**Article**

# Recombination facilitates adaptive evolution in rhizobial soil bacteria

Maria Izabel A. Cavassim[1,2,+,*], Stig U. Andersen[2], Thomas Bataillon[1], and Mikkel Heide Schierup[1,*]

**Author affiliations:**

[1] Bioinformatics Research Centre, Aarhus University, Aarhus, 8000, Denmark

[2] Department of Molecular Biology and Genetics, Aarhus University, Aarhus, 8000, Denmark

[+]Present address: Department of Ecology and Evolutionary Biology, University of California, Los Angeles, 90095, United States

**[*]Corresponding authors:**

Maria Izabel A. Cavassim, izabelcavassim@gmail.com

Mikkel H. Schierup, mheide@birc.au.dk

**Author Contributions**

**Conceptualization:** M.I.A.C., T.B., and M.H.S.; **Methodology:** M.I.A.C., T.B., and M.H.S.; **Formal Analysis:** M.I.A.C.; **Investigation:** M.I.A.C., T.B., and M.H.S.; **Resources:** S.U.A., and M.H.S.; **Data curation:** M.I.A.C., T.B., and M.H.S.; **Writing - Original Draft:** M.I.A.C.; **Writing - Review and Editing:** M.I.A.C., S.U.A., T.B., and M.H.S.; **Visualization:** M.I.A.C., T.B, and M.H.S. **Supervision:** T.B. and M.H.S.; **Project administration:** S.U.A. and M.H.S.; **Funding acquisition:** S.U.A. and M.H.S.

**Competing interest**

The authors declare that they have no competing interests.

**Classification:** Biological Sciences and Evolution

**Keywords:** Adaptive evolution, rhizobium, recombination, beneficial mutations

**This PDF file includes:**

**Main Text**

**Figures 1 to 3**

**Tables 1 to 2**

# Abstract

Homologous recombination is expected to increase natural selection efficacy by decoupling the fate of beneficial and deleterious mutations and by readily creating new combinations of beneficial alleles. Here, we investigate how the proportion of amino acid substitutions fixed by adaptive evolution ($\alpha$) depends on the recombination rate in bacteria. We analyze 3086 core protein-coding sequences from 196 genomes belonging to five closely-related species of the genus *Rhizobium*. These genes are found in all species and do not display any signs of introgression between species. We estimate $\alpha$ using the site frequency spectrum (SFS) and divergence data for all pairs of species. We evaluate the impact of recombination within each species by dividing genes into three equally sized recombination classes based on their average level of intragenic linkage disequilibrium. We find that $\alpha$ varies from 0.07 to 0.39 across species and is positively correlated with the level of recombination. This is both due to a higher estimated rate of adaptive evolution and a lower estimated rate of non-adaptive evolution, suggesting that recombination both increases the fixation probability of advantageous variants and decreases the probability of fixation of deleterious variants. Our results demonstrate that homologous recombination facilitates adaptive evolution measured by $\alpha$ in the core genome of prokaryote species in agreement with studies in eukaryotes.


# Significance statement

Whether intraspecific homologous recombination has a net beneficial or detrimental effect on adaptive evolution is largely unexplored in natural bacterial populations. We address this question by evaluating polymorphism and divergence data across the core genomes of 196 bacterial sequences—belonging to five closely related species of the genus *Rhizobium*. We show that the proportion of amino acid changes fixed due to adaptive evolution ($\alpha$) increases with an increased recombination rate. This correlation is observed both in the interspecies and intraspecific comparisons. By using a population genetics approach our results demonstrate that homologous recombination directly impacts the efficacy of natural selection in the core genome of prokaryotes, as previously reported in eukaryotes.

# Main text

# Introduction

Genetic recombination is expected to facilitate adaptive evolution by increasing the fixation probability of adaptive mutations and decreasing the probability of fixation of deleterious mutations (McVean and Charlesworth 2000). This is because recombination decouples the fate of adaptive and deleterious variants, decreasing the amount of selective interference throughout the genome (Hill and Robertson 1966; Felsenstein 1974). Selective interference—also termed the Hill-Robertson (HR) effect—is, therefore, strongest in regions of the genome where recombination is low (McVean and Charlesworth 2000). The HR effect is predicted to cause: (i) a reduction in the number of neutral polymorphisms, (ii) the accumulation of slightly deleterious polymorphisms, and (iii) a decrease in the probability of fixation of advantageous alleles (see (Charlesworth et al. 2009)). By mitigating the HR effect, homologous recombination is expected to increase the percentage of amino acid substitutions that are due to adaptive evolution ($\alpha$).

The parameter $\alpha$ can also be viewed as the relative proportion between the rate of amino acid changes fixed by positive selection ($\omega_a$) and the rate of non-adaptive amino acid changes relative to neutral ($\omega_{na}$) (as: $\alpha = \omega_a / (\omega_a + \omega_{na})$; (Galtier 2016; Moutinho et al. 2020)). Distinguishing between $\omega_a$ and $\omega_{na}$ allows us to test more precisely two expectations of the effect of increased homologous recombination: overall genes with higher recombination rates should experience more efficient purifying selection (and hence lower $\omega_{na}$) and increased probability of fixation for beneficial mutations (a higher $\omega_a$).

Empirical evidence based on population genomics data supports theory with a positive correlation between recombination and $\alpha$ reported in diverse species of eukaryotes, including flies (*Drosophila melanogaster*; (Campos et al. 2014; Castellano et al. 2016)), fungi (*Zymoseptoria tritici* (Grandaubert et al. 2019)), plants (*Arabidopsis thaliana* (Moutinho et al. 2019)), and non-model animal species (Galtier 2016; Moutinho et al. 2020).

Whereas recombination is ubiquitous and mandatory for the reproductive success of most eukaryotes (Page and Hawley 2003), this is not the case for prokaryotes. Nevertheless, many studied prokaryotes show high rates of genetic exchange (Didelot and Maiden 2010) and it is therefore of interest to explore whether such recombination also facilitates adaptive evolution in prokaryotes. Here we study how rates of adaptive evolution and intraspecific homologous recombination co-vary in a species complex of *Rhizobium leguminosarum* responsible for

97 nitrogen fixation in white clover (*Trifolium repens*) nodules. We have previously reported the full
98 genomic sequence of 196 isolates (Cavassim et al. 2020). Of 22115 orthologous gene groups
99 identified among the 196 strains, 4204 genes are present in all isolates (the core genome).
100 Although substantial adaptive evolution might be attributed to accessory genes through their gains
101 and losses via horizontal gene transfer (HGT) (Young et al. 2006; Tian et al. 2010; Porter et al.
102 2017; Cavassim et al. 2020), we focus here on genes vertically inherited and examine how much
103 variation in rates of homologous recombination explains variation in $\alpha$ and its components ($\omega_a$
104 and $\omega_{na}$).
105 Our previous analyses (Cavassim et al. 2020) showed that the 196 strains cluster into five
106 closely related species (2-5% average nucleotide divergence), with horizontal gene transfer
107 between these species only affecting the nitrogen fixation genes and a few well-defined genomic
108 regions—that we exclude in the present analysis. This species complex thus offers a unique
109 opportunity among prokaryotes to estimate the rates of fixation of amino acid changes by adaptive
110 evolution from isolates sampled from natural populations—enabling multiple comparisons of
111 polymorphism and divergence patterns among species. Our analyses provide evidence that the
112 rate of adaptive protein evolution increases with the recombination rate in this species complex.
113

114 # Results & Discussion

115 To estimate the proportion of adaptive evolution ($\alpha$) across this *Rhizobium* species
116 complex and study how $\alpha$ covaries with intraspecific recombination rate estimates, we restricted
117 analyses to polymorphism data from regions of the core genome without evidence of recent
118 interspecies HGT.
119 Of the eighteen species observed within this species complex (Young et al. 2021), we
120 have collected genomic data for five species (gsA-gE). Across all five species (196 strains,
121 gsA:32, gsB:32, gsC:112, gsD:5, gsE:11) (**Fig. S1**), a total of 22115 orthologous gene groups
122 were previously identified (Cavassim et al. 2020); of those, 4204 genes are present in all strains
123 (core genes). Most core genes are found in the large chromosome (3304 genes), but some are
124 located in the chromids (Rh01, Rh02) and one of the plasmids (Rh03) (see Harrison, et al., 2010,
125 and Cavassim et al. 2020). The chromosome, chromids, and the plasmid are hereafter referred
126 to as genomic compartments. To assess the effect of intraspecies homologous recombination on
127 adaptive evolution using a high-quality dataset we filtered out genes that showed evidence of
128 recent interspecies HGT or unexpectedly high rates of nucleotide diversity (see Methods) (**Fig.**

129 **S2**), leaving a total of 3086 genes (total alignment length: 3091179) and 334040 variable sites for
130 analysis (**Fig. S3**).
131 First, we estimated nucleotide diversity, intragenic linkage disequilibrium (LD), and the site
132 frequency spectrum (SFS) (see Methods) within each species (**Fig. 1a-c**). The average nucleotide
133 diversity, $\pi$, an estimator of $2N_e\mu$ in haploids, is significantly different among genomic
134 compartments (**Fig. 1a**, **Table S1**). Across the species, $\pi$ differs by up to a factor of 4.5 (gsA:
135 0.018, gsB:0.0045, gsC:0.0140, gsD:0.00512, gsE:0.008), with the most polymorphic species
136 being gsA and the least gsB. If we assume similar mutation rates among these closely related
137 species, nucleotide diversity differences reflect interspecies differences in long-term effective
138 population size, $N_e$.
139 When recombination occurs, we expect that levels of non-random association between
140 pairs of alleles, quantified by measures such as $r^2$ (see Methods), decay with genomic distance
141 (LD decay). To evaluate the recombination rate differences among the five species, we used
142 within-species polymorphism data and computed the average intragenic LD decay for each gene
143 in each species. We observed a rapid decay of LD within the first 1000 base pairs for all species,
144 suggesting substantial amounts of within-species homologous recombination (**Fig. 1b**). The
145 slower decay observed in species gsB either reflects a lower per generation recombination rate
146 or a smaller effective population size ($N_e$). The latter is consistent with the low level of nucleotide
147 diversity measured in gsB. To reliably estimate interspecies differences in $r^2$, we used genes with
148 at least ten informative sites within each species while also excluding variants only found in one
149 strain (singletons) and evaluated their $r^2$ distributions separately (**Fig. 1c**). As expected, the
150 species with the most striking LD decay (gsC) has the lowest $r^2$ (median $r^2$: 0.248), and the
151 opposite is also true (gsD, median $r^2$: 1.00). In summary, these species can be ranked by their
152 recombination levels, from the most recombining to the least, as follows: gsC (median $r^2$: 0.248)
153 > gsA (median $r^2$: 0.341) > gsE (median $r^2$: 0.561) > gsB (median $r^2$: 0.651) > gsD (median $r^2$:
154 1.00).
155 Next, we computed the folded site frequency spectrum (SFS) of synonymous and
156 nonsynonymous mutations within each species. Overall, both synonymous and nonsynonymous
157 SFSs differ from the "L" shaped patterns (many rare alleles and fewer frequent alleles) expected
158 in a stationary population at mutation-selection-drift equilibrium (**Fig. 1d**). The observed excess
159 of intermediate frequency SNPs indicates the presence of population structure in some of the
160 species. The effect of population structure is particularly evident in gsC, and this excess is likely
161 driven by strains isolated from French soils (**Fig. S4**). Differences among species suggest distinct

demographic histories, with gsC showing an SFS compatible with population expansion and gsA with population decline (Pool et al. 2010).

Using the counts of polymorphism in synonymous and nonsynonymous SFS within each species, we can estimate the overall strength of purifying selection via $pi_N/pi_S$. The strength of purifying selection ranks species similarly to their average recombination rate, with more recombining species showing stronger purifying selection (individual $pi_N/pi_S$ sorted by recombination rate are: gsC = 0.037, gsA = 0.039, gsE = 0.051, gsB = 0.057, gsD = 0.07). This observation is in line with the theoretical expectation of a positive effect of recombination on the overall efficacy of natural selection. We also observed an excess of rare nonsynonymous relative to synonymous variants (**Fig. 1e**), consistent with the segregation of nonsynonymous variants under weak purifying selection (Ohta 1976). Rare nonsynonymous variants are often deleterious ($s \sim 1/N_e$) (Hughes et al. 2003; Hughes 2005). Because deleterious variants contribute substantially to polymorphism but rarely to divergence (Fay et al. 2001; Charlesworth and Eyre-Walker 2006), their presence in the genomes, if not controlled for, will lead to an underestimation of $\alpha$ (Eyre-Walker and Keightley 2009).

We used GRAPES (Galtier 2016) to estimate the distribution of fitness effects (DFE) (Eyre-Walker and Keightley 2007) and the proportion of adaptive evolution ($\alpha$) from polymorphism and divergence data while accounting for the presence of deleterious mutations. This approach uses the site frequency distribution of both synonymous and nonsynonymous SFS counts to estimate the DFE while accounting for the effect of demography. The significant amount of shared polymorphism among species (**Table S2**) makes it difficult to reliably call ancestral and derived states (Schneider et al. 2011; Tataru et al. 2017). Accordingly, we chose to estimate the DFE and $\alpha$ using the folded SFSs (Galtier 2016). To determine the model of the DFE that best fit our data, we used a variety of DFE distribution models (**Table S3**). The DFE models we tested differ by the classes of mutations (deleterious, beneficial, and neutral) included in each DFE model and how fitness effects are distributed within these classes. When using Akaike's Information Criterion (AIC) to select the best DFE model, we found the GammaZero model overall provides the best fit to the SFS data (**Fig. S5**). This model assumes the existence of weakly deleterious nonsynonymous mutations, modeled as a continuous Gamma distribution (Galtier 2016).

The proportion of adaptive evolution was first computed between all combinations of "mirror" species ($\alpha_{species1\ species2}$), in which "species 2" is used as outgroup (divergence) for "species 1" (polymorphism) and vice-versa ($\alpha_{species2\ species1}$). This yielded twenty combinations in total. Because "mirror" species share an identical history of divergence, their $\alpha$ estimates can be

195 considered as "biological replicates" (Galtier 2016) (**Table 1**). Except for the comparison between
196 gsA and gsB, in which differences between $\alpha_{gsA\ gsB}$ and $\alpha_{gsB\ gsA}$ exceeded 0.09, the overall
197 discrepancy in the values estimated between mirror species does not exceed 0.1. Using each
198 species' focal polymorphism data, we calculated four $\alpha$ estimates by comparing it to the
199 divergence counts of the remaining species (**Table 1**). The most recombining species (gsC) is
200 observed to have the highest $\alpha$ across all outgroups used, while the least recombining species
201 (gsD) had the lowest $\alpha$ in three out of the four cases.

202 We then investigated whether intraspecies differences in recombination rate affect the
203 amount of adaptive evolution ($\alpha$) estimated. For each species, we split genes into three
204 recombination classes based on their average $r^2$ values and computed $\alpha$ for each class using
205 the GammaZero model (**Fig. S6, Table S4**). Because we only kept genes with at least ten
206 informative sites, the number of genes evaluated across species was different (**Fig. 1c**). For most
207 species comparisons (gsA, gsB, gsC, and gsE), there is a decrease in the proportion of adaptive
208 evolution with a reduction in recombination (increase in $r^2$) (**Fig. 2**). Except for cases in which we
209 used gsD polymorphisms to estimate $\alpha$ where all comparisons were non-significant, all the other
210 species pairwise comparisons led to at least one significant difference (based on non-overlapping
211 confidence intervals (CI)) between recombination classes. The low sample size of gsD (5 strains)
212 and its skewed LD distribution (**Fig. 1c**) may have reduced statistical power to discriminate among
213 recombination classes and to estimate $\alpha$ reliably.

214 We further assessed the significance of the pattern reported here by permuting, 200 times,
215 across recombination classes (see Methods). Except for simulations in which gsD polymorphisms
216 were used, all the other simulations led to significant differences (p-value $\leq$ 0.05) among the two
217 most extreme classes of recombination (**Fig. S7**).

