Published online 21 January 2022 *Nucleic Acids Research, 2022, Vol. 50, No. 11* ***6211–6223***
https://doi.org/10.1093/nar/gkac025

Protoplast fusion in *Bacillus* species produces frequent, unbiased, genome-wide homologous recombination

Delyana P. Vasileva[1,2,†], Jared C. Streich[1,2,†], Leah H. Burdick[1,†], Dawn M. Klingeman[1], Hari B. Chhetri[1,2], Christa M. Brelsford[3], J. Christopher Ellis[1,2], Dan M. Close[1], Daniel A. Jacobson[1,2,*] and Joshua K. Michener[1,2,*]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA, [2]Center for Bioenergy Innovation, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA and [3]Geospatial Science and Human Security Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA

Received October 05, 2021; Revised January 05, 2022; Editorial Decision January 05, 2022; Accepted January 07, 2022

ABSTRACT

In eukaryotes, fine-scale maps of meiotic recombination events have greatly advanced our understanding of the factors that affect genomic variation patterns and evolution of traits. However, in bacteria that lack natural systems for sexual reproduction, unbiased characterization of recombination landscapes has remained challenging due to variable rates of genetic exchange and influence of natural selection. Here, to overcome these limitations and to gain a genome-wide view on recombination, we crossed *Bacillus* strains with different genetic distances using protoplast fusion. The offspring displayed complex inheritance patterns with one of the parents consistently contributing the major part of the chromosome backbone and multiple unselected fragments originating from the second parent. Our results demonstrate that this bias was in part due to the action of restriction–modification systems, whereas genome features like GC content and local nucleotide identity did not affect distribution of recombination events around the chromosome. Furthermore, we found that recombination occurred uniformly across the genome without concentration into hotspots. Notably, our results show that species-level genetic distance did not affect genome-wide recombination. This study provides a new insight into the dynamics of recombination in bacteria and a platform for studying recombination patterns in diverse bacterial species.

INTRODUCTION

Homologous recombination in the form of uptake and integration of DNA from exogenous sources has played a profound role in shaping microbial evolution and speciation (1). However, genetic transfer and recombination are rare in natural bacterial populations and thus difficult to characterize in detail. While a number of computational methods have been developed to estimate the relative rates and distribution of recombination events based on genome sequences of extant bacteria (2), these analyses are confounded by historical selection on recombinant strains. Direct measurements of recombination rates on a genome-wide scale are technically challenging because recombination patterns can be significantly affected by efficiencies and mechanistic specificities of DNA transfer. To date, most experimental estimates of recombination parameters have been conducted by transformation of naturally competent bacteria (3–7). Several works have characterized the recombination landscapes of chimeric transconjugant genomes generated by exchange of multiple large chromosomal fragments between bacteria through some unconventional conjugal mechanisms like mycobacterial distributive conjugal transfer and mycoplasma chromosomal transfer (8–12). A recent study has also determined the genomic signatures of Hfr (high frequency of recombination)-mediated chromosomal transfer in interspecies hybrids between *Escherichia coli* and *Salmonella enterica* (13). Although these studies have provided an invaluable insight into the genetics of recombination, we still do not fully understand how features of the genomic environment affect intensity of recombination around the chromosome.

Recombination on a genome-wide scale in bacteria can be achieved by protoplast fusion (14). In this classical

[*]To whom correspondence should be addressed. Tel: +1 865 576 7957; Fax: +1 865 576 8646; Email: michenerjk@ornl.gov
Correspondence may also be addressed to Daniel A. Jacobson. Tel: +1 865 574 6134; Fax: +1 865 241 2869; Email: jacobsonda@ornl.gov
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.
Present address: Dan M. Close, 490 Biotech, Inc., Knoxville, TN 37996, USA.

genetic engineering method, bacterial cells are stripped of their outer layer and chemically fused together, allowing recombination between the parental chromosomes. Originally used for routine genetic manipulation, protoplast fusion has been widely adopted as a strategy to generate microorganisms with improved phenotypes for biotechnological applications by combining beneficial alleles from different strains and even species (15,16). For instance, combinatorial shuffling of complete genomes by recursive fusion of protoplast populations has been employed to engineer multigenic traits for which the underlying molecular mechanisms are poorly understood, such as tolerance to stress conditions and production of diverse metabolites (17–19). Multiple crossover events are generally assumed to occur across the entire genome during this process, giving rise to mosaic chromosomes with unique phenotypic potential, analogous to meiotic recombination products in sexually reproducing organisms. Surprisingly, the exact nature of the chromosomal rearrangements resulting from large-scale shuffling experiments has received little attention and to date there are few studies reporting detailed analyses of sequenced bacterial shuffled genomes (20–22). Furthermore, due to strong selective pressure for the desired phenotypes, these analyses could not capture the full extent of recombination occurring between the parental chromosomes in protoplast fusants.

Mosaic genomes generated by DNA shuffling provide a unique source to investigate the genomics of recombination. In this work, we generated recombinant progeny from protoplast fusion between pairs of *Bacillus* strains with varying degrees of nucleotide identity. We built fine-scale recombination maps using next-generation sequencing and developed a computational pipeline to gain a deeper insight into how genomic sequence parameters affect dynamics of recombination events. Our results revealed that protoplast fusion generates multiple recombination events distributed across the genome with bias toward one of the parents and no other regional biases. While core features of homologous recombination decrease with increasing genetic distance (23,24), in our study we showed that the genome-wide outcomes were unaffected by large differences in nucleotide identity between parental strains. This work might aid in a better understanding of bacterial evolution in natural systems as well as provide potential insights into the use of genome shuffling for improving cellular function.

MATERIALS AND METHODS

Strains and media

Bacillus strains used in this work are summarized in Table 1. *Bacillus subtilis* subsp. *subtilis* 168 (25–27), *B. subtilis* subsp. *subtilis* RO-NN-1 (28,29), *B. subtilis* subsp. *subtilis* NCIB3610 (26,30), *B. spizizenii* TU-B-10 (28,31,32), and *B. velezensis* FZB42 (33,34), BKK34900 and BKE13180 (35) strains were obtained from the Bacillus Genetic Stock Center (BGSC). *Bacillus mojavensis* RO-H-1 (28,29) was obtained from the American Type Culture Collection. Parental cell lines were initially grown in low-salt LB medium (casein digest peptone 10 g/l, NaCl 5 g/l, yeast extract 5 g/l) with antibiotic selection as appropriate. All cells were grown at 37°C with liquid cultures kept at 250 rpm rotation. Antibiotics used include kanamycin sulfate (50 μg/ml) and erythromycin (20 μg/ml) used in concert with lincomycin (12.5 μg/ml). During the shuffling procedure, cells were washed and maintained in SMM buffer (36) consisting of 0.5 M sucrose, 20 mM $MgCl_2$ and 20 mM maleic acid. Polyethylene glycol (PEG) buffer to induce protoplast fusion consisted of SMM buffer supplemented with 35% PEG 6000 (Alfa Aesar, Heysham, UK) and 10 mM $CaCl_2$ (37). Newly shuffled cells were plated on DM3 recovery medium (38) without selection and subsequently plated on minimal medium (MM) (39) or LB agar plates. Media were supplemented with antibiotics as described earlier and with tryptophan (400 μM), histidine (300 μM) and methionine (1 mM) as needed for various auxotrophic strains.

Strain construction

All oligonucleotide primers used in this study for construction and verification of mutant strains are listed in Supplementary Table S1. The allele replacement constructs for generation of RO-NN-1 histidine and methionine auxotrophic strains were amplified from genomic DNA of strains BKK34900 (168 Δ*hisB*::*kan*) and BKE13180 (168 Δ*metE*::*erm*) (35). Strain RO-NN-1 was then transformed with the gene targeting fragments by natural competence, following standard protocols (35). Transformed strains were selected using LB medium containing the appropriate antibiotic.

A double mutant strain of RO-NN-1, containing both Δ*hisB*::*kan* and Δ*metE*::*erm*, was constructed by genome shuffling as described later.

The allele replacement constructs for deletion of the genes encoding restriction subunit of type I restriction–modification system (*hsdR*) and type IV restriction endonuclease (*mrr*) were generated by splicing by overlap extension PCR. The upstream and downstream regions of genes *hsdR* and *mrr* and an *erm* cassette flanked by loxP sites were amplified using genomic DNA from RO-NN-1 and BKE13180 (35), respectively. The joined PCR products were introduced into RO-NN-1 to generate strains RO-NN-1 Δ*hsdR*::*erm* and RO-NN-1 Δ*mrr*::*erm*. The *erm* cassette was subsequently removed by Cre recombinase expressed on pDR244 as described previously (35). RO-NN-1 Δ*hisB*::*kan* Δ*hsdR* and RO-NN-1 Δ*hisB*::*kan* Δ*mrr* strains were constructed in a second event of gene replacement.

All strains constructed in this study were verified by whole-genome resequencing. We note that our RO-NN-1 Δ*hisB*::*kan* Δ*metE*::*erm* parental strain contained six recombined fragments ranging from 10 to 1548 bp originating from *B. spizizenii* TU-B-10. These fragments were not present in the original strain obtained from BGSC. They were excluded from all further bioinformatic analyses.

