RESEARCH ARTICLE

Figures and figure supplements

Complex fitness landscape shapes variation in a hyperpolymorphic species

Anastasia V Stolyarova *et al*

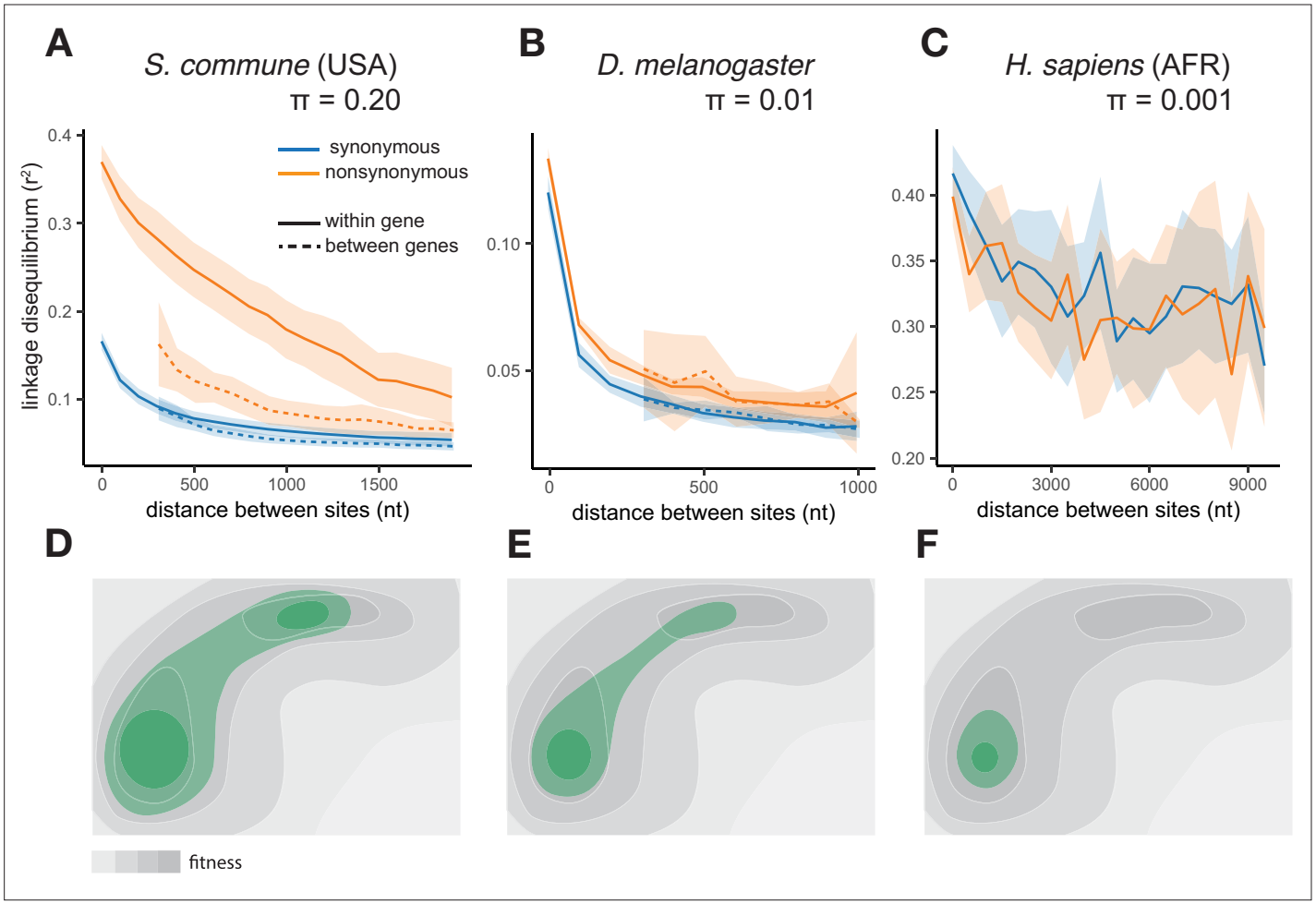


Figure 1. The efficiency of epistatic selection in populations with different levels of genetic diversity. (**A–C**) LD in natural populations for SNPs with MAF >0.05. (**A**) USA population of *S. commune*, (**B**) Zambian population of *D. melanogaster*, (**C**) African superpopulation of *H. sapiens*. Filled areas in (**A**)-(**C**) indicate SE of LD calculated for each chromosome or scaffold separately. (**D–F**) A hyperpolymorphic population (**D**) may occupy a sizeable chunk of a complex fitness landscape, leading to pervasive positive epistasis, while variation within less polymorphic populations (**E and F**) is confined to smaller, and approximately linear, portions of the landscape, so that no strong epistasis and LD can emerge. The area of the landscape covered by the population is shown in green.

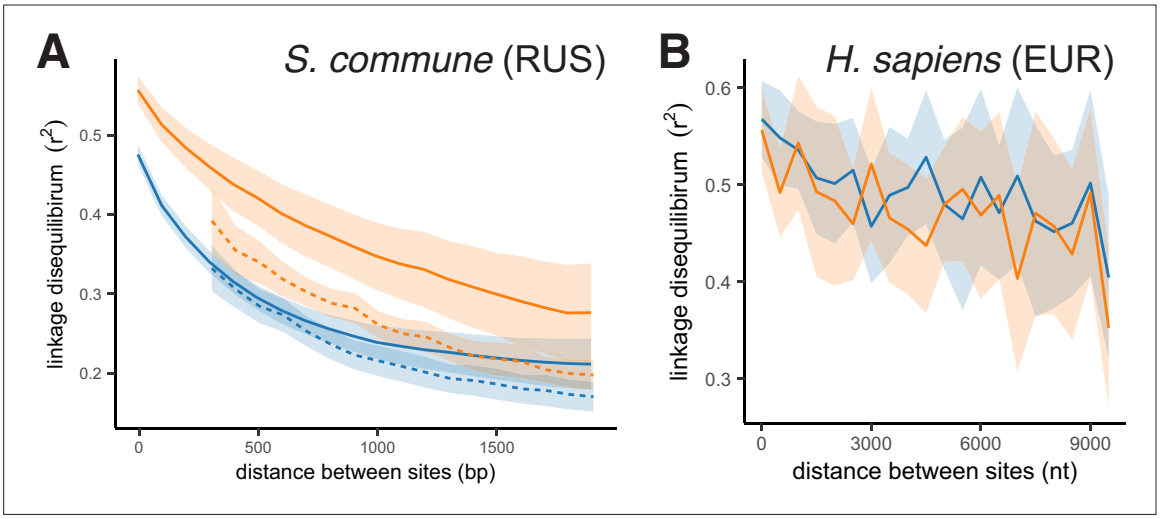


Figure 1—figure supplement 1. The efficiency of epistatic selection in populations with different levels of genetic diversity. LD between nonsynonymous SNPs is shown in orange, and LD between synonymous SNPs is shown in blue. (**A**) Russian population of *S. commune*, (**B**) European super-population of *H. sapiens*. Solid lines indicate LD between pairs of SNPs located within the same gene; dashed lines correspond to pairs of SNPs located in different genes. Only SNPs with minor allele frequency > 0.05 are analysed. Filled areas indicate SE of LD calculated for each chromosome (for human) or scaffold (for *S. commune*) separately.

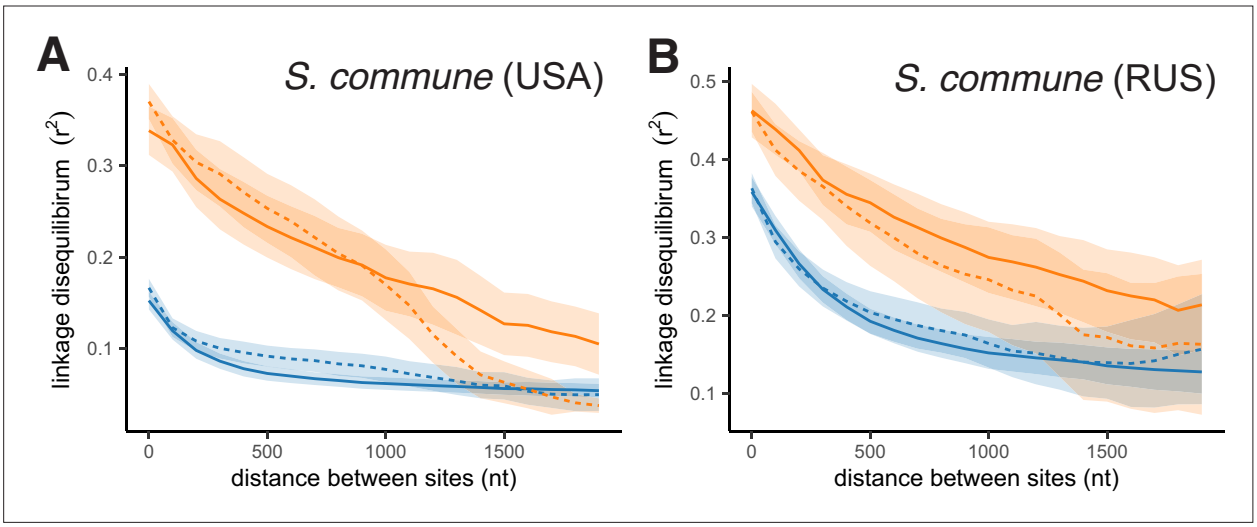


Figure 1—figure supplement 2. Linkage disequilibrium within and between exons in *S. commune*. LD between nonsynonymous SNPs is shown in orange, and LD between synonymous SNPs is shown in blue. Solid lines indicate LD between pairs of SNPs located within the same exon of the gene; dashed lines correspond to pairs of SNPs located in different exons of the gene. (**A**) USA population of *S. commune*, (**B**) RUS population of *S. commune*. Only SNPs with minor allele frequency > 0.05 are analysed. Filled areas indicate SE of LD calculated for each scaffold separately.

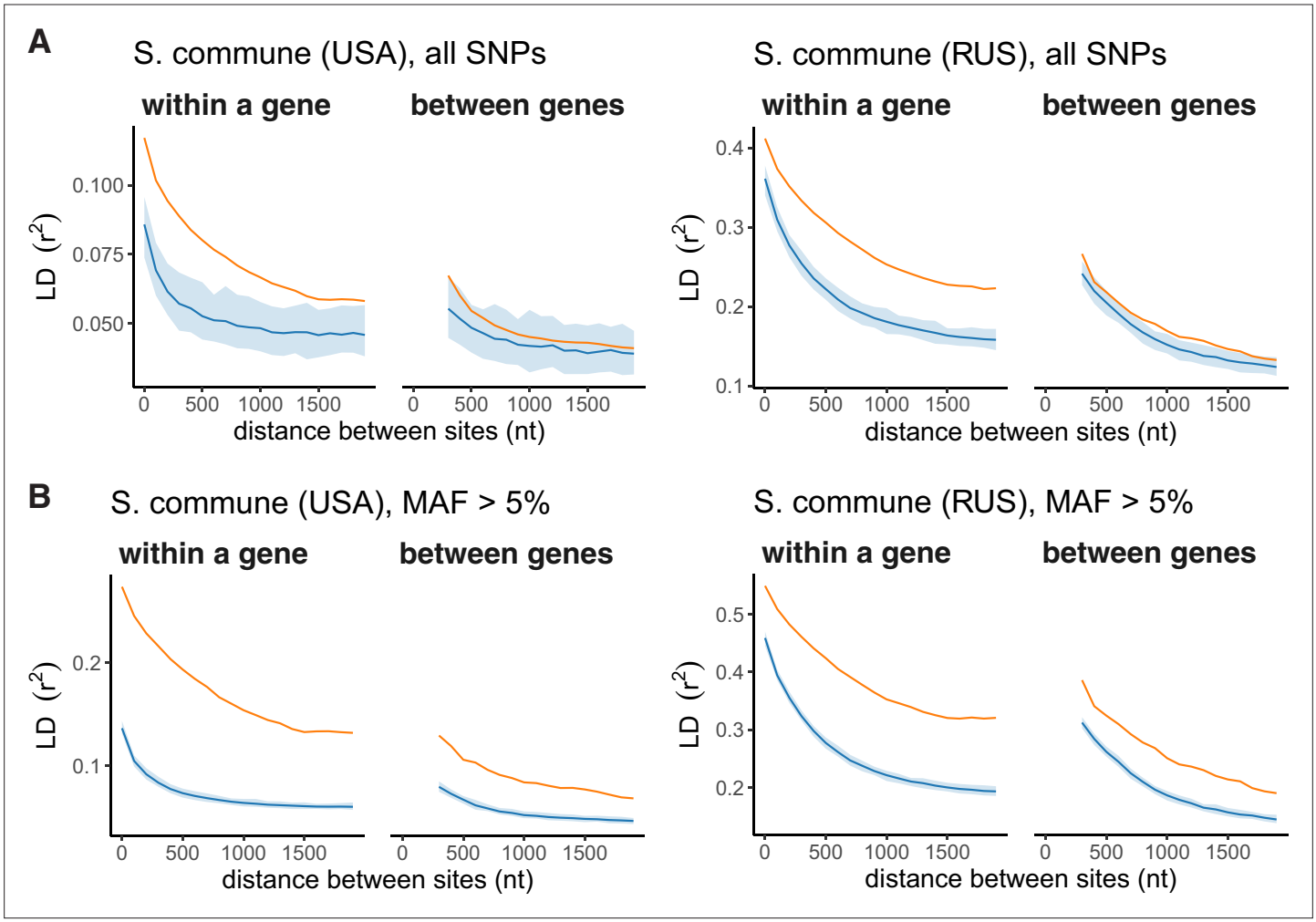


Figure 1—figure supplement 3. Comparison of LD_{nonsyn} and LD_{syn} in *S. commune* populations with exact matching of both MAFs and distance. For each possible minor allele count and nucleotide distance, the number of corresponding pairs of nonsynonymous variants and LD_{nonsyn} between them is calculated. Then, the same number of synonymous variants on the same nucleotide distance and with the same minor allele count is randomly chosen to calculate LD_{syn}. Subsampling is performed for 100 times. Filled areas show 95% intervals of LD_{syn} in the subsamples. (**A**) All SNPs, (**B**) SNPs with MAF >0.05.

