Title: Toward quantifying the adaptive role of bacterial pangenomes during environmental perturbations

Running Title

The role of the pangenome in adaptation

Authors Full Names and Affiliations

Roth E. Conrad[1,5], Tomeu Viver[3,5], Juan F. Gago[3], Janet K. Hatt[4], Stephanus N. Venter[2], Ramon Rosselló-Móra[3*], Konstantinos T. Konstantinidis[4*]

[1]Ocean Science & Engineering, School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA

[2] Department of Biochemistry, Genetics and Microbiology, and Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa

[3]Marine Microbiology Group, Department of Animal and Microbial Biodiversity, Mediterranean Institutes for Advanced Studies (IMEDEA, CSIC-UIB), Esporles, Spain

[4]School of Civil & Environmental Engineering, Georgia Institute of Technology, Atlanta Georgia, USA

[5]These authors contributed equally to this work.

Corresponding Author Info

* Correspondence should be addressed to Konstantinos Konstantinidis (kostas@ce.gatech.edu), and Ramon Rossello-Mora (rossello-mora@uib.es).

Competing interest

The authors declare no competing interest.

Abstract

Metagenomic surveys have revealed that natural microbial communities are predominantly composed of sequence-discrete, species-like populations but the genetic and/or ecological processes that maintain such populations remain speculative, limiting our understanding of population speciation and adaptation to perturbations. To address this knowledge gap, we sequenced 112 *Salinibacter ruber* isolates and 12 companion metagenomes from four adjacent saltern ponds in Mallorca, Spain that were experimentally manipulated to dramatically alter salinity and light intensity, the two major drivers of this ecosystem. Our analyses showed that the pangenome of the local *Sal. ruber* population is open and similar in size (~15000 genes) to that of randomly sampled *Escherichia coli* genomes. While most of the accessory (non-core) genes were isolate-specific and showed low *in situ* abundances based on the metagenomes compared to the core genes, indicating that they were functionally unimportant and/or transient, 3.5% of them became abundant when salinity (but not light) conditions changed and encoded for functions related to osmoregulation. Nonetheless, the ecological advantage of these genes, while significant, was apparently not strong enough to purge diversity within the population. Collectively, our results provide an explanation for how this immense intra-population gene diversity is maintained, which has implications for the prokaryotic species concept.

Introduction

Our understanding of the intraspecific diversity of prokaryotes is based largely on the comparative analyses of collections of isolates. Since these isolates typically originate from a variety of samples, habitats, and times, they often show varying fitness backgrounds and genomic adaptations specific to the local conditions at the time and place of isolation. Accordingly, the number of nonredundant genes (i.e., the pangenome) within many of the species formed by such isolates appears to increase continuously with the addition of each new isolate (i.e., the pangenome is open), and thus is quite large e.g., >30 000 or more genes than the human genome. This is especially the case for free-living, ecologically versatile species, contrasting with obligate symbionts, and other species of narrow ecological niche, that tend to have smaller or closed pangenomes (1, 2). Pangenomes are comprised of core and accessory genes (2-5). Core genes are shared by all or almost all (>90% of the total) genomes of a species and account for the general ecological and phenotypic properties of the species. Accessory genes, also referred to as auxiliary, dispensable, variable, or flexible genes, are present in only one or a few genomes of a species and can be further divided into strain-specific (isolate-specific), rare, or common genes based on the fraction of genomes found to contain the gene. While this phenomenon is well documented, it is still unclear whether results from the comparison of isolates acquired from different habitats and samples translate well to the diversity within natural populations; that is, a population of conspecific strains co-existing in the same environment or sample.

The emerging picture from culture-independent metagenomic surveys of microbial communities is that bacteria and archaea predominantly form species-like, sequence-discrete populations with intraspecific genome sequence relatedness typically ranging from ~95% to

~100% genome-aggregate nucleotide identity (ANI) depending on the population considered i.e., populations having a more recent bottleneck or genome sweep event show lower levels of intraspecific diversity (6-8). In contrast, ANI values between distinct (interspecific) populations are typically lower than 90%. Sequence-discrete populations have been commonly found in many different habitats including marine, freshwater, soils, sediment, human gut, and biofilms, and are typically persistent over time and space (9-12) indicating that they are not ephemeral but long-lived entities. Sequence-indiscrete populations (or, in other words, a continuum of ANI values) are rarely encountered in these previous studies, and when found are almost always attributable to the mixing of distinct habitats such as the mixing of water from different depths in the ocean water column during upwelling events (9, 12, 13). Consistent with the metagenomics perspective, a recent analysis of all available isolate genomes of named species (n ~ 90 000) revealed a similar bimodal distribution in ANI values; that is, a small number of genome pairs show 85-95% ANI relative to pairs showing either >95% or <85% ANI (i.e., an "ANI gap" or "discontinuity") (14). These data reveal that a similar genetic (sequence) discontinuity is characteristic of both naturally occurring populations as well as classified (named in accordance with the bacteriological code) species comprising genomes of isolates. It remains to be tested, however, if functional (gene content) diversity patterns are also similar between naturally occurring populations and pure culture collections. Furthermore, it is equally important to elucidate the gene content dynamics of local populations to better understand the underlying evolutionary processes that shape species-like, sequence-discrete populations and maintain coherent species-like genomic structure on a global scale [reviewed in (6, 15)].

More specifically, quantifying the extent of gene content variation (i.e., the accessory pangenome) within natural microbial populations is important to better understand the metabolic and ecological plasticity of a population and how accessory genes facilitate adaptation to environmental perturbations. One prevailing hypothesis is that non-core gene diversity is largely neutral or ephemeral resulting from random genetic drift and a lack of competition among members of the population that is strong enough to lead to complete dominance of the member(s) carrying the genes in question (16). A competing hypothesis is that co-occurring subpopulations may accumulate substantial and ecologically important (non-neutral) gene content differences that enable, for instance, differentiated affinity for the same substrate, and thus are on their way to parapatric or sympatric speciation (15, 17, 18). The experimental data to test these hypotheses rigorously and quantitatively for a natural population are currently lacking, a gap that the present study aimed to fulfill.

Metagenome-assembled genomes (MAGs) obtained from environmental samples using population genome binning techniques have been used to study sequence-discrete populations. However, verifying the purity, completeness, and accuracy of these MAGs is challenging (19-23). Furthermore, even with high quality MAGs, the extent to which the recovered gene content represents the pangenome of a population remains speculative, and can sometimes be low (24). MAGs cannot fully capture the total standing gene content variation of a natural population due to i) limitations in short-read assembly of hypervariable or genomic repeat regions, ii) low coverage of rare or accessory genes iii) high coverage of conserved regions shared across multiple populations, and iv) challenges in accurately grouping assembled contigs into MAGs during population genome binning (12, 24, 25). Although a few longitudinal shotgun metagenomic studies have attempted to quantify the genetic variation within natural microbial populations, their primary focus has been on single nucleotide polymorphisms (SNPs) (e.g., allelic variation) rather than gene content variation (7, 26-28). A few studies have also shown that gene content can fluctuate within a population as an effect of the dominance of different strains (7, 12, 25) by querying isolate genomes or MAGs against time or spatial series metagenomes, but to our knowledge no study has quantified gene content diversity within natural populations or the existence of distinct (co-occurring) subpopulations based on shared gene content. For this, representative isolate genomes and/or whole genomes obtained through single-cell techniques (29, 30) need to be combined with MAGs and shotgun metagenomes, as performed in our study.

Solar salterns are semi-artificial environments used for harvesting salt for human consumption, and they harbor reduced microbial diversity driven primarily by environmental stressors, most notably sunlight intensity and high salt concentrations (31). *Salinibacter ruber* represents the major component of the bacterial fraction of salterns and is commonly isolated from hypersaline habitats globally (32). Hence, we used solar salterns in Mallorca, Spain as our experimental system, and focused on *Sal. ruber* to quantify the intraspecific gene content variation of its naturally occurring population on a local scale by combining metagenomic sampling with extensive isolate culturing efforts on the same samples. Samples were collected during a multi-stressor mesocosm experiment wherein salterns were stressed by different light intensity and salinity exposure regimes. Specifically, one control and three experimental ponds were filled with the same pre-concentrated inlet brines (Fig. S1). Apart from the inlet brines coming from the same source, the experimental ponds were isolated (no brine flow between ponds) and challenged by: i) sunlight intensity alterations using a shading mesh to cover a previously unshaded pond and ii) uncover a previously shaded pond, resulting in ~37-fold reduction or increase in sun irradiation, respectively (31), and, iii) abrupt changes in salt concentration from ~34% (salt saturation or precipitation level) to ~12% by dilution with freshwater over a period of four hours ((33), Fig. S1). The *Sal. ruber* populations from each of the four ponds were observed for one-month post-treatment by sampling 207 *Sal. ruber* isolates and 12 whole-community shotgun metagenomes. Metagenomes were sequenced from three time points: time-zero (Z), one-week (W), and one-month (M). Isolates from all ponds were sequenced at time-zero and one-month with an additional sampling day for the unshaded-shaded pond (ii) and the diluted pond (iii). The diluted pond had re-established salt-saturation conditions by natural evaporation at the end of the one-month experiment. See figure S1 for more details. Herein, we report the observed gene content diversity and the relative *in situ* gene abundance of the local *Sal. ruber* population during ambient (control) and experimentally altered environmental conditions.

Results

Sampling the local *Sal. ruber* population

To characterize the intraspecific diversity of the local *Sal. ruber* population *in situ*, we isolated 207 strains during a one-month time period from four adjacent saltern ponds at 'Es Trenc' in Mallorca, Spain (Fig. S1). Based on MALDI-TOF MS and RAPD signatures (Fig. S2), we selected 123 non-clonal isolates (i.e., strains with different RAPD profiles) for genome sequencing. 54 genomes were collected across the four ponds at time-zero (Z) and 54 genomes were collected across the four ponds at one-month (M). An additional 5 genomes were collected at two-days post dilution from the diluted pond and 10 genomes were collect at one-week (W) from the unshaded-shaded pond. See figure S1 for more details. After genome assembly, we selected 112 draft genomes that were determined to be free of contamination and of sufficient quality compared to previously completed *Sal. ruber* genomes for subsequent analyses (Sup.

Excel File 1). Our genomes had an average of 268 (stdev=60) contigs per assembly, an average N50 value of 25 369 bps (stdev= 7 357) and an average sequencing depth of 15X (stdev=5). The mean genome sequence length was 3 828 264 bps (stdev=140 342) with an average of 65.8% G+C content (stdev=0.3%) and an average of 3 369 unfiltered open reading frames per genome. All the *Sal. ruber* draft genomes had one 16S rRNA gene copy of 1 535 bps in length.

