Bridging Chromosomal Architecture and Pathophysiology of *Streptococcus pneumoniae*

Antonio J. Martín-Galiano[1,*], María J. Ferrándiz[1], and Adela G. de la Campa[1,2]

[1]Unidad de Genética Bacteriana, Centro Nacional de Microbiología, Instituto de Salud Carlos III, Madrid, Majadahonda, Spain

[2]Unidad de Genética Bacteriana, Presidencia, Consejo Superior de Investigaciones Científicas, Madrid, Spain

*Corresponding author: E-mail: mgaliano@isciii.es.

Accepted: December 22, 2016

Abstract

The chromosome of *Streptococcus pneumoniae* is organized into topological domains based on its transcriptional response to DNA relaxation: Up-regulated (UP), down-regulated (DOWN), nonregulated (NR), and AT-rich. In the present work, NR genes found to have highly conserved chromosomal locations (17% of the genome) were categorized as members of position-conserved nonregulated (pcNR) domains, while NR genes with a variable position (36% of the genome) were classified as members of position-variable nonregulated (pvNR) domains. On average, pcNR domains showed high transcription rates, optimized codon usage, and were found to contain only a small number of RUP/BOX/SPLICE repeats. They were also poor in exogenous genes but enriched in leading strand genes that code for proteins involved in primary metabolism with central roles within the interactome. In contrast, pvNR genes coding for cell wall proteins, paralogs, virulence factors and immunogenic candidates for protein-based vaccines were found to be overrepresented. DOWN domains were enriched in genes essential for infection. Many UP and DOWN domain genes were seen to be activated during different stages of competence, whereas pcNR genes tended to be repressed until the competence was switched off. Pneumococcal genes appear to be subject to a topology-driven selection pressure that defines the chromosomal location of genes involved in metabolism, virulence and competence. The pcNR domains are interleaved between UP and DOWN domains according to a pattern that suggests the existence of macrodomain entities. The term "topogenomics" is here proposed to describe the study of the topological rules of genomes and their relationship with physiology.

Key words: interactome, topoisomerase, topological domain, virulence factor, X-state.

Introduction

The availability of a number of genomic sequences per species and genome-scale data (transcriptomics, interactomics, and metabolomics) ushered the postgenomic age. The answers to biological questions beyond those that can be provided by individual gene analysis can now be sought through the use of different technologies till now used only in an independent fashion. One such question is how chromosomes are topologically organized.

The bacterial chromosome is compacted by three orders of magnitude (Holmes and Cozzarelli 2000); this allows genomes of typically 1–10 Mb (with all genetic instructions to sustain life) to be accommodated within cells just 1–2 μm long. Chromosome topology depends on DNA supercoiling that must be able to show homeostatic responses to challenges such as osmotic stress, growth phase-dependent changes, topoisomerases-targeting drugs, and even circadian cycles (Woelfle et al. 2007). Supercoiling dynamics must therefore be precisely managed.

Chromosome topology is controlled by a four-tier system. The first tier involves the action of DNA topoisomerases, such as gyrase, topoisomerase I (Topo I) and topoisomerase IV (Topo IV), which actively regulate the degree of supercoiling by introducing or removing DNA turns (Champoux 2001). The second involves a number of nucleoid-associated proteins (NAPs) (Wang et al. 2013) that together form a functional network that maintains the topology of the DNA via its bending, wrapping, or bridging (Dorman 2013). In addition, NAPs regulate transcription by constraining supercoils. The third tier involves the local curvature of the DNA affecting the transcription of key genes, such as promoters that regulates genes

coding for topoisomerases (Balas et al. 1998). Finally, the fourth involves the organization of the chromosome into domains with intrinsic topological behavior (Ferrándiz et al. 2010, 2016; Higgins et al. 1996; Postow et al. 2004; Sinden and Pettijohn 1981; Worcel and Burgi 1972). These control layers are in permanent crosstalk, providing the required dynamics and precision of chromosome supercoiling.

The complex feedback control over the chromosome architecture has, however, prevented a reasonable understanding of how its topology is reached (Wright et al. 2007). Differences in gene pool, genome size, NAPs, topoisomerases, and AT content make the topological rules governing the chromosome of an organism difficult to extrapolate beyond the limits of its genus. For example, three levels of intradomain hierarchy—macro, middle, and micro—have been reported for chromosomal domains in *Escherichia coli* but the ≥40 NAPs, 5 Mb genome and ~50% AT content of this bacterium may be very different in others, making extrapolations risky.

In *Streptococcus pneumoniae*, changes in DNA supercoiling imposed by treatment with sublethal doses of GyrB and Topo I inhibitors have revealed topology-reactive domains (Ferrándiz et al. 2010, 2014, 2016). Indeed, the topology of this bacterium's chromosome involves transcriptional gene clusters that show concerted reactions to the relaxation caused by novobiocin (NOV), which targets GyrB (Ferrándiz et al. 2010). Together with GyrA, GyrB makes up gyrase, the only enzyme able to introduce negative supercoils (Gellert et al. 1976). Four types of domains, defined by their expression behavior in response to NOV have been identified: Up-regulated (UP), down-regulated (DOWN), nonregulated (NR), and flanking (Ferrándiz et al. 2010). Flanking domains are adjacent to DOWN domains; they are especially AT-rich and encode nonessential functions, and were suggested to have a role as structural DNA. It is possible that prompt relaxation due to their high AT content reduces transcription in nearby regions under conditions of topological stress.

The genus *Streptococcus* contains some of the most important human and anthropozoonotic pathogens, including *S. pneumoniae*, *Streptococcus pyogenes*, and *Streptococcus suis*. All of these carry a number of virulence factors such as adhesins, anti-phagocytic capsules, and toxins. The chromosomes of these species are of similar size and AT content, so they may share topological features. In addition, most, if not all, streptococcal species harbor competence systems responsible for the acquisition of foreign DNA (Johnsborg et al. 2007). *S. pneumoniae* has an open pangenome with every new strain sequenced adding new genes (Donati et al. 2010). It is reasonable to suppose that the long-term permanence of any novel genetic material is only secure if the changes in the topological order maintain the supercoiling balance. New, horizontally acquired virulence factors need to be placed in appropriate topological areas to be properly transcribed and functionally efficient.

Given the above, it is of interest to understand the relationships between chromosomal topology, central physiology and pathobiology of streptococci. The present work, which involved a comprehensive computational genomic analysis, reveals genes coding for most core biosynthetic proteins in *S. pneumoniae* that fall into a novel type of domain nonreactive to DNA relaxation. Such domains are here named position-conserved nonregulated (pcNR) domains. However, genes coding for virulence factors are more common in the remaining position-variable NR (pvNR) zones. These findings suggest the existence of a pressure that selects against the perturbation of supercoiling in areas where metabolic and virulence factors genes are found.