218 The parameter $\alpha$ can also be viewed as the relative proportion between the rate of amino
219 acid changes fixed by positive selection ( $\omega_a$) and the rate of non-adaptive amino acid changes
220 ($\omega_{na}$): $\alpha = \omega_a / (\omega_a + \omega_{na})$. Thus, an increase in $\alpha$ with recombination could be due to either an
221 increase in the rate of adaptive substitutions, a decrease in the rate of non-adaptive substitutions,
222 or both. Figure 3 shows that $\omega_a$ increases with recombination rate whereas $\omega_{na}$ decreases with
223 recombination rate for most combinations and that the quantitative effects are almost equal in
224 magnitude. Thus, classes of genes evolving under higher recombination rates exhibited lower
225 rates of non-adaptive substitution and increased rates of fixation of adaptive variation. This
226 matches the predictions from selective interference theory (Felsenstein 1974; Barton 1994;
227 McVean and Charlesworth 2000).

228 To evaluate the robustness of these results, we computed two alternative measures of
229 recombination (R/θ, and D'). We then made new recombination classes and evaluated how each
230 recombination measure correlated with $\alpha$. R/θ measures the importance of recombination (R),
231 relative to mutation (θ) across sequences (Vos and Didelot 2009; Didelot and Wilson 2015), while
232 D' measures LD between sites using a different metric than $r^2$ (see methods). Although the
233 distributions of R/θ and D' across genes are different than that of $r^2$ (**Fig. 1c, Fig. S8**), these three
234 measures are not independent (Pearson's correlation between $r^2$and R/θ or D' ranged from 0.19
235 to 0.70) (**Fig. S8**). For most species comparisons, the trend between $\alpha$ and recombination
236 remains consistent: the higher the amount of recombination (measure by R/θ or D'), the higher $\alpha$
237 is (**Fig. S9-S10**). The mean rates of adaptive evolution ($\omega_a$) obtained among classes based on D'
238 or on R/θ estimates also generally agree with those estimated using $r^2$ (**Fig. S11-12**).
239 The genomes of the Rhizobium species comprise three genomic compartments
240 (chromosomes, chromids, and plasmids) that may have undergone different selection regimes.
241 We tested that by building distinct SFSs for each genomic compartment (see Methods). The mean
242 $a$ estimate differs slightly among genomic compartments with higher $a$ observed in core genes
243 sampled from the chromosome and the smallest plasmid (Rh03) (**Table 2**), however, the sampling
244 variance of estimates is large, and differences observed are not statistically significant. We also
245 evaluated the recombination rate among these genomic units (**Fig. S13**), recombination is
246 heterogeneous within and across species, as observed for nucleotide diversity (**Fig 1a**). We
247 conclude that we are underpowered to detect differences in the effects of adaptive evolution as a
248 function of recombination between these genomic compartments.
249 In this study, we applied a methodology that estimates the proportion of amino acid
250 changes that have been fixed by positive selection ($\alpha$) while also estimating the individual
251 components of $\alpha$ ($\omega_a$ and $\omega_{na}$). We have only included genes that are present in all sampled
252 genomes—the so-called core genome. Because the methodology used requires summarizing the
253 data by building an SFS, the analyses cannot be readily extended to the accessory genome—
254 given that many accessory genes are a result of HGT (Popa and Dagan 2011), a DFE computed
255 from this source will not necessarily reflect the DFE of the studied species. The accessory genome
256 represents roughly 40% of the genome of these species (Cavassim et al., 2020), and likely also
257 contributes to adaptive evolution (e.g., acquired symbiotic ability (Cavassim et al., 2020; Kumar
258 et al., 2016)) as also shown in other studies (e.g., antibiotic resistance in *Staphylococcus aureus*
259 (Harris et al. 2010); or adaptation to a new ecological niche (Ochman et al. 2000; Wiedenbeck
260 and Cohan 2011)).

261 It has been previously shown experimentally that both recombination via plasmid-
262 mediated gene transfer (conjugation) or via transformation can accelerate bacterial adaptation in
263 populations of *E. coli* (Cooper 2007) and *Helicobacter pylori* (Baltrus et al. 2008). These studies
264 are in line with previous simulation studies (Cohen et al. 2005; Levin and Cornejo 2009).
265 Homologous recombination was shown to accelerate adaptation when incorporated in
266 simulations, including mutation and selection, at rates typical of species like *E. coli*, *H.*
267 *influenzae*, *B. subtilis* (Levin and Cornejo 2009). Cohen et al., 2005 studied recombination in the
268 context of a simple fitness landscape model with evolution implemented as a continuous Markov
269 process and observed a drastic speed up on the rate of adaptive evolution with increased
270 population sizes. Cooper 2007 concluded that recombination only had a positive effect on
271 adaptation when beneficial mutations were abundant in the population—implying that standing
272 genetic variation, possibly driven by higher mutation rate ($\mu$) or higher effective population size
273 ($N_e$), may be crucial for recombination to be useful. Assuming that mutation rate ($\mu$) is similar
274 among these sibling species, then the observed adaptive differences among them may reflect
275 differences in effective population size, recombination rate, or a combination of both (Cohen et
276 al. 2005; Arnold et al. 2018).

# 277 Conclusion

278 We have found that five bacterial species within the species complex *R. leguminosarum*
279 display different yet high levels of recombination. The estimates of $\alpha$ ranged between 0.07 and
280 0.39 among species. These estimates are lower than those based on 410 orthologs observed in
281 *E. coli* (0.58, CI=0.45, 0.68) but close to estimates from *S. enterica* (0.34, CI=0.14, 0.50)
282 previously reported (Charlesworth and Eyre-Walker 2006)—however, in this study population
283 fluctuations were not accounted for.
284 Levels of recombination correlate—both across and within species—with higher amounts
285 of adaptive evolution estimated either as the rate of adaptive substitutions ($\omega_a$) or as the
286 proportion of amino acid changes that have been fixed by positive selection ($\alpha$). For instance, the
287 most recombining species (gsC) consistently exhibited the largest $\alpha$, independent of the outgroup
288 used. Within each species, we also find a positive correlation between intragenomic
289 recombination rate and $\alpha$. This is both due to a higher estimated rate of adaptive evolution ($\omega_a$)
290 and a lower estimated rate of non-adaptive evolution ($\omega_{na}$), suggesting that recombination both
291 increases the fixation probability of advantageous variants and decreases the probability of

292 fixation of deleterious variants. These findings are robust to the measure of recombination ($r^2$,
293 *R*/*θ*, and D') used to define classes and the choice of outgroup used for computing divergence.
294 Despite variation in recombination rate among genomic compartments, we did not observe
295 significant differences in adaptive evolution among them.

296 The positive association between amounts of homologous recombination on $\alpha$ we report
297 here is in line with population genetic studies conducted in vertebrates (Galtier 2016; Moutinho et
298 al. 2020) and invertebrates (Presgraves 2005; Betancourt et al. 2009; Arguello et al. 2010;
299 Mackay et al. 2012; Campos et al. 2014; Grandaubert et al. 2019); it is also in line with
300 experimental and simulation studies of adaptive evolution in prokaryotes (Cohen et al. 2005;
301 Cooper 2007; Baltrus et al. 2008; Levin and Cornejo 2009). It points to recombination being a
302 general facilitator of adaptive evolution across the tree of life.
303

# 304 Material and methods

## 305 Identification of orthologous genes

306 We previously isolated and sequenced 196 strains from white clover (*Trifolium repens*)
307 root nodules harvested in Denmark, France, and the UK. To identify a set of orthologous genes
308 shared across strains, we followed the methods outlined in Cavassim et al., 2020. Briefly, the
309 strains were previously subjected to whole-genome shotgun sequencing using 2×250bp Illumina
310 paired-end reads (Illumina, USA). Genomes were assembled using SPAdes (v. 3.6.2, (Bankevich
311 et al. 2012)) and assembled further, one strain at a time, using a custom Python script (Jigome,
312 available at https://github.com/izabelcavassim/Rhizobium_analysis/tree/master/Jigome).

313 From the assembled genomes (Cavassim et al. 2020), we predicted protein-coding
314 sequences using prokka (Seemann 2014) (v1.12); this resulted in a total of 1468264 protein-
315 coding sequences. To predict orthologous genes from these sequences, we used Proteinortho
316 (Lechner et al. 2014; Seemann 2014) (v5.16b) with default parameters except for enabling the
317 synteny flag. We identified a total of 22115 orthologous gene groups, including a total of 17911
318 orthologous observed in at least two strains (accessory genes), and 4204 orthologous found in
319 all 196 strains (core genes).

320 Orthologous gene groups were aligned using clustalo (Sievers et al. 2011) (v.1.2.0) in a
321 codon-aware manner. To determine the genetic relationship among all 196 strains, we previously
322 calculated their pairwise average nucleotide identity (ANI) across 305 conserved orthologous
323 gene alignments (Cavassim et al. 2020). Under the 95% ANI threshold that delineates species
324 boundaries (Konstantinidis et al. 2006), we demonstrated that these 196 Rhizobium strains

325 constitute five distinct *R. leguminosarum* species (gsA, gsB, gsC, gsD, and gsE) (**Fig. S1**). To
326 ensure that we had a high-quality orthologous dataset for extracting segregating sites, we filtered
327 it further (see below).
328

## 329 Filtering out orthologous gene groups with evidence of interspecies 330 HGT or misassigned orthologous gene groups

331 To evaluate the impact of intraspecific homologous recombination on adaptive evolution,
332 we excluded genes that showed signals of recent HGT across the five species analyzed. We have
333 previously developed and applied a phylogenetic method to quantify HGT (introgression score)
334 (Cavassim et al. 2020). This method evaluates the possible number of shifts from one species to
335 another in a given phylogenetic tree. The pipeline takes a gene tree as an input and traverses the
336 tree—using the depth-first search approach—searching deeper in the tree whenever possible.
337 Once the tip is reached the species classification for that given strain is stored. A list containing
338 the species in order of search is collected for the entire tree, the introgression score is then
339 computed as the number of shifts from one species to another in the list minus the set of species
340 plus 1.
341 We previously showed that most of the core genes shared among the present species
342 respect the species-tree topology (introgression score = 0) (Cavassim et al. 2020). The exceptions
343 are genes sitting in the symbiosis conjugative plasmids, and two chromosomal islands
344 (introgression score > 7). To ensure that we were only analyzing high-quality gene alignments,
345 with no evidence of misassigned orthologous gene groups and with little evidence of HGT, we
346 imposed some restrictions. We only accepted genes that passed the following criteria: (i) were
347 present in every strain (196 strains), (ii) with a nucleotide diversity ($\pi$) below 0.1 (see **Fig. S2a**),
348 (iii) identifiable replicon origin (chromosome and chromids), (iiii) and with an introgression score
349 ≤ 3. A total of 3086 out of 4204 core genes were kept, and of these, 2550 genes were found in
350 the chromosome, 288 genes in chromid Rh01, 160 genes in chromid Rh02, and 88 genes in
351 plasmid Rh03.
352

## 353 Variant calling

354 To identify single nucleotide polymorphisms (SNPs) along with our high-quality set of core
355 genes, we evaluated each gene codon-aware alignment using a custom python script

356 https://github.com/izabelcavassim/Popgen_bacteria. For a given core gene alignment and
357 position, we first counted the number of unique nucleotides (A, C, T, G). Only sites containing two
358 unique nucleotides were considered variable sites (bi-allelic SNPs). SNP matrices were then built
359 and encoded as follows: major alleles were encoded as 1 and minor alleles as 0. The nucleotide
360 diversity ($\pi$), gene length, and the distributions of segregating sites across core genes are
361 described in **Fig. S2** (**b-d**).

## 362 Transition transversion rate bias (kappa) and expected number of 363 synonymous and nonsynonymous sites

364 Because transitions are more often synonymous at third codon positions than are
365 transversions, to correctly identify the expected number of synonymous ($Lps$) and
366 nonsynonymous sites ($Lpn$), we first estimated the average transition/transversion rate bias
367 (kappa) (Ina 1995) across species. To this end, we followed the methods described in (Yang and
368 Nielsen 2000) and used two classes of sites: fourfold-degenerate sites at the third codon positions
369 and nondegenerate sites. Mutations at the fourfold-degenerate sites are synonymous, and
370 therefore kappa at those sites should reflect only the mutational bias. All mutations at non-
371 degenerate sites are nonsynonymous and were also used to estimate kappa. We computed an
372 average kappa by combining these two classes based on equations 8, 9, 10, and 11 of (Yang and
373 Nielsen 2000). These equations have been implemented within the CodonSeq class in Biopython
374 (Cock et al. 2009) (private function "_count_site_YN00() "), and these private functions were
375 adapted to our dataset.
376 To estimate a common kappa for each gene alignment (including all species and strains),
377 we averaged estimates from pairwise analyses across 50 randomly chosen strains. The kappa
378 distribution has a mean of 5.6 and a median of 5.20 (**Fig. S3a**), we used the median to compute
379 the expected number of synonymous and nonsynonymous sites. To this end, we followed the
380 methods described by (Ina 1995) and modified by (Yang and Nielsen 2000)—also implemented
381 within Biopython. A total of 284742 synonymous, 49298 nonsynonymous sites were counted (**Fig.
382 S3b-c**).
383

## 384 Divergence sites and shared polymorphisms

385 For each pair of species (a focal and an outgroup), we evaluated their variable sites and
386 computed the number of shared synonymous ($pS$) and nonsynonymous ($pN$) polymorphisms.
387 Given a bi-allelic SNP (0 and 1), we considered shared polymorphic sites as sites for which both

388 alleles (0,1) were segregating in both species (**Table S2**). We restricted the estimates of
389 divergence to those sites for which we had variable sites across species. We classified
390 synonymous ($d_S$) and nonsynonymous divergent sites ($d_N$) as those sites in which we observed
391 fixed differences between a focal species and an outgroup.
392

## 393 Calculating the folded Site Frequency Spectrum

394 One can infer the distribution of fitness effects from Site Frequency Spectrum (SFS) data
395 (Eyre-Walker and Keightley 2009). Because of the amount of shared polymorphism among the
396 present species (**Table S2**), it becomes problematic to confidently distinguish ancestral from
397 derived polymorphisms (Hernandez et al. 2007). Therefore, we chose to estimate the DFE using
398 a method that uses the folded site frequency spectrums (SFS) of synonymous and
399 nonsynonymous sites (Galtier 2016). To this end, we built the folded synonymous and
400 nonsynonymous site frequency spectrums by tabulating the observed counts of the minor allele
401 frequencies. The synonymous and nonsynonymous SFSs, and the divergence counts, were then
402 used to estimate the DFE and the proportion of adaptive substitutions ($\alpha$) across pairs of species.
403

## 404 Calculating the strength of purifying selection

405 The strength of purifying selection was measured as the ratio of nucleotide diversity at
406 nonsynonymous ($pi_N$) and synonymous sites ($pi_S$). For each gene and class of polymorphisms
407 (synonymous and nonsynonymous) nucleotide diversity was computed as: $\pi = \sum_1^m (2pq)/Lp$,
408 in which $p$ and $q$ are the allele frequencies, and Lp is the expected number of synonymous ($Lps$)
409 or nonsynonymous positions ($Lpn$) along the gene. We use the median of the $pi_N/pi_S$ distribution
410 among genes as a proxy for the strength of purifying selection per species.
411

## 412 Estimation of adaptive and non-adaptive non-synonymous
## 413 substitutions rates

414 Fitted parameters of the DFE were used to compute the expected $d_N/d_S$ under the
415 different models, which was compared to the observed $d_N/d_S$ to estimate the adaptive substitution
416 rate ($\omega_A$); non-adaptive substitution rate ($\omega_{NA}$), and the proportion of adaptive substitutions ($\alpha$)
417 with $\omega A = \alpha\ d_N/d_S$ and $\omega_{NA} = (1-\alpha)\ d_N/d_S$.
418 To account for potential departures of the SFS from demographic equilibrium (assuming
419 the Wright-Fisher model)—possibly driven by changes in the effective population size or by

population structure—the method uses nuisance parameters to correct for these SFS distortions (Eyre-Walker et al. 2006). The different DFE models were compared using the Akaike Information Criterion (AIC) (Akaike 1992).

## Recombination rate estimates

To estimate the recombination rate per gene per species, we used three approaches: two based on the degree of association (or linkage disequilibrium) between pairs of alleles in a sample of haplotypes ($r^2$and $D'$), and a third approach, ClonalFrameML (R/θ) (Vos and Didelot 2009; Didelot and Wilson 2015), which relies on the maximum likelihood inference to detect recombination events that disrupt a clonal pattern of inheritance in bacterial genomes.

### (1) linkage disequilibrium ($r^2$)

Intragenic linkage disequilibrium (LD) measures the correlation between pairs of alleles with genomic distance in a gene. Here we used Pearson's $r^2$ correlation measure.

