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

Background *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.

Results We take advantage of this structure to perform a genome-wide screen for coadapted genetic elements. The great majority of interactions that we detect are between accessory genes, many involved in carbohydrate transport and metabolism, providing evidence for a plug and play-like architecture, in which elements have evolved to be functional immediately on arrival in a new genome. The most complex co-adaptations we identify include hundreds of core genome SNPs as well as accessory genome elements and involve genes encoding lateral flagella and cell wall biogenesis, implying that several genetically distinct strategies have evolved for colonizing surfaces.

Conclusions Our results provide evidence that, as in human relationships, coadaptation involves progressively increasing levels of commitment, with the most involved interactions becoming irreversible and presaging speciation.

Keywords

Vibrio parahaemolyticus, coadaptation, ecotype, epistasis

Background

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.

Epistasis between adaptive loci has important consequences for how evolution proceeds and the pattern of phenotypic diversity within individual populations. For example, in populations of 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 for a population to maintain distinct genetic strategies together 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. This means that it is possible to detect coadaptation based on detecting non-random associations between alleles [11].

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 25 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. Our approach and results provide a starting point for systematically investigating fitness landscapes in natural populations, suggests that organisms progressively modify their own landscape through coadaptation and highlights the fundamental importance of lateral flagella variation to *V. parahaemolyticus* 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 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). Interacting SNPs were substantially enriched for non-synonymous variants, which

116 is consistent with natural selection being the force generating the linkage disequilibrium we
117 detected.
118
119 The largest network (network 1) accounted for the majority of interacting SNPs as well as a
120 significant fraction of accessory genome elements (Table 1). Hierarchical clustering of the
121 interacting elements in this network based on *P* values (Figure 1c) revealed a complex pattern
122 of interactions but with differentiated sub-networks and we defined three large interaction
123 groups (IGs), IG1, IG2 and IG3 within it, placing remaining interactions within IG4. IG4-93
124 are displayed in Figure 2, 3, and Figure S2, while IG1-3 are shown separately in Figures 4
125 and 5.
126
127 We compared our results for core genome interactions with those obtained by SuperDCA, a
128 method that uses Direct Coupling Analysis to identify causal interactions [12], using the
129 default settings for the algorithm. To make the results as directly comparable as possible we
130 used the same 198 isolates that were used for the Fisher exact test. The most important
131 discrepancy is that although IG1 involves perfect associations, with *P* values as low as $1.4\times$
132 10^{-30}, the coupling strengths are lower than for the other groups we identified and none
133 appear amongst the top 5,000 couplings. This discrepancy is due the large number of SNPs
134 involved with similar association patterns, which means that coupling values are distributed
135 between them (Figure 1d, Figure S3). Excluding IG1 SNPs, there is a strong correlation
136 between SuperDCA coupling strengths and Fisher exact test *P* values (Figure 1e). And at a
137 stringent cutoff of $10^{-2.2}$, SuperDCA identifies the same multilocus interactions as Fisher
138 exact test does at $P < 10^{-10}$, with a few SNPs excluded (Figure 1f). The significance
139 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.

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-redundancy" 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).

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 10th 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

164 VppAsia isolates and others, with the island present in 52% of VppAsia strains and 90% of
165 others.
166
167 Genome islands are often associated with transmission mechanisms such as phage and
168 plasmids [13]. In our data, one example is IG17 which contains 4 genes annotated as being
169 phage related and a further 19 hypothetical proteins (Figure 3). In the reference strain, 16 of
170 the genes occur in a single block VP1563-VP1586, while 7 genes are present elsewhere in the
171 genome. 56% of the 469 strains had none of the genes, while 36% had more than 12 of them
172 and 8% had between 1 and 5, which presumably represent remnants of an old phage
173 infection. Only one gene (VP1563) in IG17 was found in appreciable frequency in strains that
174 had none of the other genes. This gene might represent cargo of the phage infection that is
175 able to persist for extended periods in the absence of infection due to a useful biological
176 function of its own.
177
178 Another common interaction is incompatibility between different accessory genome
179 elements, which is found in total 27 of the interaction groups (Figure 2). For example, all of
180 the 469 strains either have the gene *yniC*, which is annotated as being a phosphorylated
181 carbohydrates phosphatase, or at least 4 out of a set of 5 genes VP0363-VP0367 that includes
182 a phosphotransferase and a dehydrogenase (IG6, Figure 3) in the same genome location. Only
183 one strain has both sets of genes. This interaction also involves a core genome SNP, in the
184 adjacent gene VP0368, which is annotated as being a mannitol repressor protein. Another
185 pattern is found in IG45 (Figure 3), where there are three genes that are mutually
186 incompatible in our data, and 12% strains have none of the three genes. Only one of the three
187 genes is annotated.
188

189 **More detailed characterization of IG1-3**

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

204

205 Based on the clustering we split the genetic variants themselves into three tiers, shown in red,
206 blue and green in Figure 4. Tier 1 variants distinguish cleanly, or nearly cleanly between
207 EG1.1 and EG1.2 strains. Tier 2 variants are similar but with more discrepancies, while Tier
208 3 variants are typically fixed or nearly fixed within one eco-group and polymorphic within
209 the other. Each of the tiers includes multiple core and accessory genome regions. Some
210 regions are represented in only one tier, while others contain two or more. For example, the
211 lateral flagellar gene cluster (Ref1-23, VPA1538-VPA1557, Figure 4c) contains 383 Tier 1
212 SNPs, 24 from Tier 2 and 47 from Tier 3, however the extent of the region spanned by Tier 1
213 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 regulators, T6SS (type VI secretion system), and cellulose synthesis related genes 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. Tier 2 contains mostly SNPs in core genome, which located at *LuxR* family transcriptional regulator and uridine phosphorylase encoding genes. 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. Furthermore, the overall strength of the associations in the IG2 Tiers we have defined are much 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) [14] (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. 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.

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).

There are in total 60 EG1.2 strains in the global collection of 1103 *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 EG1.2 strains were isolated from clinical samples, including wound and stool, representing a lower proportion than for EG1.1 isolates (453/1043), including if the two pandemic lineages 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 93 interaction groups. 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 7).

Most interactions between core and accessory genomes are casual

A recent debate about whether the accessory genome evolves neutrally [15-17] 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.

288

289 Unsurprisingly, given the extensive literature highlighting the importance of genomic islands
290 to functional diversity of bacteria [13], the most common form of adaptation detected by our
291 screen is the coinheritance of accessory genome elements located in the same region of the
292 genome. Based on a minimum size threshold of 3 kb, we find 52 (56%) such interactions, the
293 largest of which is 57 kb (Table S3).

294

295 Previous approaches to detecting genome islands emphasize traits associated with horizontal
296 transmission, for example based on differences in GC content with the core genome, the
297 presence of phage-related genes or other markers of frequent horizontal transfer [18]. Our
298 approach, based simply on co-occurrence identifies a wider range of coinherited units and
299 suggests that many islands have functions related to carbohydrate metabolism.

300

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

305

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

313

314 **Sometimes it makes sense to go steady**

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

326

327 IG2, is an example of a complex coadaptation involving multiple core and accessory genome
328 regions. A large majority of strains in our dataset (439/469) either carry a cluster of genes
329 encoding a T6SS, or a cluster encoding cellulose biosynthesis genes, but few strains have
330 genes from both clusters (Figure 5a). Cells uses the T6SS to inject toxins into nearby bacteria
331 [19] and cellulose production to coat themselves in a protective layer [20]. Incompatibility
332 might have a functional basis, for example because cellulose production prevents the T6SS
333 functioning efficiently, or an ecological one, for example because cells that attack others do
334 not need to defend themselves.

