Genomic organization underlying deletional robustness in bacterial metabolic systems

Sayed-Rzgar Hosseini[a,b,1] and Andreas Wagner[a,b,c,1]

[a]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, CH-8057 Zurich, Switzerland; [b]The Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; and [c]The Santa Fe Institute, Santa Fe, NM 87501

Edited by Peter Schuster, University of Vienna, Vienna, and approved May 16, 2018 (received for review October 1, 2017)

Large-scale DNA deletions and gene loss are pervasive in bacterial genomes. This observation raises the possibility that evolutionary adaptation has altered bacterial genome organization to increase its robustness to large-scale tandem gene deletions. To find out, we systematically analyzed 55 bacterial genome-scale metabolisms and showed that metabolic gene ordering renders an organism's viability in multiple nutrient environments significantly more robust against tandem multigene deletions than expected by chance. This excess robustness is caused by multiple factors, which include the clustering of essential metabolic genes, a greater-than-expected distance of synthetically lethal metabolic gene pairs, and the clustering of nonessential metabolic genes. By computationally creating minimal genomes, we show that a nonadaptive origin of such clustering could in principle arise as a passive byproduct of bacterial genome growth. However, because genome randomization forces such as translocation and inversion would eventually erode such clustering, adaptive processes are necessary to sustain it. We provide evidence suggesting that this organization might result from adaptation to ongoing gene deletions, and from selective advantages associated with coregulating functionally related genes. Horizontal gene transfer in the presence of gene deletions contributes to sustaining the clustering of essential genes. In sum, our observations suggest that the genome organization of bacteria is driven by adaptive processes that provide phenotypic robustness in response to large-scale gene deletions. This robustness may be especially important for bacterial populations that take advantage of gene loss to adapt to new environments.

deletional robustness | genome organization | metabolic systems | essential genes | horizontal gene transfer

Bacterial genomes evolve highly dynamically. On the one hand, they expand through gene gain mechanisms such as horizontal gene transfer (HGT) (1). On the other hand, they contract via large-scale gene loss (2). Large-scale gene deletion events were first documented in obligate pathogens and symbionts (3, 4), but later comparative genomic studies showed that they are surprisingly pervasive in bacterial genomes in general (5, 6). Importantly, bacterial genomes experience a well-known general bias toward DNA deletion; that is, genome size reduction events prevail over genome size expansion events (7, 8). Moreover, according to experimental evolution studies, extensive gene loss by large-scale deletions can readily occur on short evolutionary timescales (9–11). Population genomic data show that the DNA deletion rate is sufficiently high that its effects would be visible to natural selection acting on bacterial genomes (*SI Appendix*, *Text S1*) (12).

Does the high incidence of large-scale gene deletions leave evolutionary signatures in bacterial genomes? We hypothesized that bacterial genomes have evolved an organization that provides robustness against the deleterious phenotypic effects of large-scale gene losses. Because large-scale deletional events typically delete multiple contiguous (linked) genes, such a robust genome organization should ensure that a tandem deletion of multiple linked genes is, on average, more tolerable than a deletion of the same number of genes randomly drawn from the genome without regard to linkage. In other words, bacterial genomes should be more robust to tandem deletion than random deletion of the same number of genes.

To validate this hypothesis, we focused on metabolic genomes, which encode the enzymes catalyzing the chemical reactions of metabolism. Compared with other biological systems, metabolism is particularly appropriate for such validation because well-established and experimentally validated computational methods are available that can predict complex phenotypes, especially a cell's viability in specific environments, from genomic information (13). What is more, well-annotated genome-scale metabolic networks with information about metabolic genes, reactions, gene-reaction association rules, and the relative genomic order of metabolic genes are available for multiple bacterial genomes (14, 15). Our analysis is based on such information from 55 genomes belonging to nine distinct species (including 46 genomes from different *Escherichia coli* strains, two genomes from different *Shigella* strains, and one genome each for seven other species) (15).

Results and Discussion

Excess Robustness to Tandem Gene Deletion. To quantify how metabolic gene order affects the phenotypic robustness of a metabolism to gene deletions, we subjected the metabolic genome of *E. coli K-12 MG1655* to two different kinds of multigene deletions. First, in tandem deletions, we deleted a given number of n metabolic genes in the order in which they occur in the *E. coli* genome. Second, in random deletions, we deleted n randomly chosen metabolic genes irrespective of their order in the genome. More specifically, for every value of n between 1 and 50, tandem deletion involved deleting all possible consecutive n-tuples of these genes (*Methods* and *SI Appendix*, Fig. S1). For

Significance

From the organismal and the anatomical levels down to the molecular level, all complex biological systems manifest astonishing organization and order that are counterintuitive and challenging to explain by evolutionary mechanisms. In this study, we focus specifically on one aspect of this biological organization: the arrangement of metabolic genes in bacterial genomes. We show that this organization ensures a substantially higher robustness to large-scale gene deletions than expected from random genomic ordering. We systematically investigate the possible evolutionary mechanisms behind the emergence of such robust organizations. Our analysis provides several lines of evidence indicating that bacteria may have gained a robust genome organization through pervasive gene loss events.

Author contributions: S.-R.H. and A.W. designed research; S.-R.H. performed research; S.-R.H. contributed new reagents/analytic tools; S.-R.H. and A.W. analyzed data; and S.-R.H. and A.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: razgar@gmail.com or andreas.wagner@ieu.uzh.ch.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1717243115/-/DCSupplemental.

