Modular gene interactions drive modular pan-genome evolution in bacteria.

Juan C. Castro[1,2], Sam P. Brown[1,2,*].

[1]School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, 30332

[2]Center for Microbial Dynamics and Infection, Georgia Institute of Technology, Atlanta, Georgia, 30332

Corresponding author: Sam P. Brown, **Address:** School of Biological Sciences, Georgia Institute of Technology, 311 Ferst Drive, Atlanta, Georgia 30332-0230, USA**, Email:** sam.brown@biology.gatech.edu

Author Contributions: J.C. and S.B. designed research; J.C. performed simulations and analyzed data; and J.C. and S.B. wrote the paper.

Competing Interest Statement: The authors declare no competing interest.

Classification: Biological Sciences (Major), Computational Biology (Minor)

Keywords: Genome evolution, Pangenome structure, Operon evolution, Gene interactions, Gene networks

This PDF file includes:

Main Text

Figures 1 to 8

Supplementary figures 1 to 4

Supplementary table 1

Abstract

Depending on the scale of observation, bacterial genomes are both organized and fluid. While individual bacterial genomes show signatures of organization (e.g. operons), pan-genomes

reveal genome fluidity, both in terms of gene content and order (synteny). Here we ask how mutational forces (including recombination and horizontal gene transfer) combine with selection and gene interactions to shape genome organization and variation both within and across strains. We first build an evolutionary simulation model to assess the impact of gene interactions on pan-genome structure. A neutral evolutionary model can produce transient co-segregation of initially linked genes but is vulnerable on longer time-scales to perturbing mutational events. In contrast, incorporation of modular gene fitness interactions can produce sustainable clusters of linked and co-segregating genes, with the network of co-segregation recapitulating the defined simulation 'ground-truth' network of gene interactions. To test our model predictions, we exploit the increasing number of closed genomes in model species to investigate gene interactions in the pan-genomes of *Escherichia coli* and *Pseudomonas aeruginosa*. Using these highly curated pan-genomes, we show that the co-segregation networks for *P. aeruginosa* and *E. coli* are modular, associate with physical linkage, and most closely map known metabolic networks (for *P. aeruginosa*) and regulatory networks (for *E. coli*). The results imply that co-segregation networks can contribute to accessory genome annotation, and more generally that gene interactions are the primary force shaping genome structure and operon evolution.

Significance Statement

Bacterial pan-genomes represent extraordinary genomic diversity yet remain relatively unexplored due to a research focus on lab reference strains. We exploit the growing availability of closed genomes to build pan-genomes where we can track the physical linkage of all genes. Through a combination of evolutionary simulations and data-analysis, we ask how mutation, selection and gene interactions combine to shape genome structural organization (linkage) and variation (co-segregation) across strains. We show that co-segregation networks are modular, associate with physical linkage, and map to metabolic (for *P. aeruginosa*) and regulatory networks (for *E. coli*). The results imply that modular gene interactions are sufficient to guide the evolution of persistent gene clusters and are the primary force shaping genome structural evolution.

Main Text

Introduction

Individual bacterial genomes show hallmarks of organization (1). For example, genes involved in related biochemical processes are often clustered together on the chromosome (2). The organization of clustered genes can be refined to the point where they are transcribed together as operons, with a shared promoter and operator sequences. Regulatory genes are also often found close to the genes they regulate. A classic example is the *lacI* repressor gene, which resides near the *lacZYA* operon in *Escherichia coli* (3).

While organized, genome structure is under constant assault from an array of mutational processes that scramble gene content and order. Genes can be gained or lost. Whole sequences of genes can be flipped (inversions), moved along a chromosome (translocation, recombination) or shuttled to a new individual (horizontal gene transfer). As a result of these perturbing forces, genomic organization is variable across bacterial isolates; genes that are clustered in one species are often reordered in another (2, 4). For example, multiple alignment of different strains of *E. coli* and *Yersinia pestis* reveal large changes in gene ordering (synteny) across the chromosome of each species (5, 6), where fragments of different length undergo rearrangement due to inversion and translocation events. In addition to variation in ordering, gene content is also variable across strains of the same species. As little as 11% of all *Escherichia coli* genes are present in all strains (7, 8) and an even smaller 1% in *Pseudomonas aeruginosa* (9). These ever-present genes are described as the core genome, and the larger set of variably present/absent genes as the accessory genome. Together, the entire set of genes across a species is termed the pan-genome.

The forces that contribute to bacterial genome organization in the face of perturbing mutational pressures have been the focus of extensive debate, leading to a menu of alternate models. The influential selfish operon model (10) views gene clustering as a selfish property of an operon, as clustering of physiologically related units enhances the probability that an entire operon will be horizontally transferred together as a functional unit, and therefore overcome selection for

the loss of multi-gene functions that are under weak or fluctuating selection. A central prediction of the selfish operon model is that operons are enriched for peripheral (versus core) metabolic functions subject to fluctuating selection, which is supported by anecdotal reports of accessory functions carried on plasmids (11). Systematic whole-genome analyses, however, have not generally supported this prediction; highlighting a greater frequency of clustering for essential genes in *E. coli* (12, 13).

Other models of operon evolution highlight the efficiency gains of sharing a single promotor. For example, by reducing the target size for loss-of-function mutations (14), or by offering more complex temporal co-ordination of co-regulated processes (15, 16). Moreover, gene clustering, even without a shared promotor, can offer efficiency gains by spatially co-localizing within the cell gene products that engage in related biochemical tasks (17–19).

In this study we propose that the models above represent special cases of a broader gene-interaction model (e.g. metabolic or regulatory connections between genes). The potential importance of gene interactions has been proposed before in the context of operon evolution (20, 21), following Fisher's (1930) foundational work on epistasis, recombination, and the evolution of linkage. Fisher reasoned that physical linkage (clustering) between genes can occur, at higher rates than expected by chance, due to positive epistatic interactions. The relevance of this logic to bacteria has subsequently been questioned (10) due to the generally low rates of homologous recombination in clonally reproducing microbes (2, 10, 23). Here we return to Fisher's model and apply it considering our broader understanding of the peculiarities of bacterial genomic organization and processes of change beyond recombination.

Our hypothesis is that gene interactions introduce selective effects on genome-modifying processes which lead to the formation of co-segregating clusters of interacting genes. We test this hypothesis via an iterative combination of evolutionary simulation models and pan-genome data resulting in clear predictions regarding the effect of gene-interaction on genomic architecture. Using closed genome data for 100s of *E. coli* and *P. aeruginosa* genomes, we identify clusters of

101 physically linked and co-segregating genes and show that the resulting co-segregation networks
102 map onto underlying gene-regulatory and metabolic gene interaction networks.

103 **Results**

104 *Simulation models (1): Neutral model of genome evolution*

105 Our overarching hypothesis is that gene interactions ultimately drive the organization of
106 genomes into modular clusters of physically linked genes. In the following sections, we develop a
107 model of pan-genome evolution, based on interactions that impact bacterial fitness. First, we
108 establish a baseline neutral model case where there are no interactions (or where gene interactions
109 have no impact on fitness). In this simplified world, genes still experience chromosomal structure,
110 with genes arranged on a circular chromosome. In this neutral world, it is still intuitively reasonable
111 to expect genome structure to persist as a result of initial conditions: genes that are initially close
112 together will tend to stay close together, as they experience a reduced likelihood of gene separation
113 via recombination or other genome modifying functions.

114 To test this neutral (no selection) hypothesis of genome inertia, we conduct *in silico*
115 experimental evolution on 100 interacting lineages of bacteria (see methods for details). In brief,
116 each lineage has an average genome size of 2000 genes, and is subject to genomic perturbations
117 of gene gain, loss, and gene rearrangement (via inversion/translocation). Gene gain for a focal
118 lineage is sourced from the pan-genome (the genomes of all lineages), therefore linking lineages
119 together via horizontal gene transfer. Our simulation tracks time in fixation events (e.g., fixation
120 within 1 lineage of a gene-loss event, see methods for simulation details).

121 In our first simulation we track a single gene pair (*x,y*) that are initially adjacent in 10% of
122 all lineages (absent in other lineages) and follow their resulting organization (linkage within
123 genomes and co-segration across genomes). In figure 1A we track linkage via the average
124 chromosomal distance d_{xy} (the number of intervening genes on the shortest chromosomal path
125 between *x* and *y*, or *d* generally for any arbitrary pair of genes), averaged across all genomes where
126 both genes are found. Initially, d_{xy} is zero, reflecting our initial conditions focused on two adjacent

genes. Figure 1A illustrates that while linked genes can persist in a state of close linkage for 100s of fixation events, given sufficient time (note log scale on *x* axis), genetic perturbations will eventually separate two focal genes. In Figure 1B we see that by 10,000 fixation events, the distribution of d_{xy} approaches a uniform distribution (Figure 1B), consistent with a simple diffusion model of genetic drift.

The distance metric d_{xy} is only applicable for lineages that contain both genes *x* and *y.* Initially, *x* and *y* were always found together (in 10% of lineages; both absent elsewhere) yet processes of gene gain and loss could potentially alter this tight pattern of co-segregation. To assess the extent to which a pair of genes co-segregate, we next estimated the mutual information (I_{xy}, or *I* for any arbitrary pair of genes) between genes *x*,*y* across the pan-genome (mutual information is analogous to correlation without the need to use real-valued variables (24, 25)). In Figure 1C we see that the initial simulation conditions produce an initial mutual information co-segregation value *I* = 0.127. Our results show that over short timeframes (less than 100 fixation events), initially linked genes maintain high mutual information, however as linkage breaks down due to inevitable neutral mutational events, so does co-segregation (Figure 1C; note *I* takes on discrete values due to the discrete nature of presence /absence values used in the pangenome characterization). At equilibrium we observe a skewed distribution of *I* values, with a peak at zero (Figure 1D). Figures 1A-D indicate that under a neutral model, gene pair distance *d* tends to drift and co-segregation peaks at zero.

Focal gene interaction model.

The absence of persistent linkage in our neutral simulation model (Figure 1) is not consistent with examples of highly conserved linkage across bacterial isolates, indicating a role for selection. For example the *peg* (polyethylene glycol-degradative) operon is conserved across multiple lineages of *Sphingomonas* (26), likewise several *E. coli* operons are conserved in different lineages (27). Therefore, to assess the role of selection in our simulation model, we develop a model to include gene interactions that impact fitness. In this model, events of gene gain, gene loss, and gene rearrangement are now potentially selectively non-neutral if they perturb a gene

interaction with defined importance for fitness. As in our previous simulation we track a single pair of genes, but now assume this pair has a positive interaction effect on fitness when found together in a genome (and set all other gene interaction effects to zero). In contrast to Figure 1, we reverse our initial conditions, so that our 2 focal genes are initially not closely linked (Figure 2A) and not co-segregating (Figure 2B).

