Genome Biology

Metabolic-network driven analysis of bacterial ecological strategies

Shiri Freilich (shiri.freilich@gmail.com)
Anat Kreimer (anat.kreimer@gmail.com)
Elhanan Borenstein (ebo@stanford.edu)
Nir Yosef (nir.yosef@gmail.com)
Roded Sharan (roded@post.tau.ac.il)
Uri Gophna (urigo@post.tau.ac.il)
Eytan Ruppin (ruppin@post.tau.ac.il)

Metabolic-network driven analysis of bacterial ecological strategies

Shiri Freilich*†[1,2] , Anat Kreimer*[3] , Elhanan Borenstein[5,6], Nir Yosef[1], Roded Sharan[1], Uri Gophna[4] & Eytan Ruppin†[1,2]

[1] The Blavatnik School of Computer Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel.

[2] School of Medicine, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel.

[3] School of Mathematical Science, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel.

[4] Department of Molecular Microbiology and Biotechnology, Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel.

[5] Department of Biological Sciences, Stanford University, Stanford, CA 94305-5020, USA.

[6] Santa Fe Institute, Santa Fe, NM 87501, USA.

*These authors contributed equally to this work

†Corresponding author

Email addresses: shiri.freilich@gmail.com; anat.kreimer@gmail.com; ebo@stanford.edu; nir.yosef@gmail.com; roded@post.tau.ac.il; urigo@post.tau.ac.il; ruppin@post.tau.ac.il.

Abstract

Background: The growth-rate of an organism is an important phenotypic trait, directly affecting its ability to survive in a given environment. Here we present the first large scale computational study of the association between ecological strategies and growth rate across 113 bacterial species, occupying a variety of metabolic habitats. Genomic data is used to reconstruct the species' metabolic-networks and habitable metabolic environments. These reconstructions are then used to investigate the typical ecological strategies taken by organisms in terms of two basic species-specific measures: *metabolic variability* - the ability of a species to survive in a variety of different environments, and *co-habitation score* vector – the distribution of other species which co-inhabit each environment.

Results: We find that growth rate is significantly correlated with metabolic variability and the level of co-habitation (i.e., competition) encountered by an organism. Most bacterial organisms adopt one of two main ecological strategies: (a) a specialized niche with little co-habitation, associated with a typical slow rate of growth versus (b) ecological diversity with intense co-habitation, associated with a typical fast rate of growth.

Conclusions: The pattern observed suggests a universal principle where metabolic flexibility is associated with a need to grow fast, possibly in the face of competition. This new ability to produce a quantitative description of the growth rate – metabolism – community relationship lays a computational foundation for the study of a variety of aspects of the communal metabolic life.

Background

Variations in growth rate are observed both within and between species, reflecting, respectively, regulatory-level and genomic-level adaptations [1-4]. Since the rate of bacterial growth is determined by metabolic factors such as the rate and yield of ATP production [5], variations in growth rate are bound to be associated with metabolic capabilities and constraints. Several examples have demonstrated, at the single species level, that growth rate is affected by the availability of environmental resources and the level of competition in a given environment [5-8]. Comparative-growth studies have pointed to several metabolic and regulatory genes that are under selective pressure for accelerated growth, e.g., genes involved in the transport of essential substrates in highly-competitive *Escherichia coli* populations [9]. However, such comparative growth studies are typically restricted to species that occupy similar ecological niches, potentially missing the impact of genomic adaptations that may vary across different niches and lifestyles. To this day, the genome-design principles underlying the association between growth rate and metabolic adaptations have not yet been established at a global, cross-species scale.

A comprehensive cross-species analysis, beyond a comparative study of organisms sharing a similar ecological niche, of genomic traits that are associated with the potential growth rates of bacterial organisms was made possible due to a recent list of minimal generation times of a wide spectrum of bacterial species [10, 11]. Previously, these doubling-times data have led to the important finding that variations between genes involved in translation and transcription influence growth rate [10, 11]. Here we focus on the influence of genomic-derived metabolic

properties. We use genomic information to generate second-order (network-based) metabolic knowledge through the reconstruction of metabolic-networks, and third-order environmental knowledge through the reconstruction of habitable metabolic environments for the species studied [12]. Then, species-specific environmental information is further exploited to estimate the level of competition encountered by each organism according to the potential ability of other species to thrive in similar habitats [13]. Through converting genomic data to environmental and communal information, this study examines factors that potentially underlie growth rates across all these levels, through the analysis of the metabolic networks and environments of 528 contemporary sequenced bacterial species, where growth rate data were available for 113 of these species (See Table S1 in Additional data file 1).