Published online June 18, 2018.

random deletions, we deleted an equivalent number of randomly chosen n-tuples of genes. Next, we mapped the eliminated genes in these deletion variants to eliminated reactions in the *E. coli* metabolism and determined the viability of each deletion variant on up to 103 carbon sources using flux balance analysis (13) (*SI Appendix*, Table S2). We computed the robustness R of the *E. coli* metabolic genome to such gene deletions as the fraction of deletion variants that retain viability on at least one of the carbon sources, either for tandem deletion (R_{tandem}) or random deletion (R_{random}).

We observed that robustness to tandem deletions is higher for all numbers $n > 1$ of deleted genes, and sometimes considerably so (Fig. 1). The same observation holds when we used a more strict definition of robustness; namely, the fraction of deletional variants that retain viability on all carbon sources on which wild-type *E. coli* is viable (*SI Appendix*, Fig. S2). We also repeated this analysis for the 54 other prokaryotic genomes and observed the same patterns in all of them (see *SI Appendix*, Fig. S3 for two examples). Moreover, the same patterns emerged when we quantified robustness as a function of the amount of deleted DNA (in kilo base pairs) instead of the number of deleted metabolic genes (*SI Appendix*, *Text S2* and Figs. S4 and S5). To quantify by how much robustness to tandem deletions is higher than to random deletions, we computed the ratio R_{tandem}/R_{random}, which we call the excess robustness under tandem deletion. For example, for deletions of 20 genes, robustness to tandem deletions is on average 3.63-fold higher than robustness to random deletion (*SI Appendix*, Fig. S6). This excess robustness increases with the number of deleted genes (*SI Appendix*, Fig. S6). In other words, gene order increases in its importance for deletional robustness as deletions become larger. Moreover, by considering robustness based on viability on single individual carbon sources, we observed that robustness to tandem deletion is more conserved among bacterial species or strains than robustness to random deletion (*SI Appendix*, Fig. S7). In addition, robustness to tandem gene deletions varied to a greater extent among carbon sources than robustness to tandem deletion (*SI Appendix*, Figs. S8 and S9).

Genomic Features Underlying Robustness. We then asked whether this substantial excess robustness to tandem deletion can be traced to specific features of genome organization. We first focused on the organization of essential metabolic genes. Because the deletion of any one essential metabolic gene is enough to cause a metabolism to lose viability in a given environment, we reasoned that the genomic organization of essential genes may help explain a genome's excess robustness to tandem gene deletions. It has been shown previously that essential genes play a key role in shaping chromosome organization (16), and that they are not uniformly distributed but clustered in bacterial genomes (17, 18). Such clustering can increase the robustness of a genome to tandem multigene deletions. If essential genes were distributed uniformly in the genome, each region of the genome would have an approximately equal chance to include at least one essential gene, whose deletion would be lethal. In contrast, if essential genes are densely packed (i.e., clustered) in some genomic regions, other regions must be depleted of essential genes. Deletions in the latter regions would be nonlethal, such that this genome organization effectively increases robustness to multigene deletions (*SI Appendix*, Fig. S10). Because our analysis uses multiple environments that differ in their carbon sources, we distinguished two types of essential metabolic genes: strictly essential genes, which are essential on all carbon sources, and conditionally essential genes, which are essential on at least one carbon source. Using Kuiper's test (19), we showed that in the vast majority of the bacterial genomes, both types of essential metabolic genes are significantly clustered (*SI Appendix*, *Text S3* and Tables S3–S5).

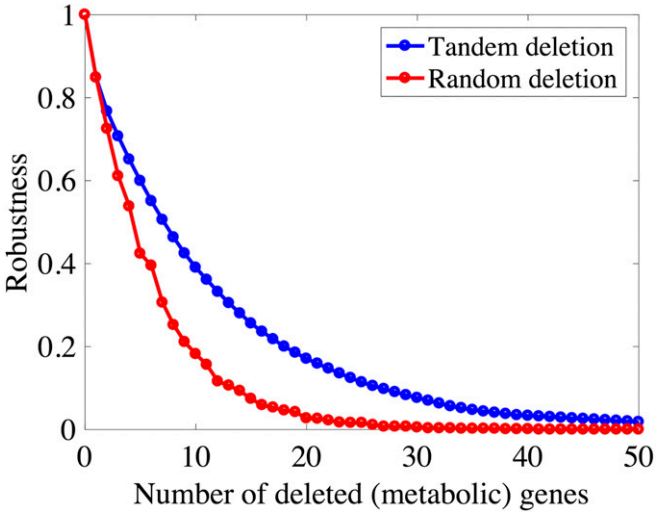

Fig. 1. Robustness to tandem deletion versus random deletion. The vertical axis shows the robustness of *E. coli K-12 MG1655* (*iJO1366*) to tandem (blue) and random (red) deletion of metabolic genes, averaged over all deletional variants we examined, as a function of the number of deleted genes (horizontal axis).

To find further genomic signatures of robustness to tandem gene deletions, we next focused on pairs of genes that are individually nonessential but jointly essential; that is, their simultaneous delection disrupts viability. Such gene pairs are also called synthetic lethals. If two synthetically lethal genes are closely linked in a genome, they are more likely to be deleted together in a tandem deletion. In contrast, if they are far away from each other, the probability that both of them are deleted in the same tandem deletion is much lower. Thus, synthetically lethal genes that are further apart than expected by chance alone could entail increasing robustness to tandem gene deletion. We refer to such synthetically lethal genes as being in repulsion.

