6-15-2021

# Minimal Genomes Simplify the Analysis of Biological Relationships in Medicine, Taxonomy, Ecology, and Astrobiology

Rafael Deliz Aguirre

# MINIMAL GENOMES SIMPLIFY THE ANALYSIS OF BIOLOGICAL RELATIONSHIPS IN MEDICINE, TAXONOMY, ECOLOGY, AND ASTROBIOLOGY

A Thesis
by
RAFAEL DELIZ-AGUIRRE

Submitted to Texas A&M International University
in partial fulfillment of the requirements
for the degree of
MASTER OF SCIENCE

December 2017

Major Subject: Biology

# MINIMAL GENOMES SIMPLIFY THE ANALYSIS OF BIOLOGICAL RELATIONSHIPS IN MEDICINE, TAXONOMY, ECOLOGY, AND ASTROBIOLOGY

A Thesis

by

RAFAEL DELIZ-AGUIRRE

Submitted to Texas A&M International University
in partial fulfillment of the requirements
for the degree of
MASTER OF SCIENCE

Approved as to style and content by:

| | |
|---|---|
| Chair of Committee, | Sebastian Schmidl |
| Committee Members, | Keith D. Combrink |
| | Michael R. Kidd |
| | Ruby A. Ynalvez |
| Head of Department, | Michael R. Kidd |

December 2017

Major Subject: Biology

# ABSTRACT

Minimal Genomes Simplify the Analysis of Biological Relationships in Medicine, Taxonomy, Ecology, and Astrobiology (December 2017)

Rafael Humberto Antonio Deliz Aguirre, B.S., Baylor University;
Chair of Committee: Dr. Sebastian Schmidl

As humans, we continuously thrive to determine meaningful biological relationships between living things on the basis of specific characteristics. Taxonomy is the science that studies these connections for the purpose of naming and grouping organisms as well as identifying logical clusters (i.e, clade). Its origin may be traced as far back as ancient Greece and beyond, and while we explore species diversity, most of our attention focuses on the biomedical and ecological applications of taxonomy. The contemporary taxonomical revolution began with Carl Linnaeus. He established the principles of binomial nomenclature and taxonomical hierarchy that we employ now. Relationships were later drawn as tree diagrams whose order described the evolutionary development of species but the branch distances didn't accurately reflect time. With Charles Darwin, genetics started to play a larger role in taxonomy, and cladograms were determined using relatedness. The rise of bioinformatics in the last century facilitated to calculate species divergence, resulting in accurate visuals of complex branch distances in the form of phylogenetic trees. Finally, the genomics revolution provided a wealth of information, but its sheer endless amount of data has been challenging to process. As a consequence, our ability to unfold the mysteries of biological evolution remain limited by technology since multi-species comparisons remain computationally intensive. To solve this problem, we used a new computational approach that is based on the analysis of organisms with small genomes to construct evolutionary relationships. Minimal genomes contain mostly the core set of genes, allowing the investigation of the origin of life, evolutionary connections, and potential antibiotic targets. By comparing genomics data of minimal genomes from all sequenced phyla, we observed that these organisms reflect the diversity of their genomically larger counterparts including GC content, proteins per megabase (Mb), and 16S rRNA relationships. Thus, minimal genomes are suitable to use in taxonomy studies.

We also compared the 16S rRNA of all species of the phylum Tenericutes as described in the *Bergey's Manual* as well as the proteomes from all mammalian *Mycoplasma* species. The Tenericutes, commonly known as "mycoplasmas," are bacteria that lack a cell wall, have notoriously small genomes, and are AT rich. Our results demonstrated that phylogenies at small scales are alarmingly contingent upon the sequence alignment algorithms that is used. In addition, comparison of the 16S rRNA of all Tenericutes revealed that these organisms are paraphyletic. Proteome alignments found computed homologs lacking. However, 16S rRNA data combined with statistics on host range, geographical distribution, and habitat (e.g., host organ system) revealed that there are common features within the clades that may be helpful for taxonomy studies.

Furthermore, our data supports the intention of other scientists to reorganize the taxonomy of the Mycoplasmatales order and its type species. Minimal genomes are therefore a source of untapped potential.

## ACKNOWLEDGEMENTS

I would like to express my sincerest appreciation to my thesis advisor, Sebastian Schmidl. His knowledge, perseverance, diplomacy, and understanding are virtues that have inspired me to excel. I would like to thank my thesis committee members, Keith D. Combrink, Michael R. Kidd, and Ruby A. Ynalvez, for their guidance, time and patience throughout the course of this research. Thanks to the College of Arts & Sciences (COAS) and the Graduate School (GS), for reviewing this thesis and travel grants.

I would also like to thank the whole Biology and Chemistry Department including all faculty, staff, and students for making my time at Texas A&M University a wonderful learning experience. I would like to especially thank Catalina I. Pislariu, Daniel J. Mott, Fernando G. Quintana, Feseha Abebe-Akele, Kameron R. Jorgensen, Lucia Cortez, Monica O. Mendez and Sang-Chul Nam. Similarly, thanks to the Department of Mathematics and Physics for research advice, including Juan H. Hinojosa, Peter S. Davis and Rohitha Goonatilake, and Roberto R. Heredia of the Department of Psychology and Communication. I'd like to also thank the people who have helped me with my research projects, including Adrian Dominguez and Jessica Perez of Environmental Health & Safety (EH&S), the Institutional Review Board (IRB) of the Texas A&M International University, Celeste Kidd and John Kilburn from the Office of Research and Sponsored Projects (ORSP), Kevin Lindberg and Thomas Mitchell of the Office of the Provost, Juanita Solis and Pablo Arenaz from the Office of the President, Rosie Benavides, the pathology laboratory and the IRB of Doctors Hospital of Laredo, Julia Perales and Rebeca Castro of the laboratory of the Laredo Health Department, and Rosaena Vero and the clinical microbiology laboratory of the Laredo Medical Center (LMC).

I would like to recognize the efforts of the undergraduates who helped me with my projects, including Alejandro D. Qiñones, Blanca L. Ortega, Carlos A. Resendiz, Carolina Alfaro, Cassandra Ramirez, Chrysantha Hernandez, Cristina E. Aguilera, Eric Villarreal, Heni Gutman, Isidro Cantu, Jose Longoria, Justin G. Tijerina, Priscilla Carlo-Polanco, and Samantha J. Jimenez, as well as my colleagues, Brian De Jesus, Maria Hinojosa, Janelle Vazquez, Travis Salinas and Yuridia Rodriguez.

My sincerest gratitude to the College of Arts & Sciences, and the Graduate School for the travel grants awarded to me, as well as the people who helped me with the application process. The work performed was done thanks to funding received from the Texas A&M International University. I'd also like to express my gratitude to the Laredo Next Generation Rotary Club and the League of United Latin American Citizens (LULAC) for their scholarship.

Many thanks to my family and friends for their encouragement. With their support, I was able to always move forward.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ABBREVIATIONS

| | |
|---|---|
| ± | mean plus/minus standard deviation |
| **AT** | Adenine plus Thymine |
| **bp** | Base Pair |
| **CRBC** | Computational Biology Research Center |
| **CRG** | Centre for Genomic Regulation |
| **DDH** | DNA-DNA Hybridization |
| **DNA** | Deoxyribonucleic Acid |
| **EMBL-EBI** | European Molecular Biology Laboratory: European Bioinformatics Institute |
| **GC** | Guanine plus Cytosine |
| **KEGG** | Kyoto Encyclopedia of Genes and Genomes |
| **LUCA** | Last Universal Common Ancestor |
| **MALDI-TOF MS** | Matrix-Assisted Laser Desorption/Ionization-Time of Flight Mass Spectrometry |
| **NCBI** | National Center for Biotechnology Information |
| **NJ** | Neighbor Joining |
| **nm** | Nanometer |
| **PPLO** | Pleuropneumoniae-Like Organisms |
| **RNA** | Ribonucleic Acid |
| **rRNA** | Ribosomal Ribonucleic Acid |
| **MAFFT** | Multiple Alignment using Fast Fourier Transform |
| **MUSCLE** | Multiple Sequence Comparison by Log-Expectation |
| **PCR** | Polymerase Chain Reaction |
| **UPGMA** | Unweighted Pair Group Method with Arithmetic Mean |

# GLOSSARY

**Cladogram:** Tree diagram of inferred or arbitrary relationships.

**Complex organic molecule:** In the astrobiological sense we employ, it means an organic molecule with multiple carbons.

**Conserved gene:** Gene present in multiple organisms that share sequence identities.

**Convergent evolution:** Independent evolution of similar features.

**Dendrograms:** Tree diagram (cladogram) where hierarchy is emphasized but not weighted.

**Essential genes:** Genes whose knockouts are lethal.

**Genome:** Set of all genes in an organism.

**Genotype:** Genetic makeup of an organism.

**Homologous genes:** Genes that share sequence and function.

**Minimal genome:** A genome with a reduced or atypically low number of genes.

**Molecular cloud:** Interstellar cloud whose increased density permits the formation of molecules.

**Mycoplasmas:** When in plural, bacteria without a cell wall of the phylum Tenericutes.

**Percent identity:** Percent of identical DNA or amino acid letters.

**Percent coverage:** Percent of the sequence that share similar or identical identities.

**Phenotype:** Visual expressions of the genes that are proteins, typically.

**Phylogram:** Tree diagram similar to the dendrogram but with calculated relationships that represent evolution.

**Prokaryote:** Domains Archaea and Bacteria; organisms that lack membrane-bound organelles.

**Proteome:** All proteins from an organism.

**Species:** We adopt *Bergey's Manual* definition, that is, >1% 16S rRNA difference if rRNA is the only resource available from that organism, or <70% DNA-DNA hybridization.

# INTRODUCTION

We will now discuss in a little more detail the struggle for existence.
— Charles Darwin's *On the Origin of the Species* [1]

Looks can be deceiving. Bacteria and humans may look nothing alike—our individual volumes are nine orders of magnitude apart—but when we look at our proteins, we find a world of parallels [2–4]. But what makes a bacterium a bacterium? Or what is the difference between Enterobacteria and *Mycoplasma* species? These questions were originally answered using phenotypic variations, visual appearances of heritable factors called genes [5,6]. With the beginning of the molecular biology revolution during the Cold War, genotypes superseded phenotypes. To better understand the nomenclature of bacteria, we shall first dive into the history of microbial taxonomy. Then, we shall explore how we can establish relationships among bacteria and lastly, we shall examine the implications of taxonomic studies beyond naming and classification.

## "Divide and Conquer:" A Brief History of Microbiology & Taxonomy

What's in a name? That which we call a rose by any other word would smell as sweet.
— William Shakespeare's *Romeo and Juliet*, Act II, Scene II [7]

### *The Dutch Golden Age*

The field of microbiology has its roots in the Netherlands in 1674. The Dutch Golden Age was at its peak when Antonie van Leeuwenhoek discovered by accident microbes, which he called *animalcules* [8,9]. Van Leeuwenhoek was a Delph tailor. As such, he used lenses for examining the quality of fabrics. Ever curious, he began to look elsewhere with his lenses. Unlike today's microscopes, van Leeuwenhoek's microscope had no interchangeable glass slides. Thus, to preserve the specimen, a new microscope had to be made. Van Leeuwenhoek, wishing to reexamine his samples and perhaps out of thriftiness considering his socio-economic status, developed a taste for lens making. His lenses became more and more powerful. Estimates place the total magnification of his lenses somewhere between 200-500 times [9,10]. In total, van Leeuwenhoek produced more than 566 lenses over his 50 years in research, though only nine survived [11]. His detailed accounts of his observations reached the ears of the Royal Society of London [12]. Overcoming his humble status as a mere merchant, van Leeuwenhoek found a place in the Royal Society of London despite his lack of knowledge of scholastic Latin and the rivalry between his native country and the British Empire due to the West Indies trade. The instrumental Society translated and disseminated his work [13], however not everything ran smoothly. Originally, van Leeuwenhoek's work was met with skepticism, and the Royal Society of London asked for

This thesis follows the journal model of the *Public Library of Science*.

witnesses to confirm his observations [14]. After satisfying the Royal Society with witnesses, channels of communication opened, and over 200 letters were exchanged. Unfortunately, the skill of making powerful lenses was a secret that van Leeuwenhoek took to the grave, and with it, microbiology to an extent. Much speculation exists on whether he grinded the lenses or made droplets by melting a glass filament [15,16]. Whichever the case may be, microbiology suffered an interruption. It took another 150 years to reach van Leeuwenhoek's magnification power again [16,17].

Half a century after van Leeuwenhoek, in 1735, the Swedish scientist Carl Linnaeus traveled to the Netherlands to advance his work on taxonomy. After quickly completing his medical degree at Harderwijk (a diploma mill at the time), Linnaeus went to Leiden [18]. There, he became acquainted with Johan Frederik Gronovius and Isaac Lawson, both who would fund Linnaeus' *Systema Naturae*, a treatise on taxonomy which established the binomial system of today. Linnaeus met with Herman Boerhaave, an esteemed physician and botanist who would later aid Linnaeus in establishing his eminence [19–22]. Having been set on the path to success at the Netherlands, Linnaeus returned to Sweden and rose through the ranks. He was appointed *rector Magnificu*s of Uppsala University in 1750, 1759, and 1772 for six-month terms [23]. In *Systema Naturae*'s 10th edition, published in Stockholm in 1758, the *animalcules* found a home in the genus *Volvox* of the kingdom Animalia, phylum Vermes, class Zoophyta [24]. It marked the first time that a microbe had a taxonomical name. This classification was made possible with earlier drawings by microscopist Henry Baker [25].

### *Developments in the Kingdom of Prussia of the German Empire*

One century later, microbiology was on the rise again in Prussia, now Germany. The word bacteria was first used by Christian Gottfried Ehrenberg in 1838, when he coined the term from the Greek word "bakteria," meaning little stick. In 1866, Ernst Haeckel united bacteria with the "blue-green algae" (i.e., cyanobacteria) to the phylum Monera of the kingdom Protista in an attempt to erase the boundaries between life and the abiotic [26,27]. Thus, Protista would include unicellular organisms and simple multicellular organisms [28]. With the exceptions of the work of Felix Dujardin and Christian Gottfried Ehrenberg, bacterial taxonomy was disorganized at this point in history, resulting in plenty of overlapping terms for the same species [29]. This was further complicated by the works of C. A. Theodor Billroth, Lev Cienkowski, Ray Lankester, Carl von Naegeli, Eugene Warming, and Wilhelm Zopf, all who hypothesized that the endless diversity of bacterial morphologies represents just different developmental phases of the same bacterium [30]. Around 1868, Ferdinand Cohn, the first person to notably classify bacteria as plants, began sorting bacterial species based on morphology, systematically classifying them as sphaerobacteria (round) under the genus *Micrococcus*, microbacteria (short rods) under the genus *Bacterium*, desmobacteria (long, filamentous rods) under the genera *Bacillus* and *Vibrio*, and spirobacteria (spiral or screw-like) under the genera *Spirillum* and *Spirochaeta* [26,31]. To this day, the simplistic terms cocci (round), bacilli (rods), and spirilli (spirals) are still used to describe bacterial morphologies [32,33].

Bacteria are difficult to observe using light microscopes without staining the cells. The Gram stain was invented near the morgues of Berlin in 1883 [34]. Hans Christian Joachim Gram, working under Carl Friedländer, accidentally discovered a technique that would make bacteria distinguishable from their host under the microscope [35]. Friedländer recognized that without this stain, bacteria could not be differentiated from human cells. This revolutionized microscopy dramatically since bacterial cells were nearly invisible before this invention. Earlier in 1863, von Waldeyer found that hematoxylin, a natural dye brought to Europe by the Spanish conquistadors, was useful in staining human nuclei but couldn't distinguish bacteria from human tissue [36]. Moreover, cell staining had been attempted as early as the 17th century, when Antonie van Leeuwenhoek would use saffron to dye muscle fibers [37]. The Gram stain would further split bacteria into two groups: the crystal violet absorbing Gram-positives and the Gram-negatives that cannot absorb the former stain [38]. To date, the Gram stain remains a valuable diagnostic tool at the clinic, competing with molecular diagnostics, and is often times the first step in the path to identify a patient's pathogen due to its quick and easy application as well as the correlation with contemporary taxonomy [39].

Robert Koch, using his own staining methods, decisively determined that bacteria cause diseases, giving rise to the germ theory, and the development of the Koch postulates (also known as the Henle-Koch postulates) in 1884: (1) correlate, (2) isolate, (3) inoculate, and (4) re-isolate and inoculate the bacterial species associated with disease [40,41]. These postulates, which provided an explanation to disease, rose from earlier concepts written by Edwin Klebs, a pathologist and disciple of Rudolf Virchow, and Jakob Henle, an academic advisor to Koch while he was training at the University of Göttingen [42]. Henle had proposed in 1840 that bacteria should be isolated and studied separately to prove if they cause disease [40,42]. With this in mind, Koch sought to isolate pure bacterial cultures between 1876 and 1884, and succeeded thanks to Fanny Hesse, one of his technicians who was also married to one of Koch's post-doctoral fellows, Walther Hesse. The Hesse family suggested to try agar instead of gelatin which melts at body temperature and had been extensively used but with poor results by Koch [43]. The work of Koch gave support to earlier claims of rival Frenchman Louis Pasteur. Between 1860 and 1864, Pasteur conducted experiments establishing cause-effect relationships between bacteria and disease, respectively, but failed to provide a mechanism, which Koch later found [44]. Koch and Pasteur shifted the field of microbiology from plain microscopy to pure culturing.

### *Defining the Prokaryotes*

Bacteria were elevated to the rank of kingdom in 1938 when Herbert Copeland proposed that the phylum *Monera* be reformed as kingdom Monera, capturing all organisms void of a nucleus [45]. Herbert Copeland's idea was an elaboration partily motivated by the works of his father, Edwin B. Copeland. Edwin Copeland argued for a third kingdom, that of bacteria, in *What is a Plant?* [46] However, Edwin Copeland did not provide enough reasons as to why merit a separation of bacteria from other kingdoms.

Three centuries had elapsed since the discovery of microbiology and a question as simple as "What are bacteria?" could not be answered definitively, not until 1962 when Cornelis B. van Niel and a former student of his, Roger Y. Stanier, published *The concept of a bacterium [47]*. This paper gave birth to the prokaryote-eukaryote dichotomy we know, and separated the two by the absence or presence of a nuclear envelope, respectively. Though the duo cite Edouard Chatton as the father of the dichotomy, there is not enough evidence pointing Chatton as the father [26,47,48]. Jan Sapp, a historian who focuses on biology, calls this misattribution a mythological tale [48].

Instead, if not Haeckel, then it was Edmund Beecher Wilson the one who argued for the separation of non-nucleated from nucleated life forms with the understanding that life forms with membrane bound nuclei were the more advanced forms [26,49]. In 1896, Wilson said bacteria may be completely different from other life forms, yet he recognized the limitations of the microscope, as described by Ernst Abbe in 1873, and was cautious enough to say "if this identification [of the nucleus] is correct" [49,50]. We further elaborate that the dimensions of *Escherichia coli* are around 1μm by 2 μm [51]. Light microscopes have a wavelength range from 380 nm to 750 nm, and a resolution of roughly 200 nm. Therefore, observing bacteria using light microscopes is challenging. The issue would be further compounded by any attempts to observe smaller structures in bacteria. Cell nuclei measure between 2-10 μm, and have a correlation with its genome size, 1bp being roughly equal to 1nm$^3$ [51]. Therefore, observing nuclei in smaller genomes, say *E. coli* str. K-12 substr. MG1655, would theoretically mean a 0.207 μm diameter nucleus, which would thus require an electron microscope to observe such fine detail and reach solid conclusions [52]. The electron microscope wasn't invented until 1931 by Ernst Ruska and Maximillion Knoll [53]. In synthesis, bacterial nuclei couldn't be confirmed before the transmission electron microscope.

It should be noted that Stanier and van Niel make no mention of the Monera nor the Copelands in their 1962 paper [47]. Most likely this was due to the objective of the paper, which was to describe cellular structures and not taxa [26]. Moreover, the duo had discussed taxonomy, including Herbert Copeland, two decades before in *Main Outlines of Bacterial Classification* [54].

Taxonomy would be remodeled again in 1969 by Robert H. Whittaker [28]. As an ecologist, he stressed function in classification. For this reason, he carved kingdom Fungi [55]. He also supported Copeland's Monera after the 1962 paper by Stanier and van Niel received widespread acceptance, and added that the Monera were not monophyletic [28]. This view was well received and persisted until 1990 [6].

The last major, widely-accepted change in taxonomy would be delivered in 1977 by Carl R. Woese and George E. Fox and accepted in 1990 [6,56]. The Cold War had pushed towards a molecular biology agenda, focusing energies in biology down to cellular structures [55]. Riding this new wave, Carl R. Woese decided to focus on the ribosome. He noted that after performing 2-D electrophoresis with an rRNA T1 RNase digest, a "fingerprint" would be produced [56]. These blots were quantified and led to one of the first quantitative phylogenetic "trees" (it was originally published as a table) [57]. With it, Woese was able to demonstrate that the prokaryotes were not a cohesive group after discovering that Archaea were a distinct group [6].

Since Woese, other taxonomic proposals at the kingdom level have emerged, notably, those of Thomas Cavalier-Smith [58–60]. Recently, Ruggiero, also collaborating with Cavalier-Smith, compiled 138 sources and united them in the Catalogue of Life [61]. Most of these new taxonomies are focused on microorganisms, have seen multiple conflicting edits, and have varying degrees of acceptance [60,62–64].

Like the ideas set forth by Linnaeus, current classification of microbes rely on characterizing the organism, biochemically nowadays. With current technologies, most dissections require a considerable number of monoclonal copies up in the millions in order for it to be studied. Despite advances in microfluidics that have allowed single-cell studies and culturing of fastidious organisms, most microbes haven't been able to be cultured, thereby experimented on [65–70]. At present, 99% of the microbial world remains unculturable and unknown, a phenomena known as the "Great Plate Count Anomaly."[71–74] The 16S sequence remains as pivotal as ever in classifying unknown life forms [75].

## Interconnecting the Living with Phylogenetics

Though this be madness, yet there is method in't.
— William Shakespeare's *Hamlet*, Act II, Scene II [76]

### *Pre-Darwinian Linnaean Phylogenetics*

The idea of classifying things could be tied to our primitive need for language, that is, to name things and organize thoughts for communication purposes [77–80]. For Carl Linnaeus, this drive was further catalyzed by Christian philosophy at the time. The son of a Christian pastor, Linnaeus was taught by his father, Nils, that everything had a name, as Genesis 2:19 says, "So from the soil Yahweh God fashioned all the wild animals and all the birds of heaven. These he brought to the man to see what he would call them; each one was to bear the name the man would give it" [23,81]. He was also taught that studying nature equated to admiring God's work [82]. Not surprisingly, the twelfth edition of *Systema Naturae* opens with an allusion to the works of God, "How countless are your works, Yahweh, all of them made so wisely! The earth is full of your creatures" (Psalm 104:24) [81,83]. This was the original motive for establishing the Linnaean taxonomic system. However, it must be noted that this method was not essentialist nor arbitrary. Linnaeus recognized that there were patterns in nature, evident in his grouping of plants according to the morphology of their sexual organs (e.g., flowers) [84–86].

### *Darwinian Phylogenetics*

Charles Darwin recognized that the patterns, or "endless forms most beautiful", Linnaeus had observed were because there had been common ancestor between the different species, and that time led to variation [1,87]. Darwin incorporated evolution into taxonomic trees, or the "Tree of Life," as he would call it *On the Origin of the Species* [1]. Darwin was aware of his limitations, and prophetically said to Huxley on a letter written on 1857, "The time will come I believe, though I shall not live to see it, when we shall have very fairly true genealogical trees of each great kingdom of nature" [88].

### *Computational Phylogenetics*

As early as the 1950s, large computers would be employed for storing biological data. However, it was not until the 1960s that it became increasingly apparent that crunching biological data by hand would be impractical. Therefore, the machine, which was becoming more accessible, entered the lab. Frederick Sanger pioneered protein sequencing. Between 1951 and 1953, he published the complete sequence of insulin [89–93]. This feat would earn Sanger the 1958 Nobel Prize in Chemistry [94]. He then attempted to sequence RNA by refining his insulin sequencing methods. Sanger succeeded in determining the 5S rRNA sequence of *E. coli*, a 120 nucleotide sequence, in 1968 [95]. After, Sanger tried DNA sequencing [96,97]. He succeeded. This accomplishment would earn him his second Nobel Prize in Chemistry in 1980 [98].

At the same time, Margaret Oakley Dayhoff, the mother and father of bioinformatics, championed the use of computers in biology early on [99,100]. Though her contributions to science are numerable (e.g., one letter amino acid codes, thermodynamic models for planetary atmospheres), of notable importance to this thesis is her use of protein sequences to deduce phylogenies back in 1966 and draw them in 1969 [101–108].

Carl Woese was able to fulfil Darwin's prophecy of "fairly true" phylogenetic trees during the 1970s. As previously discussed, Woese originally used blots of rRNA digests to establish phylogenetic relationships among bacteria [56]. Eventually, DNA sequences coding for the 16S rRNA were used [109]. The benefits of the 16S are numerable [110]. As a core gene, it is ubiquitous, with sufficiently conserved regions to permit the design of universal primers [111–115]. The 16S has retained its function. It is long enough (1,500 bp) to permit comparisons [116]. It allows for molecular sequences and not morphologies, fossil records nor biochemistries to determine relationships [6]. However, it does not account for horizontal gene transfer [117]. Horizontal gene transfer has also been noted within segments of the 16S, that is, homologous recombination within the 16S has been observed [118]. There are also multiple copies of the 16S within the same genome in the majority (75%) of bacteria with considerable differences above the species threshold (1.3% different) in a minority (3.4%) of the duplicate cases, or 2.5% overall [119]. Nevertheless, the 16S rRNA remains the gold standard for drawing quick phylogenies, perhaps more appropriately termed the "silver" standard if we consider DNA-DNA hybridization [109,116,120–124].

Other ubiquitous, conserved core genes have been proposed for drawing phylogenies, that is, relationship trees. Not surprisingly, most of these alternatives serve as targets for broad-spectrum antibiotics due to their conserved features. To name a few genes used in phylogenies that also relate to antibiotics are the following: rifamycins target RNA polymerase gene *rpoB*; quinolones target DNA topoisomerase gene *par*C and

DNA gyrase A and B genes *gyrA* and *gyrB*, respectively; kirromycin, enacyloxin, pulvomycin, and GE2270 (also called as MDL 62,879) target Elongation Factor Tu gene *tuf* [125–127]. The 16S itself also has drugs that target it, including tetracyclines and aminoglycosides [128]. Therefore, phylogenetics and drug discovery in the form of broad-spectrum antibiotics go hand in hand.

The most notable alternative to the 16S has been the *rpoB* gene which encodes for the β subunit of the RNA polymerase, a molecule responsible for transcription [129,130]. The *rpoB* gene, suggested as both, a supplement and an alternative, offers the benefit of having only one copy per genome [130–133]. The sequence, whose length is around 3411-4185 bp, offers better resolution at the species level than the 16S rRNA, which is around 1541 bp long [130,134]. The integrity of the RNA polymerase can be validated *in silico* through the detection of nonsense mutations (premature stop codons) after being electronically translated. However, a major drawback of *rpoB* is that no universal PCR primers have been found because it lacks conservation beyond the phylum level. Unfortunately, as Woese notes, there is "No consistent organismal phylogeny has emerged from the many individual protein phylogenies so far" [135].

A more radical approach to phylogenetics would be to think outside sequences. Matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) data is growing, mainly due to its ability to resolve strains. At present, most trees drawn with MALDI-TOF MS are focused on species. Mass spectrometry trees using digests have been explored and deemed plausible [136]. However, they remain to be evaluated beyond the family rank.

## Defining Parameters in Computational Phylogenetics

Comparing nucleic acid or protein sequences is at the core of current biology. Its goal is to identify similarities for deducing relationships as well as functional and structural similarities.

### *What is a Species?*

The definition of a species is of contentious debate in biological taxonomy and philosophical ontology, ranging from the conjecture that species are a made-up concept down to the Darwinian belief that species exist in nature [137–156]. Many, including Woese, note that horizontal gene transfer and homologous recombination make "fuzzy" species [137–141]. Others note that defining organismal species depends on the definition of the word "species" [152,153]. Mayden has identified over 22 definitions of species [155,157].

Darwin was aware of the species definition issue, and wrote to Hooker in 1856, noting that there are different notions of what a species is, perhaps because it is indefinable [88]. *On the Origin of the Species* (1859), he noted that the term species is arbitrary, but that there is an greater picture, like stars forming constellations [1].

When it comes to the 16S rRNA, various thresholds based on percent identity (i.e., percent similar), all ≥95%, have been suggested for defining species [110,116,158,159]. These studies have used an almost universally agreed 70% threshold for the labor intensive DNA-DNA hybridization (DDH), the "real" gold-standard (versus rRNA) for species identification, and *in silico* DDH or whole-genome sequences comparisons as a reference points [160–162]. *Mycobacterium chelonae* and *Mycobacterium abscessus* have 16S sequences that are 99% similar, yet exhibit 35% similarity in DDH [160].

*Performing Alignments*

The question of how do we determine percent identities in sequences must be then considered for it has a great impact on the classification of species. Various methods and "providers" exist for comparing DNA, RNA, and protein sequences. As previously discussed, these have been employed in the study of taxonomy and phylogenies since the late 1960s by Margaret O. Dayhoff. Starting with the The European Molecular Biology Laboratory: The European Bioinformatics Institute (EMBL-EBI) based in Germany, this institute offers web servers for conducting multiple sequence alignments (MSAs) using Clustal Omega, Kalign, Multiple Alignment using Fast Fourier Transform (MAFFT), Multiple Sequence Comparison by Log-Expectation (MUSCLE), MView, and T-Coffee [163,164]. The Computational Biology Research Consortium based in Japan offers their authored MAFFT [165,166]. The United States' National Center for Biotechnology Information (NCBI) Genome Workbench may be combined with MUSCLE as provided by the source, Edgar, to produce trees and alignments [167,168]. The Centre for Genomic Regulation (CRG) based in Barcelona offers T-Coffee as well.

Multiple sequence alignments assume that there is homology between the input sequences. They use sum of pair scores to decide which alignment is best. This matrix aggregates identical letters and penalize gaps and mismatches. Broadly speaking, there are global methods that are highly accurate but slow, and there are heuristic methods that represent a compromise between quality and speed. Our aforementioned MSA methods are heuristic, meaning that they compute scores in blocks. In terms of pairwise alignment methods, there are global and local approaches. Global pairwise alignments attempt to align two sequences from start to end. The local pairwise alignment, as the name implies, aligns two sequences by identifying highly similar subsets. At present, most MSA methods rely on progressive methods, meaning a combination of these two approaches [169]. The first program to use progressive methods was Clustal Omega [170]. Alternatively, star alignments find the most similar sequence to all others, and then uses it as an anchor for spotting the differences between the other sequences.

When A, B, and C are compared, three alignments may be performed: A-B, A-C, and B-C. A-B and A-C alignments imply an indirect B-C alignment that may not be similar to a direct B-C comparison, thus association fallacies are of concern. Looking for consistent patterns within associations, T-Coffee was developed [171]. To refine the alignment, iterations may be performed. This was the reasoning behind the creation of MUSCLE [168] and MAFFT [172]. A drawback to them is that errors are propagated, especially with gaps, and like the butterfly effect, minuscule differences may lead to completely different products.

