The landscape of coadaptation in *Vibrio parahaemolyticus*

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

1 State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing 100071, China;
2 Institute for Nutritional Sciences, Chinese Academy of Sciences, Shanghai 200031, China
3 School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China
4 University of Bath, Bath, Somerset, United Kingdom

[#]These authors contributed equally to the article
* Corresponding authors: D. F. (danielfalush@googlemail.com) or Y. C. (cuiyujun.new@gmail.com)

Abstract

Vibrio parahaemolyticus is a human gastrointestinal pathogen that thrives in warm brackish waters and is unusual amongst bacteria in having a population structure that approximates panmixia. We take advantage of this structure to perform a genome-wide screen for coadapted genetic elements. 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. 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.

Introduction

35 Biological life is made possible by the cooperative action of thousands of genes. Much
36 insight in to the way that genes interact with each other to produce complex phenotypes has
37 been gained by experiments such as systematic knock out screens in yeast (Costanzo et al.
38 2016), *Escherichia coli* (Babu et al. 2011) and other organisms and clonal evolution in
39 controlled environments (Tenaillon et al. 2012, Jerison and Desai 2015), but it is challenging
40 for any laboratory experiment to characterize the complexity of real environments or the
41 fitness effects of natural polymorphisms. Genes in the same genome co-adapt to each other
42 progressively, with the course of evolution depending on the variants that segregate in the
43 population as well as the fitness landscape, which may itself change over time. Therefore, to
44 understand the role interactions between genes play in the evolution of complex phenotypes,
45 laboratory experiments need to be complemented by studies of natural variation.
46 One approach to identifying fitness interactions is to identify statistical associations between
47 genetic variants segregating within natural populations (Campo et al. 2008, Kryazhimskiy et
48 al. 2011, Cui et al. 2015, Figliuzzi et al. 2016, Skwark et al. 2017, Puranen et al. 2018). If
49 variants are found together more frequently than expected, this might indicate that they
50 function together to improve the fitness of the organism. There are three major challenges in
51 its application. The first is obtaining a large enough sample to achieve statistical significance
52 to detect associations, given the enormous number of potential interactions are possible
53 between variants throughout the genome. The second is that it is difficult to separate
54 historical contingency from adaptation. Are genetic variants found together in the same
55 organisms because they are beneficial together or simply because they have been inherited
56 together? The third is that most methods are based on detecting pairwise interactions, but
57 selection on quantitative traits or other complex genetic interactions might also be common
58 (Kuzmin et al. 2018).

59 Here we address the first two of these challenges by studying patterns of linkage
60 disequilibrium in a panmictic bacterial population, in which recombination has been
61 extensive enough to largely scramble genetic variation (Yang et al. 2018, manuscript in
62 preparation). *Vibrio parahaemolyticus* lives in coastal waters and causes gastroenteritis in
63 humans and economically devastating diseases in farmed shrimps and is capable of
64 replicating in less than 10 minutes in appropriate conditions (Makino et al. 2003), while the
65 doubling time in the wild of the related bacteria *Vibrio cholerae* has been estimated as
66 slightly over an hour (Gibson et al. 2018). Approximately 0.017% of the genome recombines
67 each year (Yang et al. 2018), while there are thousands of cell divisions per year, meaning
68 that for any given pair of interacting loci there are hundreds of thousands of generations of
69 selection, on average, between recombination events affecting either one of them. As a result,
70 even epistatic interactions that involve only small selective coefficients s (for example $s = 1.0$
71 $\times 10^{-4}$) can create an imprint on the genome in the form of strong linkage disequilibrium
72 (Arnold et al. 2018), that can potentially be detectable with relatively modest sample sizes.
73 We have taken a conservative approach to identifying statistical associations, rigorously
74 filtering the set of genomes used in our discovery dataset to eliminate any hint of population
75 structure. We performed more than 25 billion Fisher exact tests between variants in the core
76 and accessory genomes, using a cut-off of $P < 10^{-10}$ with the aim of assembling a
77 comprehensive list of common genetic variants that have strong linkage disequilibrium
78 between them due to fitness interactions. We find that the great majority of interactions
79 involve small numbers of accessory genome elements, with surprisingly little involvement of
80 the core genome. However, we also identify three complex multi-locus interactions, in which
81 core and accessory genomes have evolved in parallel to create coadapted gene complexes
82 with distinct strategies. Our approach and results provide a starting point for systematically
83 investigating fitness landscapes in natural populations, suggests that organisms progressively

84 modify their own landscape through coadaptation and highlights the fundamental importance
85 of lateral flagella variation to *V. parahaemolyticus* ecology.
86
87 **Results**
88 **Detection and characterization of interaction groups**
89 To avoid false positives due to population structure, we restricted our initial analysis to the
90 strains from the Asia population, VppAsia, within our global collection of 1,103 isolates and
91 iteratively removed isolates until there was no sign of clonal structure (Methods), leading to a
92 discovery dataset of 198 strains (Figure 1a, Figure S1). We performed a Fisher exact test of
93 associations between all pairwise combinations of 151,957 SNP variants within the core and
94 14,486 accessory genome elements. As has been observed previously (Cui et al. 2015), most
95 of strong associations occurred between sites within 3 kb on the chromosome (Figure 1b). In
96 order to exclude associations that arise only due to physical linkage, we excluded all sets of
97 associations that spanned less than 3 kb, including between accessory genome elements. This
98 left us with 452,849 interactions with $P < 10^{-10}$, which grouped into 90 networks of
99 associated elements, all of which involved at least one accessory-genome element, with 8
100 also including core genome SNPs and 38 of which included multiple genome regions. In total
101 these networks included 1,936 SNPs in 110 core genes and 1593 accessory genome elements.
102 Interacting SNPs were substantially enriched for non-synonymous variants (Table 1), which
103 is unsurprising given the role of natural selection in generating associations.
104 The largest network (network 1) accounted for the majority of interacting SNPs as well as a
105 significant fraction of accessory genome elements (Table 1). Hierarchical clustering of the
106 interacting elements in this network based on *P* values (Figure 1c) revealed a complex pattern
107 of interactions but with differentiated sub-networks and we defined three large interaction
108 groups (IGs), IG1, IG2 and IG3 within it, placing remaining interactions within IG4. IG4-93

109 are displayed in Figure 2 and 3 and Figure S2, while IG1-3 are shown separately in Figures 4
110 and 5.
111 We compared our results for core genome interactions with those obtained by SuperDCA, a
112 method that uses Direct Coupling Analysis to identify causal interactions (Puranen et al.
113 2018), using the default settings for the algorithm. The most important discrepancy is that
114 although IG1 involves perfect associations, with *P* values as low as 1.4×10^{-30}, the coupling
115 strengths are lower than for the other groups we identified and none appear amongst the top
116 5,000 couplings. This discrepancy is due the large number of SNPs involved with similar
117 association patterns, which means that coupling values are distributed between them (Figure
118 1d, Figure S3). Excluding IG1 SNPs, there is a strong correlation between SuperDCA
119 coupling strengths and Fisher exact test *P* values (Figure 1e). And at a stringent cutoff of 10^{-}
120 $^{2.2}$, SuperDCA identifies the same multilocus interactions as Fisher exact test does at $P <$
121 10^{-10}, with a few SNPs excluded (Figure 1f). The significance thresholds for both the Fisher
122 exact test and SuperDCA could be relaxed to identify a substantially larger number of true-
123 positive hits, at the likely expense of some false ones, but we do not investigate these
124 associations further here.
125 For IG1-4, COG classes M and N, cell wall biosynthesis and cell motility, are substantially
126 overrepresented, relative to their overall frequency in the genome (Figure 2c, Fisher exact
127 test, $P < 0.01$). Amongst the other interaction groups, class G, encoding carbohydrate
128 transport and metabolism are substantially overrepresented (Fisher exact test, $P < 0.01$),
129 particularly amongst groups involving incompatibilities. There are also differences in GC
130 content between accessory genes in IGs and others (Figure 2c), with IGs having higher mean
131 values, especially in compatibility IGs.
132 **Structure of variation within interaction groups**

133 For each interaction group, we investigated how genetic variation was structured within the
134 *V. parahaemolyticus* population using a larger "non-redundancy" set of 469 strains, which
135 includes isolates from all four of the populations identified in (Yang et al. 2018), but excludes
136 closely related isolates, differing at less than 2,000 SNPs. We performed hierarchical
137 clustering of the strains based on the interaction group variants (Figure 3, 4, 5, S2).
138 The most common type of interaction group is a single accessory genome region of between
139 3 kb (IG53) and 57 kb (IG41), of which there are 52. For example, IG5 consists of 10 genes
140 in a single block. 9 code for various functions related to carbohydrate metabolism and
141 transport while the 10th is a transcriptional regulator (Figure 3 and Table S2). Four strains
142 have 9 out of the 10 genes but otherwise the genes are either all present or all absent in every
143 strain. Interestingly, there appears to be a difference in frequency between VppAsia isolates
144 and others, with the island present in 52% of VppAsia strains and 90% of others.
145 Genome islands are often associated with transmission mechanisms such as phage and
146 plasmids (Dobrindt et al. 2004). In our data, one example is IG17 which contains 4 genes
147 annotated as being phage related and a further 19 hypothetical proteins (Figure 3). In the
148 reference strain, 16 of the genes occur in a single block VP1563-VP1586, while 7 genes are
149 present elsewhere in the genome. 56% of the 469 had none of the genes, while 36% had more
150 than 12 of them and 8% had between 1 and 5, which presumably represent remnants of an old
151 phage infection. Only one gene (VP1563) in IG17 was found in appreciable frequency in
152 strains that had none of the other genes. This gene might represent cargo of the phage
153 infection that is able to persist for extended periods in the absence of infection due to a useful
154 biological function of its own.
155 Another common interaction is incompatibility between different accessory genome
156 elements, which is found in total 27 of the interaction groups (Figure 2). For example, all of
157 the 469 strains either have the gene yniC, which is annotated as being a phosphorylated