To find out whether synthetically lethal genes are in repulsion, we created pairwise deletions of all nonessential metabolic genes in all 55 prokaryotic genomes and determined their viability. In this analysis, we distinguished again between two types of synthetically lethal genes. The first comprises strictly synthetically lethal gene pairs, whose joint deletion is lethal in all carbon source environments we consider. The second comprises conditionally synthetic lethal gene pairs, whose deletion is lethal in at least one but not all environments. We determined the distance between two strictly synthetically lethal metabolic genes as the number of metabolic genes that lie between them. In the majority of genomes (41 of 55; 74.54%), at least 50 genes lie between all strictly synthetic lethal gene pairs (*SI Appendix*, Table S6), and the paucity of strictly synthetically lethal gene pairs with a distance below 50 is statistically significant (*SI Appendix*, Table S7; Fisher's exact test). This repulsion is also visible in a circos plot of the *E. coli* genome (Fig. 2*A*), and it disappears after random genome shuffling (Fig. 2*B*). No short-range synthetic lethal interactions exist in the *E. coli* genome, but in the randomized genome, such interactions are abundant (Fig. 2 *B* and *C*). Similar patterns exist in other species (*SI Appendix*, Figs. S11 and S12). The same does not hold for conditionally lethal gene pairs (*SI Appendix*, Tables S8 and S9) and for some bacterial species with small metabolic genome sizes (*SI Appendix*, Fig. S13).

The role of nonessential genes in robustness to gene deletions may not be restricted to pairs of such genes, but could be extended to three or more genes that are individually nonessential but jointly essential. Unfortunately, the number of possible combinations of such synthetically lethal n-tuples of genes is too large for exhaustive analysis. However, if such genes are in repulsion, genomes might be enriched in long clusters of genes that harbor no essential genes, and whose joint deletion is not lethal. This is indeed the case (see *SI Appendix*, *Text S4*, Figs. S14 and S15, and Tables S10 and S11 for more details).

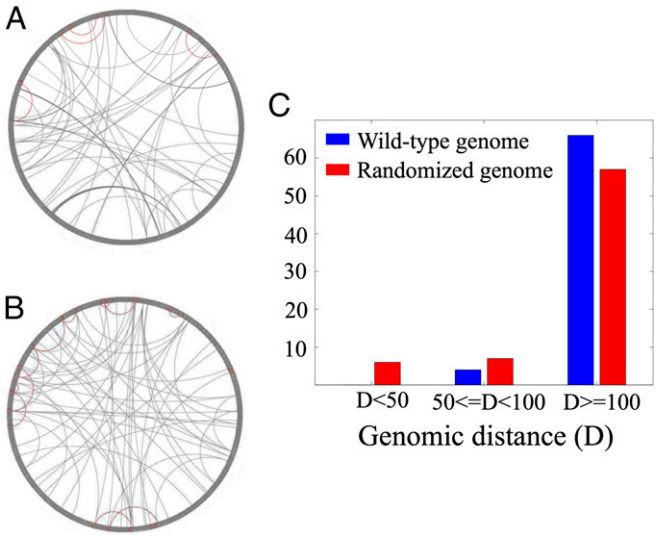

Fig. 2. Repulsion of synthetic lethal genes. (*A*) Circos plot of the *E. coli K-12 MG1655* (*iJO1366*) genome, in which metabolic genes are arranged according to their order in the genome. An arc connects two genes if they form an unconditionally synthetic lethal pair. An arc is colored red if the genomic distance (in number of intervening genes) between two synthetic lethal gene pairs is less than 100. (*B*) Same as *A*, but for randomized gene order. Note the many more short-ranged synthetic lethality interactions after gene order randomization. (*C*) Barplot of the genomic distance (in number of intervening genes) between unconditionally synthetic lethal metabolic gene pairs in the wild-type (blue) and randomized (red) *E. coli K-12 MG1655* (*iJO1366*) genome. Note the lack of short-distance synthetic lethal pairs with fewer than 50 intervening genes in the wild-type genome (Fisher's exact-test, $P = 0.0094$ and adjusted $P = 0.0306$; See *SI Appendix*, Table S7). It is important to mention that we used a distance of 50 genes as a criterion to evaluate the dispersion of synthetic lethal gene pairs, because according to *SI Appendix*, Table S6, in the genomes that show dispersion of synthetic lethals, the minimum distance between (strictly) lethal gene pairs was almost 50.

Nonadaptive or Adaptive Origins of Robust Genome Organization. The features of bacterial genome organization we just described ensure higher robustness to gene deletion, and they may have originated as adaptations to gene deletion. However, there are also several alternative possibilities. The simplest is that they originated nonadaptively. To find out whether this may be the case, we first examined the following nonadaptive scenario. It is inspired by the concept of a minimal genome, which has been used by multiple researchers as a model for the genome of early DNA-based life forms (20–23). In a minimal metabolic genome, no one gene can be removed without destroying viability, so every gene is essential. A minimal genome is basically a single cluster of essential genes. If present-day genomes evolved from minimal genomes largely by the insertion of genes, then the observed present-day clustering of essential genes might be a mere remnant of their clustering in the minimal genome, and thus a nonadaptive byproduct of evolutionary genome growth. To validate this hypothesis, we used a previously established algorithm (24) (*Methods*) that serially deletes individual genes to generate minimal (metabolic) genomes that can sustain life on a given carbon source. We then reinserted the missing metabolic genes step by step in random locations until we had reinstated a genome with the same number and identity of genes, but a different gene order. Applying this method to the genomes of three bacterial species showed that the extent of essential gene clustering is similar to that observed in wild-type genomes (*SI Appendix*, *Text S5* and Figs. S16–S21). Thus, a simple nonadaptive process could, in principle, explain the extent of essential gene clustering observed in modern genomes.

