Masterarbeit

16S rDNA sequencing of bacterial pathogens on Illumina MiniSeq: Primer optimization for human pathogenic bacteria and short read lengths

eingereicht von

Dr. med. univ. Franz Pühringer

zur Erlangung des akademischen Grades

Master of Science

(Msc)

an der

Medizinischen Universität Graz

ausgeführt am

Institut für Humangenetik

unter der Anleitung von Betreuer

Priv. Doz. Mag. Dr. phil. Karl Kashofer

St. Konrad, 25.06.2020

Eidesstattliche Erklärung

Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst habe, andere als die angegebenen Quellen nicht verwendet habe und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Acknowledgements

I would like to thank **Evelyn Lang**, whose bachelor's thesis at the FH Gesundheitsberufe OÖ (University of Applied Sciences for Health Professions Upper Austria) I supervised, and who also carried out the pipetting work of the library preparation for the sequencing of the bacterial cultures.

I thank my staff in the molecular pathology laboratory, the biomedical analysts (BMA's) **Karin Merkle**, **Penka Lechner** and, above all, **Regina Stitz** for the support of our bachelor student and professional exchange.

It is in great parts thanks to Regina that we have been able to expand the molecular pathology laboratory in the institute and make it what it is today. I would also like to thank her for this. And last but not least, that we were able to complete this extra-occupational postgraduate course together - which was certainly a relief for both of us.

Many thanks also to **Ellen Heitzer** who kindly made the Library Preparation Protocol available to me.

I am indebted to **Karl Kashofer** for supervising this master's thesis.

Above all however, it is of primary concern for me to thank my wife **Erika Johanna** for the patience that she showed when there was little time left for the family. It enabled me to spend a considerable amount of my free time on my studies in addition to my job. It would not have been possible without her encouragement and support.

Table of Contents

Abbreviations

DNA	deoxyribonucleic acid
e.g.	for example (lat. 'exempli gratia')
FFPE	formalin-fixed paraffin-embedded (tissue)
i.e.	that is (lat. 'id est')
mRNA	messenger RNA
mtDNA	mitochondrial DNA
MT-RNR1	mitochondrially encoded 12S rRNA, coding for mitochondrial SSU-rRNA
NGS	Next Generation Sequencing
PCR	polymerase chain reaction
ptDNA	plastid DNA
RNA	ribonucleic acid
RNR1	RNA, Ribosomal 45S Cluster 1), coding for cytoplasmic SSU-rRNA
rRNA	ribosomale RNA
s-rRNA	small subunit ribosomal RNA (human)
SSU	small subunit (of ribosome)
SSU-rRNA	small subunit ribosomal RNA
tRNA	transfer RNA

Glossary

Alignment, align

arrangement of homologous nucleotides or amino acids of sequences (DNA, RNA or protein) in vertical columns to identify regions of similarity or dissimilarity; for missing residues gaps are inserted

Bootstrap values

indicate how many times (%) the same branch was observed when repeating the phylogenetic reconstruction on a re-sampled set of the data

Eukaryotes

organisms with a real cell nucleus: protists, plants, fungi, animals; usually containing also mitochondria

Genus, Pl.: Genera

taxonomic category, including one or more closely related species

Phylogeny

evolutionary history of life

today, phylogenetic trees of life are usually generated from DNA sequences using appropriate computer programs

Prokaryotes

organisms without a nucleus: bacteria and archaea

Species

basic taxonomic category (however, there is no generally accepted definition of the concept of species)

scientific species names are composed of generic name [always upper case] and actual species name [always lower case], e.g. *Escherichia coli*

Taxonomy

classification, e.g. phylogenetic systematics of life

Figures

Tables

Zusammenfassung

Die Identifizierung bakterieller Erreger wird gewöhnlich, wenn eine Kultivierung nicht möglich ist – z. B. bei Formalin fixiertem Paraffin eingebettetem (FFPE) Gewebe – mittels 16S-rDNA-Sequenzierung durchgeführt. Die hierfür verwendeten Primer, die in hoch konservierten Regionen des 16S-rRNA Gens binden, sollten für human-pathogene Arten optimiert sein (was bisher nicht der Fall war). Außerdem sollten sie für die vorliegende Arbeit so gewählt werden, dass eine 16S-rDNA-Sequenzierung auch mit dem vorhandenen Illumina MiniSeq (Leselänge nur 2x150 bp) möglich war.

Es wurden daher 16S-rDNA-Sequenzen nahezu sämtlicher (2363) humanpathogener Bakterienarten aus NCBI heruntergeladen, aligniert, Sequenzierartefakte nach Möglichkeit eliminiert und unter Berücksichtigung von Konservierungsgrad der Primerbindungsstellen, Amplikonlänge, Schmelztemperatur und optimaler Annealingtemperatur für humanpathogene Bakterien optimierte Primer entworfen. Anschließend wurden die Primer sowohl mit der Sequenzdatenbank human-pathogener Bakterien wie auch mit der 'SILVA ribosomal RNA gene database' abgeglichen.

Der Sequenzvergleich mit humaner 18S- und 12S-rDNA (zytoplasmatischer bzw. mitochondrialer Ribosomen) zeigt, dass mit den vorgestellten Primerpaaren lediglich eine Coamplifikation von 18S-rDNA in V3 sowie 12S-rDNA in V3-V4 gut möglich ist (üblicherweise nur wenn keine bakterielle DNA in der Probe enthalten ist).

Die Funktionalität der neuen Primer wurde zunächst an Bakterienkulturen erfolgreich getestet, die generierten Sequenzen mit QIIME2 und BLAST ausgewertet und mit der massenspektrometrisch erfolgten Determination der Bakterien verglichen.

Insgesamt passen alle Primer für 75,4% der humanpathogenen Arten exakt. Die Passgenauigkeit der einzelnen Primer beträgt 92,3-98,77% bzw. 96,74-99,96% (1 mismatch allowed). Bei Verwendung aller vorgeschlagener Primerpaare sollten sämtliche humanpathogene Arten amplifizierbar sein.

Ein Vergleich mit der 'SILVA ribosomal RNA gene database' zeigt, dass die Primer/Primerpaare im Mittel für 3,5 bzw. 6,4 % mehr humanpathogene Arten exakt passen.

Es traten je nach verwendetem Primerpaar in sehr unterschiedlicher Frequenz (0,11 - 90,17%) Artefakte auf Grund von Primerdimeren, Primerpaardimeren oder primer

hairpins auf (relativ kurze Sequenzen, deren fehlendes Signal am Ende auf Grund der Illumina 2-dye chemistry als poly-G string gelesen wird).
Abschließend wurde die homologe small-subunit-rRNA ursprünglicher Eukaryoten in den Datensatz bakterieller 16S-rRNA integriert. Der resultierende phylogenetische Stammbaum der Bacteria stützt eine Abstammung der Mitochondrien von einem gemeinsamen Vorfahren mit den Rickettsiales. Eine enge Verwandtschaft der Mitochondrien als Endosymbionten mit den Rickettsiales als obligat intrazelluläre (Endo-)Parasiten erscheint aus evolutionärer Sicht durchaus plausibel.

Abstract

Bacterial pathogens are usually identified by 16S rDNA sequencing when culture is not possible – e.g. from formalin fixed paraffin embedded (FFPE) tissue. The primers used for this, binding in highly conserved regions of the 16S rRNA gene, ought to be optimized for human pathogenic species (which has not previously been the case). For the present work they needed also be properly placed allowing 16S rDNA sequencing using the available Illumina MiniSeq (read length 2x150 bp).

Therefore, 16S rDNA sequences of almost all (2363) human pathogenic bacterial species were downloaded from NCBI, aligned, and sequencing artifacts were eliminated where possible. Primers optimized for human pathogenic bacteria were designed, taking into account the degree of preservation of the primer binding sites, amplicon length, melting temperature and optimal annealing temperature. The primers were then compared both with the sequence database of human pathogenic bacteria and with the 'SILVA ribosomal RNA gene database'.

Sequence comparison with human 18S and 12S rDNA (cytoplasmic or mitochondrial ribosomes) shows that with the primer pairs presented, only co-amplification of 18S rDNA in V3 and 12S rDNA in V3-V4 is likely possible (usually only in samples not containing any bacterial DNA).

The functionality of the new primers was successfully tested on bacterial cultures. The sequences generated were evaluated with QIIME2 and BLAST and compared with the mass spectrometric determination of the bacteria.

Altogether, all primers match exactly for 75.4% of the human pathogenic species. The accuracy of the individual primers is 92.3-98.77% and 96.74-99.96% (1 mismatch allowed). When using all proposed primer pairs, all human pathogenic species should be amplifiable.

A comparison with the 'SILVA ribosomal RNA gene database' shows that the primers / primer pairs match on average for 3.5 and 6.4% more human pathogenic species.

Depending on the primer pair used, artifacts occurred at very different frequencies (0.11 - 90.17%) due to primer dimers, primer pair dimers or primer hairpins (relatively short sequences, the missing signal at the end interpreted as poly-G string due to the Illumina 2-dye chemistry).

Finally, the homologous small-subunit rRNA of primitive eukaryotes was integrated into the data set of bacterial 16S rRNA. The resulting phylogenetic tree of Bacteria supports a lineage of the mitochondria from a common ancestor with the Rickettsiales. A close relationship of the mitochondria as endosymbionts with the Rickettsiales as obligate intracellular (endo-)parasites seems quite plausible from an evolutionary point of view.

Keywords

16S rRNA; small subunit ribosomal RNA; Next Generation Sequencing; human pathogenic bacteria; 12S rRNA; mitochondrial DNA; endosymbiotic theory; Rickettsiales.

Introduction

In histological sections of formalin fixed paraffin embedded tissue, there are often inflammatory changes, the etiology of which is initially unclear. This is particularly true for inflammatory changes in lymph nodes with characteristic nodular changes (epitheloid cell granulomas) (Figure 1). In addition to bacterial pathogens (e.g. tuberculosis, leprosy, tularaemia, cat scratch disease, syphilis), the etiology of these granulomatous lymphadenitides also includes protozoa (e.g. leishmaniasis), fungi (e.g. histoplasmosis), inorganic materials (e.g. berylliosis), or is largely unknown (e.g. sarcoidosis).

The best way to identify bacterial pathogens in formalin-fixed paraffin-embedded (FFPE) tissue is by 16S rDNA sequencing[1]. The gene for the bacterial 16S rRNA, which codes for the ribosomal RNA of the small subunit (SSU) of the ribosome, is sequenced. Since protein synthesis takes place on the ribosomes and proteins (alongside DNA and RNA) are part of the basic components of the cell machinery, without which life seems inconceivable, certain regions of the 16S rRNA are highly conserved offering themselves as primer binding sites for amplification (PCR) and sequencing. Highly variable regions can be found between the conserved regions, the nucleotide sequence of which mostly enables the determination of bacteria to species level. Representatives of some genera, on the other hand, can only be assigned to higher taxonomic categories (genus or family level) (Martínez-Porchas et al., 2016), at least if only one variable region is sequenced.

Since protein synthesis is so fundamental to life, ribosomes are of course also found in eukaryotes, both in the cytoplasm and in mitochondria and chloroplasts. Correspondingly, homologous ribosomal RNAs are encoded both in the nuclear genome and in the mitochondrial and plastid genome, differing from the bacterial 16S rRNA in nucleotide number and sedimentation rate (Table 1).

As a Next Generation Sequencer (Illumina MiniSeq) is available at the Institute for Pathology and Molecular Pathology at the Salzkammergut-Klinikum Vöcklabruck primarily for mutation analysis of oncological samples, it was obvious to use this for 16S rDNA sequencing as well.

[1]) As DNA is sequenced anyway, the term 16S **rDNA** sequencing is correct, although the gene (DNA) codes for a ribosomal RNA (**rRNA**)!

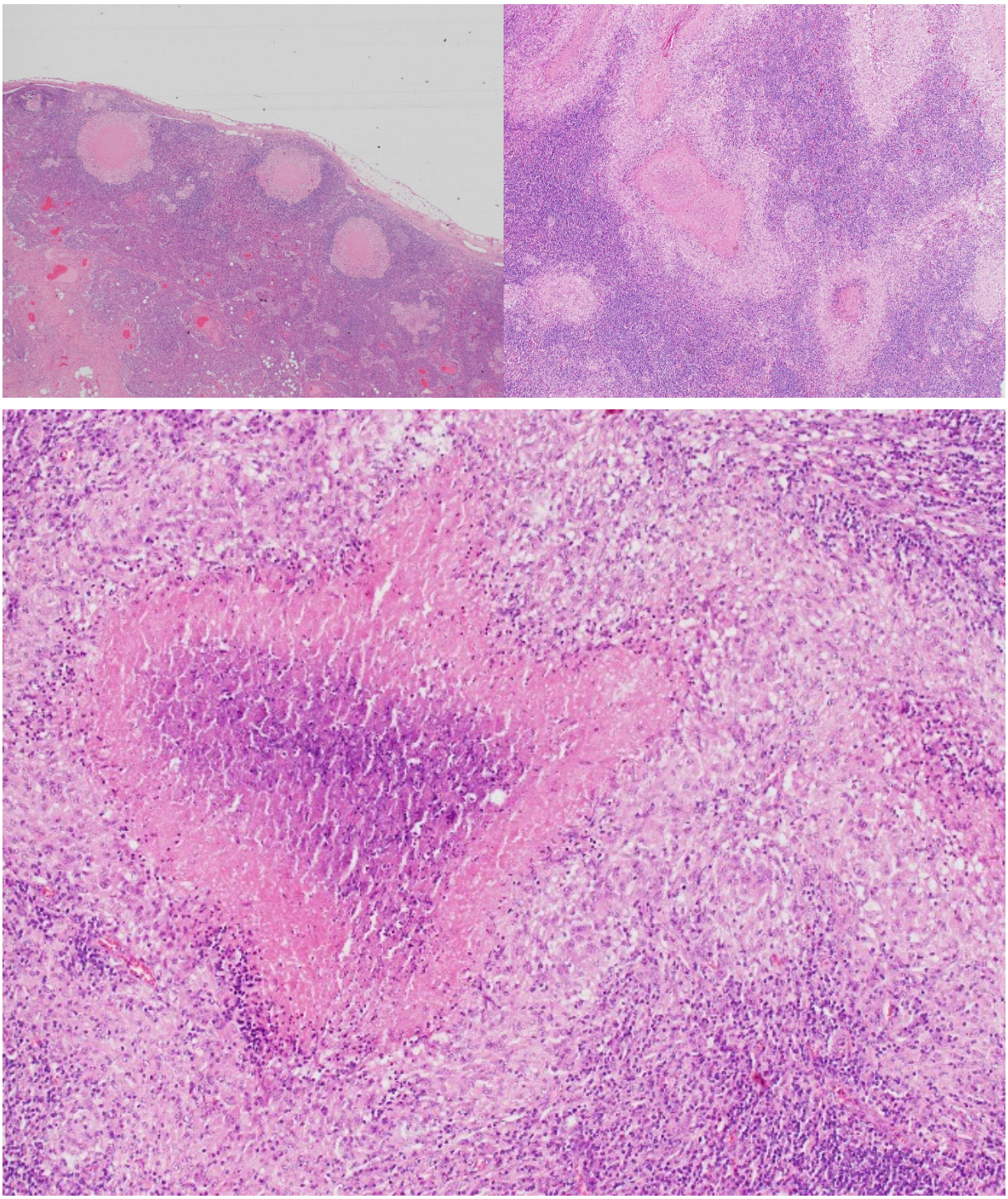

Figure 1: Epitheloid cell granulomatous lymphadenitis (tularaemia, rabbit fever)
Magnification: 12x, 24x, and 120x; The epitheloid cell granulomas with central necrosis in an inguinal lymph node are caused by the bacterium *Francisella tularensis* (detected with 16S rDNA sequencing); Institute of Pathology and Molecular Pathology, Salzkammergut-Klinikum Vöcklabruck.

However, no kit is available for this device for this application, since the read length (max. 2x150 bp) is only half as long as with the Illumina MiSeq (v3 reagents, 2x300 bp). Thus, 16S rDNA sequencing had to be re-established on the MiniSeq.

	homologous rRNA (SSU rRNA)	nucleotides
Prokaryotes	16S rRNA	ca. 1450
Eukaryotes (cytoplasmic)	18S rRNA	1869 (*Homo sapiens*) 1863 (*Reclinomonas americana*) 1800 (*Saccharomyces cerevisiae*) 1808 (*Arabidopsis thaliana*)
Mitochondria	12S rRNA 18S rRNA	954 (*Homo sapiens*) 1595 (*Reclinomonas americana*) 1648 (*Saccharomyces cerevisiae*) 1935 (*Arabidopsis thaliana*)

Table 1: homologous rRNA's of the small subunit (SSU) of the ribosomes

Reference sequences: *Homo sapiens*: NR_145819.1, NC_012920.1; *Reclinomonas americana*: AY117417.1, NC_001823.1; *Saccharomyces cerevisiae*: NR_132213.1, NC_027264.1; *Arabidopsis thaliana*: NR_139968.1, NC_037304.1

Material and methods

As 16S rDNA sequencing is mainly established for environmental sample materials (waters, soil, etc.) or microbiome analyzes (e.g. microbial composition of faeces) and the primers were mainly designed for cultivable species, a comprehensive (*In silico*) evaluation of the usability for human pathogenic bacteria has so far been lacking for the primers used (Sambo et al., 2018).