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

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

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

179 Based on the clustering we split the genetic variants themselves into three tiers, shown in red,
180 blue and green in Figure 4. Tier 1 variants distinguish cleanly, or nearly cleanly between
181 EG1.1 and EG1.2 strains. Tier 2 variants are similar but with more discrepancies, while Tier
182 3 variants are typically fixed or nearly fixed within one eco-group and polymorphic within

183 the other. Each of the tiers includes multiple core and accessory genome regions. Some
184 regions are represented in only one tier, while others contain two or more. For example, the
185 lateral flagellar gene cluster (Ref1-23, VPA1538-VPA1557, Figure 4c) contains 383 Tier 1
186 SNPs, 24 from Tier 2 and 47 from Tier 3, however the extent of the region spanned by Tier 1
187 SNPs is smaller, so that for example flhA and flhB, at the center of the gene cluster only
188 contain Tier 3 SNPs, while lafK and lafA to the left and filS to the right contain multiple
189 fixed differences between eco-groups. In other regions of this genome cluster, Tier 1 SNPs
190 are flanked by those from other tiers.
191 IG2 includes 548 SNPs and 130 accessory genes which located in 26 regions (Figure 5a),
192 including a LuxR family transcriptional regulators, T6SS (type VI secretion system), and
193 cellulose synthesis related genes that had been identified in our previous research (Cui et al.
194 2015). Here with increased statistical power with more genome sequences, we found
195 additional interacting loci. Tier 1 contains the T6SS and the cellulose synthesis related genes,
196 reflecting their strong incompatibility, with nearly all EG2.1 strains containing a T6SS and all
197 those in EG2.2 containing genes annotated as part of the biosynthesis cluster. Tier 2 contains
198 mostly SNPs in core genome, which located at LuxR family transcriptional regulator and
199 uridine phosphorylase encoding genes. Tier 3 contains 85% of all variants in IG2, which
200 encoded genes that functional related with biogenesis of elements in cell membrane,
201 carbohydrate transport and metabolism, and transcriptional regulators. Furthermore, the
202 overall strength of the associations in the IG2 Tiers we have defined are much weaker than
203 those for IG1: while there are a handful of core genome regions where EG2.1 regions are
204 differentiated (Figure 5a), none of these is as clear-cut as the differentiated regions of IG1.
205 The loci of IG3 only covered eight genome regions, which contained 75 SNPs and 65
206 accessory genes, most of which are found in a single genome regions which includes the core
207 gene TonB3 (Tanabe et al. 2011)(Figure 5b). Clustering based on variants of IG3 revealed a

208 stair-like structure of variation. Tier 1 variants differentiate EG3.1 from other ecogroups and
209 include multiple SNPs in TonB3 and in the lateral flagellar gene cluster. Tiers 2 and 3
210 variants differentiate between EG3.2, EG3.3 and EG3.4, which are progressively more
211 different from EG3.1, both in terms of accessory genome complement and core genome
212 SNPs. Many of the accessory genes specific to EG3.4 are annotated to have functions
213 associated with cell wall biogenesis.
214
215 **Phenotypic differences between EG1.1 and EG1.2**
216 We performed a preliminary analysis of the phenotypic differences underlying EG1 by
217 determining the motility, growth rate, and biofilm formation ability (Figure 6) of in 7 EG1.1
218 and 4 EG1.2 strains on laboratory media. We failed to observe differences in swimming or
219 swarming capability under the conditions tested but EG1.2 strains revealed significantly
220 higher biofilm formation ability and faster growth rate than EG1.1 strains (Figure 6b and 6c),
221 and they revealed rough colony morphology, also a indication of increased biofilm formation,
222 under low salinity (1% NaCl) culture condition (Figure 6c).
223 There are in total 60 EG1.2 strains in the global collection of 1103 *V. parahaemolyticus*
224 strains. All but one, a VppUS2 isolate, is from the VppAsia population, with the majority (n=
225 48, 80%) in this study coming from routine surveillance on food related environmental
226 samples, including fish, shell fish, and water used for aquaculture. The strains revealed no
227 clear geographical clustering pattern in China, as they can be isolated from all six provinces
228 that under surveillance. Notably, only 4 EG1.2 strains were isolated from clinical samples,
229 including wound and stool, representing a lower proportion than for EG1.1 isolates
230 (453/1043), including if the two pandemic lineages are removed (268/798), suggesting this
231 eco-group has low virulence potential in humans.
232

233 **Discussion**

234 Bacterial traits such as pathogenicity, host-specificity and antimicrobial resistance naturally
235 attract human attention but less obvious or even cryptic traits might be more important in
236 determining the underlying structure of microbial populations. Studies of coadaptation based
237 on genome sequencing of thousands of isolates have the potential to provide new insight into
238 the ecological forces shaping natural diversity, and how that variation is assembled by
239 individual strains to overcome the manifold challenges involved in colonizing specific niches,
240 such as the human gastrointestinal tract. In other words, these studies provide a unique
241 opportunity to see the world from the point of view of a bacterium.

242 We performed a genome wide scan for coadaptation in *V. parahaemolyticus*, performing
243 pairwise tests for interactions amongst genetic variants and then clustered the significant
244 pairwise interactions into 93 interaction groups. Our results highlight the central role of
245 lateral motility in structuring ecologically significant variation within the species. We also
246 find evidence that interactions move through progressive stages, analogous to differing
247 degrees of commitment within human relationships, namely casual, going steady, getting
248 married and moving out together (Figure 7).

249 **Most interactions between core and accessory genomes are casual**

250 A recent debate about whether the accessory genome evolves neutrally (Vos and Eyre-
251 Walker 2017, Andreani, Hesse, and Vos 2017, Shapiro 2017) highlighted how little we know
252 about the functional importance of much of the DNA in bacterial chromosomes. Using the
253 same statistical threshold to assess significance, our interaction screen identifies many more
254 examples of coadaptation between different accessory genome elements than of interactions

255 between the core and accessory genomes or within the core genome, implying that natural
256 selection has a central role in determining accessory genome composition.

257 Unsurprisingly, given the extensive literature highlighting the importance of genomic islands
258 to functional diversity of bacteria (Dobrindt et al. 2004), the most common form of
259 adaptation detected by our screen is the coinheritance of accessory genome elements located
260 in the same region of the genome. Based on a minimum size threshold of 3 kb, we find 52
261 (56%) such interactions, the largest of which is 57 kb (Table S3).

262 Previous approaches to detecting genome islands emphasize traits associated with horizontal
263 transmission, for example based on differences in GC content with the core genome, the
264 presence of phage-related genes or other markers of frequent horizontal transfer (Langille,
265 Hsiao, and Brinkman 2010). Our approach, based simply on co-occurrence identifies a wider
266 range of coinherited units and suggests that many islands have functions related to
267 carbohydrate metabolism.

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

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

278 functional without requiring adaptation of either the element itself or the rest of the bacterial
279 genome.

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

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

291 IG2, is an example of a complex coadaptation involving multiple core and accessory genome
292 regions. A large majority of strains in our dataset (439/469) either carry a cluster of genes
293 encoding a T6SS, or a cluster encoding cellulose biosynthesis genes, but few strains have
294 genes from both clusters (Figure 5a). Cells uses the T6SS to inject toxins into nearby bacteria
295 (Salomon et al. 2013) and cellulose production to coat themselves in a protective layer
296 (Tischler and Camilli 2004). Incompatibility might have a functional basis, for example
297 because cellulose production prevents the T6SS functioning efficiently, or an ecological one,
298 for example because cells that attack others do not need to defend themselves.

299 **Marriage changes everything**

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

308 Many of the accessory-genes and SNPs are in perfect or near perfect disequilibrium with
309 SNPs in the flagellar gene cluster. These include loci encoding flagellar genes, T2SS and
310 other membrane transport elements. There are also 285 loci (27%, Tier 3) in weaker
311 disequilibrium, typically because they are polymorphic in one of the ecogroups. Many of
312 variants are likely to represent more recently evolved coadaptation. Some of these genes are
313 also associated with flagella or the T2SS related function but also encompass a broader range
314 of functional categories, including cell division and amino acid transport and metabolism
315 (Table S2).

316 Our laboratory phenotype experiments (Figure 6) suggest that biofilm formation is likely to
317 be a key trait underlying the different ecological strategies of EG1.1 and EG1.2 but the
318 variation in phenotypic response at different salinity levels and the absence of measurable
319 difference in swarming behavior despite the large genetic difference within the lateral flagella
320 genes highlight some of the manifold difficulties of interpreting natural variation using
321 phenotypes measured under laboratory conditions.

322 Despite the extensive differences that have accumulated between EG1.1 and EG1.2, there is
323 no evidence of restricted gene flow in most of the genome (Figure 4b), and even within the

324 flagellar gene cluster strongly differentiated regions are separated by a weakly differentiated
325 one (Figure 4c), implying that the coadaptation is being maintained by selection in the face of
326 frequent recombination. Initial divergence in flagellar function is likely to have led to
327 ecological differentiation, which led to bacteria having different nutritional inputs or
328 requirements and a broadening of the functional categories undergoing divergent selection.

