The landscape of coadaptation in *Vibrio parahaemolyticus*

Yujun Cui[1*#], Chao Yang[1,2#], Hongling Qiu[3#], Hui Wang[3,4], Ruifu Yang[1], Daniel Falush[5*]

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

2 Shenzhen Centre for Disease Control and Prevention, Shenzhen, China;

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

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

5 Center for Microbes, Development and Health, Institute Pasteur of Shanghai, Chinese Academy of Sciences

[#]These authors contributed equally to the article

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

Abstract

Investigating fitness interactions in natural populations remains a considerable challenge. We take advantage of the unique population structure of *Vibrio parahaemolyticus*, a bacterial pathogen of humans and shrimp, to perform a genome-wide screen for coadapted genetic elements. We identified 90 interaction groups involving 1,560 coding genes. 82 of these interaction groups are between accessory genes, many of which have functions related to carbohydrate transport and metabolism. Only 8 interaction groups involve both core and accessory genomes. The largest includes 1,540 SNPs in 82 genes and 338 accessory genome elements, many involved in lateral flagella and cell wall biogenesis. The interactions have a complex hierarchical structure encoding at least four distinct ecological strategies. Preliminary experiments imply that the strategies influence biofilm formation and bacterial growth rate in vitro. One strategy involves a divergent profile in multiple genome regions, implying that strains have irreversibly specialized, while the others involve fewer genes and are more plastic. Our results imply that most genetic alliances are ephemeral but that increasingly complex strategies can evolve and eventually cause speciation.

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 transmission genetics, i.e. how genetic material is passed from one generation to the next, and population structure. For example, in outbreeding animals, mating mixes up variation every 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 because recombination happens at lower rate and therefore breaks up coadapted loci more slowly. 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. 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. Consequently, 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).

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 (10). As a result, the population of *V. parahaemolyticus* 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 a complex multi-locus interaction, in which core and accessory genomes have evolved in

parallel to create coadapted gene complexes encoding at least four distinct strategies. We show that it is possible to characterize these strategies, both at the level of strains and at the level of the genetic variants encoding them, using hierarchical clustering. 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.

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 (8) 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 35 of which included multiple genome regions. In total these networks included 1,873 SNPs in 100 core genes and 1,460 accessory genome elements (Table 1). Interacting SNPs were substantially enriched for non-synonymous variants, which is consistent with natural selection being the force generating the linkage disequilibrium we detected.

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 SNPs associated with EG1a (see below) 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 (Figure 1c, Figure S2). This discrepancy is due the large number of SNPs involved with similar association patterns, which means that coupling values are distributed between them. Excluding EG1a SNPs, there is a strong correlation between SuperDCA coupling strengths and Fisher exact test *P* values (Figure 1c). 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 1d). 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 1c).

The largest interaction group 1 (IG1) accounted for the majority of interacting SNPs (82%) as well as a significant fraction of accessory genome elements (23%), while IG2-IG90 generally consisted of a small number of interacting SNPs (0 - 297) and accessory genes (2 - 128). IG2-IG90 are displayed in Figure 2, 3, Table S1 and Figure S3, while IG1 is shown separately in Figures 4 and Table S2. For IG1, 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, with IGs having higher mean values, especially in compatibility IGs (Figure 2c).

Organization 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, and Figure S3). We also used the criterion used by ARACNE and SpydrPick (15, 16) 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 (IG50) and 57 kb (IG38), of which there are 52 (Figure 2). For example, IG2 consists of 10 genes in a single block (VPA1700-VPA1709). Nine of the genes code for various functions related to carbohydrate metabolism and transport while the 10th is a transcriptional regulator (Figure 3 and Table S1). 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 (17). In our data, one example is IG14 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 23 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 IG14 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 or is transmitted by other mechanisms.

Another common interaction is incompatibility between different accessory genome elements, which is found in total 24 of the interaction groups (Figure 2). For example, in IG3 (Figure 3), 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 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 IG42 (Figure 3), where there are three genes (polar flagellar sheath protein and hypothetical proteins) that are mutually incompatible in our data, and 12% strains have none of the three genes.

Detailed characterization of the largest interaction group

IG1 involves a large number of pairwise interactions, presenting a challenge for interpretation. We filtered the results to remove putatively-non-causal interactions using ARACNE (15), but this only reduced the number of interactions from 414,785 to 103,241, which is still far too many to interpret (Table S2). However, hierarchical clustering revealed that strains fall into four distinct "ecogroups" (EGs) based on IG1 variants (Figure 4a, b). Three of these groups, EG1a, EG1b and EG1d have a large number of variants (50-965) that are more-or-less strongly associated with them, and we used the clustering to sort these variants into tiers, using with "Tier 1" (T1) corresponding to the variants that showed the strongest association for each EG (Figure 4, Table 2, Table S2). Associated variants also show other patterns, for example, "Other 1" (O-1) variants revealed consistent pattern in EG1a and EG1b but distinct with other ecogroups; O-2 variants are largely consistent in EG1a and EG1d but different with EG1b and EG1c; O-3 variants are polymorphic in EG1b and EG1c but mostly fixed in EG1a and EG1d (Figure 4a, b).

27% (509/1879) of the IG1 SNPs localize to the lateral flagella gene cluster (VPA1538-VPA1557). The variants are scattered widely in the heatmap in many different configurations, with SNPs from the cluster showing strong associations with EG1a and EG1b and other patterns (Figure 4a, Table S2). The lateral flagellar locus is responsible for motility on surfaces and the large number of different configurations, most involving multiple non-synonymous SNPs, suggests that adaptation at the locus has resulted several motility phenotypes, which are biologically distinct enough to have driven coadaptation in many other genes. A phylogenetic analysis of variation at the lateral flagellar locus incorporating data from *Vibrio* genus shows that the locus has been inherited vertically within the species but with a substantially elevated rate of evolution in the EG1a version of the locus (Figure S4), which suggests that this version of the cluster has evolved a substantially new function.

EG1a strains are highly diverged compared to other EGs in the number of core SNPs (831 SNPs in 47 genes) and accessory genes (134 genes). There are 114 fixed or nearly fixed non-synonymous differences (a-T1 and a-T2) at the lateral flagellar gene cluster, implying a substantial and strongly selected divergence of biological function at the locus. Additionally, EG1a strain has a particularly complex polymorphic genomic block (Ref2-13, 82 kb, Figure 4d) with substantial variation in gene content between strains, which in total constitutes for a large fraction of the gene content difference (60 genes). Among this genomic block, 14 genes were cell wall biosynthesis genes (COG M) and 5 genes are annotated as having Type II secretion related function.

EG1b is also associated with a large block, containing 34 accessory genes (Ref2-04), as well as a smaller number of lateral flagellar gene cluster SNPs and only one further accessory gene and SNPs in a single gene, 12 of the accessory genes are cell wall biosynthesis genes (COG M). EG1c has no clear defining features, there are 7 genes (Tier O-4, Table S1) that are found at high frequency (67%) in EG1c while being rare elsewhere (5%), but none come close to representing a fixed difference. Four of these genes are also annotated as being transferases.

Tier O-1 includes 16 SNPs in lateral flagellar genes (VPA1548 and VPA1550) and 20 SNPs in the gene *TonB* (VP0163), which was associated with the energy transduction (18). These SNPs are strongly associated, despite being on different chromosomes, suggesting a specific functional interaction. Tier O-2 includes 18 SNPs in two core genes encoding *LuxR* family transcriptional regulators (VPA1446-VPA1447), which is associated with c-di-GMP signaling and biofilm formation (19). Within Tier O-2, ARACNE identified an accessory

gene group_3560 as being a driver locus, with a particularly large number of causal interactions. This gene codes for a polysaccharide biosynthesis/export protein and retains interactions with 9 genome blocks after filtering (Table S1). Tier O-3 contains SNPs that are at high frequency in strains containing the group_3560 gene but are rare in the rest of the dataset. The SNPs are in four genes coding multidrug resistance protein (VP0038), hypothetical protein (VP0039), lipase (VPA0859) and long-chain fatty acid transport protein (VPA0860) respectively.

