RESEARCH

Open Access

Nutrition or nature: using elementary flux modes to disentangle the complex forces shaping prokaryote pan-genomes

Daniel R. Garza[1,2*], F. A. Bastiaan von Meijenfeldt[3], Bram van Dijk[4], Annemarie Boleij[5], Martijn A. Huynen[1] and Bas E. Dutilh[1,6,7]

Abstract

Background: Microbial pan-genomes are shaped by a complex combination of stochastic and deterministic forces. Even closely related genomes exhibit extensive variation in their gene content. Understanding what drives this variation requires exploring the interactions of gene products with each other and with the organism's external environment. However, to date, conceptual models of pan-genome dynamics often represent genes as independent units and provide limited information about their mechanistic interactions.

Results: We simulated the stochastic process of gene-loss using the pooled genome-scale metabolic reaction networks of 46 taxonomically diverse bacterial and archaeal families as proxies for their pan-genomes. The frequency by which reactions are retained in functional networks when stochastic gene loss is simulated in diverse environments allowed us to disentangle the metabolic reactions whose presence depends on the metabolite composition of the external environment (constrained by "nutrition") from those that are independent of the environment (constrained by "nature"). By comparing the frequency of reactions from the first group with their observed frequencies in bacterial and archaeal families, we predicted the metabolic niches that shaped the genomic composition of these lineages. Moreover, we found that the lineages that were shaped by a more diverse metabolic niche also occur in more diverse biomes as assessed by global environmental sequencing datasets.

Conclusion: We introduce a computational framework for analyzing and interpreting pan-reactomes that provides novel insights into the ecological and evolutionary drivers of pan-genome dynamics.

Keywords: Pan-genome evolution, Reactomes, Genome-scale metabolic models, Prokaryote evolution, Gene frequency distribution

Background

In the evolution of microbial genomes, genes are gained and lost by mutations, insertions, deletions, duplications, and horizontal gene transfers (HGTs) [1–4]. As a result of these processes, gene content varies significantly even between closely related genomes [3, 5, 6]. Diverse gene repertoires give rise to microbial pan-genomes, which are defined as the complete set of non-redundant genes harbored by any monophyletic group of microbes [7]. The genes in a pan-genome exhibit a frequency distribution that can be estimated by comparing the gene content of many representative genomes from the same taxonomic level. Some genes are found in all representative genomes (called "core genes"), in only one or a few genomes (called "cloud genes"), or in an intermediate fraction of the genomes (called "shell genes"). The frequency

*Correspondence: danielriosgarza@gmail.com

[1] Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Centre, Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands
Full list of author information is available at the end of the article

distribution of genes likely reflects the deterministic and stochastic drivers of evolution [8], but currently, there are insufficient theoretical frameworks that allow us to relate the empirical frequency of genes in pan-genomes with the ecologic and evolutionary processes that shape microbial genomes [9, 10].

Recent studies have attempted to transfer the concepts from population genetics to interpret the frequency distribution of genes in microbial pan-genomes [7, 8, 10, 11]. This has been justified since pan-genomes, instead of populations, can be viewed as the key units of prokaryote evolution [3]. But the frequency of genotypes used by population genetics theory reflects different evolutionary forces than the frequency of genes in pan-genomes. While the former reflects mutation rates and effective population sizes, the latter reflects HGT rates and genomic adaptations to diverse environmental conditions [5, 12–16]. Currently, there is a need to develop more realistic frameworks to model and explain gene frequency in pan-genomes.

An important factor to consider when developing models for the distribution of genes in pan-genomes is that individual genes are not equally accessible to all genomes. In general, the probability that a recipient microbial genome will be capable of integrating foreign DNA increases exponentially with an increase in the similarity between the donor DNA and the recipient chromosome [17]. As a result, closely related genomes share more genes than distantly related ones [6, 18].

Another important factor to consider is that the acquisition of new genes by genomes is counterbalanced by the frequent loss of genes [3, 19–21]. Gene loss is majorly a clock-like process, where genes under weak or no selection tend to be inactivated by random mutation and lost by deletion [19, 22, 23]. This process is widely observed across microbial genomes and virtually all species with genomes smaller than 2 Mb evolved from ancestors with substantially larger genomes [23–25]. Gene loss is also the major source of genomic variation of intracellular parasites that do not undergo extensive HGT [26] and of bacteria that are adapted to stable and nutrient-rich environments, such as host-associated microbiomes [24]. Based on the high rates of gene loss, it is reasonable to assume that genes that are not under selective pressure may eventually be lost.

The frequency distributions of the genes within pan-genomes fit mathematical functions with regular and universal shapes [9, 27–29]. One example is the asymmetric U-shape that is observed for a broad range of prokaryote groups [30, 31]. Under this distribution, core and cloud genes are more frequent compared to shell genes. Pan-genome studies commonly conflate the frequency of a gene with its essentiality. In this view, core genes are considered essential under any condition, while the increasingly rare genes are considered increasingly dispensable. But the characteristic U-shape distribution of gene frequency also emerges from simple neutral models that do not attribute different selective advantages to different genes [32–34].

The commonly used neutral and non-neutral models of pan-genome evolution are simplified 'bag-of-genes' models that do not explicitly consider gene functions and their interactions [9, 32, 34]. These models provide important insights into the evolutionary dynamics of pan-genomes, but ignore the functional forces driving gene frequencies and, more importantly, do not provide a mechanistic interpretation for the variation in gene content. In nature, selection acts on the phenotype, and microbes exhibit complex phenotypes that result from the combined action of multiple gene products. In many cases, phenotypes are dependent on the interactions of gene products with the environment.

A promising approach to integrating the functions and interactions of genes into models of genome evolution is to use the genome-encoded metabolic reactions, the reactome, as a proxy for the gene content of genomes. Reactions from the reactome can be integrated into a functional network (also referred to as a genome-scale metabolic model (GSMM)) that represents the genotype-to-phenotype map [35–38]. The molecular components (protein-encoding genes) of reactomes are readily inferred from microbial genome sequences [36, 39, 40]. Similar to pan-genomes, pan-reactomes can be defined as the complete set of non-redundant metabolic reactions that are harbored by a monophyletic microbial group. The pan-reactome may also be subdivided into the core, cloud, and shell pan-reactomes based on reaction frequency. Networks derived from pan-reactomes are capable of simulating complex phenotypes, such as the conversion of energy and matter from diverse environmental metabolites into sugars, nucleic acids, lipids, and proteins [36].

Here we used pan-reactomes as models to simulate and understand patterns in pan-genomes. To incorporate realistic features of pan-genome evolution, we used pan-reactomes as proxies for pools of genes that are accessible to related strains by HGT. We then modeled alternative routes of gene loss by sampling minimal functional reaction sets in diverse environment compositions. These minimal functional reaction sets have similar properties as the previously defined elementary flux modes (EFMs) used to identify functional pathways in reactomes [41], thus, we termed our sets panEFMS (pan-reactome elementary flux modes) and used them to distinguish two important drivers of reaction frequencies in pan-reactomes, which we refer to as 'nutrition' and 'nature'.

The frequency of reactions that are driven by nutrition depends on the environment composition, while the frequency of reactions that are driven by nature does not depend on the environment composition. Our framework mechanistically disentangles environment-driven from environment-independent reactions and uses their distribution in panEFMs to build a model that predicts the metabolite preferences of pan-reactomes from their environment-driven reactions. We applied this model to the pan-reactomes of 46 bacterial and archaeal families, allowing us to assess the patterns of microbial genome evolution that result from the function and interaction between metabolic genes.

Results

Functional reactomes of individual strains are samples from a pan-reactome

Our model consists of directed bipartite graphs of reactions and metabolites (Fig. 1A) derived from reactomes of related organisms (Fig. 1B) that together form a pan-reactome (Fig. 1C). To illustrate the model in a tractable way, we use a toy model to illustrate how environment-independent and environment-driven processes together shape the frequency distribution of reactions in pan-reactomes (Fig. 1, Additional file 6: Table S1). We expand the detailed explanation of the model to the pan-reactome of the *Aeromonadaceae* family that was built based on the metabolic reactions encoded in the genomes of 135 strains that belong to different *Aeromonadaceae* species (Additional file 7: Table S2). This family was chosen as an illustrative example applied to a natural pan-reactome in contrast to the artificial reactions found in the toy model. The same analysis was performed for the pan-reactomes of forty-six other bacterial and archaeal families (Additional file 7: Table S2, see further results sections). The *Aeromonadaceae* pan-reactome network contains 1796 reactions and 292 environment compounds (Additional file 8: Table S3). These reactomes take up compounds from the external environment (MX_e metabolites in the toy model in Fig. 1) and convert them to synthesize metabolites required for biomass production (M10_i, M11_i, and M12_i). As described in the Introduction, evolutionary processes sample subsets of reactions from this pan-reactome to generate reactomes of individual strains. Strain reactomes are considered functional if they can synthesize all the biomass compounds and do not accumulate by-products. Metabolite M2 in the first reactome of Fig. 1B is an example of a byproduct that needs to be exported by reaction R10. Three examples of viable reactomes are shown in Fig. 1B and others can be formed. For the family *Aeromonadaceae*, significantly more viable reactomes can be formed than those of the 135 sequenced strains (Additional file 7: Table S2), which together defined the family-level pan-reactome. Thus, a pan-reactome defines a space of potential functional reactomes, some of which are realized by actual strains in nature.