329 How can the difference between IG1 and the other interaction groups in both the number of
330 associations and their strength be explained? *V. parahaemolyticus* is ubiquitous in shellfish in
331 warm coastal waters, within which it occurs at densities of around 1,000 cells per gram, so a
332 back of the envelope calculation suggests there are likely to be substantially more than 10^{15}
333 bacteria in the VppAsia population. The species also has a high estimated effective
334 population size (Cui et al. 2015, manuscript in preparation) and has strong codon bias, which
335 is often argued to be evidence that even tiny selective coefficients can drive adaptation (Sharp
336 et al. 2005). Furthermore, recombination only breaks up linkage disequilibrium between loci
337 slowly. Therefore, weaker and more variable patterns of association within IG2 and IG3 than
338 in IG1 is unlikely to be a simple consequence of the ineffectiveness of selection and is instead
339 likely to reflect complexity in the fitness landscape.

340 Strains gain flexibility by being able to switch between or modulate genomically encoded
341 strategies by homologous recombination. For example, expression of the T6SS might be
342 essential for survival in crowded habitats but detrimental in sparsely populated ones.
343 Recombinants between EG2.1 and EG2.2 at IG2 loci might represent transient adaptation of
344 strains to their immediate environment or long-term adaptation of intermediate strategies.
345 Furthermore, the phenotypic consequences of IG2 variants can depend on other loci in the
346 genome, such as IG1 variants, which is likely to reduce the strength of associations within
347 IG2. Crucially, the evolution of promiscuity is self-reinforcing because the presence of strains

348 using multiple strategies in the population also favors the presence of accessory genes and
349 core gene haplotypes that are fit on multiple backgrounds.

350 On the other hand, an absence of intermediate genotypes in the population can favor the
351 evolution of fastidiousness, with particular accessory genes and haplotypes becoming
352 essential components of some genetic backgrounds but deleterious on others. A likely
353 scenario is that differences between IG1.1 and IG1.2 at a lateral flagellar gene made
354 recombinants between the two versions of the gene inviable and also created divergent
355 selection at a handful of other loci that was largely independent of the external environment
356 or of interactions with other genes. The evolution of fastidiousness, like the evolution of
357 promiscuity, can be self-reinforcing, and might have led to progressive increase in the
358 differentiation of EG1.2 strains from EG1.1 until the coadaptation of IG1.2 variants to each
359 other and of IG1.1 variants to each other became more-or-less irreversible, like marriage in
360 England prior to the reign of King Henry VIII.

361 **Coadapted gene complexes as speciation triggers**

362 Running the tape forward, it is easy to envisage the number of coadapted regions of the
363 genome within IG1 undergoing progressive enlargement, until the entire genome becomes
364 differentiated. As coadapted regions become more numerous, the proportion of
365 recombination events between ecogroups that are maladaptive will increase, which might
366 prompt the evolution of mechanistic barriers to genetic exchange between them.

367 Mechanisms by which new bacterial species arise are frequently discussed in the literature
368 (Shapiro et al. 2012, Fraser et al. 2009, Falush et al. 2006) but there is currently little data on
369 how the process unfolds. IG1 is of interest both as an example of an intermediate stage of
370 divergence, prior to speciation, and because it suggests that substantial adaptive divergence

371 between gene pools can precede any barriers to genetic exchange, other than natural selection
372 at the loci involved. This – unique to our knowledge – example is exciting because the
373 distinct signature of selection should make it possible to dissect the genetic basis of
374 coadaptation in unprecedented detail. Broadly similar patterns of differentiation including
375 "genomic islands of speciation" have been observed for example between ecomorphs of
376 cichlid fishes (Malinsky et al. 2015), but the evolution of ecomorphs has been facilitated by
377 fish preferring to mate with similar individuals, which will have also inevitably lead to some
378 level of differentiation at neutral loci throughout the genome.

379 **Conclusions**

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

386 Most bacteria have population structure that deviates more markedly from panmixia
387 (manuscript in preparation). In some species this is likely due to smaller effective population
388 sizes, lower recombination rates or mechanistic barriers to genetic exchange between strains.
389 However, coadaptation can itself generate genome-wide linkage disequilibrium that might be
390 difficult to distinguish from clonal or population structure. Because the linkage
391 disequilibrium associated with IG1 is highly localized within the genome, it can, on careful
392 inspection be clearly be attributed to selection, but in other bacteria patterns are likely to be
393 less straightforward, making it challenging to understand to whether adaptive processes drive
394 population structure, or vice versa. Natural selection is the jewel of evolution but

395 distinguishing it from other processes requires in depth understanding of the relevant biology
396 in addition to suitable data and statistical methods.

397

398 **Materials and Methods**

399 **Genomes used in this work**

400 Totally 1,103 global *V. parahaemolyticus* genomes were used in this work (Table S1), which
401 also were analyzed in our another study (Yang et al. 2018). To reduce clonal signals, we
402 firstly made a "non-redundancy" dataset of 469 strains, in which no sequence differed by less
403 than 2,000 SNPs in the core genome. They were attributed to 4 populations, VppAsia (383
404 strains), VppX (43), VppUS1 (18) and VppUS2 (21) based on fineSTRUCTURE result
405 (Lawson et al. 2012). We then focused on VppAsia which has more strains, to generate a
406 genome dataset in which strains represent a freely recombining population. We selected 386
407 genomes from 469 non-redundancy genome dataset, including all the 383 VppAsia genomes
408 and 3 outgroup genomes which were randomly selected from VppX, VppUS1 and VppUS2
409 population, respectively. These 386 genomes were used in Chromosome painting and
410 fineSTRUCTURE analysis (Lawson et al. 2012) as previously described (Cui et al. 2015).
411 Initial fineSTRUCTURE result revealed multiple clonal signals still exist, thus we selected
412 one representative genome from each clone, combined them with the remaining genomes and
413 repeated the process. After 14 iterations, we got a final dataset of 201 genomes with no trace
414 of clonal signals, involving 198 VppAsia genomes that were used in further analysis.
415 The copy probability value of each strain at each SNP were generated by Chromosome
416 painting with "-b" option, and the average copying probability value of a given strain group
417 (e.g. EG1.2) at each SNP was used in Figure 4, 5 and Figure S4.

418 **Variation detection, annotation and phylogeny**

419 We re-called SNPs from 198 VppAsia genomes by aligning the assembly against reference
420 genome (RIMD 2210633) using Mummer (Delcher, Salzberg, and Phillippy 2003) as
421 previous described (Cui et al. 2015). Totally 565,466 bi-allelic SNPs were identified and
422 151,957 bi-allelic SNPs with minor allele frequency > 2% were used in coadaptation
423 detection. We re-annotated all the assemblies using Prokka (Seemann 2014), and the
424 annotated results were used in Roray (Page et al. 2015) to identify the pan-genome and gene
425 presence/absence, totally 41,053 pan-genes were found and 14,486 accessory genes (present
426 in > 2% and < 98% strains) were used in coadaptation detection. The pan-gene protein
427 sequences of Roary were used to blast (blastp) against COG and KEGG database to get
428 further annotation.
429 The Neighbour-joining trees were built by using the TreeBest software
430 (http://treesoft.sourceforge.net/treebest.shtml) based on sequences of concatenated SNPs, and
431 were visualized by using online tool iTOL (Letunic and Bork 2016).
432 **Detection of coadapted loci**
433 Totally 151,957 bi-allelic SNPs and 14,486 accessory genes identified from 198 independent
434 VppAsia genomes were used in coadaptation detection by two methods. Firstly we used
435 Fisher exact test to detect the linkage disequilibrium of each SNP-SNP, SNP-accessory gene,
436 and accessory gene-gene pair. Presence or absence of an accessory gene was considered as its
437 two alleles. Each variant locus (SNP or accessory gene) has 2 alleles, major and minor, of
438 which major represents the allele shared by majority of isolates. For each pair of loci X and
439 Y, the number of combinations between X_{major}-Y_{major}, X_{major}-Y_{minor}, X_{minor}-Y_{major}, X_{minor}-Y_{minor}
440 were separately counted and used in the contingency table to calculate the Fisher exact test *P*
441 value. It took 3 days to finish all the coadaptation detection in a computer cluster using 21
442 cores and 2 Gb memory. We also used SuperDCA (Puranen et al. 2018) to detect the
443 coadaptation between SNPs for comparison. It is based on direct coupling analysis (DCA)

444 model and has a much faster calculation speed compared with previous DCA methods.
445 However, it still took 25 days to finish the detection by using 32 cores and 86 Gb memory.
446 We removed coadaptation pairs with distance less than 3 kb to minimize the influence of
447 physical linkage. All identified SNPs in this study were located in the core genome, therefore
448 the physical distance between SNP pairs can be calculated according to their position in the
449 reference genome. To defined the distance between accessory genes, and between SNP and
450 accessory gene, we mapped the sequence of accessory genes against available 19 complete
451 maps of the *V. parahaemolyticus* genomes to acquire their corresponding position, and then
452 the gene that failed to be found in complete reference genomes were then mapped to the draft
453 genomes. If the accessory genes pair or SNP- accessory gene pair was found located in a
454 same chromosome or same contig of a draft genome, then the distance between paired
455 variants could be counted according to their position in the chromosome or contig. The
456 distance between paired variants that located in different chromosome or contig was counted
457 as larger than 3 kb and such pairs were kept in further analysis. Circos (Krzywinski and
458 Schein 2009) was used to visualize the networks of coadaptation SNPs in Figure 1f and
459 Figure S3.

460 **Lateral flagellar gene cluster region in *Vibrio* genus**

461 To identify the homologous sequences of *V. parahaemoluticus* lateral flagellar gene cluster
462 (VPA1538-1557) in the *Vibrio* genus, we downloaded all available *Vibrio* genome sequences
463 in NCBI, then aligned the nucleotide sequence of lateral flagellar gene cluster of *V.*
464 *parahaemolyticus* (NC_004605 1639906-1657888) against *Vibrio* genome dataset
465 (excluding *V. parahaemolyticus*) by using blastn. Totally 46 *Vibrio* genomes revealed above
466 than 60% coverage on lateral flagella region in *V. parahaemolyticus* genome and was used in
467 phylogeny rebuilding. We also included three randomly selected strains from EG1.1 and

468 EG1.2 respectively for comparison. In total 3,000 SNPs were identified in this region and
469 were used for NJ tree construction.

