BMC Genomics

RESEARCH ARTICLE Open Access

Comparative genomics of Bacteria commonly identified in the built environment

Nancy Merino[1,2], Shu Zhang[3,4], Masaru Tomita[5,6] and Haruo Suzuki[5,6*]

Abstract

Background: The microbial community of the built environment (BE) can impact the lives of people and has been studied for a variety of indoor, outdoor, underground, and extreme locations. Thus far, these microorganisms have mainly been investigated by culture-based methods or amplicon sequencing. However, both methods have limitations, complicating multi-study comparisons and limiting the knowledge gained regarding in-situ microbial lifestyles. A greater understanding of BE microorganisms can be achieved through basic information derived from the complete genome. Here, we investigate the level of diversity and genomic features (genome size, GC content, replication strand skew, and codon usage bias) from complete genomes of bacteria commonly identified in the BE, providing a first step towards understanding these bacterial lifestyles.

Results: Here, we selected bacterial genera commonly identified in the BE (or "Common BE genomes") and compared them against other prokaryotic genera ("Other genomes"). The "Common BE genomes" were identified in various climates and in indoor, outdoor, underground, or extreme built environments. The diversity level of the 16S rRNA varied greatly between genera. The genome size, GC content and GC skew strength of the "Common BE genomes" were statistically larger than those of the "Other genomes" but were not practically significant. In contrast, the strength of selected codon usage bias (S value) was statistically higher with a large effect size in the "Common BE genomes" compared to the "Other genomes."

Conclusion: Of the four genomic features tested, the S value could play a more important role in understanding the lifestyles of bacteria living in the BE. This parameter could be indicative of bacterial growth rates, gene expression, and other factors, potentially affected by BE growth conditions (e.g., temperature, humidity, and nutrients). However, further experimental evidence, species-level BE studies, and classification by BE location is needed to define the relationship between genomic features and the lifestyles of BE bacteria more robustly.

Keywords: Built environment, Bacteria, Diversity, Genomic features, Genome size, GC content, Replication strand skew, Codon usage bias

Background

The microbial community of the built environment (BE) is an important player in human-microbe interactions. As such, in order to build urban environments that benefit human well-being, it is necessary to study the relationship between the BE and microbial communities. As of 2016, about 54% of the world's population is living in urban areas [1], and by 2050, this number is expected to increase to 66% [2]. Moreover, people spend about 87% of their time indoors and about 6% in cars [3], suggesting that the indoor microbial community can play an important role in the lives of individuals. In fact, the indoor microbial community has already been shown to affect occupant health (e.g., respiratory health [4] and asthma [5]), including adverse effects on mental health [6], and can be influenced by building design (e.g., ventilation), occupants, and usage [7–9]. In turn, individuals can easily influence the surrounding microbial community with their own personal microbiome, especially through physical contact [10–12] and movement [13], leaving a microbial fingerprint in the built environment [9, 14, 15]. The microbial community of the BE also

* Correspondence: haruo@sfc.keio.ac.jp
[5]Faculty of Environment and Information Studies, Keio University, Fujisawa, Kanagawa 252-0882, Japan
[6]Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0035, Japan
Full list of author information is available at the end of the article

extends to the outdoor (e.g., green roofs [16] and parks [17]), underground (e.g., transit systems [18–20]), and extreme environments (e.g., cleanrooms [21] and space [21, 22]).

The BE microbiome is slightly influenced by environmental conditions, mainly temperature, humidity, and lighting [23–28]. Several other building parameters have been tested previously (e.g., room pressure, CO_2 concentration, surface material) but were not found to play a significant role in the microbial community composition [29, 30]. Moisture levels are widely known to affect microbial abundances and activity, especially when water damage occurs (e.g., flooded homes had higher abundances of *Penicillium* [31]). However, many indoor built environments are largely devoid of water and nutrients, and it is likely that geographical location, on the scale of cities or even at larger scales [32], plays a more important role in the microbiome composition [30].

The relationship between humans and microorganisms in the BE has moved from investigations limited to culture-based methods to approaches involving next-generation sequencing. One of the first publications on an indoor microbial community occurred in 1887 [33], which expounded a positive correlation between the presence of indoor microorganisms and death rate. Since the advent of high-throughput sequencing, several studies have used amplicon sequencing to gain more information about the microbial community of the BE, including the ribosomal RNA region (e.g., 16S rRNA) for Bacteria and Archaea and the internal transcribed spacer (ITS) region for Fungi [29]. The microbial communities of a variety of locations have been analyzed, such as clean rooms [21], operating rooms [34], plumbing systems [35], universities [36], and transit systems [18–20]. While these studies have enhanced our understanding of the relationship between humans, microorganisms, and the built environment [25, 29, 37], there are limitations to amplicon sequencing, including bias with sequencing primers, targeted amplicon region, DNA extraction protocols, and sequencing platforms [38], which make multi-study comparisons difficult.

Improving our understanding of microbial communities in the BE can be achieved by analyzing draft or complete genomes derived from genomic and metagenomic studies [39]. There have been several published genomes of bacteria collected from the BE, such as *Dermacoccus nishinomiyaensis* [40], *Arthrobacter* sp. [41], and *Gordonia* sp. [42], among others [43–53]. These data provide detailed information on individual bacterial genomes and can be indicative of a bacteria's lifestyle or ecological niches [54, 55]. For example, comparative genomics of *Lactobacillus* species, a common microorganism in the human vagina which is mostly absent from other habitats, revealed that the genomes of the vaginal species were smaller with lower GC (guanine and cytosine) content compared to the non-vaginal species [56]. The observed genome size reduction suggests that the vaginal *Lactobacillus* species has "some degree of adaptation to a host-dependent lifestyle" and is commonly observed in symbiotic microorganisms [56]. However, the individual organismal genome information (e.g., genome size and nucleotide composition) has not been investigated in depth for microorganisms in the BE.

In the present study, we performed genome sequence analyses for bacteria that have been commonly identified in BEs, and focused on genomic features, including genome size, GC content, replication strand skew, and codon usage bias. This information could be useful for the characterization of the microbial members present in BEs, and in the future, these basic features might be useful to help predict the microorganisms likely to adapt to BE conditions.

Results

Bacteria commonly identified in the built environment

Built environments (BEs) are occupied by various microorganisms and are also important transitions that link the natural world, humans, and the urban environment. The indoor microbiome has already been shown to influence human health [4–6], and a building's design and operation can play a major role in the spread of microorganisms, including pathogens [25]. For example, air and water via ventilation and plumbing systems, respectively, are major routes for microbial dispersal throughout a BE [25]. Since BEs are designed to improve the lives of the individuals cohabiting them, it is important to understand the relationship between the BEs and the microorganisms therein.

In this study, we selected 28 bacterial genera that have been commonly identified in the BE at the genera level from 54 publications (Additional file 1: Table S1–S2), ranging from various locations around the world (Additional file 2: Figure S1) and covering four major BE locations (indoor, outdoor, underground, and extreme), several sub-locations (e.g. hospital, residential, recreation, space, subway, and cleanroom), climates, and 3 sample types (surface, air, and water) (Table 1, Additional file 1: Table S3-S5). The International Space Station (ISS) is included as a built environment located in space (or low Earth orbit), and the microorganisms observed in this location would be affected by microgravity and increased radiation. The list of common BE bacterial genera ("Common BE genera") was obtained by selecting genera that have been identified in over 10% of the total publications ($n \geq 6$ publications) and

Table 1 Locations in the BE where "Common BE genera" were identified. The locations where "Common BE genera" were identified are listed for the 28 genera. This list is based on the 54 publications used for this study (see Additional file 1: Table S2)

Bacterial Genera	Environment Type in BE	Ref
Acinetobacter	Clinical (e.g., hospitals), Residential (e.g., bathroom), Extreme (e.g., spacecraft, cleanroom. ISS), Subway (e.g., underground touchscreens). Public recreation (e.g., gym), Hotel bathroom, Office workspace, University (e.g., classroom)	[18, 30, 32, 36, 133–144]
Arthrobacter	Extreme (e.g, cleanroom, ISS), Residential dust, Subway air	[21, 22, 141, 145–149]
Bacillus	Clinical (e.g., hospitals), Residential (e.g., bathroom), Extreme (e.g., spacecraft, cleanroom, ISS), Subway, Public recreation (e.g., gym), Hotel bathroom, Office workspace	[18, 21, 22, 133, 135, 136, 138, 139, 141–143, 147, 148, 150–156]
Bradyrhizobium	Extreme (e.g., spacecraft, cleanroom, ISS), Residential (e.g., wall surfaces), Clinical (e.g., hospital bathroom), Office workspace, Hotel bathroom	[135, 136, 139, 141, 150, 156–159]
Brevundimonas	Clinical (e.g., hospital), Extreme (e.g., spacecraft, cleanroom, ISS), Subway, University classroom	[12, 18, 21, 22, 135, 141, 143, 157, 160]
Burkholderia	Extreme (e.g., spacecraft, cleanroom, ISS), Residential (e.g., bathroom), Clinical (e.g., hospital), Hotel, bathroom Public recreation (e.g., park, gym)	[112, 136, 137, 141, 146, 149, 155, 161]
Clostridium	Residential (e.g., kitchen), Extreme (e.g., cleanroom, ISS), Subway	[20, 21, 138, 141, 154, 156, 162]
Corynebacterium	Clinical (e.g., hospitals), Residential (e.g., dust), Extreme (e.g., spacecraft, cleanroom, ISS), Subway (e.g., ticketing machines, underground touchscreens), Office, workspace University (e.g., classroom, dormitory)	[10, 12, 20, 21, 32, 36, 133, 135, 138, 140–145, 150, 153–156, 158, 163–166]
Delftia	Extreme (e.g., spacecraft, cleanroom, ISS), Clinical (e.g., hospital)	[21, 135, 137, 139, 141, 146, 156, 157, 160, 163]
Enterobacter	Extreme (e.g., spacecraft, ISS), Subway (e.g., outdoor and underground surfaces), University (e.g., classroom,)	[18, 135, 137, 156, 164, 167]
Enterococcus	Extreme (e.g., cleanroom, ISS), Clinical (e.g., hospital), Subway (e.g., outdoor and underground surfaces), Public recreation (e.g., park)	[18, 30, 137, 139, 142, 153, 161]
Escherichia	Clinical (e.g., hospitals), Residential (e.g., kitchen, bathroom), Extreme (e.g., ISS), Subway (e.g., passenger area), Public recreation (e.g., gym), Hotel bathroom	[136, 137, 153, 157, 160, 162, 168, 169]
Kocuria	Residential (e.g., indoor surface), Extreme (e.g., cleanroom, ISS), Subway (e.g., underground air), Clinical (e.g., hospitals)	[21, 22, 112, 138, 142–144, 147, 148, 167]
Lactobacillus	Clinical (e.g., nursing home), Residential (e.g., indoor air, surface dust), Extreme (e.g., cleanroom, ISS), Subway (e.g., touchscreens), Office workspace) University (e.g., classroom, dormitory, bathroom)	[10, 12, 13, 21, 30, 32, 36, 141, 144, 145, 150, 156, 165, 166]
Methylobacterium	Clinical (e.g., hospitals), Residential (e.g., bathroom), Extreme (e.g., spacecraft, cleanroom, ISS), Subway (e.g., touchscreens), Office (e.g., dust), University (e.g. door handle), Hotel bathroom	[13, 21, 22, 32, 36, 135, 136, 139, 141, 144, 146, 154, 157, 163, 167, 170]
Microbacterium	Extreme (e.g., spacecraft, cleanroom, ISS), Subway (e.g., underground air)	[21, 22, 135, 139, 147, 148]

