The variome of pneumococcal virulence factors and regulators

Gustavo Gámez1,2,3,4*, Andrés Castro1,2†, Alejandro Gómez-Mejia1,3, Mauricio Gallego1,2, Alejandro Bedoya1,2, Mauricio Camargo1 and Sven Hammerschmidt3

Abstract

Background: In recent years, the idea of a highly immunogenic protein-based vaccine to combat Streptococcus pneumoniae and its severe invasive infectious diseases has gained considerable interest. However, the target proteins to be included in a vaccine formulation have to accomplish several genetic and immunological characteristics, (such as conservation, distribution, immunogenicity and protective effect), in order to ensure its suitability and effectiveness. This study aimed to get comprehensive insights into the genomic organization, population distribution and genetic conservation of all pneumococcal surface-exposed proteins, genetic regulators and other virulence factors, whose important function and role in pathogenesis has been demonstrated or hypothesized.

Results: After retrieving the complete set of DNA and protein sequences reported in the databases GenBank, KEGG, VFDB, P2CS and Uniprot for pneumococcal strains whose genomes have been fully sequenced and annotated, a comprehensive bioinformatic analysis and systematic comparison has been performed for each virulence factor, standalone regulator and two-component regulatory system (TCS) encoded in the pan-genome of S. pneumoniae. A total of 25 S. pneumoniae strains, representing different pneumococcal phylogenetic lineages and serotypes, were considered. A set of 92 different genes and proteins were identified, classified and studied to construct a pan-genomic variability map (variome) for S. pneumoniae. Both, pneumococcal virulence factors and regulatory genes, were well-distributed in the pneumococcal genome and exhibited a conserved feature of genome organization, where replication and transcription are co-oriented. The analysis of the population distribution for each gene and protein showed that 49 of them are part of the core genome in pneumococci, while 43 belong to the accessory-genome. Estimating the genetic variability revealed that pneumolysin, enolase and Usp45 (SP_2216 in S. p. TIGR4) are the pneumococcal virulence factors with the highest conservation, while TCS08, TCS05, and TCS02 represent the most conserved pneumococcal genetic regulators.

Conclusions: The results identified well-distributed and highly conserved pneumococcal virulence factors as well as regulators, representing promising candidates for a new generation of serotype-independent protein-based vaccine(s) to combat pneumococcal infections.

Keywords: Variome, Virulence factors, Two component systems, Streptococcus Pneumoniae

* Correspondence: gustavo.gamez@udea.edu.co
† Equal contributors
1 Genetics, Regeneration and Cancer Research Group (GRC), University Research Centre (SIU), Universidad de Antioquia (UdeA), Calle 70 # 52 - 21, 050010 Medellín, Colombia
2 Basic and Applied Microbiology Research Group (MICROBA), School of Microbiology, Universidad de Antioquia (UdeA), Calle 70 # 52 - 21, 050010 Medellín, Colombia
Full list of author information is available at the end of the article

© The Author(s). 2018 Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
**Background**

*Streptococcus pneumoniae*, also known as the pneumococcus, is a Gram-positive, α-hemolytic and facultative aerobic bacterium. This microorganism is normally found as a harmless commensal in the upper respiratory tract of humans. Pneumococci have a great epidemiological importance due to their high impact on public health, causing more than one and a half million of deaths per year around the world [1]. *S. pneumoniae* is the main etiologic agent of community-acquired pneumonia. However, this is not its only clinical manifestation, because other kind of diseases such as otitis media, sinusitis, septicemia and meningitis are also caused by this pathogen and associated with high mortality rates [2].

Given the particular biochemical and molecular features of *Streptococcus pneumoniae* (Gram-positive, catalase-negative, optochin-sensitive and bile-soluble bacteria), its identification process in the laboratory is relatively simple. Nevertheless, the great molecular, biochemical and immunological diversity of its capsule and other antigens such as choline-binding proteins make them one of the hardest bacterial pathogens to face because of its variability [3, 4]. The “Quelling Reaction”, developed over 100 years ago by Neufeld, allows the specific and reliable identification of each one of the >94 serotypes that have been discovered up to date. The capsular polysaccharide is the sine qua non virulence factor, however the pathogenic potential of serotypes may vary and similarly, the frequencies or prevalence varies from one geographic region to the other [5]. Despite this, the capsule is not the only factor required to induce disease by *S. pneumoniae*. In fact, the surface of the pneumococcus is decorated by various proteins, which have been already associated with its high pathogenic potential. In addition, their interaction level with the host cellular receptors has been proved, exhibiting crucial pathogenic functions such as adhesion, colonization, breaching tissue barriers and immune evasion [6].

An important group of regulatory proteins of great interest are the histidine kinases (HK), located in the bacterial surface and functioning as the sensors of two-component regulatory systems (TCS). The sensing of environmental signals via TCS, regulates the genetic expression of cellular processes that are of great importance such as natural competence, antibiotic resistance, adaptation to different environmental situations, surface proteins expression, and others [7, 8]. In general, TCS are composed of a histidine kinase, a membrane protein sensing the extracellular signals and transmitting these signals to a cytoplasmatic regulator/effect protein referred to as response regulator (RR). This happens via the HK autophosphorylation and a subsequent trans-phosphorylation process. In *Streptococcus pneumoniae*, 13 TCS and one orphan RR have been identified [7].

The relevance of the cellular, physiological and pathogenic functions that these pneumococcal proteins fulfill, have aroused a great scientific and biotechnological interest, given their potential pharmaceutical applications as vaccine candidates [9]. Nowadays, the antibiotic treatment of the infections caused by the pneumococcus is often complicated due to the increase of antibiotic resistance [10]. Furthermore, prevention by the use of the pneumococcal polysaccharide vaccines and/or pneumococcal conjugated vaccines only helps to control the disease caused by some of the serotypes and has an indirect impact on colonization [9]. Thus, there is an urge to define more global and effective strategies for the treatment and/or prevention, and to fight the pneumococcus and its local and invasive diseases. Consequently, the idea of a protein-based vaccine has taken great importance in the last years. However, in order to be considered or included in a recombinant vaccine formulation, a bacterial protein has to fulfill specific criteria such as: (1) playing an important role in the bacterial fitness and/or pathogenesis of *S. pneumoniae*, (2) possessing a wide distribution among the circulating strains and clinical isolates, (3) exhibiting a major conservation at its genetic and protein sequence, (4) being immunogenic, (5) demonstrating protective activity in experimental assays, and (6) having favorable physicochemical properties for expression and purification of its recombinant products.

*Streptococcus pneumoniae* is a pathogen exhibiting a fratricide behavior and an enormous capacity for natural competence, acquiring foreign genetic material and integrating it into its genome [11]. These processes, in addition to the mutation rates [12, 13], greatly stimulate the horizontal gene transfer with other microorganisms, and explains pneumococcal genetic variability and genome plasticity [14, 15]. This model of pneumococcal population evolution, where recombination highly out-passes mutation, is also caused by the relatively high numbers of repetitive sequences in the genome thereby facilitating the incorporation of foreign DNA in the chromosome [15–18]. In consequence, these events contribute to structural reorganizations, and influence the presence or absence of protein-encoding genes in different subsets of the global pneumococcal population, making them highly heterogeneous from the core- and pan-genomic point of view [15]. Likewise, the generation and fixation of particular changes in the genome affect the mutation rates, which in turn influence the evolution and conservation of genes and contribute to adaptive changes that potentially lead to an increased virulence and a more complex interaction with the host [19].

