Deep genome annotation of the opportunistic human pathogen *Streptococcus pneumoniae* D39

Jelle Slager[1,#], Rieza Aprianto[1,#], Jan-Willem Veening[2,*]

[1] Molecular Genetics Group, Groningen Biomolecular Sciences and Biotechnology Institute, Centre for Synthetic Biology, University of Groningen, Nijenborgh 7, 9747 AG Groningen, the Netherlands
[2] Department of Fundamental Microbiology, Faculty of Biology and Medicine, University of Lausanne, Biophore Building, CH-1015 Lausanne, Switzerland

ORCID:
J. Slager (orcid.org/0000-0002-8226-4303),
R. Aprianto (orcid.org/0000-0003-2479-7352),
J.-W. Veening (orcid.org/0000-0002-3162-6634)

* To whom correspondence should be addressed. Tel: +41 (0)21 6925625; Email: Jan-Willem.Veening@unil.ch; Twitter: @JWVeening
# These authors should be regarded as joint First Authors

ABSTRACT

A precise understanding of the genomic organization into transcriptional units and their regulation is essential for our comprehension of opportunistic human pathogens and how they cause disease. Using single-molecule real-time (PacBio) sequencing we unambiguously determined the genome sequence of *Streptococcus pneumoniae* strain D39 and revealed several inversions previously undetected by short-read sequencing. Significantly, a chromosomal inversion results in antigenic variation of PhtD, an important surface-exposed virulence factor. We generated a new genome annotation using automated tools, followed by manual curation, reflecting the current knowledge in the field. By combining sequence-driven terminator prediction, deep paired-end transcriptome sequencing and enrichment of primary transcripts by Cappable-Seq, we mapped 1,015 transcriptional start sites and 748 termination sites. Using this new genomic map, we identified several new small RNAs (sRNAs), riboswitches (including twelve previously misidentified as sRNAs), and antisense RNAs. In total, we annotated 92 new protein-encoding genes, 39 sRNAs and 165 pseudogenes, bringing the *S. pneumoniae* D39 repertoire to 2,151 genetic elements. We report operon structures and observed that 9% of operons lack a 5'-UTR. The genome data is accessible in an online resource called PneumoBrowse (https://veeninglab.com/pneumobrowse) providing one of the most complete inventories of a bacterial genome to date. PneumoBrowse will accelerate pneumococcal research and the development of new prevention and treatment strategies.

INTRODUCTION

Ceaseless technological advances have revolutionized our capability to determine genome sequences as well as our ability to identify and annotate functional elements, including transcriptional units on these genomes. Several resources have been developed to organize current knowledge on the important opportunistic human pathogen *Streptococcus pneumoniae*, or the pneumococcus (1–3). However, an accurate genome map with an up-to-date and extensively curated genome annotation, is missing.

The enormous increase of genomic data on various servers, such as NCBI and EBI, and the associated decrease in consistency has, in recent years, led to the Prokaryotic RefSeq Genome Re-annotation Project. Every bacterial genome present in the NCBI database was re-annotated using the so-called Prokaryotic Genome Annotation Pipeline (PGAP, (4)), with the goal of increasing the quality and consistency of the many available annotations. This Herculean effort indeed created a more consistent set of annotations that facilitates the propagation and interpolation of scientific findings in individual bacteria to general phenomena, valid in larger groups of organisms. On the other hand, a wealth of information is already available for well-studied bacteria like the pneumococcus. Therefore, a separate, manually curated annotation is essential to maintain oversight of the current knowledge in the field. Hence, we generated a resource for the pneumococcal research community that contains the most up-to-date information on the D39 genome, including its DNA sequence, transcript boundaries, operon structures and functional annotation. Notably, strain D39 is one of the workhorses in research on pneumococcal biology and pathogenesis. We analyzed the genome in detail, using a

combination of several different sequencing techniques and a novel, generally applicable analysis pipeline (**Figure 1**).

Using Single Molecule Real-Time (SMRT, PacBio RS II) sequencing, we sequenced the genome of the stock of serotype 2 *S. pneumoniae* strain D39 in the Veening laboratory, hereafter referred to as strain D39V. This strain is a far descendant of the original Avery strain that was used to demonstrate that DNA is the carrier of hereditary information ((5), **Supplementary Figure S1**). Combining Cappable-seq (6), a novel sRNA detection method and several bioinformatic annotation tools, we deeply annotated the pneumococcal genome and transcriptome.

Finally, we created PneumoBrowse, an intuitive and accessible genome browser (https://veeninglab.com/pneumobrowse), based on JBrowse (7). PneumoBrowse provides a graphical and user-friendly interface to explore the genomic and transcriptomic landscape of *S. pneumoniae* D39V and allows direct linking to gene expression and co-expression data in PneumoExpress (BioRxiv: https://doi.org/10.1101/283739). The reported

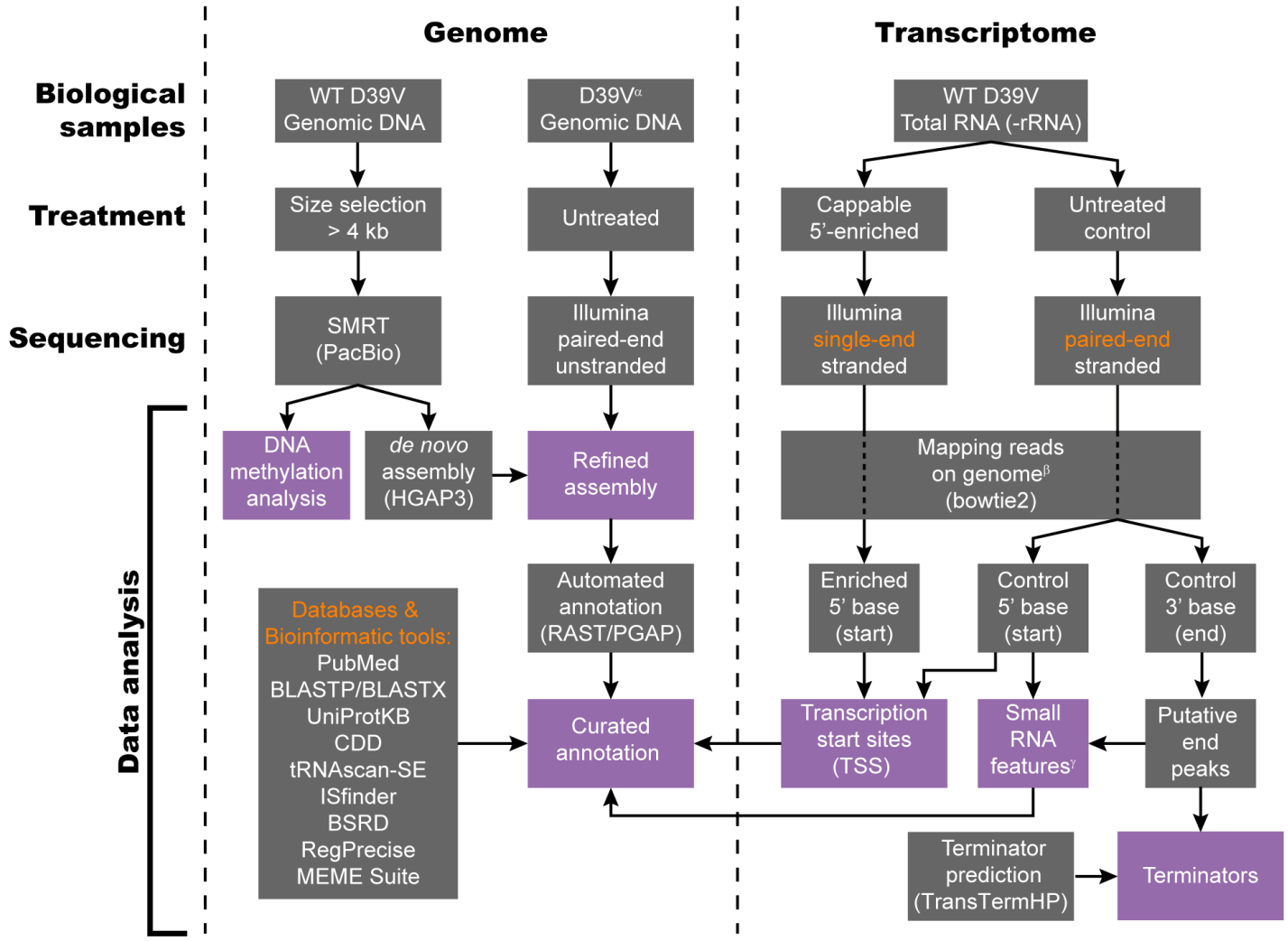

Figure 1. Data analysis pipeline used for genome assembly and annotation. Left. DNA level: the genome sequence of D39V was determined by SMRT sequencing, supported by previously published Illumina data (8, 23). Automated annotation by the RAST (11) and PGAP (4) annotation pipelines was followed by curation based on information from literature and a variety of databases and bioinformatic tools. Right. RNA level: Cappable-seq (6) was utilized to identify transcription start sites. Simultaneously, putative transcript ends were identified by combining reverse reads from paired-end, stranded sequencing of the control sample (i.e. not 5'-enriched). Terminators were annotated when such putative transcript ends overlapped with stem loops predicted by TransTermHP (20). Finally, local fragment size enrichment in the paired-end sequencing data was used to identify putative small RNA features. [α]D39 derivative (*bgaA*::P_{ssbB}-*luc*; GEO accessions GSE54199 and GSE69729). [β]The first 1 kbp of the genome file was duplicated at the end, to allow mapping over FASTA boundaries. [γ]Analysis was performed with only sequencing pairs that map uniquely to the genome.

annotation pipeline and accompanying genome browser provide one of the best curated bacterial genomes currently available and may facilitate rapid and accurate annotation of other bacterial genomes. We anticipate that PneumoBrowse will significantly accelerate the pneumococcal research field and hence speed-up the discovery of new drug targets and vaccine candidates for this devastating global opportunistic human pathogen.

MATERIALS AND METHODS

Culturing of *S. pneumoniae* D39 and strain construction

S. pneumoniae was routinely cultured without antibiotics. Transformation, strain construction and preparation of growth media are described in detail in the **Supplementary Methods**. Bacterial strains are listed in **Supplementary Table S1** and oligonucleotides in **Supplementary Table S2**.

Growth, luciferase and GFP assays

Cells were routinely pre-cultured in C+Y medium until an OD_{600} of 0.4, and then diluted 1:100 into fresh medium in a 96-wells plate. All assays were performed in a Tecan Infinite 200 PRO at 37°C. Luciferase assays were performed in C+Y with 0.25 mg·ml^{-1} D-luciferin, sodium salt and signals were normalized by OD_{595}. Fluorescence signals were normalized using data from a parental *gfp*-free strain. Growth assays of *lacD*-repaired strains were performed in C+Y with either 10.1 mM galactose or 10.1 mM glucose as main carbon source.

DNA and RNA isolation, primary transcript enrichment and sequencing

S. pneumoniae chromosomal DNA was isolated as described previously (8). A 6/8 kbp insert library for SMRT sequencing, with a lower cut-off of 4kbp, was created by the Functional Genomics Center Zurich (FGCZ) and was then sequenced using a PacBio RS II machine.

D39V samples for RNA-seq were pre-cultured in suitable medium before inoculation (1:100) into four infection-relevant conditions, mimicking (i) lung, (ii) cerebral spinal fluid (CSF), (iii) fever in CSF-like medium, and (iv) late competence (20 min after CSP addition) in C+Y medium. Composition of media, a detailed description of conditions and the total RNA isolation protocol are described in the accompanying paper by Aprianto et al. (BioRxiv: https://doi.org/10.1101/283739). Isolated RNA was sent to vertis Biotechnologie AG for sequencing. Total RNA from the four conditions was combined in an equimolar fashion and the pooled RNA was divided into two portions. The first portion was directly enriched for primary transcripts (Cappable-seq, (6)) and, after stranded library preparation according to the manufacturer's recommendations, sequenced on Illumina NextSeq in single-end (SE) mode. The second RNA portion was rRNA-depleted and sequenced on Illumina NextSeq in paired-end (PE) mode. RNA-seq data was mapped to the newly assembled genome using Bowtie 2 (9).

De novo assembly of the D39V genome and DNA modification analysis

Analysis of SMRT sequencing data was performed with the SMRT tools analysis package (DNA Link, Inc., Seoul, Korea). *De novo* genome assembly was performed using the Hierarchical Genome Assembly Process (HGAP3) module of the PacBio SMRT portal version 2.3.0. This resulted in two contigs: one of over 2 Mbp with 250-500x coverage, and one of 12 kbp with 5-25x coverage. The latter, small contig was discarded based on its low coverage and high sequence similarity with a highly repetitive segment of the larger contig. The large contig was circularized manually and then rotated such that *dnaA* was positioned on the positive strand, starting on the first nucleotide. Previously published Illumina data of our D39 strain (GEO accessions GSE54199 and GSE69729) were mapped on the new assembly, using *breseq* (10), to identify potential discrepancies. Identified loci of potential mistakes in the assembly were verified by Sanger sequencing, leading to the correction of a single mistake. DNA modification analysis was performed using the 'RS_Modification_and_Motif_Analysis.1' module in the SMRT portal, with a QV cut-off of 100. A PCR-based assay to determine the absence or presence of plasmid pDP1 is described in **Supplementary Methods**.

Automated and curated annotation

The assembled genome sequence was annotated automatically, using PGAP ((4), executed October 2015) and RAST ((11), executed June 2016). The results of both annotations were compared and for each discrepancy, support was searched. Among the support used were scientific publications (PubMed), highly similar features (BLAST, (12)), reviewed UniProtKB entries (13) and detected conserved domains as found by CD-Search (14). When no support was found for either the PGAP or RAST annotation, the latter was used. Similarly, annotations of conserved features in strain R6 (NC_003098.1) were adopted when sufficient evidence was available. Finally, an extensive literature search was performed, with locus tags and (if available) gene names from the old D39 annotation (prefix: 'SPD_') as query. When identical features were present

in R6, a similar search was performed with R6 locus tags (prefix 'spr') and gene names. Using the resulting literature, the annotation was further refined. Duplicate gene names were also resolved during curation.