Figure 2A shows that distance decreases with simulation time, and although stochastic effects can cause instances where distance is high, these are transient events, and the distribution of distances after 200,000 fixation events is stably low, with a mode of zero (Fig 2B). The shortest time to reach a distance of zero (i.e., the focal pair becoming adjacent genes) is ~40,000 fixation events (Fig 2A) which translates to around 400 fixation events per genome in the collection. Figure 2C shows that during the simulation values of I_{xy} steadily increases, because I_{xy} is a global measurement of the linkage of a pair of genes *xy* in the genome collection it is less susceptible to stochastic effects. After ~15,000 fixation events (150 per genome events) the value of *I* stabilizes around 0.27 and remains the same for the course of the simulation.

Genome network interaction model.

Figure 2 illustrates that given selection on one gene pair in isolation, close linkage (low d_{xy}) and high co-segregation I_{xy} between these the genes *x,y* can emerge, despite ongoing perturbing forces of gene loss, gain, translocation, etc. Yet in real-world genomes, movement of two positively interacting genes together might in turn pull apart other positively interacting genes. To assess how evolution proceeds given multiple gene interactions, we again extend our simulation model to incorporate *networks* of gene interactions, spanning both random network models (28) and modular interaction network models (small world (29) and preferential attachment (30) models). Specifically, we hypothesize that modular gene interaction networks will drive repeated clustering of interacting genes at a pangenome level.

On a pairwise scale, repeated clustering of interacting genes implies that clustered (low *d*) gene pairs will tend to co-segregate more often (higher *I*), leading to a predicted negative

correlation between *d* and *I*. Figure 3A-D plots *d* against *I* under the various interaction models, including the 'no interaction' neutral case (Figure 3A). In both the neutral and random interaction case we can reject the hypothesis of a negative association between *d* and *I* (neutral model: $r = -2.3X10^{-3}$, $p = 0.75$; random interaction model, $r = -1.4X10^{-3}$; p = 0.84). In contrast, when we introduce modular networks, we find support for a negative association (small world model: $r = -0.05$; $p = 6.2X10^{-3}$; preferential attachment model: $r = -0.12$; p = 0.01).

While the modular simulations (Figures 3C, D) reveal significant statistical support for a negative association, the effect size is relatively weak and the distribution of simulation results appears to be sparse for higher values of *I*, indicating a potential problem of under-sampling. We note that panels 3C, D summarize the behavior of approximately 10 million gene pairs, so we are not under-sampling. We conjecture that the sparse boundary reflects that the simulated data is a composite of a large 'background' of non-interacting pairs (see neutral simulation case) overlaid with a smaller set of interacting gene pairs that are biased towards high *I* and low *d* values (Figure 2). Consistent with this conjecture, the gene interaction networks were all sparse (around 0.1% of the possible interactions are included in the network). To further test this conjecture, we identify values of *I* that are outside of the central distribution of *I* values ('outlier *I* values') and repeat our statistical tests of association between *d* and outlier values of *I*. We then go on to assess whether outlier values of *I* can be used to recover the underlying architecture of gene interactions in our defined simulation system.

To isolate 'outlier' values of *I* from a background of neutral interactions we use mean absolute deviation (MAD) to establish a threshold above which values of *I* are to be considered outliers (red lines, Figure 3). We find that for our neutral model (Figure 3A) ($r = -3.4X10-4$; p = 0.97) and our random network model (Figure 3B) ($r = -4.1X10^{-3}$; p = 0.67) outliers have no correlation with distance. Whereas in all models of modular gene interaction we find a significant and more substantial negative correlation between outlier values of *I* and *d*. (Small world network, Figure 3C: $r = -0.33$; $p = 2.6X10^{-12}$; Preferential attachment network, Figure 3D: $r = -0.62$; $p = 5.2X10^{-16}$).

Gene co-segregation network can recover the underlying defined 'ground truth' gene interaction network in our simulation model

Our pan-genome evolutionary simulations with modular gene interactions (Figures 3C,D) indicates that evolution will produce networks of co-segregating (high *I*) and physically clustered (low *d*) genes. Following our arguments above, we hypothesize that the networks of co-segregating genes will recapitulate the underlying modular gene interaction networks, which in this simulation world are defined and accessible as a 'ground truth' case.

To test this hypothesis, we first recovered the underlying gene interaction network over which the fitness effects of our simulation are based on (our ground truth). Next, we built a network based on the outlier *I* values of co-segregation (the co-segregation network, see methods). In both networks nodes represent genes, and edges represent interaction effects. To measure similarity between the two networks we used Hamming distance (i.e., number of changes a network must undergo to be identical to the other; vertical blue lines in Figure 4). To place these similarity metrics in context, we asked how often we would see the same degree of network similarity (Hamming distance), given a randomly re-sampled gene interaction network. By re-sampling randomly 1000 times from networks with the same underlying algorithm and parameters as the specific 'ground truth' case (see methods) we are able to assess the extent to which the observed network similarity diverges from chance expectations. Figure 4 shows that modular gene interaction networks such as small world or preferential attachment produce co-segregation networks that recover with high fidelity the ground truth network (particularly for preferential attachment, where the observed Hamming distance is in the 1st percentile of the reference distribution). In contrast, a random gene-interaction network does not produce a discernable signal in the resulting co-segregation network. The results in Figure 4 indicate that outlier co-segregation values can be a reliable predictor of underlying gene interactions, given a modular organization of gene interactions. We further assess the extent to which repeated simulations identify the same outliers. By fixing the ground truth network and repeating the evolutionary simulation 100 times we find that the repeatability of 'ground truth' discovery is maximal, given modular fitness networks (Suppl Table S1).

Gene clusters

Our simulation results to date examine pairwise measures of genome structure (pairwise distance *d* and pairwise co-segregation *I*), which does not directly capture the extent of larger-scale gene clustering. In a final simulation analysis, we define a gene cluster as a set of genes that are *repeatedly* adjacent to each other, across a set of genomes. In Figure 5 we plot distributions of cluster sizes, as a function of different threshold values for 'repeatability' (defined by prevalence – i.e., the fraction of the genomes in which a cluster is observed). In the absence of gene interactions, we find no clusters larger than gene pairs, even with a permissive prevalence cuttoff of 40% (Figure 5A). In contrast, all our simulations with gene interaction effects show cluster sizes following a geometric distribution, with abundant small clusters and rare large clusters (Figure 5B-D). Unsurprisingly, when gene interactions follow a random network model, clusters are small (Figure 5B). In contrast, small world and preferential attachment networks can lead to bigger cluster sizes (Figure 5C, D).

Model assessment using pan-genomes of closed P. aeruginosa and E coli genomes.

Our simulation models produce several predictions that we now test using genome data on the model species *Escherichia coli* and *Pseudomonas aeruginosa*. We examine genetic linkage, cluster formation and their relationship to genetic networks to evaluate our model predictions

To capture information on chromosomal linkage, our analyses require the use of completed genomes only. Our results are based on 329 *E. coli* and 179 *P. aeruginosa* closed genomes from the NCBI database that met our criteria for inclusion (see methods). For each individual genome we predicted the coding sequences (CDS), finding on an average 5654 CDS for *E. coli* and 6159 for *P. aeruginosa*. We then compared the predicted CDS using BLAST reciprocal best match to determine orthologous groups (OGs, i.e., genes with a common evolutionary origin within a species) with Markov clustering. Given our focus on completed genomes only, the estimated pan-genomes (38731 OGs across all genomes of *E. coli* and 16795 for *P. aeruginosa*) are smaller than estimates using all available sequence data (9, 31). Because most of the genes are found in the

accessory genome, a matrix of presence/absence of genes in each genome is highly sparse and bears the characteristic histogram U shape (Figure S1, Baumdicker, Hess, & Pfaffelhuber, 2012; Collins & Higgs, 2012; Lobkovsky, Wolf, & Koonin, 2013). A key aspect of our analysis is to examine the extent of co-segregation among gene pairs, which we can only calculate for variably present accessory genes. We therefore focus in this study on the accessory genome of *E coli* (5035 genes) and *P. aeruginosa* (4973 genes).

Testing cluster formation predictions

Our previous simulation results (Figure 5A) predict that under the neutral model clusters of size 2 are rare and larger clusters are effectively absent. To quantify the proportional abundance of clustered genes in the neutral simulation model, we find that 0.02% of the pan-genome are found in persistent gene clusters of size 2 or more (measured with prevalence cutoff of 0.4, purple line in Figure 5A). In contrast, the selection models allow for larger and geometrically distributed clusters, particularly under modular interactions. Specifically, we found the proportion of clustered genes to be 0.9%, 6% and 8% under the random, small-world and preferential attachment models, respectively (Figure 5B-D).

To assess these predictions against data, we examine the cluster distribution in our pangenome data collections, using the same prevalence thresholds and defining adjacency as a set of genes that are within 10KB of each other. Like our simulated results, cluster size approximates a geometric distribution (Figure 6A). Indicating that large gene clusters although present are rare in nature. Combining large and small clusters together we find that the total fraction of persistently clustered genes (7% for *E. coli* and 12% for *P. aeruginosa*) is in agreement with our modular network simulation models.

Next, we used experimentally validated operons as a reference for the distribution of cluster size in a single genome. We consulted the RegulonDB for *E. coli* (35, 36) which includes experimentally validated operons for the strain K12. The study on the transcriptome of *P. aeruginosa* by Wurtzel *et al*, (37) was used to estimate operons of genes in the strain PA14. In the

latter an operon is defined as a set of genes transcribed together. In both cases the distribution of sizes approximates a geometric distribution (Figure 6B) in agreement with a modular network model (Figure 5C,D).

