# 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 1**

# 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 *Rhizobium leguminosarum* species. We find that $\alpha$ varies from 0.07 to 0.39 across species and is positively correlated with the level of recombination. We then 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. Generally, we found a significant increase in $\alpha$ with an increased recombination rate. 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 recombination facilitates adaptive evolution not only in eukaryotes, but also in prokaryotes. Adaptive evolution could thus be a selective force that universally promotes recombination.


# Significance statement

Whether 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 196 bacterial genome sequences of five closely-related *Rhizobium leguminosarum* species. 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. These results suggest that homologous recombination directly impacts the efficacy of natural selection in prokaryotes, as it has been shown previously to be 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 (1). This is because recombination decouples the fate of adaptive and deleterious variants, decreasing the amount of selective interference throughout the genome (2, 3). Selective interference—also termed as the Hill-Robertson (HR) effect—is expected to be strongest in regions of the genome where recombination is low (4). The HR effect is expected 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 (5)). Homologous recombination is expected to mitigate the HR effect and increase the percentage of amino acid substitutions that are due to adaptive evolution ($\alpha$).

Empirical evidence based on population genomics data supports these expectations with a positive correlation between recombination and $\alpha$ reported in diverse species of eukaryotes, including flies (*Drosophila melanogaster*, (6, 7)), fungi (*Zymoseptoria tritici (8)*), plants (*Arabidopsis thaliana* (9)), and non-model animal species (10, 11).

Whereas recombination is ubiquitous in eukaryotes, this is not the case for prokaryotes. Nevertheless, most studied prokaryotes show high rates of genetic exchange (12) and it is therefore of interest to explore whether such recombination also facilitates adaptive evolution in prokaryotes. Here we investigate adaptive evolution and recombination 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 (13). Our analyses showed that they cluster into five closely related species (2-5% divergent) with horizontal gene transfer only affecting the nitrogen fixation genes and a few well-defined genomic regions. 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 provide evidence that the rate of adaptive protein evolution increases with the recombination rate in this species complex.

# Results & Discussion

To estimate the proportion of adaptive evolution ($\alpha$) across this *Rhizobium* species complex, and study how $\alpha$ covaries with recombination rate estimates, we restricted analyses to polymorphism data from regions of the core genome without evidence of horizontal gene transfer (HGT).

Across all five species (196 strains, gsA:32, gsB:32, gsC:112, gsD:5, gsE:11), a total of 22115 orthologous genes were previously identified (13); of those, 4204 genes are present in all strains (core genes). Most core genes are found in the large chromosome (3304 genes), but some were located in the chromids (Rh01, Rh02) and in one of the plasmids (Rh03) (see (14), and (13)). The chromosome, chromids, and the plasmid are hereafter referred to as genomic compartments. We filtered out genes that showed evidence of interspecies HGT or unexpectedly high rates of nucleotide diversity (see Methods) (**Fig. S1**), leaving a total of 3086 genes (total alignment length: 3091179) and 334040 variable sites for analysis (**Fig. S2**).

First, we estimated nucleotide diversity, intragenic linkage disequilibrium (LD), and the site frequency spectrum (SFS) (see Methods) within each species (**Fig. 1a-c**). The average nucleotide diversity, $\pi$, an estimator of $2N_e\mu$ in haploids, is significantly different among genomic compartments (**Fig. 1a, Table S1**). Across the species, $\pi$ differs by up to a factor of 4.5 (gsA: 0.018, gsB:0.0045, gsC:0.0140, gsD:0.00512, 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 non-random association between pairs of alleles, quantified by $r^2$ (see 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 1000 base pairs for all species, suggesting substantial amounts of within-species homologous recombination (**Fig. 1b**). 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 ten informative sites within each species and evaluated their $r^2$ distributions separately (**Fig. 1c**). As expected, the species with the most striking LD decay (gsC) has the lowest $r^2$ (median $r^2$: 0.248), and the opposite is also true (gsD, median $r^2$: 0.7131).

