Synteny in Bacterial Genomes: Inference, Organization and Evolution

Ivan Junier
Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain; Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain

Olivier Rivoire
CNRS/UJF-Grenoble 1, LIphy UMR 5588, Grenoble, 38402, France

Genes are not located randomly along genomes. Synteny, the conservation of their relative positions in genomes of different species, reflects fundamental constraints on natural evolution. We present approaches to infer pairs of co-localized genes from multiple genomes, describe their organization, and study their evolutionary history. In bacterial genomes, we thus identify synteny units, or "syntons", which are clusters of proximal genes that encompass and extend operons. The size distribution of these syntons divide them into large syntons, which correspond to fundamental macro-molecular complexes of bacteria, and smaller ones, which display a remarkable exponential distribution of sizes. This distribution is "universal" in two respects: it holds for vastly different genomes, and for functionally distinct genes. Similar statistical laws have been reported previously in studies of bacterial genomes, and generally attributed to purifying selection or neutral processes.

Here, we perform a new analysis based on the concept of parsimony, and find that the prevailing evolutionary mechanism behind the formation of small syntons is a selective process of gene aggregation. Altogether, our results imply a common evolutionary process that selectively shapes the organization and diversity of bacterial genomes.

The position of genes along genomes affect their function and evolution. In bacteria, functional constraints are thus responsible for the concentration of highly expressed genes near the origin of replication, and the clustering of co-functional genes into operons of co-regulated genes. Similarly, evolutionary constraints on gene order are evidenced in bacteria by the highly variable rates of recombination over different chromosomes regions, and by the propensity of co-localized genes to be co-displaced through horizontal transfer [1]. Thus, while genes may be lost, gained, duplicated and rearranged during evolution, the comparison of evolutionary related species shows a remarkable stability of genomic organization [2, 3].

This conservation of genomic organization, often referred to as "synteny", raises three questions: (i) How to infer, from a comparison of multiple genomes, the pairs of genes that remain close-by across multiple genomes? (ii) How to describe, beyond pairwise relationships, the organization of conserved properties of co-localization? (iii) How to explain, from an evolutionary perspective, the origin of this organization?

Following upon several previous studies [4–10], we tackle these questions through a multi-genome comparative analysis, using as input the over one thousand complete bacterial genomic sequences that are now available. Our approach to answering (i) yields, at each given phylogenetic level, a list of pairs of genes that remain close-by across multiple genomes. In the spirit of [6], we address (ii) by projecting these pairs on individual genomes to define synteny units, or syntons, as clusters of mutually proximal genes. We find that these syntons correspond to functional features of bacterial genomes encompassing operons, and that the distribution of their size partition them in two classes: large syntons, associated with fundamental functions of bacterial cells, and smaller ones, which in many cases do not correspond to previously defined genomic units. These smaller syntons, however, follow a remarkable statistical property, not previously reported, with their sizes being exponentially distributed.

Finally, in response to (iii), we note that two types of evolutionary processes can account for the conservation of genomic organizations over different clades: (1) an incomplete dislocation of gene order from a common ancestral genome [11, 12], or (2) a selective aggregation of genes to form new clusters. These two processes are not exclusive, and their balance may for instance explain the conservation of the larger syntons [10]. Both processes are also compatible with the exponential size distribution for the smaller syntons; however, we provide evidence that the accumulation of genes into expanding syntons is the dominating process, a result in contrast with previous models that considered disintegration as the driving force of synteny evolution [8].

Inference of synteny

Inferring synteny from multiple genomes presents several difficulties: (a) the classification of genes into orthology classes, which is a notoriously difficult problem [13]; (b) an highly non-uniform sampling of genomes, both because natural genomes are phylogenetically related and because sequencing efforts have not been distributed evenly across strains and species; (c) the definition of a non-ambiguous criterion for assessing significant conservation of proximity between genes. Our approach specifically addresses each of these difficulties: (a) given
an initial partition of genes into orthology classes, we use the inferred properties of synteny as a guide to iteratively refine it, and to eventually test the consistency of our results; (b) one parameter in the analysis allows for the investigation of syntenic properties at different phylogenetic depths, and can thus mitigate biases from closely related genomes; (c) our thresholds of statistical significance are defined for any given rate of false positive discovery of conserved proximities.

The input data is a set of \( M \) genomes whose genes are partitioned into \( N \) orthologous classes. Specifically, we consider here a set of \( M = 1108 \) bacterial genomes annotated in terms of \( \pi \) phylogenetic depths, and can thus mitigate biases from for the investigation of syntenic properties at different levels of our results; (b) one parameter in the analysis allows for the investigation of syntenic properties at different phylogenetic levels, thus providing an information that is sure of divergence of gene contexts (Figs. S1-S2). Varying this parameter, the evolutionary distance \( \delta \) below which two genomes are considered to be similar (Fig. 1A).

As we are interested in synteny, we take here for \( \delta \) a measure of divergence of gene contexts (Figs. S1-S2). Varying \( \delta \) allows us to perform the same analysis at different phylogenetic levels, thus providing an information that is either generic to many bacteria (large \( \delta \)), or specific to a small subset of them (small \( \delta \)).

For each pair of COGs \( ij \), we define its relative distance in a genome as the minimal distance, in base pairs, between its respective genes (this distance is formally \( \infty \) if one of the COGs is not represented in the genome). The distribution of this distance across all genomes is computed by taking into account the \( \delta \)-dependent genome weights. We then assign a \( p \)-value \( \hat{\pi}_{ij} \) to the pair \( ij \) by comparing this distribution with that from a null model where genes are distributed independently and uniformly across \( M' \) genomes (Fig. 1A Methods and Fig. S4).