Table 1 Locations in the BE where "Common BE genera" were identified. The locations where "Common BE genera" were identified are listed for the 28 genera. This list is based on the 54 publications used for this study (see Additional file 1: Table S2) *(Continued)*

Bacterial Genera	Environment Type in BE	Ref
Micrococcus	Clinical (e.g., hospitals), Residential (e.g., indoor air, surface), Extreme (e.g., spacecraft, cleanroom, ISS), Subway (e.g., underground air)	[20, 22, 32, 112, 133, 135, 138, 142–144, 147, 148, 153, 155, 167]
Mycobacterium	Clinical (e.g., hospitals), Residential (e.g., indoor air, surface), Extreme (e.g., cleanroom), Subway (e.g., outdoor air), Hotel (e.g., showerhead), Public recreation (e.g., gym)	[134, 136, 137, 141, 143, 144, 159, 163, 171–173]
Neisseria	Clinical (e.g., hospitals), Residential (e.g., dust), Extreme (e.g., ISS), Hotel (e.g., showerhead), Public recreation (e.g., gym), Office workspace	[136, 145, 150, 154, 156, 158]
Paenibacillus	Extreme (e.g., space station, ISS), Subway (e.g., underground air)	[21, 22, 135, 147, 148, 156]
Prevotella	Residential (e.g., wall surface, dust), Extreme (e.g., ISS), Office workspace, University (e.g., dormitory)	[21, 32, 145, 150, 158, 165]
Propionibacterium	Clinical (e.g., nursing home), Residential (e.g., kitchen, bathroom), Extreme (e.g., cleanroom, space station), Subway (e.g., indoor air), University (e.g., classroom, door handle)	[20, 21, 32, 36, 133, 143, 144, 151, 154, 156, 157, 164–166]
Pseudomonas	Clinical (e.g., hospitals), Residential (e.g., kitchen, bathroom), Extreme (e.g., cleanroom, space station, ISS), Subway (e.g., underground air), University (e.g., door handle), Hotel (e.g., showerhead), Public recreation (e.g., gym), Office (workspace)	[18, 21, 22, 30, 36, 112, 134–137, 141–144, 148–151, 153, 156, 163, 172, 174, 175]
Ralstonia	Clinical (e.g., hospitals), Residential (e.g., indoor air), Extreme (e.g., cleanroom, space station, ISS)	[135, 139, 141, 146, 149, 152, 157, 167]
Sphingomonas	Clinical (e.g., hospitals), Residential (e.g., bathroom), Extreme (e.g., cleanroom, space station, ISS), Subway (e.g., ticketing machines, underground touchscreens), University (e.g., classroom), Hotel (e.g., showerhead), Public recreation (e.g., gym, park, parking lot), Office (e.g., dust)	[13, 21, 22, 32, 36, 134–138, 140, 141, 144, 146, 149, 154, 156, 157, 159, 161, 163, 164, 170]
Staphylococcus	Clinical (e.g., hospitals), Residential (e.g., bathroom), Extreme (e.g., cleanroom, space station, ISS), Subway (e.g., air), University (e.g. classroom), Hotel (e.g., showerhead), Public recreation (e.g., gym), Office workspace	[12, 13, 20–22, 30, 32, 36, 112, 133, 135–145, 148, 151–156, 160, 163, 164, 166–168, 171, 173]
Stenotrophomonas	Clinical (e.g., hospitals), Extreme (e.g. cleanroom, space station, ISS), Subway (e.g. ticketing machines, underground touchscreens)	[18, 21, 22, 141, 149, 157, 160]
Streptococcus	Clinical (e.g., hospitals), Residential (e.g., bathroom, wall surface), Extreme (e.g., cleanroom, ISS), Subway (e.g., indoor air, touchscreens), University (e.g., classroom, door handle), Hotel (e.g., showerhead), Public recreation (e.g., gym), Office (e.g., dust, workspace)	[12, 13, 32, 36, 133, 136, 137, 139, 141, 142, 144, 145, 150, 151, 153, 154, 156, 158, 160, 164, 166, 171, 173]

have at least one completed genome in the NCBI RefSeq database ($n = 28$ genera) (Additional file 1: Table S1). The "Common BE genera" and their identified locations in the BE are summarized in Table 1.

From the 54 publications used in this study, many of the "Common BE genera" (Table 1) were identified around the world (Additional file 2: Figure S1). For example, *Acinetobacter* was found in five countries, spanning eight different climates, and in the ISS. Unsurprisingly, all 28 genera had some association with humans, as analyzed by MetaMetaDB (Additional file 1: Table S6) [57], further demonstrating the influence that humans have on the BE microbiome [29, 37]. Due to the limitations of this study, the prevalence of these "Common BE genera" cannot yet be associated with BE selection pressures. For example, while there are several other human-associated genera (e.g., *Haemophilus, Veillonella, Alistipes, Rothia*), the microbial community abundances could be affected by different abundance levels and shedding rates across the human body. Other limitations are listed in the section "Robustness and limitations."

Diversity among common BE genera

To assess the diversity of the "Common BE genera," we calculated the mean distance (Dmean) between all pairs of taxa within each genus based on 16S rRNA gene sequences available in the LTP datasets of the SILVA v128 release [58]. The SILVA database was selected over other 16S rRNA databases (e.g. Greengenes [59, 60] and RDP [61]) due to greater alignment quality [62] and because it is continuously updated [63]. The Dmean was also selected over the phylogenetic diversity index (PD) [64, 65] because it is less affected by the number of taxa (N) available in the LTP database, as demonstrated by a smaller Pearson correlation coefficient ($r = 0.0017$) between N and Dmean compared to N and PD ($r = 0.7248$) (Additional file 2: Figure S2).

The Dmean for each "Common BE genus," with $n > 2$ in the LTP database ranged from 0.005 (*Ralstonia*) to 0.038 (*Clostridium*) with a median value of 0.015 (Fig. 1, Additional file 1: Table S7), suggesting, for example, that taxa within *Ralstonia* are relatively more closely related than those in *Clostridium*. In comparison, the Dmean for genera not commonly found in the BE (850 genera)

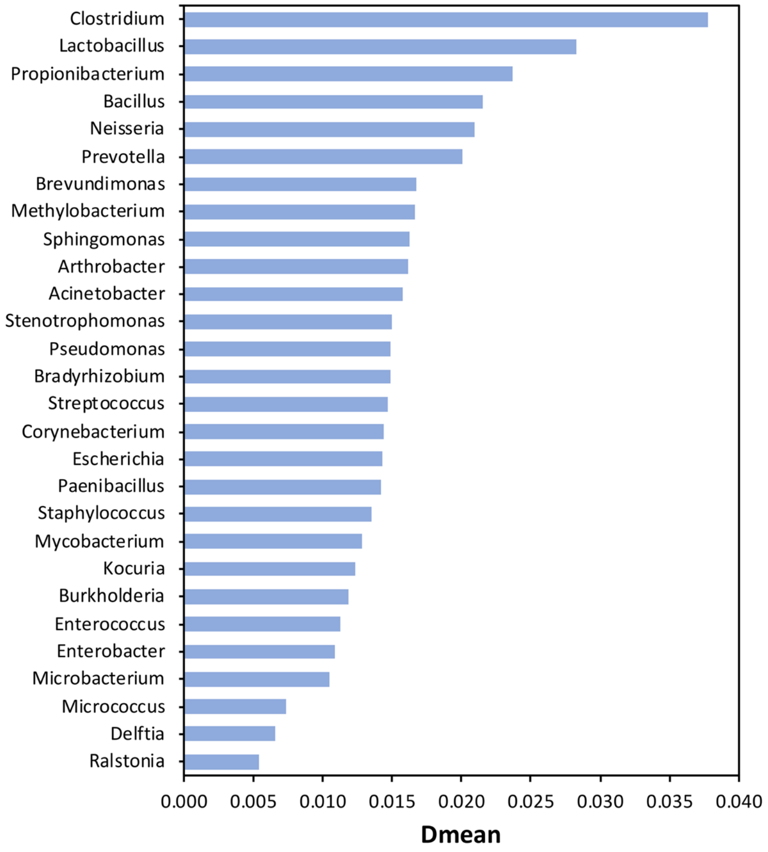

Fig. 1 Diversity levels in 16S rRNA gene sequences for each bacterial genus commonly found in the built environment. The mean distance (Dmean) between all pairs of bacteria was used as a diversity index [58]

ranged from 0 (*Stigmatella*) to 0.115 (*Salinibacter*) with a median value of 0.016. A Wilcoxon rank sum test, which compared the Dmean values between the two groups (28 genera versus 850 genera), was not statistically significant (*p*-value = 0.28). This indicates that there was insufficient evidence to conclude that there was a significant difference in intra-genus diversity between "Common BE genera" and "Other genera." However, the 16S rRNA gene has its limitations (e.g., sequence heterogeneity [66] and horizontal gene transfer [67]), even though it is widely used as a molecular clock to understand evolution [67–70]. Intragenus variations in genomic features (genome size, GC content, GC skew, and codon usage bias) can reflect the level of diversity among taxa within each of the "Common BE genus."

Genome size, GC content, and GC skew

We compared the genomic features (genome size, GC content, GC skew, and codon usage bias) of 2580 complete prokaryotic genomes from the NCBI RefSeq database, in which 717 genomes are from bacteria commonly identified in the BE ("Common BE genera") and 1863 other genomes ("Other genera") (Additional file 1: Table S8-S9). The "Other genomes" have not been identified in at least six publications (equivalent to 10% of the publications used for this study).

Genomic features, including genome size, GC content, and GC skew, can provide information about the bacterial lifestyle as well as phylogeny [54]. For example, genome size can reflect genome streamlining, symbiosis, or genome expansion [71, 72]. GC content has been shown to relate to both the phylogeny and ecological adaptations of a microbial species, as demonstrated by Reichenberger and co-workers [73]. GC content can range from 15 to 75% and can be influenced by environmental factors such as temperature [74], oxygen levels [75], and nucleotide availability [76]. Furthermore, GC skew, as quantified by the GC skew index (GCSI), measures the strength of replication strand skew [77] and could indicate variation in mutational and selective pressures between leading and lagging strands of DNA replication [78]. Indeed, the leading strand tends to be biased with G and T while the lagging strand is rich in A and C [79]. Strand composition bias has been shown to especially occur in obligate intracellular microorganisms that permanently live within a host, resulting in the loss of some DNA repair genes and the accumulation of mutations [80]. Replication, repair, and transcription enzymes are thought to influence strand composition, where different