Species gsB and gsE have similar LD decay and $r^2$distributions (**Fig. 1b-c**). We tested whether their average $r^2$ was statistically different. Based on the total number of shared genes (3086), the average $r^2$ of gsB was not significantly different from the average $r^2$ of gsE (Wilcox's test, p-value=0.10). All the other pairwise comparisons were statistically significant (**Table S2**). Thus, we considered gsB and gsE to have similar recombination rates. 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 $r^2$: 0.293) > gsB (median $r^2$: 0.48) and gsE (median $r^2$: 0.43) > gsD (median $r^2$: 0.71).

Next, we computed the folded site frequency spectrum (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. 1d**). The observed excess of intermediate frequency SNPs indicates 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 (**Fig. S3**). Differences among species suggest distinct demographic histories, with gsC showing an SFS compatible with population expansion and gsA with population decline.

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 $pi_N/pi_S$ are gsA = 0.039, gsB = 0.057, gsC = 0.037, gsD = 0.07, gsE = 0.051). 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 (15). Rare nonsynonymous variants are often deleterious ($s \sim 1/N_e$) (16, 17). Because deleterious variants contribute substantially to polymorphism but rarely to divergence (18, 19), their presence in the genomes, if not controlled for, will lead to an underestimation of $\alpha$ (20).

We used GRAPES (10) to estimate the distribution of fitness effects (DFE) (21) 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 S3**) makes it difficult to reliably call ancestral and derived states (22, 23). Accordingly, we chose to

estimate the DFE and $\alpha$ using the folded SFSs (10). To determine the model of the DFE that best fit our data, we used a variety of DFE distribution models (**Table S4**). 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. S4**). This model assumes the existence of weakly deleterious nonsynonymous mutations, modeled as a continuous Gamma distribution (10).

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 considered as "biological replicates" (10) (**Table 1**). Except for the comparison between gsA and gsB, in which differences between $\alpha_{gsA\ gsB}$ and $\alpha_{gsB\ gsA}$ exceeded 0.1, the overall discrepancy in the values estimated between mirror species does not exceed 0.1. Using each species' focal polymorphism data, we calculated four $\alpha$ 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 $\alpha$ across all outgroups used, while the least recombining species (gsD) had the lowest $\alpha$ in ¾ of the cases.

We then investigated whether intraspecies differences in recombination rate affect the amount of adaptive evolution ($\alpha$). For each species, we split genes into three recombination classes based on their average $r^2$ values and computed $\alpha$ for each class using the GammaZero model (**Fig. S5, Table S5**). Because we only kept genes with at least ten informative sites, the number of genes evaluated across species was different (**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 $r^2$). Except for cases in which we used gsD and gsB polymorphisms to estimate $\alpha$, all the other species pairwise comparisons led to at least one significant difference (based on non-overlapping CI's) between recombination classes. We further assessed the significance of the pattern reported here by permuting—200 times—across recombination classes (see Methods). Except for simulations in which gsD and gsB polymorphisms were used, all the other simulations led to significant differences (p-value <= 0.025) among the two most extreme classes of recombination (**Fig. S6**).

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

($\omega_{na}$): $\alpha = \omega_a / (\omega_a + \omega_{na})$. Thus, an increase in $\alpha$ with recombination could be due to either an increase in the rate of adaptive substitutions, a decrease in the rate of non-adaptive substitutions, or both. Figure 3 shows that $\omega_a$ increases with recombination rate whereas $\omega_{na}$ 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 non-adaptive substitution and increased rates of fixation of adaptive variation. This is exactly as predicted from selective interference theory (1, 3, 4).

To evaluate the robustness of these results, we computed an alternative measure of recombination (R/θ). We then made new recombination classes and evaluated its correlation with $\alpha$. R/θ measures the importance of recombination (R), relative to mutation (θ) across sequences (24, 25). Although the per gene estimates of R/θ are less variable than that of $r^2$ (**Fig. S7**), these two measures are not independent (Pearson correlation ranged from 0.21 to 0.44) (**Fig. S7**). For most species comparisons, the trend between $\alpha$ and recombination is still consistent: the higher R/θ, the higher $\alpha$ is (**Fig. S8**).

