Microbiology (2016), **162**, 1346–1359 **DOI** 10.1099/mic.0.000326

Novel genomic rearrangements mediated by multiple genetic elements in *Streptococcus pyogenes* M23ND confer potential for evolutionary persistence

Yun-Juan Bao,[1] Zhong Liang,[1,2] Jeffrey A. Mayfield,[1,2] William M. McShan,[3] Shaun W. Lee,[1,4] Victoria A. Ploplis[1,2] and Francis J. Castellino[1,2]

Correspondence
Francis J. Castellino
fcastell@nd.edu

[1]W. M. Keck Center for Transgene Research, University of Notre Dame, Notre Dame, IN 46556, USA

[2]Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN 46556, USA

[3]Department of Pharmaceutical Sciences, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, USA

[4]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556, USA

Symmetric genomic rearrangements around replication axes in genomes are commonly observed in prokaryotic genomes, including Group A *Streptococcus* (GAS). However, asymmetric rearrangements are rare. Our previous studies showed that the hypervirulent invasive GAS strain, M23ND, containing an inactivated transcriptional regulator system, *covRS*, exhibits unique extensive asymmetric rearrangements, which reconstructed a genomic structure distinct from other GAS genomes. In the current investigation, we identified the rearrangement events and examined the genetic consequences and evolutionary implications underlying the rearrangements. By comparison with a close phylogenetic relative, M18-MGAS8232, we propose a molecular model wherein a series of asymmetric rearrangements have occurred in M23ND, involving translocations, inversions and integrations mediated by multiple factors, *viz.*, rRNA-*comX* (factor for late competence), transposons and phage-encoded gene segments. Assessments of the cumulative gene orientations and GC skews reveal that the asymmetric genomic rearrangements did not affect the general genomic integrity of the organism. However, functional distributions reveal re-clustering of a broad set of CovRS-regulated actively transcribed genes, including virulence factors and metabolic genes, to the same leading strand, with high confidence (p-value ~10^{-10}). The re-clustering of the genes suggests a potential selection advantage for the spatial proximity to the transcription complexes, which may contain the global transcriptional regulator, CovRS, and other RNA polymerases. Their proximities allow for efficient transcription of the genes required for growth, virulence and persistence. A new paradigm of survival strategies of GAS strains is provided through multiple genomic rearrangements, while, at the same time, maintaining genomic integrity.

Received 22 April 2016
Accepted 20 June 2016

INTRODUCTION

In previous investigations, we sequenced and analysed critical features of the genome (Bao *et al.*, 2014) and the transcriptome (Bao *et al.*, 2015) of an *emm23* isolate of a hypervirulent *Streptococcus pyogenes* (*S. pyogenes*) strain, M23ND, the lethality of which was predominantly induced by the inactivated and dysfunctional virulence sensor gene (CovS–) in comparison with the wild-type gene (CovS+).

The sequencing data for the transcriptome of M23ND are available in the GEO repository with series accession GSE67533.

Three supplementary tables and two supplementary figures are available with the online Supplementary Material.

The genome of M23ND underwent complex genomic rearrangements that include multiple DNA translocations, inversions and integrations. As a result, the genomic structure of M23ND exhibits distinct characteristics which were not found in other known sequenced *S. pyogenes* genomes (Bao *et al.*, 2014). First, the *dif*-like replication terminus (*ter*) and all four prophages are located in the first half-circle of the genome, which resulted in asymmetrical replication forks and an imbalanced replichore; second, the terminal 100 kb region was inverted *via* homologous recombination; and third, key virulence genes essential for its lethality were translocated within the genome. The availability of complete genomic information of a variety of *S. pyogenes* strains makes it possible to carry out genomic-scale comparisons. Numerous studies revealed that the genomic structures of *S. pyogenes* are highly syntenic among different strains, except for M12-HKU16 (Tse *et al.*, 2012), M5-Manfredo (Holden *et al.*, 2007) and M3-SSI-1 (Nakagawa *et al.*, 2003), all of which contain an X-shaped inversion (or X-alignment) that is symmetric around the replication axis, *ori/ter*. These inversions are mediated by homologous sequences, such as transposons (Tse *et al.*, 2012) or rRNA-*comX*, the latter comprising an rRNA operon and the competence gene, *comX* (Holden *et al.*, 2007; Nakagawa *et al.*, 2003). This type of symmetric inversion is not uncommon and has been analogously observed in organisms from multiple lineages (Eisen *et al.*, 2000), including *Salmonella* (Liu & Sanderson, 1995), *Chlamydia* (Read *et al.*, 2000) and *Helicobacter* (Zawilak *et al.*, 2001).

Remarkably, the symmetric changes retain the genomic organization properties, i.e. gene orientation bias, base composition bias and gene clustering patterns along the chromosome. The characteristics in genome organization could be under different inter-correlated selection forces, such as chromosomal replication, gene expression and probably gene regulation. Therefore, the frequently observed symmetric rearrangement events were thought to be naturally selected by those factors in the evolutionary processes (Eisen *et al.*, 2000; Lobry, 1996; Rocha *et al.*, 1999; Rocha, 2002; Tillier & Collins, 2000a). For example, base composition bias and gene orientation bias were found in many bacterial genomes between the two replication strands, *viz.*, the leading strand and lagging strand (Lobry, 1996; Rocha *et al.*, 1999). In those cases, the leading strand usually displayed a positive GC skew, or sometimes a TA skew, while the lagging strand is the opposite. The bias in base composition is closely related with the asymmetric replication machinery between the two strands: the leading strand replicates continuously, whereas the lagging strand replicates in short fragments, forming Okazaki fragments (Okazaki *et al.*, 1967; Sakabe & Okazaki, 1966). The different replication mechanisms may exert differential mutation or selection pressures on the two strands and, thus, contribute to the biased base composition (Rocha, 2002; Tillier & Collins, 2000a). The bias in gene orientation is represented by the predominant distribution of genes in the leading strand, relative to the lagging strand. In *Bacillus subtilis*, as much as 75 % of the genes are concentrated in the leading strand (Kunst *et al.*, 1997). Since the location of genes on the leading strand allows gene replication in the same direction as gene transcription, gene orientation bias was proposed to be a selection consequence to minimize head-on collisions between the replication machinery and RNA polymerases (Brewer, 1988; French, 1992; Helmrich *et al.*, 2013). Specifically, the highly expressed genes tend to cluster in the leading strand, conferring potential advantages for smooth replication and transcription. The actively transcribed ribosomal genes and ribosomal proteins were frequently found to locate proximally in the leading strand (Brewer, 1988; Merrikh *et al.*, 2012).

The clustering of transcriptionally active genes in the chromosome may also be complicated by their regulation mechanisms, in addition to the selection by replication and transcription. Recently, hypothetical models were proposed for prokaryotic and eukaryotic genomes to form highly folded loops such that the actively transcribed genes in distal chromosomal regions are able to physically co-localize in the same transcription factories in the nucleoid of prokaryotes or the nucleus for eukaryotes (Cook, 2002; Papantonis & Cook, 2013). The transcription factories are compartmentalized foci in the nucleoids or nuclei, containing polymerases, transcription factors and other complexes in combination to facilitate efficient recruitment and regulated transcription of target genes (Chakalova *et al.*, 2005). The formation of genome looping and recruitment efficiency are expected to be substantially influenced by gene organization in the chromosome, though the mechanisms have not yet been elucidated (Cournac & Plumbridge, 2013).

In this communication we present evidence that the invasive hypervirulent *S. pyogenes* strain, M23ND, exhibits extensive asymmetric rearrangements around the replication axis, *via* homologous recombination, that are induced by multiple factors. These factors include inversions mediated by rRNA-*comX*, translocations due to transposons, and inversions brought about by phage repeat elements. These rearrangements combined to shape the unique genomic architecture of M23ND, and serve as examples needed to be considered in characterizing the genomic architectures. We also examined the influence of those asymmetric arrangements on the genomic organization of this microorganism, compared to that of symmetric inversion, and the driving forces underlying the rearrangement processes from the viewpoint of DNA replication, gene transcription and regulation.