335

336 **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 other membrane transport elements. There are also 285 loci (27%, Tier 3) in weaker disequilibrium, typically because they are polymorphic in one of the eco-groups. 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 6) suggest that biofilm formation is likely to be a key trait underlying the different ecological strategies of EG1.1 and EG1.2, but 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.

362 Despite the extensive differences that have accumulated between EG1.1 and EG1.2, there is
363 no evidence of restricted gene flow in most of the genome (Figure 4b), and even within the
364 flagellar gene cluster strongly differentiated regions are separated by a weakly differentiated
365 one (Figure 4c), implying that the coadaptation is being maintained by selection in the face of
366 frequent recombination. Initial divergence in flagellar function is likely to have led to
367 ecological differentiation, which led to bacteria having different nutritional inputs or
368 requirements and a broadening of the functional categories undergoing divergent selection.
369
370 How can the difference between IG1 and the other interaction groups in both the number of
371 associations and their strength be explained? *V. parahaemolyticus* is ubiquitous in shellfish in
372 warm coastal waters, within which it occurs at densities of around 1,000 cells per gram, so a
373 back of the envelope calculation suggests there are likely to be substantially more than 10^{15}
374 bacteria in the VppAsia population. The species also has a high estimated effective
375 population size [10, 11] and has strong codon bias, which is often argued to be evidence that
376 even tiny selective coefficients can drive adaptation [21]. Furthermore, recombination only
377 breaks up linkage disequilibrium between loci slowly. Therefore, weaker and more variable
378 patterns of association within IG2 and IG3 than in IG1 is unlikely to be a simple consequence
379 of the ineffectiveness of selection and is instead likely to reflect complexity in the fitness
380 landscape.
381
382 Strains gain flexibility by being able to switch between or modulate genomically encoded
383 strategies by homologous recombination. For example, expression of the T6SS might be
384 essential for survival in crowded habitats but detrimental in sparsely populated ones.
385 Recombinants between EG2.1 and EG2.2 at IG2 loci might represent transient adaptation of
386 strains to their immediate environment or long-term adaptation of intermediate strategies.

387 Furthermore, the phenotypic consequences of IG2 variants can depend on other loci in the
388 genome, such as IG1 variants, which is likely to reduce the strength of associations within
389 IG2. Crucially, the evolution of promiscuity is self-reinforcing because the presence of strains
390 using multiple strategies in the population also favors the presence of accessory genes and
391 core gene haplotypes that have high or intermediate fitness on multiple backgrounds.
392
393 On the other hand, an absence of intermediate genotypes in the population can favor the
394 evolution of fastidiousness, with particular accessory genes and haplotypes becoming
395 essential components of some genetic backgrounds but deleterious on others. A likely
396 scenario is that differences between IG1.1 and IG1.2 at a lateral flagellar gene made
397 recombinants between the two versions of the gene inviable and also created divergent
398 selection at a handful of other loci that was largely independent of the external environment
399 or of interactions with other genes. The evolution of fastidiousness, like the evolution of
400 promiscuity, can be self-reinforcing, and might have led to progressive increase in the
401 differentiation of EG1.2 strains from EG1.1 until the coadaptation of IG1.2 variants to each
402 other and of IG1.1 variants to each other became more-or-less irreversible, like marriage in
403 England prior to the reign of King Henry VIII.
404
405 **Coadapted gene complexes as speciation triggers**
406 Running the tape forward, it is easy to envisage the number of coadapted regions of the
407 genome within IG1 undergoing progressive enlargement, until the entire genome becomes
408 differentiated. As coadapted regions become more numerous, the proportion of
409 recombination events between eco-groups that are maladaptive will increase, which might
410 prompt the evolution of mechanistic barriers to genetic exchange between them.
411

Mechanisms by which new bacterial species arise are frequently discussed in the literature [22-24] but there is currently little data on how the process unfolds. IG1 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 [25], 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

437 highly localized within the genome, it can, on careful inspection be clearly be attributed to
438 selection, but in other bacteria patterns are likely to be less straightforward, making it
439 challenging to understand to whether adaptive processes drive population structure, or vice
440 versa. Natural selection is the jewel of evolution but distinguishing it from other processes
441 requires in depth understanding of the relevant biology in addition to suitable data and
442 statistical methods.
443
444 **Materials and Methods**
445 **Genomes used in this work**
446 Totally 1,103 global *V. parahaemolyticus* genomes were used in this work (Table S1), which
447 also were analyzed in our other studies [8, 10]. To reduce clonal signals, we firstly made a
448 "non-redundancy" dataset of 469 strains, in which no sequence differed by less than 2,000
449 SNPs in the core genome. They were attributed to 4 populations, VppAsia (383 strains),
450 VppX (43), VppUS1 (18) and VppUS2 (21) based on fineSTRUCTURE result [26]. We then
451 focused on VppAsia which has more strains, to generate a genome dataset in which strains
452 represent a freely recombining population. We selected 386 genomes from 469 non-
453 redundancy genome dataset, including all the 383 VppAsia genomes and 3 outgroup genomes
454 which were randomly selected from VppX, VppUS1 and VppUS2 population, respectively.
455 These 386 genomes were used in Chromosome painting and fineSTRUCTURE analysis [26]
456 as previously described [11]. Initial fineSTRUCTURE result revealed multiple clonal signals
457 still exist, thus we selected one representative genome from each clone, combined them with
458 the remaining genomes and repeated the process. After 14 iterations, we got a final dataset of
459 201 genomes with no trace of clonal signals, involving 198 VppAsia genomes that were used
460 in further analysis.
461

462 The copy probability value of each strain at each SNP was generated by Chromosome
463 painting with "-b" option, and the average copying probability value of a given strain group
464 (e.g. EG1.2) at each SNP was used in Figure 4, 5 and Figure S4.
465
466 **Variation detection, annotation and phylogeny**
467 We re-called SNPs for 198 VppAsia genomes by aligning the assembly against reference
468 genome (RIMD 2210633) using MUMmer [27] as previous described [11]. Totally 565,466
469 bi-allelic SNPs were identified and 151,957 bi-allelic SNPs with minor allele frequency > 2%
470 were used in coadaptation detection. We re-annotated all the assemblies using Prokka [28],
471 and the annotated results were used in Roray [29] to identify the pan-genome and gene
472 presence/absence, totally 41,053 pan-genes were found and 14,486 accessory genes (present
473 in > 2% and < 98% strains) were used in coadaptation detection. The pan-gene protein
474 sequences of Roary were used to blast (blastp) against COG and KEGG database to get
475 further annotation.
476
477 The Neighbour-joining trees were built by using the TreeBest software
478 (http://treesoft.sourceforge.net/treebest.shtml) based on sequences of concatenated SNPs, and
479 were visualized by using online tool iTOL [30].
480
481 **Detection of coadapted loci**
482 Totally 151,957 bi-allelic SNPs and 14,486 accessory genes identified from 198 independent
483 VppAsia genomes were used in coadaptation detection by two methods. Firstly we used
484 Fisher exact test to detect the linkage disequilibrium of each SNP-SNP, SNP-accessory gene,
485 and accessory gene-gene pair. Presence or absence of an accessory gene was considered as its
486 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 [12] to detect the coadaptation between SNPs, using the same subset of 198 strains to make the analysis as comparable as possible. It is based on direct coupling analysis (DCA) model 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.

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 chromosome or contig was counted as larger than 3 kb and such pairs were kept in further analysis. Circos [31] 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.

