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,*]

[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, CA

*__Corresponding authors:__ E-mails: izabelcavassim@gmail.com; mheide@birc.au.dk.

Associate editor: Daniel Falush

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 (α) depends on the recombination rate in bacteria. We analyze 3,086 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 α 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 α 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 nonadaptive 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 α in the core genome of prokaryote species in agreement with studies in eukaryotes.

Key words: **adaptive evolution, rhizobium, recombination, beneficial mutations.**

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 (α) 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.

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 1) a reduction in the number of neutral polymorphisms, 2) the accumulation of slightly deleterious polymorphisms, and 3) 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 (α).

The parameter α can also be viewed as the relative proportion between the rate of amino acid changes fixed by positive selection (ω_a) and the rate of nonadaptive amino acid changes relative to neutral (ω_{na}) (as: $\alpha = \omega_a/(\omega_a + \omega_{na})$) (Galtier 2016; Moutinho et al. 2020). Distinguishing between ω_a and ω_{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 ω_{na}) and increased probability of fixation for beneficial mutations (a higher ω_a).

Empirical evidence based on population genomics data supports theory with a positive correlation between recombination and α 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 nonmodel 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 nitrogen fixation in white clover (*Trifolium repens*) nodules. We have previously reported the full genomic sequence of 196 isolates (Cavassim et al. 2020). Of 22,115 orthologous gene groups identified among the 196 strains, 4,204 genes are present in all isolates (the core genome). Although substantial adaptive evolution might be attributed to accessory genes through their gains and losses via horizontal gene transfer (HGT) (Young et al. 2006; Tian et al. 2010; Porter et al. 2017; Cavassim et al. 2020), we focus here on genes vertically inherited and examine how much variation in rates of homologous recombination explains variation in α and its components (ω_a and ω_{na}).

Our previous analyses (Cavassim et al. 2020) showed that the 196 strains cluster into five closely related species (2–5% average nucleotide divergence), with HGT between these species only affecting the nitrogen fixation genes and a few well-defined genomic regions––that we exclude in the present analysis. This species complex thus offers a unique opportunity among prokaryotes to estimate the rates of fixation of amino acid changes by adaptive evolution from isolates sampled from natural populations––enabling multiple comparisons of polymorphism and divergence patterns among species. Our analyses demonstrate that the rate of adaptive protein evolution increases with the recombination rate in this species complex.

Results and discussion

To estimate the proportion of adaptive evolution (α) across this *Rhizobium* species complex and study how α covaries with intraspecific recombination rate estimates, we restricted analyses to polymorphism data from regions of the core genome without evidence of recent interspecies HGT.

Of the 18 species observed within this species complex (Young et al. 2021), we have collected genomic data for five species (gsA–gE). Across all five species (196 strains, gsA: 32, gsB: 32, gsC: 112, gsD: 5, gsE: 11) (Supplementary fig. S1), a total of 22,115 orthologous gene groups were previously identified (Cavassim et al. 2020); of those, 4,204 genes are present in all strains (core genes). Most core genes are found in the large chromosome (3,304 genes), but some are located in the chromids (Rh01, Rh02) and in one of the plasmids (Rh03) (see Harrison et al. 2010; Cavassim et al. 2020). The chromosome, chromids, and the plasmid are hereafter referred to as genomic compartments. To assess the effect of intraspecies homologous recombination on adaptive evolution using a high-quality data set, we filtered out genes that showed evidence of recent interspecies HGT or unexpectedly high rates of nucleotide diversity (see Materials and Methods) (Supplementary fig. S2), leaving a total of 3,086 genes (total alignment length: 3091179 bp) and 334040 variable sites for analysis (Supplementary fig. S3).

First, we estimated nucleotide diversity, intragenic linkage disequilibrium (LD), and the site frequency spectrum (SFS) (see Materials and Methods) within each species (fig. 1*a*–*c*). The average nucleotide diversity, π, an estimator of $2N_e\mu$ in haploids, is significantly different among genomic compartments (fig. 1*a* and Supplementary table S1). Across the species, π differs by up to a factor of 4.5 (gsA: 0.018, gsB: 0.0045, gsC: 0.0140, gsD: 0.00512, and gsE: 0.008), with the most polymorphic species being gsA and the least gsB. If we assume similar mutation rates among these closely related species, nucleotide diversity differences reflect interspecies differences in long-term effective population size, N_e.

When recombination occurs, we expect that levels of nonrandom association between pairs of alleles, quantified by measures such as r^2 (see Materials and Methods), decay with genomic distance (LD decay). To evaluate the recombination rate differences among the five species, we used within-species polymorphism data and computed the average intragenic LD decay for each gene in each species. We observed a rapid decay of LD within the first 1,000 base pairs for all species, suggesting substantial amounts of within-species homologous recombination (fig. 1*b*). The slower decay observed in species gsB either reflects a lower per generation recombination rate or a smaller effective population size (N_e). The latter is consistent with the low level of nucleotide diversity measured in gsB. To reliably estimate interspecies differences in r^2, we used genes with at least 10 informative sites within each species while also excluding variants only found in one strain (singletons) and evaluated their r^2 distributions separately (fig. 1*c*). As expected, the species with the most striking LD decay (gsC) has the lowest r^2 median (median r^2: 0.248) and the opposite is also true (gsD, median r^2: 1.00). In summary, these species can be ranked by their recombination levels, from the most recombining to the least, as follows: gsC (median r^2: 0.248) > gsA (median r^2: 0.341) > gsE (median r^2: 0.561) > gsB (median r^2: 0.651) > gsD (medianr^2: 1.00).