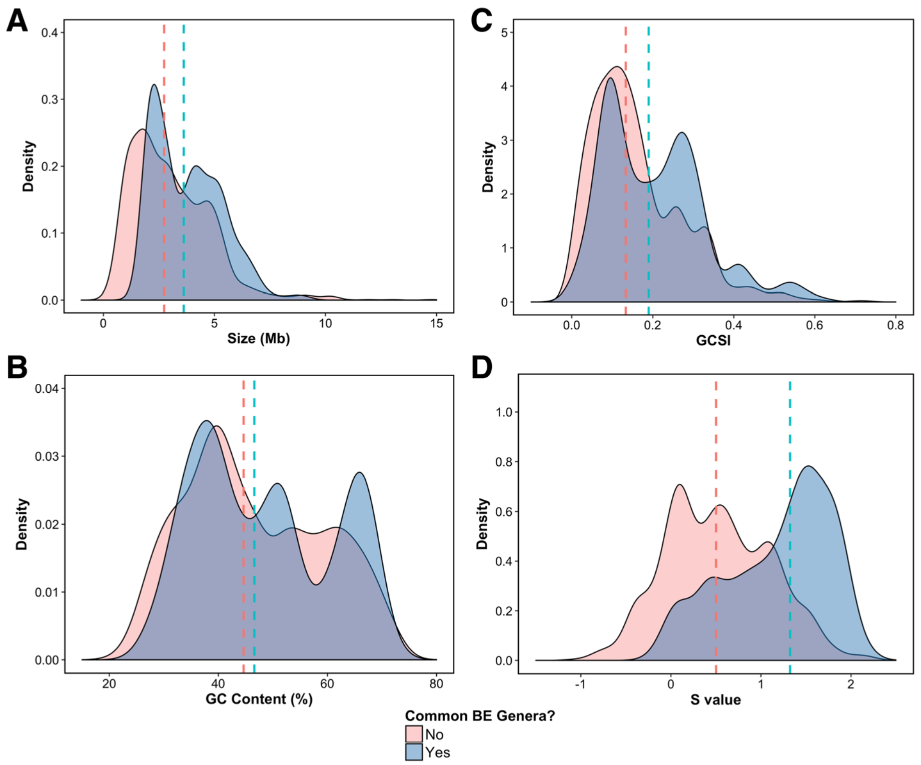

Fig. 2 Density plots comparing "Common BE genomes" against "Other genomes." **a** Genome size, **b** GC Content (%), **c** GCSI, and **d** S value. The dashed lines indicate the median value for "Common BE genomes" (blue) and "Other genomes" (pink). The S value is the only significant genomic feature when comparing "Common BE genomes" against "Other genomes"

genes are involved in transcribing the leading and lagging strand [81]. Each enzyme will have different mutational and selective pressures, and thus, GCSI informs DNA repair capabilities and provides insight into the metabolism and lifestyle of bacteria [81].

The "Common BE genomes" tended to have larger genome sizes (1.30–9.21 Mb, median 3.62 Mb) (Fig. 2a), higher GC contents (27.4–73.0%, median 46.6%) (Fig. 2b), and higher GCSI (0.007–0.629, median 0.19) (Fig. 2c) compared to the "Other genomes". Among the 717 "Common BE genomes," the bacterium, *Clostridium perfringens* strain 13 (NC_003366), had the highest GCSI value (0.629) and exhibited a clear GC skew, especially around 1.4 Mb (Additional file 2: Figure S3A), while *Methylobacterium* sp. 4–46 (NC_010511) had the lowest GCSI value (0.007) with indiscernible GC skew (Additional file 2: Figure S3B). The median for all three features of the "Common BE genomes" was higher than that of the "Other genomes" (1863 genomes; Size = 2.74 Mb; GC content = 44.6%; GCSI = 0.133) (Fig. 2). While these differences were statistically significant based on the Wilcoxon rank sum test q-value (genome size = 1.68e-31; GC content = 0.002; GCSI = 6.46e-17), further analysis using the Cliff's delta effect size (genome size = 0.3, GC content = 0.079, GCSI = 0.215) demonstrated negligible (< 0.147) or small (< 0.33) thresholds when comparing

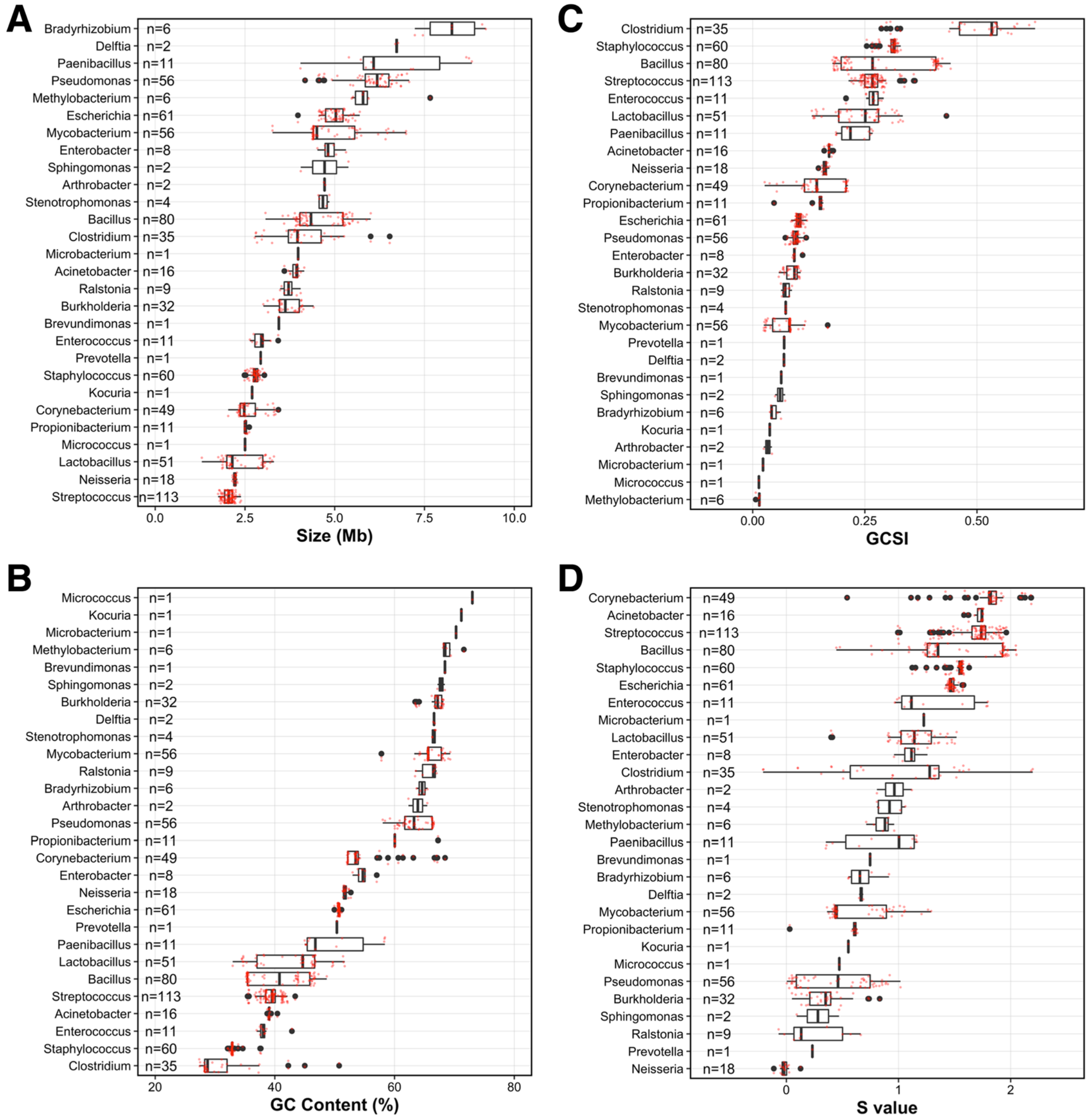

Fig. 3 Distribution of each genome feature for the "Common BE genera." Box-and-whisker plots displaying the distributions of genomic features for each "Common BE genus" based on a five number summary (minimum, 25th percentile, median, 75th percentile, and maximum). Outliers are plotted as black filled circles. Red circles denote individual genomes. Genomic features are as follows: **a** Genome size, **b** GC Content (%), **c** GCSI, and **d** S value

the "Common BE" and "Other" genomes. Similar results were observed when categorizing by environments (MetaMetaDB) (Additional file 2: Figure S4–S6). Moreover, each genome feature may cover a wide range (Fig. 3a-c), depending on the BE genus.

Codon usage bias

The genetic code of each "Common BE genus" can also provide information about codon usage bias, which has further implications on evolutionary processes, such as selection, mutation [82], and even horizontal gene transfer [83–85]. Many amino acids can be encoded by more than one codon, also known as synonymous codons, due to the redundancy of the genetic code, and there is generally a preference for one synonymous codon over another [86]. The pattern of synonymous codon usage can vary between organisms (e.g., some organisms use a set of synonymous codons more frequently) and across genes within a genome [82, 87]. It is hypothesized that codons are selected based on their impact on translation, influencing bacterial growth [88, 89], and that codon usage bias can be derived from highly expressed genes [90, 91]. Several studies have demonstrated that codon usage bias correlates with bacterial growth rates, likely suggesting a selection towards efficient translation machinery [87, 89, 92, 93]. Codons may also be selected to optimize protein production speed [94]. For example, the codon usage bias of *Salmonella enterica* serovar Typhimurium, a fast-growing bacterium, correlates well with gene expression levels [87]. Thus, it is imperative to determine the codon usage bias in order to further surmise the lifestyles of bacteria that have been commonly identified in the BE.

Here, we determined the strength of selected codon usage bias (S value) (Fig. 2d), as discussed by Sharp and co-workers [87]. The S value is based on a comparison of codon usage between constitutively highly expressed genes and the entire genome (see Methods for details) [87]. The median S value of the "Common BE genomes" (1.32) was higher than that of the "Other genomes" (0.50), with a large effect size (Cliff's delta of 0.574). Moreover, the Wilcoxon rank sum test provided a significant result with a q-value of 1.22e-111, suggesting that the S value could be more indicative of the type of bacteria commonly observed in the BE compared to other genomic features described previously (genome size, GC content, and GC skew).

Further categorization of the environments (MetaMetaDB) indicates that the S value is stronger for the "Common BE genomes" observed with the human microbiome, as compared to the other "Common BE genomes" (Additional file 1: Table S10 and Additional file 2: Figure S7). Among the 517 "Common BE genomes" for which species were categorized according to environments in MetaMetaDB, the S value tended to be lower in compost-associated "Common BE genomes" than in the other "Common BE genomes" (Cliff's delta = − 0.647; q-value = 1.01e-21). In contrast, the median S value for the "Common BE genomes" also associated with the category "human" by MetaMetaDB (*n* = 454; median S value = 1.45) was higher than that for the other "Common BE genomes" (*n* = 63; median S value = 0.71). The difference was large based on the effect size (Cliff's delta = 0.516) and was statistically significant based on the Wilcoxon rank sum test (q-value = 2.53e-10). This trend is also true when examining only the top bacterial genera found in the human microbiome (list taken from Lloyd-Price J, Mahurkar A, et al. [95]). The top human microbiome genera that are also commonly found in the BE (*n* = 301 genomes; median S value = 1.50) had significantly higher S values compared to those not commonly found in the BE (*n* = 28 genomes; median S value = 1.08) with a medium effect size (Cliff's delta of 0.451) and a q-value of 0.0009. This suggests that the human and BE microbiome are interconnected, with bacterial genera trending towards larger S values. However, the limitations of this study (see section "Robustness and limitations") cannot associate the "Common BE genera" with BE selection pressures.

When examining each "Common BE genus," the S value was found to cover a wide range (e.g., *Enterococcus*, *Mycobacterium*, and *Bacillus*) (Fig. 3d). Future reports of BE microbial communities could help to resolve the importance of the S value by accurately identifying taxa to the species level and by unifying metadata collection and method protocols. Indeed, the S value has been shown to vary across species, especially for those that are not closely related [96]; e.g., *Clostridium* has the largest S value range (Fig. 3d) and also has the largest Dmean (0.038) (Fig. 1).