EG1d was identified in our previous analysis (11). In the current dataset, it is differentiated from other strains at 5 SNPs in a transmembrane gene (VPA1081) and 45 accessory genes. A 23 gene block encoding T6SS (type VI secretion system, VP1391-VP1420, Ref1-10, d-T1) is present or partly present in all EG1d strains and absent or largely absent the other three EGs. Most EG1a-c strains instead encode a block of genes that are annotated as being cellulose synthesis related (Ref2-01, d-T2). It is notable, however, that there are no SNPs associated specifically with EG1d in the lateral flagellar genome region.

Chromosome painting (Figure 4d) shows evidence for sharp peaks of differentiation of different EGs around the IG1 loci identified by our coadaptation scan, especially for EG1a strains, but with little evidence for differentiation elsewhere. This pattern is qualitatively distinct either for that observed between geographically differentiated populations or for clonally related strains, which show higher-than-expected copying probability throughout the genome but without sharp peaks (Figure S5). These results are consistent with there being a common gene pool shared by all of the ecogroups within each *V. parahaemolyticus* population, with ecogroup structure being maintained by natural selection rather than barriers to gene flow.

Phenotypic differences of EG1a strains

We performed a preliminary investigation of the phenotypic differences underlying EG1a by determining the motility, growth rate, and biofilm formation ability (Figure 5) of in 11 strains (4 EG1a, 1 EG1b, 2 EG1c, 4 EG1d) on laboratory media. Because there are many unique variations in the lateral flagella gene cluster of EG1a isolates, we expected a different motility ability between EG1a and other isolates, but we failed to observe differences in swimming or swarming capability under the conditions tested (Figure 5a). However, EG1a strains revealed faster growth rate and significantly higher biofilm formation ability than EG1b-d strains (Figure 5b, c), and they revealed rough colony morphology, also an indication of increased biofilm formation, under low salinity (1% NaCl) culture condition (Figure 5c).

There are in total 60 EG1a strains in the global collection of 1,103 *V. parahaemolyticus* strains. All but one, a VppUS2 isolate, is from the VppAsia population, with the majority (n= 48, 80%) in this study coming from routine surveillance on food related environmental samples, including fish, shellfish, and water used for aquaculture. The strains revealed no clear geographical clustering pattern in China, as they can be isolated from all six provinces that under surveillance. Notably, only 4 EG1a strains were isolated from clinical samples, including wound and stool, representing a lower proportion than for EG1b-d isolates (453/1043), including if the two major pathogenic clonal lineages (CG1 and CG2) (8) are removed (268/798), suggesting this eco-group has low virulence potential in humans.

Discussion

Bacterial traits such as pathogenicity, host-specificity and antimicrobial resistance naturally attract human attention, but less obvious or even cryptic traits might be more important in

determining the underlying structure of microbial populations. Studies of coadaptation based on genome sequencing of thousands of isolates have the potential to provide new insight into the ecological forces shaping natural diversity, and how that variation is assembled by individual strains to overcome the manifold challenges involved in colonizing specific niches, such as the human gastrointestinal tract. In other words, these studies provide a unique opportunity to see the world from the point of view of a bacterium.

We performed a genome wide scan for coadaptation in *V. parahaemolyticus*, performing pairwise tests for interactions amongst genetic variants and then clustered the significant pairwise interactions into 90 interaction groups. Our analysis demonstrated that genome wide epistasis scans can be used successfully to identify diverse interactions involving both core 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 the SNPs associated with EG1a. 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 commitment within human relationships, namely casual, going steady, getting married and moving out together (Figure 6).

Most interactions between core and accessory genomes are casual

A recent debate about whether the accessory genome evolves neutrally (20-22) highlighted how little we know about the functional importance of much of the DNA in bacterial chromosomes. Using the same statistical threshold to assess significance, our interaction screen identifies many more examples of coadaptation between different accessory genome elements than of interactions between the core and accessory genomes or within the core genome, implying that natural selection has a central role in determining accessory genome composition.

Unsurprisingly, given the extensive literature highlighting the importance of genomic islands to functional diversity of bacteria (17), the most common form of adaptation detected by our screen is the coinheritance of accessory genome elements located in the same region of the genome. Based on a minimum size threshold of 3 kb, we find 52 (56%) such interactions, the largest of which is 57 kb (Table S1).

Previous approaches to detecting genome islands emphasize traits associated with horizontal transmission, for example based on differences in GC content with the core genome, the presence of phage-related genes or other markers of frequent horizontal transfer (23). Our approach, based simply on co-occurrence identifies a wider range of coinherited units and suggests that many islands have functions related to carbohydrate metabolism.

Amongst interactions not involving physical linkage, the most common is incompatibility of different accessory genome elements, representing 27% (24/90) of our IGs. For example, two different versions of a phosphorous-related pathway, one involving one gene, the other involving 5 (IG3, 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 8 different interaction groups involving core and accessory genome regions. One simple example included a regulatory gene VP0368 and an accessory genome element (IG3, 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. The difficulty of evolving the trait de novo is likely to slow down the rate at which it is gained and loss, which in turn makes further coadaptation at additional genes more likely.

IG1d, is an example of a complex coadaptation involving multiple core and accessory genome regions. A large majority of strains in our dataset (439/469, EG1d) either carry a cluster of genes encoding a T6SS, or a cluster encoding cellulose biosynthesis genes (EG1a-c), but few strains have genes from both clusters (Figure 4a). Cells uses the T6SS to inject toxins into nearby bacteria (24) and cellulose production to coat themselves in a protective layer (25). 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. Strains that do not have the T6SS use at least three distinct sets of strategies, corresponding to ecogroups EG1a, EG1b and EG1c, encoded by alleles that are rare or absent in EG1d strains.

Marriage changes everything

EG1a variants differ from the other interaction groups in our coadaptation screen in both the number of associated regions and the strength of the associations. The interaction group include 454 SNPs in the lateral flagellar gene cluster (VPA1538-1557, 18 kb, Figure 4c), a further 917 core genome SNPs in 62 genes and 152 accessory genes in 35 clusters. Many of the variants represent fixed or nearly fixed differences between EG1a and other strains (a-T1 and a-T2). These include loci encoding flagellar genes, T2SS and other membrane transport elements. There are also 285 loci (27%, a-T3) in weaker disequilibrium, typically because they are polymorphic in EG1a or the other EGs. Many of variants are likely to represent more recently evolved coadaptation. Some of these genes are also associated with flagella or the T2SS related function but also encompass a broader range of functional categories, including cell division and amino acid transport and metabolism (Table S2).

Our laboratory phenotype experiments (Figure 5) suggest that biofilm formation is likely to be a key trait underlying the different ecological strategies of EG1a and other EGs. However, the variation in phenotypic response at different salinity levels, and the absence of measurable difference in swarming behavior despite the large genetic difference within the lateral flagella genes, highlight some of the manifold difficulties of interpreting natural variation using phenotypes measured under laboratory conditions.

Despite the extensive differences that have accumulated between EG1a strains and the other EGs, there is no evidence of restricted gene flow in most of the genome (Figure 4d), and even within the flagellar gene cluster, strongly differentiated regions are separated by a weakly differentiated one (Figure 4c), implying that the coadaptation is being maintained by

selection in the face of frequent recombination. Initial divergence in flagellar function is likely to have led to ecological differentiation, which led to bacteria having different nutritional inputs or requirements and a broadening of the functional categories undergoing divergent selection.