Next, we computed the folded SFS of synonymous and nonsynonymous mutations within each species. Overall, both synonymous and nonsynonymous SFSs differ from the "L" shaped patterns (many rare alleles and fewer frequent alleles) expected in a stationary population at mutation–selection–drift equilibrium (fig. 1*d*). The observed excess of intermediate frequency single-nucleotide polymorphism (SNPs) indicates

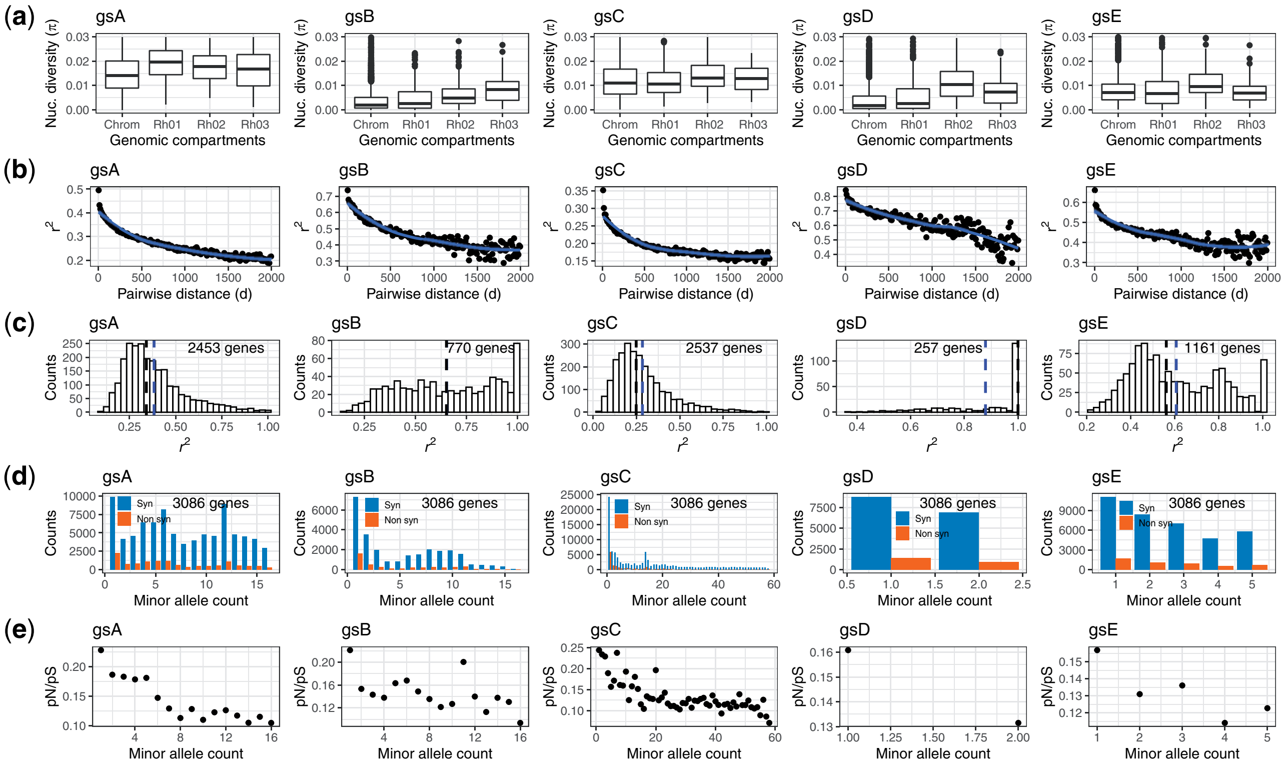

FIG. 1. Population genetics parameters across five species. (*a*) Nucleotide diversity (π) across 3,086 genes distributed along with genomic compartments (chromosome, chromids: Rh01, Rh02, and plasmid: Rh03). To exclude outliers only genes with $\pi \leq 0.03$ are shown. (*b*) Intragenic LD measured via the decay of r^2 for all core genes (3,086 genes). The curve fitting line (in blue) is from a local regression method (loess). (*c*) LD (r^2) distribution across genes. Only genes with at least 10 segregating sites were kept and singletons were excluded (gsA: 2,453 genes, gsB: 770, gsC: 2,537, gsD: 257, and gsE: 1,161). The black and blue dashed lines correspond to the median and mean r^2, respectively. (*d*) SFS counts of synonymous and nonsynonymous sites by minor allele count based on all core genes (3,086 genes). (*e*) The ratio of nonsynonymous to synonymous polymorphisms by minor allele count.

the presence of population structure in some of the species. The effect of population structure is particularly evident in gsC, and this excess is likely driven by strains isolated from French soils (Supplementary 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 (see Materials and Methods). 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, and 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. 1*e*), 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 α (Eyre-Walker and Keightley 2009).

