



**HAL**  
open science

# Statistical methods for leveraging high-dimensional data from high-throughput measurements in vaccine clinical development

Boris P. Hejblum

► **To cite this version:**

Boris P. Hejblum. Statistical methods for leveraging high-dimensional data from high-throughput measurements in vaccine clinical development. Methodology [stat.ME]. Université de Bordeaux, 2024. tel-04633105

**HAL Id: tel-04633105**

**<https://hal.science/tel-04633105>**

Submitted on 3 Jul 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# HABILITATION À DIRIGER DES RECHERCHES

École doctorale Sociétés, Politique, Santé Publique

Option Santé Publique – mention Biostatistique

**Boris HEJBLUM**

Chargé de Recherche à l'Inserm

*Statistical methods for leveraging high-dimensional  
data from high-throughput measurements  
in vaccine clinical development*

Soutenance publique le 7 mai 2024

## Membres du jury :

Anne-Laure BOULESTEIX	Professeure, Université Ludwig-Maximilians à Munich	Rapporteure
Julie JOSSE	Directrice de Recherche, Inria	Examinatrice
Pierre NEUVIAL	Directeur de Recherche, CNRS	Rapporteur
Franck PICARD	Directeur de Recherche, CNRS	Rapporteur
Rodolphe THIÉBAUT	Professeur, Université de Bordeaux	Président



# Contents

<b>Remerciements</b>	<b>7</b>
<b>1 Background</b>	<b>9</b>
1.1 The rise of vaccinomics . . . . .	9
1.2 Differential expression analysis of transcriptomic data . . . . .	10
1.2.1 Microarrays . . . . .	11
1.2.2 Bulk RNA-seq . . . . .	11
1.2.3 Single-cell RNA-seq . . . . .	12
1.3 Flow-cytometry data for cellular phenotyping . . . . .	12
1.4 Methodological challenges & research positioning . . . . .	15
<b>2 Methods for Differential Expression Analysis of RNA-seq data</b>	<b>19</b>
2.1 A working linear model for heteroscedastic gene expression . . . . .	19
2.2 Statistical tests that avoid distributional assumptions . . . . .	20
2.2.1 A versatile variance component score test . . . . .	21
2.2.2 Conditional Independence Testing for scRNA-seq . . . . .	22
2.3 Gene set approaches . . . . .	25
<b>3 Learning cellular population proportions</b>	<b>27</b>
3.1 Unsupervised clustering for cytometry data . . . . .	28
3.1.1 Clustering . . . . .	28
3.1.2 Bayesian mixtures . . . . .	28
3.1.3 A binary tree algorithm . . . . .	30
3.2 Supervised automated gating . . . . .	31
3.2.1 Semi-supervised automated gating . . . . .	32
3.2.2 Optimal transport for accelerated supervised gating . . . . .	32
3.3 Post-clustering inference . . . . .	33
3.4 Application to FCS data from the T-cell panel HIPC study . . . . .	36
<b>4 Connections with observational studies, EHR data and surrogate marker evaluation</b>	<b>43</b>
4.1 Electronic Health Records (EHR) . . . . .	43
4.1.1 Similarities between EHR and transcriptomic data . . . . .	43

4.1.2	Monitoring infections leveraging EHR from hospital data-warehouse . . . . .	44
4.1.3	Predicting COVID-19 hospitalization leveraging EHR . . . . .	45
4.2	Considering gene expression as a potential surrogate marker . . . . .	47
4.2.1	Evaluation of high-dimensional surrogate markers for a binary treatment . . . . .	48
4.2.2	Evaluation of gene expression as a surrogate marker for antibody response to Ebola infection in an observational study . . . . .	51
<b>5</b>	<b>Software dissemination and reproducible research in biostatistics</b>	<b>55</b>
5.1	Reproducible research . . . . .	55
5.1.1	Definitions . . . . .	55
5.1.2	Open science & biostatistics specificities . . . . .	57
5.2	Personal stories about reproducing other’s research . . . . .	59
5.2.1	Identifying unreported confounding batch effects . . . . .	59
5.2.2	Semi-synthetic data and confounding bias . . . . .	61
<b>6</b>	<b>Directions for future research</b>	<b>71</b>
6.1	Multi-modal data integration to enhance measurement resolution . . . . .	71
6.2	Incorporating prior biological knowledge to enhance results robustness and overcome limited sample sizes . . . . .	73
6.3	Leveraging multiple scales to enhance population generalization . . . . .	74
6.4	High-dimensional surrogate construction to enhance clinical relevance . . . . .	74
6.5	Computational efficiency to enhance numerical scalability . . . . .	76
	<b>Curriculum Vitæ</b>	<b>77</b>
	Research experience . . . . .	77
	Education . . . . .	77
	Teaching experience . . . . .	78
	Scientific supervision . . . . .	79
	Grants & funding . . . . .	80
	Patents . . . . .	81
	Software development & maintenance . . . . .	81
	Active international research collaborations . . . . .	83
	Outreach activities . . . . .	83
	Research visits abroad . . . . .	83
	Scientific evaluation . . . . .	84
	Reviewer . . . . .	84
	Academic responsibilities . . . . .	85
	Selected communications . . . . .	85
	List of scientific publications . . . . .	89

## CONTENTS

Published articles . . . . .	89
Books . . . . .	95
Preprints . . . . .	96
<b>Bibliography</b>	<b>97</b>
<b>Glossary &amp; acronyms</b>	<b>117</b>
Glossary . . . . .	117
Acronyms . . . . .	117

## CONTENTS

# Remerciements

Je remercie tout d’abord Anne-Laure Boulesteix, Pierre Neuvial et Franck Picard d’avoir accepté de rapporter ce mémoire. Je remercie également Julie Josse et Rodolphe Thiébaud d’avoir accepté de faire partie de mon jury.

Je remercie ensuite tous les étudiants que j’ai eu la chance d’encadrer, en particulier dans le cadre de leur doctorat : Soufiane, Marine, Paul, Benjamin, Thomas, Kalidou, Arthur, Annesh et bientôt Sara. Je remercie également mes collègues qui m’ont accompagné dans ces encadrements : Hélène Jacqmin-Gadda, Cécile Delcourt, Denis Agniel, Rodolphe Thiébaud, Jérémie Bigot et Xavier Hinaut. Les travaux que je synthétise et articule dans ce mémoire sont le fruit de nos collaborations.

Je remercie également mes mentors scientifiques, Daniel Commenges, à nouveau Rodolphe, François Caron, Tianxi Cai et Hélène Jacqmin-Gadda. C’est à votre contact que j’ai appris le métier de chercheur, et que je continue d’apprendre. Chacun de vous m’a fait grandir, tant humainement que scientifiquement.

I also want to thank my international collaborators, on top of Tianxi and Denis already mentioned above, and especially Damien Chaussabel, Lin Lin and Layla Parast. Beyond the pleasure of meeting familiar faces for sharing new ideas (and a good meal) at conferences and workshops around the world, you are the embodiment of the global and human endeavor that is scientific research.

Je remercie mes collègues de l’équipe SISTM – et en particulier mon directeur d’équipe, notre assistante administrative Sandrine Darmigny pour son aide inestimable, Mélanie Prague et Laura Richert qui coordonnent les 2 autres rouages de notre fameuse “roue” – ainsi que mes collègues de notre équipe soeur, l’équipe BIOSTAT, avec une mention toute particulière pour le bureau des *dudes surfeurs*. Je remercie aussi Cécile Proust-Lima pour ses conseils toujours avisés et Pierre Joly pour son écoute, ainsi que Cécilia Samieri pour ses questions stimulantes. Plus largement, je remercie l’ensemble de mes collègues du centre *Bordeaux Population Health*. Je mesure régulièrement ma chance de travailler dans un environnement aussi riche et si bienveillant.

Je tiens à remercier ici l’ensemble du Vaccine Research Institute, et en particulier Yves Lévy, Aurélie Wiedemann, Hakim Hocini et Véronique Godot. Nos discussions, parfois animées, sont toujours des moments d’échanges précieux.



## REMERCIEMENTS

Je remercie aussi les participants des essais cliniques pour la confiance et l'espoir qu'ils placent dans la recherche scientifique, et pour qui j'ai le plus grand respect.

Enfin je remercie ma famille pour leur amour et leur soutien au quotidien (et notamment durant la rédaction de ce mémoire).

# 1 Background

## *High-throughput data in vaccine clinical research*

### 1.1 The rise of vaccinomics

Since the first sequencing of a human genome at the turn of the XXI<sup>st</sup> century ([International Human Genome Sequencing Consortium, 2001](#)), there has been a surge in high-throughput data collection. Innovative technologies keep pushing the boundaries on our capacity to monitor biological processes, often generating high-throughput measurements. Such techniques include next-generation sequencing of the transcriptome (bulk RNA-seq) alongside single-cell techniques (thanks to microfluidics) such as flow-cytometry, mass cytometry and single-cell RNA-seq (scRNA-seq). The output data are frequently denoted “-omics”, such as proteomics (measurements of proteins), metabolomics (measurements of metabolites), genomics (measurements of the DNA, including methylation, SNPs or also copy numbers), transcriptomics (measurements of gene expression through RNA), etc. These omics’ data all have one feature in common: they are high-dimensional. High-dimension creates a methodological challenge in its own right for traditional statistics ([Giraud, 2021](#)), and has spurred numerous specific developments ([Bühlmann and van de Geer, 2011](#)).

Although industrial tooling typically reduces the cost of high-throughput measurements once they are initially released, they often remain expensive. This generally prevents their wide adoption in large trials or cohorts (with some notable exceptions, in particular for genomics). Yet, many biomedical studies are leveraging these high-throughput technologies to include one or several omics data collection ([Thiébaud et al., 2014](#)). In particular, clinical studies of the immune response, like vaccine trials, now routinely feature those in their early phases (either pre-clinical, phase I or phase II with their limited sample size). This gave rise to the field of vaccinomics, i.e. the modeling of the immune response through the generation and analysis of big, high-dimensional, complex data-sets, with the hope to improve both our understanding of the human immune system, and our capacity to predict vaccine responses ([Poland et al., 2008](#); [Poland and Oberg, 2010](#); [Poland et al., 2011](#); [Oberg et al., 2015](#)).

In this context, two high-throughput measurements have proven particularly valuable to deepen our understanding of the complex biological processes underpinning the immune response, and to generate new hypotheses on the underlying mechanisms that could drive and accelerate new vaccine candidate developments: i) genome-wide transcriptomics, ii) flow-cytometry. The first are high-dimensional omics data, with several thousands of gene expressions measured simultaneously. The second are big data, in the sense that one biological sample (for one individual at one time point) will feature hundred of thousands of cells characterized according to several dozens of surface and intra-cellular markers. These measurements are currently being used to characterize the complex dynamics of the immune responses to candidate vaccines (or following a natural infection). Their analysis warrants the use of dedicated statistical methods to accommodate the specifics of these data.

## 1.2 Differential expression analysis of transcriptomic data

Gene expression is a dynamic process at the root of the metabolic cascade. It is based on DNA transcription followed by translation. While there are many scientific questions that can be studied through the analysis of gene expression measurements, I focus on Differential Expression Analysis (DEA). DEA aims at identifying which genes are differentially expressed according to different experimental conditions. Most often this entails comparing gene-wise expression between two conditions, such as vaccinees against placebo or infected against healthy participants. In vaccine trials, we frequently need to go beyond this simple two-group comparison setting because of repeated measurements during the trial, important covariates requiring model adjustment (such as age or sex), and comparisons across multiple vaccine arms (with different vaccine candidates, doses and injection schedules).

Numerous technologies exist to study this mechanism, and the methods for the analysis of transcriptomic data are linked to the technology used for their generation. Over the past two decades, the two main technologies for measuring gene expression across the whole genome were microarray chips, and RNA-seq (Lowe et al., 2017). RNA-seq is more recent and more precise, and in many applications it has now replaced microarrays, but many data remain available from microarrays.

### 1.2.1 Microarrays

Microarray technology enables comprehensive genome-wide messenger Ribonucleic Acid (mRNA) measurements by utilizing microscopic spots on silicon (or glass) surfaces containing specific labeled sequences of DNA. These sequences are designed to hybridize and bind specifically to their complementary, leveraging the principle that a nucleic acid sequence would uniquely pair with its complement. This method allows for the profiling of tens of thousands of transcripts simultaneously (after reverse transcription of the extracted mRNA), offering insights into gene expression of the whole-genome.

Because microarrays eventually quantify RNA in a biological sample from which RNA was originally extracted by measuring a fluorescence, it yields continuous data as a measure of gene expression. This means that the two main statistical challenges for the analysis of microarray data are i) their normalization ii) their high-dimension. Indeed microarrays, as many other high-throughput technologies, are quite sensitive to external factors. Their measure can be heavily influenced by technical conditions. They thus require careful normalization procedures to ensure comparability of the measurements across samples (Shi et al., 2010), and ideally a careful randomization of the samples for their processing in order to avoid any confusion bias between biological or experimental factors and technical variations (also known as “batch effects”, see Leek et al., 2010, for instance).

After proper normalization of microarray data, DEA can be performed using tools such as linear regressions for each transcript separately. High-dimensionality can be tackled by borrowing information across genes for stabilizing variance estimates with an empirical Bayes moderated  $t$ -test (Smyth, 2004), and through multiple testing correction (Benjamini and Hochberg, 1995).

### 1.2.2 Bulk RNA-seq

Bulk RNA-seq – as opposed to single-cell RNA-seq (see next Section 1.2.3) – denotes the sequencing of all the mRNA extracted from a biological sample using so-called “next-generation” sequencing technologies. It has largely replaced microarray as default technology for measuring gene expression. Indeed, this technology is more versatile allowing applications such as *de novo* discovery of transcripts, targeted immune repertoire sequencing, and it is also more precise (quantification is less sensitive to the expression level) and less ambiguous compared to microarrays.

The processing of RNA-seq data involves several necessary steps before the DEA (Conesa et al., 2016). Notably, once sample sequences have passed quality control checks, sequence reads are mapped to a reference genome (for the corresponding organism, in our case the human genome). This step, called “alignment”, yields a so-called count matrix, where the number of reads successfully mapped

has been counted for each gene (Jin et al., 2017). Hence, RNA-seq assays inherently generate count data. For statistical methods, this count nature of the data induces heteroscedasticity, even if pseudo-alignment methods are used (Srivastava et al., 2020), which further complicates DEA methods for RNA-seq data.

Three methods stand out as the most commonly used in practice for DEA of bulk RNA-seq data: `edgeR` (Robinson et al., 2010), `DESeq2` (Love et al., 2014), and `limma-voom` (Law et al., 2014) (respectively 18,805, 33,562, and 2,802 citations in PubMed as of March 4<sup>th</sup>, 2024). `edgeR` and `DESeq2` both rely on the assumption that gene counts from RNA-seq measurements follow a Negative Binomial distribution, while `limma-voom` is based on a weighted linear model and assumes resulting test statistics follow a normal distribution.

### 1.2.3 Single-cell RNA-seq

Thanks to advances in microfluidics, scRNA-seq allows to measure the gene expression at the cellular level. scRNA-seq makes it possible to simultaneously measure gene expression levels at the resolution of singular cells, enabling a refined definition of cell types and states across hundreds or even thousands of cells at once. Single-cell technology significantly improves on bulk RNA-seq, which measures the average expression of a set of cells, thus mixing the information from various cell types with distinct expression profiles. New biological questions, such as the detection of different cell types or cellular response heterogeneity, can be explored thanks to scRNA-seq, promising to enhance in turn our overall understanding of the features of a cell within its microenvironment (Eberwine et al., 2014).

Several methodological challenges arise from the sequencing of the genetic material of individual cells like in transcriptomics (see Lähnemann et al. (2020) for a thorough and detailed review). Traditional bulk RNA-seq approaches have been applied to single-cell data sets (Soneson and Robinson, 2018; Wang et al., 2019), ignoring the distinctive features of scRNA-seq data. Notably, the latter display large proportions of observed zeros (i.e. “dropouts”), due either to biological processes or technical limitations Lähnemann et al., 2020. Currently, there is no consensus on how to perform adequate DEA on scRNA-seq data, and there is a surge of many different methods being proposed in the literature (e.g. Ozier-Lafontaine et al., 2023, or Yi et al., 2024).

## 1.3 Flow-cytometry data for cellular phenotyping

As highlighted in the previous subsection, the investigation of single cell biology is crucial for improving our understanding of the immune system (De Rosa et al., 2001; Perfetto et al., 2004; Stubbington et al., 2017). Over the past few

## 1 BACKGROUND

decades, and preceding the advent of scRNA-seq, Flow Cytometry (FCM), a high-throughput technology that simultaneously quantifies various cell-surface and intracellular markers at the individual cell level, has become one of the most widely used techniques for single-cell measurements in many immunological studies and clinical trials, and in particular in vaccine trials. This is due to its ability to quantitatively monitor complex cellular immune responses, such as cell phenotype, activation or maturation status, intracellular cytokine or other effector molecule concentrations. This cellular information is critical for the understanding of the immune system, for the development of effective vaccines, and for the discovery of diagnostic or prognostic biomarkers in clinical trials (Darrah et al., 2007; Corey et al., 2015; Lin et al., 2015; Seshadri et al., 2015). Historically, blood cells used to be evaluated manually using a microscope. The flow cytometer, invented by Fulwyler (1965), revolutionized the field by combining optical and computer techniques to automatically measure a tremendous amount of cells in a sample within a very short period of time. Briefly, FCM is a high-throughput, laser-based single-cell technique for measuring the individual cell surface and intracellular marker molecules. The cells (typically from a blood or tissue sample) is first stained with one or more fluorochromes that have been made specific to the cell surface or intracellular proteins of interest; also known as markers. Then FCM measures the cell light scattering and fluorescent intensities. The former provides the information about the cell size and its morphology, and the latter are related to the amount of fluorochrome in the cell or attached at its surface. The higher the fluorescent intensities, the more expressed the corresponding molecular marker. Shapiro (2005) provides a comprehensive introduction to FCM for instance.