However, this simplistic model of genome evolution has several flaws. First, although the minimal genome approach is popular (20–23), the minimal genomes it creates may not approximate the genome organization of early cells. Second, genome evolution involves many more processes, including ongoing gene deletions and duplications. A similar analysis that includes such processes shows that gene deletions enhance the clustering of essential genes further, whereas gene duplications reduce it somewhat (*SI Appendix*, *Text S5* and Figs. S16–S21). This observation, which corroborates the findings of a previous model (18), implies that adaptive processes in which selective pressure is imposed by gene deletions can also contribute to the clustering of essential genes. Thus, the origin of their clustering may not be purely nonadaptive. Finally, and most important, frequently occurring genome rearrangement processes (25, 26) such as translocation and inversion will cause clusters of essential genes to erode (*SI Appendix*, *Text S6* and Figs. S22–S26). Thus, even though gene insertion or some other nonadaptive mechanism may have originally created essential gene clustering, other, adaptive mechanisms are needed to maintain such clustering.

Coregulation and Robust Genome Organization. We next examined whether the genomic organization of metabolic genes may be purely a product of adaptation to gene deletion, a byproduct of adaptation to other selective forces, or a combination of both. In doing so, we focused on several alternatives for adaptation to gene deletion. They revolve around the organization of bacterial genes into operons.

We first examined the possibility that coregulation of functionally similar genes within operons may fully account for the excess robustness to tandem deletions. Not only do operons frequently harbor multiple essential genes, which is an important source of essential gene clustering (*SI Appendix*, *Text S7*, Figs. S27–S31, and Tables S12–S17), but operonic genes are also frequently functionally related (ref. 27 and *SI Appendix*, Fig. S32). That is, they belong to the same linear metabolic pathway or the same functional subsystem of a metabolic network (27, 28). Such functionally related genes are likely to be clustered in an operon because it is advantageous to coregulate them (29, 30). If the metabolic pathway or functional subsystem they belong to is nonessential, then many of its genes may also be nonessential, such that an organism will be robust to the tandem deletion of these genes. In other words, if the colocalization of functionally related genes, which is driven by coregulation, can account for all observed excess tandem robustness, then the coregulation of such genes is the likely ultimate cause of robustness to tandem deletions.

To find out whether this is the case, we systematically analyzed operon structures in bacterial genomes with the aid of the Database for prOkaryotic Operons (DOOR) (31, 32), a comprehensive database for operon information. As expected from the functional relatedness of operonic genes, tandem gene deletion affects fewer distinct metabolic pathways or functional subsystems than random deletion (*SI Appendix*, Figs. S33 and S34).

However, using a partial randomization of wild-type genomes (*SI Appendix*, *Text S8*) that keeps essential gene clustering unchanged (type I randomization) or that leaves metabolic pathway or subsystem organization intact (type II randomization), we were able to de-convolve the effect of the organization of metabolic pathway or subsystems from that of the organization of essential genes. To quantitatively compare the effect of these two types of partially randomized organizations, we determined which fraction of the excess robustness to tandem deletion is preserved after each type of genome randomization. That is, we computed $(R_{\text{partial}}(n) - R_{\text{random}}(n))/(R_{\text{tandem}}(n) - R_{\text{random}}(n))$, where $R_{\text{partial}}(n)$ refers to robustness to deletion of n metabolic genes after partial genome randomization of either type. We observed that a consistently higher fraction of the excess robustness to tandem deletion is preserved when essential gene clustering is preserved (type I) than when metabolic pathway or subsystem organization is preserved (type II; Fig. 3 and *SI Appendix*, Figs. S35–S41). This indicates that the clustering of essential genes explains more of the excess robustness to tandem gene deletions than the number of affected metabolic pathways or subsystems (*SI Appendix*, *Text S8* and Figs. S35–S41). A related analysis shows that the repulsion of synthetic lethal gene pairs