We used GRAPES (Galtier 2016) to estimate the distribution of fitness effects (DFEs) (Eyre-Walker and Keightley 2007) and the proportion of adaptive evolution (α) 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 also accounting for the effect of demography. The significant amount of shared polymorphism among species (Supplementary 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 α 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 (Supplementary 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 that the GammaZero model overall provides the best fit to the SFS data (Supplementary fig. S5). This model assumes the existence of weakly deleterious nonsynonymous

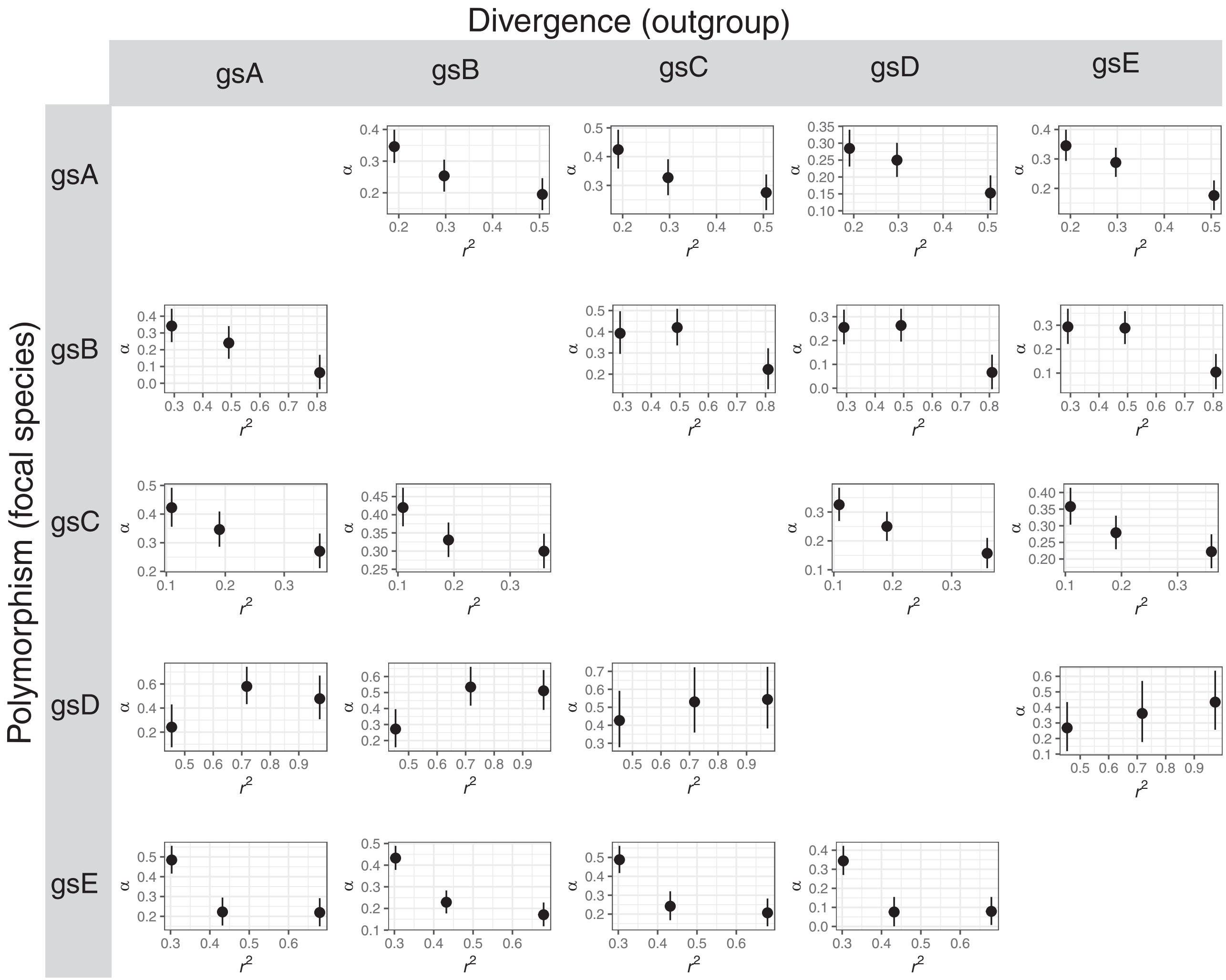

Fig. 2. The proportion of adaptive evolution (α) by classes of recombination. For each pairwise estimates of α, the polymorphism data from one species (left in title) is compared against the divergence counts of an outgroup (right in title), and vice versa. Results are divided into three equally sized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

Table 1. The Proportion of Adaptive Evolution (α) across Pairs of Species.

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)	–

The α estimates were computed based on the best-fitting DFE model (GammaZero) (Supplementary table S3). For each pairwise estimate of α ($\alpha_{\text{species1 species2}}$), the polymorphism data from a focal species (rows) is compared against the divergence counts of an outgroup (columns), and vice-versa ($\alpha_{\text{species2 species1}}$). Confidence intervals (CIs) are displayed in brackets and numbers in parentheses represent the α ranking (in decreasing order) by outgroup (by column).

mutations, modeled as a continuous Gamma distribution (Galtier 2016).

The proportion of adaptive evolution was first computed between all combinations of "mirror" species ($\alpha_{\text{species1 species2}}$), in which "species 2" is used as outgroup (divergence) for "species 1" (polymorphism) and vice versa ($\alpha_{\text{species2 species1}}$). This yielded 20 combinations in total. Because "mirror" species share an identical history of divergence, their α estimates can be considered as "biological replicates" (Galtier 2016) (table 1). Except for the comparison between gsA and gsB, in which differences between $\alpha_{\text{gsA gsB}}$ and $\alpha_{\text{gsB gsA}}$ exceeded 0.09, the overall discrepancy in the values estimated between mirror species does not exceed 0.1. Using each species' focal polymorphism data, we calculated four α estimates by comparing it to the divergence counts of the remaining species (table 1). The most recombining species (gsC) is observed to have the highest α across all outgroups used, while the least recombining species (gsD) had the lowest α in three out of the four cases.