Given the large number of pairs \( ij \) under study (\( \sim 10^7 \)), some of the \( p \)-values \( \hat{\pi}_{ij} \) are borne out to be very small, even under the null model. One more step is therefore required to set a threshold of significance for these \( p \)-values. This is achieved by comparing the empirical distribution \( f(\pi) \) of \( \hat{\pi}_{ij} \) with its distribution under the null model, \( f_0(\pi) \). The fraction of false positives when calling significant the pairs \( ij \) with \( \hat{\pi}_{ij} < \pi^* \) can be estimated as \[ q = \int_0^{\pi^*} f_0(\pi) d\pi / \int_0^\infty f(\pi) d\pi \] (Methods). A given false discovery rate \( q \) (the fraction of false positives given \( \pi^* \)), here taken at 5\%, thus selects a threshold of significance \( \pi^* \). This procedure corresponds to applying to synteny properties the approach of [17], with a simpler but more stringent criterion justified by the small fraction of true positives.

### Organization of synteny

The structure of the relationships of conserved proximity can first be analyzed on individual genomes, in line with the work of Rogozin et al. To this end, we assess whether each pair \( ij \) of COGs found to be significantly proximal, is indeed close-by in the particular genome \( g \), using \( \hat{x}_{ij} \) as a characteristic distance (Methods and Fig. S5). This defines a genome-based network of proximity, where the nodes are genes and the links relations of conserved proximity in \( g \). To identify synteny units from this network, we rely on a property of transitivity: for \( ijk \) to be considered as a unit, all three pairs \( ij \), \( ik \) and \( jk \) must be linked. This corresponds to relevant subnetworks being fully inter-connected, so-called “cliques” in graph theory (Fig. 2A). We thus define the “syntons” of a genome as the maximal cliques of its network of proximity, i.e., the cliques that are not strictly contained in any other. The syntons of a genome hence consist of maximal sets of genes that are proximal in the genome as well as in a significant number of other genomes.

Syntons are defined by comparing the positions of genes between species, without taking into account promoters or terminators. Yet, while not necessarily consisting of contiguous genes, syntons are related to operons. As shown in Fig. 2 for the \( E. coli \) genome [18], the partitions into syntons and operons share many of their boundaries: besides, operons are rarely found in two different syntons. These results indicate that syntons comprise operons but extend beyond them, just as operons are clusters of pathway-related operons [19] and statistically correlated genes [20].

The size distribution of syntons is particularly informative: as shown in Fig. 3 it displays a critical size that defines two distinct types of syntons. Syntons above this size (whose exact value depends on the phylogenetic depth \( \delta \) of the analysis) are found to contain genes encoding the building blocks of the fundamental molecular complexes of bacterial cells, including the translation/transcription machinery, the ATP-producing respiratory complex, the cell division complex, the cell envelope biogenesis and the flagellum machinery (Table S1). We refer to these syntons as syntons of type B (with "B" for ‘basic building blocks’).

After removing these syntons, the size distribution of the remaining syntons follows an exponential law,
\( \rho(\sigma) \sim e^{-\alpha \sigma} \), with an exponent \( \alpha \) that varies slightly with the phylogenetic depth \( \delta \) at which the analysis is performed; the larger \( \delta \), the larger \( \alpha \) is (Fig. 3B), reflecting the fact that smaller and smaller contexts are recognized as conserved when considering wider and wider phylogenetic ranges (\( \alpha \) is inversely related to the mean size of syntons). At a given \( \delta \), however, the exponent \( \alpha \) is nearly the same for all genomes, irrespectively of phylogenetic distances or genome lengths, and despite the fact that synton compositions may have little overlap (Fig. 4). Randomizing the COGs confirms that the exponential law is a \textit{bona fide} property of the data, and not a necessary consequence of the methods (Fig. S6). Hereafter, we refer to the small, exponentially distributed syntons, as syntons of type A (with ‘A’ for ‘aggregating’, see below).

Synteny units may also be defined for the “pan-genome” that comprises all bacterial genomes, by considering a pan-network of proximity where the nodes are COGs and the links connect pairs of COGs \( ij \) for which \( \pi_{ij} \leq \pi^* \) (as in Fig. 1B). From this pan-network, partially represented in Fig. 5A, we may invoke again transitivity to define as synteny units its maximal cliques. Since forming a clique in this pan-network is, however, only a necessary condition for a set of genes to be proximal in a particular genome, we call them “pan-syntons” only when they are a synton in at least one specific genome.

The number of pan-syntons thus defined is larger than the number of syntons in any genome, which allows for broader statistical tests. For instance, we verify in Fig. 3B that the clustering of COGs into pan-syntons is consistent with their annotation into 24 functional categories \( [14] \). From their size distribution, the same distinction can again be made between type B pan-syntons, associated with macro-complexes, and smaller type A pan-syntons, which sizes are again exponentially distributed. This distribution is preserved when the threshold of statistical significance is varied (Fig. S7). It also does not appear to be associated with any particular subclass of genes (Fig. S8). For the same reason that the exponent \( \alpha \) decreases with decreasing \( \delta \) (Fig. 3B), \( \alpha \) for the pan-genome (Fig. 4C) is smaller than the typical \( \alpha \) for a genome studied at a same phylogenetic depth \( \delta \): the syntons in individual genomes are indeed included in pan-syntons. The fact that all these networks, either global or genome-based, obey, up to a scaling factor, to the same statistical law (Fig. 5D), may reflect the "fractal structure of the gene universe" \( [21] \).

Finally, we note that \( \alpha \) could depend on the number of genomes considered for the analysis, with a larger number of genomes possibly yielding a smaller exponent. Structural constraints on synteny properties \( [20] \), together with the limited number of COGs shared by many bacteria \( [2] \), should, however, limit the maximal size of type A syntons, and hence put a bound on \( \alpha \). Repeating our analysis for smaller sets of bacterial genomes yields results consistent with these arguments (Fig. S9).

**Evolution of synteny**

The organization of genomes changes during evolution as a result of gene losses, acquisitions by duplication or horizontal transfers, and transpositions. The finding of an exponential law for the sizes of small syntons strongly constrains the way in which these different factors must interplay. Yet, qualitatively very different scenarios are consistent with this law: (i) as for the Boltzmann law in statistical mechanics, it may reflect an equilibrium with conserved mean size of the syntons; (ii) it may result from the disintegration of an ancestral genome; or (iii) from a process of aggregation of genes.

Scenario (ii) underlies previously proposed models for rearrangements of genes in eukaryotes \( [11] \) and bacteria \( [12] \). In the simplest such model, two consecutive genes are disrupted at a constant rate \( r_d \), leading to a probability \( e^{-\alpha T s} \) for an initial gene to preserve its integrity across a period of time \( T \).

