Pan-genomics: Unmasking the gene diversity hidden in the bacteria species.

Luis David Alcaraz[*]

Correspondence:
lalcaraz@iecologia.unam.mx
Laboratorio Nacional de Ciencias
de la Sostenibilidad. Departamento
de Ecología de la Biodiversidad.
Instituto de Ecología. Universidad
Nacional Autonóma de México,
AP 70-275 Ciudad Universitaria,
04510 México D.F., México
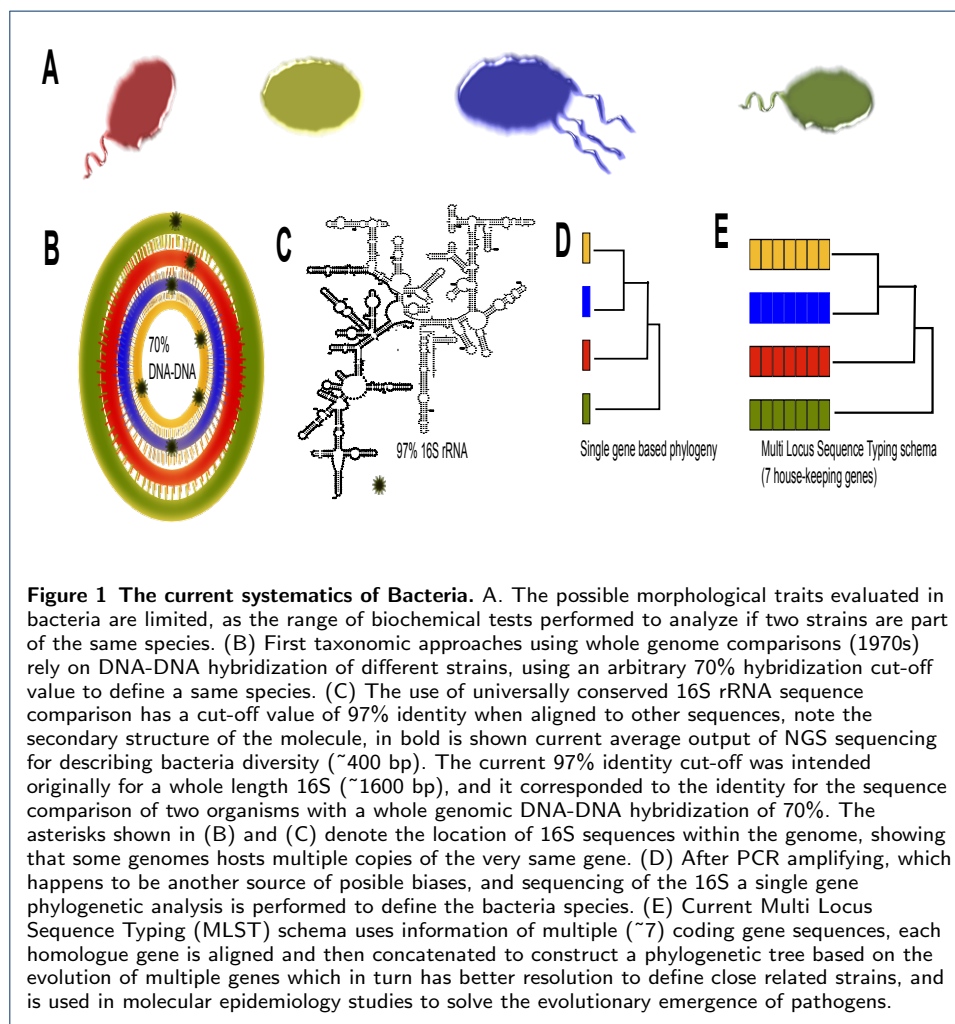[*]To whom correspondence should
be adressed

Abstract

The dramatic increase in genome sequencing during the last years has changed our ideas about bacterial diversity, from single gene to whole community DNA surveys; we have learned that nature's lagest gene repository resides in bacteria resides in bacteria. Comparison of bacteria genomes has contributed to understand the flexibility in size and gene content as well as the gene movement due to gene family expansions and Horizontal Gene Transfer. Bacteria species are currently defined by means of 16S rRNA sequence comparisons and some limited phenotypic traits. There is an ongoing debate about the biological and evolutive significance of the bacteria species, and thus the need to refine the definition of it using the most of the genomic shared information across any taxonomic range. When comparing multiple genomes of related strains we can distinguish a set of common shared features which are known as the core genome. In the other hand, the set of strain specific genes are known as accessory genome. The accessory and core genome conform the total of the genetic composition, and are known as pan-genome. Here we present the possibilities using pan-genomics as a workhorse to describe both taxonomical and functional diversity within bacteria.

