Evolutionary assembly patterns of prokaryotic genomes

Maximilian O. Press,1 Christine Queitsch,1 and Elhanan Borenstein1,2,3

1Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA; 2Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195, USA; 3Santa Fe Institute, Santa Fe, New Mexico 87501, USA

Evolutionary innovation must occur in the context of some genomic background, which limits available evolutionary paths. For example, protein evolution by sequence substitution is constrained by epistasis between residues. In prokaryotes, evolutionary innovation frequently happens by macrogenomic events such as horizontal gene transfer (HGT). Previous work has suggested that HGT can be influenced by ancestral genomic content, yet the extent of such gene-level constraints has not yet been systematically characterized. Here, we evaluated the evolutionary impact of such constraints in prokaryotes, using probabilistic ancestral reconstructions from 634 extant prokaryotic genomes and a novel framework for detecting evolutionary constraints on HGT events. We identified 8228 directional dependencies between genes and demonstrated that many such dependencies reflect known functional relationships, including for example, evolutionary dependencies of the photosynthetic enzyme RuBisCO. Modeling all dependencies as a network, we adapted an approach from graph theory to establish chronological precedence in the acquisition of different genomic functions. Specifically, we demonstrated that specific functions tend to be gained sequentially, suggesting that evolution in prokaryotes is governed by functional assembly patterns. Finally, we showed that these dependencies are universal rather than clade-specific and are often sufficient for predicting whether or not a given ancestral genome will acquire specific genes. Combined, our results indicate that evolutionary innovation via HGT is profoundly constrained by epistasis and historical contingency, similar to the evolution of proteins and phenotypic characters, and suggest that the emergence of specific metabolic and pathological phenotypes in prokaryotes can be predictable from current genomes.

A fundamental question in evolutionary biology is how present circumstances affect future adaptation and phenotypic change (Gould and Lewontin 1979). Studies of specific proteins, for example, indicate that epistasis between sequence residues limits accessible evolutionary trajectories and thereby renders certain adaptive paths more likely than others (Weinreich et al. 2006; Gong et al. 2013; de Visser and Krug 2014; Harms and Thornton 2014). Similarly, both phenotypic characters (Ord and Summers 2015) and specific genetic adaptations (Conte et al. 2012; Christin et al. 2015) show strong evidence of parallel evolution rather than convergent evolution. That is, a given adaptation is more likely to repeat in closely related organisms than in distantly related ones. This inverse relationship between the repeatability of evolution and taxonomic distance implies a strong effect of lineage-specific contingency on evolution, also potentially mediated by epistasis (Orr 2005).

Such observations suggest that genetic adaptation is often highly constrained, and the present state of an evolving system can impact future evolution. Yet, the aforementioned studies are limited to small data sets and specific genetic pathways, and a more principled understanding of the rules by which future evolutionary trajectories are governed by the present state of the system is still lacking. For example, it is not known whether such adaptive constraints are a feature of genome-scale evolution or whether they are limited to finer scales. Moreover, the mechanisms that underlie observed constraints are often completely unknown. Addressing these questions is clearly valuable for obtaining a more complete theory of evolutionary biology, but more pressing, is essential for tackling a variety of practical concerns, including our ability to combat evolving infectious diseases or engineer complex biological systems.

Here, we address this challenge by analyzing horizontal gene transfer (HGT) in prokaryotes. HGT is an ideal system to systematically study genome-wide evolutionary constraints because it involves gene-level innovation, occurs at very high rates relative to sequence substitution (Nowell et al. 2014; Puigbò et al. 2014), and is a principal source of evolutionary novelty in prokaryotes (Gogarten et al. 2002; Jain et al. 2003; Lerat et al. 2005; Puigbò et al. 2014). Clearly, many or most acquired genes are rapidly lost due to fitness costs (van Passel et al. 2008; Baltrus 2013; Soucy et al. 2015), indicating that genes retained in the long term are likely to provide a selective advantage. Moreover, not all genes are equally transferable (Jain et al. 1999; Sorek et al. 2007; Cohen et al. 2011), and not all species are equally receptive to the same genes (Smillie et al. 2011; Soucy et al. 2015). However, differences in HGT among species have been attributed not only to ecology (Smillie et al. 2011) or to phylogenetic constraints (Popa et al. 2011; Nowell et al. 2014), but also to interactions with the host genome (Jain et al. 1999; Cohen et al. 2011; Popa et al. 2011). Indeed, studies involving single genes or single species support the influence of genome content on the acquisition and retention of transferred genes (Pal et al. 2005; Sorek et al. 2007; Iwasaki and Takagi 2009; Chen et al. 2011; Press et al. 2013; Iwasaki and Takagi 2009; Chen et al. 2011; Press et al. 2013; Smillie et al. 2011; Soucy et al. 2015).
Johnson and Grossman 2014). For example, it has been demonstrated that the presence of specific genes facilitates integration of others into genetic networks (Chen et al. 2011), and genes are more commonly gained in genomes already containing metabolic genes in the same pathway (Pal et al. 2005; Iwasaki and Takagi 2009). However, to date, a systematic, large-scale analysis of such dependencies has not been presented. In this paper, we therefore set out to characterize a comprehensive collection of genome-wide HGT-based dependencies among prokaryotic genes, analyze the obtained set of epistatic interactions, and identify patterns in the evolution of prokaryotic genomes.

