The landscape of coadaptation in *Vibrio parahaemolyticus*

Yujun Cui[1*#], Chao Yang[1#], Hongling Qiu[2#], Hui Wang[2,3], Ruifu Yang[1], Daniel Falush[4*]

1 State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing 100071, China;

2 Institute for Nutritional Sciences, Chinese Academy of Sciences, Shanghai 200031, China

3 School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

4 University of Bath, Bath, Somerset, United Kingdom

[#]These authors contributed equally to the article

* Corresponding authors: D. F. (danielfalush@googlemail.com) or Y. C. (cuiyujun.new@gmail.com)

Abstract

Vibrio parahaemolyticus is a human gastrointestinal pathogen that thrives in warm brackish waters and is unusual amongst bacteria in having a population structure that approximates panmixia. We take advantage of this structure to perform a genome-wide screen for coadapted genetic elements. There are 93 groups of coadapted genome fragments were identified, with total length of 1.5 Mb and involved 1,703 coding genes. The great majority of interactions (85%) that we detect are between accessory genes, many involved in carbohydrate transport and metabolism. The rarity of interactions involving the core genome provides evidence for a plug and play-like architecture, in which elements have evolved to be functional immediately on arrival in a new organism. 12% of interaction were observed between core and accessory genome regions. The most complex interactions we identify include hundreds of core genome SNPs as well as accessory genome elements, which are organized into a hierarchical structure that implies progressive coadaptation. These interactions involve genes encoding lateral flagella and cell wall biogenesis, implying that several genetically distinct strategies have evolved for colonizing surfaces. The extensively coadapted genetic elements identified in this study indicated that as in human relationships, coadaptation involves progressively increasing levels of commitment, with the most involved interactions becoming irreversible and presaging speciation. Our approach provides new insight into how selection for phenotypic diversity shapes genetic variation within species.

Introduction

The importance of coadaptation to evolution was recognized by Darwin in the 6th edition of Origin of Species, where he wrote: "In order that an animal should acquire some structure specially and largely developed, it is almost indispensable that several other parts should be modified and coadapted" (1). As Darwin's argument implies, complex phenotypic innovation

require adaptation at multiple genes and it is inevitable that some of the changes involved will be costly on the original genetic background, implying epistasis - i.e. non-additive fitness interactions - between adaptive loci.

The consequences of epistasis for the evolution of phenotypic diversity depends on the transmission genetics and the population structure of the species in question. For example, in outbreeding animals mating mixes up variation each generation, with the result that genes only increase in frequency if they have high average fitness across genetic backgrounds (2). Consequently, extensive linkage disequilibrium due to natural selection is rare (3) and it is difficult to maintain dissimilar genetic strategies concurrently in the same population unless the strategies are encoded by a small number of loci. This means that the coadaptation necessary for extensive phenotypic diversification can only take place when facilitated by barriers to gene flow, such as geographical separation, mate choice or the suppression of recombination for example by inversion polymorphisms (4, 5). This feature also makes it difficult to study the process of complex coadaptation without temporally sampled genetic data, which remains rare despite the advent of technology for sequencing ancient DNA.

In bacterial populations, mutations do not need to have high average fitness across all genetic backgrounds to reach a substantial frequency in the population. For example, *Vibrio parahaemolyticus* lives in coastal waters and causes gastroenteritis in humans and economically devastating diseases in farmed shrimps. It is capable of replicating in less than 10 minutes in appropriate conditions (6), while the doubling time in the wild of the related bacteria *Vibrio cholerae* has been estimated as slightly over an hour (7). Approximately 0.017% of the genome recombines each year (8), implying that there are approximately 50 million generations on average between recombination events at a given genetic locus. As a

result, mutations that are beneficial only on specific backgrounds have a chance to rise to high frequency on those backgrounds even if they are harmful on others. Epistatic interactions that involve only small selective coefficients *s* (for example $s = 1.0\times10^{-4}$) can create an imprint on the genome in the form of strong linkage disequilibrium (9). These arguments imply that extensive complex coadaptation can potentially accumulate within a single population.

Although recombination happens slowly on the timescale of bacterial generations, the Asian population of *V. parahaemolyticus* has had a large effective population size for at least the last 15,000 years, or approximately 130 million bacterial generations, over which time recombination has been extensive enough to have almost completely scrambled genetic variation (10). As a result, the population is unusual amongst bacteria in that there is approximate linkage equilibrium between most loci greater than 3 kb apart on the chromosome (10). This feature increases the power of tests for interaction based on identifying non-random associations, which relies on identifying the same combination of alleles on independent genetic backgrounds and can therefore be confounded by clonal or population structure unless this is appropriately controlled for.

We perform a systematic screen for coadaptation in the core and accessory genome, based on a larger sample of genomes than Cui et al. 2015 (11), here for the first time performing a screen for the co-occurrence of accessory genome elements. We have taken a conservative approach to identifying statistical associations, rigorously filtering the set of genomes used in our discovery dataset to eliminate any hint of population structure. We performed more than 14 billion Fisher exact tests between variants in the core and accessory genomes, using a cut-

off of $P < 10^{-10}$ with the aim of assembling a comprehensive list of common genetic variants that have strong linkage disequilibrium between them due to fitness interactions.

We find that the great majority of interactions involve small numbers of accessory genome elements, with surprisingly little involvement of the core genome. However, we also identify three complex multi-locus interactions, in which core and accessory genomes have evolved in parallel to create coadapted gene complexes with distinct strategies. We demonstrate the value of hierarchical clustering in characterizing these interactions. Our results demonstrate that *V. parahaemolyticus* have progressively modified their own fitness landscape through coadaptation and demonstrate the fundamental importance of lateral flagella variation to their ecology. Taken together, our results provide a starting point for systematically investigating fitness landscapes in natural populations while highlighting several methodological challenges.

Results

Detection and characterization of interaction groups

To avoid false positives due to population structure, we restricted our initial analysis to the strains from the Asia population, VppAsia, within our global collection of 1,103 isolates (Table S1) and iteratively removed isolates until there was no sign of clonal structure (Methods), leading to a discovery dataset of 198 strains (Figure 1a, Figure S1). We performed a Fisher exact test of associations between all pairwise combinations of 151,957 SNP variants within the core and 14,486 accessory genome elements. As has been observed previously (11), most of strong associations occurred between sites within 3 kb on the chromosome (Figure 1b). In order to exclude associations that arise only due to physical linkage, we excluded all sets of associations that spanned less than 3 kb, including between

accessory genome elements. This left us with 452,849 interactions with $P < 10^{-10}$, which grouped into 90 networks of associated elements, all of which involved at least one accessory-genome element, with 8 also including core genome SNPs and 38 of which included multiple genome regions. In total these networks included 1,936 SNPs in 110 core genes and 1,593 accessory genome elements (Table 1, Table S2 and Table S3). Interacting SNPs were substantially enriched for non-synonymous variants, which is consistent with natural selection being the force generating the linkage disequilibrium we detected.

The largest network (network 1) accounted for the majority of interacting SNPs as well as a significant fraction of accessory genome elements (Table 1). Hierarchical clustering of the interacting elements in this network based on *P* values (Figure 1c) revealed a complex pattern of interactions but with differentiated sub-networks and we defined three large interaction groups (IGs), IG1, IG2 and IG3 within it, placing remaining interactions within IG4. IG4-93 are displayed in Figure 2, 3, Table S3 and Figure S2, while IG1-3 are shown separately in Figures 4, 5 and Table S2.