Genome shuffling

Cells for genome shuffling were grown in selective liquid media overnight, and then diluted 100-fold the following morning. Once cultures reached an OD_{600} between 0.4 and 0.6, 5 ml was pelleted by centrifugation for 5 min at 8000 × *g* and 25°C and washed three times in 1 ml SMM buffer. DNase I (5 μg/ml) was added to the SMM buffer (SMMD)

Table 1. Strains used in this study

Strain	Genotype	Phenotype	References
B. subtilis subsp. *subtilis* 168	*trpC2*	trp^-	(25–27)
B. subtilis subsp. *subtilis* RO-NN-1	Wild type	Wild type	(28,29)
B. subtilis subsp. *subtilis* NCIB3610	Wild type	Wild type	(26,30)
B. spizizenii TU-B-10	Wild type	Wild type	(28,31,32)
B. mojavensis RO-H-1	Wild type	Wild type	(28,29)
B. velezensis FZB42	Wild type	Wild type	(33,34)
BKE13180	168 Δ*metE*::*erm*	trp^- met^- erm^R	(35)
BKK34900	168 Δ*hisB*::*kan*	trp^- his^- kan^R	(35)
JMB194	RO-NN-1 Δ*hisB*::*kan* Δ*mrr*	his^- kan^R	This work
JMB195	RO-NN-1 Δ*hisB*::*kan* Δ*hsdR*	his^- kan^R	This work
JMB1	RO-NN-1 Δ*hisB*::*kan*	his^- kan^R	This work
JMB3	RO-NN-1 Δ*metE*::*erm*	met^- erm^R	This work
JMB60	RO-NN-1 Δ*hisB*::*kan* Δ*metE*::*erm*	his^- met^- kan^R erm^R	This work
JMB12–JMB29 (Figure 1B/2A, blue)	168 Δ*metE*::*erm* × RO-NN-1 Δ*hisB*::*kan*	Wild type	This work
JMB6–JMB11, JMBP3A1–JMBP3A12 (Figure 1C/2A, orange)	168 Δ*metE*::*erm* × RO-NN-1 Δ*hisB*::*kan*	his^- met^- kan^R erm^R	This work
JMBP3C4–JMBP3D5 (Figure 1D/2A, pink)	168 Δ*hisB*::*kan* × RO-NN-1 Δ*metE*::*erm*	his^- met^- kan^R erm^R	This work
JMB219–JMB225 (Figure 3A)	168 Δ*metE*::*erm* × RO-NN-1 Δ*hisB*::*kan* Δ*hsdR*	his^+ met^+	This work
JMB204–JMB210 (Figure 3B)	168 Δ*metE*::*erm* × RO-NN-1 Δ*hisB*::*kan* Δ*mrr*	his^+ met^+	This work
JMB239–JMB240 (Figure 3C)	168 Δ*metE*::*erm* × RO-NN-1 Δ*hisB*::*kan* Δ*mrr*	his^- met^- kan^R erm^R	This work
JMBP4D1–JMBP4E4 (Supplementary Figure S5, dark blue)	RO-NN-1 Δ*hisB*::*kan* Δ*metE*::*erm* × NCIB3610	his^- kan^R	This work
JMBP4E5–JMBP4F8 (Supplementary Figure S5, light blue)	RO-NN-1 Δ*hisB*::*kan* Δ*metE*::*erm* × NCIB3610	met^- erm^R	This work
JMB61–JMB75 (Supplementary Figure S5, red)	RO-NN-1 Δ*hisB*::*kan* Δ*metE*::*erm* × TU-B-10	his^- kan^R	This work
JMB76–JMB89 (Supplementary Figure S5, yellow)	RO-NN-1 Δ*hisB*::*kan* Δ*metE*::*erm* × TU-B-10	met^- erm^R	This work
JMB135, JMB138, JMB147, JMB148, JMB154, JMB155, JMB226–JMB238 (Supplementary Figure S5, green)	RO-NN-1 Δ*hisB*::*kan* Δ*metE*::*erm* × RO-H-1	met^- erm^R	This work

after initial wash steps to prevent natural transformation via DNA transport between cells. Complete protoplast formation was accomplished by resuspending washed cells in 1 ml SMM buffer with 1 mg/ml lysozyme followed by incubation at 37°C for 1 h. Five hundred microliters from each parental cell line was mixed together after protoplasting and centrifuged for 20 min at 2000 × *g* at 12°C. These mixed pools were washed once in SMMD buffer, resuspended in 400 μl PEG buffer and incubated at room temperature for 20 min. Cells were again washed in SMMD buffer and resuspended in 100 μl SMMD buffer with 1% bovine serum albumin added. The protoplast suspension was then spotted on DM3 regeneration media without selection and incubated overnight at 37°C. The following day, lawns of regenerated cells were scraped from the regeneration plates, suspended in MM and plated to selective media for isolation of recombinant strains. The parental cell lines showed similar protoplast regeneration efficiencies with DM3 regeneration medium: *B. subtilis* subsp. *subtilis* RO-NN-1 (5.2×10^{-1}), *B. subtilis* subsp. *subtilis* NCIB3610 (4.8×10^{-1}), *B. spizizenii* TU-B-10 (2.0×10^{-1}) and *B. mojavensis* RO-H-1 (4.0×10^{-1}).

Strain isolation and sequencing

Individual strains were isolated either by plating serial dilutions or by streaking to individual colonies on selective media. Single colonies were then picked and restreaked to selective plates before being grown to saturation in selective liquid media. Genomic DNA was isolated using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. DNA for PacBio sequencing was isolated using the same method, but multiple samples were combined and concentrated to obtain higher concentrations. To achieve this, 1/10 combined sample volume of 3 M sodium acetate was added to pooled DNA, followed by 2.5 volumes of 100% ethanol. This was mixed and incubated at −80°C for 30 min. Precipitated DNA was then pelleted by centrifugation at 14 000 rpm for 20 min at 4°C, washed with 70% ethanol and allowed to air dry. DNA was then resuspended in 1/10 TE buffer and stored at −20°C until being shipped on dry ice to the University of Maryland for PacBio sequencing.

For strain resequencing, Nextera XT libraries (Illumina, San Diego, CA) were generated from purified DNA of isolated strains according to the manufacturer's protocol (15031942 v05), stopping after library validation. Final libraries were validated on an Agilent Bioanalyzer (Agilent, Santa Clara, CA) using a DNA7500 chip and concentration was determined on an Invitrogen Qubit (Waltham, MA) with the broad range double-stranded DNA assay. Barcoded libraries were pooled and prepared for sequencing following the manufacturer's recommended protocol (15039740 v10, Standard Normalization). One paired-end sequencing run (2 × 301) was completed on an Illumina MiSeq instrument (Illumina, San Diego, CA) using v3 chemistry. Illumina resequencing of strains generated by crossing RO-NN-1Δ*hisB*::*kan* Δ*metE*::*erm* with wild-type isolates was performed commercially (SNPsaurus, Eugene, OR). PacBio sequencing was performed by the University of Maryland Institute for Genome Sciences (Baltimore, MD).

Variant calling

The average sequencing coverage was 80–100×, which provided high confidence in calling nucleotide changes. Fastq files from sequencing were first processed with Trimmomatic for phred base pair quality. Reads that lost a paired read from phred filtering were removed. Reads that were shorter than 38 bp were removed to reduce the quantity of nonuniquely mapping reads. Individuals were independently run through a variant calling pipeline using software current at the time the project started: BWA v0.7.17, SAMtools v1.8, Picard v2.20.8, GATK v3.8.0, VCFtools v0.1.15, BCFtools v1.9, PLINK v1.9.0 and in-house R scripts (40–45). Reads were aligned through BWA MEM to generate .sam files (Sam files). SAMtools was then used to create compressed .bam files (Bam files) for further processing. Bam files were then parsed by SAMtools for uniquely mapping reads to a single locus, while multi-loci mapping reads were removed. SAMtools was next used to order reads by their individual genome mapping coordinate and their read groups replaced. After removing non-mappable reads, and remaining reads ordered and properly annotated, Bam files were scanned for duplicate calls with SAMtools and then were indexed via Picard. Polished Bam file reads were run through GATK HaplotypeCaller as haploids with '-ploidy 1'. BCFtools was used to filter low-coverage variants, requiring a minimum read depth of 12 to confirm the variant. GATK's HaplotypeCaller function will only annotate the most common variant in haploid organisms, and since sequencing errors are rare, only variants with several reads ($\geq$20) are marked in VCF files. Variants were also BCFtools filtered for a genotype quality of $P < 0.1 \times 10^{-6}$ to ensure the chance of a false variant was <1:100 000. A random subset of individuals was then scanned by eye to check for variants in low-coverage areas, that no low-coverage variants were marked and no biallelic states were present. Final bioinformatic analysis was done in R v3.5.0 using PLINK ped/map file format.

Genomes of RO-NN-1 Δ*hisB*::*kan* Δ*metE*::*erm* × RO-H-1 shuffled strains were too divergent for read mapping with BWA and were instead analyzed using Geneious Prime v. 11.1.3. Sequence reads were mapped to the RO-NN-1 and RO-H-1 reference sequences. Single-nucleotide polymorphisms (SNPs) were identified and boundaries of the genomic fragments inherited from each parent were then determined manually.