Results
PGCE inference
We first set out to detect pairs of genes for which the presence of one gene in the genome promotes the gain of the other gene (though not necessarily vice versa) (Fig. 1). Such “pairs of genes with conjugated evolution” (PGCEs) represent putative epistatic interactions at the gene level and may guide genome evolution. To this end, we obtained a collection of 634 prokaryotic genomes annotated by KEGG (Kanehisa et al. 2012) and linked through a curated phylogeny (Dehal et al. 2010). For each of the 5801 genes that varied in presence across these genomes, we reconstructed the probability of this gene’s presence or absence on each branch of the phylogenetic tree using a previously introduced method (Cohen and Pupko 2010), as well as the probability that it was gained or lost along these branches using a simple heuristic (Methods). We confirmed that genes’ presence/absence was robust to the reconstruction method used (99.5% agreement between reconstruction methods used) (Methods). As expected (Mira et al. 2001), gene loss was more common than gene gain for most genes (Supplemental Fig. S1; Supplemental Text). We additionally confirmed that inferred gains of several genes of interest were consistent with gains inferred by an alternative HGT inference method (Methods; Supplemental Text; Supplemental Table S1). From the reconstructions, we estimated the frequency with which each gene was gained in the presence of each other gene, and followed previous studies (Maddison 1990; Cohen et al. 2012) in using parametric bootstrapping (Supplemental Fig. S2) to detect PGCEs—gene pairs for which one gene is gained significantly more often in the presence of the other (Supplemental Fig. S3; Supplemental Text). In total, we identified 8415 PGCEs. We finally applied a transitive reduction procedure to discard potentially spurious PGCEs, resulting in a final network containing 8228 PGCEs connecting a total of 2260 genes (Supplemental Figs. S4, S5; Supplemental Text). A detailed description of the procedures used can be found in Methods, and the final list of PGCEs is supplied as Supplemental File S1.

PGCEs represent biologically relevant dependencies
Comparing this final set of PGCEs to known biological interactions, we confirmed that the obtained PGCEs represent plausible biological dependencies. For example, genes sharing the same KEGG Pathway annotations were more likely to form a PGCE (Fig. 2A), as were genes linked in an independently derived network of bacterial metabolism (Fig. 2B; Levy and Borenstein 2013). Moreover, PGCEs often linked genes in functionally related pathways (Supplemental Fig. S6; Supplemental Text). We similarly identified specific examples in which PGCEs connected pairs of genes with well-described functional relationships. One such example is the PGCE connecting rbsL and rbsS (sometimes written rbcL/rbcS), two genes that encode the large and small subunits of the well-described photosynthetic enzyme ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCO), respectively. The rbsL subunit alone has carboxylation activity in some bacteria, but the addition of rbsS increases enzymatic efficiency, consistent with its PGCE dependency on rbsL (Fig. 3A; Andersson and

Figure 2. PGCEs are enriched for biologically meaningful interactions. (A) The observed number of PGCE edges connecting genes in the same pathway (dashed line), compared to the expected distribution obtained from 1000 rewired networks with identical degree distributions. (B) The observed number of PGCE edges that also appear in a bacteria-wide metabolic network, compared to the expected distribution.
Figure 3. The phylogenetic history of rbsL, urtA, and rbsS. The presence of each gene in each branch in the phylogenetic tree is illustrated with a colored circle, with the circle's diameter scaled to denote the probability of presence. (A) rbsL and rbsS evolutionary histories. (B) urtA and rbsS evolutionary histories. The long branch leading to Archaea (bottommost clade) was reduced in size for graphical purposes.

Multiple additional genes were found to promote rbsS gain (88 PGCEs in total) (Supplemental Table S2), many of which, as expected, are associated with carbon metabolism. Other genes in this set, however, unexpectedly implicated nitrogen acquisition, as well as other pathways (Supplemental Table S3), in promoting rbsS gain. For example, all components of the Urt urea transport complex had a PGCE link with rbsS, as shown by the reconstructed phylogenetic history of urtA and rbsS (Fig. 3B). This strict dependency could reflect nitrogen’s role as a rate-limiting resource for primary production in phytoplankton and other photosynthetic organisms (Eppley and Peterson 1979; Sohm et al. 2011). In comparing the reconstructions from which urtA-rbsS and rbsL-rbsS dependencies were inferred, we further observed that rbsS is gained only in lineages in which both urtA and rbsL were previously present. This indicates that although both rbsL and urtA may be necessary for the acquisition of rbsS, neither rbsL nor urtA are independently sufficient for the acquisition of rbsS. Other PGCEs may interact in similarly complex fashions in controlling the acquisition of genes, and thus such relationships may be gene-specific and involve a variety of biological mechanisms that may be difficult to generalize. For further analyses, we therefore focused on analyzing large-scale patterns of PGCE connectivity and on exploring how the dependencies between various genes structure the relationships between functional pathways.