# Conclusion

We have found that five bacterial species within the species complex *R. leguminosarum* display different yet high levels of recombination. The estimates of $\alpha$ 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 earlier estimates from *S. enterica* (0.34, CI=0.14, 0.50) previously reported (19).

Levels of recombination correlate—both across and within species—with higher amounts of adaptive evolution measured either as the rate of adaptive substitutions ($\omega_a$) or as the proportion of amino acid changes which have been fixed by positive selection ($\alpha$). For instance, the most recombining species (gsC) consistently exhibited the largest $\alpha$, independent of the outgroup used. Within each species, we also find a positive correlation between intragenomic recombination rate and $\alpha$, as well as for $\omega_a$. These findings are robust to the measure of recombination ($r^2$and *R/θ*) and the choice of outgroup used for computing divergence.

The positive effect of homologous recombination on $\alpha$ we report here is in line with population genetic studies conducted in vertebrates (10, 11) and invertebrates (7, 8, 26–29). It points to recombination being a general facilitator of adaptive evolution across the tree of life—possibly being a selective force for the existence of recombination in prokaryotes in the first place.

# 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 (13). Briefly, the strains were previously subjected to whole-genome shotgun sequencing using 2×250bp Illumina paired-end reads (Illumina, USA). Genomes were assembled using SPAdes (v. 3.6.2, (30)) 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 (13), we predicted protein-coding sequences using prokka (31) (v1.12); this resulted in a total of 1468264 protein-coding sequences. To predict orthologous genes from these sequences, we used Proteinortho (31, 32) (v5.16b) with default parameters except for enabling the synteny flag. We identified a total of 22115 orthologous, including a total of 17911 orthologous observed in at least two strains (accessory genes), and 4204 orthologous found in all 196 strains (core genes).

Genes were then aligned using clustalo (33) (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 (13). Under the 95% ANI threshold that delineates species boundaries (34), we demonstrated that these 196 Rhizobium strains constitute five distinct *R. leguminosarum* species (gsA, gsB, gsC, gsD, and gsE). To ensure that we had a high-quality orthologous dataset for extracting segregating sites, we filtered it further (see below).

## Confident core orthologous genes

By developing and applying a phylogenetic method to quantify HGT (introgression score), we previously showed that most of the core genes shared among the present species respect the species-tree topology (introgression score = 0) (13). 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 little evidence of HGT, we imposed some restrictions. We only accepted genes that passed the following criteria: (i) were present in every strain (196 strains), (ii) with a nucleotide diversity ($\pi$) below 0.1, (iii) identifiable replicon origin (chromosome and chromids), (iiii) and with an introgression score <= 3 (**Fig. S1a**). A total of 3086 out of 4204 core genes were kept, and of these, 2550 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 single nucleotide polymorphisms (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, 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. Gaps were replaced by the site mean. The nucleotide diversity ($\pi$), gene length, and the distributions of segregating sites across core genes are described in **Fig. S1** (**b-d**).

## Transition transversion rate bias (kappa) and synonymous and nonsynonymous counts

Because transitions are more often synonymous at third codon positions than are transversions, to correctly identify the expected number of synonymous ($Lps$) and nonsynonymous counts ($Lpn$), we first estimated the average transition/transversion rate bias (kappa) (35) across species. To this end, we followed the methods described in (36) and used two classes of sites: fourfold-degenerate sites at the third codon positions and nondegenerate sites. Mutations at the fourfold-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, 9, 10, and 11 of Yang and Nielsen 2000 (36). These equations have been implemented within the CodonSeq class in Biopython (37) (private function _count_site_YN00() ), and these private functions were adapted to our dataset.