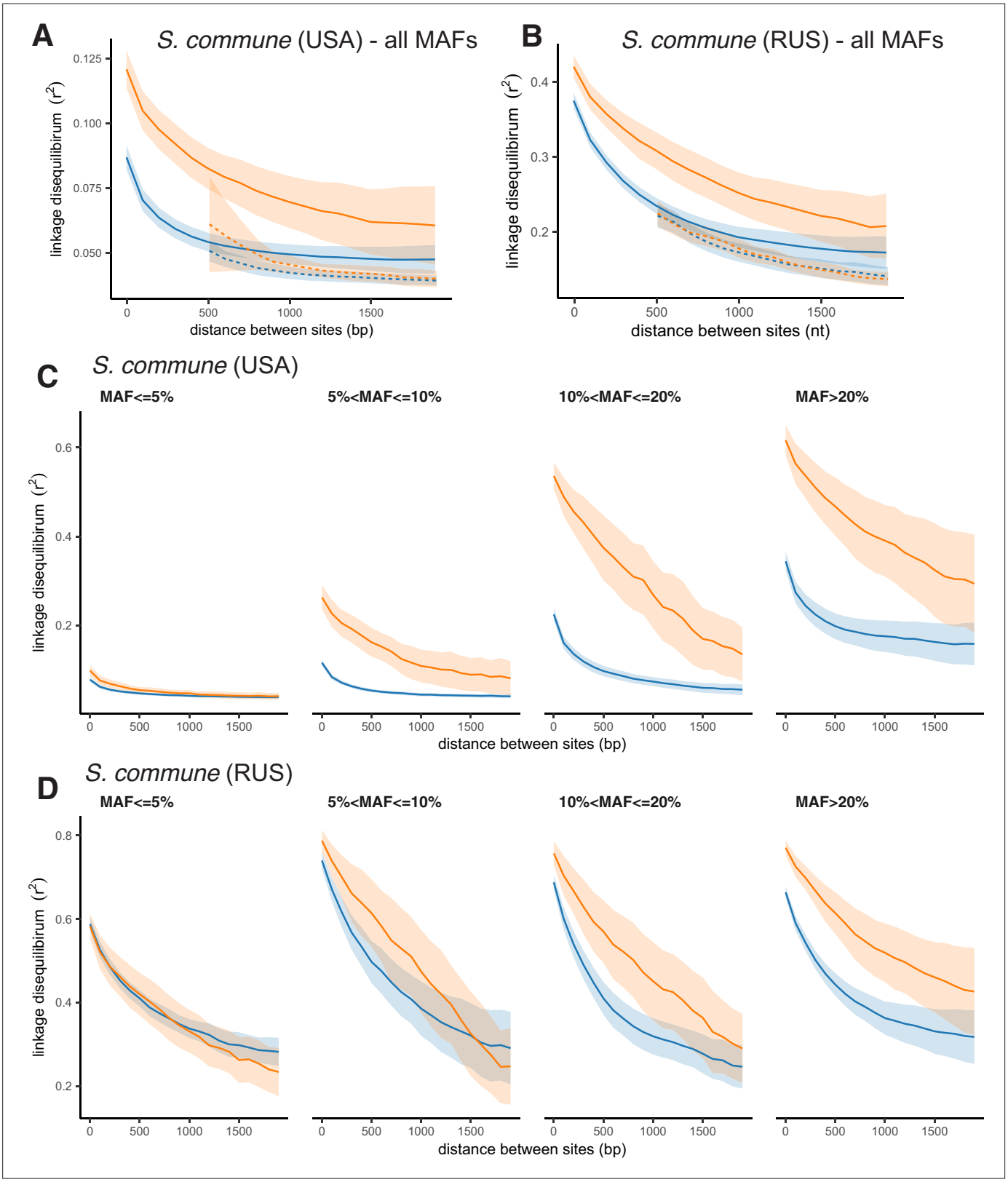


Figure 1—figure supplement 4. LD between SNPs with different MAF in *S. commune*. LD between nonsynonymous SNPs is shown in orange, and LD between synonymous SNPs is shown in blue. Filled areas indicate SE of LD calculated for each scaffold separately. (**A**, **B**) LD between all pairs of SNPs pooled together. Solid lines indicate LD between pairs of SNPs located within the same gene; dashed lines correspond to pairs of SNPs located in different genes. (**C**, **D**) Pairs of SNPs split by MAF.

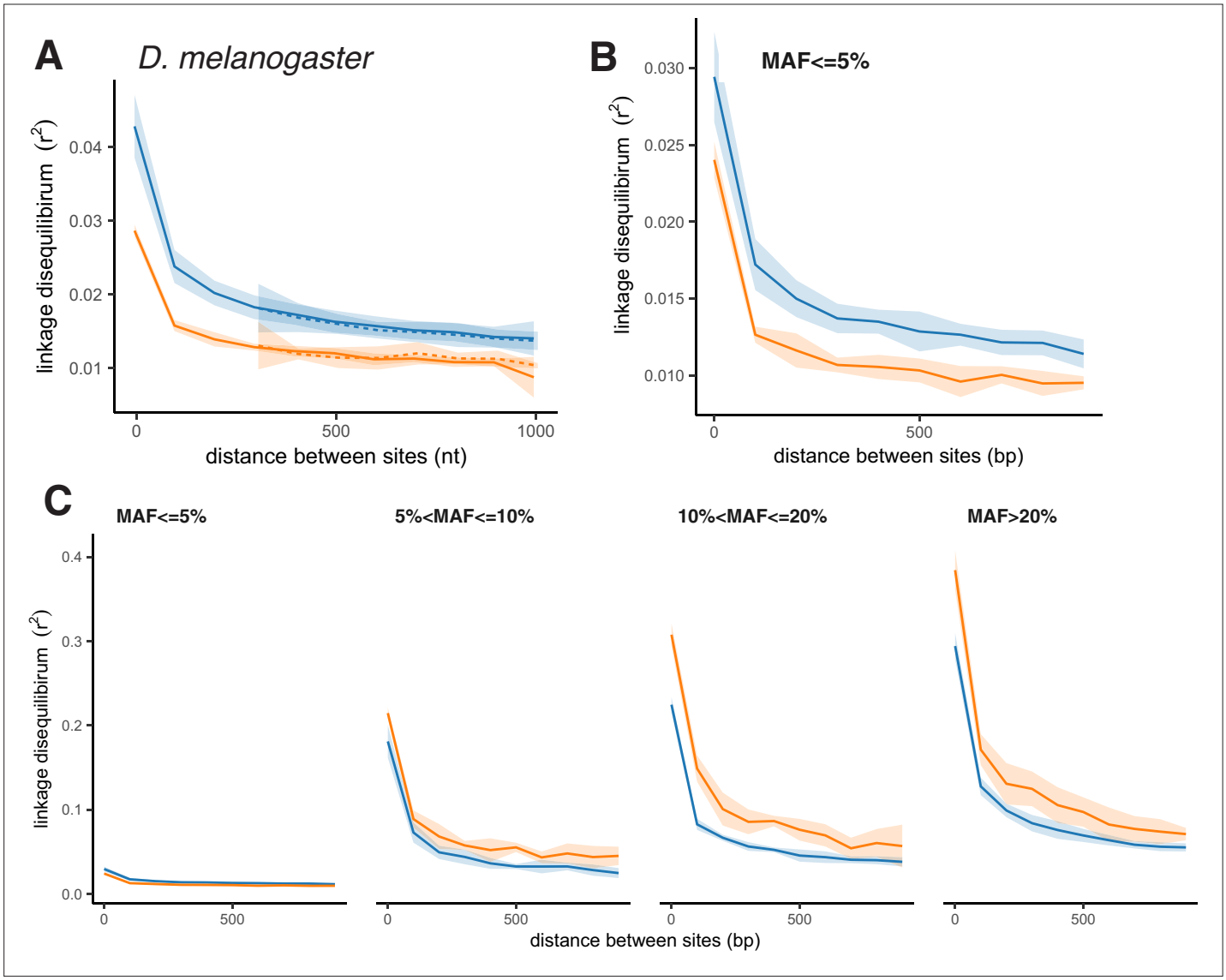


Figure 1—figure supplement 5. LD between SNPs with different MAF in *D. melanogaster*. LD between nonsynonymous SNPs is shown in orange, and LD between synonymous SNPs is shown in blue. Filled areas indicate SE of LD calculated for each chromosome separately. (**A**) LD between all pairs of SNPs pooled together. Solid lines indicate LD between pairs of SNPs located within the same gene; dashed lines correspond to pairs of SNPs located in different genes. (**B**) Pairs of SNPs with MAF <0.05 (large scale). (**C**) Pairs of SNPs split by MAF.

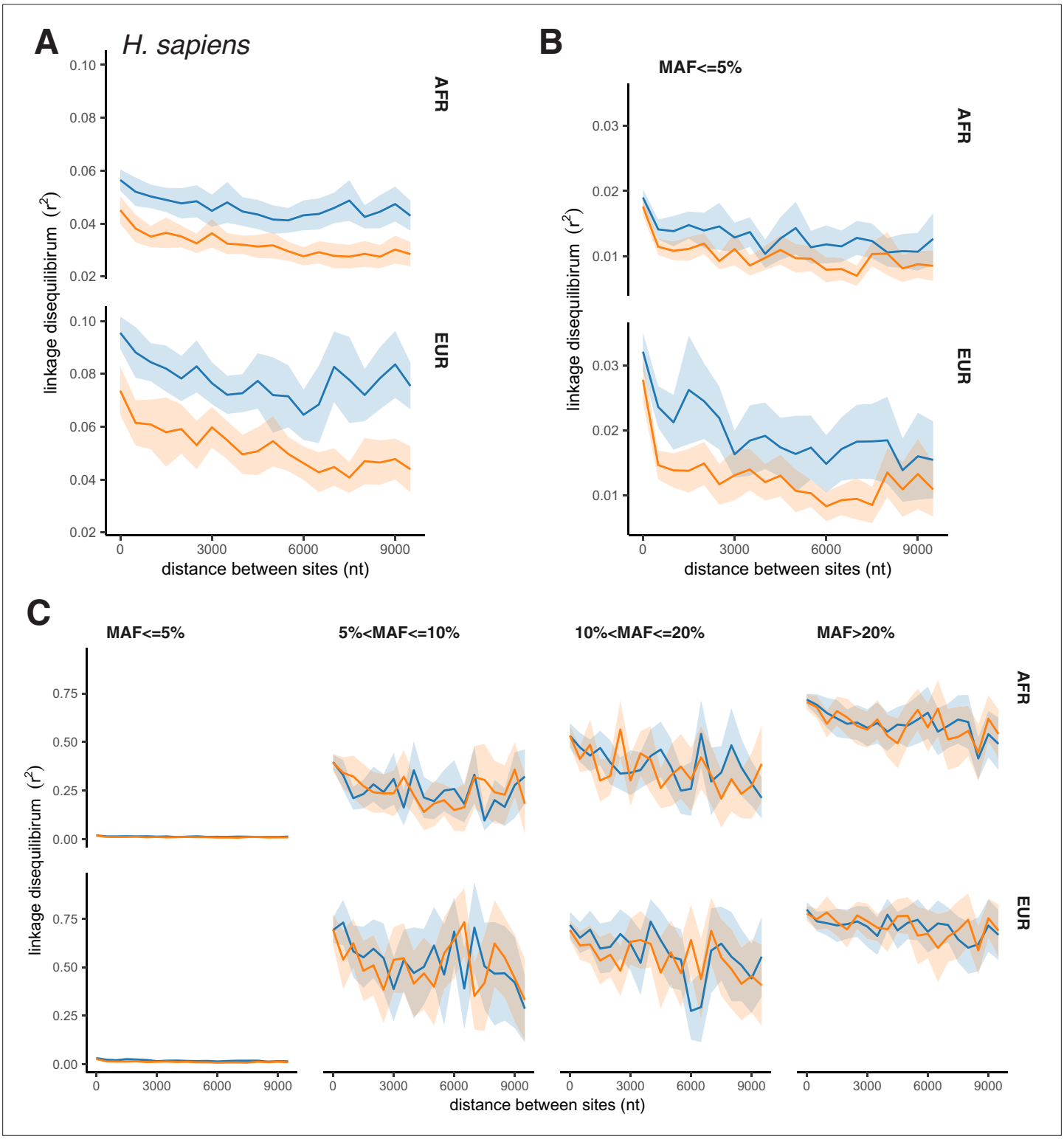


Figure 1—figure supplement 6. LD between SNPs with different MAF in *H. sapiens*. LD between nonsynonymous SNPs is shown in orange, and LD between synonymous SNPs is shown in blue. Filled areas indicate SE of LD calculated for each chromosome separately. (**A**) LD between all pairs of SNPs pooled together. Solid lines indicate LD between pairs of SNPs located within the same gene; dashed lines correspond to pairs of SNPs located in different genes. (**B**) Pairs of SNPs with MAF <0.05 (large scale). (**C**) Pairs of SNPs split by MAF.