Database of human pathogenic bacteria

In the first step, a database (*MS Access*) of all human pathogenic bacteria was generated that was as complete as possible. The basis for this was initially "Classification of Bacterial Pathogens" (van Belkum, 2011). There were also some non-bacterial taxa (Archaea: 3; Eukaryota: 6 [Chlorophyta: 1, Fungi: 4, Metamonada: 1])[2] that have been excluded. Archaea (which were separated from the Bacteria in 1990 (Woese et al., 1990), but also belong to the prokaryotes) have no reliable evidence that they are pathogenic to humans (Aminov, 2013).

In this database, the taxonomic nomenclature was successively updated and the taxonomic categories (i.e. Domain, Phylum, Class, Subclass, Order, Suborder, Family, Genus, Species) supplemented as necessary. The basis for this was first the NCBI Taxonomy Database and UniProt website (UniProt, 2008), and finally the „List of Prokaryotic names with Standing in Nomenclature“ (Euzeby, 1997, Parte, 2014, Parte, 2018). Of 2643 taxa, 227 proved to be synonyms. 21 names could not be assigned to a known taxon. This left 2395 human pathogenic bacterial taxa in the database (**S1**).

Alignment of the 16S rDNA of human pathogenic bacteria

Then 16S rDNA sequences of all human pathogenic bacteria were downloaded from NCBI[3] (1 sequence per taxon) and stored in the database. Of only 32 taxa a sequence was not available. Thus, 16S rDNA sequences of 98.7% of all human pathogenic bacteria are present. This is not surprising since bacteria have very few morphological and biochemical characteristics, and therefore DNA sequences are preferred for phylogenetic classification of the taxa. The “sequence accession no.” of the 16S rRNA

[2]) Archaea: *Anaeroflexus maritimus*, *Ferroplasma acidarmanus*, *Thermococcus eurythermalis*, *Halococcus morrhuae*; Chlorophyta: *Prototheca wickerhamii*; Metamonada: *Lophomonas* sp.; Fungi: *Kloeckera apiculata* (= *Hanseniaspora uvarum*), *Apiosporina morbosa*, *Rhodotorula mucilaginosa*, *Issatchenkia orientalis*

[3]) Search term: e.g. ("Escherichia coli"[Organism] AND (16S rRNA[All Fields] OR 16S ribosomal RNA[All Fields]) AND ("0"[SLEN] : "1700"[SLEN]))

gene is therefore also mentioned for the type strain in “Bergey's Manual of Systematic Bacteriology” (Garrity (ed.), 2001-2012) and the „List of Prokaryotic names with Standing in Nomenclature“ (Euzeby, 1997, Parte, 2014, Parte, 2018).

The downloaded sequences were then aligned with the software MUSCLE (MUltiple Sequence Comparison by Log-Expectation), implemented in MEGA7 (Molecular Evolutionary Genetics Analysis 7.0) (Edgar, 2004b, Edgar, 2004a, Kumar et al., 2016). The result was pretty disillusioning (even after increasing the gap penalty to -800). Not only did the highly variable sequence areas prove (not unexpectedly) to be de facto not alignable, numerous obvious sequencing artifacts additionally appeared in the highly conserved areas. This was partially due to the fact that the longest sequences had been preferred when downloading (in the hope of covering the entire gene if possible); but these were often the ones with the most artifacts.

The required elimination of sequencing artifacts and optimization of the alignment was essentially manual work and exceedingly time-consuming. It was carried out using the PhyDE® (Phylogenetic Data Editor) software (Müller et al., 2010).

Sequencing artifacts

Sequence artifacts were particularly suspected in:

- insertion and deletion of individual nucleotides in conserved regions
- transition / transversion of individual nucleotides in highly conserved regions *(esp. C>T and G>A, which are often caused by deamination)*

In these cases, if possible, numerous 16S sequences of the corresponding taxon were aligned and checked whether the change only occurs in one or a few sequences, which was usually the case and proved the presence of an artifact, which was then corrected in the alignment (**S4**) and documented in the database (in a separate column).

If only one sequence of a taxon was available, the plausibility of the above changes was assessed on the basis of a comparison with species of the same genus or (in the case of species-poor genera) closely related genera.

Phylogenetic tree

In order to be able to assess the relationships, a phylogenetic tree was created based on the conserved areas of the 16S rRNA gene using the software FastTree (Price et al., 2009, Price et al., 2010) and RAxML (Randomized Axelerated Maximum Likelihood) (Stamatakis, 2014, Silvestro and Michalak, 2011) (Figure 2, **S2**). FastTree uses a minimum evolution algorithm and is very fast (and exact), RAxML (maximum

likelihood) much slower (but more precise). In contrast, distance-based methods such as neighbor joining are “quick & dirty”. In any case, they quickly produce a tree, but this often differs significantly from reality (Knoop and Müller, 2009).

Reference sequence of the type strain

If taxa appeared in an unusual place in the phylogenetic family tree (in contrast to the established phylogeny), their 16S sequence was replaced, if possible, with the reference sequence of the type strain in order to rule out a misdetermination (which by definition is impossible for type strains, because the type strain is that basis for the definition and naming of a taxon (Lapage et al., 1992)). However, it often turned out that the determination was not incorrect, but that the taxon had meanwhile (usually based on 16S sequence data) been transferred to another (correct) location in the phylogenetic system.

Small subunit ribosomal RNA (SSU rRNA, 16S rRNA)

Even before the genetic code had been completely deciphered, it was recognized that invaluable evidence of the evolutionary history of life on earth has been preserved in the molecules, especially nucleic acids and proteins (Zuckerkandl and Pauling, 1965).

Three domains: Bacteria, Archaea, Eukaryota

The importance of the nucleic acid sequence of ribosomal RNAs, in particular that of the small ribosomal subunit for phylogenetics, had soon been recognized and also contributed decisively to the splitting of life into 3 domains: Bacteria, Archaea, and Eukaryota (Woese and Fox, 1977, Woese et al., 1990, Fox et al., 1980, Gray et al., 1984). After the publication of the first (almost) complete bacterial 16S rRNA gene sequence from *Escherichia coli* (Brosius et al., 1978) and the polymerase chain reaction (PCR) (Mullis et al., 1986), primers for the 16S rRNA gene were published in rapid succession (Lane et al., 1985, Chen et al., 1989, DeLong et al., 1993, Watanabe et al., 2001), usually to determine the species composition of prokaryotes in environmental samples, but also to clarify the relationship (phylogeny) of bacteria and archaea. Furthermore, these primers were also evaluated in summary (Baker et al., 2003, Klindworth et al., 2013).

Variable and conserved regions

The basic three-dimensional structure of the rRNA of the small ribosomal subunit (SSU rRNA) was already clarified in 1981 (Woese et al., 1983, Gray et al., 1984, Stiegler et

al., 1981). This results in a sequence of conserved and variable regions of the sequence, with conserved regions predominantly in the area of central helical structures (stem-loops, hairpins), but variable regions on the periphery of the molecule (Yang et al., 2016b) (Figure 3).

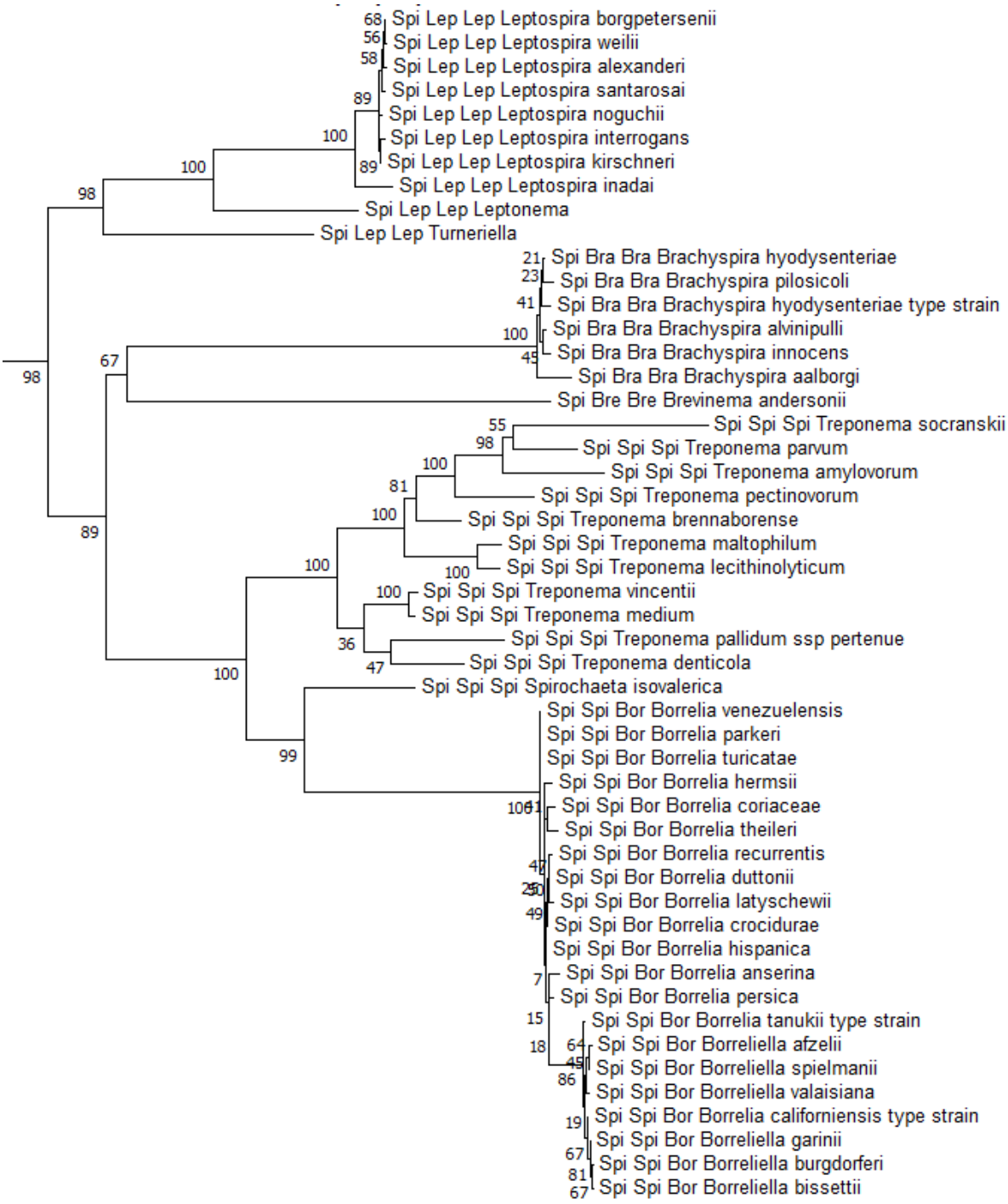

Figure 2: Bacteria - phylogenetic tree, detail (Spirochaetia)

orders Leptospirales, Brachyspirales, Brevinematales, and Spirochaetales;
created with RAxML based on conserved regions of the 16S rRNA gene (**S2**);
values shown are bootstrap values.

V1-V9	Domain	*stem-loops/hairpins*	conserved	semi-conserved	variable	alignment
V1 (1-103)	A; 5'	h1-3;4--5	8-62			8-62
A (h1), 5' (h4-6)	5'	**h6**		63-67	68-103	**63-67**
V2	5'	h6a,7-8		104-179		104-179
(104-312)	5'	**h9-10**,7			180-241	**219-241**
5' (h6a-12)	5'	h11-12		242-312		242-312
V3	5'	h13-15,4	313-405			313-405
(313-496)	5'	h16		406-436		406-436
5' (h13-17)	5'	**h17**			437-496	**437-450**
V4	5'; A	h18;3,19	497-574			497-574
(497-762)	C	**h20-22**			575-672	575-672
5' (h18), A (h3, h19), C (h20-h23a)	C	h23,23a,22,20		673-762		673-762
	C	h24	763-820			763-820
V5 (763-913)	C	h25-**26**			821-858	**821-835, 852-858**
C (h24-h26a), A (h19, h27)	C	h26a,25		859-877		859-877
V5/V6	A; 3'M	h19,27,2,28;29-31	878-984			878-984
V6 (914-1044)	3'M	h32		985-996		985-996
A (h2, h28), 3'M (h29, h31, h33)	3'M	**h33**			997-1044	
V7	3'M	h34-37	1045-1113			1045-1113
(1045-1173)	3'M	**h38-39**			1114-1140	**1114-1131**
3'M (h35-h40)	3'M	h39,40		1141-1173		1141-1173
V8 (1174-1389)	3'M	h34,32	1174-1241			1174-1241
3'M (h34, h32, h30, h41-h43)	3'M	**h41**			1242-1294	1242-1294
V8/V9	3'M; A; 3'm	h42,29,43;28;44(prox.)	1295-1407			1295-1407
V9	3'm	h44 (prox.)		**1408-1434**		**1408-1434**
(1390-1491)	3'm	**h44a (dist.)**			1435-1464	**1435-1448**
3'm (h44-h45)	3'm	h44		1465-1490		1465-1490
	3'm	h45	1491-1540			1491-1540

Table 2: Variable and conserved regions of the SSU rRNA

(Numbering of the nt according to the reference sequence of *E. coli* (J01859.1); delimitation of the regions: (Yang et al., 2016b); naming of the domains and stem-loops / hairpins: (Petrov et al., 2014, Gulen et al., 2016a); Delimitation of variable, conserved and alignable regions: own data (see **S3**, **S4**).

However, highly variable and highly conserved regions in the molecule are often closely adjacent, e.g. h17 (variable, V3) and h18 (highly preserved, V4). The most strictly conserved across all 3 domains (Bacteria, Archaea, Eukaryota) are the nt (505)515-536 in h18 (5‘ domain) and 1390-1407 in h28/44 (A/3‘m domain). They border directly on the core domain (A) or even have a share in it. With the primer combination S-D-Bact-0515-a-S-16 + S-*-Univ-1392-a-A-15, approximately 87.7/78.9/94.3 % of all Bacteria/Archaea/Eukaryota can be detected (Klindworth et al., 2013).

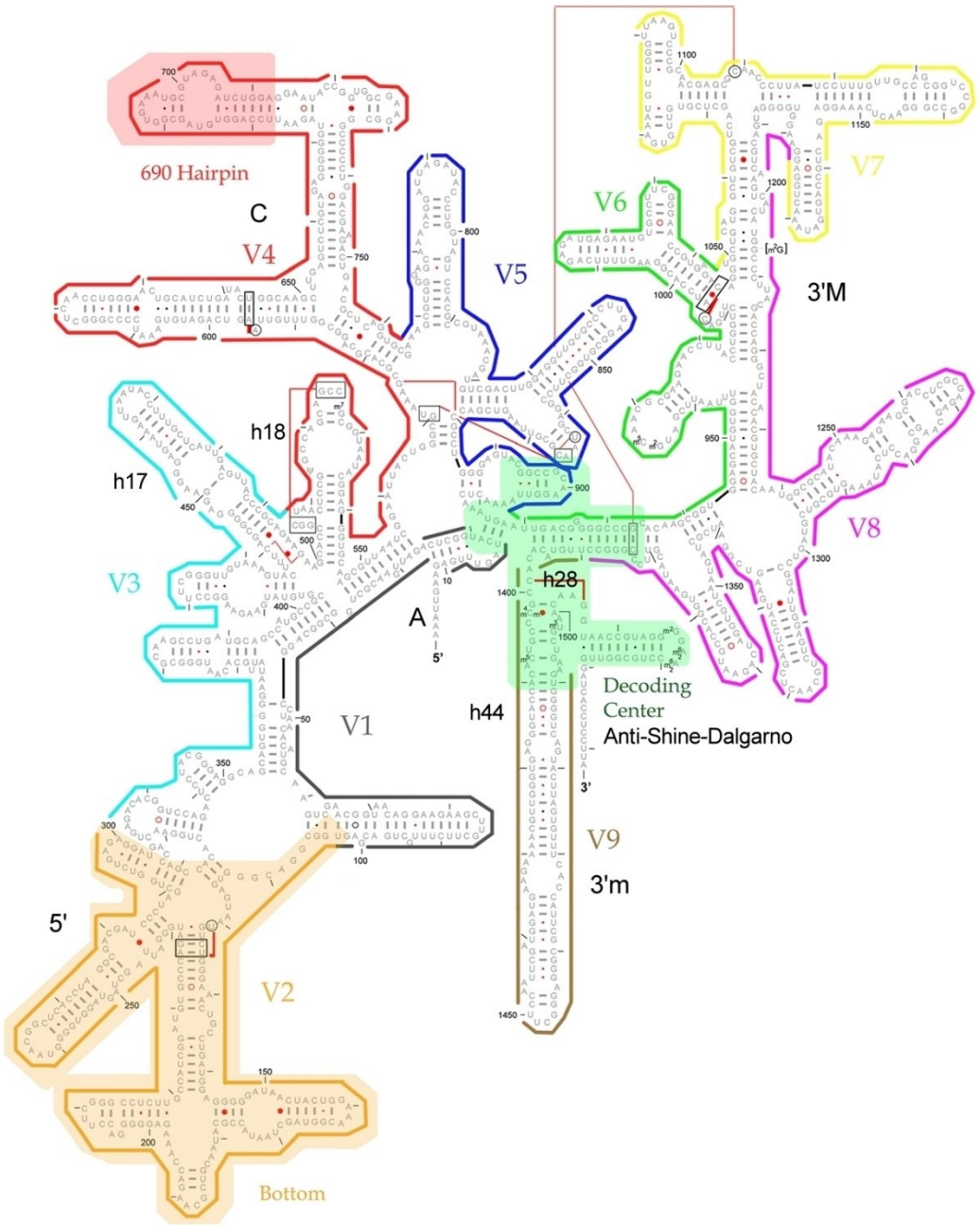

Figure 3: 2D structure of the 16S rRNA gene

Individual regions (V1-V9) are color coded; four peripheral domains (5', C, 3'M, 3'm) radiate from a central core domain (A). Modified from (Yang et al., 2016a) and (Gulen et al., 2016b).