Determination of phenotypes

Bacteria strains. In the phenotype experiments, totally 11 strains were randomly selected respectively from two EGs that defined by IG1 variants, including 7 EG1.1 strains (B1_10, 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). 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.

537 **Growth curve.** *V. parahaemolyticus* strains in 96-well plate were cultured overnight at 30 °C
538 in 3% NaCl-LB broth. The optical density of each culture was adjust to an OD_{600} of 0.6.Then
539 1 ml of each culture was inoculated 100 ml of 3% NaCl-LB broth in a 96-well plate and
540 cultured at 30 °C. The growth of each culture were measured every 1 hour at the optical
541 density of 600 nm using Multiskan Spectrum.
542
543 **Biofilm formation.** *V. parahaemolyticus* strains were cultured overnight at 30 °C in 3%
544 NaCl-LB broth. 2 µl of each overnight culture was inoculated to 100 µl of 3% NaCl-LB broth
545 in a 96-well plate and cultured at 30 °C for 24 h statically. The supernatant was discarded and
546 each well was washed once with sterile phosphate-buffered saline (PBS). 0.1% Crystal violet
547 (wt/vol) was added to each well and incubated at room temperature for 30 min. The crystal
548 violet was decanted, and each well was washed once with sterile PBS. Crystal violet that
549 stained biofilm was solubilized with dimethylsulfoxide (DMSO), and then measured at the
550 optical density of 595 nm using Multiskan Spectrum (Thermo Scientific).
551
552 **Acknowledgements**
553 This work is supported by the National Key Research & Development Program of China (No.
554 2017YFC1601503, 2016YFC1200100, and 2017YFC1200800), Sanming Project of
555 Medicine in Shenzhen (No. SZSM201811071), the National Natural Science Foundation of
556 China (No. 31770001) and the Key Research Program of the Chinese Academy of Sciences
557 (No. ZDRW-ZS-2017-1). D.F. is funded by a Medical Research Council Fellowship as part
558 of the MRC CLIMB consortium for microbial bioinformatics (grant number
559 MR/M501608/1).
560
561 **Author Contributions**

562 D. F., Y. C., and R. Y. designed the study and coordinated the project; Y. C., C. Y., and D. F.
563 analyzed the data; H. Q. and H. W. performed phenotype experiments; D. F. and Y. C. wrote
564 the manuscript. All authors approved the final version of the manuscript.
565

566 **Competing Financial Interests statement**

567 No
568

569 **References**

570 1. Darwin C: **The origin of species. 6th.** John Murray, London; 1859.
571 2. Neher RA, Shraiman BI: **Statistical genetics and evolution of quantitative traits.** *Reviews of*
572 *Modern Physics* 2011, **83:**1283.
573 3. Pritchard JK, Przeworski M: **Linkage disequilibrium in humans: models and data.** *The*
574 *American Journal of Human Genetics* 2001, **69:**1-14.
575 4. Dobzhansky T: **1937Genetics and the origin of species.** *NewYork: Columbia University*
576 *PressDobzhanskyGenetics and the origin of species1937* 1970.
577 5. Wallace B: **Coadaptation revisited.** *J Hered* 1991, **82:**89-95.
578 6. Makino K, Oshima K, Kurokawa K, Yokoyama K, Uda T, Tagomori K, Iijima Y, Najima M,
579 Nakano M, Yamashita A, et al: **Genome sequence of Vibrio parahaemolyticus: a pathogenic**
580 **mechanism distinct from that of V cholerae.** *Lancet* 2003, **361:**743-749.
581 7. Gibson B, Wilson DJ, Feil E, Eyre-Walker A: **The distribution of bacterial doubling times in**
582 **the wild.** *Proc Biol Sci* 2018, **285**.
583 8. Yang C, Pei X, Wu Y, Yan L, Yan Y, Song Y, Coyle N, Martinez-Urtaza J, Quince C, Hu Q, et al:
584 **Recent mixing of Vibrio parahaemolyticus populations.** *bioRxiv* 2018.
585 9. Arnold BJ, Gutmann MU, Grad YH, Sheppard SK, Corander J, Lipsitch M, Hanage WP: **Weak**
586 **Epistasis May Drive Adaptation in Recombining Bacteria.** *Genetics* 2018, **208:**1247-1260.
587 10. Yang C, Cui Y, Didelot X, Yang R, Falush D: **Why panmictic bacteria are rare.** *bioRxiv* 2018.

588 11. Cui Y, Yang X, Didelot X, Guo C, Li D, Yan Y, Zhang Y, Yuan Y, Yang H, Wang J, et al: **Epidemic**
589 **Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen Vibrio**
590 **parahaemolyticus.** *Mol Biol Evol* 2015, **32:**1396-1410.
591 12. Puranen S, Pesonen M, Pensar J, Xu YY, Lees JA, Bentley SD, Croucher NJ, Corander J:
592 **SuperDCA for genome-wide epistasis analysis.** *Microb Genom* 2018.
593 13. Dobrindt U, Hochhut B, Hentschel U, Hacker J: **Genomic islands in pathogenic and**
594 **environmental microorganisms.** *Nat Rev Microbiol* 2004, **2:**414-424.
595 14. Tanabe T, Funahashi T, Okajima N, Nakao H, Takeuchi Y, Miyamoto K, Tsujibo H, Yamamoto
596 S: **The Vibrio parahaemolyticus pvuA1 gene (formerly termed psuA) encodes a second**
597 **ferric vibrioferrin receptor that requires tonB2.** *FEMS Microbiol Lett* 2011, **324:**73-79.
598 15. Vos M, Eyre-Walker A: **Are pangenomes adaptive or not?** *Nat Microbiol* 2017, **2:**1576.
599 16. Andreani NA, Hesse E, Vos M: **Prokaryote genome fluidity is dependent on effective**
600 **population size.** *ISME J* 2017, **11:**1719-1721.
601 17. Shapiro BJ: **The population genetics of pangenomes.** *Nat Microbiol* 2017, **2:**1574.
602 18. Langille MG, Hsiao WW, Brinkman FS: **Detecting genomic islands using bioinformatics**
603 **approaches.** *Nat Rev Microbiol* 2010, **8:**373-382.
604 19. Salomon D, Gonzalez H, Updegraff BL, Orth K: **Vibrio parahaemolyticus type VI secretion**
605 **system 1 is activated in marine conditions to target bacteria, and is differentially regulated**
606 **from system 2.** *PLoS One* 2013, **8:**e61086.
607 20. Tischler AD, Camilli A: **Cyclic diguanylate (c-di-GMP) regulates Vibrio cholerae biofilm**
608 **formation.** *Mol Microbiol* 2004, **53:**857-869.
609 21. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE: **Variation in the strength of selected**
610 **codon usage bias among bacteria.** *Nucleic Acids Res* 2005, **33:**1141-1153.
611 22. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, Polz MF, Alm EJ:
612 **Population genomics of early events in the ecological differentiation of bacteria.** *Science*
613 2012, **336:**48-51.

