Pangenome analysis of bacteria and its application in metagenomics

Dissertation zur Erlangung des

naturwissenschaftlichen Doktorgrades

der Julius-Maximilians-Universität Würzburg

vorgelegt von

Oleksandr Maistrenko

(Geburtsort: Kyiv, Ukraine)

Würzburg, 2020

Eingereicht am: 08.07.2020

Mitglieder der Promotionskommission:

Vorsitzender: ……..

Gutachter: Prof. Dr. Peer Bork

Gutachter: Prof. Dr. Thomas Dandekar

Tag des Promotionskolloquiums: ……..…

Doktorurkunde ausgehändigt am: …….…….

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die Dissertation: „**Bakterielle Pan-Genome und ihre Anwendungen in der Metagenomik**“, eigenständig, d. h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen, als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Weiterhin erkläre ich, dass bei allen Abbildungen und Texten bei denen die Verwertungsrechte (Copyright) nicht bei mir liegen, diese von den Rechtsinhabern eingeholt wurden und die Textstellen bzw. Abbildungen entsprechend den rechtlichen Vorgaben gekennzeichnet sind sowie bei Abbbildungen, die dem Internet entnommen wurden, der entsprechende Hypertextlink angegeben wurde.

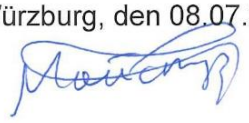
Affidavit

I hereby declare that my thesis entitled: „**Pangenome analysis of bacteria and its application in metagenomics**” is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore I verify that the thesis has not been submitted as part of another examination process neither in identical nor in similar form.

Besides I declare that if I do not hold the copyright for figures and paragraphs, I obtained it from the rights holder and that paragraphs and figures have been marked according to law or for figures taken from the internet the hyperlink has been added accordingly.

Würzburg, den 08.07.2020



Signature PhD-student

Summary

The biosphere harbors a large quantity and diversity of microbial organisms that can thrive in all environments. Estimates of the total number of microbial species reach up to 10^{12}, of which less than 15,000 have been characterized to date. It has been challenging to delineate phenotypically, evolutionary and ecologically meaningful lineages such as for example, species, subspecies and strains. Even within recognized species, gene content can vary considerably between sublineages (for example strains), a problem that can be addressed by analyzing pangenomes, defined as the non-redundant set of genes within a phylogenetic clade, as evolutionary units.

Species considered to be ecologically and evolutionary coherent units, however to date it is still not fully understood what are primary habitats and ecological niches of many prokaryotic species and how environmental preferences drive their genomic diversity. Majority of comparative genomics studies focused on a single prokaryotic species in context of clinical relevance and ecology. With accumulation of sequencing data due to genomics and metagenomics, it is now possible to investigate trends across many species, which will facilitate understanding of pangenome evolution, species and subspecies delineation.

The major aims of this thesis were 1) to annotate habitat preferences of prokaryotic species and strains; 2) investigate to what extent these environmental preferences drive genomic diversity of prokaryotes and to what extent phylogenetic constraints limit this diversification; 3) explore natural nucleotide identity thresholds to delineate species in bacteria in metagenomics gene catalogs; 4) explore species delineation for applications in subspecies and strain delineation in metagenomics.

The first part of the thesis describes methods to infer environmental preferences of microbial species. This data is a prerequisite for the analyses performed in the second part of the thesis which explores how the structure of bacterial pangenomes is predetermined by past evolutionary history and how is it linked to environmental preferences of the species. The main finding in this subchapter that habitat preferences explained up to 49% of the variance for pangenome structure, compared to 18% by phylogenetic inertia. In general, this trend indicates that phylogenetic inertia does not limit evolution of pangenome size and diversity, but that convergent evolution may overcome phylogenetic constraints. In this project we show that core genome size is associated with higher environmental ubiquity of species. It is likely this is due to the fact that species need to have more versatile genomes and most necessary genes need to be present in majority of genomes of that species to be highly prevalent. Taken together these

findings may be useful for future predictive analyses of ecological niches in newly discovered species.

The third part of the thesis explores data-driven, operational species boundaries. I show that homologous genes from the same species from different genomes tend to share at least 95% of nucleotide identity, while different species within the same genus have lower nucleotide identity. This is in line with other studies showing that genome-wide natural species boundary might be in range of 90-95% of nucleotide identity. Finally, the fourth part of the thesis discusses how challenges in species delineation are relevant for the identification of meaningful within-species groups, followed by a discussion on how advancements in species delineation can be applied for classification of within-species genomic diversity in the age of metagenomics.

Zusammenfassung

Die Biosphäre beherbergt eine große Zahl verschiedener Mikroorganismen, die fast alle bekannten Lebensräume besiedeln können. Die Gesamtzahl mikrobieller Spezies liegt Schätzungen zu Folge bei bis zu 10^{12}, von denen jedoch bis heute erst 15.000 beschrieben worden sind. Die Beschreibung von phänotypisch, evolutionsbiologisch und ökologisch kohärenten Spezies, Sub-Spezies oder Stämmen stellt Forscher vor konzeptionelle Herausforderungen. Selbst innerhalb anerkannter Spezies kann die Kombination einzelner Gene oft stark variieren. Diese Beobachtung ist die Grundlage der Analyse von Pan-Genomen. also der Konstellation originärer Gene innerhalb einer Abstammunsglinie, als evolutionsbiologische Einheiten.

Spezies entsprechen prinzipiell ökologisch und evolutionär kohärenten Einheiten, jedoch sind die primären Habitate und ökologischen Nischen vieler prokaryotischer Spezies bis heute nur unzureichend beschrieben, insbesondere mit Blick auf den Einfluss ökologischer Präferenzen auf die Evolution von Genomen. Die Mehrheit vergleichender genomischer Studien untersucht einzelne prokaryotische Spezies mit Bezug auf deren klinische oder ökologische Relevanz. Aufgrund der wachsenden Verfügbarkeit genomischer Daten ist es nun jedoch möglich, vergleichende Studien über Speziesgrenzen hinweg durchzuführen, um allgemeine Prinzipien der Evolution von Pan-Genomen, Spezies und Sub-Spezies zu untersuchen.

Die wesentlichen Ziele der vorliegenden Arbeit waren 1) die Annotation von Habitatpräferenzen prokaryotischer Spezies und Stämme; 2) die Quantifizierung des Einflusses von Umwelt und Evolutionsgeschichte (Phylogenie) auf die genomische Diversität von Prokaryoten; 3) die Bestimmung natürlicher Schwellenwerte der Genomsequenzähnlichkeit zwischen Spezies, auch anhand von Genkatalogen; 4) die Untersuchung der Abgrenzung zwischen Spezies, Sub-Spezies und Stämmen mithilfe metagenomischer Daten.

Im ersten Teil der Arbeit werden Methoden zur Bestimmung ökologischer Präferenzen mikrobieller Spezies beschrieben. Die so gewonnenen Daten dienen in der Folge als Grundlage für die Quantifizierung von Umwelt- und evolutionsgeschichtlichen Einflüssen auf die Struktur und Evolution bakterieller Pan-Genome im zweiten Teil der Arbeit. Ein zentrales Ergebnis dieser Untersuchung war, dass bis zu 49% der strukturellen Varianz in Pan-Genomen durch Habitatpräferenzen erklärt werden kann, im Gegensatz zu lediglich 18% durch phylogenetische Trägheitseffekte. Dies zeigt, dass die Größe und Diversität von Pan-Genomen nicht phylogenetisch limitiert ist, insbesondere in Fällen von konvergenter Evolution. Große Kern-Genome sind ferner mit einer weiten ökologischen Verbreitung von Spezies assoziiert; eine

mögliche Erklärung ist, dass weit verbreitete Spezies vielseitigere Genome mit mehr notwendigen Genen besitzen, die ein Überleben in vielfältigen Umgebungen ermöglichen. Die vorgelegte Arbeit kann weiterhin einen Beitrag zur Vorhersage ökologischer Profile neu beschriebener Spezies leisten.

Im dritten Teil der Arbeit werden datenbezogene, operationelle Definition von Spezies-Grenzen untersucht. Es konnte gezeigt werden, dass Gene verschiedener Genome innerhalb derselben Spezies normalerweise mindestens 95% Ähnlichkeit der Nukleotidsequenz aufweisen, während die Ähnlichkeit zwischen Spezies desselben Genus geringer ausfällt. Dieser Wert liegt im Rahmen früherer Schätzungen. Der vierte Teil der Arbeit beschreibt abschließend die Herausforderungen bei der Bestimmung von evolutionären Linien innerhalb von Spezies und diskutiert anschließend, wie konzeptionelle Entwicklungen in dieser Frage für die Klassifizierung und Quantifizierung von Diversität anhand metagenomischer Daten genutzt werden kann.

Acknowledgments

I would like to thank my supervisor Peer Bork for providing me the opportunity to do PhD in his laboratory in the extremely productive environment; for teaching me how to “sell” scientific results; how to identify important questions; and how to think about the big picture while being focused on the details.

I thank the Thesis Advisory Committee: Thomas Dandekar, Georg Zeller and Nassos Typas for valuable input on progress of my projects and career building.

I would like to acknowledge all colleagues with whom I had honor and pleasure to work on various projects and to learn from: Daniel Mende, Mechthild Luetge, Sebastian Schmidt, Supriya Khedkar, Thea Van Rossum, Simone Li, Renato Alves, Ece Kartal, Alejandro Murillo, Pamela Ferretti, Luis Pedro Coelho, Falk Hildebrand, Jaime Huerta-Cepas, Daniel Machado, Michael Kuhn.

Special thanks for:

- advice with statistical analysis to Bernd Klaus, Lucas Moitinho-Silva, Michael Kuhn;
- help with computational infrastructure to Yan Yuan, Ivica Letunic, Wasiu Akanni and Anthony Fullam;
- proofreading my annual progress reports, comments on the presentations, figures and thesis to Sebastian, Thea, Supriya, Pamela, Ece, Chris, Simone;
- all the random interesting, weird and fun daily conversations to office 107: Lucas, Anna, Simone, Sebastian, Luis, Ece, Renato, Pamela, Sander, Alejandro and all coffee-drinkers;
- for entertainment to UT and ping-pong players;

I also would like to acknowledge previous supervisors and colleagues who influenced my scientific mindset.

- Peter Bergholz from NDSU (Fargo, USA) for teaching me comparative genomics and bioinformatics;
- Entire Drosophila Lab at Taras Shevchenko National University of Kyiv (Ukraine) and especially Iryna Kozeretska, Svitlana Serga and Andrii Rozhok for showing me how creative, highly productive and cooperative academic environment can be;
- Yuliya Luchakivska from Institute of Cell Biology and Genetic Engineering (Kyiv, Ukraine) for introducing me to science.

Table of content

List of Figures

List of Tables

List of publications

Publications resulting directly from Doctoral studies:

Maistrenko, O.M., Mende, D.R., Luetge, M., Hildebrand, F., Schmidt, T.S.B., Li, S.S., Rodrigues, J.F.M., von Mering, C., Pedro Coelho, L., Huerta-Cepas, J., et al. (2020). Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. The ISME Journal 14, 1247–1259.

Van Rossum, T., Ferretti, P., **Maistrenko, O.M.**, and Bork, P. (2020). Diversity within species: interpreting strains in microbiomes. Nature Reviews Microbiology 1–16.

Mende, D.R., Letunic, I., **Maistrenko, O.M.**, Schmidt, T.S.B., Milanese, A., Paoli, L., Hernández-Plaza, A., Orakov, A.N., Forslund, S.K., Sunagawa, S., et al. (2020). proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. Nucleic Acids Res 48, D621–D625.

Coelho, L.P., Alves, R., Cantalapiedra, C.P., Giner-Lamia, J., Schmidt, T.S., Mende, D., Orakov, A., Letunic, I., Hildebrand, F., Rossum, T.V., Forslund, S. K., Khedkar, S., **Maistrenko, O M.**, et al. Towards the biogeography of prokaryotic genes. In preparation.

Other publications before and during Doctoral studies:

Machado, D., **Maistrenko, O.M.**, Andrejev, S., Kim, Y., Bork, P., Patil, K.R., and Patil, K.R. (2020). Polarization of microbial communities between competitive and cooperative metabolism. BioRxiv 2020.01.28.922583. In revision in Nature Ecology & Evolution

Gora, N.V., Serga, S.V., **Maistrenko, O.M.**, Ślęzak-Parnikoza, A., Parnikoza, I.Yu., Tarasiuk, A.N., Demydov, S.V., and Kozeretska, I.A. (2020). Climate Factors and Wolbachia Infection Frequencies in Natural Populations of Drosophila melanogaster. Cytol. Genet. 54, 189–198.

Kozeretska, O.I., **Maistrenko, O.M.**, Serga, S.V., Dombrovskyi, I.V., Ostapchenko, L.I., Demydov, S.V., and Kozeretska, I.A. (2020). Allele frequencies for 15 forensic STR loci in a population sample from the Kyiv region, Ukraine. Australian Journal of Forensic Sciences 52, 387–392.

Schmidt, T.S., Hayward, M.R., Coelho, L.P., Li, S.S., Costea, P.I., Voigt, A.Y., Wirbel, J., **Maistrenko, O.M.**, Alves, R.J., Bergsten, E., et al. (2019). Extensive transmission of microbes along the gastrointestinal tract. ELife 8, e42693.

Yasinskyi, Y., Protsenko, O., **Maistrenko, O.**, Rybalchenko, V., Prylutskyy, Yu., Tauscher, E., Ritter, U., and Kozeretska, I. (2019). Reconciling the controversial data on the effects of C60 fullerene at the organismal and molecular levels using as a model Drosophila melanogaster. Toxicology Letters 310, 92–98.

Serga, S.V., Dombrovskyi, I.V., **Maistrenko, O.M.**, Ostapchenko, L.I., Demydov, S.V., Krivda, R.G., and Kozeretska, I.A. (2017). Allele frequencies for 15 STR loci in the Ukrainian population. Forensic Science International: Genetics 29, e40–e41.

Maistrenko, O.M., Serga, S.V., Vaiserman, A.M., and Kozeretska, I.A. (2016). Longevity-modulating effects of symbiosis: insights from Drosophila–Wolbachia interaction. Biogerontology 17, 785–803.

Maistrenko, O.M., Serga, S.V., Vaiserman, A.M., and Kozeretska, I.A. (2015). Effect of Wolbachia Infection on Aging and Longevity-Associated Genes in Drosophila. In Life Extension: Lessons from Drosophila, A.M. Vaiserman, A.A. Moskalev, and E.G. Pasyukova, eds. (Springer International Publishing), pp. 83–104.

Maistrenko, O.M., Luchakivska, Yu.S., Zholobak, N.M., Spivak, M.Ya., and Kuchuk, M.V. (2015). Obtaining of the transgenic Heliantus tuberosus L. plants, callus and "hairy" root cultures able to express the recombinant human interferon alpha-2b gene. Cytol. Genet. 49, 308–313.

Serga, S.V., **Maistrenko, O.M.**, Rozhok, A.I., Mousseau, T.A., and Kozeretska, I.A. (2015). Colonization of a temperate-zone region by the fruit fly Drosophila simulans (Diptera: Drosophilidae). Can. J. Zool. 93, 799–804.

Serga, S., **Maistrenko, O.**, Rozhok, A., Mousseau, T., and Kozeretska, I. (2014). Fecundity as one of possible factors contributing to the dominance of the wMel genotype of Wolbachia in natural populations of Drosophila melanogaster. Symbiosis 63, 11–17.

Abbreviations

HGT – Horizontal gene transfer

ANI – Average nucleotide identity

SNP – Single nucleotide polymorphism

PATRIC – Pathosystems Resource Integration Center

GMGC – Global Microbial Gene Catalog

PCs – Principal components

dN/dS – ratio of non-synonymous / synonymous substitutions

COG – cluster of orthologous groups

Chapter 1. Introduction

1.1. Diversity and function of microbiome

Microbial organisms inhabit almost every environment on Earth and have a significant impact on biogeochemical cycles of entire biosphere, specific ecosystems within it and the wellbeing of individual macroorganisms (Stolz 2017; Lovelock 1979; McFall-Ngai et al. 2013). Microorganisms are key components of nutrient, energy and information (horizontal gene transfer and molecular signaling) cycles across ecosystems. Prokaryotes are primary producers as well as decomposers of organic matter and a source of nutrients for other trophic levels. Carbon, nitrogen, sulfur, phosphor and other nutrients' cycles depend on bacteria and their viruses (Delgado-Baquerizo et al. 2017). There are at least 10^4 known microbial species with sequenced genomes, while the predicted number of microbial species might be as low as 10^5 or as high as 10^{12} in the entire biosphere (Locey and Lennon 2016; Editorial 2011).

The community complexity of microbial ecosystems varies enormously, from single-species ecosystems such as in the lithosphere (Chivian et al. 2008), to highly complex soil communities with thousands of species per gram of soil (Torsvik, Goksoyr, and Daae 1990; Raynaud and Nunan 2014). Similarly, microbial population densities cover a wide range, from 10^4-10^6 cells/gram in lithosphere (Cockell 2011), ocean – 10^5 cells/g (Nimnoi and Pongsilp 2020) to soils – 10^8 - 10^{10} cells/g (Christensen, Hansen, and Sørensen 1999) and 10^{11} cells/g in the mammalian gastrointestinal tract (Sender, Fuchs, and Milo 2016). Such density, to some extent, leads to close interactions between prokaryotic species and eukaryotic organisms in the ecosystem. Formally, bacteria have been ecologically classified as either free-living and host-associated. Host-associated species tend to form close symbiotic interactions with macroorganisms (also more generically called "host" organism when cannot be referred to as macroorganism), for example plants and animals. Due to these symbiotic relationships, the host-symbiont unit (also referred to as meta-organism) is considered a fundamental unit for many biological processes (De Bary 1879; McFall-Ngai et al. 2013). Now it is acknowledged that interactions between species (macroorganism and its symbionts) form complex network which resulted in the shift in understanding what is the unit of development, disease/health, aging, lifespan determination, and natural selection (McFall-Ngai et al. 2013; Schmidt, Raes, and Bork 2018; Cho and Blaser 2012; Zilber-Rosenberg and Rosenberg 2008; B. B. Finlay et al. 2019; Maistrenko et al. 2016; 2015). Importantly, host-associated microbial species can have distinct genomic features compared to free-living bacteria (for example, genome size, functional gene content, discussed in more details in Subchapter 1.2.). Both free-living and host-associated microbes frequently exchange genetic material across clade boundaries, in a

process called horizontal gene transfer (HGT). Patterns and outcomes of HGT in host-associated and free-living species might differ. In particular host-associated species tend to share and lose genes over time due to relaxed environmental selective pressure, while free-living might accumulate genes. HGT essentially leads to circulation of information within and between ecosystems and species. In consequence, microbial lineages are interconnected genetically within and across clade boundaries, and theoretically species have access to nearly-infinite pools of genes (Baumdicker, Hess, and Pfaffelhuber 2012; Lapierre and Gogarten 2009), leading to an enormous gene content diversity within and between species. The pangenome concept addresses this extensive variation in gene content among members of the same lineage. A pangenome is the non-redundant set of genes present in group of genomes (Tettelin et al. 2005). All domains of life – Prokaryotes, Viruses and Eukaryotes have pangenomes, leading to challenges to delineate species especially in Prokaryotes due to high rates of recombination and HGT between phylogenetically distant clades. One of the aims of microbiome sciences is to catalog all genes, explore their functional potential, distribution (across species, strains and samples/biomes) and identify how environmental properties and other factors drive expansion or reduction of species' or ecosystem's gene pools and how these gene pools are linked to functioning of ecosystems and associated with phenotypes of individual organisms. This thesis, in part tackles, some of these questions. Specifically, I improve annotations of environmental preferences of microbial species (Subchapter 3.1.); investigate how gene content diversity in prokaryotes is driven by the environmental preferences (Subchapter 3.2.); search for universal nucleotide identity threshold for species delineation at gene level (Subchapter 3.3.) and discuss how problems in species delineation are inherited into identification of meaningful within species groups (Subchapter 3.4.).

1.2. Pangenomes in context of microbial ecology and evolution

Bacterial and archaeal genomes/pangenomes are "open" entities that experience gene gain (via horizontal gene transfer (HGT), duplications and *de novo* origin) and gene loss (Puigbò et al. 2014). Genome/pangenome features (size and diversity) are further shaped by varying strength of selection in combination with the genetic drift (Bentley and Parkhill 2004). Responses to these evolutionary factors depend, among other parameters, on the effective population size of species which is defined at least partially by habitat preferences (MartÃ-nez-Cano et al. 2015; Batut et al. 2014). Habitat preferences are defined as set of habitats where species are observed. Ubiquity is defined (sometimes also referred as environmental range (Barberán et al. 2014)) as a count of habitats in which species is present (B. J. Finlay and Clarke 1999; Fenchel and Finlay 2004; Whitfield 2005; O'Malley 2008). Genome size in general

known to be affected by habitat preferences. For example, some of the free-living bacteria in soil tend to have extremely large genomes (Guieysse 2012) while free-living bacteria in marine habitat (Ghai et al. 2013; Giovannoni, Cameron Thrash, and Temperton 2014) and intracellular symbionts tend to have the smallest ones (Hessen et al. 2010; Lynch 2006). Obligate symbiotic species tend to have small pangenomes – almost equal to the genome size. Soil-associated and some highly abundant marine bacteria tend to have the largest pangenomes (Rouli et al. 2015).

Passive dispersal in complex habitats such as soil causes an increase of genome/pangenome sizes possibly due to range expansion in the combination with the constant acquisition of new genes via HGT and their increase in frequency across the population due to bottlenecks (Choudoir et al. 2017). It has been shown that HGT can lead to an increase of genome sizes (Paquola et al. 2018), while, for example, the deletion bias (higher deletion than insertions/duplication rates) in bacteria leads to a reduction (or preservation) of genome size (Nilsson et al. 2005). In marine bacteria, genome streamlining and large pangenome are likely adaptive and a consequence of natural selection and incomplete selective sweeps (Grote et al. 2012; Giovannoni, Cameron Thrash, and Temperton 2014). In obligate symbionts, small genomes and pangenomes are rather result of drift caused by reduced effective population size (McCutcheon, McDonald, and Moran 2009b). However, most species are abundant in multiple habitats and host association or free-living state can be facultative which leads to multidirectional pressures on genome structure evolution. Broader ecological niches and higher ubiquity tend to be associated with larger and more functionally versatile genomes (Cobo-Simón and Tamames 2017). Highly ubiquitous species tend to accumulate more accessory genes (in particular due to HGT) to deal with environmental factors and to interact and compete with other species (Juhas et al. 2009; Bobay and Ochman 2017b).

Past exposure of a species to different habitats is likely to set constraints on habitat preferences and the evolution of phenotype traits and genome/pangenome features. Phylogenetic relatedness is predictive of some of microbial phenotypic traits (Goberna and Verdú 2016) because they tend to be somewhat phylogenetically conserved (e.g. spore-formation, Gram-staining, cell shape, photosynthesis, methanogenesis (A. C. Martiny, Treseder, and Pusch 2013; J. B. H. Martiny et al. 2015; Barberán et al. 2017)). Closely related species tend to share more genes, in other words, gene content similarity follows phylogeny reconstructed with 16S rRNA genes (Snel, Bork, and Huynen 1999; Konstantinidis and Tiedje 2005). Functional gene content contains niche and clade signal (Zhang and Sievert 2014), consequently phylogenetic relatedness and genome functionality are mildly predictive of species ubiquity and genome size (Barberán et al. 2014; Tamames et al. 2016). Moreover, habitat and phylogeny shape nucleotide composition of prokaryotes (Reichenberger et al. 2015;

Hellweger, Huang, and Luo 2018). Namely, GC content (Foerstner et al. 2005) and codon usage are affected by habitat (Roller et al. 2013). Habitat preferences are also phylogenetically predetermined (von Mering et al. 2007) and dispersal capability also varies across different taxa (Choudoir et al. 2018; Delgado-Baquerizo et al. 2018).

In conclusion, genomic parameters and phenotypic traits exhibit phylogenetic and environmental association. At the same time habitat preferences are also phylogenetically predetermined. Since, closely related species have possibly speciated in similar environment they would harbor similar genomes because of phylogenetic inertia and have alike habitat preference. Species from distant clades with similar habitat preferences will evolve similar features due to convergent evolution. For example, species in major phylogenetic groups that explore host-associated habitat, convergently, harbor reduced genomes. Also, at least 130 genera in 13 phyla evolved large genome size independently (Guieysse 2012). Given that there are many (10^5 – 10^{12}) unknown microbial species, predictive analysis of ecological niche is highly important. In particular, it will be possible to predict host-association and cultivation requirements of species from metagenome samples using (meta-)pangenomics approach, as well as to understand general principles of evolution and functioning of ecosystems at microscale. In this thesis I explore to what extent environmental preferences and phylogenetic inertia can explain pangenome structure (Subchapter 3.2.). This study provides fundamental framework for future predictive analysis of ecological niches and environmental preferences of unknown prokaryotic genomes and species.

