nature genetics

Incorporation of Evolutionary Constraint Improves Genomic Prediction of Hybrid Phenotypes

Jinliang Yang[1,2], Sofiane Mezmouk[1,2,3], Andy Baumgarten[4], Edward S. Buckler[5], Katherine E. Guill[6], Michael D. McMullen[6,7], Rita H. Mumm[8], and Jeffrey Ross-Ibarra[1,9,10]

Manuscript intended for *Nature Genetics*, September 2, 2015

Complementation of deleterious alleles has long been proposed as a major contributor to the hybrid vigor observed in offspring of inbred parents. We test this hypothesis using evolutionary measures of sequence conservation to ask whether incorporating information about putatively deleterious alleles can inform genomic selection (GS) models and improve phenotypic prediction. We measured a number of agronomic traits in both the inbred parents and hybrids of an elite maize partial diallel population. We resequenced the parents of the population, using genomic evolutionary rate profiling (GERP) to identify constrained sites across more than 86 Mb sites across the genome. We identifed haplotype blocks using an identity-by-decent (IBD) analysis and scored these blocks on the basis of segregating putatively deleterious variants. As a result, incorporating sequence conservation improves prediction accuracies in a five-fold cross-validation experiment for several traits *per se* as well as heterosis for those traits. These results provide strong empirical support for the simple complementation model of heterosis, and demonstrates the utility of incorporatin functional annotation and its potential usage in phenotypic prediction and plant breeding.

[1]Department of Plant Sciences, University of California, Davis, CA 95616, USA
[2]These authors contributed equally to this work
[3]Current address: KWS SAAT AG, Grimsehlstr. 31, 37555 Einbeck, Germany
[4]DuPont Pioneer, Johnston, IA 50131, USA
[5]US Department of Agriculture, Agricultural Research Service, Ithaca, NY 14853, USA
[6]US Department of Agriculture, Agricultural Research Service, Columbia, MO 65211, USA
[7]Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA
[8]Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[9]Center for Population Biology and Genome Center, University of California, Davis, CA 95616, USA
[10]Correspondence should be addressed to J.R-I. (rossibarra@ucdavis.edu).

The phenomenon of heterosis or hybrid vigor has been observed across many species, from yeast[1] to plants[2] and vertebrates[3]. Hybrid vigor is particularly important in agriculture, where hybrid breeding is fundamental to the production of a number of crops including both rice and maize . A number of hypotheses have been put forth to explain the phenomenon, including gene dosage[4], overdominance[5–7], and epistasis[8,9]. Complementation of recessive deleterious alleles, however, remains the simplest genetic explanation[10], and one that is supported by considerable empirical evidence[11,12]. *i'm not sure these are all appropriate. huang2015genomic for example i think doesn't actually measure heterosis, they just do GWAS on hybrids. and frascaroli if I read it right says that heterosis for grain yield is mostly overdominance (dominance $>$ 1).*

While this can be possible for model species with genome-wide SNP data such as Drosophila and Arabidopsis[??], it is difficult for bacteria that usually have strong and complex clonal population structure. Therefore, several *ad hoc* methods have been developed[??], which are designed to minimize the effect of demography. In practice, these methods focus on very recent variations and distinguish their causal mechanisms (*i.e.,* mutation vs. homologous recombination), from which the relative rate of homologous recombination and mutation is estimated. Because such recent events are indeed robust to demographic effects, application of these methods has provided improved estimates using large-scale sequence variation data, such as those made available by the multilocus sequence typing (MLST) project, in which short fragments in seven housekeeping genes were resequenced for a large sample from various species[?].