We then investigated whether intraspecies differences in recombination rate affect the amount of adaptive evolution (α) estimated. For each species, we split genes into three equally sized recombination classes based on their average r^2 values and computed α for each class using the GammaZero model (Supplementary fig. S6 and Supplementary table S4). Because we only kept genes with at least 10 informative sites, the number of genes evaluated across species was different (see fig. 1c). For most species comparisons (gsA, gsB, gsC, and gsE), there is a decrease in the proportion of adaptive evolution with a reduction in recombination (increase in r^2) (fig. 2). Except for cases in which we used gsD polymorphisms to estimate α, where all comparisons were nonsignificant, all the other species pairwise comparisons led to at least one significant difference (based on nonoverlapping confidence intervals [CIs]) between recombination classes. The low sample size of gsD (five strains) and its skewed LD distribution (fig. 1c) may have reduced statistical power to discriminate among recombination classes and to estimate α reliably.

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

The parameter α can also be viewed as the relative proportion between the rate of amino acid changes fixed by positive selection (ω_a) and the rate of nonadaptive amino acid changes (ω_{na}): $\alpha = \omega_a/(\omega_a + \omega_{na})$. Thus, an increase in α with recombination could be due to either an increase in the rate of adaptive substitutions, a decrease in the rate of

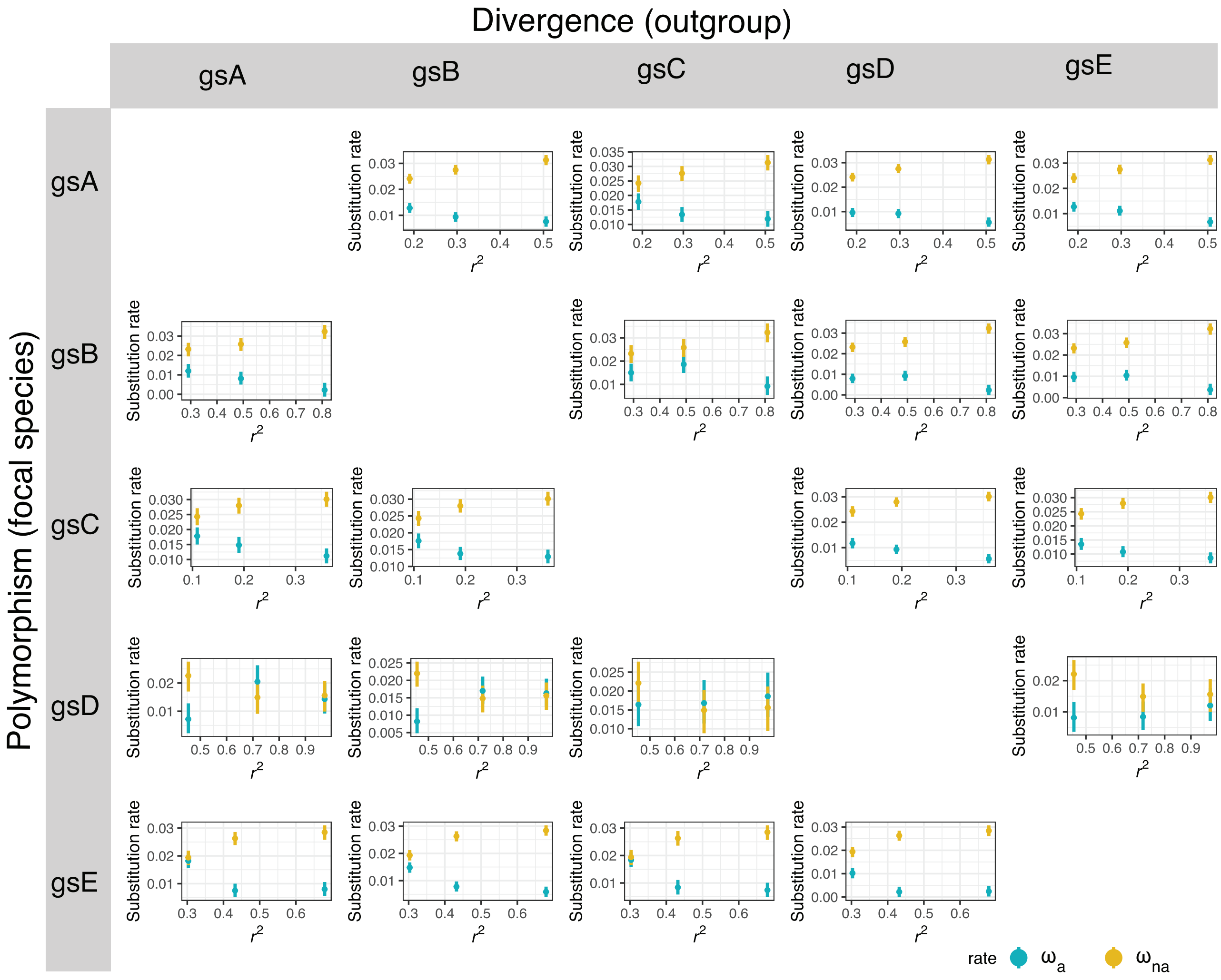

Fig. 3. The rates of adaptive (ω_a) and non − adaptive (ω_{na}) evolution by classes of recombination. For each pairwise estimates of ω_a (in blue) and ω_{na} (in yellow), the polymorphism data from one species are compared against the divergence counts of an outgroup, and vice-versa. Results are divided 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 ω_a and ω_{na} is observed in most pairwise comparisons. The rate estimates (ω_a, ω_{na}) and their associated confidence intervals (CIs) were obtained using the best fitting DFE model (GammaZero).

nonadaptive substitutions, or both. Figure 3 shows that ω_a increases with recombination rate whereas ω_{na} decreases with recombination rate for most combinations and that the quantitative effects are almost equal in magnitude. Thus, classes of genes evolving under higher recombination rates exhibited lower rates of nonadaptive substitution and increased rates of fixation of adaptive variation. This matches the predictions from selective interference theory (Felsenstein 1974; Barton 1994; McVean and Charlesworth 2000).