1.3. Delineation of species and strains in microbiome

Quantifying and classifying the biodiversity are important topics in classic biological fields such as botany, zoology, microbiology and virology. “Species” is considered to be the central unit of biodiversity. However, species delineation has been a problem in biology in general and in microbiology in particular. It has been questioned whether species are real evolutionary and ecologically meaningful and coherent units and do they have a natural boundary. More than 20 species concepts have been proposed to address these questions (Wilkins 2003; Mayden 1997; Hey 2001). In prokaryotes, the applicability of any species concept has long been considered to be limited, because they were developed for animals and plants. However, it has since become evident that some defining aspects of established species concepts are indeed approximated in microbes, in particular the so-called Biological Species Concept and phylogenetic species concepts (Konstantinidis, Ramette, and Tiedje 2006). The Biological Species Concept defines species as a group of individuals that can interbreed resulting in viable offspring (Queiroz 2005). In bacteria, homologous recombination between

cells provides an analogous mechanism. According to phylogenetic species concepts, individuals form coherent genomic clusters that are characterized by distinctive phenotypic properties (Nixon and Wheeler 1990). These concepts predict that there is a decline of homologous recombination and decline of gene turnover between different species (Shapiro et al. 2012). Significant decline in recombination rates will result in diversification of population into new species. This process will be accompanied by decline of nucleotide identity because of accumulation of distinct mutation patterns. Recently, it has been demonstrated at the genomic level that species in this sense may exist at least in some bacterial clades: (1) using core genome recombination rates (Bobay and Ochman 2017a), (2) FastANI comparisons (Jain et al. 2018), (3) gene content similarity (Moldovan and Gelfand 2018), (4) genomic ANI in metagenomes (Olm et al. 2020). Moreover, in accordance with phylogenetic coherence, species in bacteria can be operationally defined by phenotype similarity, DNA-DNA hybridization, similarity in 16S rRNA and/or in universal-single copy marker genes, as well as gene content and nucleotide identity (Konstantinidis and Tiedje 2005). Approximately ≥70% similarity in DNA–DNA hybridization corresponds to ≥95% of ANI (in core genome) and ≥96% in universal marker genes (Klappenbach et al. 2007; Konstantinidis, Ramette, and Tiedje 2006). Due to the fast accumulation of newly sequenced genomes the most scalable and efficient way to delineate species is based on the ANI of 40 (Mende et al. 2013) or 100 universal marker genes (Parks et al. 2018). This operational delineation has resulted in the reclassification of species in some cases into multiple distinct species, or in other cases combining of several species into complexes with potentially many clades/subspecies within the species (Parks et al. 2018; Mende et al. 2020). ANI thresholds are a useful operational way to define species, yet they probably are not universal for all species (Konstantinidis and Tiedje 2005). Also existence of species is questioned by other publications for various important reasons, such as that there is not enough of sampling of strains within species, incomplete sampling of different species, technical and conceptual biases, true rates of HGT and recombination are much higher compared to existing estimates (Doolittle and Zhaxybayeva 2009; Doolittle 2012; Rocha 2018). In the Subchapter 3.3. I investigate which threshold of nucleotide identity is an optimal and natural for delineating species boundaries in Prokaryotes across all genes in pangenomes. This particular gap in knowledge is important to address to find optimal threshold and apply it for gene clustering in metagenomics gene catalogs.

Due to existence of large genomic diversity between genomes of the same species, classification of ecologically, phenotypically, clinically meaningful within-species groups is another topic that is related to delineation of species in Prokaryotes. Many comparative genomics and metagenomics studies pursue in-depth exploration of within-species diversity (S.

Nayfach and Pollard 2015). However, application of terminology and concepts from older biological fields, such as microbiology, remains imprecise due technical and biological reasons. Challenges in delineation of species are also inherited into the delineation of meaningful and natural groups at intraspecific levels. Many intraspecific categories that are recognized by the International Code of Nomenclature of Prokaryotes are delineated using a combination of genomic and phenotype features (Charles T. Parker, Brian J. Tindall 2019; Brenner, Staley, and Krieg 2015), however this information is difficult to obtain, for example, within the framework of metagenomics studies. In Subchapter 3.4. of this thesis I discuss the magnitude of the problem and some of the solutions.

1.4. Pangenomes, gene catalogs and their application in metagenomics

Comparative genomics of whole genomes obtained from isolates (Tettelin et al. 2005; Medini et al. 2005) and metagenomic samples (Qin et al. 2010; L. Xiao et al. 2016; 2015) showed that the gene content of all organisms is highly variable between and within species and samples. Because a lot of genes perform similar functions and are homologs to each other, both sequencing of isolated strains in culture (whole genome shotgun sequencing) and whole genome shotgun metagenomics had developed conceptually similar redundancy reduction (also called dereplication or gene clustering) methods that enabled identification of those homologous groups across genomes and samples. This led to introduction of pangenomes and gene catalogs. Non-redundant sets of genes constructed on the metagenomic samples are called gene catalogs. Non-redundant sets of genes constructed on genomes (or isolates) are called pangenomes. Pangenomes and gene catalogs are useful “tools” to characterize the gene content of any taxonomic group and metagenomics samples, respectively. In particular, it is possible to characterize what fraction of pangenome of every taxonomic group is present in the specific samples and environments which in turn enables to perform microbiome genome-wide association studies. Such approach can help to identify genomic variants associated with, for example, microbiome-related disease, variants involved in antibiotic resistance and other phenotypes (Garud and Pollard 2020). For performing successful analysis using pangenomes and gene catalogs it is important to take into account technical and conceptual differences in underlying infrastructure of pipelines to construct them. There are >30 pipelines to construct pangenomes which usually rely on blastp search of best-hits between genomes or cd-hit clustering and sometimes others methods (J. Xiao et al. 2015; G. S. Vernikos 2020). In pangenomics, thresholds for gene clustering vary in range 30% - 95% of protein identity in different tools/pipelines and studies potentially resulting in difficulties for comparative interpretation of results from different projects (G. S. Vernikos 2020; Page et al. 2015). Gene

catalogs built using Blast-like algorithm – BLAT which is more scalable on bigger datasets than typical pangenome construction algorithms (Kent 2002; L. Xiao et al. 2015; Qin et al. 2010). Main aim of pangenomic analysis is to characterize all orthologous groups present in the species and investigate their occurrence patterns across genomes in context of phenotypes. Metagenomic gene catalogs are typically built with purpose to use them for metagenomics profiling and SNP-calling, with nucleotide identity threshold for gene clustering close to 95% (Qin et al. 2010; L. Xiao et al. 2016; 2015). Since pipelines for mapping reads that include, for example, bwa read aligner (Li and Durbin 2009) are limited in performance by 5-10% diversity threshold, it is not worth to use lower thresholds to cluster genes in gene catalogs that will be used for recruiting reads (Bush et al. 2020). Only recently it has been proven (in this thesis and several other works, discussed in Subchapter 1.3. and 3.3.) that 95% ANI might be also the optimal threshold from biological perspective for clustering genes because it is close to the natural species boundary.

Recently, distinction between pangenome and gene catalog approaches has been practically eliminated because species-centered assembly of genomes from metagenomes is now possible and pangenomes are increasingly becoming a tool to investigate metagenomic samples (Delmont and Eren 2018). A comprehensive meta-pangenomic study will include some or all of the following steps 1) assembly and binning of metagenomic sequencing reads into metagenome assembled genomes, 2) construction of pangenomes using monoisolate whole genomes and metagenome-assembled genomes, 3) construction of gene catalog for all genes that were not possible to link to species; 4) mapping of reads to the resulting pangenomes / gene catalogs to identify the presence of genes in individual samples. This approach enables a species-centered annotation of functions in community or ecosystem, investigates population structure of bacteria at intraspecific level, and as a result, enables metagenomic genome-wide association studies to link, for example, host disease phenotype and microbiome abundances, prevalence and genomic signatures (Wang and Jia 2016; Delmont and Eren 2018; Denef 2018; Scholz et al. 2016; Garud and Pollard 2020).

Metagenomics does not yet provide full breadth of coverage of individual genomes of species in microbial community but provides better estimate of overall diversity of uncultivated species and strains (Teeling and Glöckner 2012). Whole genome sequencing approach provides sufficient breadth of coverage for a single isolate, yet, it fails to capture broad strain-level diversity of known and unknown, cultivable and uncultivable microorganisms. To benefit from advantages of both approaches, several studies have used pangenome (and/or gene catalog) of species (instead of single references) to investigate variation in gene content between human microbiomes (Ma et al. 2020; Zhu et al. 2015), investigate strain-level structure of *Escherichia*

coli O104:H4 outbreaks and ecotypes of marine communities from different regions (Scholz et al. 2016; Stephen Nayfach et al. 2016; Delmont and Eren 2018). Mapping of metagenomic reads on pangenome of species allows to estimate which part of known pangenome "space" of species is present in microbial community. With advances of metagenomic assembly of genomes it is now also possible to add new genes to pangenomes from metagenomics samples, however this comes with the cost of quality decline because metagenomic genomes might be contaminated with other species (Shaiber and Eren 2019). This information leads to better understanding of the ecological niche and pathogenic capacity of species in particular ecosystem and potentially enables to identify habitat-specific accessory genes for species that are present in many habitats (Zhang and Sievert 2014).

In summary, it is important to note that the pangenomes and the gene catalogs are complementary approaches. Each of these approaches can be used to improve each another. In this thesis I use pangenomes to investigate which nucleotide identity threshold can be used for clustering genes in metagenomics gene catalogs so that each gene cluster belongs to the same species (Subchapter 3.3.).

1.5. Outline of the thesis

In the Subchapter 3.1. of the thesis I improve and systematically characterize annotations and metadata about ecological niches and environmental preferences of microbial species.

In the Subchapter 3.2. I investigate how environmental preferences of species characterized in Subchapter 3.1. and phylogenetic inertia affect pangenome structure. Results show that ecological niche is highly important in defining pangenome structure.

In the Subchapter 3.3. I use pangenomes from Subchapter 3.2. to delineate species at the gene level in metagenomics data. This subchapter shows that the 95% nucleotide identity is biologically meaningful boundary for delineating species in bacteria and can be applied for construction of metagenomics gene catalogs.

In the Subchapter 3.4. I explore how traditional microbiology, pangenomes and knowledge in evolutionary biology might be useful for delineation of meaningful groups within species when studying metagenomics data.

Chapter 2. Methods

Methods Subchapters 2.1. - 2.5. are published as part of the manuscript "Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity" in the ISME journal; Subchapter 2.6. are methods for unpublished results; Subchapter 2.7. are methods from the manuscript "Towards the biogeography of prokaryotic genes" which is in preparation/revision.

2.1. Habitat database building

Habitat metadata of genomes were obtained from the Global Microbial Gene Catalog (http://gmgc.embl.de), Microbe Atlas Project database (https://microbeatlas.org), (Wattam et al. 2014), and the Pathosystems Resource Integration Center (PATRIC) database, resulting in the high quality comprehensive annotation of species to one or many habitats (maximum possible number is 83 habitats, see Supplementary Table 1 for definitions of habitats, Supplementary Table 2 for list of genomes, Supplementary Table 3 for habitat annotations for each species). The annotations from PATRIC database were manually curated using a predefined list of key-words (Supplementary Table 1). The species was considered to be present in the biome from Global Microbial Gene Catalog if at least 10 genes from the pangenome where present in the samples from that biome. To annotate environmental preferences using the Microbe Atlas Project dataset, we extracted 16S rRNA genes (filtered for at least 50% of the entire gene length) from the original GenBank files or, we re-annotated the rRNA genes using barrnap (Seemann n.d.) for missing annotations. Obtained 16S rRNA sequences were then mapped to the Microbe Atlas Project reference database using MAPseq (Matias Rodrigues et al. 2017) to link species clusters to Operational Taxonomic Units (OTUs) at 98% sequence similarity. Species were considered to be associated with the habitats in the Microbe Atlas Project only if Fisher's Exact Tests was significant after Benjamini-Hochberg correction for multiple testing ($p \leq 0.05$). Ubiquity was estimated as the sum of all positive associations across all habitats in the Microbe Atlas Project dataset. The final annotations are available as Supplementary Table 3.

2.2. Genomic data

In this study (Subchapter 3.2.), I used 155 species, in total 7,104 genomes. All species (specI clusters) were consistently delineated using 40 single-copy universal marker genes (Mende et al. 2013), obtained from the proGenomes database version 1 (Mende et al. 2017) (see Supplementary Table 2). This removes biases resulting from differing species definitions

in distinct research projects. To increase the reliability of further analysis, (i) we included only high-quality genomes with 300 or fewer contigs, (ii) only one genome from any pair of genomes was retained for downstream analysis when pairwise nucleotide identity in the core genome was 100% and pairwise gene content overlaps (Jaccard index) > 99%, (iii) we used only species that contained at least 10 high-quality genomes in the proGenomes database version 1 after filtering to remove highly similar genomes (Mende et al. 2017). I also created two confirmatory extended datasets that included all species in proGenomes version 1 and proGenomes version 2 (for which <10 genomes were sequenced but habitat preferences were still annotated). The first confirmatory dataset represents the proGenomes database version 1 (the same database underlying the pangenome dataset) consisting of 4,582 species with 24,223 genomes. The second dataset includes almost the entire proGenomes2 database which consists of 10,100 species with 84,022 genomes (Mende et al. 2020). For this confirmatory analyses, I computed only the average genome size for each species within each of the datasets.

2.3. Pangenome reconstruction

Pangenomes for the 155 species were constructed using the Roary pipeline (Page et al. 2015). Raw contig files of genomes were first annotated using Prokka (Seemann 2014). We identified homologous gene clusters at an amino acid identity threshold of 80% (Fedrizzi et al. 2017; Iraola et al. 2017; Batty et al. 2018; Kavvas et al. 2018). Pangenome and core genome curves (rarefaction curves) were generated using 30 random permutation orders of adding genomes (Supplementary Figure 1) (similar to the approach in the GET_HOMOLOGUES pangenome pipeline (Contreras-Moreira and Vinuesa 2013)). Fitting of non-linear regressions was performed in R statistical programming language v.3.3.2 (R Core Team 2018) using the "nls package" (Baty et al. 2015). The total number of genes in the pangenome of a species, the number of new genes that are added per genome and the total number of core genes were modeled using equations below {1}, {2} and {3} respectively to estimate the openness of pangenomes (Tettelin et al. 2005; Medini et al. 2005).

{1} $G = kN^{\gamma} + c$,

{2} $G = kN^{-\alpha}$,

{3} $G = ke^{-N*\gamma} + c$,

where G – number of genes; N – genome number that is added to analysis; k, c, - constants; α and γ – saturation coefficients. When $\gamma \leq 0$ in equation {1} – pangenome is closed (saturated); $0 << \gamma \leq 1$ – pangenome is open. When $\alpha < 1$ in {2} – pangenome is open, $\alpha > 1$ – pangenome is closed.

The thresholds for classification of pangenome components were defined in the following way: core genes – present in all strains; extended core – present in > 90% of genomes; cloud genes – present in < 15% (includes unique genes in pangenome); the remaining part of pangenome were considered "shell" genes (Supplementary Figure 2). These thresholds are based on default parameters of the Roary pipeline (Page et al. 2015), although we readjusted the extended core threshold to 90%, as suggested by the distribution frequency of genes within the pangenomes in our dataset (Supplementary Figure 2). The R package "micropan" (Snipen and Liland 2015) was used to compute genomic fluidity (Kislyuk et al. 2011), Chao's lower bound for gene content in the pangenome (Chao 1987) and Heaps' alpha (equation {2}) (Tettelin et al. 2005). Functional distance between strains within each pangenome was estimated as Jaccard distance based on EggNog v4.5 annotations (Huerta-Cepas et al. 2016) of pangenome gene clusters. 23 parameters (21 pangenome features, the number of conspecific isolates and species ubiquity) were compared using Spearman's rank correlation. This comparison was performed to investigate the associations between sample sizes, components of the pangenomes, the saturation parameters (γ and α) in the equations {1}, {2}, {3}, genome fluidity, functional distance and average core genome nucleotide identity (see Supplementary Table 1 for definitions of pangenome features). To obtain unbiased estimates of core and pangenome sizes we calculated average core and pangenome sizes across 30 random combinations of 9 genomes for each species (also see Supplementary Table 1 for definitions and Supplementary Figure 1). Hierarchical clustering of a subset of pangenome features was performed on absolute values of pairwise Spearman Rho values as displayed in Figure 3.1.2a.

2.4. Phylogenetic signal and phylogenetic generalized least squares

An approximate maximum likelihood phylogenetic tree of all 155 species was generated using the *ete-build* concatenation workflow "clustalo_default-trimal01-none-none" and "sptree-fasttree-all" from ETE Toolkit v3.1.1 (Huerta-Cepas, Serra, and Bork 2016), using protein sequences of 40 conserved nearly-universal single copy marker genes (Mende et al. 2013; Sorek et al. 2007; Ciccarelli et al. 2006) and default parameters in the ClustalOmega aligner (Sievers et al. 2011) and FastTree2 (Price, Dehal, and Arkin 2010) with the JTT model (Jones, Taylor, and Thornton 1992).

To estimate the phylogenetic signal of genomic traits, we used the R package "phylosignal" (Keck et al. 2016) with Pagel's Lambda (Pagel 1999), following guidelines for phylogenetic signal analysis (Münkemüller et al. 2012; Symonds and Blomberg 2014)

(Supplementary Figure 3). The “Caper” R-package was used for phylogenetic generalized least squares regression (Orme 2013).

2.5. Quantification of explained variance in pangenome features

The phylogenetic tree was used to calculated cophenetic distance matrix. The cophenetic distance matrix and the binary habitat association matrix (83 habitat annotations originating from Microbe Atlas, Global Microbial Gene Catalog and proGenoems1/2) were each decomposed using the “FactoMineR” R package (Lê, Josse, and Husson 2008). PCs were selected using the “broken stick” model (Borcard, Gillet, and Legendre 2011). Only the first 5 phylogenetic principal components, accounting for ~80% of phylogenetic variance, and 10 habitat PCs, accounting for ~50% of habitat variance, were used for variance partitioning. The first two principal components for phylogenetic and habitat matrices decompositions are visualized in Supplementary Figure 4 and Supplementary Figure 5. To reduce the influence of different sample sizes (see equations {4}, {5}, {6}), the number of genomes that was used for each species were included as an additional predictor variable. The fraction of the variance explained by habitat and phylogeny were estimated using the CAR metric which performs a decorrelation of predictors (Zuber and Strimmer 2011) implemented in the “care” and “relaimpo” R-package (Groemping 2006) with the following models using function “calc.relimp” from “relaimpo” R-package:

{4} Pangenome feature = number of genomes in each species + 5 phylogenetic PCs + 10 habitat PCs

{5} Pangenome feature = number of genomes in each species + genome size + 5 phylogenetic PCs + 10 habitat PCs

{6} Pangenome feature = number of genomes in each species + core genome nucleotide diversity + 5 phylogenetic PCs + 10 habitat PCs

To investigate robustness of the estimates of explained variance, I fitted the model from equation {4} in 1,000 row-wise permutations of the first 10 habitat PCs and 5 phylogenetic PCs to ensure that the actual habitats and phylogeny data explain a higher fraction of the variance than randomized models (Supplementary Figure 6).

2.6. Estimation of natural selection and recombination rates

Rate of nonsynonymous and synonymous substitutions(dN/dS) was estimated using PAML pipeline (Yang 2007). Due to memory and time limitations, 30 species with more than 80 genomes were randomly subsampled to contain only 80 genomes. Estimation of the recombination rates between strains within each species was performed using FastGear software (Mostowy et al. 2017). For downstream analysis, we only used the estimate of number recent recombination events.

2.7. Investigating species boundary for global microbial gene catalog

Genes were predicted and annotated in Prokka. Blastn (Nucleotide-Nucleotide BLAST 2.2.29+) searches were performed on 107 species (specI clusters) which belong to 32 genera. Each species had at least 10 genomes. Species that contained more than 20 genomes were randomly down-sampled to 20 genomes. I performed blastn search of all genes in each genome against other genomes of the same species or against genomes form species from the same genus. Nucleotide identity in the Figure 3.3.1 is the average of all identities of gene matches in the pair of genomes. In total I performed 14,686 pairwise genome-comparisons within species and 51,368 comparisons between species within Genera.

Chapter 3. Results and discussion

3.1. Habitat resource development for genomes and metagenomes

This subchapter describes construction of microbial habitat preference resource whose utility is described in subsequent chapters of this thesis. In this project I performed (1) extraction of 16S rRNA sequences from isolates genomes, (2) curation of isolation sources of genomes in proGenomes2 database, and (3) inferring habitat preferences of species using genes from pangenomes linked to genes in metagenomic gene catalog derived from known biomes. Habitat preference data is published in part with proGenomes2 database (Mende et al. 2020).

Mende DR, Letunic I, **Maistrenko OM**, et al (2020) proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Research*

3.1.1 Introduction

With an exponential accumulation of sequenced microbial genomes and discovery of numerous new species it is ever so important to appropriately catalogue and curate the associated isolates' metadata (for example, year of isolation, isolation source, habitat preferences, disease association, Gram staining and others). Availability of metadata enables large scale comparative analyses in genomics and microbial ecology, databases such as ENA (Amid et al. 2020), NCBI, IMG (Chen et al. 2019), GOLD (Mukherjee et al. 2018), PATRIC (Wattam et al. 2014), ProTraits (Brbić et al. 2016) and others make this possible. However, for the convenience of the data depositor metadata submission in most databases is optional, unstructured and flexible (or sometimes too specific or not specific), making any downstream comparative analysis challenging due to incomplete and unclassifiable metadata. One such problem concerns isolation source (or habitat preferences) of sequenced genomes and species, wherein routinely the place/habitat from which microbial species is isolated is missing. Along with this biological and technical limitations such as biased sampling of cultivable host-associated or disease-associated species and strains further limit our understanding of real ecological niche and environmental preference of a species and thus hinder comparative studies.

The main aim of this subchapter is unbiased cataloguing and curation of habitat data. The resource made available in this subchapter enables identification of primary and secondary habitats of species, study of habitat connectivity via gene transfer, exploration of species dispersal, identification of origin of pathogenic species and association studies that link habitats to antimicrobial resistance and virulence capacities of species. In particular, in this thesis

proGenomes2 environmental data was used for studying impact of phylogenetic and environmental constraints on pangenome structure and effects of habitat preferences on recombination rate and natural selection (Subchapter 3.2.).

3.1.2. Environmental preferences of microbial species

To address the problem of non-universal and non-standardized metadata information, I first screened for optimal groups of habitats terms and categories that can be assigned universally to as many genomes and species as possible which included broad categories such as aquatic, soil/terrestrial and host-associated and subcategories, for example freshwater, sediment and disease-associated. Then I curated the isolation source using set of keywords, for an extended list of these keywords refer to Supplementary Table 1. For example, to assign strain or species to "aquatic" habitat following key-words in PATRIC database metadata were used: "aquatic", "marine", "lake", "spring", "river" and so on. As a result, ~60,000 out of >80,000 genomes were assigned to at least one of habitat and phenotype entries, and >8000/12221 species getting at least one habitat annotation. The host-associated habitat annotations as a result of bias in sequencing studies were overrepresented as compared to soil and aquatic habitats (Figure 3.1.1a, Figure 3.1.1b). This categorization also allowed inclusion of habitat sub-categories that provided better habitat characterization, since aquatic habitat could now include freshwater as a sub-category (Figure 3.1.1b).

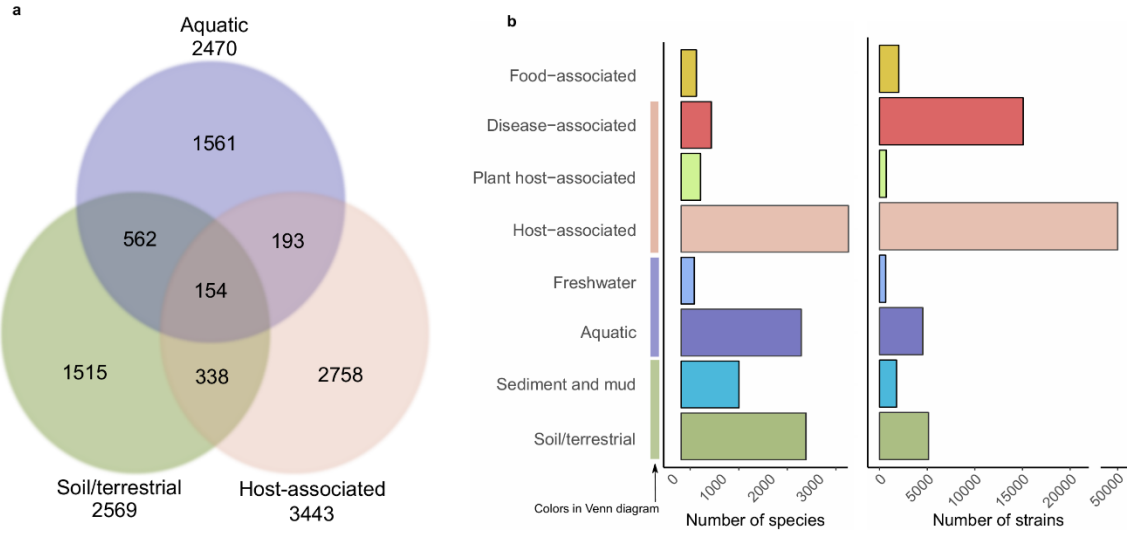


Figure 3.1.1. Summary of environmental preferences annotations of 12221 species, >87000 isolates in curated PATRIC database metadata. (Figure and description to the figure are in part published in (Mende et al. 2020) in the Nucleic Acids Research journal, 2020 Oxford University Press)

3.1.3. Discussion

It is well known that many host-associated species might inhabit multiple habitats and their habitats can be further classified as their primary or temporary/secondary niches (Leibold et al. 2004). Hence, inferring species' niche only from isolation source does not always lead to exhaustive and accurate understanding of ecological niche of a species. To address this problem, we extracted 16S rRNA gene sequences from whole genome sequencing data and screened for them in the 16S rRNA environmental metagenomics datasets with known environment/habitat annotation. We also screened for individual genes from pangenomes in the metagenomic gene catalog which enabled to link individual genes to biomes. General patterns of environmental preferences inferred from these different sources showed overall agreement and complemented with each other (Figure 3.1.2.). We found that host-associated and free-living habitats (terrestrial/soil and aquatic) tend to correlate negatively, implicating that free-living and host-associated lifestyles are to some extent mutually exclusive. However, this trend does not hold for invertebrate animals, they tend to share microbial species with aquatic and soil environments, indicating that symbiotic bacteria in those groups of animals might be transient and temporary (Whitaker et al. 2016; Hammer, Sanders, and Fierer 2019). Thus, the comprehensive approach presented here enables accurate habitat characterization and applications in the fields of genomics and metagenomics.