Due to these molecular events and their importance, there is a need to fully and globally understand the genetic heterogeneity and variability among the different pneumococcal strains/serotypes (variome), and to get a
deeper and detailed molecular understanding of the different physiological and pathogenic mechanisms that this microorganism uses to cause severe and life-threatening diseases. Definitely, obtaining this knowledge will allow to identify potential pharmaceutical targets for new antimicrobial therapies. By the recognition of their conservation and distribution degree among pneumococcal strains, this will confirm protein candidates for vaccines. However, despite the availability of a high number of completely sequenced genomes and the importance to analyse the genetic differences among pneumococci, only a few studies have focused on studying its variability from a global perspective, similarly as the Human variome databases do [20]. To date only the “Microbial Variome Database” [21], which possesses and organizes the available information of the variome of the two Gram-negative bacterial species *Escherichia coli* and *Salmonella enterica*, is providing such information for microorganisms. Remarkably, there are no open-source data of this nature for any Gram-positive bacterial genome. Hence, this study focused on the construction of the first *S. pneumoniae* Variome model, starting with the identification of all allelic and protein variants, a mutation and distribution analysis (presence and absence) of the virulence factors and regulators, among a set of pneumococcal strains that possess a fully sequenced and annotated genome.

**Methods**

**Definition of the study population set and determination of the optimal representation of the entire population of pneumococci**

The search and selection of the *Streptococcus pneumoniae* strains for the analysis in this study was done using the microbial database of the “National Center for Biotechnology Information” NCBI (http://www.ncbi.nlm.nih.gov/genome) [22]. Likewise, in order to ensure an optimal representation of the global pneumococcal population, a genomic BLAST of 8290 available *S. pneumoniae* genomes was carried out. In brief, DNA alignments, employing the tool “Microbial Nucleotide BLAST” [23], that can be found in the website http://blast.ncbi.nlm.nih.gov/blast.cgi, were performed for all the currently reported draft or complete sequenced genomes. The comparative data was then employed to construct a DNA-based Phylogenetic Tree (dendrogram), by using the Genome Tree Report Tool of the NCBI (ncbi.nlm.nih.gov/genome/tree/176). Afterwards, the file containing the dendrogram, constructed for the 8290 strains, was downloaded from the NCBI database. Finally, the dendrogram file was viewed, analyzed and adapted in order to generate circular, slanted and/or rectangular cladograms, by using the online NCBI Tool “Tree Viewer 1.17.0”, which is available online at the website: ncbi.nlm.nih.gov/projects/treeview (Fig. 1).

**Definition of the virulence factors and two-component regulatory systems to be studied in *S. pneumoniae***

The search and selection of genes and proteins widely known as virulence factors or gene encoding factors possessing a proven interaction with the human host was done by an exhaustive bioinformatic screening in the database “Virulence Factors Database - VFDB” [24], available at the website http://www.mgc.ac.cn/VFs. Additionally, the virulence factors and proteins involved in interactions with the host were confirmed and completed by a systematic review of the literature [14, 25]. The common names of each one of the selected virulence factors were then introduced in the database UNIPROT [26], available at http://www.uniprot.org/, with the aim of obtaining the locus tag for *S. pneumoniae* TIGR4 genome/strain. In addition, the genes encoding the HK or RR of the pneumococcal TCS were identified by using the
database Prokaryotic Two-Component Systems - P2CS [27], available at the website http://www.p2cs.org/index.php. Likewise, the corresponding locus tag for \textit{S. pneumoniae} TIGR4 genome / strain, of each one of the histidine kinases genes (\textit{hk}) and response regulator genes (\textit{rr}), were also recovered from the same database.

Chromosomal localization of the virulence factor and two-component regulatory systems genes in \textit{S. pneumoniae}

The chromosomal location of all the genes in the genome of \textit{S. pneumoniae} TIGR4 and the construction of the genomic maps, in linear or circular representation, was done by using the software SnapGene\textsuperscript{®} (GSL Biotech), available at http://www.snapgene.com. In brief, the studied genomes of \textit{S. pneumoniae} were imported through its corresponding access code in GenBank (ie: NC\_003028.3 for TIGR4). Then, the chromosomal location of each virulence factor gene, and the factors involved in the interaction with the host and the genes encoding for proteins of simple or two-component regulatory systems were identified. Finally, the lineal maps for the scale genomic localization for the virulence factors and the circular maps for the genomic periphery of the genes that form the two-component regulatory systems were constructed.

Distribution of the virulence factors and two-component regulatory systems in the different strains of \textit{S. pneumoniae}

The identification of the genetic and protein sequences of interest to perform the comparative analysis was done, having as reference the codes (Locus Tag) in the genomes of \textit{S. pneumoniae} TIGR4 and/or R6 in the database Kyoto Encyclopedia of Genes and Genomes – KEGG [28], available at http://www.kegg.jp/kegg/. Once every gene of interest was established in the database, a series of comparisons (BLASTs) were performed using the GenomeNet [29], available at http://www.genome.jp/ , using only the fully sequenced and annotated genomes of \textit{S. pneumoniae}. For the nucleotide sequences the search was performed using the program BLASTN 2.2.29+, which uses nucleotide vs nucleotide alignments based on a punctuation matrix BLOSUM62 [23, 30]. In the same way, the search was done for the amino acid sequences using the program BLASTP 2.2.29+ [31, 32], that performs amino acids vs amino acids alignments based on a similar matrix. Once the BLAST was finalized for each virulence factor, the list was purged using as selection criteria genes with an expectancy value: e-Value = 0. The inclusion of genes with an e-value >0 was done by direct visual inspection of the alignments to check that it was indeed the same sequence. By having defined the list with the genes and proteins that fulfilled the selection criteria, it was defined to which strains of \textit{S. pneumoniae} they belong. All the DNA and protein sequences were downloaded and stored in an organized way using the fasta format.

Genetic variability (variome) of the virulence factors and two-component regulatory systems among the different pneumococcal strains

The multiple comparative alignments of pneumococcal sequences were done using the web tool MultAlin [33], available at http://multalin.toulouse.inra.fr/multalin/, for which an identity matrix 1–0 was used to assign a penalty even for the slightest change in the nucleotides or amino acids sequences, covering substitution, deletions, insertions and variations in the length. From these analyses, the number of allelic and protein variants were determined for each gene according to the registry value assigned by the program to each sequence, where equal sequences have the same registry value, while different sequences possess different values. The results of the alignments were manually curated and stored for further analysis. Finally, the precise determination of the total