Co-segregation versus linkage

Figures 3 A-D predicts a negative relationship between *d* and outlier values of *I*, given modular gene interactions (Figures 3C, D). Turning to our genomic data (Figure 7), we see a predominance of low *I* values, regardless of chromosomal distance *d*. A simple correlation of *I* and *d* values reveals a significant yet small negative association ((Figure 7A *E. coli:* Pearson correlation r =-0.02*;* p = 0.003*.* (Figure 7B *P. aeruginosa: r* = -0.012*;* p = 0.001*)*. To analyze the associations with *d* for outlier values of *I,* we proceed with the same data analyses as in our simulation models. After identifying outliers via the median absolute deviation (38) we find a more substantial negative association between outlier *I* and *d* values (Figure S2; *E. coli: r* = -0.34*;* p = $2.2X10^{-6}$*. P. aeruginosa: r* = -0.23*;* p = $1.5X10^{-6}$*)*, consistent with our modular network simulation results. Figure 7 shows a pattern of sparse outlier values of *I* that resembles our modular interaction models (Figure 3C, D), demonstrating a bias towards the highest measures of co-segregation all occurring among genes in close association (see Figure S2 for a higher resolution plot of this region).

Defining the underlying gene interaction mechanism in E. coli and P. aeruginosa.

In light of the simulation benchmarking results in Figure 4, we next turn to our measured outlier *I* values for *E. coli* and *P. aeruginosa* and ask: what is the network structure of observed *I* values, and does this structure recapitulate candidate gene-interaction networks, namely published metabolic networks and regulatory networks for these two model organisms?

In our simulation models we had defined 'ground truth' gene interaction networks. In our empirical analysis, we turn to candidate interaction networks, beginning with established gene regulatory networks for *E. coli* and *P. aeruginosa* (39–42), involving 1258 accessory genes from *E. coli* and 987 for *P. aeruginosa*. We next assess the similarity of this network with our co-segregation network via Hamming distances, and again contrast this observed network distances with

randomized gene interaction network expectations (Figure 8A, B). The Hamming distance to the regulatory network is in the 1st percentile for *E. coli* and 25th percentile for *P. aeruginosa* (Fig 8 A, B).

In addition to evaluating regulatory networks, we performed the same analyses for established metabolic networks for both species (43, 44), which contain 1538 accessory genes of *E. coli* and 1278 genes of *P. aerugiunosa*. In this case we consider an edge between two genes if they participate in the same metabolic reaction or an adjacent reaction (i.e., the products of one reaction are reactants in the adjacent). We observe the Hamming distance for *E. coli* to be around the 10th percentile and less than the 1st percentile for *P. aeruginosa* (Fig 8 C, D).

To further test these conclusions, we assessed the extent to which the entire co-segregation network (not just outlier values of *I*) can serve as a predictor gene interaction network structure, for both metabolic and regulatory gene interaction networks (AUROC analyses, Supplementary Figure S3 A,B). Consistent with Figure 8, we find that the *E. coli* co-segregation network is most predictive of the *E. coli* regulatory gene network (AUROC = 0.86) and the *P. aeruginosa* co-segregation network is most predictive of the *P. aeruginosa* metabolic network (AUROC = 0.96). Altogether, the results in Figures 8 and S3 show that co-segregating genes in the accessory genome are more closely associated with direct regulatory links in *E. coli* (Figure 8A, S3A) and direct metabolic links in *P. aeruginosa* (Figure 8D, S3D).

Discussion

Our pan-genomic results reveal that co-segregating genes (high *I*) tend to be located closer on the genome (low *d*), for two model bacterial systems (Figure 7). While physical linkage provides an intuitive explanation for this result, our neutral evolutionary simulations show that initial linkage provides only a transient support for this association (Figure 1). Introducing selection on gene interactions into our simulation models can produce persistent linkage and co-segregation for a focal gene pair (Figure 2). When gene interactions are distributed modularly across the genome, we see the emergence of persistent clusters of chromosomally linked and co-segregating genes

(Figure 3, 5), producing patterns of co-segregation that can reveal the underlying defined gene interaction network (Figure 4). Consistent with our simulation results, we find that *E. coli* and *P. aeruginosa* pan-genomes contain clusters of linked and co-segregating genes (Figures 6,7). Assessing our gene co-segregation networks for these two species against candidate gene interaction models we find support for pan-genome evolution primarily driven by metabolic (*P. aeruginosa*) and regulatory (*E. coli*) gene interactions (Figures 8, S3).

The identification of gene interactions is a major topic in biology and has resulted in numerous algorithms to define interactions from comparative genome data (45–50). A key differentiating factor in our work is the use of bacterial closed genomes, which allows us to analyze the dependency of genome structure on gene interactions. Our results have broad implications in both applied and basic biology contexts. In the context of basic biology, our results indicate how pan-genome structure is causally connected to gene interaction networks. On an applied scale, this causal connection can be used to aid gene function annotation. High co-segregation values can be taken as evidence for positive gene interactions, providing a tool to infer functional links from genes of unknown function to genes that are already annotated (e.g., established regulatory or metabolic functions). This approach can be viewed as an extension to the common 'guilt by association' framework (51), now leveraging repeated associations on the pan-genome scale. Current functional network models, including our baseline regulatory and metabolic networks examined above, are heavily biased towards the gene content of model strains(36, 42–44, 52). Our approach offers a path to infer and extend network structures beyond the well-studied domain of reference genomes.

We note that our results are potentially dependent on several limiting assumptions, for example we currently only examine the impact of positive versus neutral fitness interactions. Although this assumption is sufficient to generate gene clusters in our simulation, avenues for future work could include examination of the effects of negative and positive interactions on cluster formation, and/or higher-order fitness interactions (i.e., pairwise effects dependent on genetic background)(47, 53, 54). By focusing on orthologous groups of genes, gene-dosing effects of gene

duplication are also omitted from our analysis (55, 56). Our focus on accessory genomes captures the large majority of the pan-genome of *E. coli* and *P. aeruginosa*, but omits the core genome (1.2 % and 4.5 % of *E coli and P.aeruginosa* pangenomes, respectively). It is reasonable to expect that core genes have interactions with the variable genome which we are not considering.

By modeling time via fixation events, we do not require parameterization of absolute rates of fixation. We do need to make assumptions however on the relative rates of distinct fixation events in our simulations. Beginning with an assumption of genome size stability (i.e. no directional change in lineage genome size through time , (54, 57)), we set rates of lineage gene gain and gene loss to be equal. Following previous work ((58–60)), we assume that genome rearrangement fixation events are substantially (three orders of magnitude) lower than lineage rates of gene gain and gene loss. While consistent with prior data and simulations, the fixation of genome rearrangement events is reported to vary widely across species (Ballouz et al. 2010). To examine the impact of varying the relative rates of fixation events, we find that our results are robust to order-of-magnitude variation in the relative rate of genome rearrangement fixation events. Specifically, relative rates of between 10^{-3} and 10^{-5} rearrangement fixations per gain (or loss) fixation result in stable cluster formation (Figure S4).

We have shown that gene interactions can introduce systematic biases to pan-genome evolution. In particular, modular gene interactions are sufficient to guide the evolution of persistent gene clusters, and we argue are the first step in the evolutionary process of operon formation. Although several other models have offered working alternatives to the clustering/operon problem, prior work lacks explicit treatment of the pan-genome. By integrating simulation and data analysis at the pan-genome scale, our work indicates that genome interactions are the central organizing principle to understand the origins and maintenance of pan-genome structure and open a new path for gene annotation across the major pool of bacterial genomic diversity – the accessory genome.

Materials and Methods

Supplementary Methods

Model of gene clustering

A collection G of 100 genomes is simulated through evolutionary time (each genome through time forms an evolutionary lineage, coupled to other lineages by processes of gene transfer). For simulation purposes a genome is an ordered set of genes, where each element represents a gene. The collection represents a polymorphic sampling of strains from the species, and a genome is a unique arrangement of this genetic pool. Each genome has an average size (*S*) of 1000 genes. Each gene is allocated to a set of genomes by sampling a beta distribution, such that the number of times each gene appears in the collection (N_g) is given by:

$$N_g = B(\alpha = 0.5, \beta = 0.5);$$

This distribution is like the one found in *E. coli* and *P. aeruginosa* pangenomes and is common in bacterial pangenomes (1). G has an associated matrix W, which describes the interactions between genes. The $W(i,j)$ element represents the associated cost or benefit to a genome of carrying both gene i and gene j. The simulation considers events of gene gain, gene loss, and genome rearrangement (i.e., genomic inversions and translocations) as independent events the genome can suffer, these events can be influenced by a selection criterion or can be free of selection. The model however does not consider the population dynamics that lead to fixation of genomic changes. It is assumed that (given enough time) a genome lineage will suffer one of the following three categories possible fixation events: gene gain, gene loss or gene inversion / translocation. Each iteration step of the model represents a fixation event for one of these processes in one of the 100 genome lineages. The simulation algorithm can be summarized as follows:

1. Initialize the set (G, W) as described above.
2. Chose an individual genome at random.
3. A fixation event occurs:
    i. Gene gain.

ii. Gene loss.

iii. Gene inversion/translocation.

4. Distance and mutual information are calculated for the genes in the set.

5. Repeat steps 2-4 until the end of the simulation.

6. Record the values and exit

Gene interaction and selection effects

Selection is only considered in step 3 of the above-described algorithm. Assumptions on the impact of selection are described for each event as follows:

For a gene gain event, the available pangenome pool is sampled, the insert size (i.e., number of genes to be inserted) is calculated according to a Poisson distribution. W is used to calculate the insertion site in the receiver genome. For each joint in the genome (i.e., edge between two genes) a weight is calculated such that

$$A_{ij} = 1 - \frac{W_{ij}}{S}$$

Where A_{ij} is the weight for the joint between genes i and j, W_{ij} is the interaction coefficient between the two genes, and S is the genome size. An insertion site is selected by sampling from the distribution of values A. We recover the neutral model (Figure 2) given W_{ij}= 1.

For a gene loss event, fragment size is selected the same way as with gene gain. A fragment to be removed is selected by calculating its weight, by considering the sum of the weights for the genes in the fragment.

$$R_{ij} = \sum_{i=1}^{I} \sum_{j=1}^{J} A_{ij}$$

The fragment to be removed is sampled form the set of all possible weights *R.*

Finally, for a genome rearrangement event. A fragment to be moved is selected same as for gene loss, and the fragment is inserted into a site selected according to gene gain.

Simulation parameterization

For all simulations we used a genome collection of 100 separate genome lineages. 1000 genes (assumed to be orthologous groups) are allocated to the collection such that the probability of each gene (P(*og*)) follows a beta distribution with positive equal shape parameters (α, β), such that $(P(og)\ Beta(\alpha = 0.5, \beta = 0.5)$. Genome size is initially 2000 genes for all genes. Fixation events rates are defined as: gene gain (P_g), gene loss (P_l), and gene rearrangement (P_r).

$$P_r = 0.001; P_g = P_l = \frac{1 - P_r}{2};$$

With $P_g = P_l$, although individual genome changes during the simulation, average genome size is constant.