We compared our results for core genome interactions with those obtained by SuperDCA, a method that uses Direct Coupling Analysis to identify causal interactions (12, 13), using the default settings for the algorithm. To make the results as directly comparable as possible we used the same 198 isolates that were used for the Fisher exact test. The most important discrepancy is that although IG1 involves perfect associations, with *P* values as low as 1.4×10^{-30}, the coupling strengths are lower than for the other groups we identified and none appear amongst the top 5,000 couplings. This discrepancy is due the large number of SNPs involved with similar association patterns, which means that coupling values are distributed between them (Figure 1d, Figure S3). Excluding IG1 SNPs, there is a strong correlation

between SuperDCA coupling strengths and Fisher exact test *P* values (Figure 1e). At a stringent cutoff of $10^{-2.2}$, SuperDCA identifies the same multi-locus interactions as Fisher exact test does at $P < 10^{-10}$, with a few SNPs excluded (Figure 1f). The significance thresholds for both the Fisher exact test and SuperDCA could be relaxed to identify a substantially larger number of true-positive hits, at the likely expense of some false ones, but we do not investigate these associations further here. We also compared our results with those obtained by SpydrPick, a model-free method based on mutual information (MI) (14). The Fisher exact test *P* value is almost perfectly correlated with the MI statistic used by SpydrPick for this data (Figure 1e).

For IG1-4, COG classes M and N, cell wall biosynthesis and cell motility, are substantially overrepresented, relative to their overall frequency in the genome (Figure 2c, Fisher exact test, $P < 0.01$). Amongst the other interaction groups, class G, encoding carbohydrate transport and metabolism are substantially overrepresented (Fisher exact test, $P < 0.01$), particularly amongst groups involving incompatibilities. There are also differences in GC content between accessory genes in IGs and others (Figure 2c), with IGs having higher mean values, especially in compatibility IGs.

Structure of variation within interaction groups

For each interaction group, we investigated how genetic variation was structured within the *V. parahaemolyticus* population using a larger "non-redundant" set of 469 strains, which includes isolates from all four of the populations identified in (8), but excludes closely related isolates, differing at less than 2,000 SNPs. We performed hierarchical clustering of the strains based on the interaction group variants (Figure 3, 4, 5, S2). We also used the criterion used by ARACNE and SpydrPick (14, 15) to remove putatively non-causal connections between

pairs of loci that were mutually connected by statistically more significant connections to third loci than they are to each other.

The most common type of interaction group is a single accessory genome region of between 3 kb (IG53) and 57 kb (IG41), of which there are 52 (Figure 2). For example, IG5 consists of 10 genes in a single block. 9 of the genes code for various functions related to carbohydrate metabolism and transport while the 10^{th} is a transcriptional regulator (Figure 3 and Table S3). Four strains have 9 out of the 10 genes but otherwise the genes are either all present or all absent in every strain. Interestingly, there appears to be a difference in frequency between VppAsia isolates and others, with the island present in 52% of VppAsia strains and 90% of others.

Genome islands are often associated with transmission mechanisms such as phage and plasmids (16). In our data, one example is IG17 which contains 4 genes annotated as being phage related and a further 19 hypothetical proteins (Figure 3). In the reference strain, 16 of the genes occur in a single block VP1563-VP1586, while 7 genes are present elsewhere in the genome. 56% of the 469 strains had none of the genes, while 36% had more than 12 of them and 8% had between 1 and 5, which presumably represent remnants of an old phage infection. Only one gene (VP1563) in IG17 was found in appreciable frequency in strains that had none of the other genes. This gene might represent cargo of the phage infection that is able to persist for extended periods in the absence of infection due to a useful biological function of its own.

Another common interaction is incompatibility between different accessory genome elements, which is found in total 27 of the interaction groups (Figure 2). For example, all of

the 469 strains either have the gene *yniC*, which is annotated as being a phosphorylated carbohydrates phosphatase, or at least 4 out of a set of 5 genes VP0363-VP0367 that includes a phosphotransferase and a dehydrogenase (IG6, Figure 3) in the same genome location. Only one strain has both sets of genes. This interaction also involves a core genome SNP, in the adjacent gene VP0368, which is annotated as being a mannitol repressor protein. Another pattern is found in IG45 (Figure 3), where there are three genes that are mutually incompatible in our data, and 12% strains have none of the three genes. Only one of the three genes is annotated.

More detailed characterization of IG1-3

For IG1, the clustering revealed that strains fall into two cleanly differentiated groups (Figure 4a). Amongst the 469 non-redundant set, there are 43 strains from VppAsia and 1 strain from VppUS2 in the minor group that we call eco-group 1.2, or EG1.2, with the remaining 425 belong to EG1.1. Chromosome painting (Figure 4b) shows evidence for sharp peaks of differentiation of EG1.2 strains around the IG1 loci but with little evidence for differentiation elsewhere (see the expected value revealed by vertical dot line). This pattern is qualitatively distinct either for that observed between geographically differentiated populations or for clonally related strains, which show higher-than-expected copying throughout the genome but without sharp peaks (Figure S4). The composition of accessory genes also differs markedly between eco-groups, in particular because EG1.2 strains has a particularly complex polymorphic region (Ref2-09, 82 kb, Figure 4b) with substantial variation in gene content between strains, which in total constitutes for a large fraction of the gene content (57 genes) not shared with EG1.1 strains. Based on the annotation of these genes (Table S2), this region might encode a type II secretion system.

After removing putatively non-causal connections using the ARACNE criteria (15), there are still too many connections to make causal inferences for IG1 (Table S4), with multiple different genome regions retaining connections to 11 different genome blocks (Figure 4a, Table S2), reflecting the large number of sites in perfect or near-perfect linkage disequilibrium with each other. However, hierarchical clustering based on IG1 SNPs helps to identify coherent patterns amongst interacting SNPs. We split the genetic variants into three tiers, shown in red, blue and green. Tier 1 variants distinguish cleanly, or nearly cleanly between EG1.1 and EG1.2 strains. Tier 2 variants are similar but with more discrepancies, while Tier 3 variants are typically fixed or nearly fixed within one eco-group and polymorphic within the other. Each of the tiers includes multiple core and accessory genome regions. Some regions are represented in only one tier, while others contain two or more. For example, the lateral flagellar gene cluster (Ref1-23, VPA1538-VPA1557, Figure 4c) contains 383 Tier 1 SNPs, 24 from Tier 2 and 47 from Tier 3, however the extent of the region spanned by Tier 1 SNPs is smaller, so that for example *flhA* and *flhB*, at the center of the gene cluster only contain Tier 3 SNPs, while *lafK* and *lafA* to the left and *filS* to the right contain multiple fixed differences between eco-groups. In other regions of this genome cluster, Tier 1 SNPs are flanked by those from other tiers.

IG2 includes 548 SNPs and 130 accessory genes which located in 26 regions (Figure 5a), including a *LuxR* family transcriptional regulator (Ref1-11), T6SS (type VI secretion system, Ref1-05), and cellulose synthesis related genes (Ref2-02) that had been identified in our previous research (11). Here with increased statistical power with more genome sequences, we found additional interacting loci. Tier 1 contains the T6SS and the cellulose synthesis related genes, reflecting their strong incompatibility, with nearly all EG2.1 strains containing a T6SS and all those in EG2.2 containing genes annotated as part of the biosynthesis cluster.

There are also SNPs in a hypothetical protein, VPA1081 (Ref1-10) that have a very similar distribution amongst strains to T6SS presence/absence, suggesting strong functional linkage. Tier 2 contains mostly SNPs in core genome, which are located at *LuxR* family transcriptional regulator and uridine phosphorylase encoding genes (Ref1-10). Tier 3 contains 85% of all variants in IG2, which encoded genes that functional related with biogenesis of elements in cell membrane, carbohydrate transport and metabolism, and transcriptional regulators.

