Title:

Pneumococcal chromosome conformation variation between epigenetic variants is driven by episomal mobile genetic elements

Author list:

Lim Tze Yee[1], Samuel T. Horsfield[1,2], Catherine M. Troman[1], Stephen D. Bentley[3], Min Jung Kwun[1,+] & Nicholas J. Croucher[1,+,*]

Affiliations:

[1] MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, W12 0BZ, UK

[2] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, UK

[3] Parasites & Microbes, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

[+] These authors contributed equally

* Correspondence to n.croucher@imperial.ac.uk

Abstract

Streptococcus pneumoniae (pneumococcus) is a genetically diverse opportunistic bacterial pathogen that expresses two phase-variable loci encoding restriction-modification systems. Comparisons of two genetically-distinct pairs of epigenetically-distinct variants, each distinguished by the arrangement of one of these loci, found the consequent change in genome-wide DNA methylation patterns was associated with differential expression of mobile genetic elements (MGEs). This relationship was hypothesised to be mediated through changes in xenogenic silencing (XS) or nucleoid organisation. Therefore the chromosomal conformation of the both variants of each isolate were characterised using Illumina Hi-C, and Oxford Nanopore Pore-C, sequencing. Both methods concurred that the organisation of the pneumococcal chromosome was generally simple, with most pairwise interactions between loci <15 kb apart. Neither found substantial evidence for higher-order structure or XS in the pneumococcal genome, with locally-elevated contact densities only evident around the origin of replication. Comparisons between the variants identified phage-related chromosomal islands (PRCIs) as the foci of differential contact densities between the variants. This was driven by copy number variation, resulting from variable excision and replication of the episomal PRCIs. However, the methods were discordant in their identification of the variant in which the PRCI was more actively replicating in both pairs. Validatory experiments demonstrated that the prevalence of circular PRCIs was not determined by DNA modification, but instead varied stochastically between colonies in both backgrounds, and was metastable in during vegetative growth. PRCI excision was inducible by mitomycin C, but independent of the presence of a phage. Yet transcriptional activation of these elements was affected by both signals, indicating transcription and replication are separately regulated.

Therefore pneumococcal MGEs do not appear to be subject to XS, resulting in heterogeneity being generated within these bacterial populations through the frequent local disruption of chromosome conformation is by the stochastic excision and reintegration of episomal elements.

Introduction

Streptococcus pneumoniae (the pneumococcus) is a gram-positive commensal bacterium and opportunistic pathogen, capable of causing diseases such as conjunctivitis, pneumonia, sepsis and meningitis [1]. It has an extensive pangenome [2–4] as a consequence of exchange through transformation [5] and the frequent integration of mobile genetic elements (MGEs) into the pneumococcal chromosome [4]. However, the acquisition of genes can be inhibited by restriction modification systems (RMSs) [6,7]. Almost all pneumococci share two phase-variable Type I RMSs: the *Spn*III RMS encoded by the inverting variable restriction (*ivr*) locus [8,9], and the *Spn*IV RMS encoded by the translocating variable restriction (*tvr*) locus [4,9]. The rapid variation of these loci can inhibit the spread of prophage within clonally-related populations [10,11]. Yet the changes also cause global alterations to the methylation of the genome. Hence the sensitivity of pneumococcal transcription to epigenomic changes means these RMSs also cause phenotypic heterogeneity between isolates [9,12].

Variation in the *Spn*III RMS has been associated with changes in capsule expression [9,13,14], human carriage [15], adhesion to human cells [16], and virulence in a mouse model [9,17]. Variation at the *tvr* locus is instead associated with variation in transformation efficiency [12,18]. However, it has not been possible to identify methylation sites that can explain the transcriptional changes through proximal effects at gene regulatory loci [9,12]. Yet methylation has broader effects on the biophysical properties of DNA that means a global change in methylation patterns could affect the conformation of the bacterial nucleoid, and alter the interactions between regulatory proteins and DNA loci [19]. Hence it is plausible that the

epigenetic effects are instead mediated through longer-distance, larger-scale changes in chromosomal conformation.

This is supported by analysis of two pneumococcal isolates, RMV7 and RMV8. Mutations disrupting the recombinase that catalyses rearrangements at the *tvr* locus enabled “locked” phase variants of each to be isolated: one representing the common, dominant arrangement ($RMV7_{domi}$ and $RMV8_{domi}$), and one expressing a rare arrangement ($RMV7_{rare}$ and $RMV8_{rare}$) [7]. These variants differ in the methylation motif targeted by the *Spn*IV RMS, but are otherwise isogenic outside the *tvr* loci. Both rare variants were more transformable than their dominant variant counterparts, and in both cases this was at least partly attributable to the altered activity of MGEs integrated into the chromosome [12,18]. RMV7 contains a cryptic integrative and conjugative element next to *rpsI* (ICE_{rpsI}) and the antibiotic-resistance transposon Tn*916*, as well as two phage-related chromosomal islands (PRCIs; alternatively known as phage-inducible chromosomal islands, or PICIs) integrated next to *dnaN* ($PRCI_{dnaN}$) and *uvrA* ($PRCI_{uvrA}$). $PRCI_{dnaN}$ was more transcriptionally active in $RMV7_{domi}$, suppressing the induction of competence for transformation through inducing stress responses [12]. RMV8 encodes a prophage (□RMV8), a PRCI integrated within the *mal* operon ($PRCI_{malA}$), and a CIPhR prophage remnant (previously annotated as $PRCI_{tadA}$) [18,20]. In $RMV8_{domi}$, reduced excision of □RMV8 resulted in greater expression of a prophage-modified RNA that suppressed competence induction [18]. However, no specific methylation sites have been identified that could explain this difference in MGE activity [12].

Hence host DNA conformation is a plausible mechanism by which epigenetic modification could modulate the activity of MGEs through xenogeneic silencing (XS) [21]. This process depends on nucleoid-associated proteins (NAPs) that preferentially bind to the AT-rich DNA that is characteristic of MGEs [22,23], and are capable of silencing their gene expression. As well as affecting MGE gene transcription [23–25], these NAPs can modulate the excision and re-integration of MGEs [26]. Such differences in XS activity could explain the inter-variant differences observed in RMV7 and RMV8. At least four families of NAPs have been identified [21,27–30]. Two are common in gram-negative *Proteobacteria*: H-NS, identified in *Escherichia coli* [31], and MvaT, characterised in *Pseudomonas* [30]. By contrast, Lsr2 has only been found in gram-positive *Actinobacteria* [29], and the Rok protein has only been characterised in *Bacillus subtilis* [32]. Hence no XS system has yet been characterised in *S. pneumoniae*. Nevertheless, the species encodes orthologues of Rok, as well as other proteins that preferentially bind AT-rich DNA [33]. Therefore the differential activity of MGEs in RMV7 and RMV8 may be attributable to changes in the patterns of NAPs associated with the genome.

The conformation of the genome, and therefore the distribution of NAPs and regions subject to XS, can be characterised using chromosome conformation capture (3C) [34]. This approach typically involves crosslinking DNA to proteins with formaldehyde, digesting the DNA with a frequently-cutting restriction enzyme, then facilitating proximity ligation to link DNA fragments that were spatially close within the cell (Fig. 1A). The subsequent use of paired-end Illumina data enables this to be assayed at a genome-wide scale using high-throughput 3C (Hi-C), which infers instances of digestion and re-ligation within the DNA separating the two reads based

on their mapping positions and orientations relative to a reference genome [35]. More recently, long-read data generated by Nanopore sequencing has enabled 3C analyses to use information from concatemers consisting of multiple DNA fragments ligated together [36,37]. Hence this Pore-C method makes it possible to infer higher-order multiway chromatin interactions, and represents an efficient method for generating data on the scale of a bacterial genome.

Therefore this study compares Hi-C and Pore-C approaches to inferring the chromosomal conformation of the dominant and rare variants of *S. pneumoniae* RMV7 and RMV8. These assays will determine whether it is feasible that the phenotypic heterogeneity associated with epigenetic variation can be attributed to changes in XS, or other alterations in the distribution of NAPs across the genome.

Results

Short-range interactions dominate pneumococcal Hi-C and Pore-C data

Three independent 3C preparations were generated for each of *S. pneumoniae* $RMV7_{domi}$, $RMV7_{rare}$, $RMV8_{domi}$ and $RMV8_{rare}$ (Table S1, S2). These were digested with *Nla*III, which can cut at sites found in the pneumococcal chromosome at intervals of ~300 bp (Table S3), and used to generate a 12-plex library for Nanopore sequencing. However, this yielded few reads, as the activity of the pores declined rapidly during the sequencing run (Fig. S1). Residual DNA from these samples was instead processed for Illumina sequencing, and successfully generated Hi-C data (Table S2). A further three independent 3C preparations were generated for each of the four variants using *Mlu*CI, which can cut at sites found in the pneumococcal

chromosome at intervals of ~125 bp (Table S3). Previous applications of the Pore-C approach to human cells found an additional protease treatment with pronase increased the efficacy of de-cross-linking, thereby slowing the rate of nanopore blocking by proteins attached to DNA [38]. Correspondingly, including this step in the sample preparation protocol enabled more efficient data generation from individual flow cells (Fig. S1; Table S2).

The three biological replicates for both methods for each of the variants were mapped to the appropriate dominant variant reference genome (Fig. S2, S3; Table S1). The necessity of mapping short read fragments meant the Pore-C data, which contains a higher rate of sequencing errors, mapped less efficiently than the Hi-C data. Nevertheless, a high proportion of the Nanopore data could be uniquely mapped to locations within the pneumococcal genome (Fig. S3). To establish the highest resolution at which analyses could be reliably conducted, the proportions of inferred mapping orientations of sequence fragments were plotted against the separation between their mapped locations if they came from the same read pair, for Illumina Hi-C data, or from the same read, for Pore-C data. This found that sequence fragments separated by less than 5 kb were typically mapped to opposite strands of the genome for Illumina Hi-C data (Fig. S4), or to the same strand of the genome, for Pore-C data (Fig. S5). These are the orientations expected for undigested DNA sequenced with these technologies, consistent with only a subset of restriction sites being cleaved during sample preparation. This is typical of 3C preparations, requiring a convergence distance to be defined as the separation at which sufficient digestion and re-ligation occurred to ensure all sequence fragment mapping orientations

become equally frequent [39]. This was estimated to be between 5 and 10 kb for both technologies (Fig. S4-5).

Plotting the overall frequency of pairs against their separation highlighted particularly strong interactions at an interval of ~5 kb (Fig. 1B), which may indicate a particular set of atypical interactions. However, plotting the locations of these high-level short-range interactions found they were distributed across the chromosome, and therefore may represent false positive interactions resulting from incompletely-digested DNA (Fig. S6). Therefore data were subsequently analysed at a resolution of 10 kb, which should be sufficiently conservative to primarily identify interactions between DNA fragments that have been digested and religated.

Therefore contact frequencies were inferred between all loci across the chromosome at a 10 kb resolution for each replicate (Fig. S7-8). Combining the data for each set of replicates revealed a strong diagonal across the symmetrical contact frequency matrix (Fig. 1C). This suggested most interactions occurred over distances <25 kb, consistent with a simple genome organisation, dominated by short-range structures. There were relatively few signals of higher-level chromosomal organisation, apart from an elevated density of off-diagonal interactions evident around the origin of replication (*ori*; Fig. 1D, S8). This was positioned ~1.6 Mb from the beginning of the RMV7 reference, resulting in vertical and horizontal stripes across the matrix (Fig. 1C). The RMV8 *ori* was at the start of the reference genome, resulting instead in a framing effect, with elevated intensities in each corner of the matrix. Zooming in on the origin of replication found boundaries between replichores that were consistent across technologies evident at ~40 kb and ~20 kb upstream of *ori* in RMV7 and

RMV8, respectively (Fig. 1D). The greater offset in RMV7 may reflect changes in organisation caused by the integration of $PRCI_{dnaN}$ ~3.5 kb downstream of *ori* [12]. Hence both Hi-C and Pore-C data produced a consistent view of the contacts that dominate the pneumococcal genome's conformation.

Differential interactions between variants span PRCIs

A genome-wide search for differences in chromosome conformation between the variants was undertaken using contact matrices that summarised contact counts within non-overlapping 10 bp loci. The DiffHiC (Fig. S9-S12) and multiHiCcompare (Fig. S13-S16) methods were used, both of which fit generalised linear models to independent biological replicate datasets to identify differences in contact frequencies [40,41]. However, they differ in their normalisation of the data [42]. Clustering of the biological replicates by the diffHiC pipeline suggested the variants could be resolved from one another by the genome-wide patterns of contacts in each comparison, except in the Hi-C analysis of the RMV7 variants (Fig. S17). Correspondingly, sites exhibiting significantly different contact frequencies between the genotypes, following Benjamini-Hochberg corrections for multiple testing, could be identified using Manhattan plots (Fig. 2A).