One of the fundamental uses of FCM is the identification and quantification of distinct cell subsets with phenotypes characterized by the density of cell surface and intracellular markers (Cossarizza et al., 2021). Technological advancements now allow FCM to measure up to 50 fluorochromes simultaneously on a single cell (Siddiqui and Livák, 2023; BD Biosciences-US, 2019). Meanwhile, Cytometry by Time-Of-Flight (CyTOF), a new concurrent to FCM and closely related technology, that is also called Mass Cytometry and which is based on ion counts, has been developed and could in theory measure up to 100 different cellular markers at once Nowicka et al. (2017). Combining many different cell surface and intracellular marker measurements is critical for identifying cellular populations: the cell subsets identified through FCM can then be tested for their functional properties. For example, the earliest uses of FCM helped to identify major cell lineages, such as T and B-cells which play a fundamental role in the immune system. As FCM now allows more and more markers to be measured, a higher resolution of immune cells profiling can be achieved. For example, we now realize that T-cells can be further distinguished into regulatory T-cells, follicular helper T-cells, and

natural killer T-cells, only to name a few. In most studies, the sample sizes of FCM data are large, reaching several millions of cells being processed from one blood draw (or other biological tissues), although, in many cases the cell subsets of interest are typically in low frequencies (e.g.  $\sim 0.01\%$  of total cells). Hence, there is a critical interest in detecting cellular heterogeneity (and especially very low frequency cell subsets), so that downstream analyses – such as association or prediction studies – can help untangle the link between cellular heterogeneity and the disease progression.

The gold standard for processing FCM data remains manual gating. It is a manual process that uses expert knowledge about the lineage, maturation and activation of cells (Roederer et al., 2004; Perfetto et al., 2004) to manually delineates cells into sequential bounded regions (called gates) on 1-D histogram or 2-D scatter plots pseudo-colored by density. Cells within the region defined by the gates are identified as a specific cell subset. A simplified example to illustrate this sequential process is the task of discriminating  $CD4^+$  T-cells, which is a type of T-cells particularly important in the adaptive immune system. A sequence of subsetting procedures could be performed. Two physical markers, forward and side light-scatter, are first used to construct a 2-D scatter plot for distinguishing lymphocytes from all the live cells. Lymphocytes can then be further partitioned based on the presence or absence of three fluorescence parameters: CD3, CD4 and CD8 cell-surface markers.  $CD4^+$  T-cells are the subclass of lymphocytes having high values of CD3 and CD4 but low value of CD8. In the case of markers for lineage, activation, exhaustion and function, it is customary to dichotomize cells being either positive (+) or negative (–) for each marker (driven by the underlying absence, or presence respectively, of the cell functionality associated with this marker), based on an appropriate negative control (or in some cases, by simply eyeballing the data).

Despite its popular usage in the analysis of (low-dimensional) cytometry data, manual gating has serious limitations including being heavily reliant on local expertise, time consuming, hard to reproduce, and cumbersome in analyzing higher dimensions (since the number of possible 1-D and/or 2-D projections that need to be examined increases rapidly with the number of markers). This partly underlies the drive for automatic cell subset identification to overcome the limitations of manual gating. Numerous methods have been developed to automatically identify and quantify cell populations from such big data (Aghaeepour et al., 2013), but those automated methods have yet to achieve gold-standard performances before immunologists are convinced that they can be widely adopted in practice.

## 1.4 Methodological challenges & Research positioning

The development of high-throughput technologies in biomedical research offers new opportunities to deepen our understanding of biology and human health. Omics data are being generated in ever-increasing quantities. But while this mass of data represents a tremendous wealth of information, its analysis is made difficult by its size (a number of observations that can be very large at the cellular level) and its characteristics such as its high-dimension and its heterogeneity. New approaches are essential to cope with this deluge of complex data and to make the most of all this information to advance knowledge. My research aims at developing rigorous and efficient methods for the analysis of large-scale, repeated biomedical data, that are readily available for the broader scientific community. To achieve this goal, I focus on taking into account high-throughput data specificities with sound statistical approaches and structuring the modeling with *a priori* external biological knowledge.

Beyond the promises of artificial intelligence currently being advertised, a real paradigm shift has taken place within Biostatistics with the emergence of a “data science” mindset. In particular, public health data science integrates biostatistics together with other fields, namely computer science (and in particular bioinformatics and medical informatics), epidemiology, and clinical medicine. As a member of the Data Science division of the LabEx Vaccine Research Institute (VRI), and thanks to my privileged links with the European consortia eboVAC<sup>1</sup> and European HIV Alliance (EHVA)<sup>2</sup>, my methodological developments are heavily influenced by the need to interpret the massive, highly-dimensional and heterogeneous data generated during HIV and Ebola vaccine trials.

In those trials, participants are more often than not followed over time, with repeated high-throughput measurements such as FCM and/or RNA-seq data. Because longitudinal data are not so common with high-throughput measurements, most methods fail to properly consider the additional correlation induced by repeated measures. On the contrary, being able to accommodate such experimental designs (fairly common in clinical trials such as vaccine trials) is paramount to my developments, and I tackle this methodological relative blind spot head-on. In particular, RNA-seq is increasingly being used to assess gene expression over time: the analysis of temporal changes in gene expression contributes to a better understanding of gene regulatory mechanisms, and in the context of vaccine to the establishment and sustainment of a humoral response.

Regardless of repeated measurements issues, inadequate Type-I error control

---

<sup>1</sup><https://www.ebovac.org/>

<sup>2</sup><https://www.ehv-a.eu/>




and inflated false positives have been previously reported for the most common DEA methods (Mazzoni et al., 2015; Rocke et al., 2015; Germain et al., 2016; Rigaiil et al., 2016; Agniel and Hejblum, 2017; Assefa et al., 2018) – even when dealing with only cross-sectional measurements. These problems actually stem from underlying modeling and parametric assumptions, that are typically not verifiable in practice, thereby leading to systematic estimation biases. As sample sizes increase and longitudinal RNA-seq data become more widespread, model inadequacy becomes more pronounced, and this problem of false positives becomes increasingly important (Gauthier et al., 2020; Li et al., 2022). Any deviation from the hypothesized distribution of test statistics will translate into ill-behaved  $p$ -values and therefore uncontrolled False Discovery rate (FDR). FDR control rests upon the entire distribution of  $p$ -values being uniform under the null hypothesis  $H_0$  (i.e. for genes that are truly not differentially expressed) – for instance using the common multiple testing correction from Benjamini and Hochberg (1995). This is a critical point in a context of multiple comparisons linked to high-dimensional data .

In the absence of a consensus on the right underlying probabilistic model for either RNA-seq or scRNA-seq data, my research focuses on deriving tailored DEA methods that avoid any distributional assumption on the data and offer a rigorous control of the Type-I error, even when the parametric modeling of the studied associations is misspecified. This focus on controlling false positives in hypotheses generating assay analyses such as whole genome DEA is important to prevent too many failures in subsequent confirmatory studies that would seek to reproduce biological association. Without it, any downstream health benefits may remain elusive, not to mention the waste of research resources.

Not controlling the FDR means getting more false positives than expected, which limits the reproducibility of study results. Similarly, FCM data are currently processed manually for estimating cellular population proportions. This process is highly variable from one operator to another (Aghaeepour et al., 2013) and thus poorly reproducible, on top of being time consuming and thus expensive. I have proposed both supervised and unsupervised approaches that attempts at delivering more reliable estimates of cellular population proportions, in relation with more general clustering developments. The challenges associated are both methodological and computational. The methodological challenges mainly come from not knowing the actual number of cellular populations present in the FCM samples, from the large scale of FCM data, from the hierarchical nature of the cellular population structure, and from the difficulty of integrating external biological knowledge about cellular populations. The computational complexity is also important, as its burden can quickly increase to yield unrealistic computational times given the large number of cells (several hundreds of thousands, if not

## 1 BACKGROUND

more) in a single FCM sample.

Many factors contribute to the reproducibility of research. In particular, tools and practices have an impact at least as important as the methodological choices. While not being at the center of my work, I have developed a keen interest for methodologies that foster and facilitates reproducibility. I strive to develop open-source software, as user-friendly as possible. Almost all of my work is implemented in  (R Core Team, 1997) as packages available from either the CRAN or Bioconductor. This increases dissemination of my methods, but also facilitates their benchmarking, and participate to good research ethics through transparency. I also try to provide scripts to reproduce my results and analyses, and whenever possible I make the data supporting my findings as openly available as possible – in biomedical research, data privacy is often limiting the extent to which individual information can be shared.

While my current scientific path has been largely influenced by the directions taken during my PhD, I purposely left out those from this synthetic memoir, which focuses instead on published research works where I had a more directive role. The remainder of this memoir is organized as follows: Chapter 2 first presents a DEA framework based on a variance component score test that can be leveraged for analyzing either RNA-seq or scRNA-seq data; Chapter 3 then tackles the automated analysis of FCM data for inference about cell-type proportions; Chapter 4 draws connection with observational studies and Electronic Health Records analysis inspired in part from my postdoctoral work; Chapter 5 highlights my interest in improving reproducible research practices (including mine); Chapter 6 finally outlines future scientific directions I wish to explore with my research.

## 1 BACKGROUND

## 2 Methods for Differential Expression Analysis of RNA-seq data

---

The main content of this chapter has been previously published in the following:

- Agniel D and Hejblum BP. Variance component score test for time-course gene set analysis of longitudinal RNA-seq data. *Biostatistics*, 18(4): 589–604, 2017. DOI: [10.1093/biostatistics/kxx005](https://doi.org/10.1093/biostatistics/kxx005).
- Gauthier M, Agniel D, Thiébaud R, and Hejblum BP. dearseq: a variance component score test for RNA-seq differential analysis that effectively controls the false discovery rate. *NAR Genomics and Bioinformatics*, 2(4): lqaa093, 2020. DOI: [10.1093/nargab/lqaa093](https://doi.org/10.1093/nargab/lqaa093).
- Gauthier M, Agniel D, Thiébaud R, and Hejblum BP. Distribution-free complex hypothesis testing for single-cell RNA-seq differential expression analysis. *bioRxiv*, 2021.05.21.445165, 2021. DOI: [10.1101/2021.05.21.445165](https://doi.org/10.1101/2021.05.21.445165).

---

The general objective of DEA is to identify genes whose expression is significantly associated with a set of clinically relevant characteristics. In my research, I have focused on a new DEA framework based on a variance component score test (Lin, 1997; Huang and Lin, 2013), a flexible and powerful test that requires few assumptions to guarantee rigorous control of Type-I and false discovery error rates. The method is suited to complex experimental designs (comparisons of multiple biological conditions, repeated or longitudinal measurements, integrated supervision by several biomarkers at once).

### 2.1 A working linear model for heteroscedastic gene expression

A large body of statistical methods have been developed to analyze microarray data. But as technology for measuring gene expression transitioned to RNA-seq, new methodological challenges arose. While microarray analysis techniques

generally assume continuity, RNA-seq produces count data and thus RNA-seq data are intrinsically heteroscedastic (even after normalization). Various approaches have been proposed to deal with these issues, mostly relying on modeling the underlying count nature of the data through the use of Poisson or negative binomial distributions (Marioni et al., 2008; Anders and Huber, 2010; Robinson et al., 2010; Law et al., 2014). Unfortunately, all of these methods make potentially restrictive assumptions about the distribution of the data. While my research focuses on methodological solutions that avoid such assumptions, I leverage mixed effects linear models to derive versatile, distribution-free, test statistics for DEA (Agniel and Hejblum, 2017; Gauthier et al., 2020, 2021).

Let  $G$  be the total number of observed genes. Let  $y_i^g$  be the normalized gene expression (any normalization can be used such as log-counts per million values) for the  $g^{\text{th}}$  gene for the  $i^{\text{th}}$  sample,  $i \in 1, \dots, n$ . To build a variance component score test statistic, we rely on the following working linear model for each gene  $g$ :

$$y_i^g = \alpha_0^g + \mathbf{X}_i \boldsymbol{\alpha}^g + \boldsymbol{\Phi}_i \boldsymbol{\beta}^g + \varepsilon_i^g, \quad (1)$$

where  $\varepsilon_i^g \sim N(0, \sigma_i^g)$ ,  $\alpha_0^g$  is the average expression of gene  $g$ ,  $\mathbf{X}_i$  is a vector of covariate values for individual  $i$  to be adjusted upon, and  $\boldsymbol{\Phi}_i$  contains the  $m$  variables for DEA, such as disease status, treatment arm, or other clinical characteristics which are to be associated with gene expression. The parameter of interest is  $\boldsymbol{\beta}^g$ : if  $\boldsymbol{\beta}^g \neq \mathbf{0}$ , then the gene is Differentially Expressed (DE) according to  $\boldsymbol{\Phi}$ . The variance of the residuals  $\varepsilon_i^g$  depends on  $i$  to model the heteroscedasticity inherent to RNA-seq data. Obviously, this individual variance cannot be estimated from a single observation. Instead, information can be borrowed across all  $G$  genes through a local linear regression (Wasserman, 2006) to estimate  $\hat{\sigma}_i^g$ , similarly to Law et al. (2014) but in a more rigorous and principled manner (Agniel and Hejblum, 2017).

Note that the model presented above is very flexible, and can be easily extended to grouped (e.g. repeated or longitudinal) data to take into account heterogeneity between individuals by adding random effects. For instance, in case of longitudinal observations indexed by  $j \in 1, \dots, n_i$ , equation (1) becomes:

$$y_{ij}^g = \alpha_0^g + \mathbf{X}_{ij}^T \boldsymbol{\alpha}^g + \boldsymbol{\Phi}_{ij}^T \boldsymbol{\beta}^g + \boldsymbol{\Phi}_{ij}^T \boldsymbol{\xi}_i^g + \varepsilon_{ij}^g, \quad (2)$$

with  $\boldsymbol{\xi}_i^g \sim N(0, \Sigma_{\boldsymbol{\xi}}^g)$  a vector of length  $m$  of individual-level random effects of the variables of interest  $\boldsymbol{\Phi}$  (Gauthier et al., 2020).

## 2.2 Statistical tests that avoid distributional assumptions

The three most popular approaches for performing DEA, respectively edgeR (Robinson et al., 2010), DESeq2 (Love et al., 2014), and limma-voom (Law et al., 2014),

all have rather strong – although different – parametric assumptions on the distribution of RNA-seq data. However, these methods’ parametric assumptions are not typically verifiable in practice. Any deviation from the hypothesized distribution of test statistics will translate into ill-behaved  $p$ -values and therefore uncontrolled FDR. [Benjamini and Hochberg \(1995\)](#) procedure, the most common out-of-the box approach for FDR control, relies on the assumption that the distribution of  $p$ -values under  $H_0$  is uniform. Therefore, even a slight deviation from strict Type-I error control can have dramatic consequences on the empirical FDR. In addition, even if Type-I error was controlled at say 5%, non-uniformity in the  $p$ -value distribution under the null hypothesis could lead to failure to control the Type-I error at lower levels (such as 1% or lower) and/or failure to control the FDR. Larger sample sizes do not always solve issues with  $p$ -value distributions and FDR control arising from violation of modeling assumptions, and can sometimes even exacerbate the problem of misspecification and its consequences.

My research focused on deriving a global DEA framework, that can be suited to various gene expression data types (and in particular RNA-seq and scRNA-seq) thanks to its lack of distributional assumption on the data. I put particular emphasis on Type-I error control and on having a uniform distribution of  $p$ -values under  $H_0$  (a necessary condition for multiple testing corrections such as the [Benjamini and Hochberg](#) procedure to work) as increases of false positives for state-of-the-art contenders are regularly reported ([Mazzoni et al., 2015](#); [Rocke et al., 2015](#); [Germain et al., 2016](#); [Rigaille et al., 2016](#); [Agniel and Hejblum, 2017](#); [Assefa et al., 2018](#); [Gauthier et al., 2020](#); [Li et al., 2022](#); [Neufeld et al., 2022a](#)).

### 2.2.1 A versatile variance component score test

Variance component tests offer the speed and simplicity of classical score tests, but potentially gain statistical power by using many fewer degrees of freedom. According to the working model (1), a gene is DE and has its expression associated with the variable(s) of interest in  $\Phi$  if  $\beta^g \neq 0$ . `dearseq` thus tests the following null hypothesis for each gene  $g$ :

$$H_0^g : \beta^g = 0. \quad (3)$$

The associated variance component score test statistic can be written as:

$$Q^g = \mathbf{q}^{gT} \mathbf{q}^g \quad \text{with} \quad \mathbf{q}^{gT} = n^{-1/2} \sum_{i=1}^n (y_i^g - \mu_i^g) \sigma_i^{g-1} \Phi_i, \quad (4)$$

where  $\mu_i$  is the conditional mean expression given the covariates  $X_i$ . Again, this formula can easily generalize to more complex experimental designs such as

grouped measurements by incorporating a random-effect covariance matrix, in which case  $\mathbf{q}^g$  becomes:

$$\mathbf{q}^{gT} = n^{-1/2} \sum_{i=1}^n (\mathbf{y}_i^g - \mu_i^g)^T \Sigma_i^{g-1} \Phi_i, \quad (5)$$

with  $\Sigma_i^g$  the covariance matrix of  $\boldsymbol{\varepsilon}_i^g$ .

