster #of interest is identified by using 16S rRNA of target
taxonomy with higher #coverage
#--gckCovTable path and filename of the output of GCKmersCovDef.R
files
#--rRNA16S path and filename to output of rRNA16S.R
#--taxonomy path and filename to output of Taxonomy.R
#--network network file
#--input variables to be used in clustering algorithm
#--minContigLeng minimum length of contigs to be considered by RF
#--targetName name of the target taxonomy of interest
#--threads number of threads to be used
#--ncomponents number of components to be extract by umap algorithm
#--Edges minimum edge value to be considered
#--VerticesDegree maximum vertices degree to be considered
#--componentSize minimum component size

```

```
#--mixedComponents maximum proportion of alternative taxonomy over  
the #total to consider the alternative as erroneous and  
correct/uniform the CC #taxonomy
```

```
    Rscript      "$SCRIPT"/SeqDex/Clustering.R      --modelOutput  
"${MODELOUTRF}" --gcCovKmersTable "${KMERSFREQ}" --rRNA16S "${RNA}"  
--taxonomy "${TAXFILE}" --network "${NETWORK}" --input "$TYPE" --  
minContigLen "$MINLENGTH" --targetName "$TRG" --threads "$THREADS"  
--ncomponents "$NCOMP" --Edges "$EDGES" --VerticesDegree "$VERTICES"  
--componentSize "$COMPONENTSIZE" --mixedComponents "$MIXEDCOMP" --  
modelType "$MDL"
```

```
    seqtk subseq ../"${2}".fasta OutputClustering.txt >  
OutputClustering.fasta
```

```
    cd ..
```

```
else
```

```
    cd RFoutput
```

```
A=$(echo "$ITER" | rev | cut -f1 -d "," | rev)
```

```
A+="OutputRF.txt"
```

```
cut -f1 "${A}" > outputRF_names.txt
```

```
    seqtk subseq ../"${2}".fasta outputRF_names.txt >  
outputRF_contigs.fasta
```

```
    cd ..
```

```
fi
```

```
elif [ $MLALG == "BOTH" ]; then
```

```
    mkdir SVMoutput
```

```
    cd SVMoutput
```

```
    echo "SVM"
```

```
#predicting taxonomic affiliation using SVM based upon output of  
#TaxonomyDef.R (for full list run Rscript SVMDef.R -h at the  
terminal)
```

```
#--gckCovTable path and filename of the output of GCKmersCovDef.R  
files
```

```
#--rRNA16S path and filename to output of rRNA16S.R
```

```
#--taxonomy path and filename to output of Taxonomy.R
```

```
#--network network file
```

```

#--cross --cost --gamma --scale parameters of SVM algorithm
#--minContigLeng minimum length of contigs to be considered by SVM
#--targetName name of the target taxonomy of interest
#--TaxaName taxonomy level of interest
#--threads number of threads to be used
#--Edges minimum edge value to be considered
#--VerticesDegree maximum vertices degree to be considered
#--componentSize minimum component size
#--mixedComponents maximum proportion of alternative taxonomy over
the #total to consider the alternative as erroneous and
correct/uniform the CC #taxonomy
#--verbose set to true to print on the terminal performance
statistics and #percentage of error committed by SVM model
    Rscript      "$SCRIPT"/SeqDex/SVM.R      --gcCovKmersTable
"${KMERSFREQ}" --rRNA16S "${RNA}" --taxonomy "${TAXFILE}" --network
"${NETWORK}" --cross "$CROSS" --cost "$COST" --gamma "$GAMMA" --
scale "$SCALE" --minContigLen "$MINLENGTH" --targetName "$ITERTRG"
--TaxaName "$ITER" --threads "$THREADS" --Edges "$EDGES" --
VerticesDegree "$VERTICES" --componentSize "$COMPONENTSIZE" --
mixedComponents "$MIXEDCOMP" --verbose "$VERBOSE"
    cd ..
    if [ $CLUSTER == "yes" ]; then
        echo "umap e dbscan on SVM"

        mkdir ClusteringOutputSVM
        cd ClusteringOutputSVM
#clustering contigs of the target taxonomy found with SVMDef.R. The
#cluster of interest is identified by using 16S rRNA of target
taxonomy #with higher coverage
#--gckCovTable path and filename of the output of GCKmersCovDef.R
files
#--rRNA16S path and filename to output of rRNA16S.R
#--taxonomy path and filename to output of Taxonomy.R
#--network network file
#--input variables to be used in clustering algorithm

```