Although the ARACNE filtering left too many interactions to be immediately useful in causal inference, it did highlight one additional gene as potentially being particularly important. The single accessory gene in Tier2, group_3560 (Ref2-10), codes for a polysaccharide biosynthesis/export protein and retains interactions with 9 genome blocks after filtering, 3 more than any other IG2 gene (Table S2). The large number of causal interactions inferred for this gene reflects its strong association with a sub-cluster of strains within EG2.2, labelled as EG2.2b (Figure 4a), which is stronger than for any other genetic element, including the SNPs in the *LuxR* transcriptional regulator genes. The remaining strains EG2.2a include all EG1.2 strains and others. These results suggest that amongst strains lacking the T6SS, there are three distinct strategies, one of which is determined by presence or absence of group_3560, the second is determined by the large number of SNPs and genes in IG1.2 discussed above, while the third is too rare in our sample to be categorized, even provisionally, but seems to be associated with presence of a handful of the genes in blocks Ref1-02 and Ref1-08.

In part due to this diversity of strategies, the overall strength of the associations in the IG2 Tiers we have defined are generally weaker than those for IG1: while there are a handful of

core genome regions where EG2.1 regions are differentiated (Figure 5a), none of these is as clear-cut as the differentiated regions of IG1.

The loci of IG3 only covered eight genome regions, which contained 75 SNPs and 65 accessory genes, most of which are found in a single genome regions which includes the core gene *TonB3* (VP0163, Ref1-01) (17) (Figure 5b). Clustering based on variants of IG3 revealed a stair-like structure of variation. Tier 1 variants differentiate EG3.1 from other eco-groups and include multiple SNPs in *TonB3* and in the lateral flagellar gene cluster (Ref1-04). Tiers 2 and 3 variants differentiate between EG3.2, EG3.3 and EG3.4, which are progressively more different from EG3.1, both in terms of accessory genome complement and core genome SNPs. Many of the accessory genes specific to EG3.4 are annotated to have functions associated with cell wall biogenesis (Ref2-01). Once again, a large number of interactions were left after ARACNE filtering (Table S2, S4), frustrating explicit causal inference using this approach.

Phenotypic differences between EG1.1 and EG1.2

We performed a preliminary analysis of the phenotypic differences underlying EG1 by determining the motility, growth rate, and biofilm formation ability (Figure 6) of in 7 EG1.1 and 4 EG1.2 strains on laboratory media. We failed to observe differences in swimming or swarming capability under the conditions tested but EG1.2 strains revealed significantly higher biofilm formation ability and faster growth rate than EG1.1 strains (Figure 6b and 6c), and they revealed rough colony morphology, also an indication of increased biofilm formation, under low salinity (1% NaCl) culture condition (Figure 6c).

289 There are in total 60 EG1.2 strains in the global collection of 1103 *V. parahaemolyticus*
290 strains. All but one, a VppUS2 isolate, is from the VppAsia population, with the majority (n=
291 48, 80%) in this study coming from routine surveillance on food related environmental
292 samples, including fish, shellfish, and water used for aquaculture. The strains revealed no
293 clear geographical clustering pattern in China, as they can be isolated from all six provinces
294 that under surveillance. Notably, only 4 EG1.2 strains were isolated from clinical samples,
295 including wound and stool, representing a lower proportion than for EG1.1 isolates
296 (453/1043), including if the two pandemic lineages are removed (268/798), suggesting this
297 eco-group has low virulence potential in humans.
298
299 **Discussion**
300 Bacterial traits such as pathogenicity, host-specificity and antimicrobial resistance naturally
301 attract human attention but less obvious or even cryptic traits might be more important in
302 determining the underlying structure of microbial populations. Studies of coadaptation based
303 on genome sequencing of thousands of isolates have the potential to provide new insight into
304 the ecological forces shaping natural diversity, and how that variation is assembled by
305 individual strains to overcome the manifold challenges involved in colonizing specific niches,
306 such as the human gastrointestinal tract. In other words, these studies provide a unique
307 opportunity to see the world from the point of view of a bacterium.
308
309 We performed a genome wide scan for coadaptation in *V. parahaemolyticus*, performing
310 pairwise tests for interactions amongst genetic variants and then clustered the significant
311 pairwise interactions into 93 interaction groups. Our analysis demonstrated that genome wide
312 epistasis scans can be used successfully to identify diverse interactions involving both core
313 and accessory genomes but also highlighted unsolved methodological challenges.

Firstly, pairwise tests should, at least in principal, have reduced statistical power compared to methods that analyze all of the data at once, such as Direct Coupling Analysis (DCA) (12). However, while there was a strong correlation between DCA and our results for core-genome interactions, DCA failed to identify the clearest, most extensive interaction in our dataset, namely IG1. DCA was designed to identify coupling interactions that take place during protein folding and implicitly entails that pairwise interaction between a given pair of sites make it less likely that other sites will interact with either of them. For many types of interaction, a prior that makes the opposite assumption seems more appropriate, for example because master regulation loci are likely to interact with many different sites. Thus, in order to develop statistical tests that exploit the full power of genomic data, new types of statistical test that search for a more diverse range of interactions would need to be developed.

Second, distinguishing direct associations – either through gene function or ecology – from those that arise due to mutual correlation with other interacting genes, is a substantial, and largely unaddressed challenge. For the complex interaction groups in our data, the criteria used by ARACNE (15) to remove interactions still left far too many interactions to be interpreted usefully as being causal. We found that hierarchical clustering organized the interactions in a manner that allowed informal interpretation but once again new statistical methodology is needed to facilitate detailed dissection of associations.

Notwithstanding the unresolved challenges, our results highlight the central role of lateral motility in structuring ecologically significant variation within the species. We also find evidence that interactions move through progressive stages, analogous to differing degrees of

338 commitment within human relationships, namely casual, going steady, getting married and
339 moving out together (Figure 7).
340
341 **Most interactions between core and accessory genomes are casual**
342 A recent debate about whether the accessory genome evolves neutrally (18-20) highlighted
343 how little we know about the functional importance of much of the DNA in bacterial
344 chromosomes. Using the same statistical threshold to assess significance, our interaction
345 screen identifies many more examples of coadaptation between different accessory genome
346 elements than of interactions between the core and accessory genomes or within the core
347 genome, implying that natural selection has a central role in determining accessory genome
348 composition.
349
350 Unsurprisingly, given the extensive literature highlighting the importance of genomic islands
351 to functional diversity of bacteria (16), the most common form of adaptation detected by our
352 screen is the coinheritance of accessory genome elements located in the same region of the
353 genome. Based on a minimum size threshold of 3 kb, we find 52 (56%) such interactions, the
354 largest of which is 57 kb (Table S3).
355
356 Previous approaches to detecting genome islands emphasize traits associated with horizontal
357 transmission, for example based on differences in GC content with the core genome, the
358 presence of phage-related genes or other markers of frequent horizontal transfer (21). Our
359 approach, based simply on co-occurrence identifies a wider range of coinherited units and
360 suggests that many islands have functions related to carbohydrate metabolism.
361

Amongst interactions not involving physical linkage, the most common is incompatibility of different accessory genome elements, representing 29% (27/93) of our IGs. For example, two different versions of phosphorous pathway, one involving one gene, the other involving 5 (IG6, Figure 3).

We propose that the rarity of interactions between core and accessory genome in our scan reflects the evolution of "plug and play"-like architecture for frequently transferred genetic elements. Accessory genome elements are more likely to establish themselves in new host genomes if they are functional immediately on arrival in many genetic backgrounds. Furthermore, from the point of view of the host bacteria, acquisition of essential functions in new environments is more likely if diverse accessory genome elements in the gene pool are immediately functional on arrival in the genome.