As this is a score test,  $\widehat{Q}^g$  is only estimated under the null hypothesis of no differential expression, and  $\widehat{\mu}_i$  can be estimated through Ordinary Least Squares (OLS). The asymptotic distribution of the test statistic  $Q$  can be shown to be a mixture of  $\chi_1^2$  random variables  $Q \rightarrow \sum_{\ell=1}^n a_\ell \chi_1^2$  where the mixing coefficients  $a_\ell$  are the eigenvalues of the covariance of  $\mathbf{q}$ . This result rests solely upon the Central Limit Theorem (CLT), and this is why `dearseq` is particularly robust to misspecification: the asymptotic distribution of  $Q$  is the same whether model (1) holds or not. Therefore, the Type-I error (and subsequent FDR) is controlled as long as the CLT is in action (meaning  $n$  is large enough). In practical simulations, convergence occurred around  $n = 40$ . In practice, the saddlepoint approximation for distributions of quadratic forms (Kuonen, 1999; Chen and Lumley, 2019) is an efficient way to compute  $p$ -values for such mixtures of  $\chi^2$ s and it is implemented in the `survey` R package (Lumley, 2004). Finally, to overcome the shortcomings of this asymptotic test in small samples, we propose to use a permutation test using the same statistic  $Q$ . Since we are in a multiple testing setting, it is of the utmost importance to carefully compute the associated  $p$ -values before applying the Benjamini and Hochberg correction according to Phipson and Smyth (2010), who also propose a correction to account for potential repetitions in the (pseudo-) random permutations.

One advantage of using a variance component score test over a regular score test is the gain in statistical power, that comes from exploiting the correlation among  $\widehat{\boldsymbol{\beta}}^g$  coefficients to potentially reduce the degrees of freedom of the test. They have been shown to have locally optimal power in some situations (Goeman et al., 2006). Another advantage is its flexibility that can accommodate random effects in the model to test mixed hypotheses. With a total of  $G$  tests, with  $G$  often greater than 10,000, the computational efficiency of the score test is extremely useful. And given this large  $G$ , it is also absolutely necessary to correct for multiple testing, for instance by using the Benjamini and Hochberg procedure.

### 2.2.2 Extension to Conditional Independence Testing for scRNA-seq data DEA

The next frontier for DEA is the Differential Expression Analysis of single-cell RNA-seq data (scRNA-seq). Sonesson and Robinson (2018) state that bulk RNA-

## 2 METHODS FOR RNA-SEQ DEA

seq DEA methods such as `dearseq` cannot be applied to single-cell data out of the box due to their zero-inflated nature. Interestingly, [Svensson \(2020\)](#) state that this zero-inflation observed in scRNA-seq data is consistent from biological variation and unlikely due to the measurement technique. On the contrary, [Hicks et al. \(2018\)](#) argue that some of this zero inflation is linked to technical variations, while [Townes et al. \(2019\)](#) discuss the role of the log-normalization in this excess of zeros. So once again, I aim at developing a general and flexible method for analyzing scRNA-seq data which do not require strong parametric assumptions.

DEA can be reformulated as a Conditional Independence Test (CIT). A Conditional Independence Test (CIT) broadens the classical independence test by testing for independence between two variables given a third one, or a set of additional variables. Two random variables  $X$  and  $Y$  are conditionally independent given a third variable  $Z$  if, and only if,  $P(X, Y | Z) = P(X | Z)P(Y | Z)$ . Performing DEA necessarily involves performing as many independent tests as there are genes. The variables of interest may be either discrete or continuous, while the number of covariates to condition upon may also increase. Consequently there is an urgent need for a CIT that is both flexible and fast. As described in [Li and Fan \(2020\)](#), many CIT have been developed previously and are readily available such as discretization-based tests. Yet, these CIT either suffer from the curse of dimensionality, or are hardly applicable with a large number of observations, or make strong restriction on the distribution of  $X$  and  $Z$ . Those limitations make these tests impractical in our context of scRNA-seq DEA.

`citcdf` ([Gauthier et al., 2021](#)) is a novel, distribution-free, and flexible approach to test the association of gene expression to one or several variables of interest (continuous or discrete) potentially adjusted for additional covariates. It is a CIT that compares conditional cumulative distribution functions (CCDFs) across conditions. CCDFs are respectively estimated through a series of logistic regressions, thus allowing to leverage a variance component score test to perform DEA. Our null hypothesis for the gene  $g$  is reformulated as:

$$H_0^g : Y^g \perp \mathbf{X} \mid \mathbf{Z}, \quad (6)$$

which is equivalent to the following using the cumulative distribution functions (CDF) with different conditioning:

$$H_0^g : F_{Y^g | \mathbf{X}, \mathbf{Z}}(y, \mathbf{x}, \mathbf{z}) = F_{Y^g | \mathbf{Z}}(y, \mathbf{z}), \quad (7)$$

where the CCDF of  $Y^g$  given  $\mathbf{X}$  and  $\mathbf{Z}$  is defined as  $F_{Y^g | \mathbf{X}, \mathbf{Z}}(y, \mathbf{x}, \mathbf{z}) = \mathbb{P}(Y^g \leq y \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ . If a group of factors is associated with the expression of a gene, the immediate consequence is that the CCDF of the gene expression would be significantly different from the marginal cumulative distribution, which overlooks this conditioning.



Now let's denote  $\mathbf{Y}^g = (Y_1^g, \dots, Y_n^g)$  an outcome vector (i.e. normalized read counts for gene  $g$  in  $n$  cells), and  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$  a  $s \times n$  matrix encoding the condition(s) to be tested along with  $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_n)$  a  $r \times n$  matrix for covariates to be taken into account. Both  $\mathbf{X}$  and  $\mathbf{Z}$  can be either continuous or discrete.  $Y_i^g \in [\zeta_{\min}, \zeta_{\max}]$  and let's consider a sequence of  $p$  ordered and regular thresholds such that:  $\zeta_{\min} \leq \omega_1 < \omega_2 < \dots < \omega_p < \zeta_{\max}$ . For each  $\omega_j$  with  $j = 1, \dots, p$ , the CCDF  $F_{Y^g|\mathbf{X},\mathbf{Z}}(\omega_j | x, z)$  can be written as a conditional expectation:

$$F_{Y^g|\mathbf{X},\mathbf{Z}}(\omega_j | \mathbf{x}, \mathbf{z}) = \mathbb{E} [\mathbb{1}_{\{Y^g \leq \omega_j\}} | \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}] = \mathbb{E} [\tilde{Y}_{ij}^g | \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}], \quad (8)$$

where  $\tilde{Y}_{ij}^g = \mathbb{1}_{\{Y_i^g \leq \omega_j\}}$  is a binary random variable that is equal to 1 if  $Y_i^g \leq \omega_j$  and 0 otherwise. These conditional expectations can be estimated through a sequence of  $p$  working models (one model for each threshold  $\omega_j$ ):

$$g \left( \mathbb{E} [\tilde{Y}_{ij}^g | \mathbf{X}_i, \mathbf{Z}_i] \right) = \beta_{0j}^g + \beta_{1j}^g \mathbf{X}_i + \beta_{2j}^g \mathbf{Z}_i, \quad \forall i = 1, \dots, n \quad (9)$$

where  $\beta_{1j}^g = (\beta_{1j1}^g, \dots, \beta_{1js}^g)$  is the vector of size  $s$  referring to the regression of  $\tilde{Y}_{ij}^g$  onto  $\mathbf{X}_i$  and  $\beta_{2j}^g$  is the vector of size  $r$  referring to the regression of  $\tilde{Y}_{ij}^g$  onto  $\mathbf{Z}_i$ . If  $\mathbf{X}$  has no link with  $Y^g$  given  $\mathbf{Z}$ , then  $\beta_{1j}^g$  will be 0. So finally, we test:

$$H_0 : \beta_{1j}^g = \mathbf{0}, \quad \forall j = 1, \dots, p \quad (10)$$

Many different test statistics can be derived for testing this null hypothesis, but a variance component test akin to the one developed above is particularly suited due to its computational efficiency and its statistical power:

$$D = n \sum_{j=1}^p \sum_{k=1}^s \beta_{1jk}^g{}^2. \quad (11)$$

The computational simplicity of the identity link  $g(y) = y$  in the models (9) allows to estimate  $\hat{D}_n = n \sum_{j=1}^p \sum_{k=1}^s \hat{\beta}_{1jk}^g{}^2$  using ordinary least squares (OLS).  $p$ -values can

then be computed by comparing the observed test statistic  $\hat{D}_n$  to the asymptotic distribution  $\sum_{j=1}^{ps} \hat{a}_j \chi_1^2$ , where  $\hat{a}_j$  are eigenvalues from a consistent estimator of the covariance matrix of a vectorized version of  $\hat{\beta}_1 = (\hat{\beta}_{11}, \hat{\beta}_{12}, \dots, \hat{\beta}_{1p})$  concatenating the  $s$  rows of  $\hat{\beta}_1$  one after another. Lastly, the large number of  $G$  tests once again requires a multiple testing correction afterwards, such as the [Benjamini and Hochberg \(1995\)](#) correction for instance.

## 2.3 Gene set approaches

While most methods for gene expression data focus on univariate differential gene expression analysis, it has been shown that GSA can be a more powerful and interpretable alternative (Subramanian et al., 2005; Hejblum et al., 2015). GSA uses *a priori* defined gene sets annotated with biological functions and investigates their potential association with biological conditions of interest. There are many different approaches to GSA, and a GSA method is typically defined by the type of hypothesis tested as well as how information across genes is aggregated. Rahmatallah et al. (2016) showed that self-contained GSA tests tend to be more powerful and more robust than competitive ones (Goeman and Bühlmann, 2007). Furthermore, some GSA tests rely on univariate gene-level statistics as a first step, aggregating them afterwards in a bottom-up enrichment approach. But when signal strength is weak, single-step top-down GSA methods relying on direct multivariate modeling are better than those enrichment based at leveraging the additional power of GSA (Hejblum et al., 2015).

### GSA for bulk RNA-seq

To study a set of  $U$  genes from a given gene set  $\mathcal{G}$  whose expression is measured in bulk, we extend the model (1) to account for multiple genes  $g = 1, \dots, U$  and potential heterogeneity inside  $\mathcal{G}$  (Ackermann and Strimmer, 2009; Hu et al., 2013; Cui et al., 2016; Hejblum et al., 2015), which then becomes:

$$\mathbf{y}_i = \boldsymbol{\alpha}_0 + \mathbf{X}_i \boldsymbol{\alpha} + \boldsymbol{\Phi}_i \boldsymbol{\beta} + \boldsymbol{\Phi}_i \boldsymbol{\xi}_i + \boldsymbol{\epsilon}_i \quad (12)$$

with  $\boldsymbol{\xi}_i \sim \mathcal{N}_{pK}(\mathbf{0}, \boldsymbol{\Sigma}_\xi)$  is a vector of gene-specific random effects of the testing variables  $\boldsymbol{\Phi}_i$ . The null hypothesis is that both the fixed effects and the variance of the gene-specific random effects of the testing variables are null:

$$H_0 : \boldsymbol{\beta} = \mathbf{0}, \boldsymbol{\Sigma}_\xi = \mathbf{0} \quad (13)$$

where the variance of the random effects being null implies the random effects themselves to be null. The variance component score test statistic is then again:

$$Q = \mathbf{q}^T \mathbf{q} \quad \text{with} \quad \mathbf{q}^T = n^{-1/2} \sum_{i=1}^n \mathbf{y}_{\mu_i}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Phi}_i, \quad (14)$$

and follows asymptotically a mixture of  $\chi_1^2$  distributions with mixing coefficients depending on the covariance of  $\mathbf{q}$  (Agniel and Hejblum, 2017).

**GSA for scRNA-seq**

Similarly, when gene expression is measured through scRNA-seq, equation (9) and its associated test statistic (11) can be extended to test a whole gene set, leveraging the following hypothesis:

$$H_0 : \beta_{1j}^g = \mathbf{0} \quad \forall j = 1, \dots, p, \forall g = 1, \dots, U. \quad (15)$$

The associated test statistics then becomes:

$$S = n \sum_{j=1}^p \sum_{k=1}^s \sum_{g=1}^U \beta_{1jk}^{g^2}, \quad (16)$$

and its limiting asymptotic distributions is still a mixture of  $\chi_1^2$  distributions with mixing coefficients that can be estimated with some additional care given to the crossed covariance terms between different genes.

### 3 Learning cellular population proportions from cytometry data

---

The main content of this chapter has been previously published in the following:

- Commenges D, Alkassim C, Gottardo R, Hejblum BP, and Thiébaud R. cytometree: a binary tree algorithm for automatic gating in cytometry analysis. *Cytometry: Part A*, 93(11): 1132–1140, 2018. DOI: [10.1002/cyto.a.23601](https://doi.org/10.1002/cyto.a.23601).
- Lin L and Hejblum BP. Bayesian mixture models for cytometry data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13: e1535, 2021. DOI: [10.1002/wics.1535](https://doi.org/10.1002/wics.1535).
- Freulon P, Bigot J, and Hejblum BP. CytOpT: Optimal Transport with Domain Adaptation for Interpreting Flow Cytometry data. *Annals of Applied Statistics*, 17(2): 1086–1104, 2023. DOI: [10.1214/22-AOAS1660](https://doi.org/10.1214/22-AOAS1660).
- Hivert B, Agniel D, Thiébaud R, and Hejblum BP. Post-clustering difference testing: valid inference and practical considerations with applications to ecological and biological data. *Computational Statistics & Data Analysis*, 107916, 2024. DOI: [10.1016/j.csda.2023.107916](https://doi.org/10.1016/j.csda.2023.107916).

---

Traditionally, FCM data are analyzed manually by drawing geometric shapes (referred to as “gates”) around populations of interest in a hierarchical series of 1-2 dimensional data visual projections. This process, known as “manual gating”, is time-consuming and highly subjective (Aghaeepour et al., 2013). Modern instruments including both flow and mass cytometers are now capable to quantify between 50 and 100 cellular markers, leading to high-dimensional data space that is impossible to exhaustively explore through manual analysis. Several supervised and unsupervised algorithms have been proposed for automatic gating of FCM data, including model-based clustering approaches (Hejblum et al., 2019; Lin and Hejblum, 2021) among others – see Aghaeepour et al. (2013). Gating thus clusters the observed cells. But this clustering of individual cells is simply a means to an

end, as the clinically relevant information from FCM data is actually the different proportions of the different cell types (Henel and Schmitz, 2007; Maecker and McCoy, 2010).

## 3.1 Unsupervised clustering for cytometry data

### 3.1.1 Clustering

Cluster analysis is ubiquitous in data science to perform data classification, data exploration, and hypothesis generation (Xu and Wunsch, 2008). Clustering aims at grouping homogeneous observations into disjoint subgroups or clusters. When multivariate data are clustered, it is common to study which variables differ between two or more of the identified clusters, in order to interpret the clustering structure and characterize specific clusters.

Despite the widespread use of clustering, there is no commonly accepted and formal definition of clusters (Hennig et al., 2015). In fact, the definition of what a cluster should be varies, depending on the context and the analysis specifics. Everitt and Hothorn (2006) presents a definition that includes only two criteria: i) homogeneity of observations within a cluster and ii) separability of observations between two different clusters. These two criteria are general enough to encompass the majority of the working definitions of clusters, and both can be quantified using various approaches such as distances or similarity metrics, shape of distribution (Steinbach et al., 2004), multimodality (Kalogeratos and Likas, 2012; Siffer et al., 2018), or distributional assumptions (Liu et al., 2008; Kimes et al., 2017).

### 3.1.2 Bayesian mixtures

Mixture modeling is an important statistical framework for performing density estimation and model-based clustering (Bouveyron et al., 2019). A general finite mixture model has the following probability density function:

$$g(y|\theta) = \sum_{k=1}^K \pi_k f(y|\theta_k), \quad (17)$$

where  $y \in \mathbb{R}^d$  is a random vector of length  $d$  representing a single cell with  $d$  measured markers,  $\pi_k$  is the mixture component probability with the constraint that  $\sum_k \pi_k = 1$ , and  $f(\cdot|\theta_k)$  denotes the multivariate density function parameterized by  $\theta_k$  for the  $k$ th mixture component. The Gaussian distribution is commonly used as the base density, but skewed and heavy-tailed distributions such as the (skew)  $t$ -distribution (Azzalini et al., 2016) can be applied directly on the un-transformed

### 3 LEARNING CELLULAR POPULATION PROPORTIONS

data to better accommodate outliers and identify biologically relevant but low probability component structures that deviate from the bulk of the FCM data. Figure 1 displays an example of the mixture of two Gaussian distributions.

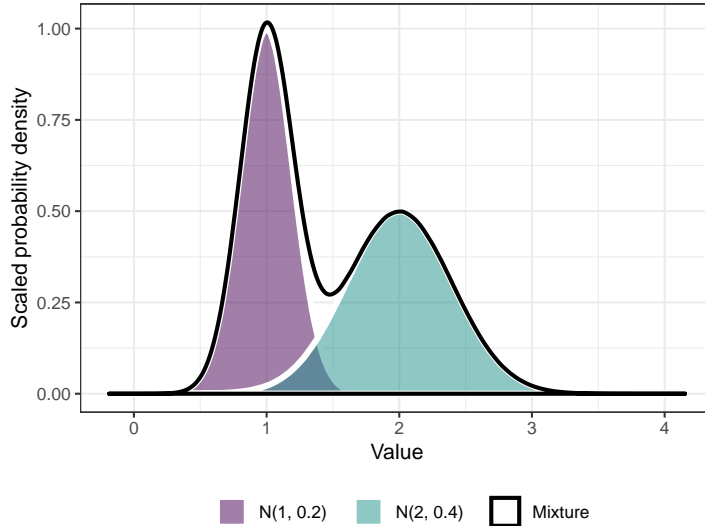


Figure 1: **Example of the probability density function of a univariate mixture of two Gaussian distributions** (Lin and Hejblum, 2021).

