A Pan-Genome Wide Association Study to Identify Genes Associated with Invasive *Streptococcus Pneumoniae*

Arash Iranzadeh

Supervisor: Prof. Nicola Mulder

Co-Supervisor: Prof. Dean Everett

The thesis is prepared for

Faculty of Health Science

University of Cape Town

in fulfillment for

Ph.D. in Bioinformatics

Computational Biology Division (CBIO)

Department of Integrative Biomedical Sciences

Institute of Infectious Disease and Molecular Medicine

Faculty of Health Sciences

University of Cape Town

2022

Acknowledgment

I want to acknowledge everyone who contributed to the successful completion of this project, especially my supervisor, Professor Nicola Mulder, for her invaluable advice and guidance and our colleagues from Malawi Liverpool Wellcome Trust Centre. Besides, I thank the Faculty of Health Science (FHS) and the Computational Biology (CBIO) group at the University of Cape Town (UCT) and appreciate my loving parents, that helped and encouraged me.

Computations were performed using facilities provided by the University of Cape Town's ICTS High-Performance Computing team: hpc.uct.ac.za. I acknowledge the Centre for High-Performance Computing (CHPC), South Africa, for providing computational resources to this research project. I thank the study participants and all involved staff at the Karonga Prevention Study.

Declaration

I, *Arash Iranzadeh*, hereby declare that the work on which this thesis is based is my original work (except where acknowledgments indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for research either the whole or any portion of the contents in any manner whatsoever.

Date: September 2022

Signed by candidate

Table of Contents

List of Figures

List of Tables

Abstract

Streptococcus pneumoniae (pneumococcus) is one of the leading causes of mortality in Africa. It asymptomatically colonizes the human nasopharynx. The invasive pneumococcal disease occurs when isolates spread to normally sterile sites such as lungs, blood, and the central nervous system. Colonization, though, does not necessarily lead to infection. Some isolates remain in the upper respiratory tract only, without causing any pathogenic symptoms. This thesis hypothesized that invasive and non-invasive isolates differ genetically. We tested this hypothesis by applying a pan-genome approach using whole-genome sequencing short reads of 1477 samples from Malawi, including those obtained from the nasopharynx of carriers (825 samples) and from the blood and cerebrospinal fluid of patients (652 samples). *In-silico* serotyping identified 56 serotypes in the cohort and statistical analysis showed that despite the vaccination, the prevalence of serotypes 1 and 12F increased amongst patients. Genomes were assembled, and a reference pan-genome for all strains was built. Short reads were aligned to the core genome, and core variants were called. The population structure was determined based on the distribution of variants in the pan-genome. Finally, genes with a significant presence in the invasive isolates were identified. Functional enrichment analysis of potential virulence genes was carried out to address how specific genes may contribute to the pathogenesis.

The findings highlighted the features of the *pneumococcus* pan-genome in Malawi. The core- and accessory-genome were characterized based on the functional analysis of genes. The core components included:

- Ribosomal subunits.
- Subunits of F-type ATP synthase.
- Enzymes that catalyze the attachment of amino acids to tRNA molecules, DNA replication, DNA repair, and homologous recombination.
- 10.13% of the core and soft-core genes were uncharacterized.

In the accessory genome, the study detected the presence of genes from Regions of Diversity (RDs), including:

- Subunits of V-type ATPases and Sodium/solute symporter from RD8a.
- Enzymes from RD3 catalyzing the capsule synthesis.
- Subunits of PsrP secY2A2 pathogenicity island from RD10.
- Genes from RD6 and RD7 involved in transposing mobile genetic elements.
- Genes from RD2 RD8b, and RD12 participating in communication and competition.
- Genes from RD4 that assemble pilins into pili and anchor pili to the cell wall.
- 53.58% of accessory genes were uncharacterized.

Most serotypes showed a similar prevalence in carriage and disease groups. However, the significant abundance of serotypes 1, 5, and 12F among patients compared to the carriage group suggested they are highly invasive with a short colonization period. These serotypes exhibited a remarkable genetic distinction from others. Their divergence included the absence and presence of several genes in their genome structure. The lack of genes from a genomic island known as RD8a was the most pronounced difference between serotypes 1, 5, and 12F compared to significantly prevalent serotypes in the nasopharynx. Genes in RD8a are involved in binding to epithelial cells and doing aerobic respiration to synthesize ATP through oxidative phosphorylation. We hypothesized that the absence of RD8a from

serotypes 1, 5, and 12F may be associated with their short duration in the nasopharynx where they need to bind to epithelial cells and access free oxygen molecules required for aerobic respiration. Given this, the amount of ATP is likely to decline in serotypes 1, 5, and 12F, causing them to harbor more phosphotransferase systems to transport carbohydrates since these transporters use phosphoenolpyruvate as the energy source instead of ATP. In conclusion, serotypes 1, 5, and 12F, the most prevalent and invasive pneumococcal strains in Malawi, showed a considerable genetic distinction from other strains that may be associated with their short colonization period and quickness to infect the blood and cerebrospinal fluid.

1 Literature review

1.1 Streptococcus pneumoniae

Streptococcus pneumoniae, S. pneumoniae, or *pneumococcus,* is an important microorganism. In 1881, *Louis Pasteur* and *George Miller Sternberg* discovered it independently, and its role in causing pneumonia was identified by the early 1880s[1]. *Pneumococcus* was the bacterium used by *Frederick Griffith* in 1928 to confirm genetic exchange in bacteria. Later, *Oswald Avery*, *Colin MacLeod*, and *Maclyn McCarty* determined DNA as the cause of bacterial transformation in the *Griffith* experiment[2]. The *pneumococcus* is a Gram-positive, facultative anaerobe, catalase-negative, and alpha-hemolytic bacterium that often colonizes large mammals and humans. The first complete genomic sequence of *pneumococcus* became available in 1997; this was the first sequenced gram-positive bacterium[3].

Pneumococcus asymptomatically colonizes the human nasopharynx. The carriage rate is influenced by factors such as age, race, smoking, viral infection, overcrowding, and winter[4,5]. Pneumococcal colonization often presents with no symptoms and does not necessarily lead to disease. However, colonization is essential for pneumococcal infection and the horizontal spread of the pathogen[6]. During colonization, pneumococci take up exogenous DNA from other species in the nasopharynx and adapt their genome structure to the surrounding environment. The communication between nasopharyngeal bacterial species is crucial for the exchange of drug resistance genes and virulence factors[7].

The symptomatic disease occurs when isolates in the nasopharynx spread to the normally sterile sites such as the ear, lung, blood, and central nervous system. Depending on the infected organ, pneumococci can cause two types of diseases: (i) non-invasive (mucosal) pneumococcal diseases such as otitis media and sinusitis, and (ii) invasive pneumococcal diseases (IPDs) such as pneumonia, bacteremia, and meningitis[8]. People with a weak immune system, such as infants, the elderly, and immunocompromised patients, are at a higher risk of infection.

The pneumococcal genome is highly diverse, and only a small fraction of genes is conserved between all strains. The isolates can exchange their genetic material with each other or with other species through bacterial recombination. Change in the pneumococcal habitat could promote the use of different combinations of genes by isolates to diversify their genome. The pathogen can respond to environmental stresses such as drugs and vaccines efficiently. The *pneumococcus* can acquire resistance to several antibiotics and switch its serotype to avoid the vaccine-induced immune response.

Due to the high rate of recombination and genetic exchange among the pneumococcal isolates, they have developed a number of virulence factors that can be categorized into three main groups: (i) pneumococcal capsule, (ii) pneumococcal surface, and (iii) secretory proteins. These factors are vital for the survival of *pneumococcus* in the nasopharynx and sterile sites[9,10].

1.1.1. Pneumococcal capsule

An essential pneumococcal virulence factor is its capsule that forms the outermost layer of the cell (Figure 1-1). It is 200-400 nm thick[11], making it more than half of the cell volume, and requires a considerable investment of energy to be synthesized. The capsular polysaccharide (CPS) is composed of saccharide repeating units such as Glucose, Ribitol, Glucoronic acid, Galactose, and Rhamnose that have

two to eight residues[12]. In most pneumococcal strains with the exception of serotype 3, the CPS is covalently attached to the peptidoglycan cell wall[13]. The pneumococcal serotype or serovar refers to the immunologic or serologic property of the capsule that is characterized by the type and the order of the monosaccharides and the type of glycosidic linkages between them in the capsule structure. Pneumococcal strains are serotyped based on their reaction with specific antisera. Those strains that do not react with any of the available antisera are called non-typeable strains.

In most cases, they are non-encapsulated strains that do not express the CPS; they can also be those strains that express a CPS with a novel and undetermined structure. To date, more than 95 distinct pneumococcal serotypes have been identified[14], and only a small number of serotypes account for most IPD. Some serotypes have dual behavior. For example, serotype 11A is harmlessly carried in young children[15] while it causes IPDs among adults[16,17]. It is known that the level of virulence and resistance to antibiotics differ among different pneumococcal serotypes[18,19]. The virulence of each serotype relates to its capacity to resist phagocytosis[9]. Pattern recognition receptors that are expressed on the surface of macrophages, such as Toll-like receptors, exhibit differences in their affinity for different serotypes[20]. The less immunogenic serotypes are often associated with virulence[21].

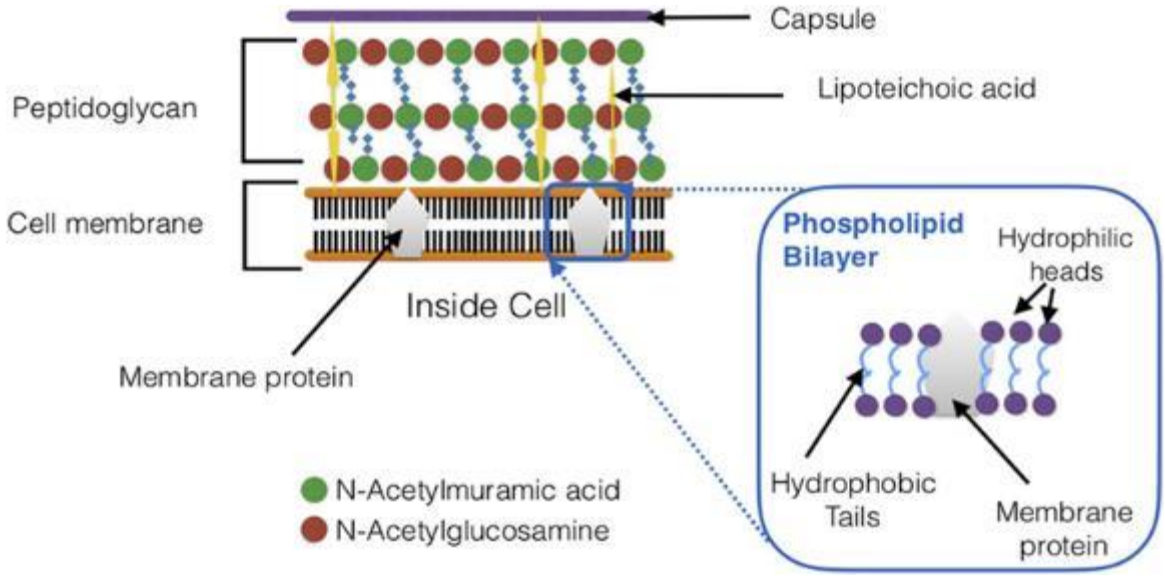


Figure 1-1. The layers that surround a *pneumococcus* cell.
The outermost layer is the polysaccharide capsule.
Image from [22]

Pneumococcal isolates change the capsule expression in different steps during colonization (capsular phase variation). Upon entering the nasopharynx, isolates increase the CPS expression, which is negatively charged and would be repulsed by the sialic acid found in the mucopolysaccharides; therefore, the capsule reduces entrapment in the nasal mucus. However, they reduce the CPS expression at the epithelial surface to expose their surface proteins on the cell wall beneath the capsule and promote adherence to the host epithelial cells[9]. Non-encapsulated pneumococci adhere to the respiratory epithelial cells better than encapsulated isolates[23]. They account for 15% of nasopharyngeal samples[24]. During disease, pneumococcal isolates enter the body's sterile sites and increase capsular expression. The thick capsule prevents immunoglobulins from interacting with the pneumococcal surface proteins. The negative charge of CPS is also important here, as it interferes with interactions with the host phagocytes[25].

Although the CPS aids the bacterium in evading the immune response during colonization and invasion[26], it is immunogenic. This capsule property was used to develop the pneumococcal conjugate vaccines (PCVs) that contain serotype-specific antigens attached to a carrier protein. The PCV causes antibody secretion and provides serotype-specific protection. Different versions of PCVs have been developed[27]. They cover the majority of invasive serotypes that are associated with the IPDs. The 7-valent vaccine PCV7, which covers seven serotypes, was licensed in 2000; it was succeeded by PCV10 and PCV13, which have been on the market since 2009. PCV13 protects against thirteen invasive serotypes 1, 3, 4, 5, 6A, 6B, 7F, 9V, 14, 18C, 19A, 19F, and 23F. In November 2011, PCV13 was introduced in *Malawi*[28]. Although the introduction of PCV has significantly reduced the burden of diseases caused by vaccine types (VT), it caused serotype replacement and has increased the carriage rate and disease incidence of non-vaccine types (NVT) [29,30]. A survey conducted in the *Karonga* district showed that PCV13 reduced VT carriage while a moderate level of serotype replacement has been observed[31]. To overcome the problem of serotype replacement, an efficient strategy is to include conserved pneumococcal virulence factors such as surface proteins in the vaccine.

The CPS expression is regulated by a cluster of genes in the *cps* locus closely linked in a cassette-like arrangement and transcribed together as an operon. The existence of different serotypes refers to genetic diversity in the *cps* locus[32]. The pneumococcal CPS can be assembled through two mechanisms: wzy-dependent assembly and synthase-dependent assembly. Except for serotypes 3 and 37, which have a different structure in their cps loci, other capsule types are assembled by the wzy-dependent mechanism. Serotypes 3 and 37 are assembled by the synthase-dependent mechanism. Serotype 37 is synthesized by a single gene called *tts*. The *cps* locus in both Wzy-dependent and Synthase-dependent is located between two genes *dexB* and *aliA.* However, *tts* is outside the region between *dexB* and *aliA*[12].

Genes involved in the wzy-dependent pathway, along with their functions, are illustrated in Figure 1-2. The first four widely conserved genes are *cpsA (wzg), cpsB (wzh, manganese-dependent phosphotyrosine-protein phosphatase), cpsC (wzd, membrane protein),* and *cpsD (wze, autophosphorylating protein-tyrosine kinase).* These genes have regulatory functions in the CPS synthesis. *CpsC is* required for *cpsD* phosphorylation, whereas *cpsB* is required to dephosphorylate *cpsD*. Genes in the central region of the *cps* locus (*wchA*, *wchJ*, *wchK*, …) are specific glycosyltransferases that synthesize serotype-specific oligosaccharides on the cytoplasmic face of the membrane. However, serotype 1 uses the first glycosyltransferase *wchA (cpsE)* for teichoic acid synthesis[12]. After these genes, a flippase called wzx transports the oligosaccharide repeat units onto the external face of the membrane. Thereafter, a polymerase called *wzy* links the units together to form the final capsular polysaccharides that will be attached to the cell wall. The wzy-dependent pathway is also called the wzx/wzy-dependent pathway.

Genes involved in the synthase-dependent pathway and their functions are illustrated in Figure 1-3. In this mechanism, a single enzyme *wchE (cpsS)* adds sugars to a lipid acceptor, then it translocates the polysaccharide chain to the external surface of the cell membrane and increases the length of the polysaccharide outside the cell. The final polysaccharide is released and inserted into the CPS structure. As stated earlier, this mechanism is used by serotype 3.

In general, for both the synthase-dependent and wzy-dependent pathways, the CPS synthesis is remarkably influenced by mutations in the *cps* locus. For instance, mutations in *cpsD* that inactivate its ATP-binding site can stop CPS production. After the deletion of *cpsB*, the amount of phosphorylated *cpsD* increases significantly and consequently decreases CPS synthesis[33]. The deletion of *cpsA* also

reduces CPS synthesis. In general, the dephosphorylated *cpsD* promotes CSP synthesis[34,35]. A single mutation in *wcrL* converts serotype 11A to serotype 11D[36]. Disrupting mutations in *galU* and *pgm* can terminate CPS synthesis[37,38]. *WcjE,* located at the end of the wzy-dependent *cps* locus, can harbor mutations that promote serotype switching[39].

As mentioned previously, some pneumococci are *non-typeable (NT*) because they do not have a capsule and are called non-encapsulated *Streptococcus pneumoniae (NESp*) strains. These strains are divided into two groups: group I are those that have the *cps* locus on the chromosome, but mutations in the *cps* locus inactivated the CPS synthesis machinery, and group II are those that lack the entire *cps* locus[40]. All non-encapsulated strains in the nasopharynx belong to group II. They seem to be able to compensate for the lack of the capsule and colonize the nasopharynx along with encapsulated isolates[41]. Strains from group II are further divided into three null capsule clades (NCC) based on the presence of *aliC*, *aliD*, and *pspK* (*nspA*):

- NCC1: $aliC^{-}$, $aliD^{-}$, $pspK^{+}$,
- NCC2: $aliC^{+}$, $aliD^{+}$, $pspK^{-}$,
- NCC3: $aliC^{-}$, $aliD^{+}$, $pspK^{-}$.

AliC and *aliD* are homologs of permease *aliB.* These three genes replace the genes in the *cps* locus and can be transferred between encapsulated and non-encapsulated isolates through the recombination of flanking homologous genes *dexB* and *aliA*[40]. After the introduction of PCVs, the carriage rate of NESp strains has been increased as they are not targeted by vaccines[42]; however, they rarely cause IPDs and in most cases, they just colonize the nasopharynx or cause mild pneumococcal diseases such as otitis media. Nonetheless, since a thick capsule can inhibit the exchange of genetic materials, NESp strains can contribute to pneumococcal genetic recombination, which is crucial for spreading antibiotic resistance genes and other virulence factors[43]. Several studies described a cycle of encapsulation and un-encapsulation among pneumococcal strains. Mutations in the *cps* locus can cease the capsule expression in the nasopharynx (group I). Lack of capsule enables the bacterium to acquire virulence and resistance genes from other isolates, however, homologous recombination leads to encapsulation again (group II or new serotype), and the resistant isolate will invade the body's sterile sites[44,45].

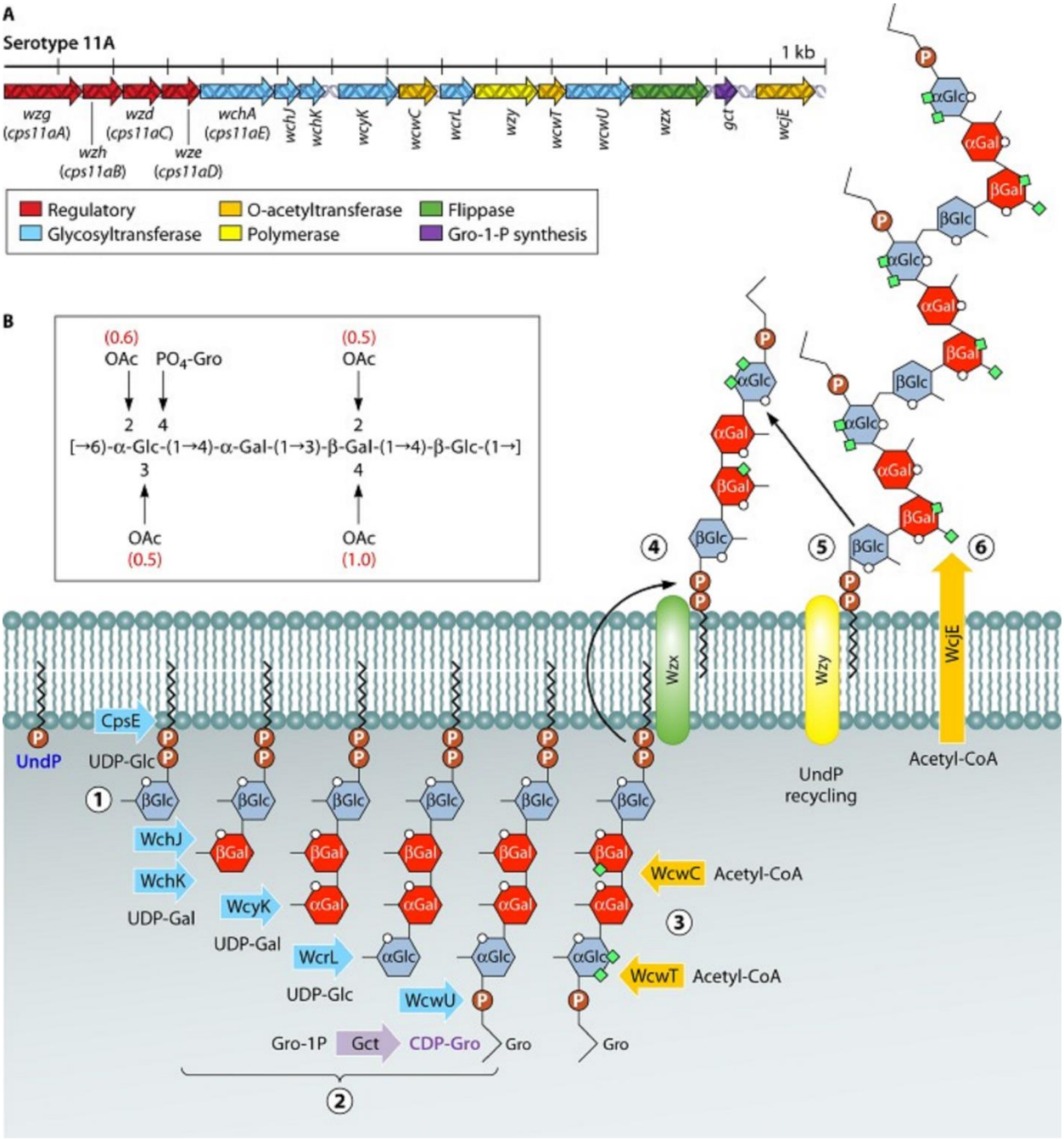


Figure 1-2. Wzy-dependent pathway for serotype 11A synthesis.
A) The order of genes in the *cps* locus. The *cps* locus is located on the chromosome between *dexB* and *aliA.* B) CPS biosynthesis. *cpsA, cpsB, cpsC,* and *cpsD* have a regulatory function. They are not shown but work together to influence the function of *wzy*. Image from[12].

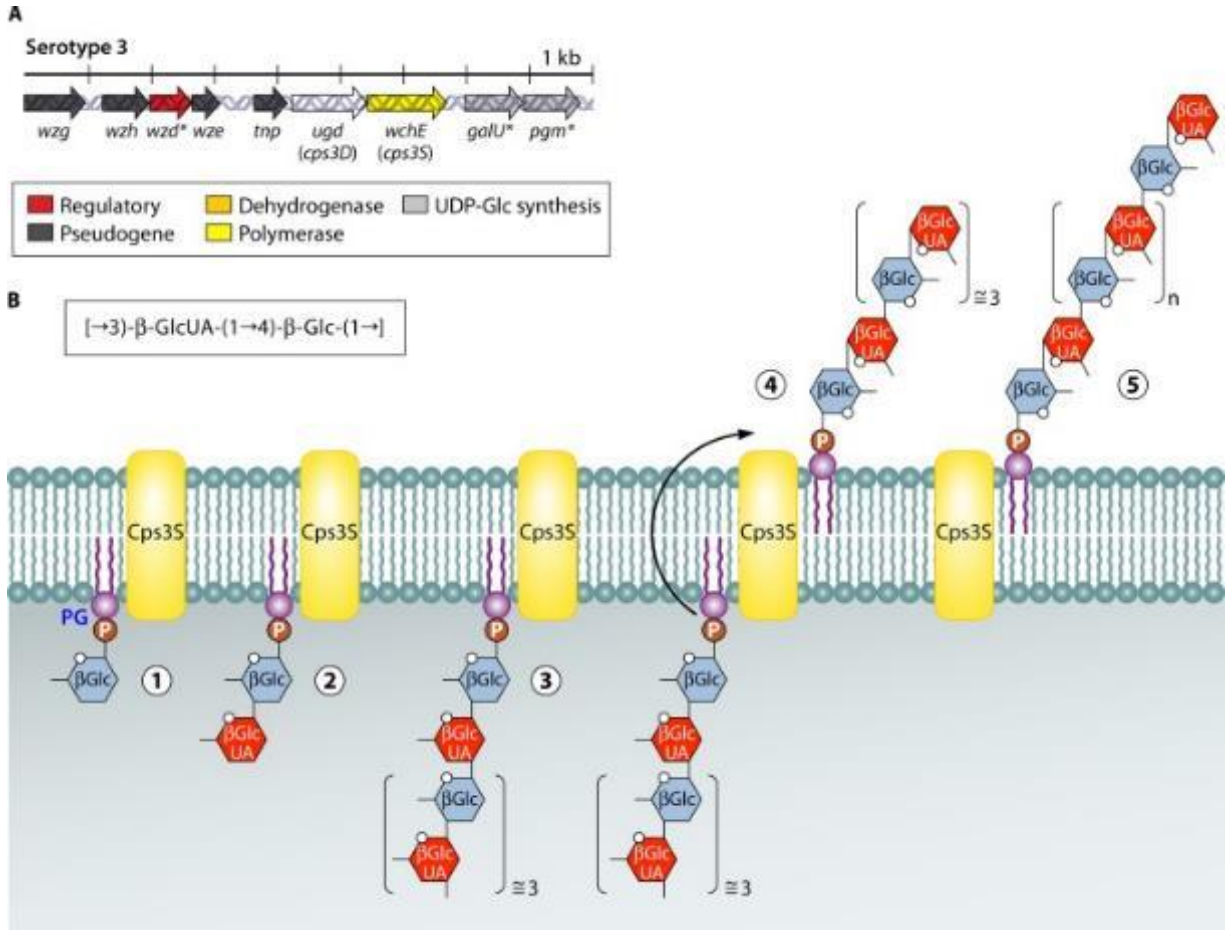


Figure 1-3. Synthase-dependent pathway for serotype 3 synthesis.
A) The order of genes in the *cps locus*. The *cps* locus is located on the chromosome between *dexB* and *aliA*. B) CPS biosynthesis. *cps3D* and *cps3S* are required for capsular synthesis, while starred genes *wzd*, *galU*, and *pgm* are unnecessary. Cps3S assembles the polysaccharide on the internal surface of the membrane. It transfers the chain to the external space of the cell, and extends it until the polysaccharide is released.
Image from [12].

1.1.2. The pneumococcal surface and secretory proteins

In addition to the capsule, many proteins promote pneumococcal colonization and pathogenesis. They assist the bacterium in interacting and adhering to the nasal epithelial cells and connective tissues, compete with other pathogens, get resistant to antibiotics, and evade the host immune system. These compounds can be secreted to the surrounding environment or expressed on the surface of the cell.

Pneumolysin *(ply)* is the secretory pneumococcal pore-forming toxic protein. It promotes pneumococcal transmission and pathogenic effects such as inflammatory responses, the induction of apoptosis, and cellular and tissue damage[46]. It is common to all serotypes and produced by almost all clinical isolates[47]. *Ply* is a cholesterol-dependent cytolysin that is synthesized in the cytoplasm and released from the cell[48], it can bind to cholesterol molecules found in the membrane of the target cells[49]. Since pneumolysin does not have a secretion signal, its release requires lysis of the cell, catalyzed by *N-acetyl-muramoil-1-alanine amidase (lytA)*[50]. Further, when *pneumococcus* is lysed within the phagosome, the released pneumolysin forms pores on the phagosome wall, allowing the pneumococcal debris to enter the cytoplasm of the host cell and cause cytokine production and an inflammatory response [51,52].

The dominant antibody in the human nasopharynx is immunoglobulin A1 (IgA1)[53]. To limit the performance of the host immune response, *pneumococcus* secretes an enzyme called the *zinc metalloprotease (Zmp)*. This protease cleaves the IgA1 molecules that have attached to the *pneumococcus* surface antigens. *Zmp* removes the FC region from antibodies and prevents the host immune system from detecting the antibody-tagged antigens on the surface of the pathogen.

As mentioned earlier, pneumococcal surface proteins are the ideal candidates for vaccine development to protect against a wide variety of strains. There are three categories of surface proteins: *lipoproteins*, choline-*binding proteins (CBPs)*, and the *LPXTG motif-containing proteins*.

The most important pneumococcal lipoproteins are *putative proteinase maturation protein A (PpmA)*, *streptococcal lipoprotein rotamase A (SlrA)*, *pneumococcal surface antigen A (PsaA), pneumococcal iron acquisition A (PiaA), and pneumococcal iron uptake A (PiaU)*. *PpmA* and *SlrA* are surface-exposed lipoproteins that are immunogenic and play roles in colonization and virulence[54,55]. They are involved in the secretion and activation of cell surface compounds[56]. *PsaA* is a part of the *psaABC* operon, it is the substrate-binding of an ATP-binding cassette (ABC*)* transport system that can bind to divalent manganese and zinc ions[57]. These ions are required for the expression of other adhesins and resistance to oxidative stress. Mutated *PsaA* causes a deficiency in the transport of these ions and reduces the pneumococcal adherence to nasopharyngeal epithelial cells[58]. Mutations in other genes of the operon (*PsaB* and *PsaC*) result in similar effects on the adherence[59]. *PiaA* and *PiuA* are parts of *pia* and *piu* operons respectively, they are the substrate-binding components of the ABC transporter systems that uptake iron. *PiaA* and *PiuA* promote pneumococcal virulence[60] and immunization with both *PiaA* and *PiuA* leads to protection against infection[61].

Phosphorylcholine (ChoP) is a compound found in the membrane and cell wall of *pneumococcus*[62]. *ChoP* is also found in the structure of the human *platelet-activating factors (PAFs*) that bind to the *receptors for platelet-activating factors (rPAF)* on the epithelial cells in the nasopharynx. Therefore, *ChoP* on the surface of the *pneumococcus* facilitates its adherence to the nasal epithelial cells through binding to *rPAF*. Another critical function of the *ChoP* is anchoring pneumococcal *choline-binding proteins (CBP)* to the cell wall. These proteins bind to the *ChoP* on the cell wall by a conserved choline-binding domain. The most important pneumococcal choline-binding proteins are *pneumococcal surface protein A (PspA), pneumococcal surface protein C (PspC), also known as choline-binding protein A (CbpA)*, and *pneumococcal autolysin (LytA)*. *PspA* protects *pneumococcus* from opsonization and the host antibacterial agents like lactoferrin. *PspC* attaches to the immunoglobulin receptors on the epithelial cells, decreases complement deposition on the pneumococcal cells, and influences the complement-mediated immunity to *Streptococcus pneumoniae*[63]. *LytA* contributes to the cell wall growth and release of pneumolysin[9].

Several pneumococcal surface proteins contain the LPXTG motif in which X can be any amino acid. They are building blocks of the pilus, and up to 20 of them are anchored to the peptidoglycan cell wall by the LPXTG motif [56]. The *sortase* enzyme encoded by *strA* assembles LPXTG proteins in the pilus structure and attaches them to the cell wall[64]. *StrA* contributes to colonization and invasive pneumococcal disease[65,66]. *NanA* is a conserved pneumococcal LPXTG-linked protein. It is a *neuraminidase (sialidase)* enzyme that promotes colonization by removing sialic acid residues from immunoglobulins, secretory components, and glycoproteins on the surface of the epithelial cells. *NanA* has also been reported as a virulence factor contributing to pneumonia and bacteraemia[67]. Other pneumococcal neuraminidases are *NanB* and *NanC* though they do not contain the LPXTG motif.

There are other critical pneumococcal proteins. *Pneumococcal adherence and virulence factor A (PavA)* and *enolase (Eno)* are expressed on the *pneumococcus* surface and attach to the extracellular matrix components in the human pharynx. Lack of *PavA* and *Eno* reduces pneumococcal infection[68,69]. Two enzymes, *peptidoglycan N-acetylglucosamine deacetylase (Pgda)* and *attenuator of drug resistance (Adr),* promote colonization as they alter peptidoglycan so that isolates become resistant to lysozymes that exist in the upper respiratory tract[70].

Hundreds of different microbial species can reside in the human nasopharynx[71]. Competition between these species is inevitable. Pneumococcal isolates compete not only with other species but also with each other in the nasopharynx. They apply two strategies for competition: they stimulate a host immune response to which they are resistant, but this immune response can clear other species, and for the intra-species competition, pneumococcal strains secrete small antimicrobial peptides called pneumococcal bacteriocins or pneumocins that have strain-specific activity and kill other members of the pneumococcal strains.

A comprehensive description of all pneumococcal proteins is beyond the scope of this study. The research would focus on the pneumococcal factors thought to be associated with the virulence of isolates in the Malawian cohort. In addition to the factors mentioned above, the host factors can also influence the level of pneumococcal virulence, as invasive isolates cause a different level of infection across different populations[72].

1.1.3. Pneumococcal epidemiology and mortality rate

In 2005, the World Health Organization (WHO) estimated that every year 1.6 million people, including one million children under five years of age, die of pneumococcal disease[73]. In 2008, globally, about 476,000 children (<5 years) died due to pneumococcal infections, out of which 247,000 deaths occurred in Africa[74]. Despite the reduction in the incidence of pneumococcal disease after the introduction of the PCV, the pneumococcal mortality rate is still high. In 2015, *pneumococcus* was estimated to be responsible for 294,000 deaths in HIV-uninfected children, 23,300 deaths in HIV-infected children, and a total of up to 515,000 deaths worldwide[75]. In 2017, WHO included *S. pneumoniae* as one of the 12 antibiotic-resistance priority pathogens that pose the greatest threat to human health[76]. In the post-PCV era, a high burden of disease and death has been reported in developing countries, mainly Sub-Saharan Africa and Asia. For example, 1,900 pneumococcal-related deaths occurred in South African children in 2012–2013[77], a much higher mortality rate was reported in Malawian children during the same time, 6903 out of 113,154 pneumoniae cases[78]. A study showed that during 2004-2009, the scale-up of national antiretroviral therapy decreased the incidence of invasive pneumococcal disease in Malawi[79].

1.2 Pan-genome and pan-genomics

For most studies in comparative genomics starting with NGS data, a reference genome is required and must be defined for data analysis. This reference genome can be:

- The genome of one strain
- The consensus sequence drawn from all strains
- A comprehensive genome that contains all genetic variants

The remarkable capability of bacteria to adapt to their environment is enabled by their ability to exchange their genetic material by homologous recombination and horizontal gene transfer. It enables bacteria to have a dynamic, adaptable, and diverse genome [80]. This genomic plasticity is even considerable across different strains of the same bacterial species, therefore, a single genome sequence cannot necessarily represent the entire range of genetic variation.

1.2.1 Pan-genome

A pan-genome is actually a type of reference genome that displays all variants, including all possible genes. In 2005 and for the first time, the term *comparative pan-genomics* became official when eight strains of *Streptococcus agalactiae* were compared [81]. Since then, the pan-genome has been defined as the following: *"For a collection of closely related strains, pan-genome is the entire gene set that exists in those strains".*

The pan-genome contains three types of genes according to their availability among strains (Figure 1-4):

- *Core genes* that are present in all strains.
- *Accessory genes* (*dispensable genes*, *variable genes,* or *adaptive genes*) that are present in some strains but not all.
- *Unique genes* (*specific genes*) that are a specific form of accessory genes present only in one strain.

The collection of core genes is called the *core-genome,* and the collection of accessory genes is called the *accessory-genome*. Therefore: The Pan-genome = Core-genome + Accessory-genome. The total gene number in the pan-genome is:

- Total genes = Core genes + Accessory genes
- Pan-genome = Core-genome + Accessory-genome

The genes in the core-genome are often the identity signals and make a species what it is. Core genes, also called *the minimal gene set,* are essential for normal cell functions such as DNA replication, transcription, and translation. The genes in the accessory genome are not necessary for basic life, at least for all conditions that bacteria encounter. The existence of these genes causes some strains to gain specific traits such as virulence and antibiotic resistance or the ability to occupy niche environments. The total number of genes in the pan-genome is usually larger than the number of genes in one single genome.

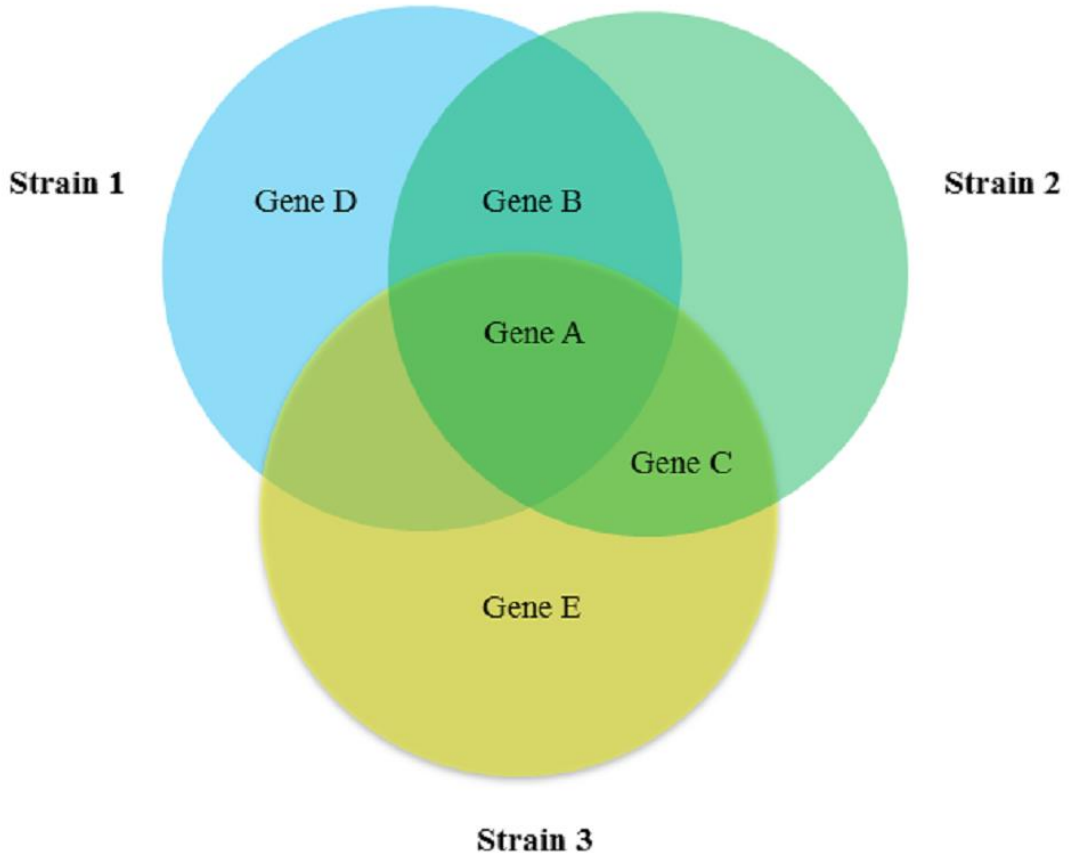


Figure 1-4. Pan-genome of three strains.
Gene A is a core gene as it exists in all strains. Gene B and C are accessory genes because they exist in two strains and not all. Gene D and E are unique genes specific to strains 1 and 3, respectively. This pan-genome has five genes A, B, C, D, and E.

In some species, the total number of genes in the pan-genome does not increase further after adding a certain number of genomes from different strains. This pan-genome that reaches a plateau is called a *closed* pan-genome. Meanwhile, for other species, every new strain adds new genes to the pan-genome. Such species have an *open* pan-genome that does not reach a plateau (Figure 1-5). Species that are dormant and live in an isolated environment often have a closed pan-genome. In contrast, metabolically active species with a diverse genome and horizontally transferred genes have an open pan-genome [82].

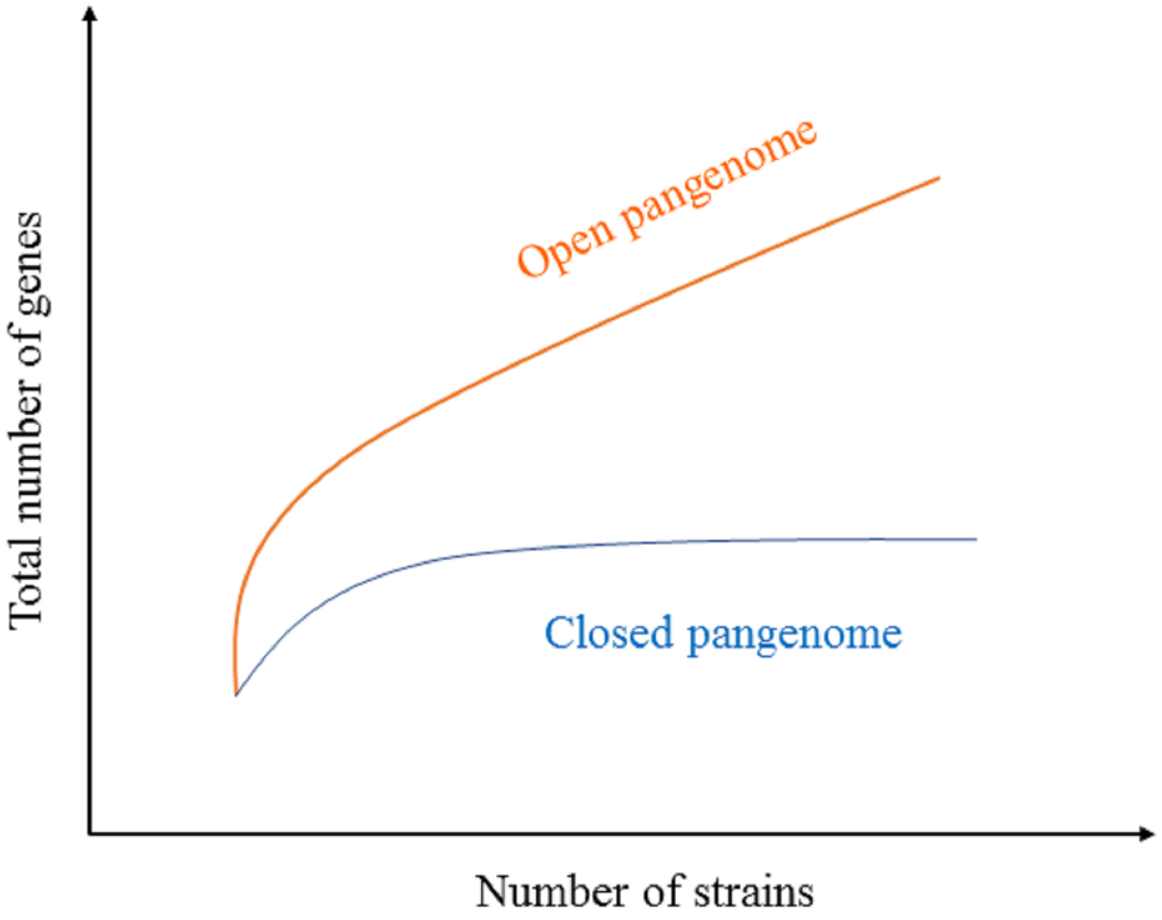


Figure 1-5. Closed and open pan-genome.
In a closed pan-genome, the number of total genes will not increase after adding a certain number of strains. In an open pan-genome, any new strain adds new genes to the pan-genome.

It is worth noting that, although the pan-genome usually has a gene-based definition which refers to the entire gene set existing in different strains of one species, it can also have a sequence-based definition which refers to all sequences found in different strains of one species. The gene-based definition considers variations at gene levels, such as gene presence/absence and gene copy number variations (CNVs), while the sequence-based description is more complete and considers all small-scale variants such as single nucleotide polymorphisms (SNPs), insertion/deletions (Indels), and structural variants in the coding and non-coding sequences. Nonetheless, the fundamental reason for the pan-genome definition is that a single individual genome is unable to show all genetic variants found in the species. Therefore, a pan-genome is a hypothetical combination of variants that do not exist in reality. Thus, a more comprehensive pan-genome definition is:

- ✓ Pan-genome is the entire set of all DNA sequences, including genes and non-coding regions found in individuals of the species.

1.2.2 Computational pan-genomics

The discipline of computational pan-genomics refers to all computational principles applied for pan-genome visualization, statistical analysis, and software development.

1.2.2.1 Pan-genome graphical representation

The main idea in pan-genomics is to replace traditional consensus and linear reference genomes with a pan-genome structure that captures all variants in one species. The two main methods to represent the pan-genome structure are:

1. A Multiple Sequence Alignment (MSA).
2. A graph data structure.

The structure of the pan-genome in Figure 1-4 is visualized in Figure 1-6 by an MSA and a graph.

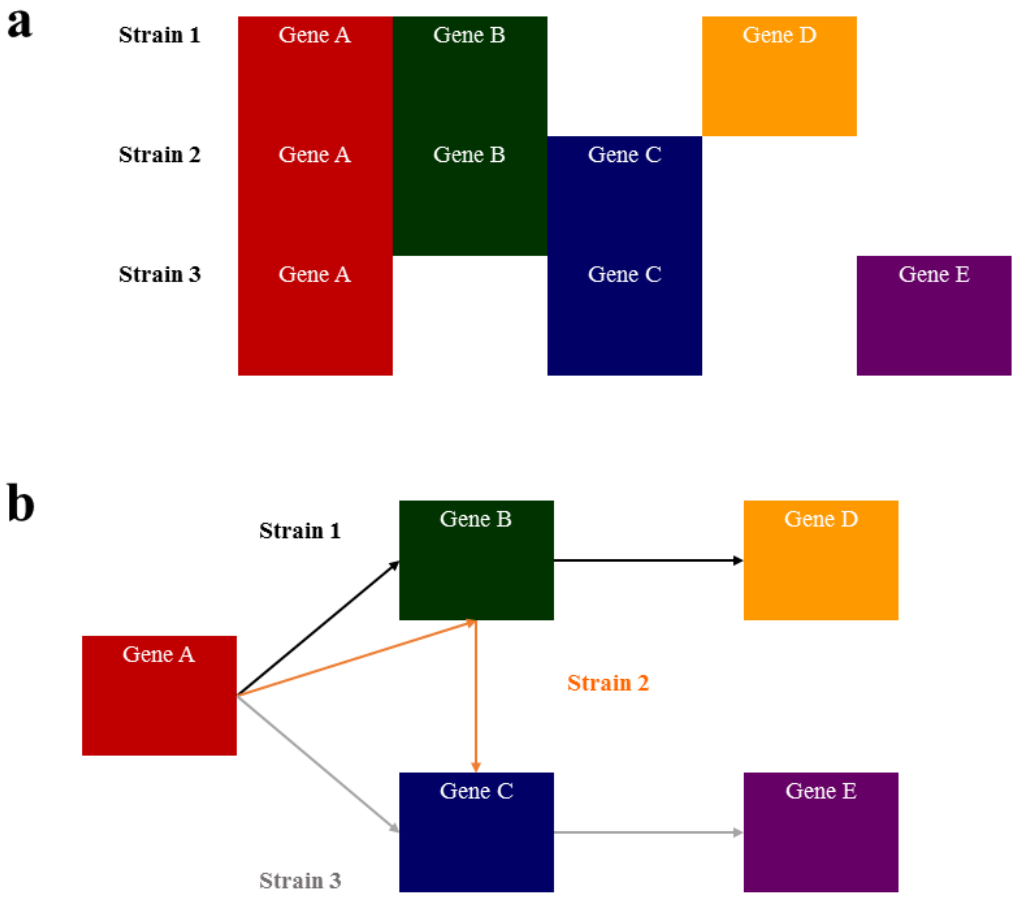


Figure 1-6. Graphical representation of a pan-genome
(a) The pan-genome is shown by an MSA. (b) a directed graph. Each rectangle is a node that represents a gene, and each arrow is a directed edge that joins two nodes. Each path, which is a combination of edges on the graph, indicates the set of genes present in one strain, the black path shows genes found in strain 1, the orange path for strain 2, and the grey path for strain 3.

Representing a pan-genome as an MSA generates a vast and sophisticated structure, the analysis of which is complicated. The reason is that each gene must be represented as many times as the number of isolates it exists in. For example, in Figure 1-6 (a), gene A is present in 3 strains and Gene B is present in two strains, therefore, Gene A appears three times and Gene B twice. Moreover, although the MSA can spot SNPs and Indels, it is not able to identify gene duplications and chromosomal structural variants. Researchers prefer graph data structures over an MSA for almost all pan-genome analyses.

In molecular biology, the terms *graph* and *network* are used interchangeably [83]. Before explaining how to represent a pan-genome as a graph data structure, it is helpful to briefly introduce the basic definitions in the *graph theory*.

Graph *G(V, E)* is a set of nodes or vertices (*V*) that are joined by a set of edges (*E*), as shown in Figure 1-6 (b). An edge that joins two nodes *u* and *v* is an *incident* on them and is denoted by *(u,v).* Two nodes joined by an edge are called adjacent nodes, and two edges joined by a node are called *adjacent edges*. The edge *(u,u)* that joins the node *u* to itself is called a *loop*. A *proper edge* is an edge that is not a loop. Two or more edges that join the same two nodes are called *multi-edges.* A complete graph is a graph where every pair of nodes is joined by an edge. A *directed edge* is an edge with a specified direction that joins a start node called a *tail node* to an end node called a *head node.* The head node is the *successor* of the tail node, and the tail node is the *predecessor* of the head node. If all edges in a graph have direction*,* the graph will be called a *directed graph* or *digraph.* A *walk* is a way of getting from one node to another through a sequence of edges. A *path* is a walk in which no vertex appears more than once. The length of the shortest path between two vertices is called the *distance* between them. A walk visiting every edge exactly once is called the *Eulerian walk* [84].

To design a graph that is able to represent all genetic variations in the pan-genome, the genome sequences of different strains are split into their substrings called *k-mers* which are subsequently arranged in a graph. A k-mer is a substring of a certain length *k* (k is a natural number) that can be obtained from the sequence *S* of length *n* while 1 <= *k* < *n*. The substring from position i to j is shown as S[i:j]. Therefore, any k-mer of sequence S is defined as below:

k-mers = S[i : i+k-1] (Inclusive) (1 <= i <= n-k+1) & (1 <= k <= n)

The total number of k-mers will be *(n-k+1).*

Example:

Sequence: S ="CGCTGAGCT"

Example of substrings: S[1:4] = "CGCT" , S[2:4] = "GCT", S[5:5] = "G"

Sequence length: n = 9

K-mer length: k = 3

Kmers = S[i:i+k-1] = S[i:i+2] (1<= i <= 7)

A total number of k-mers of length 3 (3-mers) obtained from sequence S of length 9:

N – K + 1 = 9 – 3 + 1 = 7

All possible 3-mers obtained from S, note that there is a repeat of "GCT":

3-mers = ["CGC", "GCT", "CTG", "TGA", "GAG", "AGC", "GCT"]

The k value is very important here and must be selected carefully. It depends on the genome length, the available computing resources, and the type of analysis. To count all k-mers in a sequence, many space-efficient algorithms have been developed, examples are *Disk Streaming of K-mers* [85], *K-Mer Counter* [86], and *Squeakr* [87].

All k-mers derived from sequence *S*, are arranged in a directed graph called a *de Bruijn graph (DBG)* denoted by *G(S,k).* This graph contains a node for each distinct k-mer of S, and two nodes *u* and *v* are connected by a directed edge *(u,v)* if:

u = S[i: i+k-1] and v = S[i+1: i+k]

The gene content of each strain is illustrated by an Eulerian walk on the graph (Figure 1-7).

To save memory and space, the DBG is compressed by merging its nodes to produce a *compressed DBG*. Two nodes *u* and *v* are allowed to be merged into a single node if:

Node u is the only predecessor of node v, and node v is the sole successor of node u. There may be multiple edges between them.

In a compressed DBG, every node (except the start node) has at least two different predecessors or its single predecessor has at least two different successors and every node (except the end node) has at least two different successors or its single successor has at least two different predecessors. A compressed DBG can be constructed by identifying maximal exact matches using a suffix tree [88], or more efficiently by a combination of FM index, compressed suffix tree and Burrows-Wheeler transform [89].

The kmer-based representation of the pan-genome in a graph has many advantages, such as simplicity, speed, and robustness. It is not always necessary to use fixed-length k-mers as the pan-genome can be arranged in acyclic and cyclic graphs [90]. The pan-genome graph is able to highlight all genetic diversities found in a species, from SNPs and Indels to gene presence and absence. It renders a compact graphical portrait of the pan-genome that characterizes the variants among individuals. Moreover, graph-based pan-genomics provides access to retrieve data and define a suitable coordinate system. It highlights the variable and conserved regions across the genomes. All genes are represented only once on the graph (no matter in how many strains they are present) and each strain is characterized by an exclusive walk on the graph (Figure 1-7).

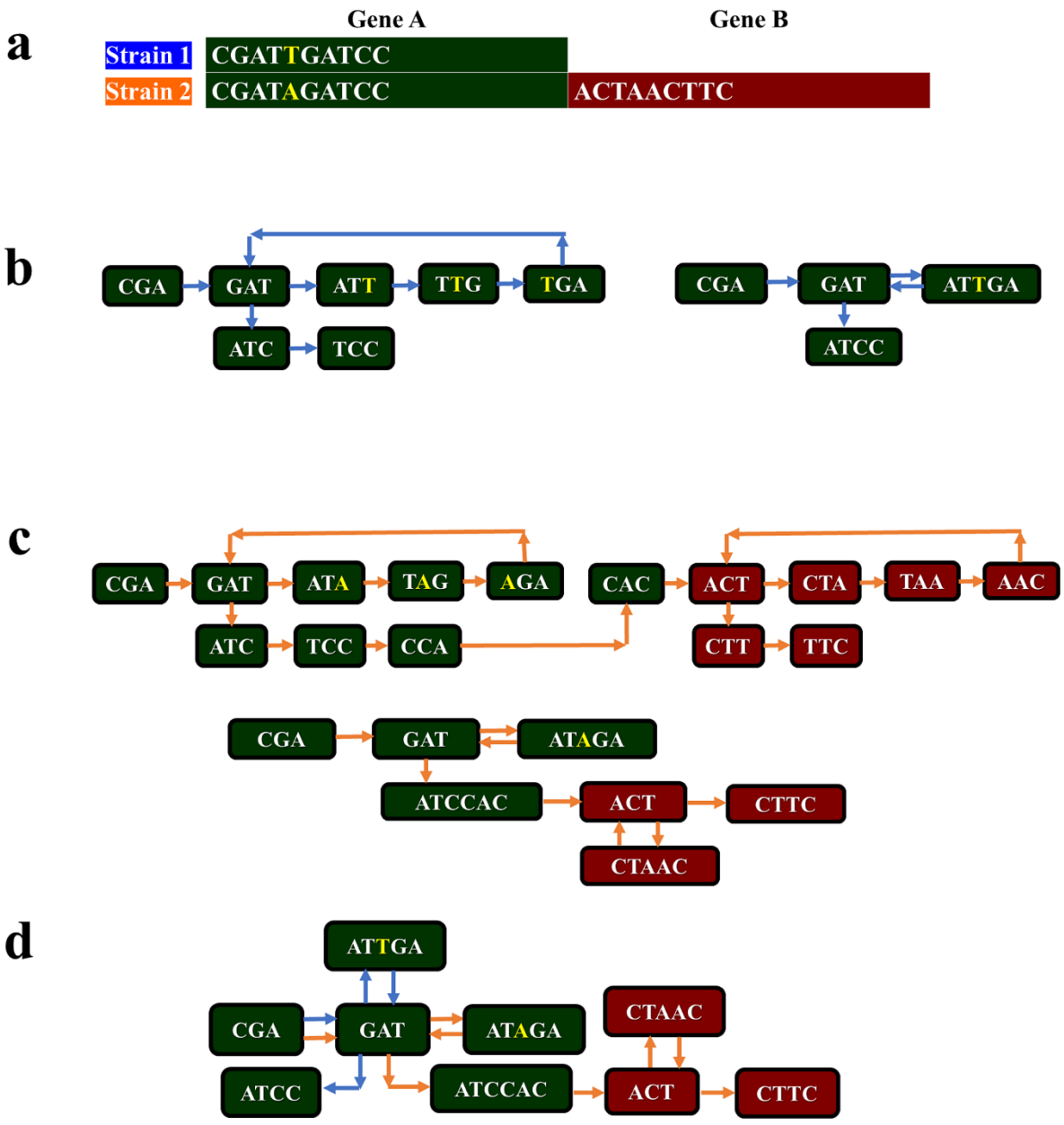


Figure 1-7. Applications of graphs in pan-genome visualization.
(a) Genomes of strains 1 and 2. Gene A shown in green, is a core gene and Gene B shown in red, is an accessory gene. There is an SNP on position 5 of Gene A highlighted in yellow. (b) The genome of strain 1 is shown by its 3-mers in the form of a DBG on the left, and a compressed DBG on the right. (c) The genome of strain 2 is shown by its 3-mers in the form of a DBG on the top, and a compressed DBG on the bottom. (d) The pan-genome of both strains is shown as a compressed DBG, the blue Eulerian walk indicates the genome of strain 1, and the orange Eulerian walk indicates the genome of strain 2, both variants in the form of SNP and gene presence/absence are identifiable on the graph, the yellow letter indicates the SNP, the green nodes show Gene A and the red nodes show Gene B.

1.2.2.2 Pan-genome computational analysis

High performance and parallel computing are necessary for pan-genome computational analysis, particularly when a high number of samples are involved in the study. The computational pipeline often needs significant RAM and storage space.

A bacterial pan-genome analysis starts with a set of *whole genome sequencing (WGS)* short reads obtained from several closely related strains, preferably from the same species. The pipeline for a pan-genome analysis has four main steps (Figure 1-8):

- Reads quality control, pre-processing, and cleaning.
- Genome assembly and annotation.
- Pan-genome construction.
- Pan-genome downstream analysis.

1.2.2.2.1 Reads quality control, pre-processing, and cleaning

The sequencing short reads are stored in standard file formats like *fastq* [91]. The quality of the reads is evaluated by tools such as *fastqc*[92], the adapter sequences are trimmed, and sequences with low quality are removed by tools like *FASTX toolki* [93]. The clean sequences that are of high quality are supplied to the genome assemblers.

1.2.2.2.2 Genome assembly and annotation

The next step is to assemble the genomes of all strains. Typically, a pan-genome analysis is useful when working with a bacterial species whose genome is highly divergent across different strains. Defining a linear reference genome for such a diverse species is difficult. Thus, *de novo assembly* [94,] which is reference-free, is desired in computational pan-genomics. The genome assembly can be achieved by a number of publicly available tools such as *VelvetOptimiser* [95] and *SOAPdenovo* [96]. For a successful pan-genome analysis, the assembled genome should be of high quality, and contigs should have a minimum length of 500 base pairs. Tools like *Quast* [97] can be used to evaluate the quality of the assembled genomes.

To define the pan-genome and determine core and accessory genes, all assembled genomes must be annotated coherently with a tool that is compatible with the pan-genome builder. Accurate assembly and annotation produce a pan-genome with high quality enabling a productive analysis. Annotated genomes are saved in standard file formats, such as *BED, GFT, GFF, and GFF3*. Many tools have been developed for bacterial gene prediction and annotation, examples include *Glimmer* [98] and *Prokka* [99]*. Prokka* is specifically designed for prokaryotic genome annotation and works based on the integration of several tools and databases such as *SignalP* [100], *Aragorn* [101], *HMMER3* [102]*, Rfam* [103] and *Infernal* [104]*. Prokka* is a well-run annotator that produces its outputs in various file formats, at least one of which will be compatible with one of the tools used for pan-genome construction. For annotation from scratch, *Pannotator* [105] is suitable, and to improve the available annotations, *eCAMBer* [106] and *Mugsy-Annotator* [107] can be applied.

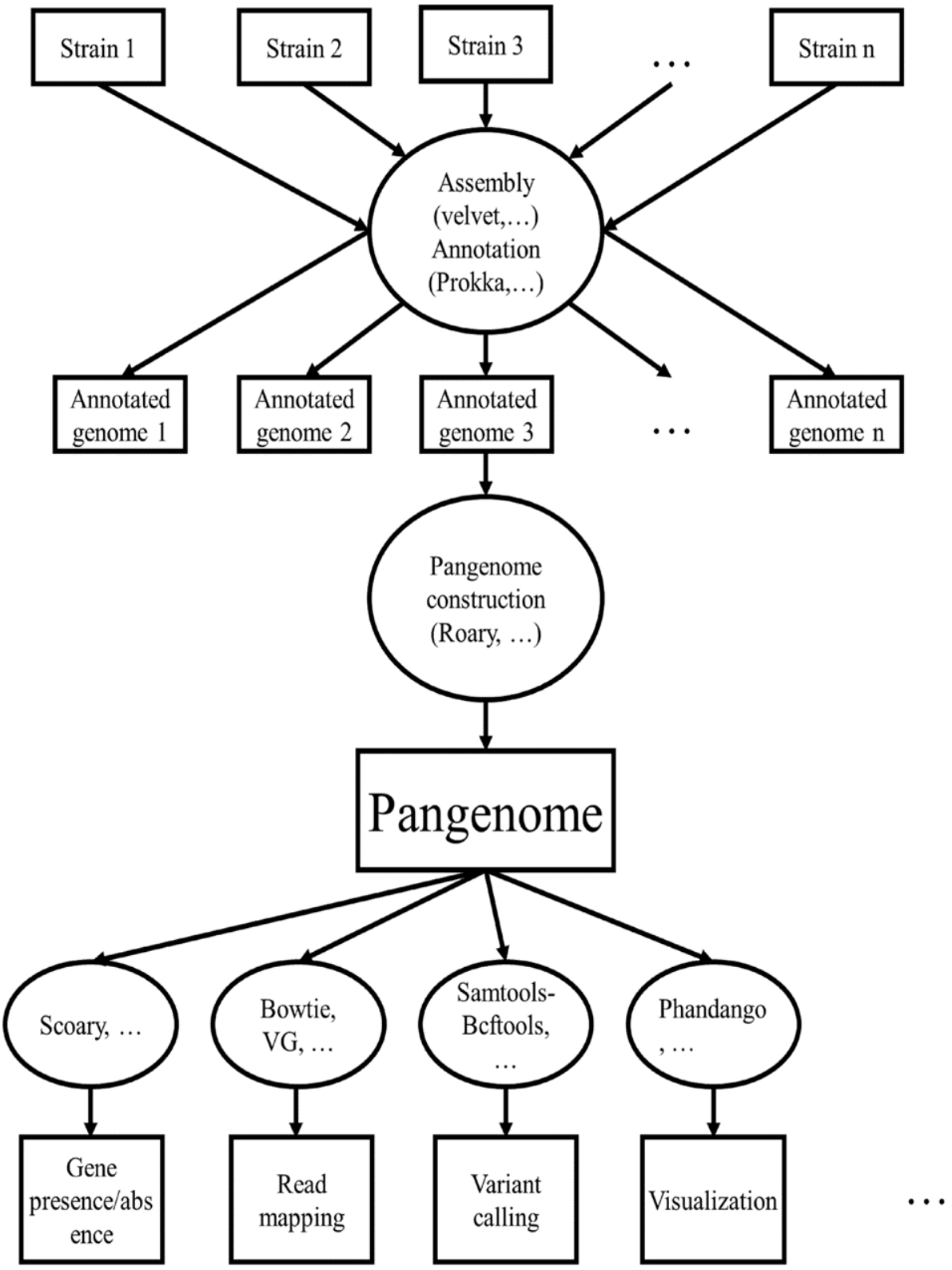


Figure 1-8. Overview of pan-genomic analysis.

1.2.2.2.3 Pan-genome construction:

As discussed earlier, the pan-genome can be defined either as a collection of genes or genome sequences from multiple strains of one species. For this reason, two types of tools have been developed for pan-genome construction and analysis [108]:

1. Gene-based tools.
2. Sequence-based tools.

To use the gene-based tools, all genomes must be annotated, and the gene content of each strain must be determined. These tools first use graph-based methods to assign orthologous genes found in strains and then construct the pan-genome. Some of the most popular gene-based tools developed so far are *EDGAR* [109], *PGAT* [110], *PGAP* [111], *PanOCT* [112], *GET_HOMOLOGOUS* [113], *PanFunPro* [114], *ITEP* [115], *PanGP* [116], *LS-BSR* [117], *Roary* [118], *micropan* [119], *piggy* [120], *BPGA*, and *pyseer* [121].

For a sequence-based pan-genome analysis, the sequences of different genomes are indexed. To increase efficiency in terms of required time and memory, graph-based methods are applied. DBG is usually employed here as the analysis does not require a reference sequence or alignment. Examples of sequence-based tools are *Panseq* [122], *Harvest* [123], *SplitMEM* [88], *TwoPaCo* [124], and *Bloom Filter Trie* [125].

1.2.2.2.4 Pan-genome downstream analysis

Most of the tools mentioned above are able to perform some downstream analysis. The downstream analysis includes tasks such as multiple sequence alignment of the core-genomes, phylogenetic tree construction, alignment of the short reads to the pan-genome, variant calling, studying of genes in different metabolic pathways, pan-genome visualization, and various statistical analysis.

The multiple sequence alignment of core-genomes is produced by the Pan-genome builder, this alignment is then used to extract the variant sites in the core-genomes, which are used to draw an initial phylogenetic tree. This tree can be colored according to the sample phenotypes and provides an overview of the association between samples from different phenotypes. *Snp-sites* [126] is a tool that is useful to extract all variant sites from the multiple sequence alignment of all the core genes. The phylogenetic tree can be drawn by considering only those variant sites in the core-genome of the different strains. However, if sufficient computing resources are available, the phylogenetic tree can be drawn directly from the alignment of the core-genomes. Tools like *Clustalw* [127] and *Fasttree* [128] are appropriate for this tree drawing. The tree coloring and visualization can be performed with the help of tools such as Evolview [129].

The software *Scoary* [130] was developed to score genes in the pan-genome for their association with an observed trait. This tool finds genes whose presence or absence are strongly associated with a phenotype and considers the influence of the population stratification. *Piggy* [120], on the other hand, is a tool that evaluates the variation in intergenic regions in bacteria. Apart from genes, the presence or absence of some intergenic regions affects the phenotypic behavior of the bacterium. Both *scoary* and *piggy* can use the output of *Roary* as their input.

To perform a pan-genome-based GWAS, the sequence of the genes in the pan-genome can be utilized as a reference to identify SNPs and Indels in each strain and determine whether they are in the core genes or accessory genes. In this case, the pan-genome is usually saved in a Fasta file in which each record represents the consensus sequence of a gene drawn from the entire population, then the short reads

are aligned to this reference sequence and variants are called. The SNPs in the core genes reflect the age of the species. To reduce the analysis workload, certain informative SNPs should be selected in a process called representative SNP selection [131]. As explained earlier, the pan-genome can be saved as a graph. Several tools have been developed to align short reads directly to the pan-genome graph, examples are *BGREAT* [132] and *VG*[1]. The pan-genome graph can also be used for reference-free variant calling [133]. For further details about the tools, their algorithms, and performance, refer to [134], [135] and [108]. Many scripts written in *R* and *Python* are available for pan-genome visualization, some of the more versatile tools for this are *Phandango* [136]*, Panx* [137]*,* and *Panviz* [138]*.*

1.3 Project motivation and scope

Prescription of antibiotics and the introduction of vaccines are the major strategies to either treat or prevent pneumococcal disease. However, pneumococcal isolates can develop resistance against antibiotics and evade the vaccine-induced immune response by modifying their genome structure through recombination. For both treatment and vaccination, the pneumococcal virulence factors must be identified. This study applied a pan-genome and GWAS approach on the whole-genome sequencing (WGS) short reads of 1477 pneumococcal samples collected from Malawi. Pneumococcal strains isolated from the nasopharynx of carriers were compared to those from the blood and cerebrospinal fluid (CSF) of patients. The **main objective of the research was to identify potential virulence genes significantly associated with invasive *Streptococcus pneumoniae***. The hypothesis was that the presence of some genes contributes to the invasiveness of pneumococcal isolates. Other factors, such as population structure and serotype of samples, were also presumed to be relevant and therefore were investigated.

This work sought to answer the followings research questions:

1. What serotypes are prevalent in Malawi, and how did this change from 1997 to 2015?
    a. How was this impacted pre- and post-vaccination?
    b. Which serotypes are the most invasive in Malawi?
2. What is the difference between the carriage and invasive isolates in terms of:
    a. Differences in core genes (SNPs and indels)?
    b. Differences in accessory genes and intergenic regions?

To answer these questions, the following tasks were accomplished:

1. In-silico identification of serotypes.
2. Determination of the effect of vaccination on serotype distribution using statistical tests.
3. Identification of serotypes with the lowest and highest invasivenesss, which is defined by the overall abundance of serotypes in the entire cohort and their prevalence among patients.
4. Assembly and building of a reference pan-genome for all strains.
5. Alignment of reads to the core-genome and calling of core SNPs.
6. Determination of the population structure based on the distribution of SNPs in the core-genome.
7. Identification of genes with a significant presence in the invasive isolates.
8. Functional enrichment analysis of potential virulence genes.
9. Discussing how genes could potentially contribute to pathogenesis.

Details of these methods and results are described in the chapters that follow.

2 Study cohort, data Quality Control (QC), and serotyping

2.1 Overview

This chapter describes the demographical characteristics and temporal distribution of pneumococcal sampling, along with the methods for sequencing and evaluating sequence data quality. In-*silico* serotyping using the sequence data was performed, and results of the distribution of serotypes across geographical locations and isolation sites are presented. The general computational methods and code available for this thesis are also summarized. The main objectives of this chapter were:

- To curate, pre-process, and prepare the data and scripts for analysis
- To identify serotypes of samples based on sequence data
- To identify the most abundant serotypes in the cohort using a frequency threshold
- To determine the distribution of serotypes among carriers and patients
- To examine the effect of vaccination on serotype distribution

2.2 Methods

2.2.1 Computational methods and codes availability

The computational methods used in this work are described in detail in the relevant chapters. An overview of the workflow of the process followed across the whole study is provided in Figure 2-1 to show how the data for each component were generated or analyzed.

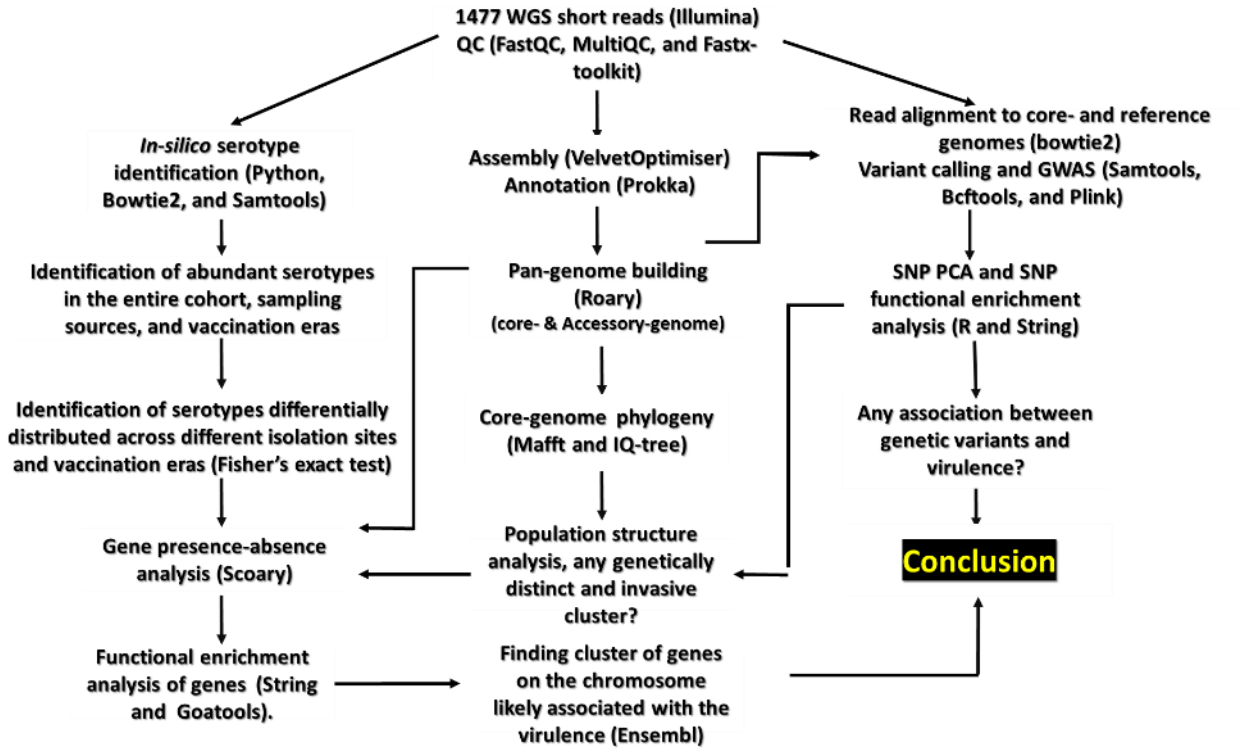


Figure 2-1. The overall data analysis workflow and the name of computational tools used in the research.

For most of the steps detailed in Figure 2-1, some custom Linux and Python scripts were written to automate the processing. Due to a large number of samples and the high memory usage of the pipeline, the data processing elements listed in Figure 2-1 were executed using the computational resources provided by the Centre for High-Performance Computing (CHPC) in South Africa.

The source code used in the study is available in Appendix 1.