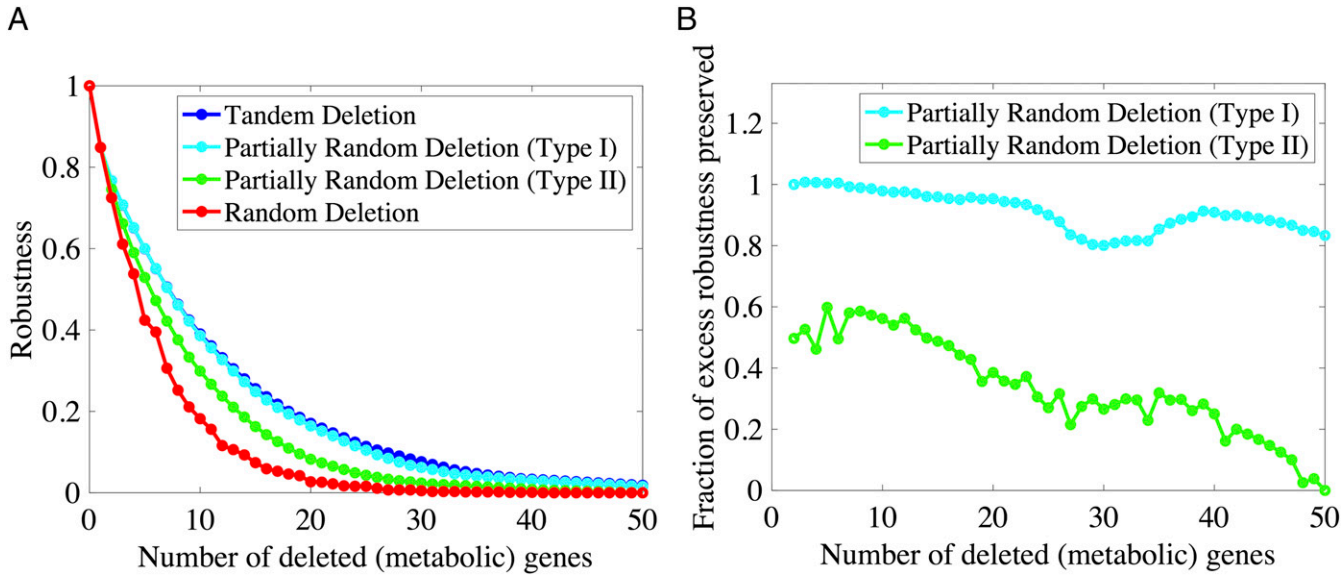

Fig. 3. Metabolic pathway organization is less important for excess robustness to tandem deletion than essential gene clustering. (*A*) Data show the robustness to tandem (blue), random (red), partially random (cyan and green for type I and II, respectively, see text) deletion of a given number of metabolic genes (horizontal axis) in *E. coli K-12 MG1655* (*iJO1366*). Each curve is obtained by a linear interpolation between the 50 data points. (*B*) Data show the fraction of excess robustness to tandem deletion that is preserved after partially random deletion of type I (cyan) and type II (green), as a function of the number of deleted metabolic genes (horizontal axis) for *E. coli K-12 MG1655* (*iJO1366*). Each curve is obtained by a linear interpolation between the 50 data points. See *SI Appendix, Text S8* for more details.

is not simply caused by the repulsion of metabolic pathways or subsystems (*SI Appendix*, *Text S9* and Tables S18 and S19). In sum, the coregulation-driven organization of functionally related genes does not suffice to explain excess robustness to tandem deletion.

We next focused our analysis on the uber-operon hypothesis (33), which asserts that operons may form larger functional units in the form of physically linked and functionally related or coregulated operons. To find out whether such uber-operons might account for all excess robustness to tandem deletions, we created tandem deletional variants using operons instead of genes as the units of deletion. We observed that the relative ordering of operons in bacterial genomes influences the clustering of essential genes to some extent, but does not have a large effect on robustness to tandem gene deletion (*SI Appendix*, *Text S10* and Figs. S42–S44).

We finally turned to the selfish operon hypothesis as a possible explanation for the excess robustness to tandem gene deletion (34). This hypothesis asserts that the organization of genes into operons is not necessarily beneficial for a host genome, but for the constituent genes, because an operon enables the spreading of its genes to new cells and species by horizontal gene transfer. The hypothesis, which has been criticized (30, 35), also predicts that horizontally transferred operons would harbor genes with peripheral (i.e., nonessential) metabolic functions (34), such that their deletion would be more tolerable than the deletion of other genes in the genome. By identifying all HGT-acquired metabolic genes in 43 of the bacterial genomes using the HGTree database (36) (*Methods*) among metabolic genes, we found no significant association between a gene being part of an operon and being horizontally acquired (*SI Appendix*, Table S20), as the selfish operon hypothesis would predict. This evidence weakens support for the selfish operon hypothesis itself and is consistent with previous work (30). More important, we also find that excess robustness to tandem deletions cannot be fully accounted for by the dispensability of operonic genes, as implied by the selfish operon hypothesis (34) (*SI Appendix*, *Text S11*, Fig. S45, and Table S21).

In sum, three prominent alternatives to the hypothesis that metabolic genome organization is an adaptation to ongoing DNA deletions cannot fully explain our key observation; namely, an excess of robustness to tandem deletions of metabolic genes.

The Importance of Horizontal Gene Transfer. We next show that HGT plays an important role in essential gene clustering, and that this role is consistent with the notion that ongoing DNA deletions help shape bacterial genome organization, and especially essential gene clustering. To appreciate the evidence, it is important to realize that HGT cannot influence the clustering of essential genes without gene deletions. The reason is that if a genome does not experience gene deletions, HGT-acquired genes cannot become essential in the environment in which they have been transferred, because in this environment, the organism is already viable without the newly transferred genes. However, in the presence of gene deletions, newly and initially nonessential HGT-acquired genes can become essential; for example, when other genes that were essential before the acquisition of the new gene or genes via HGT undergo deletion. We note that this line of reasoning applies only to strictly essential genes (i.e., essential in all environments), not conditionally essential genes, because HGT-acquired genes can become essential in new environments without the need for gene deletions (37).

We identified all HGT-acquired metabolic genes in 43 of our bacterial genomes, using the HGTree database (36) (*Methods*), and noticed that essential metabolic genes, and in particular strictly essential ones, are more likely to have been acquired by HGT than other metabolic genes (*SI Appendix*, Fig. S46 and Tables S22 and S23). This is consistent with a recent experimental study showing that HGT-acquired genes are frequently indispensable (38). What is more, HGT-acquired essential metabolic genes are significantly clustered in most genomes. We found this out by partitioning a genome's set of strictly essential genes into two groups: those acquired by HGT and those not acquired by HGT (36). Then, we quantified the clustering of genes in each group separately, using Kuiper's test, which showed that HGT-acquired essential metabolic genes are significantly clustered in 39 (90.7%) of 43 of our bacterial genomes. In contrast, non-HGT-acquired essential metabolic genes are significantly clustered only in three genomes (6.98%; *SI Appendix*, Fig. S47 and Table S24). This contribution of HGT to the clustering of essential genes is independent from that of operons because among the essential metabolic genes, there is no significant association between being part of an operon and being horizontally acquired (*SI Appendix*, Fig. S48 and Tables S25 and S26).