When dealing with closely related sequences, DNA works best at finding the differences. Inversely, when dealing with distantly related sequences, comparing amino acids is the best approach. To improve results, scoring may be adjusted, combined, and weighted. Gap penalties are also a useful parameter as once a gap is declared in an iteration, it is always a gap. Adjustments to these include penalizing gaps in hydrophobic regions differently from hydrophilic regions.

### *Drawing Trees*

Plotting phylogenetic trees has been done since 1801 [173]. Nowadays, trees are done using a variety of methods, including whole-genome alignment trees [174] and aggregate ribosomal alignments [175]. Quantitatively speaking, there are numerable ways to represent relationships. Maximum parsimony, based on Occam's razor also known as the "law" of parsimony, focuses on the simplest tree that has the least number of common ancestors. It may be computed in a variety of ways. When only a handful of organisms are present, a brute-force search may be performed, testing for all possible trees. For more complicated trees, heuristic searches are conducted. In either case, its final tree is meaningful, but time consuming. When a draft is desired, distance matrix methods such as neighbor joining (NJ) are desirable for their speed [176].

## **Going Small with Minimal Genomes**

### *Natural Minimal Genomes*

There are two kinds of minimal genomes. A minimal genome may be interpreted as the smallest genome size or it could mean a self-sustaining autotroph with the least amount of genes. We opted for the former definition: small genome size, regardless if it is parasitic or not. We opted to study these because determining conclusively that the minimal genomes represent a compendium of larger genomes. They offer the advantage of naturally selecting core genes through a process called reductive evolution. Owing to their reduced genome size, they are far less computationally intensive. For this reason, we decided to focus our studies on minimal genomes.

The minimal genome approach could have perceived drawbacks. More specifically, Drake (1991) noticed that the genome size and mutation rates are inversely proportional, that is, minimal genomes are highly mutated in microbes [177–180]. This is known as "Drake's rule." However, Lynch suggests that it might be an incomplete picture since it does not fit well with eukaryotes nor prokaryotes, unless bacteriophages are counted in [181,182].

Comparative studies on minimal genomes elute the most conserved genes. Genes identified this way may certainly be used for a multitude of fields. Essential, conserved genes are oftentimes used as antibiotic targets in drug discovery, as previously highlighted. Inversely speaking, identifying non-conserved genes could help us tap into new antibiotics. Genes present in a phyla can serve as markers and identify species and strains, just like *rpoB*, thereby cost-effectively resolving taxonomic issues. Conserved genes can also be used for deducing the genome of protobionts, including the Last Universal Common Ancestor (LUCA), and help explain abiogenesis. In line with astrobiology, knowing what genes are necessary to constitute life could aid in space exploration by searching for said genes elsewhere and test the panspermia hypothesis. Noting that the abundance of the chemical elements are conserved throughout the Universe, cosmic convergent evolution is a possibility [183]. We have already observed evidence that demonstrate that amino acids, including glycine, are found in space in places like the Moon, comets, space dust, meteoroids and Enceladus (Saturn's moon) [184–189]. Other astronomically complex organic molecules such as alcohols, carboxylic acids, esters, nitriles, and epoxides have also been observed in molecular clouds and protoplanetary disks [190–193].

### *The Tenericutes, Otherwise Known as the Mycoplasmas*

The Tenericutes, commonly called the mycoplasmas, have notoriously small genomes. The first acknowledgement of the existence of them was in 1898, when Nocard and Roux isolated the first Tenericute, bovine *Mycoplasma mycoides* [194]. For half a century, they were known as pleuropneumoniae-like organisms (PPLO), and were thought to be taxonomically unique. Klieneberger (1930) suggested that they were simply bacteria lacking cell walls living symbiotically with other, walled bacteria, and that mycoplasmas were not as unique as previously thought [195,196]. The conflicting views were reconciled in the 1960s, when the guanine-plus-cytosine (GC) content assays and DNA-DNA hybridization assays showed that mycoplasma were indeed unique, and electron microscope images showed that they lack a cell wall. Dienes and Edsall (1937) detected the first mycoplasma isolated from humans in a Bartholin's (vaginal) gland abscess, possibly *Mycoplasma hominis [197]*. By the 1950s, a group of genitourinary mycoplasmas now known as *Ureaplasma* had been described. Eaton *et al.* (1944) isolated *Mycoplasma pneumoniae*, then known as Eaton agent [198,199]. To date, this "maiden-name" is still used in some circles. From the 1950s to the 1960s, field studies proved that Eaton agent caused lower respiratory tract infections. However, it was considered to be viral up until antibiotics were shown to be effective against it. Marmion and Goodburn (1961) suggested that the Eaton agent was a PPLO [200]. Chanock *et al.* (1963) cultured the agent on cell-free medium and proposed the current taxonomic designation, *M. pneumoniae* [201].

Phylum Tenericutes only has one class, Mollicutes. All share a lack of cell wall as a hallmark. The term mycoplasmas has been synonymous with bacteria without a cell wall. Yet, there are other bacteria that lack a cell wall, but this is believed to be a transient state for non-Tenericutes. Most are fastidious organisms of unusually small dimensions and genome size. *Mycoplasma genitalium* is one of the organisms with the smallest dimension and fewest genes in the world. They were believed to have evolved through reductive evolution from Gram-positive Firmicutes, hence their minimal genomes. They are part of the normal flora of many vertebrates, living commensally among them. Occasionally, they become parasitic when they invade other organs that they're not normally part of. The reason behind this remains unknown. A few other are known to be saprotrophic (eat decaying matter). Colonies are transparent in solid media, appearing like a fried egg under the microscope. There are five orders under the Mollicutes, most have only one family under them and two genus. The exception are Entoplasmatales, which have two families under them. Entoplasmatalaes. Below are the different sub-categories (**Fig. 1**). Hereinafter, the Tenericutes information displayed was obtained from the *Bergey's Manual* [202].

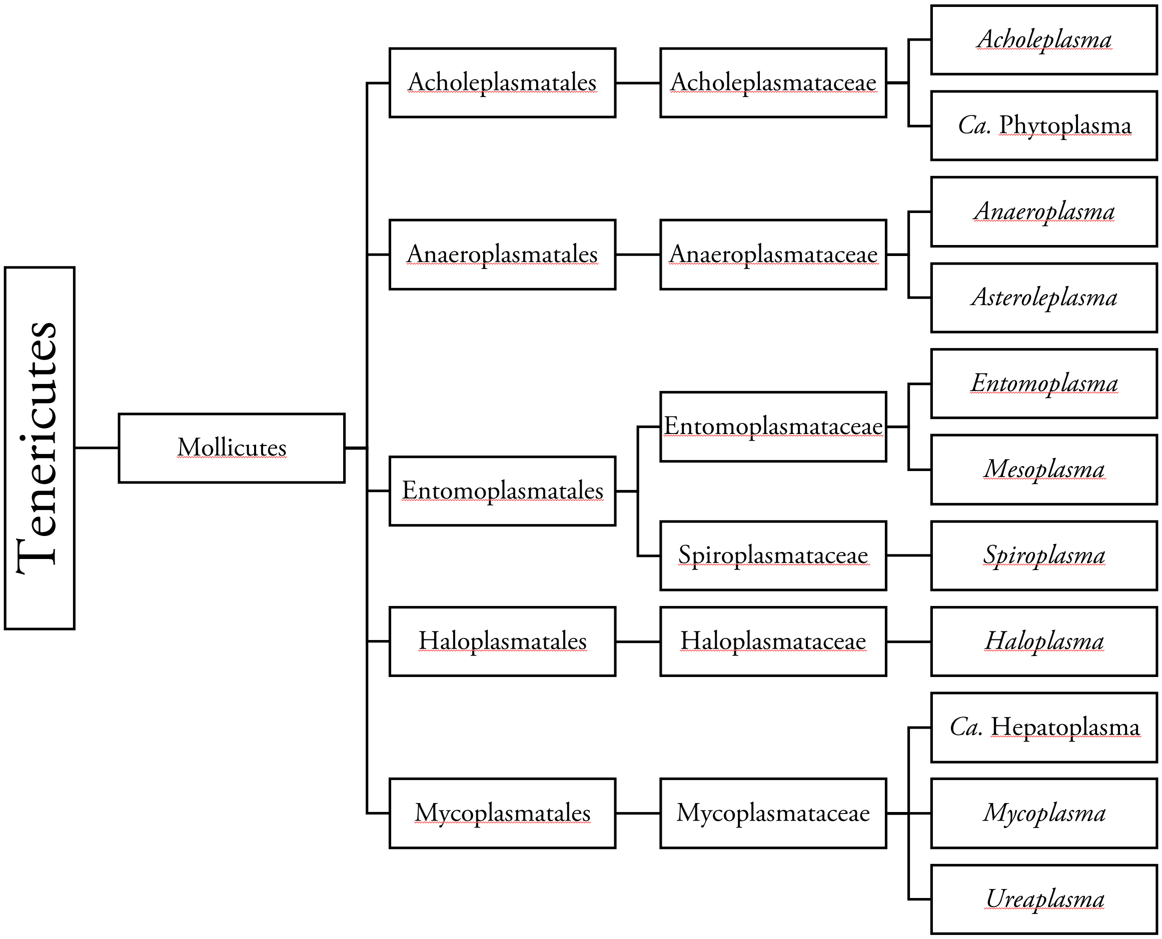


**Fig. 1. The Tenericutes cladogram**. From left to right, phyla, class, order, family, and genus at the extreme right. The Tenericutes are paraphyletic, from genera to orders. The phylum type species, *Mycoplasma mycoides*, is genetically an Entomoplasmatales. Branch distances not used. Cladogram tree drawn using Microsoft Word.

The type order are the Mycoplasmatales. The bases UGA code for tryptophan, unlike most organisms which UGA represents a stop codon. This exception to the universal codon code has been seen in other organisms with small genomes, but there is no concrete evidence that this has to do with adaptation [203]. They require sterol to grow. Most infect vertebrates. Morphologically, they are spherical to filamentous. Its only family is Mycoplasmataceae. There are three genera under the Mycoplasmataceae. *Mycoplasma* is the type genus, and *Mycoplasma mycoides* the type species. These are non-motile and found mostly in vertebrates, except *Mycoplasma iowae* and *Mycoplasma equigenitalium*. Around 133 *Mycoplasma* species exist, with another 19 candidates pending confirmation. Thirteen are known to affect humans, two affect primates in general and *Mycoplasma haemofelis* is believed to also affect humans. *Mycoplasma* exhibit a high degree of host specificity. *Ureaplasma* are highly similar to *Mycoplasma*. They are able to hydrolyze urea. Lastly, *Candidatus* Hepatoplasma is present in blood.

Acholeplasmatales is another order under the Mollicutes, and they only have one family, Acholeplasmataceae. They require no sterol for growing, hence their name, "a", no, "chole", sterol. The bases UGA code for a stop codon in the Acholeplasmatales. Genus *Acholeplasma* affects animals. It is non-motile and can produce fatty acids from acetate. *Candidatus* Phytoplasma is a proposed genus that affects plants, typically the phloem. It is nutritionally fastidious, hence its candidate designation. They have a characteristic 16S DNA sequence that sets them apart: 5'-CAAGAYBATKATGTKTAGCYGGDCT-3'. Many have plasmids.

Anaeroplasmatales is the third order we'll discuss, which only has family Anaeroplasmataceae. They are strictly anaerobe, hence their name. Genus *Anaeroplasma* requires sterol, is non-motile and like many anaerobes, most ferment. The other genus is *Asteroplasma* which requires no sterol and has a higher GC content.

Order Entoplasmatales infect arthropods and plants. Codon UGA transcribes tryptophan. Family Entomoplasmataceae are non-helical, non-motile. Their phylogeny overlaps with *Mycoplasma*. Its type genus, *Entoplasma*, requires sterol for growth, and is believed to arise in plants due to insect transmission. These bacteria are filterable. Mesoplasma require no sterol, but need polyoxyethylene sorbitan to be cultured. *Mesoplasma pleciae* may be an *Acholeplasma* because UGG codes for tryptophan and it has no known pathogenecity. The other family under Entoplasmatales are the Spiroplasmataceae. As its name implies, their morphology is spiral. Its only genus *Spiroplasma* is composed of motile bacteria. Its colonies look diffused in solid media. They are pathogenic and localized in insect guts and plant surfaces. They are also susceptible to viruses. Interestingly, the disort sex ratios in *Drosophila*.

### *Synthetic Minimal Genomes*

There are various projects for developing synthetic bacteria. They have made extensive use of minimal genomes. The J Craig Venter Institute developed *Mycoplasma laboratorium* from *Mycoplasma capricolum* and *M. mycoides* [204,205]. The *Mycoplasma* offers the advantage that it already has a small genome and tools are readily available for its manipulation [206]. Another ongoing project for creating artificial life is miniBacillus, which uses *Bacillus subtilis* as the backbone for a minimal genome [207].

## Unifying Concepts

If an elderly but distinguished scientist says that something is possible, he is almost certainly right; but if he says that it is impossible, he is very probably wrong.

— Arthur C. Clarke [208]

*Philosophy of Language and Its Implications in Taxonomy*

Contemporary biological taxonomy started out as a contemplation of the works of God. Nowadays, it is a field at the core of biology and clinical medicine. Whereas taxonomy, like language, is socially constructed by men, nature is built on evolution. Therefore, we must be cautious and say that at present, no taxonomical system, word or visual representation may be able to faithfully reproduce the relationships that exist among living organisms. There is no ideal word or language capable of containing all concepts tied to an organism into its name in this physical world at this time, not unless we consider philosophical idealism or computer simulations, in which case, bacteria are computer codes.

Like computer coding, we have to focus taxonomic nomenclature on intent. What is it that we are trying to say when we name bacteria? At present, most of this means rRNA correlations. In the near future, that may mean whole-genome sequence correlations, reviving DNA-DNA hybridization with the *in silico* twist. Whichever the case, there is still a need for "adjectives" such as *bla*$_{\text{CTX-M-15}}$ to indicate the presence of a β-lactamase (*bla*) first identified in Munich, (M) prevalent everywhere but in Spain that confers resistance to third generation β-lactams like ceftriaxone (CTX), to organisms like *Klebsiella pneumoniae* [209,210]. Lengthy, but cryptic descriptors such as these are commonly encountered in clinical microbiology and represent our best attempt yet at describing things in biomedicine.

Words are image projections in our brains, as postulated by philosopher Ludwig Wittgenstein and later confirmed by neurobiology [211]. Words also represent relationships. For example, the words chess, checkers, poker, baseball, and basketball all evoke the concept of games. Zooming in, we notice that these also possess strategy, patience, rules, and competitions with winners and losers. We must think of organisms and taxonomy in a similar fashion, as networks of relationships and common features, beyond the scope of hierarchies and cladograms.

Taxonomy is in a way biologically ingrained. Neurobiological experiments have shown visual categorization in the brain of nonhuman primates [212–215]. This could suggest a biological need for categorizing. Language, representing the images in our brains, therefore reflects these broad neural categorizations.

## Aims

We used bacteria with the smallest genome size for our concept of minimal bacterial genomes, and not autotrophic cells that are self-sufficient and require the least from their environment. We wanted to determine which genes are present in most living organisms [205,216–219], and not how many genes are required to sustain life. Morowitz (1984) had suggested to use *Mycoplasma* spp. as model organisms for life in the form of minimal metabolism [220]. However, tt quickly became apparent through our independent analysis of proteomes and 16S rRNA sequences that the *Mycoplasma* spp. are paraphyletic [221]. Therefore, since we found few sequenced species grouping together in clades, we opted for the big picture. As a result, we examined the most minimal genomes from each bacterial phylum recognized by the *Bergey's Manual.* This novel method takes advantage of reductive evolution, which eliminates all but the most essential genes, while accounting for the rich diversity of lifeforms found in domain Bacteria.

Comparative genomics (*in silico*) methods like ours offer cost-effective solutions for identifying core and essential genes [205,217,218,222]. Its predictive abilities may be combined with *in vitro* methods for further validation and site-directed mutagenesis. Other alternatives to identify essential genes exist, namely conservative and replicative transposon mutagenesis, antisense RNA, suicide plasmids [218], and CRISPR [223]. Their advantage is tempered by the time investments and financial resources needed to implement them.

This project focused solely on the viability of comparative genomics for obtaining core and essential genes in the context of minimal genomes. Several observations and conjectures exist for minimal genomes, including that they're biased towards becoming AT rich [203,224,225], and that their mutation rates are higher [177].

We hypothesized that minimal genomes share features with their larger counterparts. We expected to observe conservation of phylogenies [175] unaffected by mutation rates, and that despite the AT bias, using bacteria from all phyla would nevertheless produce GC contents ranging between 25-75% [226,227].

# RESULTS

## Minimal Genomes

We studied the most minimal genomes (n=131) in twenty-nine (29) of thirty (30) phyla recognized by the *Bergey's Manual*. The remaining phylum has no sequenced organisms. At least one species was selected from each phyla. We sampled 4.5± 2.5 species per phylum. The average genome size of our samples were 2.32± 1.44 Mb. Descriptive statistics are on **Table 1**.

The smallest characterized genome in our study belongs to *Mycoplasma parvum* (564kb) of the Tenericutes (i.e., mycoplasmas) which also had the least amount of genes at 568, and the largest minimal genome in a phyla belonged to *Sphingobacterium* sp. ML3W (5.33 Mb) of the Sphingobacteria which was the runner-up in terms of most genes (4530). *Gemmatirosa kalamazoonesis* KBS708 of the Gemmatimonadetes had the most genes for a minimal genome in a phyla (4567). The lowest GC content belonged to *Blattabacterium* sp. (Blaberus giganteus) of the Bacteroidetes (25.7%) followed closely by *Mycoplasma parvum* str. Indiana (27.0%), and the highest GC content was registered for Phycisphaera mikurensis NBRC 102666 of the Planctomycetes (73.3%) followed by *Gemmatirosa kalamazoonesis* KBS708 of the Gemmatimonadetes (72.6%). More details on the most minimal sequenced genome for each phyla is available on **Table 2**. The correlation between number of genes and genome size is consistent with previous findings (**Table 3**, **Fig. 2**). We compared genome size to the number of coding proteins per Mb, GC content, and GC content to number of proteins and number of coding proteins per Mb (**Fig. 3**). We drew trees to include phyla, GC content and genome size (**Fig. 4**). Most phyla shared clades in the tree. Minimal genome histograms were drawn, focusing on genome size, gene count, GC content, and gene size (**Fig. 5**).

## The Tenericutes

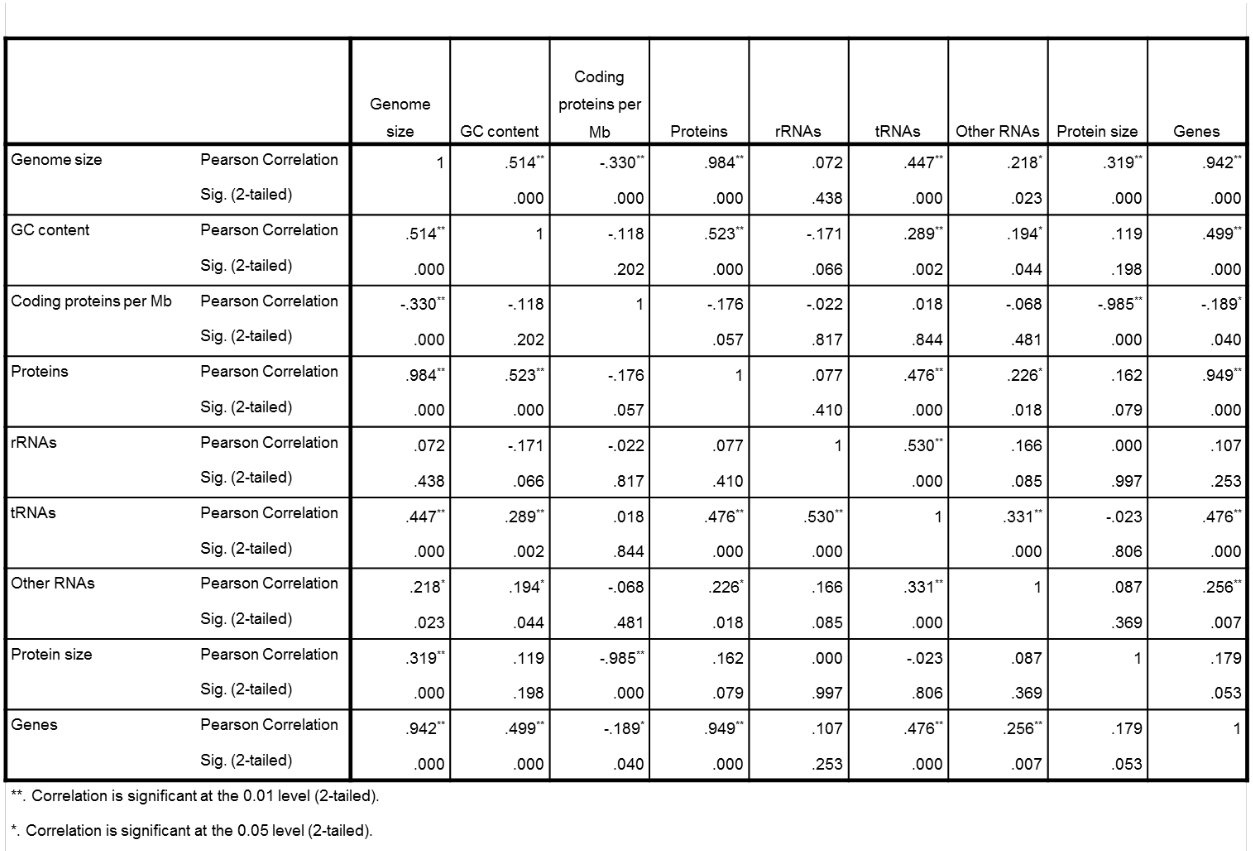
We studied most Tenericutes species (n=275) with 16S rRNA sequences available. Species discovered

| | N | Range | Minimum | Maximum | Mean | | Std. Deviation |
|---|---|---|---|---|---|---|---|
| | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic |
| Genome size (Mb) | 118 | 6.92 | .56 | 7.48 | 2.3218 | .13313 | 1.44613 |
| GC Content (%) | 118 | 47.60 | 25.70 | 73.30 | 45.4220 | 1.16876 | 12.69599 |
| Coding proteins per Mb | 118 | 496.60 | 570.25 | 1066.85 | 886.6746 | 8.00888 | 86.99870 |
| Genes | 118 | 6291.00 | 2.00 | 6293.00 | 2074.3559 | 107.14881 | 1163.93398 |
| Proteins | 118 | 5717.00 | 483.00 | 6200.00 | 2017.5424 | 105.42665 | 1145.22657 |
| rRNAs | 117 | 25.00 | .00 | 25.00 | 6.4444 | .41637 | 4.50372 |
| tRNAs | 118 | 88.00 | .00 | 88.00 | 43.8051 | 1.09965 | 11.94522 |
| Other RNAs | 109 | 5.00 | .00 | 5.00 | 2.0000 | .11363 | 1.18634 |

**Table 1**. **Descriptive statistics for minimal genomes**. Computed with IBM SPSS.

| Organism | Phyla | Accession | Size (Mb) | G+C (%) | Protein | rRNA | tRNA | Other RNA | Gene | Pseudo-gene |
|---|---|---|---|---|---|---|---|---|---|---|
| *Mycoplasma parvum str. Indiana* | Tenericute | NC_022575.1 | 0.56 | 27.0 | 526 | 0 | 0 | 0 | 568 | 0 |
| *Blattabacterium sp. (Blaberus giganteus)* | Bacteroidetes | NC_017924.1 | 0.63 | 25.7 | 569 | 3 | 33 | 1 | 610 | 4 |
| *Borreliella chilensis* | Spirochaetia | CP009910 | 0.90 | 28.5 | 805 | 5 | 31 | 1 | 846 | 4 |
| *Candidatus Xiphinematobacter sp. Idaho Grape* | Verrucomicrobia | NZ_CP012665.1 | 0.92 | 47.7 | 799 | 3 | 45 | 2 | 864 | 15 |
| *Tropheryma whipplei TW08/27* | Actinobacteria | NC_004551.1 | 0.93 | 46.3 | 833 | 3 | 51 | 1 | 889 | 1 |
| *Chloracidobacterium thermophilum B* | Acidobacteria | NC_016025.1 | 1.01 | 61.2 | 778 | - | 4 | - | 811 | 29 |
| *Chlamydia avium 10DC88* | Chlamydiae | NZ_CP006571.1 | 1.04 | 36.9 | 831 | 3 | 39 | 1 | 937 | 63 |
| *Uncultured Termite group 1 bacterium* | Elusimicrobia | NC_020419.1 | 1.13 | 35.2 | 761 | 3 | 45 | - | 809 | - |
| *Dialister pneumosintes* | Firmicutes | NZ_CP017037.1 | 1.27 | 35.2 | 1164 | 16 | 53 | 4 | 1251 | 14 |
| *Sneathia amnii* | Fusobacteria | NZ_CP011280.1 | 1.33 | 28.4 | 1234 | 10 | 36 | 1 | 1296 | 15 |
| *Dehalococcoides mccartyi CG5* | Chloroflexi | NZ_CP006951.1 | 1.36 | 47.2 | 1392 | 3 | 46 | 1 | 1449 | 6 |
| *Candidatus Atelocyanobacterium thalassa isolate ALOHA* | Cyanobacteria | NC_013771.1 | 1.44 | 31.1 | 1133 | 6 | 37 | 4 | 1217 | 37 |
| *Thermocrinis albus DSM 14484* | Aquificae | NC_013894.1 | 1.50 | 46.9 | 1567 | 3 | 43 | 1 | 1626 | 12 |
| *Caldisericum exile AZM16c01* | Caldiserica | NC_017096.1 | 1.56 | 35.4 | 1483 | 3 | 46 | 3 | 1550 | 15 |
| *Thermodesulfobacterium geofontis OPF15* | Thermodesulfobacteria | NC_015682.1 | 1.63 | 30.6 | 1608 | 3 | 46 | 2 | 1682 | 23 |
| *Thermotoga naphthophila RKU-10* | Thermotogae | NC_013642.1 | 1.81 | 46.1 | 1768 | 3 | 46 | 2 | 1856 | 37 |
| *Thermus thermophilus HB8* | Deinococcus-Thermus | NC_006461.1 | 1.85 | 69.5 | 1908 | 6 | 47 |  | 1980 | 19 |
| *Thermanaerovibrio acidaminovorans DSM 6589* | Synergistetes | NC_013522.1 | 1.85 | 63.8 | 1730 | 9 | 50 | 4 | 1821 | 28 |
| *Dictyoglomus turgidum DSM 6724* | Dictyoglomi | NC_011661.1 | 1.86 | 34.0 | 1742 | 6 | 46 | 1 | 1865 | 70 |
| *Chlorobium phaeovibrioides DSM 265* | Chlorobi | NC_009337.1 | 1.97 | 53.0 | 1765 | 3 | 45 | 3 | 1830 | 14 |
| *Calditerrivibrio nitroreducens DSM 19672* | Deferribacteres | NC_014758.1 | 2.16 | 35.8 | 2030 | 6 | 40 | 2 | 2092 | 14 |
| *Leptospirillum ferriphilum ML-04* | Nitrospirae | NC_018649.1 | 2.41 | 54.6 | 2361 | 6 | 49 | 1 | 2462 | 45 |
| *Desulfurispirillum indicum S5* | Chrysiogenetes | NC_014836.1 | 2.93 | 56.1 | 2616 | 9 | 37 | 2 | 2705 | 41 |
| *Chthonomonas calidirosea T49* | Armatimonadetes | NC_021487.1 | 3.44 | 54.6 | 2807 | 4 | 46 | 1 | 2908 | 50 |
| *Phycisphaera mikurensis NBRC 102666* | Planctomycetes | NC_017080.1 | 3.80 | 73.3 | 2955 | 3 | 46 | 2 | 3063 | 57 |
| *Fibrobacter succinogenes subsp. succinogenes S85* | Fibrobacteres | NC_017448.1 | 3.84 | 48.0 | 3078 | 9 | 58 | 1 | 3159 | 13 |
| *Gemmatirosa kalamazoonesis KBS708* | Gemmatimonadetes | CP007128 | 5.31 | 72.6 | 4513 | 6 | 48 | - | 4567 | - |
| *Sphingobacterium sp. ML3W* | Sphingobacteria | NZ_CP009278.1 | 5.33 | 37.0 | 4359 | 25 | 88 | 1 | 4530 | 57 |

**Table 2. Most minimal sequenced genome per phylum.** Data sorted in Microsoft Excel.

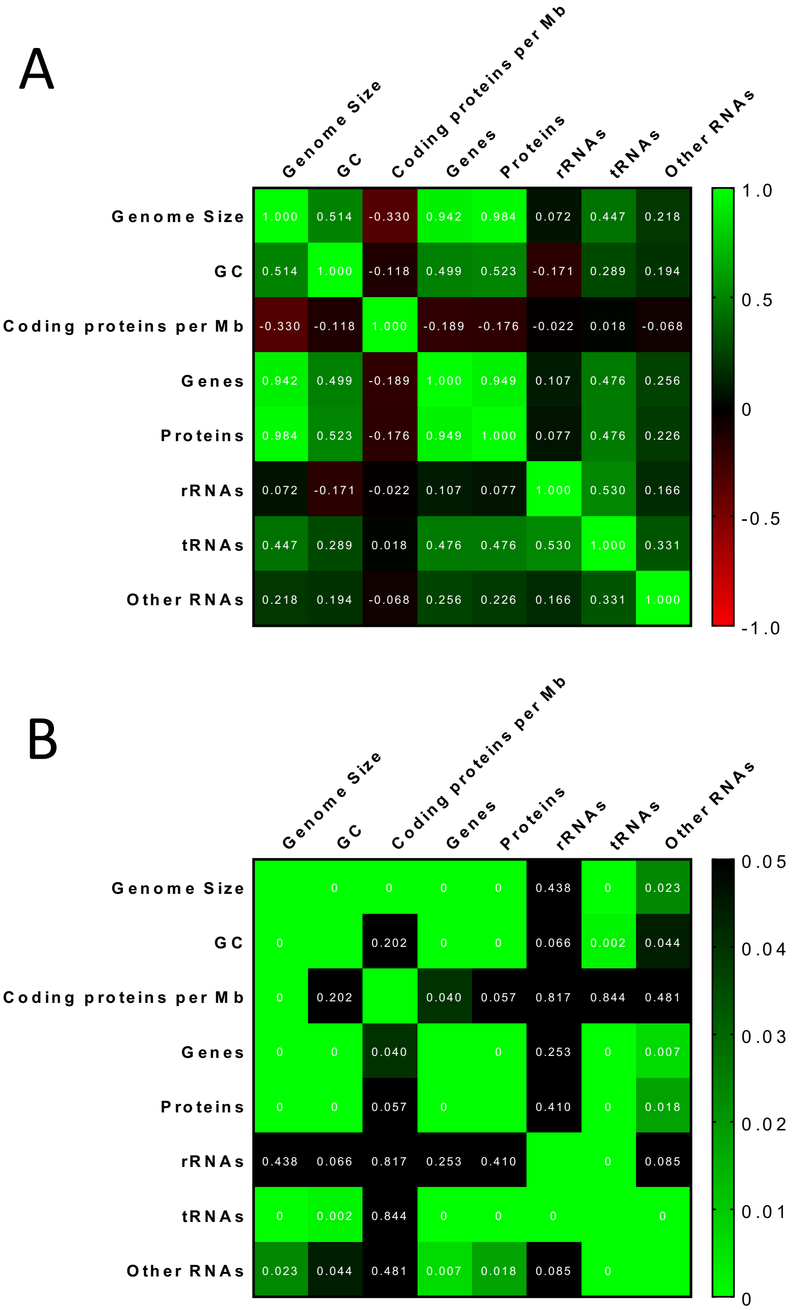


|  |  | Genome size | GC content | Coding proteins per Mb | Proteins | rRNAs | tRNAs | Other RNAs | Protein size | Genes |
|---|---|---|---|---|---|---|---|---|---|---|
| Genome size | Pearson Correlation | 1 | .514** | -.330** | .984** | .072 | .447** | .218* | .319** | .942** |
|  | Sig. (2-tailed) |  | .000 | .000 | .000 | .438 | .000 | .023 | .000 | .000 |
| GC content | Pearson Correlation | .514** | 1 | -.118 | .523** | -.171 | .289** | .194* | .119 | .499** |
|  | Sig. (2-tailed) | .000 |  | .202 | .000 | .066 | .002 | .044 | .198 | .000 |
| Coding proteins per Mb | Pearson Correlation | -.330** | -.118 | 1 | -.176 | -.022 | .018 | -.068 | -.985** | -.189* |
|  | Sig. (2-tailed) | .000 | .202 |  | .057 | .817 | .844 | .481 | .000 | .040 |
| Proteins | Pearson Correlation | .984** | .523** | -.176 | 1 | .077 | .476** | .226* | .162 | .949** |
|  | Sig. (2-tailed) | .000 | .000 | .057 |  | .410 | .000 | .018 | .079 | .000 |
| rRNAs | Pearson Correlation | .072 | -.171 | -.022 | .077 | 1 | .530** | .166 | .000 | .107 |
|  | Sig. (2-tailed) | .438 | .066 | .817 | .410 |  | .000 | .085 | .997 | .253 |
| tRNAs | Pearson Correlation | .447** | .289** | .018 | .476** | .530** | 1 | .331** | -.023 | .476** |
|  | Sig. (2-tailed) | .000 | .002 | .844 | .000 | .000 |  | .000 | .806 | .000 |
| Other RNAs | Pearson Correlation | .218* | .194* | -.068 | .226* | .166 | .331** | 1 | .087 | .256** |
|  | Sig. (2-tailed) | .023 | .044 | .481 | .018 | .085 | .000 |  | .369 | .007 |
| Protein size | Pearson Correlation | .319** | .119 | -.985** | .162 | .000 | -.023 | .087 | 1 | .179 |
|  | Sig. (2-tailed) | .000 | .198 | .000 | .079 | .997 | .806 | .369 |  | .053 |
| Genes | Pearson Correlation | .942** | .499** | -.189* | .949** | .107 | .476** | .256** | .179 | 1 |
|  | Sig. (2-tailed) | .000 | .000 | .040 | .000 | .253 | .000 | .007 | .053 |  |