Table 1

| \textit{S. pneumoniae} Strain | Serotype | \# of Genes | NCBI Annotation |
|-------------------------------|----------|-------------|----------------|
| D39                           | 2        | 2069        | NC\_008533.1   |
| R6                            | No Capsule | 1967      | NC\_003098.1   |
| TIGR4                         | 4        | 2228        | NC\_003028.3   |
| INV104                        | 1        | 2003        | NC\_017591.1   |
| AP200                         | 11A      | 2284        | NC\_014494.1   |
| JJA                           | 14       | 2235        | NC\_012466.1   |
| ATCC 700669                   | 23F      | 2224        | NC\_011900.1   |
| INV200                        | 14       | 2113        | NC\_017593.1   |
| CGSP14                        | 14       | 2276        | NC\_010582.1   |
| G54                           | 19F      | 2186        | NC\_011072.1   |
| gamPN10373                    | 1        | 2226        | NC\_018630.1   |
| P1031                         | 1        | 2254        | NC\_012467.1   |
| SPN034156                     | 3        | 1956        | NC\_021006.1   |
| SPN994039                     | 3        | 1974        | NC\_021005.1   |
| SPN990383                     | 3        | 1974        | NC\_021026.1   |
| SPN034183                     | 3        | 1985        | NC\_021028.1   |
| DXC141                        | 3        | 2037        | NC\_017592.1   |
| 670-6B                        | 6B       | 2430        | NC\_014498.1   |
| A026                          | 19F      | 2153        | NC\_022655.1   |
| Taiwan19F-14                  | 19F      | 2205        | NC\_012469.1   |
| ST556                         | 19F      | 2219        | NC\_017769.1   |
| TCHB8431/19A                  | 19A      | 2355        | NC\_014251.1   |
| SPNA45                        | 3        | 1921        | NC\_018594.1   |
| 70,585                        | 5        | 2323        | NC\_012468.1   |
| Hungary19A 6                  | 19A      | 2402        | NC\_010380.1   |
| Virulence Factors | Protein Name | Function and/or Pathogenic Role |
|-------------------|--------------|---------------------------------|
| LPxTG - Proteins  | BgaA         | β-Galactosidase Enzyme           |
|                   | EndoD        | Endo-β-N-Acetylglucosaminidase D |
|                   | PclA         | Pneumococcal Collagen-Like Protein |
|                   | SpGH101      | Endo-α-N-Acetylgalactosaminidase |
|                   | StrH         | β-N-Acetyhexosaminidase Enzyme   |
|                   | NanA         | Neuraminidase A                   |
|                   | PfbA         | Plasmin- and Fibronectin-Binding Protein A |
|                   | PrtA         | Subtilisin-Like Serine Protease |
|                   | PavB         | Pneumococcal Adherence and Virulence Protein B |
|                   | KsgA         | Dimethyladenosine Transferase    |
|                   | SpuA         | Alkaline Amylopullullanase       |
|                   | HysA         | Hyaluronate Lyase                |
|                   | SP_1492      | Cell Wall Surface Anchor Protein Family, Mucin-Binding Protein |
|                   | ZmpA         | Zinc Metalloprotease A, IgA1 Enzyme and Immune Evasion and Colonization |
|                   | ZmpB         | Zinc Metalloprotease B           |
|                   | ZmpC         | Zinc Metalloprotease C           |
|                   | ZmpD         | Zinc Metalloprotease D, IgA1 Paralog Protease |
|                   | PsrP         | Pneumococcal Serine-Rich Repeats Protein |
|                   | RrgA         | Pilus-1 Tip Protein (Adhesin)    |
|                   | RrgB         | Pilus-1 Backbone Protein         |
|                   | RrgC         | Pilus-1 Anchore Protein          |
|                   | PitA         | Pilus-2 Subunit, Ancillary Protein |
| Choline-Binding Proteins (CBPs) | LytA | Autolysin (N-Acetyl-Muramoyl-L-Alanine Amidase) Autolytic Enzyme, Cell Wall Digestion and Autolysis |
|                   | LytB         | Endo-β-N-Acetylglucosaminidase   |
|                   | LytC         | Lysozyme (1,4-β-N-AcetylMuramidase) |
|                   | Pce          | Choline-Binding Protein E, Phosphorylcholine Esterase Phosphorylcholine Esterase Enzyme, Adherence, Colonization and Cellular Metabolism |
|                   | PcpA         | Pneumococcal Choline-Binding Protein A Protection Against Lung Infection and Sepsis |
|                   | PspA         | Pneumococcal Surface Protein A Cellular Metabolism and Immune Evasion |
|                   | PspC         | Pneumococcal Surface Protein C, Choline-Binding Protein A Adherence, Immune Evasion, Colonization and Invasion |
|                   | CbpC         | Choline-Binding Protein C Virulence |
|                   | CbpD         | Choline-Binding Protein D Colonization |
|                   | CbpF         | Choline-Binding Protein F Virulence |
|                   | CbpG         | Choline-Binding Protein G Adherence and Colonization |
|                   | CbpI         | Choline-Binding Protein I Virulence |
|                   | CbpJ         | Choline-Binding Protein J Virulence |
|                   | SP_0667      | Pneumococcal Surface Protein (Putative Lysozyme) Virulence |
Table 2  Function or pathogenic role of the virulence factors and two-component regulatory systems of *S. pneumoniae* (Continued)

| Lipoproteins                  | GlqQ  | Glutamine Transporter | Cellular Metabolism |
|-------------------------------|-------|-----------------------|---------------------|
| PiaA                          | Iron-Compound ABC Transporter | Peptidyl-Proline Isomerase (PPIase) Enzyme |
| PiaA                          | Iron-Compound ABC Transporter | Peptidyl-Proline Isomerase (PPIase) Enzyme |
| PiaA                          | Streptococcal Lipoprotein Rotamase A, Peptidyl-Proline Isomerase (PPIases) Enzyme |
| PiaA                          | Peptide Permease Enzyme, Manganese ABC Transporter, Manganese-Binding Lipoprotein | Immun Evasion |
| PiaA                          | Foldase Protein PrsA, Proteinase Maturation A | Adherence, Immune Evasion, Strain-Specific Colonization and Evasion of Phagocytosis |
| AliA                          | Oligopeptide ABC Transporter | Adherence |
| PhtA                          | Pneumococcal Histidine Triad A | Adherence and Immune Evasion |
| PhtB                          | Pneumococcal Histidine Triad B | Adherence and Immune Evasion |
| PhtD                          | Pneumococcal Histidine Triad D | Adherence and Immune Evasion |
| PhtE                          | Pneumococcal Histidine Triad E | Adherence and Immune Evasion |

| Non-Classical Surface-Exposed Proteins | Eno  | Enolase (2-Phosphoglycerate Dehydratase) | Glycolytic Enzyme, Adherence and Colonization |
|---------------------------------------|------|------------------------------------------|----------------------------------------|
| GAPDH                                 | Glyceraldehyde-3-Phosphate Dehydrogenase | Glycolytic Enzyme, Adherence and Colonization |
| HtrA                                  | High-Temperature Requirement A, Serine Protease (Heat Shock Protein) | Serine Protease Enzyme |
| PavA                                  | Pneumococcal Adherence and Virulence Protein A | Adherence, Immune Evasion, Colonization and Translocation |
| Pbp1B                                 | Penicillin-Binding Protein 1B | Antibiotic Resistance |
| SkpP                                  | Serine/Threonine Protein Kinase | Cellular Metabolism and Fitness |
| Usp45                                 | PcsB, Secreted 45-KDa Protein | Virulence |
| NanB                                  | Neuraminidase B | Virulence |
| PavA                                  | Pneumococcal Protective Protein A, Non-Heme Iron-Containing Ferritine | Virulence |
| Fic-Like                              | Fic-Like Cell Filamentation Protein | Putative Cytotoxicity |
| 6PGD                                 | 6-Phosphogluconate Dehydrogenase | Virulence |
| Ply                                  | Pneumolysin | Cytolytic Toxin, Adherence, Immune Evasion, Invasion, Dissemination and Complement Activation |

| Regulators | RlA | Pathogenicity Island rIA Transcriptional Regulator | Virulence |
|------------|-----|-----------------------------------------------------|------------|
| MgrA       | MgrA Family Transcriptional Regulator | Virulence |
| MerR       | MerR Family Transcriptional Regulator | Virulence |
| PsaR       | Iron-Dependent Transcriptional Regulator | Virulence |

| Histidine Kinases (HKs) | HK01 | Sensor Histidine Kinase | Virulence |
|-------------------------|------|------------------------|------------|
| HK02                    | Sensor Histidine Kinase, VicK | Antibiotic Resistance, Virulence and Fitness |
| HK03                    | Sensor Histidine Kinase, LiaS | Antibiotic Resistance and Stress Protection |
| HK04                    | Sensor Histidine Kinase, PnpS | Genetic Competence, Fitness, Immune Evasion |
| HK05                    | Sensor Histidine Kinase, CiaH | Antibiotic Resistance, Genetic Competence and Pathogenesis |
| HK06                    | Sensor Histidine Kinase | Colonization and Invasion |
| HK07                    | Sensor Histidine Kinase, YesM | Fitness |
| HK08                    | Sensor Histidine Kinase, SaeS | Pathogenesis and Fitness |
| HK09                    | Sensor Histidine Kinase | Virulence |
| HK10                    | Sensor Histidine Kinase, VncS | Antibiotic Resistance |
| HK11                    | Sensor Histidine Kinase | Biofilm Formation |
mutations, synonymous and nonsynonymous was done using the software DnaSP V.5.1 [34, 35], available at http://www.ub.edu/dnasp/. There, all the sequences found for a determined gene were introduced and the calculations were performed for the corresponding type of mutation as mentioned before.