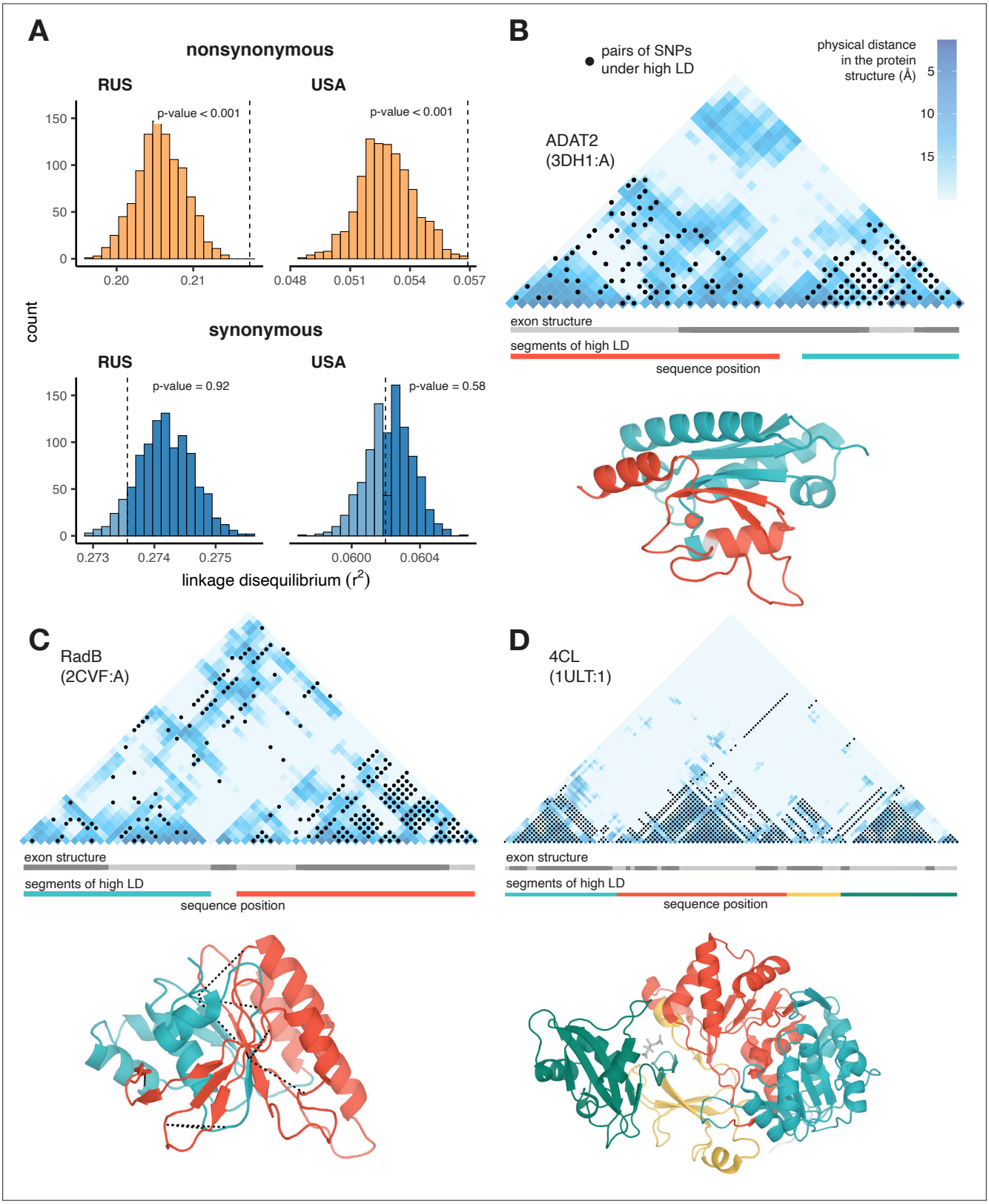


Figure 2. Excessive LD between physically interacting protein sites. (**A**) Within pairs of SNPs that correspond to pairs of amino acids that are colocalized within 10 Å in the protein structure, the LD is elevated between nonsynonymous, but not between synonymous, variants. Dashed lines show the average LD between colocalized sites. Permutations were performed by randomly sampling pairs of non-interacting SNPs while controlling for genetic distance between them, measured in amino acids; pairs of SNPs closer than 5 aa were excluded. (**B–D**) Examples of proteins with LD patterns matching their

Figure 2 continued on next page

Figure 2 continued

three-dimensional structures. Heatmaps show the physical distance between pairs of sites in the protein structure; only positions carrying biallelic SNPs are shown. Black dots correspond to pairs of sites with high LD (>0.9 quantile for the gene). Dashed lines in (**c**) structure show high LD between physically close SNPs from different segments of high LD. In these examples, LD is calculated in the Russian population of *S. commune*.

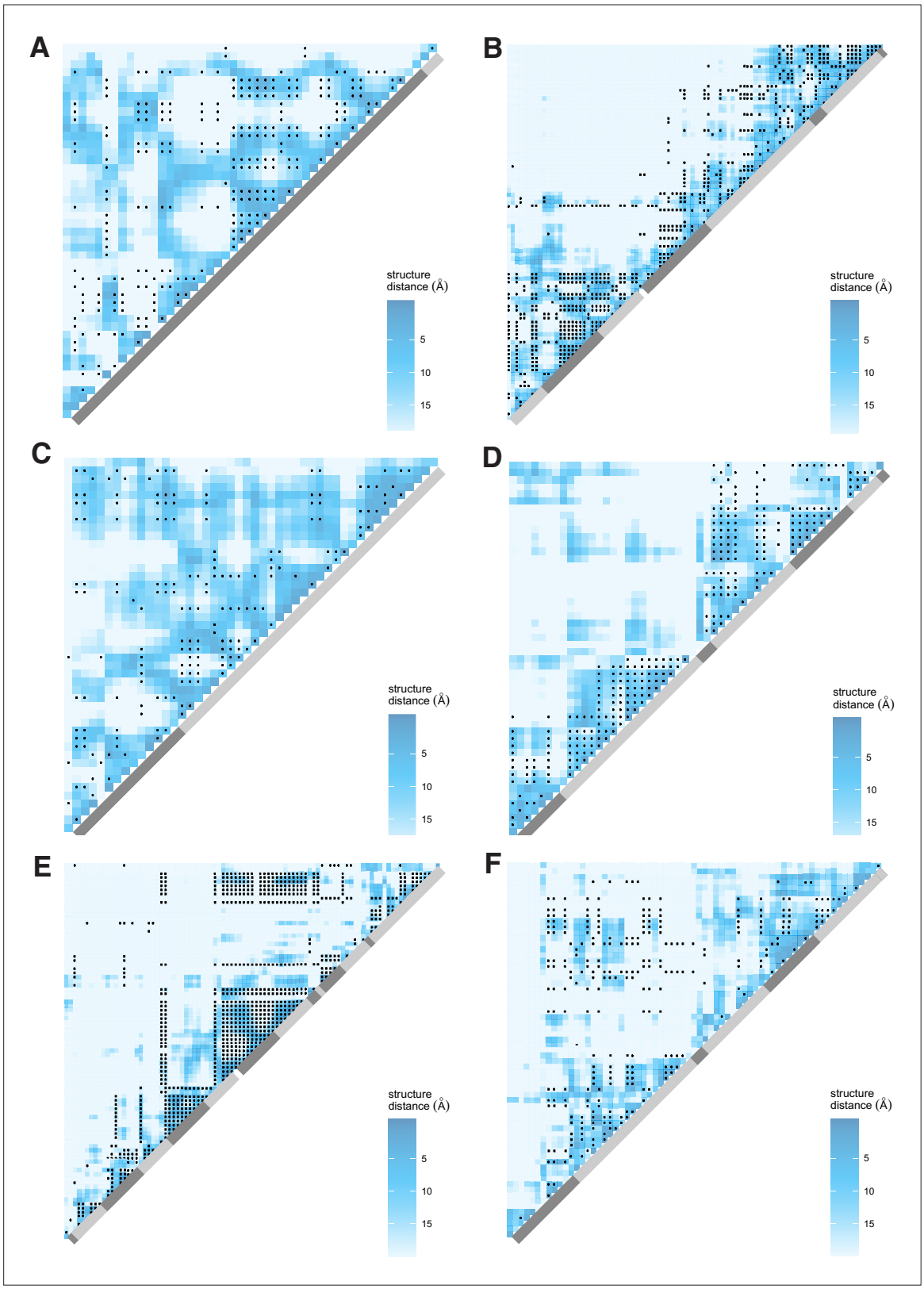


Figure 2—figure supplement 1. Examples of proteins with LD patterns matching the three-dimensional structure in the RUS population of *S. commune*. Heatmaps show the physical distance between pairs of sites in the protein structure; only positions carrying biallelic SNPs are shown. Black dots correspond to pairs of sites with high LD (>0.9 quantile for the gene). Grey regions indicate the exon structure of the genes. (**A**) cog1523 (5Y1B:A); (**B**) cog2779 (1SXJ:B); (**C**) cog5375 (1RGI:G); (**D**), cog5725 (1TA3:B); (**E**) cog18092 (4QJY:A); (**F**) cog7878 (4TYW:A). LD statistics and p-values for each gene are listed in ***Appendix 3—table 1***.

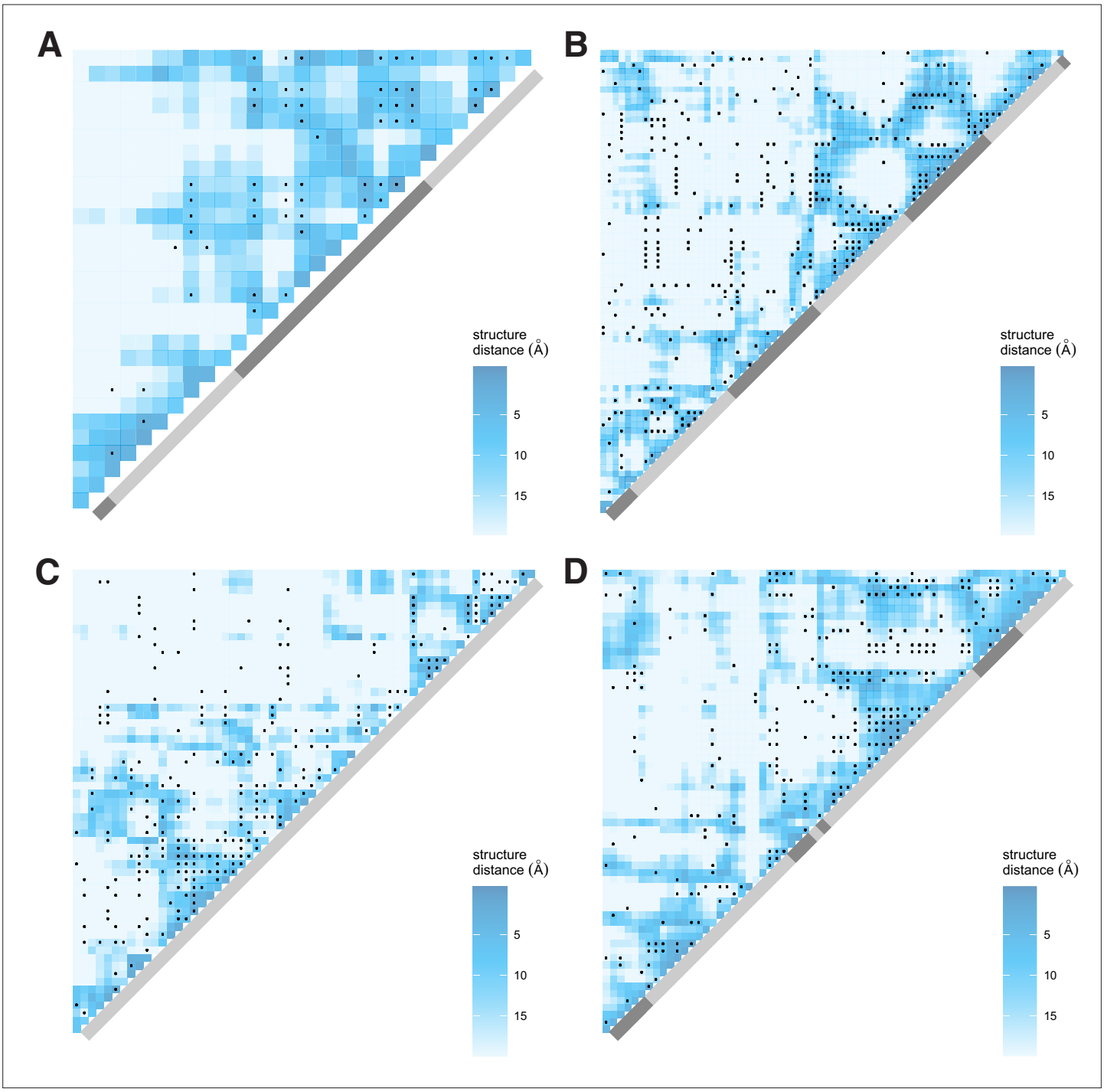


Figure 2—figure supplement 2. Examples of proteins with LD patterns matching the three-dimensional structure in the USA population of *S. commune*. Heatmaps show the physical distance between pairs of sites in the protein structure; only positions carrying biallelic SNPs are shown. Black dots correspond to pairs of sites with high LD (>0.9 quantile for the gene). Grey regions indicate the exon structure of the genes. (**A**) cog1536 (6AHR:E); (**B**) cog5725 (1TA3:B); (**C**) cog8253 (6F87:A); (**D**) cog9241 (1YCD:A). LD statistics and p-values for each gene are listed in ***Appendix 3—table 1***.