METHODS

Bacterial strains and genomic sequences. GAS strain M23ND was originally isolated from a severe streptococcal infection and designated as strain *Sv* (ATCC21059). The complete genome is accessible with GenBank accession CP008695. Genomic sequences of other *S. pyogenes* strains used for comparative purposes were obtained from the NCBI Database, with RefSeq accession NC_003485 for M18-MGAS8232 and NC_009332 for M5-Manfredo. Genomic comparisons were performed using the BLAST (Zhang *et al.*, 2000) and MUMmer

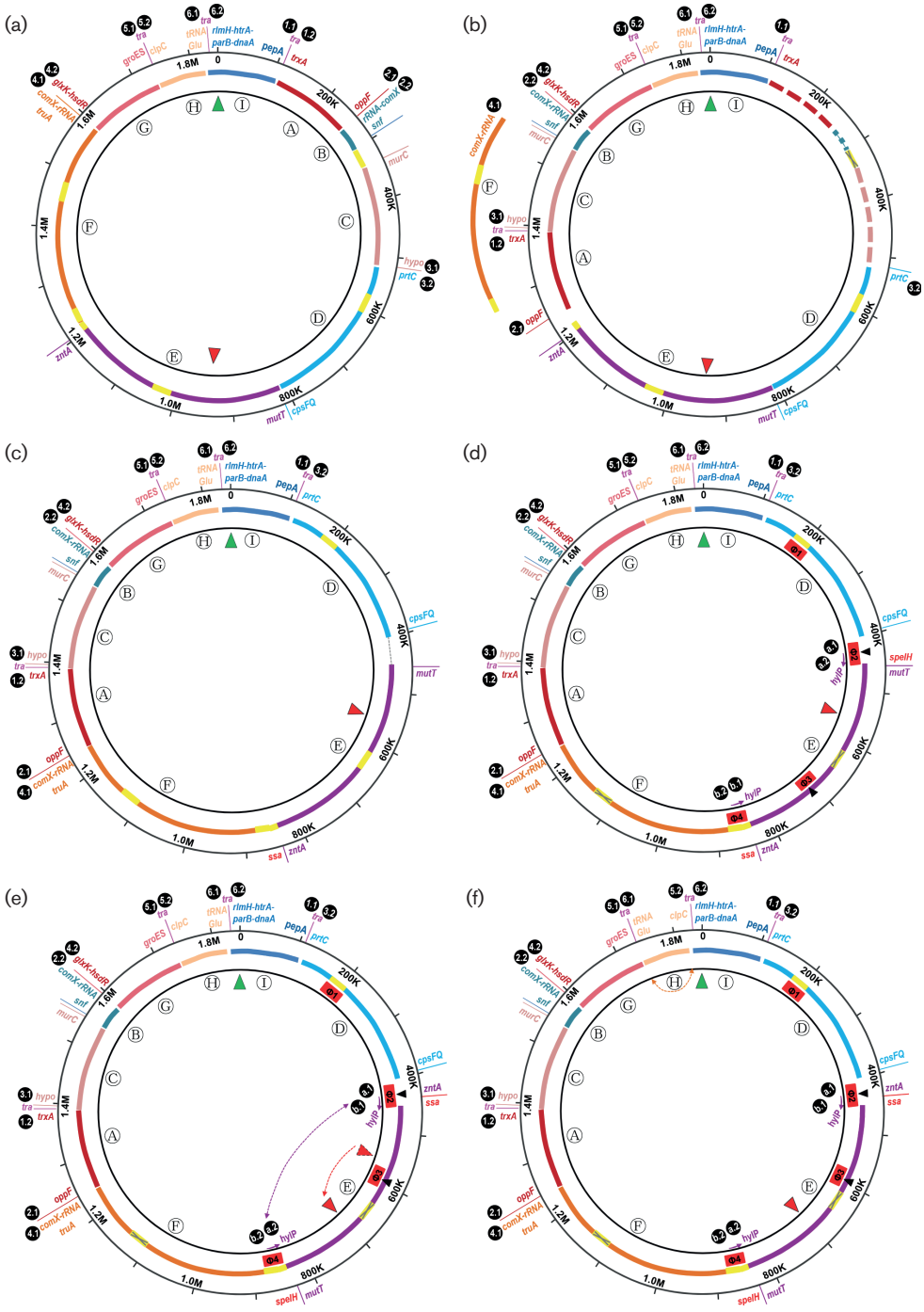

Fig. 1. Outline of major rearrangement events that occurred in M23ND. The overall rearrangements involve eight segments (A–H), forming six breakpoints (1–6) at repeat sequences of rRNA-*comX* or transposon *tra*, and two breakpoints (a–b) at phage repeat elements of *hylP*. (a) The rearranged segments and breakpoints are mapped to the reference genome of M18-MGAS8232. (b) Segments A-B-C were inversely translocated to the position of Segment F *via* rRNA-*comX* and transposon *tra*. (c) Segments F-E-D were moved towards the replication origin *via* another copy of rRNA-*comX* and transposon *tra*; this translocation brings the replication terminus *ter* site from position ~956 kb to ~560 kb. (d) A novel phage M23ND.Φ2 was integrated at the location of around 400 kb. This phage contains a ~2.4 kb inverted repeat of the phage hyaluronate lyase gene fragment (*hylP*) in another phage M23ND.Φ4. (e) The inverted repeats of *hylP* mediated an inversion of segment E in between the two *hylP* fragments, such that the *ter* site was carried backward to the current position at ~702 kb. (f) Segment H at the last 100 kb was inverted *in situ*, mediated by the transposons at the two ends. Both sides of each breakpoint are indicated by, e.g. 1.1 and 1.2. The genes near the breakpoints are shown. The phage regions are indicated with arcs in yellow, and the

phages absent from M23ND in comparison with M18-MGAS8232 are crossed. The inversions are indicated by dotted arrow curves. The replication origin (*ori*) and terminus (*ter*) are indicated by green and red triangles, respectively.

packages (Delcher *et al.*, 2002) and graphically viewed with Mauve (Darling *et al.*, 2004) and ACT (Carver *et al.*, 2008).

Strain culture and mitomycin C-driven induction of phages. M23ND was grown in Todd-Hewitt broth supplemented with 0.2 % yeast extract (THY) at 37 °C overnight. The cell cultures were diluted at 1 : 10 in 50 ml fresh THY media, and incubated at 37 °C until $A_{600\ nm}$ = 0.2. The culture was divided into two 25 ml aliquots with one grown to stationary phase ($A_{600\ nm}$ = 1.0) and the other treated with mitomycin C for 3 h (0.2 μg ml^{-1}). The resulting cells were harvested by centrifugation at 10 000 rpm for 5 min and treated with lysozyme/proteinase K and lysis buffer (100 nM Tris, 5 nM EDTA, 0.2 % SDS, 200 nM NaCl, pH 8.5). The DNA was extracted with phenol/chloroform/isoamyl alcohol (25 : 24 : 1, v/v/v).

PCR assays for the presence of rearrangements. The rearrangements caused by rRNA-*comX*, IS1239, and phage-encoded hyaluronate lyase (*hylP*) gene fragment DNAs, were verified by designing primers to amplify the regions covering the repeat elements and breakpoints. PCR amplifications of large repeat elements rRNA-*comX* (~7 kb) were performed using LA-Taq (Takara). Phage induction was tested by PCR primers designed to amplify the regions *attB* and *attP* in the gDNA of non-treated bacterial cells and cells treated with mitomycin C. The site-specific primers for these amplifications are provided in Table S1 (available in the online Supplementary Material). The PCR products were further identified by Sanger sequencing.

Transcriptomic expression profiling of M23ND regulated by CovRS. The details for the derivation of transcriptomic data for M23ND regulated by the two-component Responder-Sensor regulator system, CovRS, have been published earlier (Bao *et al.*, 2015). The M23ND isolate with inactivated CovS (CovS–) and its complemented isogenic CovS strain (CovS+) were grown to mid-log phase (LP) and stationary phase (SP) for RNA isolation and subsequent transcriptomic profiling.

Calculation of cumulative gene orientation and GC skew. The calculation of cumulative gene orientation and GC skew was based on the method described in Tillier & Collins (2000a) with minor modification. Gene orientations were assigned +1 for genes coding in the sense strand and −1 in the antisense strand. The cumulative gene orientations were normalized by dividing each value by the number of genes considered. The GC skew (G−C)/(G+C) was calculated with a window size of 1000 and a window step of 500. To normalize the cumulative GC skew, each value was multiplied by the number of steps and divided by the total number of nucleotides considered for calculation.

RESULTS

A proposed model of the rearrangement events in M23ND

In order to better understand the mechanisms and consequences of the genomic rearrangements that occurred in M23ND, we delineated the possible molecular processes of the rearrangement events by comparing the M23ND genome with that of M18-MGAS8232, a phylogenetic close relative of M23ND (Smoot *et al.*, 2002). A linear representation of the genomic architecture comparison between the two genomes is shown in Fig. S1. From the comparison, eight rearranged segments (A–H in Fig. S1), involving six breakpoints (1-6), are identified at repeat sequences (rRNA-*comX* or transposons), as well as two breakpoints (a–b) at phage repeat elements. Based on the locations and sequences at the breakpoints, we propose a model of the genomic reorganizations that occurred during the evolution of M23ND (Fig. 1). In this model, at least five major molecular changes account for the overall rearrangement profile: (1) Segments A-B-C (Fig. 1a) were inversely translocated to the position of segment F, mediated by rRNA-*comX* and transposon *tra* (Fig. 1b); (2) Segments F-E-D (Fig. 1a) were moved towards the replication origin (*ori*) by another copy of rRNA-*comX* and transposon *tra*; meanwhile the replication terminus *(ter)* was carried from the position ~956 kb to ~560 kb (Fig. 1c); (3) A novel phage, M23ND.Φ2, was integrated at a location around 400 kb. This phage contains a ~2.4 kb inverted repeat of the gene fragment of *hylP* in another phage, M23ND.Φ4 (Fig. 1d). (4) The inverted repeats of *hylP* mediated an inversion of the large segment E between the phage repeat elements *hylP*. This inversion resulted in the movement of the replication terminus backward to its current position at ~702 kb, which made the two replichores less imbalanced (Fig. 1e); (5) Segment H was inverted *in situ* by the transposon *tra* at the two ends of the segment (Fig. 1f). While the overall rearrangement events seem clear, their temporal order is uncertain due to the limited number of known genomes of GAS, as well as multiple intermediate recombinations that may have occurred during the evolution.

