

# Ageing and Sex Differences in Human Pneumococcal Carriage

minimal version

Fernando Marcon Passos

March 2021



# Contents

<b>1</b>	<b>Background Knowledge</b>	<b>7</b>
1.1	Pneumonia and the Pneumococcal diseases . . . . .	7
1.2	Streptococcus pneumoniae . . . . .	8
1.3	Human Protection to Pneumococcal Diseases . . . . .	8
1.4	The Effects of Sex and Ageing on Human Immunity . . . . .	9
1.5	The Experimental Human Pneumococcal Carriage Model . . . . .	10
1.6	Systems Vaccinology . . . . .	10
<b>2</b>	<b>Research Questions</b>	<b>13</b>
<b>3</b>	<b>Project Plan</b>	<b>15</b>
<b>4</b>	<b>Methods</b>	<b>17</b>
4.1	Overall Experimental Design, Data and Analysis . . . . .	18
4.2	Proposed Solution . . . . .	20
4.3	Transcriptomics Analysis Pipeline . . . . .	22
4.4	Network Analysis . . . . .	24
<b>5</b>	<b>Data Driven Analysis of Susceptibility to Pneumococcal Carriage</b>	<b>25</b>
5.1	Q1: Is developing carriage random? . . . . .	25
<b>6</b>	<b>Definitions</b>	<b>31</b>

<b>A Essential Literature</b>	<b>37</b>
A.1 Thesis . . . . .	37
A.2 Articles . . . . .	37
<b>B Pneumococcal Diseases and Immunity</b>	<b>39</b>
B.1 Controlled Human Infection and Rechallenge with <i>Streptococcus pneumoniae</i> Reveals the Protective Efficacy of Carriage in Healthy Adults . . . . .	39
B.2 Inflammation induced by influenza virus impairs innate control of human pneumococcal carriage . . . . .	40
<b>C Sex Differences</b>	<b>43</b>
<b>D Ageing</b>	<b>45</b>
D.1 The Hallmarks of Aging . . . . .	45
D.2 Therapies . . . . .	46
D.3 Senescent Cells . . . . .	47
D.4 The Role of Senescent Cells in Aging . . . . .	48
<b>E Bioinformatics</b>	<b>51</b>
E.1 Bioinformatics Essentials: . . . . .	51
E.2 Differential Expression Analysis . . . . .	55
E.3 Network Analysis . . . . .	63
E.4 Multidimensional scaling . . . . .	66
E.5 Next-Generation Sequencing . . . . .	67
E.6 OMICS . . . . .	73
E.7 Principal Component Analysis . . . . .	77
E.8 MixOmics . . . . .	81
E.9 Normalization . . . . .	101
E.10 Time-Series Analysis . . . . .	102
E.11 Survival analysis . . . . .	109
E.12 Data Integration . . . . .	113
E.13 Single-Cell RNA-Seq . . . . .	120
E.14 Reproducible Data Science . . . . .	120

E.15 machine learning vs DE Analysis . . . . .	124
E.16 Dealing with Missing Data . . . . .	124
E.17 Feature Selection . . . . .	126
E.18 Exploratory Data Anaysis (EDA) . . . . .	126
E.19 Dimensionality Reduction . . . . .	139
E.20 Data Science Skills . . . . .	140
E.21 Cross-Normalization . . . . .	141
E.22 Clustering . . . . .	141
<b>F Systems Biology</b>	<b>143</b>
F.1 Systems Thinking . . . . .	143
F.2 Systems Biology . . . . .	152
F.3 Systems Vaccinology . . . . .	161
F.4 Systems Immunology . . . . .	162
F.5 Artificial Biology . . . . .	162
<b>G Resources</b>	<b>167</b>
G.1 The Human Cell Atlas . . . . .	167
G.2 reproducible data analysis . . . . .	170
G.3 Databases . . . . .	173
<b>H Statistics &amp; Probability</b>	<b>175</b>
H.1 Statistics . . . . .	175
H.2 Probability . . . . .	175
<b>I Virology</b>	<b>177</b>
<b>J Goals and Questions (Original)</b>	<b>179</b>
J.1 Goals . . . . .	179
J.2 Research Questions . . . . .	179
J.3 Specific Questions . . . . .	181
J.4 Technical Questions . . . . .	182
J.5 Sex Effect on Spn Carriage . . . . .	182
J.6 Ageing Effect on Spn Carriage . . . . .	182

Infectious disease is one of the few genuine adventures left in the world. The dragons are all dead and the lance grows rusty in the chimney corner....About the only sporting proposition that remains unimpaired by the relentless domestication of a once free living human species is the war against those ferocious little fellow creatures, which lurk in the dark corners and stalk us in the bodies of rats, mice, and all kinds of domestic animals, which fly and crawl with the insects and waylay us in our food and drink and even in our love.

*Hans Zinsser, 1935*

### **Goal**

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.

### **Outline**

1. Role of Sex in Immune Responses to Pneumococcal Carriage
2. Role of Ageing in Immune Responses to Pneumococcal Carriage
3. Role of Sex and Ageing in the Immunity to Pneumococcal Carriage

# Chapter 1

## Background Knowledge

### 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 immuno-compromised 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 phagocytosed. 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  $\gamma$  and IL-6) through the activation of transcription factors like NF- $\kappa$ B (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  $\gamma$ , that can enhance intracellular killing by macrophages (Periselmanis, 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  $\gamma$  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

# Research Questions

What are the main sex-related factors in healthy adults that

**1. What are the main host factors that predispose susceptibility to pneumococcal carriage in young healthy adults?**

1. What are the main host factors that can predict the development of carriage? *Data:* rna-seq, Luminex, flow, microbiome *Analysis:* Data Integration, Feature selection, Machine Learning, Network analysis
2. What are the main perturbations and gene expression signatures associated to development of carriage? *Data:* rna-seq *Analysis:* DEGs, Co-Expression, MDP, Characteristic Direction, Pathway Analysis, Network analysis
3. What are the main perturbations and immunological response signatures associated to development of carriage? *Data:* luminex, flow *Analysis:* DEGs, Co-expression, MDP, Clustering, Data Integration, Network analysis
4. How does the microbiome affect the development of carriage? *Data:* microbiome *Analysis:* Microbiome analysis, Network analysis
5. 1. How the topology of systems biology network change between individuals that develop carriage? *Data:* all data *Analysis:* Data Integration, Network analysis

**2. How does sex affect such susceptibility?**

1. Is there any perturbation at hormonal levels during carriage development? *Data:* hormone levels *Analysis:* Statistics
2. How the hormonal levels is related with gene expression signatures associated to development of carriage? *Data:* hormone levels, rna-seq *Analysis:* Data Integration, Statistics, Network analysis

3. How are hormonal levels related to immunological responses? Data: hormone levels, rna-seq, luminex, flow cytometry Analysis: Data Integration, Statistics, Network analysis
3. **How ageing can affect the susceptibility to pneumococcal carriage in healthy subjects?**  
Same questions as (1.)
4. **How vaccination affects the transcriptional and immunological signatures of elderly individuals during pneumococcal carriage?**  
Same questions as (1.), with a special focus at the signatures found in (3.)

## Chapter 3

# Project Plan

1. profile the overall immune response of adults to Spn infection;
  - *Data types*: RNA-Seq, Luminex, Spn density, Flow cytometry
  - *Data sets*: Pilot, LAIV2; baseline and after Spn inoculation
2. profile the overall immune response of adults to Spn infection, stratified by sex;
  1. Identify sex-specific components and/or behaviours in immune response profiles
3. profile the overall immune response of elderlies to Spn infection
  1. Identify age-specific components and/or behaviours in immune response profiles among elderlies and adults
  2. profile the overall immune response of elderlies to Spn infection, stratified by sex;
  3. Identify sex-specific components and/or behaviours in immune response profiles among elderlies
  4. Identify age-specific components and/or behaviours in immune response profiles among elderlies and adults of same-sex
    - *Data types*: RNA-Seq, Luminex, Spn density, Flow cytometry
    - *Data sets*: AGES; all timepoints
4. Integrate different cohorts and data types by the main components affected by sex and age to Spn infection
  - *Data types*: RNA-Seq, Luminex, Spn density, Flow cytometry, hormone levels

- *Data sets:* Pilot, LAIV1, LAIV2, AGES; baseline and after Spn inoculation

For items 1 to 3, the major analysis to be performed consists in:

- *Preprocessing:* Raw data will be preprocessed accordingly to its standard pipelines, transformed in log2 values and/or normalized.
- *Quality Control and Exploratory Data Analysis:* Quality control, outlier detection and evaluation of overall data variability will be performed using standard dimensionality reduction techniques, clusterization methods and further displayed using several plot types. Variables/samples with more than 25% of missing values will be removed and the remaining missing data will be imputed using unsupervised algorithms.
- *Detection of perturbed features:* use of statistical tests to select features with statistically significant increase/decrease of overall levels after Spn inoculation when compared to its baseline values
- *Detection of co-regulated features:* construct network of features based on their overall levels similarity and detect cluster of highly correlated features. Their behavior will be assessed by estimating the average cluster levels and compared between groups by time-points.
- *Pathway Analysis:* gene lists from both previous analysis of RNA-Seq data will be enriched for Reactome pathways to infer underlying biological mechanisms involved in each comparison.
- *Component Selection:* For each cohort and data type, major components interrelated to either immune response to Spn inoculation, sex or ageing will be selected for final integration analysis.

The last item will be performed using different data integration frameworks on previous selected components, such as mixOmics, SIMON and MOFA.



## Chapter 4

# Methods

### DATA COLLECTED:

- Nasal cytokines (Dessi)
- Nasal cell flow (Simon & Jesus)
- Nasal cell RNAseq (Fernando & Dessi)
- Anti-CPS Baseline and D29 (Hugh)
- MSD Baseline and D29
- T cell ICS data ? Need to track down original data

### EXPERIMENTS REMAINING:

- Nasal Wash MPO – neutrophil function
- Re-challenge cohort - Anti-CPS Baseline and D29 and rechallenge
- Serum OPK assays 6B(in-house) vs more serotypes (Goldblatt)
- B cell phenotyping – for senescence and memory B cell populations Same as for SPN3 with a few more markers
- B cell ELISPOT same protocol as SPN3

## 4.1 Overall Experimental Design, Data and Analysis

### 4.1.1 Experimental Designs and Data Measurements

To describe and identify the immunological mechanisms involved in the development of Spn 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 Spn 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.

### 4.1.2 Overall Data Types and Preprocessing

Study	Code	Data Availability
MUCOSAL	015	RNASeq Microbiology Nasal Cell Immunophenotyping Nasosorp Cytokines
LAIV1	016	RNASeq Microbiology Nasal Cell Immunophenotyping Nasosorp Cytokines
LAIV2	019	RNASeq Microbiology Nasal Cell Immunophenotyping Nasosorp Cytokines Hormones
AGES	018	Antibody MSD Nasal Cell Immunophenotyping Nasosorp Cytokines Neutrophil Activity RNASeq

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 pre-processed 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. & Bitten-court a, 2010) on fastq files before alignment and MultiQC (Ewels, Magnusson, Lundin, & K  ller, 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.

### 4.1.3 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  $\text{padj} < 0.05$  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 corrrplot (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.

## 4.2 Proposed Solution

### 4.2.1 Overall Data Analysis

#### 4.2.1.1 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. Differentially 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  $\text{padj} < 0.01$  the ones with  $\text{log2FC} > 0$  are classified as UP,  $\text{log2FC} < 0$  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.

#### 4.2.1.2 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.). These 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  $|\text{Rho}| \geq 0.7$ , 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.

## 4.3 Transcriptomics Analysis Pipeline

### 4.3.1 Quality Control

- Density plot
- Clustering
- MA plot
  - microarray - `affycoretools::maplot()` - <https://rdrr.io/bioc/t/man/maplot.html>
  - rnaseq - `DESeq2::plotMA(dds)`
- Distribution (Normality) test
- PVCA
- MDP
- RLE
- Multidimensional scaling (MDS)
- Bi-clustering

- PCA
- IPVCA
- mixOmics

#### 4.3.2 Biomarkers

- DEG Analysis
  - Microarray
    - \* distribution assumptions evaluation
- RNA-Seq
  - DESeq2
  - edgeR
  - limma
- DE analysis:
  - limma
  - t-test
  - ANOVA
  - ranking methods
    - \* see other methods and approaches
  - *p-value* distribution analysis
  - pathway stabilization for DEGs thresholding
- Bi-clustering
  - [https://rpubs.com/crazyhottommy/PCA\\_MDS](https://rpubs.com/crazyhottommy/PCA_MDS)

#### 4.3.3 Pathway Analysis

- ORA
  - Network Analyst
  - Enrichr
  - StringR
- GSEA
- IPA

### 4.3.4 Pipeline

INPUT: - Raw expression data - RNA-Seq: counts - Microarray: .CEL, raw expression table (Illumina) - phenodata

Parameters:

- split\_proportion = c(#train, #test, #validation)
- response\_var = ""
- preliminary\_filter = c(T,F)

OUTPUT: - biomarkers list - my performance assessment - plots

ALGORITHM - Read data - test tables equivalence - split data in train, test and validation sets - (Optional) preliminary filter (e.g, low variance, low mean) - Biomarker detection - Biomarkers: biomarker list extraction - Biomarkers performance test in train/test dataset - Biomarkers performance test in validation dataset - plots

### 4.3.5

## 4.4 Network Analysis

### 4.4.1 Co-expression Analysis

- WGCNA / CEMiTool
- ARACNE
- Topological analysis



## Chapter 5

# Data Driven Analysis of Susceptibility to Pneumococcal Carriage

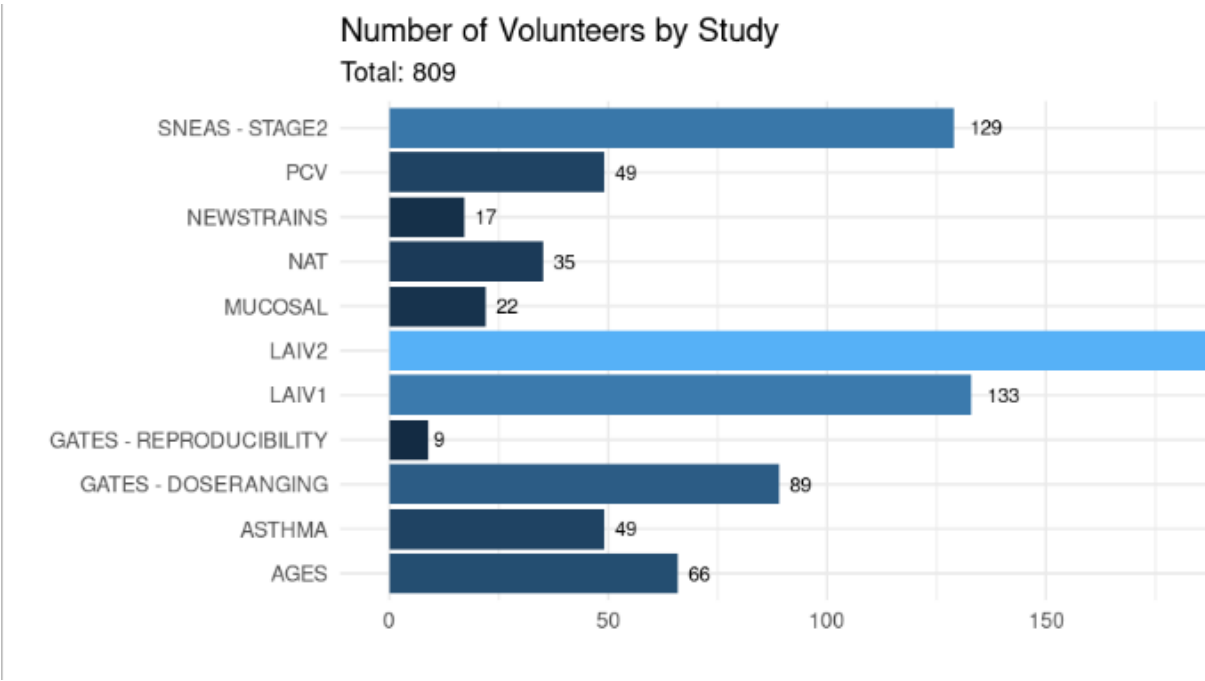
In order to understand the development of Spn carriage it is necessary to know if carriage is some kind of random process. in other words, it is necessary to know if a given subject has the same chance of developing carriage as getting heads in a coin flip after being inoculated by Spn. So, the first question is to know if the probability of developing carriage is different than 0.5.

### 5.1 Q1: Is developing carriage random?

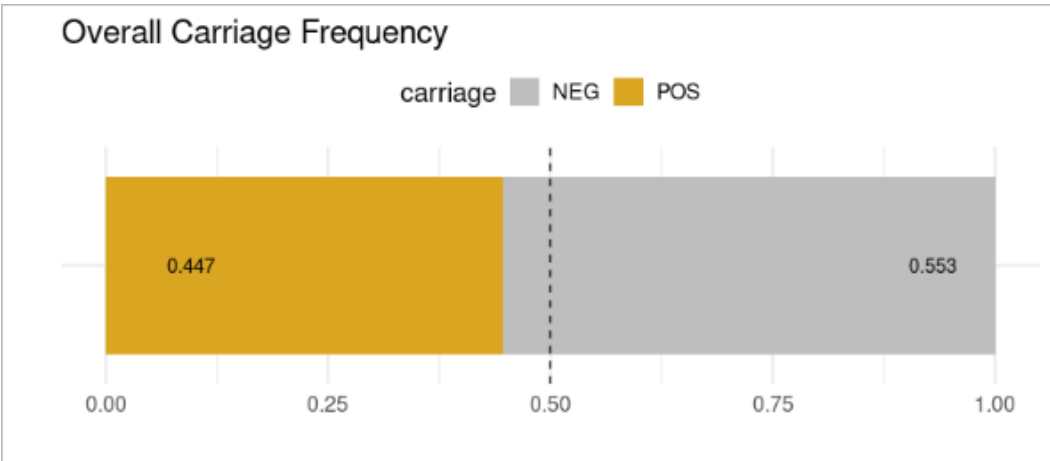
If carriage development is indeed a random process, a given subject would have a probability of 0.5 of becoming carriage positive (POS), as there are only two possible mutually exclusive outcomes: positive (POS) or negative (NEG).

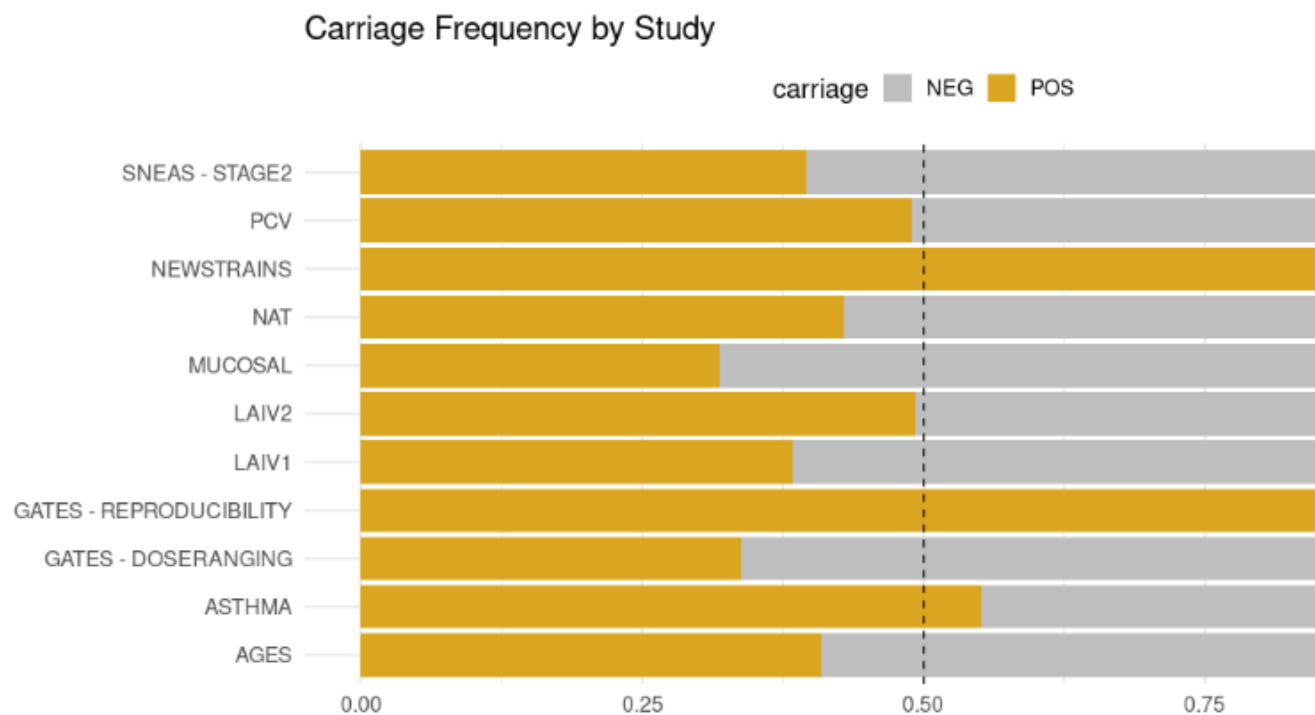
The null hypothesis is that the probability of developing carriage is equal to not developing carriage,  $P(\text{POS}) = P(\text{NEG})$ . The alternative hypothesis is that the probability of developing carriage is different than the probability of not developing carriage,  $P(\text{POS}) > P(\text{NEG})$  or  $P(\text{POS}) < P(\text{NEG})$ .

To test these hypothesis, we can estimate such probabilities based on the frequency of volunteers that became carriage POS in the whole EHPC consortium. EHPC provides the density of Spn present in the nasal mucosa after its inoculation from a total of 809 subjects that took part in several studies, as showed in the following graph (Fig 2.1).

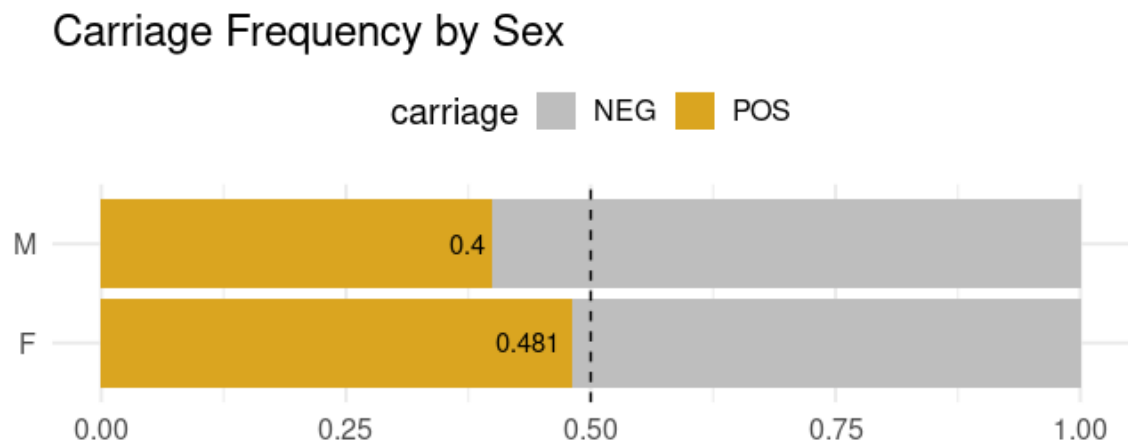


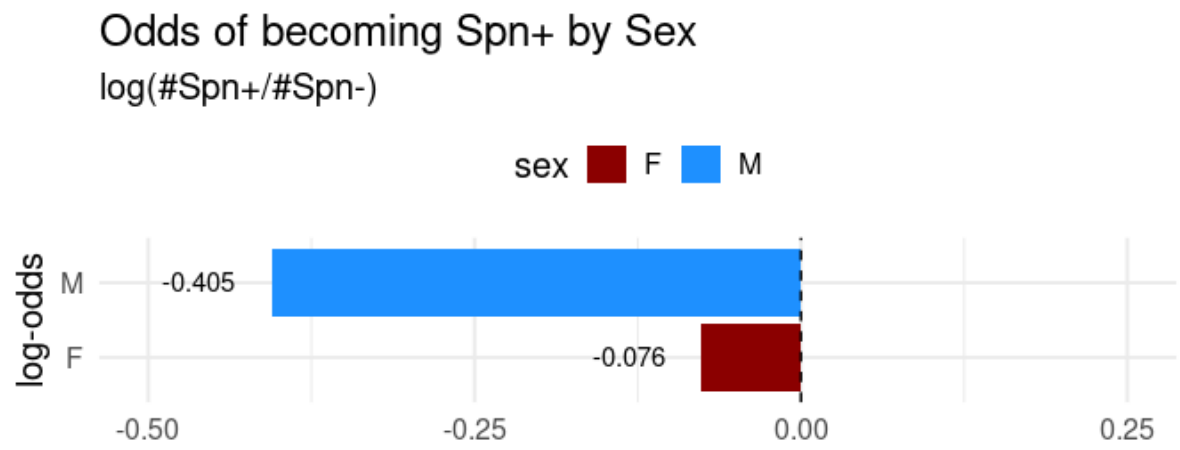
Based on this data, we arrive at an overall probability equal to 0.447 of developing carriage and a probability  $< 0.5$  in the majority of the studies, as depicted in the following figures.