For our model of interactions, a network is defined for all genes (*G*). We use three network generated models:

-A random network according to the Erdős–Rényi random graph model (2). With edge sampling parameter p =0.0125.

-A small-world network according to the Watts–Strogatz model (3), with an expectation of $1.2X10^6$ edges and an average degree (*K*) of 3.

- A preferential attachment network according Barabási–Albert model (4), with a degree distribution parameter (γ) equal to 3 and a growth rate parameter of 1.2.

This parameter selection ensured we have networks with similar number of nodes and edges such that the simulations are comparable. Because both the Watts–Strogatz and the Barabási–Albert generate sparse networks we use a small value of p in our random network. Specifically small world networks require that the number of edges *I* is smaller than the number of nodes (*n*) such that, $n \gg e \gg ln\ (e)$ (3). While preferential attachment networks have a distribution of degree following a power law, they generate sparse network with , $n \gg e$ (5).

Co-segregation and chromosomal distance

A presence/absence matrix was built for the selected genes (OGs). In this numeric matrix (1 representing presence and 0 representing absence), each gene is treated as a random variable

and its probability distribution is given by its prevalence in the collection. For each gene pair mutual information is calculated according to:

$$I_{xy} = \sum_{y \in X} \sum_{x \in X} p\,(x,y) log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

Where X and Y are two separate genes, $p(x)$, $p(y)$ are the marginal probabilities for gene X and Y respectively (e.g., $p(x)$probability of finding gene X in a single genome), and $p(x,y)$ is the joint probability distribution of gene X and Y respectively.

Chromosomal distance (*d*) is calculated between all pairs of selected genes. Such that:

$$d_{xy} = \frac{\sum_{g=1}^{G} min D_g(x,y)}{G}$$

Where $D_g(x,y)$ is the set of 2 distances in the circular chromosome between genes x and y for the gth genome and G is the total number of genomes in the collection.

Outlier detection and network inference

We used median absolute deviation (MAD) to determine outlier measurements of co-segregation according to (6):

$$MAD(I) = M(|I_i - M(I)|)$$

Where *I* is the set of measurements of co-segregation, and *C* is a consistency constant. A given value is considered an outlier if it has an absolute deviation from the maximum deviation larger than 2.

A network is inferred using outlying values of co-segregation. Where for a gene pair *x* and *y,* an edge is drawn between them if their co-segregation (I_{xy}) is larger than the MAD cutoff. Comparison between networks is performed by calculating the Hamming distance that is; the number of transformations (i.e., edge additions and subtractions necessary for the two networks to be identical, this is expressed as:

$$H_{ij} = \left(E(G_i) \cup E\left(G_j\right) \right) - 2E\left(G_i \cap G_j\right)$$

Where $E(G_i)$ is the number of edges in network i, $E(G_j)$ is the number of edges in network j.

E. coli and P. aeruginosa genome processing

To avoid sampling issues in the gene content analysis only completed genomes were included, this includes sequences labeled as either “Complete genome” (which includes the chromosome and plasmid information) and Chromosome in the database. Additionally, we removed any genome that had gaps in sequences or unspecified nucleotides (N). We found 179 *Pseudomonas aeruginosa* and 329 *Escherichia coli* genomes from the NCBI database that met our criteria for inclusion.

Coding sequences for each genome were predicted using GeneMarkS v.1.11 (7). For each pair of genomes reciprocal protein BLAST v.2.10.1 (8) was performed. Reciprocal best matches are used to determine orthologous groups (OGs) with Markov clustering, using and inflation value of 1.5 (9). We exclude core genes (i.e., OGs present in all the genomes in the collection), and OGs with a single instance in the collection to produce a dataset of accessory genes (the variable genome).

Cluster size calculation

To calculate cluster size, we selected 4 different prevalence thresholds (0.4, 0.5, 0.6, 0.7) and selected all genes with prevalence equal or higher in our genome collection. For all gene pairs we consider a cluster if the distance between them is below our selection criteria (10Kbp for bacterial genomes, 0 gene for simulated genomes) so that adjacent and overlapping genes small intergenic. For each pair, above our prevalence threshold, their distance is calculated to all other genes a cluster is determined according to the same criteria. This process is repeated iteratively until a cluster size of 10.

Acknowledgments

We thank Peng Qiu, Marvin Whiteley, Gabriel Perron, Tim Read, Rohan Mehta, the Brown lab and members of the Center for Microbial Dynamics and Infection for valuable feedback on the manuscript. We thank the NIH (5R21AI156817-02) and the Cystic Fibrosis Foundation (BROWN21P0) for funding this project.

References

1. D. J. W. Brocken, M. Tark-Dame, R. T. Dame, The organization of bacterial genomes: Towards understanding the interplay between structure and function. *Curr. Opin. Syst. Biol.* **8**, 137–143 (2018).

2. S. Ballouz, A. R. Francis, R. Lan, M. M. Tanaka, Conditions for the Evolution of Gene Clusters in Bacterial Genomes. *PLoS Comput. Biol.* **6**, e1000672 (2010).

3. P. H. von Hippel, A. Revzin, C. A. Gross, A. C. Wang, Non-specific DNA binding of genome regulating proteins as a biological control mechanism: I. The lac operon: equilibrium aspects. *Proc. Natl. Acad. Sci. U. S. A.* **71**, 4808–4812 (1974).

4. W. C. Lathe, B. Snel, P. Bork, Gene context conservation of a higher order than operons. *Trends Biochem. Sci.* **25**, 474–479 (2000).

5. M. Eppinger, *et al.*, Genome sequence of the deep-rooted Yersinia pestis strain angola reveals new insights into the evolution and pangenome of the plague bacterium. *J. Bacteriol.* **192**, 1685–1699 (2010).

6. M. Eppinger, M. K. Mammel, J. E. Leclerc, J. Ravel, T. A. Cebula, Genomic anatomy of Escherichia coli O157:H7 outbreaks. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 20142–20147 (2011).

7. N. T. Perna, *et al.*, Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature* **409**, 529–533 (2001).

8. M. Touchon, *et al.*, Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. *PLoS Genet.* **5**, e1000344 (2009).

9. L. Freschi, *et al.*, The Pseudomonas aeruginosa Pan-Genome Provides New Insights on Its Population Structure, Horizontal Gene Transfer, and Pathogenicity. *Genome Biol. Evol.* **11**, 109–120 (2019).

10. J. G. Lawrence, J. R. Roth, Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters. *Genetics* **143**, 1843–1860 (1996).

11. D. J. Rankin, E. P. C. Rocha, S. P. Brown, What traits are carried on mobile genetic elements, and why. *Heredity (Edinb).* **106**, 1–10 (2011).

12. C. Pál, L. D. Hurst, Evidence against the selfish operon theory. *Trends Genet.* **20**, 232–234 (2004).

13. C. Pál, *et al.*, Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**, 667–670 (2006).

14. G. Fang, E. P. Rocha, A. Danchin, Persistence drives gene clustering in bacterial genomes. *BMC Genomics* **9**, 4 (2008).

15. M. N. Price, K. H. Huang, A. P. Arkin, E. J. Alm, Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res.* **15**, 809–819 (2005).

16. M. N. Price, A. P. Arkin, E. J. Alm, The life-cycle of operons. *PLoS Genet.* **2**, 0859–0873 (2006).

17. R. E. Svetic, C. R. MacCluer, C. O. Buckley, K. L. Smythe, J. H. Jackson, A metabolic force for gene clustering. *Bull. Math. Biol.* **66**, 559–581 (2004).

18. G. Kolesov, Z. Wunderlich, O. N. Laikova, M. S. Gelfand, L. A. Mirny, How gene order is influenced by the biophysics of transcription regulation. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 13948–13953 (2007).

19. F. J. Martin, J. O. McInerney, Recurring cluster and operon assembly for Phenylacetate degradation genes. *BMC Evol. Biol.* **9**, 1–11 (2009).

20. W. F. Bodmer, P. A. Parsons, Linkage and Recombination in Evolution. *Adv. Genet.* **11**, 1–100 (1963).

21. F. W. Stahl, N. E. Murray, The evolution of gene clusters and genetic circularity in microorganisms. *Genetics* **53**, 569–576 (1966).

22. R. A. Fisher, *The genetical theory of natural selection.* (Clarendon Press, 1930)

23. M. Vos, X. Didelot, A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* **3**, 199–208 (2009).

24. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (2005).

25. J. Pensar, *et al.*, Genome-wide epistasis and co-selection study using mutual information. *Nucleic Acids Res.* **47**, e112–e112 (2019).

26. A. Tani, *et al.*, Structure and conservation of a polyethylene glycol-degradative operon in sphingomonads. *Microbiology* **153**, 338–346 (2007).

27. G. M. Hagelsieb, J. Collado-Vides, Operon Conservation from the Point of View of Escherichia coli, and Inference of Functional Interdependence of Gene Products from Genome Context. *In Silico Biol.* **2**, 87–95 (2002).

28. P. Erdos, A. Rényi, On the evolution of random graphs. *Struct. Dyn. Networks* **9781400841356**, 38–82 (2011).

29. D. J. Watts, S. H. Strogatz, Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).

30. R. Albert, A.-L. Barabási, Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).

31. A. G. Decano, T. Downing, An Escherichia coli ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. *Sci. Rep.* **9**, 17394 (2019).

32. F. Baumdicker, W. R. Hess, P. Pfaffelhuber, The Infinitely Many Genes Model for the Distributed Genome of Bacteria. *Genome Biol. Evol.* **4**, 443–456 (2012).

33. R. E. Collins, P. G. Higgs, Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Mol. Biol. Evol.* **29**, 3413–3425 (2012).

34. A. E. Lobkovsky, Y. I. Wolf, E. V. Koonin, Gene frequency distributions reject a neutral model of genome evolution. *Genome Biol. Evol.* **5**, 233–242 (2013).

35. H. Salgado, *et al.*, RegulonDB (version 2.0): a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res.* **27**, 59–60 (1999).

36. A. M. Huerta, H. Salgado, D. Thieffry, J. Collado-Vides, RegulonDB: A database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res.* **26**, 55–59 (1998).

37. O. Wurtzel, *et al.*, The Single-Nucleotide Resolution Transcriptome of Pseudomonas aeruginosa Grown in Body Temperature. *PLOS Pathog.* **8**, e1002945 (2012).

