OPEN ACCESS Freely available online

PLOS ONE

Metabolic Proximity in the Order of Colonization of a Microbial Community

Varun Mazumdar[1], Salomon Amar[1,2], Daniel Segrè[1,3]*

1 Bioinformatics Program, Boston University, Boston, Massachusetts, United States of America, **2** Center for Anti-Inflammatory Therapeutics; Boston University Goldman School of Dental Medicine, Boston, Massachusetts, United States of America, **3** Department of Biology and Department of Biomedical Engineering, Boston University, Boston, Massachusetts, United States of America

Abstract

Microbial biofilms are often composed of multiple bacterial species that accumulate by adhering to a surface and to each other. Biofilms can be resistant to antibiotics and physical stresses, posing unresolved challenges in the fight against infectious diseases. It has been suggested that early colonizers of certain biofilms could cause local environmental changes, favoring the aggregation of subsequent organisms. Here we ask whether the enzyme content of different microbes in a well-characterized dental biofilm can be used to predict their order of colonization. We define a metabolic distance between different species, based on the overlap in their enzyme content. We next use this metric to quantify the average metabolic distance between neighboring organisms in the biofilm. We find that this distance is significantly smaller than the one observed for a random choice of prokaryotes, probably reflecting the environmental constraints on metabolic function of the community. More surprisingly, this metabolic metric is able to discriminate between observed and randomized orders of colonization of the biofilm, with the observed orders displaying smaller metabolic distance than randomized ones. By complementing these results with the analysis of individual vs. joint metabolic networks, we find that the tendency towards minimal metabolic distance may be counter-balanced by a propensity to pair organisms with maximal joint potential for synergistic interactions. The trade-off between these two tendencies may create a ''sweet spot'' of optimal inter-organism distance, with possible broad implications for our understanding of microbial community organization.

Citation: Mazumdar V, Amar S, Segrè D (2013) Metabolic Proximity in the Order of Colonization of a Microbial Community. PLoS ONE 8(10): e77617. doi:10.1371/journal.pone.0077617

Editor: Jens Kreth, University of Oklahoma Health Sciences Center, United States of America

Received June 8, 2013; **Accepted** September 3, 2013; **Published** October 30, 2013

Funding: This research was partially funded by the United States Department of Energy, grant DE-SC0004962, and by NIH grants R01DE15345, R01HL6801 and R0115989. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: dsegre@bu.edu

Introduction

In many natural environments, bacteria and other microorganisms are part of spatially structured ecosystems, and engage in complex interactions, involving the exchange of nutrients and chemical signals [1,2]. Such communities provide their members with protection from environmental perturbations, and allow for effective utilization of available resources. Modifications of the environment, such as a change in diet in a human host, can cause shifts in the composition of a microbial community. In turn, the collective metabolic activity of the community itself can substantially modify the environment, and set the stage for transitions between health and disease states. While 16S rRNA studies [3] and metagenomic DNA sequencing [4] have been very helpful in providing provide global snapshots of the composition and biological functions of a community, a big gap still exists in our understanding of the forces that shape specific interactions between different organisms. Systems biology approaches have started to provide valuable insight into the metabolic basis of interactions between different species in elementary [5] and complex [6] microbial ecosystems. However, a lot is still unknown on how properties of individual species give rise to global ecosystem organization.

Here, we address this problem by presenting an intermediate-scale approach to elucidate the role of metabolism in determining the order of colonization in a microbial community. Our approach captures the complexity of how eleven species spatially organize in a biofilm, by using a mathematical description of metabolism that lies in between the detailed quantitative power of stoichiometric models [7], and the coarse enrichment analyses typically obtained from metagenomic studies [4,8–11]. We focus specifically on one of the most intensively characterized biofilm systems, the dental biofilm, which plays a crucial role in tooth and gum diseases. Oral pathogens such as *Porphyromonas gingivalis* are also known to be able to enter the blood stream, possibly causing cardiovascular disease. The collection of bacteria present on the surface of teeth, anchored to the salivary pellicle, has a specific spatial structure, which has been mapped and investigated in detail [12–14]. The biofilm structure is made up of a number of different species, which aggregate together over time to form a complex structure. The aggregation process is not random [15]; instead it appears to be a repeatable sequential process mediated by bacterial adhesins that allow organisms to aggregate to surfaces and other bacteria by binding to specific receptor moieties (Fig. 1A). The initial colonizers are capable of binding to salivary pellicle receptors, and the subsequent organisms proceed to bind to the initial colonizers [12,14,16]. Late colonizers, such as *P. gingivalis*, which has been linked to periodontal disease, are found in the final layer of this complex structure [17]. Some steps of the colonization process can lead to mutual exclusion between closely related