```

#--minContigLeng minimum length of contigs to be considered by SVM
#--targetName name of the target taxonomy of interest
#--threads number of threads to be used
#--ncomponents number of components to be extract by umap algorithm
#--Edges minimum edge value to be considered
#--VerticesDegree maximum vertices degree to be considered
#--componentSize minimum component size
#--mixedComponents maximum proportion of alternative taxonomy over
the #total to consider the alternative as erroneous and
correct/uniform the CC #taxonomy

        Rscript      "$SCRIPT"/SeqDex/Clustering.R      --modelOutput
"${MODELOUTSVM}"    --gcCovKmersTable      "${KMERSFREQ}"    --rRNA16S
"${RNA}"    --taxonomy "${TAXFILE}"    --network "${NETWORK}"    --input
"$TYPE"    --minContigLen "$MINLENGTH"    --targetName "$TRG"    --threads
"$THREADS"    --ncomponents "$NCOMP"    --Edges "$EDGES"    --VerticesDegree
"$VERTICES"    --componentSize "$COMPONENTSIZE"    --mixedComponents
"$MIXEDCOMP"    --modelType "$MDL1"

        seqtk  subseq  ../"${2}".fasta  OutputClustering.txt  >
OutputClusteringSVM.fasta
        cd ..
    else
        cd SVMoutput
        A=$(echo "$ITER" | rev | cut -f1 -d "," | rev)
        A+="OutputSVM.txt"
        cut -f1 "${A}" > outputSVM_names.txt
        seqtk  subseq  ../"${2}".fasta  outputSVM_names.txt  >
outputSVM_contigs.fasta
        cd ..
    fi
    echo "RF"
    mkdir RFoutput
    cd RFoutput
#predicting taxonomic affiliation using SVM based upon output of
#TaxonomyDef.R (for full list run Rscript RFDef.R -h at the terminal)

```

```

#--gckCovTable path and filename of the output of GCKmersCovDef.R
files
#--rRNA16S path and filename to output of rRNA16S.R
#--taxonomy path and filename to output of Taxonomy.R
#--network network file
#--minContigLeng minimum length of contigs to be considered by RF
#--targetName name of the target taxonomy of interest
#--TaxaName taxonomy level of interest
#--ntree number of tree to be grown in the forest
#--replace if sampling cases of RandomForest should be done with or
without #replacement
#--threads number of threads to be used
#--Edges minimum edge value to be considered
#--VerticesDegree maximum vertices degree to be considered
#--componentSize minimum component size
#--mixedComponents maximum proportion of alternative taxonomy over
the #total to consider the alternative as erroneous and
correct/uniform the CC #taxonomy
#--verbose set to true to print on the terminal performance
statistics and #percentage of error committed by RF model
    Rscript "$SCRIPT"/SeqDex/RF.R --gcCovKmersTable "${KMERSFREQ}"
--rRNA16S "${RNA}" --taxonomy "${TAXFILE}" --network "${NETWORK}" -
-minContigLen "$MINLENGTH" --targetName "$ITERTRG" --TaxaName
"$ITER" --replace "$REPLACE" --ntree "$NTREE" --threads "$THREADS"
--Edges "$EDGES" --VerticesDegree "$VERTICES" --componentSize
"$COMPONENTSIZE" --mixedComponents "$MIXEDCOMP" --verbose
"$VERBOSE"
    cd ..
    if [ $CLUSTER == "yes" ]; then
        echo "umap e dbscan on RF"
        mkdir ClusteringOutputRF
        cd ClusteringOutputRF
#clustering contigs of the target taxonomy found with RFDef.R. The
cluster #of interest is identified by using 16S rRNA of target
taxonomy with higher #coverage

```

```

#--gckCovTable path and filename of the output of GCKmersCovDef.R
files
#--rRNA16S path and filename to output of rRNA16S.R
#--taxonomy path and filename to output of Taxonomy.R
#--network network file
#--input variables to be used in clustering algorithm
#--minContigLeng minimum length of contigs to be considered by RF
#--targetName name of the target taxonomy of interest
#--threads number of threads to be used
#--ncomponents number of components to be extract by umap algorithm
#--Edges minimum edge value to be considered
#--VerticesDegree maximum vertices degree to be considered
#--componentSize minimum component size
#--mixedComponents maximum proportion of alternative taxonomy over
the #total to consider the alternative as erroneous and
correct/uniform the CC #taxonomy

```