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

**Table 3. Pearson correlations in minimal genomes.**



|  | Genome Size | GC | Coding proteins per Mb | Genes | Proteins | rRNAs | tRNAs | Other RNAs |
|---|---|---|---|---|---|---|---|---|
| Genome Size | 1.000 | 0.514 | -0.330 | 0.942 | 0.984 | 0.072 | 0.447 | 0.218 |
| GC | 0.514 | 1.000 | -0.118 | 0.499 | 0.523 | -0.171 | 0.289 | 0.194 |
| Coding proteins per Mb | -0.330 | -0.118 | 1.000 | -0.189 | -0.176 | -0.022 | 0.018 | -0.068 |
| Genes | 0.942 | 0.499 | -0.189 | 1.000 | 0.949 | 0.107 | 0.476 | 0.256 |
| Proteins | 0.984 | 0.523 | -0.176 | 0.949 | 1.000 | 0.077 | 0.476 | 0.226 |
| rRNAs | 0.072 | -0.171 | -0.022 | 0.107 | 0.077 | 1.000 | 0.530 | 0.166 |
| tRNAs | 0.447 | 0.289 | 0.018 | 0.476 | 0.476 | 0.530 | 1.000 | 0.331 |
| Other RNAs | 0.218 | 0.194 | -0.068 | 0.256 | 0.226 | 0.166 | 0.331 | 1.000 |

|  | Genome Size | GC | Coding proteins per Mb | Genes | Proteins | rRNAs | tRNAs | Other RNAs |
|---|---|---|---|---|---|---|---|---|
| Genome Size |  | 0 | 0 | 0 | 0 | 0.438 | 0 | 0.023 |
| GC | 0 |  | 0.202 | 0 | 0 | 0.066 | 0.002 | 0.044 |
| Coding proteins per Mb | 0 | 0.202 |  | 0.040 | 0.057 | 0.817 | 0.844 | 0.481 |
| Genes | 0 | 0 | 0.040 |  | 0 | 0.253 | 0 | 0.007 |
| Proteins | 0 | 0 | 0.057 | 0 |  | 0.410 | 0 | 0.018 |
| rRNAs | 0.438 | 0.066 | 0.817 | 0.253 | 0.410 |  | 0 | 0.085 |
| tRNAs | 0 | 0.002 | 0.844 | 0 | 0 | 0 |  | 0 |
| Other RNAs | 0.023 | 0.044 | 0.481 | 0.007 | 0.018 | 0.085 | 0 |  |

**Fig. 2**. **Pearson correlations between minimal genomes metrics**. (A) Positive correlations in green, and negative correlations in red. (B) The p-values are in green, darker shades indicating larger p-values. Genome size is significantly (α=0.05) correlated with all parameters (**Fig. 3A, 3B, 3E**), except rRNA counts. Similarly, GC content is significantly correlated with all parameters (**Fig. 3C-3E**), except coding proteins per Mb and rRNA counts. rRNA counts were only significantly correlated with tRNA counts. Values obtained from IBM SPSS.

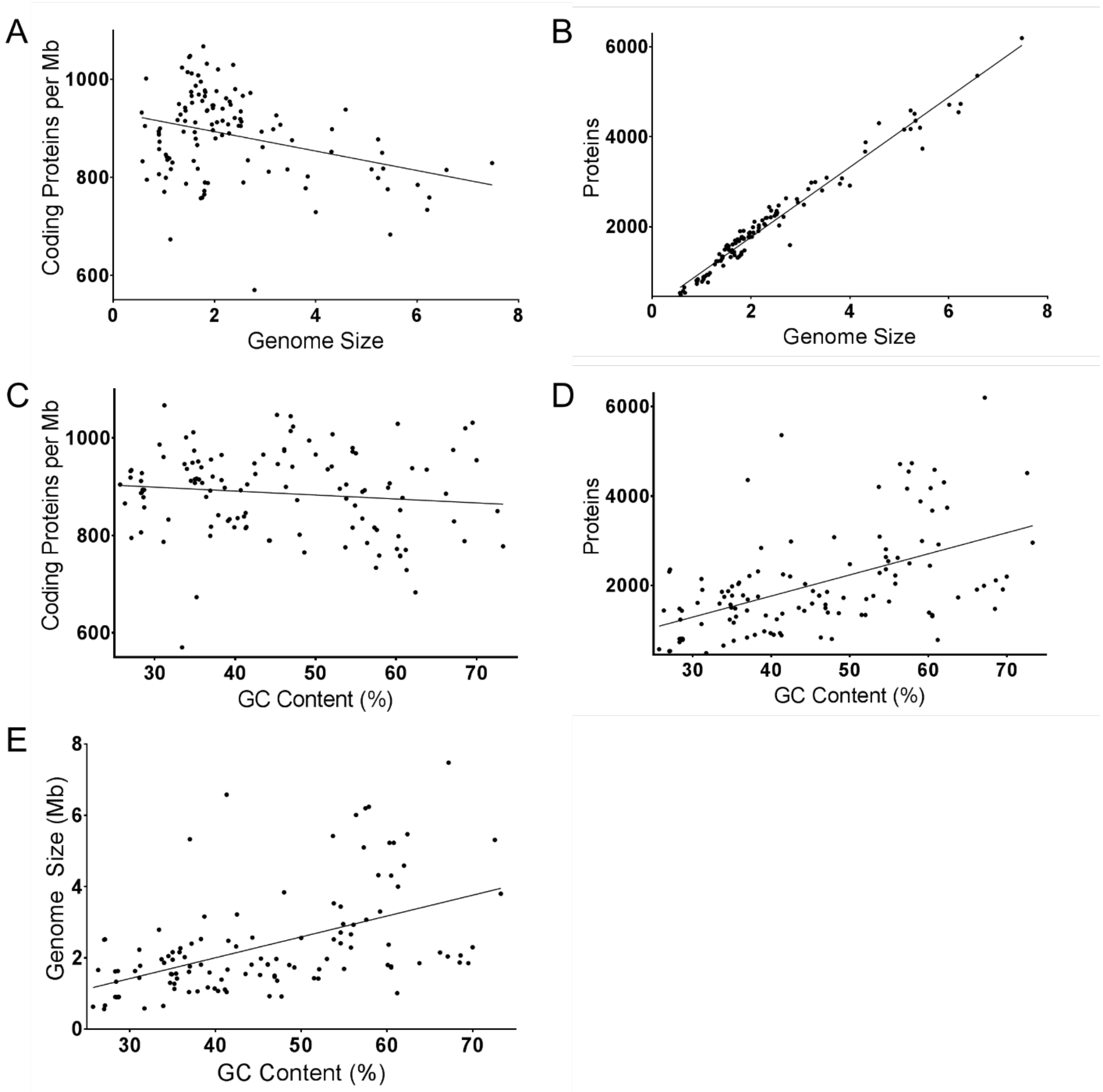


**Fig. 3**. **Plots of statistical correlations with linear regression lines**. (A) Smaller bacterial genomes have more coding proteins per Mb ("DNA fragment") than larger bacterial genomes (p=0.0003, $R^2$=0.1087, y=-19.83x+932.7). (B) The larger the genome, the more proteins. (C) A higher GC content does not mean less coding proteins per Mb (p=0.2023, $R^2$=0.01398, y=-0.8101x+923.5). (D) Nevertheless, GC content and protein count are statistically correlated (p<<0.0001, $R^2$=0.274, y=47.22x-127.1). (E) Smaller genomes have an AT bias (p<0.0001, $R^2$=0.2637, y=0.05849x-0.3351). Plots drawn with GraphPad Prism. Detailed correlation statistics are on **Fig. 2B**.

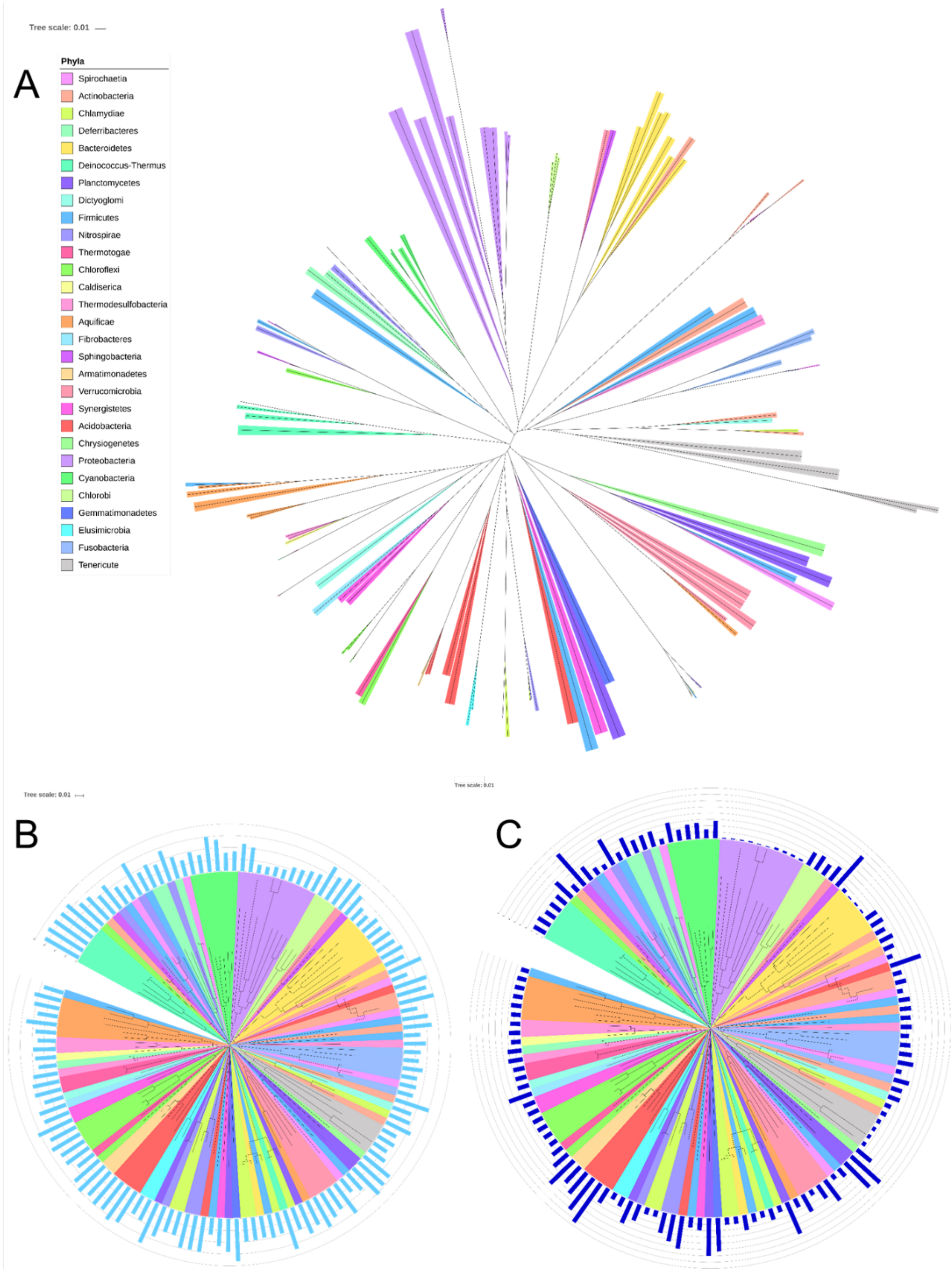


**Fig. 4. Trees of minimal bacterial genomes based on 16S rRNA alignments**. Input sequences were at least 1275 bp long, and the average was 1450 bp. (A) Unrooted phylogram. Phyla are together in cohesive clades. (B) GC contents are diverse in the minimal bacterial genomes from each phyla. Grey lines at every 20% increment. (C) Minimal genomes genome size. Grey lines at every Mb. Alignments done in EMBL using MAFFT. Trees drawn using iTOL.

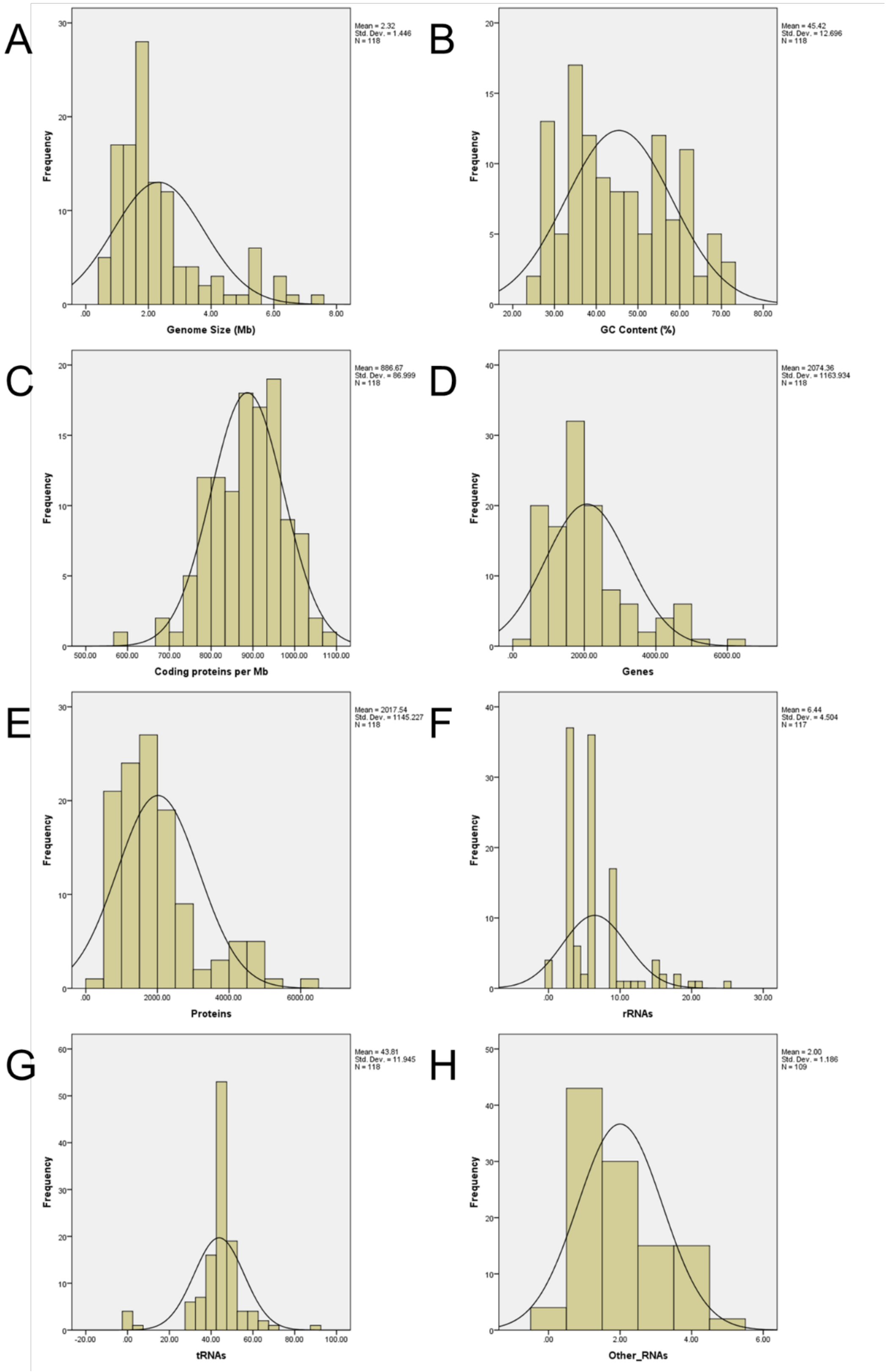


**Fig. 5**. **Histogram plots for minimal genome data**. Normal curve drawn on top of histogram. Histograms and normal curves plotted using IBM SPSS.

after 2013 were not included. We also excluded 16S sequences shorter than 1.1kb. We drew trees to illustrate in which continent the species was discovered, host classification, genus, and order (**Fig. 6**), localization of species within host (**Fig. 7A**), GC content (**Fig. 7B**), genome size (**Fig. 7C**).

Categorical data was illustrated in pie charts (**Fig. 8**), and compared to quantitative data (**Fig. 9**). Half (49%) of the discovered Tenericutes species were isolated from mammals, followed by plants (18%). Tenericutes are rarely found on environmental samples. A histogram of the genome size and GC content was also performed (**Fig. 10A, 10B**). These were compared to each other (**Fig. 10C**). Species discovery followed an exponential trend, reaching plateau this decade (**Fig. 10D**). We observed clades within the Tenericutes tree not consistent with its naming, and therefore drew an unrooted tree (**Fig. 11**).

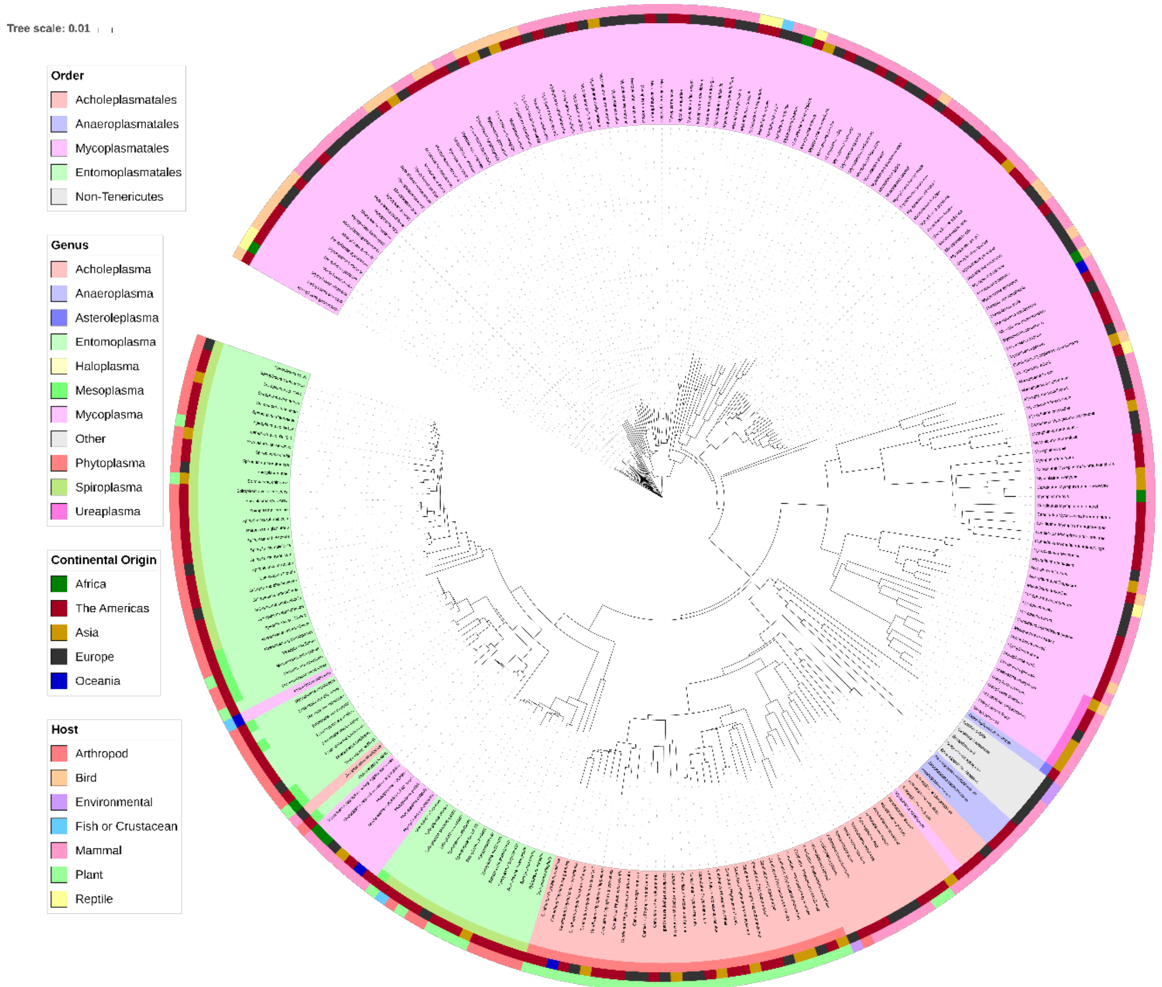


**Fig. 6**. **The Tenericutes order, genus, continental discovery place, and hosts**. Tree represents 16S rRNA alignment using EMBL's MAFFT. Innermost ring contains the phylogenetic trees with representative branch distances in circular form. The next ring contains the species name, and its shade is the organism order. Mycoplasmatales are paraphyletic. The next ring is the genus, which further suggests no cohesive nomenclature. The second to last ring is the continent it was isolated from, and the last one is the host. Tree drawn in iTOL.

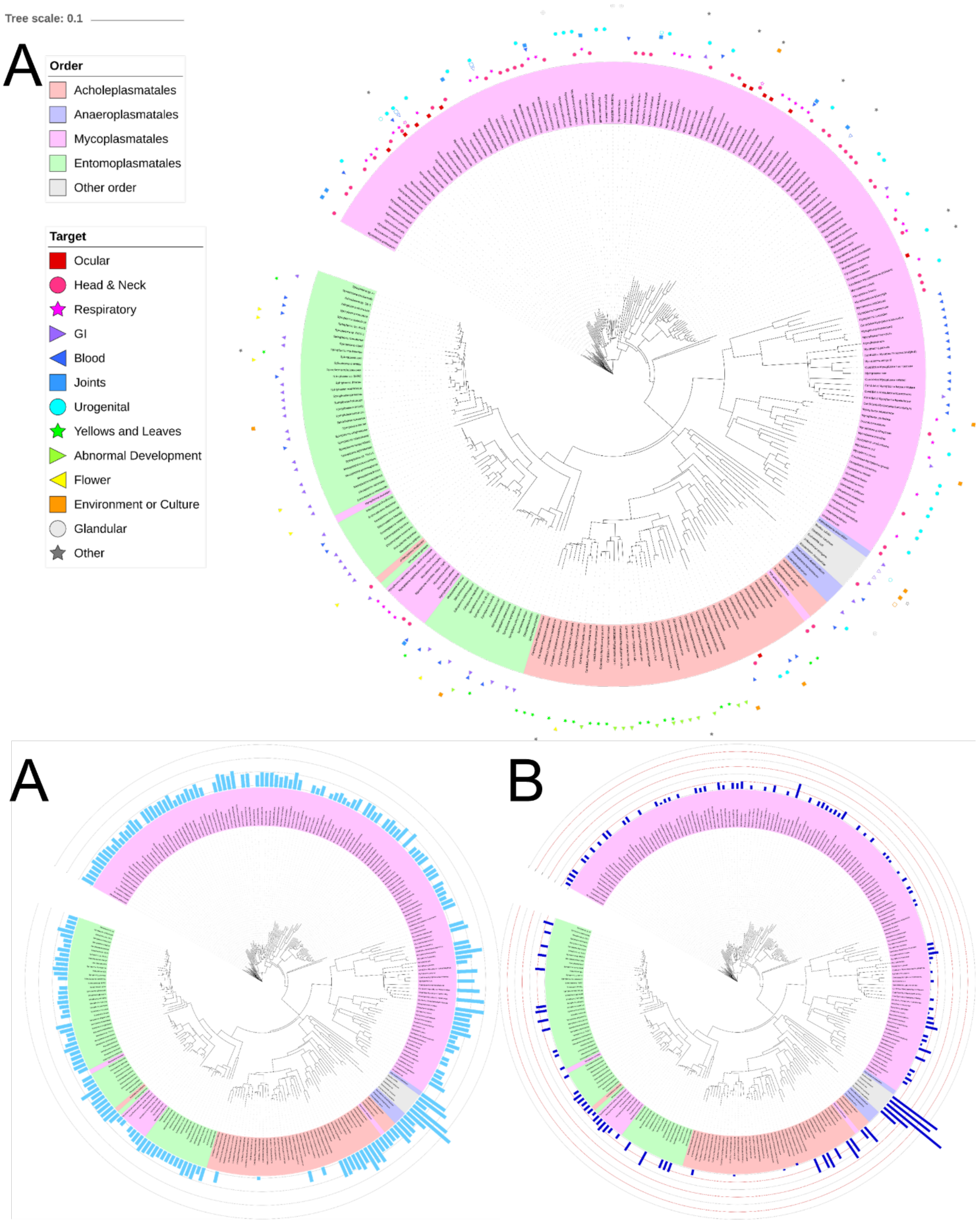


**Fig. 7. Phylogenetic trees of the Tenericutes**. Based on 16S rRNA data. (A) Localization of Tenericutes within hosts. GI refers to gastrointestinal system.(B) GC content in Tenericutes. Zeroed at 19%. Grey lines are increments of 20%. (C) Tenericutes genome size within phylogram. Zeroed at 0.5Mb. Grey lines at 0.5Mb and red lines every 1Mb. Alignments done in EMBL MAFFT. Trees drawn using iTOL.

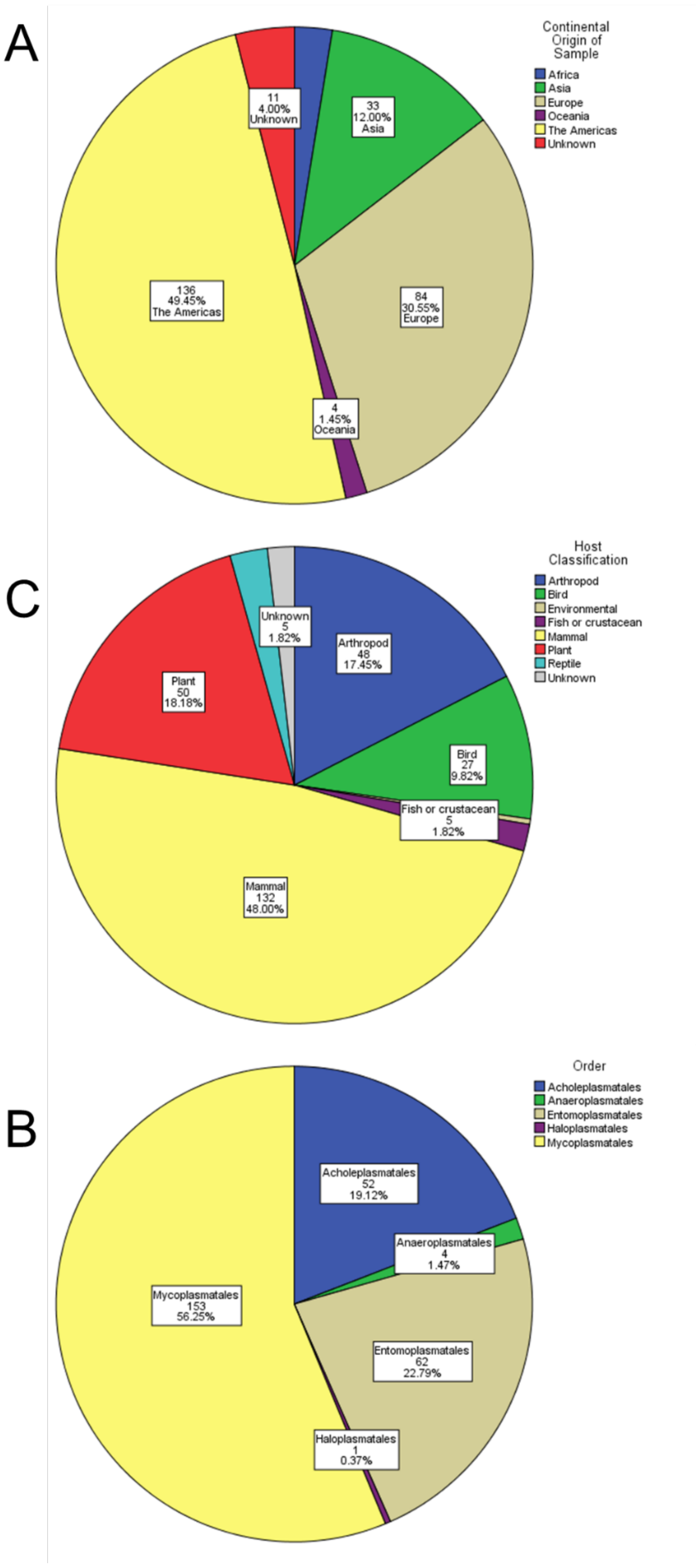


**Fig. 8**. **Pie charts of qualitative Tenericutes data**. (A) Continental origin, (B) host, and (C) order. Pie charts computed using IBM SPSS.

