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.

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

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 define gene co-segregation networks in the pan-genomes of *Escherichia coli* and *Pseudomonas aeruginosa*. Using these highly curated pan-genomes, we identify modular clusters of physically linked and co-segregating genes and show that the resulting co-segregation networks map onto underlying gene-regulatory and metabolic gene interaction networks. 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.

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

59 the accessory genome. Together, the entire set of genes across a species is termed the
60 pan-genome.

61 The forces that contribute to bacterial genome organization in the face of
62 perturbing mutational pressures have been the focus of extensive debate, leading to a
63 menu of alternate models. The influential selfish operon model (10) views gene clustering
64 as a selfish property of an operon, as clustering of physiologically related units enhances
65 the probability that an entire operon will be horizontally transferred together as a functional
66 unit, and therefore overcome selection for the loss of multi-gene functions that are under
67 weak or fluctuating selection. A central prediction of the selfish operon model is that
68 operons are enriched for peripheral (versus core) metabolic functions subject to fluctuating
69 selection, which is supported by anecdotal reports of accessory functions carried on
70 plasmids (11). Systematic whole-genome analyses, however, have not generally
71 supported this prediction, highlighting a greater frequency of clustering for essential genes
72 in *E. coli* (12,13).

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

79 In this study we propose that the models above represent special cases of a
80 broader gene-interaction model (e.g., metabolic or regulatory connections between
81 genes). The potential importance of gene interactions has been proposed before in the

82 context of operon evolution (20,21), following Fisher's (22) foundational work on epistasis,
83 recombination, and the evolution of linkage. Fisher reasoned that physical linkage
84 (clustering) between genes can occur, at higher rates than expected by chance, due to
85 positive epistatic interactions. The relevance of this logic to bacteria has subsequently
86 been questioned (10) due to the generally low rates of homologous recombination in
87 clonally reproducing microbes (2,10,23). Here we return to Fisher's model and apply it
88 considering our broader understanding of the peculiarities of bacterial genomic
89 organization and processes of change beyond recombination.

90 Our hypothesis is that gene interactions introduce selective effects on genome-
91 modifying processes which lead to the formation of co-segregating, physically linked
92 clusters of interacting genes. We test this hypothesis via an iterative combination of
93 evolutionary simulation models and pan-genome data. Using closed genome data for
94 100s of *E. coli* and *P. aeruginosa* genomes, we identify clusters of physically linked and
95 co-segregating genes and show that the resulting co-segregation networks map onto
96 underlying gene-regulatory and metabolic gene interaction networks. The results imply
97 that modular gene interactions are sufficient to guide the evolution of persistent gene
98 clusters and are the primary force shaping genome structural evolution.

99

100 **Results**

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

102 Our overarching hypothesis is that gene interactions ultimately drive the
103 organization of genomes into modular clusters of physically linked genes. In the following

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

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

120 In our first simulation we track a single gene pair (x, y) that are initially adjacent in
121 10% of all lineages (absent in other lineages) and follow their resulting organization
122 (physical linkage within genomes and co-segregation across genomes). In figure 1A we
123 track linkage via the average chromosomal distance d (the number of intervening genes
124 on the shortest chromosomal path between x and y, averaged across all genomes where
125 both genes are found). Initially, d is zero, reflecting our initial conditions focused on two
126 adjacent genes. Figure 1A illustrates that while linked genes can persist in a state of close
127 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* approaches a uniform distribution (Figure 1B), consistent with a simple diffusion model of genetic drift.

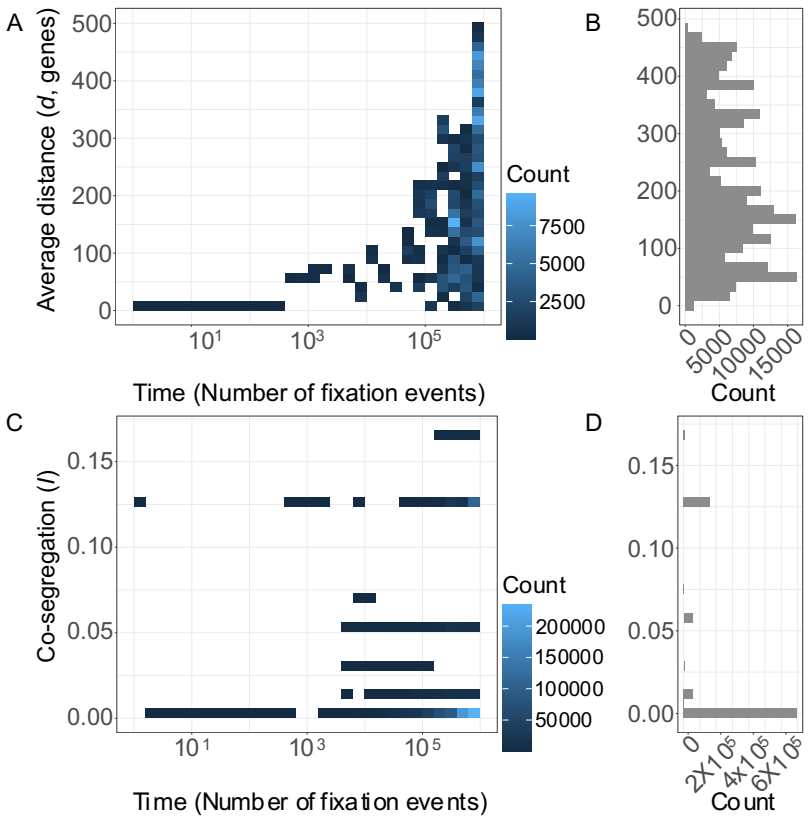


Figure 1. Neutral genomic evolution does not lead to stable co-segregation or clustering in an evolutionary simulation model. An evolutionary simulation of 100 lineages (average 2000 genes per lineage) undergoing neutral events of gene gain and loss and rearrangement. **A.** Average distance *d* between two initially adjacent genes *x* and *y* ($d(0) = 0$) through simulation time (in fixation events, note log time-scale). **B.** After ~5×10^4 fixation events, the distribution of distance *d* approaches a uniform distribution **C.** Co-segregation between a pair of genes in the absence of selection goes down ($I(0) = 0.127$)

139 and fluctuates near a mode of zero. **D.** After 10^3 fixation events, modal co-segregation (*I*)
140 = 0.