Each gene genotype matrix (containing a minimal set of ten single nucleotide polymorphisms (SNPs)) was first normalized as follow: let $N$ denote the total number of individuals, and $M$ the total number of SNPs, the full gene genotype matrix ($X$) has dimensions $N \times M$ with genotypes encoded as 0's and 1's for the $N$ haploid individuals. Each column $S_i$ ($i$ = 1, …, $M$) of the $X$ matrix is a vector of SNP information of size $N$. To compute LD, we discarded SNPs found only in one sample (singletons). We then applied a Z-score normalization to each SNP vector by subtracting the vector by its mean and dividing it by its standard deviation ($\frac{S_i - \bar{S_i}}{\sqrt{var(S_i)}}$), resulting in a vector with mean 0 and variance 1. The linkage disequilibrium was then calculated as a function of distance $d$ (maximum 1000 base pairs apart) and was computed as the average LD of pairs of SNPs $d$ base pairs away from each other. The calculations were done in the following way:

$$Cor(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sqrt{var(X_i)var(X_j)}}$$

$$r^2 = Cor(X_i, X_j)^2$$

In which $j > i$ and $X_i$ is composed of the genotypes of all individuals of a given species for SNP position $i$ in the genotype matrix. $X_j$is formed of the genotypes of all individuals of the same species for position $j$ in the genotype matrix, and$d = j - i$ with $d \leq 1000$ base pairs. Results were then summarized into bins of 100 base pairs apart; for each bin, a mean $r^2$ was computed and then averaged to a singular $r^2$ value.

### (2) linkage disequilibrium (D')

452 The average linkage disequilibrium within genes was also measured by D' (Lewontin 1964) as
453 follow: given two locus $i$ and $j$ (with alleles $A$ and $a$, observed in locus $i$ and alleles $B$ and $b$
454 observed in locus $j$), we first computed the frequency of all possible haplotype combinations
455 ($f_{AA}, f_{AB}, f_{Ab}, f_{ab}$) and allele frequencies ($f_A$, $f_B$, $f_a$, and $f_b$). The coefficient D was then computed
456 as: $D_{AB} = f_{AB} - f_A \cdot f_B$ and D' was computed by standardizing |D| by its maximum possible value
457 as: $D' = \frac{D}{min(f_A f_b, f_a f_B)}, if\ D > 0$, or $D' = \frac{-D}{min(f_A f_B, f_a f_b)}, if\ D < 0$. An average value per gene was
458 stored and computed similarly to the $r^2$ statistics (see above).

459 **(3) ClonalFrameML**

460 To estimate the changes in the clonal phylogeny by recombination (R), relative to mutation
461 (θ) (R/θ), we used the software ClonalFrameML (Vos and Didelot 2009; Didelot and Wilson 2015).
462 For each species, we first concatenated all core gene alignments (3086 genes) to build the
463 starting phylogenetic species tree using a maximum likelihood approach (Raxml-ng (Kobert et al.
464 2014; Stamatakis 2014)). We then input each phylogenetic tree within each gene alignment to
465 estimate R/θ.
466

467 ## Calculating the significance levels between recombination classes

468 To test whether differences in $\alpha$ among recombination classes were statistically significant
469 across species comparisons, we conducted a non-parametric test by shuffling genes among
470 recombination classes (200 permutations) and recording the amplitude of differences between $\alpha$
471 estimates ($\Delta_\alpha = max_\alpha - min_\alpha$). We calculated a p-value by comparing the observed $\Delta_\alpha$ against
472 the simulated $\Delta_\alpha$ distribution.
473

474 ## Estimation of adaptive substitutions by genomic compartments

475 To estimate adaptive evolution ($\alpha$) among genomic compartments we first down-sampled
476 the genes from genomic compartments (chromosome, Rh01, Rh02) to reach the size of the
477 smallest genomic compartment (Rh03). Due to the paucity of data, we chose to compute $\alpha$ using
478 polymorphism data from the most polymorphic species (gsA) and contrasted it against each
479 outgroup (gsB-gsE). The $\alpha$ estimates and their associated confidence intervals were obtained
480 using the GammaZero DFE model within GRAPES.
481

482 # Data sharing plans

- Code generated for this study can be found at https://github.com/izabelcavassim/Popgen_bacteria.
- The data that support the findings of this study are available in the INSDC databases under Study/BioProject ID PRJNA510726 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA510726/).
- Accessions numbers are from SAMN10617942 to SAMN10618137 consecutively.
- Orthologous gene alignments and SNP matrices are available on FigShare (file Data.zip): https://doi.org/10.6084/m9.figshare.11568894.v5


# Funding information

This work was funded by grant no. 4105-00007A from Innovation Fund Denmark (S.U.A. and M.H.S.).

# Acknowledgments

The authors would like to thank industrial partners DLF Trifolium, SEGES, and Legume Technology Ltd. for their contribution to the field trials. The authors would also like to thank J. Peter W. Young, Paula Tataru, Bjarni Vilhjálmsson, and Marjolaine Rousselle for their helpful discussions about this work.


# References


Akaike H. 1992. Information Theory and an Extension of the Maximum Likelihood Principle. *Springer Series in Statistics* [Internet]:610–624. Available from: http://dx.doi.org/10.1007/978-1-4612-0919-5_38

Arguello JR, Roman Arguello J, Zhang Y, Kado T, Fan C, Zhao R, Innan H, Wang W, Long M. 2010. Recombination Yet Inefficient Selection along the Drosophila melanogaster Subgroup's Fourth Chromosome. *Molecular Biology and Evolution* [Internet] 27:848–861. Available from: http://dx.doi.org/10.1093/molbev/msp291

Arnold BJ, Gutmann MU, Grad YH, Sheppard SK, Corander J, Lipsitch M, Hanage WP. 2018. Weak Epistasis May Drive Adaptation in Recombining Bacteria. *Genetics* 208:1247–1260.

Baltrus DA, Guillemin K, Phillips PC. 2008. Natural transformation increases the rate of adaptation in the human pathogen Helicobacter pylori. *Evolution* 62:39–49.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to

522 Single-Cell Sequencing. *Journal of Computational Biology* [Internet] 19:455–477. Available from:
523 http://dx.doi.org/10.1089/cmb.2012.0021

524 Barton NH. 1994. The reduction in fixation probability caused by substitutions at linked loci. *Genetical*
525 *Research* [Internet] 64:199–208. Available from: http://dx.doi.org/10.1017/s0016672300032857

526 Betancourt AJ, Welch JJ, Charlesworth B. 2009. Reduced effectiveness of selection caused by a lack of
527 recombination. *Curr. Biol.* 19:655–660.

528 Campos JL, Halligan DL, Haddrill PR, Charlesworth B. 2014. The relation between recombination rate
529 and patterns of molecular evolution and variation in Drosophila melanogaster. *Mol. Biol. Evol.*
530 31:1010–1028.

531 Castellano D, Coronado-Zamora M, Campos JL, Barbadilla A, Eyre-Walker A. 2016. Adaptive Evolution Is
532 Substantially Impeded by Hill-Robertson Interference in Drosophila. *Mol. Biol. Evol.* 33:442–455.

533 Cavassim MIA, Moeskjær S, Moslemi C, Fields B, Bachmann A, Vilhjálmsson BJ, Schierup MH, W Young
534 JP, Andersen SU. 2020. Symbiosis genes show a unique pattern of introgression and selection
535 within a species complex. *Microb Genom* [Internet] 6. Available from:
536 http://dx.doi.org/10.1099/mgen.0.000351

537 Charlesworth B, Betancourt AJ, Kaiser VB, Gordo I. 2009. Genetic recombination and molecular
538 evolution. *Cold Spring Harb. Symp. Quant. Biol.* 74:177–186.

539 Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. *Mol. Biol. Evol.*
540 23:1348–1356.

541 Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F,
542 Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular
543 biology and bioinformatics. *Bioinformatics* 25:1422–1423.

544 Cohen E, Kessler DA, Levine H. 2005. Recombination Dramatically Speeds Up Evolution of Finite
545 Populations. *Physical Review Letters* [Internet] 94. Available from:
546 http://dx.doi.org/10.1103/physrevlett.94.098102

547 Cooper TF. 2007. Recombination speeds adaptation by reducing competition between beneficial
548 mutations in populations of Escherichia coli. *PLoS Biol.* 5:e225.

549 Didelot X, Maiden MCJ. 2010. Impact of recombination on bacterial evolution. *Trends Microbiol.* 18:315–
550 322.

551 Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial
552 genomes. *PLoS Comput. Biol.* 11:e1004041.

553 Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nature Reviews*
554 *Genetics* [Internet] 8:610–618. Available from: http://dx.doi.org/10.1038/nrg2146

555 Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of
556 slightly deleterious mutations and population size change. *Mol. Biol. Evol.* 26:2097–2108.

557 Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino
558 acid mutations in humans. *Genetics* 173:891–900.

559 Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics*
560 158:1227–1234.

561 Felsenstein J. 1974. The evolutionary advantage of recombination. *Genetics* 78:737–756.

562 Galtier N. 2016. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis.
563 *PLoS Genet.* 12:e1005774.

564 Grandaubert J, Dutheil JY, Stukenbrock EH. 2019. The genomic determinants of adaptive evolution in a
565 fungal pathogen. *Evol Lett* 3:299–312.

566 Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N,
567 Lindsay JA, et al. 2010. Evolution of MRSA during hospital transmission and intercontinental spread.
568 *Science* 327:469–474.

569 Hernandez RD, Williamson SH, Bustamante CD. 2007. Context dependence, ancestral misidentification,
570 and spurious signatures of natural selection. *Mol. Biol. Evol.* 24:1792–1800.

571 Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genetical Research*
572 [Internet] 8:269–294. Available from: http://dx.doi.org/10.1017/s0016672300010156

573 Hughes AL. 2005. Evidence for abundant slightly deleterious polymorphisms in bacterial populations.
574 *Genetics* 169:533–538.

575 Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M. 2003. Widespread purifying
576 selection at polymorphic sites in human protein-coding loci. *Proc. Natl. Acad. Sci. U. S. A.*
577 100:15754–15757.

578 Ina Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions.
579 *J. Mol. Evol.* 40:190–226.

580 Kobert K, Flouri T, Aberer A, Stamatakis A. 2014. The Divisible Load Balance Problem and Its Application
581 to Phylogenetic Inference. *Lecture Notes in Computer Science* [Internet]:204–216. Available from:
582 http://dx.doi.org/10.1007/978-3-662-44753-6_16

583 Konstantinidis KT, Ramette A, Tiedje JM. 2006. The bacterial species definition in the genomic era.
584 *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 361:1929–1940.

585 Lechner M, Hernandez-Rosales M, Doerr D, Wieseke N, Thévenin A, Stoye J, Hartmann RK, Prohaska
586 SJ, Stadler PF. 2014. Orthology detection combining clustering and synteny for very large datasets.
587 *PLoS One* 9:e105015.

588 Levin BR, Cornejo OE. 2009. The population and evolutionary dynamics of homologous gene
589 recombination in bacterial populations. *PLoS Genet.* 5:e1000601.

590 Lewontin RC. 1964. The Interaction of Selection and Linkage. I. General Considerations; Heterotic
591 Models. *Genetics* 49:49–67.

592 Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM,
593 Cridland JM, et al. 2012. The Drosophila melanogaster Genetic Reference Panel. *Nature* 482:173–
594 178.

595 McVean GA, Charlesworth B. 2000. The effects of Hill-Robertson interference between weakly selected
596 mutations on patterns of molecular evolution and variation. *Genetics* 155:929–944.

597 Moutinho AF, Bataillon T, Dutheil JY. 2020. Variation of the adaptive substitution rate between species
598 and within genomes. *Evolutionary Ecology* [Internet] 34:315–338. Available from:
599 http://dx.doi.org/10.1007/s10682-019-10026-z

600 Moutinho AF, Trancoso FF, Dutheil JY. 2019. The Impact of Protein Architecture on Adaptive Evolution.
601 *Mol. Biol. Evol.* 36:2013–2028.

602 Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial

603 innovation. *Nature* 405:299–304.

604 Ohta T. 1976. Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theor.*
605 *Popul. Biol.* 10:254–275.

606 Page SL, Hawley RS. 2003. Chromosome choreography: the meiotic ballet. *Science* 301:785–789.

607 Peter W Harrison, Ryan P J Lower, Nayoung K D Kim, J Peter W Young. 2010. Introducing the bacterial
608 "chromid": not a chromosome, not a plasmid. *Trends Microbiol.* 18:141–148.

609 Pool JE, Hellmann I, Jensen JD, Nielsen R. 2010. Population genetic inference from genomic sequence
610 variation. *Genome Research* [Internet] 20:291–300. Available from:
611 http://dx.doi.org/10.1101/gr.079509.108

612 Popa O, Dagan T. 2011. Trends and barriers to lateral gene transfer in prokaryotes. *Curr. Opin. Microbiol.*
613 14:615–623.

614 Porter SS, Chang PL, Conow CA, Dunham JP, Friesen ML. 2017. Association mapping reveals novel
615 serpentine adaptation gene clusters in a population of symbiotic Mesorhizobium. *ISME J.* 11:248–
616 262.

617 Presgraves DC. 2005. Recombination enhances protein adaptation in Drosophila melanogaster. *Curr.*
618 *Biol.* 15:1651–1656.

619 Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. 2011. A method for inferring the rate of
620 occurrence and fitness effects of advantageous mutations. *Genetics* 189:1427–1437.

621 Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069.

622 Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J,
623 et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using
624 Clustal Omega. *Mol. Syst. Biol.* 7:539.

625 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
626 phylogenies. *Bioinformatics* [Internet] 30:1312–1313. Available from:
627 http://dx.doi.org/10.1093/bioinformatics/btu033

628 Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of Distribution of Fitness Effects and
629 Proportion of Adaptive Substitutions from Polymorphism Data. *Genetics* 207:1103–1119.

630 Tian CF, Young JPW, Wang ET, Tamimi SM, Chen WX. 2010. Population mixing of Rhizobium
631 leguminosarum bv. viciae nodulating Vicia faba: the role of recombination and lateral gene transfer.
632 *FEMS Microbiology Ecology* [Internet]:no – no. Available from: http://dx.doi.org/10.1111/j.1574-
633 6941.2010.00909.x

634 Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. *ISME*
635 *J.* 3:199–208.

636 Wiedenbeck J, Cohan FM. 2011. Origins of bacterial diversity through horizontal genetic transfer and
637 adaptation to new ecological niches. *FEMS Microbiol. Rev.* 35:957–976.

638 Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic
639 evolutionary models. *Mol. Biol. Evol.* 17:32–43.

640 Young JPW, Crossman LC, Johnston AWB, Thomson NR, Ghazoui ZF, Hull KH, Wexler M, Curson ARJ,
641 Todd JD, Poole PS, et al. 2006. The genome of Rhizobium leguminosarum has recognizable core
642 and accessory components. *Genome Biol.* 7:R34.

