EUDES GUILHERME VIEIRA BARBOSA

ON THE POWER AND LIMITS OF COMPUTATIONAL FUNCTIONAL GENOMICS FOR BACTERIAL LIFESTYLE PREDICTION

Tese apresentada ao Programa de Pós-Graduação em Bioinformática, Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do título de Doutor em Bioinformática.

Orientador: Prof. Dr. Vasco Ariston de Carvalho Azevedo
Co-orientador: Prof. Dr. Jan Baumbach

BELO HORIZONTE

2016

EUDES GUILHERME VIEIRA BARBOSA

ON THE POWER AND LIMITS OF COMPUTATIONAL FUNCTIONAL GENOMICS FOR BACTERIAL LIFESTYLE PREDICTION

Tese apresentada ao Curso Programa de Pós-Graduação em Bioinformática, Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do título de Doutor em Bioinformática.

Orientador: Prof. Dr. Vasco Ariston de Carvalho Azevedo
Co-orientador: Prof. Dr. Jan Baumbach

BELO HORIZONTE

2016

FICHA CATALOGRÁFICA

Vieira Barbosa, Eudes Guilherme.

On The Power and Limits of Computational Functional
Genomics for Bacterial Lifestyle Prediction– Belo Horizonte, 2016.

Nº de páginas 159

Área de concentração: Bioinformática.

Orientador: Prof. Dr. Vasco Ariston de Carvalho Azevedo.

Tese de doutorado – Programa de Pós-Graduação em Bioinformática, Universidade Federal de Minas Gerais.

1.Bioinformatics; 2. Machine Learning; 3. Actinobacteria; 4. Lifestyle.

ACKNOWLEDGMENTS

First of all, I would like to thank professors Vasco Ariston de Carvalho Azevedo and Jan Baumbach for given me this opportunity and supporting me in the most various ways during this time. It definitely wasn't an easy journey, so I'm very thankful for their patience.

I'm also very grateful for my family's support, specially my mom and my aunt "Ju". Without these two women it would be impossible for me to pursuit a life in Science.

Special thanks go to all my colleagues from the Laboratory of Cellular and Molecular Genetics (Brazil) and from the Computational Biology group (Denmark); also, to Sebastian Böcker's group in Jena, where I spent my exchange time.

I also would like to thank all my friends from Brazil for putting up with the distance and time; and my friends in Denmark for making life "hyggeligt". Further, I'm also deeply thankful for the support of two important Brazilian and Danish institutions, respectively the "Buteco da Bio" and the "Fredagsbaren".

Finally, I thank all gods that ever existed, exist and one day shall be invented.

“Don’t panic!”
Douglas Adams – The Hitchhiker's Guide to the Galaxy.

SUMMARY

FIGURE LIST

RESUMO

ON THE POWER AND LIMITS OF COMPUTATIONAL FUNCTIONAL GENOMICS FOR BACTERIAL LIFESTYLE PREDICTION

Bactérias são organismos ubíquos; elas estão presentes onde quer que a vida seja possível. Diferentes bactérias são capazes de se ajustar a diversos estilos de vida, por exemplo, elas podem estar associadas a hospedeiros ou ter um estilo de vida livre. Portanto, esses organismos devem possuir um grande e variado arsenal genômico para lidar com diferentes condições ambientais. Nós desenvolvemos duas abordagens ara investigar o repertório genético que talvez esteja associado a um estilo de vida. Ambas combinam analises evolutivas das sequencias com aprendizado estatístico (Random Forest com seleção de variáveis, ajuste de modelo e analise de robustez). Inicialmente, nós procuramos por genes homólogos que pudessem distinguir entre diferentes classes de patogenia de Actinobactérias. Nós incluímos 240 actinobacterias classificadas em quatro classes de patogenia: patógenos humanos (HP), patógenos de amplo espectro (BP), patógenos oportunistas (OP), e não patogênicos (NP). Essencialmente, nós encontramos genes homólogos que podem computacionalmente distinguir entre patógenos e não patogênicos. Além disso, nos demonstramos um claro limite na diferenciação entre patógenos oportunistas de ambos não patogênicos e patógenos. Patógenos humanos talvez não possam ser diferenciados de bactérias anotadas como de amplo espectro baseando-se apenas em um pequeno numero de genes ortólogos, uma vez que, muitos patógenos humanos podem também apresentar uma ampla variedade de hospedeiros mas não ter a devida anotação. Por ultimo, nós introduzimos a ferramenta LiSSI (LifeStyle-Specific-Islands) para facilitar a identificação de componentes genéticos que possam facilitar na adaptação de bactérias a um nicho especifico. O pipeline da ferramenta é uma extensão da nossa abordagem anterior. Resumidamente, nossa estratégia procura identificar sequencias conservadas de genes homólogos (ilhas) em genomas, e identificar as ilhas características de cada estilo de vida. Para ilustrar as suas principais funcionalidades, nós expandimos a nossa busca de apenas classes de patógenos para também incluir tolerância a oxigênio atmosférico (aeróbico, anaeróbico, facultativo) e habitat (solo e aquático). Essencialmente, nós descobrimos que ilhas parecem ter um peso menor na classificação. Aparentemente há pouca conservação da ordem genética entre as espécies bacterianas, sendo que genes individuais são mais úteis para classificação. Concluindo, nós demonstramos que mesmo na era pós-genômica e a despeito das tecnologias de sequenciamento de próxima geração,

nossa habilidade de chegar a conclusões efetivas permanecem bem limitadas. Além disso, nós apresentamos LiSSI, um ferramenta de bioinformática para identificação de assinaturas genéticas ou ilhas (sequencias conservadas de genes homólogos) para distinguir estilos de vida bacterianos.

ABSTRACT

ON THE POWER AND LIMITS OF COMPUTATIONAL FUNCTIONAL GENOMICS FOR BACTERIAL LIFESTYLE PREDICTION

Bacteria are ubiquitous organisms; they can be found wherever life is possible. Distinct bacteria are able to coop with highly diverse lifestyles; for instance, they can be classified as host associated or free living. Therefore, these organisms must possess a large and varied genomic arsenal to withstand different environmental conditions. To investigate the genetic repertoire that might be associated with a given lifestyle, we developed two approaches. Both methodologies combine evolutionary sequence analysis with statistical learning methods (Random Forest with feature selection, model tuning and robustness analysis). Initially, we searched for homologous gene sets that could distinguish Actinobacterial pathogenicity classes. We included 240 Actinobacteria classified to four pathogenicity classes: human pathogens (HP), broad-spectrum pathogens (BP), opportunistic pathogens (OP), and non-pathogens (NP). Essentially, we found homologous gene sets that computationally distinguish pathogens from non-pathogens. We further show a clear limit in differentiating opportunistic pathogens from both non-pathogens and pathogens. Human pathogens may also not be distinguished from bacteria annotated as broad-spectrum pathogens based on a small set of orthologous genes only, as many human pathogens could target a broad range of mammals but have not been annotated accordingly. Finally, to facilitate the identification of genomic features that might influence bacterial adaptation to a specific niche, we introduce LifeStyle-Specific-Islands (LiSSI). The LiSSI pipeline is an expansion of our previous strategy. In summary, our strategy aims to identify conserved consecutive homology sequences (islands) in genomes and to identify the most discriminant islands for each lifestyle. To illustrate the main functionalities, we expanded our search from exclusively pathogenic classes to include tolerance to atmospheric oxygen (aerobe, anaerobe, facultative) and habitat (soil and aquatic). Essentially, we found that islands seem to carry less weight in the classification performance. It seems that gene order is poorly conserved among bacterial species, which might make individual genes more useful as classifiers. In conclusion, we illustrate that even in the post-genome era and despite next-generation sequencing technology, our ability to efficiently deduce real-world conclusions, such as pathogenicity classification, remains quite limited. Further, we introduce LiSSI, a bioinformatics pipeline, in order to identify signature genes or islands (conserved consecutive homology sequences) that distinguish bacterial lifestyles.

1 BACKGROUND

1 BACKGROUND

For 30 years, sequencing technologies based on Sanger chemistry dominated the market. Although Sanger methodology had undergone numerous improvements over the years, gene cloning techniques were still necessary to obtain genomic DNA sequences. Therefore, the time and cost required to obtain a complete genome sequence remained high. Moreover, the capacity of parallel sequencing was quite limited (Shendure, Mitra et al. 2004, Shendure, Porreca et al. 2005, Richardson 2010). Next-generation sequencing (NGS) platforms made it possible to sequence complete prokaryotic genomes using massively parallel sequencing more rapidly and at a lower cost (Shendure, Porreca et al. 2005, Munroe and Harris 2010).

As with any methodology, NGS presents its own drawbacks. It generates large numbers of reads, but considerably smaller and, therefore, less informative than those produced by Sanger methodology. The length of the reads makes it difficult to completely assemble a genome using exclusively computational tools (Miller, Koren et al. 2010, Klassen and Currie 2012). The main limitation of short-read assembly methods is their inability to resolve repetitive regions of the genome without paired libraries (Miller, Koren et al. 2010). The assembly of repetitive regions was an important issue even before the introduction of NGS platforms; shorter reads only made the problem worse.

In 2001, Kececioglu and Yu argued about the impossibility of correctly assembling genomic regions that contain identical copies of a sequence (Kececioglu and Ju). Usually, long DNA repeats are not exact copies. They contain small differences that could, in principle, permit their correct assembly. Nevertheless, a major difficulty arises from sequencing errors. Assembly software must accept imperfect sequencing alignments to avoid missing genuine connections between sequences (Miller, Koren et al. 2010). With the smaller length of reads plus the inherent sequencing error, it is difficult to separate true differences within repeated sequences from sequencing errors.

A study by Phillippy and collaborators revealed that the majority of contig ends in draft genomes were associated with repeated regions (Phillippy, Schatz et al. 2008). They concluded that it was possible to categorize the majority of mis-assembly events into two general classes: i) repeat collapse or expansion and ii) sequence rearrangement and inversion. Each of these classes exhibits specific mis-assembly signatures: the first class results from incorrect assembly in repetitive regions, including fewer or additional copies; the second class results from the rearrangement of multiple repeated copies, which is caused by the insertion of a read

between them. The second class may be considered more influential because, if not fixed, it might be interpreted as a real biological rearrangement event (Ricker, Qian et al. 2012, Soares, Abreu et al. 2012). If the assembler cannot resolve the region between two genomic fragments, a gap is formed. Gaps may occur due to: i) an intrinsic characteristic of the sequencing platform that leads to incomplete or incorrect information or ii) the inability of an assembly algorithm to handle regions of low complexity or repeated DNA (Chain, Grafham et al. 2009, Pop 2009, Tsai, Otto et al. 2010). The process of identifying and closing these gaps is quite laborious and requires additional manual intervention.

In recent years, approaches using hybrid assemblies have been developed to facilitate the genome assembly process. These techniques take advantage of the high-quality reads of second-generation sequencers (e.g., the Illumina Genome Analyzer) and the longer read lengths of third-generation sequencers (e.g., SMRT sequencers by Pacific Biosciences and the Ion Torrent PGM) (Bashir, Klammer et al. 2012, Ribeiro, Przybylski et al. 2012). Although empirically logical, this type of approach was not facilitated by the lack of integration between sequencers. Virtually no bioinformatics system has been developed to integrate reads from different sequencers into a single assembly (Diguistini, Liao et al. 2009, Bashir, Klammer et al. 2012). This newly developed approach aims to reduce the amount of manual intervention needed to complete a genome sequence by using a hybrid approach to resolve repetitive regions.

1.1 VALUE OF A NEWLY SEQUENCED GENOME

Given the success of various whole-genome sequencing projects over the last decade, we have nowadays thousands of bacterial genome sequences available, for instance, with NCBI (Coordinators 2014). After assembly, post-processing and annotation also require a high level of bioinformatics support. Essentially, one utilizes the evolutionary conservation of the genetic repertoire to predict the genes' function through sequence similarity comparisons, for instance, by integrating the popular BLAST software (Altschul, Madden et al. 1997) into special purpose genome annotation platforms, such as CoryneRegNet (Baumbach and Apeltsin 2008, Pauling, Rottger et al. 2012), GenDB (Meyer, Goesmann et al. 2003) or RAST (Aziz, Bartels et al. 2008), just to mention a few. With the emergence of the so-called next-generation sequencing technology, the available data sets exploded such that we have >61,000 sequencing projects at NCBI, with 5,107 whole-genome bacterial sequences available (NCBI web site, Mai 22, 2016). Figure 1 depicts the growth of

genome deposits in GenBank from 2005, when NGS sequencers were introduced, to 2016.

In this section, I introduce a discussion about the “scientific value” of a newly sequenced genome and the amount of insight it can provide. Thereafter, I review the main characteristics of the bacterial genomes, how they might influence evolution and the ability to coop with different lifestyles. Further, I introduce a discussion over the balance between genome conservation and gene novelty.

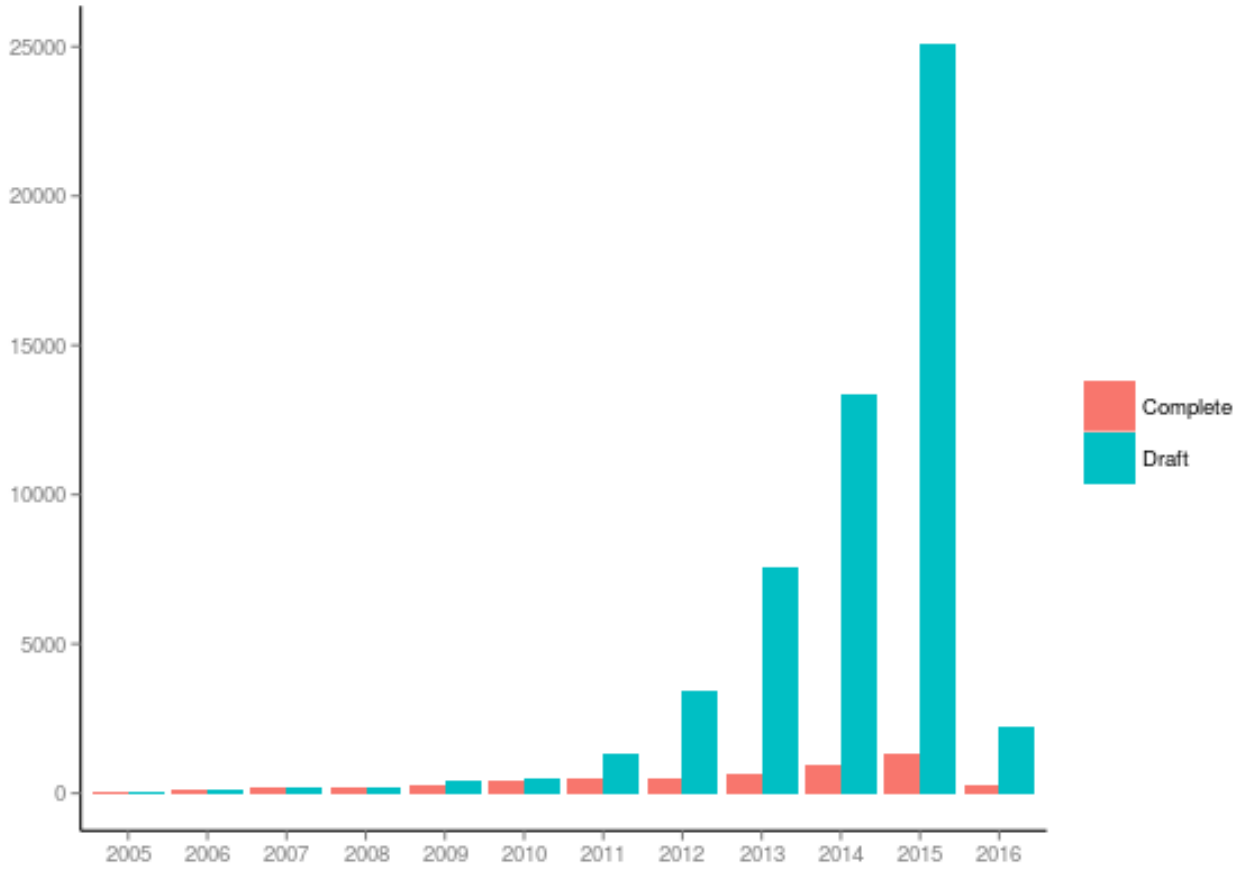

FIGURE 1 – GENBANK GENOME DEPOSITS: 2005-2016. NOTE THAT THE NUMBER OF COMPLETE GENOME DEPOSITS GROWS IN A LINEAR WAY, WHILE DRAFT (PARTIAL INFORMATION) GROWS EXPONENTIALLY. INFORMATION AS AVAILABLE ON MAI 2016.

1.1.1 PUBLICATION IMPACT

The value of a newly sequenced genome can be assessed using many different metrics. If publications are considered the main “currency” within the scientific community, there has been a considerable decrease in the value of new sequences over the last four decades.

The introduction of Sanger methodology in 1977 was one of the main landmarks in the early stages of the genomic era (Sanger, Nicklen et al. 1977). During the first years of using Sanger sequencing, a sequence of no more than 1,000 nucleotides was sufficient for a work to be accepted in a journal such as Cell (current

impact factor: 32.40) or Nature (current impact factor: 36.28) (de Boer, Gilbert et al. 1979, Nakamura and Inouye 1979, Porter, Barber et al. 1979). In 1980, the shotgun DNA sequencing methodology was introduced, enabling the sequencing of longer DNA fragments (Porter, Barber et al. 1979). Complete bacterial operons were sequenced and published in journals such as Molecular Microbiology (current impact factor: 5.01) and Proceedings of the National Academy of Sciences (PNAS - current impact factor: 9.68) (Porter, Barber et al. 1979, Postle and Good 1983, Overduin, Boos et al. 1988).

A combination of DNA sequencing improvements and the newly developed TIGR Assembler (Sutton, White et al. 1995) culminated in the publication of the first complete bacterial genomes in 1995. Papers containing the complete nucleotide sequences of *Haemophilus influenzae* Rd (1,830,137 base pairs) and *Mycoplasma genitalium* (580,070 base pairs) were both published in Science (current impact factor: 31.20) (Fleischmann, Adams et al. 1995, Fraser, Gocayne et al. 1995). Almost 20 years later, a paper containing the sequence of a prokaryotic genome alone may be published in the Genome Announcement section of the Journal of Bacteriology (current impact factor: 3.825) or in Standards in Genomic Sciences (SIGS - current impact factor: 3.167). A recent article by Smith even refers to the not-so-distant “death” of the “genome paper”, noting that the space for genome publication may soon come to an end (Smith 2013).

The publication impact of newly sequenced genomes decreased following DNA sequencing improvements, and the reason is no mystery. High-impact journals only publish groundbreaking original scientific research or results of outstanding scientific importance. To produce a higher-impact publication, more information must be extracted from genomes. For instance, several genomes may be examined in a comparative genomic analysis or pangenomic study (Medini, Donati et al. 2005, Soares, Silva et al. 2013), or an analysis may focus on the presence or absence of specific markers or on small differences between DNA sequences (Ricker, Qian et al. 2012, Jakobsen, Hansen et al. 2013). In this context, the genome becomes a stepping stone to the main goal, the comparative analysis. As the basis of the analysis, the genome sequence remains important. Nevertheless, it may not be of sufficient importance for one to undertake the painstaking task of completing the genome sequence.

1.1.2 IMPACT ON VACCINE DEVELOPMENT

The increasing amount of available genomic information was expected to boost the development of vaccines. In an attempt to measure the impact of genomic information on this field, Prachi and collaborators (Prachi, Donati et al. 2013) analyzed all the patent applications that contained genomic information. They observed that there was an enormous increase in such applications shortly after the first complete genomes were released, but since 2002, there has been a continuous decrease. The authors attributed this decrease to more stringent legal requirements, which call for empirical evidence to complement *in silico* data.

The initial increase in patent applications containing genomic information was related to the development of a new paradigm in vaccine development. In 2000, Rappouli described the “reverse vaccinology” (RV) concept, in which he proposed inverting the traditional process of antigen identification (Rappuoli 2000). Instead of identifying the antigenic components of a pathogenic organism using serological or biochemical methods, RV uses the organism’s genome to predict all of its protein antigens. RV approaches mainly focus on secreted proteins because they are more likely to induce immune responses. Secreted proteins are involved in several processes that modulate the host-pathogen relationship, such as cell adhesion and invasion, as well as resistance to stress conditions (Stavrinides, McCann et al. 2008, Simeone, Bottai et al. 2009, Wooldridge 2009). Over the years, several methodologies have been developed to predict secreted proteins and to evaluate their potential immunological properties.

In 2010, Vaxign was released as the first vaccine design tool with a web interface (http://www.violinet.org/vaxign/). Vaxign allows users to submit their own sequences to perform vaccine target predictions. The Vaxign predictions have been consistent with existing reports for organisms such as *Mycobacterium tuberculosis* and *Neisseria meningitides* (He, Xiang et al. 2010). Another vaccine design tool is MED (Mature Epitope Density), it attempts to select the more promising vaccine targets by identifying proteins with higher concentrations of epitopes (Santos, Pereira et al. 2013). There are also tools exclusively for protein epitope prediction, such as Immune Epitope Analysis (http://tools.immuneepitope.org/main/) and Vaxitope (http://www.violinet.org/vaxign/vaxitop/index.php).

Due to the fact that a large number of bacterial genomes are already available, RV is quite accessible and inexpensive. Nevertheless, as has been previously discussed (Tettelin 2009, Seib, Zhao et al. 2012), the expectations for RV techniques do not correspond to reality. The relatively small number of vaccines developed using this methodology indicates that other factors play a major role in the

host immunological response (Wirth, Hildebrand et al. 2008, Donati and Rappuoli 2013).

1.1.3 IMPACT ON ANTIBACTERIAL DISCOVERY

The period between the 1930s and the 1960s is known as the “golden age” of antibiotic discovery (Walsh 2003, Mills 2006). During this period, most of the known classes of antibiotics were discovered. These discoveries involved screening natural products regardless of their mechanisms of action. After most of the low-hanging fruits were harvested, the rate of antibacterial discovery decreased, culminating in a slowdown beginning in the 1990s (Silver 2011).

Hopes for turning this void into a rapid acceleration accompanied the completion of the first bacterial genome sequences. The goal was to use comparative genomic analysis to identify potential targets present in a desirable spectrum (e.g., the bacteria responsible for upper respiratory tract infections) (Mills 2006, Pucci 2006). It was naive to assume that having the genome sequences would be sufficient for this level of discovery; a possible drug target must undergo numerous stages from discovery to human clinical tests, and it is not possible to develop drugs for all potential targets (Pucci 2006, Payne, Gwynn et al. 2007). Nevertheless, the prospect of exploring hundreds of potential targets revived the interest of pharmaceutical companies.

After some years of trials, several companies ended their target-based programs due to lack of productivity. Despite reports of multi-resistant bacterial strains, the efforts to discover new antibacterial targets were again reduced (Projan 2003, Bush, Courvalin et al. 2011). Although genomics has not been able to reverse the lack of new antibiotic development, it has significantly improved screening methodologies. Genomics has facilitated high-throughput drug campaigns, which are being used to determine the mechanisms of action of antibacterial compounds and bacterial resistance mechanisms (Mills 2006).

1.2 BACTERIAL GENOME

One of the most distinctive characteristics of the Bacteria group is the lack of a membrane isolating the genetic material from the cytoplasm. Instead, Bacteria present a region known as nucleoid (or genophore), where all or most of the genetic material and its associated molecules are located (Griffiths 2005). The packing of the genetic material around the nucleoid must address two potentially conflicting aspects.

Not only must it compact the DNA within the cell, it must also allow for access of genes for expression and regulation, plus, rapid genome replication (Dorman 2013).

The bacterial genome is simple and tightly packed with genes. Bacterial genomes are small and vary by more than one order of magnitude, ranging from approximately 500 thousand to 10 million bases (Ochman and Davalos 2006). Due to several processes, including rearrangements, gene duplication or loss, and horizontal gene transfer, bacterial genomes are extremely variable in terms of gene repertoires. Conversely, their structural features are highly conserved (Ochman, Lawrence et al. 2000, Rocha 2008). Valens and colleagues (Valens, Penaud et al. 2004) described six distinct structural zones in the *E. coli* chromosome. Their results showed that DNA interactions, and subsequently rearrangements, were restricted to sub-regions of the DNA. That might suggest that chromosome structuring is a potential constrain for genome evolution (Esnault, Valens et al. 2007).

Given the limited amount of space available in bacterial genomes, the process of gene gain is generally counterbalanced by gene loss. In the following portions of the text, I will review the several processes that lead to gene gain, focusing mainly on horizontal gene transfer, and briefly describe the gene loss events. Finally, I will close this topic with a discussion of the balance between genome conservation and gene novelty.

1.2.1 GENE GAIN

There are several mechanisms that can lead to gene gain among bacteria: transformation, in which the bacteria incorporates extracellular DNA to the genome; transduction, in which the exogenous DNA is packaged in a bacteriophage; and conjugation, in which the DNA is transferred by mating (Griffiths 2005). These processes are generally labelled as lateral gene transfer (LGT), to differentiate them from the generational (vertical) transfer of genes (Soares, Abreu et al. 2012). In all three mechanisms, the donor DNA is delivered and incorporated in the recipient's cell genome. There is growing evidence that LGT plays a major role in bacterial genome evolution, leading to environmental adaptation and speciation (Ochman, Lawrence et al. 2000, Soares, Abreu et al. 2012). In many cases, the transferred pieces of DNA have a considerable length, containing several genes and are called genomic islands (GIs) (Waack, Keller et al. 2006, Soares, Abreu et al. 2012). Although there is no biological evidence to support this claim, the community has established that a GI has at least 8 genes or 8 kilobases (Langille, Hsiao et al. 2010).

GIs create an unusual similarity between the donor and the recipient strain. They retain sequence characteristics of the donor genome, such as GC content, codon usage and/or di- and tri-nucleotide distribution. In addition, we often observe the remains of translocatable elements, transfer origins of plasmids or known attachment sites to integrases adjacent to regions identified as GIs (Ochman, Lawrence et al. 2000, Waack, Keller et al. 2006, Soares, Abreu et al. 2012).

GIs may be classified according to their genomic content: symbiotic Islands, which might be involved in bacteria and Leguminosae plant family association (Barcellos, Menna et al. 2007); resistance Islands, which have genes related to antibiotic resistance (Krizova and Nemec 2010); metabolic Islands, which have genes associated with secondary metabolic biosynthesis (Tumapa, Holden et al. 2008); pathogenic Islands, which have a high concentration of genes related to virulence or pathogenicity and are involved in the re-emergence of several pathogens (Dobrindt, Janke et al. 2000).

There are genomic barriers to LGT: donor-recipient similarity, ecological and functional. Popa and colleagues showed in (Popa, Hazkani-Covo et al. 2011) that most of the detected LGTs occur between closely related species from the same taxonomic group. In a subsequent study, the same group showed that clusters of densely connected donors and recipients are quite similar in terms of GC content. This finding indicates that a biological barrier for gene acquisition from donors of dissimilar genomic GC content exists (Popa and Dagan 2011). The ecological barrier relates to the distance between organisms, because conjugation and transformation are influenced by the donor-recipient distance. For conjugation to occur, both organisms must be close enough for the formation of the conjugation tunnel. While transformation depends on DNA stability in the environment in order to occur. Both observations further suggest that most transfers occur within habitats (Popa and Dagan 2011). The final barrier to DNA acquisition is functional. As the bacterial genome has limited size and it is continually passing through a dynamic process of incorporating genomic material, sequences with little or no contribution to cell fitness are more likely to disappear again (Ochman, Lawrence et al. 2000, Popa and Dagan 2011).

1.2.2 GENE LOSS

The process of gene loss frequently involves the formation of pseudogenes as an intermediary step. The term pseudogene designates sequences that present high similarity with functional genes as well as genetic defects that preclude the

formation of functional products. The genetic defects can be frameshifts or the insertion of premature stop codons (Gerstein and Zheng 2006).

Given the characteristics of pseudogenes, the likelihood of finding one in a bacterial genome are consider to be fairly low, if any are to be found at all. That view started to change in 2001, when the complete genome sequence of *Mycobacterium leprae* was released (Eiglmeier, Parkhill et al. 2001). Only 49.5% of the genome contains coding regions. In the remaining part, 27 consist of identifiable pseudogenes; 23.5% represent non-coding regions that might correspond to regulatory sequences or the remains of pseudogenes that are too degraded to be identified.

Pseudogenes can be mainly created by three processes: inactivation of duplicated sequences, inactivation of unique sequences and unsuccessful horizontal gene transfer (Liu, Harrison et al. 2004). In prokaryotic genomes, there is a continuous process of pseudogene creation, decay and eventual removal from the genome due to the accumulated mutations. The fact that closed related species and strains share few pseudogenes suggests that the time spam between gene inactivation and removal from the genome is fairly short (Liu, Harrison et al. 2004, Lerat and Ochman 2005, Kuo and Ochman 2010). In 2010, Kuo and Ochman found evidence that degraded genes might be actively removed from the genome through an adaptive process (Kuo and Ochman 2010). The sequences could indeed be harmful for the organism due to the high transcriptional and translational costs of a non-functional protein and/or the generation of toxic products.

1.2.3 BALANCE BETWEEN GENOME CONSERVATION AND GENE NOVELTY

Bacterial chromosome architecture is subject to a balance between genetic novelty and stability of the gene arrangement in the chromosome. While genetic novelties have great influence in adaptation, the introduction of new genes tends to disrupt the chromosome organization. The trade-off between these two processes depends on bacterial niche and lifestyle (Rocha 2004). Furthermore, gene order conservation usually involves two categories of genes: rare and persistent, where the mechanisms that led to each kind are not identical. In summary, conservation cannot be explained in all instances by operons and lateral gene transfer (Fang, Rocha et al. 2008).

Throughout the years, several models were developed to explain gene order conservation (see (Lawrence 1999) for a review). The latest models are the Co-regulation Model (CM) and the Selfish Operon Model (SOM). CM is based on the

observation that genes that are found close together on the chromosome can be regulated efficiently. Therefore, genes involved in the same metabolic pathway or the same protein complex would present selective advantages when clustered. This model leads to the conclusion that operons are the origin of the cluster organization in bacterial chromosomes. The main problem with CM is that it fails to explain the selective advantages of gene proximity while co-transcription still not possible. SOM is based on lateral gene transfer. The model states that if a set of genes provides equivalent fitness (independent of their position), physical proximity provides an advantage to the genes themselves. In this case, clustered genes present advantage against spread ones while being transferred. Therefore, genes can be gradually moved close together even before co-transcription is possible (Lawrence and Roth 1996, Pál and Hurst 2004).

1.3 LIFESTYLES

Different environments, habitats, energy sources, and niches (short: lifestyles) require particular characteristics from bacterial species that will survive, reproduce and proliferate. Hence, one can observe various genome-sizes and mobile DNA elements associated with different lifestyles (Ochman and Davalos 2006, Newton and Bordenstein 2011). In this section, I will review the particular characteristics of organisms associated with different lifestyles in terms of pathogenicity, oxygen consumption, habitat, and growth temperature.

1.3.1 PATHOGENICITY

Usually, pathogens have smaller genomes with a bias towards gene loss. This bias is essentially explained by the abundance of intermediate metabolic compounds provided by the host. Thus, several metabolic pathways are no longer under selective pressure and thus no longer subjected to a process of decay and elimination from the genome (Moran 2002). The bias may further be explained by genetic drift, as pathogens usually require a small inoculum to infect a new host. This may lead to a population size reduction, where even useful genes may be lost by chance (Ochman and Davalos 2006).

Opportunistic pathogens are organisms that are usually not associated with diseases but can become pathogenic for individuals with compromised immune systems (Berg, Eberl et al. 2005). There are two possible explanations for the presence of virulence factors in organisms that are not pathogenic per se. First,

some gene products that allow for pathogenicity behaviour confer advantages to free-living organisms as well (Casadevall 2006). Genes associated with antibiotic resistance, for instance, are commonly found in bacteria living in areas of intense microbiological activity (Casadevall 2006). Second, a bacterium may be an animal pathogen with a host as yet to be discovered. This alternative is known as “cryptic pathogenesis” (Casadevall and Pirofski 2007).

In contrast to pathogens, non-pathogens cannot depend on a stable environment and the abundance of nutrients provided by the host. Soil-dwelling bacteria, for instance, must quickly adapt to extreme conditions such as exposure to sunlight and dehydration. Furthermore, they must be able to handle different or even constantly changing sources of nutrients (Casadevall 2006, Casadevall and Pirofski 2007, Görke and Stülke 2008, Rohmer, Hocquet et al. 2011). Therefore, these organisms must possess a larger genomic arsenal to withstand varying environmental conditions.

1.3.2 OXYGEN CONSUMPTION

The presence of atmospheric oxygen is a limiting factor for bacterial growth; specifically, oxygen levels cannot exceed those found in a bacterium's native habitat (Imlay 2013). Above these levels bacteria are subject to decrease in population growth – and ultimately death – due to the harmful effects of oxidation caused by superoxide and hydrogen peroxide in cellular component (Gutteridge 1994, Imlay 2013). During oxidative stress, lipids are the major target, leading to alterations in membrane fluidity and potentially disrupting membrane-bound proteins. Further, modifications in proteins can lead to conformational changes and consequently loss of function. Finally, another main target is the DNA, leading to single- or double-strand breaks and in extreme cases blocking replication by cross-linking the DNA to other molecules (Sies and Menck 1992, Cabiscol, Tamarit et al. 1999). Regarding oxygen tolerance, bacteria can be divided into three broad groups: aerobes, facultative and anaerobes.

Aerobes are defined as organisms that require atmospheric oxygen conditions (roughly 20%) to achieve optimal growth. The overhead associated with an oxidative environment is compensated by enabling aerobic respiration, a pathway substantially more efficient than fermentation (Poole and Cook 2000). Aside from the presence of a metabolic pathway that can use oxygen as the final electron acceptor, other features are ubiquitous among these organisms, such as enzymes that

degrade peroxide (catalases and peroxidases) (Pahl and Baeuerle 1994, Imlay 2013). Other metabolic features are also expected to be found to prevent oxidative agents formation, plus, mechanisms to repair oxidative damage and eliminate damaged molecules (Gutteridge 1994).

Facultative organisms can grow in atmospheric oxygen conditions or in the absence of oxygen. To perform both cellular respiration and fermentation, these organisms most pay the costly price of having both metabolic systems. This disadvantage is compensated for by the diversity of habitats that these organisms can occupy, including habitats with rapidly changing oxygen conditions (Unden, Becker et al. 1995). To sense the availability of oxygen and control the switch from respiration to fermentation, these organisms present a diversity of transcriptional regulators (e.g., FNR) (Unden, Becker et al. 1995).

Anaerobic organisms are defined as organisms that can tolerate at most low amounts of atmospheric oxygen and are not capable of performing cellular respiration. Organisms of this class lack the mechanisms for cellular respiration and to protect the cellular components against oxidative damage (Morris and Schmidt 2013). It is not clear which genes might be either exclusive or essential for this class of organism (Müller-Herbst, Wüstner et al. 2014).

1.3.3 HABITAT

Bacteria can also be classified according to the habitat in which they can be found. In a broad sense, bacteria can be found in the soil, freshwater or in marine habitats; where an incredibly high abiotic and biotic set of conditions can be found. For instance, these habitats can be further divided into oligotrophs, environments with low level of nutrients, and copiotrophs, environments rich in nutrients (Koch 2001). Although it is highly unlikely to find single traits that define such broad classes (Livermore, Emrich et al. 2014), we opt not to explore these subdivisions.

Soil bacteria present an enormous diversity; the richness of “species” that can be found in a gram of soil ranges from approximately 26 thousand to 8 million (Gans, Wolinsky et al. 2005, Roesch, Fulthorpe et al. 2007). Regardless of the methodological disagreements that might lead to different estimations, the complexity and diversity of this environment is undeniable. Thus, substantial efforts have been made to identify genetic features that might explain why some taxons are more abundant in particular types of soil (Fierer, Bradford et al. 2007, Barberán, Ramirez et al. 2014).

Aquatic environments (freshwater and marine) present continuous gradients for nutrients, oxygen concentration, and luminosity. Thus, similarly to soil bacteria, attempts to find discriminative genetic features for these habitats also stumble upon their underlying complexity (Lauro, McDougald et al. 2009, Livermore, Emrich et al. 2014). Furthermore, Livermore and colleagues (Livermore, Emrich et al. 2014) showed that there is a significant overlap between freshwater genetic features and those found in both soil and marine organisms. The authors justify their findings in the fact that freshwater is an intermediate step in between the two extremes, soil and marine.

1.3.4 TEMPERATURE

Bacteria can also be classified according to their optimal growth temperature, in particular when researchers are interested in bacteria growing at the range extremes: thermophiles and psychrophiles. Given the extreme conditions faced by these organisms it is expected to find severe modifications in their proteins and metabolic pathways.

Thermophilic organisms have an optimal growth temperature ranging from above 60°C to almost the point of ebullition (Wang, Cen et al. 2015). Their genomes are all smaller than 4 Mb and tend to have a higher GC content than non-thermophilic organisms (Wu, Zhang et al. 2012, van Noort, Bradatsch et al. 2013). Thermophilic genomes are highly influenced by lateral gene transfer (Wang, Cen et al. 2015). A study with species of *Caldanaerobacter subterraneus* revealed that lateral gene transfer plays a major role in their genome, corresponding to roughly 45-60% of the genes (Sant'Anna, Lebedinsky et al. 2015).

Proteins from thermophilic organisms have special characteristics; they are usually smaller and present fewer protein family members when compared to their homologous counterparts (Wang, Cen et al. 2015). Plus, there are studies that indicate that adaptation to high temperatures is concentrated on proteins with catalytic and regulatory activities (Gu and Hilser 2009).

Psychrophilic organisms are metabolically active at temperatures below 5°C; there is evidence that DNA replication occurs at temperatures lower than -20°C (Margesin and Feller 2010, Tuorto, Darias et al. 2014). The fact that 80% of Earth's environments are permanently below 5°C degrees (most oceans, areas within the Arctic Circle, montane regions, among others) makes psychrophiles geographically widely distributed (De Maayer, Anderson et al. 2014). Psychrophilic genomes present a high level of redundancy, with multiple copies of tRNA species for all amino acids

and a large number of chaperones (Math, Jin et al. 2012). It is also possible to find genes associated with antifreeze proteins, which are responsible for lowering the freezing point (Celik, Drori et al. 2013).

Proteins from psychrophilic organisms do not suffer the same damaging conditions as thermophilic proteins, which partially explains their diversity (Margesin and Feller 2010). Furthermore, their enzymes present a higher level of flexibility to decrease activation energy and increase the substrate conversion rate (Rodrigues and Tiedje 2008).

2 STATE OF THE ART

2 STATE OF THE ART

2.1 PROTEIN HOMOLOGY IDENTIFICATION

Clustering is a computer science method that partitions data objects into groups such that the objects share common traits; elements within the groups are more similar to each other than to objects from other groups. In our case, clustering is used to identify functionally related proteins, i.e., homologous proteins. Through the years, several methods have been developed and applied to address this issue, for instance: k-means, affinity propagation, Markov clustering, and FORCE, as well as transitivity clustering (Enright and Ouzounis 2000, Enright, Van Dongen et al. 2002, Paccanaro, Casbon et al. 2006, Frey and Dueck 2007, Wittkop, Emig et al. 2010).

Transitivity clustering is based on exact and heuristic algorithms for solving the Weighted Transitive Graph Projection (WTGP) problem, also known as weighted graph cluster editing (Rahmann, Wittkop et al. 2007). WTGP starts by creating an all-vs.-all BLAST of all proteins under analysis and then transforms the matrix into a weighted and undirected graph, where the nodes correspond to the biological entities (genes/proteins) and the weighted edges correspond to the similarities. Given a similarity threshold (density parameter), it removes the edges below this cut-off and seeks to transform the (potentially) intransitive graph into a disjoint set of cliques with minimal cost (based on a similarity function) for edge additions/deletions. This problem is NP-hard but guarantees that the average similarity inside the clusters is below the threshold, while the average similarity between objects from different clusters is above the threshold.

It is worth noting that Transitivity Clustering scales approximate quadratically with the input, which renders it unsuitable for low cut-off values (density parameter) in large data-sets. In this case, one should consider alternative greedy algorithms, such as UCLUST (Edgar 2010) and CD-HIT (Fu, Niu et al. 2012). UCLUST is based on an algorithm that combines k-mer lookup and compressed alphabets, where similarities can be identified in linear time. The authors claim that this improves alignment speed without considerable loss of accuracy (Edgar 2004). On the other hand, CD-HIT uses a word filtering algorithm; it can determine if two sequences share a certain similarity without aligning them. One advantage of CD-HIT is that it can make unlimited use of the computer’s RAM and can be parallelized (Li and Godzik 2006, Fu, Niu et al. 2012).

2.2 GENETIC ISLAND IDENTIFICATION

2.2.1 CONSERVATION BASED

The unusual similarity between the donor and the recipient strain for a specific region of their genomes is one of the features that can be utilized for GI identification. Although sequence comparison and phylogenetic distribution analyses are useful for detecting LGT, the DNA sequences themselves provide intrinsic information to determinate if a region is originated either from vertical (ancestral) or lateral gene transfer. Most of the few existing approaches for GI identification rely on such intrinsic characteristics of the sequence: deviations in GC content, codon usage and di- and tri-nucleotide distributions. They usually have a high rate of false-negative and false-positive predictions (undetected islands with similar sequence composition and clusters of highly expressed genes that are falsely identified as islands, respectively). Some tools also scan for the remains of translocatable elements, transfer origins of plasmids or integrase attachments sites in regions close to the potential GI (Ochman, Lawrence et al. 2000). The following three tools are most frequently applied nowadays: IslandPath (Waack, Keller et al. 2006), IslandPicker (Langille, Hsiao et al. 2008), PIPS (Soares, Abreu et al. 2012), and GiPSy (Soares, Geyik et al. 2015).

2.2.2 DE NOVO

Several algorithms have been developed to identify islands (conserved consecutive homology sequences) in bacterial genomes using a *de novo* approach (Landau, Parida et al. 2005, Böcker, Jahn et al. 2009, Jahn 2011). The main challenge faced by these algorithms is that islands are not perfectly conserved. Therefore, the tools must accept approximate islands with, for instance, gene deletion or inversion. To address that issue a tool named Gecko was developed (Jahn 2011); it identifies islands in a large number of genomes. It utilizes a strategy based on reference occurrences; it sets one genome as reference and detects approximate islands in all other genomes (the procedure is repeated for all analysed genomes). It uses three main parameters: maximum distance between islands (i.e., deletions or insertions), minimum island size and minimum number of genomes that the island is supposed to be present. Furthermore, Gecko estimates the statistical significance of the island, i.e., the probability of observing a given island with an equivalent or higher degree of conservation in multiple genomes. It assumes as a null hypothesis that the gene order is random (Jahn, Winter et al. 2013).

2.2.3 CLASSIFICATION AND FEATURE SELECTION USING RANDOM FOREST

In 2001, Breiman presented the Random Forest (RF) methodology; in the original paper, he defined random forests as "a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" (Breiman 2001). In summary, RF generates many de-correlated trees (weak learns) and aggregates their results to create a strong predictor or classifier. The main idea behind this approach is to use the diversity of the base learns to explore possible hypotheses; namely, each tree will be constructed using a different bootstrapped sample of the data, and each node will be split using the best predictor among a randomly chosen subset. In the case of a classifier, the final class label prediction is based on majority vote; while a regression averages the trees' output (Breiman 2001, Liaw and Wiener 2002, Rogers and Gunn 2006).

One of RF's main characteristics is the use of bagging to construct the predictors. It uses bootstrapping which leaves out approximately one-third of the data for a given tree (Hastie, Tibshirani et al. 2005). This portion of the data is known as "out-of-bag" (OOB), and it allows the evaluation of a given subset without the use of test sets (Rogers and Gunn 2006). The OOB can be used to estimate the error by checking the mean decrease in accuracy, it compares OOB observations with permuted OOB observation to report a given variable importance measure. The idea is that replacing a feature that contributes to correct predictions due to noise should noticeably decrease the accuracy, while the performance of an irrelevant feature should not be affected by being replaced by noise (Svetnik, Liaw et al. 2004, Archer and Kimes 2008, Kursa 2014).

Another important intrinsic importance measure is the Gini Index. The Gini Index reflects the node impurity for the splits using a given feature, i.e., the homogeneity of the two descendent nodes. The descent nodes are compared to the original node and attributed a value between zero (homogeneous) and one (heterogeneous). The final Gini Index value is given by the sum and normalization of all decreases in all splits, where higher values indicate features that led to higher purity nodes [76, 77]. However, a possible source of concern regarding Gini Index is that it may be biases, for instance, towards features with more categories or with fewer missing values (Breiman, Friedman et al. 1984, Strobl, Boulesteix et al. 2007, Sandri and Zuccolotto 2012).

Given RF's characteristic random exploration of features, it can also be utilized for feature selection (FS) (Rogers and Gunn 2006). While dealing with FS

researchers ultimately have one out of two possible goals: finding the “all relevant” or the “minimal optimal” subset. The first approach aims at finding genes for subsequent studies, thus, it is acceptable to include genes that are correlated or that present similar molecular functions. The former approach aims at classification performance, thus, the goal is to find the smallest subset of genes that performs comparable to the whole dataset (Díaz-Uriarte and De Andres 2006, Kursa 2014). In 2007, Nilsson and colleagues proved the intractability of the “all relevant” problem and also that a backward elimination algorithm is sufficient to find asymptotically optimal solutions for the “minimal optimal” problem (Nilsson, Peña et al. 2007).

There are essentially three types of algorithms available for FS: the filter model, the wrapper model, and the hybrid model (Yu and Liu 2003, Hapfelmeier and Ulm 2013). The filter model avoids using any learning algorithm by using general characteristics of the data for filtering. The wrapper model uses a learning algorithm to select the features. It usually provides better results than the filter model, but, as for each subset it creates a new classifier, it tends to be computationally more expensive. The hybrid model combines the previous models; it starts by using a learning method to create a reference quality measure using the whole data-set, followed by successive interactions where the least important features are removed (Yu and Liu 2003, Hapfelmeier and Ulm 2013).

In this project, I chose to use a methodology based on the hybrid model, **varSelRF** (Diaz-Uriarte 2007, Diaz-Uriarte 2009). It is available as an R package, and it was conceived to successively and aggressively remove non-important features. As with other hybrid models, it computes features importance only once (based on OOB error). This approach might lead to overfitting, which is why the error is also assessed using the .632+ predictor (Hapfelmeier and Ulm 2013). The .632+ bootstrap method takes this name from the fact each sample will on average contain roughly 0.632 distinct observations. Generally speaking, it is a “smoothed version” of the cross-validation, with important advantages: reduced variability of error rate prediction and assessment of the variability for the estimated parameters (Efron and Tibshirani 1997). At the end, the tool returns a very small set of features that preserves the classification accuracy. Also, due to their redundancy, highly correlated features are removed from the final result. Nevertheless, as is the case of the other available tools, varSelRF cannot guarantee stability or multiplicity of the selected genes (Diaz-Uriarte 2007, Diaz-Uriarte 2009).

In conclusion, RF has several advantages that have made it a popular methodology in life sciences and made it suitable for this project. **I)** RF doesn’t require predictor transformation, facilitating the interpretability of the results. That is a

great advantage when compared to other methods, such as the Support Vector Machine or Neural Network, which are fairly useful for the proposed classification but do not easily allow for assessment of the most important features (Breiman 2001, Díaz-Uriarte and De Andres 2006, Archer and Kimes 2008, Kursa 2014). Also, as was already mentioned, it has intrinsic importance measures (e.g., Gini Index). **II)** RF provides good support for FS due to both its random exploration of the data as well as its importance measures (Rogers and Gunn 2006). In our case, we were particularly interested in the smallest subset (putative homologous genes or islands) that could define a given lifestyle. **III)** RF has three main parameters: the number of variables randomly sampled at each node (mtry), the minimal size of terminal nodes (nodesize), and the number of trees in the forest (ntree). According to (Díaz-Uriarte and De Andres 2006), changes in these parameters usually have “negligible effects”, for instance, it was demonstrated that the default setting of mtry is already quite sensible and often a good choice in terms of the OOB error rate. Further, the number of trees should be large enough to stabilize the statistic of interest, although is worth noting that the time required to run all computations increases approximately in a linear way to the increase in the number of trees (ntree) (Svetnik, Liaw et al. 2004, Díaz-Uriarte and De Andres 2006). **IV)** RF deals comparably well with p >> n datasets (in our case: number of clusters/islands >> number of lifestyles). This class of problem is known to be associated with instability and the low statistical power of certain methods. Specifically, several genes may present the same information level leading to datasets full of highly redundant feature (Archer and Kimes 2008, Kursa 2014).

3 HOMOLOGOUS GENE ANALYSIS

3 HOMOLOGOUS GENE ANALYSIS

To understand the genetic repertoire distinctive for lifestyle-specific adaptation processes during evolution, we use the phylum Actinobacteria and important pathogenicity classes for a case study. We concentrate our efforts on Actinobacteria due to the fact the group contains well-studied microbial species of high importance in medicine, biotechnology and environmental research. Further, Actinobacteria is one of the largest clades of bacteria, and species from this group emerged various lifestyles and populate diverse habitats [148, 149].

We consider all fully sequenced actinobacterial species that classify into one of the four pathogenicity lifestyle classes: (HP) exclusively human pathogenic; (BP) broad-spectrum pathogenic (mammals, including humans); (OP), opportunistic pathogenic (bacteria that usually do not cause disease in a healthy host); and (NP) non-pathogens (e.g., soil inhabitants and gastrointestinal tract inhabitants). Related work concentrates on the identification of virulence (i.e., pathogenicity-specific) genes. Here, we are more fine-grained and distinguish between four different classes. In contrast to existing studies [150, 151], we also find non-pathogenicity specific genes.

In this section, we investigated the power and limits of using genetic features to predict pathogenic lifestyles in these bacteria. Therefore, we specifically hypothesize:

(H1) Pathogens (HPs and BPs) possess specific pathogenicity signature genes not present in non-pathogens (NPs) but in most, preferably all, pathogens.

(H2) Similarly, most opportunistic pathogens (OPs) possess specific pathogenicity signature genes that are not present in non-pathogens (NPs).

(H3) Broad-spectrum and exclusively human pathogens (BPs and HPs) cannot be distinguished from each other due to a prospective observation bias: while HPs have only human as host, BPs have dozens of possible hosts such that we may assume that HPs might as well be BPs although they have never been classified as such (lack of resources).

(H4) There is no intrinsic genomic characteristic of opportunistic pathogens (OPs) compared to pathogens (HPs and BPs), as all of them need to interact with host cells such that small short-term mutations are likely to play a more dominant role in order to survive the immune system [46].

3.1 METODOLOGY

3.1.1 GENOMES AND PATHOGENICITY CLASSES

We selected all 240 completely sequenced actinobacterial genomes that belonged to one of the four pathogenicity classes: HP, BP, OP, or NP. All lifestyles were manually curated by scanning the literature. We exclude symbiotic and plant pathogens as well as not fully sequenced species for this study. This resulted in 68 HPs, 27 BPs, 22 OPs and 123 NPs. The whole-genome annotation was downloaded from NCBI and 926,573 coding gene DNA sequences were extracted and stored in FASTA format. For the complete list of species and respective pathogenicity classification see S. Table 1 in Appendix A.

3.1.2 HOMOLOGY DETECTION

We first performed computational homology detection following a protocol suggested recently by Röttger *et al.* in [152]. It was used to obtain clusters of homologous gene products in actinobacteria of the so-called CNMR sub-classes using a combination of BLAST and Transitivity Clustering [153]. We followed the same steps as in [152] and applied BLAST [154] to our 926,573 protein-coding genes all-versus-all (E-value cutoff of 0.01) to obtain a pairwise similarity matrix. In this matrix, the similarity values were converted into the –log10 of the best achieved pairwise BLAST E-value. An E-value of 10^{-53} for two proteins A and B would consequently result in a similarity of 53 between them: *similarity*(A,B) = 53. Transitivity Clustering transformed this similarity matrix into a weighted similarity graph, where genes and similarities were considered as nodes and weighted edges, respectively. The software used a similarity cutoff (so-called density parameter) and removed all edges below this value. Afterwards, the potentially intransitive graph were transformed into a transitive one by adding and removing edges with minimal edge modification costs (Weighted Cluster Editing problem, see [153] for details). Transitivity Clustering ensures that the average similarity between clusters is below the cutoff while the average similarity between genes from the same cluster is above the threshold [155]. The methodology has proven robust for predicting clusters of homologous genes and proteins based on pairwise BLAST results. In accordance with Röttger *et al.* [152], a similarity threshold of 48 is most reasonable for actinobacterial species, which corresponds to an E-value cutoff of 10^{-48}. We therefore applied Transitivity Clustering to the BLAST results of the 926,573 gene sequences and clustered them with a density parameter of 48 into 227,412 groups of

homologous genes. The size of groups ranged from 1 to 557, whereas 98.7% of them contained less than 50 genes, which is in accordance with the actinobacterial genomic diversity (Figure 2). We subsequently removed all clusters of sizes <5 and finally end up with 28,627 groups of homologous genes for 240 actinobacteria. Note that with these steps we followed precisely the microbial homology detection protocol from Röttger *et al.* [152].

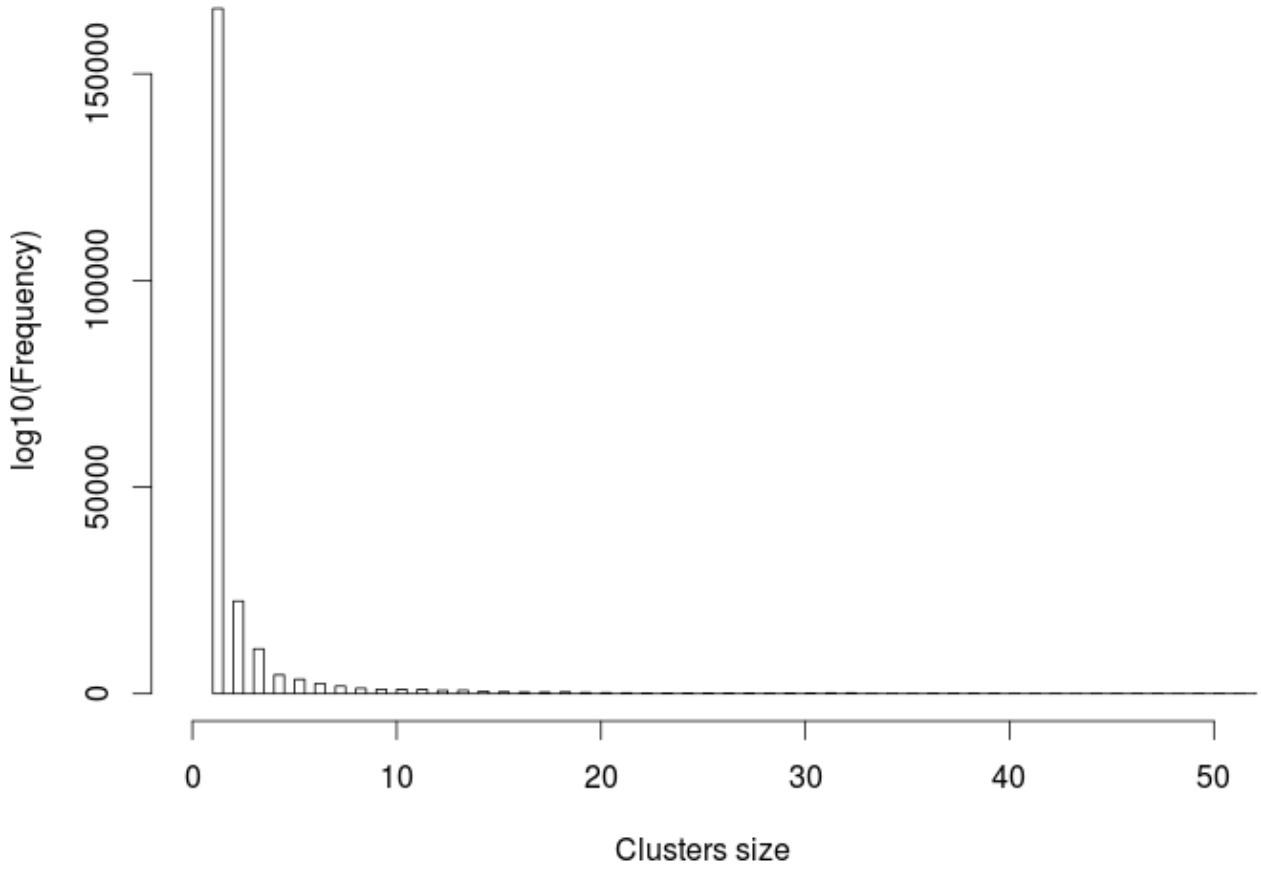

FIGURE 2 – CLUSTER SIZE DISTRIBUTION.

3.1.3 STATISTICAL LEARNING OF LIFESTYLE-SPECIFIC GENES

Scripts for the statistical learning software environment R were developed to inspect the distribution of the clusters of homologous genes amongst the different lifestyles, pathogenicity classes in our case, and to identify those that are distinctive for them

We started with a visual analysis depicted exemplarily for hypothesis H1 in Figure 3. The aim is to scan for the genetic repertoire distinctive for pathogens (HP and BP) or non-pathogens (NP), respectively. We therefore investigate the joint distribution of the homologous gene clusters between the two classes: HP+BP vs.

NP. Considering the large number of points to be plotted (28,627 gene clusters), the R library Hexbin was used. Hexbin refined and facilitated the visualization by plotting fixed-size hexagonal bins colored based on the density of points in a given area of the graph. This allowed us to inspect the joint distributions of the genetic repertoire of different sets of organisms from different lifestyle classes (see legend of Figure 3).

Clusters of homologous genes close to the axis tails in the joint distribution plots are highly class-specific and not species-specific. As depicted in Figure 3, there is no such cluster, neither close to the tail of the x-axis (NP-specific) nor to the y-axis (HP+BP-specific). Consequently, there is no single homologous gene that is specific for either of the two classes.