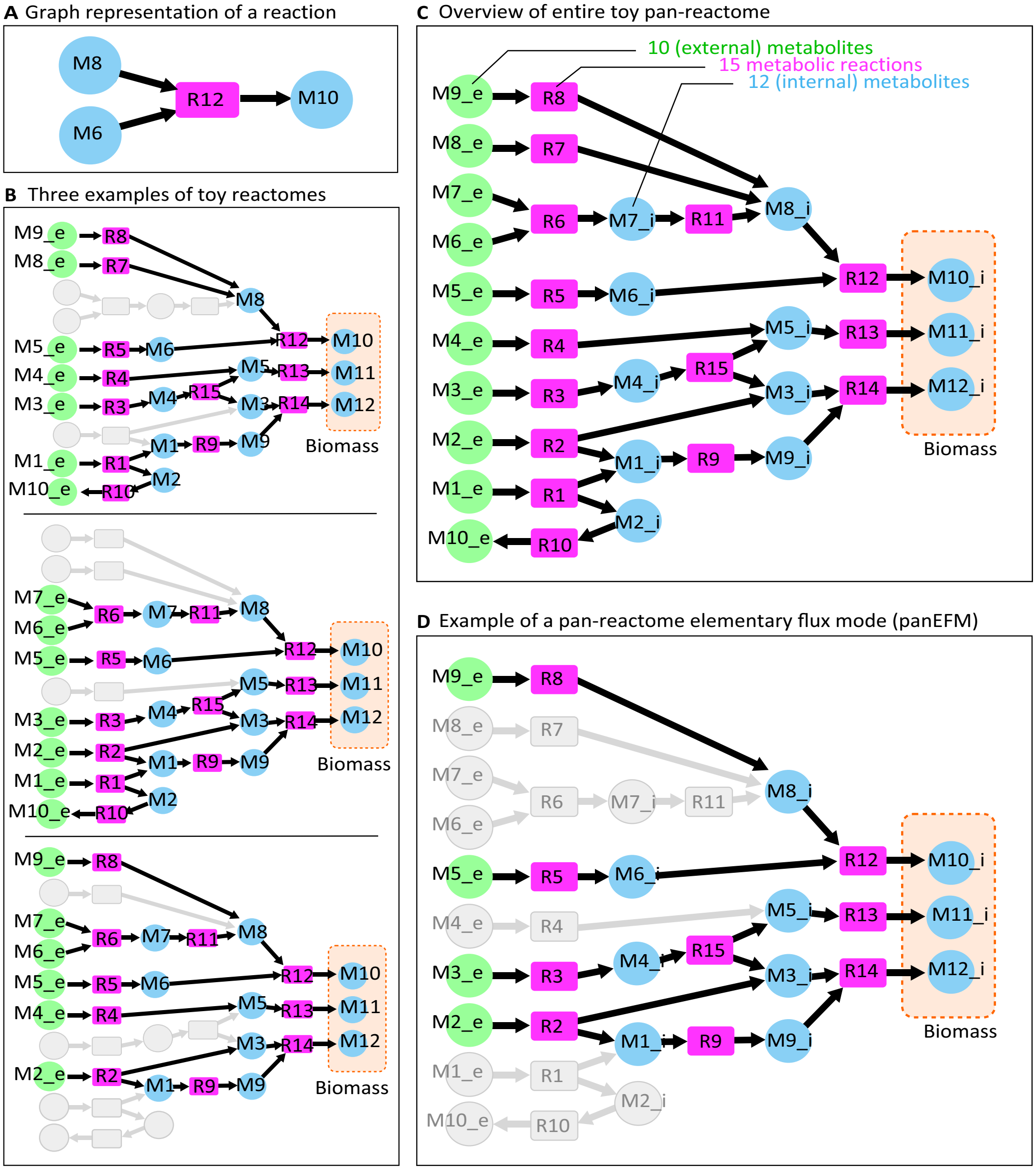
While evolutionary processes constrain the functional reactomes that can be sampled from a pan-reactome pool, the reactions that are selected in practice depend on the environment (environment-driven reactions, ‘nutrition’) and the structure of the metabolic network (environment-independent reactions, ‘nature’). Using the toy model as an example, in Fig. 1C reactions R5, R9, R12, R13, and R14 are environment-independent reactions that are always required for biomass production, they are irreplaceable in the synthesis of essential biomass precursors and are thus necessarily present in each functional reactome. In complex natural networks, some environment-independent reactions may also depend on the presence of other reactions in the network, as we will see below. In contrast, the network may synthesize the biomass precursor M8_i by either using reactions R8, R7, or a combination of R6 and R11. In principle these three metabolic routes are equivalent, but the presence of external metabolites determines which ones are functional and these reactions are considered environment-driven.

Pan-reactome elementary flux modes (panEFMs) predict reaction frequencies in the pan-reactome

To explore the space of possible reactomes within a group of organisms that comprise a pan-reactome, we modelled functional reactomes under the hypothesis that evolution tends to lose non-essential reactions. A single panEFM consists of a group of reactions that together are functional in a defined environment, but the removal of any reaction makes the network non-functional. An example of a panEFM is shown in Fig. 1D (set S1). The number and composition of panEFMs depend on the

(See figure on next page.)

Fig. 1 Toy model. **A** Example of a metabolic reaction. Reactants and products are depicted as circles and reactions as rectangles, respectively. Reaction directionality is indicated by the arrows. **B** Three functional reactomes derived from the toy model, each capable of synthesizing the biomass compounds M10_i, M11_i, and M12_i from the environmental precursors (‘MX_e’ compounds depicted with green circles). **C** Pan-reactome network aggregates reactions from the different reactomes into a single network. The “_e” and “_i” termination of metabolites denote external and internal metabolites, respectively. **D** An example of a panEFM. Each reaction in this network is essential since its removal would impair the synthesis of the biomass components. **E** Collection of all nine possible panEFMs that can be created from the reactions in this toy pan-reactome in a rich environment. Dark squares denote the presence of reactions. The frequency of reactions across the collection of panEFMs is shown in the last row


A Graph representation of a reaction
M8
M6
R12
M10
B Three examples of toy reactomes
M9_e
M8_e
M7_e
M6_e
M5_e
M4_e
M3_e
M2_e
M1_e
M10_e
R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
M11
M12
Biomass
C Overview of entire toy pan-reactome
10 (external) metabolites
15 metabolic reactions
12 (internal) metabolites
M1_i
M2_i
M3_i
M4_i
M5_i
M6_i
M7_i
M8_i
M9_i
M10_i
M11_i
M12_i
Biomass
D Example of a pan-reactome elementary flux mode (panEFM)

E All nine panEFMs that exist in a rich environment (i.e. when all metabolites are available)

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15
S1		■	■		■			■	■			■	■	■	■
S2	■		■		■		■		■	■		■	■	■	■
S3		■		■	■			■	■			■	■	■	
S4		■	■		■		■		■			■	■	■	■
S5		■		■	■	■			■		■	■	■	■	
S6		■		■	■		■		■			■	■	■	
S7	■		■		■	■			■	■	■	■	■	■	■
S8		■	■		■	■			■		■	■	■	■	■
S9	■		■		■			■	■	■		■	■	■	■
Freq.	.33	.66	.66	.33	1.00	.33	.33	.33	1.00	.33	.33	1.00	1.00	1.00	.66

Fig. 1 (See legend on previous page.)

specific environment, with rich environments having more panEFMs. For example, our toy model has nine panEFMs (sets S1–9) in a rich environment where all external metabolites are available. The panEFMs from the pan-reactome allow us to generate an expected frequency of each reaction in the pan-reactome in the context of a defined metabolic environment. For example, given a rich environment, the bottom row of Fig. 1E shows the expected frequency of reactions in the lineage represented by the toy pan-reactome, given that in this model there is no selective advantage of using one pathway over another to synthesize a specific metabolite.

Due to the combinatorial explosion involved in extracting all possible panEFMs from large pan-reactomes, for bacterial and archaeal families we used a random sampling approach to approximate the space of all possible panEFMs. We first determined that fewer than 200 panEFMs sampled in different random environments were sufficient for convergence to an average reaction frequency distribution with 99% reproducibility and a mean-squared error approaching zero (Additional file 1: Fig. S1). To reach this conclusion, we randomly sampled an increasing number of panEFM sets across 1000 random environments, performing such sampling independently twice. We tested if each time the reaction frequency distribution converged to the same values. Depending on the sample size, the frequency distributions asymptotically converge to the same values and are already nearly identical for sample sizes that are greater than 200 (Additional file 1: Fig. S1). To be safe, we used sample sizes of 1000. We thus generated one million panEFMs for each family, including 1000 panEFMs sampled in each of 1000 different random environments (see Methods).