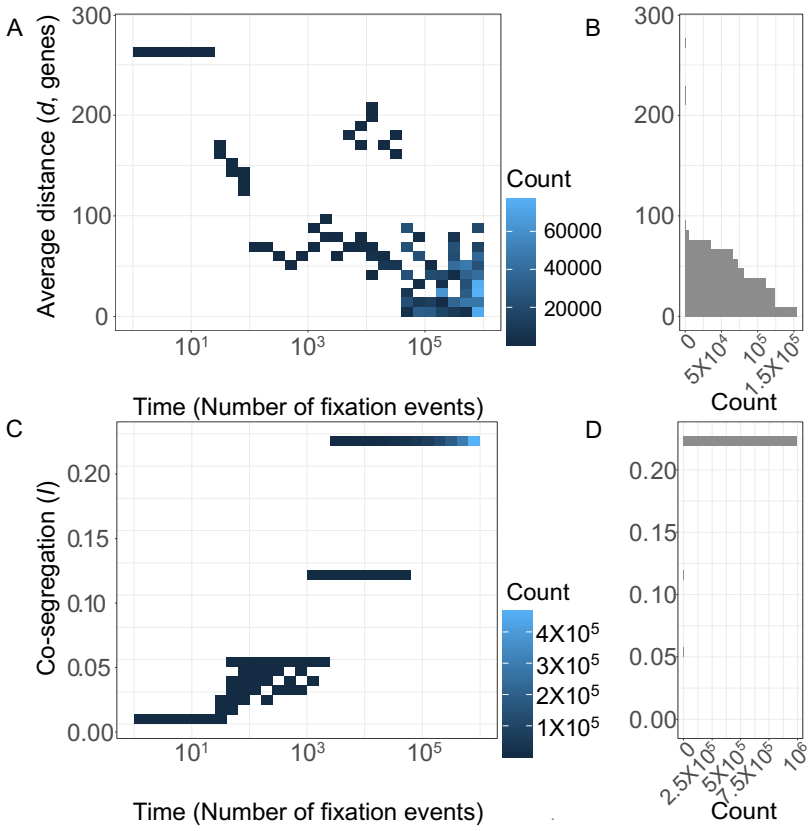
141 The distance metric *d* is only applicable for lineages that contain both genes *x* and
142 *y.* Initially, *x* and *y* were always found together (in 10% of lineages; both absent elsewhere)
143 yet processes of gene gain and loss could potentially alter this tight pattern of co-
144 segregation. To assess the extent to which a pair of genes co-segregate, we next
145 estimated the mutual information (*I*) between genes *x*, *y* across the pan-genome (mutual
146 information is analogous to correlation without the need to use real-valued variables
147 (24,25)). In Figure 1C we see that the initial simulation conditions produce an initial mutual
148 information co-segregation value *I* = 0.127. Our results show that over short timeframes
149 (less than 100 fixation events), initially linked genes maintain high mutual information,
150 however as linkage breaks down due to inevitable neutral mutational events, so does co-
151 segregation (Figure 1C; note *I* takes on discrete values due to the discrete nature of
152 presence /absence values used in the pangenome characterization). At equilibrium we
153 observe a skewed distribution of *I* values, with a peak at zero (Figure 1D). Figures 1A-D
154 indicate that under a neutral model, gene pair distance *d* tends to drift and co-segregation
155 *I* peaks at zero.

156

157 *Focal gene interaction model.*

158 The absence of persistent linkage in our neutral simulation model (Figure 1) is not
159 consistent with examples of highly conserved linkage across bacterial isolates, indicating
160 a role for selection. For example, the *peg* (polyethylene glycol-degradative) operon is
161 conserved across multiple lineages of *Sphingomonas* (26), likewise several *E. coli*

162 operons are conserved in different lineages (27). Therefore, to assess the role of selection
163 in our simulation model, we develop a model to include gene interactions that impact
164 fitness. In this model, events of gene gain, gene loss, and gene rearrangement are now
165 potentially selectively non-neutral if they perturb a gene interaction with defined
166 importance for fitness. As in our previous simulation we track a single pair of genes, but
167 now assume this pair has a positive interaction effect on fitness when found together in a
168 genome (and set all other gene interaction effects to zero). In contrast to Figure 1, we
169 reverse our initial conditions, so that our 2 focal genes are initially not closely linked (Figure
170 2A) and not co-segregating (Figure 2B).



171

172 **Figure 2. Positive gene interaction leads to close linkage and high co-segregation.**
173 An evolutionary simulation of 100 lineages (average 2000 genes per lineage) with a

174 positive defined interaction between two initially non-adjacent genes ($d(0) = 270$). **A.**
175 Average distance between two interacting genes *d* decreases through the course of the
176 simulation and reaches 0 after ~40000 fixation events (note log scale on x axis). **B.** After
177 10^6 fixation events, the modal genetic distance is zero. **C.** Co-segregation *I* between two
178 interacting genes increases during the simulation (from $I(0) = 0$). **D.** After 10^6 fixation
179 events, modal co-segregation $I = 0.27$.

180 Figure 2A shows that distance *d* decreases with simulation time, and although
181 stochastic effects can cause instances where distance is high, these are transient events,
182 and the distribution of distances after 200,000 fixation events is stably low, with a mode of
183 zero (Fig 2B). The shortest time to reach a distance of zero (i.e., the focal pair becoming
184 adjacent genes) is ~40,000 fixation events (Fig 2A) which translates to around 400 fixation
185 events per genome lineage in the collection. Figure 2C shows that during the simulation
186 values of *I* steadily increase. (Because *I* is a global measurement of the linkage of a pair
187 of genes *xy* in the genome collection it is less susceptible to stochastic effects). After
188 ~15,000 fixation events (150 per genome events) the value of *I* stabilizes around 0.27 and
189 remains the same for the course of the simulation (Figure 2D).