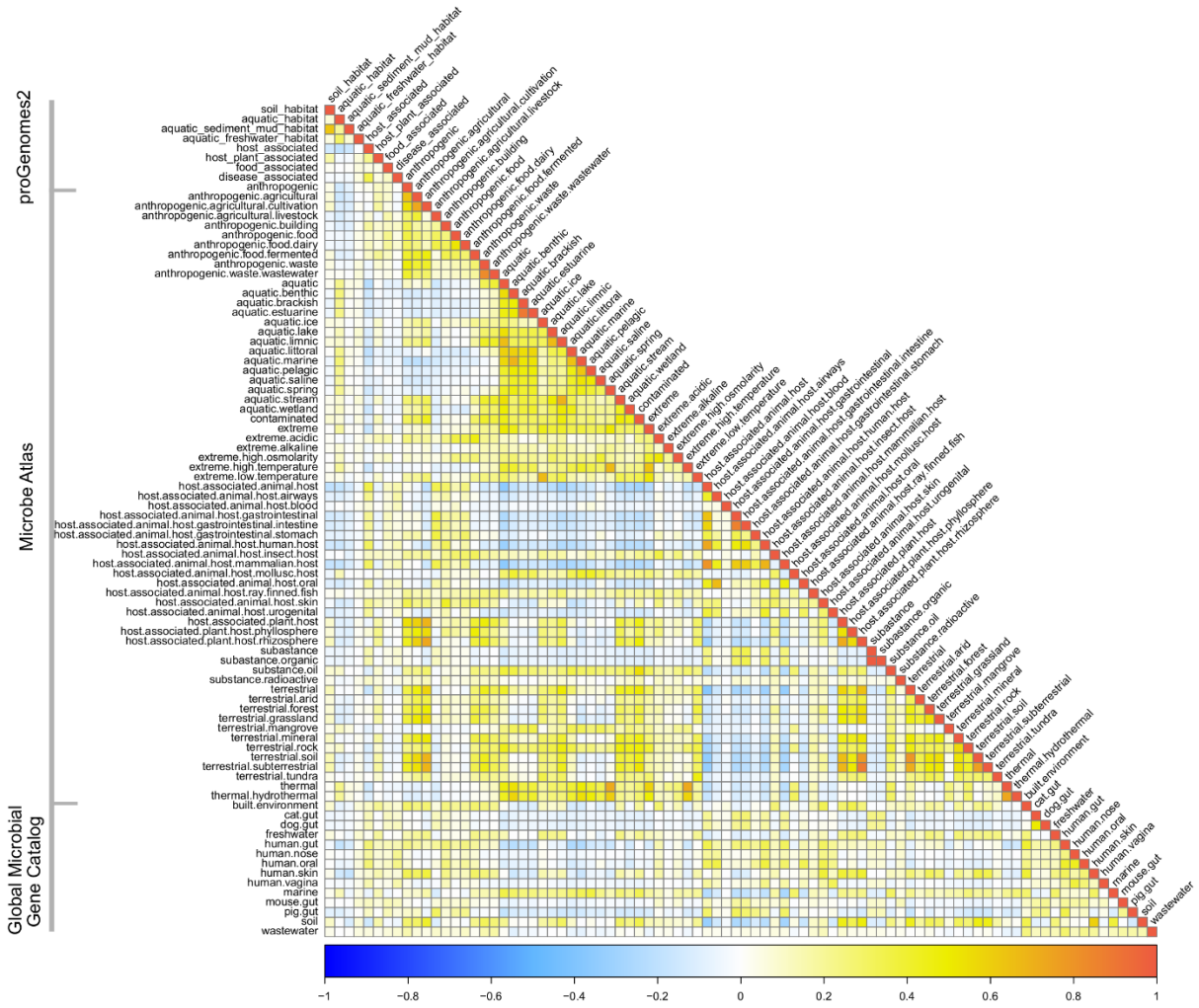


Figure 3.1.2. Pearson correlation of habitats (based on species' associations with them) from three sources used in the project (curated PATRIC habitat data, 16S rRNA based microbial atlas and Global Microbial Gene Catalog).

3.1.4. Conclusion and future outlook

Advances in database development and curation are important for further progress in microbial ecology and other fields. Comprehensive databases will enable confirmatory analysis of previously observed trends on limited datasets, discovery of new trends and predictive analysis. High quality habitat annotations will help investigation of circulation of genetic material via HGT in ecosystems and species, dispersal of species and specific strains, sources of emerging pathogens and antibiotic resistance or virulence genes. Some of the topics are tackled in this thesis, for example, in Subchapter 3.2. I ask a fundamental question about the extent to which habitat preferences drive expansion of species gene pools (pangenomes) in bacteria.

3.2. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity

This subchapter presents the main result of the thesis. I investigated the impact of environmental preferences on pangenome structure and how phylogenetic inertia limits expansion or reduction of size and diversity of pangenomes. Pangenomes were constructed by Daniel Mende and me. Curation of environmental metadata was performed by Simone Li, Sebastian Schmidt, Luis Coelho and me. I performed majority of the downstream analysis, interpretation of data, preparation of figures and writing the manuscript. Natural selection and recombination rates were calculated by Mechthild Luetge and Falk Hildebrand (this data is unpublished). Majority of the results described in this subchapter were published in:

Maistrenko OM, Mende DR, Luetge M, et al (2020) Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. The ISME Journal.

3.2.1. Introduction

Across sequenced microbial species there is approximately a 100-fold variation in genome size, at extremes: the smallest known genomes are of endosymbiotic bacteria *Candidatus Tremblaya princeps* with 116 protein coding genes and genome length 139Kbp (McCutcheon and von Dohlen 2011), and *Ca. Nasuia deltocephalinicola* with 137 protein coding genes and genome length 112Kbp (Bennett and Moran 2013), the largest known genome is of the soil bacterium *Sorangium cellulosum* which encodes >9000 protein-coding genes and is >13Mbp in size (Schneiker et al. 2007; Han et al. 2013). Genomes, when compared within species, can harbor from several new genes up to more than 300 new genes per genome (McInerney, McNally, and O'Connell 2017a). This tremendous variation in the gene content is described using the pangenome concept. The pangenome is a non-redundant set of all genes, gene clusters or homologous groups found in a certain (group of isolates/genomes of a) taxonomic group, e.g. species (Tettelin et al. 2005; Medini et al. 2005; G. Vernikos et al. 2015). Pangenomes are subdivided into core genes - genes that are present in almost all isolates and accessory genes - genes that are missing in many isolates. Remaining genes within a pangenome are called accessory (100% < extended core > ~95% > shell > ~15% < cloud). If every new genome adds new accessory genes, pangenome is called "open" and when most of genes from a species can be cataloged with few genomes, pangenome is called "closed" (Lapierre and

Gogarten 2009). Inference about pangenome openness is performed using pangenome saturation curves (Supplementary Figure 1). It is still unclear how much different habitat preferences and species ubiquity affect pangenome features such as sizes of core and accessory genomes ("shell" and "cloud"), pangenome openness, genome fluidity and core genome nucleotide diversity. In other words, the relative contributions of phylogenetic conservation and habitat preference to the evolution of pangenome features remains unexplored. Taking advantage of accumulated whole genome sequencing and habitat preference data, I present a meta-analysis study covering pangenomes of 155 species spread over 10 prokaryotic phyla with the aim of a) investigating relationship between pangenome features and ubiquity of species and b) quantifying the amount of variation in pangenome features explained by phylogenetic inertia and habitat preferences.

3.2.2. Overview of pangenome features. Ubiquity of species is related to core genome size

Each species in the analysis contained at least 10 and up 200 high quality and diverse genomes (Figure 3.2.1). We annotated habitat preferences for each species using the PATRIC database (Wattam et al. 2014), the Microbial Atlas Project Database (https://beta.microbeatlas.org/) (JF Matias Rodrigues et al, *in preparation)* and a Global Microbial Gene Catalog (http://gmgc.embl.de/) (LP Coelho et al, *in preparation),* resulting in annotations of a total of 83 habitats. Ubiquity of species was estimated as the sum of all positive associations (Benjamini-Hochberg-corrected Fisher's Exact Tests, $p \leq 0.05$) with each habitat in the Microbial Atlas Project dataset. Typically, ubiquity of species varied from 1 up to 30 habitats (out of 60 possible).

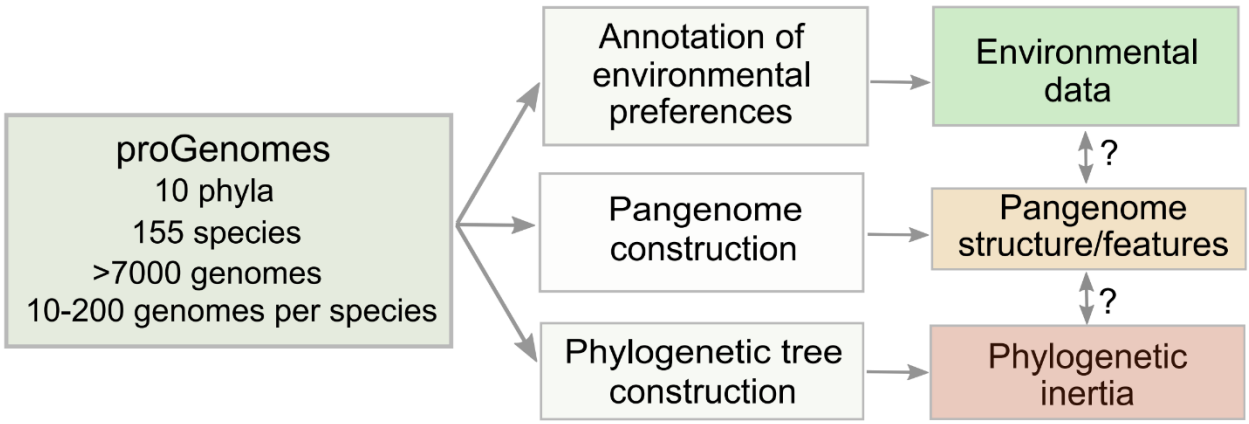


Figure 3.2.1. Study design. We used the proGenomes database version 1 (Mende et al. 2017) of high-quality genomes to compute pangenomes (using the Roary pipeline) and pangenome

features. Species were assigned to their preferred habitats using three databases: PATRIC, Microbe Atlas Project and Global Microbial Gene Catalog (see Methods). As many pangenome features are interdependent or affected by sampling size bias, I used a multivariate analysis framework to disentangle habitat properties from phylogenetic inertia. This allows for the quantification of environmental and phylogenetic factors that impact diversity within species. To construct the phylogenetic tree, we used the concatenated protein sequences of 40 conserved universal marker genes which were aligned using the ClustalOmega aligner (default parameters). The tree was constructed using FastTree2 (JTT model). (Figure and description to the figure are published in (Maistrenko et al. 2020) in the ISME journal, 2020 Springer Nature Limited).

I estimated 22 pangenome features for 155 species originating from 10 Phyla (Figure 3.2.1, Figure 3.2.2a). We observed effect of sample size on different pangenome features such as core genome saturation and sizes of cloud, unique, extended core and pangenome. Because of this sampling bias, for downstream analysis we selected less biased features according to correlational analysis (Figure 3.2.2a) and included sample size as a predictor into the regression model. To get unbiased estimates of core and pangenome size we used pangenome and core genome size calculated at 9 genomes (see Supplementary Figure 1) from random permutation orders of genomes generated as suggested in GET_HOMOLOGUES tool (Contreras-Moreira and Vinuesa 2013).

Overall the observed variation in genome size and genome fluidity was high - the largest genome in our dataset was the *Streptomyces sp* complex (species cluster -specI cluster 244 in proGenomes version 1) 7174±383 genes and the smallest was the *Tropheryma whipplei* (species cluster 1011) 833±27 genes. The genome fluidity varied from 0.01896±0.0091 in *Chlamydia trachomatis* up to 0.324±0.043 in *Bacteroides fragilis/ovatus* species complex (Supplementary Table 3). These genome fluidity estimates are in line with previous reports, suggesting that species complexes and larger genomes are more diverse in their gene content (Kislyuk et al. 2011; Andreani, Hesse, and Vos 2017). Another more qualitative way to infer size of the pangenome is to estimate the coefficient “alpha” from Heaps law equation (equation [2] in Methods). This coefficient allows to tell if pangenome is “open” (nearly infinite) or “closed”. This way I show that six pangenomes of host-associated species were “closed”: *Neisseria gonorrhoeae*, *Mannheimia haemolytica*, *Methanobrevibacter smithii*, *Lactobacillus acidophilus*, *Staphylococcus aureus* and *Streptococcus thermophilus*. In general, it is expected

that host-associated bacteria can have "closed" pangenomes because they are affected by genome reduction trend due to relaxation of selection (Boscaro et al. 2017).

Ubiquity of species is potentially linked to pangenome features such as pangenome saturation. In other words, it is expected that highly ubiquitous species have larger pangenomes that are less saturated. Indeed, we observed a weak effect of species ubiquity on the saturation of pangenomes ("alpha" in Heap's law model (equation [2] in Methods) (Table 1) and on the core genome size (Table 1 and Figure 3.2.2b) but not on any other pangenome features after correcting for phylogenetic effect. This result suggest that a larger core genome may be important to facilitate persistence/proliferation in multiple habitats. Previously, it has been shown that there is a weak positive relationship between the ubiquity of species and genome size. Species that are present in many habitats contain more genes to interact with the environment (Cobo-Simón and Tamames 2017). Our observation suggests that all genes that are needed for a species to be highly ubiquitous (present in many habitats) have to be present in the core genome rather than in accessory (in other words, most of isolates have to have genes that facilitate ubiquity, at least according to our dataset). Some of previous studies showed that, for example, Cronobacter species tended to contain relatively large core genome encoding genes that enable plant association and virulence in human (Grim et al. 2013). In our dataset, functional analysis revealed that highly ubiquitous species were enriched with genes involved in Lipid Metabolism (I) and in agreement with previous studies, enriched with Secondary metabolites biosynthesis genes (Q) (Figure 3.2.2c) (Cobo-Simón and Tamames 2017). However, phylogenetic regression did not support an enrichment of any of these functional categories among core genes of highly ubiquitous species in our dataset.

Table 1. Phylogenetic generalized least squares of core genome size and pangenome saturation (Heaps alpha) and ubiquity of species.

Dependent variables	Independent variables	Estimate	s.e.	t-value	p-value	Lambda	Lambda, confidence interval
Core genome	Intercept	38.9	9.1	4.27	3.392e-05	0.98	0.96, 0.99
	Ubiquity	0.26	0.078	3.36	0.000984		
Saturation (alpha coefficient)	Intercept	0.89	5.9680e-02	14.94	0.00000	0.317	0, 0.71
	Ubiquity	-0.0055	1.6247e-03	-3.37	0.00095		

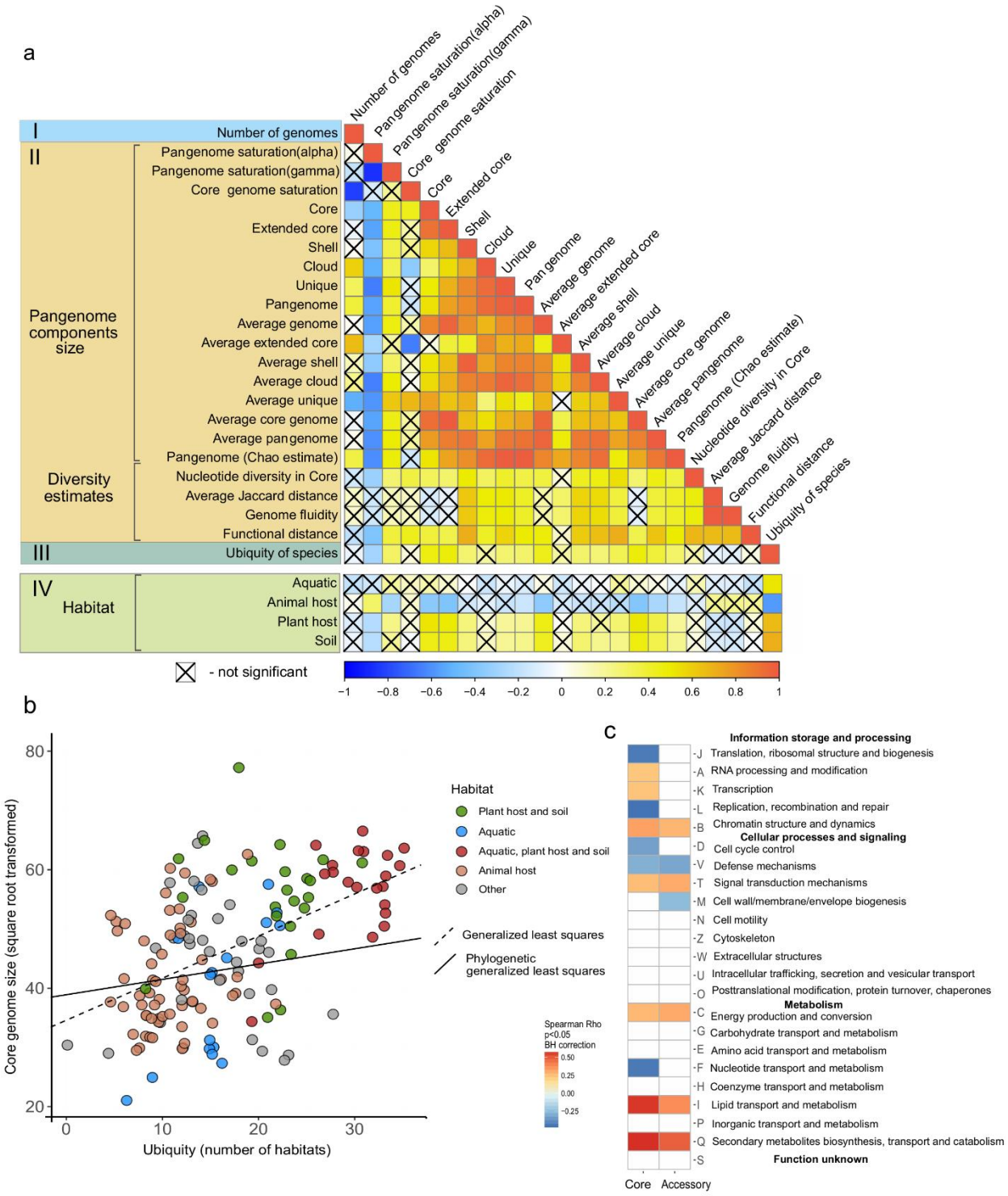


Figure 3.2.2. Relationship between different pangenome features. **a.** Correlation matrix between (I) the number of conspecific genomes used to estimate pangenome features, (II) 21 pangenome features, (III) the ubiquity of species as an environmental feature computed from habitat preference of strains, and (IV) major habitat groups from the Microbial Atlas project. The heatmap visualizes Spearman Rho values for correlations between sample size (I), 21 pangenome features (II) and species ubiquity (III). Four major habitats (aquatic, animal host,

plant host, soil (IV)) were correlated to the (I) number of conspecific genomes, (II) pangenome features and (III) ubiquity via point-biserial correlation. Statistical significance of correlations was determined using adjusted p-values (using Benjamin-Hochberg correction) < 0.05. Raw correlation data is in Supplementary Table 4. **b.** Effect of ubiquity on core genome size and functional content. Species ubiquity (number of habitats a species was assigned to), a habitat feature, is linked to core genome sizes after correction for phylogenetic effect (Phylogenetic generalized least squares, p-value=0.00005, λ=0.98 (95% C.I. 0.957, 0.992), partial R-square (for ubiquity coefficient) 0.09, see also Supplementary Table 5). **c.** Correlation of ubiquity with the relative frequency of functional categories (COG categories assigned by EggNog v4.5) in core and accessory genomes. Species of high ubiquity tend to encode more proteins involved in Lipid Metabolism (I) and Secondary Metabolite Biosynthesis (Q). (Figure and description to the figure are published in (Maistrenko et al. 2020) in the ISME journal, 2020 Springer Nature Limited).

3.2.3. Effect of habitat and phylogeny on pangenome features

Since pangenome features are correlated with each other (Figure 3.2.2a) we selected for variance partitioning only conceptually different and the least biased features as determined by correlation to number of genomes (sample size) (see Figure 3.2.3 select list of features). Pangenome features were modeled as a function of the a) number of genomes used to compute the feature, b) principal components of multidimensional scaling of the cophenetic distance matrix of the phylogenetic tree and c) principal components of multidimensional scaling of the binary matrix of species' associations with habitats. The phylogenetic effect and habitat preferences were thus represented in the linear model by 5 and 10 principal components, respectively (selected using a "broken stick" model approach):

Pangenome feature ~ Number of genomes + 5 phylogenetic PCs + 10 habitat PCs.

Both habitat and phylogenetic effects explained variance in all selected features (Figure 3.2.3, Supplementary Table 6). Habitat preferences tended to have stronger effects on pangenome features than phylogenetic inertia. The accessory genome size was less affected by habitat preference of species than core genome size. Variance in pangenome features explained by randomly permutated scores from habitat and phylogenetic principal components did not exceed variance explained by the actual data (except for genome fluidity for which randomized habitat and phylogenetic PCs explained more variance than actual). Phylogenetic component

of variance explained more variance compared to randomized data only in core and average genome sizes (Supplementary Figure 6). Phylogenetic conservation of pangenome features in our dataset was also confirmed with test for phylogenetic signal – Pagel's Lambda estimate (Pagel 1999). The strongest phylogenetic effects were observed for the average core, pangenome and genome sizes. Accessory genome (shell and cloud) size and average functional distances between genomes within species were less affected by phylogeny (Supplementary Figure 3). This is in line with results on random data were only two pangenome features (core genome size and average genome size) had more variance explained by phylogenetic component than random data (Supplementary Figure 6).

Stronger phylogenetic conservation of core genome is likely due to the fact that core genome consists of essential genes that are under stronger negative selection pressure (Bolotin et al. 2016; Rodriguez-Valera and Ussery 2012; Bohlin et al. 2017), which leads to vertical "heritability" of its content and size from ancestral species to descendants during speciation events. "Heritability" of core genome size at macroevolutionary scale (above species level) breaks-down when closely related species occupy different ecological niches. For example, *Burkholderia rhizoxinica* which is an endosymbiont of the fungus *Rhizopus microsporus* has 3,878 genes (Lackner et al. 2011) while soil free-living (and sometimes plant pathogen) *Burkholderia gladioli* has 7,410 genes (Seo et al. 2011). Even within similar niche there is specialization in genome size, e.g. obligate symbiont *Treponema pallidum* has 1,031 genes (Fraser et al. 1998) and *Treponema denticola* – 2,767 genes (Seshadri et al. 2004). In this case, difference is attributed to genome reduction and lineage-specific expansions with horizontal gene transfer (Seshadri et al. 2004).

Differential preferences in habitat lead to differences in core genome sizes across Prokaryotic species (divergent evolution), while in other cases the same habitats promote evolution of the similar core genome sizes (convergent evolution) (Figure 3.2.4). In our dataset, large core genome evolved in at least four Phyla (Proteobacteria, Actinobacteria, Spirochaetes and Firmicutes) in species associated with soil (Figure 3.2.4) (Kruskal-Wallis test, chi-squared = 32.194, df = 1, p-value = 1.395e-08). Aquatic species showed higher nucleotide diversity of core genome, however habitat and phylogenetic signals were less pronounced for this feature (Figure 3.2.4) (Kruskal-Wallis test, chi-squared = 25.69, df = 1, p-value = 4.01e-07). This observed trend of high diversity in species from aquatic habitats is consistent with well-studied aquatic species, such as *Prochloroccocus* and *Synechococcus* (Biller et al. 2014).

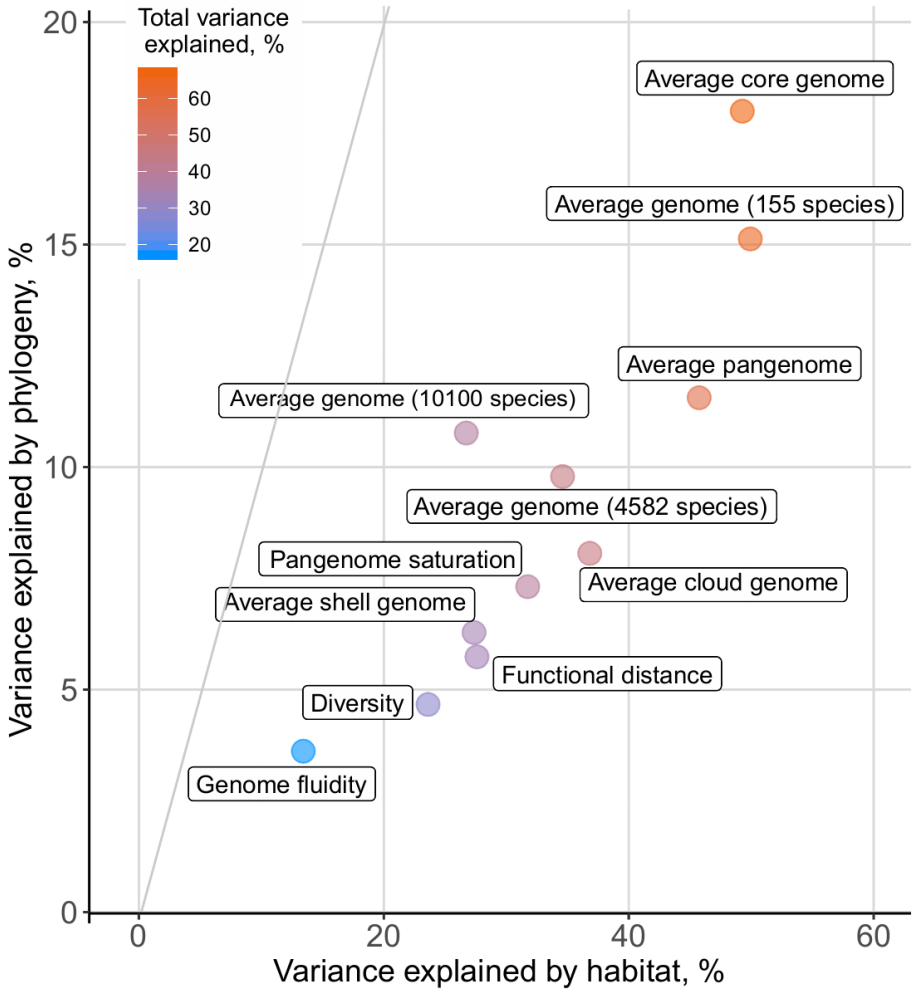


Figure 3.2.3. Partitioning of variance in pangenome features explained by phylogenetic inertia and habitat preferences (R-square(car score)) based on model {1} from Figure **3.2.5**. (Figure and description to the figure are published in (Maistrenko et al. 2020) in the ISME journal, 2020 Springer Nature Limited).