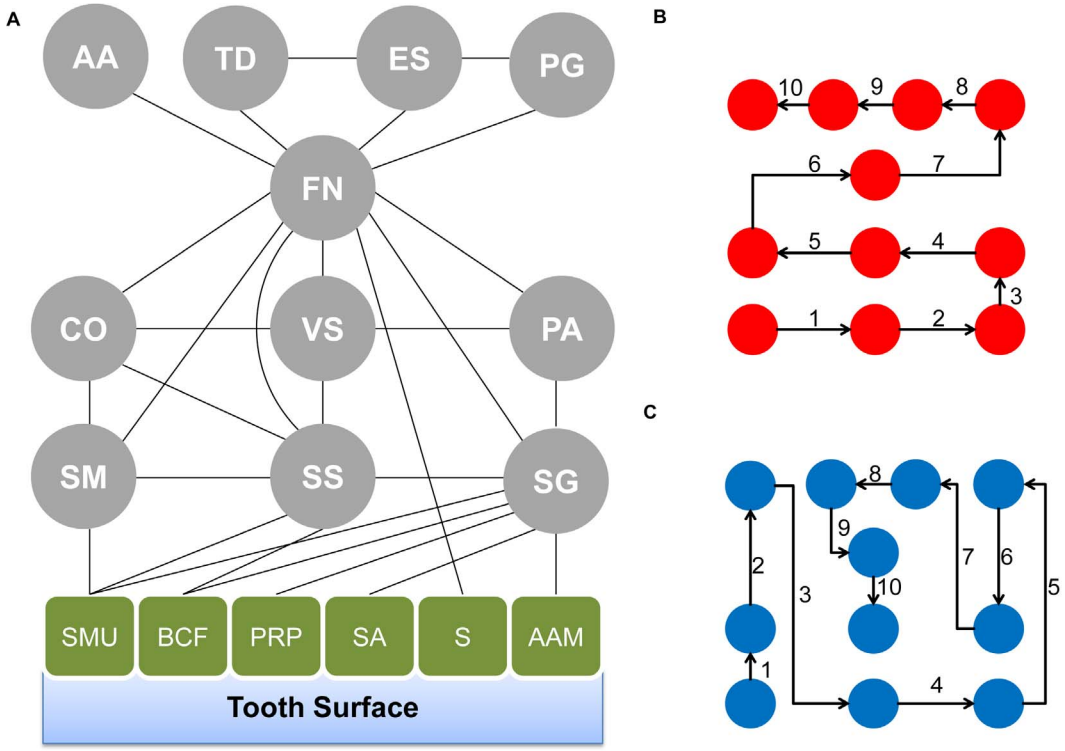

Figure 1. A simplified model of dental biofilm. (A) Rectangular nodes represent components of the salivary pellicle while circular nodes represent organisms in the biofilm. Lines represent known interactions (often mediated by adhesin molecules) between different components of the biofilm. The organism abbreviations are as follows: Layer 1: SM-*Streptococcus mitis*, SS-*Streptococcus sanguinis* and SG-*Streptococcus gordonii*. Layer 2: CO-*Capnocytophaga ochracea*,VS-*Veionella* (represented by *Veillonella parvula*), and PA-*Propionibacterium acnes*. Layer 3: FN-*Fusobacterium nucleatum*. Layer 4: AA-*Aggregatibacter actinomycetemcomitans*, TD-*Treponema denticola*, ES-*Eubacterium* (represented by *Eubacterium eligens*), and PG-*Porphyromonas gingivalis*. The salivary receptors have the following abbreviations: SMU-Sialylated mucins, BCF-Bacterial cell fragment, PRP-Proline rich protein, SA-Salivary agglutinin, S- Statherin and AAM-Alpha amylase. (B) A schematic representation of one of the many possible step-wise orders of colonization that conforms to the layered organization inferred from the literature, i.e. is such that the path that walks through the different species is monotonically departing from the salivary pellicle upwards. In our calculations of inter-species metabolic distances, we average the distances between any two species connected by a segment. This calculation is performed for all paths that reflect the order of colonization, giving rise to the distributions shown in Fig. 2. (C) A schematic representation of one of the many possible randomized orders of colonization that do not follow the order of the literature-derived layers; for the 11 organisms present in the biofilm there are 11! possible permutations.
doi:10.1371/journal.pone.0077617.g001

species (e.g. streptococci), leading to drastically different macroscopic disease-related outcomes [18].

In this work we test the hypothesis that metabolism is a predictor, and potentially a major driving force, of the order of colonization in the oral biofilm. Specifically, based on the individual inter-species interactions mapped by Kolenbrander ([12][19], Fig. 1) we quantify the overlap in metabolic functions between adjacent organisms, and compare the distribution of such overlaps to the one obtained for randomized biofilms or for random assemblages of bacteria. We find that the real biofilm is characterized by a significantly larger overlap in metabolic functions between adjacent species, relative to randomized biofilm compositions and structures. Specific metabolic pathways can be associated with the different layers, providing a snapshot of the gradient of metabolic requirements across the biofilm. The observed tendency towards maximal metabolic overlap is likely counteracted by an opposite trend driven by the synergistic advantage of combining the metabolic capabilities of sufficiently different species. In all, these findings suggest that an optimal tradeoff between resource sharing and functional synergy may constitute a fundamental property of structured microbial communities. Our approach, much more detailed than broad functional enrichment studies, but much less demanding than stoichiometric flux balance models, should be broadly applicable to other microbial ecosystems, where spatial or temporal order matters.

Results

Metabolic Proximity among Different Layers in the Oral Biofilm

We ask whether the order of colonization in the human dental biofilm may reflect a quantitative principle of microbial ecosystem organization. While the structure of the biofilm from the Kolenbrander model (Fig. 1A, and [12]) reflects known mutual binding between adjacent species, we develop our analysis based on the premise that such binding effects reflect fundamental adaptations to environmental gradients and mutual metabolic exchange between species. Hence, we analyze the biofilm structure in terms of mutual distances between adjacent organisms. 16S rRNA-based distances are standard practice when estimating the similarity of organisms without dealing with the complexities of whole genome alignment [3,10]. The assumption that 16S rRNA is conserved allows investigators to ascertain evolutionary relationships between organisms. Here, however, we wish to utilize a metabolic, rather than an evolutionary distance [17] (See

Methods). Such a metabolic distance will provide a method to gauge the difference in biochemical functions between different species. It is important to note that a metabolic metric can be used to compare biochemical abilities of different organisms without constructing full-fledged genome-scale stoichiometric models, such as the ones built for several microbial species [20–26], including the oral pathogen *P. gingivalis* [17]. We expect that organisms with similar enzyme profiles (based on the above metric) will have a comparable ability to utilize and process metabolites from their environment. The specific metabolic distance we use in this work is a standard metric (Jaccard's distance, $\mathcal{J}$) gauging the degree of dissimilarity between the sets of enzymes present in the two organisms (see Methods).

As a first step towards ascertaining the validity of this premise, one can test whether the 11 organisms that are in the dental biofilm are on average closer to each other than 11 randomly chosen prokaryotes from the KEGG database [27–30]. To this end, 1000 random groups of 11 organisms were chosen from the list of KEGG prokaryotes and the sum of pairwise metabolic distances were calculated for each permutation (11 factorial potential orders) for a given random group. The average of all permutation scores was calculated for each random group. We found that the groups of 11 organisms belonging to the oral biofilm model tend to have a smaller average of pairwise metabolic distances ($\mathcal{J}$) than 1000 randomly selected groups of prokaryotes (Fig. 2). This is not unexpected, as the organisms inhabit the same niche in the human body and hence must have some underlying metabolic similarity to cope with the common environment. It is interesting to note that the mean distance value between organisms of the dental biofilm is far from being close to a global minimum, when compared to its position in the distribution. This could be due to the fact that organisms with similar metabolic requirements will compete for the same environmental niche and probably would not be found in close proximity to each other within a complex multi-species biofilm.

The next question we ask is whether a special pattern of inter-species metabolic distances can be observed between adjacent organisms in the layered structure of the oral biofilm. In Fig. 1A we present a simplified version of the model presented in [12]. The simplified model contains only organisms whose genome has been sequenced, and whose annotation is available in KEGG. We translate the Kolenbrander map into a set of possible orders of colonization by assuming that an organism can join the biofilm only if it can bind to an organism that that is already present in the biofilm, or to an environmental anchor point. In this way, we determine 864 orders in which the bacteria may join, consistently with the reduced Kolenbrander model (Fig. 1B). The 864 orders ($3!\times3!\times4!$) come from all permutations that allow organisms to be placed in their correct layers. Conversely, there would be 39,916,800 (i.e. 11!) orders which disregard the network of experimentally known interactions that constitute the layered model (Fig. 1A). For each given order of colonization, we compute the average Jaccard's distance between metabolic compositions of adjacent organisms ($<\mathcal{J}>$, see Methods). Fig. 2 shows that the distribution of average distances for the orders of colonization compatible with the Kolenbrander model is markedly shifted to the left relative to the background distribution of random orders ($P<2\cdot10^{-7}$). This implies that spatially adjacent organisms in the biofilm have a larger number of common metabolic enzymes relative to randomly chosen pairs. In order to ascertain the overall robustness of this result we repeated the analysis upon different types of perturbations. In particular, we tried to omit from the calculation the *Streptococci* species, which are phylogenetically and metabolically very close to each other, and might therefore bias the result towards high significance. Despite removing these organisms, we still found a statistically significant p-value ($P<1.2\cdot10^{-4}$). Furthermore we verified that the results are not too sensitive to removal of specific enzymes. In fact, we found that the result is still significant when up to 60% of the enzymes used in the analysis are removed (Fig. S1). The "minimal metabolic distance" criterion apparently satisfied by the non-random orders of colonization may be indicative of the way dental biofilm is thought to form. Each group of organisms creates a micro-environment that is conducive for the next set of organisms. Each new organism joining the biofilm can take advantage of the micro-environment generated by organisms that precede it, provided that the inter-species metabolic distance is small enough to make this build-up favorable.