Materials and Methods

Calculation of the Normalized Location Dispersion Index

Homologs between species were found using Get_Homologues software (Contreras-Moreira and Vinuesa 2013), employing the default parameters. For *S. pneumoniae* R6 genes with homologues in at least 10 species, the distance to the origin was calculated in degrees (ranging from 0° to 180°) in the 10 genomes and the average calculated. The location dispersion index (LDI) was defined as the standard deviation of these averages. For R6 genes with homologues in more than 10 species, only the 10 species with the most similar homologs, in terms of BLAST scores, were taken into account in LDI calculations. When these rough LDI values were mapped according to their distance to the origin (using a window of 11 genes and a step of 10 genes), and the top 20% was selected, a strong correlation was seen between LDI and distance to origin. This relationship followed a two-order polynomial adjustment: $y = -2E{-}5x^2 + 0.039x + 4.38$, $r^2 = 0.73$ (supplementary fig. S1, Supplementary Material online). The rough LDI values were therefore normalized via their division by this baseline to render the final normalized LDI (nLDI) values. An nLDI value of 1 denotes the average genome value. pcNR zones were considered as those containing ≥25 contiguous genes in which ≥44% of those genes showed normalized nLDI values ≤1 (χ^2 test $P \leq 0.01$, the null hypothesis is: The cluster contains 29% of genes with nLDI < 1, like the average of the genome), and which in addition were not included in clusters affected by NOV treatment.

Data Acquisition

Genomic information was obtained from NCBI resources (NCBI Resource Coordinators 2016). Transcriptional units predicted with Price algorithm (Price et al. 2005) were downloaded from http://www.microbesonline.org. A transcriptional unit was considered included in a pcNR domain if the cluster contained ≥50% of its genes. A transcriptional unit was considered as positionally conserved if ≥50% of its genes showed nLDI values <1. Paralogs were obtained using BLAST employing cut-offs of ≥ 60% identity over ≥80% of the protein length. Gene ontology data were obtained from GO webpage (The Gene Ontology Consortium 2015). Lateral-transferred genes were downloaded from the horizontal gene transfer (HGT)-DB

(García-Vallve et al. 2000). Essential genes were identified by a former study involving transposon insertion libraries (van Opijnen and Camilli 2012). BOX, RUP, and SPRITE repeats were annotated using the software provided by Croucher et al. (2011). The condon adaptation index (CAI) of a gene was calculated using the algorithm of Sharp and Li (1987) with optimal codon usage of genes encoding ribosomal proteins for *S. pneumoniae* (Martín-Galiano et al. 2004). Protein–protein interactions (PPIs) were downloaded from STRING database (Szklarczyk et al. 2015) applying a score threshold $\geq$0.7. Virulence factors were downloaded from VFDB (Chen et al. 2012). Signature-tagged mutagenesis (STM) data were obtained from the original papers in which they appeared (Chen et al. 2008; Molzen et al. 2011). ANTIGENome information was acquired from the original manuscript (Giefing et al. 2008).

Results

Position-Conserved Nonregulated Domains: A Novel Type of Topological Domain

Investigations were made into whether the genes present in *S. pneumoniae* R6 topological domains, and their relative positions, were conserved across other strains of the same species. Finding this to be the case would be indicative of a selective pressure for maintaining the location of these genes in particular positions. A nLDI, normalized Location Dispersion Index, value was calculated. This parameter quantifies the position deviation of a given gene with respect to the replication origin and relative to its homologues in several genomes, and was further normalized by distance to origin (See Material and Methods). Genes showing nLDI values <1 tend to locate in stable positions on genomes regardless other biases imposed by the distance to origin of the genomic area they occupy. Therefore, a low nLDI value is suggestive of selective pressure operating on the gene to maintain a given position. The nLDI values across *S. pneumoniae* genomes were very small, because synteny is highly conserved in this species. Therefore, the nLDI was recalculated considering representative strains of 25 species of the *Streptococcus* genus (supplementary table S1, Supplementary Material online) considering that gene order is relatively conserved within this taxon as happens with Gammaproteobacteria (Sobetzko et al. 2012). It was assumed that chromosomal topology would be also roughly conserved across all these species because they share a substantial amount of their gene pools (Lefebure and Stanhope 2007), similarly, tight genome length ranges (1.8–2.4 Mb), and a similar AT content (56–64%). A total of 1216 genes coding proteins in *S. pneumoniae* R6, with homologs in $\geq$10 species, were considered in the analysis. The calculated nLDI values showed discriminatory differences. A total of 571 genes (28.0%) had nLDI values <1.

Several genes from UP and DOWN domains were found to be present in most streptococci and in comparable positions (nLDI < 1), and probably constitute the cores of the topological-reactive domains. In fact, the lowest nLDI values, that is, highest position conservation, was calculated for 40 genes around the origin, indicating this zone is extremely conserved in topological terms. Moreover, seven clusters with conserved positions (average nLDI: 0.880, 340 genes, 16.6% of the

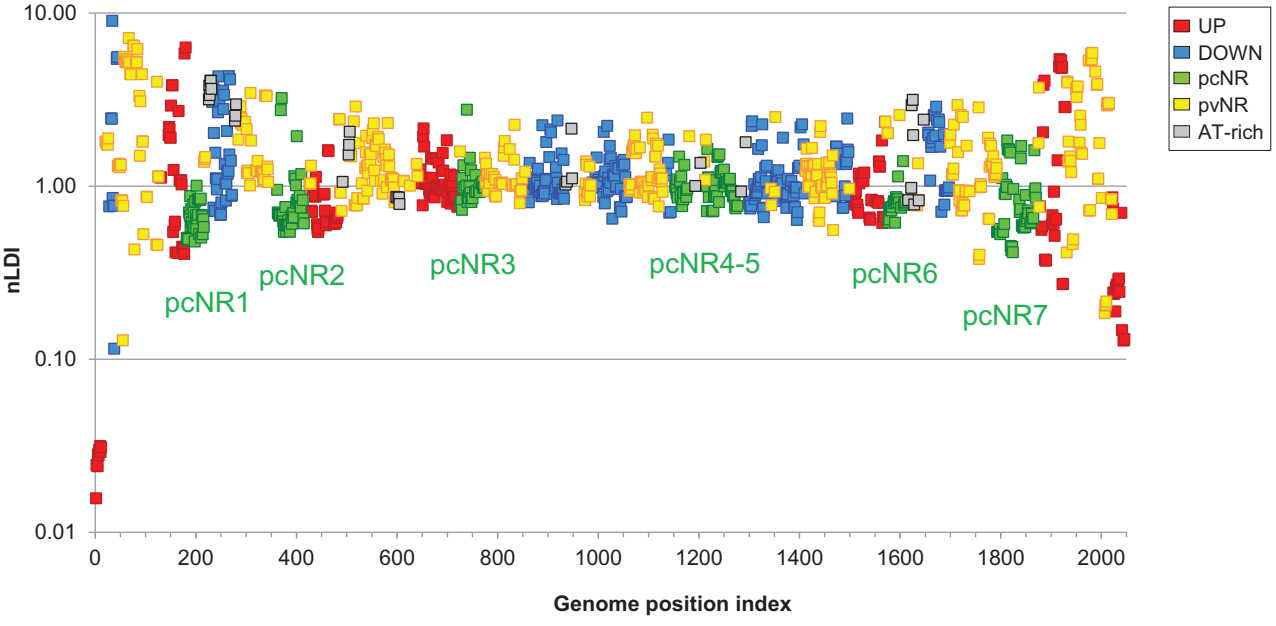

Fig. 1.—Location comparison of NOV-reactive and pcNR domains. Dots indicate nLDI values of *S. pneumoniae* R6 genes in the whole *Streptococcus* genus. Gene indexing start from the replication origin.