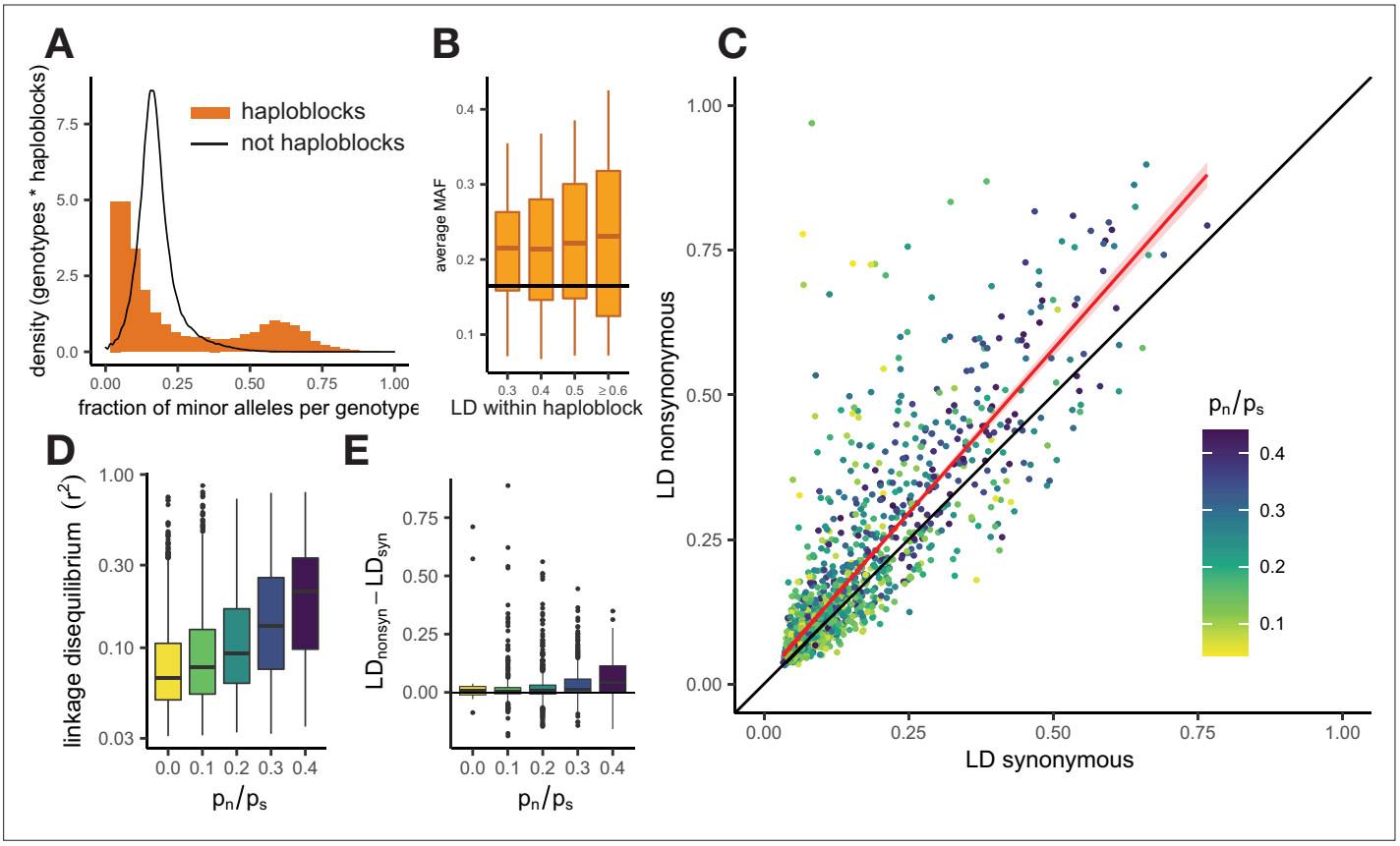


Figure 3. Patterns of linkage disequilibrium in the USA population of *S. commune*. (**A**) Distribution of the fraction of polymorphic sites that carry minor alleles in a genotype within haploblocks. Black line shows the distribution of fraction of minor alleles in genotypes in non-haploblock regions. (**B**) Distributions of the average MAF within a haploblock for haploblocks with different average values of LD. The average MAF in non-haploblock regions is shown as a horizontal black line for comparison. (**C**) LD between nonsynonymous and synonymous SNPs within individual genes. Linear regression of LD_{nonsyn} on LD_{nsyn} is shown as the red line. To control for the gene length, only SNPs within 300 nucleotides from each other were analyzed. Genes with fewer than 100 such pairs of SNPs were excluded. (**D,E**) The positive correlation between p_n/p_s of the gene and its average LD (**D**) or the difference between LD_{nonsyn} and LD_{syn} (**E**). Here, the data on the USA population of *S. commune* are shown; similar patterns in the Russian population are shown in ***Figure 3—figure supplement 4***.

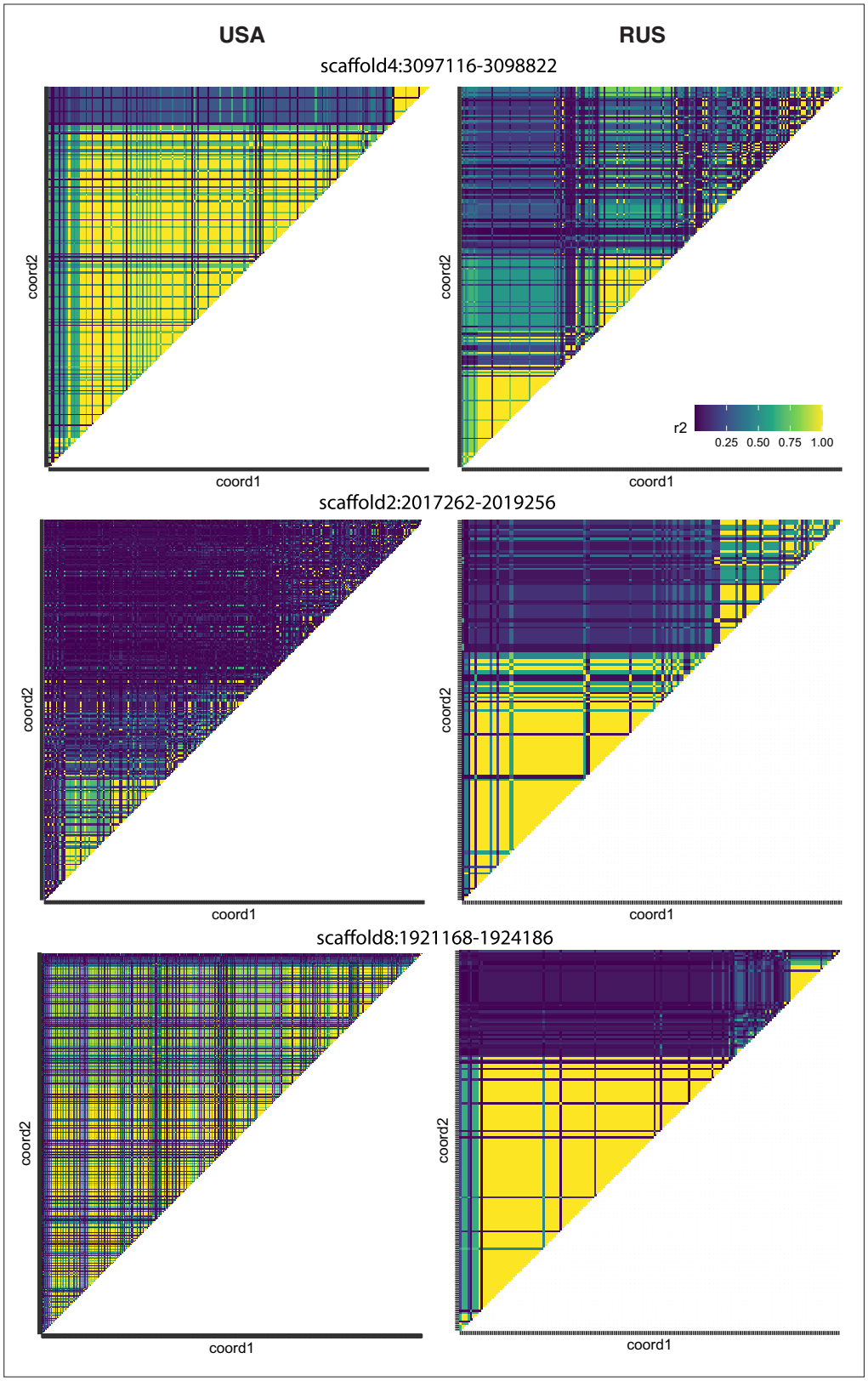


Figure 3—figure supplement 1. Examples of haploblocks in two populations of *S. commune*. The heatmaps show LD between polymorphic SNPs in the same genomic regions in the USA and RUS populations of *S. commune*. Only biallelic polymorphic sites with minor allele frequency >1 are shown, the number of such sites can differ between populations.

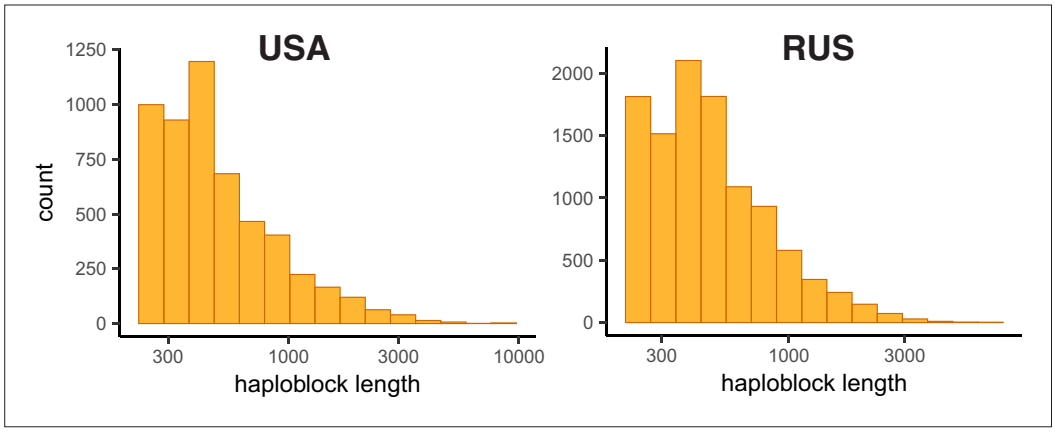


Figure 3—figure supplement 2. Distribution of haploblock lengths (nt) in the two populations of *S. commune*.

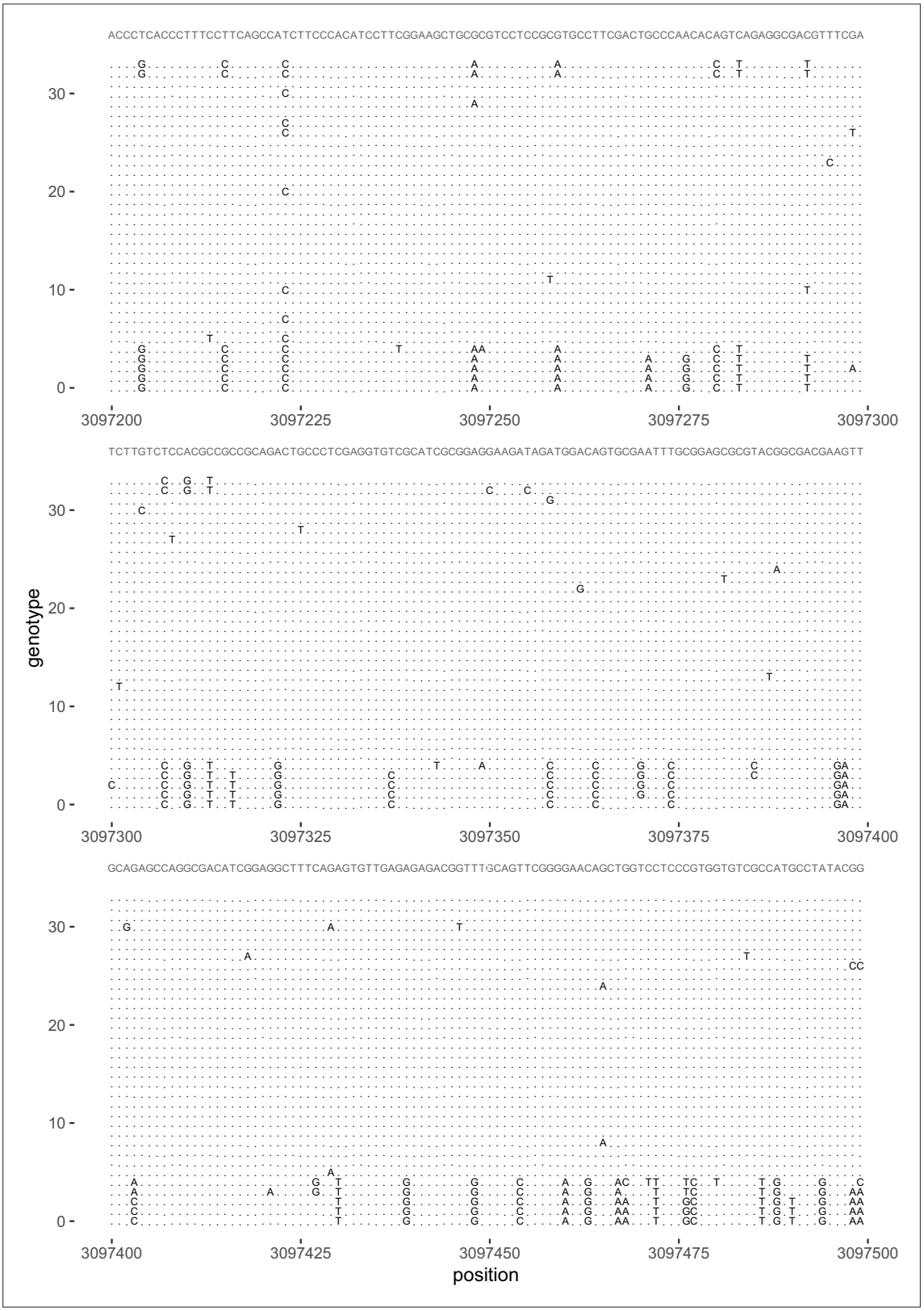


Figure 3—figure supplement 3. Example of the *S. commune* alignment within a haploblock. Region 3097200–3097500 of scaffold 4 in the USA population of *S. commune* is shown. The top line shows the consensus sequence based on 34 genotypes; dot indicates match with the consensus.