However, a serious problem still remains when clonal structure is very strong, as illustrated in **Fig. 1**. To demonstrate the point, we assumed a very simple model with multiple clonal subpopulations, among which migration is allowed (**Fig. 1a**). When the migration rate is low, it is considered that the population has strong clonal structure, while individuals in the entire population behave as if it is panmictic with frequent mi-

Figure 1 (**a-b**) Demonstrating the performance of Feil et al.'s method[??] for estimating the relative rate of homologous recombination (g/μ). The effect of population (clonal) structure is investigated using the model illustrated in (**a**), which considers 50 subpopulations that share an ancestral population. Migration among the subpopulations is allowed, and the migration rate represents the level of population structure. It is shown that g/μ is highly underestimated when the migration rate is small, while g/μ is well estimated when the entire population is less structured (**b**). (**c**) Distribution of the estimates of the relative recombination rate (g/μ) in 48 bacterial species, summarized by Vos & Didelot[?]. Note that we use this data set[?] based on Didelot & Falush's method[?] because estimates by Feil et al.'s method are available for far fewer species[?]. We verified the correlation of the estimates obtained by the two methods, as we show later.

gration. We assumed that the rates of homologous recombination (g) and mutation (μ) are identical, so that the true relative rate (g/μ) is 1. Then, genotype data that are similar to the MLST data were generated by coalescent simulations using the ms software[?], and the relative rate of homologous recombination was estimated by using Feil et al.'s method[??]. We found that the rate was well estimated under the assumption of frequent migration, yet was highly underestimated under a strong clonal structure. Further, it seems that the method has very limited power to detect recombination when the population is highly structured (**Fig. 1b**). This is because recombination can be detected only when the exchanged sequences have accumulated sufficient nucleotide variation. In each clonal population, where the genomic sequences of all individuals are very similar, most recombination events are not detectable. This reasoning also holds for the method of Didelot and Falush[?], which preferentially detects recombined regions from highly diverged clones. Therefore, we hypothesize that the estimated rates of homologous recombination thus far could be much lower than the true rates in natural populations, particularly for species with strong clonal structure.

In this study, we set out to verify this hypothesis by re-evaluating the homologous recombination rate in bacteria, mainly focusing on a species with strong clonal structure, given that, for such species, the extent of underestimation in the homologous recombination rate would be significant (**Fig. 1**). Based on this rationale, we chose *Staphylococcus aureus* as a model organism. *S. aureus* is a major pathogen that is associated with serious community-acquired and nosocomial diseases[??]. Methicillin-resistant strains (MRSA) of this species are well recognized because they are resistant to multiple antibiotics, including methicillin, and their infection could cause serious diseases[???]. Due to these medical concerns, *S. aureus* is one of the species for which genomic sequences of multiple individuals are available; whole-genome sequences for this species are available for more than ten strains, including MRSA and MSSA (methicillin-susceptible *S. aureus*) strains (**Supplementary Table 1**). It has long been believed that *S. aureus* has very strong clonal structure[????]. Previous analyses indicated that recombination rate in *S. aureus* is very low. The ratio of the homologous recombination rate to the mutation rate (g/μ) in this species was estimated to be about 0.067[?] by Feil et al.'s method[??]. Narra and Ochman[?] classified 23 bacterial species into three categories (high, medium, and low) according to the recombination rate estimated by Feil et al.'s method, and placed *S. aureus* in the 'low' category. Vos & Didelot[?] summarized estimates of the relative homologous recombination rate (g/μ) for 48 bacterial species using the method of Didelot and Falush[?]. The estimate for *S. aureus* was g/μ = 0.1, which was the second lowest among the 48 species (see **Fig. 1c** for the distribution of estimates for the 48 species). Thus, *S. aureus* provides an appropriate model organism for our purpose of evaluating the homologous recombination rate in a highly structured population. If our hypothesis turned out to be correct, the low estimates for this species would not be true, but rather a mere result of underestimation due to the strong population structure.

In this study, we used a population genetics approach that would provide a more accurate estimate (robust to demography) as long as reasonable demography was assumed. We first inferred the demographic history and then estimated the rate of homologous recombination conditional on the inferred demography. While this approach has been applied quite successfully to model eukaryotic species with abundant SNP data[??], its application to bacteria with complex population structures is not straightforward. Therefore, when applied to *S. aureus*, rather than using all strains, we focused only on a subset, which allowed us to use the coalescent framework for analyses and simulations. To this end, we show that the relative rate of recombination is about 10 times larger than the previous estimates. Our results imply that bacteria could undergo much higher rates of homologous recombination than previously thought, postulating an important role for this process in bacterial genome evolution. To validate this postulation, we investigated the evolutionary significance of homologous recombination. It is widely accepted that recombination is evolutionarily advantageous because it can break down linkage between loci and subsequently create new combinations of alleles[?????], an argument that is frequently used to explain the existence of two sexes. As such, the common assumption that the evolution of asexual species, such as bacteria, is largely clonal and does not utilize recombination, puts such species at an evolutionary disadvantage in at least two scenarios. In the first scenario, known as "clonal interference" of beneficial mutations, when two beneficial mutations arise simultaneously in different individuals, it is impossible that both of them fix in the population because when one fixes, the other has to be lost[????]. In the second scenario, deleterious mutations likely fix in the population. For example, through the "hitchhiking" effect, when a beneficial mutation fixes by positive selection, linked deleterious muta-