CDS pseudogenes were detected by performing a BLASTX search against the NCBI non-redundant protein database, using the DNA sequence of two neighboring genes and their intergenic region as query. If the full-length protein was found, the two (or more, after another BLASTX iteration) genes were merged into one pseudogene.

Furthermore, sRNAs and riboswitches, transcriptional start sites and terminators, transcription-regulatory sequences and other useful features (all described below) were added to the annotation. Finally, detected transcript borders (TSSs and terminators) were used to refine coordinates of annotated features (e.g. alternative translational initiation sites). Afterwards, the quality of genome-wide translational initiation site (TIS) calls was evaluated using 'assess_TIS_annotation.py' (15).

All publications used in the curation process are listed in the **Supplementary Data**.

Conveniently, RAST identified pneumococcus-specific repeat regions: BOX elements (16), Repeat Units of the Pneumococcus (RUPs, (17)) and *Streptococcus pneumoniae* Rho-Independent Termination Elements (SPRITEs, (18)), which we included in our D39V annotation. Additionally, ISfinder (19) was used to locate Insertion Sequences (IS elements).

Normalized start and end counts and complete coverage of sequenced fragments

The start and (for paired-end data) end positions of sequenced fragments were extracted from the sequence alignment map (SAM) produced by Bowtie 2. The positions were used to build strand-specific, single-nucleotide resolution frequency tables (start counts, end counts and coverage). For paired-end data, coverage was calculated from the entire inferred fragment (i.e. including the region between mapping sites of mate reads). Start counts, end counts and coverage were each normalized by division by the summed genome-wide coverage, excluding positions within 30 nts of rRNA genes.

Identification of transcriptional terminators

Putative Rho-independent terminator structures were predicted with TransTermHP (20), with a minimum confidence level of 60. Calling of 'putative coverage termination peaks' in the paired-end sequencing data of the control library is described in detail in **Supplementary Methods**. When such a peak was found to overlap with the 3'-poly(U)-tract of a predicted terminator, the combination of both elements was annotated as a high-confidence (HC) terminator. Terminator efficiency was determined by the total number of fragments ending in a coverage termination peak, as a percentage of all fragments covering the peak (i.e. including non-terminated fragments).

Detection of small RNA features

For each putative coverage termination peak (see above), fragments from the SAM file that ended inside the peak region were extracted. Of those reads, a peak-specific fragment size distribution was built and compared to the library-wide fragment size distributions (see **Supplementary Methods**). A putative sRNA was defined by several criteria: (i) the termination efficiency of the coverage termination peak should be above 30% (see above for the definition of termination efficiency), (ii) the relative abundance of the predicted sRNA length should be more than 25-fold higher than the corresponding abundance in the library-wide distribution, (iii) the predicted sRNA should be completely covered at least 15x for HC terminators and at least 200x for non-HC terminators. The entire process was repeated once more, now also excluding all detected putative sRNAs from the library-wide size distribution. For the scope of this paper, only predicted sRNAs that did not significantly overlap already annotated features were considered.

A candidate sRNA was annotated (either as sRNA or riboswitch) when either (i) a matching entry, with a specified function, was found in RFAM (21) and/or BSRD (22) databases; (ii) the sRNA was validated by Northern blotting, either in previous studies or by us (**Supplementary Figure S5**, **Supplementary Methods**); or (iii) at least two transcription-regulatory elements were detected (i.e. transcriptional start or termination sites, or sigma factor binding sites).

Transcription start site identification

Normalized start counts from 5'-enriched and control libraries were compared. Importantly, normalization was performed excluding reads that mapped within 30 bps of rRNA genes. An initial list was built of unclustered TSSs, which have (i) at least 2.5-fold higher normalized start counts in the 5'-enriched library, compared to the control library, and (ii) a minimum normalized start count of 2 (corresponding to 29 reads) in the 5'-enriched library. Subsequently, TSS candidates closer than 10 nucleotides were clustered, conserving the candidate with the highest start count in

the 5'-enriched library. Finally, if the 5'-enriched start count of a candidate TSS was exceeded by the value at the nucleotide immediately upstream, the latter was annotated as TSS instead. The remaining, clustered TSSs are referred to as high-confidence (HC) TSSs. To account for rapid dephosphorylation of transcripts, we included a set of 34 lower confidence (LC) TSSs in our annotation, which were not overrepresented in the 5'-enriched library, but that did meet a set of strict criteria: (i) normalized start count in the control library was above 10 (corresponding to 222 reads), (ii) a TATAAT motif (with a maximum of 1 mismatch) was present in the 5-15 nucleotides upstream, (iii) the nucleotide was not immediately downstream of a processed tRNA, and (iv) the nucleotide was in an intergenic region. If multiple LC-TSSs were predicted in one intergenic region, only the strongest one was annotated. If a HC-TSS was present in the same intergenic region, the LC-TSS was only annotated when its 5'-enriched start count exceeded that of the HC-TSS. TSS classification and prediction of regulatory motifs is described in **Supplementary Methods**.

Operon prediction and leaderless transcripts

Defining an operon as a set of genes controlled by a single promoter, putative operons were predicted for each primary TSS. Two consecutive features on the same strand were predicted to be in the same operon if (i) their expression across 22 infection-relevant conditions was strongly correlated (correlation value > 0.75 (BioRxiv: https://doi.org/10.1101/283739)) and (ii) no strong terminator (>80% efficient) was found between the features. In a total of 69 leaderless transcripts, the TSS was found to overlap with the translation initiation site of the first encoded feature in the operon.

PneumoBrowse

PneumoBrowse (https://veeninglab.com/pneumobrowse) is based on JBrowse (7), supplemented with plugins *jbrowse-dark-theme* and *SitewideNotices* (https://github.com/erasche), and *ScreenShotPlugin*, *HierarchicalCheckboxPlugin* and *StrandedPlotPlugin* (BioRxiv: https://doi.org/10.1101/212654). Annotated elements were divided over five annotation tracks: (i) genes (includes pseudogenes, shown in grey), (ii) putative operons, (iii) regulatory features, including TSSs and terminators, (iv) repeats, and (v) other features. Additionally, full coverage tracks are available, along with start and end counts.

RESULTS

De novo assembly yields a single circular chromosome

We performed *de novo* genome assembly using SMRT sequencing data, followed by polishing with high-confidence Illumina reads, obtained in previous studies (8, 23). Since this data was derived from a derivative of D39, regions of potential discrepancy were investigated using Sanger sequencing. In the end, we needed to correct the SMRT assembly in only one location. The described approach yielded a single chromosomal sequence of 2,046,572 base pairs, which was deposited to GenBank (accession number CP027540).

D39V did not suffer disruptive mutations compared to ancestral strain NCTC 7466

We then compared the newly assembled genome with the previously established sequence of D39 (D39W, (24)), and observed similar sequences, but with some striking differences (**Table 1**, **Figure 2A**). Furthermore, we cross-checked both sequences with the genome sequence of the ancestral strain NCTC 7466 (ENA accession number ERS1022033), which was recently sequenced with SMRT technology, as part of the NCTC 3000 initiative. Interestingly, D39V matches NCTC 7466 in all gene-disruptive discrepancies (e.g. frameshifts and a chromosomal inversion, see below). Most of these sites are characterized by their repetitive nature (e.g. homopolymeric runs or long repeated sequences). Considering the sequencing technology used, these differences are likely to be the result of misassembly in D39W, rather than sites of true biological divergence. On the other hand, discrepancies between D39V and the ancestral strain are limited to SNPs, with unknown consequences for pneumococcal fitness. It seems plausible that these polymorphisms constitute actual mutations in D39V, emphasizing the dynamic nature of the pneumococcal genome. Notably, there are two sites where both the D39W and D39V assemblies differ from the ancestral strain. Firstly, the ancestral strain harbors a mutation in *rrlC* (SPV_1814), one of four copies of the gene encoding 23S ribosomal RNA. It is not clear if this is a technical artefact in one of the assemblies (due to the large repeat size in this region), or an actual biological difference. Secondly, we observed a mutation in the upstream region of *cbpM* (SPV_1248) in both D39W and D39V.

D39W coordinate(s)	D39V coordinate(s)	Locus	Change	Consequence	Note
80663-83343	**80663-83799**	SPD_0080 (*pavB*)	Repeat expansion (6x>7x)	Repeat expansion in PavB (6x>7x)	
174318	174774	SPD_0170 (*ruvA*)	G>A	RuvA V52V (GTG>GTA)	
297022	**297479**	SPD_0299-300	+T	SPD_0299 and SPD_0300 shifted into same coding frame	
303240	**303697**	SPD_0306 (*pbp2x*)	A>G	PBP2X N311D (AAT>GAT)	
458088-462242	458545-462699	SPD_0450-55 (*hsdRMS, creX*)	Multiple rearrangements	HsdS type A>F	(Manso et al. 2014)
462212	458575	SPD_0453 (*hsdS*)	A>G	Imperfect > perfect inverted repeat	Inside rearranged region
675950	676407	SPD_0657 → / → SPD_0658 (*prfB*)	C>A	Intergenic (+163/-51 nt)	In 5' UTR of *prfB*
775672	776129	SPD_0764 (*sufS*)	G>A	SufS G318R (GGA>AGA)	
816157	**816615**	SPD_0800[β]	+G	Frameshift (347/360 nt)	
901217-1062944	**901675-1063403**	SPD_0889-1037	Inversion	Swap of 3' ends of *phtB* (SPD_1037) and *phtD* (SPD_0889)	
934443	1030177	SPD_0921 (*ccrB*)	A>G[α]	CcrB Q286R (CAG>CGG)	
951536	**1013083**[γ]	SPD_0942	+C[α]	Frameshift (198/783 nt)	
1035166	929453	SPD_1016 (*rexA*)	C>A[α]	RexA A961D (GCT>GAT)	
1080119	**1080577**	SPD_1050 (*lacD*)	ΔT	Frameshift (159/981 nt)	
1171761	**1172219**	SPD_1137	C>G	H431Q (CAC>CAG)	
1256812	**1257270**	SPD_1224 (*budA*) ← / → SPD_1225	ΔA	Intergenic (-100/-42 nt)	
1256937	1257394[γ]	SPD_1225	G>T	R28L (CGC>CTC)	
1672084	**1672541**	SPD_1660 (*rdgB*)	G>A	RdgB T117I (ACA>ATA)	
1676516	1676973	SPD_1664 (*treP*)	C>T	TreP G359D (GGC>GAC)	
1787708	1788165	SPD_1793	C>T	A2V (GCA>GTA)	
1977728	1978185	SPD_2002 (*dltD*)	C>A	DltD V252F (GTC>TTC)	
2022372	2022829	SPD_2045 (*mreC*)	A>G	MreC S186P (TCT>CCT)	

Table 1. Differences between old and new genome assembly. The genomic sequences of the old (D39W, CP000410) and new (D39V, CP027540) genome assemblies were compared, revealing 14 SNPs, 3 insertions, and 2 deletions. Additionally, a repeat expansion in *pavB*, several rearrangements in the *hsdS* locus and, most strikingly, a 162 kbp (8% of the genome) chromosomal inversion were observed. Finally, both sequences were compared to the recently released PacBio sequence of ancestral strain NCTC 7466 (ENA accession number ERS1022033). For each observed difference between the old and new assembly, the variant matching the ancestral strain is displayed in boldface. [α]Locus falls within the inverted *ter* region and the forward strain in the new assembly is therefore the reverse complementary of the old sequence (CP000410). [β]Region is part of a larger pseudogene in the new annotation. [γ]Only found in one of two D39 stocks in our laboratory.

Several SNPs and indel mutations observed in D39V assembly

Fourteen single nucleotide polymorphisms (SNPs) were detected upon comparison of D39W and D39V assemblies. One of these SNPs results in a silent mutation in the gene encoding RuvA, the Holliday junction DNA helicase, while another SNP was located in the 5'-untranslated region (5'-UTR) of *prfB*, encoding peptide chain release factor 2. The other twelve SNPs caused amino acid changes in various proteins, including penicillin-binding protein PBP2X and cell shape-determining protein MreC. It should be noted that one of these SNPs, leading to an arginine to leucine change in the protein encoded by SPV_1225 (previously SPD_1225), was not found in an alternative D39 stock from our lab (**Supplementary Figure S1**). The same applies to an insertion of a cytosine causing a frameshift in the extreme 3'-end of SPV_0942 (previously SPD_0942; **Supplementary Figure S2L**). All other differences found, however, were identified in both of our stocks and are therefore likely to be more widespread. Among these differences are four more indel mutations (insertions or deletions), the genetic context and consequences of which are shown in **Supplementary**

Figure S2. One of the indels is located in the promoter region of two diverging operons, with unknown consequences for gene expression. Secondly, we found an insertion in the region corresponding to SPD_0800 (D39W annotation). Here, we report this gene to be part of a pseudogene together with SPD_0801 (annotated as SPV_2242). Hence, the insertion probably is of little consequence. Thirdly, a deletion was observed in the beginning of *lacD*, encoding an important enzyme in the D-tagatose-6-phosphate pathway, relevant in galactose metabolism. The consequential absence of functional LacD may explain why the inactivation of the alternative Leloir pathway in D39 significantly hampered growth on galactose (25). We repaired *lacD* in D39V and, as expected, observed restored growth on galactose (**Supplementary Figure S3**). Finally, we observed a thymine insertion that caused SPD_0299 and SPD_0300 to be shifted into the same coding frame and form a single 1.9 kb long CDS (SPV_2142). Since the insertion was found in a homopolymeric run of thymines and the assemblies of NCTC 7466 and D39V match, it seems plausible that instead of a true indel mutation, this actually reflects a sequencing error in the D39W assembly.