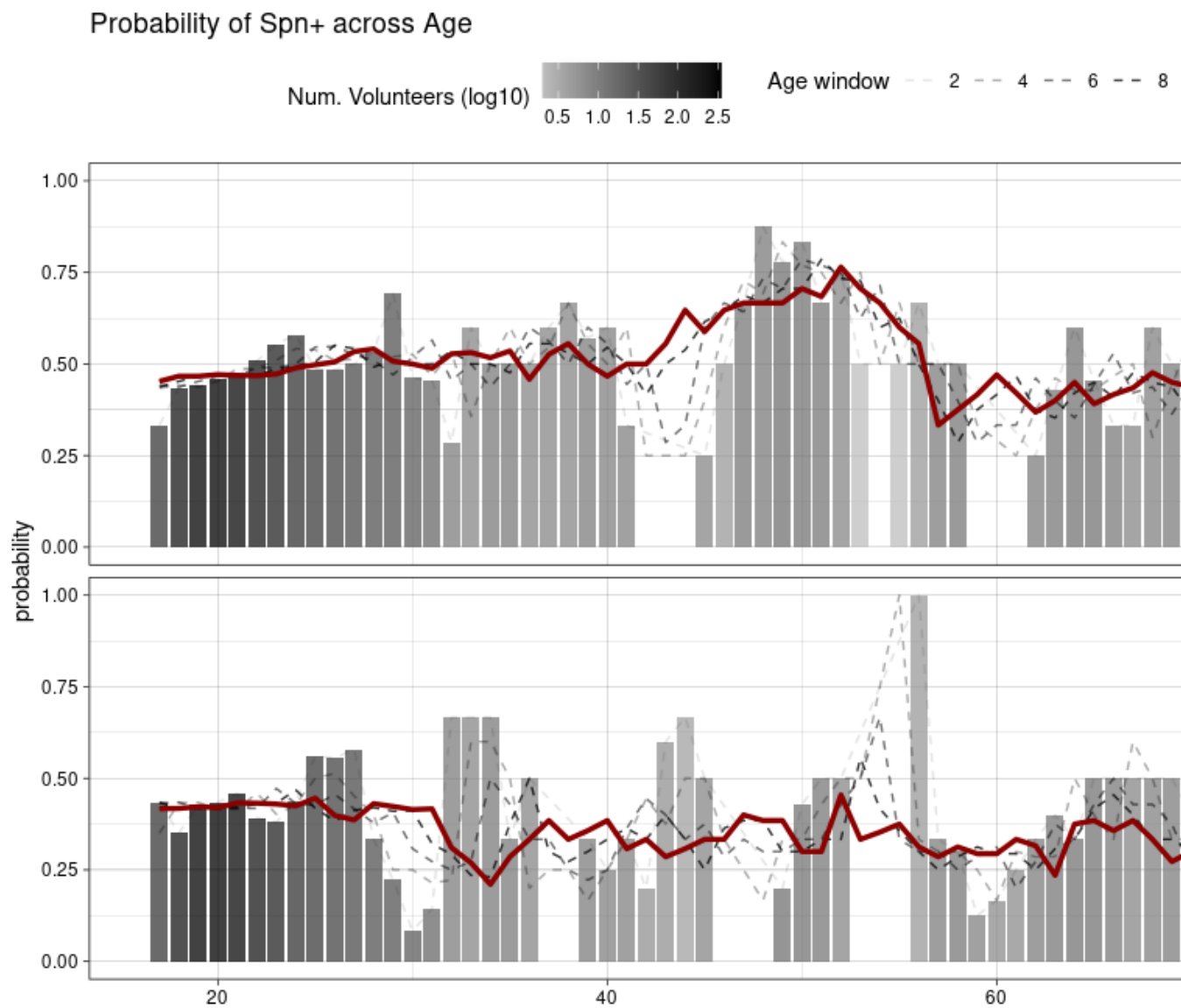




Across all studies women have a higher probability of developing carriage than men (Fig. 2.4), where men have a clear lower (log) odds ratio than women (Fig. 2.5).









## Chapter 6

# Definitions

**cross-over design** is a repeated measurements design such that each experimental unit (patient) receives different treatments during the different time periods, i.e., the patients cross over from one treatment to another during the course of the trial

**w. r. t** with respect to, with regard to

**Dependent/outcome variable** A variable whose value depends upon independent variable, is what is being measured in an experiment or evaluated in a mathematical equation. In the equation  $a = b/c$ , the dependent variable,  $a$ , is determined by the values of  $b$  and  $c$ .

**Response variables** The response variable is the variable you are measuring and trying to explain. When you have a response variable, it is always paired with one or more explanatory variables. The explanatory variable(s) drives change in the response variable.

**Latent variables** As opposed to observable variables, are variables that are not directly observed but are rather inferred (through a mathematical model) from other variables that are observed (directly measured).

**Covariance** is a measure of the joint variability of two random variables. Covariance provides a measure of the strength of the correlation between two or more sets of random variates.

$$Cov(x, y) = \sum_{i=1, \dots, N} (x_i - \mu_x)(y_i - \mu_y) / N$$

**Linear combination** A linear combination is an expression constructed from a set of terms by multiplying each term by a constant and adding the results (e.g. a linear combination of  $x$  and  $y$  would be any expression of the form  $ax + by$ , where  $a$  and  $b$  are constants).

**Independent Component Analysis (ICA)** Is a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements, or signals. In the model, the data variables are assumed to be linear or nonlinear mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables are assumed non-gaussian and mutually independent, and they are called the independent components of the observed data. These independent components, also called sources or factors, can be found by ICA.

**bipartite graphs** also called a bigraph, is a set of graph vertices decomposed into two disjoint sets such that no two graph vertices within the same set are adjacent.

**Discriminant Analysis** i.g., the prediction of group membership from the levels of continuous predictor variables

**Partial Least Squares (PLS) regression** Partial least squares (PLS) regression is a technique that reduces the predictors to a smaller set of uncorrelated components and performs least squares regression on these components, instead of on the original data. PLS regression is especially useful when your predictors are highly collinear, or when you have more predictors than observations and ordinary least-squares regression either produces coefficients with high standard errors or fails completely. PLS does not assume that the predictors are fixed, unlike multiple regression. This means that the predictors can be measured with error, making PLS more robust to measurement uncertainty.

**Multicollinearity (collinearity)** is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. In this situation the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors. That is, a multivariate regression model with collinear predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others.

Multicollinearity occurs when your model includes multiple factors that are correlated not just to your response variable, but also to each other. In other words, it results when you have factors that are a bit redundant.

Multicollinearity increases the standard errors of the coefficients. Increased standard errors in turn means that coefficients for some independent variables may be found not to be significantly different from 0. In other words, by overinflating the standard errors, multicollinearity makes some variables statistically insignificant when they should be significant.



Without multicollinearity (and thus, with lower standard errors), those coefficients might be significant.

Severe multicollinearity is a major problem, because it increases the variance of the regression coefficients, making them unstable. The more variance they have, the more difficult it is to interpret the coefficients.

Warning Signs of Multicollinearity :

- A regression coefficient is not significant even though, theoretically, that variable should be highly correlated with Y.
- When you add or delete an X variable, the regression coefficients change dramatically.
- You see a negative regression coefficient when your response should increase along with X.
- You see a positive regression coefficient when the response should decrease as X increases.
- Your X variables have high pairwise correlations.

Ways to measure multicollinearity: Variance Inflation Factor (VIF): assesses how much the variance of an estimated regression coefficient increases if your predictors are correlated:

- VIF is equal to 1, there is no multicollinearity among factors
- VIF is greater than 1, the predictors may be moderately correlated
- VIF between 5 and 10, indicates high correlation that may be problematic
- VIF above 10, you can assume that the regression coefficients are poorly estimated due to multicollinearity.

Dealing with Multicollinearity (VIF near or above 5):

Remove highly correlated predictors from the model. If you have two or more factors with a high VIF, remove one from the model. Because they supply redundant information, removing one of the correlated factors usually doesn't drastically reduce the R-squared. Consider using stepwise regression, best subsets regression, or specialized knowledge of the data set to remove these variables. Select the model that has the highest R-squared value.

Use Partial Least Squares Regression (PLS) or Principal Components Analysis, regression methods that cut the number of predictors to a smaller set of uncorrelated components.

**Multivariate/Multivariable Analysis** (linear) observation and analysis of more than one statistical outcome variable at a time.

Multivariate is used for the analysis with multiple outcomes/dependent variables, whereas multivariable is used for the analysis with multiple explanatory/independent variables. However, they are sometimes used interchangeably in the literature as not many researchers are aware of the difference.

A simple linear regression model has a continuous outcome and one predictor, whereas a multiple or multivariable linear regression model has a continuous outcome and multiple predictors (continuous or categorical). A simple linear regression model would have the form  $y = a + xb + e$ . By contrast, a multivariable or multiple linear regression model would take the form  $y = a + x_1b_1 + x_2b_2 + \dots + x_kb_k + e$ , where  $y$  is a continuous dependent variable,  $x$  is a single predictor in the simple regression model, and  $x_1, x_2, \dots, x_k$  are the predictors in the multivariable model.

Multivariate, by contrast, refers to the modeling of data that are often derived from longitudinal studies, wherein an outcome is measured for the same individual at multiple time points (repeated measures), or the modeling of nested/clustered data, wherein there are multiple individuals in each cluster. A multivariate linear regression model would have the form

**Orthogonal transformation** An orthogonal transformation is a linear transformation  $T: V \rightarrow V$  which preserves a symmetric inner product: preserves lengths of vectors and angles between vectors

**Principal Component Analysis** Linear transformations of independent variables of high-dimensional data into lower-dimensionality space, seeking the best directions in the data that account for most of the variability

Uses: Mainly used for dimensionality reduction and feature extraction in large datasets

Others: PCA transforms and project high-dimensional data in a low-dimensional space

**principal components** artificial variables that are linear combinations of the original variables

**ill-posed problem** According to Jacques Hadamard, mathematical models of physical phenomena should have the properties that:

1. a solution exists
2. the solution is unique
3. the solution's behavior changes continuously with the initial conditions

If the problem is well-posed, then it stands a good chance of solution on a computer using a stable algorithm. If it is not well-posed, it needs to be re-formulated for numerical treatment. Typically this involves including

additional assumptions, such as smoothness of solution. This process is known as regularization.

Even if a problem is well-posed, it may still be ill-conditioned, meaning that a small error in the initial data can result in much larger errors in the answers. Problems in nonlinear complex systems (so called chaotic systems) provide well-known examples of instability. An ill-conditioned problem is indicated by a large condition number.

**regularization** is the process of adding information in order to solve an ill-posed problem or to prevent overfitting

**overfitting** is the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably. An overfitted model is a statistical model that contains more parameters than can be justified by the data. The essence of overfitting is to have unknowingly extracted some of the residual variation (i.e., the noise) as if that variation represented underlying model structure.

**underfitting** occurs when a statistical model cannot adequately capture the underlying structure of the data. An underfitted model is a model where some parameters or terms that would appear in a correctly specified model are missing. Underfitting would occur, e.g., when fitting a linear model to non-linear data, tending to have poor predictive performance.

**LASSO** *Least Absolute Shrinkage and Selection Operator*. Regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. Lasso was introduced in order to improve the prediction accuracy and interpretability of regression models by altering the model fitting process to select only a subset of the provided covariates for use in the final model rather than using all of them

Is a regression method that involves penalizing the absolute size of the regression coefficients. By penalizing (or equivalently constraining the sum of the absolute values of the estimates) you end up in a situation where some of the parameter estimates may be exactly zero. The larger the penalty applied, the further estimates are shrunk towards zero. This is convenient when we want some automatic feature/variable selection, or when dealing with highly correlated predictors, where standard regression will usually have regression coefficients that are 'too large'.

Generalizations of lasso:

- Elastic net
- Group lasso
- Fused lasso
- Quasi-norms and bridge regression

- Adaptive lasso

**Factor Analysis** Factor Analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. Is a useful tool for investigating variable relationships for complex concepts such as socioeconomic status, dietary patterns, or psychological scales. Allows researchers to investigate concepts that are not easily measured directly by collapsing a large number of variables into a few interpretable underlying factors.

The key concept of factor analysis is that multiple observed variables have similar patterns of responses because they are all associated with a latent (i.e. not directly measured) variable. In every factor analysis, there are the same number of factors as there are variables. Each factor captures a certain amount of the overall variance in the observed variables, and the factors are always listed in order of how much variation they explain.

The eigenvalue is a measure of how much of the variance of the observed variables a factor explains. Any factor with an eigenvalue  $> 1$  explains more variance than a single observed variable. The relationship of each variable to the underlying factor is expressed by the so-called factor loading. Since factor loadings can be interpreted like standardized regression coefficients, one could also say that the variable income has a correlation of 0.65 with Factor 1. This would be considered a strong association for a factor analysis in most research fields.

# Appendix A

## Essential Literature

### A.1 Thesis

- Adler's PhD Thesis → Pneumococcal nasopharyngeal colonisation in adults
- Elena's PhD Thesis → The human lung immune responses to nasopharyngeal pneumococcal colonisation in healthy adults
- Fernando's Master Thesis

### A.2 Articles

#### A.2.1 Pneumococcal Diseases

- The immunological mechanisms that control pneumococcal carriage
- *Streptococcus pneumoniae*: transmission, colonization and invasion
- The Pneumococcus: Epidemiology, Microbiology, and Pathogenesis
- Inflammation induced by influenza virus impairs human innate immune control of pneumococcus
- Pneumococcal colonization impairs mucosal immune responses to Live Attenuated Influenza Vaccine in adults
- Experimental human pneumococcal carriage models for vaccine research
- Innate and adaptive nasal mucosal immune responses following experimental human pneumococcal colonization

### **A.2.2 Sex Differences**

- Sex Differences in the Blood Transcriptome Identify Robust Changes in Immune Cell Proportions with Aging and Influenza Infection

### **A.2.3 Ageing**

- The Hallmarks of Aging

### **A.2.4 Systems Biology**

### **A.2.5 Bioinformatics and Methodologies**

- Meta-analysis and the science of research synthesis
- SIMON, an Automated Machine Learning System, Reveals Immune Signatures of Influenza Vaccine Responses

## Appendix B

# Pneumococcal Diseases and Immunity

### B.1 Controlled Human Infection and Rechallenge with *Streptococcus pneumoniae* Reveals the Protective Efficacy of Carriage in Healthy Adults

- High rates of disease and death in the elderly are associated with low carriage prevalence.
- Objective: apply an experimental human pneumococcal carriage model to investigate the immunizing effect of a single carriage episode;
- Carriage increased both mucosal and serum IgG levels to pneumococcal proteins and polysaccharide, resulting in a fourfold increase in opsonophagocytic activity;
- passive transfer of postcarriage sera from colonized by a heterologous strain in a murine model of invasive pneumococcal pneumonia. These levels were significantly higher than the protection conferred by either precarriage sera (30%) or saline (10%);
- Responses elicited by a single experimentally induced subsequent carriage and mice against invasive pneumococcal disease by passive transfer of sera from colonized individuals.
- We have found no relation of baseline serum IgG and carriage outcome after challenge.

- Intranasal exposure to bacteria boosted serum IgG levels to several pneumococcal proteins, with carriage-positive subjects showing the greatest magnitude of response.
- Persistent increased IgG to proteins was observed in carriage-positive subjects at 5 weeks after inoculation and not in those without carriage.
- The boosting effect of exposure is temporary and therefore different than responses induced by persistent carriage. This increased and sustained carriage acquisition after rechallenge up to 11 months after clearance of the first carriage episode.

## **B.2 Inflammation induced by influenza virus impairs innate control of human pneumococcal carriage**

2know:

- secondary vs primary bacterial pneumonia
- upper respiratory tract
- human type 6B pneumococcal challenge model
- degranulation of neutrophils
- role of neutrophils
- role of monocytes
- within-household Spn transmission
- Live Attenuated Influenza Vaccine
- murine models
- Th17-dependent recruitment of neutrophils
- Type I interferons
- scavenger receptor MARCO
- double-blinded controlled randomized clinical trial



## *B.2. INFLAMMATION INDUCED BY INFLUENZA VIRUS IMPAIRS INNATE CONTROL OF HUMAN PNEUM*

- tretavalent inactivated influenza vaccine
- nasal lining fluid
- rhinovirus
- coronavirus
- respiratory syncytial virus
- parainfluenzavirus
- luminal neutrophils
- myeloperoxidase (marker for neutrophil degranulation)
- Nanostring expression analysis
- 80,000 CFU per nostril
- lytA qPCR
- systematic LAIV dispensing error
- Neutrophil opsonophagocytic killing

Primary endpoint:

- the occurrence of pneumococcal colonisation determined by the presence of pneumococcus in nasal

Secondary endpoint:

- density of pneumococcal colonisation in NW at each time point following pneumococcal inoculation
- the area under the curve of pneumococcal colonisation density following pneumococcal inoculation
- immunological mechanisms associated with altered susceptibility to pneumococcus following LAIV.
- Flow cytometry analysis:
  - nasal cells
  - whole blood

- Neutrophil opsonohagocytic killing
- Luminex analysis of nasal lining fluid or stimulated nasal cells
- RNA extraction and sequencing
  
- Illumina Hiseq4000, 20M reads, 100 paired-end reads
  
- Nanostring
  
- Purified blood neutrophils

## Appendix C

### Sex Differences



# Appendix D

## Ageing

### D.1 The Hallmarks of Aging

We have two different ages:

- chronological age

- number of years since you were born

- biological age

- the time-dependent decline of your body's function and appearance

Chronological and biological age correlate with each other, so the signs of aging appear around a similar chronological age in most people.

But often people exhibit signs of biological aging at very different rates.

Extreme examples can be found in progeroid conditions-congenital disorders that cause the signs of biological aging to begin at a very young age.

In recent years, scientists studying the molecular and cellular processes that govern these changes and their variation in individuals have identified nine interconnected “hallmarks of aging”. Determined mainly by our genetics, but modulated by environmental factors, each of these nine hallmarks contributes to the damage that occurs with age and ultimately drives age-associated pathologies.

1. Genomic Instability
2. Telomere attrition

3. Epigenetic alterations
4. Loss of proteostasis
5. Deregulated nutrient sensing
6. Mitochondrial dysfunction
7. Cellular senescence
8. Stem cell exhaustion
9. Altered intercellular communication

## D.2 Therapies

- CR mimetics

may improve nutrient sensing.

- senolytics

a class of drug that removes senescent cells.

- quercetin (The Achilles'heel of senescent cells: from transcriptome to senolytic drug  
senolytic treatment using quercetin significantly improved vasomotor function in t

It has been demonstrated that senescent cells can be cleared selectively by targeting a

Zhu, Yi et al. - The Achilles' Heel of Senescet Cells: From Transcriptome to Senolytic

Chang, Jianhui et al. - Clearance of Senescent Cells by ABT263 Rejuvenates Aged Hematopoietic

Discovery of Piperlongumine as a Potential Novel Lead for the Development of Senolytic

Identification of a Novel Senolytic Agent, Navitoclax, Targeting the Bcl-2 Family of

Baar et al. (Targeted Apoptosis of Senescent Cells Restores Tissue Homeostasis in Respiratory

The molecule in the Baar et al. study instead functions by disrupting the interaction between

## D.3 Senescent Cells

Senescent cells are no longer capable of cell division and they do not support the tissue they are part of. Instead they secrete a cocktail of harmful pro-inflammatory chemical signals that inhibit tissue repair and drive chronic inflammation.

Normally senescent cells are removed and recycled by the immune system but as we age this too begins to decline and more and more senescent cells escape this housekeeping process. It is at that point that cellular senescence ceases being beneficial and protecting us and becomes a driver of the aging process.

Good citizens but bad neighbors - Non-dividing cells would not be a problem themselves but unfortunately the secreted pro-inflammatory chemicals they express also encourage nearby healthy cells to enter the same senescent state.

Collectively this cocktail is known as the Senescent-Associated Secretory Phenotype (SASP)

SASP:

(The senescece-associated secretory phenotype: the dark side of tumor suppression)

- inhibits a number of important cellular processes
- prevents effective tissue repair
- contributes to chronic background inflammation
- is implicated in the onset of age-related diseases (Inflammtory networks during cellular se

Senescent cells contribute to a second hallmark of aging: altered intercellular communication. This is the age-associated low-grade chronic inflammation many researchers call “inflammaging” and is another hallmark of the aging process.

Inflammaging

is caused by:

- infectious burden
- cell debris
- excessive activation of the NF-kB protein complex (regulator of the immune response)
- senescent cell SASP

interferes with intracellular signalling

contributes to the loss of regenerative capacity in stem cells and tissues

Senescent cells normally destroy themselves via a programmed process called Apoptosis and they are also removed by the immune system, however the immune system weakens with age and increasing numbers of these senescent cells escape this process and build up. By the time people reach old age significant numbers of these senescent cells have accumulated in the body and inflammation and damage to surrounding cells and tissue.

## D.4 The Role of Senescent Cells in Aging

- profound chromatin and secretome changes
- tumor-suppressor activation

replicative senescence -> Hayflick - this particular type of senescence is linked to telomere attrition, a process that leads to chromosomal instability and promotes tumorigenesis, supporting the original hypothesis that senescence guards against unrestricted growth of damaged cells.

the physiological relevance of cellular senescence extends beyond tumour suppression into biological processes such as:

- embryonic development
  - 10--12
- wound healing 13
- tissue repair 14
- organismal ageing 15,16

Causes and effector pathways of senescence

Stresses that can induce senescence:

- telomere erosion
- DNA lesions
- ROS



What they all have in common is that they activate the DNA damage response (DDR), a signal

- Activated oncogenes

Oncogenic Ras acts through overexpression of Cdc6 and suppression of nucleotide metabolism

- senescence caused by E2F3 activation or c-Myc inhibition is DDR-independent and involves p16
- BRAF (V600E) is also DDR-independent and induces senescence through a metabolic mechanism
- various tumour suppressors trigger a senescent growth arrest when inactivated, including RB

Of these, RB inactivation engages the DDR (26), whereas the others are DDR-independent and

- Prolonged exposure to interferon- $\beta$  also induces senescence, demonstrating that chronic mitogenic
- epigenetic, nucleolar and mitotic spindle stresses.

genome-wide chromatin decompression by exposure to histone deacetylase inhibitors triggers

A key target of epigenetic stressors that promote senescence may be the INK4a/ARF locus,

- Senescence can also be elicited by suboptimal expression of proteins implicated in spindle

A notable species-specific difference is that senescence pathways of murine cells are more dependent on p19Arf than senescence in human cells (27).

Senescence is a multi-step evolving process

Acute vs chronic senescence

Senescence of post-mitotic cells

Senescence in aging and age-related disease

Senescent-cell clearance and future directions



# Appendix E

## Bioinformatics

### E.1 Bioinformatics Essentials:

- Programming
  - R basics
    - variables
    - basic operations
    - logic operations
    - functions
    - data reading and writing
    - data frame manipulations
    - merging
    - Workflow and Projects
    - Programming Skills
  - Pipes
  - Functions
  - Vectors
  - Iteration (loops)
  - Graphs
- 3. Important (frequent) Types of Data
  - Relational data
    - <https://r4ds.had.co.nz/relational-data.html>
    - will give you tools for working with multiple interrelated datasets.
  - Strings:
    - <https://r4ds.had.co.nz/strings.html>
    - will introduce regular expressions, a powerful tool for manipulating strings.
  - Factors
    - <https://r4ds.had.co.nz/vectors.html#factors-1>

are how R stores categorical data. They are used when a variable has a fixed

- Dates and times
  - <https://r4ds.had.co.nz/dates-and-times.html#dates-and-times>
  - will give you the key tools for working with dates and date-times.
- Bash
- Python
- Statistics:
  - distributions
  - T test
  - ANOVA
  - correlation
  - regression
  - PCA
  - Multivariate Analysis
  - Bayesian statistics
- Linear Algebra (very basics, mostly to understand PCA)
- Exploratory Data Analysis (EDA)
  - Olhar:
    - [https://rpubs.com/crazyhotommy/PCA\\_MDS](https://rpubs.com/crazyhotommy/PCA_MDS)
    - [http://girke.bioinformatics.ucr.edu/GEN242/mydoc\\_Rclustering\\_3.html](http://girke.bioinformatics.ucr.edu/GEN242/mydoc_Rclustering_3.html)
    - [https://rstudio-pubs-static.s3.amazonaws.com/93706\\_e3f683a8d77244a5b993b20ad6278f4](https://rstudio-pubs-static.s3.amazonaws.com/93706_e3f683a8d77244a5b993b20ad6278f4)
  - 1. Data Wrangling
    - Data Import (<https://r4ds.had.co.nz/data-import.html>)
    - Tidy data (<https://r4ds.had.co.nz/tidy-data.html>)
  - 2. Data Understand Cycle
    - (Transform -> Visualise -> Model -> Transform -> ...)
    - Variation
    - Outliers
    - Covariation
    - Patterns and Models
  - 3. Communicate
    - Data Visualisation
    - Low-Dimensional Data Visualisation
      - Continuous-Continuous
      - Continuous-Categorical
      - Categorical-Categorical
    - High-Dimensional Data Visualisation
      - Principal Component Analysis (PCA)

## Multidimensional Scaling (MDS)

- Bioinformatics Data Retrieval
  - GEO
  - ArrayExpress
  - TGAC
- Microarray preprocessing:
  - Main technologies: Illumina, Affymetrix and Agilent
  - Quality control
    - ArrayQualityMetrics
    - other technologies
  - Normalization
    - RMA
    - quantile
  - Batch analysis:
    - PCA
    - PVCA
    - RLE
  - Missing values:
    - data imputation
    - variable/sample removal assessment
- Basic Analysis
  - Differential Gene Expression Analysis
    - limma
  - Clustering
    - hierarchical clustering
    - k-means
    - density-based clustering
  - Co-expression Analysis
    - WGCNA
    - CEMITool
  - Enrichment Analysis
    - ORA
    - GSEA
    - arbitrary search for Consistent Enrichment Gene Analysis (asCEGAS)
    - more advanced ones
- Machine learning:

- Regression:
  - linear regression
  - polynomial regression
  - ridge/lasso/elastic net regression
  - SVM
  - Decision Trees
  - Random Forest
  - neural networks
- Classification:
  - logistic regression
  - SVM
  - Decision Tress
  - Random Forest
  - neural networks
- Model performance evaluation
- Variable Importance
- Feature Selection
  - forward/backward selection
- Dimensionality Reduction
  - Feature Selection/elimination
  - Feature Engineering/extraction
- Data Integration
  - MixOmics
- loose things
  - scree plot
  - Oranges
  - Regression Plots
  - gradient descent
  - heteroskedasticity
  - Independent Component Anaysis
  - Factor Analysis
  - Discriminant Analysis

External Resources:

Books:

- Livro Draghici (microarray analysis bible)

## E.2 Differential Expression Analysis

Differential gene expression analysis

[https://www.youtube.com/watch?v=5tGCBW3\\_0IA](https://www.youtube.com/watch?v=5tGCBW3_0IA)

Normalization

Dispersion estimation

Log fold change estimation

Statistical testing

Filtering

Multiple testing correction

— NORMALIZATION

for comparing gene expression between (groups of) samples, normalize for

- library size (number of reads obtained)
- RNA composition effect

The number of reads for a gene is also affected by transcript length and GC content

- When studying differential expression you assume that they stay the same

“FPKM and TC are ineffective and should be definitely abandoned in the context of differential analysis”

”In the presence of high count genes, only DESeq and TMM (edgeR) are able to maintain a reasonable false positive rate without any loss of power

Do NOT use RPKM/FPKM for differential expression analysis!