Figure 2 Genome-wide pattern of homologous recombination in *S. aureus*. (**a**) An NJ tree of the 12 strains in *S. aureus* based on the distance matrix of all synonymous SNPs. The 12 strains were classified into 5 groups, named A, B, C, D and E. (**b**) The proportions of different tree-shapes. (**c**) The distribution of tree-shapes across the genome. (**d**) Local patterns of gene trees in three representative regions with length 10 kb. Gray boxes represent coding genes. For each gene, an NJ tree was constructed if no evidence for recombination was detected by the four-gamete test[?], otherwise the gene region was divided into blocks at putative recombination breakpoints and a tree is shown for each block.

tions also fix[??]. The effect of background selection can be maximized with no recombination because of a high proportion of individuals having strongly deleterious mutations, causing a serious reduction in the effective population size. As a consequence, deleterious mutations have a high chance to fix by genetic drift[?]. An extreme case of this would be the situation known as Muller's ratchet; if the deleterious mutation rate is so high that all individuals have at least one deleterious mutation, then at least one of them will eventually fix in the population[??].

It may be possible that bacteria can overcome, by utilizing homologous recombination, the aforementioned evolutionary disadvantage scenarios, but this possibility has not been systematically investigated due to the lack of knowledge on the rate of recombination in natural populations. Instead, a view to the contrary has been provided: since homologous recombination affects very limited genomic regions, it may not have sufficient power to relax those evolutionary disadvantages[?]. The role of homologous recombination has gained more appreciation, but its evolutionary advantages are still under debate, and the arguments are mainly based on experimental evolution studies[????]. To address this issue, here we use simulations as an alternative approach. By simulating bacterial genome evolution with advantageous and deleterious mutations and homologous recombination, we investigated the fixation rates of both types of mutations. Based on the simulation results, we discuss the bacterial population's ability to avoid evolutionary disadvantages by means of homologous recombination.

RESULTS

Evaluating the homologous recombination rate in *S. aureus*

The whole-genomic sequences of 12 strains of *S. aureus* were obtained from the NCBI database (ftp://ftp.ncbi.nih.gov/genomes/). *S. aureus* has a circular genome, and the genome sizes of the 12 strains range from 2.74Mb to 2.91Mb. These strains were isolated from humans, except for RF122, which was isolated from a bovine. The 12 strains consist of 8 methicillin-resistance (MRSA) and 4 methicillin-susceptible (non-MRSA) strains (**Supplementary Table 1**). The genomes were aligned and 65,412 SNPs were identified in 2.3 Mb of well-aligned regions, which cover more than 80 % of the entire *S. aureus* genome (on average, $\sim$2.8 Mb; **Supplementary Methods** online). The average pairwise nucleotide differences per site is $\pi = 0.00847$ for all sites, and $\pi_S = 0.0247$ for synonymous sites (**Supplementary Methods** online).

Estimation of the homologous recombination rate from polymorphism data is difficult without knowing the demographic history of the population. This is especially true for bacterial populations that usually have strong clonal structure[????]. Indeed, the relationship of the 12 strains is very different from that expected when the sample is from a panmictic population. **Fig. 2a** shows an NJ tree based on the distance matrix of all synonymous SNPs. This tree represents the genome-wide average of tree structure of the 12 strains, and indicates that the 12 strains can be classified into 5 groups, named A, B, C, D and E. The three groups, A, B and C, are more closely related and consist of most strains (10/12). This pattern is consistent with previous inferences of the history of this species using the MLST data[??]. Cooper & Feil[?] grouped *S. aureus* strains into two major clades, and our groups A, B and C belong to one and E belongs to the other (D, RF-122, is not included). Each of the first three groups (A, B, and C) consists of multiple strains with almost identical genomes (nucleotide difference is < 0.0003), which is in agreement with the sampling history; for example, JH1 and JH9 in group C were isolated from a single patient. COL and USA300 belong to group A; it is known that the former is a strain that was isolated about 50 years ago, and that the latter is recently derived from the former[?].