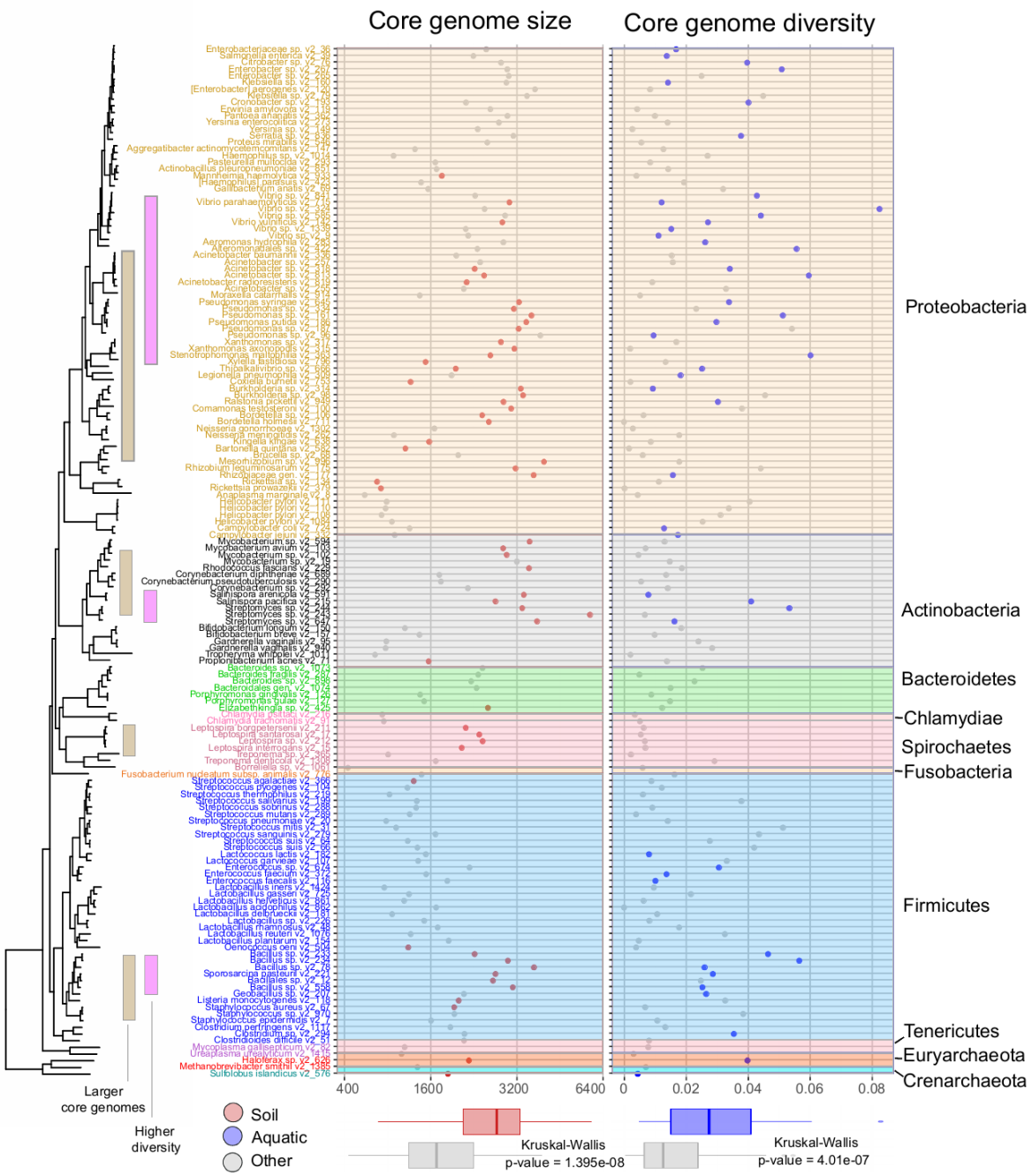


Figure 3.2.4. Phylogenetic tree of 155 microbial species with scatterplots of core genome size and average nucleotide diversity of core genomes. Soil-associated species tend to have larger core genomes (marked in red in the left scatter plot), aquatic species tend to be more diverse (marked in blue in right scatter plot). Tree labels and background of scatter plots are colored by their taxonomic annotations (phylum). Bottom panel: Relationships between habitats and core genome size and average nucleotide diversity of core genomes. (Figure and description to the figure are published in (Maistrenko et al. 2020) in the ISME journal, 2020 Springer Nature Limited).

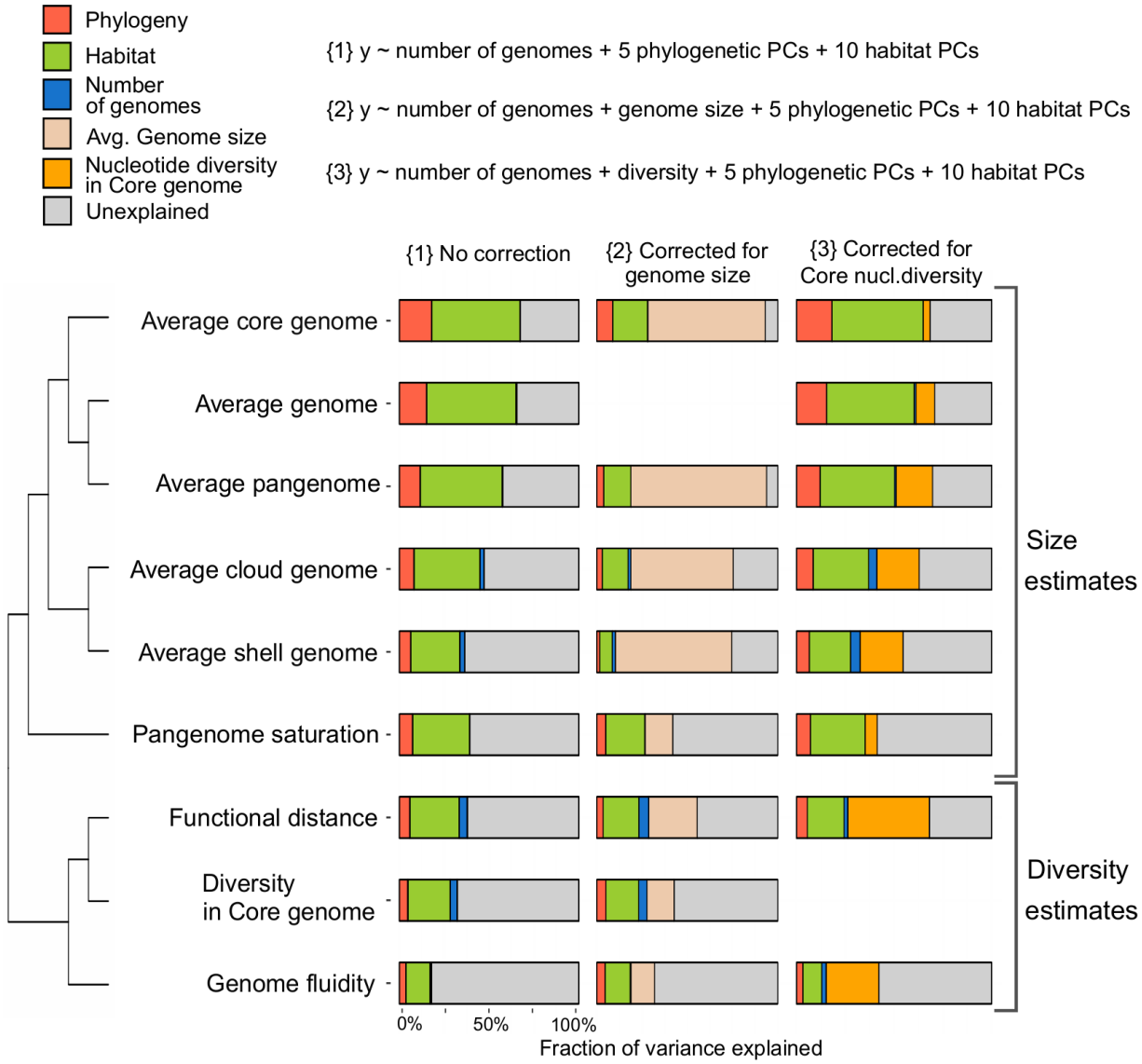


Figure 3.2.5. Clustering of a subset of nine pangenome features based on their pairwise correlation strengths. Horizontal stacked charts present amount of variance explained by various predictors (number of genomes, phylogeny and habitat represented by their principal components (PCs), and genome size or diversity). The first set of stacked charts ("no correction") shows variance explained in pangenome features by the number of genomes used to compute pangenome features as well as species' phylogeny and habitat preferences; the second and the third sets of stacked charts represent the amount of variance explained (see Methods) by the same set of predictors when correcting for genome size or nucleotide diversity in core genome respectively. Size and diversity estimates form distinct feature groups. (Figure and description to the figure are published in (Maistrenko et al. 2020) in the ISME journal, 2020 Springer Nature Limited).

It is important to take into account biases and interconnectivity between pangenome features because every feature potentially co-evolves with others. For example, large pangenome size is, probably, more likely to be seen in species with large genome. I fitted regression models that incorporated genome size or nucleotide diversity (Figure 3.2.5 equation 2 and 3). As a result, we observed that habitat and phylogeny are still explaining variance in the pangenome features even when genome size and diversity of core genome were included in the model as predictors (Figure 3.2.5). This means that there is still direct effect of habitat preferences on pangenomes size that cannot be reduced to average genome size. In general, diversity estimates are less explained by habitat preferences of species and phylogenetic inertia, compared to most of the size estimates, because they are likely reflecting spatio-temporal(microevolutionary) variation of sub-populations within species due to local adaptation and/or genetic drift (Shapiro et al. 2012). For example, pangenome analysis of 274 *Vibrio cholera* isolates revealed that pangenome structure could be attributed to variation of accessory genome (in particular mobile genetic elements) in isolates around the world and different sampling time (1910-2010 years range)(Dutilh et al. 2014).

3.2.4. Reproducibility of the observations on larger sample of species

Our observation of explained variance relies on trends of different habitats to increase or decrease genome/pangenomes sizes and diversity. Some species of bacteria are known exceptions from such trends, for example, endosymbiotic Ca. *Hodgkinia cicadicola* and Ca. *Sulcia muelleri* have high GC content in contrary to what is expected due to GC->AT mutation bias (McCutcheon, McDonald, and Moran 2009a; 2009b). Also, genome reduction in symbionts is not always true, for example, in *Ca. Endobugula sertula* (symbiont of Bryozoa) and some endohyphal bacteria genomes remain relatively large (Miller et al. 2016; Baltrus et al. 2017). Furthermore, a lot of prokaryotic species diversity is yet unknown and undiscovered (Whitman, Coleman, and Wiebe 1998; Staley and Konopka 1985) (see also Subchapter 3.3.). Hence, it is likely that more exceptions from general trends will be found which might lead to decline of explanatory power of habitat preferences and phylogeny in pangenome features. Despite this potential limitations we still observed that in larger datasets of 4,582 species from proGenomes version 1 (Mende et al. 2017) and 10,100 species from proGenomes2 (Mende et al. 2020) amount of variance explained for average genomes size by habitat was 34.6% and 26.7% and by phylogeny was 9.8% and 10.8%, respectively (raw data is provided in

Supplementary Table 7 and Supplementary Table 8). Based on this we conclude that our observations are likely to hold when more undiscovered species are considered and consequently these results in combination with functional content of genomes can be used for predicting ecological niche of species, for example, free-living or host-associated niches.

3.2.5. Discussion

Why do organisms have pangenome and are they adaptive are important and debated topics in genomics. In the debate about adaptive role of pangenome (McInerney, McNally, and O'Connell 2017a; Andreani, Hesse, and Vos 2017; Shapiro 2017; McInerney, McNally, and O'Connell 2017b; Vos and Eyre-Walker 2017; Lobkovsky, Wolf, and Koonin 2013; Baumdicker, Hess, and Pfaffelhuber 2012; Sung et al. 2012) focus on size estimates conclude that pangenomes are adaptive (McInerney, McNally, and O'Connell 2017a), while focus on diversity measures (effective population size and genome fluidity) conclude that pangenomes are rather neutral (Andreani, Hesse, and Vos 2017). Interestingly, these two sets of features, diversity estimates (core genome nucleotide diversity, functional distance, genome fluidity) and size estimates, cluster into two subgroups of pangenome features (Figure 3.2.5) based on pairwise correlations (from Figure 3.2.2a). It is possible that these two groups of features reflect different aspects of pangenome structure and should be interpreted together in the debate about adaptive role of pangenomes.

In our dataset we were able to show that accessory genes are under strong purifying selection however selection is slightly relaxed compared to core genome (Supplementary Figure 7). Phylogenetic inertia and habitat associations of pangenome features' sizes do partially explain why pangenome size could be perceived as adaptive. Habitat is likely to drive pangenome evolution by affecting dispersal (which in turn affects recombination rate and chances to acquire new genes), genetic drift and natural selection depending on the effective population size of species (Bobay and Ochman 2018). Since, number of recombination events is negatively associated with dN/dS, the dispersal of adaptive variants is potentially favored through recombination (Supplementary Figure 8a). We also observed association of number of recombination events with the rate of synonymous substitutions (dS) which is likely proportional to effective population size (Supplementary Figure 8b). Also genomic fluidity was positively associated with the number of recombination events implying that species with higher gene diversity and gene turnover have higher recombination rate and possibly bigger population size (Supplementary Figure 8c). However, at the same time pangenome diversity (or "openness") estimated using saturation curves did not have any association with recombination rate which might be a result of technical biases (Supplementary Figure 8d).

There was also no effect of habitat preferences on recombination rate and only moderate effect of habitat preferences on selection strength (Supplementary Figure 9). These observations might support the hypothesis that pangenome expansion is driven by gene turnover with incomplete elimination of acquired genes and natural selection favoring accessory genes, most likely, in a context dependent way (Choudoir et al. 2017). In conclusion, both phylogeny and habitat explain variation of pangenome features (even after correcting for genome size and diversity). Habitat preference had stronger effect on evolution of size of pangenome features compared to phylogenetic inertia. Only core genome size and total genome size are robustly exhibiting dependence on phylogenetic inertia ("heritable" during speciation). It is likely that a combination of selective and random processes contribute to expansion of pangenomes, possibly through dispersal that is followed by acquisition of new genes whose fate (retaining or elimination) in population is determined via drift and selection. Drift likely plays bigger role in accessory genome which is under slightly relaxed purifying selection compared to core genome. Future studies in this direction have to investigate patterns of selection and recombination at individual gene level to show which functions of the species drive its adaptation and habitat preferences.

3.2.6. Conclusion and future outlook

Exhaustive analysis of pangenome features in context of habitat preferences revealed that the core genome size is affected by species ubiquity and is potentially associated with breadth of niche in microbial species. Core and average genome sizes are affected strongly by phylogeny and habitat preferences of species. Up to 65% of variance in pangenome features can be explained by habitat and phylogeny combined. At macroevolutionary scale core and mean genome sizes tend to stay more conserved than other features. 44.4% of variance in genome size of 4,582 species is explained with habitat and phylogeny. These results together with previous studies indicate that it will be possible to perform predictive analysis of ecological niches of newly discovered uncultivated bacterial species

3.3. Estimating a species boundary in metagenomic global microbial gene catalog using pangenomes

This subchapter investigates the optimal nucleotide identity threshold for delineation of species boundaries in prokaryotes at gene level, which can be applied to metagenomics data when constructing metagenomic gene catalogs. For this project I performed blastn analysis of genes in pangenomes to find the nucleotide identity range between orthologs within the same species and between different species within the same genus. This work resulted in the co-authorship in the following manuscript:

Luis Pedro Coelho, Renato Alves, Álvaro Rodríguez del Río, Pernille Neve Myers, Carlos P. Cantalapiedra, Joaquín Giner-Lamia, Thomas Sebastian Schmidt, Daniel Mende, Askarbek Orakov, Ivica Letunic, Falk Hildebrand, Thea Van Rossum, Sofia K. Forslund, Supriya Khedkar, **Oleksandr M. Maistrenko**, et al. Towards the biogeography of prokaryotic genes. In preparation.

3.3.1. Introduction

The existence of species in Prokaryotes has been under debate for several decades (Konstantinidis and Tiedje 2005). A species can be defined as a coherent cluster of isolates (strains) with similar phenotype that can recombine and exchange genes (Konstantinidis, Ramette, and Tiedje 2006). In line with this definition, it is predicted that reduction of the ratio between recombination to mutation (r/m) events below 0.25 can enable subpopulations to diverge freely resulting in eventual speciation (González-Torres et al. 2019). Recombination in general tends to reduce diversity between subpopulations and species. For this reason, decline in the recombination rate between species will be likely proportional to a nucleotide identity threshold that can be potentially almost-universally applied for many species. However, these important theoretical advances in our understanding of the species concept in Prokaryotes have limited application for the construction of gene catalogs from metagenomic data. Because majority of previous studies focused on genome-wide thresholds (identity calculated on coding sequences and non-coding/intergenic regions together) (Jain et al. 2018) or only marker genes (Mende et al. 2013). It is important to know the optimal thresholds of the nucleotide identity at gene level to group genes into homologous groups at species level without knowing to which species those genes belong. This is a common situation in metagenomics, however robust data to support such an ANI threshold was missing from the field. Here I address this question using species-specific pangenomes to find operational and natural species boundaries at gene level.

These boundaries can be used in metagenomics, where the species to which a gene belongs is often unknown. I then use this species boundary to estimate the global number of prokaryotic species from a global microbial gene catalog and compare these results to past predictions.

3.3.2. Identification of a natural species boundary based on gene similarity in Prokaryotes

I compared the similarity between genes from the same species to the similarity between genes from different species within the same genus. To do this, I built redundant pangenomes for 107 species and calculated the ANI using blastn between all genes both within pangenomes and between pangenomes of different species. From this data, I observed that genes from the same species tended to have ANI greater than 95% while genes from different species within the same genus tended to have ANI values less than 95%. This separation supports the hypothesis of existence of species with a gene-based ANI boundary around 95% (Figure 3.5.1). In particular, this result is in line with evidence that species exist judging from recombination rates in core genomes (Bobay and Ochman 2017a), pangenome structure and gene sharing (Moldovan and Gelfand 2018), average nucleotide identity (>90% ANI) computed from k-mer similarity approximation for shared genomic islands (Jain et al. 2018) and finally in this study at gene level via identity of blastn matches between genes within species and between species.

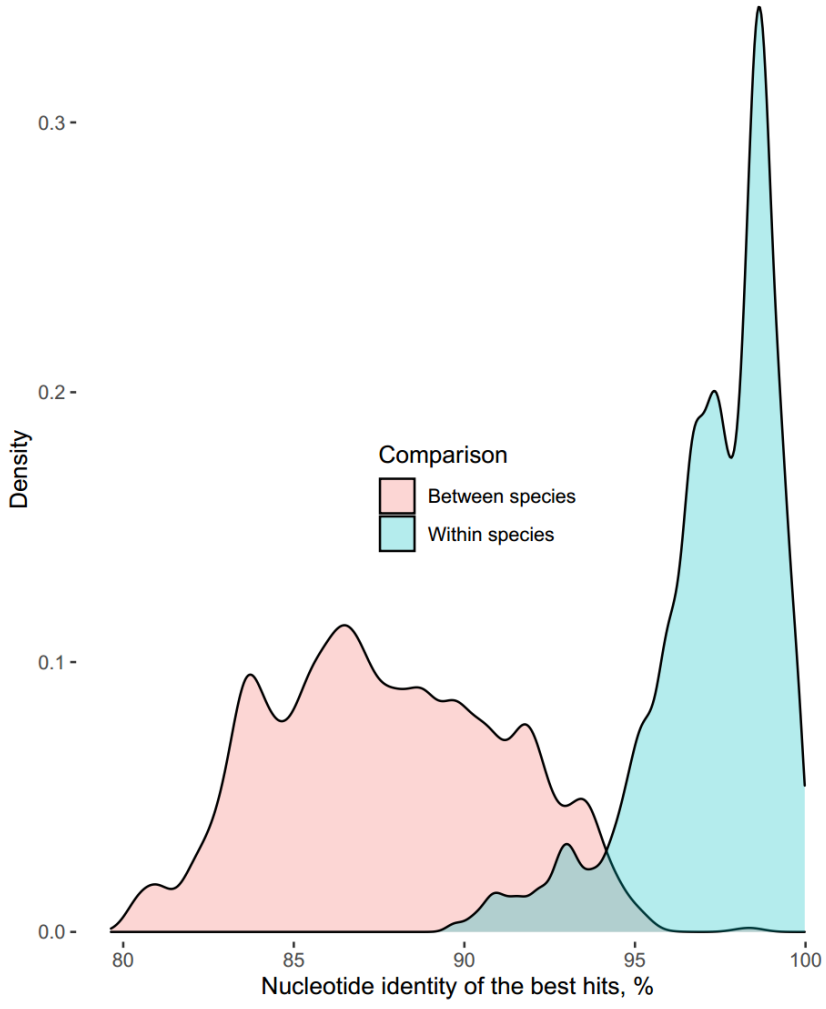


Figure 3.3.1. A 95% nucleotide identity threshold is a proxy for species in Prokaryotes. Shown is nucleotide identity of the closest gene homolog within the same species (blue) or within the same genus (pink).

3.3.3. Discussion

We compiled a global microbial gene catalog from over 13,000 metagenomic samples from all major habitats (marine, freshwater, host-associated, and soil). This resulted in a catalog of 2.3 x 10^9 open reading frames. Applying 95% ANI threshold on this data resulted in 302,655,267 gene clusters (7.4-fold redundancy reduction). These estimates allow us to speculate about the total global number of microbial species and the number that have yet to be isolated and characterized. Assuming that average microbial genome size is 3,600 genes (diCenzo and Finan 2017), then the total number of species in the studied metagenomic samples would be 85,000 (302,655,267 GMGC gene clusters / 3,600 average genome size). This number of species is close to lower bound estimate of the predicted number of microbial species, which is approximately 100,000 (Editorial 2011). In contrast, the number of known species that have been sequenced so far is estimated to be 10,000-15,000 (Parks et al. 2018; Mende et al. 2020). It is important to take into account that robustness in the predictions of the number of species of Prokaryotes is limited by incomplete recovery of genomes from metagenomics samples and bias in favor of highly abundant species; differences in species delineation approaches; imprecise average genome size; and exceptions to the 95% ANI species boundary threshold. Despite these limitations several studies discovered the large number of species-level clusters from metagenomic data (Milanese et al. 2019; Pasolli et al. 2019). With accumulation of metagenomics data, it is possible that we will be able to describe the majority of highly prevalent, highly abundant microbial species using co-assembly and isolation with improved cultivation methods. However, if the species number is indeed close 10^{12} it will be difficult to characterize all of them. Finally, due to the large gene content dissimilarity between strains in the same species (see Subchapter 3.2.) it is unlikely that it will be possible to catalog all genes of every species, thus, many pangenomes will appear “open”.

3.3.4. Conclusion and future outlook

At the moment, there is an overwhelming evidence for the existence of species clusters in Prokaryotes that is supported by different methods. Results from this project together with recent papers published independently from our project support predictions about existence of species boundary in Prokaryotes. In future attempts to catalog and delineate all species in the

biosphere it will be important to continue to question and test species concepts with more high-quality genomes, more species and different methods and algorithms. There is a possibility that gaps in ANI between known species will be filled by newly discovered strains. Future studies in this direction need to address problems of contamination in reference genomes, incorporate bigger diversity of species and incorporate HGT which will result in a comprehensive nearly-universal species model of Prokaryotes.

3.4. Assessing within-species diversity and delineating subspecies from metagenomic data

This subchapter explores challenges and possible solutions when searching for meaningful groups within species in metagenomes, and how analysis of gene content variation can result in a powerful tool to further advance this investigation. For this project, I performed an analysis to quantify the association between core genome nucleotide identity and gene content similarity between genomes of the same species. I also wrote some of the discussion on how genetic variation arises and what are challenges and advances in the delineation of subspecies. The results and discussion in part have been published in the following manuscript:

Van Rossum T, Ferretti P, **Maistrenko OM**, Bork P (2020) Diversity within species: interpreting strains in microbiomes. Nature Reviews Microbiology.

3.4.1. Introduction

Systematic identification of within-species variation is a relatively well-established field and generally relies on a set of standard genomic analysis tools (Bush et al. 2020). However, classification of within-species variation is challenging due to issues in species classification that are inhered in within-species analysis, as well as conceptual limitations and terminological confusion and redundancy.

The growing attention towards the analysis of within-species diversity in the metagenomic field led to an increasing confusion in the terminology. In particular, there is an issue of using of different definitions for what is considered a strain and a subspecies in traditional microbiology and metagenomics. In traditional culture-based microbiology, “strain” is defined as the “descendants of a single isolation in pure culture and usually is made up of a succession of cultures ultimately derived from an initial single colony” (Brenner, Staley, and Krieg 2015). Metagenomics often does not involve isolation of pure cultures of species and strains and consequently is limited in estimating phenotypes. This could potentially lead to naming every variant as a new strain, without knowing if those variants differ significantly in their phenotype. It is recommended that the null hypothesis about any variant is that it is neutral for phenotype (Rocha 2018). Such an assumption raises the question of whether all variants and which variants should be considered when naming new strains and species. These aspects represent considerable challenges when trying to apply terms from classic microbiology, such as strain and subspecies, in the metagenomic context because of absence of experimental validation and appropriate metadata. In this subchapter I discuss (1) what is the magnitude of

within-species diversity; (2) what are challenges in delineating within-species variants, and (3) why gene content is gaining relevance in the attempt to delineate meaningful within-species groups. I also discuss how the development and reapplication of the theoretical framework of species delineation will help advancing within-species analysis in metagenomics.