- Reads (or fragments) per kilobase per million mapped reads.
- Normalizes for gene length and library size:
- 20kb transcript has 400 counts, library size is 20 million reads

$$\Rightarrow \text{RPKM} = (400/20)/20 = 1$$

- 0.5 kb transcript has 10 counts, library size is 20 million reads

$$\Rightarrow \text{RPKM} = (10/0.5)/20 = 1$$

- RPKM/FPKM can be used only for reporting expresison values, not for testing differential expression
- In DE analysis raw counts are needed to assess the measurement precision correctly

— NORMALIZATION BY edgeR and DESeq

= Aim to make normalized counts for non-differentially expressed genes similar between samples

- do not aim to adjust count distributions between samples

= Assume that

- Most genes are not differentially expressed
- Differentially expressed genes are divided equally between up- and down-regulation

= Do not transform data, but use normalization factors within statistical testing

Normalizatio by edgeR/DESeq2 - how?

= DESeq2

- take geometric mean of gene's counts across all samples
- divide gene's counts in a sample by the geometric mean
- take median of these ratios -> sample's normalization factor (applied to read counts)

= edgeR

- Select as reference the sample whose upper quantile is closest to the mean upper quartile
- log ratio of gene's counts in sample vs reference -> M value
- take weighted trimmed mean of M-values (TMM) -> normalization factor (applied to library sizes)
- trim: exclude genes with high counts or large differences in expression
- weights are from the delta method on binomial data



### E.2.1 Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq

Unknowns:

- flow cytometry

In biotechnology, flow cytometry is a laser or impedance-based, biophysical technology employed in

- impedance

Electrical impedance is the measure of the opposition that a circuit presents to a current wh

- precision & recall

precision (also called positive predictive value) is the fraction of retrieved instances that are  
proportion of retrieved set that are in fact relevant

$$P = \Pr(\text{relevant} \mid \text{retrieved}) = TP / (TP + FP)$$

intuition: how much junk did we give to the user?

recall - fraction of all relevant documents that were found

$$R = \Pr(\text{retrieved} \mid \text{relevant}) = TP / (TP + FN)$$

intuition: how much of the good stuff did we miss?

also known as sensitivity, is the fraction of relevant instances that are retrieved

Recall in this context is also referred to as the true positive rate or sensitivity, and precision

Accuracy involves how close you come to the correct result and your accuracy improves with tools

Precision is how consistently you can get that result using the same method

- Expression Modeler

swiftly - rapidamente

paucity - escassez

thawed - descongelado

analytes - analitos (Analito é uma substância ou componente químico, em uma amostra, q

- PPLR

# 1. Background

- estimating expression from short sequence reads poses unique problems such as ac
- the optimal

workflow for a given application remains a subject of

intensive investigation.

- The three major steps of differential expression analysis by RNA-Seq are:
  - + alignment of reads to an annotated genome (or less commonly, ab initio recon
  - + expression modeling to obtain gene-level and/or transcript-level expression c
  - + statistical analysis to identify differentially expressed genes or transcript
- the ultimate evaluation of any given tool must take into consideration the sampl
- We find that

different RNA-Seq analysis workflows differ widely in

their performance, as assessed by recall, or the propor-

tion of reference-identified genes that were also identi-

fied by the given workflow, and precision, or the

proportion of genes identified by the workflow that were

also identified by the reference.

- Many workflows per-

form equally well, but are calibrated differently with re-

spect to favoring higher recall or precision, with an

inverse relationship between these parameters.

- we recommend that the selection of a given approach be guided by the tolerance of downstream applications for type I and type II errors.

## 2. Methods

### 2.1 Samples

### 2.2 RNA sequencing

### 2.3 Read alignment, expression modeling, and differential expression identification

all code are available at <<https://github.com/cckim47/kimlab/tree/master/rnaseq>>

Reads were aligned to release GRCh37 of the human genome.

Reads were aligned with:

- Bowtie2
- HISAT2
- Kallisto
- Salmon
- Sailfish
- SeqMap
- STAR
- TopHat2

Gene and transcript expression was estimated with:

- BitSeq
- cufflinks
- htseq

- IsoEM
- Kallisto
- RSEM
- rSeq
- Sailfish
- Salmon
- STAR
- Stringtie
- eXpress

Expression matrices for differential expression input were generated using custom s

Differentially expressed genes or transcripts were identified with Ballgown, baySe

Of these, all but Ballgown, BitSeq, NBPSeg, SAMSeq, and Sleuth used intrinsic

For Sailfish and Salmon, outputs were converted to a Sleuth-ready format using

For Kallisto, Sailfish, Salmon, and BitSeq, transcript-level values were conde

For all differential expression analyses performed at the transcript-level, signif

All software was run at a detection level of alpha of 0.05, FDR of 0.05, or PPLR i

## 2.4 Preparation of reference datasets

series matrix files were:

1. downloaded from the NCBI Gene Expression Omnibus
2. log 2 transformed if necessary
3. full-quantile normalized

Smyth GK. Linear models and empirical bayes methods for assessing differen

4. analyzed for statistically significant gene expression between classical and

To reduce bias introduced by a single statistical method, we employed two approaches:

- Significance Analysis of Microarrays (SAM) [58] with a false discovery rate of 0.05. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation experiment. *Biostatistics* 2001;2:217-226.
- limma [59, 60], with a BH-adjusted p-value of 0.05.

Kim CC, Falkow S. Significance analysis of lexical bias in microarray data. *Bioinformatics* 2003;19:103-110.

Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey VJ, eds. *Bioinformatics and Computational Biology*. Springer, 2003:397-420.

Performance of the workflows against both SAM and limma were compared to one another using the same reference dataset.

Statistical method used to generate the data; as such, we chose to use the genes at the top of the list.

## 2.5 Quantification of recall and precision

Because absolute recall and precision values are influenced by the repertoire of analytes that are used, we used relative values.

Recall was calculated as the number of significant genes in the intersection of the test RNA-Seq dataset and the reference dataset, divided by the number of genes identified as significant in the reference dataset.

Precision was calculated as the number of significant genes in the intersection of the test RNA-Seq dataset and the reference dataset, divided by the number of genes identified as significant in the test RNA-Seq dataset.

## 3. Results and discussion

### 3.1 Generation of a real-world RNA-Seq dataset for benchmarking

### 3.2 Overview of empirical testing

We found that performance of various RNA-Seq workflows was remarkably consistent across all five reference datasets.

We note, however, that these reference datasets are also subject to the inherent biases of the RNA-Seq technology.

### 3.3 Differential influence of workflow stages

In general, more significant genes were observed when evaluations were performed at the transcript level than at the gene level.

We observed substantial variability in the number of differentially expressed genes identified across the five reference datasets.

Beyond the overall variation, two trends were apparent when the number of genes identified was compared across the five reference datasets.

1. the differential expression tool had a larger impact on the number of genes identified than the reference dataset.

Consequently, the coefficient of variation of the medians was largest for

2. differential expression tools varied in their robustness to different inputs.

We also evaluated performance of the workflows by calculating recall (intersecting

for both precision and recall, the largest effects were observed in workflows

### 3.4 heterogeneity in performance characteristics of different workflows

Recall across the workflows was highly correlated with the number of genes identified.

The relative rankings of the workflows, ordered by absolute recall value, tended to

For gene-level predictions, a subset of workflows using SAMseq exhibited the highest

for transcript-level predictions, workflows using baySeq and NBPSseq exhibited the

However, there were exceptions to these rules, depending on the choice of read

Precision was highly inversely correlated with the number of genes predicted across

Rankings were generally consistent regardless of which reference dataset was used,

For gene-level predictions, a subset of workflows using NOISeqBIO exhibited the

### 3.5 Performance tradeoff

the workflows employing NOISeqBIO that exhibit the highest precision were also among

An investigation of the relationship between precision and recall revealed that the

This held true for both gene- and transcript-level analysis, was true regardless

the differential expression step had the greatest impact on the performance of each

Specific tools that tended to track along this linear tradeoff were Ballgown, DESeq2

baySeq and EBseq consistently deviated the furthest.

SAMseq, one tool with a nonparametric approach, has been highlighted as a high performer.

NOISeqBIO, the other tested differential expression tool that assumes a nonparametric

Of the differential expression methods tested, baySeq and EBseq are the most similar.

All three linear model workflows perform well and track along the linear precision/recall trajectory. Using BitSeq as the expression modeler tended to result in identification of large numbers of differentially expressed genes compared to other tools. DESeq2 and edgeR were more conservative than BitSeq. DESeq2 was the one exception, with the number of differentially expressed genes within range of other tools. We note that BitSeq was unusual in that its most prevalent estimated expression count value was 1. Using STAR as the read aligner, most notably with Ballgown as the differential expression tool, resulted in a higher number of differentially expressed genes. The selection of a specific workflow should be largely influenced by the tolerance of a specific tool for low expression values. Our findings reflect a defined set of parameters, such as read length, sequencing coverage, and read depth. Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. Comparative assessment of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 2013;14:1-12. Sonesson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 2013;14:1-12. Importantly, when selecting a pipeline it is essential to consider not only the specific tool

#### 4. Conclusions

### E.3 Network Analysis

#### E.3.1 2008 - GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function

- A new algorithm that is as accurate as the leading methods while capable of predicting protein function in real-time.
- use a fast heuristic algorithm, derived from ridge regression, to integrate multiple functional annotations.
- gene function prediction through extrapolation of the functional properties of known genes.
- Genes with similar patterns of expression, synthetic lethality, or chemical sensitivity often have similar functions.
- function tends to be shared among genes whose gene products interact physically, are part of the same pathway, or are co-expressed.
- Computational analyses have also revealed shared function among genes with similar phylogenetic relationships.

- more accurate predictions can be made by combining multiple heterogeneous sources of
- guilt-by-association principle.

### E.3.2 2010 - The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function

Questions

Points

- input: gene list
- GeneMANIA extends the user's list with genes that are functionally similar, or have s
- Users interested in prioritizing genes for planning a functional screen can use GeneM
- it assigns weights to data sets based on how useful they are for each query.
  - > Individual datasets are represented as networks, and in the basic algorithm, ea
  - > GeneMANIA's adaptive weighting methods also detect and down-weight redundant net
- Organisms and identifiers:
  - > support six organism:
    - yeast (*Saccharomyces cerevisiae*)
    - worm (*Caenorhabditis elegans*)
    - fly (*Drosophila melanogaster*)
    - mouse (*Mus musculus*)
    - *Arabidopsis thaliana*
    - human (*Homo sapiens*)
  - > 747 data sets:
    - 276 co-expression networks, from GEO;



- 232 physical interaction, from BioGRID;
- 24 genetic interaction, from BioGRID;
- 14 co-localization, from Pathway Commons;
- 5 pathway, from Pathway Commons;
- 176 predicted protein domain information, from I2D;
- 12 shared protein domain information), from I2D;
- \> support standard genes symbols:
  - Ensembl, Entrez, UniProtKB, and RefSeq database identifiers; and unique gene synonyms.
  - Since we use Ensembl as a primary identifier source, we do not recognize ambiguous gene symbols.
- Users can upload their own data sets
- Network weighting methods
  - By default the GeneMANIA prediction server uses one of two different adaptive network weighting methods:
    - For longer lists, GeneMANIA uses the basic weighting method (GeneMANIA Entry-1 or assign equal weights to all edges)
    - GeneMANIA learns from longer gene lists, allowing a gene list-specific network weighting
    - For shorter lists, GeneMANIA uses a similar principle to weight networks, but tries to avoid overfitting
    - The user may choose other adaptive and non-adaptive weighting methods in the advanced options
    - The two non-adaptive methods are the most conservative options and work well on small gene lists
    - Network weights can also be assigned based on how well they reproduce GO co-annotation patterns
    - The annotation-based weighting may slightly inflate weights for networks on which current annotations are present
- Determining the network weights
  - The constructed composite network is a weighted sum of individual data sources;
  - each edge (link) in the composite network is weighted by the corresponding individual data source
  - The network weights are non-negative, sum to 100% and reflect the relevance of each data source

- Given the composite network, we use label propagation (12) to score all non-
- These scores are used to rank the genes. The score assigned to each gene ref.
- Other gene function prediction programs
  - N-Browser
    - functions as a Java web start, which is less convenient for the casual user
  - bioPIXIE
    - provide users with a fixed network (or networks) to query, built by incor
  - MouseNet
    - provide users with a fixed network (or networks) to query, built by incor
  - STRING
    - gives users little choice about which functional association network data
  - Functional Coupling (FunCoup)
    - assigns weights using a naive Bayes framework that cannot detect redund

## E.4 Multidimensional scaling

- Multidimensional scaling (MDS) is a means of visualizing the level of similarity of
- It refers to a set of related ordination techniques used in information visualization
- It is a form of non-linear dimensionality reduction.
- An MDS algorithm aims to place each object in N-dimensional space such that the betw
  - Each object is then assigned coordinates in each of the N dimensions.
  - The number of dimensions of an MDS plot N can exceed 2 and is specified a priori
  - Choosing N=2 optimizes the object locations for a two-dimensional scatterplot

## E.5 Next-Generation Sequencing

Technologies:

- Illumina (Solexa) sequencing
- Roche 454 sequencing
- Ion torrent: Proton/PGM sequencing
- SOLiD sequencing

Illumina sequencing

In Illumina sequencing, 100-150bp reads are used.

454 sequencing

Can sequence much longer reads than Illumina. Like Illumina, it does this by sequencing multi

Ion Torrent: Proton/PGM sequencing

Unlike Illumina and 454, Ion torrent and Ion proton sequencing do not make use of optical sig

The fragment is ~200bp.

Like 454, the slide is flooded with a single species of dNTP, along with buffers and polymera

### E.5.1 RNA-Seq Analysis

#### E.5.1.1 STAR

##### 1. Basic Workflow

##### 1. Generate genome indexes files

FASTA + GTF -> genome indexes

##### 2. Mapping reads to the genome

genome indexes + RNA-Seq reads (FASTA/FASTQ) -> :

- alignments (BAM/SAM)

- mapping summary statistics
- splice junctions
- unmapped reads
- signal (wiggle) tracks
- ...

## E.5.2 microbiome data analysis

REFs:

- An introduction to the downstream analysis with R and phyloseq  
<https://micca.readthedocs.io/en/latest/phyloseq.html>
- Microbiota Analysis in R  
[https://rstudio-pubs-static.s3.amazonaws.com/268156\\_d3ea37937f4f4469839ab6fa2](https://rstudio-pubs-static.s3.amazonaws.com/268156_d3ea37937f4f4469839ab6fa2)

### E.5.2.1 microbiome data analysis - mixOmics

Supervised Analysis and Selection of Discriminative OTUs with sPLS-DA

PLSDA

1. run the perf function with a PLS-DA model with no variable selection.
    - 1.1 try to find a good PCs  $\rightarrow$  ncomp
  2. assess the performance of the PLSDA on ncomp components
    - 2.1 decrease in the classification error rate  $\rightarrow$  increase in classification performance
    - 2.2 plot indicates a increase/decrease in the classification error rate from one component to another
- BER: Balanced Error Rate
    - should be considered when we have an unbalanced number of samples per group.
    - Are the number of samples per group is similar?

- if yes, then both overall and BER should be overlapping.
- Where the performance reaches its best (ncomp)?
  - the of PC w/ the best performance should be used for a final PLSDA model

## sPLSDA

### 1. Tuning sPLS-DA

#### 1.1 Parameters to choose in sPLS-DA:

- the number of variables to select (keepX)
- the number of components (ncomp)

#### 1.2 To do this use function tune.splsda()

- needs to be performed prior to the sPLS-DA analysis to choose the parameters on a grid
- make sure to:
  - choose the appropriate M fold cross-validation
  - provide sufficient nrepeat in the evaluation model, except for 'loo' where it can o
  - also check the stability of the features selected during the cross-validation proce
- may show some convergence issues for some of the cases, it is ok for tuning

### 2. Selecting best of PC:

- 2.1 look in the graph and see if the addition of components increases/decreases the value BER
- 2.2 choose the best number of components
- 3.2 select keepX

### 3. run classic sPLS-DA w/ best\_ncomp

### 4. Evaluating sPLS-DA

- 4.1 classification performance of the sPLS-DA multilevel model wit perf()

## 4.2 perf()

- output:
  - mean error rates per component
  - type of distance
- Here do not hesitate to increase the number of repeats for accurate estimation

## OTU selection and plots

## 1. Variable importance

## 1.1 selectVar(res, comp = comp\_x)\\$value

- outputs:
  - the first selected OTUs
  - their coefficient from the loading vector (value.var)
    - absolute coefficient value: indication of the importance of the OTU
    - coefficient sign: indicates positive/negative correlations between the

## 1.2 Combine the variable importance from selectVar() with their stability

- stability: how often were they selected across the different CV runs

## 2. Contribution plots - plotLoadings()

- displays:
  - the abundance of each OTU (large abundance = large absolute value)
  - in which body site they are the most abundant for each sPLS-DA component.
- They need to be interpreted in combination with the sample plot to:
  - understand the similarities between body sites
  - to answer 'which bacteria characterise those body sites?'

3. Clustered Image Map - `cim()`

- A heatmap will also help understanding the microbial signature.
- We represent clustered image maps (with Euclidian distance, Ward linkage set by default) for
- The abundance values that are displayed are the normalised, log ratio transformed values.
- All OTUs selected by the sPLS-DA model are displayed, other options can include a specific

**E.5.2.2 preprocessing**

- Raw data

- counts [0 - inf]

- 0 values:

- Total Sum Scaling normalisation is OK with but not the log ratio transformation

- apply offset of 1: `data.raw = data.raw + 1`

- the offset will not circumvent the zero values issue, as after log ratio transformation we

- check if the offset was applied:

```
sum(which(data.raw == 0))
```

- pre-filtering

- pre-filtering out OTUs with low percentage of reads in relation with the total amount of reads

```
keep.otu = which(colSums(data)*100/(sum(colSums(data))) > percent)
```

- check library size

- ensure that:

- the number of counts for each sample is relatively similar

- there is no obvious outlier sample. Those samples may also appear as outliers in PCA plots

- Normalisation

- Because of uneven sequencing depths, library sizes often differ from one sample to another.
- Two types of scaling / normalisation currently exist to accommodate for library size differences:
  - TSS (Total Sum Scaling) normalisation which needs to be followed by log ratio transformation.
  - CSS (Cumulative Sum Scaling) normalisation followed by log transformation.
- TSS
  - Is a popular approach to accommodate for varying sampling and sequencing depth.
  - In TSS the variable read count is divided by the total number of read counts in each sample:
    - each variable read count is divided by the total number of read counts:
    - ```
TSS.divide = function(x){ x/sum(x) }
```
    - function is applied to each row (i.e. each sample):
    - ```
data.TSS = t(apply(data.filter, 1, TSS.divide))
```
  - results in compositional data (or proportions) that are restricted to a space where the sum of all components is 1.
  - TSS normalisation reflects relative information, and the resulting normalised data is compositional.
  - Using standard statistical methods on such data may lead to spurious results and therefore the data must be further transformed.
  - to circumvent this issue, we transform the compositional data using log ratios such as:
    - ILR (Isometric Log Ratio) transformation
    - CLR (Centered Log Ratio) transformation
- Log ratio transformation



- ILR and CLR transformations are implemented directly into our multivariate methods `pca`, `pls`
- Which log ratio transformation to use?
  - According to Filmozer et al [2]:
    - ILR transformation is best for a PCA analysis.
    - for a PLS-DA and sPLS-DA analysis, `logratio = 'CLR'` is necessary for OTU selection
    - Generally speaking, a PCA with either TSS+CLR or TSS+ILR may not make much difference
    - We generally prefer the 'CLR' log ratio transformation as it is faster and can be used with more data
- What if I want to apply another method (multivariate or univariate) on the ILR or CLR data?
  - In that case use our external function `"logratio.transfo"`
  - Make sure you apply it to the `TSS(data.raw +1)` first, it will be easier than having to transform the data after
  - The log ratio transformation is crucial when dealing with proportional data!, unless the data is already log transformed
- CSS & Log Transformation
  - CSS normalisation was specifically developed for sparse sequencing count data by Paulson et al
  - CSS can be considered as an extension of the quantile normalisation approach and consists of normalising the data by the total sum of counts
  - CSS corrects the bias in the assessment of differential abundance introduced by TSS and, as a result, allows for a more accurate comparison of the data
  - Therefore, for CSS normalised data, no ILR transformation is applied as we consider that the data is already log transformed

## E.6 OMICS

### E.6.1 `integrOmics` an R package to unravel relationships between two omics datasets

`integrOmics` efficiently performs integrative analyses of two types of 'omics' variables that are measured on the same samples. It includes:

- a regularized version of canonical correlation analysis to enlighten correlations between two datasets
- a sparse version of partial least squares (PLS) regression that includes simultaneous variable selection

## BACKGROUND

- the simultaneous analysis of two datasets is an important task to better understand the biological system
  - integration of 'omics' data will provide a better understanding of biological systems
  - challenges:
    - computational issues because of the "large p, small n" problem, e.g. canonical correlation analysis (CCA) is not applicable
    - give interpretable results, i.e. to answer the following questions:
      - (i) which variables from both types are related to each other
      - (ii) which relevant variables provide more insight into the biological experiment
- > The solution is to perform variable selection while combining the two types of data
- to address this problem, they developed two approaches:
- a regularized version of CCA to overcome computational issues in CCA when  $p \gg n$
  - a variant of partial least squares (PLS) regression (Wold, 1966) called sparse PLS

## METHODS AND IMPLEMENTATION

CCA and PLS are both exploratory approaches which enable the integration of two datasets

- CCA maximizes the correlation between linear combinations of the variables from each dataset
- PLS maximizes the covariance

Vinod (1976) and González et al. (2008) introduced  $l_2$  penalties on the covariance matrix

PLS circumvents this ill-conditioned matrices issue by performing local regressions

Both approaches seek for:

- (i)  $p$ - and  $q$ -dimensional weight vectors, called canonical factors or loading vectors
- (ii)  $n$ -dimensional vectors, called score or latent vectors

In order to give interpretable results and remove noisy variables, Lê Cao et al. (2008)

- > as a result, many coefficients in these vectors are set to zero, which naturally

Two types of analysis were proposed in sPLS:

- regression analysis for a causal relationship between the two datasets
- canonical analysis for a reciprocal relationship similar to a CCA framework

#### SOFTWARE FEATURES

- The  $Q^2$  criterion (Tenenhaus, 1998) can be computed to determine the number of components to choose
- The root mean square error prediction can be used to choose the optimal number of variables to choose
- the user can also estimate the predicted value of a new sample in the model
- regularization parameters in rCCA can be tuned using cross-validation.
- missing values of each dataset can be efficiently imputed with a singular value decomposition

Visualization outputs:

- scatter plots of the score (latent) vectors from the first dimensions allow the user to identify
- further, the (selected) variables can be represented by projecting them on correlation circles
- enables the inference of large-scale association networks between the two datasets with the

Versatility of integrOmics:

- rCCA and sPLS have been successfully applied in various biological contexts where  $p > n$
- sPLS: integrate gene expression with metabolite expression, clinical chemistry or fatty acid
- canonical rCCA/sPLS: relate physicochemical measurements with sensory variables or to relate

### E.6.2 Visualising associations between paired “omics” data sets

Major challenge with the integration of omics data -> extraction of discernable biological meaning from multiple omics data

Multivariate approaches:

- Aim: unravel the correlation structure between two sets of data measured on the same samples

- achieve dimension reduction by summarizing the data into a small number of components
- exploiting coexpression between disparate types of biological measures instead of

Other methods:

- clustering techniques
  - simple criteria matching: order the variables according to fold-change or un-
  - self-organizing maps, use Euclidian distances. However, they are known to en-
  - comprehensively compare all variables against each other using a similarity m-

Correlation Circle plots

- Correlation Circle plots were primarily used for PCA outputs to visualise the re-
- enables a graphical examination of the relationships between variables and variat
- the coordinates of the variables are obtained by calculating the correlation betw
- Because variables are usually centered and standardized, the correlation between
- variables can be represented as vectors (see Figure 1(b)) and the relationship (C

The inner product is defined as the product of the two vectors lengths and their co

- if the angle is sharp, the correlation is positive
- if the angle is obtuse the correlation is negative
- if the angle is right the correlation is null.
- The centered and standardized variables are projected onto the space spanned by t
- For variables closely located to the origin, it means that some information can b

Relevance Networks

- A conceptually simple approach for modelling net-like correlation structures betw
- was introduced by [10] as a tool to study associations between pair of variables
- This method generates a graph where nodes represent variables, and edges represen

- Since the relevance networks are visual representations of the correlations between variables
- The Relevance Network is built in a simple manner:
  1. the correlation matrix is inferred from the data
  2. for every estimated correlation coefficients exceeding (in absolute value) a prespecified threshold
- We will thus display Relevance Networks through the use of bipartite graph (or bigraph), where nodes represent genes and edges represent correlations
- Instead of computing the Pearson correlation coefficients between each pair of variables as in the case of the correlation matrix, the Relevance Network is built by computing the partial correlation coefficients between each pair of variables, controlling for the other variables in the network

Advantages:

- ability to simultaneously represent positive and negative correlations, which are missed by the correlation matrix
- ability to represent genes in several pathways, and, most importantly for our purpose, to represent the same gene in different pathways
- Limitations:
  - it requires extensive computing resources as mentioned by [26] to compute the comprehensive Relevance Network

#### Clustered Image Maps

The similarity matrix represented by the CIM is the same as in the relevance networks described above

## E.7 Principal Component Analysis

Is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

If there are  $n$  observations with  $p$  variables, then the number of distinct principal components is  $\min(n-1, p)$

- This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors (each being a linear combination of the variables and containing  $n$  observations) are an uncorrelated orthogonal basis set.
- PCA can be done by eigenvalue decomposition of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after

a normalization step of the initial data. The normalization of each attribute consists of mean centering – subtracting each data value from its variable’s measured mean so that its empirical mean (average) is zero – and, possibly, normalizing each variable’s variance to make it equal to 1;

component/factor scores: the transformed variable values corresponding to a particular data point

loadings: the weight by which each standardized original variable should be multiplied to get the component score

- PCA is sensitive to the relative scaling of the original variables.