For RMV8, the only loci that were consistently identified as significantly differing between the two variants by all four comparisons were within, or adjacent to, $PRCI_{malA}$ [18]. The interacting loci, and the fold differences in contact frequency, were plotted against the annotation of this MGE (Fig. 2B). The hits were either between the flanking region and the element, or between the two flanks. However, while the spatial distribution and magnitude of the effects was similar across all

comparisons, the Hi-C data suggested higher contact frequencies in $RMV8_{rare}$, whereas the Pore-C data suggested higher contact frequencies in $RMV8_{domi}$. To rule out sample misidentification as a cause of this discrepancy, it was verified that the patterns of contact frequencies at the *tvr* locus matched the arrangements characteristic of the variants, and that a simple analysis of contact frequency ratios within $PRCI_{malA}$ supported the genome-wide analysis results (Fig. S18). Detailed analysis of the Nanopore sequencing read lengths mapping to this region did not suggest this was an artefact of the Pore-C methodology preferentially sequencing shorter DNA fragments originating from the MGE (Fig. S19). Therefore the Hi-C and Pore-C analyses concurred that there is variation in the contact frequency within $PRCI_{malA}$, but disagreed on the variant in which the contact density was highest.

For RMV7, the analyses of Hi-C data were consistent in identifying $PRCI_{dnaN}$ [12] as the region differing most significantly between the variants (Fig. 2A). This MGE was also the location of the most significant difference in contact frequencies between the variants in the diffHiC comparison of Pore-C datasets, in addition to being the site of multiple significant differences in the multiHiCcompare analysis of the same data (Fig. 1D). The detailed plot of these points relative to the $PRCI_{dnaN}$ annotation demonstrated these hits were either within the element, or between the element and its flanking regions (Fig. 2B). While the differences were again of similar magnitude and distribution across the two methods, the Hi-C and Pore-C data were also discordant in identifying the variant in which the contact frequency was highest. In this comparison, the Hi-C data suggested the contact densities were higher in $RMV7_{domi}$, whereas the Pore-C data implied they were higher in $RMV7_{rare}$. Therefore multiple significant differences in contact frequencies between both pairs of variants

were within or flanking PRCIs, although the methods did not reach a consensus on the nature of the altered pattern of interactions.

Changes in PRCI contact frequency do not reflect changes in XS

To test whether a more specific, biologically-informed model could identify the nature of the differences more precisely, a generalised linear mixed-effects model was fitted to the four combinations of sequencing methodology and genotype at a resolution of 10 kb using a zero-inflated Poisson distribution with a logarithmic link function (see Methods; Fig. S20-4). The fitted models were able to explain a substantial fraction of the variation in contact frequencies across loci and datasets (Fig. S25). The effect of distance between loci was estimated to be consistent between the genotypes, but contact frequencies declined more rapidly with distance in the Hi-C data, suggesting Pore-C data may be more efficient at detecting long-range interactions (Fig. 3A).

To test whether any of the results may be an artefact of the restriction enzymes used to generate the samples, the model also estimated the sensitivity of the inferred contact frequencies to the density of sites targeted by these endonucleases (Fig. 3B). With the Pore-C data, there was no evidence of any strong relationship between restriction sites and inferred contact density. With Hi-C data, a higher density of *Nla*III sites was associated with an increased sensitivity for detecting contacts, which is consistent with the experimental protocol. However, there was an inverse relationship between the inferred contact density and the density of *MluC*I sites, despite this enzyme not being involved in sample preparation. Given the AT-rich nature of the *MluC*I sequence motif (AATT), such an observation may be attributable to Illumina sequencing's bias towards DNA with a neutral GC content, which is not

shared with Nanopore data [43]. However, contact frequencies did not exhibit an overall inverse relationship with GC content. Rather, contact frequencies increased with rising GC content in RMV7, but decreased in RMV8 (Fig. 3C). Both trends were more extreme for the Hi-C data than the Pore-C data. This suggests the Illumina data may be more affected by localised regions of extremely AT-rich sequence, rather than the mean GC content averaged over a 10 kb locus. Nevertheless, no consistent evidence could be found of a pneumococcal XS system that preferentially condensed AT-rich sequences [44].

A more specific test for XS was quantifying the contact density across the MGEs within each of the genomes, relative to the rest of the chromosome. For RMV7, neither the Hi-C or Pore-C data found any evidence of elevated contact densities across the identified MGEs, with the confidence intervals of the level of contact density relative to the rest of the genome always spanning one (Fig. 3D). When comparing the variants, differences in the density of contacts were only inferred for PRCI$_{dnaN}$. In further agreement with the differential contact analyses (Fig. 2B), the contact frequencies were higher for PRCI$_{dnaN}$ in RMV7$_{\text{rare}}$ for the Pore-C data, but lower in the Hi-C data. For RMV8, the Illumina Hi-C data again found no evidence of the RMV8 or PRCI$_{malA}$ MGEs being associated with elevated contact densities, although there was some evidence of higher contact frequencies within the latter in the Pore-C data. However, the Pore-C data also suggested that contact frequencies within PRCI$_{malA}$ were reduced in the rare variant, suggesting there was no consistent evidence for increased contact densities across both variants. By contrast, the Illumina Hi-C data suggested that PRCI$_{malA}$ was associated with higher contact densities in RMV8$_{\text{rare}}$. Therefore the model outputs were consistent with the

differential contact analysis in both genotypes, and there was no evidence of elevated contact densities within the pneumococcal MGEs that could provide evidence of XS.

To test whether there was evidence of XS at any location within MGEs, locus-specific effects on contact frequencies were estimated for the sites within and flanking PRCI$_{dnaN}$ and PRCI$_{malA}$ (Fig. 3E). There was no evidence of heterogeneity within the elements, although the size of PRCI$_{malA}$ meant it contained only a single locus, making the elevated contact density inferred for this element in the Pore-C data highly susceptible to confounding by locus-level effects (Fig. 3D). There was also little variation between the variants at these sites (Fig. 3F), which is consistent with the differences being captured by the MGE-level variables (FIg. 3D). However, variation was evident in the flanking regions, the direction of which mirrored the alterations in the elements themselves (Fig. 3D). Therefore the differential contact densities at the PRCIs was not caused by changing levels of XS-driven condensation, but instead reflected altered interactions between the MGEs and the proximal regions of the host chromosome.

Contact frequency changes are associated with altered PRCI activity

Such an observation suggested the differences between the variants may reflect changes in excision, replication and activation of the PRCIs. Given the conflicting data between the Hi-C and Pore-C methods, previously-collected RNA-seq data were used to identify the variants in which the MGEs were more active. PRCI$_{dnaN}$ had previously been identified as being transcribed at a significantly higher level in RMV7$_{domi}$ than RMV7$_{rare}$ during early exponential phase using both RNA-seq (Fig.

4A) and quantitative reverse transcriptase PCR (qRT-PCR) data [12]. Analogously, the PRCI$_{malA}$ genes were generally more highly expressed in RMV8$_{rare}$ than RMV8$_{domi}$ (Fig. 4B), although these differences were not consistent enough to reach genome-wide significance [18]. However, no difference was observed when the *tvr* loci of the variants were removed (Fig. 4C), indicating the differences might be associated with epigenetic modifications [18].

To validate whether transcription of PRCI$_{malA}$ did differ between the variants, qRT-PCR was used to assay the expression of *int*, encoding the element's integrase, and *dnaC*, encoding a DNA replication protein, during early exponential (OD_{600} = 0.2) and mid-exponential (OD_{600} = 0.5) phase growth (Fig. 4D). Although no substantial difference was evident in early exponential phase, the transcription of the replication gene rose to be ~4-fold higher in RMV8$_{rare}$ than RMV8$_{domi}$ during mid-exponential phase, consistent with the higher expression of these genes in the same variant in the RNA-seq data (Fig. 4B). This suggested that PRCI$_{dnaN}$ and PRCI$_{malA}$ were less frequently transcribed in RMV7$_{rare}$ and RMV8$_{domi}$, respectively. Hence transcription, as well as contact densities, differed between epigenetic variants at these PRCIs.

The consistently detectable levels of gene expression, which changed in response to growth, were also consistent with an absence of XS. Instead, these transcriptional signals indicated that the PRCIs were likely to be excising and replicating, which would explain the differences in contacts within these elements, and between them and their flanking regions [45]. Therefore qPCR assays were used to measure the ratios of excised to integrated elements, and the copy number of circularised elements within the cell (Fig. 4E,F). In RMV7, the ratio of excised to integrated

PRCI$_{dnaN}$ was significantly higher in RMV7$_{rare}$, although there was no substantial difference in the number of circularized elements within the cell. In RMV8, both the ratio of excised to integrated elements, and the frequency of circularised PRCI$_{malA}$ genomes, were significantly higher in RMV8$_{domi}$. Therefore the altered contact frequency patterns between variants likely represented changes in the topological configuration of the elements.

For such a topological change to fully explain the significant differences of opposing magnitude in the Hi-C and Pore-C comparisons, it would need to be possible for differences in PRCI circularization to be metastable [46]. Under such circumstances, a variant may exhibit levels of PRCI circularization and replication that are stable over enough bacterial generations such that separate biological cultures grown in parallel have consistent properties. Yet the conflicting results between analyses demands that it is also possible that re-isolation of the variant produces a genotype with a different level of PRCI activity that is similarly persistent.

To test this, three independent colonies were cultured for each variant of RMV7 and RMV8. Each of these cultures was frozen, used to inoculate overnight growth, then diluted into three parallel tubes of fresh media until they reached exponential phase growth. Measurements of the level of excision, and extracellular circularization, were generally consistent across the parallel triplicate cultures inoculated from the same isolate (Fig. 5A). However, there was substantial variation between cultures derived from different colonies for RMV8. These results also contrasted with the previous qPCR assays, which were performed on the same isolate used for the Pore-C analysis (Fig. 4F). For RMV7, the results were more consistent across the colonies

representing each variant (Fig. 5A). However, $RMV7_{domi}$ was generally associated with the higher median copy number across the replicates in this experiment, whereas this value was higher for $RMV7_{rare}$ in the previous assay (Fig. 4F). Therefore PRCI copy number levels vary stochastically between even isogenic isolates, but are metastable during culturing in liquid media.

PRCI excision and expression are separately regulated

The variation in contact frequency and copy number between isolates of the same variant contrasted with the consistent differences in PRCI expression between RNA-seq and qRT-PCR data experiments (Fig. 4B,D) [12]. Correspondingly, previous passages of RMV7 genotypes, in which *tvr* loci with differing arrangements were independently introduced, found phenotypic and expression differences were stable over multiple days [12]. The topological arrangement of $PRCI_{dnaN}$ was assayed in the same samples (Fig. 5B). This demonstrated substantial variation in the integration-excision dynamics of this element across isolates of both variants. Hence differences in PRCI expression may be stable, despite stochastic variation in the topology of the element.

This implied the excision and transcription of PRCIs may be differentially regulated, consistent with PRCI gene expression being highest in the variants in which these elements were most frequently integrated into the chromosome (Fig. 4D,F). This was tested using RMV8, which carries both the □RMV8 prophage and $PRCI_{malA}$ [18]. $RMV8_{rare}$ cells were exposed to competence stimulating peptide (CSP) and mitomycin C (MMC), both of which activate the □RMV8 prophage [20]. These assays were conducted with $RMV8_{rare}$ and a mutant in which □RMV8 had been

replaced with a chloramphenicol acetyltransferase (*cat*) resistance marker, to ascertain whether the PRCI's activity was affected by the prophage (Fig. 5C). Both the level of excision, and the concentration of circularised PRCI elements, rose in response to MMC, but not CSP. This behaviour was independent of the presence of a prophage within the same cell. This suggests the excision of PRCI$_{malA}$ is controlled by the MGE itself in response to similar stimuli that activate prophage.

A qRT-PCR assay was then used to test whether MMC also activated PRCI$_{malA}$ *int* and *dnaC* gene expression in both RMV8$_{rare}$ and RMV8$_{rare}$ RMV8::*cat* (Fig. 5D). An inconsistent rise in the transcription of both genes was recorded in MMC supplemented cultures, correlating with the elevated copy number of the element (Fig. 5C). However, independently of treatment with MMC, transcription levels were consistently lower in the absence of the prophage (Fig. 5D). This effect was greatest for the *dnaC* replication gene, the median expression of which was 4.8-fold higher in the absence of RMV8 in unsupplemented media, and 3.4-fold higher in the absence of RMV8 in MMC-supplemented cultures. This is consistent with the PRCI's dependence on the intact prophage for its ability to transmit between host cells [47]. Therefore expression of the PRCI is regulated by additional factors, associated with the presence of prophage, that do not affect excision.