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 (**Fig. S2a**), 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) (35) and modified by Yang and Nielsen 2000 (36)—also implemented within Biopython. A total of 284742 synonymous, 49298 nonsynonymous sites were counted (**Fig. S2b-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 (**Table S3**). 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 Site Frequency Spectrum

One can infer the distribution of fitness effects from Site Frequency Spectrum (SFS) data (20). Because of the amount of shared polymorphism among the present species (**Table S3**), it becomes problematic to confidently distinguish ancestral from derived polymorphisms (38). Therefore, we chose to estimate the DFE using a method that uses the folded site frequency spectrums (SFS) of synonymous and nonsynonymous sites (10). To this end, we built the folded synonymous and nonsynonymous site frequency spectrums 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 ($\alpha$) 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.

## Calculating the significance levels between recombination classes

To test whether differences among recombination classes were statistically significant across species comparisons, we conducted a non-parametric test by shuffling genes among recombination classes (200 permutations) and recording the amplitude of differences between $\alpha$

estimates ($\Delta_\alpha = max_\alpha - min_\alpha$). We calculated a p-value by comparing the observed $\Delta_\alpha$ against the simulated $\Delta_\alpha$ distribution.

## Estimation of adaptive substitutions ($\alpha$)

Fitted parameters of the DFE were used to compute the expected $d_N/d_S$ under the different models, which was compared to the observed $d_N/d_S$ to estimate the adaptive substitution rate ($\omega_A$); non-adaptive substitution rate ($\omega_{NA}$), and the proportion of adaptive substitutions ($\alpha$) 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 (39). The different DFE models were compared using the Akaike Information Criterion (AIC) (40).

## Recombination rate estimates

To estimate the recombination rate per gene per species, we used two approaches: one based on the degree of association (or linkage disequilibrium) between pairs of alleles in a sample of haplotypes ($r^2$), and another, ClonalFrameML (R/θ) (24, 25), which relies on the maximum likelihood inference to detect recombination events that disrupt a clonal pattern of inheritance in bacterial genomes.

### (1) Intragenic linkage disequilibrium

Intragenic linkage disequilibrium (LD) measures the correlation between pairs of alleles with genomic distance. 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 $N$. The first step of the calculation was to apply a Z-score normalization to each SNP vector by subtracting the vector by its mean and dividing it by its standard deviation ($\frac{S_i - \overline{S_i}}{\sqrt{var(S_i)}}$), 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) ClonalFrameML

To estimate the changes in the clonal phylogeny by recombination (R), relative to mutation (θ) (R/θ), we used the software ClonalFrameML (24, 25). For each species, we first concatenated all core gene alignments (3086 genes) to build the starting phylogenetic species-tree using a maximum likelihood approach (Raxml-ng (41, 42)). We then input each phylogenetic tree within each gene alignment to estimate R/θ.

## Data sharing plans

Code generated for this study can be found at
https://github.com/izabelcavassim/Popgen_bacteria

## 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


1. N. H. Barton, The reduction in fixation probability caused by substitutions at linked loci. *Genetical Research* **64**, 199–208 (1994).

2. W. G. Hill, A. Robertson, The effect of linkage on limits to artificial selection. *Genetical Research* **8**, 269–294 (1966).

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

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

5. B. Charlesworth, A. J. Betancourt, V. B. Kaiser, I. Gordo, Genetic recombination and molecular evolution. *Cold Spring Harb. Symp. Quant. Biol.* **74**, 177–186 (2009).

6. D. Castellano, M. Coronado-Zamora, J. L. Campos, A. Barbadilla, A. Eyre-Walker, Adaptive Evolution Is Substantially Impeded by Hill-Robertson Interference in Drosophila. *Mol. Biol. Evol.* **33**, 442–455 (2016).

7. J. L. Campos, D. L. Halligan, P. R. Haddrill, B. Charlesworth, The relation between recombination rate and patterns of molecular evolution and variation in Drosophila melanogaster. *Mol. Biol. Evol.* **31**, 1010–1028 (2014).

8. J. Grandaubert, J. Y. Dutheil, E. H. Stukenbrock, The genomic determinants of adaptive evolution in a fungal pathogen. *Evol Lett* **3**, 299–312 (2019).