470 **Determination of phenotypes**

471 **Bacteria strains.** In the phenotype experiments, totally 11 strains were randomly selected
472 respectively from two EGs that defined by IG1 variants, including 7 EG1.1 strains (B1_10,
473 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).
474 The strains stored at -80℃ were inoculated in the thiosulfate citrate bile salts sucrose agar
475 (TCBS) plates by streak plate method. Five clones for each strains were inoculated again in
476 another TCBS plate and then cultured overnight at 30℃ in 3% NaCl-LB broth overnight and
477 used for the following assays.

478 **Motility assays.** Five clones for each strain were cultured overnight at 30 ℃ and then
479 inoculated in the swimming plate (LB media containing 0.3% agar) and swarming plate (LB
480 agar with 3% NaCl). The swimming ability was recorded by measuring the diameter of
481 colony after 24 hours at 30 ℃. And the swarming ability was recorded after 72 hours at
482 24 ℃.

483 **Growth curve.** *V. parahaemolyticus* strains in 96-well plate were cultured overnight at 30°C
484 in 3% NaCl-LB broth. The optical density of each culture was adjust to an OD_{600} of 0.6.Then
485 1 ml of each culture was inoculated 100 ml of 3% NaCl-LB broth in a 96-well plate and
486 cultured at 30 °C. The growth of each culture were measured every 1 hour at the optical
487 density of 600 nm using Multiskan Spectrum.

488 **Biofilm formation.** *V. parahaemolyticus strains* were cultured overnight at 30°C in 3%
489 NaCl-LB broth. 2 μl of each overnight culture was inoculated to 100 μl of 3% NaCl-LB broth
490 in a 96-well plate and cultured at 30°C for 24 h statically. The supernatant was discarded and
491 each well was washed once with sterile phosphate-buffered saline (PBS). 0.1% Crystal violet

492 (wt/vol) was added to each well and incubated at room temperature for 30 min. The crystal
493 violet was decanted, and each well was washed once with sterile PBS. Crystal violet that
494 stained biofilm was solubilized with dimethylsulfoxide (DMSO), and then measured at the
495 optical density of 595 nm using Multiskan Spectrum (Thermo Scientific).
496

497 **Acknowledgements**

498 This work is supported by the National Key Research & Development Program of China (No.
499 2016YFC1200100, 2017YFC1200800, and 2017YFC1601503), Sanming Project of
500 Medicine in Shenzhen (No. SZSM201811071), the National Natural Science Foundation of
501 China (No. 31770001) and the Key Research Program of the Chinese Academy of Sciences
502 (No. ZDRW-ZS-2017-1). D.F. is funded by a Medical Research Council Fellowship as part
503 of the MRC CLIMB consortium for microbial bioinformatics (grant number
504 MR/M501608/1).
505

506 **Author Contributions**

507 D. F., Y. C., and R. Y. designed the study and coordinated the project; Y. C., C. Y., and D. F.
508 analyzed the data; H. Q. and H. W. performed phenotype experiments; D. F. and Y. C. wrote
509 the manuscript. All authors approved the final version of the manuscript.
510

511 **Competing Financial Interests statement**

512 No
513
514
515

516 **References**

517 Andreani, N. A., E. Hesse, and M. Vos. 2017. "Prokaryote genome fluidity is dependent on effective
518 population size." *ISME J* 11 (7):1719-1721. doi: 10.1038/ismej.2017.36.
519 Arnold, B. J., M. U. Gutmann, Y. H. Grad, S. K. Sheppard, J. Corander, M. Lipsitch, and W. P. Hanage.
520 2018. "Weak Epistasis May Drive Adaptation in Recombining Bacteria." *Genetics* 208
521 (3):1247-1260. doi: 10.1534/genetics.117.300662.
522 Babu, M., J. J. Diaz-Mejia, J. Vlasblom, A. Gagarinova, S. Phanse, C. Graham, F. Yousif, H. Ding, X.
523 Xiong, A. Nazarians-Armavil, M. Alamgir, M. Ali, O. Pogoutse, A. Pe'er, R. Arnold, M. Michaut,
524 J. Parkinson, A. Golshani, C. Whitfield, S. J. Wodak, G. Moreno-Hagelsieb, J. F. Greenblatt,
525 and A. Emili. 2011. "Genetic interaction maps in Escherichia coli reveal functional crosstalk

among cell envelope biogenesis pathways." *PLoS Genet* 7 (11):e1002377. doi: 10.1371/journal.pgen.1002377.

Campo, D. S., Z. Dimitrova, R. J. Mitchell, J. Lara, and Y. Khudyakov. 2008. "Coordinated evolution of the hepatitis C virus." *Proc Natl Acad Sci U S A* 105 (28):9685-90. doi: 10.1073/pnas.0801774105.

Costanzo, M., B. VanderSluis, E. N. Koch, A. Baryshnikova, C. Pons, G. Tan, W. Wang, M. Usaj, J. Hanchard, S. D. Lee, V. Pelechano, E. B. Styles, M. Billmann, J. van Leeuwen, N. van Dyk, Z. Y. Lin, E. Kuzmin, J. Nelson, J. S. Piotrowski, T. Srikumar, S. Bahr, Y. Chen, R. Deshpande, C. F. Kurat, S. C. Li, Z. Li, M. M. Usaj, H. Okada, N. Pascoe, B. J. San Luis, S. Sharifpoor, E. Shuteriqi, S. W. Simpkins, J. Snider, H. G. Suresh, Y. Tan, H. Zhu, N. Malod-Dognin, V. Janjic, N. Przulj, O. G. Troyanskaya, I. Stagljar, T. Xia, Y. Ohya, A. C. Gingras, B. Raught, M. Boutros, L. M. Steinmetz, C. L. Moore, A. P. Rosebrock, A. A. Caudy, C. L. Myers, B. Andrews, and C. Boone. 2016. "A global genetic interaction network maps a wiring diagram of cellular function." *Science* 353 (6306). doi: 10.1126/science.aaf1420.

Cui, Y., X. Yang, X. Didelot, C. Guo, D. Li, Y. Yan, Y. Zhang, Y. Yuan, H. Yang, J. Wang, J. Wang, Y. Song, D. Zhou, D. Falush, and R. Yang. 2015. "Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen Vibrio parahaemolyticus." *Mol Biol Evol* 32 (6):1396-410. doi: 10.1093/molbev/msv009.

Delcher, A. L., S. L. Salzberg, and A. M. Phillippy. 2003. "Using MUMmer to identify similar regions in large sequence sets." *Curr Protoc Bioinformatics* Chapter 10:Unit 10 3. doi: 10.1002/0471250953.bi1003s00.

Dobrindt, U., B. Hochhut, U. Hentschel, and J. Hacker. 2004. "Genomic islands in pathogenic and environmental microorganisms." *Nat Rev Microbiol* 2 (5):414-24. doi: 10.1038/nrmicro884.

Falush, D., M. Torpdahl, X. Didelot, D. F. Conrad, D. J. Wilson, and M. Achtman. 2006. "Mismatch induced speciation in Salmonella: model and data." *Philos Trans R Soc Lond B Biol Sci* 361 (1475):2045-53. doi: 10.1098/rstb.2006.1925.

Figliuzzi, M., H. Jacquier, A. Schug, O. Tenaillon, and M. Weigt. 2016. "Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1." *Mol Biol Evol* 33 (1):268-80. doi: 10.1093/molbev/msv211.

Fraser, C., E. J. Alm, M. F. Polz, B. G. Spratt, and W. P. Hanage. 2009. "The bacterial species challenge: making sense of genetic and ecological diversity." *Science* 323 (5915):741-6. doi: 10.1126/science.1159388.

Gibson, B., D. J. Wilson, E. Feil, and A. Eyre-Walker. 2018. "The distribution of bacterial doubling times in the wild." *Proc Biol Sci* 285 (1880). doi: 10.1098/rspb.2018.0789.

Jerison, E. R., and M. M. Desai. 2015. "Genomic investigations of evolutionary dynamics and epistasis in microbial evolution experiments." *Curr Opin Genet Dev* 35:33-9. doi: 10.1016/j.gde.2015.08.008.

Kryazhimskiy, S., J. Dushoff, G. A. Bazykin, and J. B. Plotkin. 2011. "Prevalence of epistasis in the evolution of influenza A surface proteins." *PLoS Genet* 7 (2):e1001301. doi: 10.1371/journal.pgen.1001301.

Krzywinski, M, and JBirol I Schein. 2009. "Circos: an information aesthetic for comparative genomics." *Genome Research* 19 (9):1639-1645.

Kuzmin, E., B. VanderSluis, W. Wang, G. Tan, R. Deshpande, Y. Chen, M. Usaj, A. Balint, M. Mattiazzi Usaj, J. van Leeuwen, E. N. Koch, C. Pons, A. J. Dagilis, M. Pryszlak, J. Z. Y. Wang, J. Hanchard, M. Riggi, K. Xu, H. Heydari, B. J. San Luis, E. Shuteriqi, H. Zhu, N. Van Dyk, S. Sharifpoor, M. Costanzo, R. Loewith, A. Caudy, D. Bolnick, G. W. Brown, B. J. Andrews, C. Boone, and C. L. Myers. 2018. "Systematic analysis of complex genetic interactions." *Science* 360 (6386). doi: 10.1126/science.aao1729.

Langille, M. G., W. W. Hsiao, and F. S. Brinkman. 2010. "Detecting genomic islands using bioinformatics approaches." *Nat Rev Microbiol* 8 (5):373-82. doi: 10.1038/nrmicro2350.