Varying repeat frequency in surface-exposed protein PavB

Pneumococcal adherence and virulence factor B (PavB) is encoded by SPV_0080. Our assembly shows that this gene contains a series of seven imperfect repeats of 450-456 bps

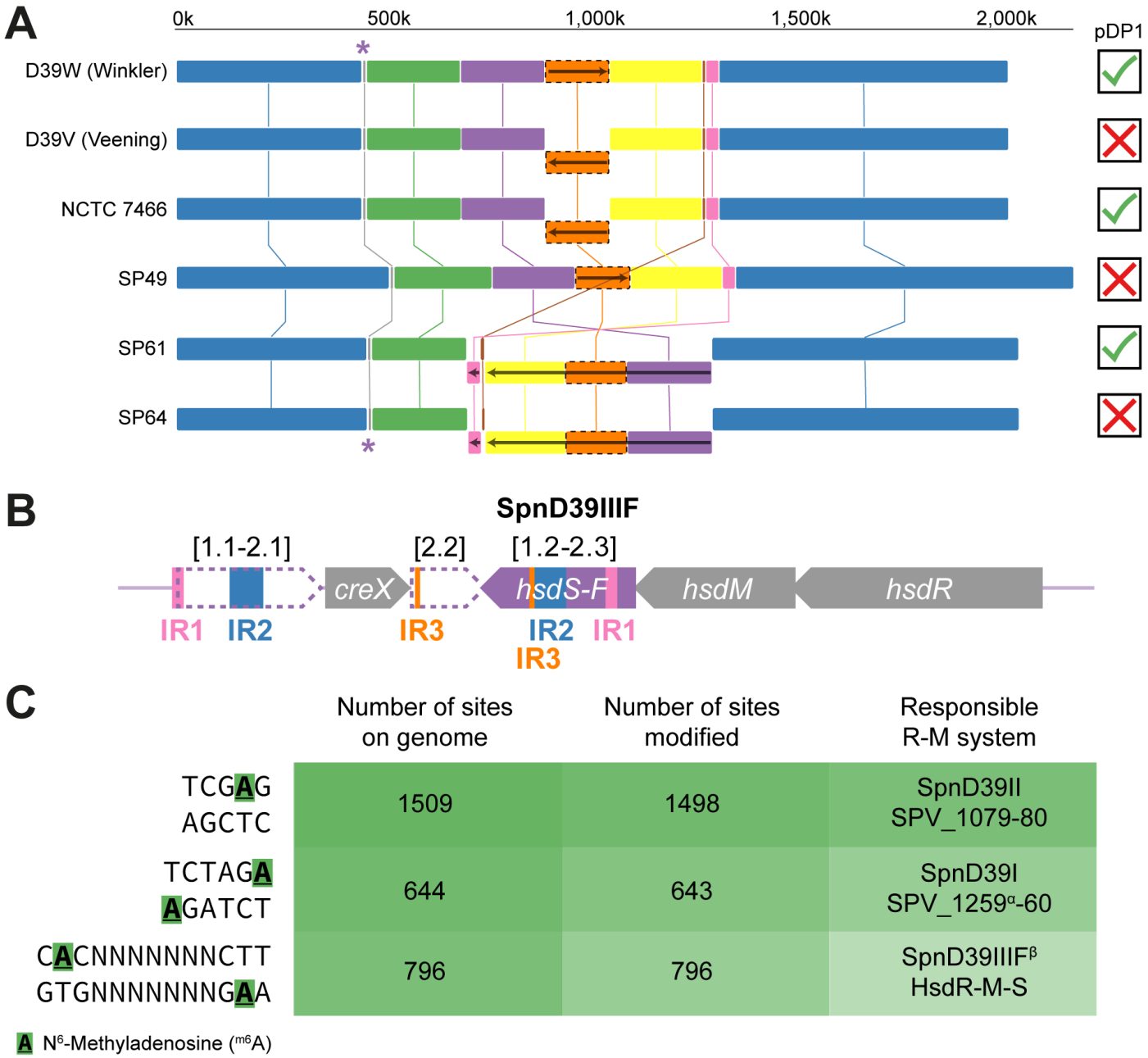

Figure 2. Multiple genome alignment. **A.** Multiple genome sequence alignment of D39W, D39V, NCTC 7466, and clinical isolates SP49, SP61, and SP64 (30) reveals multiple *ter*-symmetrical chromosomal inversions. Identical colors indicate similar sequences, while blocks shown below the main genome level and carrying a reverse arrow signify inverted sequences relative to the D39W assembly. The absence/presence of the pDP1 (or similar) plasmid is indicated with a cross/checkmark. Asterisks indicate the position of the *hsdS* locus. **B.** Genomic layout of the *hsdS* region. As reported by Manso et al. (26), the region contains three sets of inverted repeats (IR1-3), that are used by CreX to reorganize the locus. Thereby, six different variants (A-F) of methyltransferase specificity subunit HsdS can be generated, each leading to a distinct methylation motif. SMRT sequencing of D39V revealed that the locus exists predominantly in the F-configuration, consisting of N-terminal variant 2 (i.e. 1.2) and C-terminal variant 3 (i.e. 2.3). **C.** Motifs that were detected to be specifically modified in D39V SMRT data. [α]SPV_1259 (encoding the R-M system endonuclease) is a pseudogene, due to a nonsense mutation. [β]Manso et al. reported the same motifs and reported the responsible methyltransferases. The observed C**A**C-N_7-CTT motif perfectly matches the predicted putative HsdS-F motif. **Table 1.** Differences between old and new genome assembly. The genomic sequences of the old (D39W, CP000410) and new (D39V, CP027540) genome assemblies were compared, revealing 14 SNPs, 3 insertions, and 2 deletions. Additionally, a repeat expansion in *pavB*, several rearrangements in the *hsdS* locus and, most strikingly, a 162 kbp (8% of the genome) chromosomal inversion were observed. Finally, both sequences were compared to the recently released PacBio sequence of ancestral strain NCTC 7466 (ENA accession number ERS1022033). For each observed difference between the old and new assembly, the variant matching the ancestral strain is displayed in boldface. [α]Locus falls within the inverted *ter* region and the forward strain in the new assembly is therefore the reverse complementary of the old sequence (CP000410). [β]Region is part of a larger pseudogene in the new annotation. [γ]Only found in one of two D39 stocks in our laboratory.

in size. Interestingly, SPD_0080 in D39W contains only six of these repeats. If identical repeat units are indicated with an identical letter, the repeat region in SPV_0080 of D39V can be written as *ABBCBDE*, where *E* is truncated after 408 bps. Using the same letter code, SPD_0080 of D39W contains *ABBCDE*, thus lacking the third repeat of element *B*, which is isolated from the other copies in SPV_0080. Because D39V and NCTC 7466 contain the full-length version of the gene, we hypothesized that D39W lost one of the repeats, making the encoded protein 152 residues shorter.

Configuration of variable *hsdS* region matches observed methylation pattern

A local rearrangement is found in the pneumococcal *hsdS* locus, encoding a three-component restriction-modification system (HsdRMS). Recombinase CreX facilitates local recombination, using three sets of inverted repeats, and can thereby rapidly rearrange the region into six possible configurations (SpnD39IIIA-F). This process results in six different versions of methyltransferase specificity subunit HsdS, each with its own sequence specificity and transcriptomic consequences (26, 27). The region is annotated in the A-configuration in D39W, while the F-configuration is predominant in D39V (**Figure 2B**). Moreover, we employed methylation data, intrinsically present in SMRT data (28), and observed an enriched methylation motif that exactly matches the putative SpnD39IIIF motif predicted by Manso et al. (**Figure 2C**).

A large chromosomal inversion occurred multiple times in pneumococcal evolution

We also observed a striking difference between D39V and D39W: a 162 kbp region containing the replication terminus was completely inverted (**Figures 2A** and **3**), with D39V matching the configuration of the ancestral NCTC 7466. The inverted region is bordered by two inverted repeats of 1.3 kb in length. We noticed that the *xerS*/dif_{SL} site, responsible for chromosome dimer resolution and typically located directly opposite the origin of replication (29), is asymmetrically situated on the right replichore in D39V (**Figure 3A**), while the locus is much closer to the halfway point of the chromosome in the D39W assembly, suggesting that this configuration is the original one and the observed inversion in D39V and NCTC 7466 is a true genomic change, rather than merely a sequencing artefact. To confirm this, we performed a PCR-based assay, in which the two possible configurations yield different product sizes. Indeed, the results showed that two possible configurations of the region exist in different pneumococcal strains; multiple D39 stocks, TIGR4, BHN100 and PMEN-14 have matching terminus regions, while the opposite configuration was found in R6, Rx1, PMEN-2 and PMEN18. We repeated the analysis for a set of seven and a set of five strains, each related by a series of sequential transformation events. All strains had the same *ter* orientation (*not shown*), suggesting that the inversion is relatively rare, even in competent cells. However, both configurations are found in various branches

of the pneumococcal phylogenetic tree, indicating multiple incidences of this chromosomal inversion. Interestingly, a similar, even larger inversion was observed in two out of three recently-sequenced clinical isolates of *S. pneumoniae* (30) **Figure 2A**), suggesting a larger role for chromosomal inversions in pneumococcal evolution.

Antigenic variation of histidine triad protein PhtD

Surprisingly, the repeat regions bordering the chromosomal inversion are located in the middle of *phtB* and *phtD* (**Figure 3A**), leading to an exchange of the C-terminal parts of their respective products, PhtB and PhtD. These are two out of four pneumococcal histidine triad (Pht) proteins, which are surface-exposed, interact with human host cells and are considered to be good vaccine candidates (31). In fact, PhtD was already used in several phase I/II clinical trials (e.g. (32, 33)). Yun et al. analyzed the diversity of *phtD* alleles from 172 clinical isolates and concluded that the sequence variation was minimal (34). However, this conclusion was biased by the fact that inverted chromosomes would not produce a PCR product in their set-up and a swap between PhtB and PhtD would remain undetected. Moreover, after detailed inspection of the mutations in the *phtD* alleles and comparison to other genes encoding Pht proteins (*phtA, phtB* and *phtE*), we found that many of the SNPs could be explained by recombination events between these genes, rather than by random mutation. For example, extensive exchange was seen between D39V *phtA* and

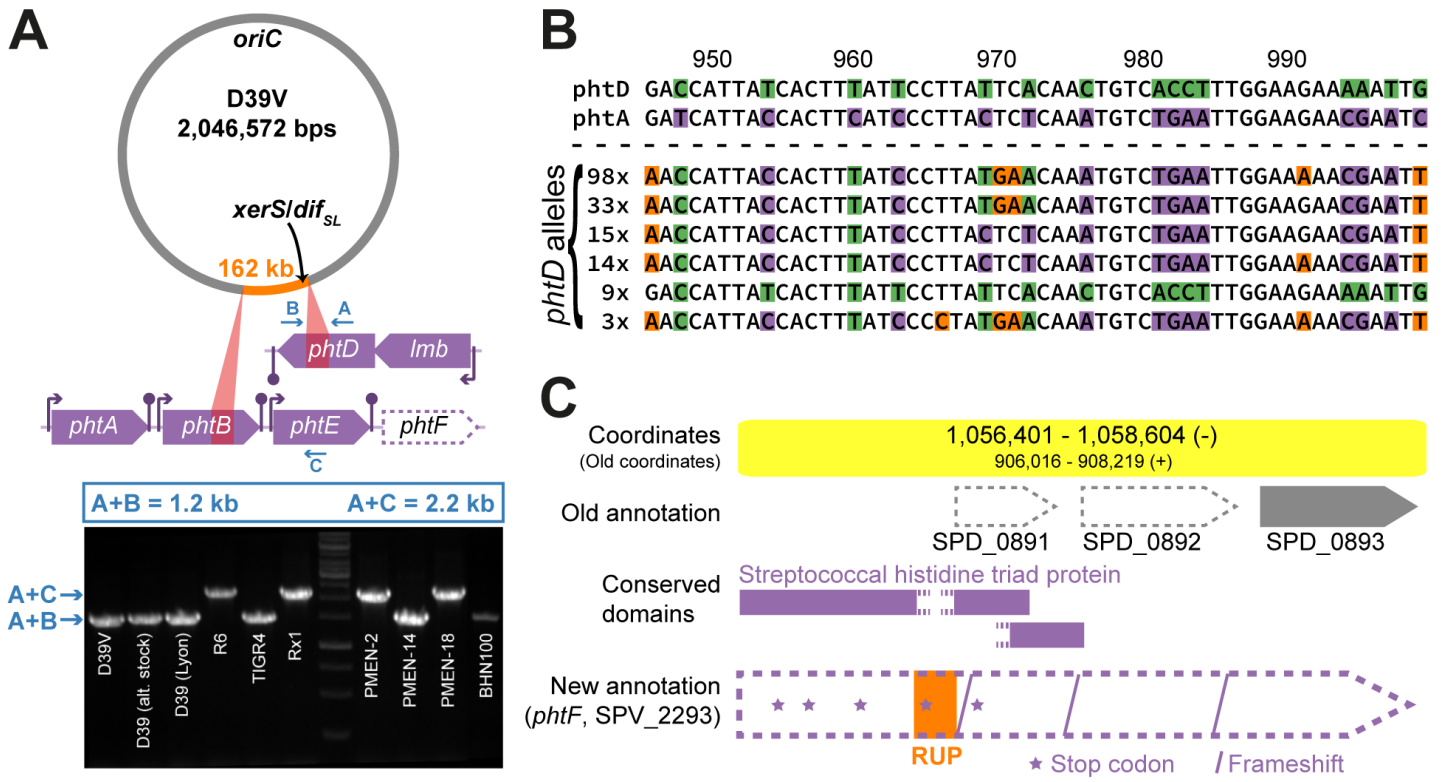

Figure 3. A large chromosomal inversion unveils antigenic variation of pneumococcal histidine triad proteins. **A.** Top: chromosomal location of the inverted 162 kb region (orange). Red triangles connect the location of the 1 kb inverted repeats bordering the inverted region and a zoom of the genetic context of the border areas, also showing that the inverted repeats are localized in the middle of genes *phtB* and *phtD*. Arrows marked with A, B and C indicate the target regions of oligonucleotides used in PCR analysis of the region. Bottom: PCR analysis of several pneumococcal strains (including both our D39 stocks and a stock from the Grangeasse lab, Lyon) shows that the inversion is a true phenomenon, rather than a technical artefact. PCR reactions are performed with all three primers present, such that the observed product size reports on the chromosomal configuration. **B.** A fragment of a Clustal Omega multiple sequence alignment of 172 reported *phtD* alleles (34) and D39V genes *phtD* and *phtA* exemplifies the dynamic nature of the genes encoding pneumococcal histidine triad proteins. Bases highlighted in green and purple match D39V *phtD* and *phtA*, respectively. Orange indicates that a base is different from both D39V genes, while white bases are identical in all sequences. **C.** Newly identified pseudogene, containing a RUP insertion and several frameshifts and nonsense mutations, that originally encoded a fifth pneumococcal histidine triad protein, and which we named *phtF*. Old (D39W) and new annotation (D39V) are shown, along with conserved domains predicted by CD-Search (14).

phtD (**Figure 3B**). Apparently, the repetitive nature of these genes allows for intragenomic recombination, causing *phtD* to become mosaic, rather than well-conserved. Finally, immediately downstream of *phtE* (**Figure 3A**), we identified a pseudogene that originally encoded a fifth histidine triad protein and which we named *phtF* (**Figure 3C**). The gene is disrupted by an inserted RUP element (see below) and several frameshifts and nonsense mutations, and therefore does not produce a functional protein. Nevertheless, *phtF* might still be relevant as a source of genetic diversity. Taken together, these findings raise caution on the use of PhtD as a vaccine target.