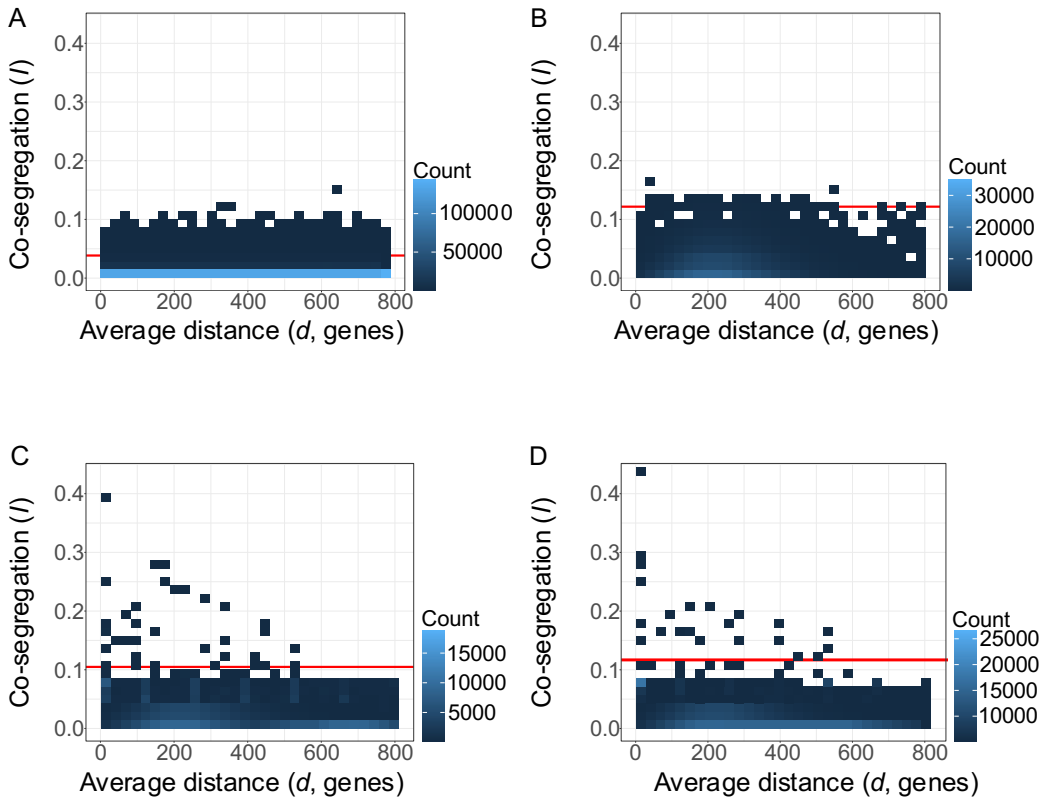
190

191 *Genome network interaction model.*

192 Figure 2 illustrates that given selection on one gene pair in isolation, close linkage
193 (low *d*) and high co-segregation *I* between these genes can emerge, despite ongoing
194 perturbing forces of gene loss, gain, translocation, etc. Yet in real-world genomes,
195 movement of two positively interacting genes together might in turn pull apart other
196 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: $r = -0.12$; $p = 0.01$).



211

212 **Figure 3. Modular gene interaction networks generate negative associations**
213 **between chromosomal distance *d* and co-segregation *I*.** Values for co-segregation
214 and distance are calculated for all pairs of genes in a collection of 100 genomes with an
215 average genome size of 2000 genes. Threshold of outlier values of *I* is denoted in red
216 (calculated with MAD (31)). **A.** Neutral model. Values of co-segregation for all pairs are
217 lower than 0.15 and distances are uniformly distributed (r = -2.3X10^{-3}, p = 0.75; outlier *I* r
218 = -3.4X10^{-4}; p = 0.97). **B.** Interaction effects distributed according to a random network
219 produce no significant linear correlation (r = -1.4X10^{-3}; p = 0.84; outlier *I* r = -4.1X10^{-3}; p =
220 0.67). **C.** A small world network produces a negative correlation (r = -0.05; p = 6.2X10^{-3};
221 outlier *I* r = -0.33; p = 2.6X10^{-12}). **D.** A preferential attachment model produces a negative
222 correlation (r = -0.12; p = 0.01; outlier *I* r = -0.62; p = 5.2X10^{-16}).

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 producing low *I* values (see neutral simulation case, Figure 3A) 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) and our random network model (Figure 3B) outliers have no correlation with distance ($r = -3.4\text{X}10^{-4}$; $p = 0.97$ and $r = -4.1\text{X}10^{-3}$; $p = 0.67$ respectively). 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: $r = -0.33$; $p = 2.6\text{X}10^{-12}$; Preferential attachment network, Figure 3D: $r = -0.62$; $p = 5.2\text{X}10^{-16}$).

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

248 Our pan-genome evolutionary simulations with modular gene interactions (Figures
249 3C,D) indicates that evolution will produce networks of co-segregating (high *I*) and
250 physically clustered (low *d*) genes. Following our arguments above, we hypothesize that
251 the networks of co-segregating genes will recapitulate the underlying modular gene
252 interaction networks, which in this simulation world are defined and accessible as a
253 'ground truth' case.

254 To test this hypothesis, we first recovered the underlying gene interaction network
255 over which the fitness effects of our simulation are based on (our ground truth). Next, we
256 built a network based on the outlier *I* values of co-segregation (the co-segregation
257 network, see methods and supplementary information). In both networks nodes represent
258 genes, and edges represent interaction effects. To measure similarity between the two
259 networks we used Hamming distance (i.e., number of changes a network must undergo
260 to be identical to the other; vertical blue lines in Figure 4). To place these similarity metrics
261 in context, we asked how often we would see the same degree of network similarity
262 (Hamming distance), given a randomly re-sampled gene interaction network. By re-
263 sampling randomly 1000 times from networks with the same underlying algorithm and
264 parameters as the specific 'ground truth' case (see methods and supplementary
265 information) we are able to assess the extent to which the observed network similarity
266 diverges from chance expectations. Figure 4 shows that modular gene interaction
267 networks such as small world or preferential attachment produce co-segregation networks
268 that recover with high fidelity the ground truth network (particularly for preferential
269 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 (Supplementary Table S1).

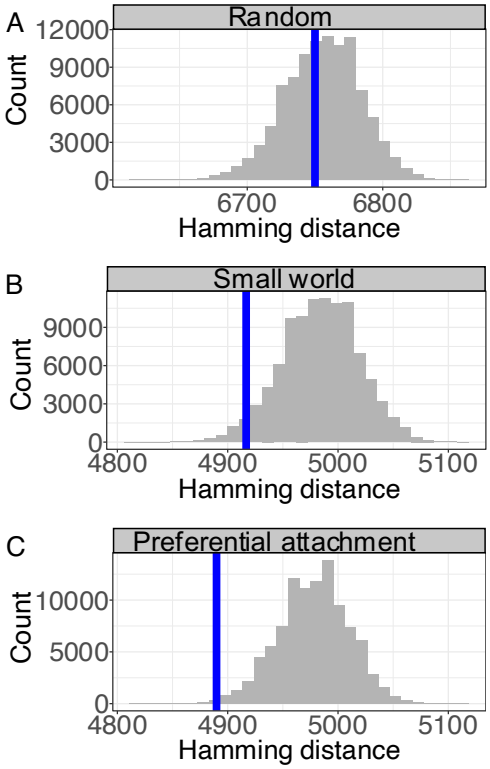


Figure 4. Co-segregation networks identify 'ground truth' gene interaction effects in simulated modular networks. Hamming distances for three different gene interaction network algorithms (ground truths) compared to the simulated values of the co-segregation network (network of outlier values of *I*). Blue lines represent Hamming

283 distances between co-segregation networks and ground truth gene interaction networks.
284 Histograms represent distributions of Hamming distances for randomly sampled gene
285 interaction networks (same number of nodes and edges, see methods and supplementary
286 information). A. Random gene interaction networks show distance close to the 50th
287 percentile compared to the background distribution. B. Small world gene interaction
288 network has a distance in the lower 10th percentile of the background distribution. C.
289 Preferential attachment gene interaction networks show a distance within the 1st
290 percentile compared to the background.