```

Rscript "$SCRIPT"/SeqDex/Clustering.R --modelOutput
"${MODELOUTRF}" --gcCovKmersTable "${KMERSFREQ}" --rRNA16S "${RNA}"
--taxonomy "${TAXFILE}" --network "${NETWORK}" --input "$TYPE" --
minContigLen "$MINLENGTH" --targetName "$TRG" --threads "$THREADS"
--ncomponents "$NCOMP" --Edges "$EDGES" --VerticesDegree "$VERTICES"
--componentSize "$COMPONENTSIZE" --mixedComponents "$MIXEDCOMP" --
modelType "$MDL2"

```

```

seqtk subseq ../"${2}".fasta OutputClustering.txt >
OutputClusteringRF.fasta
cd ..
else
cd RFoutput
A=$(echo "$ITER" | rev | cut -f1 -d "," | rev)
A+="OutputRF.txt"
cut -f1 "${A}" > outputRF_names.txt
seqtk subseq ../"${2}".fasta outputRF_names.txt >
outputRF_contigs.fasta

```

```
cd ..
```

```
fi
```

```
fi
```

# Appendix 6 – Chapter 4 Supplementary material

Appendix Table 14 – list of the genomes used in the phylogenomic and functional analyses

NCBI code	Name
GCA_000283535.1	<i>Pseudogulbenkiania</i> NH8B
GCA_000812585.1	<i>Chromobacterium piscinae</i>
GCA_001458475.1	<i>Kingella kingae</i>
GCA_001808185.1	<i>Jeongeupia</i> USM3
GCA_001809035.1	<i>Eikenella</i> HMSC061C02
GCA_001855575.1	<i>Chromobacterium sphagni</i>
GCA_002002905.1	<i>Aquaspirillum</i> LM1
GCA_002108505.1	<i>Neisseria dumasiana</i>
GCA_002216145.1	<i>Neisseria</i> 10023
GCA_002237445.1	<i>Neisseria</i> KEM232
GCA_002327085.1	<i>Neisseria</i> 10022
GCA_002848345.1	<i>Chromobacterium</i> ATCC53434
GCA_002863305.1	<i>Neisseria perflava</i>
GCA_002892535.1	<i>Neisseriaceae</i> bacterium DSM100970
GCA_002951955.1	<i>Vitreoscilla</i> C1
GCA_003013245.1	<i>Neisseria iguanae</i>
GCA_003201855.1	<i>Rivicola pingtungensis</i>
GCA_003269145.1	<i>Microvirgula</i> AG722
GCA_003325475.1	<i>Chromobacterium</i> IIBBL1121
GCA_003325495.1	<i>Chromobacterium</i> IIBBL2741
GCA_003355495.1	<i>Crenobacter</i> K1W11S77
GCA_003443515.1	<i>Chromobacterium rhizoryzae</i>
GCA_003538525.1	<i>Neisseriales</i> UBA11063
GCA_003633895.1	<i>Vogesella indigofera</i>
GCA_003952345.1	<i>Iodobacter</i> H11R3
GCA_003963245.1	<i>Neisseriaceae</i> AWTP126
GCA_004328945.1	<i>Aquitalea</i> USM4
GCA_004341385.1	<i>Uruburuella suis</i>
GCA_004363805.1	<i>Paludibacterium purpuratum</i>
GCA_004919095.1	<i>Crenobacter</i> GY70310
GCA_005048205.1	<i>Chitiniphilus</i> HX215
GCA_900090205.1	<i>Vogesella</i> LIG4
GCA_900094895.1	<i>Snodgrassella</i> R53583
GCA_900230205.1	<i>Alysiella filiformis</i>
GCA_900322255.1	<i>Prolinoborus fasciculus</i>
GCA_900451175.1	<i>Kingella potus</i>
GCA_900451195.1	<i>Iodobacter fluviatilis</i>
GCA_900636515.1	<i>Neisseria animalis</i>
GCF_000005845.2	<i>Escherichia coli</i>
GCF_000006845.1	<i>Neisseria gonorrhoeae</i>
GCF_000007705.1	<i>Chromobacterium violaceum</i>
GCF_000008805.1	<i>Neisseria meningitidis</i>
GCF_000011545.1	<i>Burkholderia pseudomallei</i>
GCF_000020105.1	<i>Neisseria gonorrhoeae</i>
GCF_000021025.1	<i>Laribacter hongkongensis</i>
GCF_000025705.1	<i>Sideroxydans lithotrophicus</i>
GCF_000090875.1	<i>Neisseria oraltaxon014</i>

GCF_000093025.1	<i>Methylothera versatilis</i>
GCF_000158615.1	<i>Eikenella corrodens</i>
GCF_000160435.1	<i>Kingella oralis</i>
GCF_000163775.2	<i>Simonsiella muelleri</i>
GCF_000173875.1	<i>Neisseria mucosa</i>
GCF_000173895.1	<i>Neisseria cinerea</i>
GCF_000173955.1	<i>Neisseria subflava</i>
GCF_000174355.1	<i>Pseudogulbenkiania ferrooxidans</i>