643 Young JPW, Moeskjær S, Afonin A, Rahi P, Maluk M, James EK, Cavassim MIA, Harun-or Rashid M,
644 Aserse AA, Perry BJ, et al. 2021. Defining the Rhizobium leguminosarum Species Complex. *Genes*
645 [Internet] 12:111. Available from: http://dx.doi.org/10.3390/genes12010111

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665 **Figures and Tables**

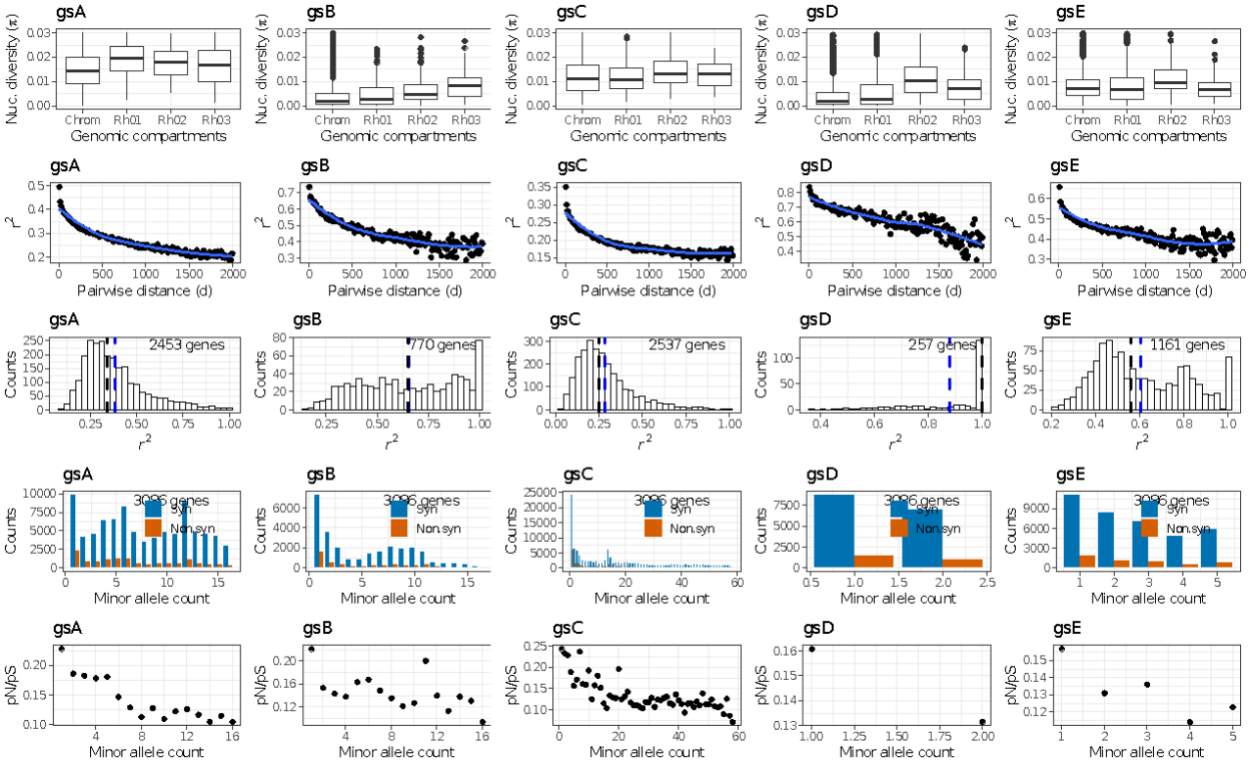


**Figure 1. Population genetics parameters across five species**. (**a**) Nucleotide diversity ($\pi$) across 3086 genes distributed along with genomic compartments (chromosome, chromids: Rh01, Rh02, and plasmid: Rh03). To exclude outliers only genes with $\pi \leq 0.03$ are shown. (**b**) Intragenic linkage disequilibrium measured via the decay of $r^2$ for all core genes (3086 genes). The curve fitting line (in blue) is from a local regression method (loess). (**c**) Linkage disequilibrium ($r^2$) distribution across genes. Only genes with at least ten segregating sites were kept and singletons were excluded (gsA:2453 genes, gsB:770, gsC:2537, gsD:257, and gsE:1161). The black and blue dashed lines correspond to the median and mean $r^2$, respectively. (**d**) Site frequency spectrum counts of synonymous and non-synonymous sites by minor allele count based on all core genes (3086 genes). (**e**) The ratio of non-synonymous to synonymous polymorphisms by minor allele count.

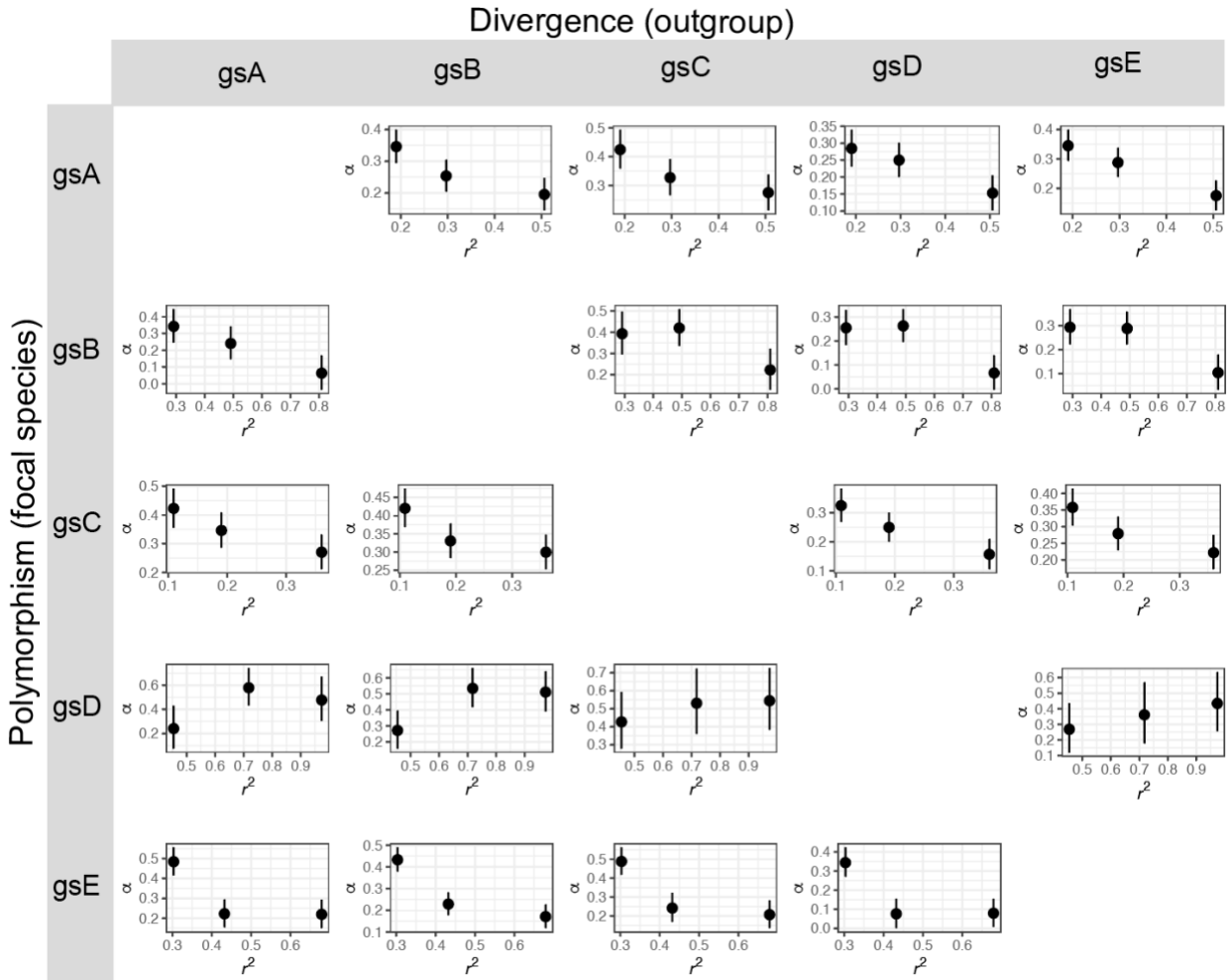


**Figure 2. The proportion of adaptive evolution ($\alpha$) by classes of recombination.** For each pairwise estimates of $\alpha$ the polymorphism data from one species (left in title) is compared against the divergence counts of an outgroup (right in title), and vice-versa. Results are divided into classes of recombination based on $r^2$(a measure that is inversely proportional to the level of recombination). The $\alpha$ estimates and their associated confidence intervals were obtained using the best fitting DFE model (GammaZero).

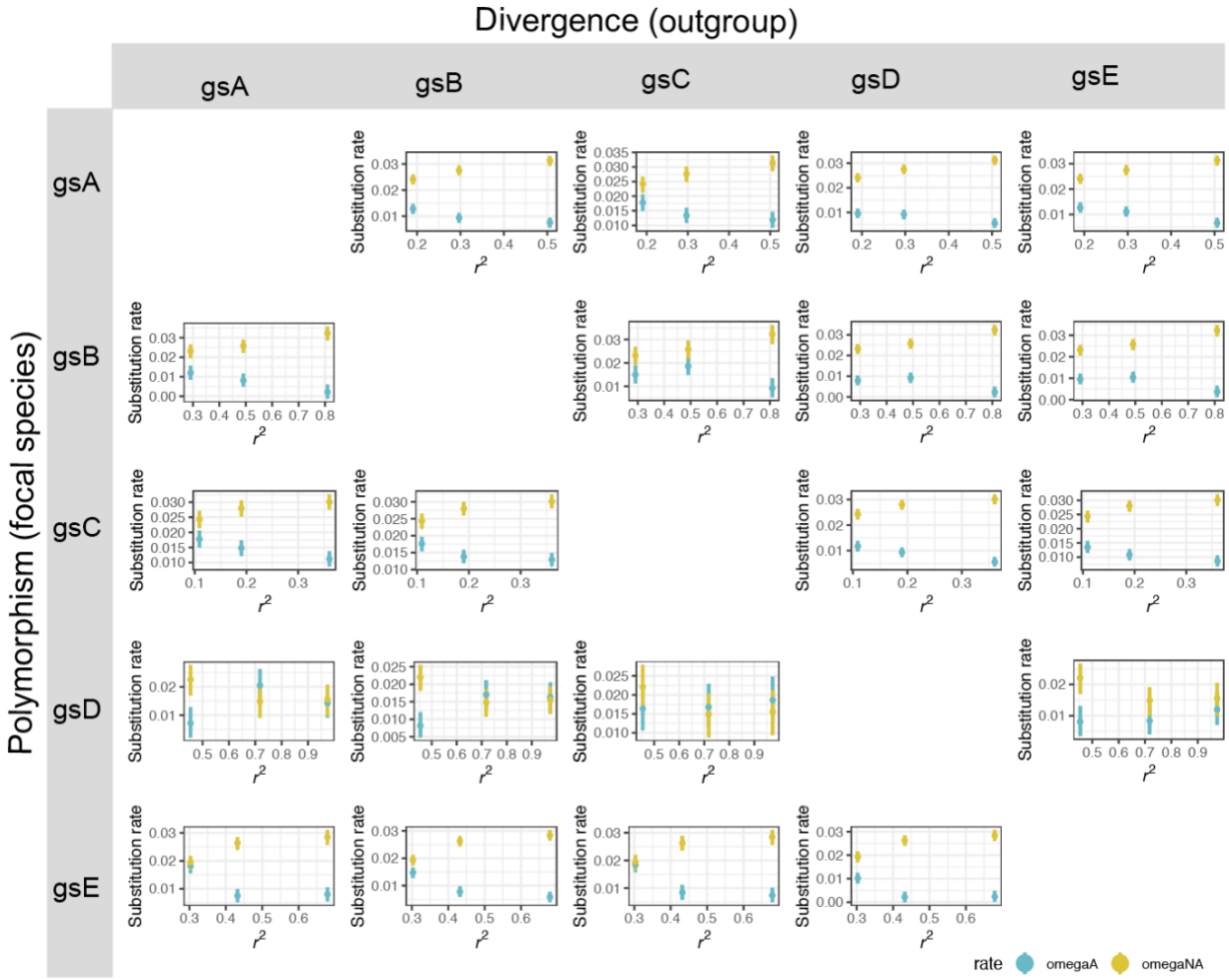


**Figure 3. The rates of adaptive ($\omega_A$) and non-adaptive ($\omega_{NA}$) evolution by classes of recombination.** For each pairwise estimates of $\omega_A$(in blue) and $\omega_{NA}$ (in yellow) the polymorphism data from one species (left in title) is compared against the divergence counts of an outgroup (right in title), and vice-versa. Results are divided into classes of recombination based on $r^2$(a measure that is inversely proportional to the level of recombination). An opposite effect of recombination on $\omega_A$ and $\omega_{NA}$ is observed in most pairwise comparisons. The estimates ($\omega_A$, $\omega_{NA}$) and their associated confidence intervals were obtained using the best fitting DFE model (GammaZero).

**Table 1. The proportion of adaptive evolution ($\alpha$) across pairs of species.** The $\alpha$ estimates were computed based on the best fitting DFE model (GammaZero) (**Table S3**).

709 For each pairwise estimate of $\alpha$ ($\alpha_{species1\ species2}$), the polymorphism data from a focal
710 species (rows) is compared against the divergence counts of an outgroup (columns), and
711 vice-versa ($\alpha_{species2\ species1}$). Confidence intervals are displayed in brackets (grey) and
712 numbers in parentheses represent the $\alpha$ ranking (in decreasing order) by outgroup (by
713 column).
714

| Polymorphism (focal) | Divergence (outgroup) | | | | |
|---|---|---|---|---|---|
| | gsA | gsB | gsC | gsD | gsE |
| gsA | - | 0.28 [0.26-0.31] (2) | 0.35 [0.33-0.39] (1) | 0.25 [0.23-0.28] (1) | 0.29 [0.27-0.32] (2) |
| gsB | 0.18 [0.16-0.21] (4) | - | 0.26 [0.24-0.29] (3) | 0.15 [0.13-0.17] (3) | 0.17 [0.15-0.19] (3) |
| gsC | 0.36 [0.33-0.39] (1) | 0.36 [0.33-0.38] (1) | - | 0.25 [0.23-0.28] (1) | 0.30 [0.27-0.32] (1) |
| gsD | 0.25 [0.22-0.28] (3) | 0.25 [0.23-0.27] (4) | 0.25 [0.22-0.28] (4) | - | 0.12 [0.10-0.15] (4) |
| gsE | 0.27 [0.24-0.30] (2) | 0.25 [0.23-0.27] (4) | 0.27 [0.24-0.30] (2) | 0.10 [0.07-0.13] (4) | - |

715
716
717
718
719
720
721
722
723
724
725
726
727 **Table 2. The proportion of adaptive evolution ($\alpha$) across genomic compartments.**
728 The $\alpha$ estimates were computed based on the best fitting DFE model (GammaZero)

729 (**Table S3**). For each genomic compartment (chrom = chromosome; Rh01 and Rh02 =
730 chromids; Rh03 = plasmid) we compared the polymorphism data from species gsA
731 against the divergence counts of an outgroup (rows). Confidence intervals are displayed
732 in brackets (grey).
733
734

| Species | Genomic compartments | | | |
|---|---|---|---|---|
| | **Chrom** | **Rh01** | **Rh02** | **Rh03** |
| **gsA** | - | - | - | - |
| **gsB** | 0.4393 [0.29-0.60] | 0.3593 [0.20-0.53] | 0.2192 [0.02-0.44] | 0.4484 [0.26-0.66] |
| **gsC** | 0.4399 [0.24-0.67] | 0.3116 [0.10-0.56] | 0.2413 [-0.03-0.57] | 0.4682 [0.26-0.71] |
| **gsD** | 0.3568 [0.19-0.54] | 0.3247 [0.17-0.50] | 0.1383 [-0.09-0.41] | 0.3444 [0.16-0.56] |
| **gsE** | 0.4013 [0.24-0.58] | 0.3650 [0.22-0.52] | 0.2235 [0.03-0.45] | 0.4203 [0.25-0.61] |

735
736

# Supplemetary Information for the manuscript: Recombination facilitates adaptive evolution in rhizobial soil bacteria

Maria Izabel A. Cavassim*, Stig U. Andersen, Thomas Bataillon, and Mikkel H. Schierup*

## Contents

* Corresponding authors:

- Maria Izabel A. Cavassim, izabelcavassim@gmail.com
- Mikkel H. Schierup, mheide@birc.au.dk

This PDF file includes:

- Figures S1 to S13
- Tables S1 to S4
- Supplementary Information References

## List of Figures

S1 Intra and interspecies diversity . . . 3
S2 Population genetics statistics of core orthologs . . . 4
S3 Transition/Transversion (kappa), Synonymous sites, and Nonsynonymous sites distributions across core genes. . . . 4
S4 Population structure and the SFS of species C (gsC) . . . 5
S5 Model selection of DFE under interspecies estimation of $\alpha$ . . . 6
S6 Model selection of DFE for intraespecific estimation of $\alpha$ . . . 7
S7 Permutations tests of significance . . . 8
S8 Distribution of R/$\theta$ and $D'$ across species and their correlation to $r^2$ . . . 9
S9 The proportion of adaptive evolution and R/$\theta$ . . . 10
S10 The proportion of adaptive evolution and $D'$ . . . 11
S11 The rate of adaptive and non-adaptive evolution relative to neutral divergence (measured by R/$\theta$) . . . 12
S12 The rate of adaptive and non-adaptive evolution relative to neutral divergence (measured by $D'$) . . . 13
S13 Recombination rate by genomic compartment . . . 14

# List of Tables

S1 Recombination variation across species and genomic compartments . . . . . 5
S2 The proportion of shared polymorphisms among species. . . . . 6
S3 Models used to compute the Distribution of Fitness Effect (DFE). . . . . 7
S4 Likelihood differences among DFE models and parameters estimated . . . . . 15