291

292 *Gene clusters*

293 Our simulation results to date examine pairwise measures of genome structure
294 (pairwise distance *d* and pairwise co-segregation *I*), which does not directly capture the
295 extent of larger-scale gene clustering. In a final simulation analysis, we define a gene
296 cluster as a set of genes that are *repeatedly* adjacent to each other, across a set of
297 genomes. In Figure 5 we plot distributions of cluster sizes, as a function of different
298 threshold values for 'repeatability' (defined by prevalence – i.e., the fraction of the
299 genomes in which a cluster is observed). In the absence of gene interactions, we find no
300 clusters larger than gene pairs, even with a permissive prevalence cuttoff of 40% (Figure
301 5A). In contrast, all our simulations with gene interaction effects show cluster sizes
302 following a geometric distribution, with abundant small clusters and rare large clusters
303 (Figure 5B-D). Unsurprisingly, when gene interactions follow a random network model,
304 clusters are small (Figure 5B). In contrast, small world and preferential attachment
305 networks can lead to bigger cluster sizes (Figure 5C, D).

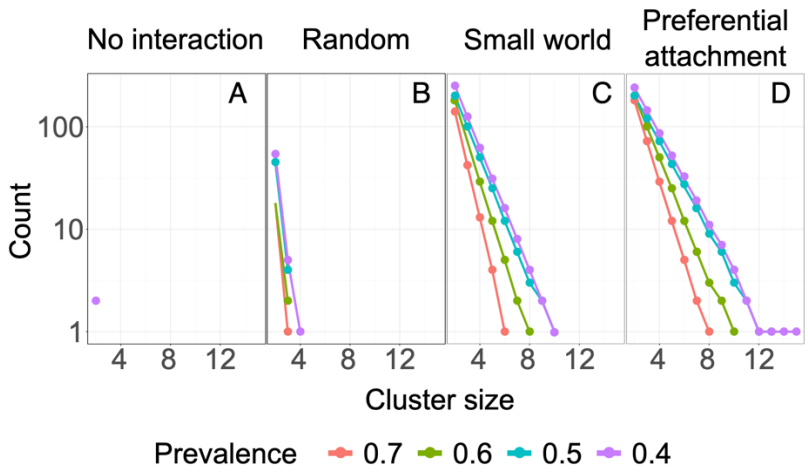


Figure 5. Modular gene interaction networks generate persistent large clusters in bacterial genomes. After evolutionary simulations reached 1 million fixation events, the number of clusters of each size is calculated for four models of gene interaction (**A**, neutral; **B**, random network; **C**, small 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.

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.

321 To capture information on chromosomal linkage, our analyses require the use of
322 completed genomes only. Our results are based on 329 *E. coli* and 179 *P. aeruginosa*
323 closed genomes from the NCBI database that met our criteria for inclusion (see methods
324 and supplementary information). For each individual genome we predicted the coding
325 sequences (CDS), finding on average 5654 CDS for *E. coli* and 6159 for *P. aeruginosa*.
326 We then compared the predicted CDS using BLAST reciprocal best match to determine
327 orthologous groups (OGs, i.e., genes with a common evolutionary origin within a species)
328 with Markov clustering. Given our focus on completed genomes only, the estimated pan-
329 genomes (38731 OGs across all genomes of *E. coli* and 16795 for *P. aeruginosa*) are
330 smaller than estimates using all available sequence data (9,32). Because most of the
331 genes are found in the accessory genome, a matrix of presence/absence of genes in each
332 genome is highly sparse and bears the characteristic histogram U shape (Figure S1,
333 Baumdicker, Hess, & Pfaffelhuber, 2012; Collins & Higgs, 2012; Lobkovsky, Wolf, &
334 Koonin, 2013). A key aspect of our analysis is to examine the extent of co-segregation
335 among gene pairs, which we can only calculate for variably present accessory genes. We
336 therefore focus in this study on the accessory genome of *E coli* (5035 genes) and *P.*
337 *aeruginosa* (4973 genes).

338

339 *Testing cluster formation predictions*

340 Our previous simulation results (Figure 5A) predict that under the neutral model
341 clusters of size 2 are rare and larger clusters are effectively absent. To quantify the
342 proportional abundance of clustered genes in the neutral simulation model, we find that
343 0.02% of the pan-genome are found in persistent gene clusters of size 2 or more

344 (measured with prevalence cutoff of 0.4, purple line in Figure 5A). In contrast, the selection
345 models allow for larger and geometrically distributed clusters, particularly under modular
346 interactions. Specifically, we found the proportion of clustered genes to be 0.9%, 6% and
347 8% under the random, small-world and preferential attachment models, respectively
348 (Figure 5B-D).

349 To assess these predictions against data, we examine the cluster distribution in
350 our pangenome data collections, using the same prevalence thresholds and defining
351 adjacency as a set of genes that are within 10KB of each other. Like our simulated results,
352 cluster size approximates a geometric distribution (Figure 6A). Indicating that large gene
353 clusters although present are rare in nature. Combining large and small clusters together
354 we find that the total fraction of persistently clustered genes (6% for *E. coli* and 12% for *P.*
355 *aeruginosa*) is in agreement with our modular network simulation models. In Figure S2 we
356 examine the sensitivity of our results to the assumption of 10KB proximity, illustrating that
357 cluster formation follows a geometric distribution in all cases (the number of clusters and
358 maximum cluster size is lower when the proximity window is smaller).

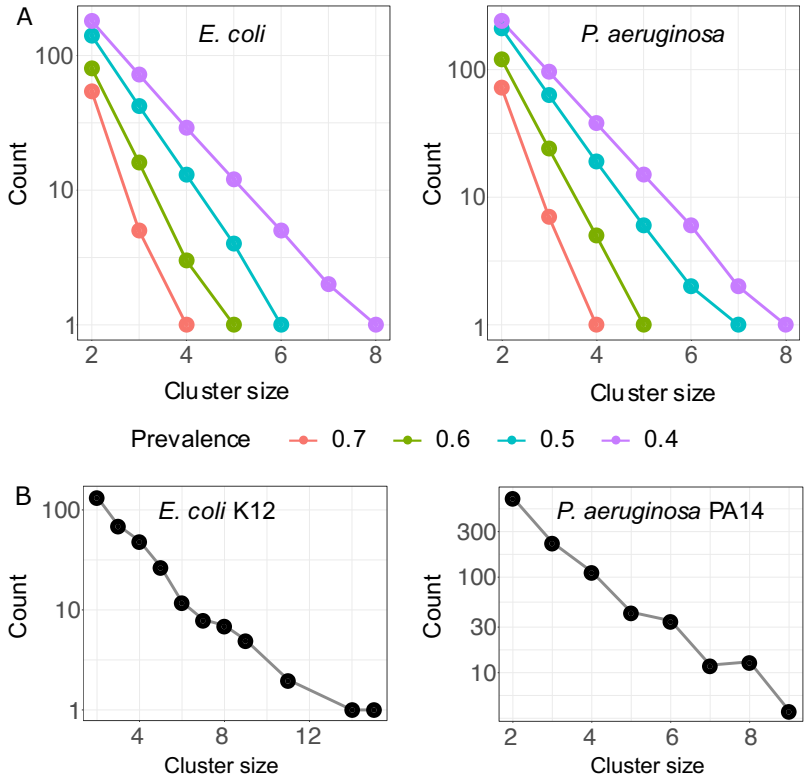


Figure 6. Persistent gene clusters are distributed geometrically across the pan-genomes *E. coli* and *P. aeruginosa*. A. Clusters of genes in the *E. coli* and *P. aeruginosa* pangenome, lines represent different prevalence for clusters in the genome collection. In both cases data follows a geometrical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