Results and Discussion

Growth rate is associated with basic genomic and environmental attributes

We first studied the association between growth rate and (i) the size of the genome, and (ii) the size of the corresponding metabolic network (Methods). Both attributes displayed a significant inverse correlation with doubling time (genome size: -0.31, metabolic network size: -0.38; Table 1), i.e., fast growth rate is typical of species with large genomes and large metabolic networks. Notably, obligatory symbionts (parasites and mutualists) are known to have both slow growth rate and small genome size [11, 14]; excluding this group from the computation, we observe no significant difference in the genome size (or network size) between slow growing and fast growing bacteria (Table 1), indicating that there is no universal link (beyond the unique properties of this group of species) between metabolic network size and

bacterial growth rate. The lack of association between growth rate and genome size was already reported in previous studies [15] where the profound effects of the translation process on growth rate were suggested to mask any influences of genome size on replication speed.

Moving to the environmental dimension, we examine the association between growth rate and two established measures of the variability of species' habitats (fraction of regulatory genes and environmental complexity estimate, Methods). Both measures yield similar results: a significant negative correlation with doubling time (-0.42 and -0.34 respectively, Table 1), i.e., fast growth rate is typical of species which exhibit ecological diversity. Though these correlations are insignificant following excluding obligatory symbionts, we still observe significant differences between fast and slow growers with respect to their fraction of regulatory genes (Table 1). While these data-driven indices track general characteristics of the environment, our goal is to focus on studying the specific relation between metabolic factors and growth. Grouping bacterial species according to their oxygen requirements (aerobic, anaerobic, and facultative), the slowest growth rate (i.e., longest mean generation time) is observed for obligatory aerobic bacteria, followed by obligatory anaerobic bacteria (Table 2). Notably, the fastest growth rate is observed for facultative bacteria (Table 2; significance over the anaerobic group: P=0.03, significance over the aerobic group: P=1.9e-6, Wilcoxon rank sum test); these bacteria can alternate between aerobic and anaerobic metabolism in accordance with their environment [5, 8], utilizing alternative metabolic pathways to maximize rate or yield and gain an advantage over competitors [7]. The growth advantage of these facultative organisms gives rise to the hypothesis that, in general, higher growth rate may be associated with increased metabolic environmental variability and flexibility.

Modeling metabolic-environmental attributes

To test this hypothesis in the absence of an appropriate large-scale data-driven index of metabolic variability, we turned to develop a computational-based one. Employing a previously developed `reverse ecology' algorithm that computes the set of metabolites that an organism extracts from its environment, we reconstructed the likely natural metabolic environments of each organism (Methods, and [12] for a comprehensive description). This provides an ensemble of environments computed for all 528 sequenced organisms, providing the broadest ecological view provided by the current data. Subsequently, the viability of each species is tested in all these environments. This is done by examining if in a given metabolic environment (i.e., a combination of metabolites) an organism can successfully expand its metabolic-network so that it produces a set of target metabolites that are essential for growth (Methods). Repeating this procedure for all species provides an *environmental viability matrix* whose rows denote the species, its columns denote the environments, and its binary entries denote whether a given species can survive in a given environment. We then computed the mean population level (no. of species per environment) across the environments populated by organisms of a given lifestyle. Reassuringly, these results are compatible with ecological knowledge (Figure 1): soil bacteria and species populating the human gut inhabit the most densely-populated environments; sparsely-populated environments are inhabitated by specialized organisms and (though to a lesser extent) by obligatory symbionts [16-19]. Notably, our data set includes a large group of obligatory symbionts (54 species in comparison to 7 terrestrial organism and 17 gut bacteria; See Tables S1 and S2 in Additional data files 1 and 2, respectively), hence indicating that the level of population of a given

environment does not reflect the prevalence of the life-style category of the species it inhabit, i.e., despite the ubiquity of obligatory symbionts in the data they tend to inhabit specialized metabolic environments. We additionally examined alternative approaches for generating other biologically-plausible sets of random metabolic environments. One such alternative approach for generating random environments is to construct 528 shuffled seed environments, i.e., maintaining an approximation of the original metabolites representation overall seeds (Methods). Sparse populations of environments populated by specialized and obligatory symbionts versus dense populations of environments inhabited by soil and gut bacteria is also observed when using this alternative collection of random environments (See Additional data file 3).

Growth rate is associated with the level of metabolic variability

The *environmental scope index* of a species (*ESI*) is defined as the fraction of environments in which it is viable. The ESI measure of metabolic variability is positively correlated with both genome size (0.4, P = 3e-6, Spearman) and metabolic network size (0.6, P = 1e-10, Spearman). It is also positively correlated with the data-driven general environmental-diversity measures examined above (fraction of regulatory genes: 0.32, P = 0.008; estimate of environmental complexity: 0.23, P = 0.01, Spearman), hence providing support to the ecological plausibility of the model. There is a significant negative correlation between the ESI and doubling time in the complete dataset (-0.25), and notably, the differences between the ESI scores of slow and fast growing bacteria remain significant also after excluding obligatory symbionts (Table 1). Thus, there is a general association between broader metabolic capacities and faster maximal growth rates, extending the initial observations concerning the fast growth rate of facultative bacteria and implying that metabolic versatility of species is