In model (17), the number of mixture components  $K$  is typically estimated through model selection (Burnham and Anderson, 2004), e.g. by minimizing Bayesian information criterion (BIC) or Akaike information criterion (AIC). Another approach for addressing this issue of estimating the number of mixture components, i.e. cell types in FCM data, is to leverage the flexibility and adaptability of non-parametric modeling. Non-parametric models scale their complexity to the amount of observations available, as illustrated in Figure 2. In this spirit, in Hejblum et al. (2019) I proposed a Bayesian non-parametric mixture model of skew  $t$ -distributions for clustering FCM data.

Multivariate mixture models rely on strong distributional assumptions that do not necessarily correspond to the reality of FCM data. Moreover, their computational cost quickly becomes important when multiple models are compared to select the best number of clusters or when more sophisticated distributions (such as skew  $t$ ) are considered. To overcome these shortcomings, I have then targeted my research on developing less parameterized methods for automated estimation of cell-type proportions from FCM data.

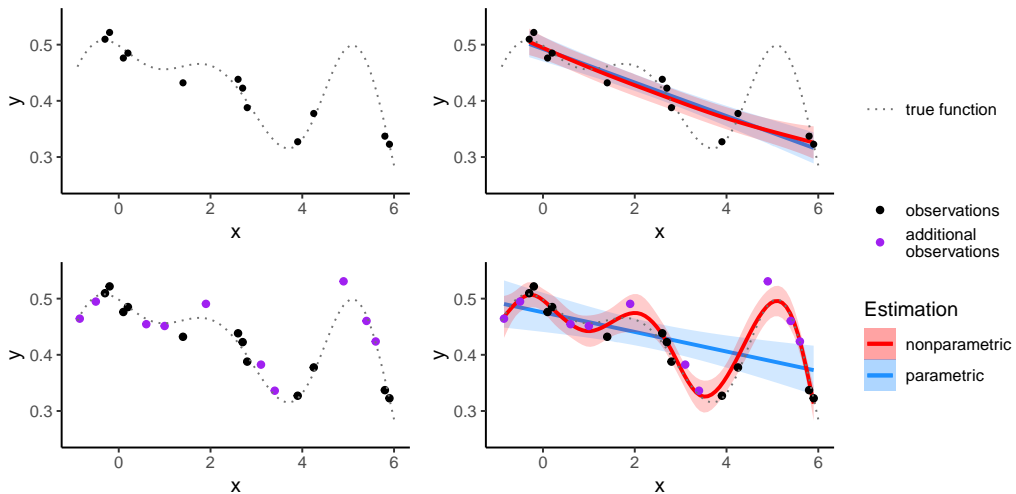


Figure 2: **Illustrating the flexibility of non-parametric modeling on a univariate regression problem by comparing linear fitting with cubic splines.** Non-parametric model complexity scales with the number of observations available.

### 3.1.3 A binary tree algorithm

Many different kinds of approaches have been proposed to automate the gating of FCM data (see [Saeys et al., 2016](#), for a review). A significant number of those have been benchmarked in the open competition set-up by the FlowCAP consortium and reported in [Aghaeepour et al. \(2013\)](#). Many of the compared algorithms presented acceptable performance on the FlowCAP benchmark data. However, no single method was uniformly superior on all data sets. Additionally, some of these methods were very computationally demanding and no method led to biologically interpretable cell populations because their cellular population labels are exchangeable.

To overcome these limitations, *cytometree* ([Commenges et al., 2018](#)) performs automated population identification, based on the construction of a binary tree whose nodes represent cell types. At each node, a univariate marker distribution is modeled by a mixture of Gaussian distributions with either just one or two components (estimated through maximum likelihood, with an EM algorithm algorithm in the case of 2 components). Then node splitting leverages model selection between those 2 mixtures, based on a normalized difference of Akaike Information Criteria (AIC) between two competing models ([Commenges et al., 2008](#)). This criterion has the advantage of being independent of the number of observations (i.e. cells), as it estimates the difference between Kullback-Leibler divergences from the

true generative distribution. At each node, all the markers are considered, and the split is performed according to the marker with the largest normalized AIC difference. The tree stops growing once there are no more markers with a difference larger than a tuning threshold. Finally, cell types are biologically labeled based on marker expression levels in post-processing the binary tree built.

`cytometree` is fast and one of its advantages is its numerical simplicity and stability. When benchmarked on data from both the FlowCAP I challenge, and the Human Immunology Project Consortium (HIPC) T-cell panel, it outperforms the best unsupervised open-source available algorithm while requiring the shortest computation time. Because it basically performs recursive thresholding of marginal densities based on the assumption that cells express or do not express certain markers, `cytometree` intrinsically assumes bimodality of the FCM markers. While this assumption is reasonable in most scientific applications, some FCM markers (e.g. functional markers) might not be truly bimodal. In this case, these markers would likely not be thresholded and thus would not be represented in the gating tree. Different cases may occur, e.g. a marker may exhibit trimodality, in which case this feature can be retrieved through the annotation process of `cytometree`. Of note, a truly “continuous” marker is not useful for distinguishing cell types. As with all unsupervised algorithms, `cytometree` has difficulties in reliably identifying small populations. For rare populations, marginal density estimates are unlikely to be clearly bimodal. In such cases, some form of *a priori* knowledge is probably necessary. Finally, it should be noted that because of the bimodality assumption `cytometree` is not adapted to gating light scatter channels (i.e. Forward Scatter Channel and Side Scatter Channel) and as such it should be applied once these have been already (manually) gated (e.g. applied to the lymphocyte population only).

`cytometree` can be extended to deal with CyTOF data. The two main differences of CyTOF data compared to FCM data is the important number of zero values coupled with an increased dimensionality. A careful transformation of the data with the  $\text{arcsinh}()$  function associated with a dynamic exclusion of zero values allows to process and annotate CyTOF data in a similarly satisfactory manner. This (unpublished) extension is called `cytoftree` and is implemented and documented in the companion `cytometree` R package.

## 3.2 Supervised automated gating

Currently, fully automated, unsupervised gating approaches still fall short to being able to completely replace manual gating. Notably, they experience difficulties to tackle the heterogeneity between different FCM data sets, to account for outlier cell events, and to discriminate cells on their morphological features (the first



gatings used to separate broad cellular families from one another based on FSC and SSC). The most hopeful avenue to speed up processing and analysis of FCM is thus tailored supervised approaches that require some initial input from the biologists, accounting for the data set as well as the scientific question specifics, before leveraging automatic data processing.

### 3.2.1 Semi-supervised automated gating

One way to build supervised automated approaches is to leverage prior biological knowledge about already known cell types and their discriminating features to guide the algorithms. In Bayesian mixture models, this translates into using informative priors that can inform the model on regions of the space where cellular clusters are expected (Hejblum et al., 2019; Lin and Hejblum, 2021). When building binary trees, as with `cytometree`, the path and order of successive markers used for splitting the cells can be forced according to prior biological knowledge. The end leaves can then be left for additional unsupervised binary partitioning, to identify rare cell sub-types if any. The same kind of supervision can also be enforced also during the post-processing annotation of clusters.

### 3.2.2 Optimal transport for accelerated supervised gating

As stated before, from a clinical perspective the relevant information in FCM data is the cell type proportions (Maecker and McCoy, 2010), and clustering of the observations is not actually required for its estimation. Cluster allocation is actually a latent instrument variable, only needed by approaches such as mixture models. With that in mind, it makes sense to directly aim for the estimation of the different cell type proportions. To that end, `CytOpT` (Freulon et al., 2023) uses regularized Optimal Transport (OT), supervised by the prior (manual) gating in one reference FCM sample. The cell type proportions from that one reference sample are “transported” in multiple other FCM data sets.

OT has recently gained interest in machine learning and statistics, thanks to approximate solvers for large dimension problems that drastically alleviated its high computational cost. OT defines a metric between two probability distributions  $\alpha$  and  $\beta$  both supported on  $\mathbb{R}^d$ . This metric can be informally defined as the lowest cost to move the mass from one probability measure, the source measure  $\alpha$ , onto the other, the target measure  $\beta$ . In our context, the reference FCM sample for which gating is known is called the “source”, while the other FCM data sets for which there is no gating are the “targets”.

For the target sample  $Y_1^t, \dots, Y_J^t$ , the segmentation into various cell types is not available, and instead the empirical target measure can be defined as:  $\hat{\beta} = \frac{1}{J} \sum_{j=1}^J \delta_{X_j^t}$ . Similarly, the empirical source measure from the source observations

### 3 LEARNING CELLULAR POPULATION PROPORTIONS

$X_1^s, \dots, X_I^s$  is defined as  $\hat{\alpha} = \frac{1}{n_X} \sum_{i=1}^{n_X} \delta_{X_i^s}$ . The knowledge of the segmentation (i.e. gating) of the source data allows to re-write  $\hat{\alpha}$  as a mixture of probability measures where each component corresponds to a known cell type:

$$\hat{\alpha} = \sum_{k=1}^K \frac{n_k}{n_X} \left( \sum_{i: X_i^s \in C_k} \frac{1}{n_k} \delta_{X_i^s} \right) = \sum_{k=1}^K \frac{n_k}{n_X} \hat{\alpha}_k, \quad (18)$$

where  $n_k = \#C_k$  is the number of cells of type  $k$  and  $\hat{\alpha}_k = \sum_{i: X_i^s \in C_k} \frac{1}{n_k} \delta_{X_i^s}$ . Namely, the component  $\hat{\alpha}_k$  is the empirical measure of the observations that belong to the known cell type  $C_k$ . Then, instead of only considering the true class proportions  $(n_1/n_X, \dots, n_K/n_X)$  in the source data set, we can re-weight the clusters in the empirical distribution as desired, borrowing ideas from domain adaptation techniques (Redko et al., 2019). Indeed, for a probability vector  $\theta = (\theta_1, \dots, \theta_K) \in [0, 1]^K$  we can define the re-weighted measure  $\hat{\alpha}(\theta)$  defined by  $\hat{\alpha}(\theta) = \sum_{k=1}^K \theta_k \hat{\alpha}_k$ . The class proportions in the target data are estimated by minimizing the regularized Wasserstein distance – accounting for possible mis-alignment of a given cell population across samples (e.g. due to technical variability) – between  $\hat{\alpha}(\theta)$  and the target empirical distribution according to  $\theta$ . Indeed, the source distribution will get closer to the target distribution as the class proportions in its re-weighted version get closer to the class proportions of the target distribution. A stochastic gradient ascent algorithm is used to solve a regularized version of this minimization problem, in the absence of a closed form solution.

Of note, regularized OT offers a natural soft assignment method, which can be used to derive a clustering of the target data set. By choosing the class with highest probability in the estimated transport plan, we can derive a clustering for each target observation. However, this requires additional calculations on top of what is strictly necessary for the sole estimation of class proportions with CytOpt.

### 3.3 Post-clustering inference

While clustering usually takes into account all variables in a data set, only a smaller set of variables can be expected to differentiate two particular clusters (i.e. separate their observations, according to the second criterion of our definition above). This question, of which variables separate clusters of individuals, is particularly relevant for high-dimensional data such as scRNA-seq data (Lähnemann et al., 2020), but also for annotation of unsupervised clustering applied for automated gating of FCM data.

Unfortunately, the current practice to identify such variables is often based on post-clustering hypothesis testing. Post-clustering inference refers to the second step of a two-step pipeline (first step is clustering, second is inference). This

### 3 LEARNING CELLULAR POPULATION PROPORTIONS

pipeline thus tests data-driven hypotheses in a process sometimes referred to as “double dipping” (Kriegeskorte et al., 2009). Without appropriate care, this double use of the data violates the requirement of *a priori* hypotheses and does not preserve the control of Type-I error enjoyed by classical inference when testing for differences between clusters.

It is always possible to cluster the data using a clustering method (even if there is no real separation of observations). The clustering then artificially creates differences between observations by dividing them into clusters. Significant differences between clusters identified through subsequent statistical testing can be artifacts originating from the first clustering step itself. Figure 3 illustrates this phenomenon in a toy example with data generated from a  $\mathcal{N}(0, 1)$  and two artificial clusters built using hierarchical clustering. Over 2,000 simulations, the t-test leads to a dramatic inflation of false positives. This behavior can be corrected by properly accounting for the clustering step in a modified test, for instance using the selective global test from Gao et al. (2024).

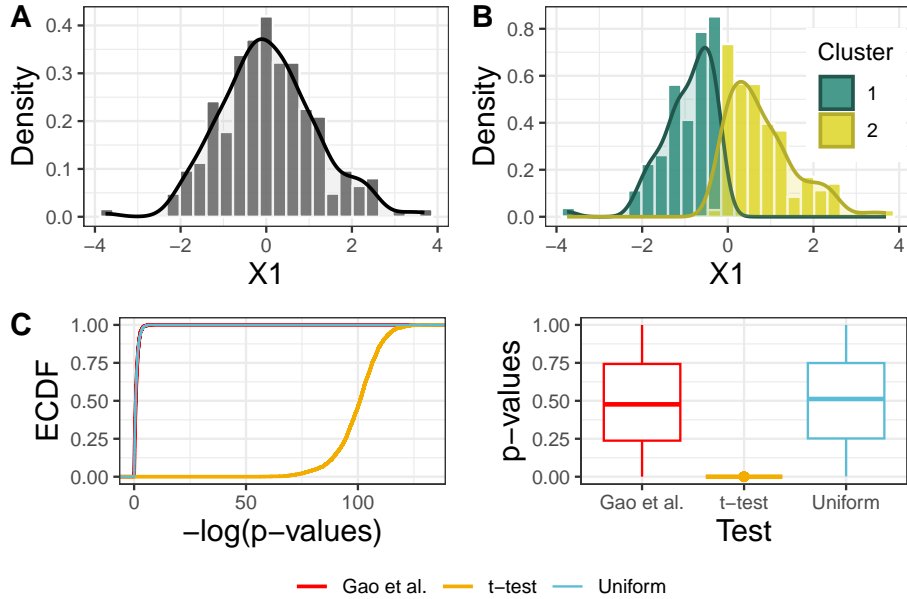


Figure 3: **Artificial differences created by clustering** (Hivert et al., 2024). **A)** Data generated according to 200 realizations of a Gaussian distribution with mean 0 and variance 1. **B)** Hierarchical clustering with Ward method and Euclidean distance is applied to build two clusters. **C)** t-test p-values and p-values given by the test proposed by Gao et al. (2024) for separating the two estimated clusters. The uniform distribution is also shown for comparison.

### 3 LEARNING CELLULAR POPULATION PROPORTIONS

Recently, [Gao et al. \(2024\)](#), [Neufeld et al. \(2022a\)](#), [Neufeld et al. \(2022b\)](#), [Chen and Witten \(2023\)](#), and [González-Delgado et al. \(2023\)](#) all proposed new developments for post-clustering inference, demonstrating the importance of this very active subject. In an effort to derive tests that take into account the first clustering step and the potential artificial differences it may introduce, I am also working on new methods for post-clustering inference. In particular, I am interested in testing the null hypothesis that a particular variable does not truly separate two of the estimated clusters, for any clustering method. This null hypothesis means that a variable can either: i) not be involved in the separation of the two clusters and unaffected by the clustering step, or ii) only be involved in this separation because the clustering applied induced artificial differences.

The selective inference test from [Gao et al. \(2024\)](#) can be extended to test for univariate separability, and also to the case where there are more than two clusters thanks to p-values aggregation for dealing with interjecting clusters ([Hivert et al., 2024](#)). As the original global test, this assumes Gaussian data. At the core of this kind of selective test lies the variance parameter. This parameter (or its equivalent in other distributions – e.g. over-dispersion in the negative binomial distributions) is assumed to be known in theory, while it is not the case in practice. While the hope is that it would be possible to use plug-in estimator instead, it is very difficult in practice to correctly estimate the covariance of the data without knowing their true clustering structure. This remains an open-question in current selective approaches to post-clustering inference. [Gao et al. \(2024\)](#) have showed type-I error control is guaranteed with an overestimated variance, at the cost of being overly conservative. A more practical compromise is to use partial variance in the univariate case, only taking into account the observation from the tested clusters. This highlights the intrinsic difficulty of variance estimation in post-selective inference, an issue tightly related to the problem of post-clustering testing itself.

Another approach at this problem is to directly test the separability of the data, without making any assumption on their distribution. This can be done by leveraging the dip test ([Hartigan and Hartigan, 1985](#)) and testing the unimodality of the data on a given variable ([Hivert et al., 2024](#)). The test leverages the limiting case of the uniform distribution, which is the unimodal distribution with the asymptotically largest dip statistic. Thus, a distribution with a dip statistic larger than that of the uniform distribution is unlikely to be unimodal. In practice, this multimodality test may often be preferred thanks to its attractive computational cost and its absence of distributional hypothesis, unless heterogeneity occurs within clusters in which case multimodality could become a poor indicator of cluster separation.

### 3.4 Application to FCS data from the T-cell panel HIPC study

The Human Immunology Project Consortium (HIPC) was developed with the aim of standardizing flow cytometry immunophenotyping in clinical studies. [Finak et al. \(2016\)](#) investigated whether automated gating could help standardizing FCM data analysis. In the T-cell panel of the HIPC Lyoplate study, seven laboratories (or centers) stained three replicates of three cryopreserved Peripheral Blood Mononuclear Cells (PBMC) samples and returned usable FCS files to the main center for manual and automated gating. T-cells were characterized across 7 cellular markers (namely CCR7, CD4, CD45RA, CD3, HLADR, CD38, and CD8). In addition, we have a reference manual gating of those cells into 8 mutually exclusive populations (2 additional gated populations – namely “CD4 Activated” and “CD8 Activated” – overlap with the other cell populations and therefore were not considered). See [Figure 4](#) for a descriptive representation of the “1228R1” sample (replicate 1 of subject 1228 processed at the Stanford laboratory) before standardization. The automated gating used a combination of algorithms including `flowDensity`, which is a supervised algorithm. Data sets are publicly available from the ImmuneSpace database ([Brusic et al., 2014](#)) and were used as part of the FlowCAP III challenge.