Disentangling the forces driving the frequency of reactions in pan-reactomes

To tease apart the forces that shape pan-reactome composition and infer to what extent the evolution of each reaction is driven by nature versus nutrition, we calculated an environment-driven score (EDS) that ranges from zero (environment-independent, frequency not dependent on the metabolic environment, nature) to one (environment-driven, frequency fully dependent on the metabolic environment, nutrition). First, we calculated the absolute difference between the predicted reaction frequency in one specific virtual environment and its mean frequency across all 1000 random virtual environments (see Methods, cf. last row in Fig. 1E), i.e. the residual. Because the latter value reflects the mean frequency overall, this residual quantifies the extent to which the frequency of a reaction is different in each specific environment. We calculated the EDS for each reaction as the scaled standard deviation of these residuals, as this reflects what extent the reaction varies across specific environments. We illustrate the EDS score in the toy model (Fig. 2A) where reactions R5, R9, R12, R13, and R14 are identified as environment-independent (EDS = 0.0), while the rest are environment-driven (EDS > 0). We used this approach to estimate EDSs for all reactions across forty-six bacterial and archaeal families (Additional file 9: Table S4). For the *Aeromonadaceae* family, we found that the top environment-driven reactions are reactions involved in the degradation of valine, leucine, and isoleucine (Additional file 9: Table S4).

Associating environment-driven reactions to environmental metabolites

The EDS score identifies reactions that are environment-driven. Next, we focused on identifying the specific metabolites that drive these frequencies. Because the frequency of reactions in a pan-reactome depends on environmental metabolites in complex ways, we correlated the predicted reaction frequency across 1000 virtual environments to the metabolite usage frequency across those same environments (see Methods). This association, illustrated in Fig. 2B and C for the toy model, quantifies how external metabolite usage explains the frequency distribution of reactions in pan-reactomes. For instance, the frequency of reaction R8 that produces the biomass precursor M8_i is positively associated with the usage of metabolite M9_e (Fig. 2C) and negatively associated with the usage of the metabolites that enable alternative routes to produce M8_i (i.e. M6_e, M7_e, and M8_e; Fig. 1C). Note that the frequencies of reactions in the pan-reactome may thus reveal metabolite availability in the environment where the lineage evolved.

Predicting metabolite usage preferences from reaction frequencies in pan-reactomes

Reaction frequencies may be readily observed in natural pan-reactomes by comparative genomics. To predict the metabolite preferences of prokaryotic families from these reaction frequencies, we trained an elastic net (EN) model on the reaction frequencies in the collection of panEFMs to predict their metabolite usage profiles across the growth-supporting virtual environments (see sections “Reaction frequencies” and “Elastic net” in Methods). We confirmed the accuracy of the EN model for the toy model and the *Aeromonadeceae* pan-reactome by predicting the metabolic niche of reactomes whose evolution was simulated in a defined environment using a Moran-like process of gain and loss of genes [42] (see “Toy model” in the Methods). The EN model accurately predicted the metabolite usage of resulting lineages of both the toy model

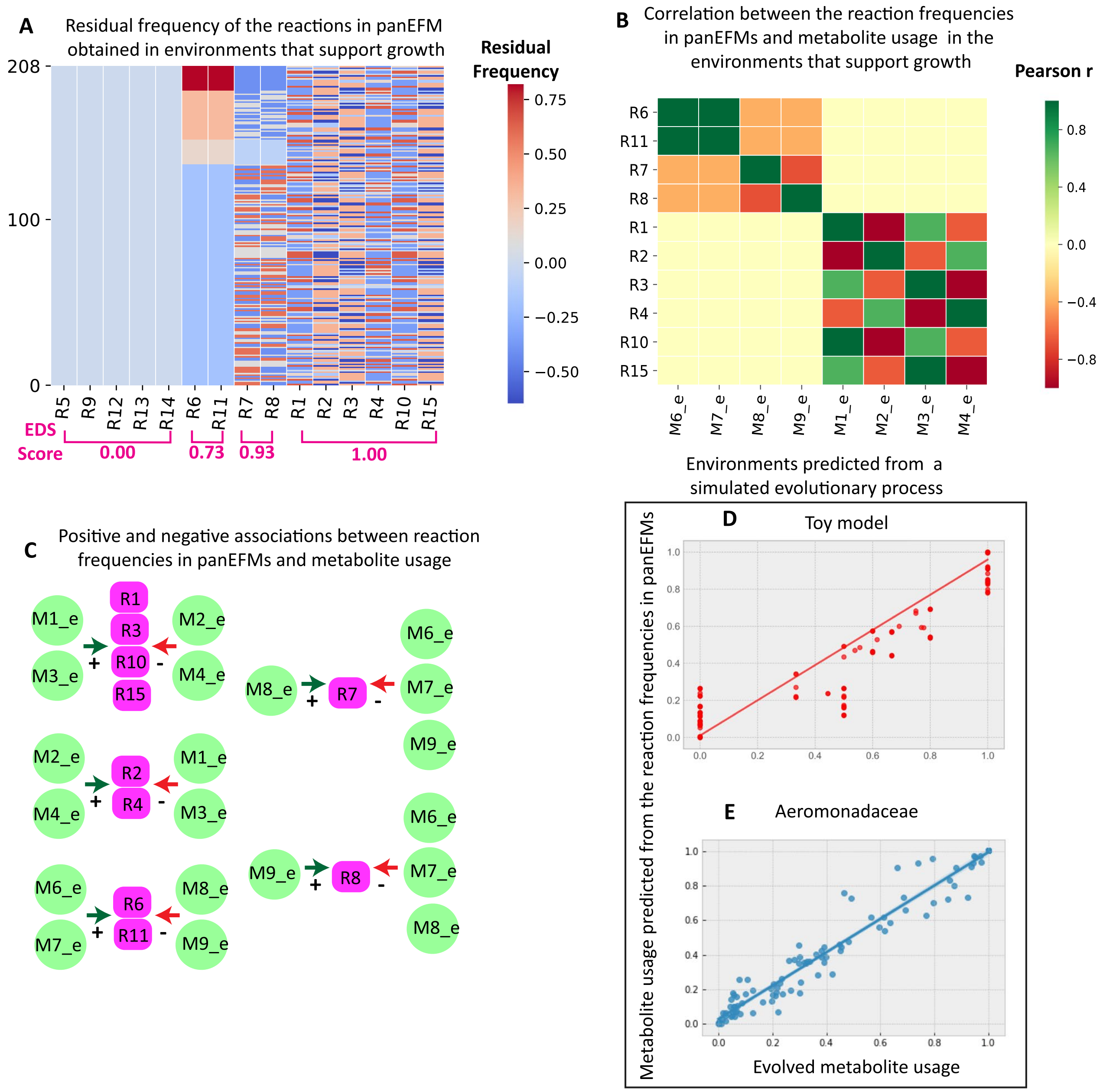


Fig. 2 Identifying environment-reaction associations. **A** Residual reaction frequencies predicted from the collection of panEFMs that exist across the 208 environments that support the growth of the toy pan-reactome (Fig. 1C). The residuals are the difference between the average frequency over the collection of panEFMs defined within each environment and the average across all environments. Reactions along the x-axis are sorted by the environment-driven score (EDS), see text for details; **B** Metabolite-reaction association matrix defined by the pairwise correlation between the metabolites and reactions with non-zeros residuals; **C** Subnetworks generated from the rows of the metabolite-reaction association matrix (shown in **B**), showing the positive (+) and negative (−) associations between metabolite usage and reaction frequencies in panEFMs; **D**, **E** Elastic net prediction of the metabolite usage that evolved in a simulation of a Moran-like process. The evolved reaction frequencies were used to predict how the resulting strains use metabolites in their environments (y-axis) and compared to their usage in the simulated environment (x-axis)

(Figs. 2D, r = 0.98, p < e−10) and the *Aeromonadaceae* pan-reactome (Figs. 2E, r = 0.98, p < e−71). Thus, we were confident that we could use the EN to predict the metabolic niche where a pan-reactome evolved based on the extant frequencies of its environment-driven reactions. As described above, these environment-driven reactions are identified by sampling panEFMs across many different environments, so the metabolite usage cannot be directly inferred from the extant reactomes (networks in Fig. 1B or the 135 *Aeromonadaceae* GSMMs, Additional file 7: Table S2) but requires the intermediate step of sampling panEFMs. We were able to compare evolved metabolite usage with the metabolite usage predicted by the EN because we simulated the evolutionary process in pre-defined environments and could then compute how the evolved reactomes utilize metabolites in these environments. Thus, we proposed an innovative approach to address the elusive question of the preferred metabolic niche of a microbial lineage from the reaction frequencies in its pan-reactome, which in turn can be readily inferred from genome sequences of related strains.

panEFMs delimit the space of possible pan-reactomes

In the following sections, we will use the framework illustrated above for the toy model and the *Aeromonadaceae* pan-reactome to analyze the pan-reactomes of 46 prokaryote families, each containing more than 24 sequenced genomes (Additional file 7: Table S2, see "Pan-reactomes" in Methods).