- is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set.

- is used to explain the variance-covariance structure of a set of variables through linear combinations.

- As an added benefit, each of the “new” variables after PCA are all independent of one another. This is a benefit because the assumptions of a linear model require our independent variables to be independent of one another.

- When to use PCA:

- do you want to reduce the number of variables, but aren't able to identify variables

- do you want to ensure your variables are independent of one another

- are you comfortable making your independent variable less interpretable?

If the answer is 'yes' to all three questions, the PCA is a good method to use.

you should have tabular data organized with  $n$  rows and  $p+1$  columns, where there is one column corresponding to your dependent variable (usually denoted  $Y$ ) and  $p$  columns corresponding to each of your independent variables (the matrix of which is usually denoted  $X$ ).

1. Separate your data into  $Y$  and  $X$ , as defined above—we’ll mostly be working with  $X$ .

2. Take the matrix of independent variables  $X$  and, for each column, subtract the mean of that column from each entry. (This ensures that each column has a mean of zero.)

3. Decide whether or not to standardize. Given the columns of  $X$ , are features with higher variance more important than features with lower variance, or is the importance of features independent of the variance? (In this case, importance means how well that feature predicts  $Y$ .) If the importance of features is independent of the variance of the features, then divide each observation in

a column by that column's standard deviation. (This, combined with step 2, standardizes each column of  $X$  to make sure each column has mean zero and standard deviation 1.) Call the centered (and possibly standardized) matrix  $Z$ .

4. Take the matrix  $Z$ , transpose it, and multiply the transposed matrix by  $Z$ . (Writing this out mathematically, we would write this as  $Z^T Z$ .) The resulting matrix is the covariance matrix of  $Z$ , up to a constant.

5. (This is probably the toughest step to follow—stick with me here.) Calculate the eigenvectors and their corresponding eigenvalues of  $Z^T Z$ . This is quite easily done in most computing packages—in fact, the eigendecomposition of  $Z^T Z$  is where we decompose  $Z^T Z$  into  $PDP^{-1}$ , where  $P$  is the matrix of eigenvectors and  $D$  is the diagonal matrix with eigenvalues on the diagonal and values of zero everywhere else. The eigenvalues on the diagonal of  $D$  will be associated with the corresponding column in  $P$ —that is, the first element of  $D$  is  $\lambda_1$  and the corresponding eigenvector is the first column of  $P$ . This holds for all elements in  $D$  and their corresponding eigenvectors in  $P$ . We will always be able to calculate  $PDP^{-1}$  in this fashion. (Bonus: for those interested, we can always calculate  $PDP^{-1}$  in this fashion because  $Z^T Z$  is a symmetric, positive semidefinite matrix.)

6. Take the eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_p$  and sort them from largest to smallest. In doing so, sort the eigenvectors in  $P$  accordingly. (For example, if  $\lambda_1$  is the largest eigenvalue, then take the second column of  $P$  and place it in the first column position.) Depending on the computing package, this may be done automatically. Call this sorted matrix of eigenvectors  $P^*$ . (The columns of  $P^*$  should be the same as the columns of  $P$ , but perhaps in a different order.) Note that these eigenvectors are independent of one another.

7. Calculate  $Z^* = ZP^*$ . This new matrix,  $Z^*$ , is a centered/standardized version of  $X$  but now each observation is a combination of the original variables, where the weights are determined by the eigenvector. As a bonus, because our eigenvectors in  $P^*$  are independent of one another, each column of  $Z^*$  is also independent of one another!

8. Finally, we need to determine how many features to keep versus how many to drop. There are three common methods to determine this, discussed below and followed by an explicit example:

- Method 1: We arbitrarily select how many dimensions we want to keep. Perhaps I want to visually
- Method 2: Calculate the proportion of variance explained (briefly explained below) for each feature
- Method 3: This is closely related to Method 2. Calculate the proportion of variance explained for each feature

Because each eigenvalue is roughly the importance of its corresponding eigenvector, the proportion of variance explained is roughly the proportion of the total variance explained by the eigenvectors.

Why does PCA work?

While PCA is a very technical method relying on in-depth linear algebra algorithms, it

- First, the covariance matrix  $ZZ^T$  is a matrix that contains estimates of how every variable is associated with every other variable.
- Second, eigenvalues and eigenvectors are important. Eigenvectors represent directions of maximum variance.
- Finally, we make an assumption that more variability in a particular direction corresponds to more important information.

Thus, PCA is a method that brings together:

- A measure of how each variable is associated with one another. (Covariance matrix.)
- The directions in which our data are dispersed. (Eigenvectors.)
- The relative importance of these different directions. (Eigenvalues.)

PCA combines our predictors and allows us to drop the eigenvectors that are relatively

Are there extensions to PCA?

- principal component regression:

where we take our untransformed  $Y$  and regress it on the subset of  $Z$  that we didn't drop.

- kernel PCA

- Kernel PCA has been demonstrated to be useful for novelty detection and image de-noising.

- use kernel methods:

- are a class of algorithms for pattern analysis, e.g. SVM

- the kernel trick means transforming data into another dimension that has a classifying hyperplane.

- to solve these task, kernel methods require only a user-specified kernel, i.e. a function that takes two data points and returns a scalar value.

PCA Assumptions:

- PCA itself is a nonparametric method, but regression or hypothesis testing after using PCA is parametric.
- We do not have to make any distributional assumptions in order to extract the Principal Components.
- All we require is that the maximizers are unit vectors and perpendicular to the previous ones.



- We do not need normality for the extraction but we definitely need the normality for hypothesis

resources:

- <http://setosa.io/ev/principal-component-analysis/>
- [https://www.youtube.com/playlist?list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE\\_ab](https://www.youtube.com/playlist?list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab)

## E.8 MixOmics

### E.8.1 DIABLO

N-Integration discriminant analysis with DIABLO

#### 1. Data Input

The data for mixDIABLO need to be set up as a list of data matrices matching the same samples.

The blocks:

are 'omics data sets, where each row matches to the same biological sample from one data set to a

The outcome Y is set as a factor for the supervised analysis. Here Y corresponds to the PAM50 cla

#### 2. Tunning

##### - Matrix design

determines which blocks should be connected to maximize the correlation or covariance between

the values may range between 0 (no correlation) to 1 (correlation to maximize) and is a symme

The design can be chosen based on:

- prior knowledge ('I expect mRNA and miRNA to be highly correlated')
- data-driven (e.g. based on a prior analysis, such as a non sparse analysis with block.p

Our experience has shown that a compromise between maximising the correlation between blocks,

##### - Number of components

fit a DIABLO model without variable selection to assess the global performance and

- keepX

- should be used to tune the keepX parameters in the block.splsda function.
- We choose the optimal number of variables to select in each data set using the tune
- the function has been set to favor the small-ish signature while allowing to obtain
- We choose the optimal number of variables to select in each data set using the tune
- the function has been set to favour the small-ish signature while allowing to obtain
- The number of features to select on each component is returned in tune.TCGA\[choice]

Alternatively, you can manually input those parameters as indicated below.

### 3. Performance

- We assess the performance of the model using perf() function.
- The method runs:
  1. block.splsda model on the pre-specified arguments input from our sgccda.res object
  2. assess the accuracy of the prediction on the left out samples.
- Outputs:
  - usual (balanced) classification error rates
  - predicted dummy variables and variates
  - stability of the selected features
  - performance based on:
    - Majority Vote (each data set votes for a class for a particular test sample)
    - weighted vote, where the weight is defined according to the correlation between
- Since the tune function was used with the centroid.dist argument, we examine the output
 

```
\> set.seed(123)for reproducibility, only when the \'cpus\' argument is not used
```

```
\> perf.diablo = perf(sgccda.res, validation = 'Mfold', M = 10, nrepeat = 10, dist = 'centroid')

\> perf.diablo lists the different outputs

\> perf.diablo\$$MajorityVote.error.rate Performance with Majority vote

\> perf.diablo\$$WeightedVote.error.rate Performance with Weighted prediction
```

#### 1. Prediction on an external test set

The predict function predicts the class of samples from a test set.

- In our specific case, one data set is missing in the test set but the method can still
- Make sure the name of the blocks correspond exactly.

prepare test set data: here one block (proteins) is missing

```
data.test.TCGA = list(mRNA = breast.TCGA\$$data.test\$$mrna, miRNA = breast.TCGA\$$data.test\$$miRNA)
```

```
predict.diablo = predict(sgccda.res, newdata = data.test.TCGA)
```

the warning message will inform us that one block is missing

```
predict.diablo list the different outputs
```

- Confusion table

- The confusion table compares the real subtypes with the predicted subtypes for a 2

```
confusion.mat = get.confusion_matrix(truth = breast.TCGA\$$data.test\$$subtype, pred = predict.diablo)
```

```
confusion.mat
```

```
get.BER(confusion.mat)
```

#### Plots

- Samples

- Arrow Plot - plotIndiv()

- arrow plot below

- projects each sample into the space spanned by the components of each block
  - the start of the arrow indicates the centroid between all data sets for a given block
  - the tips of the arrows the location of that sample in each block
  - Such graphic highlight the agreement between all data sets at the sample level
  - Variables
    - Circle Plot - `plotVar()`
      - The correlation circle plot highlights the contribution of each selected variable
      - `plotVar` displays the variables from all blocks, selected on component 1 and 2
      - Clusters of points indicate a strong correlation between variables
    - Circos Plot - `circosPlot()`
      - The circos plot represents the correlations between variables of different types
      - Several display options are possible, to show within and between connexions
      - The circos plot is built based on a similarity matrix, extended to the case of multiple blocks
    - Relevance Network Plot - `network()`
      - Another visualisation of the correlation between the different types of variables
      - is also built on the similarity matrix
      - Each color represents a type of variable. A threshold can also be set using `threshold`
      - sometimes the output may not show with Rstudio because of margin issues. The `write.graph` function can be used to save the network in a .gml format
      - The network can be saved in a .gml format to be input into the software Cytoscape
- ```

\> library(igraph)

\> my.network = network(sgccda.res, blocks = c(1,2,3), color.node = c('darkred','darkblue','darkgreen'))

\> write.graph(my.network\$$R, file = "myNetwork.gml", format = "gml")

```
- Variable Importance Plot- `plotLoadings()`

- visualizes the loading weights of each selected variables on each component and each data set
- The color indicates the class in which the variable has the maximum level of expression
- Samples and Variables
  - Clustered Image Map Plot - cimDIABLO()
    - is a clustered image map specifically implemented to represent the multi-omics molecular data
- Diagnostic
  - DIABLO Diagnostic Plot - plotDIABLO()
    - is a diagnostic plot to check whether the correlation between components from each data set is high
    - The colors and ellipses related to the sample subtypes and indicate the discriminative power
    - `plotDiablo(sgccda.res, ncomp = 1)`
    - Are the first components from each data set highly correlated to each other?
- Performance
  - AUC Plot - auROC()
    - An AUC plot per block
    - Refer to [5] for the interpretation of such output as the ROC and AUC criteria are not directly comparable

## E.8.2 EXPLORATORY DATA ANALYSIS WITH mixOMICS

### SINGLE-OMICS

PCA

IPCA

### DOUBLE-OMICS

rCCA

sPLS

## MULTI-OMICS

## DIABLO

## MINT

## SINGLE-OMICS

## PRINCIPAL COMPONENT ANALYSIS [PCA]

- Jolliffe, 2005
- is primarily used to explore one single type of 'omics' data
- identify the largest sources of variation
- a mathematical procedure that uses orthogonal linear transformation of data from
- the first component explains as much of the variability in the data as possible,
- only the PCs which explain the most variance are retained.
- \> This is why choosing the number of dimensions or components (ncomp) is crucial
- in mixOmics, PCA is numerically solved in two ways:
  1. with singular value decomposition (SVD) of the data matrix:
    - is the most computationally efficient way
    - also adopted by most softwares and prcomp
  2. with the Non-linear Iterative Partial Least Squares (NIPALS) in the case of
- input data should be centered and possibly (sometimes preferably) scaled so that
  - this is specially advised in the case where the variance is not homogeneous a
  - by default, the variables are centered and scaled in the function.
- choosing the optimal parameters
  - we can obtain as many dimensions as the minimum between the samples and variables
  - however, the goal is to reduce the complexity of the data:

- summarize the data in fewer underlying dimension
- the of PC to retain is therefore crucial when performing PCA.
- The function `tune.pca` will plot the barplot of the proportion of explained variance

#### sparse Principal Component Analysis (sPCA)

- Shen and Huang, 2008
- is an unsupervised and exploratory technique
- is based on singular value decomposition and is appropriate to deal with large data set
- in `mixOmics`, 'sparsity' is achieved via LASSO penalizations.
- sPCA is useful to remove some of the non informative variables in PCA and can be used to
- the of variables to select on each PC must be input by the user (`keepX`)
- tuning sPCA `keepX` based on the amount of explained variance is difficult (the less var

#### INDEPENDENT PRINCIPAL COMPONENT ANALYSIS [IPCA]

Deal w/ some PCA Limitations:

- PCA assumes that gene expression follows a multivariate normal distribution
  - > recent studies have demonstrated that microarray gene expression follow instead a
- PCA decomposes the data based on the maximization of its variance
  - > in some cases, the biological question may not be related to the highest variance
- combines the advantages of both PCA and Independent Component Analysis (ICA)
- It uses ICA as a denoising process of the loading vectors produced by PCA to better highlight
- The algorithm is as follows:
  1. the original data matrix is centered (by default)
  2. PCA is used to reduce dimension and generate the loading vectors
  3. ICA (FastICA) is implemented on the loading vectors to generate independent loading ve

- 4. The centered data matrix is projected on the independent loading vectors to
- offers a better visualization of the data than ICA and with a smaller number of c
- Choosing the optimal parameters
  - the of variables to select is still an open issue
  - Yao et al (2012) proposed to use the Davies Bouldin measure
    - is an index of crisp cluster validity
    - this index compares the within-cluster scatter with the between-cluster s
- Kurtosis
  - the kurtosis measure is used to order the loading vectors to order the I
  - is a good post hoc indicator of the number of components to choose, as a

PLS-DA

Multilevel

Missing Values

### E.8.3 - Independent Component Analysis (ICA)

- Resources:
  - <[http://arnauddeleorme.com/ica\\_for\\_dummies/](http://arnauddeleorme.com/ica_for_dummies/)>
- independent component analysis (ICA) is a computational method for separating a mult.
- is a statistical and computational technique for revealing hidden factors that underl
- ICA defines a generative model for the observed multivariate data, which is typically
- ICA can be seen as an extension to principal component analysis and factor analysis.
- ICA is a technique to separate linearly mixed sources.
- is quite robust to different degrees of noise



- in theory, ICA can only extract sources that are combined linearly
- Steps
  - Whitening the data
    - a first step in many ICA algorithms is to whiten (or sphere) the data: this means that
      - Why do that:
        - A geometrical interpretation is that it restores the initial "shape" of the data
    - after whitening, the variance on both axis is now equal and the correlation of the projection is zero
    - The whitening process is simply a linear change of coordinate of the mixed data. Once the data is whitened, the variance on both axis is now equal and the correlation of the projection is zero
  - The ICA algorithm
    - Intuitively you can imagine that ICA rotates the whitened matrix back to the original (non-whitened) space
    - By rotating the axis and minimizing Gaussianity of the projection, ICA is able to recover the original sources
    - ICA can deal with an arbitrary high number of dimensions:
      - Let's consider 128 EEG electrodes for instance. The signal recorded in all electrodes is a 128xN matrix (N is the number of samples)
      - What we call ICA components is the matrix that allows projecting the data in the independent components space
      - when we talk about independent components, we usually refer to two concepts:
        - Rows of the S matrix which are the time course of the component activity
        - Columns of the W-1 matrix which are the scalp projection of the components
  - ICA properties:
    - ICA can only separate linearly mixed sources.
    - Since ICA is dealing with clouds of point, changing the order in which the points are plotted has no effect
    - Changing the channel order (for instance swapping electrode locations in EEG) has also no effect
    - Since ICA separates sources by maximizing their non-Gaussianity, perfect Gaussian sources cannot be separated
    - Even when the sources are not independent, ICA finds a space where they are maximally independent

### E.8.4 Independent Principal Component Analysis (IPCA)

Limitations when using PCA:

- PCA assumes that gene expression follows a multivariate normal distribution and
- PCA decomposes the data based on the maximization of its variance. In some cases
- IPCA combines the advantages PCA and Independent Component Analysis (ICA):
  - see Yao et al., 2012 [1]
  - It uses ICA as a denoising process of the loading vectors produced by PCA to bet
- IPCA offers a better visualization of the data than ICA and with a smaller number of
- IPCA results in better clustering of biological samples on graphical representations
- both IPCA and PCA rank the variables in similar same order of importance, the largest
- choosin the number of components:

The kurtosis measure is used to order the loading vectors to order the Independent  
sIPCA

- Sparse Independent Principal Component Analysis (IPCA) combines the advantages of
- The use of a sparse IPCA would be more appropriate to interpret the results as th
- Choosing the number of variables to select:
  - is still an open issue.
  - In our paper we proposed to use the Davies Bouldinmeasure which is an index o
  - This index compares the within-cluster scatter with the between-cluster s

REF:

- Yao F., Coquery J., Lê Cao K.-A. (2012) Independent Principal Component Analysis
- Comon P: Independent component analysis, a new concept? Signal Process 1994, 36:1

- Hyvärinen A, Oja E: Independent Component Analysis: Algorithms and Applications. Neural Netw

### E.8.5 mixOmics - An R package for omics feature selection and multiple data integration\_\_2017

univariate statistical analysis:

- ANOVA
- linear models
- t-tests

Cons:

- ignores relationships between the different features
- may miss crucial biological information

(biological features act in concert to modulate and influence biological systems and signalling)

multivariate statistical analysis:

- model features as a set
- can provide a more insightful picture of a biological system
- can complement the results obtained from univariate methods

mixOmics:

- data exploration, dimension reduction and visualization of integrated omics data sets
- multivariate projection-based methodologies
- supervised analysis
- DIABLO

- enables the integration of the same biological N samples measured on different 'omics platforms
- other types of N-integration:

- often performed by concatenating all the different 'omics data sets [13], which
- combine the molecular signatures identified from separate analyses of each 'omics data set
- MINT
  - enables the integration of several independent data sets or studies measured on the same platform
  - With P-integration, statistical methods are often sequentially combined to accommodate different data types

#### Data input

- normalized continuous data
- we recommend pre-filtering the data to less than 10K predictors per data set, for example

#### Multivariate projection-based methods

- All multivariate approaches listed are projection-based methods whereby samples are projected onto a lower-dimensional space
- Unsupervised
  - Principal Component Analysis --- based on NonLinear Iterative Partial Least Squares
  - Independent Component Analysis [19]
  - Partial Least Squares regression---PLS, also known as Projection to Latent Structures
  - multi-group PLS [21]
  - regularised Canonical Correlation Analysis---rCCA [22])
  - regularised Generalised Canonical Correlation Analysis---rGCCA based on a PLS algorithm
- Supervised
  - PLS-Discriminant Analysis---PLS-DA [24--26]
  - GCC-DA [11]
  - multi-group PLS-DA [12]

### E.8.6 Partial Least Squares regression (PLS regression)

- Partial least squares (PLS) regression is a technique that reduces the predictors to a smaller number of latent variables

- Unlike least squares regression, PLS can fit multiple response variables in a single model. PLS
- is a statistical method that bears some relation to principal components regression
- instead of finding hyperplanes of maximum variance between the response and independent variables
- Partial least squares Discriminant Analysis (PLS-DA) is a variant used when the Y is categorical
- PLS is used to find the fundamental relations between two matrices (X and Y), i.e. a latent variable
- PLS regression is particularly suited when the matrix of predictors has more variables than observations
- Extensions:
  - Orthogonal Projections to Latent Structures (OPLS)
    - continuous variable data is separated into predictive and uncorrelated information.
    - This leads to improved diagnostics, as well as more easily interpreted visualization.
    - However, these changes only improve the interpretability, not the predictivity, of the model
    - Similarly, OPLS-DA (Discriminant Analysis) may be applied when working with discrete variables
- Partial least squares regression extends multiple linear regression without imposing the restriction of orthogonality
- Is probably the least restrictive of the various multivariate extensions of the multiple linear regression
- Can be used as an exploratory analysis tool to select suitable predictor variables and to identify important response variables

#### Basic Model

- As in multiple linear regression, the main purpose of partial least squares regression is to find the best fit
- Both principal components regression and partial least squares regression produce factor scores
- As in multiple linear regression, the main purpose of partial least squares regression is to find the best fit
- Both principal components regression and partial least squares regression produce factor scores

#### Algorithms:

- NIPALS algorithm
  - Herman Wold

The standard algorithm for computing partial least squares regression components

- The algorithm reduces the number of predictors using a technique similar to p
- SIMPLS ALGORITHM
- de Jong, 1993

### E.8.7 Variable selection for generalized canonical correlation analysis

Canonical correlation analysis (CCA) is a standard approach to studying the relationships between two blocks of variables only, and numerous L1 and/or L2 regularized extensions of the CCA have been proposed when the number of variables  $p_j$  exceeds the number of observations  $n$  for any  $j$ th block (e.g. Vinod, 1976; Waaijenborg and others, 2008; Parkhomenko and others, 2009; Le Cao and others, 2009;

Witten and others, 2009; Lykou and Whittaker, 2010; Hardoon and Shawe-Taylor, 2011).

Regularized generalized canonical correlation analysis (RGCCA):

- proposed in Tenenhaus and Tenenhaus (2011)
- is a framework for studying associations between more than 2 blocks
- The aim of RGCCA is to extract the information which is shared by the  $J$  blocks of variables

Biomedical data are known to be measurements of intrinsically parsimonious processes. In order to account for this parsimony and to improve the interpretability of the resulting RGCCA model, an important issue is to identify subsets of variables from each block which are active in the relation between connected blocks.

- This variable selection step can be achieved by adding, within the RGCCA optimization

### E.8.8 Unravelling “omics” data with the mixOmics R package (ppt)

Issues with integrative systems biology

- Unlimited quantity of data
- $n \ll p$  problem
- data from multiple sources
- Efficient and biologically relevant statistical methodologies are needed to combine the information

Biological Questions:

- Single Omics Analysis
  - do we observe a "natural" separation between the different groups of patients?
  - Can we identify potential biomarker candidates predicting the status of the patients?
- Integrative Omics Analysis
  - Can we identify a subset of correlated genes and proteins from matching data sets?
  - Can we predict the abundance of a protein given the expression of a small subset of genes?
  - Do two matching omics data set contain the same information?