The above analysis demonstrates that adjacent species in the correct order of colonization tend to minimize their mutual metabolic enzyme distance. This finding does not rule out the possibility that similar effects might be observable based on alternative metrics that take into account other (non-metabolic) gene categories. In other words, is the order of colonization pattern reported predominantly a metabolic effect, or does it reflect a general inter-species distance? To address this question, we performed the same analysis shown in Fig. 2 for a large set of non-metabolic genes (see Methods). As shown in Fig. 3A, the correct and random order distributions are still significantly different ($P<0.0023$), but their separation is much less dramatic than what is found with metabolic enzymes ($P<2\cdot10^{-7}$). Given this result, we can infer that metabolism is one of the most important factors in oral biofilm organization, more so than other classes of genes as a whole.

Finally, to confirm that the pattern observed is not strongly dependent on the metric used, we verified that an alternative, widely used metric can discriminate between correct vs. randomized order of colonization. In particular, we calculated the amount of Shannon information that each organisms adds to the system relative to the previous organism in the order of colonization (see Methods). As with the distance calculation, the overall score is the sum of the pairwise added information values for a given order of colonization. Using a 2-sample Komogrov-Smirnov test [31], we found that literature-informed orders are significantly ($P<4.5\cdot10^{-6}$) smaller than randomized distributions (Fig. 3B). An interesting aspect of the information content metric is that, in contrast to the metabolic distance defined above, it can also capture directionality, i.e. discriminate between a colonization order that starts with the *Streptococci* layer (the first, pellicle-bound layer of the biofilm, Fig. 1A) and one that ends with *Streptococci* layer.

If, as suggested by the above results, metabolism is a fundamental determinant of the order of colonization, we would expect to be able to find that specific metabolic pathways can be associated with different layers of the biofilm. Indeed, by performing a GSEA (Gene Set Enrichment Analysis) for metabolic functions (see Methods and Dataset S1), we found that gradients of metabolic functionalities span the different layers (Fig. 4). The pathways that displayed significant enrichment are: arginine and proline metabolism, biosynthesis of alkaloids, carbon fixation, glyoxylate and dicarboxylate metabolism, glycine, serine and threonine metabolism, nitrogen metabolism, porphyrin and chlorophyll metabolism, pentose and glucuronate interconversions, propanoate metabolism, pyruvate metabolism, terpenoid backbone biosynthesis, and tricarboxylic acid cycle. Enzymes related to carbohydrate and proline metabolism are enriched at the initial colonizer stage, possibly allowing for utilization of available carbohydrates as a source of energy that is present in the

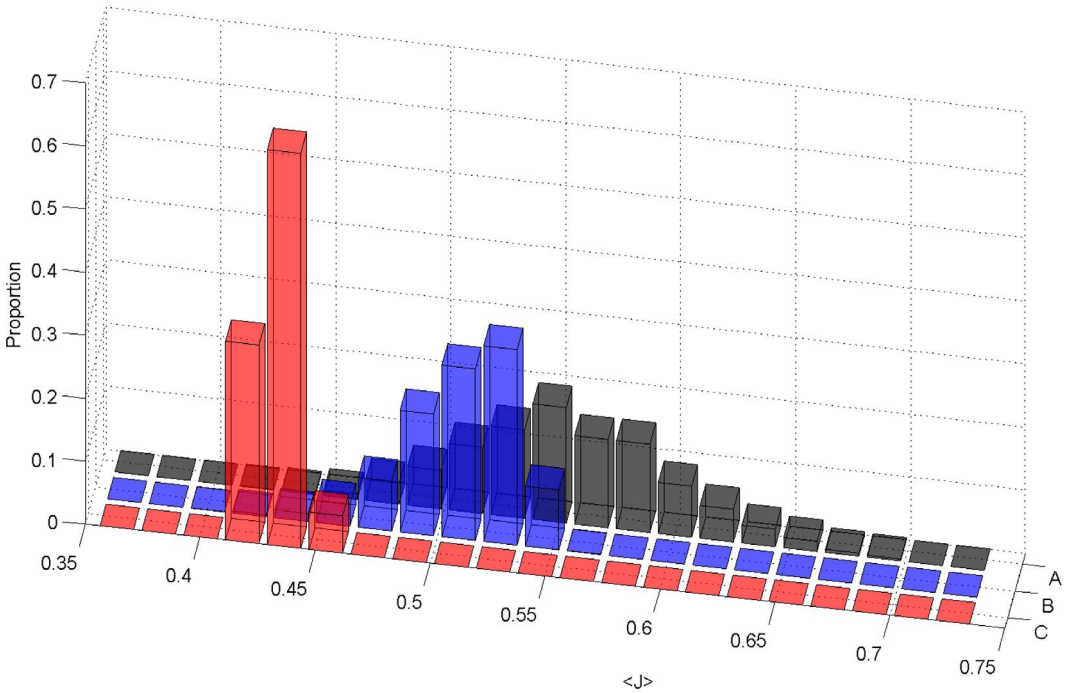

Figure 2. Metabolic distance distributions for correct and randomized orders of colonization. Distributions of average pairwise Jaccard's distance are compared across different computational realizations of the 11-species biofilm. In particular, we show in red (C) the distribution of average pairwise distances between the 11 organisms for all paths that reflect the layered structure of the Kolenbrander map (see Fig. 1B). In blue (B) we show the distribution obtained for all possible random orders that do not necessarily reflect the layered order of colonization (e.g. path shown in Fig. 1C). The last distribution (grey, A) is obtained from choosing in random order 11 random prokaryotes from the KEGG database.
doi:10.1371/journal.pone.0077617.g002