better associated with their growth rate then other, more general, environmental characteristics. This result, as all other reported correlations, remains valid when using the alternative collection of random environments described above (Table 1; Methods). The negative correlation between ESI and duplication time is also maintained in species which are evenly distributed among different habitat types and taxonomic groups (See additional data file 3). *Pseudomonas aeruginosa*, an organism with a high ESI score (See Table S1 in Additional data file 1) provides an example for fast growth rate in a generalist, possessing broad metabolic capabilities that allow it to successfully grow in diverse environments [20]. However, the association between fast growth and metabolic flexibility is not at all obvious, as one may assume that living in a specific niche habitat would enable an organism to specialize and adapt towards a high-growth solution. Indeed, *Desulfotalea psychrophila,* an organism with a low ESI score (See Table S1 in Additional data file 1), provides such an example. It is a sulfate-reducing extremophilic bacterium, thriving in extreme conditions (cold arctic sediments), and exhibiting metabolic and environmental specialization [21].

Growth rate is associated with the level of co-habitation

If metabolic specialization does not preclude fast growth, how then can we explain the slow growth of most specialists? An arising hypothesis is that such organisms face weak competition. Conversely, organisms that occupy a large variety of metabolic environments face a larger number of co-inhabiting species, which in turn may exert selection pressure for maintaining higher growth rates. To test this hypothesis we used the *co-habitation score (CHS)* vector (deduced from the *environmental viability matrix*), denoting the number of species that co-populate each viable environment of a given species. This vector can serve as an indication of the level of competition

encountered by a species in its habitats. We focus on each species' most populated niche (maximal-*CHS*) and most sparsely populated one (minimal-*CHS*). The minimal-CHS score is not significantly correlated with either ESI or doubling time. In contrast, the maximal-*CHS* exhibits a significant inverse relationship with duplication time (Table 1), i.e., faster growth rates are observed in richly populated, competitive environments. The maximal-*CHS* also displays a highly marked positive correlation with metabolic variability (*P*-value < 1e-3, computed by comparing to random; Supporting Information). That is, a species' metabolic flexibility tends to erode when it populates only sparsely populated, non-competitive environments. This result remains valid when using an alternative collection of random environments (Table 1; Methods). The negative correlation between Maximal-CHS and duplication time is also maintained in species which are evenly distributed among different habitat types and taxonomic groups (See Additional data file 3). The relevance of maximal- and minimal-*CHS* scores to growth rate can be put in a biological context by considering the lifestyle of the pathogen *Staphylococcus aureus*: inside a host-cell (where no competition with other bacterial species is encountered) it exhibits a far slower growth rate than in the more competitive environment of the human skin [22, 23].

Delineating major ecological strategies

To delineate potential major ecological strategies we grouped the bacterial species according to their location on the ESI-CHS plane (Figure 2). As can be expected from the tight association between the environmental scope and co-habitation scores, the large majority of all species fall within the low ESI - low maximal-CHS and high ESI - high maximal-CHS diagonal groups, exhibiting two different but equally popular ecological strategies: (a) a specialized niche with little

competition versus (b) ecological diversity with intense competition, with the latter group displaying faster growth rates (*P*=0.02, Wilcoxon rank sum test). *Escherichia coli*, a generalist capable of fast growth is an example of the first group, while *Mycobacterium leprae*, an obligatory intracellular pathogen with highly specialized nutritional demands and an exceptionally slow growth [24, 25] is an example of the second group (Figure 2). However, some organisms exhibit different ecological approaches (Figure 2). In some bacterial species tight adaptation to a specific niche (low scope) does not involve escaping competition (high maximal-CHS score). The oral bacterium *Fusobacterium nucleatum* is an example of a species whose metabolism is adapted to a specific, though non-exclusive, niche [26]. In contrast, the last and smallest group includes species with a relatively high environmental scope but exclusive habitats. Members of this group exhibit a faster growth rate than the low scope/low max-CHS group (*P*=0.05, Wilcoxon rank sum test). As an example, the intracellular pathogen *Legionella pneumophila* has a duplication time close to hundred times faster than *Mycobacterium leprae*. Whereas *M. leprae* possesses highly specific metabolic requirements that limit its ability to exploit the resources in the host cell [25], *L. pneumophila* exhibits a more generic metabolism, scavenging the host cell for both sugars and amino-acids, exhausting the latter's resources [27]. Accordingly, *L. pneumophila* is the causative agent of an acute disease where *M. leprae* causes a long-lasting chronic disease, requiring tight adaptation to co-existence within the host cell.