Genomic diversity of the local *Sal. ruber* population based on 112 isolates

ANI vs. shared genome fraction analysis among all 112 genomes revealed a second closely related yet distinct population (n=10) around 95% ANI to the primary *Sal. ruber* population (n=102) (Fig. 1A). While this secondary cluster appears to be a divergent *Sal. ruber*-like clade based on our ANI analysis and maximum likelihood trees from 16S rRNA gene and single copy protein-coding genes (SCGs) (Figs. 1A & S3), it was initially identified as *Sal. ruber* using MALDI-TOF MS and RAPD analyses due to the highly similar spectra that made the strains of the new lineage to appear scattered among the true *Sal. ruber* (Fig. S2). Genomes of each population cluster share greater than 97.5% ANI among themselves with roughly a 3% ANI discontinuity and 5% difference in the shared genome fraction between them (Fig. 1A). We repeated this analysis using unassembled reads mapped to the assembled genomes in order to sidestep any potential biases resulting from assembly vs. assembly comparison of draft (incomplete) genomes containing hundreds of contigs and found essentially the same results (Fig. S4). We focused the remaining analyses on the primary *Sal. ruber* population due to its larger number of isolate genomes and the fact that the ten *Sal. ruber* genomes from the NCBI database fell within this primary population (Fig. 1A). Since the genomes from NCBI were isolated from various sites across the globe and are representative of the broader species level diversity, this finding (e.g., Fig. 1A) suggests that our results from the locally sampled population of isolates in our collection may be transferable to other salterns where *Sal. ruber* represents a dominant species. The ANI values among the members of the primary population averaged approximately 98% and the shared genome fraction averaged approximately 80%, revealing that, while these genomes share high sequence identity, about 20% of the gene content differs in pairwise comparisons (Figs. 1A & S4). These findings revealed substantial intraspecific sequence (e.g., ANI) variation within the local population equivalent to that of the broader *Sal. ruber* species population based on the 10 available genomes from NCBI or other species with several sequenced representatives (14).

A maximum likelihood phylogeny of the full length 16S rRNA gene sequences carried by the isolate genomes or their concatenated set of 106 SCGs confirmed the ANI-based results. The two phylogenies showed that the minor *Sal. ruber*-like population formed a single diverging clade and that the ten NCBI genomes were dispersed throughout the primary clade (Fig. S3). In addition, the lack of defined subclades within the primary clade in terms of individual solar ponds (spatial) and time of sampling over the one-month sampling period (temporal) from which the genomes originated suggested that the local *Sal. ruber* population was largely homogeneous over space and time (Fig. S3). Likewise, the interspersed placement of the complete NCBI genomes within this primary clade indicated that genomic diversity at the local scale is representative of the current sequenced diversity captured by complete genomes collected at a more global scale (Fig. S3), although it should be noted that it remains unclear how well the NCBI genomes capture (or not) the global *Sal. ruber* genomic diversity. Collectively, these results indicated that our draft genome collection represents the extant cultivatable genomic diversity found within the local *Sal. ruber* population and probably within the broader species population as well.

Pangenome structure of the local *Sal. ruber* population

To assess genomic diversity at the gene level, we randomly selected 100 genomes from the primary *Sal. ruber* population (n=102) and analyzed the pangenome structure by calculating the empirical, non-redundant gene rarefaction curve while tracking new gene additions and gene class counts. We estimated the total pangenome (i.e., the number of non-redundant genes) of the local *Sal. ruber* population to be open ($\gamma = 0.36$; the γ parameter reflects the slope of the curve that represents the total non-redundant genes (2, 3)) consisting of 12 666 genes in total (Fig. 2A upper panel; Sup. Excel File 2). Each additional *Sal. ruber* genome added 98 new genes to the pangenome, on average, with a persistent mean addition of 48 new genes at n = 100 after 1 000 permutations of the order that genomes were added to the rarefaction curve (Fig. 2A, lower panel). The exponential decay model fit to these data estimated that the new gene ratio (number of new genes per genome added to the pangenome / number of genes in genome) reached a lower asymptotic value of $\Omega = 2.2\%$ of genes per genome (the Ω parameter estimates the lower asymptote of the decay curve (2, 3)), although the empirical values were measured to extend below this with a mean value of 1.7% at n = 100. These data show that nearly 2% of the gene content in any *Sal. ruber* genome sampled by our collection consisted of unique genetic material and that because of this, the total gene content diversity remained under-sampled even after sampling 100 genomes collected from the same source water (Fig. 2A; Sup. Excel File 2). Accordingly, we calculated that the pangenome of the local *Sal. ruber* genome collection was composed of 4 830 (~38% of total) isolate-specific genes and 5 587 (~44%) rare or common genes distributed among the members of the population, with the core genes making up the remaining 18% (Fig. 2A; Sup. Excel File 2).

While the total accessory genome remained unsaturated by sequencing, we estimated that there were 2 249 universally shared genes comprising the *Sal. ruber* core genome, or about 78% of a *Sal. ruber* genome (2 249 / 2 888) represented conserved, shared genetic material (Fig. 2A). These results were congruent with the ANI vs. shared genome fraction analysis as well (Fig. 1A). The persistence of the core genome was also validated by the extremely narrow empirical confidence intervals and by the exponential decay model reaching a similar lower asymptote limit at Ω = 2 248 after only 30 genome additions (Fig. 2A; Sup. Excel File 2). A cladogram based on the presence/absence of accessory genes revealed a similar overall clade structure to the core gene phylogeny, although several genomes clustered differently between the two trees (Figs. 3 & S5). Both trees indicated the existence of coexisting subpopulations within the *Sal. ruber* population based on the recovery of three -or more- distinct subclades (Figs. 3 & S5). Consistent with this view, analysis using evolutionary read placement of metagenomic reads to reference *Sal. ruber rpoB* gene sequences indicated the presence of subpopulations (or genotypes) that fluctuated in abundance relative to each other across the different sampling times (Fig. S6).

Pangenome structure compared to other model species populations

Using the same pipeline, with additional random trials to calculate empirical confidence intervals, we estimated the pangenome of several, phylogenetically and physiological diverse, model bacterial species including *Escherichia coli*, *Bacillus thuringiensis*, *Salmonella enterica*, *Mycobacterium tuberculosis*, and *Pseudomonas aeruginosa* (Figs. 1B-E, 2B-C, 4 & 5; Sup.

Excel File 2). For these comparisons, we chose genomes to emulate the ANI distribution observed within the primary *Sal. ruber* population (~98% ANI, on average) in order to avoid the known effect of higher gene content conservation among genomes with higher ANI (genetic) relatedness as noted previously (34). Our selection process generated intraspecific ANI distributions similar to the primary *Sal. ruber* population centered around ~98.5% ANI (Fig. 1). We found the pangenome of *E. coli* to be open (γ = 0.34 for draft genomes and γ = 0.31 for complete genomes) and similar in size to the primary *Sal. ruber* population pangenome (Figs. 2 & 4A-B; Sup. Excel File 2). Annotation results for the total pangenome were similar overall between *Sal. ruber* and *E. coli*, but, as expected for a well-studied model species, *E. coli* had fewer genes annotated as hypothetical or uncharacterized (Fig. S7; Sup. Excel File 3). Compared to *E. coli*, *Sal. ruber* had a comparable number of isolate-specific genes despite the smaller size of the *Sal. ruber* genome (2 888 vs. 4 118 genes per genome, on average). This was evident in a larger new gene ratio estimate for the *Sal. ruber* genomes, and in the number of isolate-specific (or rare) genes compared to the core genes (Figs. 4C & 5A-D), especially after normalizing by the pangenome or genome size for each species (Fig. 4D & 5E-L). Results for the other model bacterial species mentioned above are also reported but not discussed further to avoid redundancy.

Another pattern revealed by our pangenome analysis was a consistent ratio of core genome size to genome size (i.e., what fraction of the total genes in the genome the core genes make up) at about 0.8-0.9, observed across species despite the variation of core gene to pangenome size ratios (Fig. 5L vs. H). This result was consistent with earlier observations based on a much smaller number of genomes per species (n ~ 10) (17). In addition, there was a general lack of genes with intermediate prevalence in the pangenome of a species between the core and rare or isolate-specific gene classes (Figs. 4C, G, K & 6B). This means that genes are either predominantly present in all genomes or in only a few genomes of the population in accordance with the idea that new beneficial gene sweeps are comparatively less common than the creation and subsequent loss of new genetic material. In any case, we found the ratio of "accessory genes / total pangenome genes" and "accessory genes in the genome / total genes in the genome" to be greater for *Sal. ruber* compared to the other species considered, indicating that a larger portion of the *Sal. ruber* genome is allocated toward gene content variation (Fig. 5E & I).

Ecological/Functional importance of rare and isolate-specific *Sal. ruber* genes

To test if any isolate-specific or rare genes could provide an ecological benefit or were instead functionally neutral and/or ephemeral, we assessed their relative *in situ* abundance in the companion metagenomes representing changes in environmental conditions. To ensure adequate sequence coverage of the *Sal. ruber* population, metagenomes were re-sequenced to 15X coverage or greater for the *Sal. ruber* genomes based on their relative abundance (Fig. S8). To estimate the sequence depth of each gene (coverage), we computed a truncated average depth using the middle 80% of the sequence base positions (TAD80) to remove outlier effects from short conserved domains or motifs and the edge effect of read mapping to the ends of contigs (i.e. the top and bottom 10% of per base sequence depth values are removed from the distribution prior to taking the average), as suggested recently (35). We then normalized the TAD80 for each gene cluster by the average whole-genome TAD80 (Fig. S8) providing a view of the relative abundance of genes in relation to the relative genome abundance (Fig. 6A). Thus, a normalized value of 1.0 indicated that a gene has an equivalent *in situ* sequencing depth (relative abundance) to a single copy core gene; a value above 1.0 indicated that the gene abundance is greater than the genome average. The resulting data revealed a strong decreasing trend in the distribution of gene abundances from the core gene class to the isolate-specific gene class (Fig. 6A). Notably, we identified that 0.7% of the isolate-specific genes (i.e., present in only one genome in our collection) and 2.8% of the rare genes (i.e., present in <20% of the genome in our collection) became abundant *in situ* during the intermediate salinity conditions (diluted pond one-week sample; 23.6% salt concentration), which followed the dilution from high (~34% salts) to low (~12% salts) salinity (Fig. 7). The isolates found to possess these genes originated from all ponds and sampling times (data not shown). They were not specific to isolates from the diluted pond only. The *Sal. ruber* population abundance in our samples did not vary more than three-fold and metagenomes were re-sequenced to provide 15X *Sal. ruber* genome coverage or greater (Fig. S8); thus, the differential gene abundance reported above cannot be attributed to possible artifacts related to low sequence coverage of the population. Such patterns were not observed for the abundance of any rare or isolate-specific genes in the light intensity or control treatment (Fig. 7C, F, & I).

Functional annotation of the abundant fraction of isolate-specific and rare genes from the intermediate-salinity metagenome revealed that several of these genes could be involved in response to osmolarity changes, gene regulation, and transport of metabolites in/out of the cell. (Fig. 7H; Sup. Excel File 4). The isolate-specific genes that peaked in abundance during intermediate salinity shared high sequence similarity to genes found on the pSR84 plasmid from *Sal. ruber* strain M8, isolated more than a decade ago and shown to be more tolerant of lower salinity conditions than the type strain of *Sal. ruber* (Strain DMS 13855 or M31) (36). These results contrasted with an over-dominance of hypothetical and mobile functions among the isolate-specific genes that did not change in abundance in the intermediate-salinity metagenomes

(Figs. 7 & S9; Sup. Excel 4), revealing a strong bias toward functions that are presumably related to the salinity perturbation. In addition, 10.5% of the core genes also showed increased relative *in situ* abundances at the intermediate salinity metagenome and the (predicted) functions encoded by these core genes (Fig. 7B) were similarly involved in osmoregulation and transport as the rare (Fig. 7E) and isolate-specific (Fig. 7H) genes mentioned above. Hence, the enrichment of (or selection for) specific functions during salinity transition was evident in different parts of the population's pangenome, and we focused our analysis on isolate-specific and rare genes because these made up a larger fraction of the pangenome (Fig. 6B). Given that core genes are typically single-copy genes in the genome, the increased abundances noted (Fig. 7B) could be due to recent horizontal transfer of these genes to/from other co-occurring populations or duplication of the genes within the genome (e.g. they are carried by multi-copy plasmids). Future work will elucidate the relevance of each of these scenarios.