PGCE network analyses reveal evolutionary assembly patterns

The rbsS-associated PGCEs described above show how PGCEs captured an assembly pattern involving multiple pathways. Therefore, we next set out to infer global evolutionary assembly patterns based on the complete set of PGCEs identified. Specifically, we used a network-based topological sorting approach (Supplemental Text) to rank all genes in the PGCE network. According to this procedure, genes without dependencies occupy the first rank, genes in the second rank have PGCE dependencies only on first rank genes, genes in the third rank have dependencies only on first and second rank genes, and so on until all genes are associated with some rank. In other words, the obtained ranking represents general patterns in the order by which genes are gained throughout evolution, with the gain of higher-ranked genes succeeding the presence of the lower-ranked genes on which they depend. Using this approach, we found that genes could be fully classified into five ranks (Fig. 4A). The first rank was by far the largest at 1593 genes (most genes do not have detectable dependencies), the second rank had 498 genes, and successive ranks showed declining membership until the last (fifth) rank, with only five genes (Supplemental Table S4).

To identify evolutionary assembly patterns from these ranks, we examined the set of genes in each rank and identified overrepresented functional
patterns in gain order. The chronology of gene acquisition reflects the overall assembly lower-ranked genes (Fig. 4B; Supplemental Fig. S8), suggesting that served that the gains of genes appearing late in the sort were over-gain (Supplemental Fig. S7B). Furthermore, as expected, we observed that the gains of genes encoding pillar proteins) and certain two-component genes. For example, we found that the first rank was enriched for flagellar system, whereas later ranks were enriched for Type III and Type IV secretion genes (many of which are homologous to or overlap with genes encoding pillar proteins) and certain two-component genes. The second rank was enriched for various metabolic processes, whereas later ranks were enriched for Type III protein secretion pathway and Type II secretion system. Rank 4 enrichments are gained, and replaced frequently in response to evolving challenges. Considering two distinct but highly homologous pilus assembly systems (rather than randomly sampling species from throughout the tree) guarantees that true predictions are based on universal PGCEs, rather than clade-specific PGCEs. For each test set, we used a model phylogeny that excluded the test subtree as a training set and inferred PGCEs based on this pruned tree (Supplemental Table S5; Supplemental Fig. S9A). We then used these inferred PGCEs to score the relative likelihood of the gain of dependent genes on each branch in the test set, based on the genome content of the branch’s ancestor (Fig. 5A; Supplemental Table S5; Supplemental Text). We used a naïve and simplistic score: the proportion of genes upon which the gained gene depends that are present in the reconstructed ancestor of each branch. In both test sets, we found that prediction quality was surprisingly high (Fig. 5B; Supplemental Fig. S9B,C), suggesting that PGCEs are taxonomically universal and statistically robust in describing relationships between genes. This predictability is consistent with the hypothesis that gene–gene dependencies constrain the evolution of genomes by HGT. More broadly, this analysis and our finding that PGCEs can predictably determine future evolutionary gains provide substantial evidence that the preponderance of parallel evolution over convergent evolution (Conte et al. 2012; Ord and Summers 2015) may be the result of specific, identifiable genetic dependencies entraining the evolutionary trajectory taken by similar genomes.

**Evolution by HGT is predictable**

The chronological ordering of ranks was relatively consistent across the tree (Fig. 4B), indicating that PGCE dependencies are universal across prokaryotes. Notably, this universality also implies that gene acquisition is predictable from genome content. Put differently, if PGCEs are universal, then PGCEs inferred in one clade of the tree are informative in making predictions about gene acquisition in a different clade. Indeed, studies of epistasis-mediated protein evolution indicate that the constriction of possible mutational paths should lead to predictability in evolution, if epistasis is sufficiently strong (Weinreich et al. 2006). To explore this hypothesis explicitly, we partitioned the tree into training and test sets (Fig. 5A). As test sets, we selected the Firmicutes phylum, and the Alphaproteobacteria/Betaproteobacteria subphyll. Choosing whole clades as test sets (rather than randomly sampling species from throughout the tree) guarantees that true predictions are based on universal PGCEs, rather than clade-specific PGCEs. For each test set, we used a model phylogeny that excluded the test subtree as a training set and inferred PGCEs based on this pruned tree (Supplemental Table S5; Supplemental Fig. S9A). We then used these inferred PGCEs to score the relative likelihood of the gain of dependent genes on each branch in the test set, based on the genome content of the branch’s ancestor (Fig. 5A; Supplemental Table S5; Supplemental Text). We used a naïve and simplistic score: the proportion of genes upon which the gained gene depends that are present in the reconstructed ancestor of each branch. In both test sets, we found that prediction quality was surprisingly high (Fig. 5B; Supplemental Fig. S9B,C), suggesting that PGCEs are taxonomically universal and statistically robust in describing relationships between genes. This predictability is consistent with the hypothesis that gene–gene dependencies constrain the evolution of genomes by HGT. More broadly, this analysis and our finding that PGCEs can predictably determine future evolutionary gains provide substantial evidence that the preponderance of parallel evolution over convergent evolution (Conte et al. 2012; Ord and Summers 2015) may be the result of specific, identifiable genetic dependencies entraining the evolutionary trajectory taken by similar genomes.