576 Lawson, D. J., G. Hellenthal, S. Myers, and D. Falush. 2012. "Inference of population structure using
577 dense haplotype data." *PLoS Genet* 8 (1):e1002453. doi: 10.1371/journal.pgen.1002453.
578 Letunic, I., and P. Bork. 2016. "Interactive tree of life (iTOL) v3: an online tool for the display and
579 annotation of phylogenetic and other trees." *Nucleic Acids Res* 44 (W1):W242-5. doi:
580 10.1093/nar/gkw290.
581 Makino, K., K. Oshima, K. Kurokawa, K. Yokoyama, T. Uda, K. Tagomori, Y. Iijima, M. Najima, M.
582 Nakano, A. Yamashita, Y. Kubota, S. Kimura, T. Yasunaga, T. Honda, H. Shinagawa, M.
583 Hattori, and T. Iida. 2003. "Genome sequence of Vibrio parahaemolyticus: a pathogenic
584 mechanism distinct from that of V cholerae." *Lancet* 361 (9359):743-9. doi: 10.1016/S0140-
585 6736(03)12659-1.
586 Malinsky, M., R. J. Challis, A. M. Tyers, S. Schiffels, Y. Terai, B. P. Ngatunga, E. A. Miska, R. Durbin, M.
587 J. Genner, and G. F. Turner. 2015. "Genomic islands of speciation separate cichlid ecomorphs
588 in an East African crater lake." *Science* 350 (6267):1493-1498. doi: 10.1126/science.aac9927.
589 Page, A. J., C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. Holden, M. Fookes, D. Falush, J. A.
590 Keane, and J. Parkhill. 2015. "Roary: rapid large-scale prokaryote pan genome analysis."
591 *Bioinformatics* 31 (22):3691-3. doi: 10.1093/bioinformatics/btv421.
592 Puranen, S., M. Pesonen, J. Pensar, Y. Y. Xu, J. A. Lees, S. D. Bentley, N. J. Croucher, and J. Corander.
593 2018. "SuperDCA for genome-wide epistasis analysis." *Microb Genom*. doi:
594 10.1099/mgen.0.000184.
595 Salomon, D., H. Gonzalez, B. L. Updegraff, and K. Orth. 2013. "Vibrio parahaemolyticus type VI
596 secretion system 1 is activated in marine conditions to target bacteria, and is differentially
597 regulated from system 2." *PLoS One* 8 (4):e61086. doi: 10.1371/journal.pone.0061086.
598 Seemann, T. 2014. "Prokka: rapid prokaryotic genome annotation." *Bioinformatics* 30 (14):2068-9.
599 doi: 10.1093/bioinformatics/btu153.
600 Shapiro, B. J. 2017. "The population genetics of pangenomes." *Nat Microbiol* 2 (12):1574. doi:
601 10.1038/s41564-017-0066-6.
602 Shapiro, B. J., J. Friedman, O. X. Cordero, S. P. Preheim, S. C. Timberlake, G. Szabo, M. F. Polz, and E.
603 J. Alm. 2012. "Population genomics of early events in the ecological differentiation of
604 bacteria." *Science* 336 (6077):48-51. doi: 10.1126/science.1218198.
605 Sharp, P. M., E. Bailes, R. J. Grocock, J. F. Peden, and R. E. Sockett. 2005. "Variation in the strength of
606 selected codon usage bias among bacteria." *Nucleic Acids Res* 33 (4):1141-53. doi:
607 10.1093/nar/gki242.
608 Skwark, M. J., N. J. Croucher, S. Puranen, C. Chewapreecha, M. Pesonen, Y. Y. Xu, P. Turner, S. R.
609 Harris, S. B. Beres, J. M. Musser, J. Parkhill, S. D. Bentley, E. Aurell, and J. Corander. 2017.
610 "Interacting networks of resistance, virulence and core machinery genes identified by
611 genome-wide epistasis analysis." *PLoS Genet* 13 (2):e1006508. doi:
612 10.1371/journal.pgen.1006508.
613 Tanabe, T., T. Funahashi, N. Okajima, H. Nakao, Y. Takeuchi, K. Miyamoto, H. Tsujibo, and S.
614 Yamamoto. 2011. "The Vibrio parahaemolyticus pvuA1 gene (formerly termed psuA)
615 encodes a second ferric vibrioferrin receptor that requires tonB2." *FEMS Microbiol Lett* 324
616 (1):73-9. doi: 10.1111/j.1574-6968.2011.02389.x.
617 Tenaillon, O., A. Rodriguez-Verdugo, R. L. Gaut, P. McDonald, A. F. Bennett, A. D. Long, and B. S.
618 Gaut. 2012. "The molecular diversity of adaptive convergence." *Science* 335 (6067):457-61.
619 doi: 10.1126/science.1212986.
620 Tischler, A. D., and A. Camilli. 2004. "Cyclic diguanylate (c-di-GMP) regulates Vibrio cholerae biofilm
621 formation." *Mol Microbiol* 53 (3):857-69. doi: 10.1111/j.1365-2958.2004.04155.x.
622 Vos, M., and A. Eyre-Walker. 2017. "Are pangenomes adaptive or not?" *Nat Microbiol* 2 (12):1576.
623 doi: 10.1038/s41564-017-0067-5.
624 Yang, Chao, Xiaoyan Pei, Yarong Wu, Lin Yan, Yanfeng Yan, Yuqin Song, Nicola Coyle, Jaime Martinez-
625 Urtaza, Christopher Quince, Qinghua Hu, Min Jiang, Edward Feil, Dajin Yang, Yajun Song,

626 Dongsheng Zhou, Ruifu Yang, Daniel Falush, and Yujun Cui. 2018. "Recent mixing of Vibrio
627 parahaemolyticus populations." *bioRxiv*. doi: 10.1101/354761.

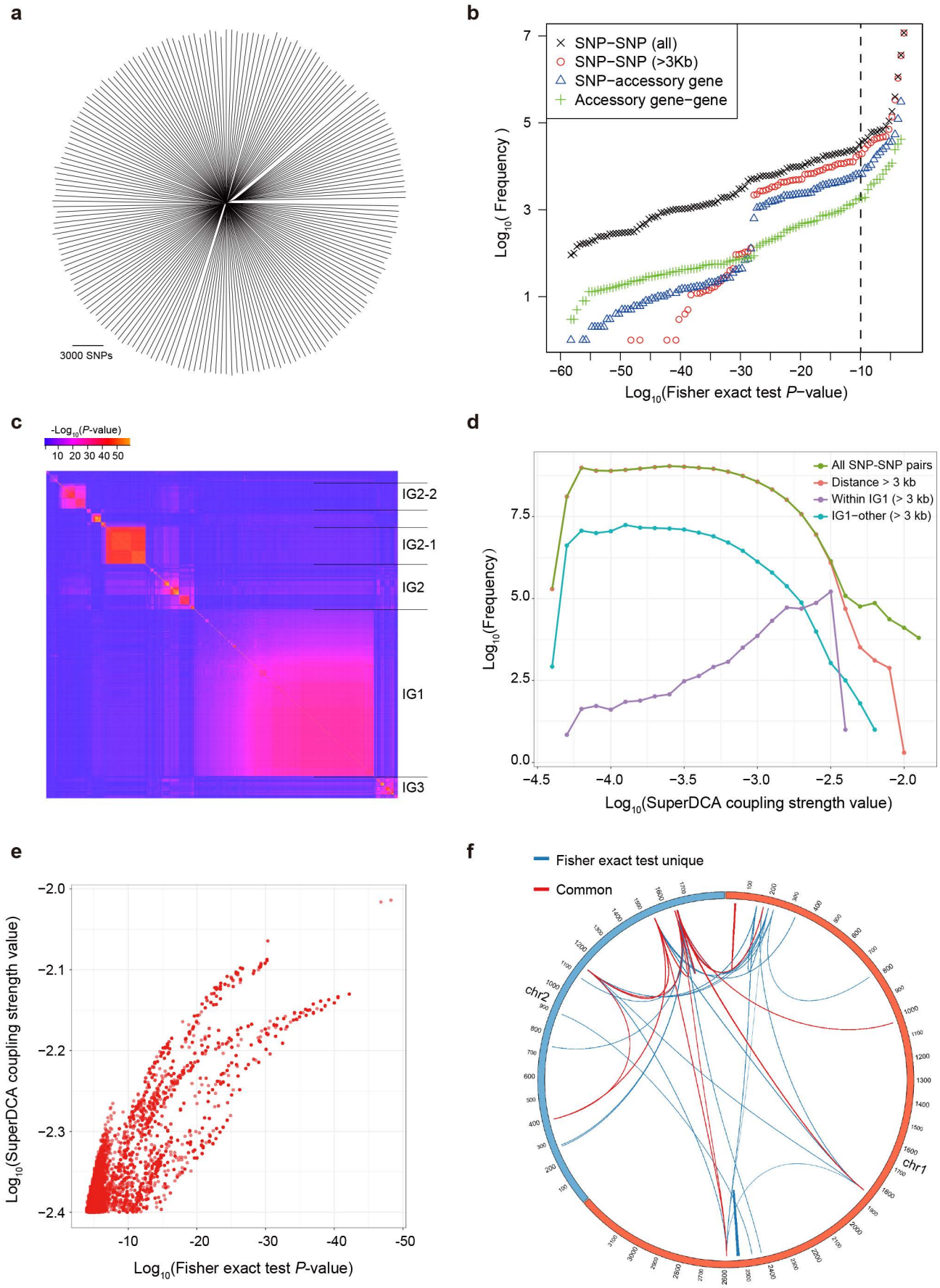


Figure 1. Detection of coadaptation loci in *V. parahaemolyticus*. (a) NJ tree of 198 VppAsia strains based on 151,957 bi-allelic SNPs (minor allele frequency > 2%) (b) Q-Q plot of Fisher exact test *P* values between genetic variants. The vertical dotted line shows the threshold $P = 10^{-10}$. (c)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