Keywords: pan-genomics; core genome; bacteria species; taxonomic diversity; functional diversity

The largest amount of life's gene functions diversity resides in bacteria. This affirmation was possible in the last decade due to rapid development of sequencing technologies, also known as Next Generation Sequencing (NGS) [1]. NGS has aided to describe huge amounts of new species at the genomic level. Despite a bias toward sequencing of human pathogens there are up to 7,411 complete sequenced bacteria genomes up to date. As well as, thousands of Whole Genome Shotgun (WGS) sequencing projects, which when taken into account they gave an approximate of $\geq$ 30,000 ongoing and available genome sequences [2]. Our current knowledge at the genes level could be summarized as genes in this planet are the ones kept by bacteria, and its exceptions (including us within exceptions). Current criteria for naming a bacterium species rely mostly on comparison of 16S rRNA gene (16S) sequences and evaluation of some phenotypic traits like fatty acid profiles, sugar uptake and assimilation, etc. The 16S threshold for delimiting a species is 97% identity of sequence conservation, lower identity values stands for different species. This cut-off value was derived from an old fashion metric when comparing genomes of different species (like *Escherichia coli* and *Salmonella* sp.) and showing a 70% whole DNA-DNA hybridization along phenotypic shared traits, when 16S arose like the gold standard for molecular phylogenetics the equivalent for a 70% DNA-DNA was

correlated with a 97% identity at 16S level [3]. There are several criticims to define bacteria species through this arbitrary cut-off criteria and its biological meaning, nonetheless the value of 16S comparisons to determine large scale evolutionary relationships is accepted universally, what is questioned is to rely only in 16S sequence comparisons for defining bacteria species [4] (see Figure 1).

Figure 1 The current systematics of Bacteria. A. The possible morphological traits evaluated in bacteria are limited, as the range of biochemical tests performed to analyze if two strains are part of the same species. (B) First taxonomic approaches using whole genome comparisons (1970s) rely on DNA-DNA hybridization of different strains, using an arbitrary 70% hybridization cut-off value to define a same species. (C) The use of universally conserved 16S rRNA sequence comparison has a cut-off value of 97% identity when aligned to other sequences, note the secondary structure of the molecule, in bold is shown current average output of NGS sequencing for describing bacteria diversity (~400 bp). The current 97% identity cut-off was intended originally for a whole length 16S (~1600 bp), and it corresponded to the identity for the sequence comparison of two organisms with a whole genomic DNA-DNA hybridization of 70%. The asterisks shown in (B) and (C) denote the location of 16S sequences within the genome, showing that some genomes hosts multiple copies of the very same gene. (D) After PCR amplifying, which happens to be another source of posible biases, and sequencing of the 16S a single gene phylogenetic analysis is performed to define the bacteria species. (E) Current Multi Locus Sequence Typing (MLST) schema uses information of multiple (~7) coding gene sequences, each homologue gene is aligned and then concatenated to construct a phylogenetic tree based on the evolution of multiple genes which in turn has better resolution to define close related strains, and is used in molecular epidemiology studies to solve the evolutionary emergence of pathogens.