38. C. Leys, C. Ley, O. Klein, P. Bernard, L. Licata, Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* **49**, 764–766 (2013).

39. M. M. Babu, S. A. Teichmann, Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res.* **31**, 1234–1244 (2003).

40. N. M. Luscombe, *et al.*, Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**, 308–12 (2004).

41. E. Galán-Vásquez, B. Luna, A. Martínez-Antonio, The Regulatory Network of Pseudomonas aeruginosa. *Microb. Inform. Exp.* **1**, 1–11 (2011).

42. E. Galán-Vásquez, B. C. Luna-Olivera, M. Ramírez-Ibáñez, A. Martínez-Antonio, RegulomePA: a database of transcriptional regulatory interactions in *Pseudomonas aeruginosa* PAO1. *Database* **2020**, 106 (2020).

43. J. D. Orth, *et al.*, A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011. *Mol. Syst. Biol.* **7**, 1–9 (2011).

44. J. A. Bartell, *et al.*, Reconstruction of the metabolic network of Pseudomonas aeruginosa to interrogate virulence factor synthesis. *Nat. Commun.* **8** (2017).

45. B. J. Arnold, *et al.*, Weak epistasis may drive adaptation in recombining bacteria. *Genetics* **208**, 1247–1260 (2018).

46. O. Brynildsrud, J. Bohlin, L. Scheffer, V. Eldholm, Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* **17**, 1–9 (2016).

47. R. S. Mehta, R. A. Petit Iii, T. D. Read, D. B. Weissman, Detecting patterns of accessory genome coevolution in bacterial species using data from thousands of bacterial genomes. *bioRxiv*, 2022.03.14.484367 (2022).

48. O. Cohen, H. Ashkenazy, D. Burstein, T. Pupko, Uncovering the co-evolutionary network among prokaryotic genes. *Bioinformatics* **28**, i389–i394 (2012).

49. F. Lassalle, P. Veber, E. Jauneikaite, X. Didelot, Automated reconstruction of all gene histories in large bacterial pangenome datasets and search for co-evolved gene modules with Pantagruel. *bioRxiv*, 586495 (2019).

50. C. Liu, B. Wright, E. Allen-Vercoe, H. Gu, R. Beiko, Phylogenetic Clustering of Genes Reveals Shared Evolutionary Trajectories and Putative Gene Functions. *Genome Biol. Evol.* **10**, 2255–2265 (2018).

51. S. Oliver, Guilt-by-association goes global. *Nat. 2000 4036770* **403**, 601–602 (2000).

52. S. H. Strogatz, Exploring complex networks. *Nature* **410**, 268–76 (2001).

53. N. H. Barton, Linkage and the limits to natural selection. *Genetics* **140**, 821–841 (1995).

54. B. J. Arnold, *et al.*, Weak Epistasis May Drive Adaptation in Recombining Bacteria. *Genetics* **208**, 1247–1260 (2018).

55. K. Kovács, L. D. Hurst, B. Papp, Stochasticity in Protein Levels Drives Colinearity of Gene Order in Metabolic Operons of Escherichia coli. *PLOS Biol.* **7**, e1000115 (2009).

56. A. Zaslaver, A. Mayo, M. Ronen, U. Alon, Optimal gene partition into operons correlates with gene functional order. *Phys. Biol.* **3**, 183–189 (2006).

57. I. Sela, Y. I. Wolf, E. V. Koonin, Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci.* **113**, 11399–11407 (2016).

58. E. Belda, A. Maya, F. J. Silva, Genome rearrangement distances and gene order phylogeny in gamma-Proteobacteria. *Mol. Biol. Evol.* **22**, 1456–1467 (2005).

59. M. Suyama, P. Bork, Evolution of prokaryotic gene order: Genome rearrangements in closely related species. *Trends Genet.* **17**, 10–13 (2001).

60. E. P. C. Rocha, Inference and Analysis of the Relative Stability of Bacterial Chromosomes. *Mol. Biol. Evol.* **23**, 513–522 (2006).

61. P. Lapierre, J. P. Gogarten, Estimating the size of the bacterial pan-genome. *Trends Genet.* **25**, 107–110 (2009).

62. C. I. Del Genio, T. Gross, K. E. Bassler, All scale-free networks are sparse (2011).

63. J. Besemer, A. Lomsadze, M. Borodovsky, GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29**, 2607–2618 (2001).

64. T. R. Gibbons, S. M. Mount, E. D. Cooper, C. F. Delwiche, Evaluation of BLAST-based edge-weighting metrics used for homology inference with the Markov Clustering algorithm (2015).

65. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–10 (1990).

Figures and Tables

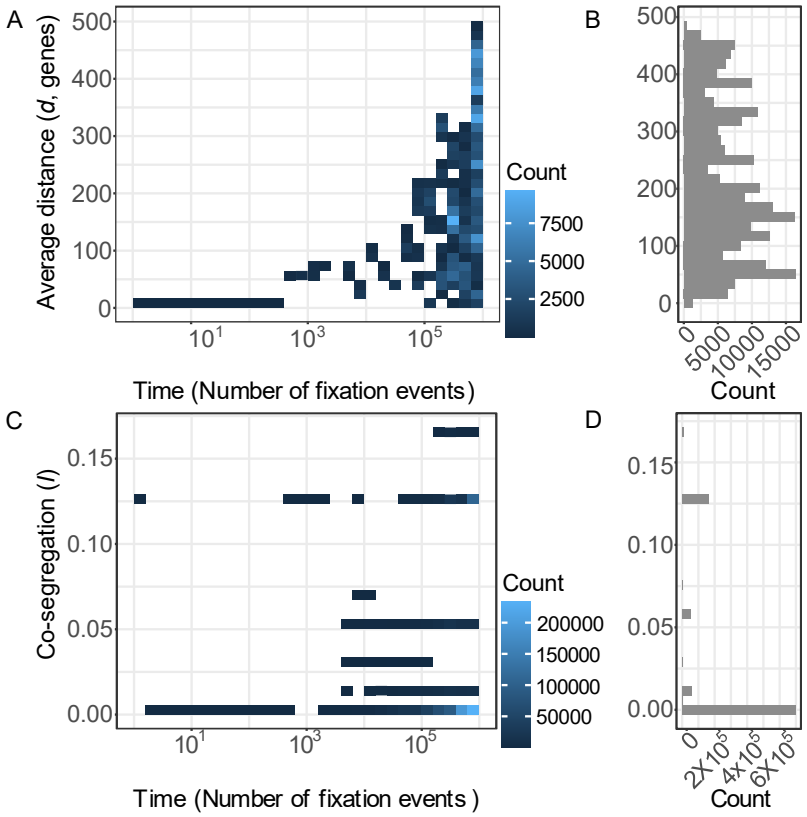
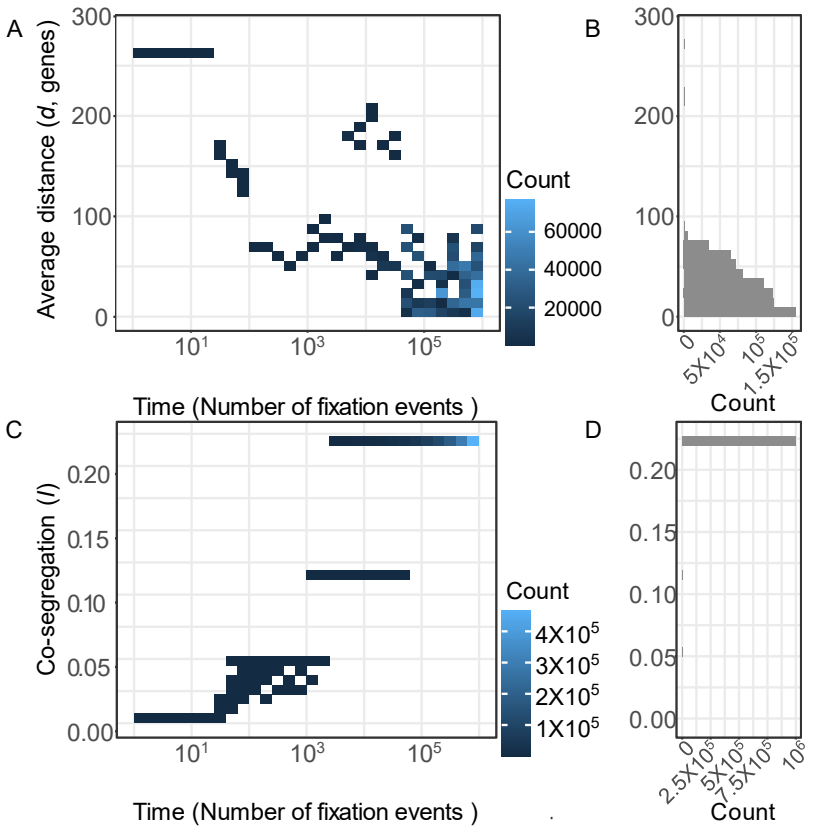


Figure 1. Neutral genomic evolution does not lead to stable co-segregation or clustering in an evolutionary simulation model. An evolutionary simulation of 100 lineages (average 2000 genes per lineage) undergoing neutral events of gene gain and loss and rearrangement. **A.** Average distance d_{xy} between two initially adjacent genes x and y ($d_{xy}(0) = 0$) through simulation time (in fixation events, note log time-scale). **B.** After ~5×10^4 fixation events, distance d is uniformly distributed **C.** Co-segregation between a pair of genes in the absence of selection goes down ($I_{xy}(0) = 0.12$) and fluctuates near a mode of zero. **D.** After 10^3 fixation events, modal co-segregation (I) = 0.



665

666 **Figure 2. Positive gene interaction leads to close linkage and high co-segregation.** An
667 evolutionary simulation of 100 lineages (average 2000 genes per lineage) with a positive defined
668 interaction between two initially non-adjacent genes ($d_{xy}(0) = 270$). **A.** Average distance between
669 two interacting genes d_{xy} decreases through the course of the simulation and reaches 0 after
670 ~30000 fixation events (note log scale on x axis). **B.** After 10^6 fixation events, the modal genetic
671 distance is zero. **C.** Co-segregation I_{xy} between two interacting genes increases during the
672 simulation (from $I_{xy}(0) = 0$). **D.** After 10^6 fixation events, modal co-segregation $I_{xy} = 0.27$.