We investigated if additional community members may also harbor the accessory genes found to fluctuate in relative abundance to determine if the genes are broadly important to the community (as opposed to just *Sal. ruber*) or if they may be horizontally transferred between community members. To evaluate the diversity of the genomic background (origin) of the *Sal. ruber* rare or isolate-specific genes that became abundant under the low-salinity condition, we assembled our metagenomic samples and searched for these genes (See Sup. Excel File 5 for metagenome assembly details). We found a variety of contigs containing similar, but not identically, copies of these genes within the 70 – 100% sequence identity range (Figs. S10 & S11). Genes of the rare class (Fig. S10) had more distant matches than genes in the isolate-specific class (Fig. S11). These results revealed that additional community members encode the genes, especially in the dilution pond one-week time point sample (salinity 23.6% NaCl), indicating that the corresponding functions may indeed be selected by intermediate salinities and that genes with functions making their way to the rare class may be more broadly useful to the community as well. We also looked at synteny and taxonomic classification of the contigs identified, and while some contigs showed possible gene synteny indicating a common origin for the corresponding genes (Fig. S12), they were generally too short for conclusive results or taxonomic placement.

Discussion

The pangenome of a local *Sal. ruber* population, sampled over a one-month period from four adjacent saltern ponds filled with the same source water, is open and similar in size to the pangenomes of *E. coli* and other species whose genomes were recovered from around the globe over the course of many years (Figs. 2 & 4). These results were somewhat unexpected given the range of samples providing the *E. coli* (and other) genomes relative to the few samples that provided the *Sal. ruber* isolates. In fact, we found that the pangenome to genome size ratio of the local *Sal. ruber* population is the largest of all species considered (Fig. 4B; Sup. Excel File 2). While pangenome sizes are known to vary between species (16, 18, 37), it is intriguing to find such extensive gene content variation within a single population and local source. These results also corroborate observations from a previous study of another large *Sal. ruber* isolate collection retrieved from a single, one drop (0.1 ml) sample that showed *Sal. ruber* to be phylogenetically homogenous at the ribosomal level yet diverse based on restriction patterns and metabolomics analysis (38). Furthermore, the *Sal. ruber* pangenome consisted primarily of core or very rare genes; few genes were found at intermediate prevalence, that is, encoded by a substantial fraction (e.g., ranging from 10% to 90% of the total) of genomes in our collection (Figs. 5 & 6B). These results imply that the majority of non-core genes may indeed be neutral and/or ephemeral as previously hypothesized and do not contribute to the major functions carried by the population. Horizontal gene transfer (HGT) and gene deletion presumably underlie these patterns. Consistent with this assumption, our comparison of the *Sal. ruber* SCGs phylogenetic tree to that of the dendrogram based on the presence or absence of gene content revealed several incongruences (Fig. 3), indicating that HGT (and gene deletion) may be common.

While the great majority of isolate-specific and rare genes remained rare as conditions changed, at least a few of them (about 3.5% of the total) were found to considerably increase in abundance during the low-salinity (dilution) perturbation (Fig. 7). Several of the latter genes were three times more abundant relative to the genome average at intermediate salinity. Notably, the predicted functions encoded by these genes were associated with environmental sensing, metabolite transport in/out of the cell, glycosyltransferases (which may be related to osmoregulation) (39), and gene regulation (Figs. 7, S10, S11 & S12; Sup. Excel File 4). Very few hypothetical or uncharacterized genes were identified among the genes showing increased abundance at intermediate salinity despite the high frequency of the former genes in the total pangenome (Fig. S9). Collectively, these results further supported our hypothesis that the identified genes involved in regulation and transport are presumably important for cell osmoregulation under low- and intermediate-salinity conditions. While these findings await further experimental validation (e.g., measure the fitness effect of the genes), they do indicate that a small fraction (3.5%) of genes identified from the isolate-specific and rare class may facilitate the *Sal. ruber* population in adapting to changes in salinity concentrations, and thus are likely not neutral or ephemeral. Unfortunately, the exact function or substrate specificity for the identified genes remains unknown as bioinformatics analysis provides only general functional prediction. Hence, we are not yet able to make specific inferences about how exactly these genes facilitate adaptation to intermediate salinity conditions. Note also that assessing the relative abundance of these genes in the time zero sample of lowest salinity (~12% salts) would not have been meaningful in this respect because the corresponding cells that carry these genes did not have enough time to adjust to the low salinity conditions and begin to grow (40, 41). Further, the great majority of the identified genes apparently do not undergo adaptive evolution since their pN/pS ratio based on metagenomic reads mapped on the genes were low, between 0.1-0.5, although slightly higher than that of the core genes (Fig. S13). These ratios indicated strong purifying selection for the identified osmoregulation-related genes and that they are already well fit for the function they carry out.

It is possible that additional isolate-specific or rare genes become important during fluctuations of other environmental variables. However, analysis of metagenomic data from the experimental manipulation of light intensity, the other major environmental factor for the saltern ecosystems (31), did not reveal any isolate-specific genes to change in abundance as we observed in the intermediate salinity samples. Notably, while *Sal. ruber* has been shown to be a heterotroph and strict aerobe, with limited substrate repertoire (40, 41), it also carries various rhodopsins, i.e. xanthorhodopsins, halorhodopsins and sensory rhodopsins (42, 43) that allow it to gain sunlight energy. Hence, the lack of rare or isolate-specific genes becoming abundant during the light intensity manipulation presumably reflects that the majority of light-specific functions are found among the core genes. Accordingly, any additional isolate-specific or rare genes of ecological importance would have to be specific to environmental parameters not measured by our work. Such parameters could include seasonal fluctuations (e.g., we sampled for one-month in August) or biotic factors such as phage predation (44). In any case, genes of such ecological importance are not expected to make up a large fraction of the pangenome based on the results reported here (e.g., most isolate-specific genes were not found to be abundant relative to the average genome relative abundance in any metagenomic datasets) for salinity and light intensity transitions, the two major drivers of the saltern ecosystem. While some of these results and interpretations echoed those in previous studies (17, 45-47), they do provide a new and more quantitative perspective into the role of biodiversity within sequence-discrete populations (and species) during environmental transition.

An emerging question based on these findings is why the intrapopulation diversity was not purged (removed) when salinity conditions changed. That is, the genomes (cells) that encode the abovementioned genes should have outcompeted the remaining genomes of the population resulting in a more clonal population and/or (sub-population) speciation. However, phylogenetic analysis of the isolates (Fig. S3) and metagenomic read placement (Fig. S6) suggests that intrapopulation diversity was maintained and, in fact, the dominant subpopulations in terms of gene content bounced back in abundance when salt-saturation conditions were re-established. Thus, we hypothesize that the ecological advantage of these genes is significant, but not strong enough to purge the intrapopulation diversity (or sweep through the population) or a much longer duration of intermediate salinities than represented by our experimental design or the typical salt cycles observed in the Mallorca salterns would have been required for a population sweep event to take place. Consistent with this working hypothesis, genomes that do not encode these genes were apparently able to survive at lower growth rates until favorable (salt-saturation) conditions returned (Figs. 7 & S6). Further, the generation time previously observed for the natural *Sal. ruber* population in similar saltern ponds is close to 30 hours (40, 41), which reveals relative slow growth compared to organisms like *E. coli* (20-30 minutes generation time) and thus, a long time is presumably required for population (and/or gene) sweep events to take place. While this hypothesis remains to be experimentally tested, it does provide a plausible explanation for the maintenance of sequence-discrete populations despite such immense intrapopulation gene content diversity, frequent HGT [Fig. 3 and Refs (12, 15, 47, 48)], and environmental transitions. That is, transient environment fluctuations could select for a subset of cells carrying specific accessory genes but the selective advantage conferred by these genes is not strong enough for the corresponding cells to outcompete the remaining co-occurring cells and dominate the population (diversity purging) within the time that the environmental fluctuations last and the strength of selection that the fluctuations impose on the population. Moreover, these results further corroborate the use of sequence-discrete populations as the important unit of microbial diversity for taxonomy as well as for future investigations for advancing taxonomy and diversity studies.

While isolate sequencing can circumvent the limitations of MAGs and short-read data in recovering the intra-population gene content diversity [discussed above and in (21, 22, 49)], isolation may provide an uneven view of the natural population due to isolation biases. However, this is unlikely to have been the case for the primary *Sal. ruber* population studied here based on several independent lines of evidence. First, the ten *Sal. ruber* genomes from NCBI, which originated from various salterns around the world, are grouped together into a single clade with the 102 genomes of the primary population in our core-genome phylogeny. This is also the case for the evolutionary placement of metagenomic reads onto the genome-based phylogeny, i.e., the great majority of reads were assigned to terminal branches (tips) of the tree as opposed to ancestral nodes; the latter would have indicated that the reads originated from abundant strains not represented by our isolates. Second, our collection of sequenced isolates represents a much larger collection (n = 207) of isolates that were first identified as *Sal. ruber* by MALDI-TOF MS whole cell analysis and dereplicated by RAPD profile analyses to avoid sequencing the same clone. Importantly, in this larger collection, there were not any major or minor subclade(s) that were not represented by the subset of the 112 isolates we sequenced (Fig. S2A). Hence, our genome collection appears to be representative of the natural *Sal. ruber* population based on the phylogenetic placement of metagenomic reads, the MALDI-TOF spectra analysis, and the RAPD profiles of a much larger collection of (non-sequenced) isolates (Figs. S6 & S2). Finally, based on the ANI vs. the shared genome fraction analysis (Fig. 1), the draft *E. coli* genome collection had a wider distribution than the complete *E. coli* genome collection; however, the mean values on both axes were nearly equal (Fig. 1B & C). The wider distribution could be a technical artifact due to the nature of incomplete versus complete genome sequences, or likely a true signal of biological diversity arising from the greater number of draft genomes available. Regardless of the exact underlying reason, the similarity in values indicates that our measurements from draft genomes provide similar estimates to complete genomes and that our pipeline was robust. Importantly, we selected draft genomes from different species to be of a similar level of completeness and ANI relatedness to avoid any systematic effect of these parameters on our results and conclusions.

Despite the abovementioned advantages, the number of genomes we sequenced was still limiting compared to the total gene content diversity as evidenced by the incomplete capture of accessory genes by our pangenome analysis. Future work could include more replicate samples, longer time series analysis, and deeper metagenomic sequencing with long-read technology for more robust results and interpretations. It would be particularly interesting to measure the fitness advantage of isolates based on their specific complement of accessory genes to directly test the hypotheses presented above related to the ecological advantage of such genes. The work presented here should serve as a guide for the number of samples and isolates to obtain, amount of sequencing to apply (Fig. S8), and what bioinformatics analyses to perform for studying the value of the diversity within natural sequence-discrete populations. It is also important to note that other species in different habitats may show different dynamics compared to those revealed here for the *Sal. ruber* population, especially with respect to what fraction of the pangenome responds to environmental change, because the perturbation (e.g., salinity change) may not translate well to the type of perturbations observed in these other habitats. Nonetheless, the striking similarities revealed between the *Sal. ruber* and *E. coli* pangenomes do indicate that, *Sal. ruber* is a useful model system for studying pangenome dynamics and the ecological/functional importance of accessory genes.