Region h18 is the most conserved of all SSU rRNA's. It can still be found unaltered in the human 18S and with only 2 mismatches also in the (human) mitochondrial 12S rRNA!

The highly conserved region in h44 at the 3' end contains the anti-Shine-Dalgarno sequence (ACCUCCU, nt 1534-1540), which is part of the ribosomal binding site (RBS) in prokaryotes and to which the complementary Shine-Dalgarno sequence located in the 5'-UTR region of the mRNA binds and thus marks the starting point of the translation (Shine and Dalgarno, 1974).

The nomenclature and numbering of the variable regions as well as the understanding of the secondary structure of the molecule has changed in the meantime (Gulen et al., 2016a, Petrov et al., 2014).

9 variable regions (V1-V9)

We still differentiate between 9 variable regions (V1-V9, from 5 'to 3'). None of them can discriminate between all bacterial species, they are differently suitable for determination and phylogenetic analyzes (Chakravorty et al., 2007, Yang et al., 2016b). V1-V2 and V6 show the greatest heterogeneity, followed by V3-V4 (Coenye and Vandamme, 2003). Chakravorty et al. found different regions best suited to discriminate between *Staphylococcus* (V1), *Mycobacterium* (V2), and *Haemophilus* spp. (V3); V2-V3 is most suitable for discrimination at the genus level, V6 for the determination of most bacterial species (incl. *Bacillus anthracis*) except Enterobacteriaceae. The other regions are less useful (V9 not examined) (Chakravorty et al., 2007). Finally, the phylogenetic resolution is best of V4-V6, the least reliable in this regard are V2 and V8 (Yang et al., 2016b). For detection and exact determination of human pathogenic bacteria from FFPE, V1-V2 and especially V6 would be preferable, whereas V4 is usually used for this in commercially available test kits.

There are no clear boundaries between conserved and variable regions. The degree of conservation is also quite different (Table 2), Gray et al. already distinguished “universally conserved“ and “semi-conserved“ (Gray et al., 1984). The variable regions can be easily assigned to the 4 large peripheral domains (5′ domain, central domain, 3′ M domain [3′ major], and 3′ m domain [3′ minor]):

V1-V3 are in the 5‘ domain, V4-V5 in the C domain, V6-V8 in the 3’M domain, and finally V9 in the 3’m domain (Table 2). Since all these domains branch off from the core domain (A), they also reach the A domain centrally (Gulen et al., 2016a).

Primer design

Spreadsheet for sequence alignments

With the aligned 2363 sequences of the bacterial 16S rRNA gene, an Excel table was created (Figure 4, **S3**) that calculates how often the 4 nucleotides occur at each position of the alignment.

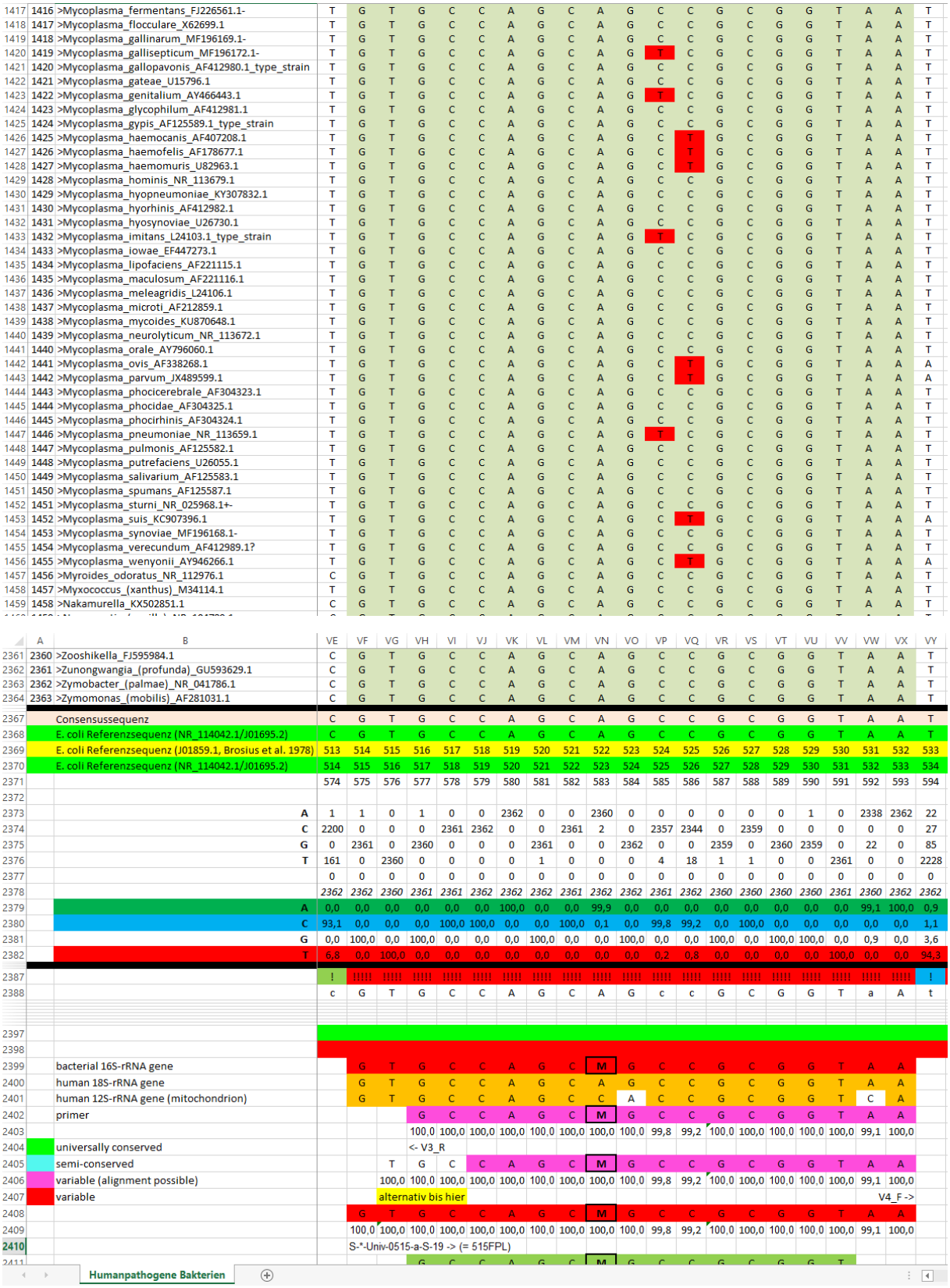

*Figure 4: Excel spreadsheet (**S3**) of human pathogenic bacteria*

above: mismatches in primer binding site V4_F in *Mycoplasma* spp. marked red;

below: primer binding site V4_F and V3_R (most conserved region of SSU rRNA!);

(explanations below on the left only refer to line 2397).

At the same time, the positions are marked where a nucleotide occurs in >99/95/90% (red/orange/green) of the cases (conserved regions), and positions where all 4 nucleotides occur are colored blue. The highly conserved regions, which are suitable as primer binding sites, are thus immediately visible.

Additionally, the numbering of the first published 16S rRNA sequence of *E. coli* (J01859.1), which is mainly used in older publications, is shown (marked in yellow), which differs from the NCBI reference sequences due to the deletion of one base (between nt 88 and 89) (green).[4]

The relatively short, highly conserved regions do not leave much scope for the primer design. It is therefore not surprising that the optimal primers found are located in the same regions as previously designed primers, which were published and evaluated in summary by Klindworth et al. (2013) (a total of 175 primers for SSU rRNA genes from bacteria, archaea, and eukaryotes). These also consistently applied to those the primer nomenclature already suggested by Alm et al. (1996), which significantly facilitates the comparison of different primers. It will also be kept here.

Since the MiniSeq sequences a maximum of 2x150 bp, primer binding sites were chosen that should not exceed an amplicon length of 300 bp. Considering that the primer sequence is also sequenced before the actual target, the primers should actually not be more than 250 bp apart so that paired-end reads overlap at least 12 bp. This was not always possible (Table 3). There is no highly conserved region between the variable regions V1 and V2 that could be used as a primer binding site, so that these two regions have to be amplified together. On the other hand, the variable regions V5-V6 and V7-V8 are sufficiently close together that they can also be amplified together; however, they then only overlap by 9 or 12 bp (in *E. coli*), which can cause problems in data analysis.

Care was taken to ensure that as few wobbles as possible were needed (which was not possible without compromise) and that the last bases at the 3' end were 100% preserved if possible; furthermore G or C (3 hydrogen bonds) should preferably be at the 3' end. The possibility of secondary structures (hairpins), primer self dimers and primer pair dimers was checked (PrimerSelect) and taken into account as far as possible, but it was usually not really avoidable.

[4]) The NCBI reference sequences (NR_114042.1, NR_024570.1, NR_112558.1) do not cover the entire 16S rRNA gene and contain some ambiguous nucleotides. The existing sequence parts match with the original sequence J01695.2, which is complete.

Primer pairs	Primer	Sequence (5'->3')	Primer length	Amplicon size	Target size	Overlap (nt)	Start (*E. coli*)	Stop (*E. coli*)	GC (%)
V1-V2		missing: 159-183							
V1-V2_F	S-D-Bact-0008-e-S-20	AGAGTTTGATYMTGGCTYAR	20	327	290	-25	8	27	50
V1-V2_R	S-D-Bact-0318-a-A-17	GDCCGTRTCTCAGTHCC	17				334	318	70,59
V3									
V3_F	S-D-Bact-0340-b-S-18	TCCTACGGGDGGCWGCAG	18	194	159	108	340	357	66,67
V3_R	S-D-Bact-0517-a-A-17	TTACCGCGGCKGCTGGC	17				533	517	70,59
V4									
V4_F	S-D-Bact-0519-a-S-15	CAGCMGCCGCGGTAA	15	288	252	14	519	533	66,67
V4_R	S-D-Bact-0786-a-A-21	GGACTACHVGGGTATCTAATC	21				806	786	47,62
V5-V6									
V5-V6_F	S-D-Bact-0785-a-S-21	GGATTAGATACCCBDGTAGTC	21	293	255	9	785	805	47,62
V5-V6_R	S-D-Bact-1061-a-A-17	CRRCACGAGCTGACGAC	17				1077	1061	64,71
V7-V8									
V7-V8_F	S-D-Bact-1063-a-S-18	CGTCAGCTCGTGYYGTGA	18	290	250	12	1063	1080	61,11
V7-V8_R	S-D-Bact-1331-a-A-22	GATTACTAGCRAHTCCRVCTTC	22				1352	1331	45,45
V8									
V8-V9_F	S-D-Bact-1176-a-S-18	AGGAAGGHGDGGAYGACG	18	177	137	125	1176	1193	61,11
V7-V8_R	S-D-Bact-1331-a-A-22	GATTACTAGCRAHTCCRVCTTC	22				1352	1331	45,45
V9									
V9_F	S-D-Bact-1390-a-S-18	TTGYACWCACYGCCCGTC	18	118	79	184	1390	1407	61,11
V8-V9_R	S-D-Bact-1487-a-A-21	TACCTTGTTACGACTTMRYCC	21				1507	1487	47,62
V8-V9		missing: 1327-1356							
V8-V9_F	S-D-Bact-1176-a-S-18	AGGAAGGHGDGGAYGACG	18	332	293	-30	1176	1193	61,11
V8-V9_R	S-D-Bact-1487-a-A-21	TACCTTGTTACGACTTMRYCC	21				1507	1487	47,62

Table 3: Primers optimized for human pathogenic bacteria
lab names and nomenclature according to Alm et al. (1996);
wobbles marked red; amplicon and target sizes given for *E. coli*;
overlap refers to primer pairs and paired-end reads (2x151 bp read length);
GC content referring to primer sequences without wobbles.

Melting temperature and optimal annealing temperature

A catalog of possible primers was created (**S5**) and, in addition to primer length (nt) and coordinates of the primer binding site (*E. coli*), the melting temperatures were determined using various programs (PrimerSelect) and online tools (NCBI PrimerBLAST, OligoCalc, eurofins). However, the calculated values are not very reliable, they spread over a very wide range (Table 4).

The product length (amplicon size, determined with NCBI Primer-BLAST) spreads - except for V4 - over a surprisingly wide range (maxima für V1-V2: 321 bp, V3: 194 bp, V4: 288 bp, V5-V6: 293 bp, V7-V8: 291 bp, V8-V9: 332 bp). The maxima mostly agree with *E. coli*.

Primer	>>	T_m^1	T_m^2	T_m^3	T_m^4	T_m^5	T_m^6	T_m^7	T_a^1	T_a^2	Primer pair dimers	Self compl.	Self 3' compl.
V1-V2													
V1-V2_F	>>	56,92	57,3	58,4	52,5	51,8	48,3	52,9	59,34	59,46	4	7.00	5.00
V1-V2_R	>>	59,35	60	59,8	51,3	54,3	50,3	58,0	58,74	58,86	4	4.00	3.00
V3													
V3_F	>>	60,76	60,5	60,8	56,3	54,9	53,4	58,9	58,28	58,72	6 !	5.00	2.00
V3_R	>>	63,26	60	59,8	57,9	54,3	58	61,6	57,68	58,12	6 !	6.00	2.00
V4													
V4_F		65,04	60,5	60,8	58,58	54,9	60,8	62,8	56,44	56,55	4	6.00	1.00
V4_R	>>	54,04	57,9	59,5	49,5	52,4	44	49,9	60,04	60,15	4	6.00	0.00
V5-V6													
V5-V6_F		54,04	57,9	59,5	49,5	52,4	44	49,9	59,69	59,86	OK	6.00	6.00
V5-V6_R	>>	57,82	57,6	57,3	53,6	51,9	47	55,2	57,89	58,06	OK	4.00	2.00
V7-V8													
V7-V8_F	>>	59,75	58,2	58,4	55,7	52,6	50,7	56,7	58,34	58,54	4	4.00	2.00
V7-V8_R	>>	56,53	58,4	60,1	52,9	53	49,5	51,8	60,14	60,34	4	4.00	2.00
V8													
V8-V9_F	>>	58,27	58,2	58,4	52,2	52,6	51,7	55,9	57,17	57,77	8 !!	2.00	2.00
V7-V8_R	>>	56,53	58,4	60,1	52,9	53	49,5	51,8	58,97	59,57	8 !!	4.00	2.00
V9													
V9_F	>>	59,97	58,2	58,4	51,88	52,6	52,2	57,1	55,02	56,57	OK	6.00	1.00
V8-V9_R	>>	57,33	57,9	59,5	51,9	52,4	48,6	52,8	56,22	57,77	OK	3.00	0.00
V8-V9													
V8-V9_F	>>	58,27	58,2	58,4	52,2	52,6	51,7	55,9	58,31	58,49	4	2.00	2.00
V8-V9_R	>>	57,33	57,9	59,5	51,9	52,4	48,6	52,8	59,51	59,69	4	3.00	0.00

Table 4: Primers optimized for human pathogenic bacteria - melting temperature (T_m) and optimal annealing temperature (T_a)

T_m^1: NCBI PrimerBLAST; T_m^2: eurofins; T_m^{3-5}: OligoCalc (T_m^3: salt adjusted; T_m^4: nearest neighbor; T_m^5: basic); T_m^6: PrimerSelect (Plasterer, 1997)*; T_m^7: Microsynth.*

T_a^1: primer without adapter sequence; T_a^2: primer with adapter sequence.