Results and discussion

“Hundreds to thousands” of S. pneumoniae strains and clinical isolates recovered from the nasopharynx, blood or cerebrospinal fluid (CSF) have been included up to date in genomic sequencing projects worldwide. However, pneumococcal strains, whose genomes are fully sequenced, annotated and publicly available, are the focus of this study. Therefore, a set of 25 pneumococcal strains were selected from the NCBI database, as population study, to perform the bioinformatic analysis needed to accomplish the construction of the variome of the virulence factors and two-component regulatory systems of Streptococcus pneumoniae (Table 1).

A Variome model of the Pneumococcal Virulence Factors and Regulators is an intraspecific study, aiming to highlight variable genetic loci on the genome of Streptococcus pneumoniae. A perfect and ultimate Variome model would be that constructed with the 100% of the genomic information correctly assessed from the entire pneumococcal population. However, the current state of the art is far away from this scenario and an optimal representation of the pneumococcal sets assessed up to date would be appropriate in order to validate these genomic analyzes. Currently, 8290 pneumococcal sequencing projects are reported as draft or complete genomes in the Genome Assembly and Annotation Report of the NCBI database. Therefore, a global genomic BLAST (DNA alignment) of those 8290 available S. pneumoniae genomes/strains was performed and a DNA-based Phylogenetic Tree was constructed by using the Genome Tree Report Tool of the NCBI. The topology of this phylogenetic tree (slanted cladogram) showed different pneumococcal lineages, where the selected set of 25 pneumococcal genomes/strains can be identified as external nodes (“well-distributed” key features highlighted in red), evidencing an optimal representation of the pneumococcal population (Fig. 1). In addition, it is important to highlight that the serotypes (1, 2, 3, 4, 5, 6B, 11A, 14, 19A, 19F and 23F), represented in this study population set, have been described as the pneumococcal types with the highest pathogenic potential, due to the high burden of invasive pneumococcal diseases (IPDs) they cause worldwide. This is the reason why the majority of them (except serotypes 2 and 11A) have been included in the pneumococcal conjugate vaccines (PCVs) currently used for immunization [1].

An initial considerable number of pneumococcal virulence factor genes were identified, by employing the database VFDB [24]. This database provided further detailed information to establish their function, pathogenic role and type of interaction with a receptor in its human host. Additionally, a systematic screening of the literature [14] did not only allow the confirmation of identified factors, but also ensured the possibility to complement the list with additional factors that have not been included in the databases. Likewise, the number of the tcs genes (27) was determined using the database

| Function or pathogenic role of the virulence factors and two-component regulatory systems of S. pneumoniae (Continued) |
|---------------------------------------------------------------|
| HK12 | Sensor Histidine Kinase, ComD | Genetic Competence |
| HK13 | Sensor Histidine Kinase, BlpH | Virulence |
| RR01 | Response Regulator | Virulence |
| RR02 | Response Regulator, VicR | Antibiotic Resistance, Virulence and Fitness |
| RR03 | Response Regulator, LiaR | Antibiotic Resistance and Stress Protection |
| RR04 | Response Regulator, PnpR | Genetic Competence, Fitness, Immune Evasion |
| RR05 | Response Regulator, CiaR | Antibiotic Resistance, Genetic Competence, Pathogenesis |
| RR06 | Response Regulator | Colonization and Invasion |
| RR07 | Response Regulator, YesN | Fitness |
| RR08 | Response Regulator, SaeR | Pathogenesis and Fitness |
| RR09 | Response Regulator | Virulence |
| RR10 | Response Regulator, VncR | Antibiotic Resistance |
| RR11 | Response Regulator | Biofilm Formation |
| RR12 | Response Regulator, ComE | Genetic Competence |
| RR13 | Response Regulator, BlpR | Virulence |
| RR14 | Response Regulator | Virulence |

The proteins are grouped by classes, depending on their surface-exposure mechanism. The names, abbreviations and function of the proteins were obtained from literature references.

Gámez et al. BMC Genomics (2018) 19:10
Page 7 of 18
Prokaryotic 2-Component Systems - P2CS [27]. In total, 92 different genes encoding 61 surface proteins, 4 stand-alone transcriptional regulators, 13 HKs and 14 RRs have been selected and included in this work for the construction of the variome, after being classified by their function and grouped according to their molecular mechanisms of surface-exposure (Table 2).

The genomes of 25 analyzed pneumococcal strains comprise genome sizes ranging from 2,024,476 bp in SPN034156 up to 2,245,615 bp in Hungary 19A-6. Likewise, the G+C content varies between 39.50% in CGSP14 and 39.90% in SPN034156. 670-6B is the strain with the highest number of genes (2430) and proteins (2352) and SPN034156 is the strain with the lowest number of genes (1956) and proteins (1799). Hence, the difference among genomes, regarding the number of genes and proteins can be up to 474 genes and 553 proteins, respectively. The overall number of genes for each

---

**Virulence Factor and Stand-Alone Regulator Genes on the S. pneumoniae TIGR4 Genome**

(Chromosome Size: 2,160,842 bp)

---

**Fig. 2** Chromosomal localization and direction of the virulence factor genes of *S. pneumoniae* TIGR4. Lineal representation of the pneumococcus genome. The arrows, drawn at scale, localize 62 of the 65 virulence factors and simple regulation genes considered in this study (pitA, pitB and zmpD are not present in the genome of TIGR4). Each color represents a different class of codified protein: blue = sortase-anchored proteins with an LPxTG cleavage motif; violet = choline-binding proteins (CBPs); green = lipoproteins, yellow = non-classical surface proteins (NCSP), and red = stand-alone regulators. This map was constructed using the Software SnapGene® (GSL Biotech; Available at snapgene.com)
pneumococcal genome evaluated here overmatches the overall number of proteins because the reported number of genes includes all the tRNA-, rRNA- and protein-encoding genes.

Considering the chromosomal localization of pneumococcal virulence factors genes, they are all distributed along the pneumococcal genome (Fig. 2). Interestingly, these genes are located in a co-oriented manner in relation with the origin of replication (oriC: 2,160,822–196). During the bidirectional replication of the genome, gene transcription must be simultaneous [36]. Hence, for the genes oriented in opposite direction to the corresponding replication fork, both molecular machineries will run into a frontal collision that might affect at least one of the processes. For replication, this phenomenon implies a genomic instability, while the gene transcription is probably inefficient. Previous studies have proven that the essential and highly constitutively expressed genes are co-oriented [36]. For the pneumococcus, 30 of the 36 genes encoding virulence factors are localized in the first half of the genome, on the forward strand, and co-oriented with the replication fork clockwise. Similarly, 21 of the 27 virulence factor genes localized on the second half of the genome, are located on the reverse strand and co-oriented with the replication fork moving anti-clockwise (Fig. 2). A similar genome organization is observed for the 27 genes that encode the TCSs in S. pneumoniae, where only one operon, the tcs04 genes (TCS04), is not co-oriented with the replication fork (Fig. 3). These data reinforce the idea that the virulence factor genes and the genes of the tcs are highly important for the pneumococcal interaction with the human host, and its pathogenic potential in processes such as adherence, colonization, invasion, immune evasion, fitness, antibiotic resistance and natural competence (Table 2).

