Adding context to the pneumococcal core genes using bioinformatic analysis of the intergenic pangenome of Streptococcus pneumoniae

Nielsen, Flemming Damgaard; Møller-Jensen, Jakob; Jørgensen, Mikkel Girke

Published in:
Frontiers in Bioinformatics

DOI:
10.3389/fbinf.2023.1074212

Publication date:
2023

Document version:
Final published version

Document license:
CC BY

Citation for pulished version (APA):
Nielsen, F. D., Møller-Jensen, J., & Jørgensen, M. G. (2023). Adding context to the pneumococcal core genes using bioinformatic analysis of the intergenic pangenome of Streptococcus pneumoniae. *Frontiers in Bioinformatics*, *3*, Article 1074212. https://doi.org/10.3389/fbinf.2023.1074212

TYPE Original Research
PUBLISHED 08 February 2023
DOI 10.3389/fbinf.2023.1074212

OPEN ACCESS

EDITED BY
Vasco Ariston De Carvalho Azevedo,
Federal University of Minas Gerais, Brazil

REVIEWED BY
Mamoon Rashid,
(KAIMRC), Saudi Arabia
Olli-Pekka Smolander,
Tallinn University of Technology, Estonia

*CORRESPONDENCE
Mikkel Girke Jørgensen,
mikkelj@bmb.sdu.dk

SPECIALTY SECTION
This article was submitted to Genomic Analysis, a section of the journal Frontiers in Bioinformatics

RECEIVED 19 October 2022
ACCEPTED 24 January 2023
PUBLISHED 08 February 2023

CITATION
Nielsen FD, Møller-Jensen J and Jørgensen MG (2023), Adding context to the pneumococcal core genes using bioinformatic analysis of the intergenic pangenome of *Streptococcus pneumoniae*.
Front. Bioinform. 3:1074212.
doi: 10.3389/fbinf.2023.1074212

Adding context to the pneumococcal core genes using bioinformatic analysis of the intergenic pangenome of *Streptococcus pneumoniae*

Flemming Damgaard Nielsen[1,2], Jakob Møller-Jensen[1] and Mikkel Girke Jørgensen[1]*

[1]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense, Denmark, [2]Department of Clinical Microbiology, Odense University Hospital, Odense, Denmark

Introduction: Whole genome sequencing offers great opportunities for linking genotypes to phenotypes aiding in our understanding of human disease and bacterial pathogenicity. However, these analyses often overlook non-coding intergenic regions (IGRs). By disregarding the IGRs, crucial information is lost, as genes have little biological function without expression.

Methods/Results: In this study, we present the first complete pangenome of the important human pathogen *Streptococcus pneumoniae* (pneumococcus), spanning both the genes and IGRs. We show that the pneumococcus species retains a small core genome of IGRs that are present across all isolates. Gene expression is highly dependent on these core IGRs, and often several copies of these core IGRs are found across each genome. Core genes and core IGRs show a clear linkage as 81% of core genes are associated with core IGRs. Additionally, we identify a single IGR within the core genome that is always occupied by one of two highly distinct sequences, scattered across the phylogenetic tree.

Discussion: Their distribution indicates that this IGR is transferred between isolates through horizontal regulatory transfer independent of the flanking genes and that each type likely serves different regulatory roles depending on their genetic context.

KEYWORDS

genomics, pangenome, intergenic region, horizontal regulatory transfer, horizontal gene transfer, computational biology

Introduction

Streptococcus pneumonia (pneumococcus) is the leading cause of sepsis, meningitis and bacterial pneumoniae in children worldwide (O'Brien et al., 2009). Widespread antibiotic resistance and the emergence of non-vaccine serotypes is making treatment increasingly difficult. These threats have led the WHO to list pneumococcus as a "priority" pathogen (O'Brien et al., 2009; Weiser et al., 2018). This clinical relevance of pneumococcus has, in part, led to great scientific interest and the publication of several thousand sequenced genomes (National Center for Biotechnology Information, 2018).

The availability of whole genome sequence (WGS) data has made it possible to study the entire pangenome of an organism rather than single isolates. A pangenome consists of the collective gene pool present in a group of organisms belonging to the same clade (Tettelin et al., 2005). The pangenome can be divided into a core genome, which constitutes genes present in all isolates and the accessory genome as the remaining genes (Tettelin et al., 2005). The pangenome of pneumococcus is considered at the extreme end of being open, that is, there is no defined limit to its pangenome as new genes are acquired continuously (Donati et al., 2010). This openness is mainly due to new genes being acquired through horizontal gene transfer (HGT) mediated by pneumococcus' natural competence (Vos, 2009; Chaguza et al., 2015).

Traditionally, pangenomes are limited to genes thereby excluding the non-coding intergenic regions (IGRs) (Page et al., 2015; Xiao et al., 2015). This focus on genes alone leaves out 15% of the genomes and ignores a significant amount of crucial genomic information as IGRs contain several biologically relevant elements such as promoters, terminators, regulatory binding sites and non-coding RNAs (Koonin et al., 2001; Dagan et al., 2008; Peters et al., 2011; McCutcheon and Moran, 2012; Ochman and Caro-Quintero, 2016; Jørgensen et al., 2020). To effectively link genotypes to phenotypes through pangenomics, IGRs must be taken into consideration, as genes have little biological function without expression.

Recently, IGRs have attracted more attention as potential drivers of evolution (Molina and Van Nimwegen, 2008; Oren et al., 2014; Thorpe et al., 2017). They persist through purifying selection, also known as negative selection, where unused or unwanted traits are removed. This persistence is true across several diverse bacterial species, in a similar fashion to that of core genes, even when major regulatory elements are excluded (Molina and Van Nimwegen, 2008; Thorpe et al., 2017). Small variations in IGRs can lead to great phenotypical impact, for instance, the inversion of a single promoter element was demonstrated to turn a commensal bacterium pathogenic (Somvanshi et al., 2012).