**Figure 5.** PGCE dependencies lead to taxonomically robust predictability of gene acquisition. (A) Workflow for predicting gene acquisition between clades of the tree. A training set is used to build a PGCE dependency model, which is then used to predict on which specific branches genes are likely to be gained (green circles), based on dependencies inferred from the training set (red and blue circles). (B) Performance of PGCEs in predicting gene acquisitions in two test sets (indicated clades of the prokaryotic tree). Areas under each curve: Firmicutes, 0.73; Alpha/Beta-proteobacteria, 0.68. The diagonal dotted line represents the performance of a purely random prediction. See also Supplemental Figure S9.

---

**Table 1. Functional groups are enriched in different ranks of the topological sort.**

<table>
<thead>
<tr>
<th>Annotation label</th>
<th>P-value</th>
<th>Enrichment ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank 1 enrichments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell motility</td>
<td>1.94 × 10⁻⁷</td>
<td>1.40</td>
</tr>
<tr>
<td>Bacterial motility proteins</td>
<td>1.85 × 10⁻¹¹</td>
<td>1.41</td>
</tr>
<tr>
<td>Type II secretion system</td>
<td>2.61 × 10⁻³</td>
<td>1.33</td>
</tr>
<tr>
<td>Two-component system</td>
<td>3.65 × 10⁻³</td>
<td>1.25</td>
</tr>
<tr>
<td>Flagellar system</td>
<td>1.01 × 10⁻⁹</td>
<td>1.43</td>
</tr>
<tr>
<td>Pilius system</td>
<td>2.11 × 10⁻⁴</td>
<td>1.38</td>
</tr>
<tr>
<td>Metabolism</td>
<td>3.37 × 10⁻²</td>
<td>0.91</td>
</tr>
<tr>
<td>Xenobiotics biodegradation and metabolism</td>
<td>1.07 × 10⁻⁴</td>
<td>0.69</td>
</tr>
<tr>
<td>Carbohydrate metabolism</td>
<td>1.20 × 10⁻⁴</td>
<td>0.84</td>
</tr>
<tr>
<td>Type IV secretion system</td>
<td>1.26 × 10⁻⁹</td>
<td>0.20</td>
</tr>
<tr>
<td>Rank 2 enrichments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolism</td>
<td>1.47 × 10⁻⁴</td>
<td>1.23</td>
</tr>
<tr>
<td>Carbohydrate metabolism</td>
<td>3.08 × 10⁻⁶</td>
<td>1.58</td>
</tr>
<tr>
<td>Rank 4 enrichments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pathogenicity</td>
<td>1.88 × 10⁻⁴</td>
<td>21.65</td>
</tr>
<tr>
<td>Conjugal transfer pilus assembly protein</td>
<td>1.08 × 10⁻⁴</td>
<td>15.01</td>
</tr>
<tr>
<td>Type III protein secretion pathway protein</td>
<td>1.88 × 10⁻⁶</td>
<td>21.65</td>
</tr>
<tr>
<td>ABC-2 type and other transporters</td>
<td>2.31 × 10⁻⁴</td>
<td>12.51</td>
</tr>
<tr>
<td>Type IV secretion system</td>
<td>1.30 × 10⁻⁷</td>
<td>8.04</td>
</tr>
</tbody>
</table>

*From a hypergeometric test. All annotations displayed are significant at a 1% false discovery rate.

The ratio of the observed proportion of genes with this label in the indicated rank to the expected proportion based on all genes in the network.

*These annotations are depleted (i.e., enrichment ratio significantly less than one) in the first rank.

Categories (Table 1). These enriched functional categories indicate that certain functional groups of genes consistently occupy specific positions in these evolutionary assembly patterns, whether in controlling other genes’ gain or in being controlled by other genes. For example, we found that the first rank was enriched for flagellar and pillar genes involved in motility, in addition to Type II secretion genes (many of which are homologous to or overlap with genes encoding pillar proteins) and certain two-component genes. The second rank was enriched for various metabolic processes, whereas later ranks were enriched for Type III and Type IV secretion systems and conjugation genes. This finding suggests that habitat commitments are made early in evolution, mediated by motility genes that could underlie the choice and establishment of physical environments. This environmental choice is followed by a metabolic commitment to exploiting the new habitat. Last, genes for interaction with the biotic complement of these habitats are gained, and replaced frequently in response to evolving challenges. Considering two distinct but highly homologous pilus assembly pathways, one (fimbrial) was enriched in a low rank and one (conjugal) was enriched in a high rank, suggesting that the specific function of the gene rather than other sequence-level gene properties drove the ranking (Supplemental Fig. S7A). We additionally confirmed that the observed rank distribution for these functions is not explained by variation in the frequency of gene gain (Supplemental Fig. S7B). Furthermore, as expected, we observed that the gains of genes appearing late in the sort were over-represented in later branches of the tree compared to the gains of lower-ranked genes (Fig. 4B; Supplemental Fig. S8), suggesting that the chronology of gene acquisition reflects the overall assembly patterns in gain order.
Discussion