How can the difference between EG1a variants and the other interaction groups in both the number of associations and their strength be explained? *V. parahaemolyticus* is ubiquitous in shellfish in warm coastal waters, within which it occurs at densities of around 1,000 cells per gram, so a back of the envelope calculation suggests there are likely to be substantially more than 10^{15} bacteria in the VppAsia population. The species also has a high estimated effective population size (10, 11) and has strong codon bias, which is often argued to be evidence that even tiny selective coefficients can drive adaptation (26). Furthermore, recombination only breaks up linkage disequilibrium between loci slowly. Therefore, weaker and more variable patterns of association found for associations other than EG1a variants is unlikely to be a simple consequence of the ineffectiveness of selection and is instead likely to reflect complexity in the fitness landscape.

Strains gain flexibility by being able to switch between or modulate genomically encoded strategies by homologous recombination. Crucially, the evolution of promiscuity is self-reinforcing because the presence of strains using multiple strategies in the population also favors the presence of accessory genes and core gene haplotypes that have high or intermediate fitness on a wide range of different genetic backgrounds.

On the other hand, an absence of intermediate genotypes in the population can favor the evolution of fastidiousness, with particular accessory genes and haplotypes becoming

essential components of some genetic backgrounds but deleterious on others. A likely scenario is that EG1a became sufficiently differentiated at the lateral flagellar gene to make recombinants between the two versions of the gene inviable and also created divergent selection at a handful of other loci that was largely independent of the external environment or of interactions with other genes. The evolution of fastidiousness, like the evolution of promiscuity, can be self-reinforcing, and might have led to progressive increase in the differentiation of EG1a strains from the remaining ones until the coadaptation of the loci to each other became more-or-less irreversible, like marriage in England prior to the reign of King Henry VIII.

Coadapted gene complexes as speciation triggers

Running the tape forward, it is easy to envisage the number of coadapted regions of the genome within EG1a undergoing progressive enlargement, until the entire genome becomes differentiated. As coadapted regions become more numerous, the proportion of recombination events between eco-groups that are maladaptive will increase, which might prompt the evolution of mechanistic barriers to genetic exchange between them.

Mechanisms by which new bacterial species arise are frequently discussed in the literature (27-29) but there is currently little data on how the process unfolds. EG1a isolates is of interest both as an example of an intermediate stage of divergence, prior to speciation, and because it suggests that substantial adaptive divergence between gene pools can precede any barriers to genetic exchange, other than natural selection at the loci involved. This – unique to our knowledge – example is exciting because the distinct signature of selection should make it possible to dissect the genetic basis of coadaptation in unprecedented detail. Broadly similar patterns of differentiation including “genomic islands of speciation” have been

observed for example between ecomorphs of cichlid fishes (30), but the evolution of ecomorphs has been facilitated by fish preferring to mate with similar individuals, which will have also inevitably 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, which also were analyzed in our other study (8). To reduce clonal signals, we firstly made a "non-redundant" dataset of 469 strains, in which no sequence differed by less than 2,000 SNPs in the core genome. They were attributed to 4 populations, VppAsia (383 strains), VppX (43), VppUS1 (18) and VppUS2 (21) based on fineSTRUCTURE result (31). We then focused on VppAsia which has more strains, to generate a genome dataset in which strains represent a freely recombining population. We selected 386 genomes from 469 non-redundant genome dataset, including all the 383 VppAsia genomes and 3 outgroup genomes which were randomly selected from VppX, VppUS1 and VppUS2 population, respectively. These 386 genomes were used in Chromosome painting and fineSTRUCTURE analysis (31) as previously described (11). Initial fineSTRUCTURE result revealed multiple clonal signals still exist, thus we selected one representative genome from each clone, combined them with the remaining genomes and repeated the process. After 14 iterations, we got a final dataset of 201 genomes with no trace of clonal signals, involving 198 VppAsia genomes that were used in further analysis (Figure S1).

The copying probability value of each strain at each SNP was generated by Chromosome painting with "-b" option, and the average copying probability value of a given strain group (e.g. EG1a) at each SNP was used in Figure 4d and Figure S5.

Variation detection, annotation and phylogeny

We re-called SNPs for 198 VppAsia genomes by aligning the assembly against reference genome (RIMD 2210633) using MUMmer (32) as previous described (11, 33). Totally

565,466 bi-allelic SNPs were identified and 151,957 bi-allelic SNPs with minor allele frequency > 2% were used in coadaptation detection. We re-annotated all the assemblies using Prokka (34), and the annotated results were used in Roray (35) to identify the pan-genome and gene presence/absence, totally 41,052 pan-genes were found and 14,486 accessory genes (present in > 2% and < 98% strains) were used in coadaptation detection. The pan-gene protein sequences of Roary were used to BLAST (BLASTP) against COG and KEGG database to get further annotation.

The Neighbour-joining trees were built by using the TreeBest software (http://treesoft.sourceforge.net/treebest.shtml) based on sequences of concatenated SNPs, and were visualized by using online tool iTOL (36).

Detection of coadapted loci

Totally 151,957 bi-allelic SNPs and 14,486 accessory genes identified from 198 independent VppAsia genomes were used in coadaptation detection by three methods. Firstly, we used Fisher exact test to detect the linkage disequilibrium of each SNP-SNP, SNP-accessory gene, and accessory gene-gene pair. Presence or absence of an accessory gene was considered as its two alleles. Each variant locus (SNP or accessory gene) has two alleles, major and minor, of which major represents the allele shared by majority of isolates. For each pair of loci X and Y, the number of combinations between X_{major}-Y_{major}, X_{major}-Y_{minor}, X_{minor}-Y_{major}, X_{minor}-Y_{minor} were separately counted and used in the contingency table to calculate the Fisher exact test *P* value. It took 3 days to finish all the coadaptation detection in a computer cluster using 21 cores and 2 Gb memory. We also used SuperDCA (13) and SpydrPick (16) to detect the coadaptation between SNPs, using the same subset of 198 strains to make the analysis as comparable as possible. SuperDCA is based on direct coupling analysis (DCA) model (12)

and has a much faster calculation speed compared with previous DCA methods. However, it still took 25 days to finish the detection by using 32 cores and 86 Gb memory. SpydrPick took one hour to finish the calculation by using 32 cores and 1 Gb memory.

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 (37) was used to visualize the networks of coadaptation SNPs in Figure 1d and Figure S2.

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 assemblies 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 EG1a and EG1b-d respectively for comparison. In total 3,000 SNPs were identified in this region and were used for NJ tree construction.

Determination of phenotypes

Bacteria strains. In the phenotype experiments, totally 11 strains were randomly selected respectively from four EGs that defined by IG1 variants, including 4 EG1a strains (B1_1, B3_1, B5_3, C3_10), 1 EG1b strain (B1_10), 2 EG1c strains (C5_2, C6_5) and 4 EG1d strains (B1_3, B2_10, B4_8, C1_5). The strains stored at -80 °C were inoculated in the thiosulfate citrate bile salts sucrose agar (TCBS) plates by streak plate method. Five clones for each strains were inoculated again in another TCBS plate and then cultured overnight at 30 °C in 3% NaCl-LB broth overnight and used for the following assays.

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

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

Biofilm formation. *V. parahaemolyticus* strains were cultured overnight at 30 °C in 3% NaCl-LB broth. 2 μl of each overnight culture was inoculated to 100 μl of 3% NaCl-LB broth in a 96-well plate and cultured at 30 °C for 24 h statically. The supernatant was discarded and each well was washed once with sterile phosphate-buffered saline (PBS). 0.1% Crystal violet (wt/vol) was added to each well and incubated at room temperature for 30 min. The crystal violet was decanted, and each well was washed once with sterile PBS. Crystal violet that stained biofilm was solubilized with dimethylsulfoxide (DMSO), and then measured at the optical density of 595 nm using Multiskan Spectrum (Thermo Scientific).