#### Unsupervised automated gating with `cytometree`

The F-measure is the harmonic mean between precision and recall, and allows to compare automated gating to reference manual gating, similarly as in [Aghaeepour et al. \(2013\)](#). An F-measure of 1 means the clustering result is a perfect reproduction of the manual gating result, and the worst value of F-measure is 0. In most cases, the  $F$ -measures obtained by `cytometree` were high with an average of 0.86, and notably better than those obtained by competing methods benchmarked in [Aghaeepour et al. \(2013\)](#). [Figure 5](#) illustrates an example of a `cytometree` binary tree obtained on a FCM sample from this T-cell panel of HIPC Lyoplate study. Of note, the variability of `cytometree` estimates was similar to that of manual gating (except for CD8 effector T-cells – this is in line with [Finak et al. \(2016\)](#), who showed that the CD8 effector T-cell subset was problematic due to poor separation between the HLA-DR<sup>-</sup> and HLA-DR<sup>+</sup> populations).

#### Supervised automated gating with `CytOpt`

Now considering only two aggregated cell types, namely CD4 and CD8 T-cells, `CytOpt` adequately retrieves the true class proportions of an unlabelled cytometry

### 3 LEARNING CELLULAR POPULATION PROPORTIONS

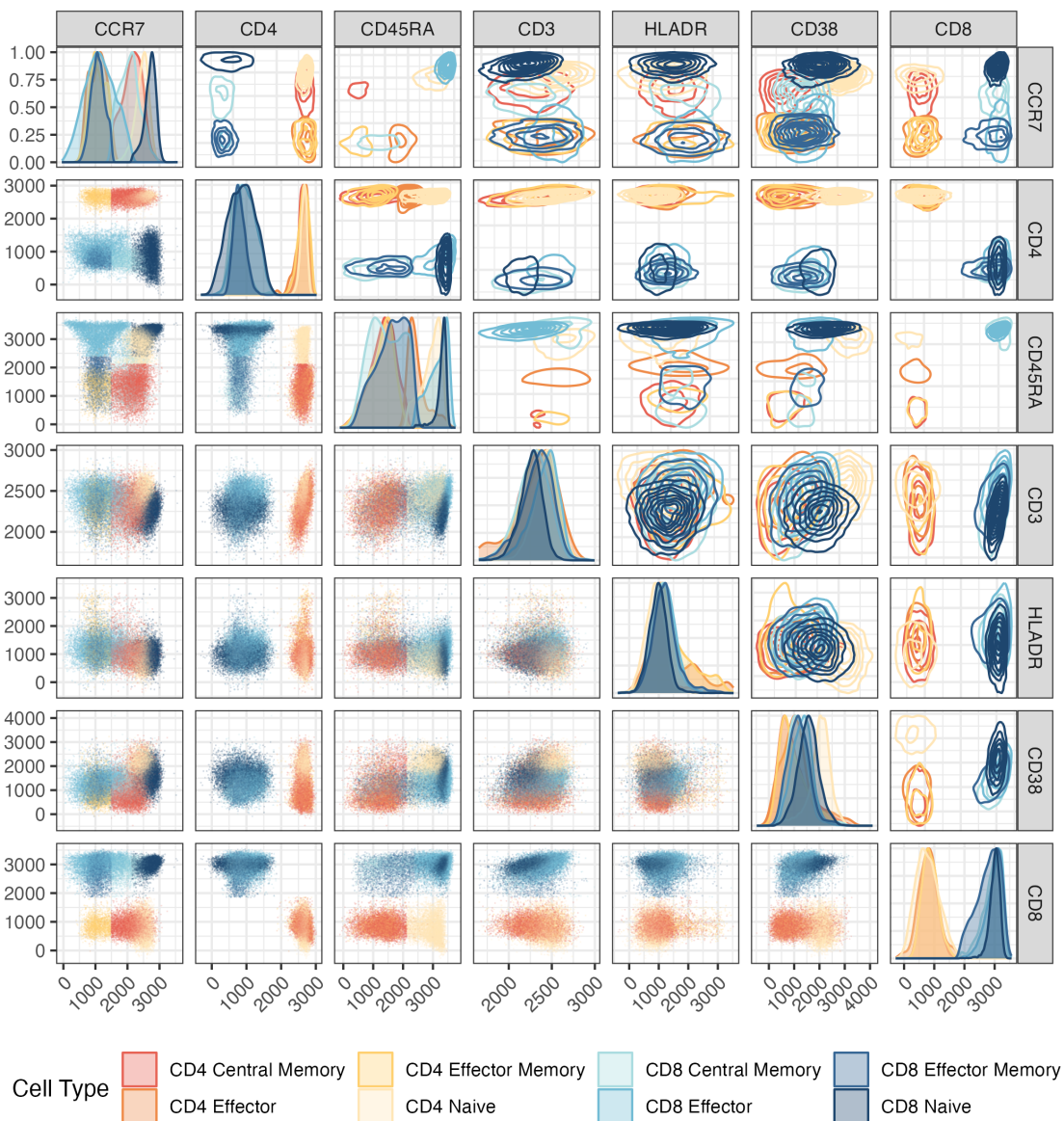


Figure 4: **Visualization of the data from one FCM sample with manual gating** (adapted from [Lin and Hejblum, 2021](#)). This FCM sample is the replicate 1 of individual 1228 processed at Stanford from the T-cell panel in the HIPC Lyoplate study. 30,427 cells are displayed before standardization of the features. Diagonal plots represent marginal densities per cell population, lower triangle plots are 2D scatter plots of gated cells, and upper triangle plots represent bi-variate densities per cell population.

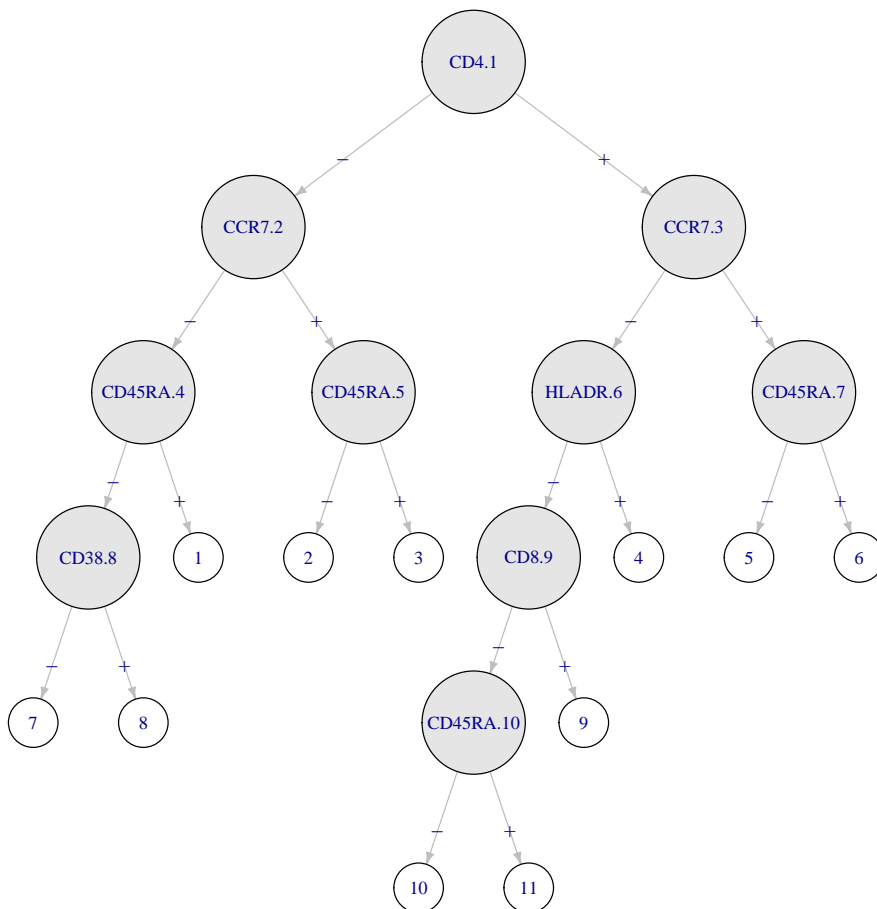


Figure 5: **Partitioning tree for the T-cells of individual 1349 (replicate 3) from the Stanford data set (Commenges et al., 2018).** Each node that has children is labeled with the marker upon which the cell-subpopulation is further split and end-leaves are arbitrarily numbered.

data set (Stanford 1349R3) from transporting the labels of a reference source (here Stanford 1228R1), as represented in Figure 6. In the source data set the CD4 cells constitute 45.1 % of the cells and the CD8 cells 54.9 % respectively, whereas in the target the CD4 cells constitute 73.9 % of the cells and the CD8 cells 26.1 % – estimated at 73.3 % and 26.7% respectively by `CytOpt`. When applied to the full T-cell panel FCM data from the HIPC Lyoplate (i.e. with  $d = 7$  markers) to estimate all cell types proportions in each of the other 61 unsegmented data sets targeted, using only the 1228R1 sample as the reference source with its manual gating available, `CytOpt` estimates fall within 5% of the true manual gating gold-standard proportion in more than 90% of the cases.

### 3 LEARNING CELLULAR POPULATION PROPORTIONS

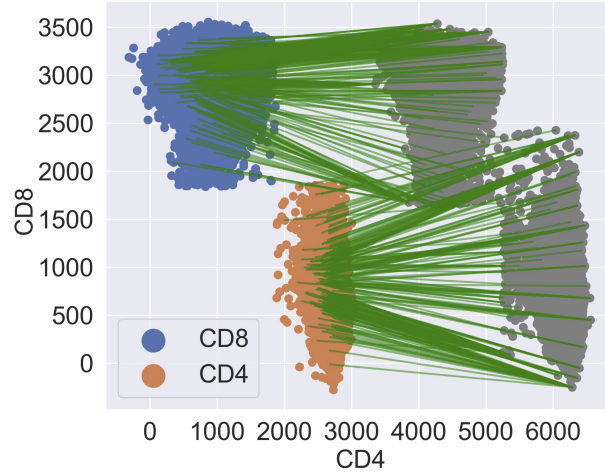


Figure 6: **Optimal transport plan between the source and target distribution from CytOpt (Freulon et al., 2023).** A green line between  $X_i^s$  and  $X_j^t$  indicates that the optimal transport plan moves some mass from  $X_i^s$  to  $X_j^t$ . For better readability, the target data set has been shifted and only 500 coefficients of the transport plan have been represented.

### Post-clustering inference

Given the size of the data and the computational burden of some of the considered tests, a subset of 5% of the cells were used to assess the ability of post-clustering inference tests to identify known specific cellular markers following data-driven clustering of the FCM data. Of note, we discarded the CD3 marker as this marker is not discriminative among T-cells (indeed, this is a marker often used to identify T-cells from other lymphocytes such as B-cells). Focusing on only 4 of the 10 gated cell types (CD8 Naive, CD8 Effector Memory, CD4 Naive and CD4 Effector Memory), a total of 1,051 cells was analyzed. Clustering was first performed with hierarchical clustering (with Ward’s method on Euclidean distances on the scaled data) with  $K = 4$  clusters, yielding an Adjusted Rand Index of 0.98 compared to the manual gating gold standard. Then, each of the 6 markers was tested as potential separator for each pair of estimated clusters and the resulting p-values for the comparison of Cluster 2 (containing 90% of CD8 Effector Memory cells) and Cluster 4 (containing 99% of CD4 Effector Memory cells) are given in Table 1. Most markers were identified as significantly separating cluster pairs. This exemplifies one of the limitations of such univariate selective inference tests in presence of correlated descriptors: because almost any couple of markers is sufficient to discriminate between the four cell sub-populations, perturbing one marker still allows recovery of the original clustering structure of the data (based on the remaining markers) leading to significance. This phenomenon is most acute when CD4 is



### 3 LEARNING CELLULAR POPULATION PROPORTIONS

estimated to contribute significantly for separating CD4 Naïve T-cells from CD4 Effector-Memory T-cells although the CD4 marker is supposed to be expressed in both cell populations. The multimodality test, which is only based on the separation between clusters, returned less significant markers, identifying only the most meaningful ones for biological annotation of the clusters (Figure 7). It was more difficult for the multimodality test to distinguish between the Naive or Effector Memory cells within CD4<sup>+</sup> or CD8<sup>+</sup> T-cells. This can be explained because CCR7 and CD45RA are not the canonical markers usually used to differentiate between those cellular subtypes – and no single specific marker of Effector Memory T-cells has even been identified so far (Saxena et al., 2019).

Table 1: **P-values for the comparison between Cluster 2 (90% of CD8 Effector Memory cells) and Cluster 4 (99% of CD4 Effector Memory cells) estimated from the HIPC data (Hivert et al., 2024).**

	Merged selective test	Multimodality test
CCR7	0.0005*	0.8611
CD4	0.0005*	0.0000*
CD45RA	0.0005*	0.9973
HLADR	0.2231	0.9960
CD38	0.0013*	0.7312
CD8	0.0005*	0.0000*

\* significant at the 5% level

### 3 LEARNING CELLULAR POPULATION PROPORTIONS

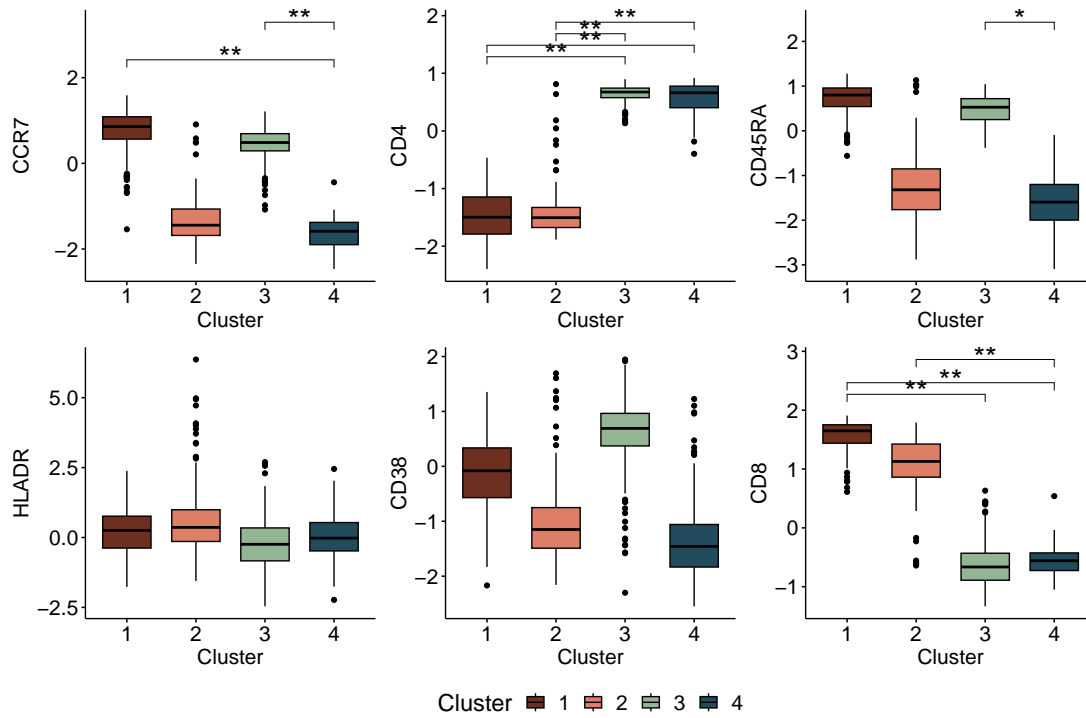


Figure 7: **Post-clustering inference with the multimodality test applied on the HIPC data allowed cells gating** (Hivert et al., 2024). Each plot gives the distribution of clusters on a surface marker. A significant separation between two clusters at the  $\alpha = 1\%$  on a marker is highlighted by \*\* and a significant separation between two clusters at the  $\alpha = 5\%$  by \*.

### 3 LEARNING CELLULAR POPULATION PROPORTIONS

# 4 Connections with observational studies, Electronic Health Records data and surrogate marker evaluation

---

The main content of this chapter has been previously published in the following:

- Ferte T, Cossin S, Schaefferbeke T, Barnette T, Jouhet V, and Hejblum BP. Automatic phenotyping of electronic health record: Phevis algorithm. *Journal of Biomedical Informatics*, 117: 103746, 2021. DOI: [10.1016/j.jbi.2021.103746](https://doi.org/10.1016/j.jbi.2021.103746).
- Ferté T, Jouhet V, Greffier R, Hejblum BP, and Thiébaud R. The benefit of augmenting open data with clinical data-warehouse EHR for forecasting SARS-CoV-2 hospitalizations in Bordeaux area, France. *JAMIA open*, ooac086, 2022. DOI: [10.1093/jamiaopen/ooac086](https://doi.org/10.1093/jamiaopen/ooac086).
- Agniel D, Hejblum BP, Thiébaud R, and Parast L. Doubly-robust evaluation of high-dimensional surrogate markers. *Biostatistics*, 24(4): 985–999, 2023. DOI: [10.1093/biostatistics/kxac020](https://doi.org/10.1093/biostatistics/kxac020).