GCF_000174655.1	<i>Neisseria sicca</i>
GCF_000175275.1	<i>Neisseria flavescens</i>
GCF_000176735.1	<i>Neisseria polysaccharea</i>
GCF_000186165.1	<i>Neisseria mucosa</i>
GCF_000190695.1	<i>Kingella denitrificans</i>
GCF_000193775.1	<i>Neisseria polysaccharea</i>
GCF_000194925.1	<i>Neisseria bacilliformis</i>
GCF_000196295.1	<i>Neisseria lactamica</i>
GCF_000213535.1	<i>Kingella kingae</i>
GCF_000220865.1	<i>Neisseria macacae</i>
GCF_000224255.1	<i>Neisseria weaveri</i>
GCF_000226875.1	<i>Neisseria shayeganii</i>
GCF_000227765.1	<i>Neisseria wadsworthii</i>
GCF_000297055.2	<i>Sulfuricella denitrificans</i>
GCF_000335715.1	<i>Vogesella</i> LIG4
GCF_000374805.1	<i>Chitiniphilus shinanonensis</i>
GCF_000376945.1	<i>Leeia oryzae</i>
GCF_000382305.1	<i>Vitreoscilla stercoraria</i>
GCF_000420525.1	<i>Aquaspirillum serpens</i>
GCF_000422925.1	<i>Paludibacterium yongneupense</i>
GCF_000425565.1	<i>Tepidiphilus margaritifer</i>
GCF_000428145.1	<i>Chitinilyticum litopenaei</i>
GCF_000428465.1	<i>Chitinimonas koreensis</i>
GCF_000428785.1	<i>Conchiformibius kuhniae</i>
GCF_000429665.1	<i>Azovibrio restrictus</i>
GCF_000429785.1	<i>Chitinibacter tainanensis</i>
GCF_000430805.1	<i>Chitinilyticum aquatile</i>
GCF_000470375.1	<i>Kingella kingae</i>
GCF_000527175.1	<i>Pseudogulbenkiania</i> MA1
GCF_000600005.1	<i>Snodgrassella alvi</i> wKB2
GCF_000620105.1	<i>Microvirgula aerodenitrificans</i>
GCF_000620145.1	<i>Deefgea rivuli</i>
GCF_000620925.1	<i>Conchiformibius steedae</i>
GCF_000695545.1	<i>Snodgrassella alvi</i> wKB29
GCF_000695565.1	<i>Snodgrassella alvi</i> wKB12
GCF_000711875.1	<i>Andreprevotia chitinilytica</i>
GCF_000711885.1	<i>Chromobacterium haemolyticum</i>
GCF_000735045.1	<i>Ferrovum myxofaciens</i>
GCF_000745895.1	<i>Stenoxybacter acetivorans</i>
GCF_000745955.1	<i>Alysiella crassa</i>
GCF_000751855.1	<i>Kingella negevensis</i>
GCF_000758475.1	<i>Chromobacterium haemolyticum</i>
GCF_000799095.1	<i>Chitinibacter</i> ZOR0017
GCF_000800415.1	<i>Neisseria meningitidis</i>
GCF_000813705.1	<i>Morococcus cerebrosus</i>
GCF_000818035.1	<i>Neisseria elongata</i>
GCF_000952105.1	<i>Chromobacterium violaceum</i>
GCF_000964065.1	<i>Aquitalea magnusonii</i>
GCF_000969645.2	<i>Janthinobacterium</i> B98

GCF_000971355.1	<i>Chromobacterium vaccinii</i>
GCF_001020585.1	<i>Chromobacterium subtsugae</i>
GCF_001027865.1	<i>Neisseria arctica</i>
GCF_001037925.1	<i>Vogesella</i> EB
GCF_001043555.1	<i>Chromobacterium</i> LK1
GCF_001043705.1	<i>Chromobacterium</i> LK11
GCF_001063455.1	<i>Morococcus cerebrosus</i>
GCF_001063965.1	<i>Neisseria bacilliformis</i>
GCF_001294205.1	<i>Amantichitinum ursilacus</i>
GCF_001302325.1	<i>Gulbenkiania mobilis</i>
GCF_001308005.1	<i>Neisseria</i> 74A18
GCF_001418035.1	<i>Gulbenkiania indica</i>
GCF_001457815.1	<i>Vitreoscilla massiliensis</i>
GCF_001515285.1	<i>Aquitalea magnusonii</i>
GCF_001515305.1	<i>Aquitalea pelogenes</i>
GCF_001517245.1	<i>Gulbenkiania indica</i>
GCF_001590725.1	<i>Snodgrassella</i> CFCC13594
GCF_001592185.1	<i>Bergeriella denitrificans</i>
GCF_001619695.1	<i>Crenobacter luteus</i>
GCF_001619865.1	<i>Chromobacterium</i> F49
GCF_001648335.1	<i>Eikenella corrodens</i>
GCF_001648345.1	<i>Eikenella</i> NML01A086
GCF_001648355.1	<i>Eikenella</i> NML02A017
GCF_001648395.1	<i>Eikenella</i> NML03A027
GCF_001648415.1	<i>Eikenella</i> NML070372