In order to identify sets homologous genes that may together distinguish the two classes (NP vs. HP+BP), we needed to scan for sets of lifestyle-distinctive homologous gene clusters that together formed a decision tree allowing us to split the two groups of species. This way, the problem turned into a statistical learning problem with 28,627 gene clusters as features and 240 data objects, which were distributed over four classes (68 HPs, 27 BPs, 22 OPs and 123 NPs). In the specific case of H1/Figure 3, we have two classes, 123 objects in a class labeled NP (non-pathogen), and 95 elements in a class labeled HP+BP (pathogen). To identify a set of signature genes (or feature genes) for the lifestyle class HP+BP, for instance, we removed all gene clusters that are found more often in non-pathogens (NP) than in pathogens (HP+BP). This refers to all clusters below the dotted line in the upper plot of Figure 3. This way, our follow-up random forest models were biased towards utilizing pathogen-specific features (i.e. the homologous genes in the bottom left plot) for classification. We followed this protocol for all four hypotheses and generated four of such joint distribution plots.

We used the R package randomForest to generate Random Forest (RF) classifiers using lifestyle-specific features. Each tree was constructed using a different bootstrapped sample of the data, and each node was split using the best predictor among a randomly chosen subset. To access a robust quality estimation of our classifier, the data was evaluated using a 5-fold cross validation. Also, this procedure was repeated five times using different cross validation sets. We could therefore analyze the robustness of the classification towards changes in the homology data sets. Furthermore, we compared the emerging RF classifiers against the predictions performed with randomized labels. By using exactly the same classification and cross validation pipeline, we aimed to classify the data not into their real classes (using the real pathogenicity labels) but we assigned each organism a random pathogenicity label instead. We may assume a drastic drop in the

classification performance when classifying the data with random labels, preferably close to that of a random classifier (50% accuracy in a two-class learning problem). This allowed us to assess the classification robustness. For all classifiers, with real labels and with random labels, ROC (receiver operating characteristics) plots were generated to inspect their performance. Four quality measures were used to evaluate the results: area under the ROC curve (AUC), accuracy (ACC), sensitivity and specificity. Figure 4 illustrates two ROC curves for hypothesis H1 (i.e. NP vs. HP+BP). The five ROC curves for classifiers learned with real labels are in dark blue solid lines, while the random label classifier ROC curves are presented in light blue dashed lines. The variation of the AUCs in the 5-fold cross validation is depicted as box plots in the figure.

To evaluate the classification performance for each hypothesis, we define three measures “Accuracy”, “Unrobustness”, and “Influence of bias”. For each biased dataset with its 5-fold cross validation, we define the following:

1. The average AUC for classifying the real data:$\overline{AUC}$
2. The average AUC for classifying the random labeled data:$\overline{AUC}_{RL}$
3. The difference between the average AUCs: $\Delta\overline{AUC}$ = $|\overline{AUC} - \overline{AUC}_{RL}|$
4. The “Accuracy” is defined as the average AUC over both biases:$\overline{AUC}_{bias1} + \overline{AUC}_{bias2}/_2$
5. The “Unrobustness” of the classification performance is defined as the mean distance of the $\overline{AUC}_{RL}$ from the best possible value of 50:$|\overline{AUC}_{RL1} + \overline{AUC}_{RL2} - 100|/_2$

   It describes how likely the RF predictor would also predict random class labels instead of the real ones.
6. The “Influence of the bias” (towards one class or the other) is defined as $|\Delta\overline{AUC}_{bias1} - \Delta\overline{AUC}_{bias2}|$ and describes how much our bias introduction influences the classification performance.
7. For two classes of pathogenicity to be well separable, we require a high “Accuracy” (close to 100) and a low “Unrobustness” (close to 0). The “Influence of bias” describes whether we find class-specific genes for both pathogenicity classes (close to 0) or not (otherwise). Table 1 summarizes our findings for each hypothesis.

The R package varSelRF was used to identify the most discriminant features for each lifestyle. The package successively eliminated the least important variables

using the so-called out-of-bag error (RF internal error estimate) as minimization criterion [146]. Afterwards, RapidMiner version 5.3.015 was used to generate decision trees (see Figures 5 and 6 for illustration) by applying the so-called Gini Index as maximization criterion (and standard values otherwise). The clusters of homologous genes used in the tree's nodes were named by using a simple majority vote while scanning all gene product descriptions of the cluster.

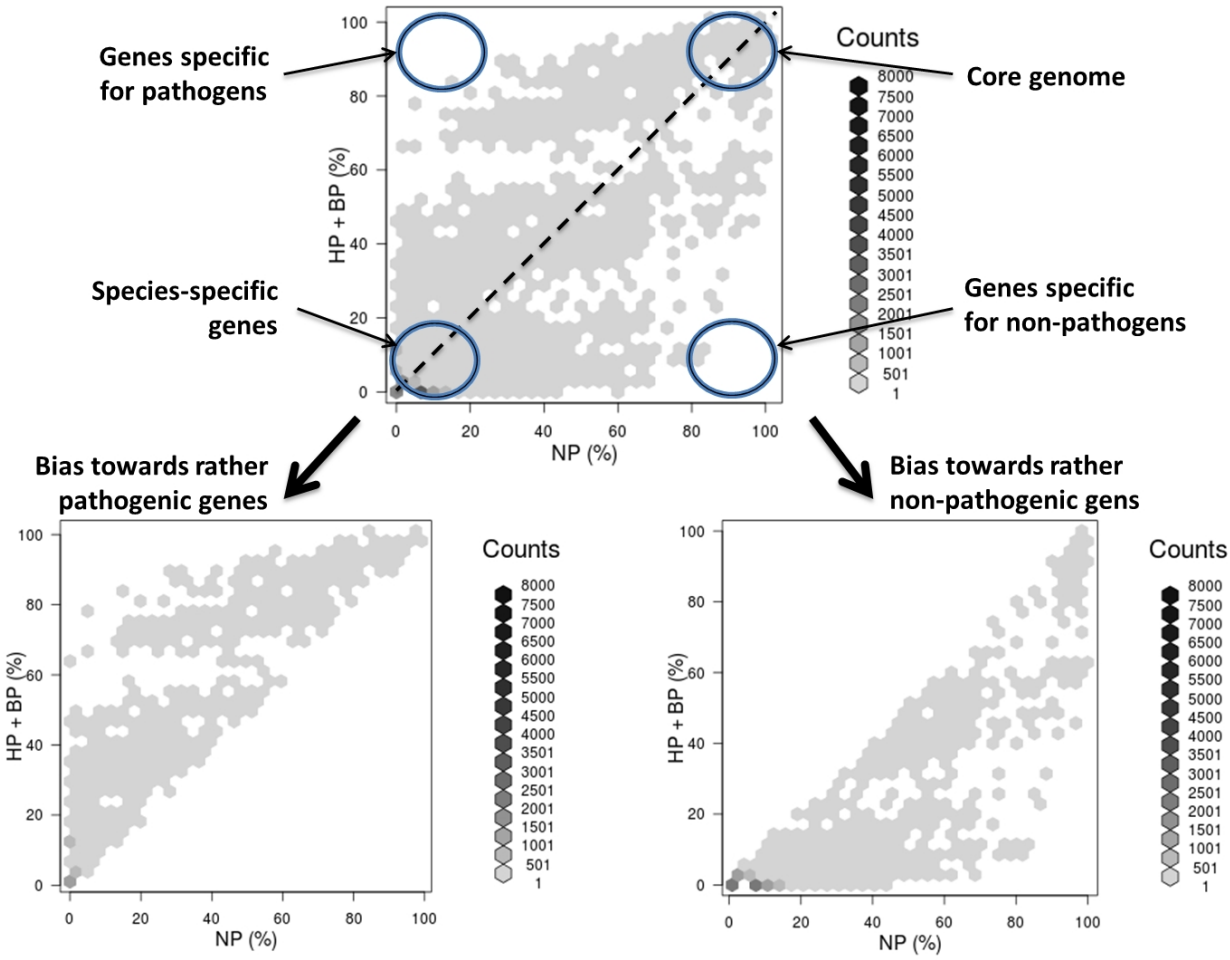

FIGURE 3 – ILLUSTRATION OF OUR BIAS INTRODUCTION STRATEGY. DISTRIBUTION OF HOMOLOGOUS GENE CLUSTERS OVER TWO LIFESTYLES (PATHOGENS VS. NON-PATHOGENS) AND ILLUSTRATION OF OUR STRATEGY TO INTRODUCE A FEATURE SELECTION BIAS INTO OUR STATISTICAL LEARNING PIPELINES. BOTH AXES IN ALL THREE PLOTS DESCRIBE THE PERCENTAGE OF SPECIES IN THE RESPECTIVE CLASS(ES), HERE HUMAN PATHOGENS (HP) AND BROAD PATHOGENS (BP) VS. NON-PATHOGENS (NP). THE COLOR-CODING OF THE HEAT MAP DEPICTS THE NUMBER OF CLUSTERS OF HOMOLOGOUS GENES THAT CERTAIN PERCENTAGES OF PATHOGENS/NON-PATHOGENS SHARE. THUS, IN THE UPPER RIGHT OF SUCH A JOINT DISTRIBUTION PLOT, WE FIND THE CORE GENOME (HOMOLOGOUS GENES PRESENT IN ALL SPECIES OF BOTH CLASSES); AND IN THE LOWER LEFT, WE SEE UNIQUE, SPECIES-SPECIFIC GENES. GENES CLOSE TO THE AXIS ARE MORE CLASS SPECIFIC. GENES CLOSE TO THE AXIS TAILS ARE HIGHLY CLASS SPECIFIC AND, THUS, THE DISTINCTIVE HOMOLOGOUS GENE CANDIDATES WE WERE HOPING TO FIND. AS THERE IS NO SINGLE SUCH GENE, WE SCANNED FOR SETS OF LIFESTYLE-DISTINCTIVE GENES. TO FIND SUCH FEATURE GENES FOR PATHOGENIC LIFESTYLES, FOR INSTANCE, WE REMOVE ALL GENES THAT ARE FOUND MORE OFTEN IN NON-PATHOGENS (NP) THAN IN PATHOGENS (HP+BP), I.E. THE GENE CLUSTERS BELOW THE DOTTED LINE IN THE UPPER PLOT, SUCH THAT OUR FOLLOW-UP MACHINE LEARNING ROUTINES ARE BIASED TOWARDS UTILIZING PATHOGENICITY-SPECIFIC FEATURES (GENES IN THE BOTTOM LEFT PLOT) FOR CLASSIFICATION.

3.2 RESULTS AND DISCUSSION

In the following, we use the classification performances to discuss our four hypotheses. For each of them, we are left with a two-class machine learning problem, which we split into two parts by introducing a pathogenicity bias (see Figure 3). This essentially leaves us with eight classifiers, whose performances we evaluate regarding their “Accuracy”, “Unrobustness” and the “Influence of bias”. Table 1 summarizes our findings, which we will discuss briefly in the following.

3.2.1 (H1) ALL PATHOGENS VERSUS NON-PATHOGENS

As expected, we were able to observe homologous genes exclusively found either in pathogens or non-pathogens (Figure 3). However, there is no group of homologous genes that is present in all or almost all (>90%) organisms of one lifestyle but not present in the other. Nevertheless, the classification of these two lifestyles shows by far the best performance on both biased data sets (Figure 4), towards pathogenic features and non-pathogenic features, with an “Accuracy” of 97.2% and an “Unrobustness” of only 0.9% (see Table 1 and its legend for definitions). The “Influence of bias” is almost inexistent (0.7%). The $\overline{AUC}$ for the pathogen classifier (NP vs HP+BP) was 96.7%, while the $\overline{AUC}$ for the non-pathogenic classifier (NP vs HP+BP) was 97.4%. These results indicate that we find both, gene sets specific for pathogens as well as gene sets specific for non-pathogens, with almost identical accuracy (remember that we introduce a bias in both directions, refer to Figure 3).

Non-pathogenic organisms must deal with constant environmental changes and different energy sources. It is expected that they present gene sets that are not necessarily found in pathogenic organisms. The best non-pathogenic discriminant features were used to generate the random tree in Figure 5. The best features were the genes “Threonine dehydratase”, “Beta-galactosidase” and “ATP-dependent DNA helicase”, respectively. Using only these three genes we may already obtain 89.9% classification accuracy. Note that all three genes are associated with metabolic processes.

In contrast, pathogens must possess genes, which support their survival under the eyes of the immune system [156, 157]. We therefore expect the existence of a set of genes encoding for membrane-associated proteins as we indeed observe in the decision tree in Figure 6. Using only the genes “Phosphate permease”, “ABC

transporter ATP-binding protein" and "transmembrane protein" for classification we already obtain 93.9% classification accuracy.

Consequently, hypothesis H1 seems to hold: We may separate at least actinobacterial species based on computational functional genomics features into pathogens and non-pathogens. Only a small set of three genes for each bias, i.e. classification direction, is sufficient to reach an approximately 90% accuracy.

TABLE 1 – SUMMARY OF OUR FINDINGS. THE TABLE SUMMARIZES THE RESULTS FOR EACH HYPOTHESIS. LIFESTYLES UNDERLINED AND BOLDED ARE THE BIASED ONES (REFER TO FIGURE 1 FOR AN ILLUSTRATION OF THE BIAS INTRODUCTION STRATEGY). $\overline{AUC}$ IS THE AVERAGE "AREA UNDER CURVE" (AUC) VALUE FOR FIVE DIFFERENT CROSS VALIDATION SUBSETS USING THE REAL LABELS. $\overline{AUC}_{RL}$ IS THE AVERAGE AUC VALUE FOR FIVE CROSS VALIDATION SUBSETS USING RANDOM LABELS. $\Delta\overline{AUC}$ IS THEIR DIFFERENCE: $|\overline{AUC} - \overline{AUC}_{RL}|$. SEE TEXT FOR DETAILS REGARDING THE CLASSIFICATION AND EVALUATION PROCEDURES. THE AVERAGE AUC FOR BOTH BIASES IS CALLED AUC AND DESCRIBES THE PREDICATION "ACCURACY". THE AUC_{RL} DESCRIBES THE "UNROBUSTNESS" OF THE CLASSIFICATION PERFORMANCE. THE AUC BIAS DESCRIBES THE "INFLUENCE OF THE BIAS" AND IS DEFINED AS $|\Delta\overline{AUC}_{bias1} - \Delta\overline{AUC}_{bias2}|$. THE "INFLUENCE OF BIAS" DESCRIBES WHETHER WE FIND CLASS-SPECIFIC GENES FOR BOTH PATHOGENICITY CLASSES (CLOSE TO 0) OR NOT (OTHERWISE).

		$\overline{AUC}$	$\overline{AUC}_{RL}$	$\Delta\overline{AUC}$	***AUC*** (Accuracy)	***AUC_{RL}*** (Unrobust-ness)	***AUC bias*** (Influence of bias)
H1	NP vs **HP+BP**	96.9	50.3	46.6	97.2	0.9	0.7
	NP vs HP+BP	97.4	51.5	45.9			
H2	OP vs **NP**	92.3	49.4	42.8	88.7	3.8	14.7
	OP vs NP	85.0	56.9	28.0			
H3	HP vs **BP**	94.5	57.4	37.1	92.2	5.4	5
	HP vs BP	89.8	47.7	42.1			
H4	OP vs **HP+BP**	91.1	61.3	29.8	91.8	10.5	2.8
	OP vs HP+BP	92.4	59.7	32.6			

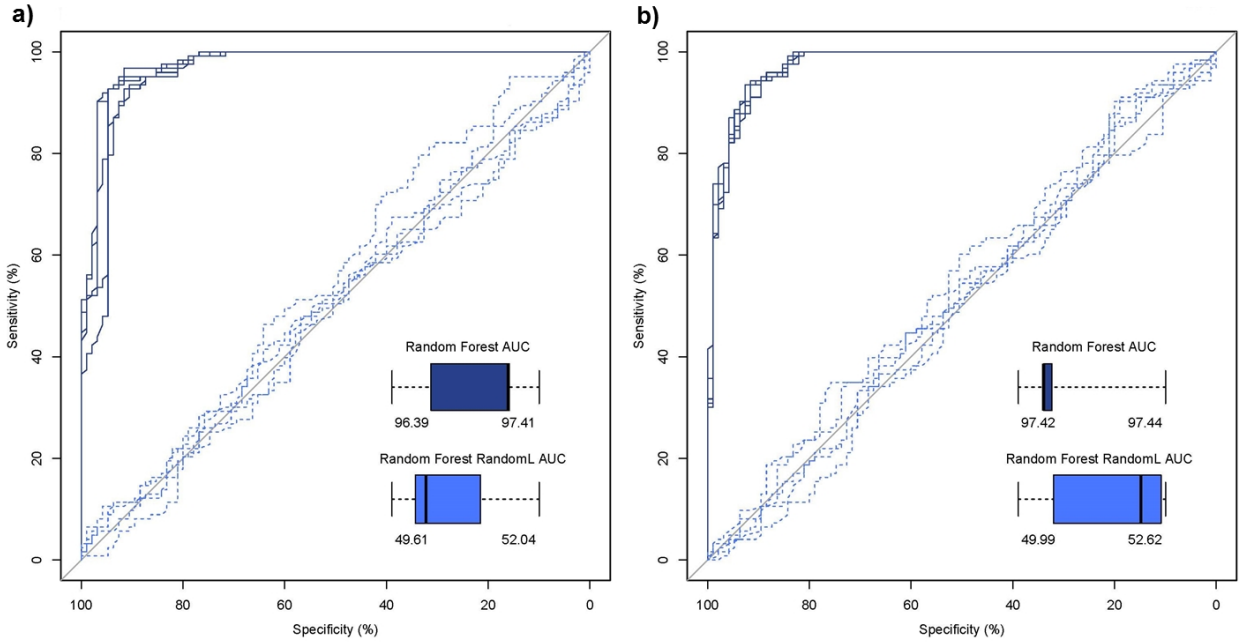

FIGURE 4 – CLASSIFICATION PERFORMANCE NON-PATHOGENS VS. PATHOGENS. ROC (RECEIVER OPERATING CHARACTERISTICS) PLOTS WERE GENERATED TO INSPECT THE PERFORMANCE OF THE CLASSIFICATION MODELS. THE DATA WAS EVALUATED FIVE TIMES USING DIFFERENT 5-FOLD CROSS VALIDATION SETS TO RECEIVE ROBUST QUALITY ESTIMATIONS OF OUR CLASSIFIERS. THE REAL LABEL CLASSIFIER CURVES ARE PRESENTED IN DARK BLUE SOLID LINES, WHILE THE RANDOM LABEL CLASSIFIER CURVES ARE GIVEN IN LIGHT BLUE DASHED LINES (THE ONES CLOSE TO THE BASE LINE). THE VARIATION OF THE AUCS (AREA UNDER CURVE) IN THE CROSS VALIDATION WAS INCLUDED IN THE FIGURE AS A BOX PLOT (BOTTOM RIGHT). THE NUMBERS BELOW EACH BOX PLOT ARE THE LOWER AND UPPER QUARTILES. A) PATHOGEN CLASSIFIER RESULTS (NP VS. HP+BP). WE BIASED THE PREDICTORS TOWARDS USING PATHOGEN-SPECIFIC GENES (SEE FIGURE 3). B) NON-PATHOGEN CLASSIFIER RESULTS (NP VS. HP+BP) WHERE THE PREDICATOR NOW WAS BIASED TO PREFER THE NON-PATHOGEN-SPECIFIC GENES. SEE TEXT FOR A FULL DESCRIPTION OF OUR MACHINE LEARNING STRATEGY AND REFER TO FIGURE 3 REGARDING THE “BIAS”.

3.2.2 (H2) OPPORTUNISTIC PATHOGENS VERSUS NON-PATHOGENS

The joint distribution between opportunist pathogens and non-pathogens reveals that most homologous gene clusters are equally present in both lifestyles, i.e. they cluster along the main diagonal (Figure 7). In contrast to H1, this data set is quite unbalanced (123 NPs vs. only 22 OPs), which might have been problematic for our classification procedure (Figure 8).

Consequently, we observe a severe “Influence of bias” (14.7%) and the lowest “Accuracy” of all four hypotheses (88.7%, see Table 1). Nevertheless, the $\overline{AUC}$ for the non-pathogen classifier (OP vs NP) was 92.3%, while the $\overline{AUC}$ for the opportunist pathogen classifier (OP vs NP) considerably drops (down to 85%). It also had the worst robustness against random labels ($\Delta\overline{AUC}$ = 28%). Note that this is still

better than a classifier using random labels. Further note that hypergeometric distribution tests would be necessary to assign a p-value to this effect.

The best opportunistic pathogen discriminant features were gene clusters with the Transitivity Clustering IDs 204058 and 217092, which are associated with the protein annotations “acyl transferase” and “transcriptional regulator”, respectively. Using only these two features we obtained 90.3% accuracy, but with 31.8% (7 out of 22) of the opportunistic pathogens being misclassified (Figure 9). The best non-pathogen discriminant features were gene clusters associated with the protein annotations “UDP-glucose 4-epimerase” and “PHP domain-containing protein”, respectively. By using only these two homologous gene clusters as features we may obtain an accuracy of 89.6%, but with 50% (11 out of 22) of the opportunistic pathogens being misclassified (Figure 10).

We cannot separate opportunistic pathogens from non-pathogens based on their gene repertoire computationally. Seen in the context of a quite unbalanced data set (with many more NP-genes than OP-genes) though, we may only carefully draw this conclusion.

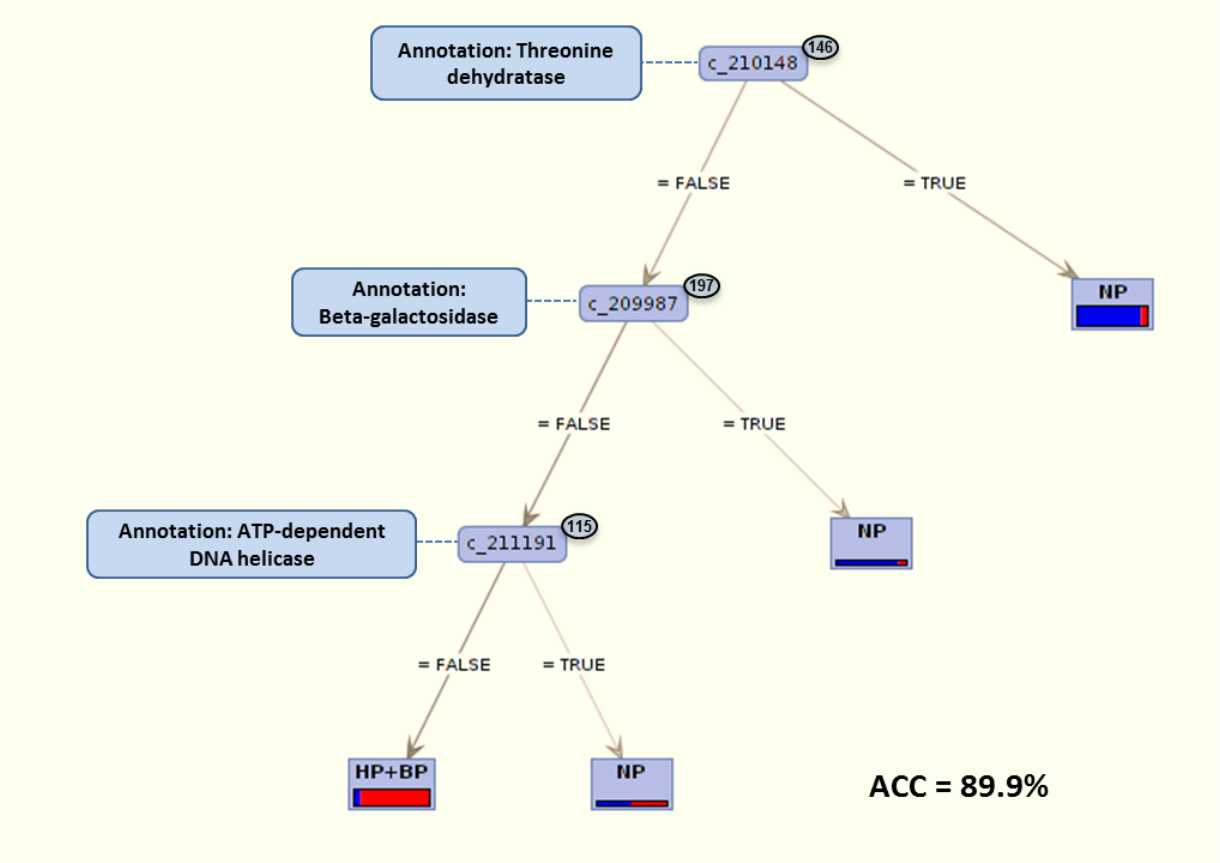

FIGURE 5 – DECISION TREE CREATED USING THE GENES MOST DISCRIMINATIVE FOR NON-PATHOGEN (NP). OUR CLASSIFICATION PIPELINE (SEE TEXT) SELECTED THE ABOVE THREE GENES AS MOST REPRESENTATIVE FOR PATHOGENS. WE LEARNED AND VISUALIZE THEM AS A SIMPLE DECISION TREE BY USING THE RAPIDMINER SOFTWARE. NODES REPRESENT GENE CLUSTERS WITH THE FOLLOWING TRANSITIVITY CLUSTERING IDS: 210148, 209987 AND 211191, WHICH ARE ASSOCIATED TO THE GENBANK ANNOTATIONS “THREONINE DEHYDRATASE” (E.G. UNIPROTKB AC: E3ERF0), “BETA-GALACTOSIDASE” (E.G. UNIPROTKB AC: D6Y6J1) AND “ATP-DEPENDENT DNA HELICASE” (E.G. UNIPROTKB AC: G0FLF9), RESPECTIVELY. THE SMALL CIRCLES CLOSE TO THE TRANSITIVITY CLUSTERING IDS INDICATE THE CLUSTER SIZE. USING ONLY THESE THREE FEATURES THE DECISION TREE ALREADY OBTAINS AN ACCURACY OF 89.9%.

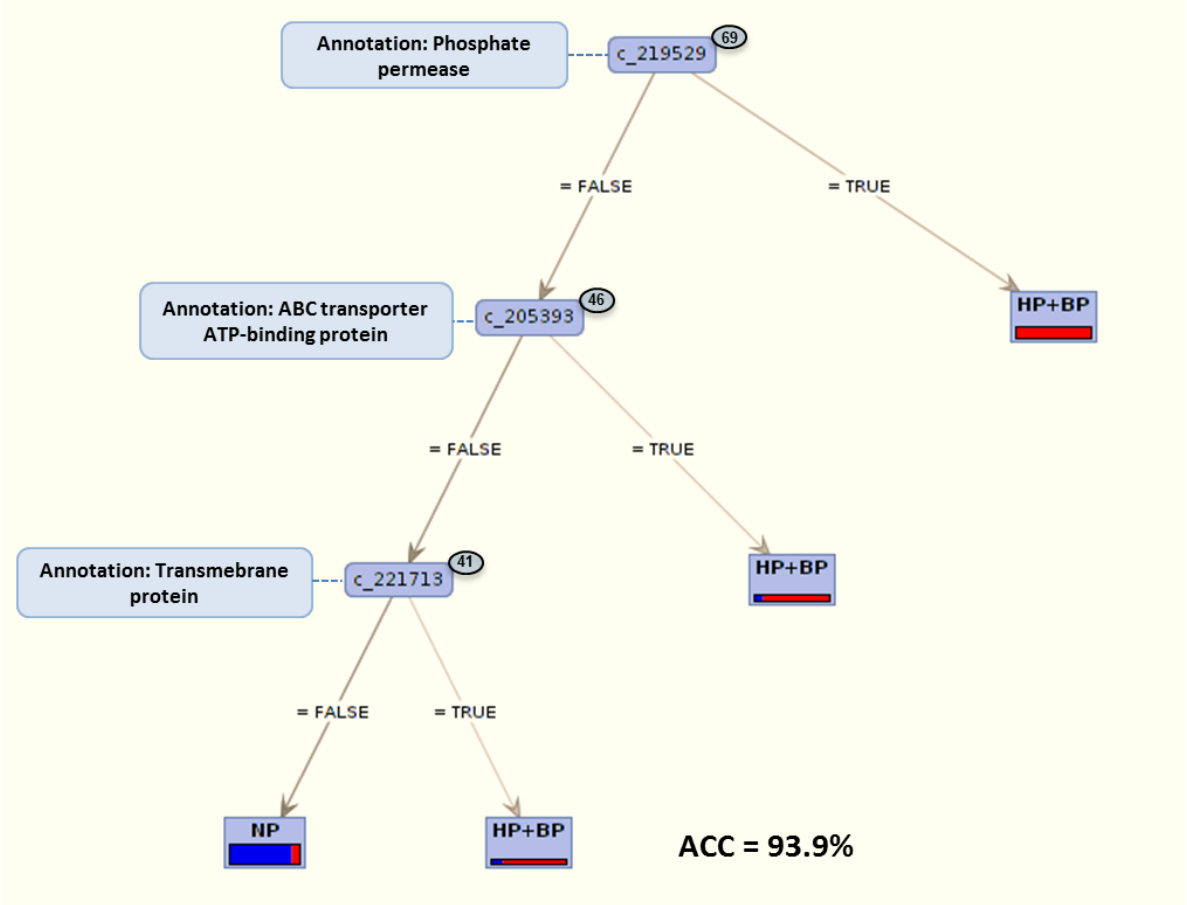

FIGURE 6 – DECISION TREE CREATED USING THE GENES MOST DISCRIMINATIVE FOR PATHOGEN (HP+BP). OUR CLASSIFICATION PIPELINE (SEE TEXT) SELECTED THE ABOVE THREE GENES AS MOST REPRESENTATIVE FOR PATHOGENS. WE LEARNED AND VISUALIZE THEM AS A SIMPLE DECISION TREE BY USING THE RAPIDMINER SOFTWARE. NODES REPRESENT GENE CLUSTERS WITH THE FOLLOWING TRANSITIVITY CLUSTERING IDS: 219529, 205393 AND 221713, WHICH ARE ASSOCIATED TO THE GENBANK ANNOTATIONS “PHOSPHATE PERMEASE” (E.G. UNIPROTKB AC: I6YD06 OR P65712), “ABC TRANSPORTER ATP-BINDING PROTEIN” (E.G. UNIPROTKB AC: D9Q9K6) AND “TRANSMEMBRANE PROTEIN” (E.G. UNIPROTKB AC: G2MY46), RESPECTIVELY. THE SMALL CIRCLES CLOSE TO THE TRANSITIVITY CLUSTERING IDS INDICATE THE CLUSTER SIZE. USING ONLY THESE THREE FEATURES THE DECISION TREE ALREADY OBTAINS AN ACCURACY OF 93.9%.

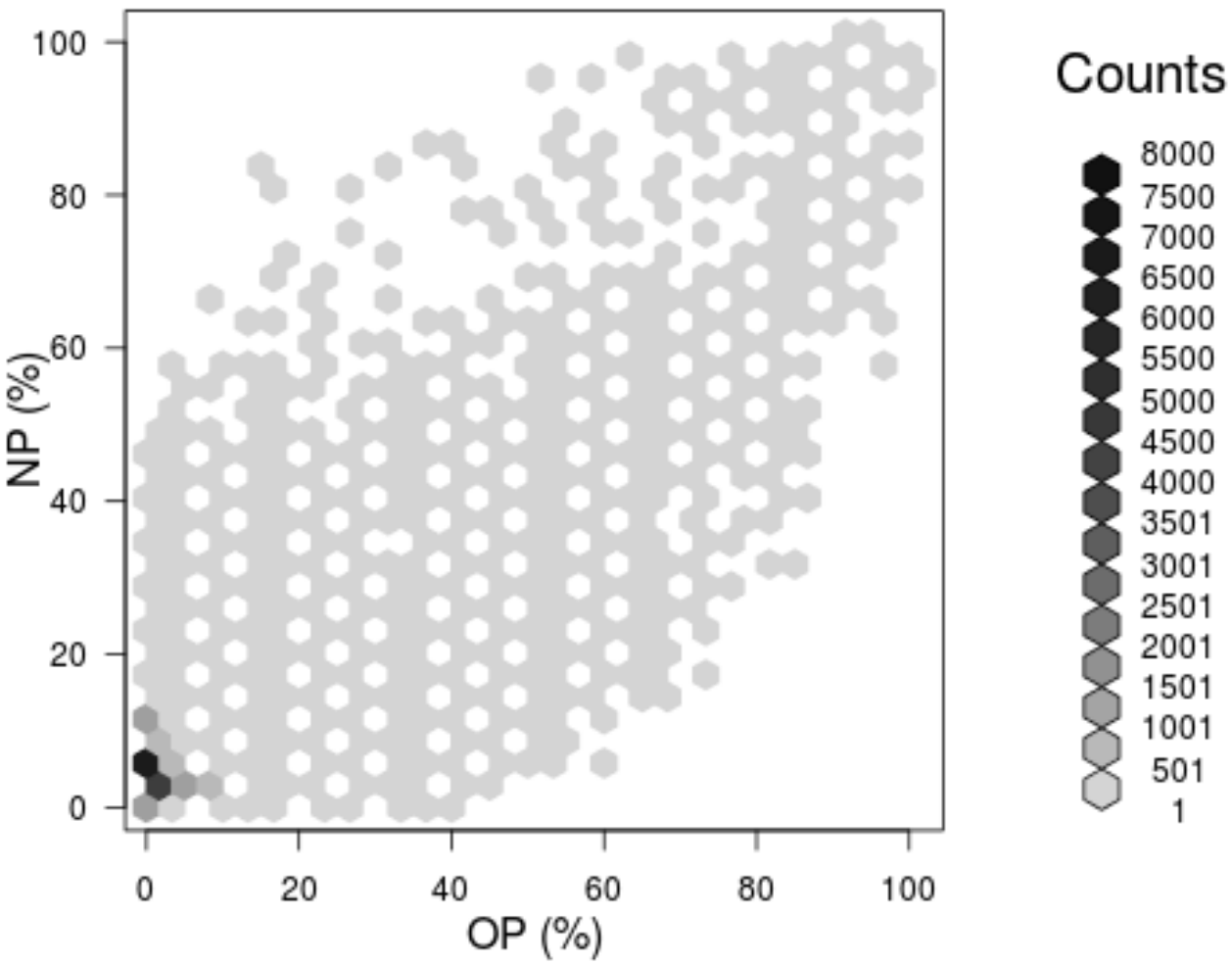

FIGURE 7 – DISTRIBUTION OF HOMOLOGOUS GENE CLUSTERS OVER TWO LIFESTYLES (OPPORTUNISTIC PATHOGENS VS. NON-PATHOGENS). BOTH AXES IN THE PLOT DESCRIBE THE PERCENTAGE OF SPECIES IN THE RESPECTIVE CLASS(ES), HERE OPPORTUNISTIC PATHOGENS (OP) VS. NON-PATHOGENS (NP). THE COLOR-CODING OF THE HEAT MAP DEPICTS THE NUMBER OF CLUSTERS OF HOMOLOGOUS GENES THAT CERTAIN PERCENTAGES OF PATHOGENS/NON-PATHOGENS SHARE. THUS, IN THE UPPER RIGHT OF SUCH A JOINT DISTRIBUTION PLOT, WE FIND THE CORE GENOME (HOMOLOGOUS GENES PRESENT IN ALL SPECIES OF BOTH CLASSES); AND IN THE LOWER LEFT, WE SEE UNIQUE, SPECIES-SPECIFIC GENES. GENES CLOSE TO THE AXIS ARE MORE CLASS SPECIFIC. GENES CLOSE TO THE AXIS TAILS ARE HIGHLY CLASS SPECIFIC.

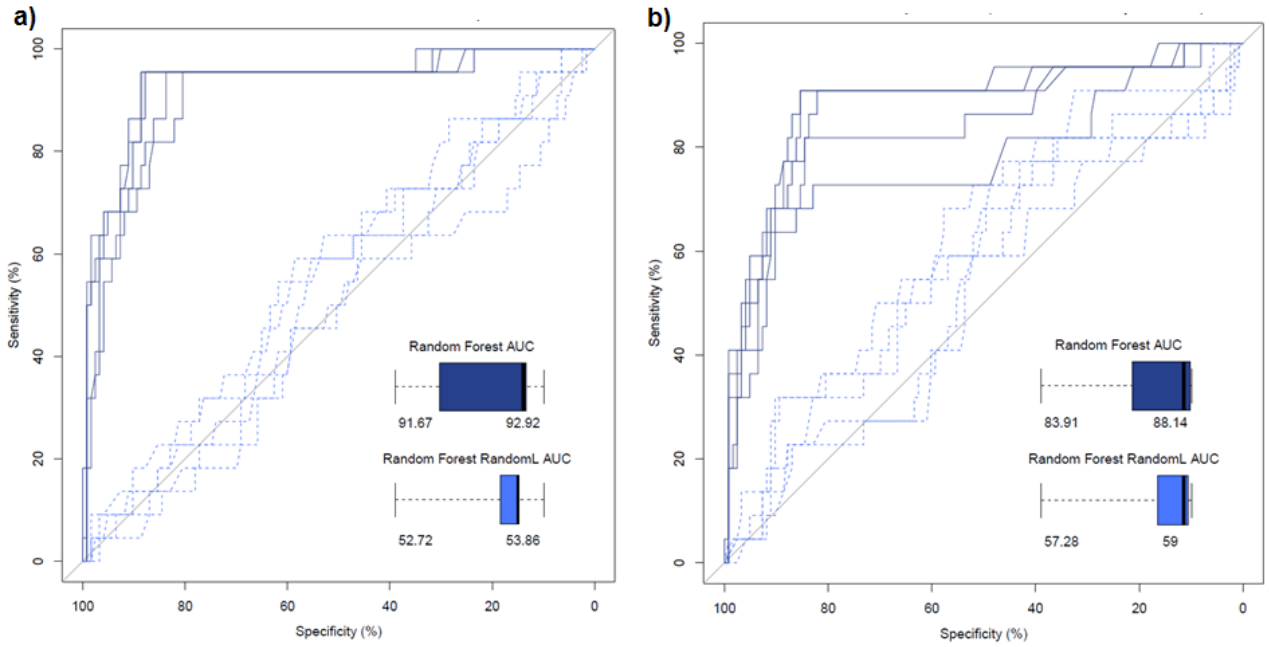

FIGURE 8 – CLASSIFICATION PERFORMANCE NON-PATHOGENS VS. OPPORTUNISTIC PATHOGENS. ROC (RECEIVER OPERATING CHARACTERISTICS) PLOTS WERE GENERATED TO INSPECT THE PERFORMANCE OF THE CLASSIFICATION MODELS. THE DATA WAS EVALUATED FIVE TIMES USING DIFFERENT 5-FOLD CROSS VALIDATION SETS TO RECEIVE ROBUST QUALITY ESTIMATIONS OF OUR CLASSIFIERS. THE REAL LABEL CLASSIFIER CURVES ARE PRESENTED IN DARK BLUE SOLID LINES, WHILE THE RANDOM LABEL CLASSIFIER CURVES ARE GIVEN IN LIGHT BLUE DASHED LINES (THE ONES CLOSE TO THE BASE LINE). THE VARIATION OF THE AUCS (AREA UNDER CURVE) IN THE CROSS VALIDATION WAS INCLUDED IN THE FIGURE AS A BOX PLOT (BOTTOM RIGHT). THE NUMBERS BELOW EACH BOX PLOT ARE THE LOWER AND UPPER QUARTILES. A) NON-PATHOGEN CLASSIFIER RESULTS (NP VS. OP). WE BIASED THE PREDICTORS TOWARDS USING NON-PATHOGEN-SPECIFIC GENES (SEE FIGURE 7). B) OPPORTUNISTIC PATHOGEN CLASSIFIER RESULTS (NP VS. OP) WHERE THE PREDICATOR NOW WAS BIASED TO PREFER THE OPPORTUNISTIC PATHOGEN-SPECIFIC GENES. SEE TEXT FOR A FULL DESCRIPTION OF OUR MACHINE LEARNING STRATEGY AND REFER TO FIGURE 3 REGARDING THE “BIAS”.

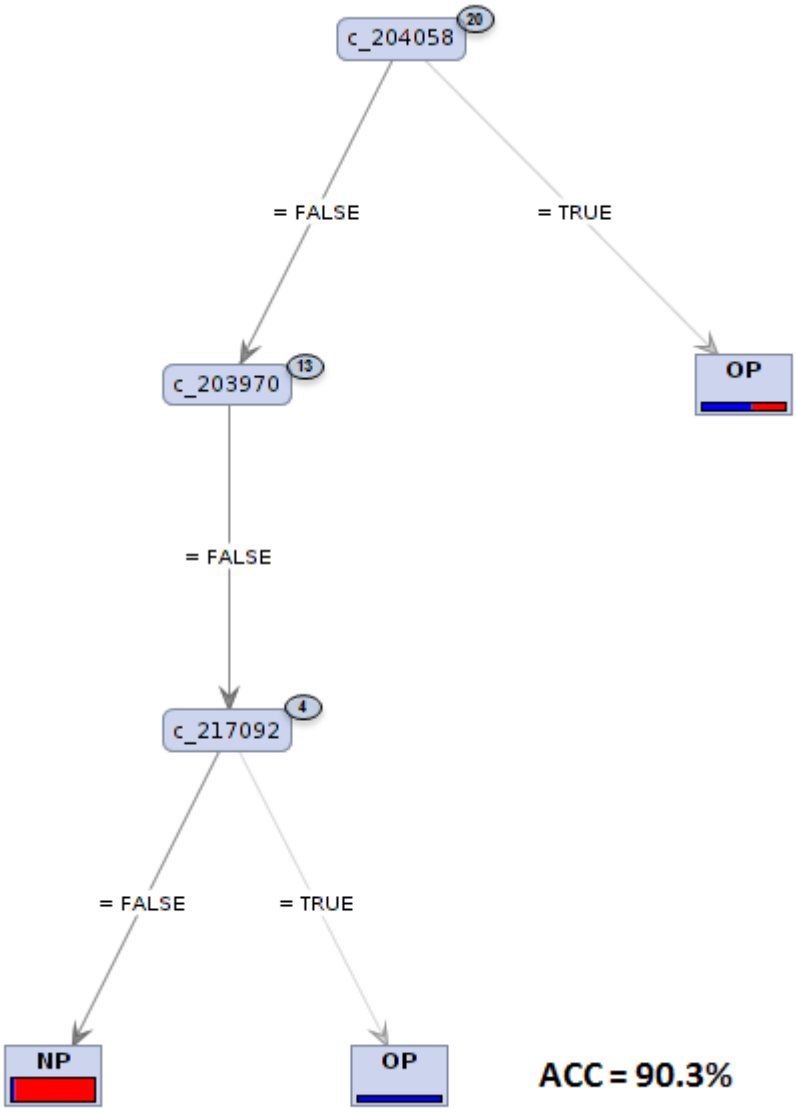

FIGURE 9 – DECISION TREE CREATED USING THE GENES MOST DISCRIMINATIVE FOR OPPORTUNISTIC PATHOGEN (OP). OUR CLASSIFICATION PIPELINE (SEE TEXT) SELECTED THE ABOVE THREE GENES AS MOST REPRESENTATIVE FOR PATHOGENS. WE LEARNED AND VISUALIZE THEM AS A SIMPLE DECISION TREE BY USING THE RAPIDMINER SOFTWARE. NODES REPRESENT GENE CLUSTERS WITH THE FOLLOWING TRANSITIVITY CLUSTERING IDS: 204058, 203970 AND 217092, WHICH ARE ASSOCIATED TO THE GENBANK ANNOTATIONS “ACYL TRANSFERASE” (E.G. UNIPROTKB AC: D0L269), “THIOESTERASE” (E.G. UNIPROTKB AC: D2NT48) AND “TRANSCRIPTIONAL REGULATOR” (E.G. UNIPROTKB AC: E3H005), RESPECTIVELY. THE SMALL CIRCLES CLOSE TO THE TRANSITIVITY CLUSTERING IDS INDICATE THE CLUSTER SIZE. USING ONLY THESE THREE FEATURES THE DECISION TREE ALREADY OBTAINS AN ACCURACY OF 90.3%.

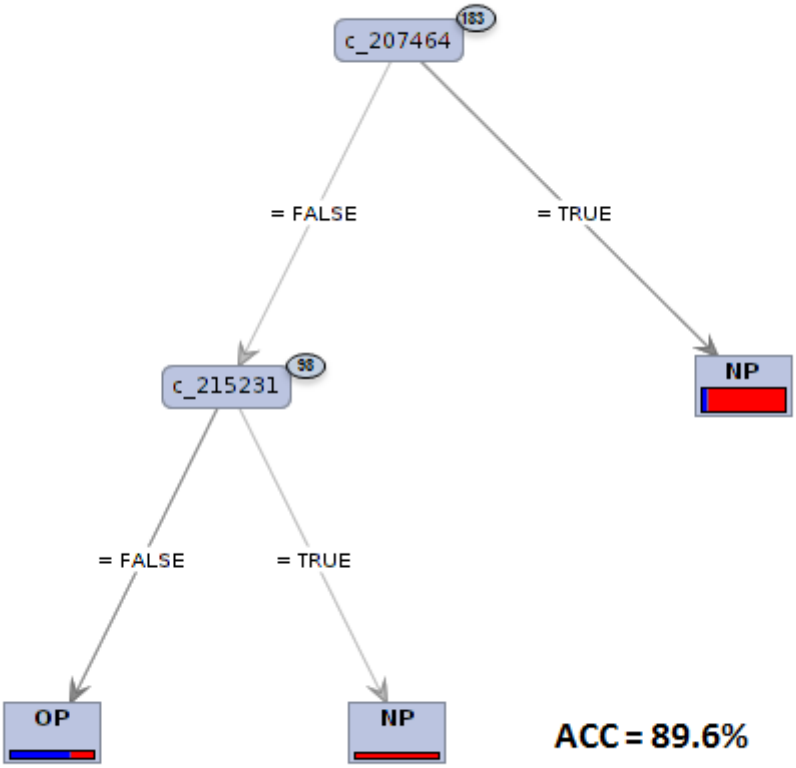

FIGURE 10 – DECISION TREE CREATED USING THE GENES MOST DISCRIMINATIVE FOR OPPORTUNISTIC NON-PATHOGEN (NP). OUR CLASSIFICATION PIPELINE (SEE TEXT) SELECTED THE ABOVE TWO GENES AS MOST REPRESENTATIVE FOR PATHOGENS. WE LEARNED AND VISUALIZE THEM AS A SIMPLE DECISION TREE BY USING THE RAPIDMINER SOFTWARE. NODES REPRESENT GENE CLUSTERS WITH THE FOLLOWING TRANSITIVITY CLUSTERING IDS: 207464 AND 215231, WHICH ARE ASSOCIATED TO THE GENBANK ANNOTATIONS “UDP-GLUCOSE 4-EPIMERASE” (E.G. UNIPROTKB AC: C6WAE7) AND “PHP DOMAIN-CONTAINING PROTEIN” (E.G. UNIPROTKB AC: C7QE58), RESPECTIVELY. THE SMALL CIRCLES CLOSE TO THE TRANSITIVITY CLUSTERING IDS INDICATE THE CLUSTER SIZE. USING ONLY THESE THREE FEATURES THE DECISION TREE ALREADY OBTAINS AN ACCURACY OF 89.6%.

3.2.3 (H3) HUMAN PATHOGENS VERSUS BROAD-SPECTRUM PATHOGENS

The joint distribution between human pathogens and broad-spectrum pathogens reveals a potentially higher separability as there are several homologous gene clusters close to the two axes (Figure 11). In contrast to H2, this data set is more balanced (68 HPs and 27 BPs). However, a closer look at the species table indicates a risk for a class-internal bias, as many broad-spectrum pathogens are *Corynebacterium pseudotuberculosis* (CP) strains (15 out of 27) or *Mycobacterium bovis* (MB) strains (5 out of 27).

Again, refer to Table 1. The “Influence of bias” is second highest (5%) and emerges from a better classification performance when the data set was biased towards using BP-specific genes. This might be due to the comparably high number of CP strains and MB strains (internal bias in the BP data set). Consequently, the data set biased towards broad-spectrum pathogens (HP vs BP) generated a higher $\overline{AUC}$ than the one biased towards human pathogens features (HP vs BP). The overall “Accuracy” is 92.2% with the second highest “Unrobustness” (5.4%). It emerges from comparably “good” results of the random classifier (57.4% for BPs). Figure 12 depicts the classification performance of the two biased data sets.

The best broad-spectrum pathogen discriminant feature gene clusters were the ones with the Transitivity Clustering IDs 1505 and 6101 (Figure 13). They are associated with “hypothetical protein” and “membrane protein”, respectively. Using only these two genes we obtain 95.6% accuracy. However, we only separate species from the two internally biased strains (CP and MB). We may consequently regard this as a data set artifact. It does not affect our conclusion, however. The Random Forest (RF) classifier we use is quite robust against such unbalanced data sets [146] and would have picked a larger feature set (i.e. more genes) if this had increased the prediction performance. We will study the effect of using a pre-processed data set (with a small number of randomly picked CP and MB strains) in the future though. The same holds for the human pathogen discriminant feature genes (IDs: 209123 and 219362; annotation: “hydrolase” and “potassium transporter”). With these two features only, the best achievable accuracy is only 74.7% (Figure 14).

Similarly to H2, we cannot separate human pathogens from broad-spectrum pathogens computationally. This makes sense as we can assume that human pathogens may infect many more hosts as have been annotated (maybe for a lack of research interest). Seen in the context of the internally unbalanced data set (with many two dominant species, which are likely to emerge two dominant feature genes) though, we may again only carefully draw this conclusion.

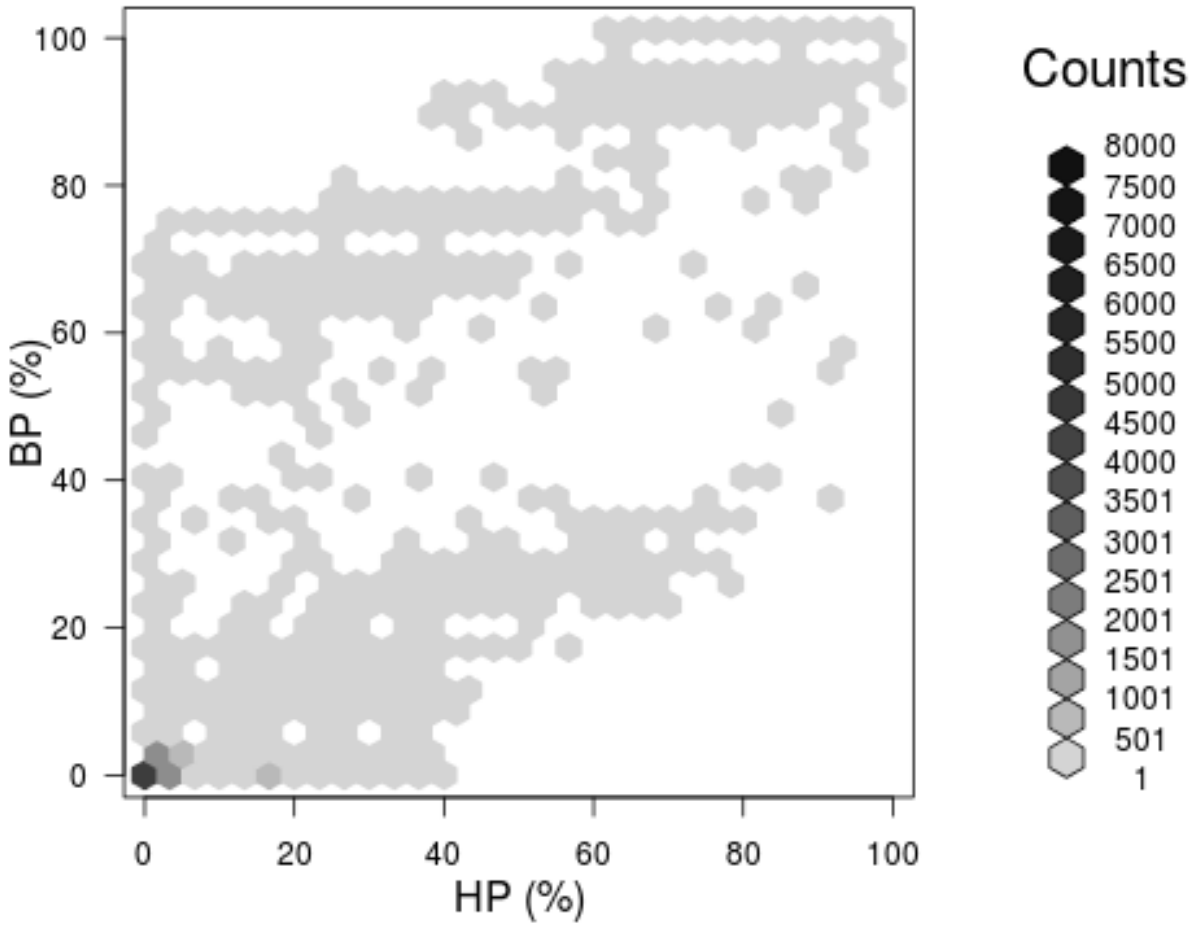

FIGURE 11 – DISTRIBUTION OF HOMOLOGOUS GENE CLUSTERS OVER TWO LIFESTYLES (HUMAN PATHOGENS VS. BROAD-SPECTRUM PATHOGENS). BOTH AXES IN THE PLOT DESCRIBE THE PERCENTAGE OF SPECIES IN THE RESPECTIVE CLASS(ES), HERE HUMAN PATHOGENS (HP) VS. BROAD-SPECTRUM PATHOGENS (BP). THE COLOR-CODING OF THE HEAT MAP DEPICTS THE NUMBER OF CLUSTERS OF HOMOLOGOUS GENES THAT CERTAIN PERCENTAGES OF PATHOGENS/NON-PATHOGENS SHARE. THUS, IN THE UPPER RIGHT OF SUCH A JOINT DISTRIBUTION PLOT, WE FIND THE CORE GENOME (HOMOLOGOUS GENES PRESENT IN ALL SPECIES OF BOTH CLASSES); AND IN THE LOWER LEFT, WE SEE UNIQUE, SPECIES-SPECIFIC GENES. GENES CLOSE TO THE AXIS ARE MORE CLASS SPECIFIC. GENES CLOSE TO THE AXIS TAILS ARE HIGHLY CLASS SPECIFIC.

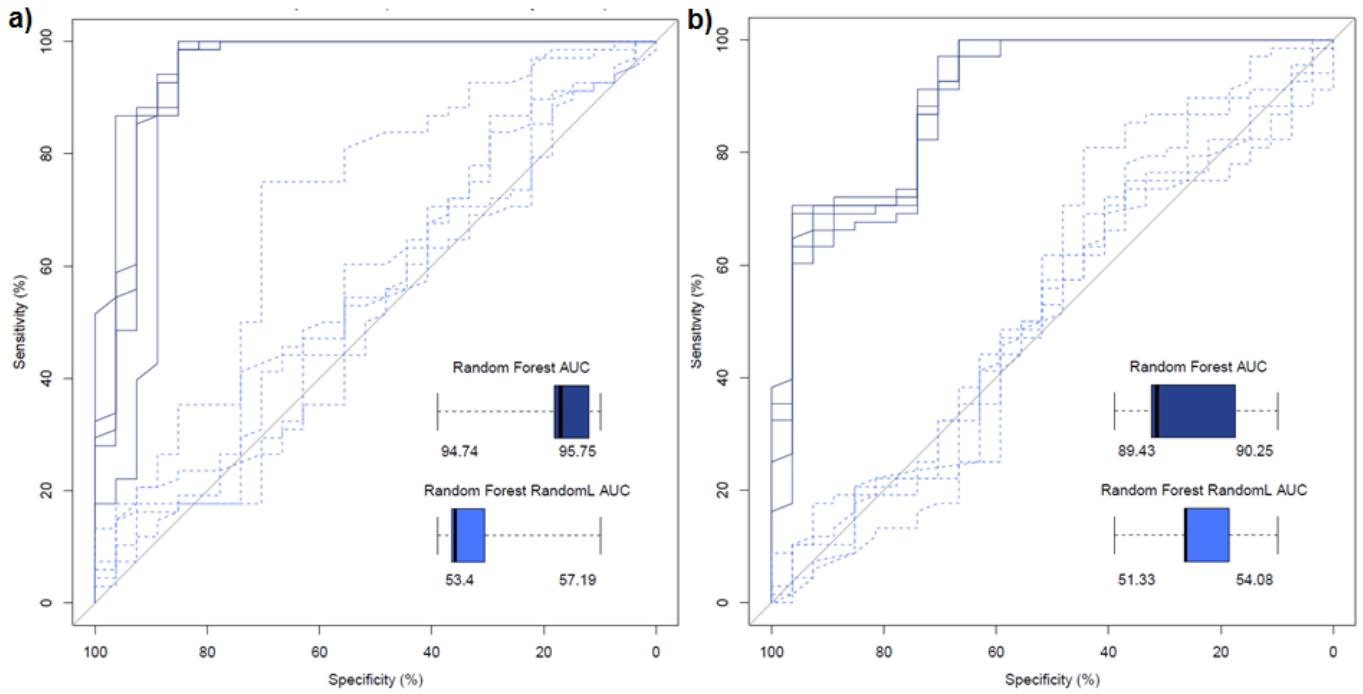

FIGURE 12 – CLASSIFICATION PERFORMANCE BROAD-SPECTRUM PATHOGENS VS. HUMAN PATHOGENS. ROC (RECEIVER OPERATING CHARACTERISTICS) PLOTS WERE GENERATED TO INSPECT THE PERFORMANCE OF THE CLASSIFICATION MODELS. THE DATA WAS EVALUATED FIVE TIMES USING DIFFERENT 5-FOLD CROSS VALIDATION SETS TO RECEIVE ROBUST QUALITY ESTIMATIONS OF OUR CLASSIFIERS. THE REAL LABEL CLASSIFIER CURVES ARE PRESENTED IN DARK BLUE SOLID LINES, WHILE THE RANDOM LABEL CLASSIFIER CURVES ARE GIVEN IN LIGHT BLUE DASHED LINES (THE ONES CLOSE TO THE BASE LINE). THE VARIATION OF THE AUCS (AREA UNDER CURVE) IN THE CROSS VALIDATION WAS INCLUDED IN THE FIGURE AS A BOX PLOT (BOTTOM RIGHT). THE NUMBERS BELOW EACH BOX PLOT ARE THE LOWER AND UPPER QUARTILES. A) BROAD-SPECTRUM PATHOGENS RESULTS (BP VS. HP). WE BIASED THE PREDICTORS TOWARDS USING BROAD-SPECTRUM PATHOGEN-SPECIFIC GENES (SEE FIGURE 11). B) HUMAN PATHOGEN CLASSIFIER RESULTS (BP VS. HP) WHERE THE PREDICATOR NOW WAS BIASED TO PREFER THE HUMAN PATHOGEN-SPECIFIC GENES. SEE TEXT FOR A FULL DESCRIPTION OF OUR MACHINE LEARNING STRATEGY AND REFER TO FIGURE 3 REGARDING THE “BIAS”.