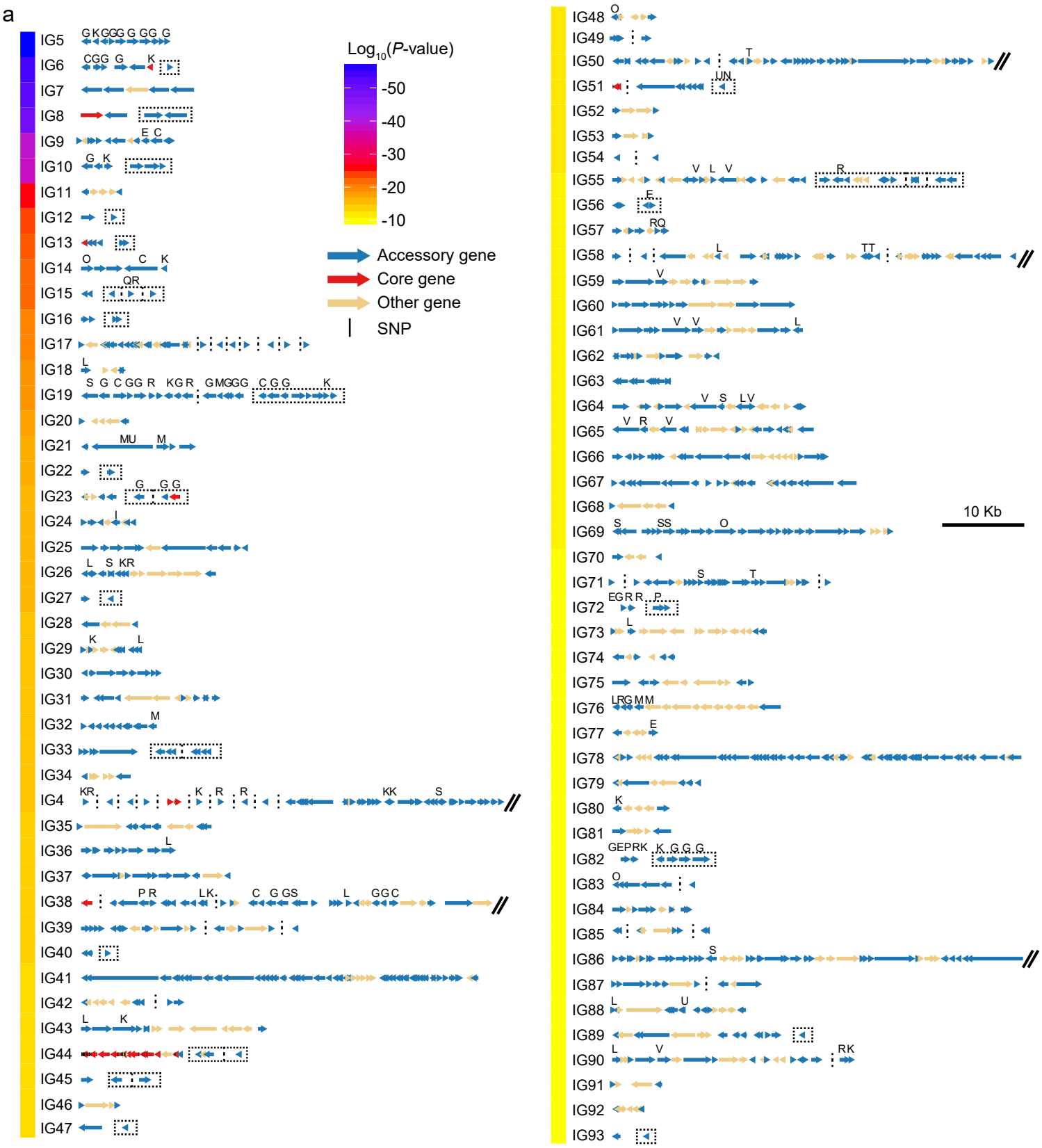
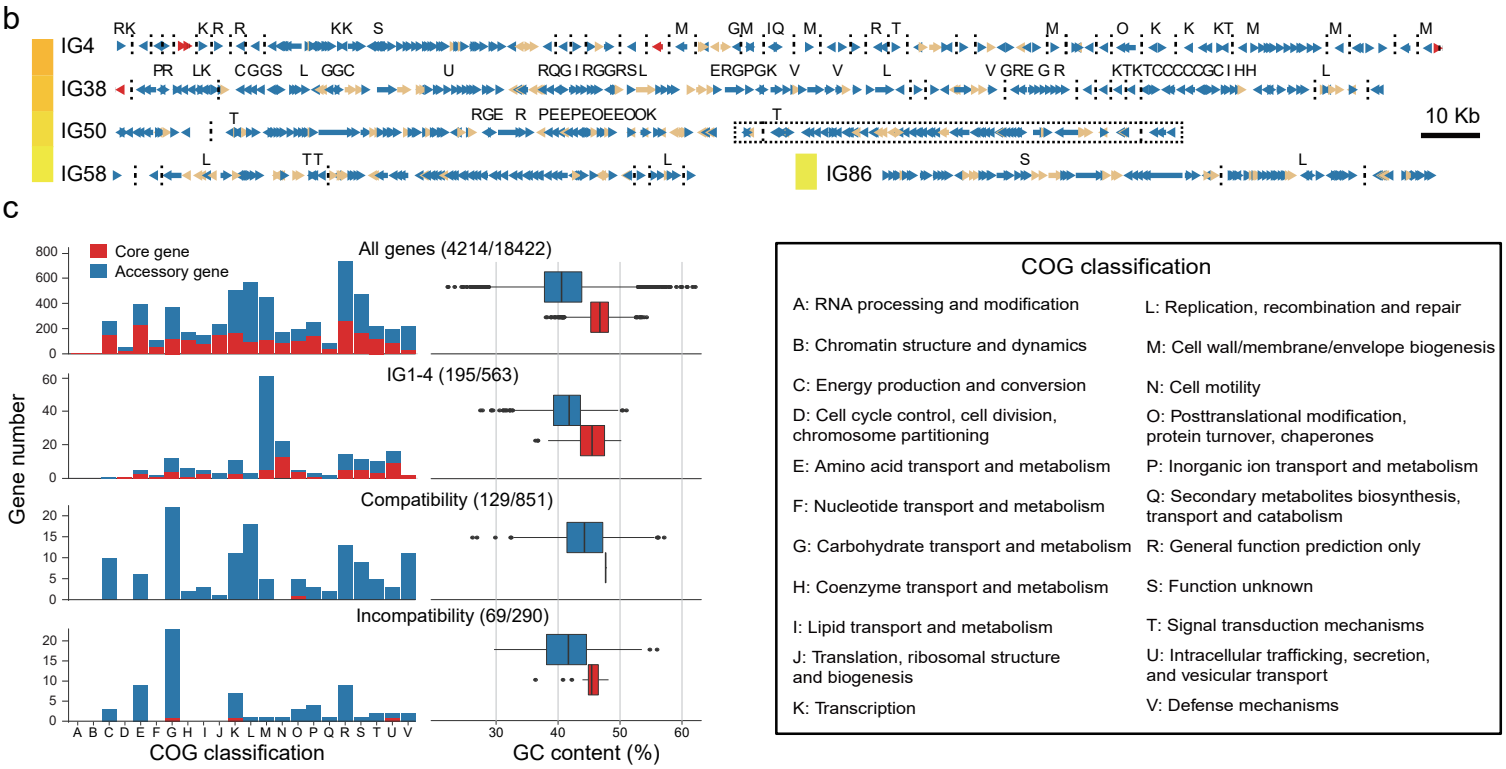


Figure 2. Landscape of coadaptation (excluding IG1-3). (a) Gene maps of different IGs. Arrows indicate genes and red for detected coadaptation core genes, blue for accessory genes and orange for genes with no coadaptation signal. Black vertical lines indicate SNPs in the interaction group. The colors 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 2 continued