614 23. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP: **The bacterial species challenge: making**
615 **sense of genetic and ecological diversity.** *Science* 2009, **323:**741-746.
616 24. Falush D, Torpdahl M, Didelot X, Conrad DF, Wilson DJ, Achtman M: **Mismatch induced**
617 **speciation in Salmonella: model and data.** *Philos Trans R Soc Lond B Biol Sci* 2006, **361:**2045-
618 2053.
619 25. Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, Miska EA, Durbin R,
620 Genner MJ, Turner GF: **Genomic islands of speciation separate cichlid ecomorphs in an East**
621 **African crater lake.** *Science* 2015, **350:**1493-1498.
622 26. Lawson DJ, Hellenthal G, Myers S, Falush D: **Inference of population structure using dense**
623 **haplotype data.** *PLoS Genet* 2012, **8:**e1002453.
624 27. Delcher AL, Salzberg SL, Phillippy AM: **Using MUMmer to identify similar regions in large**
625 **sequence sets.** *Curr Protoc Bioinformatics* 2003, **Chapter 10:**Unit 10 13.
626 28. Seemann T: **Prokka: rapid prokaryotic genome annotation.** *Bioinformatics* 2014, **30:**2068-
627 2069.
628 29. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane
629 JA, Parkhill J: **Roary: rapid large-scale prokaryote pan genome analysis.** *Bioinformatics*
630 2015, **31:**3691-3693.
631 30. Letunic I, Bork P: **Interactive tree of life (iTOL) v3: an online tool for the display and**
632 **annotation of phylogenetic and other trees.** *Nucleic Acids Res* 2016, **44:**W242-245.
633 31. Krzywinski M, Schein JI: **Circos: an information aesthetic for comparative genomics.**
634 *Genome Research* 2009, **19:**1639-1645.

635

636 **Figure legends**

637 **Figure 1. Detection of coadaptation loci in *V. parahaemolyticus*.** (a) NJ tree of 198
638 VppAsia strains based on 151,957 bi-allelic SNPs (minor allele frequency > 2%) (b) Q-Q plot
639 of Fisher exact test *P* values between genetic variants. The vertical dotted line shows the

640 threshold $P = 10^{-10}$. (c) Hierarchical clustering of interacting variants (column and row) in the
641 largest network based on Fisher exact test *P* value. IG2-1 and IG2-2 were integrated into IG2.
642 (d) Distribution of SuperDCA coupling strength value. (e) Correlation between Fisher exact
643 test *P* value and SuperDCA coupling strength. (f) Overlap of strong linked SNP sites detected
644 by Fisher exact test ($P < 10^{-10}$, excluding IG1 variants) and SuperDCA (coupling strength >
645 $10^{-2.2}$). Red for interacted SNP pairs detected by both methods, blue for SNP pairs detected
646 only by Fisher exact test. Only SNP/accessory gene pairs with distance > 3 kb were included
647 in (c,e and f).
648
649 **Figure 2. Landscape of coadaptation (excluding IG1-3).** (a) Gene maps of different IGs.
650 Arrows indicate genes and red for detected coadaptation core genes, blue for accessory genes
651 and orange for genes with no coadaptation signal. Black vertical lines indicate SNPs in the
652 interaction group. The colors of the bar on the left indicates average linkage strength of the
653 loci in each IG. Vertical dotted lines were used to split compatible genes with physical
654 distance larger than 3 kb, or genes located in different contigs, chromosomes and strains.
655 Dotted rectangles indicate incompatible genes. IGs with genome block length larger than 60
656 kb are broken by double slash and shown in (b) after zooming out. COG classification labels
657 are shown above the genes. (c) COG classification and GC content of all the genes used in
658 detection (top) and of different types of coadaptation genes. Red for core genes and blue for
659 accessory genes. The first number in brackets is the number of genes with COG annotation
660 and the second is the total number of genes in the category.
661
662 **Figure 3. Representative interaction groups.** Hierarchical clustering of 469 non-
663 redundancy strains (columns) based on coadaptation loci (rows) of 4 representative IGs.
664 Colors of the heatmap indicate the status of genetic variants, with light orange/orange for two

665 alleles of a SNP, light yellow/brown for absence and presence of the accessory genes. Bar
666 colors below the tree on the top indicate the populations of strains according to the legend.
667 Function summary of involved genes is shown on the top of each heatmap.

668

669 **Figure 4. Interaction group 1.** (a) Hierarchical clustering of 469 non-redundancy strains
670 (columns) based on coadaptated loci (rows) of IG1. Color scheme of the heatmap is the same
671 as in Figure 3. Branch colors of the tree on the top indicate the populations of strains
672 according to the legend. The black bars below the tree indicate the strains used in
673 experiments. Branch colors of the tree on the left indicate the tiers of coadaptation loci, with
674 red, blue and green for tier 1, 2 and 3, respectively. The same colors are used for tiers of
675 coadaptation in panel (b) and (c). Colors of the bar on the right indicate the genome position
676 of coadaptation loci, which is corresponding to the bar colors in panel (b). (b) Gene map of
677 coadaptation genome blocks of IG1. Two reference genomes were used to show coadaptation
678 variants in IG1. The left curve indicates the value of average copy probability of EG1.2
679 strains that copied from themselves throughout the genome of Ref1, with vertical dotted line
680 indicates the expect value (number of EG1.2 strains /number of all strains). The bars indicate
681 the reference genomes and different chromosomes are separated by a horizontal short bar.
682 The labels of coadaptation genome blocks are shown above them and are corresponding to
683 information in Table S2. Arrows and vertical lines separately indicate genes and SNPs, which
684 were colored according to different coadaptation tiers, and core genes were colored grey and
685 genes with no coadaptation signal were light orange. COG classification labels are shown
686 above the genes. (c) The distribution of coadaptation SNPs in the lateral flagellar gene cluster
687 region (VPA1538-1557). The top indicates the gene organization of lateral flagellar gene
688 cluster. Light orange rectangles show the accessory genome region. The histograms indicate
689 the distribution of SNPs along the gene cluster, with colors of bars indicate coadaptation tiers.

690

691 **Figure 5. Interaction group 2 (a) and 3 (b).** Layout and colors are the same as in Figure 4.
692 EG1.2 and EG2.1 strains are shown below the tree on the top for comparison (black bars).
693 The left curves indicate the average copy probability value of EG2.1 (a) and EG3.1 (b) strains
694 copy from themselves throughout the reference genome, respectively. Twenty randomly
695 selected SNPs separately from IG2-1 and IG2-2 were used in hierarchical clustering to
696 minimize the influence of number of variants.