The data

$n$  = patients and  $p, q$  = features

- Single Omics Analysis
  - one omic data set  $X(n \times p)$
  - for a supervised analysis,  $Y$  vector indicating the class of the patients
- Integrative Omics Analysis
  - two matching omics data sets (measured on the same patients)
  - $X(n \times p)$  and  $Z(n \times q)$

Multivariate analysis:

linear multivariate approaches enable:

- Dimension reduction
- to handle multicollinear, irrelevant, missing values
- to capture experimental and biological variation

#### MixOmics

- exploration and integrative analysis of high dimensional biological data sets
- focus is on:
  - Data integration
  - Variable selection
  - Interpretable graphical outputs
- The exploratory and integrative approaches are:
  - flexible and can answer various types of questions
  - can highlight the potential of the data
  - enable to generate new hypotheses to be further investigated

#### Analysis:

##### One data set

##### Unsupervised:

Multivariate approach: PCA \| IPCA

Internal variable selection: sPCA \| sIPCA

Graphical outputs: sample and variable plots

##### Supervised:

Multivariate approach: PLS-DA

Internal variable selection: sPLS-DA

Graphical outputs: sample and variable plots



Two matching data sets

canonical mode:

Multivariate approach: PLS \rCCA

Internal variable selection: sPLS canonical

Graphical outputs: sample and variable plots

regression mode:

Multivariate approach: PLS

Internal variable selection: sPLS regression

Graphical outputs: sample and variable plots

Missing values

imputation: NIPALS

Future work:

- Cross-platform comparison
- Integration of multiple data sets (unsupervised and supervised)
- Time-course experiments

Single Omics Analysis

Principal Component Analysis (PCA)

seek the best directions in the data that account for most of the variability

-> principal components: artificial variables that are linear combinations of the original variables

-  $c$  is a linear function of the elements of  $X$  having maximal variance

-  $v$  is called the associated loading vector

The new PCs form a vectorial subspace of dimension  $\leq p$

- approximate representation of the data points in a lower dimensional space

Problem: interpretation difficult with very large number of (possibly) irrelevant variables

- unsupervised approach

#### sparse Principal Component Analysis (sPCA)

The principal components are linear combinations of the original variables, variables

- computes the sparse loading vectors to remove irrelevant variables using lasso procedure

#### Independent Principal Component Analysis

- assumes non Gaussian data distribution (!= PCA)
- 'blind source' signal separation
- seeks for a set of independent components (!= PCA)
- combines the advantages of both PCA and ICA
- the PCA loadings are transformed via ICA to obtain independent loading vectors and independent components
- sparse IPCA also developed to select the variables contributing to the independent components
- Yao, F. Coquery, J. and Lê Cao, K-A. 2012 Independent Principal Component Analysis
- unsupervised approach

#### PLS-Discriminant Analysis

- Similarly to Linear Discriminant Analysis, classical PLS-DA looks for the best combination of variables
- supervised approach
- sPLS-DA searches for discriminative variables that can help separating the samples
- evaluation of the discriminative power of the selected variables using external validation
- Lê Cao K-A., Boitard S. and Besse P. (2011) Sparse PLS Discriminant Analysis: bioinformatics

#### Integrative Omics Analysis

aims:

- unravel the correlation structure between two data sets
- select co-regulated biological entities across samples
- Partial Least Squares regression maximises the covariance between each linear combination (
- sparse PLS has been developed to include variable selection from both data sets
- two modes are proposed to model the relationship between the two data sets: regression and
- sPLS:

Sample plot:

- aims at selecting correlated variables across the same samples by performing a multivariate
- regression: explain the protein abundance w. r. t. the gene expression relationship
- the latent variables (components) are determined based on the selected genes and proteins
- unsupervised approach

Variable plot:

- relevance networks are bipartite graphs directly inferred from the sPLS components
- other insightful graphical outputs:
  - correlation circle plots
  - clustered image maps
- González I., Lê Cao K.-A., Davis, M.D. and Déjean S. Visualising association between proteomic and transcriptomic data
- canonical mode:
  - selects correlated variables across the same samples and highlights the correlation structure
- Arrow plot: highlight the similarities between 2 data sets.

Cross-over desing

One data set repeated measurements (n x p)

## supervised

## 1 level

Multivariate approach

multilevel PLS-DA

Internal variable selection

multilevel sPLS-DA

Graphical outputs: sample and variable plots

## 2 levels

Multivariate approach

multilevel PLS-DA

Internal variable selection

multilevel sPLS-DA

Graphical outputs: sample and variable plots

Two matching data sets repeated measurements ( $n \times p$ ) & ( $n \times q$ )

## canonical, 1 level

Multivariate approach

multilevel PLS

Internal variable selection

multilevel sPLS

Graphical outputs: sample and variable plots

## PCA

sPCA - sparse Principal Component Analysis

ICA - Independent Component Analysis

IPCA - Independent Principal Component Analysis

Yao, F. Coquery, J. and Lê Cao, K-A. 2012 Independent Principal Component Analysis for biological

LDA: Linear Discriminant Analysis

PLS-DA: PLS - Discriminant Analysis

sPLS-DA: sparse PLS Discriminant Analysis

Lê Cao K-A., Boitard S. and Besse P. (2011) Sparse PLS Discriminant Analysis: biologically relevant

Graphical methods:

- sample plot
- variable plot
- correlation circle plots
- clustered image maps

to know:

## E.9 Normalization

### E.9.1 The Universal exPression Code (UPC) algorithm consists of two main steps:

- i. for each platform, linear statistical models correct for background noise by modeling the genomic background
- ii. estimates of transcriptional activation are calculated using a two-component mixture model, where the background is modeled as a mixture of two components: a uniform background and a transcriptional activation component.

QUESTIONS:

1. how do you correct something using linear models?
2. how the genomic base composition and length of target regions are incorporated in the correction?
3. what are two-component mixture model?

Let  $Y_i$  denote the unnormalized expression measurement for gene  $i$ .

Assume that

$$Y_i = (1 - \Delta_i) Y_{1i} + \Delta_i Y_{2i},$$

where:

- $Y_{1i}$  = random variable from the 'background' distribution for the gene;
- $Y_{2i}$  originates from the "background-plus-signal" distribution
- $\Delta_i$  is an unobserved indicator variable that is equal to 1 if gene  $i$  is active

QUESTIONS:

- what is the expectation-maximization (EM) algorithm?

The UPC value for gene  $i$ , denoted  $P_i$ , is given by the expected value of  $\Delta_i$ , given that the parameters  $\pi$ ,

## E.10 Time-Series Analysis

### E.10.0.1 Basics Time Series Analysis in R

- <http://luthuli.cs.uiuc.edu/~daf/courses/cs-498-daf-ps/lecture%2018%20-%20time%20series>
- <https://www.datacamp.com/community/tutorials/time-series-r>
- <https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling>
- [https://www.stat.pitt.edu/stoffer/tsa4/R\\_toot.htm](https://www.stat.pitt.edu/stoffer/tsa4/R_toot.htm)
- <https://nwfsc-timeseries.github.io/atsa-labs/sec-ts-time-series-plots.html>

### E.10.0.2 Time Series Similarity Measures

DTW (Dynamic Time Warping)

- [http://www.phon.ox.ac.uk/jcoleman/old\\_SLP/Lecture\\_5/DTW\\_explanation.html](http://www.phon.ox.ac.uk/jcoleman/old_SLP/Lecture_5/DTW_explanation.html)
- <https://pdfs.semanticscholar.org/57e1/704fd41ac85f57e68d75576645d7496c4e55.pdf>
- [http://seninp.github.io/assets/pubs/senin\\_dtw\\_litreview\\_2008.pdf](http://seninp.github.io/assets/pubs/senin_dtw_litreview_2008.pdf)
- <https://cran.r-project.org/web/packages/dtwclust/vignettes/timing-experiments.html>
- <https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html>

MIC (Maximal Information Coefficient)

DTW-MIC

Time Series Similarity Measure Comparison

- <[http://www.earthmapps.io/pubs/2011\\_Lhermitte\\_A%20comparison%20of%20time%20series.pdf](http://www.earthmapps.io/pubs/2011_Lhermitte_A%20comparison%20of%20time%20series.pdf)>

### E.10.1 Time-Series Clustering

Cluster analysis is a task which concerns itself with the creation of groups of objects, where each group is called a cluster. Ideally, all members of the same cluster are similar to each other, but are as dissimilar as possible from objects in a different cluster.

For static data, clustering methods are usually divided into:

- partitioning
- hierarchical
- density-based
- grid-based
- model-based

In the context of time-series, dimensionality of a series is related to time, and it can be understood as the length of the series.

então quando eu faço o calpsamento por idade, estou reduzindo a dimensionalidade

A single time-series object may be constituted of several values that change on the same time scale, in which case they are identified as multivariate time-series.

There are many techniques to modify time-series in order to reduce dimensionality, and they mostly deal with the way time-series are represented.

Time-series clustering is a type of clustering algorithm made to handle dynamic data.

most important elements to consider:

- (dis)similarity or distance measure
- prototype extraction function (if applicable)

- clustering algorithm
- cluster evaluation

In most cases, algorithms developed for time-series clustering take static clustering algorithms and either modify the similarity definition or the prototype extraction function to an appropriate one, or apply a transformation to the series so that static features are obtained. Therefore, the underlying basis for the different clustering procedures remains approximately the same across clustering methods.

The most common clustering methods for time-series are:

- hierarchical
- partitional and fuzzy

Classification of time-series clustering algorithms by Aghabozorgi et al. (2015), based on the way they treat the data and how the underlying grouping is performed:

1. whether the whole series, a subsequence, or individual time points are to be clustered
2. if the clustering itself may be shape-based, feature-based or model-based.

In this context, it is common to utilize the Dynamic Time Warping (DTW) distance as dissimilarity measure. The calculation of the DTW distance involves a dynamic programming algorithm that tries to find the optimum warping path between two series under certain constraints.

#### DISTANCE MEASURES

1. Dynamic Time Warping - DTW
2. Global Alignment Kernel - GAK
3. Shape-Based Distance - SBD

#### TIME-SERIES PROTOTYPES

1. Mean and median
2. Partition around medoids



3. DTW barycenter averaging
4. Shape extraction
5. Soft-DTW centroid
6. Fuzzy-based prototypes

#### TIME-SERIES CLUSTERING ALGORITHMS

1. Hierarchical clustering
2. Partitional clustering
  - 2.1 TADPole clustering
  - 2.2 k-Shaped clustering
3. Fuzzy clustering

#### CLUSTER EVALUATION

- cluster validity indices (CVIs)
  - Arbelaitz et al.(2013) and Wang and Zhang (2007);
- CVIs can be either tailored to crisp or fuzzy partitions.
- crisp:
- internal, external or relative

### E.10.2 Time-Series Options

- linear models

Linear models and empirical bayes methods for assessing differential expression in microarray

- empirical Bayes

A multivariate empirical Bayes statistic for replicated microarray time course data

- fuzzy algorithms

Mfuzz: a software package for soft clustering of microarray data.

- Bayesian approaches

An improved empirical bayes approach to estimating differential gene expression in

Desktop Java Applications:

- Short Time-series Expression Miner (STEM)

STEM: a tool for the analysis of short time series gene expression data

- Bayesian Analysis of Time Series (BATS)

BATS: a Bayesian user-friendly software for analyzing time series microarray exper

- GenT Warper

Gene Time Expression Warper: a tool for alignment, template matching and visualiza

- EDGE

EDGE: extraction and analysis of differential gene expression.

- regression-based (maSigPro)

maSigPro: a method to identify significantly differential expression profiles in time-

- multivariate approaches (ASCA-genes)

Discovering gene expression patterns in time course microarray experiments by ANOVA-SC

- specific methodologies for functional and gene-set enrichment analysis (maSig-Fun, PCA-maSigFun and ASCA-functional)

Functional assessment of time course microarray data

- Serial Expression Analysis (SEA)

About time-series:

time series data are often mathematically nonstationary and show autocorrelation, creating nonlinearities or discontinuities in the data, which limit the use of many statistical techniques that assume fixed or Gaussian probability distributions (47).

47.

A further complication of analyzing time series data is that expression patterns in pairs of genes are often not monotonically correlated, rendering inappropriate commonly used nonparametric tests of association, such as the Spearman rank correlation (48, 49).

48.

49.

Time-series analysis methods:

- Dynamic Time Warping
- Independent Component Analysis
- Mutual information-based network analysis

Dynamic Time Warping (DTW)

DTW creates this distance measure by locally compressing or stretching (warping) one trace to

DTW provides high power to detect differences in time series by accounting for the trajectory

DTW distance increased linearly with random noise and in proportion to the length of random t

- linear models

Linear models and empirical bayes methods for assessing differential expression in microarray

- empirical Bayes

A multivariate empirical Bayes statistic for replicated microarray time course data

- fuzzy algorithms

Mfuzz: a software package for soft clustering of microarray data.

- Bayesian approaches

An improved empirical bayes approach to estimating differential gene expression in microarray

Desktop Java Applications:

- Short Time-series Expression Miner (STEM)

STEM: a tool for the analysis of short time series gene expression data

- Bayesian Analysis of Time Series (BATS)

BATS: a Bayesian user-friendly software for analyzing time series microarray exper-

- GenT Warper

Gene Time Expression Warper: a tool for alignment, template matching and visualization

- EDGE

EDGE: extraction and analysis of differential gene expression.

- regression-based (maSigPro)

maSigPro: a method to identify significantly differential expression profiles in time-

- multivariate approaches (ASCA-genes)

Discovering gene expression patterns in time course microarray experiments by ANOVA-SC

- specific methodologies for functional and gene-set enrichment analysis (maSigFun, PCA-maSigFun and ASCA-functional)

Functional assessment of time course microarray data

- Serial Expression Analysis (SEA)

About time-series:

time series data are often mathematically nonstationary and show autocorrelation, creating nonlinearities or discontinuities in the data, which limit the use of many statistical techniques that assume fixed or Gaussian probability distributions (47).

47.

A further complication of analyzing time series data is that expression patterns in pairs of genes are often not monotonically correlated, rendering inappropriate commonly used nonparametric tests of association, such as the Spearman rank correlation (48, 49).

48.

49.

Time-series analysis methods:

- Dynamic Time Warping
- Independent Component Analysis
- Mutual information-based network analysis

Dynamic Time Warping (DTW)

DTW creates this distance measure by locally compressing or stretching (warping) one trace to

DTW provides high power to detect differences in time series by accounting for the trajectory

DTW distance increased linearly with random noise and in proportion to the length of random t

## E.11 Survival analysis

<https://www.youtube.com/watch?v=fTX8GghbBPc&list=PLRW9kMvtNZOjXOgq4XTBTqdLb89YW1Ygc>

ALIAS: duration/transition/failure time/time-to-event analysis

Survival analysis set up

- subjects are tracked until an event happens (failure) or wwe lose then frin the sample (censored)
- we are interested in how long they stay in the sample (survival)
- we are also interested in their risk of failure (hazard rates)

Survival analysis features

- the dependent variable is duration (time to event or time to being censored) so it is a combination of
  - time variable = length of time until the event happended or as lond as they are in the stud
  - the event variable: 1 if the event happened or 0 if the event has not yet happened

- instead of an event variable, a censor variable can be defined The censored variable is the time at which the observation is censored
- hazard rate: is the probability that the event will happen at time  $t$  given that the event has not happened before time  $t$
- Hazard rates usually change over time.
- the probability of defaulting on a loan may be low in the beginning but increases over time

Extensions of the basic survival analysis

- + Multiple occurrences of event (multiple observations per individual)
  - borrower may have repeated restructuring of the loan
  - firm may adopt technology in some year but not others
- + More than one type of event (include codes for events e.g., 1, 2,3,4)
  - borrower may default (one type of event) the loan earlier (a second type of event)
  - = firms may adopt different types of technologies
- + Two groups of participants
  - the effect of two types of educational programs on technology adoption rates
- + Time-varying covariates
  - borrower's income may have changed during the study which caused the default.
- + Discrete instead of continuous transition times
  - events are measured in intervals (such as every month)
- + there may be different starting times - we need to measure time from the beginning time

## SURVIVAL, HAZARD, AND CUMULATIVE HAZARD FUNCTIONS

- + the dependent variable duration is assumed to have a continuous probability distribution  $f(t)$
- + the probability that the duration time will be less than  $t$  is:
 
$$F(t) = \text{Prob}(T \leq t) = \int_0^t f(s) ds$$
- + Survival function is the probability that the duration will be at least  $t$ :

$$S(t) = 1 - F(t) = \text{Prob}(T \geq t)$$

+ Survival function is the probability that the duration will be at least t:

$$S(t) = 1 - F(t) = \text{Prob}(T \geq t)$$

+ Hazard rate is the probability that the duration will end after time t, given that it has lasted t:

$$\lambda(t) = f(t)/S(t)$$

+ the hazard rate is the probability that an individual will experience the event at time t while surviving up to time t

## NONPARAMETRIC MODELS

nonparametric estimation is useful for descriptive purposes and to see the shape of the hazard over time

- Think about the shapes of the hazard function and survival function over time

source: <https://www.datacamp.com/community/tutorials/survival-analysis-R>

In this type of analysis (Survival Analysis) the time to a specific event is of interest and two (or more) groups of patients are compared with respect to this time.

Three core concepts can be used to derive meaningful results from such a dataset:

+ The Statistics behind Survival Analysis

- Kaplan-Meier Method and Log-Rank Test

- Cox Proportional Hazards Models

This type of analysis can answer questions such as the following:

- do patients benefit from therapy regimen A as opposed to regimen B?

- Do patients' age and fitness significantly influence the outcome?

- Is residual disease a prognostic biomarker in terms of survival?

Terminology:

- event: is the pre-specified endpoint of your study (e.g. death, disease recurrence)

- censoring: refers to incomplete data. All patients who do not experience the "event"
  - right-censoring
    - fixed or random type I censoring
    - type II censoring
- Covariates: aka explanatory or independent variables in regression analysis, are variables

Statistics:

- Kaplan-Meier Method and Log Rank Test

Kaplan-Meier estimator:

0 independently described by Edward Kaplan and Paul Meier, and conjointly published

- is a non-parametric statistic
- allows us to estimate the survival function

OBS: makes sense to use non-parametric statistic since survival data has a skewed distribution

- this statistic gives the probability that an individual patient will survive past a certain time point
  - at  $t = 0$ , the Kaplan-Meier estimator is 1 and with  $t$  going to infinity, the estimator goes to 0
  - in theory, with an infinitely large dataset and  $t$  measured to the second, the estimator is exact
- assumption: the probability of surviving past a certain time point  $t$  is equal to the product of the probabilities of surviving past each time point up to  $t$ 
  - $S(t) = p.1 \times p.2 \times \dots \times p.t$ , with:

$S(t)$  = the survival probability at time  $t$

$p.1$  = proportion of all patients surviving past the first time point

$p.2$  = proportion of all patients surviving past the second time point

... and so forth until time point  $t$  is reached.

- IMPORTANT: starting with  $p.2$  and up to  $p.t$ , you take only those patients into account who survived past the first time point
  - => thus,  $p.2, p.3, \dots, p.t$  are proportions that are conditional on the previous time points



- in practice, you want to organize the survival times in order of increasing duration first.
- Censored patients are omitted after the time point of censoring, so they do not influence
- Log-rank test:
  - Can use to compare survival curves of two groups.
  - is a statistical hypothesis test that tests the null hypothesis that survival curves of two groups are equal.
  - a certain probability distribution, namely a chi-squared distribution, can be used to derive p-values.
  - compares two Kaplan-Meier survival curves, which might be derived from splitting a patient group into two groups.
- Cox Proportional Hazards Models
  - $h(t)$ : hazard function
  - describes the probability of an event or its hazard  $h$  (survival in this case) if the subject is at risk at time  $t$ .
  - it measures the instantaneous risk of death
  - you need the hazard function to consider covariates when you compare survival of patient groups.
  - are derived from the underlying baseline hazard functions of the patient populations in question.
  - does not assume an underlying probability distribution but it assumes that the hazards of two groups are proportional.
  - allow to include covariates
  - forest plot. It shows so-called hazard ratios (HR) which are derived from the model for all covariates.
  - $HR > 1$  indicates an increased risk of death (according to the definition of  $h(t)$ ) if a subject belongs to the group with the higher hazard.

## E.12 Data Integration

### E.12.1 Sparse canonical methods for biological data integration: application to a cross-platform study

ABSTRACT

Results:

We compare the results obtained with two other sparse or related canonical correlation

- CCA with Elastic Net penalization (CCA-EN)
- Co-Inertia Analysis (CIA)
  - does not include a built-in procedure for variable selection
  - requires a two-step analysis
- There is a lack of statistical criteria to evaluate canonical correlation methods

Conclusions:

- sPLS and CCA-EN selected highly relevante genes and complementary findings from t
- These two approaches were found to bring similar results, although they highlight
- They outperformed CIA that tended to select redundant information

## BACKGROUND

Few approaches exists to deal with high-throughput data sets:

linear multivariate models:

Partial Least Squares regression (PLS, [1])

Canonical Correlation Analysis (CCA, [2])

Problems:

- are often limited by the size of the data set (ill-posed problems, CCA)
- the noisy and the multicollinearity characteristics of the data (CCA)
- lack of interpretability (PLS)
- PLS has often been criticized for its lack of theoretical justifications. Much w

Advantages:

1. because they allow for the compression of the data into 2 to 3 dimensions f
2. because their resulting components and loading vectors capture dominant and

#### Canonical correlation framework

- there is either no assumption on the relationship between the two sets of variables (exploratory)
- When applying canonical correlation-based methods, most validation criteria used in a regression framework
- Some sparse associated integrative approaches have recently been developed to include a priori information
- Approaches:
  - penalized CCA adapted w/ Elastic Net (CCA-EN[10])
  - Co-Inertia Analysis (CIA[11])

This study propose to apply a sparse canonical approach called "sparse PLS" (sPLS) for the integrative analysis

- provides variable selection of two-block data sets in a one step procedure, while integrating prior information
- Methodological aspects and evaluation of sPLS in a regression framework:

[9]

#### Canonical correlation-based methods

focus on two-block data matrices:  $X(n \times p)$  and  $Y(n \times q)$ , where  $p$  and  $q$  are of two types, measured and unmeasured

- Prior biological knowledge on these data allows us to settle into a canonical framework, i.e. CCA
- the large number of variables may affect the exploratory method, due to numerical issues (e.g. multicollinearity)

Three types of multivariate methods: CCA, PLS, CIA

#### CCA

##### Canonical Correlation Analysis

- studies the relationship between two sets of data
- the CCA  $n$ -dimensional score vectors  $(X_{ah}, Y_{bh})$  come in pair to solve the objective function
- the aim of CCA is to simultaneously maximize  $\text{cov}(X_{ah}, Y_{bh})$  and minimize the variances of  $X_{ah}$  and  $Y_{bh}$

#### PLS

- Partial Least Squares regression

- based on the simultaneous decomposition of  $X$  and  $Y$  into latent variables and
- The latent variables methods (e.g. PLS, Principal Component Regression) assume
- These latter may correspond to some biological underlying phenomena which are
- Like CCA, the PLS latent variables are linear combinations of the variables,
- In contrary to CCA, the loading vectors ( $a_h$ ,  $b_h$ ) are interpretable and can be
- Many PLS algorithms exist:
  - for different shapes of data (SIMPLS, [18], PLS1 and PLS2 [1], PLS-SVD [19])
  - different aims:
    - predictive, like PLS2
    - modelling, like PLS-mode A, see [10,20,21]
- In this study we especially focus on a modelling aim ("canonical mode") between

#### CCA-EN

- proposed by [10]
- sparse penalized variant of CCA using Elastic Net [8,22] for a canonical frame
- Elastic Net: combines the advantages of the ridge regression, that penalizes
- However, when  $p + q$  is very large, the resolution of the optimization problem

#### sparse PLS

- proposed by [9]
- sparse PLS approach (sPLS) based on a PLS-SVD variant, so as to penalize both
- sparsity can then be introduced by iteratively penalizing  $a_h$  and  $b_h$  with a soft

#### CIA

- Co-Inertia analysis
- introduced by [11], applied to ecological data

- first application to biological data [12]
- suitable for a canonical framework, as it is adapted for a symmetric analysis.
- It involves analyzing each data set separately either with principal component analyses
- This results in two sets of axes, where the first pair of axes are maximally co-variant

#### Differences between the approaches

- profoundly differ in their construction and aims
- CCA-EN looks for canonical variate pairs  $(X_{a,h}, Y_{b,h})$ , such that a penalized version
  - This explains why a non monotonic decreasing trend in the canonical correlation can
- On the other hand, sPLS (canonical mode) and CIA aim at maximizing the covariance between
- However, here CIA is based on the construction of two Correspondence Analyses, whereas

#### Parameters tuning

- In CCA-EN, the authors proposed to tune the penalty parameters for each dimension, such
  - In practice, they showed that the correlation did not change much when more variables
  - Therefore, an appropriate way of tuning the parameters would be to choose instead to
    - Thus, depending on the aim of the study (focus on few genes or on groups of genes)
      - When focusing on groups of genes (e.g. pathways, transcription factor targets)
    - The same strategy will be used for sPLS (see also [9] where the issue of tuning
  - No other parameters than the number of selected variables is needed in CIA either

#### Outputs

##### Samples

Samples are represented with the scores or latent variable vectors, in a superimposed

1. show how samples are clustered, based on their biological characteristics
2. measure if both data sets strongly agree according to the applied approach

- each sample is indicated using an arrow.