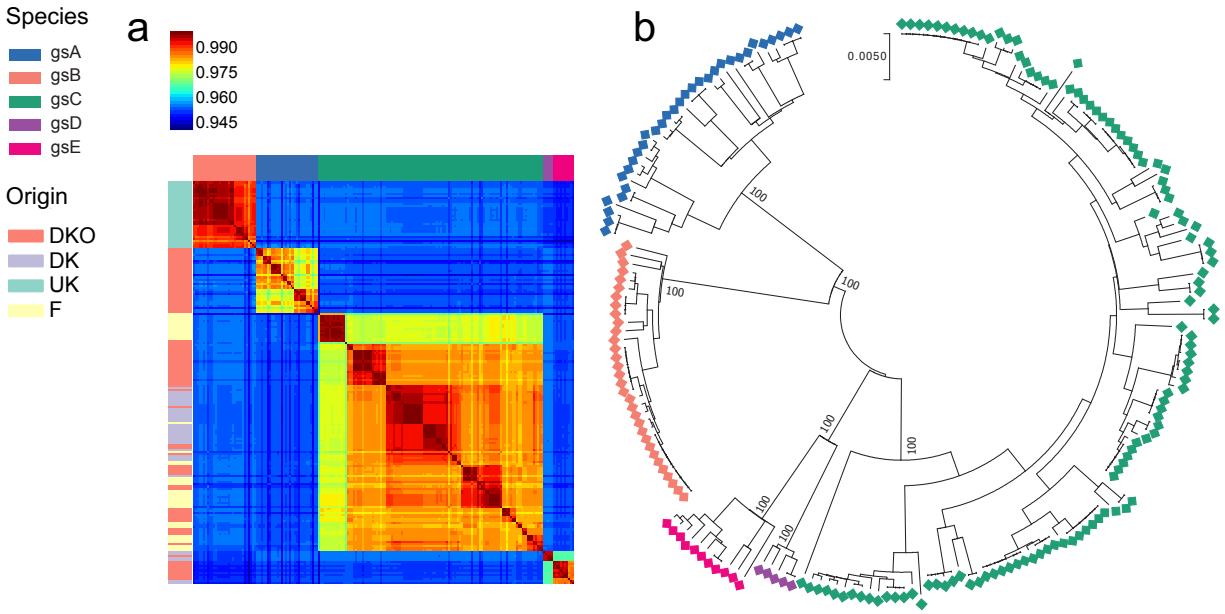


Figure S1: Pairwise average nucleotide identity (ANI) estimates (196x196) for concatenated sequences of 305 housekeeping genes (**Fig. A**, see Cavassim et al., 2020 for details). The top x-axis coloured bars represent the species classification (gsA-gsE) based on 95 percent ANI threshold (Konstantinidis et al. 2006). The y-axis corresponds to the sampling origin of strains and are indicated by coloured bars (Denmark organic fields in red (DKO), Denmark conventional fiels in purple (DK), France in yellow (F), and United Kingdom in green (the UK)). Species phylogeny based on the concatenation of 305 housekeeping genes using the neighbourjoining method (**Fig. B**). Tips are coloured by species (gsA-gsE) and the numbers in nodes indicate the bootstrap value.

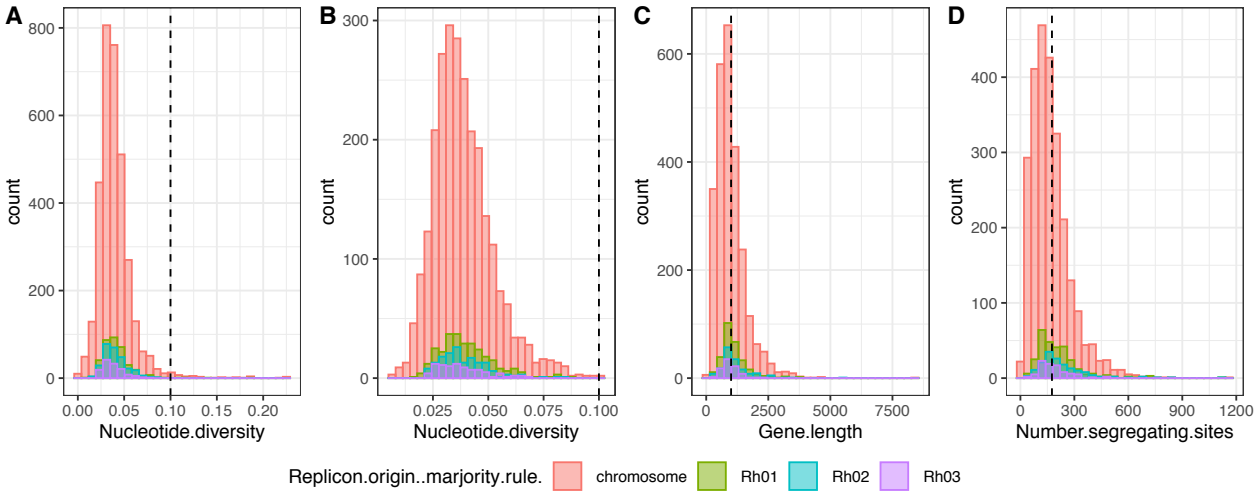


Figure S2: **Population genetics statistics of core orthologs**. The nucleotide diversity ($\pi$) distribution, gene length and the number of segregating sites based on a total of 4202 core genes. To compute polymorphism and divergence counts we required high quality gene-alignments with little evidence of horizontal gene transfer. The nucleotide diversity distribution of core genes (**Fig. A**) was used to impose a nucleotide diversity threshold (<=0.10). Other thresholds were also set (see Material and Methods), and a total of 1116 genes were discarded (**Fig. B**). The remaining genes were on average 1000 base pairs long (**Fig. C**) with an average of 175 segregating sites (across all 196 strains) (**Fig. D**).

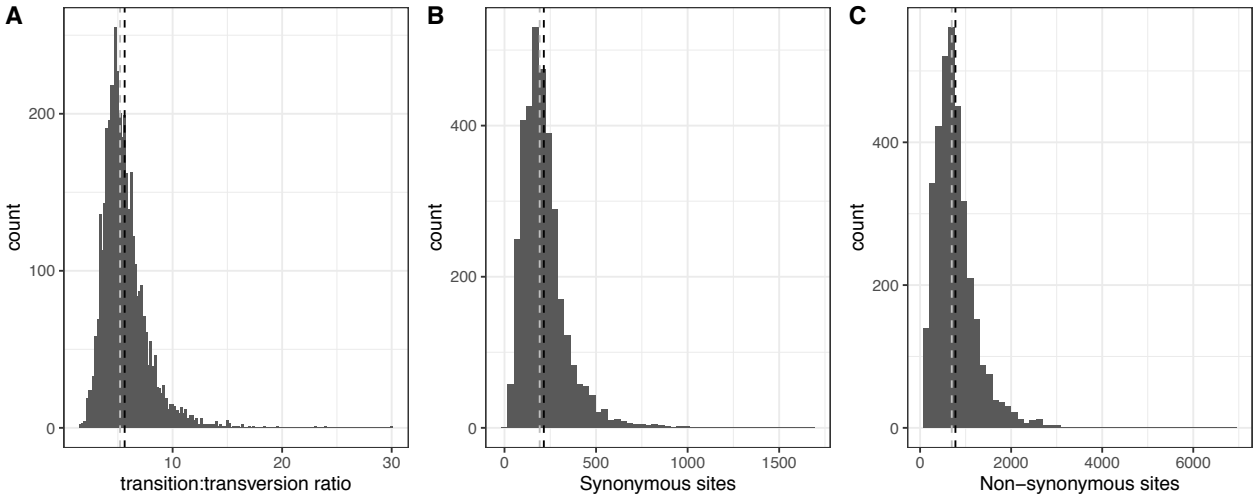


Figure S3: **Transition/Transversion (kappa), Synonymous-sites, and Nonsynonymous sites distributions across core genes**. To calculate the expected number of synonymous and nonsynonymous sites we first computed the transition/transversion ratio (kappa) across 3086 genes (**Fig. A**). Black and grey dashed lines correspond to the mean and median kappa values, respectively. Synonymous and nonsynonymous sites were computed based on the kappa median (5.20), and the distributions of counts for both site types are shown in **Fig. B-C**.

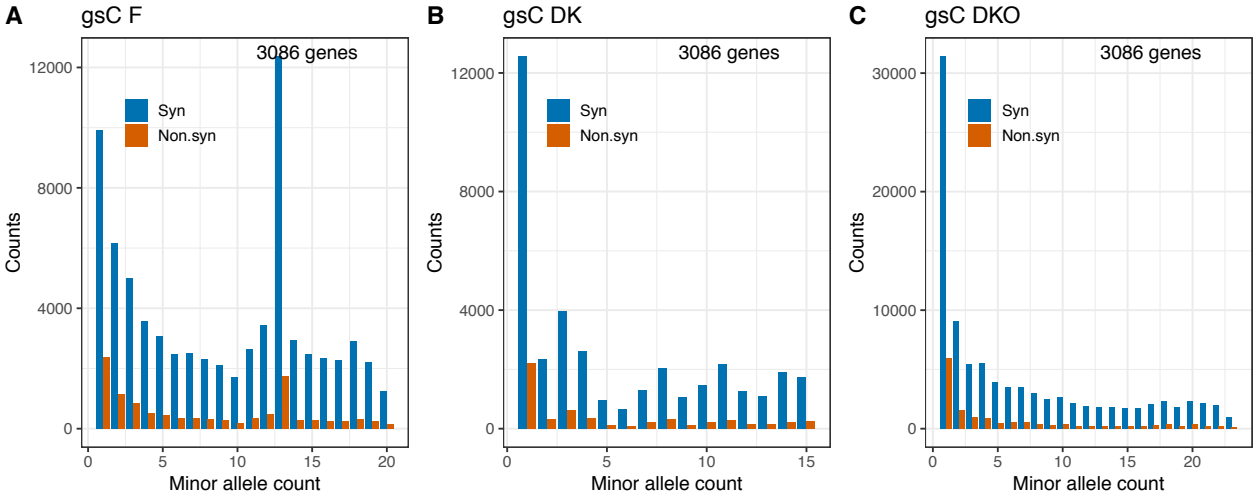


Figure S4: **Population structure and the SFS of species C (gsC)**. Isolates from gsC were found in different countries (France, **Fig. A**, and in Denmark, **Fig. B-C**). The gsC strains found in Denmark were isolated from soils subjected to different agricultural treatments (breeding field trials (gsC DK, **Fig. B**) and organic fields (gsC DKO, **Fig. C**)). The shape of SFS varies across different origins, with an evident deviation from the "L" shaped patterns (many rare alleles and fewer frequent alleles) observed in strains sampled from France (**Fig. A**).

Table S1: **The recombination rate variation across species and their genomic compartments** (chromosome, chromids (Rh01 and Rh02) and plasmid (Rh03)).

| Species | Genomic compartments | Mean $r^2$ | Median $r^2$ | Gene counts | Max Median $r^2$ differences |
|---|---|---|---|---|---|
| gsA | Chrom | 0.334 | 0.298 | 2060 | 0.04 |
| gsA | Rh01 | 0.327 | 0.283 | 271 | - |
| gsA | Rh02 | 0.285 | 0.267 | 154 | - |
| gsA | Rh03 | 0.354 | 0.312 | 82 | - |
| gsB | Chrom | 0.551 | 0.517 | 777 | 0.26 |
| gsB | Rh01 | 0.629 | 0.583 | 119 | - |
| gsB | Rh02 | 0.357 | 0.323 | 101 | - |
| gsB | Rh03 | 0.360 | 0.322 | 64 | - |
| gsC | Chrom | 0.303 | 0.266 | 2054 | 0.165 |
| gsC | Rh01 | 0.276 | 0.236 | 255 | - |
| gsC | Rh02 | 0.138 | 0.126 | 147 | - |
| gsC | Rh03 | 0.113 | 0.101 | 81 | - |
| gsD | Chrom | 0.747 | 0.777 | 418 | 0.28 |
| gsD | Rh01 | 0.754 | 0.804 | 77 | - |
| gsD | Rh02 | 0.556 | 0.521 | 94 | - |
| gsD | Rh03 | 0.668 | 0.611 | 36 | - |
| gsE | Chrom | 0.448 | 0.417 | 1351 | 0.21 |
| gsE | Rh01 | 0.670 | 0.578 | 122 | - |
| gsE | Rh02 | 0.452 | 0.423 | 121 | - |
| gsE | Rh03 | 0.657 | 0.632 | 50 | - |

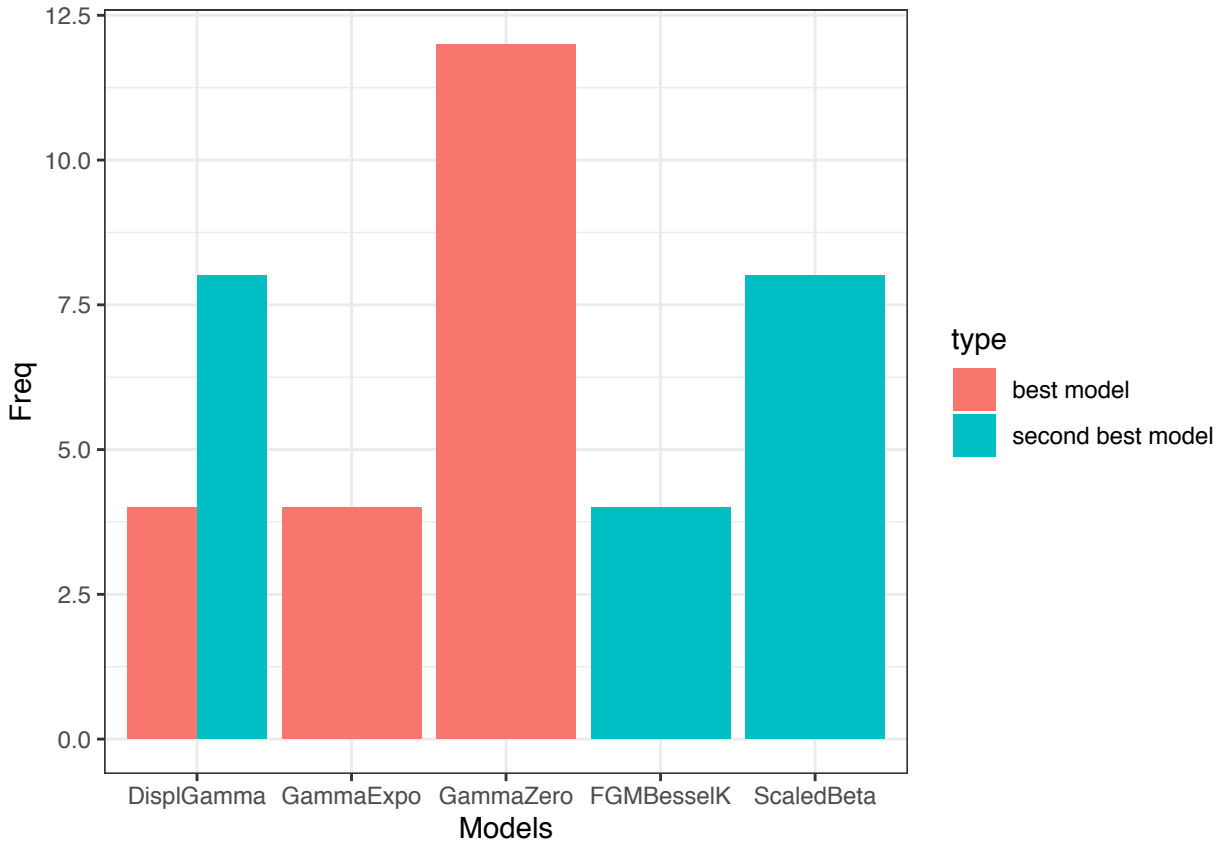


Figure S5: **Model selection of DFE models for the between species estimation of $\alpha$.** To estimate $\alpha$ an assumption of the shape of the distribution of fitness effects (DFE) is required. To determine the model of the DFE that best fit the data, a variety of DFE distribution models were tested, with assumptions described in the Sup. Table 3. For each model and each SFS, the Akaike's information criterion (AIC) was computed. The best model was defined as the model that had most often the lowest AIC. The frequency of the best and second best models are shown. The GammaZero model was found to be the best model 12 out of 20 times.