Identification of homologous recombination breakpoints at the rRNA-*comX* repeats

The ComX transcription factor for late competence genes, of length 486 bp, is in the same locus with a ribosomal operon rRNA, of length ~5.6 kb, containing 16S rRNA, tRNA-Ala, 23S rRNA, 5S rRNA, tRNA-Asn and tRNA-Arg (Lee & Morrison, 1999). This locus has two copies distributed on both sides of the replication axis, *ori/ter*, in quasi-symmetry in most *S. pyogenes* genomes. Its symmetric distribution induces X-shaped reciprocal inversion across the *ori/ter* axes in multiple clinical strains, such as M5-Manfredo (Holden *et al.*, 2007) and M3-SSI-1 (Nakagawa *et al.*, 2003). The breakpoints caused by symmetric inversion were shown to be located downstream from the two *comX* genes in M3-SSI-1 (Nakagawa *et al.*, 2003), and resulted in a large inversion of 1.4 Mbp of the total ~1.8 Mbp of the genome. Through a comparative study of the genomes of M23ND and M18-MGAS8232, we show that rRNA-*comX* inversely translocated segments A-B-C to the position of F in M23ND (Fig. 1a, b), and that the rearrangements are highly asymmetric around the replication axis involving translocation and inversion (Fig. 2a). The rearrangements

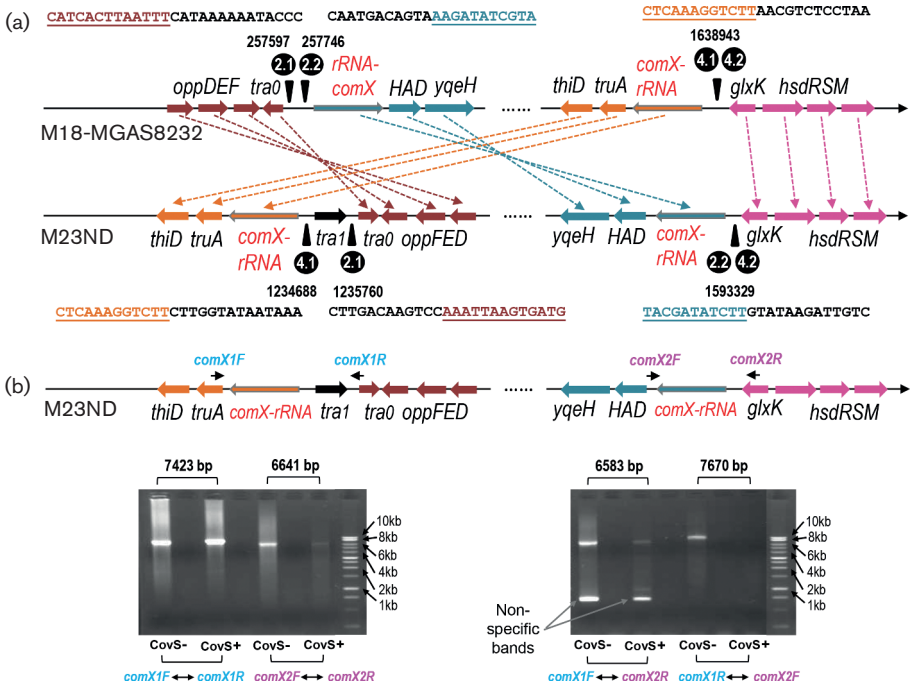

Fig. 2. Local view of the rearrangements mediated by rRNA-*comX* in M23ND in comparison with M18-MGAS8232. (a) Two copies of rRNA-*comX* facilitated a horizontal translocation and a reverse translocation at site 4.1 and 2.2, respectively. An additional breakpoint (2.1) was detected between upstream 16S rRNA and downstream *oppF* due to insertion of the interrupted transposon, *tra1*. The sequences of this transposon are mosaic and cannot be unambiguously classified. The gene blocks are indicated by arrows of the same colour and the rearrangements by dotted lines. The block of rRNA-*comX* is framed with grey lines. The black arrowheads indicate the sites of breakpoints. The breakpoint sequences and locations are shown accordingly. The sequences are coloured the same as the neighbouring gene arrows. Identical breakpoint sequences are shown in the same colours with underlining. (b) Two primer pairs, *comX1F/comX1R* and *comX2F/comX2R*, were designed to amplify the regions covering rRNA-*comX* and the breakpoints to verify the rearrangements. The PCR amplifications for the swapped primer combination, *comX1F/comX2R* and *comX1R/comX2F*, were also performed to examine the presence of the genomic structure without the rearrangement. The appropriate PCR fragments from both of the genomic structures were observed to be suggestive of the reversibility (or incompleteness) of the rearrangement. The amplification was performed in both the CovS-mutant strain (CovS−) and the isogenic CovS-intact strain (CovS+). The gene name rRNA-*comX* is highlighted in red.

mediated by rRNA-*comX* were confirmed by designing long-PCR primers to amplify the region covering rRNA-*comX* and its breakpoints. Interestingly, the PCR fragments were also detected from amplification of the rRNA-*comX* regions in a newly created genomic configuration with the rearrangements reversed using swapped primer pairs (Fig. 2b). This indicates that both of the genomic structures with and without the rearrangements co-exist in the population.

Identification of homologous recombination breakpoints at transposons

In addition to rRNA-*comX*, transposon represents another major genomic element that is responsible for facilitating the genomic recombinations found in M23ND. The reorganization of segments A, C, D and H was implemented by transfer of a transposon at four *loci* forming four breakpoints (Fig. 1a at junctions 1, 3, 5 and 6). By comparison with M18-MGAS8232, we identified the precise break locations and illustrated the recombination architectures at the specific *loci* (Fig. 3a–c). It was found that the breakpoint sequences on the transposon sides are identical, and the transposon sequences herein are homologous, thus representing a homologous recombination mechanism of the transposon (Fig. 3a–c). Sequence comparisons revealed that the transposons shared high similarity with the insertion sequence (IS) class IS1239, and can be confidently classified as IS1239 (Berge *et al.*, 1998).

Notably, the inversion of segment H towards the last 100 kb of the chromosome, mediated by two additional inverted copies of IS1239, is unusual in that it occurred in the same replichore and should be usually counter-selected by natural selection. The intra-replichore rearrangements induced by transposons were previously observed in another human pathogen, *Yersinia pestis* (Darling *et al.*, 2008; Parkhill *et al.*, 2001) and a hyperthermophilic archaea, *Pyrococcus furiosus*