Analysis of clustering of recombination fragments

Genomic proximity between recombination segments in shuffled strains was assessed by using Moran's *I* test. Individual strains with <10 recombination events were excluded from the analysis. Distance matrix was created between the endpoints of the individual recombined regions in each shuffled strain. Then, using the *Moran.I* function of the *ape* package in R, the null hypothesis that there is no autocorrelation (clustering) of recombination fragments was tested. To assess the significance of the Moran's *I* test results, a 95% confidence interval based on the empirical distribution of the test statistic [−log(*P*-value)] was created by randomly permuting the dataset 999 times. Data points were randomly assigned to the insert positions to create each permutation set and Moran's *I* test was performed on each of those permuted datasets. In addition, to assess the overall significance of the clustering test across the shuffled populations, Fisher's exact test using the *fisher.test* function (with simulate.p.value = TRUE) in R was performed on the number of significant and nonsignificant Moran's *I* test results.

Genomic feature analysis

Parent detection and filtering. Within each shuffled population, variants were first called against each parent reference genome. However, in every recombinant population except one, the parent strain RO-NN-1 remained the major contributor to offspring genomes, and thus was used for all further variant calling and genome analysis. In each population, variants were encoded as '0' for RO-NN-1 and '2' for the minor parent. Variants called in both parents at a single position are likely sequencing errors that arose during laboratory processing or DNA sequencing. Markers not present in at least one offspring were also removed. Any variant found in one parent and one individual was kept for recombination and insertion analysis methods. Lists for differential variants between parents were used for permutation testing (described later).

Insert size and frequency. After variant encoding, insert size was calculated based on the number of base pairs between continuous variants from the minor parents. The positions and lengths of insertions from minor parents were calculated for each individual by counting continuous strings of nonreference variants and their distance in bases along the genome. Features of shuffled genomes were visualized by the R package 'BioCircos' and standard plotting libraries (46). The quantity of insertions per strain was similarly quantified by totaling each individual's strings of markers originating from the minor parents. The distributions were tested for normality using the Shapiro–Wilk tests. The means and standard deviations were compared with *t*-tests, *F*-tests, Wilcoxon tests and Kolmogorov–Smirnov tests for significance.

Population-level genome feature analysis. Read mapping statistics were calculated using VCFtools and the RO-NN-1 reference genome as the major parent (41). Read depth per shuffle was calculated using '–depth' for population-level read depth. Likewise, VCFtools function '–mean-depth' was used for broad read depth and '–site-depth' was used for variant sequence depth per individual in each shuffle. Site mean depth was calculated by VCFtools '–site-mean-depth' function to obtain per sample mean sequence depth.

Permutation testing against genome features. Recombined positions in the genome were examined against other extractable genomic features. IGV v2.3.5 was used for genomic feature extraction of a known methylation motif (GAYGNNNNNNCTT) and GC content (47). Additionally, known gene positions within the major parent RO-NN-1 were also used in testing variants involved in insertion detection. In each permutation test, a two-stage random number generator was used: the first seed number to create a list of random numbers that was then used to create a second set of random numbers each used once in a single iteration within tests. In each test, positions of genomic features were compared to randomly generated lists of genomic positions to test whether insertions between parents have statistical significance to SNPs/variants, methylation motifs or GC content. Each test was run against 1000 randomly generated subsets to create a P-value significance level of 0.0001. Iteration subsets of random test positions were based on the number of features detected. For instance, 1066 methylation motifs exist in the RO-NN-1 genome; thus, for each iteration, 1066 random positions were used.

Methylation to insertion testing. To investigate whether methylation sites are closer than random to insertion sites, we compared 'distance in base pairs from methylation motifs to random positions' to 'base pair distance of motifs to insertion sites'. A list of randomly generated genome positions was created to draw subsets per iteration equal to the number of insertion events per population. In each iteration, the distance to a methylation was calculated to a randomly drawn genome position to create a distribution of randomly drawn base pair lengths. Then, subsets of the 1066 known methylation motifs were drawn per iteration and base pair distance was calculated to the nearest 5′ or 3′ end of an insertion event (see Supplementary Figure S1).

Insertion events to random position testing. Insertion events could be biased toward specific positions within the genome. To test this, we generated two lists of random genome positions and calculated base pair distance between pairs of positions. For each random position in dataset 1, we determined the distance to the closest randomly drawn position in the second random set. Then, we randomly drew positions in the genome and calculated distance to the nearest 5′ or 3′ insertion event.

GC content permutation testing. Two similar tests for GC content correlation to insertion positions were implemented. One test examined unidirectional outward GC content away from insertion sites: from the 5′ insertion, then examining increasing windows beforehand (3′ to 5′) and the 3′ end of the insertion expanding forward (5′ to 3′). GC content was measured by percent GC at increasing increments through exponentially increasing windows of 2^n bases, $n = 2$–12 (from 2^2 to 2^{12}; 4–4096 bases). The same test was performed on randomly generated insertions, unique to each iteration, and the percent GC was calculated using the same exponential scan pattern as variants. To generate a list of random insertions with comparable insertion lengths, random markers were chosen from a list of known variant sites between the two parents as the 5′ end. To get a comparable 3′ marker as the insertion switch point, actual insertion sizes were randomly drawn and assigned to 5′ variants and the closest 3′ differential variant was chosen in either direction, thus creating the most similar possible insertion size to an observed insertion size. Generation of *in silico* variants required the use of the R package 'ecodist' (48). A very similar test was performed scanning GC content, but in both 5′ and 3′ directions from insertion ends (scanning away and into the insertion markers). A smaller set of windows was used since the chance of double counting GC content exists within the boundaries of *in silico* simulated insertions. When building simulated insertions, insertion sizes that were <1024 were removed. Thus, window sizes considered were 4^n, $n = 1$–4 (4–256 bases). Limiting the GC content scan within insertion sites to 256 bases means that up to 50% of the insertion site was scanned for %GC content.

Wavelet analysis for population features and a range of complexities. Wavelet transforms can analyze signal-based data by expanding two-dimensional data into three-dimensional space at varying scales to reveal otherwise cryptic patterns. The underlying theory of wavelet analysis is to overlay an organized specific wave of designated length and area over a signal series to find differences in area annotated as coefficients. Wavelets can find patterns or quantify 'how much of a peak' is present at a region of a signal that is not immediately obvious to the human eye, and scanned at varying scales/window sizes of data (49,50). Within this study, we implemented a continuous wavelet transform using the Ricker wavelet as the mother wavelet to identify regions of the genome with differing characteristics of recombinant loci and potential hot and cold locations across recombinant populations. Ricker wavelets are ideal for this scan type since they target one specific location relative only to immediate upstream and downstream signals; they are composed of three parts with a total area of zero (two negative peaks with area $= -0.5$ flanking a single positive peak with area $= 1$). Below is the wavelet transform that returns the wavelet coefficients $W(s, \tau)$ that are calculated across scales (s) and translation along the genome as τ (shifts across the x-axis) (51).

The resulting coefficients will indicate at specific scales the *quantity of peak* present. Wavelet analysis was performed using the R statistical programming language 3.5.0 (https://www.r-project.org/) and the 'wmtsa' package was used for wavelet transform analysis (52). Genomic data were encoded as '0' and '2' and only for relevant positions such binary transition states collapsed only to varying positions does not lend well to signal processing; thus, variant data were modified in two ways (50). First, all variant positions were summed across the population to a single vector and spread out to their actual position, where absence of a variant was annotated as a zero. Second, data were binned down to ∼4010 data points (100× reduction) depending on the genome marker positions of each population. Once the data were transformed to amenable wavelet analysis qualities, the locations with differing areas to the mean with either higher or lower than expected values were revealed.

RESULTS AND DISCUSSION

Genomic consequences of genome shuffling

To determine the genome-wide effects of protoplast fusion, we crossed *B. subtilis* 168 and RO-NN-1. These strains are in the same subspecies and have ∼98% average nucleotide identity (ANI) in shared genes (53). Successful genome shuffling is typically assessed through simultaneous selection for markers present in both parents. To make this strategy more flexible, we replaced biosynthetic genes that are essential for growth in MM with antibiotic resistance markers (Figure 1A). This approach allows selection for any of four potential allele combinations. We chose *hisB* and *metE* as biosynthetic genes, since these gene deletions produce known auxotrophies (35) and the genes are roughly opposite in the genome, separated by 2.2 and 2.0 Mb.

We performed two reciprocal crosses of 168 Δ*hisB*::*kan* (‘168 HK’) × RO-NN-1 Δ*metE*::*erm* (‘RO-NN-1 ME’) and RO-NN-1 Δ*hisB*::*kan* (‘RO-NN-1 HK’) × 168 Δ*metE*::*erm* (‘168 ME’). We then selected recombinant strains containing either both mutant alleles (Δ*hisB*::*kan* Δ*metE*::*erm*, ‘DR’) or both wild-type alleles (*his*$^+$ *met*$^+$, ‘WT’). Eighteen recombinant strains from each combination of shuffle and selection were isolated and resequenced using short-read sequencing. Regrettably, the 168 HK × RO-NN-1 ME prototrophic pool was contaminated by other prototrophic isolates and therefore was not analyzed further. To identify large-scale genome rearrangements, we also sequenced two parental and four recombinant strains using long-read sequencing. The genome sequences of the recombinant strains were then analyzed computationally to determine the genetic contributions from each parent.