RNA-seq data and PCR analysis show loss of cryptic plasmid from strain D39V

Since SMRT technology is known to miss small plasmids in the assembly pipeline, we performed a PCR-based assay to check the presence of the cryptic pDP1 plasmid, reported in D39W (24, 35). To our surprise, the plasmid is absent in D39V, while clearly present in the ancestral NCTC 7466, as confirmed by a PCR-based assay (**Supplementary Figure S4**). Intriguingly, a BLASTN search suggested that *S. pneumoniae* Taiwan19F-14 (PMEN-14, CP000921), among other strains, integrated a degenerate version of the plasmid into its chromosome. Indeed, the PCR assay showed positive results for this strain. Additionally, we selected publicly available D39 RNA-seq datasets and mapped the sequencing reads specifically to the pDP1 reference sequence (Accession AF047696). The successful mapping of a significant number of reads indicated the presence of the plasmid in strains used in several studies ((36), SRX2613845; (37), SRX1725406; (26), SRX472966). In contrast, RNA-seq data of D39V (8, 23, BioRxiv: https://doi.org/10.1101/283739) contained zero reads that mapped to the plasmid, providing conclusive evidence that strain D39V lost the plasmid at some stage (**Supplementary Figure S1**). Similarly, based on Illumina DNA-seq data, we determined that of the three clinical isolates shown in **Figure 2A**, only SP61 contained a similar plasmid (30).

Automation and manual curation yield up-to-date pneumococcal functional annotation

An initial annotation of the newly assembled D39V genome was produced by combining output from the RAST annotation engine (11) and the NCBI prokaryotic genome annotation pipeline (PGAP, (4)). We, then, proceeded with exhaustive manual curation to produce the final genome annotation (see **Materials and Methods** for details). All annotated CDS features without an equivalent feature in the D39W annotation or with updated coordinates are listed in **Supplementary Table S3**. Examples of the integration of recent research into the final annotation include cell division protein MapZ (38, 39), pleiotropic RNA-binding proteins KhpA and KhpB/EloR (40, 41) and cell elongation protein CozE (42).

Additionally, we used tRNAscan-SE (43) to differentiate the four encoded tRNAs with a CAU anticodon into three categories (**Supplementary Table S4**): tRNAs used in either (i) translation initiation or (ii) elongation and (iii) the post-transcriptionally modified tRNA-Ile2, which decodes the AUA isoleucine codon (44).

Next, using BLASTX ((12), **Materials and Methods**), we identified and annotated 165 pseudogenes (**Supplementary Table S5**), two-fold more than reported previously (24). These non-functional transcriptional units may be the result of the insertion of repeat regions, nonsense and/or frameshift mutations and/or chromosomal rearrangements. Notably, 71 of 165 pseudogenes were found on IS elements (19), which are known to sometimes utilize alternative coding strategies, including programmed ribosomal slippage, producing a functional protein from an apparent pseudogene. Finally, we annotated 127 BOX elements (16), 106 RUPs (17), 29 SPRITEs (18) and 58 IS elements (19).

RNA-seq coverage and transcription start site data allow improvement of annotated feature boundaries

Besides functional annotation, we also corrected the genomic coordinates of several features. First, we updated tRNA and rRNA boundaries (**Supplementary Table S4**), aided by RNA-seq coverage plots that were built from deduced paired-end sequenced fragments, rather than from just the sequencing reads. Most strikingly, we discovered that the original annotation of genes encoding 16S ribosomal RNA (*rrsA-D*) excluded the sequence required for ribosome binding site (RBS) recognition (45). Fortunately, neither RAST or PGAP reproduced this erroneous annotation and the D39V annotation includes these sites. Subsequently, we continued with correcting annotated translational initiation sites (TISs, start codons). While accurate TIS identification is challenging, 45 incorrectly annotated start codons could be identified by looking at the relative position of the corresponding transcriptional start sites (TSS/+1, described below). These TISs were corrected in the D39V annotation (**Supplementary Table S3**). Finally, we evaluated the genome-wide quality of TISs using a statistical model that

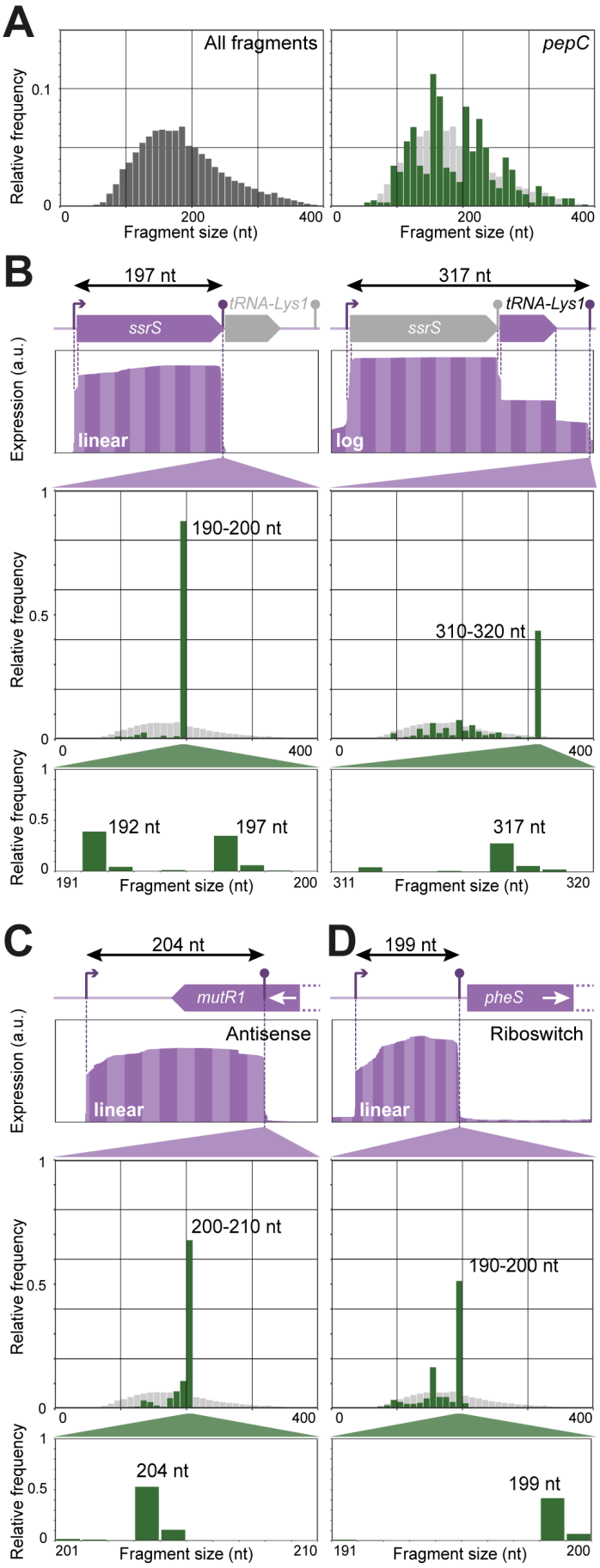

Figure 4. Detection of small RNA features. **A.** Size distributions of entire sequencing library (left) and of fragments ending in the terminator region of *pepC* (right), as determined from paired-end sequencing. Because *pepC* (1.3 kbp) is much longer than the typical fragment length in Illumina sequencing, its size distribution reflects random fragmentation of the full-length transcript and serves as a negative control. **B.** Detection of *ssrS* (left) and joint *ssrS*/*tRNA-Lys1* (right) transcripts. Top: coverage plots. Due to high abundance of *ssrS*, coverage is shown on log-scale in the right panel. Bottom: size distributions (bin sizes 10 and 1) reveal 5'-processed and -unprocessed *ssrS* (left) and full-length *ssrS*/*tRNA-Lys1* transcripts (right). **C.** Detection of an sRNA antisense to the 3'-region of *mutR1*. Top: coverage plots. Bottom: size distribution (bin sizes 10 and 1). **D.** Detection of a riboswitch structure upstream of *pheS*. Top: coverage plots. Bottom: size distribution (bin sizes 10 and 1).

compares the observed and expected distribution of the positions of alternative TISs relative to an annotated TIS (15). The developers suggested that a correlation score below 0.9 is indicative of poorly annotated TISs. In contrast to the D39W (0.899) and PGAP (0.873) annotations, our curated D39V annotation (0.945) excels on the test, emphasizing our annotation's added value to pneumococcal research.

Paired-end sequencing data contains the key to detection of small RNA features

After the sequence- and database-driven annotation process, we proceeded to study the transcriptome of *S. pneumoniae*. We pooled RNA from cells grown at four different conditions (BioRxiv: https://doi.org/10.1101/283739), to maximize the number of expressed genes. Strand-specific, paired-end RNA-seq data of the control library was used to extract start and end points and fragment sizes of the sequenced fragments. In **Figure 4A**, the fragment size distribution of the entire library is shown, with a mode of approximately 150 nucleotides and a skew towards larger fragments. We applied a peak-calling routine to determine the putative 3'-ends of sequenced transcripts. For each of the identified peaks, we extracted all read pairs that were terminated in that specific peak region and compared the size distribution of that subset of sequenced fragments to the library-wide distribution to identify putative sRNAs (See **Materials and Methods** for more details). We focused on sRNA candidates that were found in intergenic regions. Using the combination of sequencing-driven detection, Northern blotting (**Supplementary Figure S5)**, convincing homology

D39V coordinates	Locus	Old locus	Gene name	Product	Note	Literature[γ]
23967-24065 (+)	SPV_2078		*ccnC*	Small regulatory RNA csRNA3		**46**, 47, 49
24632-24707 (+)	SPV_0026	SPD_0026	*scRNA*	scRNA	RNA component of signal recognition particle	82
29658-29873 (+)	SPV_2081		*srf-01*	ncRNA of unknown function	Not validated experimentally	
39980-40186 (+)	SPV_2084		*srf-02*	ncRNA of unknown function		47, **50**
41719-41886 (+)	SPV_2086		*srf-03*	ncRNA of unknown function	Contains BOX element	16, 84
132229-132308 (+)	SPV_2107		*srf-04*	ncRNA of unknown function		**48**, 85
149645-149877 (+)[β]	SPV_2119		*srf-05*	ncRNA of unknown function		**48**, 85
150711-150914 (-)	SPV_2120		*srf-06*	ncRNA of unknown function	Antisense to *mutR1* (SPV_0144); not validated experimentally	
212734-212881 (+)	SPV_2125		*ccnE*	Small regulatory RNA csRNA5	Involved in stationary phase autolysis	**46**, 47, 49, **50**
231599-231691 (+)	SPV_2129		*ccnA*	Small regulatory RNA csRNA1		**46**, 47, **48**, 50, 85
231786-231883 (+)	SPV_2130		*ccnB*	Small regulatory RNA csRNA2		**46**, 47
232279-232354 (+)[β]	SPV_2131		*srf-07*	Type I addiction module antitoxin, Fst family		50, 51
234171-234264 (+)	SPV_2133		*ccnD*	Small regulatory RNA csRNA4	Involved in stationary phase autolysis	**46**, 47
282194-282479 (+)	SPV_2139		*srf-08*	ncRNA of unknown function		49, **50**
344607-345007 (+)	SPV_0340	SPD_0340	*rnpB*	Ribonuclease P RNA component		86
440763-440842 (+)	SPV_2167		*srf-09*	23S methyl RNA motif		49
508697-508842 (+)	SPV_2185		*srf-10*	ncRNA of unknown function		50
587896-587989 (+)	SPV_2200		*srf-11*	ncRNA of unknown function		47, 49, **50**
742022-742141 (-)[α]	SPV_2226		*srf-12*	ncRNA of unknown function	Implicated in competence	**49**
781595-781939 (+)	SPV_0769	SPD_0769	*ssrA*	tmRNA		47, **49**, 50, 82
826260-826587 (+)	SPV_2247		*srf-13*	ncRNA of unknown function		47, 49, 50, **this study**
863157-863283 (+)	SPV_2258		*srf-14*	L20 leader		47, **48**, 49, 50, 85
963342-963439 (-)	SPV_2270		*srf-15*	L21 leader		47, 49
1037649-1037755 (-)	SPV_2291		*srf-16*	ncRNA of unknown function		50
1051910-1052049 (-)	SPV_2292		*srf-17*	*asd* RNA motif		**48**, 49, 50, 82, 87
1079561-1079658 (-)	SPV_2300		*srf-18*	ncRNA of unknown function	Similar to *S. aureus* RsaK	83, **this study**
1170746-1170923 (+)[β]	SPV_2317		*srf-19*	ncRNA of unknown function		50
1216304-1216425 (-)	SPV_2330		*srf-20*	L10 leader		47, **48**, **49**, 50, 82
1528520-1528643 (-)	SPV_2378		*srf-21*	ncRNA of unknown function	Not validated experimentally	
1548804-1549088 (-)	SPV_2383		*srf-22*	ncRNA of unknown function		50
1598326-1598522 (+)	SPV_2392		*ssrS*	6S RNA		47, **49**
1759778-1759868 (-)	SPV_2421		*srf-23*	Lacto-*rpoB* leader		87
1873736-1873781 (-)	SPV_2433		*srf-24*	ncRNA of unknown function	Not validated experimentally	49
1892857-1893007 (-)	SPV_2436		*srf-25*	ncRNA of unknown function		47, **49**
1949385-1949547 (+)	SPV_2442		*srf-26*	ncRNA of unknown function	Not validated experimentally	
1973172-1973570 (-)[α]	SPV_2447		*srf-27*	Type I addiction module antitoxin, Fst family	Not detected in this study due to its size; annotation based on sequence similarity	51
1973509-1973913 (-)	SPV_2449		*srf-28*	Type I addiction module antitoxin, Fst family		51
2008242-2008356 (-)	SPV_2454		*srf-29*	ncRNA of unknown function	Similar to *L. welshimeri* LhrC; not validated experimentally	
2020587-2020685 (-)	SPV_2458		*srf-30*	ncRNA of unknown function		**This study**