Sometimes it makes sense to go steady

When an accessory genome element with an important protein coding function arrives in a new genome, it is likely that some optimization of gene regulation will be possible, coordinating the expression of the gene with others in the genome. We found 11 different interaction groups involving core and accessory genome regions. One simple example included a regulatory gene VP0368 and an accessory genome element (IG6, Figure 3). In this example, it is feasible for the core genome SNP and the associated accessory element to be transferred together between strains in a single recombination event. Where coadaptation involves two or more separate genome regions, this makes assembling fit combinations more difficult and is likely to slow down the rate at which strains gain and lose the accessory genome elements involved. This higher degree of fidelity in turn makes further coadaptation at additional genes more likely.

IG2, is an example of a complex coadaptation involving multiple core and accessory genome regions. A large majority of strains in our dataset (439/469) either carry a cluster of genes encoding a T6SS, or a cluster encoding cellulose biosynthesis genes, but few strains have genes from both clusters (Figure 5a). Cells uses the T6SS to inject toxins into nearby bacteria (22) and cellulose production to coat themselves in a protective layer (23). Incompatibility might have a functional basis, for example because cellulose production prevents the T6SS functioning efficiently, or an ecological one, for example because cells that attack others do not need to defend themselves. The evolution of dissimilar strategies has led to differentiation in gene/SNP frequencies in a large number of regions, including the variants in IG1 and IG3, that presumably also represent functional or ecological coadaptations to these two distinct strategies.

Marriage changes everything

IG1 differs from the other interaction groups in our scan in both the number of associated regions and the strength of the associations. The interaction group include 454 SNPs in the 19 gene 18 kb lateral flagellar gene cluster (VPA1538-1557, Figure 4c), a further 917 core genome SNPs in 62 genes and 152 accessory genes in 35 clusters. Strains cluster cleanly into two groups, the more common group being designated EG1.1 and the rarer EG1.2. Comparison with closely related species shows that the polymorphisms distinguishing EG1.1 and EG1.2 have evolved *de novo* within *V. parahaemolyticus*, with the EG1.2 variant undergoing faster evolution (Figure S5).

Many of the accessory genes and SNPs are in perfect or near perfect disequilibrium with SNPs in the flagellar gene cluster. These include loci encoding flagellar genes, T2SS and

412 other membrane transport elements. There are also 285 loci (27%, Tier 3) in weaker
413 disequilibrium, typically because they are polymorphic in one of the eco-groups. Many of
414 variants are likely to represent more recently evolved coadaptation. Some of these genes are
415 also associated with flagella or the T2SS related function but also encompass a broader range
416 of functional categories, including cell division and amino acid transport and metabolism
417 (Table S2).
418
419 Our laboratory phenotype experiments (Figure 6) suggest that biofilm formation is likely to
420 be a key trait underlying the different ecological strategies of EG1.1 and EG1.2, but the
421 variation in phenotypic response at different salinity levels and the absence of measurable
422 difference in swarming behavior, despite the large genetic difference within the lateral
423 flagella genes, highlight some of the manifold difficulties of interpreting natural variation
424 using phenotypes measured under laboratory conditions.
425
426 Despite the extensive differences that have accumulated between EG1.1 and EG1.2, there is
427 no evidence of restricted gene flow in most of the genome (Figure 4b), and even within the
428 flagellar gene cluster strongly differentiated regions are separated by a weakly differentiated
429 one (Figure 4c), implying that the coadaptation is being maintained by selection in the face of
430 frequent recombination. Initial divergence in flagellar function is likely to have led to
431 ecological differentiation, which led to bacteria having different nutritional inputs or
432 requirements and a broadening of the functional categories undergoing divergent selection.
433
434 How can the difference between IG1 and the other interaction groups in both the number of
435 associations and their strength be explained? *V. parahaemolyticus* is ubiquitous in shellfish in
436 warm coastal waters, within which it occurs at densities of around 1,000 cells per gram, so a

437 back of the envelope calculation suggests there are likely to be substantially more than 10^{15}
438 bacteria in the VppAsia population. The species also has a high estimated effective
439 population size (10, 11) and has strong codon bias, which is often argued to be evidence that
440 even tiny selective coefficients can drive adaptation (24). Furthermore, recombination only
441 breaks up linkage disequilibrium between loci slowly. Therefore, weaker and more variable
442 patterns of association within IG2 and IG3 than in IG1 is unlikely to be a simple consequence
443 of the ineffectiveness of selection and is instead likely to reflect complexity in the fitness
444 landscape.
445
446 Strains gain flexibility by being able to switch between or modulate genomically encoded
447 strategies by homologous recombination. For example, expression of the T6SS might be
448 essential for survival in crowded habitats but detrimental in sparsely populated ones.
449 Recombinants between EG2.1 and EG2.2 at IG2 loci might represent transient adaptation of
450 strains to their immediate environment or long-term adaptation of intermediate strategies.
451 Furthermore, the phenotypic consequences of IG2 variants can depend on other loci in the
452 genome, such as IG1 variants, which is likely to reduce the strength of associations within
453 IG2. Crucially, the evolution of promiscuity is self-reinforcing because the presence of strains
454 using multiple strategies in the population also favors the presence of accessory genes and
455 core gene haplotypes that have high or intermediate fitness on multiple backgrounds.
456
457 On the other hand, an absence of intermediate genotypes in the population can favor the
458 evolution of fastidiousness, with particular accessory genes and haplotypes becoming
459 essential components of some genetic backgrounds but deleterious on others. A likely
460 scenario is that differences between IG1.1 and IG1.2 at a lateral flagellar gene made
461 recombinants between the two versions of the gene inviable and also created divergent

462 selection at a handful of other loci that was largely independent of the external environment
463 or of interactions with other genes. The evolution of fastidiousness, like the evolution of
464 promiscuity, can be self-reinforcing, and might have led to progressive increase in the
465 differentiation of EG1.2 strains from EG1.1 until the coadaptation of IG1.2 variants to each
466 other and of IG1.1 variants to each other became more-or-less irreversible, like marriage in
467 England prior to the reign of King Henry VIII.
468
469 **Coadapted gene complexes as speciation triggers**
470 Running the tape forward, it is easy to envisage the number of coadapted regions of the
471 genome within IG1 undergoing progressive enlargement, until the entire genome becomes
472 differentiated. As coadapted regions become more numerous, the proportion of
473 recombination events between eco-groups that are maladaptive will increase, which might
474 prompt the evolution of mechanistic barriers to genetic exchange between them.
475
476 Mechanisms by which new bacterial species arise are frequently discussed in the literature
477 (25-27) but there is currently little data on how the process unfolds. IG1 is of interest both as
478 an example of an intermediate stage of divergence, prior to speciation, and because it
479 suggests that substantial adaptive divergence between gene pools can precede any barriers to
480 genetic exchange, other than natural selection at the loci involved. This – unique to our
481 knowledge – example is exciting because the distinct signature of selection should make it
482 possible to dissect the genetic basis of coadaptation in unprecedented detail. Broadly similar
483 patterns of differentiation including “genomic islands of speciation” have been observed for
484 example between ecomorphs of cichlid fishes (28), but the evolution of ecomorphs has been
485 facilitated by fish preferring to mate with similar individuals, which will have also inevitably
486 lead to some level of differentiation at neutral loci throughout the genome.

Conclusions

In *V. parahaemolyticus*, it has been possible to distinguish clearly between adaptive processes, reflecting fitness interactions between genes and neutral ones, reflecting clonal and population structure. This has allowed us to provide a description of the landscape of coadaptation, involving multiple simple interactions and a small number of complex ones. We have focused on interactions that generate strong linkage disequilibrium, but weaker and more complex polygenic ones also have the potential to provide biological insight.