The scenario (iii) of aggregation has, to our knowledge, not been previously considered. In one of its simplest instantiations, rearrangements lead isolated genes, formally forming syntons of unit size, to join the neighborhood of an existing synton of size \( k \), and, when this rearrangement confers a selective advantage, it is fixed and generate a new synton of size \( k + 1 \). If \( N_k \) denotes the number of syntons of size \( k \) in a particular genome, this simple model is described by

\[
\begin{align*}
\frac{d t N_1}{d t} &= \Phi - r_a (1 + \rho_1) N_1, \\
\frac{d t N_k}{d t} &= r_a \rho_1 N_{k-1} - r_a \rho_1 N_k \quad (k \geq 2),
\end{align*}
\]

where \( r_a \) is an aggregation rate, accounting for both rearrangement and fixation, \( \rho_1 = N_1 / \sum_{k \geq 1} N_k \) the density of isolated genes, and \( \Phi \) a flux of gene innovation, standing for gene duplications, horizontal transfers or \textit{de novo} gene births. Provided that \( r_a \) and \( \Phi \) vary slowly enough compared to the composition of syntons, this generic model leads to an exponential distribution of cluster sizes.

The evolution of the densities \( \rho_k = N_k / \sum_{l \geq 1} N_l \) of \( k \)-clusters is indeed given by

\[
\begin{align*}
\frac{d t \rho_1}{d t} &= \phi - (r_a + \phi) \rho_1, \\
\frac{d t \rho_k}{d t} &= r_a \rho_1 \rho_{k-1} - \phi \rho_k \quad (k \geq 2),
\end{align*}
\]

where \( \phi \equiv \Phi / \sum_{l \geq 1} N_l \). Assuming that the synteny rates \( r_a(t) \) and the flux \( \phi(t) \) have their own dynamics on a time scale longer than the time scale of the dynamics described by Eqs. (3)-(4), the variables \( \rho_k(t) \) reach stationary values before \( \phi \) and \( r_a \) undergo any noticeable changes. This adiabatic approximation reduces the number of parameters from 2 (\( r_a \) and \( \phi \)) to 1 (\( r \equiv \phi / r_a \)), and gives as a stationary solution an exponential distribution of synton sizes,

\[
\rho_k = \frac{r}{(1 + r)^k}, \quad \text{with} \quad r = \frac{\phi}{r_a}.
\]

[raw_text]
Without seeking to infer a precise model for syton formation, we may identify the nature of the prevalent scenario, disintegration (ii) or aggregation (iii), by comparing genomic contexts and invoking a principle of parsimony. This principle has been invoked in several studies of genome evolution [17, 22, 23], but, at variance with these previous works, we do not rely here on the reconstruction of a global species tree. Instead, we shall consider triplets of equidistant genomes.

Starting with a pair of genomes \((g_1, g_2)\) sharing a common gene \(i\), if \(g_1\) has gene \(j\) in the context of \(i\) but \(g_2\) not, two parsimonious explanations are conceivable: \(j\) was next to \(i\) in the last common ancestor and \(g_2\) underwent a disaggregation, or it was not, and \(g_1\) underwent an aggregation (Fig. 5A). We may attempt to estimate the corresponding probabilities \(p_A\) of aggregation (\(j\) joining the context of \(i\)) and \(p_D\) of disaggregation (\(j\) leaving the context of \(i\)) over all pairs of genomes separated by a given phylogenetic distance \(\delta_s\) (here measured from sequence similarity; Methods). These probabilities are indeed related to the fractions \(f_0, f_1\) and \(f_2\) of those pairs of genomes that both contain a COG \(i\), and for which a neighbor \(j\) of \(i\) in the pan-network of proximity (Fig. 5A) is, respectively, not in the context of \(i\) in any of the two genomes, present in one of them, or present in both. The relation, however, also involves the probability \(q\) for the ancestor to have \(j\) in the context of \(i\) and, therefore, cannot define uniquely \(p_A\) and \(p_D\) (Fig. 5A).

This indeterminacy is lifted when considering triplets of equidistant genomes (Fig. 5B; practically, two distances \(\delta_s\) are considered equal when they differ by \(\Delta \delta_s \leq 0.012\)). If, besides, conditioning to the presence of a \(k\)-clique instead of conditioning to the presence of a single COG \(i\) (corresponding to \(k = 1\)), we can estimate the rates of aggregation and disaggregation of a gene as a function of the size \(k\) of the group of co-localized genes that it is joining or leaving. The results of this analysis, shown in Fig. 5C, indicate that \(p_A\) is larger than \(p_D\) for \(k > 2\), all the more that \(\delta_s\) is large. The same analysis can be repeated without the condition that \(i\) and \(j\) must be significantly conserved; in this case, compared to \(p_D\), \(p_A\) is negligible for all values of \(\delta_s\) (Fig. S11), in agreement with the observation that, except for a small subset of them, genome neighborhoods are poorly conserved among distantly related bacteria [5]. For sytons, however, we conclude that aggregations dominate over disaggregations, thus ruling out an equilibrium-like distribution à la Boltzmann as an explanation for the exponential distribution.

**Discussion**

We presented statistical approaches to infer conserved proximal relationships between genes, identify the relevant units that they are forming, and deduce the nature of evolutionary process behind their formation. Our approaches can be extended to other aspects of genomic organization, and, in the case of the analysis over triplets of equidistant genomes, to other evolutionary processes, such as horizontal transfers or gene duplications. Accounting for these processes will refine the model of syton formation beyond the dichotomy aggregation/disaggregation. At this stage, we note that while our results imply the existence of a selective pressure for aggregating genes to stick together, they do not reveal its nature. We can only point out that the probability of gene aggregation, \(p_A\), appears to be nearly independent of the size of the syntenic unit already present in the genomes (\(k\) in Fig. 6C).