PRCIs do not cause a substantial fitness cost to their host cells

To test whether this regulation of PRCI activity limited their cost to their host cell, the growth of mutants in which PRCIs had been replaced with a Janus cassette [48] was compared to the corresponding unmodified genotypes in unsupplemented media (Fig. 6A). Neither RMV7$_{rare}$ PRCI$_{dnaN}$::Janus nor RMV8$_{rare}$ PRCI$_{malA}$::Janus exhibited

detectable changes in growth in unsupplemented media relative to the variant from which they were derived. By contrast, removal of the prophage RMV8 substantially increased the growth of $RMV8_{rare}$ (Fig. 6B), consistent with the behaviour of a similar RMV8 prophage null mutant [20]. The deletion of $PRCI_{malA}$ from this non-lysogenic mutant caused no additional change. Therefore PRCIs cause a substantially smaller impairment of host pneumococcus growth than prophage in the absence of an activating stimulus.

Growth assays were then used to test whether this changed when PRCIs were induced to excise from the chromosome with MMC during early exponential growth. Both $RMV7_{rare}$ and the corresponding $PRCI_{dnaN}$::Janus grew similarly in the presence of MMC, consistent with the MGE not reducing host cell fitness upon excision (Fig. 6C). Although MMC did cause a dose-dependent reduction in the density of $RMV8_{rare}$ cultures, this sensitivity was substantially reduced in a non-lysogenic RMV8::*cat* mutant (Fig. 6D). By contrast, removal of $PRCI_{malA}$ did not affect MMC sensitivity in either the wild type or non-lysogenic RMV8 cells. Therefore stimulating the excision and circularisation of PRCIs neither adversely affect pneumococcal host cell growth, nor interfered with the lysis of cells following prophage induction.

Discussion

This application of both Hi-C and Pore-C methods to epigenetic variants of two *S. pneumoniae* genotypes provides an overview of the pneumococcal chromosome's organisation. Both methods concurred that the chromosomal contacts were dominated by short-range interactions, consistent with a simple, locally-condensed structure. Neither method found strong evidence of higher-level organisation, except

for a more spatially-diffuse distribution of contacts surrounding the origin of replication (Fig. 1C). This contrasted with observations in *Bacillus subtilis* [49], *E. coli* [50,51], *Caulobacter crescentus* [52] and *Mycoplasma pneumoniae* [53]. It may be that the relatively small pneumococcal chromosome, with strong coding biases on each replichore [54], is capable of orchestrating concurrent transcription and regulation without substantial higher-order chromosomal folding.

This simple structure may reflect the apparent paucity of NAPs in pneumococci relative to *E. coli* and *B. subtilis*, with only the histone-like protein HU (or HlpA) characterised in *S. pneumoniae* [55–58]. The lack of characterised orthologues of many of the proteins involved in XS is consistent with the absence of condensed regions that form boundaries between large domains in other species [49]. Nevertheless, the epigenetic variants of RMV7 and RMV8 were each distinguished by variable contact density at the sites of PRCI integration (Fig. 2A). However, the approximately two-fold increases in contact density at these elements (Fig. 2B) was consistent with the *attP* locus of the circularised form having a similar copy number to the chromosome itself (Fig. 4E, Fig. 5A), effectively doubling the presence of the PRCIs. Hence the alternative explanation of condensation by XS was unlikely. It is possible that the high density of short-range contacts identified chromosome-wide in our analyses prevents the identification of greater condensation of loci. Yet the absence of XS also explains the absence of any detectable increase in contact density in AT-rich regions, or within other MGEs (Fig. 3C, D). Furthermore, a lack of XS is consistent with the continued expression, excision and activation of these PRCIs (Fig. 4, 5), and the larger RMV8 prophage (Fig. 6), even in the absence of inducing stimuli [18,20]. Such a difference in genome biology from *B. subtilis* and *E.*

coli suggests a fundamentally distinct interaction between MGEs and host cells in these species. One possible explanation is that XS proteins have an important role in controlling the activity of plasmids [59]. Such elements are common in *E. coli* [60], found in ~10% of *B. subtilis* isolates [61], but rarer in *S. pneumoniae* [4]. Hence pneumococci may have adopted a different strategy for reducing the fitness cost or transmissibility of their primarily chromosomally-integrated MGEs.

The activity of PRCI$_{malA}$ caused the only consistent significant difference in contact frequencies between the variants of RMV8 (Fig. 2A). This contrasted with the previous RNA-seq analysis, which only identified *comC* and RMV8 as sites of significantly-differing transcriptional variation between RMV8$_{domi}$ and RMV8$_{rare}$ [18]. However, the excision of RMV8 is linked to its increased expression, typically resulting in cell lysis [20]. By contrast the excision of PRCI$_{malA}$ is reversible, without substantial fitness cost to the host (Fig. 6), meaning these episomes are always likely to be hotspots of variation in copy number, and therefore contact density. Additional significant differences in contact density were also identified across the genome in the Pore-C data, with greater variation in RMV7 than RMV8. This is consistent with the more widespread differences in transcription between RMV7$_{domi}$ and RMV7$_{rare}$ than between the RMV8 variants [12,18]. Hence alterations in chromosomal conformation remains a feasible mechanism to explain how distinct patterns of epigenetic DNA modification can affect transcription. However, this observed correlation cannot establish causation without higher-resolution 3C data, and further experimental work that can change both the DNA methylation patterns, and the organisation of the genome.

However, the DNA modifications could not account for the differences in excision dynamics, which were variable between different isolates of the same variant (Fig. 5A). This diversity likely represents the PRCIs engaging in “bet hedging” behaviour [12,62], as integrated elements are stably inherited following chromosomal replication, but are at risk of deletion by homologous recombination [63]. Hence the vertical inheritance of PRCIs is likely to be maximised by existing as both chromosomally-integrated and episomal forms. That the *attB*/*attL* ratios were below one suggested most *attB* sites contained a chromosomally-integrated form (Fig. 4D), ensuring they will be passed to daughter cells in clonally-replicating populations. However, the *attP* levels were in the range 0.3-5, implying that circular forms were also widespread, meaning the PRCI can surivive even if the chromosomally-integrated copy is deleted by recombination. Therefore these episomal dynamics can help explain the stable inheritance of PRCIs by strains over decades, in contrast to the transient associations between prophage and host genotypes [4].

This accounts for why it is adaptive for the PRCIs to increase their excision in response to the detection of RecA-coated nucleoprotein filaments (Fig. 5C), which are generated by MMC, and mediate homologous recombination [20]. Yet PRCI excision does not cause the same fitness cost to the cell as prophage activation, because it cannot lyse the host in the absence of an intact “helper” phage (Fig. 5D). Therefore it is adaptive for expression to be regulated in response to prophage activity, as was apparent in RMV8. This distinction between the regulation of excision and activation means the former can occur stochastically, even in the absence of a prophage, as in RMV7. The observed epigenetic inheritance of a variable state suggests that the levels of excision are determined by the

concentration of regulatory molecules being passed on from mother to daughter cells [64], although no mechanistic details were analysed in this study. Hence the discrepancy between the PRCI copy numbers detected in the Hi-C and Pore-C data is indicative of these genomes being dynamic systems, the behaviour of which is determined by metastable regulatory switches, which cannot by fully described by a single sequence alone.

Hence Pore-C will reveal additional heterogeneity within species, strains, and even clonal bacterial populations, with identical genomes being differentiated by their topology and copy number of individual elements. Although our analysis used data from two flow cells, one was sufficient to identify the key differences between the studied genotypes, meaning this method can be implemented by any laboratory with the capacity to use sequencing platforms on the scale of a MinION. Hence future applications of this method to species with a greater diversity of replicons, NAPs and XS systems promises to aid our understanding of the diversity of bacterial genome biology across species.

Methods

Culturing of *S. pneumoniae*

All *S. pneumoniae* strains (Table S1) were cultured at 35°C in a 5% CO_2 atmosphere. Culturing on solid media used Todd-Hewitt broth (Sigma-Aldrich) supplemented with 0.5% yeast extract (Sigma-Aldrich), 1.5% agar (Sigma-Aldrich) and 200 U ml^{-1} catalase (Sigma-Aldrich). To select desired mutant genotypes, media were additionally supplemented with antibiotics: streptomycin (Sigma-Aldrich) or

kanamycin (Sigma-Aldrich) at 400 µg ml^{-1}, or chloramphenicol (Sigma-Aldrich) at 4 µg ml^{-1}.

Unless stated otherwise, liquid culturing utilised a 'mixed' media consisting of Todd-Hewitt broth supplemented with 0.5% yeast extract and Brain Heart Infusion broth (Sigma-Aldrich) mixed at a 2:3 ratio [12]. For comparing PRCI dynamics between variants, samples were collected during mid-exponential growth, at an optical density at 600 nm (OD_{600}) of ~0.6. For measuring the PRCI responses to stimuli, either unsupplemented media, MMC (850 ng), or competence stimulating peptide (CSP; 8.75 µg) was added to 7 ml of bacteria grown to an OD_{600} of ~0.2. The added CSP was appropriate for inducing competence in each genotype: CSP1 for RMV7, and CSP2 for RMV8. The cultures were then incubated at 35ºC with 5% CO_2 for 2 h. Subsequently, 5 ml of culture was used for RNA extraction, and 2 ml was centrifuged at 3220 *g* for 10 min to produce a cell pellet for DNA extraction.

To measure growth curves, 2 x 10^4 cells from titrated frozen glycerol stocks were grown in mixed liquid media in 96-well microtiter plates at 35ºC with 5% CO_2 for 20 h. For all experiments measuring bacterial sensitivity to inducing stimuli, 200 µl of bacterial cultures with an OD_{600} between 0.15 and 0.25 were mixed with MMC (10 ng), CSP (75 ng), or H_2O_2 (2 ng), then grown in microtiter plates at 35ºC with 5% CO_2 for 20 h. The OD_{600} was measured at 30 min intervals using a FLUOstar Omega microplate reader (BMG LABTECH). Three or four replicate wells were assayed for each tested genotype and condition in each experiment.

DNA purification and PCR amplification

Following overnight incubation, bacteria were pelleted by centrifugation at 3220 *g* for 10 min. Supernatants were discarded, and bacterial pellets were then resuspended in 480 µl of lysis buffer (Promega) and 120 µl of 30 mg ml^{-1} lysozyme (Sigma-Aldrich). Samples were incubated at 35ºC for 45 min and centrifuged for 2 min at 8000 *g*. Bacterial genomic DNA was then extracted using Wizard Genomic DNA Purification Kit (Promega), following the manufacturer's instructions. For experiments requiring subsequent PCR product purification, 1000 ng of genomic DNA, 25 µl of 2x DreamTaq Master Mix (Thermo Fisher), nuclease-free water (Thermo Fisher) and 2 µl each of the 10 µM forward and reverse primers listed in Table S4 were used for PCR amplification in a total reaction volume of 50 µl. Otherwise, 500 ng of genomic DNA, 7.5 µl of 2x DreamTaq Master Mix (Thermo Fisher), nuclease-free water and 1 µl each of a 10 µM forward and reverse primers (Table S2) were combined in a total reaction volume of 15 µl.

Construction of the *S. pneumoniae* RMV8 PRCI$_{malA}$ mutant

To generate the PRCI$_{malA}$::Janus construct, the ~1 kb region flanking upstream (UP) and downstream (DOWN) of PRCI$_{malA}$ were amplified by PCR using primers that inserted restriction enzyme sites on the internal sides of each region. The UP region was generated with primers malA_KO_UP_For and malA_KO_UP_ApaI, whereas the DOWN region was generated with primers malA_KO_DOWN_For_BamHI and malA_KO_DOWN_Rev. The Janus cassette was amplified with the corresponding primers with added restriction sites (Table S4). Purified PCR amplicons were then obtained via gel electrophoresis with a 1% agarose gel dyed with SYBR Safe (Thermo Fisher) in Tris/Borate/EDTA (TBE) buffer, followed by gel extraction with GenElute Gel Extraction Kit (Sigma-Aldrich).

Following extraction, purified PCR amplicons were digested at 37ºC for 2 h with the appropriate restriction enzyme: ApaI (Promega) for UP, and BamHI (Promega) for DOWN. Digested UP and DOWN products were then mixed with the digested Janus cassette at a 2:2:1 ratio and incubated with T4 DNA ligase (New England Biolabs) at room temperature (RT) overnight. The ligation mix was then used as the template for a touchdown PCR, and the amplified PRCI$_{malA}$::Janus construct was purified following separation by electrophoresis on a 1% agarose gel, and extracted using a QIAquick Gel & PCR Cleanup kit (Qiagen). To transform recipient bacteria (RMV8$_{rare}$ or RMV8$_{rare}$ RMV8*::cat*) with the PRCI$_{malA}$::Janus construct, a bacterial culture (1 ml) at an OD_{600} of ~0.2 was incubated with 5 µl of 500 mM $CaCl_2$ (Sigma-Aldrich), 1250 ng of CSP2, and 1 µg of the PRCI$_{malA}$::Janus construct at 35°C for 2 h. Samples were then spread onto solid media supplemented with antibiotics (400 µg ml^{-1} of either kanamycin or streptomycin) to select for resistant transformants that grew after incubation at 35ºC with 5% CO_2 for at least 16 h.