We here focus on the relationship among three groups, A, B and C, because they should be ideal for estimating the recombination rate. We can consider that the coalescent patterns of the ancestral lineages of the three groups are very similar to that expected in a panmictic population undergoing extensive recombination because of two reasons. First, the three possible coalescent patterns, ((A, B), C), ((B, C), A) and ((A, C), B), appear with nearly identical frequencies. We here define the "major topology" such that A, B and C are more closely related, as illustrated in **Fig. 2b**. The major topology includes those with all three possible coalescent patterns for the A-B-C trio. It does not specify the coalescent patterns of D, E and the ancestor of A-B-C, because this relationship cannot be resolved due to the lack of an outgroup. In the well-aligned regions, there are 1,788 coding genes. For each gene, an NJ tree was constructed (genes with very few SNPs were excluded, and those consist of about 20% of the 1,788 genes). **Fig. 2b** shows that more than 80 % of those trees are consistent with the major topology, among which the proportion of three patterns are very similar, though ((A, B), C) is slightly more common than the other two. Second, the shape of the gene tree changes gene by gene, that is, the three types of trees distribute almost randomly along the chro-

Figure 3 (**a**) The demographic model for the A-B-C trio used in this study. (**b**) The distributions of the synonymous nucleotide divergence between groups A-C (green boxes in the upper panel), B-C (blue boxes in the upper panel) and A-B (red triangles in the lower panel). The expected distributions under the inferred demography are shown by gray lines. (**c**) The decay of linkage disequilibrium (LD). The horizontal and vertical axes represent the distances (kb) between SNPs and the probability of tree-shape compatibility, respectively. The gray circles represent the observation (the proportion of compatible trees for SNP pairs given distance). Distances were binned, and each circle represents the proportion of compatible trees in the same bin. The red line represents the expected decay of LD with the estimated rate, $\hat{G}$ = 0.006, $1/\hat{q}$ = 10 kb.

Figure 4 Decay of LD (r^2) in various bacterial species. In addition to *S. aureus* (**a**), eight species (*Bacillus cereus*, *Campylobacter jejuni*, *Clostridium botulinum*, *Escherichia coli*, *Helicobacter pylori*, *Salmonella enterica*, *Streptococcus pneumoniae*, and *Streptococcus pyogenes*) were analyzed (**b**∼**i**). In all species, LD decays quite dramatically and saturates in several kb. The saturated level of LD ($LD_{\infty,obs}$) is roughly shown by a red broken line, which is usually higher than the level expected between completely unlinked SNPs ($LD_{\infty,exp}$, blue broken line), possibly due to population structure. (**j**) $LD_{\infty,obs}/LD_{\infty,exp}$ in seven species, for which estimates of the recombination rate (g/μ) are available in Narra and Ochman[?]. (**k**) $LD_{\infty,obs}/LD_{\infty,exp}$ in all nine species, plotted against estimates of g/μ by Didelot & Falush's method[?] (data from Vos & Didelot[?]).

mosome (**Fig. 2c**). A more detailed view is shown in **Fig. 2d**. A tree is shown for a gene if no evidence for recombination is detected by the four-gamete test[?], otherwise the gene region is divided into blocks at putative recombination breakpoints and a tree is shown for each block. We found recombination breakpoints on average every 1.6 kb across the genome. These two major observations indicate that the ancestral lineages of A, B, and C seem to meet in a presumably large ancestral population quite recently, and extensive homologous recombination makes the coalescent pattern of the trio in each gene nearly random. The slight excess of ((A, B), C) over ((B, C), A) and ((A, C), B) indicates that the population split of A and B is slightly younger than that of C and the ancestor of A and B.