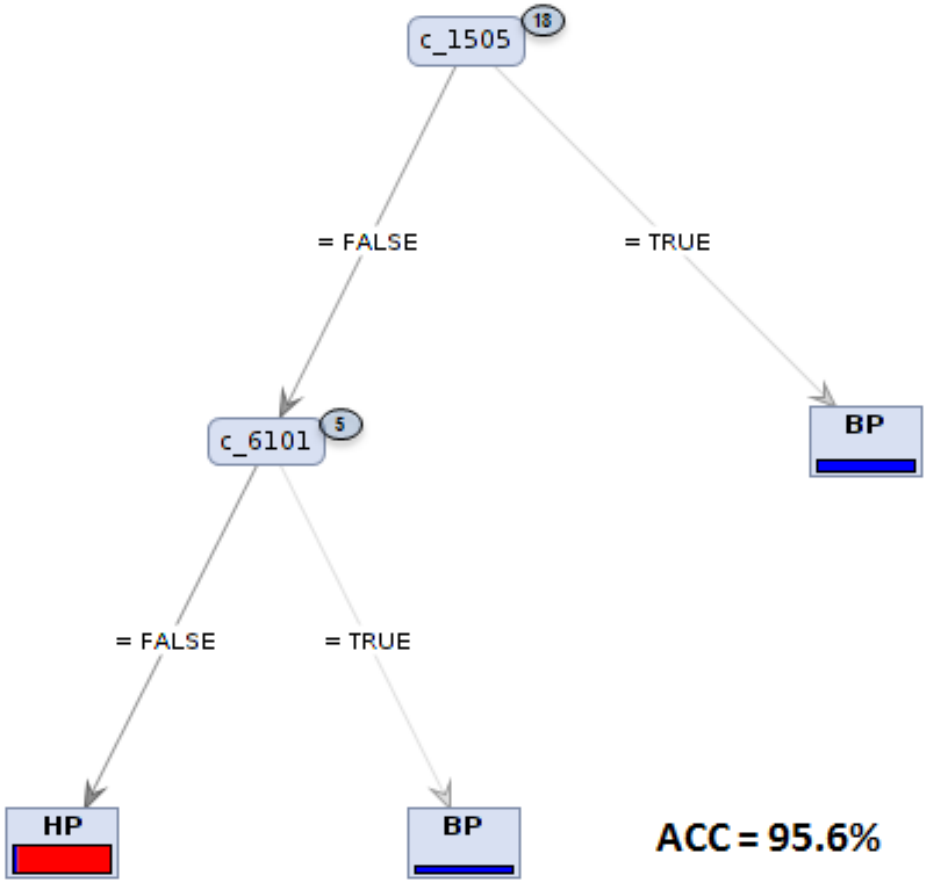

FIGURE 13 – DECISION TREE CREATED USING THE GENES MOST DISCRIMINATIVE FOR BROAD-SPECTRUM PATHOGEN (BP). OUR CLASSIFICATION PIPELINE (SEE TEXT) SELECTED THE ABOVE TWO GENES AS MOST REPRESENTATIVE FOR PATHOGENS. WE LEARNED AND VISUALIZE THEM AS A SIMPLE DECISION TREE BY USING THE RAPIDMINER SOFTWARE. NODES REPRESENT GENE CLUSTERS WITH THE FOLLOWING TRANSITIVITY CLUSTERING IDS: 1505 AND 6101, WHICH ARE ASSOCIATED TO THE GENBANK ANNOTATIONS “HYPOTHETICAL PROTEIN” (E.G. UNIPROTKB AC: I7HCH9) AND “MEMBRANE PROTEIN” (E.G. UNIPROTKB AC: M1IJT1), RESPECTIVELY. THE SMALL CIRCLES CLOSE TO THE TRANSITIVITY CLUSTERING IDS INDICATE THE CLUSTER SIZE. USING ONLY THESE THREE FEATURES THE DECISION TREE ALREADY OBTAINS AN ACCURACY OF 95.6%.

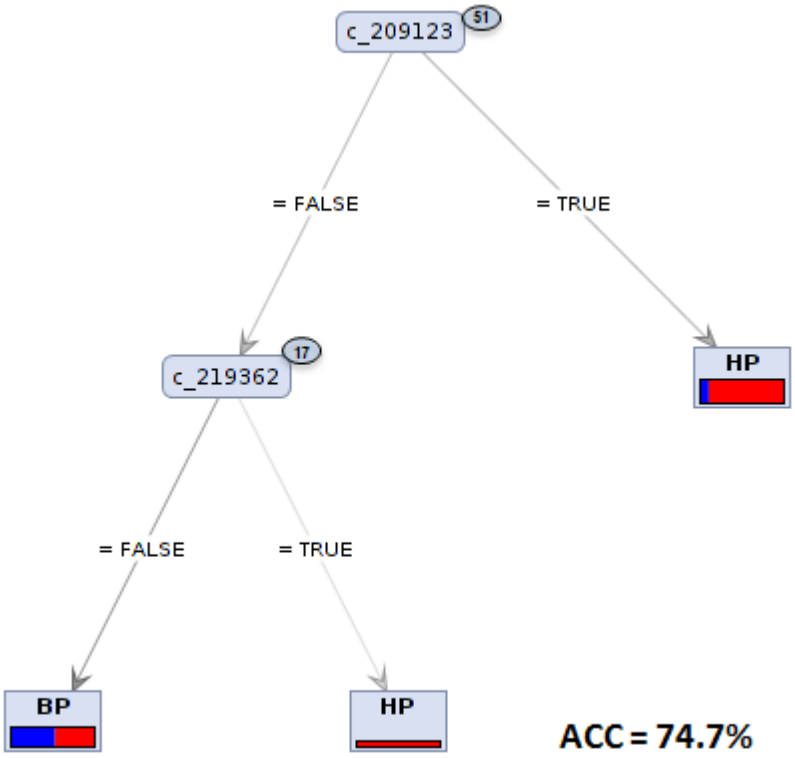

FIGURE 14 – DECISION TREE CREATED USING THE GENES MOST DISCRIMINATIVE HUMAN PATHOGEN (HP). OUR CLASSIFICATION PIPELINE (SEE TEXT) SELECTED THE ABOVE TWO GENES AS MOST REPRESENTATIVE FOR PATHOGENS. WE LEARNED AND VISUALIZE THEM AS A SIMPLE DECISION TREE BY USING THE RAPIDMINER SOFTWARE. NODES REPRESENT GENE CLUSTERS WITH THE FOLLOWING TRANSITIVITY CLUSTERING IDS: 209123 AND 219362, WHICH ARE ASSOCIATED TO THE GENBANK ANNOTATIONS "HYDROLASE" (E.G. UNIPROTKB AC: F2GEE4) AND "POTASSIUM TRANSPORTER" (E.G. UNIPROTKB AC: G0DW68), RESPECTIVELY. THE SMALL CIRCLES CLOSE TO THE TRANSITIVITY CLUSTERING IDS INDICATE THE CLUSTER SIZE. USING ONLY THESE THREE FEATURES THE DECISION TREE ALREADY OBTAINS AN ACCURACY OF 74.7%.

3.2.4 (H4) OPPORTUNISTIC PATHOGENS VERSUS ALL PATHOGENS

The joint distribution between opportunistic pathogens (OP) and all pathogens (HP and BP) is similar to the one from H2, with most homologous genes equally present in both lifestyles clustered around the main diagonal (Figure 15). We have a slightly unbalanced data set with 95 pathogens (68 HPs and 27 BPs) and only 22 opportunistic genomes.

Table 1 shows a moderate "Influence of bias" (2.8%). Although the overall "Accuracy" appears quite high (91.8%), this likely results from overfitting, as can be seen from the by far highest "Unrobustness" (10.5%). It emerges from comparably "good" results of the random classifier (approximately 60% for both biases).

Consequently, none of the classifier provides results considerably better than a random classifier. Figure 16 depicts the classification performance of the two biased data sets.

Nevertheless, if we construct decision trees based on the best separating feature genes, we may achieve an all-pathogen-specific tree (IDs: 221680 and 219390, annotations: “hypothetical protein” and “coenzyme PQQ synthesis protein”). Using only these two features we may obtain 80.5% accuracy, but only one opportunistic pathogen was correctly classified (*Cryptobacterium curtum* DSM 15641). The best opportunistic pathogen discriminant features were gene clusters with the following IDs: 219449 and 217696, which are associated with “major facilitator superfamily” and “iron permease”, respectively. Using only these two features we obtained 90.5% accuracy, with 86.3% (19 out of 22) opportunist pathogens being correctly classified (see Figure 17 and Figure 18). In summary, hypothesis H4 clearly holds: There is no robust feature gene set that separates opportunistic pathogens from all other pathogens significantly better than a random gene set. This seems reasonable, as both occupy the same niche but opportunists only cause (subjectively perceived) symptoms if the host’s immune system is compromised, for instance, in cases of co-infection, pregnancy, weak immune system, etc.

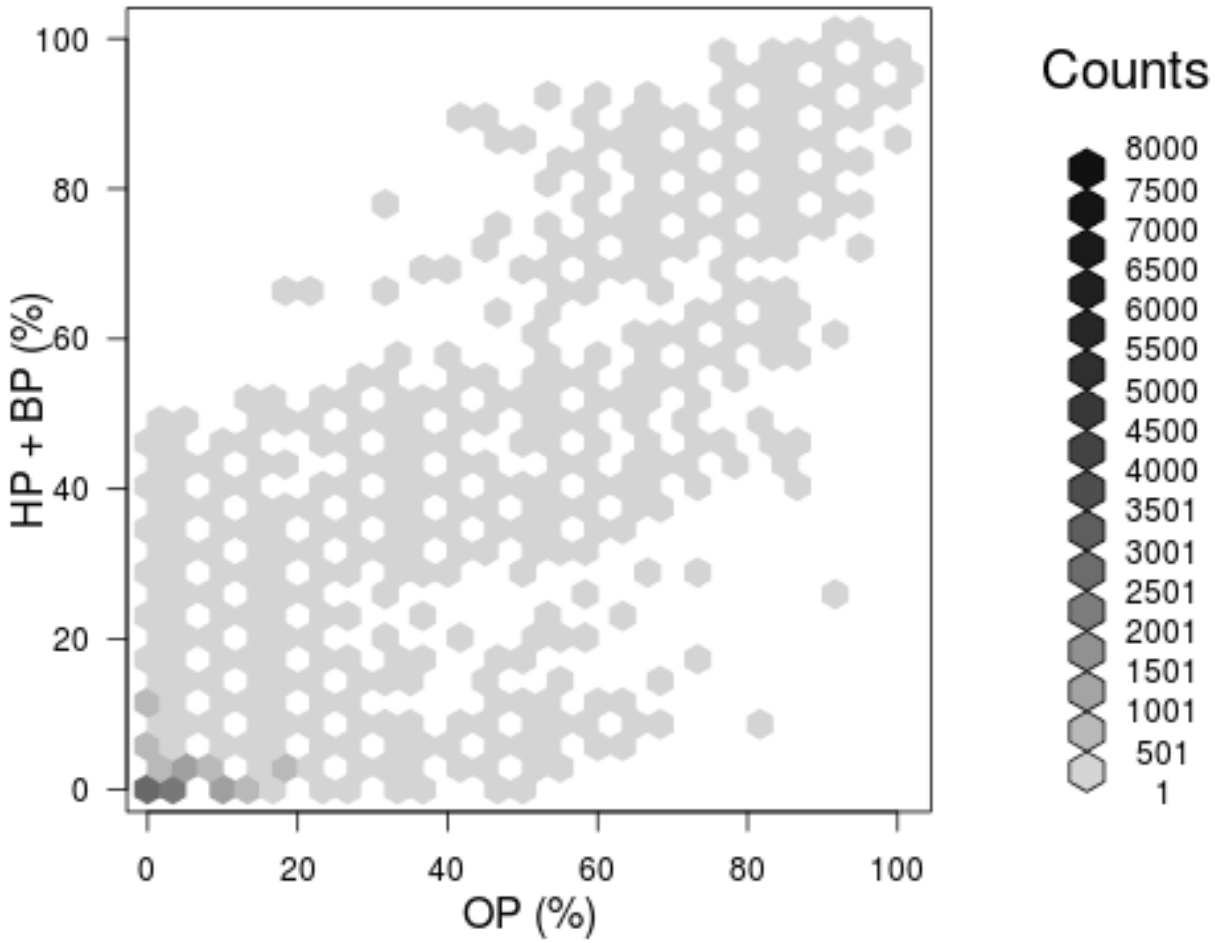

FIGURE 15 – DISTRIBUTION OF HOMOLOGOUS GENE CLUSTERS OVER TWO LIFESTYLES (OPPORTUNISTIC PATHOGENS VS. ALL PATHOGENS). BOTH AXES IN THE PLOT DESCRIBE THE PERCENTAGE OF SPECIES IN THE RESPECTIVE CLASS(ES), HERE OPPORTUNISTIC PATHOGENS (HP) VS. HUMAN PATHOGENS (HP) AND BROAD-SPECTRUM PATHOGENS (BP). THE COLOR-CODING OF THE HEAT MAP DEPICTS THE NUMBER OF CLUSTERS OF HOMOLOGOUS GENES THAT CERTAIN PERCENTAGES OF PATHOGENS/NON-PATHOGENS SHARE. THUS, IN THE UPPER RIGHT OF SUCH A JOINT DISTRIBUTION PLOT, WE FIND THE CORE GENOME (HOMOLOGOUS GENES PRESENT IN ALL SPECIES OF BOTH CLASSES); AND IN THE LOWER LEFT, WE SEE UNIQUE, SPECIES-SPECIFIC GENES. GENES CLOSE TO THE AXIS ARE MORE CLASS SPECIFIC. GENES CLOSE TO THE AXIS TAILS ARE HIGHLY CLASS SPECIFIC.

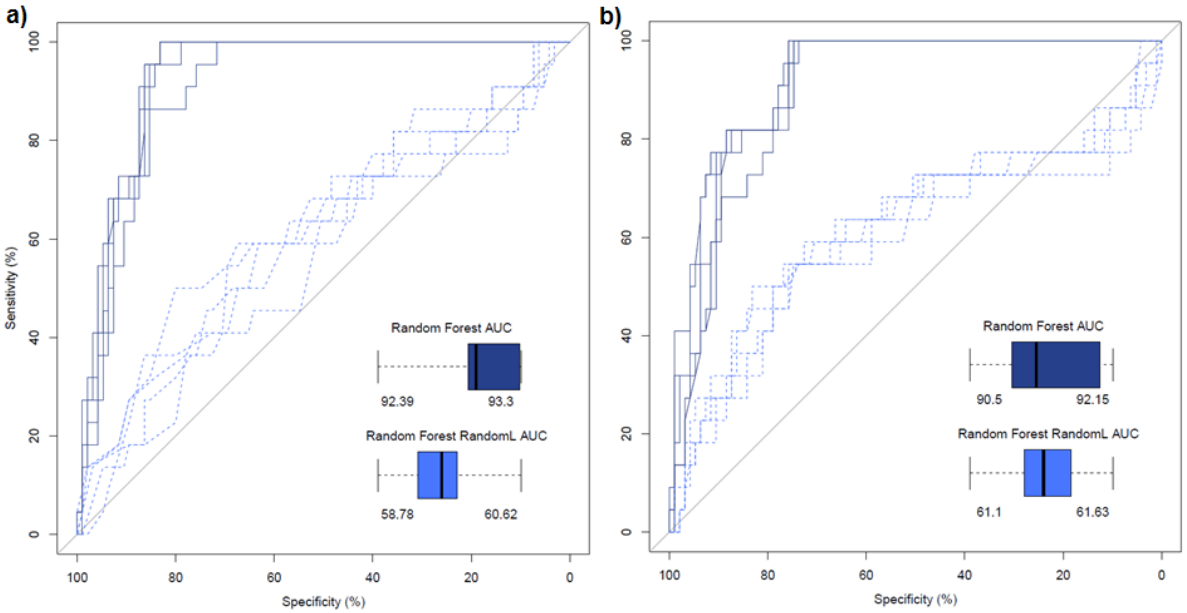

FIGURE 16 – CLASSIFICATION PERFORMANCE OPPORTUNISTIC PATHOGENS VS. ALL PATHOGENS. ROC (RECEIVER OPERATING CHARACTERISTICS) PLOTS WERE GENERATED TO INSPECT THE PERFORMANCE OF THE CLASSIFICATION MODELS. THE DATA WAS EVALUATED FIVE TIMES USING DIFFERENT 5-FOLD CROSS VALIDATION SETS TO RECEIVE ROBUST QUALITY ESTIMATIONS OF OUR CLASSIFIERS. THE REAL LABEL CLASSIFIER CURVES ARE PRESENTED IN DARK BLUE SOLID LINES, WHILE THE RANDOM LABEL CLASSIFIER CURVES ARE GIVEN IN LIGHT BLUE DASHED LINES (THE ONES CLOSE TO THE BASE LINE). THE VARIATION OF THE AUCS (AREA UNDER CURVE) IN THE CROSS VALIDATION WAS INCLUDED IN THE FIGURE AS A BOX PLOT (BOTTOM RIGHT). THE NUMBERS BELOW EACH BOX PLOT ARE THE LOWER AND UPPER QUARTILES. A) OPPORTUNISTIC PATHOGENS RESULTS (OP VS. BP+HP). WE BIASED THE PREDICTORS TOWARDS USING PATHOGEN-SPECIFIC GENES (SEE FIGURE 15). B) HUMAN PATHOGEN CLASSIFIER RESULTS (OP VS. BP+HP) WHERE THE PREDICATOR NOW WAS BIASED TO PREFER THE OPPORTUNISTIC PATHOGEN-SPECIFIC GENES. SEE TEXT FOR A FULL DESCRIPTION OF OUR MACHINE LEARNING STRATEGY AND REFER TO FIGURE 3 REGARDING THE “BIAS”.

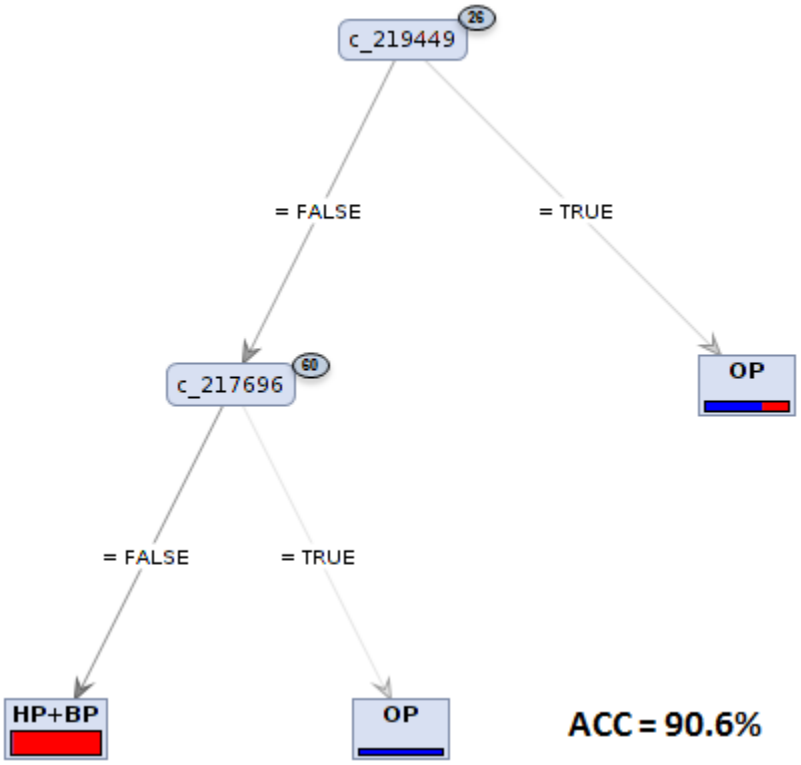

FIGURE 17 – DECISION TREE CREATED USING THE GENES MOST DISCRIMINATIVE OPPORTUNISTIC PATHOGEN (OP). OUR CLASSIFICATION PIPELINE (SEE TEXT) SELECTED THE ABOVE TWO GENES AS MOST REPRESENTATIVE FOR PATHOGENS. WE LEARNED AND VISUALIZE THEM AS A SIMPLE DECISION TREE BY USING THE RAPIDMINER SOFTWARE. NODES REPRESENT GENE CLUSTERS WITH THE FOLLOWING TRANSITIVITY CLUSTERING IDS: 219449 AND 217696, WHICH ARE ASSOCIATED TO THE GENBANK ANNOTATIONS “MAJOR FACILITATOR SUPERFAMILY PERMEASE” (E.G. UNIPROTKB AC: D7BLX7) AND “IRON PERMEASE FTR1” (E.G. UNIPROTKB AC: A0JRG0), RESPECTIVELY. THE SMALL CIRCLES CLOSE TO THE TRANSITIVITY CLUSTERING IDS INDICATE THE CLUSTER SIZE. USING ONLY THESE THREE FEATURES THE DECISION TREE ALREADY OBTAINS AN ACCURACY OF 90.6%.

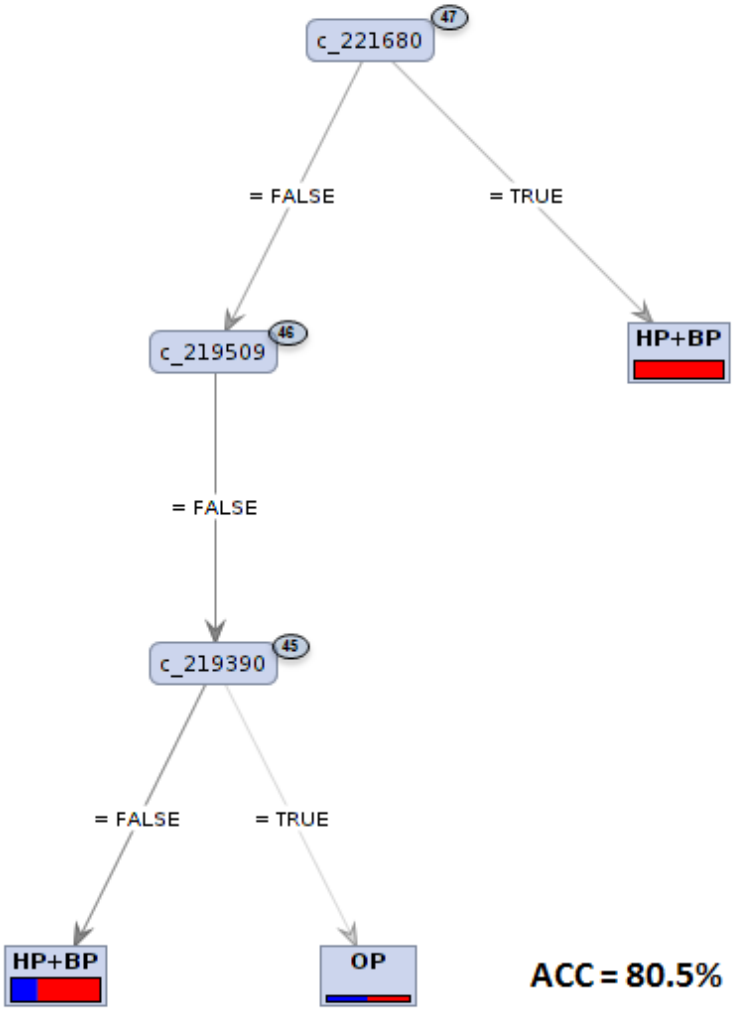

FIGURE 18 – DECISION TREE CREATED USING THE GENES MOST DISCRIMINATIVE ALL PATHOGEN (HP+BP). OUR CLASSIFICATION PIPELINE (SEE TEXT) SELECTED THE ABOVE TWO GENES AS MOST REPRESENTATIVE FOR PATHOGENS. WE LEARNED AND VISUALIZE THEM AS A SIMPLE DECISION TREE BY USING THE RAPIDMINER SOFTWARE. NODES REPRESENT GENE CLUSTERS WITH THE FOLLOWING TRANSITIVITY CLUSTERING IDS: 221680, 219509 AND 219390, WHICH ARE ASSOCIATED TO THE GENBANK ANNOTATIONS "HYPOTHETICAL PROTEIN" (E.G. UNIPROTKB AC: Q6A698), "HYPOTHETICAL PROTEIN" (E.G. UNIPROTKB AC: Q6A698) AND "COENZYME PQQ SYNTHESIS PROTEIN E" (E.G. UNIPROTKB AC: I3QX75), RESPECTIVELY. THE SMALL CIRCLES CLOSE TO THE TRANSITIVITY CLUSTERING IDS INDICATE THE CLUSTER SIZE. USING ONLY THESE THREE FEATURES THE DECISION TREE ALREADY OBTAINS AN ACCURACY OF 80.5%.

3.3 SECTION CONCLUSION

The aim of this section was to demonstrate the limited power, even of state-of-the-art bioinformatics pipelines, to fully automatically predict important bacterial lifestyles utilizing genomic information only. We illustrate and quantify the boundaries we face when trying to deduce a certain microbial pathogenicity class from the genomic repertoire, at least in the

case of Actinobacteria. We showed that we find signature genes that differentiate pathogens from non-pathogens. When trying to classify the different pathogenicity lifestyles though, it appears that too many external factors may unbalance our data sets such that we cannot be sure if we see, for instance, a strain-specific or a lifestyle-specific gene. Even in the post-genome era, and even for supposedly simple questions, our ability to efficiently deduce real-world conclusions from large-scale next-generation sequencing remains quite limited.

4 LIFESTYLE-SPECIFIC-ISLANDS

4 LIFESTYLE-SPECIFIC-ISLANDS

In this section, I will introduce and show applications for LifeStyle-Specific-Islands (LiSSI). Similarly to our previous approach, LiSSI combines evolutionary sequence analysis with statistical learning methods (Random Forest with feature selection, model tuning and robustness analysis). Plus, we included an intermediate step for island detection and an additional one for functional classification of the features (Figure 19). In summary, our strategy aims to identify conserved consecutive homology sequences (islands) in genomes and to identify the most discriminant islands for a given lifestyle.

LiSSI comes as a natural follow-up to our previous approach. Instead of solely analysing individual genes, we aim to study the evolution of genome organization. To address island detection, we included Gecko in our pipeline (for a description see “State-of-the-art” section). To address functional classification of the selected features, we relied on a BioJava [158] module to implement a Pfam search [159]. Pfam stores protein families and is used to identify conserved protein domains. Further, there is also an implementation to perform a BLAST search [160] against NCBI.

4.1 IMPLEMENTATION

LiSSI was implemented in Java and R. Java was used to generate the graphical user interface and in file manipulation operations, while R was used for the statistical analysis. LiSSI has a simple layout (Figure 20), to increase usability the analysis steps are presented as a wizard dialog. A description of the steps can be found below.

Load genomes: The first step is to load the sets of genomes associated with the lifestyles under analysis. There are three options to load the genomes: “Select from local folder”, “Download from GenBank” or a combination of both. To use local files, two sets of genomes are expected to be found in distinct directories. To use sets of genomes from NCBI, a list of all fully sequenced genomes available will be downloaded and displayed in the next step. To combine the two options, simply use one after the other.

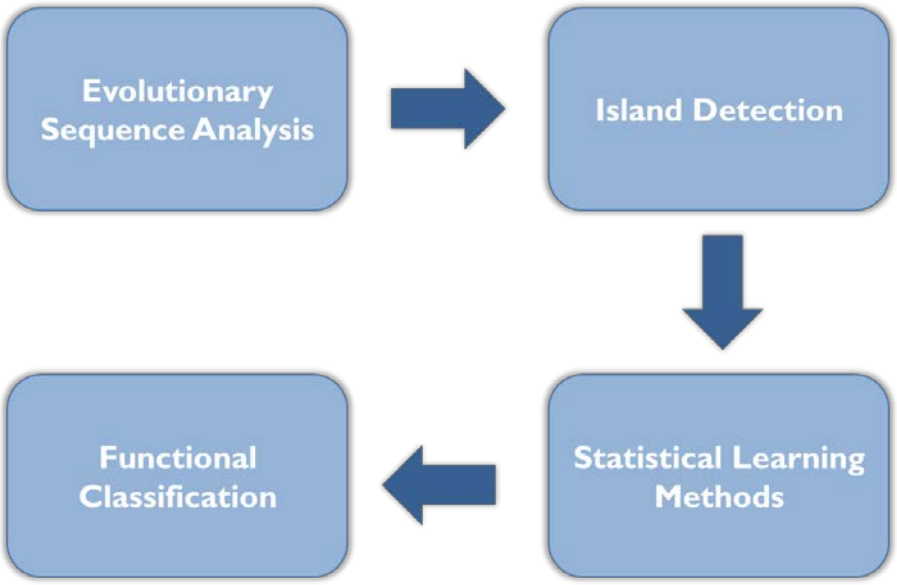

FIGURE 19 – LISSI PIPELINE. LISSI IS DIVIDED TO FOUR MODULES. A STANDARD RUN INVOLVES: THE DEFINITION OF GROUPS OF PUTATIVE HOMOLOGOUS GENES (EVOLUTIONARY SEQUENCE ANALYSIS), FOLLOWED BY ISLAND DETECTION AND IDENTIFICATION OF THE MOST DISCRIMINANT ISLANDS FOR A GIVEN LIFESTYLE (STATISTICAL LEARNING METHODS). FURTHER, FUNCTIONAL CLASSIFICATION CAN BE USED TO SEARCH FOR PROTEIN DOMAINS IN THE SELECTED GENES/ISLANDS. OPTIONALLY, THE TOOL CAN BE USED WITHOUT ISLAND DETECTION. IN THIS CASE, IT WILL REPORT PUTATIVE HOMOLOGOUS GENES THAT ARE MAINLY ASSOCIATED WITH A GIVEN LIFESTYLE.

Select genomes: The second step is to confirm the selected genomes. Basic information about the genomes loaded in the previous step will be displayed. If locally stored genomes were selected, they will be automatically displayed in the tables associated with each lifestyle. Alternatively, if NCBI genomes were selected, a list of available genomes will be displayed.

Parameters: The third and final step is displayed in Figure 20B. All parameters must be defined for Transitivity Clustering, Gecko and Random Forest. Also, it is possible to include previously generated results.

LiSSI returns the results as they are being generated. The description of the Results tab can be found below.

Clustering: It summarizes the results found during the homology detection step. It is divided in “Summary” and “Distribution”. In the Summary, it is possible to find basic information about the homology detection process, such as time required to perform BLAST and Transitivity Clustering, as well as information about the cluster size distribution. The Distribution panel contains a histogram with the cluster size distribution.

Classification: It contains all the main graphs associated with the classification process. The “Joint Distribution” depicts the distribution of the genetic features (either homologous genes or islands) among the two lifestyles. It contains a slide bar that allows for a more or less refined view of the distribution. The remaining tabs contain the ROC plots for three data-sets: the full data-set, the data-set with a bias towards class “one” (i.e., all features that were mainly found in organisms of hypothetical lifestyle “one”), and the data-set with a bias towards class “two” (i.e., all features that were mainly found in organisms of hypothetical lifestyle “two”). Each ROC plot displays the classification performance using real labels (dark-blue solid line) and using random labels (light-blue dashed line). Also, the distribution of AUC values for the distinct runs are represented as box-plots, where the values for the second and third quartiles are expressed below them. For an example, please see Figure 21.

Feature Selection: It contains the decision trees generated after feature selection. Similarly to the Classification tab, it contains decision trees for three data-sets: the full data-set, the data-set with a bias towards class “one”, and the data-set with a bias towards class “two”. By clicking in the nodes it is possible to access more information about the underling genetic feature (homologous genes or islands) or run follow up analysis (Pfam or BLAST). For an example please see Figure 22.

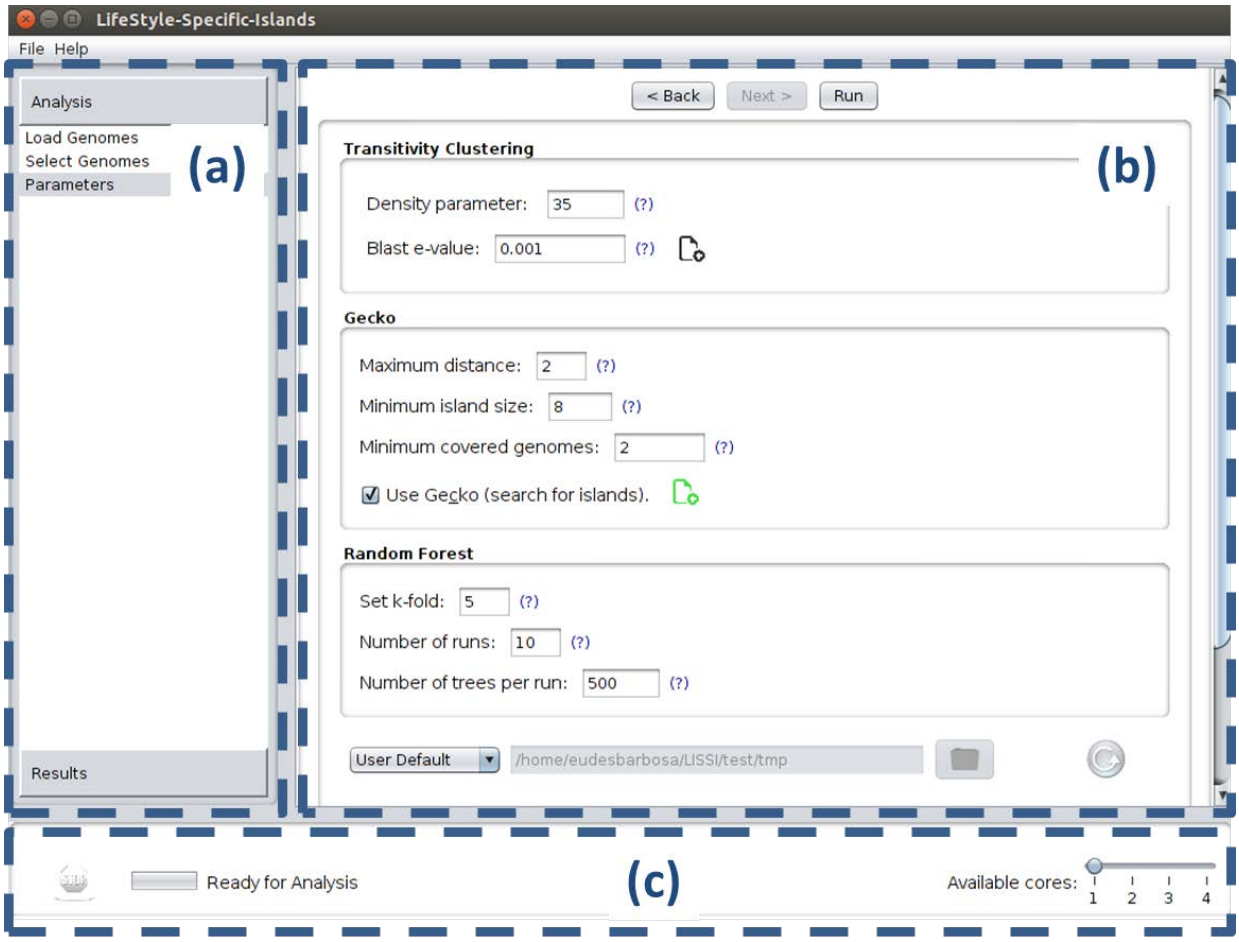

FIGURE 20 – LISSI LAYOUT. LISSI HAS THREE MAIN PARTS: A) A SELECTION MENU WITH A TAB FOR ANALYSIS AND RESULTS; B) THE MAIN PANEL, WHERE ALL INSTRUCTIONS AND RESULTS WILL BE DISPLAYED; AND C), THE PROGRESS PANEL WITH THE STATUS OF THE PROCESS CURRENTLY BEING EXECUTED. THE MAIN PANEL IS SHOWING THE LAST STEP BEFORE THE EXECUTION OF THE ANALYSIS. THERE ALL PARAMETERS ARE DEFINED, AND IT IS POSSIBLE TO INCLUDE PREVIOUSLY GENERATED RESULTS FOR TRANSITIVITY CLUSTERING AND GECKO.

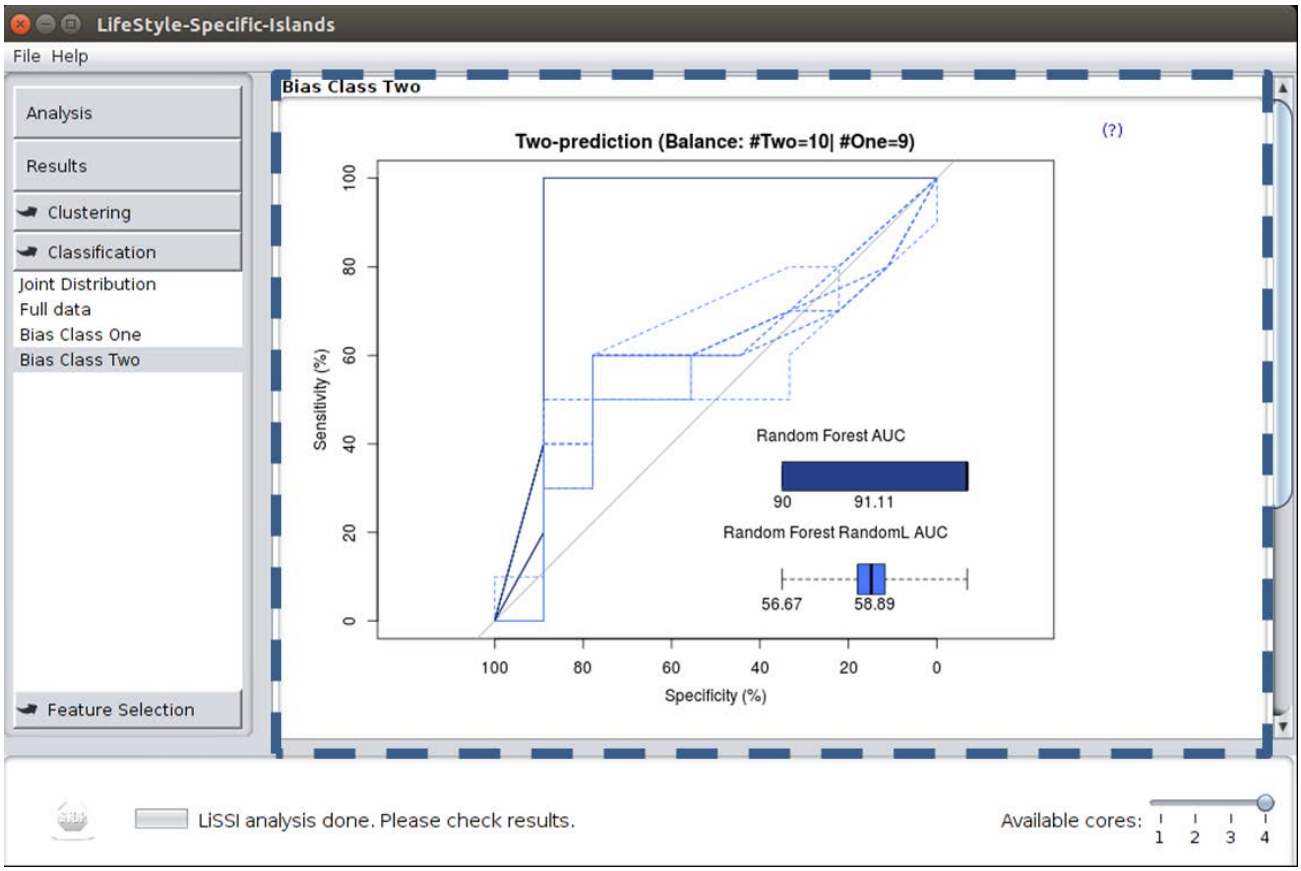

FIGURE 21 – CLASSIFICATION PERFORMANCE BETWEEN TWO HYPOTHETICAL LIFESTYLES: “ONE” AND “TWO”. THE ROC (RECEIVER OPERATING CHARACTERISTICS) PLOTS GENERATED TO INSPECT THE PERFORMANCE OF THE CLASSIFICATION MODELS ARE HIGHLIGHTED. THE DATA WAS EVALUATED FIVE TIMES USING DIFFERENT 5-FOLD CROSS-VALIDATION SETS TO ASSESS THE ROBUSTNESS OF THE CLASSIFIERS. THE REAL LABEL CLASSIFIER CURVES ARE PRESENTED AS DARK-BLUE SOLID LINES, WHILE THE RANDOM LABEL CLASSIFIERS ARE DEPICTED AS LIGHT-BLUE DASHED LINES (THE ONES CLOSE TO THE BASELINE). THE VARIATION OF THE AUCS (AREA UNDER CURVE) IN THE CROSS-VALIDATION WAS INCLUDED IN THE FIGURE AS A BOX-PLOT (BOTTOM RIGHT). THE NUMBERS BELOW EACH BOX-PLOT ARE THE LOWER AND UPPER QUARTILES.

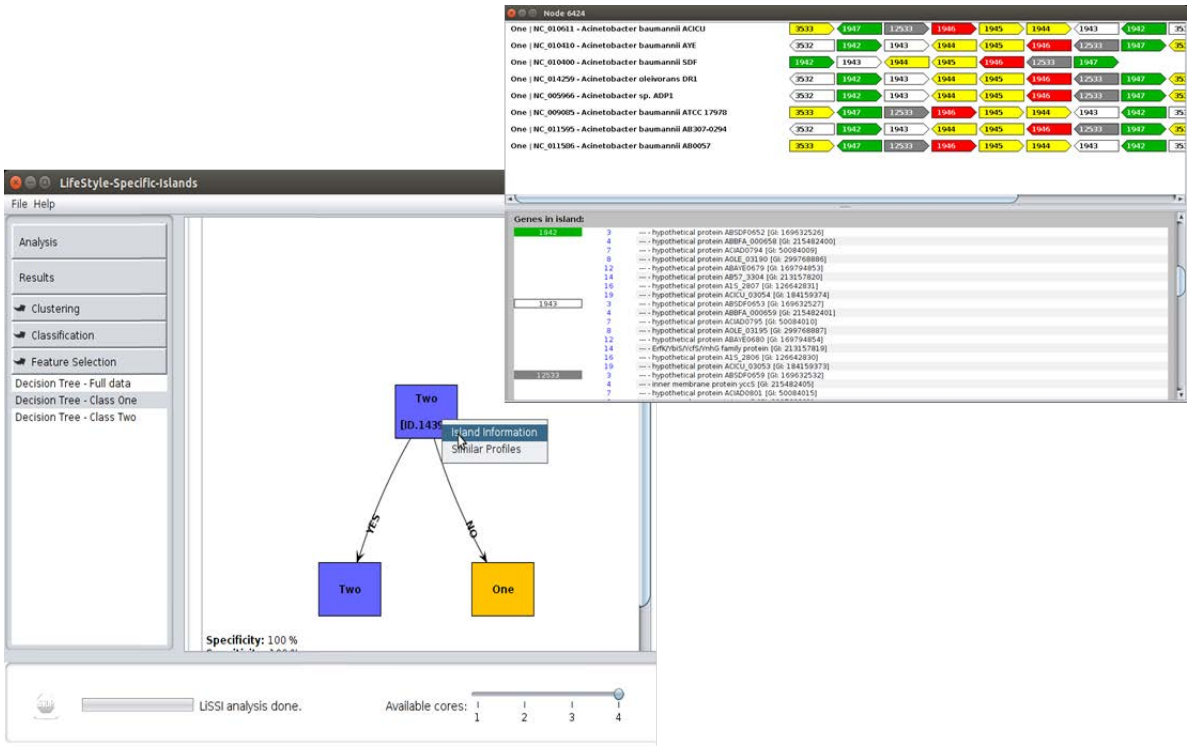

FIGURE 22 – DECISION TREE CREATED USING THE MOST DISCRIMINATIVE ISLANDS FOR THE HYPHOTHETICAL LIFESTYLE “ONE”. OUR CLASSIFICATION PIPELINE (SEE TEXT) SELECTED THE ABOVE ISLAND AS THE MOST REPRESENTATIVE FOR HYPOTHETICAL LIFESTYLE “ONE”. NODES CONTAINING AN IDENTIFIER REPRESENT A GENETIC FEATURE, IN THIS CASE, AN ISLAND. BY CLICKING ON THE NODES IT IS POSSIBLE TO VISUALIZE THE ISLAND STRUCTURE, AS WELL AS THE GENE CONTENT AND THE GENOMES THAT PRESENT IT.

4.2 VALIDATION

We created different sets of artificial data to validate the results generated by the different modules. We mainly focus on the tool ability to detect the presence of a given island in the genomes and the impact it might have on the classification performance.

4.2.1 ARTIFICIAL GENOMES

We started our analysis of the tool by checking if it was indeed capable of detecting the presence of a genomic island. Further, we were interested in the percentage of organisms that a given island would have to be present to impact the classification performance. Therefore, we create two hypothetical bacterial lifestyles, denoted as lifestyles “Alpha” and “Omega”. Both Alpha and Omega were composed

of a set of 100 genomes with 100 genes each, where a genomic island could be present or not. The islands were randomly positioned in the genomes.

Each of the genes was arbitrarily assigned to a group of homologous genes based on being part of an island or not. To ensure that each lifestyle contained only a single island, all genes that did not belong to an island received unique identifiers, i.e., no such identifier was used more than once in all 200 genomes. A small subset of 16 identifiers was selected to describe genes in each of the lifestyles' islands (eight per lifestyle). These restrictions can be observed in Figure 23, where nearly all clusters are singletons and only the islands' clusters have more than 75 genes. Further, the islands were created to simulate the variability found in nature. Thus, roughly 20% of islands presented the full length (eight genes), 45% presented a single deletion (seven genes) and 35% presented two deletions (six genes). Figure 24 depicts the variability among the islands included in the artificial genomes.

We chose to investigate the impact of the classification performance when 10, 25, 50, 75, 90 or 100% of the genomes in one of the two lifestyles contained the island. For that, 36 comparisons were performed with all possible combinations of percentages among the lifestyles. We followed the default LiSSI run, skipping the bias introductions, because they were not applicable. The island detection parameters were set as: minimum number of genomes equals two; minimum size equals eight; and, maximum indels equals two. The classification parameters were set as: ten runs using different five-fold cross-validation sets, growing 50 trees per run. The number of trees was kept low since we were not expecting a lot of variation with such low number of features.

In all cases, the pipeline was capable of detecting the island in all organisms in which it was inserted, and only the inserted islands were reported. The observed trends were similar in all comparisons and are summarized in Figure 25. As expected, classification performance increased with the percentage of organisms that possessed the island. Plus, if a single island is significantly present in a set of genomes, apparently it does not matter if the other set has a distinguishable island.

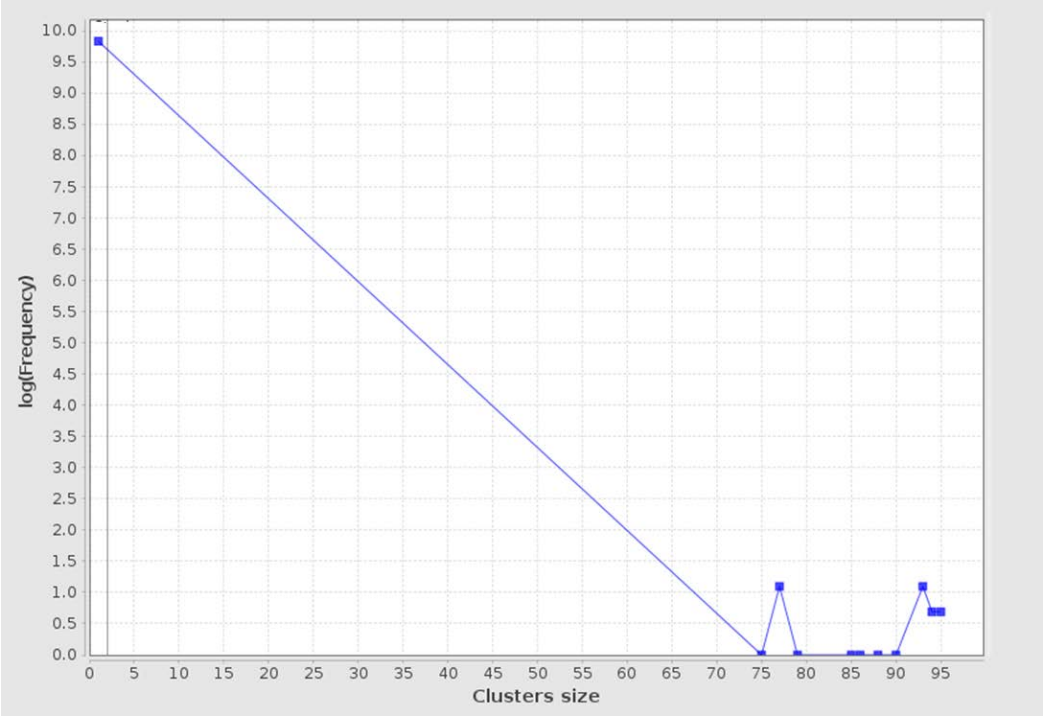

FIGURE 23 – CLUSTER SIZE DISTRIBUTION FOR ARTIFICIAL GENOMES. IN THIS PARTICULAR EXAMPLE, ALL ALPHA AND OMEGA GENOMES CONTAIN AN ISLAND. SINGLETONS REPRESENT GENES THAT ARE *NOT* PART OF AN ISLAND, WHILE CLUSTERS WITH MORE THAN 75 GENES REPRESENT GENES IN ISLANDS.

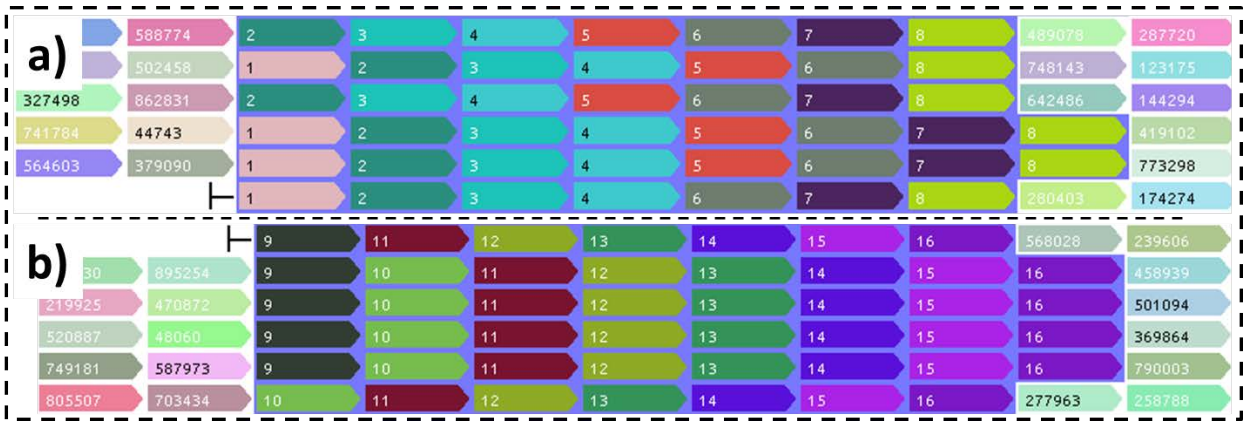

FIGURE 24 – ARTIFICIAL ISLANDS. EXAMPLE OF VARIATION IN GENE CONTENT FOUND IN THE ISLANDS INCLUDED IN THE ARTIFICIAL GENOMES: A) ALPHA ISLANDS, B) OMEGA ISLANDS.

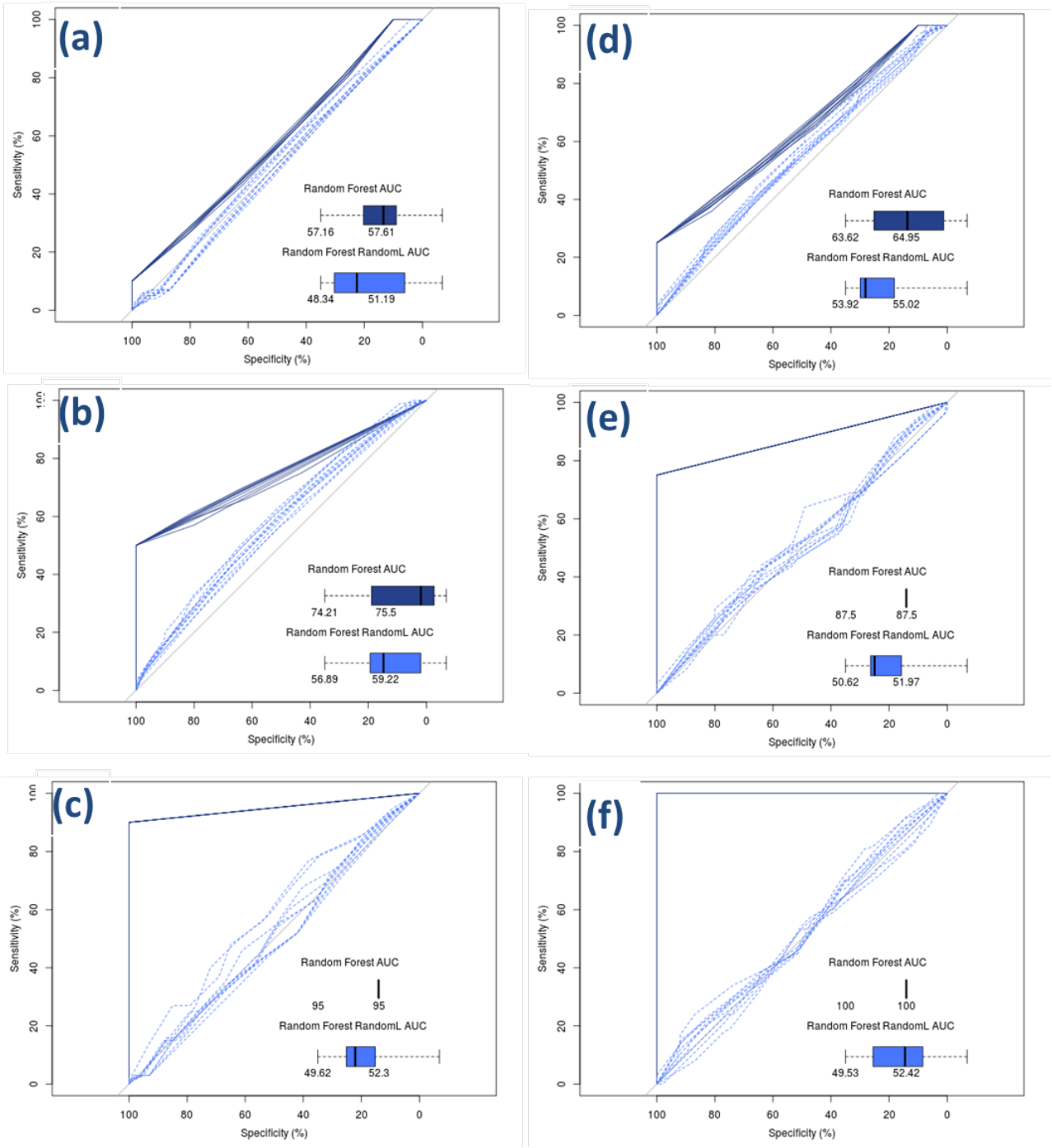

FIGURE 25 – ARTIFICIAL LIFESTYLES CLASSIFICATION PERFORMANCE. THE GRAPHS SHOW THE CLASSIFICATION PERFORMANCE FOR DIFFERENT SETS OF ALPHA AND OMEGA GENOMES. AS EXPECTED, CLASSIFICATION PERFORMANCE INCREASES WITH THE NUMBER OF GENOMES THAT CONTAIN THE ISLAND. IN ALL COMPARISONS 10% OF ALPHA GENOMES HAVE THE ISLAND, WHILE IN OMEGA THEY ARE PRESENT IN: A) 10%, $\overline{AUC}$ = 57.4%; B) 25%, $\overline{AUC}$ = 64.3%; C) 50%, $\overline{AUC}$ = 74.8%; D) 75%, $\overline{AUC}$ = 87.5%; E) 90%, $\overline{AUC}$ = 95.0%; F) 100%, $\overline{AUC}$ = 100%.

4.2.2 MODIFIED GENOMES

To evaluate if the whole pipeline – from homology detection to feature selection – was working properly, we analysed phylogenetically distant organisms instead of lifestyles. Therefore, we selected 10 genomes from the genus *Listeria* and

12 genomes from the genus *Corynebacterium*. Given the evolutionary proximity between organisms of the same genus, a high level of synteny was expected. Thus, we would not be classifying based on lifestyles but rather on phylogenetic proximity. For the complete list of species, see S. Table 2 in **Appendix A**.

To ensure the presence of at least one discriminative island for each lifestyle and to evaluate our homology detection method, we also tested a scenario that included an exogenous island. The genes were extracted from two phylogenetically distant organisms in the hope that the genes in the inserted islands were completely unrelated to the native genes. Genes from a pathogenic island were extracted from *Escherichia albertii* (accession number: NZ_CP007025), a potential human enteric pathogen; metabolic genes were extracted from *Caldicellulosiruptor owensensis* (accession number: NC_014657), a thermophilic organism. The complete list of included genes from *E. albertii* and *C. owensensis* can be found in Table 2 and Table 3, respectively.

We followed the default LiSSI run for the real genomes as well as for the modified ones. The island detection parameters were set as: minimum number of genomes equals four; minimum size equals eight; and, maximum indels equals one. The classification parameters were set as: ten runs using different 5-fold cross-validation sets, growing 500 trees per run.

TABLE 2 – GENES IN PATHOGENIC ISLANDS FOUND IN *ESCHERICHIA ALBERTII*.