2.2.2 Isolation of pneumococcal samples

Pneumococcal samples used in the study were isolated from patients across three hospitals in Malawi: i) Queen Elizabeth Central Hospital (QECH) in Blantyre (Southern Region), ii) Karonga Central Hospital in Karonga (Northern Region), and iii) Kamuzu Central Hospital in Lilongwe (Central Region) (Figure 2-2). The majority (98%) of the samples were collected from patients visiting QECH during 1997-2015 and from households in the Karonga district during 2009-2014 as part of two household carriage studies[139]. A small number of samples were collected from Kamuzu Central Hospital and archived at the Malawi-Liverpool Wellcome Trust Clinical Research Programme Centre and were included in the analysis. For the Karonga samples, the study was conducted in the area covered by the Karonga Health and Demographic Surveillance System (HDSS) in northern Malawi[140].

ation of Karonga, Lilongwe, and Blantyre in Malawi.
Most of the samples were collected from Blantyre in the South (highlighted in red) and Karonga in the North (highlighted in red), which are geographically distant. A small number of samples were collected from Lilongwe, the capital city of Malawi (indicated by the star in the center).
Image from *WorldAtlas*: https://www.worldatlas.com/maps/malawi

2.2.3 Whole-genome sequencing (WGS) and QC

Archived samples were sequenced under the Global Pneumococcal Sequencing (GPS) project at the Sanger Institute in the United Kingdom (GPS :: Global Pneumococcal Sequencing Project (pneumogen.net)) and Pneumococcal African Genomics Consortium (PAGe). The bacterial DNA was

extracted using the QIAamp DNA mini kit, QIAgen Biorobot (Qiagen, Hilden, Germany). The kit is designed to extract mitochondrial, bacterial, and viral DNA. The kit is fully automated and able to remove all contaminants and inhibitors and rapidly extract high-quality DNA. The DNA sequencing was accomplished at the Wellcome Trust Sanger Institute (UK) by Illumina Genome Analyzer II, and HiSeq platforms (Illumina, CA, USA) using 125 paired-end whole genome sequencing runs. The Illumina HiSeq platform enables researchers to sequence many of samples simultaneously. It is cost-effective, and its deep sequencing technology generates data with a high coverage ideal to identify SNPs, indels, and genomic recombination. The final clean data were stored in a pair of FASTQ files containing the forward and reverse strands for each sample. The quality of the reads was evaluated by FastQC[92] and MultiQC[141]. Reads with low quality or contamination were filtered by the Fastx toolkit[93].

After sequencing, the first step of the data analysis pipeline is to ensure that the sequence files are clean and high quality. FastQC was used, which is one of the most well-known and user-friendly software tools for sequence data QC. The advantage of FastQC is that it can be automated on different operating systems (Windows, Mac, and Linux) for thousands of samples and produce several graphs and tables in HTML format. It reports the quality of samples from different perspectives, such as sequencing errors, length of reads, adapter contamination, and sequence duplication. However, FastQC produces one report per sample, therefore, thousands of reports were generated during the research. To create a summary report showing an overview of the quality of the entire dataset, MultiQC was used. The main purpose of using MUltiQC was to represent the quality of many samples in a single report. To clean the short reads with low quality and contamination, the Fastx tool kit was used. This tool is a combination of Linux shell scripts that perform a variety of data cleaning and processing. Sequencing data coming off the sequencing machines are stored in Fastq files. Files with Fastq format contain the base call of short reads and information about the quality of sequencing that refers to the probability of calling a base incorrectly. The chance of inaccuracy is low at the beginning of reads and gets higher toward the end of reads as sequencing machines are able to read a certain length of DNA fragments. The Linux command-line scripts in the Fastx toolkit were utilized to remove barcodes and noise, remove adapters, and filter sequences based on quality.

2.2.3 *In-silico* serotyping

After cleaning the sequence data, Bowtie2 was used to align the short reads to a reference file in Fasta format containing the sequences of the cps locus from different serotypes. Bowtie2 is ultrafast, accurate, memory-efficient, and able to do both local and global alignment. Since the cps sequences in the reference file are short and similar to each other, and in order to involve all characters from the sequencing reads, the global alignment with a sensitive mode was applied. This causes aligners to align reads to a correct cps reference sequence. For each sample, the mean depth of alignment to each cps reference sequence was calculated by Samtools. The depth of alignment is defined by the number of reads aligned to a reference normalized by the length of the reference sequence. Samtools is written in C language and specifically designed to deal with the alignment files. It is fast, memory-efficient, accurate, and user-friendly. All steps stated above were coded and encapsulated into a python script that facilitated the automation, which overcomes the main issue of working with thousands of samples. Furthermore, for each sample, the python script saved the depth of alignment (alignment rate) to each cps sequence in a list and reported the cps sequence with the highest alignment rate as the serotype of each sample.

2.2.4 Assessment of serotype distribution

As described above, the serotype of samples was determined by aligning the short reads to pneumococcal capsular polysaccharide (CPS) reference sequences using Bowtie version 2.3.4.1[142]. The serotype of each sample was identified based on the maximum read alignment rate calculated by an in-house python script and Samtools version 1.7[143], and the frequency of each serotype was then calculated. Any serotype with a relative frequency greater than 5% was considered as an abundant serotype in each isolation site (sampling source).

The difference between the distribution of serotypes across carrier vs. patient groups and pre-PCV vs. post-PCV eras can be investigated using two types of statistical association tests, including (i) the Chi-squared test and (ii) the *Fisher's exact* test. Both methods work by defining a contingency table that holds integer counts of each categorical variable (serotypes) in different groups (carriers vs patients and pre-PCV13 vs. post-PCV13). However, the procedures applied in these methods are different. The Chi-squared test works based on an approximation and is suitable for application on a large number of samples when the frequency values in the contingency table are greater than 20. Comparatively, Fisher's exact test calculates the test's significance precisely without needing a large number of samples. It is suitable when frequency values in the contingency table are small (< 20) for more than 20% of samples[144]. The numbers of some serotypes in the groups defined by the association test, such as carriers/patients and pre-/post-PCV13, were low in the research presented here. Therefore, Fisher's exact test was deemed the better choice for the statistical association test than the Chi-squared test. An important issue that must be addressed here is the problem of multiple testing that occurs when there are a set of hypotheses to be tested simultaneously. Here, 56 serotypes have been identified in the dataset (see results), and one null hypothesis had to be tested for each serotype. When the number of hypotheses increases, the probability of observing at least one significant result just due to the chance increases. Hence, the significance of the results for each test must be adjusted according to the number of hypotheses. To overcome the multiple testing problem, the p-value threshold is reduced (for instance, from 0.05 to 0.01 or 0.001), or the reported p-value is adjusted by methods such as Bonferroni. However, these methods are very conservative and helpful when there are hundreds or thousands of hypotheses. For lower numbers, such as 56 serotypes identified in this research, the Bonferroni adjustment can simultaneously reduce the number of false positives and true positives. Thus, the False Discovery Rate (FDR) method was applied here. This method is less conservative and only considers the significant results for adjustment (not all hypotheses as considered by the Bonferroni). The p-value that is adjusted by the FDR method is also called the q-value. In the FDR method, the *Benjamini-Hochberg (BH)* procedure is applied:

- The null hypotheses are sorted in ascending order according to the unadjusted p-values. For each hypothesis, the value of "$p(i) = i \times \alpha/m$" is calculated where i is the index of the hypothesis in the ascending array, m is the total number of hypotheses, and α is the significant level that is the expected proportion of false discoveries. Any null hypothesis whose p-value is less than $p(i)$ is rejected.

Therefore, Fisher's exact test was applied to identify serotypes whose frequencies changed significantly after introducing PCV13 in 2011. Additionally, statistical tests on the presence-absence of serotypes in the patient and carrier groups were carried out in order to identify which serotypes are common between carriers and patients, which are prevalent only amongst carriers, and which are specifically

frequent in the patient group. The term “invasiveness” for different serotypes was defined based on the significance of their presence in the patient and carrier groups. Each abundant serotype with a significant presence in the patient group was considered as a serotype with a high invasiveness, while abundant serotypes with a significant presence amongst carriers were assumed to have a low invasiveness. Serotypes with the highest significance of presence in the patient group were considered to have the highest invasiveness. Conversely, serotypes with the highest significance of presence in the carrier group were considered to have the lowest invasiveness. The significance level for the assessment of serotype distribution was 0.01 after the p-value was adjusted by FDR.

2.3 Results

2.3.1 Overview of the dataset

The demographical and clinical characteristics of the samples are shown in Table 2-1.

Table 2-1. Demographical and clinical characteristics of 1477 pneumococcal samples collected in Malawi between 1997-2015.

Characteristics	Categories	Nasopharynx	Blood	CSF
Age (in years)	< 5	538	165	141
	5-19	109	42	60
	20-40	60	67	50
	> 40	7	24	7
	Missing	111	70	26
Sex	Female	401	131	111
	Male	313	122	122
	Missing	111	115	51
City	Blantyre	169	357	259
	Karonga	656	0	0
	Lilongwe	0	0	23
	Missing	0	11	2
Year		2009-2014	1997-2015	2000-2015

The temporal distribution of samples based on the sampling sources is illustrated in Figure 2-3. Of note is that all samples from Karonga were collected from the nasopharynx of carriers. Therefore, it was not possible to pursue an in-depth analysis of geographical site serotype distribution (invasive versus

carriage) as these would be biased by body site (source) sampling. Also, to highlight the actual differences between samples from the carriage group to those from the patient group in the following chapters, it was necessary to ensure that there is no difference between nasopharyngeal samples from Karonga and Blantyre and that they can be considered as part of the same cluster.

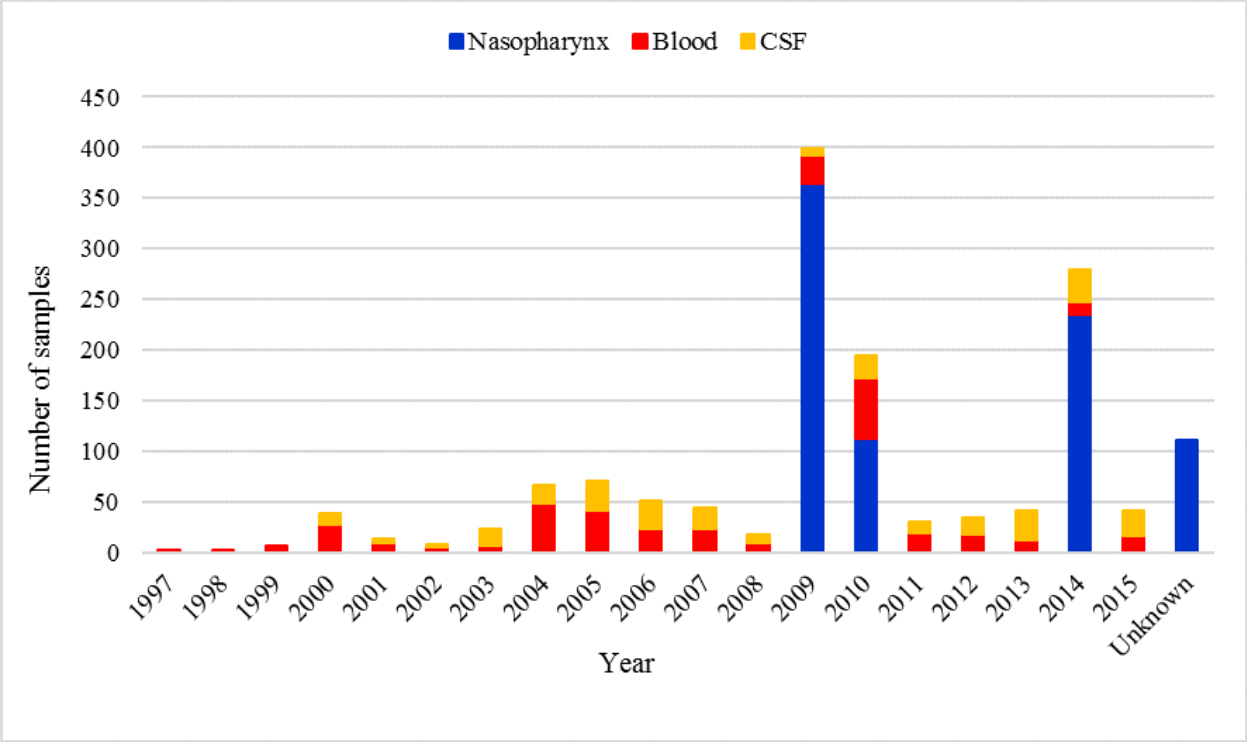


Figure 2-3. Temporal distribution of sampling.
The number of samples isolated from the nasopharynx of carriers, the blood of bacteremia patients, and the cerebrospinal fluid of meningitis patients is indicated in blue, red, and orange.

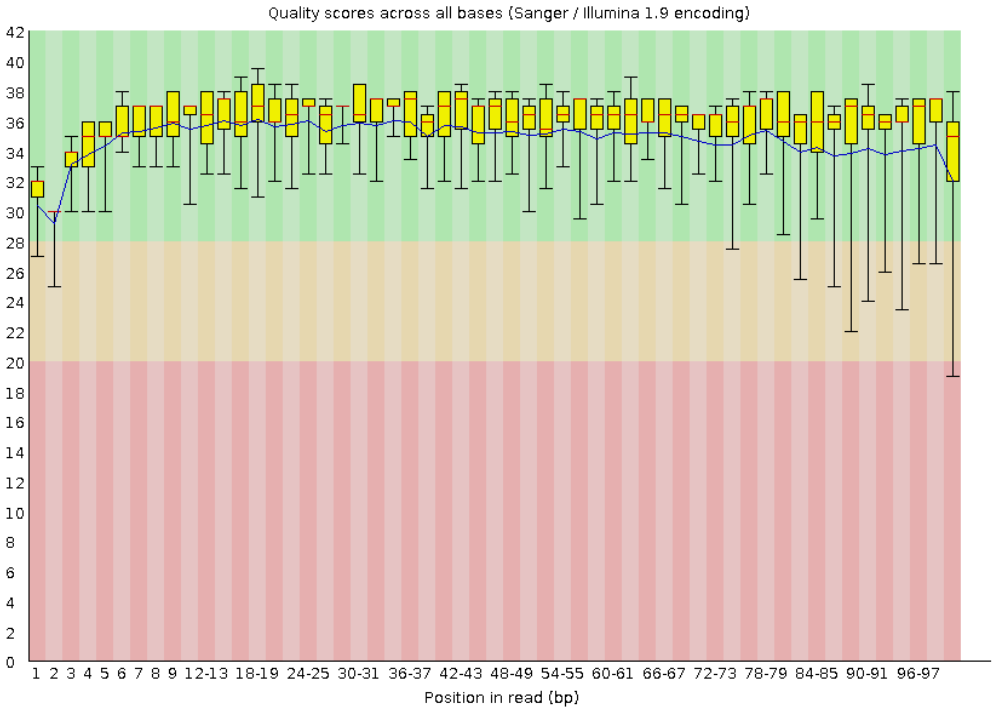
In total, 1477 samples, including 825 samples from the nasopharynx of carriers (56% of the cohort), 368 samples from the blood of bacteremia patients (25% of the cohort), and 284 samples from the cerebrospinal fluid (CSF) of meningitis patients (19% of the cohort) were sequenced. Here, the term "sterile sites" is used to refer to blood and CSF, and "invasive isolates" are those samples obtained from sterile sites.

2.3.2 WGS QC

The quality of the paired-end sequencing short reads was high and acceptable for almost all samples. The QC step included the assessment of per-base sequencing quality (evaluation of the sequencing quality at each position in short reads separately), per-sequence quality score (the overall sequencing quality of short reads), per-base sequencing content (the percentage of A, C, G, and T at each position in short reads), GC distribution over all sequences, per-bases ambiguity content, sequence length distribution, sequence duplication levels, and adapter content. An example of the Fastqc outputs for one

sample with remnant adapter sequences is shown in Figure 2-4. The adapter sequences were trimmed by Fastx toolkit.

a



b

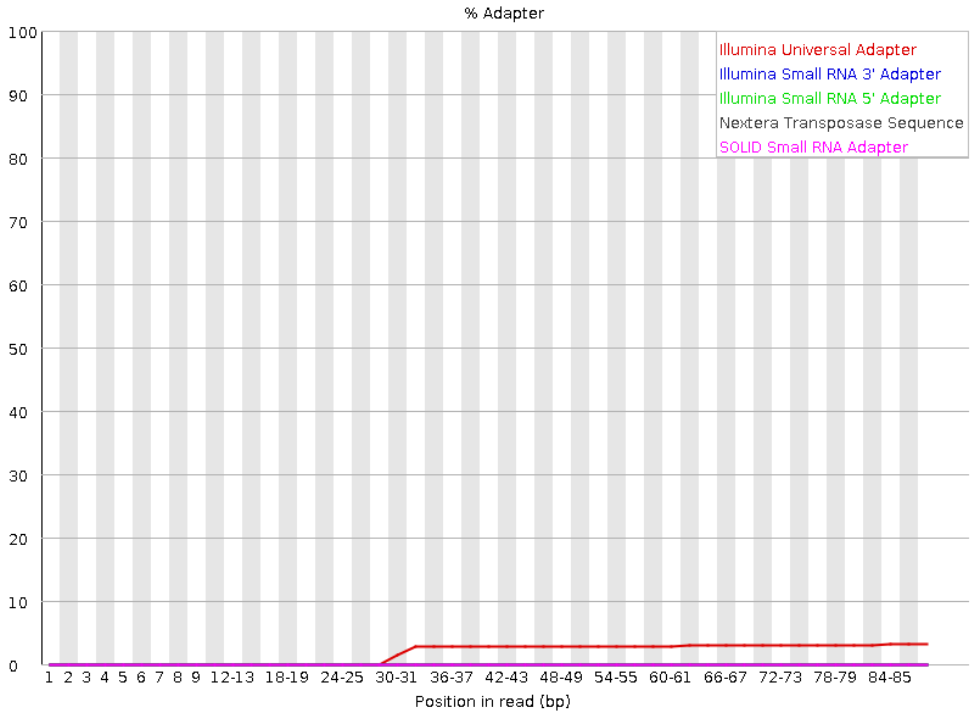
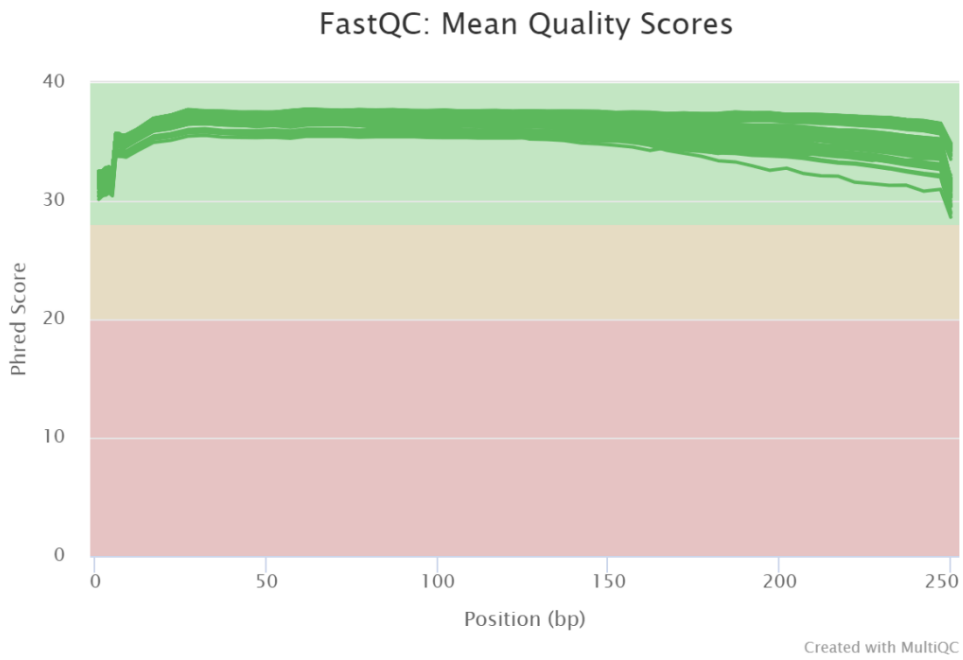


Figure 2-4. The quality of the forward strands obtained from one of the samples from CSF of meningitis patients (ERR316678). The sample contains 36457 reads with a mean length of 100. (a) The vertical boxplots represent the distribution of the sequencing quality at each position in the reads. Most of reads have a good sequencing quality falling in the green area. (b) The remnant of the Illumina adapter is observed in about 5% of reads after position 30.

The aggregated forms of the Fastqc outputs for the entire dataset produced by Multiqc are shown in Figure 2-5.

a



b

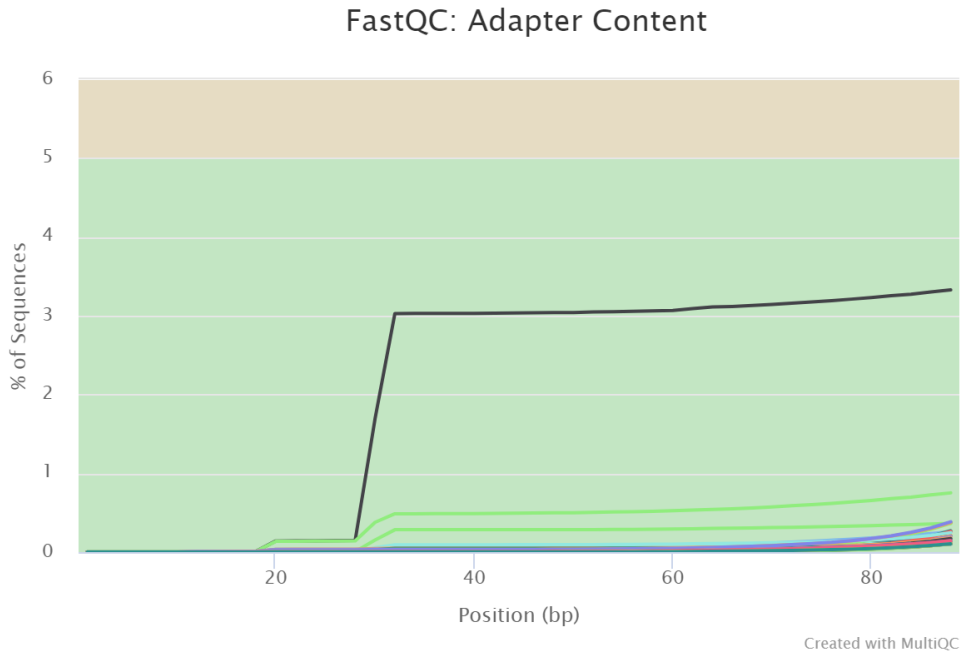


Figure 2-5. Graphs produced by Multiqc summarizing the outputs of the Fastqc for the entire dataset. (a) Each line represents a sample, numbers on the horizontal axis represent positions on the sequencing reads, and numbers on the horizontal axis represent the mean of the sequencing quality at each position. (b) The adapter content found in samples. Few samples have the remnant of the Illumina adapters in less than 5% of reads. The adapter sequences were trimmed from samples using the Fastx toolkit.

2.3.3 In-silico serotyping

In total, isolates within the dataset available comprised 56 serotypes, and irrespective of the sampling sources and collection dates, serotypes 1 (9%), 5 (8%), 6B (7%), 23F (6%), and 19F (6%) were the most abundant serotypes present. 66% of samples were obtained before the vaccine rollout in 2011, and 27% were from the post-PCV13 era. 56% of samples were isolated from the nasopharynx of carriers, 25% from the blood of bacteremia patients, and 19% from the CSF of meningitis patients (Figure 2-6).

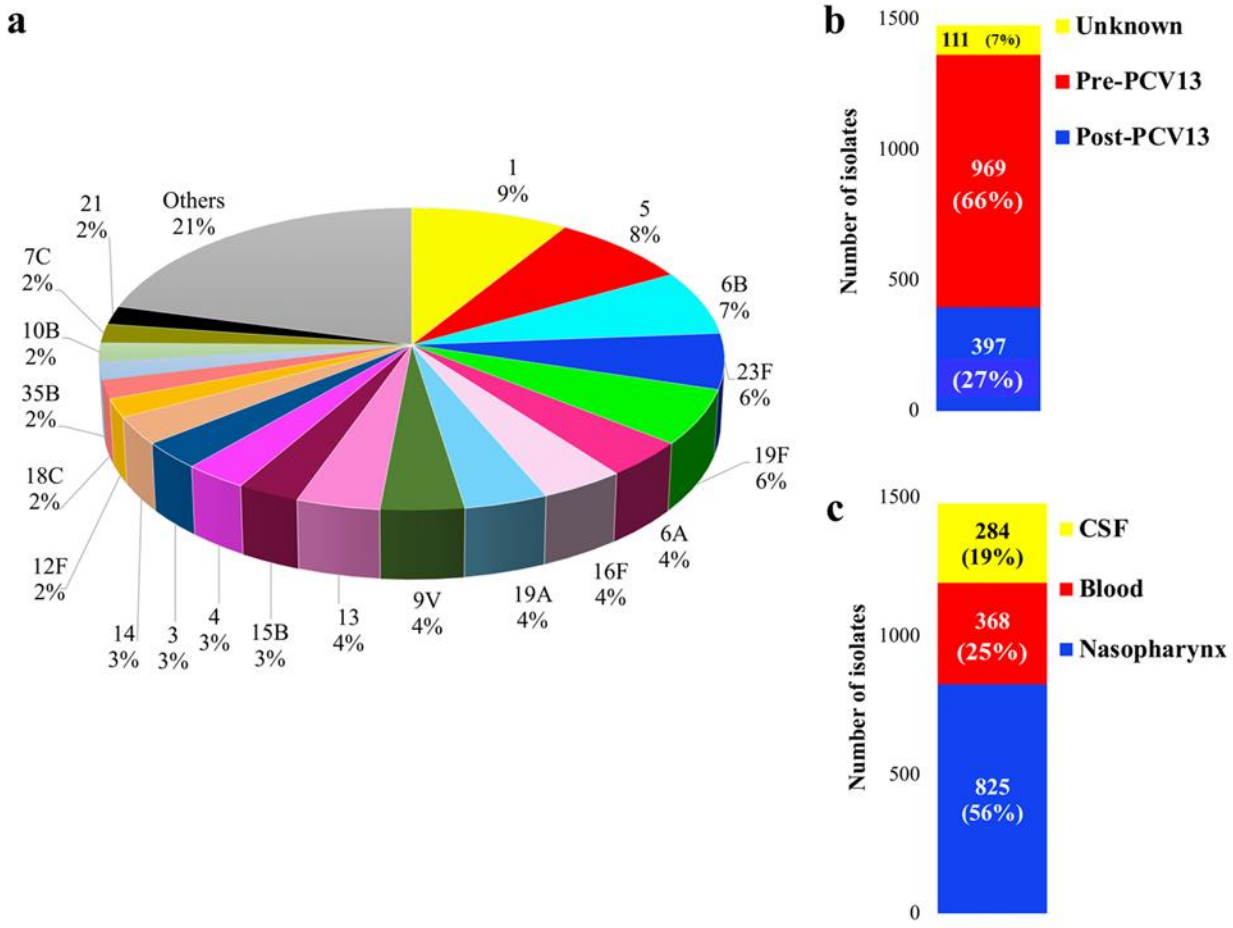


Figure 2-6. Characteristics of the 1477 pneumococcal isolates used in the study.
(a) The relative frequency of serotypes is presented. Samples were assigned to 56 serotypes. For each sample, in-silico serotyping was achieved by aligning its short reads to the pneumococcal capsular polysaccharide (CPS) reference sequences and calculating the maximum alignment rate. (b) Frequency of isolates in the pre- and post-PCV13 eras in Malawi. (c) Frequency of isolates obtained from each sampling source.

The temporal distribution of serotypes in Malawi is illustrated in Figure 2-7. The majority of samples were collected from 2009 onwards. The five most abundant serotypes in Malawi are the vaccine types covered by PCV13. Regardless of the sampling frequency each year, serotype 1 has persistently circulated in the country since 2000. There had also been a continuous presence of serotypes 19F and 23F in Malawi. The abundance of serotype 6B decreased slightly after 2011, however, the reduction in the frequency of serotype 5 is considerable.

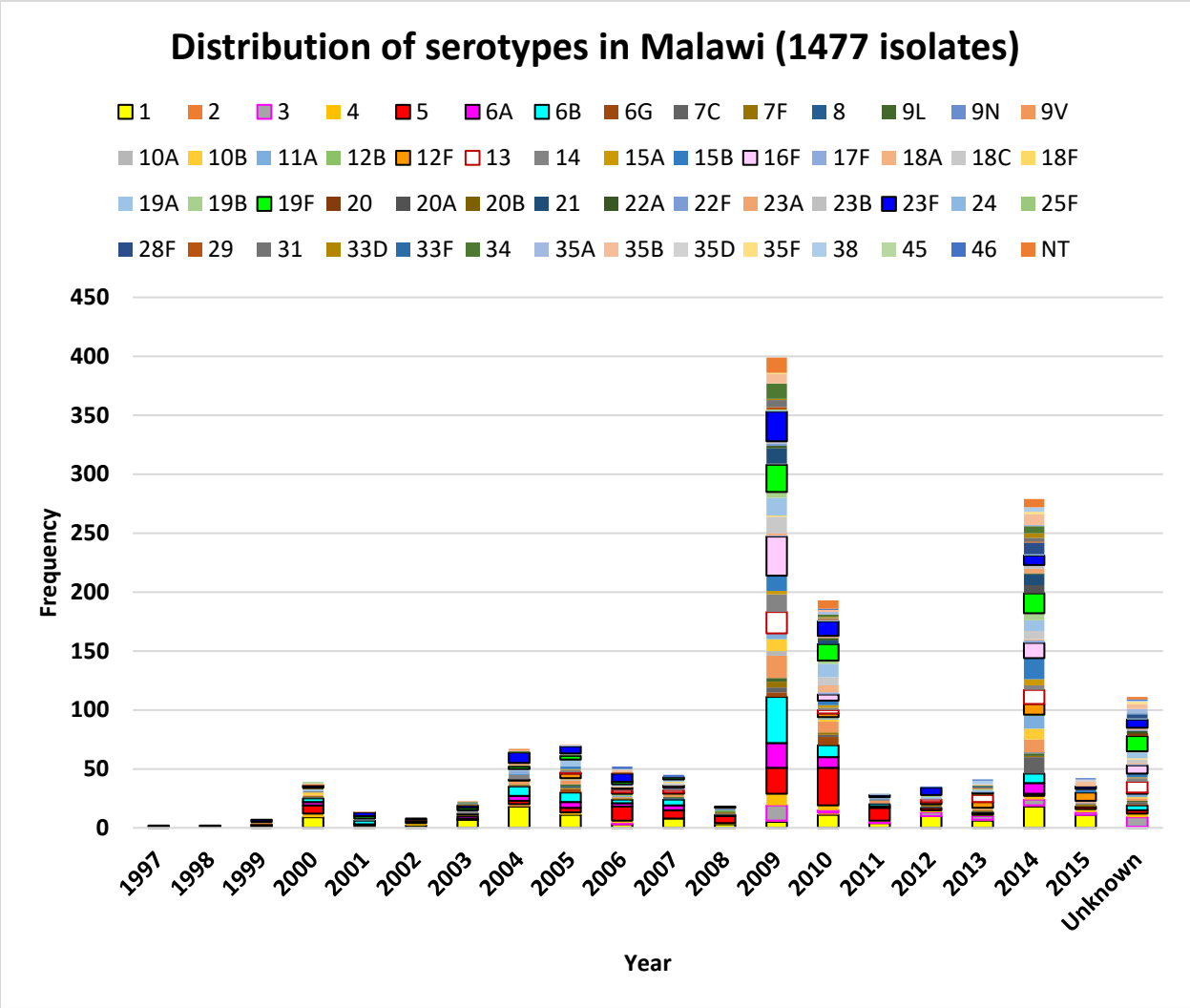


Figure 2-7. Distribution of serotypes in Malawi between 1997-2015.

The distance between Blantyre in the south of Malawi and Karonga in the north of Malawi is approximately 830 kilometers. Given the distance, it was useful to look at the temporal distribution of serotypes in these two cities separately as each of them can be considered as a distinct geographical location. The Blantyre cohort includes isolates obtained from both carriers and patients, while isolates in the Karonga cohort were obtained only from carriers (Table 2-1).

The pattern of serotype distribution in Blantyre is similar to the entire cohort. Continuous prevalence of serotype 1 and reduction in the frequency of serotype 5 after 2011 were also observed in the Blantyre dataset (Figure 2-8).

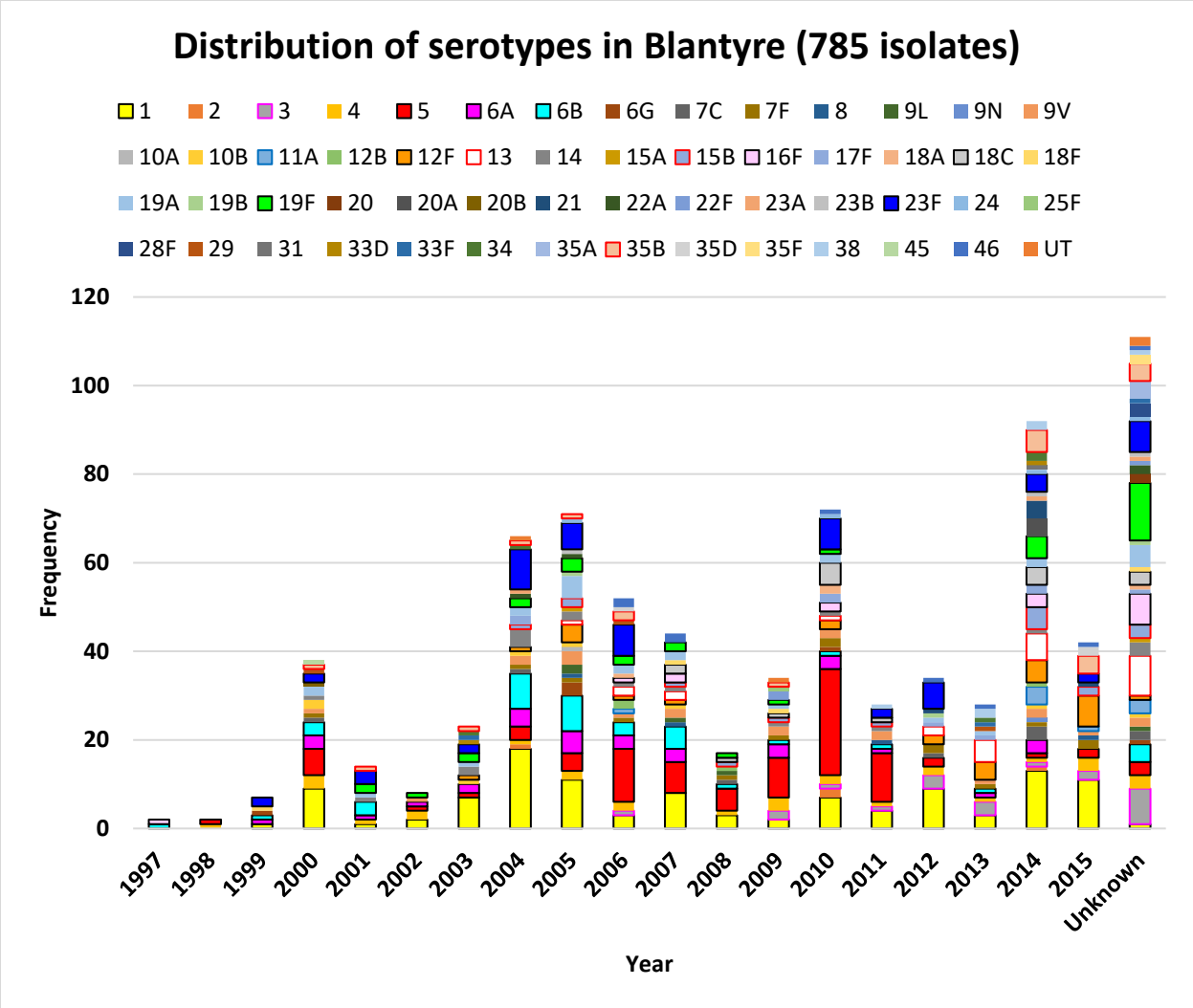


Figure 2-8. Distribution of serotypes in Blantyre in the South of Malawi.

Interestingly, serotypes are not equally distributed amongst carriers and patients in Blantyre. Serotypes 1 and 5 are infrequent amongst carriers in Blantyre (Figure 2-9). The problem here is that the sampling date of isolates from carriers (nasopharyngeal isolates) in Blantyre was only either from 2014 or unknown, which again biases the temporal results of serotype distribution. Serotype 5 was not found amongst carriage samples obtained in 2014, but it was present in the patient samples isolated before 2011, one may assume that this serotype is invasive, and vaccination could have reduced its prevalence in Malawi. Serotype 1 was present amongst patients after 2011 and totally absent from the carrier group. The most plausible explanation for this could be the high virulence of serotype 1 and its ability to escape vaccine-induced immunity.

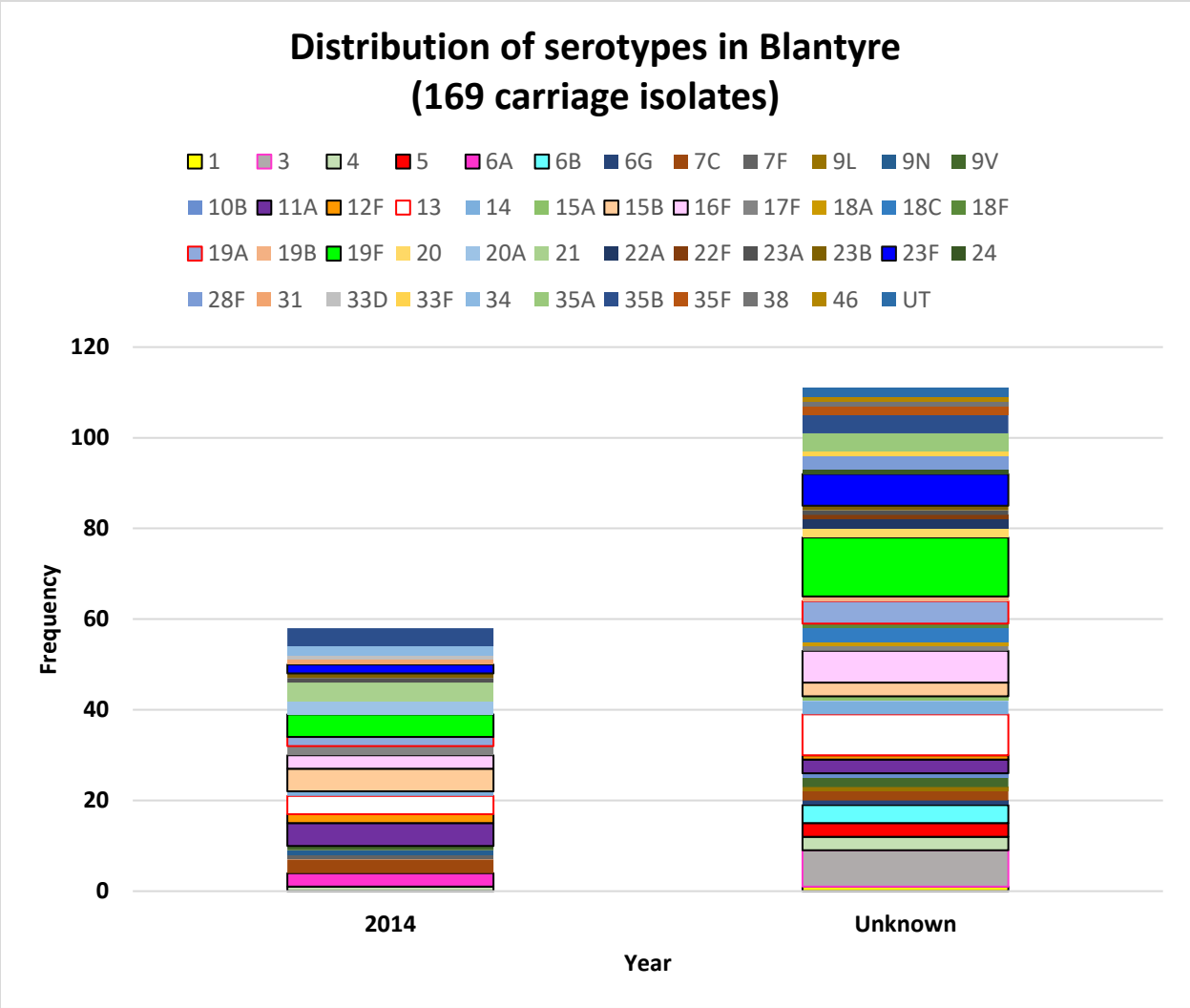


Figure 2-9. Distribution of serotypes amongst carriage group in Blantyre in the South of Malawi.

There were 656 samples from Karonga obtained only from the nasopharynx of carriers (Figure 2-10). These samples were collected from both pre- and post-PCV13 eras. The pattern of serotype distribution in the carriage groups from Karonga is not very different from that of Blantyre. Serotype 5 is only observed in the group of isolates obtained in the early sampling in 2009. Serotype 1 is not present in carriers from Karonga.

There was a small number of samples from Lilongwe isolated only from patients, and all samples were collected after 2011. As with Blantyre and Karonga, serotype 5 is not present after 2011 and therefore is not present in the Lilongwe samples, but serotype 1 is again the most dominant. Another prevalent serotype in Lilongwe is 12F which was also observed amongst patients from Blantyre.

Serotypes 6B, 19F, and 23F are present in both the Blantyre and Karonga cohort and across the nasopharynx, blood, and CSF isolates. To gain a clear understanding of vaccination impact and virulence of serotypes, statistical analysis of serotype distribution was performed next.

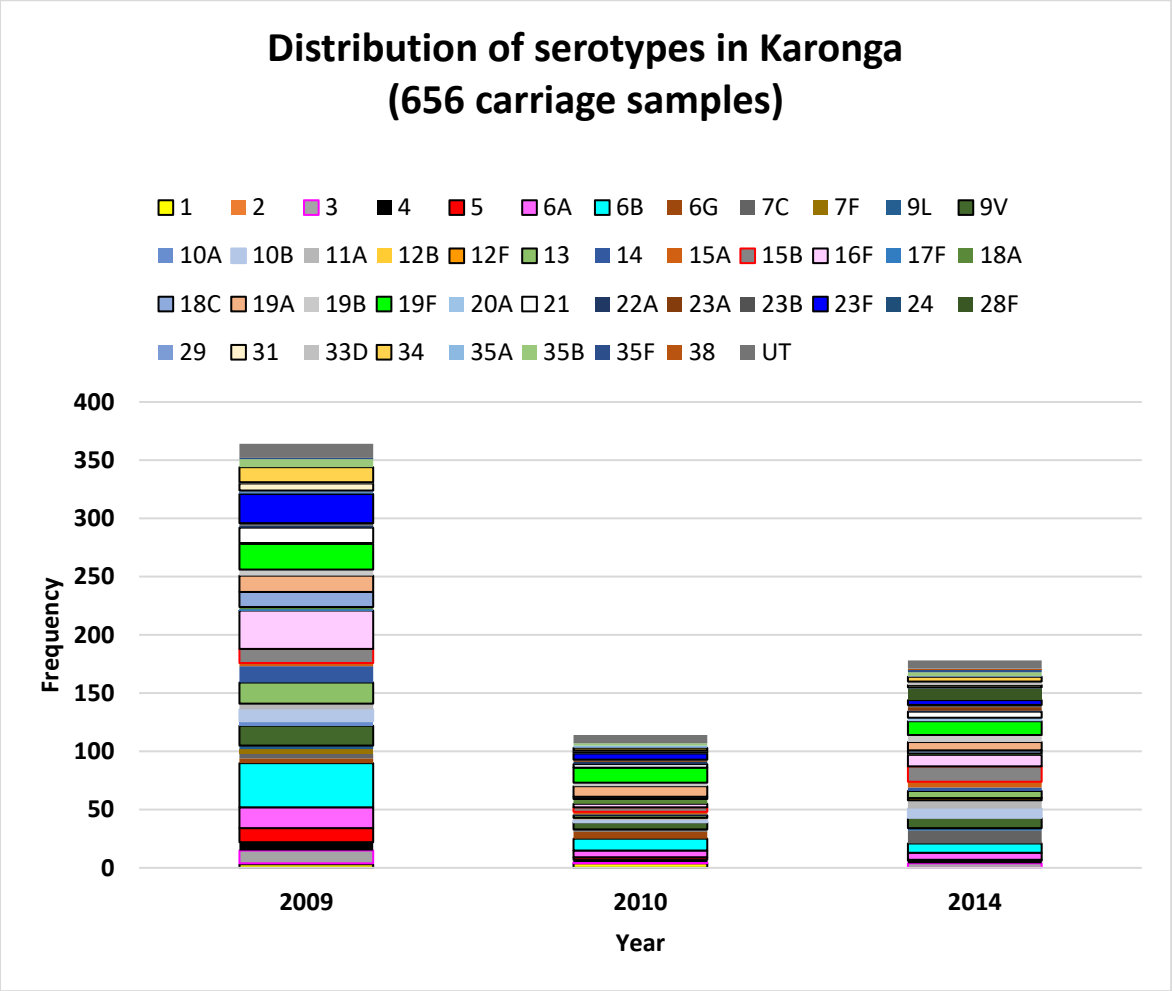


Figure 2-10. Distribution of serotypes in Karonga in the North of Malawi.

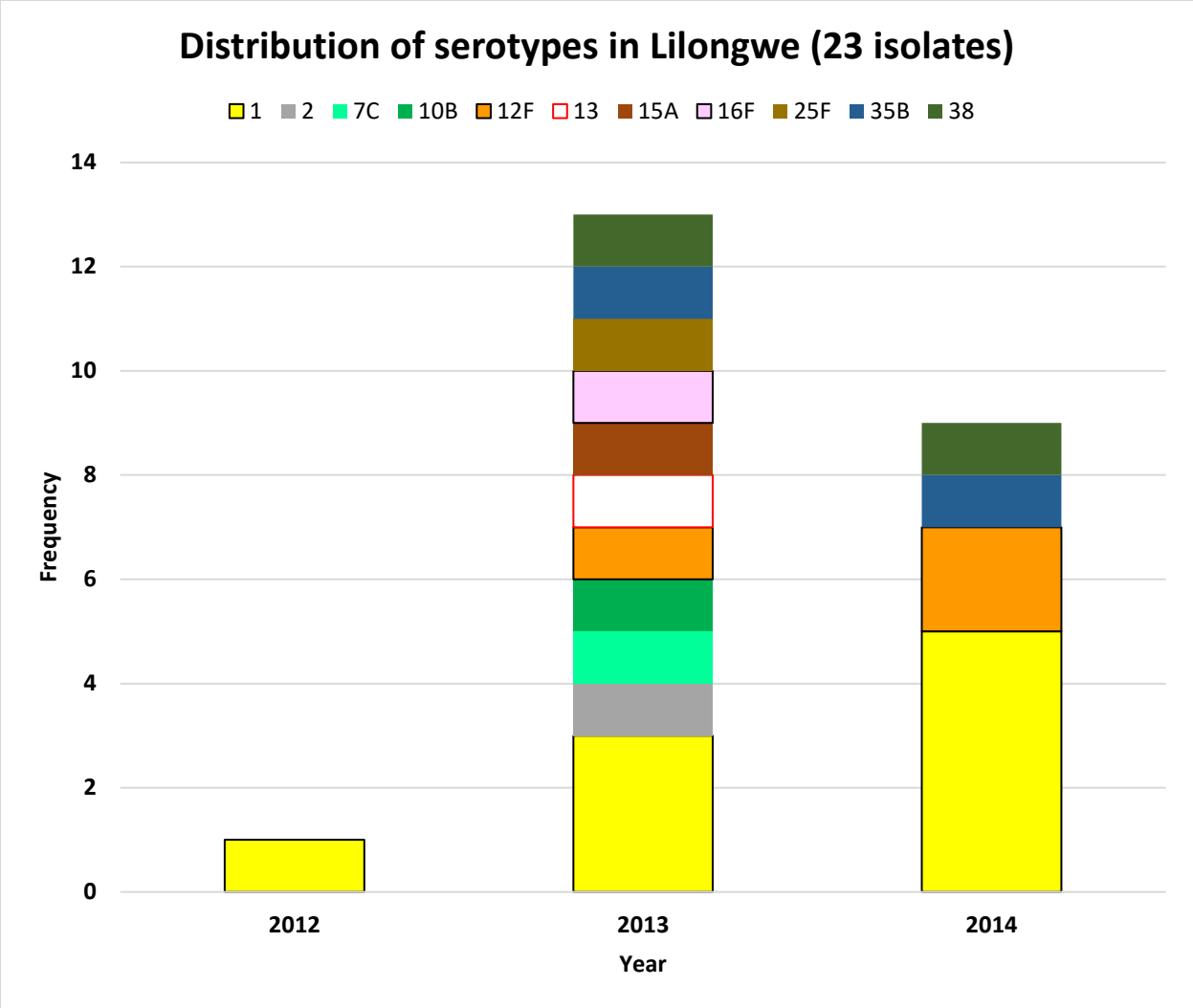
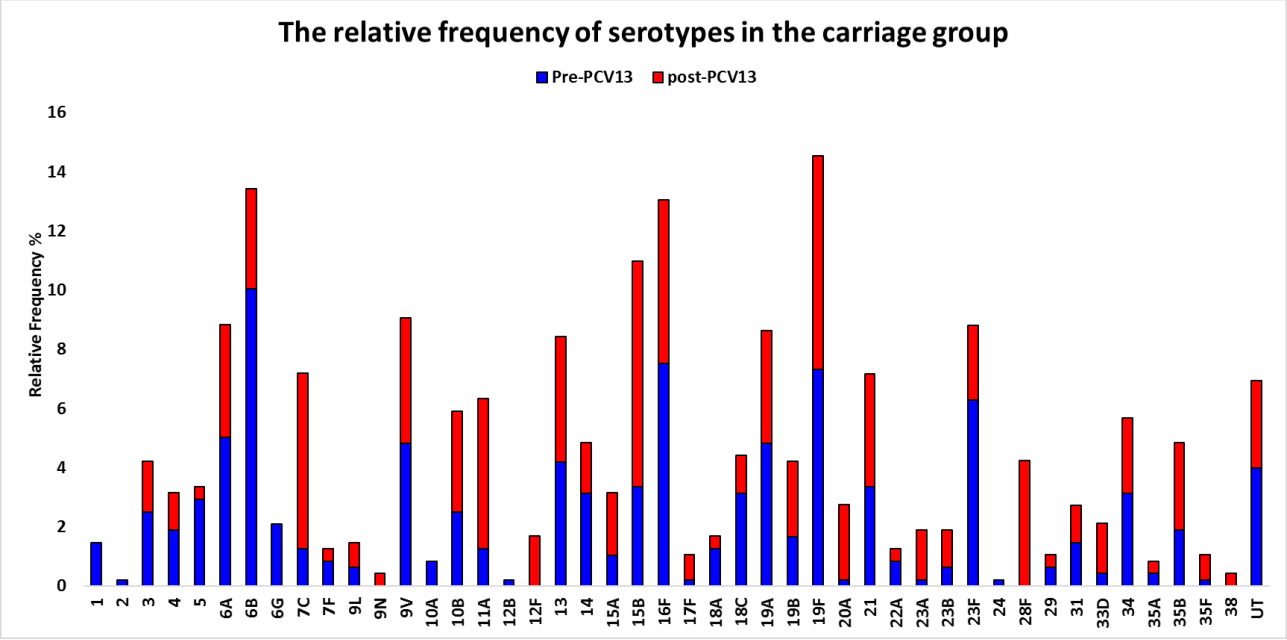


Figure 2-11. Distribution of serotypes in Lilongwe in the Center of Malawi.

2.3.4 Effect of vaccination on serotype distribution

Although the information on whether the individuals from which the samples were taken were vaccinated or not was unavailable, it was possible to analyze changes pre and post the introduction of the vaccine in 2011. An overall view of the serotype distribution in Malawi before (pre-pCV13) and after (post-PCV13) introduction of PCV13 in Malawi in 2011 is shown in Figure 2-12.



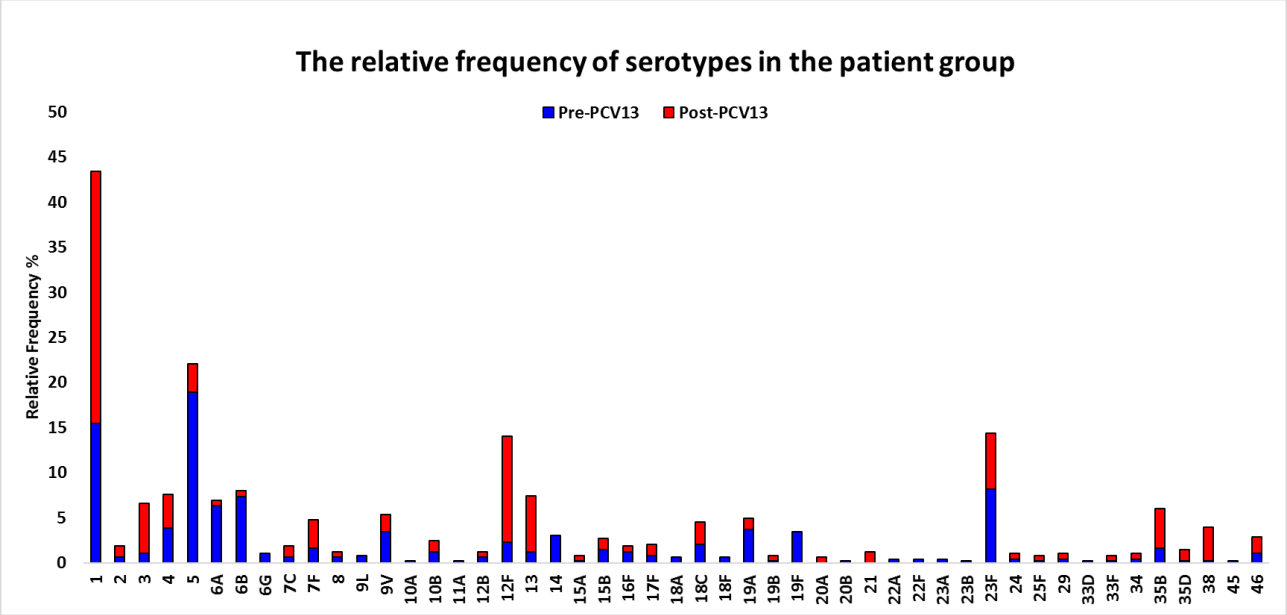


Figure 2-12. Serotype distribution in Malawi before and after 2011.
Columns indicate the relative frequency of serotypes before and after vaccination in 2011 in blue and red, respectively.

In the carrier population, introducing the vaccine into the country did not change the prevalence of vaccine-type 19F, but it significantly decreased the colonization rate of vaccine-type 6B ($p < 0.001$). However, a significant increase in the abundance of non-vaccine types 7C, 11A, 20A, and 28F was identified in the post-PCV13 group ($p < 0.001$) (Figure 2-13). In the carriage group, serotypes 9N, 12F, 28F, and 38 were found only in the post-PCV3 group (Figure 2-12).

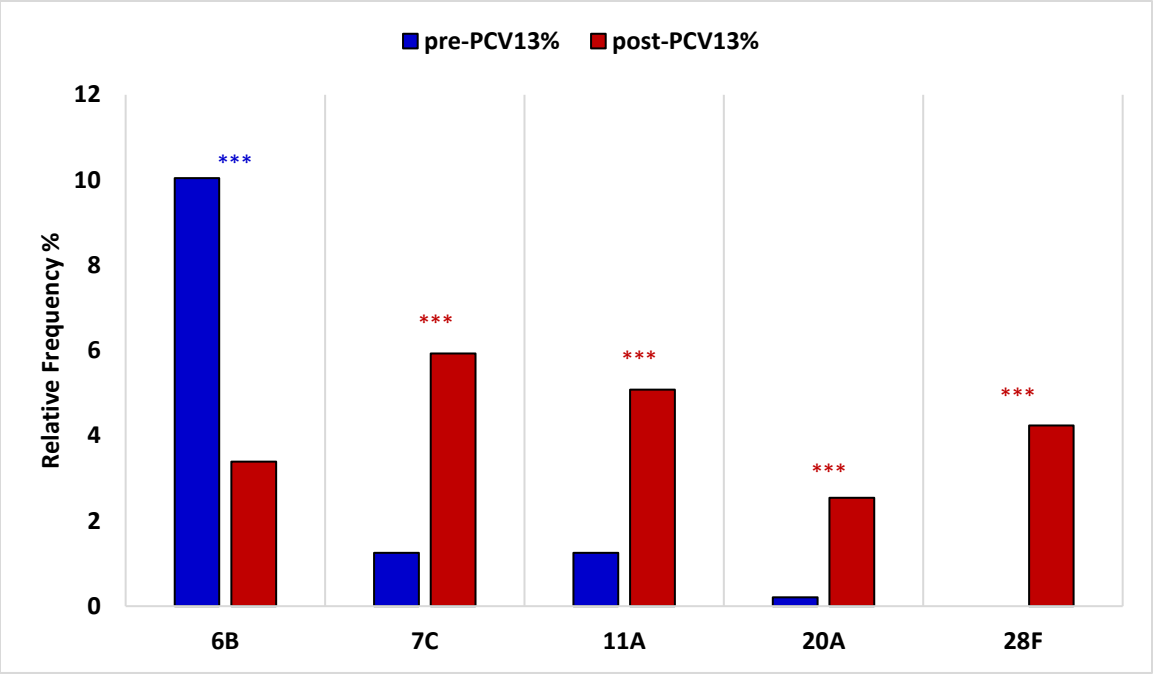


Figure 2-13. Serotypes with a significant difference between the pre- and post-PCV13 frequencies in the carriage group ($p < 0.001$).

In the patient group, although the prevalence of vaccine types 5, 6A, 6B, and 19F in the post-PCV13 group significantly decreased ($p < 0.001$), a significantly higher prevalence of vaccine types 1 and 3 and non-vaccine types 12F, 13, and 38 were observed ($p < 0.001$) (Figure 2-14). Moreover, the vaccine type 23F was among the dominant serotypes in both pre- and post-PCV13 groups.

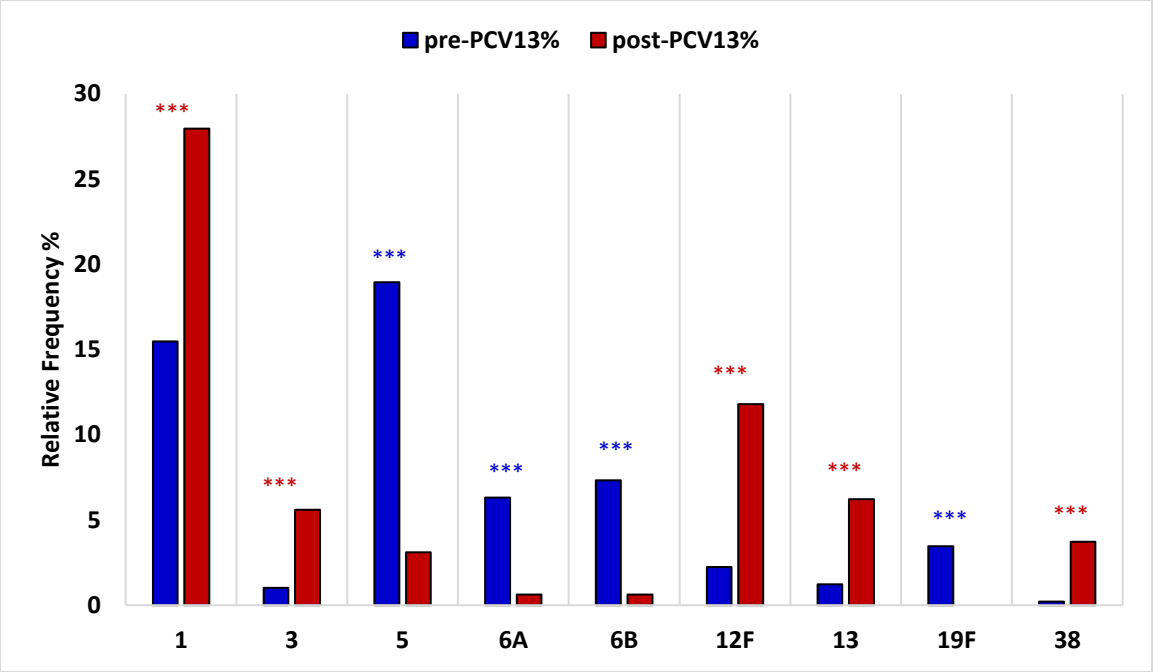


Figure 2-14. Serotypes with a significant difference between pre- and post-PCV13 frequencies in the patient group ($p < 0.001$).

2.3.5 Differentially distributed serotypes between carriers and patients

The results of the statistical analysis of the distribution of serotypes across sampling locations are summarized in Figure 2-15. Amongst the carriage isolates, serotypes 19F (7.88%), 6B (7.27%), 16F (6.79%), and 23F (5.21%) were the most abundant. Serotype 5 (20.38%) followed by 1 (16.58%), 23F (8.42%), and 6B (5.98%) dominated the blood samples. Similarly, we identified the dominance of serotypes 1 (21.13%), 5 (8.1%), 12F (7.04%), 23F (6.69%), 6A (5.63%), and 6B (5.28%) in the samples isolated from the CSF. Of all abundant serotypes, serotypes 1, 5, and 12F have the most significant presence among patients ($P < 0.001$). The term "significant invasive serotypes" refers to serotypes 1, 5, and 12F. Comparatively, serotypes 16F and 19F have a significantly high prevalence only among the carriers implying that they might have the lowest invasiveness ($p < 0.001$) (Figure 2-16). Other abundant serotypes such as 6A, 6B, and 23F have a similar frequency among the carriers and patients (Figure 2-15 and Table 2-2).

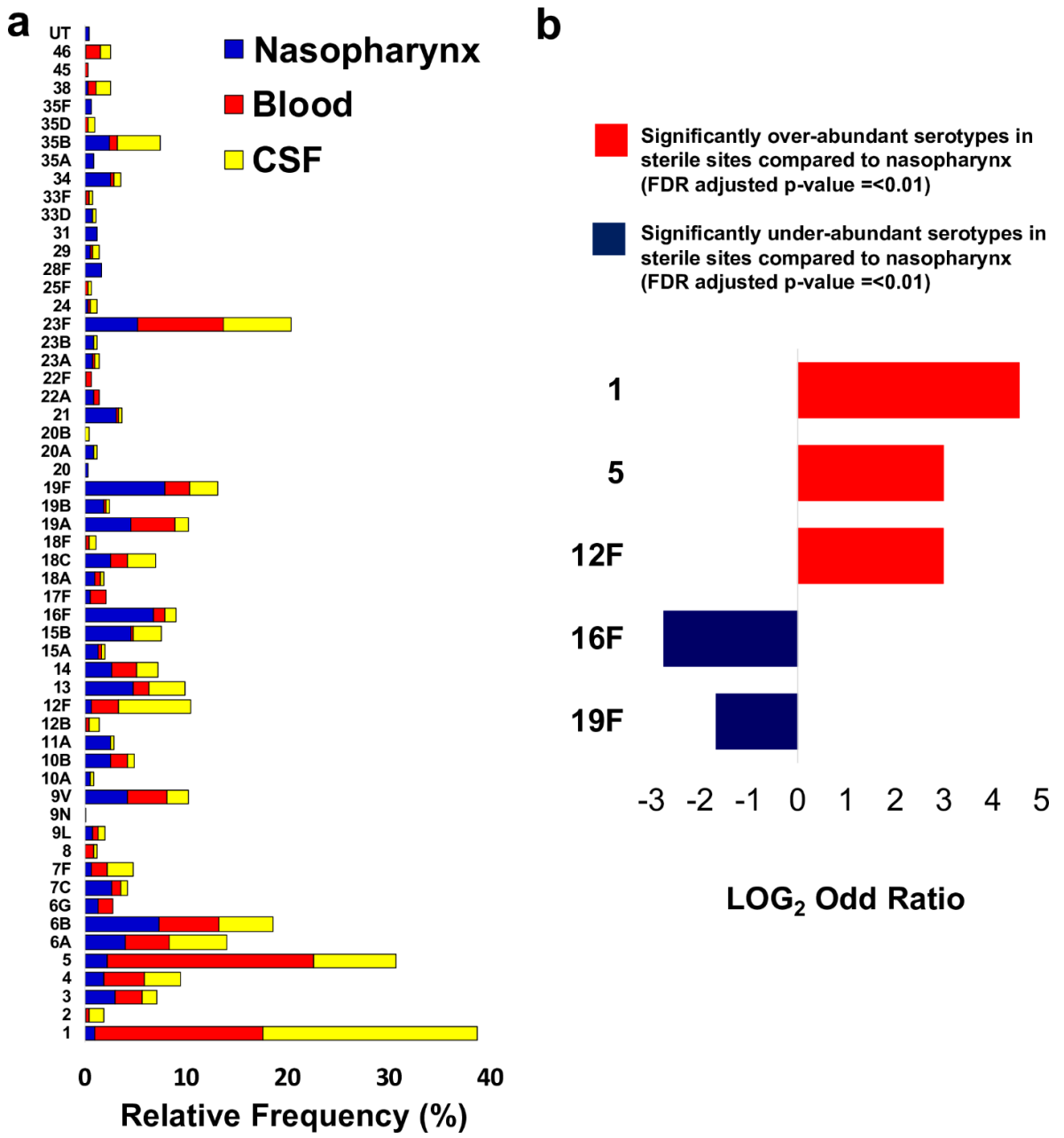


Figure 2-15. The distribution of the 56 pneumococcal serotypes assigned to 1477 samples from Malawi. (a) The relative frequency of each serotype in the nasopharynx of carriers, the blood of bacteremia patients, and the cerebrospinal fluid of meningitis patients are shown in blue, red, and yellow, respectively (UT: Un-Typeable). (b) The transformed odds ratio of the significant abundant serotypes in the patient and carrier groups. Fisher's exact test was applied to identify serotypes with a significant abundance among carriers and patients (nasopharynx and sterile sites) at the significance level of 0.01.

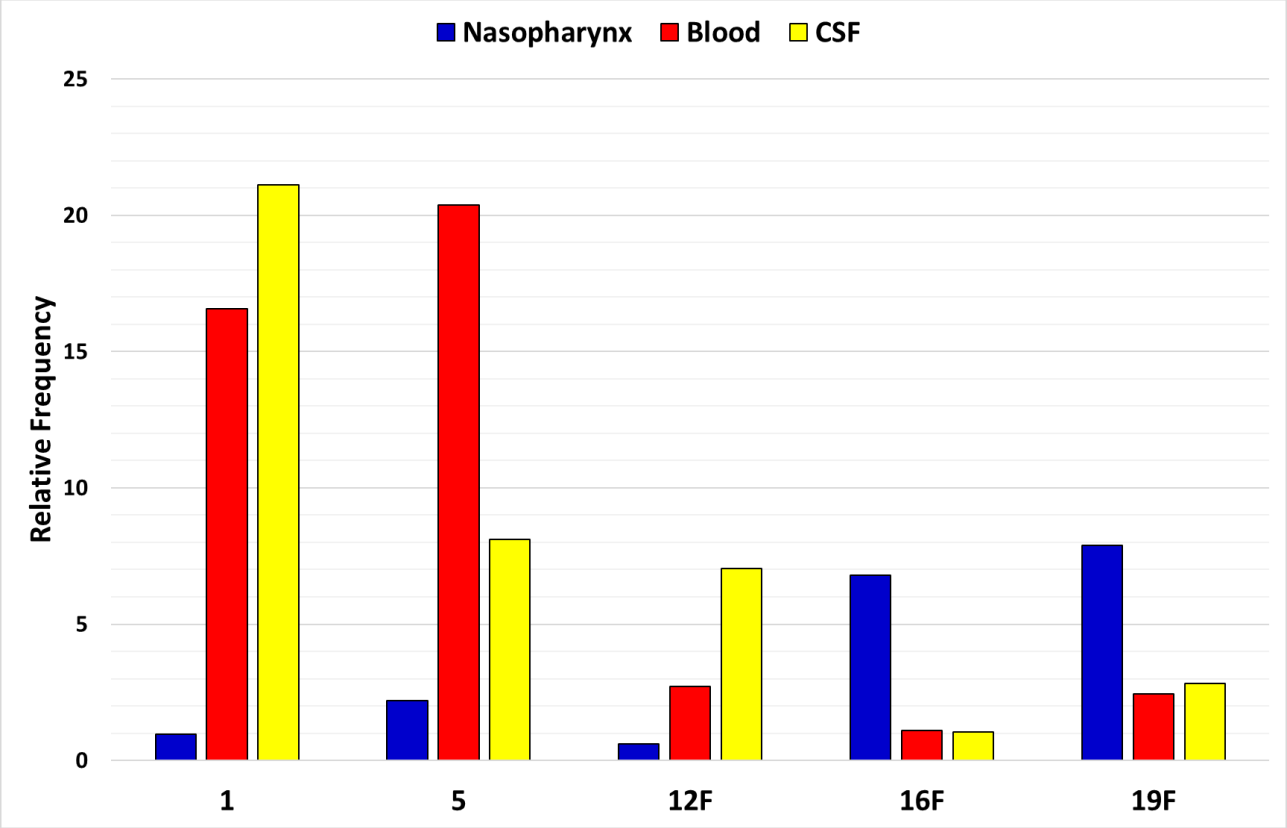


Figure 2-16. Abundant serotypes differentially distributed across the isolation sites (nasopharynx, blood, and CSF) with a p-value < 0.001.

Table 2-2 Statistical analysis of serotype distribution across specimen sources.
Abundant serotypes with a relative frequency greater than 5% in the carriers and patient groups are highlighted in green and red, respectively. P-values less than 0.01 are highlighted in yellow. Of abundant serotypes, serotypes 1, 5, and 12F are significantly prevalent among patients. Serotypes 16F and 19F are significantly prevalent among carriers. Serotypes 6B and 23F are frequent in both groups of carriers and patients.

Serotype	Frequency among carriers %	Frequency among patients %	P-value
1	0.969697	18.55828	1.96E-34
2	0.121212	0.766871	0.186633
3	2.909091	2.147239	0.553182
4	1.818182	3.834356	0.073906
5	2.181818	15.03067	3.97E-19
6A	4	4.907975	0.559777
6B	7.272727	5.674847	0.417863
6G	1.333333	0.766871	0.486148
7C	2.666667	0.766871	0.025664
7F	0.606061	1.993865	0.064075
8	0.606060	0.613497	0.109554
9L	0.727273	0.613497	1
9N	0.121212	0	1
9V	4.242424	3.067485	0.433858
10A	0.484848	0.153374	0.54027
10B	2.545455	1.226994	0.183085
11A	2.545455	0.153374	0.000367
12B	0.121212	0.613497	0.319889
12F	0.606061	5.102773	5.29E-06
13	4.727273	2.453988	0.079864
14	2.666667	2.300613	0.808228
15A	1.333333	0.306748	0.129913
15B	4.484848	1.380368	0.004652
16F	6.787879	1.07362	2.66E-07
17F	0.484848	0.920245	0.509729
18A	0.969697	0.460123	0.515704
18C	2.545455	2.147239	0.808228
18F	0.121212	0.460123	0.486148
19A	4.484848	3.067485	0.319889
19B	1.818182	0.306748	0.025664
19F	7.878788	2.607362	6.06E-05
20	0.242424	0	0.612052
20A	0.848485	0.153374	0.183085
20B	0.606060	0.153374	0.559777
21	3.030303	0.306748	0.000353
22A	0.848485	0.306748	0.486148
22F	0.121212	0.306748	0.680323

Tale 3-2 (Cont'd)

Abundant serotypes with a relative frequency greater than 5% in the carriers and patient groups are highlighted in green and red, respectively. P-values less than 0.01 are highlighted in yellow. Of abundant serotypes, serotypes 1, 5, and 12F are significantly prevalent among patients. Serotypes 16F and 19F are significantly prevalent among carriers. Serotypes 6B and 23F are frequent in both groups of carriers and patients.