RNA purification and reverse transcription

A 5 ml sample of a bacterial culture was mixed with an equal volume of RNAprotect (Qiagen) and incubated at room temperature for 5 min. RNA samples were extracted using the SV Total Isolation System (Promega) according to the manufacturer's instructions, with the exception that the initial extraction of RNA required treating cell cultures with 480 µL 30 mg mL^{-1} lysozyme (Promega) for 30 min at 35 °C. The purified RNA was treated with amplification-grade DNase I (Invitrogen), then used in reverse transcription reactions using the First-Strand III cDNA synthesis kit (Invitrogen). Each reaction used 0.2 µg RNA, 100 units of the SuperScript III reverse

transcriptase, 1 µL of 100 µM random hexamer primers (Thermo Fisher Scientific) and 1 µL of 10 mM dNTP mix (Bioline). After a 5 min annealing period at 25 °C, reverse transcription was conducted at 50 °C for 30 min, then 55°C for 30 min. The reverse transcriptase was then heat inactivated at 70 °C for 15 min.

Quantitative PCR

The oligonucleotides used for qPCR were designed to generate amplicons that were 150-200 bp in length, and are listed in Table S4. Template cDNA was diluted in a 1:25 ratio with nuclease-free water (Qiagen). Each 15 µL reaction used 3.75 µL of DNA template (either genomic DNA or cDNA), 0.75 µL of 10mM forward and reverse primer solutions (Invitrogen), 7.5 µL of PowerUp™ SYBR™ Green Master Mix (Thermo Fisher Scientific) and 2.25 µL of DNase-free and RNase-free water. Reactions were run using MicroAmp™ Optical 96-well reaction plates (Thermo Fisher Scientific) and the QuantStudio™ 7 Flex Real-Time PCR System (Applied Biosystems). The mixtures were initially heated to 50 °C for 2 min. Subsequently, 40 amplification cycles were run, each consisting of denaturation at 95 °C for 15 s, followed by annealing and elongation at 60 °C for 1 min. The purity of the amplicon at the end of the reaction was assessed through a melt curve analysis.

The expression of $PRCI_{malA}$ genes was quantified using the ΔΔCt method. The *rpoA* gene was used as the reference gene, against which target gene expression levels were normalised to yield a ΔCt value [12]. In all experiments, three technical replicate measurements were made for each of three biological replicates. The ΔCt values were calculated as the difference between the mean Ct of the target gene and the reference gene across technical replicates within the same biological

replicate. The ΔΔCt values were then calculated relative to the reference ΔCt, which was that of $RMV8_{rare}$ at an OD_{600} of 0.2. The fold gene expression difference between genotypes was then quantified as $2^{-\Delta\Delta Ct}$.

For the quantification of PRCI dynamics, the amplicons produced across the integration site, and the *rpoA* reference gene, were generated from genomic DNA (Table S2). The concentration of each purified PCR product was measured using a Qubit Broad Range Kit and Qubit 4 Fluorometer (Thermo Fisher). The Ct values for copy numbers of $3x10^3$, $3x10^4$, $3x10^5$ and $3x10^6$ of these products were used to generate a standard curve to convert the Ct values from experiments into absolute copy numbers of DNA molecules.

Construction of Hi-C and Pore-C sequencing libraries

Three independent cultures of each of the four variants (*S. pneumoniae* $RMV7_{domi}$, $RMV7_{rare}$, $RMV8_{domi}$ and $RMV8_{rare}$) were grown to mid-exponential phase, corresponding to an OD_{600} of ~0.60. The samples processed for Illumina Hi-C and Pore-C analyses were cultured independently. To crosslink DNA and proteins, cells were mixed with 1% v/v formaldehyde (Sigma-Aldrich) for 30 min at RT, followed by a 30 min incubation at 4ºC. The formaldehyde was then quenched through the addition of 1 ml 2.5 M glycine (Sigma-Aldrich), which was incubated at RT for 5 min, then at 4ºC for 15 min. Fixed cells were then collected by centrifugation at 3220 *g* for 10 min at 4ºC, flash frozen on dry ice, and stored at -80 ºC, to permit consistent processing of all samples in parallel.

Frozen pellets were thawed on ice, washed with 5 ml water, and centrifuged at 3220 *g* for 10 min. Pellets were then resuspended in 50 µl 30 mg ml^{-1} lysozyme, followed by incubation at 37ºC for 30 min. Both 10% v/v sodium dodecyl sulphate (SDS; Fisher Bioreagents) and 10% v/v Triton-X (Sigma-Aldrich) were added, followed by 10 min RT shaking incubation, then 10 min incubation at 55ºC. A 400 µl sample of the lysed cells was then combined with 130 µl of digestion mix, comprising NEB CutSmart buffer (NEB), 10% v/v Triton-X, and 1 U $µl^{-1}$ of either *Nla*III (NEB; Illumina Hi-C library preparation), or *MluC*I (NEB; Pore-C library preparation). The chromatin was digested at 37ºC for 3 h. The *Nla*III enzyme was then heat inactivated through a 65ºC incubation for 20 min. The less stable *Mlu*CI enzyme did not require inactivation.

Digested chromatin was then ligated through the addition of 460 µL ligation mix, comprising ligation buffer (NEB), 0.1 µg $µl^{-1}$ recombinant albumin (Promega), and 20 U $µl^{-1}$ T4 DNA ligase (NEB). This mixture was incubated at RT for 1 h, followed by a 4ºC incubation for 48 h. After ligation, the DNA-protein crosslinks were reversed through incubation of the mixture with 500 µl 20% Tween-20 (Sigma-Aldrich), 100 µl 10% v/v SDS, and 100 µl 20 µg $µl^{-1}$ proteinase K (Sigma-Aldrich) at 65ºC overnight. For the Pore-C library preparation, another round of protein denaturation was performed by adding 118 µl 20 mg ml^{-1} pronase (Sigma-Aldrich) and 235 µl 10% v/v SDS, followed by a 63ºC incubation for 1 h.

Following de-cross-linking, DNA was purified with 2.5 ml of a chilled 25:24:1 ratio mixture of phenol:chloroform:isoamyl alcohol (Sigma-Aldrich), followed by separation by centrifugation at 3220 *g* for 30 min at 4ºC. DNA was then precipitated from the

aqueous phase of each sample using 20 µl 5 M NaCl (Sigma Aldrich), 200 µl 3 M sodium acetate (EMD Millipore Corporation) and 6 ml isopropanol (Sigma-Aldrich). These mixtures were incubated at -80ºC overnight, then centrifuged at 3220 *g* for 30 min at 4ºC. The resulting pellets were washed with 1 ml of chilled 70% ethanol (VWR Chemicals BDH), followed by a second centrifugation (3220 *g*, 15 min, 4ºC). All remaining ethanol was removed through pipetting and drying at 65ºC heat block for 15 min. The DNA was the resuspended in 100 µl nuclease-free water.

To enrich for longer molecules likely to contain multiple ligated DNA fragments, a 100 µl sample of each preparation was mixed with 50 µl of KAPA Pure Beads (Roche) through vortexing, followed by incubation together at RT for 10 min. The beads were then magnetically captured, and the supernatant discarded. The beads were washed twice using 200 µl 80% ethanol and dried at RT for 4 min. DNA was then eluted from the beads through a 10 min incubation in 100 µl of pre-heated nuclease-free water at 37ºC. The concentration and size distribution of the DNA molecules was checked using an Agilent 2200 Tape Station (Agilent Technologies) and NanoDrop 1000 (Thermo Fisher).

DNA sequencing

Illumina Hi-C data were generated at the Wellcome Sanger Institute. Libraries were prepared using standard approaches [65], enriching for a fragment size of approximately 450 bp. Libraries corresponding to the three biological replicates for each of the four variants were then combined into a 12-plex library, which was sequenced on a single lane of the Illumina Novaseq 6000 platform. This generated 152 nt paired end reads.

Pore-C data were generated at Imperial College London using the native 12-plex barcoding library preparation kit (Oxford Nanopore Technologies, ONT; product SQK-NBD114.24) to process ~400 ng of DNA from each of the three biological replicates for each of the four variants. Each DNA sample was individually treated to ensure fragments were blunt ended with the NEBNext FFPE DNA Repair Mix (NEB) and the NEBNext Ultra II End Repair/dA-tailing Module (NEB). Samples then underwent native barcode ligation with NEB Blunt/TA Ligase Master Mix, before being pooled together. The NEBNext Quick Ligation Module (NEB) was then used to attach sequencing adapters. The completed libraries were then purified from the mixture using the short fragment buffer to obtain DNA fragments of all sizes. The concentration of the final DNA library was measured on a Qubit 4 fluorometer (Thermo Fisher). Two flow cells (ONT product FLO-MIN114) were each used to sequence 15 fmol of the final library using the MinION platform, recording data with a minimum quality score of 9 using high-accuracy basecalling with Guppy version 6.4.6 (ONT).

Bioinformatic processing

The Illumina Hi-C data were processed using version 2.1.0 of the hic workflow [66] implemented using Nextflow version 24.02.0 [67]. This workflow filtered the Illumina reads using FastQC [68], then mapped reads to the appropriate reference genome (Table S1) using bowtie2 [69], and generated raw contact maps using iced, as implemented within the HiC-Pro pipeline [70]. Genome-wide contact maps between loci were then generated at resolutions of 100, 500, 1000, 5000, 10000 and 50000

bp using cooler [71]. A quality control report was then generated with MultiQC [72], with additional visualisation files produced with juicer [73].

The Pore-C data were processed using version 1.1.0 of the wf-pore-c workflow [74] implemented using Nextflow version 23.10.1 [67]. This workflow mapped reads to the appropriate reference genome (Table S1) using minimap2 [75] with the non-default settings "-x map-ont -k 7 -w 5". These alignments were then processed with Pore-c-py [74] and pairtools [39] to generate raw contact maps, which were processed with cooler [71] and juicer [73], as for the Hi-C data.

Differential analysis of contact matrices

Analyses of differential contacts used matrices calculated at a resolution of 10 kb, with all self interactions along the matrix diagonals removed, and the *ivr* and *tvr* loci excluded, as they contained known rearrangements.

The diffHiC package [40] was applied using the recommended workflow [76], with modifications to parameter settings as appropriate for these datasets. Loci with fewer than five contacts across all replicates in each comparison were first removed, as there was low power to identify differential contacts at such positions. The counts across the filtered dataset were then normalised across replicates within each comparison between variants using a cyclic LOESS process, run for ten iterations. The biological coefficient of variation was then calculated across different contact densities, which informed the dispersion of the negative binomial generalised linear model that was fitted to the data. This used a tagwise, rather than trended, dispersion calculation, as there was not a strong relationship between variance and

contact densities in some datasets. Differential contact densities between loci in the two variants were then identified from this model using quasi F-tests. The resulting *p* values were subject to a Benjamini-Hochberg correction for multiple testing to generate *q* values.

Similarly, analysis with the multiHiCcompare package [41] used a modified version of the recommended workflow [77]. This started by first both removing any loci for which the minimum frequency across replicates was zero. The remaining data were normalised using a cyclic LOESS process run for ten iterations. Significant differences were inferred through using a negative binomial exact test, as implemented in edgeR [78], with a Benjamini-Hochberg correction for multiple testing.

Statistical modelling of contact densities

The 10 kb resolution contact matrices were analysed using a generalised linear mixed effects model with the structure:

$$\begin{gathered}\text{Contact frequency} \sim \text{Distance category} + \\ \text{Distance from origin} + \text{GC} + Mlu\text{CI} + Nla\text{III} + \\ \text{Locus classification} * \text{Variant} + \\ (1|\text{Replicate}) + (\text{Variant}|\text{Position})\end{gathered}$$

The terms all represent fixed effects, except those with the '|' notation, which denotes a grouping variable for a random effect. Asterisks indicate an interaction between fixed effects. A zero-inflated logarithmic Poisson link function was used to fit

the model with the glmmTMB package [79]. Parameter estimates reproducibly converged using the default settings. However, attempts analyse the data using an equivalent model with a negative binomial link function, which would account for any overdispersion in the data, failed to converge on robust parameter estimates. Analysis of parameter estimates used the sjPlot package [80], model fits were assessed with the performance package [81], and data processing used the tidyverse packages [82].

Each interaction was categorised by the distance between the loci: <12.5 kb (ie., neighbouring 10 kb loci); 12.5-25 kb; 25- 50 kb, and >50 kb. Only loci within 100 kb of each other were considered, to ensure parameter estimation was feasible, and avoid the model fitting primarily to the noisy, sparse interactions between distant loci.

Loci were also classified based on the annotation of the corresponding reference genomes. In both genomes, the four rRNA operons were annotated, as the difficulty of mapping to these relatively GC-rich repeats reduces the detection of contacts in these loci. In RMV7, four MGEs were annotated: PRCI$_{dnaN}$, PRCI$_{uvrA}$, ICE$_{rpsI}$, and Tn*916* [12]. In RMV8, two MGEs were annotated: prophage □RMV8 and PRCI$_{malA}$ [18]. These classifications were fitted separately for each variant, to enable intervariant differences to be inferred at these loci.