697

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

701

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

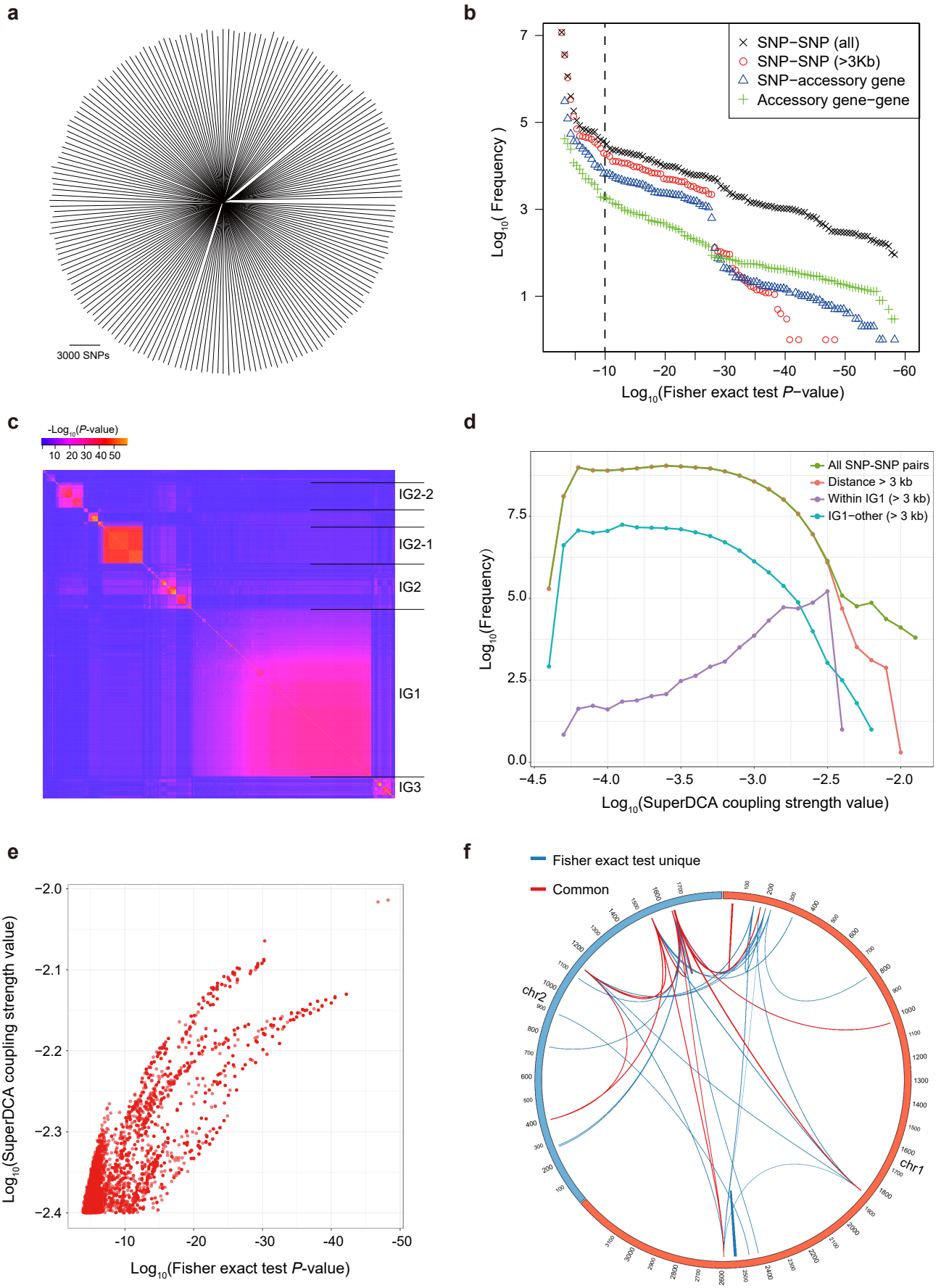
Table1. Summary of interactions detected in coadaptation screen.

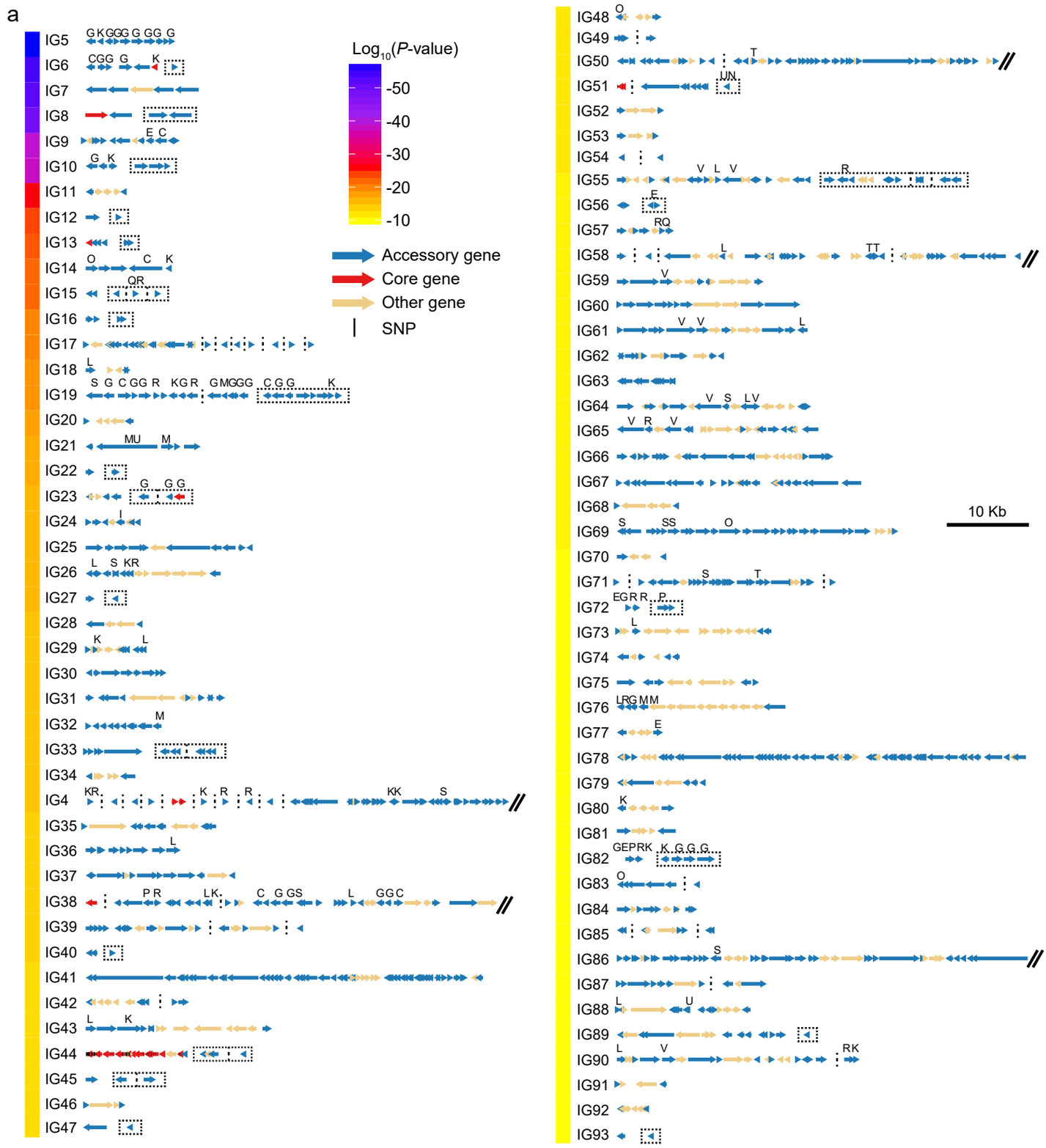
	Total*	Network 1					IG5-93	Frequency***
		IG1	IG2	IG3	IG4	Frequency**		
SNP-SNP pair	2.3×10^{10}	283632	4398	1150	6	0.00%	22751	0.00%
SNP-Accessory gene pair	2.2×10^{9}	104096	7319	2446	112	0.01%	1188	0.00%
Accessory gene-gene pair	2.1×10^{8}	7999	2823	973	1692	0.01%	12264	0.01%
SNP number	151957	917	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%