genome) were detected for NR genes, here referred to as pcNR (fig. 1). The remaining NR genes (average nLDI: 1.620, 731 genes, 35.8% of the genome) were termed pvNR. The analysis was repeated at the level of transcriptional units, considering operons and monocistronic genes as atomic units. Among 957 transcriptional units, 157 were located in pcNR clusters. The occurrence of transcriptional units with nLDI < 1 values in pcNR clusters was 55.4% (ranging 48–80% per cluster), ~2.5-fold higher than in the remaining genome, 22.7% ($P = 2 \times 10^{-23}$, χ^2 test). Interestingly, pcNR domains were located symmetrically at regular intervals (~200, 400, and 800 kb) from the replication origin. In addition, they were interleaved between UP, DOWN, and pvNR domains, suggesting a potential higher-order macrostructural unit above the domain level. The former flanking group was widened to all those AT-rich zones and renamed as "AT-rich" domains. This accounted for 13 clusters (209 ORFs, table 1) containing ≥10 genes. Of these genes, ≥75% had ≥62%AT and the whole AT-rich domain had on average %AT ≥ 64% (χ^2 $P < 0.001$, considering that only 29.3% of the *S. pneumoniae* R6 genes have ≥62%AT). Any other remaining section in the genome with high AT content were prevalently assigned to UP or DOWN domains or did not fulfil the required criteria. It is important to mention that pcNR clusters 4 and 5 are very close to one another and may in fact be one, although split by a AT-rich domain.

HGT is a primary source of evolution in prokaryotes, but introduction of new genetic material at random may perturb chromosomal topology. In *S. pneumoniae* R6, up to 12.1% of the genome is considered HGT-acquired (http://genomes.urv.cat/HGT-DB/). However, only 2.3% ($P = 2 \times 10^{-9}$; χ^2 test, compared with the remaining genome) of DOWN and 2.1% ($P = 2 \times 10^{-9}$) of pcNR genes were so-acquired (fig. 2). Thus, genes of the pcNR domains contribute to the conserved metabolic core, which has been present in the genome because speciation. In contrast, a large amount of genes (66.5%) in AT-rich domains ($P < 1 \times 10^{-100}$) were predicted to have been acquired by HGT. These data support the idea that AT-rich domains probably act as structural (Ferrándiz et al. 2010) or parasitic DNA, which agrees with their low transcriptional level and annotated functions (Ferrándiz et al. 2010, 2014, 2016).

pcNR Domains are Enriched in Genes Involved with Central Metabolism

A substantial fraction of pcNR genes were found to encode proteins with important roles in central metabolism (table 2). Analyses of several genetic features were performed in pcNR domains and compared with the results for the remaining topological classes. The number of pcNR genes in the lagging strand was 15.6% ($P = 0.03$; χ^2 analysis), significantly lower than the average in the remaining *S. pneumoniae* genome (22.3%). A total of 87.9% of pcNR genes encoded proteins containing Pfam domains, a value higher than that recorded for the average of the remaining genome (79.4%, $P = 2 \times 10^{-4}$) and for those in AT-rich domains (63.6%, $P = 2 \times 10^{-12}$). The genes of the pcNR domains also had few paralogs (arising through recent gene duplication) ($P = 0.04$), unlike those of the pvNR zones ($P = 1 \times 10^{-4}$). Accordingly, the fraction of essential pcNR genes was notably higher than those seen for UP, pvNR, and AT-rich domain genes (fig. 3).

The mRNA levels for each gene in exponential growth cultures not subjected to topological stress were estimated by RNA-Seq. These data were recorded in a previous study performed at our laboratory (Ferrándiz et al. 2016) and are available at the Gene Expression Omnibus repository (http://www.ncbi.nlm.nih.gov/geo) via accession number GSE77748. Reads were normalized by gene size (kb) using the reads per kilobase per million mapped reads (RPKM) (Mortazavi et al. 2008). The genes of pcNR domains showed high transcription levels ($P = 7 \times 10^{-28}$, two-tailed Student's *t*-test, compared with

Table 1
Statistical Measures of AT-Rich Domains

AT-Rich Domain	Spr Begin	Spr End	Number of Genes	Number of Genes %AT>62	% Genes %AT>62	Average %AT	*P*-Value
ATr-1	104	120	17	13	76.5	67.0	2×10^{-5}
ATr-2	223	231	9	9	100.0	65.6	3×10^{-6}
ATr-3	274	284	11	11	100.0	66.6	3×10^{-7}
ATr-4	348	360	13	10	76.9	64.1	2×10^{-4}
ATr-5	491	506	16	12	75.0	64.7	6×10^{-5}
ATr-6	599	611	13	10	76.9	64.7	2×10^{-4}
ATr-7	933	972	40	38	95.0	68.0	7×10^{-20}
ATr-8	1188	1210	23	22	95.6	66.1	3×10^{-12}
ATr-9	1280	1293	14	12	85.7	67.7	3×10^{-6}
ATr-10	1614	1630	17	13	76.5	64.6	2×10^{-5}
ATr-11	1639	1651	13	12	92.3	67.7	6×10^{-7}
ATr-12	1762	1774	13	10	76.9	67.0	2×10^{-4}
ATr-13	1965	1974	10	10	100.0	65.5	9×10^{-7}

NOTE.—The *P*-value was calculated by χ^2 test.

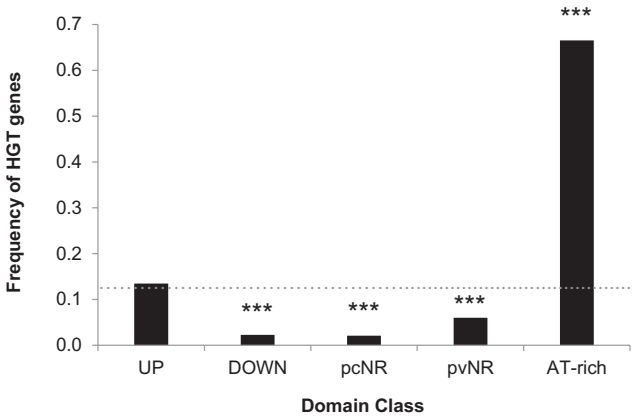

FIG. 2.—HGT abundance and domain class. The dashed lines indicate the genome average. Statistical significance: ***$P \leq 0.001$.

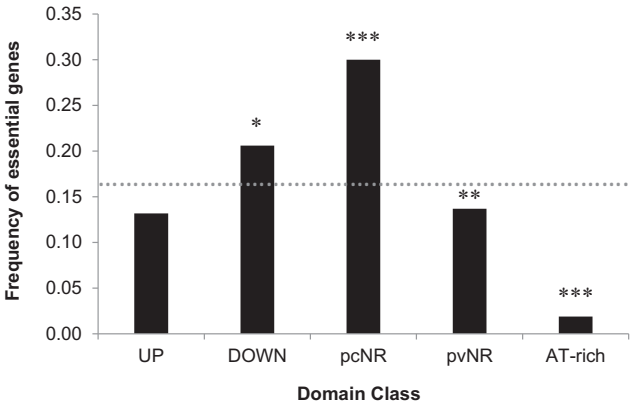

FIG. 3.—Fraction of essential genes in different domains. The dashed lines indicate the genome average. Statistical significance is indicated as in fig. 2. Statistical significance: * $P \leq 0.05$; ** $P \leq 0.01$; ***$P \leq 0.001$.

the data from the other four domain classes) being, for instance, 3.2 times the number of transcripts recorded for pvNR genes. Although this was particularly noticeable in pcNR1 and pcNR2 domains, most pcNR clusters also produced larger mRNA amounts than adjacent domains (fig. 4). This indicates that strong transcription of pcNR genes is a generic pattern. In contrast, the RPKM values for the AT-rich genes were low ($P = 4 \times 10^{-5}$) 2.9 times below the pvNR gene average. In fact, 74% AT-rich genes had RPKM < 1,000.