To obtain more detailed insights into the demographic history of the ancestral lineages of the three groups, we estimated the ancestral population sizes and divergence times assuming a simple model illustrated in **Fig. 3a**. According to the theories in refs.[??], we estimated the demographic parameters involved in the model (**Supplementary Methods** online). As expected, we found that the A-B-C trio shared a very large ancestral population; our maximum likelihood (ML) estimate of the population mutation rate is $\hat{\theta}_2 = 2\hat{N}_2\mu = 0.0105$, where N_2 and μ are the effective population size and mutation rate per site, respectively, so that $\hat{N}_2$ is estimated to be 5.3×10^7 if $\mu = 10^{-10}$ is assumed[?]. $\hat{t_2}$ was estimated to be 3.5×10^6 generations (200 years if 1 generation per 30 min is assumed[?]), which corresponds to only 7% of the mean coalescent time in the ancestral population. After the split of AB and C, A and B shared an ancestral population of size N_1, which was estimated to be $0.569\hat{N}_2$, and the time of population split between A and B was estimated to be $\mu\hat{t}_1 \approx 0$. **Fig. 3b** shows that the distributions of observed divergences among the A-B-C trio are in excellent agreement with the expectations under the inferred demographic model, indicating that the inferred model well represents the coalescent process among the A-B-C trio. It should be noted that this estimated demography does not necessarily reflect the geographic distribution of the sampled strains, because this species could have undergone recent extensive migration together with human migration. This is why MSSA476 and MW2, which were sampled in the UK and US[?], respectively, belong to the same group, B.

Conditional on this estimated demography, the rate of homologous recombination was estimated from the decay of linkage disequilibrium (LD) along distance. Representatives of the three groups, A, B and C, provide the minimum sample size to detect recombination when an outgroup (D or E) is available. Following the method of Ruderfer *et al.*[?], we used 5,289 SNPs at which the allelic configuration of {A, B, C, D, E} $\in \{\{1,1,0,0,0\},\{0,1,1,0,0\},\{1,0,1,0,0\}\}$, where 0 and 1 represent two variable nucleotides. For these sites, it is very likely that 0 is the ancestral allelic state; therefore, the tree shape can be parsimoniously inferred (*i.e.,* ((A, B), C), ((B, C), A) and ((A, C), B) are given for $\{1,1,0,0,0\}$, $\{0,1,1,0,0\}$, and $\{1,0,1,0,0\}$, respectively). It is expected that the probability of tree-shape compatibility for a pair of completely linked sites is 1 and this probability decreases as the recombination rate between the two sites increases. When the two sites are completely unlinked, the probability is expected to be 0.34. Thus, the decrease of the probability of tree-shape compatibility against distance is analogous to the decay of LD. **Fig. 3c** shows the average probability of tree-shape compatibility, which is given by a decreasing function of distance. It was found that the probability decreases dramatically and becomes close to the theoretical minimum when the distance is larger than 5 kb.

Our observation suggests extensive homologous recombination along the chromosome. Coalescent simulations were performed to estimate the rate of homologous recombination. A homologous recombination event is modeled such that the process is analogous to allelic gene conversion[??]. We take the parameter g to represent the initiation rate of a transferring event per site per generation, and G to be the population rate, $G = 2Ng$, where N is the effective population size. The elongation of the converted tract starts at the initiation site and is terminated at a constant rate, q. Therefore, the tract length follows a geometric function with mean $1/q$, and the two parameters G

and q determine the decay function (**Supplementary Methods** online). We found that $\hat{G}$ = 0.006 with $1\hat{/}q \geq 10$ kb explains the observation very well (**Fig. 3c**, **Supplementary Methods** online).

Our estimate of the ratio of the recombination to mutation rate turns out to be 0.6 (*i.e.,* G/θ or g/μ), which is roughly 10 times higher than previous estimates (0.067-0.1)[??]. However, if our new estimate is placed in the distribution of previous estimates based on the method of Didelot and Falush[?] (**Fig. 1c**), the value (0.6) is still low; roughly three quarters of species have higher estimates of the recombination rate. Furthermore, if underestimation of the previous methods can be applied to species with estimates as low as that of *S. aureus* (a quarter of species in the left of the distribution in **Fig. 1c**), we suspect that there is substantial variation in the homologous recombination rate among species, and that *S. aureus* is still in a category of species with very low rate of recombination.