To evaluate the robustness of these results, we computed two alternative measures of recombination (R/θ, and D'). We then made new recombination classes and evaluated how each recombination measure correlated with α. R/θ measures the importance of recombination (R) relative to mutation (θ) across sequences (Vos and Didelot 2009; Didelot and Wilson 2015), while D' measures LD between sites using a different metric than r^2 (see Materials and Methods). Although the distributions of R/θ *and* D' across genes are different than that of r^2 (fig. 1*c* and Supplementary fig. S8), these three measures are not independent (Pearson's correlation between $r^2$and R/θ or D' ranged from 0.19 to 0.70) (Supplementary fig. S8). For most species comparisons, the trend between α and recombination remains consistent: the higher the amount of recombination (measure by R/θ or D'), the higher α is (Supplementary figs. S9 and S10). The mean rates of adaptive evolution (ω_a) obtained among classes based on D' or on R/θ estimates also generally agree with those estimated using r^2 (Supplementary figs. S11 and S12).

The genomes of the *Rhizobium* species comprise three genomic compartments (chromosomes, chromids, and plasmids) that may have undergone different selection regimes. We tested that by building distinct SFSs for each genomic compartment (see Materials and Methods). The mean *a* estimate differs slightly among genomic compartments with higher α observed in core genes sampled from the chromosome and the smallest plasmid (Rh03) (table 2). However, the sampling variance of estimates is large, and differences observed are not statistically significant. We also evaluated the rate of recombination among these genomic units (Supplementary fig. S13), recombination is heterogeneous within and across species, as observed for nucleotide diversity (fig. 1*a*). We conclude that we are underpowered to detect differences in the effects of adaptive evolution as a function of recombination between these genomic compartments.

In this study, we applied a methodology that estimates the proportion of amino acid changes that have been fixed by positive selection (α) while also estimating the individual components of α (ω_a and ω_{na}). We have only included genes that are present in all sampled genomes––the so-called core genome. Because the methodology used requires summarizing the data by building an SFS, the analyses cannot be readily extended to the accessory genome––given that many accessory genes are a result of HGT (Popa and Dagan 2011), a DFE computed from this source will not necessarily reflect the DFE of the studied species. The accessory genome represents roughly 40% of the genome of these species (Cavassim et al. 2020), and likely also contributes to adaptive evolution (e.g., acquired symbiotic ability [Kumar et al. 2015; Cavassim et al. 2020]) as also shown in other studies (e.g., antibiotic resistance in *Staphylococcus aureus* [Harris et al. 2010]; or adaptation to a new ecological niche [Ochman et al. 2000; Wiedenbeck and Cohan 2011]).

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

Table 2. The Proportion of Adaptive Evolution (α) across Genomic Compartments.

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 to 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 to 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]

The α estimates were computed based on the best-fitting DFE model (GammaZero) (Supplementary table S3). For each genomic compartment (chrom = chromosome; Rh01 and Rh02 = chromids; Rh03 = plasmid), we compared the polymorphism data from species gsA against the divergence counts of an outgroup (rows). Confidence intervals are displayed in brackets.

Conclusion

We have found that five bacterial species within the species complex *Rhizobium leguminosarum* display different yet high levels of recombination. The estimates of α ranged between 0.07 and 0.39 among species. These estimates are lower than those based on 410 orthologs observed in *E. coli* (0.58, CI = 0.45–0.68) but close to estimates from *Salmonella enterica* (0.34, CI = 0.14–0.50) previously reported (Charlesworth and Eyre-Walker 2006)––however, in this study population, fluctuations were not accounted for.

Levels of recombination correlate––both across and within species––with higher amounts of adaptive evolution estimated either as the rate of adaptive substitutions (ω_a) or as the proportion of amino acid changes that have been fixed by positive selection (α). For instance, the most recombining species (gsC) consistently exhibited the largest α, independent of the outgroup used. Within each species, we also find a positive correlation between intragenomic recombination rate and α. This is both due to a higher estimated rate of adaptive evolution (ω_a) and a lower estimated rate of nonadaptive evolution (ω_{na}), suggesting that recombination both increases the fixation probability of advantageous variants and decreases the probability of fixation of deleterious variants. These findings are robust to the measure of recombination (r^2, R/θ, and D') used to define classes and the choice of outgroup used for computing divergence. Despite variation in recombination rate among genomic compartments, we did not observe significant differences in adaptive evolution among them.

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

Material and Methods

Identification of Orthologous Genes

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

From the assembled genomes (Cavassim et al. 2020), we predicted protein-coding sequences using prokka (Seemann 2014) (v1.12); this resulted in a total of 1468264 protein-coding sequences. To predict orthologous genes from these sequences, we used Proteinortho (Lechner et al. 2014; Seemann 2014) (v5.16b) with default parameters except for enabling the synteny flag. We identified a total of 22,115 orthologous gene groups, including a total of 17,911 orthologous observed in at least two strains (accessory genes), and 4,204 orthologous found in all 196 strains (core genes).

