Sex and Ageing Effects on Pneumococcal Carriage

Fernando Marcon Passos

2021-04-01

Contents

Preface

Put abstract here!

Chapter 1

Introduction

1.1 Pneumonia and the Pneumococcal diseases

Pneumococcal diseases (PD) constitutes several infectious diseases sharing the same causal factor: *Streptococcus pneumoniae* (Spn). Examples of such illnesses can range from harmless infections in the middle ear (*otitis*) and sinus (*sinusitis*) to more serious and life-threatening pneumonia, meningitis and sepsis. Despite most of such infections be quite common and mild, complications may occur, increasing the severity, originating serious health problems and impairing the chances of survival. Such a scenario is more likely to occur to the parcels of the population with higher exposure to risk factors, vulnerabilities or compromised overall health. Indeed, 90% of deaths caused by pneumonia occurs in countries with low- to middle-income levels, the worst scenario situated in African countries. This might be associated with the high number of people immunocompromised due to HIV infection.

Pneumonia is one of the most common infectious diseases worldwide, known for its significant mortality and frequent need for intensive care support, once respiratory insufficiency and involvement of multiple organs are common complications (Leoni & Rello, 2017) . Especially among children, elderly and patients with pre-morbid conditions, is associated with a high risk of mortality (Fine et al., 1996) .

The burden of pneumonia to the human population can be grasped by the fact that it is the leading cause of death in children worldwide, killing more children bellow 5 years than any other diseases, and that its causative agent is enlisted as one of the 12 priority pathogens since 2017 according to the World Health Organization (WHO, 2017). Furthermore, community-acquired pneumonia (CAP), the most common type of pneumonia, is ranked as the third major cause of death and sepsis in developed countries (Niederman et al., 2001).

1.2 Streptococcus pneumoniae

Streptococcus pneumoniae (Spn) is the leading causes of pneumococcal diseases, being responsible for a variety of related infections. These pathogens are Gram-positive bacteria, coated by a polysaccharide capsule which protect them from being phagocyted. More than 90 chemically and immunologically distinct types were already described (Brugger et al., 2016), several of which can cause several infections types of pneumococcal diseases (O'Brien et al., 2009).

A significant proportion of the human population is carriers of Spn in their nasal mucosa, with up to 27-65% of children and less 10% of the adults. Even though they are opportunistic pathogens (Orihuela et al., 2009), its transmission is highly dependent on the carriage, which is a commensal type of relationship. The success of transmission requires close contact with a carrier(s), with the chances of success increase if the carriers are very young or during drier and colder seasons due to increased fluid secretions in the airways, and also because it's more likely to occur in conjunction with viral infections of the upper respiratory tract (URT).

Once transmitted to the airways of a new host by several possible means, they can spread by the throat to both upper and lower respiratory tracts. Depending on its spreading patterns, can be classified into noninvasive and invasive types. The first one colonizes the mucosal superficies of URTs, being less serious because it remains on the outsides of major organs or the blood. Can cause inflammation of the middle ear (otitis media) and sinus (sinusitis) (O'Brien et al., 2009). However, the second and worst type can invade major organs, causing bacteremia, inflammation of the meninges (meningitis) and even life-threatening infection responses by the body (sepsis) and lung diseases (pneumonia). Sometimes might also infect the bones (osteomyelitis) and joints (arthritis).

1.3 Human Protection to Pneumococcal Diseases

The host defense mechanisms against pneumococcus pulmonary infection involve both innate and adaptive immune systems, the first one being rapid but unspecific and the second one being specific but slower.