It is tempting to explain the clustering of HGT-acquired essential genes by the recent observation that genes usually transfer into a small number of chromosomal hotspots (39). However, if we consider all horizontally transferred genes (both essential and nonessential ones) in an analysis of gene clustering, we find that horizontally transferred genes are not significantly clustered, but uniformly distributed in a majority of the genomes (*SI Appendix*, Table S27). The likely reason is that many of the genes in the HGTree database (36) have not been very recently transferred (*Methods*), such that genome rearrangement events had enough time to randomize their location.

Genome rearrangements would in general lead to the dispersion of gene clusters. That clusters of essential horizontally transferred genes persist is thus all the more remarkable and points to the selective advantage such clusters provide. Coregulation is not likely to be the sole cause of this advantage because of the independence of HGT-acquired essential genes from operonic essential genes (*SI Appendix*, Fig. S48 and Tables S25 and S26). In contrast, DNA deletions do not only provide a means to render horizontally transferred genes essential, as we discussed earlier, but also are sufficiently frequent to exert substantial selection pressure for a robust genome organization (*SI Appendix*, *Text S1*), which points to their involvement in maintaining such clusters.

Conclusions. We have shown that the ordering of metabolic genes in bacterial genomes provides phenotypic robustness against deleterious effects of large-scale gene deletions. This robustness can endow bacterial populations with the flexibility to survive large-scale gene deletion events, which could potentially help them adapt to new environments (particularly in pathogenic species) (40–43). Underlying this excess robustness is a nonrandom distribution of both essential and nonessential metabolic genes, which is manifested as a clustering of essential genes and repulsion of synthetic lethal genes. Although a genome growth process starting from minimal genomes shows that clustering of essential genes could in principle have nonadaptive origins, other ongoing genome rearrangement processes would erode such clustering. The data we analyzed here suggests that in the face of such processes, essential gene clustering may be maintained through a joint advantage of the coregulation of functionally similar genes and of the robustness to multigene deletions that such clustering provides. Horizontal gene transfer, together with ongoing gene deletions, plays an important role in maintaining essential gene clustering.

Methods

Bacterial Genome-Scale Metabolic Networks. We used 55 reconstructed bacterial genome-scale metabolic networks from the Biochemical Genetic and Genomic (BiGG) database (15), which provides comprehensive information about biochemical reactions, metabolites, metabolic genes, and gene reaction association rules for each bacterial species. We ordered the genes in each species based on their genomic location, as obtained from the RefSeq microbial genome database (44). We used the R-package Sybil (45) to parse the BiGG models.

Phenotype Prediction from Genomic Information. We focus our analyses on a qualitative definition of metabolic phenotypes; that is, on whether a given metabolism is viable or inviable in a given minimal chemical environment (medium) that contains only a single carbon source. More specifically, we consider a genotype viable if it can produce all essential biomass precursors from the resources in this medium. We use Flux Balance Analysis (FBA; *SI Appendix*, *Text S12*) (13) to predict viability on 103 minimal environments (*SI Appendix*, *Text S13*).

To systematically examine viability after deleting a metabolic gene or a set of metabolic genes, we used gene-reaction association rules for each species obtained from the BiGG database (15). On the basis of these rules, we translated metabolic gene deletions into deleted reactions. For more than 10% of reactions, genes and reactions do not show a one-to-one association. Some reactions are catalyzed by one or more enzymatic complexes, which may be encoded by more than one metabolic gene. In this case, deletion of a single gene whose product participates in a given enzymatic complex is enough to inactivate the complex (i.e., a Boolean AND function of gene presence/absence determines whether a reaction can be catalyzed). Other reactions can be catalyzed independently by multiple enzymatic complexes. In this case, all complexes need to be inactivated by deletion of individual genes to eliminate a reaction from the metabolic network (corresponding to a Boolean OR function of complex activity/inactivity). Finally, some gene products may participate in multiple enzymatic reactions, such that deletion of a single gene would eliminate multiple reactions. We took these associations into account when translating gene deletions into reaction deletions. After any one such deletion, we determined with FBA whether the resulting metabolic network is still viable in any one environment. The associated C++ codes are available through a public GitHub repository at https://github.com/RzgarHosseini/EMETNET.

Quantification of Robustness to Multiple Gene Deletions. To quantify the robustness of a given genome (metabolism) with n metabolic genes to "tandem deletions" of length l genes in a given environment (carbon source), we considered all possible (n) deletional variants in each of which l consecutive metabolic genes are deleted (*SI Appendix*, Fig. S1). For each deletional variant, we determined the reactions to be deleted from the wild-type metabolic network, based on the gene-reaction association rules (15). Subsequently, we determined the metabolic viability of each variant by FBA and quantified the robustness to tandem deletion as the fraction of deletional variants that retain viability on the given carbon source.

To quantify the robustness of a given genome (metabolism) with n metabolic genes to a "random deletion" of length l, in a given environment, we generated the same number n of deletional variants as for tandem deletions (*SI Appendix*, Fig. S1). In each of these variants, l randomly chosen metabolic genes in the genome are deleted (irrespective of their genomic location). We quantified robustness to random gene deletion with the same procedure described earlier, as the fraction of random deletional variants that retain viability on the carbon source. The associated C++ codes are available through a public GitHub repository at https://github.com/RzgarHosseini/EMETNET/tree/master/BIGG.

Quantification of Gene Essentiality. To determine whether a metabolic gene is essential for viability on a given carbon source, we removed the corresponding reaction or reactions from the wild-type metabolic network and determined viability using FBA. For each bacterial genome, we determined the essentiality of every metabolic gene in every environment on which the wild-type metabolism is viable. We consider a metabolic gene as strictly essential in a given genome if its deletion results in losing viability on all carbon sources on which the wild-type metabolism is viable, and we consider a metabolic gene as conditionally essential if its deletion abolishes viability on at least one carbon source. Note that strictly essential genes are a subset of conditionally essential genes.