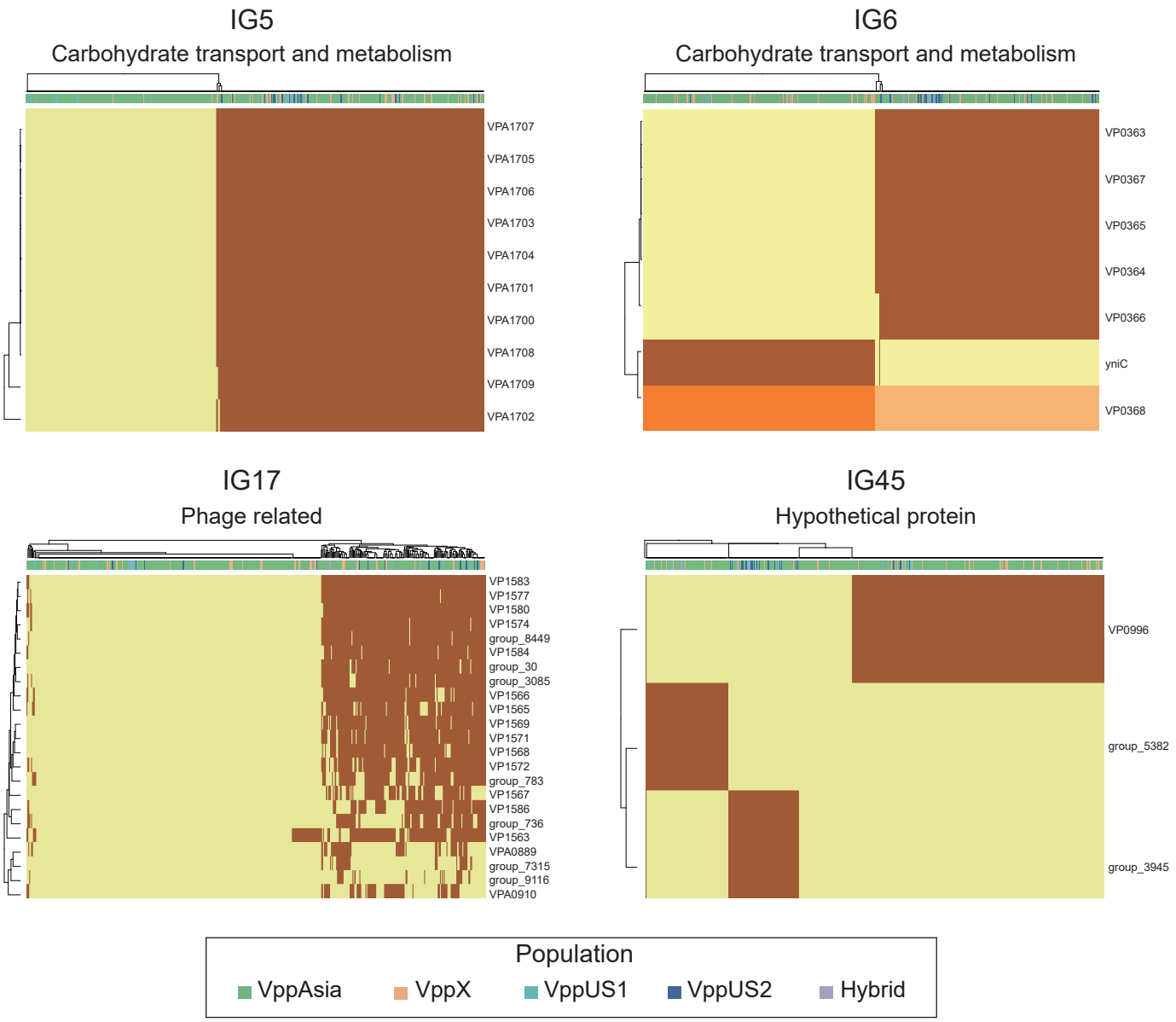
* All of the variations used in coadaptation screen.

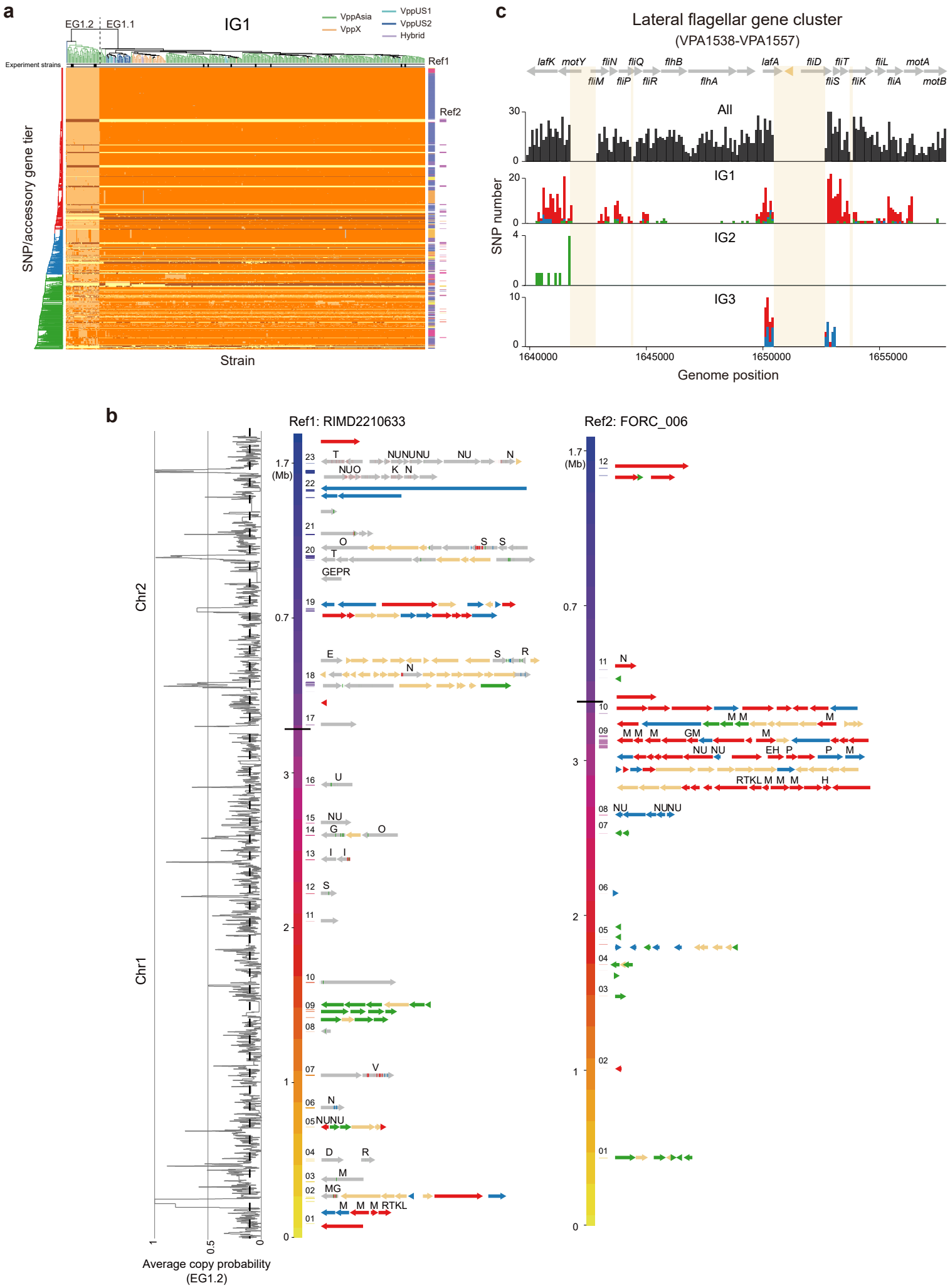
** Ratio of the number of variations in Network 1 to the total number of variations.