IGRs may also undergo genetic recombination, a term coined *horizontal regulatory transfer* (HRT) (Ragan and Beiko, 2009; Matus-Garcia et al., 2012). HRT can occur with the flanking genes of the IGR, but in some cases, the IGRs are transferred independently of the genes they regulate (Oren et al., 2014). As much as 32% of the core regulatory regions in *E. coli* and 51% of the overall core IGRs are thought to have been acquired in this manner indicating that HRT is indeed common (Oren et al., 2014). Another aspect of HRT is regulatory switching where one IGR is replaced with another non-homologous IGR. This leads to two or more conserved IGRs occupying the same genomic space across different isolates of the same species (Ragan and Beiko, 2009; Matus-Garcia et al., 2012; Somvanshi et al., 2012; Oren et al., 2014; Thorpe et al., 2018). As much as 13% of the IGRs within the core genome of *E. coli* have undergone regulatory switching (Oren et al., 2014). Thus, IGRs seemingly contribute to greater variation in the core genome than genes themselves, thereby challenging the view of the bacterial core genome as being relatively stable (Oren et al., 2014; Caicedo-Montoya et al., 2021; Hyun et al., 2022).

In this study we map the complete core genome of pneumococcus and compare the nature of genes and IGRs against each other in the pangenome. We find a clear linkage between core genes and core IGRs, but core genes are associated with different IGRs, indicating that the pneumococcal core genome is less stable than previously thought. Additionally, we identify any potential regulatory switching events within this core genome. To our knowledge we are the first to identify the complete core genome of pneumococcus, both coding and non-coding.

Results

In this study, we map the first complete pangenome of pneumococcus, spanning both genes and IGRs. The identified intergenic core genome is provided in Supplementary Appendix SA1. Additionally, we screen for any regulatory switching events present within the core genome.

Many intergenic regions are universally conserved across all pneumococcal isolates

We created a pangenome of 84 different pneumococcal isolates, spanning both genes and the non-coding IGRs. To put the nature of the pneumococcal IGR pangenome into perspective, we performed the same analysis for *S. aureus*. Both species may colonize the human upper respiratory tract, both are opportunistic pathogens and both possess open pangenomes, making them prime candidates for comparison (Laux et al., 2019). The analysis shows that the otherwise non-coding IGRs of both species are conserved in a similar manner to genes across the pangenome, although the number of unique genes outnumber the number of unique IGRs in both species (Figure 1).

While the proportion of core genes roughly scales relative to the size of the genome (pneumococcus 2.1 Mbp/*S. aureus* 2.8 Mbp) the proportion of core IGRs relative to genome size is lower in pneumococcus (Figure 1). However, pneumococcus seemingly compensates for the lower number of unique IGRs by having multiple copies of several core IGRs in each genome. On average, each core IGR is present 1.23 times in each pneumococcal genome compared to 1.03 times in *S. aureus*. Each core gene is present 1.08 times in each pneumococcal genome and 1.02 times in *S. aureus*, this indicates that the high copy number of pneumococcal core IGRs is quite unusual.

Core genes and IGRs constitute the majority of each genome

The average pneumococcal genome has 79% of its genes as core genes and 66% of its IGRs as core IGRs. The average *S. aureus* genome is comparatively close to that observed in pneumococcus, here core genes constitute 79% and core IGRs 68% of each genome (Table 1).

Despite pneumococcus having fewer unique core IGRs relative to genome size than *S. aureus*, as stated earlier, their copy number is higher in each genome, thus the percentage of core IGRs per genome is roughly equivalent in the two species (Table 1). The higher copy number of core IGRs in pneumococcus is also illustrated by the fact that 669 unique core IGRs exist in the pneumococcal core genome (Figure 1) but on average each genome has 817 core IGRs (Table 1).

IGRs are more likely to be unique to a few isolates than genes

The number of unique IGRs in the pneumococcal pangenome increases with the number of isolates analyzed in a similar manner to the number of unique genes (Figure 2A). Overall, fewer unique IGRs are present in the pangenome than genes, part of this is due to the exclusion of IGRs of <30 bp in length, which are most often intraoperonic (Thorpe et al., 2018).

Most IGRs are either present in almost all pneumococcus isolates or unique to only a few, that is, they are either very common or very rare (Figure 2B). Pneumococcus genes show a similar distribution