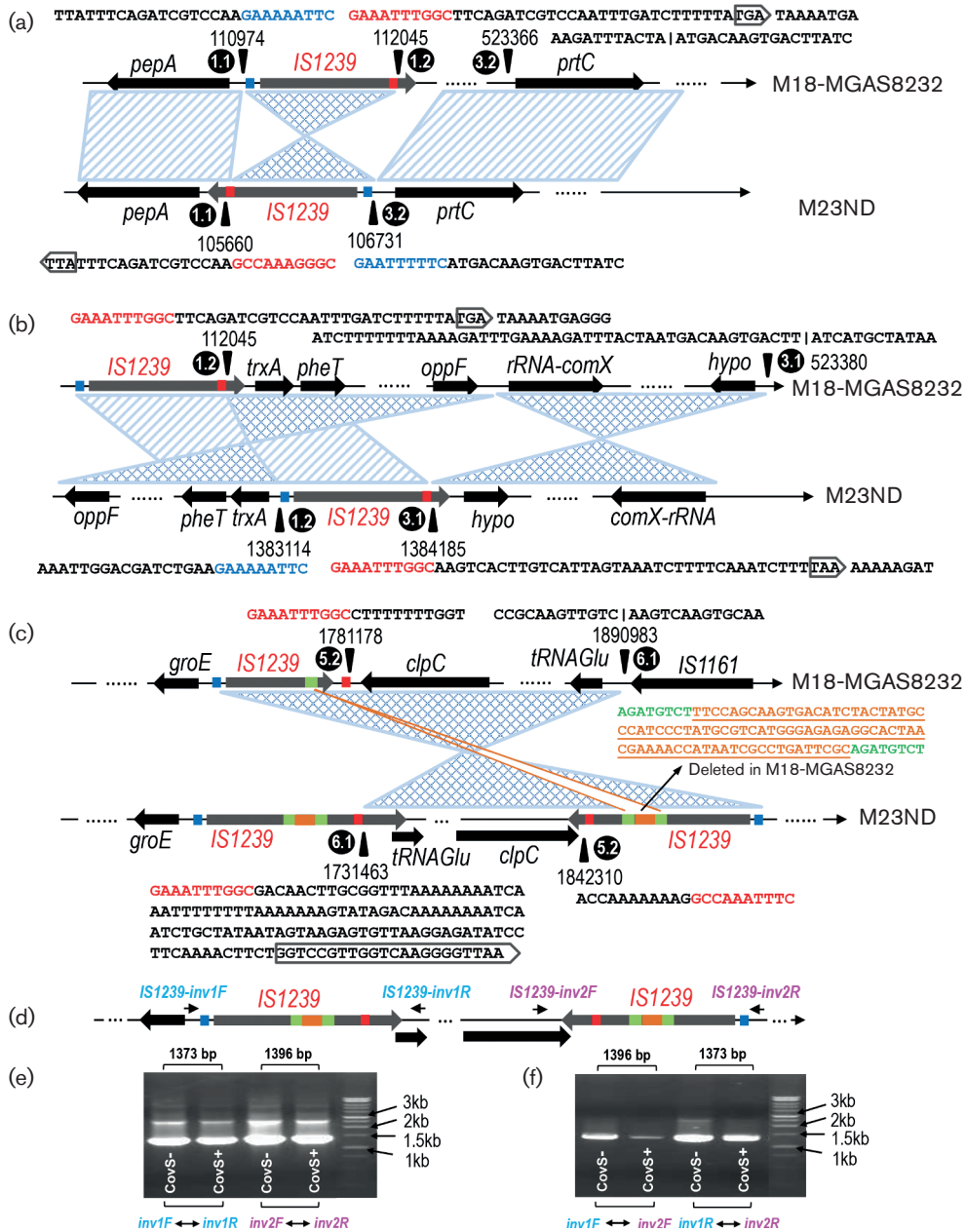

Fig. 3. Local views of the recombination and architectures mediated by IS1239 in M23ND compared to M18-MGAS8232. (a) IS1239 with breakpoints at 1.1 and 3.2. (b) IS1239 with breakpoints at 3.1 and 1.2. (c) IS1239 with breakpoints at 5.2 and 6.1, mediating the inversion of segment H in the last 100 kb of the genome. One copy of IS1239 was replaced by another transposon, IS1161, at breakpoint 6.1 in M18-MGAS8232, possibly due to the different rearrangement at this point in the two strains. IS1239 is indicated by grey arrows with gene names in red. The neighbouring genes are indicated by black arrows. The horizontal alignments between genomic blocks are shown by oblique lines and reverse alignments by cross-lines. The black arrowheads indicate the sites of breakpoints. The breakpoint sequences on both sides of IS1239 are represented by blue and red boxes and their sequences/locations are shown accordingly. The unfilled arrows in the sequences indicate the stop codon of the transposon. IS1239 at breakpoint 5.2 in M18-MGAS8232 contains a short deletion (orange box) *via* excision at the direct repeats, AGATGTCT (blue box). The deleted sequences are shown and indicated. (d) Two primer pairs were designed to amplify the fragments covering IS1239 and the flanking regions to verify the induced inversion at the last 100 kb. (e) The PCR fragments of length ~1.3 kb were amplified specifically for *IS1239-inv1* and *IS1239-inv2*. (f) The PCR reaction tests for the swapped primer combinations *IS1239-inv1F*/*IS1239-inv2F* and *IS1239-inv1R*/*IS1239-inv2R* were also performed. Observation of the proper bands suggests that the inversion of segment H is reversible. The amplification was performed in both the CovS-mutant strain (CovS−) and the isogenic CovS-intact strain (CovS+).

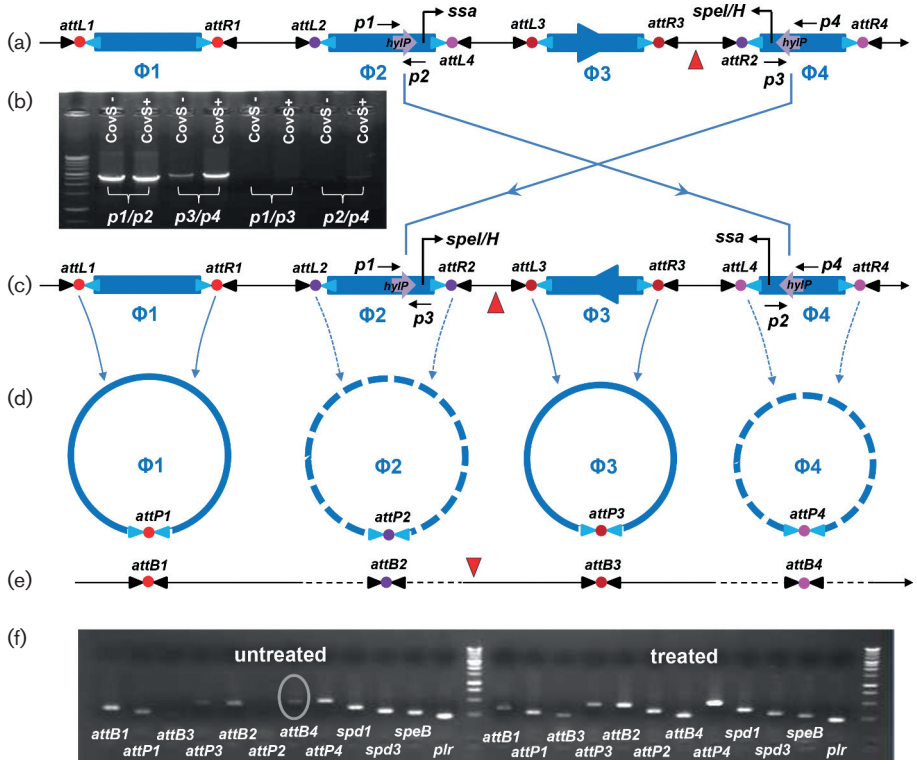

Fig. 4. Genomic architecture of the inversion induced by two inverted repeats encoding the gene fragment of *hylP* in M23ND.Φ2 and M23ND.Φ4. (a) Primers *p1, p2, p3* and *p4* were designed to amplify the two repeat regions around *hylP*. (b) PCR products were obtained for *p1/p2* and *p3/p4*, confirming the inversion induced by repeat sequences of *hylP*. Observation of the weak bands for *p1/p3* and *p2/p4* suggests the reversibility (or incompleteness) of the inversion. (c) The inversion occurred between the inverted sequences of *hylP,* causing movement of the *ter* site toward the centre of the genome and the exchange of the virulence cassettes, *viz., ssa* and *speI/H*. (d,e). Exchange of the *attR* site of M23ND.Φ2 with the *attL* site of M23ND.Φ4 would theoretically make the two phages non-inducible and prevent the rejoining of *attB* sites in the bacterial chromosome. On the other hand, M23ND.Φ1 and M23ND.Φ3 are proposed to be induced with two ends rejoined at *attP* sites and the bacterial chromosome closed at *attB* sites. (f) Primers were designed to amplify the regions around *attB* and *attP* with the phage-encoded virulence genes, *spd1* and *spd3* and chromosome-encoded genes, *speB* and *plr*, as controls. M23ND.Φ1 and M23ND.Φ3 are inducible with or without mitomycin C treatment. M23ND.Φ2 and M23ND.Φ4 are shown also to be inducible without and with treatment with mitomycin C owing to the reversibility of the region between the phage repeat element *hylP*. A false PCR product is circled with a grey line, and was shown to be irrelevant by clone sequencing.

(Diruggiero *et al.*, 2000; Zivanovic *et al.*, 2002). The inversion was confirmed by PCR amplification of IS1239 and its flanking regions (Fig. 3d, e). Again, we also detected the PCR fragments from amplification of the IS1239 regions after specifically reverting the inversion (Fig. 3d, f).

The short transposable element or insertion sequences, with the IS1239 as an example herein, are able to mediate complicated rearrangements in addition to their simple transpositions or acquisition of accessory genes (Mahillon & Chandler, 1998). IS elements have been found to induce large inversions *via* IS3 (Komoda *et al.*, 1991) and tandem duplications *via* IS200 (Haack & Roth, 1995). However, the current study of IS-driven, large-scale asymmetric genomic rearrangements in M23ND is, to our knowledge the first such report in *S. pyogenes*.