Case study: *Mycobacterium*

As a case study for one of the "Common BE genera", we further discuss *Mycobacterium* and describe how the four genomic features can be used to surmise the potential lifestyle of bacteria. *Mycobacterium*, a genus with well-known pathogenic species (e.g., *Mycobacterium tuberculosis* and *Mycobacterium bovis*), has one of the largest genome size ranges from 3.3 Mb [*Mycobacterium leprae* Br4923 (NC_011896)] to 7.0 Mb [*Mycobacterium smegmatis* strain MC2 155 (NC_008596)] with a median of 4.5 Mb (Fig. 3a). *Mycobacterium* has been found in several locations, including hospitals, therapy pools, showerheads, water-damaged homes, and cleanrooms (Table 1). One of the major factors determining the presence of *Mycobacterium* in water-damaged homes may be due to transmission from human and pet occupants [32]. The GC content in *Mycobacterium* was relatively high (57.8–69.3%) compared to other "Common BE genera" (27.4–73.0%) (Fig. 3b), where the outlier group (57.8%) was the species *M. leprae* (Additional file 1: Table S8). The smaller genome size and lower GC content of *M. leprae*, an obligate pathogen, are a result of genome reduction which has been well documented [97]. The GCSI ranged from 0.025 [*M. avium* subsp. *paratuberculosis* K-10 (NC_002944); Additional file 2: Figure S8A] to 0.167 [*M. leprae* Br4923 (NC_011896); Additional file 2: Figure S8B]. The S value for *Mycobacterium* ranged from 0.36–1.30, suggesting that either the growth rate of different *Mycobacterium* species present in the BE varies drastically or that some *Mycobacterium* species have more "volatile" codons, as discussed below. For example, *M. tuberculosis* and *M. leprae* have S values in the lower range (0.36–0.45) and also have slow generation times of ~ 1 and 14 d, respectively [87, 98, 99]. In comparison, one of the highest S values (1.3) corresponded to *M. abscessus*, which has a generation time of 4–5 h [100].

Discussion

Genomic features relation to the potential lifestyle of bacteria commonly identified in the built environment

To further understand the 28 "Common BE genera," we analyzed four genomic features: genome size, GC content, GC skew, and codon bias. While our study based itself on the results of previous studies to retrieve the "Common BE genera," we aimed to demonstrate the potential of using genomic features to provide insight into microbial lifestyles and to describe the trends found in the "Common BE genera" [54]. The "Common BE genomes" tended to have larger genome sizes, higher GC contents, higher GCSI, and larger S values compared to the "Other genomes." While the differences for all the genomic features were statistically significant based on the Wilcoxon rank sum test, further analysis by the Cliff's delta effect size demonstrated that the S value is likely a more important genomic feature for bacteria commonly identified in the BE compared to the "Others" analyzed in this study.

This initial analysis could help begin to surmise certain lifestyles of the bacteria commonly found in the BE. For example, the S value has implications on the growth rates of bacteria [89] found in the BE, which may be higher than those found in other environments, and could also be related to higher levels of gene expression [90, 91]. A stronger preference for codon usage bias in the "Common BE genera" may have resulted from a of long-term relationship with humans (e.g., genome reduction in bacteria was associated with the "Neolithic revolution" [101] and "Common BE genera" were found on nineteenth century documents [102, 103]) but further analysis is needed.

Moreover, the preference for certain codons may be related to either directional mutation or specific selection [104]. In the case of directional mutation, it is hypothesized that some codons are more prone to mutation, resulting in lower S values [87]. For example, *Mycobacterium tuberculosis*, one of the "Common BE genera" and pathogen with S values (0.41–0.45) below the "Common BE" and "Other" genome medians (Fig. 3d), has more "volatile" codons relating to antigens, surface proteins, or antibodies which are likely to mutate more than other codons [105]. These help *M. tuberculosis* prevent host-immune system interactions [105]. As for specific selection, it is thought to lead to efficient translation processes and accurate protein synthesis due to the use of more frequent codons by highly expressed genes [104]. This can be a reflection of an organism's adaptation to an environment, and it is likely that the "Common BE genomes" share "synchronized regulation mechanisms of translational optimization" [106]. Indeed, this has been shown for 11 distinct metagenomes from various environments [106], where, for example, microorganisms living with an abundant food source (whale fall carcass) have translationally optimized genes for energy production and conversion.

The trend towards larger S values in the "Common BE genera" also suggests that these genera can inhabit a wide range of environments [107]. The "Common BE genera" must also contend with chemicals derived from the daily use of personal care and household products (e.g., avobenzone from sunscreen, laureth sulfate from shampoo, and amlodipine from medication used to treat high blood pressure), in addition to human-derived chemicals (e.g., acyl glycerols, which make up the membrane of human cells) [108–110]. For example, *Propionibacterium* has been shown to metabolize triglyceride triolein, a human acylated glycerol, and was found to be co-localized with acylated glycerols on the human body [108]. Since these chemicals can be found in the BE and may be associated with an occupant's chemical signature [109], future studies are needed to determine how these chemicals may affect the BE microbial community composition (e.g., rural vs. urban environments, change in a product's formula, etc.).

While not as important as the S value in this study, larger genome sizes could be attributed to the incorporation of regulatory and secondary metabolic genes [72], which may be important for survival in the BE (e.g., aromatics degradation and regulation to environmental stresses). Indeed, the top three major functional pathways annotated for the microbial community found in ambulances were 1) biosynthesis of cofactors, prosthetic groups, and electron carriers, 2) secondary metabolites biosynthesis, and 3) aromatics compound degradation [111].

Robustness and limitations

This study demonstrates the potential of using the four genomic features (genome size, GC content, GCSI, and S value) to surmise the lifestyle of bacteria. The "Common BE genera" selected in this study have only been commonly identified by culture-based and amplicon-based sequencing studies, which have limitations as described in the Introduction. Although the "Common BE genera" have been detected in multiple BE studies (≥ 6), these bacteria may not be active in the BE. Moreover, although this study is based on completed genomes from the NCBI RefSeq database, the genomes could have been derived from environments not related to the BE. Thus, the conclusions derived from this study serve as a hypothesis for the potential lifestyles of commonly identified BE bacterial genera. Further studies are needed to accurately determine the typical BE genera and the association of BE genera with BE selection pressures.

It is important to note that the results remained similar when different data sets were compared (Additional file 1: Table S9). We tested the robustness to the composition of the genome data set by testing different subsets of bacteria (e.g., phyla of *Proteobacteria*, *Firmicutes*, and *Actinobacteria*), and also by randomly selecting one representative for species that have multiple strains sequenced. Of the four genomic features (genome size, GC content, GCSI, and S value), only the S value showed consistent results and tended to be higher in the "Common BE genera" compared to the "Others." This indicates that the selected codon usage bias tends to be stronger in the "Common BE genera" than in the "Other genera," regardless of the datasets used, and that our results were less affected by biases in the available sequenced genomes. We also tested different numbers of publications ($n = 1, 2, 3, 4, 5,$ and 6) to select for BE genera. The corresponding numbers of the selected "Common BE genomes" were 1208, 1029, 922, 825, 739, and 717. Even when genera observed in at least 1 out of 54 publications were defined as the "Common BE genera," the median S value for the "Common BE genomes" (1.14) was higher than that for the "Other genomes" (0.35) with a large effect size (Cliff's delta of 0.548), and the Wilcoxon rank sum test returning significant result with q-value of 2.59e-126. This is consistent with the results obtained by larger numbers of publications ($n > 1$) to define the "Common BE genera." Thus, selected codon usage bias tends to be larger in the "Common BE genomes" than in the "Other genomes," regardless of the genome data set used and criteria to define BE genera.

Our selection of the 28 common bacterial genera is likely biased towards the genera found in certain locations (e.g. fewer publications sampling outdoors and subways compared to indoors and extreme; more publications sampling locations with mild temperate climates) (Additional file 1: Table S3–S5) and sampling type (e.g., fewer publications conducted microbial community analysis of water samples compared to surface and air samples) (Additional file 1: Table S3). In addition, 16S rRNA amplicon sequencing was the dominant method used to determine the microbial community amongst the 54 publications used in this study. Some publications also conducted culture-based studies (e.g. study on airborne bacteria in Tokyo [112]). This introduces bias from the range of protocols used across publications, including sample collection methods (e.g. swab, wipe, air, and storage method), DNA extraction methods, primers used, 16S rRNA target region (e.g. V3–V4, V4, V6–V8), and sequencing methods [113–115]. With advances in sequencing for 16S rRNA (e.g., full-length [116]), genomes, and metagenomes (e.g., longer contigs, accurate base calling) and increased global research collaboration (e.g., MetaSUB [117]), more specific classification of BE microorganisms can be obtained at the species level, allowing for more accurate descriptions in future studies.

After obtaining the 28 "Common BE genera," we then used the NCBI RefSeq database to obtain completed genomes. Another level of bias arises from using sequenced genomes from the public database (e.g., towards medically and industrially important microorganisms), although there are ongoing "efforts to expand the bacterial and archaeal reference genomes…to maximize sequence coverage of phylogenetic space" [118]. However, this study aimed to demonstrate the capability of using genomic features to characterize the "Common BE genera," providing a first step towards understanding the potential lifestyles of these bacteria. As more genomes from the BE microbial community are sequenced (e.g., efforts by the MetaSUB International Consortium [117]), much more accurate analyses can be carried out to appropriately examine the microbial lifestyles based on genomic features and functional annotation.

Conclusions

Twenty-eight bacterial genera were selected to represent the bacteria commonly identified in the BE. Although geographical location, temperature, and humidity are important factors in shaping the BE microbial composition, many of the "Common BE genera" were identified around the world. All the genera have also been observed in the human microbiome. Here, we used genomic features to demonstrate the potential of understanding the lifestyle of bacteria from the genome. Together, the genome size, GC content, and GC skew for the "Common BE genomes" showed trends similar to (were not strongly deviated from) those for the entire data set of completed prokaryotic genomes analyzed obtained from the NCBI database. On the other hand, the strength of selected codon usage bias (S value) for the "Common BE genomes" tended to be significantly higher than that of the "Other genomes." As such, the S value could be indicative of bacterial growth rates, gene expression, and other evolutionary processes that may play a role in the bacteria present in the BE. Further insights could be gained through more BE studies analyzing locations with fewer publications (e.g., rural, tropical climates, and outdoor), identifying microbial communities at the species-level, and by minimizing cross-study biases.

Methods

Selection of common BE bacterial genera, metadata, and genome sequence data

Bacteria commonly identified in the BE are listed in Additional file 1: Table S1 and Table 1. Since most currently available BE studies conducted 16S rRNA amplicon sequencing, the identification was largely limited to the genus level. In this study, 54 total publications (published between 2003 and 2017) were compiled with metadata, including the bacterial genera, BE location identified, sample type, temperature (°C), humidity (%), and approximate climate (Additional file 1: Table S2). These publications either conducted 16S rRNA amplicon sequencing or isolated bacteria from the BE. If the temperature or humidity was not described by the publication, the average over a certain period of time (either the timeframe stated in the publication or the publication year) was obtained from online sources (see Additional file 1: Table S2 for references and timeframe). In order to obtain climate level assignment, the Köppen climate classification scheme was implemented (1981–2010) by determining the closest latitude and longitude to a publication's described study location [119] (Additional file 1: Table S4). In order to identify the "Common BE genera," we selected for bacterial genera which were identified in more than about 10% of the publications ($n \geq 6$ publications) and had at least one genome sequenced in the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov) RefSeq database [120, 121] (Additional file 1: Table S8) ($n = 28$ genera). These were denoted as "Common BE genomes" or "Common BE genera" while the bacterial genera not selected were denoted as "Other genomes" or "Other genera." Based on this criterion, 28 genera were retained (Additional file 1: Table S1).