Further complications with the use of 16S as a tool to define species are that current species conceptual frame was intended for sexual organisms, inheriting their genomes in a vertical direction. With bacteria and their promiscuity things go complicated, bacteria have capabilities to perform Horizontal Gene Transfer (HGT) and recombination of genes varies from clonal lineages to highly recombinant even named panmictical (highly promiscuous) strains [5]. HGT can operate from single genes to whole genomic islands which are a plus in highly selective environments (think about antibiotic resistance mechanisms, etc.). The difficulties come to a dead-end when it comes to define species in bacteria, if the plasticity of gene movement in these organisms is forgotten. We have to cope with Taxonomic Operational Units (OTUs) as our closest proxy to define the bacteria species, more with a need of a working unit rather than following its biological or evolutionary significance. The OTUs

are defined as clusters of aligned 16S rRNA sequences having at least 97% identity amongst them [6]. One of the currently finest strategies, without genome sequencing is to define close related bacteria strains by means multiple gene alignments and compare to define close related strains, this is called Multi Locus Sequence Typing and Multi Locus Sequence Analysis (MLST and MLSA, respectively). The rationale behind MLST is to use several (~7) conserved genes interdispersed in the genomes to avoid the chance of genetic linkage, by amplifying, sequencing, aligning and concatenating the sequences, put them in a single artificial sequence to maximize the amount of genetic information parsed into the substitution model and thus propose a phylogenetic hypothesis that helps to discriminate between close related strains [7].

Adding up complexity layers, we are just recently noticing that we were missing huge amounts of bacteria diversity out there, for some environments we only knew about 1% of the estimated diversity in part because of the difficulties to culture bacteria in Petri dishes [8]. Parallel to the advancement of sequencing technologies the sequencing of environmental DNA and thus the genomes of uncultivated bacteria are being developed, which is known as metagenomics. Metagenomics can be performed on virtually any environment to study both functional and taxonomical diversities [9]. The metagenomes taxonomical diversity is mostly conducted via Polymerase Chain Reaction (PCR) amplification of 16S rRNA libraries of the studied community. Whole Genome Shotgun Metagenomics has also been developed and with this insight we can know about both community taxonomic and metabolic diversity. Metagenomics has been applied to a wide range of environments to know the microbes associated to them, and which is named the microbiome. The studied microbiomes and their environments are diverse and go from acid mine drainages, soils, oil spills, sea water, plants, and animals [9, 10, 11, 12, 13].
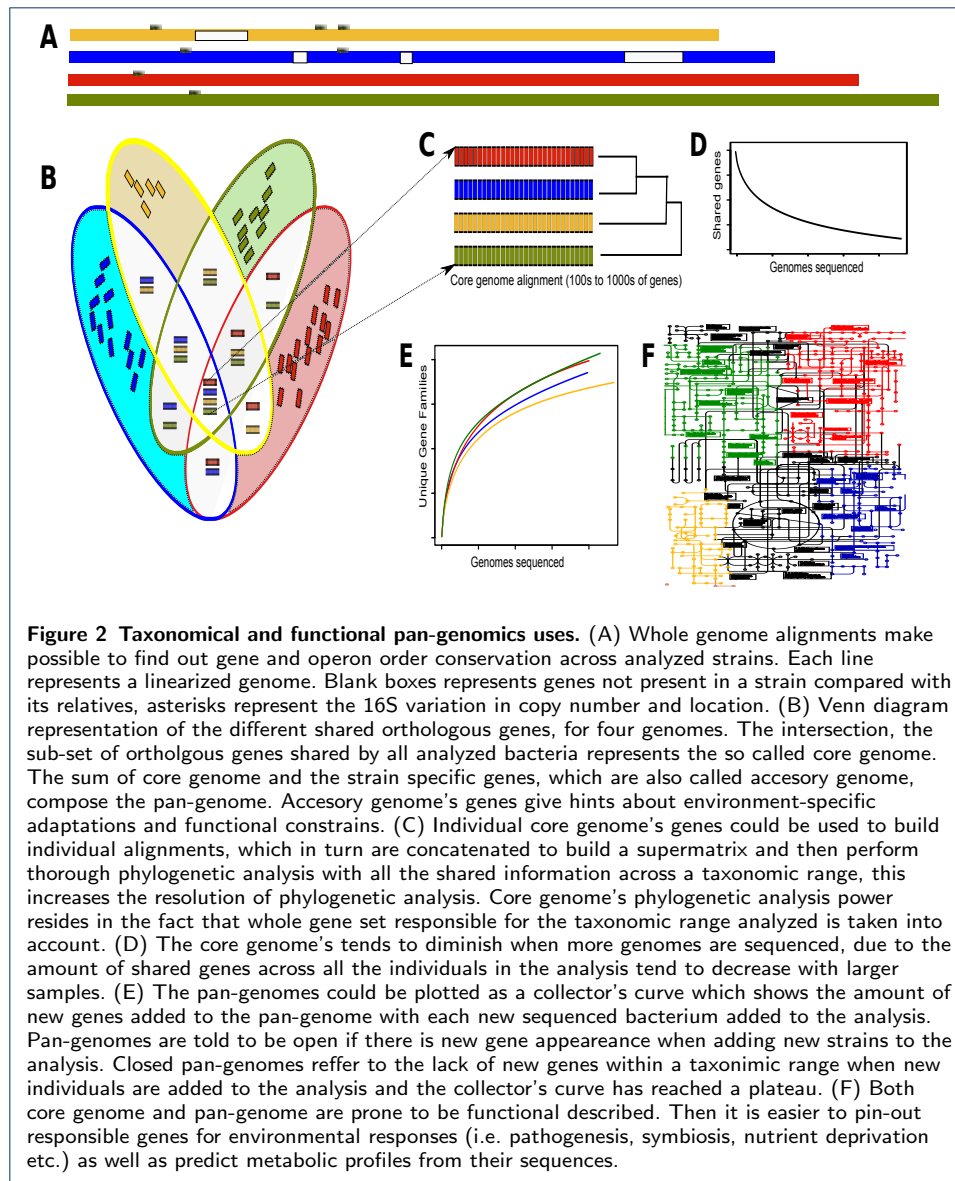
The vast majority of current microbial diversity studies rely only on 16S rRNA amplicon sequencing and comparison. At the end of the study one ends up with a large dataset of OTUs and bunches of multivariate analysis. The final goal is trying to correlate, and in the best cases associate one particular trait (i.e. disease, pollution resistance, etc.) with a particular set of OTUs. One major source of error for these studies has to do with the sampling, which usually lacks direct replicas and studies across time; this is getting better due to the reducing costs of mass sequencing. Additional experiment complications with 16S involves variable copy number across different genomes [14] biased PCR due to the primer design template [15], varying sequence lengths result of the current technologies (100 - 1,000 base pairs) and using the same threshold as if the sequence was full length 16S, etc.