---

## 4.1 Electronic Health Records (EHR)

### 4.1.1 Similarities between EHR and transcriptomic data

Statistical challenges raised by high-dimension, zero inflation, repeated observations and large volume of data are also present in EHR. On these data, phenotyping algorithms, i.e. probabilistic identification of a phenotype of interest, help to better identify certain events such as chronic conditions, infections or vaccination. The aim is to easily create cohorts of patients with certain health profiles, to facilitate the exploration of many questions in clinical research. While I participated in the development of statistical methods involving high-dimensional phenotyping leveraging data from EHR either in English or in French, those are not immediately

linked to vaccine development, but instead focus mostly on the identification of chronic health conditions, such as rheumatoid arthritis (Liao et al., 2017; Sinnott et al., 2018; Zhang et al., 2021).

To delve deeper into the comparison, EHR are a very rich but also a very noisy source of health information, not unlike gene expression. From a methodological standpoint, they share a lot of common features, starting with their high-dimension (e.g. there are more than 69,000 ICD-10 diagnosis codes and at least as many procedure codes) or the important number of zeros. Thus multiple testing and computational constraints are important in this context as well to produce effective inference. Furthermore, diagnosis codes are also count data, and they similarly range across multiple scales with different levels and aggregates, thus being well suited to grouped testing for so-called phenome-wide studies.

That being said, the generative process behind EHR data is quite different from the one of transcriptomics', as they are not from high-throughput measurements. In addition, those data are first and foremost collected as healthcare data, and research is not part of their primary intent. As such, while most of the omic data I study come from vaccine clinical trials, EHR data are usually observational which further complicates result interpretation.

#### 4.1.2 Monitoring infections leveraging EHR from hospital data-warehouse

Contrary to chronic conditions, most infectious diseases targeted by vaccines are only temporary (with the notable exception of HIV), and once identified by the healthcare system and treated adequately. Automatic phenotyping of patients from EHR usually relies on either rule-based algorithms specifically designed with clinicians, or on supervised models trained on annotated patient data sets. Such algorithms are limited because their development is disease specific, must be (re-)started from scratch for every new disease and demand a lot of clinician expertise time. In addition, portability and generalization to new databases (e.g. different hospitals) can often fail, requiring once again the process to be reiterated in the new institution. Hripcsak and Albers (2013) defined high-throughput phenotyping as an approach that "should generate thousands of phenotypes with minimal human intervention". More recently, several unsupervised frameworks have been proposed as they require neither manual chart review nor complex rule definitions to classify phenotypes, and thus allow automated high-throughput phenotyping (Agarwal et al., 2016; Halpern et al., 2016; Yu et al., 2018; Waghlikar et al., 2020). The main limitations of those frameworks is that they all consider phenotyping at the patient level, while neglecting the timing of illness onset and cure.

Yet, studying acute diseases such as infections (that can occur repeatedly) requires an increased resolution for phenotyping. Phenotyping at the finer scale of the hospital visit allows to precisely take into account the dynamic evolution of patient’s conditions. **PheVis** (Ferte et al., 2021) accumulates past information to provide an up-to-date estimation of a phenotype probability at any given visit. This accumulation of previous information from EHR can be tuned to match the condition duration, making **PheVis** a versatile tool suitable for both chronic conditions as well as acute infections. Briefly, **PheVis** combines ICD-10 codes together with medical concepts extracted from clinical notes (thanks to NLP), incorporating past information through a user-tunable exponential decay. This creates a silver-standard surrogate of the medical condition of interest. Then variable selection (through elastic-net logistic regression) and pseudo-labeling (using random-forest) are performed, leveraging extreme values of this silver-standard. Finally, a logistic regression model is estimated on those noisy labels to provide an interpretable parametric predictor of the occurrence probability for a given medical condition at each visit.

Applied to identify active tuberculosis infection (an acute disease which usually lasts between 6 to 12 months) in a study cohort of 11,461 rheumatology patients, **PheVis** obtained a good Area Under the Receiving Operator Curve of 0.987 [0.983 ; 0.990] but a more nuanced Area Under the Precision Curve of 0.299 [0.198 ; 0.403] against a manual-chart review gold-standard. Both results represent improvements over the state-of-the-art methods, and this highlights the current limitations for leveraging EHR to study acute conditions and transient events such as infectious diseases.

### 4.1.3 Predicting COVID-19 hospitalization leveraging EHR

During the latest pandemic of Coronavirus disease 2019 (COVID-19), I participated in a broad local effort in providing short-term forecast of the COVID-19 hospitalization burden in terms of needed hospital beds. Aggregated data from Severe Acute Respiratory Syndrome Corona Virus 2 (SARS-CoV-2) and weather public database were joined together with data from the data-warehouse of the Bordeaux Hospital ranging from 2020-05-16 to 2022-01-17 (88 weeks) in order to predict the number of hospitalized patients in the Bordeaux Hospital 7 and 14 days later (displayed in Figure 8). This contained information about hospitalizations, RT-PCRSARS-CoV-2 diagnoses, weather humidity indexes, vaccine rates, SARS-CoV-2 variants, emergency units activity, ambulance service activity, vaccine and majority variant. Several feature engineering transformations (namely 7-day average, minimum and maximum, as the first derivatives over the last 3, 7, 10 and 14 days), to amount to a total of 2,990 potential predictors, were fed into a linear model with elastic-net penalty (Zou and Hastie, 2005). Of note, predictors were

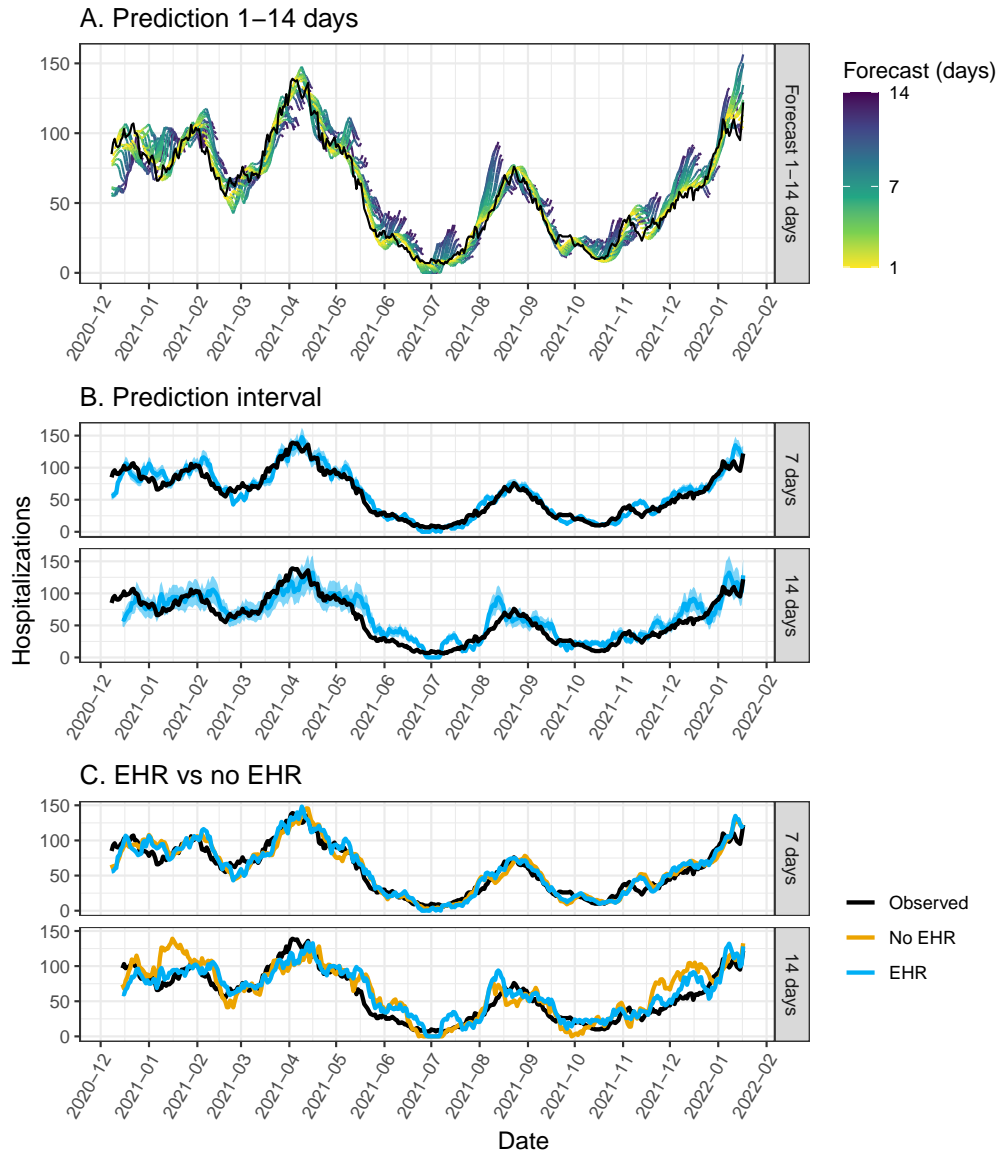


Figure 8: **Hospitalization forecasts up to 14 days at the Bordeaux hospital** (Ferté et al., 2022). **A.** model predictions from 1 to 14 days. **B.** prediction intervals of the forecast using an *ad hoc* rule of respectively 20% and 40% prediction interval at 7 and 14 days (which has better coverage percentage than bootstrapped prediction intervals). **C.** comparison of model predictive performance when information from the EHR data are included or not.

smoothed using local polynomial regression with a span of 21 days to account for outliers and weekly variations.

Figure 8 displays the good performances of the model forecast, except (i) in December 2020 during a nosocomial cluster at Bordeaux Hospital, (ii) in the end of March 2021 where April decrease is anticipated 2 weeks earlier, and (iii) during summer 2021 and winter 2021-2022 where hospitalizations are overestimated due to RT-PCR diagnoses massive increase not associated with a similar hospitalization increase, likely a combined consequence of the vaccination campaign and the spread of the Omicron (B.1.1.529) variant.

While this kind of approach is of prime interest for informing public health decision and hospital organization, it is relatively limited to fully capture the vaccine impact on the dynamics of the epidemic, let alone further understand and evaluate the effectiveness of vaccination against the SARS-CoV-2. Alternatively, more mechanistic approaches, in the spirit of [Collin et al. \(2023\)](#), can provide more insights.

## 4.2 Considering gene expression as a potential surrogate marker

At the core of methods leveraging EHR data lies the issue of correctly identifying medical information from healthcare databases. A problem that can also be viewed as the construction of good surrogate marker from the biostatistician standpoint. Surrogate markers aim at being validated proxys that can safely replace a clinical endpoint in a study. When evaluating the effectiveness of a vaccine or a treatment, a policy, or an intervention, the desired measure of efficacy may be expensive to collect, not routinely available, or may take a long time to occur. In these cases, it is sometimes possible to identify a surrogate outcome that can more easily, quickly, or cheaply capture the effect of interest.

Because it plays a central role in protein production and largely determines cellular function, gene expression (i.e. the amount of mRNA transcribed from DNA) may hold the key to a fast, reliable biomarker of immune response following an infection or a vaccine. Gene expression is a dynamic biological process whose functioning and variation can be related to numerous diseases and phenotypes, and transcriptomics data has been used to clarify immune response mechanisms in previous vaccine studies ([Querec et al., 2009b](#); [Obermoser et al., 2013a](#)). The fundamentality of gene expression in cellular processes makes it uniquely valuable for early immune response assessment ([Kennedy and Poland, 2011](#); [Oberg et al., 2015](#); [Rao et al., 2019](#); [Weiner et al., 2019](#)). Changes in gene expression could precede more traditional measures of immune function (like antibody production



or T-cell activity) by weeks or months.

### 4.2.1 Evaluation of high-dimensional surrogate markers for a binary treatment

There are numerous methods for evaluating the strength of surrogate markers in the context of a single univariate surrogate marker measured in the course of a randomized clinical study. However, quantifying the utility of surrogate markers when the dimension of the surrogate grows remains challenging.

The estimator in [Freedman et al. \(1992\)](#), itself based on the pioneering work of [Prentice \(1989\)](#), allows to draw a connection between quantifying the utility of a surrogate marker and the most fundamental tools of causal inference: namely, methods for estimating the average treatment effect. Despite the extensive links made between mediation and surrogacy ([Taylor et al., 2005](#); [Joffe and Greene, 2009](#)), this connection had remained absent from the surrogate marker literature. In [Agniel et al. \(2023\)](#), we show how this unlocks the evaluation of a set of surrogate markers that may be high-dimensional with robust and efficient estimation of average treatment effects, thanks to state-of-the-art methods for incorporating flexible machine learning and sparse high-dimensional models.

#### Notations

Let  $A$  denote a binary treatment, and let the primary outcome of the study be  $Y$ . Let there be a vector  $\mathbf{S}$  of potential surrogate information, and let  $\mathbf{X}$  be a vector of pre-treatment covariates. The primary quantity of interest is the treatment effect on the outcome:

$$\Delta = \mathbb{E}\{Y^{(1)} - Y^{(0)}\}, \quad (19)$$

where  $Y^{(a)}$  is the potential or counterfactual outcome that would have been observed if treatment were  $A = a$ , possibly contrary to fact. Similarly let  $\mathbf{S}^{(a)}$  be the potential/counterfactual value the vector of surrogates would take if  $A = a$ . Let the data observed in the current study be  $n$  i.i.d. realizations  $(\mathbf{X}_i, A_i, \mathbf{S}_i, Y_i)_{i=1, \dots, n}$ .

To evaluate a surrogate’s usefulness, one can use the Proportion of Treatment effect Explained (PTE) by  $\mathbf{S}$  that is defined as:

$$R_{\mathbf{S}} = (\Delta - \Delta_{\mathbf{S}})/\Delta = 1 - \Delta_{\mathbf{S}}/\Delta, \quad (20)$$

with  $\Delta_{\mathbf{S}}$  the residual treatment effect, or the treatment effect that remains after controlling for the surrogate information. While many measures have been proposed for this purpose (see [Parast et al., 2016](#), for a recent overview), the PTE has been used frequently since [Prentice \(1989\)](#), notably for its ease of interpretation. In particular, if  $(\mathbf{S}^{(0)}, \mathbf{S}^{(1)})$  are independent of  $(Y^{(0)}, Y^{(1)})$  conditionally on

$\mathbf{X}$ , then  $\Delta_{\mathbf{S}} = \Delta$  and  $R_{\mathbf{S}} = 0$ . In contrast, if all of the treatment effect can be attributed to  $\mathbf{S}$ , then  $\Delta_{\mathbf{S}} = 0$  and  $R_{\mathbf{S}} = 1$ .

$\Delta_{\mathbf{S}}$  is usually defined in terms of a particular reference distribution, e.g. taking  $\Delta_{\mathbf{S}} = \mathbb{E} \{ \psi_1(\mathbf{S}) - \psi_0(\mathbf{S}) | A = 1 \}$ , to be the residual treatment effect among the treated group, with  $\psi_a$  a surrogate transformation (Price et al., 2018). This choice of reference distribution is often arbitrary. Instead, this choice can be avoided by defining:  $\Delta_{\mathbf{S}}$  the average treatment effect conditional on the distributions of  $\mathbf{S}^{(0)}$  and  $\mathbf{S}^{(1)}$  both being equal to the distribution of  $\mathbf{S}$  (all conditionally on  $\mathbf{X}$ ):

$$\Delta_{\mathbf{S}} = \mathbb{E} \{ \psi_1(\mathbf{X}, \mathbf{S}) - \psi_0(\mathbf{X}, \mathbf{S}) \}. \quad (21)$$

This formulation has deep connections to mediation analysis (without requiring some of their usual restrictive assumptions), i.e.  $\Delta_{\mathbf{S}}$  is a function of conditional natural direct effects and  $\Delta - \Delta_{\mathbf{S}}$  is a function of conditional natural indirect effects (Joffe and Greene, 2009; VanderWeele, 2013), when the assumptions for identifying those effects are met.

Using causal inference tools does not require all of those assumptions necessary for mediation – importantly, here there is no requirement that all confounders of the surrogate-outcome relationship are measured and included in the study – because the aims of mediation and surrogate marker evaluation are different (VanderWeele, 2013). The aim of identifying a mediator is determining whether the effect of treatment operates through the mediator itself, e.g., through some biological pathway. Often, a good surrogate marker is similarly conceptualized as a variable through which the treatment operates, but this is not necessarily required; a variable can be a good surrogate if it captures the treatment effect on the outcome, even if the treatment effect does not operate through the variable itself (sometimes called a non-mechanistic correlate of protection, see Plotkin and Gilbert 2012).

### Assumptions

First,  $\Delta$  must be different from 0, otherwise the goals of identifying surrogate markers are practically and theoretically not meaningful. The three typical assumptions of treatment effect estimation are also needed: consistency, positivity, and no unmeasured confounding. Specifically, the observed values of  $\mathbf{S}$  and  $Y$  when  $A = a$  are assumed to be identical to their counterfactuals  $\mathbf{S}^{(a)}$  and  $Y^{(a)}$  such that:

$$\mathbf{S} = \mathbf{S}^{(1)}A + \mathbf{S}^{(0)}(1 - A) \text{ and } Y = Y^{(1)}A + Y^{(0)}(1 - A). \quad (22)$$

Furthermore  $\mathbf{X}$  is assumed to contain all confounders of the effects of  $A$  on  $\mathbf{S}$  and  $Y$  (i.e. treatment  $A$  is as good as randomized, conditional on the covariates  $\mathbf{X}$ ):

$$\{ \mathbf{S}^{(0)}, \mathbf{S}^{(1)}, Y^{(0)}, Y^{(1)} \} \perp\!\!\!\perp A \mid \mathbf{X}. \quad (23)$$

In addition, two forms of positivity ensure that individuals in the two study arms are not too different from one another. First, the usual positive probability of receiving either treatment for some  $\epsilon_1 > 0$ ,  $\mathbb{P}\{\epsilon_1 < e_1(\mathbf{X}) < 1 - \epsilon_1\} = 1$ , and a related assumption that further conditions on the surrogates,  $\mathbb{P}\{\epsilon_2 < \pi_1(\mathbf{X}, \mathbf{S}) < 1 - \epsilon_2\} = 1$ , for some  $\epsilon_2 > 0$  where  $e_1(\mathbf{S}, \mathbf{X}) = \mathbb{P}(A = 1 | \mathbf{X})$  and  $\pi_1(\mathbf{S}, \mathbf{X}) = \mathbb{P}(A = 1 | \mathbf{S}, \mathbf{X})$ . Notably these two positivity conditions ensure that the conditional distribution of the counterfactual surrogates under treatment and control cannot be too different from one another, i.e., ensuring overlap between both. When the treatment has a large effect on  $\mathbf{S}$ , this additional overlap requirement may be suspect – e.g., there may be some values of  $\mathbf{s}$  such that the density function  $f_{\mathbf{S}^{(1)}}(\mathbf{s} | \mathbf{X} = \mathbf{x})$  approaches 0 and thus  $\pi_1(\mathbf{x}, \mathbf{s})$  also approaches 0.