To further understand the potential associated environments of each BE genus, we used MetaMetaDB (data by November 6, 2014 at http://mmdb.aori.u-tokyo.ac.jp) (Additional file 1: Table S6) [57]. MetaMetaDB is a database to search for the possible habitats a microorganism could live in and was made by collecting 16S rRNA sequences. Hits for environmental categories for each common BE genus was based on an identity threshold of 97%, corresponding to the species taxonomic level. Environmental categories on MetaMetaDB are based on the classification used by the NCBI taxonomy, which include categories such as aquatic, soil, human, compost, and more. While these categories are not well-defined and controlled (e.g., there are several categories for human, including human, human gut, human oral, human skin, and others), we used MetaMetaDB to gain insight into the associated environments of each BE genus.

RefSeq chromosome sequence accessions with the NC_ prefix were obtained from the NCBI prokaryotic genome list (ftp://ftp.ncbi.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt), and complete sequences of prokaryotic chromosomes (GenBank format [122]) were downloaded with the RefSeq accessions using E-utilities on 2018-01-27. In cases where the organism has multiple replicons (chromosomes and plasmids), only the largest chromosome was used for the analysis as a representative replicon of the organism. The final data set included 2580 prokaryotic genomes (142 Archaea and 2438 Bacteria), including 717 genomes of bacteria belonging to the 28 genera commonly found in the BE ("Common BE genomes") and 1863 other prokaryotic genomes ("Other genomes"). The 717 "Common BE genomes" belonged to 4 phyla: *Firmicutes* (370), *Proteobacteria* (222), *Actinobacteria* (123), and *Bacteroidetes* (2). The 1863 "Other genomes" belonged to 644 genera from 36 phyla, including *Proteobacteria* (875), *Firmicutes* (192), *Actinobacteria* (115), and *Chlamydiae* (110). The "Common BE genomes" and "Other genomes" were linked to the 18 environmental categories in MetaMetaDB: Aquatic, Biofilm, Compost, Food, Freshwater, Hot_springs, Human, Human_gut, Human_lung, Human_nasal_pharyngeal, Human_oral, Human_skin, Marine, Rhizosphere, Rock, Root, Sediment, and Soil. Complete listings of the genomes used in this study, along with the genomic features, are shown in Additional file 1: Table S8.

Bacterial diversity

To measure the genetic diversity among taxa within a genus, the mean distance (Dmean) between all pairs of bacteria was calculated [58]. The genetic distance between a pair of bacteria was calculated with the K80 model using the 'dist.dna' function of the 'ape' package of R (https://cran.r-project.org/web/packages/ape) [123]. We used a nucleotide sequence alignment of the 16S rRNA genes in 'The All-Species Living Tree' Project (https://www.arb-silva.de/projects/living-tree/) [124]. LTP datasets based on SILVA release 128 were downloaded from the Download page [125].

Genomic features

Genome size

The total number of nucleotides (A + T + G + C) was calculated from the whole nucleotide sequence of each chromosome.

GC content (%)

The relative frequency (percentage) of guanine and cytosine (G + C)/(A + T + G + C) was calculated from the whole nucleotide sequence of each chromosome.

GC skew index (GCSI)

The asymmetry in nucleotide composition between leading and lagging strands of DNA replication is represented by GC skew (C-G)/(C + G). The strength of GC skew was measured by the GC skew index or GCSI [126] with a window number of 4096. This fixed window number was used to prevent any effects from biased nucleotide composition in coding regions and is based on an average gene length of 1 kb and a genome size of 2–4 Mb [126]. The GCSI values can range from 0 (no GC skew) to approximately 1 (strong GC skew).

Strength of selected codon usage bias (S value)

As a measure of translationally selected codon usage bias, the S value was calculated for each chromosome, as described in Sharp and co-workers [87] and Vieira-Silva and Rocha [89], using the codon usage for four amino acids, Phe (TTC and TTT), Tyr (TAC and TAT), Ile (ATC and ATT), and Asn (AAC and AAT). The two codons are recognized by the same tRNA species, and the C-ending codon is recognized more efficiently than T-ending codon. The S value is based on a comparison of codon usage within these synonymous groups between constitutively highly expressed genes (those encoding ribosomal proteins and translation elongation factors) and the entire genome [87, 89].

Statistical analyses

We performed several statistical analyses to compare the values of the genomic features (genome size, GC content, GCSI, and S value) between two groups of genomes: e.g., “Common BE genomes” versus “Other genomes”; and MetaMetaDB environment-associated “Common BE genomes” (e.g., “Human”) versus other “Common BE genomes” (e.g., not associated with “Human”).

Wilcoxon rank sum test

We performed the Wilcoxon rank sum test (also called Mann-Whitney U test) as a non-parametric statistical hypothesis test to compare the values between two groups [127]. The *p*-value obtained by the statistical test was adjusted for multiple comparisons by controlling for the false discovery rate (FDR) [128]. An FDR adjusted p-value (q-value) of 0.05 was used as a threshold for statistical significance.

Cliff’s delta effect size

We calculated Cliff’s delta statistic as a non-parametric effect size to estimate the degree of overlap between two distributions [129]. A Cliff’s delta of 0.0 indicates the group distributions overlap completely, whereas a 1.0 or – 1.0 indicates the absence of overlap between the two groups. A positive Cliff’s delta close to 1.0 indicates that the genomic feature values tended to be higher in the “Common BE genomes” than in the “Other genomes.” A negative Cliff’s delta close to – 1.0 indicates that the genomic feature values tend to be lower in the “Common BE genomes” than in the “Other genomes.” Three thresholds were used to determine the magnitude: $|d| < 0.147$ “negligible,” $|d| < 0.33$ “small,” and $|d| < 0.474$ “medium” or “large” [130]. These thresholds are used for two normal distributions [136], equivalent to the original thresholds used by Cliff (1993) [135] to scale the effect size indices to observable phenomena.

Software

Genome sequence analyses (e.g., calculating genome size, GC content, GCSI, and S value) were performed using the G-language Genome Analysis Environment version 1.9.1 (http://www.g-language.org) [131]. Statistical computing and graph drawing were conducted with R version 3.3.3 (https://www.R-project.org/) [132].

Additional files

Additional file 1: Table S1. Selection of bacterial genera commonly identified in the built environment. Bacterial genera identified in 54 publications were compiled (see Table S2) and commonly identified genera were selected. All bacterial genera identified in more than about 10% of the publications ($n \geq 6$ publications) with at least one complete reference genome on the NCBI RefSeq database were used in this study ($n = 28$ genera). **Table S2.** Metadata for each reference. 54 publications were compiled, including metadata for location, sub-locations, bacterial genera identified, sample type, climate (Table S4 and S5), temperature (° C), and humidity (%). If temperature or humidity was not described by the publication, the average over a certain period of time (either the timeframe stated in the publication or the publication year) was obtained from online sources. **Table S3.** Publication count for each “Common BE Bacterial Genus” by macro-Level BE location. Macro-level BE Locations included indoor, outdoor, underground, and extreme. Further division by type of sample is also depicted, including surface (S), air (A), water (W). Darker orange color indicates more references identified the genera in the macro BE location and sample type while lighter orange color indicates fewer references. The total number of references for each location and genera are also shown. **Table S4.** Köppen climate classification. Köppen climate classification was used to identify the climate for each publication’s study location. Only the climate assignment between 1981 and 2010 was used for this study. Abbreviation descriptions, latitude, and longitude values are listed. **Table S5.** Publication count for each "Common BE Bacterial Genus" by climate. The climate was identified for each publication's study location based on the closest Köppen latitude and longitude values and correlated with the Köppen ID (see Table S4 for Köppen assignment). For publications describing general locations (e.g., only provided a U.S. state name), a central location in the region was chosen for latitude and longitude. Publications without location specifics were not included, and publications in space were separated out to "Space" category. Darker orange color indicates more references identified the genera in the macro BE location and sample type while lighter orange color indicates fewer references. The total number of references for each location and genera are also shown. **Table S6.** MetaMetaDB environmental category assignment for each "Common BE Bacterial Genus." MetaMetaDB is a database to search for the possible habitats a microorganism could live in and was made by collecting 16S rRNA sequences. Environmental categories for each "Common BE bacterial genus" were based on the identity threshold of 97%, corresponding to the species taxonomic level. Every species for each "Common BE genus" is listed with the corresponding environmental category, where "Y" indicates that the species has been previously identified in the category and "N" indicates the species has not been identified in the category. "Hits" indicates the number of 16S rRNA sequences used by the database. **Table S7.** Mean distance (Dmean) between all pairs of bacterial species for each "Common BE Bacterial Genus." The Dmean was used to describe the genetic diversity among species within a genus. The genetic distance between a pair of bacteria was calculated with the K80 model using the 'dist.dna' function of the 'ape' package of R (https://cran.r-project.org/web/packages/ape). We used a nucleotide sequence alignment of the 16S rRNA genes in 'The All-Species Living Tree' Project (https://www.arb-silva.de/projects/living-tree/). LTP datasets based on SILVA release 128 were downloaded from Archive (https://www.arb-silva.de/no_cache/download/archive/living_tree/LTP_release_128/). Bacterial genera for which 3 or more taxa (*N* > 2) were available at LTP_release_128 were included in the 16S rRNA diversity analysis. **Table S8.** Genome information. Genome features reported include size (Mb), GC content (%), GCSI (GC skew index), and S value (strength of selected codon usage). A genus was deemed BE if observed in at least 6 publications out of 54. The column "BE" shows the number of references that identified the genera. **Table S9.** Robustness of the study. The genome data set used in this study was tested over two levels: 1) different subsets of bacteria (e.g., Phyla of Proteobacteria, Firmicutes, and Actinobacteria) and also randomly selecting one representative for species that have multiple strains sequenced, and 2) testing different numbers of publications (*n* = 1, 2, 3, 4, 5, and 6) to select for BE genera. **Table S10.** Genomic feature statistical analysis for each MetaMetaDB selected environmental category. Each genomic feature per MetaMetaDB environmental category was analyzed to determine statistical significance between the "Common BE genomes" associated with an environment and the "Common BE genomes" not associated. Significance is indicated by q-value < 0.05 and large effect size by Cliff's delta |d| > 0.474. (XLSX 3660 kb)

Additional file 2: Figure S1. Map of publications used in this study. The 54 publications used in this study are mapped by the closest Köppen latitude and longitude values in order to assign Köppen climate IDs by color (Table S4) (Shades of purple = Dry; Shades of green = Tropical; Shades of grey = Snow; Shades of red/orange = mild temperate). The size of the circle indicates the number of common BE bacterial genera (*n* = 28) identified in the publication. Publications not plotted on the map are those from the International Space Station. **Figure S2.** Descriptive statistics of diversity indices (N, Dmean, PD). Plots of diversity levels between taxa within each genus based on 16S rRNA gene sequences, with scatter plots below the diagonal, histograms on the diagonal, and the Pearson correlation coefficient (Corr) above the diagonal. The diversity levels for each genus were represented by three indices: the number of taxa (N), mean distance (Dmean) between all pairs of taxa, and phylogenetic diversity (PD). **Figure S3.** GC skew plots for *Clostridium perfringens* strain 13 (A) and *Methylobacterium* sp. 4–46 (B). G-language Genome Analysis Environment version 1.9.1 (http://www.g-language.org) was used to generate the GC skew plot. **Figure S4.** Genome size (Mb) distribution among MetaMetaDB selected environmental categories. A boxplot showing the distribution of genome sizes within each "Common BE genus" associated with an environment (purple) compared to the "Common BE genera" not associated (red). **Figure S5.** GC content (%) distribution among MetaMetaDB selected environmental categories. A boxplot showing the distribution of GC content within each "Common BE genus" associated with an environment (purple) compared to the "Common BE genera" not associated (red). **Figure S6.** GCSI distribution among MetaMetaDB selected environmental categories. A boxplot showing the distribution of GCSI within each "Common BE genus" associated with an environment (purple) compared to the "Common BE genera" not associated (red). **Figure S7.** S value distribution among MetaMetaDB selected environmental categories. A boxplot showing the distribution of S value within each "Common BE genus" associated with an environment (purple) compared to the "Common BE genera" not associated (red). **Figure S8.** GC skew plots for *Mycobacterium avium* subsp. *paratuberculosis* K-10 (A) and *Mycobacterium leprae* Br4923 (B). G-language Genome Analysis Environment version 1.9.1 (http://www.g-language.org) was used to generate the GC skew plot. (PDF 7950 kb)

Abbreviations

BE: Built environment; GC: Guanine and cytosine; GCSI: GC skew index; S value: Strength of selected codon usage bias; Dmean: Mean distance between all pairs of bacteria as a diversity index; PD: Phylogenetic diversity

Acknowledgements

We gratefully thank Professor Christopher E. Mason from Weill Cornell Medicine and the MetaSUB International Consortium for their support of our microbial community of the BE research projects. Computational resources were provided by the Data Integration and Analysis Facility, National Institute for Basic Biology. We would like to thank Editage (www.editage.jp) for English language editing.