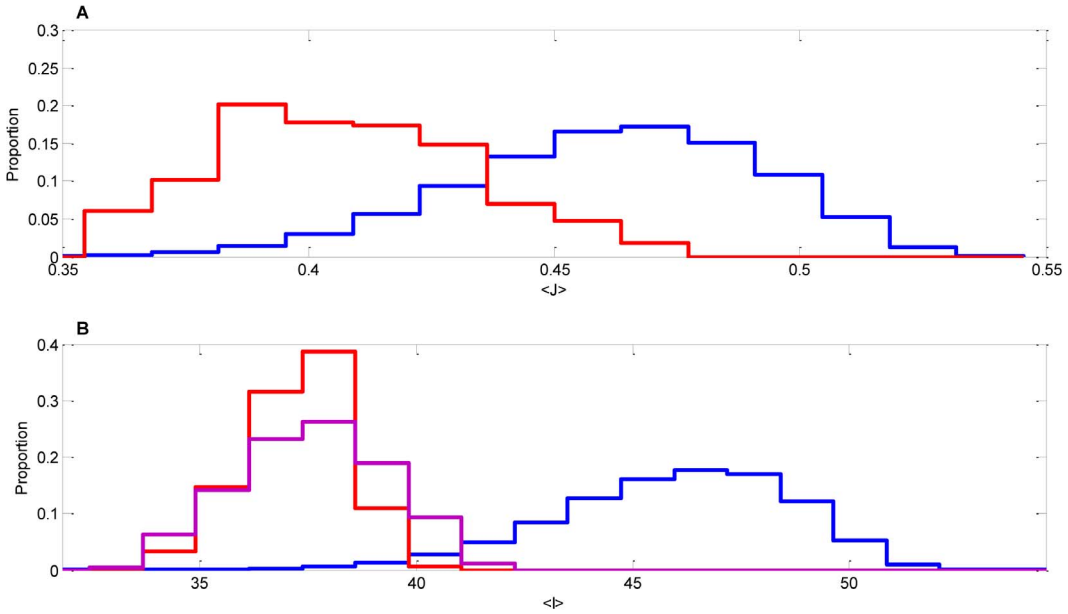

Figure 3. Distributions of alternative metrics for correct and randomized orders of colonization. (A) Similar to what shown in Fig. 1B and Fig. 2, we computed inter-species distance between organisms along paths that respect (red) or do not respect (blue) the layered order of colonization of the Kolenbrander map. Here, however, as opposed to Fig. 2, we compute the Jaccard distance between two species based on their profiles of non-enzyme genes (as identifiable through KEGG KO numbers). (B) The correct and incorrect orders of colonization are compared based on an information metric, rather than on the Jaccard distance. In walking along a colonization order path from one organism to the next, we compute (in Nats) the amount of information added due to the presence of previously absent enzymes. The added information for each pair of adjacent organisms is summed to form the added information score, along paths that respect (red) or do not respect (blue) the layered order of colonization. The purple distribution is obtained by computing the added information scores for orders of colonization that reflect the layered structure, but walk through it in reverse order (i.e. from the outer layer downwards towards the salivary pellicle).
doi:10.1371/journal.pone.0077617.g003

saliva [32]. The second biofilm layer contains both propionate and TCA pathways. Both require, as input, compounds such as lactate, a byproduct of carbohydrate metabolism which in turn is converted into cytotoxic byproducts [33]. Additionally, butyrate is known to affect gingival epithelial cells inducing apoptosis in sufficient concentrations [34]. Apoptosis of tissues provides organisms with a highly enriched food source [35]. Fumarase and succinate dehydrogenase, are both enriched in this layer of the biofilm. Both enzymes provide a pathway for proline metabolism byproducts to be funneled into energy production which takes advantage of proline catabolism enrichment in the previous biofilm layer. The third layer of the biofilm is enriched for porphyrin metabolism, a pathway that is essential to *Porphyromonas gingivalis*. The organism has a well-known requirement for heme and has been implicated in a number of disease processes such as chronic periodontitis [36]. It is however not capable of producing heme for itself, and must therefore scavenge it from the environment. Correspondingly, we found enrichment for enzymes related to heme production, in particular along a pathway that converts L-glutamate to 5-aminolevulinate, a precursor of heme. In the fourth layer of the colonization process, we find enrichment for nitrogen-related and TCA-cycle genes. This specific combination could reflect amino acids from tissue degradation being shunted into cellular metabolism via entry points within the TCA cycle.

Determination of Optimal Metabolic Overlap Using Flux Modes

The above analysis of metabolic distances between adjacent organisms in the oral biofilm demonstrates that the correct order of colonization displays close to minimal average metabolic distance relative to randomized orders. Another way of formulating this principle is that, within a biofilm, organisms next to each other will be as close as possible in terms of their metabolic enzyme content. Taken to the extreme, this principle would suggest that biofilms may be preferably composed of rather similar and uniform species (compatibly with the biofilm size and environmental gradients). This is likely not the case, both because organisms metabolically too close to each other may engage in fierce competition for survival [37], and because there may be a physiological advantage to pairing organisms that are neither too close nor too distant from each other.

To formulate a specific hypothesis about this last scenario, we evaluated the metabolic potential of conjoined metabolic networks as a function of their metabolic distance, using elementary flux modes. Elementary flux modes analysis identifies all minimal non-zero flows through a metabolic network [38]. At a first approximation, the number of elementary flux modes can be thought of as an estimate of the size of the space of possible paths through a metabolic network. Here, we sought to estimate the increase in the number of such paths for a pair of interacting metabolic networks (e.g. two bacterial species), relative to the metabolic capabilities of isolated networks, as a function of the similarity between the two networks. The intuition is that when two networks are very similar to each other, there is little added benefit in combining them with each other. At the opposite extreme, if two networks are too different from each other they will ''speak different metabolic languages'' and barely be able to build significant synergistic pathways. In between these two extremes, there may be an inter-species metabolic distance that provides a maximal synergistic benefit.

Indeed, upon computing a metabolic synergy score for randomly generated pairs of metabolic networks with a given Jaccard's distance between each other (see Methods), we found that the mean score displays a distinct profile as a function of inter-network metabolic distance (Fig. 5). In particular, there is a peak at a Jaccard's coefficient of 0.2. This means that having 33.3% reaction overlap between the component networks is optimal, in the sense that it generates the maximum number of useable balanced pathways (elementary flux modes). The range of Jaccard's coefficients (i.e. metabolic similarity) within which synergistic interaction is expected extends up to approximately $\mathcal{J}c = 0.55$ (i.e. $\mathcal{J} = 0.45$). This means that metabolic networks with a distance below $\mathcal{J} = 0.45$ will have little potential for increased biochemical capabilities through metabolic cross-talk. Interestingly, this Jaccard's distance is very close to the average of the