Table S2: **The proportion of shared polymorphisms ('shared poly') among species**. Pn, and Ps refer to the total number of non-synonymous and synonymous polymorphisms, respectively.

| Species1 | Species2 | Pn (sp1) | Ps (sp1) | Pn (sp2) | Ps (sp2) | Shared non poly | Shared syn poly | Prop shared poly (sp1) | Prop shared poly (sp2) |
|---|---|---|---|---|---|---|---|---|---|
| gsA | gsB | 12974 | 86783 | 4457 | 26682 | 406 | 5686 | 0.061 | 0.196 |
| gsA | gsC | 12974 | 86783 | 18700 | 107889 | 1774 | 21531 | 0.234 | 0.184 |
| gsA | gsD | 12974 | 86783 | 2325 | 15724 | 239 | 3271 | 0.035 | 0.194 |
| gsA | gsE | 12974 | 86783 | 5069 | 37227 | 412 | 6027 | 0.065 | 0.152 |
| gsB | gsC | 4457 | 26682 | 18700 | 107889 | 543 | 6843 | 0.237 | 0.058 |
| gsB | gsD | 4457 | 26682 | 2325 | 15724 | 61 | 967 | 0.033 | 0.057 |
| gsB | gsE | 4457 | 26682 | 5069 | 37227 | 100 | 1869 | 0.063 | 0.047 |
| gsC | gsD | 18700 | 107889 | 2325 | 15724 | 291 | 3758 | 0.032 | 0.224 |
| gsC | gsE | 18700 | 107889 | 5069 | 37227 | 513 | 7213 | 0.061 | 0.183 |
| gsD | gsE | 2325 | 15724 | 5069 | 37227 | 124 | 1618 | 0.097 | 0.041 |

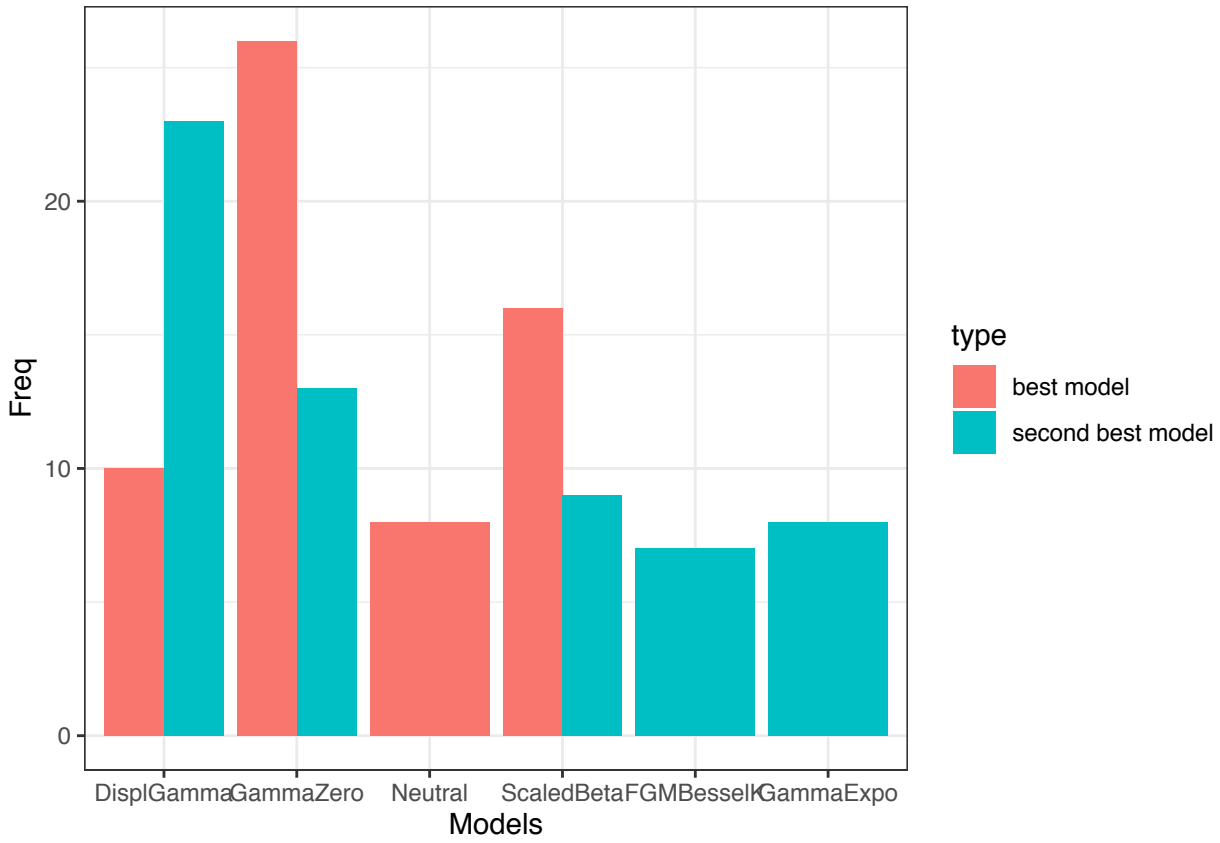


Figure S6: **Model selection of DFE for intraespecific estimation of** $\alpha$. The Akaike's information criterion (AIC) of competing models were computed (see **Fig. S4** legend). The GammaZero model was found to be the best model 30 out of 60 times.

Table S3: **Modelling the distribution of fitness effects (DFE)**. Description of the models used to compute the distribution of fitness effects and alpha using the software Grapes (Galtier et al., 2016).

| Model | Assumptions |
|---|---|
| Neutral | alpha = 1 − [(pN/pS)/(dN/dS)], considers a discrete DFE, with a class of neutral mutations and a class of strongly deleterious mutations, which do not contribute to polymorphism or divergence. |
| GammaZero | It assumes the existence of weakly deleterious nonsynonymous mutations, modeled as a continuous, negative Gamma distribution. |
| GammaExpo | Builds upon GammaZero by adding a proportion of weakly advantageous mutations assumed to be exponentially distributed. |
| GammaGamma | Builds upon GammaZero by adding a proportion of weakly advantageous mutations assumed to be gamma distributed. |
| DisplGamma | ScaledBeta model is similar to the Neutral model in including a class of strongly deleterious mutations, but assumes a Beta-shaped distribution of weak-effect mutations |
| ScaledBeta | Uses a Beta-shaped distribution of mild negative and positive effects |
| FGMBesselK | DFE derive from two distinct versions of Fisher's geometric model (see Lourenco et al., 2011) |

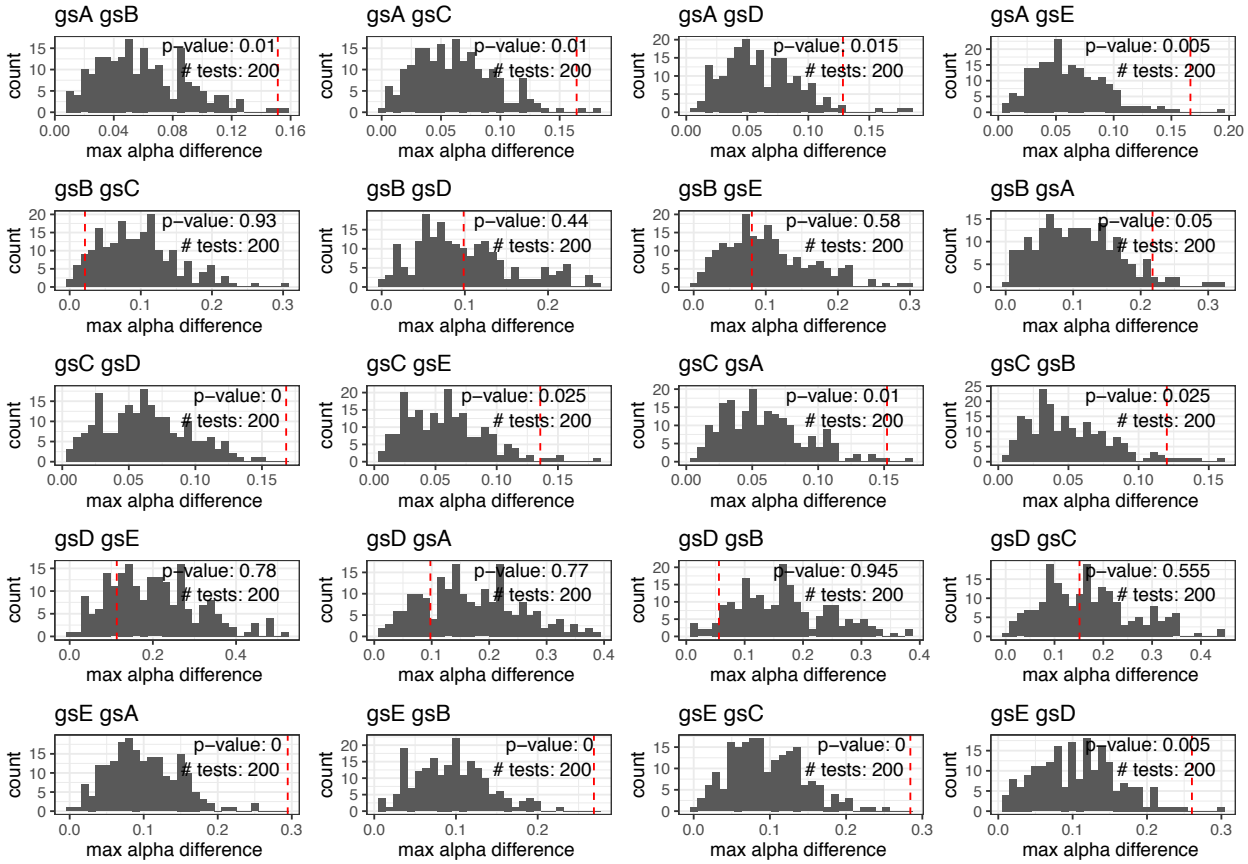


Figure S7: **Null distributions of intragenic differences in $\alpha$ across recombination classes**. For each pairwise estimates of $\alpha$ ($\alpha_{species1species2}$), the polymorphism data from a focal species (species 1) is compared against the divergence counts of an outgroup (species 2), and vice-versa ($\alpha_{species2species1}$)). We simulated a null-distribution by permuting the recombination classes of genes (200 times), and for each permuted dataset computing the magnitude of differences between $\alpha$ estimates across recombination classes ($\Delta_\alpha = max_\alpha$ - $min_\alpha$). The $\Delta_\alpha$s distributions (histograms) were compared against each observed $\Delta_\alpha$ (red dashed line) (see **Fig. 3**). P-values were computed as the number of simulations above or equal to the observed $\Delta_\alpha$ (red dashed lines).

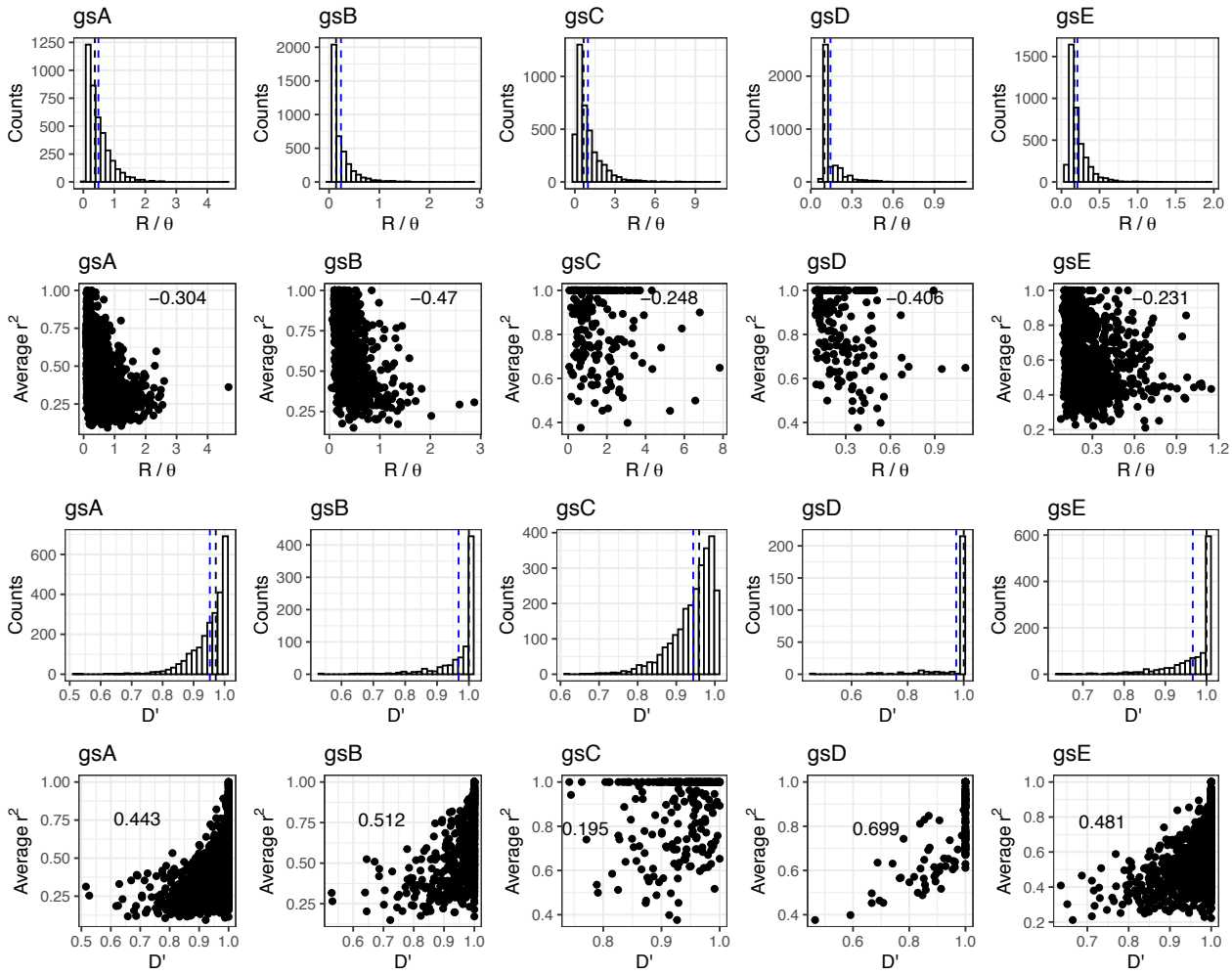


Figure S8: **Distribution of R/$\theta$ and $D$ across species and their correlation to $r^2$.** R/$\theta$ (relative importance of Recombination (R) to mutation ($\theta$)) was computed using the software ClonalFrameML (Didelot and Wilson., 2015) by analizing all core orthologs (3086 genes). The distribution of R/$\theta$ across species is shown (first row), blue and black dashed lines correspond to median and mean R/$\theta$ values, respectively. The Pearson correlation between R/$\theta$ and $r^2$ is shown for each species (second row)). $D'$ distributions across species are shown in the third row, and the Pearson's correlation between $D'$ and $r^2$ metrics are shown in the fourth row.