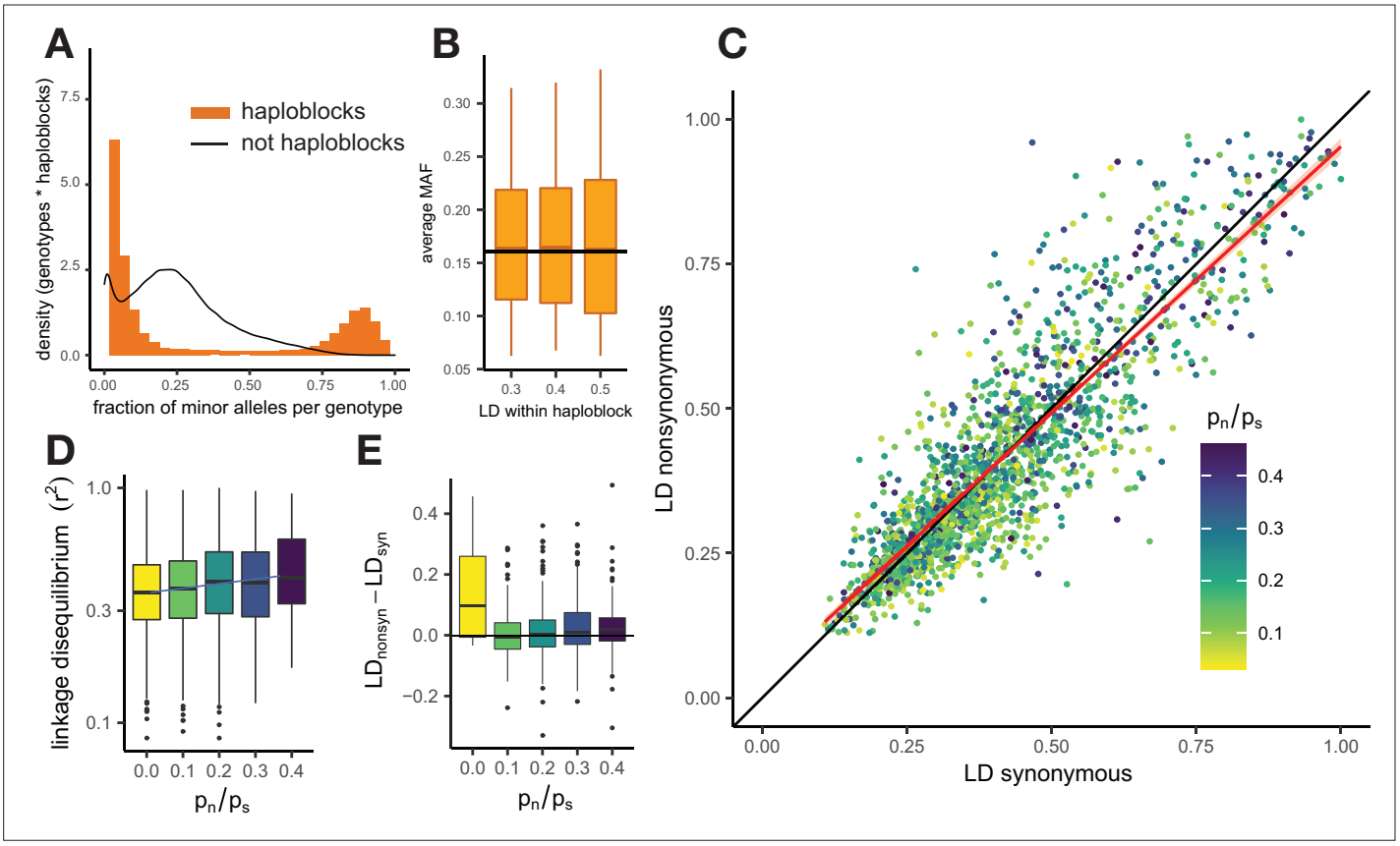


Figure 3—figure supplement 4. Patterns of linkage disequilibrium in the RUS population of *S. commune*. (**A**) Bimodal distribution of the fraction of polymorphic sites carrying minor alleles per genome within the haploblocks. Each count corresponds to a genotype within a haploblock. Black line shows the background distribution of minor alleles in the non-haploblock regions. (**B**) The increased average minor allele frequency within haploblocks as compared to the non-haploblock regions (dashed line, t-test p-value <2e-16). (**C**) LD between nonsynonymous and synonymous SNPs within single genes. Each dot represents an individual gene. Linear regression of LD_{nonsyn} over LD_{nsyn} is shown as the red line. To control for the gene length, only SNPs within 300 bp from each other were analyzed. Genes with fewer than 100 such pairs of SNPs were excluded. (**D,E**) The positive correlation between p_n/p_s of the gene and its average LD (Spearman correlation p-value = 4e-16) (**D**) or the difference between LD_{nonsyn} and LD_{syn} (Spearman correlation p-value = 2e-5) (**E**).

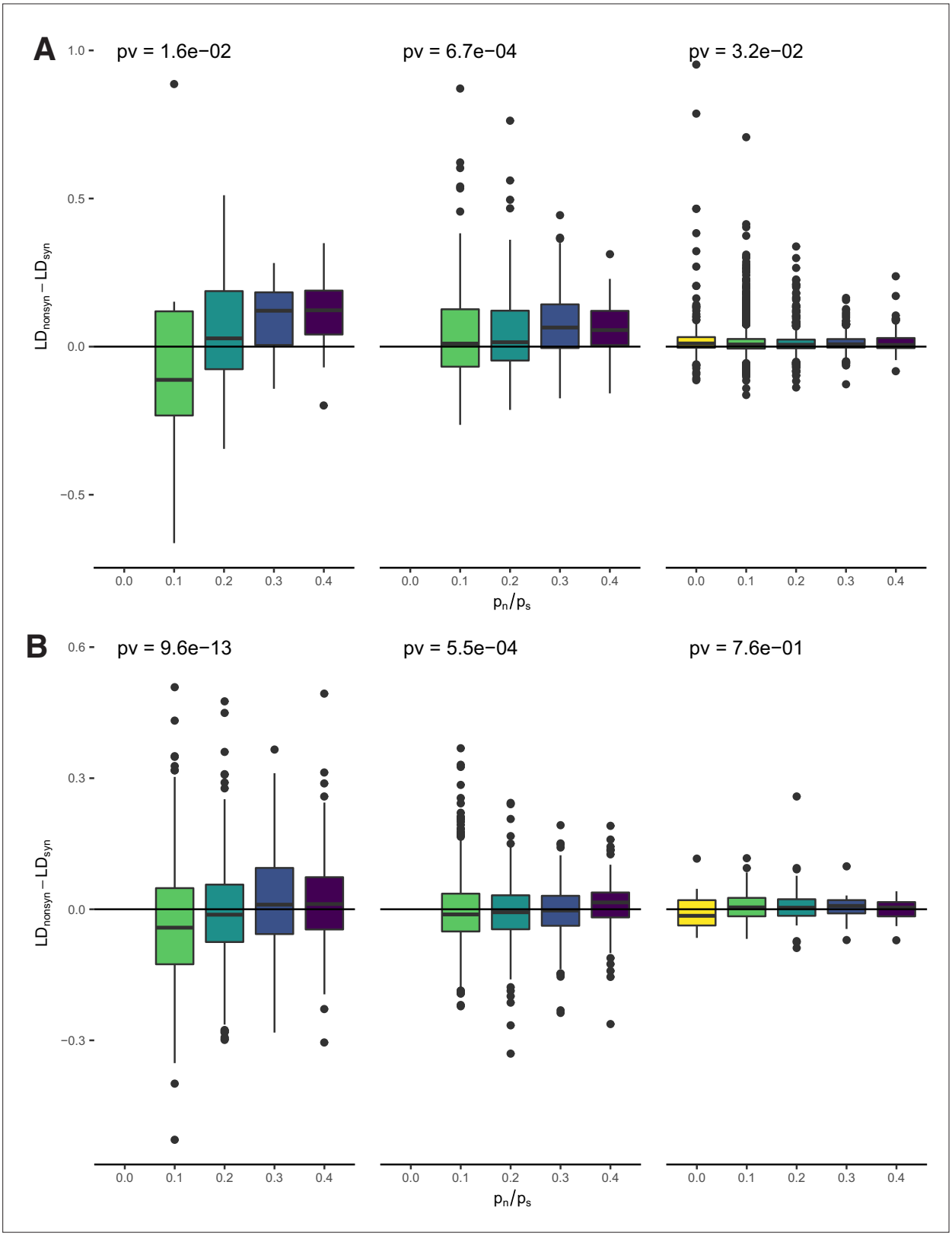
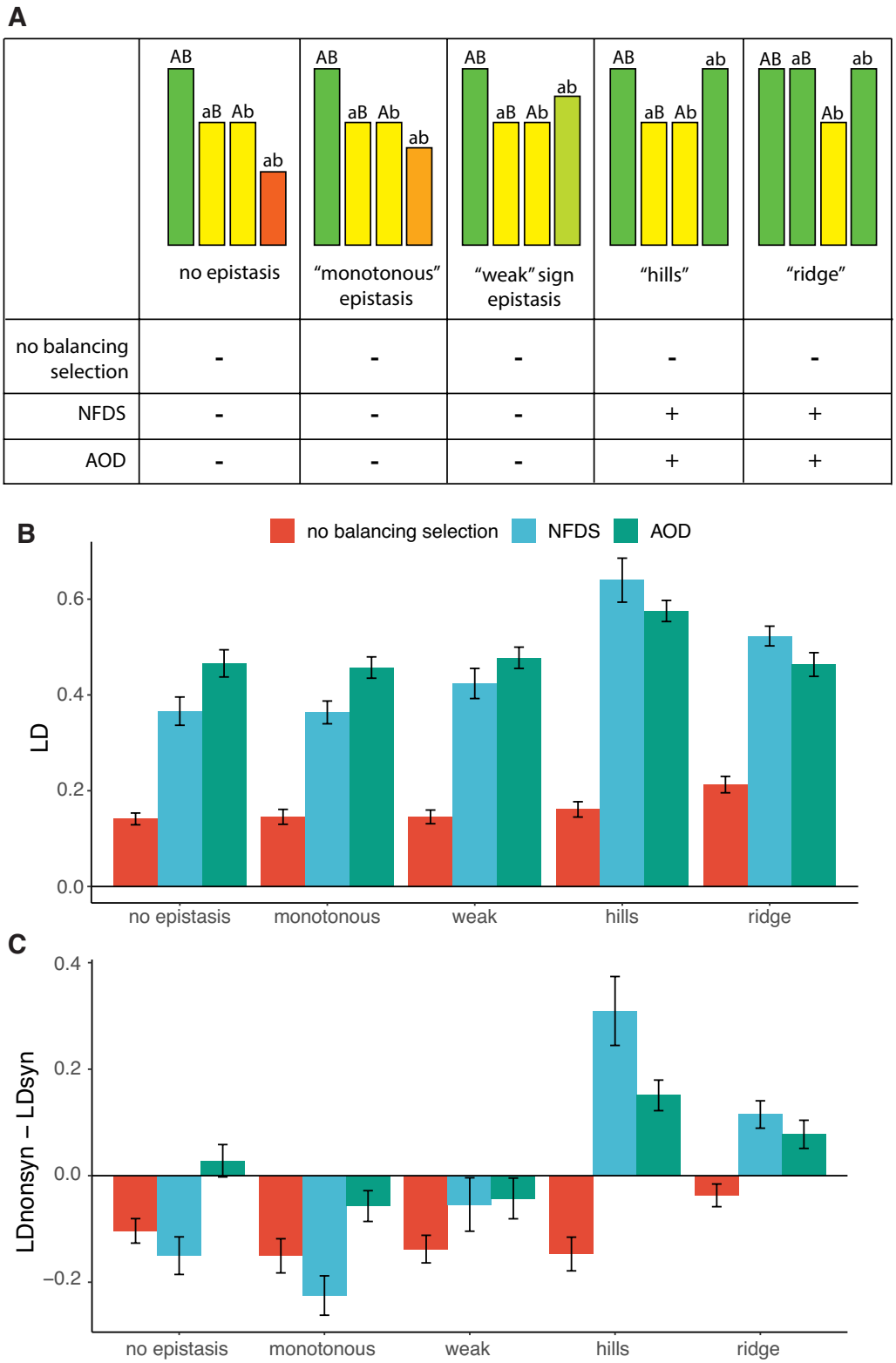


Figure 3—figure supplement 5. Comparison of LD_{nonsyn} and LD_{syn} in the genes of *S. commune*. (**A**) The USA population, (**B**) the RUS population. The genes are stratified by their average LD (the panels) and by the p_n/p_s. Only pairs of SNPs within 300 bp from each other are analyzed; genes with less than 100 such pairs of nonsynonymous or synonymous SNPs are excluded. Spearman correlation p-values are shown.



	no epistasis	"monotonous" epistasis	"weak" sign epistasis	"hills"	"ridge"
no balancing selection	-	-	-	-	-
NFDS	-	-	-	+	+
AOD	-	-	-	+	+

Figure 3—figure supplement 6. The difference between LD_{nonsyn} and LD_{syn} under pairwise epistasis and balancing selection. (**A**) The excess of LD_{nonsyn} over LD_{syn} under different models of epistasis between two deleterious mutations A → a and B → b without balancing selection and in the presence of negative frequency-dependent selection (NFDS) or associate overdominance (AOD) acting in the linked sites. The height of columns shows

Figure 3—figure supplement 6 continued on next page

Figure 3—figure supplement 6 continued

fitness of the corresponding genotypes. (+) indicate simulations where the excess of LD_{nonsyn} is reproduced. (**B**) The average LD in the simulations. (**C**) The difference between LD_{nonsyn} and LD_{syn} in the simulations.