The optimal annealing temperature was calculated using the following formula (Mülhardt, 2013):

$T_a^{opt} = 0,3 * T_m^{Primer} + 0,7 * T_m^{Produkt} - 14,9$

(T_a^{opt} = optimal annealing temperature, T_m = melting temperature (of primer or product))

The values of T_m and T_a are colored depending on the temperature (lower: green, higher: red).

Primer pair dimers for primers including adapter (linker) sequences!

Primer combinations

The most suitable primer combinations for the various variable regions were then selected on the basis of the amplicon length, melting temperatures and the optimal annealing temperature (Table 3, 4). A comparison with the primer catalog by

Klindworth et al. (2013) shows that only one of the selected primers is included (V5-V6_R), some bind in regions that are not represented there (V8-V9_F, V7-V8_R, V9_F). Otherwise they differ at least in wobble positions or primer lengths.

However, the mostly used standard primers for 16S rDNA sequencing (515F and 806R) are virtually identical to the primers presented here for the variable region V4, they only differ in length and should amplify human pathogenic bacteria just as well, the additional or omitted bases fitting 100%.

Extended primer catalog

The primer catalog of Klindworth et al. (2013) was extended by the primers presented here, some minor inaccuracies and errors corrected (**S6**). The primers for Archaea have not been checked.

Coamplification of human 18S (cytoplasmic) and 12S rDNA (mitochondrial)

Since DNA of the eukaryote Homo sapiens is always contained in clinical samples, possible coamplification of the homologous human 18S and 12S rDNA is of great importance. 18S rRNA is an integral part of the SSU in cytoplasmic ribosomes, whereas 12S rRNA is in mitochondrial ribosomes.

The sequence comparison shows that with the primer pairs presented, a co-amplification in the regions V1-V2, V5-V6, V7-V8, V8, V9, and V8-V9 is not possible. **However, the co-amplification of 18S-rDNA in V3 and 12S-rDNA in V3-V4 is quite possible** (forward primer fitting about half [3'], reverse primer totally matching). Less likely (but not excluded) is a co-amplification of 12S rDNA in V3, 18S in V3-V4 as well as 18S and 12S rDNA in V4 (**S7**).

Due to competitive effects, co-amplification of human SSU rDNA occurs virtually only if no bacterial SSU rDNA is contained in the sample (Kashofer, pers. comm.)

16S rDNA sequencing of bacterial cultures on Illumina MiniSeq™

The functionality of the primers presented here was first tested as part of a bachelor's thesis[5] on bacterial cultures. 17 bacterial cultures of different species[6] were selected

[5]) carried out by Evelyn Lang, supervised by the author; bachelor theses of the FH Gesundheitsberufe OÖ are not published

[6]) *Staphylococcus epidermidis*, *S. aureus*, *S. saprophyticus*, *Enterococcus faecium*, *Nocardia farcinica*, *Eikenella corrodens*, *Moraxella catarrhalis*, *Pasteurella multocida*, *Pseudomonas aeruginosa*, *Haemophilus influenzae*, *Stenotrophomonas maltophilia*, *Klebsiella pneumoniae*, *Citrobacter koseri*, *Campylobacter coli*, *Bacteroides fragilis*, *B. thetaiotaomicron*

at random, which had been determined at species level by mass spectrometry using Bruker Maldi-TOF.

DNA extraction

Automated DNA extraction from bacterial cultures was carried out using Magnetic Silica particles on a bioMérieux NucliSENS® easyMAG® platform. The eluate was diluted to a final concentration of 0,05 ng/µl DNA (~ 25.000 template molecules/µl).

Library Preparation

Next-generation sequencing (NGS) library preparation involves generating a collection of DNA fragments for sequencing.

Amplicon-based NGS libraries typically involve 2 PCR steps: The fragments are amplified by PCR using target specific primers with an attached linker sequence at the 5'-end. After purification (magnetic beads) adapters and index sequences (for pooling) are attached in a second PCR, the primers being complementary to the linker sequence of the 1st PCR (Figure 5).

After a second cleaning step the library is ready for normalization, pooling, and sequencing (protocol **S8**).

The adapter sequences at each end of the library fragments are binding to complementary adapter sequences on the flow cell, enabling bridge amplification and cluster generation. The index sequences (usually 6-8 nt long) are allowing for pooling of different samples and sequencing them together on a flow cell. The linker sequences additionally serve as binding site for sequencing primers (index sequencing and target sequencing).

1. Target-specific PCR

PCR primers specific for the genomic regions of interest (target region) contain a linker sequence attached to the 5'-end:

PRIMER	SEQUENCES
Forward	TCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNNNNNNNNNNN
Reverse	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNNNNNNNNNNNNN

The 1st PCR was carried out using **4µl bacterial genomic DNA (= 0,2 ng** ~ 100.000 template molecules**)** and FastStart™ High Fidelity PCR System (Merck) according to manufacturer's protocol (H_2O 15.75 µl, 10x Buffer 2.5 µl, 10mM dNTPs 0.5 µl, 10µM amplicon-specific primers F+R 1 µl each, FastStart HiFi Polymerase (5U/µl) 0.25 µl). The PCR protocol (total reaction volume 25 µl) included an initial denaturation step (94°C, 3 min) for activation of hot start polymerase, 25 cycles of denaturation (94°C,

15 sec), annealing (60°C, 45 sec), and extension (72°C, 1 min); no final extension step before cooling (4°C).

- **Purification**

Purification of the PCR product was carried out using Agencourt AMPure XP beads according to manufacturer's protocol.

2. Indexing PCR/library amplification

PCR primers are composed (5'->3') of adapter (P5/P7) sequence, index (i5/i7) sequence, and linker sequence (as above):

PRIMER	SEQUENCES
Forward	AATGATACGGCGACCACCGAGATCTACAC[i5comrev]ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Reverse	CAAGCAGAAGACGGCATACGAGAT[i7comrev]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Forward (P5_I_Rd1SP); Reverse (P7_I_Rd2SP); Rd1/2SP = Read 1/2 Sequencing Primer

The 2nd PCR was carried out using **5µl of PCR product** and the same reagents and protocol (but H_2O 14.75 µl).

product after 1st PCR:
TCTTTCCCTACACGACGCTCTTCCGATCT[FORWARD][AMPLICON][R_COMPL]AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
AGAAAGGGATGTGCTGCGAGAAGGCTAGA[F_COMPL][A_COMPLE][ESREVER]TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG

product after 2nd PCR:
AATGATACGGCGACCACCGAGATCTACAC[i5comrev]ACACTCTTTCCCTACACGACGCTCTTCCGATCT[FORWARD][AMPLICON][R_
COMPL]AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7forwar]ATCTCGTATGCCGTCTTCTGCTTG
TTACTATGCCGCTGGTGGCTCTAGATGTG[i5revers]TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA[F_COMPL][A_COMPLE][ES
REVER]TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7comfor]TAGAGCATACGGCAGAAGACGAAC

product after 2nd PCR (indices D501+D701:
AATGATACGGCGACCACCGAGATCTACAC[AGGCTATA]ACACTCTTTCCCTACACGACGCTCTTCCGATCT[FORWARD][AMPLICON][R_
COMPL]AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[ATTACTCG]ATCTCGTATGCCGTCTTCTGCTTG
TTACTATGCCGCTGGTGGCTCTAGATGTG[TCCGATAT]TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA[F_COMPL][A_COMPLE][ES
REVER]TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[TAATGAGC]TAGAGCATACGGCAGAAGACGAAC

Figure 5: Structure of the library for paired-end sequencing on Illumina platforms
sequences on Minus strand in *italics*.
Oligonucleotide sequences © 2019 Illumina, Inc. All rights reserved.

- **Purification**

as above

- **Determination of product concentration**

The Product concentration was determined by fluorescence spectroscopy (Qubit® 2.0, broad range assay; Life Technologies, Grand Island, NY). A Bioanalyzer has not been available at that time.

- **Normalization and pooling**

All samples were normalized to a final concentration of 4 ng/µl with resuspension buffer (Illumina).

Sequencing

Samples for 16S rDNA sequencing usually were run together with routine clinical samples for analysis of tumor mutations and gene fusions.

1 µl of each normalized 16S library was used for sequencing on the Illumina MiniSeq platform (GenerateFastQ mode).

Data analysis

Bioinformatics analysis was performed with QIIME 2 2019.7 (Bolyen et al., 2019). Raw sequence data had already been demultiplexed by MiniSeq Local Run Manager, and was quality-filtered using the q2-demux plugin followed by denoising with DADA2 (Callahan et al., 2016) (via q2-dada2). All amplicon sequence variants (ASVs) were aligned with mafft (Katoh et al., 2002) (via q2-alignment) and used to construct a phylogeny with fasttree2 (Price et al., 2010) (via q2-phylogeny). Taxonomy was assigned to ASVs using the q2-feature-classifier (Bokulich et al., 2018) classify-sklearn and fit-classifier-naive-bayes taxonomy classifier against the sequences of human-pathogenic bacteria presented here (Pedregosa et al., 2011). (Scripts **S9**-**S11**)

The generated bacterial sequences were then compared with the NCBI database using BLAST (Altschul et al., 1990, Altschul et al., 1997, Zhang et al., 2000, Morgulis et al., 2008). If forward and reverse read overlapped less than 12 bp and the reads were not merged by QIIME2, this was done separately with the two reads, otherwise with the merged reads.

After trimming the primer sequences (which are sequenced initially in the read), there was virtually always a series of sequence entries that matched 100% with the generated sequence.

The results are presented in the next chapter.

Results

In silico analysis of the selected primer pairs

The selected primers were analyzed *in silico*.

Primer matches / mismatches

A total of 581 of 2363 species (24.6%) have mismatches with the selected primers. Among them are also some more important genera such as *Actinomyces*, *Corynebacterium*, *Nocardia*, *Rickettsia (prowazekii)*, *Bordetella (bronchiseptica)*, several Enterobacteriaceae (incl. *Klebsiella pneumoniae*, *Morganella morganii*), *Coxiella (burnetii)*, *Legionella*, *Xanthomonas*, *Acinetobacter*, *Pseudomonas*, *Francisella (tularensis)*, *Leptospira*, *Borrelia*, *Treponema*, *Mycoplasma*, and *Ureaplasma*.

The fit of the primers is between 92.3 and 98.77% (perfect match; average 96.31%) and 96.74-99.96% (1 mismatch allowed; average 99.35%) (Table 5; Figure 6) , values between 88.45 and 95.73% (mean 93.26%) or 95.51-99.66% (mean 98.39%) apply to the primer pairs (Table 6; Figure 7).

Mismatches are usually limited to single primers, mostly only 1 or 2 mismatches are present (48 or 27.9%).

If sequencing the 3 variable regions V4 + V5-V6 + V7-V8 with the primer pairs presented here, one can be pretty sure to use a suitable primer pair for each human-pathogenic bacterium[7]. There are mismatches for a few taxa that usually never occur in clinical routine: *Sulfurihydrogenibium*, *Bellilinea*, *Leptolinea*, *Dialister*, *Brachyspira*, *Opitutus*, *Coraliomargarita*, *Akkermansia*, *Rubritalea*, *Prosthecobacter*, *Verruco-microbium.* In general, there are also primer pairs where the mismatches only appear at the 5' end, so that they are almost certainly usable as well.

Mismatches in all 6 primer combinations (V1-V2, V3, V4, V5-V6, V7-V8, V8-V9) are only found in *Anaerolinea thermophila* and numerous *Mycoplasma* spp.; the former only has a mismatch at the 5' end of the forward primer in V3, the relevant *Mycoplasma* spp. at the 5 'end of the reverse primer of V5-V6; all of them should therefore be amplifiable as well.

Comparison with SILVA ribosomal RNA gene database

In order to be able to assess whether there are actually more significant differences in the area of the primer binding sites in the bacterial 16S rRNA gene between clinical

[7]) V4 is usually used for 16S rDNA sequencing, and V6 is according to Chakravorty et al. well suited for most bacterial species (incl. *Bacillus anthracis*) except Enterobacteriaceae.

and environmental samples, the proportion of taxa (database entries) with primer matches (perfect matches / 1 mismatch allowed) in the data set of human-pathogenic bacteria presented here was compared with the SILVA ribosomal RNA gene database (Quast et al., 2012, Yilmaz et al., 2013).

Both the individual primers (using the TestProbe function) and the primer pairs (using TestPrime) were tested against the SILVA database (SSU r132, Sequence Collection RefNR) (Klindworth et al., 2013) (Table 5, 6).

The difference in primer matches (Table 5) between the two databases fluctuates between -0.9 and -8.2% (mean -3.5%), with the exception of primer V8-V9_R, where there results an essentially lower percentage of primer matches in the SILVA database. The reason is not entirely clear. It should be noted that this region near the 3' end of the 16S rRNA gene is poorly covered in both databases (in the database of human-pathogenic bacteria by approx. 16% compared to V4_R), but this cannot explain the unusually large difference of over 30%.

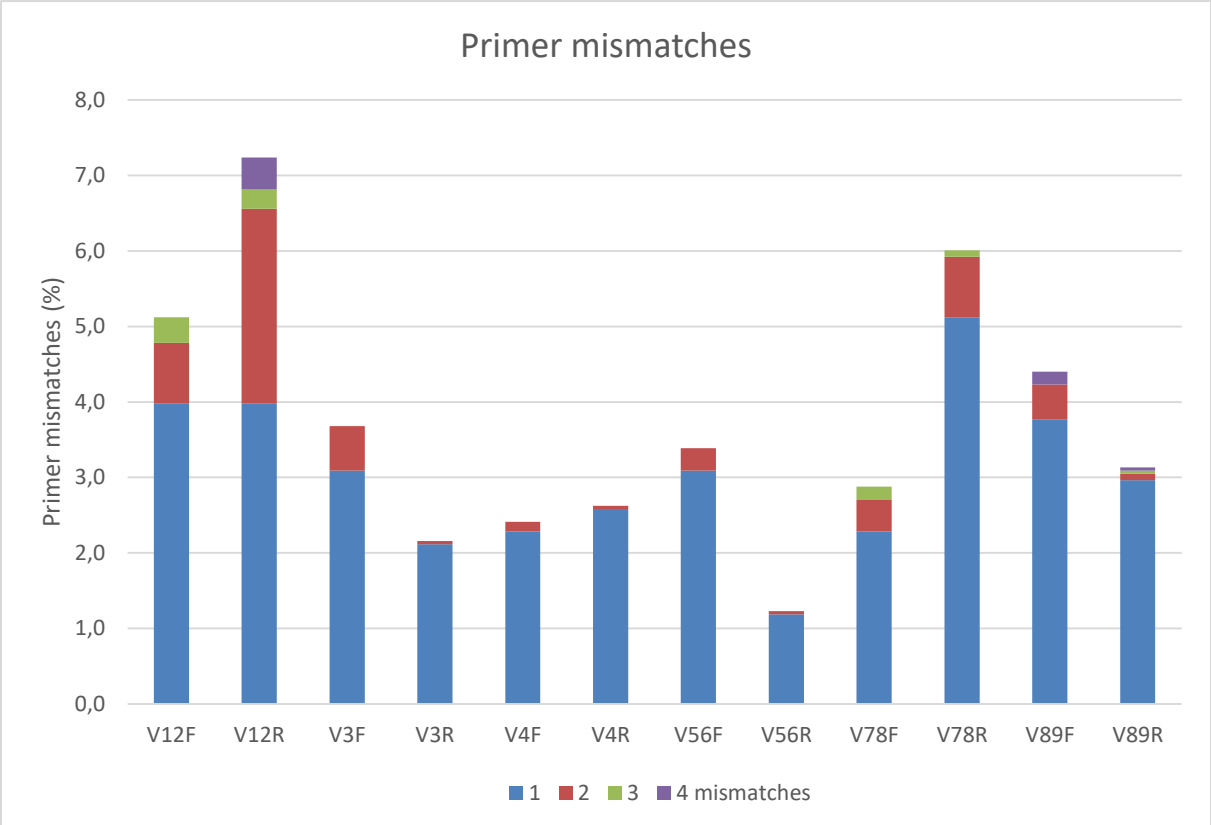

Figure 6: Primer mismatches of human pathogenic species

percentage of species with 1-4 mismatches per primer is color-coded;

V12F = V1-V2_F etc.