Sequencing results revealed a strong asymmetry in recombination, with one of the parental strains (RO-NN-1 ME or RO-NN-1 HK) contributing the majority of the chromosome of every progeny (Figure 1B–D). All recombinant strains carried the selected marker flanked by different amounts of DNA, ranging from 1 to 76 kb, that originated from the second parent (168 HK or 168 ME). In addition, we detected extensive unselected variation across their genomes with multiple unrelated regions of recombination. Within a single strain, these unselected recombined regions were not distributed evenly around the chromosome, but instead tended to be clustered in tracts within a relatively small region of the genome (Figure 1B–D, Supplementary Figures S2 and S3, and Supplementary Table S2). Similar mosaic inheritance patterns have been observed after conjugation and natural transformation in *B. subtilis* and several other bacterial species (3,4,8,12,54). The precise mechanism creating these genomic signatures is unclear, but it has previously been suggested that localized clustering of recombination events might result from noncontiguous integration of a single imported large donor DNA fragment (3). In this work, however, the entire chromosomes of RO-NN-1 and 168 were brought together by protoplast fusion, suggesting that frequent local recombination is not an artifact of DNA transport limitations.

Interestingly, genomes of the DR progeny generated by fusion of 168 ME and RO-NN-1 HK protoplasts showed lower levels of complexity compared to the other two progeny populations (Figures 1B–D and 2A, and Supplementary Tables S3 and S4). Individual 168 ME × RO-NN-1 HK WT (blue) and 168 HK × RO-NN-1 ME DR (pink) recombinant chromosomes contained a median of 23 and 21 separate 168-derived genome segments, respectively, while 168 ME × RO-NN-1 HK DR (orange) genomes contained a median of 1 fragment originating from 168 ME (Figure 2A). The two sets of 168 ME × RO-NN-1 HK progeny (blue and orange) result from the same fusion and regeneration process, simply plated on different selective media. DR progeny were selected on LB with both antibiotics, while WT progeny were selected on MM. We could not detect significant differences in fitness between any of the parental strains under either growth condition. Thus, the genetic background details of each cross affect the recombination process in subtle but important ways and further analyses will be required to identify the factors that have promoted enrichment of recombinant populations with different levels of heterogeneity.

Recombination fragment sizes in the three progeny populations were broadly distributed, including short inserts and large (>60 kb) multigene fragments (Figure 2B). On average, fragments from strain 168 replaced 4.5% (168 ME × RO-NN-1 HK WT), 0.6% (168 ME × RO-NN-1 HK DR) and 3.0% (168 ME × RO-NN-1 HK DR) of the chromosome of RO-NN-1 (Supplementary Table S3). Long-read sequencing of four progeny strains did not identify any large chromosomal rearrangements. Selection ensured that one marker from the 168 parent was necessarily present in the recombinant progeny, and this marker was often integrated as part of a >4 kb segment. The high frequency of recombinant segments of ∼4 kb (Figure 2B, red bars) is partly due to this bias.

Recombination can only be detected when it results in a genetic change. The chromosomes of strains 168 and RO-NN-1 have an average of one genetic variant approximately every 50 bp, which allows identification of parental genomic contributions at a similar resolution. Small recombination events, particularly for closely related strains such as these, may not result in any genetic change and will go undetected. Therefore, these measurements provide a lower bound on the recombination rate. Similarly, we chose to calculate the minimum insert size, based on the shortest length between recombined variants (Supplementary Figure S2). The true length of exchanged DNA may, in some cases, be several fold longer. The frequent single base recombination events likely result from recombination of larger physical DNA fragments (up to ∼1 kb in length) that only alter a single nucleotide (Figure 2B and Supplementary Figure S4), while a substantial number of small recombination events are missed entirely because they do not introduce any nucleotide changes. Finally, we cannot rule out that some genetic changes may be the result of two or more adjacent recombination events, and it is difficult to precisely confirm the directionality of frequent fine-scale recombination events. In all cases, we have chosen the most parsimonious explanation.

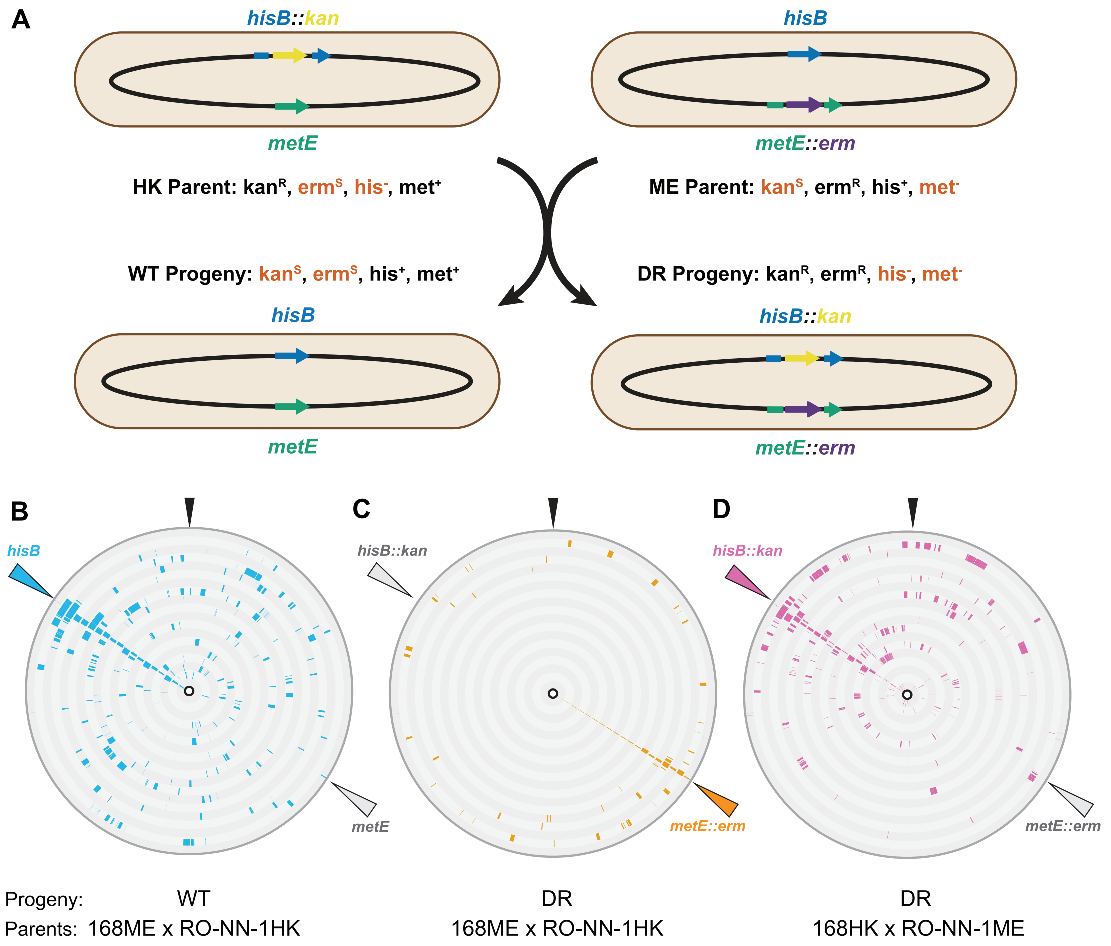
A
hisB::kan
metE
hisB
metE::erm
HK Parent: kanR, ermS, his-, met+
ME Parent: kanS, ermR, his+, met-
WT Progeny: kanS, ermS, his+, met+
DR Progeny: kanR, ermR, his-, met-
hisB
metE
hisB::kan
metE::erm
B
C
D
hisB
metE
hisB::kan
metE::erm
hisB::kan
metE::erm
Progeny: WT
Parents: 168ME x RO-NN-1HK
DR
168ME x RO-NN-1HK
DR
168HK x RO-NN-1ME

Figure 1. Analysis of genome shuffling in *B. subtilis*. (**A**) Replacing amino acid biosynthesis genes with antibiotic resistance markers allows flexible identification of recombinant progeny following genome shuffling. Crossing mutants of 168 and RO-NN-1 yielded prototrophic (**B**) and double-resistant (**C**, **D**) progeny. Each concentric circle represents a different resequenced individual from this cross. The colored bars indicate sequences recombined from strain 168, with the remaining genomic sequence coming from RO-NN-1. Orange, blue, pink and gray arrows indicate locations of selection markers. Black arrows indicate the origin of replication. WT: wild type; DR: double resistant.