Table 2. All annotated small RNA features. Coordinates shown in boldface represent small RNA features not previously reported in *S. pneumoniae*. [α]Not detected in this study, exact coordinates uncertain. [β]Alternative terminator present. [γ]Studies containing Northern blot validation are highlighted in boldface.

with previously validated sRNAs, and/or presence of two or more regulatory features (e.g. TSSs and terminators, see below), we identified 63 small RNA features. We annotated 39 of these as sRNAs (**Table 2**) and 24 as riboswitches (**Supplementary Table S6**).

Until now, several small RNA features have been reliably validated by Northern blot in *S. pneumoniae* strains D39, R6 and TIGR4 (46–50). Excluding most validation reports by Mann et al. due to discrepancies found in their data, 34 validated sRNAs were conserved in D39V. Among the 63 here-detected features, we recovered and refined the coordinates of 33 out of those 34 sRNAs, validating our sRNA detection approach.

One of the detected sRNAs is the highly abundant 6S RNA (**Figure 4B**, left), encoded by *ssrS*, which is involved in transcription regulation. Notably, both automated annotations (RAST and PGAP) failed to report this RNA feature. We observed two different sizes for this feature, probably corresponding to a native and a processed transcript. Interestingly, we also observed a transcript containing both *ssrS* and the downstream tRNA gene. The absence of a TSS between the two genes, suggests that the tRNA is processed from this long transcript (**Figure 4B**, right).

Other detected small transcripts include three type I toxin-antitoxin systems as previously predicted based on orthology (51). Unfortunately, previous annotations omit these systems. Type I toxin-antitoxin systems consist of a toxin peptide (SPV_2132/SPV_2448/SPV_2450) and an antitoxin sRNA (SPV_2131/SPV_2447/SPV_2449). Furthermore, SPV_2120 encodes a novel sRNA that is antisense to the 3'-end of *mutR1* (SPV_0144), which encodes a transcriptional regulator (**Figure 4C**) and might play a role in controlling the production of MutR1, a putative transcriptional regulator (52).

The pneumococcal genome contains at least 24 riboswitches

Several small RNA fragments were located upstream of protein-encoding genes, without an additional TSS in between. This positioning suggests that the observed fragment may be a terminated riboswitch, rather than a functional RNA molecule. Indeed, when we compared expression profiles of 5'-UTRs (untranslated regions) and the gene directly downstream, across infection-relevant conditions (BioRxiv: https://doi.org/10.1101/283739), we found several long UTRs (>100 nt) with a significantly higher average abundance than their respective downstream genes (**Figure 5A**). Such an observation may suggest conditional termination at the end of the putative riboswitch. We queried RFAM and BSRD (22) databases with the 63 identified small RNA features, which allowed us to immediately annotate 22 riboswitches (**Supplementary Table S6**).

Furthermore, we found a candidate sRNA upstream of *pheS* (SPV_0504), encoding a phenylalanyl-tRNA synthetase component. Since Gram-positive bacteria typically regulate tRNA levels using so-called T-box leaders (53), we performed a sequence alignment between nine already identified T-box leaders and the *pheS* leader. Based on these results, we concluded that *pheS* is indeed regulated by a T-box leader (**Figure 4D**).

Finally, we identified a putative PyrR binding site upstream of *uraA* (SPV_1141), which encodes uracil permease. In a complex with uridine 5'-monophosphate (UMP), PyrR binds to 5'-UTR regions of pyrimidine synthesis operons, causing the regions to form hairpin structures and terminate transcription (54). This mechanism was already shown to regulate the expression of *uraA* in both Gram-positive (55) and Gram-negative (56) bacteria. Combining descriptions of PyrR in other species and sequence similarity with 3 other identified PyrR binding sites, we annotated this small RNA feature as a PyrR-regulated riboswitch, completing the 24 annotated riboswitches in D39V. To validate these annotations, we transcriptionally integrated a gene encoding firefly luciferase (*luc*) behind the four putative PyrR-regulated operons (*pyrFE*, *pyrKDb*, *pyrRB-carAB, uraA*), along with four negative controls (*pyrG*, *pyrDa-holA*, *pyrH*, *ung-mutX-pyrC*) that do contain genes involved in pyrimidine metabolism but lack a putative PyrR-binding site (**Figure 5B**). As expected, expression of operons lacking a PyrR riboswitch is not affected by increasing concentrations of uridine (for example in *pyrH-luc*, **Figure 5C**), while the putative PyrR-regulated operons are strongly repressed (for example in *pyrRB-carA-luc*, **Figure 5D**). Finally, we tested other intermediates from uridine metabolism and observed a similar trend when uracil was added instead. Surprisingly, UMP exhibited a marginal effect on the expression of the *pyrR* operon, with much weaker repression than observed

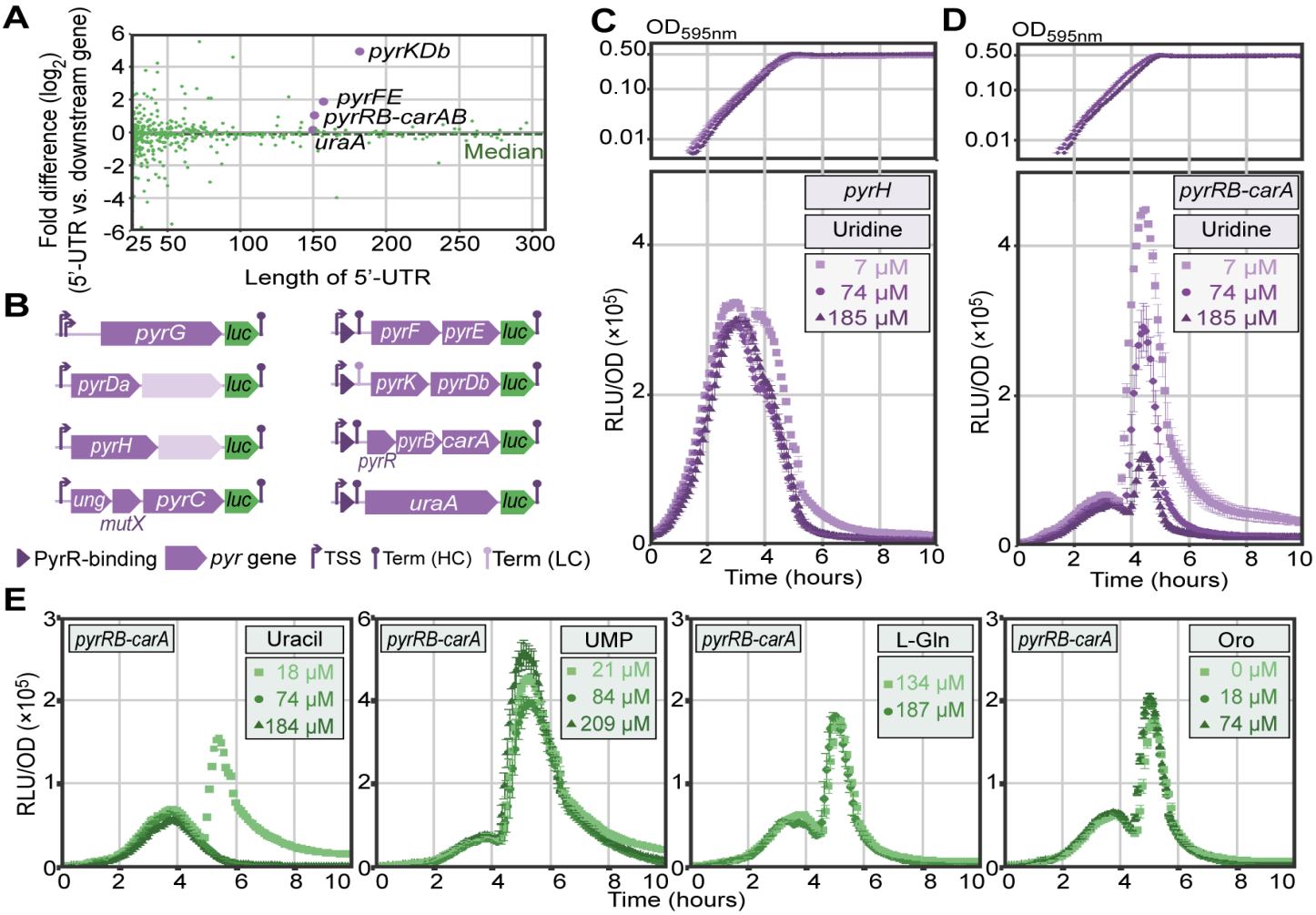

Figure 5. Long 5'-UTRs and validation of riboswitches. **A.** Mean ratio (^{2}log) of 5'-UTR expression to that of the corresponding downstream gene. This fold difference was measured across 22 infection-relevant conditions (BioRxiv: https://doi.org/10.1101/283739) and plotted against 5'-UTR length. Among the overrepresented 5'-UTRs are several pyrimidine metabolism operons (*pyrKDb*, *pyrFE*, *pyrRB-carAB*). **B.** Firefly *luc* integration constructs used to assay the response to pyrimidine metabolites. Left: operons containing pyrimidine-related genes but lacking an upstream PyrR binding site. Right: pyrimidine operons with a predicted PyrR binding site (dark purple triangles). **C-D.** Growth (top) and normalized luciferase activity (RLU/OD, bottom) of *pyrH-luc* (**C**) and *pyrRB-carA-luc* (**D**) cells with varying uridine concentrations. **E.** Normalized luciferase activity (RLU/OD) of *pyrRB-carA-luc* cells with varying concentrations of (from left to right) uracil, uridine 5'-monophosphate, L-glutamine and orotic acid.

with comparable uridine concentrations. The response might indicate that UMP is not efficiently imported by the cell, resulting in uridine starvation and high expression of the operon. Other intermediates in the pyrimidine metabolism, L-glutamine and orotic acid, did not incite an observable effect on *pyrR* expression.

Combining putative termination peaks and predicted stem-loops to annotate terminators

Aside from the customary annotation of gene and pseudogene features, we complemented the annotation with transcriptional regulatory elements including TSSs, terminators and other transcription regulatory elements. First, we set out to identify transcriptional terminators. Since *S. pneumoniae* lacks the Rho factor (57), all of its terminators will be Rho-independent. We combined the previously generated list of putative termination peaks with a highly sensitive terminator stem-loop prediction by TransTermHP (20) and annotated a terminator when a termination peak was found within 10 bps of a predicted stem loop. The genome-wide distribution of the 748 annotated terminators is shown in **Figure 6A**. Terminator characteristics (**Supplementary Figure S6**) resembled those observed in *B. subtilis* and *E. coli* (58). We further compared the number of sequenced fragments ending at each termination peak with the number of fragments covering the peak without being terminated. This allowed us to estimate the termination efficiency of each

terminator in the genome, which is listed for all terminators in **Supplementary Table S7** and visible in PneumoBrowse (see below).

Direct enrichment of primary transcripts allows precise identification of transcription start sites

We determined TSSs using a novel technique (Cappable-seq, (6)): primary transcripts (i.e. not processed or degraded) were directly enriched by exploiting the 5'-triphosphate group specific for primary transcripts, in contrast with the 5'-monophosphate group that processed transcripts carry. We then sequenced the 5'-enriched library, along with a non-enriched control library, and identified 981 TSSs with at least 2.5 fold higher abundance in the enriched library than in the untreated control library. Taking into account the possibility of rapid 5'-triphosphate processing, we added 34 (lower confidence, LC) TSSs that were not sufficiently enriched, yet met a set of strict criteria (see **Materials and Methods** for details). The genomic distribution of these 1,015 TSSs is shown in **Figure 6A**. Importantly, our data could reproduce previously experimentally determined TSSs (59, 60).