The sequencing technologies are getting cheaper and increasingly accurate, facts that had allowed analysis of whole genomic variation within the very same bacterium species. The pan-genome concept arose when comparing *Streptococcus agalactiae* strains who accomplished all the current taxonomical and clinical criteria to be part of the very same species, producing the same symptomatic illness and hosting the very same 16S rRNA sequences [16]. But, when comparing the genome sequences of the *S. agalactiae* isolated from different patients against the reference genome it was totally unexpected to find out that each strain shared about 20% of the genes. In bacteria genomes, there are a high density of coding genes with small intergenic

spaces and lower amount of repetitive DNA when comparing with eukaryotes. So, in bacteria differences in genome size correlates directly with coding sequences, the larger of your bacterial genome, the more functions you can potentially perform with the genes coded in your genome. *S. agalactiae*'s variation in conservancy of genes are huge, when comparing different eukaryote species, for instance divergence in shared coding sequences across different species like chimpanzees and humans does not go further than 1.23% [17]. Thinking about the differences of 20% in a single bacterium, suposed to be the very same species and finding this difference within the same species is astonishing. As stated above, the sum of the shared and strain unique genes across all the compared genomes is called pan-genome, which in turn can be divided in core genome and accessory genome. In some cases, like the *S. agalactiae*, there is a predicted chance to get new genes for each new sequenced strain, this is called an open pan-genome. For other groups the dynamics are different and there are not predicted new genes for new sequenced strains like is the case for the *Bacillus cereus*; this is called a closed pan-genome [16, 18, 19].