The first line of defence consists mostly in physical barriers, such as mucus, and also phagocytic and inflammatory responses elicited by the respiratory tract (Wilson et al., 2015). The pathogen is recognized through pattern recognition receptors (PRRs), which can identify important microbial structures like Pathogen Associated Molecular Patterns (PAMPs) and Danger-Associated Molecular Patterns (DAMPs) (Koppe, Suttorp, & Opitz, 2012). Not only the well-studied Toll-like receptors (TLRs, Geijtenbeek & Gringhuis, 2009) but also other non-TLR families of innate receptors play critical roles in innate

sensing of pathogens. These are C type lectin-like receptors (Geijtenbeek & Gringhuis, 2009), nucleotide-binding oligomerization domain-like receptors (Ting, Duncan, & Lei, 2010), and retinoic acid-inducible gene I (RIG-I)-like receptors (Wilkins & Gale, 2010).

These receptors, when activated, leads to increased production of inflammatory cytokines (e.g., TNF , IFN / and IL-6) through the activation of transcription factors like NF-kB (Opitz, van Laak, Eitel, & Suttorp, 2010). This activation, coupled with the recognition of Spn by macrophages, increases the inflammatory process, further recruiting other inflammatory cells (Opitz, Van Laak, Eitel, & Suttorp, 2010). Recruited neutrophils are also of extreme importance for the first host response to the pathogen. This inflammatory process in the lungs leads to a systemic response, with increased serum levels of components from the complement system and the ultimate activation of the adaptive system.

This activation is required to the complete clearance of Spn in humans given that it can prevent pneumonia and other invasive diseases (Martinot et al., 2014; McCool, Cate, Moy, & Weiser, 2002), where both humoral (antibody) and cellular mediate responses have equal importance in controlling colonization and development of invasive disease. Despite the polysaccharide capsule being recognized as an antigen, bacterial proteins are also capable of eliciting a protective response by recruiting T lymphocytes, which induces B cell antibody response and memory B cells generation (Coutinho & Möller, 1973).

The cellular mediated response by T lymphocytes are subdivided into Th1, Th2, Th17 and regulatory T cells, all of which are of major importance to Spn infection, especially the CD4-positive helper T cells. Th1 cells secrete important cytokines, like IFN , that can enhance intracellular killing by macrophages (Periselneris, 2014). Th17 cells have a fundamental role against extracellular pathogens by secreting IL-17 cytokine, increasing the recruitment of neutrophils and macrophages to the infection site. Moreover, has suggestive cross-serotype protection due to its ability in recognizing protein antigens (Moffitt et al., 2011). The anti-inflammatory role of regulatory T cells enables the control of the immune response to self and foreign particles by secreting IL-10 and reducing IFN production, which has been demonstrated to prevent bacteria penetration of the epithelium to reach the bloodstream, which can cause septicemia during pneumonia (Neill et al., 2012). Although the role of the Th2 responses in pneumococcal infection is not well comprehended, it is known that they secrete cytokines involved in antibody production, eosinophils activation and inhibition of phagocytes function.

1.4 The Effects of Sex and Ageing on Human Immunity

1.5 The Experimental Human Pneumococcal Carriage Model

Anyone can be infected with Spn; however, not everybody can develop pneumococcal carriage: some individuals have a higher susceptibility than others (Ferreira, Jambo, & Gordon, 2011). This can be due to several factors, both intrinsic as extrinsic to the host biology. Understanding what makes someone more susceptible than others to develop carriage is central to develop appropriate vaccines. Although several of those risk factors are well known, not all of them are truly understood and others remain to be uncovered. Further investigation is mandatory to uncover the essential factors as well as how they orchestrated with our immune system and Spn infection, as well when interacting with the influenza virus and vaccines.

To such, the Experimental Human Pneumococcal Carriage (EHPC) model was developed (Ferreira, Jambo, & Gordon, 2011). It is a framework to safely explore the mechanisms and dynamics of the disease: human volunteers are inoculated intranasally with a serotype 6B strain leading to successful nasopharyngeal colonization in approximately half of the subjects (Niederman et al., 2001). The research is mainly focused on the overall host responses in the nasal mucosa to the pneumococcal carriage, specific responses of B and T cell in the lungs and blood, as well the host-pathogen interactions. Such information is obtained from new methods of mucosal nasal sampling, that are used to investigate cellular responses to carriage.