Beyond these specific examples, the characterization of growth strategies based on the intricate interplay between the ESI-maximal-CHS values suggests that when taken together these values can be used for predicting growth strategies. Using ESI and maximal-CHS values retrieved from the 113 species which participate in the

original analysis, we trained an SVM classifier which assigns bacterial species into one of two extreme growth classes (either fast or slow, Methods). We tested the generalization ability of the classifier in a cross validation setting obtaining an average ROC score of 0.75 (Methods). We then obtained growth rate data for *Parachlamydia* UWE25 [11], an endosymbiont of amoeba from the chlamydiae group, and for *Bacillus thuringiensis*, a widely distributed bacteria [28]; for both species growth rate data were not included in our analysis so far. *Parachlamydia* UWE25 is an obligate intracellular bacterium that exhibits reduced central metabolic and biosynthetic pathways, and is auxotrophic for most amino acids and nucleotides [29]. In accordance with the lifestyle of this organism, we compute for this species a low environmental diversity and low competition scores (ESI: 0.002, maximal-CHS: 0, See Table S1 in Additional data file 1). For the ubiquitous *Bacillus thuringiensis* we compute high environmental diversity and high competition scores (ESI: 0.09, maximal-CHS: 59, See Table S1 in Additional data file 1). We applied the SVM classifier to characterize the growth rates of these species, in accordance with experimental data *Bacillus thuringiensis* falls into the fast-growing category and *Parachlamydia* UWE25 falls into the slow-growing category (corresponding to the experimentally observed doubling times of 40 minutes [30] and 48 hours respectively [11],).

Conclusions

In summary, this paper presents the first large-scale rigorous computational exploration of the ecological strategies taken by a species in association with its growth rate, lifestyle and metabolic-capabilities. Several limitations of this analysis should be mentioned: first, the estimation of the growth environments and their

viability is based on a topological network-based computation, which is obviously a first-approximation model of the underlying biology. Second, co-habitation involves additional facets of interspecies interactions beyond competition, notably, cooperation and symbiosis. Bearing those in mind, the metabolic environmental model correctly captures several patterns already observed in data-driven biological habitats, testifying to its ecological plausibility.

We find that growth rate is significantly positively correlated with both the span of the environments and the level of competition that bacterial species encounter in their metabolic habitats. The model points at two main ecological strategies suggesting a universal principle where metabolic flexibility is associated with a need to grow fast, possibly in face of competition. The new ability to produce a quantitative description of the growth rate – metabolism – community relationship lays a computational foundation for the study of a variety of aspects of the communal metabolic life. With the growing recognition that bacteria should be better studied in the context of their ecological niche and communities, future computational approaches should take into account the complex interrelationships between organisms. Such approaches are likely to become increasingly helpful for studying various aspects of microbial life within naturally-occurring ecological habitats.

Materials and Methods

Dataset

Metabolic data were collected from KEGG [31] (release 46) for 528 bacterial organisms. Out of these species, information describing the maximal doubling time was available for 113 species (downloaded from [11]). We constructed the metabolic

networks of bacterial organisms following the approach outlined in [32]. To download networks see Additional data file 3.

Construction of metabolic environments

Metabolic growth environments were inferred using the seed algorithm developed by [12]. This algorithm predicts the set of exogenously acquired compounds, given the metabolic network. We additionally examined alternative approaches for generating other biologically-plausible sets of random metabolic environments. One such alternative approach for generating random environments is to construct 528 shuffled seed environments, i.e., maintaining an approximation of the original metabolites representation overall seeds. Each metabolite in the original seed environments is randomly assigned to the shuffled environments where its representation over all environments is 1.05 times in comparison to its original representation. That is, if a certain metabolite has, for example, 20 appearances over all seeds, then it is randomly assigned to 21 out of 528 environments. This process is repeated for each seed metabolite. The 1.05 ratio of appearances between original and shuffled environments was chosen as it allows a similar level of inhabitation of the *environmental viability matrix*, described below (See Additional data file 3).

Characterizing bacterial environments

Fraction of regulatory genes values were taken from [33], describing the fraction of transcription factors out of the total number of genes in the organism, an indicator of environmental variability [19]. Environmental complexity estimated values were obtained from [19], where the natural environments of 117 bacterial species were categorized based on the NCBI classification for bacterial lifestyle [34] and ranked according to the complexity of each category (1- obligatory symbyonts; 2-

specialized; 3- aquatic; 4- facultative host-associated; 5- multiple; 6- terrestrial species [19]). Annotations for environmental complexity were available only for 68 species from the 113 species for which doubling time was available. To validate the reliability of these annotations we manually searched the literature. In two cases we changed the original annotation (from multiple to terrestrial, See Table S3 in Additional data file 4). In addition, we searched the literature for annotations for the remaining 45 species (See Table S3 in Additional data file 4). Together with retrieving the annotations as described above, classification of species into habitats was also done by looking for the presence of species from the dataset in environmental samples. Occurrence of species from our dataset in environmental samples was inferred according to the results of a BLAST search [35] of 16s RNA sequences from the 528 species in the analysis against env_nt – a comprehensive collection of sequences from environmental samples (downloaded on February 2009). We find that in the large majority of cases (30/33) experimental findings support the literature-based annotations (See Additional data file 3).

All parameters retrieved/computed for the species in the analysis are provided in See Table S1 in Additional data file 1.