Most bacteria have population structure that deviates more markedly from panmixia (10). In some species this is likely due to smaller effective population sizes, lower recombination rates or mechanistic barriers to genetic exchange between strains. However, coadaptation can itself generate genome-wide linkage disequilibrium that might be difficult to distinguish from clonal or population structure. Because the linkage disequilibrium associated with IG1 is highly localized within the genome, it can, on careful inspection be clearly be attributed to selection, but in other bacteria patterns are likely to be less straightforward, making it challenging to understand to whether adaptive processes drive population structure, or vice versa. Natural selection is the jewel of evolution but distinguishing it from other processes requires in depth understanding of the relevant biology in addition to suitable data and statistical methods.

Materials and Methods

Genomes used in this work

Totally 1,103 global *V. parahaemolyticus* genomes were used in this work (Table S1), which also were analyzed in our other studies (8, 10). To reduce clonal signals, we firstly made a

512 "non-redundant" dataset of 469 strains, in which no sequence differed by less than 2,000
513 SNPs in the core genome. They were attributed to 4 populations, VppAsia (383 strains),
514 VppX (43), VppUS1 (18) and VppUS2 (21) based on fineSTRUCTURE result (29). We then
515 focused on VppAsia which has more strains, to generate a genome dataset in which strains
516 represent a freely recombining population. We selected 386 genomes from 469 non-
517 redundant genome dataset, including all the 383 VppAsia genomes and 3 outgroup genomes
518 which were randomly selected from VppX, VppUS1 and VppUS2 population, respectively.
519 These 386 genomes were used in Chromosome painting and fineSTRUCTURE analysis (29)
520 as previously described (11). Initial fineSTRUCTURE result revealed multiple clonal signals
521 still exist, thus we selected one representative genome from each clone, combined them with
522 the remaining genomes and repeated the process. After 14 iterations, we got a final dataset of
523 201 genomes with no trace of clonal signals, involving 198 VppAsia genomes that were used
524 in further analysis.
525
526 The copy probability value of each strain at each SNP was generated by Chromosome
527 painting with "-b" option, and the average copying probability value of a given strain group
528 (e.g. EG1.2) at each SNP was used in Figure 4, 5 and Figure S4.
529
530 **Variation detection, annotation and phylogeny**
531 We re-called SNPs for 198 VppAsia genomes by aligning the assembly against reference
532 genome (RIMD 2210633) using MUMmer (30) as previous described (11). Totally 565,466
533 bi-allelic SNPs were identified and 151,957 bi-allelic SNPs with minor allele frequency > 2%
534 were used in coadaptation detection. We re-annotated all the assemblies using Prokka (31),
535 and the annotated results were used in Roray (32) to identify the pan-genome and gene
536 presence/absence, totally 41,052 pan-genes were found and 14,486 accessory genes (present

537 in > 2% and < 98% strains) were used in coadaptation detection. The pan-gene protein
538 sequences of Roary were used to BLAST (BLASTP) against COG and KEGG database to get
539 further annotation.
540
541 The Neighbour-joining trees were built by using the TreeBest software
542 (http://treesoft.sourceforge.net/treebest.shtml) based on sequences of concatenated SNPs, and
543 were visualized by using online tool iTOL (33).
544
545 **Detection of coadapted loci**
546 Totally 151,957 bi-allelic SNPs and 14,486 accessory genes identified from 198 independent
547 VppAsia genomes were used in coadaptation detection by three methods. Firstly we used
548 Fisher exact test to detect the linkage disequilibrium of each SNP-SNP, SNP-accessory gene,
549 and accessory gene-gene pair. Presence or absence of an accessory gene was considered as its
550 two alleles. Each variant locus (SNP or accessory gene) has two alleles, major and minor, of
551 which major represents the allele shared by majority of isolates. For each pair of loci X and
552 Y, the number of combinations between X_{major}-Y_{major}, X_{major}-Y_{minor}, X_{minor}-Y_{major}, X_{minor}-Y_{minor}
553 were separately counted and used in the contingency table to calculate the Fisher exact test *P*
554 value. It took 3 days to finish all the coadaptation detection in a computer cluster using 21
555 cores and 2 Gb memory. We also used SuperDCA (13) and SpydrPick (14) to detect the
556 coadaptation between SNPs, using the same subset of 198 strains to make the analysis as
557 comparable as possible. SuperDCA is based on direct coupling analysis (DCA) model (12)
558 and has a much faster calculation speed compared with previous DCA methods. However, it
559 still took 25 days to finish the detection by using 32 cores and 86 Gb memory. SpydrPick
560 took one hour to finish the calculation by using 32 cores and 1 Gb memory.
561

We removed coadaptation pairs with distance less than 3 kb to minimize the influence of physical linkage. All identified SNPs in this study were located in the core genome, therefore the physical distance between SNP pairs can be calculated according to their position in the reference genome. To define the distance between accessory genes, and between SNP and accessory gene, we mapped the sequence of accessory genes against available 19 complete maps of the *V. parahaemolyticus* genomes to acquire their corresponding position, and then the gene that failed to be found in complete reference genomes were then mapped to the draft genomes. If the accessory genes pair or SNP- accessory gene pair was found located in a same chromosome or same contig of a draft genome, then the distance between paired variants could be counted according to their position in the chromosome or contig. The distance between paired variants that located in different chromosomes or contigs was counted as larger than 3 kb and such pairs were kept in further analysis. Circos (34) was used to visualize the networks of coadaptation SNPs in Figure 1f and Figure S3.

Lateral flagellar gene cluster region in *Vibrio* genus

To identify the homologous sequences of *V. parahaemoluticus* lateral flagellar gene cluster (VPA1538-1557) in the *Vibrio* genus, we downloaded all available *Vibrio* genome sequences in NCBI, then aligned the nucleotide sequence of lateral flagellar gene cluster of *V. parahaemolyticus* (NC_004605 1639906-1657888) against *Vibrio* genome dataset (excluding *V. parahaemolyticus*) by using BLASTN. Totally 46 *Vibrio* genomes revealed above than 60% coverage on lateral flagella region in *V. parahaemolyticus* genome and was used in phylogeny rebuilding. We also included three randomly selected strains from EG1.1 and EG1.2 respectively for comparison. In total 3,000 SNPs were identified in this region and were used for NJ tree construction.

587 **Determination of phenotypes**

588 **Bacteria strains.** In the phenotype experiments, totally 11 strains were randomly selected
589 respectively from two EGs that defined by IG1 variants, including 7 EG1.1 strains (B1_10,
590 B1_3, B2_10, B4_8, C1_5, C5_2, C6_5) and 4 EG1.2 strains (B1_1, B3_1, B5_2, C3_10).
591 The strains stored at -80 °C were inoculated in the thiosulfate citrate bile salts sucrose agar
592 (TCBS) plates by streak plate method. Five clones for each strains were inoculated again in
593 another TCBS plate and then cultured overnight at 30 °C in 3% NaCl-LB broth overnight and
594 used for the following assays.

595

596 **Motility assays.** Five clones for each strain were cultured overnight at 30 °C and then
597 inoculated in the swimming plate (LB media containing 0.3% agar) and swarming plate (LB
598 agar with 3% NaCl). The swimming ability was recorded by measuring the diameter of
599 colony after 24 hours at 30 °C. And the swarming ability was recorded after 72 hours at 24 °
600 C.

601

602 **Growth curve.** *V. parahaemolyticus* strains in 96-well plate were cultured overnight at 30 °C
603 in 3% NaCl-LB broth. The optical density of each culture was adjust to an OD_{600} of 0.6.Then
604 1 ml of each culture was inoculated 100 ml of 3% NaCl-LB broth in a 96-well plate and
605 cultured at 30 °C. The growth of each culture were measured every 1 hour at the optical
606 density of 600 nm using Multiskan Spectrum.