As the DNA-protein crosslinks were fixed in mid-exponential phase, the detected contact frequencies were higher near the origin of replication, as some cells would have initiated, but not completed, chromosomal replication. Hence the number of detected contacts was expected to change with distance from the origin.

The additional fixed effects analysed the properties of each DNA locus, which were calculated using biopython [83]. The effect of the distribution of *Nla*III and *Mlu*CI sites was accounted for through calculating the mean number of restriction sites across the two interacting loci, and dividing this value by ten. The GC content was calculated as the mean proportion of bases in the two interacting loci, which was scaled upwards by a factor of ten. Similarly, the distance to the origin was calculated as the proportion of the genome between the site and the origin of replication, and this value was scaled upwards by a factor of ten when included in the model. These adjustments were necessary for numerical stability of the model fitting process.

The identity of the biological replicate was included as a random effect, to correct for differences in the amount of sequence data generated for each dataset. A second random effect estimated the corrected level of contact frequency for each locus in the dominant variant, and the difference in binding to each of these loci in the rare variant.

Data Availability

All raw genetic data used in this study are available from the ENA with the accession codes listed in Table S1 and S2. Processed files generated by bioinformatic pipelines are available from FigShare (https://figshare.com/account/home#/projects/227616). Experimental results and sample information are available from the Github repository https://github.com/nickjcroucher/RMV_PoreC.

Code Availability

All code used in this analysis is available from the Github repository https://github.com/nickjcroucher/RMV_PoreC.

Acknowledgements

We thank Prof. Juanma Vaquerizas for helpful discussions about Hi-C data generation and analysis, and the Bespoke team at the Wellcome Sanger Institute for generating the Hi-C sequencing libraries. This work was supported by a Sir Henry Dale fellowship jointly funded by Wellcome and the Royal Society (grant 104169/Z/14/A), Wellcome (grant 206194), the European Molecular Biology Laboratory, and the UK Medical Research Council and Department for International Development (grant MR/R015600/1).

Competing interests

NJC has consulted for Antigen Discovery Inc., Merck and Pfizer, and has received an investigator-initiated award from GlaxoSmithKline. S.D.B. has consulted for Pfizer and Merck.

References

1. Croucher NJ, Løchen A, Bentley SD. Pneumococcal Vaccines: Host Interactions, Population Dynamics, and Design Principles. Annu Rev Microbiol. 2018;72: 521–549. doi:10.1146/annurev-micro-090817-062338
2. Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, Powell E, et al. Comparative Genomic Analyses of Seventeen *Streptococcus pneumoniae* Strains: Insights into the Pneumococcal Supragenome. J Bacteriol. 2007;189: 8186 LP – 8195. doi:10.1128/JB.00690-07
3. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli S V, et al. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol. 2010;11: R107–R107. doi:10.1186/gb-2010-11-10-r107
4. Croucher N, Coupland PG, Stevenson AE, Callendrello A, Bentley SD, Hanage WP. Diversification of bacterial genome content through distinct mechanisms over different timescales. Nat Commun. 2014;5: 5471. doi:10.1038/ncomms6471
5. D'Aeth JC, van der Linden MPG, McGee L, de Lencastre H, Turner P, Song J-H, et al. The role of interspecies recombination in the evolution of antibiotic-resistant pneumococci. Elife. 2021;10: e67113. doi:10.7554/eLife.67113
6. Johnston C, Martin B, Granadel C, Polard P, Claverys JP. Programmed Protection of Foreign DNA from Restriction Allows Pathogenicity Island Exchange during Pneumococcal Transformation. PLoS Pathog. 2013;9: e1003178. doi:10.1371/journal.ppat.1003178
7. Kwun MJ, Oggioni MR, De Ste Croix M, Bentley SD, Croucher NJ. Excision-reintegration at a pneumococcal phase-variable restriction-modification locus

drives within- and between-strain epigenetic differentiation and inhibits gene acquisition. Nucleic Acids Res. 2018;46: 11438–11453. doi:10.1093/nar/gky906

8. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, et al. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science (80-). 2001;293: 498–506. doi:10.1126/science.1061217

9. Manso AS, Chai MH, Atack JM, Furi L, De Ste Croix M, Haigh R, et al. A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. Nat Commun. 2014;5: 5055. doi:10.1038/ncomms6055

10. Kwun MJ, Oggioni MR, Bentley SD, Fraser C, Croucher NJ. Synergistic Activity of Mobile Genetic Element Defences in *Streptococcus pneumoniae*. Genes (Basel). 2019;10: 707.

11. Furi L, Crawford LA, Rangel-Pineros G, Manso AS, De Ste Croix M, Haigh RD, et al. Methylation warfare: interaction of pneumococcal bacteriophages with their host. J Bacteriol. 2019; JB.00370-19. doi:10.1128/JB.00370-19

12. Kwun MJ, Ion A V, Oggioni MR, Bentley SD, Croucher NJ. Diverse regulatory pathways modulate bet hedging of competence induction in epigenetically-differentiated phase variants of *Streptococcus pneumoniae*. Nucleic Acids Res. 2023;51: 10375–10394. doi:10.1093/nar/gkad760

13. Li J, Li JW, Feng Z, Wang J, An H, Liu Y, et al. Epigenetic Switch Driven by DNA Inversions Dictates Phase Variation in *Streptococcus pneumoniae*. PLoS Pathog. 2016;12: e1005762. doi:10.1371/journal.ppat.1005762

14. Durmort C, Ercoli G, Ramos-Sevillano E, Chimalapati S, Haigh RD, Croix MDS, et al. Deletion of the Zinc Transporter Lipoprotein AdcAII Causes

Hyperencapsulation of Streptococcus pneumoniae Associated with Distinct Alleles of the Type I Restriction-Modification System. MBio. 2020;11: 10.1128/mbio.00445-20. doi:10.1128/mbio.00445-20

15. De Ste Croix M, Mitsi E, Morozov A, Glenn S, Andrew PW, Ferreira DM, et al. Phase variation in pneumococcal populations during carriage in the human nasopharynx. Sci Rep. 2020;10: 1803. doi:10.1038/s41598-020-58684-2

16. Phillips ZN, Trappetti C, Van Den Bergh A, Martin G, Calcutt A, Ozberk V, et al. Pneumococcal Phasevarions Control Multiple Virulence Traits, Including Vaccine Candidate Expression. Microbiol Spectr. 2022;10: e0091622. doi:10.1128/spectrum.00916-22

17. Agnew HN, Atack JM, Fernando ARD, Waters SN, van der Linden M, Smith E, et al. Uncovering the link between the SpnIII restriction modification system and LuxS in *Streptococcus pneumoniae* meningitis isolates. Front Cell Infect Microbiol. 2023;13: 1177857. doi:10.3389/fcimb.2023.1177857

18. Kwun MJ, Ion A V, Cheng H-C, D'Aeth JC, Dougan S, Oggioni MR, et al. Post-vaccine epidemiology of serotype 3 pneumococci identifies transformation inhibition through prophage-driven alteration of a non-coding RNA. Genome Med. 2022;14: 144. doi:10.1101/2022.09.21.508813

19. Buitrago D, Labrador M, Arcon JP, Lema R, Flores O, Esteve-Codina A, et al. Impact of DNA methylation on 3D genome structure. Nat Commun. 2021;12: 3243. doi:10.1038/s41467-021-23142-8

20. Kwun MJ, Ion A V, Apagyi KJ, Croucher NJ. Chromosomal curing drives prophage evolution and bacterial epidemiology. bioRxiv. 2024; 2024.08.13.607808. doi:10.1101/2024.08.13.607808

21. Duan B, Ding P, Navarre WW, Liu J, Xia B. Xenogeneic Silencing and

Bacterial Genome Evolution: Mechanisms for DNA Recognition Imply Multifaceted Roles of Xenogeneic Silencers. Mol Biol Evol. 2021;38: 4135–4148. doi:10.1093/molbev/msab136

22. Dorman CJ. H-NS, the genome sentinel. Nat Rev Microbiol. 2006;5: 157–161. doi:10.1038/nrmicro1598

23. Navarre WW. The Impact of Gene Silencing on Horizontal Gene Transfer and Bacterial Evolution. Advances in Microbial Physiology. Elsevier; 2016. pp. 157–186. doi:10.1016/bs.ampbs.2016.07.004

24. Wiechert J, Filipchyk A, Hünnefeld M, Gätgens C, Brehm J, Heermann R, et al. Deciphering the Rules Underlying Xenogeneic Silencing and Counter-Silencing of Lsr2-like Proteins Using CgpS of *Corynebacterium glutamicum* as a Model. MBio. 2020;11: e02273-19. doi:10.1128/mBio.02273-19

25. Navarre WW, McClelland M, Libby SJ, Fang FC. Silencing of xenogeneic DNA by H-NS—facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. Genes Dev . 2007;21: 1456–1471. doi:10.1101/gad.1543107

26. Seah NE, Warren D, Tong W, Laxmikanthan G, Van Duyne GD, Landy A. Nucleoprotein architectures regulating the directionality of viral integration and excision. Proc Natl Acad Sci U S A. 2014/08/11. 2014;111: 12372–12377. doi:10.1073/pnas.1413019111

27. Perez-Rueda E, Ibarra JA. Distribution of putative xenogeneic silencers in prokaryote genomes. Comput Biol Chem. 2015;58: 167–172. doi:10.1016/j.compbiolchem.2015.06.007

28. Winardhi RS, Fu W, Castang S, Li Y, Dove SL, Yan J. Higher order oligomerization is required for H-NS family member MvaT to form gene-

silencing nucleoprotein filament. Nucleic Acids Res. 2012/07/13. 2012;40: 8942–8952. doi:10.1093/nar/gks669

29. Qu Y, Lim CJ, Whang YR, Liu J, Yan J. Mechanism of DNA organization by *Mycobacterium tuberculosis* protein Lsr2. Nucleic Acids Res. 2013/04/10. 2013;41: 5263–5272. doi:10.1093/nar/gkt249

30. Tendeng C, Soutourina OA, Danchin A, Bertin PN. MvaT proteins in Pseudomonas spp.: a novel class of H-NS-like proteins. Microbiology. 2003;149: 3047–3050. doi:10.1099/mic.0.c0125-0

31. Wang W, Li G-W, Chen C, Xie XS, Zhuang X. Chromosome organization by a nucleoid-associated protein in live bacteria. Science. 2011;333: 1445–1449. doi:10.1126/science.1204697

32. Seid CA, Smith JL, Grossman AD. Genetic and biochemical interactions between the bacterial replication initiator DnaA and the nucleoid-associated protein Rok in *Bacillus subtilis*. Mol Microbiol. 2017/01/03. 2017;103: 798–817. doi:10.1111/mmi.13590

33. Solano-Collado V, Hüttener M, Espinosa M, Juárez A, Bravo A. Mga*Spn* and H-NS: Two Unrelated Global Regulators with Similar DNA-Binding Properties. Front Mol Biosci. 2016;3: 60. doi:10.3389/fmolb.2016.00060

34. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing Chromosome Conformation. Science (80-). 2002;295: 1306–1311. doi:10.1126/science.1067799

35. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326: 289–293. doi:10.1126/science.1181369

36. Ulahannan N, Pendleton M, Deshpande A, Schwenk S, Behr JM, Dai X, et al. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. bioRxiv. 2019. doi:10.1101/833590

37. Li Z, Long Y, Yu Y, Zhang F, Zhang H, Liu Z, et al. Pore-C simultaneously captures genome-wide multi-way chromatin interaction and associated DNA methylation status in Arabidopsis. Plant Biotechnol J. 2022;20: 1009–1011. doi:10.1111/pbi.13811

38. Zhong J-Y, Niu L, Lin Z-B, Bai X, Chen Y, Luo F, et al. High-throughput Pore-C reveals the single-allele topology and cell type-specificity of 3D genome folding. Nat Commun. 2023;14: 1250. doi:10.1038/s41467-023-36899-x

39. Open2C, Abdennur N, Fudenberg G, Flyamer IM, Galitsyna AA, Goloborodko A, et al. Pairtools: From sequencing data to chromosome contacts. PLOS Comput Biol. 2024;20: e1012164. Available: https://doi.org/10.1371/journal.pcbi.1012164

40. Lun ATL, Smyth GK. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. BMC Bioinformatics. 2015;16: 258. doi:10.1186/s12859-015-0683-0

41. Stansfield JC, Cresswell KG, Dozmorov MG. multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments. Bioinformatics. 2019;35: 2916–2923. doi:10.1093/bioinformatics/btz048

42. Stansfield JC, Tran D, Nguyen T, Dozmorov MG. R Tutorial: Detection of Differentially Interacting Chromatin Regions From Multiple Hi-C Datasets. Curr Protoc Bioinforma. 2019;66: e76. doi:10.1002/cpbi.76

43. Browne PD, Nielsen TK, Kot W, Aggerholm A, Gilbert MTP, Puetz L, et al. GC bias affects genomic and metagenomic reconstructions, underrepresenting

GC-poor organisms. Gigascience. 2020;9: giaa008. doi:10.1093/gigascience/giaa008

44. Ali SS, Xia B, Liu J, Navarre WW. Silencing of foreign DNA in bacteria. Curr Opin Microbiol. 2012;15: 175–181. doi:https://doi.org/10.1016/j.mib.2011.12.014

45. Wang S, Lee S, Chu C, Jain D, Kerpedjiev P, Nelson GM, et al. HiNT: a computational method for detecting copy number variations and translocations from Hi-C data. Genome Biol. 2020;21: 73. doi:10.1186/s13059-020-01986-5

46. Newby J, Chapman J. Metastable behavior in Markov processes with internal states. J Math Biol. 2014;69: 941–976. doi:10.1007/s00285-013-0723-1

47. Martínez-Rubio R, Quiles-Puchalt N, Martí M, Humphrey S, Ram G, Smyth D, et al. Phage-inducible islands in the Gram-positive cocci. ISME J. 2017;11: 1029–1042. doi:10.1038/ismej.2016.163

48. Sung CK, Li H, Claverys JP, Morrison DA. An *rpsL* cassette, Janus, for gene replacement through negative selection in *Streptococcus pneumoniae*. Appl Environ Microbiol. 2001;67: 5190–5196.