Identifying effects of genomic features on recombination

We next sought to identify genomic properties that might have influenced recombination. We hypothesized that differential methylation patterns in the two parents might bias recombination directionality and localization. Methylation analysis of our PacBio sequencing data confirmed the known GAGGAC methylation motif in strain 168 (55) and identified a different motif, AAGNNNNNNCRTC, in RO-NN-1. The DnmA methyltransferase in strain 168 is not associated with a cognate restriction enzyme, and methylation in this strain is instead thought to influence transcriptional regulation (55). Conversely, strain RO-NN-1 encodes both a putative type I restriction enzyme targeting unmethylated DNA and a putative type IV restriction enzyme targeting methylated DNA. We therefore hypothesized that biased inheritance of RO-NN-1 genomic DNA in recombinant progeny is due to asymmetric enzymatic cleavage of the 168 chromosome in fused protoplasts. To test this hypothesis, we constructed two RO-NN-1 mutant strains lacking the individual type I (HsdR) and type IV (Mrr) restriction machineries and evaluated the effect on genome-wide recombination during protoplast fusion (Figure 3). We crossed RO-NN-1 HK Δ*hsdR* and RO-NN-1 HK Δ*mrr* strains with 168 ME and selected for prototrophic and double-resistant progeny from each cross. Protoplast fusion between RO-NN-1 HK Δ*hsdR* and 168 ME produced viable prototrophic progeny but failed to yield double-resistant progeny despite multiple attempts. In the prototrophic progeny, inactivation of HsdR did not affect the outcome of the genome shuffling experiments and strain RO-NN-1 still contributed the major portion of the offspring chromosomes (Figure 3A). Interestingly, protoplast fusion between RO-NN-1 HK Δ*mrr* and 168 ME produced prototrophic progeny that displayed a 168 backbone with multiple incorporated RO-NN-1-derived recombined segments and double-resistant progeny with most of the genomic DNA originating from RO-NN-1 (Figure 3B and C). These results suggested that the RO-NN-1 type IV restriction–modification system affects the directionality of recombination in fused protoplasts, but other factors in both parental strains are involved in this process.

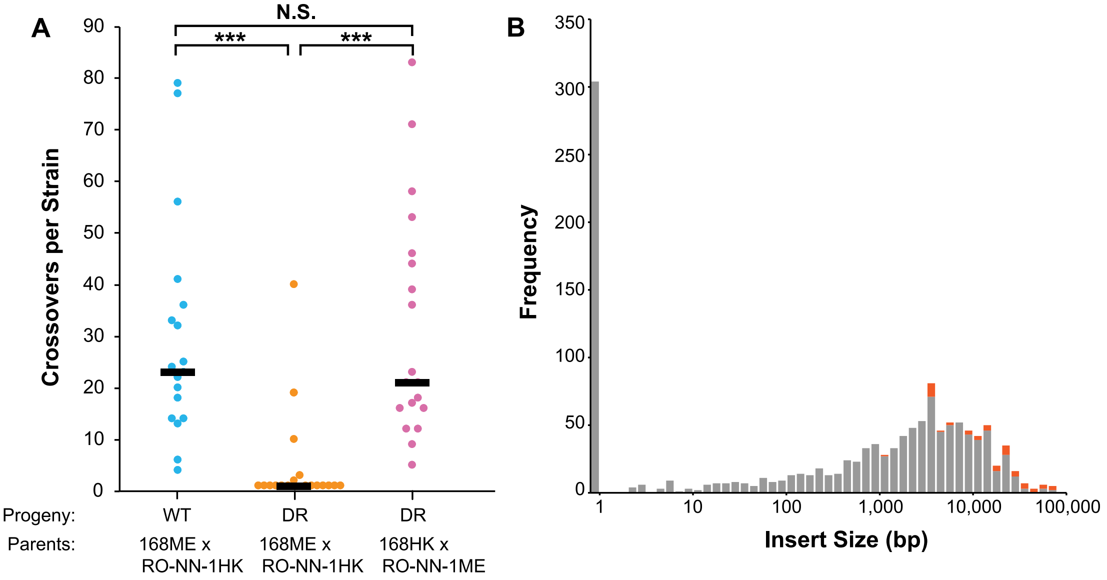

Figure 2. Analysis of recombination frequency and size. (**A**) The number of recombination events was calculated for each strain in a given pool, representing each strain by a single data point. The median value for each pool is shown with a black line. Shapiro–Wilk tests were used to test distribution normality of number of recombination events showing that each population had non-normally distributed data (Supplementary Table S3). *t*-tests, *F*-tests, Kolmogorov–Smirnov tests and Wilcoxon tests were used to compare variances and distribution means between the three populations (Supplementary Table S4). (**B**) The distribution of recombination fragment lengths is shown for all three populations combined. Fragments containing the selection marker are indicated in red. WT: wild type; DR: double resistant. ***$P < 0.001$; N.S., not significant.

To investigate more subtle influences on recombination, we examined the correlation between local recombination frequencies and several features of the genomic context, including proximity to methylation sites, local GC content and SNP density (Figure 4). We performed these analyses using all recombination regions in the 168 ME × RO-NN-1 HK prototrophic progeny where we identified extensive genetic variability (Figure 1B). First, we hypothesized that double-strand breaks caused by DNA restriction might trigger increased homologous recombination near the restriction site. To test this hypothesis, we calculated distances between the boundaries of the recombination segments and the nearest methylation sites. Comparison of experimental data to the equivalent measurement for randomly permuted recombination regions showed no significant differences (Figure 4A). Thus, the methylation landscapes of the parental strains seem to affect directionality of recombination, but they do not appear to determine the endpoints of recombination events around the chromosomes of the offspring.

Next, we analyzed GC richness of the homologous regions surrounding the recombination segments in the shuffled genomes and compared these results to GC content of regions flanking randomly permuted recombination sites. We did not reveal any bias in recombination frequency toward GC content for the tested window sizes (2–512 bp, increasing 2^n; Figure 4B shows a 256-bp window) (Figure 4B). These findings suggested that, at a fine scale, this aspect of the genomic environment might not be relevant for distribution of recombination events across the chromosome.

To evaluate the correlation between local sequence similarity and positioning of recombination events around the chromosome, we next estimated SNP densities in regions flanking the recombination segments, in a similar fashion to the analysis of GC content. Rates of transformation by homologous recombination decrease exponentially as a function of local sequence divergence (23,56,57). As a result, we hypothesized that recombination would occur more frequently in regions of high local sequence identity. However, the local SNP frequency near recombination junctions did not differ significantly from random chromosomal locations in the tested window sizes (between 2 and 512 bp; 256 bp shown in Figure 4C). In fact, fewer recombination events at regions of very high local sequence identity were observed than would be predicted by chance, likely because these recombination events rarely produced changes in nucleotide sequence. Similar to GC content, local SNP frequency does not appear to cause a significant bias in genome-wide distribution of recombination events.

Patterns of recombination have been extensively investigated in eukaryotes. Distribution of meiotic recombination events across eukaryotic chromosomes is nonrandom. It has become clear that recombination predominantly occurs in specific regions of the genome known as recombination hotspots, the localization of which is dictated by binding of the histone methyltransferase PRDM9 to specific DNA motifs (58–60). In bacteria, recombination is less organized and knowledge about the intensity of crossover events around the chromosome is limited. Since the genomic position and relevant length scale of a potential recombination hotspot are not known *a priori*, we used continuous wavelet transform analysis to simultaneously analyze average recombination frequencies of all potential positions and lengths (Figure 4D). We identified the known hotspot at the selection marker from the minor parent (*hisB*) but could not detect any other biases. Wavelet analysis could also poten-