Besides the inference of evolutionary processes, our results have implications for the functional interpretation of genomic sequences. Syntenic properties are indeed commonly used to discriminate paralogs [21], based on the premise that similarity of context correlates with similarity of function. Genomic contexts can thus refine the partition into COGs, which is based on sequences only, by defining smaller clusters of genes with analogous contexts [23]. Following this line, we derived from our results a set of contextually refined COGs, which we call cCOGs (Methods). Repeating our analysis with the cCOGs as fundamental units consistently reinforces the results found from the COGs. All the cliques of the corresponding pan-network of proximal relationships now consist of genes that are co-localized in at least one genome: pan-syntons can therefore be defined as maximal cliques of the pan-network of proximity with no further condition. From the size distribution of syntons, the same distinction can be made between syntons of types A and B, and the exponential distribution for the syntons of type A is only more significant (Fig. S12).

**Conclusion**

In conclusion, we presented a statistical study of multiple bacterial genomes that leads to the identification of novel units of synteny, called sytons. Identifying the relevant units of coevolution between genes is an essential step towards the rationale design of de novo functional genomes [26]. In addition, the exponential distribution of syton sizes may be added to the list of statistical "laws of genome evolution" [27, 32]. Previous examples could be explained by neutral processes and/or purifying selection [33]. The exponential distribution of syton sizes is remarkable for being driven by a process of positive selection, an on-going accretion of genes.

Genomic features other than synteny, such as for instance the co-occurrence of orthologous genes, may be analyzed along the same lines [34]. As these features may result from an evolutionary dynamics of a different nature than the aggregative process leading to sytons, the
relevant subunits of the network of conserved properties may not be its maximal cliques and/or their sizes may not be exponentially distributed. Identifying those relevant conserved units, studying their properties, and inferring the evolutionary mechanisms behind their formation are avenues for future studies of bacterial evolution.

MATERIALS

Data set – Sequenced bacterial genomes and COG annotations were downloaded from NCBI, yielding an initial data set of $M_0 = 1432$ genomes and $N = 4467$ COGs. We removed genomes with size below 500 kb or with less than 60% of genes annotated by COGs to obtain the $M = 1108$ genomes used in our analysis.

Inter-genome distances and genome weights – A measure of distance $D_{gh}$ between pairs of genomes is defined from the divergence of contexts of 10 genes known to be vertically inherited in bacterial genomes \(1\) (see Suppl. Info. for details). The number $M_{ij}(x)$ of genomes in which genes $i$ and $j$ are at distance $d_{ij} \leq x$ is computed as $M_{ij}(x) = \sum_x \omega_x (d_{ij} \leq x)$, with genome weights defined by $\omega_x = 1/\{h : D_{gh} < \delta\}$ \(2\); this weighting procedure defines an effective number of genomes as $M' = \sum_x \omega_x$ (Fig. 4A). An alternative measure $\delta_x$ of inter-genome distance is defined from sequence similarity by considering the same 10 genes, and computing the fraction of amino acids that they have in common after aligning them. $\delta_x$ and $\delta$ are correlated (Fig. S2), but because sequences are more conserved than contexts, $\delta_x$ is more appropriate for large divergences.

Significance of proximity – Assuming a uniform distribution of genes along a circular genome of length $L$, the probability of observing a distance $d$ less than $xL/2$ between 2 given genes is just $x$. In the null model, the number $M_{ij}(x)$ of genomes with $d \leq x$ thus follows a binomial law $B(M, x)$. The probability $\pi_{ij}(x)$ of observing $M_{ij}(x)$ events is therefore $\pi_{ij}(x) = I_x(M_{ij}(x), M' - M_{ij}(x) + 1)$, where $I_x(m, n)$ is the regularized incomplete beta function. The least likely and therefore most significant distance $\hat{x}_{ij}$ between a given pair of genes $ij$, is the one minimizing $\pi_{ij}(x)$, which defines a distance $\hat{x}_{ij}$ and an associated $p$-value $\hat{\pi}_{ij} = \pi_{ij}(\hat{x}_{ij})$.

Under the null model, the distribution of $y_{ij} = -\ln \hat{\pi}_{ij}$ is found to have an exponential tail, $f_0(y) \sim e^{-ay}$, with an exponent $a$ depending on $M'$ (Fig. 4 and Suppl. Info.). Given a threshold of significance $\pi^*$, we compute the fraction $\sigma_\pi$ of significant pairs, with $\hat{\pi}_{ij} \leq \pi^*$, and estimate the fraction of false positive pairs as $\sigma_{fp} = \int_{-\infty}^{\ln \pi^*} f_0(y) \sim (\pi^*)^a$. Imposing a false discovery rate $q = \sigma_{fp}/\sigma_\pi$ thus determines the threshold of significance $\pi^*$.

To account for the fact genomes may have several chromosomes, be non-circular and have different lengths, we formally circularize linear chromosomes and normalize them to a common length $L$ by setting all distances exceeding $L/2$ to $L/2$. The distance between genes on distinct chromosomes is also set to $L/2$. We take $L = 500$ kb, but our results are not sensitive to the exact value of this cutoff.

To treat pairs of COGs $ij$ with multiple copies (genes), we fix a gene $g_i$ in $i$, count the number $n$ of genes in $j$ at distance less than $xL/2$ and compute the probability of the event as $p(x) = 1 - (1 - x)^n$. The analysis is then performed as for $n = 1$ with $\pi_{g_i,j}(x)$ now standing for $\pi_{g_i,j}(p(x))$, thus defining $\hat{\pi}_{g_i,j}$. We then define $\pi_{ij}$ as the most significant observation when considering successively each gene $g_i$ in $i$, i.e., $\pi_{ij} = \min_{g_i \in C}{\pi_{g_i,j}}$. Different numbers of genes in $i$ and $j$ may imply $\pi_{ij} \neq \pi_{ji}$. Pairs of proximal COGs are identified by requiring that both $\pi_{ij} \leq \pi^*$ and $\pi_{ji} \leq \pi^*$.

Syntons and pan-syntons – For a given gene in a given genome, we build the maximal set of genes (including the gene itself) that are fully interconnected in the pan-network of proximity (clique) and close-by along the genome (Fig. 2A); two genes are considered to be close-by if they are separated by less than 50 kb (or less than $\hat{x}_{ij}$ if $\hat{x}_{ij} > 50$ kb, which occurs rarely, Fig. S4; taking a smaller value than 50 kb does not affect the results). In cases where several syntons of same size are possible, we take the synton with most significant score, computed by summing the $-\ln(\hat{\pi}_{ij})$ over each pair of genes.