Core genome phylogenetic analysis are the next level of the MLST schema, comparing whole shared genome information coded in the genomes of interest strains. Core genome phylogenomics is done by comparing all the shared (orthologous) genes amongst all the compared species, then align each one of the genes and then concatenating the alignments to build a supermatrix, which in turns feeds a phylogenetic reconstruction; this approach is known as core genome phylogenomics [20] (see Figure 2). The usual parameters of classic molecular evolution, like nucleotide diversity and synonymous/non-synonymous ratios, could be inferred from the core genome alignments. The core genomes could be defined at varying taxonomical depths and could be used to analyze shared gene features from species, genus, family, order, class, and phylum. The upper level of taxonomical resolution (*i.e.* phylum), has the fewer shared genes expected and the lower taxonomical hierarchy (*i.e.* species) is expected to have the larger amount of shared genes. The extra bonus of getting core genomes is that we are able to build molecular functions profiles with the conserved genes across a taxonomic range and find out gene functions responsible for the group cohesion. For example, core genome analysis can aid to find for the expected genes for a enterobacteria like *E. coli* or a sporulating genus like *Bacillus* [21, 19]. Core genome analysis can be helpful to analyze particular phenotypic features like the core genome for any shared trait (i.e. sporulation, heat resistance, antibiotic degradation, etc.) when comparing the shared genes conservancy profile for multiple species dealing with the same environmental challenges [19].

The in-depth study of the core genome sheds light over relevant evolutionary questions, like what are the conserved genes across a taxonomic range, its cut-off similarity values and what are the functional gene constrains of this conservancy. Based on the central dogma of molecular biology one would expect that genes coding for the core machinery of replication (DNA), transcription (RNA) and translation (proteins) would be universally conserved as well as some other house-keeping genes. Phylogenetic reconstructions relying in the whole genome are maybe closer to explain the organism evolutionary history, rather than individual gene genealogies. Important applications also arise, like the development of strain specific vaccines based on knowledge of the variation within the conserved genes of a species [22].

Figure 2 Taxonomical and functional pan-genomics uses. (A) Whole genome alignments make possible to find out gene and operon order conservation across analyzed strains. Each line represents a linearized genome. Blank boxes represents genes not present in a strain compared with its relatives, asterisks represent the 16S variation in copy number and location. (B) Venn diagram representation of the different shared ortholgous genes, for four genomes. The intersection, the sub-set of ortholgous genes shared by all analyzed bacteria represents the so called core genome. The sum of core genome and the strain specific genes, which are also called accesory genome, compose the pan-genome. Accesory genome's genes give hints about environment-specific adaptations and functional constrains. (C) Individual core genome's genes could be used to build individual alignments, which in turn are concatenated to build a supermatrix and then perform thorough phylogenetic analysis with all the shared information across a taxonomic range, this increases the resolution of phylogenetic analysis. Core genome's phylogenetic analysis power resides in the fact that whole gene set responsible for the taxonomic range analyzed is taken into account. (D) The core genome's tends to diminish when more genomes are sequenced, due to the amount of shared genes across all the individuals in the analysis tend to decrease with larger samples. (E) The pan-genomes could be plotted as a collector's curve which shows the amount of new genes added to the pan-genome with each new sequenced bacterium added to the analysis. Pan-genomes are told to be open if there is new gene appeareance when adding new strains to the analysis. Closed pan-genomes reffer to the lack of new genes within a taxonimic range when new individuals are added to the analysis and the collector's curve has reached a plateau. (F) Both core genome and pan-genome are prone to be functional described. Then it is easier to pin-out responsible genes for environmental responses (i.e. pathogenesis, symbiosis, nutrient deprivation etc.) as well as predict metabolic profiles from their sequences.

The presence and absence patterns of the accessory genome observed throughout a set of bacteria being compared could be the result of gene loss or gene acquisition through Horizontal Gene Transfer (Figure 2). The importance of gene acquisition/loss ratio is yet to be investigated but there are some examples like the one examining the Achaea *Sulfolobus ilsandicus* and the importance of analyzing its pan-genome to determine strain and even location specific genes and their dynamics [23]. The building of the pan-genome is helpful to have a full inventory of the metabolic capabilities of a given group of organisms. Differences in the unique genes of close related bacteria could be a partial answer of local adaptation to particular life styles or niches (i.e. free-living, host-associate, virulence, etc.).