Computing the environmental viability matrix, Environmental Scope Index (ESI) and Co-habitation Score (CHS)

As a measure for species viability we constructed a list of 65 compounds termed "target metabolites" (See Table S4 in Additional data file 5), which are most likely to be essential for growth in most species [36-38]. A species-specific target metabolite list is formed by the intersection between the target metabolites and the metabolites that each species produces. We then tested the viability of each species over the set of

528 metabolic growth environments. Given a specific organism and an environment, an organism is considered viable in this environment if all its essential target metabolites are produced – this is examined by using a network expansion algorithm [39] that outputs an activated metabolic sub-network, and verifying that the expanded subnetwork produces all target metabolites. This process yields the *environmental viability matrix* whose rows denote the species and columns denote the metabolic environments, and binary entries denote the corresponding viability. From this matrix, the scope (*ESI)* and co-habitation (*CHS*) scores for each species are deduced: The *ESI* of a species is defined as its fraction of viable environments. The *CHS* vector of a species records how many viable organisms populate each of its viable environments.

All software used for the analysis will be provided upon request from the authors.

Constructing an SVM classifier

We partitioned the organisms according to their doubling time: fast and slow growers are these whose duplication time is shorter and longer in at least one standard deviation from the mean (0.48 and 9 hours respectively, See Figure S6 in Additional data file 3). Species with intermediate values were excluded from the analysis. For the remaining species, ESI and maximal-CHS values were used for training a Support Vector Machine (SVM) classifier with a linear kernel [40]. We estimate the accuracy of the classifier using a ten-fold cross validation. In this procedure, the organisms are randomly partitioned into 10 distinct set; then the class labels (slow or fast) in each set are predicted by a classifier trained on the rest of the sets. We repeated this procedure 50 times, and report the mean and standard deviation of the receiver operating characteristic (ROC) curve [41]. Our quality metric is the area under this curve (the ROC score).

Abbreviations

CHS: co-habitation score; ESI: environmental scope index of a species.

Authors' contributions

SF designed and performed research, analyzed the data and drafted the manuscript. AK performed research, analyzed the data and drafted the manuscript. EB contributed analysis tools. NY performed research. UG analyzed the data. RS analyzed the data. ER conceived and designed research and wrote the paper. All authors discussed the results and commented on the manuscript.

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 is a table listing genomic and ecological attributes for the 113 species in the analysis (Table S1). Additional data file 2 is a table listing the NCBI annotations and description of environmental sample for species that can be identified in an environmental sample (Table S2). Additional data file 3 includes supplementary Notes, Figures, detailed description of all tables in Additional data files and a Table (Table S5). Additional data file 4 is a table listing the original and manually curated values of environmental complexity (Table S3). Additional data file 5 is a table listing the biomass target metabolites (Table S4). Additional data file 6 is a table listing genomic and ecological attributes for the 528 species in the metabolic analysis (Table S6).

References

1. Berg OG, Kurland CG: **Growth rate-optimised tRNA abundance and codon usage**. *J Mol Biol* 1997, **270**:544-550.
2. Dong H, Nilsson L, Kurland CG: **Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates**. *J Mol Biol* 1996, **260**:649-663.
3. Klappenbach JA, Dunbar JM, Schmidt TM: **rRNA operon copy number reflects ecological strategies of bacteria**. *Applied and environmental microbiology* 2000, **66**:1328-1333.
4. Mager WH, Planta RJ: **Coordinate expression of ribosomal protein genes in yeast as a function of cellular growth rate**. *Mol Cell Biochem* 1991, **104**:181-187.
5. Novak M, Pfeiffer T, Lenski RE, Sauer U, Bonhoeffer S: **Experimental tests for an evolutionary trade-off between growth rate and yield in E. coli**. *Am Nat* 2006, **168**:242-251.
6. Hansen SR, Hubbell SP: **Single-nutrient microbial competition: qualitative agreement between experimental and theoretically forecast outcomes**. *Science (New York, NY* 1980, **207**:1491-1493.
7. Helling RB: **Speed versus efficiency in microbial growth and the role of parallel pathways**. *J Bacteriol* 2002, **184**:1041-1045.
8. Pfeiffer T, Schuster S, Bonhoeffer S: **Cooperation and competition in the evolution of ATP-producing pathways**. *Science (New York, NY* 2001, **292**:504-507.
9. Lenski RE, Mongold JA, Sniegowski PD, Travisano M, Vasi F, Gerrish PJ, Schmidt TM: **Evolution of competitive fitness in experimental populations of E. coli: what makes one genotype a better competitor than another?** *Antonie Van Leeuwenhoek* 1998, **73**:35-47.
10. Rocha EP: **Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization**. *Genome Res* 2004, **14**:2279-2286.
11. Couturier E, Rocha EP: **Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes**. *Mol Microbiol* 2006, **59**:1506-1518.
12. Borenstein E, Kupiec M, Feldman MW, Ruppin E: **Large-scale reconstruction and phylogenetic analysis of metabolic environments**. *Proceedings of the National Academy of Sciences of the United States of America* 2008.
13. Janga SC, Babu MM: **Network-based approaches for linking metabolism with environment**. *Genome biology* 2008, **9**:239.
14. Joseph B, Goebel W: **Life of Listeria monocytogenes in the host cells' cytosol**. *Microbes Infect* 2007, **9**:1188-1195.
15. Mira A, Ochman H, Moran NA: **Deletional bias and the evolution of bacterial genomes**. *Trends Genet* 2001, **17**:589-596.
16. Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI: **Host-bacterial mutualism in the human intestine**. *Science,* 2005, **307**:1915-1920.
17. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM:

Comparative metagenomics of microbial communities. *Science* 2005, **308**:554-557.

18. Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, Lapidus A, Grigoriev I, Richardson P, Hugenholtz P, Kyrpides NC: **Use of simulated data sets to evaluate the fidelity of metagenomic processing methods**. *Nat Methods* 2007, **4**:495-500.

19. Parter M, Kashtan N, Alon U: **Environmental variability and modularity of bacterial metabolic networks**. *BMC Evol Biol* 2007, **7**:169.

20. Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, Hickey MJ, Brinkman FS, Hufnagle WO, Kowalik DJ, Lagrou M, Garber RL, Goltry L, Tolentino E, Westbrock-Wadman S, Yuan Y, Brody LL, Coulter SN, Folger KR, Kas A, Larbig K, Lim R, Smith K, Spencer D, Wong GK, Wu Z, Paulsen IT, Reizer J, Saier MH, Hancock RE, Lory S *et al*: **Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen**. *Nature* 2000, **406**:959-964.

21. Rabus R, Ruepp A, Frickey T, Rattei T, Fartmann B, Stark M, Bauer M, Zibat A, Lombardot T, Becker I, Amann J, Gellner K, Teeling H, Leuschner WD, Glöckner FO, Lupas AN, Amann R, Klenk HP: **The genome of Desulfotalea psychrophila, a sulfate-reducing bacterium from permanently cold Arctic sediments**. *Environ Microbiol* 2004, **6**:887-902.

22. Vesga O, Groeschel MC, Otten MF, Brar DW, Vann JM, Proctor RA: **Staphylococcus aureus small colony variants are induced by the endothelial cell intracellular milieu**. *J Infect Dis* 1996, **173**:739-742.

23. von Eiff C: **Staphylococcus aureus small colony variants: a challenge to microbiologists and clinicians**. *Int J Antimicrob Agents* 2008, **31**:507-510.

24. Gomez-Valero L, Rocha EP, Latorre A, Silva FJ: **Reconstructing the ancestor of Mycobacterium leprae: the dynamics of gene loss and genome reduction**. *Genome Res* 2007, **17**:1178-1185.

25. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, Mungall K, Basham D, Brown D, Chillingworth T, Connor R, Davies RM, Devlin K, Duthoy S, Feltwell T, Fraser A, Hamlin N, Holroyd S, Hornsby T, Jagels K, Lacroix C, Maclean J, Moule S, Murphy L, Oliver K, Quail MA *et al*: **Massive gene decay in the leprosy bacillus**. *Nature* 2001, **409**:1007-1011.

26. Kapatral V, Anderson I, Ivanova N, Reznik G, Los T, Lykidis A, Bhattacharyya A, Bartman A, Gardner W, Grechkin G, Grechkin G, Zhu L, Vasieva O, Chu L, Kogan Y, Chaga O, Goltsman E, Bernal A, Larsen N, D'Souza M, Walunas T, Pusch G, Haselkorn R, Fonstein M, Kyrpides N, Overbeek R: **Genome sequence and analysis of the oral bacterium Fusobacterium nucleatum strain ATCC 25586**. *J Bacteriol* 2002, **184**:2005-2018.

27. Molofsky AB, Swanson MS: **Differentiate to thrive: lessons from the Legionella pneumophila life cycle**. *Mol Microbiol* 2004, **53**:29-40.

28. Broderick NA, Raffa KF, Handelsman J: **Midgut bacteria required for Bacillus thuringiensis insecticidal activity**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**:15196-15199.

29. Horn M, Collingro A, Schmitz-Esser S, Beier CL, Purkhold U, Fartmann B, Brandt P, Nyakatura GJ, Droege M, Frishman D, Rattei T, Mewes HW,

Wagner M: **Illuminating the evolutionary history of chlamydiae**. *Science* 2004, **304**:728-730.

30. Donovan WP, Tan Y, Slaney AC: **Cloning of the nprA gene for neutral protease A of Bacillus thuringiensis and effect of in vivo deletion of nprA on insecticidal crystal protein**. *Applied and environmental microbiology* 1997, **63**:2311-2317.

31. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic acids research* 2000, **28**:27-30.

32. Cano DA, Pucciarelli MG, Martinez-Moya M, Casadesus J, Garcia-del Portillo F: **Selection of small-colony variants of Salmonella enterica serovar typhimurium in nonphagocytic eucaryotic cells**. *Infection and immunity* 2003, **71**:3690-3698.

33. Madan Babu M, Teichmann SA, Aravind L: **Evolutionary dynamics of prokaryotic transcriptional regulatory networks**. *J Mol Biol* 2006, **358**:614-633.