9. A. F. Moutinho, F. F. Trancoso, J. Y. Dutheil, The Impact of Protein Architecture on Adaptive Evolution. *Mol. Biol. Evol.* **36**, 2013–2028 (2019).

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

11. A. F. Moutinho, T. Bataillon, J. Y. Dutheil, Variation of the adaptive substitution rate between species and within genomes. *Evolutionary Ecology* **34**, 315–338 (2020).

12. X. Didelot, M. C. J. Maiden, Impact of recombination on bacterial evolution. *Trends Microbiol.* **18**, 315–322 (2010).

13. M. I. A. Cavassim, *et al.*, Symbiosis genes show a unique pattern of introgression and selection within a Rhizobium leguminosarum species complex. *Microbial Genomics* **6** (2020).

14. Harrison, *et al.*, Introducing the bacterial “chromid”: not a chromosome, not a plasmid. *Trends Microbiol.* **18**, 141–148 (2010).

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

16. A. L. Hughes, *et al.*, Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 15754–15757 (2003).

17. A. L. Hughes, Evidence for abundant slightly deleterious polymorphisms in bacterial populations. *Genetics* **169**, 533–538 (2005).

18. J. C. Fay, G. J. Wyckoff, C. I. Wu, Positive and negative selection on the human genome. *Genetics* **158**, 1227–1234 (2001).

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

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

21. A. Eyre-Walker, P. D. Keightley, The distribution of fitness effects of new mutations. *Nature Reviews Genetics* **8**, 610–618 (2007).

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

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

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

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

26. D. C. Presgraves, Recombination enhances protein adaptation in Drosophila melanogaster. *Curr. Biol.* **15**, 1651–1656 (2005).

27. A. J. Betancourt, J. J. Welch, B. Charlesworth, Reduced effectiveness of selection caused by a lack of recombination. *Curr. Biol.* **19**, 655–660 (2009).

28. J. R. Arguello, *et al.*, Recombination Yet Inefficient Selection along the Drosophila melanogaster Subgroup's Fourth Chromosome. *Molecular Biology and Evolution* **27**, 848–861 (2010).

29. T. F. C. Mackay, *et al.*, The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**, 173–178 (2012).

30. A. Bankevich, *et al.*, SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **19**, 455–477 (2012).

31. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

32. M. Lechner, *et al.*, Orthology detection combining clustering and synteny for very large datasets. *PLoS One* **9**, e105015 (2014).

33. F. Sievers, *et al.*, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

34. K. T. Konstantinidis, A. Ramette, J. M. Tiedje, The bacterial species definition in the genomic era. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 1929–1940 (2006).

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

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

37. P. J. A. Cock, *et al.*, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

38. R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol. Biol. Evol.* **24**, 1792–1800 (2007).

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

40. H. Akaike, Information Theory and an Extension of the Maximum Likelihood Principle. *Springer Series in Statistics*, 610–624 (1992).

41. K. Kobert, T. Flouri, A. Aberer, A. Stamatakis, The Divisible Load Balance Problem and Its Application to Phylogenetic Inference. *Lecture Notes in Computer Science*, 204–216 (2014).

42. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

# Figures and Tables

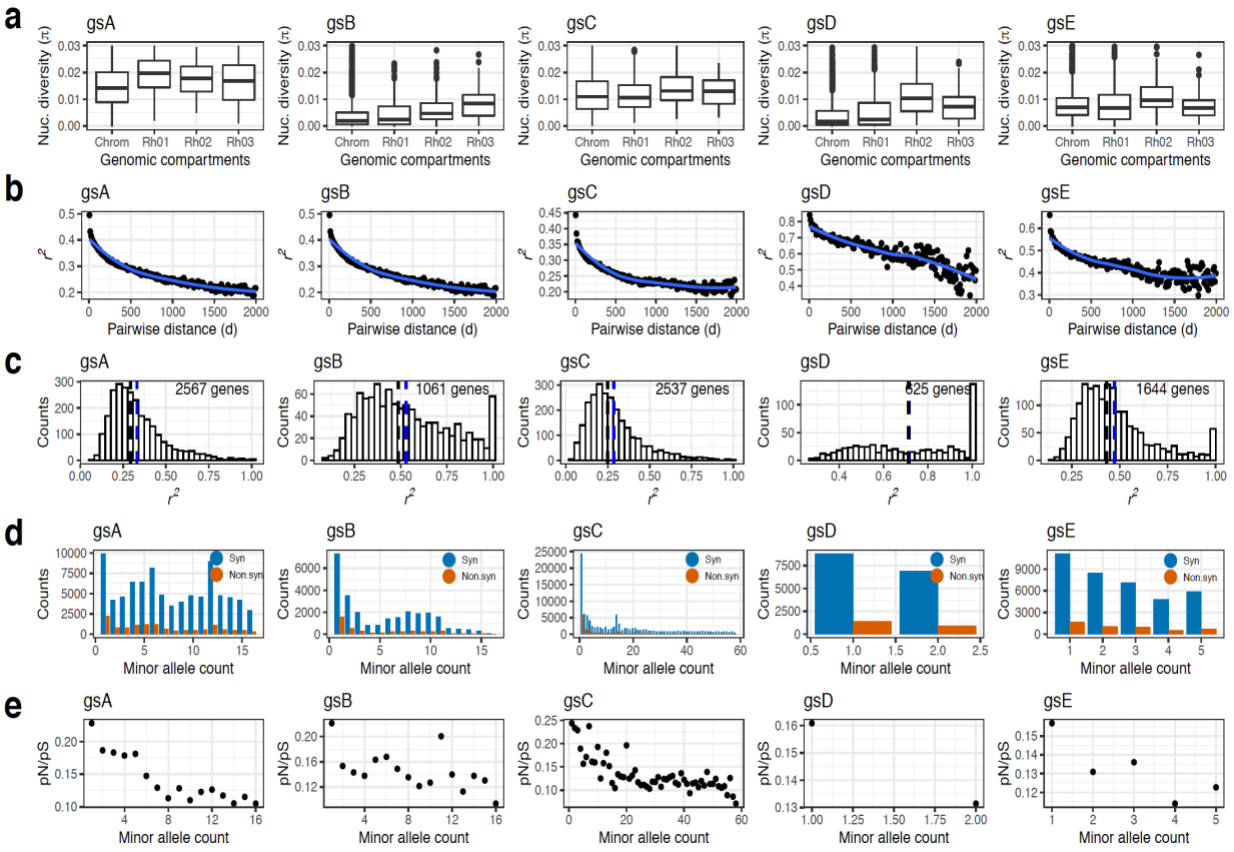


**Figure 1. Population genetics parameters across five species**. (**a**) Nucleotide diversity ($\pi$) across 3086 genes distributed along with genomic compartments (chromosome and chromids: Rh01, Rh02, and Rh03). To exclude outliers only genes with $\pi$ $<= 0.03$ are shown. (**b**) Intragenic linkage 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 (gsA: 2567 genes, gsB: 1061, gsC: 2537, gsD: 625, and gsE: 1644). 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.