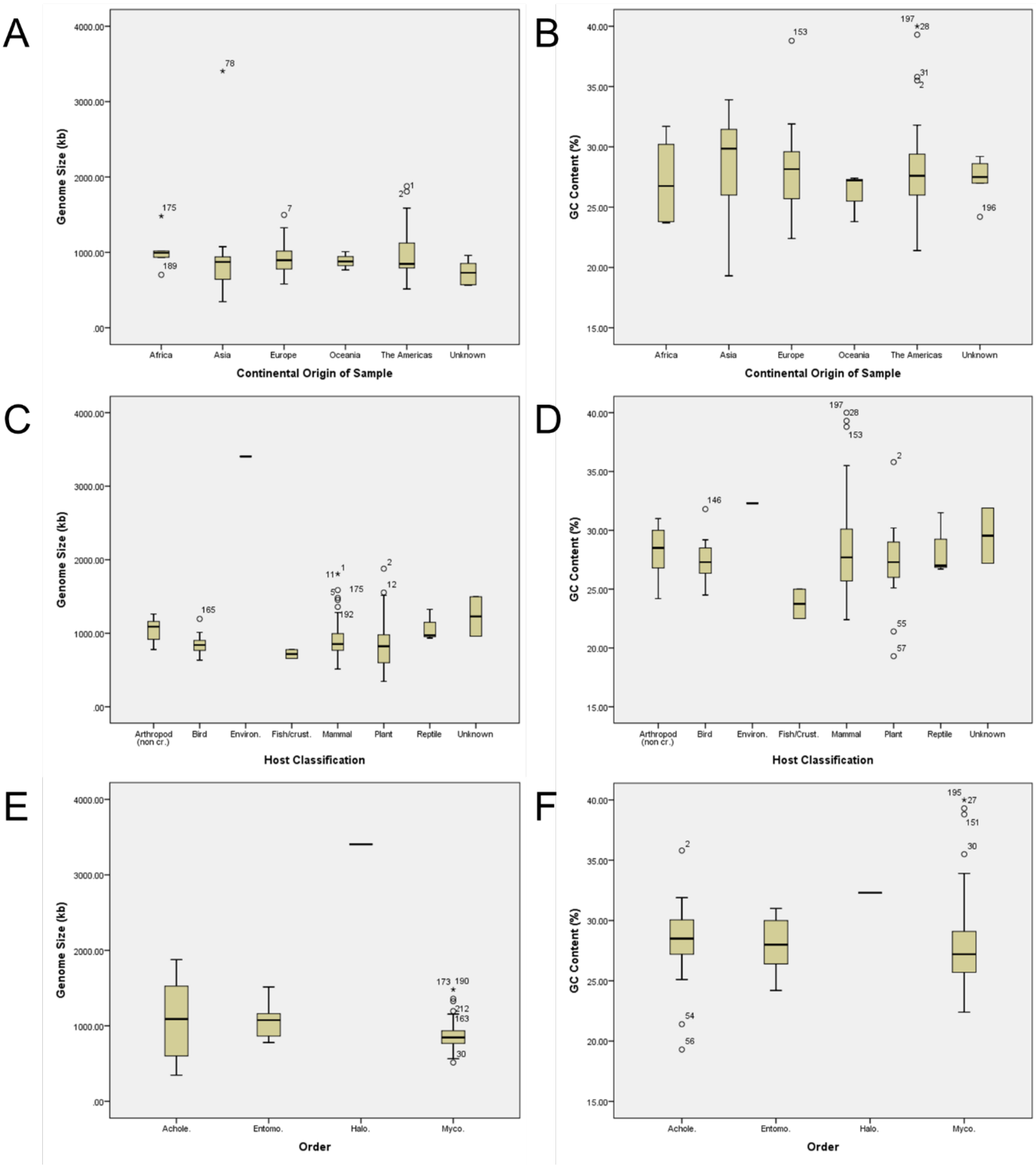


**Fig. 9. Box plots of Tenericutes data**. X-axis displays in (A)(B) Continental origin of samples, (C)(D) host classification, and (E)(F) order. Y-axis displays in (A)(C)(E) genome size and in (B)(D)(F) GC content. (A) African Tenericutes have less genome size variance, and Asian Tenericutes are generally smaller in genome size. (B) African Tenericutes have less GC contents, while Asian Tenericutes have higher GC contents. (D) Fish/crustacean Tenericutes have lower GC contents, while arthropods have the highest. (E) Mycoplasmatales have genome sizes with the least variance while Acholeplasmatales have the most variance. (F) Mycoplasmatales generally have less GC content, while Acholeplasmatales have the most. Box plots drawn using IBM SPSS.

| | N | Range | Minimum | Maximum | Mean | | Std. Deviation | Variance |
|---|---|---|---|---|---|---|---|---|
| | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Statistic |
| GC Content (%) | 207 | 20.70 | 19.30 | 40.00 | 28.1478 | .22429 | 3.22702 | 10.414 |
| Genome Size (kb) | 128 | 3058.54 | 345.97 | 3404.50 | 958.0176 | 30.46574 | 344.68053 | 118804.664 |
| Valid N (listwise) | 122 | | | | | | | |

**Table 4**. **Descriptive statistics of the Tenericutes**. GC content and genome size are correlated. Pearson correlation of .191, p=.037, significant at the 0.05 level (2-tailed). Statistics run on IBM SPSS.

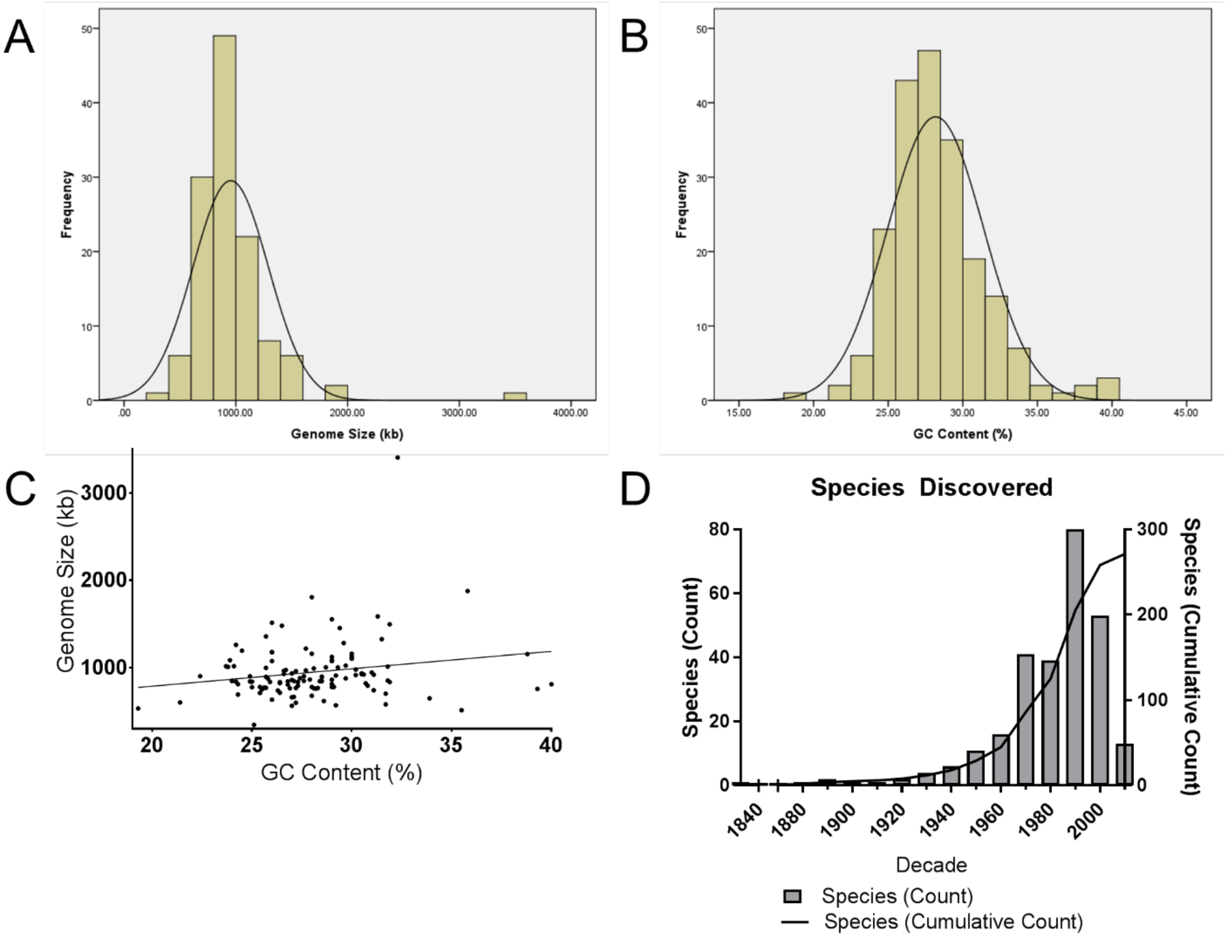


**Fig. 10**. **Quantitative Tenericutes data**. (A)(B) display histograms with normal curves on top. Normal distribution observed in (A) genome size and (B) GC content. (C) GC content compared to genome size in Tenericutes. Smaller genomes are AT biased. (D) Decade Tenericutes species were discovered. Despite the advances in genomics, discoveries have suffered a hiatus since the 2000s. (A)(B) computed in IBM SPSS, and (C)(D) in GrapPad Prism.

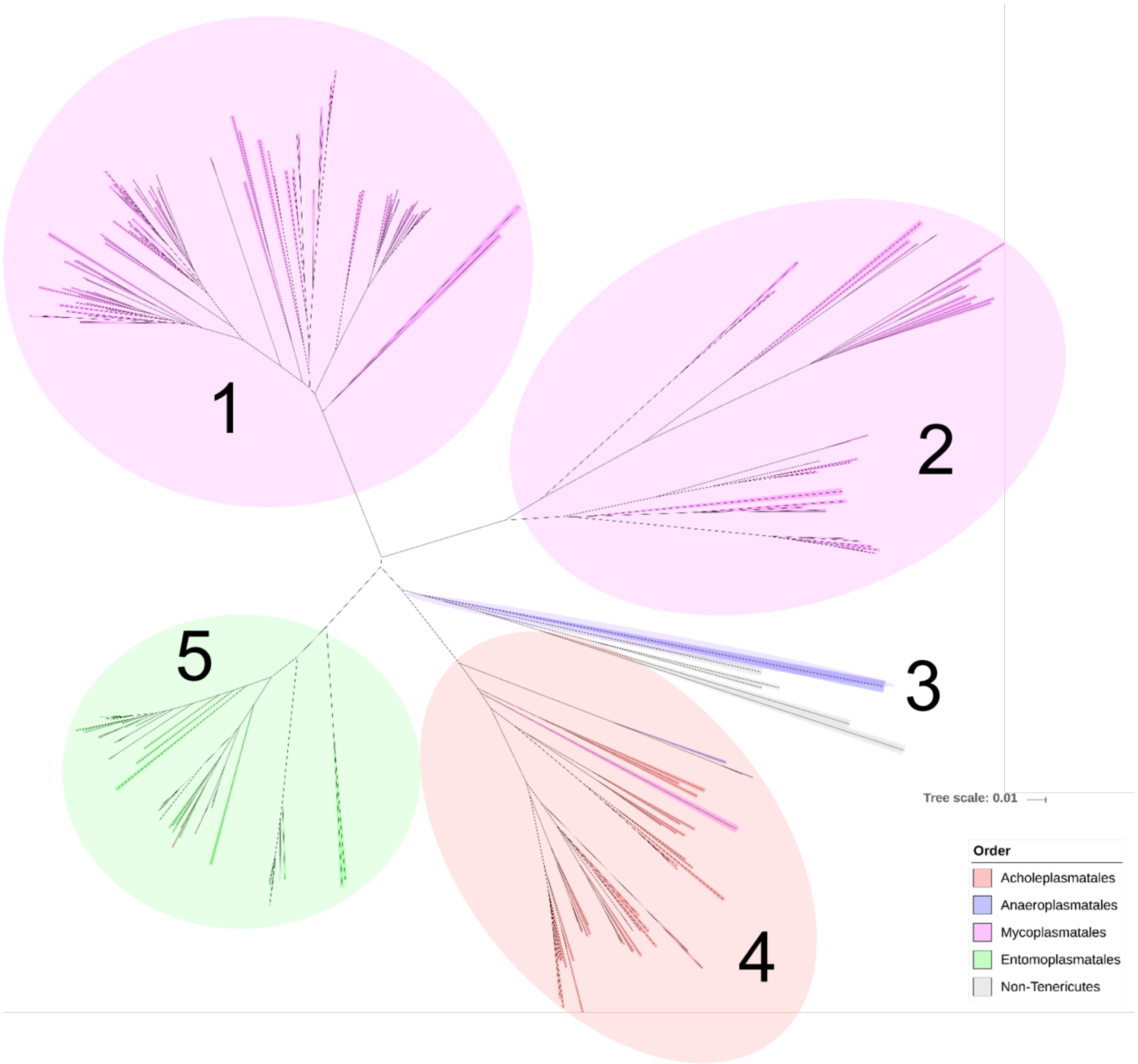


**Fig. 11. Unrooted Tenericutes Tree with clades circled**. Five clades were observed and numbered from 1-5. Clade 2 has *Mycoplasma* spp. that infect the blood. Clade 3 contains extremophiles. Clade 4 contains Tenericutes that use the universal codon table. Clade 5 contains the type species, *Mycoplasma mycoides*. Based on phylogenetics, the Tenericutes have nomenclatures that not reflective its evolutionary history. Instead, nomenclature is arbitrary, especially for the Mycoplasmatales. Tree reflects 16S rRNA alignment done in EMBL MAFFT. Tree drawn in iTOL.

Whole-genome alignments for human-pathogenic Tenericutes revealed relative close-relatedness (**Fig. 12A**). However, whole-genome alignments for mammalian Tenericutes displayed extensive horizontal gene transfer events (**Fig. 12B**). Proteome alignments place certain Tenericutes distant from the rest, notably *Acholepasma*. This is consistent with 16S data.

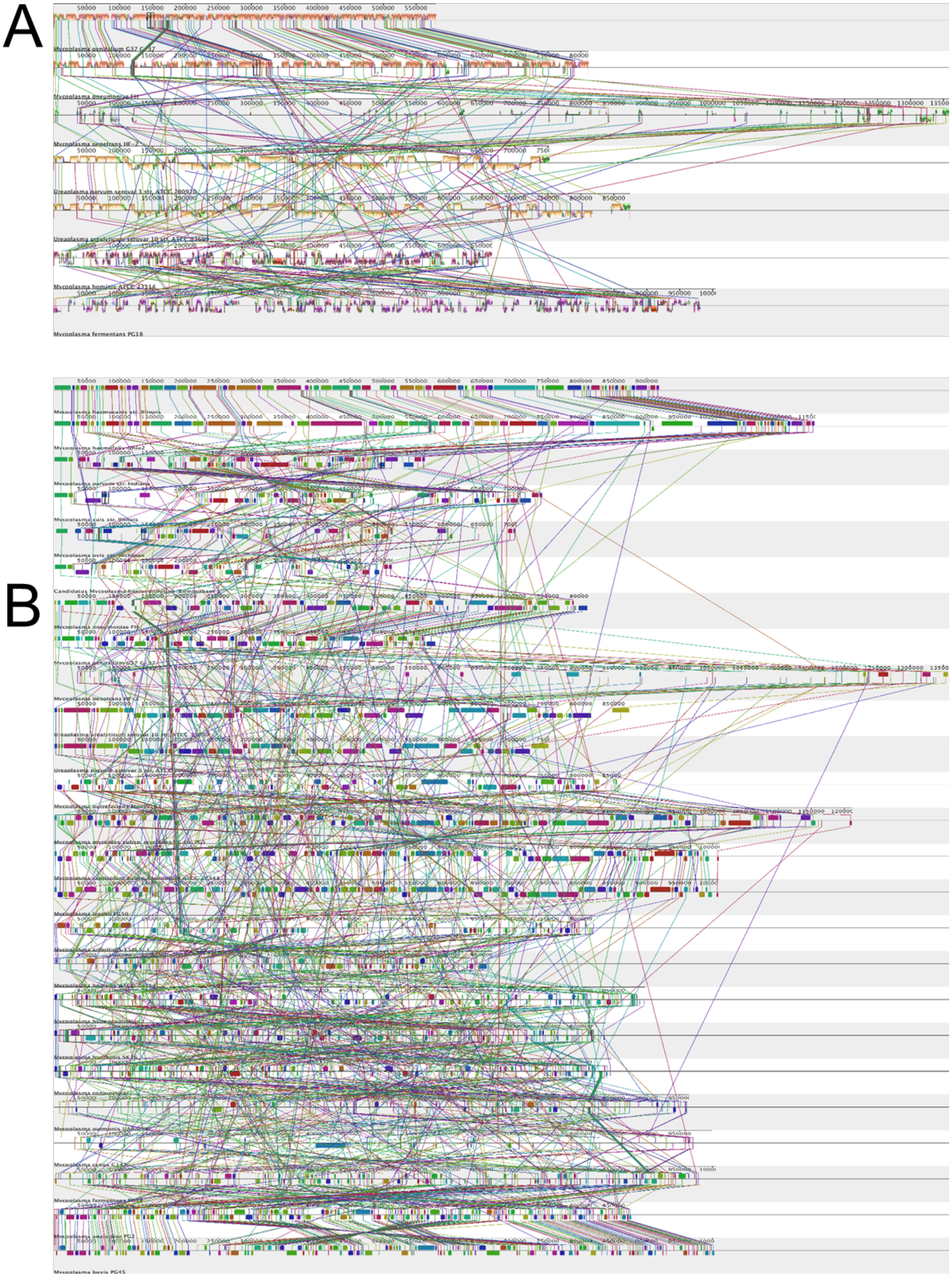

A
Mycoplasma genitalium G37 G-37
Mycoplasma pneumoniae FH
Mycoplasma penetrans HF-2
Ureaplasma parvum serovar 3 str. ATCC 700970
Ureaplasma urealyticum serovar 10 str. ATCC 33699
Mycoplasma hominis ATCC 23114
Mycoplasma fermentans PG18
B
Mycoplasma haemocanis str. Illinois
Mycoplasma bovis PG45

**Fig. 12. Mauve proteome alignments**. Either the Mycoplasmatales are distantly related or their nomenclature is not representative of its genetic relationships. (A) Human-pathogenic Mycoplasmatales proteome alignment. (B) Mammalian Mycoplasmatales proteome alignments. Mauve is a software from the Darling Lab.

Codon usage was analyzed to compare AT rich Tenericutes with UGA codons transcribing for tryptophan against other organisms and mitochondria (**Figures 13-15**). Yeast mitochondria have AUA transcribing for methionine (Met, M); CUU, CUC, CUA, CUG are threonine (Thr, T); CGA, CGC are absent. Vertebrate mitochondria have AGA and AGG transcribing for a stop codon (Ter, *), AUA codes for methionine (Met, M), and UGA codes for tryptophan (Trp, W). In human (*Homo sapiens*) mitochondria, AUA and AUU are alternative initiation codons. In mice mitochondria (*Mus musculus*), AUA, AUU, and AUC are alternative initiation codons.

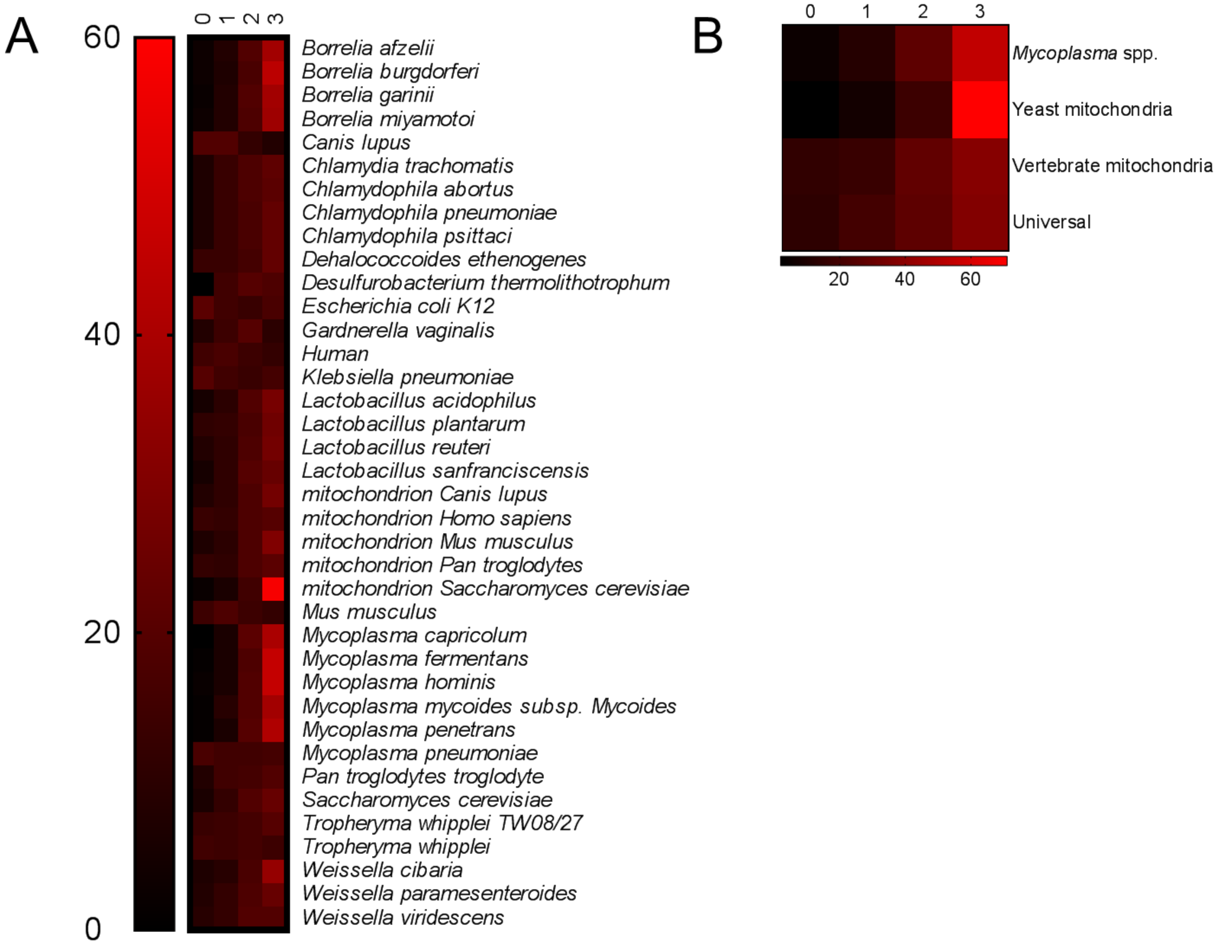


**Fig. 13**. **Heatmap of the number of AT bases in codons**. Horizontal numbers 1-3 indicate number of AT bases in each codons. Red shades indicates more codons containing that number of AT. A uniform shade like in human indicates equal representation of AT and CG in protein codons. *Mycoplasma* spp. have more red in column 3, indicating that most expressed proteins use 3 AT bases in each codon (3bp) each. *Borrelia* spp. display a similar AT-rich pattern. (A) AT content by species. (B) Genetic code by AT content heatmap. Red indicates percentage of codons, x-axis indicates genetic code and y-axis indicates number of bases that are AT. Heatmap drawn using GrapPad Prism.

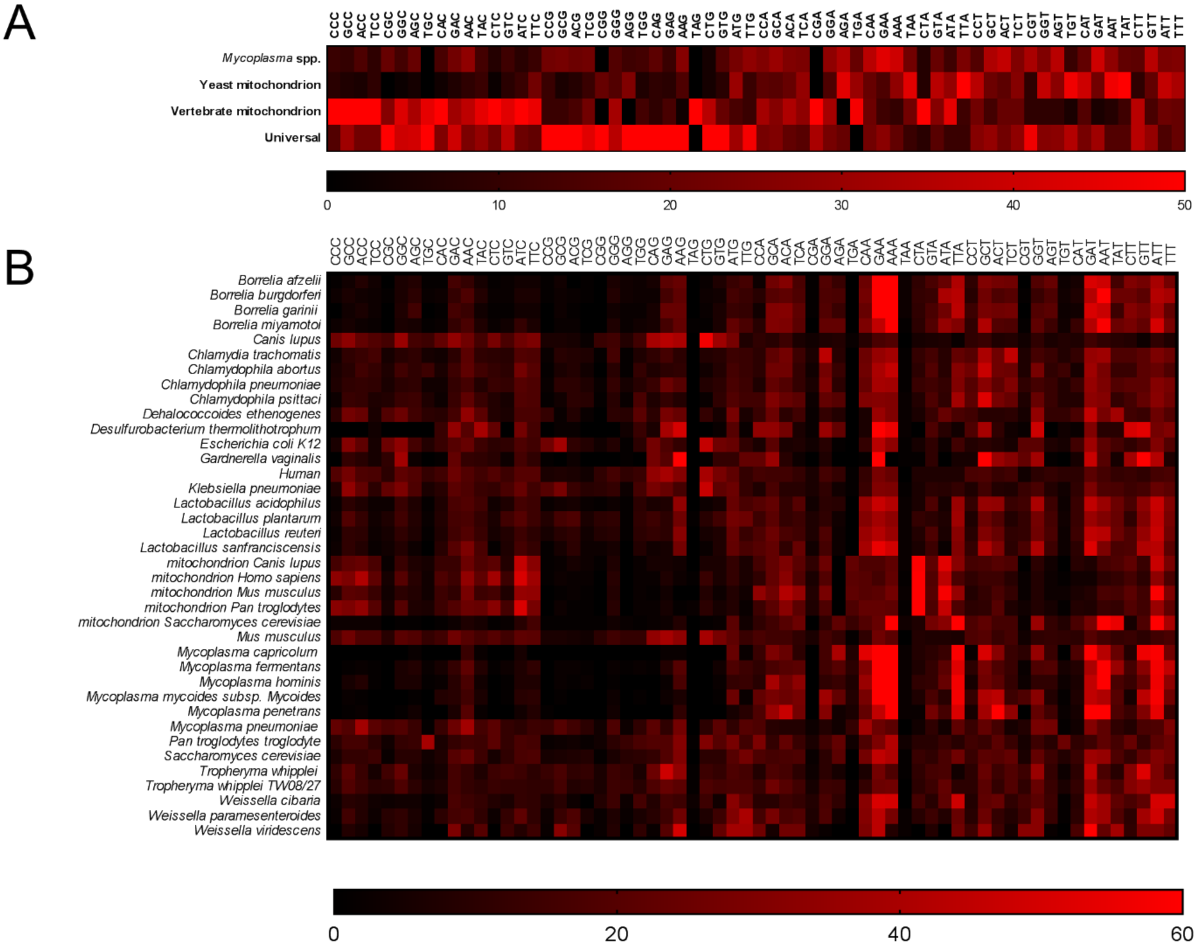


**Fig. 14**. **Codon use heatmap**. Codons sorted by hydrogen bonding, nucleobases appearing in the following order: cytosine (C), guanine (G), adenine (A) and thymine (T). Codons sorted from right to left, the first sorterer being the third (i.e., wobble, 3' end of the mRNA) position of the codon, then the middle or second position, and lastly, the first and leftmost codon position (5' end of the mRNA). Whereas darker shades indicate that the codon is least frequently used, redder shades indicate higher frequencies of that codon. Legend indicates codons used per thousand coding basepairs. (A) Organisms using the universal codon table are somewhat GC biased. Though mitochondria and some Teneritutes share UGA coding for tryptophan, there is no AT bias for codons in vertebrate mitochondria. However, yeast mitochondria are AT biased. To enhance contrast, probabilities greater or equal to 50 codons per thousand basepairs were pure red. (B) *Borrelia* spp. and *Mycoplasma* spp. are AT biased at the wobble position, thus from the left to the middle of the chart, most codons near black, while redder shades extend from the middle to the right. For (B), codons greater or equal to 60 codons per thousand basepairs are indicated in red. The threshold was applied to enhance visibility. The heatmap was drawn using GrapPad Prism.

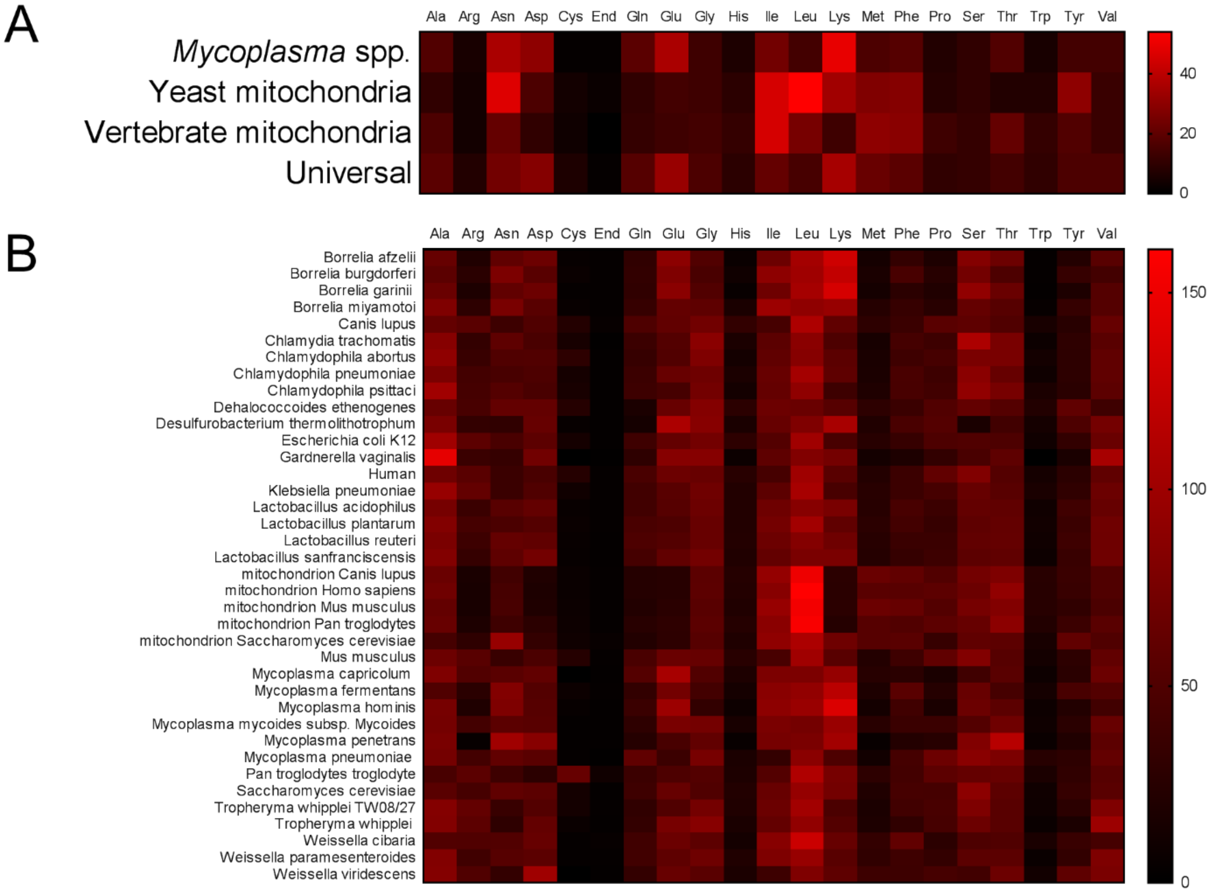


**Fig. 15**. **Heatmap of amino-acid use based on codons**. Amino-acid frequency per thousand amino-acids. Darker shades indicate rarer amino-acids. (A) Even though *Mycoplasma* spp. and mitochondria have an additional codon for tryptophan (Trp), UGA, there is no drastic increase in tryptophan. Groups by genetic codon tables. (B) Organisms sharing common genetic codon tables have similar profiles. Heatmaps drawn in GraphPad Prism.

### Other Comparative Studies on the 16S rRNA and Proteomes

From a bioinformatics perspective, there are nearly endless ways DNA sequences may be analyzed. We compared alignment methods (i.e., algorithms) such as MAFFT, MUSCLE and T-Coffee offered through different servers like EMBL for tree rendering purposes (**Fig. 16**). It appears that phylogenetic trees, or the evolutionary narrative to an extent, vary from to alignment method to alignment method. The 16S rRNA narrative of T-Coffee is most similar with the proteome data (**Figures 16, 17**). We also drew an unrooted tree to include all of our minimal genomes, labeled as Bacteria or Tenericutes, representative species from the Archaea and Eukarya domains (**Fig. 18**).