23A	0.727273	0.306748	0.590018
23B	0.848485	0.153374	0.183085
23F	5.212121	7.668712	0.16692
24	0.242424	0.460123	0.750435
25F	0.606060	0.306748	0.342197
28F	1.575758	0	0.005192
29	0.484848	0.460123	1
31	1.212121	0	0.015253
33D	0.727273	0.153374	0.274783
33F	0.121212	0.306748	0.680323
34	2.545455	0.460123	0.007386
35A	0.848485	0	0.068171
35B	2.424242	2.300613	1
35D	0.606060	0.460123	0.183085
35F	0.606061	0	0.170856
38	0.242424	1.07362	0.129913
45	0.606060	0.153374	0.559777
46	0.121212	1.032204	0.049967
UT	0.363636	0	0.430506

2.4 Discussion

PCV13 protects against 13 pneumococcal types including 1, 3, 4, 5, 6A, 6B, 7F, 9V, 14, 18C, 19A, 19F, and 23F. In this study, any serotype with a frequency greater than 5% is reported as a frequent serotype. The dataset in this research shows that during the pre-PCV13 era from 1997 to 2011, the frequent nasopharyngeal types obtained from healthy carriers were 6B (10.04%), 16F (7.53%), 19F (7.32%), 23F (6.28%), and 6A (5.02%). However, the prevalence of vaccine types 6B (3.39%), 23F (2.54%), and 6A (3.81%) decreased after 2011 (the reduction in the frequency of serotype 6B was significant, $p < 0.001$). During the post-PCV13 period from 2012 to 2015, the most prevalent serotypes collected from carriers were 15B (7.63%), 19F (7.2%), 7C (5.93%), 16F (5.51%), and 11A (5.08%). The non-vaccine type 16F and vaccine type 19F are frequent in both pre- and pot-PCV13 eras, and the non-vaccine types 15B, 7C, and 11A became dominant after 2011. However, a significant increase ($p < 0.01$) was observed in the post-PCV13 frequency of serotypes 7C (5.93%), 11A (5.08%), 20A (2.54%), and 28F (4.24%). These serotypes may represent the instance of serotype replacement in the carriage group. However, the frequency of serotypes 20A and 28F was < 5% in the post-PCV13 era. Of note is that serotypes 12F and 28F were present in the carriage group only after 2011. About 4% of pre- and post-PCV13 carriage isolates were un-typable; these are likely not capsulated isolates. The un-encapsulated isolates were not observed

amongst patients as all samples in the sterile sites should be capsulated to evade the host immune response.

In the patient group, the findings of this dataset suggest that vaccination in Malawi significantly reduced the burden of disease caused by serotypes 5, 6A, 6B, and 19F ($p < 0.01$). However, it was not successful in preventing infection from all VTs. The statistical analysis shows that the IPD incidence caused by vaccine types 1 and 3 significantly increased after 2011 ($p < 0.01$). Serotype 1, the most abundant and invasive serotype in the cohort, remained significantly dominant in the post-PCV13 era amongst patients (27.95%). This emphasizes the role of serotype 1 in IPD development in Malawi, despite vaccination. Serotype 1 is prone to be epidemic, and it often causes outbreaks in a population. Apart from VTs, the non-vaccine types 12F (11.80%), 13 (6.21%), and 38 (3.73%) significantly caused IPDs after 2011 ($p < 0.01$). Serotypes 1 and 12F are dominant in the CSF of meningitis patients and should be considered as serotypes of concern since they show the ability to infect the central nervous system and remained prevalent in the post-PCV13 era. It is worth noting that there are limitations in this study as the number of samples is low, sampling is biased, and it is not clear whether people from whom the post-PCV13 samples were collected received the vaccine or not. A dataset with larger and geographically evenly distributed samples may answer these questions and assist in fully exploring the effect of vaccination in Malawi. A study conducted in 2020 that investigated pneumococcal samples between 2015 to 2018 concluded that within 3.6–7.1 years after Malawi's 2011 PCV13 introduction, there was still high residual VT carriage in the country[145]. The current study supports this, and these results are not surprising as no catch-up program was implemented for vaccination.

The significant presence of serotypes 1, 5, and 12F among patients together with their scarcity in the carrier group, can be interpreted as their high invasiveness, suggesting a shorter period of nasopharyngeal colonization. Serotypes 1 and 5 are also the most abundant serotype across the entire cohort, though in the temporal analysis in this research, it was evident that serotype 5 prevalence decreased over time. The findings of this study support previous studies, which have shown serotypes 1, 5, and 12F are the major cause of invasive pneumococcal diseases in Malawi[146]. Serotypes 1 and 5 are known as invasive vaccine types that can infect all age groups and cause severe IPDs[147]. Serotype 1 is genetically distinct between different geographical regions[148] and is known as the leading cause of pneumococcal meningitis in Africa[149,150].

In contrast, serotypes 16F and 19F were predominantly found among carriers implying that they may have the lowest invasiveness in the cohort. The low frequency of serotype 19F is likely due to the effect of vaccination as its prevalence significantly decreased after 2011 amongst patients. Most of the other serotypes, including frequent vaccine types 6A, 6B, and 23F, are common between carriers and patients, which means they cause invasive pneumococcal disease but are likely to colonize the nasopharynx before they infect the sterile sites. The frequency of serotypes 6A and 6B significantly declined in the patient group after 2011. Serotype 23F remained dominant in carriage and patient groups before and after 2011.

Given that pneumococcal virulence remarkably depends on the serotype of isolates, the issue that must be discussed is why several serotypes are shared across both nasopharynx and sterile sites. Two potential scenarios could explain the ubiquitous presence of some serotypes in both nasopharynx and sterile sites. The first scenario is related to the duration of the pneumococcal colonization which is a known prerequisite for virulence. Abundant serotypes 6B and 23F with a similar frequency amongst the

carriers and patients possibly need to colonize the upper respiratory tract before entering the bloodstream. In contrast, the significant invasive serotypes 1, 5, and 12F do not colonize for long and more quickly enter the sterile sites. Therefore, serotypes 6B, and 23F may also be very invasive, but they need a more extended period of nasopharyngeal colonization than serotypes 1, 5, and 12F. Thus, at a random collection time, serotypes 6B and 23F are found in both carriage and patient groups, but serotypes 1, 5, and 12F are only found amongst patients due to their short colonization period. The second possible scenario could be the differential gene expression pattern of shared genes in the ubiquitous serotypes. Although the type-specific *cps* genes were identified in the isolates of both nasopharynx and sterile sites, the expression pattern of these genes could be varied in each isolation site, which would explain the invasiveness of isolates in the sterile sites. Several studies described a cycle of encapsulation and un-encapsulation among the pneumococcal strains. Isolates benefit from mutations in the cps locus to either cease or re-start the capsule expression[40]. The lack of a capsule at the epithelial surface enables the bacterium to expose its surface proteins on the cell wall underneath the capsule and promote adherence to the host epithelial cells. It has been estimated that 15% of isolates in the upper respiratory tract are unencapsulated and adhere to the respiratory epithelial cells more efficiently than encapsulated isolates[23,24]. Lack of capsule also facilitates acquiring virulence and resistance genes from other isolates. During disease, the thick capsule prevents immunoglobulins from interacting with the pathogen surface proteins, meanwhile, the negatively charged CPS interferes with the function of the host phagocytes[25,45]. Taken together, the presence of the *cps* locus in the genome of isolates assigned to the same serotype does not necessarily reflect the encapsulation of all cells.

3 Genome assembly, pan-genome construction, and phylogeny

3.1 Overview

In this chapter, the clean raw sequencing data was used as the input to construct the pan-genome, which acted as the foundation component of the analysis used to answer the questions theorized in this research study. The computational pipeline included the *de novo* assembly of genomes, followed by a quality assessment. The assembled genomes were computationally annotated to produce the input files for the pan-genome construction. The pan-genome formed the reference sequence representing the high level of diversity in the *pneumococcus* genome. Core genes from the reference sequence were used to undertake a phylogenetic analysis on the dataset to determine the genetic similarity between samples. The following chapters used the pan-genome as a reference file for variant calling and gene presence-absence statistical analysis. The main objectives of this chapter were:

- Assembly and annotation of the genomes
- Pan-genome construction and identification of the core and accessory genes
- Phylogenetic analysis
- Functional enrichment analysis of the pan-genome components

3.2 Methods

3.2.1 Genome assembly

Based on the current technology limitations, sequencing platforms are able to sequence only a certain length of the DNA molecule and require multiple copies of DNA fragments. Therefore, the sequencing output files contain millions of overlapping short reads. Genome assembly takes the sequencing reads generated by the sequencer and puts them back together to reconstruct the sample's genome. This process is computationally expensive. The algorithms developed for genome assembly must consider differences between eukaryotic and prokaryotic genome structures.

There are two main methods to assemble genomes from sequencing files:

- Reference-based genome assembly
- *De novo* genome assembly

In the reference-based assembly, the reference genome of the organism is used as a guide to join the short reads. For each species, the reference genome is a complete genome with high quality constructed from previously sequenced samples of that species. The presence of the reference genome in this method reduces the complexity and cost of computing. Therefore, this method is applied to assemble large genomes that are not genetically diverse such as *Homo sapiens*.

De novo assembly is used when no suitable reference genome is available for the study samples. Short reads are aligned and assembled by assessing the overlaps between them. A pan-genome is often applied to portray genetic variation in one organism and is mainly beneficial to investigate bacterial

species with highly divergent genomes across different strains, as defining a linear reference genome for such a diverse species is difficult. Thus, *de novo* assembly, which is reference-free, is desired in computational pan-genomics.

The *pneumococcus* is a genetically diverse bacterium. The variation can exist in the presence or absence of entire genes resulting in each strain having its combination of genes within an individual genome. As a result, for each pneumococcal strain, a separate reference genome has been defined[151]. Given this, there isn't a unique linear reference genome appropriate for the multiple *S. pneumoniae* strains. Therefore, this research employed the reference-free *de novo* assembly method to assemble pneumococcal genomes. The *pneumococcus* genome is not very large, with a length of ~2 million base pairs carrying 2043 coding and 73 non-coding genes[152], therefore, *de-novo* genome assembly is practical for this genome and does not require much computing memory.

The main issue related to *de novo* assembly is how to assemble repetitive genome regions without a reference guide. Repetitive regions refer to loci in the genome composed of tandem repeats and have low complexity. Modern technologies such as paired-end sequencing and long-read sequencing can solve this problem. As mentioned in this study, sequencing data was produced by Illumina machines in the format of paired-end sequencing Fastq files that are desired to resolve the issue of the repetitive regions.

Several well-documented and efficient tools such as Canu[153], Flye[154], Hinge[155], and Velvet[156] are available for *de novo* assembly of bacterial genomes. Canu, Flye, and Hinge are specifically designed to assemble long-read sequences using the modified Smith-Waterman algorithm. Alternatively, Velvet employs *de Bruijn* graphs, a good solution for assembling short reads sequences[157], making it a helpful tool for this research. Velvet is open-source, widely-used, one of the most-cited assemblers for bacterial genomes, and best suited to Illumina sequence reads[158]. Velvet first converts short reads to k-mers using a hash table (k-mers refers to all substrings of length k from a string of length n if $k < n$), then assembles overlapping k-mers into contigs through a *de Bruijn* graph construction. The hash value and the value of k are crucial for optimal assembly. VelvetOptimiser is the tool that automates the optimization and selects the best value for parameters used by Velvet (Figure 3-1)[156].

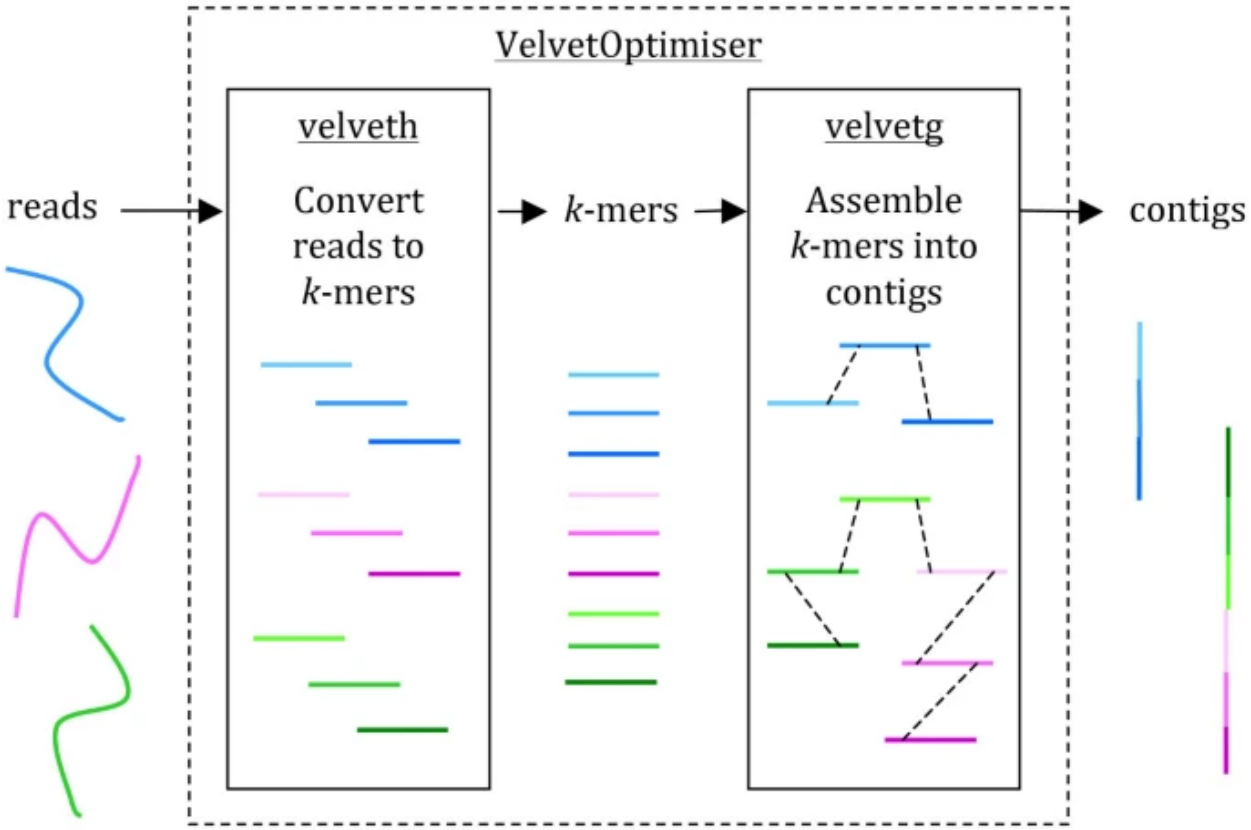


Figure 3-1. Genome assembly using Velvet and VelvetOptimiser.
Image from [158]

VelvetOptimiser version 2.2.5[95] was used to assemble sequencing short reads. The assembler was set to produce contigs longer than 500 base pairs because this length would improve the quality of the pan-genome structure. Contigs shorter than 500 base pairs cause misannotation of genomes resulting in skewed gene clustering during pan-genome construction. The most critical parameter for a Velvet run is the hash value (k-mer size) which should be an odd number shorter than the average length of reads in the sequencing file. An optimal hash value would often be between half read and average read length minus 10[156]. Since the average size of reads was 125, the starting hash value was set to 61, and the end hash value was set to 115. VelvetOptmiser examined all odd numbers between 61 and 115 for the genome assembly and selected the best hash value leading to an assembly with the highest coverage and N50. N50 refers to the length when the collection of contigs of that length or longer covers at least half of the assembly. The quality of the assembled genomes was evaluated by Quast[97] The advantage of Quast is that it is able to assess an assembled genome when the reference genome is not available. It produces several summary tables and plots.

The source code for genome assembly and the assessment of assembled genomes is available in Appendix 1.

3.2.2 Genome Annotation

The assembled genomes were annotated using Prokka version 13.1[99]. Prokka is a Linux command-line tool implemented in Perl, which is freely available and is designed to annotate bacterial and viral genomes. Prokka fully annotates a draft bacterial genome in about 10 minutes on a typical desktop

computer. The fragmented draft *de novo* contigs produced by VelvetOptimiser were used as the input for the genome annotation. Prokka relies on external tools to predict genetic features with assigned coordinates and labels. These tools include Prodigal[159] to recognize genes and coding sequences, SignalP[100] to find signal peptides and secretory proteins, Aragorn[101] that identifies transfer RNAs (tRNAs) genes, RNAmmer[160] to annotate ribosomal RNAs (rRNAs), and Infernal[104] to explore non-coding RNAs (nRNAs). In addition to the default Prokka configured databases, a customized *pneumococcus* genus database was built using 23 well-annotated finished pneumococcal genomes available from the National Centre for Biotechnology Information (NCBI) genome database (Table 3-1).

Table 3-1. NCBI reference genomes used to build the Prokka database.

NCBI Reference Sequence	**Definition**
NC_003028.3	*Streptococcus pneumoniae* TIGR4, complete genome.
NC_003098.1	*Streptococcus pneumoniae* R6 chromosome, complete genome.
NC_008533.1	*Streptococcus pneumoniae* D39, complete genome.
NC_010380.1	*Streptococcus pneumoniae* Hungary19A-6, complete genome.
NC_010582.1	*Streptococcus pneumoniae* CGSP14, complete genome.
NC_011072.1	*Streptococcus pneumoniae* G54, complete genome.
NC_011900.1	*Streptococcus pneumoniae* ATCC 700669 complete genome.
NC_012466.1	*Streptococcus pneumoniae* JJA, complete genome.
NC_012467.1	*Streptococcus pneumoniae* P1031, complete genome.
NC_012468.1	*Streptococcus pneumoniae* 70585, complete genome.
NC_012469.1	*Streptococcus pneumoniae* Taiwan19F-14, complete genome.
NC_014251.1	*Streptococcus pneumoniae* TCH8431/19A, complete genome.
NC_014494.1	*Streptococcus pneumoniae* AP200, complete genome.
NC_014498.1	*Streptococcus pneumoniae* 670-6B, complete genome.
NC_017591.1	*Streptococcus pneumoniae* INV104 genome.
NC_017592.1	*Streptococcus pneumoniae* OXC141 complete genome.
NC_017593.1	*Streptococcus pneumoniae* INV200 genome.
NC_017769.1	*Streptococcus pneumoniae* ST556, complete genome.
NC_018630.1	*Streptococcus pneumoniae* gamPNI0373, complete genome.
NC_021005.1	*Streptococcus pneumoniae* SPN994039 draft genome.
NC_021006.1	*Streptococcus pneumoniae* SPN034156 draft genome.
NC_021026.1	*Streptococcus pneumoniae* SPN994038 draft genome.
NC_021028.1	*Streptococcus pneumoniae* SPN034183 draft genome.

Prokka produces its outputs in various file formats listed in Table 3-2. At least one of the Prokka output files is compatible with the tools used for pan-genome construction. In this study, the full annotations of the draft assembled genomes produced by Prokka in Gff3 format were used to build the pan-genome. This Gff3 format includes annotation and the sequences of contigs.

The source code for annotating draft assembled genomes is available in Appendix 1.

Table 3-2. The list of Prokka output files.

Suffix	Description of file contents
.fna	Fasta file of original input contigs (nucleotide)
.faa	Fasta file of translated coding genes (protein)
.ffn	Fasta file of all genomic features (nucleotide)
.fsa	Contig sequences for submission (nucleotide)
.tbl	Feature table for submission
.sqn	Sequin editable file for submission
.gbk	Genbank file containing sequences and annotations
.gff	GFF 3 file containing sequences and annotations
.log	Log file of Prokka processing output
.txt	Annotation summary statistics

Table from [99].

3.2.3 Pan-genome and pan-IGR construction

All annotation files in the GFF3 format were used to build the pan-genome using Roary version 3.12.0[161]. Roary is a suitable tool for building large-scale pan-genomes that include thousands of samples. Its algorithm is designed to use computing resources efficiently without compromising the quality of the outcomes. For instance, 128 samples can be analyzed in under 1 hour using 1 GB of RAM and a single processor. A similar analysis performed by other methods would take weeks and hundreds of GB of RAM[162]. It applies graph-based methods to assign orthologous genes found in strains. The summary of Roary's algorithm is as follows:

- It converts coding sequences into amino acid sequences.
- It clusters predicted proteins using CD-hit[163].
- It performs an all against all comparison using Blastp[164].
- It does a final clustering by MCL[165] and reports the final clusters.

The advantage of Roary is that it produces several output files that can be used as the input for software in downstream analysis. It also creates tab-delimited files such as a gene presence-absence matrix that R can easily use to visualize the pan-genome matrix and perform Principal Component Analysis (PCA).

A pan-genome contains the entire genome (genes and intergenic regions) in the set of strains from the same or closely related species. To accurately perform a phylogenetic analysis and determine the population structure, genes and intergenic regions (IGRs) in the pan-genome must be investigated concurrently. To explore the variation in the IGRs, Piggy[166] was used. Piggy emulates Roary, but it works on IGRs instead of genes. It extracts IGRs from annotation files and uses the flanking gene names and their orientations produced by Prokka and Roary to name the IGRs. It also provides information about the presence of IGRs in the isolates. Analysis of IGRs and genes together provides insight into the patterns of genes and IGRs distribution in the cohort.

Roary was run to perform the core gene alignment with Mafft version 7.313[167]. The minimum percentage identity for Blastp was set to 95%. Piggy was run using default parameters with the size of IGRs to extract set to 30-1000. Each gene or IGR that existed in 100% of samples was considered as a core gene or core IGR, and those that were not core but present in more than 95% of samples were

regarded as a soft-core gene or soft-core IGR. The rest were considered as accessory genes or accessory IGRs.

The source code for the pan-genome and pan-IGR constructions using Roary and Piggy is available in Appendix 1.

3.2.4 Phylogenetics

The multiple sequence alignment of core genes produced by Mafft during the pan-genome construction was used as input into the IQ-TREE version 2[168] to draw a phylogenetic tree. IQ-TREE is open-source software, available as either a web server or a standalone Linux command-line tool.

It applies the Maximum Likelihood (ML) method for phylogenetic inference. This software has high performance due to its algorithms for model selection, tree search, and a novel ultrafast bootstrap approximation. Bootstrapping is a technique to assess the accuracy of phylogenetic trees. Many artificial alignments (hundreds or thousands) are generated during bootstrapping by randomly picking columns from the multiple sequence alignment of samples. For each artificial alignment, a phylogenetic tree is built. Then, the frequency of each phylogenetic feature in thousands of trees is calculated. The higher frequency causes more confidence to be assigned to the phylogenetic feature. IQTREE is compatible with multicore CPUs and parallel programming to speed up analysis and supports large datasets with thousands of sequences. In terms of memory usage, computing times, and likelihood maximization, IQ-TREE is more efficient than other popular ML phylogenetic software such as RAxML and PhyML[169]. The resulting tree from the pneumococcal sequences was saved in the Newick format, mid-point rooted, and subsequently submitted to iTOL[170] for visualization. Newick is a text format that describes a phylogenetic tree as a string. In a Newick file, samples are grouped using parentheses, and the branch lengths are represented using colons followed by numbers. iTOL is a web-based tool that interactively displays phylogenetic trees and provides a graphical interface for tree manipulation, annotation, and mapping of various features of samples onto the tree. The annotated tree was utilized to group pneumococcal isolates according to their collection times, serotypes, isolation sites, and geographical locations to provide insight into the population structure. Determination of the population structure contributes to the identification of the potential drivers of the genetic variations in the entire cohort.

The source code for the ML tree construction using IQ-TREE is available in Appendix 1.

3.2.5 Pan-genome/IGRs visualization and primary sample clustering

The R package Nonnegative Matrix Factorization (NMF)[171,172] was used to visualize the pan-genome/IGRs matrices and create a general overview of gene and IGR distributions. It produced the heatmaps for the gene/IGR presence-absence matrices and clustered samples according to the distribution of genes/IGRs in the cohort. NMF is an unsupervised learning method that has been applied in several algorithms in bioinformatics. The core concept of NMF is to factorize a big matrix into two matrices with no negative value that is easier to inspect. NMF can identify how samples are grouped by computing a dendrogram from hierarchical clustering using the distance method with a small set of marker features. The package plots high-quality heatmaps with a detailed legend and unlimited annotation tracks for columns and rows.

The source code to visualize the pan-genome/IGRs matrices using NMF is available in Appendix 1.

3.2.6 Functional and pathway enrichment analysis

The core and accessory sequences were submitted to STRING webtool version 11.5[173] to visualize the protein-protein interaction network and perform functional enrichment analysis. STRING is a free online tool with a simple interface. It can also be accessed from R/Bioconductor. Its database is a network that integrates information from all publicly available protein-protein interaction databases. STRING's computational pipeline can predict direct (physical) and indirect (functional) interactions between proteins. It allows users to upload sequences, visualize the network, and perform enrichment analysis on the input data. It has a well-curated database of both known and predicted protein-protein interactions. The STRING database currently contains information from 5090 organisms, including 24,584,628 proteins. Interactions in STRING are the most comprehensive and are derived from 5 sources, including (i) genomic context predictions, (ii) high-throughput lab experiments, (iii) co-expression, (iv) automated text mining, and (v) previous knowledge in databases. Functional enrichment analysis in STRING uses information from classification systems such as the Kyoto Encyclopedia of Genes and Genomes (KEGG)[174] and Protein families database (Pfam)[175]. The tool entitled "Multiple Sequences" was selected, and *Streptococcus pneumoniae TIGR4* was chosen as the reference organism. STRING reports the associated Gene Ontology (GO) terms with an FDR less than 0.05. STRING also assigns a "Strength" value that describes how large the enrichment effect is for each GO term. It is the ratio between the number of annotated proteins in the submitted sequences with a GO term to the number of proteins expected to be annotated with this term in a random network of the same size.

The STRING IDs of sequences were submitted to the ShinyGO[174] web page, an online gene-set enrichment analysis tool, to acquire information about the pathways in which the genes of interest were involved. ShinyGO benefits from the Shiny framework (an R package designed to build interactive web apps) to access several R/Bioconductor packages. ShinyGO applies the *Chi-squared* test for the statistical analysis. It integrates gene annotations from public databases such as InterPro[176], Ensembl[177], KEGG[178], and STRING for 59 plants, 256 animals, 115 archeal, and 1678 bacterial species (November 2021). *Streptococcus pneumoniae (TIGR4)* was selected as the reference organism for the pathway enrichment analysis. Pathways with a p-value of less than 0.01 were reported as enriched. The p-value was adjusted by the FDR method adapted from the hypergeometric test, indicating how likely the results were obtained by chance. For each pathway, a "Fold Enrichment" value was defined as the percentage of genes in the list belonging to a pathway, divided by the corresponding percentage in the background. Fold enrichment showed how genes of a specific pathway were over-represented in the study set.

3.3 Results

3.3.1 Genome assembly and annotation

The average optimized assembly hash value was 96, longer than half of the length of forward and reverse reads (125 bp). The average N50 was 113,986. The mean assembled genome size was calculated at 2,116,779 nucleotides with a standard deviation of 106,481. This length was near the previously reported *pneumococcus* genome size. All samples had a GC content within the Gaussian distribution with a mean of 40%, equal to the reported GC content of *S. pneumoniae* estimated between 39% - 40%[151] (Figure 3-2). Prokka successfully annotated all draft assembled genomes.

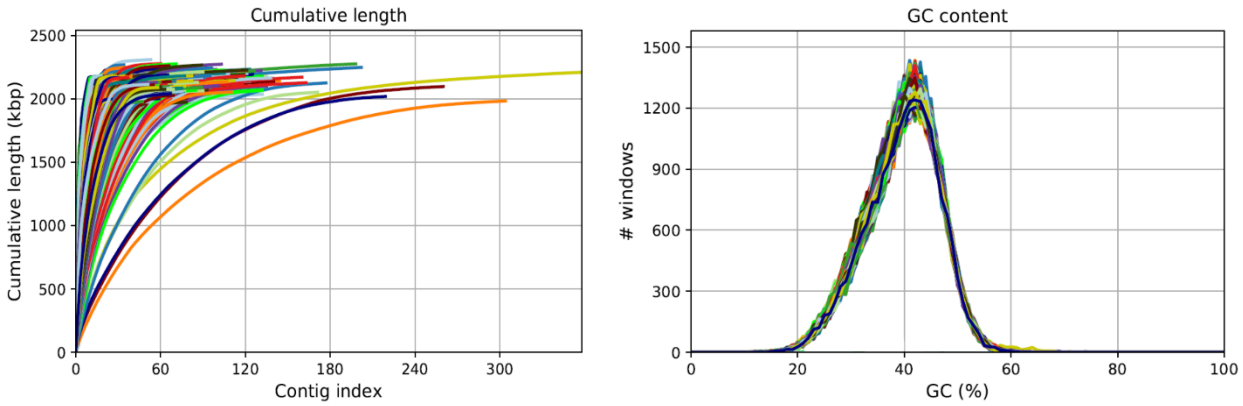


Figure 3-2. Assembly quality assessment of 1477 pneumococcal samples.
(a) Cumulative length plot shows the growth of contig lengths. On the x-axis, contigs are ordered from the largest to the smallest. The y-axis gives the size of the x largest contigs in the assembly. (b) GC content plot shows the distribution of GC content in the contigs. The x value is the GC percentage (0 to 100 %). The y value is the number of non-overlapping 100 bp windows in which the GC content equals x %[179].

3.3.2 Pan-genome/IGRs

The length of the pan-genome (the entire gene set of all samples) was 5,175,290 base pairs. The total number of genes in the pan-genome was 6,803, including 729 core, 821 soft-core, and 5,253 accessory genes (Figure 3-3 and Appendix 2). The overall GC content of the pan-genome was 38.60%. The pan-genome was open, which means the number of genes increased unrestricted when the sample size grew (Figure 3-4). The core-genome's length and overall GC content were 727,753 base pairs and 41.82%, respectively. The size of the accessory-genome was 4,447,537, and the GC content was 38.01%.

The Pan-genome contains the whole set of genes only. Pan-IGR refers to the entire non-coding regions of all samples, which had a total length of 6,619,419 base pairs with a GC content of 36.33%. Typically, IGRs compose 10–15% of a single *pneumococcus* genome. Nonetheless, the size of pan-IGRs was longer than the pan-genome length, meaning diversity in the intergenic regions was much higher than in the coding sequences. The total number of IGRs was 16,147, including 48 core, 626 soft-core, and 15,473 accessory IGRs (Figure 3-3, Appendix 3). The lengths of core- and accessory-IGRs were 40,203 and 6,579,216, with the GC content of 37.83% and 36.46%, respectively.

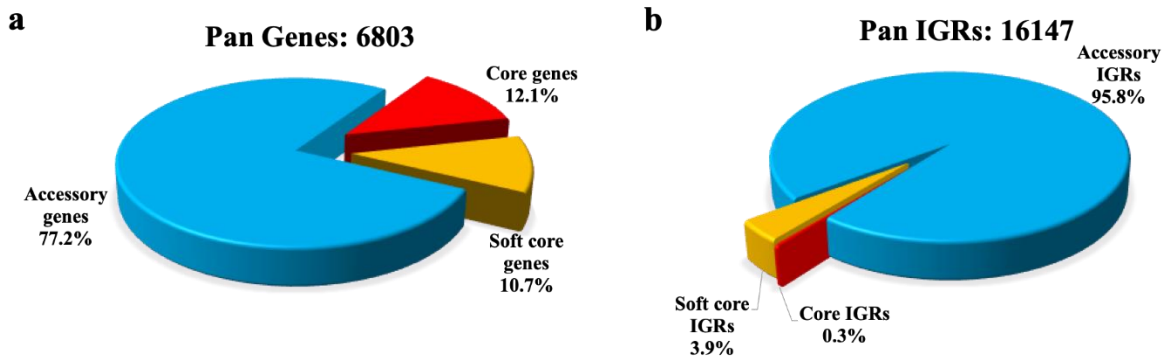


Figure 3-3. Pan-genome of 1477 pneumococcal isolates from Malawi.
The pan-genome was constructed using annotation files from all samples, assuming that proteins with at least 95% of identity belong to the same cluster. The total number of genes in the pan-genome was 6803, including 729 core, 821 soft-core, and 5253 accessory genes. The total number of IGRs was 16147, including 48 core, 626 soft-core, and 15473 accessory IGRs. Core genes and IGRs were indicated in red.

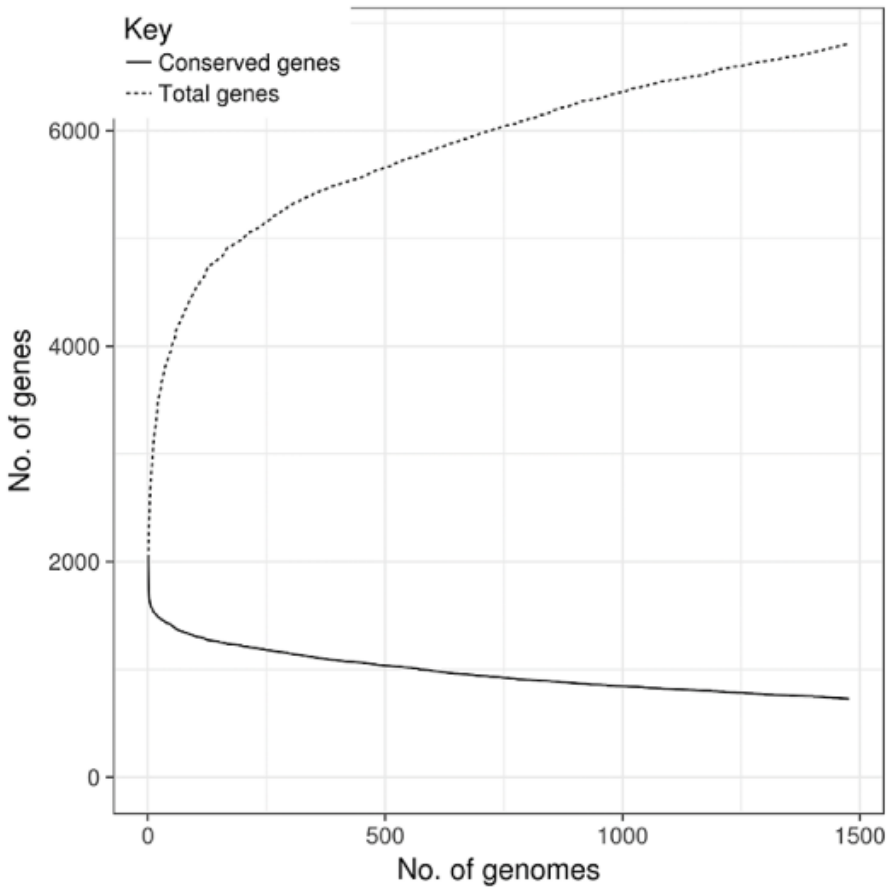


Figure 3-4. Pneumococcal pan-genome was open.
The number of total genes increased unlimitedly when the sample size grew. The horizontal axis shows the number of genomes, and the vertical axis shows the total number of genes in the pan-genome. The solid line indicates the number of core genes. The dashed line represents the total number of genes.

The results showed that the GC content of the coding regions was slightly higher than the intergenic regions. The number of IGRs increased faster than the number of genes when more genomes were added to the pan-genome (Figure 3-5). This is due to the higher diversity amongst IGRs than in the coding regions. Coding regions are more constrained in variation as changes could lead to altered or impaired protein function.

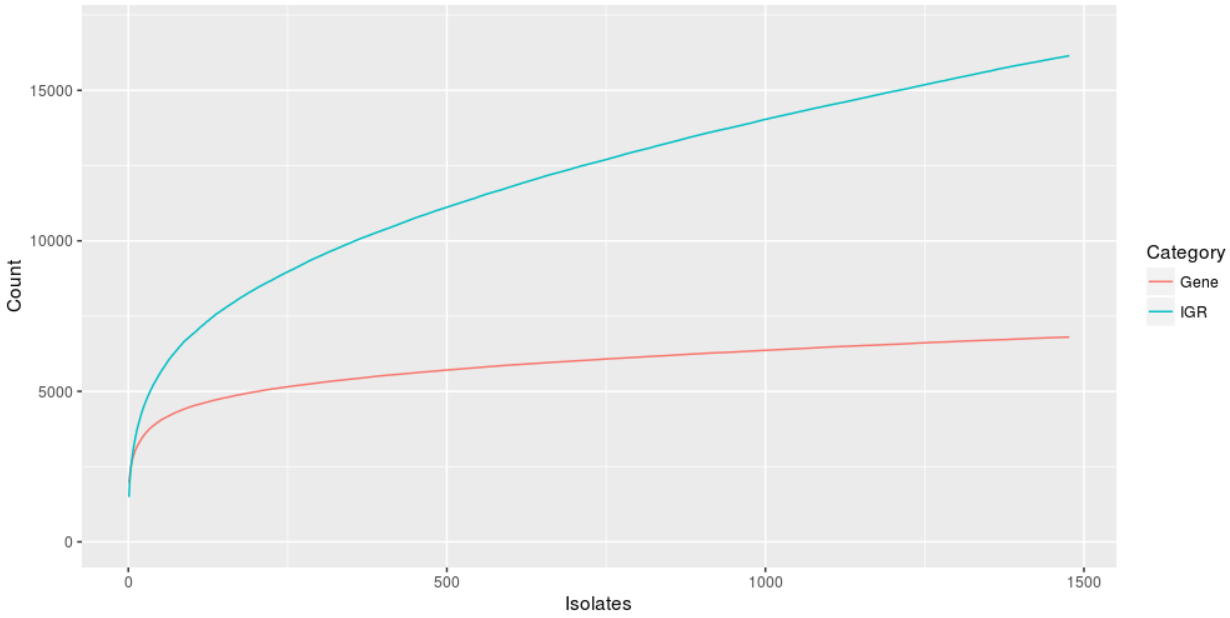


Figure 3-5. Gene and IGR accumulation.

3.3.3 Phylogenetics

The phylogenetic tree drawn from the multiple sequence alignment of core genes is shown in Figure 3-6. This tree demonstrates similarity in the core-genome and illustrates how samples clustered. Serotypes with the source-based p-value < 0.05 (1, 5, 11A, 12F, 15B, 16F, 19F, and 21) or relative frequency > 0.05 (6A, 6B, and 23F) are highlighted on the tree. The rings around the tree display the following features: serotypes, specimen sources (isolation sites), PCV13 eras, and geographical locations (city). The tree revealed that serotypes assigned to isolates explained the core-genome variations better than other features (isolation sites, cities, and collection dates). The greatest distinctions were especially evident for the significant invasive serotypes 1, 5, and 12F as they clustered into unique clades. Serotypes 16F and 19F were also separated from other serotypes, but the distinction in these serotypes was not as noticeable as in serotypes 1, 5, and 12F. Interestingly, although 12F is invasive, it did not cluster near 1 and 5.

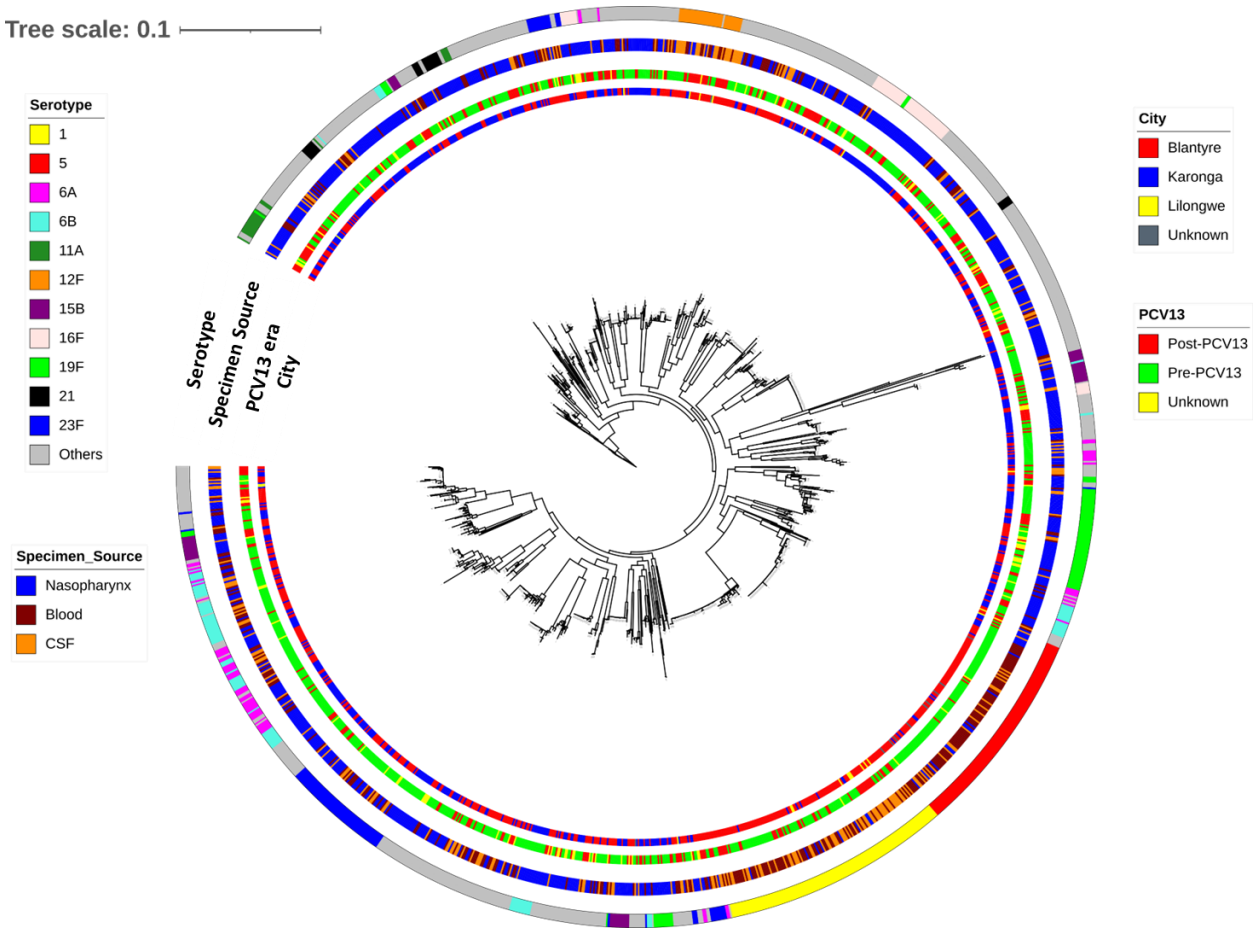


Figure 3-6. The phylogenetic tree drawn from the core-genome alignment.
The tree was built using the ML method and based on the multiple sequence alignment of core genes. Colors on the inner rings show the geographical location (city) and PCV13 era. Color codes on the outer rings indicate the specimen sources (isolation sites) and the serotype of samples. In addition to the significant serotypes 1, 5, 12F, 16F, and 19F, other abundant serotypes, including 6A, 6B, and 23F, as well as serotypes with the source-based p-value < 0.05, including 21, 11A, and 15B, were also highlighted on the tree.

The phylogenetic tree drawn from the multiple sequence alignment of the core intergenic regions is shown in Figure 3-7. The pattern of genetic similarity in the core IGRs was similar to the core genes. The significant invasive serotypes 1, 5, and 12F clustered on the tree as distinct clades implying they were genetically different from other serotypes in both genes and intergenic regions. Serotypes 1, 5, and 12F were also separate from each other, with other serotypes interspersed between them. The level of distinction for serotypes 16, 19F, and 23F was also high within intergenic regions.

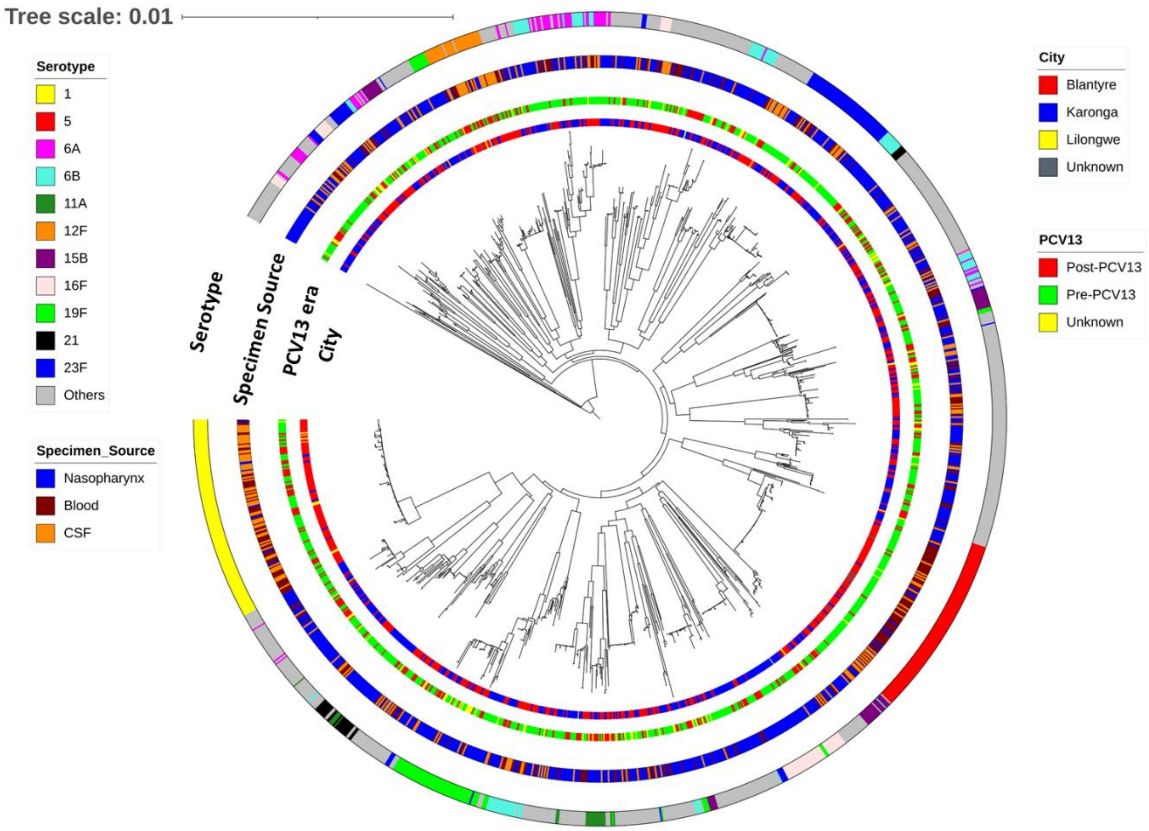


Figure 3-7. The phylogenetic tree drawn from the core-IGRs alignment.
The tree was built using the ML method and based on the multiple sequence alignment of core genes. Colors on the inner rings show the geographical location (city) and PCV13 era. Color codes on the outer rings indicate the specimen sources (isolation sites) and the serotype of samples. In addition to the significant serotypes 1, 5, 12F, 16F, and 19F, other abundant serotypes, including 6A, 6B, and 23F, as well as serotypes with the source-based p-value < 0.05, including 21, 11A, and 15B, were also highlighted on the tree.

3.3.4 Distribution of genes and IGRs

The distribution of genes in the pan-genome is shown in Figure 3-8. The hierarchical clustering showed that the serotype of samples was the main driver of the large-scale gene presence-absence variants in the pan-genome. In other words, specificity in the gene content of samples was not well explained by the specimen sources, PCV13 eras, and geographical locations. In particular, serotypes 1, 5, 12F, 16F, 19F, and 23F were recognizable clusters on the heatmap and had a set of accessory genes that were not frequent in other serotypes. As stated in chapter 2, these serotypes have a frequency greater than 5%. Serotypes 1, 5, and 12F were significantly present in the blood and CSF, and serotypes 16F and 19F were significantly present in the nasopharynx.

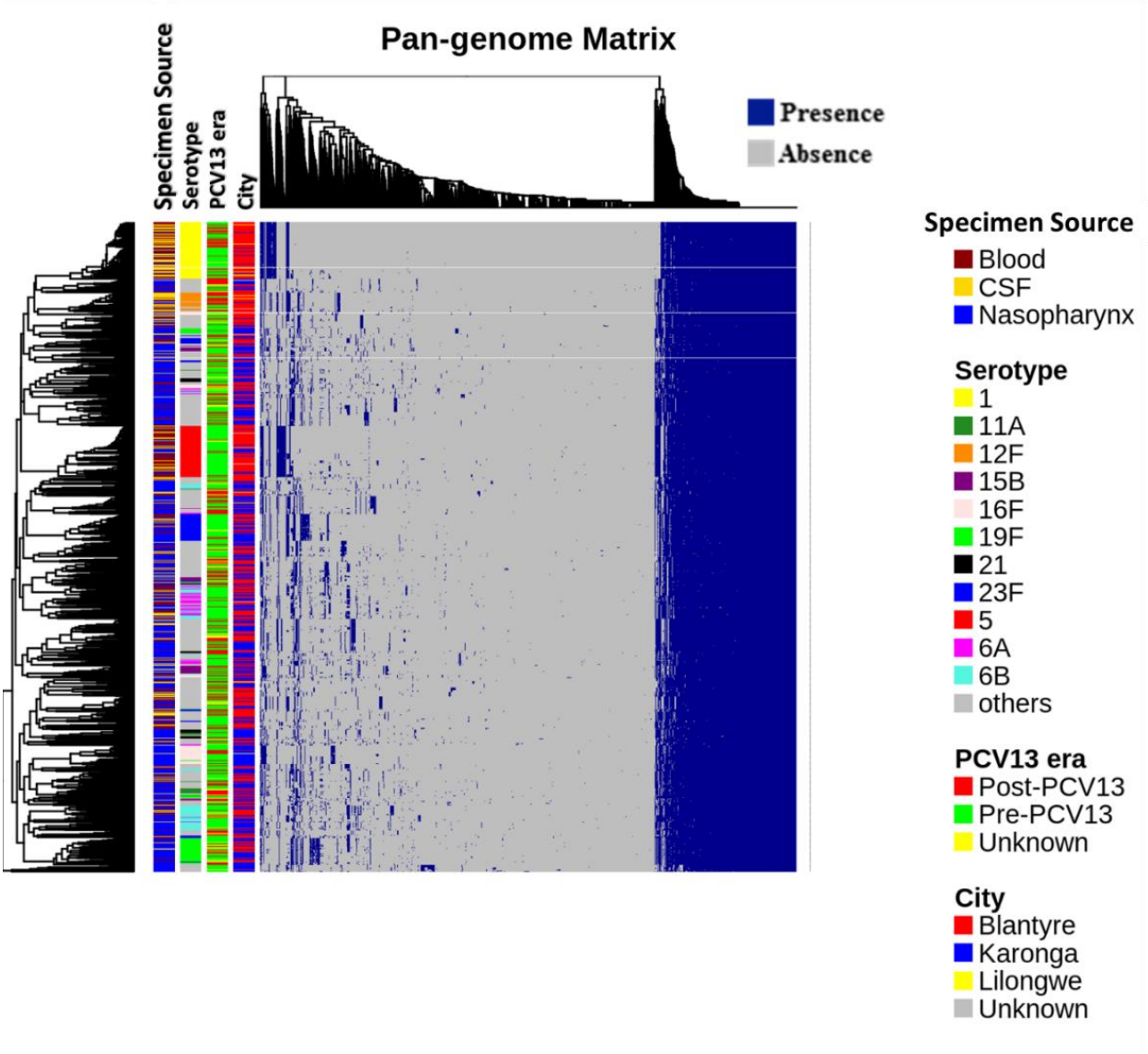


Figure 3-8. The pan-genome matrix shown as a gene presence-absence heatmap.
The heatmap represents the hierarchical unsupervised clustering of samples based on the distribution of genes in the pan-genome. Each row is a sample, and each column is a gene. A blue dot denotes the presence of each gene. On the right side of the heatmap, the large blue block represents core genes present in all samples. The left side of the heatmap shows the accessory genome. The small blue blocks on the left side are accessory genes present only in specific serotypes. In addition to the significant serotypes 1, 5, 12F, 16F, and 19F, other abundant serotypes, including 6A, 6B, and 23F, as well as serotypes with the source-based p-value < 0.05, including 21, 11A, and 15B, were also highlighted on the heatmap.

The distribution of intergenic regions in the pan-IGR is shown in Figure 3-9. Serotypes 1, 5, 12F, 16F, 19F, and 23F have the most remarkable divergence in the accessory IGRs. For both the pan-genome and pan-IGR, the most evident separation from other serotypes belongs to serotypes 1 and 5.

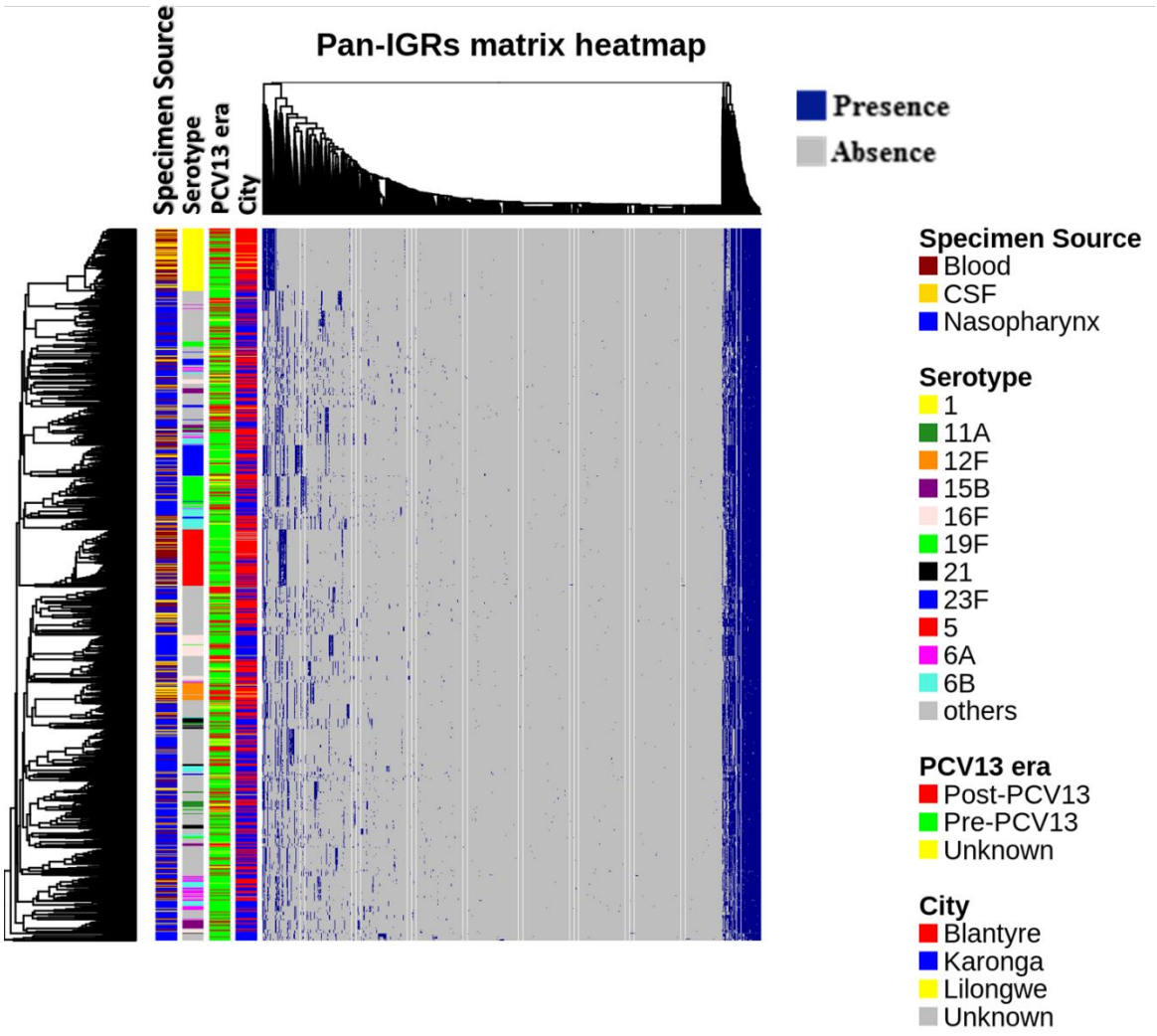


Figure 3-9. The pan-IGR shown as an IGR presence-absence heatmap.
The heatmap represents the hierarchical unsupervised clustering of samples based on the distribution of genes in the pan-IGR. Each row is a sample, and each column is an IGR. A blue dot denotes the presence of each IGR. On the right side of the heatmap, the large blue block represents core IGRs present in all samples. The left side of the heatmap shows the accessory genome. The small blue blocks on the left side are accessory IGRs present only in specific serotypes. In addition to the significant serotypes 1, 5, 12F, 16F, and 19F, other abundant serotypes, including 6A, 6B, and 23F, as well as serotypes with the source-based p-value < 0.05, including 21, 11A, and 15B, are also highlighted on the heatmap.

3.3.5 Functional enrichment of core and accessory genes

Of 729 core genes predicted by Roary, 7.27% (n=53) were reported as hypothetical proteins. 99.31% (n=722) of the core genes were mapped to 722 annotated proteins in the STRING database (one-to-one relationship). STRING reported a set of enriched GO terms in the core-genome listed in Table 3-3. The core genes of Malawian samples significantly contributed to the biological processes at the cellular level

described by GO:0009987. This GO term describes processes in a single cell and communication between cells. Of genes annotated to GO:0009987, half existed in the core-genome of samples from Malawi. Based on the reported molecular functions, core genes mostly had binding functions. In terms of cellular components, core genes mainly encode the subunits of the intracellular organelles.

Table 3-3. The functional enrichment analysis of core genes performed by STRING. GO terms, including biological process, molecular function, and cellular component with FDR < 0.05, were reported.

Biological Process (Gene Ontology)				
GO-term	*description*	*count in network*	*strength*	*false discovery rate*
GO:0009987	Cellular process	207 of 423	0.16	0.0170
Molecular Function (Gene Ontology)				
GO-term	*description*	*count in network*	*strength*	*false discovery rate*
GO:0003676	Nucleic acid binding	82 of 130	0.27	0.0316
GO:1901363	Heterocyclic compound binding	138 of 266	0.18	0.0321
GO:0097159	Organic cyclic compound binding	138 of 267	0.18	0.0321
GO:0005488	Binding	162 of 327	0.16	0.0321
Cellular Component (Gene Ontology)				
GO-term	*description*	*count in network*	*strength*	*false discovery rate*
GO:0110165	Cellular anatomical entity	177 of 332	0.2	0.00025
GO:0005622	Intracellular	141 of 260	0.2	0.0018

The STRING enrichment analysis provided an overview of the enriched GO terms in the core-genome. The pathway enrichment analysis performed by ShinyGO was used to complement the STRING results (Table 3-4). The enriched pathways contained many genes in common. In general, genes participating in DNA replication and repair, homologous recombination, oxidative phosphorylation, and protein biosynthesis were over-represented in the core-genome of samples from Malawi. The network of interactions between proteins from the significantly enriched pathways was depicted in Figure 3-10. The tightness of the network reflects the close direct (physical) and indirect (functional) interactions between core genes. Ribosomal subunits (red nodes) and aminoacyl-tRNA synthetase enzymes that attach amino acids to tRNA (blue nodes) perform gene translation. Silver nodes include a diverse range of enzymes catalyzing DNA replication, repair, and homologous recombination. In summary, the red, blue, and silver nodes were involved in gene expression and translation.

The most noteworthy thing about the network was the presence of the green nodes, including SP_1509 (*atpG*), SP_1510 (*atpA*), SP_1511 (*atpH*), SP_1512 (*atpF*), SP_1513 (*atpB*), and SP_1514 (*atpE*). These genes encode the subunits of *F-type ATP synthase (F1F0-ATPase),* a membrane-bound ion transporter that harnesses energy from the H^+/Na^+ gradient to drive ATP synthesis[180]. Another type of transmembrane ATP synthase in *pneumococcus* is *V-type ATP synthase (V1V0-ATPase),* which differs from *F-type ATP synthase* in its structure[181]. According to the results, the core-genome did not contain any components of the *V-type ATP synthase*.

Table 3-4. The pathway enrichment analysis of core genes sorted by fold enrichment.

Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathways (click for details)
1.4E-11	98	157	1.8	translation, and DNA binding
2.5E-11	103	170	1.8	translation, and DNA binding
7.8E-11	108	184	1.7	cellular macromolecule metabolic process, and ribonucleoside triphosphate biosynthetic process
1.8E-10	77	119	1.9	translation, and ribonucleoside triphosphate biosynthetic process
1.4E-09	71	110	1.9	translation, and ATP synthesis coupled proton transport
8.2E-07	55	87	1.9	Ribosome, and ATP synthesis coupled proton transport
4.3E-06	44	67	1.9	Ribosome, and translation factor activity, RNA binding
1.9E-05	48	78	1.8	Ribosome, and translation factor activity, RNA binding
9.9E-05	39	62	1.9	Ribosome
1.0E-04	36	56	1.9	Ribosome
7.4E-04	32	51	1.8	Ribosome
2.4E-03	15	19	2.3	Aminoacyl-tRNA biosynthesis

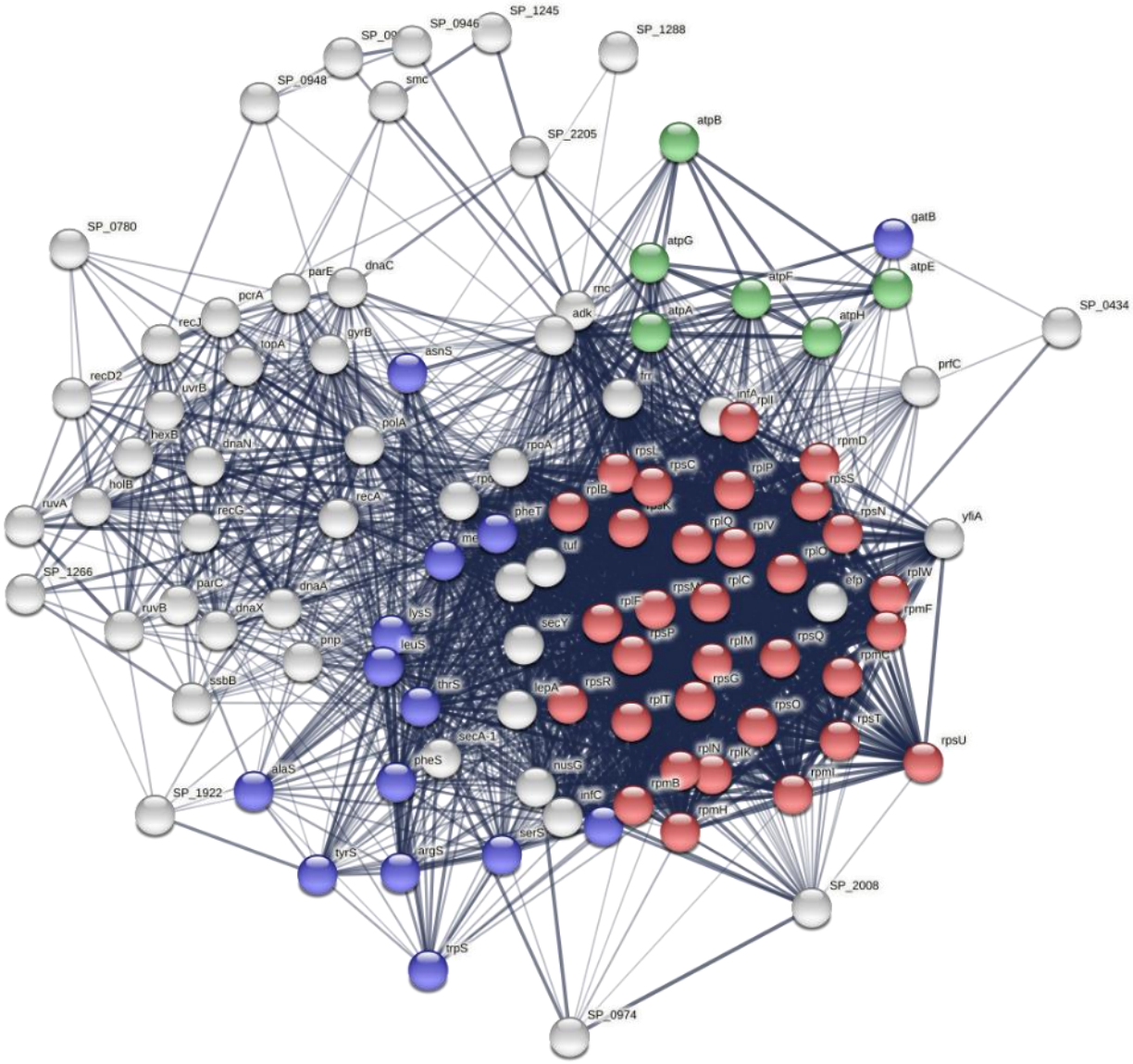


Figure 3-10. Network of interactions between proteins catalyzing the enriched pathways in the core-genome.
Each node represents a core protein. Edges present direct (physical) and indirect (functional) interactions between proteins. The thickness of edges indicates the confidence about the interaction predicted by STRING. Blue nodes catalyze the attachment of amino acids to tRNA molecules. Red nodes are ribosomal subunits. Silver nodes perform DNA replication, DNA repair, and homologous recombination. Green nodes are subunits of F-type ATP synthase.

Of 820 soft-core genes predicted by Roary, 12.68% (n=104) were hypothetical proteins, and 94.63% (n=776) were mapped to 776 proteins in the STRING database (one-to-one relationship). The functional enrichment analysis by STRING and ShinyGO did not find any enriched GO term or pathway in the list of soft-core genes.

Of 5253 accessory genes predicted by Roary, 53.58% (n=2815) were reported as hypothetical proteins. Of all accessory genes, only 24.27% (n=1275) were mapped to 620 proteins in the STRING database, 7.27% (n=382) were mapped to 382 proteins (one-to-one relationships) and 17% (n=893) mapped to 238 recurring proteins (many-to-one relationships). The highest number of repeats belonged to the following proteins:

- SP_1771 (*glyA*), to which 31 accessory genes were mapped. SP_1771 is a glycosyltransferase encoded in a cluster of genes known as Region of Diversity 10 (RD10) and involved in glycosylation of serine-rich repeat protein (PsrP)[182].
- SP_1770 (*glyB*), to which 18 accessory genes were mapped. SP_1770 is another glycosyltransferase adjacent to SP_1771 in RD10, also contributing to the glycosylation of PsrP[182].
- SP_1056 (Tn5252 relaxase) from RD6, to which 18 accessory genes were mapped. Relaxase is a mobilization protein required for the horizontal transfer of genes and plasmids through bacterial conjugation. SP_1056 forms the relaxation complex or relaxosome with the help of other enzymes[183].
- Other proteins with high redundancy in the accessory genome were different versions of Zinc metalloproteases (a virulence factor), bacteriocins (competition peptides), MutT proteins (a housecleaning enzyme), and capsular polysaccharide biosynthesis proteins.

The redundancy in the accessory-genome could be due to the existence of the paralogous genes. Paralogs are homologous genes existing in the same species with similar functions (not necessarily identical). They result from gene duplication or horizontal transfer of genes. Of note was that the paralogous genes did not exist in the core-genome. There was only one version of each core gene in the pan-genome, implying less intra-diversity amongst the core genes than in the accessory-genome (as expected).

The functional enrichment analysis of the accessory genes performed by STRING did not detect any enriched GO terms. This was likely because the accessory-genome comprised a considerable proportion of the pan-genome (77.23%), causing the accessory genes to participate in most biological processes carried out by the annotated genes from the background set. STRING only reported one local network, including genes that contribute to oxidative phosphorylation and transmembrane helix proteins (Table 3-5). Oxidative phosphorylation is associated with the function of *V-type ATP synthases* encoded by genes in the accessory-genome.

Table 3-5. The functional enrichment analysis of core genes performed by STRING. One local network cluster with FDR < 0.05 was reported.

Local network cluster (STRING)				
cluster	*description*	*count in network*	*strength*	*false discovery rate*
CL:3558	Mixed, incl. oxidative phosphorylation, and transmembrane helix	92 of 141	0.35	0.0045

The pathway enrichment analysis can provide more detail about the function of the pneumococcal accessory genes in Malawi. ShinyGO identified the significant mixed pathways listed in Table 3-6. In contrast with the core-genome, the number and variation of enriched pathways in the accessory-genome were higher, involving many uncharacterized proteins.

Table 3-6. The pathway enrichment analysis of accessory genes sorted by FDR.

Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathways (click for details)
2.0E-10	54	80	2.3	mixed, incl. Transmembrane helix, and Helix-turn-helix
2.8E-10	58	90	2.2	mixed, incl. Transmembrane helix, and Helix-turn-helix
7.3E-09	52	82	2.2	mixed, incl. Nucleotide-diphospho-sugar transferases, and Membrane
6.8E-08	17	17	3.4	mixed, incl. Oxidative phosphorylation, and Sodium/solute symporter
1.4E-07	29	38	2.6	mostly uncharacterized, incl. Helix-turn-helix, and CAAX prenyl protease 2
5.5E-07	15	15	3.4	mixed, incl. capsule organization, and UDP-N-acetylglucosamine 2-epimerase
1.6E-06	16	17	3.2	mixed, incl. Glycosyl transferase family 8, and Translocation
1.6E-06	25	33	2.6	mostly uncharacterized, incl. Helix-turn-helix, and CAAX prenyl protease 2
2.3E-06	33	50	2.3	mostly uncharacterized, incl. Glycosyl transferase family 8, and Translocation
2.4E-06	21	26	2.8	mixed, incl. Glycosyl transferase family 8, and Translocation
7.5E-06	25	35	2.4	mostly uncharacterized, incl. Tetratricopeptide-like helical domain superfamily, and Type 2 lantibiotic, SP_1948 family
7.5E-06	25	35	2.4	mostly uncharacterized, incl. capsule organization, and Glycosyltransferase subfamily 4-like, N-terminal domain
1.4E-05	27	40	2.3	mostly uncharacterized, incl. Glycosyl transferase family 8, and Translocation
2.1E-05	29	45	2.2	mostly uncharacterized, incl. Helix-turn-helix, and Type 2 lantibiotic, SP_1948 family
3.9E-05	19	25	2.6	mixed, incl. Tetratricopeptide-like helical domain superfamily, and Type 2 lantibiotic, SP_1948 family
7.2E-05	24	36	2.3	mixed, incl. Oxidative phosphorylation, and YhcH/YjgK/YiaL family
9.3E-05	10	10	3.4	Oxidative phosphorylation
9.3E-05	10	10	3.4	mixed, incl. Glycosyl transferase family 8, and Translocation
5.3E-04	14	18	2.7	mixed, incl. Toxin-antitoxin system, and CAAX prenyl protease 2
5.4E-04	16	22	2.5	mixed, incl. Sortase family, and Immunoglobulin-like fold
9.5E-04	8	8	3.4	mixed, incl. UDP-N-acetylglucosamine 2-epimerase, and Glycosyltransferase subfamily 4-like, N-terminal domain
1.3E-03	14	19	2.5	mixed, incl. Tetratricopeptide-like helical domain superfamily, and Type 2 lantibiotic, SP_1948 family
2.4E-03	10	12	2.9	mixed, incl. ABC transporter transmembrane region, and Recombinase
2.4E-03	16	24	2.3	mostly uncharacterized, incl. Toxin-antitoxin system, and CAAX prenyl protease 2
2.4E-03	10	12	2.9	mixed, incl. CAAX prenyl protease 2, and Bacterial mobilisation
2.7E-03	7	7	3.4	mixed, incl. Sodium/solute symporter, and N-acetylneuraminate lyase
4.8E-03	19	32	2	mostly uncharacterized, incl. MFS transporter superfamily, and ABC-2 family transporter protein

Considering both FDR and fold enrichment, the most significantly enriched pathways in the accessory-genome were:

1. **Oxidative phosphorylation and Sodium/solute symporter** including the following genes:

- SP_1315 (*ntpD*), SP_1316 (*ntpB*), SP_1317 (*ntpA*), SP_1318 (*ntpG*), SP_1319 (*ntpC*), SP_1320 (*ntpE*), SP_1321 (*ntpK*), and SP_1322 (*ntpI*), SP_1323, and SP_1324 that belong to operon RD8a1 in RD8. They encode subunits of *V-type proton/sodium ATPases* that produce ATP from ADP in the presence of an H^+ or Na^+ gradient across the membrane[184]. Unlike the core-genome, the components of *F-type ATP synthase* were not found in the accessory-genome.
- SP_1325, SP_1326, SP_1327, SP_1328, SP_1329, SP_1330 (*nanE*), and SP_1331 from operon RD8a2 in RD8. These genes encode the subunits of the *Sodium/solute symporter* that imports solutes (mostly carbohydrates). Symporter means the channel transports solute and co-solute (in this case Na^+) in the same direction using the energy stored in an inwardly directed sodium gradient. The energy provided by the Sodium gradient is named the Sodium Motive Force (SMF). The SMF is generated by *V-type proton/sodium ATPases*[185]*.* Genes in RD8a1 and RD8a2 operons work together to produce ATP and import Carbohydrates.

2. **Capsule organization, and UDP-N-acetylglucosamine 2-epimerase**, including SP_0346, SP_0347, SP_0348, SP_0349, SP_0350, SP_0351, SP_0352, SP_0353, SP_0354, SP_0355, SP_0356, SP_0357, SP_0358, SP_0359, and SP_0360 from RD3. This region is also known as the *cps* locus and is responsible for capsule biosynthesis.
3. **Glycosyltransferase family 8 and translocation**, including SP_1757, SP_1758, SP_1759, SP_1760, SP_1761, SP_1762, SP_1763, SP_1765, SP_1767, SP_1768, SP_1770, SP_1771, and SP_1772. These genes are from RD10, which is known as *PsrP-secY2A2 pathogenicity island* and is responsible for synthesizing and exporting PsrP. SP_1757 (*gtfB*), SP_1758 (*gtfA*), SP_1765 (*glyF*), SP_1767 (*glyD*), SP_1768 (*nss*), SP_1770 (*glyB*), and SP_1771 (*glyA*) are glycosyltransferases. SP_1759 (*secA2*), SP_1760 (*asp3*), SP_1761 (*asp2*), SP_1762 (*asp1*), and SP_1763 (*secY2*) are secretory components. SP_1772 is the longest gene in RD10 that encodes PsrP. A study showed that *gtfA* and *gtfB* are necessary for PsrP-mediated pneumococcal virulence[186].
4. **Helix-turn-helix and CAAX prenyl protease 2 (mostly uncharacterized)** containing several genes that encode helix-turn-helix proteins. Genes in this pathway were:
    - SP_1332 to SP_1351 from RD8b (RD8b1: SP_1332- 1337, RD8b2: SP_1338-1344, RD8b3: SP_1345-1351) that are mostly uncharacterized. However, SP_1336 is a cytosine-specific DNA methylase that specifically methylates the C-5 carbon of cytosines in DNA.
    - SP_1129-1147 from RD7, most of which are uncharacterized. SP_1129 is an integrase/recombinase putatively required to insert mobile genetic elements.
    - SP_1048-1056 from RD6 or Pneumococcal Pathogenicity Island 1 (PPI1). This region contains genes involved in the movement of Tn5252 transposon and genes that encode the components of the type II toxin-antitoxin system. SP_1051 (*pezT*) is the toxic component, and SP_1050 (pezA) is the antitoxic component. PezT is a toxic compound secreted by pneumococci for competition, and PezA protects the host cell by neutralizing the toxic effect of the cognate toxin PezT. These proteins have a CAAX motif with putative bacteriocin-related functions. C is a cysteine residue, each A is an aliphatic residue, and X is any C-terminal amino acid with different substrate specificity[187].