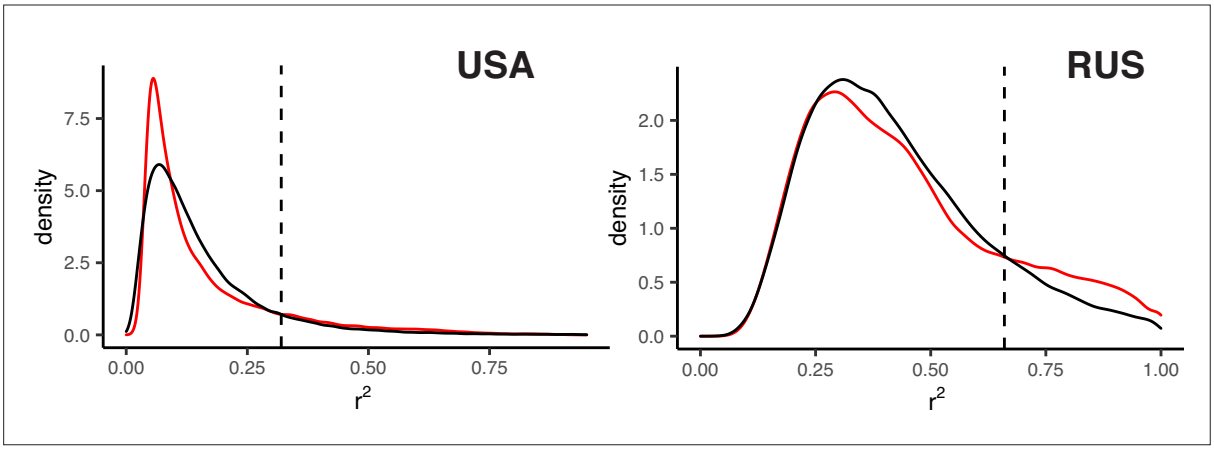


Figure 3—figure supplement 7. Criteria for haploblocks in *S. commune*. Red lines show the distribution of LD (r^2) in windows of 250 nucleotides in two populations. Black line corresponds to the lognormal distribution with the same mean and variance. The windows with LD higher than the threshold value defined as the intersection point of the two lines (dashed) are attributed to haploblocks.

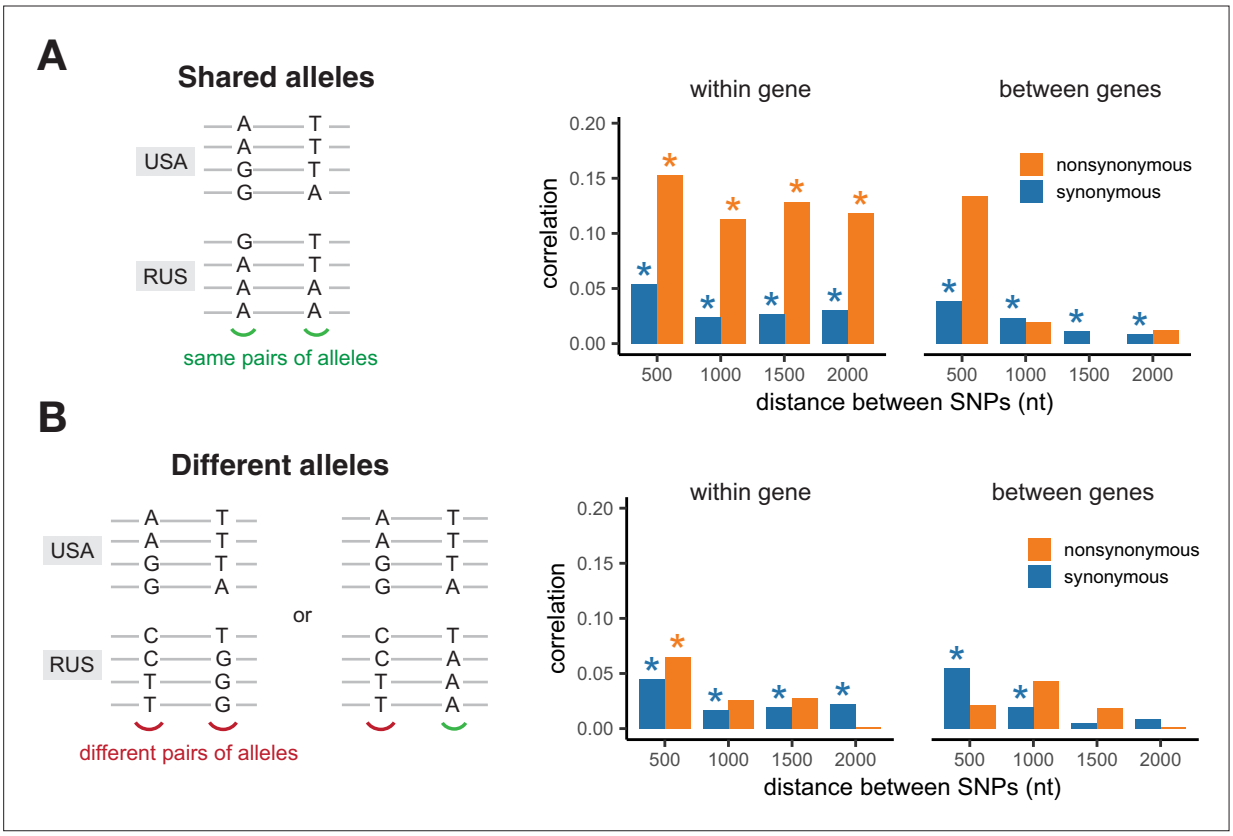


Figure 4. Correlation of LD values between pairs of shared SNPs in the two *S. commune* populations. (**A**) Pairs of SNPs with the same alleles in both sites, (**B**) pairs of SNPs differing by at least one allele. Asterisks indicate Spearman correlation p-values <0.001.

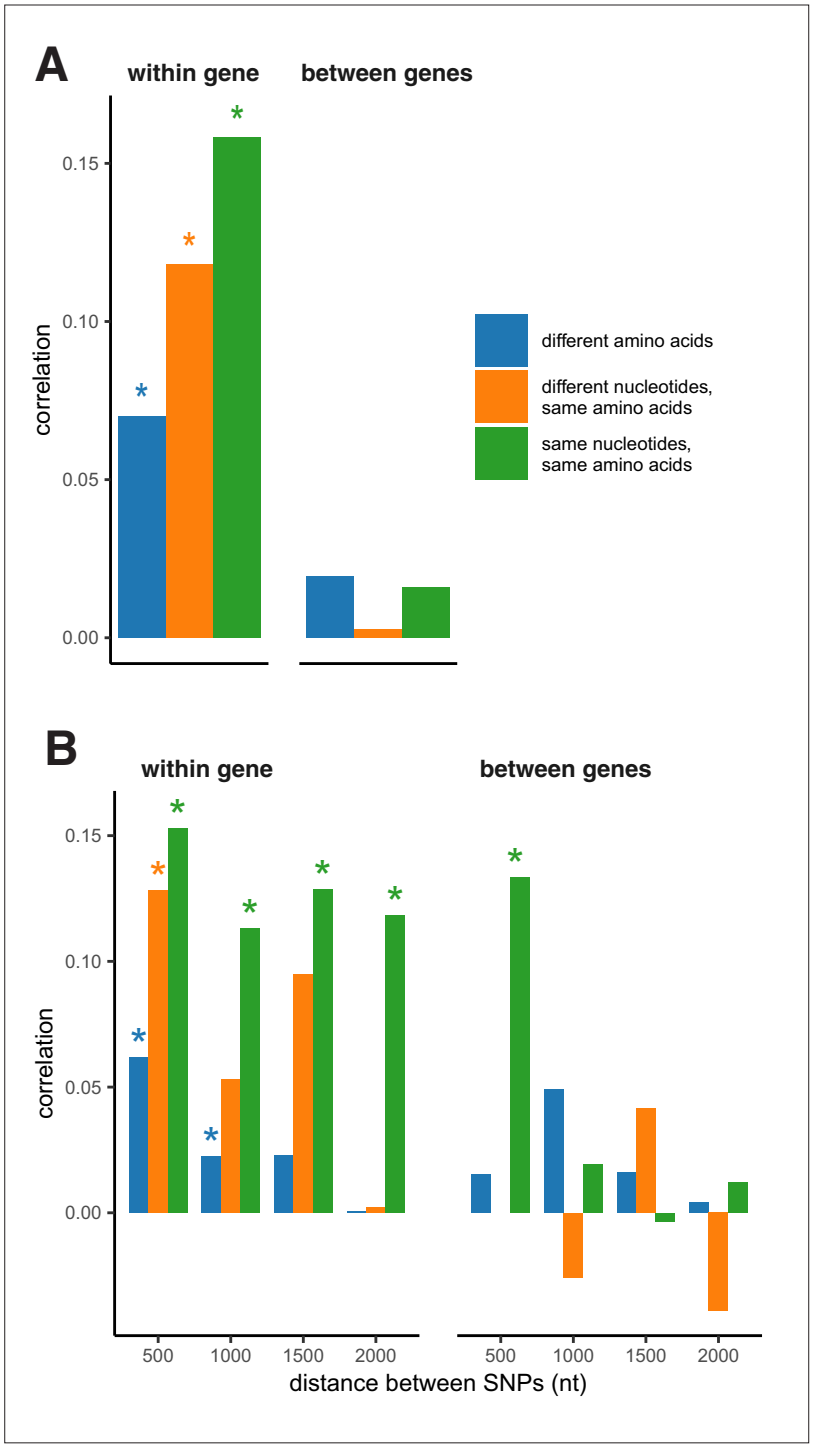


Figure 4—figure supplement 1. Association of LD values between pairs of shared nonsynonymous SNPs encoding the same amino acids in the two *S. commune* populations. (**A**) All pairs of SNPs pooled together. Pair of SNPs is considered to carry different alleles if at least one allele differs in at least one site. (**B**) Pairs of SNPs stratified by distance between them. Asterisks indicate Spearman correlation p-values <0.01.