Acknowledgements

This work is supported by the National Key Research & Development Program of China (No. 2017YFC1601503 and 2018YFC1603902), National Key Program for Infectious Diseases of China (No. 2018ZX10101003 and 2018ZX10714-002), Sanming Project of Medicine in Shenzhen (No. SZSM201811071), the National Natural Science Foundation of China (No. 31770001) and the Key Research Program of the Chinese Academy of Sciences (No. ZDRW-ZS-2017-1). D.F. is funded by a Medical Research Council Fellowship as part of the MRC CLIMB consortium for microbial bioinformatics (grant number MR/M501608/1).

Author Contributions

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

Conflict of interest

The authors declare that they have no conflict of interest.

References

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. The ISME journal. 2019;13(10):2578-88.
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, Arnold B, MacAlasdair N, Kuronen J, Tonkin-Hill G, et al. Genome-wide epistasis and co-selection study using mutual information. Nucleic acids research. 2019;47(18):e112.
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. 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.
17. Dobrindt U, Hochhut B, Hentschel U, Hacker J. Genomic islands in pathogenic and environmental microorganisms. Nature reviews Microbiology. 2004;2(5):414-24.
18. Kuehl CJ, Crosa JH. The TonB energy transduction systems in Vibrio species. Future microbiology. 2010;5(9):1403-12.
19. Ferreira RB, Chodur DM, Antunes LC, Trimble MJ, McCarter LL. Output targets and transcriptional regulation by a cyclic dimeric GMP-responsive circuit in the Vibrio parahaemolyticus Scr network. Journal of bacteriology. 2012;194(5):914-24.
20. Vos M, Eyre-Walker A. Are pangenomes adaptive or not? Nature microbiology. 2017;2(12):1576.
21. Andreani NA, Hesse E, Vos M. Prokaryote genome fluidity is dependent on effective population size. The ISME journal. 2017;11(7):1719-21.
22. Shapiro BJ. The population genetics of pangenomes. Nature microbiology. 2017;2(12):1574.

23. Langille MG, Hsiao WW, Brinkman FS. Detecting genomic islands using bioinformatics approaches. Nature reviews Microbiology. 2010;8(5):373-82.
24. 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.
25. Tischler AD, Camilli A. Cyclic diguanylate (c-di-GMP) regulates Vibrio cholerae biofilm formation. Molecular microbiology. 2004;53(3):857-69.
26. 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.
27. 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.
28. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. The bacterial species challenge: making sense of genetic and ecological diversity. Science. 2009;323(5915):741-6.
29. Falush D, Torpdahl M, Didelot X, Conrad DF, Wilson DJ, Achtman M. Mismatch induced speciation in Salmonella: model and data. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2006;361(1475):2045-53.
30. Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, et al. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. Science. 2015;350(6267):1493-8.
31. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS genetics. 2012;8(1):e1002453.
32. Delcher AL, Salzberg SL, Phillippy AM. Using MUMmer to identify similar regions in large sequence sets. Current protocols in bioinformatics. 2003;Chapter 10:Unit 10 3.
33. Yang C, Zhang X, Fan H, Li Y, Hu Q, Yang R, et al. Genetic diversity, virulence factors and farm-to-table spread pattern of Vibrio parahaemolyticus food-associated isolates. Food microbiology. 2019;84:103270.
34. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-9.
35. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691-3.
36. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic acids research. 2016;44(W1):W242-5.
37. Krzywinski M, Schein JI. Circos: an information aesthetic for comparative genomics. Genome Research. 2009;19(9):1639-45.

Figure legends

Figure 1. Detection of coadaptation loci in *V. parahaemolyticus*. (a) NJ tree of 198 VppAsia strains based on 151,957 bi-allelic SNPs (minor allele frequency > 2%) (b) Q-Q plot of Fisher exact test *P* values between genetic variants. The vertical dotted line shows the threshold $P = 10^{-10}$. (c) Correlation between Fisher exact test *P* value and SuperDCA coupling strength (red and blue), and between Fisher exact test *P* value and SpydrPick mutual information (green). (d) Overlap of strong linked SNP sites detected by Fisher exact test (*P* <

10^{-10}, excluding EG1a SNPs) and SuperDCA (coupling strength > $10^{-2.2}$). Red for interacted SNP pairs detected by both methods, blue for SNP pairs detected only by Fisher exact test.

Figure 2. Landscape of coadaptation. (a) Gene maps of different IGs. Arrows indicate genes and red for detected coadaptation core genes, blue for accessory genes and orange for genes with no coadaptation signal. Black vertical lines indicate SNPs in the interaction group. The colors of the bar on the left indicates average linkage strength of the loci in each IG. Vertical dotted lines were used to split compatible genes with physical distance larger than 3 kb, or genes located in different contigs, chromosomes and strains. Dotted rectangles indicate incompatible genes. IGs with genome block length larger than 60 kb are broken by double slash and shown in (b) after zooming out. COG classification labels are shown above the genes. (c) COG classification and GC content of all the genes used in detection (top) and of different types of coadaptation genes. Red for core genes and blue for accessory genes. The first number in brackets is the number of genes with COG annotation and the second is the total number of genes in the category.

Figure 3. Representative interaction groups. Hierarchical clustering of 469 non-redundant strains (columns) based on coadaptation loci (rows) of 4 representative IGs. Colors of the heatmap indicate the status of genetic variants, with light orange/orange for two alleles of a SNP, light yellow/brown for absence and presence of the accessory genes. Bar colors below the tree on the top indicate the populations of strains according to the legend. Function summary of involved genes is shown on the top of each heatmap. Arcs on the right indicate the causal links after ARACNE flitering, colors and the width of the arcs scale with the *P* values.

Figure 4. The largest interaction group (IG1) of in *V. parahaemolyticus*. (a, b) Hierarchical clustering of 469 non-redundant strains (columns) based on coadaptated loci (rows) of IG1. Color scheme of the heatmap is the same as in Figure 3. Colors of bars below the tree on the top indicate the populations of strains according to the legend. (c) The distribution of coadaptation SNPs in the lateral flagellar gene cluster region (VPA1538-1557). The top indicates the gene organization of lateral flagellar gene cluster. Light orange rectangles show the accessory genome region. The histograms indicate the distribution of SNPs along the gene cluster, with colors of bars indicate coadaptation tiers. (d) Gene map of coadaptation genome blocks of IG1. Four reference genomes were used to show coadaptation variants. The labels of coadaptation genome blocks are shown above them and are corresponding to information in Table S2. Arrows and vertical lines separately indicate genes and SNPs, which were colored according to different coadaptation tiers, and core genes were colored grey and genes with no coadaptation signal were light orange. COG classification labels are shown above the genes. The bottom left curves indicate the value of average copying probability of different EGs that copied from themselves throughout the genome.

Figure 5. Phenotypes of strains from different EGs. (a) Swimming and swarming ability. (b) Growth curve. (c) Biofilm formation and colony morphology.