5. **Tetratricopeptide-like helical domain superfamily, type 2 lantibiotic, and SP_1948 family (mostly uncharacterized)** containing genes from RD2 (SP_0163-0168), RD6 (SP_1057-1063), and RD12 (SP_1947-1954). Genes from RD6 (PPI1) were common with the previous pathway explained above. Other genes were from:
    - RD2 mainly made up of uncharacterized proteins and the following:
        - SP_0163 (PlcR), a transcription factor with a tetratricopeptide-repeat helical domain. The tetratricopeptide repeat is a structural motif consisting of 34 amino acid tandem repeats. PlcR is activated upon binding its cognate quorum-sensing signaling peptide (PapR) on a tetratricopeptide-repeat domain[188]. Quorum sensing is a communication mechanism between pneumococci to regulate biofilm formation, virulence factor expression, competition, and metabolism[189].
        - SP_0165, a flavoprotein enzyme thought to play a role in spore heat resistance[190].
        - SP_0166, a pyridoxal-dependent decarboxylase, and SP_0168, a putative macrolide efflux protein. They contribute to the diffusion of bacteriocins out of the cell.
    - RD12 composed of the subunits of a toxin secretion ABC transporter and genes involved in lantibiotic biosynthesis (SP_1948 and SP_1954). SP_1948, also known as *lant_SP_1948,* consists of a 20-residue block of conserved sequence and represents several type 2 lantibiotic-type bacteriocins distinct from other pneumococcal antibiotics[191]. Lantibiotics are a family of antimicrobial peptides known as bacteriocins that contain amino acids lanthionine and methyllanthionine in their sequences. Lantibiotic genes exist in the bacterial genome and cluster with other genes on the chromosome that facilitate their secretion to the extracellular space[192]. In RD12, the lantibiotic-encoding genes cluster with a toxin secretion ABC transporter. Several genes in RD12 have remained uncharacterized.
6. **Transmembrane helix, and Helix-turn-helix,** this pathway included genes from RD2, RD3, RD6, and RD12, whose functions were described above.
7. **Sortase family, and Immunoglobulin-like,** composed of:
    - Genes from RD4, including SP_0462, SP_0463, and SP_0464 that encode surface proteins, together with SP_0466, SP_0467, and SP_0468 that are putative sortases. Sortase refers to a group of enzymes in gram-positive bacteria that catalyze the assembly of pilins into pili and the anchoring of pili and other surface proteins to the cell wall[193,194]. The sortases are essential proteins for pathogenic bacteria and could be good targets for new antibiotics[195].
    - SP_1038 (uncharacterized protein), SP_1039 (uncharacterized protein), and SP_1040 (site-specific recombinase) involved in DNA recombination.
    - SP_1433 (transcriptional regulator), SP_1434 (ABC transporter, ATP-binding/permease protein), SP_1435 (ABC transporter, ATP-binding protein), SP_1436 (energy-coupling factor transport system substrate-specific component), SP_1437 (energy-coupling factor transport system permease protein), SP_1438 (energy-coupling factor transport system atp-binding protein), and SP_1440 (uncharacterized protein) that encode the subunits of ABC transporters.

Other significant pathways in Table 3-6 had common genes with the pathways mentioned above.

The network of protein-protein interactions between accessory genes contributing to the enriched pathway is depicted in Figure 3-11. The network was obviously not as tight as the network of interactions between core genes with fewer direct and indirect interactions. It could be due to more independence between accessory genes and the presence of more uncharacterized proteins compared to the core genome.

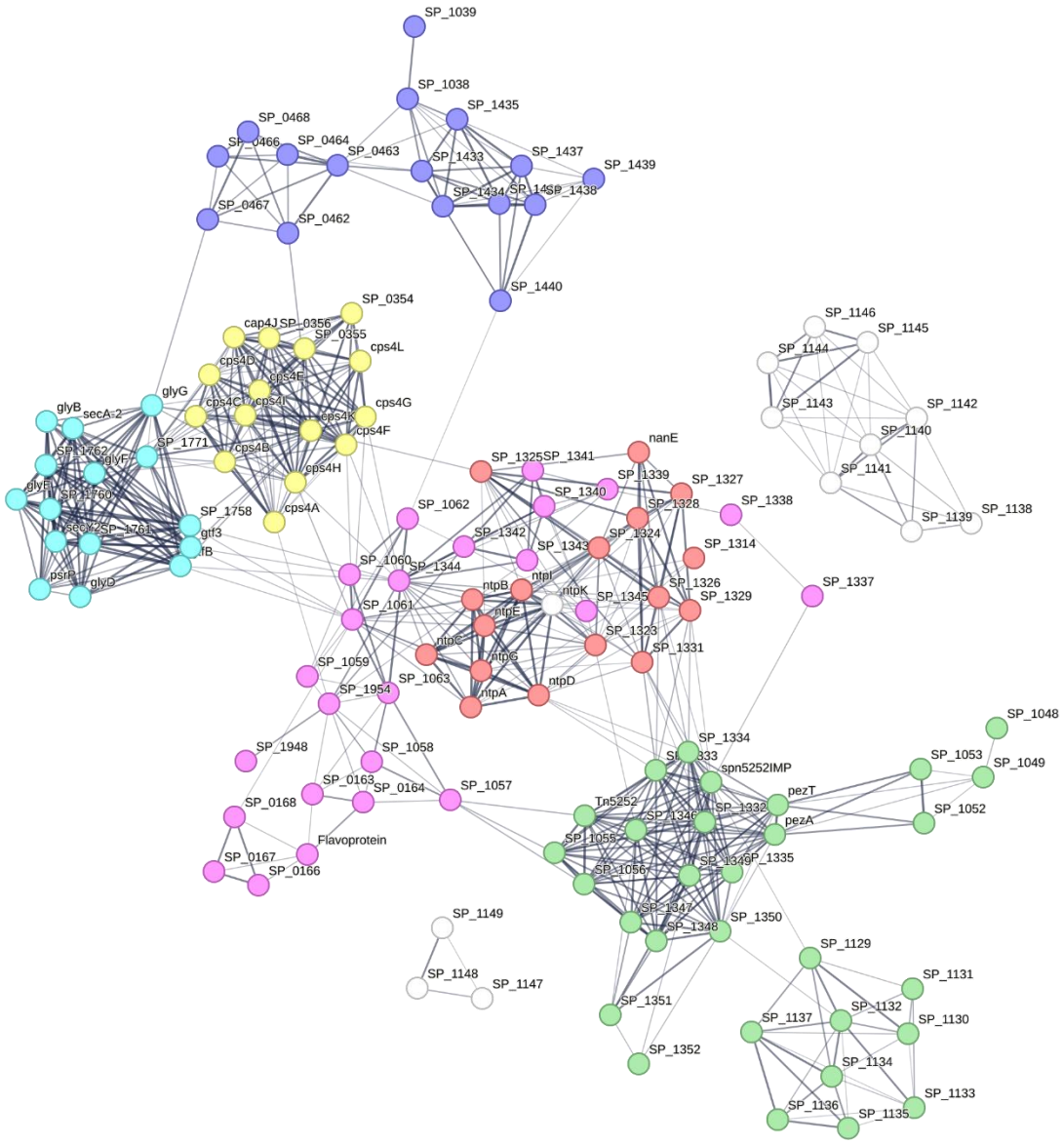


Figure 3-11. Network of interactions between proteins catalyzing the enriched pathways in the accessory-genome. Each node represents a core protein. Edges present direct (physical) and indirect (functional) interactions between proteins. The thickness of edges indicates the confidence about the interaction predicted by STRING. Red nodes are the subunits of *V-type ATPases* and *Sodium/solute symporter* from RD8a. Yellow nodes are enzymes from RD3 that synthesize the capsule. Blue nodes are subunits of *PsrP secY2A2 pathogenicity island* from RD10. Green nodes are from RD6 and RD7 involved in transposing mobile genetic elements. Nodes in pink are from RD2, RD8b, and RD12 participating in communication and competition. Nodes indicated in purple are a mixture of proteins from RD4 that assemble pilins into pili and anchor pili to the cell wall and the subunits of an ABC transporter. White nodes are mostly uncharacterized proteins and not from RDs.

The results showed that the accessory-genome was composed of genes contributing to various and mixed pathways.

3.4 Discussion

The genome sequence of a single pneumococcal strain is approximately two megabases and encodes about 2200 genes[196]. However, the *pneumococcus* genome is highly diverse across different strains. The level of diversity is higher in thirteen recombinogenic regions known as regions of diversity[197,198]. Researchers use various terms to name these regions. Here, we used the terminology suggested by Obert *et al.* [199], calling the regions RDs numbered from RD1 to RD13. These regions contain atypical GC content, several operons, and mobile genetic elements. RDs encode virulence determinants and distinguish invasive from non-invasive strains[200].

The pneumococcal pan-genome is open due to the remarkable genetic diversity between pneumococci. Any additional sample adds new genes to the entire gene set in an open pan-genome. The number of genes in an open pan-genome is directly correlated with the number of genomes. For instance, according to the *Pan-genome Analysis & Exploration* (*panX)* webpage, a pan-genome of 52 pneumococcal samples consists of about 4000 genes, with 31% being core[201]. In this study, using 1477 samples, the total number of genes was 6803, including 729 core genes (11%). As observed, the diversity increases as the number of genomes grows. The advantage of an open pan-genome is that it supplies an unlimited gene repertoire for pneumococci to change their genome structure through recombination and horizontal transfer of genes to diversify and respond to stresses efficiently.

Core and soft-core genes comprised a small proportion of the Malawian pneumococcal pan-genome (22.77%). They have been conserved between more than 95% of samples for eighteen years, from 1997 to 2015. Therefore, the functions of core genes are most likely essential for cell survival, making them potential targets for drug design and vaccine development. On the other hand, accessory genes comprise a considerable proportion of the pan-genome (77.23%). The accessory-genome is the flexible part of the pan-genome; unlike the core genes, the presence of accessory genes in the genome of isolates is not always necessary and depends on the environmental conditions that the bacterium faces. The importance of the accessory genes is their potential involvement in specific traits such as virulence and resistance to antibiotics. The main goal of this research was to identify accessory genes associated with pneumococcal invasiveness in Malawi.

The functional analysis of genes identified differences between core- and accessory-genomes. Firstly

1. The proportion of hypothetical proteins in the accessory-genome (53.58%) was much higher than in the soft-core- (12.68%) and core-genome (7.27.%). By definition, hypothetical or uncharacterized proteins are sequences predicted to be expressed from open reading frames *in silico*, however, no experimental evidence is available to prove their expression and determine their molecular functions *in vivo*. Although some hypothetical proteins could result from computational errors, studies have shown that they make up a substantial fraction of proteomes in prokaryotes and eukaryotes[202].
2. Furthermore, the fraction of genes mapped to the STRING database was also of note, 99.31% of the core and 94.63% of the soft-core genes were mapped to the STRING database, but this value was only 24.27% for the accessory-genome.

From the functional analysis results, we can conclude that the functions of most components in the accessory-genome are still unknown, whereas more is known about the core genes, which generally perform housekeeping functions. This affects our understanding of pneumococcal pathogenesis mechanisms since accessory genes have a different prevalence in different strains and could be potential virulence factors. More effort is required to discover the functions of the hypothetical and uncharacterized pneumococcal genes and proteins.

Secondly, the gene duplication observed in the accessory-genome did not exist in the core-genome. There was only one version of each core gene in the pan-genome, while some accessory genes had different versions. Roary assigned proteins with 95% amino acid similarity to the same cluster. The putative paralogs in the accessory-genome were those genes put in different clusters by Roary but were mapped to the same protein in the STRING database. If the paralogous accessory genes were treated as a single gene, the frequency of that gene would be higher in the pan-genome than any of the paralogous genes. For example, 31 accessory genes were mapped to SP_1771 in the STRING database. These 31 accessory genes could be considered as paralogs in the pan-genome. Some of these paralogs were present in only two samples, whereas some existed in 999 samples, but all of them unified as a single gene were present in 1068 samples. Therefore, caution must be taken during the gene presence-absence statistical analysis if an accessory gene is present only in a particular invasive strain. In this case, the presence of its paralogs with the same function in non-invasive strains must be investigated before reporting it as a virulence factor.

Thirdly, the biological pathways that core and accessory genes carry out differ. Core genes significantly participated in the basic cellular functions. In particular, they encoded the ribosomal subunits and genes contributing to transcription and translation processes. The type of biological processes associated with the accessory genes was more diverse. They particularly encoded membrane proteins and proteins involved in various pathways such as carbohydrate metabolism, ATP synthesis, toxin secretion, and competition. In short, while core genes encoded the conserved genes active in intracellular organelles in the cytoplasm (such as the ribosome), the accessory genes were significantly enriched in enzymes and surface proteins.

Lastly, the most notable difference was the presence of RDs in the accessory- and their absence from the core-genome. Many genes (n=146) from all RDs (RD1 through RD13) were identified in the accessory-genome. In contrast, only one gene from RD1 was found in the core-genome (SP_0073, which encodes an uncharacterized protein). As stated earlier, RDs are areas in the genome with a high divergence across the pneumococcal strains, hence, we do not expect to find RDs in the core-genome. RDs are important because they include several genes that work together in operons. Therefore, for the gene presence-absence statistical analysis (presented later), we can expect to find some RDs specifically present in a set of invasive or non-invasive samples. These may highlight some of the biological pathways contributing to the invasiveness since the overall functions of RDs have been characterized.

In addition to finding the difference between the core- and accessory-genome, the phylogenetic analysis to determine the population structure provided input for the computational pipelines described in the following chapters. The phylogenetic tree revealed which samples were genetically similar and clustered together. The similarity of samples in the core-genome was investigated using multiple sequence alignments of the concatenated sequences of the core genes from all samples. The genetic similarity was explored in the accessory-genome by visualizing the pan-genome matrix and seeing how samples

clustered based on the accessory gene distribution. The results showed that the stratifications of the core- and accessory-genome depend more on the serotype of samples than other factors such as collection time, geographical location, and isolation site. Therefore, serotypes of isolates explained the genetic diversity most effectively and are thus the best option to use as the primary trait of samples to identify the virulence factors.

The significant invasive serotypes 1, 5, and 12F, accounting for 38% of patient samples, were the most divergent strains. They were explicitly characterized as separate clades on the core-genome/IGRs phylogenetic trees and pan-genome/IGRs matrix heatmaps. Serotypes 1 and 5 were the most prevalent amongst patients and the most frequent in the entire Malawian cohort. Considering the gene distribution shown by the pan-genome heatmaps, serotypes 1 and 5 had the highest number of accessory genes that appeared to be absent from other serotype genomes. These specific genes appeared as the discrete blue blocks on the pan-genome/IGRs matrix heatmaps (Figure 3-8 and Figure 3-9). Serotype 12F clustered close to serotype 1 on the pan-genome matrix heatmap. This could be because the gene content of serotype 12F was more like serotype 1 than serotype 5. However, it must be noted that the relative frequency of serotype 12F in the population was only 2.37%, lower than the frequency of serotype 1 (8.73%) and 5 (7.85). Since vaccination decreased the frequency of serotype 5 after 2011, we could conclude that serotype 1 has remained the most persistent and abundant invasive strain in Malawi.

To find the genetic variants such as SNPs, Indels, and genes associated with the virulence, we hypothesized that the genetic variations specifically present in the genome structure of serotypes 1, 5, and 12F may have caused them to have a low colonization rate and become invasive, and this is explored further in the next chapter.

Other distinct clusters on the trees and heatmaps were serotypes 16F, 19F, and 23F. Serotypes 16F and 19F accounted for 14.67% of isolates in the nasopharynx, and they were significantly present in the carrier group. Serotype 23F was abundant in both the nasopharynx (5.21%) and blood (7.67%). Serotype 6B was also abundant in the nasopharynx (7.27%) and blood (5.67%), but it did not separate on the trees and heatmaps like serotype 23F. This may be due to the particular characteristic of serotype 23F, which is known to be a multidrug-resistant clone.

In summary, according to the results in this chapter, serotypes 1, 5, and 12F, which had the highest frequency and invasiveness in Malawi, showed the most significant distinction in their genome structure. Another invasive serotype is 12F, which was not the most frequent but had a high invasiveness, and also showed a high level of genetic distinction on the phylogenetic tree. This suggests that the high genetic divergence of these serotypes may be associated with their virulence. In the following chapters, the results in this chapter were considered to define the trait or groupings of samples for GWAS and gene presence-absence analysis. The main aims of the studies in chapters 4 and 5 were:

- **Chapter 4:** The distribution of small-scale variants (SNPs and Indels) in core genes was investigated to determine the conserved and mutated parts of the core-genome, and a GWAS analysis was performed to identify serotypes with the highest genetic distinction and the variants likely correlated with pneumococcal virulence.

- **Chapter 5:** The distribution of large-scale variants (gene presence-absence) in the accessory-genome was examined to identify genes associated with invasive *Streptococcus pneumoniae*. This chapter sought to answer the project's central question of identifying the putative virulence genes. The importance of RDs was taken into account since they were detected in the accessory-genome with sets of operons contributing to well-described biological pathways.

4 Analysis of small-scale variants in the core-genome

4.1 Overview

Pneumococcus, like other bacteria, continues to evolve not only because of manifold evolutionary pressures such as the host immune system and host genetic background but also increased contact with different host species [10,203]. In addition to large-scale genomic variations such as gene presence and absence, numerous studies showed that SNPs and indels are among clinically meaningful genetic variations for *pneumococcus* [204]. For example, using joint sequencing of human and *pneumococcus* genomes, Lees *et al*. concluded that genetic variation of host and pathogen plays a role in invasive pneumococcal disease so that human variation explains almost half of variation in susceptibility to pneumococcal meningitis and one-third of the variation in severity whereas pneumococcal genetic variation explains 70% of invasive potential but has no effect on severity[205]. Jindal *et al*. suggested that drug resistance can occur through the occurrence of SNPs in *pneumococcus* and identified 31 new potential drug-resistance conferring SNPs [206]. Also, Arends *et al.* identified 79 different missense SNPs in the capsules (*cps*) of 19A serotype isolates (number of samples was 338). They showed significant differences between isolates in nucleotide sugar content and capsule shedding[207]. High-throughput genome sequencing approaches and bioinformatic tools have paved the way for the in-depth identification and study of bacterial pathogen genomic mutational dynamics. We leveraged these approaches for the comparative study of small-scale variations between invasive and non-invasive serotypes.

The core genome is the conserved part of the pan-genome composed of genes present in all samples. Diversity in the core-genome exists in the form of SNPs and indels, which are types of small-scale variants. Although core genes exist in all samples, they may mutate across different strains. In this chapter, the distribution of variants in the core-genome was investigated in the first stage to identify any core gene that was highly conserved and did not mutate over time. Since these conserved core genes have been retained in the pan-genome from 1997 to 2015 and have not mutated during this time, their roles must be vital for cell survival. Therefore, the functional enrichment analysis of the conserved core genes can determine which biological processes are the most fundamental for pneumococci.

In the second stage, the population structure was explored in more detail using the PCA method based on the distribution of variants in the core-genome. This section complemented the population structure analysis performed in the previous chapter. Accurate determination of the population structure is a critical prerequisite for obtaining sensible results from a GWAS analysis because it shows how samples cluster inside the cohort. After determining the population structure, we ran several GWA studies to identify variants potentially involved in specific traits. The trait of samples was defined according to the relation between clusters and invasiveness. The GWAS applies a statistical test to analyze mutations in the samples with different characteristics and links the mutations to features, such as pathogenesis.

The main objectives of the research in this chapter were to determine:

- The most conserved part of the pan-genome. Core genes that are highly conserved between all strains could provide a list of potential targets for drug design.
- What biological functions the conserved core genes perform in the cell.

- The list of mutated core genes that bear small-scale variants (SNPs and Indels).
- The distribution of small-scale variants in the core-genome.
- How samples cluster based on their core-genome structure.
- Which SNPs and Indels might be associated with invasiveness.

The objectives were addressed by following these steps:

- Alignment of the sequencing reads to the core-genome built by Roary in chapter 3.
- Identification of the genetic variants from the alignment files.
- Extraction and functional analysis of the most conserved core genes.
- Identification of the core genes that have mutated across different strains (variable core genes).
- Alignment of the sequencing short reads to a complete and well-annotated *pneumococcus* reference genome that contains all RDs (TIGR4).
- Identification and annotation of variants (SNPs and Indels) from the alignment files.
- PCA of variant distribution in the core genome to identify how samples cluster.
- Defining traits for GWAS based on the cluster of samples that were genetically distinct.
- GWAS to identify the significant mutations associated with the invasiveness.
- Functional analysis of the significant variants identified in the GWAS.

4.2 Methods

4.2.1 Read alignment, QC, and variant calling

For genomic data, variants are saved in files with the *Variant Call Format* (VCF), a tab-separated text format developed to store and annotate sequence variations such as SNPs and Indels[208]. In this study, variants were called from two reference files:

1. The first reference file was generated by extracting core sequences from the pan-genome FASTA file using SeqKit[209]. Variants were detected by aligning the short reads to this FASTA file and stored in VCF format. The distribution of variants was investigated to identify the most conserved part of the core-genome. The advantage of the reference file used in this method was that it contained all core clusters predicted by Roary. However, since the reference genome produced by this method was composed of predicted gene clusters from Roary, accurate annotation of the resulting VCF file was not feasible.
2. To obtain a fully annotated VCF file, variants were also called from a complete reference genome of *Streptococcus pneumoniae* (TIGR4, https://www.ncbi.nlm.nih.gov/nuccore/AE005672.3). TIGR4 is the representative reference genome of serotype 4 that is well-annotated and includes most regions of pneumococcal genetic diversity, such as RDs. This method was suitable for annotating the VCF file using available bioinformatics tools, predicting the effect of variants on genes and proteins, and doing a functional enrichment analysis.

For both reference genomes mentioned above, the following procedure was applied for variant calling. Short reads were aligned to the reference files by Bowtie version 2.3.5[210]. Bowtie is a short-read aligner that is fast and memory efficient. It compresses and indexes the reference genome with a Burrows-Wheeler transformation (BWT)[211]. BWT, also called Block-Sorting compression, is a method for text

compression. BWT re-orders the text and allows compression by algorithms such as move-to-front coding and run-length[212]. Indexing the reference genome makes the alignment fast and memory efficient. The forward and reverse reads had an average length of 125 base pairs from a paired-end sequencing platform. The maximum fragment length for alignment was set to 1000 base pairs to cover the distance between forward and reverse reads (inner distance) and the size of the adapters. Alignment files were saved in the Sequence Alignment/Map (SAM) format and sorted according to the position of mapped reads on the reference genome using Samtools version 1.9[213]. Samtools is a suite of programs designed to manipulate and analyze the alignment files. The mate-related flags were added to the alignment files to tag and remove secondary alignments and mitigate the effects of the PCR amplification. Secondary alignments occur when a short read is mapped to more than one place in the reference genome. Only concordant reads with mapping quality greater than 30 were extracted for variant calling. Concordant reads were aligned to the reference genome with an expected distance and orientation.

SNPs and indels were called by Bcftools version 1.9[214] with the ploidy set to 1 since the *pneumococcus* genome is haploid. Any variant with a missing genotype (the type of variants that could not be determined for some samples), quality less than 50, or depth less than 5000 were excluded from the analysis. Bcftools was also used to sort variants in the VCF files according to their position on the chromosome and split the multi-allelic variants into the bi-allelic variations (variant normalization).

To annotate the variants in the second VCF file (created by mapping reads to TIGR4), SnpEff[215] was used. SNPs and Indels were annotated based on the criteria listed in Table 4-1:

Table 4-1. Criteria applied by SnpEff to predict the effect of variants.

Impact	**Meaning**	**Example**
HIGH	The variant is assumed to have a high (disruptive) impact on the protein, probably causing protein truncation, loss of function, or triggering nonsense-mediated decay.	Stop gained, Frameshift variant
MODERATE	A non-disruptive variant that might change protein effectiveness.	missense_variant, inframe_deletion
LOW	Assumed to be mostly harmless or unlikely to change protein behavior.	synonymous_variant
MODIFIER	Usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact.	exon_variant, downstream_gene_variant

Table from SnpEff manual.

The source code for the read alignment, QC, and variant calling is available in Appendix 1.

4.2.2 Identification and analysis of the conserved and mutated core genes

SNPs and Indels with minor allele frequency (MAF) greater than 0.05 were considered as mutations in the core-genome. To statistically compare the differences in the mutational patterns of core genes, a two-sided binomial test was applied based on the method previously described to identify significantly over-or under-mutated genes in pathogens[216]. Briefly, the frequency of mutations for each core gene was calculated and then compared to the expected frequency based on the ratio of the length of that gene to the total size of the core-genome. After multiple testing correction using the FDR method, the p-values and odds ratio for each core gene were determined using the basic functions in R. Genes with

adjusted p-values < 0.01 and odds ratio > 1 were considered as significantly mutated (polymorphic). Genes with adjusted p-values < 0.01 and odds ratio < 1 were regarded as significantly conserved.

4.2.3 Population structure analysis and identification of subpopulations

The population structure refers to differences within a population that cause deviation from the expected allele frequencies across the individuals in the population. The analysis of population structure is crucial for case-control studies such as a GWAS analysis. It is essential to determine how individuals in the population cluster into subpopulations based on the similarity in their genome structure, as the population structure can cause spurious associations between genotypes and phenotypes in the GWAS analysis. For example, in a disease study, if a high percentage of the control samples are Caucasian and most of the disease cases are Asian, the significant presence of an SNP in the case group would not necessarily be associated with the pathogenesis as the polymorphism might be related to the ethnicity.

Several methods have been developed for population structure analysis. These methods have generally been categorized into two main approaches: parametric and nonparametric. Nonparametric methods are more appropriate for large datasets due to their advantage of not making an assumption on genetic data, efficiency in computational cost, and effectiveness in handling high-dimensional data. The nonparametric methods are further classified into *Dimensional Reduction* and *Distance-based* methods[217].

Principal Component Analysis or *PCA* is a dimensional reduction method for population structure analysis, which maps high dimensional data to a low dimension space and then clusters data. SNPs are the most widely used markers to investigate the variation of DNA sequences between groups of individuals in a population. For the population structure analysis, one vector was assigned to each sample in this study. The vector dimension was equal to the number of SNPs that passed the thresholds applied during the variant calling. The presence and absence of SNPs in each sample were represented by 1 and 0. The R Bioconductor package MixOmics[218] was used to perform the PCA of SNP distribution in the core-genome. The analysis detected the distinct subgroups of samples that could potentially be any feature of the samples, such as isolation sites and serotypes.

The source code for the analysis of population structure is available in Appendix 1.

4.2.4 GWAS analysis

To implement the statistical association test between genotypes and traits of samples, Plink version 1.9[219] was used. Plink is a genotype-phenotype association analysis toolkit that executes statistical tests (such as *Fisher's Exact* test and *Chi-Squared* test) to measure the association between genotypes (SNPs and Indels) and a particular phenotype (e.g., invasiveness). It reports the variants with a significant association with a specific trait represented by a p-value. Since any GWAS analysis considers a set of statistical inferences simultaneously, Plink performs the multiple-comparison correction by applying methods such as Bonferroni and FDR.

The objective of the GWAS analysis in this chapter was to identify the virulence-associated variants. At first glance, the simplest way to identify the virulence-associated SNPs and Indels was to run the statistical tests between samples obtained from the nasopharynx of carriers and those obtained from the blood and CSF of the patients. The hypothesis was that some of the specific differences between the genetic makeup of nasopharyngeal and invasive isolates are linked to pathogenesis. Although this

method could potentially identify some putative virulence-associated variants, it could not thoroughly explain the difference between non-invasive and invasive samples for two reasons:

- Although all isolates from blood and CSF were invasive because they caused illness, the nasopharyngeal population was presumably a mixture of non-invasive and invasive isolates since it was unclear which isolates caused disease after the collection time.
- The population structure was highly stratified, so the comparison across the isolation sites was skewed by the significant presence of serotypes 1 and 5 in the blood and CSF. Therefore, the trait of samples must be defined in agreement with the population stratification and the invasiveness of samples.

To address the above considerations, the second way to identify the virulence-associated SNPs and Indels was to consider the serotype distribution of samples. Serotypes 1, 5, and 12F were significantly present in the blood and CSF and were genetically distinct (Figure 3-6). Therefore, mutations in their genomes could likely be associated with their invasiveness.

The GWAS analysis was conducted using methods to identify the small-scale variants (SNPs and indels) that best describe both diversity in the population and virulence of samples. In the first step, the nasopharyngeal samples were compared to those in blood and CSF. Then, the comparison was repeated after excluding the significant invasive serotypes to determine to what extent these serotypes skewed the results in the first step. Secondly, each of the significant invasive serotypes was independently compared to the rest of the samples. Considering the population structure (Figure 4-9), the number and significance of variants in serotypes 1 and 5 were expected to be very high compared to other strains.

In summary, the GWAS analysis was conducted as follows:

- Comparison across specimen sources:
    - Nasopharyngeal isolates vs. invasive isolates.
    - Nasopharyngeal isolates vs. invasive isolates (with serotypes 1, 5, and 12F excluded).
- Comparison across serotypes:
    - Serotype 1 vs. others (serotypes 5 and 12F were excluded).
    - Serotype 5 vs. others (serotypes 1 and 12F were excluded).
    - Serotype 12F vs. others (serotypes 1 and 5 were excluded).

The fully annotated VCF file, created by mapping the short reads to the TIGR4 reference file, was used as the input file for the association test. The haploid pneumococcal genes were treated as the human mitochondria in Plink. Genotypes with a Bonferroni-corrected p-value less than $5e^{-8}$ were considered as significant variants. The Bonferroni adjustment was suitable for the multiple testing correction since the genome was haploid and very diverse and required an over-correct method to correct the p-values. The Bonferroni method divides the original significance level by the number of tests.

The source code for the GWAS analysis is available in Appendix 1.

4.2.5 Gene functional enrichment analysis

Core-genome and GWAS analyses identified sets of genes, including conserved and polymorphic core genes as well as the lists of genes bearing the significant variants. The functional enrichment analysis was applied to gain insight from those gene sets. The main objective was to determine if the lists of

genes were enriched with specific functional categories and components of particular pathways. It was essential to understand what classes of genes or proteins with specific functions were over-represented or under-represented in the list of interesting genes in contrast with a background set such as a pan-genome. The presence or absence of genes that contribute to a particular pathway might be linked to a trait of samples, such as invasiveness.

For the functional enrichment analysis of genes, the sets of genes were submitted to STRING and ShinyGO, as explained in section 3.2.6.

4.3 Results

4.3.1 Variant statistics, conserved and mutated core-genome

After alignment of reads against TIGR4 reference genome, a total of 31914 SNPs and 206 Indels were identified in the core-genome. Transitions were 3.2 times more frequent than transversions. Conversion between C and T was the most common SNP observed in the core-genome (Figure 4-1).

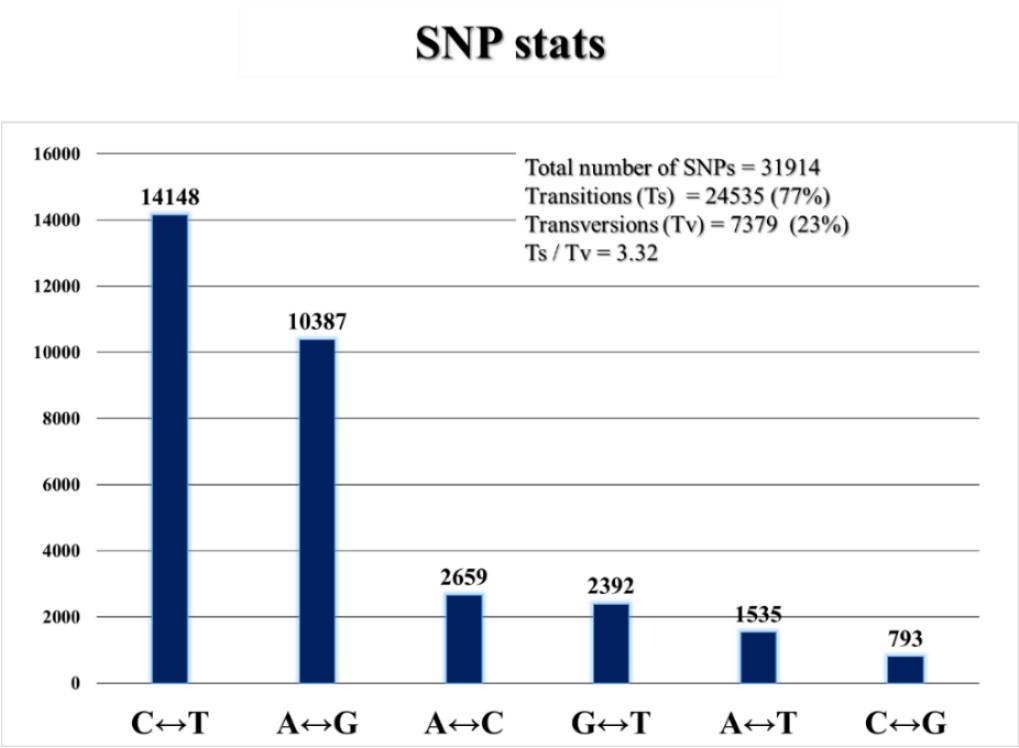


Figure 4-1. Frequency of SNPs in the core-genome.
The transition between C and T is the most frequent SNP in the core -genome.

All core genes harbored at least one mutation in their sequences. However, mutations were significantly less frequent in 14.4% of the core-genome (p-value < 0.01), including 105 genes (conserved core genes). In contrast, the number of mutations was significantly high in 15.64% of the core-genome (p-value < 0.01), including 114 genes (mutated or polymorphic core genes). The distributions of p-values (log-transformed) for the conserved and polymorphic core genes are depicted in Figure 4-2. The most

significant conserved and polymorphic genes (defined as described in the methods) in the core-genome appeared as the outliers of the blue and red boxplots in Figure 4-2. In total, there were six outliers in the list of conserved and eleven outliers in the list of polymorphic core genes. The results of statistical tests and lists of conserved and polymorphic core genes are available in Appendix 4.

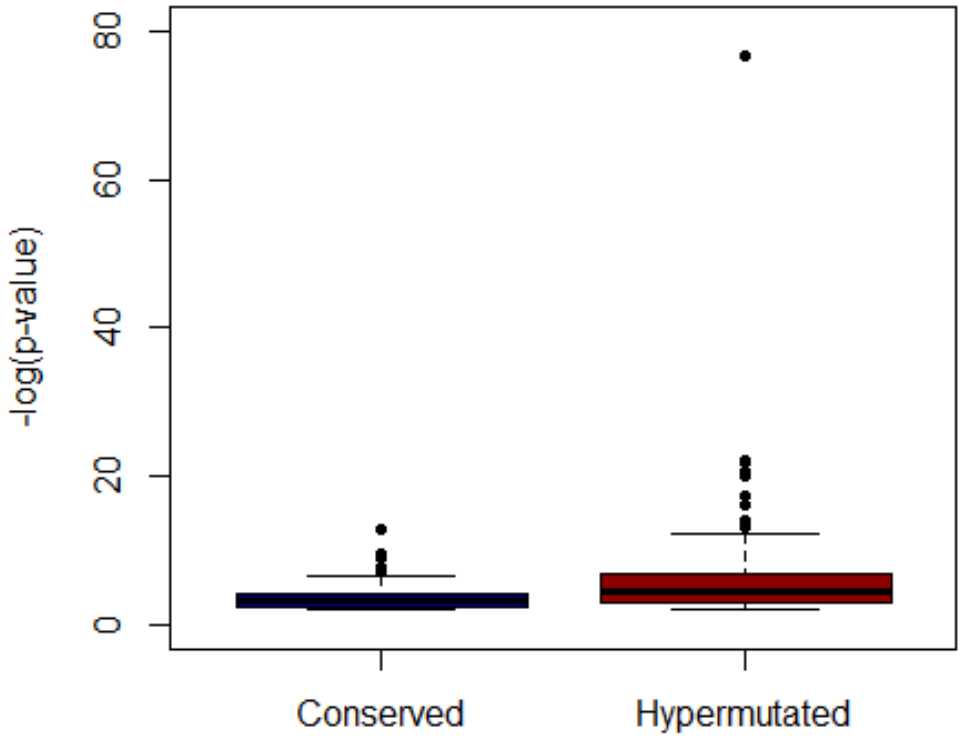


Figure 4-2. The boxplots of the p-value distribution of the 105 conserved and 114 polymorphic core genes. There were six outliers in the list of conserved and eleven outliers in the list of polymorphic genes. One polymorphic core gene was much more significant than others.

The most conserved core genes (outliers of the blue boxplot in Figure 4-2) are listed in Table 4-2. Gene *rpoB* on the top of the list is the beta subunit of the DNA-dependent RNA polymerase that catalyzes the transcription of DNA into RNA. Gene *ropA,* which encodes the alpha subunits of the DNA-dependent RNA polymerase, was also among the conserved core genes with a q-value of 3.6×10^{-3}. There is only one form of DNA-dependent RNA polymerase in prokaryotes, but there are three types of this enzyme in eukaryotes. Based on the results from the Malawian cohort, gene *rpoB* is the most conserved component of the pan-genome, and the RNA polymerase is highly conserved, as expected, given its important function in transcription.

The second gene in the list was SP_0251 that encodes formate acetyltransferase, a key enzyme of anaerobic glucose metabolism that converts pyruvate and coenzyme A (CoA) into acetyl-CoA and formate, which is a step in the fermentation of glucose[220]. SP_0251 has a glycyl radical in its conserved region. The glycyl radical enzymes are inactive in aerobic conditions and interconverted into an active

form only in anaerobic conditions when the stable radical binds to the specific glycine residue at the C-terminal region[221]. The importance of the glycyl radical enzymes is their role in glucose metabolism. SP_0251 is one of these enzymes highly conserved in the core-genome.

There are two other genes in Table 4-2, including *amiA* and SP_1241, that encode the binding proteins of the oligopeptide ABC transporter system. Genes *amiB* (1.7×10^{-4}) and *amiC* (1.7×10^{-4}), that encode the permease subunits in the ABC transport system, were also found to be highly conserved in the core-genome. The ABC transporters exist in eukaryotes and prokaryotes and consist of a membrane transport channel (permease) and a solute binding protein with a high affinity for a specific compound. The binding proteins deliver the external molecule such as carbohydrates to the permease to import it into the cell. In *pneumococcus* and other Gram-positive bacteria, the binding proteins of the ABC transporters are bound to the external surface of the cell membrane[222]. This study showed that the components of the oligopeptide ABC transporter that transfer oligopeptides across the membrane were highly conserved. The solute binding protein of the system showed the highest level of conservation.

Another conserved core pneumococcal gene was *parC* (*topoisomerase iv subunit),* which relaxes the DNA molecule during DNA segregation and cell division. Based on the information in the STRING database, evidence suggests functional links between *parC* and *rpoB*. These two genes are co-expressed in other organisms, such as *Staphylococcus aureus* and *Pseudomonas aeruginosa*. They were both conserved core genes in the Malawian samples, likely due to the functional relation between these two in *S. pneumoniae*.

Finally, the last outlier above the blue boxplot in Figure 4-2 was SP_0892*,* which encodes the R subunit of the type I restriction-modification system. The complex comprises three subunits, including restriction (R), methylation (M), and DNA sequence recognition (S) domains. The M and S subunits recognize specific sequences, and The R subunit cuts the DNA at random, remotely from the recognition site[223]. Interestingly, the results showed that in the pneumococcal samples from Malawi, the R and M subunits of the type I restriction enzyme were encoded by conserved core genes, while the S subunit was an accessory component and not present in all samples. Studies showed that R and M subunits are conserved motifs of type I restriction enzymes[224]. The presence of the S motif in the accessory-genome may change specificity.

Table 4-2. The most conserved genes in the core-genome.

STRING ID	**Description**	**p-value**
SP_1961 (*rpoB*)	DNA-dependent RNA polymerase	1.55E-13
SP_0251	Putative formate acetyltransferase	2.17E-10
SP_1891 (*amiA*)	Oligopeptide-binding protein	1.14E-09
SP_0855 (*parC*)	Topoisomerase IV subunit	2.37E-08
SP_1241	Amino acid-binding protein/permease	4.30E-08
SP_0892	Type I restriction-modification system	6.12E-08

The overall network of the physical (direct) and functional (indirect) associations between all conserved core genes is shown in Figure 4-3. In general, the network is composed of genes that contribute to cell growth, cell maintenance, cell division, and gene expression. The most conserved core gene (*rpoB*) interacts with many nodes in a tight cluster of genes involved in transcription and translation. Other

genes in the network were primarily engaged in carbohydrate metabolism, cell growth, and cell division. These are all processes involved in the core functioning of the cell, so unsurprisingly, they are conserved.

In addition to core genes listed in Table 4-2, examples of the other most significant conserved core genes in the network were:

- *DnaA (Chromosomal replication initiator protein dnaa), dnaK (chaperone protein DnaK), dnaN (beta sliding clamp) and dnaX (DNA polymerase III subunit gamma)* contributing to DNA replication.
- *GuaB (Inosine-5'-monophosphate dehydrogenase)* playing a vital role in cell growth regulation.
- *FtsA (Cell division protein FtsA) and ftsZ (Cell division protein FtsZ)* that control the timing and location of cell division.
- *ClpE (ATP-dependent Clp protease ATP-binding subunit ClpE),* which protects cells from stress by regulating the aggregation and denaturation of cellular structures.

There were also several genes in the network contributing to the translation process:

  - *RpsA (ribosomal protein S1)* that encodes ribosomal protein S1.
  - *FusA (Elongation factor G)* that catalyzes the ribosomal translocation step during translation elongation.
  - *Tuf (translation elongation factor tu)* that promotes the binding of tRNA to the ribosome during protein biosynthesis.
  - Genes such as *thrS (threonine tRNA ligase)*, *aspS (aspartate tRNA ligase)*, *serS (serine tRNA ligase),* and *proS (proline tRNA ligase)* catalyzing amino acid attachment to tRNA molecules.

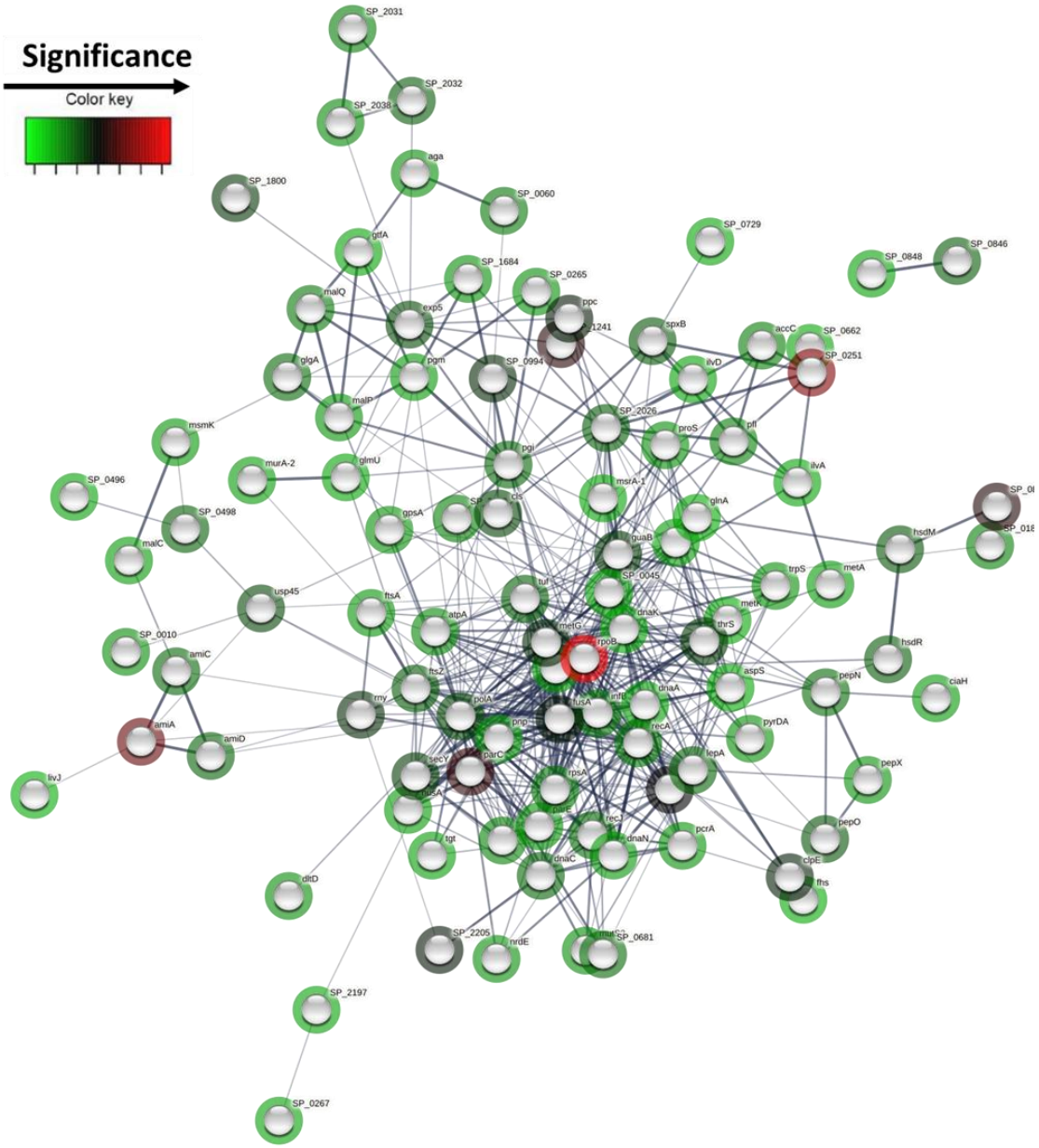


Figure 4-3. The network of interactions between conserved core proteins.
Each node represents a core protein. Edges present direct (physical) and indirect (functional) interactions between proteins. The thickness of edges indicates the confidence of the interaction predicted by STRING. The level of the gene conservation was illustrated by the node colors showing the significance of genes increasing from green to black and red. The more reddish, the more conserved. The reddest node is *rpoB*.

The pathway enrichment analysis of the most conserved core genes showed that these genes encode proteins involved in the central dogma of molecular biology. The names of the enriched pathways in the conserved part of the core genome are listed in Table 4-3 and shown in Figure 4-4.

Table 4-3. The pathway enrichment analysis of conserved core genes.

Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathway
7.5E-04	5	6	16.9	Transcription factor, GTP-binding domain
4.4E-03	4	5	16.2	Translation elongation factor EFTu-like, domain 2
4.4E-03	4	5	16.2	Tr-type G domain, conserved site
9.5E-03	3	3	20.2	Elongation factor EFG, domain V-like
9.5E-03	3	3	20.2	Aminoacyl-tRNA synthetase, class II (G/ P/ S/T)
9.5E-03	5	11	9.2	Small GTP-binding protein domain

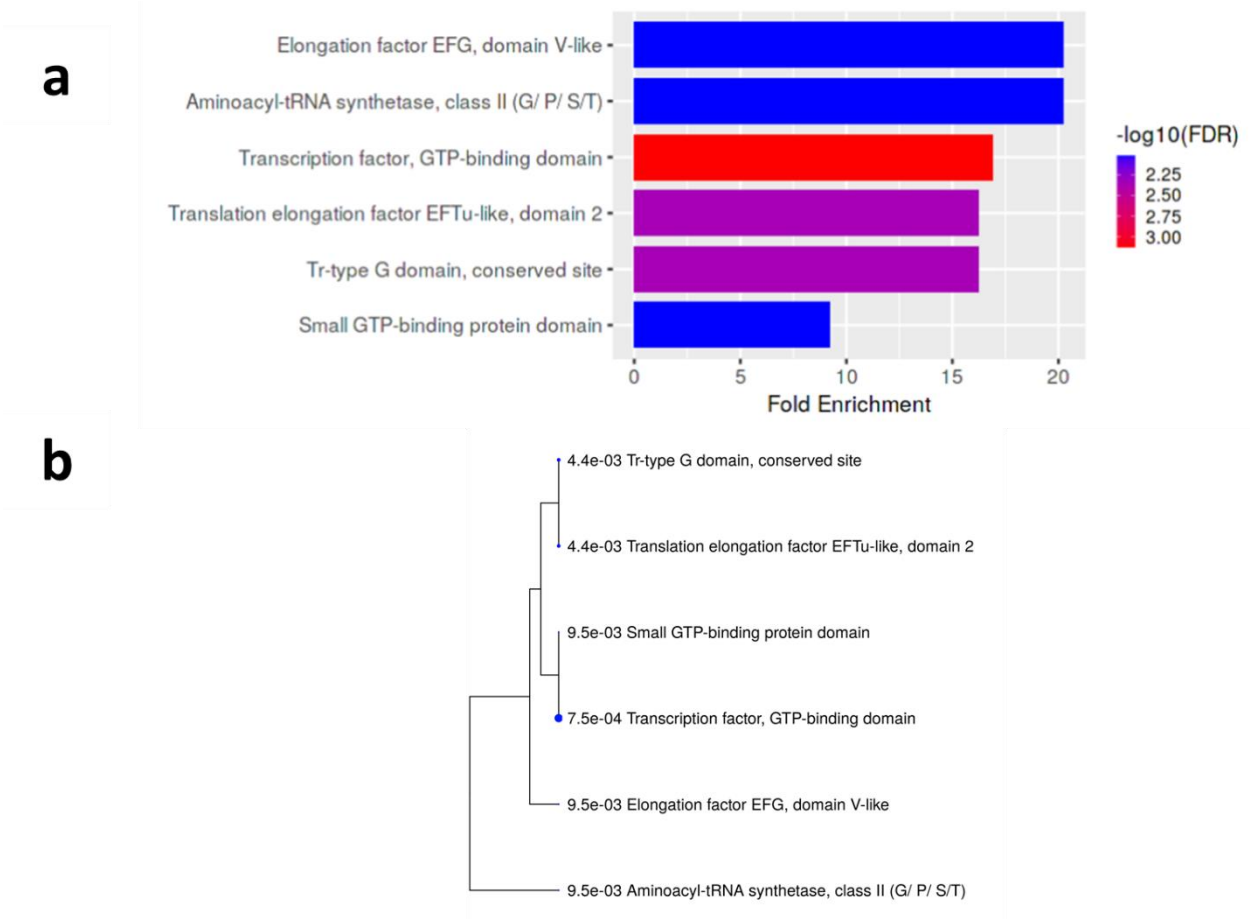


Figure 4-4. The pathway enrichment analysis of conserved core genes.
(a) Visualization of the pathway enrichment analysis of the conserved core genes. The enriched pathways were sorted by the fold enrichment and colored according to their significance. (b) A clustering tree illustrating the correlation between significant pathways. Pathways with many shared genes were clustered together.

The network of genes involved in the enriched pathways listed in Table 4-3 is shown in Figure 4-5. According to the information from STRING, all of these genes are involved in the translation process. Nodes *serS*, *proS*, and *thrS* are aminoacyl-tRNA synthetases that catalyze the attachment of serine,

proline, and threonine to their corresponding tRNA molecules. Gene *infB* is a translation initiation factor (if-2) required to initiate protein synthesis, protect tRNA from hydrolysis, and promote its binding to the 30S ribosomal subunits. Moreover, this protein is involved in GTP hydrolysis during the formation of the 70S ribosomal complex. Genes *tuf* and SP_0681 are translation elongation factors that catalyze the binding of aminoacyl- tRNA to the A-site of ribosomes during protein biosynthesis. Gene *fusA* is the translation elongation factor that promotes ribosomal translocation during translation elongation. Gene *lepA* encodes the elongation factor 4 that regulates the accuracy of protein synthesis by controlling ribosome movement.

According to the STRING report, the functional links between proteins in Figure 4-5 included neighborhood in the genome, gene fusion, co-occurrence, and co-expression.

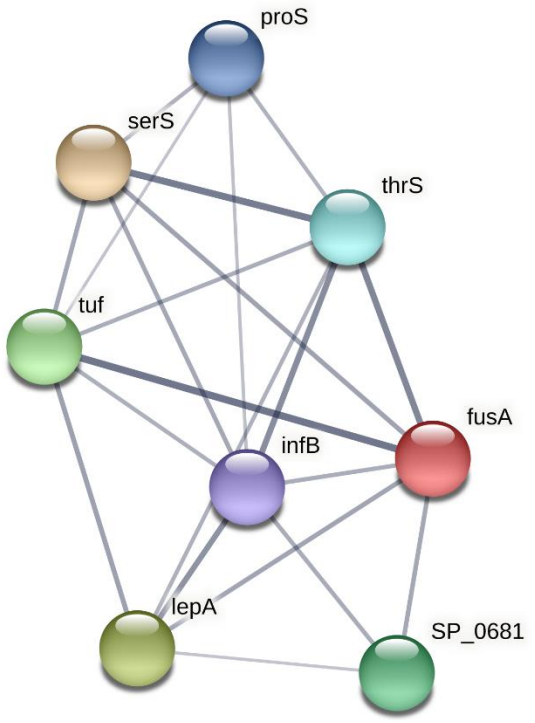


Figure 4-5. The network of interactions between conserved core genes involved in the significantly enriched pathways. The STRING webpage generated the figure using genes involved in the pathways from Table 4-3. Each node represents a conserved core protein. Edges present direct (physical) and indirect (functional) interactions between proteins. The thickness of edges indicates the confidence about the interaction predicted by STRING. The network consists of translation initiation factor (*infB*), translation elongation factors (*tuf, fusA*, and SP_0681), translation fidelity factor (*lepA*), and genes that catalyze the attachment of amino acids to tRNA (*serS*, *proS*, and *thrS*).

The eleven outliers of the significant polymorphic core genes (outliers of the red boxplot in Figure 4-2) are listed in Table 4-4. The mutation rate in SP_1247 (*smc*) was remarkably higher than other core genes. This gene has a length of 3540 base pairs with 367 polymorphic loci. It encodes the chromosome segregation protein required for chromosome condensation and partitioning. SMC is a family of ATPase proteins found in all organisms. As its name implies, it functions in chromosomal interactions such as condensation, recombination, DNA repair, and epigenetic silencing of gene expression[225]. The sequence

homology of SMC in eukaryotic species is confined to amino- and carboxyl-terminal domains suggesting that the regions between two terminals are more prone to mutations[226].

For the pneumococcal samples from Malawi, although the polymorphic loci were evenly distributed along the *smc* gene, the multiallelic loci were not found in the regions of the gene that encode the amino- and carboxyl-terminal domain (Figure 4-6). Studies showed that in *Bacillus subtilis,* mutations in *smc* slow down cell growth[227]. In this study, to find an association between mutations in *smc* and the trait of samples, a GWAS analysis was required to determine if *smc* harbors any virulence-related mutations.

Table 4-4. The most polymorphic genes in the core-genome.

STRING ID	**Description**	**p-value**
SP_1247 (*smc*)	Chromosome partition protein	1.93E-77
SP_2066 (*thrC*)	Threonine synthase	7.75E-23
SP_0254 (*leuS*)	Leucine tRNA ligase	1.29E-22
SP_1670 (*murF*)	UDP-N-acetylmuramoyl tripeptide D-alanyl-D-alanine ligase	2.07E-21
SP_1380	Uncharacterized protein	9.54E-21
SP_0377 (*cbpC*)	Choline binding protein C	3.24E-18
SP_0127	Uncharacterized potein	3.97E-18
SP_0581 (*pheT*)	Phenylalanine tRNA ligase beta subunit	8.77E-17
SP_0948	Phosphate starvation-inducible protein	9.37E-15
SP_1623 (*exp7*)	P-type ATPase metal cation transport	2.20E-14
SP_0454	Uncharacterized protein	7.84E-14

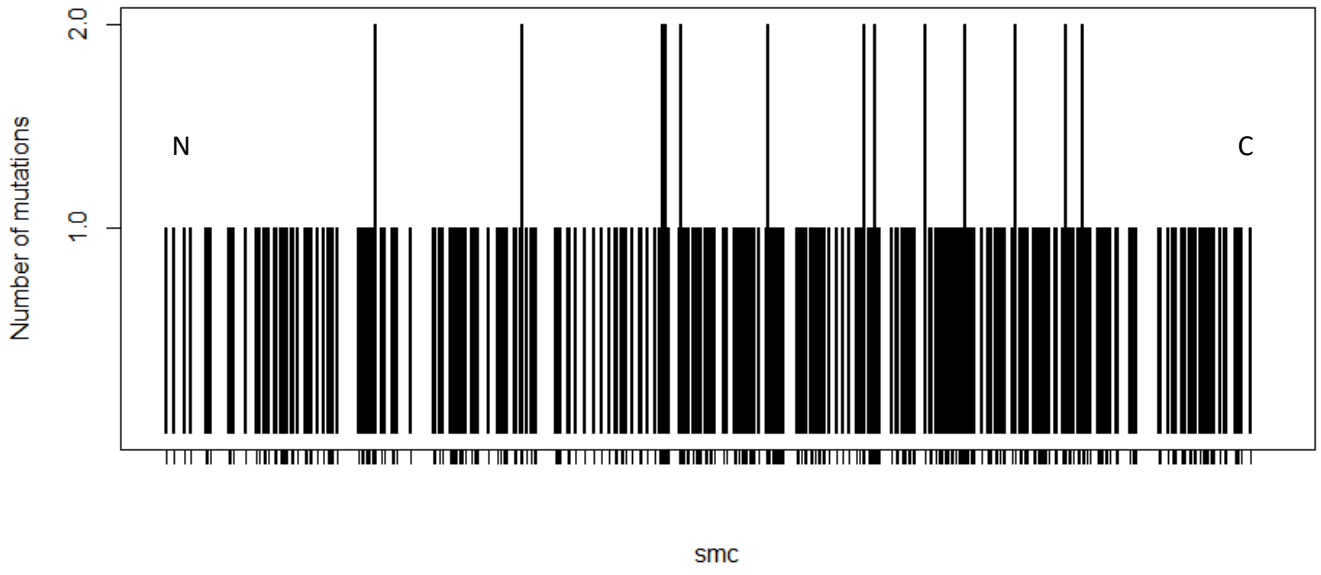


Figure 4-6. Distribution of the polymorphic sites in gene *smc*.
The mutation rate is higher in the middle of the gene. Loci with more than one mutation do not exist in loci near the ends of the gene that encode the amino- and carboxyl-terminal domain of the protein.

Gene *thrC* is the second record in Table 4-4. It catalyzes the final step of the threonine synthesis, a polar, uncharged, essential amino acid found in peptide linkages in proteins.

There are also two polymorphic tRNA ligases in Table 4-4, *leuS* and *pheT,* which attach leucine and phenylalanine onto the corresponding tRNA. Since some of the other tRNA ligases were identified as the conserved core genes in the previous section, including those that attach proline, serine, tryptophan, aspartate, and threonine, the level of conservation in tRNA ligases may depend on the type of amino acids that the ligase enzymes attach to the tRNA molecules.

MurF is another polymorphic core gene in Table 4-4. MurF belongs to the family of Mur enzymes (MurA-F) that regulate different steps of peptidoglycan biosynthesis. According to the results of this study, the level of polymorphism is only high in the MurF enzyme. In gram-positive bacteria such as *S. pneumoniae, murF* catalyzes the final and critical step of peptidoglycan synthesis in cell wall formation. Studies suggest that *murF* may also be involved in controlling cell division. Since its function is unique to bacteria, it is a putative target for developing antibacterial agents. However, its polymorphic structure may enable MurF to acquire drug resistance easily[228].

The next polymorphic core gene in Table 4-4 was *cbpC,* which encodes choline-binding protein C (CbpC). There are twelve types of choline-binding proteins (CBPs) in *S. pneumoniae*. CBPs are a family of surface proteins that bind to the phosphorylcholine residues in the cell wall and help the pathogen attach to the surface of the host's cells and thus play a major role in colonization. CbpC is one of the most abundant choline-binding proteins in *pneumococcus*[229]. In addition to *cbpC* (p-value = 3.24×10^{-18}), *cbpE* (p-value = 2.06×10^{-5}) was also in the list of polymorphic core genes. Studies showed that mutations in *cbpE* are associated with the reduced colonization rate of invasive strains 4 and 6 in the nasopharynx[230]. Other pneumococcal choline-binding proteins (*cbpA*, *cbpF*, *cbpI*, and *cbpJ*) were found in the accessory-genome, and none were identified in the list of conserved core genes, suggesting that CBPs are not conserved components of the pan-genome.

Another gene on the list, SP_0948, has not been fully annotated yet, but is a kind of ATPase induced by phosphate starvation and belongs to the phosphate regulon family[231].

Gene *exp7* in Table 4-4 encodes an ATPase that uses ATP energy to catalyze cation (H+, Na+, K+, Mg2+, Ca2+, and Ag+) uptake. There are several reported mutations in P-type ATPases in different species[232]. Mutation in P-type ATPase reduces virulence In *Listeria monocytogenes*[233]*.* For this research, a GWAS analysis could determine whether mutations in *exp7* were linked to any particular characteristic of the pathogen, such as colonization rate and virulence.

The network of interactions between the polymorphic core proteins is shown in Figure 4-7. The network was obviously not as tight as the interactions between conserved core proteins. This is likely due to the more functional independence and the variety of biological pathways in which polymorphic core genes participate. Genes in the cluster on the left side of the network primarily contribute to different metabolic pathways. The most significant polymorphic genes in this cluster were:

- Gene *glmS* catalyzing the first step in the hexosamine pathway, the products of which are precursors for the biosynthesis of macromolecules that contain amino sugars.
- Genes *purF* and *purK* that have roles in purine metabolism.

- Genes *trpE* and *trpB* that are neighbors in the *pneumococcus* genome. These genes promote the biosynthesis of anthranilate and L-tryptophan.
- Gene *SP_0128* that acetylates the N-terminal alanine of ribosomal protein S18.
- Gene *manA* that participates in capsular polysaccharide biosynthesis and D-mannose metabolism.
- Gene *ssbB* encoding the single-strand DNA-binding protein, which plays a vital role in DNA replication, recombination, and repair. It binds to a single strand of DNA and a set of enzymes and recruits them together during DNA metabolism

Genes in the cluster on the right side of the network in Figure 4-7 were mostly uncharacterized proteins. Submission of polymorphic core genes to ShinyGO did not identify any significantly enriched pathways.

In summary, there was a direct correlation between the level of conservation and the number or connectedness of protein-protein interactions. The network of interactions between all core proteins is depicted in Figure 4-8. Most of the conserved core genes are indicated in green in the center of the network with many interactions. In contrast, the polymorphic core genes lie mainly around the outside of the network with fewer interactions.

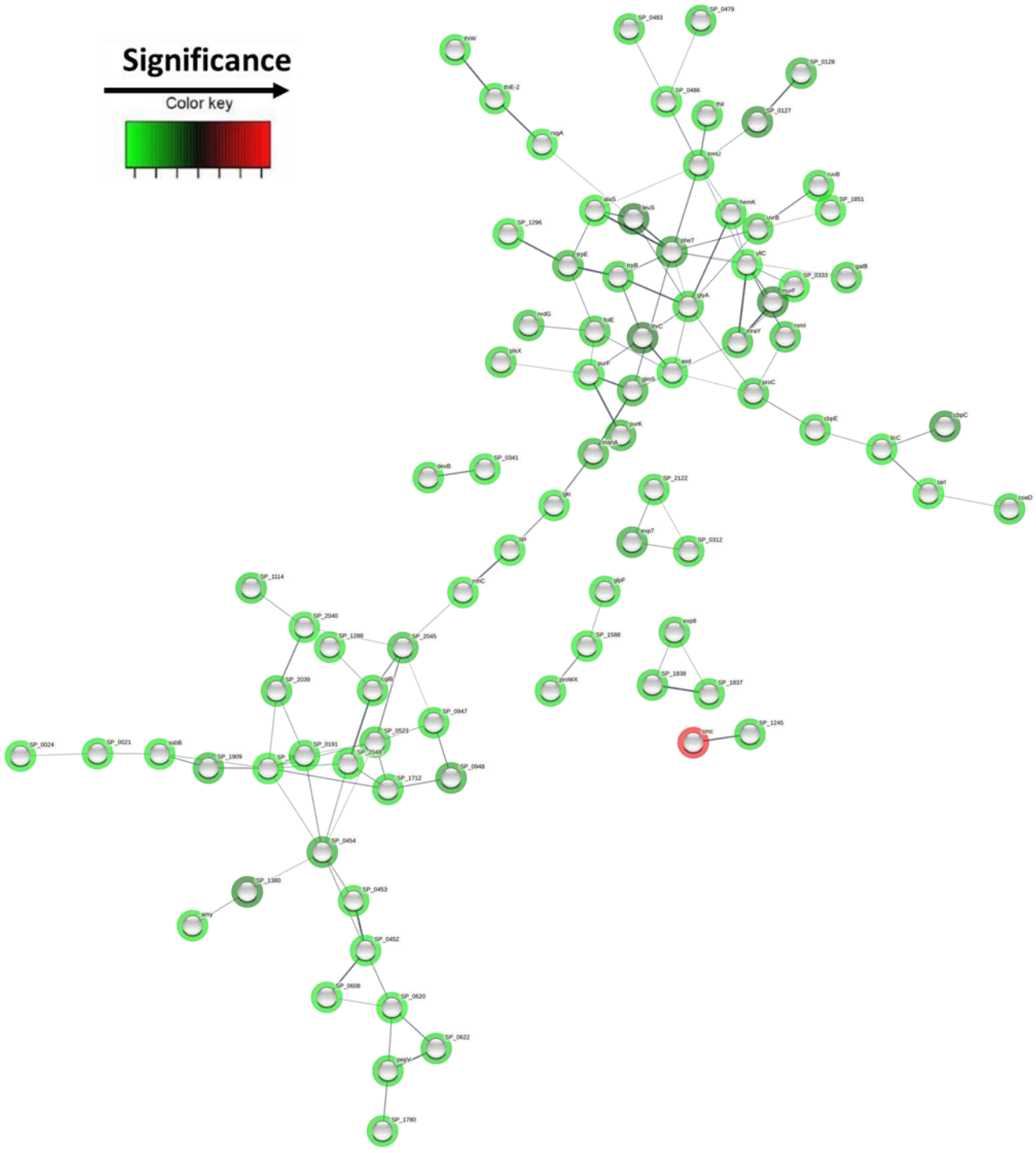


Figure 4-7. The network of interactions between polymorphic core proteins.
Each node represents a core protein. Edges present direct (physical) and indirect (functional) interactions between proteins. The thickness of edges indicates the confidence of the interaction predicted by STRING. The level of the variation was illustrated by the node colors showing the significance of genes increasing from green to black and red. The more reddish, the more polymorphic. The reddest node is *smc*.

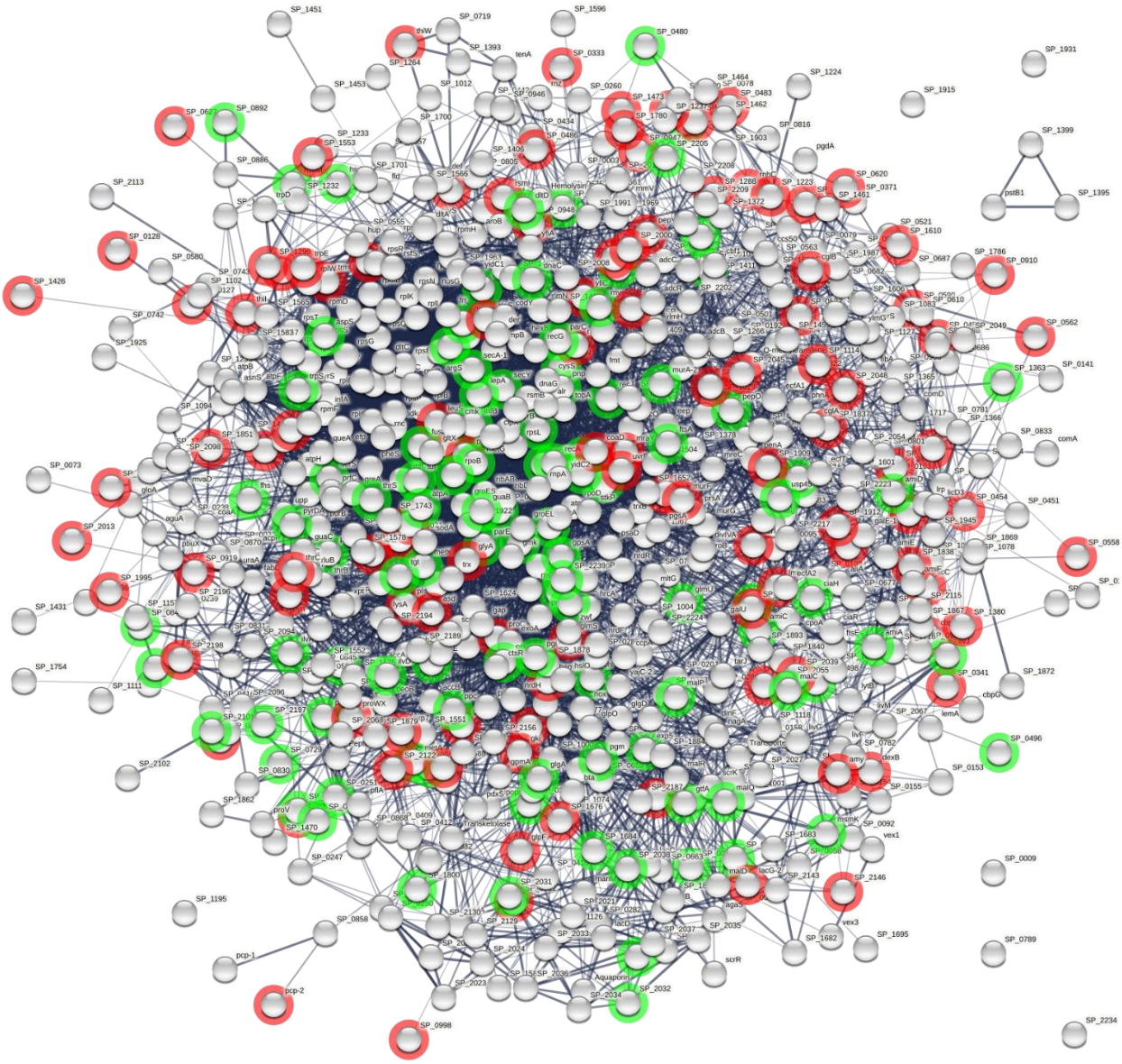


Figure 4-8. The network of interactions between all core proteins.
Each node represents a core protein. Edges present direct (physical) and indirect (functional) interactions between proteins. The thickness of edges indicates the confidence of the interaction predicted by STRING. Conserved core genes are shown in green, and polymorphic core genes are highlighted in red.

The conserved and polymorphic parts of the core-genome were characterized above, however, a GWAS analysis should be helpful to determine the relationship between genotype and phenotype of the samples. In the next step, samples were clustered according to their genetic similarity. Then statistical tests were applied to identify the association between the genetic variants and the trait of the isolates, such as invasiveness.

4.3.2 Population structure

The first step in a GWAS analysis is to categorize samples into control and case groups. Isolates in the population should cluster in a PCA based on their genetic distinction. Statistical analysis of the variant distribution between case and control groups would identify the association between genotypes and phenotypes. The objective of this section was to identify SNPs and Indels associated with pneumococcal virulence in Malawi. At first glance, a comparison between the carriage and patient groups could answer this question because, at the time of collection, the isolates from the carriage group did not cause disease, but those isolated from patients did. However, this method can introduce false positives and false negatives. The reason is that it was unclear which isolates from the nasopharynx of carriers progressed to disease after the date of sample collection. Meanwhile, other factors influencing the population structure, such as collection times, serotypes, and geographical locations, must be considered.

The PCA using the variant distribution determined which characteristic of samples, including isolation sites, serotypes, collection times, or geographical locations, had the highest impact on the population structure. A vector representing the presence or absence of 32120 variants (31914 SNPs and 206 Indels) was created for each sample. PCA was applied to reduce the dataset's dimensionality to observe the main clusters and outliers in the population. It used the vectors' coordinates to calculate the genetic similarity between samples.

As previously described by the phylogenetic analysis (Figure 3-6), diversity in the core-genome was well-explained by the serotype of samples. Separate clades on the phylogenetic tree represented different serotypes. The PCA of the variant distribution in the core-genome gave concordant results, illustrated in Figure 4-9, that showed four main clusters in the population. However, most samples (92%) belonged to three of them.