Phage repeat elements induced a large inversion

It is well documented that a homologous recombination of inter- and intra-strain phage genomes plays key roles in genetic diversification of phages (Hatfull, 2008), through which highly mosaic structures are formed across phage genomes (Banks *et al.*, 2004; Bessen *et al.*, 2015; Green *et al.*, 2005; McShan & Ferretti, 1997). The sites of homologous recombination between phages are usually located in regions of homology, such as the *holin* (Nakagawa *et al.*, 2003) and the *tail fibre* genes (Canchaya *et al.*, 2002). In the genome of M23ND, a large inversion was observed at around 441 000–850 000 bp, bracketed by two inverted repeat regions, encoded respectively by two phages, *viz.*, M23ND.Φ2 and M23ND.Φ4, when compared with M18-MGAS8232 or other strains. The repeat region codes for a

~2.4 kb fragment of the *hylP* gene (hyaluronate lyase), which is conserved in multiple phages and, therefore, may act as another hot spot for phage recombination. For verification of the large inversion, we carried out PCR amplifications by designing primer pairs upstream and downstream from each of the two repeat regions (*p1, p2, p3* and *p4* in Fig. 4a). The amplified PCR products, of length ~3 kb, were obtained from *p1/p2* and *p3/p4* amplifications (Fig. 4b), thus explicitly confirming the inversion induced by the phage repeat elements of *hylP*. We also noted that there are comparatively very weak bands for amplicons with primers *p1/p3* and *p2/p4* which indicates that both genotypes coexist in the population, with inversion as the dominant form.

It is noteworthy that the recombination sites between the two copies of the *hylP* fragment in M23ND are not equidistant from the *ter* site, revealing the asymmetric properties of the inversion around the replication axis. Rather, the inversion has caused the movement of the *ter* site toward the centre of the circular genome by ~160 kb, thereby making the replichores less unbalanced (Fig. 4c). This, at least in part, can be a driving force of the inversion.

Inversion induced by phage repeat elements is reversible

In order to further confirm the inversion induced by phage repeat elements and its reversibility, we examined the inducibility of the phages encoded by M23ND. Phage induction can be stimulated by DNA damage or host enzymatic systems, which allows phage excised from the host bacterial chromosome to form a circularized free phage genome through a Campbell mechanism of homologous recombination between paired integration sites, *attL* and *attR* (Campbell, 1992) (Fig. 4c, d). The recombination results in the re-joining of the junction to form *attP* in phage genomes and *attB* in bacterial chromosomes (Fig. 4d, e).

The inversion induced by the phage repeat elements in M23ND violates the pairing between *attL* and *attR* by exchanging *attR* at the right end of M23ND.Φ2 and *attL* at the left end of M23ND.Φ4, such that the two phages may not be excised, and, hence, become non-inducible (Fig. 4a). M23ND.Φ1 and M23ND.Φ3 are not affected by this inversion and may theoretically be inducible to form free phage particles (Novick *et al.*, 2010; Scott *et al.*, 2008). To test this hypothesis, phage induction was conducted without and with treatment of M23ND with mitomycin C, which is a commonly used DNA damaging agent (Roberts & Roberts, 1975). For this, we initially identified the integration sites on both ends of each phage (*attL/attR*) and the rejoined junctions of the bacterial (*attB*) and phage genomes (*attP*) after excision (McShan *et al.*, 1997) (Fig. 4d, e and Table S2). Subsequently, PCR amplifications were performed with primers designed around the sites of *attB* and *attP* using the primers listed in Table S1. Proper PCR products were obtained for *attB/attP* from M23ND.Φ1 and M23ND.Φ3 (*attB1/attP1* and *attB3/attP3*), as expected (Fig. 4f). It is observed that M23ND.Φ1 and M23ND.Φ3 were induced not only by mitomycin C, but also spontaneously, thus reflecting the mixture of the integrated and excised phage DNA in the natural population. Interestingly, the PCR products of correct size were also amplified for *attB* and *attP* from M23ND.Φ2 and M23ND.Φ4 (*attB2/attP2* and *attB4/attP4*) without treatment with mitomycin C, suggesting that the large inversion mediated by the phage elements, *hylP*, is reversible, such that the paired *attL* and *attR* sites can be recovered. This finding is consistent with the observation of the weak bands for primers *p1/p3* and *p2/p4* in Fig. 4b. The reversibility is further supported by the proper PCR products for *attB/attP* from the cells treated with mitomycin C, which makes the signals more significant (Fig. 4f, right panel). The specificity of the products was additionally confirmed by DNA sequencing. We notice that mitomycin C not only induced phage excision, but was also involved in reverting the large inversion between *hylP* (Fig. S2). Therefore, the final PCR products are the combined contributions of phage excision following spontaneous reversion of the region between *hylP* and mitomycin C-induced reversion of the same region. The two contributions generated the same set of PCR products, therefore demonstrating that inversion between the *hylP* is reversible, and both orientations of the fragment exist in the same bacterial population, with the inverted structure as the stable one.

Genomic integrity and functional gene distribution are not affected by the rearrangements in M23ND

It has been shown that many prokaryotic genomes exhibit quasi-antisymmetric distribution of gene orientation and base composition between the two replication strands, *viz.*, the leading strand and lagging strand (Eisen *et al.*, 2000; Rocha *et al.*, 1999; Tillier & Collins, 2000a). The antisymmetry (or quasi-antisymmetry) was proposed to be probably due to differential mutation or selection pressures on the two replication strands (Lobry, 1996; Rocha, 2002; Tillier & Collins, 2000b). Based on these observations, it was postulated that the symmetric rearrangement events are more common because they retain the disparity of gene orientation and base composition between the two strands and, hence, are naturally selected by a mechanism associated with genome replication in the evolutionary processes (Tillier & Collins, 2000b). It is important to assess how the asymmetric rearrangements in M23ND would influence the disparity of gene orientation and base composition, and more interestingly, whether this disparity will be retained in the genome of M23ND.

The cumulative gene orientation and GC skew was calculated along the genomes of M23ND, M18-MGAS8232, which contains no large-scale rearrangements, and M5-Manfredo, which has a large symmetric inversion across the replication axis. The calculation of the cumulative gene orientation curves and GC skew curves for each of the genomes reveals V-shaped curves depicting an explicit

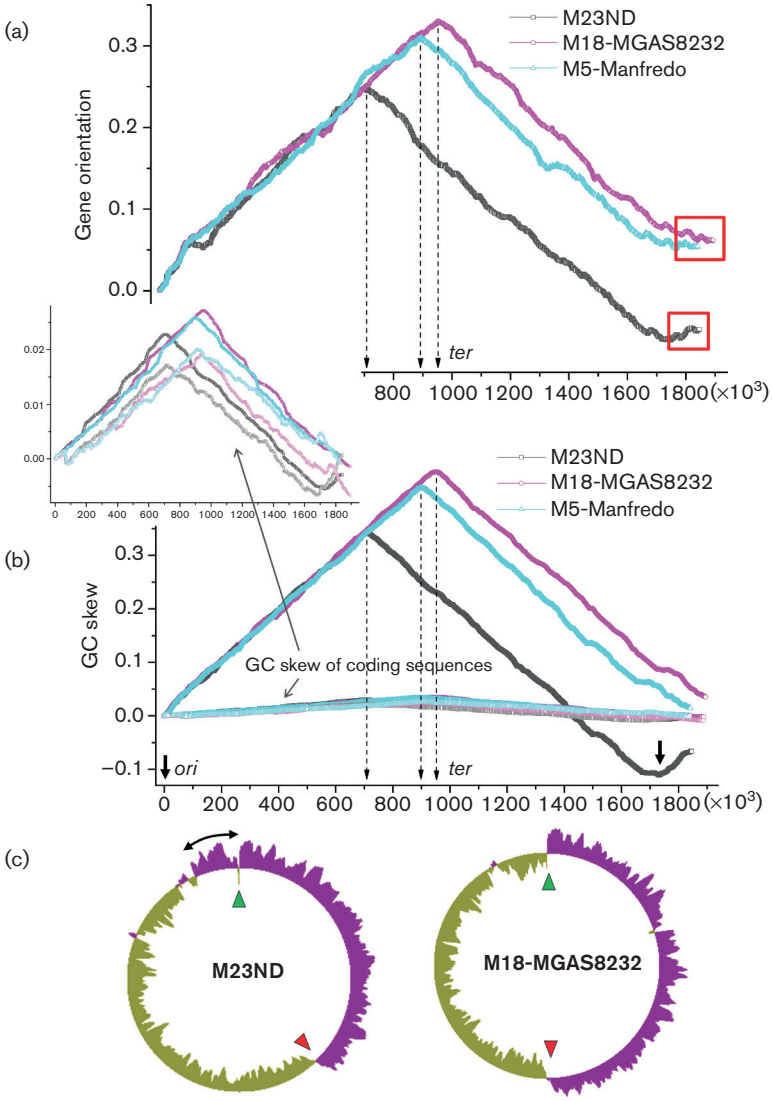

Fig. 5. Changes in genomic properties by genomic rearrangements in M23ND in comparison with M18-MGAS8232 and M5-Manfredo. (a) The cumulative gene orientations along the whole genome. (b) The cumulative GC skews along the whole genome. The inset shows a zoomed view of the cumulative GC skews for the coding sequences, with the light colour for the third codon of coding genes. The inflection points appear at the *ter* site in the cumulative gene orientation curves and GC skew curves and are indicated by vertical dotted lines. The 100 kb tail in the cumulative gene orientation curves exhibits a flat trend (framed by a red rectangle) in comparison with the slope in other regions. (c) Absolute GC skews in the circularized genomic view. The replication origin and terminus are indicated by green and red triangles, respectively. The inversion towards the last 100 kb in M23ND (double-headed arrow) recovered the anti-symmetric distribution of GC skew by shifting the switching point of the cumulative GC skew.