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* (36,37) which includes experimentally validated operons for the strain K12. The study on the transcriptome of *P. aeruginosa* by Wurtzel *et al*, (38) was used to estimate operons of genes in the strain PA14. In the latter an operon is defined as a set of genes transcribed

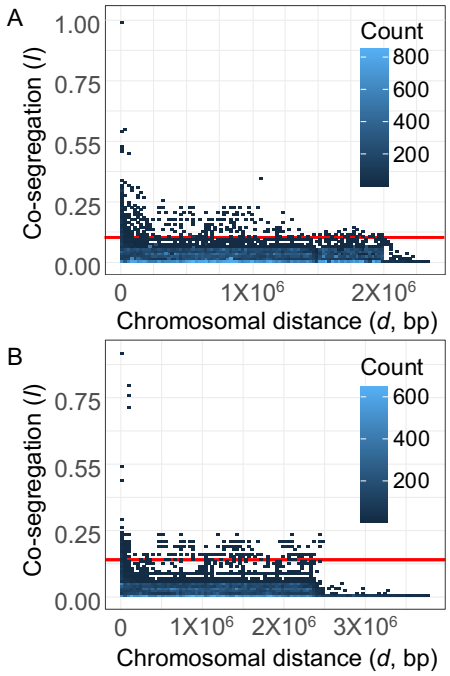
371 together. In both cases the distribution of sizes approximates a geometric distribution
372 (Figure 6B) in agreement with a modular network model (Figure 5C, D).

373

374 *Co-segregation versus linkage*

375 Figures 3 C, D predict a negative relationship between *d* and outlier values of *I*, given
376 modular gene interactions. Turning to our genomic data (Figure 7), we see a
377 predominance of low *I* values, regardless of chromosomal distance *d*. A simple correlation
378 of *I* and *d* values reveals a significant yet small negative association (Figure 7A *E. coli:*
379 Pearson correlation r =-0.02*;* p = 0.003*.* (Figure 7B *P. aeruginosa:* r = -0.012*;* p = 0.001*)*.
380 To analyze the associations with *d* for outlier values of *I,* we proceed with the same data
381 analyses as in our simulation models. After identifying outliers via the median absolute
382 deviation (31) we find a more substantial negative association between outlier *I* and *d*
383 values (Figure S3; *E. coli:* r = -0.34*;* p = $2.2X10^{-6}$*. P. aeruginosa:* r = -0.23*;* p = $1.5X10^{-6}$*)*,
384 consistent with our modular network simulation results. Figure 7 shows a pattern of sparse
385 outlier values of *I* that resembles our modular interaction models (Figure 3C, D),
386 demonstrating a bias towards the highest measures of co-segregation all occurring among
387 genes in close association (see Figure S3 for a higher resolution plot of this region). Our
388 statistical analysis of Figure 7 data treats values of *d* and *I* as statistically independent,
389 however the underlying gene pair data is derived from bacterial genomes that show
390 variable degrees of recent common ancestry. To address the potentially confounding
391 impact of phylogenetic distance, we repeated our analysis controlling for the effect of
392 phylogenetic relationships. Specifically, for each gene pair we calculated the average
393 phylogentic distance among strains that share this gene pair and added the distance

394 metric as a model covariate (see methods and supplementary information for details).
395 Correcting for phylogenetic distance produces only minor changes to our results. *E. coli,*
396 all data*:* Pearson correlation *r* =-0.02*;* p = 0.005. *P. aeruginosa all data: r* = -0.008*;* p =
397 0.0001. *E. coli,* outlier *I* values*: r* = -0.32*;* p = $1.5X10^{-6}$. *P. aeruginosa* outlier *I* values*: r* = -
398 0.19*;* p = $5.6X10^{-5}$).



399

400 **Figure 7. Outlier values of co-segregation *(I*) are negatively correlated with**
401 **chromosomal distance (*d*) across *E. coli* and *P. aeruginosa* pan-genomes.** Each data
402 point refers to an individual gene pair from among the 5035 accessory genes of *E. coli* (**A**)
403 or the 4973 accessory genes of *P. aeruginosa* (**B**). For each gene pair (*x*,*y)*, we calculate
404 average chromosomal distance *d* (across all instances of co-localization within a genome)
405 and co-segregation *I.* Across all gene pairs, we find a small negative correlation between
406 *I* and *d* (*E. coli: r* =-0.02*;* p = 0.003*. P. aeruginosa: r* = -0.012*;* p =0.001*)*. Across gene
407 pairs with outlier values of co-segregation (values above the red line: *E. coli: I* > 0.093. *P.*

408 *aeruginosa*: $I > 0.012$; see Figure S3) we find a stronger negative correlation (*E. coli:* $r = -$
409 0.34; $p = 2.2X10^{-6}$. *P. aeruginosa:* $r = -0.23$; $p = 1.5X10^{-6}$).

410

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

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

417 In our simulation models we had defined 'ground truth' gene interaction networks
418 (Figure 4). In our empirical analysis, we turn to candidate interaction networks, beginning
419 with established gene regulatory networks for *E. coli* and *P. aeruginosa* (39–42), involving
420 1258 accessory genes from *E. coli* and 987 for *P. aeruginosa*. We next assess the
421 similarity of this network with our co-segregation network via Hamming distances, and
422 again contrast this observed network distances with randomized gene interaction network
423 expectations (Figure 8A, B). The Hamming distance to the regulatory network is in the 1st
424 percentile for *E. coli* and 25th percentile for *P. aeruginosa* (Fig 8 A, B).