The main goal of the 16S amplicon studies is to have a diversity inventory of a particular environment and try to associate OTUs with particular functions, using the 16S OTU as a proxy for the metabolic diversity. Predictions trying to connect

a particular OTU with metabolic functions, for example a gut environment is dominated by say *E. coli* related OTUs, then taking a single reference genome of a single strain would be an naive guess if it fails to consider that we can have up to 30% in presence-absence of coding genes, and that particular genes in the accessory genomes are the probable environment restricted genes. However, several attempts are being performed now trying to use the large amount of 16S massive sequencing from virtually any environment and infer metabolic diversity and link it to phylogenetic distances [24, 25]. Another approach in trying to gain insights into the metabolic diversity when only having 16S sequences could be to have pan-genomes catalogues for all the known groups of bacteria with sequenced genomes, generating confidence intervals based on gene presence-absence within particular bacteria.

The 16S gene databases are among the most prolific ones. The understanding of what we know actually about bacteria diversity is mostly in debt with 16S sequence analysis. For sure 16S analysis has been useful and will keep that way when studying unknown environments and when the goal is to have a first glimpse about complexity of the community structure. For pathogens, diagnosis and management 16S analysis was outdated some time ago and the need to develop rapid and accurate methods for resolving close relative type strains derived into the MLST analysis. With the current pace of sequencing technologies development is urgent to redefine the minimum standards when defining bacteria diversity. The new bacteria diversity standards are likely to require core and pan-genomics analysis to define the bacteria taxa, as well as understanding local dynamics for pan-genomics at each taxonomic unit.

The beauty of not knowing a precise way to describe a species should not be taken as a pitfall for microbiology. With bacteria, we are dealing with the main repository of genes and biological functions that have allowed microbes to be the major players in our world, from biogeochemical cycles, energy harvesting and cycling and thus making life for all the other being forms possible. The species concept, developed for when you are well behaved and transmit your genes in vertical form seems a little rigid when dealing with bacteria and their tremendous capabilities of transforming and sharing genes in a happy and promiscuous way. We just need to refine our vision and take into account the internal variability for genes and thus functions of each related bacteria, and praise it, develop new indexes (like Genome Similarity Score [19]) that take into account the whole set of shared features when comparing bacteria, along sides phylogenetic traditional ways. If we are aware of the current utilitarian bacterium species concept and that we understand that some of the major traits of a bacterium, like pathogenesis, are likely to occur in close related working units, call them species or OTUs, we can cope with that. But being unaware of the huge functional diversity connected to what we already call a bacterium species is nonsense nowadays.