overview of the disparity of gene orientation and GC skew between the two replication strands in M23ND, analogous to the other two genomes (Fig. 5a, b). The apparent gene orientation bias in M23ND generates a ratio of 77 % : 23 % of the genes transcribed in the replication direction *vs.* the counter-replication direction. This is comparable to the 80 % : 20 % for M18-MGAS8232, and *75 % : 25 %* for *Bacillus subtilis* (Kunst *et al.*, 1997). In this regard, the genome rearrangements in M23ND do not generally affect the overall genomic properties. Specifically, a short inversion at the final 100 kb region, manifested by a smaller inflection point in the cumulative GC skew in M23ND, results in a ~100 kb shift of the switching point of the cumulative GC skew (Fig. 5c). The shift is not uncommon in prokaryotic genomes and could be induced by diverse rearrangements (Andersson *et al.*, 1998; Grigoriev, 2000). The inverted100 kb region exhibits less biased gene orientation, with a flatter trend in the curve of the cumulative gene orientation due

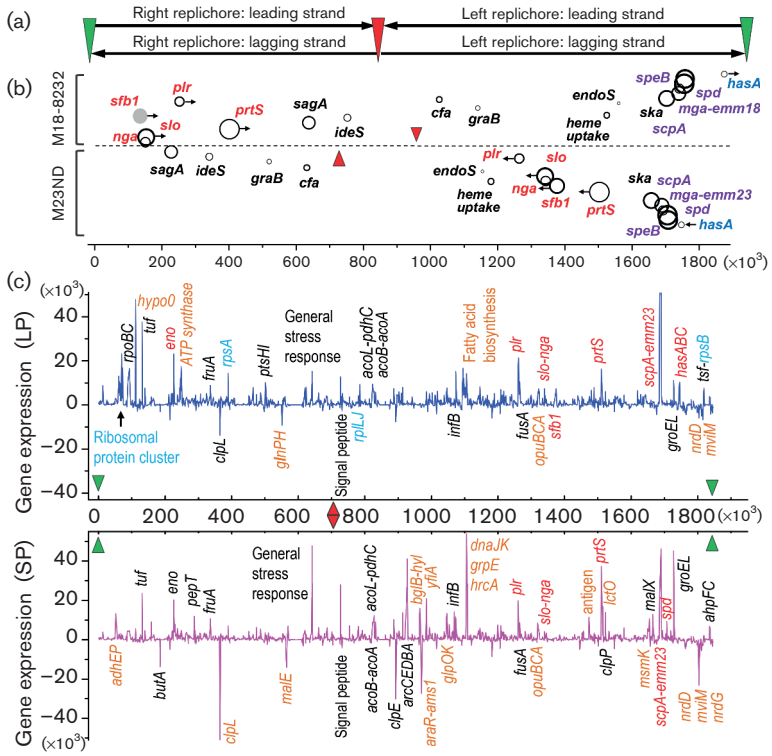

Fig. 6. Re-clustering of virulence genes and global expression profile for the actively transcribed genes in M23ND. (a) A schematic representation of the leading strand and lagging strand on each replichore is shown for a typical bidirectional replicating bacterial chromosome. The *ori*/*ter* sites are indicated by green and red triangles, respectively. (b) The genomic relocation of the virulence genes associated with the pathogenesis of *S. pyogenes* M23ND in comparison with M18-MGAS8232. The *ter* sites for each genome are indicated by red triangles. A re-clustering pattern is observed. The genes *plr, sfb1, prtS, nga* and *slo*, shown in red with black arrows for the transcription orientation, were relocated to the opposite leading strand and clustered with the *in situ* genes *speB, spd, mga* regulon (shown in purple) and *hasABC* operon (shown in blue). The *hasABC* operon has been carried to the leading strand from the lagging strand by the inversion in the last 100 kb. The gene *sfb1* is not encoded by M18-MGAS8232, and therefore the grey-filled circle represents the average position of *sfb1* in the genomes that code for this gene, *viz.,* M6, M4, M14, M12, M59 and M28. The positions of the genes are artificially offset along the y-axis for better representation. The bubble sizes are proportional to the length of genes (except the gene regulons). (c) The gene expression profile is compiled for the mid-log phase (LP) and stationary phase (SP). The expression levels are shown as read counts normalized to the total sequencing capacity. The sign of the expression represents the co-direction or anti-direction of gene replication relative to transcription orientation. A clustering pattern is observed for a broad set of genes on the leading strand in the left replichore (the region between *ter* and *ori*), including virulence factors (red) and metabolic genes (orange), most of which were shown to be abundantly expressed and proposed to be regulated by CovRS. The highly expressed ribosomal genes are highlighted in light blue. The details of functions and expression levels of the genes are described in Table S3.

to the comparable gene densities between the leading strand and lagging strand (40 % : 60 %).

We also examined the changes of global functional gene distribution in M23ND in comparison with M18-MGAS8232. The distribution density of broad functional categories in M23ND does not show significant changes from M18-MGAS8232, i.e. amino acids and their derivatives, nucleosides and nucleotides, fatty acids and lipids, carbohydrates and metabolism. The average expression levels of the functional genes do not exhibit bias between the two replication strands by examining the transcriptome profile of M23ND (data not shown).

Gene reorganization may reflect a functional selection for the growth and pathogenesis of M23ND

We narrowed the targets to the genes relevant to GAS virulence for investigation of the changes in gene reorganization induced by the rearrangement events in M23ND. We observed that the genes responsible for the hypervirulence

of M23ND are altered in their organization patterns. The hyaluronic acid capsule operon, *hasABC*, was carried to the leading strand from the lagging strand by the inversion of the last 100 kb (Fig. 6a, b). The operon was previously shown to be abundantly regulated in M23ND by the inactivated CovS, and is necessary for GAS invasiveness and survival. The strand transfer transcribes the highly expressed *hasABC* operon in the direction of replication and potentially avoids the collision between RNA polymerases and replication complexes, and hence is likely to confer an advantage for smooth replication and transcription at the local site (Brewer, 1988).

On the other hand, the genomic rearrangements also caused the translocation of the virulence genes *plr, nga, slo, prtS* and *sfb1* from the right replichore to the left replichore (Fig. 6a, b). The translocation resulted in the re-localization of these genes to the opposite leading strand and clustering with several other *in situ* key virulence factors of GAS, *viz., mga, emm23, scpA, spd* and *hasABC*. The examination of our transcriptome data of M23ND (Bao *et al.*, 2015) showed that the clustered genes are all transcribed in high abundance and supposedly regulated by the two-component CovRS regulator system, except *plr*. Therefore, we further examined the CovRS-regulated expression profile of all transcribed genes in addition to the virulence genes. It is observed that the clustering of actively transcribed genes involves a broad set of abundantly transcribed genes including virulence genes and metabolic genes, both of which were shown to be under the regulation of CovRS (Fig. 6c and Table S3). The clustering is statistically significant, with a Bonferroni multiple test adjusted *p*-value of ~10–10 using Fisher's exact test.

Overall, the composite of the asymmetric rearrangements that occurred in M23ND exhibits the evidence of bringing the highly expressed CovRS-regulated genes to be co-localized in one of the leading strands. Given that CovRS is a global transcriptional regulator that may regulate 10%–18% of all GAS genes (Graham *et al.*, 2002; Horstmann *et al.*, 2015), we propose that a transcription factory-like machinery containing CovRS and other RNA polymerase complexes may exist, and the co-localization of the abundantly regulated genes reflects a selection for spatial proximity to CovRS allowing for efficient transcription of the genes required for growth and pathogenesis in challenging human niches (Bartlett *et al.*, 2006). Like eukaryotes, the chromosomes of prokaryotes could also be compactly packed *via* loops and coils in the nucleoid and bring distantly localized genes into closer proximity in the nucleoid, thereby facilitating long-range interactions with the molecular complexes required for transcription (Osborne *et al.*, 2004; Wang *et al.*, 2011). However, the packaging model of prokaryotic nucleoids has not yet been fully elucidated (Bartlett *et al.*, 2006; Cook, 2002).