Reaction frequencies of the collection of panEFMs obtained across random simulated environments reflect an evolutionary landscape of reactomes that could be derived from the family-level pan-reactome pools. Notably, we observed that this landscape exhibited clear family-specific clusters when projected in two dimensions (Fig. 3A). The reaction frequencies of pan-reactomes derived from the sequenced genomes in a family (see "Reaction frequencies" in Methods), here referred to as the natural pan-reactome reaction frequencies (large points in Fig. 3A), were generally found to lie within these clusters, which were composed of the frequency of reactions on panEFMs sampled across random environments. Thus, our approach of sampling a stochastic distribution of pan-reactomes represented the observed (realized) pan-reactome within this evolutionary landscape.

This landscape also reflects a metabolite usage landscape that is based on the frequency that metabolites are used by the sampled panEFMS across random simulated environments. The metabolite usage landscape also exhibits family-specific clusters (Fig. 3B) and the family-specific metabolite usage profiles that were predicted

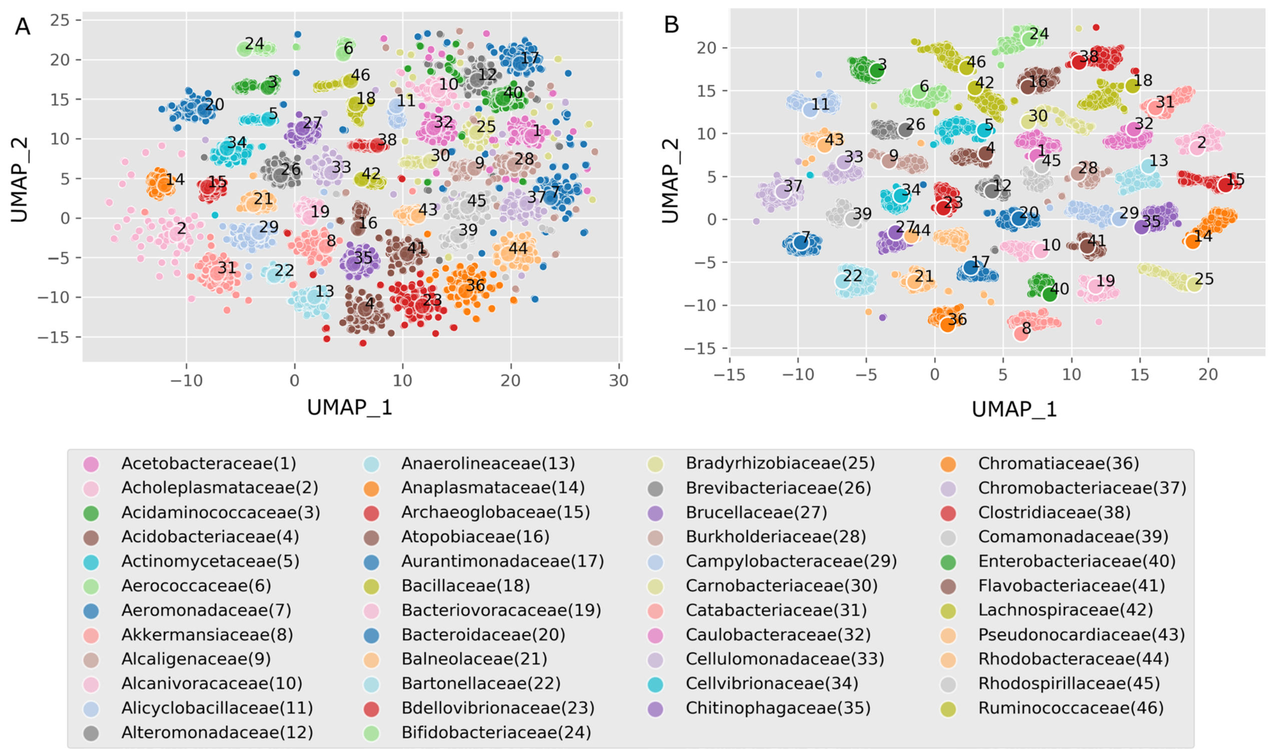


Fig. 3 Evolutionary landscape of possible pan-reactome reaction frequencies and metabolite usage profiles based on sampling panEFMs in 1000 random environments. **A** UMAP projection of reaction frequencies in the collection of panEFMs sampled from different prokaryotic families (Table S2). Each smaller point represents the reaction frequency distribution calculated from 1000 panEFMs sampled in one random environment. The large dots are the frequencies observed in the natural pan-reactomes. **B** UMAP projection of the metabolite usage profiles obtained from the same panEFMs projected in A. The large dots are the elastic net (EN) predictions of these profiles that were predicted from the natural pan-reactomes reaction frequencies. The ENs were trained on the sampled panEFMs (Table S5)

with an EN model (as explained above for the toy model and *Aeromonadaceae* pan-reactome) also behave as observed values (realizations) of the stochastic distribution covered by this evolutionary landscape (large points in Fig. 3B).

The observed strong separation of prokaryotic families reflects family-specific differences in reaction content and metabolite preferences. A bias may also be expected from the way we sampled panEFMs by defining strict family-specific pan-reactome reaction pools. In reality, reaction (gene) pools are not strictly confined to a family and horizontal gene transfer between different families may alleviate the separation between family-level pan-reactomes, although we note that all reactions observed within the sequenced strains of a family were already included in the pan-reactome definition.

To get a better idea of how the reaction frequencies of panEFMS sampled in random virtual environments compare to the natural reaction frequencies, we compared their distributions using kernel density plots (Additional file 2: Fig. S2). Although panEFMs were sampled in random virtual environments, we found that reactions with a high frequency in panEFMs are often universal among the strains of a pan-reactome (Dense regions in the top right corner of Additional file 2: Fig. S2). In contrast, there is significant variability in the frequency of reactions that are rare among panEFMs (averages usually close to 0.50, see Additional file 2: Fig. S2). Thus, our computational approach to sampling panEFMs captures at least some of the dynamics of the natural pan-genomes, which we summarized by the following scenario: (i) Reactions with high frequency in panEFMs are environment-independent (EDS = 0, Additional file 3: Fig. S3, average of 6.6(±2) % of total reactions) and universally essential (Additional file 2: Fig. S2) since across many environments it is statistically unlikely for evolution to form functional reactomes without them; (ii) Reactions with intermediate frequency in panEFMs are variable in pan-reactomes and are enriched for environment-driven reactions (EDS > 0, Additional file 3: Fig. S3, average of 50.5(±6) % of total reactions) since only in a fraction of the sampled environments it is statistically unlikely for evolution to form functional reactomes without them; (iii) Reactions with low frequency in panEFMs have high variability in their frequencies that are usually distributed as a U-shape (Additional file 2: Fig. S2, average of 48.9 (±6) % of total reactions). Their presence or absence are not captured by the panEFMs and their frequency could very well behave as the "bag of genes" models explained in the introduction [9, 32–34, 43], although we did not explore this any further.

Predicting metabolite-reaction associations in the pan-reactomes of 46 prokaryote families

Similar to what we observed for the toy model and the *Aeromonadaceae* pan-reactome, most reactions have a similar predicted frequency in panEFMs across all environments (points that fall in the dense diagonal region of Additional file 4: Fig. S4, average Pearson $r^2 = 0.99$; $p < e-10$), with some reactions exhibiting a significant environment-driven variation in frequency, quantified by the EDS and illustrated by the points that fall outside of dense diagonal line in Additional file 4: Fig. S4. To illustrate, we identified the two reactions that had the highest average EDS scores across all prokaryote families (Additional file 9: Table S4): Dihydroxy hydrolase (EC 4.2.1.9) and pyruvate decarboxylase (EC 2.2.1.6). Both reactions catalyze steps in the synthesis of the three branched-chain amino acids (L-isoleucine, L-valine, and L-leucine) and are universally shared across bacterial reactomes [44]. These reactions are also part of the pantothenate and coenzyme A (CoA) biosynthesis pathway, where the product of the dihydroxy hydrolase (3-Methyl1-2-oxobutanoic acid) can either be used for the synthesis of L-valine or the synthesis of 2-Dehydropantoate, a precursor for pantothenate and subsequently CoA. Pantothenate and CoA are connected to the biosynthesis of several amino acids, which explains why reactions upstream of their synthesis would be essential or not depending on the availability of these amino acids in the external environment.