Refinement of COG annotation: cCOGs – We iteratively partition the COGs into subsets of “contextually refined COGs” (cCOGs) by splitting them based on the synton to which they belong in each given genome. Each COG $i$ is thus independently reannotated by iterating the following steps:

1. Identification of COGs with conserved genomic proximity with respect to $i$.
2. Identification of the cliques $c$ to which $i$ belongs in the resulting network of proximal pairs of COGs. The cliques correspond a priori to incompatible genomic contexts, i.e., contexts found in different genomes.
3. The maximal cliques $c$ to which $i$ belongs are scored independently in each genome $g$ by $\sigma_{g}^{(c)} = \sum_{j \in c} -\ln(\pi_{ij})$, where the sum is over the COGs $j$ that are both in clique $c$ and in the context of $i$ in $g$. In every genome, the clique with the best score is associated to $i$, which is hence partitioned in cCOGs (see Fig. S12A).
4. A consensus genomic context is computed for each $i$ annotated in the previous step. COGs not annotated in the previous step are then annotated using the best-
matching context of these COGs. We stop the iterations when 99% of the COGs have their annotation unchanged in two successive iterations. For simplicity, the identification of the context of a COG, or of a cCOG, is always done using the original COGs. Note also that because the identification of maximal cliques is computationally demanding for large pan-networks, we identify the maximal cliques of the subnetwork associated to every COG, rather than the maximal cliques in the full network (see Fig. S12A).

At the most generic phylogenetic depth ($\delta = 0.91, M' = 10$), we take as input the original COGs. For smaller values of $\delta$, we start from the cCOGs obtained at a value of $\delta$ slightly larger. Eventually, we thus obtain many more cCOGs than COGs and, hence, larger pan-networks of proximity: \(~\sim 5200\ cCOGs\) with \(1300\) proximal pairs at $\delta = 0.91\ (M' = 10), \sim 48000\ cCOGs\) with \(131000\) pairs at $\delta = 0.68\ (M' = 304)$. In any case, the size distribution of (pan-)syntons always leads to two types of syntons, with an exponential distribution for type A syntons (Fig. S12B).

**ACKNOWLEDGMENTS**

We thank Toni Gabaldón, Bahram Houchmandzadeh and Kim Reynolds for helpful comments. I.J. thanks François Képès for his input at an early stage of this work. This work was supported by a Novartis grant (to I.J.) and by ANR grant ‘CoEvolInterProt’ (to O.R.).

---

[1] J G Lawrence. Gene organization: selection, selfishness and serendipity. *Anna. Rev. Microbiol.*, 57:419–40, 2003.
[2] E V Koonin and Y I Wolf. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucl acids res*, 36:6688, 2008.
[3] E P C Rocha. The Organization of the Bacterial Genome. *Annu Rev Genetics*, 42:211, 2008.
[4] W C Lathe, B Snel, and P Bork. Gene context conservation of a higher order than operons. *Trends in biochemical sciences*, 25:474, 2000.
[5] J Tamames. Evolution of gene order conservation in prokaryotes. *Genome biology*, 2(6), 2001.
[6] I B Rogozin, K S Makarova, J Murvai, E Czabarka, Y I Wolf, R L Tatusov, L A Szekely, and E V Koonin. Connected gene neighborhoods in prokaryotic genomes. *Nucleic acids research*, 30(10):2212–2223, 2002.
[7] B Snel, P Bork, and M A Huynen. Genomes in flux: the evolution ofarchaeal and proteobacterial gene content. *Genome Research*, 12(1):17–25, 2002.
[8] E P C Rocha. Inference and Analysis of the Relative Stability of Bacterial Chromosomes. *Molecular Biology and Evolution*, 23(3):513–522, 2005.
[9] M A Wright, P Kharchenko, G M Church, and D Segrè. Chromosomal periodicity of evolutionarily conserved gene pairs. *Proc Nat Acad Sc*, 104:10559, 2007.
[10] G Fang, E P C Rocha, and A Danchin. Persistence drives gene clustering in bacterial genomes. *BMC Genomics*, 9:4, 2008.
[11] J H Nadeau and B A Taylor. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc Nat Acad Sc*, 81:814, 1984.
[12] E P C Rocha. Order and disorder in bacterial genomes. *Current Opinion in Microbiology*, 7(5):519–527, 2004.
[13] T Gabaldón and E V Koonin. Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, 14(5):360–366, 2013.
[14] R L Tatusov, M Y Galperin, D A Natale, and E V Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucl acids res*, 28:33, 2000.
[15] F Morcos, A Pagnani, B Lunt, A Bertolino, D Marks, C Sander, R Zecchina, J N Onuchic, T Hwa, and M Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Nat Acad Sc*, 108:E1293, 2011.
[16] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B*, page 289, 1995.
[17] J D Storey and R Tibshirani. Statistical significance for genomewide studies. *Proc Nat Acad Sc.*, 100:9440, 2003.
[18] S Gama-Castro and et al. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (GenomS Uner). Nucl acids res res, 39:D98, 2011.
[19] Y Yin, H Zhang, V Olman, and Y Xu. Genomic arrangement of bacterial operons is constrained by biological pathways encoded in the genome. *Proc Nat Acad Sc*, 107:6310, 2010.
[20] I Junier, J Hérisson, and F Képès. Genomic Organization of Evolutionarily Correlated Genes in Bacteria: Limits and Strategies. *J Mol Biol*, 419:369, 2012.
[21] E V Koonin, The logic of chance: the nature and origin of biological evolution. *FT Press*, 2011.
[22] B G Mirkin, T I Fenner, M Y Galperin, and E V Koonin. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology*, 3(1):2, 2003.
[23] S K Kummerfeld and S A Teichmann. Relative rates of gene fusion and fission in multi-domain proteins. *Trends in genetics: TIG*, 21(1):25–30, 2005.
[24] C Von Mering, M A Huynen, D Jaeggi, S Schmidt, P Bork, and B Snel. STRING: a database of predicted functional associations between proteins. *Nucl acids res acids research*, 31:258, 2003.
[25] A P Yelton, B C Thomas, S L Simmons, P Wilmes, A Zemla, M P Thelen, N Justice, and J F Banfield. A Semi-Quantitative, Synteny-Based Method to Improve Functional Predictions for Hypothetical and Poorly Annotated Bacterial and Archaeal Genes. *PLoS Computational Biology*, 7(10):e1002230, 2011.
[26] F Képès, B C Jester, T Lepage, N Rafei, B Rosu, and I Junier. The layout of a bacterial genome. *FEBS Letters*, 586:2043, 2012.
[27] N M Lascombe, J Qian, Z Zhang, T Johnson, and M Gerstein. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome biology*, 3(8):1–0040.7, 2002.
[28] E van Nimwegen. Scaling laws in the functional content of genomes. *Trends in Genetics*, 19(9):479–484, 2003.

