Comparative Genomics Revealed Genus Specific Encoding of Amino Acids by Tri-Nucleotide SSRs in Human Pathogenic Streptococcus and Staphylococus Bacteria

Sahil Mahfooz
Babasaheb Bhimrao Ambedkar University
Jitendra Narayan
CSIR Institute of Genomics & Integrative Biology
Ruba Mustafa Elsaid Ahmed
University of Hail
Amel Bakri Mohammed El Hag
University of Hail College of Medicine
Nuha Abdel Rahman Khalil Mohammed
University of Hail College of Medicine
Mohd Adnan Kausar (adnankausar1@gmail.com)
University of Hail https://orcid.org/0000-0002-8931-9290

Research Article

Keywords: Phase variation, simple sequence repeats, human pathogens, genetic relationship

Posted Date: January 6th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1214328/v1

Abstract

Pathogenic bacteria use phase variation of surface molecules and other characteristics as a significant adaptation mechanism. Repetitive sequences made up of numerous identical repeat units can be found in many phase variable genes. Here, we investigated the frequency and distribution of long-SSRs in 15 human pathogenic *Staphylococcus, Streptococcus*, and *Enterococcus* bacteria. Long-SSRs were found to be distributed differently in the genic and intergenic sequences. In the genic sequences, 61.3 SSRs were discovered on average, while 16.2 SSRs were found in the intergenic regions. *Staphylococci* exhibited the highest frequency of SSRs, followed by *Enterococcus*, and *Streptococci* had the lowest frequency of SSRs. Higher A+T content was found to be the best predictor of long-SSR in these human pathogens. Tetranucleotide repeats predominated in intergenic regions, while trinucleotide repeats predominated in genic regions. In human pathogenic *Streptococcus* and *Staphylococcus* bacteria, genus-specific encoding of amino acids by tri-nucleotide SSRs was observed. A genetic relationship between these human pathogenic bacteria was derived based on the presence of SSRs in the housekeeping genes and compared to the phylogeny generated based on the 16S ribosomal RNA gene.

1. Introduction

Microorganisms play a decisive role in our lives. They are responsible for many of the food products that we consume, and at the same time, they are the cause of many human diseases. Infections caused by gram-positive bacteria (GPB) are a common cause of significant illness in individuals. *Streptococcus* is a genus of gram-positive bacteria that includes a number of clinically relevant species that cause a wide range of infections in humans and animals, each with its own symptoms and course (Krismer, et al. 2017). Streptococci can colonise human and animal mucous surfaces and are considered opportunistic pathogens, meaning they can cause acute infections under certain conditions. Streptococcal species such as *Streptococcus pyogenes* and *Streptococcal pneumonia* are extremely virulent and cause serious infections such as pneumonia, necrotizing fasciitis, sepsis, and meningitis, whereas others (for example, *Streptococcal mutans, Streptococcal sanguis, Streptococcal agalactiae*, and *Streptococcal anginosus*) cause endocarditis, abscesses (Kosecka-Strojek, et al. 2019).

Species of *Staphylococcus* bacteria, which are found in the natural microbiota of humans and animals, also act as pathogens. *Staphylococcus aureus*, the group's main pathogen, causes a wide range of clinical infections in humans, including bacteremia, endocarditis, and infections associated with invasive medical devices. (Rossi, et al. 2020). Meanwhile, coagulase-negative staphylococci (CoNS), particularly *S. haemolyticus* and *S. epidermidis*, have emerged as recurrent causal agents of nosocomial infections, particularly those involving indwelling devices (Becker, et al. 2014). Similarly, gram positive species of *Enterococcus* are abundantly found in the intestines of humans. The two most common bacteria responsible for infection in humans are *E. faecalis* and *E. faecium* (Safdar and Armstrong 2019).

The pathogenicity of these microorganisms is influenced by a variety of factors. The ability of microbial pathogens to persist in human hosts protected by adaptive immunity is a major biological challenge.

Antigenic variation of surface components is one technique used by microbial parasites to evade host immune responses during long-term colonization. (Woolhouse, et al. 2001). Although microbial populations are continually changing in all colonised habitats, many successful long-term colonists have mastered the barrier of producing antigenic variations (Krinos, et al. 2001).

Bacterial pathogens use polymerase slippage of simple sequence repeats (SSRs) to generate phenotypic variation and increased fitness through localised hypermutation. SSRs are made up of tandem repetitions of short oligonucleotides that have functional and/or structural characteristics that set them apart from other DNA sequences (Mrazek, et al. 2007). The SSR distribution is not random, and they can be found all over the genome. SSRs can be found in the 30-UTR, 50-UTR, protein-coding, and non-coding sections of the genome (Mahfooz, et al. 2015). SSRs are often utilised as genetic markers since they are extremely changeable and exceptionally polymorphic (Datta, et al. 2010). SSR types are expected to alter transcriptional activity and have a functional role in the evolution of gene regulation (Geng, et al. 2020) and SSRs are thought to be different in various taxa or different areas of the same taxon (Song, et al. 2021).

In bacteria, a large number of the SSR tracts are present, with motifs spanning from one to six nucleotides (Gur-Arie, et al. 2000), there are certain SSRs that are exceptionally long (more than 12 nucleotides). These long SSRs are found within the reading frame or promoter of a class of genes known as contingency loci, whose functions are typically involved in direct interactions with the host (Moxon, et al. 2006). Since the genome sequences of most of the pathogenic bacteria are available, one can easily analyse the distribution of these loci among the human pathogens to get an insight into their evolution as pathogens. Hence, we wish to analyse the frequency and distribution of these long SSRs among the three most common genera of human pathogens species. Furthermore, we would like to examine in which genes these long SSRs are present and how the presence of these long repeats could affect the transcription. Finally, we would like to elucidate a genetic relationship based on the presence and absence of these long SSRs in certain housekeeping genes.