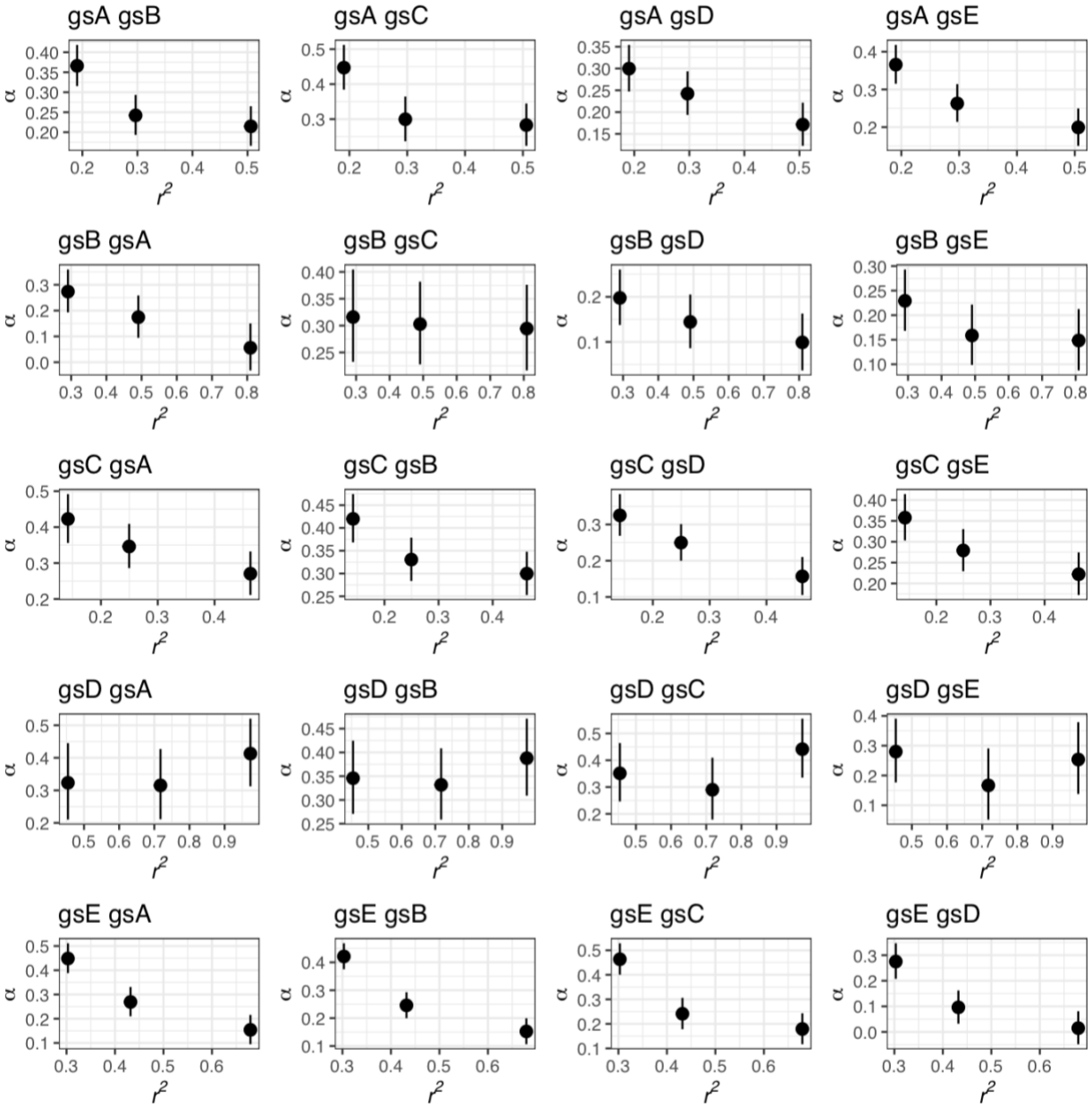


**Figure 2. The proportion of adaptive evolution ($\alpha$) by classes of recombination.** For each pairwise estimates of $\alpha$ the polymorphism data from one species (left in title) is compared against the divergence counts of an outgroup (right in title), and vice-versa. Results are divided into 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).