The abundance of transcription-repressive short repeats in different domains could be a factor indicating gene expression level. Three kinds of repeat elements—BOX, RUP, and SPRITE—have been described in *S. pneumoniae*, all associated with repressive control of expression (Croucher et al. 2011). A total of 58 BOX, 61 RUP, and 18 SPRITE elements were detected either within R6 coding sequences or near (within 200 bp) their start codons. They were evenly distributed over the R6 chromosome. The exceptions were the pcNR domains, which contained roughly half each kind of repeat compared with the remaining genome (combined $P = 2 \times 10^{-3}$; χ^2 test) whereas the pvNR domains contained 1.4-fold more repeats (combined $P = 7 \times 10^{-3}$) (fig. 5). Short repeats had an AT content of 61%, 69% and 68%, respectively, although AT-rich zones are not specially enriched in them. Despite no association being found between gene function and presence of short repeats when no topological criteria were applied (Croucher et al. 2011), the pcNR areas showed low tolerance to such elements.

Two additional parameters were calculated to assess the potential protein abundance and the relevance of the coding genes. The first one, the codon adaptation index, CAI, quantifies the occurrence of efficient isocodons for translation

Table 2

pcNR Domains, Statistical Measures, and Mean Functions

pcNR Domain	Spr Begin	Spr End	Number of Genes[a]	Number of Genes nLDI<1[a]	% Genes nLDI<1[a]	*P*-value[a]	Median nLDI	Mean Functions
pcNR1	182	216	35 (5)	32 (4)	91.4 (80.0)	6×10^{-17}(2×10^{-3})	0.669	Translation
pcNR2	362	415	54 (29)	39 (18)	72.2 (62.1)	4×10^{-13}(4×10^{-7})	0.679	Fatty acid biosynthesis, translation, branched-amino acid biosynthesis
pcNR3	726	774	49 (22)	22 (11)	44.9 (50.0)	8×10^{-3}(2×10^{-3})	0.978	Diverse
pcNR4	1153	1177	25 (14)	15 (9)	60.0 (64.3)	4×10^{-4}(2×10^{-4})	0.948	Diverse (Pyrimidine biosynthesis)
pcNR5	1215	1277	63 (25)	28 (12)	44.4 (48.0)	3×10^{-3}(3×10^{-3})	0.872	Shikimate pathway, N-acetyl-glucosamine metabolism
pcNR6	1577	1611	35 (24)	31 (13)	52.5 (54.2)	6×10^{-7}(8×10^{-4})	0.798	DNA replication, RNA metabolism, aromatic amino acid biosynthesis
pcNR7	1793	1870	77 (38)	44 (20)	57.1 (52.6)	1×10^{-8}(1×10^{-5})	0.650	Competence, translation, sugar metabolism and transport

NOTE.—The *P*-value was calculated by χ^2 test.

[a]Numbers in brackets correspond to data from transcriptional units.

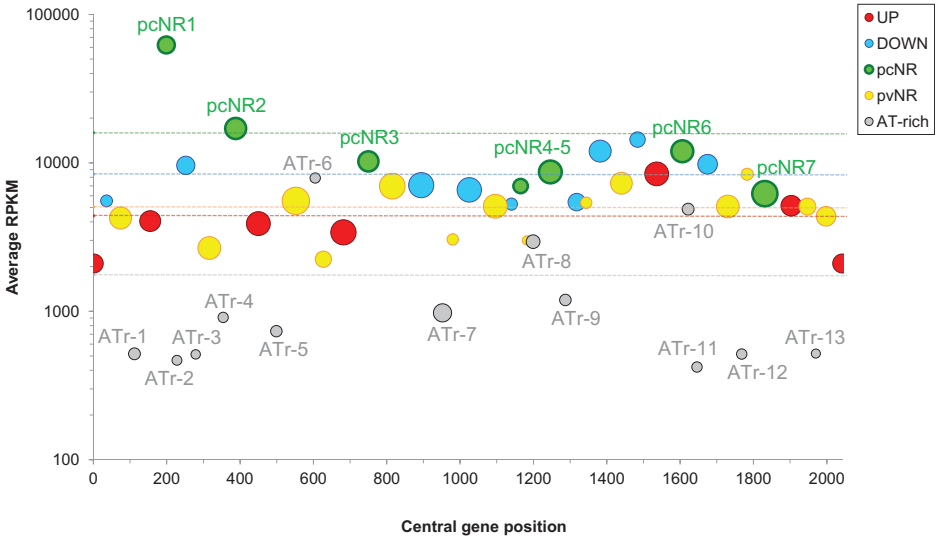

Fig. 4.—Abundance of RNA transcripts according to domain type. RNA-Seq data are expressed as RPKM values. The positions of the domains in the genome are represented by the position of the central gene (central gene position, *X*-axis). The size of the circles is proportional to the number of genes in the domain. Only pvNR clusters with at least 10 members are shown. Horizontal lines in the corresponding color indicate the average RPKM value for each domain class. ATr: AT-rich.

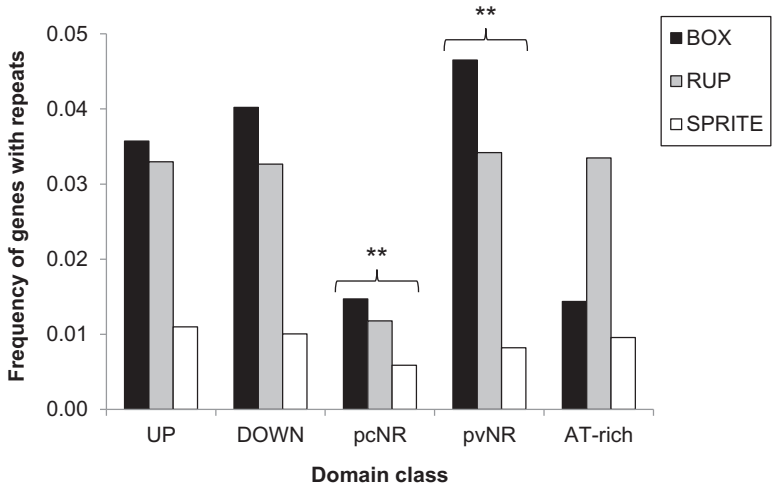

Fig. 5.—Frequency of repeat sequences. Statistical significance: $^{**}P \leq 0.01$.