Interpretation of  $R_{\mathbf{S}}$  as the PTE depends on it actually being a proportion, lying between 0 and 1. The two following conditions together ensure that  $0 \leq R_{\mathbf{S}} \leq 1$  (assuming without loss of generality that  $\Delta > 0$ ). First, to ensure that  $\Delta \geq \Delta_{\mathbf{S}}$ :

$$\int \{e_1(\mathbf{x})\psi_1(\mathbf{x}, \mathbf{s}) + e_0(\mathbf{x})\psi_0(\mathbf{x}, \mathbf{s})\} d\{F_{\mathbf{X}, \mathbf{S}^{(1)}}(\mathbf{x}, \mathbf{s}) - F_{\mathbf{X}, \mathbf{S}^{(0)}}(\mathbf{x}, \mathbf{s})\} \geq 0 \quad (24)$$

requires that a propensity-weighted mixture of the two conditional mean functions  $e_1(\mathbf{x})\psi_1(\mathbf{x}, \mathbf{s}) + e_0(\mathbf{x})\psi_0(\mathbf{x}, \mathbf{s})$  is larger when  $\mathbf{s}$  takes values from the distribution of the counterfactual surrogates under treatment than if it took values from the distribution under control. Second, to ensure that  $\Delta_{\mathbf{S}} \geq 0$

$$\int \{\psi_1(\mathbf{x}, \mathbf{s}) - \psi_0(\mathbf{x}, \mathbf{s})\} dF_{\mathbf{X}, \mathbf{S}}(\mathbf{x}, \mathbf{s}) \geq 0 \quad (25)$$

meaning that the residual treatment effect is in the same direction as the overall treatment effect  $\Delta$ .

### Estimation & inference

$\Delta$  and  $\Delta_{\mathbf{S}}$  can actually be identified in terms of average treatment effects, one of which conditions on the surrogates ( $\Delta_{\mathbf{S}}$ ) and one which does not ( $\Delta$ ). Robust estimators of  $\Delta_{\mathbf{S}}$  and  $\Delta$  can then be derived as:

$$\widehat{\Delta}_{\mathbf{S}} = n^{-1} \sum_{i=1}^n \left[ \frac{A_i Y_i - \{A_i - \widehat{\pi}_1(\mathbf{X}_i, \mathbf{S}_i)\} \widehat{\mu}_1(\mathbf{X}_i, \mathbf{S}_i)}{\widehat{\pi}_1(\mathbf{X}_i, \mathbf{S}_i)} - \frac{(1 - A_i) Y_i - \{1 - A_i - \widehat{\pi}_0(\mathbf{X}_i, \mathbf{S}_i)\} \widehat{\mu}_0(\mathbf{X}_i, \mathbf{S}_i)}{\widehat{\pi}_0(\mathbf{X}_i, \mathbf{S}_i)} \right], \quad (26)$$

$$\widehat{\Delta} = n^{-1} \sum_{i=1}^n \left[ \frac{A_i Y_i - \{A_i - \widehat{e}_1(\mathbf{X}_i)\} \widehat{m}_1(\mathbf{X}_i)}{\widehat{e}_1(\mathbf{X}_i)} - \frac{(1 - A_i) Y_i - \{1 - A_i - \widehat{e}_0(\mathbf{X}_i)\} \widehat{m}_0(\mathbf{X}_i)}{\widehat{e}_0(\mathbf{X}_i)} \right], \quad (27)$$

because (22) implies that  $\psi_a(\mathbf{X}, \mathbf{S}^{(a)}) = \mu_a(\mathbf{X}, \mathbf{S})$ . Thus,  $R_{\mathbf{S}}$  can be estimated as:

$$\widehat{R}_{\mathbf{S}} = 1 - \widehat{\Delta}_{\mathbf{S}} / \widehat{\Delta}, \quad (28)$$

with  $\mu_a(\mathbf{x}, \mathbf{s}) = \mathbb{E}(Y | \mathbf{X} = \mathbf{x}, A = a, \mathbf{S} = \mathbf{s})$ ,  $m_a(\mathbf{X}) = \mathbb{E}(Y | \mathbf{X} = \mathbf{x}, A = a)$ ,  $e_a(\mathbf{x}) = \mathbb{P}(A = a | \mathbf{X})$ , and where their estimation, along with those of  $\pi_a(\mathbf{X}, \mathbf{S})$ , can be quite general. In practice, these can be estimated using the *Super Learner* (Van der Laan et al., 2007) which finds an optimal combination of a set of candidate models or learners. Additionally, a sample-splitting scheme avoids placing restrictive conditions on the estimation of the nuisance functions (Chernozhukov et al., 2017).

Under these settings,  $\widehat{R}_{\mathbf{S}}$  will converge at the parametric  $n^{-\frac{1}{2}}$  rate and be asymptotically normal, as long as  $\widehat{\mu}_a$ ,  $\widehat{\pi}_a$ ,  $\widehat{m}_a$ , and  $\widehat{e}_a$  converge fast enough. While this is the case in many settings in low dimensions, this convergence in high-dimensions is more difficult to ensure without further restrictions. If only  $\mathbf{S}$  is high-dimensional, then a typical approach is to specify as a sparse linear model for  $\mu_a$  and a sparse logistic regression model for  $\pi_a$ . This is sufficient so long as  $\mathbf{X}$  is low-dimensional and  $e_1(\mathbf{x})$  may be estimated nonparametrically (i.e., without being restricted to the class of sparse linear models). However, if  $\mathbf{X}$  is also high-dimensional, sparse logistic models for  $e_1(\mathbf{x})$  and  $\pi_1(\mathbf{x}, \mathbf{s})$  may not in general be compatible with one another because of the non-collapsibility of logistic regression (Guo and Geng, 1995) – unless  $\mathbf{S} \perp\!\!\!\perp A | \mathbf{X}$  (which would imply  $R_{\mathbf{S}} = 0$ ) or  $\mathbf{S} \perp\!\!\!\perp \mathbf{X} | A$  (which would imply that  $\mathbf{X}$  does not confound the relationship between  $A$  and  $\mathbf{S}$ ). Simulation results in Agniel et al. (2023) suggest that an ensemble approach with the *Super Learner* can still provide good performances in such cases.

### 4.2.2 Evaluation of gene expression as a surrogate marker for antibody response to Ebola infection in an observational study

The general approach we proposed in Agniel et al. (2023) to evaluate surrogate markers can be applied in randomized experiments or in observational studies, and can be used regardless of the dimensionality of the surrogates. It is robust in that the PTE of the surrogates has been defined without reference to any models, and machine learning approaches like *Super Learner* can be used to very flexibly

estimate nuisance functions. Thus, it can be applied to the evaluation of gene expression as a surrogate marker.

The concentration of binding antibodies is often used as the primary outcome of interest in studies of Ebola vaccine efficacy, being itself a surrogate of vaccine efficacy as measured by the effect on the incidence of infections (Rozenendaal et al., 2020). Because gene expression is the means by which DNA is turned into RNA and eventually proteins, it is associated to cellular function. Thus, the establishment of the humoral immune response may be captured by changes in gene expression as suggested by early works on systems vaccinology (Li et al., 2014). Furthermore, gene expression changes may occur days or even weeks before traditional measures of immune function (Rechtien et al., 2017). Genome-wide expression data offer the opportunity to look at various pathways which constitutes potential surrogate markers. In this study, observational data on long-term Ebola survivors and healthy controls shed light on the possibility of gene expression’s use as a surrogate for antibody response to Ebola virus, inspired by the study of potential surrogates of protection among Ebola disease survivors (Sullivan et al., 2009).

In total, 26 Ebola survivors of the 2013-2016 Ebola outbreak in West Africa were recruited from the Postebogui cohort (Etard et al., 2017) as well as 33 healthy donors as described in Wiedemann et al. (2020), each of whom had expression for 29,624 genes quantified from whole blood RNA-seq (publicly available from the Gene Expression Omnibus repository with accession code GSE143549) as well as the measured concentration of Immunoglobulin G antibodies specific to Ebola nucleoprotein. Figures 9 and 10 represent the gene expression data and the Antibodies distribution respectively. Clearly, this is a setting where the number of potential surrogate markers (the genes) is substantially larger than the sample size. Propensity and surrogate scores ( $\hat{e}$  and  $\hat{\pi}$ ) were truncated at 0.05 and 0.95 to prevent instability due to extreme weights, and  $\mathbf{X}$  included age and sex. Candidate learners included for  $\mu_1, \mu_0, m_1, m_0$  were the lasso, ridge regression, ordinary least squares, support vector machines, and random forests, and for  $\pi$  and  $e$  were the lasso, logistic regression, linear discriminant analysis, quadratic discriminant analysis, support vector machines, and random forests.

Ebola survivors were estimated to have a much higher abundance of Ebola-specific antibodies ( $\hat{\Delta} = 3,998$ ,  $SE = 851.5$ ). The residual treatment effect was estimated at  $\hat{\Delta}_{\mathcal{S}} = 3,242$  with  $SE = 727.8$ , and the proportion of the difference explained by gene expression was estimated at  $\hat{R}_{\mathcal{S}} = 0.1890$ , with a  $SE$  of 0.07923. Thus, a large part of the humoral immune response cannot be explained by the differences in gene expression. Of note, this assumes no unmeasured confounding factors, although it cannot be guaranteed in such a real-life context where survivors and healthy volunteers are two selected populations. If unmeasured confounding

#### 4 CONNECTIONS WITH EHR AND SURROGATE MARKER EVALUATION

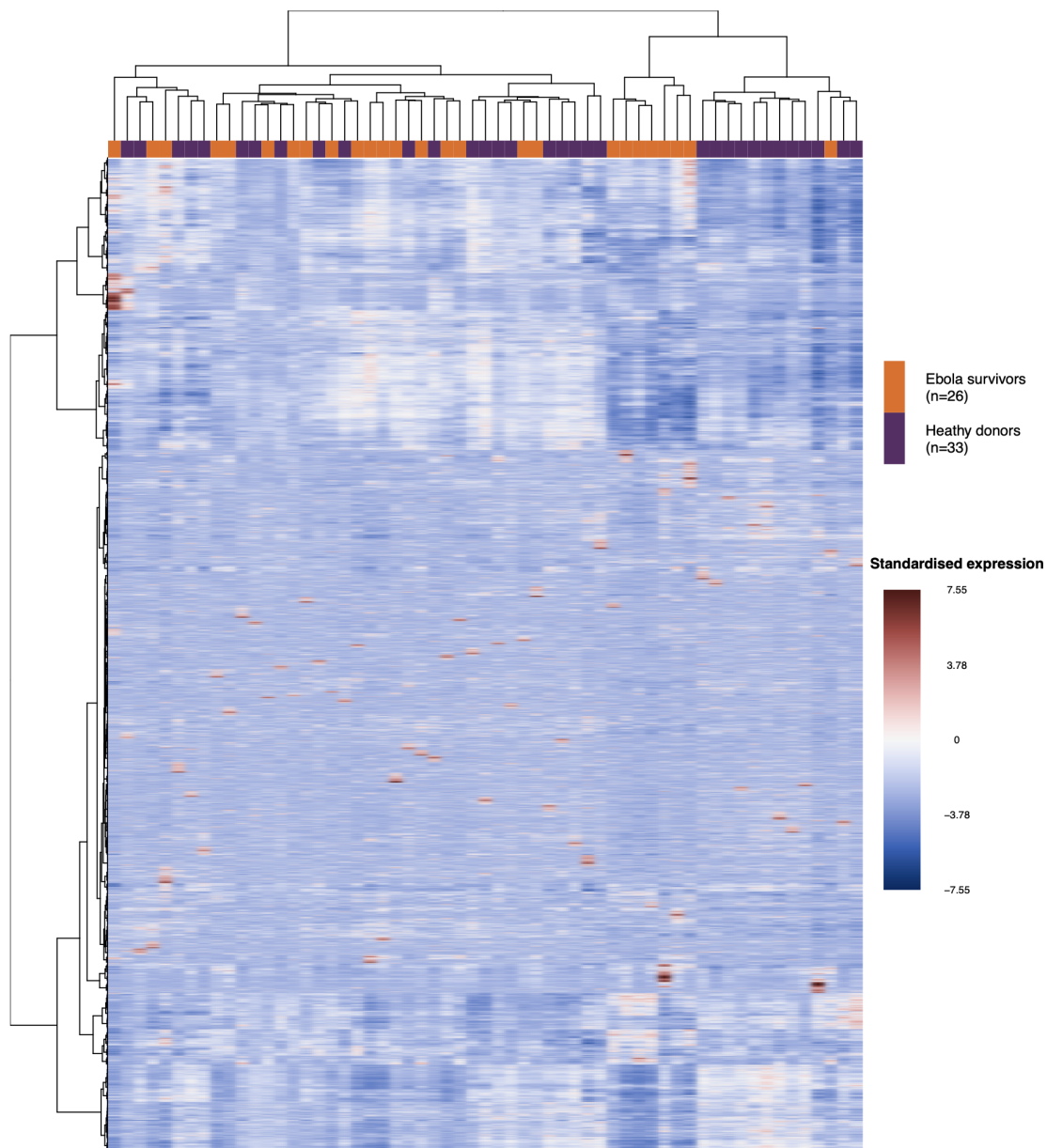


Figure 9: **Heatmap of the scaled expression of the 29,624 aligned genes from RNA-seq measurement.** See GSE143549 study on Gene Expression Omnibus repository for more information. The blurry impression is due to the necessary rasterization of several matrix elements into only one pixel, given that there are not enough pixels available on most screens to display as many elements.

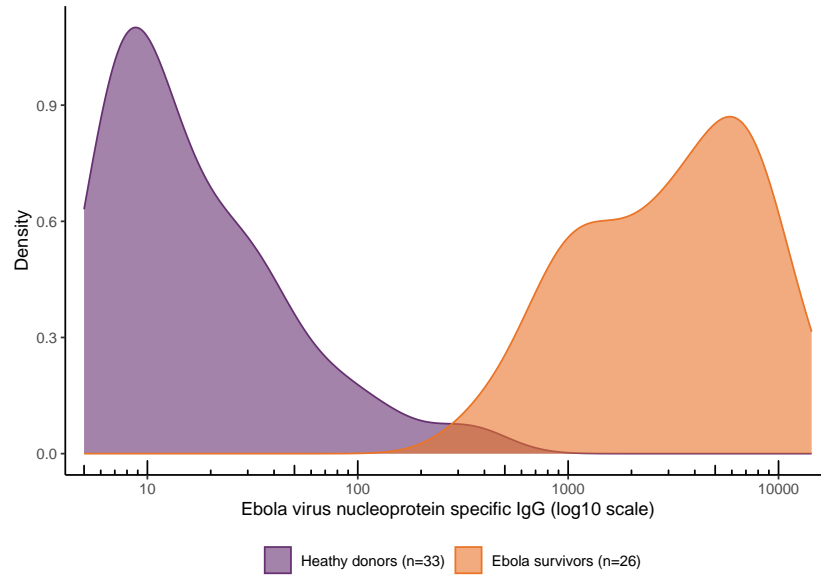


Figure 10: **Scaled probability density estimation of antibody measurements for both Ebola survivors and healthy donors.** Overlap between the two groups is minimal.

inflated both  $\hat{\Delta}$  and  $\hat{\Delta}_{\mathbf{S}}$  roughly equally, this could have the effect of artificially deflating  $\hat{R}_{\mathbf{S}}$ . Another explanation for the low estimated  $\hat{R}_{\mathbf{S}}$  could be that, while gene expression measured shortly after infection may potentially be a good surrogate, measuring it long after infection (as in this study), does not capture the treatment effect as well. Measurement error in  $\mathbf{S}$  could also deflate the PTE. Importantly, one other potential violation of the assumptions is that about one third of the observations have truncated surrogate scores (i.e.  $\hat{\pi}_{-k}(\mathbf{X}_i, \mathbf{S}_i) < 0.05$  or  $\hat{\pi}_{-k}(\mathbf{X}_i, \mathbf{S}_i) > 0.95$ ). This suggests that positivity might be (nearly) violated.

# 5 Software dissemination and reproducible research in biostatistics

---

The main content of this chapter has been previously published in the following:

- Desquilbet L, Granger S, Hejblum BP, Legrand A, Pernot P, and Rougier N. *Vers une recherche reproductible : Faire évoluer ses pratiques*. Urfist de Bordeaux, 2019. ISBN 979-10-97595-05-0. URL <https://rr-france.github.io/bookrr/>.
  - Hejblum BP, Kunzmann K, Lavagnini E, Hutchinson A, Robertson DS, Jones SC, and Eckes-Shephard AH. Realistic and robust reproducible research for biostatistics. *Preprints*, 2020060002, 2020. DOI: [10.20944/preprints202006.0002.v1](https://doi.org/10.20944/preprints202006.0002.v1).
- Hejblum BP, Ba K, Thiébaud R, and Agniel D. Neglecting normalization impact in semi-synthetic RNA-Seq data simulation generates artificial false positives. *bioRxiv*, page 2022.05.10.490529, 2022. DOI: [10.1101/2022.05.10.490529](https://doi.org/10.1101/2022.05.10.490529).