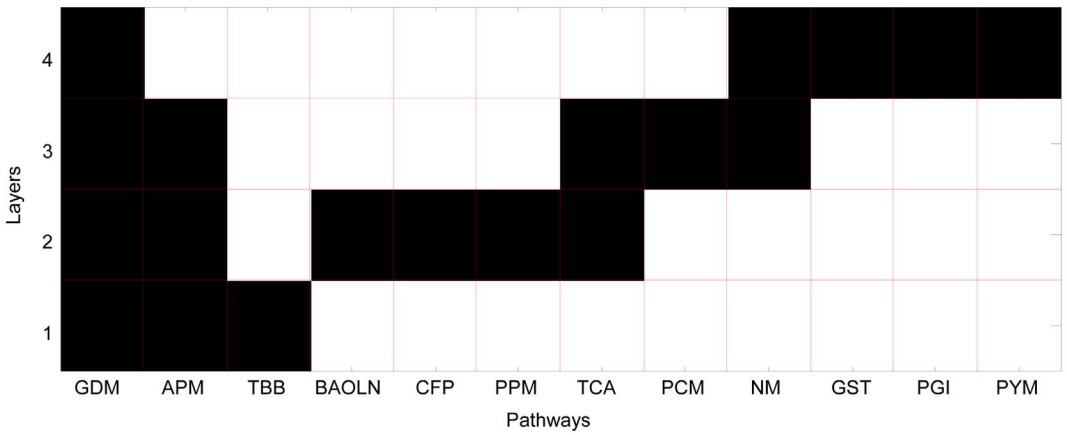

Figure 4. Metabolic pathway enrichment across layers. Based on the enzyme content of the different species found in different layers of the biofilm (with layers labeled from 1 to 4, see Fig. 1), one can estimate whether any given layer is enriched for specific metabolic functions. Enzyme and pathway enrichments for each layer are computed based on a standard GSEA algorithm. Black boxes in the pathways-by-layers matrix denote enrichment of a particular KEGG pathway in a given layer. The pathway abbreviations are as follows: APM-Arginine and proline metabolism, BAOLN-Biosynthesis of alkaloids derived from ornithine lysine and nicotinic acid, CFP-Carbon fixation pathways in prokaryotes, GDM-Glyoxylate and dicarboxylate metabolism, GST-Glycine, serine and threonine metabolism, NM-Nitrogen metabolism, PCM-Porphyrin and chlorophyll metabolism, PGI-Pentose and glucuronate interconversions, PPM-Propanoate metabolism, PYM-Pyruvate metabolism, TBB-Terpenoid backbone biosynthesis, and TCA-Tricarboxylic acid cycle.
doi:10.1371/journal.pone.0077617.g004

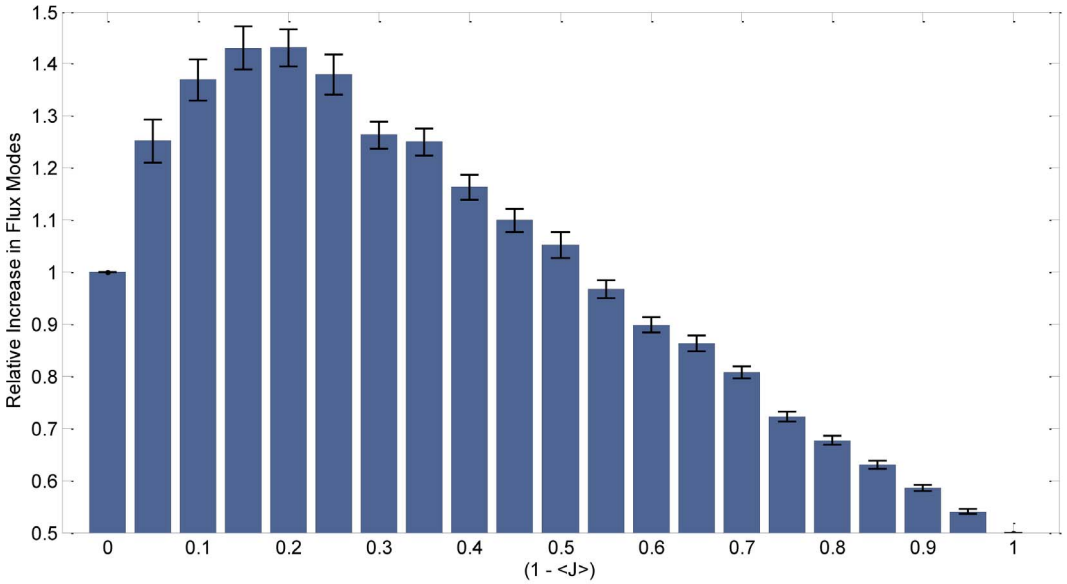

Figure 5. Expected synergy between metabolic networks as a function of metabolic distance. The synergy is computed as the count of elementary flux modes (pathways) that are feasible for a metabolic network that is the union of two networks with a given Jaccard's distance from each other, normalized to the count of elementary flux modes of the constituent networks. The count of elementary flux modes can be thought of as an estimate of the number of distinct metabolic tasks that the network can perform, i.e. its versatility. Hence, the graph shows how the versatility of two conjoined networks relative to the constituent networks is maximal for an intermediate Jaccard's distance between such networks. 100 random paired networks were generated for each of several possible Jaccard's distances. Bar heights reflect the average normalized increase in the number of elementary flux modes, whereas error bars represent the standard error of the mean.
doi:10.1371/journal.pone.0077617.g005

distance among adjacent species in oral biofilm (Fig. 2). A possible interpretation of this result is that the players in the community encounter a tradeoff between maximizing their metabolic overlap, and still not losing the benefit of possible synergistic interactions (i.e. going *beyond* $\mathcal{J}=0.45$, i.e. $\mathcal{J}_C=0.55$).

Discussion

We have addressed the question of whether the spatio-temporal organization of a biofilm can be understood in terms of the differential metabolic properties of individual organisms relative to their neighboring organisms. We implemented a simplified model of the spatial organization of the biofilm, based on experimental evidence of individual pairwise interactions between species. This abstraction of the colonization process enabled us to discretize the order of succession, and systematically investigate all potential permutations of organisms in a linear fashion. A more complex model could more realistically capture inter-species dynamics in physical space [15], without the limitations imposed by taking into account only pairwise interactions and a step-by-step "walk" through the different layers of the biofilm. However, our simplified approach overcomes in an effective way the combinatorial complexity of multi-species networks, and takes advantage of the available pairwise interaction data. We found that metabolic similarity is a highly informative indicator of vicinity in the biofilm. The metabolic structure of the biofilm is reflected in the existence of multiple layers enriched for specific biochemical pathways. This structure lends credence to the idea that each layer contributes to a gradient of metabolic properties, causing environmental modifications that pave the way for subsequent layers of bacteria. We cannot exclude the possibility that the observed effect might be just a result of metabolically similar organisms adapting to environmentally present gradients (e.g. abundance of oxygen). However, the fact that next-to-minimal metabolic distance is significantly associated with binding between organisms in the biofilm suggests this metabolic similarity is truly reflective of particular inter-species interactions. The current work is limited to the eleven organisms of the Kolenbrander map whose annotated sequence was publically available (in KEGG) at the time of the analysis. Future extensions could include additional organisms, benefit from improved genome annotation approaches [39] and gradually move to more mechanistic models of microbe-microbe interactions, such as ecosystem-level flux balance models [40]. In addition, while the current evidence we use for a putative order of colonization is based on a collection of multiple *in vitro* individual pairwise interactions, more comprehensive *in vivo* measurements [19] could in the future be used as a more accurate baseline for testing hypotheses.