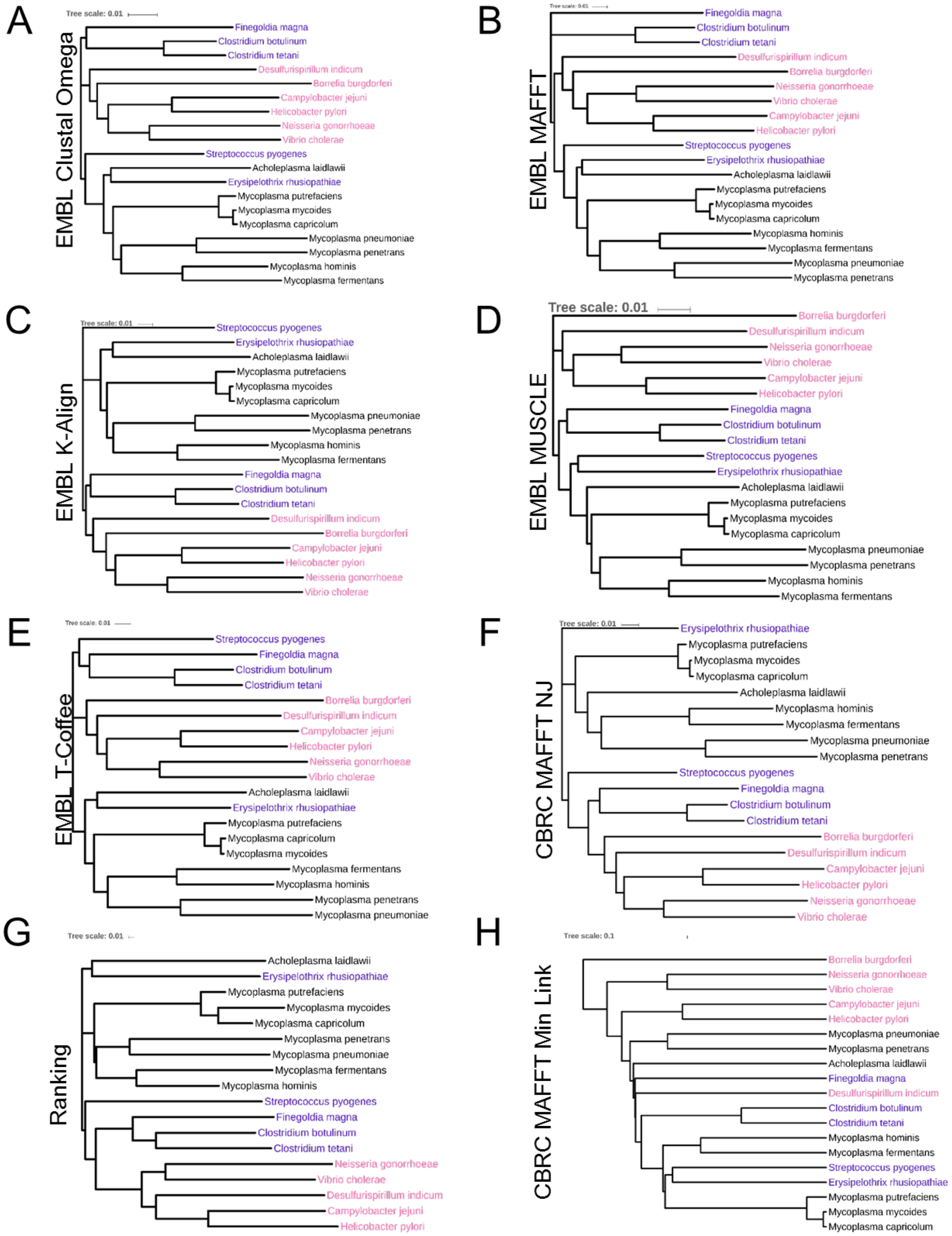
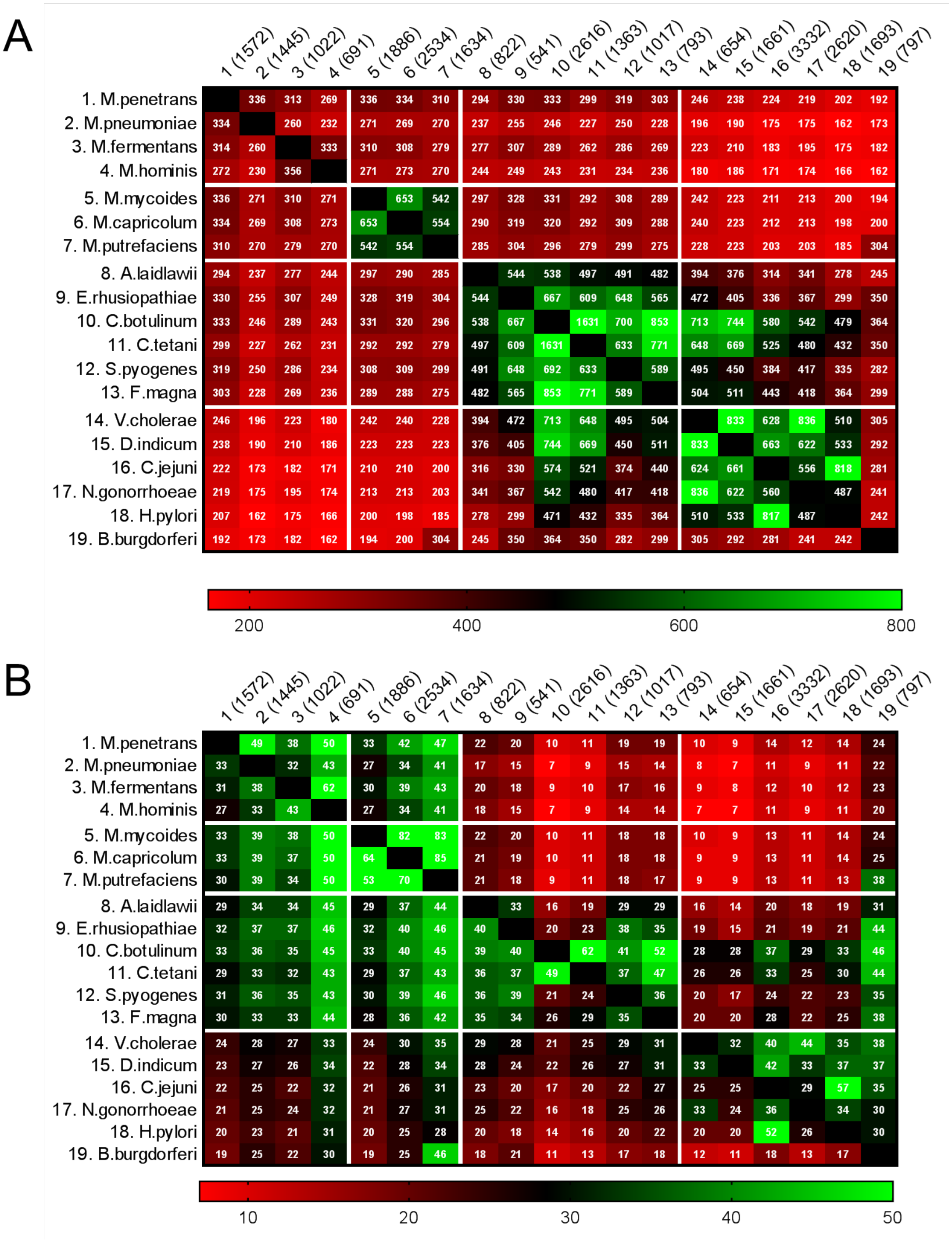


**Fig. 16**. **Multiple sequence alignment comparisons**. Different phylogenies drawn using different algorithms (e.g., MAFFT, MUSCLE), tree methods (e.g., neighbor-joining), and service providers (e.g., EMBL, CBRC). Alignments run as indicated. Trees drawn with iTOL.



| | 1 (1572) | 2 (1445) | 3 (1022) | 4 (691) | 5 (1886) | 6 (2534) | 7 (1634) | 8 (822) | 9 (541) | 10 (2616) | 11 (1363) | 12 (1017) | 13 (793) | 14 (654) | 15 (1661) | 16 (3332) | 17 (2620) | 18 (1693) | 19 (797) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. M.penetrans | | 336 | 313 | 269 | 336 | 334 | 310 | 294 | 330 | 333 | 299 | 319 | 303 | 246 | 238 | 224 | 219 | 202 | 192 |
| 2. M.pneumoniae | 334 | | 260 | 232 | 271 | 269 | 270 | 237 | 255 | 246 | 227 | 250 | 228 | 196 | 190 | 175 | 175 | 162 | 173 |
| 3. M.fermentans | 314 | 260 | | 333 | 310 | 308 | 279 | 277 | 307 | 289 | 262 | 286 | 269 | 223 | 210 | 183 | 195 | 175 | 182 |
| 4. M.hominis | 272 | 230 | 356 | | 271 | 273 | 270 | 244 | 249 | 243 | 231 | 234 | 236 | 180 | 186 | 171 | 174 | 166 | 162 |
| 5. M.mycoides | 336 | 271 | 310 | 271 | | 653 | 542 | 297 | 328 | 331 | 292 | 308 | 289 | 242 | 223 | 211 | 213 | 200 | 194 |
| 6. M.capricolum | 334 | 269 | 308 | 273 | 653 | | 554 | 290 | 319 | 320 | 292 | 309 | 288 | 240 | 223 | 212 | 213 | 198 | 200 |
| 7. M.putrefaciens | 310 | 270 | 279 | 270 | 542 | 554 | | 285 | 304 | 296 | 279 | 299 | 275 | 228 | 223 | 203 | 203 | 185 | 304 |
| 8. A.laidlawii | 294 | 237 | 277 | 244 | 297 | 290 | 285 | | 544 | 538 | 497 | 491 | 482 | 394 | 376 | 314 | 341 | 278 | 245 |
| 9. E.rhusiopathiae | 330 | 255 | 307 | 249 | 328 | 319 | 304 | 544 | | 667 | 609 | 648 | 565 | 472 | 405 | 336 | 367 | 299 | 350 |
| 10. C.botulinum | 333 | 246 | 289 | 243 | 331 | 320 | 296 | 538 | 667 | | 1631 | 700 | 853 | 713 | 744 | 580 | 542 | 479 | 364 |
| 11. C.tetani | 299 | 227 | 262 | 231 | 292 | 292 | 279 | 497 | 609 | 1631 | | 633 | 771 | 648 | 669 | 525 | 480 | 432 | 350 |
| 12. S.pyogenes | 319 | 250 | 286 | 234 | 308 | 309 | 299 | 491 | 648 | 692 | 633 | | 589 | 495 | 450 | 384 | 417 | 335 | 282 |
| 13. F.magna | 303 | 228 | 269 | 236 | 289 | 288 | 275 | 482 | 565 | 853 | 771 | 589 | | 504 | 511 | 443 | 418 | 364 | 299 |
| 14. V.cholerae | 246 | 196 | 223 | 180 | 242 | 240 | 228 | 394 | 472 | 713 | 648 | 495 | 504 | | 833 | 628 | 836 | 510 | 305 |
| 15. D.indicum | 238 | 190 | 210 | 186 | 223 | 223 | 223 | 376 | 405 | 744 | 669 | 450 | 511 | 833 | | 663 | 622 | 533 | 292 |
| 16. C.jejuni | 222 | 173 | 182 | 171 | 210 | 210 | 200 | 316 | 330 | 574 | 521 | 374 | 440 | 624 | 661 | | 556 | 818 | 281 |
| 17. N.gonorrhoeae | 219 | 175 | 195 | 174 | 213 | 213 | 203 | 341 | 367 | 542 | 480 | 417 | 418 | 836 | 622 | 560 | | 487 | 241 |
| 18. H.pylori | 207 | 162 | 175 | 166 | 200 | 198 | 185 | 278 | 299 | 471 | 432 | 335 | 364 | 510 | 533 | 817 | 487 | | 242 |
| 19. B.burgdorferi | 192 | 173 | 182 | 162 | 194 | 200 | 304 | 245 | 350 | 364 | 350 | 282 | 299 | 305 | 292 | 281 | 241 | 242 | |

| | 1 (1572) | 2 (1445) | 3 (1022) | 4 (691) | 5 (1886) | 6 (2534) | 7 (1634) | 8 (822) | 9 (541) | 10 (2616) | 11 (1363) | 12 (1017) | 13 (793) | 14 (654) | 15 (1661) | 16 (3332) | 17 (2620) | 18 (1693) | 19 (797) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. M.penetrans | | 49 | 38 | 50 | 33 | 42 | 47 | 22 | 20 | 10 | 11 | 19 | 19 | 10 | 9 | 14 | 12 | 14 | 24 |
| 2. M.pneumoniae | 33 | | 32 | 43 | 27 | 34 | 41 | 17 | 15 | 7 | 9 | 15 | 14 | 8 | 7 | 11 | 9 | 11 | 22 |
| 3. M.fermentans | 31 | 38 | | 62 | 30 | 39 | 43 | 20 | 18 | 9 | 10 | 17 | 16 | 9 | 8 | 12 | 10 | 12 | 23 |
| 4. M.hominis | 27 | 33 | 43 | | 27 | 34 | 41 | 18 | 15 | 7 | 9 | 14 | 14 | 7 | 7 | 11 | 9 | 11 | 20 |
| 5. M.mycoides | 33 | 39 | 38 | 50 | | 82 | 83 | 22 | 20 | 10 | 11 | 18 | 18 | 10 | 9 | 13 | 11 | 14 | 24 |
| 6. M.capricolum | 33 | 39 | 37 | 50 | 64 | | 85 | 21 | 19 | 10 | 11 | 18 | 18 | 9 | 9 | 13 | 11 | 14 | 25 |
| 7. M.putrefaciens | 30 | 39 | 34 | 50 | 53 | 70 | | 21 | 18 | 9 | 11 | 18 | 17 | 9 | 9 | 13 | 11 | 13 | 38 |
| 8. A.laidlawii | 29 | 34 | 34 | 45 | 29 | 37 | 44 | | 33 | 16 | 19 | 29 | 29 | 16 | 14 | 20 | 18 | 19 | 31 |
| 9. E.rhusiopathiae | 32 | 37 | 37 | 46 | 32 | 40 | 46 | 40 | | 20 | 23 | 38 | 35 | 19 | 15 | 21 | 19 | 21 | 44 |
| 10. C.botulinum | 33 | 36 | 35 | 45 | 33 | 40 | 45 | 39 | 40 | | 62 | 41 | 52 | 28 | 28 | 37 | 29 | 33 | 46 |
| 11. C.tetani | 29 | 33 | 32 | 43 | 29 | 37 | 43 | 36 | 37 | 49 | | 37 | 47 | 26 | 26 | 33 | 25 | 30 | 44 |
| 12. S.pyogenes | 31 | 36 | 35 | 43 | 30 | 39 | 46 | 36 | 39 | 21 | 24 | | 36 | 20 | 17 | 24 | 22 | 23 | 35 |
| 13. F.magna | 30 | 33 | 33 | 44 | 28 | 36 | 42 | 35 | 34 | 26 | 29 | 35 | | 20 | 20 | 28 | 22 | 25 | 38 |
| 14. V.cholerae | 24 | 28 | 27 | 33 | 24 | 30 | 35 | 29 | 28 | 21 | 25 | 29 | 31 | | 32 | 40 | 44 | 35 | 38 |
| 15. D.indicum | 23 | 27 | 26 | 34 | 22 | 28 | 34 | 28 | 24 | 22 | 26 | 27 | 31 | 33 | | 42 | 33 | 37 | 37 |
| 16. C.jejuni | 22 | 25 | 22 | 32 | 21 | 26 | 31 | 23 | 20 | 17 | 20 | 22 | 27 | 25 | 25 | | 29 | 57 | 35 |
| 17. N.gonorrhoeae | 21 | 25 | 24 | 32 | 21 | 27 | 31 | 25 | 22 | 16 | 18 | 25 | 26 | 33 | 24 | 36 | | 34 | 30 |
| 18. H.pylori | 20 | 23 | 21 | 31 | 20 | 25 | 28 | 20 | 18 | 14 | 16 | 20 | 22 | 20 | 20 | 52 | 26 | | 30 |
| 19. B.burgdorferi | 19 | 25 | 22 | 30 | 19 | 25 | 46 | 18 | 21 | 11 | 13 | 17 | 18 | 12 | 11 | 18 | 13 | 17 | |

**Fig. 17**. **Computed homolog heatmaps**. (A) Number of genes in common between Tenericutes and other reference bacteria. In parenthesis is the number of proteins in the reference organism. Cutoff at 800. Separation by *Mycoplasma* clades, Gram-positive and Gram-negative. (B) Percent of genes in common between Tenericutes and other reference bacteria. Heatmap color cutoff at 50%. Separation by *Mycoplasma* clade, Gram-positive and Gram-negative. Homologs are from CoreGenes 3.5, recorded in Microsoft Excel, and heatmap drawn with GraphPad Prism.

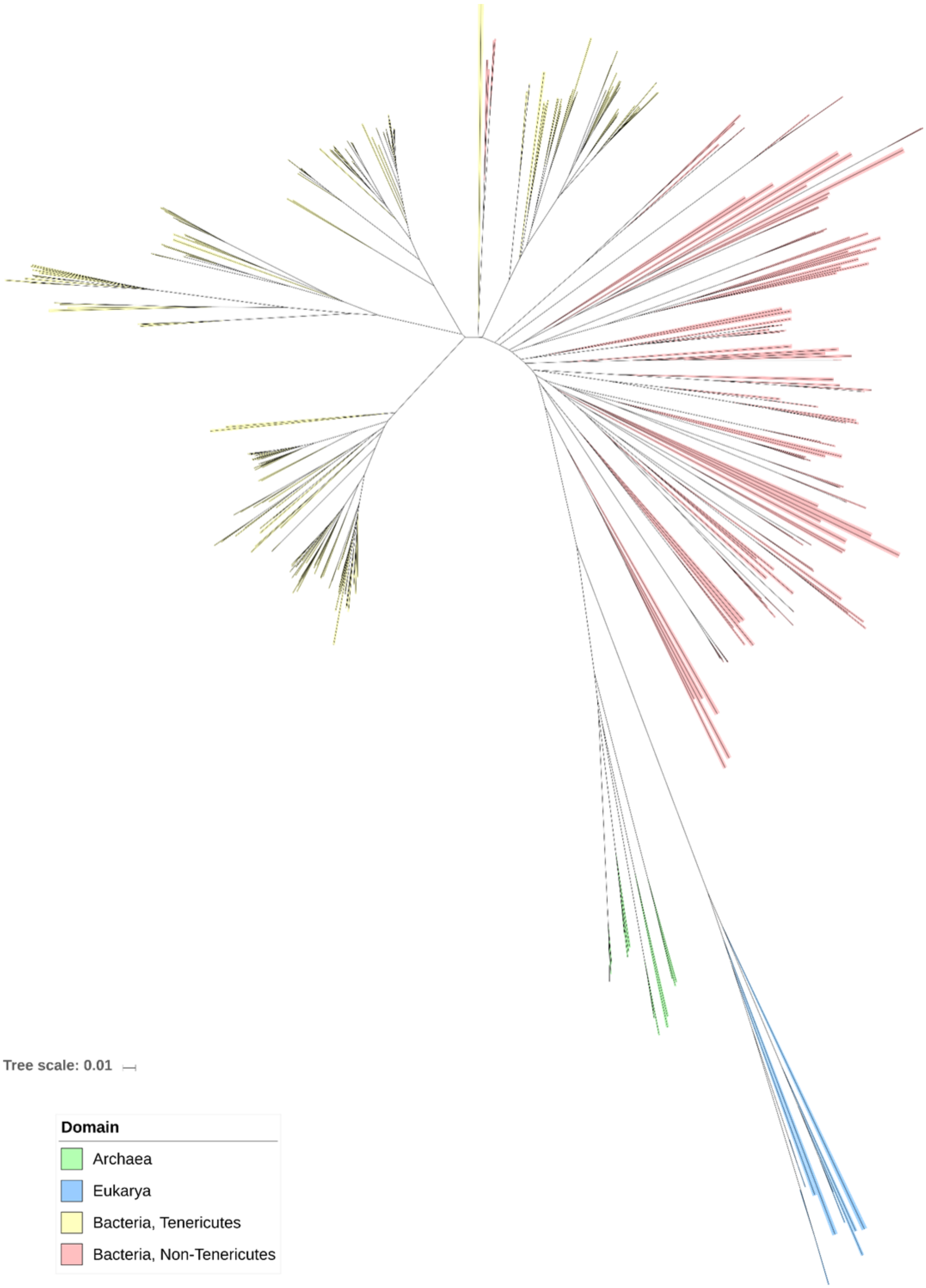


**Fig. 18**. **Unrooted tree of all domains**. In red are the minimal genomes, and in yellow, the Tenericutes. The Tenericutes are phylogenetically distinct. Aligned in EMBL's MAFFT, and tree drawn in iTOL

Phylum Tenericutes displayed less variable regions than the minimal genomes of domain Bacteria (**Figures 19**, **20**). We also tested 16S rRNA virtual universal primers for trimming sequences (**Table 5**). Variable region 4 and 6 (V4, V6) contain the most conserved sequences that are ideal for universal primers. This was followed by variable regions 3, 5 (V3, V5). Inversely, variable region 1, 7, 8, and 9 (V1, V7-V9) are sub-optimal. Nevertheless, primer 27F of the V1 and 1492R of the V9 remain popular choices for detection of microbial communities and medical diagnosis.

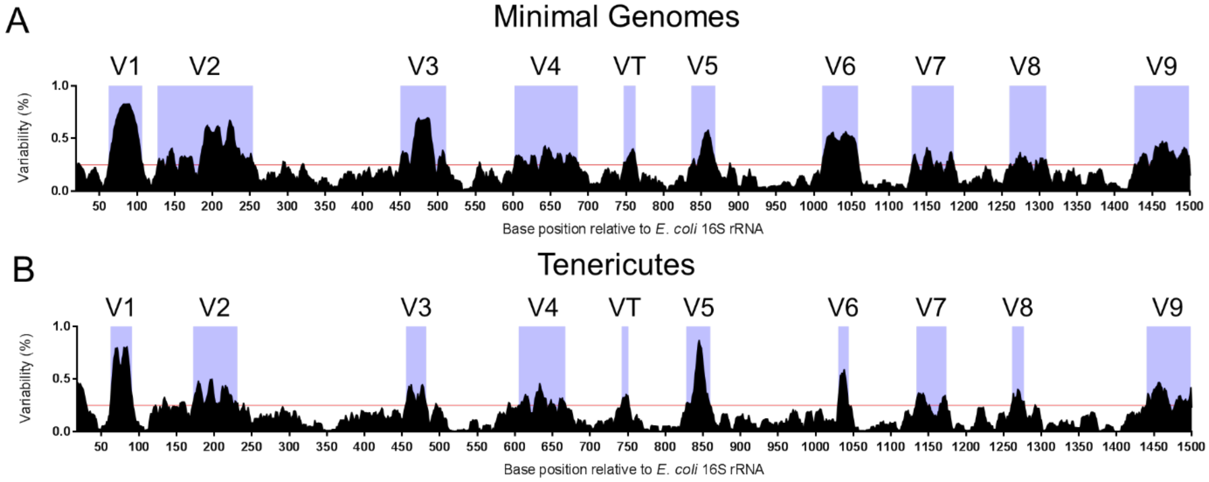


**Fig. 19**. **Histograms of the most frequent base at each position**. The alignment results used for drawing the phylogenetic trees were the same ones used for computing these variable regions displayed above. The peaks indicate variance, higher peaks (hills) indicating more variance and valleys (dips) indicating conserved regions. Variable regions are ideal for determining the identity of the organism, while conserved regions are best suited for primer binding. Variance is the moving average per 10bp periods of percent variability. Red line is at 25% variability. Blue shade indicates variable region (≥25% variability). Variable Transition (VT) region is a transition we found consistently between V4 and V5. V1 contains the highest variance by percentage, whereas V2 represents its largest extension. More details in **Table 6**. The alignment used for drawing this variability histogram was EMBL's MAFFT with 100 iterations. Alignments were converted to numbers and separated into cells. *Escherichia coli* was used as a reference, and only bases aligning with *E.coli* were kept. Modes at each base position, counts of the mode at each base position, and moving averages of 10 bp periods were calculated in Microsoft Excel. (A) Variable regions in minimal genomes. (B) Tenericutes. Tables were transposed and exported to GraphPad Prism.

**Minimal Genomes**

| Variable Region | *E.coli* Range | N (bp) | Gaps (bp) | Mean | | Std. Deviation | Q1 | Q2 | Q3 | Area under the curve (Mean*N) | Area under the curve (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic |
| V1 | 62-107 | 46 | 6 | 0.627 | 0.035 | 0.235 | 0.567 | 0.675 | 0.824 | 28.9 | 0.141 |
| V2 | 127-254 | 128 | 38 | 0.379 | 0.021 | 0.232 | 0.170 | 0.438 | 0.576 | 48.6 | 0.238 |
| V3 | 450-511 | 62 | 18 | 0.444 | 0.033 | 0.261 | 0.210 | 0.549 | 0.682 | 27.5 | 0.135 |
| V4 | 602-686 | 85 | 35 | 0.301 | 0.023 | 0.216 | 0.070 | 0.342 | 0.496 | 23.8 | 0.116 |
| VT | 747-763 | 17 | 8 | 0.279 | 0.048 | 0.199 | 0.101 | 0.327 | 0.435 | 4.74 | 0.023 |
| V5 | 837-869 | 33 | 11 | 0.379 | 0.037 | 0.212 | 0.208 | 0.442 | 0.556 | 6.44 | 0.032 |
| V6 | 1011-1059 | 49 | 7 | 0.466 | 0.025 | 0.173 | 0.392 | 0.523 | 0.587 | 22.8 | 0.112 |
| V7 | 1130-1186 | 57 | 29 | 0.263 | 0.026 | 0.200 | 0.085 | 0.230 | 0.442 | 15.0 | 0.073 |
| V8 | 1260-1309 | 50 | 24 | 0.265 | 0.029 | 0.208 | 0.035 | 0.279 | 0.436 | 1.32 | 0.006 |
| V9 | 1426-1499 | 74 | 19 | 0.341 | 0.020 | 0.176 | 0.240 | 0.327 | 0.481 | 25.2 | 0.123 |

**Tenericutes**

| Variable Region | *E.coli* Range | N (bp) | Gaps (bp) | Mean | | Std. Deviation | Q1 | Q2 | Q3 | Area under the curve (Mean*N) | Area under the curve (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic |
| V1 | 62-91 | 30 | 6 | 0.599 | 0.069 | 0.377 | 0.349 | 0.572 | 0.988 | 18.6 | 0.160 |
| V2 | 172-239 | 68 | 25 | 0.330 | 0.025 | 0.207 | 0.105 | 0.383 | 0.490 | 21.4 | 0.184 |
| V3 | 455-482 | 28 | 13 | 0.323 | 0.043 | 0.228 | 0.176 | 0.322 | 0.468 | 9.1 | 0.078 |
| V4 | 605-667 | 63 | 29 | 0.282 | 0.031 | 0.247 | 0.012 | 0.297 | 0.486 | 17.8 | 0.153 |
| VT | 742-751 | 10 | 4 | 0.267 | 0.062 | 0.197 | 0.107 | 0.299 | 0.442 | 2.7 | 0.023 |
| V5 | 828-860 | 33 | 11 | 0.459 | 0.057 | 0.330 | 0.156 | 0.465 | 0.742 | 6.0 | 0.052 |
| V6 | 1030-1044 | 15 | 10 | 0.393 | 0.115 | 0.443 | 0.031 | 0.148 | 0.992 | 5.9 | 0.051 |
| V7 | 1134-1174 | 41 | 19 | 0.260 | 0.031 | 0.197 | 0.082 | 0.254 | 0.375 | 10.6 | 0.091 |
| V8 | 1261-1277 | 17 | 8 | 0.300 | 0.059 | 0.244 | 0.012 | 0.340 | 0.508 | 5.1 | 0.044 |
| V9 | 1440-1499 | 60 | 26 | 0.317 | 0.025 | 0.190 | 0.172 | 0.285 | 0.490 | 19.0 | 0.164 |

**Table 5. Descriptive statistics for the variable regions**. There were a total of ten variable regions identified in this study. The parameters used for determining variable regions were variability of moving averages of ≥25% and gaps of ≤10bp. The longest variable region was V2. Based on the area under the curve, V2 also encodes the most information for a variable region, containing 18-24% of all variability within variable regions. The runner-up for most information in a variable region was in V1. A primer pair spanning from V1 to V3 will contain 51.4% of all variance in minimal genomes and 42.2% of variance in Tenericutes. The shortest variable region was the Variable Transition (VT) region. The two most universal primers 515FB and 926R (**Table 6**) cover region V4 to V5. These represent 17.1% and 22.8% of all variance in minimal genomes and Tenericutes, respectively.