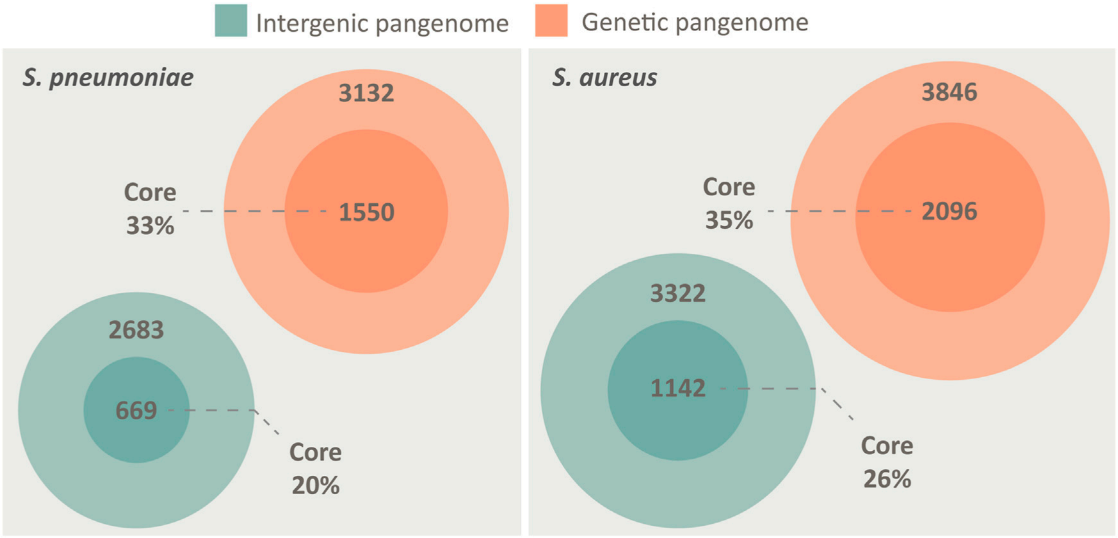

FIGURE 1
The pangenome of *S. pneumoniae* and *S. aureus*, spanning both intergenic regions (green) and genes (orange), illustrated by Venn diagrams. Both species possess a core genome of both IGRs and genes, defined as being present in >95% of isolates. The pangenomes are constructed from 84 unique genomes of each species. *S. pneumoniae* has a core genome of 1,550 genes and 669 IGRs, while an accessory genome of 3,132 genes and 2683 IGRs. *S. aureus* has a core genome of 2096 genes and 1,142 IGRs, while an accessory genome of 3,846 genes and 3,322 IGRs.

TABLE 1 The number of genes and IGRs in the core and accessory genome in selected genomes and across the collected pangenome, as well as the percentage of genes and IGRs that are core.

Species/isolate	Core genes	Core IGRs	Accessory genes	Accessory IGRs	Percentage core genes pr. genome (%)	Percentage core IGRs pr. genome (%)
S. pneumoniae (Species average)	1,670	817	448	425	78.93	65.83
S. pneumoniae D39	1,672	810	352	388	82.61	67.61
S. pneumoniae R6	1,674	809	345	388	82.91	67.59
S. pneumoniae Tigr4	1703	820	447	447	79.21	64.72
S. aureus (Species average)	2131	1,170	570	544	78.98	68.28

across the pangenome, though a larger proportion of IGRs are confined to only a few isolates than genes.

Pneumococcus retains more unique genes than IGRs within its pangenome (Figure 2A), and most unique IGRs are only found in single isolates, making them rare (Figure 2B). This scarcity of unique IGRs could indicate that IGRs experience a higher evolutionary selection threshold than genes, thereby lowering the likelihood of a newly acquired IGR of spreading to more isolates through HRT.

Double regulatory regions are more common in the core genome

IGRs can be categorized according to the orientation of their flanking genes. IGRs that are downstream of two convergently transcribed genes are considered non-regulatory (NR), IGRs that are upstream one gene and downstream another gene are considered single regulatory (SR) and IGRs that are between two divergently transcribed genes are considered double regulatory (DR) (Figure 3).

Looking at the distribution of the IGR types across the pneumococcal pangenome, NR and DR regions are rare compared to SR IGRs (Figure 3). DR regions also constitute a greater relative proportion of the core IGRs than seen in the accessory genome.

Core IGRs are linked to core genes

Next, we analyzed the degree of linkage between core IGRs and core genes, that is, how often a core IGR is directly upstream a core gene. IGRs and their flanking genes were identified and any IGRs directly upstream the start codon of a gene was selected. The status of the IGR/gene pairs as accessory or core genome was then assessed and the ratio of each combination calculated. On average 81% of core genes