elongation in the gene sequence. CAIs are related with the translation rate and mRNA levels (Martín-Galiano et al. 2004), and, consequently, to the amount of protein produced. The second one, the PPIs indicates the number of predicted functional relationships of a given protein with other polypeptides and, therefore, denote the connectivity of a protein within the protein network. PPIs provide a rough estimate of the importance of the protein in cell physiology, because the protein network is very sensitive to the removal of highly connected nodes (Jeong et al. 2001). The average values for CAI and PPIs for *S. pneumoniae* were 0.346 ± 0.120 and 29.5 ± 37.1, respectively. Genes in pcNR domains showed a distribution shifted toward higher CAIs ($P = 5 \times 10^{-6}$; two-tailed *t*-test) and PPIs 1.6 times higher than those of the remaining genome ($P = 2 \times 10^{-13}$). These results may be biased by pcNR1, which is composed by genes of ribosomal proteins. However, the PPIs for genes in the pcNR domains were very different, even in low CAI ranges (<0.3 and 0.3–0.4, with $P = 0.008$ and 1×10^{-6}, respectively) (fig. 6). These results reinforce the notion that genes in pcNR domains are more involved in the central metabolic network than those in pvNR domains, irrespective of their abundance. In stark

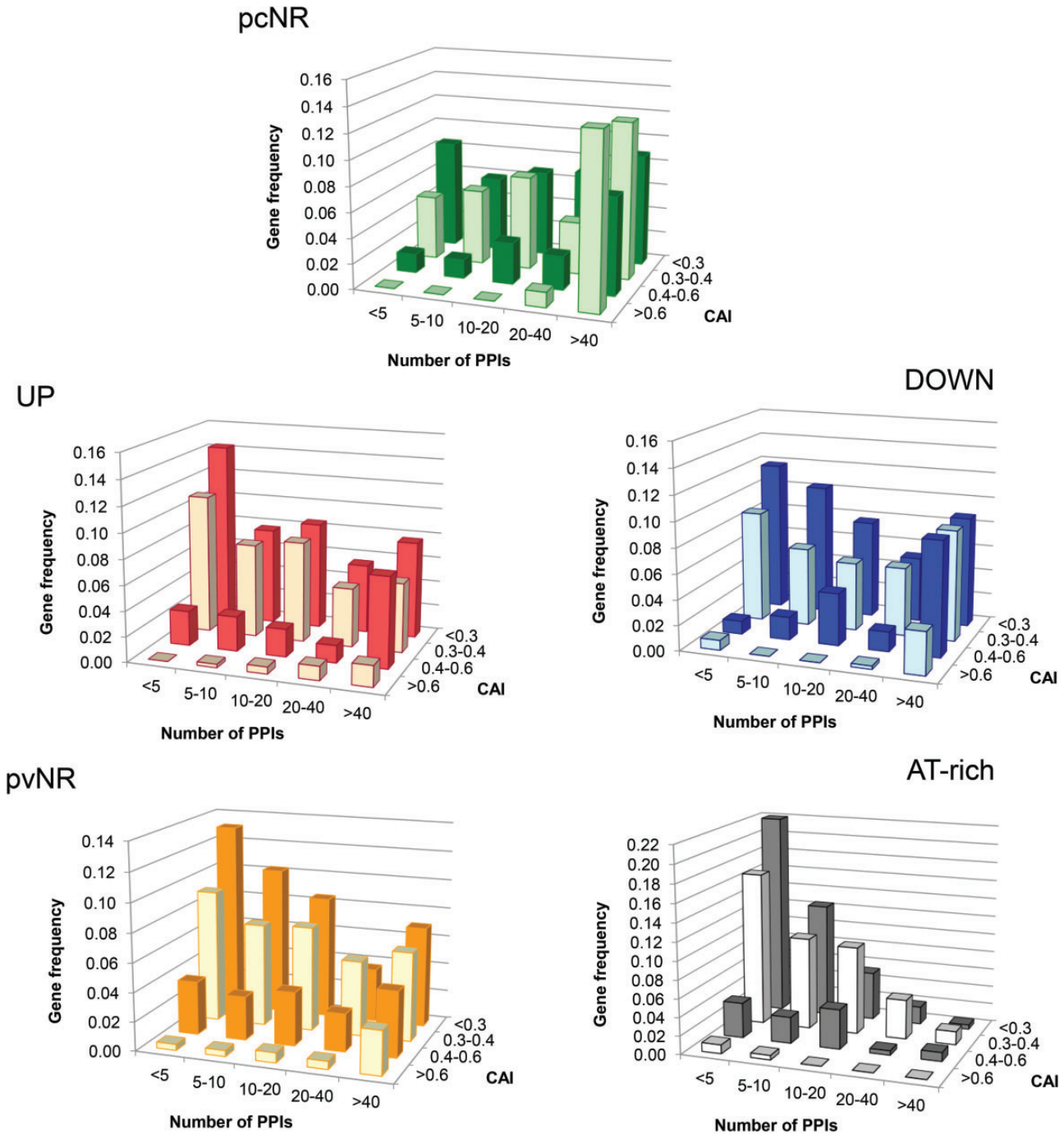

FIG. 6.—Number of PPIs vs CAI in different domain types. Dark/light bar colors within charts were used simply for reasons of clarity.

contrast, AT-rich genes appear to play little or no role in the central metabolic network; their PPIs reached roughly one third of the average of the remaining genome ($P = 1 \times 10^{-15}$). Overall, data suggest that function, expression, essentiality and stability of the genomic position are connected issues as previously claimed for *Dickeya dadantii* and *Escherichia coli* (Jiang et al. 2015; Sobetzko et al. 2012).

pcNR Genes Are Not Commonly Involved in Pathogenesis

The relationship between virulence and chromosomal topology was examined via the inspection of three kinds of gene. First, genuine virulence factors were examined. The pvNR domains produced more of these factors than the remaining genome ($P = 2 \times 10^{-11}$; χ^2 test) (fig. 7*A*). In contrast, pcNR and UP zones produced relatively few ($P = 3 \times 10^{-2}$ and