In Figure 4-9, the PCA results were visualized in two-dimensional space. As illustrated in panel *a*, the isolation sites (specimen sources) could not explain the population stratification because there was a mixture of samples from different isolation sites (the nasopharynx, blood, and CSF) in each distinct cluster. Therefore, at a particular time, a random set of samples from the nasopharynx is not genetically different from a set of samples from sterile sites.

Panel *b* showed that the distinct clusters were serotypes 1 and 5, demonstrating that these two serotypes have a distinction in the core-genome and harbor a set of mutations that separate them from other isolates. The separation was more evident for the cluster of serotype 1, with the most considerable distance from different serotypes. Thus, serotype 1 had the highest divergence from other serotypes in the core-genome. Chapter 2 showed that serotypes 1 and 5 were the most frequent strains in Malawi. Additionally, they had a low carriage rate and a significant presence in the patient group (Figure 2-15). One hypothesis is that the distinction in the core-genome of serotypes 1 and 5 might be associated with their invasiveness.

Panel *c* demonstrates the effect of vaccination on bacterial evolution. The results from samples collected in this study showed that vaccination did not cause any significant mutations in the core-genome of the dataset. Samples were assigned to pre- and post-PCV13 categories based on their collection times before and after introducing the vaccine in Malawi in November 2011. There were samples from both pre- and post-PCV13 eras in all of the four main clusters identified by the PCA,

suggesting that the vaccination program did not cause any genetic divergence in the cohort. However, more investigation is required to explore the effect of vaccination as it is not clear which samples collected after November 2011 were obtained from a person who got vaccinated. Moreover, the population was not big enough to draw a solid conclusion.

In panel *d*, the four subpopulations were colored according to the geographical locations of isolates which also did not seem responsible for describing the population structure. The dataset in this study was entirely biased for the location of the samples, about 95% of isolates collected from patients were from Blantyre, and none of them were from Karonga. On the other hand, about 15% of nasopharyngeal isolates were collected from Blantyre, and about 85% were from Karonga. Nonetheless, panel *d* showed that the structure of the core-genome was not affected by the geographical locations of samples; samples from different cities had a similar core-genome.

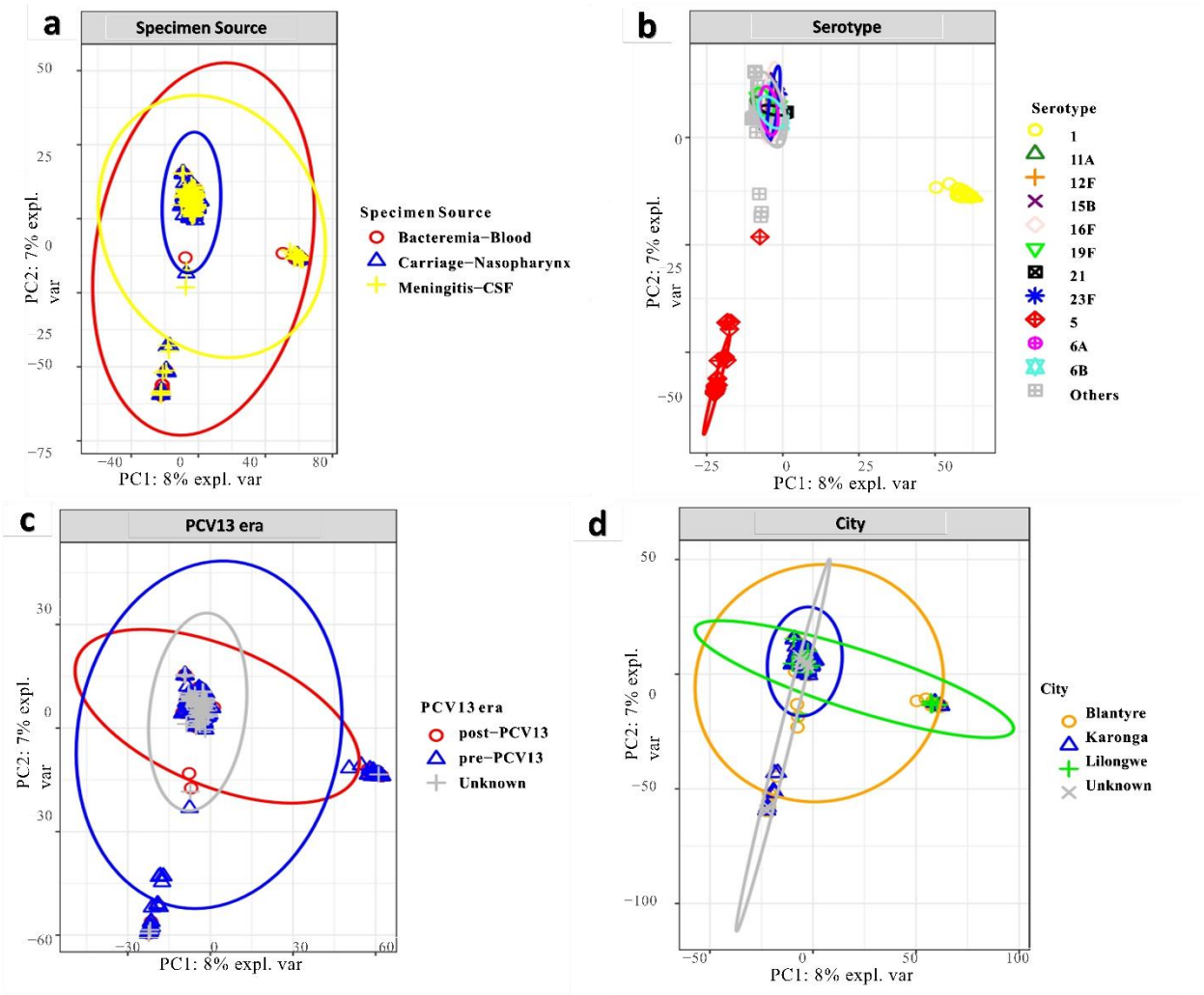


Figure 4-9. The PCA is based on the distribution of SNPs in the core-genome.
The figure shows the effect of specimen sources (in panel *a*), serotypes (in panel *b*), vaccination eras (in panel *c*), and geographical location (in panel *d*) on the core variant profile of pneumococcal isolates. Genetic diversity in the core-genome was explained well by serotypes. Serotypes 1 and 5 were evidently distinct.

In conclusion, serotypes 1 and 5 had the highest divergence from other serotypes in the core-genome and bore mutations that did not exist in other strains. That is why they appear as separate clusters on the PCA plots (Figure 4-9). As shown in chapter 2, serotypes 1 and 5 had the highest invasiveness in Malawi, therefore, some core mutations in serotypes 1 and 5 could potentially be linked to their invasiveness.

4.3.3 GWAS analysis

The simplest way to identify the potential virulence-associated SNPs and Indels is to run statistical tests between the nasopharynx and sterile sites. Isolates from blood and CSF were considered to be invasive because they caused disease. The hypothesis is that some mutations in their genomes were linked to pathogenesis that enable invasiveness. Although this method could potentially identify some putative virulence-associated variants, it may not fully characterize the association between variants and virulence. As stated earlier in this chapter, the carriage group most likely contained both non-invasive and invasive isolates; it was unclear which nasopharyngeal isolates progressed to disease after sample collection. Another issue was the population structure, which was highly stratified by serotypes 1 and 5. Thus, the comparison results across the isolation sites would be skewed by the significant presence of serotypes 1 and 5 in the blood and CSF. A second hypothesis could be that since serotypes 1, 5, and 12F were significantly abundant in the blood and CSF and were genetically distinct, mutations in their genomes could be associated with their short colonization period and invasiveness.

The GWAS analysis was conducted using both hypotheses to identify all putative small-scale variants (SNPs and indels) that best describe diversity in the population and virulence of samples. In the first step, the nasopharyngeal samples were compared to those in blood and CSF. Then, the comparison was repeated after excluding the significant invasive serotypes to determine how much these serotypes skewed the results in the first step. In the next step, each significant invasive serotype was independently compared to the remaining samples. Considering the population structure (Figure 4-9), the number and significance of variants in serotypes 1 and 5 would be expected to be very high compared to other strains.

To recap, the GWAS analysis design was defined as follows:

1. Comparison across isolation sites (location-based GWAS analysis):
    a. Nasopharyngeal isolates vs. invasive isolates.
    b. Nasopharyngeal isolates vs. invasive isolates (serotypes 1, 5, and 12F were excluded).
2. Comparison across serotypes (serotype-based GWAS analysis):
    a. Serotype 1 vs. others (serotypes 5 and 12F were excluded).
    b. Serotype 5 vs. others (serotypes 1 and 12F were excluded).
    c. Serotype 12F vs. others (serotypes 1 and 5 were excluded).

4.3.3.1 Identification of the significant SNPs

4.3.3.1.1 Nasopharyngeal vs. invasive

Comparison between the carriage and patient groups identified 1816 significant SNPs. The Manhattan plot illustrating the significance and chromosomal locations of SNPs is shown in Figure 4-10.

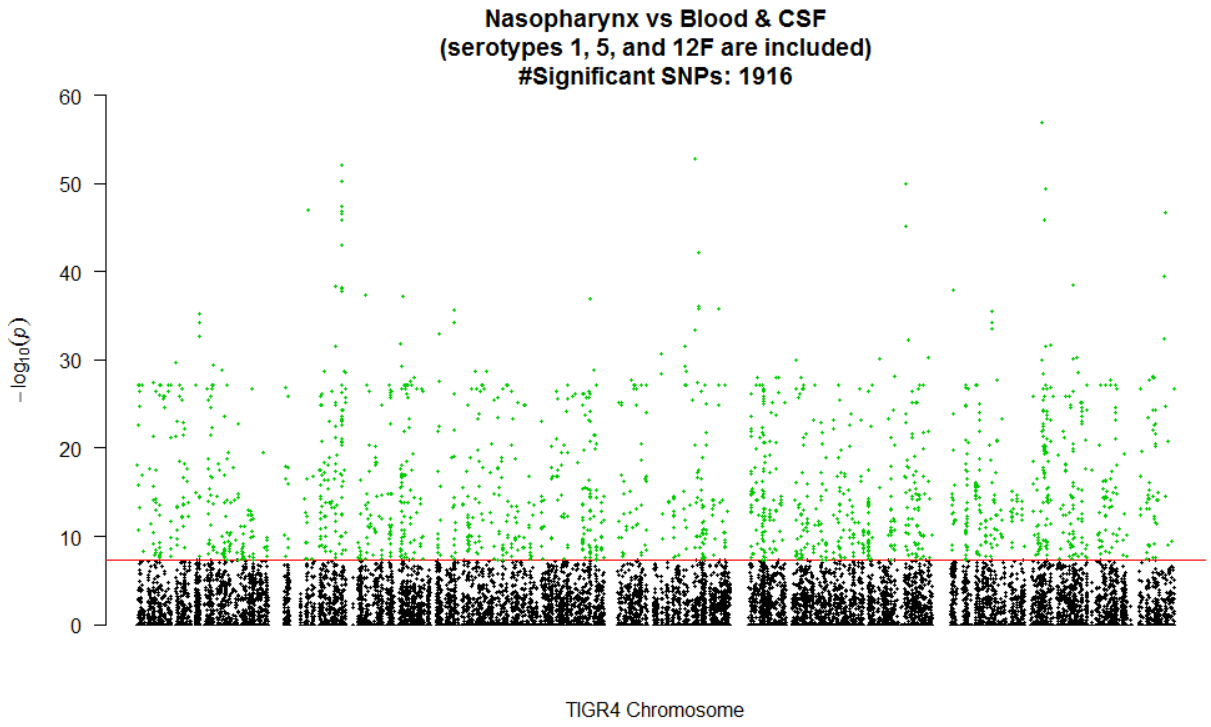


Figure 4-10. Manhattan plot of SNPs, nasopharyngeal vs. invasive isolates (serotypes 1, 5, and 12F were included). Significant SNPs are shown in green above the significance line indicated in red, representing the significance level of $5e^{-8}$.

Considering the genetic distinction of serotypes 1, 5, and 12F and their significant presence in the blood and CSF, many of the significant SNPs shown in Figure 4-10 most likely belonged to these serotypes. Therefore, the results were skewed and may not represent the real difference between the entire carriage and patient groups. To determine how much the significant invasive serotypes biased the analysis, the test between the carriage and patient groups was repeated after excluding serotypes 1, 5, and 12F. Interestingly, only one significant SNP was identified (Table 4-5). This number was much lower than the number of SNPs identified when the significant invasive serotypes were included in the test (1916 significant SNPs). As illustrated in Figure 4-10 and Figure 4-11, the significance level of SNPs also decreased after excluding serotypes 1, 5, and 12F. The only SNP was a missense substitution with a moderate effect on the protein function:

- A transition from (C -> T) in SP_0375 *(gnd),* converting Histidine to Tyrosine at position 301 of the protein. Gene SP_0375 (*gnd*) encodes 6-phosphogluconate dehydrogenase (6PGDH), an enzyme that catalyzes a chemical reaction in the center of the pentose phosphate pathway at the crossroad of many metabolic pathways.

This SNP was not identified in the significant invasive serotypes 1, 5, and 12F. However, serotypes 1 and 5 harbored another missense mutation at position 153 of the SP_0375 protein converting Alanine to Serine. This gene may carry different mutations in different strains.

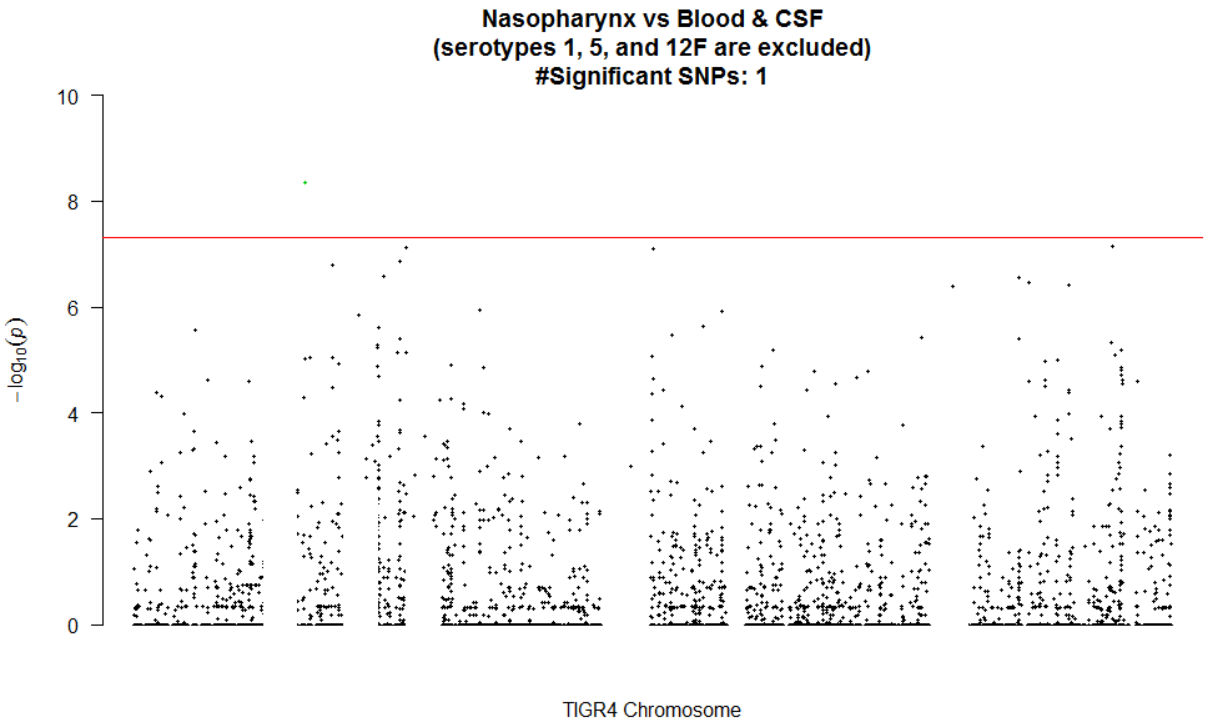


Figure 4-11. Manhattan plot of SNPs, nasopharyngeal vs. invasive isolates (serotypes 1, 5, and 12F were excluded). Significant SNPs are shown in green above the significance line indicated in red, representing the significance level of 5e-8.

Table 4-5. Significant SNP identified in invasive samples (excluding serotypes 1, 5, and 12F).

CHR:POS:REF:ALT	P-value	Annotation	Putative Effect	Gene ID	AA substitution
TIGR4-Chr:354623:C:T	4.354e-09	missense_variant	MODERATE	SP_0375	p.His301Tyr

The PPP and Glycolysis constitute the core of carbon metabolism in bacteria. PPP allows prokaryotes to use sugars such as ribose as the carbon source, which is essential for bacteria to modulate their metabolism during infection in the human body[234]. The missense SNP in SP_0375 could potentially affect the PPP since the gene is located in the center of the pathway (Figure 4-12). However, more experimental evidence is required to confirm the potential effect.

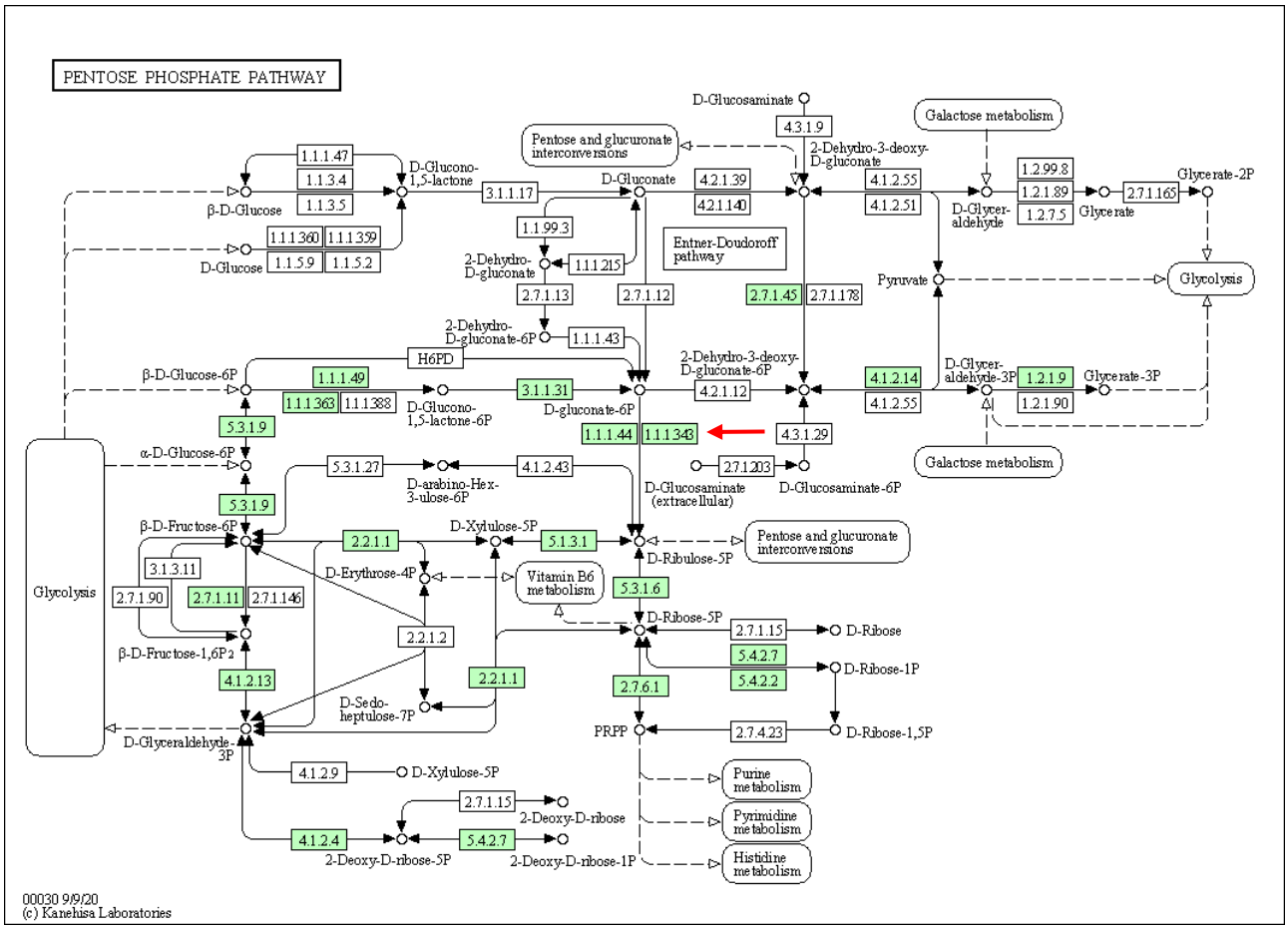


Figure 4-12. The pentose phosphate pathway (PPP).
The pneumococcal background genes are indicated in green. The enzyme commission ID for genes SP_0375 is 1.1.1343.

In summary, comparing whole carriage and patient groups identified putative virulence-related variants, but many of these were likely overpredicted when the abundant serotypes were included or underpredicted when these were excluded. A better strategy was to perform a serotype-based association analysis, taking out serotypes 1, 5, and 12F and comparing them one by one to the rest of the cohort. The serotype-based analysis identified a larger number of variants with a higher significance than the comparison between the carriage and patient groups. Due to the high invasiveness of serotypes 1, 5, and 12F, the hypothesis was that some of the significant variants in these serotypes were likely associated with their virulence.

4.3.3.1.2 Serotypes 12F vs. Others (serotypes 1 and 5 were excluded)

The number of significant SNPs in serotype 12F was 2011 (Figure 4-13 and Appendix 5). The significance of SNPs identified by the serotype-based GWAS analysis was much higher than those identified by the location-based GWAS analysis; compare the scale of the y-axis in Figure 4-11 and Figure 4-13. This reflected the genetic distinction of serotype 12F, one of the significantly present strains in the central nervous system. Of 2011 significant SNPs in serotype 12F, two SNPs were disruptive with a high impact on protein function, 474 SNPs were missense with a moderate effect, and the rest were either synonymous or located upstream or downstream of genes with a low predicted impact.

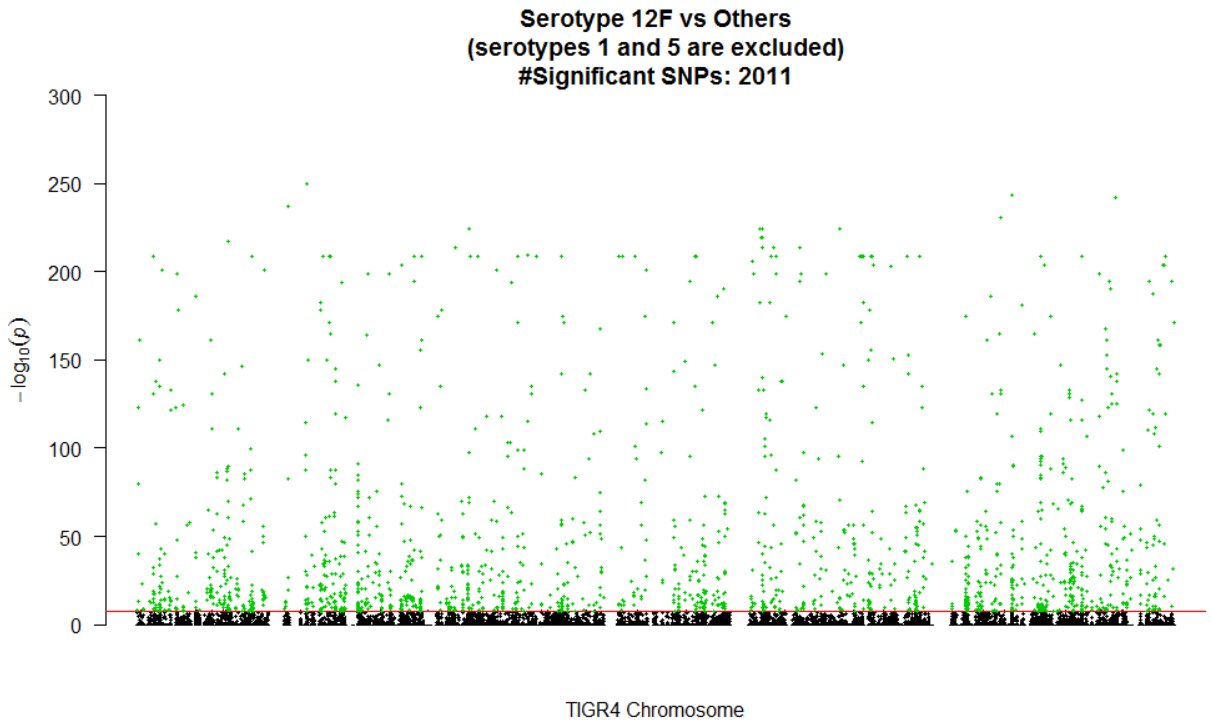


Figure 4-13. Manhattan plot of SNPs, serotype 12F vs. others (serotypes 1 and 5 are excluded).
Significant SNPs are shown in green above the significance line indicated in red, representing the significance level of $5e^{-8}$.

The two significant disruptive SNPs in serotype 12F are listed in Table 4-6. These SNPs caused the following genes to most likely become nonfunctional.

1. SP_0610: Encodes the ATP-binding protein of an amino acid ABC transporter.
2. SP_1245: Encodes a protein from the Cof family. Most of the proteins from this family are uncharacterized. However, some of them are involved in detoxification and amino acid biosynthesis[235].

The disruptive SNPs in serotype 12F both converted Glutamine to a stop codon.

Table 4-6. Significant SNPs in serotype 12F with predicted disruptive impact on the function of genes.

CHR:POS:REF:ALT	P-value	Annotation	Gene ID	AA substitution
TIGR4-Chr:576833:C:A	2.49E-209	Stop gained	SP_0610	p.Glu115*
TIGR4-Chr:1178118:C:A	5.85E-46	Stop gained	SP_1245	p.Glu28*

In addition to the disruptive SNPs that have deleterious effects, the significant missense SNPs could potentially affect the functions of genes. Although missense mutations are not disruptive, they could change protein effectiveness. The most significant missense SNPs in serotype 12F were:

1. The most significant missense SNP (p-value = 6.64e-251) was a transversion from (C -> G) in SP_0374 *(mapZ),* converting Proline to Alanine at position 44 of the protein. SP_0374 encodes the mid-cell anchored protein Z (MapZ). This protein forms ring structures at the cell equator and regulates cell division by marking the division site[236]. Cell division is crucial for pneumococcal biofilm formation and colonization. The mutation in MapZ could potentially

influence the cell division machinery and colonization rate in serotype 12F. However, experimental evidence is required to confirm the mutation effect.

2. A transition (G -> A) in SP_1908 (*ssbB*) converting Leucine to Phenylalanine at position 101 of the protein (p-value = 5e-244). In section 4.3.1, gene *ssbB* was identified as a significant polymorphic core gene harboring a significantly high number of mutations. This gene encodes a single-stranded DNA-binding protein that prevents re-annealing of the DNA strands during DNA replication and recombination. Its role is important during transformation and competence. A study found *ssbB* mutants to be less effective at homologous recombination[237]. Based on the low colonization rate of serotype 12F in Malawi, the recombination and competence strategy may differ for this serotype, and mutations in *ssbB* might be associated with this.

The 474 missense SNPs were in 257 genes, which means some genes harbored multiple missense SNPs in their sequences. Genes bearing the maximum number of significant missense SNPs in serotype 12F were among the most polymorphic core genes identified previously (see Table 4-4), including:

- SP_2066 had 10 missense SNPs in its structure; this gene was identified as one of the most polymorphic core genes in the pan-genome. SP_2066 or *thrC* encodes threonine synthase involved in threonine biosynthesis.
- SP_1380 with 8 missense SNPs in its sequences. This gene encodes an uncharacterized protein, and its function is unclear.

According to the functional enrichment analysis, genes carrying significant missense mutations in serotype 12F participate in a set of metabolic pathways listed in Table 4-7. The most considerable fold enrichment belonged to the tRNA metabolic process (Figure 4-14). Of note was that the significant pathways in Table 4-7 shared several genes. The “cellular aromatic compound metabolic process” pathway included genes from other pathways. This was because of the hierarchy of pathways; for instance, the tRNA metabolic process is a subtype of RNA metabolic process, a type of nucleic acid metabolic process.

Table 4-7. The functional enrichment analysis of genes harboring the significant missense SNPs in serotype 12F.

Enrichment FDR	**nGenes**	**Pathway Genes**	**Fold Enrichment**	**Pathways**
5.7E-02	13	41	2.6	tRNA metabolic process
7.8E-02	14	48	2.4	ncRNA metabolic process
9.0E-02	19	77	2	RNA metabolic process
2.8E-02	28	118	2	nucleic acid metabolic process
2.8E-02	40	188	1.8	nucleobase-containing compound metabolic process
2.8E-02	43	209	1.7	heterocycle metabolic process
2.8E-02	43	212	1.7	cellular aromatic compound metabolic process

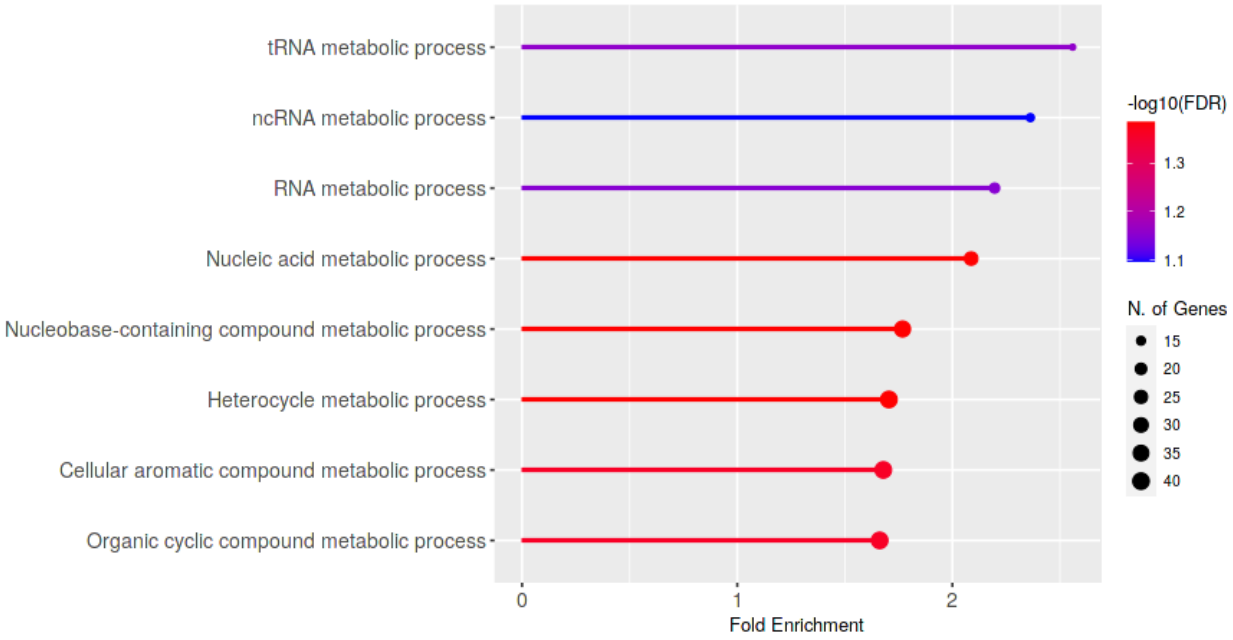


Figure 4-14. The significant pathways involving genes bearing the significant missense SNPs in Serotype 12F.

4.3.3.1.3 Serotypes 5 vs. Others (serotypes 1 and 12F were excluded)

The number of significant SNPs in serotype 5 was 4633 (Figure 4-15 and Appendix 6). Five SNPs were disruptive and had a high impact on protein function. 967 SNPs were missense and had a moderate impact. The rest of the SNPs were either synonymous or located upstream and downstream of genes with a low predicted impact. The number and significance of SNPs were double in serotype 5 compared to serotype 12F, reflecting the higher abundance and genetic distinction of serotype 5.

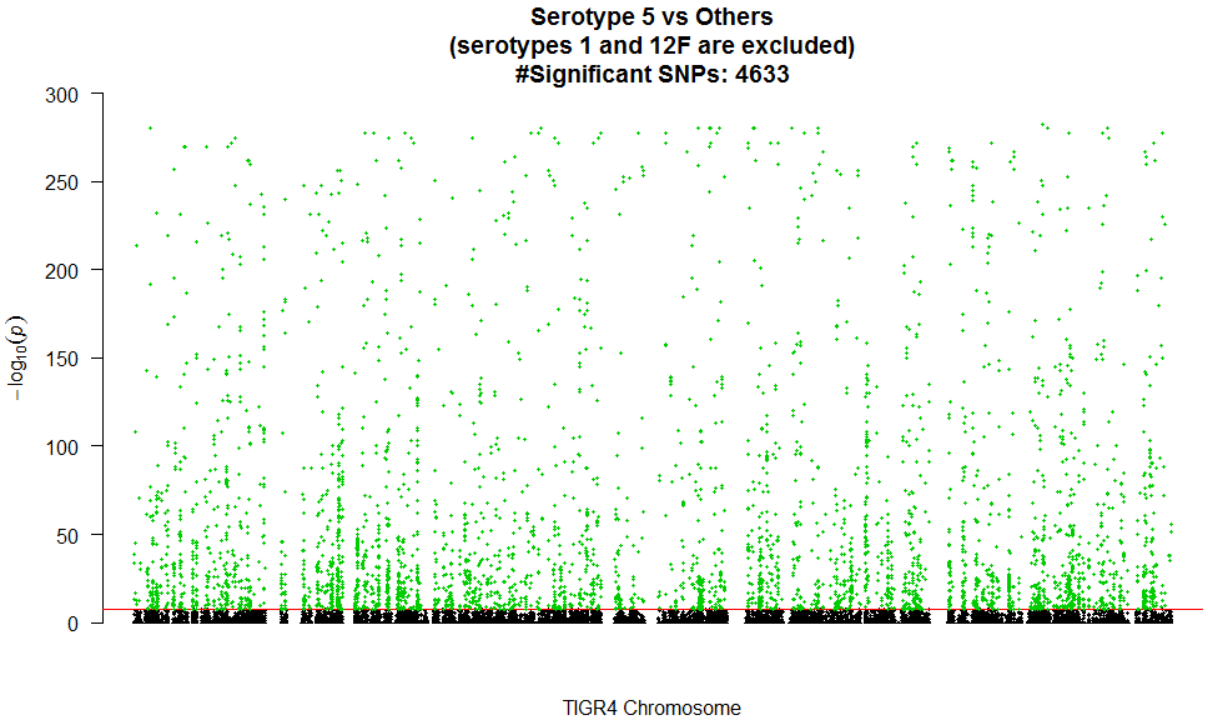


Figure 4-15. Manhattan plot of SNPs, serotype 5 vs. others (serotypes 1 and 12F are excluded).
Significant SNPs are shown in green above the significance line indicated in red, representing the significance level of 5e-8.

Despite the similarities between the mutation profiles in serotypes 5 and 12F, they were not completely identical. There were five SNPs with a disruptive impact on the function of genes in serotype 5, whereas there were two in serotype 12F. The five disrupting SNPs in serotype 5 are listed in Table 4-8.

Table 4-8. Significant SNPs in serotype 5 with disruptive impact on the function of genes.

CHR:POS:REF:ALT	P-value	Annotation	Gene ID	AA substitution
TIGR4-Chr:863725:G:A	1.32E-256	Stop gained	SP_0910	p.Trp88*
TIGR4-Chr:219907:C:T	1.68E-203	Stop gained	SP_0250	p.Gln439*
TIGR4-Chr:754282:G:A	1.59E-61	Start lost	SP_0801	p.Met1?
TIGR4-Chr:942444:C:T	1.48E-47	Stop gained	SP_1001	p.Gln214*
TIGR4-Chr:781877:G:A	3.27E-14	Stop gained	SP_0833	p.Trp57*

Most of the disrupted genes in serotype 5 were uncharacterized:

- SP_0910: Uncharacterized protein.
- SP_0250: A subunit in a PTS transporter.
- SP_0801: Uncharacterized protein.
- SP_1001: Amino acid permeases protein, an integral membrane protein involved in transporting amino acids into the cell.
- SP_0833: Uncharacterized protein.

The most significant missense SNPs in serotype 12F were in genes SP_0374 and SP_1908, but these genes harbored synonymous SNPs in serotype 5. Instead, the most significant missense SNPs in serotype 5 were:

- A transversion from C to A in gene SP_1988 converting Arginine to Serine at position 272 of the protein (p-value = 1.946E-283). This gene encodes a transmembrane putative immunity protein, a type of bacteriocin secreted by pneumococci.
- A transversion from G to T in gene SP_1266 (*dprA*) converting Leucine to Methionine at position 146 of the protein (p-value = 9.786E-281). SP_1266 encodes DNA processing protein A (DprA), a member of the recombination-mediator protein family, involved in bacterial transformation.

The 967 significant missense SNPs in serotype 5 existed in 370 genes, implying that some core genes mutated in multiple sites in serotype 5. As observed in serotype 12F, the polymorphic core genes SP_1380 and SP_2066 harbored 22 and 16 significant missense SNPs in serotype 5, respectively. These were the maximum numbers of missense SNPs existing in a single gene in serotype 5.

The functional enrichment analysis of the mutated genes was required to gain insight into the biological processes these genes perform in the cell. The enrichment analysis identified the biological processes listed in Table 4-9. Despite the differences between the SNP profiles of serotypes 5 and 12F, the functional enrichment analysis of mutated genes identified the same enriched pathways in both strains, which means the types of genes affected by SNPs in both serotypes were similar. However, the number of genes contributing to the enriched pathways was slightly higher in serotype 5 (Figure 4-16).

Table 4-9. The functional enrichment analysis of genes harboring the significant missense SNPs in serotype 5.

Enrichment FDR	**nGenes**	**Pathway Genes**	**Fold Enrichment**	**Pathways**
1.1E-03	13	21	3.6	tRNA aminoacylation for protein translation
1.1E-03	13	21	3.6	amino acid activation
1.1E-03	13	21	3.6	tRNA aminoacylation
1.7E-03	19	41	2.7	tRNA metabolic process
1.7E-03	21	48	2.5	ncRNA metabolic process
6.0E-03	37	118	1.8	nucleic acid metabolic process
8.9E-03	46	161	1.6	small molecule metabolic process
6.0E-03	53	188	1.6	nucleobase-containing compound metabolic process
5.4E-03	58	209	1.6	heterocycle metabolic process
6.0E-03	58	212	1.6	cellular aromatic compound metabolic process
6.8E-03	58	214	1.6	organic cyclic compound metabolic process

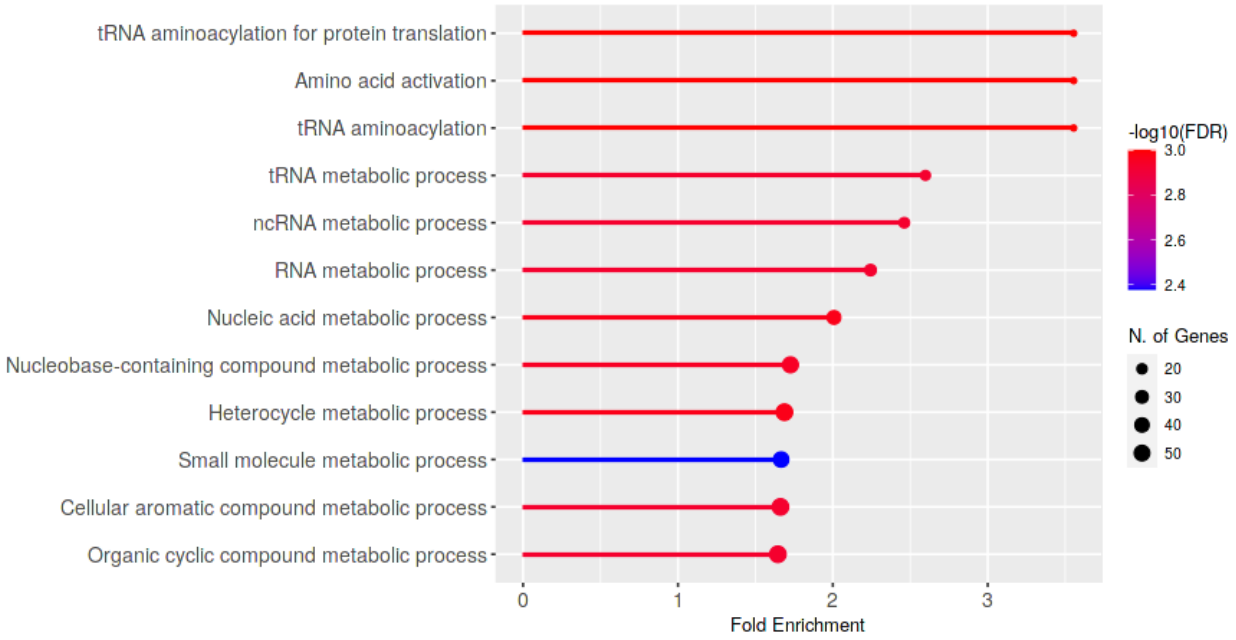


Figure 4-16. The significant pathways involving genes bearing the significant missense SNPs in Serotype 5.

4.3.3.1.4 Serotypes 1 vs. Others (serotypes 5 and 12F were excluded)

The number of significant SNPs identified in serotype 1 was 4891, the maximum compared to serotypes 5 and 12F (Figure 4-17 and Appendix 7), which is likely to be associated with the highest prevalence and genetic distinction of serotype 1. Of 4891 significant SNPs, five were disruptive and had a high impact on protein function, and 1004 SNPs were missense and had a moderate effect. The rest were either synonymous or located upstream and downstream of genes with a low impact.

In chapter 2, serotype 1 was identified as the most abundant and the most significant invasive strain (p-value 1.96E-34) in Malawi (Table 2-2). Its high genetic divergence from other serotypes may be associated with its properties, such as its shortest colonization period, quickness to infect the sterile site, and persistent dominance in Malawi until 2015. In total, 50 invasive serotypes were detected in the blood and CSF. However, most of these strains (47/50) had a low frequency among patients (< 5%). The abundant serotypes in the patient group (frequency > 5%) were serotypes 1 (frequency = 15.65%), 5 (frequency = 11.25%), and 23F (frequency = 5.43%). Serotype 23F had a similar frequency in the carriage group (5.21%), while the frequencies of serotypes 1 and 5 were significantly lower in the nasopharynx (serotype 1 = 0.97% and serotypes 5 = 2.18%). Despite the highest frequency of serotype 1 in the patient group, it had the lowest frequency among carriers, implying its highest invasiveness. Moreover, the main difference between serotypes 1 and 5 was the significant increase in the prevalence of serotype 1 after vaccination, whereas the prevalence of serotype 5 significantly decreased in the post-PCV13 era. Serotype 1 was the only abundant and invasive strain whose prevalence increased after vaccination.

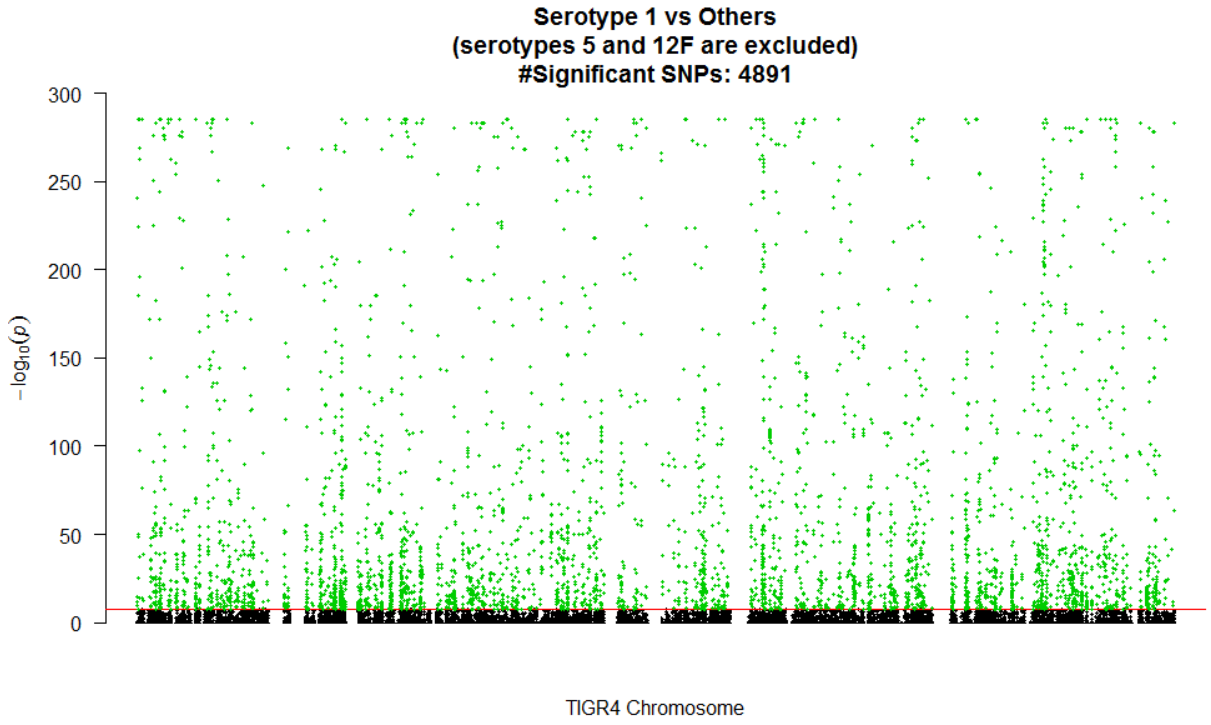


Figure 4-17. Manhattan plot of SNPs, serotype 1 vs. others (serotypes 5 and 12F are excluded).
Significant SNPs are shown in green above the significance line indicated in red, representing the significance level of 5e-8.

The five disruptive SNPs in serotype 1 with a high impact on the function of genes are listed in Table 4-10. Four genes were affected by these SNPs in serotype 1:

- SP_0833: Encoding an uncharacterized protein harboring two significant disruptive SNPs. This gene is most likely inactive in serotype 1.
- SP_0830: Encoding an uncharacterized protein.
- SP_1245: Encoding a protein from the Cof family. Most of the proteins from this family are uncharacterized. However, some of them are involved in detoxification and amino acid biosynthesis[235]. The disruptive mutation in SP_1245 was also observed in serotype 12F.
- SP_0801: Encoding an uncharacterized protein. The disruptive mutation in SP_0801 was also observed in serotype 5.

The majority of genes that bear disruptive SNPs in serotype 1 were uncharacterized.

Table 4-10. Significant SNPs in serotype 1 with disruptive impact on the function of genes.
There are two disruptive SNPs in genes SP_0833.

CHR:POS:REF:ALT	P-value	Annotation	Gene ID	AA substitution
TIGR4-Chr:782211:C:T	2.21E-269	Stop gained	SP_0833	p.Gln169*
TIGR4-Chr:780048:G:A	9.60E-184	Stop gained	SP_0830	p.Trp137*
TIGR4-Chr:1178118:C:A	4.73E-122	Stop gained	SP_1245	p.Glu28*
TIGR4-Chr:754282:G:A	3.79E-67	Start lost	SP_0801	p.Met1?
TIGR4-Chr:781877:G:A	1.27E-14	Stop gained	SP_0833	p.Trp57*

The most significant missense SNPs in serotype 1 were:

- A transversion from C to A and a transition from C to T in gene SP_0045 converting Histidine to Asparagine at position 590 and Proline to Serine at position 706 of the protein (p-value = 2.92E-286). This gene encodes a phosphoribosylformylglycinamidine synthase involved in the purine biosynthetic process.
- A transition from C to T in gene SP_0046 (*purF*), converting Alanine to Serine at position 69 of the protein (p-value = 2.92E-286). SP_0046 is also involved in purine biosynthesis.

The 1004 missense SNPs in serotype 1 were in 375 genes, implying that some core genes mutated in multiple sites. As observed in serotypes 5 and 12F, the polymorphic core genes SP_1380 and SP_2066 harbored 28 and 15 significant missense SNPs, which were the maximum numbers of missense SNPs in a single gene in serotype 1.

The functional analysis of genes bearing missense SNPs in serotype 1 identified a set of pathways in Table 4-11 and shown in Figure 4-18. Although most of the enriched pathways in serotypes 1, 5, and 12F were similar, the carboxylic acid metabolic process was detected only in serotype 1. This may be because serotype 1 uses carboxylic acid derivatives in its capsular polysaccharide structure[238]. There are three acidic carboxyls in the sugar units of serotype 1, while in the capsule structure of serotype 5 and 12F, the carboxyl groups do not exist[239].

Table 4-11. The functional enrichment analysis of genes harboring the significant missense SNPs in serotype 1.

Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathways
1.5E-04	14	21	3.8	tRNA aminoacylation for protein translation
1.5E-04	14	21	3.8	amino acid activation
1.5E-04	14	21	3.8	tRNA aminoacylation
5.2E-04	20	41	2.8	tRNA metabolic process
5.7E-04	31	82	2.1	carboxylic acid metabolic process
5.7E-04	31	82	2.1	oxoacid metabolic process
8.9E-04	31	85	2.1	organic acid metabolic process
9.3E-04	39	118	1.9	nucleic acid metabolic process
6.2E-04	50	161	1.8	small molecule metabolic process
6.5E-04	56	188	1.7	nucleobase-containing compound metabolic process
5.2E-04	62	209	1.7	heterocycle metabolic process
5.7E-04	62	212	1.7	cellular aromatic compound metabolic process
6.2E-04	62	214	1.6	organic cyclic compound metabolic process
9.0E-04	98	389	1.4	cellular process

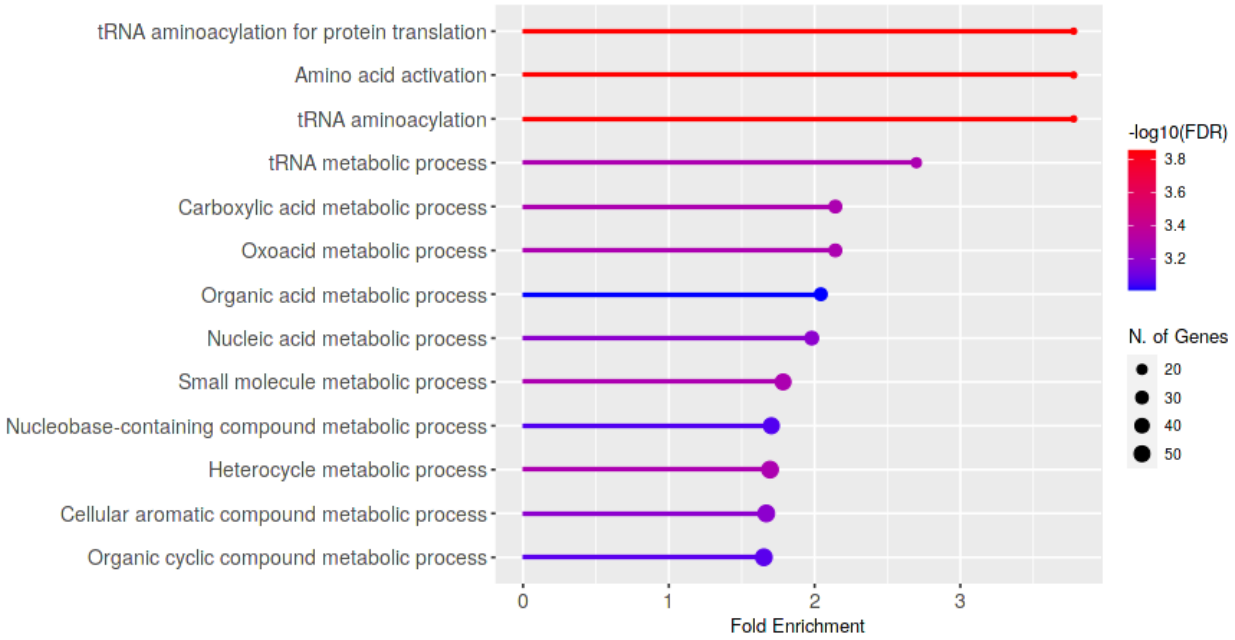


Figure 4-18. The significant pathways involving genes bearing the significant missense SNPs in Serotype 1.

The proportion of nonsynonymous SNPs commonly present in the significant invasive serotypes is of interest. Despite the difference between the SNP profiles in serotypes 1, 5, and 12F, some characteristics of mutated core genes in these serotypes were analogous. The overlaps between significant nonsynonymous SNPs in serotypes 1, 5, and 12F are shown in Figure 4-19.

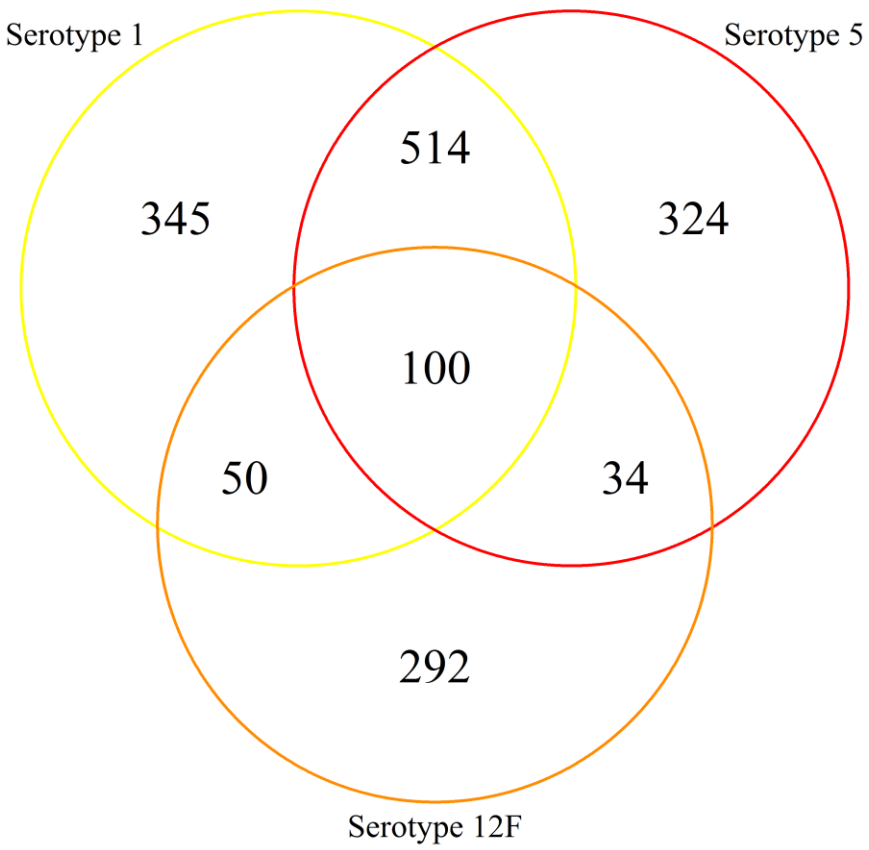


Figure 4-19. The Venn diagram showing the overlap between significant nonsynonymous SNPs in serotypes 1, 5, and 12F.

The 100 common SNPs between the significant invasive serotypes could represent the core SNPs characteristic of invasiveness. The SNPs with the highest significance in all three significant invasive serotypes existed in the following genes:

- SP_1697 (*recG*): A DNA helicase that has a critical role in DNA recombination and repair.
- SP_2058 (*tgt*): This gene encodes queuine tRNA ribosyltransferase and contributes to tRNA modification.
- SP_2045: A site-specific DNA methyltransferase that transfers a methyl group to either N-6 of adenine or C-5 or N-4 of cytosine. DNA methylation is a control mechanism for many biological processes such as gene expression.
- SP_0668 (*gki*): A housekeeping gene that phosphorylates glucose. This gene is involved in the glycolytic process and carbohydrate metabolism.
- SP_0588 (*pnp*): Polyribonucleotide nucleotidyltransferase with RNA binding function involved in mRNA degradation.
- SP_1207 (*xseA*): Exodeoxyribonuclease with DNA binding function involved in the breakdown of DNA.

As stated above, the most significant nonsynonymous SNPs in serotypes 1, 5, and 12F were in the genes contributing to DNA and RNA metabolism. The functional enrichment analysis of all mutated genes in the significant invasive serotypes identified the tRNA metabolic process as an enriched biological

process (Figure 4-20). The tRNA metabolic pathway includes many genes from the aminoacyl-tRNA synthetase (AaRS) family. These genes were mutated in significant invasive serotypes. AaRSs are universal enzymes found in all three kingdoms of life. They catalyze the attachment of an amino acid to its cognate tRNA. The role of AaRSs is not limited to tRNA processing; they also contribute to several biological processes, such as RNA splicing and transcriptional regulation. Due to the considerable divergence between eukaryotic and prokaryotic AaRSs, they have been recognized as suitable targets for antimicrobial drug development[240]. Inhibition of prokaryotic AaRSs can selectively halt protein biosynthesis resulting in attenuation of bacterial growth. An example of an AsRS-based antibiotic is *Mupirocin*, licensed in 1987. Despite its effectiveness against some species resistant to other types of antibiotics, such as *Methicillin-resistant Staphylococcus aureus*, mupirocin resistance phenotypes have been identified in 1993[241]. The high number of mutated AaRSs in serotypes 1, 5, and 12F may be associated with their mechanism to gain resistance against antibiotics. However, this hypothesis needs to be tested through further experiments.

Of importance was that some of AaRSs such as *serS*, *proS*, and *thrS* that catalyze the attachment of serine, proline, and threonine to the tRNA molecules were fully conserved in the pan-genome (see Figure 4-5). On the other hand, AsRSs such as *hisS*, *lysS*, *tyrS*, *leuS* that catalyze the attachment of histidine, lysine, tyrosine, and leucine harbored significant SNPs in serotypes 1, 5, and 12F.

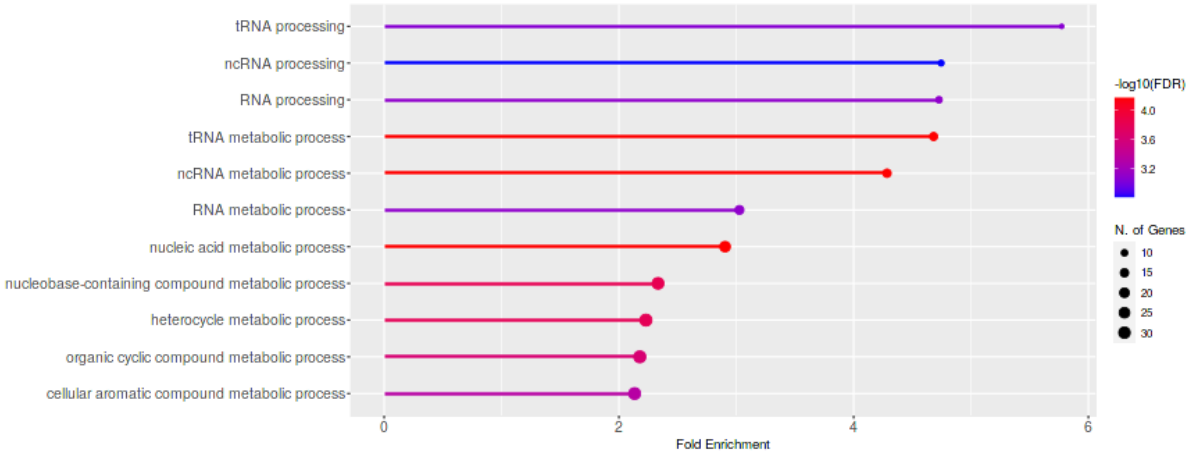


Figure 4-20. The enrichment analysis of genes bearing the common significant missense SNPs in serotypes 1, 5, and 12F.

61.34% of the significant nonsynonymous SNPs in serotype 12F (292/476) were unique and were not detected in serotypes 1 and 5. This number was 31.12% (345/1009) and 34.95% (324/972) in serotypes 1 and 5, respectively (Figure 4-19). However, the unique SNPs may not reflect the actual differences between the mutation profiles of the significant invasive serotypes since many of these SNPs from these serotypes were located on the same genes but in different positions. Instead of unique SNPs, it was more helpful to identify genes that harbored missense variants in only one of the significant invasive serotypes and not in the other two. This could provide information about the type of genes affected by mutations in each serotype. The number of genes bearing significant unique missense variants in serotypes 1, 5, and 12F was 30, 33, and 27, respectively.

The functional enrichment analysis of these genes did not identify any pathways. Genes exclusively bearing the most significant nonsynonymous SNPs in serotypes 1, 5, and 12F were:

- In serotype 1:
    - SP_0830: Uncharacterized protein. This gene harbored a disruptive SNP in serotype 1, whereas it did not have any SNP in serotypes 5 and 12F.
    - SP_0158: Bears a missense SNP in serotype 1, but did not contain any SNPs in serotypes 5 and 12F. It encodes a putative NrdI-like protein involved in the cellular protein modification process.
- In serotype 5:
    - SP_0910: Uncharacterized protein. This gene harbored a disruptive SNP in serotype 5 but did not bear any significant SNPs in serotypes 1 and 12F.
    - SP_1451: A phosphatase protein from Cof family. The biological process that this protein is involved in is not well understood. Some of the proteins from Cof family are involved in detoxification and protein biosynthesis.
    - SP_2041 (*yidC2*): A membrane protein insertase required for the insertion and folding of proteins into the membrane.
- In serotype 12F:
    - SP_0374 (*mapZ*): This gene contained nonsynonymous SNPs only in serotype 12F. In serotypes 1 and 5, it only harbored synonymous SNPs. SP_0374 encodes a cell division protein that marks the future cell division site.
    - SP_1908 (*ssbB*): In section 4.3.1, SP_1908 was identified as a significant polymorphic core gene harboring a significantly high number of mutations. However, this gene carried nonsynonymous SNPs only in serotype 12F but harbored several synonymous SNPs in serotypes 1 and 5. SP_1908 encodes the single-stranded DNA-binding protein that prevents re-annealing of the DNA strands during DNA replication and recombination. This protein plays an essential role during transformation and competence[237].

The similarities and differences between the SNP profiles of serotypes 1, 5, and 12F could reflect that although the genome structures of the significant invasive serotypes are similar from some perspectives, they still differ in other respects.

4.3.3.2 Identification of significant indels

Indels refer to insertions and deletions in the genome structures. They can be found either in non-coding regions located upstream or downstream of genes or inside the coding sequences. Indels with a high impact are insertions and deletions inside coding sequences of genes when the number of inserted or deleted base pairs is not a multiple of three, the indel would be a frameshift variant that disrupts the reading frames in the gene structure. The function of frameshift variants is similar to disruptive SNPs since they render active genes unfunctional.

4.3.3.2.1 Nasopharyngeal vs. invasive isolates

After excluding the significant invasive serotypes from the association test, no significant indels were identified between the carriage and patient groups. The distribution of indels in the population was similar to that of SNPs. There is no significant difference between nasopharyngeal and invasive samples if serotypes 1, 5, and 12F are ignored. However, the serotype-based GWAS analysis identified significant indels in the significant invasive serotypes.

4.3.3.2.2 Serotypes 12F vs. Others (serotypes 1 and 5 were excluded)

There were three significant frameshift indels with a high impact in serotype 12F. These indels were found in three genes listed in Table 4-12.

Table 4-12. Significant indels in serotype 12F with a disruptive impact on the function of genes.

CHR:POS:REF:ALT	P-value	Annotation	Gene ID	AA substitution
TIGR4-Chr:1898322:GA:G	4.19E-63	Frameshift variant	SP_1995	p.Ser91fs
TIGR4-Chr:1374280:TA:T	8.30E-42	Frameshift variant	SP_1457	p.Asn240fs
TIGR4-Chr:727113:A:ATGAAG	8.40E-39	Frameshift variant	SP_0768	p.Glu147fs

The function of genes that harbored significant indels are as follows:

- SP_1995: Uncharacterized protein.
- SP_1457: 23S rRNA (guanosine2251-2'-O)-methyltransferase that belongs to RNA methyltransferase TrmH family.
- SP_0768 (*rlmN*): 23S rRNA (adenine2503-C2)-methyltransferase This gene methylates both 23S rRNA and tRNAs.

SP_1457 and SP_0378 methylate nucleotides in non-coding RNAs such as rRNA and tRNA[242]. They are involved in controlling the gene translation process. Deactivating these genes in serotype 12F may alter gene expression in this serotype.

4.3.3.2.3 Serotypes 5 vs. Others (serotypes 1 and 12F were excluded)

There were three significant frameshift indels with a high impact in serotype 5. These indels were found in three genes listed in Table 4-13.

Table 4-13. Significant indels in serotype 5 with a disruptive impact on the function of genes.

CHR:POS:REF:ALT	P-value	Annotation	Gene ID	AA substitution
TIGR4-Chr:1501699:G:GT	3.2E-285	Frameshift variant	SP_1598	p.Gln250fs
TIGR4-Chr:780040:GATTTTCTGGGGGAA:G	2E-190	Frameshift variant	SP_0830	p.Trp137fs
TIGR4-Chr:1485134:T:TC	9.3E-148	Frameshift variant	SP_1580	p.Ser86fs

The function of genes that harbored significant indels are as follows:

- SP_1598: Putative phosphomethylpyrimidine kinase likely involved in thiamine (vitamin B1) biosynthetic process. Vitamin B1 is an essential cofactor for several key enzymes in carbohydrate metabolism[243].
- SP_0830: Uncharacterized protein.
- SP_1580 (*msmK*): An ATP-binding protein that belongs to an ABC transporter.

Genes disrupted by frameshift indels in serotype 5 contribute to carbohydrate metabolism.

4.3.3.2.4 Serotypes 1 vs. Others (serotypes 5 and 12F were excluded)

There were three significant frameshift indels with a high impact in serotype 1. These indels were found in three genes listed in Table 4-14.

Table 4-14. Significant indels in serotype 1 with a disruptive impact on the function of genes.

CHR:POS:REF:ALT	P-value	Annotation	Gene ID	AA substitution
TIGR4-Chr:1890924:CA:C	4.7E-288	Frameshift variant	SP_1988	p.Phe558fs
TIGR4-Chr:350142:A:ACCCC	4.1E-135	Frameshift variant	SP_0371	p.Lys158fs
TIGR4-Chr:383089:GT:G	5.91E-54	Frameshift variant	SP_0403	p.Lys289fs

- SP_1988: An integral membrane protein that spans the membrane seven times. Although this protein is suggested to be an immunity protein, its exact role is unclear.
- SP_0371: Uncharacterized protein.
- SP_0403 (*rnhC*): This gene encodes an endonuclease that specifically degrades the mRNA of RNA-DNA hybrids. This protein exists in all three kingdoms of living organisms.

As observed in serotype 12F, a gene involved in nucleic acid metabolism (SP_0403) bears a significant indel in serotype 1. Serotypes 1 and 12F are dominant in the central nervous system and may need to adapt their biology to the CSF environment. The indel in SP_0403 may affect gene expression and translation since this gene is involved in mRNA degradation, however, experimental work is needed to confirm this assumption.

The significant frameshift indel in SP_1988 in serotype 1 may be associated with the conditions in the sterile sites. As described in section 4.3.3.1.3, the most significant missense SNP in serotype 5, was also detected in SP_1988. Serotype 5 is dominant in the blood, and the pathogen may modify the structure of the immunity protein SP_1988 in different environments.

4.4 Discussion

The main objectives of the analysis in this chapter were:

- To determine the most conserved part of the pan-genome.
- To detect the most genetically distinct strains using the PCA of SNP distribution.
- To perform a GWAS analysis and identify SNPs and indels differentiating invasiveness from the carriage.

14.4% of the core genome, including 105 genes, were highly conserved with a significantly low number of polymorphic sites. On the other hand, 15.64% of the core genome, including 114 genes, were significantly polymorphic. This means only 1.5% of the pan-genome (105 out of 6803 pan genes) was fully conserved both in their presence and sequences. This result again emphasized the high plasticity in the *pneumococcus* pan-genome since more than 98% of genes represented at least one kind of diversity in the pan-genome, including:

1. Being a polymorphic core gene present in all samples but bearing small-scale mutations (SNPs or indels).
2. Being an accessory gene present in some samples (not all).

What is the importance of identifying the most conserved core genes?

Firstly, the conserved core genes can be considered as putative targets for future drugs. The Malawian isolates were collected over a long time, from 1997 to 2015. As a highly recombinogenic bacterium, *pneumococcus* has a pan-genome with a dynamic structure. Nonetheless, the conserved core genes have been retained in all samples and did not mutate for 18 years. These genes' persistent presence and conservation in the pan-genome without any mutation implies their essentiality for cell survival. Therefore, the conserved core genes could provide a list of potential targets for drug design (Table 4-2). Theoretically, any hypothetical drug that interferes with the function of these genes could be a candidate for further research. The RNA polymerase subunits (encoded by *rpoA* and *rpoB*) and the ribosomal protein S1 (encoded *rpsA*) were highly conserved in the pan-genome and might be ideal targets for novel medications. For drug design, it is crucial to know which conserved core genes work together and perform a fundamental biological function in the cell. The pathway enrichment analysis identified conserved core genes that encode three families of the elongation factors (EF-Tu, EF-G, and EF-4). These proteins bring the aminoacyl-tRNA to the ribosome and regulate the movement of the messenger RNA through the protein synthesis machinery. Some antibiotics have already been designed to target EFs, such as *Kirromycin* and *Fusidic acid*, which inhibit EF-Tu and EF-G, respectively. This study found some EFs and aminoacyl-tRNA synthetases conserved in the Malawian pan-genome for a long time. Any strategy that interferes with the function of these proteins can potentially deactivate the entire gene translation process in the cell. It is worth noting that some of the conserved core genes would also likely be well conserved in other bacteria in the human microbiome, which may not be advantageous to target. Thus, the medication should have the specificity to target the conserved pneumococcal genes selectively.

Secondly, some of the conserved core genes that encode membrane proteins are worth considering for vaccine development:

- SP_1241: A subunit of an amino acid ABC transporter.
- SP_1889 (*amiD*): A subunit of an oligopeptide ABC transporter.
- SP_1890 (*amiC*): A subunit of an oligopeptide ABC transporter.
- SP_1891 (*amiA*): A subunit of an oligopeptide ABC transporter.
- SP_0230 (*secY*): The central subunit of the protein translocation channel SecYEG. SecY is an integral membrane protein that interacts with the signal sequences of secretory proteins as well as with SecE and SecG components.

More experimental work is required to determine whether the human immune system recognizes these proteins and secretes any antibodies to tag them.

The number of known protein-protein interactions was greater between conserved core genes than the polymorphic ones. This was because a higher percentage of the polymorphic core genes remained uncharacterized (21%) compared to the conserved core genes (0.9%). This result was similar to what was found from the comparison between core- and accessory- genomes in chapter 3, as a higher frequency of accessory genes were uncharacterized compared with the core genes. This means that knowledge about the variable part of the pan-genome is not as complete as the conserved part. The functional enrichment analysis of the polymorphic core-genome did not detect any enriched biological processes or pathways. However, based on the manual investigation, many of the polymorphic core genes were found involved in various metabolic pathways, such as glycolysis-gluconeogenesis, citrate

cycle (TCA cycle), inositol phosphate metabolism, fructose metabolism, pentose phosphate pathway, ascorbate metabolism, purine metabolism, pyruvate metabolism, and amino acid biosynthesis.

As described in chapter 2, 56 serotypes were identified in the dataset. Serotypes 1, 5, and 12F were only frequent in the patient group, so they were labeled as significant invasive serotypes. Serotypes 16F and 19F were only frequent in the carriage group, and the other 51 serotypes were common between carriers and patients (common serotypes). The PCA of SNPs in the core-genome determined the population structure that provided a guideline for the GWAS analysis in this chapter and the gene presence-absence analysis in the next chapter. The SNP-based PCA suggested that the samples did not cluster genetically according to their isolation sites (nasopharynx, blood, and CSF). Instead, the genetically distinct clusters in the population were serotypes 1 and 5. Therefore, any GWAS analysis comparing serotypes 1 and 5 to other strains would identify the maximum number of significant SNPs and indels. In chapter 5, a gene-based PCA was applied based on the distribution of genes in the accessory-genome. Theoretically, the patterns of genetic separation in the core- and accessory-genome are similar, therefore, as observed for SNPs and indels (the small-scale variants in the core-genome), the maximum number of significant genes (the large-scale variants in the accessory-genome) would belong to serotypes 1 and 5.

It was especially crucial to analyze the core and accessory-genome of serotype 1 because of its characteristics:

- Serotype 1 was the most abundant strain in the entire cohort.
- Serotype 1 had the lowest frequency in the carriage group and the highest frequency in the patient group.
- Serotype 1 showed the most significant distinction in the core-genome with the maximum number of significant SNPs and would likely also show this in the accessory-genome with the highest number of significant genes.