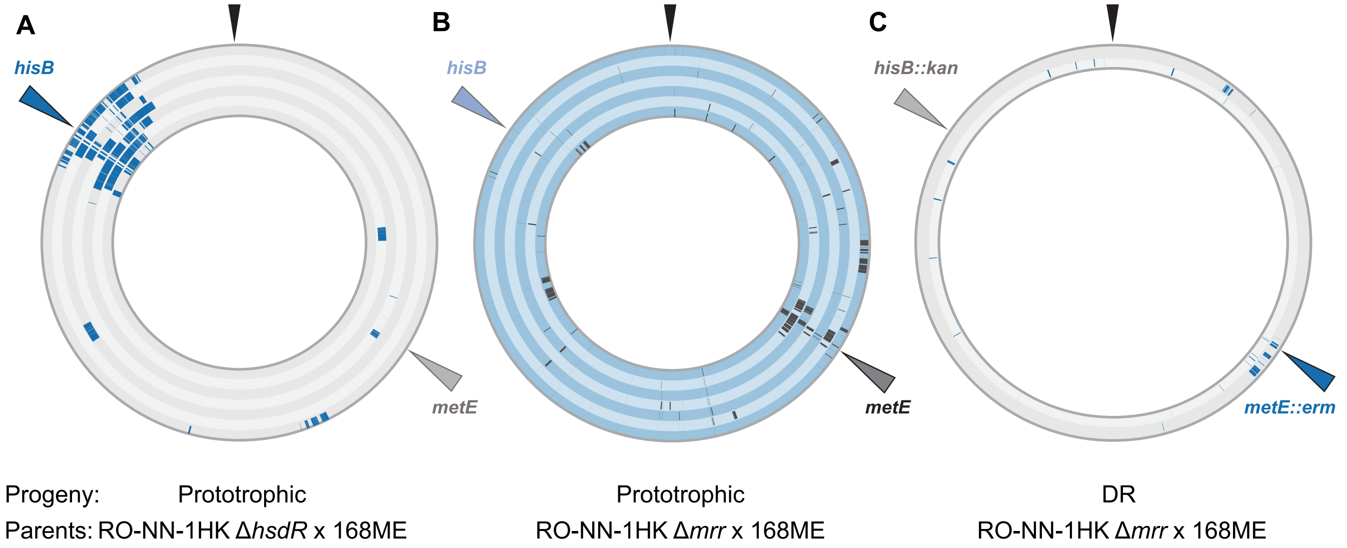

Figure 3. Restriction affects directionality of recombination in fused protoplasts. *Bacillus subtilis* RO-NN-1 ΔhisB::kan strains lacking the individual type I (HsdR) and type IV (Mrr) restriction machineries were crossed with *B. subtilis* 168 ΔmetE::erm. Prototrophic and double-resistant progeny were selected. (**A**) RO-NN-1 ΔhisB::kan ΔhsdR × 168 ΔmetE::erm prototrophic progeny displayed an RO-NN-1 backbone. (**B**, **C**) Protoplast fusion between RO-NN-1 ΔhisB::kan Δmrr and 168 ΔmetE::erm produced prototrophic progeny with a 168 backbone and double-resistant progeny with an RO-NN-1 backbone. Concentric circles represent resequenced individuals from each cross. In RO-NN-1 ΔhisB::kan ΔhsdR × 168 ΔmetE::erm prototrophic and RO-NN-1 ΔhisB::kan Δmrr × 168 ΔmetE::erm double-resistant strains, blue bars indicate recombined regions originating from strain 168, with the remaining of the genome sequences coming from RO-NN-1. In RO-NN-1 ΔhisB::kan Δmrr × 168 ΔmetE::erm prototrophic strains, gray bars indicate recombined regions originating from strain RO-NN-1, with the rest of the genome coming from strain 168. Dark blue and dark gray arrows show location of the selection markers. Black arrows indicate origin of replication. DR: double resistant.

tially detect regions with lower recombination frequencies than expected by chance. One known cold spot is located at the other selection marker (*metE*), but this location was not identified in our analysis. A higher average recombination frequency would be necessary to accurately detect cold spots.

Taken together, our analyses of the impact of genomic context on recombination rates did not reveal positive or negative associations. Other factors might also affect distribution of crossover events across the genome. Bacterial chromosomal DNA is organized into a compact structure called nucleoid by the cooperative action of DNA supercoiling and nucleoid-associated proteins (61). Several lines of evidence suggested that recombination could be affected by chromosome architecture. For example, analysis of site-specific recombination between regions scattered over the chromosome in *E. coli* demonstrated that intramolecular recombination between different nucleoid macrodomains is highly restricted (62). The nucleoid in *B. subtilis* is organized into three distinct topological domains (63). Thus, some regions of the parental genomes might be randomly and temporarily more accessible for recombination in protoplast fusants, which could explain the higher frequencies of local crossover events. Future investigation of the effect of DNA topology on recombination using the computational methods developed in this study might provide a mechanistic insight into genome-wide recombination patterns in bacteria.

Effects of genetic distance on recombination

Genetic distance is an important physical factor that affects rates of recombination. To better understand the role of nucleotide identity in recombination parameters, we shuffled RO-NN-1 Δ*hisB*::*kan* Δ*metE*::*erm* with wild-type strains *B. subtilis* subsp. *subtilis* NCIB3610 (the parent of strain 168, with 98% ANI), *B. spizizenii* TU-B-10 (93% ANI) and *B. mojavensis* RO-H-1 (87% ANI). We were unable to generate recombinants using *B. velezensis* FZB42 (78% ANI). In each successful example, we selected for strains from both potential recombinant genotypes, either Δ*hisB*::*kan metE*$^+$ or Δ*metE*::*erm hisB*$^+$ (Supplementary Figure S5). After several attempts, we failed to isolate viable Δ*hisB*::*kan metE*$^+$ recombinants from the RO-NN-1 Δ*hisB*::*kan* Δ*metE*::*erm* × RO-H-1 cross, which we hypothesize was due to genetic incompatibility. Previous studies of natural transformation in *B. subtilis* demonstrated that frequency of recombination decreases exponentially as a function of genetic divergence, likely because of inefficient pairing between donor and recipient DNA sequences with increasing number of nucleotide mismatches (23). Moreover, detailed analyses of recombination at the *rpoB* locus indicated that the size of recombined regions decreases significantly with increasing genetic divergence (24). Our sequencing data showed no correlation between genetic distance and number of recombination events per strain (Figure 5A, Supplementary Figure S5 and Supplementary Table S3). Furthermore, we did not detect significant differences in size distribution of the recombined segments (Figure 5B). These fragments were clustered in tracts along the genomes of the recombinant strains and their localization was not affected by proximity to methylation sites, local GC content and nucleotide identity (Supplementary Figures S3, S5 and S6). In addition, except for the selection markers, we did not detect recombination hotspots in any

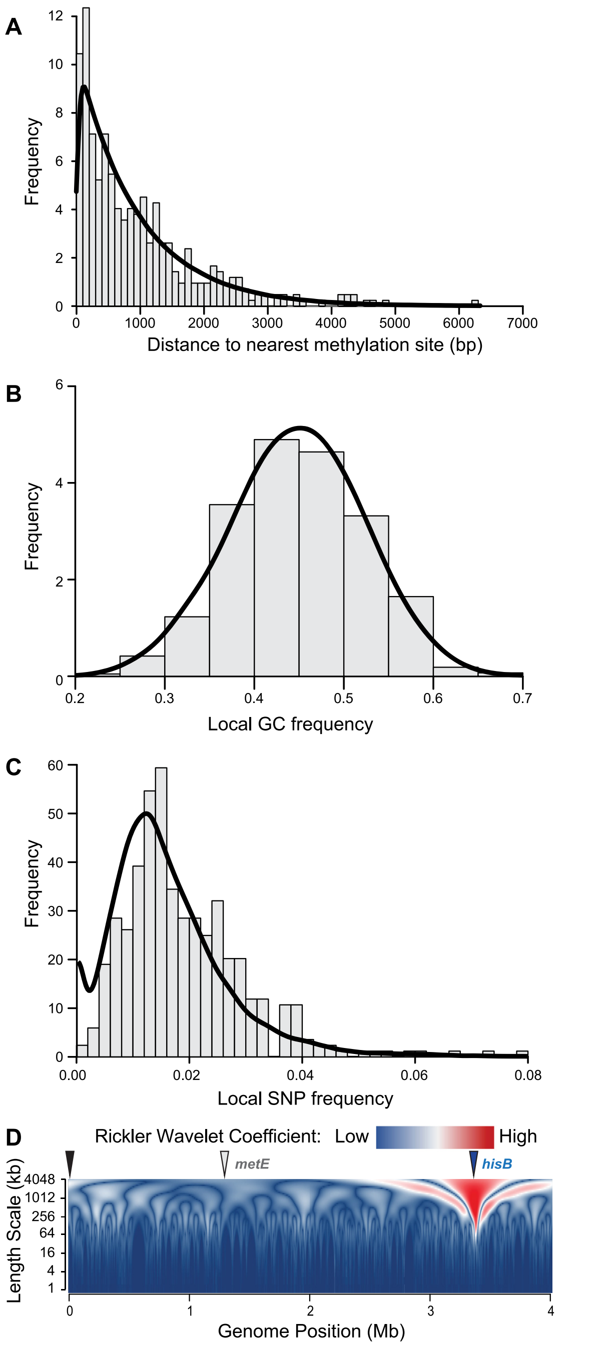

Figure 4. Genomic features do not affect distribution of recombination events around the chromosome. (**A**–**C**) Genome properties were calculated for the complete set of 560 recombination sites in 168 ME × RO-NN-1 HK prototrophic progeny (gray histograms) and equivalent randomly permuted recombination sites (black lines). Features analyzed are (**A**) distance between the boundary of a recombination site and the nearest methylation site, (**B**) GC frequency in a 256-bp window spanning the recombination boundary and (**C**) SNP frequency in the same 256-bp window. Differences between actual and permuted distributions were not significant. (**D**) Population-level recombination was analyzed across the genome using a continuous wavelet transform analysis with Ricker wavelets. The wavelet coefficient is plotted for each combination of genomic position and length scale. High wavelet coefficients indicate deviations from the baseline at a particular combination of position and length scale. Genomic positions of the selection markers are indicated; this population selected for recombination at the *hisB* marker and against recombination at the *metE* marker. Only the recombination hotspot at *hisB* is evident.