The analysis of the distribution of genes associated with virulence and host-pathogen interactions among the studied pneumococcal strains revealed that only 26 of the 65 genes considered here are present in the all 25 strains. These genes encode for products involved in different functions such as cell wall hydrolysis, ABC transporters and structural proteins implied in the adherence to host tissue, the so-called adhesins. Interestingly, after preliminar inspection (by locus tag, identifier names and/or product sizes) of the datasets and supplementary material reported by van Tonder and colleagues in 2017, only a few of the pneumococcal virulence factors (PspC, KsgA, and 4 hypothetical lipoproteins) and regulators (RR04, HK08, RR08, RR09, RR10) were found in the pneumococcal “supercore” genomic list of 303 genes, based on the analysis of 3121 pneumococci recovered from healthy individuals from four different subsets of the global pneumococcal population [15]. These findings, if confirmed after deeper analysis of the datasets based on sequence comparison, may indicate that pneumococcal pathogenesis is a much more complex process than thought before. While most of the genes have a single copy in the genome, the lytA gene, encoding the major pneumococcal autolysin, is found also in two and even three copies in 13, and 2 strains, respectively. This is most likely due to the multiple integration of prophages in the chromosomal DNA [37] (Table 3). In strain SPNA45, the gene gnd, encoding the enzyme 6-phosphogluconate dehydrogenase, is duplicated and fused with a second copy of its downstream neighbor gene, which encodes the orphan response regulator (rr14). The remaining 39 of the 65 virulence factor genes were found to belong to the accessory genome, presenting different degrees of absence in the 25 strains. Thus, all these genes are not essential but are beneficial for fitness and pathogenesis. Striking examples are the genes encoding the Pilus-1 and Pilus-2 structures that have been identified to mediate adherence, contribute to virulence and promote invasion [38–42]. These genes are located on pathogenicity islands (PAI) and these islands contain also the genes required for cell surface anchoring and regulation [38–41]. Remarkably, strains like ST556, Taiwan19F-14 and TCH8431/19A, were detected here as positive for both types of pili (1 and 2). Among the other genes with restricted presence in some strains it is important to mention that they encode for sortase-anchored proteins.
| Virulence Factors | TIGR4 | D39 | RE/70 | 3583B | Hungary | ST/19 | SPN994038 | SPN994039 | SPN994085 | GWC/141 | SPN994056 | ST/55 | Taiwan/59/19 | A026 | IN4200 | CG3P14 | IN4/104 | G54 | AP2000 | P1031 | gam/PN00375 | SPN4A5 | ATCC | JJA | 700669 |
|-------------------|-------|-----|-------|-------|---------|-------|------------|------------|------------|---------|------------|-------|-----------|-------|--------|-------|--------|-----|--------|--------|-----------|--------|------|-----|-------|
| **UpReg** - Proteins |       |     |       |       |         |       |            |            |            |         |            |       |           |       |        |       |        |     |        |        |           |        |     |     |       |
|                    |       |     |       |       |         |       |            |            |            |         |            |       |           |       |        |       |        |     |        |        |           |        |     |     |       |
| **Choline- Binding Proteins (CBPs)** |       |     |       |       |         |       |            |            |            |         |            |       |           |       |        |       |        |     |        |        |           |        |     |     |       |
| Table 3 Distribution of the virulence factor and regulation genes of *S. pneumoniae* |       |     |       |       |         |       |            |            |            |         |            |       |           |       |        |       |        |     |        |        |           |        |     |     |       |
The present table shows the absence (0) or presence (1, 2 or 3) of each considered genes in the 25 strains selected for this study. The number (1, 2 or 3) indicates the amount of copies of each gene in the genome. lytA is the only factor with more than one copy per genome. In the strain SPNA45, the gene gnd was found duplicated (2) and fused with a duplication of its neighbor gene (rr14) downstream. In the gene nanA of TIGR4 (1) a shift in its ORF was found. However, it has also been reported that NanA is expressed in this pneumococcal strain. Gene defective copies (genes with any alteration in their primary DNA sequences) are depicted in bold and italics: In the SPNA45 strain ply is fused with a copy of lytA, and pspA is defective in the ATCC700669 pneumococcal strain.
Table 4: Distribution of the genes that conform the two-component systems in *S. pneumoniae*

| Two-Component Systems (TCSs) | TIGR4  | D39  | R6    | 70,585 Hong 19A | SPN94018 | SPN94039 | SPN034183 | GXTC141 | SPN034156 | ST536 | Taiwan 19F-146 | TCH8431/19A | A036 | INV200 | C63 | AP100 | PT1031 | gamP N0373 | SPN45 | ATCC 700669 | JIA 670686 |
|------------------------------|--------|------|-------|-----------------|----------|----------|-----------|---------|-----------|--------|-----------------|------------|------|--------|----|-------|---------|-------------|--------|-------------|------------|
| Histidine Kinases (HKs)      |        |      |       |                 |          |          |           |         |           |        |                 |            |      |        |    |       |         |             |        |             |            |
| hk01                         | 1      |      | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| hk02                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| hk03                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| hk04                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| hk05                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| hk06                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| hk07                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| hk08                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| hk09                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| hk10                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| hk11                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| hk12                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| hk13                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| Response Regulators (RRs)    |        |      |       |                 |          |          |           |         |           |        |                 |            |      | 1       |    |       |         |             |        |             |            |
| rr01                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| rr02                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| rr03                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| rr04                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| rr05                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| rr06                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| rr07                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| rr08                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| rr09                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| rr10                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| rr11                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| rr12                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| rr13                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |
| rr14                         | 1      | 1    | 1     | 1               | 1        | 1        | 1         | 1       | 1         | 1      | 1               | 1          |      | 1       |    | 1     |         |             |        |             |            |

The table shows the absence (0) or presence (1 or 2) of each gene considered in the 25 strains selected for this study. The number (1 or 2) indicates the amount of copies per gene in each genome. rr14 is the only gene with more than one copy per genome (2), which is actually fused with its neighbor gene (gnd) upstream. In bold and italics, three genes are observed (hk01, hk12 and rr04) in two different strains which might have some alteration (insertion or deletion) in their primary DNA and protein sequence. The two-component system NisK-NisR of the pneumococcus is rare, of the 25 strains analyzed in this study only in the strain 70,585 was found.
Table 5: Analysis of the Variome of the virulence factor genes of *S. pneumoniae*