Our analysis focused on a specific microbial community in which the order of colonization is manifested both in the chronological sequence of events leading to the full biofilm, as well as in the final spatial architecture of the biofilm itself. We envisage that this analysis could be extended to other microbial ecosystem with a similar spatio-temporal organization, such as biofilms on catheters and medical instruments [41–44] or microbial mats in hot springs and desert environments [45,46]. However, the approach we proposed is not limited to communities with a well-defined or known spatial structure, and could be extended to analyze purely temporal orders of colonization in microbial ecosystems whose biomass is found largely in a planktonic phase, or whose detailed spatial structure is not easily observable (e.g., in the gut microbiome [10,47]). Metagenomic sequencing projects frequently produce 16S rRNA population composition data, which in a longitudinal study provides us with changes in population composition over time. Combining population data with enzymatic profiles from KEGG would make it possible to test whether metabolic proximity is significantly predictive of temporal species-to-species shifts in an ecosystem.

An interesting outcome of our analysis is the hypothesis that multiple counteracting forces may ultimately determine, at the evolutionary scale, an optimal steady state genomic and spatial configuration of different species in a biofilm. Close metabolic proximity seems to be one desirable criterion for spatial vicinity, motivated by uniformity of environmental conditions, and by multiple chances for metabolic cross-feeding. At the same time, organisms which are metabolically too close to each other would likely compete for common metabolic resources. In addition, as we found in Fig. 5, they would have minimal chance for true synergism, i.e. for the capacity to contribute novel metabolic capabilities to the group as a whole. The specific Jaccard's distances at which the optimum occurs are likely dependent on the specific topology of the underlying reaction networks utilized, and may not be universal (see Methods). However, in our synergy calculations based on elementary flux modes, we retain a similar degree distribution and reaction topology as would be seen in the bacterial species of the oral biofilm. It will be interesting to explore the possibility that the general shape of the curve observed in Fig. 5 could be derived analytically.

Finally, our finding poses an interesting evolutionary chicken and egg dilemma: did the observed metabolic proximity pattern precede or follow the emergence of specific binding affinities between receptors and ligands across species? On one hand, energy and food-related requirements may be hypothesized to dictate the emergence of a biofilm structure. Subsequent adaptations could have optimized inter-species binding interactions to facilitate the formation of an efficient nutrient and energy flow. Conversely, we cannot rule out the alternative possibility that metabolic proximity may have arisen between organisms with a tendency to bind to each other, e.g. through horizontal gene transfer, or by forcing each other to face specific selective pressures. This may be an interesting challenge for future research, in which the experimental investigation of evolving symbiotic system [48] could be complemented by computational studies of evolutionary rates in genomic sequences [49–51].

Methods

Parsing an Experimental Map of the Oral Biofilm Structure

The present analysis uses the biofilm organization map that is presented in [12], which is based on the collection of several individual experimental papers. This model (to which we will refer as the Kolenbrander model) contains 22 organisms, 11 of which were sequenced and annotated [27–30] at the time of our analysis. The 11 organisms for which data is available are listed in Fig. 1A. The map also includes known connections between some organisms and the salivary pellicle, from which we assume that the biofilm starts developing. In the Kolenbrander model, reproduced in simplified form in Fig. 1A, nodes correspond to species (with the exception of *Veionella* and *Eubacterium*, which were reported only at the genus level). Organisms for which there was no KEGG data were omitted from our analysis. An edge between two nodes in the network of Fig. 1A denotes a documented capacity of the two corresponding biofilm constituents to bind to each other.

Calculation of Enzyme Based Distances

The KEGG database contains information that describes the number and type of enzymes present in an organism's genome [27–30]. Each enzymatic function is associated with an Enzyme Commission (EC) number [52]. In this case, we are not interested in the abundance of any particular enzyme or in its substrate/product stoichiometry, but simply in the presence or absence of such enzyme in a given organism's genome. Given this information, a binary vector $\boldsymbol{S^{(A)}}$ can be defined to describe the enzyme composition for an organism A, with component $S^{(A)}{}_i = 1$ if enzyme i is present in organism A, and $S^{(A)}{}_j = 0$ otherwise. We then evaluate the difference between the metabolic profiles of two organisms by computing the Jaccard's distance $\mathcal{J}(A,B)$ between $\boldsymbol{S^{(A)}}$ and $\boldsymbol{S^{(B)}}$, defined as follows:

$$J(A,B) = 1 - \frac{|\boldsymbol{S^{(A)}} \cap \boldsymbol{S^{(B)}}|}{|\boldsymbol{S^{(A)}} \cup \boldsymbol{S^{(B)}}|}$$

If A and B have the same metabolic enzymes, then $\mathcal{J} = 0$. If they have no enzyme in common, it is $\mathcal{J} = 1$. This metric will also be used to quantify metabolic similarity between species, in the form of the Jaccard's coefficient ($\mathcal{J}_c = 1 - \mathcal{J}$) [53].

Calculation of Non-metabolic Distances

In addition to enzyme content, the KEGG database includes data on the presence of different categories of non-metabolic genes. Using this data we can generate binary vectors $\boldsymbol{S^{(A)}}$ and $\boldsymbol{S^{(B)}}$, that represent the non-metabolic gene content for organisms A and B, just as was done for metabolic distance. We then evaluate the difference between the non-metabolic profiles of two organisms by computing their Jaccard's distance $\mathcal{J}(A,B)$.

Calculation of Added Information

All organisms in the community contribute to an overall super-set of enzymes that represents the metabolic potential of the community. In examining the gradual build-up of the oral biofilm, we can ask how much novelty is introduced by each new organism joining an increasingly complex ecosystem. This can be achieved by calculating the amount of metabolic information added to the current super-set upon introducing a new organism to the biofilm. If we call N_i the number of organisms in which enzyme i is present ($N_i \in \{1, 2, \ldots, N_{organisms}\}$, with $N_{organisms} = 11$), then the probability to find a given enzymatic function i in the whole biofilm is $P_i = N_i / N_{organisms}$. For an ordered pair of organisms (A,B), we can identify the set $K(A,B)$ of all EC numbers k such that $S^{(A)}{}_k = 0$ and $S^{(B)}{}_k = 1$. The information added when organism B is added to organism A is then computed as the Shannon information content of all enzymes that are currently added and were not present in the prior organisms, i.e.:

$$\Delta I(A,B) = -\sum_{k \in K(A,B)} P_k \ln P_k$$

$\Delta I(A,B)$ represents the amount of Shannon information (relative to the overall abundance of EC numbers in the entire biofilm), added by an organism B upon colonization on top of an organism A. Note that ΔI is not symmetric, i.e., in general, $\Delta I(A,B) \neq \Delta I(B,A)$. Hence, this added information metric allows us to distinguish between orders of colonization that would be indistinguishable using the Jaccard's metric $\mathcal{J}$ defined above.