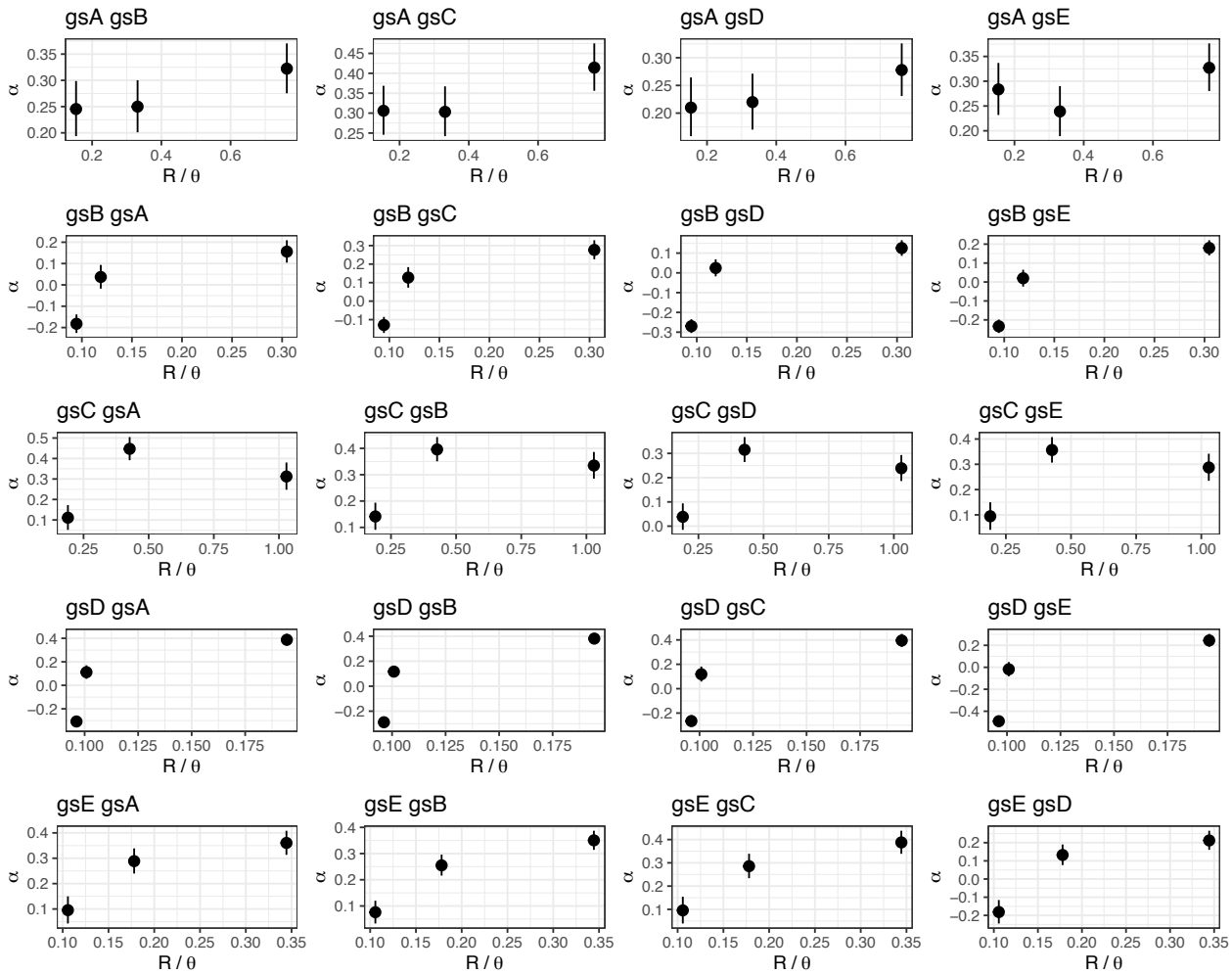


Figure S9: **The proportion of adaptive evolution by classes of R/$\theta$.** The proportion of adaptive evolution ($\alpha$) is computed for each class of recombination (classes of recombination are based on averages values of R/$\theta$). For each pairwise estimates of $\alpha$ (species1 species2), the polymorphism data from a focal species (species 1) is compared against the divergence counts of an outgroup (species 2), and vice-versa. The $\alpha$ estimates (and their associated confidence intervals) were obtained using the best fitting DFE model (GammaZero).

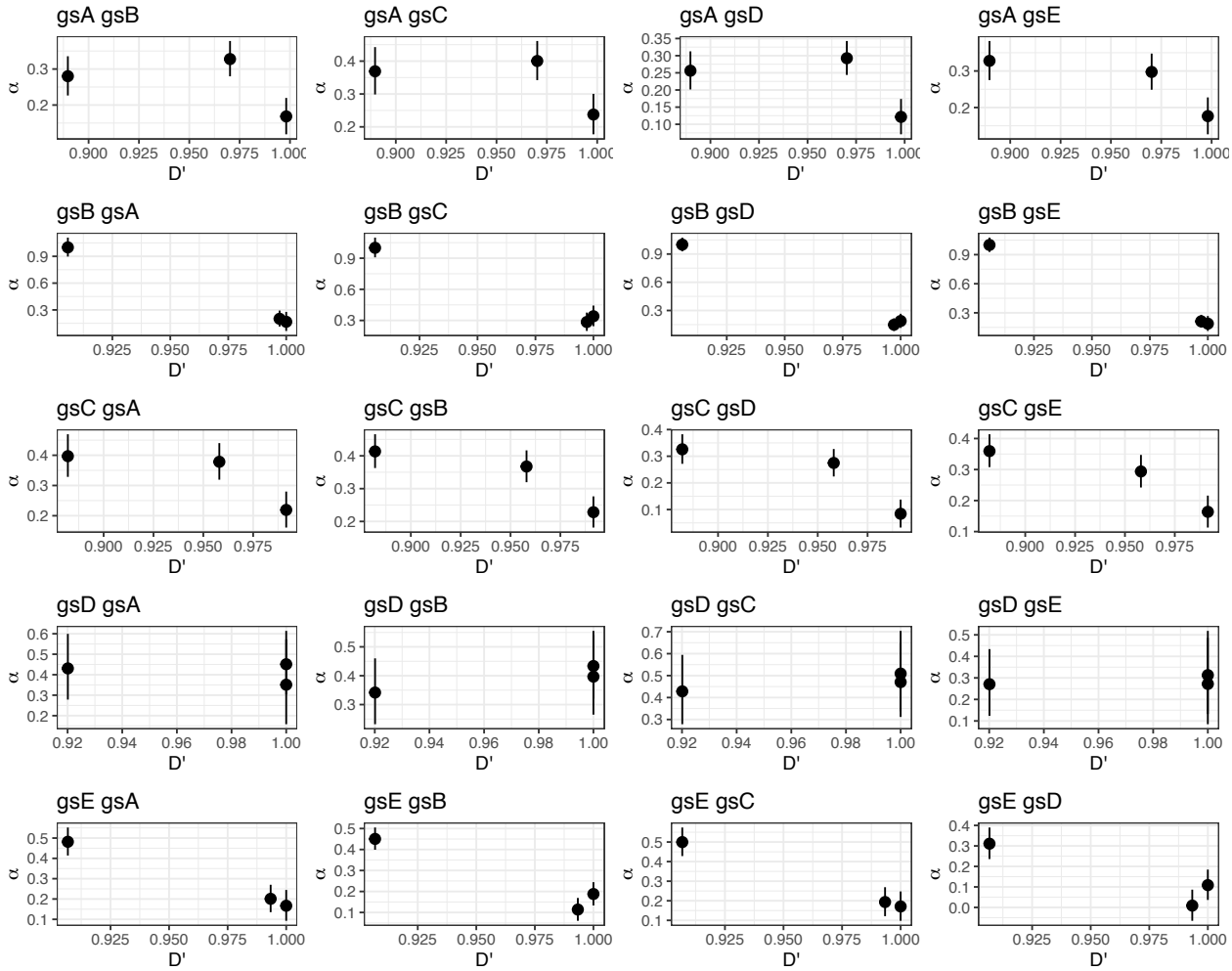


Figure S10: **The proportion of adaptive evolution by classes of $D'$.** The proportion of adaptive evolution ($\alpha$) is computed for each class of recombination (based on $D'$). For each pairwise estimates of $\alpha$ (species1 species2), the polymorphism data from a focal species (species 1) is compared against the divergence counts of an outgroup (species 2), and vice-versa. The $\alpha$ estimates (and their associated confidence intervals) were obtained using the best fitting DFE model (GammaZero).

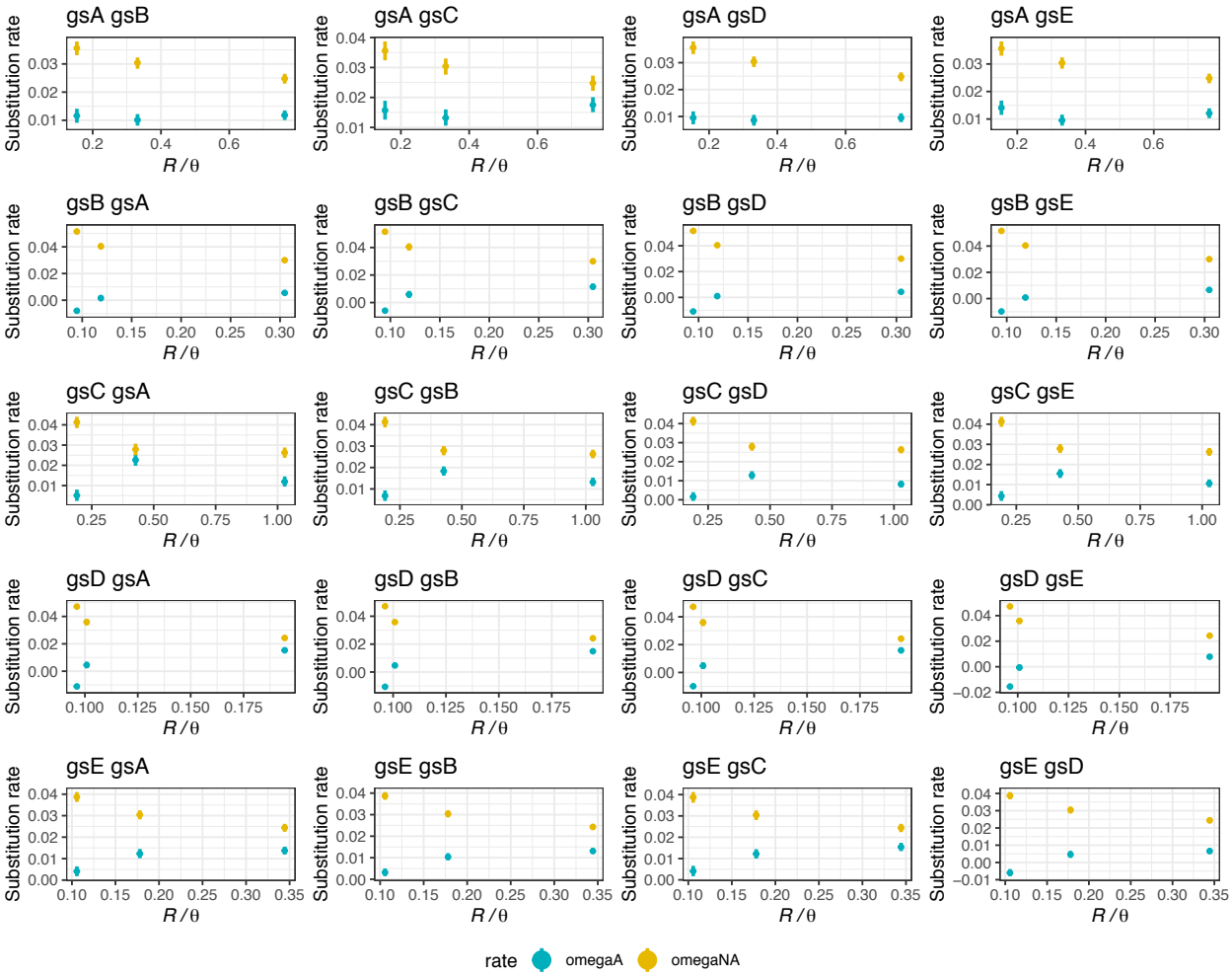


Figure S11: The rate of non-adaptive substitution relative to neutral divergence ($\omega_{NA}$), and the rate of adaptive evolution $\omega_A$ were computed for each recombination class (measured by R/$\theta$). For each pairwise estimates of $\omega_{NA}$ and $\omega_A$, the polymorphism data from one species (species 1) is compared against the divergence counts of an outgroup (species 2), and vice-versa. The $\omega_{NA}$ and $\omega_A$ estimates (and their associated confidence intervals) were obtained using the GammaZero model.)

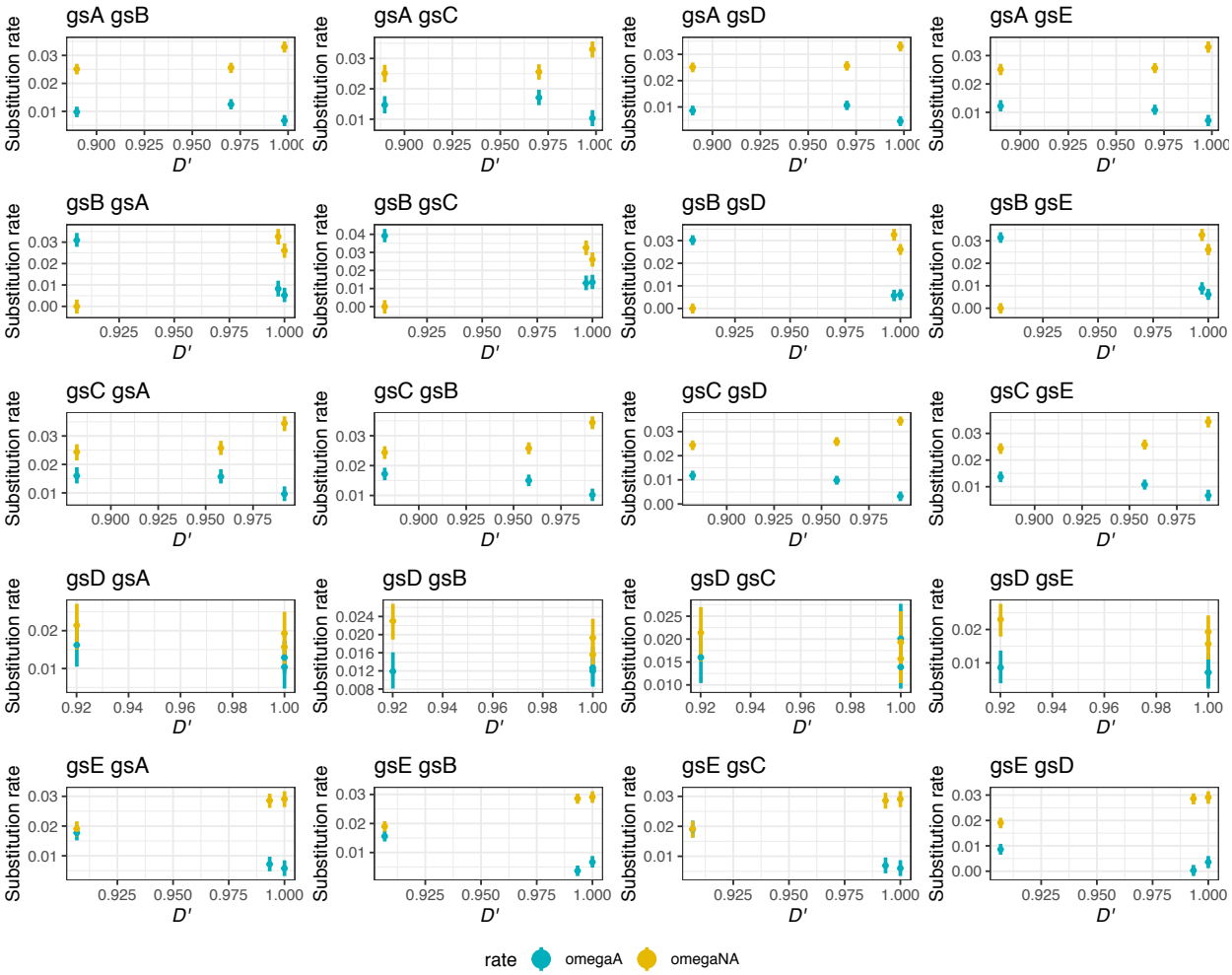


Figure S12: The rate of non-adaptive substitution relative to neutral divergence ($\omega_{NA}$), and the rate of adaptive evolution $\omega_A$ were computed for each recombination class (measured by $D'$). For each pairwise estimates of $\omega_{NA}$ and $\omega_A$, the polymorphism data from one species (species 1) is compared against the divergence counts of an outgroup (species 2), and vice-versa. The $\omega_{NA}$ and $\omega_A$ estimates (and their associated confidence intervals) were obtained using the GammaZero model.