3.4.2. Magnitude of within-species diversity

It is established that more closely related strains share more genes compared to more distantly related ones within the same species (Sheppard et al. 2013; Dillon et al. 2019). In other words, there are species with tight cohesion of core genome SNP diversity and accessory gene content. Gene and SNP contents diversify together gradually and in a linked manner. However, some species have highly flexible secondary chromosomes (e.g. Vibrio spp.) (Heidelberg et al. 2000) and in some species nearly half of the genome is represented by plasmids, for example, in Borrelia (Brisson et al. 2012). Such properties of genomes can promote breakdown of the association between gene content and SNP similarity of strains of the same species. To date, there has been no explicit estimates of genomic dissimilarity of pangenomes using the same method across many species (using only high quality genomes, and species classified with the same method). To quantify this relationship across many species systematically, I gathered reference genomes for 155 species (at least 10 genomes per species) and calculated the core genome ANI and gene content dissimilarity among genomes for each species. In Figure 3.4.1, I show that the trend of association of gene content and core genome identity across species indeed exists with relatively high support based on Spearman correlation (Spearman Rho = 0.57, $p < 2.2e-16$). In the figure it is also visible that the magnitude of variation in the gene content (the highest density of points is in the range of 60%-100% Jaccard index of gene content overlap) is bigger than in nucleotide identity (the highest density of points is in the range 97%-100% of nucleotide identity). In part due to this difference in the distribution of values, a considerable number of strains are similar in their core genome but dissimilar in the gene content. Such large difference together with the observation that gene content diversifies earlier than SNP content might indicate in favor of the idea that gene content is more important for ongoing adaptions and subspeciation across many species (Andreani, Hesse, and Vos 2017).

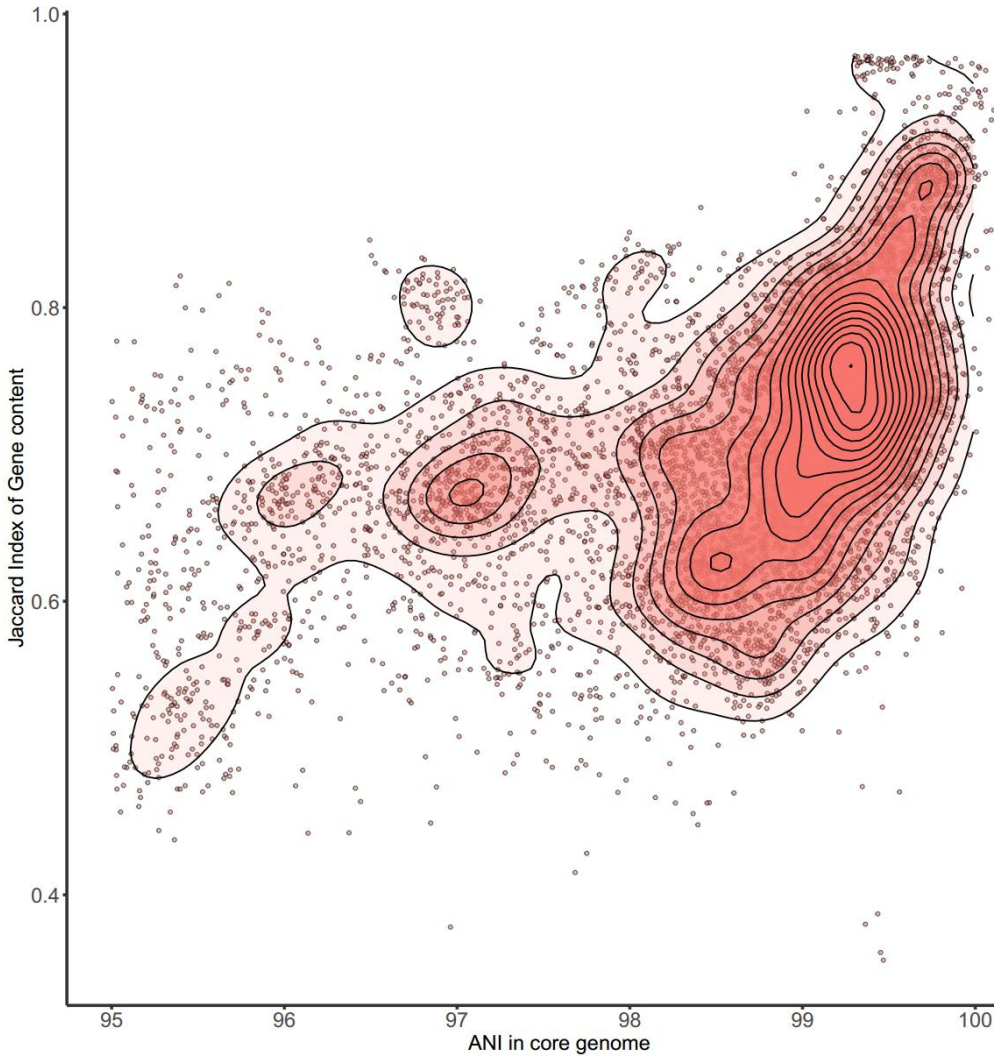


Figure 3.4.1. Scatterplot of average gene content overlap and average core genome identity between strains. Each point is a pairwise comparison of one isolate genome versus all other conspecific isolate genomes. The data is from 155 bacterial species, each with at least 10 sequenced isolate genomes. Opacity of red-coloured topographical overlay indicates density of points. The plot shows the relationship between the similarity of the core genome, measured by average nucleotide identity (ANI), versus the similarity of gene content, measured by Jaccard Index. Genomes with higher similarity between their core gene sequences tend to have more genes in common (Spearman correlation R=0.57, $p < 2.2e-16$). However, high ANI does not necessarily imply highly similar gene content, with many genomes with over 99% core genome ANI having less than 70% of genes in common. Most within-species ANI values are greater than 97%, the few data points below 95% ANI are not shown (83% and 4% of data points, respectively). (Data for this figure originates from Subchapter 3.2. published in (Maistrenko et al. 2020), figure and its caption was adapted from "Van Rossum, T., Ferretti, P., Maistrenko, O.M., Bork, P. Diversity within species: interpreting strains in microbiomes. Nat Rev Microbiol

(2020). https://doi.org/10.1038/s41579-020-0368-1, Springer Nature, 2020 Springer Nature Limited").

3.4.3. Discussion

Genomic variation emerges through constant occurrence of mutations due to errors in replication, repair and recombination, resulting in diversification of species. This variation is shaped and distributed in population(s) by genetic drift, selection and dispersal/HGT/gene flow into new genomes and environments(O'Donnell, Langston, and Stillman 2013; Radman, Taddei, and Matic 2000). All these forces combined together result in high genetic diversity and lead to pangenomes. In the broad and more abstract sense, a pangenome is defined as a set of all possible variants: gene content variation and other structural variants, indels, SNPs in a sample of genomes or in any taxonomic group (Marschall et al. 2016). A combination of evolutionary factors can result in emergence of distinctive subgroups within a species. While species are the results of long-term evolution and specialization, within-species variations are often novel mutations that are neutral and/or indistinguishable in phenotype and might be quickly removed from the population. This brings up the first conceptual problem: where to set the boundaries of how much variation to consider as relevant in within-species analysis. Traditional multi-locus sequence typing (MLST) based on few loci might result in lack of resolution. At the same time, multi-locus sequence typing based on too many loci (up to whole genome multilocus sequence typing (wgMLST)) can result in reclassification of existing types, adding more confusion (Pearce et al. 2018; Achtman et al. 2012). Metagenomics can potentially introduce even more confusion by introducing additional sources of multi-strain diversity from the same sample. This is because metagenomics theoretically enables recovery of several strains of the same species simultaneously. Consequently, even more diversity can be characterized with metagenomics.

The second conceptual problem is which type of variation is more important for delineating within-species groups: whether to prefer single nucleotide polymorphisms over gene content or vice versa. It is likely that both variation types are important, but even very high (>99%) nucleotide similarity can correspond to very low (70%) gene content overlap (as shown in the Figure 3.4.1). Without phenotype data it is possible to speculate whether on average across many species if gene content variations will be more important for pathogenesis and ecological niche functions than single nucleotide polymorphisms. Such speculation is backed-up by the fact that accessory genes can be transferred as part of genomic islands and

operons introducing new ready-to-use functional pathways into recipient strains. Various studies support this hypothesis, arguing that gene content change is a primary evolutionary response (Dutilh et al. 2014; Andreani, Hesse, and Vos 2017).

The third conceptual and technical problem is that some of the variation is not necessarily due to true variants, because they might be polymerase chain reaction errors during library preparation and sequencing. Presence and absence of genes has to be cautiously evaluated when dealing with assembled genomes. For example, metagenomics-assembled genomes are likely to have low completeness and high contamination levels with fragments from other species or the same species but different strains (Bowers et al. 2017; Shaiber and Eren 2019). These problems can potentially lead to overestimations of the gene content diversity. To mitigate and address these problems various pipelines set a nucleotide identity threshold above which variation is considered to be biologically meaningful. Qualities of the metagenomic assemblies are evaluated to exclude additional genes originating from other species.

The fourth problem is the lack of knowledge about natural universal boundaries for subspecies and other intraspecific groups for a majority of species (Brenner, Staley, and Krieg 2015). In principle, delineation of subspecies could rely on conceptually similar approaches as species delineation, such as detectable decline in frequency of HGT, recombination and ANI between groups of strains. However, due to high recombination rates, actual phylogeny might be difficult to reconstruct, which limits the potential to search for meaningful clusters within species (Shapiro et al. 2012).

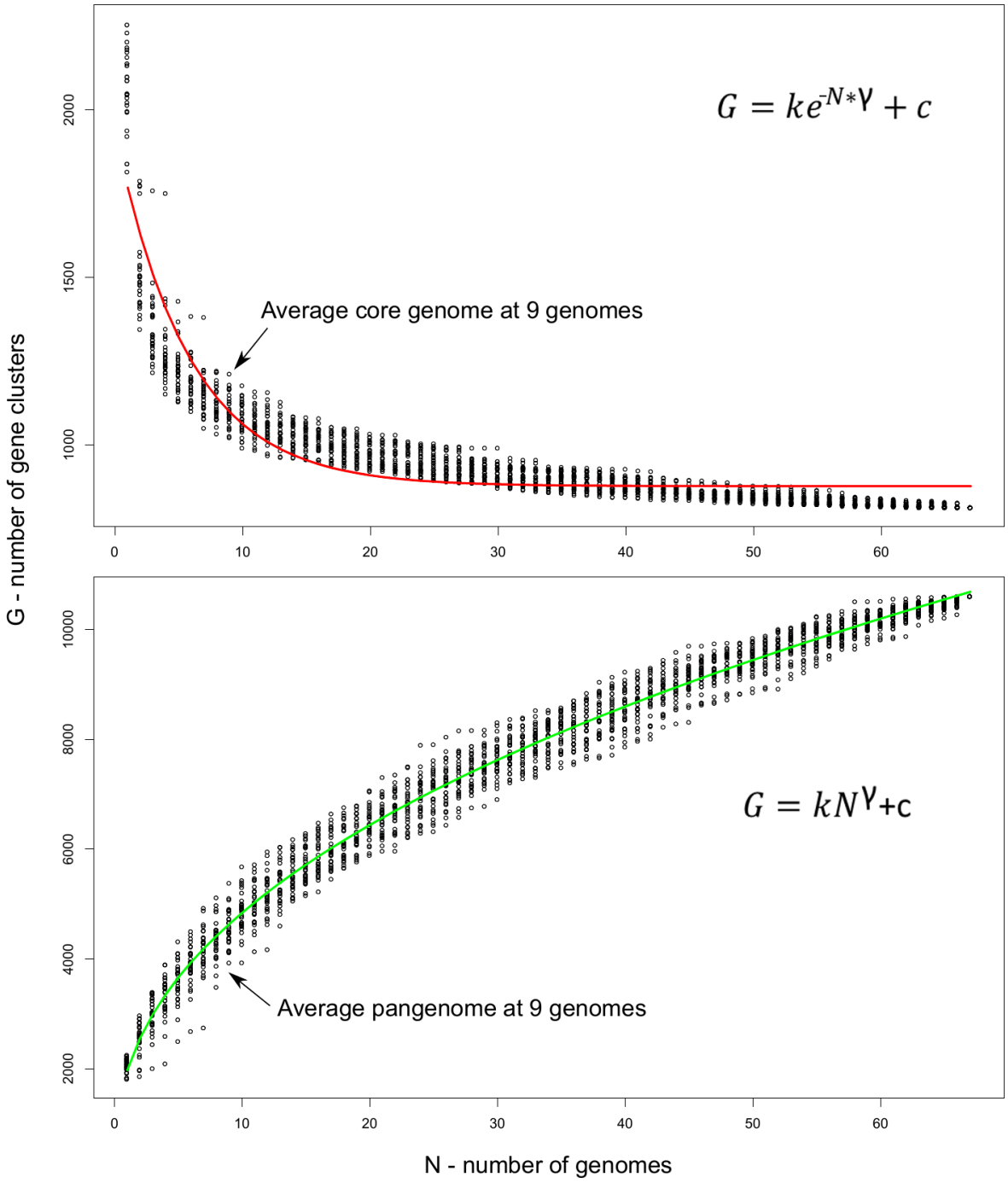
The practical solution would be to refer to new genomes (isolate-based sequencing or metagenomics assembled genomes) in a context-dependent way in each project and use the closest terms from classical microbiology. In other words, naming of within-species variants is rather a convenience than an evolutionary meaningful delineation.

Additional progress might be achieved by incorporating the concepts of genotype space and pangenomes (Medini et al. 2005; Tettelin et al. 2005) into metagenomics (metapangenomics) (Delmont and Eren 2018). This integration will enable viewing of all genomic variation as a nearly-continuous, nearly-infinite genotype space. In such a scenario, every genotype is characterized by the set of coordinates in multidimensional space of all possible genotypes. There are several efforts moving in this direction, such as the attempt to upgrade the human genome reference by including population diversity as a pangenome graph (Llamas et al. 2019). However, with the continuous increase in sequencing data this approach might become not computationally feasible to apply for all microbial species, including isolate-based and metagenomics-based whole genome sequencing data.

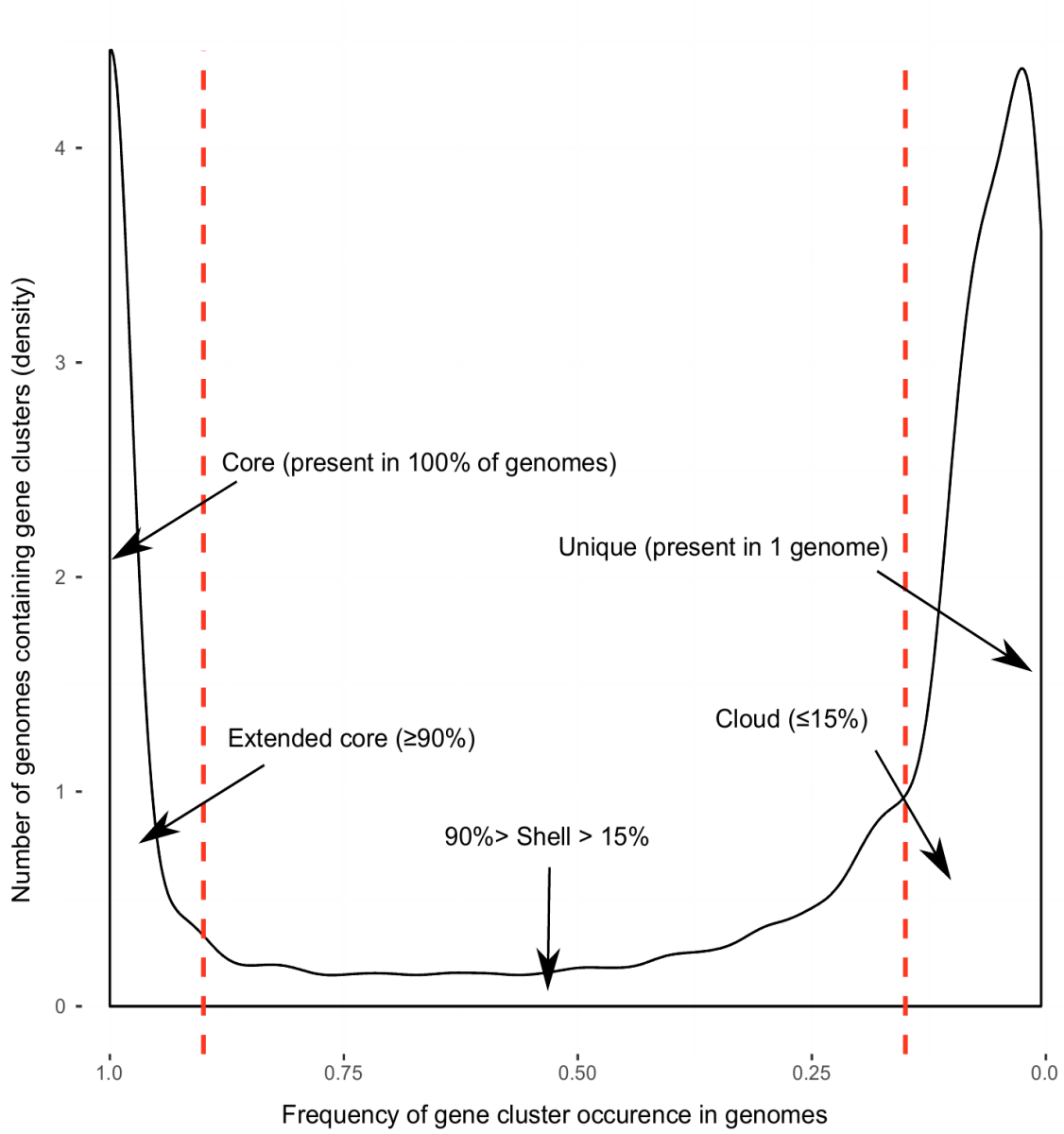
3.4.4. Conclusion and future outlook

Gene content diversity is becoming a primary focus of species and within-species level analysis in genomics and metagenomics. Metapangenomics analysis, with careful application of known terminology from classic microbiology, will ensure a coherent development of the field. Investigating within-species diversity systematically across species (Figure 3.4.1) provides a foundation for building a better understanding on how to delineate subspecies and within-species groups. In future studies, it is important to continue to search for biologically meaningful within-species groups that might highlight ongoing speciation. In particular, combining metadata about species' and strains' environmental preferences, antibiotic resistance and virulence can provide insights into what drives the within-species groups' diversification in the settings of metagenomics and microbiome genome-wide association studies.

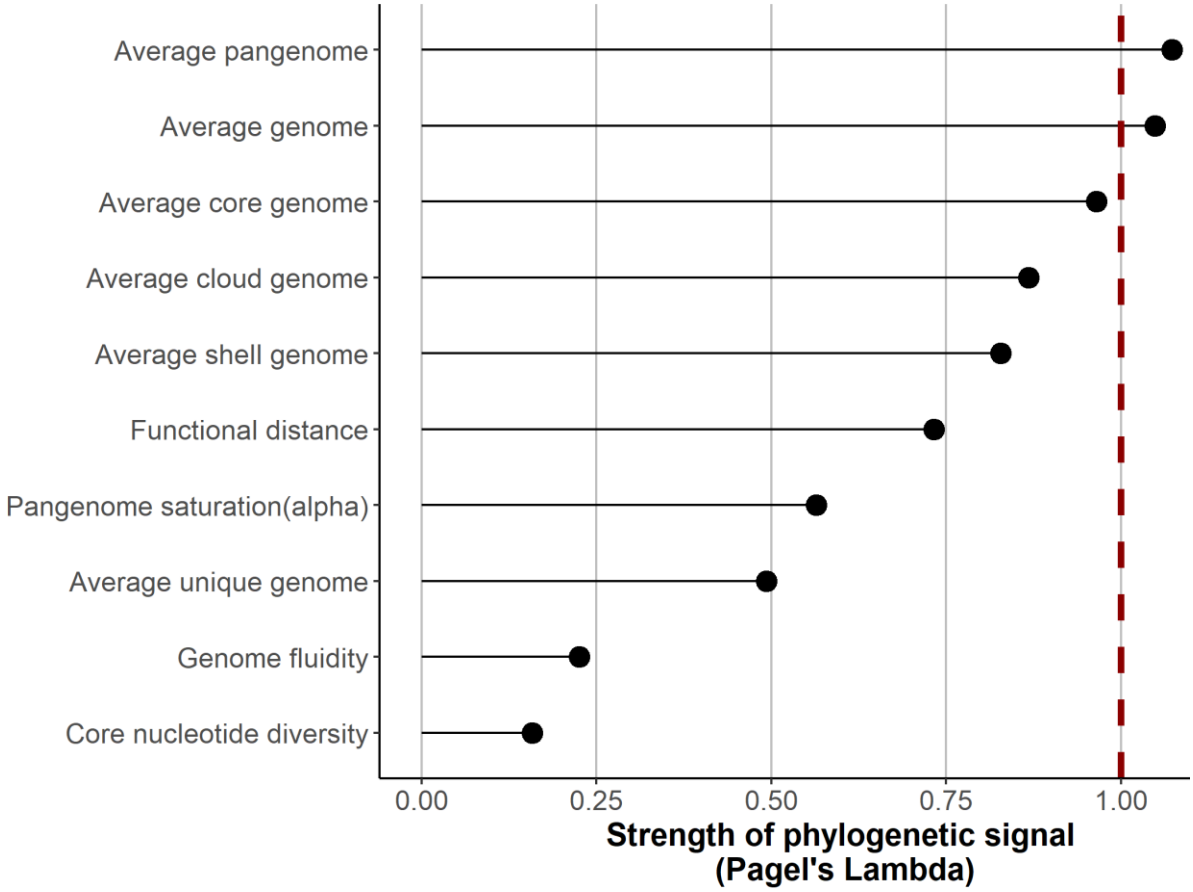
Appendix



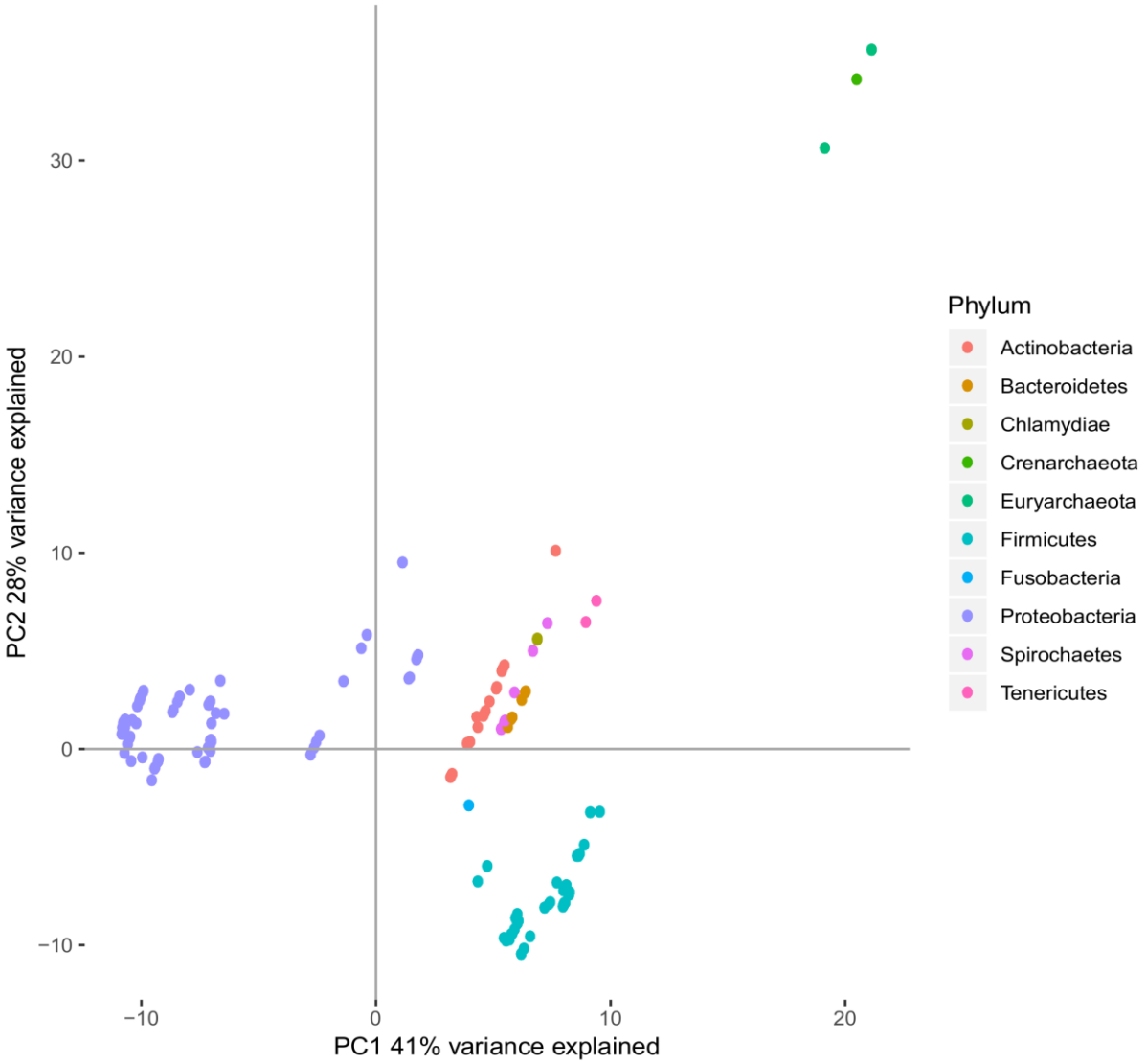
Supplementary Figure 1. Examples of saturation curves. Core- and pangenome size at a given number of (randomly chosen) genomes from a species. Fits to displayed formulas shown in red (core genome) and green (pangenome). (Figure and description to the figure are published in (Maistrenko et al. 2020) in the ISME journal, 2020 Springer Nature Limited).