Gene Identifier	Name	Product
446960572	---	secretion system apparatus protein SsaV
643603877	---	T3SS regulator Mpc
643603880	---	type III secretion system protein SepZ
643603884	---	type III secretion apparatus protein
446638846	---	secretion system apparatus lipoprotein EscJ
446986427	---	type III secretion system protein SepD
643603890	ssaC	outer membrane secretin SsaC
446009609	---	hypothetical protein

TABLE 3 – METABOLIC GENES FOUND IN *CALDICELLULOSIRUPTOR OWENSENSIS*, A THERMOPHILIC ORGANISM.

Gene Identifier	Name	Product
503177880	---	protein-tyrosine phosphatase
503177881	---	arsenic resistance protein ArsB
503177882	---	MBL fold metallo-hydrolase
503177883	---	hypothetical protein
754099456	---	amino acid ABC transporter ATPase
503177885	---	glutamine ABC transporter permease
503177886	---	transporter substrate-binding protein
503177887	---	S-layer protein

4.2.2.1 CLASSIFICATION

Figure 26 and Figure 27 summarize the classification performance. As expected, the classifiers based on both *Corynebacterium* and *Listeria* biases worked well. In both cases, the classification using the real labels had AUC equal to 100% in all runs; while the classification using random labels had AUC oscillating little above 50%. Further, this was valid for data-sets with and without insertion of the exogenous island. The main difference between the two cases was the decrease in the $\overline{AUC}$ for the random label classifiers. The reduction was particularly noticeable for the bias towards *Corynebacterium*, where it dropped from 62.5% to 55.4%. Nevertheless, these variations could be due to the random nature of the process.

The decision trees generated for both data-sets presented similar results, correctly classifying nearly all genomes using a single island. Further, none of the trees for the data-set with the exogenous islands were based on the inserted islands. Given the phylogenic distance between the two geni, it was expected that many genomic regions (islands) would differ between groups, where the exogenous islands would carry the same information as many other features. The gene content of the discriminate islands was not further investigated since it was relevant to the evaluation.

4.2.2.2 HOMOLOGY DETECTION

Transitivity Clustering, our homology detection method, correctly assigned all instances of the genes from the exogenous islands to the correct groups. Contrarily to what was expected, some of the genomes indeed contained genes closed related to the ones from the exogenous islands. That was the case for gene 446960572 from

the pathogenic island (*E. albertii*) and genes 754099456, 503177885, and 503177886, associated with membrane transport in the metabolic "island" (*C. owensensis*). We checked the alignments of the putative homologous sequences for all the previous cases to confirm that they were correct; Figure 28 depicts one of such alignments.

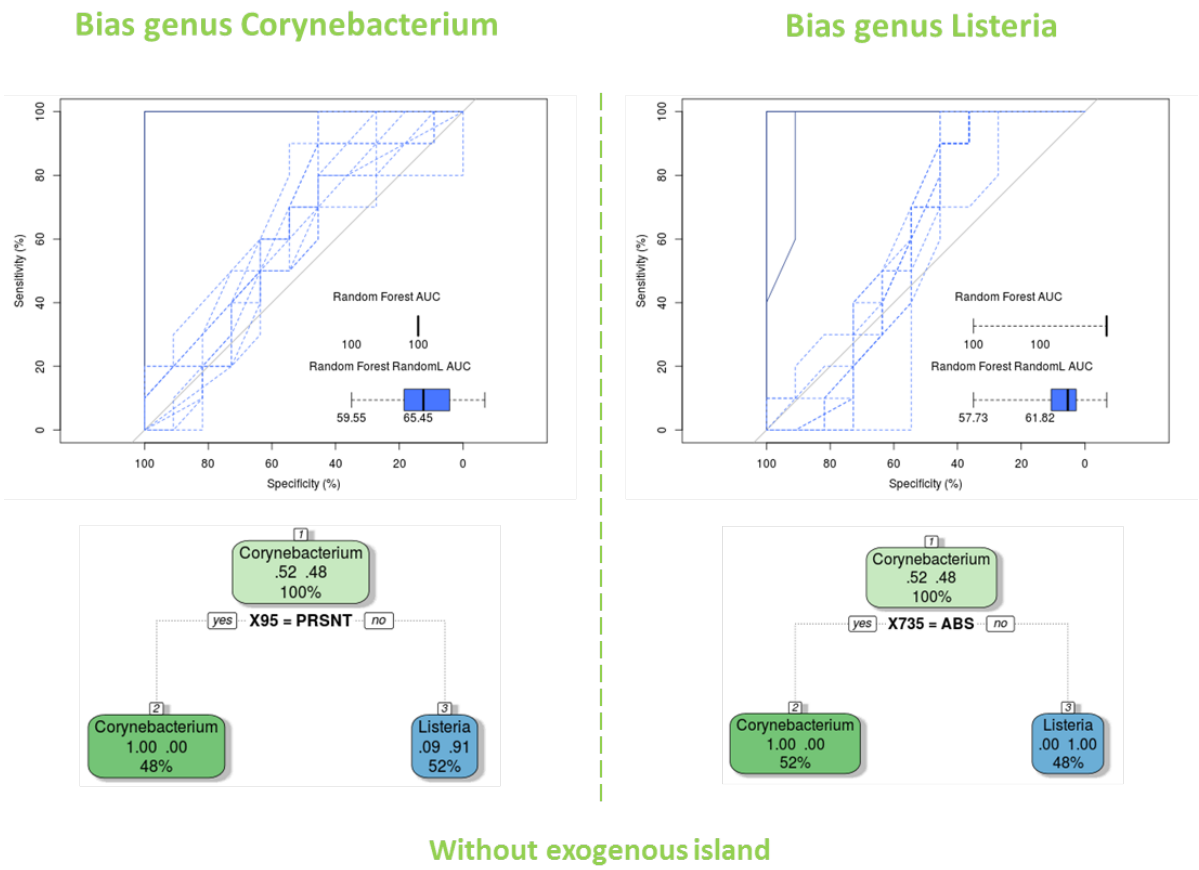

FIGURE 26 – SUMMARY OF THE CLASSIFICAITON PERFORMANCE FOR THE DATA-SET WITHOUT THE EXOGENOUS ISLANDS. BOTH BIASES PRESENTED PERFECT CLASSIFICATION USING REAL LABELS (DARK SOLID LINE) AND $\overline{AUC}$ SLIGHTLY ABOVE 50% FOR RANDOM LABELS. ALSO, IN BOTH CASES A SINGLE ISLAND WAS ENOUGH TO CORRECTLY CLASSIFY NEARLY ALL GENOMES IN THE DECISION TREE. ABS STANDS FOR "ABSENT" AND PRSNT FOR "PRESENT". LEFT: ROC PLOT AND DECISION TREE GENERATED WITH FEATURES BIASED TOWARDS *CORYNEBACTERIUM*. RIGHT: ROC PLOT AND DECISION TREE GENERATED WITH FEATURES BIASED TOWARDS *LISTERIA*.

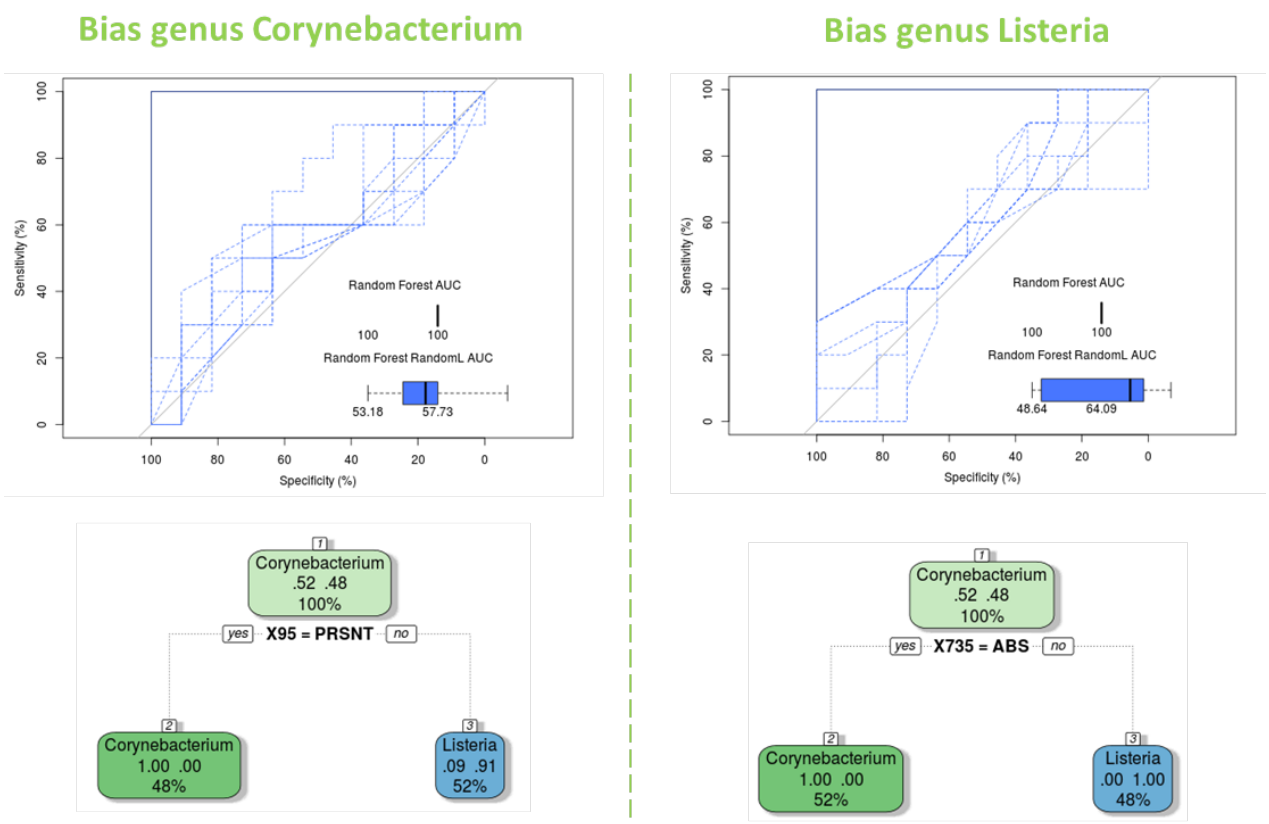

FIGURE 27 – SUMMARY OF THE CLASSIFICAITON PERFORMANCE FOR THE DATA-SET WITH THE EXOGENOUS ISLAND. BOTH BIAS PRESENTED PERFECT CLASSIFICATION USING REAL LABELS (DARK SOLID LINE), AND $\overline{AUC}$ SLIGHTLY ABOVE 50% FOR RANDOM LABELS. IN BOTH CASES, A SINGLE ISLAND WAS ENOUGH TO CORRECTLY CLASSIFY NEARLY ALL GENOMES IN THE DECISION TREE. ABS STANDS FOR “ABSENT” AND PRSNT FOR “PRESENT” LEFT: ROC PLOT AND DECISION TREE GENERATED WITH FEATURES BIASED TOWARDS *CORYNEBACTERIUM*. RIGHT: ROC PLOT AND DECISION TREE GENERATED WITH FEATURES BIASED TOWARDS *LISTERIA*.

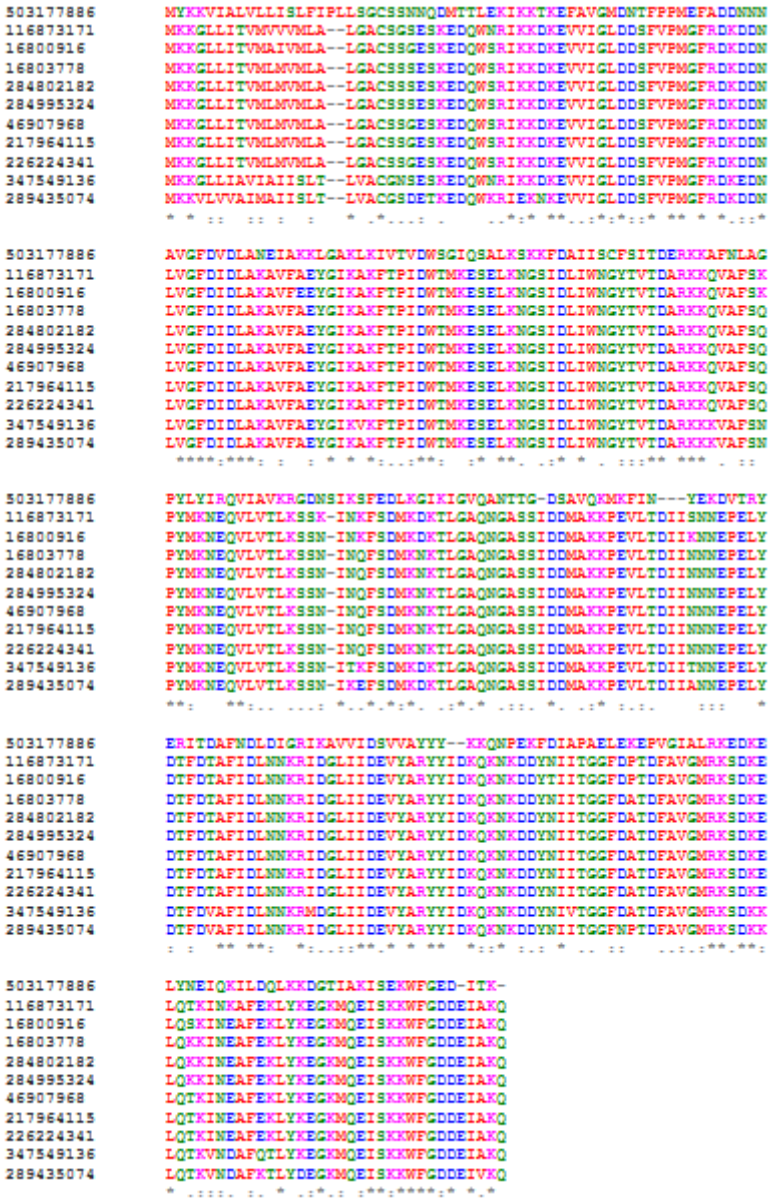

FIGURE 28 – SEQUENCE ALIGNMENT FOR SEQUENCES ASSOCIATED WITH PROTEIN 503177886. THE CLUSTER FOR PROTEIN 503177886 CONTAINED ELEVEN PROTEINS IN TOTAL. THE EXTENSION OF THE ALIGNMENT AMONG THE PROTEINS SUPPORT THE CLUSTERING RESULTS. THE ALINGMENT WAS GENERATED USING THE DEFAULT PARAMETER OF THE TOOL MUSCLE, AVAILABLE AT HTTP://WWW.EBI.AC.UK/TOOLS/MSA/MUSCLE/.

4.2.3 REAL DATA

After the validation rounds, we continued to evaluate the tool using real data. In this case, we selected all fully sequenced genomes from Actinobacteria that have information available about their lifestyles. In this section, we described the methodology we applied and the results we obtained.

4.2.3.1 METODOLOGY

4.2.3.1.1 GENOMES AND LIFESTYLES

We selected all 202 completely sequenced Actinobacterial genomes that belonged to at least one of the following lifestyles: aerobes (AE), anaerobes (AN), facultative (FA), soil (SO), aquatic (AQ), non-pathogenic (NP), and pathogenic (PA). The annotations for habitat and oxygen tolerance were extracted from fusionDB [161], while the pathogenicity annotations were extracted from our previous work [162]. In total we had 63 AE, 23 AN, 9 FA, 34 SO, 9 AQ, 112 NP, and 87 PA. The whole-genome annotation was downloaded from NCBI for the complete list of species and respective pathogenicity classification see S. Table 3 in **Appendix A**.

4.2.3.1.2 PARAMETERS

We followed the default LiSSI run for all comparisons. The homology detection parameter was set as 35, the lower bound of the interval of reasonable values for Actinobacterial species [152]. The island detection parameters were set as: minimum number of genomes equals two; minimum size equals eight; maximum indels equals two. The classification parameters were set as: ten runs using different 5-fold cross-validation sets, growing 500 trees per run.

4.2.3.1.3 FUNCTIONAL CLASSIFICATION

We followed two approaches to classify the genes found in our analysis. For the homologous gene analysis, we searched for conserved protein domains and families using Pfam (http://pfam.xfam.org/). For the islands analysis, we adapted the approach described in [125] and searched for similar genes in different databases. We restrain our search to virulence, resistance and metabolic databases. We used E-value cut-off of 10^{-6} and similarity of at least 50%.

4.2.3.1.4 PATHOGENICITY

As a sanity check, we started our analysis using a data-set similar to the one used in our previous approach (see “Homologous sequence analysis”). Instead of trying to separate the organism into pathogenicity sub-classes, we combined all

pathogens – generally speaking, all associated with mammal hosts – and we tried to distinguish them from non-pathogens. The hope was to find similar results and to compare the impact of the use of islands instead of homologous genes for classification and feature selection.

4.2.3.1.4.1 NON-PATHOGENIC VS. PATHOGENIC

Similarly to our previous results, we were able to observe homologous genes exclusively found in either pathogens or non-pathogens. We were also able to observe islands exclusively found in one of the two classes (Figure 29). As was expected, the number of features was dramatically reduced for the analysis using islands. We found 375,427 distinct homologous genes, where 317,751 where mainly present in non-pathogens and 57,676 in pathogens. The situation is exactly the opposite for islands. Most of the 465 islands are mainly present in pathogens (386); the remaining 79 are mainly present in non-pathogens. Further, there is no island that is present in more than 35% of the organisms (including non-pathogens and pathogens).

The classification results were fairly different for homologous genes and islands (Figure 30). The analysis using homologous genes followed the same trend as previous observed (see “Homologous sequence analysis”). The classifiers had good performance for both non-pathogen bias ($\overline{AUC}$ = 94.16%, Figure 30B) and pathogen bias ($\overline{AUC}$ = 93.91%, Figure 30C), as well as for the classifiers using the full data-set ($\overline{AUC}$ = 94.81%, Figure 30A). These results indicate that we find both, gene sets specific for pathogens, as well as gene sets specific for non-pathogens, with almost identical accuracy. On the other hand, the scenario is fairly different for the analysis using islands. Overall, the classification performance dropped, where the non-pathogen bias performed poorly ($\overline{AUC}$ = 63.61%, Figure 30E), and the pathogen bias was worse than its homologous genes counterpart ($\overline{AUC}$ = 88.84%, Figure 30F).

The most discriminative homologous genes were used to create the decision trees in Figure 31. For the bias towards pathogens, the selected clusters were: 9811 (76 genes, associated with Ribosomal_S7 domain), 31894 (30 genes, associated with GMC_oxred_N domain), 149120 (38 genes, associated with MraZ family), 274546 (11 genes, associated with Ribosomal_L5 domain) and 281756 (4 genes, associated with ABC_tran domain). The Pfam results can be observed in S. Table 4. For the bias towards non-pathogens, the selected cluster identifiers were: 1025 (28 genes, associated with Adenylsucc_synt domain), 1565 (27 genes, associated with Thiolase_N and Thiolase_C domains), 1704 (25 genes, associated with HTH_18

domain), 1851 (94 genes, associated with different RNA_pol domains), 4318 (64 genes, associated with ABC_tran domain), 8006 (32 genes, associated with ABC_tran_Xtn and ABC_tran domains), 12433 (11 genes, associated with different ThiC domains), 13007 (25 genes, associated with IGPD family), 14351 (47 genes, associated with GcpE family), 23225 (74 genes, associated with Ribosomal_S7 domain), 28316 (10 genes, associated with RNase_PH and RNase_PH_C domains), 29574 (10 genes, associated with GTP_cyclohydro2 and GTP_CH_N domains), and 35107 (100 genes, associated with Ribosomal_L34 Family). The Pfam results can be observed in S. Table 5.

The most discriminative islands were used to create the decision trees in Figure 32. For the bias towards pathogens, the selected island identifiers were: 70890 (length 8, present in 24 organisms), 21890 (length 8, present in 27 organisms), 56705 (length 17, present in 10 organisms), and 95033 (length 56, present in 10 organisms). For the bias towards non-pathogens, the selected island identifiers were: 170 (length 19, present in 26 organisms) and 32195 (length 20, present in 12 organisms). Table 4 describes the amount of genes associated with metabolic, resistance and virulence that are present in the islands.

TABLE 4 – DESCRIPTION OF THE MOST DISCRIMINATIVE ISLANDS FOR NON-PATHOGENS AND PATHOGENS. "LENGTH" REPRESENTS THE AMOUNT OF CONSECUTIVE GROUPS OF HOMOLOGOUS SEQUENCES; WHILE THE "HOMOLOGOUS SEQUENCES" REPRESENT THE SUM OF ALL GENES IN THE GROUPS OF HOMOLOGOUS SEQUENCES.

Island	Length	Homologous sequences	Metabolic	Resistance	Virulence
70890	8	435	236	0	0
21890	8	311	172	0	0
56705	17	242	23	0	21
95033	56	2080	554	22	149
170	19	1071	0	0	0
32195	20	2179	0	0	0

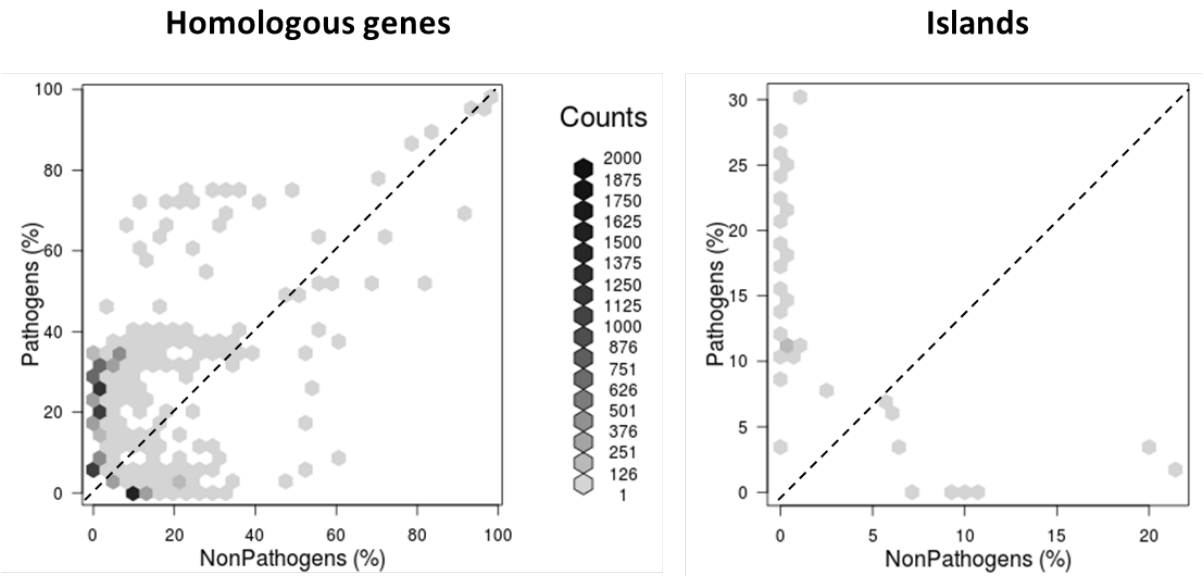

FIGURE 29 – DISTRIBUTION OF GENETIC FEATURES OVER TWO LIFESTYLES (PATHOGENS VS. NON-PATHOGENS). BOTH AXES IN THE PLOT DESCRIBE THE PERCENTAGES OF SPECIES IN THE RESPECTIVE CLASS(ES), HERE PATHOGENS (PA) AND NON-PATHOGENS (NP). THE COLOR-CODING OF THE HEAT MAP DEPICTS THE NUMBER OF CLUSTERS OF HOMOLOGOUS GENES THAT CERTAIN PERCENTAGES OF PATHOGENS/NON-PATHOGENS SHARE. LEFT: HOMOLOGOUS GENE DISTRIBUTION. RIGHT: ISLAND DISTRIBUTION.

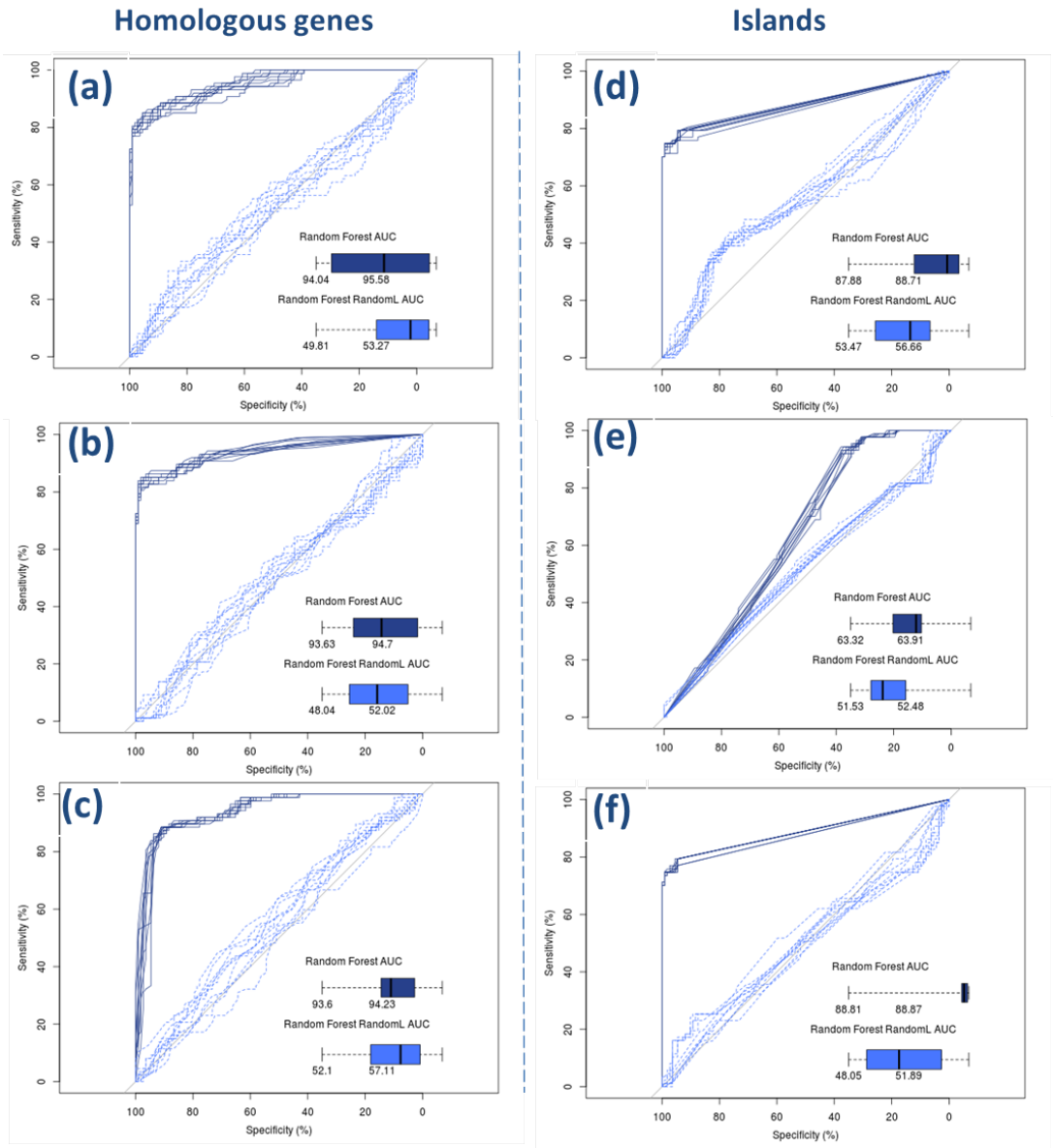

FIGURE 30 – CLASSIFICATION PERFORMANCE, PATHOGENS VS. NON-PATHOGENS. FOR EACH ROC PLOT, THE REAL LABEL CLASSIFIER CURVES ARE PRESENTED IN DARK-BLUE SOLID LINES, WHILE THE RANDOM LABEL CLASSIFIERS ARE PRESENTED AS LIGHT-BLUE DASHED LINES (THE ONES CLOSE TO THE BASELINE). THE VARIATION OF THE AUCS (AREA UNDER CURVE) IN THE CROSS-VALIDATION WAS INCLUDED IN THE FIGURE AS A BOX-PLOT (BOTTOM RIGHT). THE NUMBERS BELOW EACH BOX-PLOT ARE THE LOWER AND UPPER QUARTILES. HOMOLOGOUS GENES: A) FULL DATA-SET ($\overline{AUC}$ = 94.81%, $\overline{AUC}_{\text{RL}}$ = 51.54%); B) BIAS NON-PATHOGEN ($\overline{AUC}$ = 94.16%, $\overline{AUC}_{\text{RL}}$ = 50.03%); C) BIAS PATHOGENS ($\overline{AUC}$ = 93.91%, $\overline{AUC}_{\text{RL}}$ = 54.06%). ISLANDS: D) FULL DATA-SET ($\overline{AUC}$ = 88.29%, $\overline{AUC}_{\text{RL}}$ = 55.06%); BIAS NON-PATHOGENS ($\overline{AUC}$ = 63.61%, $\overline{AUC}_{\text{RL}}$ = 52.00%); BIAS PATHOGENS ($\overline{AUC}$ = 88.84%, $\overline{AUC}_{\text{RL}}$ = 49.97%).

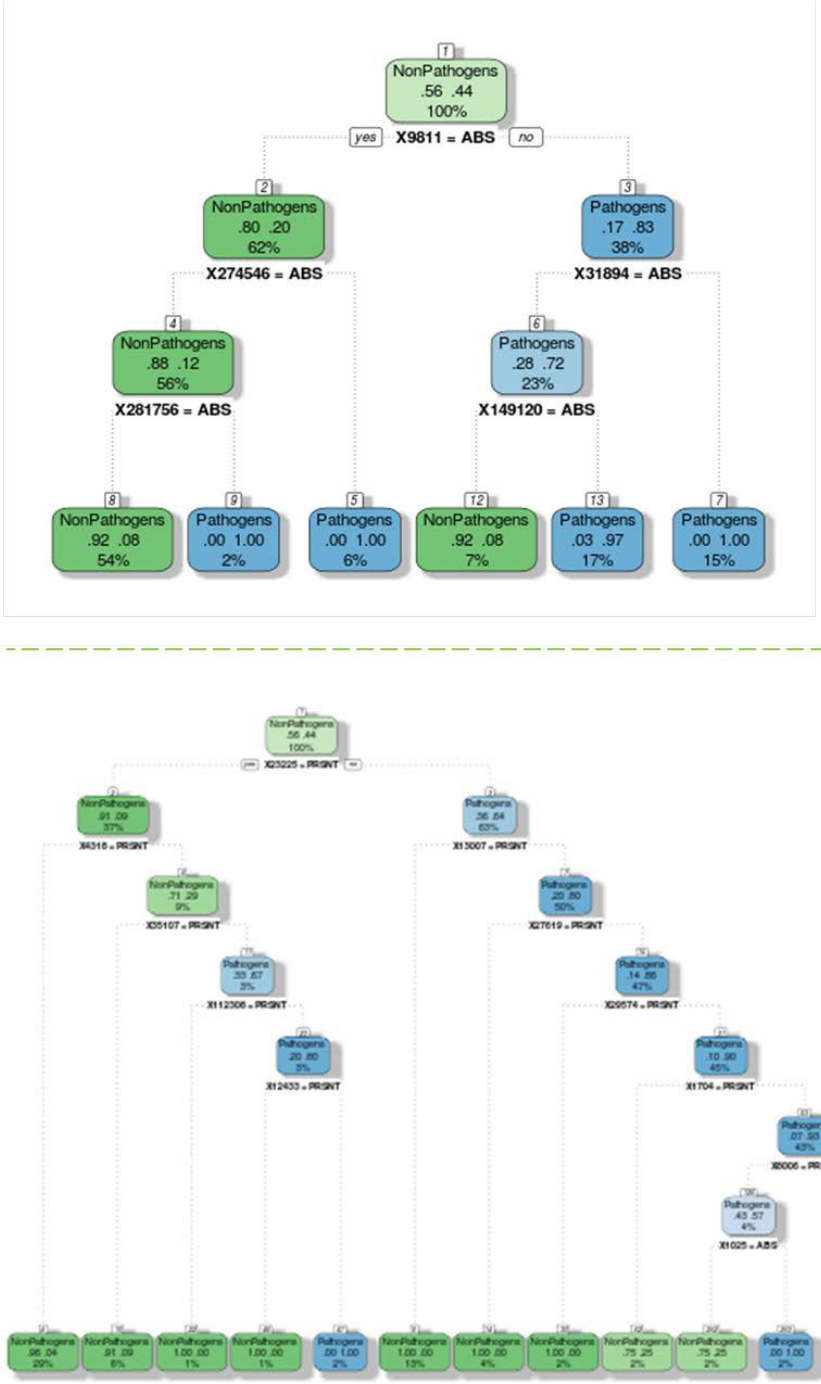

FIGURE 31 – DECISION TREES FOR HOMOLOGOUS GENES (NON-PATHOGENS VS. PATHOGENS). DECISION TREES CREATED USING THE MOST DISCRIMINATIVE FEATURES FOR BOTH BIASES. ABS STANDS FOR “ABSENT” AND PRSNT FOR “PRESENT”. TOP: DECISION TREE FOR PATHOGENS (ACCURACY: 94.8, PRECISION: 91.8%). BOTTOM: DECISION TREE FOR NON-PATHOGENS (ACCURACY: 96.7%, PRECISION: 94.9%).

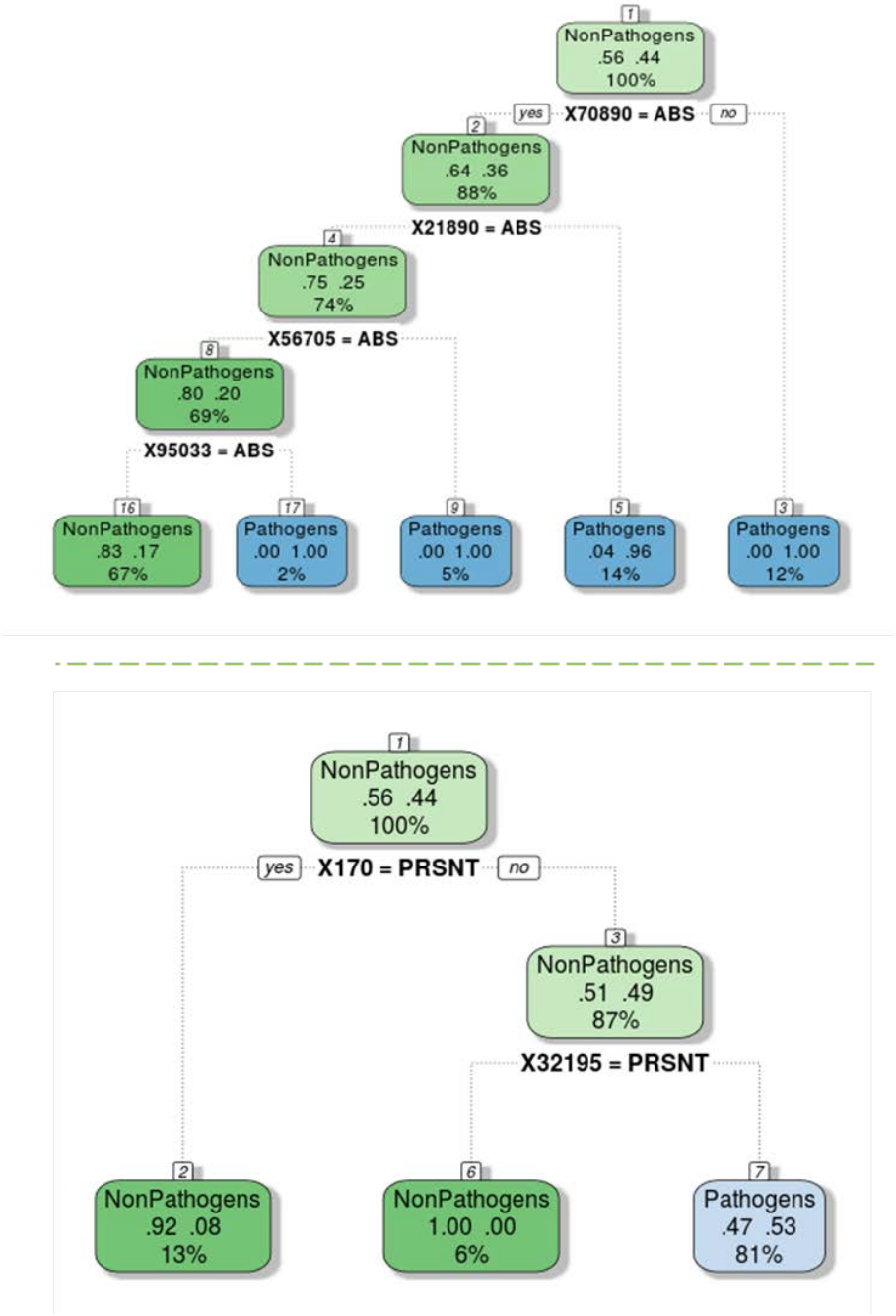

FIGURE 32 – DECISION TREES FOR ISLANDS (NON-PATHOGENS VS. PATHOGENS). DECISION TREES CREATED USING THE MOST DISCRIMINATIVE FEATURES FOR BOTH BIASES. ABS STANDS FOR “ABSENT” AND PRSNT FOR “PRESENT”. TOP: DECISION TREE FOR PATHOGENS (ACCURACY: 88.0%, PRECISION: 72.8%). BOTTOM: DECISION TREE FOR NON-PATHOGENS (ACCURACY: 60.9%, PRECISION: 31.5%).

4.2.3.1.4.2 OXYGEN TOLERANCE

We followed our analysis by trying to identify genetic features associated with different classes of atmospheric oxygen tolerance, namely: aerobe, facultative, and anaerobe. Given the metabolic differences regarding the tolerance to oxidation and

cellular respiration, we expect to find sets of genes and hopefully islands that distinguish all lifestyles.

4.2.3.1.4.3 AEROBE VS. ANAEROBE

We found 335,532 distinct homologous genes, where 198,529 where mainly present in aerobes and 28,974 in anaerobes. The situation was the same for islands, where most of the 181 islands were mainly present in aerobes (107); the remaining 74 were mainly present in anaerobes. The distribution of homologous sequences and islands can be observed in Figure 33.

Again, the classification results were fairly different for homologous genes and islands (Figure 34). The classifiers had good performance for both aerobe bias ($\overline{AUC}$ = 92.48%, Figure 34B) and anaerobe bias ($\overline{AUC}$ = 99.15%, Figure 34C), as well as for the classifier using the full data-set ($\overline{AUC}$ = 95.15%, Figure 34A). These results indicate that we find both gene sets specific for aerobes as well as gene sets specific for anaerobes, with almost identical and high accuracy. On the other hand, the scenario is fairly different for the analysis using islands. Overall, the classification performance dropped, where the aerobe bias performed poorly ($\overline{AUC}$ = 66.51%, Figure 34E), and the anaerobe bias was worse than its homologous gene counterpart ($\overline{AUC}$ = 78.26%, Figure 34F).

The most discriminative homologous genes were used to create the decision trees in Figure 35. For the bias towards aerobes, the selected clusters were: 6725 (70 genes, associated with ClpB_D2-small, zf-C4_ClpX and AAA_2 domains), 4030 (90 genes, associated with several RNA_pol domains), and 2456 (55 genes, associated with ABC_tran). The Pfam results are presented in S. Table 6. For the bias towards anaerobes, the selected cluster identifiers were: 28912 (14 genes, associated with HSP70 family), 19398 (4 genes, associated with Ribosomal_S19 domain), 98168 (6 genes, associated with Terminase_4 family), and 2543 (3 genes, not associated with any domain or family). The Pfam results can be observed in S. Table 7.

The most discriminative islands were used to create the decision trees in Figure 36. It was not possible to create any meaningful decision tree for the aerobe bias. For the bias towards anaerobes, the selected island identifier was: 15 (length 19, present in 13 organisms). Table 5 describes the amount of genes associated with metabolic, resistance and virulence that are present in the island.

TABLE 5 – DESCRIPTION OF THE MOST DISCRIMINATIVE ISLAND FOR ANAEROBES. “LENGTH” REPRESENTS THE AMOUNT OF CONSECUTIVE GROUPS OF HOMOLOGOUS SEQUENCES; WHILE THE “TOTAL GENES” REPRESENT THE SUM OF ALL GENES IN THE GROUPS OF HOMOLOGOUS SEQUENCES.

Island	Length	Homologous sequences	Metabolic	Resistance	Virulence
15	19	2140	162	7	18

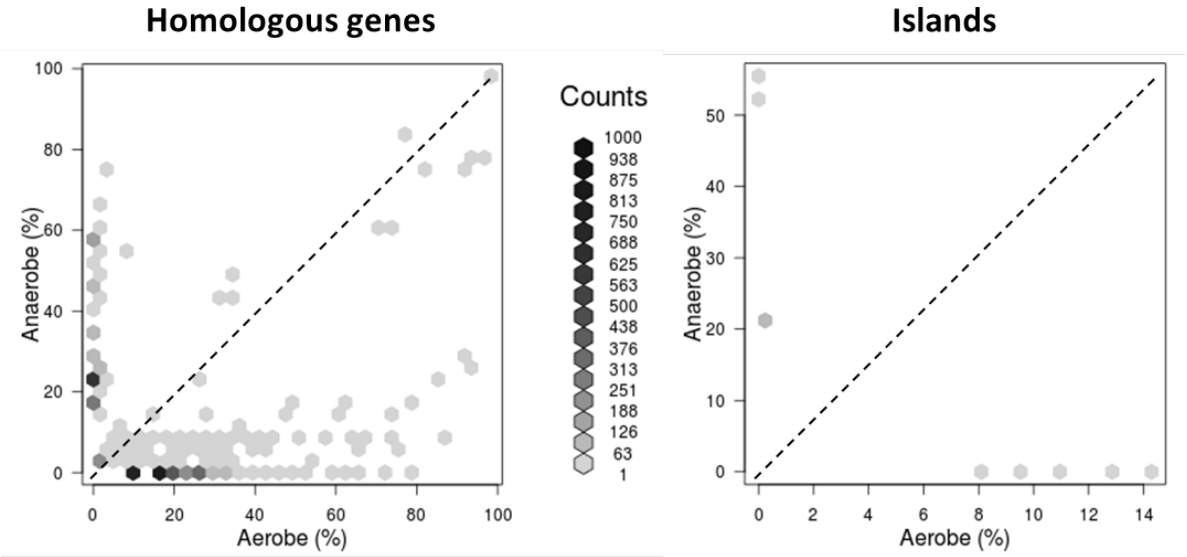

FIGURE 33 – DISTRIBUTION OF GENETIC FEATURES OVER TWO LIFESTYLES (AEROBIC VS. ANAEROBIC). BOTH AXES IN THE PLOT DESCRIBE THE PERCENTAGE OF SPECIES IN THE RESPECTIVE CLASS(ES), HERE AEROBES (AE) AND ANAEROBES (AN). THE COLOR-CODING OF THE HEAT MAP DEPICTS THE NUMBER OF CLUSTERS OF HOMOLOGOUS GENES THAT CERTAIN PERCENTAGES OF PATHOGENS/NON-PATHOGENS SHARE. LEFT: HOMOLOGOUS GENES DISTRIBUTION. RIGHT: ISLANDS DISTRIBUTION.

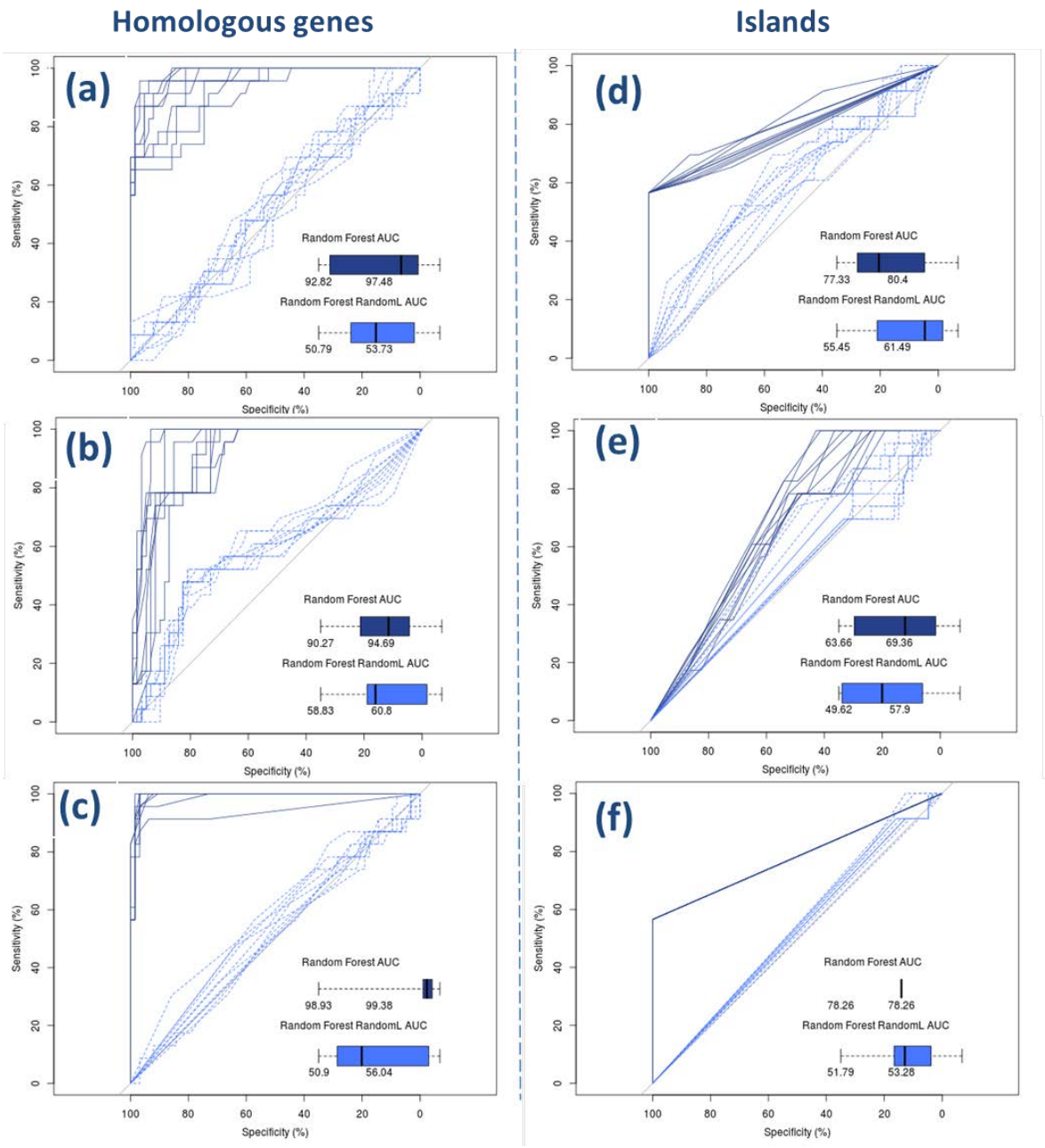

FIGURE 34 – CLASSIFICATION PERFORMANCE, AEROBE VS. ANAEROBE. FOR EACH ROC PLOT, THE REAL LABEL CLASSIFIER CURVES ARE PRESENTED IN DARK-BLUE SOLID LINES, WHILE THE RANDOM LABEL CLASSIFIER ARE IN LIGHT-BLUE DASHED LINES (THE ONES CLOSE TO THE BASELINE). THE VARIATION OF THE AUCS (AREA UNDER CURVE) IN THE CROSS-VALIDATION WAS INCLUDED IN THE FIGURE AS A BOX-PLOT (BOTTOM RIGHT). THE NUMBERS BELOW EACH BOX-PLOT ARE THE LOWER AND UPPER QUARTILES. HOMOLOGOUS GENES: A) FULL DATA-SET ($\overline{AUC}$ = 95.15%, $\overline{AUC}_{\text{RL}}$ = 52.26%); B) BIAS AEROBE ($\overline{AUC}$ = 92.48%, $\overline{AUC}_{\text{RL}}$ = 59.81%); C) BIAS ANAEROBE ($\overline{AUC}$ = 99.15%, $\overline{AUC}_{\text{RL}}$ = 53.47%). ISLANDS: D) FULL DATA-SET ($\overline{AUC}$ = 78.86%, $\overline{AUC}_{\text{RL}}$ = 58.47%); BIAS AEROBE ($\overline{AUC}$ = 66.51%, $\overline{AUC}_{\text{RL}}$ = 53.76%); BIAS ANAEROBE ($\overline{AUC}$ = 78.26%, $\overline{AUC}_{\text{RL}}$ = 52.53%).

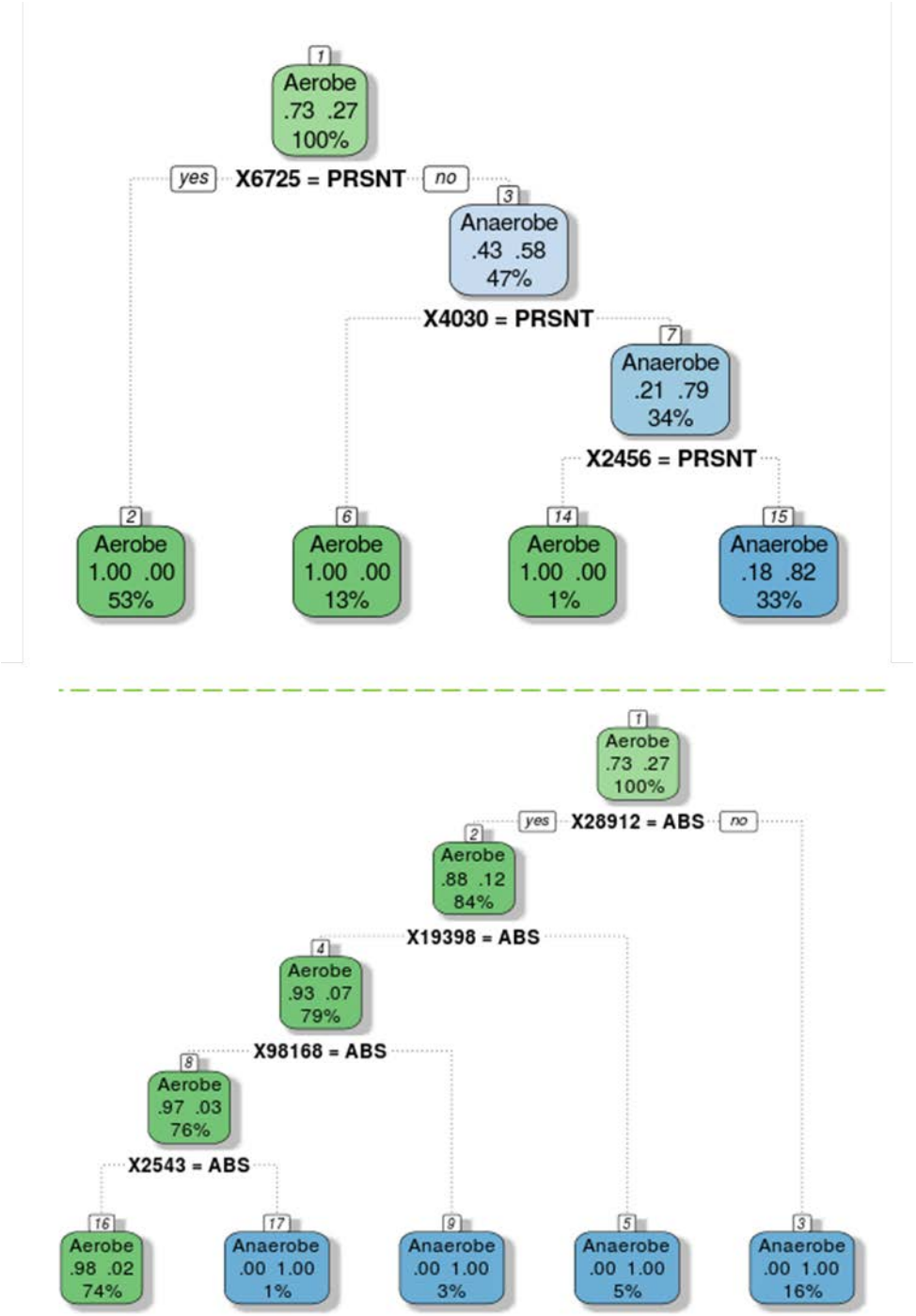

FIGURE 35 – DECISION TREES FOR HOMOLOGOUS GENES (AEROBES VS. ANAEROBES). DECISION TREES CREATED USING THE MOST DISCRIMINATIVE FEATURES FOR BOTH BIASES. ABS STANDS FOR “ABSENT” AND PRSNT FOR “PRESENT”. TOP: DECISION TREE FOR AEROBES (ACCURACY: 94.0%, PRECISION: 91.7%). BOTTOM: DECISION TREE FOR ANEROBES (ACCURACY: 98.0%, PRECISION: 92.6%).

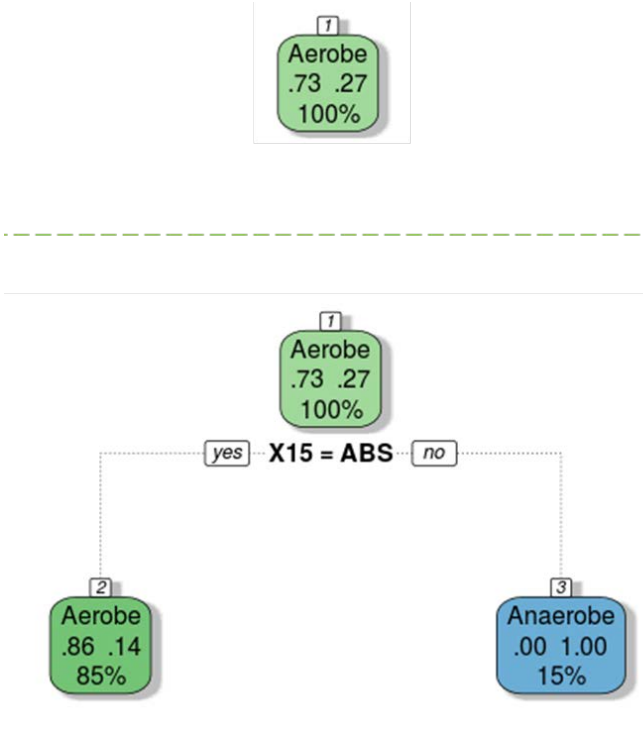

FIGURE 36 – DECISION TREES FOR HOMOLOGOUS GENES (AEROBES VS. ANAEROBES). DECISION TREES CREATED USING THE MOST DISCRIMINATIVE FEATURES FOR BOTH BIASES. ABS STANDS FOR "ABSENT" AND PRSNT FOR "PRESENT". TOP: NO MEANINFUL DECISION TREE WAS GENERATED FOR AEROBES. BOTTOM: DECISION TREE FOR ANEROBES (ACCURACY: 88.0%, PRECISION: 55.5%).

4.2.3.2 AEROBE VS. FACULTATIVE

In our data-set, we have 63 aerobes and only nine facultative organisms. Given the highly unbalanced number, we opt not to compare these two groups. Even if we were able to retrieve distinctive genetic features from these lifestyles, they simply might not be meaningful.

4.2.3.3 ANAEROBE VS. FACULTATIVE

We found 46,635 distinct homologous genes, with 28,368 mainly present in anaerobes and 18,267 in facultatives. The situation was the same for islands, where most of the 142 islands are mainly present in anaerobes, 109; and the remaining 33 in facultatives. The distribution of homologous sequences and islands can be observed in Figure 37.

Again, the classification results differed for homologous genes and islands (Figure 38). The classifiers had good performance for both anaerobes bias ($\overline{AUC}$ = 92.6%, Figure 38B) and facultatives bias ($\overline{AUC}$ = 96.2%, Figure 38C), as well as for the classifier using the full data-set ($\overline{AUC}$ = 96.5%, Figure 38A). These results

indicate that we find both, gene sets specific for anaerobes, as well as gene sets specific for facultatives, with almost identical and high accuracy. The scenario differed for the analysis using islands. The anaerobe bias performed is comparable to the one using homologous genes ($\overline{AUC}$ = 96.8%, Figure 38E), while the anaerobe bias worse ($\overline{AUC}$ =72.6 %, Figure 38E). Additionally, the classification using the full data-set has an odd performance, with big AUCs for both real labels ($\overline{AUC}$ = 96.2%) and random labels ($\overline{AUC}_{\mathrm{RL}}$ = 83.0%). This result reduces the confidence that the selected features are indeed meaningful to discriminate the two lifestyles.

The most discriminative homologous genes were used to create the decision trees in Figure 39. For the bias towards anaerobes, the selected clusters were: 1449 (17 genes, associated with ABC_tran and ABC_tran_Xtn domains) and 45 (15 genes, associated with Ribosomal_L33 family). The Pfam results can be observed in S. Table 8. For the bias towards facultatives, the selected cluster identifier was 6075 (10 genes, associated with Gp_dh_N and Gp_dh_C domains). The Pfam results can be observed in S. Table 9.

The most discriminative islands were used to create the decision trees in Figure 40. It was not possible to create any meaningful decision tree for the anaerobe bias. For the bias towards facultatives, the selected island identifiers were: 3900 (length 8, present in 7 organisms) and 3460 (length 17, present in 3 organisms). Table 6 presents the numbers of genes associated with metabolic, resistance and virulence that were present in the island.

TABLE 6 – DESCRIPTION OF THE MOST DISCRIMINATIVE ISLANDS FOR FACULTATIVES. “LENGTH” REPRESENTS THE AMOUNT OF CONSECUTIVE GROUPS OF HOMOLOGOUS SEQUENCES; WHILE THE “TOTAL GENES” REPRESENT THE SUM OF ALL GENES IN THE GROUPS OF HOMOLOGOUS SEQUENCES.