As explained above, we used the frequency of reactions in panEFMs to train an EN model that predicts the metabolite niches of family-specific pan-reactomes from their natural reaction frequencies—these predictions are analogous to the predictions obtained from the Moran-like process that was applied to the *Aeromonadaceae* pan-reactome (Fig. 2D) except that reaction frequencies are now derived from their actual distribution in the pan-reactome (Additional file 7: Table S2) rather than from a simulated evolutionary process. In both cases, the model, was trained on the frequency of panEFMs sampled across random environments. The predicted metabolic niches are summarized in Additional file 10: Table S5. Most reactomes require inorganic ions, such as Ca, Cl, Mn, Zn, K, Mg, and Fe, and some organic molecules such as heme are also widely required [45]. Different pan-reactomes require specific metabolites, making them distinguishable when projected in lower-dimensional space (large points in Fig. 3B). A more detailed characterization of metabolite preferences in pan-genomes can be of interest in future studies aimed at explaining the metabolic basis of genome evolution events.

panEFMs are mechanistic predictors of patterns in pan-reactome shape, size, and distribution

To further explore the evolutionary signals that can be extracted from family-specific pan-EFMs we compared pan-reactomes and panEFMs using multiple variables (Table 1). We selected variables that are commonly used in pan-genome studies [7, 14, 46]. Figure 4 displays the pairwise correlations between all variables across the 46 prokaryote families, detailed in Additional file 11: Table S6. Larger pan-reactomes contain a larger set of reactions that may be integrated into panEFMs [Pearson $r = 0.94$; adj. $p < e{-}20$; variable 'pan(pEFMs)']. Notably, this variable correlates better with the size of the "shell" [Pearson $r = 0.85$; adj. $p < e{-}11$; variable 'shell(Reactomes)'] than with the size of the "core" [Pearson $r = 0.24$; adj. $p = 0.154$; variable 'core(Reactomes)'] or the "cloud" [Pearson $r = 0.38$; adj. $p = 0.016$; variable 'cloud(Reactomes)'] of the pan-reactome. We also observed a significant spread in the average size of panEFMs (Additional file 5: Fig. S5), which correlates with the average reactome size of pan-reactomes [Pearson $r = 0.82$; adj. $p < e{-}10$; variable 'size(Reactomes)'] and all the other variables have similar correlations to the average panEFM size as they have with the pan-reactome size (Fig. 4).

Correlation of panEFMs with global environmental sequencing data

Next, we evaluated whether the panEFMs and pan-reactome variables correlate with the niche breadth of prokaryote families inferred from global environmental sequencing datasets (Fig. 4; Table 1; variable 'NicheBreadth'). We inferred niche breadth scores for all families from thousands of metagenomic datasets that quantify the diversity of the environments where each family is found, and are a proxy for the breadth of their niche preferences across the planet (Von Meijenfeldt et al., manuscript in preparation, see Methods section "Niche breadth"). First, we found that the diversity between the reaction frequencies of panEFMs that were sampled from different environments positively and significantly correlated with the niche breadth of the bacterial and archaeal families [Pearson $r = 0.42$; adj. $p < e{-}02$; variable 'diversity(pEFMs)']. This confirmed that families whose strains occur in diverse environments tend to have more diverse environment-driven reactions than families whose strains occur in uniform environments. Notably, families with a high panEFM fluidity, which is the pan-genome analog to mutation rates [46] and measures the average dissimilarity in the reaction content between random pairs of reactomes across all environments, did not have a significantly higher niche breadth [Pearson $r = 0.23$; adj. $p = 0.18$; variable 'fluidity(pEFMs)']. This shows that families whose pan-reactome is capable of differentially adapting to different simulated environments may be observed in more diverse metagenomic datasets than families whose pan-reactome similarly adapts to different environments. The two other variables that significantly correlated with the niche

Table 1 Variables that were used to compare the panEFMs and pan-reactomes of 46 prokaryote families (Fig. 4, Table S6)

Variable	Description
NicheBreadth	Predicted niche breadth from global environmental sequencing datasets (see Methods)
diversity(panEFMs)	The diversity between reaction frequency of panEFMs sampled in different virtual environments (Average squared pairwise Euclidean distance)
fluidity(panEFMs)	The average dissimilarity between panEFMs independently of the random environments in which it was sampled
pan(panEFMs)	Total reactions that are included in at least one of the panEFMs sampled in different virtual environments
pan(Reactomes)	The number of reactions found in the pan-reactome of a prokaryote family
size(panEFMs)	The average size of panEFMs sampled in different virtual environments
size(Reactomes)	The average size of the natural reactomes from a prokaryote family
core(panEFMs)	The number of reactions that are present in at least 98% of the panEFMs sampled in different virtual environments
core(Reactomes)	The number of reactions present in at least 98% of the natural reactomes from a prokaryote family
shell(panEFMs)	The number of reactions that are present in 3 to 98% of all the panEFMs sampled in different virtual environments
shell(Reactomes)	The number of reactions that are present in 3 to 98% of the natural reactomes from a prokaryote family
cloud(panEFMs)	The number of reactions present in up to 3% of the panEFMs sampled in different virtual environments
cloud(Reactomes)	The number of reactions present in up to 3% of the natural reactomes from a prokaryote family
diversity(Metabs)	The diversity between metabolite usage profiles of panEFMs sampled in different virtual environments (Average squared pairwise Euclidean distance)
EnvDReacs	The number of reactions with an environment-driven score (EDS) EDS significantly > 0 (adj. $p < 0.05$ on a Z-test)
EnvDMetabs	The number of metabolites that are significantly associated with reactions with a non-zero EDS

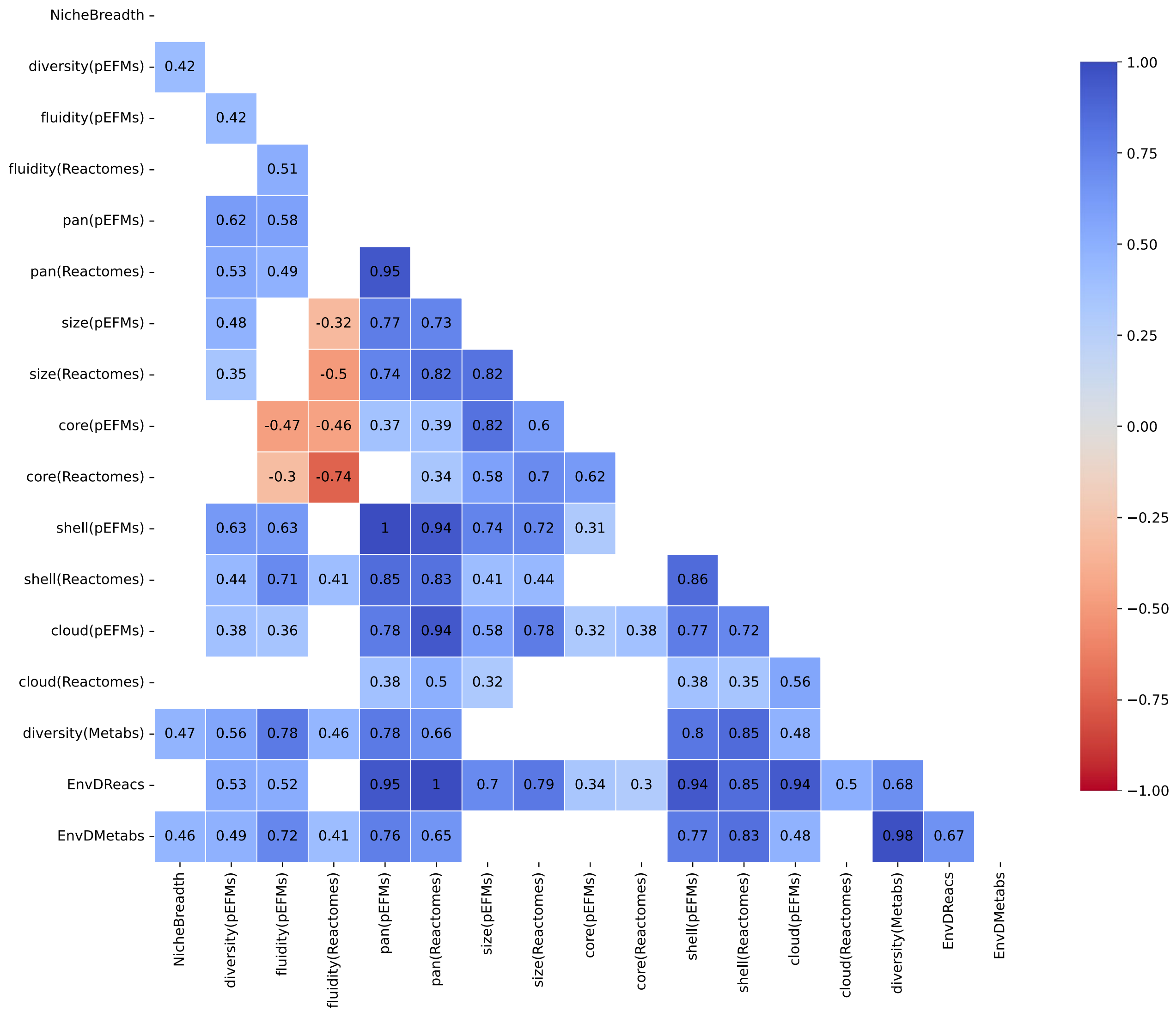


Fig. 4 Correlation of the variables measured from panEFMs with reactomes and metagenomes across 46 prokaryotic families. Only significant values are shown. A description of the variables is available in Table 1. Detailed Pearson correlation values and adjusted p-values are available in Table S6

breadth, namely (i) the number of metabolites that are significantly associated with environment-driven reactions (reactions with and EDS significantly > 0, Pearson r = 0.46, adj. p = 0.003; variable ‘EnvDMetabs’) and (ii) the diversity of metabolite usage profiles between panEFMs sampled in different environments (Pearson r = 0.47; adj. p < e−03; variable ‘diversity(Metabs)’) further indicate that reactomes of families with larger niche breadths not only have more diverse environment-driven reactions but can also use more diverse metabolite compositions when constrained to diverse environments.