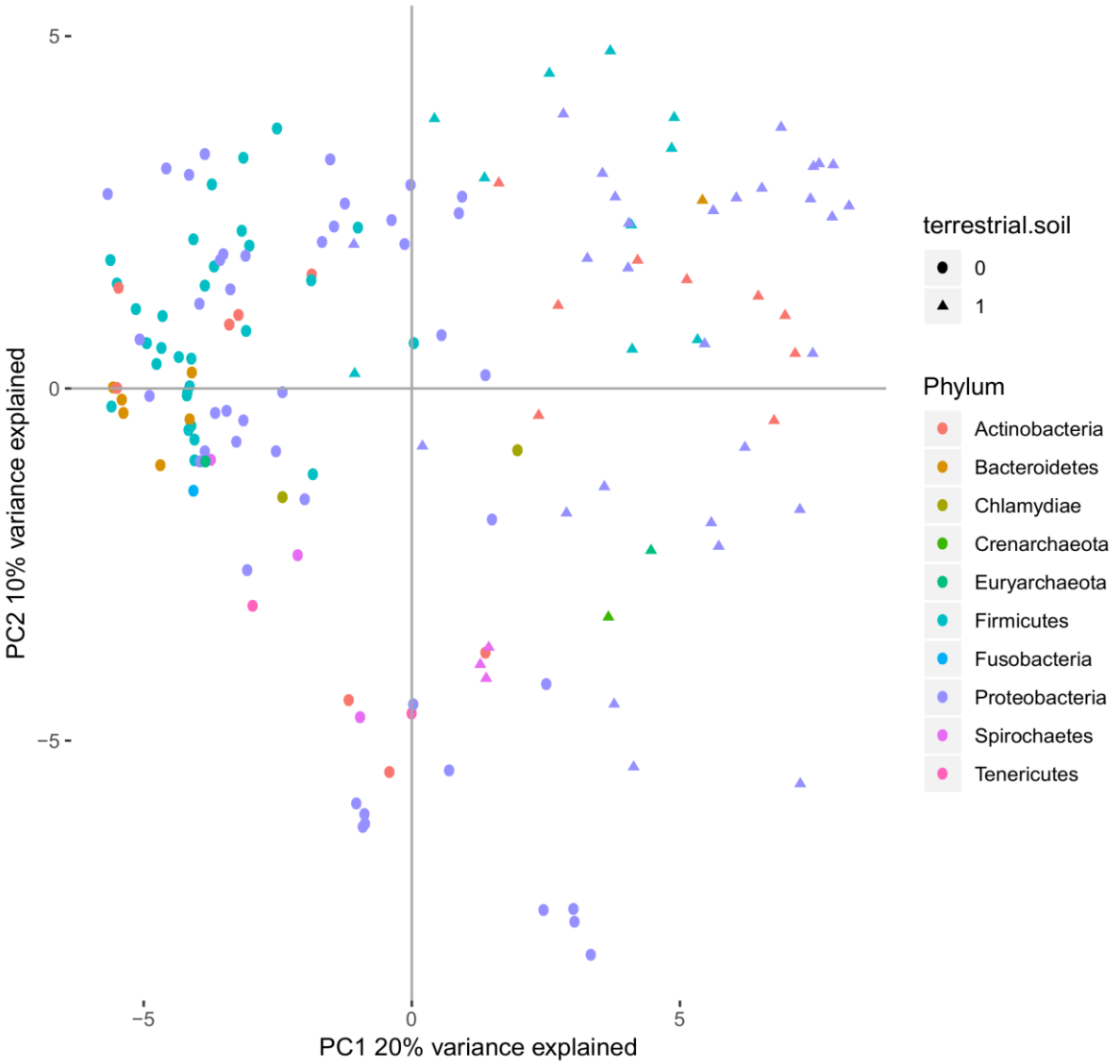
Supplementary Figure 2. Thresholds for pangenome components. Gene frequency distribution displayed in black. Dash lines represent boundaries between extended core and shell genome; shell and cloud genome. (Figure and description to the figure are published in (Maistrenko et al. 2020) in the ISME journal, 2020 Springer Nature Limited).



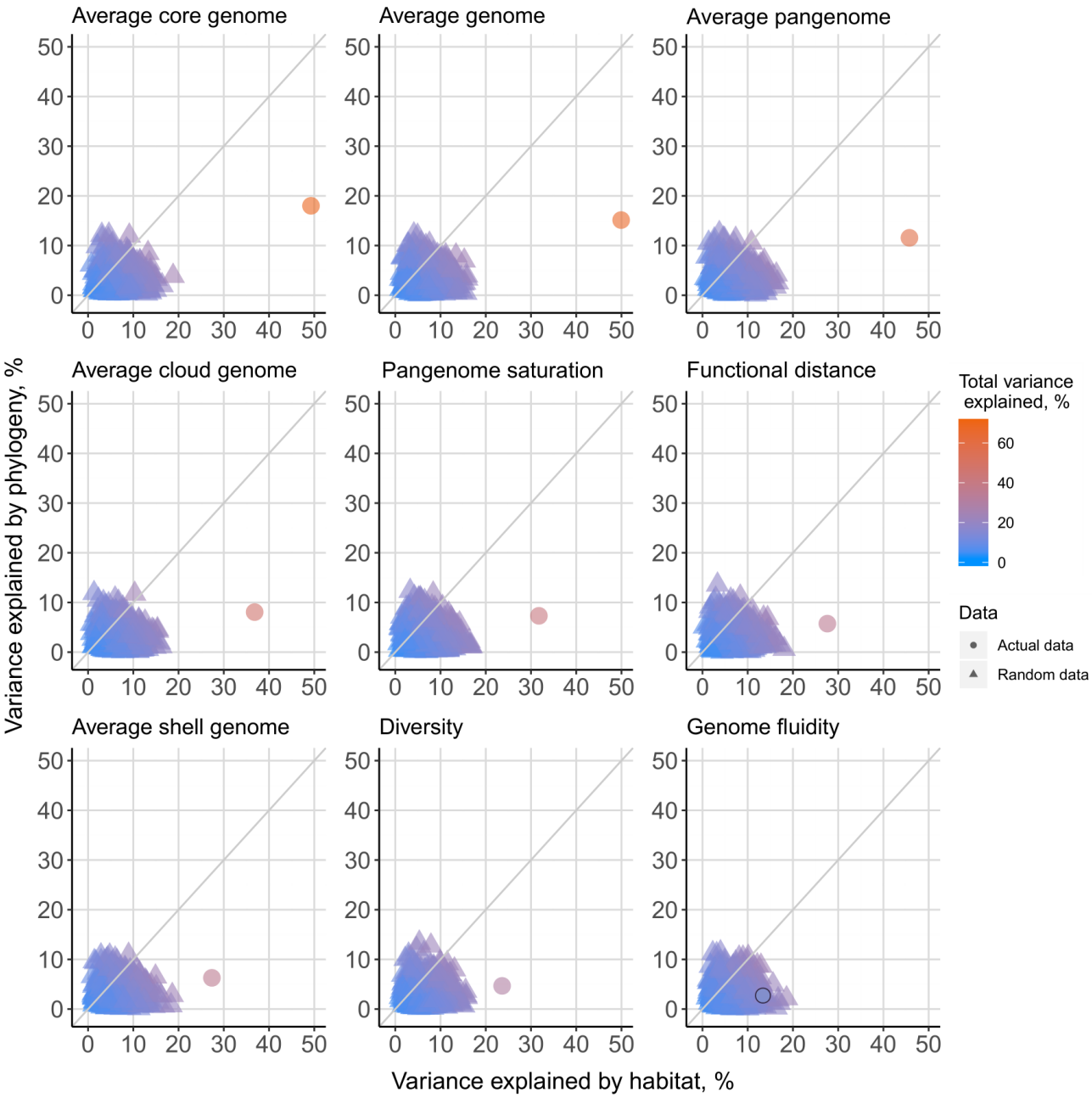
Supplementary Figure 3. Phylogenetic signal of 10 genomic characteristics across 155 species of Prokaryotes. When Pagel's λ approximates to 1 – trait manifests phylogenetic signal (marked with dash line). (Figure and description to the figure are published in (Maistrenko et al. 2020) in the ISME journal, 2020 Springer Nature Limited).



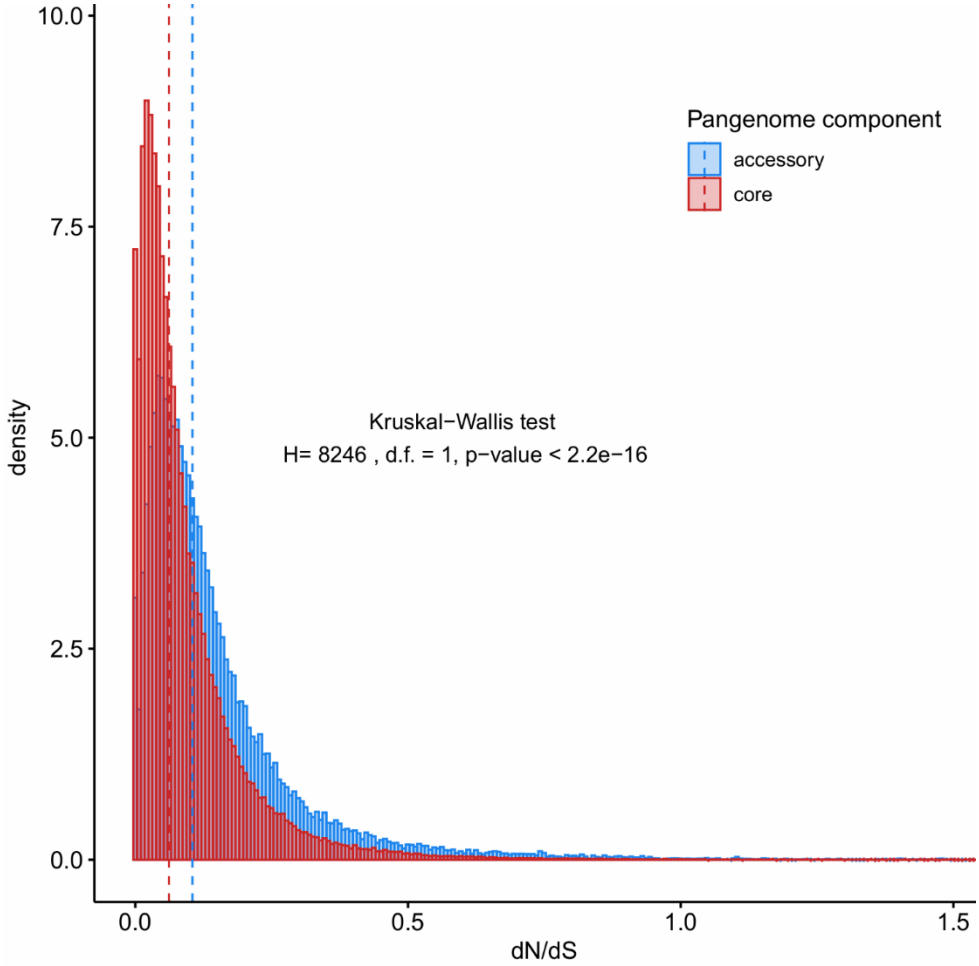
Supplementary Figure 4. Biplot of PCA (PC1 and PC2) using cophenetic distances observed in the phylogenetic tree reconstructed from 155 species used in this study. (Figure and description to the figure are published in (Maistrenko et al. 2020) in the ISME journal, 2020 Springer Nature Limited).



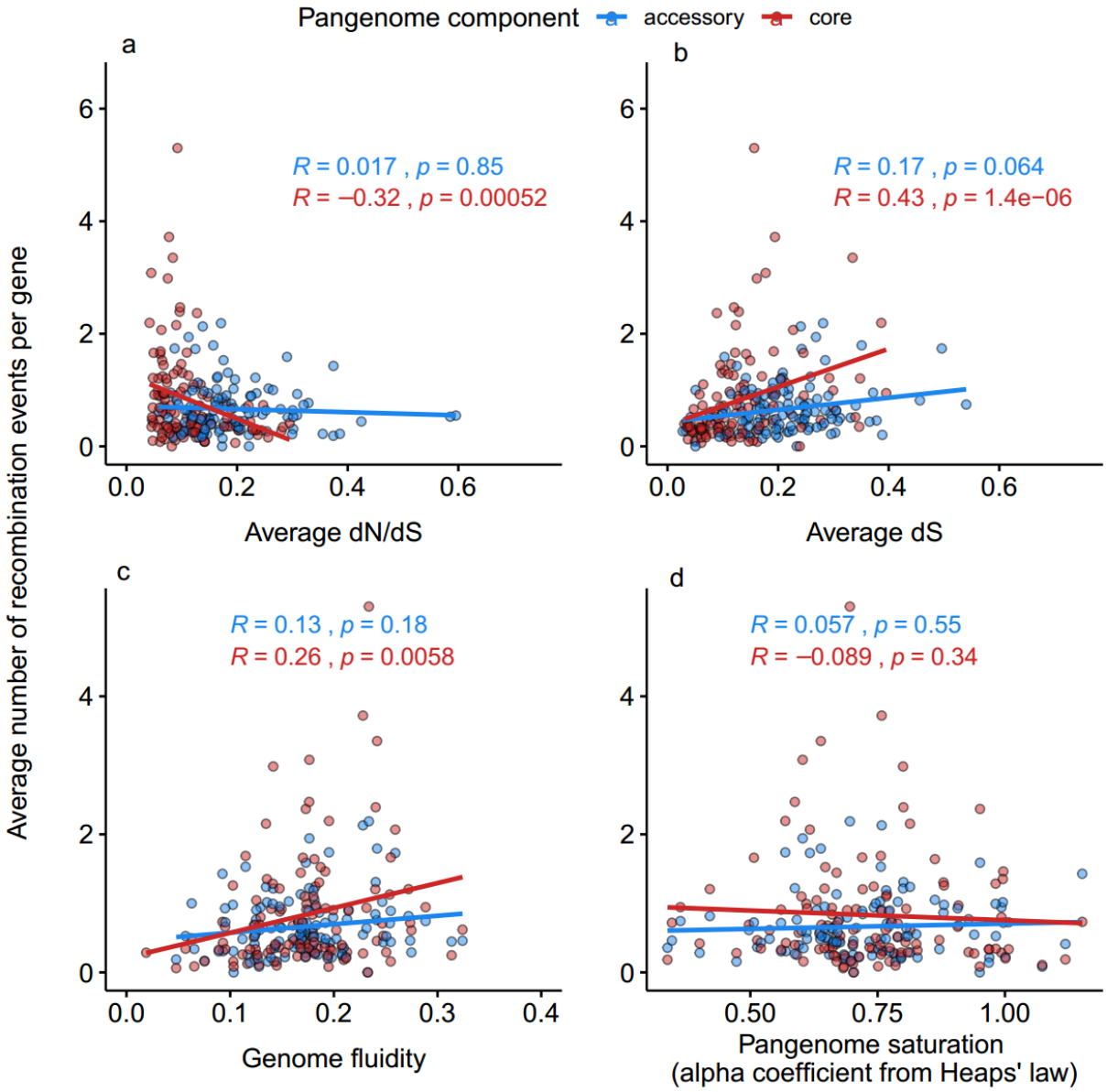
Supplementary Figure 5. Biplot of the first 2 principal components of the decomposition of the habitat-association matrix (0 - not associated with soil habitat, 1 - associated with soil habitat). (Figure and description to the figure are published in (Maistrenko et al. 2020) in the ISME journal, 2020 Springer Nature Limited).



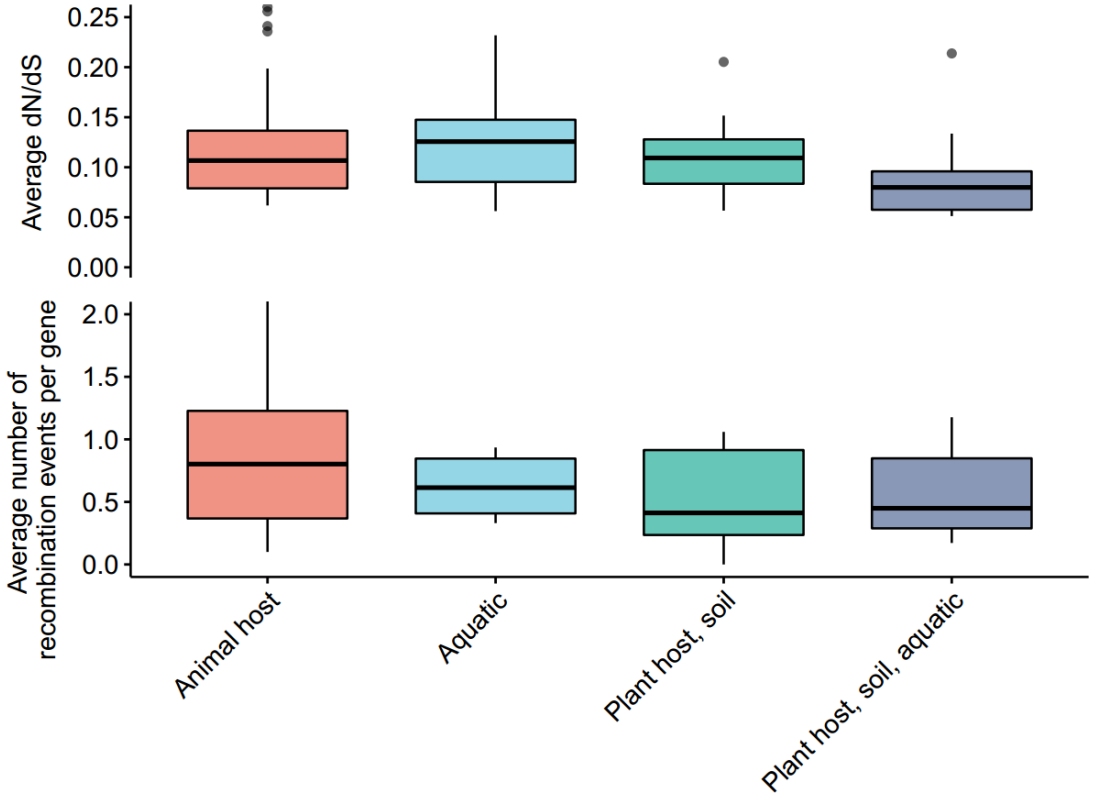
Supplementary Figure 6. Randomized phylogenetic and habitat PCs explain a smaller fraction of the variance than actual data for the core genome size and the average genome. (Figure and description to the figure are published in (Maistrenko et al. 2020) in the ISME journal, 2020 Springer Nature Limited).



Supplementary Figure 7. Accessory genome is under relaxed purifying selection compared to core genome (unpublished data).



Supplementary Figure 8. Recombination rates. Association of recombination events in core and accessory components of pangenome with (a) average ratio of non-synonymous to synonymous substitutions – dN/dS, (b) average fraction of synonymous substitutions – dS, (c) genome fluidity, and (d) pangenome saturation. Association between variables was tested with Spearman rank correlation ("Rho" and their p-values are presented on figures for component of pangenome) (unpublished data).



Supplementary Figure 9. Association of recombination rate and dN/dS with habitat preferences. Global test shows difference for dN/dS by habitat Kruskal-Wallis chi-squared = 8.3923, df = 3, p-value = 0.03856, post-hoc pairwise test does not show difference. Recombination rates are not affected by habitat preferences in this dataset. Kruskal-Wallis chi-squared = 5.3574, df = 3, p-value = 0.1474 (unpublished data).

Supplementary Table 1

(Supplementary tables and descriptions are published in (Maistrenko et al. 2020) in the ISME journal, 2020 Springer Nature Limited).

Supplementary Table 1. Definitions of pangenome features and other terms used in the manuscript; key-words used for habitat annotation from PATRIC database.	
Term	Definition
Number of genomes (sample size)	Number of genomes (also referred as sample size) that were used to calculate pangenome features for each species
Pangenome saturation (alpha)	Absolute value of alpha in equation [1] in methods
Pangenome saturation (gamma)	Coefficient gamma from equation [2] in methods
Core genome saturation	Coefficient gamma from equation [3] in methods
Pangenome	Total number of protein coding gene clusters estimated present in any number of genomes for each species (all genomes of a given species used to calculate this)
Core	Number of protein coding gene clusters found in every genome in a given species (all genomes of a given species used to calculate this)
Extended core	Total number of protein coding gene clusters in the extended core genome of a given species (all genomes of a given species used to calculate this)
Shell	Total number of protein coding gene clusters in the shell genome fo a given species (all genomes of a given species used to calculate this)
Cloud	Total number of protein coding gene clusters in the cloud genome for a given species (all genomes of a given species used to calculate this)
Unique	Total number of unique protein coding gene clusters found in a given species (all genomes of a given species used to calculate this)
Average genome	Mean number of protein coding gene clusters averaged across all genomes of each species

Average extended core	Mean number of protein coding gene clusters in the extended core averaged across all genomes of each species
Average shell	Mean number of protein coding gene clusters in the shell genome averaged across all genomes of each species
Average cloud	Mean number of protein coding gene clusters in the cloud genome averaged across all genomes of each species
Average unique	Mean number of unique protein coding gene clusters averaged across all genomes of each species
Average core genome	Mean number of protein coding gene clusters in the core genome averaged across 30 random combination of 9 genomes (for each species)
Average pangenome	Mean number of protein coding gene clusters in the core genome averaged across 30 random combination of 9 genomes (for each species)
Pangenome size (Chao lower bound estimate)	Chao lower bound estimate of the possible number of genes in a given species' pangenome (calculated using the corresponding function of the "micropan" R-package)
Diversity	Mean nucleotide divergence (1 - nucleotide identity) of the core genome calculated across all pairs of genomes within a given species
Average Jaccard distance	Average Jaccard distance (1 - Jaccard index) calculated across all pairs of genomes within a given species
Genome fluidity	Genomic fluidity is calculated as the ratio of unique gene families over the sum of gene families averaged over randomly chosen pairs of genomes from within a given species of *N* genomes
Functional distance	COG-based average Jaccard distance between isolates/strains within a given species
Ubiquity of species	Ubiquity is the sum of all positive associations (Benjamini-Hochberg-corrected Fisher's Exact Tests, $p \leq 0.05$) of a species with habitats in the Microbial Atlas Project dataset. In other words, ubiquity shows with how many habitats a certain species was associated.
Key-words for habitat annotation "soil"	["terrestrial biome", "Terrestrial", "Soil", "soil", "Rhizosphere", "rhizosphere", "plant root", "Plant root", "root nodule", "Root nodule", "root nodules", "rhizosphere soils", "Tailings", "Rhizospheric", "rhizospheric", "Sand", "soil sample", "sediment",

	"Sediment", "Sludge", "sludge", "mud", "Mud", "Sand"]
Key-words for habitat annotation "aquatic"	["Aquatic", "aquatic", "marine", "Marine", "water", "Water", "Sea water", "Fresh water", "sea water", "fresh water", "Pond", "pond", "river", "River", "lake", "Lake", "Ocean", "ocean", "creek", "Creek", "waterfall", "Waterfall", "Hot spring", "Hot springs", "hot springs", "Hot springs", "hot spring", "oceanic", "Oceanic", "sea ", "sea-", "Wastewater", "wastewater", "Rice paddies"]
Key-words for habitat annotation "host-associated"	["HostAssociated", "Host-associated", "Host Associated", "host", "Host", "Rhizosphere", "rhizosphere", "Plants", "plant root", "Plant root", "root nodule", "Root nodule", "root nodules", "rhizosphere soils", "Rhizospheric", "rhizospheric", "Symbiotic", " skin", "Zoonotic", "rumen", "livestock-assoicated habitat", "Feces", "nasopharynx", "Blood", "sputum", "blood", "patient", "CSF", "stool", "feces", "Bodily fluid", "nares", "BLOOD", "Respiratory system"]
Key-words for habitat annotation "food"	["food", "Food", "FOOD", "milk", "Milk", "cheese", "Cheese", "fermented", "meat", "Meat", "frozen peas", "dairy product", "dairy products", "Dairy", "koumiss", "burger", "yogurt", "Burger", "Yogurt", "wine", "Wine", "Champagne", "Cider", "champagne", "cider", "bread ", "Bread ", "liver paste", "liquor", "sashimi", "beef", "Beef", "Seafood", "seafood", "silage", "ground turkey", "chicken breast", "sourdough", "fermented soybean", "fermented", "fermentation", "Fermented", "Fermentation"]

Supplementary tables 2-8 are provided in online version of thesis. Tables can be accessed also at the ISME journal were the manuscript is published ("Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity", https://doi.org/10.1038/s41396-020-0600-z).

Supplementary Table 2. List of isolates of 155 species used in the analysis
Supplementary Table 3. Taxonomy, pangenome features, habitat metadata, principal components used for variance quantification
Supplementary Table 4. Pairwise correlations between pangenome features and habitat preferences of 155 species.
Supplementary Table 5. (or Supplementary Table 6 in the manuscript). Phylogenetic generalized least squares (pangenome features ~ ubiquity fit (y~x)).
Supplementary Table 6. (or Supplementary Table 5 in the manuscript). Quantities of variance explained by habitat preferences and phylogenetic inertia.

Supplementary Table 7. Average genome size and habitat preferences of species in proGenomes1 database.
Supplementary Table 8. Average genome size and habitat preferences of species in proGenomes2 database.

References

Achtman, Mark, John Wain, François Xavier Weill, Satheesh Nair, Zhemin Zhou, Vartul Sangal, Mary G. Krauland, et al. 2012. "Multilocus Sequence Typing as a Replacement for Serotyping in Salmonella Enterica." *PLoS Pathogens* 8 (6). https://doi.org/10.1371/journal.ppat.1002776.