[29] Y I Wolf, P S Novichkov, G P Karev, E V Koonin, and D J Lipman. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Nat Acad Sc*, 106:7273, 2009.

[30] N Molina and E van Nimwegen. Scaling laws in functional genome content across prokaryotic clades and lifestyles. *Trends in genetics: TIG*, 25(6):243–247, 2009.

[31] J Grilli, B Bassetti, S Maslov, and M Cosentino Laggomarsino. Joint scaling laws in functional and evolutionary categories in prokaryotic genomes. *Nucl acids res*, 40:530, 2012.

[32] T Y Pang and S Maslov. Universal distribution of component frequencies in biological and technological systems. *Proc Nat Acad Sc*, 110:6235, 2013.

[33] E V Koonin. Are There Laws of Genome Evolution? *PLoS Comp Biol*, 7:e1002173, 2011.

[34] I Junier and O Rivoire. (in preparation).

[35] D R Zeigler. Gene sequences useful for predicting relatedness of whole genomes in bacteria. *Int. J. Syst. Evol. Microbiol.*, 53:1893–1900, 2003.

[36] G Palla, I Derényi, I Farkas, and T Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814, 2005.
FIG. 1: (A) A weight \( w_g \) is defined for each genome \( g \) that is inversely proportional to the number of other genomes at distance less than \( \delta \) from it. This distance is defined from the divergence of gene contexts between genomes, and \( M' = \sum_g w_g \) gives the effective number of genomes at this phylogenetic level (in this illustration, 2 at \( \delta = 0.9 \) and 3 at \( \delta = 0.5 \)). For a given \( \delta \), we compute for every pair of COGs \( ij \) the effective number \( M_{ij}(x) \) of genomes for which the genomic distance \( d_{ij} \) separating their genes is less than \( x \). This number is converted into a \( p \)-value, \( P[M_{ij}(x)] \), by considering a null model where gene positions are independently and uniformly distributed in \( M' \) genomes (Methods). Finally, the significance of the co-localization between \( i \) and \( j \) is defined by the minimum \( \hat{\pi}_{ij} \) of these \( p \)-values over the distances \( x \). (B) The tail of the distribution of \(- \log(\hat{\pi}_{ij})\) is exponential under the null model (black dots). The empirical distributions (red and blue) clearly deviate from it for small values of \( \hat{\pi}_{ij} \) (large values of \(- \log(\hat{\pi}_{ij})\)). We select a threshold of significance \( \pi^* \) (green) such that the false discovery rate, estimated from the areas below the two curves (Methods), is less than 5%. For smaller \( \delta \), more relationships of conserved proximity are thus detected, as depicted here by the (pan-)networks of proximity, where nodes represent COGs and links relationships of proximity (the red links at \( \delta = 0.5 \) are those already identified at \( \delta = 0.9 \)).
FIG. 2: (A) A synton is a maximal set of genes that are proximal both globally across genomes (dotted lines) and in a particular genome (full lines). The case of three different genomes is depicted here, which each have different syntons. In the last case, the genome has two overlapping syntons because the blue and red genes have no conserved proximity (the four genes do not form a clique in the pan-network). (B) Comparison between synton and operon organizations in *Escherichia coli*, here for $\delta = 0.53$ ($M' = 469$). Only a small fraction (77kb) of the *E. coli* genome is represented, with the operons in the upper band and the syntons in the lower bands (two bands are used for clarity; overlapping syntons have the same colors). By definition, operons are made of consecutive genes; in contrast, syntons may intermingle (as it is the case here for the yellow and dark grey syntons). The extreme boundaries of the operons are, however, often common to the synton splittings and synton separations (indicated by thin black lines). Many apparent mismatches come from an absence of COG annotation (red lines). (C) Statistical significance of these observations (see Fig. S5 for other values of $\delta$). The first histogram shows the expected number of common boundaries between operons and syntons after a random rotation of the operon organization (hence keeping all other features intact), with the orange arrow indicating the observed value. The second histogram shows the fraction of operons containing genes in at least two different syntons, with again the orange arrow pointing to the observed value. These graphs indicate that syntons are genomic features encompassing operons.
FIG. 3: (A) For a given $\delta$ (here = 0.83), the size distribution of syntons is similar for genomes that are phylogenetically far apart, or of vastly different sizes (e.g. *Mycoplasma hyopneumoniae*, 900 kb long genome, green points, versus *Catenulispora acidiphila*, 10.5 Mb long, orange points). These distributions can be divided into two parts: above a critical size (around 10 here), the (type B) syntons contain genes encoding the building blocks of the fundamental molecular complexes of bacterial cells. Below this critical size, the (type A) syntons have their sizes distributed according to an exponential law. (B) The exponent $\alpha$ of the exponential is nearly identical for all genomes, but varies with the phylogenetic depth $\delta$ at which conservation is estimated.