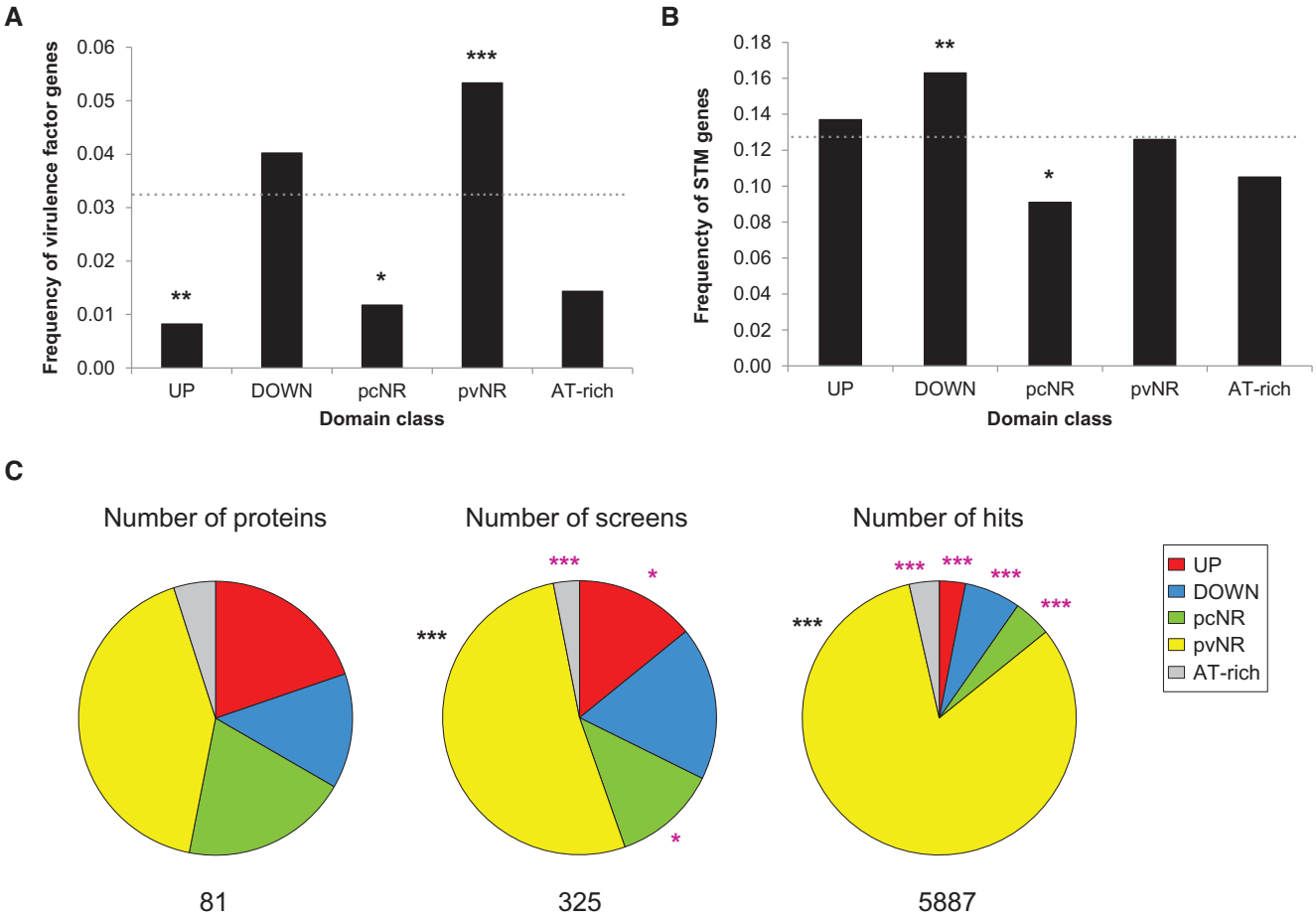

Fig. 7.—Relationship between topological domains and pathogenesis, infection and immunogenicity. (*A*) Virulence factors. (*B*) Essential genes for infection by STM. (*C*) ANTIGENome hits. The dashed lines indicate the genome average. Statistical significance: $*P \leq 0.05$; $**P \leq 0.01$; $***P \leq 0.001$ Asterisks indicate significant gene increments (in black) or decrements (in purple) according to domain type respect to the remaining genome average.

5×10^{-3}, respectively); only about a quarter of the number recorded for the pvNR domains were thus involved. Second, genes contributing to infection, as determined by STM (Hensel et al. 1995), were evaluated. For this, the results of experiments on intranasal colonization, meningitis and otitis media were considered. Many DOWN genes have been deemed essential ($P = 5 \times 10^{-3}$; χ^2 test) in animal models (fig. 7*B*) according to these studies. pcNR genes were again less involved ($P = 4 \times 10^{-2}$), indicating that these genes are more important under physiological than under virulence conditions. Third, genes coding for proteins triggering an immunogenic response in humans that might be used in protein-based serotype-independent vaccination, were evaluated. The pneumococcal ANTIGENome—the whole set of proteins quantitatively inducing the production of antibodies—has been determined for the TIGR4 strain using patient antisera (Giefing et al. 2008). A total of 81 equivalent genes were found present in the R6 strain. These represent the data from 325 positive screens and a total of 5,887 ANTIGENome hits (fig. 7*C*). The pvNR domains contained the genes encoding 11 out of the 12 proteins most immunogenic (showing $\geq$109 hits) and globally accounted for 85.4% of all ANTIGENome hits. Thus, pvNR returned >6-fold the number of hits than any other domain class, even when normalized in proportion to the number of genes ($P = < 1 \times 10^{-100}$; χ^2 test).

Cellular location is another important factor in protein function (The Gene Ontology Consortium 2015). The pvNR domains contained more genes coding for extracellular proteins or proteins anchored in the cell wall ($P = 2 \times 10^{-3}$; χ^2 test), and the pcNR domains contained fewer ($P = 1 \times 10^{-2}$) (fig. 8*A*). In addition, the UP domains were enriched in multi-transmembrane protein-coding genes ($P = 7 \times 10^{-5}$) (fig. 8*B*).

Lastly, competence was inspected. When competent, pneumococcus is able to acquire exogenous DNA, including antibiotic-resistance genes, and recombine it into its

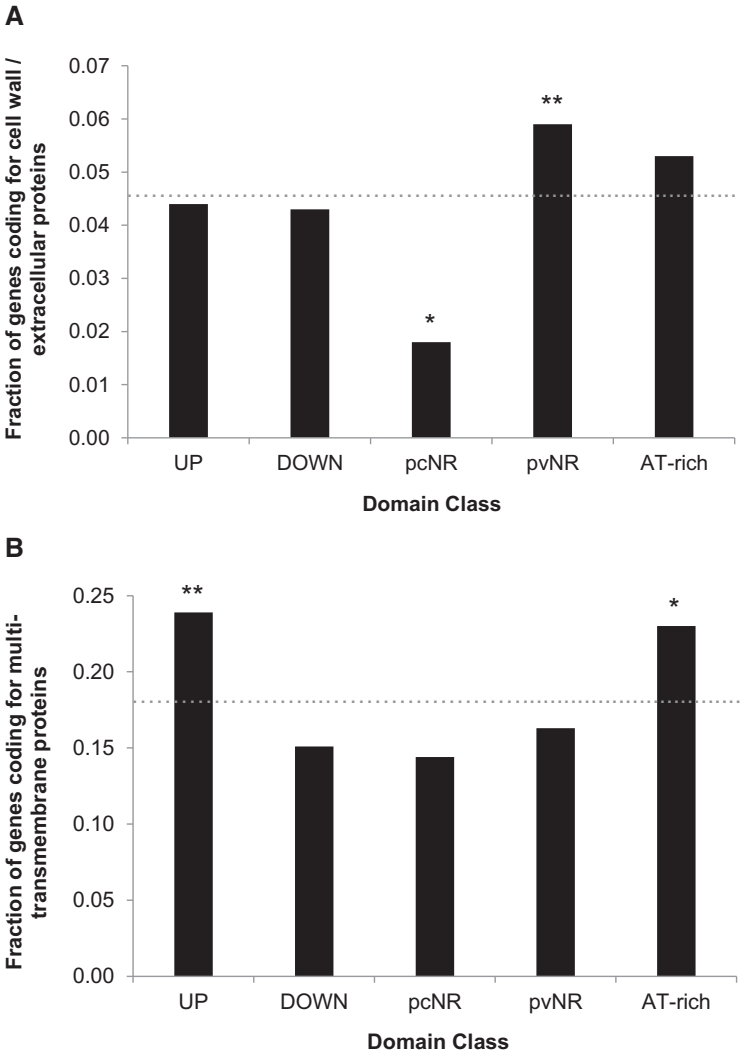

Fig. 8.—Predicted cellular location of proteins in different domains. (*A*) Additive value for cell wall and extracellular proteins according to LocateP. (*B*) LocateP prediction of transmembrane proteins with several transmembrane helices. The dashed lines indicate the genome average. Statistical significance: $*P \leq 0.05$; $**P \leq 0.01$.

chromosome. Global transcriptome time-course analysis in the presence of the competence-stimulator peptide revealed four kinds of genes to be responsible for this property: Early, late, delayed, and repressed (Peterson et al. 2004). Distinct topological classes were enriched with respect each competence stage (fig. 9). UP domain genes provided a substantial amount of early and delayed genes compared with the rest of the genome ($P = 5 \times 10^{-4}$ and 3×10^{-28}, respectively; χ^2 test) whereas DOWN genes dominated the late stage ($P = 1 \times 10^{-9}$), and pcNR genes the repressed ($P = 1 \times 10^{-11}$).