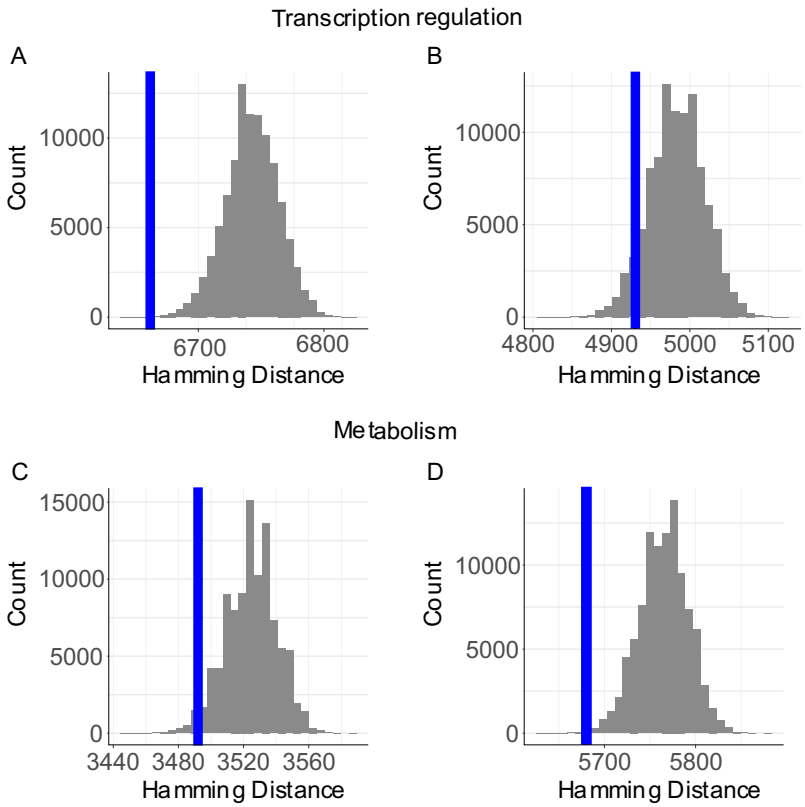


Figure 8. Co-segregation networks capture properties of both regulatory and metabolic networks in bacteria. Hamming distance between a co-segregation network obtained from the pan-genomes of 329 genomes from *E. coli* (panels A, C) and 179 genomes of *P. aeruginosa* (B, D) and their respective transcription 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.

In addition to evaluating regulatory networks, we performed the same analyses for established metabolic networks for both species (43,44), which contain 1538 accessory

437 genes of *E. coli* and 1278 genes of *P. aerugiunosa*. In this case we consider an edge
438 between two genes if they participate in the same metabolic reaction or an adjacent
439 reaction (i.e., the products of one reaction are reactants in the adjacent). We observe the
440 Hamming distance for *E. coli* to be around the 10th percentile and less than the 1st
441 percentile for *P. aeruginosa* (Fig 8 C, D).

442 To further test these conclusions, we assessed the extent to which the entire co-
443 segregation network (not just outlier values of *I*) can serve as a predictor gene interaction
444 network structure, for both metabolic and regulatory gene interaction networks (AUROC
445 analyses, Supplementary Figure S4 A, B). Consistent with Figure 8, we find that the *E.*
446 *coli* co-segregation network is most predictive of the *E. coli* regulatory gene network
447 (AUROC = 0.86) and the *P. aeruginosa* co-segregation network is most predictive of the
448 *P. aeruginosa* metabolic network (AUROC = 0.96). Altogether, the results in Figures 8 and
449 S3 show that co-segregating genes in the accessory genome are more closely associated
450 with direct regulatory links in *E. coli* (Figure 8A, S3A) and direct metabolic links in *P.*
451 *aeruginosa* (Figure 8D, S3D). The genomic data on co-segregation networks underlying
452 Figures 7 and 8 is available as Supplementary Table S2 and via the Brown lab repository
453 (https://github.gatech.edu/jcastro37/genome_evolution).

454

455 **Discussion**

456 Our pan-genomic results reveal that co-segregating genes (high *I*) tend to be
457 located closer on the genome (low *d*), for two model bacterial systems (Figure 7). While
458 physical linkage provides an intuitive explanation for this result, our neutral evolutionary
459 simulations show that initial linkage provides only a transient support for this association

460 (Figure 1). Introducing selection on gene interactions into our simulation models can
461 produce persistent linkage and co-segregation for a focal gene pair (Figure 2). When gene
462 interactions are distributed modularly across the genome, we see the emergence of
463 persistent clusters of chromosomally linked and co-segregating genes (Figure 3, 5),
464 producing patterns of co-segregation that can reveal the underlying defined gene
465 interaction network (Figure 4). Consistent with our simulation results, we find that *E. coli*
466 and *P. aeruginosa* pan-genomes contain clusters of linked and co-segregating genes
467 (Figures 6,7). Assessing our gene co-segregation networks for these two species against
468 candidate gene interaction models we find support for pan-genome evolution primarily
469 driven by metabolic (*P. aeruginosa*) and regulatory (*E. coli*) gene interactions (Figures 8,
470 S3).

471 The identification of gene interactions is a major topic in biology and has resulted
472 in numerous algorithms to define interactions from comparative genome data (45–50). A
473 key differentiating factor in our work is the use of bacterial closed genomes, which allows
474 us to analyze the dependency of genome structure on gene interactions. Our results have
475 broad implications in both applied and basic biology contexts. In the context of basic
476 biology, our results indicate how pan-genome structure is causally connected to gene
477 interaction networks. On an applied scale, this causal connection can be used to aid gene
478 function annotation. High co-segregation values can be taken as evidence for positive
479 gene interactions, providing a tool to infer functional links from genes of unknown function
480 to genes that are already annotated (e.g., established regulatory or metabolic functions).
481 This approach can be viewed as an extension to the common 'guilt by association'
482 framework (51), now leveraging repeated associations on the pan-genome scale. Current
483 functional network models, including our baseline regulatory and metabolic networks

484 examined above, are heavily biased towards the gene content of model strains(37,42–
485 44,52). Our approach offers a path to infer and extend network structures beyond the well-
486 studied domain of reference genomes.

487 We note that our results are potentially dependent on several limiting assumptions,
488 for example we currently only examine the impact of positive versus neutral fitness
489 interactions. Although this assumption is sufficient to generate gene clusters in our
490 simulation, avenues for future work could include examination of the effects of negative
491 and positive interactions on cluster formation, and/or higher-order fitness interactions (i.e.,
492 pairwise effects dependent on genetic background)(50,53,54). By focusing on orthologous
493 groups of genes, gene-dosing effects of gene duplication are also omitted from our
494 analysis (55,56). Our focus on accessory genomes captures the large majority of the pan-
495 genome of *E. coli* and *P. aeruginosa*, but omits the core genome (1.2 % and 4.5 % of *E*
496 *coli and P.aeruginosa* pangenomes, respectively). It is reasonable to expect that core
497 genes have interactions with the variable genome which we are not considering.