Figure 6. Overview of four stages of coadaptation. Circles indicate bacterial strains within a population. Stars indicate SNPs, with red and green indicating the two alleles. Blue rectangles indicate accessory genes or genome islands. Arrows indicate the transitions between stages. Casual interactions involve genes and SNPs coming and going on all genetic backgrounds. Steady interactions involve particular genes that are associated with each other

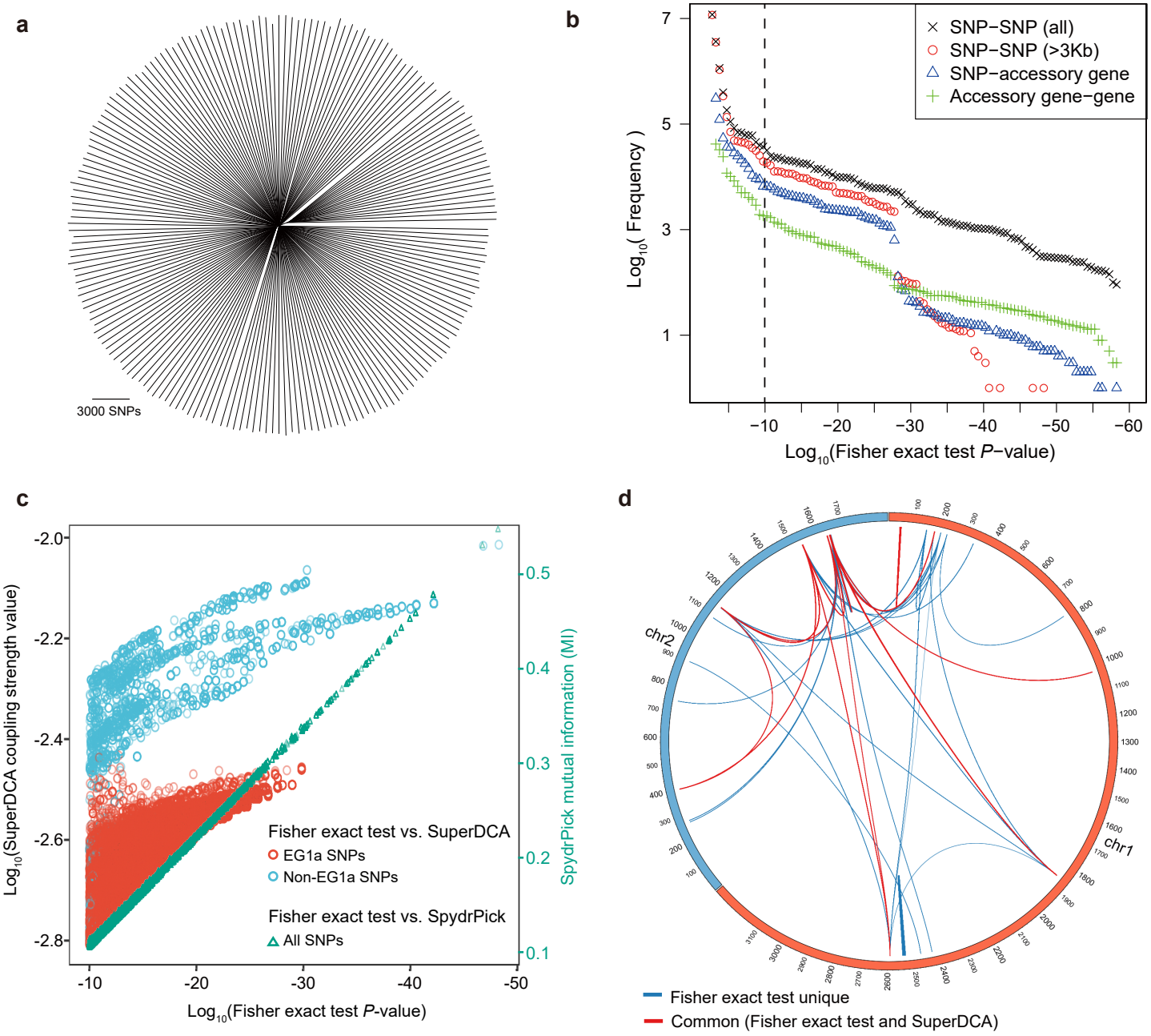
but with frequent exceptions due to ongoing genetic flux and coadaptation with other loci. Married interactions involve a core of fastidiously associated loci with other loci that lead to further co-adaptation in multiple genome regions. The horizontal line in the fourth stage indicates a barrier to gene flow entailing speciation.

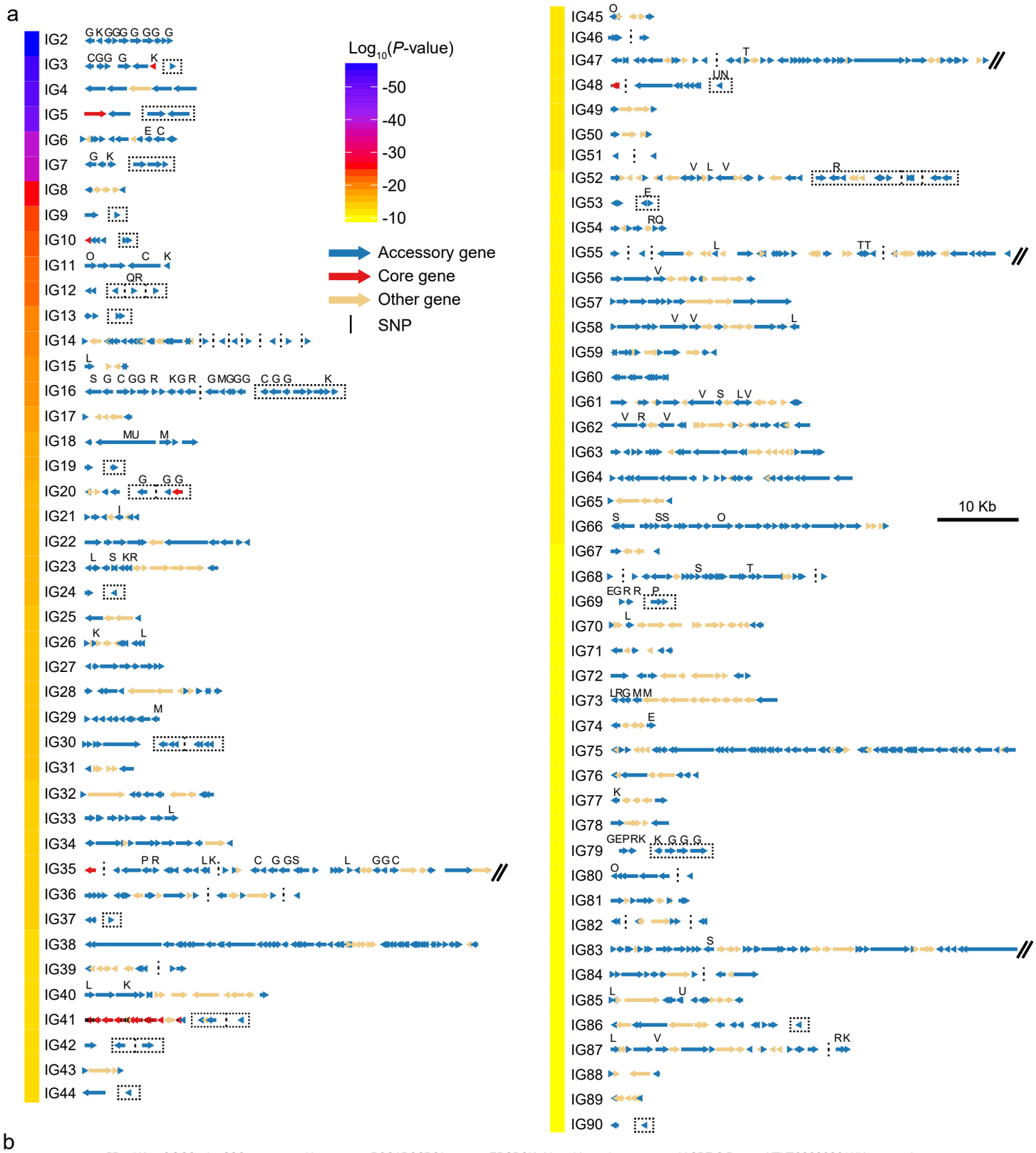
Table1. Summary of interactions detected in coadaptation screen.