Funding

NM received the Earth-Life Science Institute Origin of Life (EON) Postdoctoral Fellowship, which is supported by a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation. Research funds were received from Keio University, the Yamagata prefectural government and the City of Tsuruoka.

Availability of data and materials

The genomes used in this study were obtained from the NCBI RefSeq database. All data analyzed during this study are included in this published article (see also Supplementary Tables and Figures).

Authors' contributions

NM, SZ, and HS contributed to analyzing the data and writing the manuscript. HS conducted bioinformatics analysis on all the genomes. MT managed bioinformatics environments and helped write the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Earth-Life Science Institute, Tokyo Institute of Technology, Ookayama, Meguro-ku, Tokyo 152-8550, Japan. [2]Department of Earth Sciences, University of Southern California, Stauffer Hall of Science, Los Angeles, CA 90089, USA. [3]Global Research Center for Environment and Energy based on Nanomaterials Science, National Institute for Material Science, 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan. [4]Section of Infection and Immunity, Herman Ostrow School of Dentistry of USC, University of Southern California, Los Angeles, CA 90089-0641, USA. [5]Faculty of Environment and Information

Studies, Keio University, Fujisawa, Kanagawa 252-0882, Japan. [6]Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0035, Japan.

Received: 13 August 2018 Accepted: 18 December 2018
Published online: 28 January 2019

References

1. The World bank. Urban population (% of total). 2018. https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS. Accessed 30 Nov 2018.
2. United Nations. World urbanization prospects: The 2014 revision, highlights. department of economic and social affairs. Population Division, United Nations. 2014. https://esa.un.org/unpd/wup/publications/files/wup2014-highlights.pdf. Accessed 30 Nov 2018.
3. Klepeis NE, Nelson WC, Ott WR, Robinson JP, Tsang AM, Switzer P, et al. The national human activity pattern survey (NHAPS): a resource for assessing exposure to environmental pollutants. J Expo Anal Environ Epidemiol. 2001;11:231–52.
4. Lynch SV, Wood RA, Boushey H, Bacharier LB, Bloomberg GR, Kattan M, et al. Effects of early-life exposure to allergens and bacteria on recurrent wheeze and atopy in urban children. J Allergy Clin Immunol. 2014;134:593–601.e512.
5. Dannemiller KC, Mendell MJ, Macher JM, Kumagai K, Bradman A, Holland N, et al. Next-generation DNA sequencing reveals that low fungal diversity in house dust is associated with childhood asthma development. Indoor Air. 2014;24:236–47.
6. Hoisington AJ, Brenner LA, Kinney KA, Postolache TT, Lowry CA. The microbiome of the built environment and mental health. Microbiome. 2015;3:60.
7. Thaler DS. Toward a microbial Neolithic revolution in buildings. Microbiome. 2016;4:14.
8. Prussin AJ, Marr LC. Sources of airborne microorganisms in the built environment. Microbiome. 2015;3:78.
9. Dunn RR, Fierer N, Henley JB, Leff JW, Menninger HL. Home life: factors structuring the bacterial diversity found within and between homes. PLoS One. 2013;8:e64133.
10. Gibbons SM, Schwartz T, Fouquier J, Mitchell M, Sangwan N, Gilbert JA, et al. Ecological succession and viability of human-associated microbiota on restroom surfaces. Appl Environ Microbiol. 2015;81:765–73.
11. Wood M, Gibbons SM, Lax S, Eshoo-Anton TW, Owens SM, Kennedy S, et al. Athletic equipment microbiota are shaped by interactions with human skin. Microbiome. 2015;3:25.
12. Meadow JF, Altrichter AE, Kembel SW, Moriyama M, O'Connor TK, Womack AM, et al. Bacterial communities on classroom surfaces vary with human contact. Microbiome. 2014;2:7.
13. Kembel SW, Meadow JF, O'Connor TK, Mhuireach G, Northcutt D, Kline J, et al. Architectural design drives the biogeography of indoor bacterial communities. PLoS One. 2014;9:e87093.
14. Meadow JF, Altrichter AE, Bateman AC, Stenson J, Brown GZ, Green JL, et al. Humans differ in their personal microbial cloud. PeerJ. 2015;3:e1258.
15. Meadow JF, Altrichter AE, Green JL. Mobile phones carry the personal microbiome of their owners. PeerJ. 2014;2:e447.
16. McGuire KL, Payne SG, Palmer MI, Gillikin CM, Keefe D, Kim SJ, et al. Digging the New York City skyline: soil fungal communities in green roofs and city parks. PLoS One. 2013;8:e58020.
17. Xu H-J, Li S, Su J-Q, Nie SA, Gibson V, Li H, et al. Does urbanization shape bacterial community composition in urban park soils? A case study in 16 representative Chinese cities based on the pyrosequencing method. FEMS Microbiol Ecol. 2014;87:182–92.
18. Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, et al. Geospatial resolution of human and bacterial diversity with city-scale metagenomics. Cell Sys. 2015;1:72–87.
19. Robertson CE, Baumgartner LK, Harris JK, Peterson KL, Stevens MJ, Frank DN, et al. Culture-independent analysis of aerosol microbiology in a metropolitan subway system. Appl Environ Microbiol. 2013;79:3485–93.
20. Leung MH, Wilkins D, Li EK, Kong FK, Lee PK. Indoor-air microbiome in an urban subway network: diversity and dynamics. Appl Environ Microbiol. 2014;80:6760–70.
21. Checinska A, Probst AJ, Vaishampayan P, White JR, Kumar D, Stepanov VG, et al. Microbiomes of the dust particles collected from the International Space Station and Spacecraft Assembly Facilities. Microbiome. 2015;3:50.
22. Mayer T, Blachowicz A, Probst AJ, Vaishampayan P, Checinska A, Swarmer T, et al. Microbial succession in an inflated lunar/Mars analog habitat during a 30-day human occupation. Microbiome. 2016;4:22.
23. Gilbert JA, Stephens B. Microbiology of the built environment. Nature Rev Microbiol. 2018;16:661–70.
24. Stephens B. What Have We Learned about the Microbiomes of Indoor Environments? mSystems. 2016;1.
25. Adams RI, Bhangar S, Dannemiller KC, Eisen JA, Fierer N, Gilbert JA, et al. Ten questions concerning the microbiomes of buildings. Build Environ. 2016;109:224–34.
26. McEldowney S, Fletcher M. The effect of temperature and relative humidity on the survival of bacteria attached to dry solid surfaces. Lett in Appl Microbiol. 1988;7:83–6.
27. Tang JW. The effect of environmental parameters on the survival of airborne infectious agents. J Royal Soc Interface. 2009;6:S737–46.
28. Mbithi JN, Springthorpe VS, Sattar SA. Effect of relative humidity and air temperature on survival of hepatitis a virus on environmental surfaces. Appl Environ Microbiol. 1991;57:1394–9.
29. Stephens B. What have we learned about the microbiomes of indoor environments? mSystems. 2016;1:e00083–16.
30. Chase J, Fouquier J, Zare M, Sonderegger DL, Knight R, Kelley ST, et al. Geography and Location Are the Primary Drivers of Office Microbiome Composition. mSystems. 2016;1:e00022–16.
31. Emerson JB, Keady PB, Brewer TE, Clements N, Morgan EE, Awerbuch J, et al. Impacts of flood damage on airborne bacteria and fungi in homes after the 2013 Colorado front range flood. Environ Sci Technol. 2015;49:2675–84.
32. Hsu T, Joice R, Vallarino J, Abu-Ali G, Hartmann EM, Shafquat A, et al. Urban transit system microbial communities differ by surface type and interaction with humans and the environment. mSystems. 2016;1:e00018–6.
33. Thos C, Haldane JS, Anderson AM. The carbonic acid, organic matter, and micro-organisms in air, more especially of dwellings and schools. Philos Trans R Soc Lond Ser B Biol Sci. 1887;178:61–111.
34. Shin H, Pei Z, Martinez KA, Rivera-Vinas JI, Mendez K, Cavallin H, et al. The first microbial environment of infants born by C-section: the operating room microbes. Microbiome. 2015;3:59.
35. Rhoads WJ, Ji P, Pruden A, Edwards MA. Water heater temperature set point and water use patterns influence *Legionella pneumophila* and associated microorganisms at the tap. Microbiome. 2015;3:67.
36. Ross AA, Neufeld JD. Microbial biogeography of a university campus. Microbiome. 2015;3:66.
37. Leung MHY, Lee PKH. The roles of the outdoors and occupants in contributing to a potential pan-microbiome of the built environment: a review. Microbiome. 2016;4:21.
38. Adams RI, Bateman AC, Bik HM, Meadow JF. Microbiota of the indoor environment: a meta-analysis. Microbiome. 2015;3:49.
39. Sangwan N, Xia F, Gilbert JA. Recovering complete and draft population genomes from metagenome datasets. Microbiome. 2016;4:8.
40. Klein BA, Lemon KP, Gajare P, Jospin G, Eisen JA, Coil DA. Draft genome sequences of *Dermacoccus nishinomiyaensis* strains UCD-KPL2534 and UCD-KPL2528 isolated from an indoor track facility. Genome Announc. 2017;5.
41. Kincheloe GN, Eisen JA, Coil DA. Draft Genome Sequence of *Arthrobacter* sp. Strain UCD-GKA (Phylum *Actinobacteria*). Genome Announc. 2017;5.
42. Koenigsaecker TM, Eisen JA, Coil DA. Draft Genome Sequence of *Gordonia* sp. Strain UCD-TK1 (Phylum *Actinobacteria*). Genome Announc. 2016;4.
43. Klein BA, Lemon KP, Faller LL, Jospin G, Eisen JA, Coil DA. Draft Genome Sequence of *Curtobacterium* sp. Strain UCD-KPL2560 (Phylum *Actinobacteria*). Genome Announc. 2016;4.
44. Coil DA, Benardini JN, Eisen JA. Draft genome sequence of *Bacillus safensis* JPL-MERTA-8-2, isolated from a Mars-bound spacecraft. Genome Announc. 2015;3.
45. Coil DA, Eisen JA. Draft Genome Sequence of *Porphyrobacter mercurialis* (sp. nov.) Strain Coronado. Genome Announc. 2015;3.
46. Betts MN, Jospin G, Eisen JA, Coil DA. Draft genome sequence of *Planomicrobium glaciei* UCD-HAM (phylum *Firmicutes*). Genome Announc. 2015;3.
47. Lymperopoulou DS, Coil DA, Schichnes D, Lindow SE, Jospin G, Eisen JA, et al. Draft genome sequences of eight bacteria isolated from the indoor environment: *Staphylococcus capitis* strain H36, S. capitis strain H65, S. cohnii strain H62, *S. hominis* strain H69, *Microbacterium* sp. strain H83, *Mycobacterium iranicum* strain H39, *Plantibacter* sp. strain H53, and *Pseudomonas oryzihabitans* strain H72. Stand Genomic Sci. 2017;12:17.