Materials and Methods

Experimental site, sampling and processing

Sal. ruber isolates and whole-community samples for shotgun-metagenomics were collected concomitantly from four adjacent solar saltern ponds in Mallorca, Spain at 'Es Trenc' (Fig. S1) at three time points over a one-month period post-treatment in August of 2012. The saltern ponds are part of a larger facility of crystallizers for salt harvest and are fed with the same inlet brines. After filling with inlet brines there was no fluid exchange between the ponds during the course of study, and there were negligible effects from rainfall in Majorca, Spain during the month of August, when the experiment was conducted. The conditions for each pond were as follows: 1) a control pond with ambient sunlight and salt-saturation conditions found at the 'Es Trenc' salterns, 2) a shaded-unshaded pond that was covered with a mesh to reduce sunlight intensity by 37-fold for 3 months prior to the experiment and then uncovered (unshaded) at time zero, 3) an unshaded-shaded pond that was kept uncovered until time zero, then covered with a mesh (shaded), and 4) a diluted pond to which freshwater was added to reduce the salinity from ~34% to ~12% in less than one hour. The mesocosm experiment was designed to test the effects of light intensity and salinity levels on the indigenous microbial communities inhabiting the salterns. Ponds #2 and #4 reached salt saturation conditions after one month during which time the microbial community dynamics re-stabilized; thus, no further sampling was performed after one month. The isolates used in this study were collected from all four ponds at time zero (just before the stressors were applied) and one-month (at the end of the experiment). Isolates were also sequenced from samples taken at two days post dilution treatment and one-week post the unshaded-shaded treatment (Fig. S1). The companion metagenomes for this study were sequenced from all four ponds sampled at one day, one week and one-month post treatment (Fig. S1). Metagenomes for the shaded-unshaded experiment were excluded in this study due to low sequence coverage of the *Sal. ruber* population in the corresponding samples. The experimental design and detailed procedures were previously described in (31), except for the inclusion of the diluted pond experiment which is described in (33) and outlined in Fig. S1. Samples were collected, processed for culturing, and resulting isolates were identified using Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry (MALDI–TOF MS) as described by (31, 32, 50) (Fig. S2A). Multiple clonal isolates were dereplicated using random amplified polymorphic DNA (RAPD) fingerprinting (50) (Fig. S2B).

DNA extraction and Sequencing

Sal. ruber isolate cultivation and DNA extraction were performed as described in (42, 51). For metagenomic DNA extraction, 25ml of brine samples were centrifuged at 13 000 rpm as detailed in (51). DNA sequencing libraries were prepared using the Nextera XT DNA library prep kit (Illumina) according to manufacturer's instructions up to the isolation of cleaned double stranded libraries. Library concentrations were determined by fluorescent quantification using a Qubit HS DNA kit and Qubit 2.0 fluorometer (ThermoFisher Scientific) and samples were run on a High Sensitivity DNA chip using the Bioanalyzer 2100 instrument (Agilent) to determine library insert sizes. Libraries were sequenced for 500-cycles (2 x 250-bp paired-end) on a MiSeq instrument (Illumina; Molecular Evolution Core facility, Georgia Institute of Technology) as recommended by the manufacturer. Additional sequencing of selected low-coverage libraries after the MiSeq sequencing was carried out on NextSeq 500 instrument (Illumina; located in the same facility) using a rapid run of 300 cycles (2 x 150-bp paired-end). Adapter trimming and demultiplexing of sequenced samples was carried out by the software available on each respective sequencing instrument.

Sequence quality control, assembly, and gene prediction

Raw reads in fastq format were evaluated with FastQC version 0.11.2 (52) in addition to quality analysis using custom Python scripts. Trimming and adapter clipping were performed using Trimmomatic version 0.39 (53) with settings ILLUMINACLIP NexteraPE-PE.fa:2:30:10:2:keepBothReads LEADING:3 TRAILING:3 MINLEN:36. Assembly was performed using SPAdes version 3.13.0 with the "--careful" flag and "-k 21,33,55,77,99,127". Gene prediction was performed using Prodigal version 2.6.3 with default settings (54). The resulting summary tables can be found in Supplementary Excel File 1.

Assessment of draft genome quality and phylogenetic analyses

The *Sal. ruber* isolate draft genomes were assembled from an average of 158 Mbps (stdev=66) of sequenced reads per isolate after adapter clipping and quality trimming (Supplemental Excel File 1). For each assembly, contigs shorter than 1 000 base pairs with supporting sequence coverage of less than 2X were removed from the assembly. The draft genomes in addition to *Sal. ruber* and *Sal. altiplanensis* genomes retrieved from NCBI were evaluated using the Microbial Genomes Atlas (MiGA) (55) to generate assembly metrics, quality scores, and all vs. all ANI scores (Supplemental Excel File 1). MiGA also identifies and extracts predicted 16S rRNA gene sequences and universal SCGs for each genome submitted. The sequences for the 16S rRNA gene, *rpoB* and the concatenated set of SCGs from each genome were aligned using Clustal Omega version 1.2.1 (56) with default settings. Maximum likelihood trees for the 16S rRNA gene and *rpoB* alignments were generated using RAxML version 8.0.19 (57) with parameters: *-m GTRGAMMA -f a -N autoMRE -p 4564821 -T 2 -x 1235*. An approximate maximum likelihood tree was generated for the concatenated SGCs using FastTree v2.1.10 (58) with default settings. The trees were drawn using either FigTree v1.4.3 (59) or iTOL version 4 (60).

Pangenome analysis

A custom pipeline was developed for the pangenome analysis using a combination of Bash and Python programming. The pipeline starts with a directory containing genomes for single species and proceeds in five parts. Part 1 selects a seed genome at random and then continues random selection of genomes without replacement until the requested number of genomes meeting the criteria is reached. The pipeline keeps a genome if it matches the seed genome above a user defined ANI value (97.5% by default to match the ANI values observed among *Sal. ruber* isolate genomes; Fig. 1). Part 2 predicts genes for each genome selected using Prodigal and removes genes shorter than 300 nucleotides in length. Part 3 runs an all vs. all ANI genome comparison using FastANI version 1.1 (14). Part 4 clusters all genes from all genomes using CD-HIT-EST version 4.7 (61) with parameters: *-c 0.9 -n 8 -G 0 -g 1 -aS 0.7 -M 10000 -d 0 -T 10*. Step 5 uses custom Python scripts to parse the CD-HIT cluster file, run permutations to calculate the pangenome statistics, fit models to the empirical data, and build graphical plots of the results. The genomes used in our analysis included 878 complete and 11,167 draft *E. coli*, 433 draft *Bacillus thuringiensis*, 3,037 draft *Salmonella enterica*, 1 865 draft *Mycobacterium tuberculosis*, and 3 264 draft *Pseudomonas aeruginosa* genomes downloaded from NCBI on June 26, 2019. To generate empirical distributions, the custom pangenome pipeline was used to run 100 random bootstrap trials for each species (other than *Sal. ruber*) with genome replacement between trials. Only one trial was run for *Sal. ruber* since only about 100 draft genomes were available and each trial used 100 genomes.

In accordance with Tettelin et. al. 2005 & 2008 (2, 3), we fit a power law model to our data to estimate the upward trajectory of the pangenome growth curve using the γ parameter, and, we fit an exponential decay model to our data to estimate the Ω parameter which represents the lower boundary of the curves for the number of core genes or new gene additions. The possible values for the γ parameter reflect an open ($0 < \gamma < 1$) or closed ($\gamma < 0$) pangenome. For this analysis, we chose 100 genomes at random from the primary *Sal. ruber* population which excluded the two draft genomes labelled as SZ05 and SM11 corresponding to the 5th and 11th isolates obtained from the shaded pond at time zero and one month respectively. We then defined gene classes by a parameter $p = \mathrm{n} / \mathrm{N}$ where n is the number of genomes carrying a gene and N is the total number of genomes (N = 100). Core genes are defined as those showing $p \geq 0.9$, common when $0.2 \leq p < 0.9$, rare when $1 / N < p < 0.2$, or isolate-specific when $p = 1 / N$. Accordingly, the accessory pangenome consists of all isolate-specific, rare, and common genes.

Estimation of *in situ* gene abundance

Quality-trimmed metagenomic reads were searched against *Sal. ruber* draft genomes separately for each genome using the *blastn* option for the "task" parameter with BLAST+ version 2.2.29 with default settings (62). Reads that found a match higher than the 95% sequence identity threshold and an (alignment length) / (query read length) greater than a threshold of 0.9 were used to calculate sequence depth (relative abundance). The resulting read depth data was truncated to the middle 80% (TAD80) of depth values (i.e., the upper and lower 10% of outliers were removed) using a custom Python script to provide TAD80 values for each genome, contig, and gene (Figs. 6, 7, & S8; Supplemental Excel File 4). The 2nd pond (shaded-unshaded) had relatively lower *Sal. ruber* abundance compared to the other ponds presumably due to the long-term shading of the pond, and was sequenced at a lower effort, which rendered the assessment of gene *in situ* abundance unreliable. Hence, this pond was used for isolation but not in the remaining bioinformatics analysis.

Gene annotations

Representative genes for each CD-HIT-EST gene cluster were annotated against both UniProt databases (SwissProt and TrEMBL release-2018_05) using the *blastp* algorithm from BLAST+ version 2.2.29 with default settings (62). Results were filtered for best match using a minimum threshold of 40% sequence identity and 50% alignment length coverage of the UniProt sequence for a match. Genes were also annotated using KofamKOALA version 2019-07-03 (63) with KEGG release 91.0 (64) and only annotations with an asterisk indicating they were above predefined thresholds for the corresponding HMM models were kept for analysis. Annotations can be found in supplemental Excel file 3 and 4.

Code Availability

The custom code for these analyses are available on GitHub: https://github.com/rotheconrad/Salinibacter_ruber_01.

Data availability

The data for this study has been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB27680 (https://www.ebi.ac.uk/ena/data/view/PRJEB27680).

Acknowledgements

The authors would particularly like to thank especially the whole team at Salinas d'Es Trenc and Gusto Mundial Balearides, S.L. (Flor de Sal d´Es Trenc) for allowing access to their facilities and for their support in performing the experiments. This work was partly funded by the US National Science Foundation, awards #1831582 and #1759831 (to KTK), and by the projects CLG2015_66686-C3-1-P, PGC2018-096956-B-C41 and RTC-2017-6405-1 of the Spanish Ministry of Science, Innovation and Universities (to RRM), which were also supported with European Regional Development Fund (FEDER) funds. RRM acknowledges the financial support of the sabbatical stay at Georgia Tech by the grant PRX18/00048 also from the Spanish Ministry of Science, Innovation and Universities. We thank Miguel Rodriguez-R and Carlos Ruiz for useful discussions on the methodology applied and ecologic implications of our results as well as the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, which enabled the computational tasks associated with this study.