GCF_001648425.1	<i>Eikenella</i> NML080894
GCF_001648435.1	<i>Eikenella</i> NML120348
GCF_001648475.1	<i>Eikenella</i> NML130454
GCF_001648495.1	<i>Eikenella</i> NML96A049
GCF_001648505.1	<i>Eikenella</i> NML97A109
GCF_001648535.1	<i>Eikenella</i> NML990057
GCF_001676875.1	<i>Chromobacterium subtsugae</i>
GCF_001811325.1	<i>Eikenella</i> HMSC071B05
GCF_001813335.1	<i>Neisseria</i> HMSC055H02
GCF_001815615.1	<i>Neisseria</i> HMSC075C10
GCF_001815685.1	<i>Neisseria</i> HMSC056A03
GCF_001855275.1	<i>Chromobacterium vaccinii</i>
GCF_001855555.1	<i>Chromobacterium sphagni</i>
GCF_001855565.1	<i>Chromobacterium amazonense</i>
GCF_002022745.1	<i>Neisseria lactamica</i>
GCF_002073715.2	<i>Neisseria sicca</i>
GCF_002081815.1	<i>Chromobacterium haemolyticum</i>
GCF_002081825.1	<i>Chromobacterium haemolyticum</i>
GCF_002088735.1	<i>Snodgrassella alvi</i> A112
GCF_002108495.1	<i>Neisseria canis</i>
GCF_002108575.1	<i>Neisseria zoodegmatis</i>
GCF_002108595.1	<i>Neisseria dentiae</i>
GCF_002108605.1	<i>Neisseria animaloris</i>
GCF_002213445.1	<i>Xenophilus</i> AP218F
GCF_002215055.1	<i>Laribacter hongkongensis</i>
GCF_002217795.2	<i>Aquitalea magnusonii</i>
GCF_002222655.1	<i>Vitreoscilla filiformis</i>
GCF_002735645.1	<i>Iodobacter</i> BJB302
GCF_002735695.1	<i>Chitinimonas</i> BJB300
GCF_002777425.1	<i>Snodgrassella alvi</i> App48
GCF_002777745.1	<i>Snodgrassella alvi</i> WF33
GCF_002777825.1	<i>Snodgrassella alvi</i> Nev42

GCF_002777855.1	<i>Snodgrassella alvi</i> wkB298
GCF_002777865.1	<i>Snodgrassella alvi</i> PEB0171
GCF_002803635.1	<i>Neisseria</i> N17716
GCF_002847985.1	<i>Neisseria perflava</i>
GCF_002863285.1	<i>Neisseria sicca</i>
GCF_900086555.1	<i>Neisseria weaveri</i>
GCF_900113545.1	<i>Neisseria elongata</i>
GCF_900115065.1	<i>Formivibrio citricus</i>
GCF_900119825.1	<i>Chitinimonas taiwanensis</i>
GCF_900143275.1	<i>Nitrosomonas cryotolerans</i>
GCF_900169325.1	<i>Kingella denitrificans</i>
GCF_900176275.1	<i>Andreprevotia lacus</i>
GCF_900177275.1	<i>Pseudogulbenkiania subflava</i>
GCF_900177895.1	<i>Kingella negevensis</i>
GCF_900187105.1	<i>Eikenella corrodens</i>
GCF_900187305.1	<i>Neisseria zoodegmatis</i>

# Appendix 7 – Other project



## Targeted genotyping by sequencing: a new way to genome profile the cat

M. Longeri\* , A. Chiodi<sup>†</sup>, M. Brilli<sup>‡,§</sup>, A. Piazza<sup>§,¶</sup>, L. A. Lyons<sup>\*\*</sup>, G. Sofronidis<sup>††</sup>,  
M. C. Cozzi\* and C. Bazzocchi<sup>\*,§,‡‡</sup>

\*Department of Veterinary Medicine, University of Milan, Milano 20133, Italy. <sup>†</sup>Department of Earth and Environmental Sciences, University of Pavia, Pavia 27100, Italy. <sup>‡</sup>Department of Biosciences, University of Milan, Milano 20133, Italy. <sup>§</sup>Paediatric Clinical Research Centre “Romeo ed Enrica Invernizzi”, University of Milan, Milano 20157, Italy. <sup>¶</sup>Department of Biomedical and Clinical Sciences “L. Sacco”, University of Milan, Milano 20157, Italy. <sup>\*\*</sup>Department of Veterinary Medicine and Surgery, College of Veterinary Medicine, University of Missouri, Columbia, MO 65211, USA. <sup>††</sup>Orivet Genetic Pet Care, Suite 102/163-169 Inkerman Street, St. Kilda, Vic. 3182, Australia. <sup>‡‡</sup>Coordinated Research Centre “EpiSoMI”, University of Milan, Milano 20133, Italy.