The significant SNPs and indels were identified in this chapter through the GWAS analysis, and the significant genes were further explored in chapter 5. The GWAS analysis aimed to identify any putative virulence-associated SNPs and indels. To achieve this goal, the population structure and the distribution of serotypes must be considered simultaneously. Any mutation in the invasive samples obtained from blood and CSF that does not exist in the nasopharyngeal population could be a candidate virulence-associated variant. However, the genetic distinction of serotypes 1 and 5 and their significant presence in the sterile sites suggest that the majority of significant variants identified by a GWAS analysis between the nasopharynx and sterile sites would most likely belong to serotypes 1 and 5. To consider the population structure and the serotype distribution, samples from blood and CSF were divided into the significant invasive serotypes (1, 5, and 12F) and the common serotypes (51 other serotypes in the blood and CSF). Serotype 12F was put in the same group as serotypes 1 and 5 due to its significant presence in the sterile sites. For the first GWAS analysis, the significant invasive serotypes were excluded, and the carriage and invasive samples were compared (location-based GWAS). In the second step, serotypes 1, 5, and 12F were separately compared to the rest of the samples (serotype-based GWAS). The location-based GWAS analysis identified the mutations in the common serotypes that are able to enter the blood and CSF. The serotype-based GWAS analysis identified mutations in the significant invasive serotypes that could be linked to their significant presence in the sterile sites and invasiveness. The effect of SNPs and indels on the functions of genes were predicted using variant annotations. Nonsense and missense

SNPs and frameshift indels are likely to affect protein function, while synonymous SNPs and indels in the non-coding regions may affect gene expression. Since the nonsense SNPs and frameshift Indels truncate proteins in the middle, they most likely deactivate them. It is worth noting that all significant SNPs and indels are not necessarily associated with virulence, they could be just linked to other factors such as the serologic (serotype) properties of isolates; however, a list of significant SNPs and indels should still contain a set of potential virulence-associated variants.

The location-based GWAS analysis identified only one significant missense SNP and no Indel in the blood and CSF compared to the nasopharynx. This indicates that if the significant invasive serotypes (1, 5, and 12F) are excluded, the rest of the samples from blood and CSF (common serotypes) will not differ too much from the nasopharyngeal isolates. The only difference is a missense substitution in gene SP_0375 (*gnd*) which encodes an enzyme in the center of the pentose phosphate pathway at the crossroad of many metabolic reactions. Therefore, the effect of SNP in *gnd* could influence several biological processes involved in pneumococcal metabolism. Further experiments are required to demonstrate the exact impact of the SNP in SP_0375.

One may ask why the pneumococcal population in the blood and CSF does not differ much from the nasopharyngeal population when the significant invasive serotypes are excluded. A couple of reasons could explain this:

1. Because invasive pneumococcal strains need to colonize the human nasopharynx for a while before entering the blood and CSF, the nasopharyngeal population should be a mixture of non-invasive and invasive strains. Therefore, the nasopharyngeal population could not be considered as a pure non-invasive collection. In the Malawian cohort, several serotypes isolated from the nasopharynx were also detected in the blood and CSF (common serotypes). The presence of invasive serotypes in the nasopharynx can result in more similarities between the genomes of the carriage and patient groups so that only one significant SNP was Identified.
2. The virulence of common serotypes may not always be significantly associated with SNPs and Indels. Other factors such as gene presence-absence, gene expression, and host conditions may cause invasiveness.

The serotype-based GWAS analysis identified 476 (in 257 genes), 972 (370 genes), and 1009 (in 376 genes) significant nonsynonymous SNPs in serotypes 12F, 5, and 1, respectively. These numbers were much greater than the one significant missense SNP identified by the location-based GWAS analysis. The location-based GWAS analysis did not identify any significant indel, but the serotype-based analysis identified three significant indels in each of serotypes 1, 5, and12F. The higher number of significant variants in serotypes 1 and 5 confirms the high dependency of the population stratification on these strains. Their short colonization period and quickness to infect blood and CSF may be associated with their genetic divergence. Serotype 1 contained the highest number of SNPs and mutated genes, implying its highest divergence may be related to its highest invasiveness.

The most polymorphic core gene was SP_1247 (*smc*), with 367 polymorphic sites (some of these polymorphisms could be sequencing errors despite the strict thresholds applied). The location-based GWAS analysis comparing nasopharynx and sterile sites did not identify any significant SNP in *smc;* however, according to the serotype-based GWAS analysis, the numbers of significant SNPs in *smc* were as follows:

- 32 significant SNPs in serotype 1, including 29 synonymous and three missense SNPs. Missense substitutions were p.Thr249Ala, p.Asn482Ser, and p.Leu879Phe, with a moderate impact on the protein's function.
- 27 significant missense SNP in serotype 5, including 24 synonymous and three missense SNPs. Missense substitutions were p.Ala159Ser, p.Thr249Ala, and p.Asn482Ser with a moderate impact on the protein's function.
- 24 significant SNPs in serotype 12F, including 22 synonymous and two missense SNPs. Missense substitutions were p.Leu808Val and p.Ala827Thr with a moderate impact on the protein's function.

The missense substitutions p.Thr249Ala and p.Asn482Ser in the SMC protein were common between serotypes 1 and 5. SMC is required for chromosome condensation and partitioning. This protein has DNA binding and ATP hydrolysis activity and is involved in DNA replication and cell division. Mutation in this protein may affect the cell division mechanism. Experimental work and protein structural analysis are required to confirm how mutations influence the function of SMC. The results showed that some mutations in SMC were significantly unique in serotypes 1, 5, and 12F.

It is important to note that being a polymorphic core gene did not necessarily mean that the gene harbored significant SNPs identified by the GWAS analysis based on the traits defined in this project. For instance, the location-based and serotype-based GWAS analyses did not identify any significant SNPs in the polymorphic gene *cbpC* (Table 4-4). In contrast, resulting from the serotype-based GWAS analysis, the polymorphic gene SP_1623 (*exp7*) harbored 49 significant SNPs in serotype 1 (43 synonymous and six missense), 75 significant SNPs in serotype 5 (61 synonymous and 14 missense), and 20 significant SNPs in serotype 12F (17 synonymous and three missense). As stated earlier, mutated *exp7* in *Listeria monocytogenes* was associated with virulence[233]. Missense SNPs in *exp7* could be considered as potential pneumococcal virulence factors, though again, experimental work is needed to confirm the effects of the substitution.

Two polymorphic core genes, SP_1380 (uncharacterized protein) and SP_2066 (catalyzing the final step in threonine synthesis), harbored the maximum number of significant nonsynonymous SNPs in serotypes 1, 5, and 12F. These two genes mutated in several positions in the significant invasive serotypes. Therefore, their functions may differ in serotypes 1, 5, and 12F in contrast with other strains. SP_1380 mutated in 8, 22, and 28 positions, and SP_2066 in 10, 16, and 15 positions in serotype 1, 5, and 12F, respectively.

Another point about the polymorphic genes was that they did not bear the most significant disruptive and missense SNPs in serotypes 1, 5, and 12F, therefore they may be flexible to accommodate changes.

Disruptive SNPs deactivated the following annotated genes in the significant invasive serotypes:

- Serotype 12F:
    - SP_0610: Involved in transporting amino acids across the membrane.
    - SP_1245: A protein from the Cof family involved in detoxification and amino acid biosynthesis.
- Serotype 5:
    - SP_0250 and SP_1001: Both contribute to transporting amino acids across the membrane.

- Serotype 1:
    - SP_1245: A protein from the Cof family involved in detoxification and amino acid biosynthesis.

According to the annotation of the genes, deleterious SNPs affected amino acid synthesis and amino acid transport in the significant invasive serotypes.

The most significant missense SNPs common between serotypes 1, 5, and 12F influenced genes such as SP_1697 (*recG*), SP_2058 (*tgt*), SP_2045, and SP_0588 (*pnp*), and SP_1207 (*xseA*). These genes contribute to DNA and RNA metabolism. In particular, the presence of aminoacyl-tRNA synthetases that attach amino acids to tRNA molecules in the list of mutated genes in serotypes 1, 5, and 12F was significant (p-value < 0.01). We know that aminoacyl-tRNA synthetases are targeted by antibiotics such as mupirocin. The stress imposed by medications may cause invasive serotypes 1, 5, and 12F to mutate these genes. However, experimental work is required to determine how variants in aminoacyl-tRNA synthetases influence the biology of pneumococci since these proteins are involved in several biological processes in the cell, such as RNA splicing and transcriptional regulation.

The functional enrichment analysis of genes mutated by SNPs and indels is critical in determining how variants change the biology of the different serotypes and are potentially related to invasiveness. However, the main limitation of the GWAS analysis was the high frequency of uncharacterized proteins that comprised 43% of the pan-genome. This caused the SNP functional enrichment analysis to be less comprehensive and informative since many genes were not annotated. Examples of these genes were SP_0830 and SP_0910 disrupted by deleterious SNPs in serotypes 1 and 5. Knowing the function of genes with different mutations is essential to determine the biological implications. In particular, better annotation of genes that bore the most significant variants in the significant invasive serotypes is needed. Again, this highlights the need for further experimental work to discover the functions of pneumococcal uncharacterized proteins.

5 Analysis of large-scale variants in the accessory-genome

5.1 Overview

Pathogenic bacteria such as *S. pneumoniae* exhibit an extremely high level of diversity in their genome as a result of various evolutionary pressures such as the host immune system, vaccines, and medicines, which suggests some of this variation may be due to selective adaptation[244]. Despite the sequence similarity in parts of the genome between isolates obtained from different body niches, known as the core-genome that often encodes housekeeping functions, there are differences between the accessory-genomes of commensal and invasive *S. pneumoniae* strains. Some accessory genes cluster into the genomic loci known as Regions of Diversity (RDs), as described in chapter 3 (Figure 3-11). The accessory genome is actually the main contribution to invasiveness and post-vaccine success of pneumococcal strains[245]. It is essential to understand the virulence-associated properties of pneumococci by comparative analysis of accessory components to understand the pathogenesis and prevent vaccine breakthroughs.

Pneumococcal colonization in the upper respiratory tract coupled with frequent exposure to other microbes in the same niche is believed to provide enough time and a significant genetic repertoire to enable the flexibility to evolve into a pathogenic form and invade blood and CSF. The most well-known strategies (some known as horizontal gene transfer) for bacterial genomic diversification and rearrangements are plasmids, bacteriophages, transposons, and genomic islands, helping bacteria improve their fitness and potentially shaping their virulence[244]. Regardless of the mechanisms of their generation, invasive accessory genes associated with pneumococcal virulence usually encode functions such as adhesion, toxicity, invasion, evasion of host immune response, and metabolic processes required for anaerobic respiration in blood and CSF[246]. The well-studied accessory genes in *S. pneumoniae* encode capsular polysaccharides, and isolates assigned to particular serotypes have a higher potential to cause invasive disease[247].

Understanding the mechanisms for the development of invasive pneumococcal disease requires large-scale genomics and epidemiological studies in which specimen source, serotype, and the complete set of accessory genes could be studied simultaneously. In the previous chapter, we discussed that despite their close genetic relationships, pneumococcal serotypes strikingly have different pathogenicity capabilities, where serotypes 1, 5, and 12F were significantly identified in sterile sites. This chapter hypothesizes that stratification in the accessory-genome follows the pattern of variation in the core-genome. A PCA of gene presence-absence profiles tested the hypothesis. Serotypes would be expected to be the main drivers of the gene distribution pattern in the accessory genome. This feature should be more remarkable for serotypes 1 and 5, meaning many accessory genes would be present (or absent) only in the genome of these serotypes. In this chapter, the presence of genes in isolates was investigated using genome-wide association methods to determine how samples were clustered based on the distribution of genes in the pan-genome. Statistical tests were applied to identify candidate genes possibly linked to pneumococcal virulence. Finally, functional analysis was performed to identify the most important biological processes that putative virulence genes carry out in the cell and how they may contribute to pneumococcal pathogenesis.

5.2 Methods

5.2.1 Population structure analysis using large-scale variants

In pan-genomics studies, the population structure can be investigated from two perspectives:

1. Based on the variation in the core-genome, SNPs are the markers used to determine how the population is stratified.
2. Based on the variation in the accessory-genome, the presence and absence of genes are used to characterize the distinct clusters.

As stated in section 4.2.3, PCA, a dimensional reduction method, was used for population structure analysis, mapping high dimensional data to a low dimension space and then clustering the data. For the core-genome-based process, one vector was assigned to each sample so that the vector dimension was equal to the number of SNPs, which was 31914. The presence and absence of SNPs in each sample were represented by 1 and 0.

In this chapter, for the accessory-genome-based analysis, the presence and absence of genes were represented by 1 and 0. The vector dimension was equal to the number of genes in the pan-genome, which was 6803. The importance of the gene-based PCA was that it identified the main drivers of variation in the accessory-genome and specified which cluster of samples had distinct gene content. This helped to address the aim of the thesis reflected in its title:

"A pan-genome wide association study to identify genes associated with invasive Streptococcus pneumonia in Malawi."

The R Bioconductor package MixOmics[218] was used for the PCA of gene distribution in the accessory-genome as applied for the core genome in section 4.2.3. The analysis detected distinct subgroups of samples that could be driven by any feature, such as isolation sites and serotypes. However, considering the results of the population structure analysis in chapter 4, the hypothesis was that serotypes 1 and 5 were most likely to form distinct clusters with a set of specific genes present (absent) in their genome structure.

The MixOmics package applies supervised learning methods to explore data and reduce dimensions to identify the most discriminant subset of biological features. The gene-based PCA results identified the potential drivers of variation in the accessory genome that would be used to define traits for the gene presence-absence analysis.

5.2.2 Gene presence-absence analysis

A pan-genome-wide association study (pan-GWAS) approach was applied to investigate the presence-absence of genes in different strains. The association between genes and phenotypic traits of isolates was examined by Scoary version 1.6.16[130]. This tool is implemented as a standalone python script that scores genes in the pan-genome for associations to the phenotypic traits while accounting for population structure. Scoary uses a series of filters to remove genes not associated with the assigned traits to save computational resources. Scoary excluded core genes from the analysis for any particular trait because core components were common in all samples and provided no information for the association test. Scoary then collapsed correlated genes into a single unit and assigned a null hypothesis

of no association to the trait for each unit. Scoary then applied *Fisher's exact* test on each gene and used Bonferroni and Benjamini–Hochberg adjustments to correct for multiple comparisons. In the next step and to consider the structure of the population, Scoary implemented the pairwise comparisons algorithm[248] using a phylogenetic tree calculated from the Hamming distances in the pan-genome matrix. The pairwise comparisons algorithm calculated the maximum and the minimum number of pairs that support an association. Finally, Scoary ran label-switching permutations to produce a single list of significant genes per trait. Every gene with a Bonferroni corrected p-value less than 0.05 was reported as significant.

5.2.3 Functional enrichment analysis of significant genes

After identifying candidate genes likely associated with invasiveness, a functional enrichment analysis was performed to gain insight into the overall functions of the putative virulence factors. The task was similar to that previously described in section 3.2.6.

The gene set enrichment analysis (GSEA) algorithm was applied using STRING[173]. Genes were ranked according to their significance and submitted to the STRING webtool. The "Proteins with Values/Ranks" tool was selected, and the *Streptococcus pneumoniae* was chosen as the reference organism. Any gene cluster that produced an FDR less than 0.05 was reported by STRING as a significant cluster. The identified clusters were filtered manually, and the final clusters were mapped to the TIGR4 genome by Ensembl[249].

For further investigation, each set of significant genes was also submitted to the ShinyGO tool, and the significant pathways were reported as described in section 3.2.6.

5.3 Results

5.3.1 Population structure analysis based on the distribution of the genes in the accessory-genome

As mentioned previously, the hypothesis to be tested was that the patterns of stratifications in the core- and accessory-genome were likely similar, which means the serotypes of samples best explain the diversity, and serotypes 1 and 5 would cluster separately from other strains.

The information provided by the gene-based population structure analysis had a crucial role in defining traits of samples for gene presence-absence statistical analysis. The PCA results illustrated in Figure 5-1 in two-dimensional space detected which groups of samples had the highest distinction in their gene content, clearly confirming the accuracy of the above hypothesis.

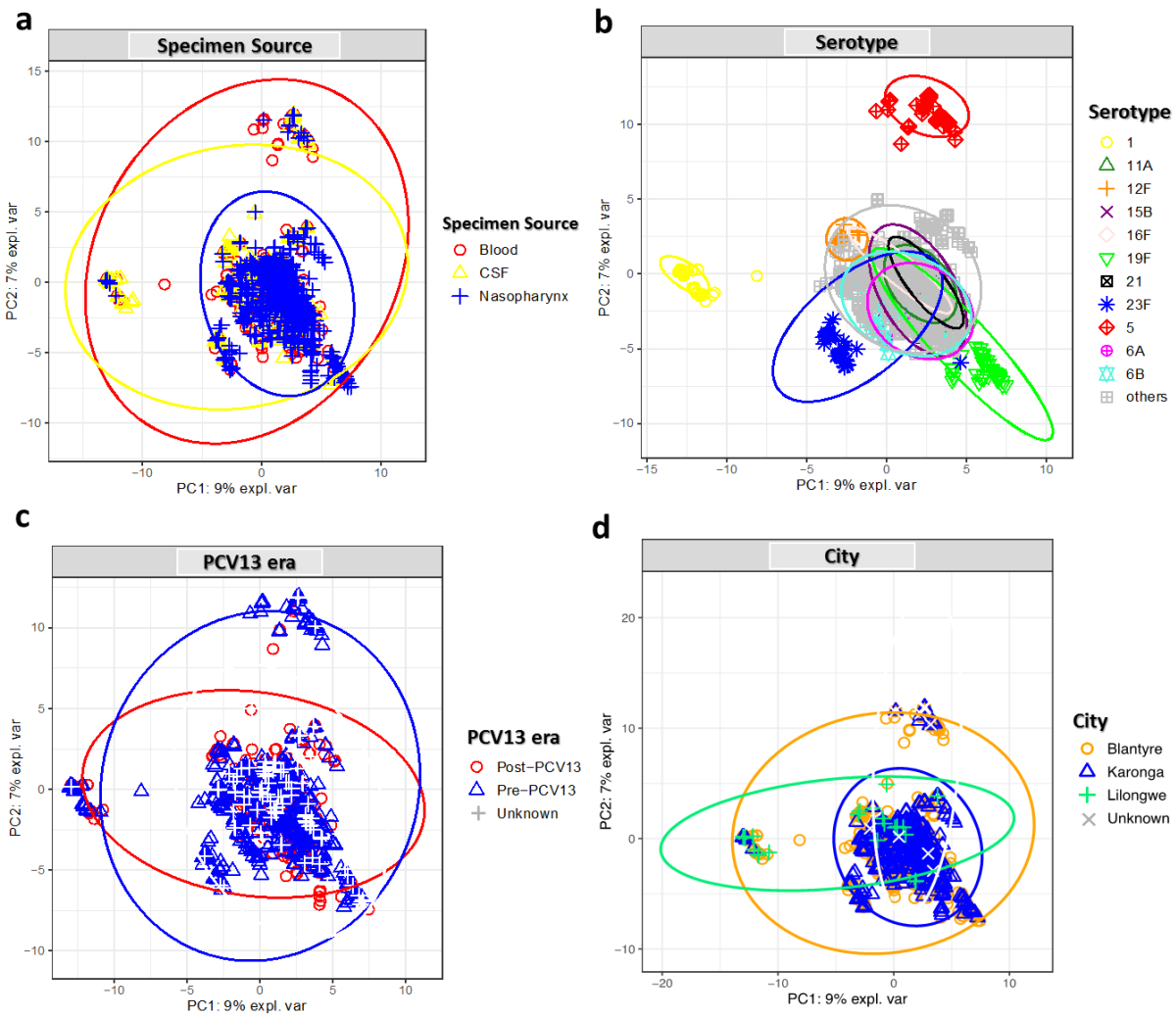


Figure 5-1. The gene-based PCA indicating the distinction between the gene content of isolates.
The effects of specimen sources, serotypes, vaccination eras, and geographical locations on the population structure are illustrated in panels a, b, c, and d, respectively. Serotypes best explain diversity in the accessory-genome. Serotypes 1 and 5 are evidently distinct.

The isolation sites (specimen sources) in panel *a* could not explain the population stratification because the largest cluster in the middle of the panel contained a mixture of samples from all isolation sites, including nasopharynx, blood, and CSF. Therefore, at the time of sampling, the set of samples from the nasopharynx was not genetically different from those from sterile sites.

Panel b verified serotypes as the best factor to describe the diversity in the accessory-genome. The separation pattern in the accessory-genome was similar to the observation in the core-genome as described in chapter 4. Serotypes 1 (yellow clusters) and 5 (red cluster) were clearly separated from other strains. The third cluster consisted of four subgroups, including serotypes 12F (orange cluster), 19F (green cluster), 23F (blue cluster), and other strains (grey cluster). Serotype 12F was significantly invasive and present in the sterile sites (as was observed with serotypes 1 and 5), while serotype 19F was significantly prevalent in the nasopharynx. Serotype 23F was an abundant and ubiquitous strain

frequently found in the nasopharynx and sterile sites. Serotype 23F is known for its resistance to multiple antibiotics.

Figure 5-1b provides vital information to define the trait of samples for the gene presence-absence analysis confirming serotypes as the best determinant of variations in the pan-genome. Nonetheless, the analysis must rigorously account for population structure and carefully curate any hits to show statistical association with the disease across the population. The separation between serotypes 1, 5, 12F, and the other serotypes may be just because these strains were hardly found in the carriage group. Therefore, these patterns may primarily reflect biases in sampling rather than genuine genetic phenomena. To address this issue, ten samples of each hyper-invasive serotypes (1, 5, and 12F) and the vaccine types (involved in PCV13) were randomly selected from the nasopharynx, blood, and CSF. The PCA of the gene distribution was repeated for the downsampled dataset and identified the same pattern of stratification so that each of serotypes 1, 5, and 12F formed a distinct group clustered distantly from other strains (Figure 5-2). All serotypes in Figure 5-2 are invasive, however, the difference between the hyper-invasive serotypes (1, 5, and 12F) and the rest is the speed of the hyper-invasive serotypes at infecting the sterile sites. This may strengthen the hypothesis that the high genetic distinction of the hyper-invasive serotypes is associated with their quickness to enter the blood and CSF.

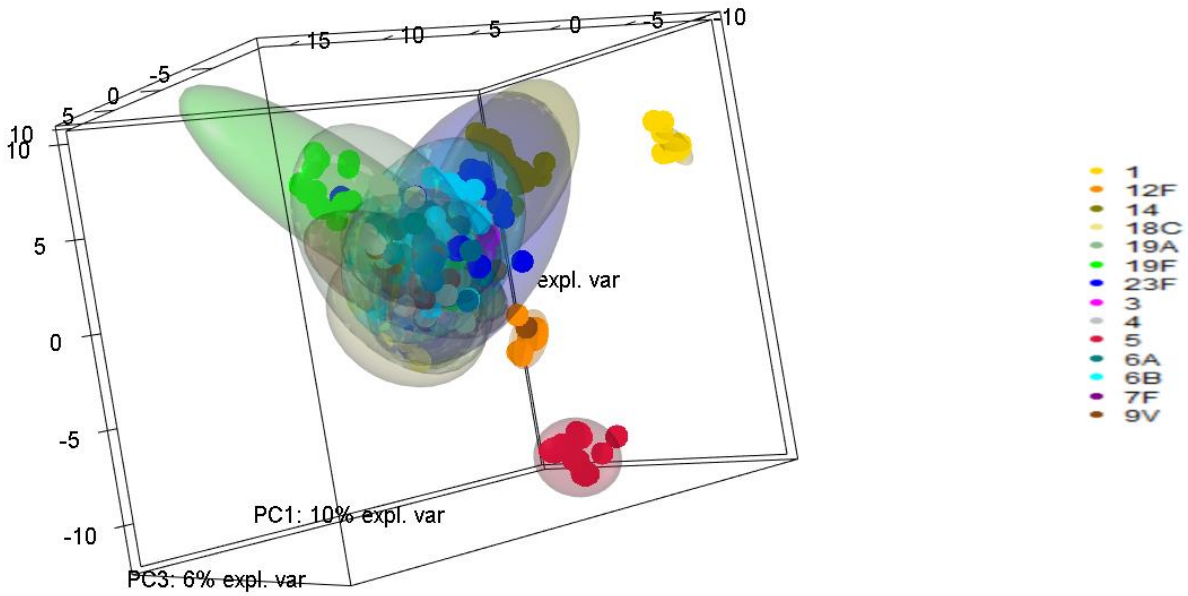


Figure 5-2. The three-dimensional PCA of the gene distribution applied to the downsampled dataset.
Ten samples were randomly selected from the nasopharynx, blood, and CSF for each serotype. The PCA was conducted using the R package MixOmics. Hyper-invasive serotypes 1, 5, and 12F clustered separately from other strains.

The population could be divided into six subpopulations, including serotypes 1, 5, 12F, 19F, 23F, and other strains. Serotypes 1, 5, and 12F represent the significant invasive, and serotype 19F represents the significant non-invasive strains. Therefore:

- Any significant genes present (or absent) in serotypes 1, 5, and 12F may be associated with virulence because these serotypes were significantly present in the patient groups with a low frequency among carriers.
- Genes significantly present (absent) in serotype 19F could likely explain its low invasiveness in Malawi.
- Any gene significantly present (absent) in serotype 23F could be investigated for its possible relevance to multi-antibiotic-resistance.

Panel *c in* Figure 5-1 demonstrated the effect of vaccination on bacterial evolution. The results from collected samples in this study showed that the vaccination did not cause any considerable change in the accessory-genome. Samples were assigned to pre- and post-PCV13 categories based on their collection times before and after introducing the vaccine in Malawi in November 2011. There were samples from both pre- and post-PCV13 eras in all of the main clusters identified by the gene-based PCA. In summary, the vaccination time did not cause any genetic divergence in the cohort. However, as suggested earlier, more research is required to correctly understand the effect of vaccination since it was not clear which samples in the post-PCV13 group were obtained from a vaccinated person. Furthermore, the population must be larger to make a solid conclusion about the effect of vaccination.

In panel *d* of Figure 5-1, the samples were colored according to their geographical location, which could not explain the population structure as efficiently as the serotype of samples. It must be emphasized again that the location distributions of the nasopharyngeal and invasive populations were not similar:

- Nasopharyngeal samples: 15% were collected from Blantyre, 85% were from Karonga, and no sample was from Lilongwe.
- Invasive samples: 95% of invasive samples were collected from Blantyre, none of them were from Karonga, and 5% were from Lilongwe

Since the gene content of the pneumococcal populations depends on several factors, including geographical locations, any significant difference between nasopharyngeal and invasive isolates might be just the difference between Karonga and Blantyre. However, Karonga and Blantyre are not very far from each other. They are just 820 kilometers apart. To address the problem and avoid introducing any batch effect during the gene presence-absence analysis, nasopharyngeal samples from Blantyre and Karonga had to be compared. If their gene pools were similar, the effect of geographical locations could be ignored, and any difference between nasopharyngeal and invasive samples could potentially be associated with the invasiveness.

5.3.2 Gene presence-absence statistical analysis

The phenotypic trait of isolates was defined mainly based on the outcomes of the population structure analysis and the main goal of the research to identify virulence genes. The dataset involved six distinct clusters, including serotypes 1, 5, 12F, 23F, 19F, and other strains. Therefore, the gene presence-absence comparison between these six clusters could identify the maximum number of significant genes. However, comparing the nasopharynx and sterile sites could still be productive since the gene-based PCA applied for the population structure analysis separated samples based on the entire gene pools of the clusters. If samples in the nasopharynx and sterile sites diverged by the presence of a few genes, the PCA would not be able to detect this and separate samples clearly in the two-dimensional space.

Therefore, the gene presence-absence analysis examined the distribution of the genes across the isolation sites (location-based) and serotypes (serotype-based) using the set of comparisons to determine any significant change in the pneumococcal accessory-genome that may be related to virulence. The following factors were considered for the evaluation of the accessory gene distribution:

1. As described in chapter 3, the presence of paralogs in the accessory-genome must be considered for the gene presence-absence investigation. Paralogs are gene duplicates with different IDs but similar functions. Their presence amongst the significant genes in two clusters of samples could skew the enrichment analysis as they do not represent the real difference between the two clusters. The paralogous genes must be excluded from the enrichment analysis as they produce misleading results, such as identifying the same enriched pathways in both clusters of samples.
2. More than half of the accessory-genome (53.58%) was composed of uncharacterized proteins. Therefore, a complete functional enrichment analysis of the significant genes might not be achievable.
3. Identification of significant genes working together in annotated operons was of particular interest in this research, especially those from RDs since these regions were known to be associated with virulence.

5.3.2.1 Location-based analysis

5.3.2.1.1 Blantyre vs. Karonga (nasopharyngeal isolates)

The carriage isolates from Karonga were compared to carriage isolates from Blantyre. The test did not detect any significant gene in either Blantyre or Karonga groups, meaning that the gene content of the carriage samples from Blantyre and Karonga was quite similar. Therefore, the geographic locations would not introduce any batch effect to the test between nasopharyngeal and invasive isolates.

5.3.2.1.2 Nasopharynx vs. sterile sites (only grey cluster)

As a mixture of many serotypes, the grey cluster in panel *b* of Figure 5-1 was split into the nasopharyngeal and invasive samples (collected from blood and CSF). The statistical test in this step investigated the difference between these two groups to identify any significant genes not highlighted by the PCA.

Four significant genes were identified in the nasopharynx samples (Table 5-1). The most significant genes were neighbors in the genome:

- SP_0537: An open reading frame from pneumococcal insertion sequence IS1381. This gene encodes a transposase enzyme required to move the insertion element[250].
- SP_0536 (*blpL*): This gene encodes an immunity protein. In the nasopharynx, pneumococci compete for resources with other bacteria. They secrete antimicrobial peptides and proteins known as bacteriocins that target other bacteria in the upper respiratory tract. However, they also secrete immunity proteins to protect against self-destruction.

The significant presence of SP_0536 and SP_0537 was likely due to the exchange of genetic elements and competition in the nasopharynx.

The next two genes in Table 5-1 (SP_0576 and SP_0577) with a lower significance were also neighbors on the chromosome and involved in carbohydrate metabolism:

- SP_0576: Encoding beta-glucoside operon transcriptional antiterminator.
- SP_0577: Encoding the phosphoenolpyruvate phosphotransferase system (PTS), beta-glucosides-specific IIABC components.

Of note was that the significant genes in the nasopharynx were within a relatively short region of the genome (32kbp), and none of them were from the previously described RDs.

Table 5-1. Significant genes in the nasopharynx with a Bonferroni corrected p-value less than 0.05.

STRING ID	Preferred name	P-value
SP_0537	SP_0537	0.000639699
SP_0536	*blpL*	0.001090243
SP_0576	*licT*	0.003230874
SP_0577	SP_0577	0.003230874

The analysis identified 15 significant genes in the sterile sites (Table 5-2).

The four genes at the top of the table with the highest significance were:

- SP_0535: Encoding an uncharacterized protein
- SP_0359 (*cpsK*): Encoding UDP-2-acetamido-2,6-beta-L-arabino-hexul-4-ose reductase, a capsular polysaccharide biosynthesis protein (CpsK).
- SP_0360 (*cpsL*): Encoding CpsL belonging to the UDP-N-acetylglucosamine 2-epimerase family and involved in capsular polysaccharide synthesis.
- SP_0358 (*capJ*): Encoding Udp-n-acetylglucosamine 4,6-dehydratase/5-epimerase, a capsular polysaccharide biosynthesis protein (CpsJ).

The *cps* genes mentioned above were from the RD3. Their significant presence in the invasive isolates could imply a higher level of capsulation in the sterile sites protecting the pathogen from the host immune response. As stated in section 1.1.1, non-encapsulated pneumococci have better adherence to the respiratory epithelial cells compared to encapsulated isolates[23].

Other annotated significant genes in the sterile sites were:

- SP_1055 (Orf9 protein) and SP_1056 (relaxase) from RD6, which were the elements of the pneumococcal conjugative transposon Tn5252.
- SP_0015 that encodes a transposase protein and contributes to genetic transposition.

As discussed in section 3.3.5, the presence of paralogs in the accessory-genome was considerable. The results in this section identified two paralogs of SP_0015, 12 paralogs of P_1055, and 18 paralogs of SP_1056 in the accessory-genome. Interestingly, all paralogs of these genes were significantly present only in invasive samples, suggesting their potential roles in virulence.

Another significant gene in the sterile sites was *ftsk,* which encodes a DNA segregation ATPase that coordinates cell division. However, the cell division controlling genes *ftsA* and *ftsZ* were identified as conserved core genes in section 4.3.1. Therefore, the mechanism of cell division might require a different set of genes in the nasopharynx and sterile sites. Studies suggest that *ftsA* and *ftsZ* are essential components for pneumococcal cell division[251]. Gene *ftsk*, however, may be less important for cell growth in the nasopharynx.

Table 5-2. Significant genes in the sterile sites with a Bonferroni corrected p-value less than 0.05.

STRING ID	Preferred name	P-value	Serotypes present in
SP_0535	SP_0535	2.19E-07	4, 6G, 6A, 6B, 7F, 7C, 14, 16F, 17F, 28F
SP_0359	cpsK	5.95E-05	4, 12B, 45, 46
SP_0360	cpsL	5.95E-05	4, 12B, 45, 46
SP_0358	capJ	5.95E-05	4, 12B, 45, 46
SP_1055	SP_1055	0.000207345	3, 4, 6A, 6G, 9L ,9V ,16F ,19B, 33D, 46
SP_0878	ftsK	0.002025695	3, 4, 6A, 6B, 9V, 10B, 14, 16F, 35A
SP_1936	SP_1936	0.004029749	4, 6A, 6B, 7F, 9V, 10B, 14, 15A, 15B, 35B
SP_1056	SP_1056	0.004935925	3, 4, 6G, 6B, 9L, 9V, 13, 15B, 16F, 46
SP_0015	IS630-Spn1	0.00527717	3, 6A, 6B, 9V, 13, 14, 16F, 18C, 19A, 35B
SP_1916	SP_1916	0.010058032	4, 6G, 6A, 6B, 12B, 18F, 18C, 24, 46
SP_1742	tmcAL	0.010672534	6A, 6G, 9L, 12B, 24, 31, 35B, 38, 46
SP_1037	SP_1037	0.026513467	4, 6A, 6B, 9V, 13, 14, 15B, 16F, 18C, 35B
SP_1221	SP_1221	0.028294187	6G, 6A, 7C, 12B, 16F, 18C, 19A, 21, 24, 46
SP_1222	SP_1222	0.028294187	6G, 6A, 7C, 12B, 16F, 18C, 19A, 21, 24, 46
SP_0817	SP_0817	0.047011811	3, 4, 6A, 6B, 9V, 14, 15B, 18C, 19A, 35B

The log-transformed p-value distributions of significant genes identified by the location-based analysis are shown in Figure 5-3. Genes are more significant in the sterile sites than in the nasopharynx. This illustration is helpful in comparing the significance level of genes identified by location-based and serotype-based analysis.

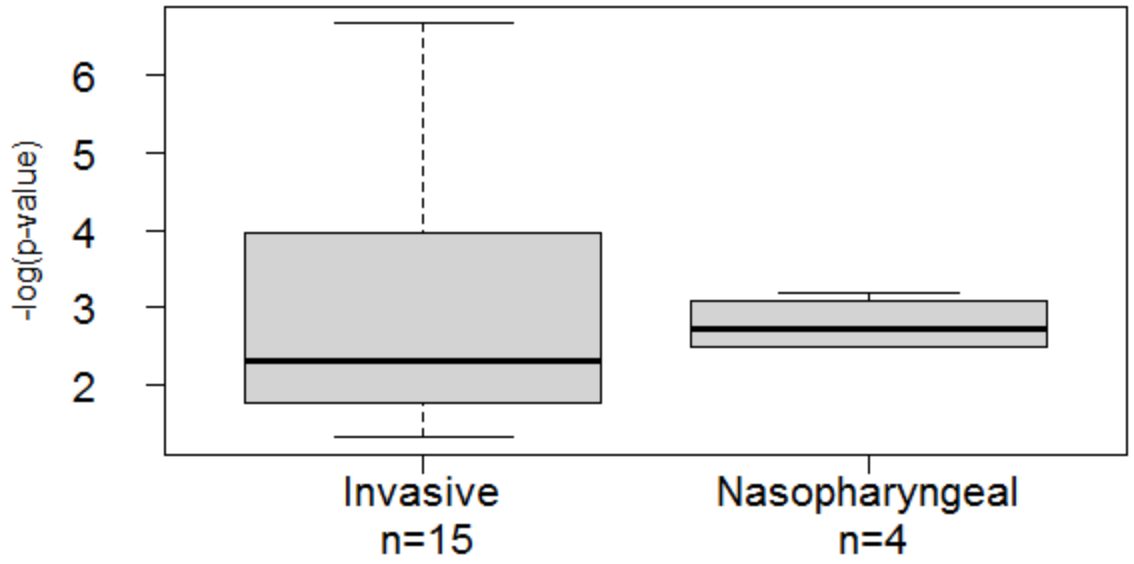


Figure 5-3. The log-transformed p-value distribution of significant genes identified by the location-based statistical analysis.

5.3.2.1.3 Blood vs. CSF

As a comparison between invasive samples inside the grey cluster in Fig 5-1b, isolates from blood were compared to those in the CSF to identify genes potentially assisting the pathogen in crossing the blood-brain barrier. However, the statistical test between blood and CSF did not identify any significant genes.

5.3.2.2 Serotype-based analysis

The core objective of the research was to identify genes associated with invasive *pneumococcus* in Malawi. Through the population structure analysis, a serotype-based analysis was determined to be the best approach to detect any potential virulence genes (Figure 5-1). The following considerations were taken into account:

- The nasopharyngeal population was most likely a mixture of non-invasive and invasive isolates. The best representatives of non-invasive isolates were serotypes 16F and 19F, with a significant presence among carriers and a low prevalence in the patient group. The population structure showed a distinction in the accessory-genome of serotype 19F compared with other nasopharynx strains.
- Isolates in the blood and CSF could be treated as invasive samples. However, the tight clustering, representing genetic homogeneity, only existed in serotypes 1, 5, and 12F, with a significant abundance in the sterile sites.

Based on the above, one solution for identifying potential virulence genes was to compare the significant invasive serotypes (1, 5, and 12F) with serotype 19F. The result would provide lists of putative virulence factors in serotypes 1, 5, and 12F, which were the most frequent and comprised about 40% of strains in the patient group. In particular, significant genes gained or lost in serotype 1 could potentially explain its rapid infection ability. Serotype 1 required special attention due to its characteristics:

- Serotype 1 was the most frequent strain, with 8% relative frequency in the entire population (nasopharynx + blood + CSF). Since 56 serotypes were detected, the expected frequency for each serotype was 1.8%, meaning serotype 1 was 4.5 times more frequent than expected.
- The most remarkable fact about serotype 1 was that it comprised less than 1% of the nasopharyngeal strains, whereas its prevalence in sterile sites was 19%. This is most likely because of its short colonization period and reported rapid invasion into the blood and CSF, particularly the central nervous system.
- The relative frequency of serotype 1 increased after the vaccination program in 2011.

It was worth noting that the accessory-genome of serotype 19F did not cluster very distantly from the grey cluster in Figure 5-1, which included the entire nasopharyngeal group. Therefore, the results should not be very different if the significant invasive serotypes (1, 5, and 12) were compared to serotype 19F or the entire carriage group. However, serotype 19F was a better representative of non-invasive isolates than the grey cluster in Figure 5-1. The grey cluster includes several serotypes with a similar abundance in the carriage and patient groups. These serotypes are most likely invasive, and their presence in the carriage group is due to their colonization period before infecting the sterile sites. But serotype 19F was an abundant strain in the nasopharynx with a significantly low frequency in the blood and CSF. Therefore, serotype 19F can be considered as a non-invasive strain.

The gene-based PCA was ineffective in differentiating the entire nasopharyngeal and invasive populations. The reason could be the presence of invasive strains in the nasopharynx, such as 6A, 6B,

and 23F. These serotypes were abundant and common between the nasopharynx and sterile sites. Their presence in the blood and CSF indicates their potential for invasiveness, while their presence in the nasopharynx could mean that they were collected during the colonization phase before infecting the sterile sites. The gene pools of the common serotypes (6A, 6B, and 23F) were similar during colonization and infection. Therefore, instead of gene presence-absence, there may be other factors, such as gene expression contributing to their invasiveness.

5.3.2.2.1 Serotype 12F vs. 19F

The statistical test identified 66 significant genes in serotype 12F and 82 significant genes in serotype 19F (Appendix 8). The p-value distributions of significant genes are illustrated as boxplots in Figure 5-4. The significance levels were slightly lower for serotype 19F.

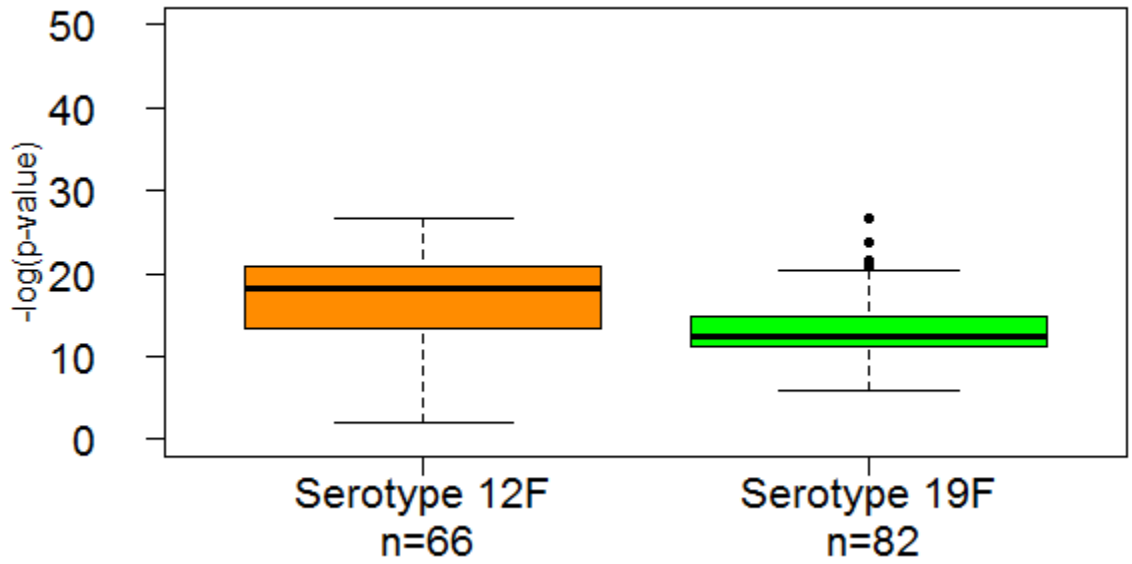


Figure 5-4. The log-transformed p-value distribution of significant genes in serotypes 12F and 19F. Five outliers in serotype 19F are shown as black dots above the green boxplot.

The network of significant genes in serotype 12F shown in Figure 5-5 highlighted the notable existence of components from RD2, RD3, RD5, RD6, and RD8b2 in serotype 12F. The function of genes in these RDs are as follows:

- Genes from RD2 are involved in quorum sensing, spore formation, and secretion of bacteriocins.
- The most significant genes in serotype 12F were from the *cps* locus (RD3). This was predictable as 87% of isolates assigned to serotype 12F were from sterile sites, whereas only 20% of samples assigned to serotype 19F were from blood and CSF. The level of capsulation is expected to be higher in serotype 12F as the capsule structure protects the pathogen from the host immune system during the infection.

- Most of the genes in RD5 have remained uncharacterized to date. The only annotated significant genes from RD5 were SP_0695, involved in thiamine biosynthesis, and SP_0698, which encodes the permease protein of an uncharacterized ABC transporter.
- Significant genes from RD6 or *Pneumococcal Pathogenicity Island 1* (PPI1) facilitate the movement of the Tn5252 transposon and the type II toxin-antitoxin system (*pezA-pezT*). Genes in this region contribute to recombination and pneumococcal bacteriocin secretion.
- Genes in RD8b2 are mostly conserved hypothetical proteins. Two annotated genes in this region are SP_1341 and SP_1342, which encode a drug efflux ABC transporter.

In addition to significant genes from RDs, the most noticeable finding was the significant presence of cellobiose and mannitol PTS transporters in serotype 12F. Genes encoding these two systems are neighbors on the chromosome. A study confirmed that deleting these PTS systems caused pneumococcal virulence to become attenuated[252]. The primary role of the PTS transporters is to uptake specific carbohydrates into the cell for metabolism.

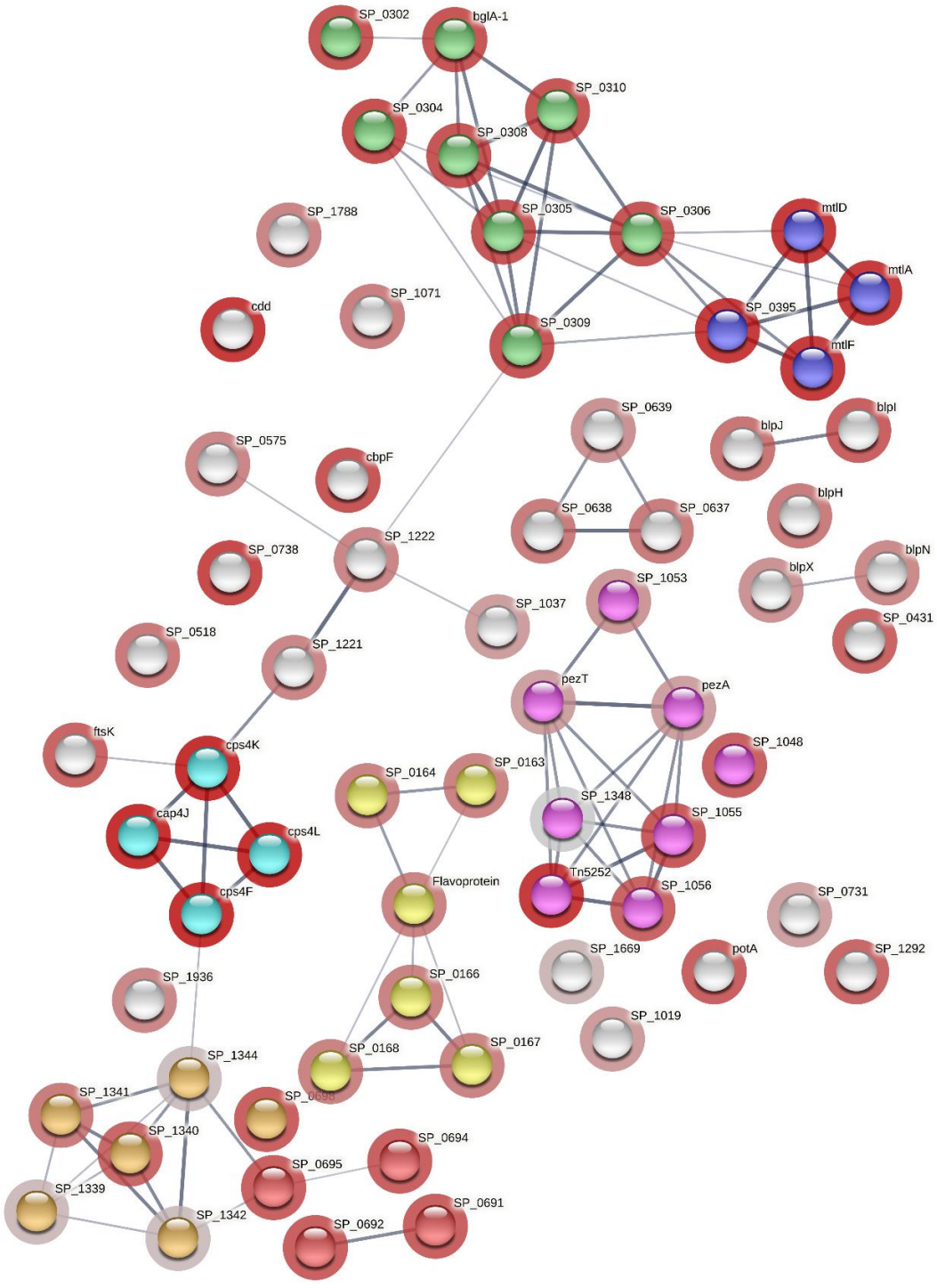


Figure 5-5. The network of interactions between significant genes in serotype 12F.
The red halo around the nodes represents gene significance, a more reddish halo represents a higher significance. The yellow cluster includes genes from RD2. Nodes indicated in light blue are genes from RD3, those nodes highlighted in red are from RD5, those in purple are from RD6, and orange nodes are from RD8b. Green and blue nodes encode the subunits of cellobiose and mannitol PTS transporters.

Based on comparing serotypes 12F and 19F, significant genes in serotype 19F characterized the gene absence from the invasive serotype 12F with respect to a non-invasive strain (19F). Therefore, when considering pathogenesis, the absence of these genes could be potentially linked to invasiveness, as virulence may be correlated with gene loss instead of gene gain. The most significant genes in serotype 19F were five outliers above the green boxplot in Figure 5-4, shown as nodes surrounded by the most intense red halo in Figure 5-6:

- SP_1273 (*licD1*): Encoding a protein from the LicD family involved in phosphorylcholine metabolism. Studies showed that LicD mutants have a reduced ability to take up choline and a decreased ability to adhere to the host cells[253]. Therefore, SP_1273 is likely necessary for nasopharyngeal colonization, and its function might not be beneficial for serotype 12F with a short colonization period. A significant absence of this gene was also seen in serotypes 1 and 5.
- SP_0570: P-loop containing nucleoside triphosphate hydrolase involved in DNA metabolism and recombination[254], and a neighbor of the below gene, SP_0569. Recombination is more likely to happen in the nasopharynx than in the sterile sites due to the higher density of pneumococcal colonies in the upper respiratory tract.
- SP_0569: DNA (cytosine-5-)-methyltransferase, a component of restriction-modification systems, serves as a tool for manipulating DNA[255]. The significant presence of this gene in serotype 19F may be required because more interaction between isolates in the nasopharynx leads to a higher probability of genetic exchange between them that requires a protective response.
- SP_1685 (*nanE2*): The information available in the STRING database showed evidence of co-occurrence between SP_1685 that encodes NanE2 (N-acetylmannosamine-6-phosphate 2-epimerase 2) and SP_1330 from RD8a that encodes NanE1 (N-acetylmannosamine-6-phosphate 2-epimerase 1). These genes function together to cleave terminal sialic acid residues from epithelial glycoconjugates, assisting the pathogen in penetrating the mucus layer and adhering to the epithelial cells in the nasopharynx. Therefore, their functions seem to be beneficial for nasopharyngeal colonization. As illustrated in Figure 5-6, all genes in RD8a were significantly absent from serotype 12F. More details about RD8a are explained below.
- SP_1310: An open reading frame (OrfA) from insertion sequence IS1381l, encoding the DNA-binding region of transposase enzymes, necessary for efficient DNA transposition.

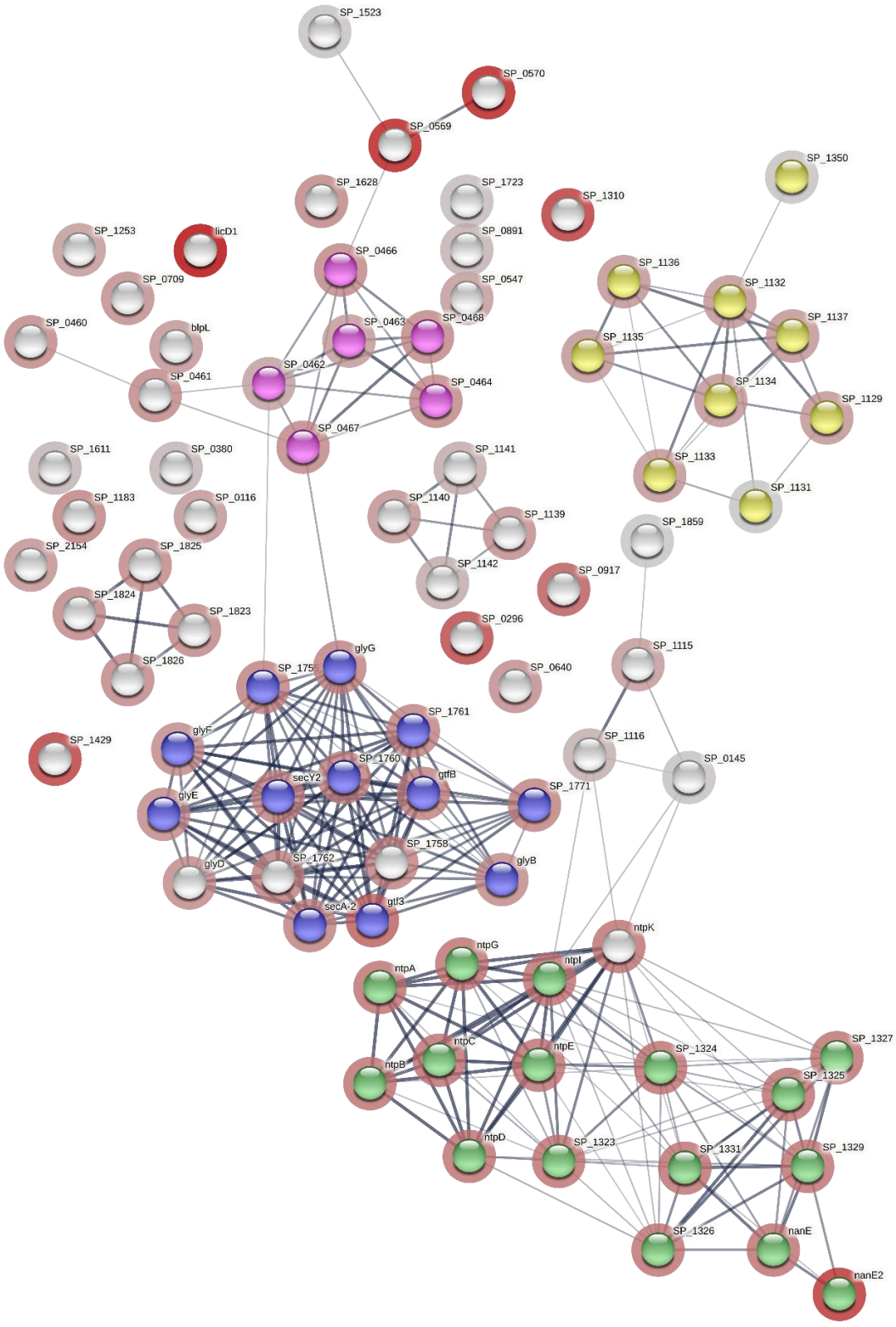


Figure 5-6. The interactions between significant genes absent from serotype 12F (present in serotype 19F).
The red halo around the nodes represents gene significance, a more reddish halo represents a higher significance. Nodes in purple are from RD4. the yellow cluster includes genes from RD7, nodes colored in green are from RD8a, nodes in the blue cluster are from RD10.

The clusters of genes from RD4, RD7, RD8a, and RD10 were significantly absent from the genome of serotype 12F in contrast with serotype 19F (Figure 5-6). In the Malawian cohort, the functions of genes from these regions appear to be essential for the colonization of the nasopharynx.

As explained earlier in section 3.3.5:

- RD4 is composed of a group of sortase enzymes that catalyze the assembly of pilins into pili and the anchoring of pili and other surface proteins to the cell wall[193,194]. The pilus is a hair-like structure associated with bacterial adhesion and colonization[256]. Due to its short colonization period, genes from RD4 that assemble the pilus were either not required in serotype 12F or its absence may be the reason for the short colonization period.
- Genes in RD7 were uncharacterized.
- RD8a is composed of RD8a1 and RD8a2 Figure 5-7):
    - SP_1315 (*ntpD*), SP_1316 (*ntpB*), SP_1317 (*ntpA*), SP_1318 (*ntpG*), SP_1319 (*ntpC*), SP_1320 (*ntpE*), SP_1321 (*ntpK*), and SP_1322 (*ntpI*), SP_1323, and SP_1324 that belong to operon RD8a1 in RD8, encode subunits of *V-type proton/sodium ATPases* that produces ATP from ADP in the presence of an H^+ or Na^+ gradient across the membrane[184].
    - SP_1325, SP_1326, SP_1327, SP_1328, SP_1329, SP_1330 (*nanE1*), and SP_1331 from operon RD8a2 in RD8a encode the subunits of the *Sodium/solute symporter* that imports solutes (mostly carbohydrates). Symporter means the channel transports the solute and co-solute (in this case, Na^+) in the same direction using the energy stored in an inwardly directed sodium gradient. The energy provided by the Sodium gradient is named the Sodium Motive Force (SMF). The SMF is generated by *V-type proton/sodium ATPases*[185]. Genes in RD8a1 and RD8a2 operons work together to produce ATP and import carbohydrates.

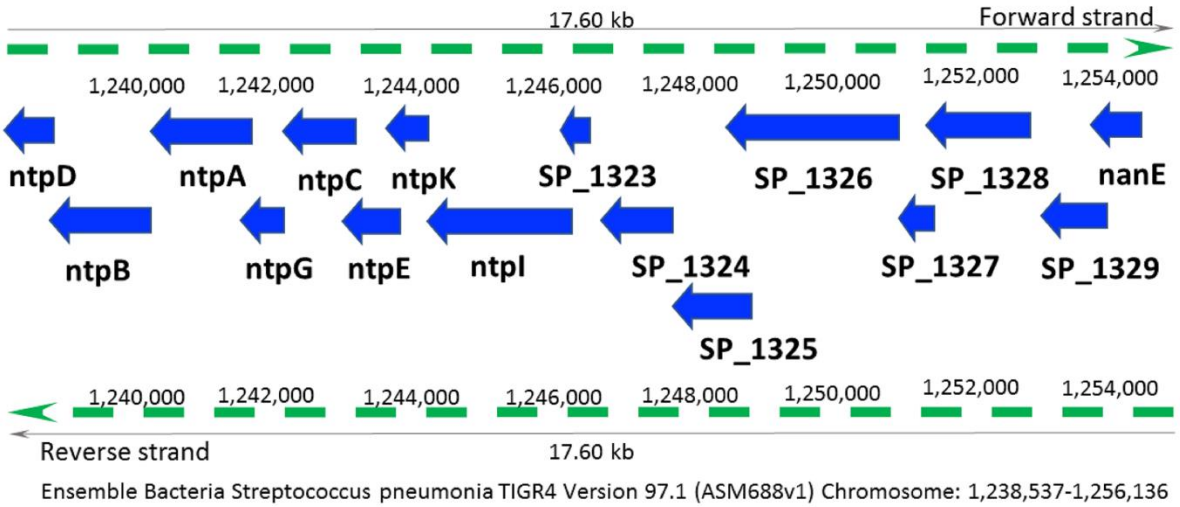


Figure 5-7. RD8a consists of RD8a1 (SP1315-1324) and RD8a2 (SP1325-SP1331).
The biological processes carried out by these genes are the transport of ions and carbohydrates across the membrane coupled with the synthesis of ATP molecules through oxidative phosphorylation.

- RD10, also known as *SecY2A2 pathogenicity island,* is responsible for the secretion of serine-rich repeat protein (PsrP). The arrangement of genes in RD10 is shown in Figure 5-8. This region contains several glycosyltransferases, including SP_1757 (*gtfB*), SP_1758 (*gtfA*), SP_1765 (*glyF*), SP_1767 (*glyD*), SP_1768 (*nss*), SP_1770 (*glyB*), and SP_1771 (*glyA*). Other genes in RD10 include SP_1759 (*secA2*), SP_1760 (*asp3*), SP_1761 (*asp2*), SP_1762 (*asp1*), and SP_1763 (*secY2*) that are secretory components. SP_1772 that is located before SP_1771 and not shown in Figure 5-8 because of its long length encodes PsrP. The presence of *gtfA* and *gtfB* is necessary for PsrP-mediated pneumococcal virulence[186]. RD10 is known as a pathogenicity island, and its presence was presumed to cause virulence[257]. However, in the Malawian dataset, this region was significantly absent from the genome of serotypes 1 and 12F.

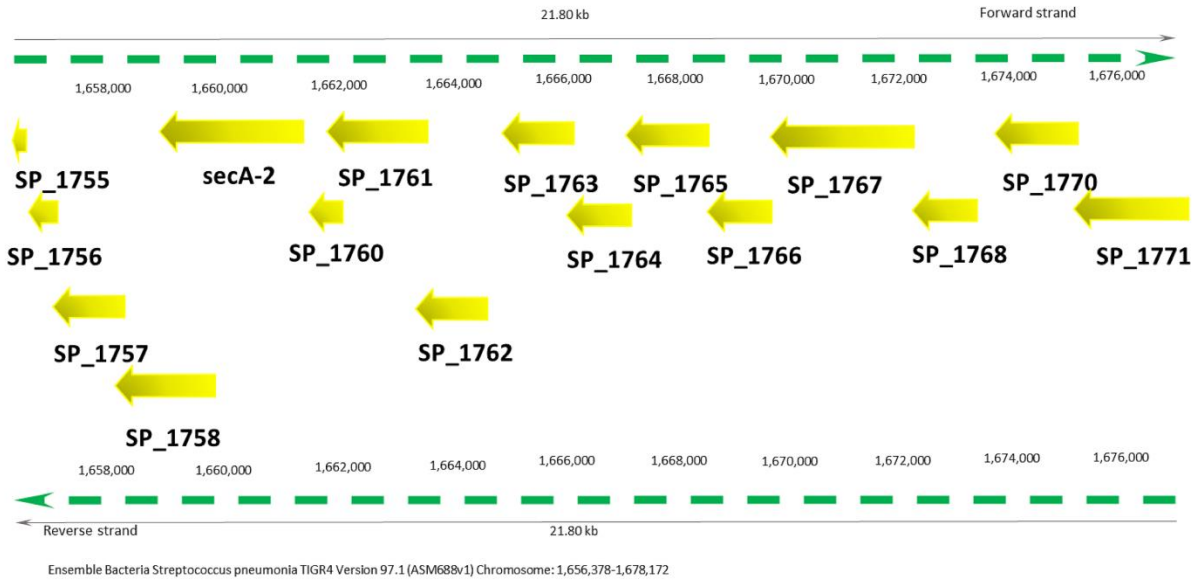


Figure 5-8. Arrangement of genes in RD10.
Genes in this region are responsible for the secretion of serine-rich repeat protein (PsrP) that enhances the binding of the pneumococci to the host cells.

5.3.2.2.2 Serotype 5 vs. serotype 19F

The statistical test identified 82 significant genes in serotype 5 and 104 significant genes in serotype 19F (Appendix 9). The p-value distributions of significant genes were illustrated as boxplots in Figure 5-9. The significance levels were similar in both serotypes 5 and 19F. In contrast to serotype 12F (Figure 5-4), the number and significance of genes present (absent) in serotype 5 were more remarkable, implying more genetic divergence of serotype 5 from the non-invasive strain 19F.

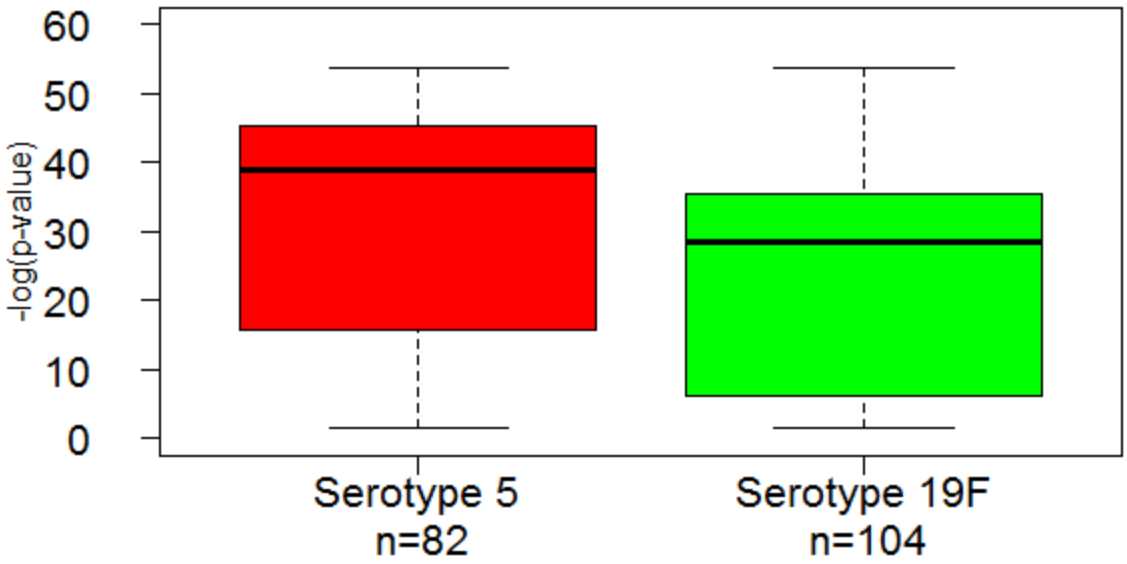


Figure 5-9. The log-transformed p-value distribution of significant genes in serotypes 5 and 19F. There is no outlier above the boxplots.

The network of interactions between significant genes in serotype 5 is shown in Figure 5-10. There were similarities and differences between significant genes present (absent) in serotypes 12F and 5. Similarities included:

- Both serotypes contained significant genes from RD3, RD6, and RD8b:
    - The presence of RD3 implied that the level of encapsulation was likely higher in both serotypes than in serotype 19F. This was expected as both serotypes quickly infect the sterile sites and must protect themselves from the host immune response.
    - Both harbored type II toxin-antitoxin system (PezA/PezT) from RD6 (PPI1). As stated in section SP_1051 (*pezT*) is the toxic component, and SP_1050 (*pezA*) is the antitoxic component. PezT is a toxic compound secreted by pneumococci for competition, and PezA protects the host cell by neutralizing the toxic effect of the cognate toxin PezT. A study highlighted the role of PezA/PezT system in producing phenotypic heterogeneity as a bacterial virulence strategy; under nutrient limitation, the PezT toxin lyses rapidly growing pneumococcal cells resulting in a heterogeneous population. The lysed cells act as a source for pneumolysin release as the main virulence factor. Moreover, the partial lysis of cells in the population results in conditions favorable to biofilm formation[258].
    - Both serotypes 5 and 12F retained genes from RD8b2. However, uncharacterized genes from RD8b3 were only observed in serotype 5.
- Both had significant genes that encode the subunits of mannitol and cellobiose PTS transporters.
- Bacteriocins BlpH, BlpI, BlpJ, BlpN, and immunity protein BlpX significantly existed in both serotypes. Bacteriocins are toxins produced by bacteria to inhibit the growth of similar or closely related strains. The bacteria producing these are protected from the effects of their bacteriocins by secreting a specific immunity protein.

The differences between the two serotypes included the following cases:

- Genes from RD2 that were significantly present in serotype 12F did not exist in serotype 5. Genes in RD2 are mostly uncharacterized, but the annotated genes in this region are involved in quorum sensing, which is a communication mechanism between pneumococci to regulate biofilm formation, virulence factor expression, competition, and metabolism[189].
- Genes from RD9 were identified in serotype 5, whereas this region was not present in serotype 12F. RD9 includes the components of the fructose and mannose PTS transporter.
- Elements of the lactose PTS transporter were significantly present only in serotype 5.

The most noticeable difference between the lists of significant genes in serotype 5 and 12F was the presence of fructose, mannose, and lactose PTS transporters in serotype 5.

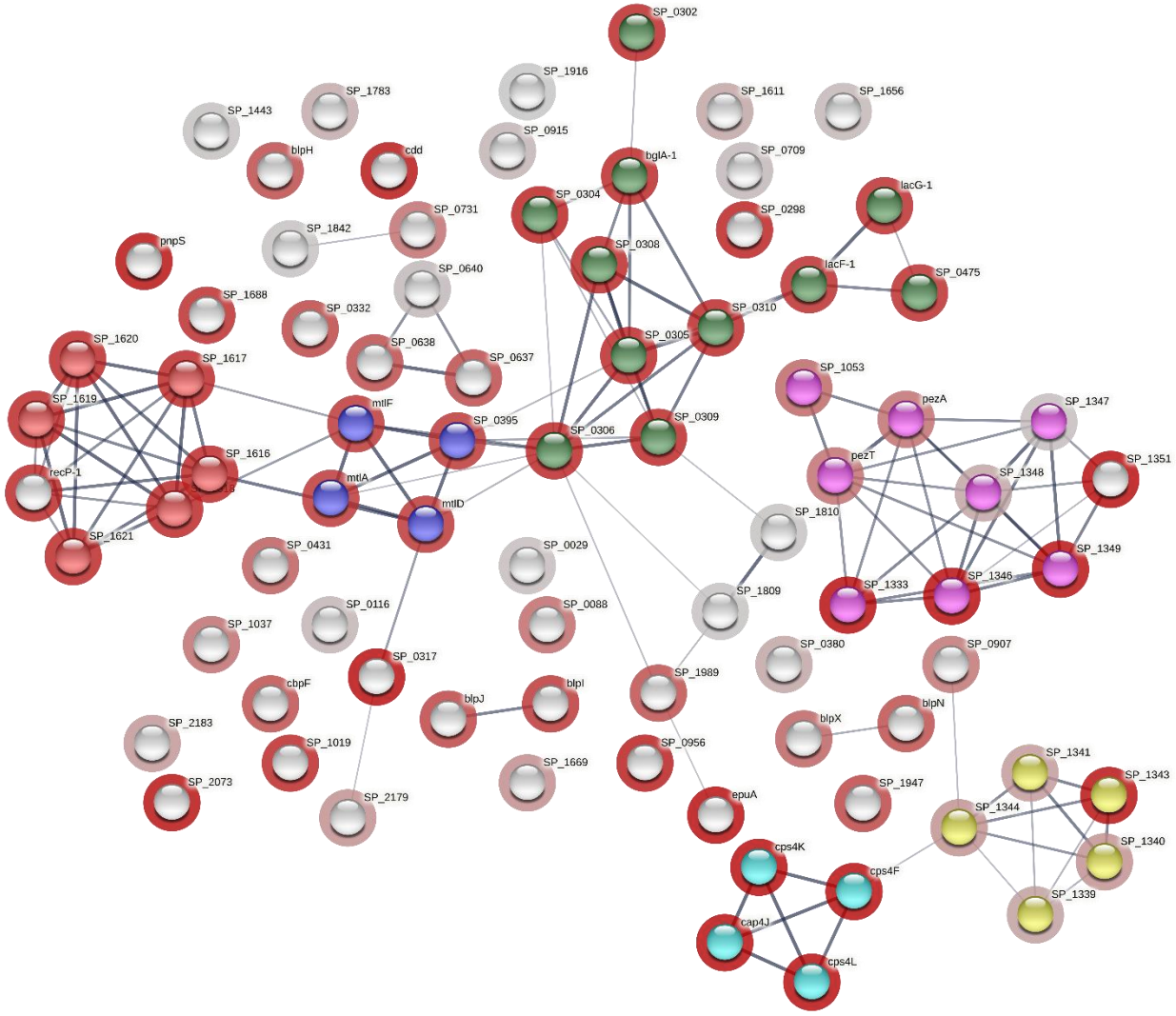


Figure 5-10. The network of interactions between significant genes in serotype 5.
The red halo around the nodes represents gene significance, a more reddish halo represents a higher significance. The yellow cluster includes uncharacterized genes from RD8b2, light blue nodes are genes from RD3, nodes in purple are from RD6 and RD8b3, and red nodes are from RD9. Nodes colored in blue encode the subunits of the mannitol PTS transporter, and those in green are the subunits of cellobiose and lactose PTS transporters.

The network of significant genes in serotype 19F (absent from serotype 5) is shown in Figure 5-11. The pattern of gene loss in serotypes 5 and 12F (compared to serotype 19F) showed similarities and differences. Both serotypes lost genes from:

- RD4, a cluster of enzymes catalyzing the pilus assembly.
- RD7, a group of uncharacterized proteins.
- RD8a, including a V-type sodium ATP synthase, oxidoreductase, neuraminidase, and associated sugar-modifying enzymes.

But in terms of gene absence, the main difference between serotype 5 and 12F was:

- Serotype 12F lost RD10, whereas serotype 5 retained this region, suggesting that RD10 was essential for the functionality of serotype 5. As previously described, RD10 is a pneumococcal pathogenicity island contributing to the biosynthesis and export of pneumococcal serine-rich repeat protein (PsrP).