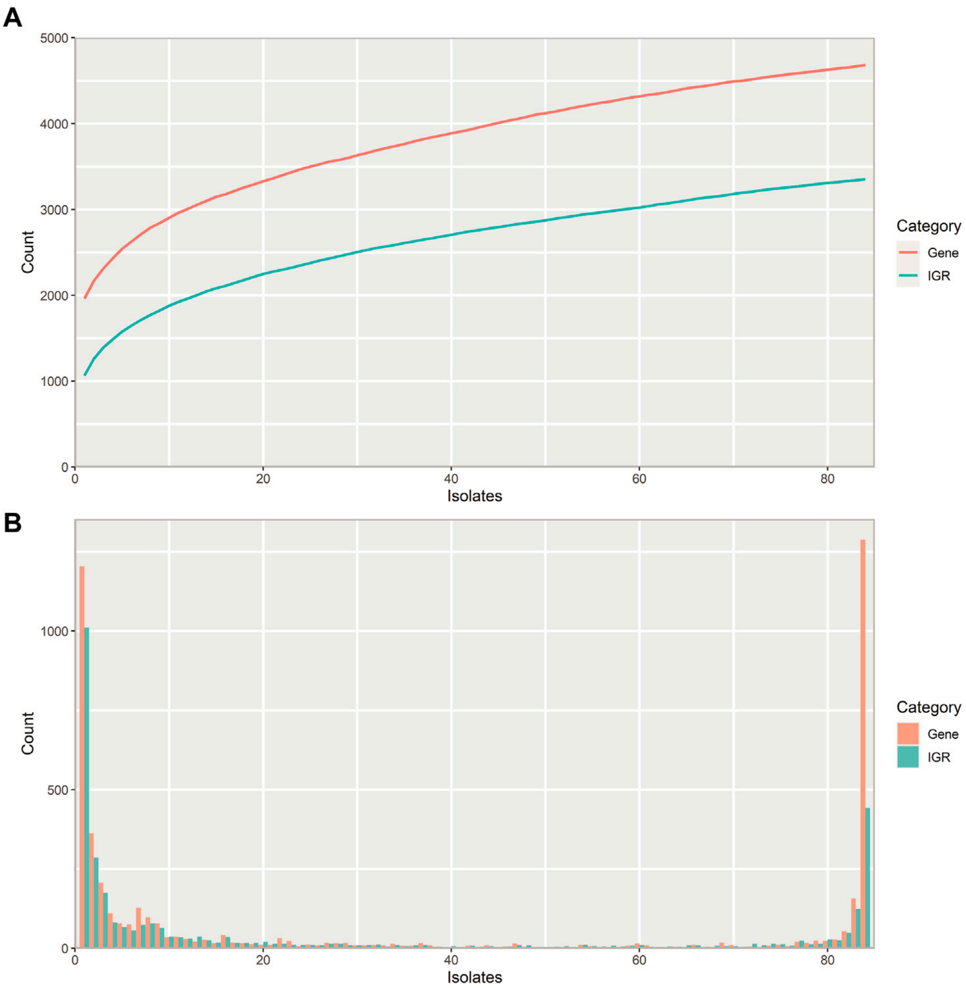

FIGURE 2
Properties of the pneumococcal pangenome and its intergenic regions (IGRs) **(A)** Number of unique intergenic regions (green) and genes (orange) as a function of the number of isolates included in the pangenome. **(B)** Distribution of unique IGRs (green) and genes (orange) across the streptococcal pangenome, illustrated with a frequency histogram (number of IGRs/genes present in the given number of isolates). Most IGRs and genes are part of the core genome or confined to a small fraction of the isolates.

in *S. pneumoniae* are associated with a core IGR, whereas only 74% of accessory genes are linked to accessory IGRs (Table 2). For comparison, the linkage of core genes to core IGRs is greater in *S. aureus* at 86%, and accessory IGRs are flanking accessory genes 82% of the time.

None of the IGRs of the capsular polysaccharide synthesis (cps) operon were found to be core IGRs, however the highly conserved flanking genes *dexB* and *aliA* were both associated with core IGRs (Appendix 1).

A single core IGR shows sign of regulatory switching

Next, we examined the IGR candidates for regulatory switching. Regulatory switching describes when one IGR is replaced by a different non-homologue IGR. The origin of these switched IGRs is not inferred in this analysis, thus they can both originate from within the isolate itself or even from a separate species. For this analysis, only switches where the IGRs share no significant sequence homology with a BLASTN were included.

We detected three switches within the pneumococcus pangenome and only one of these is flanked by core genes. We designated the core switched IGR as csIGR (Table 3). While the two versions of the csIGR are highly conserved on their own, with both having a nucleotide identity of >99% amongst themselves, aligning the two versions with each other results in an insignificant nucleotide identity of 57%. These results were manually confirmed with a blastn and confirmed that all pneumococcal isolates always have one of these two csIGRs but only in a single copy and always between the same flanking genes.

The flanking genes were both single copy core genes and were identified in the common lab strains *S. pneumoniae* D39 and TIGR4 (Figure 4). These two strains have distinct csIGR types, with D39 having csIGR1 and TIGR4 having csIGR2 (Figure 4). Interestingly, rather than flanking an operon, the IGRs are predicted to sit in the middle of an operon. Little is known about the flanking genes, other than their status as single copy core genes

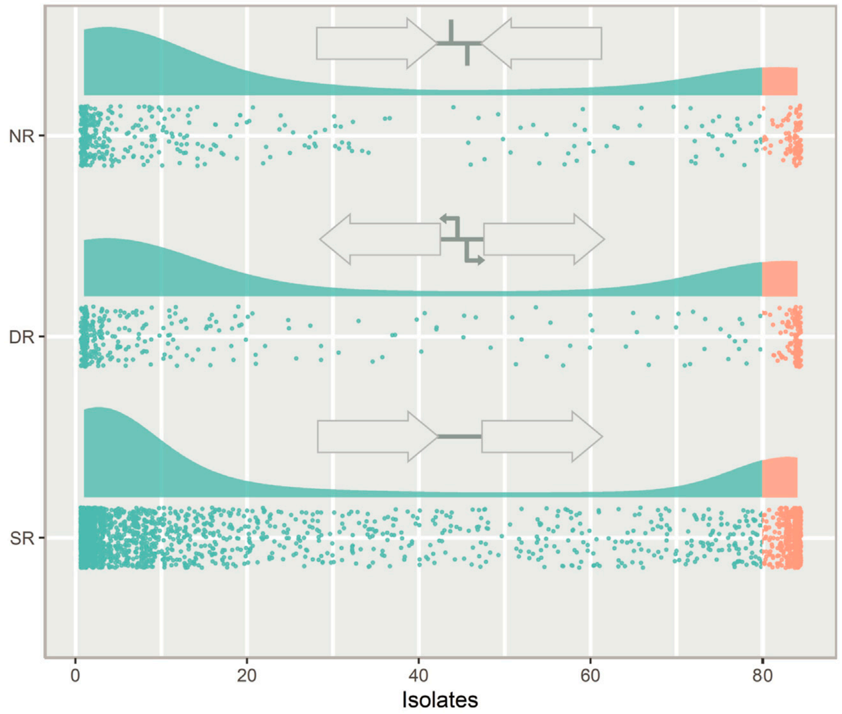

FIGURE 3
The types of intergenic regions (IGRs) and their distribution in the pneumococcal pangenome, illustrated with a raincloud plot. Each point is a unique IGR of that type plotted against the number of isolates in the pangenome it is present in. Cloud areas are scaled relative to the size of each dataset. Core IGRs are present in >95% strains (orange) and accessory IGRs are present in <95% of isolates (green). The IGRs are categorized according to the orientation of their flanking genes. If the flanking genes are pointing in the same direction the IGR is categorized as single regulatory (SR), if they face towards the IGR, it is categorized as non-regulatory (NR) and if they face away from the IGR, it is categorized as double regulatory (DR).

TABLE 2 Genes and their upstream IGR was analyzed for their distribution in the pangenome. Listed are the percentage core and/or genes with a core and/or accessory IGR immediately upstream.