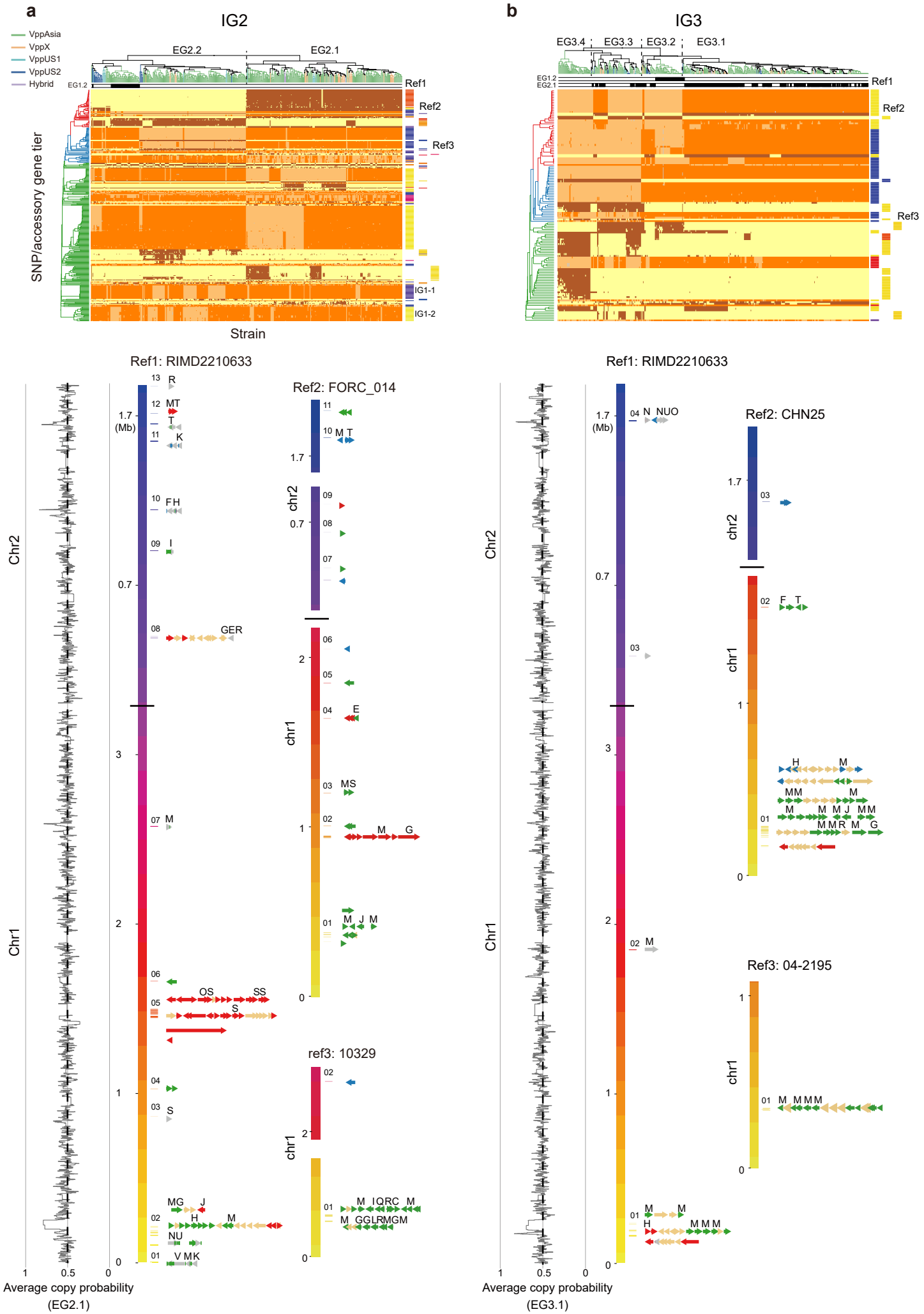
*** Ratio of the number of variations in IG5-93 to the total number of variations.

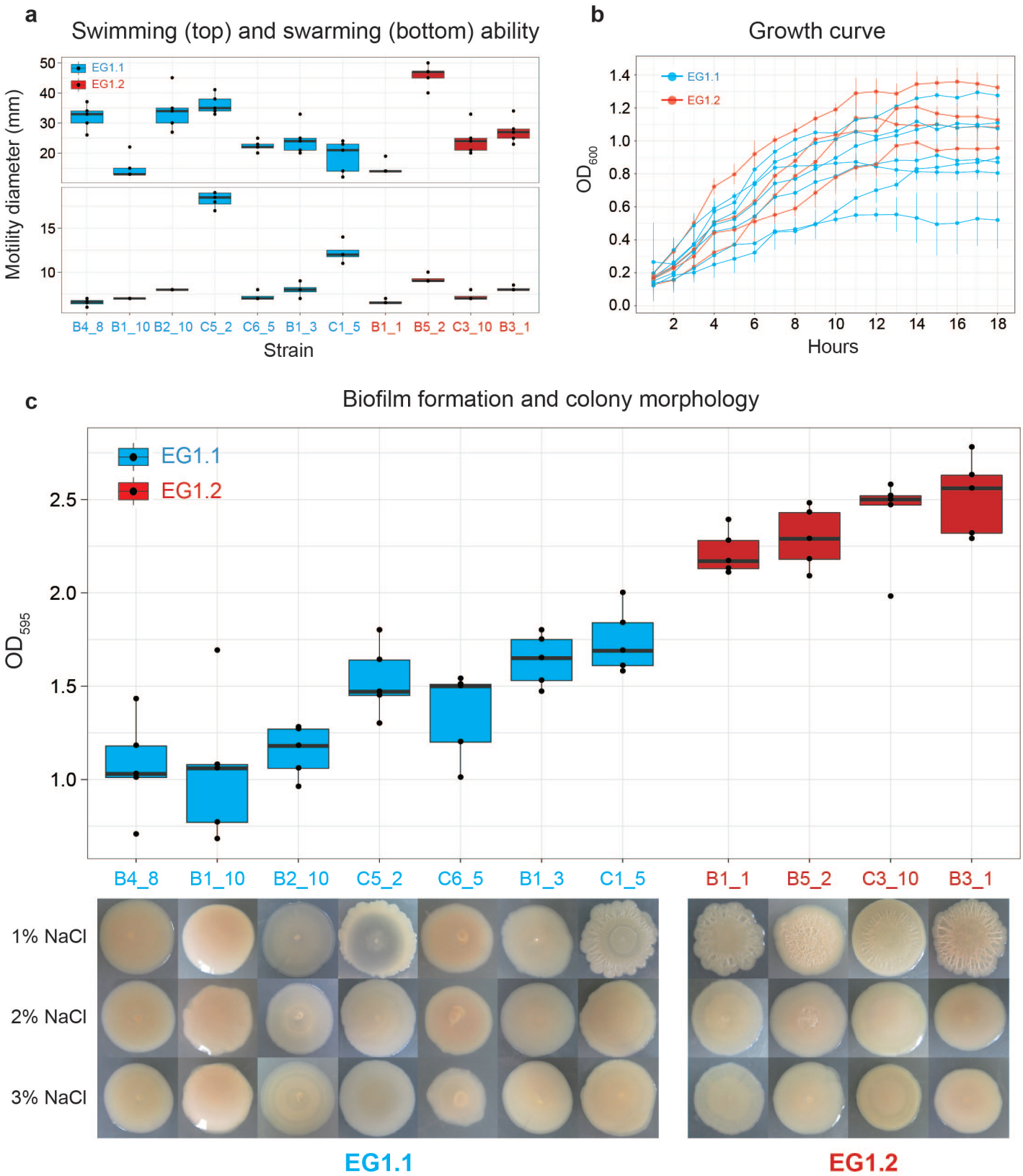

a
3000 SNPs
b
SNP−SNP (all)
SNP−SNP (>3Kb)
SNP−accessory gene
Accessory gene−gene
Log10(Frequency)
Log10(Fisher exact test P−value)
c
-Log10(P-value)
10 20 30 40 50
IG2-2
IG2-1
IG2
IG1
IG3
d
All SNP-SNP pairs
Distance > 3 kb
Within IG1 (> 3 kb)
IG1−other (> 3 kb)
Log10(Frequency)
Log10(SuperDCA coupling strength value)
e
Log10(SuperDCA coupling strength value)
Log10(Fisher exact test P-value)
f
Fisher exact test unique
Common
chr1
chr2