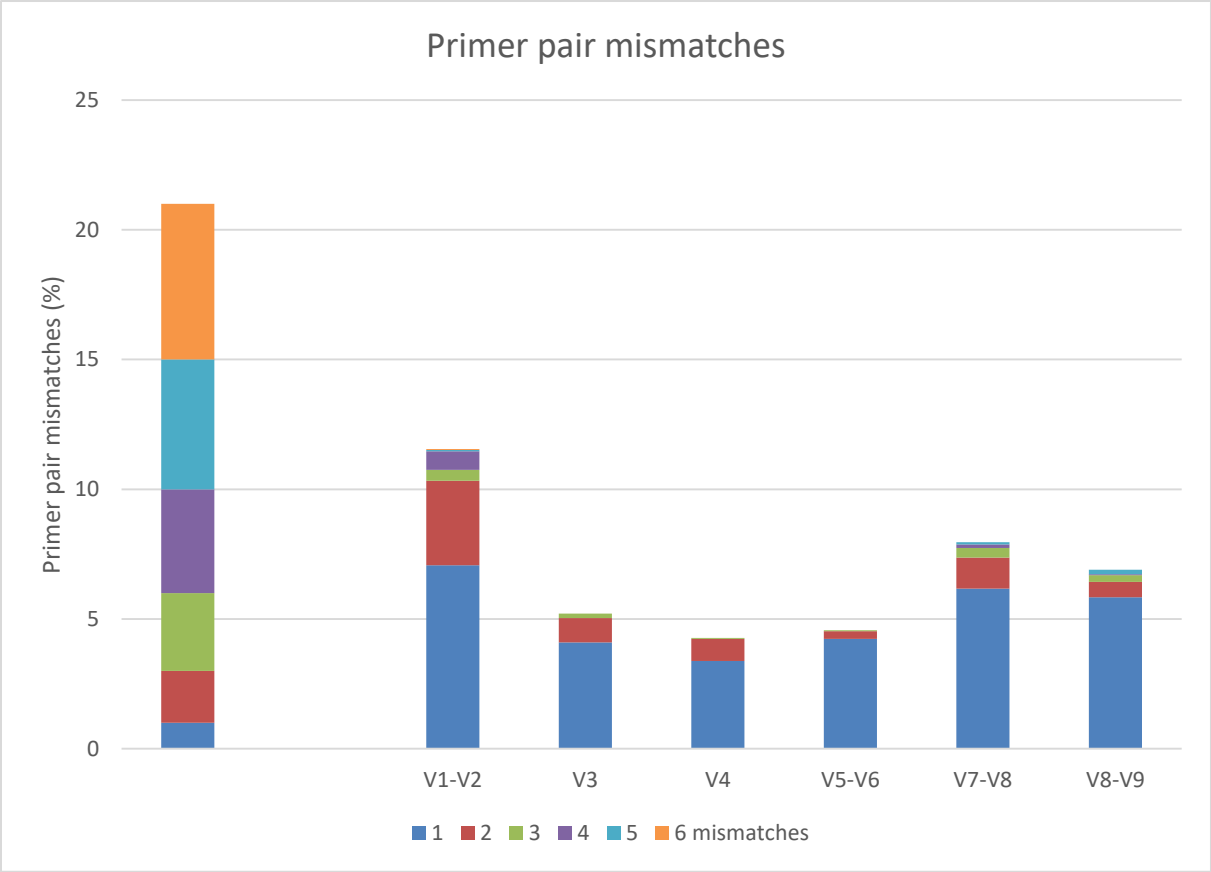

Figure 7: Primer pair mismatches of human pathogenic species

percentage of species with 1-6 mismatches per primer pair is color-coded;

V1-V2 = V1-V2_F + V1-V2_R etc.

In order to come closer to an explanation or to optimize this primer, it was determined to which nucleotides the mismatches are distributed if 1 mismatch is allowed (Figure 8Figure 8). However, there is little scope for primer optimization in this area, since the primer already contains 3 wobbles (especially close to the 3' end).

A	389	1665	0	439	386	0	0	77	27	62	1388	1901	0	0	226	0	0	47	33	130	0
C	21	499	1054	0	47	4	5	15	54	0	128	107	83	92	0	16	9	33	82	193	41
G	0	0	0	295	0	74	81	0	18	242	0	2195	82	62	25	94	168	0	0	42	104
T	99	217	28	0	0	16	15	25	0	70	17	0	36	23	17	29	28	28	56	0	178
rc	G	G	Y	R	M	A	A	G	T	C	G	T	A	A	C	A	A	G	G	T	A

Figure 8: Primer mismatches (V8-V9_R) to the SILVA database

Below is the sequence of the plus strand (reverse complement to the primer sequence)!

Positions with most mismatches are marked red (separated according to different nucleotides).

If one compares the primer pair matches (Table 6), there are somewhat clearer differences between -3.0 and -10.8% (mean -6.4%); throughout, the primers (as expected) are less suited to the sequences in the SILVA database than to the human pathogenic bacteria.

		V12F		V12R		V3F		V3R		V4F		V4R	
	mm	path.	Silva	path.	Silva	path.	Silva	path.	Silva	path.	Silva	path.	Silva
Archaea			0,00%		12,61%		0,00%		58,06%		57,87%		90,27%
Bacteria	0	94,88%	86,70%	92,76%	88,36%	96,32%	92,30%	97,84%	94,41%	97,59%	94,22%	97,38%	90,81%
Eukaryota			0,00%		0,01%		0,00%		91,08%		90,86%		0,01%
Archaea			0,05%		21,36%		3,06%		97,45%		97,30%		96,93%
Bacteria	0-1	98,86%	93,79%	96,74%	94,88%	99,41%	97,49%	99,96%	97,93%	99,87%	97,80%	99,96%	96,97%
Eukaryota			0,04%		0,01%		0,11%		96,37%		96,35%		0,34%
	0	2242		2192		2276		2312		2306		2301	
	1	94		94		73		50		54		61	
	2	19		61		14		1		3		1	
	3	8		6									
	4			10									
	tot.	121		171		87		51		57		62	
	diff.		-8,2%		-4,4%		-4,0%		-3,4%		-3,4%		-6,6%
			-5,1%		-1,9%		-1,9%		-2,0%		-2,1%		-3,0%

		V56F		V56R		V78F		V78R		V89F		V89R	
	mm	path.	Silva	path.	Silva	path.	Silva	path.	Silva	path.	Silva	path.	Silva
Archaea			91,07%		2,09%		2,38%		0,00%		14,58%		20,92%
Bacteria	0	96,61%	91,20%	98,77%	96,01%	97,12%	95,38%	93,99%	85,88%	95,60%	92,08%	96,87%	60,62%
Eukaryota			0,05%		0,01%		0,00%		0,01%		0,01%		0,00%
Archaea			97,01%		24,74%		55,19%		16,51%		18,65%		43,41%
Bacteria	0-1	99,70%	96,92%	99,96%	98,67%	99,41%	98,49%	99,11%	95,38%	99,37%	98,08%	99,83%	66,00%
Eukaryota			0,62%		6,08%		0,01%		0,01%		0,02%		0,46%
	0	2283		2334		2295		2221		2259		2289	
	1	73		28		54		121		89		70	
	2	7		1		10		19		11		2	
	3					4		2				1	
	4									4		1	
	tot.	80		29		68		142		104		74	
	diff.		-5,4%		-2,8%		-1,7%		-8,1%		-3,5%		-36,2%
			-2,8%		-1,3%		-0,9%		-3,7%		-1,3%		-33,8%

Table 5: 16S rRNA gene primer matches compared (human pathogenic - SILVA database) primer matches (top: perfect matches; bottom: 1 mismatch allowed) in database of human-pathogenic bacteria and SILVA ribosomal RNA gene database (SSU r132, Sequence Collection RefNR); number of primer mismatches (mm) of human pathogenic bacteria per primer (binding site); difference in primer matches (perfect matches / 1 mismatch allowed) between the two databases.

		V1-V9	V1-V2		V3		V4		V5-V6		V7-V8		V8-V9	
	mm		path.	Silva	path.	Silva	path.	Silva	path.	Silva	path.	Silva	path.	Silva
Archaea				0,00%		0,00%		53,20%		1,80%		0,00%		1,60%
Bacteria	0	75,41%	88,45%	77,60%	94,79%	88,20%	95,73%	86,60%	95,43%	88,00%	92,04%	82,90%	93,10%	56,70%
Eukaryota				0,00%		0,00%		0,00%		0,00%		0,00%		0,00%
Archaea				0,00%		2,90%		94,60%		23,80%		15,30%		3,40%
Bacteria	0-1	87,22%	95,51%	89,20%	98,90%	95,90%	99,11%	95,00%	99,66%	95,80%	98,22%	94,90%	98,94%	65,30%
Eukaryota				0,00%		0,10%		0,30%		0,00%		0,00%		0,00%
Diff.	0	1782	2090		2240		2262		2255		2175		2200	
	1	279	167		97		80		100		146		138	
	2	162	77		22		20		7		28		14	
	3	47	10		4		1		1		9		6	
	4	46	17								3		1	
	5	14	1								2		4	
	6	11	1											
	7	6												
	8	3												
	9	5												
	10													
	11	4												
	12													
	13													
	14	2												
	15	1												
	16	1												
	tot.	581	273		123		101		108		188		163	
diff.	0			-10,8%		-6,6%		-9,1%		-7,4%		-9,1%		-36,4%
	1			-6,3%		-3,0%		-4,1%		-3,9%		-3,3%		-33,6%

Table 6: 16S-rRNA gene primer pairs compared (human pathogenic - SILVA database)
primer pair matches (top: perfect matches; bottom: 1 mismatch allowed) in database of human pathogenic bacteria and SILVA ribosomal RNA gene database (SSU r132, Sequence Collection RefNR);
Number of primer pair mismatches (mm) of human pathogenic bacteria;
Difference in primer matches (perfect matches / 1 mismatch allowed) between the two databases.

Primer specificity

The primer combinations were also tested with Primer-BLAST against RNA reference sequences from *Homo sapiens*. There were no disturbing matches with the exception of the already known possible coamplification of human 18S rRNA with primers for V4.

16S rDNA sequencing of bacterial cultures

Selected primer pairs (V1-V2, V3, V4, V5-V6, V7-V8, V8-V9)

First, it was tested whether *Pseudomonas aeruginosa* (the most DNA extract of this pathogen was present) can be amplified and sequenced with all selected primer combinations (Table 3). This was easily possible (Table 7).

	forward						reverse					
Primer	**input**	**denoised**	**non-chim.**	**f**	**f(Ψ)**	**%(Ψ)**	**input**	**denoised**	**non-chim.**	**r**	**r(Ψ)**	**%(Ψ)**
V1-V2	277846	266175	57093	27376	29662	52,00%	277846	259059	38700	6127	32528	84,15%
V3	408080	391386	219201	10494	208707	95,21%	408080	373247	182100	17901	164199	90,17%
V4	169666	166779	165500	164946	114	0,07%	169666	160097	104579	104068	119	0,11%
V5-V6	232284	221811	86127	9193	76934	89,33%	232284	221359	84275	33401	50874	60,37%
V7-V8	187128	180974	102840	96969	5871	5,71%	187128	180113	33112	30331	2781	8,40%
V8-V9	230285	223256	84189	76677	7474	8,88%	230285	217272	148050	134620	13365	9,03%

						BLAST	*P. aeruginosa*			*Pseudomonas* sp.		
Primer	**det (f)**	**len(f)**	**len(r)**	**len(f+r)**	**overlap**	**% id.**	**f**	**r**	**f+r**	**f**	**r**	**f+r**
V1-V2	g_P.;s_aeruginosa	131	134	284	-19	100%	97	86	86	1	1	1
V3	g_P.;s_aeruginosa	133	134	159	108	100%	83	84	93	8	7	4
V4	g_P.;s_aeruginosa	136	130	252	14	100%	93	51	91	1	32	1
V5-V6	g_Pseudomonas	130	134	255	9	100%	84	71	93	4	18	0
V7-V8	f_Pseudomonadaceae	133	129	250	12	100%	85	73	86	10	21	9
V8-V9	g_Pseudomonas	133	130	293	-30	100%	73	87	73	21	4	21

Table 7: Sequencing of Pseudomonas aeruginosa with the selected primer pairs;
DADA2 (denoising); f = forward, r = reverse, Ψ = pseudo homopolymer, non-chim. = non-chimeric;
len = length (bp), % id. = % identity; det = determination by QIIME feature classifier.
The number of matches to *P. aeruginosa* and to other *Pseudomonas* spp. in BLAST is given. The analysis of the forward reads alone provides almost as good results as those of the (joined) paired-end reads.

However, one could already see here (Table 8), that depending on the primer pair used very short sequences were generated in very different frequencies (0.11 - 90.17%), which ended up with 8-10 A's and finally a poly-G string of different lengths (hereinafter referred to as G-pseudo homopolymers). The poly-A string is apparently attached terminally by the polymerase. Since the Illumina MiniSeq uses only 2 fluorochromes, a missing signal is interpreted as G (Illumina, 2016) (Pichler et al., 2018).
However, the pseudo-homopolymers only werde noticed when evaluated as single-end reads, since they do not have a terminally overlapping sequence in paired-end reads and are therefore filtered out throughout.

41671-18-V3 forward

TCCTACGGGDGGCWGCAG forward primer
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTACCGCGGCKGCTGGC reverse primer
GCCAGCMGCCGCGGTAAAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC reverse primer (reverse complement)
Index D703: CGCTCATT (reverse complement)
*CAAGCAGAAGACGGCATACGAGAT**AATGAGCG**GTGACTGGAGTTCAGACGTGT**GCTCTTCCGATCT***
Rev_Index1 (P7_I_Rd2SP)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGCTCATTATCTCGTATGCCGTCTTCTGCTTG
Rev_Index1(reverse complement)

Primer pair dimer:
TCCTACGGGD**GG**CWGCAG
| ||||||
GCCAGCMGCCGCGGTAAAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGCTCATTATCTCG
TATGCCGTCTTCTGCTTG

>95ff7c83badf1437f8d349ef359a14c5 (36.964 reads)
TCCTACGGGTGGCTGCAGCCGCGGTAAAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGCTCATTATCTCG
TATGCCGTCTTCTGCTTGAAAAAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
G
>c21e3e822a1748de49c23eb190a6d6db (665 reads)
TCCTACGGGTGGCTGCAGGCCCCGGATCCACTGGAGCAGTGCCAGCCGCCGCGGTAAAGATCGGAAGAGCACACG
TCTGAACTCCAGTCACCGCTCATTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAGGGGGGGGGGGGGGGGG
G
NG_060501.1 Homo sapiens Sharpr-MPRA regulatory region 7840 (LOC112577585) on chromosome 1

346-V12 forward

AGAGT**T**TGAT**YM**T**GGC**TYAR forward primer
G**GD**AC**T**GAGA**YA**C**GGH**C**AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC** reverse primer (rc)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGATTCCATCTCGTATGCCGTCT
TCTGCTTG
Index D704: GAGATTCC (reverse complement)

primer hairpin:
5'TCTTTCCCTACACGACGCTCTTCCGATC┐ TCTTTCCCTACACGACGCTCTTCCGATC┐
||| | | | || | -> ||| | | | || | |
3' RAYTCGGTMYTAGTTTGAGAT┘ AGAAAGGRAYTCGGTMYTAGTTTGAGAT┘
TCTTTCCCTACACGACGCTCTTCCGATCTAGAGTTTGATYMTGGCTYARGGAAAGA

>e2389d71c0a7f89893e218a7d4f2c1c4
AGAGTTTGATTATGGCC**AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC**GAGATTCCATCTCGTATGCCGTCT
TCTGCTTGAAAAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
G
AGAGTTTGATTCTGGCTTAGGGAAAGAGGGACTGAGATACGGTC**AGATCGGAAGAGCACACGTCTGAACTCCAGT**
CACGAGATTCCATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
G

Figure 9: Origin of G-pseudo-homopolymers

(short sequences as result of primer dimers, primer pair dimers, primer hairpins, insertion of short human DNA-sequences; poly-A string attached by polymerase; lack of sequencing signal at the end interpreted as poly-G strings ['pseudo-homopolymers'] according to the Illumina MiniSeq 2-dye chemistry)

It was obvious that these short sequences can be explained as primer dimers, primer pair dimers or as a result of the formation of primer hairpins. This was tested as an example on sequences with primers that have a particularly high rate of G-pseudo-homopolymers (Figure 9), and could be confirmed.