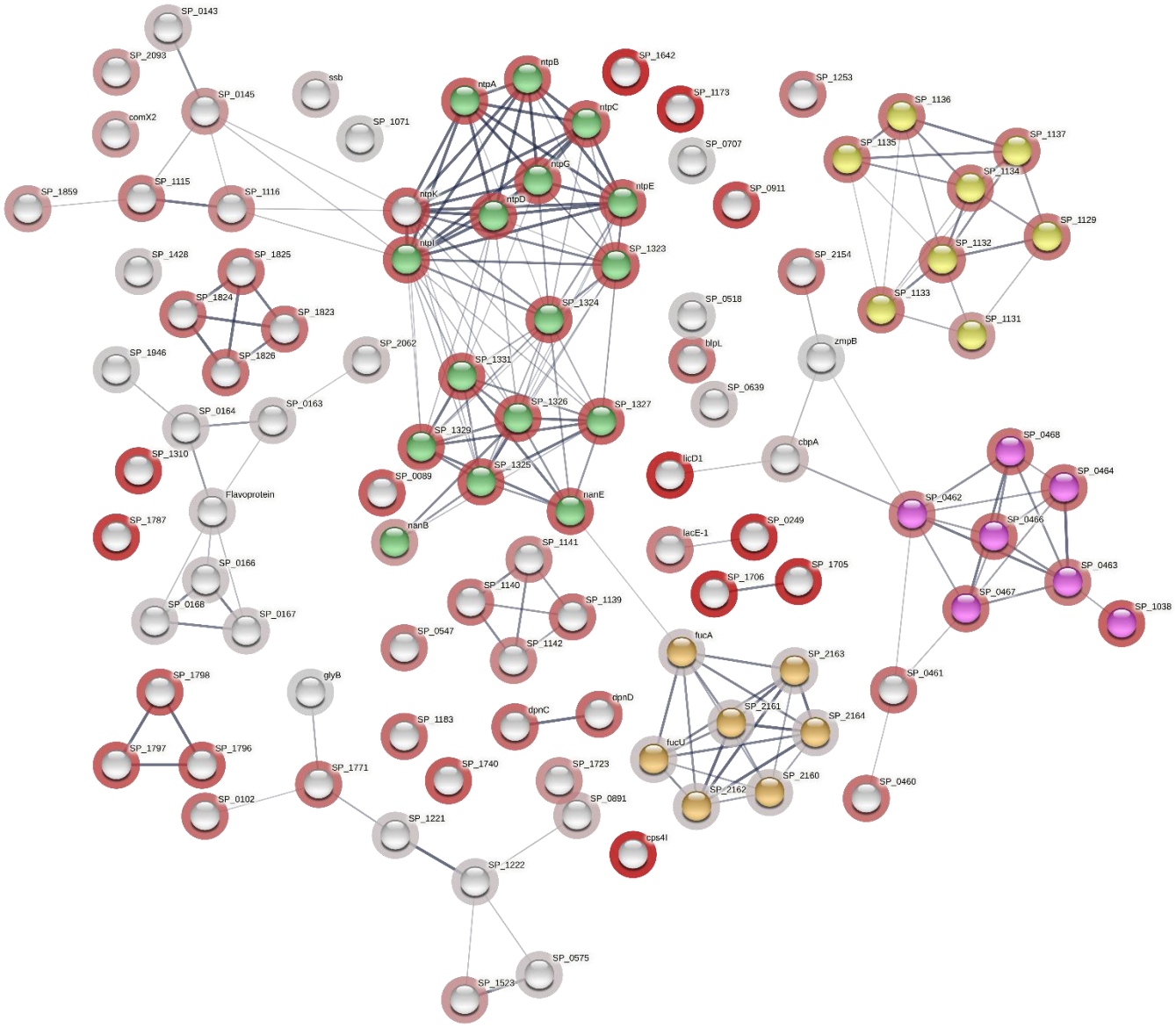


Figure 5-11. The network interactions between significant genes absent from serotype 5 (present in serotype 19F). The red halo around the nodes represents gene significance, a more reddish halo represents a higher significance. The nodes in purple are from RD4, the yellow cluster includes genes from RD7. Green nodes are from RD8a, and nodes in the orange cluster are from RD13.

5.3.2.2.3 Serotype 1 vs. 19F

The gene presence-absence analysis identified 68 significant genes in serotype 1 and 148 significant genes in serotype 19F (Appendix 10). The p-value distributions of the significant genes were illustrated as boxplots in Figure 5-12. The significance levels of gene gain and loss in serotype 1 were similar.

The levels of gene significance in serotypes 1 and 5 were almost identical. The number of genes gained was greater in serotype 5 (82 > 68), while the number of genes lost was larger in serotype 1 (148 > 104). In summary, the number of significant genes in serotype 1 (68 + 148 = 216) was the maximum amongst

the significant invasive serotypes (serotype 5: 82 + 104 = 186, serotype 12F: 82 + 66 = 148). This highlights the highest genetic distinction observed for serotype 1 in the Malawian cohort.

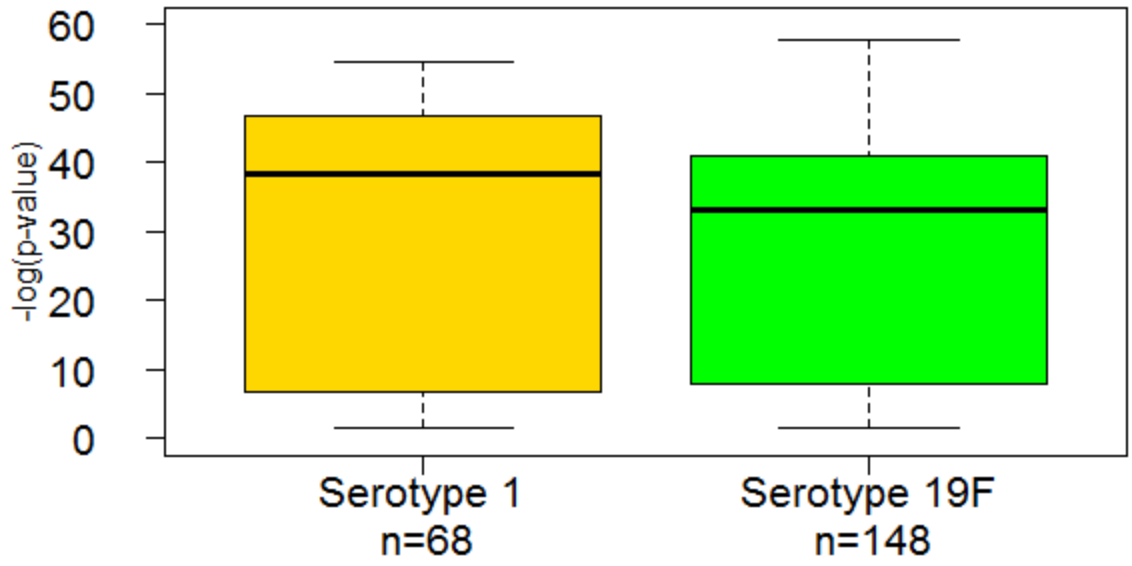


Figure 5-12. The log-transformed p-value distribution of significant genes in serotypes 1 and 19F. There is no outlier above the boxplots.

The network of interactions between significant genes in serotype 1 is shown in Figure 5-13. The most significant annotated genes were:

- SP_0844 (*cdd)*: Cytidine deaminase catalyzing the hydrolysis of cytidine into uridine and ammonia. This gene was also amongst the most significant genes in serotypes 5 and 12F. Cytidine deaminase is a metalloprotein and needs to bind to zinc for its catalytic activity. The biological processes this enzyme is involved in have not been fully characterized.
- Genes from RD6 (PPI1), including Orf9 and Orf10 from the Tn5252 insertion element as well as components of type II toxin-antitoxin (*pezA*-*pezT*).
- Uncharacterized genes from RD8b3, including SP_1347, SP_1348, SP_1349, and SP_1351.
- Genes from RD9 encoding the subunits of the PTS fructose transporter.
- Genes encoding the subunits of cellobiose and lactose transporters.

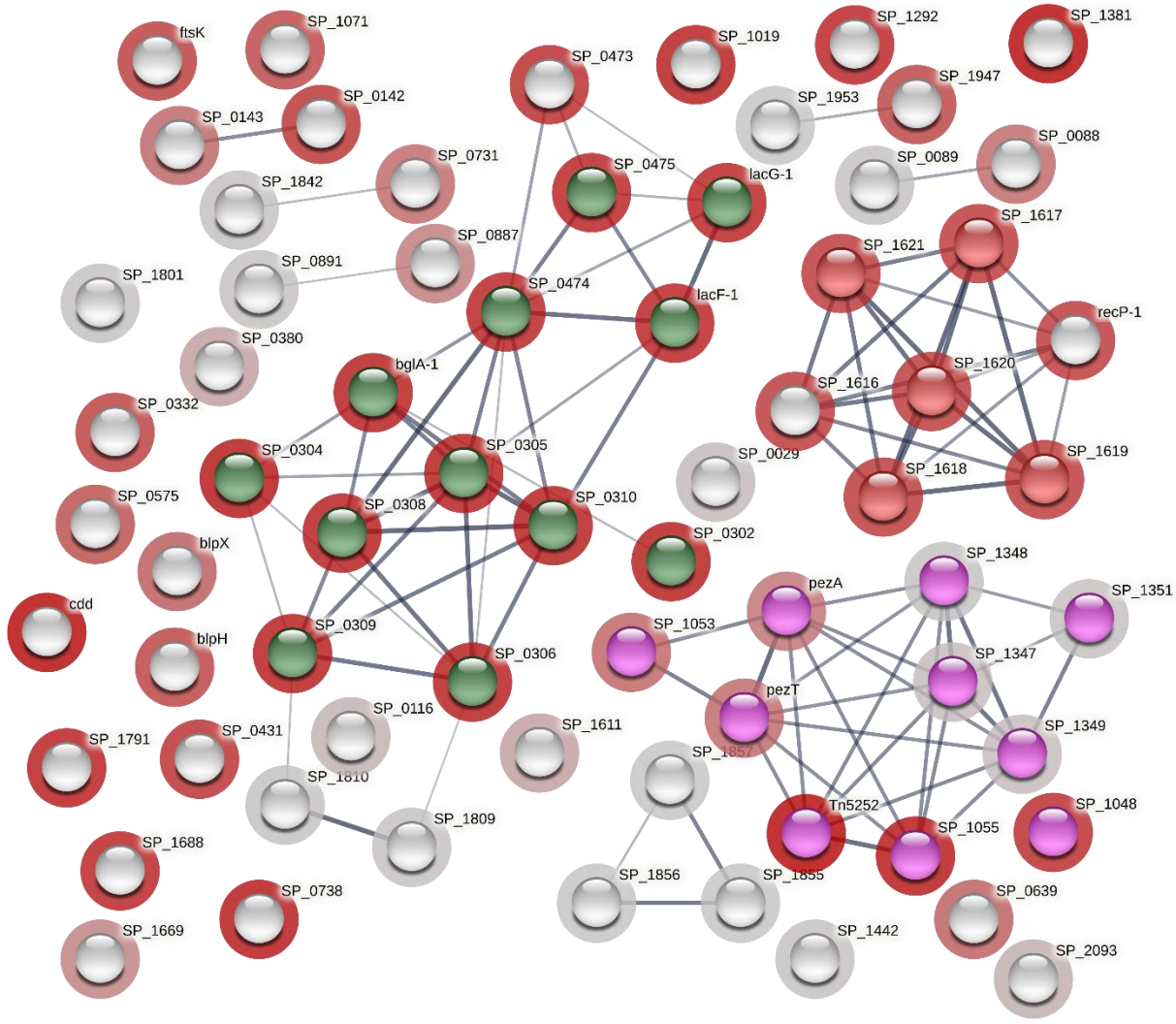


Figure 5-13. The network interactions between significant genes in serotype 1.
The red halo around the nodes represents gene significance, a more reddish halo represents a higher significance. Nodes in purple are from RD6 and RD8b3, those from RD6 are more significant. Nodes colored in red are from RD9, while those in green are the subunits of cellobiose and lactose PTS transporters.

The network of interactions between genes significantly absent from serotype 1 (present in serotype 19F) is depicted in Figure 5-14 and some of these genes are described below:

- Genes from RD10 and RD8a (RD8a1 and RD8a2) were missing from the serotype 1 genome. These two RDs were also significantly absent from the genome of serotype 12F. However, only RD8a was not present in serotype 5, this serotype retained genes from RD10.
- Genes from RD7 and RD4 were significantly missing from the genome of serotype 1. The loss of genes from these two regions was also observed in serotypes 5 and 12F.
- Genes from RD8b2 were significantly absent from serotype 1, whereas they were conserved in serotypes 5 and 12F. RD8b2 is composed of genes that have not been annotated to date.

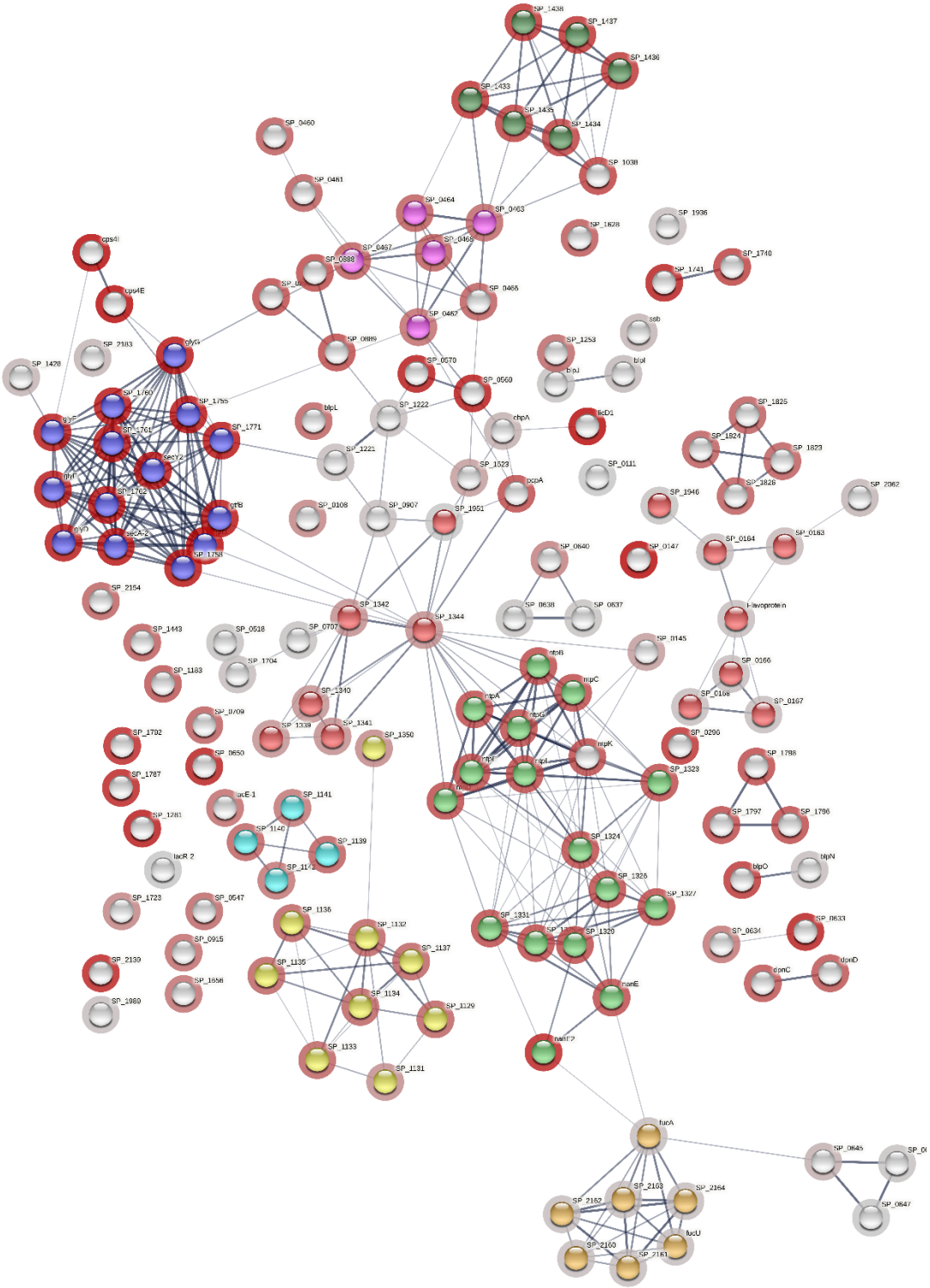


Figure 5-14. The network interactions between significant genes absent from serotype 1 (present in serotype 19F). The red halo around the nodes represents gene significance, a more reddish halo represents a higher significance. The red nodes show genes from RD2 and RD8b2, those in purple are from RD4, the yellow cluster includes genes from RD7, green nodes are from RD8a, nodes in the blue cluster are from RD10, and the orange nodes are from RD13. Nodes in dark green are the components of an uncharacterized ABC transporter.

Since the common characteristic of the significant invasive serotypes was their high invasiveness, It would be important to know what they shared in their gene profiles, including gene gain and loss. The Venn diagrams illustrated in Figure 5-15 showed that 19 significant genes were jointly present, and 50 genes were commonly absent from the genomes of the significant invasive serotypes. The number of gene losses was more remarkable than gene gains.

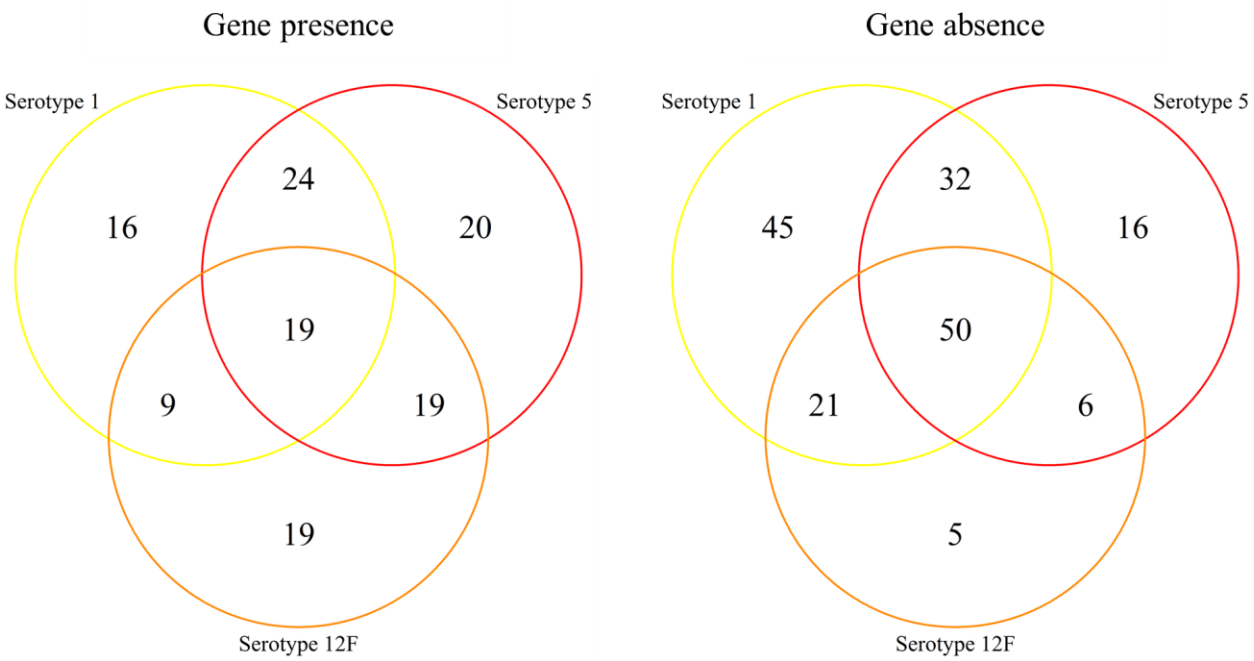


Figure 5-15. Venn diagrams showing the overlap between significant gene sets in serotypes 1, 5, and 12F

The pathways identified by the enrichment analysis of the significant genes jointly present in serotypes 1, 5, and 12F are listed in Table 5-3. Some enriched pathways in Table 5-3 contain multiple significant shared genes. There were two distinct annotated pathways in the table, including:

- Cellobiose, sucrose, and lactose metabolism. These carbohydrates are transported by specific PTS systems encoded by SP_0302, SP_0303, SP_0304, SP_0305, SP_0306, SP_0308, SP_0309, and SP_0310.
- PezA-PezT toxin-antitoxin system involving SP_1050, SP_1051, and SP_1053 from RD6 (PPI1) and SP_1348.

The genes stated above contributed to the significant pathways listed in Table 5-3. In addition to these genes, other annotated significant genes present in serotypes 1, 5, and 12F were:

- SP_0844 (*cdd*): Cytidine deaminase, a metalloprotein with an uncharacterized function.
- Bacteriocins BlpH and BlpX.
- SP_1669: Encoding MutT, which is a house-cleaning enzyme. It must be emphasized that MutT is encoded by paralogs in the pan-genome. For instance, other accessory genes, such as SP_ 0817 and SP_1235, also encode MutT.

The other jointly present genes in the significant invasive serotypes were uncharacterized.

Table 5-3. The enriched pathways in the list of significant genes jointly present in serotypes 1, 5, and 12F.

Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathways (click for details)
2.3E-07	8	44	20.3	mixed, incl. Starch and sucrose metabolism, and lactose catabolic process
7.1E-06	4	6	74.6	mixed, incl. PQ-loop repeat, and Dimeric alpha-beta barrel
1.5E-04	4	12	37.3	mixed, incl. Phosphoenolpyruvate-dependent sugar phosphotransferase system, EIIA 2, and Phosphotransferase system, EIIB component, type 2
3.1E-04	4	15	29.8	Starch and sucrose metabolism
5.4E-04	4	18	24.9	mixed, incl. Toxin-antitoxin system, and CAAX prenyl protease 2
1.5E-03	4	24	18.6	mostly uncharacterized, incl. Toxin-antitoxin system, and CAAX prenyl protease 2
1.5E-03	4	25	17.9	Starch and sucrose metabolism, and lactose catabolic process
2.0E-03	2	2	111.8	Toxin-antitoxin system
3.3E-03	4	32	14	Starch and sucrose metabolism, and lactose catabolic process
3.3E-03	4	33	13.6	mostly uncharacterized, incl. Helix-turn-helix, and CAAX prenyl protease 2
5.3E-03	4	38	11.8	mostly uncharacterized, incl. Helix-turn-helix, and CAAX prenyl protease 2

The pathways enriched in the set of commonly absent genes from the genomes of significant invasive serotypes are listed in Table 5-4. There were overlaps between biological processes in Table 5-4; the distinct pathways were as follows:

- Any pathway with the “Oxidative Phosphorylation” or “Sodium/Solute Symporte” keywords involved genes from RD8a (RD8a1 + RD8a2). The absence of RD8a from the significant invasive serotypes had the highest significance and fold enrichment.
- Genes from RD4 were significantly absent from the genomes of the significant invasive serotypes. They contribute to pathways with the “Sortase” keyword that catalyze the assembly of pilins into pili.
- Pathways involving uncharacterized proteins with CAAX and TOBE domains consisted of genes from RD7.

In summary, RD4, RD7, and RD8a were missing from all significant invasive serotypes. It was worth noting that genes from RD10 were not present in serotypes 1 and 12F whereas this region was retained in serotype 5.

Table 5-4. The enriched pathways in the list of significant genes commonly absent from serotypes 1, 5, and 12F.

Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathways (click for details)
6.4E-25	16	17	40	mixed, incl. Oxidative phosphorylation, and Sodium/solute symporter
1.0E-16	16	36	18.9	mixed, incl. Oxidative phosphorylation, and YhcH/YjgK/YiaL family
4.1E-16	10	10	42.5	Oxidative phosphorylation
1.2E-10	16	81	8.4	mixed, incl. MetI-like superfamily, and Oxidative phosphorylation
1.5E-10	12	38	13.4	mostly uncharacterized, incl. Helix-turn-helix, and CAAX prenyl protease 2
8.7E-09	6	7	36.4	mixed, incl. Sodium/solute symporter, and N-acetylneuraminate lyase
4.5E-08	5	5	42.5	ATPase, alpha/beta subunit, nucleotide-binding domain, active site, and V-type ATPase subunit c
4.5E-08	5	5	42.5	Sortase family, and Fimbrial isopeptide formation D2 domain
1.7E-07	6	10	25.5	Sortase family, and Immunoglobulin-like fold
2.9E-06	8	33	10.3	mostly uncharacterized, incl. Helix-turn-helix, and CAAX prenyl protease 2
6.3E-06	4	5	34	mixed, incl. MgtC family, and Cyclic phosphodiesterase
6.3E-06	4	5	34	Phage integrase, N-terminal SAM-like domain, and Helix-turn-helix
6.3E-06	4	5	34	mixed, incl. Coiled coil
3.1E-05	6	22	11.6	mixed, incl. Sortase family, and Immunoglobulin-like fold
1.0E-04	6	27	9.4	mixed, incl. Sortase family, and Immunoglobulin-like fold
3.0E-04	4	11	15.5	mostly uncharacterized, incl. MgtC family, and Pectin lyase fold
1.8E-03	4	17	10	mixed, incl. TOBE domain, and Bacterial extracellular solute-binding protein

5.3.2.2.4 Serotype 23F vs. others

Serotype 23F is known for its resistance to several antibiotics. Comparing its genome to other serotypes (the grey cluster in Figure 5-1) could determine whether there are any putative resistance genes in the genome of serotype 23F from Malawi. The gene presence-absence analysis identified 119 significant genes present and 85 genes absent from the genome of serotype 23F (Appendix 11). The p-value distributions of significant genes were illustrated as boxplots in Figure 5-16. The significance levels and the number of gene gains were greater than gene losses in serotype 23F (see boxplots in Figure 5-16). The outliers above the blue boxplots in Figure 5-16 were uncharacterized proteins.

In contrast with serotypes 1 and 5, the significance of genes up to the third quartile of the boxplot in serotype 23F (Figure 5-16) was lower, implying its lower genetic distinction compared to serotypes 1 (Figure 5-12) and 5 (Figure 5-9).

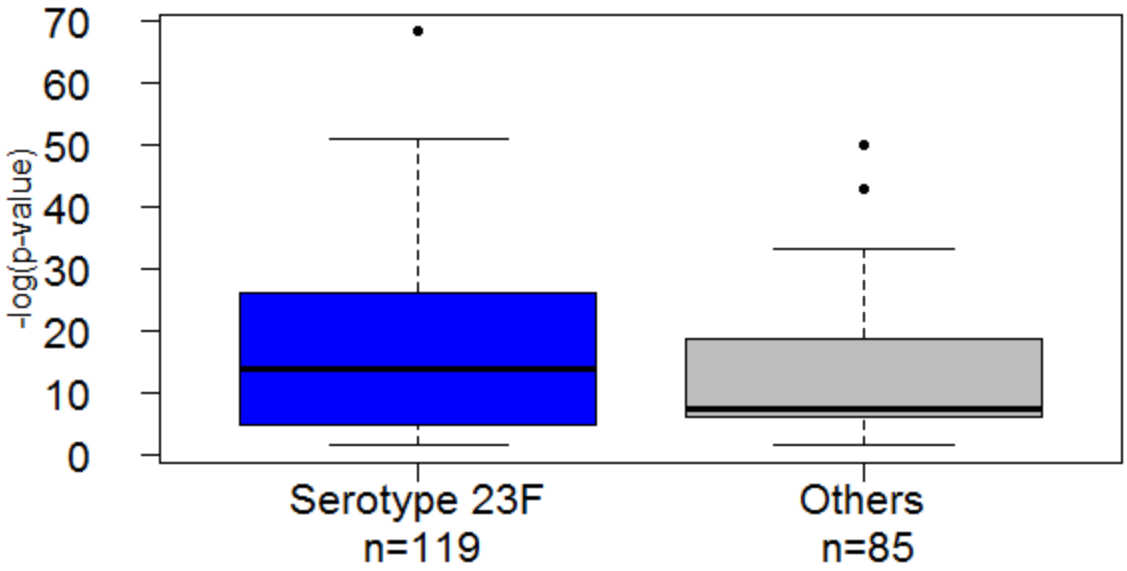


Figure 5-16. The log-transformed p-value distribution of significant genes in serotype 23F. Outliers are shown as black dots above the boxplots.

In serotype 23F, the significantly present genes could include potential antibiotic resistance factors. Also, since this serotype was abundant in both the carriage and patient groups, it could represent an invasive strain that colonizes the nasopharynx longer than serotypes 1 and 5 before infecting the sterile sites. The network of interactions between significant genes in serotype 23F is illustrated in Figure 5-17. As expected, serotype 23F shared similarities with the significant invasive serotypes (1, 5, and 12F) and the non-invasive strain 19F.

As identified in serotypes 1, 5, and 12F:

- The components of the PTS transporters from RD9 and genes involved in the lactose and fructose metabolisms were also significantly present in serotype 23F. These genes were indicated in green and red in Figure 5-17. However, the significance of these genes was lower in serotype 23F than in the significant invasive serotypes (concluded by observing the reddish intensity of halos in Figure 5-10, Figure 5-13, and Figure 5-17 around the green and red nodes).
- Uncharacterized genes from RD8b2 and RD8b3 were also identified in serotype 23F.

The similarity between 23F and 19F included the presence of genes from:

- RD4, assembling pilins into pili and anchoring pili to the cell wall.
- RD7, genes in this region have remained uncharacterized to date.
- RD13, this region contributes to fucose uptake.
- Genes encoding the subunits of the Iron ABC transporter.

In summary, serotype 23F harbored a combination of genes from the significant invasive serotypes and the non-invasive strains such as 19F. An intra-comparison between the serotype 23F samples from the nasopharynx and sterile sites (the patient versus carrier groups within serotype 23F) did not identify any significant genes present (absent) in each group. This means that the presence of genes in serotype 23F

did not depend on the isolation sites. In other words, the gene content of serotype 23F samples obtained from carriers was not different from the genome of serotype 23F samples isolated from the patients.

The distinction in serotype 23F was the presence of particular genes from RD6 that were not present in any of the significant invasive serotypes (1, 5, and12F) or the non-invasive serotype 19F. RD6 (PPI1) is a pathogenicity island that contains 20 genes starting from SP_1046 through SP_1065. Studies showed that the gene content of RD6 has considerable variation among different *S. pneumoniae* strains[259]. The gene presence-absence analysis in this research identified the *PezA-PezT* toxin-antitoxin system from RD6 as being jointly present in serotypes 1, 5, and 12F, including SP_1050, SP_1051, and SP_1053. However, in serotype 23F, other genes from RD6, including SP_1057, SP_1058, SP_1059, SP_1060, SP_1061, SP_1062, and SP_1063 were significantly present. These genes are shown as orange nodes in Figure 5-17, and are absent from serotypes 1, 5, 12F, and 19F. Of these genes, the annotated components were:

- SP_1057, encoding the quorum-sensing regulator PlcR, a transcription factor containing tetratricopeptide repeat regions and is activated upon binding its cognate signaling peptide PapR on a tetratricopeptide repeat-type regulatory domain[188]. PlcR, is a quorum-sensing regulator that controls gene expression according to population density. PlcR functions as a virulence factor in gram-positive bacteria[260].
- SP_1062 and SP_1063, encoding the subunits of an unannotated ABC transporter.

In terms of finding the resistance genes in serotype 23F, the mapping of significant genes in this strain to the Comprehensive Antibiotic Resistance Database (CARD) did not identify any BLAST hits.

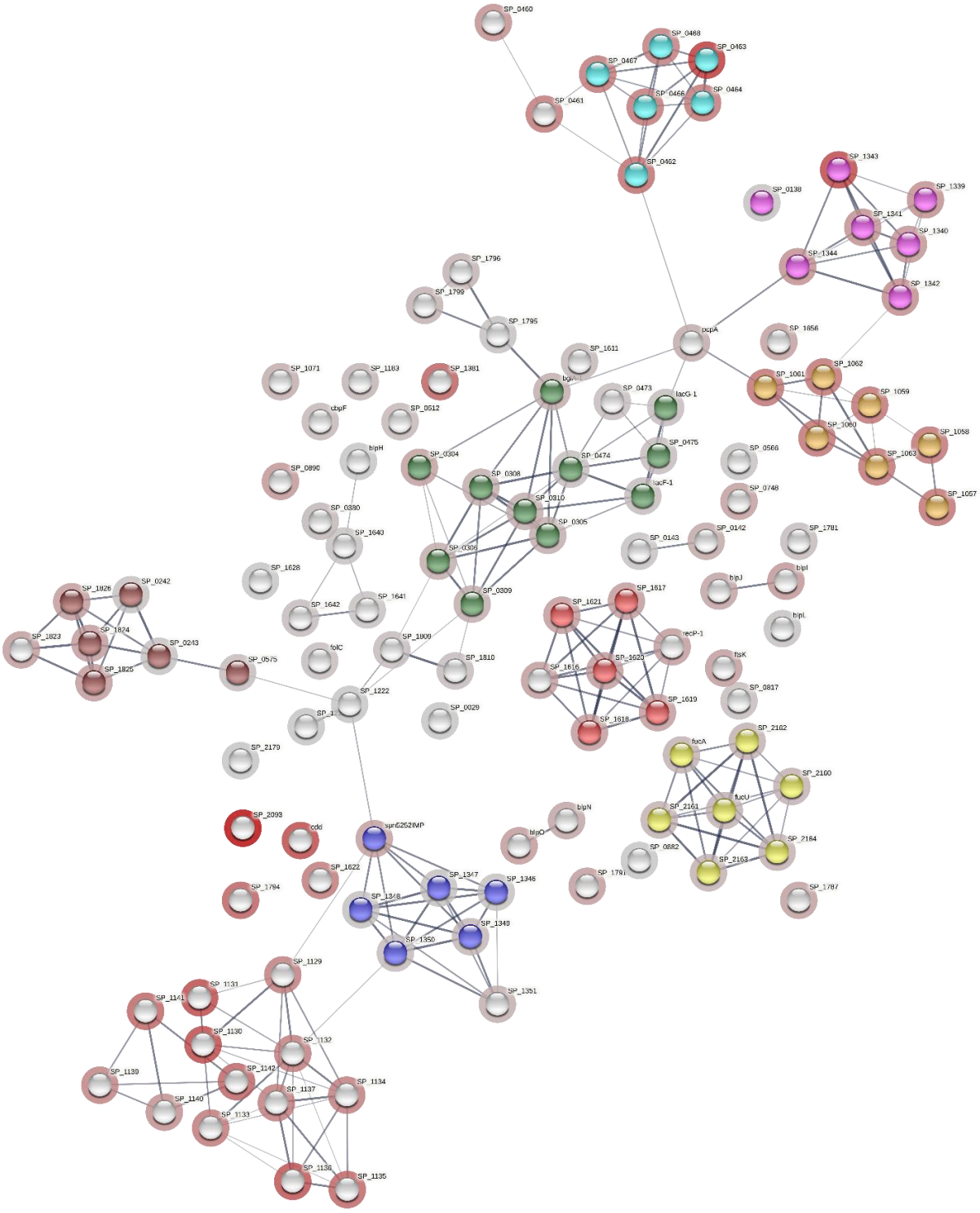


Figure 5-17. The network of significant genes in serotype 23F compared to other serotypes.
The red halo around the nodes represents the significance. A more reddish halo represents a higher significance. The cluster in sky blue contains genes from RD4, genes in orange are from RD6, nodes in the silver cluster on the left bottom of the figure are genes from RD7, lavender nodes are from RD8b2, and dark blue components are from RD8b3. Nodes colored in red are from RD9, and the yellow cluster is composed of genes from RD13. Nodes in the green are the subunits of cellobiose and lactose PTS transporters. Brown nodes encode the subunits of the Iron ABC transporter.

The network of interactions between significant genes absent from serotype 23F is depicted in Figure 5-18. The most significant genes were nodes with the most bluish halo around them, which were five outliers above the grey boxplot in Figure 5-16. Of these five, four were uncharacterized:

1. SP_0449: Uncharacterized protein.
2. SP_0911: Uncharacterized protein.
3. SP_1669: MutT family protein.
4. SP_2089: Uncharacterized protein.
5. SP_0650: Uncharacterized protein.

The pattern of gene losses in serotype 23F was similar to the significant invasive serotypes. As observed in serotypes 1, 5, and 12F, genes from RD8a and RD10 were absent from serotype 23F. However, the significance of the missing genes was lower in serotype 23F than in the significant invasive serotypes.

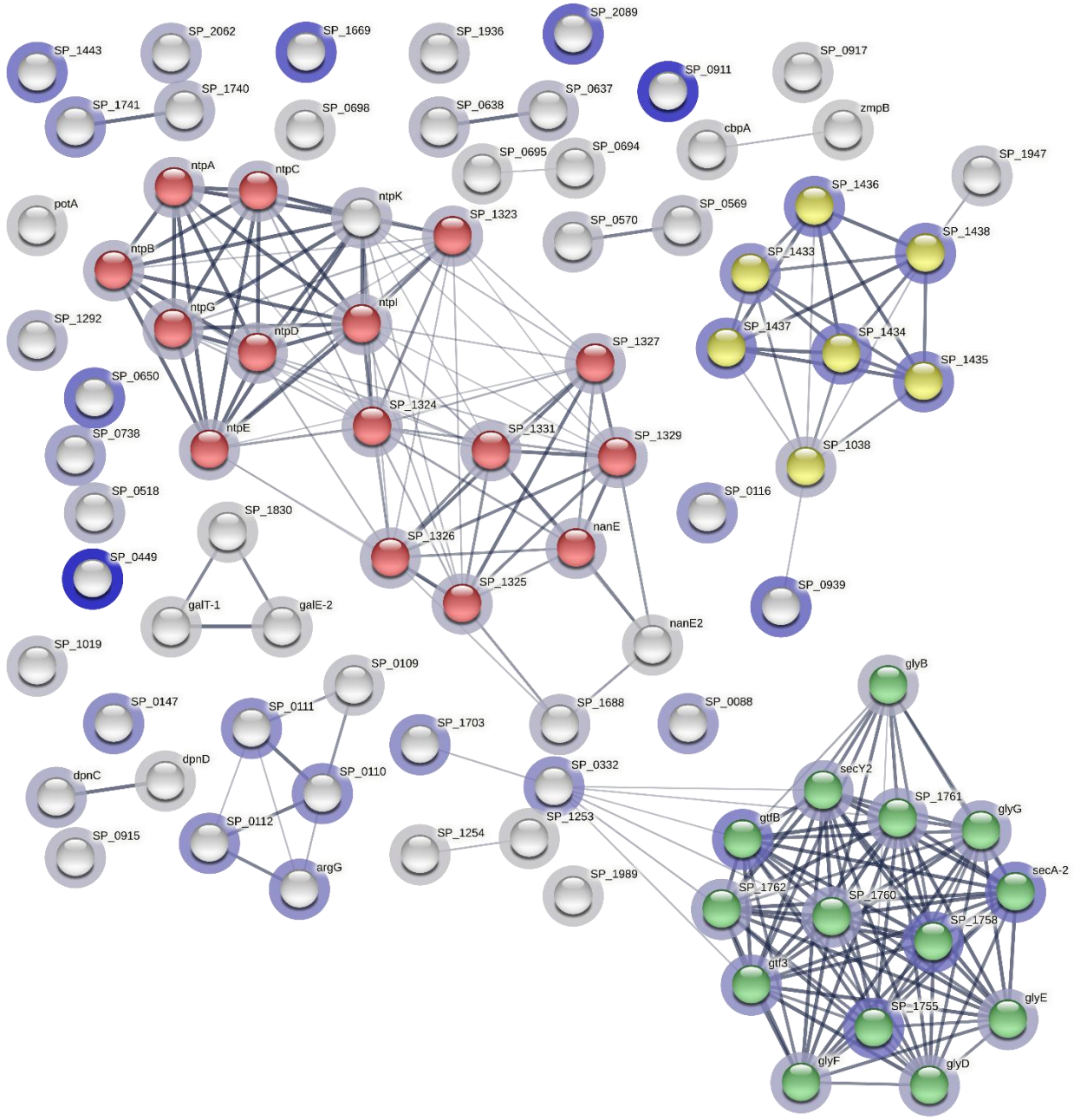


Figure 5-18. The network of significant genes absent from serotype 23F compared to other serotypes.
The blue halo around the nodes represents the significance, a more bluish halo is more significant. The red cluster includes genes from RD8a. The green cluster is composed of genes from RD10. Genes in the yellow cluster encode an ABC transporter.

5.4 Discussion

The analysis in this chapter addressed the central aim of the research to identify genes contributing to pneumococcal pathogenesis in Malawi. It is important to note that the presence of genes is just one of several virulence factors which could be impacted by gene expression and host conditions. It would be useful to have transcriptome data between different serotypes to probe this hypothesis further.

The results in this chapter agreed with the outcome of the GWAS analysis in chapter 4, concluding that the genetic variation in the core and accessory genomes followed the same distribution pattern, suggesting that the isolation sites (nasopharynx, blood, and CSF) could not fully describe pneumococcal genetic diversity in Malawi. This may be due to the fact that the nasopharyngeal population was not homogeneous and most likely contained both non-invasive and invasive samples. It was unknown which sample from the nasopharynx would progress to disease after the collection time. Rather, the genetic diversity in the pan-genome was well explained by serotypes significantly present in the sterile sites, especially serotypes 1 and 5, representing the most abundant and invasive strains in Malawi.

After excluding serotypes 1, 5, 12F, and 19F from the statistical tests, other samples from the nasopharynx and sterile sites did not show much difference in their gene content. Only a few genes with a low level of significance were identified as more prevalent in the sterile sites, including:

- Genes from the *cps* locus in RD3 responsible for encapsulation, including *NAD-dependent epimerase* and *UDP-N-acetylglucosamine 2-epimerase*. Epimerases are involved in synthesizing complex carbohydrates that are essential compounds for the bacterial cell wall. These enzymes are potential therapeutic targets for the treatment of bacterial infection[261].
- Elements of the pneumococcal conjugative transposon Tn5252 from RD6 (PPI1) responsible for transposition.

Genes from RD3 and RD6 were also significantly present in serotypes 1, 5, and 12F, supporting the theory that these regions contribute to the pathogenesis of the most invasive serotypes in the sterile sites, whether these serotypes were significantly present in the blood and CSF or common between carriers and patients.

The population structure was a key factor for the gene presence-absence analysis. The significant invasive serotypes had the greatest distinction in their genome structure. The distributions of the significant invasive serotypes across different isolation sites were as follows:

- Serotype 1 was the most abundant in the entire cohort, with an overall relative frequency of 8.7 %. Despite the high frequency of serotype 1, it comprised less than 1% of the nasopharyngeal population, whereas its relative frequencies in the blood and CSF were 17% and 21%, respectively.
- Serotype 5 was the second most abundant serotype in the entire cohort, with an overall relative frequency of 7.8%. It comprised 2% of nasopharyngeal samples, and its relative frequencies in the blood and CSF were 20% and 8%, respectively.
- The overall frequency of serotype 12F was 2.4%, comprising only 0.6% of the nasopharyngeal population and 2.7% and 7% in blood and CSF, respectively.

In conclusion, serotype 1 was dominant in the blood and CSF though it was more abundant in the CSF, serotype 5 was more frequent in the blood, and serotype 12F had a higher abundance in the CSF. These three serotypes most likely had a low colonization rate which could explain why their frequency was significantly low among carriers, i.e., in the nasopharynx. The population structure analysis highlighted the distinction in their gene content, which could be related to their short period of colonization and urgency to infect the blood and CSF. Any genes that were significantly present (absent) in serotypes 1, 5, and 12F could be candidate virulence factors. However, further experimental investigation is required to evaluate the accuracy of the computational analysis results.

Some of the significant genes were neighbors and work together in operons. The most important finding of the gene loss profiles was the absence of RD8a from the genome of serotypes 1, 5, and 12F. As stated earlier, RD8a is composed of two operons, including RD8a1 (SP_1315-1324) and RD8a2 (SP_1325-1331). The biological function associated with RD8a is ATP synthesis through the oxidative phosphorylation by the *V-Type ATP synthase* (encoded by genes from RD8a1) using energy from the sodium gradient produced by the *Sodium/Solute symporter* (encoded by genes in RD8a2). This suggests that the function of RD8a was either not required for serotypes 1, 5, and 12F, which have a low colonization rate in Malawi, or resulted in this low colonization rate. However, RD8a was conserved in serotype 19F, a non-invasive strain with genetic homogeneity. One possibility is that non-invasive serotypes that stay in the nasopharynx benefit from the presence of RD8a. The functions of genes in this region support this assumption. RD8a harbors genes such as *neuraminidase*, *nanA,* and *nanE* that cleave terminal sialic acid residues from glycoconjugates and mucoglycans available on the surface of the human epithelial cells. This process exposes the receptors on the surface of the epithelial cells and therefore assists pneumococci in penetrating the mucus layer and colonizing the human nasopharynx. It also allows the usage of cleaved sialic acid as a carbon source since free carbohydrates are scarce in the upper respiratory tract[262]. During colonization, secretion of the pneumococcal toxins elevates the level of sodium ions (Na^+) in the nasopharynx[263]. The *Sodium/Solute symporter* encoded by RD8a2 imports extra sodium ions along with a wide variety of substrates into the *pneumococcus* cell[264]. The *ntp* genes in RD8a1 encode the *V-type sodium ATP synthase* that pumps the sodium ions out of the cell[184] and uses the sodium-motive force (SMF) for ATP synthesis[265].

Thus, the level of ATP synthesis was presumably higher in serotype 19F in contrast with the significant invasive serotypes 1, 5, and 12F, which lack RD8a. Although RD8a was present in non-invasive serotypes such as 19F in Malawi, a study in the United States showed that the presence of RD8a was associated with the virulence of serotypes 6B and 14 in this country[199]. RD8a was also conserved in serotype 6B in Malawi, which was an abundant strain in the carrier and patient groups. Since serotype 6B was frequent in both groups, it most likely was able to colonize the nasopharynx longer than serotypes 1, 5, and 12F before infecting the sterile sites. Therefore, the relation between the presence of RD8a and invasiveness probably affects or depends on the period of colonization. We found that invasive serotypes with a short colonization period (1, 5, and 12F) did not retain RD8a in their genomes. However, the presence of RD8a seemed necessary for strains with a high colonization rate, such as 19F, probably because these strains access free oxygen molecules in the upper respiratory tract to perform oxidative phosphorylation by genes encoded in RD8a and produce ATP. RD8a genes also assist pneumococci in importing carbohydrates obtained from the mucus in the nasopharynx.

Despite the similarities between serotypes 1, 5, and 12F, the most important difference between these serotypes was the presence of genes from RD10. This region was absent from the genomes of serotypes

1 and 12F, which were dominant in the CSF. However, RD10 was conserved in serotype 5, which was significantly abundant in the blood, and serotypes 16F and 19F which were significantly present in the nasopharynx.

RD10 has homology to the *gspB-sceA2/Y2* system in *Streptococcus Gordonii* that facilitates the secretion of the *General Secretion pathway Protein B* (GSPB), which is the virulence driver of infective endocarditis[266]. RD10 is a pathogenicity island in the pneumococcus genome with an atypical GC content. The primary function of this island is the correct synthesis and export of *Pneumococcal Serine-Rich Protein* (PSRP), which is the homolog of GSPB in *Streptococcus Gordonii.* RD10 genes transport PSRP to the bacterial cell surface. PSRP promotes the adhesion of serotypes 16F and 19F to epithelial cells in the nasopharynx and serotype 5 to erythrocytes in the blood[267,268]. The presence of the secA2/Y2-like component should also facilitate the export of pneumolysin (PLY) that enhances adhesion to the host cell and survival in the blood[269,270]. Our research showed that RD10 was not present and maybe not required for pneumococcal serotypes 1 and 12F, mostly found in the central nervous system.

In addition to the gene loss patterns in serotypes 1, 5, and 12F (removal of RD8a and RD10), this study highlighted a notable fact about the gene presence profile in these serotypes. The statistical tests and the functional analysis of genes showed the significant presence of the PTS transporters in serotypes 1, 5, and 12F. The primary role of the PTS transporters is taking up carbohydrates from the extracellular environment. Pneumococci can ferment up to 30 types of carbohydrates that are imported mainly by two types of membrane transporters, including ATP-binding cassette (ABC) transporters and phosphotransferase system (PTS) transporters[271].

The major differences between ABC and PTS transporters are:

1. ABC transporters derive energy from ATP molecules, while PTS transporters use phosphoenolpyruvate (PEP) as an energy source.
2. ABC transporters do not modify the imported substrate, but PTS transporters phosphorylate the incoming sugar upon transport.
3. ABC transporters require more energy than PTS transporters, albeit they can transport longer and more complicated carbohydrates[222].

Unlike isolates in the nasopharynx, serotypes 1, 5, and 12F have access to more simple and free host dietary carbohydrates in the blood and the central nervous system. Meanwhile, the absence of RD8a from serotypes 1, 5, and 12F could decrease the level of aerobic ATP synthesis in these serotypes compared to nasopharyngeal strains. PTS transporters use PEP and do not consume ATP for carbohydrate transport across the membrane. Due to a lower ATP synthesis level in serotypes 1, 5, and 12F, they likely prefer to use PTS transporters to uptake sugars such as fructose and lactose. They may benefit from retaining genes that encode PTS transporters to regulate their metabolism, enabling them to survive in the blood and CSF. Conserved significant genes in serotypes 1, 5, and 12F that encode PTS transporters are potential therapeutic targets for future IPD treatment in Malawi. An *in vivo* study would be required to investigate further the biological functions and processes that the PTS genes perform in the cell and their suitability as targets. It must be emphasized that besides sugar uptake, PTS transporters regulate several pathways in bacteria, such as gene expression and communication between cells. Thus, the phenotypic effects of the PTS transporters should not be limited just to their ability to import carbohydrates[272].

The main challenge for the functional analysis of the significant genes was the presence of many uncharacterized proteins in the pan-genome. Almost one-third of significant genes in serotypes 1, 5, and 12F were hypothetical proteins. These genes were predicted open reading frames with no assigned function *in vivo.* Therefore, no further conclusions were drawn on their roles in pneumococcal pathogenesis. These uncharacterized genes could be ideal targets for future functional studies.

6 Conclusion

This work aimed to identify the genetic differences between pneumococcal samples obtained from the nasopharynx of carriers (n = 825) and blood and cerebrospinal fluid (CSF) of patients (n = 652) suffering from bacteremia and meningitis. Samples were collected from three cities in Malawi, including Blantyre, Lilongwe, and Karonga, between 1997 and 2015. The main objective of the research was to determine whether there were genetic differences between non-invasive and invasive pneumococci. Since the pneumococcal genome is highly diverse across different strains, we applied a pangenome approach to include all genetic variations in the analysis. The study included the following steps:

- Since the pneumococcal serotype is the main virulence determinant, the first step was an investigation of the serotype distribution. Some serotypes were expected to be more prevalent in the patient groups. The analysis identified the most invasive and persistent serotypes in Malawi with the highest frequency in the patient group and an increased prevalence after the vaccination program in Malawi in 2011.
- Construction of a pan-genome to determine the core and accessory genes. The advantage of the pan-genome is its ability to cover all genome diversities between samples. Core genes could be considered as putative therapeutic targets since they have been conserved among all samples over a long time (1997-2015) due to their vital roles in cell survival. The accessory-genome includes components that may be significantly present/absent in the invasive samples.
- Analysis of the population structure and identification of genetically distinct strains. This step was the prerequisite to defining traits of samples for the GWAS analysis.
- The distribution of small-scale variants (SNPs and Indels) in the core-genome was explored. The GWAS analysis in this step identified the most significant mutations in the invasive strains.
- The distribution of large-scale variations (gene presence-absence) in the accessory-genome was analyzed. The gene presence-absence analysis identified genes significantly present (or absent) in the invasive serotypes.

To understand the serotype distribution in Malawi, the collection time must be considered, especially with regard to the time of the pneumococcal conjugative vaccine (PCV) rollout in Malawi, which was in November 2011. The dataset could be divided into pre- and post-PCV13 samples. PCV13 involves 13 serotypes, including 1, 3, 4, 5, 6A, 6B, 7F, 9V, 14, 18C, 19A, 19F, and 23F. These serotypes are called vaccine types. The analysis showed that PCV13 significantly reduced the frequency of vaccine type 6B (p-value < 0.01) amongst carriers. In contrast, there was a significant increase in the frequency of non-vaccine types 7C, 11A, 20A, and 28F (p-value < 0.01). These serotypes could represent instances of serotype replacement in the nasopharynx induced by the vaccine exposure. In the patient group, PCV13 significantly reduced the burden of disease caused by vaccine types 5, 6A, 6B, and 19F ($p < 0.01$), however, the incidence of invasive pneumococcal diseases (IPDs) caused by vaccine types 1 and 3 and non-vaccine types 12F, 13, and 38 significantly increased after vaccination ($p < 0.01$).

Studies in different countries showed that the pneumococcal serotype prevalence depends on several factors such as the age of participants, collection times, and geographical locations:

- In 2007, before the routine use of PCV in South Africa, serotypes 1, 3, 14, and 19A were the most frequent in patients of all ages[273].
- In 2010, data from 26 European countries in the post-PCV7 era showed that serotypes 19A, 1, 7F, 3, 14, 22F, 8, 4, 12F, and 19F were the most invasive among children under five years of age and adults older than 65[281].
- In Japan, during the post-PCV7 era from April 2010 to March 2013, there was an increase in the incidence of IPDs caused by serotypes 19A, 15A, 15B, 15C, and 24 among children <18 years of age[274].
- Data from Asian countries, including Korea, China, Malaysia, Singapore, Philippines, and Thailand, showed that from 2012 to 2017, the IPD-causing serotypes in adults above 50 years were 19F, 19A, and 3. Meanwhile, non-vaccine types 11A, 15A, 35B, and 23A have emerged as invasive after vaccination[275].
- According to articles published from 1 January 2000 to 21 August 2019, the most frequent IPD-causing serotypes in Iran were 23F, 19F, 19A, 6A/B, 9V, and 11A[276].
- In Uruguay in 2012 and after PCV13 vaccination, vaccine-type 3 and non-vaccine types 12F, 8, 24F, 22F, 24A, 15C, 9N, 10A, and 33 were the most frequent and invasive among children under five years of age[277].
- In Canada (Ontario), after PCV13 implementation in 2010 and until 2017, a growing incidence of IDPs caused by serotypes 3 and 22F was reported[278].
- In the United Kingdom, in 2020, vaccine types 3 and 19A continue to circulate among children five years after PCV13 introduction[279].

Serotypes 3 and 19F were among the IPD-causing strains in most countries stated above. However, in Malawi:

- Serotype 3 was not abundant, comprising only 2.57% of the entire population (38/1477). Of samples assigned to serotype 3, 63% (24/38) were isolated from the nasopharynx, and 37% (14/38) were from the blood and CSF. This serotype was more prevalent amongst carriers.
- In this research, any serotype with an overall frequency greater than 5% was considered as an abundant strain. Serotype 19F was abundant, with a frequency of 5.5% in the entire population (82/1477). Of samples assigned to serotype 19F, 77% (65/84) were obtained from the nasopharynx, and 23% (17/84) were from sterile sites. Serotype 19F represented a vaccine-type significantly present in the carriage group (p-value < 0.01). The noteworthy fact about serotype 19F was that this strain was not present among patients after 2011. Vaccination may reduce the disease burden caused by 19F, while this strain could still colonize the nasopharynx.

In summary, Serotypes 3 and 19F were not a big challenge in Malawi. Instead, serotypes 1, 5, and 12F were identified as significantly abundant amongst the patient (p-value < 0.01), which supported previous research finding that serotypes 1, 5, and 12F are the primary cause of IPDs in Malawi[146,280]. The difference between serotype distribution in Malawi and other countries may originate from several factors, such as the time of vaccine rollout, race, and immune status of people. The time of vaccine rollout is not the same in different geographical areas. Since vaccination changes the prevalence of pneumococcal strains in circulation, different vaccination times provide scope for different pneumococcal serotypes to dominate or emerge in different regions. The population's race and immune status could also influence the serotype distribution in different countries. The impact of race may be due to genes differentially expressed between ethnic groups that cause the difference in antibody

response[281]. This research was conducted on a black population. The rate of immunosuppressive disorders (such as HIV infection) is also a significant driver of pneumococcal serotype distribution. Therefore, the serotype distribution would not be expected to be similar in different countries. However, consideration of sampling time needs to be also taken into account. All samples in this study were collected before 2015, seven years ago, and the serotype distribution may have changed over this period. A more recent and larger dataset would be useful to investigate the current state of pneumococcal serotypes in Malawi.

Serotypes 1 and 5 had the highest abundance in Malawi, as observed in most African countries. Despite the high frequency of serotypes 1 and 5 in the entire cohort, their prevalence in the carriage group was significantly lower, most likely due to their low nasopharyngeal colonization rate and their speed in infecting the sterile sites. Serotypes 1, 5, and 12F were considered as the significant invasive serotypes with the highest invasiveness (hyper-invasive serotypes or the significant invasive serotypes). However, their relative frequency and temporal distribution were not the same. The relative frequency of serotype 12F (2.37%) was lower than serotype 5 (7.65%) and 1 (8.73%). As mentioned, serotype 1 was the only abundant invasive strain with a persistent dominance in Malawi. There was a significant increase in the frequency of serotype 1 after 2011 (post-PCV13), which could suggest vaccination did not have the impact as would be expected and did not reduce the burden of IPDs caused by serotype 1 in Malawi. Serotype 1 is known as the most prevalent invasive serotype in Africa[282]. A recent study demonstrated the increased pathogenicity of pneumococcal serotype 1 after the introduction of the PCV13 in 2011[283].

In contrast, the burden of disease caused by serotype 5 significantly decreased after the introduction of the PCV13, which could be associated with the vaccine's effect. The prevalence of serotype 12F was low before the vaccination program (0.74%), but its prevalence increased after vaccination (1.6%). Serotype 12F can be considered as an emerging invasive strain in the post-PCV13 era. However, the potential limitations or biases in the sampling process or priority in this project must be considered. Again, it must be emphasized that a larger and more recent dataset is required to more thoroughly investigate the serotype distribution and the effect of PCV13 in Malawi.

Two abundant serotypes 16F and 19F had a significant presence amongst carriers. Serotype 16F is a non-vaccine type with a low frequency in the disease group before and after the vaccine rollout. However, serotype 19F is a vaccine type, 30% of samples assigned to serotype 19F were collected from patients in the pre-PCV13 era, but after PCV13 introduction, serotype 19F was only collected from carriers. Vaccination may protect against 19F efficiently.

Other abundant serotypes in Malawi were 6B and 23F. These two serotypes are vaccine types involved in PCV13, and their remarkable characteristic was their similar frequencies amongst carriers and patients. Vaccination did not change the carriage rate of 6B and 23F and IPDs caused by 23F. The equivalent number of serotypes 6B and 23F in the carriage and disease groups could imply that despite their invasiveness, these two serotypes need to colonize the nasopharynx longer than serotypes 1, 5, and 12F. Therefore, the carriage group could not be considered as a pure non-invasive collection. The presence of several serotypes with similar frequencies in the nasopharynx and sterile sites (e.g., 6B and 23F) could mean that the nasopharyngeal population contained invasive serotypes that were collected during their colonization phase. It was unclear which nasopharyngeal samples progressed to disease after the collection time. Therefore, serotypes such as 16F and 19F with a significant prevalence only

amongst carriers could represent the non-invasive cohort better than the whole nasopharyngeal collection.

A single pneumococcal genome carries about 2200 genes[152]. The pan-genome of 1477 genomes in this study consists of 6809 genes. Core and soft-core genes constitute only 22.77% of the pan-genome reflecting diversity across the pneumococcal genomes. Core genes participate in essential cellular functions such as gene transcription and translation. Analysis of the SNP distribution in the core-genome assisted in finding the most conserved part of the pan-genome. Unsurprisingly, core genes contributing to DNA replication, cell division, and translation were highly conserved. Their roles in the cell were likely very crucial. Therefore, any compounds that can interfere with the functions of these genes could be an ideal case study for drug design against IPD.

On the other hand, most components in the pan-genome were accessory genes (77.23%). Their presence is not always necessary and depends on environmental conditions. The accessory genome codes for enzymes and membrane proteins. The importance of the accessory genes is their potential involvement in specific traits such as virulence and resistance to antibiotics. Indeed, the accessory genome represents regions in the chromosome with a higher level of diversity. Some of these regions are known as regions of diversity that have been characterized and numbered from RD1 to RD13[284].

Following pan-genome construction and identification of the core and accessory components, the population structure was investigated to determine any clusters of samples showing divergence in their genomes and then attempted to link the genetic distinction to the invasiveness. The population structure was determined based on the distribution of SNPs (small-scale variants) in the core-genome (SNP-based phylogeny) and the distribution of genes (large-scale variants) in the accessory-genome (gene-based PCA). Serotypes 1, 5, and 12F appeared as the monophyletic clusters on the phylogenetic tree. The genetic distinction of serotypes 1, 5, and 12F (hyper-invasive serotypes) was higher than other abundant serotypes such as 6B, 19F, and 23F that appeared as multiple clusters on the phylogenetic tree. The gene-based PCA also showed serotypes 1 and 5 clustered distantly from other strains, while a moderate level of separation was observed for serotypes 12F, 19F, and 23F.

Based on the population structure analysis, it was evident that the isolation sites (nasopharynx, blood, and CSF) could not fully explain the genetic plasticity in the population. As stated earlier, several abundant serotypes, such as 6B and 23F were common between the nasopharynx, blood, and CSF with a similar frequency. These strains are known for their invasiveness and are involved in PCV13. Their presence in the nasopharynx does not mean that they are non-invasive. They might have been collected during colonization before they entered the blood and CSF and caused disease. Therefore, the nasopharyngeal collection did not separate from blood and CSF, most likely due to the presence of invasive serotypes in the nasopharynx. Another reason to justify the genetic similarity between carriage and disease pneumococci is the possibility of gene expression changes rather than the genome structure. Common abundant serotypes in the carriage and disease groups, such as 6B and 23F, may change their gene expression instead of their genome structure when they enter the blood and CSF. A gene expression analysis would assist in investigating this assumption.

In summary, the isolation sites and vaccine rollout were not the main drivers of the genetic diversity in the pan-genome. Indeed, the serotype of samples was the strongest determinant of the population structure. The hyper-invasive serotypes showed the maximum core and accessory distinction from other strains. Since these serotypes were significantly present in the disease group, their genetic divergence

may be associated with the disease. Nonetheless, the analysis must rigorously account for population structure and carefully curate any hits to show statistical association with the disease across the population. The near-perfect separation of serotypes 1 and 5 may be associated with their significantly low abundance in the carriage group. Therefore, these patterns may primarily reflect biases in sampling rather than true genetic phenomena. To address the issue, all vaccine types were downsampled, and the population structure analysis was repeated. The downsampling included ten samples randomly selected from the nasopharynx, blood, and CSF for each vaccine type. The PCA of gene distribution identified the same stratification pattern so that each hyper-invasive serotype formed a single distinct group while other serotypes clustered together. Again, serotypes 1 and 5 were grouped distantly from other strains.

Typically, serotypes are separated from each other since they carry different serotype-defining capsule genes. The issue that must be addressed here is why hyper-invasive serotypes cluster more distinctly in contrast with other serotypes. One hypothesis could be that the highest distinction of the hyper-invasive serotypes is associated with their short colonization period and significant presence in the disease group.

The GWAS analysis explored the distribution of small-scale variants in the core-genome (SNPs and Indels) and large-scale variants in the accessory-genome. Comparing the nasopharynx and sterile sites would identify the significant variants that distinguish carriage and patient groups in a serotype-independent manner. However, the population structure suggested that the most significant variants belong to the hyper-invasive serotypes. Of note was that the significant presence of the hyper-invasive serotypes in the blood and CSF would skew the results to explain the difference between the carriage and patient groups. Therefore, the statistical tests were applied across the isolation sites (location-based GWAS analysis) and serotypes (serotype-based GWAS analysis) as follows:

- For the location-based GWAS analysis, the hyper-invasive serotypes were excluded, and the analysis tested samples from the nasopharynx against those obtained from the sterile sites.
- For the serotype-based GWAS analysis, the hyper-invasive serotypes were tested against strains with the lowest invasiveness (16F and 19F). This analysis may address the distinction of the hyper-invasive serotypes in association with their invasiveness.

The location-based GWAS of the small-scale variants identified a significant missense SNP in gene SP_0375 (*gnd*), which encodes the enzyme 6-phosphogluconate dehydrogenase that functions in the center of the pentose phosphate pathway. Mutation in this enzyme can potentially influence several metabolic pathways. Experimental work is required to determine the consequence of the *in-silico* results and characterize how a mutation in *gnd* affects the metabolic pathways.

The location-based GWAS of the large-scale variants (gene presence-absence) identified immunity proteins such as SP_0536 (*blpL*) significantly present in the nasopharyngeal isolates. The significant presence of the immunity protein in the nasopharynx may be associated with the competence of bacteria in the respiratory tract. On the other hand, SP_1055 and SP_1056 from RD6 were significantly present in the blood and CSF. RD6 is a pneumococcal pathogenicity island with genes known for infection involvement.

The serotype-based GWAS of the small-scale variants identified hundreds of significant missense SNPs in the hyper-invasive serotypes (as predicted by the population structure). Mutated genes in serotypes 1,

5, and 12F significantly contribute to DNA and RNA metabolism. A lot of significant missense SNPs were detected in genes that catalyze the attachment of amino acids to tRNA molecules. The common missense SNPs between serotypes 1, 5, and 12F were in genes that encode DNA helicase, DNA methyltransferase, exodeoxyribonuclease, tRNA ribosyltransferase, and nucleotidyltransferase.

The serotype-based GWAS of the large-scale variants found several significant genes located next to each other and working as operons in the hyper-invasive serotypes. These genes could explain the difference between the biology of the significant invasive serotypes and non-invasive strains (e.g., 16F and 19F). The thesis has described the biological processes these genes carry out in the cell. Some key points are highlighted below.

Regarding gene absence profiles, serotypes 1, 5, and 12F did not bear genes from RD8a encoding the subunit of the *V-type ATP synthases* that perform aerobic ATP synthesis through oxidative phosphorylation. Genes in RD8a also promote carbohydrate metabolism and the binding of pneumococci to the epithelial cells lining the human upper respiratory tract. This finding strengthened the assumption about the short colonization period of serotypes 1, 5, and 12F before infecting the blood and CSF since oxidative phosphorylation requires free oxygen molecules that are more available in the upper respiratory tract. However, it must be emphasized that the *in-silico* results must be confirmed *in vivo* through experimental validation.

In terms of gene presence profiles, a common characteristic of the significant invasive serotypes was the conservation of genes that encode the subunits of the PTS cellobiose, fructose, and lactose transporters. Serotypes 1, 5, and 12F spend more time in the sterile sites (blood and CSF) and may benefit from the PTS transporters for carbohydrate metabolism because:

- In the absence of RD8a in serotypes 1, 5, and 12F, the level of ATP synthesis through oxidative phosphorylation may decrease. PTS transporters use phosphoenolpyruvate as the source of energy instead of ATP to transport compounds across the membrane. Therefore, serotypes 1, 5, and 12F could benefit from PTS transporters without RD8a and a possible shortage of ATP.
- The significant PTS transporters in serotypes 1, 5, and 12F import dietary carbohydrates such as fructose and lactose that are more frequent in the blood and CSF than in the nasopharynx. Thus, the role of PTS transporters in importing these kinds of sugars may not be beneficial for non-invasive strains such as 16F and 19F.

The genome structure of serotypes 1, 5, and 12F was not entirely similar. The main difference between these strains was the presence of RD10, which was absent from the genome of serotypes 1 and 12F but fully conserved in serotypes 5, 16F, and 19F. Genes in RD10 are involved in the secretion of the pneumococcal serine-rich protein that enhances the binding of the pneumococci to the epithelial cells in the nasopharynx and erythrocytes in the blood. Therefore, abundant serotypes in the nasopharynx (serotypes 16F and 19F) and blood (serotype 5) may benefit from RD10, whereas this region may not be beneficial for serotypes 1 and 12F that dominate the central nervous system.

The finding about serotypes 16F and 19F suggest that the gene pools of these strains from the nasopharynx, blood, and CSF were the same. Their gene content also did not change after vaccination. Serotypes 16F and 19F, regardless of their locations in the human body (nasopharynx, blood, and CSF) and their collection time (pre- and post-PCV13), conserved RD8a. As stated earlier, the function of genes in RD8a can promote nasopharyngeal colonization. This could mean that, unlike serotypes 1, 5, and 12F,

serotypes 16F and 19F need to colonize the nasopharynx for a longer period of time (as their abundance in the nasopharynx suggests). The presence of genes from RD8a was the most significant difference between the least (16F and 19F) and most (1, 5, and 12F) invasive pneumococcal strains in Malawi. This finding could also be considered as the most important novel finding of the research.

In addition to genes from RD8a and RD10, other significant genes identified by the gene presence-absence analysis were as follows:

- Genes in RD7 that catalyze the assembly of pilins into pili. The presence of pili on the surface of the pathogen promotes its attachment to epithelial cells in the upper respiratory tract. RD7 was significantly absent from serotypes 1, 5, and 12F.
- The type of genes that encode immunity proteins was different in the nasopharyngeal and invasive isolates. The components of the *pezA-pezT* system from RD6 (PPI1) were significantly present in samples from blood and CSF (patient group), whereas the immunity protein *blpI* was significantly present in the nasopharyngeal samples.

In this research, the serotype distribution and genetic differences between pneumococcal strains were studied. The main objective of the work was to address any association between the genome structure of samples and their pathogenesis. The research identified serotypes 1, 5, and 12F with the highest invasiveness and the lowest colonization rate. Serotype 1 was the most persistent invasive strain in Malawi (until 2015). The study's novelty described the high genetic distinction of serotypes 1, 5, and 12F that was not observed in other strains in Malawi. The hypothesis was that the divergence in their genomes was likely associated with their short colonization period and high invasiveness. Their genetic distinction involved many SNPs and genes in their genome structure. Significantly, the absence of RD8a in these serotypes was noteworthy. This, along with the absence of RD7, may be the cause of the short colonization period in the nasopharynx. The vital fact to be considered is that the presence of SNPs or genes is probably essential for virulence but not necessarily sufficient. It must also be determined whether SNPs affect the function of proteins or whether the significant genes are expressed or not.

7 Future work

With the information gained from this research, future work could include the experimental verification of results and the analysis of other possible virulence factors such as the expression of genes. Core genes identified in this research can represent drug target candidates. They perform vital roles in the cell, however, these genes may be also conserved in other members of the human microbiota. A microbiome analysis could address which core genes are only conserved in *S. pneumoniae to* be targeted by the new drugs.

Genes in RD8a and RD10 that are specifically absent and present in the invasive serotypes 1, 5, and 12F can be investigated in future research. It is important to know how the deletion of RD8a could affect the colonization capacity of nasopharyngeal strains and how the pathogenesis changes if RD10 is eliminated from invasive serotypes.

Apart from the gene presence-absence, a transcriptomic study will help identify the differentially expressed genes under different conditions. The gene expression analysis may determine why isolates from the same serotype are present in both carriage and patient groups. Other potential factors involved in developing IPDs are host-related such as race, age, smoking, and immune system. We know that infants, the elderly, and immunocompromised patients are more vulnerable to developing IPDs. Some pneumococcal strains that are harmless to a healthy person might be able to cause severe symptoms in people at high risk. In summary, we recommend careful consideration of multiple factors in future work to fully understand the biology and virulence of *Streptococcus pneumoniae,* including possibly genotyping the hosts. A proper sampling method is also essential to ensure complete and unbiased results.

The outcome of the research was summarized as a research article and submitted to a peer-review journal.

References

1. Watson, D. A., Musher, D. M., Jacobson, J. W. & Verhoef, J. A Brief History of the Pneumococcus in Biomedical Research: A Panoply of Scientific Discovery. *Clin. Infect. Dis.* (1993). doi:10.1093/clinids/17.5.913

2. Avery, O. T. & Macleod, C. M. Studies on the Chemical Inducing Nature Types of the Substance Transformation. *J. Exp. Med.* (1944).

3. Tuomanen, E. I. Microbiology and pathogenesis of Streptococcus pneumoniae. at <https://www.uptodate.com/contents/microbiology-and-pathogenesis-of-streptococcus-pneumoniae#H11090376.>

4. Le Polain De Waroux, O., Flasche, S., Prieto-Merino, D. & Edmunds, W. J. Age-dependent prevalence of nasopharyngeal carriage of streptococcus pneumoniae before conjugate vaccine introduction: A prediction model based on a meta-analysis. *PLoS One* **9,** (2014).

5. Backhaus, E., Berg, S., Andersson, R., Ockborn, G., Malmström, P., Dahl, M., Nasic, S. & Trollfors, B. Epidemiology of invasive pneumococcal infections: Manifestations, incidence and case fatality rate correlated to age, gender and risk factors. *BMC Infect. Dis.* (2016). doi:10.1186/s12879-016-1648-2

6. Bogaert, D., De Groot, R. & Hermans, P. W. M. Streptococcus pneumoniae colonisation: The key to pneumococcal disease. *Lancet Infect. Dis.* (2004). doi:10.1016/S1473-3099(04)00938-7

7. Dowson, C. G., Coffey, T. J. & Spratt, B. G. Origin and molecular epidemiologY of penicillin-binding-protein-mediated resistance to β-lactam antibiotics. *Trends Microbiol.* (1994). doi:10.1016/0966-842X(94)90612-2

8. Drijkoningen, J. J. C. & Rohde, G. G. U. Pneumococcal infection in adults: Burden of disease. *Clin. Microbiol. Infect.* (2014). doi:10.1111/1469-0691.12461

9. Kadioglu, A., Weiser, J. N., Paton, J. C. & Andrew, P. W. The role of Streptococcus pneumoniae virulence factors in host respiratory colonization and disease. *Nat. Rev. Microbiol.* (2008). doi:10.1038/nrmicro1871

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

11. Sorensen, U. B. S., Blom, J., Birch-Andersen, A. & Henrichsen, J. Ultrastructural localization of capsules, cell wall polysaccharide, cell wall proteins, and F antigen in pneumococci. *Infect. Immun.* (1988).

12. Geno, K. A., Gilbert, G. L., Song, J. Y., Skovsted, I. C., Klugman, K. P., Jones, C., Konradsen, H. B. & Nahm, M. H. Pneumococcal capsules and their types: Past, present, and future. *Clin. Microbiol. Rev.* (2015). doi:10.1128/CMR.00024-15

13. Sørensen, U. B. S., Henrichsen, J., Chen, H. C. & Szu, S. C. Covalent linkage between the capsular polysaccharide and the cell wall peptidoglycan of Streptococcus pneumoniae revealed by immunochemical methods. *Microb. Pathog.* (1990). doi:10.1016/0882-4010(90)90091-4

14. Ben-Shimol, S., Givon-Lavi, N., Kotler, L., van der Beek, B. A., Greenberg, D. & Dagan, R. Post-13-valent pneumococcal conjugate vaccine dynamics in young children of serotypes included in candidate extended-spectrum conjugate vaccines. *Emerg. Infect. Dis.* **27,** 150–160 (2021).