607

608 **Biofilm formation.** *V. parahaemolyticus* strains were cultured overnight at 30 °C in 3%
609 NaCl-LB broth. 2 μl of each overnight culture was inoculated to 100 μl of 3% NaCl-LB broth
610 in a 96-well plate and cultured at 30 °C for 24 h statically. The supernatant was discarded and

611 each well was washed once with sterile phosphate-buffered saline (PBS). 0.1% Crystal violet
612 (wt/vol) was added to each well and incubated at room temperature for 30 min. The crystal
613 violet was decanted, and each well was washed once with sterile PBS. Crystal violet that
614 stained biofilm was solubilized with dimethylsulfoxide (DMSO), and then measured at the
615 optical density of 595 nm using Multiskan Spectrum (Thermo Scientific).
616

617 **Acknowledgements**

618 This work is supported by the National Key Research & Development Program of China (No.
619 2017YFC1601503, 2016YFC1200100, and 2017YFC1200800), Sanming Project of
620 Medicine in Shenzhen (No. SZSM201811071), the National Natural Science Foundation of
621 China (No. 31770001) and the Key Research Program of the Chinese Academy of Sciences
622 (No. ZDRW-ZS-2017-1). D.F. is funded by a Medical Research Council Fellowship as part
623 of the MRC CLIMB consortium for microbial bioinformatics (grant number
624 MR/M501608/1).
625

626 **Author Contributions**

627 D. F., Y. C., and R. Y. designed the study and coordinated the project; Y. C., C. Y., and D. F.
628 analyzed the data; H. Q. and H. W. performed phenotype experiments; D. F. and Y. C. wrote
629 the manuscript. All authors approved the final version of the manuscript.
630

631 **Conflict of interest**

632 The authors declare that they have no conflict of interest.
633

634 **References**

635 1. Darwin C. The origin of species. 6th. John Murray, London; 1859.

2. Neher RA, Shraiman BI. Statistical genetics and evolution of quantitative traits. Reviews of Modern Physics. 2011;83(4):1283.
3. Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. The American Journal of Human Genetics. 2001;69(1):1-14.
4. Dobzhansky T. 1937Genetics and the origin of species. NewYork: Columbia University PressDobzhanskyGenetics and the origin of species1937. 1970.
5. Wallace B. Coadaptation revisited. The Journal of heredity. 1991;82(2):89-95.
6. Makino K, Oshima K, Kurokawa K, Yokoyama K, Uda T, Tagomori K, et al. Genome sequence of Vibrio parahaemolyticus: a pathogenic mechanism distinct from that of V cholerae. Lancet. 2003;361(9359):743-9.
7. Gibson B, Wilson DJ, Feil E, Eyre-Walker A. The distribution of bacterial doubling times in the wild. Proceedings Biological sciences. 2018;285(1880).
8. Yang C, Pei X, Wu Y, Yan L, Yan Y, Song Y, et al. Recent mixing of Vibrio parahaemolyticus populations. bioRxiv. 2018.
9. Arnold BJ, Gutmann MU, Grad YH, Sheppard SK, Corander J, Lipsitch M, et al. Weak Epistasis May Drive Adaptation in Recombining Bacteria. Genetics. 2018;208(3):1247-60.
10. Yang C, Cui Y, Didelot X, Yang R, Falush D. Why panmictic bacteria are rare. bioRxiv. 2018.
11. Cui Y, Yang X, Didelot X, Guo C, Li D, Yan Y, et al. Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen Vibrio parahaemolyticus. Molecular biology and evolution. 2015;32(6):1396-410.
12. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences. 2011;108(49):E1293-E301.
13. Puranen S, Pesonen M, Pensar J, Xu YY, Lees JA, Bentley SD, et al. SuperDCA for genome-wide epistasis analysis. Microbial genomics. 2018.
14. Pensar J, Puranen S, MacAlasdair N, Kuronen J, Tonkin-Hill G, Pesonen M, et al. Genome-wide epistasis and co-selection study using mutual information. BioRxiv. 2019:523407.
15. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al., editors. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC bioinformatics; 2006: BioMed Central.
16. Dobrindt U, Hochhut B, Hentschel U, Hacker J. Genomic islands in pathogenic and environmental microorganisms. Nature reviews Microbiology. 2004;2(5):414-24.
17. Tanabe T, Funahashi T, Okajima N, Nakao H, Takeuchi Y, Miyamoto K, et al. The Vibrio parahaemolyticus pvuA1 gene (formerly termed psuA) encodes a second ferric vibrioferrin receptor that requires tonB2. FEMS microbiology letters. 2011;324(1):73-9.
18. Vos M, Eyre-Walker A. Are pangenomes adaptive or not? Nature microbiology. 2017;2(12):1576.
19. Andreani NA, Hesse E, Vos M. Prokaryote genome fluidity is dependent on effective population size. ISME J. 2017;11(7):1719-21.
20. Shapiro BJ. The population genetics of pangenomes. Nature microbiology. 2017;2(12):1574.
21. Langille MG, Hsiao WW, Brinkman FS. Detecting genomic islands using bioinformatics approaches. Nature reviews Microbiology. 2010;8(5):373-82.
22. Salomon D, Gonzalez H, Updegraff BL, Orth K. Vibrio parahaemolyticus type VI secretion system 1 is activated in marine conditions to target bacteria, and is differentially regulated from system 2. PloS one. 2013;8(4):e61086.
23. Tischler AD, Camilli A. Cyclic diguanylate (c-di-GMP) regulates Vibrio cholerae biofilm formation. Molecular microbiology. 2004;53(3):857-69.
24. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. Variation in the strength of selected codon usage bias among bacteria. Nucleic acids research. 2005;33(4):1141-53.
25. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, et al. Population genomics of early events in the ecological differentiation of bacteria. Science. 2012;336(6077):48-51.

687 26. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. The bacterial species challenge: making
688 sense of genetic and ecological diversity. Science. 2009;323(5915):741-6.
689 27. Falush D, Torpdahl M, Didelot X, Conrad DF, Wilson DJ, Achtman M. Mismatch induced
690 speciation in Salmonella: model and data. Philosophical transactions of the Royal Society of London
691 Series B, Biological sciences. 2006;361(1475):2045-53.
692 28. Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, et al. Genomic islands of
693 speciation separate cichlid ecomorphs in an East African crater lake. Science. 2015;350(6267):1493-
694 8.
695 29. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense
696 haplotype data. PLoS genetics. 2012;8(1):e1002453.
697 30. Delcher AL, Salzberg SL, Phillippy AM. Using MUMmer to identify similar regions in large
698 sequence sets. Current protocols in bioinformatics. 2003;Chapter 10:Unit 10 3.
699 31. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics.
700 2014;30(14):2068-9.
701 32. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale
702 prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691-3.
703 33. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and
704 annotation of phylogenetic and other trees. Nucleic acids research. 2016;44(W1):W242-5.
705 34. Krzywinski M, Schein JI. Circos: an information aesthetic for comparative genomics. Genome
706 Research. 2009;19(9):1639-45.