Combined, our findings provide substantial evidence to suggest that gene acquisitions in bacteria are governed by genome content through numerous gene-level dependencies. Our ability to detect these underlying dependencies is clearly imperfect, owing to various data and methodological limitations (Supplemental Text; Supplemental Fig. S3). Therefore, in reality, the complete dependency network is likely much denser than that described above and includes numerous dependencies and constraints that our approach may not be able to detect. Consequently, our estimates should be considered as a lower bound on the extent of gene–gene interactions, and accordingly, the predictability of HGT.

Notably, even considering such caveats, our observations dramatically expand our knowledge of the constraints on HGT. Previous studies of such constraints demonstrated that genes frequently acquired by HGT tend to occupy peripheral positions in biological networks, are often associated with specific cellular functions, and are phylogenetically clustered (Jain et al. 1999; Cohen et al. 2011). These observations suggested that properties of transferred genes are also important determinants of HGT regardless of recipient genome content (Jain et al. 1999; Cohen et al. 2011; Gophna and Ofran 2011) and that the acquisition of certain genes is clade-specific (Andam and Gogarten 2011; Popa et al. 2011). In contrast, our analysis demonstrates the importance of recipient genome content in influencing the propensity of a new gene to be acquired. In fact, to some extent, properties previously reported as determining the general “acquirability” of genes across all species may reflect an average constraint across genomes. By also considering variation in genomes acquiring genes, our analysis focused on specific biological effects, whose strengths may vary from genome to genome.

Importantly, our model that gene acquisition is affected by recipient genome content is consistent with the observed enrichment of HGT among close relatives, which presumably have similar genome content (Gogarten et al. 2002; Andam and Gogarten 2011; Popa and DaGan 2011; Popa et al. 2011). This taxonomic clustering of innovation by HGT is also in agreement with previous studies that demonstrated that phenotypic and genetic parallel evolution is more common than convergent evolution, regardless of recipient genome content (Jain et al. 1999; Cohen et al. 2011). These observations suggested that properties of transferred genes are also important determinants of HGT regardless of recipient genome content (Jain et al. 1999; Cohen et al. 2011; Gophna and Ofran 2011) and that the acquisition of certain genes is clade-specific (Andam and Gogarten 2011; Popa et al. 2011). This taxonomic clustering of innovation by HGT is also in agreement with previous studies that demonstrated that phenotypic and genetic parallel evolution is more common than convergent evolution, potentially due to the effects of historical contingency (Gould and Lewontin 1979; Conte et al. 2012; Christin et al. 2015; Ord and Summers 2015). However, in contrast to other studies, we present direct evidence that the mechanism by which contingency controls evolution is epistasis. Furthermore, the universality of PGCEs shows that the constraints underlying the effect of contingency operate outside the context of parallel evolution.

Put differently, since each phylum-level clade is subject to an independent evolutionary trajectory, it is unlikely that the same dependency patterns would repeat solely due to parallel evolution. Moreover, our ability to predict where exactly along the tree gains of a specific gene are likely to occur (Fig. 5B) suggests that PGCEs successfully capture how variation in the genomic content (even among closely related species) affects future gain events. Such PGCE specificity therefore indicates that observed dependencies are not a trivial byproduct of prevalent gene transfer events among taxonomically closely related genomes (e.g., due to homologous recombination constraints) (Popa et al. 2011). Nonetheless, the relative contributions of each of these various processes governing the assembly of prokaryotic genomes (and the evolution of complex systems in general) clearly deserve future study.

Although our analysis revealed several intriguing patterns, the precise interpretation of some of these patterns remains unclear. For instance, the observed correspondence of topological ranks of genes to chronology suggests that evolutionary age is a potential contributor to such ranking, especially considering that our reconstructions likely lack many genes that have not been retained in any extant genomes. However, the biological plausibility and statistical robustness of PGCEs demonstrated above strongly argue that the observed evolutionary patterns are the result of constraint-inducing dependencies. Future work may therefore aim to quantify the trade-off between functional and chronological determinants in apparent evolutionary constraints.

Finally, we demonstrate the predictability of genomic evolution by horizontal transfer from current genomic content. As stated above, this finding also suggests that such dependencies are fairly universal across the prokaryotic tree. Our approach was designed specifically to understand the PGCE network’s significance and universality, rather than predict gene acquisition. It is likely that an approach specifically engineered for gene acquisition prediction would substantially outperform our approach. The estimates of predictability of genomic evolution presented here are accordingly quite conservative.