The re-clustering of the CovRS-regulated genes to the same leading strands in M23ND can arise because of several possibilities: (1) the CovRS-associated transcription complex may not lie at the centre of the nucleoid; therefore the actively transcribed genes are concentrated in a relatively narrow chromosomal region in order to be within the reach of the transcription complex; (2) the loops or coils may be restricted to specific lengths, such that some active genes can only be within the reach of the complexes by moving to specific chromosomal locations; (3) the circularization of the bacterial chromosomes and supercoils formed at the replication forks may restrict the packaging of the nucleoid; hence, the distal genes move to a region with a relaxed level of chromosomal folding. Based on these hypotheses, we propose that the actively transcribed genes may benefit from re-clustering in the same leading strand with enhanced proximity to the CovRS-associated transcription complex, and may represent a functional selection for efficient transcription of the genes required for growth and pathogenesis in that particular CovS-mutant strain. In actuality, another group of abundantly transcribed genes encoding ribosomal proteins in prokaryotes are commonly clustered proximately in the rightward-moving leading strand. This is another example of evolutionary selection for clustering in the same locus in the chromosome in addition to the selection for co-directional transcription and replication (Brewer, 1988). A similar distribution pattern for ribosomal proteins is also observed in M23ND (genes highlighted with light blue in Fig. 6c).

DISCUSSION

Genomic rearrangements mediated by homologous recombination are known to be diversification and adaptation strategies in bacteria. In the current study, we have presented unique architectures of genomic rearrangements in the hypervirulent GAS strain, M23ND, showing that the recombination events are highly asymmetric. This feature makes it distinct from other GAS genomes and allows a deeper understanding of the biological events underlying survival strategies in virulent bacteria.

Based on comparative studies, we propose in this communication a model of the major rearrangement events that may have occurred in the M23ND genome and identify the locations and sequences of the rearrangement junctions precisely. We also note that it is not possible for us to determine unambiguously the temporal order of the rearrangement events due to the limited number of known genomes of GAS and the lack of a complete understanding of the phylogenetic history. Multiple intermediate recombination events may have occurred, and alternative models may exist, depending on the temporal nature of the events. Previous consensus hypotheses proposed that the quasi-symmetric genomic rearrangements pivotal to the replication axis are much more pervasive probably due to replication-selection pressure, and many rounds of quasi-symmetric rearrangements may result in a significant randomized gene reordering. However, the present large-scale genomic reorganization is not likely to be the accumulated consequence of many small quasi-symmetric rearrangements. The combined complex of

multiple recombination events has likely reshaped the genome of M23ND in a discrete manner.

The genome of M23ND seems in a state of dynamic motion. The translocation induced by rRNA-*comX*, the inversion induced by the transposon IS1239 and the large inversion by phage repeat elements are shown to be reversible, based on PCR verification and phage induction. This indicates that M23ND is a heterogeneous population with one dominant and stable genomic organization probably under selective pressure. The heterogeneity in virulent strains has been reported in an epidemic GAS strain M12-HKU (Tse *et al.*, 2012) and a β-lactam antibiotic resistant *Staphylococcus aureus* strain Mu50Ω (Cui *et al.*, 2012). We note that we were unable to isolate a single living colony with any recombination event reverted. The genotypes and morphogenesis of M23ND were not changed by cultures or passages during a long period of time. This indicates that the current genome architecture is stable and has been adaptively fixed during evolution.

Regardless of the mediator of the homologous recombination, the overall consequence reveals the integrity of the genome as well as processes that select for genetic advantages from the diverse recombination events. We observe that the genome reorganization caused the co-localization of a set of actively transcribed genes to one of the leading strands, with a highly significant *p*-value ($\sim 10^{-10}$). We further hypothesize that the genes were clustered under selection pressures to share the same transcription complex containing at least the global transcriptional regulator, CovRS, for GAS pathogenesis. In this regard, the collective multifaceted genome rearrangements suggest a selection advantage for efficient transcription of genes required for growth, virulence and persistence in specific host environments, without affecting the overall genomic integrity. Specifically, the repositioning of the virulence genes *viz., sfb1, nga, slo* and *prtS* allowed their clustering with *speB*, the *mga* regulon, and the *hasABC* operon, the appropriate regulation of which is indispensable for the virulence of *S. pyogenes* strains. Similar relocations of *prtS*, which are essential for resistance to neutrophil killing, were also induced from inversions in strains causing epidemic outbreaks or severe diseases, *viz.*, M12-HKU16 (Tse *et al.*, 2012), M3-SSI-1 (Nakagawa *et al.*, 2003) and M5-Manfredo (Holden *et al.*, 2007). Those observations indicate that the replication- and transcription-associated chromosomal rearrangements are dependent on the bacterial phenotypes or host infectious niches, and may signify recently emergent events. It is noteworthy that the *mga* regulon and *speB* are uniformly localized in the left-moving leading strand proximal to the replication origin in all the GAS strains. Their precise regulation and proper expression are essential for the full virulence of GAS, and the conserved locations may result from a long-term evolutionary consequence. In conclusion, we have provided herein a comprehensive study of the genomic properties of a variety of genomic rearrangements that have influenced the current evolutionary status of a virulent strain of *S. pyogenes*. We offer this as an example of complex rearrangements that shape the genome architecture.

ACKNOWLEDGEMENTS

This work was supported by HL013423 from the NIH.

REFERENCES

Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Pontén, T., Alsmark, U. C., Podowski, R. M., Näslund, A. K., Eriksson, A. S., Winkler, H. H. & other authors (1998). The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature* **396**, 133–140.

Banks, D. J., Porcella, S. F., Barbian, K. D., Beres, S. B., Philips, L. E., Voyich, J. M., DeLeo, F. R., Martin, J. M., Somerville, G. A. & other authors (2004). Progress toward characterization of the group A Streptococcus metagenome: complete genome sequence of a macrolide-resistant serotype M6 strain. *J Infect Dis* **190**, 727–738.

Bao, Y., Liang, Z., Booyjzsen, C., Mayfield, J. A., Li, Y., Lee, S. W., Ploplis, V. A., Song, H. & Castellino, F. J. (2014). Unique genomic arrangements in an invasive serotype M23 strain of Streptococcus pyogenes identify genes that induce hypervirulence. *J Bacteriol* **196**, 4089–4102.

Bao, Y. J., Liang, Z., Mayfield, J. A., Lee, S. W., Ploplis, V. A. & Castellino, F. J. (2015). CovRS-regulated transcriptome analysis of a hypervirulent M23 strain of Group A *Streptococcus* pyogenes provides new insights into virulence determinants. *J Bacteriol* **197**, 3191–3205.

Bartlett, J., Blagojevic, J., Carter, D., Eskiw, C., Fromaget, M., Job, C., Shamsher, M., Trindade, I. F., Xu, M. & other authors (2006). Specialized transcription factories. *Biochem Soc Symp*, 67–75.

Berge, A., Rasmussen, M. & Björck, L. (1998). Identification of an insertion sequence located in a region encoding virulence factors of Streptococcus pyogenes. *Infect Immun* **66**, 3449–3453.

Bessen, D. E., McShan, W. M., Nguyen, S. V., Shetty, A., Agrawal, S. & Tettelin, H. (2015). Molecular epidemiology and genomics of group A Streptococcus. *Infect Genet Evol* **33**, 393–418.

Brewer, B. J. (1988). When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. *Cell* **53**, 679–686.

Campbell, A. M. (1992). Chromosomal insertion sites for phages and plasmids. *J Bacteriol* **174**, 7495–7499.

Canchaya, C., Desiere, F., McShan, W. M., Ferretti, J. J., Parkhill, J. & Brussow, H. (2002). Genome analysis of an inducible prophage and prophage remnants integrated in the Streptococcus pyogenes strain SF370. *Virology*. **302**, 245–258.

Carver, T., Berriman, M., Tivey, A., Patel, C., Böhme, U., Barrell, B. G., Parkhill, J. & Rajandream, M. A. (2008). Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24**, 2672–2676.

Chakalova, L., Debrand, E., Mitchell, J. A., Osborne, C. S. & Fraser, P. (2005). Replication and transcription: shaping the landscape of the genome. *Nat Rev Genet* **6**, 669–677.

Cook, P. R. (2002). Predicting three-dimensional genome structure from transcriptional activity. *Nat Genet* **32**, 347–352.

Cournac, A. & Plumbridge, J. (2013). DNA looping in prokaryotes: experimental and theoretical approaches. *J Bacteriol* **195**, 1109–1119.

Cui, L., Neoh, H. M., Iwamoto, A. & Hiramatsu, K. (2012). Coordinated phenotype switching with large-scale chromosome flip-flop inversion observed in bacteria. *Proc Natl Acad Sci U S A* **109**, E1647–E1656.