Layer Specific Pathway Enrichment

To determine possible layer-specific metabolic pathway enrichment in the oral biofilm, we first calculate the proportion of organisms in a given layer that contain a given enzyme. This calculation uses the same organism-specific binary vector $\boldsymbol{S^{(A)}}$ defined above. If the set of organisms in biofilm layer x is defined

as L_x, then a biofilm layer profile ($\boldsymbol{B}^{(x)}$), describing the occurrence of each enzyme in any given organism in layer x can be calculated as follows:

$$\boldsymbol{B}^{(x)} = \frac{\sum_{i \in L_X} \boldsymbol{S}^{(i)}}{|L_x|}$$

These profiles can be used to estimate the enrichment of the different layers for specific metabolic processes. This is achieved by using gene set enrichment analysis (GSEA). In addition to the $\boldsymbol{B}^{(x)}$ profiles for the four different layers (Fig. 1A), the GSEA algorithm utilizes a binary mapping matrix $K_{(i,j)}$, where $K_{(i,j)} = 1$ if enzyme i is present in pathway j. This matrix maps enzymes to corresponding KEGG metabolic pathways. We next look for specific enzyme and pathway enrichment in a given layer relative to other layers within the dental biofilm. The enrichment calculation is performed using a standard GSEA application [54]. Pathways with a nominal p-value of 0.05 or less and an FDR of less than 0.25 were chosen (with FDR accounting for multiple testing biases).

Construction of Randomized Paired Networks

In order to estimate the metabolic benefit derived from the cooperation of two species, we perform an analysis of elementary flux modes in appropriately modified versions of the *E. coli* FBA model [21]. In particular, we used the *E. coli* FBA model (cytoplasm reactions only) as an initial main network (of size R_{TOT}) to generate random metabolic networks with a degree distribution and topology similar to that of real metabolic networks. Random networks are generated in pairs, with a specified metabolic similarity (Jaccard's coefficient, J_c) between them. The algorithmic pipeline to generate such pairs of networks proceeds as follows:

(i) Out of the main source network of size R_{TOT}, we choose a subnetwork (the "source network") of size $R_s = P \times R_{TOT}$, where P is a given percent coverage, chosen in order to guarantee tractability of the elementary flux modes calculations. The standard value used throughout this work is P = 0.8.

(ii) Given the desired degree of metabolic overlap (J_c) between the two networks, and the size of the source network (R_s), we compute the size R_{int} of the set N^{int} of reactions that should be in common between the two networks. This number is simply

$$Rint = Jc \times Rs$$

(iii) We select two random sets (N^1, N^2) of non-overlapping reactions from the source network. Each of these reaction sets is chosen to have size

$$Rn = (Rs - Rint)/2$$

(iv) We use the sets of reactions N^{int}, N^1 and N^2 to build reaction sets for the two desired randomized networks, and for their union. These reaction sets for individual networks ($C^{(1)}$, $C^{(2)}$) and for their joint combination ($C^{(1,2)}$) are defined as follows:

$$C^{(1)} = N^1 \cup N^{int}$$

$$C^{(2)} = N^2 \cup N^{int}$$

$$C^{(1,2)} = N^1 \cup N^2 \cup N^{int}$$

(v) The reaction sets $C^{(1)}$ $C^{(2)}$ and $C^{(1,2)}$ are mapped to stoichiometric matrices M^1 and M^2 and $M^{(1,2)}$ respectively. When generating the combined stoichiometric matrix $M^{(1,2)}$, we need to make sure that no additional overlap (in addition to the chosen N^{int} reactions) is introduced between the two random networks. This is achieved by ascribing new names (i.e. new stoichiometric matrix rows) to the metabolites contributed by N^1 and N^2 which are not already present in N^{int}.

Computation of Degree of Metabolic Synergy Using Elementary Flux Modes

For the stoichiometric matrices described in the previous paragraph, the number of elementary flux modes (EFM) is calculated using the *efmtool* software [38]. For each stoichiometric matrix passed to it, *efmtool* returns an EFM matrix describing the various elementary modes through the network. The number of columns of the matrix corresponds to the number of EFMs possible for the given network. A normalized score estimating the increase in the number of EFMs obtained upon conjoining two networks can be computed as follows:

$$\Delta EFM = \frac{EFM(1,2)}{EFM(1) + EFM(2)}$$

Where $EFM(1,2)$ is the number of elementary flux modes from the joint network as defined by $M^{(1,2)}$, while $EFM(1)$ and $EFM(2)$ represent the numbers of elementary flux modes generated by the constituent randomized stoichiometric matrices M^1 and M^2 respectively. ΔEFM represents the increase in the number of flux modes, relative to the constituent networks.

Supporting Information

Figure S1 Sensitivity analysis of our metabolic approach for recapitulating the order of colonization, upon gradual removal of information. The distributions of pairwise metabolic distances for correct (literature-informed) and randomized orders of colonization are plotted for different percentages of enzymes removed from the dataset. Between 20 and 80 percent of enzymes were removed.
(PDF)

Dataset S1 This file contains tables listing the presence and absence of different KEGG metabolic pathways within the 11 organisms used in this study.
(XLSX)

Acknowledgments

We are grateful to members of the Segrè lab for helpful feedback and discussions.

Author Contributions

Conceived and designed the experiments: VM SA DS. Performed the experiments: VM. Analyzed the data: VM DS. Wrote the paper: VM DS.