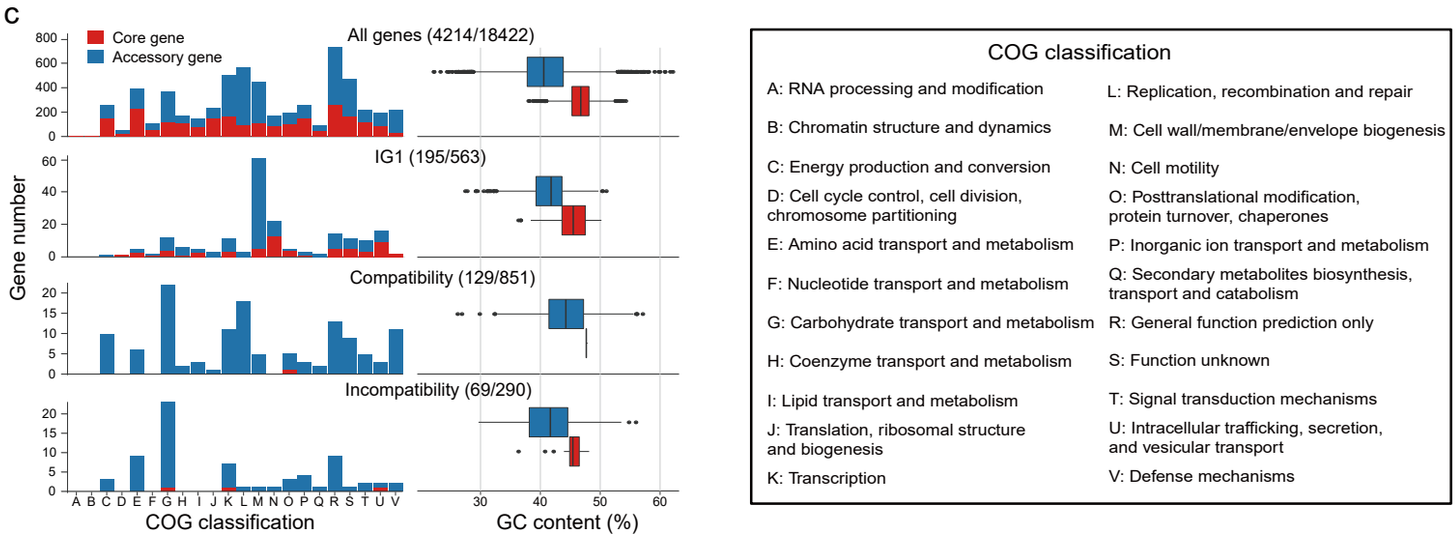
	Total number	IG1		IG2-90	
		Number	Fraction	Number	Fraction
SNP-SNP pair	2.3×10^{10}	289186	0.00%	22751	0.00%
SNP-Accessory gene pair	2.2×10^{9}	113973	0.01%	1188	0.00%
Accessory gene-gene pair	2.1×10^{8}	13487	0.01%	12264	0.01%
SNP number	151957	1540	1.01%	333	0.22%
Synonymous (Syn)	117541	1084	0.92%	226	0.19%
Nonsynonymous (NonSyn)	23673	379	1.60%	107	0.45%
NonSyn/Syn	0.20	0.35		0.47	
Core gene	3936	82	2.25%	18	0.56%
Accessory gene	14486	338	2.33%	1122	7.75%

Table 2. Summary of interaction group 1 variants.

Tier	Core SNPs	Accessory genes
a-T1	520 SNPs (359 Syn, 137 NonSyn) in 21 genes of 8 blocks, 14 genes encoding lateral flagellar	66 genes in 10 blocks, 11 COG M genes, 5 T2SS genes
a-T2	121 SNPs (68 Syn, 43 NonSyn) in 18 genes of 10 blocks, 10 genes encoding lateral flagellar	22 genes in 10 blocks, 2 COG M genes, 3 T2SS genes
a-T3	190 SNPs (137 Syn, 38 NonSyn) in 44 genes of 18 blocks, 17 genes encoding lateral flagellar	46 genes in 13 blocks, 5 COG M genes, 3 COG NU genes, 4 T2SS genes
b-T1	12 SNPs (5 Syn, 2 NonSyn) lin 3 genes of 2 blocks, 2 genes encoding lateral flagellar	33 genes in 2 blocks, 12 COG M genes
b-T2	25 SNPs (15 Syn, 10 NonSyn) in 3 genes of 1 block encoding lateral flagellar	8 genes in 2 blocks, 1 COG M gene
d-T1	5 SNPs (3 syn, 2 nonsyn) in 1 gene encoding transmembrane	31 genes in 5 blocks, 23 genes (1 block) encoding T6SS
d-T2	0 SNP	14 genes in 4 blocks, 8 genes (1 block) encoding cellulose synthase
O-1	36 SNPs (28 Syn, 10 NonSyn) in 3 genes of 2 blocks, 2 genes encoding lateral flagellar, 1 *TonB* gene	6 genes in 4 blocks, 1 COG H gene
O-2	27 SNPs (18 Syn, 1 NonSyn) in 3 genes in 1 block, 2 genes encoding *LuxR* family transcriptional regulator	2 genes in 1 block, COG M &T
O-3	368 SNPs (269 Syn, 97 NonSyn) in 4 genes of 2 blocks, encoding multidrug resistance protein, lipase and long-chain fatty acid transport protein	0 gene
O-4	0 SNPs	7 genes in 5 blocks, 4 genes encoding transferase
Others	236 SNPs (182 Syn, 41 NonSyn) in 48 genes, 8 genes encoding lateral flagellar	102 genes, 15 COG M genes,