Species/isolate	Core gene: core IGR (%)	Core gene: accessory IGR (%)	Accessory gene: core IGR (%)	Accessory gene: accessory IGR (%)
S. pneumoniae (Species average)	80.92	19.08	26.45	73.55
S. pneumoniae D39	81.06	18.94	30.77	69.23
S. pneumoniae R6	80.97	19.03	30.39	69.61
S. pneumoniae Tigr4	80.38	19.62	26.52	73.48
S.aureus (Species average)	86.19	13.81	17.62	82.38

TABLE 3 The two versions of the csIGR in-between the single copy core genes. csIGR1 is present in 32 of the isolates analyzed and is highly conserved across the genomes with an average nucleotide identity of 99.49%. csIGR2 is present in 52 of the strains analyzed and is likewise highly conserved with an average nucleotide identity of 99.28%. Both IGRs have roughly the same length in base pairs.

	Length	SNPs	Nuc_identity (%)	Length_identity (%)	No.isolates
csIGR1	215	1	99.49	99.53	32
csIGR2	214	1	99.28	99.05	52

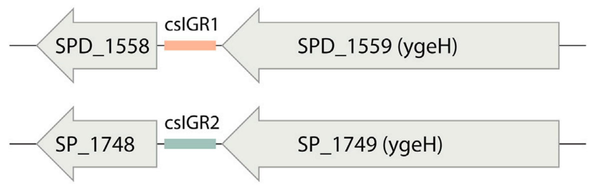

FIGURE 4
The core switched intergenic region (csIGR) in *S. pneumoniae* D39 and Tigr4. Each type of csIGR is represented in these strains, with D39 having csIGR1 (orange) and TIGR4 having csIGR2 (green). In D39, csIGR1 is flanked by SPD_1558 and SPD_1559. In TIGR4, csIGR2 is flanked by the genes SP_1748 and SP_1749.

found in this study. SPD_1559/SP_1749 is considered essential in pneumococcus and is a homologue to *ygeH*, a gene involved in biogenesis of the 30 S ribosome subunit (Pek et al., 2007; Liu et al., 2017). The sequence of both csIGR types is provided in Supplementary Appendix SA2. Interestingly, neither of the csIGR types were confined to a specific phylogenetic cluster of pneumococci (Figure 5). The fact that the csIGR types are spread across the phylogenetic tree indicates that their distribution is due to HRT.

We performed a pangenome wide association study to see if any genes within the accessory genome were significantly co-occurring with the csIGR alleles across the pangenome. However, no genes were exclusively associated with neither of the csIGR types. Both sequences were also screened for promoters, riboswitches and homology to known regulatory RNAs with no significant hits. However, the translated RNA sequence of both sequences was predicted to form significant secondary structures, the significance of which remains to be elucidated. The predicted secondary structures are provided in Supplementary Appendix SA2.