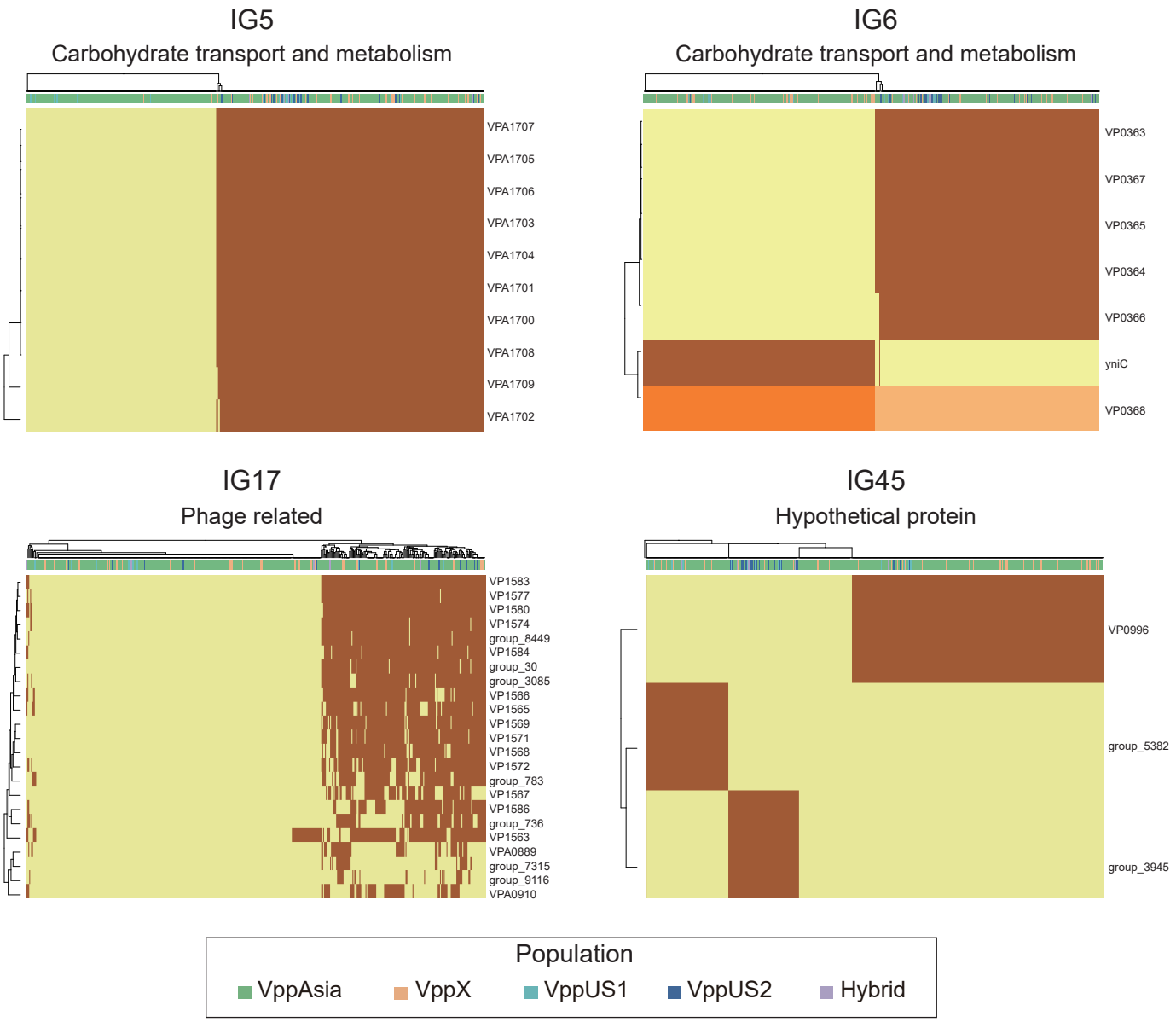


Figure 3. Representative interaction groups. Hierarchical clustering of 469 non-redundancy strains (columns) based on coadaptation loci (rows) of 4 representative IGs. Colors of the heatmap indicate the status of genetic variants, with light orange/orange for two alleles of a SNP, light yellow/brown 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.

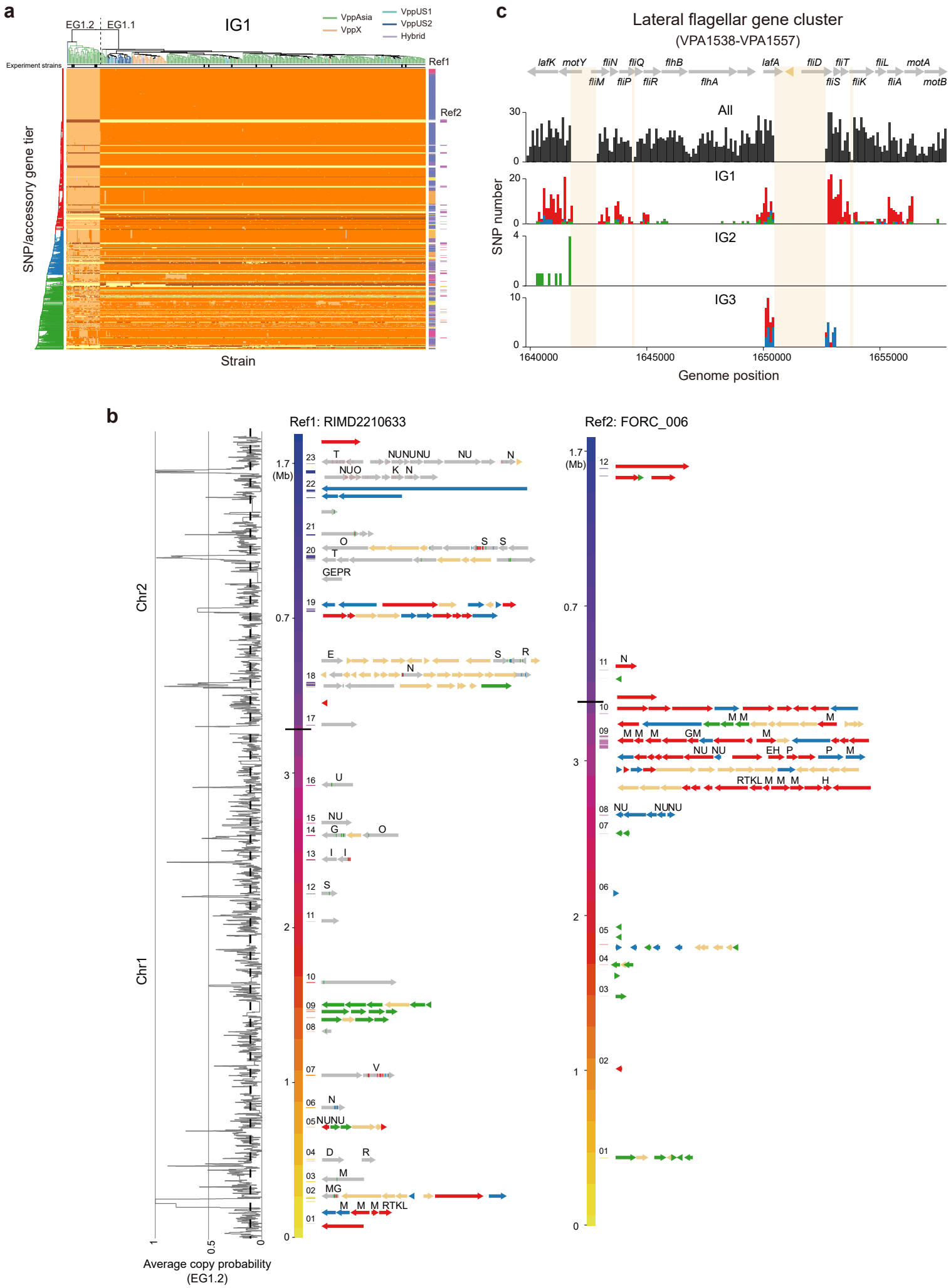


Figure 4. Interaction group 1.

Figure 4 continued

Figure 4. Interaction group 1. (a) Hierarchical clustering of 469 non-redundancy strains (columns) based on coadaptated loci (rows) of IG1. Color scheme of the heatmap is the same as in Figure 3. Branch colors of the tree on the top indicate the populations of strains according to the legend. Branch colors of the tree on the left indicate the tiers of coadaptation loci, with red, blue and green for tier 1, 2 and 3, respectively. The same colors are used for tiers of coadaptation in panel (b) and (c). Colors of the bar on the right indicate the genome position of coadaptation loci, which is corresponding to the bar colors in panel (b). (b) Gene map of coadaptation genome blocks of IG1. Two reference genomes were used to show coadaptation variants in IG1. The left curve indicates the value of average copy probability of EG1.2 strains that copied from themselves throughout the genome of Ref1, with vertical dotted line indicates the expect value (number of EG1.2 strains /number of all strains). The bars indicate the reference genomes and different chromosomes are separated by a horizontal short bar. 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. (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.