498 By modeling time via fixation events, we do not require parameterization of
499 absolute rates of fixation, and therefore of the strength of selection. Our model captures
500 two limit cases for the strength of selection: (1) No selection, where drift is the driving
501 evolutionary force (Figure 1) and (2) selection dominating drift (Figures 2-5). In this second
502 scenario, increasing strength of selection translates to more rapid fixation events, leaving
503 the underlying model behavior unchanged when time is measured in fixation events
504 (Figures 1-5). We do need to make assumptions however on the relative rates of distinct
505 fixation events in our simulations. Following from an assumption of genome size stability
506 (i.e. no directional change in lineage genome size through time, (54,57)), we set rates of
507 lineage gene gain and gene loss to be equal. Following previous work ((58–60)), we

508 further assume that genome rearrangement fixation events are substantially (three orders
509 of magnitude) lower than lineage rates of gene gain and gene loss. While consistent with
510 prior data and simulations, the fixation of genome rearrangement events is reported to
511 vary widely across species (Ballouz et al. 2010). To examine the impact of varying the
512 relative rates of fixation events, we find that our results are robust to order-of-magnitude
513 variation in the relative rate of genome rearrangement fixation events. Specifically, relative
514 rates of between 10^{-3} and 10^{-5} rearrangement fixations per gain (or loss) fixation result in
515 stable cluster formation (Figure S5).

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

525

526 **Methods summary**

527 An evolutionary simulation model was created to analyze pan-genome evolution in 100
528 evolving lineages (linked by horizontal gene transfer) across the course of 10^6 fixation
529 events. Each genome is an ordered arrangement of OGs, where the distribution of OGs
530 approximates the observed U-shaped distribution in bacterial genomes (61). Simulations

span neutral and selection models, where gene interactions are distributed according to one of three network models: Random (28), Small-world (29) or Preferential attachment (30). All three models are parametrized to generate sparsely populated networks (62). Physical distances *d* and co-segregation values *I* (mutual information) are calculated for all pairs of genes during the course of evolutionary simulations. Outlier values of co-segregation are estimated using median absolute deviation (MAD, 38). Model predictions were tested using a collection of complete genomes from *E. coli* and *P. aeruginosa*. OGs were predicted from predicted coding sequences (63), using BLAST reciprocal best matches and Markov Clustering (64, 65). Distances *d* for all pairs of genes were calculated in base pairs (bp) likewise co-segregation *I* was calculated for all pairs. Detailed descriptions of computational models and methods are provided in the supplementary information files.

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.

548 **References**

549 1. Brocken DJW, Tark-Dame M, Dame RT. The organization of bacterial genomes:
550 Towards understanding the interplay between structure and function. Curr Opin
551 Syst Biol. 2018;8:137–43.

552 2. Ballouz S, Francis AR, Lan R, Tanaka MM. Conditions for the Evolution of Gene
553 Clusters in Bacterial Genomes. Eisen JA, editor. PLoS Comput Biol. 2010 Feb
554 12;6(2):e1000672.

555 3. von Hippel PH, Revzin A, Gross CA, Wang AC. Non-specific DNA binding of
556 genome regulating proteins as a biological control mechanism: I. The lac operon:
557 equilibrium aspects. Proc Natl Acad Sci U S A. 1974;71(12):4808–12.

558 4. Lathe WC, Snel B, Bork P. Gene context conservation of a higher order than
559 operons. Trends Biochem Sci. 2000;25(10):474–9.

560 5. Eppinger M, Worsham PL, Nikolich MP, Riley DR, Sebastian Y, Mou S, et al.
561 Genome sequence of the deep-rooted Yersinia pestis strain angola reveals new
562 insights into the evolution and pangenome of the plague bacterium. J Bacteriol.
563 2010;192(6):1685–99.

564 6. Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA. Genomic anatomy of
565 Escherichia coli O157:H7 outbreaks. Proc Natl Acad Sci U S A.
566 2011;108(50):20142–7.

567 7. Perna NT, Plunkett G, Burland V, Mau B, Glasner JD, Rose DJ, et al. Genome
568 sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature. 2001
569 Jan;409(6819):529–33.

570 8. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al.
571 Organised Genome Dynamics in the Escherichia coli Species Results in Highly
572 Diverse Adaptive Paths. Casadesús J, editor. PLoS Genet. 2009 Jan
573 23;5(1):e1000344.

574 9. Freschi L, Vincent AT, Jeukens J, Emond-Rheault JG, Kukavica-Ibrulj I, Dupont
575 MJ, et al. The Pseudomonas aeruginosa Pan-Genome Provides New Insights on
576 Its Population Structure, Horizontal Gene Transfer, and Pathogenicity. Genome
577 Biol Evol. 2019;11(1):109–20.

578 10. Lawrence JG, Roth JR. Selfish Operons: Horizontal Transfer May Drive the
579 Evolution of Gene Clusters. Genetics. 1996 Aug 1;143(4):1843–60.

580 11. Rankin DJ, Rocha EPC, Brown SP. What traits are carried on mobile genetic
581 elements, and why. Heredity. 2011;106(1):1–10.

582 12. Pál C, Hurst LD. Evidence against the selfish operon theory. Trends in Genetics.
583 2004;20(6):232–4.

13. Pál C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD. Chance and necessity in the evolution of minimal metabolic networks. Nature. 2006;440(7084):667–70.

14. Fang G, Rocha EP, Danchin A. Persistence drives gene clustering in bacterial genomes. BMC Genomics. 2008 Dec 7;9(1):4.

15. Price MN, Huang KH, Arkin AP, Alm EJ. Operon formation is driven by co-regulation and not by horizontal gene transfer. Genome Res. 2005;15(6):809–19.

16. Price MN, Arkin AP, Alm EJ. The life-cycle of operons. PLoS Genet. 2006;2(6):0859–73.

17. Svetic RE, MacCluer CR, Buckley CO, Smythe KL, Jackson JH. A metabolic force for gene clustering. Bull Math Biol. 2004;66(3):559–81.

18. Kolesov G, Wunderlich Z, Laikova ON, Gelfand MS, Mirny LA. How gene order is influenced by the biophysics of transcription regulation. Proc Natl Acad Sci U S A. 2007;104(35):13948–53.

19. Martin FJ, McInerney JO. Recurring cluster and operon assembly for Phenylacetate degradation genes. BMC Evol Biol. 2009;9(1):1–11.

20. Bodmer WF, Parsons PA. Linkage and Recombination in Evolution. Adv Genet. 1963;11(C):1–100.

21. Stahl FW, Murray NE. The evolution of gene clusters and genetic circularity in microorganisms. Genetics. 1966;53(3):569–76.

22. Fisher RA. The genetical theory of natural selection. [Internet]. Oxford: Clarendon Press; 1930.

23. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. ISME Journal. 2009;3(2):199–208.

24. Cover TM, Thomas JA. Elements of Information Theory. Vol. 2, Elements of Information Theory. 2005. 1–748 p.

25. Pensar J, Puranen S, Arnold B, MacAlasdair N, Kuronen J, Tonkin-Hill G, et al. Genome-wide epistasis and co-selection study using mutual information. Nucleic Acids Res. 2019 Oct 10;47(18):e112–e112.

26. Tani A, Charoenpanich J, Mori T, Takeichi M, Kimbara K, Kawai F. Structure and conservation of a polyethylene glycol-degradative operon in sphingomonads. Microbiology (N Y). 2007 Feb 1;153(2):338–46.