Orthologous gene groups were aligned using clustalo (Sievers et al. 2011) (v.1.2.0) in a codon-aware manner. To determine the genetic relationship among all 196 strains, we previously calculated their pairwise average nucleotide identity (ANI) across 305 conserved orthologous gene alignments (Cavassim et al. 2020). Under the 95% ANI threshold that delineates species boundaries (Konstantinidis et al. 2006), we demonstrated that these 196 *Rhizobium* strains constitute five distinct *R. leguminosarum* species (gsA, gsB, gsC, gsD, and gsE) (Supplementary fig. S1). To ensure that we had a high-quality orthologous data set for extracting segregating sites, we filtered it further (see below).

Filtering Out Orthologous Gene Groups with Evidence of Interspecies HGT or Misassigned Orthologous Gene Groups

To evaluate the impact of intraspecific homologous recombination on adaptive evolution, we excluded genes that showed signals of recent HGT across the five species analyzed. We have previously developed and applied a phylogenetic method to quantify HGT (introgression score) (Cavassim et al. 2020). This method evaluates the possible number of shifts from one species to another in a given phylogenetic tree. The pipeline takes a gene tree as an input and traverses the tree––using the depth-first search approach––searching deeper in the tree whenever possible. Once the tip is reached the species classification for that given strain is stored. A list containing the species in order of search is collected for the entire tree, the introgression score is then computed as the number of shifts from one species to another in the list minus the set of species plus 1.

We previously showed that most of the core genes shared among the present species respect the species-tree topology (introgression score = 0) (Cavassim et al. 2020). The exceptions are genes sitting in the symbiosis conjugative plasmids and two chromosomal islands (introgression score > 7). To ensure that we were only analyzing high-quality gene alignments, with no evidence of misassigned orthologous gene groups and with little evidence of HGT, we imposed some restrictions. We only accepted genes that passed the following criteria: 1) were present in every strain (196 strains), 2) with a nucleotide diversity (π) below 0.1 (see Supplementary fig. S2*a*), 3) identifiable replicon origin (chromosome and chromids), 4) and with an introgression score $\leq$ 3. A total of 3,086 out of 4,204 core genes were kept, and of these, 2,550 genes were found in the chromosome, 288 genes in chromid Rh01, 160 genes in chromid Rh02, and 88 genes in plasmid Rh03.

Variant Calling

To identify SNPs along with our high-quality set of core genes, we evaluated each gene codon-aware alignment using a custom python script https://github.com/izabelcavassim/Popgen_bacteria. For a given core gene alignment and position, we first counted the number of unique nucleotides (A, C, T, and G). Only sites containing two unique nucleotides were considered variable sites (bi-allelic SNPs). SNP matrices were then built and encoded as follows: major alleles were encoded as 1 and minor alleles as 0. The nucleotide diversity (π), gene length, and the distributions of segregating sites across core genes are described in Supplementary fig. S2*b*–*d*.

Transition Transversion Rate Bias (Kappa) and Expected Number of Synonymous and Nonsynonymous Sites

Because transitions are more often synonymous at third codon positions than are transversions, to correctly identify the expected number of synonymous (*Lps*) and nonsynonymous sites (*Lpn*), we first estimated the average transition/transversion rate bias (kappa) (Ina 1995) across species. To this end, we followed the methods described in (Yang and Nielsen 2000) and used two classes of sites: 4-fold-degenerate sites at the third codon positions and nondegenerate sites. Mutations at the 4-fold-degenerate sites are synonymous, and therefore kappa at those sites should reflect only the mutational bias. All mutations at nondegenerate sites are nonsynonymous and were also used to estimate kappa. We computed an average kappa by combining these two classes based on equations (8)–(11) of Yang and Nielsen (2000). These equations have been implemented within the CodonSeq class in Biopython (Cock et al. 2009) (private function "_count_site_YN00()"), and these private functions were adapted to our data set.

To estimate a common kappa for each gene alignment (including all species and strains), we averaged estimates from pairwise analyses across 50 randomly chosen strains. The kappa distribution has a mean of 5.6 and a median of 5.20 (Supplementary fig. S3*a*), we used the median to compute the expected number of synonymous and nonsynonymous sites. To this end, we followed the methods described by (Ina 1995) and modified by (Yang and Nielsen 2000)—–also implemented within Biopython. A total of 284,742 synonymous, 49,298 nonsynonymous sites were counted (Supplementary fig. S3*b and c*).

Divergence Sites and Shared Polymorphisms

For each pair of species (a focal and an outgroup), we evaluated their variable sites and computed the number of shared synonymous (*pS*) and nonsynonymous (*pN*) polymorphisms. Given a bi-allelic SNP (0 and 1), we considered shared polymorphic sites as sites for which both alleles (0,1) were segregating in both species (Supplementary table S2). We restricted the estimates of divergence to those sites for which we had variable sites across species. We classified synonymous (d_S) and nonsynonymous divergent sites (d_N) as those sites in which we observed fixed differences between a focal species and an outgroup.