In this model was observed that previous colonization prevented recolonization with the same strain (Niederman et al., 2001), where all recruited subjects had detectable Immunoglobulin G (IgG) levels to Spn antigens before its inoculation. However, anti-pneumococcal IgG levels did not correlate with the success of the colonization (Niederman et al., 2001). In contrast, antibodies to the Spn surface protein (PspA) were only detected in subjects that developed carriage after challenge (Niederman et al., 2001) and correlated with prevention of successful colonization, suggesting that anti-protein antibody may prevent colonization (McCool et al., 2002). IL-17 secreting CD4+ cells specific to Spn were also found in the lung before exposure to the 6B strain, increasing their levels to 8- and 17-fold in the blood and bronchoalveolar fluid 17of volunteers successfully colonized (Neill et al., 2012).

1.6 Systems Vaccinology

The immune system can sense and control external threats through several signalling systems. Such mechanisms can confer protection by an orchestrated cascade of events originating from the infection environment to the lowest molecular levels, where the components are perturbed by the incoming signal. The perturbed components then process and integrate the signal into response, which

is reversely propagated in a bottom-up fashion towards the environment. Different immune responses have their own transcriptional patterns or molecular signatures rapidly induced after the challenge. These molecular signatures correlate with and predict further protective immune responses, which means that its characterization can be a great strategy to uncover molecular mechanisms or even prospectively determine vaccine efficacy (Pulendran, Li, & Nakaya, 2010).

Despite the apparent conceptual simplicity, there are more than 26,000 genes in our genomes and the challenge of a pathogen in the host body could perturb the expression of a substantial fraction of them (Pulendran, Li, & Nakaya, 2010), a fraction that could still present extremely difficult if using traditional reductionist approaches. Systemic approaches, however, turn this problem tractable by providing us with a global picture of the biological response to such challenges (Pulendran, 2014). With new technologies for measuring the behaviour of different layers of biological systems, such as genes, molecules, cells and so forth, and coupled with the latest advances in computational and mathematical tools for dealing with such complexity, we have an unprecedented opportunity to understand the fundamental features involved in such questions (Pulendran, 2014; Pulendran, Li, & Nakaya, 2010).

Systems biology is an interdisciplinary approach that systematically describes the complex interactions between all the parts in a biological system, to elucidate biological rules underlying the behaviour of the biological system (Pulendran, Li, & Nakaya, 2010, Kitano, 2002). Data are collected for different components simultaneously, representing different levels of the system under different perturbations or temporal stage. They are further integrated to generate a model that could describe or predict the behaviour of such mechanisms during different scenarios, aiming to the comprehensive understanding of biological network nature (Ideker et al., 2001; Kitano, 2002; Pulendran et al., 2010).

Such networks are reductive representations lower functional units of complex biological processes, such as nutrient signalling, immune response and energy production, being able to receive, compute, integrate and communicate information from the inside (genome) to the outside (environment) of a biological system (Ideker et al., 2001; Kitano, 2002; Pulendran et al., 2010). Through these models, is possible to evaluate their nature by its dynamics, robustness and plasticity upon the influence of a defiant environment.

This approach has two broad applications in the context of the infectious diseases, with distinct methodologies and rationales: scientific discovery and prediction of immunogenicity and efficacy of vaccines. For instance, to comprehend mechanisms of innate and adaptive immunity in various organisms (Aderem & Hood, 2001; Haining et al., 2008; Haining & Wherry, 2010; Kaech, Wherry, & Ahmed, 2002; N. Subramanian, Torabi-Parizi, Gottschalk, Germain, & Dutta, 2015; Wherry et al., 2007; Zak & Aderem, 2009), including humans (Aderem & Hood, 2001; N. Subramanian et al., 2015), and for identification of infectious diseases biomarkers with diagnostic purposes (Chaussabel et al., 2008; Lee et al., 2008; Otaegui et al., 2009; Ramilo et al., 2007).