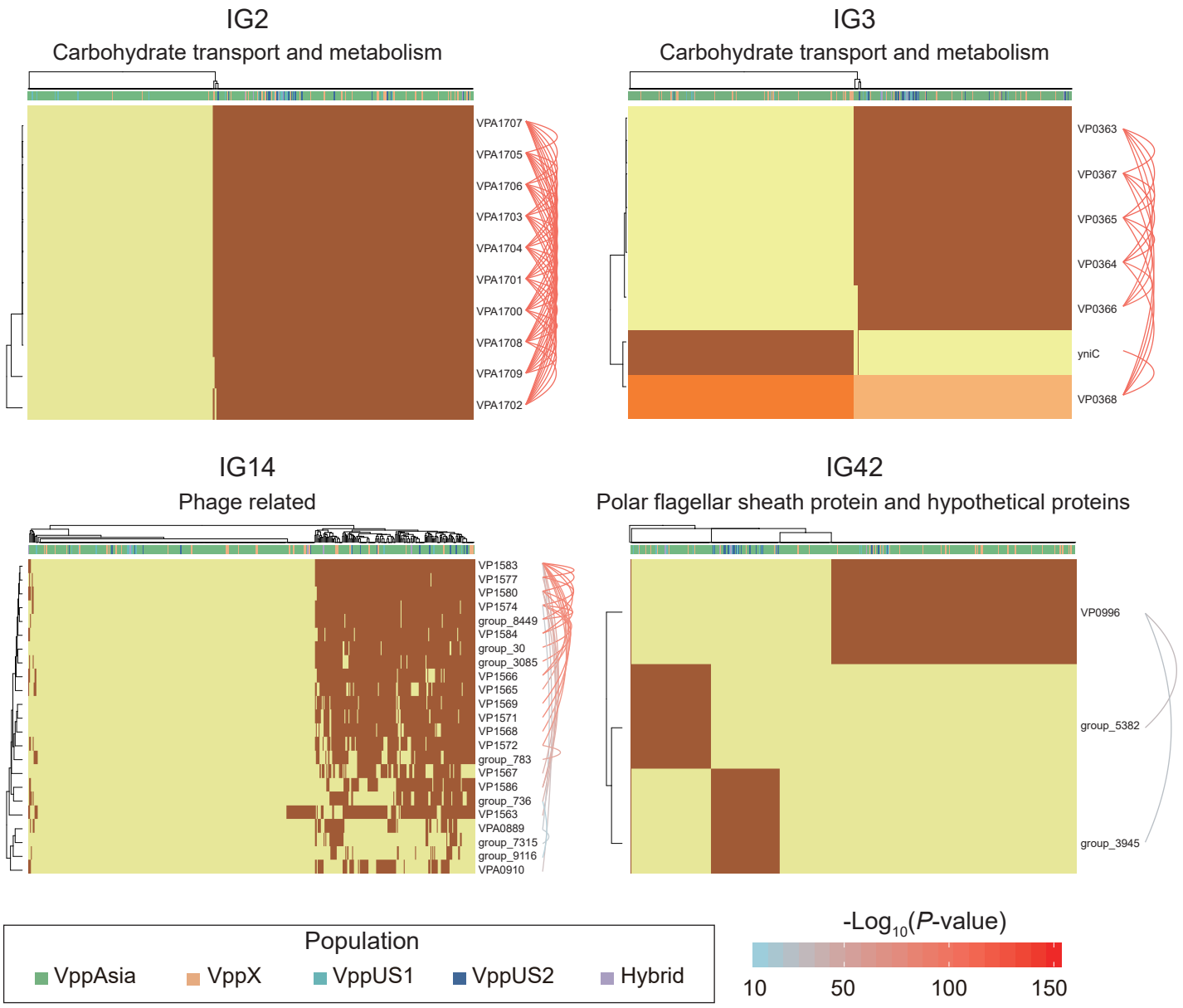

a
3000 SNPs
b
SNP−SNP (all)
SNP−SNP (>3Kb)
SNP−accessory gene
Accessory gene−gene
Log10(Frequency)
Log10(Fisher exact test P−value)
c
Log10(SuperDCA coupling strength value)
SpydrPick mutual information (MI)
Fisher exact test vs. SuperDCA
EG1a SNPs
Non-EG1a SNPs
Fisher exact test vs. SpydrPick
All SNPs
Log10(Fisher exact test P-value)
d
chr1
chr2
Fisher exact test unique
Common (Fisher exact test and SuperDCA)