This idea is supported by investigating the decay of LD in genomic sequence data for multiple species. **Fig. 4** shows the patterns of LD decay (measured by r^2[?]) in various bacterial species, for which multiple genomic sequences are available, including *S. aureus* (**Fig. 4a**). There is a major difference between **Fig. 3c** and **Fig. 4a**, although both focus on the decay of linkage between SNPs. In **Fig. 3c**, LD decays and saturates at the value expected assuming free recombination between SNPs. On the other hand, the level of saturated LD (denoted by $LD_{\infty,obs}$, red dashed line in **Fig. 4**) is higher than expected under the free recombination assumption (denoted by $LD_{\infty,exp}$, blue dashed line in **Fig. 4**). This should be because of the effect of demography; population (or clonal) structure inflates the genome-wide average level of LD[?]. In other words, the degree of the elevation of the observed saturated LD ($LD_{\infty,obs}/LD_{\infty,exp}$) could be mainly explained by the extent of population structure. It is predicted that the degree of underestimation of the recombination rate by previous methods could be large for species with high $LD_{\infty,obs}/LD_{\infty,exp}$, while the recombination rate could be well estimated for species with $LD_{\infty,obs}/LD_{\infty,exp} \sim 1$. **Fig. 4** compares the patterns of LD decay in nine species including *S. aureus*, showing that $LD_{\infty,obs}/LD_{\infty,exp}$ is highly variable. It has been known that *Escherichia coli* and *Bacilus cereus* have relatively strong population structures, while the population of *H. pylori* could be well mixed[??]. This seems to be well reflected in our observation in **Fig. 4**. **Figs. 4j, k** show that species with low estimates of recombination rate by Feil et al's and Didelot and Falush's methods are likely to have high $LD_{\infty,obs}/LD_{\infty,exp}$. Thus, it is suggested that previous estimates of recombination rate would be strongly associated with the extent of population structure. Therefore, we would predict that species that were thought to have low recombination rates might have higher rates than previously thought. Thus, previous underestimation of recombination rate could not be specific to *S. aureus*.

Therefore, we conclude that our estimate of the homologous recombination rate may be close to the lower boundary in bacteria. In the following, we assess the advantageous roles of this level of homologous recombination in adaptive genome evolution.

Assessing evolutionary advantages of homologous recombination

To investigate the beneficial effect of recombination, forward simulations of the evolution of a bacterial population were performed. The purpose of the simulations is to examine the fixation processes of adaptive and deleterious mutations in a circular genome, which does not undergo meiotic crossing-over but homologous recombination. The model assumes that each individual has a circular genome with size $L = 2 \times 10^6$-bp, in which advantageous and deleterious mutations arise at rates U_A and U_D per genome per generation, respectively. The fitness effects of advantageous and deleterious mutations follow exponential distributions with means $\bar{s}_A$ and $\bar{s}_D$, respectively. $U_A = \{10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}\}$, $U_D = 10^{-4}$, $\bar{s}_A = 0.01$ and $\bar{s}_D = \{0, -0.001, -0.01\}$ were assumed based on empirical estimates (**Supplementary Methods**). Because the bacterial population size would be highly variable, we considered a wide range of $N = \{10^4, 10^5, 10^6, 10^7\}$ (see **Supplementary Methods** online for our choice of parameters).

For each parameter set, a long simulation run was performed to accumulate a large number of adaptive and deleterious substitutions (**Supplementary Methods** online), from which we investigated (i) how the interference among competing adaptive mutations is relaxed by homologous recombination and (ii) how homologous recombination can prevent simultaneous fixations of deleterious mutations. For (i), we focused on K_A, the substitution rate of advantageous mutations per genome per generation (**Fig. 5a**). In each panel, the population size is fixed, and the substitution rates (K_A) obtained from the simulations are plotted. Obviously, K_A is given by an increasing function of the adaptive mutation rate (U_A), and there is also a positive correlation with the rate of homologous recombination. In each panel, we also show the upper and lower limits of K_A (**Supplementary Methods** online). The former assumes free recombination among all adaptive mutations, though this is not a realistic situation in any organism, including eukaryotes. The lower limit was obtained by simulations assuming the entire chromosome is completely linked. As expected, results of all simulations with homologous recombination fall within the range defined by the upper and lower limits (**Fig. 5a**), and as the recombination rate increases, K_A increases, because the interference among competing adaptive mutations is relaxed.