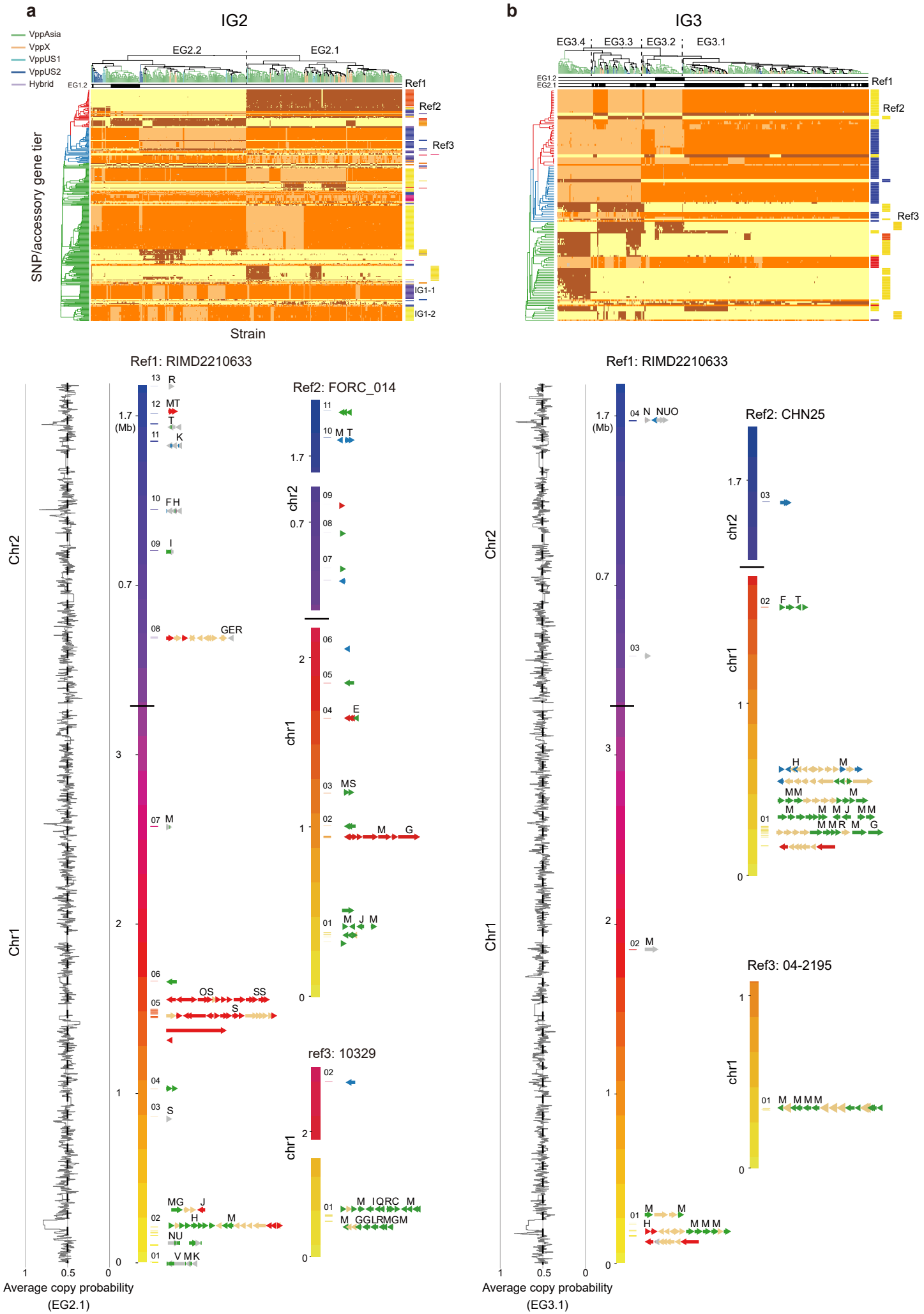


Figure 5. Interaction group 2 (a) and 3 (b). Layout and colors are the same as in Figure 4. EG1.2 and EG2.1 strains are shown below the tree on the top for comparison (black bars). The left curves indicate the average copy probability value of EG2.1 (a) and EG3.1 (b) strains copy from themselves throughout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

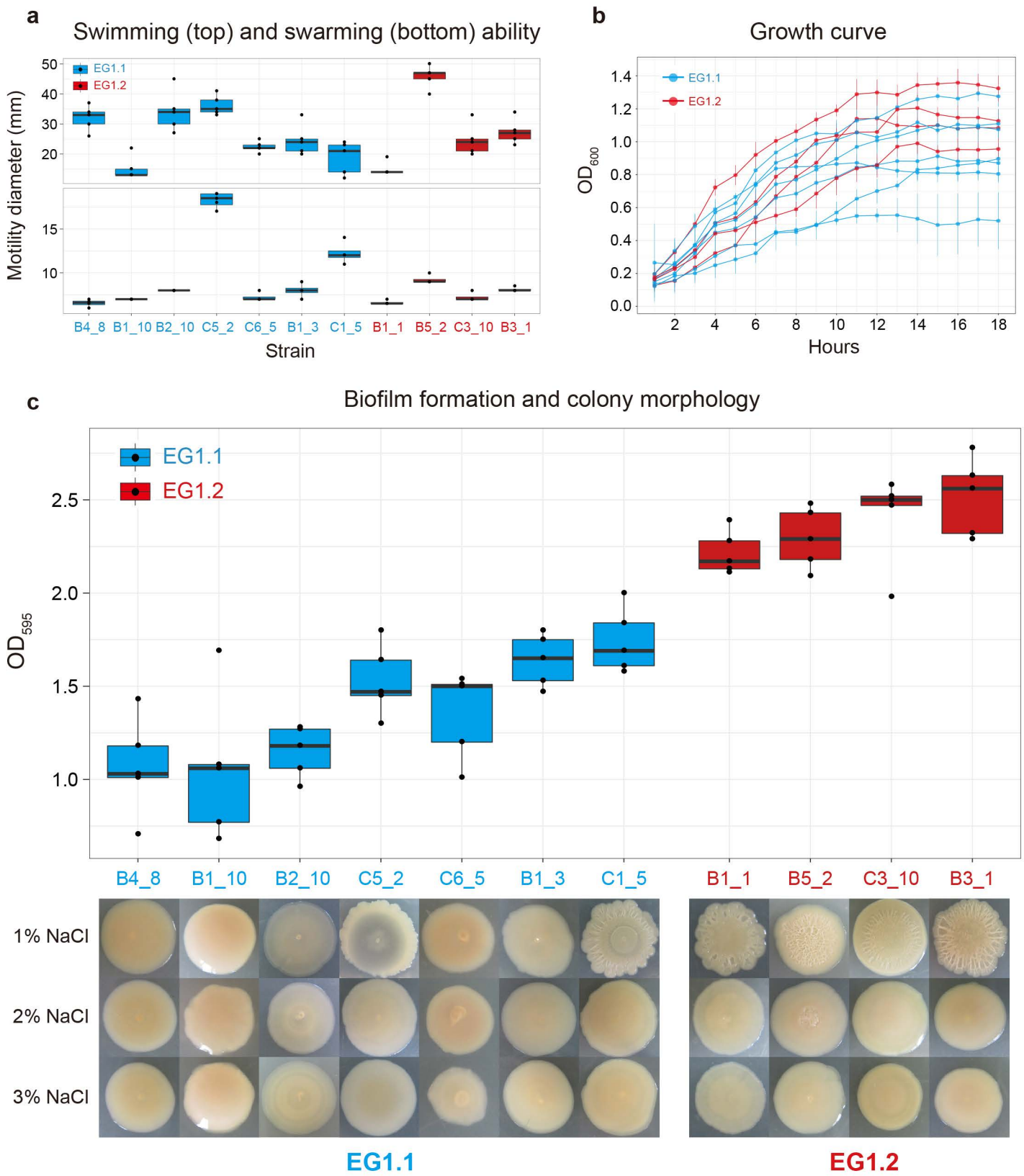


Figure 6. Phenotypes of 7 EG1.1 and 4 EG1.2 strains. (a) Swimming and swarming ability. (b) Growth curve. (c) Biofilm formation and colony morphology. Strains used in experiments were randomly selected and were marked in Figure 4a.

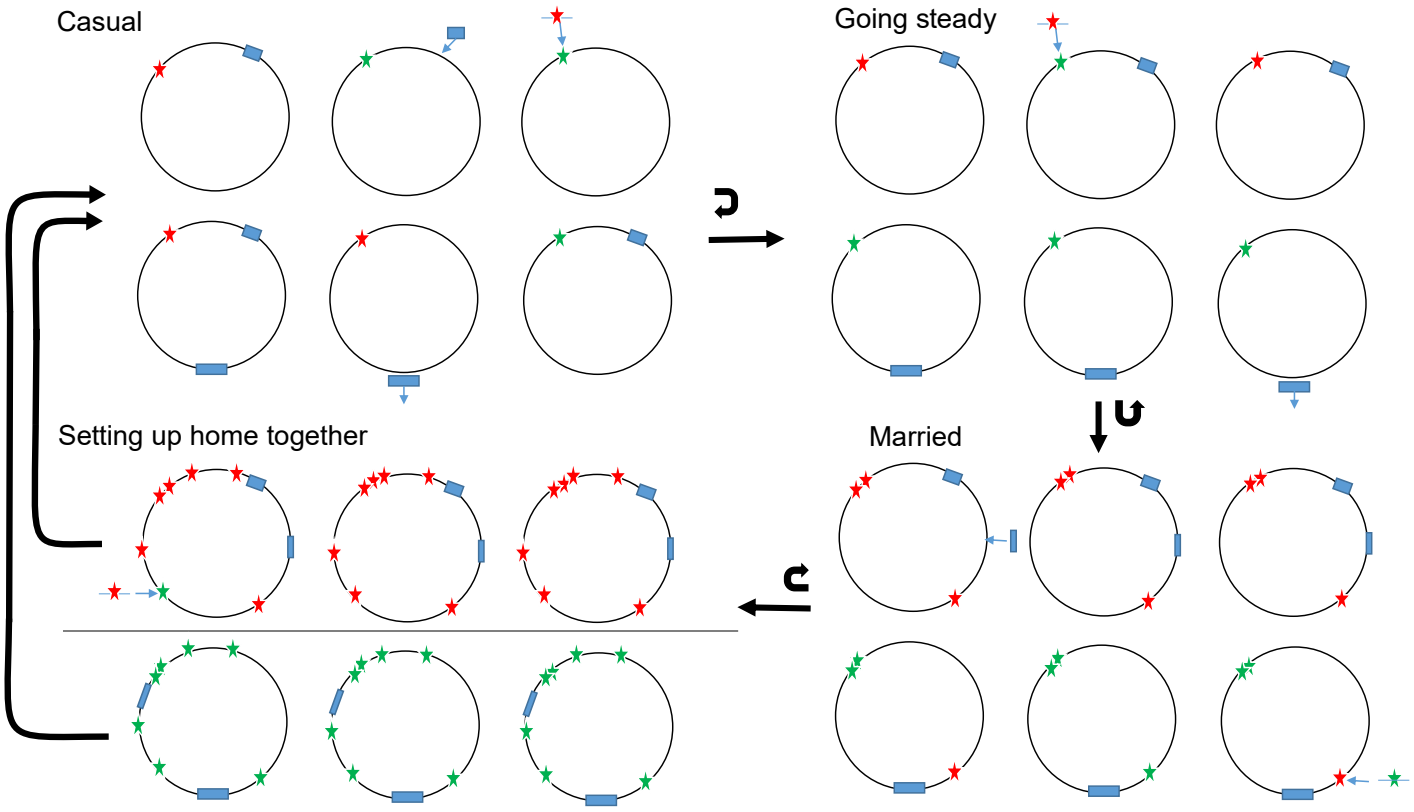


Figure 7. Overview of four stages of coadaptation. Circles indicate bacterial strains within a population. Stars indicate SNPs, with red and green indicating the two alleles. Blue rectangles indicate accessory genes or genome islands. Arrows indicate the transitions between stages. Casual interactions 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*	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.