2. Material And Methods

2.1 Genome sequences. Whole-genome and coding sequences of 15 human pathogenic bacteria were downloaded in fasta format from National Center for Biotechnology Information (www.ncbi.nlm.nih.gov). This study did not include plasmid DNA sequences. On the basis of current annotations, genic and intergenic areas within the genomes were determined. Of the 15 bacteria, 4 belong to the genus *Staphylococcus*, 9 belongs to the genus *Streptococcus,* and the remaining 2 were *Enterococcus* bacteria (Table 1). Data regarding G+C content and the number of genes were taken from the information provided along with the genome sequences.

2.2 Simple Sequence Repeat (SSR) mining:

Prokaryotic organisms have a higher number of short repeats, which typically range from two to four repeat units in length. Mononucleotide repeats are the most common among them, followed by

trinucleotide repeats. Sequencing errors result in a large number of mononucleotide repeats. As a result, mononucleotide repeats were excluded, and di-to hexa-nucleotide repeats longer than 12 nucleotides were designated as long SSRs. In this case, a di-nucleotide pattern must appear at least six times, a tri-nucleotide motif must appear four times, a tetra-nucleotide motif must appear three times, and penta-, and hexa-nucleotide motifs must appear three times each. SSR mining was carried out with the help of SSR locator software (da Maia, et al. 2008). This software can mine SSRs as well as create primers and simulate PCR reactions against various databases. Each SSR's frequency, relative abundance (RA), and relative density (RD) were computed. The total number of SSRs and the total length (bases) covered by each of the SSRs were divided by the entire size of the sequence to determine RA and RD (Mb). R was used to do Pearson correlation analysis and Principal Component Analysis (PCA), and R Studio was used to create a graph.

2.3 Statistical analysis:

The presence or absence of repetitive motifs in human pathogens' housekeeping genes was used to produce binary data, which was then analysed using the SIMQUAL approach to generate Jaccard's similarity coefficients with NTSYSpc software version 2.1 (Rohlf 1998). Using the Unweighted Paired Group Method of Arithmetic Averages (UPGMA) method and SAHN clustering, these similarity coefficients were utilised to create a dendrogram demonstrating genetic relatedness among the species. The polymorphism information content (PIC) was determined as previously described (Botstein, et al. 1980). The probability that two randomly chosen copies of a gene represent different alleles within a population is defined as PIC. The following formula was used to calculate the PIC value:

$$\mathbf{PIC}_i = 1 - \sum_{j=1}^{n} \mathbf{Pij}^2$$

where P_{ij} represents the frequency of the j^{th} pattern for marker i, and summation extends over n patterns.

The ClustalW program in the MEGA 5.2 software was also used to create a phylogenetic tree of the 16S rRNA gene. The phylogenetic tree was built using the neighbor-joining algorithm and 1000 replicates of bootstrap analysis.

2.4 Gene enrichment analysis

Gene enrichment analysis was performed on the genes that contain long SSRs. This could aid in the discovery of biological pathways that are rich in repeats. One pathogenic microorganism with the greatest SSR count was chosen among Staphylococcus, Streptococcus, and Enterococcus bacteria for this study. We employed a custom protocol that included three main steps: 1) detailed description of a gene list derived from "omic" data; 2) computational determination of statistically enriched pathways; and 3) visualisation and analysis of the results. We used PANZER (Protein ANNotation with Z-scoRE) for the first and second steps, a completely automated tool for functional annotation of unknown functional

prokaryotic and eukaryotic proteins. The programme accepts protein sequences and predicts functional descriptions (DE) and GO classifications. In the final phase, visual analysis was performed using visualisation tools provided by ShinyGO v0.61 (http://bioinformatics.sdstate.edu/gonovel). It offers a graphical web tool that can assist in gaining actionable information and access to KEGG and STRING for route diagrams and protein-protein interaction networks.

3. Results

3.1 Staphylococcus has the maximum relative abundance and density of long SSRs

To analyse the presence of long SSRs in human pathogenic bacteria, we scanned the presence of these SSRs in different species of Staphylococcus, Streptococcus, and Enterococcus bacteria. Among all, a total of 1286 long SSRs were identified, with an average of 77.5 SSRs/species. Differential distribution of SSRs was observed in the genic and intergenic regions. An average of 61.3 SSRs were found in the genic sequences, whereas 16.2 SSRs were located in the intergenic regions. The maximum frequency of SSRs was identified among Staphylococci (Avg=117.5), which was followed by Enterococcus (Avg.=100), *Streptococci* had the least abundance of SSRs (Avg.=54.7). Individually, *Staphylococcus epidermidis* has the highest frequency of long SSR (125), closely followed by *Staphylococcus haemolyticus* (123), and least in Streptococcus mitis (45). We observed a statistically significant correlation between the genome size and the frequency of SSRs (r= 0.83, p=0.0001), which means that genome size could influence the frequency of SSRs. Hence, in order to minimise this, we have calculated the relative abundance and density of SSRs. With updated calculations, *Staphylococcus epidermidis* remained at the top with the highest RA and RD (50.6 and 631.5), but *Streptococcus suis* became the least (22.3 and 335.8), displacing *Streptococcus mitis* (23.4 and 305.2) (Table 2). We further calculated the percentage of repeats in the genomes. Again, the maximum percentage of repeats (0.063) was observed in *S. epidermidis*, whereas *Streptococcus mutans* had the lowest (0.029).

Among the classes of SSRs, tetranucleotide repeats were preferred in the intergenic regions (67.1.2%), whereas genic regions were dominated by trinucleotide repeats (49%). The second most frequent motif class in the intergenic region was tetranucleotide repeats (17.6%), while it was tetranucleotide repeats (44.4%) in the genic sequences. Hexanucleotide SSRs were the least preferred (4.8%) in the intergenic regions, whereas genic regions avoided the presence of di-nucleotide SSRs (0.5%). *Staphylococcus haemolyticus* harbours the maximum SSRs as tetranucleotide repeats in both genic as well as intergenic regions (Supplementary table 1).