| Virulence Factor Genes | Length Gene (bp) | Protein (a.a.) | Mutations Overall | Synonymous | Non-Synonymous | Variants Alleles | Protein Analyzed Sequences |
|------------------------|------------------|----------------|-------------------|------------|---------------|-----------------|--------------------------|
| **lytA**               | sp_1937 / spr1754| 957            | 318               | 208        | 154           | 54              | 25 20 42              |
| **gnd**                | sp_0375 / spr3035| 1425           | 474               | 82         | 75            | 7               | 19 11 26               |
| **strH**               | sp_0057 / spr0057| 3939           | 1312              | 87         | 32            | 55              | 21 21 25               |
| **lytB**               | sp_0965 / spr0867| 1977           | 658               | 132        | 84            | 48              | 22 20 25               |
| **endoD**              | sp_0498 / spr0040| 4980           | 1659              | 139        | 70            | 69              | 20 20 25               |
| **nanA**               | sp_1693 / spr1536| 3108           | 1035              | 460        | 282           | 178             | 20 19 25               |
| **tgaA**               | sp_0648 / spr0565| 6702           | 2233              | 415        | 247           | 168             | 19 18 25               |
| **spGH101**            | sp_0368 / spr0328| 5304           | 1767              | 348        | 238           | 110             | 18 18 25               |
| **glnQ**               | sp_0609 / spr0534| 765            | 254               | 40         | 22            | 18              | 17 17 25               |
| **phtE**               | sp_1004 / spr0908| 3120           | 1039              | 56         | 28            | 28              | 20 16 25               |
| **pbb18**              | sp_2099 / spr1909| 2466           | 821               | 81         | 16            | 65              | 17 16 25               |
| **pce**                | sp_0930 / spr0831| 1884           | 627               | 148        | 80            | 68              | 16 16 25               |
| **pavA**               | sp_0966 / spr0868| 1656           | 551               | 40         | 18            | 22              | 17 15 25               |
| **cbpD**               | sp_2201 / spr2006| 1347           | 448               | 49         | 28            | 21              | 18 14 25               |
| **piuA**               | sp_1872 / spr1687| 966            | 321               | 22         | 7             | 15              | 14 12 25               |
| **hrA**                | sp_2239 / spr2045| 1182           | 393               | 12         | 8             | 4               | 14 10 25               |
| **pcaA**               | sp_1546 / spr1402| 630            | 209               | 19         | 9             | 10              | 13 10 25               |
| **mgaA**               | sp_1800 / spr1622| 1482           | 493               | 21         | 15            | 6               | 13 9 25                |
| **pplA**               | sp_0771 / spr0679| 804            | 267               | 16         | 8             | 8               | 12 9 25                |
| **stkP**               | sp_1732 / spr1577| 1980           | 659               | 52         | 46            | 6               | 18 7 25                |
| **gap**                | sp_2012 / spr1825| 1008           | 335               | 14         | 12            | 2               | 13 7 25                |
| **psaA**               | sp_1650 / spr1494| 930            | 309               | 53         | 41            | 12              | 13 6 25                |
| **psaR**               | sp_1638 / spr1480| 651            | 216               | 10         | 5             | 5               | 9 6 25                 |
| **piaA**               | sp_1032 / spr0934| 1026           | 341               | 9          | 2             | 7               | 7 6 25                 |
| **usp45**              | sp_2216 / spr2021| 1179           | 392               | 9          | 6             | 3               | 8 4 25                 |
| **eno**                | sp_1128 / spr1036| 1305           | 434               | 17         | 16            | 1               | 12 2 25                |
| **spaA**               | sp_0268 / spr0247| 3843           | 1280              | 121        | 77            | 44              | 18 18 24               |
| **prtA**               | sp_0641 / spr0561| 6423           | 2140              | 436        | 272           | 164             | 18 16 24               |
| **nanB**               | sp_1687 / spr1531| 2094           | 697               | 45         | 20            | 25              | 18 16 24               |
| **phbA**               | sp_1833 / spr1652| 2127           | 708               | 209        | 80            | 129             | 15 15 24               |
| **pppA**               | sp_1572 / spr1430| 537            | 178               | 69         | 43            | 23              | 16 12 24               |
| **ply**                | sp_1923 / spr1739| 1416           | 471               | 20         | 19            | 1               | 14 2 24                |
| **pavB**               | sp_0082 / spr0075| 2574           | 857               | (--)       | (--)          | (--)            | 20 17 23               |
| **phdD**               | sp_1003 / spr0907| 2520           | 839               | 604        | 331           | 273             | 17 17 23               |
| **zmpB**               | sp_0664 / spr0581| 5646           | 1881              | (--)       | (--)          | (--)            | 15 15 23               |
| **psoA**               | sp_0117 / spr0121| 2235           | 744               | (--)       | (--)          | (--)            | 18 17 22               |
| **cbpC**               | sp_0377 / spr0337| 1023           | 340               | 176        | 87            | 89              | 15 14 22               |
| **ksgA**               | sp_1992 / spr1806| 666            | 221               | 36         | 8             | 28              | 14 14 22               |
| **ppmA**               | sp_0981 / spr0884| 942            | 313               | 9          | 6             | 3               | 9 6 22                 |
| **hysA**               | sp_0314 / spr0286| 3201           | 1066              | 91         | 41            | 50              | 18 17 20               |
| **lytC**               | sp_1573 / spr1431| 1473           | 490               | 80         | 20            | 60              | 17 16 20               |
| **pspC**               | sp_2190 / spr1995| 2082           | 693               | (--)       | (--)          | (--)            | 19 19 19               |
| **aliA**               | sp_0366 / spr0327| 1986           | 661               | 67         | 50            | 17              | 17 17 19               |
or choline-binding proteins (CBPs), as well as histidine triad proteins (pht genes). These gene products are associated with different processes of bacterial fitness and pathogenesis (Tables 3 and 2) [6, 43, 44]. Regarding the distribution and data of the analyzed strains for the TCS most of them were found in the 25 pneumococcal strains. Exceptions are presented by the TCS07 and TCS12, which contribute to fitness and competence, respectively [7, 45]. These TCS are absent in a couple of strains (Table 4). In some other strains genes like hko1, hko2, and rr04, presented incomplete sequences, an artefact leading to truncated and hence non-functional proteins/regulators (Tables 3 and 2). Interestingly, only the genes encoding the hko8, rr08, rr09, rr06 and rr04 were found to belong to the “supercore” genomic set of genes reported by van Tonder et al., in 2017 [15], indicating the important role these highly conserved and well-distributed regulatory proteins play in the pneumococcus and in its interplay with the environment.

The estimation of the variability for each individual virulence factor and pneumococcal regulator (at the DNA and protein level) allowed the construction of a partial variome for the analysed 25 pneumococcal strains. Briefly, the variome takes into consideration the estimation of (1) the presence, absence or the number of copies of genes in the different strains, (2) the number of total synonymous and nonsynonymous mutations, and (3) the number of allelic and protein variants explaining the variability for each factor. The results summarized in Tables 5 and 6, contain the data for the genes and proteins associated to virulence and host-pathogen interaction, and also the data for the stand-alone and TCS regulators. Specifically there are some identified factors with the best distribution and highest evolutionary conservation, These were (1) the ply gene encoding the sole pneumococcal cytolysin and cytoxin pneumolysin [46], (2) the enolase, which encodes the enzyme enolase (2-phosphoglycerate dehydratase) and has an essential function in the metabolism [47], but also interacts specifically with plasmin(ogen) and is therefore involved in fibrinolytic processes, adherence and virulence, and (3) the pcsB (Usp45) gene, which encodes for a 45-kDa secreted and immunogenic protein that is involved in cell division and stress response [48]. As for the mutations, these three proteins presented a minor number of changes, in comparison with others proteins that were also analyzed. The variome of the TCS (Table 6) allowed to conclude that the most conserved genes from the evolutionary point of view, are the genes hko5 and rr05 of

| Analysis of the Variome of the virulence factor genes of S. pneumoniae (Continued) |
|---------------------------------|-----------------|---------------|
| **SP_0667** sp_0667 / sp0583 | 999             | 332           |
| **merR** sp_0739 / sp0649     | 741             | 246           |
| **pcpA** sp_2136 / sp1945     | 1866            | 621           |
| **SP_1492** sp_1492 / sp1345 | 609             | 202           |
| **cbpG** sp_0390 / sp0349     | 858             | 285           |
| **zmpA** sp_1154 / sp1042     | 6015            | 2004          |
| **cbpJ** sp_0378 / Absent     | 999             | 332           |
| **nanC** sp_1326 / Absent     | 2223            | 740           |
| **phtA** sp_1175 / sp1061     | 2409            | 802           |
| **phtB** sp_1174 / Absent     | 2460            | 819           |
| **cbpF** sp_0391 / sp0351     | 1023            | 340           |
| **zmpD** Absent / Absent      | 5238            | 1745          |
| **rsgA** sp_0462 / Absent     | 2682            | 893           |
| **rsgC** sp_0464 / Absent     | 1182            | 393           |
| **rsgB** sp_0463 / Absent     | 1998            | 665           |
| **rflA** sp_0461 / Absent     | 1530            | 509           |
| **pitA** Absent / Absent      | 1770            | 589           |
| **pitB** Absent / Absent      | 1233            | 410           |
| **psrP** sp_1772 / Absent     | 14,331          | 4776          |
| **zmpC** sp_0071 / Absent     | 5571            | 1856          |
| **cbpl** sp_0069 / Absent     | 636             | 211           |