27. Hagelsieb GM, Collado-Vides J. Operon Conservation from the Point of View of Escherichia coli, and Inference of Functional Interdependence of Gene Products from Genome Context. In Silico Biol. 2002 Jan 1;2(2):87–95.

28. Erdos P, Rényi A. On the evolution of random graphs. The Structure and Dynamics of Networks. 2011 Oct 23; 9781400841356:38–82.

621 29. Watts DJ, Strogatz SH. Collective dynamics of ‘small-world’ networks. Nature.
622 1998 Jun 4;393(6684):440–2.

623 30. Albert R, Barabási AL. Statistical mechanics of complex networks. Rev Mod Phys
624 [Internet]. 2002 Jan;74(1):47–97.

625 31. Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: Do not use
626 standard deviation around the mean, use absolute deviation around the median. J
627 Exp Soc Psychol. 2013 Jul 1;49(4):764–6.

628 32. Decano AG, Downing T. An Escherichia coli ST131 pangenome atlas reveals
629 population structure and evolution across 4,071 isolates. Sci Rep. 2019 Dec
630 22;9(1):17394.

631 33. Baumdicker F, Hess WR, Pfaffelhuber P. The Infinitely Many Genes Model for the
632 Distributed Genome of Bacteria. Genome Biol Evol. 2012;4(4):443–56.

633 34. Collins RE, Higgs PG. Testing the infinitely many genes model for the evolution of
634 the bacterial core genome and pangenome. Mol Biol Evol. 2012;29(11):3413–25.

635 35. Lobkovsky AE, Wolf YI, Koonin E V. Gene frequency distributions reject a neutral
636 model of genome evolution. Genome Biol Evol. 2013;5(1):233–42.

637 36. Salgado H, Santos A, Garza-Ramos U, Van Helden J, Díaz E, Collado-Vides J.
638 RegulonDB (version 2.0): a database on transcriptional regulation in Escherichia
639 coli. Nucleic Acids Res. 1999 Jan 1;27(1):59–60.

640 37. Huerta AM, Salgado H, Thieffry D, Collado-Vides J. RegulonDB: A database on
641 transcriptional regulation in Escherichia coli. Nucleic Acids Res. 1998 Jan
642 1;26(1):55–9.

643 38. Wurtzel O, Yoder-Himes DR, Han K, Dandekar AA, Edelheit S, Greenberg EP, et
644 al. The Single-Nucleotide Resolution Transcriptome of Pseudomonas aeruginosa
645 Grown in Body Temperature. PLoS Pathog. 2012 Sep;8(9):e1002945.

646 39. Babu MM, Teichmann SA. Evolution of transcription factors and the gene
647 regulatory network in Escherichia coli. Nucleic Acids Res. 2003;31(4):1234–44.

648 40. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M. Genomic
649 analysis of regulatory network dynamics reveals large topological changes.
650 Nature. 2004 Sep 16;431(7006):308–12.

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

653 42. Galán-Vásquez E, Luna-Olivera BC, Ramírez-Ibáñez M, Martínez-Antonio A.
654 RegulomePA: a database of transcriptional regulatory interactions in
655 *Pseudomonas aeruginosa* PAO1. Database. 2020 Dec 1;2020:106.

656 43. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive
657 genome-scale reconstruction of Escherichia coli metabolism-2011. Mol Syst Biol.
658 2011;7(535):1–9.

659 44. Bartell JA, Blazier AS, Yen P, Thøgersen JC, Jelsbak L, Goldberg JB, et al.
660 Reconstruction of the metabolic network of Pseudomonas aeruginosa to
661 interrogate virulence factor synthesis. Nat Commun. 2017;8.

662 45. Arnold BJ, Gutmann MU, Grad YH, Sheppard SK, Corander J, Lipsitch M, et al.
663 Weak epistasis may drive adaptation in recombining bacteria. Genetics. 2018 Mar
664 1;208(3):1247–60.

665 46. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in
666 microbial pan-genome-wide association studies with Scoary. Genome Biol. 2016
667 Nov 25;17(1):1–9.

668 47. Cohen O, Ashkenazy H, Burstein D, Pupko T. Uncovering the co-evolutionary
669 network among prokaryotic genes. Bioinformatics. 2012 Sep 15;28(18):i389–94.

670 48. Lassalle F, Veber P, Jauneikaite E, Didelot X. Automated reconstruction of all
671 gene histories in large bacterial pangenome datasets and search for co-evolved
672 gene modules with Pantagruel. bioRxiv. 2019 Mar 24;586495.

673 49. Liu C, Wright B, Allen-Vercoe E, Gu H, Beiko R. Phylogenetic Clustering of Genes
674 Reveals Shared Evolutionary Trajectories and Putative Gene Functions. Genome
675 Biol Evol. 2018 Sep 1;10(9):2255–65.

676 50. Mehta RS, Petit RA, Read TD, Weissman DB. Detecting patterns of accessory
677 genome coevolution in Staphylococcus aureus using data from thousands of
678 genomes. BMC Bioinformatics 2023 24:1 [Internet]. 2023 Jun 9 [cited 2023 Jun
679 12];24(1):1–16.

680 51. Oliver S. Guilt-by-association goes global. Nature 2000 403:6770. 2000 Feb
681 10;403(6770):601–2.

682 52. Strogatz SH. Exploring complex networks. Nature. 2001 Mar 8;410(6825):268–76.

683 53. Barton NH. Linkage and the limits to natural selection. Genetics.
684 1995;140(2):821–41.

685 54. Arnold BJ, Gutmann MU, Grad YH, Sheppard SK, Corander J, Lipsitch M, et al.
686 Weak Epistasis May Drive Adaptation in Recombining Bacteria. Genetics. 2018
687 Mar 1;208(3):1247–60.

688 55. Kovács K, Hurst LD, Papp B. Stochasticity in Protein Levels Drives Colinearity of
689 Gene Order in Metabolic Operons of Escherichia coli. PLoS Biol. 2009
690 May;7(5):e1000115.

691 56. Zaslaver A, Mayo A, Ronen M, Alon U. Optimal gene partition into operons
692 correlates with gene functional order. Phys Biol. 2006 Sep 1;3(3):183–9.

693 57. Sela I, Wolf YI, Koonin E V. Theory of prokaryotic genome evolution. Proceedings
694 of the National Academy of Sciences. 2016 Oct 11;113(41):11399–407.

695 58. Belda E, Maya A, Silva FJ. Genome rearrangement distances and gene order
696 phylogeny in gamma-Proteobacteria. Mol Biol Evol. 2005 Jun;22(6):1456–67.

59. Suyama M, Bork P. Evolution of prokaryotic gene order: Genome rearrangements in closely related species. Trends in Genetics. 2001;17(1):10–3.

60. Rocha EPC. Inference and Analysis of the Relative Stability of Bacterial Chromosomes. Mol Biol Evol. 2006 Mar 1;23(3):513–22.