The determinism and predictability of evolutionary patterns therefore appear to be an outcome not only of intramolecular epistasis in proteins or phylogenetic constraints, but also of genome-wide interactions between genes. This suggests that the evolution of medically, economically, and ecologically important traits in prokaryotes depends on ancestral genome content and is hence at least partly predictable, potentially informing research in the epidemiology of infectious diseases, bioengineering, and biotechnology.

Methods

All mathematical operations and statistical analyses were performed in R 2.15.3 (R Core Team 2016). Probabilistic ancestral reconstructions were obtained using the gainLoss program (Cohen and Popko 2010). Phylogenetic simulations and plots were performed with the APE library (Paradis et al. 2004). Network analyses and algorithms were implemented using either the graph (Csárdi and Nepusz 2006) or NetworkX (https://networkx.lanl.gov) libraries, and visualized using Cytoscape v3.1.1 (Shannon et al. 2003).

Phylogeny

We used a precomputed phylogenetic tree (Dehal et al. 2010) as a model of bacterial evolution. We mapped all extant organisms in this tree to organisms in the KEGG database by their NCBI genome identifiers and pruned all tips that did not directly and uniquely map to KEGG. This yielded a phylogenetic tree connecting 634 prokaryotic species. For analyses involving subtrees of this phylogenetic tree, we used iTOL (Letunic and Bork 2011) to extract subtrees.

Inferring phylogenetic histories for genes

We used the gainLoss v1.266 software (Cohen and Popko 2010), a set of presence/absence patterns of orthologous genes from KEGG (Kanehisa et al. 2012), and the phylogenetic tree described above to infer (1) the probabilities of presence and absence of genes at internal nodes of the tree; (2) gain and loss rates of each gene; and (3) tree branch lengths within a single model. Specifically, in running gainLoss, we assumed a stationary evolutionary process, with gene gain and loss rates for each gene modeled as a mixture of three rates
Evolutionary assembly of prokaryotic genomes

drawn from gamma distributions defined based on overall initial presence/absence patterns. A complete list of parameters used for gainLoss runs is given in the Supplemental Text and as Supplemental File S2. The gainLoss log file for the principal run on the full tree is also included as Supplemental File S3. Based on these models, we obtained a probabilistic ancestral reconstruction based on stochastic mapping for each of 5801 genes that were present in at least one species and absent in at least one species and filtered out genes that were found to be gained less than twice throughout the tree, yielding 5031 genes, which we further analyzed.

Inferring gains and presence of genes on branches
To focus on gain events with strong support and where the gained gene is retained (rather than gain events in which the gene is subsequently lost along the same branch), we used a simple model for computing the probability of different evolutionary gain/loss scenarios based on gainLoss ancestral reconstructions rather than directly using gainLoss gain inferences (Supplemental Text). Specifically, we assumed that unobserved gains and losses are not relevant, and evolutionary scenarios are defined by the states at the ancestor and descendant nodes of each branch (regardless of branch length). With these assumptions, we used the probabilities of presence and absence of each of 5031 genes at each node and tip on the tree to compute the probability of each branch undergoing each scenario: (1) gain (absent in ancestor and present in descendant); (2) presence (present in both ancestor and descendant); and (3) loss (present in ancestor and absent in descendant) (Supplemental Text). For a gene X on a branch with ancestor A and descendant B, we assume the following:

1. Pr(X present on branch) = Pr(X present in A \cap X present in B) = Pr(X present in A) \times Pr(X present in B);
2. Pr(X gained on branch) = Pr(X absent in A \cap X present in B) = Pr(X absent in A) \times Pr(X present in B); and
3. Pr(X lost on branch) = Pr(X present in A \cap X absent in B) = Pr(X present in A) \times Pr(X absent in B).

Note again that these probability estimates are distinct from those obtained by using the gainLoss continuous-time Markov chain on the same ancestral reconstruction, which consider also hypothetical gains that are not retained and are thus not relevant to our analysis (Supplemental Text).

Robustness analysis of reconstruction method
We used a maximum-parsimony reconstruction as inferred by gainLoss to benchmark the accuracy of the gainLoss reconstruction by stochastic mapping. In this analysis, only internal node reconstructions were considered, as tip reconstructions (for which the states are known) are not informative about algorithm performance. Since the maximum-parsimony reconstruction is binary (presence/absence) and the stochastic mapping reconstruction is probabilistic, for purposes of comparison we rounded the probabilities of the stochastic mapping reconstruction to obtain a presence/absence reconstruction (i.e., a probability >0.5 denotes presence and ≤0.5 denotes absence). We computed the agreement between the two reconstructions as the percentage of internal node reconstructions that agree on the state of the gene.