- The start of the arrow indicates the location of the sample in the X

- Thus, short (long) arrows indicate if both data sets strongly agree

#### Variables

Variables are represented on correlation circles, as previously proposed by

Correlations between the original data sets and the score or latent variable

Only the selected variables in each dimension are represented.

This type of graphic not only allows for the identification of interactions

#### Cross-platform study

Data sets and relevance for canonical correlation analysis

The Ross Data Set

The Staunton Data set

Application of the three sparse canonical correlation-based methods

## RESULTS AND DISCUSSIONS

#### How to assess the results?

- Canonical correlation-based methods are statistically difficult to assess\>

1. they do not fit into a regression/prediction framework, meaning that the pr

2. because in many two-block biological studies, the number of samples  $n$  is ve

- This is why graphical outputs are important to help analyze the results

#### Link between two-block data sets

Variance explained by each component

Correlations between each component

Interpretation of the observed cell line clusters

- Graphical representation of the samples

- Hierarchical clustering of the samples

Interpretation of the observed genes clusters

- Graphical representation of the genes

- Analysis of the gene lists

- Analysis of the gene lists with IPA

- Over-represented biological functions

- Canonical pathways

- Networks

CONCLUSION

CIA

CIA does not propose a built-in variable selection procedure and requires a two-step analysis

However, the loadings or weight vectors obtained were not orthogonal, in contrary to CCA-EN and

CCA-EN

CCA-EN first captured the main robust effect on the individuals that was present in the two datasets

This explains why the canonical correlations do not monotonically decrease. The only difference is

sPLS

We found that sPLS made a good compromise between all these approaches. It includes variable

Based on the present study, we would primarily recommend the use of CCA-EN or sPLS when gene selection is an issue. Like CCA-EN, sPLS includes a built-in variable

selection procedure but captured subtle individual effects. Therefore, these two approaches may differ when computing the first axes. All approaches are easy to use and fast to compute. These approaches would benefit from the development of an R package to harmonize their inputs and outputs so as to facilitate their use and their comparison.

## E.13 Single-Cell RNA-Seq

<https://hemberg-lab.github.io/scRNA.seq.course/biological-analysis.html#pseudotime-analysis>

### E.13.1 Analysis

single-cell RNA-Seq PACKAGES: - Conos (Clustering on Network of Samples)  
+ <https://github.com/hms-dbmi/conos> - It's a package to wire together large collections of single-cell RNA-seq datasets. - It focuses on uniform mapping of homologous cell types across heterogeneous sample collections.

- SCDE
  - + <http://hms-dbmi.github.io/scde/index.html>
  - implements a set of statistical methods for analyzing single-cell RNA-seq data
  - Single cell error modeling
    - fits individual error models for single cells using counts derived from single cells
  - Differential expression analysis
    - compares groups of single cells and tests for differential expression, taking into account the error model
  - Pathway and gene set overdispersion analysis
    - contains pagoda routines that characterize aspects of transcriptional heterogeneity
- PAGODA
  - <https://github.com/hms-dbmi/pagoda2>
  - <http://pklab.med.harvard.edu/nikolas/pagoda2/frontend/current/pagodaURL/index.html>
  - <http://pklab.med.harvard.edu/scde/pagoda.links.html>
- VELOCITY
  - + <http://velocity.org/>
  - analysis of expression dynamics in single cell RNA seq data.
  - enables estimations of RNA velocities of single cells by distinguishing unspliced and spliced isoforms
  - <http://velocity.org/velocity.py/tutorial/cli.html>

TUTORIALS:

<https://hemberg-lab.github.io/scRNA.seq.course/cleaning-the-expression-matrix.html>

## E.14 Reproducible Data Science

Reproducible Data Science w/ R



<<https://resources.rstudio.com/rstudio-conf-2019/a-guide-to-modern-reproducible-data-science-with>

Research compendia

"...We introduce the concept of a compendium as both a container for the different elements t

Research compendium principles

- stick with the conventions of your peers
- Keep data, methods and outputs separate
- Specify your computational environment as clear as you can

Key components you'll need for sharing a compendium:

License + VCS + Metadata + Archive

compendium DESCRIPTION file

Type: Compendium

Package: pomdpintro

Version: 0.1.0

Depends: nimble, tidyverse, sarsop, MDPtoolbox

Suggests: extrafont, hrbrthemes, Cairo, ggthemes

Remotes: boettiger-lab/sarsop

Packaging your analysis as a compendium gives you access to powerfull developer tools

Small compendia

COMPENDIUM

\\

\\|--- DESCRIPTION

\\

\\|--- LICENSE

```

\|
\|---  Readme.md
\|
\|---  data/
\|      \|
\|      '----- Mydata.csv
\|
'---    analysis/
        \|
        '---  Report.Rmd

```

#### Components of a compendium

##### Data (Small -\> Medium)

How does one manage small to medium data in the context of a research compendium?

##### Small data

Put small data inside packages, especially if you ship a methods package

CRAN = \< 5 mb

37% of the 13K packages on CRAN have some form of data.

Leveraging Github releases to share medium sized files

piggyback

attach large [data] files to Github repositories

[github.com/ropensci/piggyback](https://github.com/ropensci/piggyback)

```
pb_new_release('user/repo', 'v0.0.5')
```

```
pb_upload('datasets.tsv.xz','user/repo')
```

Access them in your scripts with

```
pb_download
```

Medium data

```
github.com/ropensci/arkdb
```

Computing environment

Its important to isolate the computing environment so that changes in software dependencies

Adding a Dockerfile to your compendium

Many ways to write a Dockerfile for your R project

```
o2r/containerit
```

Binder

```
mybinder.org
```

Binder is an open source project that is designed to make it really easy to share and

Git + Docker + RStudio

Workflows

Include a workflow to manage relationships between data output and code

drake

general purpose workflow manager & pipeline toolkit for reproducibility and high-performance

```
github.com/ropensci/drake
```

No cumbersome Makefiles

Vast arsenal of parallel computing options

Visualize dependency graph and estimate run times

Convenient organization of output

RESEARCH COMPENDIUM

<<https://research-compendium.science>>

<<https://github.com/research-compendium/research-compendium.github.io>>

<<http://inundata.org/talks/rstd19/#/>>

<<https://github.com/karthik/rstudio2019>>

<<https://biostats.bepress.com/cgi/viewcontent.cgi?article=1001&context=bioconductor>>

## E.15 machine learning vs DE Analysis

<<https://www.biostars.org/p/305532/>>

## E.16 Dealing with Missing Data

MISSING DATA

- Resources

<<https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-data/>>

- DATACAMP

- "The best thing to do with missing data is to not have any" - Gertrude Mary Cox

- Missing data can have unexpected effects on your analysis

- Bad imputation can lead to poor estimates and decisions

- missing data = NA, Not Available

- R consider NaN (Not a Number) as NA

- missingness summaries

- `any_na(x)`

- `are_na(x)`
- `n_miss(x)`
- `prop_miss(x)`
  
- basic summaries:
  - `n_miss`
  - `n_complete`
- dataframe summaries:
  - `miss_var_summary`
  - `miss_case_summary`
- missing data tabulations
  - `miss_var_table`
  - `miss_case_table`
- `miss_var_span()`
- `miss_var_run()` find repeated patterns of missing data in a run

Visualizations

- `viss_miss()`
- `gg_miss_var()`
- `gg_miss_case()`
- `gg_miss_upset()`
- `gg_miss_fct()`
- `gg_miss_span()`

## E.17 Feature Selection

### E.17.1 Features and feature engineering

The goals of a good feature are to simultaneously vary with what matters and be invariant with what does not.

A natural question is whether or not we can select good features automatically. This problem is known as feature selection. There are many methods that have been proposed for this problem, but in practice, very simple ideas work best. It does not make sense to use feature selection in these small problems, but if you had thousands of features, throwing out most of them might make the rest of the process much faster.

spectral feature selection r

<https://cran.r-project.org/web/packages/sparcl/index.html>

[https://www.researchgate.net/post/What\\_is\\_the\\_best\\_unsupervised\\_method\\_for\\_feature\\_subset\\_selection](https://www.researchgate.net/post/What_is_the_best_unsupervised_method_for_feature_subset_selection)

<https://stats.stackexchange.com/questions/108743/methods-in-r-or-python-to-perform-feature-selection-in-unsupervised-learning>

<https://www.google.co.uk/search?client=opera&q=unsupervised+variable+selection+r&sourceid=opera&ie=UTF-8&oe=UTF-8>

<https://www.datacamp.com/community/tutorials/introduction-t-sne>

[https://cran.r-project.org/web/packages/tsna/vignettes/tsna\\_vignette.html](https://cran.r-project.org/web/packages/tsna/vignettes/tsna_vignette.html)

## E.18 Exploratory Data Analysis (EDA)

Iterative cycle:

1. Generate questions about your data
2. Search for answers by visualising, transforming, and modelling your data
3. Use what you learn to refine your questions and/or generate new questions

Your goal during EDA is to develop an understanding of your data. The easiest way to do this is to use questions as tools to guide your investigation. When you ask a question, the question focuses your attention on a specific part of your dataset and helps you decide which graphs, models, or transformations to make.

EDA is fundamentally a creative process. And like most creative processes, the key to asking quality questions is to generate a large quantity of questions. It is difficult to ask revealing questions at the start of your analysis because you do not know what insights are contained in your dataset. On the other hand, each new question that you ask will expose you to a new aspect of your data and increase your chance of making a discovery.

However, two types of questions will always be useful for making discoveries within your data. You can loosely word these questions as:

1. What type of variation occurs within my variables?
2. What type of covariation occurs between my variables?

#### Variation

Variation is the tendency of the values of a variable to change from measurement to measurement.

Every variable has its own pattern of variation, which can reveal interesting information. The be

Questions about histograms:

- Which values are the most common? Why?
- Which values are rare? Why? Does that match your expectations?
- Can you see any unusual patterns? What might explain them?

Clusters of similar values suggest that subgroups exist in your data. To understand the subgroups

- How are the observations within each cluster similar to each other?
- How are the observations in separate clusters different from each other?
- How can you explain or describe the clusters?
- Why might the appearance of clusters be misleading?

#### Outliers

It's good practice to repeat your analysis with and without the outliers. If they have minimal ef

#### Covariation

If variation describes the behavior within a variable, covariation describes the behavior between

## Visualisation and Plots

histogram - for several histogram, use `geom_freqpoly()`

density - is the count standardised so that the area under each frequency polygon is one.

### Two categoriacal variables

`geom_count()`

`geom_tile()` If the categorical variables are unordered, you might want to use the s

### Two continuous variables

- `geom_point()`

- For large datasets: `geom_bin2d()` and `geom_hex()` divide the coordinate plane in

- Another option is to bin one continuous variable so it acts like a categorical v

- `geom_boxplot(mapping = aes(group = cut_width(carat, 0.1)))`

- Another approach is to display approximately the same number of points in ea

## Patterns and models

### Patterns

Patterns in your data provide clues about relationships. If a systematic relationsh

Could this pattern be due to coincidence (i.e. random chance)?

How can you describe the relationship implied by the pattern?

How strong is the relationship implied by the pattern?

What other variables might affect the relationship?

Does the relationship change if you look at individual subgroups of the data?

Patterns provide one of the most useful tools for data scientists because they rev

### Models



Models are a tool for extracting patterns out of data.

It's hard to understand the relationship between two variables when they are tightly related.

It's possible to use a model to remove the very strong relationship between them so we can explore.

This can be done by fitting a model that predicts one variable from another and then computing residuals.

```
library(modelr)

mod <- lm(log(price) ~ log(carat), data = diamonds)

diamonds2 <- diamonds %>%

  add_residuals(mod) %>%

  mutate(resid = exp(resid))

ggplot(data = diamonds2) +

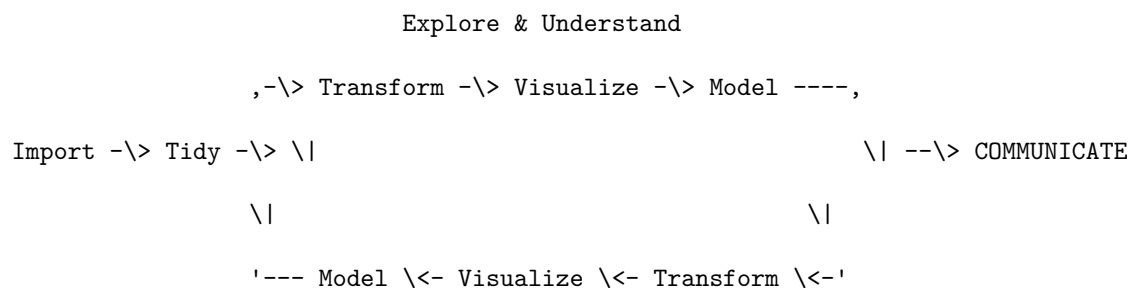
  geom_point(mapping = aes(x = carat, y = resid))
```

Once you've removed the strong relationship between carat and price, you can see what you can learn from the data.

```
ggplot(data = diamonds2) + geom_boxplot(mapping = aes(x = cut, y = resid))
```

## Exploratory Data Analysis

Diagram from: R for Data Science



## INTERACTIVE GRAPHICS AUGMENT EXPLORATION

Interactive graphics can augment exploratory analysis, but are only practical when we can interactively explore the data.

iDENTIFY STRUCTURE THAT OTHERWISE GOES MISSING

sEARCH FOR INFORMATION QUICKLY WITHOU FULLY SPECIFIED QUESTIONS

mULTIPLE LINKED VIEWs are the optimal framework for posing queries about data

Diagnose and understand models

See: visual (Majumder et al 2013) and post-selection (Berk et al 2013) inference frames

<<https://plotly-book.cpsievert.me>> \ | <<https://plotly-r.com/introduction.html>>

Quantile Quantile Plots (qq-plots)

To corroborate that a theoretical distribution, for example the normal distribution, is

Quantiles are best understood by considering the special case of percentiles. The p-tl

HOW TO DISPLAY DATA BADLY

Karl W. Broman - <<http://kbroman.org/pages/talks.html>>

General principles

The aims of good data graphics is to display data accurately and clearly. According to

Display as little information as possible.

Obscure what you do show (with chart junk).

Use pseudo-3D and color gratuitously.

Make a pie chart (preferably in color and 3D).

Use a poorly chosen scale.

Ignore significant figures.

Displaying data well

In general, you should follow these principles:

Be accurate and clear.

Let the data speak.

Show as much information as possible, taking care not to obscure the message.

Science not sales: avoid unnecessary frills (esp. gratuitous 3D).

In tables, every digit should be meaningful. Don't drop ending 0's.

Some further reading:

ER Tufte (1983) The visual display of quantitative information. Graphics Press.

ER Tufte (1990) Envisioning information. Graphics Press.

ER Tufte (1997) Visual explanations. Graphics Press.

WS Cleveland (1993) Visualizing data. Hobart Press.

WS Cleveland (1994) The elements of graphing data. CRC Press.

A Gelman, C Pasarica, R Dodhia (2002) Let's practice what we preach: Turning tables into graphs. The American Statistician 56:121-130

NB Robbins (2004) Creating more effective graphs. Wiley.

Nature Methods columns

## CORRELATIONS

For two given highly correlated technical replicates, To examine how well the second vector repr

These are referred to as Bland-Altman plots, or MA plots in the genomics literature, and we will

### Misunderstanding Correlation (Advanced)

The use of correlation to summarize reproducibility has become widespread in, for example, genomics. Despite its English language definition, mathematically, correlation is not necessarily informative with regards to reproducibility. Here we briefly describe three major problems.

The most egregious related mistake is to compute correlations of data that is not approximated by bi-variate normal data. As described above, averages, standard deviations and correlations are popular summary statistics for two-dimensional data because, for the bivariate normal distribution, these five parameters fully describe the distribution. However, there are many examples of data that are not well approximated by bivariate normal data. Gene expression data, for example, tends to have a distribution with a very fat right tail.

The standard way to quantify reproducibility between two sets of replicated measurements, say  $x_1, \dots, x_n$

and  $y_1, \dots, y_n$

, is simply to compute the distance between them:

$$i=1 \text{ to } n \quad d_i = \sqrt{(x_i - y_i)^2}$$

This metric decreases as reproducibility improves and it is 0 when the reproducibility is perfect. Another advantage of this metric is that if we divide the sum by  $N$ , we can interpret the resulting quantity as the standard deviation of the  $d_1, \dots, d_N$ .

If we assume the  $d$  average out to 0. If the  $d$

can be considered residuals, then this quantity is equivalent to the root mean squared error (RMSE), a summary statistic that has been around for over a century. Furthermore, this quantity will have the same units as our measurements resulting in a more interpretable metric.

Another limitation of the correlation is that it does not detect cases that are not reproducible due to average changes. The distance metric does detect these differences. We can rewrite:

$$\sum_{i=1}^n (x_i - y_i)^2 = \sum_{i=1}^n [(x_i - \bar{x}) - (y_i - \bar{y}) + (\bar{x} - \bar{y})]^2$$

with  $\bar{x}$

and  $\bar{y}$

the average of each list. Then we have:

$$\sum_{i=1}^n (x_i - y_i)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2 + (\bar{x} - \bar{y})^2 + 2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

For simplicity, if we assume that the variance of both lists is 1, then this reduces to:

$$\sum_{i=1}^n (x_i - y_i)^2 = 2 + (\bar{x} - \bar{y})^2 - 2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

with

the correlation. So we see the direct relationship between distance and correlation. However, an important difference is that the distance contains the term  $(\bar{x} - \bar{y})^2$

and, therefore, it can detect cases that are not reproducible due to large average changes.

Yet another reason correlation is not an optimal metric for reproducibility is the lack of units. To see this, we use a formula that relates the correlation of a variable with that variable, plus what is interpreted here as deviation:  $x$

and  $y = x + d$ . The larger the variance of  $d$ , the less  $x + d$  reproduces  $x$ . Here the distance metric would depend only on the variance of  $d$  and would summarize reproducibility. However, correlation depends on the variance of  $x$  as well. If  $d$  is independent of  $x$

, then

$$\text{cor}(x, y) = \frac{1}{\sqrt{1 + \text{var}(d)/\text{var}(x)}}$$

This suggests that correlations near 1 do not necessarily imply reproducibility. Specifically, irrespective of the variance of  $d$

, we can make the correlation arbitrarily close to 1 by increasing the variance of  $x$ .

Robust summaries and log transformation

The normal approximation is often useful when analyzing life sciences data. However, due to the c  
n statistics we refer to these type of points as outliers. A small number of outliers can throw c

The median

The median, defined as the point having half the data larger and half the data smaller, is a

The median absolute deviation

The median absolute deviation (MAD) is a robust summary for the standard deviation. It is def

$1.4826 \times \text{median}\{|X_i - \text{median}(X_i)|\}$

The number 1.4826

is a scaling factor such that the MAD is an unbiased estimate of the standard deviation. Noti

Spearman correlation

Earlier we saw that the correlation is also sensitive to outliers.

The Spearman correlation follows the general idea of median and MAD, that of using quantiles.

So if these statistics are robust to outliers, why would we ever use the non-robust version?

We also note that there is a large statistical literature on Robust Statistics that go far beyond

Symmetry of log ratios

Ratios are not symmetric.

Rank tests

Wilcoxon Rank Sum Test

We learned how the sample mean and SD are susceptible to outliers. The t-test is based on the

The basic idea is to 1) combine all the data, 2) turn the values into ranks, 3) separate them

### E.18.1 Workflow: EDA for Transcriptomics

#### 1. Select meaningful variables from pheno data

- check for pvca assumptions for pheno vars
- identify categorical variables with constant values across all samples or specific factor
- identify numerical variables with unique/constant values across all samples
- give a warning for numerical variables with unique values, in case they are some sort of factor
- identify variables with NA values or with NA percentage above some given threshold (0.1)

#### 2. Density plot

##### Microarray

Density plots give you an idea about the signal distribution across a chip.

The default density plot function in the limma package (`plotDensities`) in Bioconductor

Arrays that have very different distributions in these images should be checked carefully

#### 3. Sample clustering

Clustering methods are commonly used to look for patterns in microarray data. However,

Careful experimental design and scheduling should result in problems such as this being

With these data, it is likely that two routes could be considered. If there were enough

- Heatmap ( $1 - \text{Cor}$ )
- Dendrograms
  - Complete linkage - Euclidean distance
    - Maximum dissimilarity between points in two sets used to determine which two
    - Often gives comparable cluster sizes.
    - Less sensitive to outliers.

- Works better with spherical distributions.
- Single linkage
  - Minimum dissimilarity between points in two sets used to determine which two sets should be joined.
  - Can handle diverse shapes.
  - Very sensitive to outliers or noise.
  - Often results in unbalanced clusters.
  - Extended, trailing clusters in which observations fused one at a time-chaining.
- Average linkage
  - Average dissimilarity between points in two sets used to determine which two sets should be joined.
  - A compromise between single and complete linkage.
  - Less sensitive to outliers.
  - Works better with spherical distributions.
  - Similar linkage: Ward's linkage. Join objects that minimize Euclidean distance / average variance.

#### 4. MA plot

- Microarray data

MAplots show the relationship of signal ratios to signal intensities, where values are usually on a log scale.

For a large chip where most data is not expressed at different levels across the treatments, the MA plot can be used to identify differentially expressed genes.

- RNA-Seq data

#### 5. PVCA

- estimate the variability of experimental effects including batch
- The PVCA approach can be used as a screening tool to determine which sources of variability (biological, technical, or experimental) are most influential in the data.
- leverages the strengths of two very popular data analysis methods:

1. principal component analysis (PCA) is used to efficiently reduce data dimension
  2. variance components analysis (VCA) fits a mixed linear model using factors of in
- Using the eigenvalues associated with their corresponding eigenvectors as weights, as
  - Although PVCA is a generic approach for quantifying the corresponding proportion of v

## 6. RLE

- RLE Plots: Visualising Unwanted Variation in High Dimensional Data  
<https://arxiv.org/pdf/1704.03590.pdf>
- are a powerful tool for visualising such variation in high dimensional data
- are particularly useful for assessing whether a procedure aimed at removing unwanted

## 7. Multidimensional Scaling (MDS)

- Alternative dimensionality reduction approach
- Represents distances in 2D or 3D space
- Starts from distance matrix (PCA uses data points)

## 8. EDA with mixOmics

### 8.1 PCA

Principal Component Analysis (Jolliffe, 2005) is primarily used to explore one single

- In mixOmics, PCA is numerically solved in two ways:
  1. With singular value decomposition (SVD) of the data matrix, which is the most
  2. With the Non-linear Iterative Partial Least Squares (NIPALS) in the case of
- Input data should be centered (center = TRUE) and possibly (sometimes preferably)

Choosing the optimal parameters

We can obtain as many dimensions (i.e. number of PCs) as the minimum between the



The number of principal Components to retain (also called the number of dimensions) is th

```
tune.pca(X, ncomp = 10, center = TRUE, scale = FALSE)
```

## 8.2 IPCA

### Independant Principal Component Analysis

In some case studies, we have identified some limitations when using PCA:

- PCA assumes that gene expression follows a multivariate normal distribution and recent
- PCA decomposes the data based on the maximization of its variance. In some cases, the b

Instead, we propose to apply Independent Principal Component Analysis (IPCA) which combines t

The algorithm of IPCA is as follows:

1. The original data matrix is centered (by default).
2. PCA is used to reduce dimension and generate the loading vectors.
3. ICA (FastICA) is implemented on the loading vectors to generate independent loading ve
4. The centered data matrix is projected on the independent loading vectors to obtain the

IPCA offers a better visualization of the data than ICA and with a smaller number of componen

### Choosing the optimal parameters

The number of variables to select is still an open issue. In Yao et al (2012) we proposed

IPCA is of class sPCA and PCA, and most of the PCA graphical methods can be applied. The defa

### Kurtosis

The kurtosis measure is used to order the loading vectors to order the Independent Princi

```
ipca.res$kurtosis
```

### 8.3 PLS-DA

#### PLS Discriminant Analysis (PLS-DA)

Partial Least Squares was not originally designed for classification and discriminant analysis.

- PLS-Discriminant Analysis (PLS-DA, Barker and Rayens, 2003) is a linear discriminant analysis.
- sparse PLS-DA (sPLS-DA) enables the selection of the most predictive or discriminative variables.

Similar to a PLS-regression mode, the tuning parameters include the number of components and the regularization parameter.

```
plsda.res <- plsda(X, Y, ncomp = 5) where ncomp is the number of components
```

We use the function `perf` to evaluate a PLS-DA model, using 5-fold cross-validation.

```
set.seed(2543) for reproducibility here, only when the 'cpus' argument is used
```

```
perf.plsda <- perf(plsda.res, validation = "Mfold", folds = 5, progressBar = FALSE)
```

```
perf.plsda.srbct$error.rate error rates
```

```
plot(perf.plsda, col = color.mixo(1:3), sd = TRUE, legend.position = "horizontal")
```

Here `ncomp = 4` with max distance seems to achieve the best classification performance.

The AUROC can also be plotted, beware that it only complements the PLSDA performance.