FIG. 4: Comparison of synton content between bacteria. The similarity in synton content between two species is defined as their number of common syntons divided by their total number of different syntons (Jaccard index). These similarities are shown here for a sample of 10 strains, showing a clear correlation with the phylogenetic relationships between species, as defined from sequence similarity and indicated by the trees.
Type A pan-syntons, below the critical size, have their sizes distributed according to an exponential law, \( \rho \) leads to distributions \( \tilde{\rho} \) that collapse into a single exponential; deviations at the largest values are within error bars (not indicated for clarity).
**FIG. 6:** Two non-exclusive mechanisms can explain the exponential size distribution of syntons: aggregation or disaggregation.

(A) Given a COG \( i \), we can compare its context in all pairs of genomes that contain it. For each COG \( j \) significantly co-localized with \( i \), we compute the fraction \( f_n \) of pairs of genomes at distance \( \delta_s \) where \( n \) out of the two genomes have \( j \) in the context of \( i \) \((n = 0, 1, 2)\). In a simple model where events of aggregation or disaggregation occur with probabilities \( p_A \) and \( p_D \) since the divergence between the two genomes, these frequencies are related to the probability \( q \) that \( j \) belongs to the context of \( i \) in the last common ancestor of the pair. As \( f_0 + f_1 + f_2 = 1 \), this gives only two independent equations for three unknown parameters. (B) The same approach, but based on triplets of equidistant genomes, now provides three independent equations for the same three unknown parameters. (C) A further generalization consists in fixing a clique of \( k \) proximal genes instead of a single gene \( i \). We thus report \( p_A \) and \( p_D \) for \( k = 3, 5, 7 \) as a function of the distance \( \delta_s \) between the genomes, and for \( \delta_s = 0.35 \) as a function of \( k \) (the error bars are standard errors). The last graph shows that \( p_A > p_D \) for \( k > 2 \). Fig. S10 extends these results to all values of \( k \in [1, 8] \) and \( \delta_s \in [0, 0.35] \).
SUPPLEMENTARY INFORMATION

Measure of context divergence

The context divergence $\delta$ between any two genomes is computed as $\delta = 1 - f$, where $f$ is the average fraction of common COGs in the context of a selection of 10 genes. The context of a gene is defined as the COGs located within 20 kb of this gene, and the 10 selected genes are associated to the COGs 126G, 173J, 202K, 2255L, 481M, 497L, 541U, 544O, 556L, 1158K. These COGs are taken from a list of genes shown to report phylogenetic distances between bacterial strains (Table 2 in [1]), with the additional constraint that they comprise a single copy in most of the 1108 genomes of our dataset.

Out of these 10 genes, only 202K shows a particularly conserved context. Comparing the context divergence obtained from two distinct sets of 5 genes, with 202K in common only, shows a good linear relation between the two estimations (Fig. S1). This self-averaging property indicates that using these 10 genes is suitable for measuring the context divergence between pairs of genomes.

Measure of sequence divergence

The same 10 genes are used to compute a measure of sequence divergence $\delta_s$ between any two genomes as $\delta_s = 1 - f_s$, where $f_s$ is the average fraction of common amino acids between the 10 genes, after alignment of their amino acid sequences. $\delta_s$ and $\delta$ are related as indicated in Fig. S2.

Distribution of $p$-values for the null model of proximity tendencies

A null model is defined by assuming that the positions of genes are drawn from an uniform distribution, independently in each of $M$ genomes. For each pair $ij$ of genes, the number of genomes for which the distance between $i$ and $j$ is smaller than $x$ is translated into a $p$-value $\pi_{ij}(x)$ (Methods). As any $p$-value, $\pi_{ij}(x)$ is uniformly distributed over the pairs $ij$ for each given $x$, or, equivalently, $y_{ij} = -\log \pi_{ij}(x)$ is exponentially distributed, $\psi(y) = e^{-y}$.

The quantity $\hat{\pi}_{ij} = \min_x \pi_{ij}(x)$ is not a $p$-value, but numerical simulations show that the tail of the distribution of $\hat{y}_{ij} = -\log \hat{\pi}_{ij}(x)$ is exponentially distributed, $\psi(\hat{y}) \sim e^{-a\hat{y}}$, with an exponent $a$ that depends on $M$ (Fig. S3). Taking for $M$ the effective number of genomes $M'$ gives this exponent as a function of the context divergence $\delta$.

Significance of the exponential distribution of type A synton sizes

Randomization of gene positions

To support the non-trivial nature of the exponential distribution of type A synton sizes, we repeated the analysis after randomly permuting the labels of the COGs of the pan-syntons of certain sizes (obtained at $\delta = 0.83$). Fig. S6 shows the results, where the blue points correspond to randomizing the pan-syntons of size 3-6 and the red points those of size 4-6. In any case, the exponential nature of the distribution is lost.

Varying false discovery rates

Using a more stringent false discovery rate than $q = 0.05$ reduces the statistics but does not affect the conclusion that pan-syntons can be divided into two types according to their size, with the size of the smaller ones being exponentially distributed (Fig. S7).

[1] Zeigler, D. R., 2003, Int. J. Syst. Evol. Microbiol. 53, 1893.
| Connected component | Biological function | Example of syntons |
|----------------------|---------------------|--------------------|
| 1005C 1007C 1008C 1009CP 1034C 1143C 1894C 1905C 377C 649C 713C 838C 839C 852C | NADH:ubiquinone oxidoreductase | 1005C 1007C 1008C 1009CP 1034C 1143C 1894C 1905C 377C 649C 713C 838C 839C 852C |
| 1181M 1589M 2001S 206D 275M 3116D 472M 707M 768M 769M 770M 771M 772D 773M 812M 849D | Cell division / Cell envelope biogenesis | 1181M 1589M 2001S 206D 275M 707M 768M 769M 770M 771M 772D 773M 812M 849D |
| 1157NU 1191K 1291N 1298NU 1317NU 1338NU 1360N 1377NU 1419N 1580N 1582N 1677NU 1684NU 1749N 1766NU 1776NT 1815N 1843N 1868N 1886NU 1987NU 2063N 3144N 3190N 455D 4786N | Flagellum | 1291N 1298NU 1317NU 1338NU 1360N 1580N 1582N 1677NU 1684NU 1749N 1766NU 1815N 1843N 1868N 1886NU 1987NU 455D 4786N |
| 100J 101J 102J 103J 1841J 185J 186J 197J 198J 199J 200J 201U 202K 203J 222J 244J 24J 250K 255J 256J 257J 361J 480J 48J 49J 50J 51J 522J 563F 59J 80J 81J 85K 86K 87J 89J 90J 91J 92J 93J 94J 96J 97J 98J 99J | Ribosome / RNA polymerase | 100J 101J 102J 103J 200J 201U 202K 203J 255J 256J 361J 563F 94J 96J 97J 98J 99J |
| 100J 1841J 185J 186J 197J 198J 199J 200J 201U 202K 203J 255J 256J 480J 48J 49J 50J 51J 522J 563F 80J 81J 85K 86K 87J 89J 90J 91J 92J 93J 94J 96J 97J 98J 99J | NADH:ubiquinone oxidoreductase | 100J 1841J 185J 186J 197J 198J 199J 200J 201U 202K 203J 255J 256J 480J 48J 49J 50J 51J 522J 563F 80J 81J 85K 86K 87J 89J 90J 91J 92J 93J 94J 96J 97J 98J 99J |