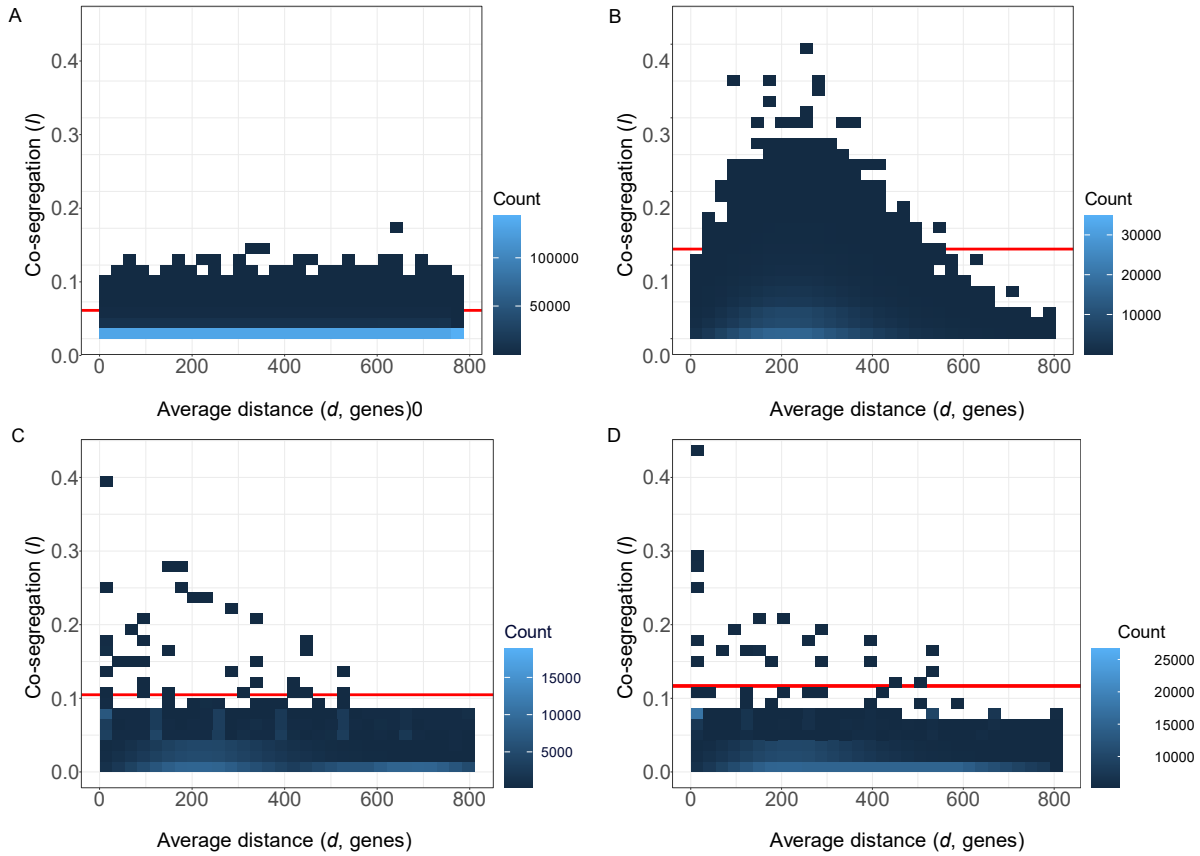


Figure 3. Modular gene interaction networks generate negative associations between chromosomal distance *d* and co-segregation *I*. Values for co-segregation and distance are calculated for all pairs of genes in a collection of 100 genomes with an average genome size of 2000 genes. Threshold of outlier values of *I* is denoted in red (calculated with MAD). **A.** Neutral model. Values of co-segregation for all pairs are lower than 0.15 and distances are uniformly distributed (r = -2.3X10^{-3}, p = 0.75; outlier MAD *I* r = -3.4X10^{-4}; p = 0.97). **B.** Interaction effects distributed according to a random network produces no significant linear correlation (r = -1.4X10^{-3}; p = 0.84; outlier MAD *I* r = -4.1X10^{-3}; p = 0.67). **C.** A small world network produces a negative correlation (r = -0.05; p = 6.2X10^{-3}; outlier MAD *I* r = -0.33; p = 2.6X10^{-12}). **D.** A preferential attachment model produces a negative correlation (r = -0.12; p = 0.01; outlier MAD *I* r = -0.62; p = 5.2X10^{-16}).

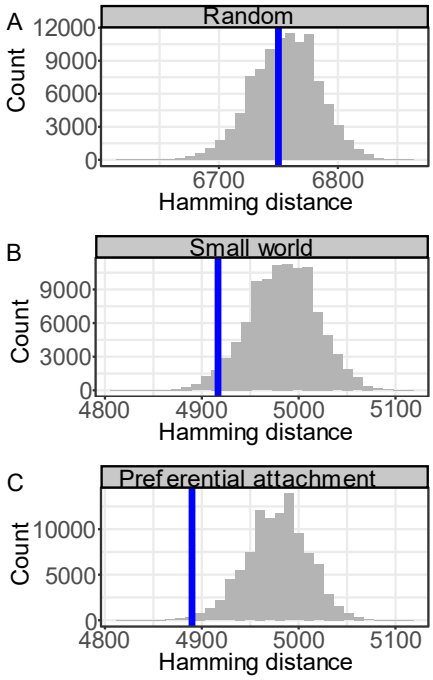


Figure 4. Co-segregation networks identify 'ground truth' gene interaction effects in simulated modular networks. Hamming distances for three different gene interaction network algorithms (ground truths) compared to the simulated values of the co-segregation network (network of outlier values of *I*). Blue lines represent Hamming distances between co-segregation networks and ground truth gene interaction networks. Histograms represent distributions of Hamming distances for randomly sampled gene interaction networks (same number of nodes and edges, see methods). A. Random gene interaction networks shows distance close to the 50th percentile compared to the background distribution. B. Small world gene interaction network has a distance in the lower 10th percentile of the background distribution. C. Preferential attachment gene interaction networks show a distance within the 1st percentile compared to the background.

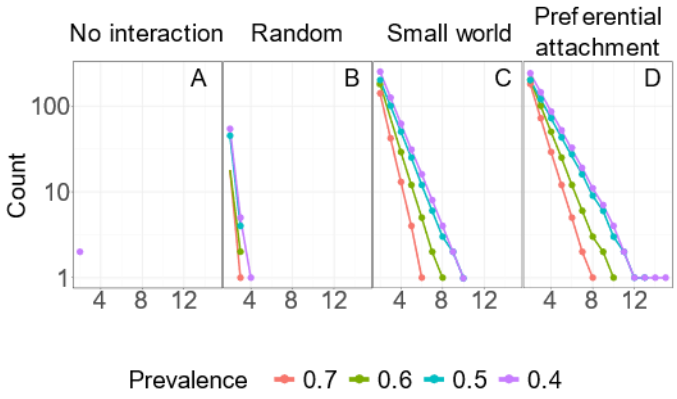
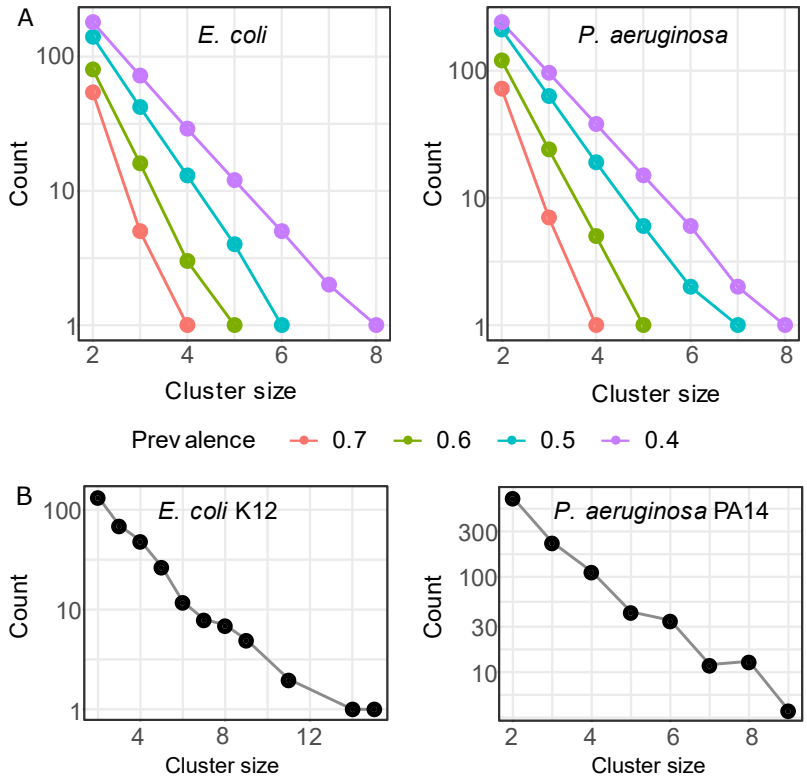


Figure 5. Modular gene interaction networks generate persistent large clusters in bacterial genomes. After evolutionary simulations reached 1 million fixation events, the number of clusters of each size is calculated for four models of gene interaction (**A**, neutral; **B**, random network; **C**, small world network; **D**, preferential attachment network). Clusters are defined as sets of genes that are repeatably observed across the simulated pan genome. Colored lines represent different prevalence thresholds for cluster definition (present in 40% to 70% of all genomes). Preferential attachment and small-world networks yield more clusters and larger cluster sizes.



704

705 **Figure 6. Persistent gene clusters are distributed geometrically across the pan-genomes *E.***
706 ***coli* and *P. aeruginosa*. A.** Clusters of genes in the *E. coli* and *P. aeruginosa* pangenome, lines
707 represent different prevalence for clusters in the genome collection. In both cases data follows a
708 geometrical distribution like the simulation results. **B.** Defined operon cluster sizes from model
709 strain genomes are distributed according to a geometric distribution for both *E. coli* and *P.*
710 *aeruginosa*.