Likewise, we call a metabolic gene strictly nonessential if its deletion does not abolish viability on any carbon source, and we indicate a metabolic gene as conditionally nonessential if its deletion does not abolish viability on at least one carbon source. Strictly nonessential genes are a subset of conditionally nonessential genes.

Quantification of the Clustering of Essential Genes in a Given Genome. We used Kuiper's test (19) to assess whether the distribution of essential genes in a given genome is uniform or not. This test is closely related to the Kolmogorov-Smirnov test, which computes the discrepancy statistics D^+ and D^- that represent the absolute sizes of the most positive and most negative differences between two cumulative probability distribution functions that are being compared. Because the Kolmogorov-Smirnov test is not invariant under cyclic transformations, it is not useful to detect clusters of genes distributed in a circular bacterial genome. Kuiper's test allows cyclic transformations while taking advantage of the D^+ and D^- test statistics.

Identification of Pairs of Synthetic Lethal Genes. For any given genome (metabolism), we identified all genes that are nonessential for viability in a given environment. Then, we examined all pairs of nonessential genes to determine whether simultaneous deletion of these genes is lethal. If yes, we consider the pair of genes as a synthetic lethal pair in this environment. We call a pair of genes that are synthetic lethal in all environments on which a wild-type metabolism is viable unconditionally synthetic lethal genes. Conversely, we call pairs of genes that are synthetically lethal in some but not all environments conditionally synthetically lethal. Finally, to identify nonessential clusters of nonessential metabolic genes, we first identified the set of adjacent nonessential genes intervening between two successive (but not adjacent) essential genes, and then we checked whether simultaneous deletion of the metabolic genes belonging to a given cluster of nonessential genes is lethal or not (see *SI Appendix*, *Text S4* for more details).

Generation of Minimal Metabolic Genome. We define a minimal metabolic genome as a set of metabolic genes of a given species that are all necessary to produce essential biomass precursors from external nutrients available in a given environment. To create a minimal genome, one needs to delete the nonessential genes step by step until no nonessential metabolic genes remain. Note that the size of minimal genome may be larger than the number of essential genes in a wild-type (full-sized) genome because after deleting a given gene, some previously nonessential genes may become essential. Moreover, genome size and gene identities in a minimal genome depend on the order of gene deletions that occur during genome reduction (24).

To generate a minimal genome from a given full-size genome, we apply a previously established stepwise stochastic algorithm (24). In each step, we remove a randomly chosen metabolic gene from the genome and determine the viability of the resulting metabolism in the given environment. If the metabolism is still viable (i.e., it can produce all biomass precursors), we accept the deletion and remove the gene from the genome; otherwise, the gene is restored to the genome. This procedure is repeated until no further genes can be deleted; that is, until all remaining metabolic genes are essential for survival in the given environment. We applied this procedure to three different genomes; namely, to *E. coli K-12 MG1655* (*iJO1366*), *Bacillus subtilis*, and *Salmonella enterica*, using glucose or acetate as carbon sources. From each genome and on each carbon source, we generated 100 different minimal genomes.

Identification of Operons in Bacterial Genomes. We used the DOOR database (31), which is a comprehensive database for prokaryotic operon information to identify metabolic genes that belong to an operon. It predicts operons based on a computational method (32) that was ranked first in an independent assessment of 14 operon prediction methods (46). For genomes with many experimentally validated operons, this method predicts operons based on a decision tree-based classifier that uses both genome-specific features, such as conserved gene neighborhood, phylogenetic profiles, and intergenic distances, and general features, such as the length ratio between a pair of adjacent genes, Gene ontology-based functional similarity between adjacent genes and the frequency of a specific DNA motif in the intergenic region. In contrast, for genomes with only limited experimental data on operons, the program applies a logistic function-based classifier using solely general genome features. The DOOR database contained operon information for 52 of our 55 bacterial genomes.

Identification of Horizontally Transferred Metabolic Genes. We used the HGTree database (36) to identify the metabolic genes that any one genome has likely obtained through horizontal gene transfer. In this database, horizontally transferred genes are predicted on the basis of a tree reconciliation method, which reconstructs approximate maximum likelihood phylogenetic trees for each orthologous gene and corresponding 16S rRNA reference species sets, and then reconciles the two trees using a maximum likelihood framework. This database harbors a more comprehensive set of HGT-acquired genes than others, because it relies on large-scale phylogenetic analysis of many distantly related bacterial species and can thus identify not only recent but also older HGT events (36). Because 43 of the 55 bacterial species we considered were included in this database, we focused this part of our analysis on 43 bacterial genomes.

ACKNOWLEDGMENTS. We thank the anonymous reviewers for their useful comments and suggestions. We acknowledge support through Swiss National Science Foundation Grants 31003A_146137 and 31003A_172887, as well as by European Research Council Advanced Grant 739874 and the University Priority Research Program in Evolutionary Biology at the University of Zurich.