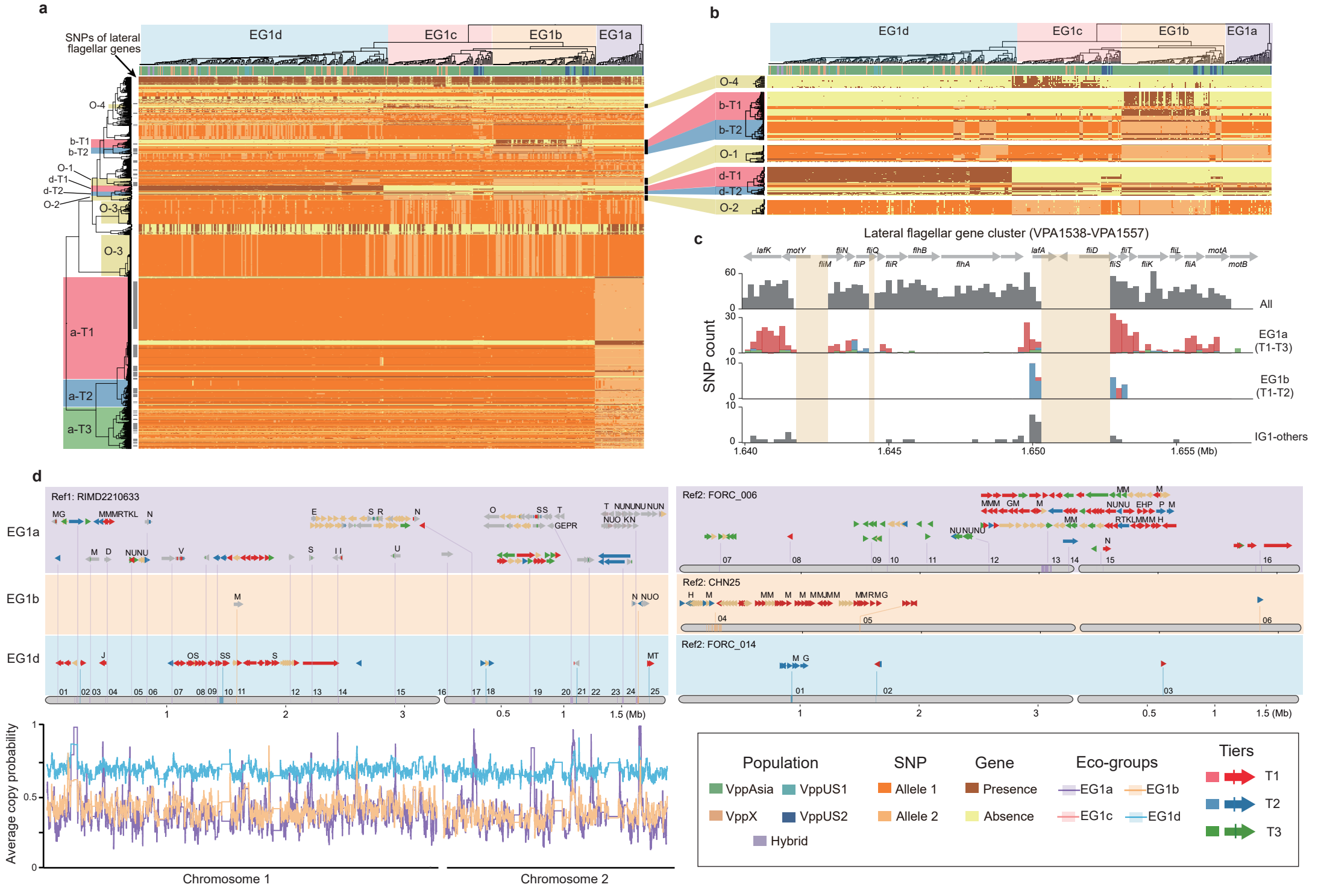


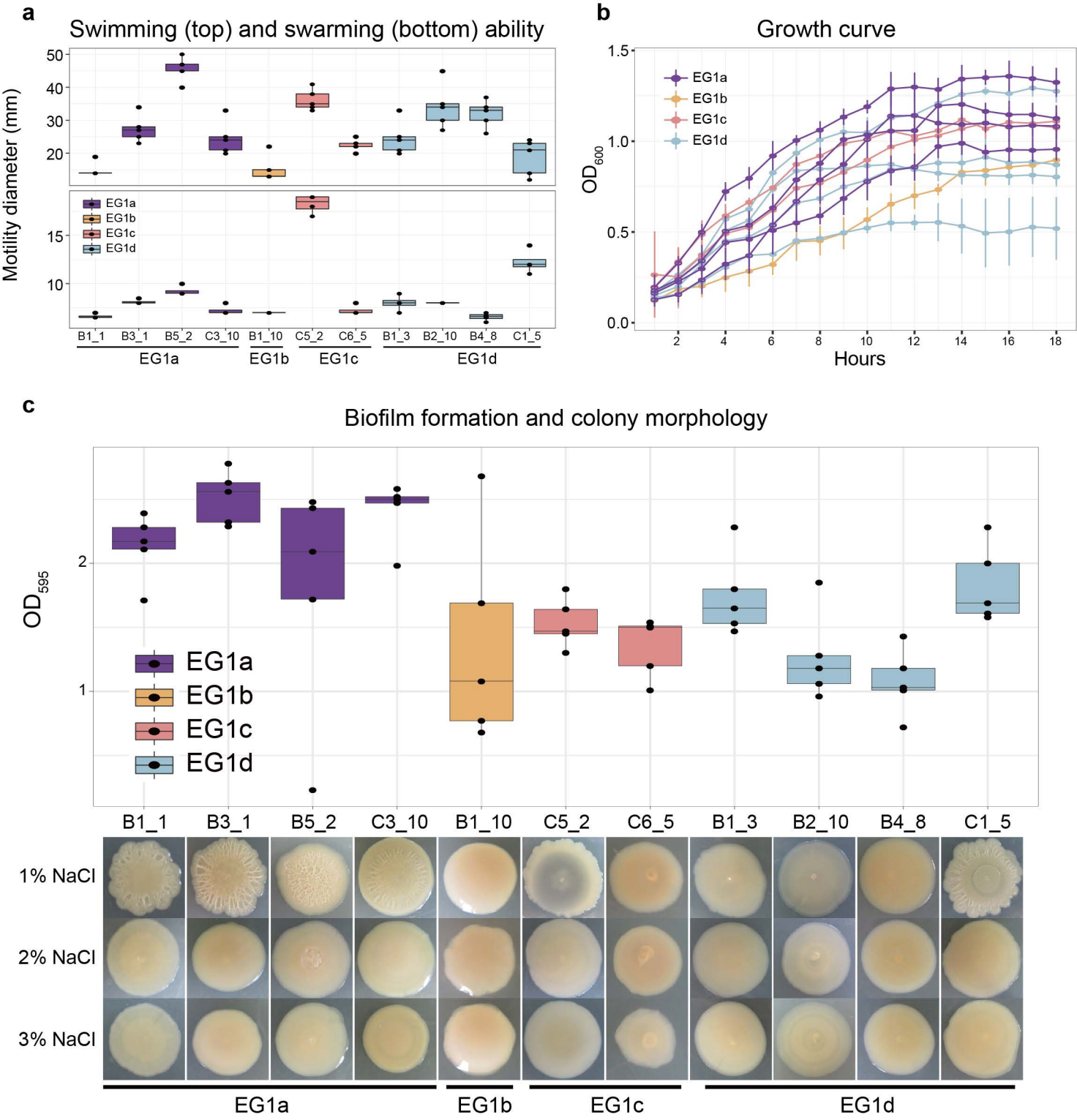


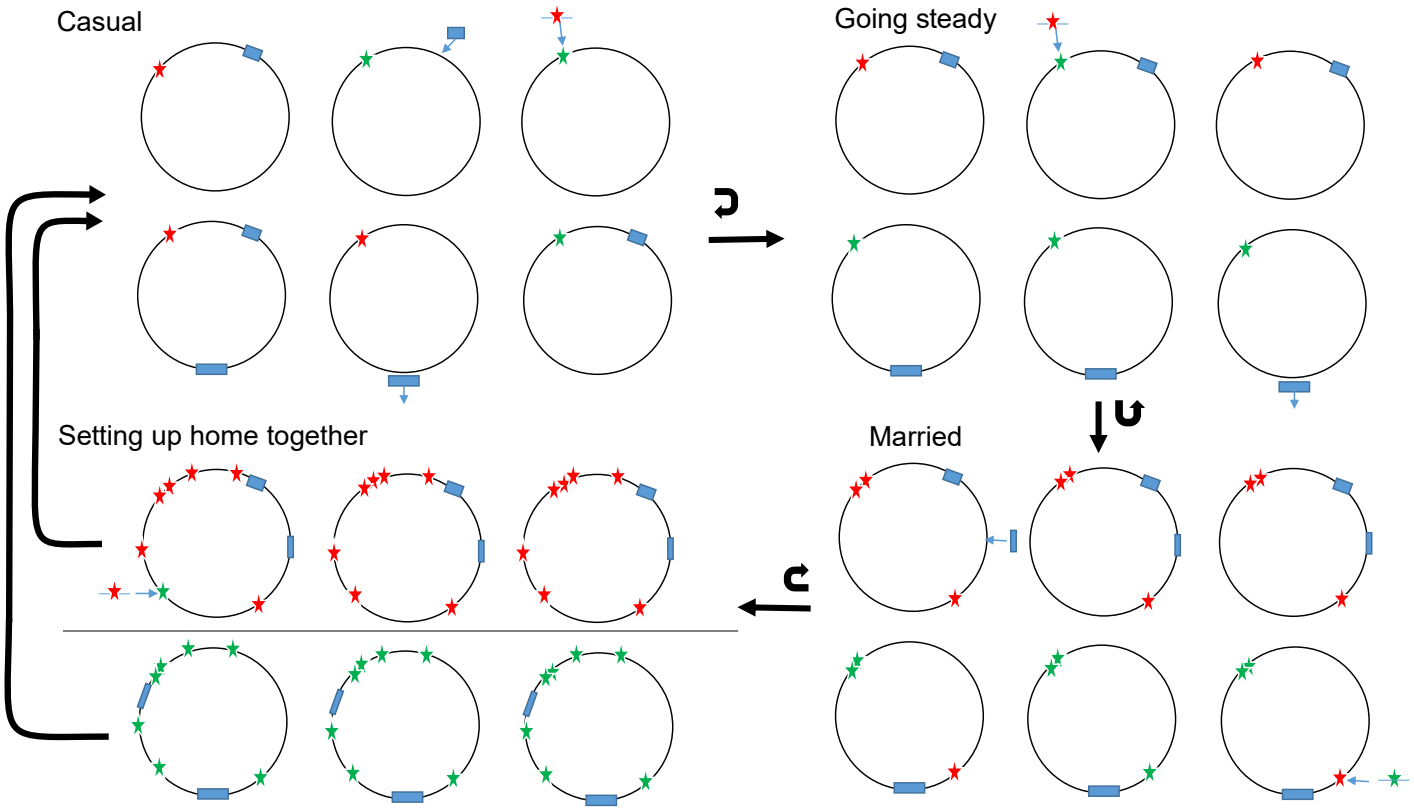
COG classification

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


IG2
Carbohydrate transport and metabolism
VPA1707
VPA1705
VPA1706
VPA1703
VPA1704
VPA1701
VPA1700
VPA1708
VPA1709
VPA1702
IG3
Carbohydrate transport and metabolism
VP0363
VP0367
VP0365
VP0364
VP0366
yniC
VP0368
IG14
Phage related
VP1583
VP1577
VP1580
VP1574
group_8449
VP1584
group_30
group_3085
VP1566
VP1565
VP1569
VP1571
VP1568
VP1572
group_783
VP1567
VP1586
group_736
VP1563
VPA0889
group_7315
group_9116
VPA0910
IG42
Polar flagellar sheath protein and hypothetical proteins
VP0996
group_5382
group_3945
Population
VppAsia
VppX
VppUS1
VppUS2
Hybrid
-Log10(P-value)
10
50
100
150


a
SNPs of lateral flagellar genes
EG1d
EG1c
EG1b
EG1a
O-4
b-T1
b-T2
O-1
d-T1
d-T2
O-2
O-3
a-T1
a-T2
a-T3
b
c
Lateral flagellar gene cluster (VPA1538-VPA1557)
lafK
motY
fliN
fliQ
flhB
lafA
fliD
fliT
fliL
motA
fliM
fliP
fliR
flhA
fliS
fliK
fliA
motB
SNP count
All
EG1a (T1-T3)
EG1b (T1-T2)
IG1-others
1.640
1.645
1.650
1.655 (Mb)
d
Ref1: RIMD2210633
Ref2: FORC_006
Ref2: CHN25
Ref2: FORC_014
EG1a
EG1b
EG1d
1.5 (Mb)
Average copy probability
Chromosome 1
Chromosome 2
Population
VppAsia
VppUS1
VppX
VppUS2
Hybrid
SNP
Allele 1
Allele 2
Gene
Presence
Absence
Eco-groups
EG1a
EG1b
EG1c
EG1d
Tiers
T1
T2
T3


a
Swimming (top) and swarming (bottom) ability
Motility diameter (mm)
EG1a
EG1b
EG1c
EG1d
B1_1
B3_1
B5_2
C3_10
B1_10
C5_2
C6_5
B1_3
B2_10
B4_8
C1_5
b
Growth curve
OD600
Hours
c
Biofilm formation and colony morphology
OD595
1% NaCl
2% NaCl
3% NaCl


Casual
Going steady
Married
Setting up home together