3.2 Tetranucleotide repeats were the most preferred classes of repeats among the human pathogens

We further estimated the RA and RD of different classes of repeats among human pathogens. Most pathogens, irrespective of their genera, prefer tetranucleotide repeats within their genomes (Table 3). The

species defying this trend were *S. epidermidis*, *S. pyogenes*, and *S. sobrinus*. It was interesting to note that *S. saprophyticus* had exactly the same RA for tetra and trinucleotide SSRs. Dinucleotide repeats were completely absent in *S. aureus*, *S. haemolyticus, S. mutans, S. salivarius*, and *S. suis.* Similarly, *S. gordonii* had no hexanucleotide repeats in its genome.

3.3 Hexanucleotides constitute the longest SSRs among pathogens

We further evaluated our data to find the longest SSRs within each genome. The majority of the genomes (73.0%) have hexanucleotides which are repeated thrice as the longest SSRs. *S. agalactiae* and *S. sobrinus* have trinucleotide repeats as their longest SSRs, along with hexanucleotide and pentanucleotide repeats (Table 4). The majority of the long SSRs were of 18 base pairs in length, with the exception of a 125 bp long repeat in *S. suis* where a pentanucleotide (GAGCA) was found to be repeated 25 times. It is noteworthy that most of the long SSRs were located in genes that are directly or indirectly related to pathogenesis.

3.4 Gene enrichment studies

Gene sequences with long SSRs from *S. haemolyticus*, *E. faecium, and S. agalactiae* were taken for gene enrichment analysis. A total of 231 SSR-containing sequences from all three species were aligned to 34 different functional pathways. The highest number of genes with repeats was located in *S. haemolyticus* (187, 47.5%), which was followed by *E. faecium* (114, 28%) and the least was observed in *S. agalactiae* (93, 23.6%). SSRs have been discovered in the genes of specific metabolic pathways in all species. Genes involved in the organic substance metabolic process, nitrogen compound metabolic process, catalytic activity, and small molecule binding, for example, were discovered in each of the three species. Some pathways have common genes for two species. For instance, repeats were discovered in genes linked to biological regulation in *S. haemolyticus*, and *E. faecium*, and some pathways have genes for only one species (a stress-responsive gene was present only in *S. haemolyticus*) (Figure 1). The most noteworthy finding in this study was the presence of SSRs in pathogenicity related genes, which validates the important role of SSRs in pathogenicity.

3.5 Genus wise amino acid was preferred

Since trinucleotide SSRs can encode amino acids if they are present in the reading frame, it would be interesting to investigate the occurrence of amino acids encoded by trinucleotide SSRs in the genic sequences of pathogenic bacteria. A total of 1820 amino acids were encoded by tri-nucleotide SSRs among all the pathogens. Isoleucine was the most abundant amino acid (11.0%), which was followed by lysine (8.2%). Glutamic acid was the third most abundant amino acid (8.1%), whereas proline encoding repeats were the least abundant (0.3%). It is evident from the data that different bacteria have their own preferences for amino acids. Isoleucine, for example, is the most abundant amino acid in two Streptococcus, two Staphylococcus, and one Enterococcus bacteria, whereas glutamic acid was highest in one bacterium from each genus (Supplemental table 2). Since there are 20 amino acids, it would be difficult to interpret each in a different organism. Hence, we further perform Principal Component Analysis

(PCA), which is a technique for lowering the dimensionality of datasets, improving interpretability while minimising information loss. PC1 has higher values for Staphylococcus bacteria as they clustered together, whereas Streptococcus bacteria clustered together with lower PC1 values. Enterococcus bacteria were grouped distantly with higher and lower PC1 values.

3.6 Motif conservation among human pathogens

We selected 13 housekeeping genes that perform important biological functions like replication, transcription, and translation-related functions to study the conservation of motifs within them. The highest concentration of long SSRs was located in the rplD (50S ribosomal protein) gene. This was followed by the infB gene, which had five SSRs, and the uvrC gene, which had three SSRs. Among the housekeeping genes, maximum conservation of motifs was recorded in the rplD gene where motif $(gtg)_4$ was found conserved within eight species of Staphylococcus and *Streptococus* bacteria. Among the genus, Staphylococcus shows more conservation as compared to its counterparts. For instance, motif $(atca)_3$ was conserved within the infB gene of *S. aureus*, *S. epidermidis*, and *S. saprophyticus*. Similarly, motif $(tga)_4$ was conserved in *S. aureus* and *S. epidermidis*, and motif $(gtc)_4$ was conserved in the *rps*O (30S ribosomal protein) gene of *S. epidermidis* and *S. haemolyticus* (Table 5). Motif conservation was also witnessed among Staphylococcus bacteria. Motif $(tga)_4$ was conserved between *S. agalactiae* and *S. mutans* in the *par*C (DNA topoisomerase 4 subunit A) gene.