Amid, Clara, Blaise T F Alako, Balavenkataraman Kadhirvelu, Tony Burdett, Josephine Burgin, Jun Fan, Peter W Harrison, et al. 2020. "The European Nucleotide Archive in 2019." *Nucleic Acids Research* 48. https://doi.org/10.1093/nar/gkz1063.

Andreani, Nadia Andrea, Elze Hesse, and Michiel Vos. 2017. "Prokaryote Genome Fluidity Is Dependent on Effective Population Size." *ISME Journal* 11 (7): 1719–21. https://doi.org/10.1038/ismej.2017.36.

Baltrus, David A., Kevin Dougherty, Kayla R. Arendt, Marcel Huntemann, Alicia Clum, Manoj Pillay, Krishnaveni Palaniappan, et al. 2017. "Absence of Genome Reduction in Diverse, Facultative Endohyphal Bacteria." *Microbial Genomics* 3 (2). https://doi.org/10.1099/mgen.0.000101.

Barberán, Albert, Hildamarie Caceres Velazquez, Stuart Jones, and Noah Fierer. 2017. "Hiding in Plain Sight: Mining Bacterial Species Records for Phenotypic Trait Information." *MSphere* 2 (4): e00237-17. https://doi.org/10.1128/mSphere.00237-17.

Barberán, Albert, Kelly S. Ramirez, Jonathan W. Leff, Mark A. Bradford, Diana H. Wall, and Noah Fierer. 2014. "Why Are Some Microbes More Ubiquitous than Others? Predicting the Habitat Breadth of Soil Bacteria." Edited by John Klironomos. *Ecology Letters* 17 (7): 794–802. https://doi.org/10.1111/ele.12282.

Bary, A. De. 1879. *Die Erscheinung Der Symbiose: Vortrag Gehalten Auf Der Versammlung Deutscher*.

Batty, Elizabeth M., Suwittra Chaemchuen, Stuart Blacksell, Allen L. Richards, Daniel Paris, Rory Bowden, Caroline Chan, et al. 2018. "Long-Read Whole Genome Sequencing and Comparative Analysis of Six Strains of the Human Pathogen Orientia Tsutsugamushi." Edited by José Reck. *PLOS Neglected Tropical Diseases* 12 (6): e0006566. https://doi.org/10.1371/journal.pntd.0006566.

Batut, Bérénice, Carole Knibbe, Gabriel Marais, and Vincent Daubin. 2014. "Reductive Genome Evolution at Both Ends of the Bacterial Population Size Spectrum." *Nature Reviews Microbiology* 12 (12): 841–50. https://doi.org/10.1038/nrmicro3331.

Baty, Florent, Christian Ritz, Sandrine Charles, Martin Brutsche, Jean Pierre Flandrois, and Marie Laure Delignette-Muller. 2015. "A Toolbox for Nonlinear Regression in R: The Package Nlstools." *Journal of Statistical Software* 66 (5): 1–21. https://doi.org/10.18637/jss.v066.i05.

Baumdicker, F., W. R. Hess, and P. Pfaffelhuber. 2012. "The Infinitely Many Genes Model for the Distributed Genome of Bacteria." *Genome Biology and Evolution* 4 (4): 443–56. https://doi.org/10.1093/gbe/evs016.

Bennett, Gordon M., and Nancy A. Moran. 2013. "Small, Smaller, Smallest: The Origins and Evolution of Ancient Dual Symbioses in a Phloem-Feeding Insect." *Genome Biology and Evolution* 5 (9): 1675–88. https://doi.org/10.1093/gbe/evt118.

Bentley, Stephen D., and Julian Parkhill. 2004. "Comparative Genomic Structure of Prokaryotes." *Annual Review of Genetics* 38 (1): 771–91. https://doi.org/10.1146/annurev.genet.38.072902.094318.

Biller, Steven J., Paul M. Berube, Debbie Lindell, and Sallie W. Chisholm. 2014. "Prochlorococcus: The Structure and Function of Collective Diversity." *Nature Reviews Microbiology* 13 (1): 13–27. https://doi.org/10.1038/nrmicro3378.

Bobay, Louis-Marie, and Howard Ochman. 2017a. "Biological Species Are Universal across Life's Domains." *Genome Biology and Evolution* 9 (3): 491. https://doi.org/10.1093/gbe/evx026.

———. 2017b. "The Evolution of Bacterial Genome Architecture." *Frontiers in Genetics* 8 (May): 72. https://doi.org/10.3389/fgene.2017.00072.

———. 2018. "Factors Driving Effective Population Size and Pan-Genome Evolution in Bacteria." *BMC Evolutionary Biology* 18 (1): 153. https://doi.org/10.1186/s12862-018-1272-4.

Bohlin, Jon, Vegard Eldholm, John H. O. Pettersson, Ola Brynildsrud, and Lars Snipen. 2017. "The Nucleotide Composition of Microbial Genomes Indicates Differential Patterns of Selection on

Core and Accessory Genomes." *BMC Genomics* 18 (1): 151. https://doi.org/10.1186/s12864-017-3543-7.

Bolotin, Evgeni, Ruth Hershberg, F. Delsuc, E. J. Douzery, and E. V. Koonin. 2016. "Bacterial Intra-Species Gene Loss Occurs in a Largely Clocklike Manner Mostly within a Pool of Less Conserved and Constrained Genes." *Scientific Reports* 6 (1): 35168. https://doi.org/10.1038/srep35168.

Borcard, Daniel, Francois Gillet, and Pierre Legendre. 2011. *Numerical Ecology with R*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-7976-6.

Boscaro, Vittorio, Martin Kolisko, Michele Felletti, Claudia Vannini, Denis H. Lynn, and Patrick J. Keeling. 2017. "Parallel Genome Reduction in Symbionts Descended from Closely Related Free-Living Bacteria." *Nature Ecology & Evolution* 1 (8): 1160–67. https://doi.org/10.1038/s41559-017-0237-0.

Bowers, Robert M., Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T. B.K. Reddy, Frederik Schulz, et al. 2017. "Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea." *Nature Biotechnology*. Nature Publishing Group. https://doi.org/10.1038/nbt.3893.

Brbić, Maria, Matija Piškorec, Vedrana Vidulin, Anita Kriško, Tomislav Šmuc, and Fran Supek. 2016. "The Landscape of Microbial Phenotypic Traits and Associated Genes." *Nucleic Acids Research* 44 (21): gkw964. https://doi.org/10.1093/nar/gkw964.

Brenner, Don J., James T. Staley, and Noel R. Krieg. 2015. "Classification of Procaryotic Organisms and the Concept of Bacterial Speciation." In *Bergey's Manual of Systematics of Archaea and Bacteria*, 1–9. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118960608.bm00006.

Brisson, Dustin, Dan Drecktrah, Christian H. Eggers, and D. Scott Samuels. 2012. "Genetics of Borrelia Burgdorferi." *Annual Review of Genetics* 46 (1): 515–36. https://doi.org/10.1146/annurev-genet-011112-112140.

Bush, Stephen J, Dona Foster, David W Eyre, Emily L Clark, Nicola De Maio, Liam P Shaw, Nicole Stoesser, Tim E A Peto, Derrick W Crook, and A Sarah Walker. 2020. "Genomic Diversity Affects the Accuracy of Bacterial Single-Nucleotide Polymorphism-Calling Pipelines." *GigaScience* 9: 1–21. https://doi.org/10.1093/gigascience/giaa007.

Chao, Anne. 1987. "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability." *Biometrics* 43 (4): 783. https://doi.org/10.2307/2531532.

Charles T. Parker, Brian J. Tindall, George M. Garrity. 2019. "International Code of Nomenclature of Prokaryotes." *International Journal of Systematic and Evolutionary Microbiology* 69 (1A): S1–111. https://doi.org/10.1099/ijsem.0.000778.

Chen, I-Min A, Ken Chu, Krishna Palaniappan, Manoj Pillay, Anna Ratner, Jinghua Huang, Marcel Huntemann, et al. 2019. "IMG/M v.5.0: An Integrated Data Management and Comparative Analysis System for Microbial Genomes and Microbiomes." *Nucleic Acids Research* 47. https://doi.org/10.1093/nar/gky901.

Chivian, Dylan, Eoin L. Brodie, Eric J. Alm, David E. Culley, Paramvir S. Dehal, Todd Z. DeSantis, Thomas M. Gihring, et al. 2008. "Environmental Genomics Reveals a Single-Species Ecosystem Deep within Earth." *Science* 322 (5899): 275–78. https://doi.org/10.1126/science.1155495.

Cho, Ilseung, and Martin J. Blaser. 2012. "The Human Microbiome: At the Interface of Health and Disease." *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3182.

Choudoir, Mallory J., Albert Barberán, Holly L. Menninger, Rob R. Dunn, and Noah Fierer. 2018. "Variation in Range Size and Dispersal Capabilities of Microbial Taxa." *Ecology* 99 (2): 322–34. https://doi.org/10.1002/ecy.2094.

Choudoir, Mallory J, Kevin Panke-Buisse, Cheryl P Andam, and Daniel H Buckley. 2017. "Genome Surfing As Driver of Microbial Genomic Diversity." *Trends in Microbiology* 164 (0): 1567–87. https://doi.org/10.1016/j.tim.2017.02.006.

Christensen, Henrik, Michael Hansen, and Jan Sørensen. 1999. “Counting and Size Classification of Active Soil Bacteria by Fluorescence In Situ Hybridization with an RRNA Oligonucleotide Probe.” *Applied and Environmental Microbiology* 65 (4): 1753–61.

Ciccarelli, Francesca D., Tobias Doerks, Christian von Mering, Christopher J. Creevey, Berend Snel, and Peer Bork. 2006. “Toward Automatic Reconstruction of a Highly Resolved Tree of Life.” *Science* 311 (5765): 1283–87.

Cobo-Simón, Marta, and Javier Tamames. 2017. “Relating Genomic Characteristics to Environmental Preferences and Ubiquity in Different Microbial Taxa.” *BMC Genomics* 18 (1): 499. https://doi.org/10.1186/s12864-017-3888-y.

Cockell, Charles S. 2011. “Life in the Lithosphere, Kinetics and the Prospects for Life Elsewhere.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 369 (1936): 516–37. https://doi.org/10.1098/rsta.2010.0232.

Contreras-Moreira, Bruno, and Pablo Vinuesa. 2013. “GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis.” *Applied and Environmental Microbiology* 79 (24): 7696–7701. https://doi.org/10.1128/AEM.02411-13.

Delgado-Baquerizo, Manuel, Angela M Oliverio, Tess E Brewer, Alberto Benavent-González, David J Eldridge, Richard D Bardgett, Fernando T Maestre, Brajesh K Singh, and Noah Fierer. 2018. “A Global Atlas of the Dominant Bacteria Found in Soil.” *Science (New York, N.Y.)* 359 (6373): 320–25. https://doi.org/10.1126/science.aap9516.

Delgado-Baquerizo, Manuel, Peter B. Reich, Amit N. Khachane, Colin D. Campbell, Nadine Thomas, Thomas E. Freitag, Waleed Abu Al-Soud, Søren Sørensen, Richard D. Bardgett, and Brajesh K. Singh. 2017. “It Is Elemental: Soil Nutrient Stoichiometry Drives Bacterial Diversity.” *Environmental Microbiology* 19 (3): 1176–88. https://doi.org/10.1111/1462-2920.13642.

Delmont, Tom O., and A. Murat Eren. 2018. “Linking Pangenomes and Metagenomes: The *Prochlorococcus* Metapangenome.” *PeerJ* 6 (January): e4320. https://doi.org/10.7717/peerj.4320.

Denef, Vincent J. 2018. “Peering into the Genetic Makeup of Natural Microbial Populations Using Metagenomics.” In , 49–75. Springer, Cham. https://doi.org/10.1007/13836_2018_14.

diCenzo, George C., and Turlough M. Finan. 2017. “The Divided Bacterial Genome: Structure, Function, and Evolution.” *Microbiology and Molecular Biology Reviews* 81 (3). https://doi.org/10.1128/mmbr.00019-17.

Dillon, Marcus M., Shalabh Thakur, Renan N. D. Almeida, Pauline W. Wang, Bevan S. Weir, and David S. Guttman. 2019. “Recombination of Ecologically and Evolutionarily Significant Loci Maintains Genetic Cohesion in the Pseudomonas Syringae Species Complex.” *Genome Biology* 20 (1): 3. https://doi.org/10.1186/s13059-018-1606-y.

Doolittle, W. Ford. 2012. “Population Genomics: How Bacterial Species Form and Why They Don’t Exist.” *Current Biology* 22 (11): R451–53. https://doi.org/10.1016/j.cub.2012.04.034.

Doolittle, W. Ford, and Olga Zhaxybayeva. 2009. “On the Origin of Prokaryotic Species.” *Genome Research* 19 (5): 744–56. https://doi.org/10.1101/gr.086645.108.

Dutilh, Bas E, Cristiane C Thompson, Ana CP Vicente, Michel A Marin, Clarence Lee, Genivaldo GZ Silva, Robert Schmieder, et al. 2014. “Comparative Genomics of 274 Vibrio Cholerae Genomes Reveals Mobile Functions Structuring Three Niche Dimensions.” *BMC Genomics* 15 (1): 654. https://doi.org/10.1186/1471-2164-15-654.

Editorial. 2011. “Microbiology by Numbers.” *Nature Reviews Microbiology*. Nature Publishing Group. https://doi.org/10.1038/nrmicro2644.

Fedrizzi, Tarcisio, Conor J. Meehan, Antonella Grottola, Elisabetta Giacobazzi, Giulia Fregni Serpini, Sara Tagliazucchi, Anna Fabio, et al. 2017. “Genomic Characterization of Nontuberculous Mycobacteria.” *Scientific Reports* 7 (1): 45258. https://doi.org/10.1038/srep45258.

Fenchel, Tom, and Bland J. Finlay. 2004. “The Ubiquity of Small Species: Patterns of Local and Global Diversity.” *BioScience* 54 (8): 777–84. https://doi.org/10.1641/0006-3568(2004)054[0777:tuossp]2.0.co;2.

Finlay, Bland J., and Ken J. Clarke. 1999. “Ubiquitous Dispersal of Microbial Species.” *Nature* 400 (6747): 828–828. https://doi.org/10.1038/23616.

Finlay, Brett B., Sven Pettersson, Melissa K. Melby, and Thomas C. G. Bosch. 2019. “The Microbiome Mediates Environmental Effects on Aging.” *BioEssays* 41 (10): 1800257. https://doi.org/10.1002/bies.201800257.

Foerstner, Konrad U, Christian von Mering, Sean D Hooper, and Peer Bork. 2005. “Environments Shape the Nucleotide Composition of Genomes.” *EMBO Reports* 6 (12): 1208–13. https://doi.org/10.1038/sj.embor.7400538.

Fraser, C M, S J Norris, G M Weinstock, O White, G G Sutton, R Dodson, M Gwinn, et al. 1998. “Complete Genome Sequence of Treponema Pallidum, the Syphilis Spirochete.” *Science (New York, N.Y.)* 281 (5375): 375–88. https://doi.org/10.1126/SCIENCE.281.5375.375.

Garud, Nandita R., and Katherine S. Pollard. 2020. “Population Genetics in the Human Microbiome.” *Trends in Genetics* 36 (1): 53–67. https://doi.org/10.1016/j.tig.2019.10.010.

Ghai, Rohit, Carolina Megumi Mizuno, Antonio Picazo, Antonio Camacho, and Francisco Rodriguez-Valera. 2013. “Metagenomics Uncovers a New Group of Low GC and Ultra-Small Marine Actinobacteria.” *Scientific Reports* 3 (1): 2471. https://doi.org/10.1038/srep02471.

Giovannoni, Stephen J, J Cameron Thrash, and Ben Temperton. 2014. “Implications of Streamlining Theory for Microbial Ecology.” *The ISME Journal* 8 (8): 1553–65. https://doi.org/10.1038/ismej.2014.60.

Goberna, Marta, and Miguel Verdú. 2016. “Predicting Microbial Traits with Phylogenies.” *The ISME Journal* 10 (4): 959–67. https://doi.org/10.1038/ismej.2015.171.

González-Torres, Pedro, Francisco Rodríguez-Mateos, Josefa Antón, and Toni Gabaldón. 2019. “Impact of Homologous Recombination on the Evolution of Prokaryotic Core Genomes.” *MBio* 10 (1): e02494-18. https://doi.org/10.1128/MBIO.02494-18.

Grim, Christopher J, Michael L Kotewicz, Karen A Power, Gopal Gopinath, Augusto A Franco, Karen G Jarvis, Qiong Q Yan, et al. 2013. “Pan-Genome Analysis of the Emerging Foodborne Pathogen Cronobacter Spp. Suggests a Species-Level Bidirectional Divergence Driven by Niche Adaptation.” *BMC Genomics* 14 (1): 366. https://doi.org/10.1186/1471-2164-14-366.

Groemping, Ulrike. 2006. “Relative Importance for Linear Regression in R: The Package Relaimpo.” *Journal of Statistical Software* 17 (1): 1–27. https://doi.org/10.18637/jss.v017.i01.

Grote, Jana, J Cameron Thrash, Megan J Huggett, Zachary C Landry, Paul Carini, Stephen J Giovannoni, and Michael S Rappé. 2012. “Streamlining and Core Genome Conservation among Highly Divergent Members of the SAR11 Clade.” *MBio* 3 (5): e00252-12. https://doi.org/10.1128/mBio.00252-12.

Guieysse, Benoit. 2012. “Metabolically Versatile Large-Genome Prokaryotes.” *Current Opinion in Biotechnology* 23 (3): 467–73. https://doi.org/10.1016/J.COPBIO.2011.12.022.

Hammer, Tobin J, Jon G Sanders, and Noah Fierer. 2019. “Not All Animals Need a Microbiome.” *FEMS Microbiology Letters* 366 (10): fnz117. https://doi.org/10.1093/femsle/fnz117.

Han, Kui, Zhi-feng Li, Ran Peng, Li-ping Zhu, Tao Zhou, Lu-guang Wang, Shu-guang Li, et al. 2013. “Extraordinary Expansion of a Sorangium Cellulosum Genome from an Alkaline Milieu.” *Scientific Reports* 3 (1): 2101. https://doi.org/10.1038/srep02101.

Heidelberg, John F., Jonathan A. Eisen, William C. Nelson, Rebecca A. Clayton, Michelle L. Gwinn, Robert J. Dodson, Daniel H. Haft, et al. 2000. “DNA Sequence of Both Chromosomes of the Cholera Pathogen Vibrio Cholerae.” *Nature* 406 (6795): 477–83. https://doi.org/10.1038/35020000.

Hellweger, Ferdi L, Yongjie Huang, and Haiwei Luo. 2018. “Carbon Limitation Drives GC Content Evolution of a Marine Bacterium in an Individual-Based Genome-Scale Model.” *The ISME Journal* 12 (5): 1180–87. https://doi.org/10.1038/s41396-017-0023-7.

Hessen, Dag O, Punidan D Jeyasingh, Maurine Neiman, and Lawrence J Weider. 2010. “Genome Streamlining and the Elemental Costs of Growth.” *Trends in Ecology & Evolution* 25 (2): 75–80. https://doi.org/10.1016/j.tree.2009.08.004.

Hey, Jody. 2001. “The Mind of the Species Problem.” *Trends in Ecology & Evolution* 16 (7): 326–29. https://doi.org/10.1016/S0169-5347(01)02145-0.

Huerta-Cepas, Jaime, François Serra, and Peer Bork. 2016. “ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data.” *Molecular Biology and Evolution* 33 (6): 1635–38. https://doi.org/10.1093/molbev/msw046.

Huerta-Cepas, Jaime, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, et al. 2016. “EggNOG 4.5: A Hierarchical Orthology Framework with Improved Functional Annotations for Eukaryotic, Prokaryotic and Viral Sequences.” *Nucleic Acids Research* 44 (D1): D286–93. https://doi.org/10.1093/nar/gkv1248.

Iraola, Gregorio, Samuel C. Forster, Nitin Kumar, Philippe Lehours, Sadjia Bekal, Francisco J. García-Peña, Fernando Paolicchi, et al. 2017. “Distinct Campylobacter Fetus Lineages Adapted as Livestock Pathogens and Human Pathobionts in the Intestinal Microbiota.” *Nature Communications* 8 (1): 1367. https://doi.org/10.1038/s41467-017-01449-9.

Jain, Chirag, Luis M Rodriguez-R, Adam M Phillippy, Konstantinos T Konstantinidis, and Srinivas Aluru. 2018. “High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries.” *Nature Communications* 9 (1): 5114. https://doi.org/10.1038/s41467-018-07641-9.

Jones, David T., William R. Taylor, and Janet M. Thornton. 1992. “The Rapid Generation of Mutation Data Matrices from Protein Sequences.” *Bioinformatics* 8 (3): 275–82. https://doi.org/10.1093/bioinformatics/8.3.275.

Juhas, Mario, Jan Roelof van der Meer, Muriel Gaillard, Rosalind M. Harding, Derek W. Hood, and Derrick W. Crook. 2009. “Genomic Islands: Tools of Bacterial Horizontal Gene Transfer and Evolution.” *FEMS Microbiology Reviews* 33 (2): 376–93. https://doi.org/10.1111/j.1574-6976.2008.00136.x.

Kavvas, Erol S., Edward Catoiu, Nathan Mih, James T. Yurkovich, Yara Seif, Nicholas Dillon, David Heckmann, et al. 2018. “Machine Learning and Structural Analysis of Mycobacterium Tuberculosis Pan-Genome Identifies Genetic Signatures of Antibiotic Resistance.” *Nature Communications* 9 (1): 4306. https://doi.org/10.1038/s41467-018-06634-y.

Keck, François, Frédéric Rimet, Agnès Bouchez, and Alain Franc. 2016. “Phylosignal: An R Package to Measure, Test, and Explore the Phylogenetic Signal.” *Ecology and Evolution* 6 (9): 2774–80. https://doi.org/10.1002/ece3.2051.

Kent, W. J. 2002. “BLAT---The BLAST-Like Alignment Tool.” *Genome Research* 12 (4): 656–64. https://doi.org/10.1101/gr.229202.

Kislyuk, Andrey O, Bart Haegeman, Nicholas H Bergman, and Joshua S Weitz. 2011. “Genomic Fluidity: An Integrative View of Gene Diversity within Microbial Populations.” *BMC Genomics* 12 (1): 32. https://doi.org/10.1186/1471-2164-12-32.

Klappenbach, Joel A., Johan Goris, Peter Vandamme, Tom Coenye, Konstantinos T. Konstantinidis, and James M. Tiedje. 2007. “DNA–DNA Hybridization Values and Their Relationship to Whole-Genome Sequence Similarities.” *International Journal of Systematic and Evolutionary Microbiology* 57 (1): 81–91. https://doi.org/10.1099/ijs.0.64483-0.

Konstantinidis, Konstantinos T, Alban Ramette, and James M Tiedje. 2006. “The Bacterial Species Definition in the Genomic Era.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 361 (1475): 1929–40. https://doi.org/10.1098/rstb.2006.1920.

Konstantinidis, Konstantinos T, and James M Tiedje. 2005. “Genomic Insights That Advance the Species Definition for Prokaryotes.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (7): 2567–72. https://doi.org/10.1073/pnas.0409727102.

Lackner, Gerald, Nadine Moebius, Laila Partida-Martinez, and Christian Hertweck. 2011. “Complete Genome Sequence of Burkholderia Rhizoxinica, an Endosymbiont of Rhizopus Microsporus.” *Journal of Bacteriology* 193 (3): 783–84. https://doi.org/10.1128/JB.01318-10.

Lapierre, Pascal, and J. Peter Gogarten. 2009. “Estimating the Size of the Bacterial Pan-Genome.” *Trends in Genetics* 25 (3): 107–10. https://doi.org/10.1016/J.TIG.2008.12.004.

Lê, Sébastien, Julie Josse, and François Husson. 2008. “**FactoMineR** : An *R* Package for Multivariate Analysis.” *Journal of Statistical Software* 25 (1): 1–18. https://doi.org/10.18637/jss.v025.i01.

Leibold, M. A., M. Holyoak, N. Mouquet, P. Amarasekare, J. M. Chase, M. F. Hoopes, R. D. Holt, et al. 2004. “The Metacommunity Concept: A Framework for Multi-Scale Community Ecology.” *Ecology Letters* 7 (7): 601–13. https://doi.org/10.1111/j.1461-0248.2004.00608.x.

Li, H., and R. Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Llamas, Bastien, Giuseppe Narzisi, Valerie Schneider, Peter A. Audano, Evan Biederstedt, Lon Blauvelt, Peter Bradbury, et al. 2019. “A Strategy for Building and Using a Human Reference Pangenome.” *F1000Research* 8 (October): 1751. https://doi.org/10.12688/f1000research.19630.1.

Lobkovsky, Alexander E., Yuri I. Wolf, and Eugene V. Koonin. 2013. “Gene Frequency Distributions Reject a Neutral Model of Genome Evolution.” *Genome Biology and Evolution* 5 (1): 233–42. https://doi.org/10.1093/gbe/evt002.

Locey, Kenneth J., and Jay T. Lennon. 2016. “Scaling Laws Predict Global Microbial Diversity.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (21): 5970–75. https://doi.org/10.1073/pnas.1521291113.

Lovelock, James E. 1979. “Gaia: A New Look At Life On Earth.”

Lynch, Michael. 2006. “Streamlining and Simplification of Microbial Genome Architecture.” *Annual Review of Microbiology* 60 (1): 327–49. https://doi.org/10.1146/annurev.micro.60.080805.142300.

Ma, Bing, Michael T. France, Jonathan Crabtree, Johanna B. Holm, Michael S. Humphrys, Rebecca M. Brotman, and Jacques Ravel. 2020. “A Comprehensive Non-Redundant Gene Catalog Reveals Extensive within-Community Intraspecies Diversity in the Human Vagina.” *Nature Communications* 11 (1): 1–13. https://doi.org/10.1038/s41467-020-14677-3.