Table S I: COG composition of type B pan-syntons, defined at $\delta = 0.83$ as pan-syntons containing more than 13 COGs. The left column indicates the composition of the four connected components that are obtained when restraining the pan-network of proximity to the set of COGs that are included in type B syntons (graphs below); the biological function associated with each component is indicated in the center. The right column provides two examples of syntons for each component (only one in the case of NADH:ubiquinone oxidoreductase). The same biological functions are found when the syntons are determined using cCOGs. In this case, we also find the F0F1-type ATP synthase machinery; together with the ubiquinone oxidoreductase, it corresponds to the core of the ATP-producing respiratory macro-molecular complex.
FIG. S 1: Comparison of context similarity $f = 1 - \delta$ computed from two different groups of 5 genes (box plot). All genome pairs formed from the 1108 genomes of our dataset are reported.

FIG. S 2: Relation between the inter-genomic distance $\delta$ based on context similarity and the inter-genomic distance $\delta_s$ based on sequence similarity (box plot). The two are correlated, but context similarity vanishes before sequence similarity and $\delta_s$ therefore describes large divergences better.
FIG. S 3: Probability density of $-\log(\hat{\pi})$ for the empirical data (orange circles) at three phylogenetic depths: $\delta = 0.91, 0.83, 0.53$. Left panels: For small enough values of $-\log(\hat{\pi})$, the density decays exponentially with $-\log(\hat{\pi})$ (red lines). The deviation from an exponential at large values indicates the conservation of co-localization. For the null model where gene positions are randomized (black squares, right panels), with as number of genomes the effective number $M'$ corresponding to $\delta$ ($M' = 10, 130, 469, \text{respectively}$), the exponential decay extends to larger values of $-\log(\hat{\pi})$.

FIG. S 4: Distribution of the characteristic distances $\hat{x}_{ij}$ (in kb) associated with $\hat{\pi}$ (see Fig. 1B in main text).
FIG. S 5: Significance of the observed number of synton boundaries common to an operon boundary and of the fraction of operons shared between more than 2 syntons when comparing with randomly rotated genomes. This figure extends Fig. 2C to different values of the phylogenetic depth $\delta$. 

$\delta = 0.91$
$M' = 10$

$\delta = 0.89$
$M' = 36$

$\delta = 0.83$
$M' = 130$

$\delta = 0.53$
$M' = 469$
FIG. S 6: Effect of randomization on the distribution of pan-synton sizes ($\delta = 0.83; M' = 130$). Red points show the original distribution of type A pan-syntons (Fig. 4 in main text). The blue and green curves show the distributions obtained after reshuffling the positions of the COGs belonging to pan-synton of sizes 3 to 6 and 4 to 6, respectively. Reshuffling all COG positions leads to a very small graph that contains only cliques of size 2 (not shown).

FIG. S 7: Distribution of pan-synton sizes at $\delta = 0.53$ ($M' = 469$) for different false discovery rates (FDRs), showing again the presence of two types of pan-syntons with an exponential decay for type A pan-syntons. The value of the exponent $\alpha$ for the exponential decay is indicated. FDRs equal to 0.05 and $10^{-4}$ correspond respectively to $\pi^* = 10^{-4}$ and $\pi^* = 10^{-6}$; an expected absence of false positives correspond to $\pi^* = 10^{-9}$. For clarity, curves have been shifted along the y-axis for FDR = $10^{-4}$ and for the absence of false positives.

FIG. S 8: Compositions of pan-syntons in terms of the four main functional classes of COGs [14], as a function of their size. For type B pan-syntons, the categories "information storage and processing" (translation/transcription machinery) and "cellular process and signaling" (flagellum, cell division machineries) are over-represented. For type A pan-syntons, all functional classes are equally represented.
FIG. S 9: Exponent $\alpha$ for the exponential distribution of the size of type A pan-syntons at $\delta = 0.53$ (FDR = 0.05), for different numbers of genomes in the dataset; datasets with $M < 1108$ genomes were obtained by resampling the original dataset of $M = 1108$ strains (rightmost point). This analysis is consistent with a saturation of the value of $\alpha$ with the number of genomes, which future analyses including more genomes will be able to confirm or infirm.

FIG. S 10: Generalization of Fig. 6C to other values of $k$.

FIG. S 11: Probabilities of aggregation ($p_A$) and disaggregation ($p_D$) obtained from genomic contexts that are not conserved. The graph displays the results for the context of synteny units that contain 5 genes ($k = 5$ in Fig. 6 of main text); for conserved contexts, the dynamics is dominated by aggregation events (see Fig. 6 of main text). Here, we see that $p_A$ is negligible for all values of $\delta_s$; the same observation holds for all sizes of the synteny units (points not represented correspond to values of $p_A$ that were found to be slightly negative).
FIG. S 12: (A) A COG $i$ is partitioned in two cCOGs ($i_{c_1}$ and $i_{c_2}$) due to the presence of two maximal cliques ($c_1$ and $c_2$) in the subnetwork associated with $i$ (left), corresponding to genome organizations found in different genomes (right). (B) Size distributions of type A pan-syntons obtained from the pan-network of proximity associated with the cCOGs.