3.7 Genetic relationship

To determine the degree of polymorphism, a similarity matrix was created using Jaccard's estimate of similarity, which is based on the probability that SSR found in one species' housekeeping genes will also be found in the housekeeping genes of another species. Over the 120 combinations, the calculated similarity coefficient ranged from 0 to 1.0, with a mean of 0.73. Staphylococcus bacteria had a higher average genetic diversity of 26 percent when compared to Streptococcus and Enterococcus bacteria, which had lower average genetic diversity of 11.0 percent and 3.0 percent, respectively. Gene rpmH (50S ribosomal protein) was the most informative (PIC =0.99), whereas rplD was the least (PIC =0.78) among all. We further constructed a dendrogram based on similarity coefficient values. The dendrogram resulted in two main clusters, A and B. Cluster A is comprised of *S. aurius* and *S. saprophiticus*, whereas the remaining species cluster together in B. B was further subdivided into two clusters, BI and BII. BI has the remaining Staphylococcus species, whereas in BII, Streptococcus and Enterococcus bacteria cluster together in two separate clades. Since this dendrogram is based on hypervariable regions of the genome, it would be interesting to see how much it differs from conserved region-based phylogeny. We further constructed a dendrogram based on the 16S ribosomal regions of the genome. The dendrogram was also divided into two main clusters, A and B. Cluster A groups all the Streptococcal bacteria together, whereas Cluster B separates Staphylococcus and Enterococcus bacteria into two different clades.

Discussion

Bacterial pathogens confront difficult conditions due to numerous unpredictable, frequently abrupt, and dynamic changes that occur in the host environment or during transmission from one host to another

(Bayliss 2009). Bacterial adaptation to their hosts entails either a system for sensing and responding to environmental changes or selecting mutation-induced variations within their contingency loci (Moxon, et al. 2006). These loci allow bacterial populations to adapt to or survive selective pressures by generating and spreading genetic variants that are "fitter," or better adapted to, a particular environment than the majority of the population. Through comparative genomics, it is now possible to find differences in genetic variants across entire genomes, and to tie those differences together to biological function, as well as to learn more about selective patterns of gene transfer and evolutionary pressures or loss, especially when it comes to virulence in pathogenic organisms (Fitzgerald and Musser 2001). We analysed the frequency and distribution of long SSRs in sequenced species of human pathogenic Staphylococcus, Streptococcus, and Enterococcus bacteria. Among the three, *Staphyloccus* has a relatively higher frequency of long repeats as compared to the others. To find an appropriate justification for higher repeats in Staphylococcus bacteria, we looked into their G+C content. In our previous reports on fungi, we observed a positive correlation between G+C content and the frequency of SSRs (Mahfooz, et al. 2017, Mahfooz, et al. 2016). However, we found a statistically significant negative correlation (r^2 = -0.83, p = 0.0001) between the two in this study. Hence, we can hypothesise that higher A+T content is a good predictor of the frequency of repeats among these human pathogenic bacteria. We found *S. epidermidis* harbouring the maximum frequency of long repeats among all the pathogenic bacteria. A close comparison between S. *epidermidis* and *S. aureus* revealed a higher percentage (~9) of genomic elements (genome islands and *Staphylococcus* cassette chromosome-like elements, insertion sequence elements, integrated prophage, integrated plasmids, and composite transposons) as compared with its closest species, *S. aureus* (~7)(Hiramatsu, et al. 2001). We can speculate that this could be the possible reason why we obtained a higher frequency of long SSRs in S. *epidermidis.*

Among the classes of SSRs, we found an abundance of tetranucleotide SSRs in the intergenic region, whereas trinucleotide SSRs were found dominant in the genic regions. Tetranucleotide repeats present in the intergenic regions are reported to modulate transcription factor binding and consequently modulate gene expression (Martin, et al. 2005). The presence of trinucleotide repeats in the genic regions is expected as it avoids frameshift mutations and these triplets could code for amino acid runs that may have specific functions in the protein structure (Metzgar, et al. 2000). Further analysis of repeat classes at the whole genome level revealed that in most of bacterial species, the RA and RD of tetranucleotide SSRs were the highest. Its occurrence in intergenic regions is fine as it regulates gene expression by binding the RNA polymerase. However, its presence in genic sequences is surprising as it could change the open reading frame. It is believed that rearrangement within these tetrameric repeats could work as a switch-on/off mechanism in phase variable genes (De Bolle, et al. 2000).

Further analysis of the data showed hexanucleotides constituted the longest SSRs in most of the species, which is expected as they contribute the highest number of repeats when compared to other classes of SSRs. We observed an unexpectedly higher repeat number (25) of a pentanucleotide repeat (gagca) that codes for a hypothetical protein in *S. suis*. We can hypothesise that *S. suis* could have acquired this repeat through horizontal gene transfer (Perna, et al. 2001).

We further examined the tri-nucleotide SSRs that have a probability of being transcribed by codons and translated in frame into amino acid residue repeats. Isoleucine was the most abundant amino acid, followed by lysine and glutamic acid. It has been reported that in bacterial physiology, branched-chain amino acids like isoleucine play a variety of roles, from promoting protein synthesis to signalling and fine-tuning the adaptation to amino acid deficiency. In some pathogenic bacteria, the response to amino acid deficiency includes activation of virulence gene expression. As a result, isoleucine aids not just infection but also evasion of host defences (Kaiser Julienne, et al.). The second most abundant amino acid, lysine, is used for protein synthesis and the peptidoglycan layer of Gram-positive bacterial cell walls. Additionally, the relevance of lysine for bacterial cell survival is highlighted by the availability of numerous biosynthetic routes for lysine synthesis in bacteria (Gillner, et al. 2013). The quantity of glutamic acid in bacteria appears to be linked mostly to tolerance to acidic environments. Food borne diseases and spoilage bacteria would be able to grow on acidic foods if they developed acid resistance. This feature is also a virulence factor, as it permits pathogens to pass past the stomach barrier's very acidic conditions, respectively (Feehily and Karatzas 2013). We wanted to explore whether any association could be found based on amino acids encoded by tri-nucleotide SSRs. To do this, we used principal component analysis (PCA), where a genus-wide clustering of *Staphylococcus* and *Streptococcus* bacteria was observed. The genus-wide clustering of Staphylococcus and Streptococcus species was expected as, in the course of evolution, positive selection tends to prefer those amino acids which are required for organism growth and survivability (Loewe and Hill 2010). The distant clustering of Enterococcus species could be due to the marked difference in the amino acids encoded by the trinucleotide SSRs.