Taxon	Vx	input	filtered	denoised	non-chimeric	reads (forward)	f(Ψ)	f(Ψ)	r(Ψ)	r(Ψ)
Eikenella corrodens	V4	188449	182503	182453	124532	33549	1054	3,05%	0	0,00%
	V7-V8					81894	7767	8,66%	3405	10,41%
	V5-V6	191134	183544	183521	67667	11021	56646	83,71%	26552	38,02%
Pasteurella multocida	V4	283940	266974	266864	256854	90640	14189	13,54%	0	0,00%
	V7-V8					70296	81549	53,71%	16248	52,70%
	V5-V6	241312	223623	222680	147334	49806	97526	66,19%	114942	69,88%
Moraxella catarrhalis	V4	350948	342719	342541	274363	65717	917	1,38%	0	0,00%
	V7-V8					200560	7055	3,40%	2301	3,05%
	V5-V6	219717	210374	210346	188360	41051	147294	78,20%	106065	71,68%
Pseudomonas aeruginosa	V4	169666	166835	166779	165500	164946	24	0,01%	0	0,00%
	V7-V8	187128	180982	180974	102840	96969	5871	5,71%	2777	8,39%
	V5-V6	232284	223232	223216	199839	38734	161039	80,61%	92893	70,69%
Haemophilus influenzae	V4	231540	226584	226380	200587	108314	855	0,78%	0	0,00%
	V7-V8					88822	2513	2,75%	878	2,72%
	V5-V6	23851	22560	22551	8636	1151	7485	86,67%	5697	57,22%
Stenotrophomonas maltophilia	V4	164690	160193	159977	130604	53048	589	1,10%	0	0,00%
	V7-V8					74683	1972	2,57%	893	3,18%
	V5-V6	276990	264346	264054	101738	11641	59496	83,64%	133487	74,63%
Bacteroides fragilis	V4	178309	166186	166113	121752	3481	788	18,46%	4424	80,64%
	V7-V8					50461	67022	57,05%	692	2,35%
	V5-V6	205772	188603	187827	107627	20981	86571	80,49%	89001	81,20%
Enterococcus faecium	V4	269770	258761	258741	130798	4077	1822	30,89%	13620	91,13%
	V7-V8					102492	22256	17,84%	3181	7,17%
	V5-V6	244065	233971	233908	209283	34429	174657	83,53%	124966	78,65%
Citrobacter koseri	V4	265188	256723	256671	181176	112613	801	0,71%	3497	4,73%
	V7-V8					57297	10349	15,30%	0	0,00%
	V5-V6	196690	188303	188260	172041	60878	111041	64,59%	117149	66,28%
Klebsiella pneumoniae	V4	63671	62867	62789	49766	30983	89	0,29%	0	0,00%
	V7-V8					17204	1438	7,71%	318	3,37%
	V5-V6	133031	130538	130533	120273	50833	69440	57,74%	55587	55,52%
Campylobacter coli	V4	105578	103220	103101	92826	15643	107	0,68%	1776	27,01%
	V7-V8					58694	18340	23,81%	229	0,94%
	V5-V6	150843	145591	145032	65401	45588	19813	30,29%	74935	62,40%
Staphylococcus epidermidis	V4	102283	101109	100933	85923	6744	78	1,14%	113	4,02%
	V7-V8					77121	1887	2,39%	314	1,26%
	V5-V6	71583	70599	70587	68758	56286	12472	18,14%	7380	12,14%
Staphylococcus saprophyticus	V4	115120	113781	113519	95872	9293	512	5,22%	0	0,00%
	V7-V8					85302	722	0,84%	557	1,83%
	V5-V6	107388	105730	105724	100461	63974	36487	36,32%	30895	32,71%
Staphylococcus aureus	V4	102650	101598	101535	88217	7371	82	1,10%	0	0,00%
	V7-V8					80403	352	0,44%	274	1,13%
	V5-V6	142128	139407	139387	129262	56277	72942	56,45%	60888	52,17%
Nocardia farcinica	V4	96115	95014	95000	94968	12862	1142	8,15%	0	0,00%
	V7-V8					77495		0,00%	365	1,48%
	V5-V6	99489	95931	95122	46193	22817	23294	50,52%	34049	60,12%
Bacteroides fragilis	V4	95787	94282	94193	75680	2309	78	3,27%	0	0,00%
	V7-V8					72124	1169	1,59%	377	1,39%
	V5-V6	138401	135711	135700	126351	60274	66012	52,27%	51938	46,51%
Bacteroides thetaiotaomicron	V4	101062	99481	99357	82084	4285	46	1,06%	0	0,00%
	V7-V8					74873	2878	3,70%	802	2,81%
	V5-V6	98512	96625	96622	89567	38766	50730	56,68%	18635	18,68%

Table 8: Sequencing of bacterial cultures (G-pseudo-homopolymers)

V4 and V7-V8 were sequenced using multiplex PCR (except *P. aeruginosa*). Shown are raw reads ('input'), forward reads after denoising and filtering with DADA2 (color coding: green ↑, red ↓), and G-pseudo-homopolymers (color coding: green ↓, red ↑). The percentage refers to all remaining reads after denoising and filtering; f = forward, r = reverse, Ψ = pseudo-homopolymer.

Selected bacteria (V4, V5-V6, V7-V8)

Taxon	Vx	input	reads (forward)	reads (reverse)	reads (merged)	len(f)	len(r)	len (f+r)	overlap
Eikenella corrodens	V4	188449	33549	10293	32629	136	130	252	14
	V7-V8		81894	29293	92346	133	129	249	13
	V5-V6	191134	11021	43281	36414	130	134	255	9
Pasteurella multocida	V4	283940	90640	56955	88593	136	130	252	14
	V7-V8		70296	14581	79200	133	129	250	12
	V5-V6	241312	49806	49540	36588	130	134	253	11
Moraxella catarrhalis	V4	350948	65717	41293	64382	136	130	252	14
	V7-V8		200560	73249	227224	133	129	250	12
	V5-V6	219717	41051	41904	25410	130	134	255	9
Pseudomonas aeruginosa	V4	169666	164946	104068	159508	136	130	252	14
	V7-V8	187128	96969	30331	118720	133	129	250	12
	V5-V6	232284	38734	38513	26232	130	134	255	9
Haemophilus influenzae	V4	231540	108314	68297	105922	136	130	252	14
	V7-V8		88822	31458	99629	133	129	250	12
	V5-V6	23851	1151	4259	3582	130	134	253	11
Stenotrophomonas maltophilia	V4	164690	53048	26061	47596	136	130	252	14
	V7-V8		74683	27149	0	133	129	251	11
	V5-V6	276990	11641	45370	1110	130	134	256	8
Bacteroides fragilis	V4	178309	3481	1062	6273	136	130	252	14
	V7-V8		50461	28727	90287	133	129	250	12
	V5-V6	205772	20981	20608	14865	130	134	251	13
Enterococcus faecium	V4	269770	4077	1325	7664	136	130	252	14
	V7-V8		102492	41188	116906	133	129	249	13
	V5-V6	244065	34429	33919	700	130	134	256	8
Citrobacter koseri	V4	265188	112613	70378	109795	136	130	252	14
	V7-V8		57297	24860	76982	133	129	250	12
	V5-V6	196690	60878	59611	36061	130	134	255	9
Klebsiella pneumoniae	V4	63671	30983	20786	32422	136	130	252	14
	V7-V8		17204	9124	28210	133	129	250	12
	V5-V6	133031	50833	44541	32090	130	134	255	9
Campylobacter coli	V4	105578	15643	4799	15458	136	130	252	14
	V7-V8		58694	24244	68483	133	129	248	14
	V5-V6	150843	45588	45146	0	130	134	259	5
Staphylococcus epidermidis	V4	102283	6744	2701	89823	136	130	252	14
	V7-V8		77121	24611	6669	133	129	249	13
	V5-V6	71583	56286	53414	28	130	134	258	6
Staphylococcus saprophyticus	V4	115120	9293	3762	9236	136	130	252	14
	V7-V8		85302	29913	99042	133	129	249	13
	V5-V6	107388	63974	63565	13	130	134	258	6
Staphylococcus aureus	V4	102650	7371	2963	7318	136	130	252	14
	V7-V8		80403	23926	93264	133	129	249	13
	V5-V6	142128	56277	55821	33	130	134	258	6
Nocardia farcinica	V4	96115	12862	2116	12726	136	130	252	14
	V7-V8		77495	24328	0	133	129	251	11
	V5-V6	99489	22817	22587	17928	130	134	255	9
Bacteroides fragilis	V4	95787	2309	2725	4224	136	130	252	14
	V7-V8		72124	26678	82923	133	129	250	12
	V5-V6	138401	60274	59731	48487	130	134	251	13
Bacteroides thetaiotaomicron	V4	101062	4285	4969	7707	136	130	252	14
	V7-V8		74873	27733	86233	133	129	250	12
	V5-V6	98512	38766	81106	31145	130	134	249	15

Table 9: Sequencing of bacterial cultures (reads, denoising)

same dataset as tab. 8; f = forward, r = reverse, len = length (bp).

Raw reads ('input'), remaining number (color-coded) and length of forward/reverse/merged reads after denoising and filtering with DADA2 (color coding: green ↑, red ↓) as well as the number of overlapping bases of forward and reverse read (yellow: <12 bp) are shown.

Taxon	Vx	+++ (f)	--- (f)	+++ (r)	--- (r)	+++ (f+r)	--- (f+r)	remark
Eikenella corrodens	V4	7		4	11	7		
	V7-V8	7	14	11	49	10	1	esp. *Kingella negevensis (22)*
	V5-V6	12	4	5	4	5		
Pasteurella multocida	V4	91	1	6	77	92		esp. *Mannheimia haemolytica (34)*
	V7-V8	99		99		99		
	V5-V6	40	45	42	42	94		esp. *Haemophilus influenzae (44)*
Moraxella catarrhalis	V4	12		11		14		
	V7-V8	17	21	17	19	17	15	
	V5-V6	15	5	17	2	16	2	
Pseudomonas aeruginosa	V4	93	2	9	26	93	2	esp. *Pseudomonas putida (12)*
	V7-V8	86	10	68	22	87	10	
	V5-V6	86	5	72	17	94	1	
Haemophilus influenzae	V4	32	8	3	80	32	8	esp. *Mannheimia haemolytica (34)*
	V7-V8	65	2	78	2	62	2	
	V5-V6	53	27	44	41	72	2	
Stenotrophomonas maltophilia	V4	44	11	45	6	44	7	
	V7-V8	23	27	47	10	49	13	
	V5-V6	39	19	42	7	44	7	
Bacteroides fragilis	V4	15		30	12	53		
	V7-V8	12	22	6	1	3	1	
	V5-V6	12	33	44	1	41	1	
Enterococcus faecium	V4	36	51	26	58	36	54	
	V7-V8	68	27	88	10	77	18	
	V5-V6	48	40	59	30	22	69	esp. *Enterococcus durans (59)*
Citrobacter koseri	V4	65	1		95	63	1	esp. *Serratia marcescens* (51), no *Citrobacter koseri*!
	V7-V8		65		61	1	72	esp. *Klebsiella pneumoniae* (16/44/52)
	V5-V6		96	1	91	1	93	esp. *Enterobacter kobei* (40) or *Escherichia coli* (67/84)
Klebsiella pneumoniae	V4	3	95	3	97	3	95	esp. *Enterobacter asburiae (46)*
	V7-V8	8	80	44	17	63	22	esp. *Escherichia coli (41)*
	V5-V6	94	3	27	52	72	11	*Klebsiella pneumoniae et al. (27+52)*
Campylobacter coli	V4	28	65	26	56	28	63	esp. *Campylobacter jejuni (62/55/62)*
	V7-V8	26	73	22	71	21	71	esp. *Campylobacter jejuni (72/67/70)*
	V5-V6	27	58	20	36	23	38	esp. *Campylobacter jejuni (53/56/37)*
Staphylococcus epidermidis	V4	5	90	5	77	10	77	esp. *Staphylococcus aureus (24/24/29)*
	V7-V8	10	80	64	3	65	3	esp. *Staphylococcus aureus (36)*
	V5-V6	5	74	50	21	51	23	esp. *Staphylococcus aureus (23)*
Staphylococcus saprophyticus	V4	8	70	1	90	23	60	esp. *Staphylococcus cohnii/aureus/cohnii* (39/29/31)
	V7-V8	14	65	14	56	31	45	esp. *Staphylococcus sciuri (25)*
	V5-V6	1	78	30	54	30	53	esp. *Staphylococcus aureus (23)*
Staphylococcus aureus	V4	24	67	31	60	33	58	
	V7-V8	37	55	48	46	62	33	
	V5-V6	25	62	88	7	89	6	
Nocardia farcinica	V4	48	34	8	41	55	38	esp. *Nocardia abscessus (15+34)*
	V7-V8	60	26	1	46	55	38	esp. *Nocardia seriolae (20+27)*
	V5-V6	2	41	2	52	42	28	esp. *Nocardia abscessus (15+39)*
Bacteroides fragilis	V4	15		30	12	53		
	V7-V8	29	20	27	1	13	1	
	V5-V6	12	33	44	1	41	1	
Bacteroides thetaiotaomicron	V4	11	1	3	39	14	1	esp. *Bacteroides fragilis (30)*
	V7-V8	16	28	17	16	21	3	
	V5-V6	2	43	12		12		esp. *Bacteroides fragilis (12)*

Table 10: Sequencing of bacterial cultures (species determination)

same dataset as tab. 8; number of species (sequences) with 100 % identity, determined by BLAST (+++ expected sp. [as determined by MALDI-TOF]; --- alternate spp.); color coding: green = predominantly correct determination, red = predominantly incorrect determination (at species level).
The analysis of the forward reads partially provides similar results as the paired-end reads.

The sequencing of bacterial cultures was straightforward and delivered the expected results.

The primer pairs used were those for the variable regions V4, V5-V6, and V7-V8 (Table 3), with the library preparation for V4 and V7-V8 being carried out in one tube (as a multiplex approach), but this was not without problems proved that the variable regions of different species of bacteria were amplified very unequally effective.

The forward and reverse reads overlap a total of 5-15 nt (V4: 14 nt, V5-V6: 5-15 nt; V7-V8: 11-14 nt overlap) (Table 9). The sequences could therefore only be partially evaluated as paired-end reads with QIIME2, since an overlap of at least 12 bp is required for this. The extent of the overlap is shown in the table. Reads with a shorter overlap had to be evaluated as single-end reads, which, despite the shorter sequence length, only partially yielded poorer results. In addition, the demultiplexed (consensus) reads could also be merged subsequently and then analyzed e.g. with BLAST (Altschul et al., 1990), provided it was clear that they came from the same bacterial species (Table 10). Thus, no information was lost here either.

Discussion

The almost complete database of 16S rDNA sequences of human pathogenic bacteria presented here provides for the first time a solid basis for assessing the accuracy of primer matches and the degree of coverage of human pathogenic bacterial species. It also makes it possible to determine which species may not be detected with a particular primer pair.

The addition of the few sequences that are still missing in the database primarily fails due to the lack of supporting material, since especially endoparasitic bacteria (including type strains!) are often lost in the culture.

Overall, all of the primers presented here fit very well, for 75.4% of the human pathogenic bacteria there is not a single mismatch with these primers. All primers match at least 92.3% of the human pathogenic species or 96.74% if one mismatch is allowed. Consequently, all human pathogenic species should be amplifiable with these primers.

Since the mostly used standard primers for 16S rDNA sequencing of the variable region V4 (515F and 806R) have proved to be already largely optimized, there should be little need for action here.

There is still scope for optimization for the primers of the other variable regions. For the exact determination of human pathogenic bacteria from FFPE, V1-V2 and especially V6 would be preferable (Coenye and Vandamme, 2003), but they are also flanked by somewhat less conserved primer binding sites (Figure 7) and most affected by the pseudo-homopolymer problem mentioned.

Artifacts due to secondary structures (primer dimers, primer pair dimers or primer hairpins) mainly occurred in the primer pairs for the amplification of the variable regions V1-V2, V3, and V5-V6, but hardly in the mostly used primers for V4 (Table 7).

The next step will be to establish and validate 16S rDNA sequencing from FFPE, although optimization steps may still be necessary here. In a few experiments, FFPE could almost only be used to sequence the usual contaminants (*Acinetobacter junii* etc.), which could also be an indication that there was very little bacterial DNA in the sample or that there was a problem with the extraction.

Finally, the analysis of the relatively short and little overlapping sequences causes problems, although the isolated evaluation of the individual (forward or reverse) reads also gives surprisingly reliable results. With paired-end reads, QIIME2 requires an

overlap of at least 12 bp, which is never the case with V1-V2 and V8-V9, and usually not with V5-V6. Only V3, V4 and (mostly) V7-V8 overlap sufficiently (Table 3, Table 7). Here, bioinformatics would be required to compare the generated sequences with the 16S rDNA sequences in the databases (human pathogenic bacteria, SILVA) and to use even short overlaps to connect forward and reverse reads of one and the same taxon (or the non-overlapping reads from V1-V2 or V8-V9 by inserting the corresponding number of N's). Reads from several variable regions could also be mapped against the reference sequence of the associated taxon in order to ensure the determination.

Furthermore, there are now available special computer programs for the optimization of 16S primers (Sambo et al., 2018), which are open source, but cannot be used entirely without bioinformatic skills. All that is required as input is a data set of 16S reference sequences and a set of primers to be optimized as a starting point. The algorithm is designed to maximize efficiency and specificity of target amplification and can be applied to any desired amplicon length. It would be very interesting to see which primer pairs this algorithm outputs as optimal for the short read length of the Illumina Miniseq.

Appendix:
Phylogenetic origin of mitochondria

Since a phylogenetic tree of Bacteria was now already available, it was obvious to integrate the homologous small-subunit rRNA gene of the mitochondria into this tree and to examine its lineage[8].

Mitochondria

Mitochondria are cell organelles that use the respiratory chain to build a proton gradient on the inner mitochondrial membrane, which drives almost the entire ATP synthesis of the organism (oxidative phosphorylation).