Strong nucleotide bias on transcriptional start sites

Analysis of the nucleotide distribution across all TSSs showed a strong preference for adenine (A, 63% vs. 30% genome-wide) and, to a lesser extent, guanine (G, 28% vs. 22%) on +1 positions (**Figure 6B**). While it has to be noted that especially upstream regions are biased by the presence of transcriptional regulatory elements (see below), we observed a general bias around the TSSs towards sequences rich in adenine and poor in cytosine (C) (**Figure 6C**). A striking exception to this rule is observed on the -1 position, where thymine (T) is the most frequently occurring base (51%) while A is underrepresented (16%). Interestingly, similar biases are absent in *Helicobacter pylori* (61), while they were observed in *Escherichia coli* (6) and *Salmonella enterica* (62), albeit to a lesser extent than in the pneumococcus.

Promoter analysis reveals regulatory motifs for the majority of transcription start sites

We defined the 100 bps upstream of each TSS as the promoter region and used the MEME Suite (63) to scan each promoter region for regulatory motifs (**Figure 6D**), thereby identifying 382 RpoD sites (σ_A, (64)), 19 ComX sites (σ_X, (65)) and 13 ComE sites (66). In addition to the complete RpoD sites, another 449 promoter regions contained only a -10 (64) or an extended -10 sequence (67). Finally, we annotated other transcription-factor binding sites, including those of CodY and CcpA, as predicted by RegPrecise (68).

Characterization of TSSs based on genomic context and putative operon definition

Subsequently, the TSSs were classified based on their position relative to annotated genomic features (61), categorizing them as primary (*P*, the only or strongest TSS upstream of feature), secondary (*S*, upstream of a feature, but not the strongest TSS), internal (*I*, inside annotated feature), antisense (*A*, antisense to an annotated feature) and/or orphan (*O*, not in any of the other categories), as shown in **Figure 6E**. Notably, we categorized antisense TSSs into three classes depending on which part of the feature the TSS overlaps: A_5 (in the 100 bps upstream), A_0 (within the feature), or A_3 (in the 100 bps downstream). Doing so, we could classify 827 TSSs (81%) as primary (pTSS), underscoring the quality of TSS calls. Finally, we defined putative operons, starting from each pTSS. The end of an operon was marked by (i) the presence of an efficient (>80%) terminator, (ii) a strand swap between features, or (iii) weak correlation of expression (<0.75) across 22 infection-relevant conditions (BioRxiv: https://doi.org/10.1101/283739) between consecutive features. The resulting 827 operons cover 1,390 (65%) annotated features (**Supplementary Table S8**) and are visualized in PneumoBrowse.

Coding sequence leader analysis reveals ribosome-binding sites and many leaderless genes

We evaluated the relative distance between primary TSSs and the first gene of their corresponding operons (**Figure 7A**). Interestingly, of the 767 operons starting with a coding sequence (or pseudogene), 80 have a 5'-UTR too short to harbor a potential Shine-Dalgarno (SD) ribosome-binding motif (leaderless; **Figure 7A**, inset). In fact, for 69 of those (9% of all operons), transcription starts exactly on the first base of the coding sequence, which is comparable to findings in other organisms (69, 70). Although the presence of leaderless operons suggests the dispensability of ribosomal binding sites (RBSs), motif enrichment analysis (63) showed that 69% of all coding sequences do have an RBS upstream. To evaluate the translation initiation efficiency of leaderless coding sequences, we selected the promoter of leaderless *pbp2A* (SPV_1821) and cloned it upstream of a reporter cassette, containing *luc* and *gfp*. The cloning approach was threefold: while *gfp* always contained an upstream RBS, *luc* contained either (i) no leader, (ii) an upstream RBS, or (iii) no leader and a premature stop codon

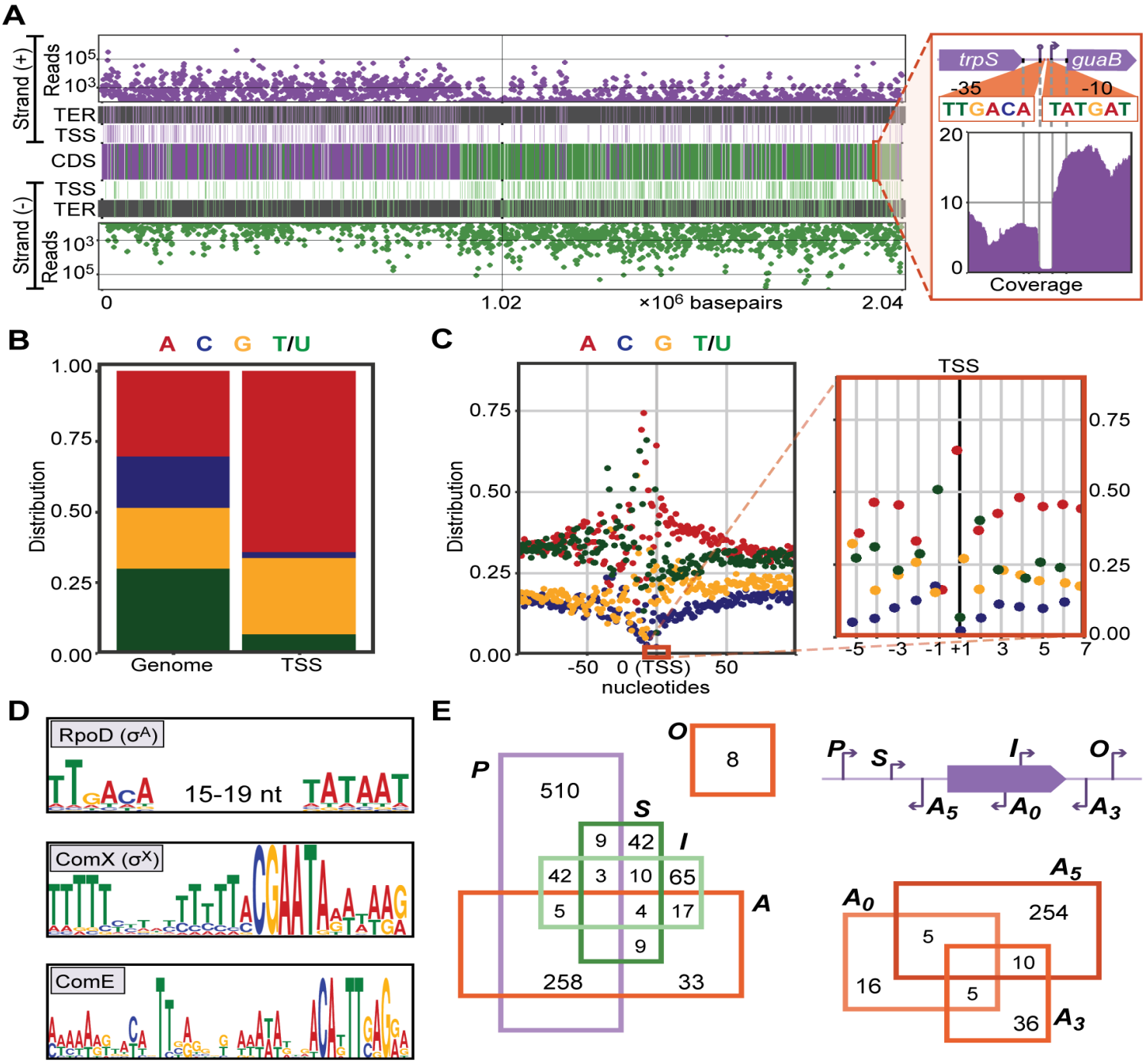

Figure 6. Characterization of transcriptional start sites. **A.** Genome-wide distributions of sequencing reads, terminators (TER) and transcriptional start sites (TSS) on the positive (top) and negative strand (bottom) and annotated coding sequences (middle) are closely correlated. Features on the positive and negative strand are shown in purple and green, respectively. The inset shows the coverage in the *trpS-guaB* locus, along with a detected terminator (84% efficient), TSS, and RpoD-binding elements (-35 and -10). **B.** Nucleotide utilization on +1 (TSS) positions, compared to genome content. **C.** Nucleotide utilization around TSSs (left: -100 to +101, right: -6 to +7). **D.** RpoD (σ_A), ComX (σ_X) and ComE binding sites found upstream of TSSs. **E.** TSSs were divided, based on local genomic context, into five classes (top right): primary (P, only or strongest TSS within 300 nt upstream of a feature), secondary (S, within 300 nt upstream of a feature, not the strongest), internal (I, inside a feature), antisense (A), and orphan (O, in none of the other classes). Results are shown in a Venn diagram (left). Antisense TSSs were further divided into 3 subclasses (bottom right): A_5 (upstream of feature), A_3 (downstream of feature), and A_0 (inside feature).

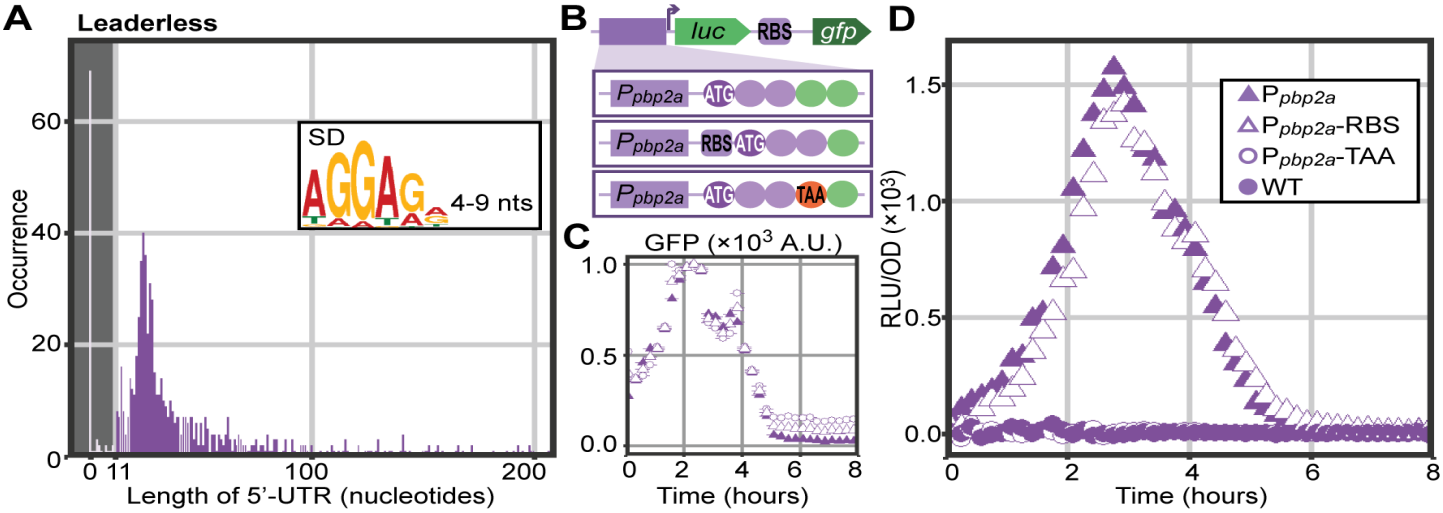

Figure 7. Leaderless coding-sequences and pseudogenes. **A.** Length distribution of 5'-untranslated regions of all 767 operons that start with a CDS or pseudogene shows 69 (9%) leaderless transcripts. The shaded area contains leaders too short to contain a potential Shine-Dalgarno (SD) motif (inset). **B.** In an ectopic locus, we cloned the promoter of leaderless gene *pbp2a* upstream of an operon containing *luc* (firefly luciferase) and *gfp* (superfolder GFP), the latter always led by a SD-motif (RBS). Three constructs were designed: (i) native promoter including the first three codons of *pbp2a*, (ii) the same as (i), but with an additional RBS in front of the coding sequence, (iii) the same as (i), but with a stop codon (TAA) after three codons. **C.** Fluorescence signal of all three constructs shows successful transcription. **D.** Normalized luminescence signals in all three strains and a wild-type control strain. Successful translation of *luc* is clear from luminescence in P_{pbp2a}-*luc* (filled triangles) and adding an RBS had no measurable effect on luciferase activity (open triangles). Integration of an early stop codon (open circles) reduced signal to wild-type level (filled circles).

(**Figure 7B**). Fluorescence signals were comparable in all three constructs, indicating that transcription and translation of the downstream *gfp* was unaffected by the different *luc* versions (**Figure 7C**). Bioluminescence assays (**Figure 7D**) showed, surprisingly, that the introduction of an upstream RBS had no effect on luciferase production. Additionally, the absence of signal in the strain with an early stop codon allowed us to rule out potential downstream translation initiation sites. This experiment provides direct evidence that leaderless genes can be efficiently translated.

PneumoBrowse integrates all elements of the deep D39V genome annotation

Finally, we combined all elements of the final annotation to compile a user-friendly, uncluttered genome browser, PneumoBrowse (https://veeninglab.com/pneumobrowse, **Figure 8**). Based on JBrowse (7), the browser provides an overview of all genetic features, along with transcription regulatory elements and sequencing coverage data. Furthermore, it allows users to flexibly search for their gene of interest using either its gene name or locus tag in D39V, D39W or R6 (prefixes 'SPV_', 'SPD_' and 'spr', respectively). Right-clicking on features reveals further information, such as its gene expression profile across 22 infection-relevant conditions and co-expressed genes (BioRxiv: https://doi.org/10.1101/283739).