1. Thomas CM, Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* 3:711–721.
2. Lynch M (2006) Streamlining and simplification of microbial genome architecture. *Annu Rev Microbiol* 60:327–349.
3. McCutcheon JP, Moran NA (2011) Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* 10:13–26.
4. Moran NA (2002) Microbial minimalism: Genome reduction in bacterial pathogens. *Cell* 108:583–586.
5. Wolf YI, Koonin EV (2013) Genome reduction as the dominant mode of evolution. *BioEssays* 35:829–837.
6. Albalat R, Cañestro C (2016) Evolution by gene loss. *Nat Rev Genet* 17:379–391.
7. Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet* 17:589–596.
8. Kunin V, Ouzounis CA (2003) The balance of driving forces during genome evolution in prokaryotes. *Genome Res* 13:1589–1594.
9. Nilsson AI, et al. (2005) Bacterial genome size reduction by experimental evolution. *Proc Natl Acad Sci USA* 102:12112–12116.
10. Lee MC, Marx CJ (2012) Repeated, selection-driven genome reduction of accessory genes in experimental populations. *PLoS Genet* 8:e1002651.
11. Koskiniemi S, Sun S, Berg OG, Andersson DI (2012) Selection-driven gene loss in bacteria. *PLoS Genet* 8:e1002787.
12. Sung W, et al. (2016) Evolution of the insertion-deletion mutation rate across the tree of life. *G3 (Bethesda)* 6:2583–2591.
13. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? *Nat Biotechnol* 28: 245–248.
14. McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol Syst Biol* 9:661.
15. King ZA, et al. (2015) BiGG models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res* 44:D515–D522.
16. Rocha EPC, Danchin A (2003) Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res* 31:6570–6577.
17. Fang G, Rocha E, Danchin A (2005) How essential are nonessential genes? *Mol Biol Evol* 22:2147–2156.
18. Fang G, Rocha EPC, Danchin A (2008) Persistence drives gene clustering in bacterial genomes. *BMC Genomics* 9:4.
19. Kuiper NH (1960) Tests concerning random points on a circle. *Indag Math* 63:38–47.
20. Mushegian A (1999) The minimal genome concept. *Curr Opin Genet Dev* 9:709–714.
21. Glass JI, et al. (2006) Essential genes of a minimal bacterium. *Proc Natl Acad Sci USA* 103:425–430.
22. Posfai G, et al. (2006) Emergent properties of reduced-genome Escherichia coli. *Science* 312:1044–1046.
23. Hutchison CA, et al. (2016) Design and synthesis of a minimal bacterial genome. *Science* 351:aad6253.
24. Pál C, et al. (2006) Chance and necessity in the evolution of minimal metabolic networks. *Nature* 440:667–670.
25. Hill CW, Gray JA (1988) Effects of chromosomal inversion on cell fitness in Escherichia coli K-12. *Genetics* 119:771–778.
26. Segall A, Mahan MJ, Roth JR (1988) Rearrangement of the bacterial chromosome: Forbidden inversions. *Science* 241:1314–1318.
27. Overbeek R, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33:5691–5702.
28. de Daruvar A, Collado-Vides J, Valencia A (2002) Analysis of the cellular functions of Escherichia coli operons and their conservation in Bacillus subtilis. *J Mol Evol* 55: 211–221.
29. Hershberg R, Yeger-Lotem E, Margalit H (2005) Chromosomal organization is shaped by the transcription regulatory network. *Trends Genet* 21:138–142.
30. Price MN, Huang KH, Arkin AP, Alm EJ (2005) Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res* 15:809–819.
31. Mao F, Dam P, Chou J, Olman V, Xu Y (2009) DOOR: A database for prokaryotic operons. *Nucleic Acids Res* 37:D459–D463.
32. Dam P, Olman V, Harris K, Su Z, Xu Y (2007) Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Res* 35:288–298.
33. Lathe WC, 3rd, Snel B, Bork P (2000) Gene context conservation of a higher order than operons. *Trends Biochem Sci* 25:474–479.
34. Lawrence JG, Roth JR (1996) Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics* 143:1843–1860.
35. Pál C, Hurst LD (2004) Evidence against the selfish operon theory. *Trends Genet* 20: 232–234.
36. Jeong H, et al. (2016) HGTree: Database of horizontally transferred genes determined by tree reconciliation. *Nucleic Acids Res* 44:D610–D619.
37. Pál C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* 37:1372–1375.
38. Karcagi I, et al. (2016) Indispensability of horizontally transferred genes and its impact on bacterial genome streamlining. *Mol Biol Evol* 33:1257–1269.
39. Oliveira PH, Touchon M, Cury J, Rocha EPC (2017) The chromosomal organization of horizontal gene transfer in bacteria. *Nat Commun* 8:841.
40. Hottes AK, et al. (2013) Bacterial adaptation through loss of function. *PLoS Genet* 9: e1003617.
41. Sokurenko EV, Hasty DL, Dykhuizen DE (1999) Pathoadaptive mutations: Gene loss and variation in bacterial pathogens. *Trends Microbiol* 7:191–195.
42. Maurelli AT, Fernández RE, Bloch CA, Rode CK, Fasano A (1998) "Black holes" and bacterial pathogenicity: A large genomic deletion that enhances the virulence of Shigella spp. and enteroinvasive Escherichia coli. *Proc Natl Acad Sci USA* 95: 3943–3948.
43. Moore RA, et al. (2004) Contribution of gene loss to the pathogenic evolution of Burkholderia pseudomallei and Burkholderia mallei. *Infect Immun* 72:4172–4187.
44. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I (2014) RefSeq microbial genomes database: New representation and annotation strategy. *Nucleic Acids Res* 42: D553–D559.
45. Gelius-Dietrich G, Desouki AA, Fritzemeier CJ, Lercher MJ (2013) Sybil–Efficient constraint-based modelling in R. *BMC Syst Biol* 7:125.
46. Brouwer RWW, Kuipers OP, van Hijum SAFT (2008) The relative value of operon predictions. *Brief Bioinform* 9:367–375.