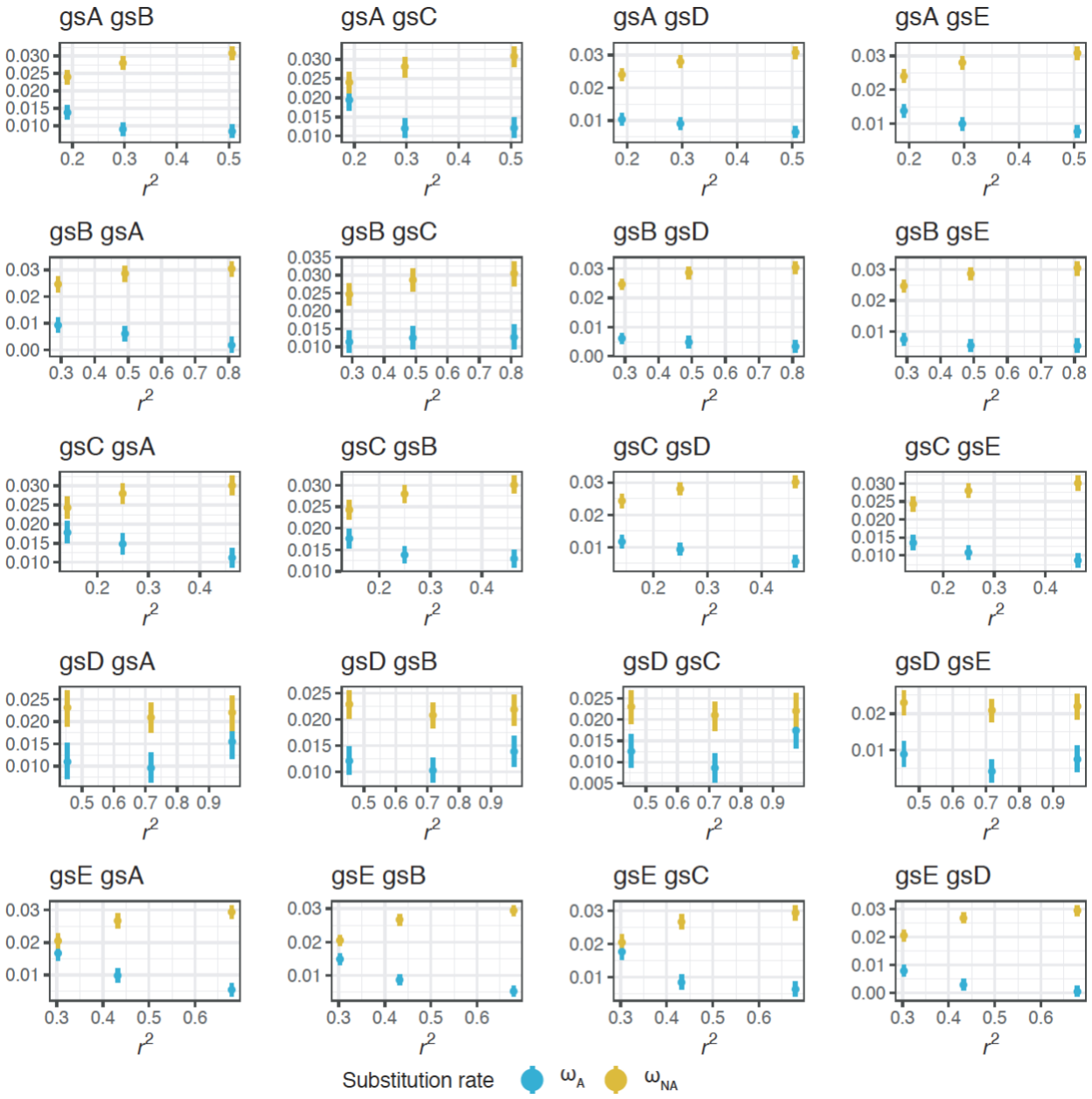


**Figure 3. The rates of adaptive ($\omega_A$) and non-adaptive ($\omega_{NA}$) evolution by classes of recombination.** For each pairwise estimates of $\omega_A$(in blue) and $\omega_{NA}$ (in yellow) the polymorphism data from one species (left in title) is compared against the divergence counts of 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 S4**). For each pairwise estimate of $\alpha$ ($\alpha_{species1\ species2}$), the polymorphism data from a focal species (rows) is compared against the divergence counts of an outgroup (columns), and vice-versa ($\alpha_{species2\ species1}$). Confidence intervals are displayed in brackets (grey) and numbers in parentheses represent the $\alpha$ ranking (in decreasing order) by outgroup (by column).

| Polymo-rphism (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) | - |