Discussion

Following the division of the pneumococcal chromosome into four types of topological domains (UP, DOWN, NR, and flanking) (Ferrándiz et al. 2010), the present results split the NR class into two new ones: pcNR and pvNR, and span the Flanking group to more peaks and called them AT-rich. The transcription of pcNR domains is unaltered by DNA relaxation and show low tolerance to changes in chromosomal location.

Gene expression is particularly high in pcNR domains, favored by an efficient codon usage and scarce repeat elements, the stable secondary structures of which reduces expression levels in nearby genes. Another important feature is that pcNR proteins have more PPIs, playing key roles in the interactome. It would be expected that evolutionary constraints forced the tropism of pcNR genes for topologically secure areas to maintain the constant provision of these central proteins, even under challenging conditions of supercoiling. It has been

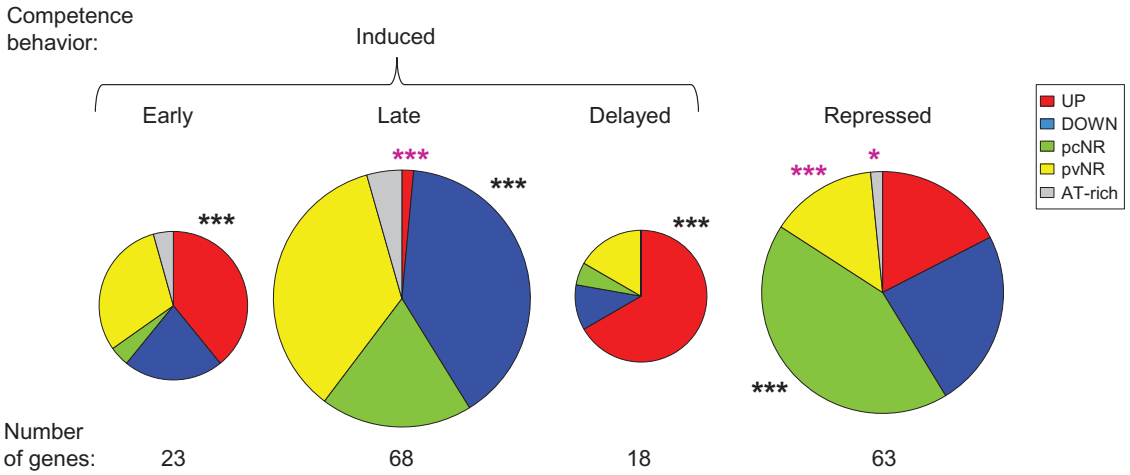

Fig. 9.—Relationship between topological domains and competence behavior. The panel covers the four kinds of competence responsive genes. The circle area is proportional to the number of genes involved. Statistical significance: * $P \leq 0.05$; ** $P \leq 0.01$; *** $P \leq 0.001$. Asterisks indicate significant gene increments (in black) or decrements (in purple) according to domain type respect to the remaining genome average.

demonstrated that changes in the location of genes can lead to alterations in the cellular physiology (Gerganova et al. 2015), which in the case of central metabolic genes, would be devastating for the cell.

Because *S. pneumoniae* is a major human pathogen, its pathobiology was re-evaluated through the lens of topogenomics. *S. pneumoniae* carries specific virulence factors. Most of these, together with the outer proteins, are encoded in pvNR zones. These areas are also enriched in immunogenic determinants representing the top 12 proteins triggering an antibody response in patients. Moreover, the pvNR domains harbor more paralogs, with double to many times more the number of genes beginning to show new functional settings, compared with the genome average. As revealed by STM, DOWN domains are enriched in genes encoding determinants for surveillance under pathogenic conditions. In contrast, pcNR zones show diminished numbers of virulence factors, STM genes and paralogs, and encode fewer cell wall/extracellular proteins. Together the present data indicate that, pcNR domains assure a constant expression of housekeeping genes whereas pvNR domains are related to the adaptation to infection. The specific virulence potential in *S. pneumoniae* lies in these pvNR, and not in the AT-rich domains despite its high HGT content. AT-rich domains may therefore constitute a source of structural or parasitic DNA.

Several intricate intersections between decreased supercoiling and competence were also found. Competence involves transient transcription changes of ~10% genome keeping to an orchestrated timing (Peterson et al. 2004). Beyond the acquisition of exogenous DNA, competence is related to a stress-resistance condition defined as the X-state (Claverys et al. 2006). Early and delayed competence genes are mainly located in UP domains, compatible with a stress situation. DOWN gene overexpression appears later. In contrast, repression affects many pcNR genes, indicating that, during the X-state, the chromosomal topology is perturbed at a potentially threatening level. This explains why growth is slowed during the narrow window of competence (Oggioni et al. 2004) and why several mechanisms have been acquired, including small untranslated RNAs and proteases, to actively terminate the X-state and promptly recover the initial topological situation (Echenique et al. 2000; Cassone et al. 2012).

Taking all these data together, a global topology theory can be envisaged in which gene positioning is far from random. Genes positively regulated by relaxation (UP genes) would be those whose expression is favored by topological stress. DOWN genes seem to be more "well-being" genes, highly expressed under favorable conditions but less so during times of topological stress. AT-rich domains accommodate large amounts of little-expressed foreign DNA with atypically high AT content, and which may sense a topological stress and modify supercoiling in the area to reduce transcription of adjacent genes, preferentially those in DOWN domains. The chromosome supercoiling structure may act as a multisensor with homeostatic capacity, adapted to react to a myriad of unfavorable conditions. Housekeeping genes show tropism for "topologically-secure" pcNR areas and thus ensure their constant strong expression.

The pcNR domains are located symmetrically around 200, 400, and 800 ORFs on both sides of the replication origin. While UP, DOWN and pcNR domains are arranged regularly over the chromosome, no precise pattern is clear at present. Difficulties in finding such a pattern may arise from the existence of longitudinal (distance to replication origin), circular (the existence of macrodomains) and adaptive forces operating to yield the definite topological landscape. In this light, the

first UP domain and the fourth DOWN domains are associated to the Ori and Ter areas, probably adapted to DNA replication. However, the remaining domains appear dispersed interleaved between others, following a cyclical arrangement likely related to transcription and, probably, the functioning of the interactome. The reduced dissemination of the transcripts, which dictates a protein concentration gradient into the cell cytoplasm, suggests a certain chromosome-centric organization of the cytoplasm's protein content (Montero et al. 2010). Topology domains may then facilitate the interaction of their products, which in the case of the pcNR genes would constitute the central core of the network. In such a scenario, chromosome topology would be connected to the structure of the PPI network.