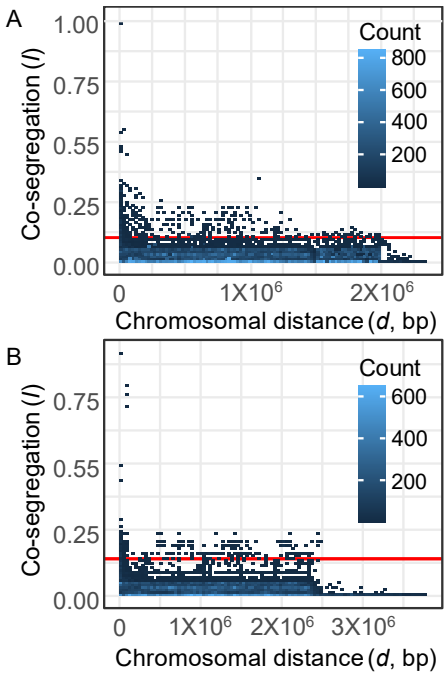


Figure 7. Outlier values of co-segregation *(I*) are negatively correlated with chromosomal distance (*d*) across *E. coli* and *P. aeruginosa* pan-genomes. Each data point refers to an individual gene pair from among the 5035 accessory genes of *E. coli* (**A**) or the 4973 accessory genes of *P. aeruginosa* (**B**). For each gene pair (*x*,*y)*, we calculate average chromosomal distance d_{xy} (across all instances of co-localization within a genome) and co-segregation I_{xy}. Across all gene pairs, we find a small negative correlation between *I* and *d* (*E. coli: r* =-0.02*;* p = 0.003*. P. aeruginosa: r* = -0.012*;* p =0.001*)*. Across gene pairs with outlier values of co-segregation (values above the red line: *E. coli: I* > 0.093. *P. aeruginosa*: *I* > 0.012; see Figure S2) we find a stronger negative correlation (*E. coli: r* = -0.34*;* p = $2.2X10^{-6}$*. P. aeruginosa: r* = -0.23; p = $1.5X10^{-6}$).

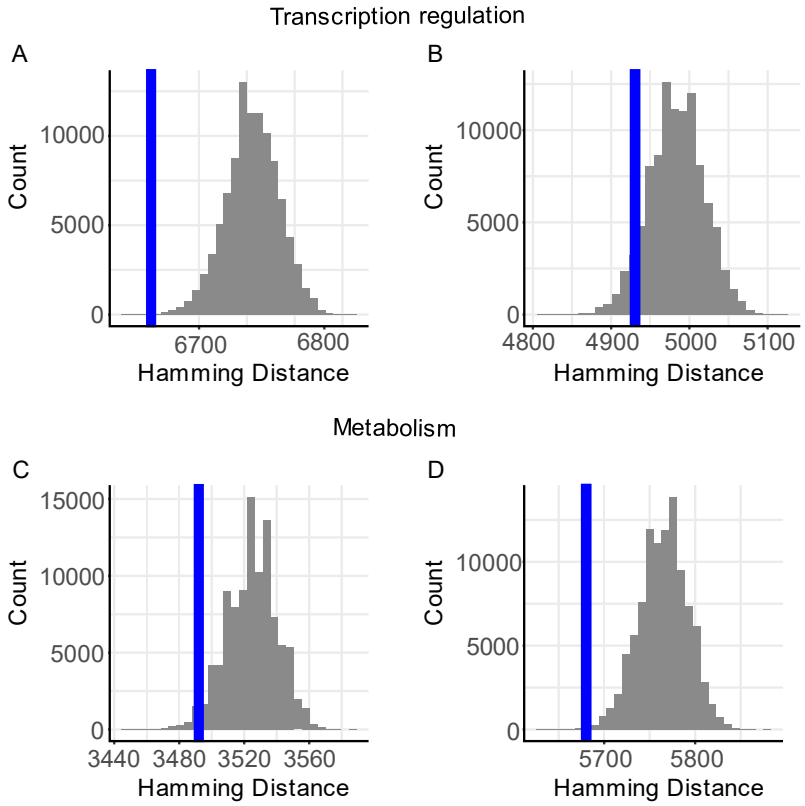
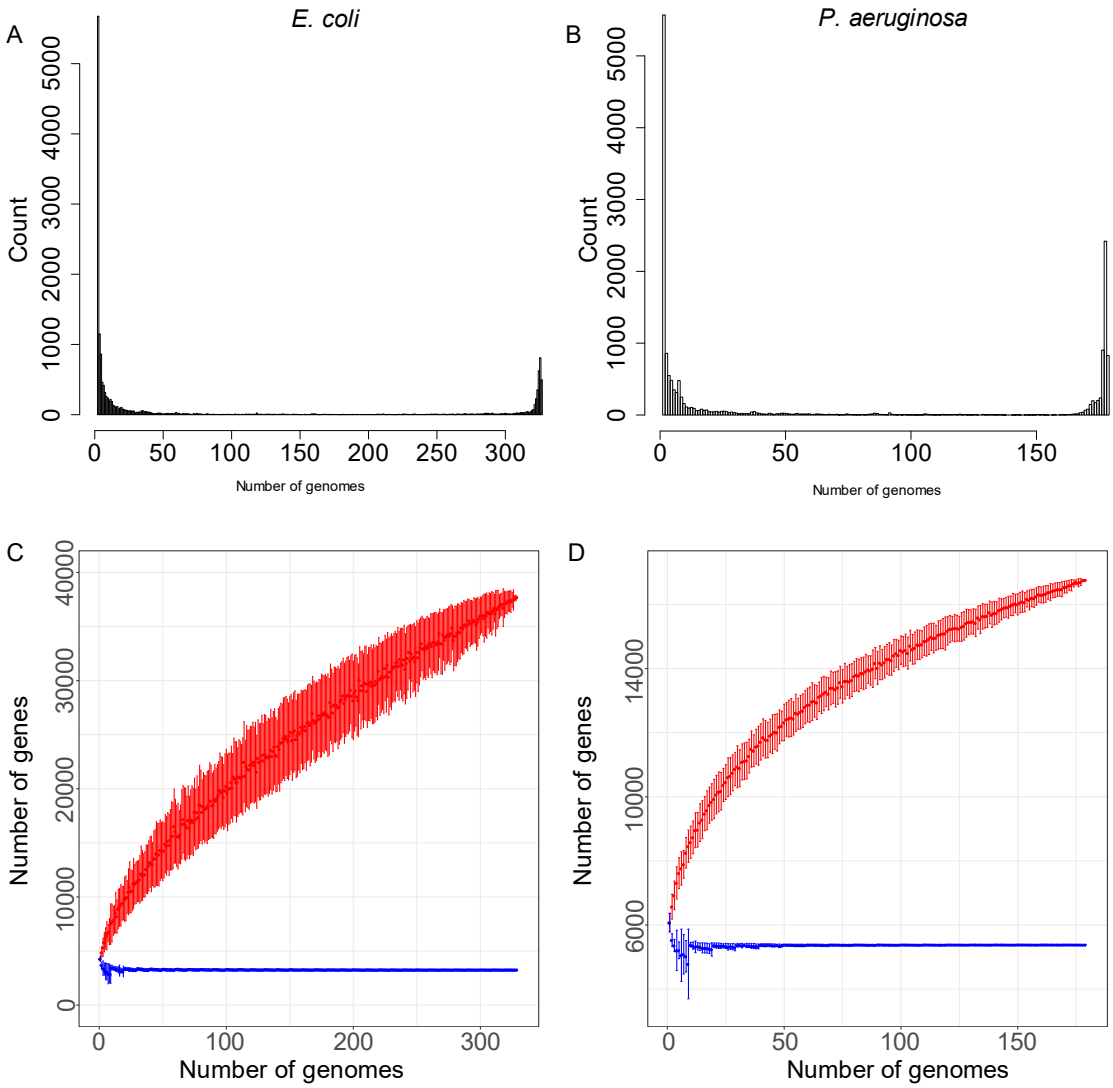


Figure 8. Co-segregation networks capture properties of both regulatory and metabolic networks in bacteria. Hamming distance between a co-segregation network obtained from the pan-genomes of 329 genomes from *E. coli* (panels A, C) and 179 genomes of *P. aeruginosa* (B, D) and their respective transcription regulatory network (A, B) and metabolic networks (C, D). Histogram data represents the distribution of comparisons of 100000 bootstraps of the networks. The vertical blue line represents the measured value. A, B. Assessment of regulatory network for *E. coli* and *P. aeruginosa* respectively. C, D. Assessment of metabolic networks for *E. coli* and *P. aeruginosa* respectively.

730 **Supplementary Information**

731



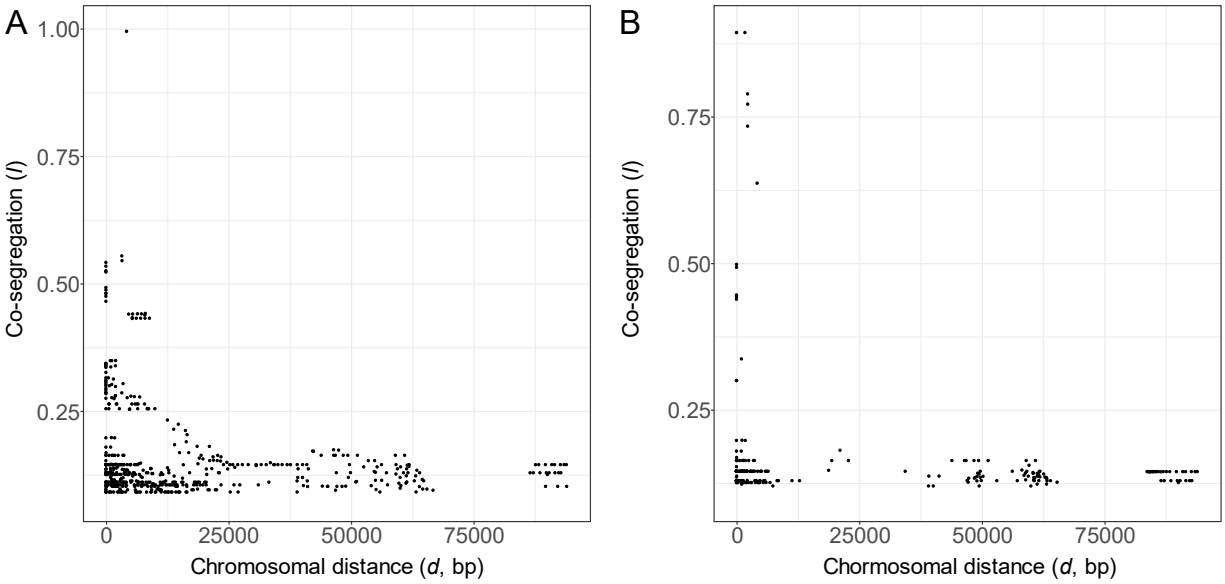
732

733 **Supplementary Figure S1. Pangenome sampling of completed genomes of *E. coli* and *P.***
734 ***aeruginosa*.** Identified genes are assigned into orthologous groups via Markov clustering.
735 Distribution of genes across genome collection of 329 *E. coli* **(A)** and 179 P. aeruginosa genomes
736 **(B)**. A significant fraction of genes is present in only one genome, with the majority of genes being
737 present in the variable genome (98% *E. coli* and 96% *P. aeruginosa*). Sampling all genes in the
738 collection is illustrated with red lines, the blue lines denote the core genome. The increasing trend

739 indicates that the pangenome is under-sampled. A total of 38731 genes are identified across all

740 genomes of *E. coli* **(C)** and 16795 for *P. aeruginosa* **(D)**.