34. Entrez Genome Project [http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi].

35. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic acids research* 1997, **25**:3389-3402.

36. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Mol Syst Biol* 2007, **3**:121.

37. Becker SA, Palsson BO: **Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation**. *BMC Microbiol* 2005, **5**:8.

38. Oh YK, Palsson BO, Park SM, Schilling CH, Mahadevan R: **Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data**. *J Biol Chem* 2007, **282**:28791-28799.

39. Ebenhoh O, Handorf T, Heinrich R: **Structural analysis of expanding metabolic networks**. *Genome Inform* 2004, **15**:35-45.

40. Boser BE, Guyon IM, Vapnik VN: **A training algorithm for optimal margin classifiers**. In: *Annual ACM Workshop on COLT: 1992; Pittsburgh*: ACM Press; 1992: 144-152.

41. Hanley JA, McNeil BJ: **The meaning and use of the area under a receiver operating characteristic (ROC) curve**. *Radiology* 1982, **143**:29-36.

42. Metabolic networks. [http://www.cs.tau.ac.il/~jonatha6/publications/113_dt_networks_html.zip]

Tables

Table 1. Correlation (P value) V. Duplication time

	Total (N=113)	§ Non obligatory symbiont only (N=77)	Significance of difference between slow grower, fast grower±	
			Total (N=113)	§ Non obligatory symbionts spc. (N=77)
Genome size (bp)	-0.30 (0.001)	-0.04 (0.7)	0.001 (S: 2695676 F: 3402099)	0.4 (S: 3614838 F: 3479053)
Network size	-0.38 (3.1e-05)	-0.13 (0.2)	0.002 (S: 326 F: 410)	0.3 (S: 408 F: 431)
Fraction of regulatory genes [33]	-0.42 (0.0004)	-0.21 (0.13)	4e-4 (S: 0.03 F: 0.05)	0.2 (S: 0.04 F: 0.05)
Estimate of environmental complexity [19]	-0.34 (2-04)	-0.07 (0.5)	1e-4 (S: 3 F: 4)	€-
ESI	-0.25 (0.008)	-0.23 (0.04)	0.03 (S: 0.006 F: 0.02)	0.06 (S: 0.008 F: 0.02)
ESI – random environments ψ	-0.47 (1,6e-07)	-0.35 (0.002)	8e-6 (S: 0.007 F: 0.03)	0.002 (S: 0.01 F0.00 4)
Max *CHS*	-0.27 (0.03)	-0.28 (0.01)	0.03 (S: 14 F: 27)	0.02 (S: 20 F: 31)
Max *CHS*– random environments ψ	-0.34 (1e-4)	-0.23 0.05	6e-4 (S: 39 F: 72)	0.01 (S: 50 F: 85)

±The two sets of data (all species, non obligatory symbionts) were divided into two bins according to species' growth rate (fast and slow). The significance between the genomic attributes studied (e.g., genome size, network size etc) was calculated with one-sided Wilcoxon rank sum test. In brackets: the mean value of the relevant attribute in the slow growing and fast growing groups.

§ According to definitions from [19] and manual curation (Methods).

€Values not computed since low-ranked estimates of environmental complexity represent the obligatory symbionts which were excluded from the analysis.

ψRandom environments are described at the Methods section.

Table 2. Typical duplication time of bacterial organisms according to their mode of respiration±

	Number of species§	**Mean duplication time**	**Median duplication time**	Mean network size
Aerobic bacteria	40	13	3	412
Anaerobic bacteria	18	5.3	1.6	318
Facultative bacteria	41	1.7	0.8	380

±Oxygen-dependence annotations were taken from [34].

§Species in this analysis are those for which duplication times are available, the metabolic network was reconstructed (Methods), and their oxygen-dependence group is one of the groups in the table (bacteria species whose oxygen-dependence annotation is "unknown" or "microaerophylic" are not shown here).

Figure legends

Figure 1. Mean co-habitation (population) levels of environments occupied by bacteria of a given life style. Annotations of lifestyle are according to [19] (Specialized, obligatory symbionts, aquatic, multiple, faculatative symbionts, and terrestrial) and according to identification of species in environmental samples (human gut, Methods). Number of environments in each lifestyle (ordered as in the figure) are 11, 81, 5, 144, 157, 38, 117 (environments can populate species of more than a single lifestyle). Bars show the standard error.

Figure 2. *Environmental scope* index versus maximal *co-habitation score*. Size of dots corresponds to duplication time – larger size corresponds to longer duplication time and slower growth rate. Color of dots corresponds to their ecological habitat (See Table S1 in Additional data file 1): red – obligatory host-associated; green – specialized; blue – aquatic; black – host-associated (non-obligatory); orange – multiple; brown – terrestrial. Legends: DT – duplication time; BL – bottom left (47 species); BR – bottom right (10 species); TL – top left (16 species); TR – top right (40 species). The plot is divided according to median values of the axes. Abbreviations: Low ESI – low maximal-CHS (BL); high ESI – low maximal-CHS (BR); high ESI – high maximal-CHS (TR); and low ESI – high maximal-CHS (TL).