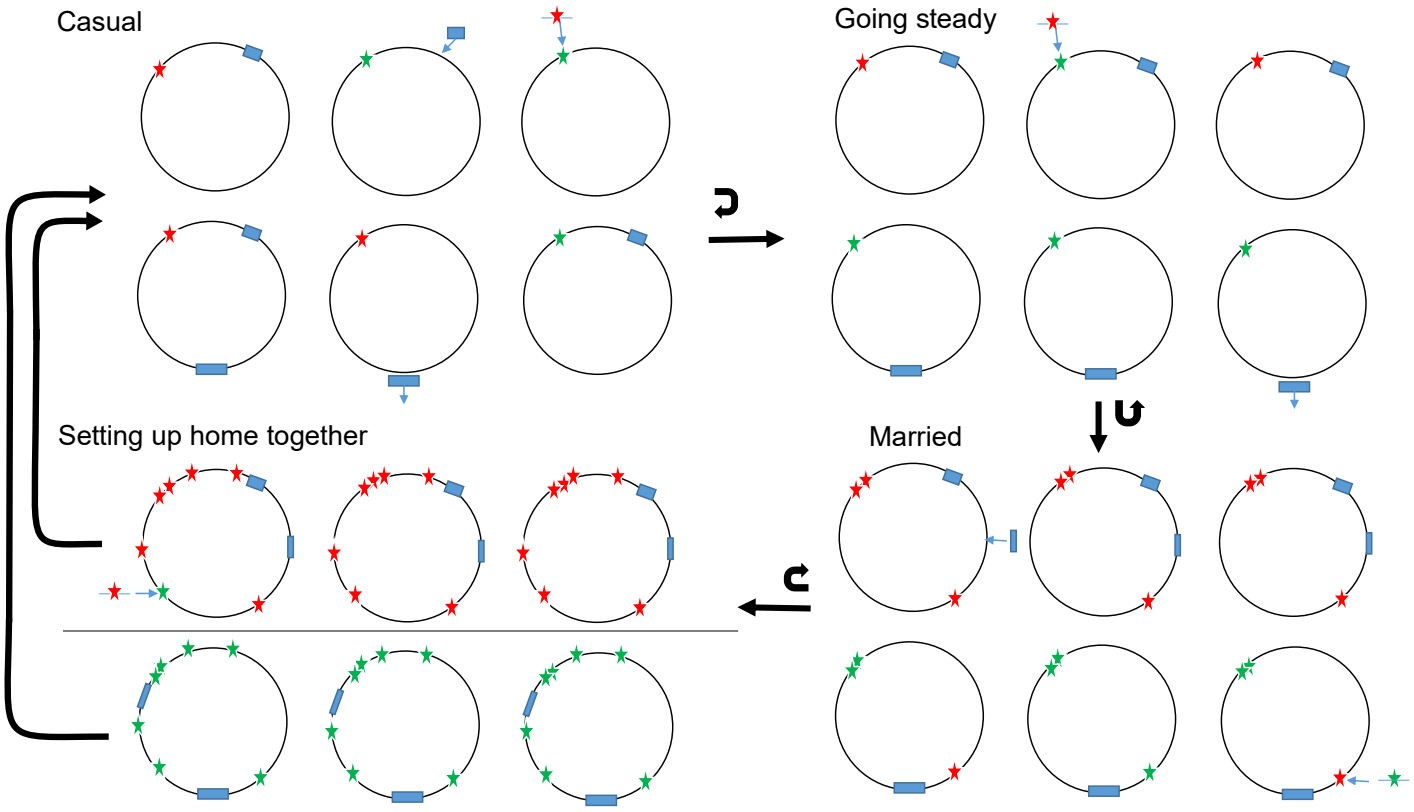

a
Log10(P-value)
-50
-40
-30
-20
-10
Accessory gene
Core gene
Other gene
SNP
10 Kb
b
10 Kb
c
Core gene
Accessory gene
All genes (4214/18422)
IG1-4 (195/563)
Compatibility (129/851)
Incompatibility (69/290)
Gene number
COG classification
GC content (%)
COG classification
A: RNA processing and modification
B: Chromatin structure and dynamics
C: Energy production and conversion
D: Cell cycle control, cell division, chromosome partitioning
E: Amino acid transport and metabolism
F: Nucleotide transport and metabolism
G: Carbohydrate transport and metabolism
H: Coenzyme transport and metabolism
I: Lipid transport and metabolism
J: Translation, ribosomal structure and biogenesis
K: Transcription
L: Replication, recombination and repair
M: Cell wall/membrane/envelope biogenesis
N: Cell motility
O: Posttranslational modification, protein turnover, chaperones
P: Inorganic ion transport and metabolism
Q: Secondary metabolites biosynthesis, transport and catabolism
R: General function prediction only
S: Function unknown
T: Signal transduction mechanisms
U: Intracellular trafficking, secretion, and vesicular transport
V: Defense mechanisms


IG5
Carbohydrate transport and metabolism
VPA1707
VPA1705
VPA1706
VPA1703
VPA1704
VPA1701
VPA1700
VPA1708
VPA1709
VPA1702
IG6
Carbohydrate transport and metabolism
VP0363
VP0367
VP0365
VP0364
VP0366
yniC
VP0368
IG17
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
IG45
Hypothetical protein
VP0996
group_5382
group_3945
Population
VppAsia
VppX
VppUS1
VppUS2
Hybrid


a
IG1
EG1.2
EG1.1
VppAsia
VppX
VppUS1
VppUS2
Hybrid
Experiment strains
Ref1
Ref2
SNP/accessory gene tier
Strain
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
All
IG1
IG2
IG3
SNP number
1640000
1645000
1650000
1655000
Genome position
b
Ref1: RIMD2210633
Ref2: FORC_006
Chr2
Chr1
1.7 (Mb)
Average copy probability
(EG1.2)


a
IG2
VppAsia
VppX
VppUS1
VppUS2
Hybrid
EG2.2
EG2.1
EG1.2
Ref1
Ref2
Ref3
IG1-1
IG1-2
SNP/accessory gene tier
Strain
Ref1: RIMD2210633
Ref2: FORC_014
ref3: 10329
Chr2
Chr1
1.7
(Mb)
0.7
chr2
chr1
Average copy probability
(EG2.1)
b
IG3
EG3.4
EG3.3
EG3.2
EG3.1
Ref1: RIMD2210633
Ref2: CHN25
Ref3: 04-2195
Average copy probability
(EG3.1)


a
Swimming (top) and swarming (bottom) ability
EG1.1
EG1.2
Motility diameter (mm)
50
40
30
20
15
10
B4_8 B1_10 B2_10 C5_2 C6_5 B1_3 C1_5 B1_1 B5_2 C3_10 B3_1
Strain
b
Growth curve
EG1.1
EG1.2
OD600
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
2 4 6 8 10 12 14 16 18
Hours
c
Biofilm formation and colony morphology
EG1.1
EG1.2
OD595
2.5
2.0
1.5
1.0
B4_8 B1_10 B2_10 C5_2 C6_5 B1_3 C1_5 B1_1 B5_2 C3_10 B3_1
1% NaCl
2% NaCl
3% NaCl
EG1.1
EG1.2


Casual
Going steady
Married
Setting up home together