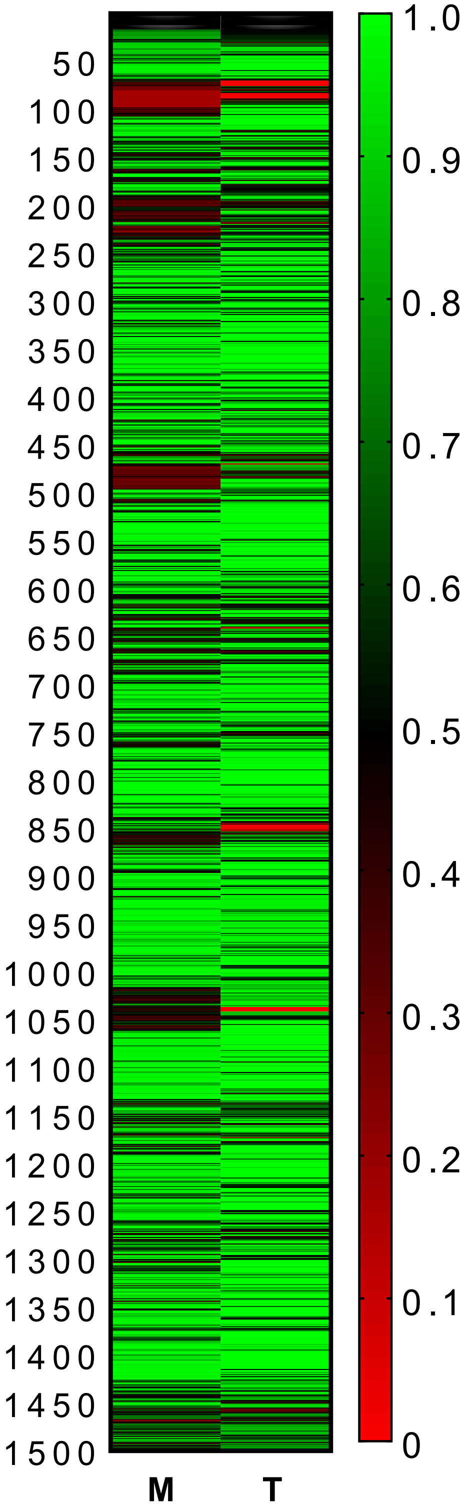


**Fig. 20**. **Variable regions heatmap**. Left (M) are the minimal genomes and to the right (T), the Tenericutes. In red are the variable regions and in green the conserved regions. Values are percentages in decimal. Base position is relative to *E. coli*.

| | | | All | | Archaea | | Bacteria | | Eukaryota | |
|---|---|---|---|---|---|---|---|---|---|---|
| Name | Sequence | Matches | Selected | Coverage | Selected | Coverage | Selected | Coverage | Selected | Coverage |
| 515FB | GTGYCAGCMGCCGCGGTAA | 587 252 | 0.11 | 0.94 | 0.92 | 0.93 | 0.94 | 0.94 | 0.89 | 0.91 |
| 926r | CCGYCAATTYMTTTRAGTTT | 583 847 | 0.10 | 0.93 | 0.90 | 0.90 | 0.93 | 0.93 | 0.92 | 0.92 |
| 515F | GTGCCAGCMGCCGCGGTAA | 578 396 | 0.10 | 0.92 | 0.56 | 0.57 | 0.94 | 0.94 | 0.89 | 0.91 |
| 533F | GTGCCAGCAGCCGCGGTAA | 564 617 | 0.10 | 0.90 | 0.01 | 0.01 | 0.94 | 0.94 | 0.89 | 0.91 |
| 519R | GWATTACCGCGGCKGCTG | 540 050 | 0.10 | 0.86 | 0.51 | 0.51 | 0.87 | 0.87 | 0.87 | 0.89 |
| 907R | CCGTCAATTCMTTTRAGTTT | 536 173 | 0.10 | 0.85 | 0.01 | 0.01 | 0.90 | 0.90 | 0.75 | 0.75 |
| 518F | CCAGCAGCCGCGGTAAT | 527 856 | 0.10 | 0.84 | - | - | 0.87 | 0.87 | 0.87 | 0.89 |
| 806R | GGACTACNVGGGTWTCTAA | 522 770 | 0.09 | 0.83 | 0.91 | 0.91 | 0.93 | 0.93 | - | - |
| 806RB | GGACTACNVGGGTWTCTAAT | 516 234 | 0.09 | 0.82 | 0.91 | 0.91 | 0.92 | 0.92 | - | - |
| IlluminaR | GACTACHVGGGTATCTAATCC | 510 061 | 0.09 | 0.81 | 0.91 | 0.91 | 0.91 | 0.91 | - | - |
| 357F | CTCCTACGGGAGGCAGCAG | 487 689 | 0.09 | 0.78 | - | - | 0.91 | 0.91 | - | - |
| 1392R | ACGGGCGGTGTGTRC | 442 347 | 0.08 | 0.85 | 0.36 | 0.70 | 0.70 | 0.85 | 0.87 | 0.93 |
| 1496R | ACGGGCGGTGTGTRCAA | 438 658 | 0.08 | 0.75 | 0.36 | 0.66 | 0.69 | 0.73 | 0.87 | 0.92 |
| 1391R | GACGGGCGGTGTGTRCA | 436 233 | 0.08 | 0.84 | 0.35 | 0.66 | 0.69 | 0.83 | 0.87 | 0.92 |
| U1390R | GACGGGCGGTGTGTRCAA | 434 411 | 0.08 | 0.74 | 0.35 | 0.64 | 0.68 | 0.72 | 0.86 | 0.92 |
| CC [F] | CCAGACTCCTACGGGAGGCAGC | 393 385 | 0.07 | 0.63 | - | - | 0.73 | 0.73 | - | - |
| 1185mR | GAYTTGACGTCATCCM | 387 512 | 0.07 | 0.63 | - | - | 0.72 | 0.72 | - | - |
| CD [R] | CTTGTGCGGGCCCCCGTCAATTC | 341 461 | 0.06 | 0.54 | - | - | 0.64 | 0.64 | - | - |
| 1381R | CGGTGTGTACAAGRCCYGRGA | 339 084 | 0.06 | 0.57 | - | - | 0.63 | 0.65 | - | - |
| 1381bR | CGGGCGGTGTGTACAAGRCCYGRGA | 330 078 | 0.06 | 0.56 | - | - | 0.61 | 0.64 | - | - |
| 895F | CRCCTGGGGAGTRCRG | 321 190 | 0.06 | 0.51 | 0.20 | 0.20 | 0.59 | 0.59 | - | - |
| 905F | TGAAACTYAAAGGAATTG | 315 907 | 0.06 | 0.50 | 0.84 | 0.84 | 0.46 | 0.46 | 0.74 | 0.74 |
| 27F | AGAGTTTGATCMTGGYTCAG | 309 429 | 0.06 | 0.50 | - | - | 0.58 | 0.58 | - | - |
| 1100R | AGGGTTGCGCTCGTTG | 308 508 | 0.06 | 0.50 | - | - | 0.57 | 0.57 | - | - |
| 16S.1100.F16 | CAACGAGCGCAACCCT | 308 508 | 0.06 | 0.50 | - | - | 0.57 | 0.57 | - | - |
| 1237F | GGGCTACACACGYGCWAC | 296 340 | 0.05 | 0.48 | - | - | 0.55 | 0.55 | - | - |
| 1492R (s) | ACCTTGTTACGACTT | 204 306 | 0.04 | 0.75 | 0.14 | 0.81 | 0.31 | 0.75 | 0.50 | 0.73 |
| 8F | AGAGTTTGATCCTGGCTCAG | 112 196 | 0.02 | 0.56 | - | - | 0.21 | 0.65 | - | - |
| 1492R | TACGGYTACCTTGTTACGACTT | 77 532 | 0.01 | 0.29 | 0.06 | 0.36 | 0.14 | 0.34 | - | 0.01 |
| 1492R (l) | GGTTACCTTGTTACGACTT | 69 727 | 0.01 | 0.26 | 0.05 | 0.26 | 0.13 | 0.31 | 0.01 | 0.02 |
| 1185aR | GAYTTGACGTCATCCA | 14 582 | - | 0.02 | - | - | 0.03 | 0.03 | - | - |
| EukA, 1A, Euk1F | AACCTGGTTGATCCTGCCAGT | 4 959 | - | 0.10 | - | - | - | - | 0.07 | 0.39 |
| IlluminaF | CCTACGGGGNGGCWGCAG | 158 | - | - | - | - | - | - | - | - |

**Table 6**. **Universal primers tested *in silico***. Test ran on SILVA using probe search (https://www.arb-silva.de/search/testprobe/). International Union of Pure and Applied Chemistry (IUPAC) nucleotide nomenclature was used for the sequences portion. 515FB located between V4 and V5 had the most hits at 587,252. The runner up was 926R at 583,847 16S or 18S rRNA matches. 926R is located between V5 and V6. Results were recorded in Microsoft Excel.

We also drew sequence logos to see which bases were present at which position of the final alignment done in MAFFT for both, the minimal genomes and the Tenericutes (**Fig. 21**). Interestingly, universal primer sequences can be seen clearly in the figures.

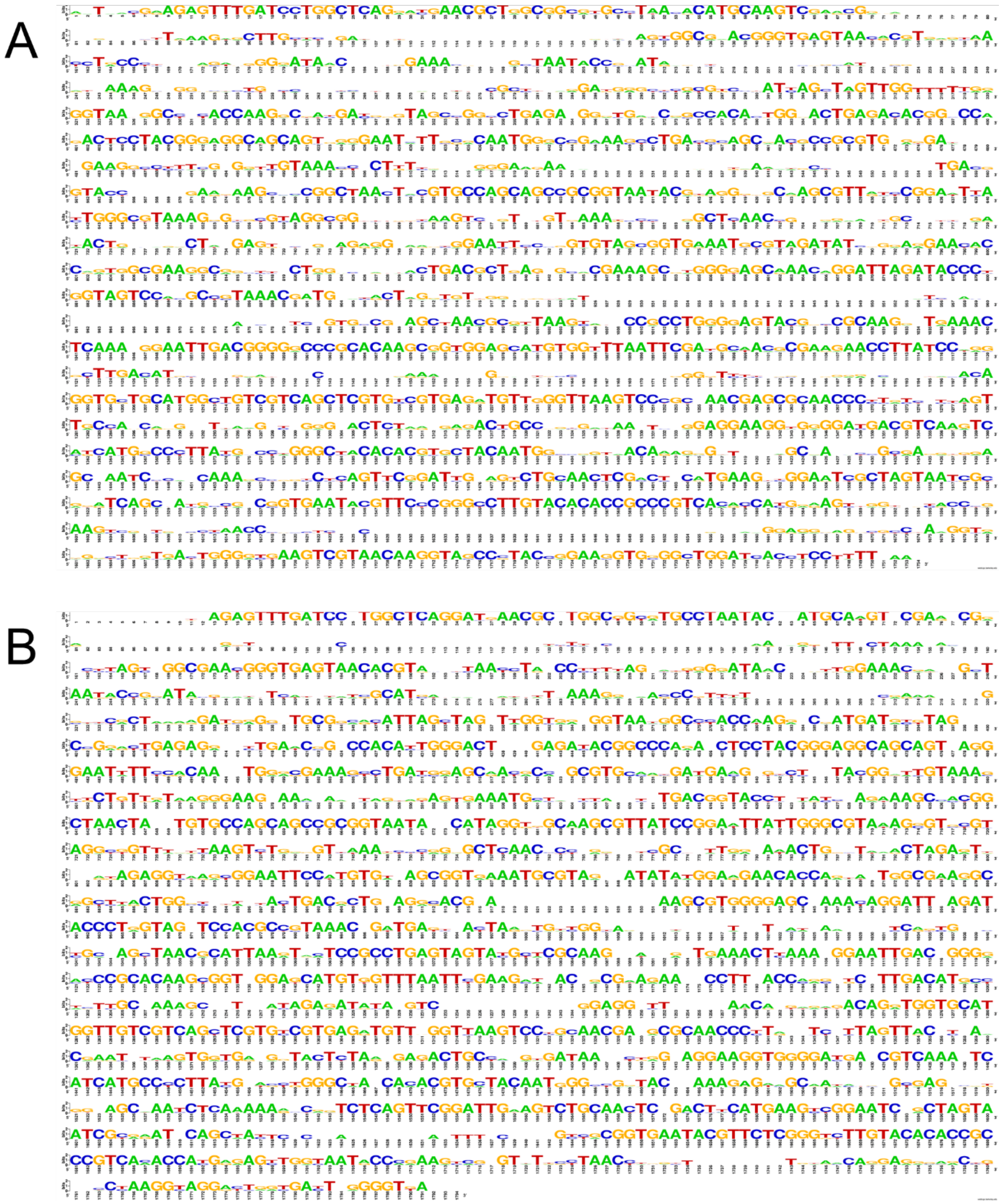


**Fig. 21**. **Sequence logos of the 16S rRNA sequence alignments**. These alignments are the same ones used for the trees and the variable regions. Thus, they were computed using MAFFT. (A) Minimal genomes alignment. (B) Tenericutes alignment. Universal primers like the 8F (5'-AGAGTTTGATCCTGGCTCAG-3') are visible. Graphics drawn with WebLogo.

# DISCUSSION

## Minimal Genomes Conserve Phylogenies

When examined at the 16S rRNA level, minimal genomes conserve phylogenetic relationships (**Fig. 4**). Most phyla were together in clades, and appear to be congruent with other phylogenetic trees [175]. This suggests minimal genomes may be suitable as representatives of their own phylas for comparative studies on conserved genes.

We observed a correlation between genome size, GC content, and gene count (**Table 3**). The smaller the genome size, the less genes and less proteins ($p<.001$). We warn that this correlation might not hold true with the inclusion of organisms that have lower coding densities (less than 800 proteins/Mb), including Eukaryotes. According to NCBI genome medians, *Saccharomyces cerevisiae* has 5,409 genes and a genome size of 12.1234 Mb for a total of 446 proteins per Mb; the numbers for *E. coli* are as follows: 4,996 genes, 5.13829 Mb, and 972 proteins per Mb. Roughly speaking, *S. cerevisiae* and *E. coli* have similar protein counts, but the genome size is halved in *E. coli*. Our average protein count per Mb was 887± 87, and the genome 2.32± 1.44 Mb (**Table 1)**. The correlation we observed between genome size and gene count was likely formed because our study was limited to domain Bacteria, hence the narrow standard deviation when compared against the standard deviation between *Saccharomyces* and *E. coli*, which is 5.13 Mb.

## I'm AT Rich! Minimal Genomes Have Lower GC Contents

Smaller genomes had noticeably lower GC contents ($p<.001$). Splitting the minimal genomes by percentiles, the lower 50th percentile of GC content (43.8% GC) had genome sizes of 1.74± 1.02 Mb and the upper 50th percentile had a genome size of 2.90± 1.58 Mb (for a histogram, see **Fig. 5**). The Tenericutes also have a reputation for being AT rich. However, this has been suggested to be feature of bacteria with small genome size and not necessarily the Tenericutes [203].

This information seems paradoxical because stop codons are biased towards adenine and thymine (AT) [228], coding regions of the genome are biased towards GC [229–232], and the smaller the genome size, the higher the coding density (coding proteins per Mb) (**Fig. 5**) [233–235]. Furthermore, longer coding mRNAs have higher GC contents [228]. High mutation rates has been associated with smaller genome size [182] and higher AT contents [236,237]. Thus, by extrapolation combined with our observations, we can assume smaller genomes have higher AT contents. Indeed, this finding has been previously reported [226,227,238,239]. However, Hildebrand, Meyer, and Eyre-Walker suggest organisms with high point mutation rates have GC to AT bias, yet AT rich organisms have AT to GC bias predominantly [237], and most organisms have AT bias [224]. Sequencing errors, infinite sites violations, mutational biases, translational selection, or biased gene conversion have been discarded [224,237].

The question remains, why be AT rich? Since free-living bacteria are GC rich, it has been suggested that competition for the resources make pathogenic bacteria select for AT [226]. Many minimal genomes are pathogenic [240], and minimal genomes tend to be AT rich [226,227,238,239]. Yet according to NCBI, the human genome has a median GC content of 40.9%; mice (*Mus musculus*), 42.4891%; chicken (*Gallus gallus domesticus*), 42.9197%; zebrafish (*Danio rerio*), 36.7687%; western clawed frog (*Xenopus tropicalis*), 40.5404%; fruitfly (*Drosophila melanogaster*), 41.9%; and plant *Arabidopsis thaliana*, 36.6%; corn (*Zea mays*), 46.7867%. There is an abundance of AT or lack of GC in the host's environment (GC content of 41.11± 3.33% for the aforementioned hosts), which it is more likely to explain the AT rich pathogens than competition with the host. Moreover, a mere 0.3% or 200g of our body is bacteria [241,242]. If known pathogenic and opportunistic bacteria are considered, then the competition would seem a bit more possible, though remotely. *E. coli* has 50.6% GC content; *Klebsiella pneumoniae*, 57.1%; *Morganella morganii*, 51%; *Pseudomonas aeruginosa*, 66.2%; *Streptococcus pyogenes*, 38.5%; *Staphylococcus aureus*, 32.8%. Again, a drop in the bucket if we consider masses. We also have to consider that the host and the minimal genomes GC contents are close (41.11± 3.33%, 45.42± 1.17%, respectively). We also have to take into account the biological composition of our microbiota, including the GC content of the microbiome.

Perhaps life respects equilibrium, hence the AT rich shift to GC and GC rich organisms to AT. After all, mutations happen at random. There is an equal probability of having AT or GC. This is further evidenced in the GC contents range of our minimal bacteria (25.70-73.30%) and the GC contents range others have found (25-75%) [224].

## Warning: Computer Codes May Not Accurately Reflect Biological Relationships

Inferred phylogenetic relationships may be contingent upon the algorithms used for multiple sequence alignments as well as the codes used for drawing the phylogenetic trees. We used nineteen sequences for quickly analyzing relationships but discovered high variation between phylogenetic trees. Therefore, we recommend more than nineteen (19) sequences for drawing trees.

As the sample size increased, we noticed that published sequences in reference databases like NCBI and SILVA are reverse-complement sequences in relation to *E. coli* after observing a deep-rooted clade whose branches were distant from the rest of the domains. A quick fix would be to employ universal primers to detect direction. We recommend the 515FB primer (5'-GTGYCAGCMGCCGCGGTAA-3'). This primer offers a 93.6% coverage of the SILVA database. It is able to detect Archaea, Bacteria, and Eukaryota. For Archaea, it selected 92% of the sequences and had a 93% coverage; for Bacteria, the percentages were 94% and 94%, respectively; Eukaryota, 89%,91.4% (**Table 6**).

Alignments improve dramatically in quality when sequences are trimmed to start and end with conserved regions, or at least kept at similar lengths if there is only one conserved region in the sequence of interest. Another benefit of universal primers is that they may be used for trimming. We employed primer 27F

(5’-AGAGTTTGATCMTGGYTCAG-3’) and the reverse-complement of the popular 1492R (5’-AAGTCGTAACAAGGT-3’). These two primers represented the best attempts at developing universal primers at the ends of the 16S, more specifically, before variable region 1 (V1) and after variable region 9 (V9).

### The Tenericutes Have Different Ancestors

The Tenericutes were noticeably apart in renderings of only minimal genomes and minimal genomes plus Tenericutes using multiple algorithms. We agree that the Tenericutes are paraphyletic [221,243]. We noted that its species are phylogenetically distant from other sequences (**Figures 4, 6, 11, 18**). We observed at least two distinct clades, and there may be as many as five clades (**Fig. 11**). Some clade like clade number two (2) has an affinity to be bloodborne. The whole-genome alignments show few similar genes (**Figures 19, 20**), as did proteome analysis (**Fig. 17**).

The most likely reason as to why “Tenericutes” exist as a taxon is because most bacteria void of a cell wall were lumped together, thus forming this phylum. The loss of a cell wall might have been the product of convergent evolution. Our proteome analysis (**Fig. 12**) suggest either excessive horizontal gene transfer or distant common ancestors.

### Stop! Why UGA Became a Tryptophan Codon in Tenericutes

The UGA (“*opal*”) switch from a stop codon to a tryptophan codon is one of the most common alterations to the “universal” codon code, especially in bacteria [244]. In two clades of the Tenericutes, UGA codes for tryptophan. Either convergent evolution happened, or these organisms speciate rapidly given the tree distances. This low GC content has led to speculation that selective pressure converted UGA from a stop codon to tryptophan in a hypothesis called ‘codon capture’ or ‘codon reassignment’ [245–247]. This does not explain how *Candidatus* Hodgkinia cicadicola, a GC rich organism, switched UGA to tryptophan [248].

Speculation exists that this codon change has to do with selective pressure since the AT rich genome would prefer UGA over UGC, yet eukaryotic mitochondria also have UGA coding for tryptophan, yet they are not AT rich [249]. For example, the mitochondrial GC content of the human mitochondria is 44.4%; mice (*Mus musculus*), 36.7%; chicken (*Gallus gallus domesticus*), 46.0%; zebrafish (*Danio rerio*), 39.6%; western clawed frog (*Xenopus tropicalis*), 42.5%; fruitfly (*Drosophila melanogaster*), 17.8%; and plant *Arabidopsis thaliana*, 44.8%; corn (*Zea mays*), 38.5%. The average GC content for the these mitochondria is 38.8± 9.1%, or 3.6% below the GC content of their whole genomes.

We believe that release factor 2, which is associated with UAA (*ochre*) and UGA (*opal*) stop codons, was lost. It is not present in *Mycoplasma* [250]. UAA, along with UAG (*amber*), is also associated with release factor 1, is commonly seen after UGA in *Mycoplasma* [251]. This was the first step on the road to UGA as a tryptophan codon. Since the tryptophan codon UGG is similar to UGA, the difference being in the wobble position, it makes sense that eventually the Tenericutes made UGA code for tryptophan. This shift may also be a defense mechanism against virus [252]. This does not resolve the shift of why Mycoplasmas prefer UGA for coding tryptophan instead of UGG, but at least, it offers an explanation for its origins.

### Who Are You? The Importance of Variable Regions in Diagnostics

Variable regions of the 16S rRNA are segments of the genetic code (DNA, RNA) which look different even for closely related species. As previously discussed, there are conserved regions in the DNA that are able to serve as binding regions for synthetic pieces of DNA. These synthetic pieces are called primers. Primers, when combined with ingredients like pure cell cultures, DNA polymerase, nucleotides, salts, heat, and water, are able to make exponentially more copies of DNA in a process that mimics nature. With enough copies, these new DNA molecules are able to be visualized with a staining solution after the molecules are separated in an agarose gel. This process is called polymerase chain reaction (PCR).

PCR has been a powerful tool in medical diagnosis and ecology for identifying bacteria. The 16S has been used because it has enough similar (conserved) sequences to permit primers to bind, yet it has also enough dissimilar (variable) sequences to tell apart bacterial identities (species). At present, researchers have been using primers to amplify the whole 16S. The problem with the said approach is that the primers used have less binding affinity than primers like 515FB (**Table 6**). In other words, many bacteria will be left out. In order to counteract this, a new trend in laboratories has been to use more conserved regions, but ensuring that variable regions are captured. In our case, we studied the 16S because we wanted to develop multiplex diagnostics. By accident, we discovered the importance of the variable regions. Primers for detecting species (species-specific primers) should be placed in these regions, while common primers should reside in conserved regions. Common primers may be used not only to reduce the number of primers used, but also as positive controls. The aforementioned separation of molecules in PCR runs logarithmically, and while we identified around eleven (11) spots where these primers could bind, that means that primers opposite of the common primer (not the positive control) would be squished together. If the DNA ran linearly within the gel, there would be no issues, but that is not the case. Thus, variable regions may be split to allow more species-specific primers only if they are placed near the common primer (again, not the positive control).

Our study found that the most variable region was V1 (0.634), followed by V6 (0.482) and V3 (0.439) for the minimal genomes, and for the Tenericutes, it was V1 (0.574), followed by V5 (0.433) and V6 (0.429). Again putting this information in the context of medical diagnostics and ecology, primer pair 27F and 1496R would provide the most information if the product were to be sequenced, but these primers have low binding with DNA when compared to the most specific primer pair, 515FB and 926R. However, the latter offer less variability information (19.7% vs 27F, 1496R). Nevertheless, it may be useful for astrobiology and for searching lifeforms. A compromise would be 515F and 1392R (41.1%) or 357F and 515RB, which are slightly less specific than 515FB and 926R, but offer enough information (33.7%). We clarify that these numbers are extrapolations of our findings. At present, we are working on gathering experimental evidence to confirm the binding ability of the primers in **Table 5** and the variability as observed in **Table 6**.

### The Last Universal Common Ancestor

Minimal genomes naturally select which genes to keep, and when combined with global transposon mutagenesis, it can offer which genes are the most essential ones for life [217]. Knowing which genes (conserved and essential) represent life can help us understand the nature of the Last Universal Common Ancestor (LUCA) [253] and the protobionts, that is, the first "organisms" or life-like organic bubbles on Earth. However, the definition of minimal genomes must be revisited when dealing with astrobiology. Many minimal genomes are pathogens [239]. That means they depend on other organisms for sustaining life. Autotrophic organisms with minimal genomes may reflect more accurately LUCA [254].

### Searching for Life Outside Earth Using Minimal Knowledge

There is a binary opposition as to the origins of Life on Earth. Either it started here or it was seeded from elsewhere. These ideas are called abiogenesis and panspermia [255], respectively. Aware that the origins of life are of contentious debate, and that the majority of biologists agree with abiogenesis [256,257], let us nevertheless dance on top of glass shards and entertain panspermia. If life was seeded from elsewhere, then our searches for extraterrestrial life should focus on the most conserved genes, which may be condensed from comparative genome studies. We suggest using minimal genomes for this purpose.

On a similar note, our space exploration activities, however unlikely, may have seeded life in Mars, Titan, and comets through manmade landers [258,259], or even less likely through orbiters such as Cassini, which performed a dive through the water plumes of Enceladus [260,261]. We reiterate that safety precautions have been taken to prevent forward-contamination (from Earth to space), namely the United Nations (UN) Space Treaty of 1967 [262], the Committee on Space Research (COSPAR) Planetary Protection Policy [263], the Office of Planetary Protection of the National Aeronautics and Space Administration (NASA) [264], and the Planetary Protection Officer of the European Space Agency (ESA) [265]. Nevertheless, genes and its products conserved here on Earth may be used for ensuring planetary preservation and testing for biological contamination in outer space [258].

Shifting our focus to abiogenesis, we ask again, *How did life on Earth start*? Filtering out panspermia, there are plenty of models and conjectures based on abiogenesis with varying degrees of acceptance [266], including RNA world [267–269], Polycyclic aromatic hydrocarbons (PAH) hypothesis [270], Graded Autocatalysis Replication Domain (GARD) model [271], Iron–sulfur world hypothesis [272], Pyrite hypothesis [273], Zinc (Zn)-world hypothesis [274], Deep sea vent hypothesis [275], Thermosynthesis [276], Clay hypothesis [277], Deep-hot biosphere model [278], Lipid world [279], Radioactive beach hypothesis [280,281], and Thermodynamic dissipation evolution [282]. While these ideas all have their differences, they all agree that terrestrial life begun here on Earth. Whichever the recipe for life may be, the ingredients are still the same.

Focusing solely on the 4-5% of the Universe that's ordinary ("baryonic" in the astronomical sense, atomic, normal) matter (composed of protons, neutrons and electrons) [283,284], the relative abundances of the chemical elements by mass fraction are around 74.91% hydrogen (H), 23.77% helium (He), and 1.33% heavy elements [285]. If we made a list of the ten (10) most common elements in the Universe [285] and in the human body [286] treating all atoms as equals (mol), five elements would be shared: H, O, C, N, and S [285,286]. If again, we measure atoms by mole abundances, comparing humans [286] and oceans [287,288], we find that we share eight out of our ten most common elements, namely H, O, C, Na, Ca, S, Cl, and K. Our body is composed mainly of water [289]. Not surprisingly, the composition of our bodies are most similar to Earth's seawater (**Fig. 23**). This suggests that the relative abundance of the elements in living things are similar to its environment.

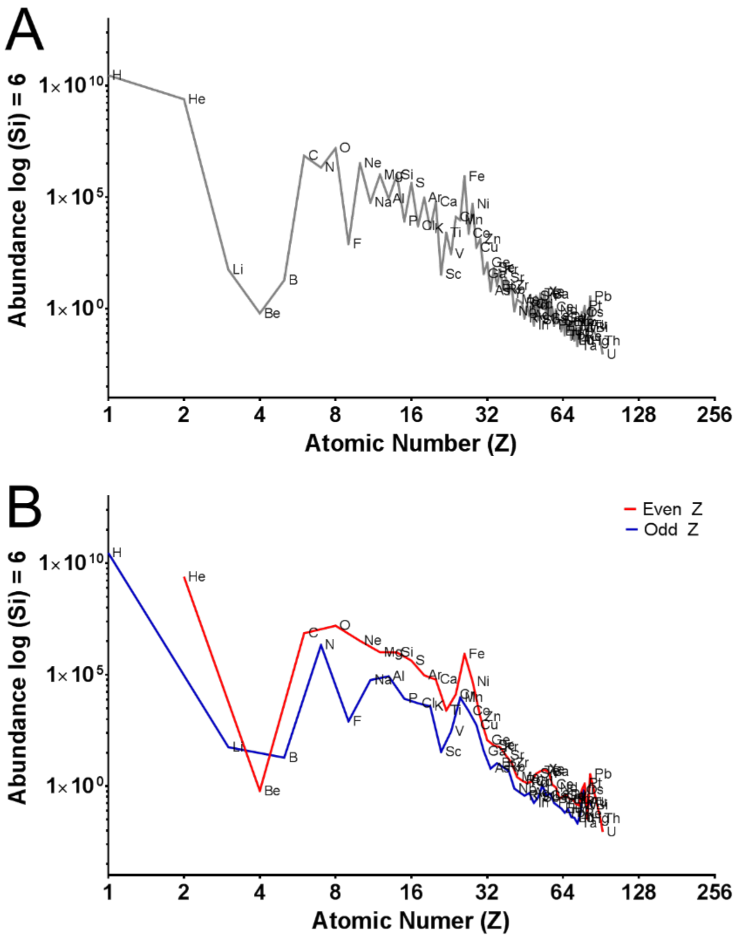


**Fig. 22. Nucleosynthesis mass by atomic number**. (A) All elements. (B) Odd atomic numbers (Z) and even atomic numbers separated. Relative abundance based on mass fraction. Graph drawn using GraphPad Prism.

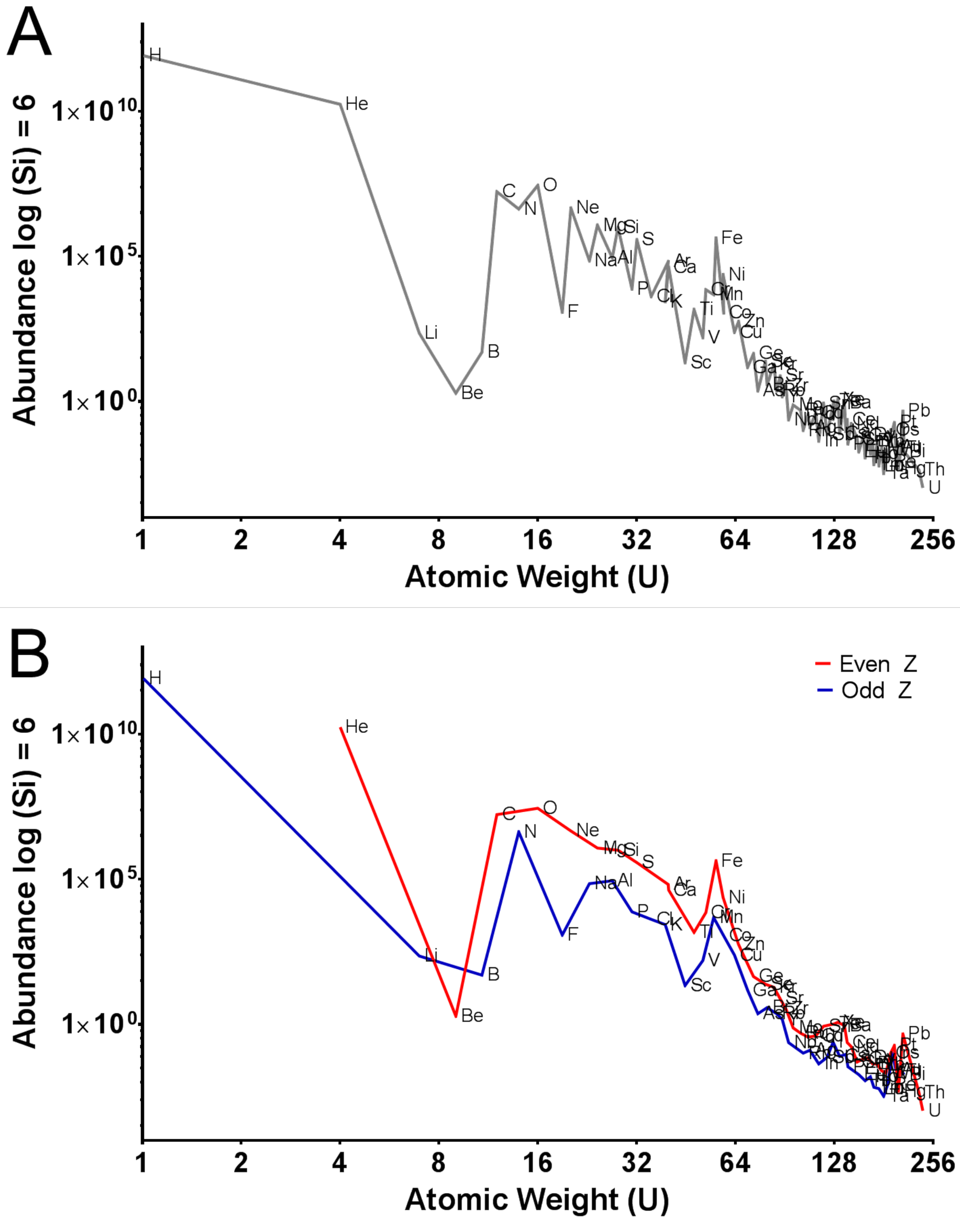


**Fig. 23. Nucleosynthesis mole by atomic weight**. (A) All elements. (B) Odd atomic numbers (Z) and even atomic numbers separated. Relative abundance based on mole fraction. Graph drawn using GraphPad Prism.