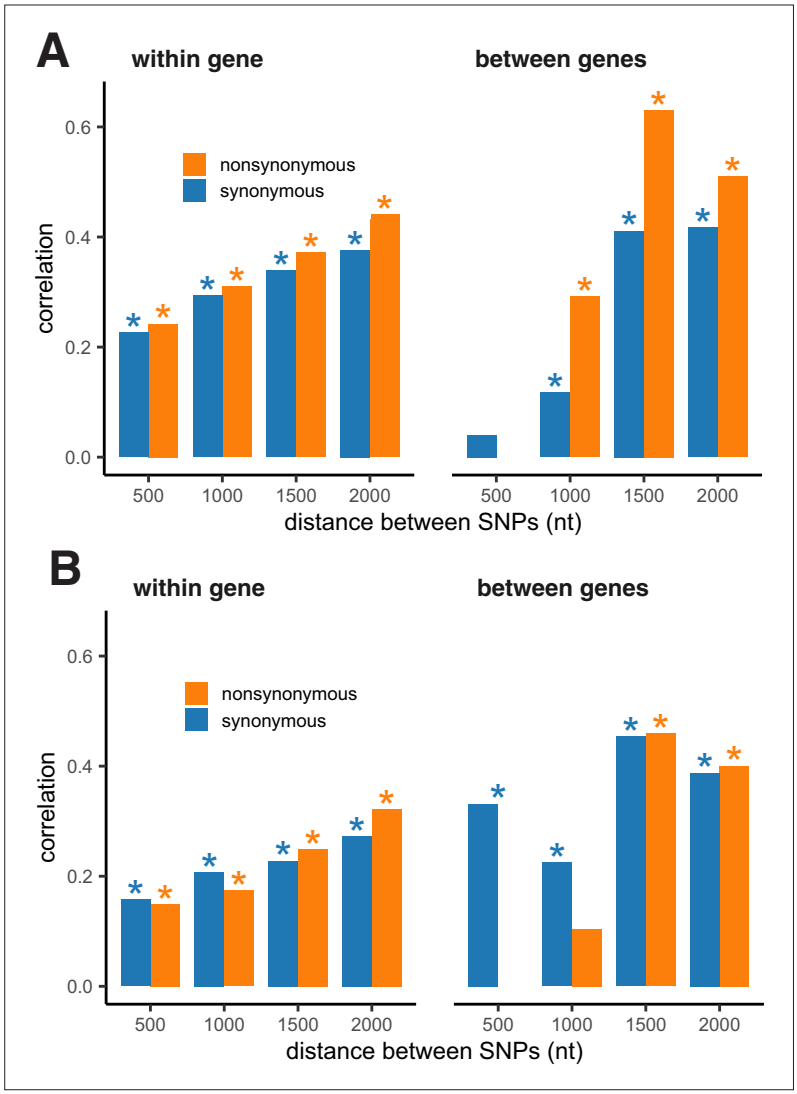
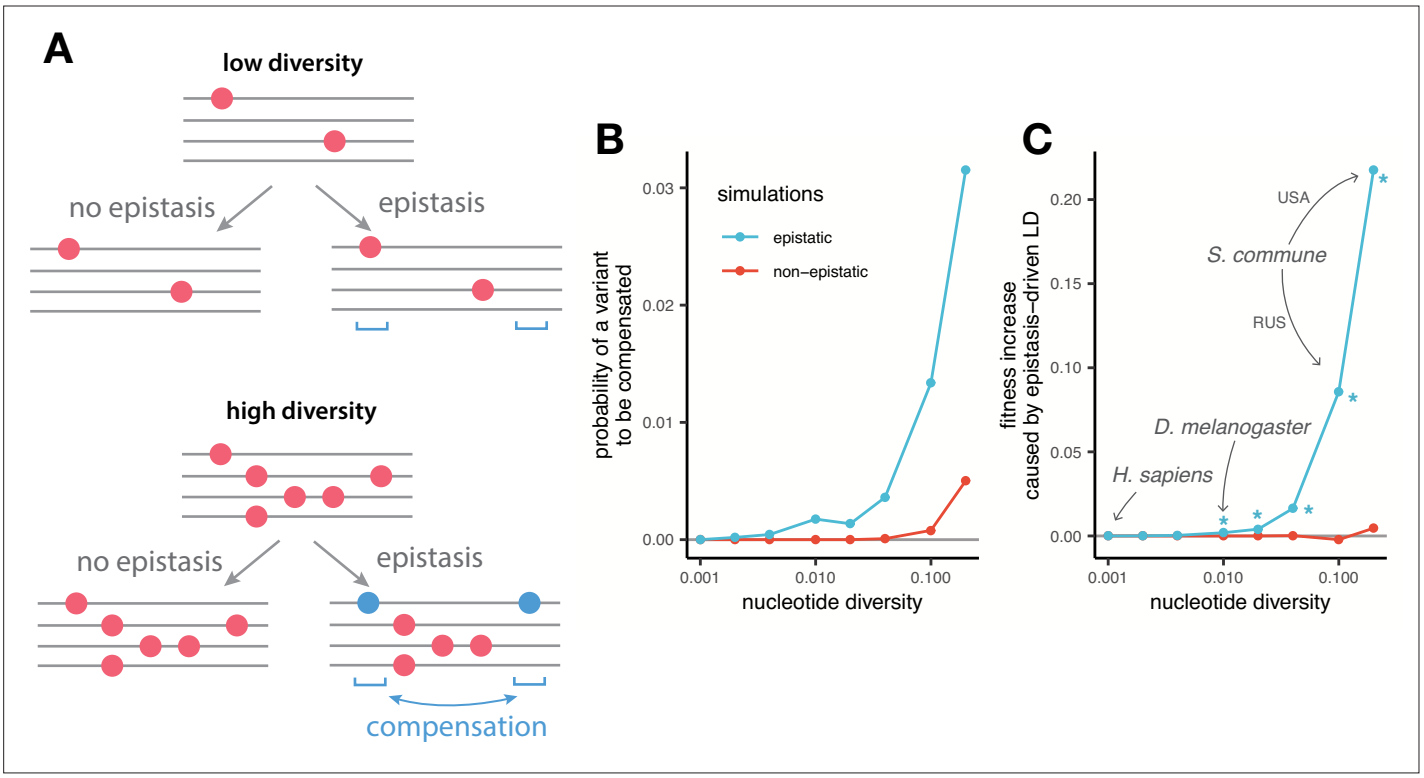
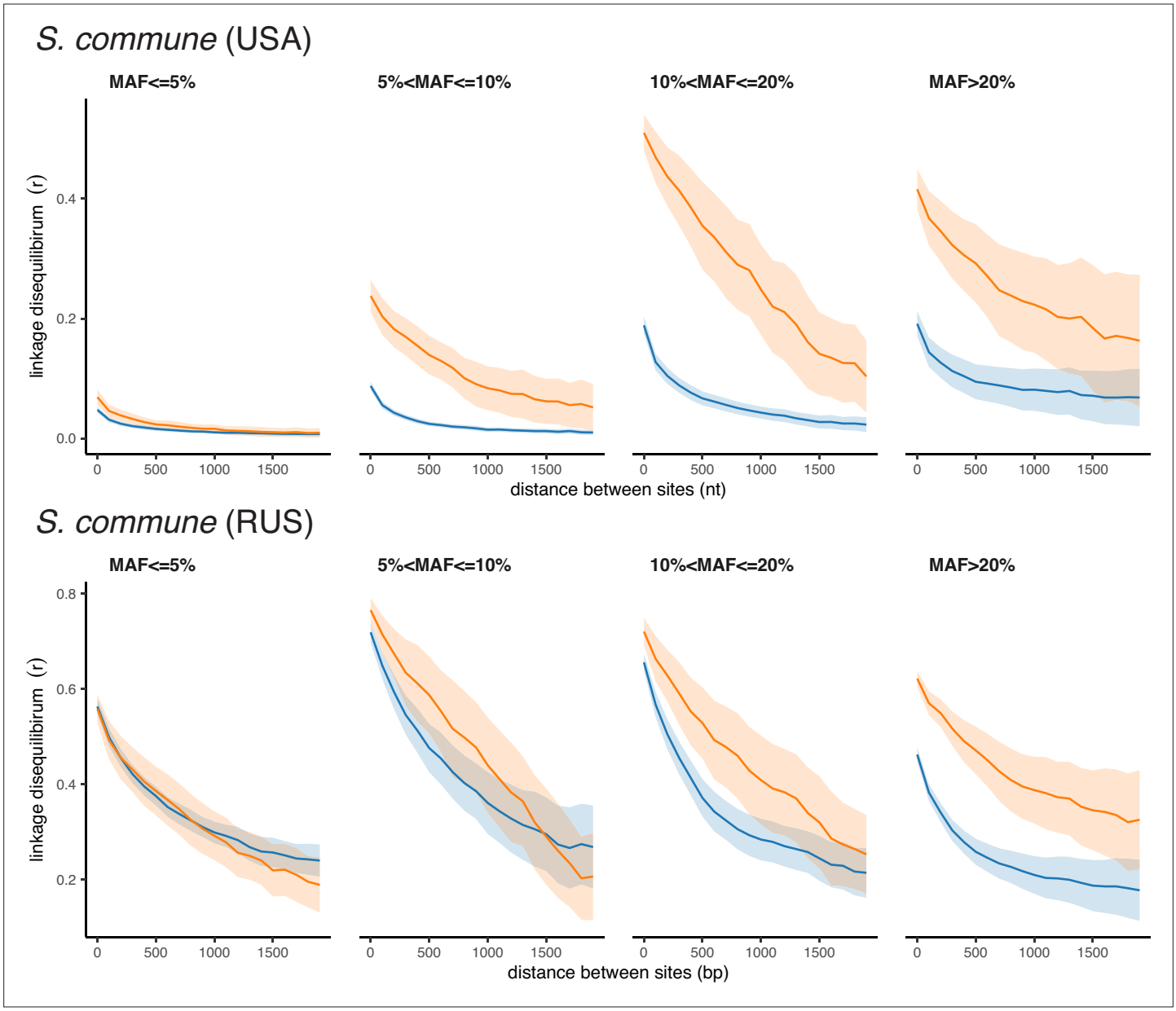


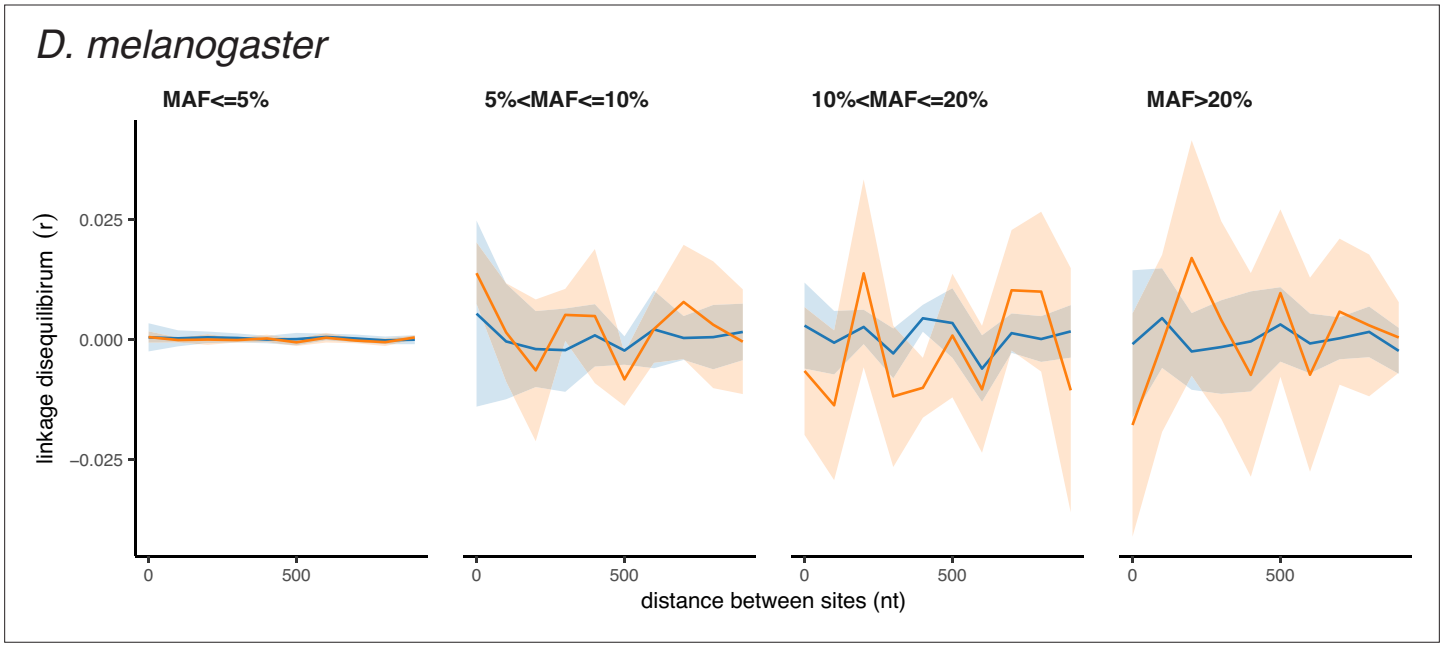
Figure 4—figure supplement 2. Association of LD values between pairs of shared SNPs within haploblocks in the two *S. commune* populations. (**A**) Pairs of SNPs with the same major and minor alleles in both sites, (**B**) pairs of SNPs differing by at least one allele. Asterisks indicate Spearman correlation p-values <0.001.



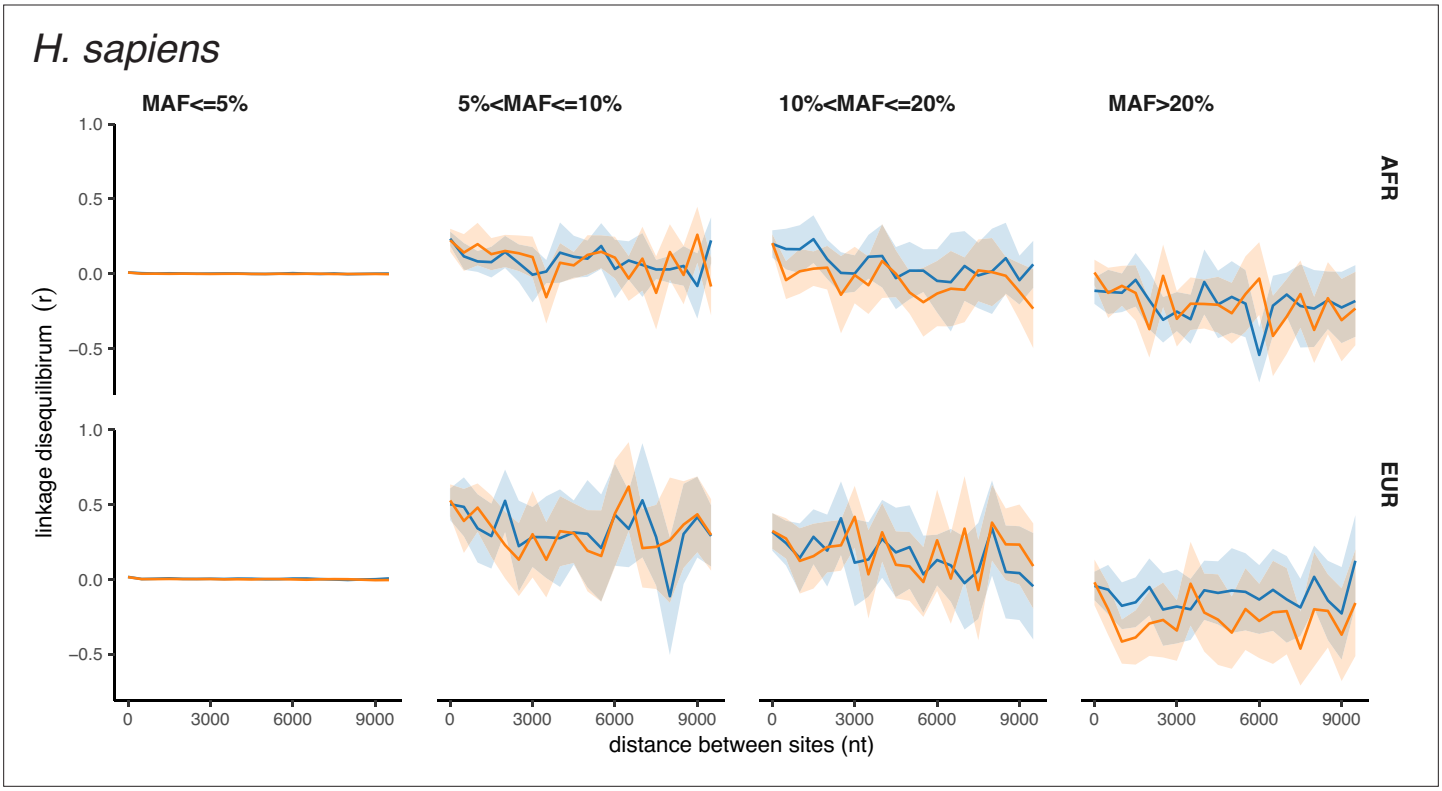
Appendix 1—figure 1. The efficiency of epistasis in populations with different levels of nucleotide diversity. (**A**) Under low nucleotide diversity, deleterious mutations (red dots) are unlikely to be compensated. If nucleotide diversity is high, epistatic selection maintains LD between SNPs in interacting sites (blue dots). (**B**) The probability that a deleterious variant is compensated by another variant within the same individual at the end of the simulation. (**C**) Increase in the mean fitness of the population caused by epistatic selection maintaining LD between favorable allele combinations. The fitness is plotted relative to that of a population consisting of individuals with uncorrelated alleles at different sites, obtained by permuting alleles among individuals. The efficiency of epistatic selection in maintaining linkage is much higher in genetically variable populations. Asterisks in (**C**) indicate significant deviation from 0 (Wilcoxon paired test p-value <0.01). Each simulation was repeated between 100 and 10,000 times depending on genetic diversity.