Calculating the Folded SFS

One can infer the DFEs from SFS data (Eyre-Walker and Keightley 2009). Because of the amount of shared polymorphism among the present species (Supplementary table S2), it becomes problematic to confidently distinguish ancestral from derived polymorphisms (Hernandez et al. 2007). Therefore, we chose to estimate the DFE using a method that uses the folded SFSs of synonymous and nonsynonymous sites (Galtier 2016). To this end, we built the folded synonymous and nonsynonymous SFSs by tabulating the observed counts of the minor allele frequencies. The synonymous and nonsynonymous SFSs, and the divergence counts, were then used to estimate the DFE and the proportion of adaptive substitutions (α) across pairs of species.

Calculating the Strength of Purifying Selection

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

Estimation of Adaptive and Nonadaptive Nonsynonymous Substitutions Rates

Fitted parameters of the DFE were used to compute the expected d_N/d_S under the different models, which was compared with the observedd_N/d_S to estimate the adaptive substitution rate (ω_a); nonadaptive substitution rate (ω_{na}), and the proportion of adaptive substitutions (α) with $\omega_a = \alpha d_N / d_S$ and $\omega_{na} = (1-\alpha) d_N/d_S$.

To account for potential departures of the SFS from demographic equilibrium (assuming 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's 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 LD) 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 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 10 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 ($i = 1, \ldots, M$) 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 $\left(\frac{S_i - \mu_i}{\sigma_i}\right)$, resulting in a vector with mean 0 and variance 1. The LD was then calculated as a function of distance d (maximum 1,000 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:

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

$$r^2 = \mathrm{Cor}(X_i, X_j)^2.$$

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 $d \leq 1{,}000$ 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′)

The average LD within genes was also measured by D' (Lewontin 1964) as follow: given two locus i and j (with alleles A and a, observed in locus i and alleles B and b observed in locus j), we first computed the frequency of all possible haplotype combinations ($f_{AA}, f_{AB}, f_{Ab}, f_{ab}$) and allele frequencies (f_A, f_B, f_a, and f_b). The coefficient D was then computed as: $D_{AB} = f_{AB} - f_A \cdot f_B$ and D' was computed by standardizing $|D|$ by its maximum possible value 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 stored and computed similarly to the r^2 statistics (see above).

(3) ClonalFrameML

To estimate the changes in the clonal phylogeny by recombination (R), relative to mutation (θ) (R/θ), we used the software ClonalFrameML (Vos and Didelot 2009; Didelot and Wilson 2015). For each species, we first concatenated all core gene alignments (3,086 genes) to build the starting phylogenetic species tree using a maximum-likelihood approach (Raxml-ng [Stamatakis 2014]). We then input each phylogenetic tree within each gene alignment to estimate R/θ.

Calculating the Significance Levels between Recombination Classes

To test whether differences in α among recombination classes were statistically significant across species comparisons, we conducted a nonparametric test by shuffling genes among recombination classes (200 permutations) and recording the amplitude of differences between α estimates ($\Delta_\alpha = \max_\alpha - \min_\alpha$). We calculated a P-value by comparing the observed Δ_α against the simulated Δ_α distribution.

Estimation of Adaptive Substitutions by Genomic Compartments

To estimate adaptive evolution (α) by genomic compartments, we first down-sampled the genes from genomic compartments (chromosome, Rh01, Rh02) to reach the size of the smallest genomic compartment (Rh03). Due to the paucity of data, we chose to compute α using polymorphism data from the most polymorphic species (gsA) and contrasted it against each outgroup (gsB–gsE). The α estimates and their associated CIs were obtained using the GammaZero DFE model within GRAPES.

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

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was funded by grant no. 4105-00007A from Innovation Fund Denmark (S.U.A. and M.H.S.). The authors would like to thank industrial partners DLF Trifolium, SEGES, and Legume Technology Ltd. for their contribution to the field trials. They would also like to thank J. Peter, W. Young, Paula Tataru, Bjarni Vilhjálmsson, and Marjolaine Rousselle for their helpful discussions about this work.

Author Contributions

M.I.A.C., T.B., and M.H.S.: conceptualization, methodology, investigation, data curation, and visualization. M.I.A.C.: formal analysis and writing—original draft. S.U.A. and M.H.S.: resources. M.I.A.C., S.U.A., T.B., and M.H.S.: writing—review and editing. T.B. and M.H.S.: supervision. S.U.A. and M.H.S.: project administration and funding acquisition.

References

Akaike H. 1992. Information theory and an extension of the maximum likelihood principle. *Springer Ser Stat* 610–624. Available from: 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. *Mol Biol Evol* 27(4):848–861. Available from: 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(3):1247–1260.

Baltrus DA, Guillemin K, Phillips PC. 2008. Natural transformation increases the rate of adaptation in the human pathogen *Helicobacter pylori*. *Evolution* 62(1):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 single-cell sequencing. *J Comput Biol* 19(5):455–477. Available from: 10.1089/cmb.2012.0021.

Barton NH. 1994. The reduction in fixation probability caused by substitutions at linked loci. *Genet Res* 64(3):199–208. Available from: 10.1017/s0016672300032857.

Betancourt AJ, Welch JJ, Charlesworth B. 2009. Reduced effectiveness of selection caused by a lack of recombination. *Curr Biol* 19(8):655–660.

Campos JL, Halligan DL, Haddrill PR, Charlesworth B. 2014. The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. *Mol Biol Evol* 31(4):1010–1028.

Castellano D, Coronado-Zamora M, Campos JL, Barbadilla A, Eyre-Walker A. 2016. Adaptive evolution is substantially impeded by Hill–Robertson interference in *Drosophila*. *Mol Biol Evol* 33(2):442–455.