We then performed functional annotation on the genes containing these long SSRs. This allowed us to figure out if these long SSRs were linked to any particular biological process. SSRs were found in the majority of the important biological function pathways.This raises the risk of non-functional proteins as a result of frameshifts, implying that these species may have evolved SSRs in coding areas to promote phase variation(Lin and Kussell 2012). Notably, SSRs were located in pathogensity genes in *Staphylococcus* and *Streptococcus* bacteria, which is direct evidence of their involvement in pathogenicity.

SSRs were found in housekeeping genes in our study; however, the majority of SSRs were tri-nucleotide motifs, with a few tetra- and hexanucleotide motifs. These housekeeping genes are linked to a variety of biological functions. The existence of SSRs in housekeeping genes is surprising because SSRs are known to be mutation hotspots(Lin and Kussell 2012), and any mutation in the housekeeping gene would be fatal. Because most SSRs are trinucleotides, their chances of causing phase variation are low. In a previous study, a significant difference was reported in the 5'-UTR while comparing the densities and repeat types of SSRs between housekeeping and tissue-specific genes. According to the report, the GC content of trinucleotide SSRs in the 5'-UTRs of housekeeping genes is higher than that of tissue-specific genes (Lawson and Zhang 2008). Motif $(gtg)_{4\ was}$ found to be conserved in the *rplD* gene in eight species of Staphylococcus and Streptococcus bacteria. GTG is also an alternative start codon in bacteria

(Hwang, et al. 2005), but its presence as repeated motifs ruled out any such role and focused on its translation to valine, which has a multifaceted physiological role in bacterial survival and pathogenicity (Kaiser Julienne, et al.).

We further attempted to construct a phylogenetic tree using the presence of SSRs, in particular housekeeping genes, as these genes are more conserved with minimal selection pressure. The highest and lowest polymorphism levels were obtained among ribosomal 50S subunit genes. The 50S ribosomal subunit is always an easy target for various antibiotics (Champney, et al. 2003), but with the presence of repetitive sequences, the bacteria may evade the binding of these antibiotics. A SSR-based phylogeny grouped Enterococcus species along with Streptococcus, whereas a 16S ribosomal region-based phylogeny grouped Enterococcus and Staphylococcus together. It was until 1984 that Enterococcus species were classified as Streptococcus (Andrewes and Horder 1906). However, with the advancement of techniques like DNA-DNA hybridisation and 16S RNA sequencing, a new genus Enterococcus was formed (Schleifer and Kilpper-Bälz 1984). We observed that SSR-based phylogeny's resolution power was low as it could not differentiate between five species of Streptococcus, which 16S ribosomal region-based phylogeny easily did.

Conclusion

In this study, we have discovered a pattern of long-SSR distribution in the genic and intergenic regions of human pathogenic Staphylococcus, Streptococcus, and Enterococcus bacteria and established a relationship between them. Human pathogens have a higher relative abundance of SSRs as compared to non-pathogenic control. The genic regions were home to the majority of the repeat motifs. As a result, the presence of microsatellites in human pathogens is not random. A novel finding in this study was the preference for genus-specific trinucleotide SSR encoded amino acids.

Declarations

Acknowledgment

This research has been funded by Research Deanship at the University of Ha’il – Saudi Arabia through project number RG-21 152.

Conflict of interest

None declared

References

1. Andrewes FW, Horder TJ (1906) A study of the streptococci pathogenic for man. The Lancet 168: 708-713 doi: https://doi.org/10.1016/S0140-6736(01)31538-6

2. Bayliss CD (2009) Determinants of phase variation rate and the fitness implications of differing rates for bacterial pathogens and commensals. FEMS Microbiol Rev 33: 504-520 doi: 10.1111/j.1574-6976.2009.00162.x
3. Becker K, Heilmann C, Peters G (2014) Coagulase-negative staphylococci. Clinical microbiology reviews 27: 870-926 doi: 10.1128/CMR.00109-13
4. Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32: 314-331 doi:
5. Champney WS, Chittum HS, Tober CL (2003) A 50S ribosomal subunit precursor particle is a substrate for the ErmC methyltransferase in *Staphylococcus aureus* cells. Curr Microbiol 46: 453-460 doi: 10.1007/s00284-002-3901-8
6. da Maia LC, Palmieri DA, de Souza VQ, Kopp MM, de Carvalho FIF, Costa de Oliveira A (2008) SSR Locator: Tool for simple sequence repeat discovery integrated with primer design and PCR simulation. Int. J. Plant Genomics 2008: 412696 doi: 10.1155/2008/412696
7. Datta S, Mahfooz S, Singh P, Choudhary AK, Singh F, Kumar S (2010) Cross-genera amplification of informative microsatellite markers from common bean and lentil for the assessment of genetic diversity in pigeonpea. Physiol Mol Biol Plants 16: 123-134 doi: 10.1007/s12298-010-0014-x
8. De Bolle X, Bayliss CD, Field D, van de Ven T, Saunders NJ, Hood DW, Moxon ER (2000) The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases. Mol Microbiol 35: 211-222 doi: 10.1046/j.1365-2958.2000.01701.x
9. Feehily C, Karatzas KA (2013) Role of glutamate metabolism in bacterial responses towards acid and other stresses. J Appl Microbiol 114: 11-24 doi: 10.1111/j.1365-2672.2012.05434.x
10. Fitzgerald JR, Musser JM (2001) Evolutionary genomics of pathogenic bacteria. Trends Microbiol 9: 547-553 doi: 10.1016/s0966-842x(01)02228-4
11. Geng H, Hao L, Cheng Y, Wang C, Huang S, Wei W, Yang R, Li H, Liu S, Yu H, Lu H (2020) Interaction between CA repeat microsatellites and HIF1α regulated the transcriptional activity of porcine IGF1 promoter. J Appl Genet 61: 105-112 doi: 10.1007/s13353-019-00529-4
12. Gillner DM, Becker DP, Holz RC (2013) Lysine biosynthesis in bacteria: a metallodesuccinylase as a potential antimicrobial target. Journal of biological inorganic chemistry : JBIC : a publication of the Society of Biological Inorganic Chemistry 18: 155-163 doi: 10.1007/s00775-012-0965-1
13. Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM, Kashi Y (2000) Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism. Genome Res 10: 62-71 doi:
14. Hiramatsu K, Cui L, Kuroda M, Ito T (2001) The emergence and evolution of methicillin-resistant Staphylococcus aureus. Trends Microbiol 9: 486-493 doi: 10.1016/s0966-842x(01)02175-8
15. Hwang SR, Garza CZ, Wegrzyn JL, Hook VY (2005) Demonstration of GTG as an alternative initiation codon for the serpin endopin 2B-2. Biochem Biophys Res Commun 327: 837-844 doi: 10.1016/j.bbrc.2004.12.053