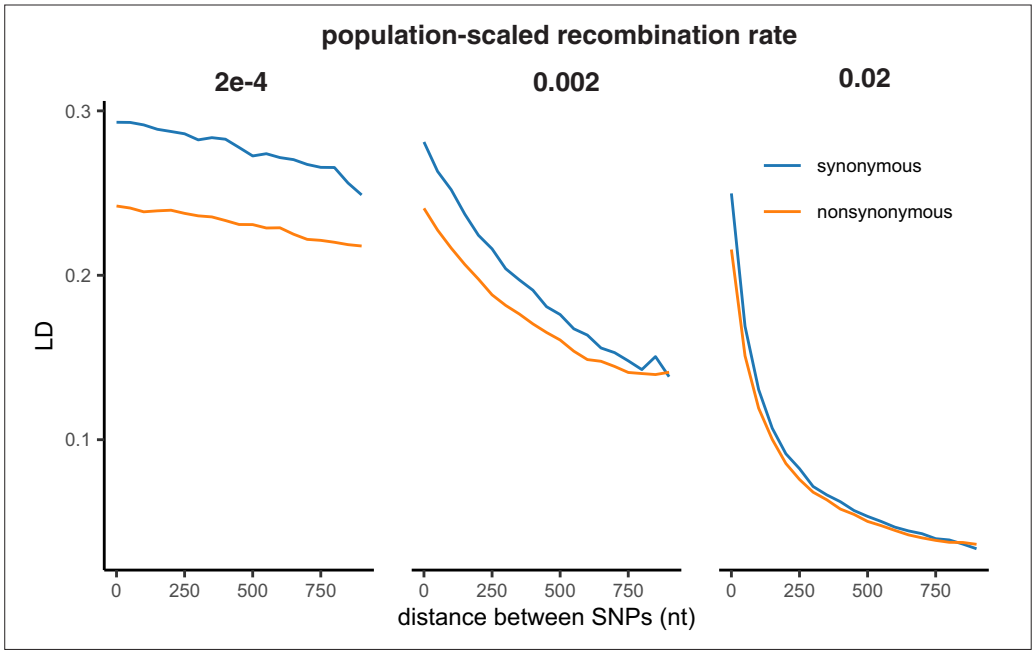
Appendix 2—figure 1. Polarized linkage disequilibrium in *S. commune*. LD between nonsynonymous SNPs is shown in orange, and LD between synonymous SNPs is shown in blue. Filled areas indicate SE of LD calculated for each scaffold separately.



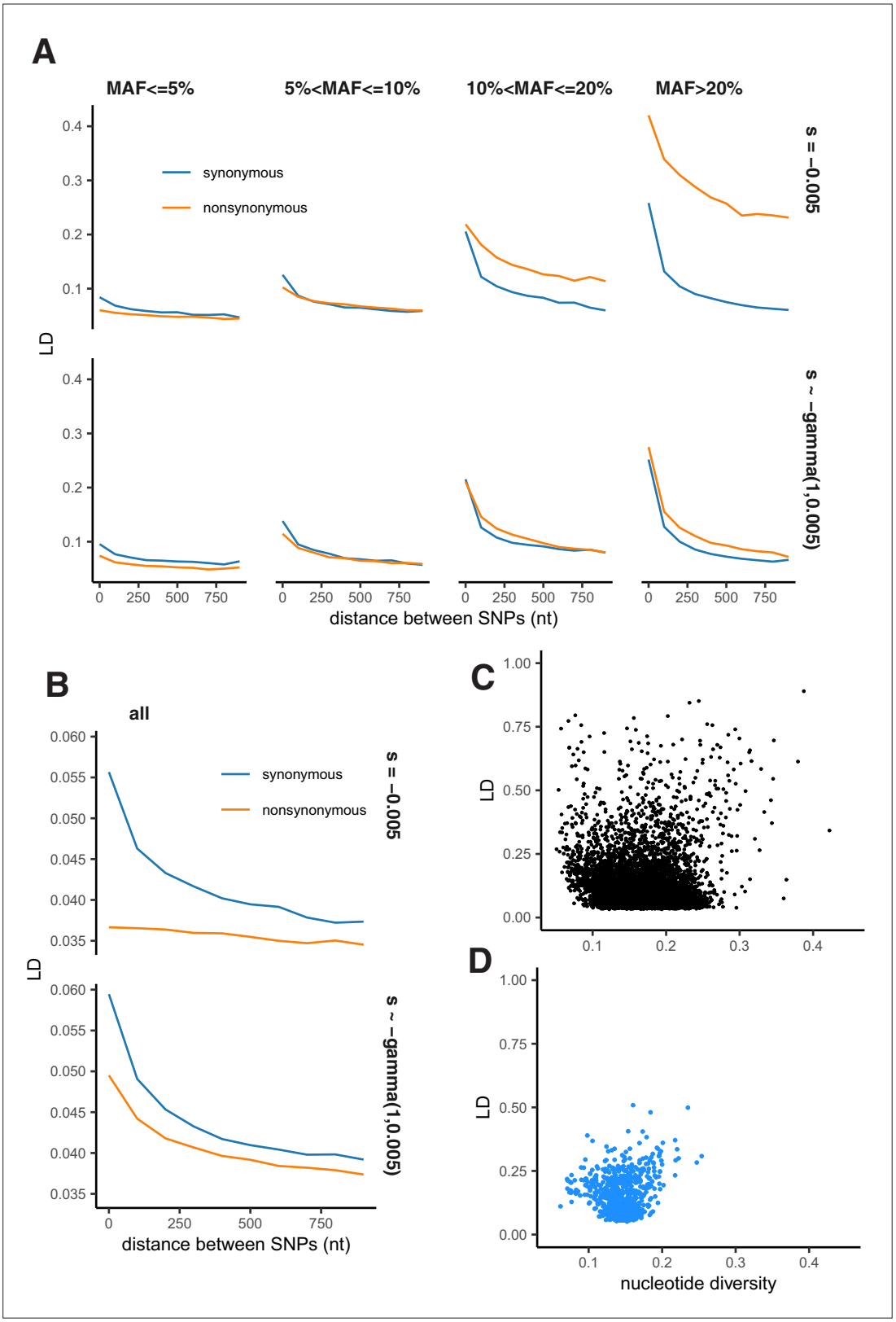
Appendix 2—figure 2. Polarized linkage disequilibrium in *D. melanogaster*. LD between nonsynonymous SNPs is shown in orange, and LD between synonymous SNPs is shown in blue. Filled areas indicate SE of LD calculated for each chromosome separately.



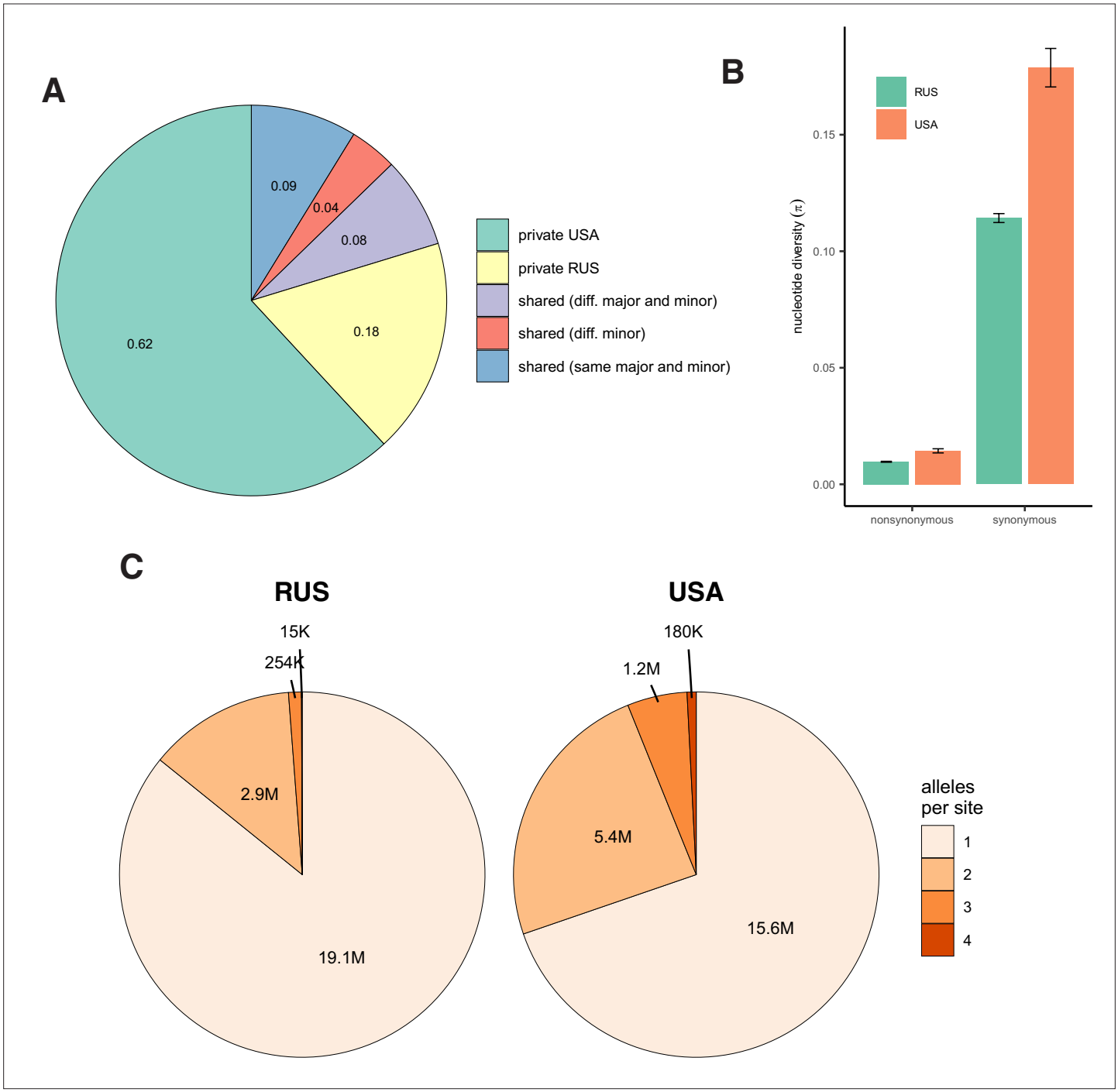
Appendix 2—figure 3. Polarized linkage disequilibrium in *H. sapiens*. LD between nonsynonymous SNPs is shown in orange, and LD between synonymous SNPs is shown in blue. Filled areas indicate SE of LD calculated for each chromosome separately.



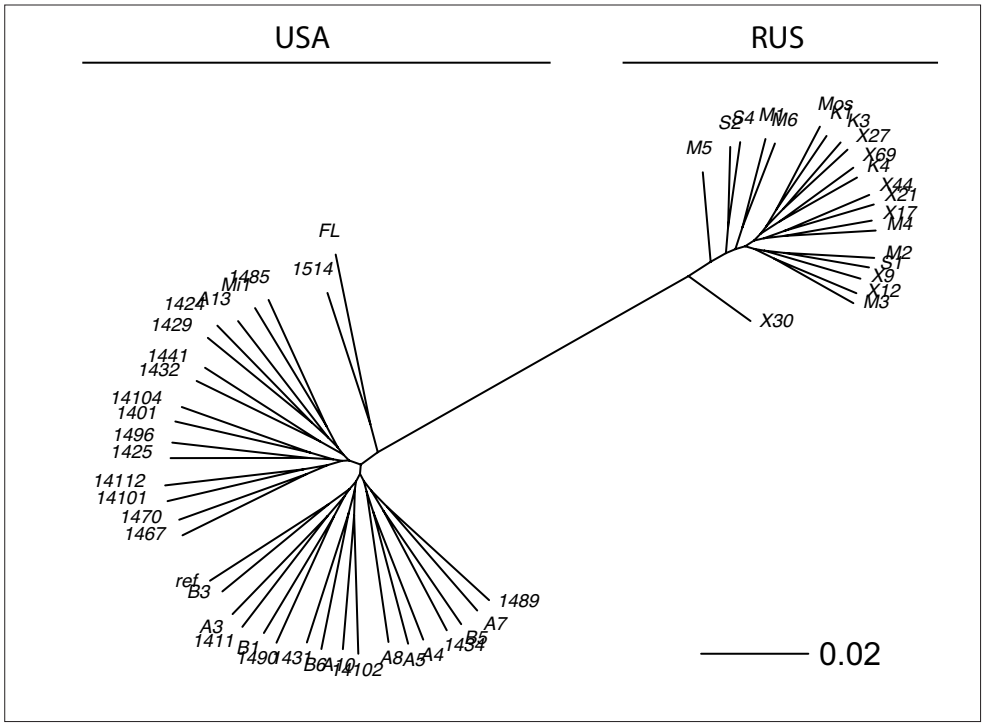
Appendix 2—figure 4. LD_{nonsyn} and LD_{syn} in simulations under weak negative selection. LD between synonymous (blue, selection coefficient $N_e s = 0$) and nonsynonymous (orange, $N_e s = -1$) variants under varying recombination rate. Only SNPs with MAF >0.05 are shown. Simulated haploid population size N=2000, sequence length L=1000 bp.



Appendix 2—figure 5. Patterns of LD in simulations under Hill-Robertson interference. (**A**) LD between nonsynonymous and synonymous pairs of SNPs split by MAF. (**B**) LD between all pairs of nonsynonymous and synonymous SNPs pooled together. (**A–B**) Simulated haploid population size N=2000, sequence length L=1000 bp. Top panels - selection coefficients of all nonsynonymous mutations are equal to –0.005 ($N_e s$ = –10); bottom panels - selection coefficients of nonsynonymous mutations are gamma-distributed with parameters rate = 1, scale = 0.005. (**C**) LD and nucleotide diversity within genes of the USA population of *S. commune* (each point represents one gene). (**D**) LD and nucleotide diversity obtained in simulations.



Appendix 3—figure 1. Patterns of nucleotide diversity in *S. commune*. (**A**) The fraction of private and shared biallelic SNPs. (**B**) Within-population nucleotide diversity at different classes of sites (measured as π without Jukes-Cantor correction). (**C**) The number of monomorphic and polymorphic sites in the multiple whole-genome alignments of *S. commune* genomes.



Appendix 3—figure 2. The reconstructed phylogeny of *S. commune*. USA and Russian populations of *S. commune* are highly divergent while having almost no within-population structure. Genetic distance is measured in nucleotide differences, the phylogeny is reconstructed based on the multiple whole-genome alignment.