49. Marbouty M, Le Gall A, Cattoni DI, Cournac A, Koh A, Fiche J-B, et al. Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. Mol Cell. 2015;59: 588–602. doi:10.1016/j.molcel.2015.07.020

50. Espeli O, Mercier R, Boccard F. DNA dynamics vary according to macrodomain topography in the *E. coli* chromosome. Mol Microbiol. 2008;68: 1418–1427. doi:10.1111/j.1365-2958.2008.06239.x

51. Lioy VS, Cournac A, Marbouty M, Duigou S, Mozziconacci J, Espéli O, et al. Multiscale Structuring of the *E. coli* Chromosome by Nucleoid-Associated and

Condensin Proteins. Cell. 2018;172: 771-783.e18. doi:10.1016/j.cell.2017.12.027

52. Le TBK, Imakaev M V, Mirny LA, Laub MT. High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. Science (80-). 2013;342: 731–734. doi:10.1126/science.1242059

53. Trussart M, Yus E, Martinez S, Baù D, Tahara YO, Pengo T, et al. Defined chromosome structure in the genome-reduced bacterium *Mycoplasma pneumoniae*. Nat Commun. 2017;8: 14665. doi:10.1038/ncomms14665

54. Croucher NJ, Walker D, Romero P, Lennard N, Paterson GK, Bason NC, et al. Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae*^{Spain23F} ST81. J Bacteriol. 2008/12/31. 2009;191: 1480–1489. doi:10.1128/jb.01343-08

55. Rode CK, Melkerson-Watson LJ, Johnson AT, Bloch CA. Type-specific contributions to chromosome size differences in *Escherichia coli*. Infect Immun. 1999;67: 230–236. doi:10.1128/IAI.67.1.230-236.1999

56. Wasim A, Gupta A, Bera P, Mondal J. Interpretation of organizational role of proteins on *E. coli* nucleoid via Hi-C integrated model. Biophys J. 2022/11/25. 2023;122: 63–81. doi:10.1016/j.bpj.2022.11.2938

57. Swetha RG, Sekar DKK, Devi ED, Ahmed ZZ, Ramaiah S, Anbarasu A, et al. *Streptococcus pneumoniae* Genome Database (SPGDB): A database for strain specific comparative analysis of *Streptococcus pneumoniae* genes and proteins. Genomics. 2014;104: 582–586. doi:10.1016/j.ygeno.2014.09.012

58. Ferrándiz M-J, Carreño D, Ayora S, de la Campa AG. HU of *Streptococcus pneumoniae* Is Essential for the Preservation of DNA Supercoiling. Front Microbiol. 2018;9: 493. doi:10.3389/fmicb.2018.00493

59. Dorman CJ. H-NS-like nucleoid-associated proteins, mobile genetic elements and horizontal gene transfer in bacteria. Plasmid. 2014;75: 1–11. doi:https://doi.org/10.1016/j.plasmid.2014.06.004

60. Denamur E, Clermont O, Bonacorsi S, Gordon D. The population genetics of pathogenic *Escherichia coli*. Nat Rev Microbiol. 2021;19: 37–54. doi:10.1038/s41579-020-0416-x

61. Zawadzki P, Riley MA, Cohan FM. Homology among nearly all plasmids infecting three Bacillus species. J Bacteriol. 1996;178: 191–198. doi:10.1128/jb.178.1.191-198.1996

62. Veening J-W, Smits WK, Kuipers OP. Bistability, epigenetics, and bet-hedging in bacteria. Annu Rev Microbiol. 2008;62: 193–210. doi:10.1146/annurev.micro.62.081307.163002

63. Croucher NJ, Mostowy R, Wymant C, Turner P, Bentley SD, Fraser C. Horizontal DNA Transfer Mechanisms of Bacteria as Weapons of Intragenomic Conflict. Barton NH, editor. PLOS Biol. 2016;14: e1002394. doi:10.1371/journal.pbio.1002394

64. Martins BMC, Locke JCW. Microbial individuality: how single-cell heterogeneity enables population level strategies. Curr Opin Microbiol. 2015;24: 104–112. doi:https://doi.org/10.1016/j.mib.2015.01.003

65. Bronner IF, Quail MA. Best Practices for Illumina Library Preparation. Curr Protoc Hum Genet. 2019;102: e86. doi:10.1002/cphg.86

66. Servant N, Ewels P, Garcia MU, Talbot A, Peltzer A, Miller E, et al. nf-core/hic: nf-core/hic v2.1.0. Zenodo; 2023. doi:10.5281/zenodo.7994878

67. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol.

2017;35: 316–319. doi:10.1038/nbt.3820

68. Andrews S. FastQC. 2023.

69. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359. doi:10.1038/nmeth.1923

70. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;16: 259. doi:10.1186/s13059-015-0831-x

71. Abdennur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics. 2020;36: 311–316. doi:10.1093/bioinformatics/btz540

72. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32: 3047–3048. doi:10.1093/bioinformatics/btw354

73. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. 2016;3: 95–98. doi:10.1016/j.cels.2016.07.002

74. EPI2ME. wf-pore-c. 2024.

75. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34: 3094–3100. doi:10.1093/bioinformatics/bty191

76. Lun ATL. DiffHiC: Differential analysis of Hi-C data User's Guide. 2017.

77. Stansfield JC, Dozmorov MG. multiHiCcompare Vignette. 2024.

78. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26: 139–140. doi:10.1093/bioinformatics/btp616

79. Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen

A, et al. {glmmTMB} Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. R J. 2017;9: 378–400. doi:10.32614/RJ-2017-066

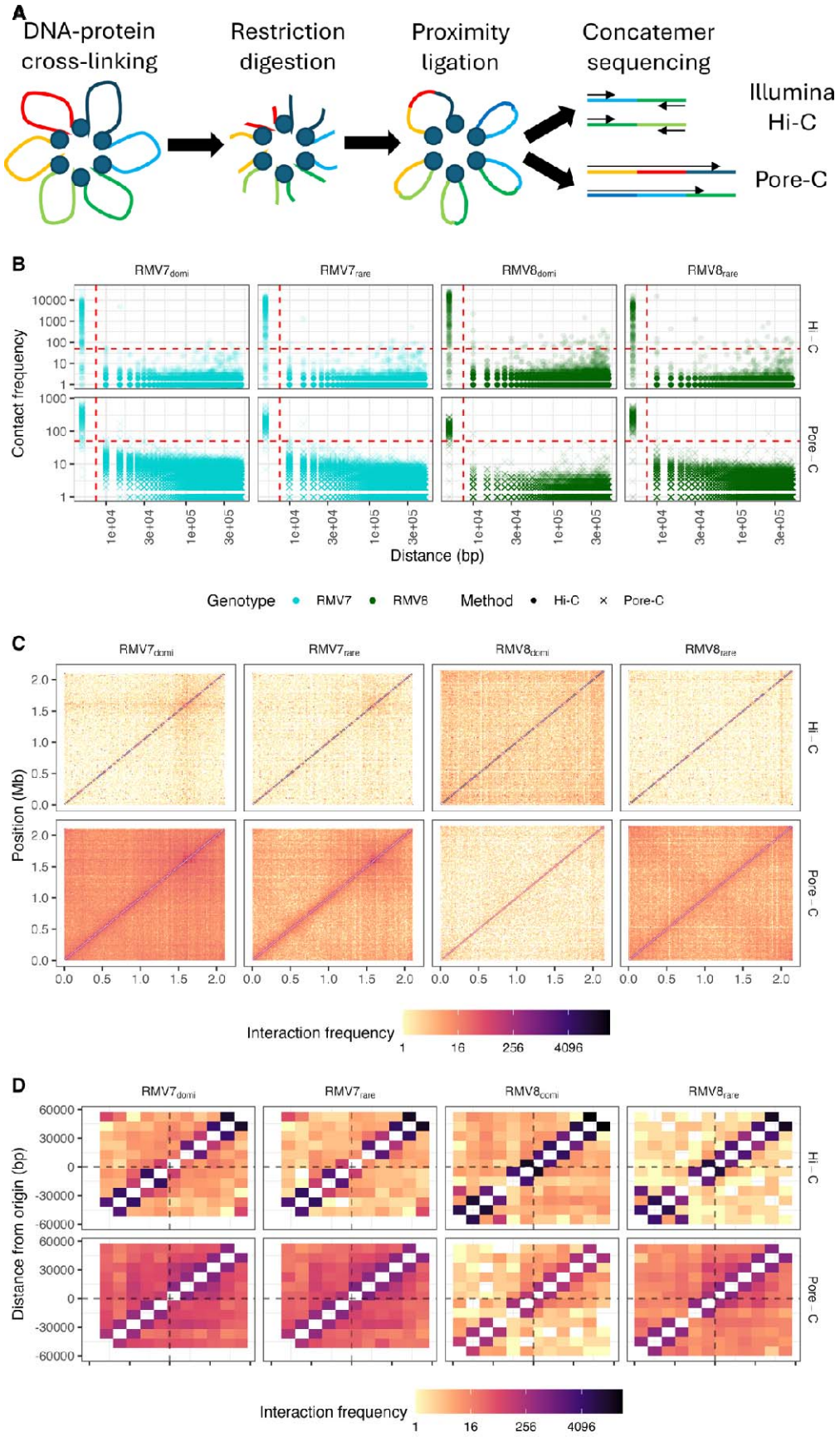
80. Lüdecke D. sjPlot: Data Visualization for Statistics in Social Science, R package version 2.4.0. R package version 2.4.0. 2017.

81. Lüdecke D, Ben-Shachar MS, Patil I, Waggoner P, Makowski D. {performance}: An {R} Package for Assessment, Comparison and Testing of Statistical Models. J Open Source Softw. 2021;6: 3139. doi:10.21105/joss.03139

82. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 2019;4: 1686. doi:10.21105/joss.01686

83. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25: 1422–1423. doi:10.1093/bioinformatics/btp163

Figures


A
DNA-protein cross-linking
Restriction digestion
Proximity ligation
Concatemer sequencing
Illumina Hi-C
Pore-C
B
RMV7domi
RMV7rare
RMV8domi
RMV8rare
Contact frequency
Hi-C
Pore-C
Distance (bp)
Genotype
RMV7
RMV8
Method
Hi-C
Pore-C
C
Position (Mb)
Interaction frequency
1
16
256
4096
D
Distance from origin (bp)
Interaction frequency
1
16
256
4096

1

Figure 1 Chromosomal conformation analysis of the epigenetic variants of two pneumococcal genotypes. (A) Diagram describing the methodology underpinning the use of high-throughput sequencing to capture chromatin conformation (3C). The final sequencing step distinguishes the Illumina (Hi-C) and Oxford Nanopore Technology (Pore-C) methods. **(B)** Relationship between the frequency of contacts between loci and their separation in the genome. Each panel represents a unique combination of sequencing technology and epigenetic variant. Each point corresponds to a detected pairwise interaction between loci in a contact matrix calculated at a 5 kb resolution for both technologies. The red dashed lines represent the thresholds used to define the closely-spaced loci between which particularly high contact densities were inferred. **(C)** Genome-wide Hi-C and Pore-C contact matrices shown at a 10 kb resolution. Each panel displays the cumulative number of contacts identified in three biological replicates for each unique combination of sequencing technology and epigenetic variant. The frequency of detected interactions between loci across the genome is represented by the colour of the corresponding cell in these symmetrical matrices. **(D)** Zoomed in matrix showing the cumulative number of contacts identified in loci surrounding the origin of replication, the position of which is marked by dashed lines. Each cell in the matrix represents data from a 10 kb locus.