16. Kaiser Julienne C, Heinrichs David E, Garsin Danielle A Branching Out: Alterations in Bacterial Physiology and Virulence Due to Branched-Chain Amino Acid Deprivation. mBio 9: e01188-01118 doi: 10.1128/mBio.01188-18
17. Kosecka-Strojek M, Sabat AJ, Akkerboom V, Kooistra-Smid AMD, Miedzobrodzki J, Friedrich AW (2019) Development of a reference data set for assigning Streptococcus and Enterococcus species based on next generation sequencing of the 16S–23S rRNA region. Antimicrobial Resistance & Infection Control 8: 178 doi: 10.1186/s13756-019-0622-3
18. Krinos CM, Coyne MJ, Weinacht KG, Tzianabos AO, Kasper DL, Comstock LE (2001) Extensive surface diversity of a commensal microorganism by multiple DNA inversions. Nature 414: 555-558 doi: 10.1038/35107092
19. Krismer B, Weidenmaier C, Zipperer A, Peschel A (2017) The commensal lifestyle of Staphylococcus aureus and its interactions with the nasal microbiota. Nat Rev Microbiol 15: 675-687 doi: 10.1038/nrmicro.2017.104
20. Lawson MJ, Zhang L (2008) Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5'-UTR region. Gene 407: 54-62 doi: 10.1016/j.gene.2007.09.017
21. Lin WH, Kussell E (2012) Evolutionary pressures on simple sequence repeats in prokaryotic coding regions. Nucleic Acids Res 40: 2399-2413 doi: 10.1093/nar/gkr1078
22. Loewe L, Hill WG (2010) The population genetics of mutations: good, bad and indifferent. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 365: 1153-1167 doi: 10.1098/rstb.2009.0317
23. Mahfooz S, Singh SP, Mishra N, Mishra A (2017) A Comparison of Microsatellites in Phytopathogenic *Aspergillus Species* in Order to Develop Markers for the Assessment of Genetic Diversity among Its Isolates. Front Microbiol 8: 1774 doi: 10.3389/fmicb.2017.01774
24. Mahfooz S, Singh SP, Rakh R, Bhattacharya A, Mishra N, Singh PC, Chauhan PS, Nautiyal CS, Mishra A (2016) A Comprehensive Characterization of Simple Sequence Repeats in the Sequenced Trichoderma Genomes Provides Valuable Resources for Marker Development. Front Microbiol 7: 575 doi: 10.3389/fmicb.2016.00575
25. Mahfooz S, Srivastava A, Srivastava AK, Arora DK (2015) A comparative analysis of distribution and conservation of microsatellites in the transcripts of sequenced Fusarium species and development of genic-SSR markers for polymorphism analysis. FEMS Microbiol Lett 36210.1093/femsle/fnv131
26. Martin P, Makepeace K, Hill SA, Hood DW, Moxon ER (2005) Microsatellite instability regulates transcription factor binding and gene expression. Proc Natl Acad Sci U S A 102: 3800-3804 doi: 10.1073/pnas.0406805102
27. Metzgar D, Bytof J, Wills C (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res 10: 72-80 doi:
28. Moxon R, Bayliss C, Hood D (2006) Bacterial Contingency Loci: The Role of Simple Sequence DNA Repeats in Bacterial Adaptation. Annual Review of Genetics 40: 307-333 doi: 10.1146/annurev.genet.40.110405.090442

29. Mrazek J, Guo X, Shah A (2007) Simple sequence repeats in prokaryotic genomes. Proc Natl Acad Sci U S A 104: 8472-8477 doi: 10.1073/pnas.0702412104
30. Perna NT, Plunkett G, 3rd, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Pósfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR (2001) Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature 409: 529-533 doi: 10.1038/35054089
31. Rohlf FJ (1998) On applications of geometric morphometrics to studies of ontogeny and phylogeny. Syst Biol 47: 147-158; discussion 159-167 doi:
32. Rossi CC, Pereira MF, Giambiagi-deMarval M (2020) Underrated Staphylococcus species and their role in antimicrobial resistance spreading. Genet Mol Biol 43: e20190065 doi: 10.1590/1678-4685-gmb-2019-0065
33. Safdar A, Armstrong D (2019) *Staphylococcus, Streptococcus, and Enterococcus*. In: Safdar A (ed) Principles and Practice of Transplant Infectious Diseases. Springer New York, New York, NY, pp. 419-445.
34. Schleifer KH, Kilpper-Bälz R (1984) Transfer of Streptococcus faecalis and Streptococcus faecium to the Genus Enterococcus nom. rev. as Enterococcus faecalis comb. nov. and Enterococcus faecium comb. nov. Int J Syst Evol Microbiol 34: 31-34 doi: https://doi.org/10.1099/00207713-34-1-31
35. Song X, Yang T, Zhang X, Yuan Y, Yan X, Wei Y, Zhang J, Zhou C (2021) Comparison of the Microsatellite Distribution Patterns in the Genomes of Euarchontoglires at the Taxonomic Level. Front Genet 1210.3389/fgene.2021.622724
36. Woolhouse ME, Taylor LH, Haydon DT (2001) Population biology of multihost pathogens. Science 292: 1109-1112 doi: 10.1126/science.1059026