Chapter 2

Objectives

The goal of this study is to evaluate how sex and ageing might affect the immune responses to Spn carriage development and its control, to understand intrinsic factors that may give protection for some carriers but increase susceptibility to diseases in others. More specifically,

This will be done by identifying the components and describing their relationships of different layers of human biology during the response to Spn, and further modelling the complexity of the overall response by integrating each level. At first, a single-omics approach will be used to upon each cohort and data type to:

- profile the overall immune response of adults to Spn infection
- profile the overall immune response of adults to Spn infection, stratified by sex
- Identify sex-specific components and/or behaviors in immune response profiles
- profile the overall immune response of elderlies to Spn infection
- Identify age-specific components and/or behaviors in immune response profiles among elderlies and adults
- profile the overall immune response of elderlies to Spn infection, stratified by sex
- Identify sex-specific components and/or behaviors in immune response profiles among elderlies
- Identify age-specific components and/or behaviors in immune response profiles among elderlies and adults of same-sex

- Select the main components affected by sex and age to Spn infection

With the main components selected, different layers of the nasal mucosa will be integrated with multi-omics approaches to describe the main components and its interactions related to sex and ageing factors during Spn infection, carriage development and influenza vaccine interactions.

Chapter 3

Methods

3.1 Overall Experimental Design, Data and Analysis

3.1.1 Experimental Designs and Data Measurements

To describe and identify the immunological mechanisms involved in the development of pneumococcal carriage this project will analyze data from different cohorts of the EHPC consortium. The datasets selected comprise of 5 different cohorts, each with an experimental design allow to identify and describe the main components involved in the response to Spn infection, development of pneumococcal carriage, its interactions with flu vaccines (TIV and LAIV), as well evaluate how these immunological responses behave between sex and age groups. Different experiments probe a specific level of the human nasal mucosa system such as gene expression, protein production and cytokine signalling, immune cell recruitment, Spn colonization and microbiota composition.

So far, 3 cohorts are fully completed: PILOT, LAIV1 and LAIV2. They are all constituted of healthy adult volunteers, with age ranging from 17 to 48 years old. The first cohort (PILOT) was designed to evaluate the immunological factors involved in the response to Spn infection, containing a relatively small set of 20 volunteers, when compared to other studies. The LAIV1 cohort, performed during the winter of 2015/2016, was developed to evaluate how influenza vaccines, such as trivalent inactivated (TIV) and live attenuated influenza (LAIV) vaccines, affects the development of carriage of 129 volunteers after being inoculated with Spn. The opposite biological question was asked in the following LAIV2 study, in which response to flu vaccines was evaluated regarding prior exposure to Spn. This cohort was performed in the winter of 2016/2017 with a total of 198 volunteers. The remaining two ongoing cohorts have also the same designs but probing the elderly population instead. The first study, namely

Elderly, 82 volunteers from both genders and within 50 and 81 years old were inoculated with Spn and samples were collected at baseline and after inoculation, similarly to PILOT study. The last cohort, still in the development stage, will add the interaction with influenza vaccines to the previous design.

Different types of experiments were performed in all cohorts, measuring multiple levels of the human nasal mucosa: bulk RNA sequencing to represent the gene expression of basal cells, multiplex cytokine assays (Luminex) to measure 30 different immune factors involved in the immune system signalling. Cytometry was performed to estimate the number of cell types present at the nasal mucosa, as well as the density of Spn after inoculation. Virus serotyping and bulk RNA sequencing of the nasal mucosa was also carried aiming to measure nasal microbiota composition.

3.1.2 Overall Data Types and Preprocessing

Only data without any preprocessing or transformation were used, for all cohorts and data types, except for data retrieved from public databases when additional information was required in a given step of the analysis (e.g., GMT files from Reactome database for pathway analysis).