741



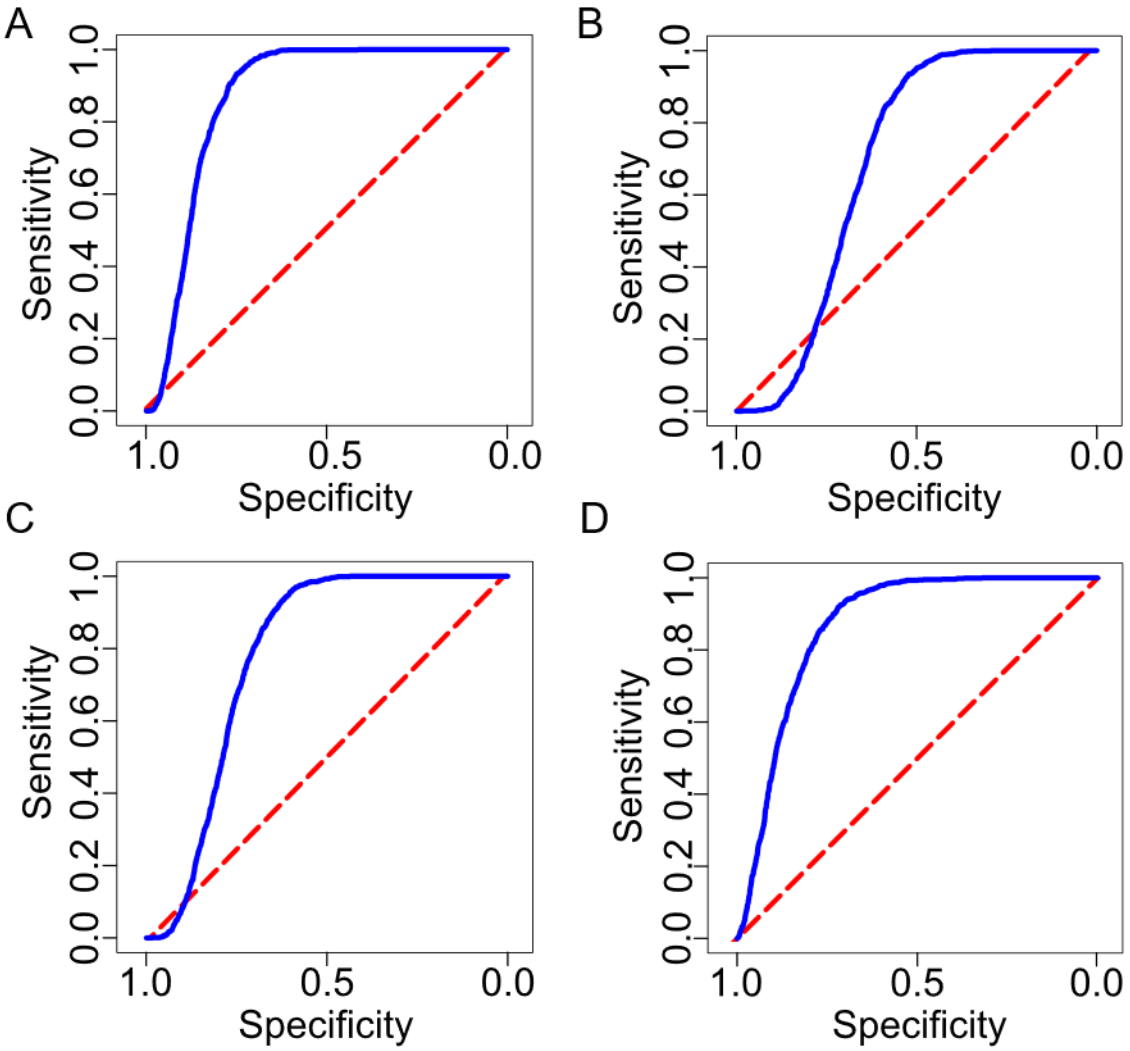
742

743 **Supplementary figure S2. Outlier values of co-segregation (*I*) are negatively correlated with**

744 **distance.** Identified outliers of *I* using median absolute deviation are plotted against chromosomal

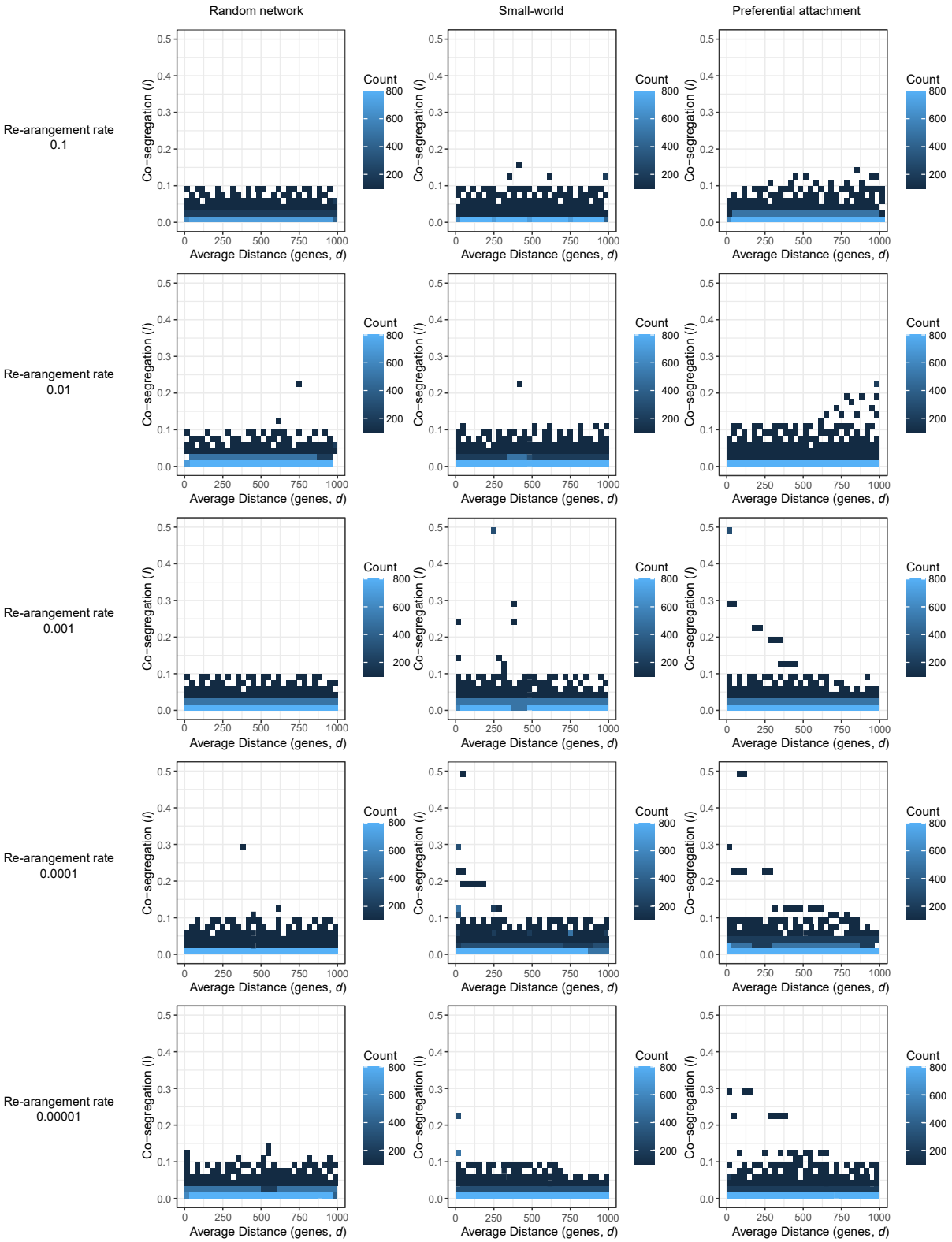
745 distance for *E. coli* **(A)** and *P. aeruginosa* **(B)**. In both cases a negative relationship between the

746 two is significant (*E. coli: r* = -0.34*;* p = $2.2X10^{-6}$*. P. aeruginosa: r* = -0.23; p = $1.5X10^{-6}$).



Supplementary Figure S3. Co-segregation is a competent predictor of regulatory and metabolic interactions in bacteria. Receiving operating characteristic (ROC) is used as a diagnostic of predictive ability for co-segregation networks obtained from the pan-genomes of 329 genomes from *E. coli* (panels **A, C**) and 179 genomes of *P. aeruginosa* (**B, D**) predicting their respective transcription regulatory network (**A, B**) or metabolic networks (**C, D**). Blue lines represent the predictive ability of co-segregation across a moving threshold in *I*. Red lines represent the expectation of a random guess. **A, B.** Assessment of regulatory network for *E. coli* (Area under ROC = 0.86) and *P. aeruginosa* (Area under ROC = 0.68) respectively. **C, D.** Assessment of

metabolic networks for *E. coli* (Area under ROC = 0.78) and *P. aeruginosa* (Area under ROC = 0.86) respectively.



758

759 **Supplementary figure S4. The relationship between *d* and *l* is affected by network**
760 **organization as well as re-arrangement rate due to translocation/inversion events.** A negative
761 relationship can be observed for both small-world and preferential attachment networks, however
762 this relationship is conditional on low re-arrangement rates.

	Average network size	Average True Positives	Average False positives	Average AUROC
Random network	14994	7521	7472	0.52
Small-world	15239	11416	3825	0.73
Preferential attachment	14974	11973	3001	0.79

Supplementary table S1. Outlier values of co-segregation (*I*) can recover the underlying interaction network. Simulations are repeated 100 times each, in each case the ground truth network and simulation parameters are fixed. Identified outliers of *I* using median absolute deviation (MAD) are used to estimate a network of interaction for the genes in the simulation. This estimate can recover around 50% of the edges in a random network model, 74.9% for a small-world model and 80.0% for a preferential attachment model.