15. Brady, A. M., Calix, J. J., Yu, J., Geno, K. A., Cutter, G. R. & Nahm, M. H. Low invasiveness of pneumococcal serotype 11A is linked to ficolin-2 recognition of O-acetylated capsule epitopes and lectin complement pathway activation. *J. Infect. Dis.* (2014). doi:10.1093/infdis/jiu195

16. Lee, S., Bae, S., Lee, K. J., Yu, J. Y. & Kang, Y. Changes in serotype prevalence and antimicrobial resistance among invasive Streptococcus pneumoniae isolates in Korea, 1996-2008. *J. Med. Microbiol.* (2013).

doi:10.1099/jmm.0.058164-0

17. Harboe, Z. B., Thomsen, R. W., Riis, A., Valentiner-Branth, P., Christensen, J. J., Lambertsen, L., Krogfelt, K. A., Konradsen, H. B. & Benfield, T. L. Pneumococcal serotypes and mortality following invasive pneumococcal disease: A population-based cohort study. *PLoS Med.* (2009). doi:10.1371/journal.pmed.1000081

18. Austrian, R. Some observations on the pneumococcus and on the current status of pneumococcal disease and its prevention. *Rev. Infect. Dis.* (1981). doi:10.1093/clinids/3.Supplement_1.S1

19. Fernebro, J., Andersson, I., Sublett, J., Morfeldt, E., Novak, R., Tuomanen, E., Normark, S. & Normark, B. H. Capsular Expression in Streptococcus pneumoniae Negatively Affects Spontaneous and Antibiotic-Induced Lysis and Contributes to Antibiotic Tolerance . *J. Infect. Dis.* (2004). doi:10.1086/380564

20. Insel, R. A. & Anderson, P. W. J. Cross-reactivity with Escherichia coli K100 in the human serum anticapsular antibody response to Haemophilus influenzae type B. *J. Immunol.* (1982).

21. Robbins, J. B., Austrian, R., Lee, C. J., Rastogi, S. C., Schiffman, G., Henrichsen, J., Makela, P. H., Broome, C. V., Facklam, R. R., Tiesjema, R. H. & Parke, J. C. Considerations for formulating the second-generation pneumococcal capsular polysaccharide vaccine with emphasis on the cross-reactive types within groups. *J. Infect. Dis.* (1983). doi:10.1093/infdis/148.6.1136

22. Brooks, L. R. K. & Mias, G. I. Streptococcus pneumoniae's virulence and host immunity: Aging, diagnostics, and prevention. *Front. Immunol.* (2018). doi:10.3389/fimmu.2018.01366

23. Talbot, U. M., Paton, A. W. & Paton, J. C. Uptake of Streptococcus pneumoniae by respiratory epithelial cells. *Infect. Immun.* (1996).

24. van der Windt, D., Bootsma, H. J., Burghout, P., Van der Gaast-de Jongh, C. E., Hermans, P. W. M. & Van Der Flier, M. Nonencapsulated Streptococcus pneumoniae resists extracellular human neutrophil elastase- and cathepsin G-mediated killing. *FEMS Immunol. Med. Microbiol.* (2012). doi:10.1111/j.1574-695X.2012.01028.x

25. Lee, C. J., Banks, S. D. & Li, J. P. Virulence, immunity, and vaccine related to streptococcus pneumoniae. *Crit. Rev. Microbiol.* (1991). doi:10.3109/10408419109113510

26. Nelson, A. L., Roche, A. M., Gould, J. M., Chim, K., Ratner, A. J. & Weiser, J. N. Capsule enhances pneumococcal colonization by limiting mucus-mediated clearance. *Infect. Immun.* (2007). doi:10.1128/IAI.01475-06

27. McCollum, E. D., Nambiar, B., Deula, R., Zadutsa, B., Bondo, A., King, C., Beard, J., Liyaya, H., Mankhambo, L., Lazzerini, M., Makwenda, C., Masache, G., Bar-Zeev, N., Kazembe, P. N., Mwansambo, C., Lufesi, N., Costello, A., Armstrong, B. & Colbourn, T. Impact of the 13-valent pneumococcal conjugate vaccine on clinical and hypoxemic childhood pneumonia over three years in central Malawi: An observational study. *PLoS One* (2017). doi:10.1371/journal.pone.0168209

28. Malawi:Analytical summary - Immunization and vaccines development - AHO. at <http://www.aho.afro.who.int/profiles_information/index.php/Malawi:Analytical_summary_-_Immunization_and_vaccines_development#cite_note-twenty-four-1>

29. Weinberger, D. M., Malley, R. & Lipsitch, M. Serotype replacement in disease after pneumococcal vaccination. *Lancet* (2011). doi:10.5455/apd.239006

30. Feikin, D. R., Kagucia, E. W., Loo, J. D., Link-Gelles, R., Puhan, M. A., Cherian, T., Levine, O. S., Whitney, C. G., O'Brien, K. L., Moore, M. R., Feikin, D. R., Link-Gelles, R., Cherian, T., Adegbola, R. A., Agocs, M., Ampofo, K., Andrews, N., Barton, T., Benito, J., Broome, C. V., Bruce, M. G., Bulkow, L. R., Byington, C. L., Camou, T., Cook, H., Cotter, S., Dagan, R., de Wals, P., Deceuninck, G., Denham, B., Edwards, G., Eskola, J.,

Fitzgerald, M., Galanakis, E., Garcia-Gabarrot, G., Garcia-Garcia, J. J., Gene, A., Gomez, B., Heffernan, H., Hennessy, T. W., Heuberger, S., Hilty, M., Ingels, H., Jayasinghe, S., Kagucia, E. W., Kellner, J. D., Klein, N. P., Kormann-Klement, A., Kozakova, J., Krause, V., Kriz, P., Lambertsen, L., Lepoutre, A., Levine, O. S., Lipsitch, M., Loo, J. D., Lopez-Vega, M., Lovgren, M., Maraki, S., Mason, E. O., McIntyre, P. B., Menzies, R., Messina, A., Miller, E., Mintegi, S., Moore, M. R., Motlova, J., Moulton, L. H., Mühlemann, K., Muñoz-Almagro, C., O'Brien, K. L., Murdoch, D. R., Park, D. E., Puhan, M. A., Reingold, A. L., Sa-Leao, R., Sanyal, A., Smith, P. G., Spanjaard, L., Techasaensiri, C., Thompson, R. E., Thoon, K. C., Tyrrell, G. J., Valentiner-Branth, P., van der Ende, A., Vanderkooi, O. G., van der Linden, M. P. G., Varon, E., Verhaegen, J., Vestrheim, D. F., Vickers, I., von Gottberg, A., von Kries, R., Waight, P., Weatherholtz, R., Weiss, S., Whitney, C. G., Yee, A. & Zaidi, A. K. M. Serotype-Specific Changes in Invasive Pneumococcal Disease after Pneumococcal Conjugate Vaccine Introduction: A Pooled Analysis of Multiple Surveillance Sites. *PLoS Med.* (2013). doi:10.1371/journal.pmed.1001517

31. Heinsbroek, E., Tafatatha, T., Phiri, A., Swarthout, T. D., Alaerts, M., Crampin, A. C., Chisambo, C., Mwiba, O., Read, J. M. & French, N. Pneumococcal carriage in households in Karonga District, Malawi, before and after introduction of 13-valent pneumococcal conjugate vaccination. *Vaccine* (2018). doi:10.1016/j.vaccine.2018.10.021

32. Bentley, S. D., Aanensen, D. M., Mavroidi, A., Saunders, D., Rabbinowitsch, E., Collins, M., Donohoe, K., Harris, D., Murphy, L., Quail, M. A., Samuel, G., Skovsted, I. C., Kaltoft, M. S., Barrell, B., Reeves, P. R., Parkhill, J. & Spratt, B. G. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet.* (2006). doi:10.1371/journal.pgen.0020031

33. Morona, J. K., Morona, R., Miller, D. C. & Paton, J. C. Mutational analysis of the carboxy-terminal (YGX)4 repeat domain of CpsD, an autophosphorylating tyrosine kinase required for capsule biosynthesis in Streptococcus pneumoniae. *J. Bacteriol.* (2003). doi:10.1128/JB.185.10.3009-3019.2003

34. Morona, J. K., Paton, J. C., Miller, D. C. & Morona, R. Tyrosine phosphorylation of CpsD negatively regulates capsular polysacchsaride biosynthesis in Streptococcus pneumoniae. *Mol. Microbiol.* (2000). doi:10.1046/j.1365-2958.2000.01808.x

35. Morona, J. K., Miller, D. C., Morona, R. & Paton, J. C. The Effect That Mutations in the Conserved Capsular Polysaccharide Biosynthesis Genes cpsA, cpsB, and cpsD Have on Virulence of Streptococcus pneumoniae . *J. Infect. Dis.* (2004). doi:10.1086/383352

36. Oliver, M. B., Jones, C., Larson, T. R., Calix, J. J., Zartler, E. R., Yother, J. & Nahm, M. H. Streptococcus pneumoniae serotype 11D has a bispecific glycosyltransferase and expresses two different capsular polysaccharide repeating units. *J. Biol. Chem.* (2013). doi:10.1074/jbc.M113.488528

37. Cieslewicz, M. J., Kasper, D. L., Wang, Y. & Wessels, M. R. Functional analysis in type Ia group B Streptococcus of a cluster of genes involved in extracellular polysaccharide production by diverse species of streptococci. *J. Biol. Chem.* (2001). doi:10.1074/jbc.M005702200

38. Mollerach, M., López, R. & García, E. Characterization of the galU gene of Streptococcus pneumoniae encoding a uridine diphosphoglucose pyrophosphorylase: a gene essential for capsular polysaccharide biosynthesis. *J. Exp. Med.* (1998).

39. Calix, J. J., Saad, J. S., Brady, A. M. & Nahm, M. H. Structural characterization of Streptococcus pneumoniae serotype 9A capsule polysaccharide reveals role of glycosyl 6-O-acetyltransferase wcjE in serotype 9V capsule biosynthesis and immunogenicity. *J. Biol. Chem.* (2012). doi:10.1074/jbc.M112.346924

40. Langereis, J. D. & de Jonge, M. I. Non-encapsulated Streptococcus pneumoniae , vaccination as a measure to interfere with horizontal gene transfer . *Virulence* (2017). doi:10.1080/21505594.2017.1309492

41. Park, I. H., Kim, K. H., Andrade, A. L., Briles, D. E., Mcdaniel, L. S. & Nahma, M. H. Nontypeable pneumococci can be divided into multiple cps types, including one type expressing the novel gene pspK.

MBio (2012). doi:10.1128/mBio.00035-12

42. Sá-Leão, R., Nunes, S., Brito-Avô, A., Frazão, N., Simões, A. S., Crisóstomo, M. I., Paulo, A. C. S., Saldanha, J., Santos-Sanches, I. & de Lencastre, H. Changes in pneumococcal serotypes and antibiotypes carried by vaccinated and unvaccinated day-care centre attendees in Portugal, a country with widespread use of the seven-valent pneumococcal conjugate vaccine. *Clin. Microbiol. Infect.* (2009). doi:10.1111/j.1469-0691.2009.02775.x

43. Marks, L. R., Reddinger, R. M. & Hakansson, A. P. High levels of genetic recombination during nasopharyngeal carriage and biofilm formation in Streptococcus pneumoniae. *MBio* (2012). doi:10.1128/mBio.00200-12

44. Andam, C. P. & Hanage, W. P. Mechanisms of genome evolution of Streptococcus. *Infect. Genet. Evol.* (2015). doi:10.1016/j.meegid.2014.11.007

45. Keller, L. E., Robinson, D. A. & McDaniel, L. S. Nonencapsulated Streptococcus pneumoniae : Emergence and Pathogenesis . *MBio* (2016). doi:10.1128/mbio.01792-15

46. Hirst, R. A., Kadioglu, A., O'Callaghan, C. & Andrew, P. W. The role of pneumolysin in pneumococcal pneumonia and meningitis. *Clin. Exp. Immunol.* (2004). doi:10.1111/j.1365-2249.2004.02611.x

47. Kalin, M., Kanclerski, K., Granstrom, M. & Mollby, R. Diagnosis of pneumococcal pneumonia by enzyme-linked immunosorbent assay of antibodies to pneumococcal hemolysin (pneumolysin). *J. Clin. Microbiol.* (1987).

48. Jedrzejas, M. J. Pneumococcal Virulence Factors: Structure and Function. *Microbiol. Mol. Biol. Rev.* (2003). doi:10.1128/mmbr.65.2.187-207.2001

49. Boulnois, G. J., Paton, J. C., Mitchell, T. J. & Andrew, P. W. Structure and function of pneumolysin, the multifunctional, thiol-activated toxin of Streptococcus pneumoniae. *Mol. Microbiol.* (1991). doi:10.1111/j.1365-2958.1991.tb01969.x

50. Walker, J. A., Allen, R. L., Falmagne, P., Johnson, M. K. & Boulnois, G. J. Molecular cloning, characterization, and complete nucleotide sequence of the gene for pneumolysin, the sulfhydryl-activated toxin of Streptococcus pneumoniae. *Infect. Immun.* (1987).

51. Davis, K. M., Nakamura, S. & Weiser, J. N. Nod2 sensing of lysozyme-digested peptidoglycan promotes macrophage recruitment and clearance of S. pneumoniae colonization in mice. *J. Clin. Invest.* (2011). doi:10.1172/JCI57761

52. Karmakar, M., Katsnelson, M., Malak, H. A., Greene, N. G., Howell, S. J., Hise, A. G., Camilli, A., Kadioglu, A., Dubyak, G. R. & Pearlman, E. Neutrophil IL-1β Processing Induced by Pneumolysin Is Mediated by the NLRP3/ASC Inflammasome and Caspase-1 Activation and Is Dependent on K + Efflux . *J. Immunol.* (2015). doi:10.4049/jimmunol.1401624

53. Wani, J. H., Gilbert, J. V., Plaut, A. G. & Weiser, J. N. Identification, cloning, and sequencing of the immunoglobulin A1 protease gene of Streptococcus pneumoniae. *Infect. Immun.* (1996).

54. Hermans, P. W. M., Adrian, P. V., Albert, C., Estevão, S., Hoogenboezem, T., Luijendijk, I. H. T., Kamphausen, T. & Hammerschmidt, S. The streptococcal lipoprotein rotamase A (SlrA) is a functional peptidyl-prolyl isomerase involved in pneumococcal colonization. *J. Biol. Chem.* (2006). doi:10.1074/jbc.M510014200

55. Overweg, K., Kerr, A., Sluijter, M., Jackson, M. H., Mitchell, T. J., De Jong, A. P. J. M., De Groot, R. & Hermans, P. W. M. The putative proteinase maturation protein A of Streptococcus pneumoniae is a conserved surface protein with potential to elicit protective immune responses. *Infect. Immun.* (2000). doi:10.1128/IAI.68.7.4180-4188.2000

56. Bergmann, S. & Hammerschmidt, S. Versatility of pneumococcal surface proteins. *Microbiology* (2006). doi:10.1099/mic.0.28610-0

57. Lawrence, M. C., Pilling, P. A., Epa, V. C., Berry, A. M., Ogunniyi, A. D. & Paton, J. C. The crystal structure of pneumococcal surface antigen PsaA reveals a metal-binding site and a novel structure for a putative ABC-type binding protein. *Structure* (1998). doi:10.1016/S0969-2126(98)00153-1

58. Romero-Steiner, S., Pilishvili, T., Sampson, J. S., Johnson, S. E., Stinson, A., Carlone, G. M. & Ades, E. W. Inhibition of Pneumococcal Adherence to Human Nasopharyngeal Epithelial Cells by Anti-PsaA Antibodies. *Clin. Vaccine Immunol.* (2003). doi:10.1128/CDLI.10.2.246-251.2003

59. Johnston, J. W., Myers, L. E., Ochs, M. M., Benjamin, W. H., Briles, D. E. & Hollingshead, S. K. Lipoprotein PsaA in virulence of Streptococcus pneumoniae: Surface accessibility and role in protection from superoxide. *Infect. Immun.* (2004). doi:10.1128/IAI.72.10.5858-5867.2004

60. Brown, J. S., Gilliland, S. M. & Holden, D. W. A Streptococcus pneumoniae pathogenicity island encoding an ABC transporter involved in iron uptake and virulence. *Mol. Microbiol.* (2001). doi:10.1046/j.1365-2958.2001.02414.x

61. Brown, J. S., Ogunniyi, A. D., Woodrow, M. C., Holden, D. W. & Paton, J. C. Immunization with components of two iron uptake ABC transporters protects mice against systemic Streptococcus pneumoniae infection. *Infect. Immun.* (2001). doi:10.1128/IAI.69.11.6702-6706.2001

62. Cundell, D. R., Gerard, N. P., Gerard, C., Idanpaan-Heikkila, I. & Tuomanen, E. I. Streptococcus pneumoniae anchor to activated human cells by the receptor for platelet-activating factor. *Nature* (1995). doi:10.1038/377435a0

63. Yuste, J., Khandavilli, S., Ansari, N., Muttardi, K., Ismail, L., Hyams, C., Weiser, J., Mitchell, T. & Brown, J. S. The effects of PspC on complement-mediated immunity to Streptococcus pneumoniae vary with strain background and capsular serotype. *Infect. Immun.* (2010). doi:10.1128/IAI.00541-09

64. Clancy, K. W., Melvin, J. A. & McCafferty, D. G. Sortase transpeptidases: insights into mechanism, substrate specificity, and inhibition. *Biopolymers* (2010). doi:10.1002/bip.21472

65. Chen, S., Paterson, G. K., Tong, H. H., Mitchell, T. J. & DeMaria, T. F. Sortase A contributes to pneumococcal nasopharyngeal colonization in the chinchilla model. *FEMS Microbiol. Lett.* (2005). doi:10.1016/j.femsle.2005.09.052

66. Paterson, G. K. & Mitchell, T. J. The role of Streptococcus pneumoniae sortase A in colonisation and pathogenesis. *Microbes Infect.* (2006). doi:10.1016/j.micinf.2005.06.009

67. Manco, S., Hernon, F., Yesilkaya, H., Paton, J. C., Andrew, P. W. & Kadioglu, A. Pneumococcal neuraminidases A and B both have essential roles during infection of the respiratory tract and sepsis. *Infect. Immun.* (2006). doi:10.1128/IAI.01237-05

68. Holmes, A. R., McNab, R., Millsap, K. W., Rohde, M., Hammerschmidt, S., Mawdsley, J. L. & Jenkinson, H. F. The pavA gene of Streptococcus pneumoniae encodes a fibronectin-binding protein that is essential for virulence. *Mol. Microbiol.* (2001). doi:10.1046/j.1365-2958.2001.02610.x

69. Bergmann, S., Rohde, M., Chhatwal, G. S. & Hammerschmidt, S. α-Enolase of Streptococcus pneumoniae is a plasmin(ogen)-binding protein displayed on the bacterial cell surface. *Mol. Microbiol.* (2001). doi:10.1046/j.1365-2958.2001.02448.x

70. Davis, K. M., Akinbi, H. T., Standish, A. J. & Weiser, J. N. Resistance to mucosal lysozyme compensates for the fitness deficit of peptidoglycan modifications by Streptococcus pneumoniae. *PLoS Pathog.* (2008). doi:10.1371/journal.ppat.1000241

71. Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I. & Dewhirst, F. E. Defining the normal bacterial flora of the oral

cavity. *J. Clin. Microbiol.* (2005). doi:10.1128/JCM.43.11.5721-5732.2005

72. Sandgren, A., Albiger, B., Orihuela, C. J., Tuomanen, E., Normark, S. & Henriques-Normark, B. Virulence in Mice of Pneumococcal Clonal Types with Known Invasive Disease Potential in Humans. *J. Infect. Dis.* (2005). doi:10.1086/432513

73. World Health Organisation. Pneumococcal conjugate vaccine for childhood immunization. *WHO position Pap. Wkly. Epidemiol. Rec.* (2007).

74. WHO | Estimates of disease burden and cost-effectiveness. *WHO* (2017). at <http://www.who.int/immunization/monitoring_surveillance/burden/estimates/en/>

75. Wahl, B., O'Brien, K. L., Greenbaum, A., Majumder, A., Liu, L., Chu, Y., Lukšić, I., Nair, H., McAllister, D. A., Campbell, H., Rudan, I., Black, R. & Knoll, M. D. Burden of Streptococcus pneumoniae and Haemophilus influenzae type b disease in children in the era of conjugate vaccines: global, regional, and national estimates for 2000–15. *Lancet Glob. Heal.* (2018). doi:10.1016/S2214-109X(18)30247-X

76. WHO. WHO priority pathogens list for R&D of new antibioticse. *http://www.who.int/en/news-room/detail/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed* (2017).

77. C., V. M., S., T., A., V. G., S., M., V., Q., C., F., J., C., S.A., M., K.L., O., K.P., K., C.G., W., C., C., Von Mollendorf, C., Tempia, S., Von Gottberg, A., Meiring, S., Quan, V., Feldman, C., Cloete, J., Madhi, S. A., O'Brien, K. L., Klugman, K. P., Whitney, C. G. & Cohen, C. Estimated severe pneumococcal disease cases and deaths before and after pneumococcal conjugate vaccine introduction in children younger than 5 years of age in South Africa. *PLoS One* (2017). doi:10.1371/journal.pone.0179905

78. Lazzerini, M., Seward, N., Lufesi, N., Banda, R., Sinyeka, S., Masache, G., Nambiar, B., Makwenda, C., Costello, A., McCollum, E. D. & Colbourn, T. Mortality and its risk factors in Malawian children admitted to hospital with clinical pneumonia, 2001-12: A retrospective observational study. *Lancet Glob. Heal.* (2016). doi:10.1016/S2214-109X(15)00215-6

79. Everett, D. B., Mukaka, M., Denis, B., Gordon, S. B., Carrol, E. D., van Oosterhout, J. J., Molyneux, E. M., Molyneux, M., French, N. & Heyderman, R. S. Ten years of surveillance for invasive streptococcus pneumoniae during the era of antiretroviral scale-up and cotrimoxazole prophylaxis in Malawi. *PLoS One* (2011). doi:10.1371/journal.pone.0017765

80. Maloy, S. in *Encycl. Biodivers. Second Ed.* (2013). doi:10.1016/B978-0-12-384719-5.00431-7

81. Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., Madupu, R., Brinkac, L. M., Dodson, R. J., Rosovitz, M. J., Sullivan, S. A., Daugherty, S. C., Haft, D. H., Selengut, J., Gwinn, M. L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K. J. B., Smith, S., Utterback, T. R., White, O., Rubens, C. E., Grandi, G., Madoff, L. C., Kasper, D. L., Telford, J. L., Wessels, M. R., Rappuoli, R. & Fraser, C. M. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial 'pan-genome'. *Proc. Natl. Acad. Sci.* **102,** 13950–13955 (2005).

82. Rouli, L., Merhej, V., Fournier, P. E. & Raoult, D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect.* **7,** 72–85 (2015).

83. Huber, W., Carey, V. J., Long, L., Falcon, S. & Gentleman, R. Graphs in molecular biology. *BMC Bioinformatics* (2007). doi:10.1186/1471-2105-8-S6-S8

84. Wilson, R. J. in *Hist. Topol.* (2006). doi:10.1016/B978-044482375-5/50018-3

85. Rizk, G., Lavenier, D. & Chikhi, R. DSK: K-mer counting with very low memory usage. *Bioinformatics* (2013).

doi:10.1093/bioinformatics/btt020

86. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx304

87. Pandey, P., Bender, M. A., Johnson, R. & Patro, R. Squeakr: An exact and approximate k -mer counting system. *Bioinformatics* (2018). doi:10.1093/bioinformatics/btx636

88. Marcus, S., Lee, H. & Schatz, M. C. SplitMEM: A graphical algorithm for pan-genome analysis with suffix skips. *Bioinformatics* (2014). doi:10.1093/bioinformatics/btu756

89. Baier, U., Beller, T. & Ohlebusch, E. Graphical pan-genome analysis with compressed suffix trees and the Burrows-Wheeler transform. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv603

90. Marschall, T., Marz, M., Abeel, T., Dijkstra, L., Dutilh, B. E., Ghaffaari, A., Kersey, P., Kloosterman, W., Makinen, V., Novak, A., Paten, B., Porubsky, D., RIVALS, E., Alkan, C., Baaijens, J., de Bakker, P. I. W., Boeva, V., Bonnal, R. J. P., Chiaromonte, F., Chikhi, R., Ciccarelli, F. D., Cijvat, R., Datema, E., Van Duijn, C. M., Eichler, E. E., Ernst, C., Eskin, E., Garrison, E., El-Kebir, M., Klau, G. W., Korbel, J. O., Lameijer, E., Langmead, B., Martin, M., Medvedev, P., Mu, J. C., Neerincx, P., Ouwens, K., Peterlongo, P., Nadia, P., Rahmann, S., Raphael, B., Reinert, K., de Ridder, D., de Ridder, J., Schlesner, M., Schulz-Trieglaff, O., Sanders, A., Sheikhizadeh, S., Shneider, C., Smit, S., Valenzuela, D., Wang, J., Wessels, L., Zhang, Y., Guryev, V., Vandin, F., Ye, K. & Schoenhuth, A. *Computational Pan-Genomics: Status, Promises and Challenges*. *bioRxiv* (2016). doi:10.1101/043430

91. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* (2009). doi:10.1093/nar/gkp1137

92. Andrews, S. *FASTQC. A quality control tool for high throughput sequence data. 2010*. *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/* (2010).

93. Gordon, A. & Hannon, G. J. Fastx-toolkit. FASTQ/A short-reads pre-processing tools. *Unpubl. http//hannonlab.cshl.edu/fastx_toolkit/* (2010).

94. Paszkiewicz, K. & Studholme, D. J. De novo assembly of short sequence reads. *Brief. Bioinform.* (2010). doi:10.1093/bib/bbq020

95. Gladman, S. & Seemann, T. VelvetOptimiser. *Free Softw. Found.* (2008). doi:10.1016/S0925-8574(99)00040-3

96. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D. W., Yiu, S. M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T. W. & Wang, J. Erratum to 'SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler' [GigaScience, (2012), 1, 18]. *Gigascience* (2015). doi:10.1186/s13742-015-0069-2

97. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* (2013). doi:10.1093/bioinformatics/btt086

98. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* (2007). doi:10.1093/bioinformatics/btm009

99. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).

100. Petersen, T. N., Brunak, S., Von Heijne, G. & Nielsen, H. SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* (2011). doi:10.1038/nmeth.1701

101. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide

sequences. *Nucleic Acids Res.* (2004). doi:10.1093/nar/gkh152

102. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* (2011). doi:10.1093/nar/gkr367

103. Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J. & Finn, R. D. Rfam 12.0: Updates to the RNA families database. *Nucleic Acids Res.* (2015). doi:10.1093/nar/gku1063

104. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* (2013). doi:10.1093/bioinformatics/btt509

105. Santos, A. R., Barbosa, E., Fiaux, K., Zurita-Turk, M., Chaitankar, V., Kamapantula, B., Abdelzaher, A., Ghosh, P., Tiwari, S., Barve, N., Jain, N., Barh, D., Silva, A., Miyoshi, A. & Azevedo, V. PANNOTATOR: An automated tool for annotation of pan-genomes. *Genet. Mol. Res.* (2013). doi:10.4238/2013.August.16.2

106. Wozniak, M., Wong, L. & Tiuryn, J. ECAMBer: Efficient support for large-scale comparative analysis of multiple bacterial strains. *BMC Bioinformatics* (2014). doi:10.1186/1471-2105-15-65

107. Angiuoli, S. V., Dunning Hotopp, J. C., Salzberg, S. L. & Tettelin, H. Improving pan-genome annotation using whole genome multiple alignment. *BMC Bioinformatics* (2011). doi:10.1186/1471-2105-12-272

108. Zekic, T., Holley, G. & Stoye, J. in *Methods Mol. Biol.* (2018). doi:10.1007/978-1-4939-7463-4_2

109. Blom, J., Kreis, J., Spänig, S., Juhre, T., Bertelli, C., Ernst, C. & Goesmann, A. EDGAR 2.0: an enhanced software platform for comparative gene content analyses. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw255

110. Brittnacher, M. J., Fong, C., Hayden, H. S., Jacobs, M. A., Radey, M. & Rohmer, L. PGAT: A multistrain analysis resource for microbial genomes. *Bioinformatics* (2011). doi:10.1093/bioinformatics/btr418

111. Zhao, Y., Wu, J., Yang, J., Sun, S., Xiao, J. & Yu, J. PGAP: Pan-genomes analysis pipeline. *Bioinformatics* (2012). doi:10.1093/bioinformatics/btr655

112. Inman, J. M., Sutton, G. G., Beck, E., Brinkac, L. M., Clarke, T. H. & Fouts, D. E. Large-Scale Comparative Analysis of Microbial Pan-genomes using PanOCT. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty744

113. Contreras-Moreira, B. & Vinuesa, P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.* (2013). doi:10.1128/AEM.02411-13

114. Lukjancenko, O., Thomsen, M. C., Voldby Larsen, M. & Ussery, D. W. PanFunPro: PAN-genome analysis based on FUNctional PROfiles. *F1000Research* (2013). doi:10.12688/f1000research.2-265.v1

115. Benedict, M. N., Henriksen, J. R., Metcalf, W. W., Whitaker, R. J. & Price, N. D. ITEP: An integrated toolkit for exploration of microbial pan-genomes. *BMC Genomics* (2014). doi:10.1186/1471-2164-15-8

116. Zhao, Y., Jia, X., Yang, J., Ling, Y., Zhang, Z., Yu, J., Wu, J. & Xiao, J. PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics* (2014). doi:10.1093/bioinformatics/btu017

117. Sahl, J. W., Caporaso, J. G., Rasko, D. A. & Keim, P. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. *PeerJ* (2014). doi:10.7717/peerj.332

118. Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A. & Parkhill, J. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31,** 3691–3693 (2015).

119. Snipen, L. & Liland, K. H. micropan: An R-package for microbial pan-genomics. *BMC Bioinformatics* (2015).

doi:10.1186/s12859-015-0517-0

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

121. Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N. & Corander, J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty539

122. Laing, C., Buchanan, C., Taboada, E. N., Zhang, Y., Kropinski, A., Villegas, A., Thomas, J. E. & Gannon, V. P. J. Pan-genome sequence analysis using Panseq: An online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics* (2010). doi:10.1186/1471-2105-11-461

123. Treangen, T. J., Ondov, B. D., Koren, S. & Phillippy, A. M. The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* (2014). doi:10.1186/s13059-014-0524-x

124. Minkin, I., Pham, S. & Medvedev, P. TwoPaCo: an efficient algorithm to build the compacted de Bruijn graph from many complete genomes. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btw609

125. Holley, G., Wittler, R. & Stoye, J. Bloom Filter Trie: An alignment-free and reference-free data structure for pan-genome storage. *Algorithms Mol. Biol.* (2016). doi:10.1186/s13015-016-0066-8

126. Keane, J. A., Page, A. J., Delaney, A. J., Taylor, B., Seemann, T., Harris, S. R. & Soares, J. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genomics* (2016). doi:10.1099/mgen.0.000056

127. Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, P., McWilliam, H., Valentin, F., Wallace, I., Wilm, A., Lopez, R., Thompson, J., Gibson, T. & Higgins, D. ClustalW and ClustalX version 2. *Bioinformatics* (2007). doi:doi:10.1093/bioinformatics/btm404

128. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* (2010). doi:10.1371/journal.pone.0009490

129. He, Z., Zhang, H., Gao, S., Lercher, M. J., Chen, W. H. & Hu, S. Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw370

130. Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* **17,** 238 (2016).

131. Hurgobin, B. & Edwards, D. SNP Discovery Using a Pangenome: Has the Single Reference Approach Become Obsolete? *Biology (Basel).* **6,** 21 (2017).

132. Limasset, A., Cazaux, B., Rivals, E. & Peterlongo, P. Read mapping on de Bruijn graphs. *BMC Bioinformatics* (2016). doi:10.1186/s12859-016-1103-9

133. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* (2012). doi:10.1038/ng.1028

134. Xiao, J., Zhang, Z., Wu, J. & Yu, J. A brief review of software tools for pangenomics. *Genomics, Proteomics Bioinforma.* (2015). doi:10.1016/j.gpb.2015.01.007

135. Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* (2015). doi:10.1016/j.mib.2014.11.016

136. Hadfield, J., Croucher, N. J., Goater, R. J., Abudahab, K., Aanensen, D. M. & Harris, S. R. Phandango: An interactive viewer for bacterial population genomics. *Bioinformatics* (2018). doi:10.1093/bioinformatics/btx610

137. Ding, W., Baumdicker, F. & Neher, R. A. panX: pan-genome analysis and exploration. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx977

138. Pedersen, T. L., Nookaew, I., Wayne Ussery, D. & Månsson, M. PanViz: Interactive visualization of the structure of functionally annotated pangenomes. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btw761

139. Heinsbroek, E., Tafatatha, T., Chisambo, C., Phiri, A., Mwiba, O., Ngwira, B., Crampin, A. C., Read, J. M. & French, N. Pneumococcal Acquisition among Infants Exposed to HIV in Rural Malawi: A Longitudinal Household Study. in *Am. J. Epidemiol.* (2016). doi:10.1093/aje/kwv134

140. Crampin, A. C., Dube, A., Mboma, S., Price, A., Chihana, M., Jahn, A., Baschieri, A., Molesworth, A., Mwaiyeghele, E., Branson, K., Floyd, S., Mcgrath, N., Fine, P. E. M., French, N., Glynn, J. R. & Zaba, B. Profile: The Karonga health and demographic surveillance system. *Int. J. Epidemiol.* (2012). doi:10.1093/ije/dys088

141. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32,** 3047–3048 (2016).

142. Langmead, B., Salzberg, S. L. & Langmead. Bowtie2. *Nat. Methods* (2013). doi:10.1038/nmeth.1923.Fast

143. Langmead, B. Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinforma.* (2010). doi:10.1002/0471250953.bi1107s32

144. Kim, H.-Y. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restor. Dent. Endod.* **42,** 152 (2017).

145. Swarthout, T. D., Fronterre, C., Lourenço, J., Obolski, U., Gori, A., Bar-Zeev, N., Everett, D., Kamng'ona, A. W., Mwalukomo, T. S., Mataya, A. A., Mwansambo, C., Banda, M., Gupta, S., Diggle, P., French, N. & Heyderman, R. S. High residual carriage of vaccine-serotype Streptococcus pneumoniae after introduction of pneumococcal conjugate vaccine in Malawi. *Nat. Commun.* **11,** 1–12 (2020).

146. Cornick, J. E., Everett, D. B., Broughton, C., Denis, B. B., Banda, D. L., Carrol, E. D. & Parry, C. M. Invasive streptococcus pneumoniae in children, Malawi, 2004-2006. *Emerg. Infect. Dis.* (2011). doi:10.3201/eid1706.101404

147. Hausdorff, W. P. The roles of pneumococcal serotypes 1 and 5 in paediatric invasive disease. *Vaccine* (2007). doi:10.1016/j.vaccine.2006.09.009

148. Cornick, J. E., Chaguza, C., Harris, S. R., Yalcin, F., Senghore, M., Kiran, A. M., Govindpershad, S., Ousmane, S., Du Plessis, M. & Pluschke, G. Region-specific diversification of the highly virulent serotype 1 Streptococcus pneumoniae. *Microb. Genomics* **1,** (2015).

149. Leimkugel, J., Adams Forgor, A., Gagneux, S., Pflüger, V., Flierl, C., Awine, E., Naegeli, M., Dangy, J.-P., Smith, T. & Hodgson, A. An outbreak of serotype 1 Streptococcus pneumoniae meningitis in northern Ghana with features that are characteristic of Neisseria meningitidis meningitis epidemics. *J. Infect. Dis.* **192,** 192–199 (2005).

150. Gessner, B. D., Mueller, J. E. & Yaro, S. African meningitis belt pneumococcal disease epidemiology indicates a need for an effective serotype 1 containing vaccine, including for older children and adults. *BMC Infect. Dis.* (2010). doi:10.1186/1471-2334-10-22

151. Swetha, R. G., Sekar, D. K. K., Devi, E. D., Ahmed, Z. Z., Ramaiah, S., Anbarasu, A. & Sekar, K. Streptococcus pneumoniae Genome Database (SPGDB): A database for strain specific comparative analysis of Streptococcus pneumoniae genes and proteins. *Genomics* (2014). doi:10.1016/j.ygeno.2014.09.012

152. Hoskins, J., Alborn, J., Arnold, J., Blaszczak, L. C., Burgett, S., Dehoff, B. S., Estrem, S. T., Fritz, L., Fu, D. J., Fuller, W., Geringer, C., Gilmour, R., Glass, J. S., Khoja, H., Kraft, A. R., Lagace, R. E., LeBlanc, D. J., Lee, L. N.,

Lefkowitz, E. J., Lu, J., Matsushima, P., McAhren, S. M., McHenney, M., McLeaster, K., Mundy, C. W., Nicas, T. I., Norris, F. H., O'Gara, M., Peery, R. B., Robertson, G. T., Rockey, P., Sun, P. M., Winkler, M. E., Yang, Y., Young-Bellido, M., Zhao, G., Zook, C. A., Baltz, R. H., Jaskunas, S. R., Rosteck, J., Skatrud, P. L. & Glass, J. I. Genome of the bacterium Streptococcus pneumoniae strain R6. *J. Bacteriol.* **183,** 5709–5717 (2001).

153. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. & Phillippy, A. M. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res.* **27,** 722–736 (2017).

154. Flye/FAQ.md at flye · fenderglass/Flye · GitHub. at <https://github.com/fenderglass/Flye/blob/flye/docs/FAQ.md>

155. Kamath, G. M., Shomorony, I., Xia, F., Courtade, T. A. & Tse, D. N. HINGE: Long-read assembly achieves optimal repeat resolution. *Genome Res.* **27,** 747–756 (2017).

156. Zerbino, D. R. Using the Velvet de novo assembler for short-read sequencing technologies. *Curr. Protoc. Bioinforma.* (2010). doi:10.1002/0471250953.bi1105s31

157. Compeau, P. E. C., Pevzner, P. A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* **29,** 987–991 (2011).

158. Edwards, D. J. & Holt, K. E. *Beginner's guide to comparative bacterial genome analysis using next-generation sequence data*. (2013). at <http://www.microbialinformaticsj.com/content/3/1/2>

159. Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W. & Hauser, L. J. *Prodigal: prokaryotic gene recognition and translation initiation site identification*. (2010). doi:10.1186/1471-2105-11-119

160. Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H. H., Rognes, T. & Ussery, D. W. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* (2007). doi:10.1093/nar/gkm160

161. Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A. & Parkhill, J. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv421

162. Roary: the pan genome pipeline. at <https://sanger-pathogens.github.io/Roary/>

163. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* (2012). doi:10.1093/bioinformatics/bts565

164. Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. & Madden, T. L. NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36,** 5–9 (2008).

165. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* (2002). doi:10.1093/nar/30.7.1575

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

167. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* (2013). doi:10.1093/molbev/mst010

168. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., Lanfear, R. & Teeling, E. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37,** 1530–1534 (2020).

169. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A. & Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37,** 1530–1534 (2020).

170. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw290

171. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* (2010). doi:10.1186/1471-2105-11-367

172. Gaujoux, R. An introduction to NMF package. *BMC Bioinformatics* (2010). doi:10.1186/1471-2105-11-367

173. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J. & Von Mering, C. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky1131

174. Ge, S. X., Jung, D., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36,** 2628–2629 (2020).

175. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C. & Eddy, S. R. The Pfam protein families database. *Nucleic Acids Res.* **32,** D138–D141 (2004).

176. Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G. A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D. H., Letunic, I., Marchler-Bauer, A., Mi, H., Natale, D. A., Necci, M., Orengo, C. A., Pandurangan, A. P., Rivoire, C., Sigrist, C. J. A., Sillitoe, I., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Wu, C. H., Bateman, A. & Finn, R. D. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49,** D344–D354 (2021).

177. Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Ridwan Amode, M., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., Gall, A., Giron, C. G., Grego, T., Guijarro-Clarke, C., Haggerty, L., Hemrom, A., Hourlier, T., Izuogu, O. G., Juettemann, T., Kaikala, V., Kay, M., Lavidas, I., Le, T., Lemos, D., Martinez, J. G., Marugán, J. C., Maurel, T., McMahon, A. C., Mohanan, S., Moore, B., Muffato, M., Oheh, D. N., Paraschas, D., Parker, A., Parton, A., Prosovetskaia, I., Sakthivel, M. P., Abdul Salam, A. I., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Steed, E., Szpak, M., Szuba, M., Taylor, K., Thormann, A., Threadgold, G., Walts, B., Winterbottom, A., Chakiachvili, M., Chaubal, A., de Silva, N., Flint, B., Frankish, A., Hunt, S. E., IIsley, G. R., Langridge, N., Loveland, J. E., Martin, F. J., Mudge, J. M., Morales, J., Perry, E., Ruffier, M., Tate, J., Thybert, D., Trevanion, S. J., Cunningham, F., Yates, A. D., Zerbino, D. R. & Flicek, P. Ensembl 2021. *Nucleic Acids Res.* **49,** D884–D891 (2021).

178. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27,** 29–34 (1999).

179. QUAST 5.0.2 manual. at <http://quast.sourceforge.net/docs/manual.html>

180. Cross, R. L. & Müller, V. The evolution of A-, F-, and V-type ATP synthases and ATPases: reversals in function and changes in the H+/ATP coupling ratio. *FEBS Lett.* **576,** 1–4 (2004).

181. Rappas, M., Niwa, H. & Zhang, X. Mechanisms of ATPases - A Multi-Disciplinary Approach. *Curr. Protein Pept. Sci.* **5,** 89–105 (2005).

182. Campbell, J. A., Davies, G. J., Bulone, V. & Henrissat, B. A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *Biochem. J.* **326,** 929 (1997).

183. Zhang, S. & Meyer, R. The relaxosome protein MobC promotes conjugal plasmid mobilization by extending DNA strand separation to the nick site at the origin of transfer. *Mol. Microbiol.* **25,** 509–516 (1997).

184. Yokoyama, K. & Imamura, H. Rotation, structure, and classification of prokaryotic V-ATPase. *J. Bioenerg. Biomembr.* (2005). doi:10.1007/s10863-005-9480-1

185. Reizer, J., Reizer, A. & Saier, M. H. A functional superfamily of sodium/solute symporters. *Biochim. Biophys. Acta* **1197,** 133–166 (1994).

186. Lizcano, A., Akula Suresh Babu, R., Shenoy, A. T., Maren Saville, A., Kumar, N., Hinojosa, C. A., Gilley, R. P., Segovia, J., Mitchell, T. J., Tettelin, H. & Orihuela, C. J. Transcriptional organization of pneumococcal psrP-secY2A2 and impact of GtfA and GtfB deletion on PsrP-associated virulence properties. (2017). doi:10.1016/j.micinf.2017.04.001

187. Pei, J., Mitchell, D. A., Dixon, J. E. & Grishin, N. V. Expansion of type II CAAX proteases reveals evolutionary origin of γ-secretase subunit APH-1. *J Mol Biol* **410,** 18–26 (2011).

188. Grenha, R., Slamti, L., Nicaise, M., Refes, Y., Lereclus, D. & Nessler, S. Structural basis for the activation mechanism of the PlcR virulence regulator by the quorum-sensing signal peptide PapR. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 1047–1052 (2013).

189. Pena, R. T., Blasco, L., Ambroa, A., González-Pedrajo, B., Fernández-García, L., López, M., Bleriot, I., Bou, G., García-Contreras, R., Wood, T. K. & Tomás, M. Relationship between quorum sensing and secretion systems. *Front. Microbiol.* **10,** 1100 (2019).

190. Daniel, R. A. & Errington, J. Cloning, DNA Sequence, Functional Analysis and Transcriptional Regulation of the Genes Encoding Dipicolinic Acid Synthetase Required for Sporulation in Bacillus subtilis. *J. Mol. Biol.* **232,** 468–483 (1993).

191. Appleyard, A. N., Choi, S., Read, D. M., Lightfoot, A., Boakes, S., Hoffmann, A., Chopra, I., Bierbaum, G., Rudd, B. A. M., Dawson, M. J. & Cortes, J. Dissecting Structural and Functional Diversity of the Lantibiotic Mersacidin. *Chem. Biol.* **16,** 490–498 (2009).

192. Qi, F., Chen, P. & Caufield, P. W. The group I strain of Streptococcus mutans, UA140, produces both the lantibiotic mutacin I and a nonlantibiotic bacteriocin, mutacin IV. *Appl. Environ. Microbiol.* **67,** 15–21 (2001).

193. Rekvig, O. P., Flægstad, T., Fredriksen, K. & Traavik, T. Stimulation of Clones Specific for dsDNA or Idiotypes of Anti-dsDNA as a Consequence of BK Virus Inoculation. *http://dx.doi.org/10.3109/08820138909057753* **18,** 657–669 (2009).

194. Mazmanian, S. K., Ton-That, H. & Schneewind, O. Sortase-catalysed anchoring of surface proteins to the cell wall of Staphylococcus aureus. *Mol. Microbiol.* **40,** 1049–1057 (2001).

195. Maresso, A. W. & Schneewind, O. Sortase as a Target of Anti-Infective Therapy. *Pharmacol. Rev.* **60,** 128–141 (2008).

196. Hiller, N. L. & Sá-Leão, R. Puzzling Over the Pneumococcal Pangenome. *Front. Microbiol.* **9,** (2018).

197. Embry, A., Hinojosa, E. & Orihuela, C. J. Regions of Diversity 8, 9 and 13 contribute to Streptococcus pneumoniae virulence. *BMC Microbiol.* (2007). doi:10.1186/1471-2180-7-80

198. Brückner, R., Nuhn, M., Reichmann, P., Weber, B. & Hakenbeck, R. Mosaic genes and mosaic chromosomes-genomic variation in Streptococcus pneumoniae. *Int. J. Med. Microbiol.* (2004). doi:10.1016/j.ijmm.2004.06.019

199. Obert, C., Sublett, J., Kaushal, D., Hinojosa, E., Barton, T., Tuomanen, E. I. & Orihuela, C. J. Identification of a candidate Streptococcus pneumoniae core genome and regions of diversity correlated with invasive pneumococcal disease. *Infect. Immun.* **74,** 4766–4777 (2006).

200. van der Poll, T. & Opal, S. M. Pathogenesis, treatment, and prevention of pneumococcal pneumonia. *Lancet* **374,** 1543–1556 (2009).

201. panX. at <https://pangenome.org/Streptococcus_pneumoniae>

202. Benso, A., Ijaq, J., Sundararajan, V. S., Chandrasekharan, M., Poddar, R. & Bethi, N. Annotation and curation of uncharacterized proteins- challenges. (2015). doi:10.3389/fgene.2015.00119

203. Ekroth, A. K. E., Gerth, M., Stevens, E. J., Ford, S. A. & King, K. C. Host genotype and genetic diversity shape the evolution of a novel bacterial infection. *ISME J.* 1–12 (2021).

204. Brown, J. S. Single nucleotide polymorphisms within the cps loci: another potential source of clinically important genetic variation for Streptococcus pneumoniae? *Infect. Immun.* IAI-00374 (2021).

205. Lees, J. A., Ferwerda, B., Kremer, P. H. C., Wheeler, N. E., Serón, M. V., Croucher, N. J., Gladstone, R. A., Bootsma, H. J., Rots, N. Y., Wijmega-Monsuur, A. J., Sanders, E. A. M., Trzciński, K., Wyllie, A. L., Zwinderman, A. H., van den Berg, L. H., van Rheenen, W., Veldink, J. H., Harboe, Z. B., Lundbo, L. F., de Groot, L. C. P. G. M., van Schoor, N. M., van der Velde, N., Ängquist, L. H., Sørensen, T. I. A., Nohr, E. A., Mentzer, A. J., Mills, T. C., Knight, J. C., du Plessis, M., Nzenze, S., Weiser, J. N., Parkhill, J., Madhi, S., Benfield, T., von Gottberg, A., van der Ende, A., Brouwer, M. C., Barrett, J. C., Bentley, S. D. & van de Beek, D. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nat. Commun.* **10,** (2019).

206. Jindal, H. M., Babu Ramanathan, C. F. Le, Gudimella, R., Manikam, R. & Devi, S. Comparative Genomic Analysis of Clinical Streptococcus Pneumoniae Isolates Reveal 31 New Unique Drug Resistant SNPs Using Whole Genome Sequencing. (2020).

207. Arends, D. W., Miellet, W. R., Langereis, J. D., Ederveen, T. H. A., van der Gaast-de Jongh, C. E., van Scherpenzeel, M., Knol, M. J., van Sorge, N. M., Lefeber, D. J. & Trzciński, K. Examining the distribution and impact of single nucleotide polymorphisms in the capsular locus of Streptococcus pneumoniae serotype 19A. *Infect. Immun.* IAI-00246 (2021).

208. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., Mcvean, G., Durbin, R. & Project, G. The variant call format and VCFtools. *Bioinforma. Appl. NOTE* **27,** 2156–2158 (2011).

209. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11,** e0163962 (2016).

210. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* (2012). doi:10.1038/nmeth.1923

211. Burrows, M., Burrows, M. & Wheeler, D. A Block-Sorting Lossless Data Compression Algorithm. *Digit. SRC Res. Rep.* 12--4 (1994). at <https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.6774>

212. Crochemore, M., Désarménien, J. & Perrin, D. A note on the Burrows-Wheeler transformation. (2005).

213. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

214. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* (2011). doi:10.1093/bioinformatics/btr509

215. Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. & Ruden, D. M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* **6,** 80–92 (2012).

216. Jaroszewski, L., Iyer, M., Alisoltani, A., Sedova, M. & Godzik, A. The interplay of SARS-CoV-2 evolution and constraints imposed by the structure and functionality of its proteins. *PLoS Comput. Biol.* **17,** e1009147 (2021).

217. Alhusain, L. & Hafez, A. M. Nonparametric approaches for population structure analysis.

doi:10.1186/s40246-018-0156-4

218. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K. A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* (2017). doi:10.1371/journal.pcbi.1005752

219. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J. & Sham, P. C. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* (2007). doi:10.1086/519795

220. Wagner, A. F. V., Frey, M., Neugebauer, F. A., Schafer, W. & Knappe, J. The free radical in pyruvate formate-lyase is located on glycine-734. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 996–1000 (1992).

221. Eklund, H. & Fontecave, M. Glycyl radical enzymes: a conservative structural basis for radicals. *Structure* **7,** R257–R262 (1999).

222. Buckwalter, C. M. & King, S. J. Pneumococcal carbohydrate transport: Food for thought. *Trends Microbiol.* (2012). doi:10.1016/j.tim.2012.08.008

223. Loenen, W. A. M., Dryden, D. T. F., Raleigh, E. A. & Wilson, G. G. Type I restriction enzymes and their relatives. *Nucleic Acids Res.* **42,** 20–44 (2014).

224. Murray, N. E., Daniel, A. S., Cowan, G. M. & Sharp, P. M. Conservation of motifs within the unusually variable polypeptide sequences of type I restriction and modification enzymes. *Mol. Microbiol.* **9,** 133–143 (1993).

225. Haering, C. H., Löwe, J., Hochwagen, A. & Nasmyth, K. Molecular Architecture of SMC Proteins and the Yeast Cohesin Complex. *Mol. Cell* **9,** 773–788 (2002).

226. Harvey, S. H., Krien, M. J. & O'connell, M. J. Structural maintenance of chromosomes (SMC) proteins, a family of conserved ATPases. (2002). at <http://genomebiology.com/2002/3/2/reviews/3003.1>

227. Britton, R. A., Chi-Hong Lin, D. & Grossman, A. D. Characterization of a prokaryotic SMC protein involved in chromosome partitioning. (1998). at <www.genesdev.org>

228. Sobral, R. G., Ludovice, A. M., De Lencastre, H. & Tomasz, A. Role of murF in cell wall biosynthesis: Isolation and characterization of a murF conditional mutant of Staphylococcus aureus. *J. Bacteriol.* **188,** 2543–2553 (2006).

229. Rosenow, C., Ryan, P., Weiser, J. N., Johnson, S., Fontan, P., Ortqvist, A. & Masure, H. R. Contribution of novel choline-binding proteins to adherence, colonization and immunogenicity of Streptococcus pneumoniae. *Mol. Microbiol.* **25,** 819–829 (1997).

230. KK, G., ER, M., C, G., EI, T. & HR, M. Role of novel choline binding proteins in virulence of Streptococcus pneumoniae. *Infect. Immun.* **68,** 5690–5695 (2000).

231. Kim, S. K., Makino, K., Amemura, M., Shinagawa, H. & Nakata, A. Molecular analysis of the phoH gene, belonging to the phosphate regulon in Escherichia coli. *J. Bacteriol.* **175,** 1316–1324 (1993).

232. A, P., A, M., A, T., K, S., R, M., S, S., L, G. & DJ, M. Common pathogenic effects of missense mutations in the P-type ATPase ATP13A2 (PARK9) associated with early-onset parkinsonism. *PLoS One* **7,** (2012).

233. MS, F. & CJ, T. Mutants in the CtpA copper transporting P-type ATPase reduce virulence of Listeria monocytogenes. *Microb. Pathog.* **22,** 67–78 (1997).

234. Stincone, A., Prigione, A., Cramer, T., Wamelink, M. M. C., Campbell, K., Cheung, E., Olin-Sandoval, V., Grüning, N.-M., Krüger, A., Alam, M. T., Keller, M. A., Breitenbach, M., Brindle, K. M., Rabinowitz, J. D. & Ralser, M. The return of metabolism: biochemistry and physiology of the pentose phosphate pathway. *Biol. Rev.* **90,** 927–963 (2015).

235. Koonin, E. V. & Tatusove, R. L. Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search. *J. Mol. Biol.* **244,** 125–132 (1994).

236. Fleurie, A., Lesterlin, C., Manuse, S., Zhao, C., Cluzel, C., Lavergne, J. P., Franz-Wachtel, M., MacEk, B., Combet, C., Kuru, E., VanNieuwenhze, M. S., Brun, Y. V., Sherratt, D. & Grangeasse, C. MapZ marks the division sites and positions FtsZ rings in Streptococcus pneumoniae. *Nat. 2014 5167530* **516,** 259–262 (2014).

237. Johnston, C., Campo, N., Bergé, M. J., Polard, P. & Claverys, J. P. Streptococcus pneumoniae, le transformiste. *Trends Microbiol.* **22,** 113–119 (2014).

238. Humpierre, A. R., Zanuy, A., Saenz, M., Garrido, R., Vasco, A. V., Pérez-Nicado, R., Soroa-Milán, Y., Santana-Mederos, D., Westermann, B., Vérez-Bencomo, V., Méndez, Y., García-Rivera, D. & Rivera, D. G. Expanding the Scope of Ugi Multicomponent Bioconjugation to Produce Pneumococcal Multivalent Glycoconjugates as Vaccine Candidates. *Bioconjug. Chem.* **31,** 2231–2240 (2020).

239. Melin, M., Trzciński, K., Antonio, M., Meri, S., Adegbola, R., Kaijalainen, T., Käyhty, H. & Väkeväinen, M. Serotype-Related Variation in Susceptibility to Complement Deposition and Opsonophagocytosis among Clinical Isolates of Streptococcus pneumoniae. *Infect. Immun.* **78,** 5252 (2010).

240. Pang, L., Weeks, S. D. & Van Aerschot, A. Aminoacyl-tRNA Synthetases as Valuable Targets for Antimicrobial Drug Discovery. *Int. J. Mol. Sci. 2021, Vol. 22, Page 1750* **22,** 1750 (2021).

241. Gregston, J., John O'neill, A. & Chopra, I. Prospects for Aminoacyl-tRNA Synthetase Inhibitors as New Antimicrobial Agents. *Antimicrob. Agents Chemother.* **49,** 4821–4833 (2005).

242. Benítez-Páez, A., Villarroya, M. & Armengod, M. E. The Escherichia coli RlmN methyltransferase is a dual-specificity enzyme that modifies both rRNA and tRNA and controls translational accuracy. *RNA* **18,** 1783–1795 (2012).

243. Müller, I. B., Bergmann, B., Groves, M. R., Couto, I., Amaral, L., Begley, T. P., Walter, R. D. & Wrenger, C. The Vitamin B1 Metabolism of Staphylococcus aureus Is Controlled at Enzymatic and Transcriptional Levels. *PLoS One* **4,** (2009).

244. Jackson, R. W., Vinatzer, B., Arnold, D. L., Dorus, S. & Murillo, J. The influence of the accessory genome on bacterial pathogen evolution. *Mob. Genet. Elements* **1,** 55–65 (2011).

245. Azarian, T., Martinez, P. P., Arnold, B. J., Qiu, X., Grant, L. R., Corander, J., Fraser, C., Croucher, N. J., Hammitt, L. L., Reid, R., Santosham, M., Weatherholtz, R. C., Bentley, S. D., O'Brien, K. L., Lipsitch, M. & Hanage, W. P. Frequency-dependent selection can forecast evolution in Streptococcus pneumoniae. *PLoS Biol.* **18,** (2020).

246. Kilian, M. & Tettelin, H. Identification of virulence-associated properties by comparative genome analysis of streptococcus pneumoniae, S. Pseudopneumoniae, S. mitis, three S. oralis subspecies, and S. infantis. *MBio* **10,** (2019).

247. Blomberg, C., Dagerhamn, J., … S. D.-T. J. of & 2009, undefined. Pattern of Accessory Regions and Invasive Disease Potential in Streptococcus pneumoniae. *academic.oup.com* at <https://academic.oup.com/jid/article-abstract/199/7/1032/849610>

248. Maddison, W. P. Testing character correlation using pairwise comparisons on a phylogeny. *J. Theor. Biol.* (2000). doi:10.1006/jtbi.1999.1050

249. Kersey, P. J., Allen, J. E., Allot, A., Barba, M., Boddu, S., Bolt, B. J., Carvalho-Silva, D., Christensen, M., Davis, P., Grabmueller, C., Kumar, N., Liu, Z., Maurel, T., Moore, B., McDowall, M. D., Maheswari, U., Naamati, G., Newman, V., Ong, C. K., Paulini, M., Pedro, H., Perry, E., Russell, M., Sparrow, H., Tapanari, E., Taylor, K.,

Vullo, A., Williams, G., Zadissia, A., Olson, A., Stein, J., Wei, S., Tello-Ruiz, M., Ware, D., Luciani, A., Potter, S., Finn, R. D., Urban, M., Hammond-Kosack, K. E., Bolser, D. M., De Silva, N., Howe, K. L., Langridge, N., Maslen, G., Staines, D. M. & Yates, A. Ensembl Genomes 2018: An integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gkx1011

250. Sa´nchez, A. R., Sa´nchez-Beato, S., Garci´a, E., Garci´a, G., Lo´pez, R., Lo´pez, L. & Garci´a, J. L. *Identification and Characterization of IS1381, a New Insertion Sequence in Streptococcus pneumoniae*. *J. Bacteriol.* **179,** (1997).

251. Mura, A., Fadda, D., Perez, A. J., Danforth, M. L., Musu, D., Rico, A. I., Krupka, M., Denapaite, D., Tsui, H. C. T., Winkler, M. E., Branny, P., Vicente, M., Margolin, W. & Massidda, O. Roles of the essential protein FtsA in cell growth and division in Streptococcus pneumoniae. *J. Bacteriol.* **199,** (2017).

252. McAllister, L. J., Ogunniyi, A. D., Stroeher, U. H. & Paton, J. C. Contribution of a Genomic Accessory Region Encoding a Putative Cellobiose Phosphotransferase System to Virulence of Streptococcus pneumoniae. *PLoS One* **7,** e32385 (2012).

253. Zhang, J. R., Idanpaan-Heikkila, I., Fischer, W. & Tuomanen, E. I. Pneumococcal licD2 gene is involved in phosphorylcholine metabolism. *Mol. Microbiol.* **31,** 1477–1488 (1999).

254. Johnston, C., Mortier-Barrière, I., Granadel, C., Polard, P., Martin, B. & Claverys, J. P. RecFOR Is Not Required for Pneumococcal Transformation but Together with XerS for Resolution of Chromosome Dimers Frequently Formed in the Process. *PLOS Genet.* **11,** e1004934 (2015).

255. Kumar, S., Cheng, X., Klimasauskas, S., Sha, M., Posfai, J., Roberts, R. J. & Wilson, G. G. The DNA (cytosine-5) methyltransferases. *Nucleic Acids Res.* **22,** 1–10 (1994).

256. Williams, L., Stapleton, F. & Carnt, N. Microbiology, lens care and maintenance. *Contact Lenses* 65–96 (2019). doi:10.1016/B978-0-7020-7168-3.00004-0

257. Middleton, D. R., Aceil, J., Mustafa, S., Paschall, A. V. & Avci, F. Y. Glycosyltransferases within the psrp locus facilitate pneumococcal virulence. *J. Bacteriol.* **203,** (2021).

258. Weigel, W. A. & Dersch, P. Phenotypic heterogeneity: a bacterial virulence strategy. *Microbes Infect.* **20,** 570–577 (2018).

259. Brown, J. S., Gilliland, S. M., Spratt, B. G. & Holden, D. W. A Locus Contained within A Variable Region of Pneumococcal Pathogenicity Island 1 Contributes to Virulence in Mice. *Infect. Immun.* **72,** 1587–1593 (2004).

260. Declerck, N., Bouillaut, L., Chaix, D., Rugani, N., Slamti, L., Hoh, F., Lereclus, D. & Arold, S. T. Structure of PlcR: Insights into virulence regulation and evolution of quorum sensing in Gram-positive bacteria. (2007). at <www.pnas.org/cgi/content/full/>

261. Allard, S. T. M., Giraud, M. F. & Naismith, J. H. Epimerases: Structure, function and mechanism. *Cell. Mol. Life Sci.* (2001). doi:10.1007/PL00000803

262. Walters, D. M., Stirewalt, V. L. & Melville, S. B. Cloning, sequence, and transcriptional regulation of the operon encoding a putative N-acetylmannosamine-6-phosphate epimerase (nanE) and sialic acid lyase (nanA) in Clostridium perfringens. *J. Bacteriol.* (1999).

263. Bakeeva, L. E., Chumakov, K. M., Drachev, A. L., Metlina, A. L. & Skulachev, V. P. The sodium cycle. III. Vibrio alginolyticus resembles Vibrio cholerae and some other vibriones by flagellar motor and ribosomal 5S-RNA structures. *BBA - Bioenerg.* (1986). doi:10.1016/0005-2728(86)90115-5

264. Saier, M. H. Families of transmembrane sugar transport proteins. *Mol. Microbiol.* (2000). doi:10.1046/j.1365-2958.2000.01759.x

265. Boyer, P. D. THE ATP SYNTHASE—A SPLENDID MOLECULAR MACHINE. *Annu. Rev. Biochem.* (2002). doi:10.1146/annurev.biochem.66.1.717

266. Bensing, B. A., Gibson, B. W. & Sullam, P. M. The Streptococcus gordonii Platelet Binding Protein GspB Undergoes Glycosylation Independently of Export. *J. Bacteriol.* (2004). doi:10.1128/JB.186.3.638-645.2004

267. Bensing, B. A. & Sullam, P. M. Transport of preproteins by the accessory Sec system requires a specific domain adjacent to the signal peptide. *J. Bacteriol.* (2010). doi:10.1128/JB.00373-10

268. Yamaguchi, M., Terao, Y., Mori-Yamaguchi, Y., Domon, H., Sakaue, Y., Yagi, T., Nishino, K., Yamaguchi, A., Nizet, V. & Kawabata, S. Streptococcus pneumoniae Invades Erythrocytes and Utilizes Them to Evade Human Innate Immunity. *PLoS One* (2013). doi:10.1371/journal.pone.0077282

269. Bandara, M., Skehel, J. M., Kadioglu, A., Collinson, I., Nobbs, A. H., Blocker, A. J. & Jenkinson, H. F. The accessory Sec system (SecY2A2) in Streptococcus pneumoniae is involved in export of pneumolysin toxin, adhesion and biofilm formation. *Microbes Infect.* (2017). doi:10.1016/j.micinf.2017.04.003

270. Wu, R. & Wu, H. A molecular chaperone mediates a two-protein enzyme complex and glycosylation of serine-rich streptococcal adhesins. *J. Biol. Chem.* (2011). doi:10.1074/jbc.M111.239350

271. Bidossi, A., Mulas, L., Decorosi, F., Colomba, L., Ricci, S., Pozzi, G., Deutscher, J., Viti, C. & Oggioni, M. R. A functional genomics approach to establish the complement of carbohydrate transporters in Streptococcus pneumoniae. *PLoS One* (2012). doi:10.1371/journal.pone.0033320

272. Saier, M. H. The Bacterial Phosphotransferase System: New Frontiers 50 Years after Its Discovery. *J. Mol. Microbiol. Biotechnol.* (2015). doi:10.1159/000381215

273. Ndlangisa, K. M., Du Plessis, M., Wolter, N., De Gouveia, L., Klugman, K. P. & Von Gottberg, A. Population snapshot of streptococcus pneumoniae causing invasive disease in South Africa prior to introduction of pneumococcal conjugate vaccines. *PLoS One* (2014). doi:10.1371/journal.pone.0107666

274. Chiba, N., Morozumi, M., Shouji, M., Wajima, T., Iwata, S. & Ubukata, K. Changes in capsule and drug resistance of pneumococci after introduction of PCV7, Japan, 2010-2013. *Emerg. Infect. Dis.* (2014). doi:10.3201/eid2007.131485

275. Kim, S. H., Chung, D. R., Song, J. H., Baek, J. Y., Thamlikitkul, V., Wang, H., Carlos, C., Ahmad, N., Arushothy, R., Tan, S. H., Lye, D., Kang, C. I., Ko, K. S. & Peck, K. R. Changes in serotype distribution and antimicrobial resistance of Streptococcus pneumoniae isolates from adult patients in Asia: Emergence of drug-resistant non-vaccine serotypes. *Vaccine* (2020). doi:10.1016/j.vaccine.2019.09.065

276. Alizadeh Chamkhaleh, M., Esteghamati, A., Sayyahfar, S., Gandomi-Mohammadabadi, A., Balasi, J., Abdiaei, H., Moradi, Y. & Moradi-Lakeh, M. Serotype distribution of Streptococcus pneumoniae among healthy carriers and clinical patients: a systematic review from Iran. *Eur. J. Clin. Microbiol. Infect. Dis.* (2020). doi:10.1007/s10096-020-03963-z

277. Gabarrot, G. G., Vega, M. L. P., Giffoni, G. P. R., Ndez, S. H., Cardinal, P., Lix, V. F., Gabastou, J. M. & Camou, T. Effect of pneumococcal conjugate vaccination in uruguay, a middle-income Country. *PLoS One* (2014). doi:10.1371/journal.pone.0112337

278. Wijayasri, S., Hillier, K., Lim, G. H., Harris, T. M., Wilson, S. E. & Deeks, S. L. The shifting epidemiology and serotype distribution of invasive pneumococcal disease in Ontario, Canada, 2007-2017. *PLoS One* (2019). doi:10.1371/journal.pone.0226353

279. Kandasamy, R., Voysey, M., Collins, S., Berbers, G., Robinson, H., Noel, I., Hughes, H., Ndimah, S., Gould, K., Fry, N., Sheppard, C., Ladhani, S., Snape, M. D., Hinds, J. & Pollard, A. J. Persistent circulation of vaccine serotypes and serotype replacement after 5 years of infant immunization with 13-valent pneumococcal conjugate vaccine in the United Kingdom. *J. Infect. Dis.* (2020). doi:10.1093/infdis/jiz178

280. Chaguza, C., Cornick, J. E., Andam, C. P., Gladstone, R. A., Alaerts, M., Musicha, P., Peno, C., Bar-Zeev, N., Kamng'ona, A. W., Kiran, A. M., Msefula, C. L., McGee, L., Breiman, R. F., Kadioglu, A., French, N., Heyderman, R. S., Hanage, W. P., Bentley, S. D. & Everett, D. B. Population genetic structure, antibiotic resistance, capsule switching and evolution of invasive pneumococci before conjugate vaccination in Malawi. *Vaccine* (2017). doi:10.1016/j.vaccine.2017.07.009

281. Kurupati, R., Kossenkov, A., Haut, L., Kannan, S., Xiang, Z., Li, Y., Doyle, S., Liu, Q., Schmader, K., Showe, L. & Ertl, H. Race-related differences in antibody responses to the inactivated influenza vaccine are linked to distinct pre-vaccination gene expression profiles in blood. *Oncotarget* **7,** 62898 (2016).

282. Chaguza, C., Cornick, J. E., Harris, S. R., Andam, C. P., Bricio-Moreno, L., Yang, M., Yalcin, F., Ousmane, S., Govindpersad, S., Senghore, M., Ebruke, C., Du Plessis, M., Kiran, A. M., Pluschke, G., Sigauque, B., McGee, L., Klugman, K. P., Turner, P., Corander, J., Parkhill, J., Collard, J. M., Antonio, M., von Gottberg, A., Heyderman, R. S., French, N., Kadioglu, A., Hanage, W. P., Everett, D. B. & Bentley, S. D. Understanding pneumococcal serotype 1 biology through population genomic analysis. *BMC Infect. Dis.* **16,** (2016).

283. Jacques, L. C., Panagiotou, S., Baltazar, M., Senghore, M., Khandaker, S., Xu, R., Bricio-Moreno, L., Yang, M., Dowson, C. G., Everett, D. B., Neill, D. R. & Kadioglu, A. Increased pathogenicity of pneumococcal serotype 1 is driven by rapid autolysis and release of pneumolysin. *Nat. Commun. 2020 111* **11,** 1–13 (2020).

284. Obert, C., Sublett, J., Kaushal, D., Hinojosa, E., Barton, T., Tuomanen, E. I. & Orihuela, C. J. Identification of a candidate Streptococcus pneumoniae core genome and regions of diversity correlated with invasive pneumococcal disease. *Infect. Immun.* **74,** 4766–4777 (2006).

Appendices

Appendix 1

Source code

https://docs.google.com/document/d/1hZrbIyoV0XNREjR-dmucKAoYJ5vvIYVs/edit?usp=sharing&ouid=111752898007117861241&rtpof=true&sd=true

Appendix 2

Pan genes

https://docs.google.com/spreadsheets/d/1DqE9DVgWVk7vUQkX4GQ1OquxuC0KQgsn/edit?usp=sharing&ouid=111752898007117861241&rtpof=true&sd=true

Pan-genome reference sequences

https://drive.google.com/file/d/1vAPutBzTq4DqWuDemsunJc1wiW-Gu78B/view?usp=sharing

Appendix 3

Pan IGRs

https://docs.google.com/spreadsheets/d/12LAxcmOeEfPuOZHkwYZYRJoXHa212Fe8/edit?usp=sharing&ouid=111752898007117861241&rtpof=true&sd=true

Pan IGR reference sequences

https://drive.google.com/file/d/1nowdPiZR9wNp_tQjVPK-fD-lJAFDSeyw/view?usp=sharing

Appendix 4

Results of binomial test to identify conserved and polymorphic core genes

https://docs.google.com/spreadsheets/d/1iWnWsMZ3VUArLHbxfradgYp6l4O8DWn0/edit?usp=sharing&ouid=111752898007117861241&rtpof=true&sd=true

Appendix 5

Significant SNPs in serotype 12F

https://docs.google.com/spreadsheets/d/1q3Dv1pyp_zIJf7Ys9W9YtzrrrR79l4sr/edit?usp=sharing&ouid=111752898007117861241&rtpof=true&sd=true

Appendix 6

Significant SNPs in serotype 5

https://docs.google.com/spreadsheets/d/1m9LKyQOH5aKyTwheHuhUApsILa-V5ma7/edit?usp=sharing&ouid=111752898007117861241&rtpof=true&sd=true

Appendix 7

Significant SNPs in serotype 1

https://docs.google.com/spreadsheets/d/1M42_hllCDbXRI7ZZMb1alkJui_TX1Fe2/edit?usp=sharing&ouid=111752898007117861241&rtpof=true&sd=true

Appendix 8

Gene presence-absence 12F vs. 19F

https://docs.google.com/spreadsheets/d/1fORlA1uTxudAn-hiPjRWjGZh5YntWdEf/edit?usp=sharing&ouid=111752898007117861241&rtpof=true&sd=true

Appendix 9

Gene presence-absence 5 vs. 19F

https://docs.google.com/spreadsheets/d/1spMmuItiMRScHuelDfoLEZmz7wekG5pG/edit?usp=sharing&ouid=111752898007117861241&rtpof=true&sd=true

Appendix 10

Gene presence-absence 1 vs. 19F

https://docs.google.com/spreadsheets/d/1g9iqjVvcgN3a8JY7e86VVpDUNplm2kWf/edit?usp=sharing&ouid=111752898007117861241&rtpof=true&sd=true

Appendix 11

Gene presence-absence 23F vs. Others

https://docs.google.com/spreadsheets/d/1ubamD-5DoTlFH_bggu4DDw5O2WqVf458/edit?usp=sharing&ouid=111752898007117861241&rtpof=true&sd=true