Raw data from RNA-Seq experiments consists of fastq files and were preprocessed with a benchmark pipeline: samples were aligned using the STAR (Dobin et al., 2013) software, transcripts counts calculated with featureCounts (Liao, Smyth, & Shi, 2014) software and, with personalized scripts in bash and R language, samples of a given cohort were merge in a single table, where each row represent a transcript and each column a sample. Both STAR and featureCounts steps were supplied with the Hg38 reference (Herrero et al., 2016; Ruffier et al., 2017) genome and features, represented by Ensembl identifiers (EnsemblID, Aken et al., 2016), were parsed so that the EnsemblID‘s version is omitted. Quality control checks were performed between each step of the preprocessing procedure, with fastQC (A. & Bittencourt a, 2010) on fastq files before alignment and MultiQC (Ewels, Magnusson, Lundin, & Käller, 2016) for all steps with no samples removed due to low quality. Counts data were normalized either with trimmed means of M-values (TMM, Smid et al., 2018) or variance stabilizing transformation (vst) (Zwiener, Frisch, & Binder, 2014)with the DESeq2 (Love, Huber, & Anders, 2014) package from Bioconductor (Ihaka & Gentleman, 1996), when necessary.

For the remaining data types, raw data were retrieved in standard comma/tab-separated formats (e.g. CSV), containing tables representing either absolute measurements, proportions or mean levels of biological variables (e.g. EGF, IL-10 for Luminex or T-Cell Percentage for flow cytometry data) in the columns, for a given volunteer and time-point in the rows. Raw values were shifted by a constant value (all values are added by the absolute minimum value of the dataset) to avoid negative/zero value and then log2 transformed. Depending on the analysis step, each variable could be further scaled by the Z-Score transformation.

Variables or features with more than 25% of missing values were excluded, with the remaining missing data imputed through an unsupervised approach algorithm implemented in the missForest (Stekhoven & Buhlmann, 2012) package.

Each dataset was further inspected for data quality, outliers and technical artefacts or batch effects, and overall variable distribution. This was done with usual exploratory data analysis (EDA), such as dimensionality reduction algorithms (e.g. Principal Component Analysis (PCA) and Independent PCA (IPCA)), clustering techniques (e.g. Hierarchical Clustering (HC)) and visualization tools provided by R packages.

3.1.3 Biological Pathways

With the results obtained from the analysis of DE and co-expression analysis of RNA-Seq data, we get new sets of variables that could potentially have some biological meaning, as explained in the previous section. For instance, after the identification of DEGs obtained in a certain comparison, we will possibly have two lists of genes that are up- and down-regulated. The same is true for the co-expression analysis, in which we will get a different list of genes for each co-expression module identified during the analysis. All those lists should then be further explored through enrichment analysis to identify which biological pathways are most probably related in order to uncover the underlying molecular mechanisms.

To identify which biological pathways could potentially be perturbed in each comparison of a DE analysis, gene set enrichment analysis (A. Subramanian et al., 2005) was performed upon the ranked log2FC values. The Reactome (Croft et al., 2011) gene set (GMT file) from Enrichr (Chen et al., 2013; Kuleshov et al., 2016) database was retrieved from Enrichr database. Also, the EnsemblIDs of each transcript was translated into Gene Symbol IDs with the biomaRt (Durinck, Spellman, Birney, & Huber, 2009) package so the transcripts could be matched between the datasets and Reactome gene set. The actual enrichment was performed with the fgsea function from fgsea (Sergushichev, 2016) package, with all parameters set as default, except for the number of permutations (nperm) that was set to 1000. The enrichment results of each comparison were merged with all comparison in each analysis, and all pathways with padj < 0.05 in at least one of the comparisons were selected to further exploratory analysis. To further explore such selected pathways, common pathways for all groups or time-points were displayed in correlation plots, from corrplot (Wei & Simko, 2017) package, with pathways represented in rows, comparison represented in columns and enrichment (NES) values as circles, coloured in a blue-to-red gradient, representing negative to positive perturbation respectively.