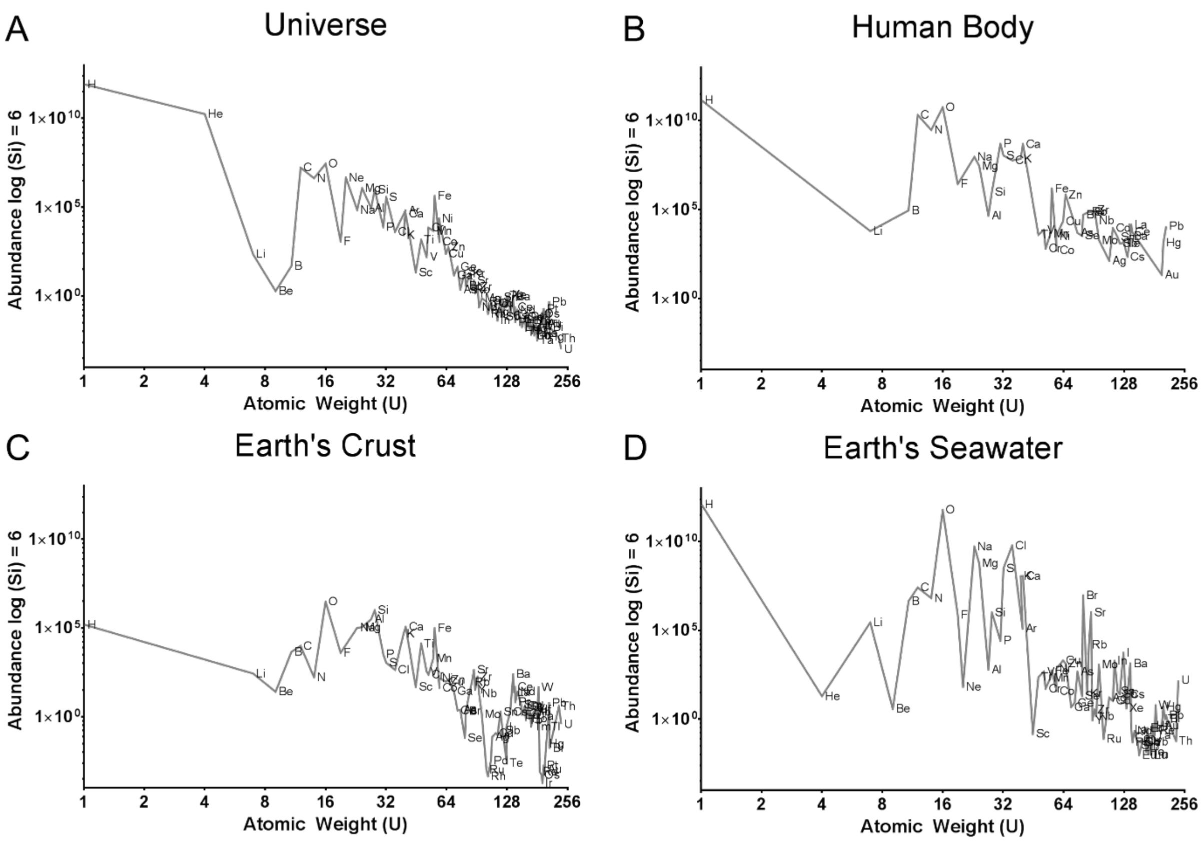


**Fig. 23**. **Relative abundance of the chemical elements by mole fraction**. (A) Universe [285], (B) human body [286], (C) is the Earth's crust [290], (D) Earth's seawater [287,288]. Humans (B) are most similar to Earth's seawater (D).

How likely is it to have a similar environment to our biosphere? While we are not going into the probabilities of another Earth, do consider that the chemical abundances in the planetary nebulae (where planets are formed) in the three spiral galaxies of our Local Group (i.e., galaxy group of over 54 galaxies which we belong to) are not that varied [291], just like the homogeneity observed in the universe through cosmic microwave background (CMB) radiation [292]. It has been suggested that the limited information we have on planetary nebulae be extended to all nebula [293]. Planetary nebulae offer the advantage of permitting spectroscopy which in turn allows us to study the chemical composition of stars (i.e., nebulae precursor) and these clouds (i.e., nebulae) [291,294]. Astronomical spectroscopy is an optical technique capable of resolving chemistries (i.e., elements, molecules) by reading the fingerprint spectral lines electromagnetic (e.g., infrared, visible light, ultraviolet) waves produce. For example, streetlights are sometimes orange because the heated sodium gas in the bulbs produce orange emissions. In a similar manner, spectrometers are able to detect the spectral lines of chemicals, either by reading the light emitted from heat (i.e., emission lines), or by reading the light absorbed from a heated object by a cold object (i.e., absorption lines). While colors may seem similar to the naked eye, spectrometers are able to break apart the frequencies of the light (i.e., photons) using diffraction gradients and prisms. The farther away an object is, the dimmer it becomes. Therefore, spectrometry has its distance limitations. This is further complicated by special relativity and motion in a phenomenon known as the Doppler effect. Celestial bodies are not stationary. Objects moving away produce red-shifts in the spectral lines because the energy in the photons lower, and objects moving closer produce blue light. Exaggerating for blueshift, this means that our orange streetlights will now look green from the Andromeda galaxy, our neighboring galaxy that's moving toward us. The Doppler effect can be corrected using parallax, but it requires patience as at least two measurements are needed to allow Earth to change positions. Try covering one eye at a time, and notice how the moon or a distant object moves positions. This is parallax. The farther the object is, the smaller the parallax until it becomes indistinguishable.

Once again, assuming that stellar nucleosynthesis offers similar relative abundances of the chemical elements throughout the Universe, then we should give serious contemplation to the possibility of cosmical convergent evolution. Like birds, bats have wings. The bat ancestor, however, was an ancient mammal that crawled using arms, not wings. We know this because the bat wing bones and the mouse arm bones are morphologically similar, and the rRNA of bats look more closely related to that of mice than to any bird. The wings may look similar, but they have different evolutionary origins. This is called convergent evolution.

Similarly, the same means by which life originated on Earth may recur elsewhere. The ingredients are already out there in space. Extra-terrestrial sugars [295], amino-acids [184–189,296], nucleobases, and water [260,297–305] have all been detected. Labs have demonstrated how these biomolecules may have originated here and across the Universe.

Oró successfully synthesized the nucleotide adenine (A) from hydrogen cyanide (HCN), ammonia ($NH_3$), and water ($H_2O$) [306]. The Miller–Urey experiments proved that amino-acids ($H_2NCHRCOOH$ in most cases) can be recreated from water ($H_2O$), methane ($CH_4$), ammonia ($NH_3$), and hydrogen ($H_2$) when combined with a dash of electric sparks [307]. Miller later left hydrogen cyanide (HCN), ammonia ($NH_3$), and water ($H_2O$) vials in freezers, water baths and other places throughout his lab for decades. After cracking the vial open, Miller and his team found seven types of amino acids and eleven nucleobases [308,309].

Minimal genomes are important to the study of cosmical convergent evolution because the most conserved genes here on Earth, the "essence of life" contained within our lipid bubbles may be encountered again on another planet, moon or other astronomical body in a surprisingly similar fashion despite the distance. The most conserved genes like the 16S can reveal priority in a sea of genes for astrobiology. Heading to the bridge poised between astrobiology and molecular biology, we must continue exploring how complex molecules like the 16S rRNA evolved because unlocking our past can help us predict our futures and transform it into a better tomorrow.

# CONCLUSION & RECOMMENDATIONS

Minimal genomes conserve phylogenies. Phyla were together in clades. Trees appeared similar to literature trees. The Tenericutes, however, don't conserve phylogenies, and need a new taxonomy. There are plenty paraphyletic cases. These relationships can lead to confusion, especially with those not familiar with the details of each species. Simplicity should trump complexity. The type species of *Mycoplasma* should be reassigned. The present *Mycoplasma* type species, *Mycoplasma mycoides*, has more in common with order Entomoplasmatales than its presently assigned order, Mycoplasmatales. There are fewer *Mycoplasma* spp. in the *M.mycoides* clade, hence less renaming is needed if this route is opted.

Examining minimal genomes, we observed that smaller genomes were AT biased. Interestingly, if we observe the smallest genome within each phyla, the range in GC content is similar to that seen in non-minimal genomes (25-75%). While the Tenericutes have small genomes and are AT rich, we believe that their altered codon table in some clades is the product of a lost release fac tor, and not from selective pressure. *Acholeplasma* spp. are Tenericutes that are AT rich, yet they have release factor 2 and use the universal codon table. Again, this highlights that the Tenericutes are not genetically cohesive, and that AT content did not drive UGA to become a tryptophan codon. Genome size and AT content are directly proportional in domain Bacteria. Smaller bacterial genomes had higher AT contents. The pattern was observed in minimal genomes and the Tenericutes when examined separately and together.

While it has been suggested that the Mycoplasmatales and Entomoplasmatales had codon UGA (a stop codon in the universal code) code for tryptophan (UGG in the universal code) because they were AT rich, we hypothesize that due to the extreme genome reduction, release factor 2 (recognizes UGA and UAA as stop codons) was lost (*RF2*) along with having a cell wall. Then, since UGG and UGA differ only on the wobble position considered, some Tenericutes made UGA a tryptophan codon.Based on our small sample size of Eukaryotes, nucleotide bias does not seem to hold true in domain Eukarya. If it were to be found that eukaryotes have a complement nucleotide pair bias, then it should be reset at each domain.

Caution must be exercised when drawing conclusions from phylogenies based on few samples (n<20). There were large discrepancies between the algorithms for the select Gram-positive, Gram-negative, and Tenericutes we studied (n=19). Of all algorithms tested, T-Coffee combined with the 16S produced a phylogenetic tree that resembled most closely our proteome analysis.

## Recommendations for Future Studies

The core and essential gene set remains to be identified in minimal bacterial genomes, and Tenericutes clades. These genes may be used in medicine as future antibiotic targets, in taxonomy and ecology as supplements to the 16S rRNA phylogeny, and in astrobiology for drawing cosmical convergent evolution and LUCA models. Given that the core genes in all Bacteria will probably determine the sequence of the last common ancestor of Bacteria, core genes in Archaea and Eukarya should be determined, and then compared with prokaryotes to estimate the LUCA genome. These comparisons could also help identify the root of the tree of life.

With further analysis, uniting features among the different Tenericutes clades may be found with the data produced as a result of this study. For example, we found a clade of *Mycoplasmas* that were mostly found in the blood of mammals. Similarly, other common characteristics may be found with more analysis. The neighbor-joining data should also be compared to maximum parsimony methods. Phylogenies should be compared with statistical methods, like the Kishino-Hasegawa test.

# MATERIALS AND METHODS

## The Tenericutes

### *Database*

A database was created in Microsoft Excel 2013 (PC) and 2016 (Mac) for all Tenericutes species, including candidate species. We obtained the order, family, genus, species, type strain name, ATCC strain name, NCBI taxonomy identification, and most 16S rRNA sequence accession numbers, all under phylum Tenericutes, from the List of Prokaryotic Names with Standing Nomenclature and the Bergey's Manual [202,310]. We also recorded the organism code (e.g., mge for *Mycoplasma genitalium*) from the Kyoto Encyclopedia of Genes and Genomes (KEGG), [311–313] the accession number from GenBank [314] or the National Center for Biotechnology Information (NCBI), and the 16S rRNA sequences, genome length, GC content, protein, and gene count from NCBI [315]. The 16S rRNA accession numbers and sequences for *Candidatus* Phytoplasma japonicum, *Candidatus* Phytoplasma solani, and *Mycoplasma ferirumnatoris* were obtained from Silva [316–319]. We obtained where the organism was isolated from, including the target organ or system, host's common and scientific name, family, order and class, and country where sample was acquired, from the literature referenced in the *Bergey's Manual*, and the American Type Culture Collection (ATCC). For compact display purposes, country names were shortened to two-letter codes using International Organization for Standardization (ISO) 3166-1 alpha-2 [320]. Countries were then classified by continent, that is, Africa, The Americas, Asia, Europe or Oceania. Countries was defined as Member States of the United Nations (UN). Taking note of the unique geography of Turkey, species west of the Bosphorous were classified as European, and those East of the Bosphorous were classified as Asian. Hosts were classified as arthropods (excluding crustaceans), birds, environmental samples, fish and crustaceans, mammals, plants or reptiles.

### *Proteome Analysis*

To determine which alignment relates more closely with the proteome, we sought to find proteome "homologies" using GeneCore 3.5 [321–323]. These are computed homologs and not actual homologs. The number of computed homologs were recorded. A matrix that doubled as a heatmap was made based on the number of homologs. Numbers were then converted to percentages based on genome size. The highest percentage from each pair was used for determining rankings of how similar a reference sequence was to the rest of the 18 sequences. Rankings were recorded as a matrix. The matrix was converted to a cladogram tree using the T-Rex web server [324].

Furthermore, whole genome alignments of Mycoplasmatales associated with human disease were done using Mauve from the Darling Laboratory [325]. We compared the results to pathogenic Enterobacteriales to confirm that paraphyletic results produce a different visual to the relatively homogeneous Enterobacteriales.

*16S rRNA Alignments*

The 16S rRNA sequences aforementioned were initially aligned using MUSCLE drive5 [167,168] running under the NCBI Genome Workbench. To ensure that only the 16S rRNA gene was present, sequences were trimmed using consensus sequences. The universal forward primer 27F and universal reverse primer 1492R were used as reference consensus sequences. Given the variety in multiple sequence alignment (MSA) methods, and to ensure their validity, we selected representative 16S rRNA sequences from Gram-positives, Gram-negatives and Tenericutes species to examine which MSA produced phylogenetic trees that reflected proteome analysis and shared characteristics (e.g., anaerobic) based on the literature. Due to the large number of sequences (ranging from 131-406 sequences), and to prevent bias, alignments were not modified manually.

We aligned these representative organisms of the Tenericutes, Gram-positives and Gram-negatives with EMBL's Clustal Omega, K-Align MAFFT, MUSCLE and T-Coffee as well as MAFFT of the Computational Biology Research Center (CBRC) using neighbor joining and minimum linkage for plotting trees.

While T-Coffee offered optimal results when n=19, this MSA is limited to a hundred input sequences. We opted for MAFFT for the Tenericutes (n=275) tree because it offered consistency. The parameters were set as 1.53 gap open penalty, 0.123 gap extension penalty, 100 tree rebuilding number, 100 iterations ("Maxiterate"), with global pair fast Fourier transforms. Trees were computed with Neighbor-joining [176] without distance corrections. Trees were drawn using the Interactive Tree of Life (iTOL) [174].

## Minimal Genomes

A second database was created in Google Sheets for the most minimal genomes recorded in NCBI across all phyla in the bacterial domain. Genome size, whole genome accessions, 16S rRNA sequences, protein counts and GC contents were all obtained from NCBI. The 16S rRNA alignments and tree drawing methods were the same as listed for the Tenericutes.

## 16S rRNA Studies

We added representative organisms from the Eukaryotes and Archaea to have something to compare with. Noting that the clade remain isolated from everything else, we examined the DNA sequence. We discovered that online databases did not reverse-complement their DNA sequences to align with other organisms, and would like to warn others to be cautious. A simple way to note if the sequences are flipped would be to use virtual universal primers, as we did. We would like to suggest sequence 5'-GTGYCAGCMGCCGCGGTAA-3' as a direction identifier. When tested in Silva, said sequence produced 587,252 matches, and covered 94% of Bacteria, 93% of Archaea and 91.4% of Eukaryotes. By far, this sequence produced the most by any universal primer found on the literature for a complete list of virtual "universal" primers tested) (**Table 5**).

Using the alignment files, we proceeded to draw histograms and heatmaps of the most frequent base. Bases not found in *E. coli* were discarded. We also used WebLogo from Berkeley to plot the bases [326,327]. The actual location of the bases in reference to *E. coli* was not computed.

### Codon Usage

Codon usage was obtained using the Codon Usage Database [328]. Values were recorded as parts-per-thousand. Codons were converted to amino-acids, and the fractions were summed to obtain amino-acid frequencies. NCBI Genome was used for determining codon tables.

### Statistics

Statistical analysis was computed in GraphPad Prism 7.03, IBM SPSS Statistics version 23, and Microsoft Excel 2016. Results list which software was used where. Descriptive statistics, correlations were conducted in SPSS. Pearson correlations were two-tailed, significance was set at $\alpha=5$.

# BIBLIOGRAPHY

1. Darwin C. On the Origin of Species by Means of Natural Selection, Or The Preservation of Favoured Races in the Struggle for Life. 1861;

2. Santarella-Mellwig R, Pruggnaller S, Roos N, Mattaj IW, Devos DP. Three-dimensional reconstruction of bacteria with a complex endomembrane system. PLoS Biol. 2013;11: e1001565;

3. Kubitschek HE. Cell volume increase in Escherichia coli after shifts to richer media. J Bacteriol. 1990;172: 94–101;

4. Nagao N, Tamaki K, Kuchiki T, Nagao M. A new gas dilution method for measuring body volume. Br J Sports Med. 1995;29: 134–8;

5. Pace NR. Problems with "Procaryote." J Bacteriol. 2009;191: 2008–10;

6. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A. 1990;87: 4576–9;

7. Shakespeare W. The Most Excellent and Lamentable Tragedy of Romeo and Juliet. The Oxford Shakespeare: Romeo and Juliet. 1599. pp.136–8;

8. Porter JR. Antony van Leeuwenhoek: tercentenary of his discovery of bacteria. Bacteriol Rev. 1976;40: 260–9;

9. Ford BJ. The clarity of images from early single-lens microscopes captured on video. Microsc Anal . 2011;25: 15–7;

10. Ford BJ. From Dilettante to Diligent Experimenter, a Reappraisal of Leeuwenhoek as microscopist and investigator. Biology History. 1992;5;

11. van Zuylen J. The Microscopes of Antoni Van Leeuwenhoek. 1981;

12. Fred EB. Antony van Leeuwenhoek: On the Three-hundredth Anniversary of his Birth. J Bacteriol. 1933;25: iv.2–18;

13. Khattab ADS. Dances with microscopes: Antoni van leeuwenhoek (1632-1723). Cytopathology. 1995;6: 215–8;

14. Dobell C, Van Leeuwenhoek A. Antony Van Leeuwenhoek and His Little Animals; Being Some Account of the Father of Protozoology and Bacteriology and His Multifarious Discoveries in These Disciplines; Andesite Press; 2015;

15. Anderson D. Still going strong: Leeuwenhoek at eighty. Antonie Van Leeuwenhoek. 2014;106: 3–26;

16. Gest H. The discovery of microorganisms by Robert Hooke and Antoni Van Leeuwenhoek, fellows of the Royal Society. Notes Rec R Soc Lond. 2004;58: 187–201;

17. James J. Van Leeuwenhoeks ontdekking van bacteriën: een blik te ver vooruit. Ned Tijdschr Geneeskd. 2004;148: 2590–4;

18. Bloch RS. Healers and Achievers: Physicians Who Excelled in Other Fields and the Times in Which

They Lived. Xlibris Corporation; 2012;

19. Rutgers J. Linnaeus in the Netherlands. TijdSchrift voor Skandinavistiek. 2008;29. http://jdbsc.rug.nl/index.php/tvs/article/download/10743/8314

20. Stöver DJH. The Life of Sir Charles Linnæus ...: To which is Added, a Copious List of His Works, and a Biographical Sketch of the Life of His Son. 1794;

21. Greenwood T. Eminent naturalists. 1886;

22. Lindeboom GA. Boerhaave and His Time. Brill Archive; 1970;

23. Anderson MJ. Carl Linnaeus: Father of Classification. Enslow Publishers, Inc.; 2009;

24. Linné C von, von Linné C, Salvii. L. Caroli Linnaei...Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. 1758;

25. Baker H. Employment for the Microscope. In Two Parts: Likewise a Description of the Microscope Used in These Experiments .... An examination of salts and saline substances, their amazing configurations and crystals, as formed under the eye of the observer .... An account of various animalcules never before described . 1753;

26. Sapp J. Two faces of the prokaryote concept. Int Microbiol. 2006;9: 163–172;

27. Haeckel E. Generelle Morphologie der Organismen: allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie. Allgemeine Anatomie der Organismen. 1866;

28. Whittaker RH. New Concepts of Kingdoms of Organisms. Science. 1969;163: 150–160;

29. Drews G. Ferdinand Cohn, a Founder of Modern Microbiology. ASM News. 1999;65: 9;

30. Charlton Bastian H. On the Great Importance from the Point of View of Medical Science of the Proof that Bacteria and their Allies are Capable of Arising De Novo. Lancet. 1903;162: 1220–4;

31. Drews G. The roots of microbiology and the influence of Ferdinand Cohn on microbiology of the 19th century. FEMS Microbiol Rev. 2000;24: 225–249;

32. Young KD. The selective value of bacterial shape. Microbiol Mol Biol Rev. 2006;70: 660–703;

33. Young KD. Bacterial morphology: why have different shapes? Curr Opin Microbiol. 2007;10: 596–600;

34. Alturkistani HA, Tashkandi FM, Mohammedsaleh ZM. Histological Stains: A Literature Review and Case Study. Glob J Health Sci. 2015;8: 72;

35. Beveridge TJ. Use of the Gram stain in microbiology. Biotechnic and Histochemistry. 2001;76: 111–8;

36. Titford M. The long history of hematoxylin. Biotech Histochem. 2005;80: 73–8;

37. Lewis FT. The introduction of biological stains: Employment of saffron by Vieussens and Leeuwenhoek. Anat Rec. 1942;83: 229–253;

38. Ueber die isolirte Färbung der Schizomyceten in Schnitt- und Trockenpräparaten von Dr. Gram,

Kopenhagen. — Fortschritte der Medicin 1884 No. 6. Ref. Dr. Becker. DMW - Deutsche Medizinische Wochenschrift. 1884;10: 234–5;

39. Shostak S. Histology's Nomenclature: Past, Present and Future. Biological Systems: Open Access. 2013;02. doi:10.4172/2329-6577.1000122

40. Evans AS. Causation and disease: the Henle-Koch postulates revisited. Yale J Biol Med. 1976;49: 175–95;

41. Grimes DJ, Jay Grimes D. Koch's Postulates — Then and Now. Microbe Magazine. 2006;1: 223–8;

42. Codell Carter K. Koch's postulates in relation to the work of Jacob Henle and Edwin Klebs. Med Hist. 1985;29: 353–374;

43. Hitchens AP, Leikind MC. The Introduction of Agar-agar into Bacteriology. J Bacteriol. 1939;37: 485–493;

44. Ullmann A. Pasteur–Koch: Distinctive Ways of Thinking about Infectious Diseases. Microbe Wash DC. 2007;2. http://antimicrobe.org/h04c.files/history/Microbe%202007%20Pasteur-Koch.pdf

45. Copeland HF. The Kingdoms of Organisms. Q Rev Biol. 1938;13: 383–420;

46. Copeland EB. What is a Plant? Science. 1927;65: 388–390;

47. Stanier RY, Niel CB. The concept of a bacterium. Archiv für Mikrobiologie. 1962;42: 17–35;

48. Sapp J. The prokaryote-eukaryote dichotomy: meanings and mythology. Microbiol Mol Biol Rev. 2005;69: 292–305;

49. Wilson EB. The cell in development and inheritance / by Edmund B. Wilson. 1896;

50. Abbe E. Beiträge zur Theorie des Mikroskops und der mikroskopischen Wahrnehmung. Archiv für Mikroskopische Anatomie. 1873;9: 413–8;

51. Milo R, Phillips R. Cell Biology by the Numbers. Garland Science; 2015;

52. Riley M, Abe T, Arnaud MB, Berlyn MKB, Blattner FR, Chaudhuri RR, et al. Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. Nucleic Acids Res. 2006;34: 1–9;

53. Freundlich MM. Origin of the Electron Microscope. Science. 1963;142: 185–8;

54. Stanier RY, Van Niel CB. The Main Outlines of Bacterial Classification. J Bacteriol. 1941;42: 437–466;

55. Hagen JB. Five Kingdoms, More or Less: Robert Whittaker and the Broad Classification of Organisms. Bioscience. 2012;62: 67–74;

56. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A. 1977;74: 5088–90;

57. Pace NR, Sapp J, Goldenfeld N. Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. Proc Natl Acad Sci U S A. 2012;109: 1011–8;

58. Cavalier-Smith T. A revised six-kingdom system of life. Biol Rev Camb Philos Soc. 1998;73: 203–66;

59. Cavalier-Smith T. Eukaryote kingdoms: seven or nine? Biosystems. 1981;14: 461–81;

60. Cavalier-Smith T. Only six kingdoms of life. Proc Biol Sci. 2004;271: 1251–62;

61. Ruggiero MA, Gordon DP, Orrell TM, Bailly N, Bourgoin T, Brusca RC, et al. A higher level classification of all living organisms. PLoS One. 2015;10: e0119248;

62. Blackwell WH. Is It Kingdoms or Domains? Confusion & Solutions. Am Biol Teach. 2004;66: 268;

63. Embley TM, Martin Embley T, Martin W. Eukaryotic evolution, changes and challenges. Nature. 2006;440: 623–30;

64. Adl SM, Simpson AGB, Lane CE, Lukeš J, Bass D, Bowser SS, et al. The revised classification of eukaryotes. J Eukaryot Microbiol. 2012;59: 429–93;

65. Yin H, Marshall D. Microfluidics for single cell analysis. Curr Opin Biotechnol. 2012;23: 110–9;

66. Kalisky T, Quake SR. Single-cell genomics. Nat Methods. 2011;8: 311–4;

67. Whitesides GM. The origins and the future of microfluidics. Nature. 2006;442: 368–73;

68. Stewart EJ. Growing unculturable bacteria. J Bacteriol. 2012;194: 4151–60;

69. Blainey PC. The future is now: single-cell genomics of bacteria and archaea. FEMS Microbiol Rev. 2013;37: 407–27;

70. Haliburton JR, Shao W, Deutschbauer A, Arkin A, Abate AR. Genetic interaction mapping with microfluidic-based single cell sequencing. PLoS One. 2017;12: e0171302;

71. Pande S, Kost C. Bacterial Unculturability and the Formation of Intercellular Metabolic Networks. Trends Microbiol. 2017;25: 349–61;

72. Rappé MS, Giovannoni SJ. The Uncultured Microbial Majority. Annu Rev Microbiol. 2003;57: 369–94;

73. Schloss PD, Handelsman J. Status of the microbial census. Microbiol Mol Biol Rev. 2004;68: 686–91;

74. Amann R, Ludwig W, Schleifer K-H. Identification and in situ detection of individual bacterial cells. FEMS Microbiol Lett. 1992;100: 45–50;

75. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol. 2014;12: 635–45;

76. Shakespeare W. The Tragedy of Hamlet, Prince of Denmark. The Oxford Shakespeare: Hamlet. 1603. pp. 138–40;

77. Arbib MA. From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. Behav Brain Sci. 2005;28: 105–24; discussion 125–67;

78. Matsuzawa T. Colour naming and classification in a chimpanzee (Pan troglodytes). J Hum Evol. 1985;14: 283–91;

79. Lonsdorf EV, Ross SR, Matsuzawa T, Goodall J. The Mind of the Chimpanzee. 2010;

80. Trabant J. New Essays on the Origin of Language. Walter de Gruyter; 2001;

81. Wansbrough H. The New Jerusalem Bible. Doubleday; 1985;

82. Buckman TR. Bibliography & Natural History: Essays Presented at a Conference Convened in June 1964. Lawrence : University of Kansas Libraries; 1966;

83. von Linné C. Systema naturae: per regna tria natura, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. 1766;

84. Winstor MP. Non-essentialist methods in pre-Darwinian taxonomy. Biology and Philosophy. 2003;18: 387–400;

85. Müller-Wille S. Collection and collation: theory and practice of Linnaean botany. Stud Hist Philos Biol Biomed Sci. 2007;38: 541–62;

86. Winsor MP. Linnaeus' Biology was not Existentialist. Ann Mo Bot Gard. 2006;93: 2–7;

87. Wyhe J v., v. Wyhe J. The complete work of Charles Darwin online. Notes Rec R Soc Lond. 2006;60: 87–9;

88. Darwin C, Burkhardt F, Smith S. The Correspondence of Charles Darwin: 1856-1857. Cambridge University Press; 1985;

89. Sanger F, Tuppy H. The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem J. 1951;49: 463–81;

90. Sanger F, Tuppy H. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem J. 1951;49: 481–90;

91. Sanger F, Thompson EOP. The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. Biochem J. 1953;53: 353–66;

92. Sanger F, Thompson EOP. The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem J. 1953;53: 366–74;

93. Stretton AOW. The first sequence. Fred Sanger and insulin. Genetics. 2002;162: 527–32;

94. Sanger F. Chemistry of Insulin: Determination of the structure of insulin opens the way to greater understanding of life processes. Science. 1959;129: 1340–4;

95. Bkownlee GG, Sanger F, Barrell BG. The sequence of 5S ribosomal ribonucleic acid. J Mol Biol. 1968;34: 379–412;

96. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74: 5463–67;

97. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94: 441–8;

98. Sanger F. Determination of nucleotide sequences in DNA. Science. 1981;214: 1205–10;

99. Moody G. Digital Code of Life: How Bioinformatics is Revolutionizing Science, Medicine, and Business. John Wiley & Sons; 2004;

100. Hagen JB. The origin and early reception of sequence databases. Methods Mol Biol. 2011;696: 61–77;

101. Dayhoff MO, Eck RV. Atlas of Protein Sequence and Structure, 1966. 1966;

102. IUPAC-IUB Commission on Biochemical Nomenclature. A one-letter notation for amino acid sequences. Tentative rules. J Biol Chem. 1968;243: 3557–9;

103. Dayhoff MO. Atlas of Protein Sequence and Structure 1967-68: Editors , Margaret O. Dayhoff, Richard V. Eck. 1968;

104. Dayhoff MO, Eck RV, Lippincott ER, Sagan C. Venus: Atmospheric Evolution. Science. 1967;155: 556–8;

105. Lippincott ER, Eck RV, Dayhoff MO, Sagan C. Thermodynamic Equilibria in Planetary Atmospheres. Astrophys J. 1967;147: 753;

106. Dayhoff MO, Scientific American. Computer Analysis of Protein Evolution. 1969;

107. Dayhoff MO, Eck RV. Paleobiochemistry. Organic Geochemistry. 1969. pp. 196–212;

108. Eck RV, Dayhoff MO. Evolution of the structure of ferredoxin based on living relics of primitive amino Acid sequences. Science. 1966;152: 363–6;

109. Weisburg WG, Barns SM, Pelletier DA, Lane DJ. 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol. 1991;173: 697–703;

110. Janda JM, Abbott SL. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. J Clin Microbiol. 2007;45: 2761–4;

111. Baker GC, Smith JJ, Cowan DA. Review and re-analysis of domain-specific 16S primers. J Microbiol Methods. 2003;55: 541–55;

112. Turner S, Pryer KM, Miao VP, Palmer JD. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. J Eukaryot Microbiol. 1999;46: 327–38;

113. Rudi K, Skulberg OM, Larsen F, Jakobsen KS. Strain characterization and classification of oxyphotobacteria in clone cultures on the basis of 16S rRNA sequences from the variable regions V6, V7, and V8. Appl Environ Microbiol. 1997;63: 2593–9;

114. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 2013;41: e1;