Darling, A. C., Mau, B., Blattner, F. R. & Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**, 1394–1403.

Darling, A. E., Miklós, I. & Ragan, M. A. (2008). Dynamics of genome rearrangement in bacterial populations. *PLoS Genet* **4**, e1000128.

Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**, 2478–2483.

Diruggiero, J., Dunn, D., Maeder, D. L., Holley-Shanks, R., Chatard, J., Horlacher, R., Robb, F. T., Boos, W. & Weiss, R. B. (2000). Evidence of recent lateral gene transfer among hyperthermophilic archaea. *Mol Microbiol* **38**, 684–693.

Eisen, J., Heidelberg, J., White, O. & Salzberg, S. (2000). Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol* **1**, research0011.0011–research0011.0019.

French, S. (1992). Consequences of replication fork movement through transcription units in vivo. *Science* **258**, 1362–1365.

Graham, M. R., Smoot, L. M., Migliaccio, C. A., Virtaneva, K., Sturdevant, D. E., Porcella, S. F., Federle, M. J., Adams, G. J., Scott, J. R. & other authors (2002). Virulence control in group A Streptococcus by a two-component gene regulatory system: global expression profiling and in vivo infection modeling. *Proc Natl Acad Sci U S A* **99**, 13855–13860.

Green, N. M., Zhang, S., Porcella, S. F., Nagiec, M. J., Barbian, K. D., Beres, S. B., LeFebvre, R. B. & Musser, J. M. (2005). Genome sequence of a serotype M28 strain of group a streptococcus: potential new insights into puerperal sepsis and bacterial disease specificity. *J Infect Dis* **192**, 760–770.

Grigoriev, A. (2000). Graphical genome comparison: rearrangements and replication origin of Helicobacter pylori. *Trends Genet* **16**, 376–378.

Haack, K. R. & Roth, J. R. (1995). Recombination between chromosomal IS200 elements supports frequent duplication formation in Salmonella typhimurium. *Genetics* **141**, 1245–1252.

Hatfull, G. F. (2008). Bacteriophage genomics. *Curr Opin Microbiol* **11**, 447–453.

Helmrich, A., Ballarino, M., Nudler, E. & Tora, L. (2013). Transcription-replication encounters, consequences and genomic instability. *Nat Struct Mol Biol* **20**, 412–418.

Holden, M. T., Scott, A., Cherevach, I., Chillingworth, T., Churcher, C., Cronin, A., Dowd, L., Feltwell, T., Hamlin, N. & other authors (2007). Complete genome of acute rheumatic fever-associated serotype M5 Streptococcus pyogenes strain manfredo. *J Bacteriol* **189**, 1473–1477.

Horstmann, N., Sahasrabhojane, P., Saldaña, M., Ajami, N. J., Flores, A. R., Sumby, P., Liu, C. G., Yao, H., Su, X. & other authors (2015). Characterization of the effect of the histidine kinase CovS on response regulator phosphorylation in group A Streptococcus. *Infect Immun* **83**, 1068–1077.

Komoda, Y., Enomoto, M. & Tominaga, A. (1991). Large inversion in Escherichia coli K-12 1485IN between inversely oriented IS3 elements near lac and cdd. *Genetics* **129**, 639–645.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessières, P., Bolotin, A. & other authors (1997). The complete genome sequence of the gram-positive bacterium Bacillus subtilis. *Nature* **390**, 249–256.

Lee, M. S. & Morrison, D. A. (1999). Identification of a new regulator in Streptococcus pneumoniae linking quorum sensing to competence for genetic transformation. *J Bacteriol* **181**, 5004–5016.

Liu, S. L. & Sanderson, K. E. (1995). Rearrangements in the genome of the bacterium Salmonella typhi. *Proc Natl Acad Sci U S A* **92**, 1018–1022.

Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* **13**, 660–665.

Mahillon, J. & Chandler, M. (1998). Insertion sequences. *Microbiol Mol Biol Rev* **62**, 725–774.

McShan, W. M. & Ferretti, J. J. (1997). Genetic diversity in temperate bacteriophages of Streptococcus pyogenes: identification of a second attachment site for phages carrying the erythrogenic toxin A gene. *J Bacteriol* **179**, 6509–6511.

McShan, W. M., Tang, Y. F. & Ferretti, J. J. (1997). Bacteriophage T12 of Streptococcus pyogenes integrates into the gene encoding a serine tRNA. *Mol Microbiol* **23**, 719–728.

Merrikh, H., Zhang, Y., Grossman, A. D. & Wang, J. D. (2012). Replication-transcription conflicts in bacteria. *Nat Rev Microbiol* **10**, 449–458.

Nakagawa, I., Kurokawa, K., Yamashita, A., Nakata, M., Tomiyasu, Y., Okahashi, N., Kawabata, S., Yamazaki, K., Shiba, T. & other authors (2003). Genome sequence of an M3 strain of Streptococcus pyogenes reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome Res* **13**, 1042–1055.

Novick, R. P., Christie, G. E. & Penadés, J. R. (2010). The phage-related chromosomal islands of Gram-positive bacteria. *Nat Rev Microbiol* **8**, 541–551.

Okazaki, R., Okazaki, T., Sakabe, K. & Sugimoto, K. (1967). Mechanism of DNA replication possible discontinuity of DNA chain growth. *Jpn J Med Sci Biol* **20**, 255–260.

Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J. A., Lopes, S. & other authors (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet* **36**, 1065–1071.

Papantonis, A. & Cook, P. R. (2013). Transcription factories: genome organization and gene regulation. *Chem Rev* **113**, 8683–8705.

Parkhill, J., Wren, B. W., Thomson, N. R., Titball, R. W., Holden, M. T., Prentice, M. B., Sebaihia, M., James, K. D., Churcher, C. & other authors (2001). Genome sequence of Yersinia pestis, the causative agent of plague. *Nature* **413**, 523–527.

Read, T. D., Brunham, R. C., Shen, C., Gill, S. R., Heidelberg, J. F., White, O., Hickey, E. K., Peterson, J., Utterback, T. & other authors (2000). Genome sequences of Chlamydia trachomatis MoPn and Chlamydia pneumoniae AR39. *Nucleic Acids Res* **28**, 1397–1406.

Roberts, J. W. & Roberts, C. W. (1975). Proteolytic cleavage of bacteriophage lambda repressor in induction. *Proc Natl Acad Sci U S A* **72**, 147–151.

Rocha, E. P., Danchin, A. & Viari, A. (1999). Universal replication biases in bacteria. *Mol Microbiol* **32**, 11–16.

Rocha, E. (2002). Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol* **10**, 393–395.

Sakabe, K. & Okazaki, R. (1966). A unique property of the replicating region of chromosomal DNA. *Biochim Biophys Acta* **129**, 651–654.

Scott, J., Thompson-Mayberry, P., Lahmamsi, S., King, C. J. & McShan, W. M. (2008). Phage-associated mutator phenotype in group A streptococcus. *J Bacteriol* **190**, 6290–6301.

Smoot, J. C., Barbian, K. D., Van Gompel, J. J., Smoot, L. M., Chaussee, M. S., Sylva, G. L., Sturdevant, D. E., Ricklefs, S. M., Porcella, S. F. & other authors (2002). Genome sequence and comparative microarray analysis of serotype M18 group A Streptococcus strains associated with acute rheumatic fever outbreaks. *Proc Natl Acad Sci U S A* **99**, 4668–4673.

Tillier, E. R. & Collins, R. A. (2000a). The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J Mol Evol* **50**, 249–257.

Tillier, E. R. M. & Collins, R. A. (2000b). Genome rearrangement by replication-directed translocation. *Nature Genetics* **26**, 195–197.

Tse, H., Bao, J. Y., Davies, M. R., Maamary, P., Tsoi, H. W., Tong, A. H., Ho, T. C., Lin, C. H., Gillen, C. M. & other authors (2012). Molecular characterization of the 2011 Hong Kong scarlet fever outbreak. *J Infect Dis* **206**, 341–351.

Wang, W., Li, G. W., Chen, C., Xie, X. S. & Zhuang, X. (2011). Chromosome organization by a nucleoid-associated protein in live bacteria. *Science* **333**, 1445–1449.

Zawilak, A., Cebrat, S., Mackiewicz, P., Król-Hulewicz, A., Jakimowicz, D., Messer, W., Gosciniak, G. & Zakrzewska-Czerwinska, J. (2001). Identification of a putative chromosomal replication origin from Helicobacter pylori and its interaction with the initiator protein DnaA. *Nucleic Acids Res* **29**, 2251–2259.

Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**, 203–214.

Zivanovic, Y., Lopez, P., Philippe, H. & Forterre, P. (2002). Pyrococcus genome comparison evidences chromosome shuffling-driven evolution. *Nucleic Acids Res* **30**, 1902–1910.

Edited by: T. Msadek