Island	Length	Homologous sequences	Metabolic	Resistance	Virulence
3900	8	99	0	0	22
3460	17	501	3	0	0

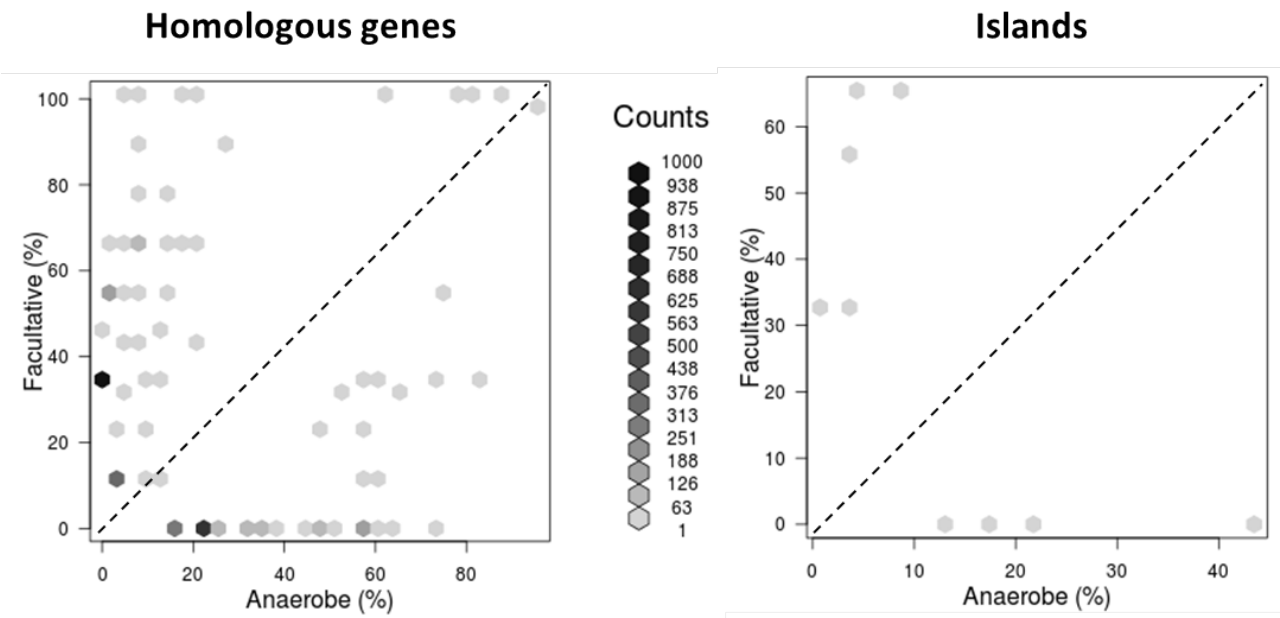

FIGURE 37 – DISTRIBUTION OF GENETIC FEATURES OVER TWO LIFESTYLES (ANAEROBES VS. FACULTATIVES). BOTH AXES IN THE PLOT DESCRIBE THE PERCENTAGE OF SPECIES IN THE RESPECTIVE CLASS(ES), HERE ANAEROBES (AN) AND FACULTATIVES (FA). COLOR-CODING OF THE HEAT MAP DEPICTS THE NUMBER OF CLUSTERS OF HOMOLOGOUS GENES THAT CERTAIN PERCENTAGES OF PATHOGENS/NON-PATHOGENS SHARE. LEFT: HOMOLOGOUS GENES DISTRIBUTION. RIGHT: ISLANDS DISTRIBUTION.

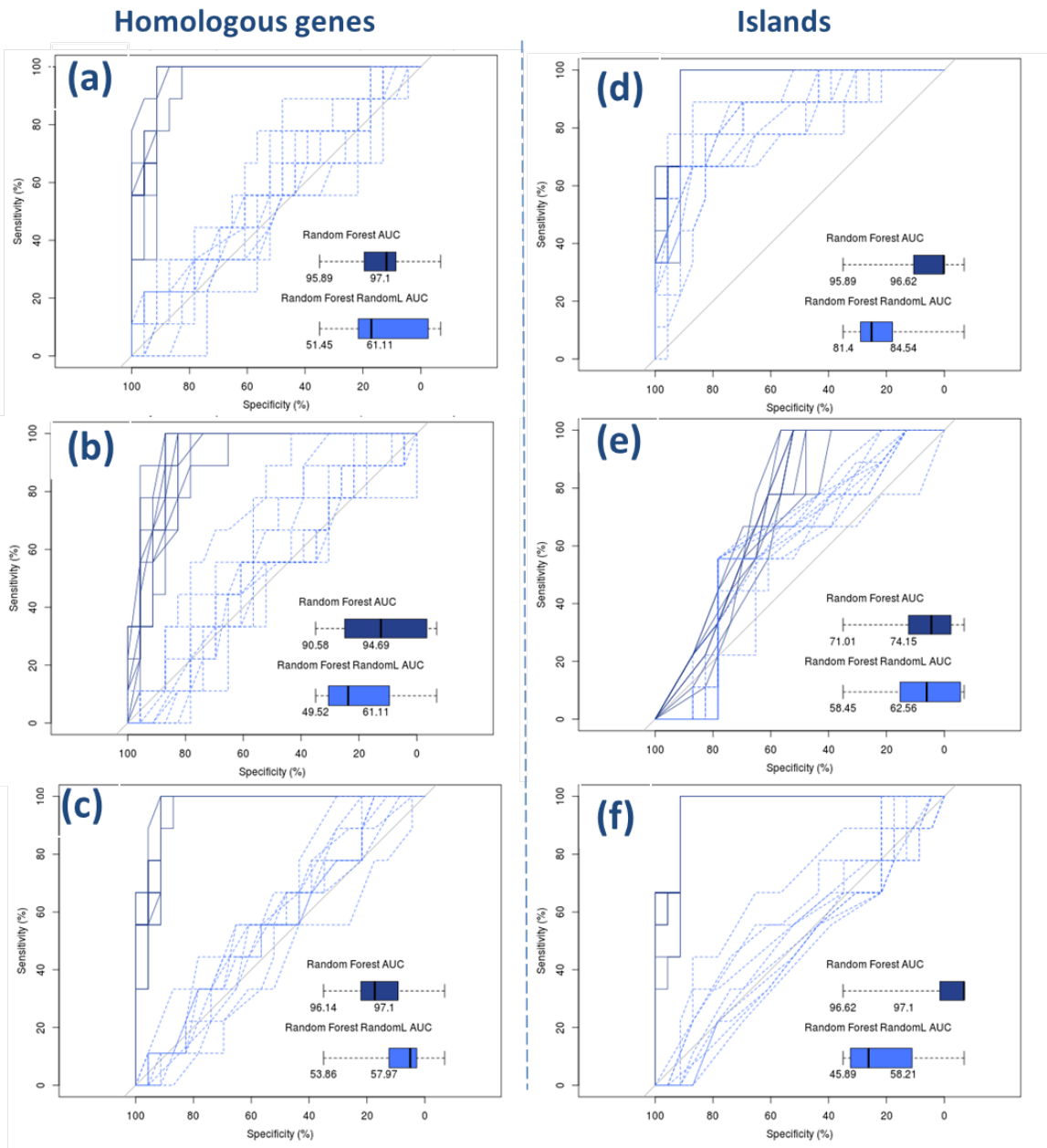

FIGURE 38 – CLASSIFICATION PERFORMANCE ANAEROBE VS. FACULTATIVE. FOR EACH ROC PLOT, THE REAL LABEL CLASSIFIER CURVES ARE PRESENTED IN DARK-BLUE SOLID LINES, WHILE THE RANDOM LABEL CLASSIFIERS ARE PRESENTED AS LIGHT-BLUE DASHED LINES (THE ONES CLOSE TO THE BASELINE). THE VARIATION OF THE AUCS (AREA UNDER CURVE) IN THE CROSS-VALIDATION WAS INCLUDED IN THE FIGURE AS A BOX-PLOT (BOTTOM RIGHT). THE NUMBERS BELOW EACH BOX-PLOT ARE THE LOWER AND UPPER QUARTILES. HOMOLOGOUS GENES: A) FULL DATA-SET ($\overline{AUC}$ = 96.5%, $\overline{AUC}_{\text{RL}}$ = 56.3%); B) BIAS ANAEROBE ($\overline{AUC}$ = 92.6%, $\overline{AUC}_{\text{RL}}$ = 55.3%); C) BIAS FACULTATIVE ($\overline{AUC}$ = 96.2%, $\overline{AUC}_{\text{RL}}$ = 55.9%). ISLANDS: D) FULL DATA-SET ($\overline{AUC}$ = 96.2%, $\overline{AUC}_{\text{RL}}$ = 83.0%); BIAS ANAEROBE ($\overline{AUC}$ = 72.6%, $\overline{AUC}_{\text{RL}}$ = 60.5%); BIAS FACULTATIVE ($\overline{AUC}$ = 96.8%, $\overline{AUC}_{\text{RL}}$ = 52.0%).

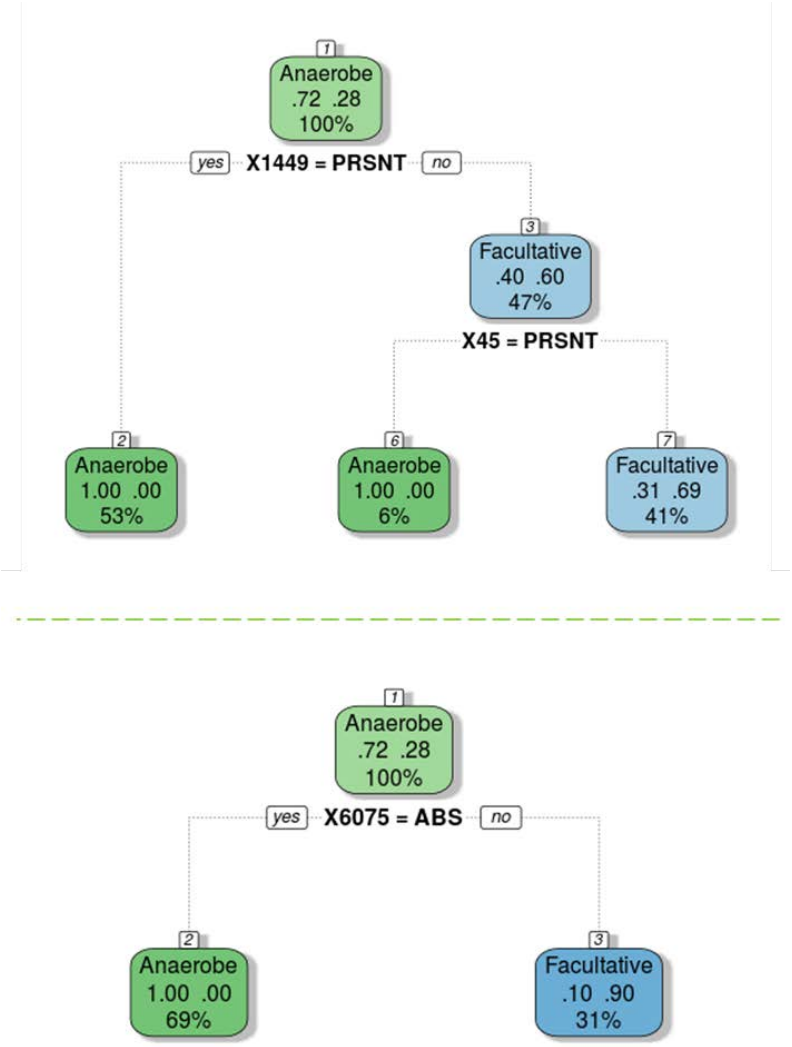

FIGURE 39 – DECISION TREES FOR HOMOLOGOUS GENES (ANAEROBES VS. FACULTATIVES). DECISION TREES CREATED USING THE MOST DISCRIMINATIVE FEATURES FOR BOTH BIASES. ABS STANDS FOR “ABSENT” AND PRSNT FOR “PRESENT”. TOP: DECISION TREE FOR ANEROBES (ACCURACY: 87.29%, PRECISION: 81.94%). BOTTOM: DECISION TREE FOR FACULTATIVES (ACCURACY: 96.9%, PRECISION: 87.09%).

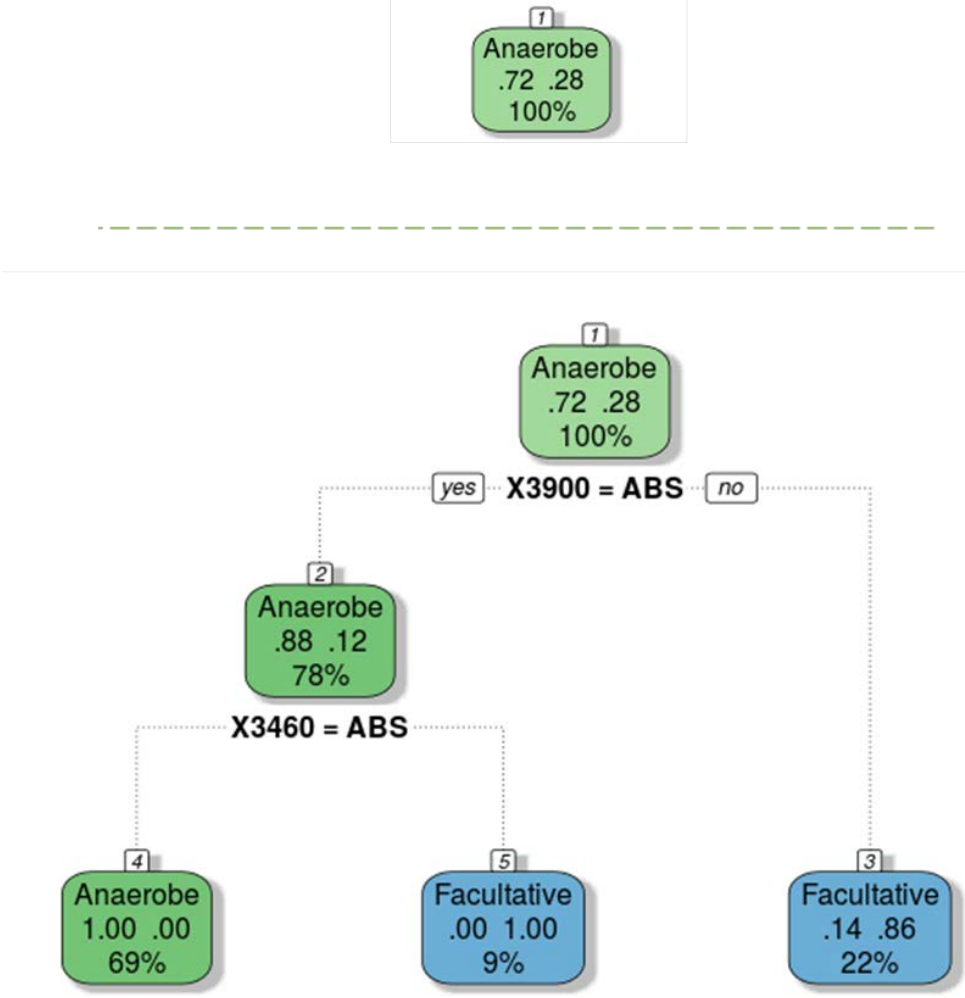

FIGURE 40 – DECISION TREES FOR HOMOLOGOUS GENES (ANAEROBES VS. FACULTATIVES). DECISION TREES CREATED USING THE MOST DISCRIMINATIVE FEATURES FOR BOTH BIASES. ABS STANDS FOR "ABSENT" AND PRSNT FOR "PRESENT". TOP: NO MEANINFUL DECISION TREE WAS GENERATED FOR ANAEROBES. BOTTOM: DECISION TREE FOR FACULTATIVES (ACCURACY: 88.92%, PRECISION: 90.32%).

4.2.4 HABITAT

Our last analysis involved organisms from different habitats, namely soil and aquatic organisms. This data-set is challenging for a couple of reasons. First, it is highly unbalanced; we had 34 soil organisms and only nine aquatics. Second, we already know from the literature that those are lifestyles that are hard to define due to their complexity, and the apparent gradient that exists from soil up to marine environments [97]. We mainly included this analysis for completion and to evaluate the limits of our predictions on a difficult data-set.

4.2.4.1 SOIL VS. AQUATIC

We found 134553 distinct homologous genes, where 107338 mainly present in soil and 27215 in aquatic organisms. Conversely,most of the 90 islands are mainly present in aquatics (55); the remaining 35 islands are found in soil. Further, no island is present in more than 30% of the organisms. The distribution of homologous sequences and islands can be observed in Figure 41.

Differently from previous cases, the classification results are similarly bad for both homologous genes and islands (Figure 42). In all comparisons the classifiers using real labels had similar performance to the ones using random labels. The classifiers had low performance for both aquatic bias ($\overline{AUC}$ = 61.2%, Figure 42B) and soil bias ($\overline{AUC}$ = 80.4% and $\overline{AUC}_{\text{RL}}$ = 60.0% Figure 38C), as well as for the classifier using the full data-set ($\overline{AUC}$ = 50.4%, Figure 42A). These results indicate that we are unlikely to find gene sets specific for aquatic and soil organisms. The scenario is similar for the analysis using islands, the aquatic bias present $\overline{AUC}$ of 54.2% (Figure 42E), and the soil bias of 58.7% (Figure 42E).

The most discriminative homologous genes were used to create the decision trees in Figure 43. For the bias towards aquatic, the selected cluster was 67953 (3 genes, not associated with any domain). For the bias towards soil, the selected cluster identifiers were: 3326 (22 genes, associated with HTH_26 domain), 6779 (42 genes, associated with ABC_tran domain), 19513 (10 genes, associated with Hexapep_2 and Hexapep repeats), 10196 (16 genes, associated with different Acyl-CoA_dh domains), and 20354 (18 genes, associated with Arabinose_Isome family and Arabinose_Iso_C domain). The Pfam results can be observed in S. Table 10. On the other hand, it was not possible to create any meaningful decision tree using the most discriminative islands for neither of the biases.

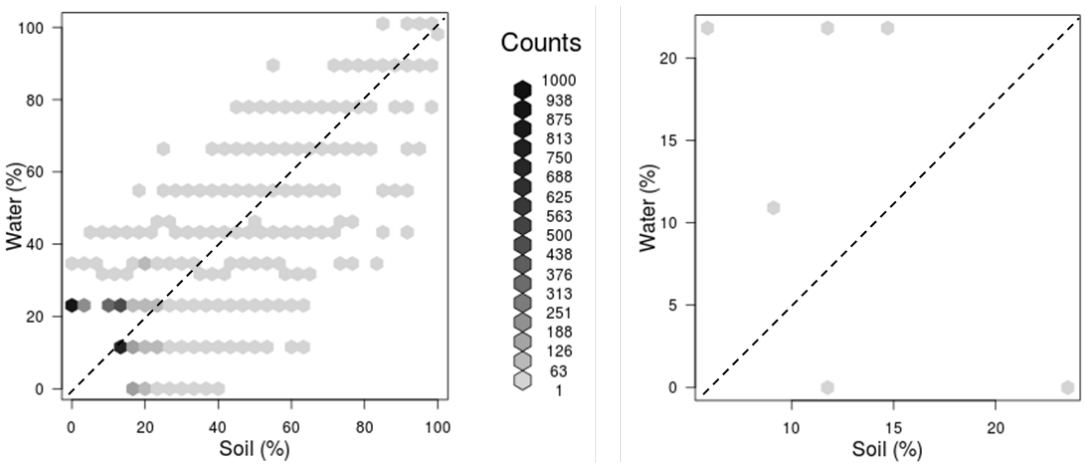

FIGURE 41 – DISTRIBUTION OF GENETIC FEATURES OVER TWO LIFESTYLES (SOIL VS. WATER/AQUATIC). BOTH AXES IN THE PLOT DESCRIBE THE PERCENTAGE OF SPECIES IN THE RESPECTIVE CLASS(ES), HERE SOIL (AN) AND WATER/AQUATIC (AQ). COLOR-CODING OF THE HEAT MAP DEPICTS THE NUMBER OF CLUSTERS OF HOMOLOGOUS GENES SHARED BY CERTAIN PERCENTAGES OF PATHOGENS/NON-PATHOGENS. LEFT: HOMOLOGOUS GENE DISTRIBUTION. RIGHT: ISLAND DISTRIBUTION.

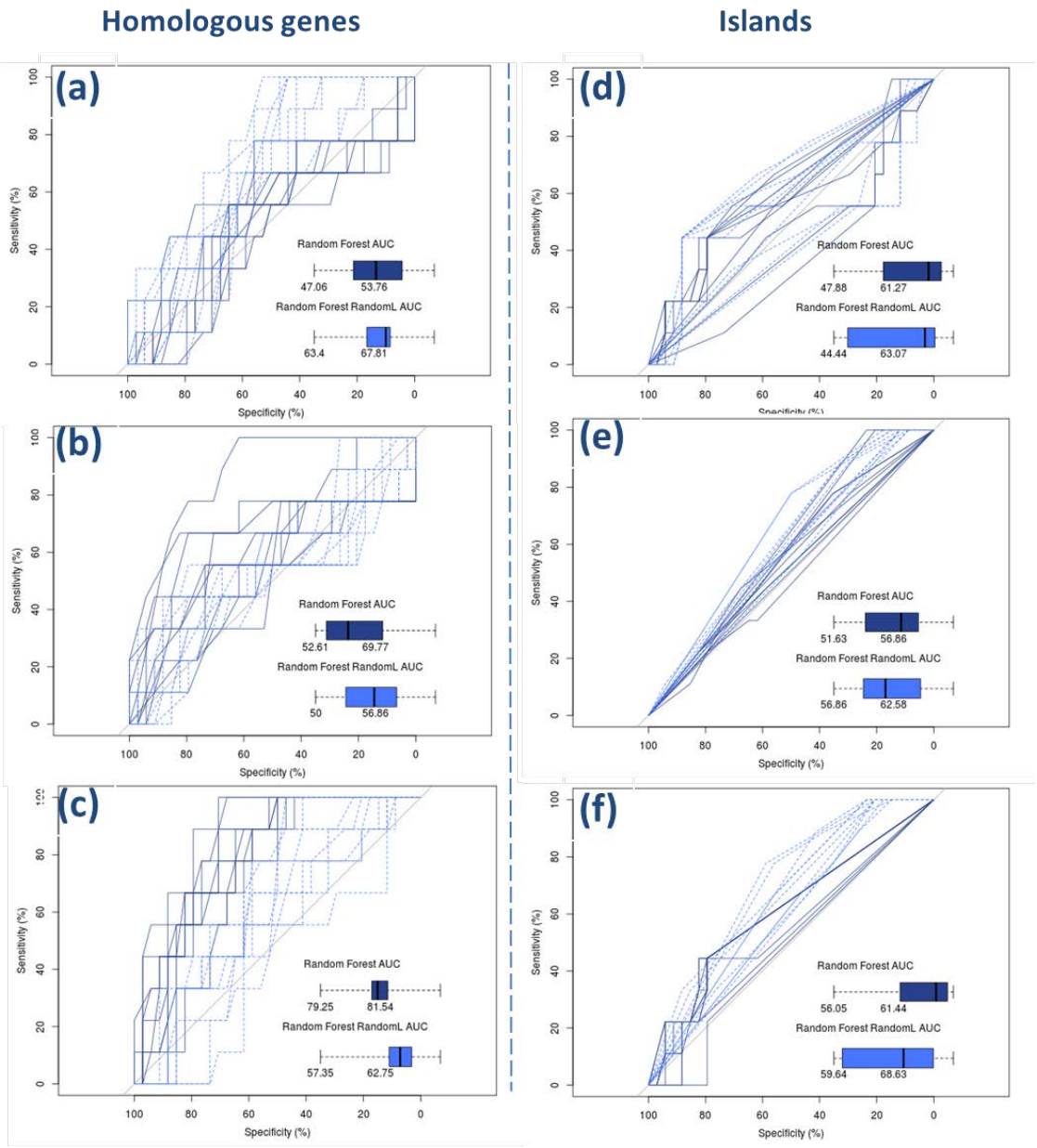

FIGURE 42 – CLASSIFICATION PERFORMANCE SOIL VS. AQUATIC. FOR EACH ROC PLOT, THE REAL LABEL CLASSIFIER CURVES ARE PRESENTED IN DARK-BLUE SOLID LINES, WHILE THE RANDOM LABEL CLASSIFIER ARE IN LIGHT-BLUE DASHED LINES (THE ONES CLOSE TO THE BASELINE). THE VARIATION OF THE AUCS (AREA UNDER CURVE) IN THE CROSS-VALIDATION WAS INCLUDED IN THE FIGURE AS A BOX-PLOT (BOTTOM RIGHT). THE NUMBERS BELOW EACH BOX-PLOT ARE THE LOWER AND UPPER QUARTILES. HOMOLOGOUS GENES: A) FULL DATA-SET ($\overline{AUC}$ = 50.4%, $\overline{AUC}_{\text{RL}}$ = 65.6%); B) BIAS AQUATIC ($\overline{AUC}$ = 61.2%, $\overline{AUC}_{\text{RL}}$ = 50.4%); C) BIAS SOIL ($\overline{AUC}$ = 80.4%, $\overline{AUC}_{\text{RL}}$ = 60.0%). ISLANDS: D) FULL DATA-SET ($\overline{AUC}$ = 54.5%, $\overline{AUC}_{\text{RL}}$ = 53.7%); BIAS AQUATIC ($\overline{AUC}$ = 54.2%, $\overline{AUC}_{\text{RL}}$ = 59.72%); BIAS SOIL ($\overline{AUC}$ = 58.7%, $\overline{AUC}_{\text{RL}}$ = 64.1%).

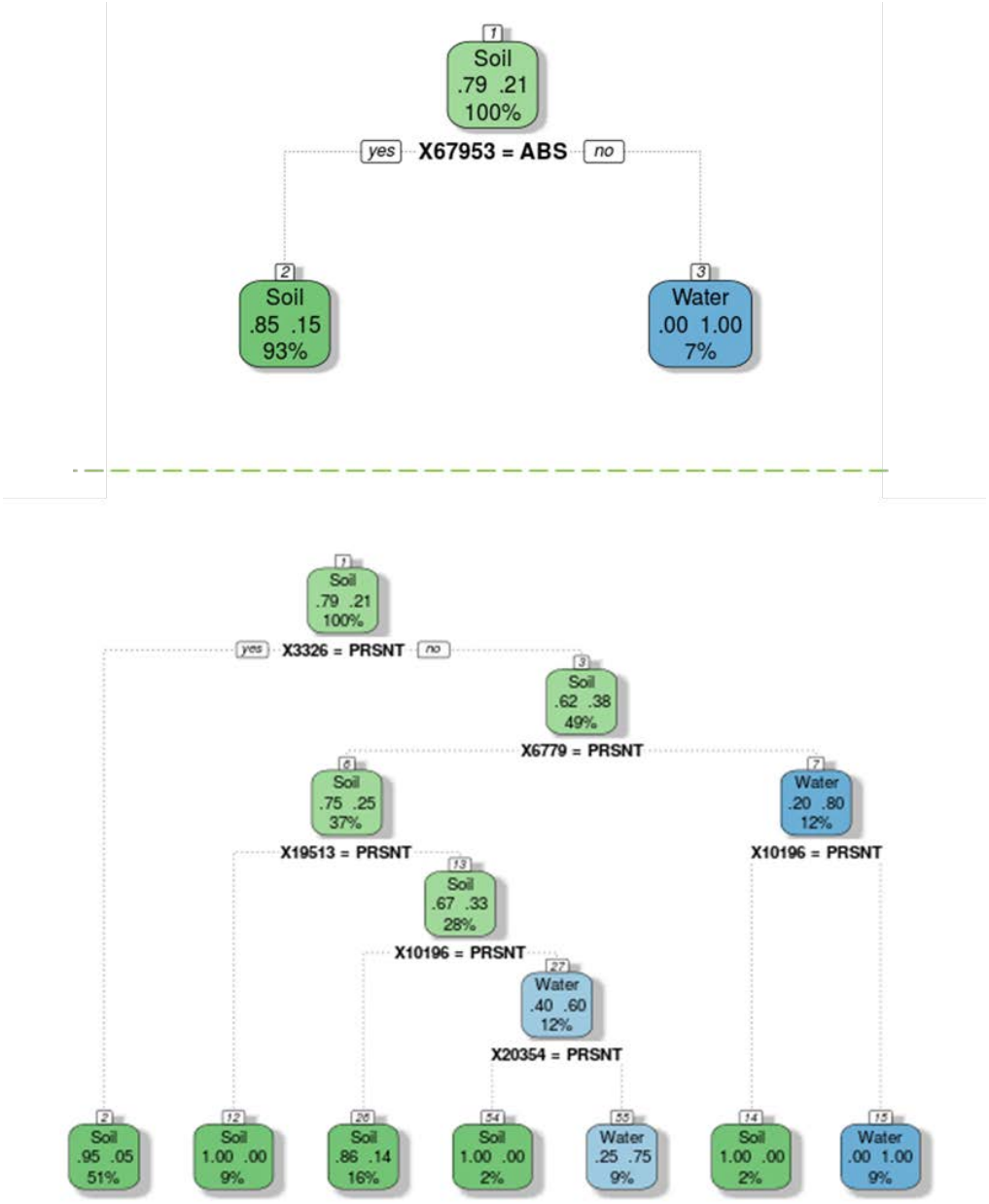

FIGURE 43 – DECISION TREES FOR HOMOLOGOUS GENES (SOIL VS. AQUATIC). DECISION TREES CREATED USING THE MOST DISCRIMINATIVE FEATURES FOR BOTH BIASES. ABS STANDS FOR “ABSENT” AND PRSNT FOR “PRESENT”. TOP: DECISION TREE FOR AQUATIC (ACCURACY: 86.0%, PRECISION: 33.3%). BOTTOM: DECISION TREE FOR SOIL (ACCURACY: 93.5%, PRECISION: 98.7%).

4.3 SECTION CONCLUSION

The aim of this section was to introduce LiSSI, a bioinformatics pipeline that can be used to identify signature genes or islands (conserved consecutive homology sequences) that distinguish bacterial lifestyles. To illustrate the tool's main features, we used different lifestyles found in Actinobacteria, namely: pathogenicity, tolerance for atmospheric oxygen, and habitat. In most cases, we were able to find signature genes and islands for these lifestyles. Nevertheless, we found that islands seem to carry less weight in the classification performance. It seems that gene order is poorly conserved among bacterial species, which might make individual genes more useful as classifiers.

5 GENERAL CONCLUSION

5 GENERAL CONCLUSION

The quote “we are drowning in information but starved for knowledge” from American author John Naisbitt seems to encapsulate the current state of genomic research. It is a well-known fact that the quantity of genomes publicly available exploded in recent decades, posing the challenge: how can we extract the most out of this treasure of data? As demonstrated throughout this thesis, this challenge is far from trivial, even for supposedly simple questions. During my PhD, I applied several approaches to identify a genetic feature (homologous sequences or islands) that could help elucidate the differences in bacterial lifestyles. My efforts revealed limitations in the availability of data, as it is not possible to find relevant lifestyle information for most of the available genomes. Further, the sequencing process outpaces the characterization process for most groups of bacteria, especially if there is no medical-veterinary interest. Also, it remains a challenge to establish which genetic features might indeed be associated with the lifestyles or simply due to phylogenetic proximity, because shared evolutionary history and shared functional traits are not necessarily independent [163].

Throughout this project, I developed methods and tools to explore the limits of computational functional genomics for bacterial lifestyle prediction. We started by developing a simple and straightforward approach to identify homologous genes that could help distinguish organisms from different pathogenicity classes. That approach was then extended to include the selection of distinguishable genomic islands, culminating in use of the LiSSI (LifeStyle-Specific-Island) tool. In both cases, the difficulty of identifying genetic features that can help explain bacterial lifestyles was clearly identified. These results point to the necessity for the further development of tools and methodologies to help expand our knowledge of bacterial genomes.

.

6 OUTLOOK

6 OUTLOOK

The main goal of my project was to develop methods and tools to help researchers analyse bacterial genomic sequences. The hope is that by providing the means to the data experts, we can help increase the amount of information that can be extracted from the already available genomic data. Thus, our goal stumbles upon the tool’s usability and friendliness. Currently, LiSSI depends on several R packages and third-party software; installation can represent a substantial impediment for the average final user. Plus, modifications in one or more of the R dependencies could disrupt the tool's function. Therefore, one of our future tasks would be to convert LiSSI into a self-contained tool. To achieve that goal, we are considering two approaches, either releasing the tool as a Virtual Machine or as a Docker. Both structures would automate application installation by providing a self-contained structure.

Further, our analyses are restricted to two lifestyles out of time. Ideally, we could include multiple lifestyles comparisons. We also plan to expand our functional classification analysis, which has thus far been restricted to conserved domain (Pfam) and similarity (NCBI BLAST) searches. It would help the interpretability of our results if we could easily summarize the gene characteristics found in both homologous sequences and islands. Also, regarding the predicted islands, it would be interesting to integrate knowledge about their origin, for instance, if they are part of an operon or, rather,laterally transferred.

7 REFERENCES

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

Archer, K. J. and R. V. Kimes (2008). "Empirical characterization of random forest variable importance measures." Computational Statistics & Data Analysis **52**(4): 2249-2260.

Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke and O. Zagnitko (2008). "The RAST Server: rapid annotations using subsystems technology." BMC Genomics **9**: 75.

Barberán, A., K. S. Ramirez, J. W. Leff, M. A. Bradford, D. H. Wall and N. Fierer (2014). "Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria." Ecology letters **17**(7): 794-802.

Barcellos, F. G., P. Menna, J. S. da Silva Batista and M. Hungria (2007). "Evidence of horizontal transfer of symbiotic genes from a Bradyrhizobium japonicum inoculant strain to indigenous diazotrophs Sinorhizobium (Ensifer) fredii and Bradyrhizobium elkanii in a Brazilian Savannah soil." Applied and environmental microbiology **73**(8): 2635-2643.

Bashir, A., A. A. Klammer, W. P. Robins, C.-S. Chin, D. Webster, E. Paxinos, D. Hsu, M. Ashby, S. Wang, P. Peluso, R. Sebra, J. Sorenson, J. Bullard, J. Yen, M. Valdovino, E. Mollova, K. Luong, S. Lin, B. LaMay, A. Joshi, L. Rowe, M. Frace, C. L. Tarr, M. Turnsek, B. M. Davis, A. Kasarskis, J. J. Mekalanos, M. K. Waldor and E. E. Schadt (2012). "A hybrid approach for the automated finishing of bacterial genomes." Nat Biotechnol **30**(7): 701-707.

Baumbach, J. and L. Apeltsin (2008). "Linking Cytoscape and the corynebacterial reference database CoryneRegNet." BMC Genomics **9**: 184.

Berg, G., L. Eberl and A. Hartmann (2005). "The rhizosphere as a reservoir for opportunistic human pathogenic bacteria." Environ Microbiol **7**(11): 1673-1685.

Breiman, L. (2001). "Random Forests." Machine Learning **45**(1): 5-32.

Breiman, L., J. Friedman, C. J. Stone and R. A. Olshen (1984). Classification and regression trees, CRC press.

Bush, K., P. Courvalin, G. Dantas, J. Davies, B. Eisenstein, P. Huovinen, G. A. Jacoby, R. Kishony, B. N. Kreiswirth, E. Kutter, S. A. Lerner, S. Levy, K. Lewis, O. Lomovskaya, J. H. Miller, S. Mobashery, L. J. V. Piddock, S. Projan, C. M. Thomas, A. Tomasz, P. M. Tulkens, T. R. Walsh, J. D. Watson, J. Witkowski, W. Witte, G. Wright, P. Yeh and H. I. Zgurskaya (2011). "Tackling antibiotic resistance." Nat Rev Microbiol **9**(12): 894-896.

Böcker, S., K. Jahn, J. Mixtacki and J. Stoye (2009). "Computation of median gene clusters." Journal of Computational Biology **16**(8): 1085-1099.

Cabiscol, E., J. Tamarit and J. Ros (1999). "Oxidative stress in bacteria and protein damage by reactive oxygen species." International Microbiology **3**(1): 3-8.

Casadevall, A. (2006). "Cards of virulence and the global virulome for humans." MICROBE-AMERICAN SOCIETY FOR MICROBIOLOGY **1**(8): 359.

Casadevall, A. and L.-a. Pirofski (2007). "Accidental virulence, cryptic pathogenesis, martians, lost hosts, and the pathogenicity of environmental microbes." Eukaryot Cell **6**(12): 2169-2174.

Celik, Y., R. Drori, N. Pertaya-Braun, A. Altan, T. Barton, M. Bar-Dolev, A. Groisman, P. L. Davies and I. Braslavsky (2013). "Microfluidic experiments reveal that antifreeze proteins bound to ice crystals suffice to prevent their growth." Proceedings of the National Academy of Sciences **110**(4): 1309-1314.

Chain, P. S. G., D. V. Grafham, R. S. Fulton, M. G. Fitzgerald, J. Hostetler, D. Muzny, J. Ali, B. Birren, D. C. Bruce, C. Buhay, J. R. Cole, Y. Ding, S. Dugan, D. Field, G. M. Garrity, R. Gibbs, T. Graves, C. S. Han, S. H. Harrison, S. Highlander, P. Hugenholtz, H. M. Khouri, C. D. Kodira, E. Kolker, N. C. Kyrpides, D. Lang, A. Lapidus, S. A. Malfatti, V. Markowitz, T. Metha, K. E. Nelson, J. Parkhill, S. Pitluck, X. Qin, T. D. Read, J. Schmutz, S. Sozhamannan, P. Sterk, R. L. Strausberg, G. Sutton, N. R. Thomson, J. M. Tiedje, G. Weinstock, A. Wollam and J. C. Detter (2009). "Genomics. Genome project standards in a new era of sequencing." Science (80-) **326**(5950): 236-237.
Coordinators, N. R. (2014). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Research **42**(D1): D7-D17.
de Boer, H. A., S. F. Gilbert and M. Nomura (1979). DNA sequences of promoter regions for rRNA operons rrnE and rrnA in E. coli, Cell Press. **17:** 201-209.
De Maayer, P., D. Anderson, C. Cary and D. A. Cowan (2014). "Some like it cold: understanding the survival strategies of psychrophiles." EMBO reports: e201338170.
Diaz-Uriarte, R. (2007). "GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest." BMC bioinformatics **8**(1): 328.
Diaz-Uriarte, R. (2009). "varSelRF: Variable selection using random forests." URL http://ligarto.org/rdiaz/Software/Software.html, R package version 0.7-1.
Díaz-Uriarte, R. and S. A. De Andres (2006). "Gene selection and classification of microarray data using random forest." BMC bioinformatics **7**(1): 1.
Diguistini, S., N. Y. Liao, D. Platt, G. Robertson, M. Seidel, S. K. Chan, T. R. Docking, I. Birol, R. A. Holt, M. Hirst, E. Mardis, M. A. Marra, R. C. Hamelin, J. Bohlmann, C. Breuil and S. J. Jones (2009). "De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data." Genome Biol **10**(9): R94.
Dobrindt, U., B. Janke, K. Piechaczek, G. Nagy, W. Ziebuhr, G. Fischer, A. Schierhorn, M. Hecker, G. Blum-Oehler and J. Hacker (2000). "Toxin genes on pathogenicity islands: impact for microbial evolution." International journal of medical microbiology **290**(4): 307-311.
Donati, C. and R. Rappuoli (2013). "Reverse vaccinology in the 21st century: improvements over the original design." Ann N Y Acad Sci **1285**: 115-132.
Dorman, C. J. (2013). "Genome architecture and global gene regulation in bacteria: making progress towards a unified model?" Nature Reviews Microbiology **11**(5): 349-355.
Edgar, R. C. (2004). "Local homology recognition and distance measures in linear time using compressed amino acid alphabets." Nucleic Acids Research **32**(1): 380-385.
Edgar, R. C. (2010). "Search and clustering orders of magnitude faster than BLAST." Bioinformatics **26**(19): 2460-2461.
Efron, B. and R. Tibshirani (1997). "Improvements on cross-validation: the 632+ bootstrap method." Journal of the American Statistical Association **92**(438): 548-560.
Eiglmeier, K., J. Parkhill, N. Honore, T. Garnier, F. Tekaia, A. Telenti, P. Klatser, K. D. James, N. R. Thomson and P. R. Wheeler (2001). "The decaying genome of Mycobacterium leprae." Leprosy review **72**(4): 387-398.
Enright, A. J. and C. A. Ouzounis (2000). "GeneRAGE: a robust algorithm for sequence clustering and domain detection." Bioinformatics **16**(5): 451-457.
Enright, A. J., S. Van Dongen and C. A. Ouzounis (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic acids research **30**(7): 1575-1584.
Esnault, E., M. Valens, O. Espéli and F. Boccard (2007). "Chromosome structuring limits genome plasticity in Escherichia coli." PLoS Genet **3**(12): e226.
Fang, G., E. P. Rocha and A. Danchin (2008). "Persistence drives gene clustering in bacterial genomes." BMC genomics **9**(1): 4.

Fierer, N., M. A. Bradford and R. B. Jackson (2007). "Toward an ecological classification of soil bacteria." Ecology **88**(6): 1354-1364.
Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick and e. al (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science (80-) **269**(5223): 496-512.
Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, J. L. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J.-F. Tomb, B. A. Dougherty, K. F. Bott, P.-C. Hu, T. S. Lucier, S. N. Peterson, H. O. Smith, C. A. Hutchison and J. C. Venter (1995). "The Minimal Gene Complement of Mycoplasma genitalium." Science (80-) **270**(5235): 397-404.
Frey, B. J. and D. Dueck (2007). "Clustering by passing messages between data points." science **315**(5814): 972-976.
Fu, L., B. Niu, Z. Zhu, S. Wu and W. Li (2012). "CD-HIT: accelerated for clustering the next-generation sequencing data." Bioinformatics **28**(23): 3150-3152.
Gans, J., M. Wolinsky and J. Dunbar (2005). "Computational improvements reveal great bacterial diversity and high metal toxicity in soil." Science **309**(5739): 1387-1390.
Gerstein, M. and D. Zheng (2006). "The real life of pseudogenes." Scientific American **295**(2): 48-55.
Griffiths, A. J. (2005). An introduction to genetic analysis, Macmillan.
Gu, J. and V. J. Hilser (2009). "Sequence-based analysis of protein energy landscapes reveals nonuniform thermal adaptation within the proteome." Molecular biology and evolution **26**(10): 2217-2227.
Gutteridge, J. M. (1994). "Biological origin of free radicals, and mechanisms of antioxidant protection." Chemico-biological interactions **91**(2-3): 133-140.
Görke, B. and J. Stülke (2008). "Carbon catabolite repression in bacteria: many ways to make the most out of nutrients." Nature Reviews Microbiology **6**(8): 613-624.
Hapfelmeier, A. and K. Ulm (2013). "A new variable selection approach using random forests." Computational Statistics & Data Analysis **60**: 50-69.
Hastie, T., R. Tibshirani, J. Friedman and J. Franklin (2005). "The elements of statistical learning: data mining, inference and prediction." The Mathematical Intelligencer **27**(2): 83-85.
He, Y., Z. Xiang and H. L. T. Mobley (2010). "Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development." J Biomed Biotechnol **2010**: 297505.
Imlay, J. A. (2013). "The molecular mechanisms and physiological consequences of oxidative stress: lessons from a model bacterium." Nature Reviews Microbiology **11**(7): 443-454.
Jahn, K. (2011). "Efficient computation of approximate gene clusters based on reference occurrences." Journal of Computational Biology **18**(9): 1255-1274.
Jahn, K., S. Winter, J. Stoye and S. Böcker (2013). "Statistics for approximate gene clusters." BMC bioinformatics **14**(Suppl 15): S14.
Jakobsen, T. H., M. A. Hansen, P. Ø. Jensen, L. Hansen, L. Riber, A. Cockburn, M. Kolpen, C. Rønne Hansen, W. Ridderberg, S. Eickhardt, M. Hansen, P. Kerpedjiev, M. Alhede, K. Qvortrup, M. Burmølle, C. Moser, M. Kühl, O. Ciofu, M. Givskov, S. J. Sørensen, N. Høiby and T. Bjarnsholt (2013). "Complete genome sequence of the cystic fibrosis pathogen Achromobacter xylosoxidans NH44784-1996 complies with important pathogenic phenotypes." PLoS ONE **8**(7): e68484.
Kececioglu, J. and J. Ju Separating repeats in DNA sequence assembly. Proceedings of the fifth annual international conference on Computational biology, ACM.

Klassen, J. L. and C. R. Currie (2012). "Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation." BMC Genomics **13**: 14.
Koch, A. L. (2001). "Oligotrophs versus copiotrophs." Bioessays **23**(7): 657-661.
Krizova, L. and A. Nemec (2010). "A 63 kb genomic resistance island found in a multidrug-resistant Acinetobacter baumannii isolate of European clone I from 1977." Journal of antimicrobial chemotherapy: dkq223.
Kuo, C.-H. and H. Ochman (2010). "The extinction dynamics of bacterial pseudogenes." PLoS Genet **6**(8): e1001050.
Kursa, M. B. (2014). "Robustness of Random Forest-based gene selection methods." BMC bioinformatics **15**(1): 1.
Landau, G. M., L. Parida and O. Weimann (2005). "Gene proximity analysis across whole genomes via pq trees1." Journal of Computational Biology **12**(10): 1289-1306.
Langille, M. G., W. W. Hsiao and F. S. Brinkman (2010). "Detecting genomic islands using bioinformatics approaches." Nature Reviews Microbiology **8**(5): 373-382.
Langille, M. G. I., W. W. L. Hsiao and F. S. L. Brinkman (2008). "Evaluation of genomic island predictors using a comparative genomics approach." BMC Bioinformatics **9**: 329.
Lauro, F. M., D. McDougald, T. Thomas, T. J. Williams, S. Egan, S. Rice, M. Z. DeMaere, L. Ting, H. Ertan and J. Johnson (2009). "The genomic basis of trophic strategy in marine bacteria." Proceedings of the National Academy of Sciences **106**(37): 15527-15533.
Lawrence, J. (1999). "Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes." Current opinion in genetics & development **9**(6): 642-648.
Lawrence, J. G. and J. R. Roth (1996). "Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters." Genetics **143**(4): 1843-1860.
Lerat, E. and H. Ochman (2005). "Recognizing the pseudogenes in bacterial genomes." Nucleic Acids Research **33**(10): 3125-3132.
Li, W. and A. Godzik (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics **22**(13): 1658-1659.
Liaw, A. and M. Wiener (2002). "Classification and regression by randomForest." R news **2**(3): 18-22.
Liu, Y., P. M. Harrison, V. Kunin and M. Gerstein (2004). "Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes." Genome biology **5**(9): R64.
Livermore, J. A., S. J. Emrich, J. Tan and S. E. Jones (2014). "Freshwater bacterial lifestyles inferred from comparative genomics." Environmental microbiology **16**(3): 746-758.
Margesin, R. and G. Feller (2010). "Biotechnological applications of psychrophiles." Environmental technology **31**(8-9): 835-844.
Math, R. K., H. M. Jin, J. M. Kim, Y. Hahn, W. Park, E. L. Madsen and C. O. Jeon (2012). "Comparative genomics reveals adaptation by Alteromonas sp. SN2 to marine tidal-flat conditions: cold tolerance and aromatic hydrocarbon metabolism." PLoS One **7**(4): e35784.
Medini, D., C. Donati, H. Tettelin, V. Masignani and R. Rappuoli (2005). "The microbial pan-genome." Curr Opin Genet Dev **15**(6): 589-594.
Meyer, F., A. Goesmann, A. C. McHardy, D. Bartels, T. Bekel, J. Clausen, J. Kalinowski, B. Linke, O. Rupp and R. Giegerich (2003). "GenDB—an open source genome annotation system for prokaryote genomes." Nucleic Acids Research **31**(8): 2187-2195.
Miller, J. R., S. Koren and G. Sutton (2010). "Assembly algorithms for next-generation sequencing data." Genomics **95**(6): 315-327.
Mills, S. D. (2006). "When will the genomics investment pay off for antibacterial discovery?" Biochem Pharmacol **71**(7): 1096-1102.

Moran, N. A. (2002). "Microbial Minimalism: Genome Reduction in Bacterial Pathogens." Cell **108**(5): 583-586.
Morris, R. L. and T. M. Schmidt (2013). "Shallow breathing: bacterial life at low O2." Nature Reviews Microbiology **11**(3): 205-212.
Munroe, D. J. and T. J. R. Harris (2010). "Third-generation sequencing fireworks at Marco Island." Nat Biotechnol **28**(5): 426-428.
Müller-Herbst, S., S. Wüstner, A. Mühlig, D. Eder, T. M. Fuchs, C. Held, A. Ehrenreich and S. Scherer (2014). "Identification of genes essential for anaerobic growth of Listeria monocytogenes." Microbiology **160**(4): 752-765.
Nakamura, K. and M. Inouye (1979). DNA sequence of the gene for the outer membrane lipoprotein of E. coli: an extremely AT-rich promoter, Cell Press. **18:** 1109-1117.
Newton, I. L. G. and S. R. Bordenstein (2011). "Correlations between bacterial ecology and mobile DNA." Curr Microbiol **62**(1): 198-208.
Nilsson, R., J. M. Peña, J. Björkegren and J. Tegnér (2007). "Consistent feature selection for pattern recognition in polynomial time." The Journal of Machine Learning Research **8**: 589-612.
Ochman, H. and L. M. Davalos (2006). "The nature and dynamics of bacterial genomes." Science **311**(5768): 1730-1733.
Ochman, H., J. G. Lawrence and E. A. Groisman (2000). "Lateral gene transfer and the nature of bacterial innovation." Nature **405**(6784): 299-304.
Overduin, P., W. Boos and J. Tommassen (1988). "Nucleotide sequence of the ugp genes of Escherichia coli K-12: homology to the maltose system." Mol Microbiol **2**(6): 767-775.
Paccanaro, A., J. A. Casbon and M. A. Saqi (2006). "Spectral clustering of protein sequences." Nucleic acids research **34**(5): 1571-1580.
Pahl, H. L. and P. A. Baeuerle (1994). "Oxygen and the control of gene expression." Bioessays **16**(7): 497-502.
Pál, C. and L. D. Hurst (2004). "Evidence against the selfish operon theory." Trends in Genetics **20**(6): 232-234.
Pauling, J., R. Rottger, A. Tauch, V. Azevedo and J. Baumbach (2012). "CoryneRegNet 6.0--Updated database content, new analysis methods and novel features focusing on community demands." Nucleic Acids Res **40**(Database issue): D610-614.
Payne, D. J., M. N. Gwynn, D. J. Holmes and D. L. Pompliano (2007). "Drugs for bad bugs: confronting the challenges of antibacterial discovery." Nat Rev Drug Discov **6**(1): 29-40.
Phillippy, A. M., M. C. Schatz and M. Pop (2008). "Genome assembly forensics: finding the elusive mis-assembly." Genome Biol **9**(3): R55.
Poole, R. K. and G. M. Cook (2000). "Redundancy of aerobic respiratory chains in bacteria? Routes, reasons and regulation." Advances in microbial physiology **43**: 165-224.
Pop, M. (2009). "Genome assembly reborn: recent computational challenges." Brief Bioinform **10**(4): 354-366.
Popa, O. and T. Dagan (2011). "Trends and barriers to lateral gene transfer in prokaryotes." Curr Opin Microbiol **14**(5): 615-623.
Popa, O., E. Hazkani-Covo, G. Landan, W. Martin and T. Dagan (2011). "Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes." Genome Res **21**(4): 599-609.
Porter, A. G., C. Barber, N. H. Carey, R. A. Hallewell, G. Threlfall and J. S. Emtage (1979). "Complete nucleotide sequence of an influenza virus haemagglutinin gene from cloned DNA." Nature **282**(5738): 471-477.
Postle, K. and R. F. Good (1983). "DNA sequence of the Escherichia coli tonB gene." Proceedings of the National Academy of Sciences **80**(17): 5235-5239.

Prachi, P., C. Donati, F. Masciopinto, R. Rappuoli and F. Bagnoli (2013). "Deep sequencing in pre- and clinical vaccine research." Public Health Genomics **16**(1-2): 62-68.
Projan, S. J. (2003). "Why is big Pharma getting out of antibacterial drug discovery?" Curr Opin Microbiol **6**(5): 427-430.
Pucci, M. J. (2006). "Use of genomics to select antibacterial targets." Biochem Pharmacol **71**(7): 1066-1072.
Rahmann, S., T. Wittkop, J. Baumbach, M. Martin, A. Truss and S. Böcker (2007). Exact and heuristic algorithms for weighted cluster editing. Comput Syst Bioinformatics Conf, Citeseer.
Rappuoli, R. (2000). "Reverse vaccinology." Curr Opin Microbiol **3**(5): 445-450.
Ribeiro, F. J., D. Przybylski, S. Yin, T. Sharpe, S. Gnerre, A. Abouelleil, A. M. Berlin, A. Montmayeur, T. P. Shea, B. J. Walker, S. K. Young, C. Russ, C. Nusbaum, I. MacCallum and D. B. Jaffe (2012). "Finished bacterial genomes from shotgun sequence data." Genome Res **22**(11): 2270-2277.
Richardson, P. (2010). "Special Issue: Next Generation DNA Sequencing." Genes **1**(3): 385-387.
Ricker, N., H. Qian and R. Fulthorpe (2012). "The limitations of draft assemblies for understanding prokaryotic adaptation and evolution." Genomics.
Rocha, E. P. (2004). "Order and disorder in bacterial genomes." Current opinion in microbiology **7**(5): 519-527.
Rocha, E. P. (2008). "The organization of the bacterial genome." Annual review of genetics **42**: 211-233.
Rodrigues, D. F. and J. M. Tiedje (2008). "Coping with our cold planet." Applied and environmental microbiology **74**(6): 1677-1686.
Roesch, L. F., R. R. Fulthorpe, A. Riva, G. Casella, A. K. Hadwin, A. D. Kent, S. H. Daroub, F. A. Camargo, W. G. Farmerie and E. W. Triplett (2007). "Pyrosequencing enumerates and contrasts soil microbial diversity." The ISME journal **1**(4): 283-290.
Rogers, J. and S. Gunn (2006). Identifying feature relevance using a random forest. Subspace, Latent Structure and Feature Selection, Springer**:** 173-184.
Rogers, J. and S. Gunn (2006). Identifying Feature Relevance Using a Random Forest. Subspace, Latent Structure and Feature Selection. C. Saunders, M. Grobelnik, S. Gunn and J. Shawe-Taylor, Springer Berlin Heidelberg. **3940:** 173-184.
Rohmer, L., D. Hocquet and S. I. Miller (2011). "Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis." Trends Microbiol **19**(7): 341-348.
Sandri, M. and P. Zuccolotto (2012). "A bias correction algorithm for the Gini variable importance measure in classification trees." Journal of Computational and Graphical Statistics.
Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-5467.
Sant'Anna, F., A. Lebedinsky, T. Sokolova, F. Robb and J. Gonzalez (2015). "Analysis of three genomes within the thermophilic bacterial species Caldanaerobacter subterraneus with a focus on carbon monoxide dehydrogenase evolution and hydrolase diversity." BMC genomics **16**(1): 757.
Santos, A. R., V. B. Pereira, E. Barbosa, J. Baumbach, J. Pauling, R. R\"ottger, M. Z. Turk, A. Silva, A. Miyoshi and V. Azevedo (2013). "Mature Epitope Density-A strategy for target selection based on immunoinformatics and exported prokaryotic proteins." BMC Genomics **14**(Suppl 6): S4.
Seib, K. L., X. Zhao and R. Rappuoli (2012). "Developing vaccines in the era of genomics: a decade of reverse vaccinology." Clinical Microbiology and Infection **18**: 109-116.
Shendure, J., R. D. Mitra, C. Varma and G. M. Church (2004). "Advanced sequencing technologies: methods and goals." Nat Rev Genet **5**(5): 335-344.
Shendure, J., G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra and G. M. Church (2005). "Accurate

multiplex polony sequencing of an evolved bacterial genome." Science (80-) **309**(5741): 1728-1732.
Sies, H. and C. F. Menck (1992). "Singlet oxygen induced DNA damage." Mutation Research/DNAging **275**(3-6): 367-375.
Silver, L. L. (2011). "Challenges of antibacterial discovery." Clin Microbiol Rev **24**(1): 71-109.
Simeone, R., D. Bottai and R. Brosch (2009). "ESX/type VII secretion systems and their role in host-pathogen interaction." Curr Opin Microbiol **12**(1): 4-10.
Smith, D. R. (2013). "Death of the genome paper." Front Genet **4**: 72.
Soares, S. C., V. A. Abreu, R. T. Ramos, L. Cerdeira, A. Silva, J. Baumbach, E. Trost, A. Tauch, R. Hirata Jr and A. L. Mattos-Guaraldi (2012). "PIPS: pathogenicity island prediction software." PloS one **7**(2): e30848.
Soares, S. C., V. A. C. Abreu, R. T. J. Ramos, L. Cerdeira, A. Silva, J. Baumbach, E. Trost, A. Tauch, R. Hirata, Jr., A. L. Mattos-Guaraldi, A. Miyoshi and V. Azevedo (2012). "PIPS: pathogenicity island prediction software." PLoS ONE **7**(2): e30848.
Soares, S. C., H. Geyik, R. T. Ramos, P. H. de Sá, E. G. Barbosa, J. Baumbach, H. C. Figueiredo, A. Miyoshi, A. Tauch and A. Silva (2015). "GIPSy: Genomic island prediction software." Journal of biotechnology.
Soares, S. C., A. Silva, E. Trost, J. Blom, R. Ramos, A. Carneiro, A. Ali, A. R. Santos, A. C. Pinto, C. Diniz, E. G. V. Barbosa, F. A. Dorella, F. Aburjaile, F. S. Rocha, K. K. F. Nascimento, L. C. Guimarães, S. Almeida, S. S. Hassan, S. M. Bakhtiar, U. P. Pereira, V. A. C. Abreu, M. P. C. Schneider, A. Miyoshi, A. Tauch and V. Azevedo (2013). "The pan-genome of the animal pathogen Corynebacterium pseudotuberculosis reveals differences in genome plasticity between the biovar ovis and equi strains." PLoS ONE **8**(1): e53818.
Stavrinides, J., H. C. McCann and D. S. Guttman (2008). "Host-pathogen interplay and the evolution of bacterial effectors." Cell Microbiol **10**(2): 285-292.
Strobl, C., A.-L. Boulesteix, A. Zeileis and T. Hothorn (2007). "Bias in random forest variable importance measures: Illustrations, sources and a solution." BMC bioinformatics **8**(1): 1.
Sutton, G. G., O. White, M. D. Adams and A. Kerlavage (1995). "TIGR Assembler: A new tool for assembling large shotgun sequencing projects." **1**: 9-19+.
Svetnik, V., A. Liaw, C. Tong and T. Wang (2004). Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. Multiple Classifier Systems, Springer**:** 334-343.
Tettelin, H. (2009). "The bacterial pan-genome and reverse vaccinology." Genome Dyn **6**: 35-47.
Tsai, I. J., T. D. Otto and M. Berriman (2010). "Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps." Genome Biol **11**(4): R41.
Tumapa, S., M. T. Holden, M. Vesaratchavest, V. Wuthiekanun, D. Limmathurotsakul, W. Chierakul, E. J. Feil, B. J. Currie, N. P. Day and W. C. Nierman (2008). "Burkholderia pseudomallei genome plasticity associated with genomic island variation." BMC genomics **9**(1): 1.
Tuorto, S. J., P. Darias, L. R. McGuinness, N. Panikov, T. Zhang, M. M. Häggblom and L. J. Kerkhof (2014). "Bacterial genome replication at subzero temperatures in permafrost." The ISME journal **8**(1): 139-149.
Unden, G., S. Becker, J. Bongaerts, G. Holighaus, J. Schirawski and S. Six (1995). "O2-sensing and O2-dependent gene regulation in facultatively anaerobic bacteria." Archives of microbiology **164**(2): 81-90.
Valens, M., S. Penaud, M. Rossignol, F. Cornet and F. Boccard (2004). "Macrodomain organization of the Escherichia coli chromosome." The EMBO journal **23**(21): 4330-4341.
van Noort, V., B. Bradatsch, M. Arumugam, S. Amlacher, G. Bange, C. Creevey, S. Falk, D. R. Mende, I. Sinning and E. Hurt (2013). "Consistent mutational paths predict eukaryotic thermostability." BMC evolutionary biology **13**(1): 1.