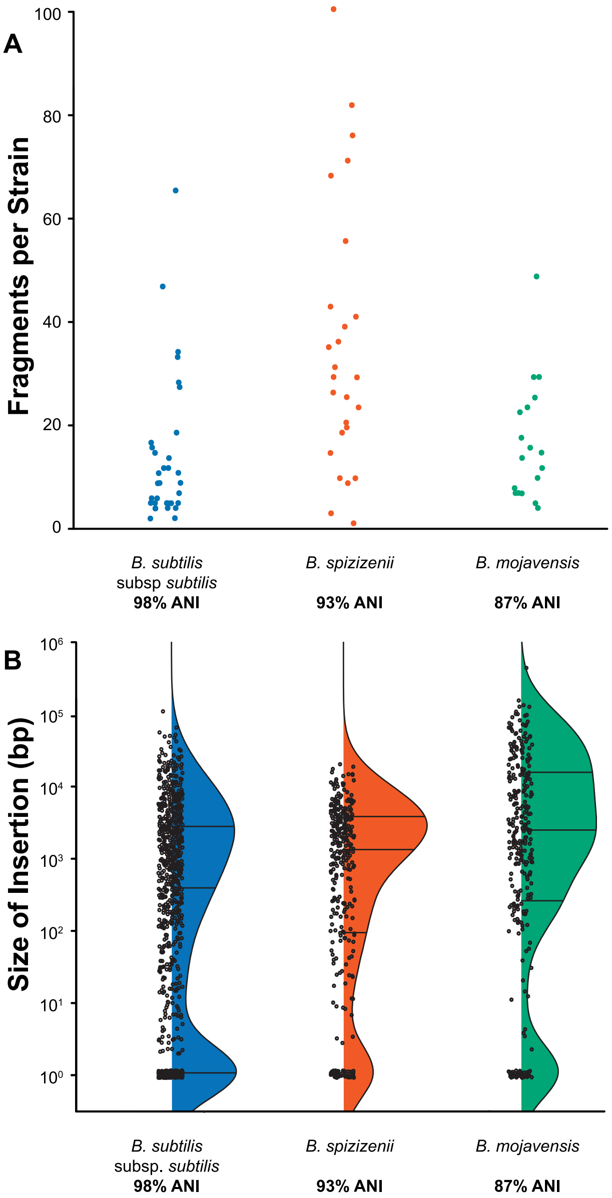

Figure 5. Protoplast fusion yields efficient homologous recombination across species boundaries. A double-resistant mutant of RO-NN-1 was crossed with prototrophic strains of varying genetic distance. No significant differences were observed in (**A**) the number of recombination events per strain or (**B**) the distribution of recombination event sizes. Horizontal lines in the violin plots show the median and interquartile range for each distribution.

of the tested recombinant populations (Supplementary Figure S6). Collectively, these results suggested that while the likelihood of recombination events decreases with increasing genetic distance, when recombination occurs between divergent strains, it creates extensive genetic diversity across the chromosome at macro- and microscale.

In this work, we investigated the genomic consequences of protoplast fusion between *Bacillus* strains. We observed substantial unselected recombination throughout the genome, for a broad range of fragment sizes. Restriction–modification systems affected the directionality of transfer, but no other factors were identified that biased the local position of recombination events. While we were unable to obtain recombinants between strains with low levels of sequence identity, genome-wide recombination was otherwise largely unaffected by variation in sequence identity between parental strains, even among strains classified as different species. It is worth noting that some of the inheritance patterns identified in this work may result from unique aspects of the recombination process during genome shuffling. Therefore, future targeted analyses will be required to delineate the specifics of the recombination mechanism in fused protoplasts. These studies might facilitate development of new strategies for rapid cell engineering. Furthermore, combined with the computational methods developed in this study, protoplast fusion might provide a tractable method for studying homologous recombination at scale with minimal selection bias.

DATA AVAILABILITY

Source code for calculating subpopulation figures is available at GitHub (https://github.com/jstreich/Bsubstilis_QTL_Project_Aug2020/blob/master/210Indv_Bact_Recombination_2021--09-23_V0.1.8D.R).

ACCESSION NUMBERS

Sequence reads from recombined progeny are available from the Sequence Read Archive under accession numbers PRJNA669142 and PRJNA766217.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Bacillus strains were provided by Daniel Zeigler and the Bacillus Genetic Stock Center, through contributions from the initial depositors. This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract No. DE-AC05-00OR22725. This research also used resources of the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory.

FUNDING

Oak Ridge National Laboratory Laboratory Directed Research and Development; U.S Department of Energy, Office of Science, Office of Biological and Environmental Research, Center for Bioenergy Innovation [to D.P.V., J.C.S., H.B.C., J.C.E., D.A.J. and J.K.M.]. Funding for open access charge: Center for Bioenergy Innovation.

Conflict of interest statement. D.P.V., J.C.S., D.A.J. and J.K.M. are inventors on a patent that has been filed based, in part, on the work reported in this manuscript. This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

REFERENCES

1. Soucy,S.M., Huang,J. and Gogarten,J.P. (2015) Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.*, **16**, 472–482.
2. Didelot,X. and Maiden,M.C.J. (2010) Impact of recombination on bacterial evolution. *Trends Microbiol.*, **18**, 315–322.
3. Croucher,N.J., Harris,S.R., Barquist,L., Parkhill,J. and Bentley,S.D. (2012) A high-resolution view of genome-wide pneumococcal transformation. *PLoS Pathog.*, **8**, e1002745.
4. Mell,J.C., Lee,J.Y., Firme,M., Sinha,S. and Redfield,R.J. (2014) Extensive cotransformation of natural variation into chromosomes of naturally competent *Haemophilus influenzae*. *G3*, **4**, 717–731.
5. Cowley,L.A., Petersen,F.C., Junges,R., Jimson D Jimenez,M., Morrison,D.A. and Hanage,W.P. (2018) Evolution via recombination: cell-to-cell contact facilitates larger recombination events in *Streptococcus pneumoniae*. *PLoS Genet.*, **14**, e1007410.
6. Bubendorfer,S., Krebes,J., Yang,I., Hage,E., Schulz,T.F., Bahlawane,C., Didelot,X. and Suerbaum,S. (2016) Genome-wide analysis of chromosomal import patterns after natural transformation of *Helicobacter pylori*. *Nat. Commun.*, **7**, 11995.
7. Power,J.J., Pinheiro,F., Pompei,S., Kovacova,V., Yüksel,M., Rathmann,I., Förster,M., Lässig,M. and Maier,B. (2021) Adaptive evolution of hybrid bacteria by horizontal gene transfer. *Proc. Natl Acad. Sci. U.S.A.*, **118**, e2007873118.
8. Gray,T.A., Krywy,J.A., Harold,J., Palumbo,M.J. and Derbyshire,K.M. (2013) Distributive conjugal transfer in mycobacteria generates progeny with meiotic-like genome-wide mosaicism, allowing mapping of a mating identity locus. *PLoS Biol.*, **11**, e1001602.
9. Derbyshire,K.M. and Gray,T.A. (2014) Distributive conjugal transfer: new insights into horizontal gene transfer and genetic exchange in mycobacteria. *Microbiol. Spectr.*, **2**, 4.
10. Dordet-Frisoni,E., Sagné,E., Baranowski,E., Breton,M., Nouvel,L.X., Blanchard,A., Marenda,M.S., Tardy,F., Sirand-Pugnet,P. and Citti,C. (2014) Chromosomal transfers in mycoplasmas: when minimal genomes go mobile. *mBio*, **5**, e01958.
11. Gray,T.A. and Derbyshire,K.M. (2018) Blending genomes: distributive conjugal transfer in mycobacteria, a sexier form of HGT. *Mol. Microbiol.*, **108**, 601–613.
12. Dordet-Frisoni,E., Faucher,M., Sagné,E., Baranowski,E., Tardy,F., Nouvel,L.X. and Citti,C. (2019) Mycoplasma chromosomal transfer: a distributive, conjugative process creating an infinite variety of mosaic genomes. *Front. Microbiol.*, **10**, 2441.
13. Bartke,K., Garoff,L., Huseby,D.L., Brandis,G. and Hughes,D. (2021) Genetic architecture and fitness of bacterial interspecies hybrids. *Mol. Biol. Evol.*, **38**, 1472–1481.
14. Hopwood,D.A., Wright,H.M., Bibb,M.J. and Cohen,S.N. (1977) Genetic recombination through protoplast fusion in *Streptomyces*. *Nature*, **268**, 171–174.
15. Petri,R. and Schmidt-Dannert,C. (2004) Dealing with complexity: evolutionary engineering and genome shuffling. *Curr. Opin. Biotechnol.*, **15**, 298–304.
16. Biot-Pelletier,D. and Martin,V.J.J. (2014) Evolutionary engineering by genome shuffling. *Appl. Microbiol. Biotechnol.*, **98**, 3877–3887.
17. Zhang,Y.-X., Perry,K., Vinci,V.A., Powell,K., Stemmer,W.P.C. and del Cardayré,S.B. (2002) Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature*, **415**, 644–646.
18. Patnaik,R., Louie,S., Gavrilovic,V., Perry,K., Stemmer,W.P.C., Ryan,C.M. and del Cardayré,S. (2002) Genome shuffling of *Lactobacillus* for improved acid tolerance. *Nat. Biotechnol.*, **20**, 707–712.
19. Magocha,T.A., Zabed,H., Yang,M., Yun,J., Zhang,H. and Qi,X. (2018) Improvement of industrially important microbial strains by

genome shuffling: current status and future prospects. *Bioresour. Technol.*, **257**, 281–289.