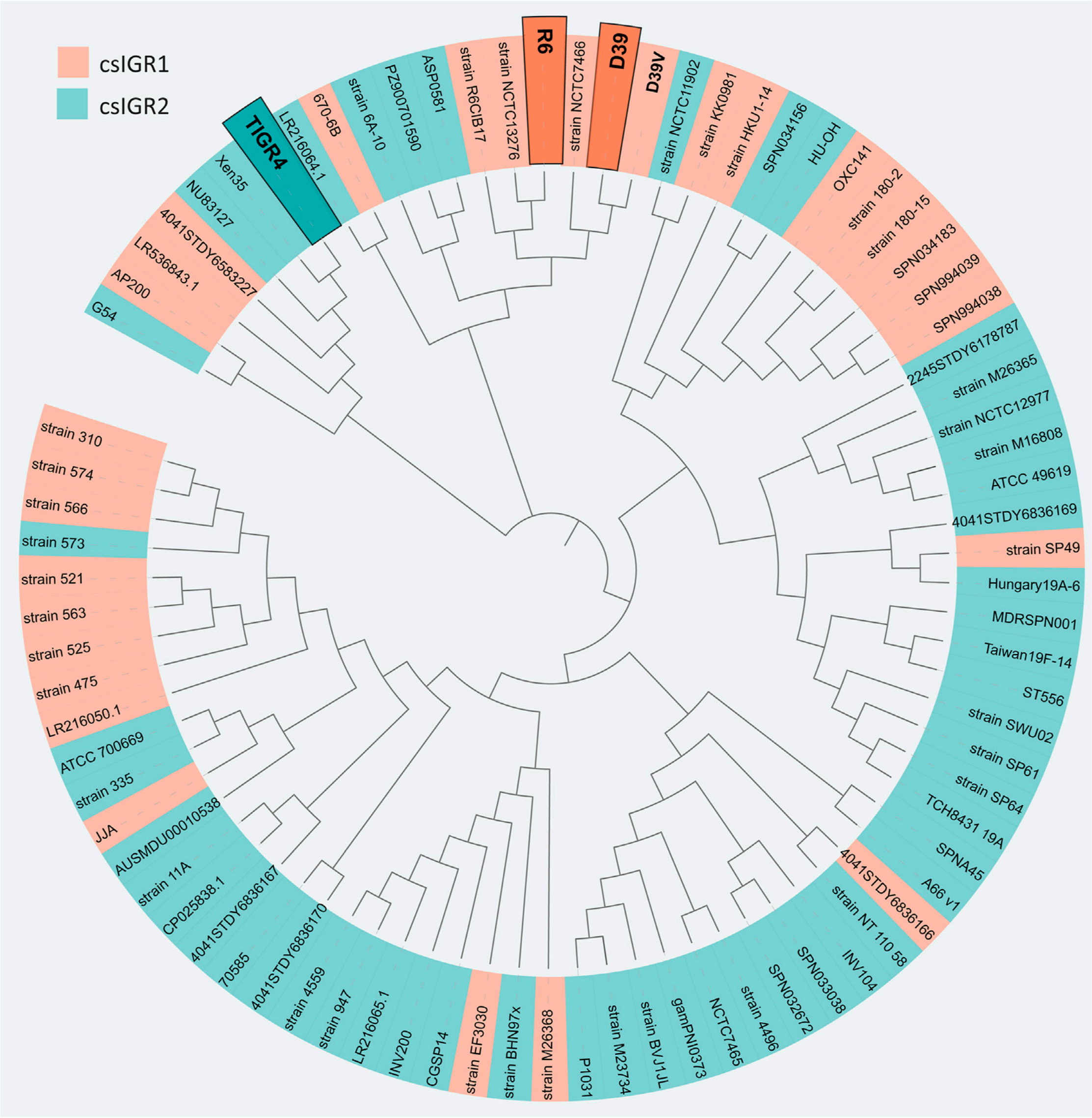

FIGURE 5
Unrooted phylogenetic tree of the 84 *S. pneumoniae* strains used in this study. The tree is based on SNPs in the core genes. The shading of each label indicates the presence of csIGR1 (orange) or csIGR2 (green). The tree was created using Roary, Fasttree and iTol. The *S. pneumoniae* strain names are specified if applicable, if no clear strain name was given the sequence ID was used.

Discussion

Here we present the first complete pangenome of pneumococcus, spanning both genes and the non-coding IGRs. A small but conserved IGR core genome in pneumococcus was identified. We find that the pneumococcal core genome consists of 1,550 unique genes and 669 IGRs, whereas the accessory genome consists of 3,132 unique genes and 2683 IGRs. The number of unique genes surpasses that of unique IGR in both cases, this is unsurprising as most intraoperonic regions in pneumococcus are less than 30 bp in length and are therefore disregarded in the analysis. This also means that most IGRs identified are associated with the flanking genes of operons i.e., the regulatory regions.

IGRs between two divergently transcribed genes are termed double regulatory (DR). These regions constituted a greater relative part of the core genome than the accessory genome. It is likely because meaningful regulation of two genes is harder to achieve than regulation of single genes, raising the selection threshold for the emergence of beneficial divergence. This increased selection pressure has previously been observed as purifying selection has been shown to be more prominent in DR regions than the other types (Molina and Van Nimwegen, 2008).

Our analysis reveals that IGRs are highly conserved in pneumococcus. On average, 66% of IGRs in any isolate is shared with all other isolates and 79% of genes in any isolate is shared with all other isolates. A similar trend is seen in *S. aureus*, however, the overall number of unique IGRs is lower in pneumococcus relative to genome size. Instead, our analysis reveals that pneumococcus has several duplicates of some core IGRs across the genome, with core IGRs on average being present 1.25 times in each isolate. This trend is not seen with its core genes and is not observed in neither the core genes nor core IGRs of *S. aureus*. This suggests that pneumococcus is more rigid with its transcriptional profile as the same regulatory regions might be repeated to a greater degree than observed in *S. aureus*.

We identify a clear linkage between the core IGRs and core genes in pneumococcus, on average 81% of core IGRs are directly upstream of a core gene. This indicates that the transcriptional regulation of the core genome in pneumococcus is mostly conserved across all isolates, but to a lesser degree than seen in *S. aureus*. However, this leaves 19% of core genes being associated with accessory IGRs, indicating some plasticity to the core genome that is otherwise viewed as stable. The greatest difference seen between the two species in this regard is that core IGRs are more often associated with accessory genes in pneumococcus. This might be explained by pneumococcus retaining multiple copies of some core IGRs, making them associated with both core and accessory genes, though this remains to the investigated.

Surprisingly, only three switches were detected in pneumococcus and only one of these was flanked by core genes. In another study, the same analysis was done on a collection of *E. coli* genomes and 61 switches were detected (Thorpe et al., 2018). This indicates that regulatory switching does not play a major role in pneumococcal disease. It is possible that regulatory switching is more prominent in *E. coli* as it inhabits a great number of different niches compared to pneumococcus (Tenaillon et al., 2010; Weiser et al., 2018).Our results show that the pneumococcal core genome is less stable than previously thought. While there is indeed a stable reservoir of highly conserved core genes, their flanking IGRs, which contain most of the regulatory regions responsible for controlling the transcription of these core genes show greater plasticity. We believe that future studies will benefit from viewing the genes as a "package" with their upstream IGR, as even core genes maintain different regulatory regions within the pneumococcal species.

Materials and methods

Genomes

All 84 complete *S. pneumoniae* genomes available from the National Center for Biotechnology Information, GenBank resource was downloaded in raw FASTA format. Additionally, 84 randomly selected *S. aureus* genomes were retrieved for comparison with *S. pneumoniae* (12/5/2021). Genomes were then annotated with Prokka (v 1.14.5), using the standard parameters of the software (Seemann, 2014). The genomes used are listed in Supplementary Appendix SA3.

Pangenome creation

Initially a pangenome of the coding sequences (CDS) was created using Roary (v3.13.0) (Page et al., 2015). Then a complementary pangenome of the IGRs was created using Piggy (v1.5), an intergenic pangenome analysis tool that emulates Roary (Thorpe et al., 2018). Some steps were taken to ensure comparability between the outputs of the software. Roary was set to cluster CDSs with -e -n (to perform alignments using MAFFT (Katoh et al., 2002)), -i 90 (90% sequence identity cut-off) and -s (to not split paralogs into separate clusters). The settings for running Piggy were set at the standard parameters of the software, except for -len_id 10 (the minimum percentage of length identity to form a cluster). The length identity was reduced for comparability with Roary, as gene clusters generated by Roary only require a sequence length identity of 120 bp for clustering CDSs, thus the len_id of 10 is recommended by the creators of Piggy for Roary consistency as IGRs are not erroneously placed into separate clusters (Koonin et al., 2001; Dagan et al., 2008; Molina and Van Nimwegen, 2008; Ragan and Beiko, 2009; Peters et al., 2011; Matus-Garcia et al., 2012; McCutcheon and Moran, 2012; Somvanshi et al., 2012; Oren et al., 2014; Page et al., 2015; Xiao et al., 2015; Ochman and Caro-Quintero, 2016; Thorpe et al., 2017; Thorpe et al., 2018; Jørgensen et al., 2020). The randomly assigned locus tags provided by Prokka were translated when necessary, by aligning the GFF files of the Genbank annotated files and Prokka output.