References

1. Stoodley P, Sauer K, Davies DG, Costerton JW (2002) Biofilms as complex differentiated communities. Annual review of microbiology 56: 187–209.
2. Donlan RM (2002) Biofilms: microbial life on surfaces. Emerging infectious diseases 8: 881–890.
3. Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner ACR, et al. (2010) The Human Oral Microbiome. Journal of Bacteriology 192: 5002–5017.
4. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Gordon JI (2010) Metagenomic Analysis in Humanized Gnotobiotic Mice. 1: 1–19.
5. Klitgord N, Segre D (2010) Environments that Induce Synthetic Microbial Ecosystems. PLoS Computational Biology 6: e1001002.
6. Borenstein E, Kupiec M, Feldman MW, Ruppin E (2008) Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proceedings of the National Academy of Sciences 105: 14482–14487.
7. Kauffman KJ, Prakash P, Edwards JS (2003) Advances in flux balance analysis. Curr Opin Biotechnol 14: 491–496.
8. Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, Romero H, Simón-Soro A, et al. (2011) The Oral Metagenome in Health and Disease. The ISME journal: 1–11.
9. Diaz-Torres ML, Villedieu A, Hunt N, McNab R, Spratt DA, et al. (2006) Determining the antibiotic resistance potential of the indigenous oral microbiota of humans using a metagenomic approach. FEMS microbiology letters 258: 257–262.
10. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic Analysis of the Human Distal Gut Microbiome. Science 312: 1355–1359.
11. Turnbaugh P, Ley R, Hamady M, Fraser-Liggett C, Knight R, et al. (2007) The Human Microbiome Project. Nature 449: 804–810.
12. Kolenbrander PE, Palmer RJ, Periasamy S, Jakubovics NS (2010) Oral Multispecies Biofilm Development and the Key Role of Cell-cell Distance. Nature reviews Microbiology 8: 471–480.
13. Rickard AH, Gilbert P, High NJ, Kolenbrander PE, Handley PS (2003) Bacterial coaggregation: an integral process in the development of multi-species biofilms. Trends in Microbiology 11: 94–100.
14. Kolenbrander PE, Andersen RN, Blehert DS, Egland PG, Foster JS, et al. (2002) Communication among oral bacteria. Microbiology and molecular biology reviews 66: 486–505.
15. Valm AM, Welch JLM, Rieken CW, Hasegawa Y, Sogin ML, et al. (2011) Systems-level analysis of microbial community organization through combinatorial labeling and spectral imaging. Proceedings of the National Academy of Sciences 108: 4152–4157.
16. Kolenbrander PE, Andersen RN (1986) Multigeneric aggregations among oral bacteria: a network of independent cell-to-cell interactions. Journal of bacteriology 168: 851–859.
17. Mazumdar V, Snitkin ES, Amar S, Segrè D (2009) Metabolic Network Model of a Human Oral Pathogen. Journal of Bacteriology 191: 74–90.
18. Li L, Guo L, Lux R, Eckert R, Yarbrough D, et al. (2010) Targeted antimicrobial therapy against Streptococcus mutans establishes protective non-cariogenic oral biofilms and reduces subsequent infection. International journal of oral science 2: 66.
19. Palmer Jr RJ, Wu R, Gordon S, Bloomquist CG, Liljemark WF, et al. (2001) [27] Retrieval of biofilms from the oral cavity. Methods in enzymology 337: 393–403.
20. Becker SA, Palsson BO (2005) Genome-scale Reconstruction of the Metabolic Network in *Staphylococcus aureus* N315: an Initial Draft to the Two-dimensional Annotation. BMC Microbiology 5: 8.
21. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, et al. (2007) A Genome-scale Metabolic Reconstruction for *Escherichia coli* K-12 MG1655 that Accounts for 1260 ORFs and Thermodynamic Information. Molecular Systems Biology 3: 121.
22. Forster J, Famili I, Fu P, Palsson BO, Nielsen J (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. Genome Res 13: 244–253.
23. Nogales J, Palsson B, Thiele I (2008) A Genome-scale Metabolic Reconstruction of Pseudomonas putida KT2440: iJN746 as a Cell Factory. BMC Systems Biology 2: 79.
24. Oliveira AP, Nielsen J, Forster J (2005) Modeling *Lactococcus lactis* Using a Genome-scale Flux Model. BMC Microbiology 5: 39.
25. Senger RS, Papoutsakis ET (2008) Genome-scale Model for *Clostridium acetobutylicum*: Part I. Metabolic Network Resolution and Analysis. Biotechnology and Bioengineering: 1036–1052.
26. Thiele I, Vo TD, Price ND, Palsson BO (2005) Expanded metabolic reconstruction of *Helicobacter pylori* (iIT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants. Journal of bacteriology 187: 5818–5830.
27. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for Representation and Analysis of Molecular Networks Involving Diseases and Drugs. Nucleic acids research 38: D355–360.
28. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, et al. (2006) From Genomics to Chemical Genomics: New Developments in KEGG. Nucleic Acids Research 34: D354–357.
29. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.
30. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29–34.
31. Massey Jr FJ (1951) The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association: 68–78.
32. Lendenmann U, Grogan J, Oppenheim F (2000) Saliva and dental pellicle-a review. Advances in Dental Research 14: 22–28.
33. Singer RE, Buckner BA (1981) Butyrate and propionate: important components of toxic dental plaque extracts. Infection and Immunity 32: 458–463.
34. Tsuda H, Ochiai K, Suzuki N, Otsuka K (2010) Butyrate, a Bacterial Metabolite, Induces Apoptosis and Autophagic Cell Death in Gingival Epithelial cells. Journal of Periodontal Research 45: 626–634.
35. Abe N, Kadowaki T, Okamoto K, Nakayama K, Ohishi M, et al. (1998) Biochemical and functional properties of lysine-specific cysteine proteinase (Lys-gingipain) as a virulence factor of *Porphyromonas gingivalis* in periodontal disease. Journal of biochemistry 123: 305–312.
36. Olczak T, Simpson W, Liu X, Genco CA (2005) Iron and Heme Utilization in *Porphyromonas gingivalis*. FEMS Microbiology Reviews 29: 119–144.
37. Ley RE, Peterson DA, Gordon JI (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell 124: 837–848.
38. Terzer M, Stelling J (2008) Large-scale Computation of Elementary Flux Modes with Bit Pattern Trees. Bioinformatics 24: 2229–2235.
39. Roberts RJ, Chang YC, Hu Z, Rachlin JN, Anton BP, et al. (2011) COMBREX: a project to accelerate the functional annotation of prokaryotic genomes. Nucleic acids research 39: D11–D14.
40. Klitgord N, Segrè D (2010) Environments that induce synthetic microbial ecosystems. PLoS computational biology 6: e1001002.
41. Fu W, Forster T, Mayer O, Curtin JJ, Lehman SM, et al. (2010) Bacteriophage cocktail for the prevention of biofilm formation by Pseudomonas aeruginosa on catheters in an in vitro model system. Antimicrobial agents and chemotherapy 54: 397–404.
42. Hall-Stoodley L, Costerton JW, Stoodley P (2004) Bacterial biofilms: from the natural environment to infectious diseases. Nature Reviews Microbiology 2: 95–108.
43. Hawser SP, Douglas LJ (1994) Biofilm formation by Candida species on the surface of catheter materials in vitro. Infection and Immunity 62: 915–921.
44. Nickel J, Ruseska I, Wright J, Costerton J (1985) Tobramycin resistance of *Pseudomonas aeruginosa* cells growing as a biofilm on urinary catheter material. Antimicrobial agents and chemotherapy 27: 619–624.
45. Canfield DE, Des Marais DJ (1991) Aerobic sulfate reduction in microbial mats. Science (New York, NY) 251: 1471.
46. Ward DM, Ferris MJ, Nold SC, Bateson MM (1998) A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. Microbiology and Molecular Biology Reviews 62: 1353–1370.
47. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, et al. (2011) Enterotypes of the human gut microbiome. Nature 473: 174–180.
48. Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, et al. (2007) Evolution of symbiotic bacteria in the distal human intestine. PLoS biology 5: e156.
49. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, et al. (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461: 1243–1247.
50. Boussau B, Karlberg EO, Frank AC, Legault BA, Andersson SGE (2004) Computational inference of scenarios for α-proteobacterial genome evolution. Proceedings of the National Academy of Sciences of the United States of America 101: 9722–9727.
51. Ochman H, Elwyn S, Moran NA (1999) Calibrating bacterial evolution. Proceedings of the National Academy of Sciences 96: 12638–12643.
52. Webb EC (1992) Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes: Academic Press.
53. Jaccard P (1908) Nouvelles recherches sur la distribution florale.
54. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene Set Enrichment Analysis: a Knowledge-based Approach for Interpreting Genome-wide Expression Profiles. Proceedings of the National Academy of Sciences of the United States of America 102: 15545–15550.