Conflict of Interest

The authors declare no conflicts of interest.

Figures

Figure 1:

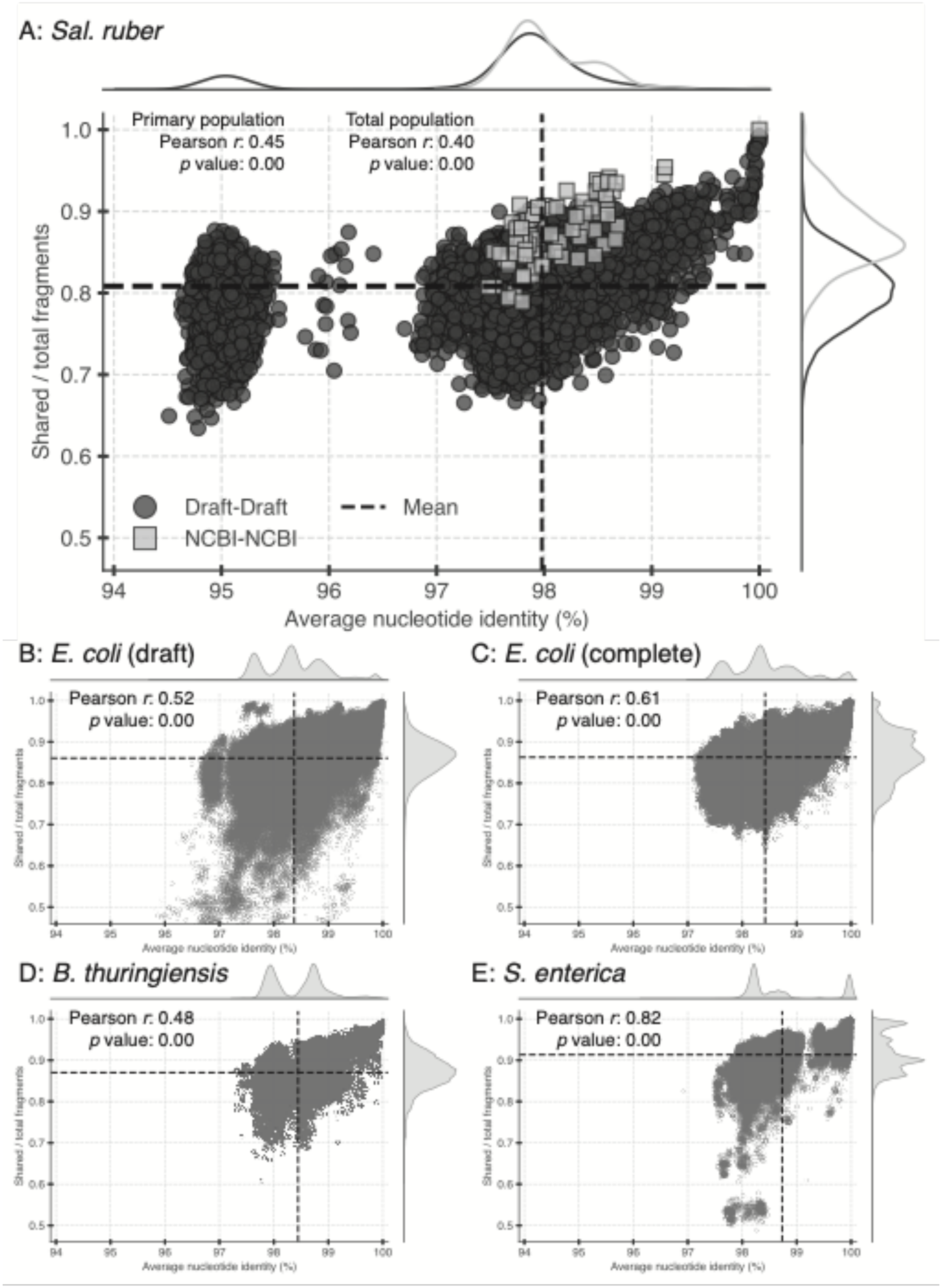
A: *Sal. ruber*
Primary population
Pearson *r*: 0.45
p value: 0.00
Total population
Pearson *r*: 0.40
p value: 0.00
Shared / total fragments
Average nucleotide identity (%)
Draft-Draft
NCBI-NCBI
Mean
1.0
0.9
0.8
0.7
0.6
0.5
94
95
96
97
98
99
100
B: *E. coli* (draft)
Pearson *r*: 0.52
p value: 0.00
C: *E. coli* (complete)
Pearson *r*: 0.61
p value: 0.00
D: *B. thuringiensis*
Pearson *r*: 0.48
p value: 0.00
E: *S. enterica*
Pearson *r*: 0.82
p value: 0.00
Shared / total fragments
Average nucleotide identity (%)

Fig. 1 Genomic diversity within selected species assessed by ANI and shared genome fraction. The shared genome fraction (y-axis) is plotted against ANI (x-axis). Correlation coefficients between these two genome variables are shown on each plot along with dashed lines for the mean values. The graphs to the top and right of each panel show the kernel density estimates for each axis. (**A**) Values representing the 112 x 112 comparisons of our *Sal. ruber* isolate draft genome collection (designated as Draft-Draft), 10 x 10 for *Sal. ruber* genomes from NBI (designated as NCBI-NCBI), and 112 x 10 for our draft genomes versus the NCBI genomes (designated as Draft-NCBI). Correlation coefficients were calculated for the primary *Sal. ruber* population (n=102) as well as for the total population (n=112). The means were calculated for the primary *Sal. ruber* population only. Note that the draft datapoints overlap with the NCBI datapoints representing complete genomes revealing no major biases by the draft nature of these genomes with respect to gene content or ANI values. Similarly, this shows that our isolate collection captures the diversity seen within the available reference genomes. (**B-E**) 100 x 100 genomic comparison results times 100 random sampling trials for (**B**) *E. coli* draft, (**C**) *E. coli* complete, (**D**) *B. thuringiensis* draft, and (**E**) *S. enterica* draft genome collections. Correlation coefficients and means were calculated for all values. Note the kernel density estimates on the perimeter of each panel (top and right) that clearly show the density distributions of the corresponding data points.

Figure 2:

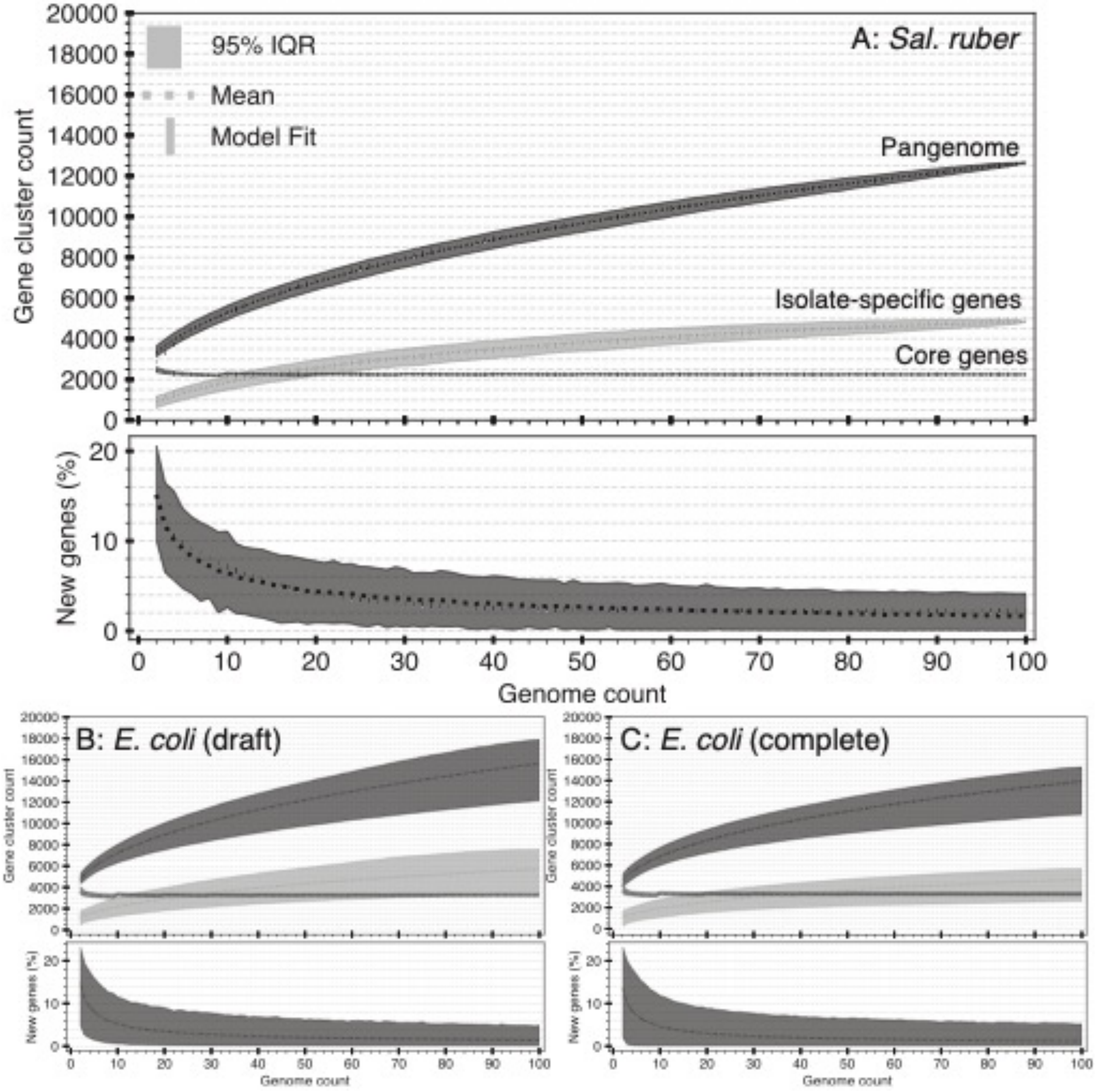

Fig. 2 Pangenome comparison of *Sal. ruber* to selected bacterial species. (**A-C**) The top panel shows the mean, 95% empirical confidence interval of permuted values, and the model fit for each of three curves showing the total non-redundant genes in the pangenome (dark gray), total number of core genes (gray), and total number of isolate-specific genes (light gray) on the y-axis plotted against the number of genomes sampled (x-axis) for (**A**) 100 draft genomes from the

primary *Sal. ruber* population, and results from all random trials for (**B**) *E. coli* draft genomes and (**C**) *E. coli* complete genomes. The lower panel shows the same calculations but for the new gene per genome ratio. The axes scales are conserved between panels. Pangenome metrics are also provided in Sup. Excel File 2.