707

708 **Figure legends**

709 **Figure 1. Detection of coadaptation loci in *V. parahaemolyticus*.** (a) NJ tree of 198
710 VppAsia strains based on 151,957 bi-allelic SNPs (minor allele frequency > 2%) (b) Q-Q plot
711 of Fisher exact test *P* values between genetic variants. The vertical dotted line shows the
712 threshold $P = 10^{-10}$. (c) Hierarchical clustering of interacting variants (column and row) in the
713 largest network based on Fisher exact test *P* value. IG2-1 and IG2-2 were integrated into IG2.
714 (d) Distribution of SuperDCA coupling strength value. (e) Correlation between Fisher exact
715 test *P* value and SuperDCA coupling strength (red), and between Fisher exact test *P* value
716 and SpydrPick mutual information (blue). (f) Overlap of strong linked SNP sites detected by
717 Fisher exact test ($P < 10^{-10}$, excluding IG1 variants) and SuperDCA (coupling strength > $10^{-2.2}$).
718 Red for interacted SNP pairs detected by both methods, blue for SNP pairs detected only
719 by Fisher exact test. Only SNP/accessory gene pairs with distance > 3 kb were included in
720 (c,e and f).

721

722 **Figure 2. Landscape of coadaptation (excluding IG1-3).** (a) Gene maps of different IGs.
723 Arrows indicate genes and red for detected coadaptation core genes, blue for accessory genes
724 and orange for genes with no coadaptation signal. Black vertical lines indicate SNPs in the
725 interaction group. The colors of the bar on the left indicates average linkage strength of the
726 loci in each IG. Vertical dotted lines were used to split compatible genes with physical
727 distance larger than 3 kb, or genes located in different contigs, chromosomes and strains.
728 Dotted rectangles indicate incompatible genes. IGs with genome block length larger than 60
729 kb are broken by double slash and shown in (b) after zooming out. COG classification labels
730 are shown above the genes. (c) COG classification and GC content of all the genes used in
731 detection (top) and of different types of coadaptation genes. Red for core genes and blue for
732 accessory genes. The first number in brackets is the number of genes with COG annotation
733 and the second is the total number of genes in the category.
734
735 **Figure 3. Representative interaction groups.** Hierarchical clustering of 469 non-redundant
736 strains (columns) based on coadaptation loci (rows) of 4 representative IGs. Colors of the
737 heatmap indicate the status of genetic variants, with light orange/orange for two alleles of a
738 SNP, light yellow/brown for absence and presence of the accessory genes. Bar colors below
739 the tree on the top indicate the populations of strains according to the legend. Function
740 summary of involved genes is shown on the top of each heatmap. Arcs on the right indicate
741 the causal links after ARACNE flitering, colors and the width of the arcs scale with the *P*
742 values.
743
744 **Figure 4. Interaction group 1.** (a) Hierarchical clustering of 469 non-redundant strains
745 (columns) based on coadaptated loci (rows) of IG1. Color scheme of the heatmap is the same
746 as in Figure 3. Branch colors of the tree on the top indicate the populations of strains

747 according to the legend. The black bars below the tree indicate the strains used in
748 experiments. Branch colors of the tree on the left indicate the tiers of coadaptation loci, with
749 red, blue and green for tier 1, 2 and 3, respectively. The same colors are used for tiers of
750 coadaptation in panel (b) and (c). Colors of the bar on the right indicate the genome position
751 of coadaptation loci, which is corresponding to the bar colors in panel (b). Arcs on the right
752 indicate the causal links as in Figure 3, the links within the genome block in (b) were
753 removed and only one link between different blocks was shown. (b) Gene map of
754 coadaptation genome blocks of IG1. Two reference genomes were used to show coadaptation
755 variants in IG1. The left curve indicates the value of average copy probability of EG1.2
756 strains that copied from themselves throughout the genome of Ref1, with vertical dotted line
757 indicates the expect value (number of EG1.2 strains /number of all strains). The bars indicate
758 the reference genomes and different chromosomes are separated by a horizontal short bar.
759 The labels of coadaptation genome blocks are shown above them and are corresponding to
760 information in Table S2. Arrows and vertical lines separately indicate genes and SNPs, which
761 were colored according to different coadaptation tiers, and core genes were colored grey and
762 genes with no coadaptation signal were light orange. COG classification labels are shown
763 above the genes. (c) The distribution of coadaptation SNPs in the lateral flagellar gene cluster
764 region (VPA1538-1557). The top indicates the gene organization of lateral flagellar gene
765 cluster. Light orange rectangles show the accessory genome region. The histograms indicate
766 the distribution of SNPs along the gene cluster, with colors of bars indicate coadaptation tiers.

767

768 **Figure 5. Interaction group 2 (a) and 3 (b).** Layout and colors are the same as in Figure 4.
769 EG1.2 and EG2.1 strains are shown below the tree on the top for comparison (black bars).
770 The left curves indicate the average copy probability value of EG2.1 (a) and EG3.1 (b) strains
771 copy from themselves throughout the reference genome, respectively. Twenty randomly

772 selected SNPs separately from IG2-1 and IG2-2 were used in hierarchical clustering to
773 minimize the influence of number of variants.

774

775 **Figure 6. Phenotypes of 7 EG1.1 and 4 EG1.2 strains.** (a) Swimming and swarming
776 ability. (b) Growth curve. (c) Biofilm formation and colony morphology. Strains used in
777 experiments were randomly selected and were marked in Figure 4a.

778

779 **Figure 7. Overview of four stages of coadaptation.** Circles indicate bacterial strains within
780 a population. Stars indicate SNPs, with red and green indicating the two alleles. Blue
781 rectangles indicate accessory genes or genome islands. Arrows indicate the transitions
782 between stages. Casual interactions involve genes and SNPs coming and going on all genetic
783 backgrounds. Steady interactions involve particular genes that are associated with each other
784 but with frequent exceptions due to ongoing genetic flux and coadaptation with other loci.
785 Married interactions involve a core of fastidiously associated loci with other loci that lead to
786 further co-adaptation in multiple genome regions. The horizontal line in the fourth stage
787 indicates a barrier to gene flow entailing speciation.

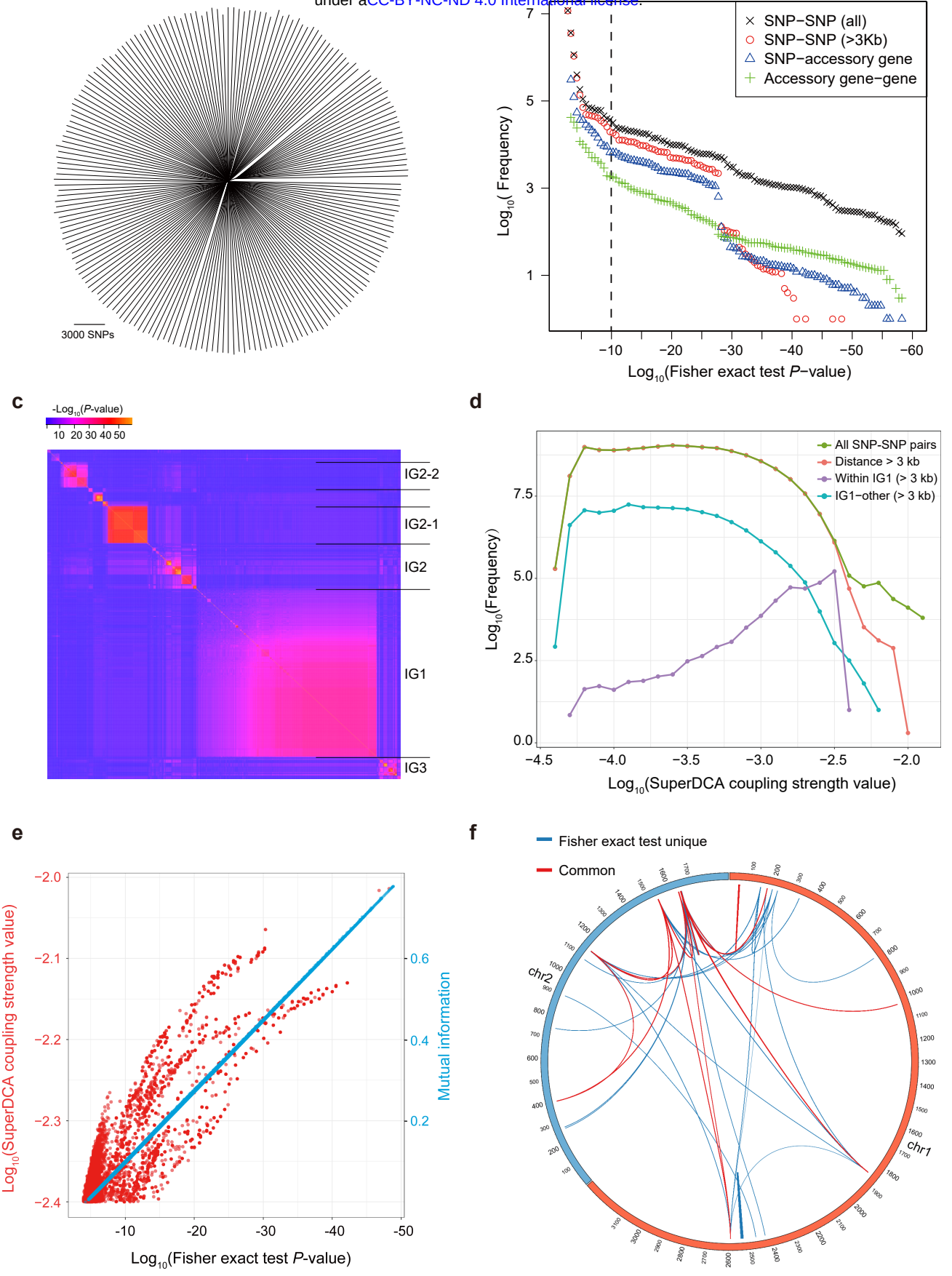
Table1. Summary of interactions detected in coadaptation screen.