20. Luna-Flores,C.H., Palfreyman,R.W., Krömer,J.O., Nielsen,L.K. and Marcellin,E. (2017) Improved production of propionic acid using genome shuffling. *Biotechnol. J.*, **12**, 1600120.
21. Wang,W., Wu,B., Qin,H., Liu,P., Qin,Y., Duan,G., Hu,G. and He,M. (2019) Genome shuffling enhances stress tolerance of *Zymomonas mobilis* to two inhibitors. *Biotechnol. Biofuels*, **12**, 288.
22. Ega,S.L., Drendel,G., Petrovski,S., Egidi,E., Franks,A.E. and Muddada,S. (2020) Comparative analysis of structural variations due to genome shuffling of *Bacillus subtilis* VS15 for improved cellulase production. *Int. J. Mol. Sci.*, **21**, 1299.
23. Zawadzki,P., Roberts,M.S. and Cohan,F.M. (1995) The log-linear relationship between sexual isolation and sequence divergence in *Bacillus* transformation is robust. *Genetics*, **140**, 917–932.
24. Carrasco,B., Serrano,E., Sánchez,H., Wyman,C. and Alonso,J.C. (2016) Chromosomal transformation in *Bacillus subtilis* is a non-polar recombination reaction. *Nucleic Acids Res.*, **44**, 2754–2768.
25. Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessières,P., Bolotin,A., Borchert,S. *et al.* (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–256.
26. Zeigler,D.R. (2011) The genome sequence of *Bacillus subtilis* subsp. *spizizenii* W23: insights into speciation within the *B. subtilis* complex and into the history of *B. subtilis* genetics. *Microbiology*, **157**, 2033–2041.
27. Burkholder,P.R. and Giles,N.H. Jr (1947) Induced biochemical mutations in *Bacillus subtilis*. *Am. J. Bot.*, **34**, 345–348.
28. Earl,A.M., Eppinger,M., Fricke,W.F., Rosovitz,M.J., Rasko,D.A., Daugherty,S., Losick,R., Kolter,R. and Ravel,J. (2012) Whole-genome sequences of *Bacillus subtilis* and close relatives. *J. Bacteriol.*, **194**, 2378–2379.
29. Cohan,F.M., Roberts,M.S. and King,E.C. (1991) The potential for genetic exchange by transformation within a natural population of *Bacillus subtilis*. *Evolution*, **45**, 1393–1421.
30. Srivatsan,A., Han,Y., Peng,J., Tehranchi,A.K., Gibbs,R., Wang,J.D. and Chen,R. (2008) High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. *PLoS Genet.*, **4**, e1000139.
31. Roberts,M.S. and Cohan,F.M. (1995) Recombination and migration rates in natural populations of *Bacillus subtilis* and *Bacillus mojavensis*. *Evolution*, **49**, 1081–1094.
32. Dunlap,C.A., Bowman,M.J. and Zeigler,D.R. (2020) Promotion of *Bacillus subtilis* subsp. *inaquosorum*, *Bacillus subtilis* subsp. *spizizenii* and *Bacillus subtilis* subsp. *stercoris* to species status. *Antonie Van Leeuwenhoek*, **113**, 1–12.
33. Chen,X.H., Koumoutsi,A., Scholz,R., Eisenreich,A., Schneider,K., Heinemeyer,I., Morgenstern,B., Voss,B., Hess,W.R., Reva,O. *et al.* (2007) Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus amyloliquefaciens* FZB42. *Nat. Biotechnol.*, **25**, 1007–1014.
34. Krebs,B., Höding,B., Kübart,S., Workie,M.A., Junge,H., Schmiedeknecht,G., Grosch,R., Bochow,H. and Hevesi,M. (1998) Use of *Bacillus subtilis* as biocontrol agent. I. Activities and characterization of *Bacillus subtilis* strains. *Z. Pflanzenkrankh. Pflanzenschutz/J. Plant Dis. Prot.*, **105**, 181–197.
35. Koo,B.-M., Kritikos,G., Farelli,J.D., Todor,H., Tong,K., Kimsey,H., Wapinski,I., Galardini,M., Cabal,A., Peters,J.M. *et al.* (2017) Construction and analysis of two genome-scale deletion libraries for *Bacillus subtilis*. *Cell Syst.*, **4**, 291–305.
36. Wyrick,P.B. and Rogers,H.J. (1973) Isolation and characterization of cell wall-defective variants of *Bacillus subtilis* and *Bacillus licheniformis*. *J. Bacteriol.*, **116**, 456–465.
37. Schaeffer,P., Cami,B. and Hotchkiss,R.D. (1976) Fusion of bacterial protoplasts. *Proc. Natl Acad. Sci. U.S.A.*, **73**, 2151–2155.
38. Chang,S. and Cohen,S.N. (1979) High frequency transformation of *Bacillus subtilis* protoplasts by plasmid DNA. *Mol. Gen. Genet.*, **168**, 111–115.
39. Spizizen,J. (1958) Transformation of biochemically deficient strains of *Bacillus subtilis* by deoxyribonucleate. *Proc. Natl Acad. Sci. U.S.A.*, **44**, 1072–1078.
40. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
41. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
42. Danecek,P. and McCarthy,S.A. (2017) BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics*, **33**, 2037–2039.
43. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A.R., Bender,D., Maller,J., Sklar,P., de Bakker,P.I.W., Daly,M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
44. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
45. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
46. Cui,Y., Chen,X., Luo,H., Fan,Z., Luo,J., He,S., Yue,H., Zhang,P. and Chen,R. (2016) BioCircos.js: an interactive Circos JavaScript library for biological data visualization on web applications. *Bioinformatics*, **32**, 1740–1742.
47. Robinson,P. and Jtel,T.Z. (2017) Integrative Genomics Viewer (IGV): visualizing alignments and variants. In: Robinson,P.N., Piro,R.M. and Jäger,M. (eds). *Computational Exome and Genome Analysis*. Chapman and Hall/CRC, Boca Raton, FL.
48. Goslee,S.C. and Urban,D.L. (2007) The ecodist package for dissimilarity-based analysis of ecological data. *J. Stat. Softw.*, **22**, 1–19.
49. Spencer,C.C.A., Deloukas,P., Hunt,S., Mullikin,J., Myers,S.R., Silverman,B., Donnelly,P., Bentley,D. and McVean,G. (2005) The influence of recombination on human genetic diversity. *PLoS Genet.*, **2**, e148.
50. Weighill,D.A. and Jacobson,D. (2017) Network metamodeling: effect of correlation metric choice on phylogenomic and transcriptomic network topology. *Adv. Biochem. Eng. Biotechnol.*, **160**, 143–183.
51. Leavey,C.M., James,M.N., Summerscales,J. and Sutton,R. (2003) An introduction to wavelet transforms: a tutorial approach. *Insight: Non-Destr. Test. Cond. Monit.*, **45**, 344–353.
52. Constantine,W., Percival,D.B. and Reinhall,P.G. (2001) Inertial range determination for aerothermal turbulence using fractionally differenced processes and wavelets. *Phys. Rev. E: Stat. Nonlin. Soft Matter Phys.*, **64**, 036301.
53. Markowitz,V.M., Chen,I.-M.A., Palaniappan,K., Chu,K., Szeto,E., Grechkin,Y., Ratner,A., Jacob,B., Huang,J., Williams,P. *et al.* (2012) IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.*, **40**, D115–D122.
54. Slomka,S., Françoise,I., Hornung,G., Asraf,O., Biniashvili,T., Pilpel,Y. and Dahan,O. (2020) Experimental evolution of *Bacillus subtilis* reveals the evolutionary dynamics of horizontal gene transfer and suggests adaptive and neutral effects. *Genetics*, **216**, 543–558.
55. Nye,T.M., van Gijtenbeek,L.A., Stevens,A.G., Schroeder,J.W., Randall,J.R., Matthews,L.A. and Simmons,L.A. (2020) Methyltransferase DnmA is responsible for genome-wide *N*6-methyladenosine modifications at non-palindromic recognition sites in *Bacillus subtilis*. *Nucleic Acids Res.*, **48**, 5332–5348.
56. Roberts,M.S. and Cohan,F.M. (1993) The effect of DNA sequence divergence on sexual isolation in *Bacillus*. *Genetics*, **134**, 401–408.
57. Majewski,J. and Cohan,F.M. (1999) DNA sequence similarity requirements for interspecific recombination in *Bacillus*. *Genetics*, **153**, 1525–1533.
58. Jeffreys,A.J. and Neumann,R. (2005) Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Hum. Mol. Genet.*, **14**, 2277–2287.
59. Jeffreys,A.J., Neumann,R., Panayi,M., Myers,S. and Donnelly,P. (2005) Human recombination hot spots hidden in regions of strong marker association. *Nat. Genet.*, **37**, 601–606.
60. Baudat,F., Buard,J., Grey,C., Fledel-Alon,A., Ober,C., Przeworski,M., Coop,G. and de Massy,B. (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, **327**, 836–840.

61. Dillon,S.C. and Dorman,C.J. (2010) Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nat. Rev. Microbiol.*, **8**, 185–195.
62. Valens,M., Penaud,S., Rossignol,M., Cornet,F. and Boccard,F. (2004) Macrodomain organization of the *Escherichia coli* chromosome. *EMBO J.*, **23**, 4330–4341.
63. Marbouty,M., Le Gall,A., Cattoni,D.I., Cournac,A., Koh,A., Fiche,J.-B., Mozziconacci,J., Murray,H., Koszul,R. and Nollmann,M. (2015) Condensin- and replication-mediated bacterial chromosome folding and origin condensation revealed by Hi-C and super-resolution imaging. *Mol. Cell*, **59**, 588–602.