Figure 3:

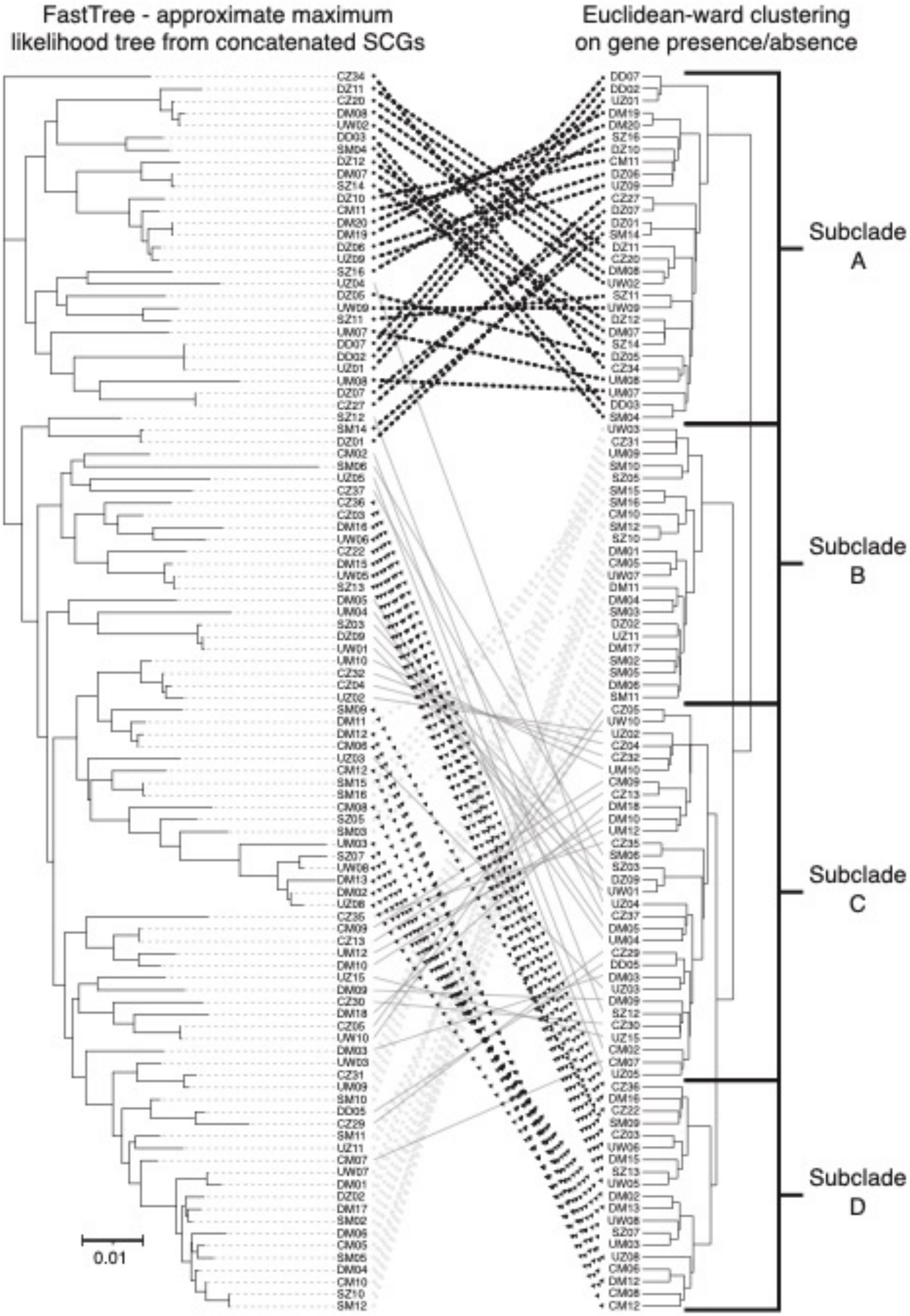
FastTree - approximate maximum likelihood tree from concatenated SCGs
Euclidean-ward clustering on gene presence/absence
CZ34
DZ11
CZ20
DM08
UW02
DD03
SM04
DZ12
DM07
SZ14
DZ10
CM11
DM20
DM19
DZ06
UZ09
SZ16
UZ04
DZ05
UW09
SZ11
UM07
DD07
DD02
UZ01
UM08
DZ07
CZ27
SZ12
SM14
DZ01
CM02
SM06
UZ05
CZ37
CZ36
CZ03
DM16
UW06
CZ22
DM15
UW05
SZ13
DM05
UM04
SZ03
DZ09
UW01
UM10
CZ32
CZ04
UZ02
SM09
DM11
DM12
CM06
UZ03
CM12
SM15
SM16
CM08
SZ05
SM03
UM03
SZ07
UW08
DM13
DM02
UZ08
CZ35
CM09
CZ13
UM12
DM10
UZ15
DM09
CZ30
DM18
CZ05
UW10
DM03
UW03
CZ31
UM09
SM10
DD05
CZ29
SM11
UZ11
CM07
UW07
DM01
DZ02
DM17
SM02
DM06
CM05
SM05
DM04
CM10
SZ10
SM12
0.01
DD07
DD02
UZ01
DM19
DM20
SZ16
DZ10
CM11
DZ06
UZ09
CZ27
DZ07
DZ01
SM14
DZ11
CZ20
DM08
UW02
SZ11
UW09
DZ12
DM07
SZ14
DZ05
CZ34
UM08
UM07
DD03
SM04
UW03
CZ31
UM09
SM10
SZ05
SM15
SM16
CM10
SM12
SZ10
DM01
CM05
UW07
DM11
DM04
SM03
DZ02
UZ11
DM17
SM02
SM05
DM06
SM11
CZ05
UW10
UZ02
CZ04
CZ32
UM10
CM09
CZ13
DM18
DM10
UM12
CZ35
SM06
SZ03
DZ09
UW01
UZ04
CZ37
DM05
UM04
CZ29
DD05
DM03
UZ03
DM09
SZ12
CZ30
UZ15
CM02
CM07
UZ05
CZ36
DM16
CZ22
SM09
CZ03
UW06
DM15
SZ13
UW05
DM02
DM13
UW08
SZ07
UM03
UZ08
CM06
DM12
CM08
CM12
Subclade A
Subclade B
Subclade C
Subclade D

Fig. 3 Clustering of *Sal. ruber* isolate genomes based on presence or absence of accessory genes reveals similar but divergent subpopulation structure compared to 16S or concatenated SCG gene tree. Comparison of the approximately-maximum-likelihood phylogenetic tree from SCGs to a dendrogram based on the presence or absence of accessory genes. Genome clustering is based on the pairwise Euclidean distance calculated from gene presence or absence using the *scipy.spatial.distance.pdistfrom* package and the Ward variance minimization algorithm from the *scipy.cluster.hierarchy.linkage* package in Python. The concatenated alignment of SCGs was stringently trimmed to remove all gap positions and contiguous non-conserved positions of 5 or greater. The lines connect the same gene across both trees. Isolates are named according to the different ponds and sampling times they were recovered from corresponding to the abbreviations in Fig. S1. So, CZ22 designates the 22nd isolate recovered from the control pond (C) at time zero (Z) and DM15 designates the 15th isolate recovered from the diluted pond (D) at time one-month (M). Note the plausible highlighted sub-clade structures consist of isolates from all ponds and sampling times.

Figure 4:

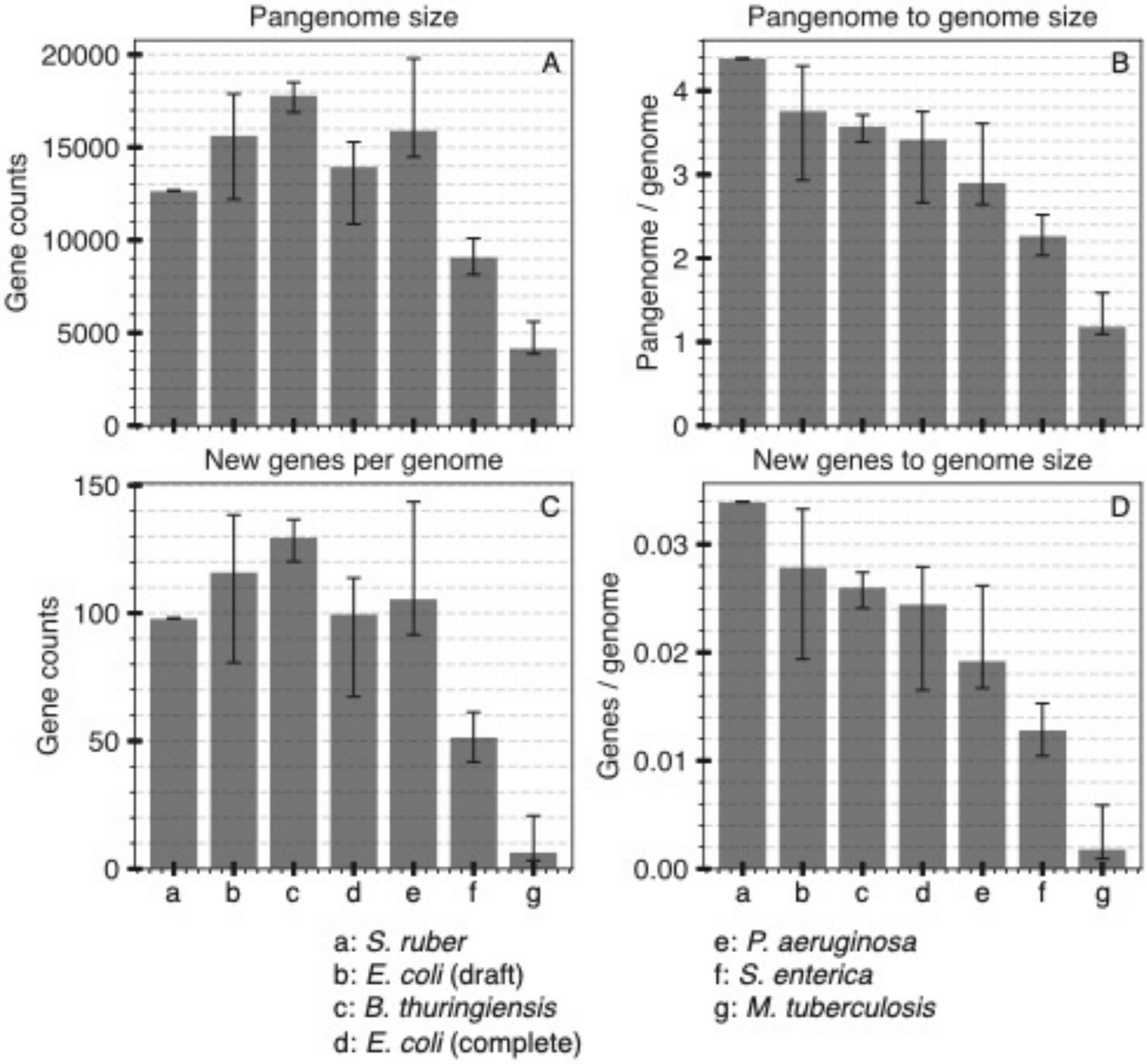

Fig. 4 Detailed comparison of pangenome metrics for selected bacterial species. Pangenome metrics were calculated from draft genome collections for multiple organisms, plus one closed genome collection for *E. coli*. Error bars show the 95% empirical confidence interval of results calculated from 100 random trials each selecting 100 genomes and running 100 permutations. *Sal. ruber* error bars are not shown because only 100 draft genomes from this experiment were available (no random trials). (**A**) Absolute count of total non-redundant genes after the addition

of 100 randomly selected genomes. (**B**) Values from (**A**) normalized by the average genome size of each species. (**C**) Mean number of new genes added to the total pangenome per genome addition. (**D**) Values from (**C**) normalized by the average genome size of each species. Pangenome metrics also provided in Sup. Excel File 2.