Core gene and core IGR linkage analysis

The linkage of core genes and core IGRs was quantified using R (v. 4.1.0). The gene_presence_absence file from Roary and the IGR_presence_absence file from Piggy was loaded as dataframes in R. For each gene and IGR cluster in the files their status as a core or accessory gene was identified and assigned. For each genome all IGRs were paired with their upstream gene. Thus, NR regions were removed from the dataset and both flanking genes for DR regions were analyzed separately, if any of the two genes were core, the DR IGR was assigned as flanking a core gene. The R code is provided in Supplementary Appendix SA4.

Switched intergenic regions analysis

For identification of switched IGRs, a separate analysis using Piggy was performed with -len_id 90 (the minimum percentage of length identity to form a cluster). This was done to perform a more strict analysis of the IGRs, as the higher threshold for forming a cluster ensured that homologue IGRs were not identified as switched IGRs (Thorpe et al., 2018). IGR switches were identified using the "gene-pair" method of Piggy, here two or more different IGR sequences that occupy the same space between a specific gene pair are analyzed. The candidate IGR sequences are then aligned with BLASTN with low complexity filtering turned off and if there are no significant matches between the IGR they are identified as "switched". If there is a significant match Piggy aligns the sequences using MAFFT and provides the nucleotide identity of the alignments.

The identified switch was validated manually with a BLASTN against all the genomes (data not shown).

Phylogenetic analysis

A phylogenetic tree of the strains included in this study was created, based on single nucleotide polymorphisms (SNPs) in the core genes. Roary was run separately with the same settings as previously mentioned with the exception of -e (Core gene alignment with PRANK) (Page et al., 2015). This produced a highly accurate alignment of the core genes within the pangenome. FastTree (v2.1.11) was then run to infer an approximately-maximum-likelihood phylogenetic tree based on SNPs within the core genes (Price et al., 2009). The resulting newick file was then visualized using iTol (v5.7) and exported to Adobe Illustrator (Letunic and Bork, 2007).

Pangenome wide association study

To identify whether any accessory genes were significantly associated with any of the csIGR alleles, a pangenome-wide association study was performed using Scoary (v1.6.16) (Brynildsrud et al., 2016). A trait matrix was created as an input for Scoary, indicating which of the two csIGRs alleles were present in which genomes. Scoary then sorted the accessory genome provided by the gene_presence_absence file from Roary, scoring each accessory gene according to their co-occurrence with each csIGR.

csIGR1 and csIGR2 analysis

To assess homology to existing regulatory RNAs, a BLASTN was performed for each csIGR against the RefSeq RNA database. Both sequences were screened for potential riboswitches using Riboswitch Finder (Bengert and Dandekar, 2004). Any potential promoter or terminator regions were screened for using BPROM and FindTerm (Softberry) (Solovyev et al., 2011). Secondary structures were predicted for both sequences using the RNA structure package available through Mathews lab, at standard parameters (Reuter and Mathews, 2010).

Data availability statement

Publicly available datasets were analyzed in this study. In total, 84 different pneumococcal genomes and 84 randomly selected *S. aureus* genomes were retrieved from the National Center for Biotechnology Information. Their accession numbers are stated in Supplementary Appendix SA3.

Author contributions

FN, MJ and JM-J conceptualized the study. FN and MJ wrote the paper. Project supervised and funded by JM-J and MJ.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2023.1074212/full#supplementary-material

SUPPLEMENTARY TABLE S1
Streptococcus pneumoniae intergenic core genome.

SUPPLEMENTARY TABLE S2
Sequences and predicted secondary structures of csIGR1 and csIGR2.

SUPPLEMENTARY TABLE S3
Genomes used in this study.

SUPPLEMENTARY TABLE S4
Code.

References

Bengert, P., and Dandekar, T. (2004). Riboswitch finder--a tool for identification of riboswitch RNAs. *Nucleic Acids Res.* 32, W154–W159. doi:10.1093/NAR/GKH352

Brynildsrud, O., Bohlin, J., Scheffer, L., and Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* 17 (1), 238. doi:10.1186/s13059-016-1108-8

Caicedo-Montoya, C., Manzo-Ruiz, M., and Ríos-Estepa, R. (2021). Pan-genome of the genus streptomyces and prioritization of biosynthetic gene clusters with potential to produce antibiotic compounds. *Front. Microbiol.* 12, 677558. doi:10.3389/FMICB.2021.677558

Chaguza, C., Cornick, J. E., and Everett, D. B. (2015). Mechanisms and impact of genetic recombination in the evolution of *Streptococcus pneumoniae*. *Comput. Struct. Biotechnol. J.* 13, 241–247. doi:10.1016/j.csbj.2015.03.007

Dagan, T., Artzy-Randrup, Y., and Martin, W. (2008). Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 105, 10039–10044. doi:10.1073/pnas.0800679105

Donati, C., Hiller, N. L., Tettelin, H., Muzzi, A., Croucher, N. J., Angiuoli, S. V., et al. (2010). Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biol.* 11, R107. doi:10.1186/gb-2010-11-10-r107

Hyun, J. C., Monk, J. M., and Palsson, B. O. (2022). Comparative pangenomics: Analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics* 23, 7. doi:10.1186/S12864-021-08223-8

Jørgensen, M. G., Pettersen, J. S., and Kallipolitis, B. H. (2020). sRNA-mediated control in bacteria: An increasing diversity of regulatory mechanisms. *Biochimica Biophysica Acta - Gene Regul. Mech.* 1863, 194504. doi:10.1016/j.bbagrm.2020.194504