A rough expected value of K_A with the estimated homologous recombination rate (*i.e.,* $\hat{G} = 0.006$) for *S. aureus* is shown by a blue arrow (**Fig. 5a**). It seems that the estimated level of homologous recombination plays a significant role in relaxing the interference, but the effect largely depends on N: The effect decreases as N increases. However, it should be noted that we estimated the recombination rate assuming a neu-

Figure 5 Partial results of the forward simulations. The effect of homologous recombination on the substitution rates of adaptive and deleterious mutations (**a**, K_A and **b**, K_D) are shown. For full results, see **Supplementary Methods** online. The black and white arrows represent the expectations assuming free recombination and complete linkage, respectively. The blue arrows represent the results with the estimated recombination rate without selection taken into account ($\hat{R}_{neu}$). The red arrows roughly show the levels with ($\hat{R}_{sel}$), an estimate taking selection into account (**Supplementary Methods** online). Red arrows are not shown when $N = 10^7$ because we did not obtain reliable simulation results for the decay of LD.

tral population, and that this assumption causes a serious underestimation when selection is operating (**Supplementary Methods** online). To correct this bias, we checked the decay of LD in our simulations with selection, and roughly inferred what recombination rate would be consistent with the decay of LD we observed in *S. aureus* (red arrow). We found that the interference can be significantly relaxed even in large populations. Thus, in *S. aureus*, it may be concluded that homologous recombination plays a crucial role in breaking down the linkage and reconstructing good pairs of mutations, resulting in efficient fixations of adaptive mutations. Although **Fig. 5a** shows the results of the simulations without deleterious mutations, we confirmed that almost identical results were obtained with deleterious mutations (**Supplementary Methods** online).

For (ii), we focused on K_D, the substitution rates of deleterious mutations. The results are well interpreted if the main reason of the fixation of deleterious mutations is the hitchhiking effect: when an adaptive mutation is fixed in the population, linked deleterious mutations could fix along. We found that in general K_D is large with a high adaptive mutation rate. The recombination rate also has a significant effect on the substitution rate. The upper limit can be obtained when there is no recombination, where a number of deleterious mutations will fix by hitchhiking, and lower limit is obtained when all mutations are unlinked (**Supplementary Methods** online). **Fig. 5b** summarizes the results of the simulations when $\bar{s}_D = -0.01$, which shows that homologous recombination reduces substitutions of deleterious mutations in comparison with the case of complete linkage (upper limit). Given the rate of homologous recombination that accounts for the observed LD decay, K_D is close to the lower limit, indicating a significant role of homologous recombination to avoid fixation of many deleterious mutations.

It should be noted that Muller's ratchet or background selection do not seem to play significant roles for fixations of deleterious mutations under our parameter settings. We assumed the deleterious mutation rate according to previous estimates from experimental evolution studies[?]. It is widely accepted that mutations with relatively strong effects ($|s| > 0.01$)[?] can be empirically detectable. It is reasonable to predict that there are a number of deleterious mutations with small effects. If such mutations are taken into account, the relative contributions of the three mechanisms (hitchhiking, Muller's ratchet and background selection) will change. The theoretical conditions to determine the relative roles of the three mechanisms will be published somewhere else.

DISCUSSION

In this study, we re-evaluated the rate of homologous recombination in *S. aureus*, because we suspected that previous estimates might be underestimated due to strong clonal structure of this species (**Fig. 1**). Indeed, our analyses of whole-genome SNP pattern revealed that the population is strongly structured as shown in **Fig. 2a**. Nevertheless, we detected evidence for recombination every ∼1.6 kb (**Fig. 2**), indicating extensive roles of homologous recombination in the genome evolution of this species.