Figure 1

Figure 1

Figure 2

Figure 2
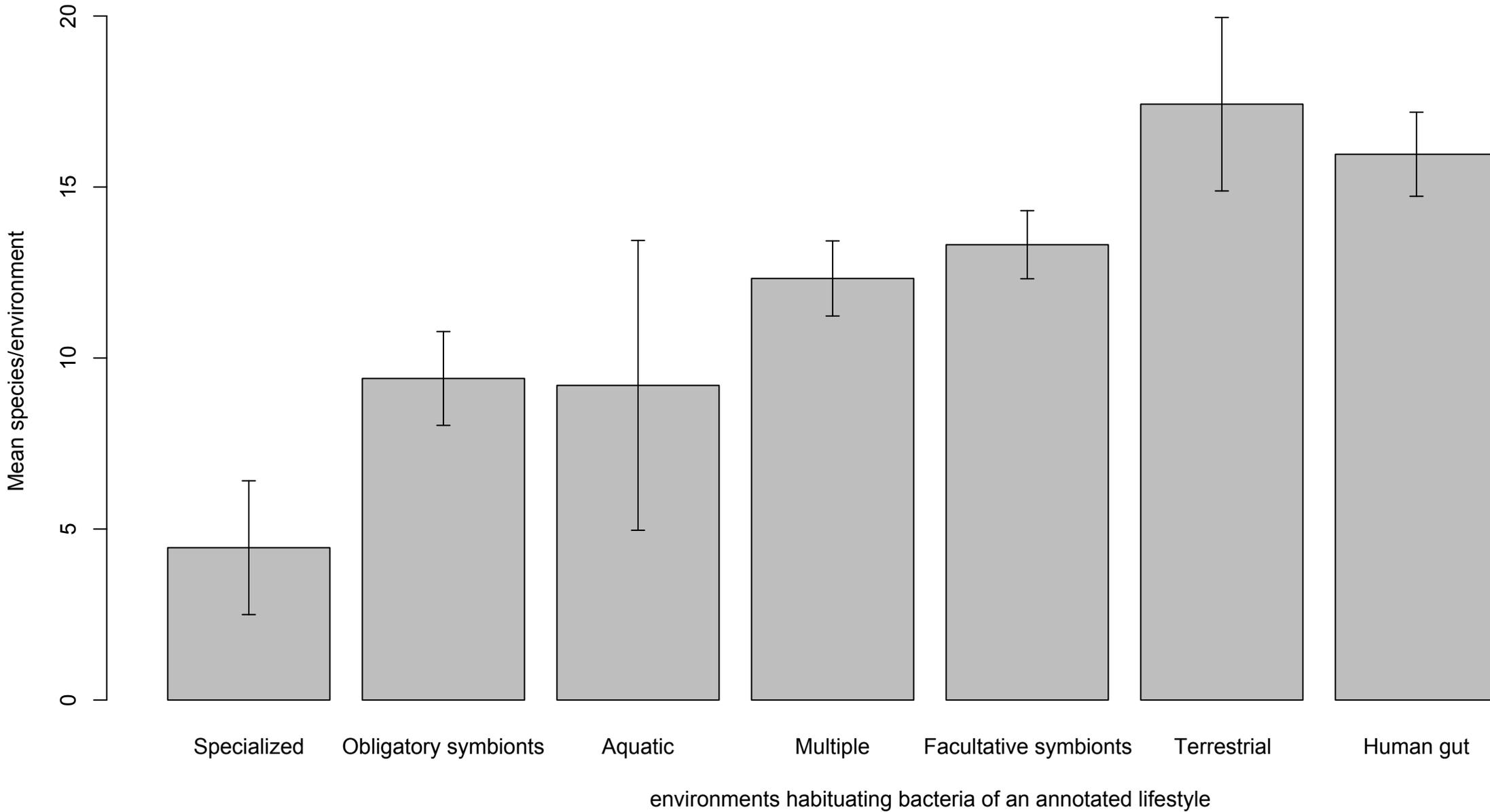
Additional files provided with this submission:

Additional file 1: additional_data_file_1.xls, 51K
http://genomebiology.com/imedia/1302404505280671/supp1.xls
Additional file 2: additional_daa_file_2.xls, 20K
http://genomebiology.com/imedia/8499871672806726/supp2.xls
Additional file 3: additional_data_file_3.pdf, 324K
http://genomebiology.com/imedia/6817513272811961/supp3.pdf
Additional file 4: additional_data_file_4.xls, 55K
http://genomebiology.com/imedia/1782579964280672/supp4.xls
Additional file 5: additiona_data_file_5.pdf, 15K
http://genomebiology.com/imedia/1160011872806740/supp5.pdf
Additional file 6: additional_data_file_6.xls, 107K
http://genomebiology.com/imedia/1382645199280673/supp6.xls