115. Větrovský T, Baldrian P. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. PLoS One. 2013;8: e57923;

116. Patel JB. 16S rRNA Gene Sequencing for Bacterial Pathogen Identification in the Clinical Laboratory. Mol Diagn. 2001;6: 313–21;

117. Gogarten JP, Peter Gogarten J, Townsend JP. Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol. 2005;3: 679–87;

118. Schouls LM, Schot CS, Jacobs JA. Horizontal transfer of segments of the 16S rRNA genes between species of the Streptococcus anginosus group. J Bacteriol. 2003;185: 7241–6;

119. Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, et al. Diversity of 16S rRNA genes within individual prokaryotic genomes. Appl Environ Microbiol. 2010;76: 3886–97;

120. Hugenholtz P, Goebel BM, Pace NR. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol. 1998;180: 4765–74;

121. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proceedings of the National Academy of Sciences. 1985;82: 6955–9;

122. Lazarevic V, Whiteson K, Huse S, Hernandez D, Farinelli L, Osterås M, et al. Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. J Microbiol Methods. 2009;79: 266–71;

123. Jones RT, Robeson MS, Lauber CL, Hamady M, Knight R, Fierer N. A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. ISME J. 2009;3: 442–53;

124. Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, et al. The phylogeny of prokaryotes. Science. 1980;209: 457–63;

125. Parmeggiani A, Nissen P. Elongation factor Tu-targeted antibiotics: four different structures, two mechanisms of action. FEBS Lett. 2006;580: 4576–81;

126. Brisse S, Verhoef J. Phylogenetic diversity of Klebsiella pneumoniae and Klebsiella oxytoca clinical isolates revealed by randomly amplified polymorphic DNA, gyrA and parC genes sequencing and automated ribotyping. Int J Syst Evol Microbiol. 2001;51: 915–24;

127. Chopra S, Reader J. tRNAs as antibiotic targets. Int J Mol Sci. 2014;16: 321–349;

128. Lewis K. Platforms for antibiotic discovery. Nat Rev Drug Discov. 2013;12: 371–87;

129. Vos M, Quince C, Pijl AS, de Hollander M, Kowalchuk GA. A comparison of rpoB and 16S rRNA as markers in pyrosequencing studies of bacterial diversity. PLoS One. 2012;7: e30600;

130. Adékambi T, Drancourt M, Raoult D. The rpoB gene as a tool for clinical microbiologists. Trends Microbiol. 2009;17: 37–45;

131. Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol. 2007;73: 278–88;

132. Dahllöf I, Baillie H, Kjelleberg S. rpoB-based microbial community analysis avoids limitations inherent in 16S rRNA gene intraspecies heterogeneity. Appl Environ Microbiol. 2000;66: 3376–80;

133. Peixoto RS, da Costa Coutinho HL, Rumjanek NG, Macrae A, Rosado AS. Use of rpoB and 16S rRNA genes to analyse bacterial diversity of a tropical soil using PCR and DGGE. Lett Appl Microbiol. 2002;35: 316–20;

134. Brosius J, Palmer ML, Kennedy PJ, Noller HF. Complete nucleotide sequence of a 16S ribosomal RNA gene from Escherichia coli. Proceedings of the National Academy of Sciences. 1978;75: 4801–5;

135. Woese C. The universal ancestor. Proc Natl Acad Sci U S A. 1998;95: 6854–9;

136. Lun ATL, Swaminathan K, Wong JWH, Downard KM. Mass Trees: A New Phylogenetic Approach and Algorithm to Chart Evolutionary History with Mass Spectrometry. Anal Chem. 2013;85: 5475–82;

137. Hanage WP, Fraser C, Spratt BG. Fuzzy species among recombinogenic bacteria. BMC Biol. 2005;3: 6;

138. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. The bacterial species challenge: making sense of genetic and ecological diversity. Science. 2009;323: 741–6;

139. Fraser C, Hanage WP, Spratt BG. Recombination and the nature of bacterial speciation. Science. 2007;315: 476–80;

140. Wolf YI, Rogozin IB, Grishin NV, Koonin EV. Genome trees and the tree of life. Trends Genet. 2002;18: 472–9;

141. Woese CR. Interpreting the universal phylogenetic tree. Proc Natl Acad Sci U S A. 2000;97: 8392–6;

142. Darmon E, Leach DRF. Bacterial Genome Instability. Microbiol Mol Biol Rev. 2014;78: 1–39;

143. Cohan FM. What are Bacterial Species? Annu Rev Microbiol. 2002;56: 457–87;

144. Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci. 2006;361: 1929–40;

145. Caro-Quintero A, Konstantinidis KT. Bacterial species may exist, metagenomics reveal. Environ Microbiol. 2012;14: 347–55;

146. Polz MF, Alm EJ, Hanage WP. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet. 2013;29: 170–5;

147. Franklin LR. Bacteria, Sex, and Systematics*. Philos Sci. 2007;74: 69–95;

148. Wayne LG, Moore WEC, Stackebrandt E, Kandler O, Colwell RR, Krichevsky MI, et al. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. Int J Syst Evol Microbiol. 1987;37: 463–4;

149. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009;106: 19126–31;

150. Petersen RH, Hughes KW. Species and Speciation in Mushrooms. Bioscience. 1999;49: 440;

151. Rosselló-Mora R, Amann R. The species concept for prokaryotes. FEMS Microbiol Rev. 2001;25: 39–67;

152. De Queiroz K. Species Concepts and Species Delimitation. Syst Biol. 2007;56: 879–86;

153. Wiley EO, Lieberman BS. Phylogenetics: Theory and Practice of Phylogenetic Systematics. John Wiley & Sons; 2011;

154. Mayr E. What Is a Species, and What Is Not? Philos Sci. 1996;63: 262–77;

155. Hey J. The mind of the species problem. Trends Ecol Evol. 2001;16: 326–9;

156. Bock WJ. Nonvalidity of the "Phylogenetic Fallacy." Syst Zool. 1969;18: 111;

157. Claridge MF. Species: The Units of Biodiversity. Taylor & Francis US; 1997;

158. Chan JZ-M, Halachev MR, Loman NJ, Constantinidou C, Pallen MJ. Defining bacterial species in the genomic era: insights from the genus Acinetobacter. BMC Microbiol. 2012;12: 302;

159. Reller LB, Weinstein MP, Petti CA. Detection and Identification of Microorganisms by Gene Amplification and Sequencing. Clin Infect Dis. 2007;44: 1108–14;

160. Kusunoki S, Ezaki T. Proposal of Mycobacterium peregrinum sp. nov., nom. rev., and Elevation of Mycobacterium chelonae subsp. abscessus (Kubica et al.) to Species Status: Mycobacterium abscessus comb. nov. Int J Syst Bacteriol. 1992;42: 240–5;

161. Auch AF, von Jan M, Klenk H-P, Göker M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci. 2010;2: 117–34;

162. Meier-Kolthoff JP, Klenk H-P, Göker M. Taxonomic use of DNA G+C content and DNA-DNA hybridization in the genomic age. Int J Syst Evol Microbiol. 2014;64: 352–6;

163. Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, et al. A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res. 2010;38: W695–9;

164. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, et al. Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res. 2013;41: W597–600;

165. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 2017.

166. Kuraku S, Zmasek CM, Nishimura O, Katoh K. aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity. Nucleic Acids Res. 2013;41: W22–8;

167. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32: 1792–7;

168. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5: 113;

169. Feng DF, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol. 1987;25: 351–60;

170. Thompson JD, Higgins DG, Gibson TJ. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22: 4673–80;

171. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000;302: 205–17;

172. Katoh K, Misawa K, Kuma K-I, Miyata T. MAFFT: a novel method for rapid multiple sequence

alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30: 3059–66;

173. Stevens PF, Augier A. Augustin Augier's "Arbre Botanique" (1801), a Remarkable Early Botanical Representation of the Natural System. Taxon. 1983;32: 203;

174. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44: W242–5;

175. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. 2016;1: 16048;

176. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4: 406–25;

177. Drake JW. A constant rate of spontaneous mutation in DNA-based microbes. Proc Natl Acad Sci U S A. 1991;88: 7160–4;

178. Shadrin AA, Parkhomchuk DV. Drake's rule as a consequence of approaching channel capacity. Naturwissenschaften. 2014;101: 939–54;

179. Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. Drift-barrier hypothesis and mutation-rate evolution. Proc Natl Acad Sci U S A. 2012;109: 18488–92;

180. Drake JW. Rates of Spontaneous Mutation: Insights Gained Over the Last Half Century. Genetics, Evolution and Radiation. 2016. pp. 77–84;

181. Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, et al. Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet. 2016;17: 704–14;

182. Lynch M. Evolution of the mutation rate. Trends Genet. 2010;26: 345–52;

183. Flores Martinez CL. SETI in the light of cosmic convergent evolution. Acta Astronaut. 2014;104: 341–9;

184. Elsila JE, Glavin DP, Dworkin JP. Cometary glycine detected in samples returned by Stardust. Meteorit Planet Sci. 2009;44: 1323–30;

185. Cronin JR, Moore CB. Amino Acid analyses of the murchison, murray, and allende carbonaceous chondrites. Science. 1971;172: 1327–9;

186. Kvenvolden K, Lawless J, Pering K, Peterson E, Flores J, Ponnamperuma C, et al. Evidence for extraterrestrial amino-acids and hydrocarbons in the Murchison meteorite. Nature. 1970;228: 923–6;

187. Cronin JR, Pizzarello S. Amino acids in meteorites. Adv Space Res. 1983;3: 5–18;

188. Elsila JE, Callahan MP, Dworkin JP, Glavin DP, McLain HL, Noble SK, et al. The origin of amino acids in lunar regolith samples. Geochim Cosmochim Acta. 2016;172: 357–69;

189. Pizzarello S, Shock E. Carbonaceous Chondrite Meteorites: the Chronicle of a Potential Evolutionary Path between Stars and Life. Orig Life Evol Biosph. 2017.

190. Belloche A, Garrod RT, Müller HSP, Menten KM, Comito C, Schilke P. Increased complexity in interstellar chemistry: detection and chemical modeling of ethyl formate and n-propyl cyanide in

Sagittarius B2(N). Astron Astrophys Suppl Ser. 2009;499: 215–32;

191. Gottlieb CA, Ball JA, Gottlieb EW, Dickinson DF. Interstellar methyl alcohol. Astrophys J. 1979;227: 422;

192. Nummelin A, Dickens JE, Bergman P, Hjalmarson A, Irvine WM, Ikeda M, et al. Abundances of ethylene oxide and acetaldehyde in hot molecular cloud cores. Astron Astrophys. 1998;337: 275–286;

193. Herbst E, van Dishoeck EF. Complex Organic Interstellar Molecules. Annu Rev Astron Astrophys. 2009;47: 427–80;

194. Nocard E. RER. Le microbe de la peripneumonie. Ann Inst Pasteur . 1898;12: 240–62;

195. Tully JG. The Emmy Klieneberger-Nobel Award lecture. Reflections on recovery of some fastidious mollicutes with implications of the changing host patterns of these organisms. Yale J Biol Med. 1983;56: 799–813;

196. Ford D. Pleuropneumonia-like organisms (PPLO): Mycoplasmataceae. E. Klieneberger-Nobel. London and New York, Academic Press, 1962, 157, XL illus., 40 s. net. Arthritis & Rheumatism. 1962;5: 437;

197. Dienes L, Edsall G. Observations on the L-Organism of Klieneberger. Exp Biol Med. 1937;36: 740–4;

198. Eaton MD, Meiklejohn G, van Herick W. Studies on the Etiology of Primary Atypical Pneumonia: A Filterable Agent Transmissible to Cotton Rats, Hamsters, and Chick Embryos. J Exp Med. 1944;79: 649–68;

199. Chanock RM, Rifkind D, Kravetz HM, Kinght V, Johnson KM. Respiratory disease in volunteers infected with Eaton agent: a preliminary report. Proc Natl Acad Sci U S A. 1961;47: 887–90;

200. Marmion BP, Goodburn GM. Effect of an organic gold salt on Eaton's primary atypical pneumonia agent and other observations. Nature. 1961;189: 247–8;

201. Chanock RM, Dienes L, Eaton MD, ff. Edward DG, Freundt EA, Hayflick L, et al. Mycoplasma pneumoniae: Proposed Nomenclature for Atypical Pneumonia Organism (Eaton Agent). Science. 1963;140: 662;

202. Brown DR. Tenericutes. Bergey's Manual of Systematics of Archaea and Bacteria. 2015. 1–2;

203. Koonin EV, Novozhilov AS. Origin and Evolution of the Universal Genetic Code. Annu Rev Genet. 2017.

204. Glass JI. Synthetic genomics and the construction of a synthetic bacterial cell. Perspect Biol Med. 2012;55: 473–89;

205. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, et al. Essential genes of a minimal bacterium. Proceedings of the National Academy of Sciences. 2006;103: 425–30;

206. Halbedel S, Stulke J. Tools for the genetic analysis of Mycoplasma. Int J Med Microbiol. 2007;297: 37–44;

207. Reuß DR, Commichau FM, Gundlach J, Zhu B, Stülke J. The Blueprint of a Minimal Cell: MiniBacillus. Microbiol Mol Biol Rev. 2016;80: 955–87;

208. Clarke AC. Profiles of the Future: Unabridged. 1962;

209. Lytsy B, Sandegren L, Tano E, Torell E, Andersson DI, Melhus Å. The first major extended-spectrum β-lactamase outbreak in Scandinavia was caused by clonal spread of a multiresistant Klebsiella pneumoniae producing CTX-M-15. APMIS. 2008;116: 302–8;

210. Cantón R, González-Alba JM, Galán JC. CTX-M Enzymes: Origin and Diffusion. Front Microbiol. 2012;3: 110;

211. Wittgenstein L. Philosophische Untersuchungen. 1953;

212. Freedman DJ, Riesenhuber M, Poggio T, Miller EK. A comparison of primate prefrontal and inferior temporal cortices during visual categorization. J Neurosci. 2003;23: 5235–46;

213. Sigala N, Logothetis NK. Visual categorization shapes feature selectivity in the primate temporal cortex. Nature. 2002;415: 318–20;

214. Freedman DJ, Assad JA. Neuronal Mechanisms of Visual Categorization: An Abstract View on Decision Making. Annu Rev Neurosci. 2016;39: 129–47;

215. Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, et al. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol. 2014;10: e1003963;

216. Gil R, Silva FJ, Pereto J, Moya A. Determination of the Core of a Minimal Bacterial Gene Set. Microbiol Mol Biol Rev. 2004;68: 518–37;

217. Hutchison CA III. Global Transposon Mutagenesis and a Minimal Mycoplasma Genome. Science. 1999;286: 2165–9;

218. Juhas M, Eberl L, Glass JI. Essence of life: essential genes of minimal genomes. Trends Cell Biol. 2011;21: 562–8;

219. Charlebois RL, Doolittle WF. Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res. 2004;14: 2469–77;

220. Hutchison CA 3rd, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, et al. Design and synthesis of a minimal bacterial genome. Science. 2016;351: aad6253;

221. Browning GF, Citti C. Mollicutes: Molecular Biology and Pathogenesis. Horizon Scientific Press; 2014;

222. Gibson DG, Benders GA, Andrews-Pfannkoch C, Denisova EA, Baden-Tillson H, Zaveri J, et al. Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. Science. 2008;319: 1215–20;

223. Doudna JA, Charpentier E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science. 2014;346: 1258096;

224. Hershberg R, Petrov DA. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. 2010;6: e1001115;

225. Lassalle F, Périan S, Bataillon T, Nesme X, Duret L, Daubin V. GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS Genet. 2015;11: e1004941;

226. Rocha EPC, Danchin A. Base composition bias might result from competition for metabolic resources. Trends Genet. 2002;18: 291–94;

227. Bentley SD, Parkhill J. Comparative Genomic Structure of Prokaryotes. Annu Rev Genet. 2004;38: 771–91;

228. Oliver JL, Marín A. A relationship between GC content and coding-sequence length. J Mol Evol. 1996;43: 216–23;

229. Pozzoli U, Menozzi G, Fumagalli M, Cereda M, Comi GP, Cagliani R, et al. Both selective and neutral processes drive GC content evolution in the human genome. BMC Evol Biol. 2008;8: 99;

230. Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biol. 2006;4: e180;

231. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in Escherichia coli. Science. 2009;324: 255–8;

232. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2010;12: 32–42;

233. Lane N, Martin W. The energetics of genome complexity. Nature. 2010;467: 929–34;

234. Patthy L. Genome evolution and the evolution of exon-shuffling — a review. Gene. 1999;238: 103–14;

235. Yacoub E, Sirand-Pugnet P, Blanchard A, Ben Abdelmoumen Mardassi B. Genome Sequence of Mycoplasma meleagridis Type Strain 17529. Genome Announc. 2015;3;

236. Hershberg R, Petrov DA. Selection on codon bias. Annu Rev Genet. 2008;42: 287–299;

237. Hildebrand F, Meyer A, Eyre-Walker A. Evidence of Selection upon Genomic GC-Content in Bacteria. PLoS Genet. 2010;6: e1001107;

238. Moran NA. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proceedings of the National Academy of Sciences. 1996;93: 2873–78;

239. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol. 2011;10: 13–26;

240. Moran NA. Microbial minimalism: genome reduction in bacterial pathogens. Cell. 2002;108: 583–6;

241. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol. 2016;14: e1002533;

242. Sender R, Fuchs S, Milo R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. Cell. 2016;164: 337–40;

243. Thompson CC, Chimetto L, Edwards RA, Swings J, Stackebrandt E, Thompson FL. Microbial genomic taxonomy. BMC Genomics. 2013;14: 913;

244. Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, et al. Stop codon reassignments in the wild. Science. 2014;344: 909–13;

245. Osawa S, Jukes TH. Codon reassignment (codon capture) in evolution. J Mol Evol. 1989;28: 271–8;

246. Oba T, Andachi Y, Muto A, Osawa S. CGG: an unassigned or nonsense codon in Mycoplasma capricolum. Proc Natl Acad Sci U S A. 1991;88: 921–5;

247. Castresana J, Feldmaier-Fuchs G, Pääbo S. Codon reassignment and amino acid composition in hemichordate mitochondria. Proc Natl Acad Sci U S A. 1998;95: 3703;

248. McCutcheon JP, McDonald BR, Moran NA. Origin of an Alternative Genetic Code in the Extremely Small and GC–Rich Genome of a Bacterial Symbiont. PLoS Genet. 2009;5: e1000565;

249. Barrell BG, Bankier AT, Drouin J. A different genetic code in human mitochondria. Nature. 1979;282: 189–94;

250. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, et al. The minimal gene complement of Mycoplasma genitalium. Science. 1995;270: 397–403;

251. Korkmaz G, Holm M, Wiens T, Sanyal S. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. J Biol Chem. 2014;289: 30334–42;

252. Lajoie MJ, Rovner AJ, Goodman DB, Aerni H-R, Haimovich AD, Kuznetsov G, et al. Genomically recoded organisms expand biological functions. Science. 2013;342: 357–360;

253. Koonin EV. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol. 2003;1: 127–36;

254. Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, et al. The physiology and habitat of the last universal common ancestor. Nat Microbiol. 2016;1: 16116;

255. Wickramasinghe C. Origin of Life and Panspermia. Cellular Origin, Life in Extreme Habitats and Astrobiology. 2012. pp. 621–49;

256. Di Giulio M. Biological evidence against the panspermia theory. J Theor Biol. 2010;266: 569–572;

257. Woese CR. Bacterial evolution. Microbiol Rev. 1987;51: 221–71;

258. Bada JL. State-of-the-art instruments for detecting extraterrestrial life. Proc Natl Acad Sci U S A. 2001;98: 797–800;

259. Rummel JD. Preventing the Forward Contamination of Mars: Concerns, Questions, and Required Actions. 2006 IEEE Aerospace Conference. 2006. doi:10.1109/aero.2006.1655744

260. Hansen CJ, Esposito L, Stewart AIF, Colwell J, Hendrix A, Pryor W, et al. Enceladus' water vapor plume. Science. 2006;311: 1422–5;

261. Parkinson CD, Liang M-C, Yung YL, Kirschivnk JL. Habitability of enceladus: planetary conditions for life. Orig Life Evol Biosph. 2008;38: 355–69;

262. United Nations. Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies. US Department of State. 1967. https://www.state.gov/t/isn/5181.htm#treaty

263. Congress WS. COSPAR Planetary Protection Policy. 2002. p. 10. http://w.astro.berkeley.edu/~kalas/ethics/documents/environment/COSPAR%20Planetary%20Protection%20Policy.pdf

264. NASA. Office of Planetary Protection. In: NASA Office of Planetary Protection. 30 Oct 2017. https://planetaryprotection.nasa.gov/

265. ESA. Planetary Protection. In: Planetary Protection [Internet]. 27 Jul 2016. http://www.esa.int/Our_Activities/Space_Science/ExoMars/Planetary_protection

266. Orgel LE. The origin of life—a review of facts and speculations. Trends Biochem Sci. 1998;23: 491–5;

267. Robertson MP. Origins of Life: Emergence of the RNA World. Wiley Encyclopedia of Chemical Biology. 2008;

268. Robertson MP, Joyce GF. The Origins of the RNA World. Cold Spring Harb Perspect Biol. 2012;4: a003608–a003608;

269. Alberts B, Johnson A, Lewis J, Walter P, Raff M, Roberts K. Molecular Biology of the Cell 4th Edition: International Student Edition. Garland Science; 2002;

270. Tielens AGGM. 25 years of PAH hypothesis. EAS Publications Series. 2011;46: 3–10;

271. Segré D, Lancet D, Kedem O, Pilpel Y. Graded autocatalysis replication domain (GARD): kinetic analysis of self-replication in mutually catalytic sets. Orig Life Evol Biosph. 1998;28: 501–14;

272. Wächtershäuser G. The case for the chemoautotrophic origin of life in an iron-sulfur world. Orig Life Evol Biosph. 1990;20: 173–6;

273. Wächtershäuser G. Pyrite Formation, the First Energy Source for Life: a Hypothesis. Syst Appl Microbiol. 1988;10: 207–10;

274. Mulkidjanian AY, Galperin MY. On the origin of life in the zinc world. 2. Validation of the hypothesis on the photosynthesizing zinc sulfide edifices as cradles of life on Earth. Biol Direct. 2009;4: 27;

275. Martin W, Baross J, Kelley D, Russell MJ. Hydrothermal vents and the origin of life. Nat Rev Microbiol. 2008;

276. Muller AWJ. Thermosynthesis as energy source for the RNA World: A model for the bioenergetics of the origin of life. Biosystems. 2005;82: 93–102;

277. Cairns-Smith AG. Chemistry and the missing era of evolution. Chemistry. 2008;14: 3830–9;

278. Gold T. The deep, hot biosphere. Proc Natl Acad Sci U S A. 1992;89: 6045–9;

279. Segré D, Ben-Eli D, Deamer DW, Lancet D. The lipid world. Orig Life Evol Biosph. 2001;31: 119–

45;

280. Adam Z. Actinides and Life’s Origins. Astrobiology. 2007;7: 852–72;

281. Dartnell L. Did life begin on a radioactive beach? New Sci. 2008;197: 8;

282. Schneider ED, Kay JJ. Life as a manifestation of the second law of thermodynamics. Math Comput Model. 1994;19: 25–48;

283. Ostriker JP. New Light on Dark Matter. Science. 2003;300: 1909–13;

284. Spergel DN. The dark side of cosmology: dark matter and dark energy. Science. 2015;347: 1100–2;

285. Lodders K. Solar System Abundances and Condensation Temperatures of the Elements. Astrophys J. 2003;591: 1220–47;

286. David M. Taylor DRW. The elemental composition of the human body. Williams DR, Taylor DM. Royal Society of Chemistry; 1995. pp. 16–25;

287. Kaye GW LTH, editor. Chemistry. Tables of physical and chemical constants: and some mathematical functions. National Physical Laboratory; 2005;

288. Haynes WM. CRC Handbook of Chemistry and Physics, 95th Edition. CRC Press; 2014;

289. Watson PE, Watson ID, Batt RD. Total body water volumes for adult males and females estimated from simple anthropometric measurements. Am J Clin Nutr. 1980;33: 27–39;

290. WebElements. Abundance in Earth’s crust: periodicity. In: WebElements. https://www.webelements.com/periodicity/abundance_crust/

291. Richer MG, McCall ML. Chemical abundances from planetary nebulae in local spiral galaxies. Proc Int Astron Union. 2015;11;

292. Hinshaw G, Larson D, Komatsu E, Spergel DN, Bennett CL, Dunkley J, et al. Nine-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results. Astrophys J Suppl Ser. 2013;208: 19;

293. Magrini L, Perinotto M, Mampaso A, Corradi RLM. Chemical abundances of Planetary Nebulae in M 33. Astron Astrophys Suppl Ser. 2004;426: 779–86;

294. Ciardullo R. Planetary Nebulae as Probes of Stellar Populations. Proc Int Astron Union. 2006;2: 325;

295. Jørgensen JK, Favre C, Bisschop SE, Bourke TL, van Dishoeck EF, Schmalzl M. Detection of the Simplest Sugar, Glycolaldehyde, in a Solar-Type Protostar with ALMA. Astrophys J. 2012;757: L4;

296. Cleaves HJ, James Cleaves H, Neish C, Callahan MP, Parker E, Fernández FM, et al. Amino acids generated from hydrated Titan tholins: Comparison with Miller–Urey electric discharge products. Icarus. 2014;237: 182–9;

297. Běhounková M, Tobie G, Čadek O, Choblet G, Porco C, Nimmo F. Timing of water plume eruptions on Enceladus explained by interior viscosity structure. Nat Geosci. 2015;8: 601–4;

298. Hansen CJ, Esposito LW, Aye K-M, Colwell JE, Hendrix AR, Portyankina G, et al. Investigation of diurnal variability of water vapor in Enceladus’ plume by the Cassini ultraviolet imaging

spectrograph. Geophys Res Lett. 2017;44: 672–7;

299. Hansen CJ, Esposito LW, Stewart AIF, Meinke B, Wallis B, Colwell JE, et al. Water vapour jets inside the plume of gas leaving Enceladus. Nature. 2008;456: 477–9;

300. Hansen CJ, Shemansky DE, Esposito LW, Stewart AIF, Lewis BR, Colwell JE, et al. The composition and structure of the Enceladus plume. Geophys Res Lett. 2011;38;

301. Maltagliati L, Montmessin F, Fedorova A, Korablev O, Forget F, Bertaux J-L. Evidence of water vapor in excess of saturation in the atmosphere of Mars. Science. 2011;333: 1868–71;

302. Roth L, Saur J, Retherford KD, Strobel DF, Feldman PD, McGrath MA, et al. Transient water vapor at Europa's south pole. Science. 2014;343: 171–4;

303. Smith KT. Water ice beneath the surface of Ceres. Science. 2017;355: 35–37;

304. Waite JH Jr, Combi MR, Ip W-H, Cravens TE, McNutt RL Jr, Kasprzak W, et al. Cassini ion and neutral mass spectrometer: Enceladus plume composition and structure. Science. 2006;311: 1419–22;

305. Hillier JK, Green SF, McBride N, Schwanethal JP, Postberg F, Srama R, et al. The composition of Saturn's E ring. Mon Not R Astron Soc. 2007;377: 1588–96;

306. Oró J. Comets and the Origin of Life on the Primitive Earth. Cellular Origin, Life in Extreme Habitats and Astrobiology. 2004. pp. 549–65;

307. Miller SL, Urey HC. Organic compound synthesis on the primitive earth. Science. 1959;130: 245–51;

308. Fox D. Did Life Evolve in Ice? Discover Magazine. 2008;

309. Parker ET, James Cleaves H, Callahan MP, Dworkin JP, Glavin DP, Lazcano A, et al. Prebiotic Synthesis of Methionine and Other Sulfur-Containing Organic Compounds on the Primitive Earth: A Contemporary Reassessment Based on an Unpublished 1958 Stanley Miller Experiment. Origins of Life and Evolution of Biospheres. 2010;41: 201–12;

310. Parte AC. LPSN—list of prokaryotic names with standing in nomenclature. Nucleic Acids Res. 2013;42: D613–6;

311. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45: D353–61;

312. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44: D457–62;

313. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 1999;27: 29–34;

314. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2014;42: D32–7;

315. Wheeler DL. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2004;33: D39–45;

316. Glöckner FO, Yilmaz P, Quast C, Gerken J, Beccati A, Ciuprina A, et al. 25 years of serving the

community with ribosomal RNA gene reference databases and tools. J Biotechnol. 2017;261: 169–76;

317. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 2007;35: 7188–96;

318. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41: D590–6;

319. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. 2013;42: D643–8;

320. ISO. ISO 3166 Country Codes. 2017. https://www.iso.org/iso-3166-country-codes.html

321. Mahadevan P, King JF, Seto D. CGUG: in silico proteome and genome parsing tool for the determination of "core" and unique genes in the analysis of genomes up to ca. 1.9 Mb. BMC Res Notes. 2009;2: 168;

322. Mahadevan P, King JF, Seto D. Data mining pathogen genomes using GeneOrder and CoreGenes and CGUG: gene order, synteny and in silico proteomes. Int J Comput Biol Drug Des. 2009;2: 100–14;

323. Zafar N, Mazumder R, Seto D. CoreGenes: a computational tool for identifying and cataloging "core" genes in a set of small genomes. BMC Bioinformatics. 2002;3: 12;

324. Boc A, Diallo AB, Makarenkov V. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. Nucleic Acids Res. 2012;40: W573–9;

325. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14: 1394–403;

326. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990;18: 6097–100;

327. Crooks GE. WebLogo: A Sequence Logo Generator. Genome Res. 2004;14: 1188–90;

328. Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res. 2000;28: 292;

## VITA

### 1. FIELD OF SPECIALIZATION

Major Biomedicine

### 2. ADDRESS

Mailing 9114 McPherson Road Suite 2509, Laredo, Texas 78045

Phone +1 (936) 666-6694

Email Rafael@DelizAguirre.com;
rafael.delizaguirre@dusty.tamiu.edu

Website Rafael.DelizAguirre.com

### 3. EDUCATION

2009 – 2014 **Bachelor of Science in Biology with a Minor in World Affairs**
Baylor University, Waco, Texas

### 4. EMPLOYMENT HISTORY

2012 – present **Medical Assistant**
Rafael J Deliz, MD, PA, Laredo, Texas

2017 **Professional Tutor for the Joint Admissions Medical Program**
Department of Biology & Chemistry, Texas A&M International University, Laredo, Texas

2014 – 2015 **Research Assistant II**
Department of Cancer Biology, University of Texas: MD Anderson Cancer Center, Houston, Texas

2014 **Student Assistant**
Center for Astrophysics, Space Physics, and Engineering Research, Waco, Texas

2009 – 2012 **Spanish Tutor**
Department of Spanish, Baylor University, Waco, Texas

### 5. RESEARCH EXPERIENCE

2016 – Present **Schmidl (Clinical Microbiology) Laboratory**
Department of Biology & Chemistry, Texas A&M International University, Laredo, Texas

2014 – 2015 **Kalluri (Matrix Biology) Laboratory**
Department of Cancer Biology, University of Texas: MD Anderson Cancer Center, Houston, Texas

2014 **Hypervelocity Impacts & Dusty Plasmas Laboratory, and the Space Science Laboratory**
Center for Astrophysics, Space Physics and Engineering Research (CASPER), Waco, Texas

2012 – 2013 **Danley (Population Genetics) Laboratory**
Department of Biology, Baylor University, Waco, Texas

2010 – 2011 **Trawick (Biochemistry) Research Group**
Department of Chemistry & Biochemistry, Baylor University, Waco, Texas

2009 – 2012 **Kebaara (Cell Biology) Laboratory**
Department of Biology, Baylor University, Waco, Texas