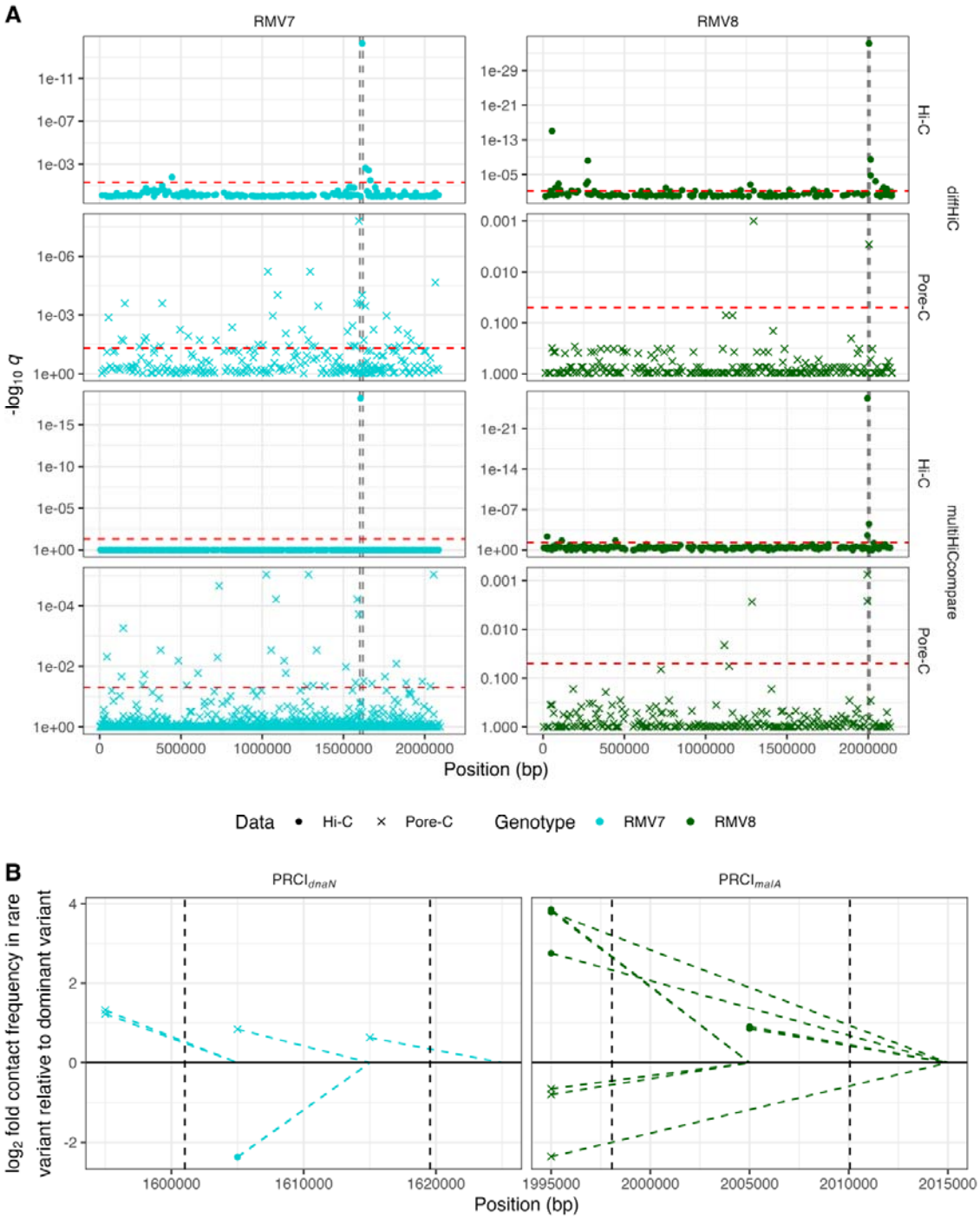


Figure 2 Genome-wide analysis of differential contact frequencies. (A) Manhattan plots comparing the contact densities between epigenetic variants. For each genotype, two different statistical methods (diffHiC and multiHiCCompare) were used to compare the three biological replicates generated using each technology (Hi-

C and Pore-C) for both epigenetic variants. The negative logarithmic *q* values, corresponding to false discovery rates following a Benjamini-Hochberg correction for multiple testing, are shown for each locus across the genome. The genome-wide level of significance is shown by the red dashed horizontal line. The vertical black dashed lines show the boundaries of the variable PRCIs ($PRCI_{dnaN}$ in RMV7, and $PRCI_{malA}$ in RMV8) within the chromosomes. **(B)** Sites of differential contact frequencies identified by all four statistical comparisons within $PRCI_{dnaN}$ in RMV7, and $PRCI_{malA}$ in RMV8. Each point represents an interaction that was found to be significantly different between the corresponding dominant and rare variants following a correction for multiple testing. Its position on the horizontal axis represents the site of the interacting locus with the lower numerical position. The dashed line connects the point to the interacting locus with the higher numerical position. The vertical position of the point represents the base two logarithmic fold difference in the rare variant relative to the dominant variant, such that values greater than zero represent a higher contact frequency in the rare variant. The shapes of the points represent the method used to generate the data.

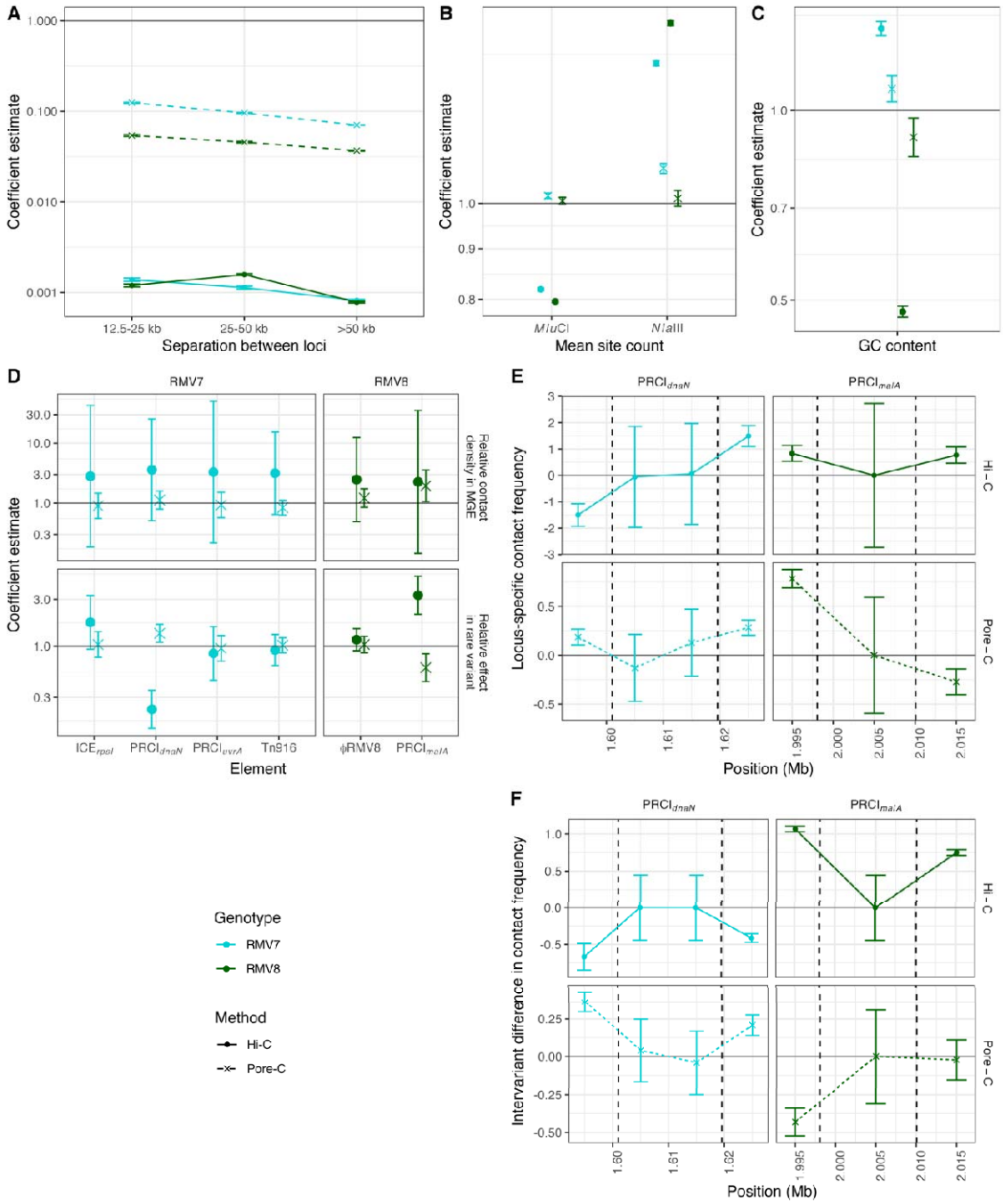


Figure 3 Statistical modelling of contact frequencies. A zero-inflated Poisson generalised linear mixed effects model was fitted to the four combinations of genotype and sequencing technology using contact matrices calculated at a resolution of 10 kb. The two genotypes are distinguished by colours, and the two technologies are distinguished by line and point styles. **(A)** Coefficients estimating

the effect of separation between loci on contact frequencies, normalised to the expected contact frequency between neighbouring loci, shown by the horizontal line at a value of one. **(B)** Coefficients estimating the effect of *Mlu*CI and *Nla*III site densities on the detection of contacts; these coefficients reflect the impact of the number of sites per 1 kb. The horizontal line at one represents the coefficient value if the density of sites had no effect on the number of inferred contacts. **(C)** Coefficients estimating the effect of the fraction of bases (scaled upwards 10-fold to aid model fitting) that are guanine or cytosine (GC) on contact frequencies. **(D)** Coefficients estimating the contact frequencies within mobile genetic elements normalised to the rest of the chromosome, shown by the horizontal line at a value of one. **(E)** Locus-specific variation in the contact frequencies across the loci encompassing the two PRCIs highlighted in Fig. 2: $PRCI_{dnaN}$ within RMV7, and $PRCI_{malA}$ within RMV8. Data are plotted at 10 kb intervals. The black vertical dashed lines indicate the boundaries of the PRCIs. **(F)** Differences between the variants at the loci shown in (E).

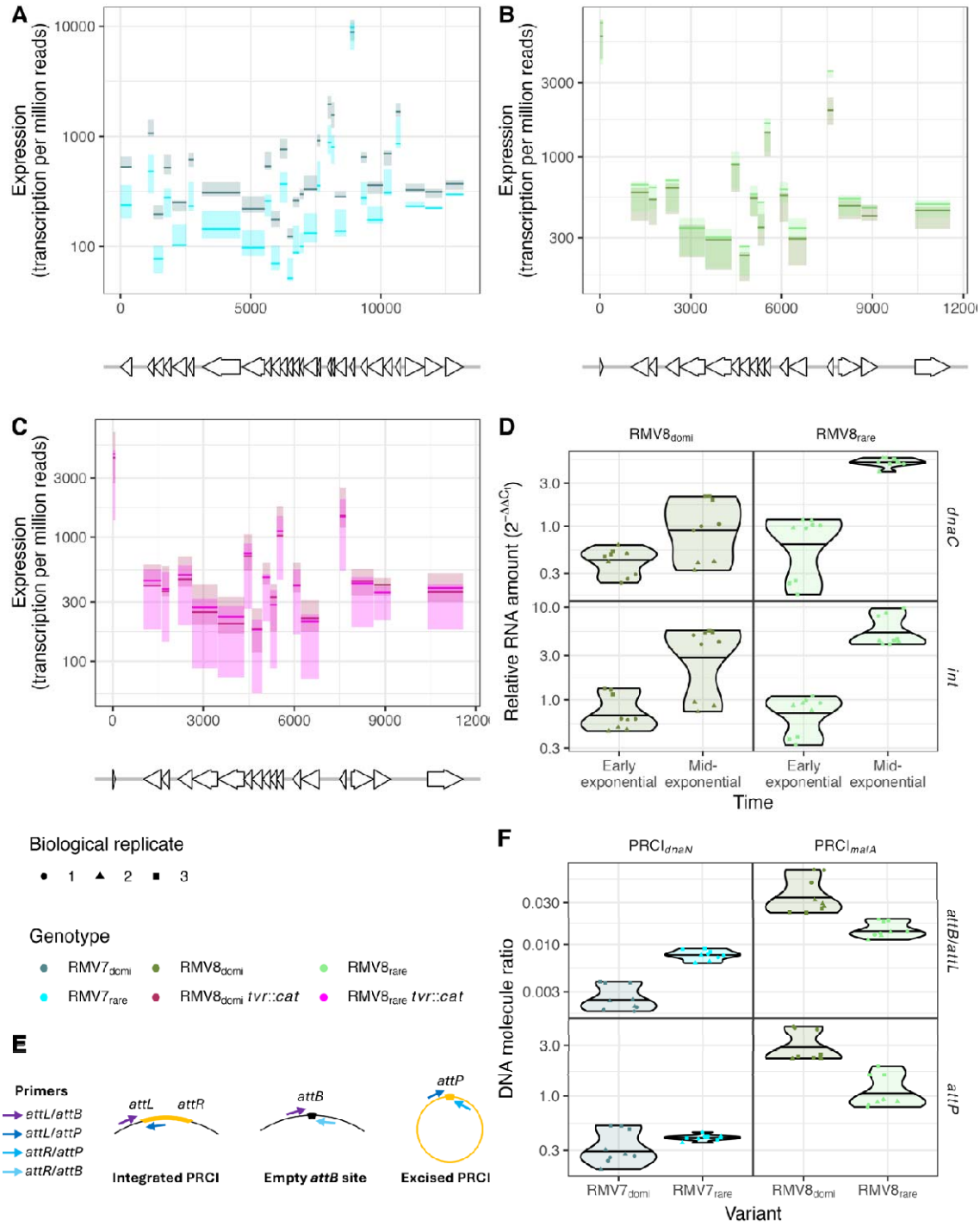


Figure 4 Differential activity of PRCIs between variants. **(A)** Quantification of PRCI$_{dnaN}$ expression in RMV7$_{domi}$ and RMV7$_{rare}$ using previously-published RNA-seq data [12]. The annotated protein coding sequences (CDSs) within the PRCI are