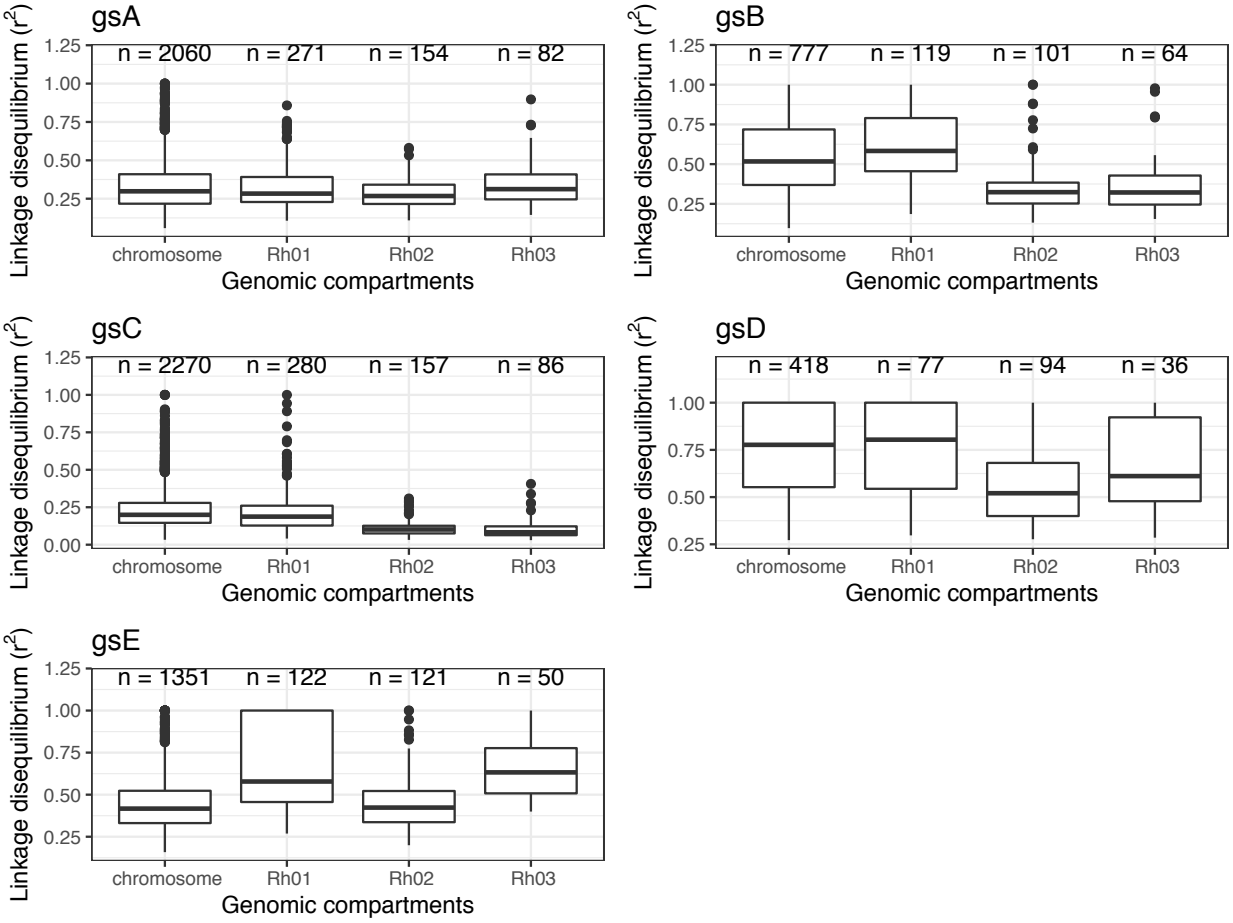


Figure S13: Recombination rate by genomic compartiment. Recombination rate by genomic compartment was comupted based on patterns of linkage disequilibrium (see Methods). Genomic compartiments were previously defined based on sequence variation of a plasmid partitioning gene (repA) that is essential for stable maintenance of nearly all plasmids in Rhizobium (Cavassim et al., 2020). These compartments correspond to the chromosome, chromids (Rh01 and Rh02), and plasmid (Rh03) (Harrison et al., 2010).

Table S4: **Likelihood differences among DFE models and parameters estimated**. Pn and Ps refer to the total number of non-synonymous and synonymous polymorphisms, respectively. Dn and Ds refers to the number of non-synonymous and synonymous substitutions. The total number of non-synonymous and synonymous sites available in the within-species refers to Lpn and Lps, respectively. The number of non-synonymous and synonymous sites observed between-species alignments is denoted by dn and ds, respectively. The models and their corresponding $\alpha$ estimates are shown for a subset of all comparisons.

| pn | Lpn | ps | Lps | dn | ds | Lds | model | X.param | alpha | lnL | AIC | geno |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 18700 | 2413139 | 107889 | 668782.2 | 5993 | 37205 | 668782.2 | Neutral | 61 | -0.076 | -1263.394 | 2648.789 | gsC gsB |
| 18700 | 2413139 | 107889 | 668782.2 | 5120 | 36783 | 668782.2 | GammaZero | 62 | 0.255 | -590.638 | 1305.276 | gsC gsD |
| 18700 | 2413139 | 107889 | 668782.2 | 5120 | 36783 | 668782.2 | ScaledBeta | 63 | 0.736 | -590.177 | 1306.354 | gsC gsD |
| 18700 | 2413139 | 107889 | 668782.2 | 5120 | 36783 | 668782.2 | DisplGamma | 63 | 0.255 | -590.638 | 1307.276 | gsC gsD |
| 18700 | 2413139 | 107889 | 668782.2 | 5120 | 36783 | 668782.2 | GammaGamma | 65 | 0.254 | -590.638 | 1311.277 | gsC gsD |
| 18700 | 2413139 | 107889 | 668782.2 | 5120 | 36783 | 668782.2 | GammaExpo | 64 | 0.243 | -595.904 | 1319.808 | gsC gsD |
| 18700 | 2413139 | 107889 | 668782.2 | 5120 | 36783 | 668782.2 | FGMBesselK | 64 | 0.631 | -714.724 | 1557.448 | gsC gsD |
| 18700 | 2413139 | 107889 | 668782.2 | 5120 | 36783 | 668782.2 | Neutral | 61 | -0.245 | -1263.310 | 2648.620 | gsC gsD |
| 18700 | 2413139 | 107889 | 668782.2 | 5322 | 36086 | 668782.2 | GammaZero | 62 | 0.297 | -590.648 | 1305.296 | gsC gsE |
| 18700 | 2413139 | 107889 | 668782.2 | 5322 | 36086 | 668782.2 | ScaledBeta | 63 | 0.751 | -590.187 | 1306.374 | gsC gsE |
| 18700 | 2413139 | 107889 | 668782.2 | 5322 | 36086 | 668782.2 | DisplGamma | 63 | 0.297 | -590.648 | 1307.296 | gsC gsE |
| 18700 | 2413139 | 107889 | 668782.2 | 5322 | 36086 | 668782.2 | GammaGamma | 65 | 0.296 | -590.648 | 1311.296 | gsC gsE |
| 18700 | 2413139 | 107889 | 668782.2 | 5322 | 36086 | 668782.2 | GammaExpo | 64 | 0.285 | -595.914 | 1319.827 | gsC gsE |
| 18700 | 2413139 | 107889 | 668782.2 | 5322 | 36086 | 668782.2 | FGMBesselK | 64 | 0.652 | -714.734 | 1557.468 | gsC gsE |
| 18700 | 2413139 | 107889 | 668782.2 | 5322 | 36086 | 668782.2 | Neutral | 61 | -0.175 | -1263.320 | 2648.639 | gsC gsE |
| 2325 | 2413139 | 15724 | 668782.2 | 5600 | 40199 | 668782.2 | GammaZero | 6 | 0.251 | -31.135 | 74.270 | gsD gsA |
| 2325 | 2413139 | 15724 | 668782.2 | 5600 | 40199 | 668782.2 | DisplGamma | 7 | 0.256 | -31.127 | 76.255 | gsD gsA |
| 2325 | 2413139 | 15724 | 668782.2 | 5600 | 40199 | 668782.2 | ScaledBeta | 7 | 0.560 | -31.127 | 76.255 | gsD gsA |
| 2325 | 2413139 | 15724 | 668782.2 | 5600 | 40199 | 668782.2 | GammaExpo | 8 | 0.074 | -31.127 | 78.255 | gsD gsA |
| 2325 | 2413139 | 15724 | 668782.2 | 5600 | 40199 | 668782.2 | FGMBesselK | 8 | 0.697 | -31.127 | 78.255 | gsD gsA |
| 2325 | 2413139 | 15724 | 668782.2 | 5600 | 40199 | 668782.2 | GammaGamma | 9 | 0.118 | -31.127 | 80.255 | gsD gsA |
| 2325 | 2413139 | 15724 | 668782.2 | 5600 | 40199 | 668782.2 | Neutral | 5 | -0.061 | -41.088 | 92.175 | gsD gsA |
| 2325 | 2413139 | 15724 | 668782.2 | 9790 | 70104 | 668782.2 | GammaZero | 6 | 0.254 | -31.693 | 75.385 | gsD gsB |
| 2325 | 2413139 | 15724 | 668782.2 | 9790 | 70104 | 668782.2 | DisplGamma | 7 | 0.258 | -31.685 | 77.369 | gsD gsB |
| 2325 | 2413139 | 15724 | 668782.2 | 9790 | 70104 | 668782.2 | ScaledBeta | 7 | 0.564 | -31.685 | 77.369 | gsD gsB |
| 2325 | 2413139 | 15724 | 668782.2 | 9790 | 70104 | 668782.2 | GammaExpo | 8 | 0.062 | -31.685 | 79.369 | gsD gsB |
| 2325 | 2413139 | 15724 | 668782.2 | 9790 | 70104 | 668782.2 | FGMBesselK | 8 | 0.697 | -31.685 | 79.369 | gsD gsB |
| 2325 | 2413139 | 15724 | 668782.2 | 9790 | 70104 | 668782.2 | GammaGamma | 9 | 0.107 | -31.685 | 81.369 | gsD gsB |
| 2325 | 2413139 | 15724 | 668782.2 | 9790 | 70104 | 668782.2 | Neutral | 5 | -0.059 | -41.645 | 93.290 | gsD gsB |
| 2325 | 2413139 | 15724 | 668782.2 | 5120 | 36783 | 668782.2 | GammaZero | 6 | 0.250 | -31.046 | 74.092 | gsD gsC |
| 2325 | 2413139 | 15724 | 668782.2 | 5120 | 36783 | 668782.2 | DisplGamma | 7 | 0.255 | -31.038 | 76.076 | gsD gsC |
| 2325 | 2413139 | 15724 | 668782.2 | 5120 | 36783 | 668782.2 | ScaledBeta | 7 | 0.560 | -31.038 | 76.076 | gsD gsC |
| 2325 | 2413139 | 15724 | 668782.2 | 5120 | 36783 | 668782.2 | GammaExpo | 8 | 0.076 | -31.038 | 78.076 | gsD gsC |
| 2325 | 2413139 | 15724 | 668782.2 | 5120 | 36783 | 668782.2 | FGMBesselK | 8 | 0.697 | -31.038 | 78.076 | gsD gsC |
| 2325 | 2413139 | 15724 | 668782.2 | 5120 | 36783 | 668782.2 | GammaGamma | 9 | 0.120 | -31.038 | 80.076 | gsD gsC |
| 2325 | 2413139 | 15724 | 668782.2 | 5120 | 36783 | 668782.2 | Neutral | 5 | -0.062 | -40.998 | 91.997 | gsD gsC |
| 2325 | 2413139 | 15724 | 668782.2 | 4713 | 39531 | 668782.2 | GammaZero | 6 | 0.125 | -31.041 | 74.081 | gsD gsE |
| 2325 | 2413139 | 15724 | 668782.2 | 4713 | 39531 | 668782.2 | DisplGamma | 7 | 0.130 | -31.033 | 76.065 | gsD gsE |
| 2325 | 2413139 | 15724 | 668782.2 | 4713 | 39531 | 668782.2 | ScaledBeta | 7 | 0.486 | -31.033 | 76.065 | gsD gsE |
| 2325 | 2413139 | 15724 | 668782.2 | 4713 | 39531 | 668782.2 | GammaExpo | 8 | -0.082 | -31.033 | 78.065 | gsD gsE |
| 2325 | 2413139 | 15724 | 668782.2 | 4713 | 39531 | 668782.2 | FGMBesselK | 8 | 0.646 | -31.033 | 78.065 | gsD gsE |
| 2325 | 2413139 | 15724 | 668782.2 | 4713 | 39531 | 668782.2 | GammaGamma | 9 | -0.030 | -31.033 | 80.065 | gsD gsE |
| 2325 | 2413139 | 15724 | 668782.2 | 4713 | 39531 | 668782.2 | Neutral | 5 | -0.240 | -40.993 | 91.986 | gsD gsE |
| 5069 | 2413139 | 37227 | 668782.2 | 5764 | 39059 | 668782.2 | GammaZero | 9 | 0.270 | -63.961 | 145.922 | gsE gsA |
| 5069 | 2413139 | 37227 | 668782.2 | 5764 | 39059 | 668782.2 | DisplGamma | 10 | 0.270 | -63.961 | 147.922 | gsE gsA |
| 5069 | 2413139 | 37227 | 668782.2 | 5764 | 39059 | 668782.2 | ScaledBeta | 10 | 0.374 | -64.157 | 148.315 | gsE gsA |
| 5069 | 2413139 | 37227 | 668782.2 | 5764 | 39059 | 668782.2 | GammaExpo | 11 | 0.232 | -63.892 | 149.784 | gsE gsA |
| 5069 | 2413139 | 37227 | 668782.2 | 5764 | 39059 | 668782.2 | FGMBesselK | 11 | 0.600 | -64.455 | 150.910 | gsE gsA |
| 5069 | 2413139 | 37227 | 668782.2 | 5764 | 39059 | 668782.2 | GammaGamma | 12 | 0.868 | -63.896 | 151.792 | gsE gsA |
| 5069 | 2413139 | 37227 | 668782.2 | 5764 | 39059 | 668782.2 | Neutral | 8 | 0.077 | -86.821 | 189.642 | gsE gsA |
| 5069 | 2413139 | 37227 | 668782.2 | 9536 | 66442 | 668782.2 | GammaZero | 9 | 0.250 | -64.478 | 146.957 | gsE gsB |
| 5069 | 2413139 | 37227 | 668782.2 | 9536 | 66442 | 668782.2 | DisplGamma | 10 | 0.250 | -64.478 | 148.957 | gsE gsB |
| 5069 | 2413139 | 37227 | 668782.2 | 9536 | 66442 | 668782.2 | ScaledBeta | 10 | 0.356 | -64.675 | 149.349 | gsE gsB |
| 5069 | 2413139 | 37227 | 668782.2 | 9536 | 66442 | 668782.2 | GammaExpo | 11 | 0.205 | -64.409 | 150.819 | gsE gsB |
| 5069 | 2413139 | 37227 | 668782.2 | 9536 | 66442 | 668782.2 | FGMBesselK | 11 | 0.588 | -64.973 | 151.945 | gsE gsB |
| 5069 | 2413139 | 37227 | 668782.2 | 9536 | 66442 | 668782.2 | GammaGamma | 12 | 0.864 | -64.413 | 152.826 | gsE gsB |
| 5069 | 2413139 | 37227 | 668782.2 | 9536 | 66442 | 668782.2 | Neutral | 8 | 0.051 | -87.338 | 190.677 | gsE gsB |
| 5069 | 2413139 | 37227 | 668782.2 | 5322 | 36086 | 668782.2 | GammaZero | 9 | 0.270 | -63.882 | 145.763 | gsE gsC |
| 5069 | 2413139 | 37227 | 668782.2 | 5322 | 36086 | 668782.2 | DisplGamma | 10 | 0.270 | -63.882 | 147.763 | gsE gsC |
| 5069 | 2413139 | 37227 | 668782.2 | 5322 | 36086 | 668782.2 | ScaledBeta | 10 | 0.374 | -64.078 | 148.156 | gsE gsC |
| 5069 | 2413139 | 37227 | 668782.2 | 5322 | 36086 | 668782.2 | GammaExpo | 11 | 0.233 | -63.813 | 149.625 | gsE gsC |

## References


- N. Galtier, Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. PLoS Genet. 12, e1005774 (2016).

- J. Lourenço, N. Galtier, S. Glémin, Complexity, pleiotropy, and the fitness effect of mutations. Evolution 65, 1559–1571 (2011).

- X. Didelot, D. J. Wilson, ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput. Biol. 11, e1004041 (2015).

- Konstantinidis, Konstantinos T., Alban Ramette, and James M. Tiedje. "The bacterial species definition in the genomic era." Philosophical Transactions of the Royal Society B: Biological Sciences 361.1475 (2006)

- Cavassim, Maria Izabel A., et al. "Symbiosis genes show a unique pattern of introgression and selection within a Rhizobium leguminosarum species complex." Microbial genomics 6.4 (2020).

- Harrison, Peter W., et al. "Introducing the bacterial ‘chromid’: not a chromosome, not a plasmid." Trends in microbiology 18.4 (2010): 141-148.