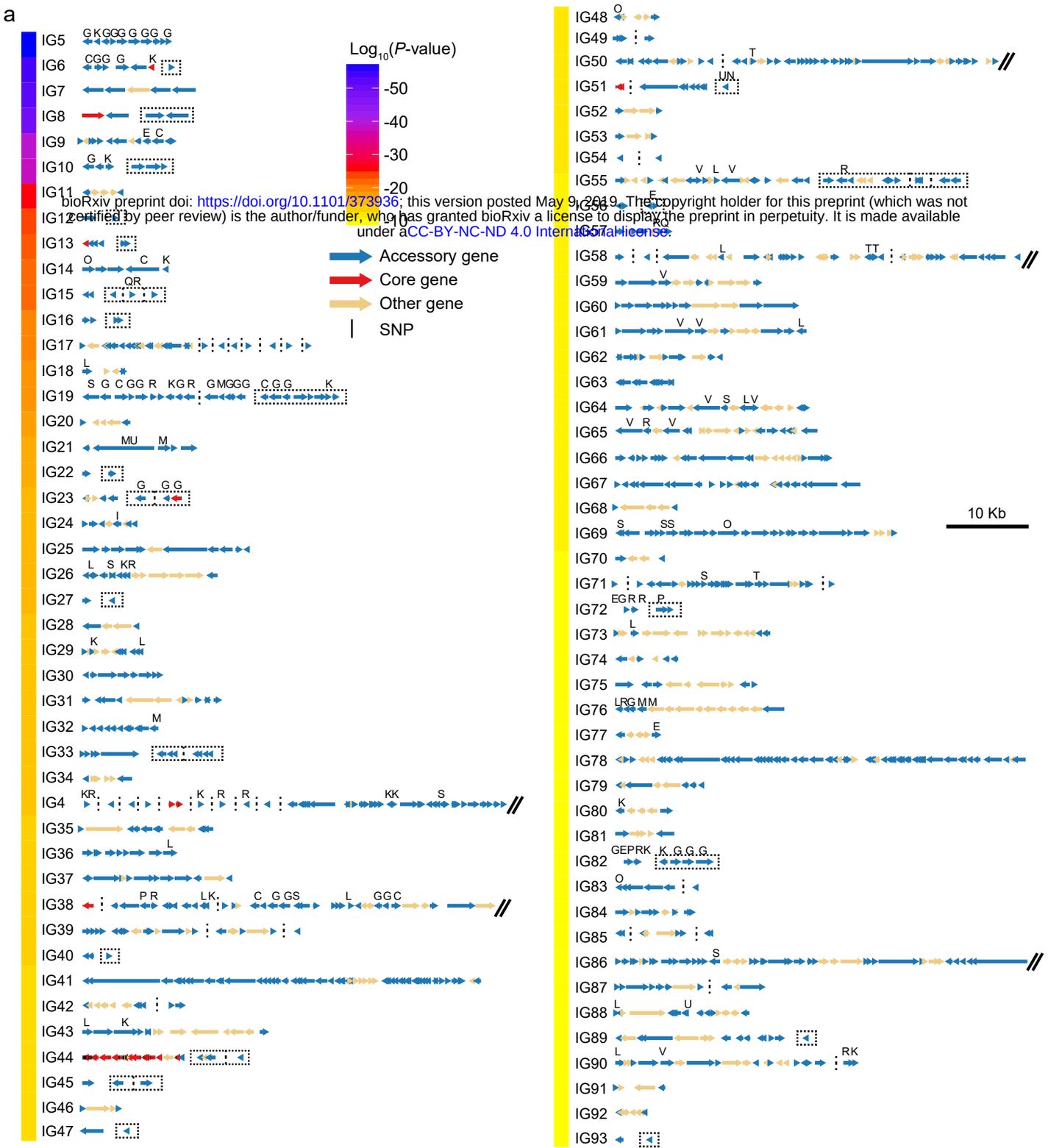
	Total*	Network 1					IG5-93	Frequency***
		IG1	IG2	IG3	IG4	Frequency**		
SNP-SNP pair	2.3×10^{10}	283632	4398	1150	6	0.00%	22751	0.00%
SNP-Accessory gene pair	2.2×10^{9}	104096	7319	2446	112	0.01%	1188	0.00%
Accessory gene-gene pair	2.1×10^{8}	7999	2823	973	1692	0.01%	12264	0.01%
SNP number	151957	917	548	75	63	1.05%	333	0.22%
Synonymous (Syn)	117541	626	409	49	38	0.95%	226	0.19%
Nonsynonymous (NonSyn)	23673	236	122	21	21	1.69%	107	0.45%
NonSyn/Syn	0.2	0.38	0.30	0.43	0.55		0.47	
Core gene	3936	62	20	6	4	2.34%	18	0.56%
Accessory gene	14486	152	130	65	124	3.25%	1122	7.75%

* All of the variations used in coadaptation screen.

** Ratio of the number of variations in Network 1 to the total number of variations.

*** Ratio of the number of variations in IG5-93 to the total number of variations.





COG classification

A: RNA processing and modification	L: Replication, recombination and repair
B: Chromatin structure and dynamics	M: Cell wall/membrane/envelope biogenesis
C: Energy production and conversion	N: Cell motility
D: Cell cycle control, cell division, chromosome partitioning	O: Posttranslational modification, protein turnover, chaperones
E: Amino acid transport and metabolism	P: Inorganic ion transport and metabolism
F: Nucleotide transport and metabolism	Q: Secondary metabolites biosynthesis, transport and catabolism
G: Carbohydrate transport and metabolism	R: General function prediction only
H: Coenzyme transport and metabolism	S: Function unknown
I: Lipid transport and metabolism	T: Signal transduction mechanisms
J: Translation, ribosomal structure and biogenesis	U: Intracellular trafficking, secretion, and vesicular transport
K: Transcription	V: Defense mechanisms