shown by the arrows at the bottom of the panel, with their orientation indicating the direction in which they are transcribed. The lines in the main panel show the level of expression estimated for each CDS for each of the three independent biological replicates as normalised transcripts per million reads. The solid line indicates the median value for each variant, and the shaded area indicates the range between the maximum and minimum values. **(B)** Quantification of PRCI$_{malA}$ expression in RMV7$_{domi}$ and RMV7$_{rare}$ using previously-published RNA-seq data [18], as displayed in panel (A). **(C)** Quantification of PRCI$_{malA}$ expression in RMV7$_{domi}$ *tvr*::*cat* and RMV7$_{rare}$ *tvr*::*cat* using previously-published RNA-seq data [18], as displayed in panel (A). These two mutants do not have functional *Spn*IV RMSs. **(D)** Quantification of PRCI$_{malA}$ expression in RMV7$_{domi}$ and RMV7$_{rare}$ using qRT-PCR assays of transcription of an integrase (*int*) and replication (*dnaC*) CDS. Three technical replicates are shown for each of three biological replicates, as indicated by the shapes of the points. The violin plot summarises their distribution, with the median indicated by the horizontal line. Measurements were made at both early and mid-exponential timepoints. **(E)** Diagram showing the primers used to quantify the integration, excision and replication of the PRCIs. **(F)** Differences in integration, excision and replication of the PRCIs between variants. Data are shown as in panel (D). For both PRCI$_{dnaN}$, in RMV7, and PRCI$_{malA}$, in RMV8, the level of excision was quantified as the ratio of *attB* sites to *attL* sites A higher ratio indicates a greater level of excision. Similarly, the levels of circularised PRCIs were quantified by measuring the concentration of *attP*.

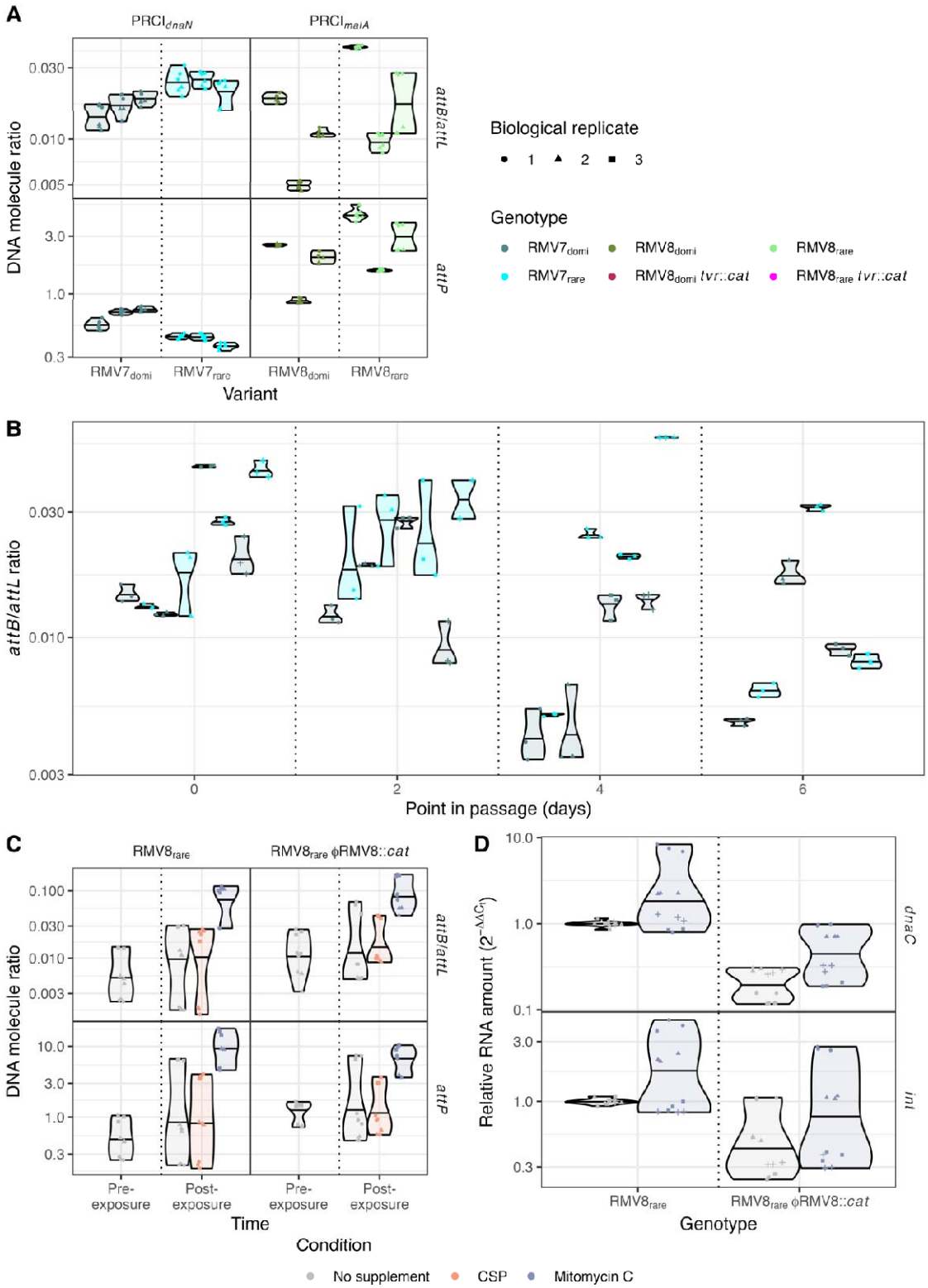


Figure 5 Metastability and inducibility of PRCI activity. **(A)** Violin plot showing the variation in PRCI topology between cultures inoculated with independently-isolated colonies of the four variants. Three biological replicate cultures were grown

from each of three independently-isolates colonies for each variant. Each violin plot represents a single colony, and describes the distribution of three technical replicate measurements for an individual biological replicate, with the shape of each point indicating the biological replicate to which they relate. The horizontal line across the violins represents the median measurement. **(B)** Violin plot showing the variation in PRCI$_{dnaN}$ excision in a six-day passage of RMV7 genotypes. Four independent mutants, in which the RMV8$_{domi}$ or RMV8$_{rare}$ *tvr* locus had been introduced into a *tvr*-null RMV8 genotype, were passaged in parallel [12]. DNA was sampled from a single biological replicate of each mutant at the start of the passage, and after two, four and six days. Each violin summarises three technical replicate measurements of a single biological replicate. **(C)** Violin plot showing the effect of inducing stimuli on the topology of PRCI$_{malA}$. The excision and replication of PRCI$_{malA}$ following exposure to competence stimulating peptide and mitomycin C were assayed two hours after exposure in early exponential growth. These experiments used both RMV8$_{rare}$, and a mutant derivative lacking the □RMV8 prophage. Data are shown as in panel (A). **(D)** Violin plot quantifying the effect of mitomycin C on PRCI gene expression. The expression of the *int* and *dnaC* genes of PRCI$_{dnaN}$ were assayed after two hours of exposure to mitomycin C, in both RMV8$_{rare}$ and a mutant derivative lacking the □RMV8 prophage. Data are shown as in Fig. 4.

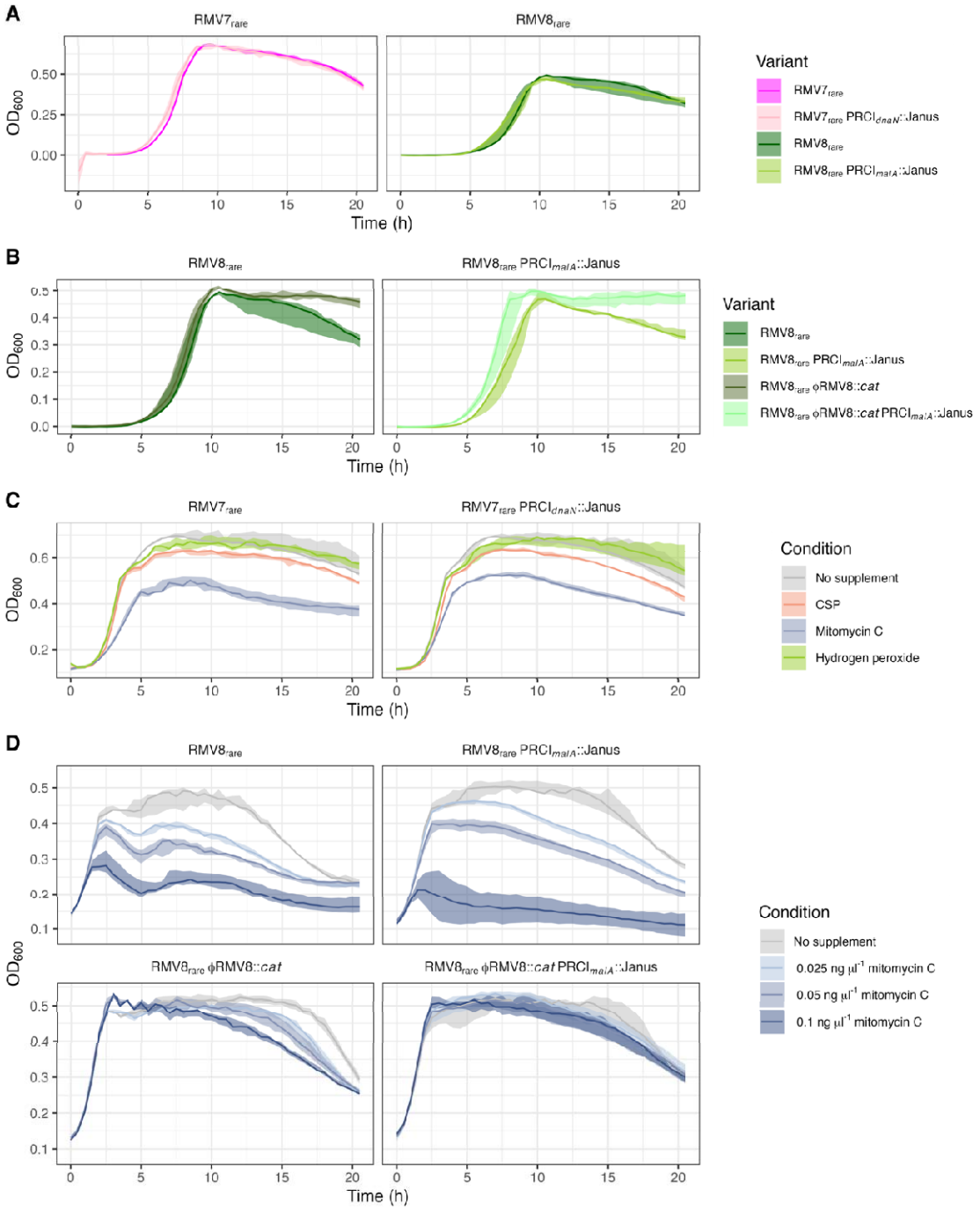


Figure 6 The impact of induced and uninduced MGEs on the growth of RMV7 and RMV8 genotypes. Each panel shows the 20 h growth curve of a set of pneumococcal genotypes, measured through assaying the optical density (OD_{600}) at 30 min intervals. In each plot, the solid line shows the median, and the shaded region shows the full range observed across three biological replicates. **(A)** A

comparison of the effect of removing PRCI$_{dnaN}$ from RMV7$_{rare}$ and removing PRCI$_{malA}$ from RMV8$_{rare}$. **(B)** A comparison of the effect of removing RMV8 from RMV8$_{rare}$ and RMV8$_{rare}$ PRCI$_{dnaN}$::Janus. **(C)** A comparison of the effect of inducing stimuli on the growth of RMV7$_{rare}$ and RMV7$_{rare}$ PRCI$_{dnaN}$::Janus. **(D)** The effect of increasing concentrations of mitomycin C on RMV8$_{rare}$ and mutant derivatives lacking RMV8 or PRCI$_{malA}$.