Walsh, C. (2003). "Where will new antibiotics come from?" Nat Rev Microbiol **1**(1): 65-70.
Wang, Q., Z. Cen and J. Zhao (2015). "The survival mechanisms of thermophiles at high temperatures: an angle of omics." Physiology **30**(2): 97-106.
Wirth, T., F. Hildebrand, C. Allix-Béguec, F. Wölbeling, T. Kubica, K. Kremer, D. van Soolingen, S. Rüsch-Gerdes, C. Locht and S. Brisse (2008). "Origin, spread and demography of the Mycobacterium tuberculosis complex." PLoS Pathogens **4**(9): e1000160.
Wittkop, T., D. Emig, S. Lange, S. Rahmann, M. Albrecht, J. H. Morris, S. Böcker, J. Stoye and J. Baumbach (2010). "Partitioning biological data with transitivity clustering." Nature methods **7**(6): 419-420.
Wooldridge, K. (2009). Bacterial secreted proteins: secretory mechanisms and role in pathogenesis, The Publisher.
Wu, H., Z. Zhang, S. Hu and J. Yu (2012). "On the molecular mechanism of GC content variation among eubacterial genomes." Biology Direct **7**: 2-2.
Waack, S., O. Keller, R. Asper, T. Brodag, C. Damm, W. F. Fricke, K. Surovcik, P. Meinicke and R. Merkl (2006). "Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models." BMC Bioinformatics **7**: 142.
Yu, L. and H. Liu (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. ICML.

APPENDIX A

S. TABLE 1 – LIST OF ORGANISMS USED IN THE "HOMOLOGOUS GENES ANALYSIS" SECTION**. ACTINOBACTERIAL SPECIES CLASSIFIED INTO FOUR PATHOGENICITY LIFESTYLE CLASSES: (HP) EXCLUSIVELY HUMAN PATHOGENIC; (BP) BROAD-SPECTRUM PATHOGENIC; (OP), OPPORTUNISTIC PATHOGENIC; AND (NP) NON-PATHOGENS.**

ID	Organism	Lifestyle
1	Acidimicrobidae_bacterium_YM16_304_uid193703	NP
2	Acidimicrobium_ferrooxidans_DSM_10331_uid59215	NP
3	Acidothermus_cellulolyticus_11B_uid58501	NP
4	Actinoplanes_missouriensis_431_uid158169	NP
5	Actinoplanes_SE50_110_uid162333	NP
6	Actinosynnema_mirum_DSM_43827_uid58951	NP
7	Amycolatopsis_mediterranei_S699_uid158689	NP
8	Amycolatopsis_mediterranei_S699_uid171830	NP
9	Amycolatopsis_mediterranei_U32_uid50565	NP
10	Amycolicicoccus_subflavus_DQS3_9A1_uid67253	NP
11	Arcanobacterium_haemolyticum_DSM_20595_uid49489	BP
12	Arthrobacter_arilaitensis_Re117_uid53509	NP
13	Arthrobacter_aurescens_TC1_uid58109	NP
14	Arthrobacter_chlorophenolicus_A6_uid58969	NP
15	Arthrobacter_FB24_uid58141	NP
16	Arthrobacter_nitroguajacolicus_Rue61a_uid174511	NP
17	Arthrobacter_phenanthrenivorans_Sphe3_uid63629	NP
18	Atopobium_parvulum_DSM_20469_uid59195	NP
19	Beutenbergia_cavernae_DSM_12333_uid59047	NP
20	Bifidobacterium_adolescentis_ATCC_15703_uid58559	NP
21	Bifidobacterium_animalis_ATCC_25527_uid162513	NP
22	Bifidobacterium_animalis_lactis_AD011_uid58911	NP
23	Bifidobacterium_animalis_lactis_B420_uid163691	NP
24	Bifidobacterium_animalis_lactis_BB_12_uid158871	NP
25	Bifidobacterium_animalis_lactis_Bi_07_uid163693	NP
26	Bifidobacterium_animalis_lactis_Bl_04_uid59359	NP
27	Bifidobacterium_animalis_lactis_BLC1_uid158867	NP
28	Bifidobacterium_animalis_lactis_CNCM_I_2494_uid158869	NP
29	Bifidobacterium_animalis_lactis_DSM_10140_uid59357	NP
30	Bifidobacterium_animalis_lactis_V9_uid158865	NP
31	Bifidobacterium_asteroides_PRL2011_uid176921	NP
32	Bifidobacterium_bifidum_BGN4_uid167988	NP
33	Bifidobacterium_bifidum_PRL2010_uid59883	NP
34	Bifidobacterium_bifidum_S17_uid59545	NP
35	Bifidobacterium_breve_ACS_071_V_Sch8b_uid158863	NP
36	Bifidobacterium_breve_UCC2003_uid193702	NP
37	Bifidobacterium_dentium_Bd1_uid43091	OP
38	Bifidobacterium_longum_BBMN68_uid60163	NP
39	Bifidobacterium_longum_DJO10A_uid58833	NP
40	Bifidobacterium_longum_F8_uid197184	NP
41	Bifidobacterium_longum_infantis_157F_uid62693	NP
42	Bifidobacterium_longum_infantis_ATCC_15697_uid159865	NP
43	Bifidobacterium_longum_infantis_ATCC_15697_uid58677	NP
44	Bifidobacterium_longum_JCM_1217_uid62695	NP
45	Bifidobacterium_longum_JDM301_uid49131	NP
46	Bifidobacterium_longum_KACC_91563_uid158861	NP
47	Bifidobacterium_longum_NCC2705_uid57939	NP
48	Bifidobacterium_thermophilum_RBL67_uid193770	NP
49	Blastococcus_saxobsidens_DD2_uid89391	NP
50	Brachybacterium_faecium_DSM_4810_uid58649	NP
51	Catenulispora_acidiphila_DSM_44928_uid59077	NP

52	Cellulomonas_fimi_ATCC_484_uid66779	NP
53	Cellulomonas_flavigena_DSM_20109_uid48821	NP
54	_Cellvibrio__gilvus_ATCC_13127_uid68143	NP
55	Conexibacter_woesei_DSM_14684_uid43467	NP
56	Corynebacterium_aurimucosum_ATCC_700975_uid59409	OP
57	Corynebacterium_callunae_DSM_20147_uid193714	NP
58	Corynebacterium_diphtheriae_241_uid83607	HP
59	Corynebacterium_diphtheriae_31A_uid84309	HP
60	Corynebacterium_diphtheriae_BH8_uid84311	HP
61	Corynebacterium_diphtheriae_C7__beta__uid84313	HP
62	Corynebacterium_diphtheriae_CDCE_8392_uid84295	HP
63	Corynebacterium_diphtheriae_HC01_uid84297	HP
64	Corynebacterium_diphtheriae_HC02_uid84317	HP
65	Corynebacterium_diphtheriae_HC03_uid84299	HP
66	Corynebacterium_diphtheriae_HC04_uid84301	HP
67	Corynebacterium_diphtheriae_INCA_402_uid83605	HP
68	Corynebacterium_diphtheriae_NCTC_13129_uid57691	HP
69	Corynebacterium_diphtheriae_PW8_uid84303	HP
70	Corynebacterium_diphtheriae_VA01_uid84305	HP
71	Corynebacterium_efficiens_YS_314_uid62905	NP
72	Corynebacterium_glutamicum_ATCC_13032_uid193708	NP
73	Corynebacterium_glutamicum_ATCC_13032_uid57905	NP
74	Corynebacterium_glutamicum_ATCC_13032_uid61611	NP
75	Corynebacterium_glutamicum_R_uid58897	NP
76	Corynebacterium_halotolerans_YIM_70093_DSM_44683_uid189953	NP
77	Corynebacterium_jeikeium_K411_uid58399	HP
78	Corynebacterium_kroppenstedtii_DSM_44385_uid59411	HP
79	Corynebacterium_pseudotuberculosis_1002_uid159677	BP
80	Corynebacterium_pseudotuberculosis_1_06_A_uid159665	BP
81	Corynebacterium_pseudotuberculosis_258_uid167260	BP
82	Corynebacterium_pseudotuberculosis_267_uid162175	BP
83	Corynebacterium_pseudotuberculosis_316_uid89381	BP
84	Corynebacterium_pseudotuberculosis_31_uid162167	BP
85	Corynebacterium_pseudotuberculosis_3_99_5_uid83609	BP
86	Corynebacterium_pseudotuberculosis_42_02_A_uid159669	BP
87	Corynebacterium_pseudotuberculosis_C231_uid159675	BP
88	Corynebacterium_pseudotuberculosis_CIP_52_97_uid159667	BP
89	Corynebacterium_pseudotuberculosis_Cp162_uid168258	BP
90	Corynebacterium_pseudotuberculosis_FRC41_uid50585	BP
91	Corynebacterium_pseudotuberculosis_I19_uid159673	BP
92	Corynebacterium_pseudotuberculosis_P54B96_uid157909	BP
93	Corynebacterium_pseudotuberculosis_PAT10_uid159671	BP
94	Corynebacterium_resistens_DSM_45100_uid50555	HP
95	Corynebacterium_ulcerans_0102_uid169879	BP
96	Corynebacterium_ulcerans_809_uid159659	BP
97	Corynebacterium_ulcerans_BR_AD22_uid68291	BP
98	Corynebacterium_urealyticum_DSM_7109_uid61639	OP
99	Corynebacterium_urealyticum_DSM_7111_uid188688	OP
100	Corynebacterium_variabile_DSM_44702_uid62003	NP
101	Cryptobacterium_curtum_DSM_15641_uid59041	OP
102	Eggerthella_lenta_DSM_2243_uid59079	HP
103	Eggerthella_YY7918_uid68707	NP
104	Frankia_EuI1c_uid42615	NP
105	Frankia_symbiont_of_Datisca_glomerata_uid46257	NP
106	Gardnerella_vaginalis_409_05_uid43211	HP
107	Gardnerella_vaginalis_ATCC_14019_uid55487	HP
108	Gardnerella_vaginalis_HMP9231_uid162045	HP
109	Geodermatophilus_obscurus_DSM_43160_uid43725	NP
110	Gordonia_bronchialis_DSM_43247_uid41403	OP
111	Gordonia_KTR9_uid174812	NP

112	Gordonia_polyisoprenivorans_VH2_uid86651	NP
113	Gordonibacter_pamelaeae_7_10_1_b_uid197167	OP
114	Intrasporangium_calvum_DSM_43043_uid61729	NP
115	Isoptericola_variabilis_225_uid67501	NP
116	Jonesia_denitrificans_DSM_20603_uid59053	BP
117	Kineococcus_radiotolerans_SRS30216_uid58067	NP
118	Kitasatospora_setae_KM_6054_uid77027	NP
119	Kocuria_rhizophila_DC2201_uid59099	NP
120	Kribbella_flavida_DSM_17836_uid43465	NP
121	Kytococcus_sedentarius_DSM_20547_uid59071	OP
122	Microbacterium_testaceum_StLB037_uid62789	NP
123	Micrococcus_luteus_NCTC_2665_uid59033	NP
124	Microlunatus_phosphovorus_NM_1_uid68055	NP
125	Micromonospora_aurantiaca_ATCC_27029_uid42501	NP
126	Micromonospora_L5_uid45895	NP
127	Mobiluncus_curtisii_ATCC_43063_uid49695	HP
128	Modestobacter_marinus_uid167487	NP
129	Mycobacterium_abscessus_uid61613	OP
130	Mycobacterium_africanum_GM041182_uid68839	HP
131	Mycobacterium_avium_104_uid57693	OP
133	Mycobacterium_bovis_AF2122_97_uid57695	BP
134	Mycobacterium_bovis_BCG_Korea_1168P_uid189029	BP
135	Mycobacterium_bovis_BCG_Mexico_uid86889	BP
136	Mycobacterium_bovis_BCG_Pasteur_1173P2_uid58781	BP
137	Mycobacterium_bovis_BCG_Tokyo_172_uid59281	BP
138	Mycobacterium_canettii_CIPT_140010059_uid70731	HP
139	Mycobacterium_canettii_CIPT_140060008_uid184829	HP
140	Mycobacterium_canettii_CIPT_140070008_uid184832	HP
141	Mycobacterium_canettii_CIPT_140070010_uid184828	HP
142	Mycobacterium_canettii_CIPT_140070017_uid184830	HP
143	Mycobacterium_chubuense_NBB4_uid168322	NP
144	Mycobacterium_gilvum_PYR_GCK_uid59421	NP
145	Mycobacterium_gilvum_Spyr1_uid61403	NP
146	Mycobacterium_indicus_pranii_MTCC_9506_uid175523	NP
147	Mycobacterium_intracellulare_ATCC_13950_uid167994	OP
148	Mycobacterium_intracellulare_MOTT_02_uid89387	OP
149	Mycobacterium_intracellulare_MOTT_64_uid89385	OP
150	Mycobacterium_JDM601_uid67369	HP
151	Mycobacterium_JLS_uid58489	NP
152	Mycobacterium_KMS_uid58491	NP
153	Mycobacterium_leprae_Br4923_uid59293	HP
154	Mycobacterium_leprae_TN_uid57697	HP
156	Mycobacterium_marinum_M_uid59423	BP
157	Mycobacterium_massiliense_GO_06_uid170732	OP
158	Mycobacterium_MCS_uid58465	NP
159	Mycobacterium_MOTT36Y_uid164001	HP
160	Mycobacterium_rhodesiae_NBB3_uid75107	HP
161	Mycobacterium_smegmatis_JS623_uid184820	OP
162	Mycobacterium_smegmatis_MC2_155_uid171958	OP
163	Mycobacterium_smegmatis_MC2_155_uid57701	OP
164	Mycobacterium_tuberculosis_Beijing_NITR203_uid197218	HP
165	Mycobacterium_tuberculosis_CCDC5079_uid161943	HP
166	Mycobacterium_tuberculosis_CCDC5180_uid161941	HP
167	Mycobacterium_tuberculosis_CDC1551_uid57775	HP
168	Mycobacterium_tuberculosis_CTRI_2_uid161997	HP
169	Mycobacterium_tuberculosis_Erdman___ATCC_35801_uid193763	HP
170	Mycobacterium_tuberculosis_F11_uid58417	HP
171	Mycobacterium_tuberculosis_H37Ra_uid58853	HP
172	Mycobacterium_tuberculosis_H37Rv_uid170532	HP
173	Mycobacterium_tuberculosis_H37Rv_uid57777	HP

174	Mycobacterium_tuberculosis_KZN_1435_uid59069	HP
175	Mycobacterium_tuberculosis_KZN_4207_uid83619	HP
176	Mycobacterium_tuberculosis_KZN_605_uid54947	HP
177	Mycobacterium_tuberculosis_RGTB327_uid157907	HP
178	Mycobacterium_tuberculosis_RGTB423_uid162179	HP
179	Mycobacterium_tuberculosis_uid185758	HP
180	Mycobacterium_tuberculosis_UT205_uid162183	HP
181	Mycobacterium_ulcerans_Agy99_uid62939	HP
182	Mycobacterium_vanbaalenii_PYR_1_uid58463	NP
183	Nakamurella_multipartita_DSM_44233_uid59221	NP
184	Nocardia_brasiliensis_ATCC_700358_uid86913	HP
185	Nocardia_cyriacigeorgica_GUH_2_uid89395	BP
186	Nocardia_farcinica_IFM_10152_uid58203	OP
187	Nocardioides_JS614_uid58149	NP
188	Nocardiopsis_alba_ATCC_BAA_2165_uid174334	NP
189	Nocardiopsis_dassonvillei_DSM_43111_uid49483	HP
190	Olsenella_uli_DSM_7084_uid51367	HP
191	Propionibacterium_acidipropionici_ATCC_4875_uid179069	NP
192	Propionibacterium_acnes_266_uid162059	HP
193	Propionibacterium_acnes_6609_uid162137	HP
194	Propionibacterium_acnes_ATCC_11828_uid162177	HP
195	Propionibacterium_acnes_C1_uid176501	HP
196	Propionibacterium_acnes_HL096PA1_uid198524	HP
197	Propionibacterium_acnes_KPA171202_uid58101	HP
198	Propionibacterium_acnes_SK137_uid48071	HP
199	Propionibacterium_acnes_TypeIA2_P_acn17_uid80735	HP
200	Propionibacterium_acnes_TypeIA2_P_acn31_uid80733	HP
201	Propionibacterium_acnes_TypeIA2_P_acn33_uid80745	HP
202	Propionibacterium_avidum_44067_uid197361	HP
203	Propionibacterium_freudenreichii_shermanii_CIRM_BIA1_uid49535	NP
204	Propionibacterium_propionicum_F0230a_uid170533	HP
205	Pseudonocardia_dioxanivorans_CB1190_uid65087	NP
208	Rhodococcus_erythropolis_PR4_uid59019	NP
209	Rhodococcus_jostii_RHA1_uid58325	NP
210	Rhodococcus_opacus_B4_uid13791	NP
211	Rothia_dentocariosa_ATCC_17931_uid49331	OP
212	Rothia_mucilaginosa_uid43093	OP
213	Rubrobacter_xylanophilus_DSM_9941_uid58057	NP
214	Saccharomonospora_viridis_DSM_43017_uid59055	HP
215	Saccharopolyspora_erythraea_NRRL_2338_uid62947	NP
216	Saccharothrix_espanaensis_DSM_44229_uid184826	NP
217	Salinispora_arenicola_CNS_205_uid58659	NP
218	Salinispora_tropica_CNB_440_uid58565	NP
219	Sanguibacter_keddieii_DSM_10542_uid40845	NP
220	Segniliparus_rotundus_DSM_44985_uid49049	OP
221	Slackia_heliotrinireducens_DSM_20476_uid59051	NP
222	Stackebrandtia_nassauensis_DSM_44728_uid46663	NP
223	Streptomyces_albus_J1074_uid196849	NP
224	Streptomyces_avermitilis_MA_4680_uid57739	NP
225	Streptomyces_bingchenggensis_BCW_1_uid82931	NP
226	Streptomyces_cattleya_NRRL_8057___DSM_46488_uid162187	NP
227	Streptomyces_cattleya_NRRL_8057_uid77117	NP
228	Streptomyces_coelicolor_A3_2__uid57801	NP
229	Streptomyces_davawensis_JCM_4913_uid193657	NP
230	Streptomyces_flavogriseus_ATCC_33331_uid40839	NP
231	Streptomyces_griseus_NBRC_13350_uid58983	NP
232	Streptomyces_hygroscopicus_jinggangensis_5008_uid89409	NP
233	Streptomyces_hygroscopicus_jinggangensis_TL01_uid189753	NP
234	Streptomyces_venezuelae_ATCC_10712_uid177080	NP
235	Streptomyces_violaceusniger_Tu_4113_uid52609	NP

236	Streptosporangium_roseum_DSM_43021_uid42521	NP
237	Thermobifida_fusca_YX_uid57703	NP
238	Thermobispora_bispora_DSM_43833_uid48999	NP
239	Thermomonospora_curvata_DSM_43183_uid41885	NP
240	Tropheryma_whipplei_TW08_27_uid57961	HP
241	Tropheryma_whipplei_Twist_uid57705	HP
242	Tsukamurella_paurometabola_DSM_20162_uid48829	OP
243	Verrucosispora_maris_AB_18_032_uid66297	NP
244	Xylanimonas_cellulosilytica_DSM_15894_uid41935	NP

S. TABLE 2 – LIST OF ORGANISMS OF THE GENUS LISTERIA AND CORYNEBACTERIUM USED TO EVALUATE LISSI.

Organism	**Accession**
Listeria monocytogenes serotype 4b str. F2365	NC_002973
Listeria monocytogenes EGD-e	NC_003210
Listeria innocua Clip11262	NC_003212
Listeria welshimeri serovar 6b str. SLCC5334	NC_008555
Listeria monocytogenes HCC23	NC_011660
Listeria monocytogenes serotype 4b str. CLIP 80459	NC_012488
Listeria monocytogenes 08-5578	NC_013766
Listeria monocytogenes 08-5923	NC_013768
Listeria seeligeri serovar 1/2b str. SLCC3954	NC_013891
Listeria ivanovii subsp. ivanovii PAM 55	NC_016011
Corynebacterium diphtheriae NCTC 13129	NC_002935
Corynebacterium glutamicum ATCC 13032	NC_003450
Corynebacterium efficiens YS-314	NC_004369
Corynebacterium glutamicum ATCC 13032	NC_006958
Corynebacterium jeikeium K411	NC_007164
Corynebacterium glutamicum R	NC_009342
Corynebacterium urealyticum DSM 7109	NC_010545
Corynebacterium aurimucosum ATCC 700975	NC_012590
Corynebacterium kroppenstedtii DSM 44385	NC_012704
Corynebacterium pseudotuberculosis FRC41	NC_014329
Corynebacterium ulcerans BR-AD22	NC_015683
Corynebacterium variabile DSM 44702	NC_015859

S. TABLE 3 – LIST OF ORGANISMS USED TO GENERATE LISSI RESULTS. **ACTINOBACTERIAL SPECIES CLASSIFIED ACCORDING TO THE LEVEL OXYGEN TOLERANCE, HABITATS, AND PATHOGENICITY. THE TABLE CONTAINS A LIST OF AEROBES (AE), ANAEROBES (AN), FACULTATIVE (FA), SOIL (SO), AQUATIC (AQ), NON-PATHOGENIC (NP), AND PATHOGENIC (PA). NOTE THAT THE SAME ORGANISM CAN RECEIVE DIFFERENT LABELS.**

Accession	Organism	Lifestyle
NC_002677	Mycobacterium leprae TN	AE
NC_002755	Mycobacterium tuberculosis CDC1551	AE
NC_002945	Mycobacterium bovis AF2122/97	AE
NC_003155	Streptomyces avermitilis MA-4680 = NBRC 14893	AE
NC_004551	Tropheryma whipplei TW08/27	AE
NC_004572	Tropheryma whipplei str. Twist	AE
NC_007333	Thermobifida fusca YX	AE

NC_008146	Mycobacterium sp. MCS	AE
NC_008148	Rubrobacter xylanophilus DSM 9941	AE
NC_008268	Rhodococcus jostii RHA1	AE
NC_008541	Arthrobacter sp. FB24	AE
NC_008578	Acidothermus cellulolyticus 11B	AE
NC_008611	Mycobacterium ulcerans Agy99	AE
NC_008699	Nocardioides sp. JS614	AE
NC_008705	Mycobacterium sp. KMS	AE
NC_008711	Arthrobacter aurescens TC1	AE
NC_008726	Mycobacterium vanbaalenii PYR-1	AE
NC_008769	Mycobacterium bovis BCG str. Pasteur 1173P2	AE
NC_009077	Mycobacterium sp. JLS	AE
NC_009142	Saccharopolyspora erythraea NRRL 2338	AE
NC_009380	Salinispora tropica CNB-440	AE
NC_009525	Mycobacterium tuberculosis H37Ra	AE
NC_009565	Mycobacterium tuberculosis F11	AE
NC_009664	Kineococcus radiotolerans SRS30216 = ATCC BAA-149	AE
NC_009953	Salinispora arenicola CNS-205	AE
NC_010572	Streptomyces griseus subsp. griseus NBRC 13350 (Streptomyces	AE
NC_010612	Mycobacterium marinum M	AE
NC_010617	Kocuria rhizophila DC2201	AE
NC_011886	Arthrobacter chlorophenolicus A6	AE
NC_011896	Mycobacterium leprae Br4923	AE
NC_012207	Mycobacterium bovis BCG str. Tokyo 172	AE
NC_012490	Rhodococcus erythropolis PR4	AE
NC_012522	Rhodococcus opacus B4	AE
NC_012669	Beutenbergia cavernae DSM 12333	AE
NC_012803	Micrococcus luteus NCTC 2665	AE
NC_012943	Mycobacterium tuberculosis KZN 1435	AE
NC_013093	Actinosynnema mirum DSM 43827	AE
NC_013124	Acidimicrobium ferrooxidans DSM 10331	AE
NC_013131	Catenulispora acidiphila DSM 44928	AE
NC_013159	Saccharomonospora viridis DSM 43017	AE
NC_013172	Brachybacterium faecium DSM 4810	AE
NC_013235	Nakamurella multipartita DSM 44233	AE
NC_013530	Xylanimonas cellulosilytica DSM 15894	AE
NC_013595	Streptosporangium roseum DSM 43021	AE
NC_013729	Kribbella flavida DSM 17836	AE
NC_013739	Conexibacter woesei DSM 14684	AE
NC_013757	Geodermatophilus obscurus DSM 43160	AE
NC_013947	Stackebrandtia nassauensis DSM 44728	AE
NC_014165	Thermobispora bispora DSM 43833	AE
NC_014211	Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111	AE
NC_014391	Micromonospora aurantiaca ATCC 27029	AE
NC_014550	Arthrobacter arilaitensis Re117	AE
NC_014666	Frankia sp. EuI1c	AE
NC_014815	Micromonospora sp. L5	AE
NC_014830	Intrasporangium calvum DSM 43043	AE
NC_015145	Arthrobacter phenanthrenivorans Sphe3	AE
NC_015312	Pseudonocardia dioxanivorans CB1190	AE
NC_015576	Mycobacterium sp. JDM601	AE

NC_015635	Microlunatus phosphovorus NM-1	AE
NC_015656	Frankia symbiont of Datisca glomerata	AE
NC_015758	Mycobacterium africanum GM041182	AE
NC_015859	Corynebacterium variabile DSM 44702	AE
NC_015957	Streptomyces violaceusniger Tu 4113	AE
NC_004307	Bifidobacterium longum NCC2705	AN
NC_006085	Propionibacterium acnes KPA171202	AN
NC_008618	Bifidobacterium adolescentis ATCC 15703	AN
NC_010816	Bifidobacterium longum DJO10A	AN
NC_011593	Bifidobacterium longum subsp. infantis ATCC 15697 = JCM 1222 = DSM	AN
NC_011835	Bifidobacterium animalis subsp. lactis AD011	AN
NC_012704	Corynebacterium kroppenstedtii DSM 44385	AN
NC_012814	Bifidobacterium animalis subsp. lactis Bl-04	AN
NC_012815	Bifidobacterium animalis subsp. lactis DSM 10140	AN
NC_013165	Slackia heliotrinireducens DSM 20476	AN
NC_013203	Atopobium parvulum DSM 20469	AN
NC_013204	Eggerthella lenta DSM 2243	AN
NC_013721	Gardnerella vaginalis 409-05	AN
NC_014215	Propionibacterium freudenreichii subsp. shermanii CIRM-BIA1	AN
NC_014218	Arcanobacterium haemolyticum DSM 20595	AN
NC_014246	Mobiluncus curtisii ATCC 43063	AN
NC_014363	Olsenella uli DSM 7084	AN
NC_014616	Bifidobacterium bifidum S17	AN
NC_014638	Bifidobacterium bifidum PRL2010	AN
NC_014644	Gardnerella vaginalis ATCC 14019	AN
NC_015052	Bifidobacterium longum subsp. infantis 157F	AN
NC_015067	Bifidobacterium longum subsp. longum JCM 1217	AN
NC_015673	Corynebacterium resistens DSM 45100	AN
NC_004369	Corynebacterium efficiens YS-314	FA
NC_006958	Corynebacterium glutamicum ATCC 13032	FA
NC_007164	Corynebacterium jeikeium K411	FA
NC_009342	Corynebacterium glutamicum R	FA
NC_013174	Jonesia denitrificans DSM 20603	FA
NC_013521	Sanguibacter keddieii DSM 10542	FA
NC_014151	Cellulomonas flavigena DSM 20109	FA
NC_014329	Corynebacterium pseudotuberculosis FRC41	FA
NC_015683	Corynebacterium ulcerans BR-AD22	FA
NC_008578	Acidothermus cellulolyticus 11B	AQ
NC_008611	Mycobacterium ulcerans Agy99	AQ
NC_009380	Salinispora tropica CNB-440	AQ
NC_009664	Kineococcus radiotolerans SRS30216 = ATCC BAA-149	AQ
NC_009953	Salinispora arenicola CNS-205	AQ
NC_010617	Kocuria rhizophila DC2201	AQ
NC_012490	Rhodococcus erythropolis PR4	AQ
NC_013124	Acidimicrobium ferrooxidans DSM 10331	AQ
NC_015312	Pseudonocardia dioxanivorans CB1190	AQ
NC_003155	Streptomyces avermitilis MA-4680 = NBRC 14893	SO
NC_006958	Corynebacterium glutamicum ATCC 13032	SO
NC_008146	Mycobacterium sp. MCS	SO
NC_008148	Rubrobacter xylanophilus DSM 9941	SO
NC_008268	Rhodococcus jostii RHA1	SO

NC_008541	Arthrobacter sp. FB24	SO
NC_008699	Nocardioides sp. JS614	SO
NC_008705	Mycobacterium sp. KMS	SO
NC_008711	Arthrobacter aurescens TC1	SO
NC_009077	Mycobacterium sp. JLS	SO
NC_009142	Saccharopolyspora erythraea NRRL 2338	SO
NC_009342	Corynebacterium glutamicum R	SO
NC_010572	Streptomyces griseus subsp. griseus NBRC 13350 (Streptomyces	SO
NC_011886	Arthrobacter chlorophenolicus A6	SO
NC_012669	Beutenbergia cavernae DSM 12333	SO
NC_012803	Micrococcus luteus NCTC 2665	SO
NC_013093	Actinosynnema mirum DSM 43827	SO
NC_013131	Catenulispora acidiphila DSM 44928	SO
NC_013172	Brachybacterium faecium DSM 4810	SO
NC_013530	Xylanimonas cellulosilytica DSM 15894	SO
NC_013595	Streptosporangium roseum DSM 43021	SO
NC_013729	Kribbella flavida DSM 17836	SO
NC_013739	Conexibacter woesei DSM 14684	SO
NC_013757	Geodermatophilus obscurus DSM 43160	SO
NC_013947	Stackebrandtia nassauensis DSM 44728	SO
NC_014151	Cellulomonas flavigena DSM 20109	SO
NC_014391	Micromonospora aurantiaca ATCC 27029	SO
NC_014666	Frankia sp. EuI1c	SO
NC_014815	Micromonospora sp. L5	SO
NC_015145	Arthrobacter phenanthrenivorans Sphe3	SO
NC_015564	Amycolicicoccus subflavus DQS3-9A1	SO
NC_015656	Frankia symbiont of Datisca glomerata	SO
NC_015859	Corynebacterium variabile DSM 44702	SO
NC_015957	Streptomyces violaceusniger Tu 4113	SO
NC_003155	Streptomyces avermitilis MA-4680	NP
NC_004307	Bifidobacterium longum NCC2705	NP
NC_004369	Corynebacterium efficiens YS-314	NP
NC_006958	Corynebacterium glutamicum ATCC 13032	NP
NC_007333	Thermobifida fusca YX	NP
NC_008146	Mycobacterium sp. MCS	NP
NC_008148	Rubrobacter xylanophilus DSM 9941	NP
NC_008268	Rhodococcus jostii RHA1	NP
NC_008541	Arthrobacter sp. FB24	NP
NC_008578	Acidothermus cellulolyticus 11B	NP
NC_008618	Bifidobacterium adolescentis ATCC 15703	NP
NC_008699	Nocardioides sp. JS614	NP
NC_008705	Mycobacterium sp. KMS	NP
NC_008711	Arthrobacter aurescens TC1	NP
NC_008726	Mycobacterium vanbaalenii PYR-1	NP
NC_009077	Mycobacterium sp. JLS	NP
NC_009142	Saccharopolyspora erythraea NRRL 2338	NP
NC_009342	Corynebacterium glutamicum R	NP
NC_009380	Salinispora tropica CNB-440	NP
NC_009664	Kineococcus radiotolerans SRS30216	NP
NC_009953	Salinispora arenicola CNS-205	NP
NC_010572	Streptomyces griseus subsp. griseus NBRC 13350 (Streptomyces	NP

NC_010617	Kocuria rhizophila DC2201	NP
NC_010816	Bifidobacterium longum DJO10A	NP
NC_011593	Bifidobacterium longum subsp. infantis ATCC 15697 = JCM 1222	NP
NC_011835	Bifidobacterium animalis subsp. lactis AD011	NP
NC_011886	Arthrobacter chlorophenolicus A6	NP
NC_012490	Rhodococcus erythropolis PR4	NP
NC_012522	Rhodococcus opacus B4	NP
NC_012669	Beutenbergia cavernae DSM 12333	NP
NC_012803	Micrococcus luteus NCTC 2665	NP
NC_012814	Bifidobacterium animalis subsp. lactis Bl-04	NP
NC_012815	Bifidobacterium animalis subsp. lactis DSM 10140	NP
NC_013093	Actinosynnema mirum DSM 43827	NP
NC_013124	Acidimicrobium ferrooxidans DSM 10331	NP
NC_013131	Catenulispora acidiphila DSM 44928	NP
NC_013165	Slackia heliotrinireducens DSM 20476	NP
NC_013172	Brachybacterium faecium DSM 4810	NP
NC_013203	Atopobium parvulum DSM 20469	NP
NC_013235	Nakamurella multipartita DSM 44233	NP
NC_013521	Sanguibacter keddieii DSM 10542	NP
NC_013530	Xylanimonas cellulosilytica DSM 15894	NP
NC_013595	Streptosporangium roseum DSM 43021	NP
NC_013729	Kribbella flavida DSM 17836	NP
NC_013739	Conexibacter woesei DSM 14684	NP
NC_013757	Geodermatophilus obscurus DSM 43160	NP
NC_013947	Stackebrandtia nassauensis DSM 44728	NP
NC_014151	Cellulomonas flavigena DSM 20109	NP
NC_014165	Thermobispora bispora DSM 43833	NP
NC_014169	Bifidobacterium longum subsp. longum JDM301	NP
NC_014215	Propionibacterium freudenreichii subsp. shermanii CIRM-BIA1	NP
NC_014318	Amycolatopsis mediterranei U32	NP
NC_014391	Micromonospora aurantiaca ATCC 27029	NP
NC_014550	Arthrobacter arilaitensis Re117	NP
NC_014616	Bifidobacterium bifidum S17	NP
NC_014638	Bifidobacterium bifidum PRL2010	NP
NC_014666	Frankia sp. EuI1c	NP
NC_014814	Mycobacterium gilvum Spyr1	NP
NC_014815	Micromonospora sp. L5	NP
NC_014830	Intrasporangium calvum DSM 43043	NP
NC_015052	Bifidobacterium longum subsp. infantis 157F	NP
NC_015067	Bifidobacterium longum subsp. longum JCM 1217	NP
NC_015125	Microbacterium testaceum StLB037	NP
NC_015145	Arthrobacter phenanthrenivorans Sphe3	NP
NC_015312	Pseudonocardia dioxanivorans CB1190	NP
NC_015434	Verrucosispora maris AB-18-032	NP
NC_015514	Cellulomonas fimi ATCC 484	NP
NC_015564	Amycolicicoccus subflavus DQS3-9A1	NP
NC_015588	Isoptericola variabilis 225	NP
NC_015635	Microlunatus phosphovorus NM-1	NP
NC_015656	Frankia symbiont of Datisca glomerata	NP
NC_015671	(Cellvibrio) gilvus ATCC 13127	NP
NC_015738	Eggerthella sp. YY7918	NP

NC_015859	Corynebacterium variabile DSM 44702	NP
NC_015957	Streptomyces violaceusniger Tu 4113	NP
NC_016109	Kitasatospora setae KM-6054	NP
NC_016111	Streptomyces cattleya NRRL 8057 = DSM 46488	NP
NC_016114	Streptomyces flavogriseus ATCC 33331	NP
NC_016582	Streptomyces bingchenggensis BCW-1	NP
NC_016906	Gordonia polyisoprenivorans VH2	NP
NC_016943	Blastococcus saxobsidens DD2	NP
NC_017093	Actinoplanes missouriensis 431	NP
NC_017214	Bifidobacterium animalis subsp. lactis BB-12	NP
NC_017215	Bifidobacterium animalis subsp. lactis CNCM I-2494	NP
NC_017216	Bifidobacterium animalis subsp. lactis BLC1	NP
NC_017217	Bifidobacterium animalis subsp. lactis V9	NP
NC_017218	Bifidobacterium breve ACS-071-V-Sch8b	NP
NC_017221	Bifidobacterium longum subsp. longum KACC 91563	NP
NC_017765	Streptomyces hygroscopicus subsp. jinggangensis 5008	NP
NC_017803	Actinoplanes sp. SE50/110	NP
NC_017866	Bifidobacterium animalis subsp. lactis B420	NP
NC_017867	Bifidobacterium animalis subsp. lactis Bi-07	NP
NC_017999	Bifidobacterium bifidum BGN4	NP
NC_018027	Mycobacterium chubuense NBB4	NP
NC_018266	Amycolatopsis mediterranei S699	NP
NC_018531	Arthrobacter sp. Rue61a	NP
NC_018581	Gordonia sp. KTR9	NP
NC_018612	Mycobacterium indicus pranii MTCC 9506	NP
NC_018720	Bifidobacterium asteroides PRL2011	NP
NC_018750	Streptomyces venezuelae ATCC 10712	NP
NC_019395	Propionibacterium acidipropionici ATCC 4875	NP
NC_019673	Saccharothrix espanaensis DSM 44229	NP
NC_020302	Corynebacterium halotolerans YIM 70093 = DSM 44683	NP
NC_020504	Streptomyces davawensis JCM 4913	NP
NC_020506	Corynebacterium callunae DSM 20147	NP
NC_020517	Bifidobacterium breve UCC2003	NP
NC_020519	Corynebacterium glutamicum K051	NP
NC_020520	Ilumatobacter coccineus YM16-304	NP
NC_020546	Bifidobacterium thermophilum RBL67	NP
NC_020895	Streptomyces hygroscopicus subsp. jinggangensis TL01	NP
NC_020990	Streptomyces albus J1074	NP
NC_021008	Bifidobacterium longum subsp. longum F8	NP
NC_002677	Mycobacterium leprae TN	PA
NC_002755	Mycobacterium tuberculosis CDC1551	PA
NC_002945	Mycobacterium bovis AF2122/97	PA
NC_004551	Tropheryma whipplei TW08/27	PA
NC_004572	Tropheryma whipplei str. Twist	PA
NC_006085	Propionibacterium acnes KPA171202	PA
NC_007164	Corynebacterium jeikeium K411	PA
NC_008611	Mycobacterium ulcerans Agy99	PA
NC_008769	Mycobacterium bovis BCG str. Pasteur 1173P2	PA
NC_009525	Mycobacterium tuberculosis H37Ra	PA
NC_009565	Mycobacterium tuberculosis F11	PA
NC_010168	Renibacterium salmoninarum ATCC 33209	PA

NC_010612	Mycobacterium marinum M	PA
NC_011896	Mycobacterium leprae Br4923	PA
NC_012207	Mycobacterium bovis BCG str. Tokyo 172	PA
NC_012704	Corynebacterium kroppenstedtii DSM 44385	PA
NC_012943	Mycobacterium tuberculosis KZN 1435	PA
NC_013159	Saccharomonospora viridis DSM 43017	PA
NC_013174	Jonesia denitrificans DSM 20603	PA
NC_013204	Eggerthella lenta DSM 2243	PA
NC_013721	Gardnerella vaginalis 409-05	PA
NC_014211	Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111	PA
NC_014218	Arcanobacterium haemolyticum DSM 20595	PA
NC_014246	Mobiluncus curtisii ATCC 43063	PA
NC_014329	Corynebacterium pseudotuberculosis FRC41	PA
NC_014363	Olsenella uli DSM 7084	PA
NC_014644	Gardnerella vaginalis ATCC 14019	PA
NC_014659	Rhodococcus equi 103S	PA
NC_015576	Mycobacterium sp. JDM601	PA
NC_015673	Corynebacterium resistens DSM 45100	PA
NC_015683	Corynebacterium ulcerans BR-AD22	PA
NC_015758	Mycobacterium africanum GM041182	PA
NC_015848	Mycobacterium canettii CIPT 140010059	PA
NC_016511	Propionibacterium acnes TypeIA2 P.acn31	PA
NC_016512	Propionibacterium acnes TypeIA2 P.acn17	PA
NC_016516	Propionibacterium acnes TypeIA2 P.acn33	PA
NC_016604	Mycobacterium rhodesiae NBB3	PA
NC_016768	Mycobacterium tuberculosis KZN 4207	PA
NC_016781	Corynebacterium pseudotuberculosis 3/99-5	PA
NC_016783	Corynebacterium diphtheriae INCA 402	PA
NC_016785	Corynebacterium diphtheriae CDCE 8392	PA
NC_016787	Corynebacterium diphtheriae HC03	PA
NC_016788	Corynebacterium diphtheriae HC04	PA
NC_016789	Corynebacterium diphtheriae PW8	PA
NC_016790	Corynebacterium diphtheriae VA01	PA
NC_016799	Corynebacterium diphtheriae 31A	PA
NC_016800	Corynebacterium diphtheriae BH8	PA
NC_016801	Corynebacterium diphtheriae C7 (beta)	PA
NC_016802	Corynebacterium diphtheriae HC02	PA
NC_016804	Mycobacterium bovis BCG str. Mexico	PA
NC_016932	Corynebacterium pseudotuberculosis 316	PA
NC_016934	Mycobacterium tuberculosis UT205	PA
NC_017031	Corynebacterium pseudotuberculosis P54B96	PA
NC_017300	Corynebacterium pseudotuberculosis 1002	PA
NC_017301	Corynebacterium pseudotuberculosis C231	PA
NC_017303	Corynebacterium pseudotuberculosis I19	PA
NC_017305	Corynebacterium pseudotuberculosis PAT10	PA
NC_017306	Corynebacterium pseudotuberculosis 42/02-A	PA
NC_017307	Corynebacterium pseudotuberculosis CIP 52.97	PA
NC_017308	Corynebacterium pseudotuberculosis 1/06-A	PA
NC_017317	Corynebacterium ulcerans 809	PA
NC_017456	Gardnerella vaginalis HMP9231	PA
NC_017462	Corynebacterium pseudotuberculosis 267	PA

NC_017522	Mycobacterium tuberculosis CCDC5180	PA
NC_017524	Mycobacterium tuberculosis CTRI-2	PA
NC_017534	Propionibacterium acnes 266	PA
NC_017535	Propionibacterium acnes 6609	PA
NC_017550	Propionibacterium acnes ATCC 11828	PA
NC_017730	Corynebacterium pseudotuberculosis 31	PA
NC_017904	Mycobacterium sp. MOTT36Y	PA
NC_017945	Corynebacterium pseudotuberculosis 258	PA
NC_018019	Corynebacterium pseudotuberculosis Cp162	PA
NC_018078	Mycobacterium tuberculosis KZN 605	PA
NC_018101	Corynebacterium ulcerans 0102	PA
NC_018142	Propionibacterium propionicum F0230a	PA
NC_018143	Mycobacterium tuberculosis H37Rv	PA
NC_018707	Propionibacterium acnes C1	PA
NC_019950	Mycobacterium canettii CIPT 140060008	PA
NC_019951	Mycobacterium canettii CIPT 140070010	PA
NC_019952	Mycobacterium canettii CIPT 140070017	PA
NC_019965	Mycobacterium canettii CIPT 140070008	PA
NC_020089	Mycobacterium tuberculosis 7199-99	PA
NC_020133	Mycobacterium liflandii 128FXT	PA
NC_020245	Mycobacterium bovis BCG str. Korea 1168P	PA
NC_020559	Mycobacterium tuberculosis str. Erdman = ATCC 35801	PA
NC_021064	Propionibacterium avidum 44067	PA
NC_021085	Propionibacterium acnes HL096PA1	PA

S. TABLE 4 – PFAM RESULTS FOR MOST DISCRIMINANT GENES FOR PATHOGENS.

<seq_id>	<hmm_name>	<type>	<bit_score>	<E-value>	<clust_id>
333918125	Ribosomal_S7	Domain	209.8	1.2e-62	9811
470157906	Ribosomal_S7	Domain	205.6	2.5e-61	9811
406032679	Ribosomal_S7	Domain	210.9	5.6e-63	9811
25027071	Ribosomal_S7	Domain	205.2	3.1e-61	9811
386739682	Ribosomal_S7	Domain	204.2	6.5e-61	9811
451943222	Ribosomal_S7	Domain	201.6	4.2e-60	9811
145294646	Ribosomal_S7	Domain	203.7	9.1e-61	9811
392414802	Ribosomal_S7	Domain	211.8	3.0e-63	9811
470173516	Ribosomal_S7	Domain	203.7	9.1e-61	9811
404216353	Ribosomal_S7	Domain	208.7	2.7e-62	9811
340795359	Ribosomal_S7	Domain	205.0	3.8e-61	9811
378719165	Ribosomal_S7	Domain	209.8	1.2e-62	9811
315446005	Ribosomal_S7	Domain	211.2	4.5e-63	9811
62389392	Ribosomal_S7	Domain	203.7	9.1e-61	9811
336326461	Ribosomal_S7	Domain	202.5	2.3e-60	9811
224989076	Ribosomal_S7	Domain	208.2	3.8e-62	9811
384514925	Ribosomal_S7	Domain	204.2	6.5e-61	9811
387139975	Ribosomal_S7	Domain	204.2	6.5e-61	9811
384508118	Ribosomal_S7	Domain	204.2	6.5e-61	9811
387137939	Ribosomal_S7	Domain	204.2	6.5e-61	9811
15840086	Ribosomal_S7	Domain	208.4	3.3e-62	9811
379714619	Ribosomal_S7	Domain	204.2	6.5e-61	9811
118616548	Ribosomal_S7	Domain	207.2	8.0e-62	9811
121636604	Ribosomal_S7	Domain	208.2	3.8e-62	9811

237786408	Ribosomal_S7	Domain	205.0	3.7e-61	9811
433625773	Ribosomal_S7	Domain	208.2	3.8e-62	9811
376250560	Ribosomal_S7	Domain	204.5	5.4e-61	9811
15828006	Ribosomal_S7	Domain	208.2	4.0e-62	9811
376286932	Ribosomal_S7	Domain	204.5	5.4e-61	9811
433633716	Ribosomal_S7	Domain	208.2	3.8e-62	9811
479054633	Ribosomal_S7	Domain	208.2	3.8e-62	9811
443489504	Ribosomal_S7	Domain	208.3	3.6e-62	9811
68536932	Ribosomal_S7	Domain	206.7	1.1e-61	9811
433640804	Ribosomal_S7	Domain	208.2	3.8e-62	9811
376256375	Ribosomal_S7	Domain	204.5	5.4e-61	9811
397672491	Ribosomal_S7	Domain	208.2	3.8e-62	9811
384503938	Ribosomal_S7	Domain	204.2	6.5e-61	9811
337290003	Ribosomal_S7	Domain	204.2	6.5e-61	9811
471336530	Ribosomal_S7	Domain	208.2	3.8e-62	9811
385990160	Ribosomal_S7	Domain	208.2	3.8e-62	9811
148660458	Ribosomal_S7	Domain	208.2	3.8e-62	9811
392431111	Ribosomal_S7	Domain	208.2	3.8e-62	9811
375294901	Ribosomal_S7	Domain	208.2	3.8e-62	9811
449062703	Ribosomal_S7	Domain	208.2	3.8e-62	9811
376253564	Ribosomal_S7	Domain	204.5	5.4e-61	9811
376247741	Ribosomal_S7	Domain	204.5	5.4e-61	9811
376242114	Ribosomal_S7	Domain	201.8	3.6e-60	9811
148821888	Ribosomal_S7	Domain	208.2	3.8e-62	9811
392399895	Ribosomal_S7	Domain	204.2	6.5e-61	9811
340625702	Ribosomal_S7	Domain	208.2	3.8e-62	9811
333989335	Ribosomal_S7	Domain	207.5	6.2e-62	9811
339630753	Ribosomal_S7	Domain	208.2	3.8e-62	9811
392385403	Ribosomal_S7	Domain	208.2	3.8e-62	9811
376289616	Ribosomal_S7	Domain	204.5	5.4e-61	9811
385806784	Ribosomal_S7	Domain	204.2	6.5e-61	9811
387877790	Ribosomal_S7	Domain	210.9	5.6e-63	9811
387135885	Ribosomal_S7	Domain	204.2	6.5e-61	9811
300857750	Ribosomal_S7	Domain	204.2	6.5e-61	9811
385997462	Ribosomal_S7	Domain	208.2	3.8e-62	9811
384510211	Ribosomal_S7	Domain	204.2	6.5e-61	9811
376283970	Ribosomal_S7	Domain	202.8	1.7e-60	9811
397653182	Ribosomal_S7	Domain	204.2	6.5e-61	9811
183981033	Ribosomal_S7	Domain	208.3	3.6e-62	9811
312140979	Ribosomal_S7	Domain	208.8	2.4e-62	9811
253797625	Ribosomal_S7	Domain	208.2	3.8e-62	9811
375292355	Ribosomal_S7	Domain	204.5	5.4e-61	9811
384506027	Ribosomal_S7	Domain	204.2	6.5e-61	9811
375287916	Ribosomal_S7	Domain	204.2	6.5e-61	9811
31791867	Ribosomal_S7	Domain	208.2	3.8e-62	9811
378770438	Ribosomal_S7	Domain	208.2	3.8e-62	9811
389849685	Ribosomal_S7	Domain	204.2	6.5e-61	9811
433629769	Ribosomal_S7	Domain	208.2	3.8e-62	9811
375137930	Ribosomal_S7	Domain	210.3	8.4e-63	9811
221230483	Ribosomal_S7	Domain	208.2	4.0e-62	9811
383313518	Ribosomal_S7	Domain	204.2	6.5e-61	9811

376292529	Ribosomal_S7	Domain	204.0	7.6e-61	9811
386741089	GMC_oxred_N	Domain	235.1	9.8e-70	31894
386741089	GMC_oxred_C	Domain	126.1	1.2e-36	31894
336325834	GMC_oxred_N	Domain	275.3	5.7e-82	31894
336325834	GMC_oxred_C	Domain	123.7	6.7e-36	31894
384516385	GMC_oxred_N	Domain	275.0	6.8e-82	31894
384516385	GMC_oxred_C	Domain	126.8	7.2e-37	31894
384509561	GMC_oxred_N	Domain	274.3	1.2e-81	31894
384509561	GMC_oxred_C	Domain	125.9	1.4e-36	31894
387141337	GMC_oxred_N	Domain	273.7	1.7e-81	31894
387141337	GMC_oxred_C	Domain	125.9	1.4e-36	31894
387139360	GMC_oxred_N	Domain	273.7	1.7e-81	31894
387139360	GMC_oxred_C	Domain	125.9	1.4e-36	31894
379716077	GMC_oxred_N	Domain	273.7	1.7e-81	31894
379716077	GMC_oxred_C	Domain	125.9	1.4e-36	31894
376252250	GMC_oxred_N	Domain	285.4	4.5e-85	31894
376252250	GMC_oxred_C	Domain	123.5	7.5e-36	31894
376288694	GMC_oxred_N	Domain	285.4	4.6e-85	31894
376288694	GMC_oxred_C	Domain	123.5	7.5e-36	31894
68536266	GMC_oxred_N	Domain	283.5	1.7e-84	31894
68536266	GMC_oxred_C	Domain	127.4	4.9e-37	31894
376258027	GMC_oxred_N	Domain	285.4	4.6e-85	31894
376258027	GMC_oxred_C	Domain	121.3	3.7e-35	31894
384505372	GMC_oxred_N	Domain	274.3	1.2e-81	31894
384505372	GMC_oxred_C	Domain	125.9	1.4e-36	31894
337291617	GMC_oxred_N	Domain	275.0	6.8e-82	31894
337291617	GMC_oxred_C	Domain	126.8	7.2e-37	31894
376249481	GMC_oxred_N	Domain	285.4	4.6e-85	31894
376249481	GMC_oxred_C	Domain	123.5	7.5e-36	31894
376255256	GMC_oxred_N	Domain	286.0	3.0e-85	31894
376255256	GMC_oxred_C	Domain	123.3	8.6e-36	31894
376243785	GMC_oxred_N	Domain	284.8	7.1e-85	31894
376243785	GMC_oxred_C	Domain	124.7	3.3e-36	31894
392401275	GMC_oxred_N	Domain	276.5	2.4e-82	31894
392401275	GMC_oxred_C	Domain	126.0	1.3e-36	31894
376291376	GMC_oxred_N	Domain	285.4	4.6e-85	31894
376291376	GMC_oxred_C	Domain	123.5	7.5e-36	31894
385808261	GMC_oxred_N	Domain	274.3	1.2e-81	31894
385808261	GMC_oxred_C	Domain	125.9	1.4e-36	31894
300859200	GMC_oxred_N	Domain	274.3	1.2e-81	31894
300859200	GMC_oxred_C	Domain	125.9	1.4e-36	31894
387137295	GMC_oxred_N	Domain	274.3	1.2e-81	31894
387137295	GMC_oxred_C	Domain	125.9	1.4e-36	31894
384511646	GMC_oxred_N	Domain	274.3	1.2e-81	31894
384511646	GMC_oxred_C	Domain	125.9	1.4e-36	31894
376285705	GMC_oxred_N	Domain	285.4	4.6e-85	31894
376285705	GMC_oxred_C	Domain	123.5	7.5e-36	31894
397654755	GMC_oxred_N	Domain	275.0	6.8e-82	31894
397654755	GMC_oxred_C	Domain	126.8	7.2e-37	31894
375294017	GMC_oxred_N	Domain	285.4	4.6e-85	31894
375294017	GMC_oxred_C	Domain	123.5	7.5e-36	31894

384507464	GMC_oxred_N	Domain	274.3	1.2e-81	31894
384507464	GMC_oxred_C	Domain	125.9	1.4e-36	31894
375289391	GMC_oxred_N	Domain	274.3	1.2e-81	31894
375289391	GMC_oxred_C	Domain	125.9	1.4e-36	31894
389851126	GMC_oxred_N	Domain	273.7	1.7e-81	31894
389851126	GMC_oxred_C	Domain	125.9	1.4e-36	31894
376294190	GMC_oxred_N	Domain	285.4	4.6e-85	31894
376294190	GMC_oxred_C	Domain	123.5	7.5e-36	31894
383314956	GMC_oxred_N	Domain	274.3	1.2e-81	31894
383314956	GMC_oxred_C	Domain	125.9	1.4e-36	31894
119869359	MraZ	Family	75.4	2.4e-21	149120
119869359	MraZ	Family	76.7	8.9e-22	149120
406030406	MraZ	Family	75.7	1.9e-21	149120
406030406	MraZ	Family	81.3	3.3e-23	149120
226360239	MraZ	Family	72.4	2.0e-20	149120
226360239	MraZ	Family	69.5	1.7e-19	149120
111018110	MraZ	Family	72.7	1.6e-20	149120
111018110	MraZ	Family	69.6	1.5e-19	149120
108800231	MraZ	Family	75.4	2.4e-21	149120
108800231	MraZ	Family	76.7	8.9e-22	149120
126435854	MraZ	Family	75.4	2.4e-21	149120
126435854	MraZ	Family	76.7	8.9e-22	149120
224990542	MraZ	Family	73.0	1.3e-20	149120
224990542	MraZ	Family	77.5	5.2e-22	149120
15841658	MraZ	Family	73.0	1.3e-20	149120
15841658	MraZ	Family	77.5	5.2e-22	149120
118618805	MraZ	Family	73.5	9.3e-21	149120
118618805	MraZ	Family	79.3	1.4e-22	149120
121638048	MraZ	Family	73.0	1.3e-20	149120
121638048	MraZ	Family	77.5	5.2e-22	149120
433627284	MraZ	Family	74.7	3.8e-21	149120
433627284	MraZ	Family	77.5	5.2e-22	149120
15827425	MraZ	Family	75.2	2.7e-21	149120
15827425	MraZ	Family	81.9	2.1e-23	149120
479056129	MraZ	Family	73.0	1.3e-20	149120
479056129	MraZ	Family	77.5	5.2e-22	149120
433635235	MraZ	Family	74.7	3.8e-21	149120
433635235	MraZ	Family	77.5	5.2e-22	149120
443491475	MraZ	Family	73.5	9.3e-21	149120
443491475	MraZ	Family	79.3	1.4e-22	149120
433642347	MraZ	Family	74.7	3.8e-21	149120
433642347	MraZ	Family	77.5	5.2e-22	149120
397674050	MraZ	Family	73.0	1.3e-20	149120
397674050	MraZ	Family	77.5	5.2e-22	149120
471338141	MraZ	Family	73.0	1.3e-20	149120
471338141	MraZ	Family	77.5	5.2e-22	149120
385991511	MraZ	Family	73.0	1.3e-20	149120
385991511	MraZ	Family	77.5	5.2e-22	149120
148661982	MraZ	Family	73.0	1.3e-20	149120
148661982	MraZ	Family	77.5	5.2e-22	149120
392432236	MraZ	Family	73.0	1.3e-20	149120