Maistrenko, Oleksandr M., Daniel R. Mende, Mechthild Luetge, Falk Hildebrand, Thomas S.B. Schmidt, Simone S. Li, João F.Matias Rodrigues, et al. 2020. “Disentangling the Impact of Environmental and Phylogenetic Constraints on Prokaryotic Within-Species Diversity.” *ISME Journal*, February, 1–13. https://doi.org/10.1038/s41396-020-0600-z.

Maistrenko, Oleksandr M., Svitlana V. Serga, Alexander M. Vaiserman, and Iryna A. Kozeretska. 2015. “Effect of Wolbachia Infection on Aging and Longevity-Associated Genes in Drosophila.” In *Life Extension: Lessons from Drosophila*, edited by Alexander M. Vaiserman, Alexey A. Moskalev, and Elena G. Pasyukova, 83–104. Healthy Ageing and Longevity. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-18326-8_4.

———. 2016. “Longevity-Modulating Effects of Symbiosis: Insights from Drosophila–Wolbachia Interaction.” *Biogerontology* 17 (5): 785–803. https://doi.org/10.1007/s10522-016-9653-9.

Marschall, Tobias, Manja Marz, Thomas Abeel, Louis Dijkstra, Bas E Dutilh, Ali Ghaffaari, Paul Kersey, et al. 2016. “Computational Pan-Genomics: Status, Promises and Challenges.” *Briefings in Bioinformatics* 19 (1): bbw089. https://doi.org/10.1093/bib/bbw089.

MartÃ-nez-Cano, David J., Mariana Reyes-Prieto, Esperanza MartÃ-nez-Romero, Laila P. Partida-MartÃ-nez, Amparo Latorre, AndrÃ©s Moya, and Luis Delaye. 2015. “Evolution of Small Prokaryotic Genomes.” *Frontiers in Microbiology* 5 (January): 742. https://doi.org/10.3389/fmicb.2014.00742.

Martiny, Adam C, Kathleen Treseder, and Gordon Pusch. 2013. “Phylogenetic Conservatism of Functional Traits in Microorganisms.” *The ISME Journal* 7 (4): 830–38. https://doi.org/10.1038/ismej.2012.160.

Martiny, Jennifer B H, Stuart E Jones, Jay T Lennon, and Adam C Martiny. 2015. “Microbiomes in Light of Traits: A Phylogenetic Perspective.” *Science (New York, N.Y.)* 350 (6261): aac9323. https://doi.org/10.1126/science.aac9323.

Matias Rodrigues, João F, Thomas S B Schmidt, Janko Tackmann, and Christian von Mering. 2017. “MAPseq: Highly Efficient k-Mer Search with Confidence Estimates, for RRNA Sequence Analysis.” Edited by Inanc Birol. *Bioinformatics* 33 (23): 3808–10. https://doi.org/10.1093/bioinformatics/btx517.

Mayden, R. L. 1997. “A Hierarchy of Species Concepts: The Denouement in the Saga of the Species Problem.”

McCutcheon, John P., Bradon R. McDonald, and Nancy A. Moran. 2009a. “Origin of an Alternative Genetic Code in the Extremely Small and GC–Rich Genome of a Bacterial Symbiont.” Edited by Ivan Matic. *PLoS Genetics* 5 (7): e1000565. https://doi.org/10.1371/journal.pgen.1000565.

McCutcheon, John P, Bradon R McDonald, and Nancy A Moran. 2009b. “Convergent Evolution of Metabolic Roles in Bacterial Co-Symbionts of Insects.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (36): 15394–99. https://doi.org/10.1073/pnas.0906424106.

McCutcheon, John P., and Carol D. von Dohlen. 2011. “An Interdependent Metabolic Patchwork in the Nested Symbiosis of Mealybugs.” *Current Biology* 21 (16): 1366–72. https://doi.org/10.1016/J.CUB.2011.06.051.

McFall-Ngai, Margaret, Michael G Hadfield, Thomas C G Bosch, Hannah V Carey, Tomislav Domazet-Lošo, Angela E Douglas, Nicole Dubilier, et al. 2013. “Animals in a Bacterial World, a New Imperative for the Life Sciences.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (9): 3229–36. https://doi.org/10.1073/pnas.1218525110.

McInerney, James O., Alan McNally, and Mary J. O’Connell. 2017a. “Why Prokaryotes Have Pangenomes.” *Nature Microbiology* 2 (4): 17040. https://doi.org/10.1038/nmicrobiol.2017.40.

McInerney, James O., Alan McNally, and Mary J. O’Connell. 2017b. “Reply to ‘The Population Genetics of Pangenomes.’” *Nature Microbiology* 2 (12): 1575–1575. https://doi.org/10.1038/s41564-017-0068-4.

Medini, Duccio, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. 2005. “The Microbial Pan-Genome.” *Current Opinion in Genetics & Development* 15 (6): 589–94. https://doi.org/10.1016/J.GDE.2005.09.006.

Mende, Daniel R., Ivica Letunic, Jaime Huerta-Cepas, Simone S. Li, Kristoffer Forslund, Shinichi Sunagawa, and Peer Bork. 2017. “ProGenomes: A Resource for Consistent Functional and Taxonomic Annotations of Prokaryotic Genomes.” *Nucleic Acids Research* 45 (D1): D529–34. https://doi.org/10.1093/nar/gkw989.

Mende, Daniel R, Ivica Letunic, Oleksandr M Maistrenko, Thomas S B Schmidt, Alessio Milanese, Lucas Paoli, Ana Hernández-Plaza, et al. 2020. “ProGenomes2: An Improved Database for Accurate and Consistent Habitat, Taxonomic and Functional Annotations of Prokaryotic Genomes.” *Nucleic Acids Research* 48 (D1): D621–25. https://doi.org/10.1093/nar/gkz1002.

Mende, Daniel R, Shinichi Sunagawa, Georg Zeller, and Peer Bork. 2013. “Accurate and Universal Delineation of Prokaryotic Species.” *Nature Methods* 10 (9): 881–84. https://doi.org/10.1038/nmeth.2575.

Mering, C. von, P. Hugenholtz, J. Raes, S. G. Tringe, T. Doerks, L. J. Jensen, N. Ward, and P. Bork. 2007. “Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments.” *Science* 315 (5815): 1126–30.

Milanese, Alessio, Daniel R. Mende, Lucas Paoli, Guillem Salazar, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Pascal Hingamp, et al. 2019. “Microbial Abundance, Activity and Population Genomic Profiling with MOTUs2.” *Nature Communications* 10 (1): 1014. https://doi.org/10.1038/s41467-019-08844-4.

Miller, Ian J, Niti Vanee, Stephen S Fong, Grace E Lim-Fong, and Jason C Kwan. 2016. “Lack of Overt Genome Reduction in the Bryostatin-Producing Bryozoan Symbiont "Candidatus Endobugula Sertula".” *Applied and Environmental Microbiology* 82 (22): 6573–83. https://doi.org/10.1128/AEM.01800-16.

Moldovan, Mikhail A., and Mikhail S. Gelfand. 2018. “Pangenomic Definition of Prokaryotic Species and the Phylogenetic Structure of Prochlorococcus Spp.” *Frontiers in Microbiology* 9 (March): 428. https://doi.org/10.3389/fmicb.2018.00428.

Mostowy, Rafal, Nicholas J Croucher, Cheryl P Andam, Jukka Corander, William P Hanage, and Pekka Marttinen. 2017. “Efficient Inference of Recent and Ancestral Recombination within Bacterial

Populations." *Molecular Biology and Evolution* 34 (5): 1167–1182. https://doi.org/10.1093/molbev/msx066.

Mukherjee, Supratim, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Hema Y Katta, Alejandro Mojica, I-Min A Chen, Nikos C Kyrpides, and T B K Reddy. 2018. "Genomes OnLine Database (GOLD) v.7: Updates and New Features." *Nucleic Acids Research* 47: 649–59. https://doi.org/10.1093/nar/gky977.

Münkemüller, Tamara, Sébastien Lavergne, Bruno Bzeznik, Stéphane Dray, Thibaut Jombart, Katja Schiffers, and Wilfried Thuiller. 2012. "How to Measure and Test Phylogenetic Signal." *Methods in Ecology and Evolution* 3 (4): 743–56. https://doi.org/10.1111/j.2041-210X.2012.00196.x.

Nayfach, S., and K. S. Pollard. 2015. "Population Genetic Analyses of Metagenomes Reveal Extensive Strain-Level Variation in Prevalent Human-Associated Bacteria." *BioRxiv*, 031757. https://doi.org/10.1101/031757.

Nayfach, Stephen, Beltran Rodriguez-Mueller, Nandita Garud, and Katherine S Pollard. 2016. "An Integrated Metagenomics Pipeline for Strain Profiling Reveals Novel Patterns of Bacterial Transmission and Biogeography." *Genome Research* 26 (11): 1612–25. https://doi.org/10.1101/gr.201863.115.

Nilsson, A I, S Koskiniemi, S Eriksson, E Kugelberg, J C D Hinton, and D I Andersson. 2005. "Bacterial Genome Size Reduction by Experimental Evolution." *Proceedings of the National Academy of Sciences of the United States of America* 102 (34): 12112–16. https://doi.org/10.1073/pnas.0503654102.

Nimnoi, Pongrawee, and Neelawan Pongsilp. 2020. "Marine Bacterial Communities in the Upper Gulf of Thailand Assessed by Illumina Next-Generation Sequencing Platform." *BMC Microbiology* 20 (1): 19. https://doi.org/10.1186/s12866-020-1701-6.

Nixon, Kevin C., and Quentin D. Wheeler. 1990. "An Amplification of the Phylogenetic Species Concept." *Cladistics* 6 (3): 211–23. https://doi.org/10.1111/j.1096-0031.1990.tb00541.x.

O'Donnell, Michael, Lance Langston, and Bruce Stillman. 2013. "Principles and Concepts of DNA Replication in Bacteria, Archaea, and Eukarya." *Cold Spring Harbor Perspectives in Biology* 5 (7): a010108. https://doi.org/10.1101/cshperspect.a010108.

Olm, Matthew R., Alexander Crits-Christoph, Spencer Diamond, Adi Lavy, Paula B. Matheus Carnevali, and Jillian F. Banfield. 2020. "Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries." *MSystems* 5 (1). https://doi.org/10.1128/msystems.00731-19.

O'Malley, Maureen A. 2008. "'Everything Is Everywhere: But the Environment Selects': Ubiquitous Distribution and Ecological Determinism in Microbial Biogeography." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 39 (3): 314–25. https://doi.org/10.1016/J.SHPSC.2008.06.005.

Orme, David. 2013. "The Caper Package: Comparative Analysis of Phylogenetics and Evolution in R. R Package Version 5.2," 1–36.

Page, Andrew J., Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T.G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. 2015. "Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis." *Bioinformatics* 31 (22): 3691–93. https://doi.org/10.1093/bioinformatics/btv421.

Pagel, M. 1999. "Inferring the Historical Patterns of Biological Evolution." *Nature* 401 (6756): 877–84. https://doi.org/10.1038/44766.

Paquola, Apuã C. M., Huma Asif, Carlos Alberto de Bragança Pereira, Bruno César Feltes, Diego Bonatto, Wanessa Cristina Lima, and Carlos Frederico Martins Menck. 2018. "Horizontal Gene Transfer Building Prokaryote Genomes: Genes Related to Exchange Between Cell and Environment Are Frequently Transferred." *Journal of Molecular Evolution*, March, 1–14. https://doi.org/10.1007/s00239-018-9836-x.

Parks, Donovan H, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2018. "A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially Revises the Tree of Life." *Nature Biotechnology* 36 (10): 996. https://doi.org/10.1038/nbt.4229.

Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, et al. 2019. “Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle.” *Cell* 176 (3): 649-662.e20. https://doi.org/10.1016/j.cell.2019.01.001.

Pearce, Madison E., Nabil Fareed Alikhan, Timothy J. Dallman, Zhemin Zhou, Kathie Grant, and Martin C.J. Maiden. 2018. “Comparative Analysis of Core Genome MLST and SNP Typing within a European Salmonella Serovar Enteritidis Outbreak.” *International Journal of Food Microbiology* 274 (June): 1–11. https://doi.org/10.1016/j.ijfoodmicro.2018.02.023.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. “FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments.” Edited by Art F. Y. Poon. *PLoS ONE* 5 (3): e9490. https://doi.org/10.1371/journal.pone.0009490.

Puigbò, Pere, Alexander E Lobkovsky, David M Kristensen, Yuri I Wolf, and Eugene V Koonin. 2014. “Genomes in Turmoil: Quantification of Genome Dynamics in Prokaryote Supergenomes.” *BMC Biology* 12 (1): 66. https://doi.org/10.1186/s12915-014-0066-4.

Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, et al. 2010. “A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing.” *Nature* 464 (7285): 59–65. https://doi.org/10.1038/nature08821.

Queiroz, Kevin de. 2005. “Ernst Mayr and the Modern Concept of Species.” *Proceedings of the National Academy of Sciences* 102 (suppl 1): 6600–6607. https://doi.org/10.1073/pnas.0502030102.

R Core Team. 2018. “R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.” 2018. http://www.r-project.org.

Radman, M, F Taddei, and I Matic. 2000. “DNA Repair Systems and Bacterial Evolution.” *Cold Spring Harbor Symposia on Quantitative Biology* 65 (January): 11–19. https://doi.org/10.1101/sqb.2000.65.11.

Raynaud, Xavier, and Naoise Nunan. 2014. “Spatial Ecology of Bacteria at the Microscale in Soil.” Edited by Francesco Pappalardo. *PLoS ONE* 9 (1): e87217. https://doi.org/10.1371/journal.pone.0087217.

Reichenberger, Erin R., Gail Rosen, Uri Hershberg, and Ruth Hershberg. 2015. “Prokaryotic Nucleotide Composition Is Shaped by Both Phylogeny and the Environment.” *Genome Biology and Evolution* 7 (5): 1380–89. https://doi.org/10.1093/gbe/evv063.

Rocha, Eduardo P C. 2018. “Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics.” Edited by Sudhir Kumar. *Molecular Biology and Evolution* 35 (6): 1338–47. https://doi.org/10.1093/molbev/msy078.

Rodriguez-Valera, Francisco, and David W Ussery. 2012. “Is the Pan-Genome Also a Pan-Selectome?” *F1000Research* 1 (September): 1–7. https://doi.org/10.12688/f1000research.1-16.v1.

Roller, Maša, Vedran Lučić, István Nagy, Tina Perica, and Kristian Vlahoviček. 2013. “Environmental Shaping of Codon Usage and Functional Adaptation across Microbial Communities.” *Nucleic Acids Research* 41 (19): 8842–52. https://doi.org/10.1093/nar/gkt673.

Rouli, L., V. Merhej, P.-E. Fournier, and D. Raoult. 2015. “The Bacterial Pangenome as a New Tool for Analysing Pathogenic Bacteria.” *New Microbes and New Infections* 7 (September): 72–85. https://doi.org/10.1016/j.nmni.2015.06.005.

Schmidt, Thomas S.B., Jeroen Raes, and Peer Bork. 2018. “The Human Gut Microbiome: From Association to Modulation.” *Cell*. Cell Press. https://doi.org/10.1016/j.cell.2018.02.044.

Schneiker, Susanne, Olena Perlova, Olaf Kaiser, Klaus Gerth, Aysel Alici, Matthias O Altmeyer, Daniela Bartels, et al. 2007. “Complete Genome Sequence of the Myxobacterium Sorangium Cellulosum.” *Nature Biotechnology* 25 (11): 1281–89. https://doi.org/10.1038/nbt1354.

Scholz, Matthias, Doyle V Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, Adrian Tett, Ardythe L Morrow, and Nicola Segata. 2016. “Strain-Level Microbial Epidemiology and Population Genomics from Shotgun Metagenomics.” *Nature Methods* 13 (5): 435–38. https://doi.org/10.1038/nmeth.3802.

Seemann, T. n.d. “GitHub - Tseemann/Barrnap: Bacterial Ribosomal RNA Predictor.” Accessed November 9, 2019. https://github.com/tseemann/barrnap.
Seemann, T. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14): 2068–69. https://doi.org/10.1093/bioinformatics/btu153.
Sender, Ron, Shai Fuchs, and Ron Milo. 2016. “Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans.” *Cell* 164 (3): 337–40. https://doi.org/10.1016/j.cell.2016.01.013.
Seo, Young-Su, Jaeyun Lim, Beom-Soon Choi, Hongsup Kim, Eunhye Goo, Bongsoo Lee, Jong-Sung Lim, et al. 2011. “Complete Genome Sequence of Burkholderia Gladioli BSR3.” *Journal of Bacteriology* 193 (12): 3149. https://doi.org/10.1128/JB.00420-11.
Seshadri, Rekha, Garry S A Myers, Hervé Tettelin, Jonathan A Eisen, John F Heidelberg, Robert J Dodson, Tanja M Davidsen, et al. 2004. “Comparison of the Genome of the Oral Pathogen Treponema Denticola with Other Spirochete Genomes.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (15): 5646–51. https://doi.org/10.1073/pnas.0307639101.
Shaiber, Alon, and A. Murat Eren. 2019. “Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories.” *MBio*. American Society for Microbiology. https://doi.org/10.1128/mBio.00725-19.
Shapiro, B. Jesse. 2017. “The Population Genetics of Pangenomes.” *Nature Microbiology* 2 (12): 1574–1574. https://doi.org/10.1038/s41564-017-0066-6.
Shapiro, B Jesse, Jonathan Friedman, Otto X Cordero, Sarah P Preheim, Sonia C Timberlake, Gitta Szabó, Martin F Polz, and Eric J Alm. 2012. “Population Genomics of Early Events in the Ecological Differentiation of Bacteria.” *Science* 336 (6077): 48–51. https://doi.org/10.1126/science.1218198.
Sheppard, Samuel K., Xavier Didelot, Guillaume Meric, Alicia Torralbo, Keith A. Jolley, David J. Kelly, Stephen D. Bentley, Martin C. J. Maiden, Julian Parkhill, and Daniel Falush. 2013. “Genome-Wide Association Study Identifies Vitamin B5 Biosynthesis as a Host Specificity Factor in Campylobacter.” *Proceedings of the National Academy of Sciences* 110 (29): 11923–27. https://doi.org/10.1073/pnas.1305559110.
Sievers, Fabian, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. “Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega.” *Molecular Systems Biology* 7 (1): 539. https://doi.org/10.1038/msb.2011.75.
Snel, Berend, Peer Bork, and Martijn A. Huynen. 1999. “Genome Phylogeny Based on Gene Content.” *Nature Genetics* 21 (1): 108–10. https://doi.org/10.1038/5052.
Snipen, Lars, and Kristian Hovde Liland. 2015. “Micropan: An R-Package for Microbial Pan-Genomics.” *BMC Bioinformatics* 16 (1): 79. https://doi.org/10.1186/s12859-015-0517-0.
Sorek, Rotem, Yiwen Zhu, Christopher J Creevey, M Pilar Francino, Peer Bork, and Edward M Rubin. 2007. “Genome-Wide Experimental Determination of Barriers to Horizontal Gene Transfer.” *Science (New York, N.Y.)* 318 (5855): 1449–52. https://doi.org/10.1126/science.1147112.
Staley, J T, and A Konopka. 1985. “Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats.” *Annual Review of Microbiology* 39 (1): 321–46. https://doi.org/10.1146/annurev.mi.39.100185.001541.
Stolz, John F. 2017. “Gaia and Her Microbiome.” *FEMS Microbiology Ecology* 93: 247. https://doi.org/10.1093/femsec/fiw247.
Sung, Way, Matthew S Ackerman, Samuel F Miller, Thomas G Doak, and Michael Lynch. 2012. “Drift-Barrier Hypothesis and Mutation-Rate Evolution.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (45): 18488–92. https://doi.org/10.1073/pnas.1216223109.
Symonds, Matthew R. E., and Simon P. Blomberg. 2014. “A Primer on Phylogenetic Generalised Least Squares.” In *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*, 105–30. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-43550-2_5.

Tamames, Javier, Pablo D. Sánchez, Pablo I. Nikel, and Carlos Pedrós-Alió. 2016. "Quantifying the Relative Importance of Phylogeny and Environmental Preferences As Drivers of Gene Content in Prokaryotic Microorganisms." *Frontiers in Microbiology* 7 (March): 433. https://doi.org/10.3389/fmicb.2016.00433.

Teeling, Hanno, and Frank Oliver Glöckner. 2012. "Current Opportunities and Challenges in Microbial Metagenome Analysis--a Bioinformatic Perspective." *Briefings in Bioinformatics* 13 (6): 728–42. https://doi.org/10.1093/bib/bbs039.

Tettelin, Hervé, Vega Masignani, Michael J Cieslewicz, Claudio Donati, Duccio Medini, Naomi L Ward, Samuel V Angiuoli, et al. 2005. "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): 13950–55. https://doi.org/10.1073/pnas.0506758102.

Torsvik, V., J. Goksoyr, and F. L. Daae. 1990. "High Diversity in DNA of Soil Bacteria." *Applied and Environmental Microbiology* 56 (3): 782–87. https://doi.org/10.1128/aem.56.3.782-787.1990.

Vernikos, G. S. 2020. "A Review of Pangenome Tools and Recent Studies." In *The Pangenome*, 89–112. Springer International Publishing. https://doi.org/10.1007/978-3-030-38281-0_4.

Vernikos, George, Duccio Medini, David R. Riley, and Herv?? Tettelin. 2015. "Ten Years of Pan-Genome Analyses." *Current Opinion in Microbiology* 23: 148–54. https://doi.org/10.1016/j.mib.2014.11.016.

Vos, Michiel, and Adam Eyre-Walker. 2017. "Are Pangenomes Adaptive or Not?" *Nature Microbiology* 2 (12): 1576–1576. https://doi.org/10.1038/s41564-017-0067-5.

Wang, Jun, and Huijue Jia. 2016. "Metagenome-Wide Association Studies: Fine-Mining the Microbiome." *Nature Reviews Microbiology*. Nature Publishing Group. https://doi.org/10.1038/nrmicro.2016.83.

Wattam, Alice R., David Abraham, Oral Dalay, Terry L. Disz, Timothy Driscoll, Joseph L. Gabbard, Joseph J. Gillespie, et al. 2014. "PATRIC, the Bacterial Bioinformatics Database and Analysis Resource." *Nucleic Acids Research* 42 (D1): D581–91. https://doi.org/10.1093/nar/gkt1099.

Whitaker, Melissa R. L., Shayla Salzman, Jon Sanders, Martin Kaltenpoth, and Naomi E. Pierce. 2016. "Microbial Communities of Lycaenid Butterflies Do Not Correlate with Larval Diet." *Frontiers in Microbiology* 7. https://doi.org/10.3389/fmicb.2016.01920.

Whitfield, John. 2005. "Biogeography: Is Everything Everywhere? Researchers Have Dug up Some Surprising Evidence Casting Doubt on the Long-Held Belief That Microbes Are Impervious to Geographic Constraints." *Science* 310 (5750): 960–62.

Whitman, W B, D C Coleman, and W J Wiebe. 1998. "Prokaryotes: The Unseen Majority." *Proceedings of the National Academy of Sciences of the United States of America* 95 (12): 6578–83. https://doi.org/10.1073/PNAS.95.12.6578.

Wilkins, John S. 2003. "How to Be a Chaste Species Pluralist-Realist: The Origins of Species Modes and the Synapomorphic Species Concept." *Biology & Philosophy* 18 (5): 621–38. https://doi.org/10.1023/A:1026390327482.

Xiao, Jingfa, Zhewen Zhang, Jiayan Wu, and Jun Yu. 2015. "A Brief Review of Software Tools for Pangenomics." *Genomics, Proteomics and Bioinformatics*. Beijing Genomics Institute. https://doi.org/10.1016/j.gpb.2015.01.007.

Xiao, Liang, Jordi Estellé, Pia Kiilerich, Yuliaxis Ramayo-Caldas, Zhongkui Xia, Qiang Feng, Suisha Liang, et al. 2016. "A Reference Gene Catalogue of the Pig Gut Microbiome." *Nature Microbiology* 1 (12): 1–6. https://doi.org/10.1038/nmicrobiol.2016.161.

Xiao, Liang, Qiang Feng, Suisha Liang, Si Brask Sonne, Zhongkui Xia, Xinmin Qiu, Xiaoping Li, et al. 2015. "A Catalog of the Mouse Gut Metagenome." *Nature Biotechnology* 33 (10): 1103–8. https://doi.org/10.1038/nbt.3353.

Yang, Ziheng. 2007. "PAML 4: Phylogenetic Analysis by Maximum Likelihood." *Molecular Biology and Evolution* 24 (8): 1586–1591. https://doi.org/10.1093/molbev/msm088.

Zhang, Ying, and Stefan M. Sievert. 2014. “Pan-Genome Analyses Identify Lineage- and Niche-Specific Markers of Evolution and Adaptation in Epsilonproteobacteria.” *Frontiers in Microbiology* 5 (March): 110. https://doi.org/10.3389/fmicb.2014.00110.

Zhu, Ana, Shinichi Sunagawa, Daniel R Mende, and Peer Bork. 2015. “Inter-Individual Differences in the Gene Content of Human Gut Bacterial Species.” https://doi.org/10.1186/s13059-015-0646-9.

Zilber-Rosenberg, Ilana, and Eugene Rosenberg. 2008. “Role of Microorganisms in the Evolution of Animals and Plants: The Hologenome Theory of Evolution.” *FEMS Microbiology Reviews* 32 (5): 723–35. https://doi.org/10.1111/j.1574-6976.2008.00123.x.

Zuber, Verena, and Korbinian Strimmer. 2011. “High-Dimensional Regression and Variable Selection Using CAR Scores.” *Statistical Applications in Genetics and Molecular Biology* 10 (1): 1–27. https://doi.org/10.2202/1544-6115.1730.