Comparison of analyzed gains to reconciliation-based HGT inference
We compared gains inferred by our method for several genes central to the PGCE network to gain events reported in a searchable database of horizontally acquired genes inferred by a sequence-based reconciliation method (Jeong et al. 2015). To this end, we classified all branches supporting a gain event for each of these genes with >50% probability by our method as “true” gains. We next searched the reconciliation database (all queries performed between January 15 and February 20, 2016) for each gene, identifying orthologous genes across 2472 genomes that exhibit HGT according to reconciliation (excluding events that occurred on branches without descendants). We manually compared descendants of the remaining events from our method with the genomes experiencing gene acquisition in the reconciliation data set to assess overlap between these two methods (see Supplemental Text).

Quantifying PGCEs
We defined a “pair of genes with conjugated evolution” (PGCE) as a gene pair (i, j) for which the presence of one gene i encourages the gain of the other, j. Considering these genes as phylogenetic characters, we therefore aim to detect pairs for which “gain” state transitions for character i are enriched on branches where character j remains in the “present” state. This problem is related to previous methods for detecting coevolution or correlation between phylogenetic characters (Maddison 1990; Huelsenbeck et al. 2003; Cohen et al. 2012). Given N branches and k genes, there are 2 N × k matrices, P and G, describing the probabilities, respectively, of presence and gain of each gene along each branch (using our model for estimating gains described above). The test statistic for a dependency between each gene pair (i, j) is the expected number of branches where the gain of gene j occurs, while conditioning on the presence of gene i (cell Cij in a k × k matrix C). Counting transitions of one character (gene j gain) given some state of another character (gene i presence) yields a standard test statistic for testing correlated evolution of binary characters on phylogenies (Maddison 1990). To compute C across N branches, we sum the conditional probabilities of the gain of gene j in the presence of gene i across the tree, i.e., the products of the two N × k matrices, P (presence) and G (gain), for each gene pair

\[ C_{ij} = \sum_{n=1}^{N} C_{nij}. \]

Entries in C which are significantly larger than a null expectation of gains represent PGCEs between the row and column genes of C.

Null distribution for PGCEs
For two independently evolving genes i and j, the counted gains of j in the presence of i, C_{ij}, will be distributed under the null hypothesis (independent evolution) as some function of the prevalence of j (the sum of P_{i}, the vector of probabilities of presence of i across branches of the tree), the expected number of branches where j is gained (the sum of G_{i}, the vector of probabilities of gains of j across nodes of the tree), and the topology and branch lengths of the tree (\tau)

\[ C_{ij} \sim f(P_{i}, G_{i}, \tau). \]

We followed previous studies (Maddison 1990; Huelsenbeck et al. 2003; Cohen et al. 2012) by approximating this null distribution via parametric bootstrapping. Specifically, we simulated the evolution of 10^3 genes along the tree using the APE library function rTraitDisc() (Paradis et al. 2004). For the gain and loss rates used in these simulations, we used gainLoss gain and loss rates estimated for the S801 empirical genes. We fit gamma distributions to these values by maximum likelihood using the function fldistri() from the MASS library (Venables and Ripley 2002). For both gains and losses, we increased the shape parameter of the gamma distribution (by a factor of 3 for gains, 1.5 for losses) to ensure that
simulated genes showed sufficiently large numbers of gains. This was necessary because parametric bootstrapping with the rates inferred by gainLoss resulted in left skewed distributions of gene gains (cf. Supplemental Fig. S2A,C,E), which were likely to confound null models. For our null models to be applicable, the distribution of simulated gene gains should be roughly similar to the distribution of gains among empirical genes (see Supplemental Fig. S2; Supplemental Text).

These simulated genes should evolve independently and thus represent a null model for PGCEs. As above, we constructed matrices representing the probabilities of presence and gain of these 10^5 genes across all of the branches of the phylogeny (Pnull and Gnull). We then multiplied these matrices of simulated genes to compute a 10^5 × 10^5 matrix Cnull of expected branch counts under a model of independence. We excluded gene pairs with Cij ≤ 1 from further analysis, as it may be difficult to distinguish between no association and a lack of statistical power for such pairs (Supplemental Fig. S3A), reducing overall power in computing false discovery rates (Bourgon et al. 2010). As a null distribution for each pair of genes i and j with Cij > 1, we used the 1000 simulated genes with prevalence closest to gene i (rows of Cnull) and the 1000 simulated genes with a number of gains closest to gene j (columns of Cnull). We used the 10^6 simulated observations in the resulting submatrix of Cnull as a null distribution for Cij. Notably, Cij includes noninteger count expectations, whereas Cnull represents integer counts (because the true reconstruction is known). Consequently, we floored values in Cij such that all counts were truncated at the decimal point. The comparison of Cij to this null distribution yields an empirical P-value; we rejected the null hypothesis of independence between genes i and j for the Cij observation at a 1% false discovery rate (P < 7 × 10^-5) (Benjamini and Hochberg 1995).