Acknowledgements

References

1. Metzker, M.L.: Sequencing technologies — the next generation. Nature Reviews Genetics **11**(1), 31–46 (2009)
2. Pagani, I., Liolios, K., Jansson, J., Chen, I.-M.A., Smirnova, T., Nosrat, B., Markowitz, V.M., Kyrpides, N.C.: The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Research **40**(Database issue), 571–579 (2012)
3. Goris, J., Konstantinidis, K.T., Klappenbach, J.a., Coenye, T., Vandamme, P., Tiedje, J.M.: DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. International Journal of Systematic and Evolutionary Microbiology **57**(Pt 1), 81–91 (2007)
4. Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G., Hanage, W.P.: The bacterial species challenge: making sense of genetic and ecological diversity. Science **323**(5915), 741–746 (2009)
5. Smith, J.M., Smith, N.H., Rourke, M.O., Spratt, B.G., O'Rourke, M.: How Clonal are Bacteria? Proceedings of the National Academy of Sciences of the United States of America **90**(10), 4384–4388 (1993)
6. Konstantinidis, K.T., Tiedje, J.M.: Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. Current Opinion in Microbiology **10**(5), 504–509 (2007)
7. Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E., Peer, Y.V., Vandamme, P., Thompson, F.L., Swings, J.: Re-evaluating prokaryotic species. Microbiology **3**(September), 733–739 (2005)
8. Vartoukian, S.R., Palmer, R.M., Wade, W.G.: Strategies for culture of 'unculturable' bacteria. FEMS Microbiology Letters **309**(1), 1–7 (2010)
9. Lundberg, D.S., Lebeis, S.L., Paredes, S.H., Yourstone, S., Gehring, J., Malfatti, S., Tremblay, J., Engelbrektson, A., Kunin, V., Rio, T.G.D., Edgar, R.C., Eickhorst, T., Ley, R.E., Hugenholtz, P., Tringe, S.G., Dangl, J.L.: Defining the core Arabidopsis thaliana root microbiome. Nature **488**(7409), 86–90 (2012)
10. Belda-Ferre, P., Alcaraz, L.D., Cabrera-Rubio, R., Romero, H., Simón-Soro, A., Pignatelli, M., Mira, A.: The oral metagenome in health and disease. The ISME Journal **1**(11), 46–56 (2011)
11. Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-pfannkoch, C., Venter, J.E., Li, K., Kravitz, S., Heidelberg, J.F., Utterback, T., Rogers, Y.-H., Falc, N, L.I., Souza, V., Bonilla-Rosso, G., Eguiarte, L.E., Karl, D.M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M.R., Strausberg, R.L., Nealson, K., Friedman, R., Frazier, M., Venter, J.C.: The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biology **5**(3), 399–431 (2007)
12. Mendes, R., Kruijt, M., de Bruijn, I., Dekkers, E., van der Voort, M., Schneider, J.H.M., Piceno, Y.M., DeSantis, T.Z., Andersen, G.L., Bakker, P.a.H.M., Raaijmakers, J.M.: Deciphering the rhizosphere microbiome for disease-suppressive bacteria. Science (New York, N.Y.) **332**(6033), 1097–100 (2011). doi:http://dx.doi.org/10.1126/science.120398010.1126/science.1203980
13. Beazley, M.J., Martinez, R.J., Rajan, S., Powell, J., Piceno, Y.M., Tom, L.M., Andersen, G.L., Hazen, T.C., Van Nostrand, J.D., Zhou, J., Mortazavi, B., Sobecky, P.a.: Microbial community analysis of a coastal salt marsh affected by the Deepwater Horizon oil spill. PLoS one **7**(7), 41305 (2012)
14. Klappenbach, J.a., Saxman, P.R., Cole, J.R., Schmidt, T.M.: rrndb: the Ribosomal RNA Operon Copy Number Database. Nucleic Acids Research **29**(1), 181–4 (2001)
15. Engelbrektson, A., Kunin, V., Wrighton, K.C., Zvenigorodsky, N., Chen, F., Ochman, H., Hugenholtz, P.: Experimental factors affecting PCR-based estimates of microbial species richness and evenness. The ISME journal **4**(5), 642–647 (2010)
16. Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., DeBoy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'connor, K.J., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., Fraser, C.M.: Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences of the United States of America **102**(39), 13950–13955 (2005)
17. The Chimpanzee Sequencing and Analysis Consortium: Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**(7055), 69–87 (2005)
18. Mira, A., Martín-Cuadrado, A.: The bacterial pan-genome: a new paradigm in microbiology. International Microbiology **10**, 45–57
19. Alcaraz, L.D., Moreno-Hagelsieb, G., Eguiarte, L.E., Souza, V., Herrera-Estrella, L., Olmedo, G.: Understanding the evolutionary relationships and major traits of Bacillus through comparative genomics. BMC Genomics **11**(1), 332 (2010)
20. Alcaraz, L.D., López-Ramirez, V., Moreno-Letelier, A., Herrera-Estrella, L., Souza, V., Olmedo-Alvarez, G.: Genomics of bacteria from an ancient marine origin: Clues to survival in an oligotrophic environment. In: Dar, I.A., Mithas, A.D. (eds.) Earth and Environmental Sciences, 1st edn., pp. 609–630. Intech, Croatia (2011)
21. Sims, G.E., Kim, S.-H.: Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs). Proceedings of the National Academy of Sciences of the United States of America **108**(20), 8329–8334 (2011)
22. Luciani, F., Bull, R.a., Lloyd, A.R.: Next generation deep sequencing and vaccine design: today and tomorrow. Trends in Biotechnology **30**(9), 443–452 (2012)
23. Reno, M., Held, N.: Biogeography of the Sulfolobus islandicus pan-genome. Proceedings of the National Academy of Sciencies of the United States of America **106**(21) (2009)
24. Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.a., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., Beiko, R.G., Huttenhower, C.: Predictive functional profiling

of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnology **31**(9), 814–21 (2013)

25. Okuda, S., Tsuchiya, Y., Kiriyama, C., Itoh, M., Morisaki, H.: Virtual metagenome reconstruction from 16S rRNA gene sequences. Nature Communications **3**, 1203 (2012)