Endosymbiotic theory

According to the very well-established endosymbiotic theory, mitochondria are derived from an Alphaproteobacterium, which was taken up by an anaerobic precursor cell of the eukaryotes (an archaeon) about 1.6 billion years ago, but was not phagocytosed (digested). Instead, through gene transfer to the nuclear genome on the one hand and an increase in energy requirements due to increasingly complex organization, enlargement of the genome and development of multicellular organisms on the other hand an endosymbiosis developed in which both partners would not be able to survive without the other.

In fact, there is now overwhelming evidence of bacterial origin of mitochondria and plastids (Alberts et al., 2017):

- Their structure largely corresponds to that of prokaryotes: they have a circular DNA, but no cell nucleus. The DNA is not 'packaged' by histones, but by histone-like proteins (HLP). The mRNA of the two organelles does not have the 5'-cap sequence typical for eukaryotes, and the polyadenylation is also missing. They have their own prokaryotic protein synthesis apparatus with 70-S (instead of 80-S) ribosomes and their own tRNA's, protein synthesis starts in both organelles with N-formylmethionine (as in the bacteria) and not with methionine as in the cytoplasm of the eukaryotes. Both organelles are sensitive to antibacterial antibiotics (e.g. tetracycline).
- The DNA sequences of the mitochondria are similar to those of the α-proteobacteria, whereas plastid DNA sequences resemble those of the cyanobacteria.

[8]) Since there are no cyanobacteria that are pathogenic to humans, an analogous trial with DNA of plastids (ptDNA: chloroplasts etc., Sadali et al., 2019) did not seem to be worthwhile.

- Primary plastids and mitochondria are surrounded by double membranes, the outer one being added during the uptake (endocytosis) of the bacterium and resembling the cell membrane of the eukaryotes. The inner one corresponds to that of bacteria (occurrence of cardiolipin, no cholesterol; also occurrence of transmembrane proteins (β-barrel proteins), which only occur in the membranes of bacteria and cell organelles).
- The membrane-bound ATPases of the bacteria and organelles are closely related to one another, just like those of the archaea and the eukaryote. There is only a more distant relationship between these two groups (Lane, 2015). Exceptions to this rule can probably be explained by horizontal gene transfer (Hilario and Gogarten, 1993).
- The cyanelles of the glaucophyta are even surrounded by a thin bacterial cell wall. Red algae and glaucophyta, like cyanobacteria, use phycobilins to capture photons in photosynthesis.
- Mitochondria and plastids multiply by division and are distributed to the daughter cells when the host cell is divided. The cell cannot regenerate them if they are accidentally lost.

The endosymbiont theory was first formulated in 1883 by the botanist Andreas F. W. Schimper to explain the formation of chloroplasts (Schimper, 1883). Ivan E. Wallin took up the idea for both mitochondria and chloroplasts in 1922 and concluded: „the author can arrive at no other conclusion than, that mitochondria are symbiotic bacteria in the cytoplasm of the cells of all higher organisms whose symbiotic existence had its inception at the dawn of phylogenetic evolution“, and chloroplasts are "bacteria or bacteria-like organisms that accepted the leisure of a symbiotic partnership in the struggle for existence" (Wallin, 1922b, Wallin, 1922a, Wallin, 1922c). Although he even published his findings in book form (Wallin, 1927), they were largely forgotten for decades until Lynn Margulis put them on a broader scientific basis with better evidence in 1967 (Sagan, 1967, Margulis, 1970).

The lineage of the mitochondria from the Alphaproteobacteria was recognized around 1985 (Yang et al., 1985), in 1994 the mitochondrial ancestor for the first time was placed near *Rickettsia*, obligatory intracellular parasites (using mitochondrial sequences of the small-subunit rRNA [*Zea mays*] or the heat-shock protein Hsp60 [14 eukaryotes from protists, plants, fungi and animals]) (Olsen et al., 1994, Viale and Arakaki, 1994), 1998 due to sequences of cytochrome b (CYTB) and cytochrome c

oxidase I (COX1) in the vicinity of *Rickettsia prowazekii*, the causative agent of epidemic typhus ('spotted fever') (Sicheritz-Ponten et al., 1998). In the same year, this lineage hypothesis was well-founded and supported by the complete genome of *R. prowazekii* (Andersson et al., 1998, Gray et al., 1999), but did not remain unchallenged in the following two decades (Thrash et al., 2011). Only recently was published in a high-ranking journal that the mitochondria developed from a proteobacterial line that branched off before the splitting of all examined alpha proteobacteria. (Martijn et al., 2018).

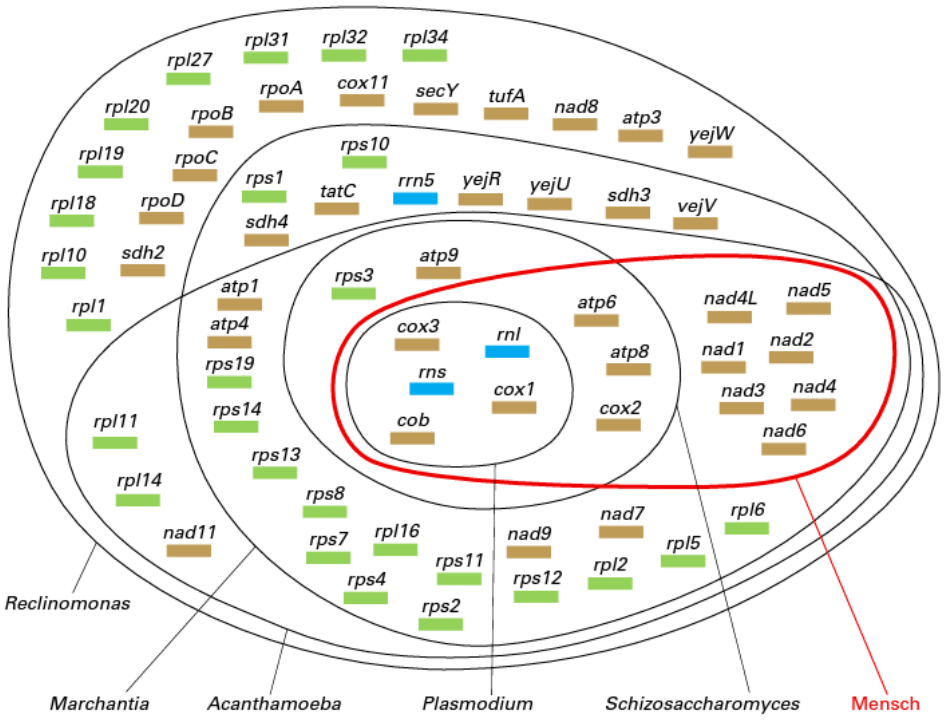

Figure 10: Comparison of mitochondrial genomes of different taxa (genes for proteins and ribosomal RNA);

blue: rRNA, green: ribosomal proteins, brown: proteins of the respiratory chain etc.

from Alberts et al., 2017, Molekularbiologie der Zelle. 6. Aufl., Wiley-VCH (Weinheim).

Primordial mitochondrial genome

A first attempt to integrate mitochondrial DNA (mtDNA) from *Homo sapiens* (and other representatives of animals and fungi) into the bacterial phylogenetic tree failed. However, the human 12S rRNA gene (MT-RNR1) is only 954 bp long (reference sequence NC_012920.1) in contrast to the on average 1550 bp long bacterial 16S rRNA gene (e.g. *Bacillus subtilis*, NR_102783.2). Not only have numerous genes been transferred to the nuclear genome, the genes remaining in the mtDNA (37 genes for

Homo sapiens: 13 protein-coding genes, 2 for rRNA and 22 for tRNA-coding genes) were in some cases greatly modified and reduced. (Alberts et al., 2017)

However, there are eukaryotes whose mitochondrial genome is still very original.

The record holders are the protists (Eukaryota: Jakobea: Jakobida) *Andalucia godoyi* and *Reclinomonas americana*, whose mitochondrial genome comprises 101 and 97 genes (Burger et al., 2013) and still uses the universal genetic code (standard code)[9], while the human mitochondrium encodes only 37 genes and that of *Plasmodium falciparum*, the causative agent of malaria tropica, only 5 genes having shifted all tRNAs to the nucleus (Figure 10). These use different mitochondrial codes for vertebrates or protozoa. Surprisingly, plant-based mitochondrial genomes also appear to be quite original. The mtDNA of the liverwort *Marchantia* encodes at least 40 proteins and 3 rRNA's (Alberts et al., 2017, Andersson et al., 1998). Plants have very large mitochondrial genomes, but their annotation is very difficult (large amounts of 'junk DNA'?), as you can see if you look at the plot ('Graphics') of the mitochondrial reference genome of the model organism *Arabidopsis thaliana* (thale cress).

Common ancestor of mitochondria and Rickettsiales

The integration of the available Jakobida, another protist and representatives of numerous plant orders (**S12**) into the data set was no problem.

In the phylogenetic family tree obtained, all mitochondrial genomes cluster together (i.e. appear monophyletic, bootstrap 100) and branch off from a common ancestor with the Rickettsiales (genera *Rickettsia*, *Orientia*, *Neorickettsia*, *Wolbachia*, *Ehrlichia*, *Aegyptianella*, *Anaplasma*) (Figure 11). A close relationship between the mitochondria as endosymbionts and the Rickettsiales as obligate intracellular (endo)parasites appears to be quite plausible from an evolutionary point of view and does not lack a certain elegance and beauty.

[9]) According to the entry in the NCBI reference sequences NC_021124.1 (*Andalucia godoyi*) and NC_001823.1 (*Reclinomonas americana*), they use transl_table = 11 (The Bacterial, Archaeal and Plant Plastid Code), which differs from the standard code only through the additional use of 4 codons (AUU, AUC, AUA, GUG) as start codons (these otherwise code for isoleucine and valine).

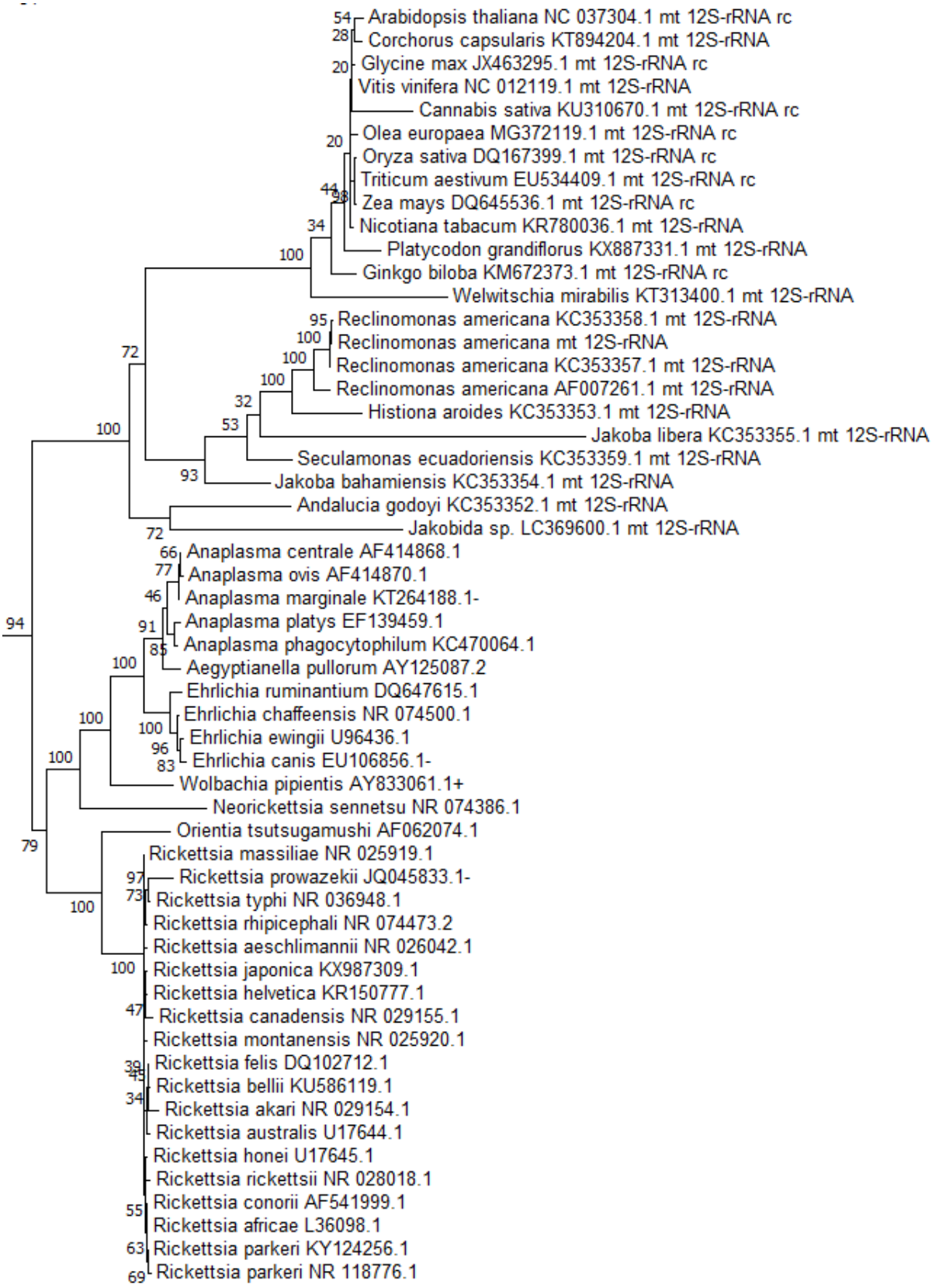

Figure 11: Section of the phylogenetic tree of bacteria and mitochondria;

(the mitochondria [above] and Rickettsiales [below] are shown); Source: Sequence data of the small subunit rRNA of human pathogenic bacteria and the mitochondria of selected types of protists and plants (with quite original mitochondrial genome); Method: RAxML (see **S13**);

mt = mitochondrial, rc = reverse complement; values are bootstrap values.

Supplementary material

(available from the author: f.puehringer@sesiidae.net, franz.puehringer@ooeg.at**)**

S1: Human pathogenic bacteria database (MS Access [58.7 MB], Table 16S rDNA; and Excel [2.17 MB])

S2: Phylogenetic tree of the domain Bacteria (created with RaxML) [226 KB]

S3: 16S-rDNA analysis, Excel table (spreadsheet for sequence alignments): aligned sequences of the bacterial 16S rRNA gene, reference sequence of *E. coli* with numbering; bases per position; degree of conservation (MS Excel) [29.7 MB] Mismatches with the primers presented here are marked in red.

S4: Alignment of the almost complete set of sequences of the 16S rRNA gene (nucleotides 8-1541; reference: E. coli) of human pathogenic bacteria, predominantly created manually using MUSCLE and PhyDE (description in text; fasta format) [3.91 MB]

S5: Catalog of possible primers with melting temperatures (MS Excel) [72 KB]

S6: 16S-rDNA primer catalog (according to Klindworth et al., 2013, corrected and supplemented; MS Access, Table 16S-rRNA primer catalog; and Excel) [30 KB]

S7: Primer matches with human 18S and 12S rRNA (MS Excel) [17 KB]

S8: 16S rDNA Library Preparation Protocol, adapted from 'Deep-Seq Library Preparation Protocol' (provided by E. Heitzer, MUG) [55 KB]

S9: QIIME2 script for Training feature classifiers with q2-feature-classifier [19 KB]

S10: QIIME2 script for QIIME2 and DADA2 analysis of Illumina paired-end reads [28 KB]

S11: QIIME2 script for QIIME2 and DADA2 analysis of Illumina single-end reads [24 KB]

S12: Species with original mitochondrial SSU rRNA (integrated in the data set; MS Excel) [11 KB]

S13: Phylogenetic tree of the domain Bacteria, including quite original mitochondrial genomes (created with RaxML) [244 KB]

RAxML was called as follows:

raxmlHPC-PTHREADS-SSE3.exe -T 3 -f a -x 995 -p 173 -N 100 -m GTRCAT -O -n S13_RaxML.tre -s C:\Medizinische_Genetik\16S-rDNA\infile.fas -w C:\Medizinische_Genetik\16S-rDNA

Literature

eurofins [Online]. Available: https://www.eurofinsgenomics.eu/en/ecom/tools/oligo-analysis/ [Accessed 25-10-2019].

Illumina, Inc. Illumina Adapter Sequences. 2019 [Online]. Available: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-adapter-sequences-1000000002694-11.pdf [Accessed 29-10-2019].

Illumina, Inc. Illumina Two-Channel SBS Sequencing Technology. High data accuracy with faster data generation. 2016 [Online]. Pub. No. 770-2013-054. Available: https://www.well.ox.ac.uk/ogc/wp-content/uploads/2017/09/techspotlight_two-channel_sbs.pdf [Accessed 22-02-2020].