Cavassim MIA, Moeskjær S, Moslemi C, Fields B, Bachmann A, Vilhjálmsson BJ, Schierup MH, Young JPW, Andersen SU. 2020. Symbiosis genes show a unique pattern of introgression and selection within a species complex. *Microb Genom* 6. Available from: 10.1099/mgen.0.000351.

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

Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. *Mol Biol Evol* 23(7):1348–1356.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11):1422–1423.

Cohen E, Kessler DA, Levine H. 2005. Recombination dramatically speeds up evolution of finite populations. *Phys Rev Lett* 94(9):098102.Available from: 10.1103/physrevlett.94.098102.

Cooper TF. 2007. Recombination speeds adaptation by reducing competition between beneficial mutations in populations of *Escherichia coli*. *PLoS Biol* 5(9):e225.

Didelot X, Maiden MCJ. 2010. Impact of recombination on bacterial evolution. *Trends Microbiol* 18(7):315–322.

Didelot X, Wilson DJ. 2015. ClonalFrameML: Efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* 11(2):e1004041.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet* 8(8):610–618. Available from: 10.1038/nrg2146.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol* 26(9):2097–2108.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173(2):891–900.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* 158(3):1227–1234.

Felsenstein J. 1974. The evolutionary advantage of recombination. *Genetics* 78(2):737–756.

Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS Genet* 12(1):e1005774.

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

Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, et al. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327(5964):469–474.

Harrison PW, Lower RPJ, Kim NKD, Young JPW. 2010. Introducing the bacterial 'chromid': not a chromosome, not a plasmid. *Trends Microbiol*. 18(4):141–148.

Hernandez RD, Williamson SH, Bustamante CD. 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol Biol Evol* 24(8):1792–1800.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res* 8(3):269–294. Available from: 10.1017/s0016672300010156.

Hughes AL. 2005. Evidence for abundant slightly deleterious polymorphisms in bacterial populations. *Genetics* 169(2):533–538.

Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M. 2003. Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proc Natl Acad Sci USA* 100(26):15754–15757.

Ina Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J Mol Evol* 40(2):190–226.

Konstantinidis KT, Ramette A, Tiedje JM. 2006. The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* 361(1475):1929–1940.

Kumar N, Lad G, Giuntini E, Kaye ME, Udomwong P, Shamsani NJ, Young JPW, Bailly X. 2015. Bacterial genospecies that are not ecologically coherent: population genomics of Rhizobium leguminosarum. *Open Biol*. 5(1):140133.

Lechner M, Hernandez-Rosales M, Doerr D, Wieseke N, Thévenin A, Stoye J, Hartmann RK, Prohaska SJ, Stadler PF. 2014. Orthology detection combining clustering and synteny for very large datasets. *PLoS ONE* 9(8):e105015.

Levin BR, Cornejo OE. 2009. The population and evolutionary dynamics of homologous gene recombination in bacterial populations. *PLoS Genet* 5(8):e1000601.

Lewontin RC. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49(1):49–67.

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The *Drosophila melanogaster* genetic reference panel. *Nature* 482(7384):173–178.

McVean GA, Charlesworth B. 2000. The effects of Hill–Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* 155(2):929–944.

Moutinho AF, Bataillon T, Dutheil JY. 2020. Variation of the adaptive substitution rate between species and within genomes. *Evol Ecol* 34(3):315–338. Available from: 10.1007/s10682-019-10026-z.

Moutinho AF, Trancoso FF, Dutheil JY. 2019. The impact of protein architecture on adaptive evolution. *Mol Biol Evol* 36(9):2013–2028.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405(6784):299–304.

Ohta T. 1976. Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theor Popul Biol* 10(3):254–275.

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

Pool JE, Hellmann I, Jensen JD, Nielsen R. 2010. Population genetic inference from genomic sequence variation. *Genome Res* 20(3):291–300. Available from: 10.1101/gr.079509.108.

Popa O, Dagan T. 2011. Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol* 14(5):615–623.

Porter SS, Chang PL, Conow CA, Dunham JP, Friesen ML. 2017. Association mapping reveals novel serpentine adaptation gene clusters in a population of symbiotic Mesorhizobium. *ISME J* 11(1):248–262.

Presgraves DC. 2005. Recombination enhances protein adaptation in Drosophila melanogaster. *Curr Biol* 15(18):1651–1656.

Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. 2011. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics* 189(4):1427–1437.

Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30(14):2068–2069.

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

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313. Available from: 10.1093/bioinformatics/btu033.

Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics* 207(3):1103–1119.

Tian CF, Young JPW, Wang ET, Tamimi SM, Chen WX. 2010. Population mixing of *Rhizobium leguminosarum* bv. *viciae* nodulating *Vicia faba*: The role of recombination and lateral gene transfer. *FEMS Microbiol Ecol* 73(3):563–576. Available from: 10.1111/j.1574-6941.2010.00909.x.

Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. *ISME J* 3(2):199–208.

Wiedenbeck J, Cohan FM. 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev* 35(5):957–976.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17(1):32–43.

Young JPW, Crossman LC, Johnston AWB, Thomson NR, Ghazoui ZF, Hull KH, Wexler M, Curson ARJ, Todd JD, Poole PS, et al. 2006. The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol* 7(4):R34.

Young JPW, Moeskjær S, Afonin A, Rahi P, Maluk M, James EK, Cavassim MIA, Rashid MH-O, Aserse AA, Perry BJ, et al. 2021. Defining the *Rhizobium leguminosarum* species complex. *Genes*. 12:111. Available from: 10.3390/genes12010111.