To quantitatively estimate the rate of homologous recombination by taking the effect of demography into account, application of population genetic theories is required. Unfortunately, the population structure of this species seems to be too complicated to apply simple population genetic theories. However, we found that the genealogical relationship among the A-B-C trio is relatively simple and provides an ideal situation for applying population genetic theories (**Figs. 2, 3**). It turned out that our population genetic analysis of genome-wide pattern of SNPs provided an estimate of the ratio of the recombination to mutation rates ($g/\mu = 0.6$) to be roughly 10 times larger than previous estimates for this species. Thus, with a demonstration using *S. aureus*, it is indicated that for estimating the homologous recombination rate, it is very important to take the effect of population structure into account.

Recent accumulation of DNA sequence data (including MLST data) revealed a wide range of the homologous recombination rates in bacteria species[??]. We here demonstrated that species that were thought to have low recombination rates might have higher rates because previous methods[???] would likely underestimate the recombination rate when the population is highly structured. This finding raised the question of how bacterial homologous recombination contributes to adaptive genome evolution. We addressed this question by simulations. We set the recombination rate in the simulations such that the decay of LD between neutral SNPs in simulated data is consistent with the observation in *S. aureus*. We found that homologous recombination plays significant roles in accelerating the rate of adaptive mutations and pursing deleterious mutations (**Fig. 5**). Because we suspect that *S. aureus* has quite a lower recombination rate in comparison with other bacterial species (**Fig. 1, 4**), it is expected that other bacterial species might take more advantages of homologous recombination. Thus, our results indicate that homologous recombination has significant contribution to adaptive genome evolution. It has been proposed that there are at least three potential advantages of bacterial homologous recombination: (i) Homologous recombina-

tion could repair damaged DNA[??]. (ii) DNA molecules that are taken up by bacterial cells could be used as a nutrient[?]. (iii) Homologous recombination could increase the efficacy of selection on linked sites[?]. We emphasize the third role, which is an underappreciated role so far[??].

METHODS

Methods and any associated references are available in a separate pdf file.

ACKNOWLEDGMENTS

This work is primarily supported by NIH and NSF grants to L.N. and H.I. S.T. and R.P.S. are research fellows of the Japan Society for the Promotion of Science (JSPS).

AUTHOR CONTRIBUTIONS

H.I. and L.N. designed this work. S.T., T.K. and R.P.S analyzed data and T.K performed simulations. S.T., L.N. and H.I. wrote the paper.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

1. Shapira, R., Levy, T., Shaked, S., Fridman, E. & David, L. Extensive heterosis in growth of yeast hybrids is explained by a combination of genetic models. *Heredity* **113**, 1–11 (2014).

2. Shull, G. H. The composition of a field of maize. *Journal of Heredity* **1**, 296–301 (1908).

3. Gama, L. T. *et al.* Heterosis for meat quality and fatty acid profiles in crosses among Bos indicus and Bos taurus finished on pasture or grain. *Meat Science* **93**, 98–104 (2013).

4. Birchler, J. A., Auger, D. L. & Riddle, N. C. In search of the molecular basis of heterosis. *The Plant Cell* **15**, 2236–2239 (2003).

5. East, E. M. Heterosis. *Genetics* **21**, 375 (1936).

6. Schwartz, D. Single gene heterosis for alcohol dehydrogenase in maize: the nature of the subunit interaction. *Theoretical and Applied Genetics* **43**, 117–120 (1973).

7. Krieger, U., Lippman, Z. B. & Zamir, D. The flowering gene single flower truss drives heterosis for yield in tomato. *Nature genetics* **42**, 459–463 (2010).

8. Minvielle, F. Dominance is not necessary for heterosis: a two-locus model. *Genetical research* **49**, 245–247 (1987).

9. Schnell, F. & Cockerham, C. Multiplicative vs. arbitrary gene action in heterosis. *Genetics* **131**, 461–469 (1992).

10. Charlesworth, D. & Willis, J. H. The genetics of inbreeding depression. *Nature reviews. Genetics* **10**, 783–96 (2009).

11. Xiao, J., Li, J., Yuan, L. & Tanksley, S. D. Dominance is the major genetic basis of heterosis in rice as revealed by qtl analysis using molecular markers. *Genetics* **140**, 745–754 (1995).

12. Frascaroli, E. *et al.* Classical genetic and quantitative trait loci analyses of heterosis in a maize hybrid between two elite inbred lines. *Genetics* **176**, 625–644 (2007).