Constructing a PGCE network

For each entry in Cij for which we observed a significant association, we recorded an edge from gene i to gene j in a network of PGCEs. To focus purely on direct interactions, we subjected this network to a transitive reduction (Hsu 1975). This reduction requires a directed acyclic graph (DAG). To identify the largest possible DAG in our PGCE network, we identified and removed the minimal set of edges inducing cycles (Supplemental Text). We performed a transitive reduction of the resulting DAG using Hsu’s algorithm (Supplemental Text; Hsu 1975).

Mapping biological information to the network

We used network rewiring (as implemented in the rewire() function of the igraph library) (Csárdi and Nepusz 2006) to generate null distributions of the PGCE network by randomly exchanging edges between pairs of connected nodes, while excluding self-edges. In each permutation, we performed SN rewiring operations, where there are N edges in the network, to ensure sufficient randomization. To estimate the relationship between the PGCE network and biological information, we calculated the number of edges shared between the PGCE network and a metabolic network of all bacterial metabolism obtained from KEGG (Kanehisa et al. 2012; Levy and Borenstein 2013), and the number of edges shared between members of the same functional pathway as defined by KEGG, in both the original and randomized networks.

To determine whether genes with certain functional annotations were more likely to associate with one another in the PGCE network, we examined the KEGG Pathway annotations of each pair of genes in the network. We counted the number of edges leading from each pathway to each other pathway and obtained an empirical P-value for this count by comparing it to a null distribution of the expected counts obtained by random rewiring as above.

Topological sorting of PGCE networks

To identify global patterns in our PGCE network, we performed topological sorting (Kahn 1962) with groupit. Topological sorting finds an absolute ordering of nodes in a directed acyclic graph (DAG), such that no node later in the ordering has an edge directed toward a node earlier in the ordering. Grouping the sort allows nodes to have the same rank in the ordering if precedence cannot be established between them, giving a unique solution. For a description of the algorithm used, see Supplemental Text.

Prediction of HGT events on branches

We used the PGCE network to predict the occurrence of specific HGT events (gene acquisitions) on the tree in the following fashion. We used two test/training set partitions, with the clades of Firmicutes and the Alpha/Beta-proteobacteria as independent test sets and the training sets as the rest of the tree without these clades. To “train” PGCE networks, we performed ancestral reconstruction of gene presence, PGCE inference, and network processing just as for the entire tree. We only attempted to predict genes with at least one PGCE dependency (“predictable” genes). We then considered each branch in the test set independently, attempting to predict whether each predictable gene was gained on that branch based on the reconstructed genome at the ancestor node. For each predictable gene-branch combination, our prediction score was the proportion of the predictable gene’s PGCE dependencies that are present in the ancestor. This is the dot product of the gene presence/absence pattern of the ancestor node (Ai across i potentially present genes) and a binary vector denoting which genes in the PGCE network the predictable gene depends on (Pj across j genes in potential PGCEs), scaled by Pj:

\[
\text{score} = \frac{\sum A_i P_j}{\sum P_j}.
\]

Note that this value ranges between 0 and 1 for each predicted gene. As true gains, we used our reconstructed gene acquisition events for each branch in the test set. We arbitrarily called any predictable gene-branch pair with Pr(gain) > 0.5 as a gain, and any predictable gene-branch pair with Pr(gain) ≤ 0.5 as no gain. We filtered out any gene-branch pair in which the gene was known to be present with Pr > 0.4, because in these cases, the gene is probably already present. We analyzed the accuracy of our prediction scores using receiver operating characteristic (ROC) analysis and by comparing scores of the gain branches to those of the no-gain branches.

Data access

Parameter and log files for principal analyses are provided as Supplemental Files S2 and S3. Data and code are provided as Supplemental File S4.

Acknowledgments

We thank members of the Borenstein and Queitsch laboratories, and Evgeny Sokurenko, Joe Felsenstein, and Willie Swanson for helpful discussions. We thank Ofir Cohen for help with
the gainLoss program. We thank Hyeon Soo Jeong for help with the HGTree database. M.O.P. was supported in part by the National Human Genome Research Institute (National Institutes of Health) Interdisciplinary Training in Genome Sciences Grant 2T2HG35-16. C.Q. is supported by National Institutes of Health New Innovator Award DP2OD008371. E.B. is supported by National Institutes of Health New Innovator Award DP2AT00780201. We thank UW Genome Sciences Information Technology Services for high-performance computing resources.

References


Baltrus DA. 2013. Exploring the costs of horizontal gene transfer.


Received September 25, 2015; accepted in revised form April 12, 2016. Evolutionary assembly of prokaryotic genomes
Evolutionary assembly patterns of prokaryotic genomes

Maximilian O. Press, Christine Queitsch and Elhanan Borenstein

*Genome Res.* 2016 26: 826-833 originally published online April 14, 2016
Access the most recent version at doi:10.1101/gr.200097.115

Supplemental Material

http://genome.cshlp.org/content/suppl/2016/05/11/gr.200097.115.DC1.html

References

This article cites 50 articles, 25 of which can be accessed free at:
http://genome.cshlp.org/content/26/6/826.full.html#ref-list-1

Creative Commons License

This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.