General pan-reactome patterns

By evaluating the correlation of multiple variables from extant reactomes and the collection of panEFMs sampled across 1000 virtual environments (Table 1 and Additional file 11: Table S6) we propose a general scenario for pan-reactome evolution. For this scenario, we consider that prokaryotic clades, such as the families described here, have access to a shared pool of reactions and that non-essential reaction are frequently lost. Some pan-reactomes have a large number of core reactions and also a large number of reactions that are present in all panEFMs (Pearson r = 0.62; adj. p < e−04; variable ‘core(pEFMs)’).

The larger the core-reactome, the less diverse are reactome pairs, measured by the fluidity (Pearson r = − 0.74; adj. p < e−07; variable 'fluidity(pEFMs)'). The shell of panEFMs and the shell of pan-reactomes are significant predictors of ecological flexibility. Large shells imply larger pan-reactomes (Pearson r = 0.99; adj. p < e−46), more fluidity (Pearson r = 0.63; adj. p < e−04), and more environment-driven reactions (Pearson r = 0.94; adj. p < e−21) and metabolites (Pearson r = 0.77; adj. p < e−08). These scenarios are non-trivial properties of the pan-reactome composition, mechanistically identifiable in the collection of panEFMs, and suggest that the shape and size of the pangenomes reflect their evolution constrained to different environments.

Discussion

Diverse microbial clades have evolved specific sets of metabolic reactions to obtain their required energy and biomass. On average, more than 50% of the genes in microbial genomes code for metabolic functions [47], and metabolic genes are often found to be horizontally transferred [48]. The different sets of metabolic reactions used in different contexts by microbes reflect patterns and mechanisms of their genome evolution. For this reason, reactomes have increasingly been used as model systems for evolutionary genomics [49–54]. Here we modelled the evolution of reaction frequency distribution in pan-reactomes to understand the forces and mechanisms driving genome evolution. We used pan-reactomes as proxies for pan-genomes because they represent complex genotype-to-phenotype maps that allow us to directly explore the effects of differential gene composition, and study the composition of genomes in the context of complex cellular phenotypes.

We developed a mechanistic evolutionary model to expose the forces that drive reaction frequency distribution. In our framework, pan-reactomes share a pool of reactions and individual reactome lineages undergo a process of extensive gene loss. We modeled this natural process of gene loss [3, 19–21], by allowing reactomes to lose all of their non-essential reactions in independent iterations across different simulated environments. This process provided an empirical distribution of the functional reactomes that can evolve from a given reaction pool. This distribution allowed us to disentangle the environment-driven (nutrition) and environment-independent (nature) reactions, build a model that can predict metabolic niches from reaction frequencies, and compare reactome patterns between prokaryote families.

Our model captures some of the essential features of pan-genome evolution but neglects some noticeable features that will likely be of great interest in future extensions. Examples of missing features are the barriers and costs associated with accommodating foreign genes [4]. An objective function that considers total protein allocation [55] could simulate these costs with a similar framework as ours, but would also require high-quality protein-reaction maps, which are currently not available for most genomes. The use of a continuous distribution for the probability of sharing reactions would also likely increase the realism of our model. We chose instead to use discrete family-level reaction pools since we started from draft GSMMs reconstructions. In our experience, the draft reconstructions from the ModelSEED platform contain adequate information to distinguish higher taxonomic levels [56, 57] but may not have enough resolution for a detailed comparison between strains from the same species or genus, particularly from non-model organisms.

Previous studies have also used stochastic reductive evolution in reactomes to assess alternative scenarios of reactome diversity [50, 53, 54]. Most of these studies were applied to the pan-reactome of the *Escherichia coli* clade [51, 53, 58] or were applied to understand patterns that emerge from the universal set of reactions [50, 54], i.e. all metabolic reactions that have been identified in prokaryotes. Here we applied stochastic reductive evolution to understand differences within and across pan-reactomes of different prokaryotic families, providing a unique systematic overview of their pan-reactome dynamics. This approach allowed us to expose the patterns in pan-reactome size, shape, and diversity that are functions of its composition. It also allowed us to predict family-specific metabolic niches that await experimental testing.

We identified patterns at two levels. At the lower level of individual reactions, we mechanistically predicted the essentiality of a reaction based on the capacity of the pan-reactome to generate functional alternatives across environments. This allowed us to identify how many and which reactions become essential in new environments. At the higher level of pan-reactomes, our framework revealed non-trivial features. While all 46 pan-reactomes were subjected to the same process of reaction loss and were under similar functional constraints, features such as size, shape, and diversity were significantly different between families. We found that these features depend closely on the composition of the pan-reactomes and are reflected in the in-silico-generated collection of panEFMs.

The composition of the pan-reactome determines patterns observed in specific organisms. With our framework, we mechanistically identified these patterns from functional reactomes. For example, some pan-reactomes form functional reactomes with a large number of core reactions. These reactomes are very similar to each other and use a small set of metabolites (Fig. 4, Additional file 11: Table S6). Other pan-reactomes form functional reactomes with a small set of core reactions and a large set of reactions of intermediate frequency (shell reactions). These require more metabolites and exhibit significantly different reaction frequencies when their pan-reactomes are challenged with different environments (Fig. 4, Additional file 1: Table S6). All these properties result from the different ways that reaction sets can assemble to form functional reactomes.

Conclusion

Reaction frequency highly depends on the global reactome functionality since the patterns that we observed in panEFMs were identified without determining specific evolutionary goals or additive adaptive values for specific reactions. This constraint of functionality shapes the sample space of possible reactomes and constraints its evolutionary potential. In a similar trend, we expect that the combinatorial functionality of genes within a gene pool is an important driver of the pan-genome composition. In other words, the question of how often a gene is found in the genomes of a prokaryote group is to some extent addressed by how often the gene is expected in functional gene sets and to some extent by the composition of its external environment. We used these connections to predict the metabolic niches of natural evolving pan-genomes and identify the forces that shape pan-genomes as important functional units of prokaryote evolution.

Methods

Reactomes

Bacterial and archaeal strains (n = 4885) from 46 taxonomic families were selected from the PATRIC database [59] (Additional file 7: Table S2). We chose to use families that had genome sequences of over 24 different species and selected one strain of each species based on the maximum completeness and minimum contamination values of their genome sequences as reported in the PATRIC metadata.

We reconstructed genome-scale metabolic models (GSMMs) for each strain using the model SEED pipeline [56] implemented in PATRIC with the Mackinac python package v.0.8.4 [60]. Each model contains a list of reactions that are predicted to be coded by the prokaryote genomes; these reactions are referred to as the reactome of the strain. In addition to the genome encoded reactions, each model has a biomass reaction consisting of the relative proportions of biomass components, such as amino acids, nucleotides, proteins, fats, co-factors, and sugars, that the reactome should be able to synthesize in a growth environment. Additionally, some reactions that were not annotated in the genomes were added to assure that the reactomes were capable of producing biomass in complete media, an approach referred to as "gap-filling" [61].

The functionality of GSMMs was assessed by flux balance analysis (FBA) [62], optimized for biomass production. Computed with cobrapy version 0.21.0. We used the flux yields on the biomass reaction as an indication of growth.

Pan-reactomes

Pan-reactomes were generated by merging the reactomes of all the strains from a given prokaryote family. Each reaction was added once. Additionally, we added exchange reactions for the compounds that have transporters in any of the reactomes (Additional file 7: Table S2), resulting in pan-reactomes with the same group of 292 exchange reactions.

Environment ball

We generated vectors of random uniform relative concentrations for the shared list of external compounds added as exchange reactions, excluding water and oxygen (n = 290). For obtaining relative uniform concentrations, we sampled a Dirichlet distribution with dimensions equal to the number of compounds (290) with uniform parameters. Samples from this distribution add to one and there is an equal probability of observing any relative concentration of any of the compounds. The resulting relative concentrations were adjusted to a constant uptake rate of water in mmol gDW^{-1} h^{-1}. Oxygen was added as a binary factor, with environments being either aerobic (containing an unconstrained amount of oxygen) or anaerobic (with zero oxygen), selected with a probability of 0.5. We generated 1000 random samples of the environment ball and used the resulting concentrations as the growth environments (Additional file 7: Table S2).