392432236	MraZ	Family	77.5	5.2e-22	149120
375296027	MraZ	Family	73.0	1.3e-20	149120
375296027	MraZ	Family	77.5	5.2e-22	149120
449064224	MraZ	Family	73.0	1.3e-20	149120
449064224	MraZ	Family	77.5	5.2e-22	149120
148823375	MraZ	Family	73.0	1.3e-20	149120
148823375	MraZ	Family	77.5	5.2e-22	149120
340627174	MraZ	Family	74.7	3.8e-21	149120
340627174	MraZ	Family	77.5	5.2e-22	149120
333990374	MraZ	Family	72.6	1.8e-20	149120
333990374	MraZ	Family	75.3	2.5e-21	149120
339632197	MraZ	Family	73.0	1.3e-20	149120
339632197	MraZ	Family	77.5	5.2e-22	149120
392386812	MraZ	Family	73.0	1.3e-20	149120
392386812	MraZ	Family	77.5	5.2e-22	149120
387875555	MraZ	Family	75.7	1.9e-21	149120
387875555	MraZ	Family	81.3	3.3e-23	149120
385998943	MraZ	Family	73.0	1.3e-20	149120
385998943	MraZ	Family	77.5	5.2e-22	149120
183983193	MraZ	Family	70.6	7.1e-20	149120
183983193	MraZ	Family	79.1	1.6e-22	149120
312140151	MraZ	Family	70.3	9.3e-20	149120
312140151	MraZ	Family	73.0	1.3e-20	149120
253798769	MraZ	Family	73.0	1.3e-20	149120
253798769	MraZ	Family	77.5	5.2e-22	149120
31793346	MraZ	Family	73.0	1.3e-20	149120
31793346	MraZ	Family	77.5	5.2e-22	149120
378771897	MraZ	Family	73.0	1.3e-20	149120
378771897	MraZ	Family	77.5	5.2e-22	149120
433631286	MraZ	Family	74.7	3.8e-21	149120
433631286	MraZ	Family	77.5	5.2e-22	149120
375141598	MraZ	Family	76.8	8.2e-22	149120
375141598	MraZ	Family	74.5	4.6e-21	149120
221229902	MraZ	Family	75.2	2.7e-21	149120
221229902	MraZ	Family	81.9	2.1e-23	149120
50843306	Ribosomal_L5	Domain	91.8	2.2e-26	274546
50843306	Ribosomal_L5_C	Domain	131.5	8.5e-39	274546
365963496	Ribosomal_L5	Domain	91.8	2.2e-26	274546
365963496	Ribosomal_L5_C	Domain	133.1	2.8e-39	274546
386070048	Ribosomal_L5	Domain	91.8	2.2e-26	274546
386070048	Ribosomal_L5_C	Domain	133.1	2.8e-39	274546
386024788	Ribosomal_L5	Domain	91.8	2.2e-26	274546
386024788	Ribosomal_L5_C	Domain	133.1	2.8e-39	274546
480328572	Ribosomal_L5	Domain	92.3	1.6e-26	274546
480328572	Ribosomal_L5_C	Domain	131.8	7.0e-39	274546
387504216	Ribosomal_L5	Domain	91.8	2.2e-26	274546
387504216	Ribosomal_L5_C	Domain	131.5	8.5e-39	274546
365965740	Ribosomal_L5	Domain	91.8	2.2e-26	274546
365965740	Ribosomal_L5_C	Domain	133.1	2.8e-39	274546
365974675	Ribosomal_L5	Domain	91.8	2.2e-26	274546
365974675	Ribosomal_L5_C	Domain	133.1	2.8e-39	274546

482890919	Ribosomal_L5	Domain	91.8	2.2e-26	274546
482890919	Ribosomal_L5_C	Domain	133.1	2.8e-39	274546
407936232	Ribosomal_L5	Domain	91.8	2.2e-26	274546
407936232	Ribosomal_L5_C	Domain	133.1	2.8e-39	274546
397670909	Ribosomal_L5	Domain	96.5	7.5e-28	274546
397670909	Ribosomal_L5_C	Domain	130.7	1.5e-38	274546
298345568	ABC_tran	Domain	101.6	4.4e-29	281756
311115247	ABC_tran	Domain	101.7	4.0e-29	281756
283782666	ABC_tran	Domain	101.8	3.7e-29	281756
385801131	ABC_tran	Domain	101.7	4.0e-29	281756

S. TABLE 5 – PFAM RESULTS FOR MOST DISCRIMINANT GENES FOR NON-PATHOGENS.

<seq_id>	**<hmm_name>**	**<type>**	**<bit_score>**	**<E-value>**	**<clust_id>**
296453271	Adenylsucc_synt	Domain	594.7	8.5e-179	1025
213691044	Adenylsucc_synt	Domain	596.4	2.5e-179	1025
23465134	Adenylsucc_synt	Domain	595.1	6.3e-179	1025
241190191	Adenylsucc_synt	Domain	591.5	7.9e-178	1025
311063586	Adenylsucc_synt	Domain	597.2	1.5e-179	1025
219682616	Adenylsucc_synt	Domain	591.5	7.9e-178	1025
322690287	Adenylsucc_synt	Domain	594.7	8.1e-179	1025
387821709	Adenylsucc_synt	Domain	591.5	7.9e-178	1025
310286699	Adenylsucc_synt	Domain	596.7	2.0e-179	1025
189440209	Adenylsucc_synt	Domain	595.7	4.3e-179	1025
470202095	Adenylsucc_synt	Domain	592.4	4.1e-178	1025
384196284	Adenylsucc_synt	Domain	596.5	2.3e-179	1025
322688272	Adenylsucc_synt	Domain	595.7	4.3e-179	1025
384190408	Adenylsucc_synt	Domain	591.5	7.9e-178	1025
479136465	Adenylsucc_synt	Domain	596.4	2.5e-179	1025
408501857	Adenylsucc_synt	Domain	593.9	1.4e-178	1025
384193190	Adenylsucc_synt	Domain	591.5	7.9e-178	1025
387820054	Adenylsucc_synt	Domain	591.5	7.9e-178	1025
384191544	Adenylsucc_synt	Domain	591.5	7.9e-178	1025
390936050	Adenylsucc_synt	Domain	596.7	2.0e-179	1025
241195597	Adenylsucc_synt	Domain	591.5	7.9e-178	1025
384202383	Adenylsucc_synt	Domain	596.0	3.5e-179	1025
384194747	Adenylsucc_synt	Domain	591.5	7.9e-178	1025
476417248	Adenylsucc_synt	Domain	596.6	2.2e-179	1025
119025085	Adenylsucc_synt	Domain	596.6	2.3e-179	1025
311113943	Adenylsucc_synt	Domain	592.4	4.2e-178	1025
283782628	Adenylsucc_synt	Domain	592.9	2.9e-178	1025
385802236	Adenylsucc_synt	Domain	592.5	3.8e-178	1025
374987306	Thiolase_N	Domain	205.9	6.3e-61	1565
374987306	Thiolase_C	Domain	158.0	7.1e-47	1565
357399689	Thiolase_N	Domain	205.3	9.6e-61	1565
357399689	Thiolase_C	Domain	163.5	1.4e-48	1565
474982847	Thiolase_N	Domain	201.0	1.9e-59	1565
474982847	Thiolase_C	Domain	163.4	1.5e-48	1565
379734280	Thiolase_N	Domain	205.3	9.5e-61	1565
379734280	Thiolase_C	Domain	156.2	2.4e-46	1565
345015724	Thiolase_N	Domain	209.3	5.7e-62	1565
345015724	Thiolase_C	Domain	158.7	4.1e-47	1565

433602606	Thiolase_N	Domain	197.1	2.9e-58	1565
433602606	Thiolase_C	Domain	156.6	1.9e-46	1565
256389852	Thiolase_N	Domain	204.9	1.3e-60	1565
256389852	Thiolase_C	Domain	160.1	1.5e-47	1565
408678551	Thiolase_N	Domain	205.8	6.7e-61	1565
408678551	Thiolase_C	Domain	161.6	5.2e-48	1565
357391619	Thiolase_N	Domain	207.1	2.6e-61	1565
357391619	Thiolase_C	Domain	158.4	5.1e-47	1565
296271076	Thiolase_N	Domain	208.5	9.6e-62	1565
296271076	Thiolase_C	Domain	161.1	7.7e-48	1565
119716062	Thiolase_N	Domain	214.5	1.4e-63	1565
119716062	Thiolase_C	Domain	161.3	6.7e-48	1565
119715607	Thiolase_N	Domain	204.5	1.7e-60	1565
119715607	Thiolase_C	Domain	147.9	9.3e-44	1565
182438243	Thiolase_N	Domain	209.5	5.0e-62	1565
182438243	Thiolase_C	Domain	161.5	5.6e-48	1565
72160840	Thiolase_N	Domain	216.7	3.2e-64	1565
72160840	Thiolase_C	Domain	148.8	4.7e-44	1565
29830055	Thiolase_N	Domain	207.4	2.2e-61	1565
29830055	Thiolase_C	Domain	163.3	1.5e-48	1565
399541802	Thiolase_N	Domain	190.8	2.5e-56	1565
399541802	Thiolase_C	Domain	161.4	6.2e-48	1565
300789922	Thiolase_N	Domain	190.8	2.5e-56	1565
300789922	Thiolase_C	Domain	161.4	6.2e-48	1565
383775945	Thiolase_N	Domain	197.5	2.2e-58	1565
383775945	Thiolase_C	Domain	160.8	9.7e-48	1565
386845900	Thiolase_N	Domain	199.6	5.3e-59	1565
386845900	Thiolase_C	Domain	155.1	5.6e-46	1565
357413021	Thiolase_N	Domain	206.9	3.1e-61	1565
357413021	Thiolase_C	Domain	162.5	2.7e-48	1565
404216924	Thiolase_N	Domain	200.3	3.2e-59	1565
404216924	Thiolase_C	Domain	150.6	1.3e-44	1565
478688558	Thiolase_N	Domain	206.9	3.1e-61	1565
478688558	Thiolase_C	Domain	162.0	3.9e-48	1565
471324675	Thiolase_N	Domain	207.3	2.3e-61	1565
471324675	Thiolase_C	Domain	163.1	1.8e-48	1565
386840638	Thiolase_N	Domain	201.0	1.9e-59	1565
386840638	Thiolase_C	Domain	163.4	1.5e-48	1565
284989508	Thiolase_N	Domain	207.6	1.8e-61	1565
284989508	Thiolase_C	Domain	157.9	7.3e-47	1565
331694798	Thiolase_N	Domain	203.6	3.2e-60	1565
331694798	Thiolase_C	Domain	152.8	2.8e-45	1565
257057158	Thiolase_N	Domain	207.4	2.2e-61	1565
257057158	Thiolase_C	Domain	164.6	6.2e-49	1565
333920017	HTH_18	Domain	76.0	1.8e-21	1704
336178627	HTH_18	Domain	79.7	1.3e-22	1704
134100241	HTH_18	Domain	76.2	1.6e-21	1704
433608317	HTH_18	Domain	78.2	3.9e-22	1704
256392357	HTH_18	Domain	76.6	1.2e-21	1704
406030684	HTH_18	Domain	79.8	1.2e-22	1704
408679220	HTH_18	Domain	75.4	2.8e-21	1704

284041959	HTH_18	Domain	77.2	8.1e-22	1704
315505936	HTH_18	Domain	79.7	1.3e-22	1704
330466671	HTH_18	Domain	78.3	3.6e-22	1704
378717943	HTH_18	Domain	82.6	1.6e-23	1704
386846991	HTH_18	Domain	81.1	4.8e-23	1704
226308449	HTH_18	Domain	78.4	3.3e-22	1704
336320851	HTH_18	Domain	74.5	5.5e-21	1704
336118453	HTH_18	Domain	78.9	2.3e-22	1704
284029129	HTH_18	Domain	76.7	1.1e-21	1704
357412230	HTH_18	Domain	82.4	1.9e-23	1704
404215004	HTH_18	Domain	77.6	5.7e-22	1704
120403990	HTH_18	Domain	79.8	1.2e-22	1704
258653074	HTH_18	Domain	80.7	6.2e-23	1704
226363116	HTH_18	Domain	74.7	4.6e-21	1704
302867661	HTH_18	Domain	79.7	1.3e-22	1704
331697072	HTH_18	Domain	80.0	1.0e-22	1704
332668815	HTH_18	Domain	74.7	4.8e-21	1704
387875907	HTH_18	Domain	79.9	1.1e-22	1704
239918173	RNA_pol_L	Domain	65.3	2.2e-18	1851
239918173	RNA_pol_A_bac	Domain	82.5	2.4e-23	1851
239918173	RNA_pol_A_CTD	Domain	95.4	1.1e-27	1851
291298742	RNA_pol_L	Domain	66.0	1.4e-18	1851
291298742	RNA_pol_A_bac	Domain	79.0	2.8e-22	1851
291298742	RNA_pol_A_CTD	Domain	87.3	3.9e-25	1851
374988973	RNA_pol_L	Domain	68.4	2.4e-19	1851
374988973	RNA_pol_A_bac	Domain	84.5	5.8e-24	1851
374988973	RNA_pol_A_CTD	Domain	89.9	5.9e-26	1851
374992568	RNA_pol_L	Domain	65.2	2.4e-18	1851
374992568	RNA_pol_A_bac	Domain	78.0	6.0e-22	1851
374992568	RNA_pol_A_CTD	Domain	88.1	2.1e-25	1851
116671486	RNA_pol_L	Domain	66.9	6.8e-19	1851
116671486	RNA_pol_A_bac	Domain	83.9	9.2e-24	1851
116671486	RNA_pol_A_CTD	Domain	91.9	1.4e-26	1851
269955462	RNA_pol_L	Domain	69.1	1.4e-19	1851
269955462	RNA_pol_A_bac	Domain	90.0	1.2e-25	1851
269955462	RNA_pol_A_CTD	Domain	90.0	5.6e-26	1851
257069480	RNA_pol_L	Domain	68.5	2.3e-19	1851
257069480	RNA_pol_A_bac	Domain	85.2	3.5e-24	1851
257069480	RNA_pol_A_CTD	Domain	86.0	9.6e-25	1851
357401195	RNA_pol_L	Domain	68.4	2.4e-19	1851
357401195	RNA_pol_A_bac	Domain	86.3	1.6e-24	1851
357401195	RNA_pol_A_CTD	Domain	89.9	5.9e-26	1851
325964132	RNA_pol_L	Domain	66.9	6.8e-19	1851
325964132	RNA_pol_A_bac	Domain	82.3	2.8e-23	1851
325964132	RNA_pol_A_CTD	Domain	91.9	1.4e-26	1851
159039805	RNA_pol_L	Domain	62.7	1.4e-17	1851
159039805	RNA_pol_A_bac	Domain	82.1	3.3e-23	1851
159039805	RNA_pol_A_CTD	Domain	88.3	1.8e-25	1851
403528104	RNA_pol_L	Domain	66.9	7.1e-19	1851
403528104	RNA_pol_A_bac	Domain	83.9	9.2e-24	1851
403528104	RNA_pol_A_CTD	Domain	91.9	1.4e-26	1851

474984117	RNA_pol_L	Domain	68.4	2.4e-19	1851
474984117	RNA_pol_A_bac	Domain	84.5	5.8e-24	1851
474984117	RNA_pol_A_CTD	Domain	89.9	5.9e-26	1851
336179754	RNA_pol_L	Domain	67.4	4.8e-19	1851
336179754	RNA_pol_A_bac	Domain	79.6	2.0e-22	1851
336179754	RNA_pol_A_CTD	Domain	92.6	8.1e-27	1851
379737614	RNA_pol_L	Domain	62.7	1.4e-17	1851
379737614	RNA_pol_A_bac	Domain	84.1	7.7e-24	1851
379737614	RNA_pol_A_CTD	Domain	89.2	9.7e-26	1851
345008598	RNA_pol_L	Domain	68.4	2.4e-19	1851
345008598	RNA_pol_A_bac	Domain	84.3	6.7e-24	1851
345008598	RNA_pol_A_CTD	Domain	89.9	5.9e-26	1851
433609629	RNA_pol_L	Domain	69.3	1.2e-19	1851
433609629	RNA_pol_A_bac	Domain	89.8	1.3e-25	1851
433609629	RNA_pol_A_CTD	Domain	90.9	2.8e-26	1851
119867181	RNA_pol_L	Domain	72.9	9.7e-21	1851
119867181	RNA_pol_A_bac	Domain	89.8	1.3e-25	1851
119867181	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
256390173	RNA_pol_L	Domain	70.6	4.9e-20	1851
256390173	RNA_pol_A_bac	Domain	86.8	1.1e-24	1851
256390173	RNA_pol_A_CTD	Domain	90.8	3.0e-26	1851
406032527	RNA_pol_L	Domain	71.9	1.9e-20	1851
406032527	RNA_pol_A_bac	Domain	90.0	1.2e-25	1851
406032527	RNA_pol_A_CTD	Domain	92.5	8.7e-27	1851
220913400	RNA_pol_L	Domain	66.9	6.8e-19	1851
220913400	RNA_pol_A_bac	Domain	82.3	2.8e-23	1851
220913400	RNA_pol_A_CTD	Domain	91.9	1.4e-26	1851
108798085	RNA_pol_L	Domain	72.9	9.7e-21	1851
108798085	RNA_pol_A_bac	Domain	89.8	1.3e-25	1851
108798085	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
408680125	RNA_pol_L	Domain	68.4	2.4e-19	1851
408680125	RNA_pol_A_bac	Domain	84.5	5.8e-24	1851
408680125	RNA_pol_A_CTD	Domain	89.9	5.9e-26	1851
357390060	RNA_pol_L	Domain	67.0	6.5e-19	1851
357390060	RNA_pol_A_bac	Domain	84.7	4.9e-24	1851
357390060	RNA_pol_A_CTD	Domain	92.2	1.1e-26	1851
126433745	RNA_pol_L	Domain	72.9	9.7e-21	1851
126433745	RNA_pol_A_bac	Domain	89.8	1.3e-25	1851
126433745	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
296130462	RNA_pol_L	Domain	66.1	1.2e-18	1851
296130462	RNA_pol_A_bac	Domain	87.9	5.2e-25	1851
296130462	RNA_pol_A_CTD	Domain	87.7	2.8e-25	1851
315501403	RNA_pol_L	Domain	63.9	5.9e-18	1851
315501403	RNA_pol_A_bac	Domain	82.1	3.3e-23	1851
315501403	RNA_pol_A_CTD	Domain	88.3	1.8e-25	1851
296268581	RNA_pol_L	Domain	70.6	4.8e-20	1851
296268581	RNA_pol_A_bac	Domain	92.9	1.4e-26	1851
296268581	RNA_pol_A_CTD	Domain	94.5	2.2e-27	1851
119718095	RNA_pol_L	Domain	66.2	1.2e-18	1851
119718095	RNA_pol_A_bac	Domain	85.8	2.2e-24	1851
119718095	RNA_pol_A_CTD	Domain	89.7	6.8e-26	1851

330470144	RNA_pol_L	Domain	63.9	5.9e-18	1851
330470144	RNA_pol_A_bac	Domain	82.1	3.3e-23	1851
330470144	RNA_pol_A_CTD	Domain	88.3	1.8e-25	1851
152964684	RNA_pol_L	Domain	67.8	3.5e-19	1851
152964684	RNA_pol_A_bac	Domain	86.8	1.1e-24	1851
152964684	RNA_pol_A_CTD	Domain	89.2	9.9e-26	1851
182436600	RNA_pol_L	Domain	68.4	2.4e-19	1851
182436600	RNA_pol_A_bac	Domain	84.5	5.8e-24	1851
182436600	RNA_pol_A_CTD	Domain	89.9	5.9e-26	1851
72163017	RNA_pol_L	Domain	69.3	1.2e-19	1851
72163017	RNA_pol_A_bac	Domain	89.0	2.4e-25	1851
72163017	RNA_pol_A_CTD	Domain	93.6	4.2e-27	1851
29826981	RNA_pol_L	Domain	68.4	2.3e-19	1851
29826981	RNA_pol_A_bac	Domain	84.3	6.6e-24	1851
29826981	RNA_pol_A_CTD	Domain	87.9	2.5e-25	1851
29831496	RNA_pol_L	Domain	68.4	2.4e-19	1851
29831496	RNA_pol_A_bac	Domain	84.5	5.8e-24	1851
29831496	RNA_pol_A_CTD	Domain	89.9	5.9e-26	1851
315445898	RNA_pol_L	Domain	72.5	1.2e-20	1851
315445898	RNA_pol_A_bac	Domain	91.4	4.1e-26	1851
315445898	RNA_pol_A_CTD	Domain	91.1	2.5e-26	1851
399534525	RNA_pol_L	Domain	71.2	3.1e-20	1851
399534525	RNA_pol_A_bac	Domain	89.3	1.9e-25	1851
399534525	RNA_pol_A_CTD	Domain	89.9	6.0e-26	1851
269796227	RNA_pol_L	Domain	70.2	6.4e-20	1851
269796227	RNA_pol_A_bac	Domain	90.3	8.8e-26	1851
269796227	RNA_pol_A_CTD	Domain	87.2	4.0e-25	1851
334336213	RNA_pol_L	Domain	69.5	1.1e-19	1851
334336213	RNA_pol_A_bac	Domain	91.6	3.7e-26	1851
334336213	RNA_pol_A_CTD	Domain	89.9	5.7e-26	1851
300782639	RNA_pol_L	Domain	71.2	3.1e-20	1851
300782639	RNA_pol_A_bac	Domain	89.3	1.9e-25	1851
300782639	RNA_pol_A_CTD	Domain	89.9	6.0e-26	1851
383775802	RNA_pol_L	Domain	66.0	1.4e-18	1851
383775802	RNA_pol_A_bac	Domain	81.8	4.1e-23	1851
383775802	RNA_pol_A_CTD	Domain	88.3	1.8e-25	1851
184200289	RNA_pol_L	Domain	67.4	5.0e-19	1851
184200289	RNA_pol_A_bac	Domain	83.8	9.4e-24	1851
184200289	RNA_pol_A_CTD	Domain	90.0	5.5e-26	1851
386845763	RNA_pol_L	Domain	66.0	1.4e-18	1851
386845763	RNA_pol_A_bac	Domain	81.8	4.0e-23	1851
386845763	RNA_pol_A_CTD	Domain	88.3	1.8e-25	1851
336319962	RNA_pol_L	Domain	69.5	1.1e-19	1851
336319962	RNA_pol_A_bac	Domain	87.2	8.5e-25	1851
336319962	RNA_pol_A_CTD	Domain	90.0	5.4e-26	1851
336116819	RNA_pol_L	Domain	66.9	7.2e-19	1851
336116819	RNA_pol_A_bac	Domain	84.7	4.9e-24	1851
336116819	RNA_pol_A_CTD	Domain	89.8	6.0e-26	1851
357411706	RNA_pol_L	Domain	68.4	2.4e-19	1851
357411706	RNA_pol_A_bac	Domain	84.5	5.8e-24	1851
357411706	RNA_pol_A_CTD	Domain	89.9	5.9e-26	1851

284033971	RNA_pol_L	Domain	63.3	9.7e-18	1851
284033971	RNA_pol_A_bac	Domain	84.8	4.6e-24	1851
284033971	RNA_pol_A_CTD	Domain	89.8	6.3e-26	1851
271962664	RNA_pol_L	Domain	67.5	4.4e-19	1851
271962664	RNA_pol_A_bac	Domain	90.8	6.4e-26	1851
271962664	RNA_pol_A_CTD	Domain	94.5	2.1e-27	1851
471323173	RNA_pol_L	Domain	68.4	2.4e-19	1851
471323173	RNA_pol_A_bac	Domain	84.5	5.8e-24	1851
471323173	RNA_pol_A_CTD	Domain	89.9	5.9e-26	1851
478690284	RNA_pol_L	Domain	68.4	2.4e-19	1851
478690284	RNA_pol_A_bac	Domain	84.6	5.2e-24	1851
478690284	RNA_pol_A_CTD	Domain	89.9	5.9e-26	1851
471327910	RNA_pol_L	Domain	68.5	2.2e-19	1851
471327910	RNA_pol_A_bac	Domain	84.1	7.5e-24	1851
471327910	RNA_pol_A_CTD	Domain	87.2	4.0e-25	1851
308178106	RNA_pol_L	Domain	66.3	1.1e-18	1851
308178106	RNA_pol_A_bac	Domain	78.7	3.7e-22	1851
308178106	RNA_pol_A_CTD	Domain	89.6	7.5e-26	1851
117927543	RNA_pol_L	Domain	69.4	1.2e-19	1851
117927543	RNA_pol_A_bac	Domain	89.2	2.0e-25	1851
117927543	RNA_pol_A_CTD	Domain	93.8	3.6e-27	1851
386841906	RNA_pol_L	Domain	68.4	2.4e-19	1851
386841906	RNA_pol_A_bac	Domain	84.5	5.8e-24	1851
386841906	RNA_pol_A_CTD	Domain	89.9	5.9e-26	1851
312199992	RNA_pol_L	Domain	64.2	4.8e-18	1851
312199992	RNA_pol_A_bac	Domain	84.8	4.7e-24	1851
312199992	RNA_pol_A_CTD	Domain	92.5	8.9e-27	1851
145596406	RNA_pol_L	Domain	62.6	1.6e-17	1851
145596406	RNA_pol_A_bac	Domain	82.0	3.5e-23	1851
145596406	RNA_pol_A_CTD	Domain	90.5	3.8e-26	1851
119961601	RNA_pol_L	Domain	66.9	7.1e-19	1851
119961601	RNA_pol_A_bac	Domain	83.9	9.2e-24	1851
119961601	RNA_pol_A_CTD	Domain	91.9	1.4e-26	1851
284992846	RNA_pol_L	Domain	63.8	6.5e-18	1851
284992846	RNA_pol_A_bac	Domain	85.4	3.0e-24	1851
284992846	RNA_pol_A_CTD	Domain	89.2	9.7e-26	1851
392414925	RNA_pol_L	Domain	72.9	9.7e-21	1851
392414925	RNA_pol_A_bac	Domain	91.4	4.0e-26	1851
392414925	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
302869942	RNA_pol_L	Domain	63.9	5.9e-18	1851
302869942	RNA_pol_A_bac	Domain	82.1	3.3e-23	1851
302869942	RNA_pol_A_CTD	Domain	88.3	1.8e-25	1851
331699145	RNA_pol_L	Domain	70.9	3.8e-20	1851
331699145	RNA_pol_A_bac	Domain	90.4	8.3e-26	1851
331699145	RNA_pol_A_CTD	Domain	89.6	7.3e-26	1851
317125840	RNA_pol_L	Domain	66.7	7.9e-19	1851
317125840	RNA_pol_A_bac	Domain	90.9	6.0e-26	1851
317125840	RNA_pol_A_CTD	Domain	89.5	7.9e-26	1851
229821593	RNA_pol_L	Domain	67.3	5.1e-19	1851
229821593	RNA_pol_A_bac	Domain	88.9	2.4e-25	1851
229821593	RNA_pol_A_CTD	Domain	87.6	3.1e-25	1851

332669545	RNA_pol_L	Domain	69.5	1.1e-19	1851
332669545	RNA_pol_A_bac	Domain	89.8	1.3e-25	1851
332669545	RNA_pol_A_CTD	Domain	90.0	5.4e-26	1851
224991873	RNA_pol_L	Domain	72.6	1.2e-20	1851
224991873	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
224991873	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
118616625	RNA_pol_L	Domain	72.9	9.6e-21	1851
118616625	RNA_pol_A_bac	Domain	87.6	6.4e-25	1851
118616625	RNA_pol_A_CTD	Domain	92.5	8.7e-27	1851
15843052	RNA_pol_L	Domain	72.6	1.2e-20	1851
15843052	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
15843052	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
121639377	RNA_pol_L	Domain	72.6	1.2e-20	1851
121639377	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
121639377	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
433628599	RNA_pol_L	Domain	72.6	1.2e-20	1851
433628599	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
433628599	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
433636548	RNA_pol_L	Domain	72.6	1.2e-20	1851
433636548	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
433636548	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
479057436	RNA_pol_L	Domain	72.6	1.2e-20	1851
479057436	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
479057436	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
256831853	RNA_pol_L	Domain	67.3	5.2e-19	1851
256831853	RNA_pol_A_bac	Domain	88.8	2.6e-25	1851
256831853	RNA_pol_A_CTD	Domain	89.5	7.9e-26	1851
443489581	RNA_pol_L	Domain	70.1	7.2e-20	1851
443489581	RNA_pol_A_bac	Domain	87.6	6.4e-25	1851
443489581	RNA_pol_A_CTD	Domain	92.5	8.7e-27	1851
397675411	RNA_pol_L	Domain	72.6	1.2e-20	1851
397675411	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
397675411	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
433643652	RNA_pol_L	Domain	72.6	1.2e-20	1851
433643652	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
433643652	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
163840858	RNA_pol_L	Domain	66.1	1.2e-18	1851
163840858	RNA_pol_A_bac	Domain	83.8	9.2e-24	1851
163840858	RNA_pol_A_CTD	Domain	91.9	1.4e-26	1851
471339522	RNA_pol_L	Domain	72.6	1.2e-20	1851
471339522	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
471339522	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
385992690	RNA_pol_L	Domain	72.6	1.2e-20	1851
385992690	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
385992690	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
148663322	RNA_pol_L	Domain	72.6	1.2e-20	1851
148663322	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
148663322	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
392433940	RNA_pol_L	Domain	72.6	1.2e-20	1851
392433940	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
392433940	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851

375297728	RNA_pol_L	Domain	72.6	1.2e-20	1851
375297728	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
375297728	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
449065568	RNA_pol_L	Domain	72.6	1.2e-20	1851
449065568	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
449065568	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
148824667	RNA_pol_L	Domain	72.6	1.2e-20	1851
148824667	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
148824667	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
340628428	RNA_pol_L	Domain	72.6	1.2e-20	1851
340628428	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
340628428	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
339633462	RNA_pol_L	Domain	72.6	1.2e-20	1851
339633462	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
339633462	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
392388060	RNA_pol_L	Domain	72.6	1.2e-20	1851
392388060	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
392388060	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
387877654	RNA_pol_L	Domain	71.9	1.9e-20	1851
387877654	RNA_pol_A_bac	Domain	90.0	1.2e-25	1851
387877654	RNA_pol_A_CTD	Domain	92.5	8.7e-27	1851
386000246	RNA_pol_L	Domain	72.6	1.2e-20	1851
386000246	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
386000246	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
183981110	RNA_pol_L	Domain	72.9	9.6e-21	1851
183981110	RNA_pol_A_bac	Domain	87.6	6.4e-25	1851
183981110	RNA_pol_A_CTD	Domain	92.5	8.7e-27	1851
253800502	RNA_pol_L	Domain	72.6	1.2e-20	1851
253800502	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
253800502	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
31794633	RNA_pol_L	Domain	72.6	1.2e-20	1851
31794633	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
31794633	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
378773238	RNA_pol_L	Domain	72.6	1.2e-20	1851
378773238	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
378773238	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
297564065	RNA_pol_L	Domain	67.0	6.6e-19	1851
297564065	RNA_pol_A_bac	Domain	88.3	3.8e-25	1851
297564065	RNA_pol_A_CTD	Domain	90.6	3.5e-26	1851
433632555	RNA_pol_L	Domain	72.6	1.2e-20	1851
433632555	RNA_pol_A_bac	Domain	89.7	1.4e-25	1851
433632555	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
375137829	RNA_pol_L	Domain	70.8	4.2e-20	1851
375137829	RNA_pol_A_bac	Domain	88.7	2.9e-25	1851
375137829	RNA_pol_A_CTD	Domain	91.1	2.4e-26	1851
374990428	ABC_tran	Domain	32.4	1.1e-07	4318
291300697	ABC_tran	Domain	48.0	1.6e-12	4318
116670746	ABC_tran	Domain	47.2	2.7e-12	4318
333920015	ABC_tran	Domain	46.8	3.6e-12	4318
325961542	ABC_tran	Domain	44.9	1.4e-11	4318
159038059	ABC_tran	Domain	39.4	7.1e-10	4318

403527406	ABC_tran	Domain	49.8	4.3e-13	4318
134100239	ABC_tran	Domain	42.7	6.9e-11	4318
474983189	ABC_tran	Domain	47.7	1.9e-12	4318
336178629	ABC_tran	Domain	40.2	4.1e-10	4318
379737926	ABC_tran	Domain	42.3	9.4e-11	4318
345013203	ABC_tran	Domain	45.1	1.3e-11	4318
433608319	ABC_tran	Domain	40.4	3.4e-10	4318
119867824	ABC_tran	Domain	36.5	5.5e-09	4318
256377796	ABC_tran	Domain	46.3	5.3e-12	4318
256392355	ABC_tran	Domain	49.9	4.0e-13	4318
220910891	ABC_tran	Domain	49.8	4.4e-13	4318
108798706	ABC_tran	Domain	36.5	5.5e-09	4318
408676145	ABC_tran	Domain	46.9	3.6e-12	4318
357388821	ABC_tran	Domain	41.5	1.6e-10	4318
284041961	ABC_tran	Domain	44.9	1.4e-11	4318
111020821	ABC_tran	Domain	45.2	1.1e-11	4318
126434307	ABC_tran	Domain	38.1	1.8e-09	4318
296131032	ABC_tran	Domain	37.9	2.1e-09	4318
315505933	ABC_tran	Domain	41.4	1.8e-10	4318
296270317	ABC_tran	Domain	49.6	5.2e-13	4318
119716904	ABC_tran	Domain	44.9	1.4e-11	4318
330466672	ABC_tran	Domain	45.1	1.3e-11	4318
152964028	ABC_tran	Domain	41.0	2.3e-10	4318
182438566	ABC_tran	Domain	41.8	1.3e-10	4318
72161961	ABC_tran	Domain	30.2	5.0e-07	4318
29828639	ABC_tran	Domain	40.4	3.4e-10	4318
399536321	ABC_tran	Domain	47.2	2.8e-12	4318
378717945	ABC_tran	Domain	48.4	1.1e-12	4318
269795145	ABC_tran	Domain	41.4	1.7e-10	4318
300784436	ABC_tran	Domain	47.2	2.8e-12	4318
383778869	ABC_tran	Domain	43.3	4.3e-11	4318
386851861	ABC_tran	Domain	39.8	5.3e-10	4318
226308447	ABC_tran	Domain	40.1	4.4e-10	4318
336321840	ABC_tran	Domain	46.5	4.7e-12	4318
336118451	ABC_tran	Domain	49.1	7.0e-13	4318
284029131	ABC_tran	Domain	40.2	3.9e-10	4318
357415145	ABC_tran	Domain	39.2	8.1e-10	4318
404215006	ABC_tran	Domain	51.2	1.7e-13	4318
271969099	ABC_tran	Domain	39.7	5.6e-10	4318
478689408	ABC_tran	Domain	39.1	8.6e-10	4318
471320288	ABC_tran	Domain	45.7	8.1e-12	4318
120403988	ABC_tran	Domain	39.0	9.7e-10	4318
258653071	ABC_tran	Domain	42.5	7.9e-11	4318
226363118	ABC_tran	Domain	48.3	1.3e-12	4318
386840978	ABC_tran	Domain	47.7	1.9e-12	4318
312194453	ABC_tran	Domain	43.5	3.8e-11	4318
145594864	ABC_tran	Domain	43.8	3.2e-11	4318
119960502	ABC_tran	Domain	48.9	8.5e-13	4318
284990237	ABC_tran	Domain	43.2	4.9e-11	4318
302867664	ABC_tran	Domain	41.2	1.9e-10	4318
331697070	ABC_tran	Domain	42.7	6.6e-11	4318

317125230	ABC_tran	Domain	41.6	1.4e-10	4318
332668813	ABC_tran	Domain	37.8	2.3e-09	4318
163840291	ABC_tran	Domain	49.1	6.9e-13	4318
257056818	ABC_tran	Domain	37.5	2.7e-09	4318
312139790	ABC_tran	Domain	42.9	6.0e-11	4318
375142066	ABC_tran	Domain	41.6	1.5e-10	4318
296454750	ABC_tran	Domain	89.7	2.1e-25	8006
296454750	ABC_tran_Xtn	Domain	53.3	2.0e-14	8006
296454750	ABC_tran	Domain	72.6	4.1e-20	8006
213691460	ABC_tran	Domain	89.3	2.8e-25	8006
213691460	ABC_tran_Xtn	Domain	53.3	2.0e-14	8006
213691460	ABC_tran	Domain	72.6	4.1e-20	8006
23466236	ABC_tran	Domain	89.3	2.8e-25	8006
23466236	ABC_tran_Xtn	Domain	53.3	2.0e-14	8006
23466236	ABC_tran	Domain	72.7	3.8e-20	8006
241190547	ABC_tran	Domain	91.6	5.6e-26	8006
241190547	ABC_tran_Xtn	Domain	51.6	6.6e-14	8006
241190547	ABC_tran	Domain	69.1	4.8e-19	8006
311064783	ABC_tran	Domain	89.1	3.2e-25	8006
311064783	ABC_tran_Xtn	Domain	53.4	1.8e-14	8006
311064783	ABC_tran	Domain	72.2	5.3e-20	8006
219682970	ABC_tran	Domain	91.5	5.8e-26	8006
219682970	ABC_tran_Xtn	Domain	51.5	6.8e-14	8006
219682970	ABC_tran	Domain	69.1	4.9e-19	8006
322691762	ABC_tran	Domain	89.3	2.8e-25	8006
322691762	ABC_tran_Xtn	Domain	53.3	2.0e-14	8006
322691762	ABC_tran	Domain	72.6	4.1e-20	8006
387822082	ABC_tran	Domain	91.6	5.6e-26	8006
387822082	ABC_tran_Xtn	Domain	51.6	6.6e-14	8006
387822082	ABC_tran	Domain	69.1	4.8e-19	8006
310287902	ABC_tran	Domain	89.1	3.2e-25	8006
310287902	ABC_tran_Xtn	Domain	53.4	1.8e-14	8006
310287902	ABC_tran	Domain	72.2	5.3e-20	8006
189440717	ABC_tran	Domain	89.0	3.4e-25	8006
189440717	ABC_tran_Xtn	Domain	53.3	2.0e-14	8006
189440717	ABC_tran	Domain	72.7	3.8e-20	8006
470203325	ABC_tran	Domain	91.2	7.0e-26	8006
470203325	ABC_tran_Xtn	Domain	52.7	2.9e-14	8006
470203325	ABC_tran	Domain	73.7	1.8e-20	8006
384197585	ABC_tran	Domain	89.7	2.1e-25	8006
384197585	ABC_tran_Xtn	Domain	53.3	2.0e-14	8006
384197585	ABC_tran	Domain	72.6	4.1e-20	8006
257064916	ABC_tran	Domain	93.9	1.0e-26	8006
257064916	ABC_tran_Xtn	Domain	46.8	2.0e-12	8006
257064916	ABC_tran	Domain	69.7	3.0e-19	8006
322689823	ABC_tran	Domain	89.3	2.8e-25	8006
322689823	ABC_tran_Xtn	Domain	53.3	2.0e-14	8006
322689823	ABC_tran	Domain	72.6	4.1e-20	8006
384190790	ABC_tran	Domain	91.6	5.6e-26	8006
384190790	ABC_tran_Xtn	Domain	51.6	6.6e-14	8006
384190790	ABC_tran	Domain	69.1	4.8e-19	8006

479135161	ABC_tran	Domain	89.3	2.8e-25	8006
479135161	ABC_tran_Xtn	Domain	53.3	2.0e-14	8006
479135161	ABC_tran	Domain	72.7	3.8e-20	8006
408501459	ABC_tran	Domain	92.3	3.3e-26	8006
408501459	ABC_tran_Xtn	Domain	51.3	7.8e-14	8006
408501459	ABC_tran	Domain	68.8	5.8e-19	8006
452892199	ABC_tran	Domain	91.6	5.6e-26	8006
452892199	ABC_tran_Xtn	Domain	51.6	6.6e-14	8006
452892199	ABC_tran	Domain	69.1	4.8e-19	8006
387820414	ABC_tran	Domain	91.6	5.6e-26	8006
387820414	ABC_tran_Xtn	Domain	51.6	6.6e-14	8006
387820414	ABC_tran	Domain	69.1	4.8e-19	8006
339445184	ABC_tran	Domain	92.3	3.3e-26	8006
339445184	ABC_tran_Xtn	Domain	52.5	3.4e-14	8006
339445184	ABC_tran	Domain	72.5	4.3e-20	8006
384191935	ABC_tran	Domain	91.6	5.6e-26	8006
384191935	ABC_tran_Xtn	Domain	51.6	6.6e-14	8006
384191935	ABC_tran	Domain	69.1	4.8e-19	8006
390937324	ABC_tran	Domain	89.1	3.2e-25	8006
390937324	ABC_tran_Xtn	Domain	53.4	1.8e-14	8006
390937324	ABC_tran	Domain	72.2	5.3e-20	8006
241195953	ABC_tran	Domain	91.6	5.6e-26	8006
241195953	ABC_tran_Xtn	Domain	51.6	6.6e-14	8006
241195953	ABC_tran	Domain	69.1	4.8e-19	8006
384200896	ABC_tran	Domain	89.7	2.1e-25	8006
384200896	ABC_tran_Xtn	Domain	53.3	2.0e-14	8006
384200896	ABC_tran	Domain	72.6	4.1e-20	8006
384195103	ABC_tran	Domain	91.6	5.6e-26	8006
384195103	ABC_tran_Xtn	Domain	51.6	6.6e-14	8006
384195103	ABC_tran	Domain	69.1	4.8e-19	8006
476418587	ABC_tran	Domain	89.7	2.1e-25	8006
476418587	ABC_tran_Xtn	Domain	53.3	2.0e-14	8006
476418587	ABC_tran	Domain	72.6	4.1e-20	8006
257784105	ABC_tran	Domain	97.0	1.1e-27	8006
257784105	ABC_tran_Xtn	Domain	52.8	2.7e-14	8006
257784105	ABC_tran	Domain	72.3	4.9e-20	8006
119025433	ABC_tran	Domain	88.7	4.1e-25	8006
119025433	ABC_tran_Xtn	Domain	52.4	3.7e-14	8006
119025433	ABC_tran	Domain	71.6	8.2e-20	8006
311114329	ABC_tran	Domain	89.9	1.9e-25	8006
311114329	ABC_tran_Xtn	Domain	50.9	1.0e-13	8006
311114329	ABC_tran	Domain	68.8	6.1e-19	8006
283783574	ABC_tran	Domain	91.3	6.7e-26	8006
283783574	ABC_tran_Xtn	Domain	51.5	6.8e-14	8006
283783574	ABC_tran	Domain	69.0	5.0e-19	8006
257791194	ABC_tran	Domain	91.2	7.0e-26	8006
257791194	ABC_tran_Xtn	Domain	49.5	2.8e-13	8006
257791194	ABC_tran	Domain	70.5	1.7e-19	8006
385802024	ABC_tran	Domain	89.8	1.9e-25	8006
385802024	ABC_tran_Xtn	Domain	50.9	1.1e-13	8006
385802024	ABC_tran	Domain	68.8	6.1e-19	8006

239917317	ThiC-associated	Domain	29.8	3.4e-07	12433
239917317	ThiC_Rad_SAM	Domain	661.6	3.8e-199	12433
116671033	ThiC-associated	Domain	34.5	1.2e-08	12433
116671033	ThiC_Rad_SAM	Domain	668.7	2.6e-201	12433
269956254	ThiC-associated	Domain	38.5	6.6e-10	12433
269956254	ThiC_Rad_SAM	Domain	665.9	1.9e-200	12433
325963718	ThiC-associated	Domain	28.8	7.3e-07	12433
325963718	ThiC_Rad_SAM	Domain	668.7	2.6e-201	12433
403527656	ThiC-associated	Domain	36.8	2.4e-09	12433
403527656	ThiC_Rad_SAM	Domain	669.9	1.2e-201	12433
220912972	ThiC-associated	Domain	31.8	8.1e-08	12433
220912972	ThiC_Rad_SAM	Domain	671.8	2.9e-202	12433
340794956	ThiC-associated	Domain	31.6	9.9e-08	12433
340794956	ThiC_Rad_SAM	Domain	663.2	1.2e-199	12433
410867847	ThiC-associated	Domain	27.9	1.4e-06	12433
410867847	ThiC_Rad_SAM	Domain	666.9	9.5e-201	12433
334336150	ThiC-associated	Domain	33.2	3.0e-08	12433
334336150	ThiC_Rad_SAM	Domain	673.4	1.0e-202	12433
308178182	ThiC-associated	Domain	34.4	1.3e-08	12433
308178182	ThiC_Rad_SAM	Domain	670.2	9.1e-202	12433
119960761	ThiC-associated	Domain	36.7	2.5e-09	12433
119960761	ThiC_Rad_SAM	Domain	669.9	1.2e-201	12433
296454464	IGPD	Family	183.4	2.5e-54	13007
213691720	IGPD	Family	184.5	1.1e-54	13007
23465859	IGPD	Family	184.5	1.1e-54	13007
241191247	IGPD	Family	186.2	3.3e-55	13007
311063937	IGPD	Family	186.2	3.4e-55	13007
219683286	IGPD	Family	186.2	3.3e-55	13007
322691490	IGPD	Family	184.5	1.1e-54	13007
387822796	IGPD	Family	186.2	3.3e-55	13007
310287072	IGPD	Family	186.6	2.4e-55	13007
189439017	IGPD	Family	184.6	1.0e-54	13007
470202582	IGPD	Family	185.0	8.0e-55	13007
384196725	IGPD	Family	183.7	2.0e-54	13007
322689533	IGPD	Family	184.5	1.1e-54	13007
384189870	IGPD	Family	186.2	3.3e-55	13007
479135455	IGPD	Family	184.6	1.1e-54	13007
408500751	IGPD	Family	188.4	7.0e-56	13007
384194244	IGPD	Family	186.2	3.3e-55	13007
387821116	IGPD	Family	186.2	3.3e-55	13007
384192660	IGPD	Family	186.2	3.3e-55	13007
390936421	IGPD	Family	186.6	2.4e-55	13007
241196653	IGPD	Family	186.2	3.3e-55	13007
384201213	IGPD	Family	184.5	1.1e-54	13007
384195809	IGPD	Family	186.2	3.3e-55	13007
476418386	IGPD	Family	183.7	2.0e-54	13007
119026145	IGPD	Family	186.1	3.5e-55	13007
291299505	GcpE	Family	425.7	1.2e-127	14351
374986168	GcpE	Family	444.9	1.7e-133	14351
291302569	GcpE	Family	437.8	2.5e-131	14351
357402054	GcpE	Family	444.1	3.1e-133	14351

357402512	GcpE	Family	444.1	3.1e-133	14351
159036943	GcpE	Family	438.8	1.2e-131	14351
474985948	GcpE	Family	430.1	5.7e-129	14351
336179477	GcpE	Family	443.2	5.5e-133	14351
474985076	GcpE	Family	457.5	2.6e-137	14351
379737111	GcpE	Family	442.6	8.7e-133	14351
345009683	GcpE	Family	444.6	2.2e-133	14351
345010495	GcpE	Family	442.5	9.5e-133	14351
433607840	GcpE	Family	438.8	1.2e-131	14351
433608801	GcpE	Family	444.7	2.0e-133	14351
256379883	GcpE	Family	443.2	5.7e-133	14351
256395545	GcpE	Family	443.0	6.5e-133	14351
256396861	GcpE	Family	451.6	1.6e-135	14351
408681068	GcpE	Family	445.6	1.0e-133	14351
357392190	GcpE	Family	448.6	1.3e-134	14351
315502480	GcpE	Family	445.7	9.7e-134	14351
296269024	GcpE	Family	452.4	8.9e-136	14351
119717426	GcpE	Family	445.2	1.4e-133	14351
330466322	GcpE	Family	445.1	1.6e-133	14351
182435614	GcpE	Family	443.7	4.1e-133	14351
29828189	GcpE	Family	450.8	2.8e-135	14351
29829103	GcpE	Family	453.4	4.6e-136	14351
399535841	GcpE	Family	454.3	2.4e-136	14351
378717535	GcpE	Family	448.2	1.7e-134	14351
300783956	GcpE	Family	454.3	2.4e-136	14351
383782164	GcpE	Family	441.8	1.6e-132	14351
386852182	GcpE	Family	444.2	2.9e-133	14351
336321187	GcpE	Family	437.0	4.5e-131	14351
336117207	GcpE	Family	442.4	1.0e-132	14351
357410869	GcpE	Family	442.3	1.1e-132	14351
271963500	GcpE	Family	447.6	2.7e-134	14351
271965181	GcpE	Family	426.5	6.9e-128	14351
478687749	GcpE	Family	444.7	1.9e-133	14351
471322212	GcpE	Family	450.2	4.1e-135	14351
471321070	GcpE	Family	443.2	5.8e-133	14351
258652395	GcpE	Family	442.3	1.0e-132	14351
117928729	GcpE	Family	445.0	1.6e-133	14351
386842867	GcpE	Family	457.5	2.6e-137	14351
386843739	GcpE	Family	430.1	5.7e-129	14351
145593900	GcpE	Family	438.6	1.4e-131	14351
302865924	GcpE	Family	443.6	4.5e-133	14351
332670038	GcpE	Family	437.4	3.3e-131	14351
256832245	GcpE	Family	438.2	2.0e-131	14351
239918205	Ribosomal_S7	Domain	200.9	6.7e-60	23225
291298706	Ribosomal_S7	Domain	201.8	3.5e-60	23225
374988924	Ribosomal_S7	Domain	199.0	2.6e-59	23225
116671527	Ribosomal_S7	Domain	199.7	1.6e-59	23225
269955431	Ribosomal_S7	Domain	207.9	4.9e-62	23225
257069516	Ribosomal_S7	Domain	202.3	2.6e-60	23225
357401145	Ribosomal_S7	Domain	201.0	6.3e-60	23225
325964171	Ribosomal_S7	Domain	201.2	5.5e-60	23225

159039838	Ribosomal_S7	Domain	198.9	2.7e-59	23225
403528139	Ribosomal_S7	Domain	201.0	6.2e-60	23225
134103268	Ribosomal_S7	Domain	199.7	1.6e-59	23225
474984083	Ribosomal_S7	Domain	200.5	9.1e-60	23225
336179786	Ribosomal_S7	Domain	200.3	1.1e-59	23225
379737646	Ribosomal_S7	Domain	199.5	1.9e-59	23225
345008561	Ribosomal_S7	Domain	200.0	1.3e-59	23225
433609664	Ribosomal_S7	Domain	199.6	1.8e-59	23225
119867055	Ribosomal_S7	Domain	211.1	4.8e-63	23225
323357407	Ribosomal_S7	Domain	204.2	6.4e-61	23225
256380602	Ribosomal_S7	Domain	199.2	2.3e-59	23225
256390131	Ribosomal_S7	Domain	195.8	2.5e-58	23225
220913439	Ribosomal_S7	Domain	201.2	5.5e-60	23225
108797958	Ribosomal_S7	Domain	211.1	4.8e-63	23225
408680090	Ribosomal_S7	Domain	201.9	3.3e-60	23225
357390110	Ribosomal_S7	Domain	199.6	1.8e-59	23225
111018920	Ribosomal_S7	Domain	211.4	4.0e-63	23225
126433621	Ribosomal_S7	Domain	211.1	4.8e-63	23225
296130499	Ribosomal_S7	Domain	207.3	7.5e-62	23225
315501437	Ribosomal_S7	Domain	196.9	1.2e-58	23225
296268546	Ribosomal_S7	Domain	208.0	4.5e-62	23225
119718144	Ribosomal_S7	Domain	199.8	1.5e-59	23225
330470177	Ribosomal_S7	Domain	196.9	1.2e-58	23225
152964652	Ribosomal_S7	Domain	204.5	5.5e-61	23225
182436638	Ribosomal_S7	Domain	199.8	1.5e-59	23225
72163049	Ribosomal_S7	Domain	207.7	5.6e-62	23225
29831461	Ribosomal_S7	Domain	199.7	1.6e-59	23225
399534490	Ribosomal_S7	Domain	199.3	2.1e-59	23225
410867106	Ribosomal_S7	Domain	204.9	3.9e-61	23225
269796261	Ribosomal_S7	Domain	206.7	1.1e-61	23225
334336182	Ribosomal_S7	Domain	207.8	5.2e-62	23225
300782604	Ribosomal_S7	Domain	199.3	2.1e-59	23225
383775769	Ribosomal_S7	Domain	191.1	7.3e-57	23225
184200258	Ribosomal_S7	Domain	204.5	5.2e-61	23225
386845730	Ribosomal_S7	Domain	194.2	8.2e-58	23225
226305251	Ribosomal_S7	Domain	210.8	6.3e-63	23225
336319920	Ribosomal_S7	Domain	209.4	1.6e-62	23225
336116738	Ribosomal_S7	Domain	203.6	9.7e-61	23225
284034006	Ribosomal_S7	Domain	198.3	4.3e-59	23225
357411742	Ribosomal_S7	Domain	199.7	1.6e-59	23225
271962628	Ribosomal_S7	Domain	202.3	2.5e-60	23225
478690249	Ribosomal_S7	Domain	200.6	8.3e-60	23225
471323209	Ribosomal_S7	Domain	199.7	1.6e-59	23225
120402291	Ribosomal_S7	Domain	212.0	2.6e-63	23225
258651356	Ribosomal_S7	Domain	197.1	1.0e-58	23225
226361020	Ribosomal_S7	Domain	212.1	2.4e-63	23225
308178141	Ribosomal_S7	Domain	198.2	4.8e-59	23225
117927511	Ribosomal_S7	Domain	201.2	5.5e-60	23225
386841873	Ribosomal_S7	Domain	200.5	9.1e-60	23225
312200024	Ribosomal_S7	Domain	197.2	9.5e-59	23225
145596438	Ribosomal_S7	Domain	195.8	2.6e-58	23225

297625773	Ribosomal_S7	Domain	205.7	2.2e-61	23225
119961568	Ribosomal_S7	Domain	201.0	6.2e-60	23225
284992878	Ribosomal_S7	Domain	199.7	1.6e-59	23225
302869976	Ribosomal_S7	Domain	196.9	1.2e-58	23225
331699180	Ribosomal_S7	Domain	198.4	4.2e-59	23225
317125877	Ribosomal_S7	Domain	204.3	6.1e-61	23225
229821627	Ribosomal_S7	Domain	207.5	6.1e-62	23225
332669512	Ribosomal_S7	Domain	209.8	1.2e-62	23225
298345484	Ribosomal_S7	Domain	209.8	1.2e-62	23225
256831823	Ribosomal_S7	Domain	207.5	6.3e-62	23225
163840914	Ribosomal_S7	Domain	200.2	1.1e-59	23225
257054466	Ribosomal_S7	Domain	204.4	5.6e-61	23225
397670953	Ribosomal_S7	Domain	203.4	1.2e-60	23225
297571939	Ribosomal_S7	Domain	205.3	2.9e-61	23225
297564033	Ribosomal_S7	Domain	203.9	8.1e-61	23225
239917361	RNase_PH	Domain	87.2	1.2e-24	28316
239917361	RNase_PH_C	Domain	40.3	2.1e-10	28316
116671127	RNase_PH	Domain	82.6	3.1e-23	28316
116671127	RNase_PH_C	Domain	39.5	3.8e-10	28316
269955862	RNase_PH	Domain	82.1	4.2e-23	28316
269955862	RNase_PH_C	Domain	33.0	4.2e-08	28316
325963801	RNase_PH	Domain	83.3	1.9e-23	28316
325963801	RNase_PH_C	Domain	37.8	1.3e-09	28316
403527765	RNase_PH	Domain	82.7	3.0e-23	28316
403527765	RNase_PH_C	Domain	38.7	7.0e-10	28316
220913058	RNase_PH	Domain	83.8	1.3e-23	28316
220913058	RNase_PH_C	Domain	39.5	3.8e-10	28316
184200648	RNase_PH	Domain	79.9	2.1e-22	28316
184200648	RNase_PH_C	Domain	39.3	4.3e-10	28316
308177068	RNase_PH	Domain	83.4	1.8e-23	28316
308177068	RNase_PH_C	Domain	40.5	1.8e-10	28316
119962182	RNase_PH	Domain	82.7	3.0e-23	28316
119962182	RNase_PH_C	Domain	38.7	7.0e-10	28316
163840218	RNase_PH	Domain	79.3	3.3e-22	28316
163840218	RNase_PH_C	Domain	35.0	9.5e-09	28316
333919998	GTP_CH_N	Family	264.3	5.6e-79	29574
333919998	GTP_cyclohydro2	Family	57.5	1.1e-15	29574
333919998	DUF1688	Family	451.6	2.0e-135	29574
119871306	GTP_CH_N	Family	264.6	4.5e-79	29574
119871306	GTP_cyclohydro2	Family	55.7	4.2e-15	29574
108802153	GTP_CH_N	Family	264.6	4.5e-79	29574
108802153	GTP_cyclohydro2	Family	55.7	4.2e-15	29574
111017280	GTP_CH_N	Family	261.5	3.8e-78	29574
111017280	GTP_cyclohydro2	Family	57.5	1.1e-15	29574
126438134	GTP_CH_N	Family	264.6	4.5e-79	29574
126438134	GTP_cyclohydro2	Family	55.7	4.2e-15	29574
315444173	GTP_CH_N	Family	261.1	5.4e-78	29574
315444173	GTP_cyclohydro2	Family	55.9	3.6e-15	29574
404215200	GTP_CH_N	Family	263.6	8.8e-79	29574
404215200	GTP_cyclohydro2	Family	52.9	2.9e-14	29574
120403956	GTP_CH_N	Family	261.8	3.1e-78	29574