### Summary

Targeted GBS is a recent approach for obtaining an effective characterization for hundreds to thousands of markers. The high throughput of next-generation sequencing technologies, moreover, allows sample multiplexing. The aims of this study were to (i) define a panel of single nucleotide polymorphisms (SNPs) in the cat, (ii) use GBS for profiling 16 cats, and (iii) evaluate the performance with respect to the inference using standard approaches at different coverage thresholds, thereby providing useful information for designing similar experiments. Probes for sequencing 230 variants were designed based on the *Felis catus* 8.0 genome. The regions comprised anonymous and non-anonymous SNPs. Sixteen cat samples were analysed, some of which had already been genotyped in a large group of loci and one having been whole-genome sequenced in the 99\_Lives Cat Genome Sequencing Project. The accuracy of the method was assessed by comparing the GBS results with the genotypes already available. Overall, GBS achieved good performance, with 92–96% correct assignments, depending on the coverage threshold used to define the set of trustable genotypes. Analyses confirmed that (i) the reliability of the inference of each genotype depends on the coverage at that locus and (ii) the fraction of target loci whose genotype can be inferred correctly is a function of the total coverage. GBS proves to be a valid alternative to other methods. Data suggested a depth of less than 11× is required for greater than 95% accuracy. However, sequencing depth must be adapted to the total size of the targets to ensure proper genotype inference.

**Keywords** DNA profiling, *Felis catus*, genotyping-by-sequencing, single nucleotide polymorphisms

### Introduction

The global pet care market size (major segments including food, veterinary care and over-the-counter products) was estimated at USD 131.7 billion in 2016 and is expected to reach USD 202.6 billion by 2025, an estimated growth of 4.9% calculated with the Compound Annual Growth Rate

(<https://www.grandviewresearch.com/press-release/global-pet-care-market>, March 2018). Cats are increasingly appreciated as pets because they are known for helping reduce stress and anxiety and for having strong interactions with humans (Hart *et al.* 2018). In this context, fancy breeds are becoming more and more popular worldwide. In pedigreed cats, studbooks can recommend a DNA-based control of both animal identity and traits of interest for enrolment and selection, together with a permanent electronic identification. The International Society of Animal Genetics (ISAG) fosters the definition and nomenclature standardization of panels of genetic markers for the identification and parentage control of domestic animals, including cats (Lipinski *et al.* 2007; <https://www.isag.us/committees.asp>). These panels are used by service laboratories for owners and

#### Address for correspondence

M. Longeri, Department of Veterinary Medicine, University of Milan, Milano 20133, Italy.  
E-mail: maria.longeri@unimi.it

M. Longeri and A. Chiodi are contributed equally.

Accepted for publication 15 July 2019

© 2019 The Authors. *Animal Genetics* published by John Wiley & Sons Ltd on behalf of Stichting International Foundation for Animal Genetics  
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.