Toy model

The toy model was generated with the reactions in Additional file 6: Table S1. Functionality was directly assessed by evaluating if the biomass components could be synthesized without accumulating by-products. Since stoichiometries were all equal to one (Additional file 6: Table S1, Fig. 1), the external environment was defined by the presence or absence of metabolites.

Moran process

We evolved populations of reactomes derived from the toy and the *Aeromonadaceae* pan-reactomes with a Moran-like process [42]. For this, we started with random functional reactomes (n = 1000) and simulated a two-step process. In the first step, a random reactome was chosen and a reaction was either deleted or inserted.

If the change resulted in a functional reactome, the process continued, otherwise, the previous reactome was restored. In the second step, two reactomes were chosen and one of them was replaced by a copy of the other, simulating a birth–death process with constant population size. After many iterations ($n = 10^6$), the different types of reactomes that persisted were selected as the evolved reactome types.

Sampling the environment-specific collection pan-reactome elementary flux mode (panEFMs)

To generate random samples from the environment-specific collection of panEFMs, we first constrained the pan-reactome to a given environment (Additional file 7: Table S2) and only proceeded if the flux on the biomass reaction was greater than zero with five significant digits. We then randomly removed reactions from the pan-reactomes and evaluated if the resulting network exhibits flux in the biomass reaction that is greater than a cutoff of 1% of the flux observed in the pan-reactome. If the biomass flux is below the cutoff, the reaction was restored to the network, otherwise, the next reaction was removed until all reactions were assessed. At the end of one iteration, the reactions that remained in the network constitute a random sample of a panEFM. Next, we randomized the order of reaction removal and repeated the process. Each randomization of the reaction order finds a random panEFMs. We sampled 1000 panEFMs for each of the 1000 metabolite concentrations in the environment ball.

Environment-driven scores (EDS)

Reaction frequency in the environment-specific collections of panEFM was represented by a matrix containing the different environments of the environment ball as rows and reactions found in a pan-reactome as columns. Similarly, environment-specific metabolic niches were summarized in a matrix with a similar structure but containing the environment ball metabolites as columns. The averages of the columns of these matrices are, respectively, the expected values of reaction frequencies and the expected values of metabolite usage across environments. Residuals were obtained by taking the difference between these expected values and the values in each row (each environment). The environment-driven scores for reactions and metabolites were defined as the standard deviation of these residuals, divided by their maximum value.

Obtaining metabolite usage profiles

The metabolic niches of the panEFMs of bacterial and archaeal families were obtained by enumerating which of the possible external compounds were imported into the metabolic network when optimizing for biomass production. For each panEFM we first obtained the environment-specific FBA solution. We then assessed the fluxes in the exchange reactions for this solution. Negative fluxes correspond to the metabolites that are effectively required to produce biomass. One metabolic niche corresponds to the set of metabolites whose transporters exhibited a negative flux in the FBA solution of a panEFM. Metabolic niches were summarized by the metabolite frequencies obtained from the 1000 random samples of panEFMs that were obtained for each environment. We thus obtained a metabolic niche for each random environment by enumerating how often each metabolite was used after sampling panEFMs from pan-reactomes.

Reaction frequencies

We restricted our analysis to reactions that had gene evidence (not gap-filled) and that could be active in a model. To define if a reaction could be active, we used flux variability analysis and excluded reactions that exhibited a flux variability of zero. We refer to the “natural reaction frequencies” as the frequencies that were observed from the reaction composition of reactomes reconstructed from the genomes of a prokaryote family, while “frequencies in panEFMs” refer to the frequency that reactions were found in the random samples of panEFMs.

Elastic net

An elastic net model was trained to predict metabolite usage from reaction frequencies (natural or resulting from simulations in the Moran process). We used the matrices described above as training sets with five-fold cross-validation. The natural reaction frequencies were used to predict evolutionary environments. To train and fit the model we used the Python 3.7 package scikit-learn version 0.22.2.

Niche breadth

For each family, we calculated its niche breadth on the scale from specialist to generalist based on its presence in a large number of publicly available environmental sequencing datasets (Von Meijenfeldt et al., https://doi.org/10.1101/2022.07.21.500953).

In short, we selected taxonomically annotated environmental sequencing projects from the MGnify dataset [63]. We selected analyses that were annotated with the 4.1 pipeline to ensure that the taxonomic profiles were comparable, removed analyses with less than 50,000 taxonomically annotated reads or ≥ 10% eukaryotic reads, chose a maximum of 1,000 samples per biome and selected 1 analysis per sample. The 22,518 selected analyses spanned 140 different biomes across a wide geographical range, containing both metagenomic, transcriptomic, and amplicon datasets (Von Meijenfeldt et al., https://doi.org/10.1101/2022.07.21.500953).

A family was considered present in a sample if its relative abundance was ≥ 1/10,000. Niche breadth was defined as the mean pairwise distance between all the samples in which a family is found, where the mean pairwise distance is defined as ½ − (Spearman's rank correlation on family level/2). Since this measure is solely based on the taxonomic content of a sample, it is independent of manually added metadata such as the biome from which it originates. A family with a low score is primarily found in samples with similar taxonomic profiles and we thus consider it a specialist, and a family with a high score is found in more dissimilar samples and is thus a generalist.

Abbreviations

EDS: Environment-driven score; EFM: Elementary flux mode; EN: Elastic net; FBA: Flux balance analysis; HGT: Horizontal gene transfer; GSMM: Genome-scale metabolic model; panEFM: Pan-reactome elementary flux mode.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12862-022-02052-3.

Additional file 1: Figure S1. Convergence of the reaction frequencies of panEFMs sampled across 1000 virtual environments (Table S2) to an average. The frequency of reactions was obtained from two random-independent (non-overlapping) groups of panEFMs defined across the env

Additional file 2: Figure S2. Distribution of pan-reactome reaction frequencies and panEFMs reaction frequencies. The x-axis contains the frequency of reactions in sampled panEFMs across random environments, while the y-axis contains the natural frequency coded by the genomes of the taxonomic families.

Additional file 3: Figure S3. Distribution of the environment-driven reaction score (EDS) and panEFMs reaction frequencies.

Additional file 4: Figure S4. Comparison of panEFMs across environments. Scatter plot of the average reaction frequency of panEFMs defined across random virtual environments and within each environment.

Additional file 5: Figure S5. Size distribution of panEFMs sampled across random virtual environments.

Additional file 6: Table S1. Reactions in the toy model

Additional file 7: Table S2. Bacterial and archaeal strains used in this study. The Aeromonadaceae strains are highlighted.

Additional file 8: Table S3. Environment compounds used in this study. Concentrations of 1000 random environments that constitute the environment ball (see Methods). These concentrations are set as upper bounds to the metabolic models of the pan-reactomes.

Additional file 9: Table S4. Environment-driven reaction scores (EDS) for all 46 prokaryote families.

Additional file 10: Table S5. Elastic net predictions of the metabolite usage by the pan-reactomes of 46 prokaryote families

Additional file 11: Table S6. Correlation of variables related to panEFM, pan-reactomes, and metagenomes

Acknowledgements

We thank Bas Teusink and Paulien Hogeweg for the useful discussion.

Author contributions

DRG: conceived and designed the analysis, performed the analysis, prepared figures and tables, and wrote the paper; FABM: contributed with data analysis tools. BVD: prepared figures, conceived and designed the analysis. AB: conceived and designed the analysis. MAH: conceived and designed the analysis. BED: conceived and designed the analysis, wrote the paper. All authors reviewed the manuscript.

Funding

DRG was supported by CNPq Science Without Borders. FABvM and BED were supported by the Netherlands Organization for Scientific Research (NWO) Vidi grant 864.14.004 and European Research Council (ERC) Consolidator grant 865694: DiversiPHI. BED was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2051 – Project-ID 390713860.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the GitHub repository, https://github.com/danielriosgarza/NutritionOrNature

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Centre, Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands. [2]Microbial Systems Biology, Laboratory of Molecular Bacteriology, Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Louvain, Belgium. [3]Department of Marine Microbiology and Biogeochemistry (MMB), NIOZ Royal Netherlands Institute for Sea Research, PO Box 59, 1790 AB Den Burg, The Netherlands. [4]Department of Microbial Population Biology, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany. [5]Department of Pathology, Radboud Institute for Molecular Life Sciences (RIMLS), Radboud University Medical Center, Geert Grooteplein-Zuid 10, 6525 GA Nijmegen, The Netherlands. [6]Theoretical Biology and Bioinformatics, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands. [7]Institute of Biodiversity, Faculty of Biology, Cluster of Excellence Balance of the Microverse, Friedrich Schiller University, Jena, Germany.