120403956	GTP_cyclohydro2	Family	54.4	1.0e-14	29574
226359803	GTP_CH_N	Family	262.3	2.3e-78	29574
226359803	GTP_cyclohydro2	Family	57.7	9.9e-16	29574
392416736	GTP_CH_N	Family	258.8	2.6e-77	29574
392416736	GTP_cyclohydro2	Family	57.5	1.1e-15	29574
374988138	Ribosomal_L34	Family	70.1	9.6e-20	35107
239918807	Ribosomal_L34	Family	71.1	4.9e-20	35107
291303917	Ribosomal_L34	Family	66.6	1.3e-18	35107
116672714	Ribosomal_L34	Family	70.3	8.4e-20	35107
470159997	Ribosomal_L34	Family	69.1	2.0e-19	35107
333922236	Ribosomal_L34	Family	70.5	7.3e-20	35107
357400703	Ribosomal_L34	Family	72.3	2.0e-20	35107
159040591	Ribosomal_L34	Family	71.2	4.4e-20	35107
269958147	Ribosomal_L34	Family	70.6	6.9e-20	35107
134103817	Ribosomal_L34	Family	62.5	2.4e-17	35107
474983485	Ribosomal_L34	Family	72.3	2.0e-20	35107
336176144	Ribosomal_L34	Family	72.0	2.6e-20	35107
379738371	Ribosomal_L34	Family	69.9	1.2e-19	35107
433610265	Ribosomal_L34	Family	72.2	2.3e-20	35107
119866065	Ribosomal_L34	Family	62.5	2.4e-17	35107
323357957	Ribosomal_L34	Family	75.7	1.7e-21	35107
256381076	Ribosomal_L34	Family	72.0	2.5e-20	35107
108802372	Ribosomal_L34	Family	62.5	2.4e-17	35107
357390829	Ribosomal_L34	Family	74.5	4.2e-21	35107
111020654	Ribosomal_L34	Family	68.9	2.3e-19	35107
126438352	Ribosomal_L34	Family	62.5	2.4e-17	35107
296131547	Ribosomal_L34	Family	72.2	2.2e-20	35107
315506968	Ribosomal_L34	Family	72.8	1.4e-20	35107
296271544	Ribosomal_L34	Family	74.3	4.9e-21	35107
330470832	Ribosomal_L34	Family	74.5	4.2e-21	35107
152968452	Ribosomal_L34	Family	69.9	1.1e-19	35107
182437492	Ribosomal_L34	Family	70.7	6.6e-20	35107
72163516	Ribosomal_L34	Family	75.0	2.9e-21	35107
29830858	Ribosomal_L34	Family	70.4	8.0e-20	35107
315446826	Ribosomal_L34	Family	67.6	6.0e-19	35107
399543042	Ribosomal_L34	Family	68.5	3.1e-19	35107
470176031	Ribosomal_L34	Family	69.1	2.0e-19	35107
25029503	Ribosomal_L34	Family	70.4	8.3e-20	35107
378720501	Ribosomal_L34	Family	69.2	1.9e-19	35107
269797068	Ribosomal_L34	Family	71.8	3.0e-20	35107
334338441	Ribosomal_L34	Family	72.4	1.9e-20	35107
300791165	Ribosomal_L34	Family	68.5	3.1e-19	35107
383783294	Ribosomal_L34	Family	73.7	7.4e-21	35107
184202003	Ribosomal_L34	Family	74.1	5.6e-21	35107
386853316	Ribosomal_L34	Family	72.5	1.8e-20	35107
226309519	Ribosomal_L34	Family	68.9	2.4e-19	35107
336322298	Ribosomal_L34	Family	72.2	2.2e-20	35107
336120989	Ribosomal_L34	Family	73.1	1.2e-20	35107
284034942	Ribosomal_L34	Family	67.4	7.0e-19	35107
357412380	Ribosomal_L34	Family	71.3	4.1e-20	35107
404217323	Ribosomal_L34	Family	66.8	1.1e-18	35107

62391947	Ribosomal_L34	Family	69.1	2.0e-19	35107
271970554	Ribosomal_L34	Family	71.6	3.3e-20	35107
478689544	Ribosomal_L34	Family	70.3	8.8e-20	35107
120407007	Ribosomal_L34	Family	67.6	6.0e-19	35107
258655511	Ribosomal_L34	Family	72.6	1.7e-20	35107
226362894	Ribosomal_L34	Family	68.9	2.3e-19	35107
145297093	Ribosomal_L34	Family	69.1	2.0e-19	35107
308179190	Ribosomal_L34	Family	66.3	1.5e-18	35107
117929368	Ribosomal_L34	Family	73.1	1.2e-20	35107
386841273	Ribosomal_L34	Family	72.3	2.0e-20	35107
312200980	Ribosomal_L34	Family	68.2	3.8e-19	35107
145597102	Ribosomal_L34	Family	71.9	2.7e-20	35107
297627573	Ribosomal_L34	Family	78.8	1.9e-22	35107
119962693	Ribosomal_L34	Family	70.1	1.0e-19	35107
284993439	Ribosomal_L34	Family	66.3	1.5e-18	35107
392419086	Ribosomal_L34	Family	66.3	1.6e-18	35107
302870730	Ribosomal_L34	Family	72.8	1.4e-20	35107
331700394	Ribosomal_L34	Family	72.3	2.1e-20	35107
317126740	Ribosomal_L34	Family	72.6	1.7e-20	35107
229822699	Ribosomal_L34	Family	71.6	3.5e-20	35107
332672293	Ribosomal_L34	Family	72.6	1.7e-20	35107
336326732	Ribosomal_L34	Family	70.7	6.3e-20	35107
224992330	Ribosomal_L34	Family	65.1	3.6e-18	35107
15843558	Ribosomal_L34	Family	65.1	3.6e-18	35107
118620079	Ribosomal_L34	Family	61.9	3.6e-17	35107
121635910	Ribosomal_L34	Family	65.1	3.6e-18	35107
121639835	Ribosomal_L34	Family	65.1	3.6e-18	35107
433629069	Ribosomal_L34	Family	65.1	3.6e-18	35107
15828470	Ribosomal_L34	Family	65.3	3.1e-18	35107
433637021	Ribosomal_L34	Family	65.1	3.6e-18	35107
479057926	Ribosomal_L34	Family	65.1	3.6e-18	35107
256833766	Ribosomal_L34	Family	74.7	3.6e-21	35107
433644114	Ribosomal_L34	Family	65.1	3.6e-18	35107
397675889	Ribosomal_L34	Family	65.1	3.6e-18	35107
163842276	Ribosomal_L34	Family	70.8	5.9e-20	35107
257057916	Ribosomal_L34	Family	70.6	6.9e-20	35107
148663791	Ribosomal_L34	Family	65.1	3.6e-18	35107
392434410	Ribosomal_L34	Family	65.1	3.6e-18	35107
375298196	Ribosomal_L34	Family	65.1	3.6e-18	35107
449066046	Ribosomal_L34	Family	65.1	3.6e-18	35107
148825132	Ribosomal_L34	Family	65.1	3.6e-18	35107
340628894	Ribosomal_L34	Family	65.1	3.6e-18	35107
333992986	Ribosomal_L34	Family	62.5	2.4e-17	35107
339633913	Ribosomal_L34	Family	65.1	3.6e-18	35107
392388517	Ribosomal_L34	Family	65.1	3.6e-18	35107
387878538	Ribosomal_L34	Family	63.4	1.2e-17	35107
386000716	Ribosomal_L34	Family	65.1	3.6e-18	35107
183985458	Ribosomal_L34	Family	61.9	3.6e-17	35107
253800974	Ribosomal_L34	Family	65.1	3.6e-18	35107
31795097	Ribosomal_L34	Family	65.1	3.6e-18	35107
378773695	Ribosomal_L34	Family	65.1	3.6e-18	35107

433633011	Ribosomal_L34	Family	65.6	2.5e-18	35107
221230947	Ribosomal_L34	Family	65.3	3.1e-18	35107
375138995	Ribosomal_L34	Family	66.0	1.9e-18	35107

S. TABLE 6 – PFAM RESULTS FOR MOST DISCRIMINANT GENES FOR AEROBES.

<seq_id>	<hmm_name>	<type>	<bit_score>	<E-value>	<clust_id>
108798706	ABC_tran	Domain	36.5	5.5e-09	2456
111020821	ABC_tran	Domain	45.2	1.1e-11	2456
258653071	ABC_tran	Domain	42.5	7.9e-11	2456
126434307	ABC_tran	Domain	38.1	1.8e-09	2456
226363118	ABC_tran	Domain	48.3	1.3e-12	2456
315505933	ABC_tran	Domain	41.4	1.8e-10	2456
296270317	ABC_tran	Domain	49.6	5.2e-13	2456
325961542	ABC_tran	Domain	44.9	1.4e-11	2456
116670746	ABC_tran	Domain	47.2	2.7e-12	2456
284041961	ABC_tran	Domain	44.9	1.4e-11	2456
119716904	ABC_tran	Domain	44.9	1.4e-11	2456
291300697	ABC_tran	Domain	48.0	1.6e-12	2456
120403988	ABC_tran	Domain	39.0	9.7e-10	2456
271969099	ABC_tran	Domain	39.7	5.6e-10	2456
152964028	ABC_tran	Domain	41.0	2.3e-10	2456
145594864	ABC_tran	Domain	43.8	3.2e-11	2456
134100239	ABC_tran	Domain	42.7	6.9e-11	2456
336178629	ABC_tran	Domain	40.2	4.1e-10	2456
159038059	ABC_tran	Domain	39.4	7.1e-10	2456
182438566	ABC_tran	Domain	41.8	1.3e-10	2456
119960502	ABC_tran	Domain	48.9	8.5e-13	2456
345013203	ABC_tran	Domain	45.1	1.3e-11	2456
312194453	ABC_tran	Domain	43.5	3.8e-11	2456
284990237	ABC_tran	Domain	43.2	4.9e-11	2456
29828639	ABC_tran	Domain	40.4	3.4e-10	2456
302867664	ABC_tran	Domain	41.2	1.9e-10	2456
119867824	ABC_tran	Domain	36.5	5.5e-09	2456
331697070	ABC_tran	Domain	42.7	6.6e-11	2456
256377796	ABC_tran	Domain	46.3	5.3e-12	2456
317125230	ABC_tran	Domain	41.6	1.4e-10	2456
256392355	ABC_tran	Domain	49.9	4.0e-13	2456
226308447	ABC_tran	Domain	40.1	4.4e-10	2456
220910891	ABC_tran	Domain	49.8	4.4e-13	2456
284029131	ABC_tran	Domain	40.2	3.9e-10	2456
336118451	ABC_tran	Domain	49.1	7.0e-13	2456
257056818	ABC_tran	Domain	37.5	2.7e-09	2456
72161961	ABC_tran	Domain	30.2	5.0e-07	2456
72162591	zf-C4_ClpX	Domain	72.7	1.4e-20	2456
72162591	AAA_2	Domain	171.1	2.0e-50	2456
72162591	ClpB_D2-small	Domain	62.7	2.3e-17	2456
239918173	RNA_pol_L	Domain	65.3	2.2e-18	4030
239918173	RNA_pol_A_bac	Domain	82.5	2.4e-23	4030
239918173	RNA_pol_A_CTD	Domain	95.4	1.1e-27	4030
271962664	RNA_pol_L	Domain	67.5	4.4e-19	4030
271962664	RNA_pol_A_bac	Domain	90.8	6.4e-26	4030

271962664	RNA_pol_A_CTD	Domain	94.5	2.1e-27	4030
224991873	RNA_pol_L	Domain	72.6	1.2e-20	4030
224991873	RNA_pol_A_bac	Domain	89.7	1.4e-25	4030
224991873	RNA_pol_A_CTD	Domain	91.1	2.4e-26	4030
118616625	RNA_pol_L	Domain	72.9	9.6e-21	4030
118616625	RNA_pol_A_bac	Domain	87.6	6.4e-25	4030
118616625	RNA_pol_A_CTD	Domain	92.5	8.7e-27	4030
269955462	RNA_pol_L	Domain	69.1	1.4e-19	4030
269955462	RNA_pol_A_bac	Domain	90.0	1.2e-25	4030
269955462	RNA_pol_A_CTD	Domain	90.0	5.6e-26	4030
257069480	RNA_pol_L	Domain	68.5	2.3e-19	4030
257069480	RNA_pol_A_bac	Domain	85.2	3.5e-24	4030
257069480	RNA_pol_A_CTD	Domain	86.0	9.6e-25	4030
315501403	RNA_pol_L	Domain	63.9	5.9e-18	4030
315501403	RNA_pol_A_bac	Domain	82.1	3.3e-23	4030
315501403	RNA_pol_A_CTD	Domain	88.3	1.8e-25	4030
253800502	RNA_pol_L	Domain	72.6	1.2e-20	4030
253800502	RNA_pol_A_bac	Domain	89.7	1.4e-25	4030
253800502	RNA_pol_A_CTD	Domain	91.1	2.4e-26	4030
121639377	RNA_pol_L	Domain	72.6	1.2e-20	4030
121639377	RNA_pol_A_bac	Domain	89.7	1.4e-25	4030
121639377	RNA_pol_A_CTD	Domain	91.1	2.4e-26	4030
296268581	RNA_pol_L	Domain	70.6	4.8e-20	4030
296268581	RNA_pol_A_bac	Domain	92.9	1.4e-26	4030
296268581	RNA_pol_A_CTD	Domain	94.5	2.2e-27	4030
308178106	RNA_pol_L	Domain	66.3	1.1e-18	4030
308178106	RNA_pol_A_bac	Domain	78.7	3.7e-22	4030
308178106	RNA_pol_A_CTD	Domain	89.6	7.5e-26	4030
15843052	RNA_pol_L	Domain	72.6	1.2e-20	4030
15843052	RNA_pol_A_bac	Domain	89.7	1.4e-25	4030
15843052	RNA_pol_A_CTD	Domain	91.1	2.4e-26	4030
291298742	RNA_pol_L	Domain	66.0	1.4e-18	4030
291298742	RNA_pol_A_bac	Domain	79.0	2.8e-22	4030
291298742	RNA_pol_A_CTD	Domain	87.3	3.9e-25	4030
325964132	RNA_pol_L	Domain	66.9	6.8e-19	4030
325964132	RNA_pol_A_bac	Domain	82.3	2.8e-23	4030
325964132	RNA_pol_A_CTD	Domain	91.9	1.4e-26	4030
183981110	RNA_pol_L	Domain	72.9	9.6e-21	4030
183981110	RNA_pol_A_bac	Domain	87.6	6.4e-25	4030
183981110	RNA_pol_A_CTD	Domain	92.5	8.7e-27	4030
117927543	RNA_pol_L	Domain	69.4	1.2e-19	4030
117927543	RNA_pol_A_bac	Domain	89.2	2.0e-25	4030
117927543	RNA_pol_A_CTD	Domain	93.8	3.6e-27	4030
119718095	RNA_pol_L	Domain	66.2	1.2e-18	4030
119718095	RNA_pol_A_bac	Domain	85.8	2.2e-24	4030
119718095	RNA_pol_A_CTD	Domain	89.7	6.8e-26	4030
116671486	RNA_pol_L	Domain	66.9	6.8e-19	4030
116671486	RNA_pol_A_bac	Domain	83.9	9.2e-24	4030
116671486	RNA_pol_A_CTD	Domain	91.9	1.4e-26	4030
148663322	RNA_pol_L	Domain	72.6	1.2e-20	4030
148663322	RNA_pol_A_bac	Domain	89.7	1.4e-25	4030

148663322	RNA_pol_A_CTD	Domain	91.1	2.4e-26	4030
159039805	RNA_pol_L	Domain	62.7	1.4e-17	4030
159039805	RNA_pol_A_bac	Domain	82.1	3.3e-23	4030
159039805	RNA_pol_A_CTD	Domain	88.3	1.8e-25	4030
312199992	RNA_pol_L	Domain	64.2	4.8e-18	4030
312199992	RNA_pol_A_bac	Domain	84.8	4.7e-24	4030
312199992	RNA_pol_A_CTD	Domain	92.5	8.9e-27	4030
152964684	RNA_pol_L	Domain	67.8	3.5e-19	4030
152964684	RNA_pol_A_bac	Domain	86.8	1.1e-24	4030
152964684	RNA_pol_A_CTD	Domain	89.2	9.9e-26	4030
145596406	RNA_pol_L	Domain	62.6	1.6e-17	4030
145596406	RNA_pol_A_bac	Domain	82.0	3.5e-23	4030
145596406	RNA_pol_A_CTD	Domain	90.5	3.8e-26	4030
336179754	RNA_pol_L	Domain	67.4	4.8e-19	4030
336179754	RNA_pol_A_bac	Domain	79.6	2.0e-22	4030
336179754	RNA_pol_A_CTD	Domain	92.6	8.1e-27	4030
31794633	RNA_pol_L	Domain	72.6	1.2e-20	4030
31794633	RNA_pol_A_bac	Domain	89.7	1.4e-25	4030
31794633	RNA_pol_A_CTD	Domain	91.1	2.4e-26	4030
297564065	RNA_pol_L	Domain	67.0	6.6e-19	4030
297564065	RNA_pol_A_bac	Domain	88.3	3.8e-25	4030
297564065	RNA_pol_A_CTD	Domain	90.6	3.5e-26	4030
182436600	RNA_pol_L	Domain	68.4	2.4e-19	4030
182436600	RNA_pol_A_bac	Domain	84.5	5.8e-24	4030
182436600	RNA_pol_A_CTD	Domain	89.9	5.9e-26	4030
119961601	RNA_pol_L	Domain	66.9	7.1e-19	4030
119961601	RNA_pol_A_bac	Domain	83.9	9.2e-24	4030
119961601	RNA_pol_A_CTD	Domain	91.9	1.4e-26	4030
72163017	RNA_pol_L	Domain	69.3	1.2e-19	4030
72163017	RNA_pol_A_bac	Domain	89.0	2.4e-25	4030
72163017	RNA_pol_A_CTD	Domain	93.6	4.2e-27	4030
345008598	RNA_pol_L	Domain	68.4	2.4e-19	4030
345008598	RNA_pol_A_bac	Domain	84.3	6.7e-24	4030
345008598	RNA_pol_A_CTD	Domain	89.9	5.9e-26	4030
284992846	RNA_pol_L	Domain	63.8	6.5e-18	4030
284992846	RNA_pol_A_bac	Domain	85.4	3.0e-24	4030
284992846	RNA_pol_A_CTD	Domain	89.2	9.7e-26	4030
29826981	RNA_pol_L	Domain	68.4	2.3e-19	4030
29826981	RNA_pol_A_bac	Domain	84.3	6.6e-24	4030
29826981	RNA_pol_A_CTD	Domain	87.9	2.5e-25	4030
29831496	RNA_pol_L	Domain	68.4	2.4e-19	4030
29831496	RNA_pol_A_bac	Domain	84.5	5.8e-24	4030
29831496	RNA_pol_A_CTD	Domain	89.9	5.9e-26	4030
302869942	RNA_pol_L	Domain	63.9	5.9e-18	4030
302869942	RNA_pol_A_bac	Domain	82.1	3.3e-23	4030
302869942	RNA_pol_A_CTD	Domain	88.3	1.8e-25	4030
148824667	RNA_pol_L	Domain	72.6	1.2e-20	4030
148824667	RNA_pol_A_bac	Domain	89.7	1.4e-25	4030
148824667	RNA_pol_A_CTD	Domain	91.1	2.4e-26	4030
331699145	RNA_pol_L	Domain	70.9	3.8e-20	4030
331699145	RNA_pol_A_bac	Domain	90.4	8.3e-26	4030

331699145	RNA_pol_A_CTD	Domain	89.6	7.3e-26	4030
339633462	RNA_pol_L	Domain	72.6	1.2e-20	4030
339633462	RNA_pol_A_bac	Domain	89.7	1.4e-25	4030
339633462	RNA_pol_A_CTD	Domain	91.1	2.4e-26	4030
317125840	RNA_pol_L	Domain	66.7	7.9e-19	4030
317125840	RNA_pol_A_bac	Domain	90.9	6.0e-26	4030
317125840	RNA_pol_A_CTD	Domain	89.5	7.9e-26	4030
184200289	RNA_pol_L	Domain	67.4	5.0e-19	4030
184200289	RNA_pol_A_bac	Domain	83.8	9.4e-24	4030
184200289	RNA_pol_A_CTD	Domain	90.0	5.5e-26	4030
256390173	RNA_pol_L	Domain	70.6	4.9e-20	4030
256390173	RNA_pol_A_bac	Domain	86.8	1.1e-24	4030
256390173	RNA_pol_A_CTD	Domain	90.8	3.0e-26	4030
229821593	RNA_pol_L	Domain	67.3	5.1e-19	4030
229821593	RNA_pol_A_bac	Domain	88.9	2.4e-25	4030
229821593	RNA_pol_A_CTD	Domain	87.6	3.1e-25	4030
220913400	RNA_pol_L	Domain	66.9	6.8e-19	4030
220913400	RNA_pol_A_bac	Domain	82.3	2.8e-23	4030
220913400	RNA_pol_A_CTD	Domain	91.9	1.4e-26	4030
284033971	RNA_pol_L	Domain	63.3	9.7e-18	4030
284033971	RNA_pol_A_bac	Domain	84.8	4.6e-24	4030
284033971	RNA_pol_A_CTD	Domain	89.8	6.3e-26	4030
108800542	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
108800542	AAA_2	Domain	171.1	2.1e-50	6725
108800542	ClpB_D2-small	Domain	60.9	8.2e-17	6725
271968523	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
271968523	AAA_2	Domain	172.1	1.0e-50	6725
271968523	ClpB_D2-small	Domain	63.6	1.2e-17	6725
224990834	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
224990834	AAA_2	Domain	170.2	3.8e-50	6725
224990834	ClpB_D2-small	Domain	61.3	6.1e-17	6725
258652075	zf-C4_ClpX	Domain	74.1	5.2e-21	6725
258652075	AAA_2	Domain	167.8	2.1e-49	6725
258652075	ClpB_D2-small	Domain	60.1	1.4e-16	6725
118618989	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
118618989	AAA_2	Domain	170.8	2.5e-50	6725
118618989	ClpB_D2-small	Domain	64.0	9.1e-18	6725
126436158	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
126436158	AAA_2	Domain	171.1	2.1e-50	6725
126436158	ClpB_D2-small	Domain	60.9	8.2e-17	6725
269955921	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
269955921	AAA_2	Domain	166.6	4.7e-49	6725
269955921	ClpB_D2-small	Domain	60.9	8.0e-17	6725
226360500	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
226360500	AAA_2	Domain	170.1	4.0e-50	6725
226360500	ClpB_D2-small	Domain	60.1	1.5e-16	6725
315504167	zf-C4_ClpX	Domain	73.4	9.0e-21	6725
315504167	AAA_2	Domain	170.3	3.5e-50	6725
315504167	ClpB_D2-small	Domain	55.9	3.0e-15	6725
253798464	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
253798464	AAA_2	Domain	170.2	3.8e-50	6725

253798464	ClpB_D2-small	Domain	61.3	6.1e-17	6725
121638340	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
121638340	AAA_2	Domain	170.2	3.8e-50	6725
121638340	ClpB_D2-small	Domain	61.3	6.1e-17	6725
296270439	zf-C4_ClpX	Domain	73.7	6.9e-21	6725
296270439	AAA_2	Domain	173.8	3.1e-51	6725
296270439	ClpB_D2-small	Domain	62.9	2.0e-17	6725
120404990	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
120404990	AAA_2	Domain	170.1	4.0e-50	6725
120404990	ClpB_D2-small	Domain	61.3	6.1e-17	6725
111018379	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
111018379	AAA_2	Domain	170.1	4.0e-50	6725
111018379	ClpB_D2-small	Domain	60.1	1.5e-16	6725
15841981	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
15841981	AAA_2	Domain	169.1	8.6e-50	6725
15841981	ClpB_D2-small	Domain	61.3	6.1e-17	6725
117927947	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
117927947	AAA_2	Domain	172.8	5.9e-51	6725
117927947	ClpB_D2-small	Domain	58.6	4.4e-16	6725
119717697	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
119717697	AAA_2	Domain	170.5	3.0e-50	6725
119717697	ClpB_D2-small	Domain	67.4	7.9e-19	6725
183983779	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
183983779	AAA_2	Domain	170.8	2.5e-50	6725
183983779	ClpB_D2-small	Domain	64.0	9.1e-18	6725
159039406	zf-C4_ClpX	Domain	73.4	9.0e-21	6725
159039406	AAA_2	Domain	171.3	1.8e-50	6725
159039406	ClpB_D2-small	Domain	54.0	1.2e-14	6725
148662292	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
148662292	AAA_2	Domain	170.2	3.8e-50	6725
148662292	ClpB_D2-small	Domain	61.3	6.1e-17	6725
312195812	zf-C4_ClpX	Domain	73.4	9.0e-21	6725
312195812	AAA_2	Domain	168.9	9.6e-50	6725
312195812	ClpB_D2-small	Domain	57.2	1.2e-15	6725
152967458	zf-C4_ClpX	Domain	73.4	9.0e-21	6725
152967458	AAA_2	Domain	171.9	1.1e-50	6725
152967458	ClpB_D2-small	Domain	65.3	3.4e-18	6725
134097948	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
134097948	AAA_2	Domain	170.1	4.1e-50	6725
134097948	ClpB_D2-small	Domain	59.2	2.7e-16	6725
145596009	zf-C4_ClpX	Domain	73.4	9.0e-21	6725
145596009	AAA_2	Domain	171.3	1.8e-50	6725
145596009	ClpB_D2-small	Domain	55.9	3.0e-15	6725
336177527	zf-C4_ClpX	Domain	73.4	9.0e-21	6725
336177527	AAA_2	Domain	169.5	6.3e-50	6725
336177527	ClpB_D2-small	Domain	56.0	2.9e-15	6725
31793638	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
31793638	AAA_2	Domain	170.2	3.8e-50	6725
31793638	ClpB_D2-small	Domain	61.3	6.1e-17	6725
182438718	zf-C4_ClpX	Domain	73.3	9.1e-21	6725
182438718	AAA_2	Domain	173.4	4.0e-51	6725

182438718	ClpB_D2-small	Domain	55.8	3.1e-15	6725
345015162	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
345015162	AAA_2	Domain	174.5	1.9e-51	6725
345015162	ClpB_D2-small	Domain	55.0	5.8e-15	6725
284990115	zf-C4_ClpX	Domain	73.4	8.8e-21	6725
284990115	AAA_2	Domain	173.4	4.0e-51	6725
284990115	ClpB_D2-small	Domain	60.0	1.5e-16	6725
15827775	zf-C4_ClpX	Domain	74.3	4.5e-21	6725
15827775	AAA_2	Domain	170.5	3.0e-50	6725
15827775	ClpB_D2-small	Domain	53.7	1.5e-14	6725
29831992	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
29831992	AAA_2	Domain	173.0	5.3e-51	6725
29831992	ClpB_D2-small	Domain	56.3	2.2e-15	6725
302869358	zf-C4_ClpX	Domain	73.4	9.0e-21	6725
302869358	AAA_2	Domain	170.3	3.5e-50	6725
302869358	ClpB_D2-small	Domain	55.9	3.0e-15	6725
119869681	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
119869681	AAA_2	Domain	171.1	2.1e-50	6725
119869681	ClpB_D2-small	Domain	60.9	8.2e-17	6725
148823657	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
148823657	AAA_2	Domain	170.2	3.8e-50	6725
148823657	ClpB_D2-small	Domain	61.3	6.1e-17	6725
331695809	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
331695809	AAA_2	Domain	170.7	2.6e-50	6725
331695809	ClpB_D2-small	Domain	60.4	1.2e-16	6725
333990090	zf-C4_ClpX	Domain	73.4	8.7e-21	6725
333990090	AAA_2	Domain	169.9	4.8e-50	6725
333990090	ClpB_D2-small	Domain	60.8	9.1e-17	6725
256375295	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
256375295	AAA_2	Domain	171.5	1.5e-50	6725
256375295	ClpB_D2-small	Domain	54.5	8.3e-15	6725
339632485	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
339632485	AAA_2	Domain	170.2	3.8e-50	6725
339632485	ClpB_D2-small	Domain	61.3	6.1e-17	6725
317124507	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
317124507	AAA_2	Domain	170.4	3.2e-50	6725
317124507	ClpB_D2-small	Domain	52.9	2.6e-14	6725
221230252	zf-C4_ClpX	Domain	74.3	4.5e-21	6725
221230252	AAA_2	Domain	170.5	3.0e-50	6725
221230252	ClpB_D2-small	Domain	53.7	1.5e-14	6725
229821077	zf-C4_ClpX	Domain	73.4	8.8e-21	6725
229821077	AAA_2	Domain	165.6	9.7e-49	6725
229821077	ClpB_D2-small	Domain	56.4	2.1e-15	6725
226307300	zf-C4_ClpX	Domain	73.4	8.9e-21	6725
226307300	AAA_2	Domain	168.6	1.2e-49	6725
226307300	ClpB_D2-small	Domain	62.9	1.9e-17	6725
257055264	zf-C4_ClpX	Domain	73.7	6.9e-21	6725
257055264	AAA_2	Domain	171.9	1.2e-50	6725
257055264	ClpB_D2-small	Domain	58.4	4.9e-16	6725
284030230	zf-C4_ClpX	Domain	73.3	9.0e-21	6725
284030230	AAA_2	Domain	168.4	1.4e-49	6725

284030230	ClpB_D2-small	Domain	61.0	7.5e-17	6725
336119295	zf-C4_ClpX	Domain	73.4	9.0e-21	6725
336119295	AAA_2	Domain	168.1	1.7e-49	6725
336119295	ClpB_D2-small	Domain	50.3	1.7e-13	6725

S. TABLE 7 – PFAM RESULTS FOR MOST DISCRIMINANT GENES FOR ANAEROBES.

<seq_id>	<hmm_name>	<type>	<bit_score>	<E-value>	<clust_id>
302336378	Ribosomal_S19	Domain	126.5	2.6e-37	19398
257065029	Ribosomal_S19	Domain	118.4	8.7e-35	19398
257791920	Ribosomal_S19	Domain	126.9	1.9e-37	19398
257784959	Ribosomal_S19	Domain	125.5	5.3e-37	19398
23465109	HSP70	Family	97.7	4.3e-28	28912
23465109	HSP70	Family	699.0	3.8e-210	28912
213691062	HSP70	Family	98.0	3.6e-28	28912
213691062	HSP70	Family	698.0	7.4e-210	28912
219683384	HSP70	Family	98.2	3.1e-28	28912
219683384	HSP70	Family	697.0	1.5e-209	28912
322688301	HSP70	Family	97.7	4.3e-28	28912
322688301	HSP70	Family	699.0	3.8e-210	28912
241191566	HSP70	Family	98.2	3.1e-28	28912
241191566	HSP70	Family	697.0	1.5e-209	28912
322690313	HSP70	Family	97.7	4.3e-28	28912
322690313	HSP70	Family	699.0	3.8e-210	28912
310288259	HSP70	Family	785.3	2.8e-236	28912
241196971	HSP70	Family	98.2	3.1e-28	28912
241196971	HSP70	Family	697.0	1.5e-209	28912
283783717	HSP70	Family	97.5	4.9e-28	28912
283783717	HSP70	Family	698.7	4.5e-210	28912
311114143	HSP70	Family	782.7	1.7e-235	28912
189440175	HSP70	Family	97.7	4.3e-28	28912
189440175	HSP70	Family	699.0	3.8e-210	28912
311065122	HSP70	Family	785.3	2.8e-236	28912
119026565	HSP70	Family	784.7	4.2e-236	28912
298345670	HSP70	Family	99.1	1.6e-28	28912
298345670	HSP70	Family	691.6	6.3e-208	28912
336326101	Terminase_4	Family	25.0	1.7e-05	98168

S. TABLE 8 – PFAM RESULTS FOR MOST DISCRIMINANT GENES FOR ANAEROBES.

<seq_id>	<hmm_name>	<type>	<bit_score>	<E-value>	<clust_id>
23466112	Ribosomal_L33	Family	90.8	4.7e-26	45
322691889	Ribosomal_L33	Family	91.0	4.2e-26	45
213693106	Ribosomal_L33	Family	91.0	4.2e-26	45
241190415	Ribosomal_L33	Family	88.6	2.3e-25	45
311115182	Ribosomal_L33	Family	93.9	5.2e-27	45
310288042	Ribosomal_L33	Family	90.9	4.5e-26	45
189440542	Ribosomal_L33	Family	91.0	4.2e-26	45
219682839	Ribosomal_L33	Family	81.2	4.8e-23	45
311064918	Ribosomal_L33	Family	90.9	4.5e-26	45
322689948	Ribosomal_L33	Family	91.0	4.2e-26	45
283782743	Ribosomal_L33	Family	93.4	7.7e-27	45

241195821	Ribosomal_L33	Family	88.6	2.3e-25	45
298346263	Ribosomal_L33	Family	85.6	2.0e-24	45
297625754	Ribosomal_L33	Family	84.4	4.7e-24	45
119025311	Ribosomal_L33	Family	87.4	5.5e-25	45
213691460	ABC_tran	Domain	89.3	2.8e-25	1449
213691460	ABC_tran_Xtn	Domain	53.3	2.0e-14	1449
213691460	ABC_tran	Domain	72.6	4.1e-20	1449
311064783	ABC_tran	Domain	89.1	3.2e-25	1449
311064783	ABC_tran_Xtn	Domain	53.4	1.8e-14	1449
311064783	ABC_tran	Domain	72.2	5.3e-20	1449
311114329	ABC_tran	Domain	89.9	1.9e-25	1449
311114329	ABC_tran_Xtn	Domain	50.9	1.0e-13	1449
311114329	ABC_tran	Domain	68.8	6.1e-19	1449
189440717	ABC_tran	Domain	89.0	3.4e-25	1449
189440717	ABC_tran_Xtn	Domain	53.3	2.0e-14	1449
189440717	ABC_tran	Domain	72.7	3.8e-20	1449
257064916	ABC_tran	Domain	93.9	1.0e-26	1449
257064916	ABC_tran_Xtn	Domain	46.8	2.0e-12	1449
257064916	ABC_tran	Domain	69.7	3.0e-19	1449
310287902	ABC_tran	Domain	89.1	3.2e-25	1449
310287902	ABC_tran_Xtn	Domain	53.4	1.8e-14	1449
310287902	ABC_tran	Domain	72.2	5.3e-20	1449
322691762	ABC_tran	Domain	89.3	2.8e-25	1449
322691762	ABC_tran_Xtn	Domain	53.3	2.0e-14	1449
322691762	ABC_tran	Domain	72.6	4.1e-20	1449
322689823	ABC_tran	Domain	89.3	2.8e-25	1449
322689823	ABC_tran_Xtn	Domain	53.3	2.0e-14	1449
322689823	ABC_tran	Domain	72.6	4.1e-20	1449
241190547	ABC_tran	Domain	91.6	5.6e-26	1449
241190547	ABC_tran_Xtn	Domain	51.6	6.6e-14	1449
241190547	ABC_tran	Domain	69.1	4.8e-19	1449
219682970	ABC_tran	Domain	91.5	5.8e-26	1449
219682970	ABC_tran_Xtn	Domain	51.5	6.8e-14	1449
219682970	ABC_tran	Domain	69.1	4.9e-19	1449
23466236	ABC_tran	Domain	89.3	2.8e-25	1449
23466236	ABC_tran_Xtn	Domain	53.3	2.0e-14	1449
23466236	ABC_tran	Domain	72.7	3.8e-20	1449
283783574	ABC_tran	Domain	91.3	6.7e-26	1449
283783574	ABC_tran_Xtn	Domain	51.5	6.8e-14	1449
283783574	ABC_tran	Domain	69.0	5.0e-19	1449
257791194	ABC_tran	Domain	91.2	7.0e-26	1449
257791194	ABC_tran_Xtn	Domain	49.5	2.8e-13	1449
257791194	ABC_tran	Domain	70.5	1.7e-19	1449
241195953	ABC_tran	Domain	91.6	5.6e-26	1449
241195953	ABC_tran_Xtn	Domain	51.6	6.6e-14	1449
241195953	ABC_tran	Domain	69.1	4.8e-19	1449
297571050	ABC_tran	Domain	98.6	3.7e-28	1449
297571050	ABC_tran_Xtn	Domain	48.8	4.7e-13	1449
297571050	ABC_tran	Domain	70.0	2.5e-19	1449
257784105	ABC_tran	Domain	97.0	1.1e-27	1449
257784105	ABC_tran_Xtn	Domain	52.8	2.7e-14	1449

257784105	ABC_tran	Domain	72.3	4.9e-20	1449
119025433	ABC_tran	Domain	88.7	4.1e-25	1449
119025433	ABC_tran_Xtn	Domain	52.4	3.7e-14	1449
119025433	ABC_tran	Domain	71.6	8.2e-20	1449

S. TABLE 9 – PFAM RESULTS FOR MOST DISCRIMINANT GENES FOR FACULTATIVES.

<seq_id>	<hmm_name>	<type>	<bit_score>	<E-value>	<clust_id>
336325636	Gp_dh_N	Domain	202.5	2.9e-60	6075
336325636	Gp_dh_C	Domain	226.3	1.2e-67	6075

S. TABLE 10 – PFAM RESULTS FOR MOST DISCRIMINANT GENES FOR SOIL.

<seq_id>	<hmm_name>	<type>	<bit_score>	<E-value>	<clust_id>
291301862	HTH_26	Domain	72.1	3.3e-20	3326
116669488	HTH_26	Domain	66.9	1.5e-18	3326
269956770	HTH_26	Domain	69.6	2.1e-19	3326
325964371	HTH_26	Domain	69.1	2.9e-19	3326
134097153	HTH_26	Domain	70.0	1.6e-19	3326
336176624	HTH_26	Domain	68.0	6.3e-19	3326
345013705	HTH_26	Domain	68.7	3.9e-19	3326
256376853	HTH_26	Domain	67.7	7.8e-19	3326
256393115	HTH_26	Domain	69.3	2.5e-19	3326
220913410	HTH_26	Domain	66.7	1.7e-18	3326
284045570	HTH_26	Domain	69.0	3.2e-19	3326
296131301	HTH_26	Domain	68.8	3.7e-19	3326
29826603	HTH_26	Domain	68.7	4.0e-19	3326
284033180	HTH_26	Domain	68.7	3.9e-19	3326
62390180	HTH_26	Domain	68.7	4.1e-19	3326
145295435	HTH_26	Domain	67.6	8.8e-19	3326
312200738	HTH_26	Domain	68.7	4.0e-19	3326
284991253	HTH_26	Domain	70.9	7.9e-20	3326
302867851	HTH_26	Domain	70.1	1.4e-19	3326
331695298	HTH_26	Domain	68.3	5.3e-19	3326
229818535	HTH_26	Domain	66.4	2.0e-18	3326
315505747	HTH_26	Domain	70.1	1.4e-19	3326
116671347	ABC_tran	Domain	126.6	8.2e-37	6779
257069406	ABC_tran	Domain	110.4	8.2e-32	6779
239918030	ABC_tran	Domain	111.9	2.8e-32	6779
291301450	ABC_tran	Domain	123.6	7.3e-36	6779
333920073	ABC_tran	Domain	124.5	3.9e-36	6779
134098328	ABC_tran	Domain	120.8	5.4e-35	6779
345010107	ABC_tran	Domain	121.2	3.9e-35	6779
119868238	ABC_tran	Domain	123.3	8.9e-36	6779
256375585	ABC_tran	Domain	122.3	1.8e-35	6779
336180055	ABC_tran	Domain	117.7	4.7e-34	6779
159037033	ABC_tran	Domain	120.3	7.7e-35	6779
345009766	ABC_tran	Domain	124.4	4.0e-36	6779
220913266	ABC_tran	Domain	125.1	2.5e-36	6779
340794576	ABC_tran	Domain	125.1	2.5e-36	6779
325964011	ABC_tran	Domain	124.6	3.5e-36	6779
108803351	ABC_tran	Domain	128.1	2.9e-37	6779

111023728	ABC_tran	Domain	120.8	5.2e-35	6779
126434729	ABC_tran	Domain	123.1	1.0e-35	6779
296129398	ABC_tran	Domain	126.1	1.2e-36	6779
111023901	ABC_tran	Domain	118.2	3.3e-34	6779
315502597	ABC_tran	Domain	119.9	1.0e-34	6779
152965466	ABC_tran	Domain	127.0	6.2e-37	6779
182435044	ABC_tran	Domain	125.4	2.0e-36	6779
182435529	ABC_tran	Domain	124.9	2.9e-36	6779
29828505	ABC_tran	Domain	123.3	8.6e-36	6779
29829026	ABC_tran	Domain	124.9	2.9e-36	6779
226306245	ABC_tran	Domain	125.0	2.7e-36	6779
284032613	ABC_tran	Domain	123.0	1.1e-35	6779
271963752	ABC_tran	Domain	121.3	3.6e-35	6779
271967565	ABC_tran	Domain	117.3	6.2e-34	6779
145295855	ABC_tran	Domain	127.6	4.1e-37	6779
312198102	ABC_tran	Domain	120.8	5.3e-35	6779
145593978	ABC_tran	Domain	123.0	1.1e-35	6779
119962732	ABC_tran	Domain	126.9	7.1e-37	6779
284992298	ABC_tran	Domain	126.0	1.3e-36	6779
302866038	ABC_tran	Domain	119.9	1.0e-34	6779
62390790	ABC_tran	Domain	127.0	6.6e-37	6779
229820929	ABC_tran	Domain	121.0	4.5e-35	6779
331697740	ABC_tran	Domain	114.9	3.4e-33	6779
257070105	ABC_tran	Domain	129.5	1.1e-37	6779
269956047	ABC_tran	Domain	122.9	1.2e-35	6779
108799123	ABC_tran	Domain	123.1	1.0e-35	6779
333919403	Acyl-CoA_dh_N	Domain	87.1	1.1e-24	10196
333919403	Acyl-CoA_dh_M	Domain	93.2	7.8e-27	10196
333919403	Acyl-CoA_dh_1	Domain	160.2	3.5e-47	10196
134097255	Acyl-CoA_dh_N	Domain	88.6	3.7e-25	10196
134097255	Acyl-CoA_dh_M	Domain	86.0	1.4e-24	10196
134097255	Acyl-CoA_dh_1	Domain	159.0	8.2e-47	10196
336180032	Acyl-CoA_dh_N	Domain	82.4	3.2e-23	10196
336180032	Acyl-CoA_dh_M	Domain	80.8	5.6e-23	10196
336180032	Acyl-CoA_dh_1	Domain	169.9	3.6e-50	10196
345013436	Acyl-CoA_dh_N	Domain	83.2	1.8e-23	10196
345013436	Acyl-CoA_dh_M	Domain	76.3	1.5e-21	10196
345013436	Acyl-CoA_dh_1	Domain	160.0	4.2e-47	10196
119866165	Acyl-CoA_dh_N	Domain	90.6	9.2e-26	10196
119866165	Acyl-CoA_dh_M	Domain	86.0	1.4e-24	10196
119866165	Acyl-CoA_dh_1	Domain	154.0	3.0e-45	10196
340795058	Acyl-CoA_dh_N	Domain	92.8	1.9e-26	10196
340795058	Acyl-CoA_dh_M	Domain	92.9	9.6e-27	10196
340795058	Acyl-CoA_dh_1	Domain	157.5	2.4e-46	10196
108797080	Acyl-CoA_dh_N	Domain	90.6	9.2e-26	10196
108797080	Acyl-CoA_dh_M	Domain	86.0	1.4e-24	10196
108797080	Acyl-CoA_dh_1	Domain	154.0	3.0e-45	10196
284042015	Acyl-CoA_dh_N	Domain	96.5	1.3e-27	10196
284042015	Acyl-CoA_dh_M	Domain	75.3	3.0e-21	10196
284042015	Acyl-CoA_dh_1	Domain	160.9	2.2e-47	10196
111020379	Acyl-CoA_dh_N	Domain	91.2	6.0e-26	10196

111020379	Acyl-CoA_dh_M	Domain	88.0	3.2e-25	10196
111020379	Acyl-CoA_dh_1	Domain	158.4	1.3e-46	10196
126432702	Acyl-CoA_dh_N	Domain	90.6	9.2e-26	10196
126432702	Acyl-CoA_dh_M	Domain	86.0	1.4e-24	10196
126432702	Acyl-CoA_dh_1	Domain	154.0	3.0e-45	10196
126436590	Acyl-CoA_dh_N	Domain	89.4	2.2e-25	10196
126436590	Acyl-CoA_dh_M	Domain	84.4	4.3e-24	10196
126436590	Acyl-CoA_dh_1	Domain	165.0	1.2e-48	10196
119714508	Acyl-CoA_dh_N	Domain	82.2	3.6e-23	10196
119714508	Acyl-CoA_dh_M	Domain	96.4	7.6e-28	10196
119714508	Acyl-CoA_dh_1	Domain	161.0	2.1e-47	10196
284029016	Acyl-CoA_dh_N	Domain	91.8	4.0e-26	10196
284029016	Acyl-CoA_dh_M	Domain	83.3	9.8e-24	10196
284029016	Acyl-CoA_dh_1	Domain	158.0	1.7e-46	10196
226305373	Acyl-CoA_dh_N	Domain	86.3	2.0e-24	10196
226305373	Acyl-CoA_dh_M	Domain	84.3	4.7e-24	10196
226305373	Acyl-CoA_dh_1	Domain	162.5	7.0e-48	10196
312197807	Acyl-CoA_dh_N	Domain	83.9	1.1e-23	10196
312197807	Acyl-CoA_dh_M	Domain	82.8	1.4e-23	10196
312197807	Acyl-CoA_dh_1	Domain	171.0	1.7e-50	10196
284990870	Acyl-CoA_dh_N	Domain	90.0	1.4e-25	10196
284990870	Acyl-CoA_dh_M	Domain	76.2	1.6e-21	10196
284990870	Acyl-CoA_dh_1	Domain	157.9	1.8e-46	10196
239917418	Hexapep	Repeat	18.8	0.00085	19513
239917418	Hexapep	Repeat	28.6	7.2e-07	19513
239917418	Hexapep_2	Repeat	21.7	0.00012	19513
257069246	Hexapep	Repeat	29.5	3.5e-07	19513
269957471	Hexapep	Repeat	28.1	1.0e-06	19513
269957471	Hexapep	Repeat	23.9	2.1e-05	19513
269957471	Hexapep	Repeat	20.3	0.0003	19513
116672594	Hexapep	Repeat	32.8	3.4e-08	19513
116672594	Hexapep	Repeat	18.4	0.0012	19513
116672594	Hexapep_2	Repeat	22.0	9.4e-05	19513
256396732	Hexapep	Repeat	32.6	3.9e-08	19513
256396732	Hexapep	Repeat	21.1	0.00017	19513
256396732	Hexapep_2	Repeat	15.4	0.01	19513
220913170	Hexapep	Repeat	38.3	6.2e-10	19513
220913170	Hexapep	Repeat	18.5	0.0011	19513
182438743	Hexapep	Repeat	23.8	2.4e-05	19513
182438743	Hexapep	Repeat	27.1	2.1e-06	19513
182438743	Hexapep_2	Repeat	15.8	0.008	19513
119718428	Hexapep	Repeat	33.7	1.7e-08	19513
119718428	Hexapep	Repeat	20.4	0.00028	19513
119718428	Hexapep_2	Repeat	23.4	3.3e-05	19513
325963916	Hexapep	Repeat	33.0	2.9e-08	19513
325963916	Hexapep	Repeat	23.2	3.6e-05	19513
325963916	Hexapep_2	Repeat	24.7	1.3e-05	19513
119963120	Hexapep	Repeat	26.2	3.9e-06	19513
119963120	Hexapep	Repeat	22.9	4.4e-05	19513
119963120	Hexapep	Repeat	22.2	7.6e-05	19513
269955211	Arabinose_Isome	Family	568.5	5.5e-171	20354

269955211	Arabinose_Iso_C	Domain	152.0	4.4e-45	20354
116668792	Arabinose_Isome	Family	569.8	2.2e-171	20354
116668792	Arabinose_Iso_C	Domain	158.8	3.6e-47	20354
325961809	Arabinose_Isome	Family	566.1	2.9e-170	20354
325961809	Arabinose_Iso_C	Domain	156.9	1.4e-46	20354
333919899	Arabinose_Isome	Family	573.7	1.4e-172	20354
333919899	Arabinose_Iso_C	Domain	156.0	2.5e-46	20354
256375997	Arabinose_Isome	Family	569.1	3.4e-171	20354
256375997	Arabinose_Iso_C	Domain	153.9	1.1e-45	20354
256393851	Arabinose_Isome	Family	556.6	2.3e-167	20354
256393851	Arabinose_Iso_C	Domain	152.6	2.9e-45	20354
220911160	Arabinose_Isome	Family	567.5	1.0e-170	20354
220911160	Arabinose_Iso_C	Domain	155.2	4.6e-46	20354
296130588	Arabinose_Isome	Family	571.8	5.4e-172	20354
296130588	Arabinose_Iso_C	Domain	156.5	1.8e-46	20354
315505465	Arabinose_Isome	Family	564.2	1.1e-169	20354
315505465	Arabinose_Iso_C	Domain	155.1	5.0e-46	20354
119714647	Arabinose_Isome	Family	573.3	1.9e-172	20354
119714647	Arabinose_Iso_C	Domain	156.4	2.0e-46	20354
271967493	Arabinose_Isome	Family	572.4	3.4e-172	20354
271967493	Arabinose_Iso_C	Domain	163.4	1.3e-48	20354
312197283	Arabinose_Isome	Family	564.0	1.3e-169	20354
312197283	Arabinose_Iso_C	Domain	151.7	5.8e-45	20354
117928081	Arabinose_Isome	Family	546.9	2.0e-164	20354
117928081	Arabinose_Iso_C	Domain	150.7	1.1e-44	20354
284988866	Arabinose_Isome	Family	571.9	5.0e-172	20354
284988866	Arabinose_Iso_C	Domain	155.8	3.0e-46	20354
119961215	Arabinose_Isome	Family	568.9	4.1e-171	20354
119961215	Arabinose_Iso_C	Domain	158.5	4.3e-47	20354
302868138	Arabinose_Isome	Family	564.2	1.1e-169	20354
302868138	Arabinose_Iso_C	Domain	155.1	5.0e-46	20354
229821751	Arabinose_Isome	Family	561.4	7.9e-169	20354
229821751	Arabinose_Iso_C	Domain	161.1	7.0e-48	20354
257070051	Arabinose_Isome	Family	578.4	5.3e-174	20354
257070051	Arabinose_Iso_C	Domain	155.2	4.7e-46	20354

APPENDIX B

PEER-REVIEWED JOURNAL PUBLICATIONS

1. Chen M, Baumbach J, Vandin F, Röttger R, Barbosa E, Dong M, Frost M, Christiansen L, Tan Q. Differentially methylated genomic regions in adult twin pairs discordant for birth-weight. Annals of Human Genetics.
2. Zeng Y, Baumbach J, Barbosa EGV, Azevedo V, Zhang C, Koblížek M. 2015. Metagenomic evidence for the presence of phototrophic Gemmatimonadetes bacteria in diverse environments. Environmental Microbiology Reports 2015.
3. Soares, Siomar C; Geyik, Hakan; Ramos, Rommel TJ; de Sá, Pablo HCG; Barbosa, Eudes GV; Baumbach, Jan; Figueiredo, Henrique CP; Miyoshi, Anderson; Tauch, Andreas; Silva, Artur: GIPSy: Genomic island prediction software. Journal of Biotechnology 2015.
4. Barbosa E, Röttger R, Hauschild A-C, Azevedo V, Baumbach J: On the limits of computational functional genomics for bacterial lifestyle prediction. Briefings in Functional Genomics 2014.
5. Barbosa EGV, Aburjaile FF, Ramos RTJ, Carneiro AR, Le Loir Y, Baumbach J, Miyoshi A, Silva A, Azevedo V. Value of a newly sequenced bacterial genome. World J Biol Chem 2014; 5(2): 161-168
6. Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, Ali A, Santos AR, Pinto AC, Diniz C: The pan-genome of the animal pathogen Corynebacterium pseudotuberculosis reveals differences in genome plasticity between the biovar ovis and equi strains. PloS one 2013, 8(1):e53818.
7. Santos AR, Pereira VB, Barbosa E, Baumbach J, Pauling J, Röttger R, Turk MZ, Silva A, Miyoshi A, Azevedo V: Mature Epitope Density-A strategy for target selection based on immunoinformatics and exported prokaryotic proteins. BMC Genomics 2013, 14(Suppl 6):S4.
8. Santos A, Barbosa E, Fiaux K, Zurita-Turk M, Chaitankar V, Kamapantula B, Abdelzaher A, Ghosh P, Tiwari S, Barve N: PANNOTATOR: an automated tool for annotation of pan-genomes. Genetics and Molecular Research 2013, 12(3):2982-2989.
9. Soares SC, Trost E, Ramos RT, Carneiro AR, Santos AR, Pinto AC, Barbosa E, Aburjaile F, Ali A, Diniz CA: Genome sequence of Corynebacterium pseudotuberculosis biovar equi strain 258 and prediction of antigenic targets to improve biotechnological vaccine production. Journal of biotechnology 2013, 167(2):135-141.
10. Ramos RTJ, Carneiro AR, Soares SdC, Santos ARd, Almeida S, Guimarães L, Figueira F, Barbosa E, Tauch A, Azevedo V: Tips and tricks for the assembly of a Corynebacterium pseudotuberculosis genome using a semiconductor sequencer. Microbial biotechnology 2013, 6(2):150-156.
11. Ali A, Soares SC, Barbosa E, Santos AR, Barh D, Bakhtiar SM, Hassan SS, Ussery DW, Silva A, Miyoshi A: Microbial Comparative Genomics: An Overview of Tools and Insights Into The Genus Corynebacterium. Journal of Bacteriology & Parasitology 2013, 4(2).
12. Silva A, Ramos RTJ, Carneiro AR, Pinto AC, de Castro Soares S, Santos AR, Almeida SS, Guimarães LC, Aburjaile FF, Barbosa EGV: Complete genome

sequence of Corynebacterium pseudotuberculosis Cp31, isolated from an Egyptian buffalo. Journal of bacteriology 2012, 194(23):6663-6664.
13. Santos AR, Carneiro A, Gala-García A, Pinto A, Barh D, Barbosa E, Aburjaile F, Dorella F, Rocha F, Guimarães L: The Corynebacterium pseudotuberculosis in silico predicted pan-exoproteome. BMC genomics 2012, 13(Suppl 5):S6.
14. Ramos RTJ, Silva A, Carneiro AR, Pinto AC, de Castro Soares S, Santos AR, Almeida SS, Guimarães LC, Aburjaile FF, Barbosa EGV: Genome sequence of the Corynebacterium pseudotuberculosis Cp316 strain, isolated from the abscess of a Californian horse. Journal of bacteriology 2012, 194(23):6620-6621.
15. Pethick FE, Lainson AF, Yaga R, Flockhart A, Smith DG, Donachie W, Cerdeira LT, Silva A, Bol E, Lopes TS: Complete genome sequence of Corynebacterium pseudotuberculosis strain 1/06-A, isolated from a horse in North America. Journal of bacteriology 2012, 194(16):4476-4476.
16. Pethick FE, Lainson AF, Yaga R, Flockhart A, Smith DG, Donachie W, Cerdeira LT, Silva A, Bol E, Lopes TS: Complete genome sequences of Corynebacterium pseudotuberculosis strains 3/99-5 and 42/02-A, isolated from sheep in Scotland and Australia, respectively. Journal of bacteriology 2012, 194(17):4736-4737.
17. Lopes T, Silva A, Thiago R, Carneiro A, Dorella FA, Rocha FS, dos Santos AR, Lima ARJ, Guimarães LC, Barbosa EG: Complete genome sequence of Corynebacterium pseudotuberculosis strain Cp267, isolated from a llama. Journal of bacteriology 2012, 194(13):3567-3568.
18. Hassan SS, Schneider MPC, Ramos RTJ, Carneiro AR, Ranieri A, Guimarães LC, Ali A, Bakhtiar SM, de Pádua Pereira U, dos Santos AR: Whole-genome sequence of Corynebacterium pseudotuberculosis strain Cp162, isolated from camel. Journal of bacteriology 2012, 194(20):5718-5719.
19. Hassan SS, Guimarães LC, Islam A, Ali A, Bakhtiar S, Ribeiro D, Rodrigues Dos Santos A, Dorella F, Pinto A, Schneider M: Complete genome sequence of Corynebacterium pseudotuberculosis biovar ovis strain P54B96 isolated from antelope in South Africa obtained by Rapid Next Generation Sequencing Technology. Standards in genomic sciences 2012, 7(2):189-199.
20. Carneiro AR, Ramos RTJ, Dall'Agnol H, Pinto AC, de Castro Soares S, Santos AR, Guimarães LC, Almeida SS, Baraúna RA, das Graças DA: Genome sequence of Exiguobacterium antarcticum B7, isolated from a biofilm in Ginger Lake, King George Island, Antarctica. Journal of bacteriology 2012, 194(23):6689-6690.
21. Ali A, Soares SC, Santos AR, Guimarães LC, Barbosa E, Almeida SS, Abreu VA, Carneiro AR, Ramos RT, Bakhtiar SM: Campylobacter fetus subspecies: Comparative genomics and prediction of potential virulence targets. Gene 2012, 508(2):145-156.
22. Stynen APR, Lage AP, Moore RJ, Rezende AM, de Cássia Ruy P, Daher N, de Melo Resende D, de Almeida SS, de Castro Soares S, de Abreu VAC: Complete genome sequence of type strain Campylobacter fetus subsp. venerealis NCTC 10354T. Journal of bacteriology 2011, 193(20):5871-5872.
23. Santos A, Ali A, Barbosa E, Silva A, Miyoshi A, Barh D, Azevedo V: The Reverse Vaccinology - A Contextual Overview. IIOAB Journal 2011, 2(4).
24. Cerdeira LT, Schneider MPC, Pinto AC, de Almeida SS, dos Santos AR, Barbosa EGV, Ali A, Aburjaile FF, de Abreu VAC, Guimarães LC: Complete genome

sequence of Corynebacterium pseudotuberculosis strain CIP 52.97, isolated from a horse in Kenya. Journal of bacteriology 2011, 193(24):7025-7026.
25. Cerdeira LT, Pinto AC, Schneider MPC, de Almeida SS, Dos Santos AR, Barbosa EGV, Ali A, Barbosa MS, Carneiro AR, Ramos RTJ: Whole-genome sequence of Corynebacterium pseudotuberculosis PAT10 strain isolated from sheep in Patagonia, Argentina. Journal of bacteriology 2011, 193(22):6420-6421.