Data of the punctual genetic variability (total mutations, synonymous and nonsynonymous + allelic and protein variants) estimated for each one of the virulence factors and simple regulators genes. The analyzed sequences depend on the presence, absence or number of copies of the genes in the different strains. The size of the sequences and loci are also shown in TIGR4 and R6, pneumococcal representative strains. Factors in bold were identified as the most conserved. (=) = mutations could not be estimated for different reasons, like repetitive sequences.
ciaR/H (tcs05). The TCS CiaRH is involved in the resistance to cefotaxime, regulation of genetic competence and increase in pathogenicity in the respiratory tract in murine models [7, 49, 50]. Meanwhile, hk02 and rr02 (WalR/K, MicA/B or VicR/K), have been associated with resistance to erythromycin and are essential for the bacterial growth. Nevertheless, the latter was proven to be due to its regulator (pcsB), and was no longer essential upon ectopic expression of PcsB [7, 48]. Pneumococcal TCS08 is involved in the genetic regulation of pilus-1 [41]. The mutation analysis showed that the response regulators exhibited a lower rate of variations in comparison to the histidine kinases, being the response regulators rr05, rr02, rr06, and rr08 the most conserved. All the results obtained in this study support the global idea of a new generation of protein-based and serotype-independent vaccines for *Streptococcus pneumoniae*. The basis is the high degree of distribution and conservation of the virulence proteins in combination with the importance of their functions and immunogenic capacities. This probably makes them ideal pharmacological targets to treat the pneumococcus and its diseases. This might be an alternative to the immunization with the conjugated serotypes, or represent a strategy to combine immunogenic and highly conserved proteins with capsular polysaccharides to generate a serotype-independent immune response.

### Conclusions

The construction of this “low-scale” Variome model for the virulence factors and regulators of *Streptococcus pneumoniae*
pneumoniae was achieved from 25 pneumococcal strains with fully sequenced and annotated genomes. According to the Molecular Phylogenetic Analysis performed on the NCBI website, this selected set of pneumococcal genomes ensured an optimal representation of the pneumococcal population (8290 strains) reported in the NCBI database up to date. Similarly, this study population set also represented an important group of highly pathogenic pneumococcal serotypes (1, 2, 3, 4, 5, 6B, 11A, 14, 19A, 19F and 23F), which have been also included in the current pneumococcal conjugate vaccine formulations (except serotypes 2 and 11A), used to prevent pneumococcal infections. A total of 92 different genes and proteins were identified, classified, and studied for the construction of the variome. The genes of the pneumococcal virulence factors and TCS, are distributed along the genome, and are located in such a manner that transcription is co-oriented with replication. The analysis of the gene distribution in this study population set showed that 26 of them were found in the 100% of the 25 pneumococcal genomes/strains (core genome), while 39 are part of the flexible genome. The estimation of the variability for each individual virulence factors, stand-alone regulator or TCS, indicated that the virulence factors with the lowest variability in the pneumococcus are pneumolysin, enolase and PcsB, while the regulators with the highest conservation are TCS05 (CiaR/H), TCS02 (VicR/K) and TCS08. Finally, all the results obtained here with the bioinformatic analysis performed, constitute the first model to compare, visualize and understand the future flood of new genomic data about the genetic variation (in terms of gene presence/absence or mutation) of pneumococcal virulence factors and regulators [51–53]. The applicability offered by this variome model, together with further population genomic analysis of pneumococci, will provide relevant information on potential targets for vaccines, supporting the idea of a new generation of protein-based formulations to combat Streptococcus pneumoniae and its disease burden.

Abbreviations

BLAST: Basic local alignment search tool; BLOSUM: Blocks substitution matrix; CBPs: Choline binding proteins; CSF: Cerebrospinal fluid; DnaSP: DNA sequence polymorphism; HK: Histidine kinase; IPs: Invasive pneumococcal diseases; KEGG: Kyoto encyclopedia of genes and genomes; MultiAlin: Multiple sequence alignment; NCBI: National center for biotechnology information; NCSP: Non-classical surface proteins; P2CS: Prokaryotic two-component systems; PAt: Pathogenicity Islands; RR: Response regulator; S. p.: Streptococcus pneumoniae; TCS: Two-Component regulatory Systems; UNIPROT: The Universal Protein Resource; Variome: Pan-genomic variability map; VFDB: Virulence factors data base

Acknowledgements

The authors thank to Prof. Vanessa Cienfuegos, School of Microbiology, University of Antioquia for her critical evaluation of this research work and manuscript. We express our acknowledgments to peer reviewers for critical review of the manuscript. Their suggestions and comments significantly improved the quality of this piece of work.

Funding

The fundings for this research work have been provided by the Committee for Development of Research at the University of Antioquia (CODY, CIEMB-097-13) in Colombia, and the DFG GKR 1870/1 (Bacterial Respiratory Infections) in Germany. Both funding sources had no involvement in the design of this study, in the collection, analysis and interpretation of data, in the writing of this manuscript, and in the decision to submit this article for publication.

Availability of data and materials

Sequence data that support the findings of this study were already-published information, retrieved from GenBank (accession numbers are provided in Table 1). All the bioinformatic-analyzed data generated here are included in this published study. However, supplementary raw information files (mainly DNA and protein sequence comparisons) are available from the corresponding author on reasonable request.

Authors’ contributions

All the authors have contributed to this research work, participating in the conception and design (GG, AC, MG, AB, SH), collection and analysis of information (GG, AC, MG, AB), discussion of results (GG, AC, MG, AB, AGM, MC, SH), manuscript draft preparation (GG, AC), and critical revision and edition (GG, AC, AGM, MC, SH) of the manuscript. GG and AC have contributed equally to this research work and manuscript. All the authors have read and approved the final version of this manuscript.

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Competing interests

The authors declare that they have no competing interests in relation with this research work and manuscript.

Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

1Genetics, Regeneration and Cancer Research Group (GRC), University Research Centre (SIU), Universidad de Antioquia (UdeA), Calle 70 # 52 - 21, 050010 Medellin, Colombia. 2Basic and Applied Microbiology Research Group (MICROBA), School of Microbiology, Universidad de Antioquia (UdeA), Calle 70 # 52 - 21, 050010 Medellin, Colombia. 3Department of Molecular Genetics and Infection Biology, Interfaculty Institute of Genetics and Functional Genomics, Center for Functional Genomics of Microbes, Ernst-Moritz-Arndt University of Greifswald, Felix-Hausdorff-Str. 8, D-17487 Greifswald, Germany. 4School of Microbiology, University of Antioquia, Bloque S - Oficina 408, Calle 70 # 52 - 21, 050010 Medellin, Colombia.