Katoh, K., Misawa, K., Kuma, K. I., and Miyata, T. (2002). Mafft: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi:10.1093/nar/gkf436

Koonin, E. V., Makarova, K. S., and Aravind, L. (2001). Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu. Rev. Microbiol.* 55, 709–742. doi:10.1146/annurev.micro.55.1.709

Laux, C., Peschel, A., and Krismer, B. (2019). *Staphylococcus aureus* colonization of the human nose and interaction with other microbiome members. *Microbiol. Spectr.* 7. doi:10.1128/microbiolspec.gpp3-0029-2018

Letunic, I., and Bork, P. (2007), Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* 23 (1), 127–128. doi:10.1093/bioinformatics/btl529

Liu, X., Gallay, C., Kjos, M., Domenech, A., Slager, J., Kessel, S. P., et al. (2017). High-throughput CRISPRi phenotyping identifies new essential genes in Streptococcus pneumoniae. *Mol. Syst. Biol.* 13, 931. doi:10.15252/msb.20167449

Matus-Garcia, M., Nijveen, H., and Van Passel, M. W. J. (2012). Promoter propagation in prokaryotes. *Nucleic Acids Res.* 40, 10032–10040. doi:10.1093/nar/gks787

McCutcheon, J. P., and Moran, N. A. (2012). Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* 10, 13–26. doi:10.1038/nrmicro2670

Molina, N., and Van Nimwegen, E. (2008). Universal patterns of purifying selection at noncoding positions in bacteria. *Genome Res.* 18, 148–160. doi:10.1101/gr.6759507

National Center for Biotechnology Information, (2018).“GenBank and WGS statistics,” Available at: https://www.ncbi.nlm.nih.gov/genbank/statistics/ (accessed Dec. 13, 2018).

O'Brien, K. L., Wolfson, L. J., Watt, J. P., Henkle, E., Deloria-Knoll, M., McCall, N., et al. (2009). Burden of disease caused by Streptococcus pneumoniae in children younger than 5 years: Global estimates. *Lancet* 374, 893–902. doi:10.1016/S0140-6736(09)61204-6

Ochman, H., and Caro-Quintero, A. (2016). “Genome size and structure, bacterial,” in *Encyclopedia of evolutionary biology* (Amsterdam: Elsevier).

Oren, Y., Smith, M. B., Johns, N. I., Kaplan Zeevi, M., Biran, D., Ron, E. Z., et al. (2014). Transfer of noncoding DNA drives regulatory rewiring in Bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 111, 16112–16117. doi:10.1073/pnas.1413272111

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., et al. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31 (22), 3691–3693. doi:10.1093/bioinformatics/btv421

Pek, C. L., Morimoto, T., Matsuo, Y., Oshima, T., and Ogasawara, N. (2007). The GTP-binding protein YqeH participates in biogenesis of the 30S ribosome subunit in Bacillus subtilis. *Genes Genet. Syst.* 82, 281–289. doi:10.1266/ggs.82.281

Peters, J. M., Vangeloff, A. D., and Landick, R. (2011). Bacterial transcription terminators: The RNA 3′-end chronicles. *J. Mol. Biol.* 412, 793–813. doi:10.1016/j.jmb.2011.03.036

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 26, 1641–1650. doi:10.1093/molbev/msp077

Ragan, M. A., and Beiko, R. G. (2009). Lateral genetic transfer: Open issues. *Philos. Trans. R. Soc. B Biol. Sci.* 364, 2241–2251. doi:10.1098/rstb.2009.0031

Reuter, J. S., and Mathews, D. H. (2010). RNAstructure: Software for RNA secondary structure prediction and analysis. *BMC Bioinforma.* 11, 129. doi:10.1186/1471-2105-11-129

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30 (14), 2068–2069. doi:10.1093/bioinformatics/btu153

Solovyev, V., and Salamov, A. (2011). “Automatic annotation of microbial genomes and metagenomic sequences,” in *Metagenomics and its applications in agriculture, biomedicine and environmental studies*. Editor R. W. Li (New York: Nova Science Publishers), 61–78.

Somvanshi, V. S., Sloup, R. E., Crawford, J. M., Martin, A. R., Heidt, A. J., Kim, K. s., et al. (2012). A single promoter inversion switches photorhabdus between pathogenic and mutualistic states. *Science* 80, 88–93. doi:10.1126/science.1216641

Tenaillon, O., Skurnik, D., Picard, B., and Denamur, E. (2010). The population genetics of commensal *Escherichia coli*. *Nat. Rev. Microbiol.* 8, 207–217. doi:10.1038/nrmicro2298

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial ‘pan-genome. *Proc. Natl. Acad. Sci. U. S. A.* 102, 13950–13955. doi:10.1073/pnas.0506758102

Thorpe, H. A., Bayliss, S. C., Hurst, L. D., and Feil, E. J. (2017). Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics*. doi:10.1534/genetics.116.195784

Thorpe, H. A., Bayliss, S. C., Sheppard, S. K., and Feil, E. J. (2018). Piggy: A rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. *Gigascience*. doi:10.1093/gigascience/giy015

Vos, M. (2009). Why do bacteria engage in homologous recombination? *Trends Microbiol.* 17, 226–232. doi:10.1016/j.tim.2009.03.001

Weiser, J. N., Ferreira, D. M., and Paton, J. C. (2018). Streptococcus pneumoniae: Transmission, colonization and invasion. *Nat. Rev. Microbiol.* 16 (6), 355–367. doi:10.1038/s41579-018-0001-8

Xiao, J., Zhang, Z., Wu, J., and Yu, J. (2015). A brief review of software tools for pangenomics. *Genomics, Proteomics Bioinforma.* 13 (1), 73–76. doi:10.1016/j.gpb.2015.01.007