Tables

Tables 1-5 are not available with this version

Figures

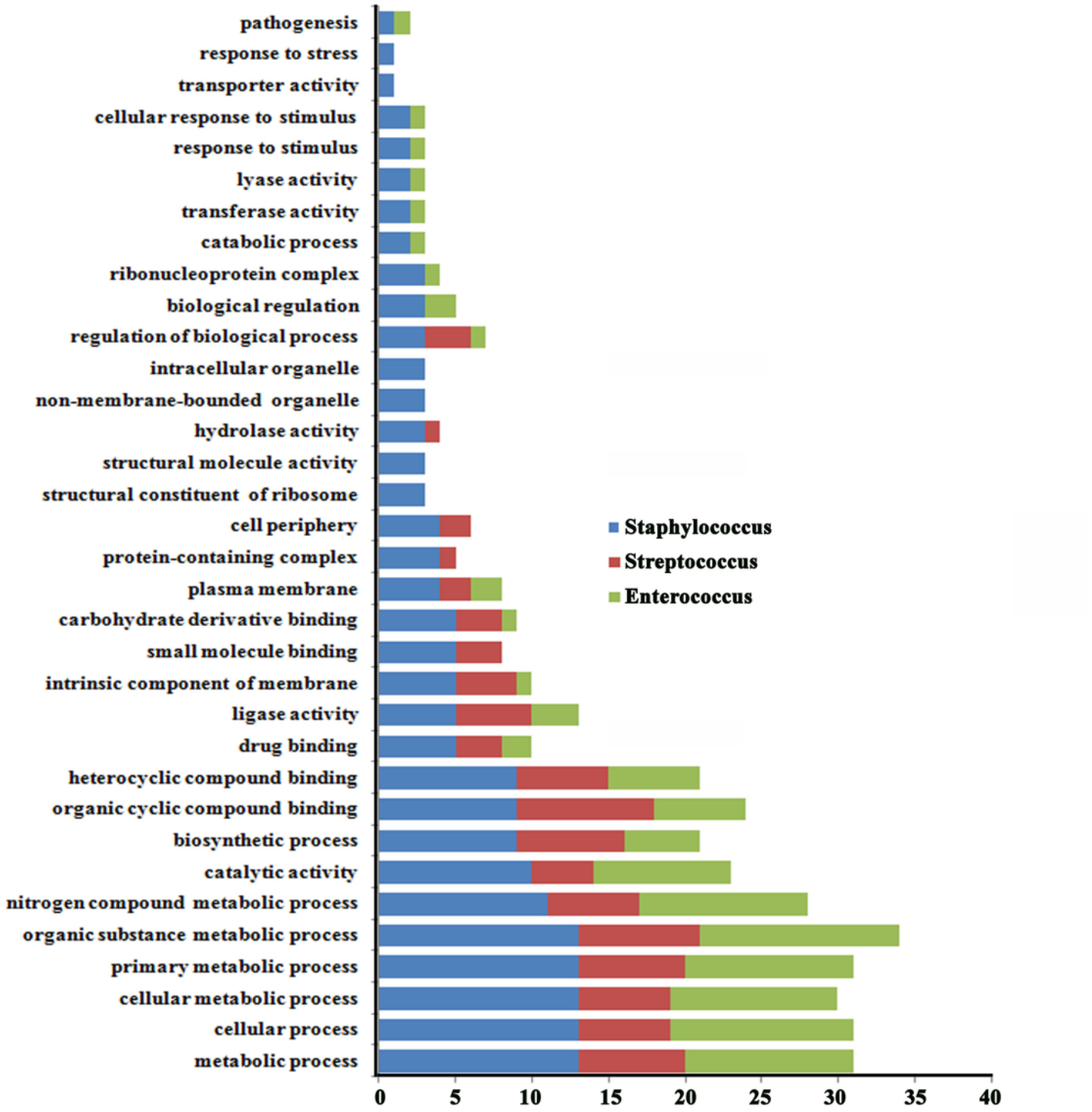

Figure 1

Figure showing gene ontology (GO) enrichment analysis of genes with long SSRs in human pathogenic *S. haemolyticus*, *E. faecium, and S. agalactiae*