Received: 3 August 2017 Accepted: 11 December 2017

Published online: 03 January 2018

References

1. World Health Organization. The global burden of disease. 2004 update. Geneva: WHO; 2008.
2. Bricard-Pappas AE, Margolis MB, Center KJ, Isaacman DJ. Streptococcus Pneumoniae: description of the pathogen, disease epidemiology, treatment, and prevention. Pharmacotherapy. 2005;25(9):1193–212.
3. Reithner R, Muenz LR, O’Brien KL. Systematic evaluation of serotypes causing invasive pneumococcal disease among children under five: the pneumococcal global serotype project. PLoS Med. 2010;7(10):e1–13.
4. Johnson HI, Deloria-Knoll M, Levine OS, Stoszek SK, Freimanis Hance L, Rotherhing R, Muenz LR, O’Brien KL. Systematic evaluation of serotypes causing invasive pneumococcal disease among children under five: the pneumococcal global serotype project. PLoS Med. 2010;7(10):1–13.
5. Jedzejas MJ. Pneumococcal virulence factors: structure and function. Microbiol Mol Biol Rev. 2001;65(2):187–207. First page, table of contents
6. Voss S, Gamez G, Hammerschmidt S. Impact of pneumococcal microbial surface components recognizing adhesive matrix molecules on colonization. Mol Oral Biol. 2012;27(4):246–56.

7. Throup JP, Korette KK, Bryant AP, Inghram KA, Chalker AF, GY, Yaraa M, Wallis NC, Brown JR, Holmes DJ, et al. A genomic analysis of two-component signal transduction in Streptococcus pneumoniae. Mol Microbiol. 2003;53(3):566–76.

8. McCluskey J, Hinds J, Husain S, Witney A, Mitchell TJ. A two-component system that controls the expression of pneumococcal surface antigen a (PsaA) and regulates virulence and resistance to oxidative stress in Streptococcus pneumoniae. Mol Microbiol. 2000;36(4):1661–75.

9. Gamez G, Hammerschmidt S. Combat pneumococcal infections: adhesins as candidates for protein-based vaccine development. Curr Drug Targets. 2012;13(3):323–7.

10. Centers for Disease Control and Prevention. Active Bacterial Core Surveillance Report, Emerging Infections Program Network, Streptococcus pneumoniae. Atlanta: CDC; 2015.

11. Eldholm V, Johnborg O, Ohnstad HS, Berg KH, Hermoso JA, Havarstein LS. Pneumococcal CspB is a murine hydrolase that requires a dual cell envelope binding specificity to kill target cells during fradicate. Mol Microbiol. 2010;76(4):905–17.

12. Donkor ES. Understanding the pneumococcus: transmission and evolution. Front Cell Infect Microbiol. 2013;3:7.

13. Feil EJ, Smith JM, Enright MC, Spratt BG. Estimating recombination parameters in Streptococcus Pneumoniae from multiclonal sequence typing data. Genetics. 2000;154(4):1439–50.

14. Donat C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV, Oggoni M, Dunning Hotopp JC, Hu ZF, Riley DR, et al. Structure and dynamics of the pan-genome of Streptococcus Pneumoniae and closely related species. Genome Biol. 2010;11(10):R107.

15. van Tonder AJ, Bray JE, Zollikofer M, Horsfield SC, Bentley SD, Hardison SR, Kristensen KG et al. Heterogeneity Among Estimates Of The Core Genome And Pan-Genome In Different Pneumococcal Populations. bioRxiv 2017, doi:https://doi.org/10.1101/133991.

16. Ara RS, Kang J, Tsushima AI, Harasaki Y, Blaser MJ. Extensive repetitive DNA element facilitates prokaryotic genome plasticity. Proc Natl Acad Sci U S A. 2003;100(23):13579–84.

17. Chevapchieva C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, Yao Z, Sun L, Shen Y, Jin Q. VFDB: a reference database for clinical isolates of Streptococcus pneumoniae. Nucleic Acids Res. 2010;38(Database issue):D115–8.

18. Donati C, Hiller NL, Croucher NJ, Angiuoli SV, Oggoni M, Dunning Hotopp JC, Hu ZF, Riley DR, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014;46(3):305–9.

19. Croucher NJ, Harris SR, Fraser CQ, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, et al. Rapid evolution of pneumococcal resistance in response to clinical interventions. Science. 2011;331(6016):430–4.

20. Sokurenko EV, Gomulkiewicz R, Dykhuizen DE. Source-sink dynamics of pneumococcal populations. bioRxiv 2018, doi:https://doi.org/10.1101/133991.

21. van Kooyk Y, Geijtenbeek TB. DC-SIGN: escape mechanism for pathogens. Nat Rev Immunol. 2003;3(9):697–709.

22. Porukova M, Moschioni M, De Angelis G, Barocci M, Sabbanwal V, Mignani V, Velton SL. Variations of pneumococcal Puls-1 expression results in vaccine escape during experimental Otis media (EOM). PLoS One. 2014;9(11):e83798.

23. Donati C, Hiller NL, Croucher NJ, Angiuoli SV, Oggoni M, Dunning Hotopp JC, Hu ZF, Riley DR, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014;46(3):305–9.

24. van Tonder AJ, Bray JE, Zollikofer M, Horsfield SC, Bentley SD, Hardison SR, Kristensen KG et al. Heterogeneity Among Estimates Of The Core Genome And Pan-Genome In Different Pneumococcal Populations. bioRxiv 2017, doi:https://doi.org/10.1101/133991.

25. van Kooyk Y, Geijtenbeek TB. DC-SIGN: escape mechanism for pathogens. Nat Rev Immunol. 2003;3(9):697–709.

26. Porukova M, Moschioni M, De Angelis G, Barocci M, Sabbanwal V, Mignani V, Velton SL. Variations of pneumococcal Puls-1 expression results in vaccine escape during experimental Otis media (EOM). PLoS One. 2014;9(11):e83798.

27. van Kooyk Y, Geijtenbeek TB. DC-SIGN: escape mechanism for pathogens. Nat Rev Immunol. 2003;3(9):697–709.

28. Porukova M, Moschioni M, De Angelis G, Barocci M, Sabbanwal V, Mignani V, Velton SL. Variations of pneumococcal Puls-1 expression results in vaccine escape during experimental Otis media (EOM). PLoS One. 2014;9(11):e83798.

29. van Kooyk Y, Geijtenbeek TB. DC-SIGN: escape mechanism for pathogens. Nat Rev Immunol. 2003;3(9):697–709.

30. Porukova M, Moschioni M, De Angelis G, Barocci M, Sabbanwal V, Mignani V, Velton SL. Variations of pneumococcal Puls-1 expression results in vaccine escape during experimental Otis media (EOM). PLoS One. 2014;9(11):e83798.

31. van Kooyk Y, Geijtenbeek TB. DC-SIGN: escape mechanism for pathogens. Nat Rev Immunol. 2003;3(9):697–709.

32. Porukova M, Moschioni M, De Angelis G, Barocci M, Sabbanwal V, Mignani V, Velton SL. Variations of pneumococcal Puls-1 expression results in vaccine escape during experimental Otis media (EOM). PLoS One. 2014;9(11):e83798.

33. van Kooyk Y, Geijtenbeek TB. DC-SIGN: escape mechanism for pathogens. Nat Rev Immunol. 2003;3(9):697–709.
52. Castro AF, Gómez-Mejia A, Gallego M, Bedoya A, Hammerschmidt S, Gámez GA. Variome of the Pneumococcal Surface-Exposed Proteins and other Virulence Factors: A Bioinformatics Analysis. [Abstract EuroPneumo-P1.27]. Pneumonia. 2015;7:17.

53. Gámez GA, Castro AF, Gómez-Mejia A, Gallego M, Bedoya A, Hammerschmidt S. Análisis Bioinformático y Construcción del Varioma de los Factores de Virulencia y Sistemas de Regulación por Dos-Componentes de Streptococcus pneumoniae. [Abstract 3rd Colombian Congress on Computational Biology and Bioinformatics-CCBCOL3]. Medellín - Colombia; 2015, Oral Presentation 129.