NCBI PrimerBLAST [Online]. Available: https://www.ncbi.nlm.nih.gov/tools/primer-blast/ [Accessed 25-10-2019].

NCBI RefSeq. Arabidopsis thaliana, reference mitochondrial genome NC_037304.1, Graphics [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/NC_037304.1?report=graph [Accessed 2019-10-19].

NCBI Taxonomy Database, [Online]. Available: https://www.ncbi.nlm.nih.gov/guide/taxonomy/ [Accessed 18-10-2019].

OligoCalc [Online]. Available: http://biotools.nubic.northwestern.edu/OligoCalc.html [Accessed 25-10-2019].

ALBERTS, B., JOHNSON, A., LEWIS, J., MORGAN, D., RAFF, M., ROBERTS, K. & WALTER, P. 2017. *Molekularbiologie der Zelle. 6. Aufl. Übersetzung herausgegeben von U. Schäfer,* Weinheim (Germany): Wiley-VCH

ALM, E. W., OERTHER, D. B., LARSEN, N., STAHL, D. A. & RASKIN, L. 1996. The oligonucleotide probe database. *Appl Environ Microbiol,* 62**,** 3557-9.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* 25**,** 3389-402.

AMINOV, R. I. 2013. Role of archaea in human disease. *Frontiers in Cellular and Infection Microbiology,* 3.

ANDERSSON, S. G., ZOMORODIPOUR, A., ANDERSSON, J. O., SICHERITZ-PONTEN, T., ALSMARK, U. C., PODOWSKI, R. M., NASLUND, A. K., ERIKSSON, A. S., WINKLER, H. H. & KURLAND, C. G. 1998. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature,* 396**,** 133-40.

BAKER, G. C., SMITH, J. J. & COWAN, D. A. 2003. Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods,* 55**,** 541-55.

BOKULICH, N. A., KAEHLER, B. D., RIDEOUT, J. R., DILLON, M., BOLYEN, E., KNIGHT, R., HUTTLEY, G. A. & GREGORY CAPORASO, J. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome,* 6**,** 90.

BOLYEN, E., RIDEOUT, J. R., DILLON, M. R., BOKULICH, N. A., ABNET, C. C., AL-GHALITH, G. A., ALEXANDER, H., ALM, E. J., ARUMUGAM, M., ASNICAR, F., BAI, Y., BISANZ, J. E., BITTINGER, K., BREJNROD, A., BRISLAWN, C. J., BROWN, C. T., CALLAHAN, B. J., CARABALLO-RODRIGUEZ, A. M., CHASE, J., COPE, E. K., DA SILVA, R., DIENER, C., DORRESTEIN, P. C., DOUGLAS, G. M., DURALL, D. M., DUVALLET, C., EDWARDSON, C. F., ERNST, M., ESTAKI, M., FOUQUIER, J., GAUGLITZ, J. M., GIBBONS, S. M., GIBSON, D. L., GONZALEZ, A., GORLICK, K., GUO, J., HILLMANN, B., HOLMES, S., HOLSTE, H., HUTTENHOWER, C., HUTTLEY, G. A., JANSSEN, S., JARMUSCH, A. K., JIANG, L., KAEHLER, B. D., KANG, K. B., KEEFE, C. R., KEIM, P., KELLEY, S. T., KNIGHTS, D., KOESTER, I., KOSCIOLEK, T., KREPS, J., LANGILLE, M. G. I., LEE, J., LEY, R., LIU, Y. X., LOFTFIELD, E., LOZUPONE, C., MAHER, M., MAROTZ, C., MARTIN, B. D., MCDONALD, D., MCIVER, L. J., MELNIK, A. V., METCALF, J. L., MORGAN, S. C., MORTON, J. T., NAIMEY, A. T., NAVAS-MOLINA, J. A., NOTHIAS, L. F., ORCHANIAN, S. B., PEARSON, T., PEOPLES, S. L., PETRAS, D., PREUSS, M. L., PRUESSE, E., RASMUSSEN, L. B., RIVERS, A., ROBESON, M. S., 2ND, ROSENTHAL, P., SEGATA, N., SHAFFER, M., SHIFFER, A., SINHA, R., SONG, S. J., SPEAR, J. R., SWAFFORD, A. D., THOMPSON, L. R., TORRES, P. J., TRINH, P., TRIPATHI, A., TURNBAUGH, P. J., UL-HASAN, S., VAN DER HOOFT, J. J. J., VARGAS, F., VAZQUEZ-BAEZA, Y., VOGTMANN, E., VON HIPPEL, M., WALTERS, W., et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol,* 37**,** 852-857.

BROSIUS, J., PALMER, M. L., KENNEDY, P. J. & NOLLER, H. F. 1978. Complete nucleotide sequence of a 16S ribosomal RNA gene from Escherichia coli. *Proc Natl Acad Sci U S A,* 75**,** 4801-5.

BURGER, G., GRAY, M. W., FORGET, L. & LANG, B. F. 2013. Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biol Evol,* 5**,** 418-38.

CALLAHAN, B. J., MCMURDIE, P. J., ROSEN, M. J., HAN, A. W., JOHNSON, A. J. & HOLMES, S. P. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods,* 13**,** 581-3.

CHAKRAVORTY, S., HELB, D., BURDAY, M., CONNELL, N. & ALLAND, D. 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods,* 69**,** 330-9.

CHEN, K., NEIMARK, H., RUMORE, P. & STEINMAN, C. R. 1989. Broad range DNA probes for detecting and amplifying eubacterial nucleic acids. *FEMS Microbiol Lett,* 48**,** 19-24.

COENYE, T. & VANDAMME, P. 2003. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiol Lett,* 228**,** 45-9.

DELONG, E. F., FRANKS, D. G. & ALLDREDGE, A. L. 1993. Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. *Limnology and Oceanography,* 38**,** 924-934.

EDGAR, R. C. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics,* 5**,** 113.

EDGAR, R. C. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res,* 32**,** 1792-7.

EUZEBY, J. P. 1997. List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet. *Int J Syst Bacteriol,* 47**,** 590-2.

FOX, G. E., STACKEBRANDT, E., HESPELL, R. B., GIBSON, J., MANILOFF, J., DYER, T. A., WOLFE, R. S., BALCH, W. E., TANNER, R. S., MAGRUM, L. J., ZABLEN, L. B., BLAKEMORE, R., GUPTA, R., BONEN, L., LEWIS, B. J., STAHL, D. A., LUEHRSEN, K. R., CHEN, K. N. & WOESE, C. R. 1980. The phylogeny of prokaryotes. *Science,* 209**,** 457-63.

GARRITY, G. M. *(ed.) 2001-2012. Bergey's Manual of Systematic Bacteriology. Vol. 1-5,* New York, Springer Verlag.

GRAY, M. W., BURGER, G. & LANG, B. F. 1999. Mitochondrial evolution. *Science,* 283**,** 1476-81.

GRAY, M. W., SANKOFF, D. & CEDERGREN, R. J. 1984. On the evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved structural core in small subunit ribosomal RNA. *Nucleic Acids Res,* 12**,** 5837-52.

GULEN, B., PETROV, A. S., OKAFOR, C. D., VANDER WOOD, D., O'NEILL, E. B., HUD, N. V. & WILLIAMS, L. D. 2016a. Ribosomal small subunit domains radiate from a central core. *Sci Rep,* 6**,** 20885.

GULEN, B., PETROV, A. S., OKAFOR, C. D., VANDER WOOD, D., O'NEILL, E. B., HUD, N. V. & WILLIAMS, L. D. 2016b. Ribosomal small subunit domains radiate from a central core. *Scientific Reports,* 6.

HILARIO, E. & GOGARTEN, J. P. 1993. Horizontal transfer of ATPase genes--the tree of life becomes a net of life. *Biosystems,* 31**,** 111-9.

KATOH, K., MISAWA, K., KUMA, K. & MIYATA, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res,* 30**,** 3059-66.

KLINDWORTH, A., PRUESSE, E., SCHWEER, T., PEPLIES, J., QUAST, C., HORN, M. & GLOCKNER, F. O. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res,* 41**,** e1.

KNOOP, V. & MÜLLER, K. 2009. *Gene und Stammbäume. Ein Handbuch zur molekularen Phylogenetik*, Elsevier.

KUMAR, S., STECHER, G. & TAMURA, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol,* 33**,** 1870-4.

LANE, D. J., PACE, B., OLSEN, G. J., STAHL, D. A., SOGIN, M. L. & PACE, N. R. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A,* 82**,** 6955-9.

LANE, N. 2015. *The vital question: energy, evolution, and the origins of complex life,* New York u.a., Norton & Company.

LAPAGE, S. P., SNEATH, P. H. A., LESSEL, E. F., SKERMAN, V. B. D., SEELIGER, H. P. R. & CLARK, W. A. 1992. (eds.) 1992. International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision. Washington (DC).

MARGULIS, L. 1970. *Origin of eukaryotic cells; evidence and research implications for a theory of the origin and evolution of microbial, plant, and animal cells on the Precambrian earth,* New Haven,, Yale University Press.

MARTIJN, J., VOSSEBERG, J., GUY, L., OFFRE, P. & ETTEMA, T. J. G. 2018. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature,* 557**,** 101-105.

MARTÍNEZ-PORCHAS, M., VILLALPANDO-CANCHOLA, E. & VARGAS-ALBORES, F. 2016. Significant loss of sensitivity and specificity in the taxonomic classification occurs when short 16S rRNA gene sequences are used. *Heliyon,* 2.

MORGULIS, A., COULOURIS, G., RAYTSELIS, Y., MADDEN, T. L., AGARWALA, R. & SCHAFFER, A. A. 2008. Database indexing for production MegaBLAST searches. *Bioinformatics,* 24**,** 1757-64.

MÜLHARDT, C. 2013. *Der Experimentator: Molekularbiologie, Genomics,* Berlin u.a., Springer Spektrum.

MÜLLER, J., MÜLLER, K., NEINHUIS, C. & QUANDT, D. 2010. *PhyDE - Phylogenetic Data Editor* [Online]. Available: http://www.phyde.de/index.html [Accessed].

MULLIS, K., FALOONA, F., SCHARF, S., SAIKI, R., HORN, G. & ERLICH, H. 1986. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol,* 51 Pt 1**,** 263-73.

OLSEN, G. J., WOESE, C. R. & OVERBEEK, R. 1994. The winds of (evolutionary) change: breathing new life into microbiology. *J Bacteriol,* 176**,** 1-6.

PARTE, A. C. 2014. LPSN - list of prokaryotic names with standing in nomenclature. *Nucleic Acids Res,* 42**,** D613-6.

PARTE, A. C. 2018. LPSN - List of Prokaryotic names with Standing in Nomenclature (bacterio.net), 20 years on. *Int J Syst Evol Microbiol,* 68**,** 1825-1829.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., #201 & DUCHESNAY, D. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.,* 12**,** 2825-2830.

PETROV, A. S., BERNIER, C. R., GULEN, B., WATERBURY, C. C., HERSHKOVITS, E., HSIAO, C., HARVEY, S. C., HUD, N. V., FOX, G. E., WARTELL, R. M. & WILLIAMS, L. D. 2014. Secondary structures of rRNAs from all three domains of life. *PLoS One,* 9**,** e88222.

PICHLER, M., COSKUN, O. K., ORTEGA-ARBULU, A. S., CONCI, N., WORHEIDE, G., VARGAS, S. & ORSI, W. D. 2018. A 16S rRNA gene sequencing and analysis protocol for the Illumina MiniSeq platform. *Microbiologyopen,* 7**,** e00611.

PLASTERER, T. N. 1997. PRIMERSELECT. Primer and probe design. *Methods Mol Biol,* 70**,** 291-302.

PRICE, M. N., DEHAL, P. S. & ARKIN, A. P. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol,* 26**,** 1641-50.

PRICE, M. N., DEHAL, P. S. & ARKIN, A. P. 2010. FastTree 2 - approximately maximum-likelihood trees for large alignments. *PLoS One,* 5**,** e9490.

QUAST, C., PRUESSE, E., YILMAZ, P., GERKEN, J., SCHWEER, T., YARZA, P., PEPLIES, J. & GLÖCKNER, F. O. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research,* 41**,** D590-D596.

SADALI, N. M., SOWDEN, R. G., LING, Q. & JARVIS, R. P. 2019. Differentiation of chromoplasts and other plastids in plants. *Plant Cell Rep,* 38**,** 803-818.

SAGAN, L. 1967. On the origin of mitosing cells. *J Theor Biol,* 14**,** 255-74.

SAMBO, F., FINOTELLO, F., LAVEZZO, E., BARUZZO, G., MASI, G., PETA, E., FALDA, M., TOPPO, S., BARZON, L. & DI CAMILLO, B. 2018. Optimizing PCR primers targeting the bacterial 16S ribosomal RNA gene. *BMC Bioinformatics,* 19.

SCHIMPER, A. F. W. 1883. Ueber die Entwicklung der Chlorophyllkörner und Farbkörper. *Botanische Zeitung,* 41**,** 105-112, 121-131, 137-146, 153-162, pl. 1.

SHINE, J. & DALGARNO, L. 1974. The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A,* 71**,** 1342-6.

SICHERITZ-PONTEN, T., KURLAND, C. G. & ANDERSSON, S. G. 1998. A phylogenetic analysis of the cytochrome b and cytochrome c oxidase I genes supports an origin of mitochondria from within the Rickettsiaceae. *Biochim Biophys Acta,* 1365**,** 545-51.

SILVESTRO, D. & MICHALAK, I. 2011. raxmlGUI: a graphical front-end for RAxML. *Organisms Diversity & Evolution,* 12**,** 335-337.

STAMATAKIS, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* 30**,** 1312-3.

STIEGLER, P., CARBON, P., EBEL, J. P. & EHRESMANN, C. 1981. A general secondary-structure model for procaryotic and eucaryotic RNAs from the small ribosomal subunits. *Eur J Biochem,* 120**,** 487-95.

THRASH, J. C., BOYD, A., HUGGETT, M. J., GROTE, J., CARINI, P., YODER, R. J., ROBBERTSE, B., SPATAFORA, J. W., RAPPE, M. S. & GIOVANNONI, S. J. 2011. Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. *Sci Rep,* 1**,** 13.

UNIPROT, C. 2008. The universal protein resource (UniProt). *Nucleic Acids Res,* 36**,** D190-5.

VAN BELKUM, A. 2011. *Classification of Bacterial Pathogens,* COGEM: Bilthoven, The Netherlands

VIALE, A. M. & ARAKAKI, A. K. 1994. The chaperone connection to the origins of the eukaryotic organelles. *FEBS Lett,* 341**,** 146-51.

WALLIN, I. E. 1922a. A Note on the Morphology of Bacteria Symbiotic in the Tissues of Higher Organisms. *J Bacteriol,* 7**,** 471-4.

WALLIN, I. E. 1922b. On the nature of mitochondria. I. Observations on mitochondria staining methods applied to bacteria. II. Reactions of bacteria to chemical treatment. *American Journal of Anatomy,* 30**,** 203-229.

WALLIN, I. E. 1922c. On the nature of mitochondria. III. The demonstration of mitochondria by bacteriological methods. IV. A comparative study of the morphogenesis of root-nodule bacteria and chloroplasts. *American Journal of Anatomy,* 30**,** 451-471.

WALLIN, I. E. 1927. *Symbionticism and the origin of species,* Baltimore,, Williams & Wilkins company.

WATANABE, K., KODAMA, Y. & HARAYAMA, S. 2001. Design and evaluation of PCR primers to amplify bacterial 16S ribosomal DNA fragments used for community fingerprinting. *Journal of microbiological methods,* 44**,** 253-262.

WOESE, C. R. & FOX, G. E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A,* 74**,** 5088-90.

WOESE, C. R., GUTELL, R., GUPTA, R. & NOLLER, H. F. 1983. Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. *Microbiol Rev,* 47**,** 621-69.

WOESE, C. R., KANDLER, O. & WHEELIS, M. L. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U S A,* 87**,** 4576-9.

YANG, B., WANG, Y. & QIAN, P.-Y. 2016a. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics,* 17.

YANG, B., WANG, Y. & QIAN, P. Y. 2016b. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics,* 17**,** 135.

YANG, D., OYAIZU, Y., OYAIZU, H., OLSEN, G. J. & WOESE, C. R. 1985. Mitochondrial origins. *Proc Natl Acad Sci U S A,* 82**,** 4443-7.

YILMAZ, P., PARFREY, L. W., YARZA, P., GERKEN, J., PRUESSE, E., QUAST, C., SCHWEER, T., PEPLIES, J., LUDWIG, W. & GLÖCKNER, F. O. 2013. The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Research,* 42**,** D643-D648.

ZHANG, Z., SCHWARTZ, S., WAGNER, L. & MILLER, W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol,* 7**,** 203-14.

ZUCKERKANDL, E. & PAULING, L. 1965. Molecules as documents of evolutionary history. *J Theor Biol,* 8**,** 357-66.