Received: 13 September 2021 Accepted: 22 July 2022
Published online: 16 August 2022

References

1. Lorenz MG, Wackernagel W. Bacterial gene transfer by natural genetic transformation in the environment. Microbiol Rev. 1994;58:563–602.
2. Paget E, Simonet P. On the track of natural transformation in soil. FEMS Microbiol Ecol. 1994;15:109–17.
3. Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC Biol. 2014;12:66.
4. Thomas CM, Nielsen KM. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol. 2005;3:711–21.
5. Hao W, Golding GB. The fate of laterally transferred genes: life in the fast lane to adaptation or death. Genome Res. 2006;16:636–43.
6. Iranzo J, Wolf YI, Koonin EV, Sela I. Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence. Nat Commun. 2019;10:5376.

7. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol. 2017;2:1–5.
8. Sela I, Wolf YI, Koonin EV. Theory of prokaryotic genome evolution. Proc Natl Acad Sci. 2016;113:11399–407.
9. Sela I, Wolf YI, Koonin EV. Assessment of assumptions underlying models of prokaryotic pangenome evolution. BMC Biol. 2021;19:27.
10. Shapiro BJ. The population genetics of pangenomes. Nat Microbiol. 2017;2:1574–1574.
11. Rocha EPC. Neutral theory, microbial practice: challenges in bacterial population genetics. Mol Biol Evol. 2018;35:1338–47.
12. Aminov RI. Horizontal gene exchange in environmental microbiota. Front Microbiol. 2011. https://doi.org/10.3389/fmicb.2011.00158.
13. Bonham KS, Wolfe BE, Dutton RJ. Extensive horizontal gene transfer in cheese-associated bacteria. Elife. 2017;6:e22144.
14. Maistrenko OM, Mende DR, Luetge M, Hildebrand F, Schmidt TSB, Li SS, et al. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. ISME J. 2020;14:1247–59.
15. Sitaraman R. Prokaryotic horizontal gene transfer within the human holobiont: ecological-evolutionary inferences, implications and possibilities. Microbiome. 2018;6:163.
16. Wiedenbeck J, Cohan FM. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev. 2011;35:957–76.
17. Dixit PD, Pang TY, Maslov S. Recombination-driven genome evolution and stability of bacterial species. Genetics. 2017;207:281–95.
18. Sela I, Wolf YI, Koonin EV. Selection and genome plasticity as the key factors in the evolution of bacteria. Phys Rev X. 2019;9: 031018.
19. Bolotin E, Hershberg R. Bacterial intra-species gene loss occurs in a largely clocklike manner mostly within a pool of less conserved and constrained genes. Sci Rep. 2016;6:35168.
20. Snel B, Bork P, Huynen MA. Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res. 2002;12:17–25.
21. Wolf YI, Koonin EV. Genome reduction as the dominant mode of evolution. BioEssays. 2013;35:829–37.
22. Sheinman M, Arkhipova K, Arndt PF, Dutilh BE, Hermsen R, Massip F. Long identical sequences found in multiple bacterial genomes reveal frequent and widespread exchange of genetic material between distant species. bioRxiv. 2020;2020.06.09.139501.
23. Wolf YI, Makarova KS, Yutin N, Koonin EV. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. Biol Direct. 2012;7:46.
24. Makarova KS, Koonin EV. Evolutionary genomics of lactic acid bacteria. J Bacteriol. 2007;189:1199–208.
25. Ochman H. Genomes on the shrink. Proc Natl Acad Sci USA. 2005;102:11959–60.
26. Moran NA, Mira A. The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. Genome Biol. 2001;2:research0054.1.
27. Koonin EV. Are There laws of genome evolution? PLoS Comput Biol. 2011;7: e1002173.
28. Mazzolini A, Gherardi M, Caselle M, Cosentino Lagomarsino M, Osella M. Statistics of shared components in complex component systems. Phys Rev X. 2018;8: 021023.
29. Pang TY, Maslov S. Universal distribution of component frequencies in biological and technological systems. Proc Natl Acad Sci USA. 2013;110:6235–9.
30. Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 2008;36:6688–719.
31. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 2009;5: e1000344.
32. Baumdicker F, Hess WR, Pfaffelhuber P. The infinitely many genes model for the distributed genome of bacteria. Genome Biol Evol. 2012;4:443–56.
33. Collins RE, Higgs PG. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. Mol Biol Evol. 2012;29:3413–25.
34. Haegeman B, Weitz JS. A neutral theory of genome evolution and the frequency distribution of genes. BMC Genomics. 2012;13:196.
35. Edwards JS, Palsson BO. Systems properties of the *Haemophilus influenza* Rd metabolic genotype. J Biol Chem. 1999;274:17410–6.
36. Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. Current status and applications of genome-scale metabolic models. Genome Biol. 2019;20:121.
37. Norsigian CJ, Fang X, Palsson BO, Monk JM. Pangenome flux balance analysis toward panphenomes. In: Tettelin H, Medini D, editors. The pangenome: diversity, dynamics and evolution of genomes. Cham (CH): Springer; 2020.
38. O'Brien EJ, Monk JM, Palsson BO. Using genome-scale models to predict biological capabilities. Cell. 2015;161:971–87.
39. Hosseini S-R, Martin OC, Wagner A. Phenotypic innovation through recombination in genome-scale metabolic networks. Proc R Soc B. 2016;283:20161536.
40. Wagner A. Metabolic networks and their evolution. In: Soyer OS, editor. Evolutionary systems biology. New York: Springer; 2012. p. 29–52.
41. Zanghellini J, Ruckerbauer DE, Hanscho M, Jungreuthmayer C. Elementary flux modes in a nutshell: properties, calculation and applications. Biotechnol J. 2013;8:1009–16.
42. Moran PP. The rate of approach to homozygosity. Ann Hum Genet. 1958;23:1–5.
43. Lobkovsky AE, Wolf YI, Koonin EV. Gene frequency distributions reject a neutral model of genome evolution. Genome Biol Evol. 2013;5:233–42.
44. Amorim Franco TM, Blanchard JS. Bacterial branched-chain amino acid biosynthesis: structures, mechanisms, and drugability. Biochemistry. 2017;56:5849–65.
45. Benson DR, Rivera M. Heme uptake and metabolism in bacteria. In: Banci L, editor. Metallomics and the cell. Dordrecht: Springer; 2013. p. 279–332.
46. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: an integrative view of gene diversity within microbial populations. BMC Genomics. 2011;12:32.
47. Griesemer M, Kimbrel JA, Zhou CE, Navid A, D'haeseleer P. Combining multiple functional annotation tools increases coverage of metabolic annotation. BMC Genomics. 2018;19:948.
48. Goyal A. Metabolic adaptations underlying genome flexibility in prokaryotes. PLoS Genet. 2018;14: e1007763.
49. Aguilar-Rodríguez J, Wagner A. Metabolic determinants of enzyme evolution in a genome-scale bacterial metabolic network. Genome Biol Evol. 2018;10:3076–88.
50. Barve A, Rodrigues JFM, Wagner A. Superessential reactions in metabolic networks. Proc Natl Acad Sci USA. 2012;109:E1121–30.
51. Pál C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD. Chance and necessity in the evolution of minimal metabolic networks. Nature. 2006;440:667–70.
52. Pang TY, Lercher MJ. Each of 3323 metabolic innovations in the evolution of *E. coli* arose through the horizontal transfer of a single DNA segment. Proc Natl Acad Sci USA. 2019;116:187–92.
53. Szappanos B, Fritzemeier J, Csörgő B, Lázár V, Lu X, Fekete G, et al. Adaptive evolution of complex innovations through stepwise metabolic niche expansion. Nat Commun. 2016;7:11607.
54. Wagner A. Evolutionary constraints permeate large metabolic networks. BMC Evol Biol. 2009;9:231.
55. Alter TB, Blank LM, Ebert BE. Protein allocation and enzymatic constraints explain *Escherichia coli* wildtype and mutant phenotypes. bioRxiv. 2020. https://doi.org/10.1101/2020.02.10.941294.
56. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol. 2010;28:977–82.
57. Seaver SMD, Liu F, Zhang Q, Jeffryes J, Faria JP, Edirisinghe JN, et al. The ModelSEED biochemistry database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. Nucleic Acids Res. 2021;49:D575–88.
58. Yizhak K, Tuller T, Papp B, Ruppin E. Metabolic modeling of endosymbiont genome reduction on a temporal scale. Mol Syst Biol. 2011;7:479.
59. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. Nucleic Acids Res. 2017;45:D535–42.
60. Mundy M, Mendes-Soares H, Chia N. Mackinac: a bridge between ModelSEED and COBRApy to generate and analyze genome-scale metabolic models. Bioinformatics. 2017;33:2416–8.
61. Pan S, Reed JL. Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries. Curr Opin Biotechnol. 2018;51:103–8.
62. Feist AM, Palsson BO. The biomass objective function. Curr Opin Microbiol. 2010;13:344–9.

63. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, et al. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. 2020;48:D570–8.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.