Figure 5:

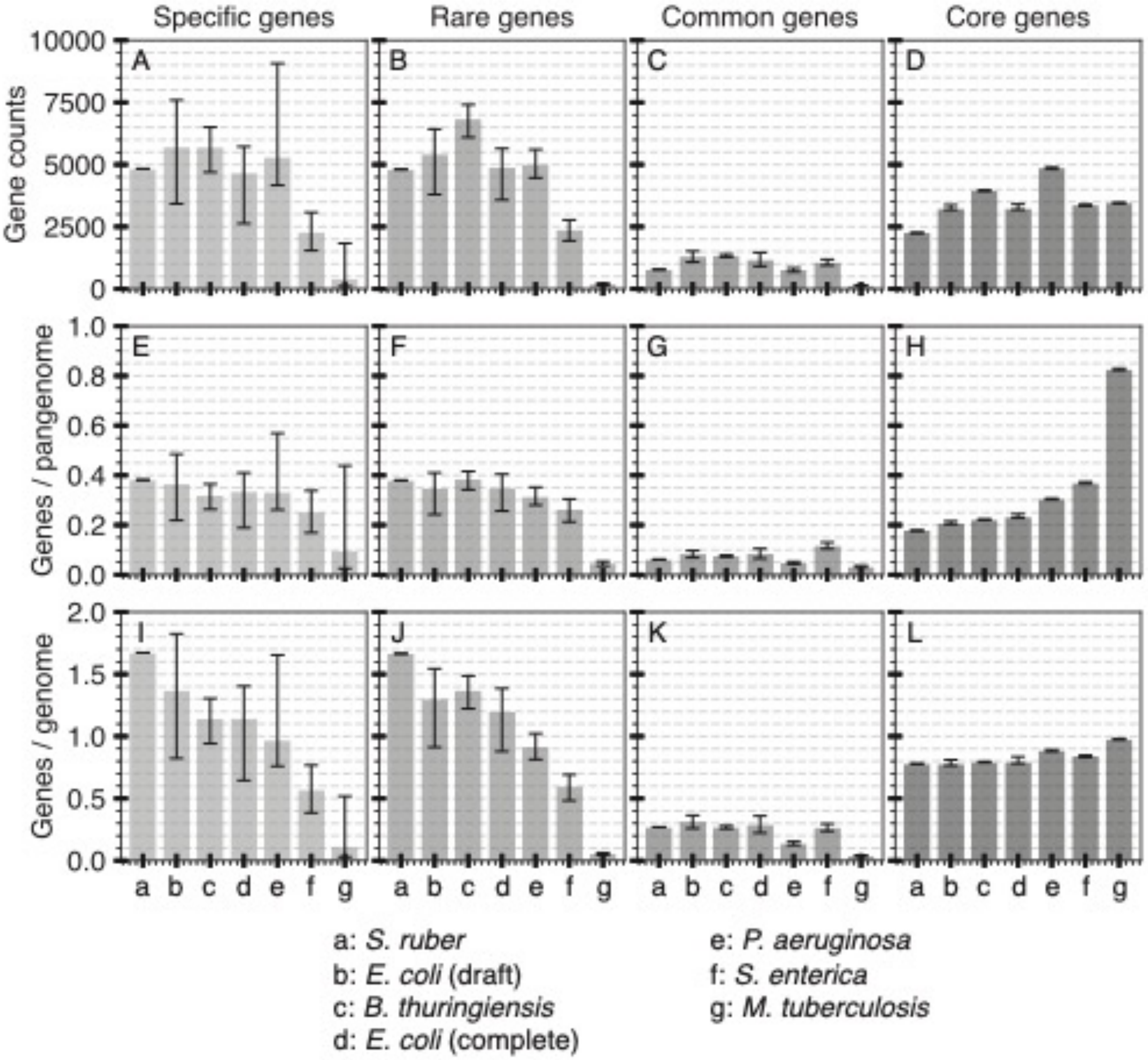

Fig. 5 Detailed comparison of pangenome metrics for selected bacterial species. Pangenome metrics were calculated from draft genome collections for multiple organisms, plus one closed genome collection for *E. coli*. Error bars show the 95% empirical confidence interval of results calculated from 100 random trials each selecting 100 genomes and running 100 permutations. *Sal. ruber* error bars are not shown because only 100 draft genomes from this experiment were available (no random trials). Columns show the contribution to the pangenome of different classes of genes (isolate-specific, rare, common, or core) based on their prevalence among the 100 genomes of the species analyzed (see text for details). (**A-D**) The absolute count of genes for each gene class are shown. (**E-H**) Counts from A-D normalized by the total size of the pangenome from (**A**). (**I-L**) Counts from A-D normalized by the average genome size of each of the species analyzed. Pangenome metrics also provided in Sup. Excel File 2.

Figure 6:

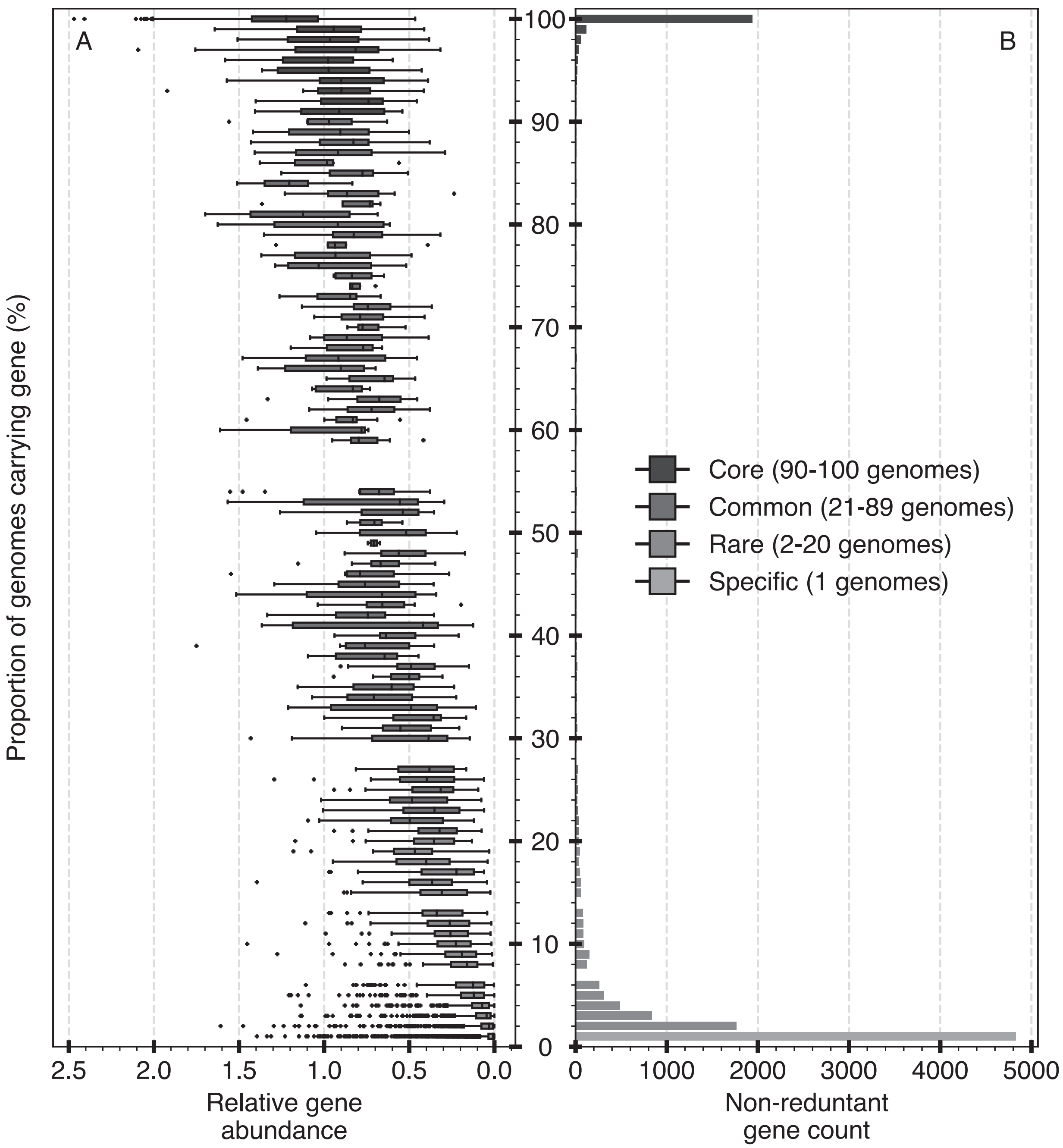

Fig. 6 Relative *in situ* abundance of different classes of *Sal. ruber* genes based on their prevalence in isolate genomes. (**A**) Box plots show the distribution of the average TAD80 (i.e., *in situ* abundance) for each non-redundant gene cluster normalized by the average whole-genome TAD80 that was computed for and averaged across three control pond (salt saturation conditions) metagenomic datasets (y-axis). A normalized value of 1.0 indicates that the gene has

an equivalent *in situ* coverage (abundance) to a single copy core gene. Genes are ordered on the x-axis based on their prevalence among 100 *Sal. ruber* genomes, e.g., a value of 1 at the bottom indicates that only 1 genome has this gene cluster (isolate-specific gene class), and values of 90 – 100 at the top indicate that 90 to 100 genomes have the gene cluster (core gene class). Coverage values for each gene cluster provided in Sup. Excel File 4. (**B**) Counts of the number of non-redundant genes of the pangenome (y-axis) assigned to each gene prevalence class (x-axis); classes are ordered as in panel A.

Figure 7:

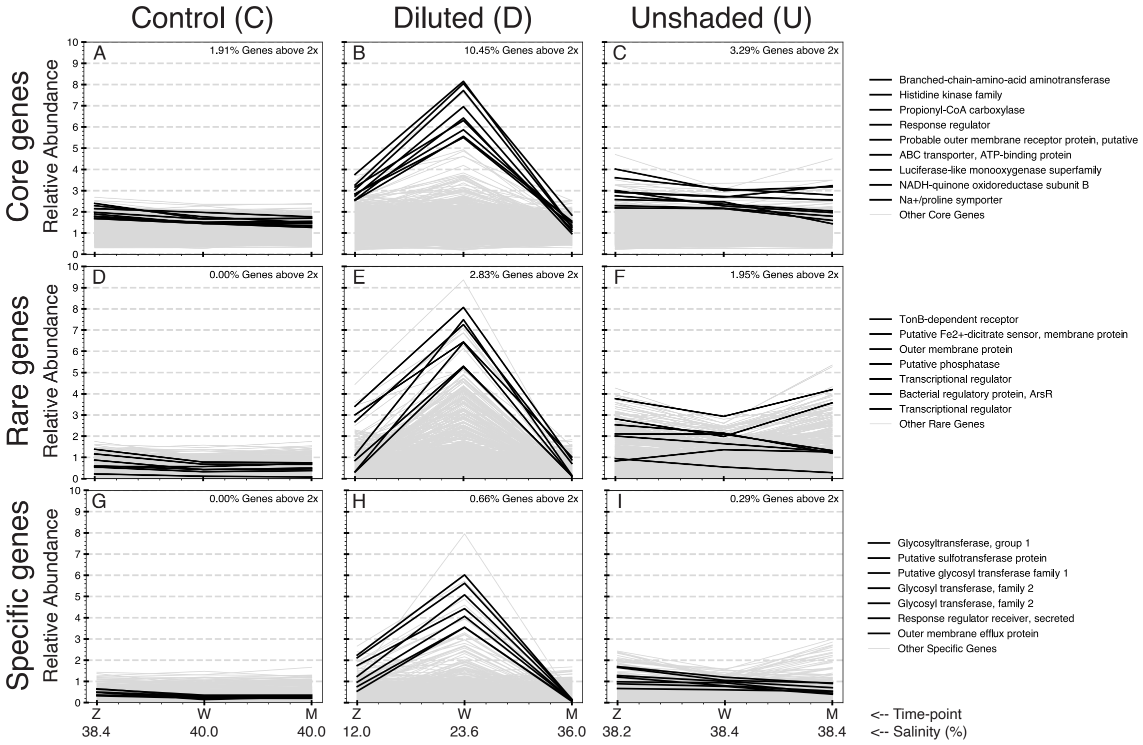

Fig. 7 *Sal. ruber* gene abundance dynamics over the one month period of the experiment. Each line represents a non-redundant gene of the *Sal. ruber* pangenome for which the average TAD80 of the gene normalized by the whole-genome TAD80 of the same sample (y-axes) is plotted against the three metagenomic sampling time points (x-axes) for each of the three separate experimental ponds outlined in Fig. S1. Therefore, the lines represent the relative gene abundance in relation to the relative genome abundance. Results are organized by core (**A-C**), rare (**D-F**), or isolate-specific (**G-I**) gene class. Results for all genes are plotted as gray lines with a handful of genes from each class, selected based on their (higher) peak in **B**, **E**, and **H** panels representing the pond with changing salinity conditions (see x-axis values), shown in black. The corresponding functional annotations of these genes are also shown (figure legend). All gene annotations provided in Sup. Excel File 4. The diluted pond salt concentration was reduced from 33.6% to 12.0% salts at time zero (Z).

References

1. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: an integrative view of gene diversity within microbial populations. BMC Genomics. 2011;12(1):1-10.
2. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol. 2008;11(5):472-7.
3. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci USA. 2005;102(39):13950-5.
4. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev. 2005;15(6):589-94.
5. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Curr Opin Microbiol. 2015;23:148-54.
6. Caro-Quintero A, Konstantinidis KT. Bacterial species may exist, metagenomics reveal. Environ Microbiol. 2012;14(2):347-55.
7. Garcia SL, Stevens SLR, Crary B, Martinez-Garcia M, Stepanauskas R, Woyke T, et al. Contrasting patterns of genome-level diversity across distinct co-occurring bacterial populations. ISME J. 2018;12(3):742-55.
8. Olm MR, Crits-Christoph A, Diamond S, Lavy A, Matheus Carnevali PB, Banfield JF. Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. mSystems. 2020;5(1).
9. Konstantinidis KT, DeLong EF. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. ISME J. 2008;2(10):1052-65.
10. Bendall ML, Stevens SL, Chan LK, Malfatti S, Schwientek P, Tremblay J, et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J. 2016;10(7):1589-601.
11. Johnston ER, Rodriguez RL, Luo C, Yuan MM, Wu L, He Z, et al. Metagenomics reveals pervasive bacterial populations and reduced community diversity across the Alaska tundra ecosystem. Front Microbiol. 2016;7:579.
12. Meziti A, Tsementzi D, Rodriguez RL, Hatt JK, Karayanni H, Kormas KA, et al. Quantifying the changes in genetic diversity within sequence-discrete bacterial populations across a spatial and temporal riverine gradient. ISME J. 2019;13(3):767-79.
13. Orellana LH, Ben Francis T, Kruger K, Teeling H, Muller MC, Fuchs BM, et al. Niche differentiation among annually recurrent coastal Marine Group II Euryarchaeota. ISME J. 2019;13(12):3024-36.
14. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9(1):5114.
15. Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol. 2014;22(5):235-47.
16. Andreani NA, Hesse E, Vos M. Prokaryote genome fluidity is dependent on effective population size. ISME J. 2017;11(7):1719-21.
17. Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. Philos Trans R Soc B. 2006;361(1475):1929-40.
18. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol. 2017;2:17040.

19. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;35(8):725-31.
20. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data. 2018;5:170203.
21. Chen LX, Anantharaman K, Shaiber A, Eren AM, Banfield JF. Accurate and complete genomes from metagenomes. Genome Res. 2020;30(3):315-33.
22. Shaiber A, Eren AM. Composite metagenome-assembled genomes reduce the quality of public genome repositories. mBio. 2019;10(3).
23. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. 2016;1:16048.
24. Meziti A, Rodriguez-R LM, Hatt JK, Peña-Gonzalez A, Levy K, Konstantinidis KT. How reliably do metagenome-assembled genomes (MAGs) represent natural populations? Insights from comparing MAGs against isolate genomes derived from the same fecal sample. Appl Environ Microbiol. 2021.
25. Meziti A, Tsementzi D, Ar Kormas K, Karayanni H, Konstantinidis KT. Anthropogenic effects on bacterial diversity and function along a river-to-estuary gradient in Northwest Greece revealed by metagenomics. Environ Microbiol. 2016;18(12):4640-52.
26. Arevalo P, VanInsberghe D, Elsherbini J, Gore J, Polz MF. A reverse ecology approach based on a biological definition of microbial populations. Cell. 2019;178(4):820-34 e14.
27. Delmont TO, Eren AM. Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. PeerJ. 2018;6:e4320.
28. Delmont TO, Kiefl E, Kilinc O, Esen OC, Uysal I, Rappe MS, et al. Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. Elife. 2019;8.
29. Berube PM, Biller SJ, Hackl T, Hogle SL, Satinsky BM, Becker JW, et al. Single cell genomes of *Prochlorococcus*, *Synechococcus*, and sympatric microbes from diverse marine environments. Sci data. 2018;5(1):1-11.
30. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. Science. 2014;344(6182):416-20.
31. Viver T, Orellana LH, Diaz S, Urdiain M, Ramos-Barbero MD, Gonzalez-Pastor JE, et al. Predominance of deterministic microbial community dynamics in salterns exposed to different light intensities. Environ Microbiol. 2019;21(11):4300-15.
32. Viver T, Cifuentes A, Diaz S, Rodriguez-Valdecantos G, Gonzalez B, Anton J, et al. Diversity of extremely halophilic cultivable prokaryotes in Mediterranean, Atlantic and Pacific solar salterns: evidence that unexplored sites constitute sources of cultivable novelty. Syst Appl Microbiol. 2015;38(4):266-75.
33. Viver T, Conrad RE, Orellana LH, Urdiain M, González-Pastor JE, Hatt JK, et al. Distinct ecotypes within a natural haloarchaeal population enable adaptation to changing environmental conditions without causing population sweeps. ISME J. 2020:1-14.
34. Konstantinidis KT, Tiedje JM. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci USA. 2004;101(9):3160-5.
35. Rodriguez-R LM, Tsementzi D, Luo C, Konstantinidis KT. Iterative subtractive binning of freshwater chronoseries metagenomes identifies over 400 novel species and their ecologic preferences. Environ Microbiol. 2020;22(8):3394-412.

36. Pena A, Teeling H, Huerta-Cepas J, Santos F, Yarza P, Brito-Echeverria J, et al. Fine-scale evolution: genomic, phenotypic and ecological differentiation in two coexisting *Salinibacter ruber* strains. ISME J. 2010;4(7):882-95.
37. Maistrenko OM, Mende DR, Luetge M, Hildebrand F, Schmidt TSB, Li SS, et al. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. ISME J. 2020;14(5):1247-59.
38. Anton J, Lucio M, Pena A, Cifuentes A, Brito-Echeverria J, Moritz F, et al. High metabolomic microdiversity within co-occurring isolates of the extremely halophilic bacterium *Salinibacter ruber*. PLOS One. 2013;8(5):e64701.
39. Luley-Goedl C, Nidetzky B. Glycosides as compatible solutes: biosynthesis and applications. Nat Prod Rep. 2011;28(5):875-96.
40. Antón J, Oren A, Benlloch S, Rodríguez-Valera F, Amann R, Rosselló-Mora R. *Salinibacter ruber* gen. nov., sp. nov., a novel, extremely halophilic member of the bacteria from saltern crystallizer ponds. IJSEM. 2002;52(2):485-91.
41. Antón J, Rosselló-Mora R, Rodríguez-Valera F, Amann R. Extremely halophilic bacteria in crystallizer ponds from solar salterns. Appl Environ Microbiol. 2000;66(7):3052-7.
42. Viver T, Orellana L, Gonzalez-Torres P, Diaz S, Urdiain M, Farias ME, et al. Genomic comparison between members of the *Salinibacteraceae* family, and description of a new species of *Salinibacter* (*Salinibacter altiplanensis* sp. nov.) isolated from high altitude hypersaline environments of the Argentinian Altiplano. Syst Appl Microbiol. 2018;41(3):198-212.
43. Oren A, Rodríguez-Valera F. The contribution of halophilic Bacteria to the red coloration of saltern crystallizer ponds. FEMS Microbiol Ecol. 2001;36(2-3):123-30.
44. Santos F, Moreno-Paz M, Meseguer I, Lopez C, Rossello-Mora R, Parro V, et al. Metatranscriptomic analysis of extremely halophilic viral communities. ISME J. 2011;5(10):1621-33.
45. Kuo CH, Ochman H. Deletional bias across the three domains of life. Genome Biol Evol. 2009;1:145-52.
46. Lane N, Martin W. The energetics of genome complexity. Nature. 2010;467(7318):929-34.
47. Vos M, Hesselman MC, Te Beek TA, van Passel MWJ, Eyre-Walker A. Rates of lateral gene transfer in prokaryotes: high but why? Trends Microbiol. 2015;23(10):598-605.
48. Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol. 2005;3(9):679-87.
49. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Droge J, et al. Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nat Methods. 2017;14(11):1063-71.
50. Munoz R, Lopez-Lopez A, Urdiain M, Moore ER, Rossello-Mora R. Evaluation of matrix-assisted laser desorption ionization-time of flight whole cell profiles for assessing the cultivable diversity of aerobic and moderately halophilic prokaryotes thriving in solar saltern sediments. Syst Appl Microbiol. 2011;34(1):69-75.
51. Urdiain M, López-López A, Gonzalo C, Busse H-J, Langer S, Kämpfer P, et al. Reclassification of *Rhodobium marinum* and *Rhodobium pfennigii* as *Afifella marina* gen. nov. comb. nov. and *Afifella pfennigii* comb. nov., a new genus of photoheterotrophic *Alphaproteobacteria* and emended descriptions of *Rhodobium*, *Rhodobium orientis* and *Rhodobium gokarnense*. Syst Appl Microbiol. 2008;31(5):339-51.

52. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.
53. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.
54. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11(1):119.
55. Rodriguez-R LM, Gunturu S, Harvey WT, Rosselló-Mora R, Tiedje JM, Cole JR, et al. The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. Nucleic Acids Res. 2018;46(W1):W282-W8.
56. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7(1).
57. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3.
58. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLOS One. 2010;5(3):e9490.
59. Rambaut A, Drummond A. FigTree version 1.4. 0. 2012.
60. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. 2007;23(1):127-8.
61. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150-2.
62. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):421.
63. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics. 2020;36(7):2251-2.
64. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30.