### 9. MDP

See:

PCA, MDS, k-means, Hierarchical clustering and heatmap for microarray data

[https://rstudio-pubs-static.s3.amazonaws.com/93706\\_e3f683a8d77244a5b993b20ad6278f](https://rstudio-pubs-static.s3.amazonaws.com/93706_e3f683a8d77244a5b993b20ad6278f)

A Tutorial Review of Microarray Data Analysis

[http://www.ub.edu/stat/docencia/bioinformatica/microarrays/ADM/slides/A\\_Tutorial\\_Microarray\\_Data\\_Analysis](http://www.ub.edu/stat/docencia/bioinformatica/microarrays/ADM/slides/A_Tutorial_Microarray_Data_Analysis)

## E.19 Dimensionality Reduction

PCA (linear)

t-SNE (non-parametric/ nonlinear)

Sammon mapping (nonlinear)

Isomap (nonlinear)

LLE (nonlinear)

CCA (nonlinear)

SNE (nonlinear)

MVU (nonlinear)

Laplacian Eigenmaps (nonlinear)

PCA is a linear algorithm. It will not be able to interpret complex polynomial relationship between features. On the other hand, t-SNE is based on probability distributions with random walk on neighborhood graphs to find the structure within the data.

A major problem with, linear dimensionality reduction algorithms is that they concentrate on placing dissimilar data points far apart in a lower dimension representation. But in order to represent high dimension data on low dimension, non-linear manifold, it is important that similar datapoints must be represented close together, which is not what linear dimensionality reduction algorithms do.

Local approaches seek to map nearby points on the manifold to nearby points in the low-dimensional representation. Global approaches on the other hand attempt to preserve geometry at all scales, i.e mapping nearby points to nearby points and far away points to far away points

It is important to know that most of the nonlinear techniques other than t-SNE are not capable of retaining both the local and global structure of the data at the same time.

tSNE

Algorithm

<https://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/>

### 10. Common Fallacies

Following are a few common fallacies to avoid while interpreting the results of t-SNE:

For the algorithm to execute properly, the perplexity should be smaller than the number of points. Also, the suggested perplexity is in the range of (5 to 50)

Sometimes, different runs with same hyper parameters may produce different results.

Cluster sizes in any t-SNE plot must not be evaluated for standard deviation, dispersion or any other similar measures. This is because t-SNE expands denser clusters and contracts sparser clusters to even out cluster sizes. This is one of the reasons for the crisp and clear plots it produces.

Distances between clusters may change because global geometry is closely related to optimal perplexity. And in a dataset with many clusters with different number of elements one perplexity cannot optimize distances for all clusters.

Patterns may be found in random noise as well, so multiple runs of the algorithm with different sets of hyperparameter must be checked before deciding if a pattern exists in the data.

Different cluster shapes may be observed at different perplexity levels.

Topology cannot be analyzed based on a single t-SNE plot, multiple plots must be observed before making any assessment.

## E.20 Data Science Skills

### Descriptive Analytics

- Analyze historical data to answer "WHAT HAS HAPPENED TILL NOW?"

### Predictive Analytics

- "WHAT WILL HAPPEN IN THE FUTURE?"

-> Mathematics

-> Statistics

-> Data handling

- knowledge of ETL (Extract Transform and Load) operations on data and experience w
- comfortable in handling data from different sources and in different formats
- Excellent knowledge of SQL
- work with structured, semi-structured and unstructured data

Bonus:

- knowledge of Big Data tools and technologies
- Experience with NoSQL databases such as HBase, Cassandra and MongoDB

-> Expert in Analysing and Visualizing the data

- Experience working with popular data analysis and visualization packages in python
- Experience with popular data analysis and visualization tools such as Tableau, Mi

-> Goog communication and storytelling skills

-> Predictive analysics:

- Artificial intelligence
- data mining
- machine learning
- statistical modeling
- exposure to popular predictive analytics tools

## E.21 Cross-Normalization

Some methods for cross-study normalisation: - combining gene expression measures across independent studies (Wang et al. or Stevens and Doerge) Wang et al. DOI: 10.1093/bioinformatics/bth381 Stevens and Doerge DOI: 10.1186/1471-2105-6-57 - combining other measures such as rank-ordering (as in RankProd) DOI: <https://doi.org/10.1093/bioinformatics/btl476> - or p-values (Rhodes et al.) PMID: 12154050 - the Bayesian approaches (e.g. Conlon et al.). DOI: 10.1186/1471-2105-7-247

Every method has caveats, issues and the more trivial the solution the more caveats.

There is a BioConductor package called MADAM which implements some meta-analysis methods including the RankProducts approaches. There are also cross-study approaches implemented in the web-based analysis tool ArrayMining.net including empirical Bayes approaches (as in ComBat), median rank score normalisation, normalised discretization, quantile discretization - references to all of these are on the website. MADAM - doi: 10.1186/1751-0473-5-3

## E.22 Clustering

### E.22.1 Clustering Evaluation

Cluster evaluation

Ref.: - <https://nlp.stanford.edu/IR-book/html/htmledition/contents-1.html> - <https://link.springer.com/article/10.1007/s40595-016-0086-9> - <http://datamining.rutgers.edu/publication/internalmeasures.pdf> - <https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf>

Goal: high intra-cluster similarity and low inter-cluster similarity

Criteria: - internal Internal validation is the other type clustering evaluation, where the evaluation of the clustering is compared only with the result itself, i.e., the structure of found clusters and their relations to each other. This is much more realistic and efficient in many real-world scenarios as it does not refer to any assumed references from outside which is not always feasible to

obtain. Particularly, with the huge increase of the data size and dimensionality as in recent applications with streaming data outputs, one can hardly claim that a complete knowledge of the ground truth is available or always valid. Internal clustering validation is based only on the intrinsic information of the data. Since we can only refer to the input dataset itself, internal validation needs assumptions about a “good” structure of found clusters which are normally given by reference result in external validation. Two main concepts, the compactness and the separation, are the most popular ones. Most other concepts are actually just combinations of variations of these two

- Compactness

The Compactness measures how closely data points are grouped in a cluster. Gro

- Separation

The Separation measures how different the found clusters are from each other. U

- external

The external validation, which compares the clustering result to a reference result.

1. Purity

Each cluster is assigned to the class which is most frequent in the cluster, and  
Bad clusterings have purity values close to 0, a perfect clustering has a purity of 1.  
High purity is easy to achieve when the number of clusters is large - in particular  
A measure that allows us to make this tradeoff is NMI.

2. Normalized mutual information (NMI)

NMI is always a number between 0 and 1.

3. Rand index

We want to assign two documents to the same cluster if and only if they are similar.  
The Rand index gives equal weight to false positives and false negatives. Separation

4. F measure

We can use the F measure measuresperf to penalize false negatives more strongly.

# Appendix F

## Systems Biology

### F.1 Systems Thinking

#### F.1.1 MODELS

ref: <<https://r4ds.had.co.nz/model-intro.html>>

The goal of a model is to provide a simple low-dimensional summary of a dataset. Ideally, the model will capture true “signals” (i.e. patterns generated by the phenomenon of interest), and ignore “noise” (i.e. random variation that you’re not interested in).

Types of models:

- Predictive models: supervised, generate predictions.
- Data discovery models: Unsupervised, these models don't make predictions, but instead help
- model basics:
- how models work mechanistically, focussing on the important family of linear models.
- general tools for gaining insight into what a predictive model tells you about your data, focus

model building:

- how to use models to pull out known patterns in real data. Once you have recognised an important

many models:

- how to use many simple models to help understand complex datasets. This is a powerful

Hypothesis generation vs. hypothesis confirmation

Traditionally, the focus of modelling is on inference, or for confirming that an hypoth

1. Each observation can either be used for exploration or confirmation, not both.
2. You can use an observation as many times as you like for exploration, but you ca

This is necessary because to confirm a hypothesis you must use data independent of the

If you are serious about doing an confirmatory analysis, one approach is to split your

1. 60% of your data goes into a training (or exploration) set. You're allowed to d
2. 20% goes into a query set. You can use this data to compare models or visualisat
3. 20% is held back for a test set. You can only use this data ONCE, to test your

This partitioning allows you to explore the training data, occasionally generating can

Note that even when doing confirmatory modelling, you will still need to do EDA. If you

## MODEL BASICS

There are 2 parts to a model:

1. Define a family of models that express a precise, but generic, pattern that you

Exemple:

- straight line:  $y = a_1 \cdot x + a_2$
- quadratic curve:  $y = a_1 \cdot x^2 + a_2$

2. Generate a fitted model by finding the model from the family that is the closes



- This takes the generic model family and makes it specific:
    - straight line:  $y = 3x + 7$
    - quadratic curve:  $y = 9x^2$
  - IMPORTANT: a fitted model is just the closest model from a family of models.
    - That implies that you have the "best" model (according to some criteria)
    - it doesn't imply that you have a good model
    - it certainly doesn't imply that the model is "true"
  - The goal of a model is not to uncover truth, but to discover a simple approximation that
  - George Box:
    - "All models are wrong, but some are useful."
    - "Now it would be very remarkable if any system existing in the real world could be
- For such a model there is no need to ask the question "Is the model true?". If "truth"

#### Predictions

- The predictions tell you the pattern that the model has captured, and the residuals tell you

#### Residuals

- It's also useful to see what the model doesn't capture, the so-called residuals which are
- Plot x vs residuals: If this looks like random noise, then suggests that our model has
- Continuous variables
- Categorical Variables
  - Generating a function from a formula is straight forward when the predictor is continuous
  - Effectively, a model with a categorical x will predict the mean value for each category

- You can't make predictions about levels that you didn't observe. Sometimes you

#### - Interactions

continuous and categorical

- What happens when you combine a continuous and a categorical variable?

There are two possible models you could fit to this data:

```
mod1 <- lm(y ~ x1 + x2, data = sim3)
```

```
mod2 <- lm(y ~ x1 * x2, data = sim3)
```

When you add variables with +, the model will estimate each effect independently.

It's possible to fit the so-called interaction by using \*. For example:

```
y ~ x1 * x2
```

is translated to

$$y = a_0 + a_1 * x1 + a_2 * x2 + a_{12} * x1 * x2.$$

- Note that whenever you use \*, both the interaction and the individual effects are estimated.
- Note that the model that uses + has the same slope for each line, but different intercepts.

#### Two Continuous

#### Transformations

You can also perform transformations inside the model formula. For example:

- Transformations are useful because you can use them to approximate non-linear relationships.
- However there's one major problem with using poly(): outside the range of the data, the model will extrapolate.

#### Other model families

This chapter has focussed exclusively on the class of linear models, which assume a linear relationship between the response and the predictors.

Generalised linear models, e.g. stats::glm(). Linear models assume that the response is a continuous variable.

Generalised additive models, e.g. `mgcv::gam()`, extend generalised linear models to incorporate non-linear effects.

Penalised linear models, e.g. `glmnet::glmnet()`, add a penalty term to the distance that penalises model complexity.

Robust linear models, e.g. `MASS::rlm()`, tweak the distance to downweight points that are very far from the line.

Trees, e.g. `rpart::rpart()`, attack the problem in a completely different way than linear models.

These models all work similarly from a programming perspective. Once you've mastered linear models, you can master these others too.

## MODEL BUILDING

We will take advantage of the fact that you can think about a model partitioning your data into parts. For very large and complex datasets this will be a lot of work. There are certainly alternative approaches. It's a challenge to know when to stop. You need to figure out when your model is good enough, and when it's not. "A long time ago in art class, my teacher told me "An artist needs to know when a piece is done."

-- Broseidon241, <<https://www.reddit.com/r/datascience/comments/4irajq>>

## MORE:

- Statistical Modeling: A Fresh Approach by Danny Kaplan, <<http://www.mosaic-web.org/go/StatisticalModeling>>
- An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, <<http://www.stat.columbia.edu/gareth/james/>>
- Applied Predictive Modeling by Max Kuhn and Kjell Johnson, <<http://appliedpredictivemodeling.com/>>

## MANY MODELS

three powerful ideas that help you to work with large numbers of models with ease:

1. Using many simple models to better understand complex datasets.
2. Using list-columns to store arbitrary data structures in a data frame. For example, this works for storing model coefficients.
3. Using the broom package, by David Robinson, to turn models into tidy data. This is a powerful idea.

Watch:

<<https://www.youtube.com/watch?v=jbkSRLYSojo>>

## F.1.2 General Systems Theory

### Chapter 1

- Systems Everywhere

‘In one way or another, we are forced to deal with complexities, with “wholes” or “systems”, in all fields of knowledge. This implies a basic re-orientation in scientific thinking.’

- On the History of Systems Theory
- Trends in Systems Theory

Applications of the analytical procedure depends on two conditions. The first is that interactions between “parts” be non-existent or weak enough to be neglected for certain research purposes. Only under this condition, can the parts be “worked out”, actually, logically, and mathematically, and then be “put together”. The second conditions is that the relations describing the behavior of parts be linear; only then is the condition of summativity give, i.e., an equation describing the behavior of the total is of the same form as the equations describing the behavior of the parts; partial processes can be superimposed to obtain the total process, etc.

These conditions are not fulfilled in the entities called systems, i.e., consisting of parts “in interaction”. The prototype of their description is a set of simultaneous differential equations, which are nonlinear in the general case. A system or “organized complexity” may be circumscribed by the existence of “strong interactions” or interactions which are “nontrivial”, i.e., nonlinear. The methodological problem of systems theory, therefore, is to provide for problems which, compared with the analytical-summative ones of classical science, are of a more general nature.

There are various approaches to deal with such problems. We intentionally use the somewhat loose expression “approaches” because they are logically inhomogeneous, represent different conceptual models, mathematical techniques, general points of view, etc.; they are, however, in accord in being “systems theories”:

- "Classical" system theory: applies classical mathematics, i.e., calculus. Its aim is
- Computerization and simulation: Sets of simultaneous differential equations as a way
- Compartment theory: the systems consists of subunits with certain boundary conditions
- Set theory: The general formal properties of systems, closed and open systems, etc.,

- Graph theory: Many systems problems concern structural or topologic properties of systems, rather than their dynamic behavior.
- Net theory: is connected with set, graph, compartment, etc. theories and is applied to such systems.
- Cybernetics: is a theory of control systems based on communication (transfer of information) between systems.
- Information theory: is based on the concept of information, defined by an expression isomorphic to the entropy of a system.
- Theory of automata: is the theory of abstract automata, with input, output, possibly trial-and-error learning.
- Game theory: it is concerned with the behavior of supposedly "rational" players to obtain maximum benefit.
- Decision theory:
- Queuing theory: concerns optimization of arrangements under conditions of crowding.

---

## Chapter 2 - The Meaning of General Systems Theory

- The Quest for a General System Theory

Surveying the evolution of modern science, we encounter a surprising phenomenon: independently of each other, similar problems and conceptions have evolved in widely different fields.

It was the aim of classical physics eventually to resolve natural phenomena into a play of elementary units governed by "blind" laws of nature. This was expressed in the ideal of the Laplacean spirit which, from the position and momentum of particles, can predict the state of the universe at any point in time.

We can ask for principles applying to systems in general, irrespective of whether they are of physical, biological or sociological nature. If we pose this question and conveniently define the concept of system, we find that models, principles, and laws exist which apply to generalized systems irrespective of their particular kind, elements and the "forces" involved.

A consequence of the existence of general system properties is the appearance of structural similarities or isomorphisms in different fields. There are correspondences in the principles that govern the behavior of entities that are, intrinsically, widely different.

- Aims of General System Theory
- Closed and Open Systems: Limitations of Conventional Physics

The principle of equifinality.

Living systems, maintaining themselves in a steady state, can avoid the increase of entropy, and may even develop towards states of increased order and organization.

- Information and Entropy

In many cases, the flow of information corresponds to a flow of energy. However, examples can easily be given where the flow of information is opposite to the flow of energy, or where information is transmitted without a flow of energy or matter. So information, in general, cannot be expressed in terms of energy. There is, however, another way to measure information, namely, in terms of decisions.

Entropy, as we have already defined, is a measure of disorder; hence negative entropy or information is a measure of order or of organization since the latter, compared to distribution at random, is an improbable state.

A great variety of systems in technology and in living nature follow the feedback scheme, and it is well-known that a new discipline, called Cybernetics, was introduced by Norbert Wiener to deal with these phenomena. The theory tries to show that mechanisms of a feedback nature are the base of teleological or purposeful behavior in man-made machines as well as in living organisms, and in social systems.

It can be shown that the primary regulations in organic systems, i.e., those which are most fundamental and primitive in embryonic development as well as in evolution, are of the nature of dynamic interaction. They are based upon the fact that the living organism is an open system, maintaining itself in, or approaching a steady state. Superposed are those regulations which we may call secondary, and which are controlled by fixed arrangements, specially of the feedback type. This state of affairs is a consequence of a general principle of organization which may be called progressive mechanization. At first, systems-biological, neurological, psychological or social - are governed by dynamic interaction of their components; later on, fixed arrangements and conditions of constraint are established which render the system and its parts more efficient, but also gradually diminish and eventually abolish its equipotentiality. Thus, dynamics is the broader aspect, since we can always arrive from general system laws to machine like function by introducing suitable conditions of constraint, but the opposite is not possible.

- Causality and Teleology
- What is Organization?

Characteristics of organization, whether of a living organism or a society, are notions like those of wholeness, growth, differentiation, hierarchical order, dominance, control, competition, etc.

- General System Theory and the Unity of Science

The total of observable events, shows structural uniformities, manifesting themselves by isomorphic traces of order in the different levels or realms.

We cannot reduce the biological, behavioral, and social levels to the lowest level, that of the constructs and laws of physics. We can, however, find constructs and possibly laws within the individual levels.

The unifying principle is that we find organization at all levels.

---

### Chapter 3 - Some System Concepts in Elementary Mathematical Consideration

- The System Concept

In dealing with complexes of “elements”, three different kinds of distinction may be made:

1. according to their number;
2. according to their species;
3. according to the relations of elements;

Summative characteristics of an element are those which are the same within and outside the complex; they may therefore be obtained by means of summation of characteristics and behavior of elements as known in isolation.

Constitutive characteristics are those which are dependent on the specific relations within the complex; for understanding such characteristics we therefore must know not only the parts, but also the relations.

If we know the total of parts contained in a system and the relations between them, the behavior of the system may be derived from the behavior of the parts.

A system can be defined as a complex of interacting elements. Interaction means that elements,  $p$ , stand in relations,  $R$ , so that behavior of an element  $p$  in  $R$  is different from its behavior in another relation,  $R'$ . If the behaviours in  $R$  and  $R'$  are not different, there is no interaction, and the elements behave independently with respect to the relations  $R$  and  $R'$ .

If we are speaking of “systems”, we mean “whole” or “unities”. Then it seems paradoxical that, with respect to a whole, the concept of competition between its parts is introduced. In fact, however, these apparently contradictory statements both belong to the essentials of systems. Every whole is based upon the competition of its elements, and presupposes the “struggle between parts” (Roux).

## F.2 Systems Biology

### Systems Biology Overview

#### doi: 10.4137/BBI.S12467 ####

In saying that we understand a biological process, we usually mean that we are able to predict future events and manipulate the process into a desired direction. Thus, biological inquiry could be viewed as an attempt to understand how a biological system transits from one state to another. In attempting to understand these transitions, a simple and frequently used approach is to compare two states of a system.

A central challenge in posed by omics data is how to navigate through the haystack of measurements (eg, differential expression between two states) to identify the needles comprised of the critical causal factors.

Network analysis is a powerful and general approach to this problem, in which the biological system is modeled as a network whose nodes represent dynamical units (eg, genes, proteins, metabolites, etc) and edges stand for links between them.

Multiple groups have been successfully using such methods to gain a systems-level understanding of biological processes and to reveal mechanisms of different diseases:

Amit I, Garber M, Chevrier N, et al. Unbiased reconstruction of a mammalian transcript

Sumazin P, Yang X, Chiu HS, et al. An extensive microRNA-mediated network of RNA-RNA in

Yang D, Sun Y, Hu L, et al. Integrated analyses identify a master microRNA regulatory n

Several recent discoveries ranging from genes that drive progression of different cancers:

Mine KL, Shulzhenko N, Yambartsev A, et al. Gene network reconstruction reveals cell cy

Chen JC, Alvarez MJ, Talos F, et al. Identification of causal genetic drivers of human

to microbes and microbial genes that cause a human illness:

Morgun A, Dzutsev A, Dong X, et al. Uncovering effects of antibiotics on the host and m

became possible because of the predictive power of network analysis. In particular, such insights would be very difficult to achieve if analysis is limited to finding differentially expressed genes and follow-up data mining of those genes.

Types of Network Analysis:



- Covariation network analysis
- semantic networks
- molecular interaction networks

Types of omics measurements that are amenable to network analysis:

- microarrays
- next generation sequencing (for genotyping, transcriptome profiling, or microbiome analysis)
- mass spectrometry-based proteomics
- metabolomics

Databases:

- Gene Expression Omnibus (GEO) and Array Express (for transcriptomics and epigenomics datasets),
- PRIDE<sup>47</sup> (for proteomics datasets)
- the Human Metabolome Database (for metabolomics datasets)
- lipid MAPS<sup>49</sup> (for lipidomics datasets)
- molecular interaction data from the BioGRID<sup>50</sup> or BioCyc databases can be used as a prior for edge

Network analysis consists of two fundamental stages: network reconstruction and network interrogation.

Integrative Networks

Integrative networks are becoming more popular under the premise that the resulting networks more

Tran LM, Zhang B, Zhang Z, et al. Inferring causal genomic alterations in breast cancer using

Chen Y, Zhu J, Lum PY, et al. Variations in DNA elucidate molecular networks that cause disease

Each type of omics measurement technology has a specific procedure for transforming the raw data

Olson NE. The microarray data analysis process: from raw data to biological significance. Ne

Anders S, McCarthy DJ, Chen Y, et al. Count-based differential expression analysis of RNA se

Hamady M, Knight R. Microbial community profiling for human microbiome projects: 1

Hernandez P, Müller M, Appel RD. Automated protein identification by tandem mass s

## 1. Network Reconstruction

the data-driven discovery or inference of the entities/nodes (transcripts, pro-teins, g

### - Normalization (pre-processing)

+ log-transformation in order to stabilize variances when measurements span orders

+ normalization to correct for sample-to-sample variation in the overall distribut.

#### - Microarray:

median normalization

quantile normalization

LOWESS normalization (Berger JA, Hautaniemi S, Jarvinen AK, Edgren H, Mitra

#### - RNA-Seq:

reads per kilobase per million mapped reads (RPKM) (Mortazavi A, Williams M

trimmed mean of M-values (TMM) (Robinson MD, Oshlack A. A scaling normaliz

Lim WK, Wang K, Lefebvre C, Califano A. Comparative analysis of microarray normal

Dillies MA, Rau A, Aubert J, et al; French StatOmique Consortium. A comprehensive

### - Discovery of Differentially expressed genes (selecting nodes)

identification of the relevant subset of variables/genes that will constitute the n

Statistical Tests:

- Welch's t-test

- moderated t-test

- permutation tests

--\> For parametric tests, accurate estimation of intra-sample-group variance s

+ locally pooled error

Jain N, Thatte J, Braciale T, Ley K, O'Connell M, Lee JK. Local-pooled-error test

+ empirical Bayes methods

Smyth GK, Michaud J, Scott HS. Use of within-array replicate spots for assessing experiments. *Bioinformatics*. 2005;21(9):2067--75.

Pan W. A comparative review of statistical methods for discovering differentially expressed

--\> Because omics data analysis typically involves tens of thousands of statistical tests, t

Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments.

- Correlation analysis for network reconstruction (finding links between nodes)

The central biological principles underlying correlation network analysis are:

1) that DEGs reflect functional changes

2) that DEGs do not work individually but interact (eg, at the protein or pathway level)

The central mathematical/statistical principle that allows us to use correlation networks for

Pearl J. *Causality: Models, Reasoning and Inference*. Vol 29. Cambridge University Press,

To reconstruct the network, the Pearson or Spearman correlation coefficient can be used to ob

correlations should be calculated within a group of samples that belong to one class/biologic

- Discriminating between direct and indirect links

Covariation gene networks in general consist of connections that result from a combination of

Pearl J. Direct and indirect effects. Paper presented at: Proceedings of the Seventeenth

For this reason, correlation networks in general have many edges that reflect indirect relati

Mathematically, direct effects can be defined as the association between two genes, holding t

Pearl J. An introduction to causal inference. *Int J Biostat.* 2010;6(2):7

methods to discriminate between direct and indirect links in covariation networks:

- partial correlation coefficient:

De La Fuente A, Bing N, Hoeschele I, Mendes P. Discovery of meaningful associations

Marbach D, Costello JC, Küffner R, et al; DREAM5 Consortium. Wisdom of crowds

- local partial correlation:

Thomas LD, Fossaluza V, Yambartsev A. Building complex networks through causal

- others:

Margolin AA, Nemenman I, Basso K, et al. ARACNE: an algorithm for the reconstruction

Jang IS, Margolin A, Califano A. hARACNe: improving the accuracy of regulatory

Barzel B, Barabási A-L. Network link prediction by global silencing of indirect

Feizi S, Marbach D, Médard M, Kellis M. Network deconvolution as a general

- Proportion of unexpected correlations (improvement of reconstruction and error evaluation)

A fundamental problem of the standard correlation network approach is that practical

- + proportion of unexpected correlations (PUC):

allows identifying and removing approximately half of false positive edges

Yambartsev A, Perlin M, Kovchegov Y, Shulzhenko N, Mine KL, Morgun A. Unexpected

The method takes into account a relation between the direction of regulatory

- Meta-analysis (improvement of reconstruction and error evaluation)

A first step for meta-analysis we apply two filters:

- 1) the same sign of statistic (mean, covariance, or correlation) throughout all

- 2) P-value thresholds across all datasets

These filters provide consistency and control for heterogeneity across datasets for a given gene. After calculating Fisher's P-values for all genes, the standard FDR procedure can be used to select genes. Several other approaches have been proposed for meta-analysis of gene expression data:

Rhodes DR, Yu J, Shanker K, et al. Large-scale meta-analysis of cancer microarray data identifies common and disease-specific gene expression patterns. *Proc Natl Acad Sci USA* 2006;103(12):13602-13607.