3.2 Proposed Solution

3.2.1 Overall Data Analysis

3.2.1.1 Identification of Perturbed Features

To first answer the most basic question, "which biological variable(s) are possibly related to Spn inoculation?", features that suffered any perturbation were detected through statistical methods which compare their sample distribution after the inoculation to its baseline levels. In RNA-Seq data, this analysis is performed with both DESeq2 and edgeR (Robinson, McCarthy, & Smyth, 2009) R packages and it is known as differential expression analysis (DEA). This perturbation is represented by the ratio of distribution means and displayed as log2 fold-change (log2FC) values. Statistical confidence is also estimated, represented by the p-values (pval), and further adjusted (padj) for multiple comparisons. Deferentially expressed genes (DEG) can be selected by setting thresholds for both log2FCs and p-values, and classified in up-, down-regulated or unperturbed according to its log2FCs direction. In this study DEGs were defined according to the following thresholds: among the genes with padj < 0.01 the ones with log2FC > 0 are classified as UP, log2FC < 0 classified as DOWN and the remaining as UNCHANGED.

Similar statistical tests were applied to the remaining data types using R core functions. These tests can be parametric (e.g. T-Test) or not (e.g. Mann-Whitney-Wilcoxon (MWW test) and are applied accordingly if most of the variables in a given dataset follow a normal distribution or not, respectively. P-values obtained from these tests were corrected for multiple comparisons with Bonferroni method (Armstrong, 2014) and variables with padj below a threshold of 0.01 were selected for further analysis.

3.2.1.2 Identification of Transcriptional Programs and Variables Cluster with Similar Patterns

Beyond identifying perturbed variables, one might want to discover groups of variables that have similar patterns throughout the time-points or between conditions. One example of this type of analysis is co-expression analysis, usually performed with transcriptomic data. Such analysis assumes that genes with similar behaviour or pattern of expression could be biological entities that are working together, potentially representing a biological program and perhaps being orchestrated by common regulators. For example, a given set of highly correlated genes could indicate that they are all member of the apoptosis mechanism and might even be regulated by the same transcriptional factor, long non-coding RNA (lincRNAs) or micro RNA (miRNA). This type of analysis is very powerful to identify major patterns in the data, being them transcriptomic data or not.

In the case of RNA-Seq data, co-expression analysis were performed using the CEMiTool (Russo et al., 2018) package, which allow the identification of co-expression modules (MOD), visualize their overall expression along samples, identify most probable biological pathways that a given MOD could be related to and also analyze how each module are behaving in each condition or time-point, in other words, identify if the given module is up- or down-regulated in each condition/time-point by analyzing if the majority of its constituents (genes/transcripts) are highly enriched among the highly positively or negatively expressed genes.

The same basic idea was also applied to the other types of data in order to identify possible relationship patterns among the biological variables. For example, identifying sets of cytokines (Luminex) or immune cell types (flow cytometry) that are highly correlated between each other, either across time-points or groups (e.g. Spn groups, vaccine type, etc.). This sets of highly correlated variables were determined through the calculation of pairwise correlations (Spearman, Rho) for all variables in each dataset. These pairwise correlations are represented in the form of a correlation matrix, with the number of rows and columns equal to the total number of variables present in the dataset. To define if two variables are connected or not, discrete values of 0s and 1s were used to represent if a connection exists, respectively. Values of 1 were assigned if their absolute correlation values were $|\text{Rho}| >= 0.7$, 0 otherwise. This discrete matrix, also known as an adjacency matrix, was used to create a graph, where the nodes represent biological variables (e.g. cytokines, cell type) and can be connected (edges) only if its adjacency value is equal to 1. Finally, variables were grouped into modules through the Louvain (Newman, 2006; Traag, Waltman, & van Eck, 2019) clustering method, a commonly used community detection algorithm and available in igraph (Csárdi & Nepusz, n.d.) package.

Chapter 4

Applications

Some *significant* applications are demonstrated in this chapter.

4.1 Example one

4.2 Example two

Chapter 5

Final Words