48. Lo JR, Lang JM, Darling AE, Eisen JA, Coil DA. Draft genome sequence of an Actinobacterium, *Brachybacterium muris* strain UCD-AY4. Genome Announc. 2013;1.
49. Bendiks ZA, Lang JM, Darling AE, Eisen JA, Coil DA. Draft Genome Sequence of *Microbacterium* sp. Strain UCD-TDU (Phylum *Actinobacteria*). Genome Announc. 2013;1.
50. Coil DA, Doctor JI, Lang JM, Darling AE, Eisen JA. Draft Genome Sequence of *Kocuria* sp. Strain UCD-OTCP (Phylum *Actinobacteria*). Genome Announc. 2013;1.
51. Holland-Moritz HE, Bevans DR, Lang JM, Darling AE, Eisen JA, Coil DA. Draft Genome Sequence of *Leucobacter* sp. Strain UCD-THU (Phylum *Actinobacteria*). Genome Announc. 2013;1.
52. Flanagan JC, Lang JM, Darling AE, Eisen JA, Coil DA. Draft genome sequence of *Curtobacterium flaccumfaciens* strain UCD-AKU (phylum *Actinobacteria*). Genome Announc. 2013;1.
53. Diep AL, Lang JM, Darling AE, Eisen JA, Coil DA. Draft Genome Sequence of *Dietzia* sp. Strain UCD-THP (Phylum *Actinobacteria*). Genome Announc. 2013;1.
54. Dutta C, Paul S. Microbial lifestyle and genome signatures. Curr Genom. 2012;13:153–62.
55. Coutinho TJD, Franco GR, Lobo FP. Homology-independent metrics for comparative genomics. Comput Struct Biotechnol J. 2015;13:352–7.
56. Mendes-Soares H, Suzuki H, Hickey RJ, Forney LJ. Comparative functional genomics of lactobacillus spp. reveals possible mechanisms for specialization of vaginal lactobacilli to their environment. J Bacteriol. 2014; 196:1458–70.
57. Yang C-C, Iwasaki W. MetaMetaDB: A database and analytic system for investigating microbial habitability. PLOS ONE. 2014;9:e87126.
58. Watve MG, Gangal RM. Problems in measuring bacterial diversity and a possible solution. Appl Environ Microbiol. 1996;62:4299–301.
59. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2011;6:610.
60. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72:5069–72.
61. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, et al. The ribosomal database project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 2009;37:D141–5.
62. Pollock J, Glendinning L, Wisedchanwet T, Watson M. The madness of microbiome: attempting to find consensus “best practice” for 16S microbiome studies. Appl Environ Microbiol. 2018. https://doi.org/10.1128/aem.02627-17.
63. Balvočiūtė M, Huson DH. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? BMC Genomics. 2017;18:114.
64. Kembel SW, Eisen JA, Pollard KS, Green JL. The phylogenetic diversity of metagenomes. PLoS One. 2011;6:e23214.
65. Faith DP. Conservation evaluation and phylogenetic diversity. Biol Conserv. 1992;61:1–10.
66. Cilia V, Lafay B, Christen R. Sequence heterogeneities among 16S ribosomal RNA sequences, and their effect on phylogenetic analyses at the species level. Mol Biol Evol. 1996;13:451–61.
67. Tsukuda M, Kitahara K, Miyazaki K. Comparative RNA function analysis reveals high functional similarity between distantly related bacterial 16Sr RNAs. Sci Rep. 2017;7:9993.
68. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev. 1995;59:143–69.
69. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci U S A. 1985;82:6955–9.
70. 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.
71. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. ISME J. 2014;8:1553–65.
72. Konstantinidis KT, Tiedje JM. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci U S A. 2004;101:3160–5.
73. Reichenberger ER, Rosen G, Hershberg U, Hershberg R. Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. Genome Biol Evol. 2015;7:1380–9.
74. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valín F, Bernardi G. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochem Biophys Res Commun. 2006;347:1–3.
75. Naya H, Romero H, Zavala A, Alvarez B, Musto H. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. J Mol Evol. 2002;55:260–4.
76. Mann S, Chen Y-PP. Bacterial genomic G + C composition-eliciting environmental adaptation. Genomics. 2010;95:7–15.
77. Arakawa K, Tomita M. The GC skew index: a measure of genomic compositional asymmetry and the degree of replicational selection. Evol Bioinformatics Online. 2007;3:159–68.
78. Necşulea A, Lobry JR. A new method for assessing the effect of replication on DNA base composition asymmetry. Mol Biol Evol. 2007;24:2169–79.
79. Frank AC, Lobry JR. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene. 1999;238:65–77.
80. Zhao H-L, Xia Z-K, Zhang F-Z, Ye Y-N, Guo F-B. Multiple factors drive replicating Strand composition Bias in bacterial genomes. Int J Mol Sci. 2015;16:23111.
81. Zhang G, Gao F. Quantitative analysis of correlation between AT and GC biases among bacterial genomes. PLoS One. 2017;12:e0171408.
82. Hershberg R, Petrov DA. Selection on codon bias. Annu Rev Genet. 2008;42:287–99.
83. Bulmer M. The selection-mutation-drift theory of synonymous codon usage. Genetics. 1991;129:897–907.
84. Kliman RM, Hey J. The effects of mutation and natural selection on codon bias in the genes of drosophila. Genetics. 1994;137:1049–56.
85. Garcia-Vallve S, Guzman E, Montero MA, Romeu A. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res. 2003;31:187–9.
86. Brule CE, Grayhack EJ. Synonymous codons: choose wisely for expression. Trends Genet. 2017;33:283–97.
87. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. 2005;33:1141–53.
88. Kurland CG. Strategies for efficiency and accuracy in gene expression. Trends Biochem Sci. 1987;12:126–8.
89. Vieira-Silva S, Rocha EP. The systemic imprint of growth and its uses in ecological (meta)genomics. PLoS Genet. 2010;6.
90. Eyre-Walker A, Bulmer M. Synonymous substitution rates in enterobacteria. Genetics. 1995;140:1407–12.
91. Sharp PM, Li WH. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol. 1987;4:222–30.
92. Sharp PM, Emery LR, Zeng K. Forces that influence the evolution of codon bias. Philos Trans Royal Soc B. 2010;365:1203–12.
93. Rocha EPC. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res. 2004;14:2279–86.
94. Brandis G, Hughes D. The selective advantage of synonymous codon usage bias in *Salmonella*. PLoS Genet. 2016;12:e1005926.
95. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded human microbiome project. Nature. 2017;550:61.
96. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2010;12:32.
97. Moran NA. Microbial minimalism: genome reduction in bacterial pathogens. Cell. 2002;108:583–6.
98. Saini V, Raghuvanshi S, Talwar GP, Ahmed N, Khurana JP, Hasnain SE, et al. Polyphasic taxonomic analysis establishes Mycobacterium indicus pranii as a distinct species. PLoS One. 2009;4:e6263.
99. Baloni P, Padiadpu J, Singh A, Gupta KR, Chandra N. Identifying feasible metabolic routes in *Mycobacterium smegmatis* and possible alterations under diverse nutrient conditions. BMC Microbiol. 2014;14:276.
100. Cortes MAM, Nessar R, Singh AK. Laboratory maintenance of *Mycobacterium abscessus*. In: Curr Protoc Microbiol: Wiley; 2005. https://doi.org/10.1002/9780471729259.mc10d01s18.
101. Mira A, Pushker R, Rodríguez-Valera F. The Neolithic revolution of bacterial genomes. Trends Microbiol. 2006;14:200–6.
102. Jacob SM, Bhagwat AM, Kelkar-Mane V. *Bacillus* species as an intrinsic controller of fungal deterioration of archival documents. Int Biodeterior Biodegradation. 2015;104:46–52.