Hwang D, Rust AG, Ramsey S, et al. A data integration methodology for systems biology. *Proc Natl Acad Sci USA* 2006;103(12):13608-13613.

#### - Differentially coexpressed gene pairs (evaluating network changes)

The networks discussed above model static correlations between genes that change their expression levels. However, the sets of edges within a gene covariation network can themselves vary from state to state.

Kostka D, Spang R. Finding disease specific alterations in the co-expression of genes. *BMC Bioinformatics* 2006;7:1-12.

It has been shown that differentially coexpressed gene pairs frequently play critical roles in biological processes. In order to search for differentially coexpressed gene pairs:

#### - differentially associated pairs (DAPs):

Skinner J, Kotliarov Y, Varma S, et al. Construct and compare gene coexpression networks. *Proc Natl Acad Sci USA* 2006;103(12):13614-13619.

#### - Other methods:

Watson M. CoXpress: differential co-expression in gene expression data. *BMC Bioinformatics* 2006;7:1-12.

#### - Integrating heterogeneous omics data types: inter-omics networks

An inter-omics network is a bipartite network in which each edge connects two nodes of different omics data types. Approaches for omics data integration generally fall into one of two modalities:

+ first (and most prevalent) is integrating different types of data generated for a given gene.

+ The other type of integration makes an edge/link between two nodes from different omics data types.

There are two different approaches to infer such interomics links/edges:

- + The first one is based on bringing into reconstruction an experimental result
- + The second approach, which infers edges between different omics, establishes

Thus, the entire reconstruction procedure consists of inference on networks of each

Le Cao KA, Gonzalez I, Dejean S. integrOmics: an R package to unravel rela- ti-

Genome-wide measurements of epi-genetic marks and transcriptome data can be combin-

elucidate mechanisms of gene regulation:

Ramsey SA, Knijnenburg TA, Kennedy KA, et al. Genome-wide histone acetylation

Shilatifard A. Chromatin modifications by methylation and ubiquitination: imp-

Ramsey, Stephen A, et al. Epigenome-guided analysis of the transcriptome of p-

Integration of gene copy number data (chromosomal aberrations) and gene expression

Mine KL, Shulzhenko N, Yambartsev A, et al. Gene network reconstruction reveal-

Akavia UD, Litvin O, Kim J, et al. An integrated approach to uncover drivers of

integration of metage-nomics data from gut microbiota with intestinal gene expres-

Morgun A, Dzutsev A, Dong X, et al. Uncovering effects of antibiotics on the

## 2. Network interrogation

-> Systematic network analysis to gain maximal insights from a biological network that has been reconstructed

- Revealing potential mechanisms of a biological process or disease

- + quais são as vias envolvidas no processo?

- + quais são os key nodes de tais vias?
- + quais são as interações entre as vias identificadas. incluindo os nodos nas redes responsav
- Which functional pathways are involved?
  - \> A principal vantagem de se fazer uma module network analysis, ao invés de simplesmente a
  - + Quais são as subredes mais densas (módulos/clusters)?
- Enrichment analysis with external data
  - \> Performed by using literature-curated, gene-centric biological knowledge bases that con
  - \> A gene can be enriched for a particular biochemical pathway, a location in a genome, or
- Key regulators of pathways/modules
 

There are 2 major complementary approaches for finding key master regulators in covariation n

  1. Using network topology properties
  2. incorporating additional data into networks that provides information about causes of

Topological Properties:

  - degree and centrality measures:
    - + betweenness centrality
    - + closeness centrality
    - + eigenvector centrality
  - \> Nodes with high betweenness centrality (the so-called bottlenecks) have been shown t
- Integrating additional information in order to find causes of regulation
  - \> By overlaying such information on a coexpression network, one can establish the directio

types of biological information:

  - + genetic variants(aberrations, mutations, gene polymorphisms, etc)

Mine KL, Shulzhenko N, Yambartsev A, et al. Gene network reconstruction reveals

Tran LM, Zhang B, Zhang Z, et al. Inferring causal genomic alterations in

Akavia UD, Litvin O, Kim J, et al. An integrated approach to uncover drivers

+ epigenetic modifications

+ transcription factors

Ramsey SA, Knijnenburg TA, Kennedy KA, et al. Genome-wide histone acetylation

Carro MS, Lim WK, Alvarez MJ, et al. The transcriptional network for mesenchymal

+ other types of gene expression regulation (e.g., miRNA)

Sumazin P, Yang X, Chiu HS, et al. An extensive microRNA-mediated network

Lewis BP, Shih I-H, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human microRNA

Yang D, Sun Y, Hu L, et al. Integrated analyses identify a master microRNA

+ binding sites (or computationally predicted binding sites) of transcription

- How the pathways interact

Dado que as redes representam modelos de mudanças globais nos sistemas biológicos,

- Revealing function of individual node in the network

network biology offers a novel way to infer functions for genes whose functions have

There are 2 major approaches that implement guilt by association for prediction of

+ direct approach:

- neighbor counting; graphic algorithm; probabilistic methods;

- they all assign a function to a node based on the functions of the neighbors

+ modular approach:

- guide the assignment of a function to a gene by the collective function of



- Network cross-species conservation

Assessment of evidence for network function

subgraphs of the novel network (and in some approaches, constituent protein sequences) are used

Alternatively, gene coexpression networks from two species can be compared in their entirety,

- What is the number of nodes needed to be perturbed in order to achieve a transition from one state to another?

Liu YY, Slotine JJ, Barabasi AL. Controllability of complex networks. *Nature*. 2011;473(7346):167--171.

- some mathematical properties observed in biological networks such as small world, scale-free, and modular

102. Piraveenan M, Prokopenko M, Zomaya A. Assortative mixing in directed biological networks. *PLoS ONE*. 2012;7(12):e44111.

103. Newman ME. The structure and function of complex networks. *SIAM Rev*. 2003;45(2):167--256.

### F.2.1 Hematopoiesis and its disorders\_a systems biology approach

## F.3 Systems Vaccinology

### F.3.1 Systems Vaccinology - Pulendran, Nakaya - 2010

Understanding the immunological mechanisms of vaccination is of paramount importance in the rational design of future vaccines against pandemics such as HIV, malaria, and tuberculosis and against emerging infections.

The innate immune sensing system is through the pattern recognition receptors (PRRs):

- Toll-like receptors (TLRs)
- C type lectin-like receptors
- nucleotide-binding oligomerization domain-like receptors
- retinoic acid-inducible gene I (RIG-I)-like receptors

Emerging evidence suggests that the nature of the DC subtype, as well as the particular PRR triggered, plays a critical role in modulating the strength, quality, and persistence of adaptive immune responses.

Systems Biology:

- The grand challenge of systems biology is to understand the biological complexity that arises from the interactions between all the parts in a biological system, with the goal of understanding the nature of biological networks.
- Describes the complex interactions between all the parts in a biological system, with the goal of understanding the nature of biological networks.
- A key goal is to understand the nature of biological networks.

Biological networks:

- access, integrate, and communicate information from the genome to the environment and the environment back to the genome.
- These networks represent, in a sense, the lowest functional units of life processes, and understanding these life processes requires understanding the nature and behavior of these networks.
- Understanding these life processes requires understanding the nature and behavior of these networks.

## F.4 Systems Immunology

## F.5 Artificial Biology

### F.5.1 Principles of Genetic Circuit Design

Principles of Genetic Circuit Design

#### ——— Need to know ——— ####

- impact of an amino acid substitution on protein thermostability (Protein folding and stability)
- distribution of flux through modified metabolic networks (Constraining the metabolic network)
- oscillators;
- PID (proportional integral derivative) controller

#### ————— ####

The production of bio-based chemicals can be improved by:

- timing gene expression at different stages of fermentation;
- turning on an enzyme only under particular conditions (e.g, high cell density);

Synthetic regulation is also important to discover natural products like pharmaceuticals.

-\> 'Silent' gene clusters: the conditions under which they are induced are unknown;

living cells could be programmed to serve as:

- therapeutic agents that correct genetic disease;
- colonize niches in the human microbiome to perform a therapeutic function;

'smart' plants that sense and adapt to environmental challenges

bacteria that organize to weave functional materials with nanoscale features

1. circuits require the precise balancing of their component regulators to generate the proper re-
2. many circuits are difficult to screen in directed-evolution experiments for correct perform-
3. there are few tools to measure circuit performance. Typically, a fluorescent reporter is used
4. synthetic circuits are very sensitive to environment, growth condi- tions and genetic context

- Transcriptional vs Post-transcriptional circuits

Transcriptional circuits maintain a common signal carrier, which simplifies the connection of cir-

Transcriptional circuits function by changing the flow of RNA polymerase (RNAP) on DNA.

Post-transcriptional circuits, including those based on protein and RNA interactions, are covered

Regulator Classes

- DNA-binding proteins

DNA-binding proteins can recruit or block RNAP to increase or decrease the flux, respectively

Many families of proteins can bind to specific DNA sequences (operators).

The simplest way to use these proteins as regulators is to design promoters with operators th

Such repressors have been built out of:

- zinc-finger proteins;
- transcription activator-like effectors;
- TetR homologs;

- phage repressors;
- LacI homologs;

Expanding protein libraries can be challenging because each repressor has to be ori

There are also several challenges in using DNA-binding proteins to build circuits.

The circuits can also be very dependent on growth rate because differences in the c

the response functions are often suboptimal and difficult to control because they l

#### - Recombinases

RNAP flux can also be altered with invertases that change the orientation of promo

Recombinases are proteins that can facilitate the inversion of DNA segments between

Site specific recombinases often mediate 'cut-and-paste' recombination, during whi

Two types of recombinases have been used to build genetic circuits:

1. tyrosine recombinases (such as Cre, Flp and FimBE);

This type require host-specific factors. These recombinases can be reversibl

2. serine integrases;

Catalyze unidirectional reactions that rely on double-strand breaks to inv

These proteins are ideal for memory storage because they flip DNA permanently, and

using recombinases can be challenging because their reactions are slow (requiring 2

Reversible recombinases can also generate mixed populations; however, this limitat

#### - CRISPRi

The CRISPRi system uses the Cas9 protein to bind to the DNA and alter transcrip

#### - Adapted RNA-IN/OUT

Selecting parts to tune the circuit response

Common failures modes from connecting circuits

Interactions between synthetic circuits and the host organism



# Appendix G

## Resources

### G.1 The Human Cell Atlas

#### Introduction

Between 1838 and 1855, Schleiden, Schwann, Remak, Virchow and others crystallized an elegant Cell Theory:

1. all organisms are composed of one or more cells;
2. that cells are the basic unit of structure and function in life;
3. and that all cells are derived from pre-existing cells;

Human physiology emerges from normal cellular functions and interactions. Human disease entails the disruption of these processes and may involve aberrant cell types and states. Genetic variants that contribute to disease typically manifest their action through impact in a particular cell type.

At a conceptual level, one challenge is that we lack a rigorous definition of what we mean by the intuitive terms “cell type” and “cell state.” Cell type often implies a notion of persistence (e.g., being a hepatic stellate cell or a cerebellar Purkinje cell), while cell state often refers to more transient properties (e.g., being in the G1 phase of the cell cycle or experiencing nutrient deprivation). But, the boundaries between these concepts can be blurred, because cells change over time in ways that are far from fully understood.

What is the Human Atlas?

At its most basic level, the Human Cell Atlas must include a comprehensive reference catalog of all human cells based on their stable properties and transient features,

as well as their locations and abundances. Yet, an atlas is more than just a catalog: it is a

map that aims to show the relationships among its elements. By doing so, it can

sometimes reveal fundamental processes—akin to how the atlas of Earth suggested

continental drift through the correspondence of coastlines.

To be useful, an atlas must also be an abstraction—comprehensively representing

certain features, while ignoring others.

A natural solution would be to describe each human

cell by a defined set of molecular markers. For example, one might describe each cell by

the expression level of each of the ~20,000 human protein-coding genes—that is, each cell

would be represented as a point in ~20,000-dimensional space. Of course, the set of

markers could be expanded to include the expression levels of non-coding genes, the

levels of the alternatively spliced forms of each transcript, the chromatin state of every

promoter and enhancer, and the levels of each protein or each post-translationally

modified form of each protein.

The imaginary Ultimate Human Cell Atlas would represent all conceivable markers in:

1. every cell in a person's body
2. every cell's spatial position (by adding three dimensions for the body axes)
3. every cell at every moment of a person's lifetime (by adding another dimension for time relating the cells by a lineage)
4. the superimposition of such atlases from every human being, annotated according to differences in health, genotype, lifestyle and environmental exposure



A cell “type” might be defined as a region or a probability distribution — either in the full-dimensional space or in a projection onto a lower-dimensional space that reflects salient features.

Cell types follow a highly stereotyped spatial patterns.

Histology: Cell neighborhood and position

Histology examines the spatial position of cells and molecules within tissues.

Development: transitions to differentiated cell types

Cells arrive at their final differentiated cell types through partly asynchronous branching pathways of development, which are driven by and reflected in molecular

changes, especially gene-expression patterns

a cell may be following multiple dynamic paths simultaneously—for example, differentiation, the cell cycle, and pathogen response — that may affect each other.

Physiology and homeostasis: cycles, transient responses and plastic states

cells are constantly undergoing multiple dynamic processes of physiological changes, including cyclical processes, such as the cell cycle and circadian rhythms; transient responses to diverse factors, from nutrients and microbes to stress; and plastic states that can be stably maintained over longer time scales, but can change. The molecular phenotype of a cell reflects a superposition of these various processes.

Disease: Cells and cellular ecosystems

disease, which invariably involves disruption of normal cellular functions, interactions with the environment, and changes in gene expression.

Single-cell information across many patients will allow us to learn about how cell states and functions vary.

proportions and states vary and how this variation correlates with genome variants, environmental factors, and clinical data.

disease course and treatment response

## G.2 reproducible data analysis

[OK] <https://resources.rstudio.com/rstudio-conf-2019/a-guide-to-modern-reproducible-d>

=\> <https://ropensci.github.io/reproducibility-guide/>

<http://grunwaldlab.github.io/Reproducible-science-in-R/>

<https://github.com/grunwaldlab/Reproducible-science-in-R>

<https://rpubs.com/minebocek/user2017-ors>

[https://www.r-bloggers.com/preview-my-new-book-introduction-to-reproducible-science-in-](https://www.r-bloggers.com/preview-my-new-book-introduction-to-reproducible-science-in-r/)

<https://datacarpentry.org/rr-workshop/>

<https://rviews.rstudio.com/2018/01/18/package-management-for-reproducible-r-code/>

[https://cartesianfaith.files.wordpress.com/2018/11/rowe-introduction-to-reproducible-](https://cartesianfaith.files.wordpress.com/2018/11/rowe-introduction-to-reproducible-science-in-r/)

<https://www.reconlearn.org/post/reproducibility.html>

[http://www.geo.uzh.ch/microsite/reproducible\\_research/post/rr-rstudio-git/](http://www.geo.uzh.ch/microsite/reproducible_research/post/rr-rstudio-git/)

<https://swcarpentry.github.io/r-novice-gapminder/02-project-intro/>

[https://www.crcpress.com/Reproducible-Research-with-R-and-R-Studio/Gandrud/p/book/978](https://www.crcpress.com/Reproducible-Research-with-R-and-R-Studio/Gandrud/p/book/9781498714444)

<https://www.tandfonline.com/doi/abs/10.1080/00031305.2017.1375986?journalCode=utas20>

<https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989?src=recsys>

<https://github.com/ropenscilabs/checkers>

<https://github.com/benmarwick/onboarding-reproducible-compedia>

<https://github.com/research-compedium/ten-simple-rules>

<https://github.com/research-compedium/how-to-read-a-research-compedium>

<https://mybinder.org>

<https://ropenscilabs.github.io/drake-manual/>

`<https://peerj.com/preprints/3159/>`

`<https://happygitwithr.com/big-picture.html>`

`<https://ropenscilabs.github.io/drake-manual/>`

`<https://plotly-r.com/introduction.html>`

`https://ironholds.org/projects/r_shiny/index.html`

`https://bioinformatics.ca/workshops/2016-exploratory-analysis-biological-data-using-r/`

`https://r4ds.had.co.nz/exploratory-data-analysis.html`

`https://ropensci.github.io/reproducibility-guide/`

`https://www.earthdatascience.org/tags/reproducible-science-and-programming/RStudio/`

`http://ohi-science.org/data-science-training/`

`http://grunwaldlab.github.io/Reproducible-science-in-R`

`https://www.jove.com/video/50115/experimental-human-pneumococcal-carriage`

analysis report

rmarkdown

`<https://rmarkdown.rstudio.com>`

`<https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf>`

LaTeX

`<https://wch.github.io/latexsheet/>`

programming

R

`<https://www.rstudio.com/wp-content/uploads/2016/02/advancedR.pdf>`

exploratory data analysis

<<https://bookdown.org/rdpeng/exdata/preface.html>>

<<https://towardsdatascience.com/simple-fast-exploratory-data-analysis-in-r-with-dataexplor>>

<<https://blog.datascienceheroes.com/exploratory-data-analysis-in-r-intro/>>

<<https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scr>>

<<https://r4ds.had.co.nz/transform.html>>

<<https://www.youtube.com/watch?v=MCJD5iJjr7Y>>

<<http://www.sthda.com/english/wiki/print.php?id=202>>

<<http://www.sthda.com/english/wiki/factoextra-r-package-easy-multivariate-data-analysis>>

<[https://www.youtube.com/watch?v=aiqGyLGrQ48&list=PLnZgp6epRBbTsZEFXi\\_p6W48HhNyqwxIu&in](https://www.youtube.com/watch?v=aiqGyLGrQ48&list=PLnZgp6epRBbTsZEFXi_p6W48HhNyqwxIu&in)>

data science

<<https://github.com/rstudio/Intro/tree/master/slides>>

<<https://www.oreilly.com/library/view/introduction-to-data/9781491915028/video192702.htm>>

bioinformatics

<<https://bioinformatics.ca/workshops/workshops-2018/>>

<<https://www.ebi.ac.uk/training/online/course/bioinformatics-terrified-0>>

<<https://www.ebi.ac.uk/training/online/course/bioinformatics-sequence-analysis-web-ser>>

<<https://www.ebi.ac.uk/training/online/course/bringing-data-life-data-management-biome>>

<<https://www.ebi.ac.uk/training/online/course/challenges-and-opportunities-virtual-res>>

systems biology

<<https://www.ebi.ac.uk/training/events/2019/systems-biology-large-datasets-biological->>

metagenomics

<<https://www.ebi.ac.uk/training/online/course/ebi-metagenomics-analysing-and-exploring>>

<<https://www.ebi.ac.uk/training/online/course/analysing-and-visualising-microbiome-der>>

data analysis certificates

<[https://www.colorado.edu/earthlab/earth-data-analytics-foundations-professional-certificate?utm\\_](https://www.colorado.edu/earthlab/earth-data-analytics-foundations-professional-certificate?utm_)

TIMESERIES

<[https://subscription.packtpub.com/book/big\\_data\\_and\\_business\\_intelligence/9781786462411](https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781786462411)>

## G.3 Databases

Genes associated with a disease:

<<http://www.disgenet.org/>>

<<https://www.malacards.org/>>

Genes associated with a tissue type:



## Appendix H

# Statistics & Probability

### H.1 Statistics

#### H.1.0.1 Statistical Tests

##### Two Categorical Variables

Checking if two categorical variables are independent can be done with Chi-Squared test of independence.

This is a typical Chi-Square test: if we assume that two variables are independent, then the test statistic follows a Chi-Square distribution.

There also exists a Crammer's V that is a measure of correlation that follows from this test.

##### Categorical vs Numerical Variables

For this type we typically perform One-way ANOVA test: we calculate in-group variance and inter-group variance.

### H.2 Probability





# Appendix I

## Virology

- Viruses infect all living things
- We regularly eat and breath billions of virus particles
- We carry viral genomes as part of our own genetic material



## Appendix J

# Goals and Questions (Original)

### J.1 Goals

#### J.1.1 Main Goal

The goal of this project is to provide a systemic description of the interplay between sex differences and ageing in human pneumococcal carriage, as well how these and other factors affect the susceptibility to pneumococcal diseases.

#### J.1.2 Specific Goals

### J.2 Research Questions

#### J.2.1 Does inflammation in the nose change with sex?

#### J.2.2 Does inflammation in the nose change with age?

- Compare baseline values between elderly(AGES) and young (LAIV2)
  - Control/test for effect of sex → this is something I think we'll have to control for in most of our tests
  - Linear regression with Age as continuous variable
    - \* *It may be that we need to try this analysis in the AGES group alone...we don't have great sampling of middle aged adults in either cohort which may bias the analysis?*

### J.2.3 Can baseline values of inflammation in the nose predict who will become colonized in the future?

- AGES study we have Days 0,2,7,9
  - *Right now they determine carriage positive as at least one day with carriage +ve*
    - \* *I wonder if there is a way to look at patterns for people who carry for more than one day, etc*
  - We can find other data sets in the lab to test this with?

### J.2.4 Do any of the cytokines of immune cell readouts (from Flow data) correlate with Spn density?

### J.2.5 Do any of the cytokines correlate with flow cytometry data?

Primarily numbers of granulocytes, monocytes, lymphocytes, B cells, T cells, CD4 T cells, CD8 T cells and MAIT?

### J.2.6 Does the induction of cytokines differ after Spn inoculation with sex?

### J.2.7 Does the induction of cytokines differ after Spn inoculation with age?

- Using Day 0 to Day 2 changes ( $\log_2(\text{FC})$ )
  - *I wanted to recreate Fig1b from the Nature Immunology paper and was wondering if you could help me by even just sending me the code and I can have a go at it?*

### J.2.8 Is the response different in natural carriers?

- I don't think we're powered enough to do this as we only have 4 natural carriers in AGES and 1 natural carrier in LAIV2
- In the Rechallenge group
  - 16 people from AGES were rechallenged 36 months later and Days 0,2,7 were collected
    - \* An "R" was just added to the end of the sample ID

- There are 4 people that are carriage positive in the rechallenge
  - \* Can we see if there is a difference in nasal inflammation or immune cells (measured by flow) between people who were carriage negative vs those that were carriage positive?

## J.3 Specific Questions

### J.3.1 Pneumococcal Carriage Load (*Spn Density*)

1. Quem tem maior chance de desenvolver *carriage*?
2. Existem outros grupos de risco?
3. Quais sao as variaveis importante p/ predizer *carriage*?
4. Existem agrupamentos de individuos baseado em seu perfil de expressao?

### J.3.2 Gene Expression (*RNA-Seq*)

#### J.3.2.1 At Baseline

1. Which genes are perturbed between C+ and C- among man/woman?
2. Which pathways are enriched between C+ and C- among man/woman?
3. Which genes can predict *carriage* in men/women?

#### J.3.2.2 Day<sub>x</sub> vs Baseline

1. Which genes are perturbed among C+ man/women?
2. Which pathways are perturbed among C+ man/women?
3. Which genes are perturbed among C- man/women?
4. Which pathways are perturbed among C- man/women?

### J.3.3 Cytokines (*Luminex*)

#### J.3.3.1 At Baseline

1. Which cytokines are perturbed between C+ and C- among man/woman?

#### J.3.3.2 Day<sub>x</sub> vs Baseline

1. Which cytokines are perturbed among C+ man/woman?
2. Which cytokines are perturbed among C- man/woman?

### J.3.4 Hormone Levels

### J.3.5 Cell Recruitment (*Flow Cytometry*)

## J.4 Technical Questions

- *any idea how we should deal with the repeated values of cytokines that we have in our datasets?*

## J.5 Sex Effect on Spn Carriage

**Central Question:** Why are human more prone to developing pneumococcal carriage, but men more susceptible to pneumococcal disease?

## J.6 Ageing Effect on Spn Carriage

**Central Question:** Why are older adults more susceptible to pneumococcal disease?

- Change in carriage status with age    Hugh found a decrease in carriage acquisition with age
  - Test carriage rates with age
  - Test the sex difference in carriage rates in AGES
- Baseline differences in inflammatory status and immune cell composition in the nose
  - Are there changes in cytokines with age?
    - \* IL-2 and FGF-basic
    - \* Sex differences?
      - EGF & IL-8
    - \* *Do these remain in younger cohorts/correlate with any other cell types?*
- How does the response to Spn change with age?
  - Baseline predictors for carriage status:
    - \* Lower CD16 MFI on Granulocytes

- **Does CD16 MFI correlate with OPK data?**
  - \* Higher IP-10 and MIG in nose
  - \* **Compare to what was found in LAIV studies**
- Kinetic responses to Spn colonization in the nose:
  - \* How do nasal cytokines change depending on Carriage, Sex and Age?
  - \* How do immune cells in the nose change depending on Carriage, Sex and Age?
  - \* How does neutrophil degranulation change depending on Carriage, Sex and Age?
  - \* How do anti-CPS antibody levels in the nose change with Carriage, Sex and Age?
- Adaptive response to Spn colonization in circulation:
  - \* How does the serum antibody response (anti-CPS and anti-protein) change depending on Carriage, Sex and Age?
  - \* Does the opsonophagocytic function of antibodies change with age?