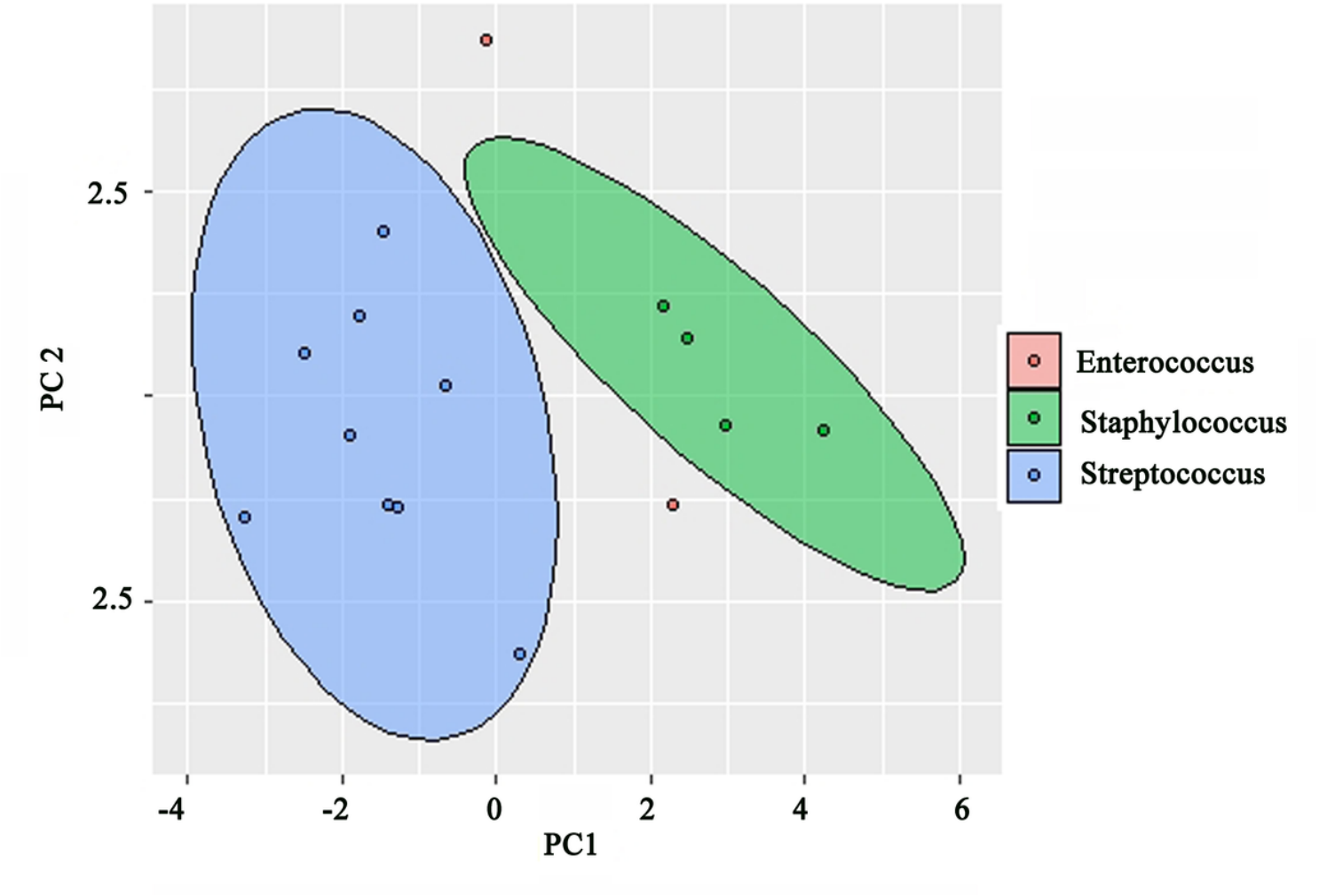

Figure 2

Figure showing close clustering of *Streptococcus* and *Staphylococus* bacteria in PCA plot on the basis of similar amino acids encoded by tri-nucleotide SSRs

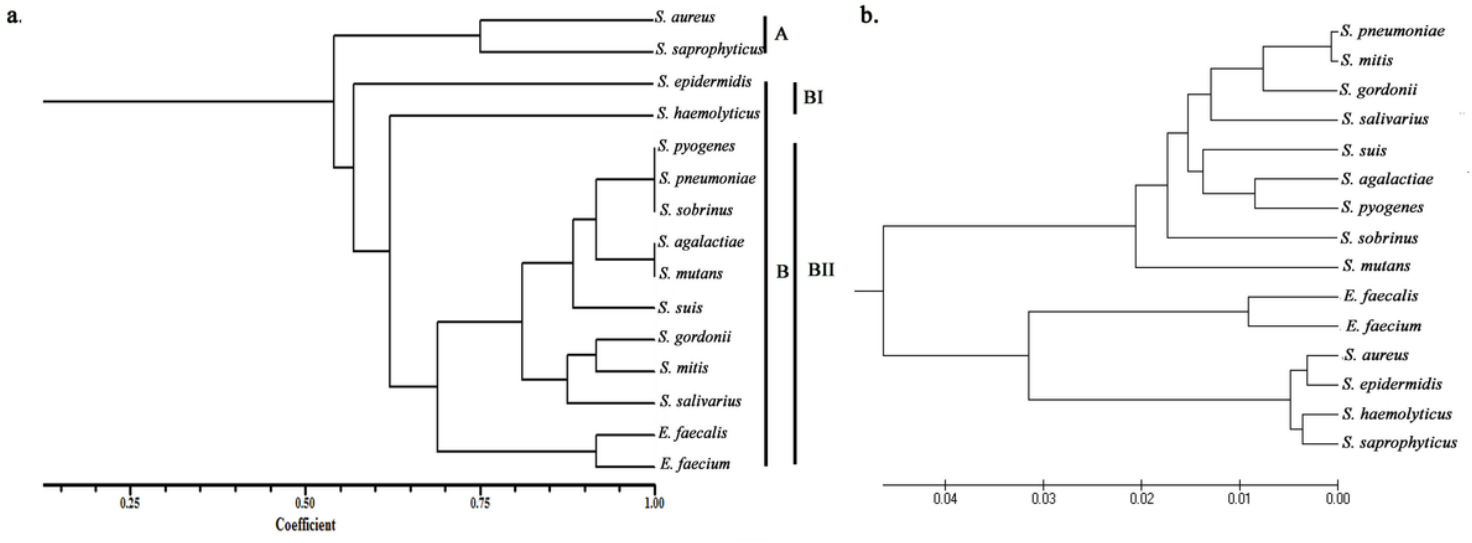

Figure 3

Figure showing phylogenetic relationship among different human pathogenic bacteria on the basis of SSRs present in housekeeping genes (**a**) 16S rRNA sequences (**b**)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytable.docx