103. Karakasidou K, Nikolouli K, Amoutzias GD, Pournou A, Manassis C, Tsiamis G, et al. Microbial diversity in biodeteriorated Greek historical documents dating back to the 19th and 20th century: a case study. Microbiology Open. https://doi.org/10.1002/mbo3.596:e00596-n/a.
104. Błażej P, Mackiewicz D, Wnętrzak M, Mackiewicz P. The impact of selection at the amino acid level on the usage of synonymous codons. G3 Genes Genom Genet. 2017;7:967–81.
105. Plotkin JB, Dushoff J, Fraser HB. Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*. Nature. 2004; 428:942–5.
106. Roller M, Lucić V, Nagy I, Perica T, Vlahoviček K. Environmental shaping of codon usage and functional adaptation across microbial communities. Nucleic Acids Res. 2013;41:8842–52.
107. Botzman M, Margalit H. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. Genome Biol. 2011; 12:R109.
108. Bouslimani A, Porto C, Rath CM, Wang M, Guo Y, Gonzalez A, et al. Molecular cartography of the human skin surface in 3D. Proc Natl Acad Sci U S A. 2015;112:E2120–9.
109. Kapono CA, Morton JT, Bouslimani A, Melnik AV, Orlinsky K, Knaan TL, et al. Creating a 3D microbial and chemical snapshot of a human habitat. Sci Rep. 2018;8:3669.
110. Bouslimani A, Melnik AV, Xu Z, Amir A, da Silva RR, Wang M, et al. Lifestyle chemistries from phones for individual profiling. Proc Natl Acad Sci U S A. 2016;113:E7645–54.
111. O'Hara NB, Reed HJ, Afshinnekoo E, Harvin D, Caplan N, Rosen G, et al. Metagenomic characterization of ambulances across the USA. Microbiome. 2017;5:125.
112. Seino K, Takano T, Nakamura K, Watanabe M. An evidential example of airborne bacteria in a crowded, underground public concourse in Tokyo. Atmos Environ. 2005;39:337–41.
113. Tremblay J, Singh K, Fern A, Kirton E, He S, Woyke T, et al. Primer and platform effects on 16S rRNA tag sequencing. Front Microbiol. 2015;6:771.
114. Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. BMC Microbiol. 2016;16:123.
115. Albertsen M, Karst SM, Ziegler AS, Kirkegaard RH, Nielsen PH. Back to basics – the influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. PLoS One. 2015;10:e0132783.
116. Burke CM, Darling AE. A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq. PeerJ. 2016;4:e2492.
117. The MetaSUB International Consortium. The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. Microbiome. 2016;4:24.
118. Mukherjee S, Seshadri R, Varghese NJ, Eloe-Fadrosh EA, Meier-Kolthoff JP, Goker M, et al. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. Nat Biotech. 2017;35:676–83.
119. Chen D, Chen HW. Using the Köppen classification to quantify climate variation and change: an example for 1901–2010. Environ Dev. 2013;6:69–79.
120. Coordinators NR. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2016;44:D7.
121. Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, et al. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res. 2017;46:D851–60.
122. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2017;45:D37–42.
123. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20:289–90.
124. Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer K-H, et al. The all-species living tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol. 2008;31:241–50.
125. SILVA. High quality ribosomal RNA databases. 2017. https://www.arb-silva.de/no_cache/download/archive/living_tree/LTP_release_128/. Nov. 30, 2018.
126. Arakawa K, Suzuki H, Tomita M. Quantitative analysis of replication-related mutation and selection pressures in bacterial chromosomes and plasmids using generalised GC skew index. BMC Genomics. 2009;10:640.
127. Weaver KF, Morales VC, Dunn SL, Godde K, Weaver PF. An introduction to statistical analysis in research: with applications in the biological and life sciences: Wiley; 2017.
128. Sundarraman S. Recent advances in biostatistics: false discovery rates, survival analysis, and related topics: World Scientific; 2011.
129. Cliff N. Dominance statistics: ordinal analyses to answer ordinal questions. Psychol Bull. 1993;114:494–509.
130. Romano J, Kromrey JD, Coraggio J, Skowronek J. Appropriate statistics for ordinal level data: should we really be using t-test and Cohen'sd for evaluating group differences on the NSSE and other surveys. In: Annual Meeting of the Florida Association of Institutional Research; 2006. p. 1–33.
131. Arakawa K, Mori K, Ikeda K, Matsuzaki T, Kobayashi Y, Tomita M. G-language genome analysis environment: a workbench for nucleotide sequence data mining. Bioinformatics. 2003;19:305–6.
132. R Project. R: A language and environment for statistical computing. . 2010. ISBN 3–900051–07–0, URL: https://wwwr-projectorg Nov. 30, 2018.
133. Aydogdu H, Asan A, Tatman OM. Indoor and outdoor airborne bacteria in child day-care centers in Edirne City (Turkey), seasonal distribution and influence of meteorological factors. Environ Monit Assess. 2010;164: 53–66.
134. Baron JL, Vikram A, Duda S, Stout JE, Bibby K. Shift in the Microbial Ecology of a Hospital Hot Water System following the Introduction of an On-Site Monochloramine Disinfection System. PLoS One. 2014;9:e102679.
135. Castro VA, Thrasher AN, Healy M, Ott CM, Pierson DL. Microbial Characterization during the Early Habitation of the International Space Station. Microb Ecol. 2004;47:119–26.
136. Feazel LM, Baumgartner LK, Peterson KL, Frank DN, Harris JK, Pace NR. Opportunistic pathogens enriched in showerhead biofilms. Proc Natl Acad Sci U S A. 2009;106:16393–9.
137. Frank DN, Wilson SS, St. Amand AL, Pace NR. Culture-Independent Microbiological Analysis of Foley Urinary Catheter Biofilms. PLoS One. 2009; 4:e7811.
138. Hwang SH, Yoon CS, Ryu KN, Paik SY, Cho JH. Assessment of airborne environmental bacteria and related factors in 25 underground railway stations in Seoul, Korea. Atmos Environ. 2010;44:1658–62.
139. La Duc MT, Nicholson W, Kern R, Venkateswaran K. Microbial characterization of the Mars Odyssey spacecraft and its encapsulation facility. Environ Microbiol. 2003;5:977–85.
140. Meadow JF, Altrichter AE, Kembel SW, Kline J, Mhuireach G, Moriyama M, et al. Indoor airborne bacterial communities are influenced by ventilation, occupancy, and outdoor air source. Indoor Air. 2014;24:41–8.
141. Moissl C, Osman S, Duc MTL, Dekas A, Brodie E, DeSantis T, et al. Molecular bacterial community analysis of clean rooms where spacecraft are assembled. FEMS Microbiol Ecol. 2007;61:509–21.
142. Novikova N, De Boever P, Poddubko S, Deshevaya E, Polikarpov N, Rakova N, et al. Survey of environmental biocontamination on board the International Space Station. Res Microbiol. 2006;157:5–12.
143. Park HK, Han JH, Joung Y, Cho SH, Kim SA, Kim SB. Bacterial diversity in the indoor air of pharmaceutical environment. J Appl Microbiol. 2014;116:718–27.
144. Wilkins D, Leung MH, Lee PK. Indoor air bacterial communities in Hong Kong households assemble independently of occupant skin microbiomes. Environ Microbiol. 2016;18:1744–53.
145. Barberán A, Dunn RR, Reich BJ, Pacifici K, Laber EB, Menninger HL, et al. The ecology of microscopic life in household dust. Proc R Soc B. 2015;282. https://doi.org/10.1098/rspb.2015.1139.
146. Bruce RJ, Ott CM, Skuratov VM, Pierson DL. Microbial surveillance of potable water sources of the International Space Station: SAE Technical Paper; 2005. http://papers.sae.org/2005-01-2886/. Accessed 16 Oct 2016
147. Coil DA, Neches RY, Lang JM, Brown WE, Severance M, Cavalier D, et al. Growth of 48 built environment bacterial isolates on board the International Space Station (ISS). PeerJ. 2016;4:e1842.
148. Dybwad M, Granum PE, Bruheim P, Blatny JM. Characterization of Airborne Bacteria at an Underground Subway Station. Appl Environ Microbiol. 2012; 78:1917–29.
149. Song B, Leff LG. Identification and characterization of bacterial isolates from the Mir space station. Microbiol Res. 2005;160:111–7.
150. Hewitt KM, Gerba CP, Maxwell SL, Kelley ST. Office Space Bacterial Abundance and Diversity in Three Metropolitan Areas. PLoS One. 2012;7: e37849.
151. Jeon YS, Chun J, Kim BS. Identification of Household Bacterial Community and Analysis of Species Shared with Human Microbiome. Curr Microbiol. 2013;67:557–63.

152. Kang Y, Nagano K. Field measurement of indoor air quality and airborne microbes in a near-zero energy house with an earth tube in the cold region of Japan. Sci Technol Built En. 2016;22:1010–23.
153. Kim KY, Kim YS, Daekeun KIM, Kim HT. Exposure level and distribution characteristics of airborne bacteria and fungi in Seoul metropolitan subway stations. Ind Health. 2011;49:242–8.
154. Lang JM, Coil DA, Neches RY, Brown WE, Cavalier D, Severance M, et al. A microbial survey of the International Space Station (ISS). PeerJ. 2017;5:e4029.
155. Pierson DL. Microbial contamination of spacecraft. Gravitational and Space Research. 2001;14 http://www.gravitationalandspacebiology.org/index.php/journal/article/view/261. Accessed 16 Oct 2016.
156. Venkateswaran K, Vaishampayan P, Cisneros J, Pierson DL, Rogers SO, Perry J. International Space Station environmental microbiome — microbial inventories of ISS filter debris. Appl Microbiol Biotechnol. 2014;98:6453–66.
157. La Duc MT, Sumner R, Pierson D, Venkateswaran K. Characterization and Monitoring of Microbes in the International Space Station Drinking Water. Vancouver, British Columbia, Canada: International Conference for Environmental Systems; 2003.
158. Ruiz-Calderon JF, Cavallin H, Song SJ, Novoselac A, Pericchi LR, Hernandez JN, et al. Walls talk: Microbial biogeography of homes spanning urbanization. Sci Adv. 2016;2:e1501061.
159. Soto-Giron MJ, Rodriguez-R LM, Luo C, Elk M, Ryu H, Hoelle J, et al. Characterization of biofilms developing on hospital shower hoses and implications for nosocomial infections. Appl Environ Microbiol. 2016;AEM: 03529–15.
160. Zhang L, Sriprakash KS, McMillan D, Gowardman JR, Patel B, Rickard CM. Microbiological pattern of arterial catheters in the intensive care unit. BMC Microbiol. 2010;10:266.
161. Mhuireach G, Johnson BR, Altrichter AE, Ladau J, Meadow JF, Pollard KS, et al. Urban greenness influences airborne bacterial community composition. Sci Total Environ. 2016;571:680–7.
162. Flores GE, Bates ST, Caporaso JG, Lauber CL, Leff JW, Knight R, et al. Diversity, distribution and sources of bacteria in residential kitchens. Environ Microbiol. 2013;15:588–96.
163. Farias PG, Gama F, Reis D, Alarico S, Empadinhas N, Martins JC, et al. Hospital microbial surface colonization revealed during monitoring of *Klebsiella* spp., *Pseudomonas aeruginosa*, and non-tuberculous mycobacteria. Antonie van Leeuwenhoek. 2017:1–14.
164. Hospodsky D, Qian J, Nazaroff WW, Yamamoto N, Bibby K, Rismani-Yazdi H, et al. Human Occupancy as a Source of Indoor Airborne Bacteria. PLoS One. 2012;7:e34867.
165. Luongo JC, Barberán A, Hacker-Cary R, Morgan EE, Miller SL, Fierer N. Microbial analyses of airborne dust collected from dormitory rooms predict the sex of occupants. Indoor Air. 2016;27:338–44.
166. Rintala H, Pitkäranta M, Toivola M, Paulin L, Nevalainen A. Diversity and seasonal dynamics of bacterial community in indoor environment. BMC Microbiol. 2008;8:56.
167. Kembel SW, Jones E, Kline J, Northcutt D, Stenson J, Womack AM, et al. Architectural design influences the diversity and structure of the built environment microbiome. ISME J. 2012;6:1469–79.
168. Medrano-Félix A, Martínez C, Castro-del Campo N, León-Félix J, Peraza-Garay F, Gerba CP, et al. Impact of prescribed cleaning and disinfectant use on microbial contamination in the home: Impact of disinfectants in the home. J Appl Microbiol. 2011;110:463–71.
169. Sinclair RG, Gerba CP. Microbial contamination in kitchens and bathrooms of rural Cambodian village households. Lett Appl Microbiol. 2011;52:144–9.
170. Kelley ST, Theisen U, Angenent LT, St. Amand A, Pace NR. Molecular Analysis of Shower Curtain Biofilm Microbes. Appl Environ Microbiol. 2004;70:4187–92.
171. Angenent LT, Kelley ST, Amand AS, Pace NR, Hernandez MT. Molecular identification of potential pathogens in water and air of a hospital therapy pool. Proc Natl Acad Sci U S A. 2005;102:4860–5.
172. Perkins SD, Mayfield J, Fraser V, Angenent LT. Potentially Pathogenic Bacteria in Shower Water and Air of a Stem Cell Transplant Unit. Appl Environ Microbiol. 2009;75:5363–72.
173. Triadó-Margarit X, Veillette M, Duchaine C, Talbot M, Amato F, Minguillón MC, et al. Bioaerosols in the Barcelona subway system. Indoor Air. 2017;27: 564–75.
174. Idi TF, Heéger Z, Vargha M, Márialigeti K. Detection of potentially pathogenic bacteria in the drinking water distribution system of a hospital in Hungary. Clin Microbiol Infect. 2010;16:89–92.
175. Oubre CM, Birmele MN, Castro VA, Venkateswaran KJ, Vaishampayan PA, Jones KU, et al. Microbial Monitoring of Common Opportunistic Pathogens by Comparing Multiple Real-Time PCR Platforms for Potential Space Applications. Am Instit Aeronaut Astronaut. 2013. https://doi.org/10.2514/6.2013-3314.