In conclusion, topological genomics—topogenomics—constitutes an alternative paradigm of genome analysis. The genome architecture plays an important role in the pathobiology and evolution of the primary human pathogen *S. pneumoniae*, and probably of other pathogens.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Alex Mira for critical reading of the manuscript. A.J.M.G. is the beholder of a Miguel Servet contract from the Spanish Ministry of Health. This work was supported by Ministry of Economy and Competitiveness [BIO2014-55462-R].

Literature Cited

Balas D, Fernández-Moreira E, de la Campa AG. 1998. Molecular characterization of the gene encoding the DNA gyrase A subunit of *Streptococcus pneumoniae*. J Bacteriol. 180:2854–2861.

Cassone M, Gagne AL, Spruce LA, Seeholzer SH, Sebert ME. 2012. The HtrA protease from *Streptococcus pneumoniae* digests both denatured proteins and the competence-stimulating peptide. J Biol Chem. 287:38449–38459.

Champoux JJ. 2001. DNA topoisomerases: structure, function, and mechanism. Annu Rev Biochem. 70:369–413.

Chen L, Xiong Z, Sun L, Yang J, Jin Q. 2012. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Res. 40:D641–D645.

Chen H, et al. 2008. Genetic requirement for pneumococcal ear infection. PLoS One 3:e2950.

Claverys JP, Prudhomme M, Martin B. 2006. Induction of competence regulons as a general response to stress in gram-positive bacteria. Annu Rev Microbiol. 60:451–475.

Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. 79:7696–7701.

Donati C, et al. 2010. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol. 11:R107.

Dorman CJ. 2013. Genome architecture and global gene regulation in bacteria: making progress towards a unified model?. Nat Rev Microbiol. 11:349–355.

Echenique JR, Chapuy-Regaud S, Trombe MC. 2000. Competence regulation by oxygen in *Streptococcus pneumoniae*: involvement of *ciaRH* and *comCDE*. Mol Microbiol. 36:688–696.

Ferrándiz MJ, Martín-Galiano AJ, Schvartzman JB, de la Campa AG. 2010. The genome of *Streptococcus pneumoniae* is organized in topology-reacting gene clusters. Nucleic Acids Res. 38:3570–3581.

Ferrándiz MJ, Arnanz C, Martín-Galiano AJ, Rodríguez-Martín C, de la Campa AG. 2014. Role of global and local topology in the regulation of gene expression in *Streptococcus pneumoniae*. PLoS One 9:e101574.

Ferrándiz MJ, et al. 2016. An increase in negative supercoiling in bacteria reveals topology-reacting gene clusters and a homeostatic response mediated by the DNA topoisomerase I gene. Nucleic Acids Res. 44:7292–7303.

García-Vallve S, Romeu A, Palau J. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. 10:1719–1725.

Gellert M, Mizuuchi K, O'Dea MH, Nash HA. 1976. DNA gyrase: an enzyme that introduces superhelical turns into DNA. Proc Natl Acad Sci U S A. 73:3872–3876.

Gerganova V, et al. 2015. Chromosomal position shift of a regulatory gene alters the bacterial phenotype. Nucleic Acids Res. 43:8215–8226.

Giefing C, et al. 2008. Discovery of a novel class of highly conserved vaccine antigens using genomic scale antigenic fingerprinting of pneumococcus with human antibodies. J Exp Med. 205:117–131.

Hensel M, et al. 1995. Simultaneous identification of bacterial virulence genes by negative selection. Science 269:400–403.

Higgins NP, Yang X, Fu Q, Roth JR. 1996. Surveying a supercoil domain by using the gamma delta resolution system in *Salmonella typhimurium*. J Bacteriol. 178:2825–2835.

Holmes VF, Cozzarelli NR. 2000. Closing the ring: links between SMC proteins and chromosome partitioning, condensation, and supercoiling. Proc Natl Acad Sci U S A. 97:1322–1324.

Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. Nature 411:41–42.

Jiang X, Sobetzko P, Nasser W, Reverchon S, Muskhelishvili G. 2015. Chromosomal "stress-response" domains govern the spatiotemporal expression of the bacterial virulence program. MBio 6:e00353–e00315.

Johnsborg O, Eldholm V, Havarstein LS. 2007. Natural genetic transformation: prevalence, mechanisms and function. Res Microbiol. 158:767–778.

Lefebure T, Stanhope MJ. 2007. Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. Genome Biol. 8:R71.

Martín-Galiano AJ, Wells JM, de la Campa AG. 2004. Relationship between codon biased genes, microarray expression values and physiological characteristics of *Streptococcus pneumoniae*. Microbiology 150:2313–2325.

Molzen TE, et al. 2011. Genome-wide identification of *Streptococcus pneumoniae* genes essential for bacterial replication during experimental meningitis. Infect Immun. 79:288–297.

Montero LP, et al. 2010. Spatial organization of the flow of genetic information in bacteria. Nature 466:77–81.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628.

NCBI Resource Coordinators. 2016. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 44:D7–19.

Oggioni MR, et al. 2004. Antibacterial activity of a competence-stimulating peptide in experimental sepsis caused by *Streptococcus pneumoniae*. Antimicrob Agents Chemother. 48:4725–4732.

Peterson SN, et al. 2004. Identification of competence pheromone responsive genes in *Streptococcus pneumoniae* by use of DNA microarrays. Mol Microbiol. 51:1051–1070.
Postow L, Hardy CD, Arsuaga J, Cozzarelli NR. 2004. Topological domain structure of the *Escherichia coli* chromosome. Genes Dev. 18:1766–1779.
Price MN, Huang KH, Alm EJ, Arkin AP. 2005. A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res. 33:880–892.
Sharp PM, Li WH. 1987. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.
Sinden RR, Pettijohn DE. 1981. Chromosomes in living *Escherichia coli* cells are segregated into domains of supercoiling. Proc Natl Acad Sci U S A. 78:224–228.
Sobetzko P, Travers A, Muskhelishvili G. 2012. Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. Proc Natl Acad Sci U S A. 109:E42–E50.
SzklarczykD, et al. 2015. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43:D447–D452.
The Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. Nucleic Acids Res. 43:D1049–D1056.
van Opijnen T, Camilli A. 2012. A fine scale phenotype-genotype virulence map of a bacterial pathogen. Genome Res. 22:2541–2551.
Croucher NJ, Vernikos GS, Parkhill J, Bentley SD. 2011. Identification, variation and transcription of pneumococcal repeat sequences. BMC Genomics 12:120.
Wang H, Liu B, Wang Q, Wang L. 2013. Genome-wide analysis of the salmonella Fis regulon and its regulatory mechanism on pathogenicity islands. PLoS One 8:e64688.
Woelfle MA, Xu Y, Qin X, Johnson CH. 2007. Circadian rhythms of superhelical status of DNA in cyanobacteria. Proc Natl Acad Sci U S A. 104:18819–18824.
Worcel A, Burgi E. 1972. On the structure of the folded chromosome of *Escherichia coli*. J Mol Biol. 71:127–147.
Wright MA, Kharchenko P, Church GM, Segre D. 2007. Chromosomal periodicity of evolutionarily conserved gene pairs. Proc Natl Acad Sci U S A. 104:10559–10564.

Associate editor: Esther Angert