DISCUSSION

Annotation databases such as SubtiWiki (71) and EcoCyc (72) have tremendously accelerated gene discovery, functional analysis and hypothesis-driven research in the fields of model organisms *Bacillus subtilis* and *Escherichia coli*. It was therefore surprising that such a resource was not yet available for the important opportunistic human pathogen *Streptococcus pneumoniae*, annually responsible for more than a million deaths (73). With increases in vaccine escape strains and antimicrobial resistance, a better understanding of this organism is required, enabling the identification of new vaccine and antibiotic targets. By exploiting recent technological and scientific advances, we now mapped and deeply annotated the pneumococcal genome at an unprecedented level of detail. We combined cutting-edge technology (e.g. SMRT sequencing, Cappable-seq, a novel sRNA detection method and several bioinformatic annotation tools), with thorough manual curation to create PneumoBrowse, an unparallelled resource that allows users to browse through the newly assembled pneumococcal genome and inspect encoded features along with regulatory

Figure 8. PneumoBrowse. A screenshot of the *def2* locus in PneumoBrowse shows the richness of available data. In the left pane, tracks can be selected. In the right pane, annotation tracks (*Putative operons, Genes, Regulatory features*) are shown along with data tracks (*Control coverage, Control first base, 5'-enriched first base*). Regulatory features annotated upstream of *def2* include a 98% efficient terminator, ComE- and RpoD-binding sites, a TSS and an RBS. Additionally, a context menu (upon right-click) provides links to external resources, such as PneumoExpress (BioRxiv: https://doi.org/10.1101/283739).

elements, repeat regions and other useful properties (**Figure 8**). Additionally, the browser provides direct linking to expression and co-expression data in PneumoExpress (BioRxiv: https://doi.org/10.1101/283739). The here-applied methods and approaches might serve as a roadmap for genomics studies in other bacteria. It should be noted that the detection of sRNAs and transcription regulatory elements, such as TSSs and terminators, relies on active transcription of these features in the studied conditions. Therefore, despite the already high feature density in the current annotation, new elements may still be elucidated in the future.

We showed that D39V strongly resembles the ancestral strain NCTC 7466, although it has lost the cryptic plasmid pDP1 and contains a number of single-nucleotide mutations. Larger discrepancies were observed with the previous D39 assembly ((24), **Table 1**). Upon comparison to several other pneumococcal strains, we identified multiple large chromosomal inversions (**Figure 2**). Although the inversion observed in D39V contained the terminus of replication, it created an asymmetry in replichore length, visible both in the location of the *dif*/Xer locus (**Figure 3A)** and the direction of encoded features (**Figure 6A**). Such inversions are likely to be facilitated by bordering repeat regions. In this context, the many identified repeat regions (i.a. BOX elements, RUPs and IS elements) may play an important role in pneumococcal evolution, by providing a template for intragenomic recombination or inversion events (74, 75).

We further unveiled the mosaic nature of pneumococcal histidine triad proteins, and especially vaccine candidate PhtD, variation of which is facilitated both by the aforementioned chromosomal inversion and more local recombination events. This observation expands the set of previously identified highly variable pneumococcal antigens (76, 77).

Despite the differences observed with NCTC 7466, D39V is genetically stable under laboratory conditions, as both the *hsdS* locus and the *ter* configuration are identical in all analyzed derivative strains (*not shown*). Additionally and importantly, D39V is virulent in multiple infection models (78–81), making it an ideal, stable genetic workhorse for pneumococcal research. Therefore, the strain will be made available to the community through the PHE National Collection of Type Cultures (NCTC).

Besides annotating 1,877 CDSs, 12 rRNAs, 58 tRNAs and 165 pseudogenes, we identified and annotated 39 sRNAs and 24 riboswitches (**Table 2, Supplementary Table S6**, **Figure 5**). We look forward to seeing future studies into the function of the newly identified sRNAs. The novel sRNA detection method employed, based on fragment size distribution in paired-end sequencing data (**Figure 4**), while already sensitive, can probably be further improved by employing sRNA-specific library techniques.

Finally, to understand bacterial decision-making, it is important to obtain detailed information about gene expression and regulation thereof. Therefore, we identified transcriptional start sites and terminators (**Figure 6)**, sigma

factor binding sites and other transcription-regulatory elements. This architectural information, complemented by expression data from PneumoExpress (BioRxiv: https://doi.org/10.1101/283739), will be invaluable to future microbiological research and can be readily accessed via PneumoBrowse.

AVAILABILITY

Python scripts used for data analysis are available at https://github.com/veeninglab/Spneu-deep-annotation. PneumoBrowse can be accessed via https://veeninglab.com/pneumobrowse. PneumoExpress can be accessed via https://veeninglab.com/pneumoexpress. All detected and annotated small RNA features were submitted to the Bacterial Small Regulatory RNA Database (BSRD).

ACCESSION NUMBERS

SMRT sequencing data used for de novo assembly was deposited to SRA (SRP063763). Unless stated otherwise, RNA-seq datasets used in this study can be found in the SRA repository (SRP133365). The D39V assembly and annotation was submitted to GenBank (CP027540).

ACKNOWLEDGEMENTS

We are grateful to A. Patrignani (Functional Genomics Center Zurich) for support in SMRT sequencing; F. Thümmler (vertis Biotechnologie AG) for support in Illumina RNA-seq; A. de Jong and S. Holsappel for (bio)informatics support; J. Donner for kindly providing DNA-seq data for the three investigated clinical isolates; and S.B. van der Meulen for his advice regarding Northern blotting.

FUNDING

Work in the Veening lab is supported by the Swiss National Science Foundation (project grant 31003A_172861; a VIDI fellowship (864.11.012) of the Netherlands Organisation for Scientific Research (NWO-ALW); a JPIAMR grant (50-52900-98-202) from the Netherlands Organisation for Health Research and Development (ZonMW); and ERC starting grant 337399-PneumoCell.

CONFLICT OF INTEREST

None declared.

REFERENCES

1. Swetha,R.G., Sekar,D.K.K., Devi,E.D., Ahmed,Z.Z., Ramaiah,S., Anbarasu,A. and Sekar,K. (2014) *Streptococcus pneumoniae* Genome Database (SPGDB): a database for strain specific comparative analysis of *Streptococcus pneumoniae* genes and proteins. *Genomics*, **104**, 582–586.

2. Jolley,K.A. and Maiden,M.C.J. (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, **11**, 595.

3. Wattam,A.R., Davis,J.J., Assaf,R., Boisvert,S., Brettin,T., Bun,C., Conrad,N., Dietrich,E.M., Disz,T., Gabbard,J.L., *et al.* (2017) Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res.*, **45**, D535–D542.

4. Tatusova,T., DiCuccio,M., Badretdin,A., Chetvernin,V., Ciufo,S. and Li,W. (2013) Prokaryotic Genome Annotation Pipeline. In *The NCBI Handbook*. National Center for Biotechnology Information (US).

5. Avery,O.T., Macleod,C.M. and McCarty,M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.*, **79**, 137–158.

6. Ettwiller,L., Buswell,J., Yigit,E. and Schildkraut,I. (2016) A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. *BMC Genomics*, **17**, 199.

7. Skinner,M.E., Uzilov,A.V., Stein,L.D., Mungall,C.J. and Holmes,I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.

8. Slager,J., Kjos,M., Attaiech,L. and Veening,J.-W. (2014) Antibiotic-induced replication stress triggers bacterial competence by increasing gene dosage near the origin. *Cell*, **157**, 395–406.

9. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

10. Barrick,J.E., Colburn,G., Deatherage,D.E., Traverse,C.C., Strand,M.D., Borges,J.J., Knoester,D.B., Reba,A. and Meyer,A.G. (2014) Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. *BMC Genomics*, **15**, 1039.

11. Overbeek,R., Olson,R., Pusch,G.D., Olsen,G.J., Davis,J.J., Disz,T., Edwards,R.A., Gerdes,S., Parrello,B., Shukla,M., *et al.* (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.*, **42**, D206–D214.

12. Boratyn,G.M., Camacho,C., Cooper,P.S., Coulouris,G., Fong,A., Ma,N., Madden,T.L., Matten,W.T., McGinnis,S.D., Merezhuk,Y., *et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.*, **41**, W29–W33.

13. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.

14. Marchler-Bauer,A. and Bryant,S.H. (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.*, **32**, W327-331.

15. Overmars,L., Siezen,R.J. and Francke,C. (2015) A novel quality measure and correction procedure for the annotation of microbial translation initiation sites. *PloS One*, **10**, e0133691.

16. Martin,B., Humbert,O., Camara,M., Guenzi,E., Walker,J., Mitchell,T., Andrew,P., Prudhomme,M., Alloing,G. and Hakenbeck,R. (1992) A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*. *Nucleic Acids Res.*, **20**, 3479–3483.

17. Oggioni,M.R. and Claverys,J.P. (1999) Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in *Streptococcus pneumoniae*. *Microbiol. Read. Engl.*, **145 (Pt 10)**, 2647–2653.

18. Croucher,N.J., Vernikos,G.S., Parkhill,J. and Bentley,S.D. (2011) Identification, variation and transcription of pneumococcal repeat sequences. *BMC Genomics*, **12**, 120.

19. Siguier,P., Perochon,J., Lestrade,L., Mahillon,J. and Chandler,M. (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.*, **34**, D32-36.

20. Kingsford,C.L., Ayanbule,K. and Salzberg,S.L. (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol.*, **8**, R22.

21. Kalvari,I., Argasinska,J., Quinones-Olvera,N., Nawrocki,E.P., Rivas,E., Eddy,S.R., Bateman,A., Finn,R.D. and Petrov,A.I. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, **46**, D335-342.

22. Li,L., Huang,D., Cheung,M.K., Nong,W., Huang,Q. and Kwan,H.S. (2013) BSRD: a repository for bacterial small regulatory RNA. *Nucleic Acids Res.*, **41**, D233-238.

23. Kjos,M., Miller,E., Slager,J., Lake,F.B., Gericke,O., Roberts,I.S., Rozen,D.E. and Veening,J.-W. (2016) Expression of *Streptococcus pneumoniae* bacteriocins is induced by antibiotics via regulatory interplay with the competence system. *PLoS Pathog.*, **12**, e1005422.

24. Lanie,J.A., Ng,W.-L., Kazmierczak,K.M., Andrzejewski,T.M., Davidsen,T.M., Wayne,K.J., Tettelin,H., Glass,J.I. and Winkler,M.E. (2007) Genome sequence of Avery's virulent serotype 2 strain D39 of *Streptococcus pneumoniae* and comparison with that of unencapsulated laboratory strain R6. *J. Bacteriol.*, **189**, 38–51.

25. Afzal,M., Shafeeq,S., Manzoor,I. and Kuipers,O.P. (2015) GalR acts as a transcriptional activator of *galKT* in the presence of galactose in *Streptococcus pneumoniae*. *J. Mol. Microbiol. Biotechnol.*, **25**, 363–371.

26. Manso,A.S., Chai,M.H., Atack,J.M., Furi,L., De Ste Croix,M., Haigh,R., Trappetti,C., Ogunniyi,A.D., Shewell,L.K., Boitano,M., *et al.* (2014) A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. *Nat. Commun.*, **5**, 5055.

27. Feng,Z., Li,J., Zhang,J.-R. and Zhang,X. (2014) qDNAmod: a statistical model-based tool to reveal intercellular heterogeneity of DNA modification from SMRT sequencing data. *Nucleic Acids Res.*, **42**, 13488–13499.

28. Clark,T.A., Murray,I.A., Morgan,R.D., Kislyuk,A.O., Spittle,K.E., Boitano,M., Fomenkov,A., Roberts,R.J. and Korlach,J. (2012) Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res.*, **40**, e29.

29. Le Bourgeois,P., Bugarel,M., Campo,N., Daveran-Mingot,M.-L., Labonté,J., Lanfranchi,D., Lautier,T., Pagès,C. and Ritzenthaler,P. (2007) The unconventional Xer recombination machinery of Streptococci/Lactococci. *PLoS Genet.*, **3**, e117.

30. Donner,J., Bunk,B., Schober,I., Spröer,C., Bergmann,S., Jarek,M., Overmann,J. and Wagner-Döbler,I. (2017) Complete genome sequences of three multidrug-resistant clinical isolates of *Streptococcus pneumoniae* serotype 19A with different susceptibilities to the myxobacterial metabolite carolacton. *Genome Announc.*, **5**, e01641-16.

31. Khan,M.N. and Pichichero,M.E. (2012) Vaccine candidates PhtD and PhtE of *Streptococcus pneumoniae* are adhesins that elicit functional antibodies in humans. *Vaccine*, **30**, 2900–2907.

32. Seiberling,M., Bologa,M., Brookes,R., Ochs,M., Go,K., Neveu,D., Kamtchoua,T., Lashley,P., Yuan,T. and Gurunathan,S. (2012) Safety and immunogenicity of a pneumococcal histidine triad protein D vaccine candidate in adults. *Vaccine*, **30**, 7455–7460.

33. Odutola,A., Ota,M.O.C., Antonio,M., Ogundare,E.O., Saidu,Y., Foster-Nyarko,E., Owiafe,P.K., Ceesay,F., Worwui,A., Idoko,O.T., *et al.* (2017) Efficacy of a novel, protein-based pneumococcal vaccine against nasopharyngeal carriage of *Streptococcus pneumoniae* in infants: A phase 2, randomized, controlled, observer-blind study. *Vaccine*, **35**, 2531–2542.

34. Yun,K.W., Lee,H., Choi,E.H. and Lee,H.J. (2015) Diversity of pneumolysin and pneumococcal histidine triad protein D of *Streptococcus pneumoniae* isolated from invasive diseases in Korean children. *PloS One*, **10**, e0134055.

35. Oggioni,M.R., Iannelli,F. and Pozzi,G. (1999) Characterization of cryptic plasmids pDP1 and pSMB1 of *Streptococcus pneumoniae*. *Plasmid*, **41**, 70–72.

36. Liu,X., Li,J.-W., Feng,Z., Luo,Y., Veening,J.-W. and Zhang,J.-R. (2017) Transcriptional repressor PtvR regulates phenotypic tolerance to vancomycin in *Streptococcus pneumoniae*. *J. Bacteriol.*, **199**, e00054-17.

37. Zheng,J.J., Sinha,D., Wayne,K.J. and Winkler,M.E. (2016) Physiological roles of the dual phosphate transporter systems in low and high phosphate conditions and in capsule maintenance of *Streptococcus pneumoniae* D39. *Front. Cell. Infect. Microbiol.*, **6**, 63.

38. Fleurie,A., Lesterlin,C., Manuse,S., Zhao,C., Cluzel,C., Lavergne,J.-P., Franz-Wachtel,M., Macek,B., Combet,C., Kuru,E., *et al.* (2014) MapZ marks the division sites and positions FtsZ rings in *Streptococcus pneumoniae*. *Nature*, **516**, 259–262.

39. Holečková,N., Doubravová,L., Massidda,O., Molle,V., Buriánková,K., Benada,O., Kofroňová,O., Ulrych,A. and Branny,P. (2014) LocZ is a new cell division protein involved in proper septum placement in *Streptococcus pneumoniae*. *mBio*, **6**, e01700-01714.

40. Zheng,J.J., Perez,A.J., Tsui,H.-C.T., Massidda,O. and Winkler,M.E. (2017) Absence of the KhpA and KhpB (JAG/EloR) RNA-binding proteins suppresses the requirement for PBP2b by overproduction of FtsA in *Streptococcus pneumoniae* D39. *Mol. Microbiol.*, **106**, 793–814.

41. Stamsås,G.A., Straume,D., Ruud Winther,A., Kjos,M., Frantzen,C.A. and Håvarstein,L.S. (2017) Identification of EloR (Spr1851) as a regulator of cell elongation in *Streptococcus pneumoniae*. *Mol. Microbiol.*, **105**, 954–967.

42. Fenton,A.K., Mortaji,L.E., Lau,D.T.C., Rudner,D.Z. and Bernhardt,T.G. (2016) CozE is a member of the MreCD complex that directs cell elongation in *Streptococcus pneumoniae*. *Nat. Microbiol.*, **2**, 16237.

43. Lowe,T.M. and Chan,P.P. (2016) tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.*, **44**, W54-57.

44. Silva,F.J., Belda,E. and Talens,S.E. (2006) Differential annotation of tRNA genes with anticodon CAT in bacterial genomes. *Nucleic Acids Res.*, **34**, 6015–6022.

45. Hui,A. and de Boer,H.A. (1987) Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.*, **84**, 4762–4766.

46. Halfmann,A., Kovács,M., Hakenbeck,R. and Brückner,R. (2007) Identification of the genes directly controlled by the response regulator CiaR in *Streptococcus pneumoniae*: five out of 15 promoters drive expression of small non-coding RNAs. *Mol. Microbiol.*, **66**, 110–126.

47. Kumar,R., Shah,P., Swiatlo,E., Burgess,S.C., Lawrence,M.L. and Nanduri,B. (2010) Identification of novel non-coding small RNAs from *Streptococcus pneumoniae* TIGR4 using high-resolution genome tiling arrays. *BMC Genomics*, **11**, 350.

48. Tsui,H.-C.T., Mukherjee,D., Ray,V.A., Sham,L.-T., Feig,A.L. and Winkler,M.E. (2010) Identification and characterization of noncoding small RNAs in *Streptococcus pneumoniae* serotype 2 strain D39. *J. Bacteriol.*, **192**, 264–279.

49. Acebo,P., Martin-Galiano,A.J., Navarro,S., Zaballos,Á. and Amblar,M. (2012) Identification of 88 regulatory small RNAs in the TIGR4 strain of the human pathogen *Streptococcus pneumoniae*. *RNA*, **18**, 530–546.

50. Mann,B., van Opijnen,T., Wang,J., Obert,C., Wang,Y.-D., Carter,R., McGoldrick,D.J., Ridout,G., Camilli,A., Tuomanen,E.I., *et al.* (2012) Control of virulence by small RNAs in *Streptococcus pneumoniae*. *PLoS Pathog.*, **8**, e1002788.

51. Fozo,E.M., Makarova,K.S., Shabalina,S.A., Yutin,N., Koonin,E.V. and Storz,G. (2010) Abundance of type I toxin–antitoxin systems in bacteria: searches for new candidates and discovery of novel families. *Nucleic Acids Res.*, **38**, 3743–3759.

52. Hendriksen,W.T., Kloosterman,T.G., Bootsma,H.J., Estevão,S., de Groot,R., Kuipers,O.P. and Hermans,P.W.M. (2008) Site-specific contributions of glutamine-dependent regulator GlnR and GlnR-regulated genes to virulence of *Streptococcus pneumoniae*. *Infect. Immun.*, **76**, 1230–1238.

53. Grigg,J.C., Chen,Y., Grundy,F.J., Henkin,T.M., Pollack,L. and Ke,A. (2013) T box RNA decodes both the information content and geometry of tRNA to affect gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 7240–7245.

54. Lu,Y. and Switzer,R.L. (1996) Transcriptional attenuation of the *Bacillus subtilis pyr* operon by the PyrR regulatory protein and uridine nucleotides in vitro. *J. Bacteriol.*, **178**, 7206–7211.

55. Martinussen,J., Schallert,J., Andersen,B. and Hammer,K. (2001) The pyrimidine operon *pyrRPB-carA* from *Lactococcus lactis*. *J. Bacteriol.*, **183**, 2785–2794.

56. Turnbough,C.L. and Switzer,R.L. (2008) Regulation of pyrimidine biosynthetic gene expression in bacteria: repression without repressors. *Microbiol. Mol. Biol. Rev. MMBR*, **72**, 266–300.

57. D'Heygère,F., Rabhi,M. and Boudvillain,M. (2013) Phyletic distribution and conservation of the bacterial transcription termination factor Rho. *Microbiol. Read. Engl.*, **159**, 1423–1436.

58. de Hoon,M.J.L., Makita,Y., Nakai,K. and Miyano,S. (2005) Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS Comput. Biol.*, **1**, e25.

59. Peters,K., Pipo,J., Schweizer,I., Hakenbeck,R. and Denapaite,D. (2016) Promoter identification and transcription analysis of penicillin-binding protein genes in *Streptococcus pneumoniae* R6. *Microb. Drug Resist. Larchmt. N*, **22**, 487–498.

60. Sánchez-Beato,A.R., López,R. and García,J.L. (1998) Molecular characterization of PcpA: a novel choline-binding protein of *Streptococcus pneumoniae*. *FEMS Microbiol. Lett.*, **164**, 207–214.

61. Sharma,C.M., Hoffmann,S., Darfeuille,F., Reignier,J., Findeiss,S., Sittka,A., Chabas,S., Reiche,K., Hackermüller,J., Reinhardt,R., *et al.* (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, **464**, 250–255.

62. Kröger,C., Dillon,S.C., Cameron,A.D.S., Papenfort,K., Sivasankaran,S.K., Hokamp,K., Chao,Y., Sittka,A., Hébrard,M., Händler,K., *et al.* (2012) The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, E1277-1286.

63. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202-208.

64. Shimada,T., Yamazaki,Y., Tanaka,K. and Ishihama,A. (2014) The whole set of constitutive promoters recognized by RNA polymerase RpoD holoenzyme of *Escherichia coli*. *PloS One*, **9**, e90447.

65. Dagkessamanskaia,A., Moscoso,M., Hénard,V., Guiral,S., Overweg,K., Reuter,M., Martin,B., Wells,J. and Claverys,J.-P. (2004) Interconnection of competence, stress and CiaR regulons in *Streptococcus pneumoniae*: competence triggers stationary phase autolysis of *ciaR* mutant cells. *Mol. Microbiol.*, **51**, 1071–1086.

66. Martin,B., Soulet,A.-L., Mirouze,N., Prudhomme,M., Mortier-Barrière,I., Granadel,C., Noirot-Gros,M.-F., Noirot,P., Polard,P. and Claverys,J.-P. (2013) ComE/ComE~P interplay dictates activation or extinction status of pneumococcal X-state (competence). *Mol. Microbiol.*, **87**, 394–411.

67. Sabelnikov,A.G., Greenberg,B. and Lacks,S.A. (1995) An extended −10 promoter alone directs transcription of the *dpnII* operon of *Streptococcus pneumoniae*. *J. Mol. Biol.*, **250**, 144–155.

68. Novichkov,P.S., Kazakov,A.E., Ravcheev,D.A., Leyn,S.A., Kovaleva,G.Y., Sutormin,R.A., Kazanov,M.D., Riehl,W., Arkin,A.P., Dubchak,I., *et al.* (2013) RegPrecise 3.0--a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics*, **14**, 745.

69. van der Meulen,S.B., de Jong,A. and Kok,J. (2016) Transcriptome landscape of *Lactococcus lactis* reveals many novel RNAs including a small regulatory RNA involved in carbon uptake and metabolism. *RNA Biol.*, **13**, 353–366.

70. Schmidtke,C., Findeiss,S., Sharma,C.M., Kuhfuss,J., Hoffmann,S., Vogel,J., Stadler,P.F. and Bonas,U. (2012) Genome-wide transcriptome analysis of the plant pathogen *Xanthomonas* identifies sRNAs with putative virulence functions. *Nucleic Acids Res.*, **40**, 2020–2031.

71. Michna,R.H., Zhu,B., Mäder,U. and Stülke,J. (2016) SubtiWiki 2.0--an integrated database for the model organism *Bacillus subtilis*. *Nucleic Acids Res.*, **44**, D654-662.

72. Keseler,I.M., Mackie,A., Santos-Zavaleta,A., Billington,R., Bonavides-Martínez,C., Caspi,R., Fulcher,C., Gama-Castro,S., Kothari,A., Krummenacker,M., *et al.* (2017) The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res.*, **45**, D543–D550.

73. O'Brien,K.L., Wolfson,L.J., Watt,J.P., Henkle,E., Deloria-Knoll,M., McCall,N., Lee,E., Mulholland,K., Levine,O.S., Cherian,T., *et al.* (2009) Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. *Lancet Lond. Engl.*, **374**, 893–902.

74. Daveran-Mingot,M.L., Campo,N., Ritzenthaler,P. and Le Bourgeois,P. (1998) A natural large chromosomal inversion in *Lactococcus lactis* is mediated by homologous recombination between two insertion sequences. *J. Bacteriol.*, **180**, 4834–4842.

75. Small,K., Iber,J. and Warren,S.T. (1997) Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nat. Genet.*, **16**, 96–99.

76. Croucher,N.J., Harris,S.R., Fraser,C., Quail,M.A., Burton,J., van der Linden,M., McGee,L., von Gottberg,A., Song,J.H., Ko,K.S., *et al.* (2011) Rapid pneumococcal evolution in response to clinical interventions. *Science*, **331**, 430–434.

77. van Tonder,A.J., Bray,J.E., Quirk,S.J., Haraldsson,G., Jolley,K.A., Maiden,M.C.J., Hoffmann,S., Bentley,S.D., Haraldsson,Á., Erlendsdóttir,H., *et al.* (2016) Putatively novel serotypes and the potential for reduced vaccine effectiveness: capsular locus diversity revealed among 5405 pneumococcal genomes. *Microb. Genomics*, **2**, 000090.

78. Jim,K.K., Engelen-Lee,J., van der Sar,A.M., Bitter,W., Brouwer,M.C., van der Ende,A., Veening,J.-W., van de Beek,D. and Vandenbroucke-Grauls,C.M.J.E. (2016) Infection of zebrafish embryos with live fluorescent *Streptococcus pneumoniae* as a real-time pneumococcal meningitis model. *J. Neuroinflammation*, **13**, 188.

79. Kjos,M., Aprianto,R., Fernandes,V.E., Andrew,P.W., van Strijp,J.A.G., Nijland,R. and Veening,J.-W. (2015) Bright fluorescent *Streptococcus pneumoniae* for live-cell imaging of host-pathogen interactions. *J. Bacteriol.*, **197**, 807–818.

80. Sorg,R.A., Lin,L., van Doorn,G.S., Sorg,M., Olson,J., Nizet,V. and Veening,J.-W. (2016) Collective resistance in microbial communities by intracellular antibiotic deactivation. *PLoS Biol.*, **14**, e2000631.

81. Aprianto,R., Slager,J., Holsappel,S. and Veening,J.-W. (2016) Time-resolved dual RNA-seq reveals extensive rewiring of lung epithelial and pneumococcal transcriptomes during early infection. *Genome Biol.*, **17**, 198.

82. Andersen,E.S., Rosenblad,M.A., Larsen,N., Westergaard,J.C., Burks,J., Wower,I.K., Wower,J., Gorodkin,J., Samuelsson,T. and Zwieb,C. (2006) The tmRDB and SRPDB resources. *Nucleic Acids Res.*, **34**, D163-168.

83. Geissmann,T., Chevalier,C., Cros,M.-J., Boisset,S., Fechter,P., Noirot,C., Schrenzel,J., François,P., Vandenesch,F., Gaspin,C., *et al.* (2009) A search for small noncoding RNAs in *Staphylococcus aureus* reveals a conserved sequence motif for regulation. *Nucleic Acids Res.*, **37**, 7239–7257.

84. Knutsen,E., Johnsborg,O., Quentin,Y., Claverys,J.-P. and Håvarstein,L.S. (2006) BOX elements modulate gene expression in *Streptococcus pneumoniae*: impact on the fine-tuning of competence development. *J. Bacteriol.*, **188**, 8307–8312.

85. Livny,J., Brencic,A., Lory,S. and Waldor,M.K. (2006) Identification of 17 *Pseudomonas aeruginosa* sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2. *Nucleic Acids Res.*, **34**, 3484–3493.

86. Täpp,J., Thollesson,M. and Herrmann,B. (2003) Phylogenetic relationships and genotyping of the genus Streptococcus by sequence determination of the RNase P RNA gene, *rnpB*. *Int. J. Syst. Evol. Microbiol.*, **53**, 1861–1871.

87. Weinberg,Z., Wang,J.X., Bogue,J., Yang,J., Corbino,K., Moy,R.H. and Breaker,R.R. (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol.*, **11**, R31.