---

## 5.1 Reproducible research

### 5.1.1 Definitions

The majority of the scientific community has an understanding of what “reproducible research” means for its own field. Yet it is hard to provide a universal definition for all disciplines, in part because the very notion of “result” is highly dependent on the research field ([Desquilbet et al., 2019](#)). For some, it is about confirming the significance of an effect; for others, it is a matter of obtaining the exact same numerical result to the exact bit. [Vandewalle et al. \(2009\)](#) provide the following definition:



“A research work is called *reproducible* if all information relevant to the work, including, but not limited to, text, data and code, is made available, such that an independent researcher can reproduce the results.”

I adhere to this definition, which is broad enough to cover many situations, and is coherent with the definition from [The Turing Way Community \(2022\)](#) articulated in Figure 11.

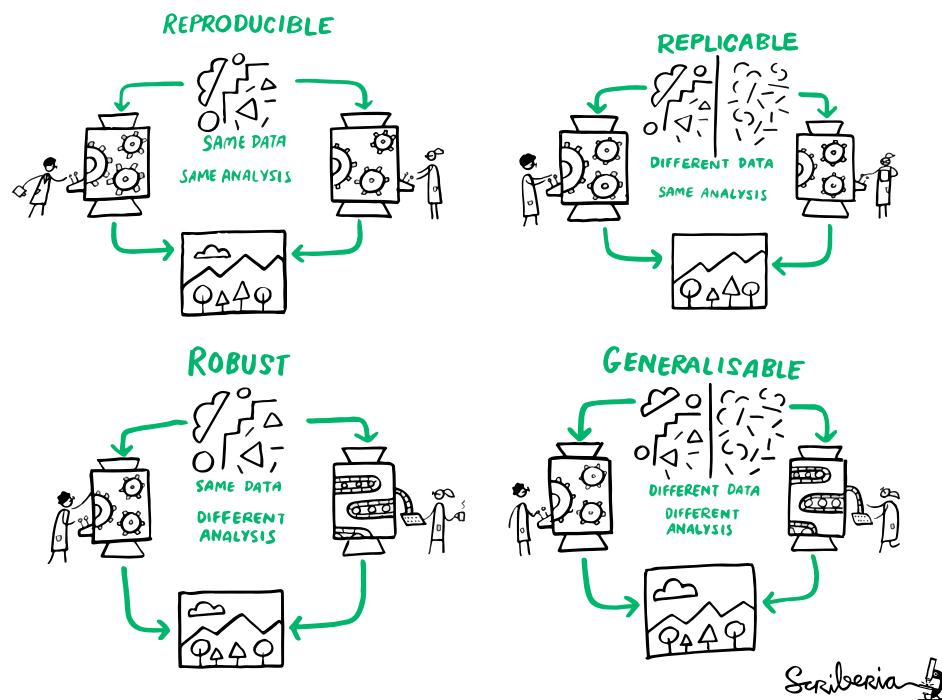


Figure 11: **Reproducible, robust, replicable, and generalizable research defined in terms of same/different code and same/new data;** CC-BY [The Turing Way Community and Scriberia \(2019\)](#).

The ultimate goal of science is to provide *generalizable* results to advance knowledge. Evidence-based scientific methods in (life) sciences are eventually about convincing colleagues, decision makers, and the public of scientific facts by reasoning based on empirical observations and analyses. Data collection and data analysis have become an integral part of this, and the ability to reliably reproduce results from data, i.e. to make an argument verifiable, is therefore essential to one’s credibility in the scientific debate. The first step toward generalization is thus to make the primary analysis *reproducible*.

### 5.1.2 Open science & biostatistics specificities

In biostatistics, reproducibility has been a long standing concern. Pharmaceutical companies currently have to file hundreds of pages documenting the various clinical evidence in favour of a new drug or treatment (often trials of various phases) to the Food and Drug Administration (FDA) or the European Medicines Agency (EMA) for instance. With the advent of modern tools for data management, version control, archiving and reporting, reproducibility and transparency of analyses have never been so accessible for biostatisticians. However, the very reasons that lead to the emergence of biostatistics as its own field warrants the need to adapt these tools to the specifics of biostatistics.

Reproducible research is not the same as Open Science, and being open is not the same thing as being reproducible – even though openness facilitates external reproducibility. This distinction is particularly important in the context of biostatistics and the biomedical research. The Open Science movement propagates the spread of knowledge through the use of “digital technologies and new forms of collaboration” (European Commission, 2018). The Organisation for Economic Co-operation and Development (OECD) definition of Open Science is even more concrete and explicitly references research data management as well:

“[Open Science refers to the efforts] *to make the primary outputs of publicly funded research results – publications and the research data – publicly accessible in digital format with no or minimal restriction.*” (OECD, 2015)

For example, preprint services such as arXiv.org, bioRxiv.org or medRxiv.org serve this goal. It should be noted that the OECD definition of Open Science does account for the fact that some research data may be subject to restrictions on sharing but that researchers are required to keep access restrictions to a minimum. This is particularly important in biomedical research where personal health-care and biological data will frequently be subject to privacy regulations such as General Data Protection Regulation (GDPR European Commission, 2018) or Health Insurance Portability and Accountability Act (HIPAA Public Law 104-191, 1996). Even research based on data that are not open can be reproducible: efforts should be made to properly separate research data from both the definition of the required computing environment and the analysis code, so that those can be shared publicly. Furthermore, upon obtaining access to the source data via a well-defined process, the results may still be reproduced, in line with Vandewalle et al. (2009) definition.



Whilst there is little reason not to release the source code used for a data analysis funded by public research agencies (regardless of the programming language used), ethical and privacy considerations generally prevent a full open data strategy (European Commission, 2018). Fortunately, although open-source software

and open data can be important ingredients in reproducible research, these are not mandatory. In some specific contexts, for instance in transcriptomics data (which have a long tradition of openness through the Gene Expression Omnibus<sup>1</sup> data archive), part of the data can be shared publicly. In such cases, the data should be provided in a non-proprietary and non-binary format to ensure cross-platform readability and posterity on a persistent archive. In most cases, however, data can only be shared privately or not shared at all. This is due to study participant consent which usually restrains further re-use of health data due to their particular personal and sensitive nature. Exceptions to this are if appropriate participant consent was sought and obtained before and during data collection for future research, data re-use and data sharing.

While data is an important component in a reproducible research workflow, it is not the only one. The current complexity of analysis pipelines in biomedical sciences, in particular for high-throughput measurement data, poses a severe challenge for the transparency and reproducibility of results. Researchers are increasingly incorporating new software into their analyses, but in a quickly evolving scientific landscape where resources for software support and maintenance are not a priority, those tools can rapidly change or even worse get deprecated.

### Code sharing, software dissemination & personal practices

As a biostatistician, I produce code rather than data. Contemporary data analysis often involves interrelated code, data sets and output files. Research compendia are tools that facilitates reproducible research by bringing together in a single virtual "place" the data, codes, protocols and documentation associated within a single research project. The simplest way to build a research compendium is to create a directory associated with the project, with sub-directories into which the objects are distributed. An explicit naming convention for objects and directories can greatly facilitate reusability. Marwick et al. (2018) propose different structures of increasing complexity depending on the scope and ambition of a data analysis.

For more methodological developments, code can actually represents the main research output. An excellent way of ensuring reproducibility of statistical method research is to provide re-usable computer code (Boulesteix et al., 2020). This is easier said than done, but fortunately, the R ecosystem provides an excellent way of doing so:  packages.  packages can be seen as a particular type of research compendium, that are especially suited to share and disseminate statistical method implementations. Bioconductor, CRAN and GitHub<sup>2</sup> are three common ways of distributing such packages, with decreasing requirements to ensure portability, ease of installation and good coding practices such as documentation (GitHub having

<sup>1</sup>GEO: <https://www.ncbi.nlm.nih.gov/geo/>

<sup>2</sup><https://github.com/>

virtually no requirements, while CRAN has already many, and Bioconductor even more so). Most of my methodological developments are available as software packages in the [R](#) language available from the CRAN or Bioconductor – with a few only available through GitHub (see the corresponding CV section [Software development & maintenance](#) page 81 for a full list).

While all the different aspects involved in reproducible research can feel overwhelming, it is important to keep in mind that reproducibility is a continuum, an asymptotic ideal to yearn for. Reproducibility is not all or nothing, and every step along the way is a step in the right direction. Even though the efforts towards reproducible research can appear time consuming, this time is well spent in light of the transparency owed to the public in the case of publicly funded research. In addition, the first person likely to seek to reproduce one’s research is your future self (either for a manuscript revision before resubmission, or for benchmarking a new approach against the previous one). So once again, time spent on making one’s research more reproducible is worth it.

## 5.2 Personal stories about reproducing other’s research

Of note, reproducible research is not a guarantee of research quality, but only of transparency. While transparency contributes to quality, bad research may be reproducible just as well. I will now elaborate on two instances where reproducible practices from others have benefited my research, and helped answer scientific questions by studying in more details results previously published by [Steinbach et al. \(2004\)](#) and [Li et al. \(2022\)](#) respectively.

### 5.2.1 Identifying unreported confounding batch effects

Like a number of high-throughput biotechnologies, gene expression measurements by microarray and by RNA-seq suffer from a high degree of sensitivity to experimental conditions, which can lead to considerable variability in measurements irrespective of the biological questions of interest. In some cases, this leads to the appearance of technical biases in the data, otherwise known as the “batch effect” ([Leek et al., 2010](#)). Gene expression is usually measured from venous blood samples in 5ml tubes. Recently, studies have succeeded in using very small volumes of finger-prick blood to sequence whole blood RNA ([Rinchai et al., 2022](#)). [Stein et al. \(2016\)](#) reanalyzed data from [Obermoser et al. \(2013b\)](#) to compare gene expression measured either in venous blood or finger-prick in an influenza and pneumococcal vaccination study. Ignoring the batch effects documented in this data set, they

found significant differences between the sampling methods. These observations contradict other findings in which the two technologies were also compared.

In total, 17 subjects had measurements from both venous blood and finger-prick. In total, those 17 subjects had 511 samples, each with 48,803 probes measuring gene expression, ranging from -7 to +28 days relative to vaccination. Figure 12 summarizes the various technical, clinical and demographics information available for each sample. This immediately highlights the entanglement between the technical variable encoding for the different flow cells and the binary variable of interest indicating which samples were extracted from finger-prick measurements and which were from venous blood.

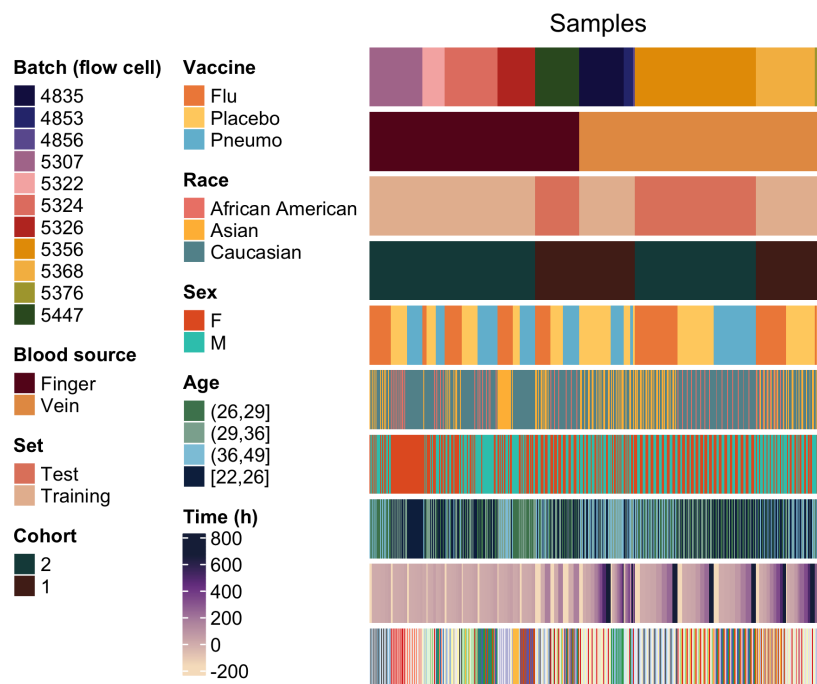


Figure 12: **Clinical, technical and demographic information available for the 511 samples from the 17 individual with both venous blood and finger-prick measurements.** The bottom line colors the 511 samples according to the subject they belong to, for which no legend is shown.

This potential confounding bias is further confirmed through multivariate descriptive analysis of the gene expression data. The raw microarray data<sup>3</sup> are normalized with a Norm-Exp background correction followed by quantile normaliza-

<sup>3</sup>GEO identifier [GSE48762](#)

tion and a  $\log_2$  transformation, following [Ritchie et al. \(2007\)](#). Figure 13 displays the first two principal components from a PCA. The batch effect of the different flow cells is clearly visible.

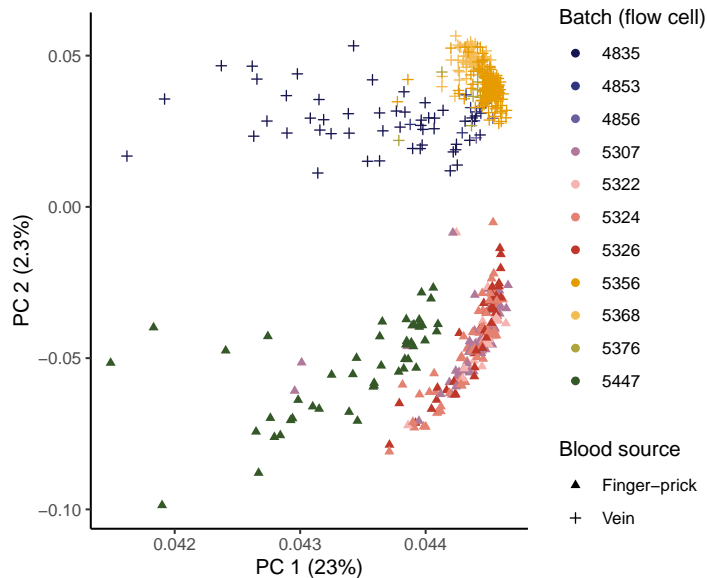


Figure 13: **First factorial plan of the normalized gene expression data.**

A usual solution when faced with such large batch effects is to apply a batch effect correction before proceeding to the DEA, such as ComBat ([Johnson et al., 2007](#)). But here, we are faced with a choice: either to adjust on the variable of interest (namely the blood source, being either finger-prick or Venous blood) or not to adjust on it. But due to the confounding between the potential batch effect of the different flow cells and this variable, there is no right decision. Adjusting slightly exacerbates the difference between the two blood sources, although this could be due only to the original batch effect between the different flow cells. On the contrary, not adjusting completely erase any difference there might be between the two blood sources. Thus, this experiment does not contain the necessary evidence to tackle the question of differential gene expression between finger-prick and venous blood. By ignoring these potential batch effects, [Stein et al. \(2016\)](#) are drawing conclusions based on questionable results that could well be deceptive.

### 5.2.2 Semi-synthetic data and confounding bias

Benchmarking different statistical methods is an arduous task ([Weber et al., 2019](#)). While simulations are a necessary tool when it comes to studying methods per-

formance and limits, and validating implementations (Morris et al., 2019), benchmarking tackles a different question. It requires comparisons based in real-world settings (akin to *in vivo* experiments to take on the analogy from Boulesteix et al., 2020). Unfortunately, the truth is rarely known in real data, especially in biomedical science. Instead, some compromise can be reached with so-called “semi-synthetic” or “realistic” simulations, i.e. simulations where additional care is given to represent real data as closely as possible while still controlling the absolute truth, either by using simulation parameters estimated on real data or by even directly starting from real data and adding some noise or perturbations (Van Mechelen et al., 2023).

Li et al. (2022) recently raised significant concerns regarding popular RNA-seq DEA methods, namely `edgeR` (Robinson et al., 2010) and `DESeq2` (Love et al., 2014), in the context of large human population sample sizes. I share those concerns, having come to similar conclusions before (Gauthier et al., 2020), as have others (Burden et al., 2014; Rocke et al., 2015). However, their findings that other methods (namely `dearseq`, `limma-voom` by Law et al., 2014, and `NOISeq` by Tarazona et al., 2015) also have increased false positive rates does not appear to be correct, and the evidence does not support their claim that the Wilcoxon rank-sum test should be preferred to these alternatives. Using the same semi-synthetic data sets as Li et al. (2022) can show that no method – including Wilcoxon test – is able to maintain the nominal level of “false discoveries” according to their definition. That is because the semi-synthetic data used for their analysis were not truly generated under  $H_0$ . Instead, the permutation scheme by which they generated the semi-synthetic data sets should be amended to actually support analysis of false positive rates under  $H_0$ . Using this amended scheme, `dearseq` outperforms other methods under these specific settings of large human population samples, and otherwise offers competitive performance, on par with the other methods.

By accessing code and data shared publicly by Li and Ge (2022), I could reproduce their Figure 2A where the empirical (“actual”) FDR is plotted against the nominal (“claimed”) FDR using semi-synthetic data generated from the full *GTEX Heart atrial appendage* ( $n=372$ ) VS *Heart left ventricle* ( $n=386$ ) original data set (with  $p=56,200$  transcripts). I also identified a discrepancy between the data they used for the Wilcoxon test and the data used for the other methods: all methods – except the Wilcoxon test – embed a normalization step before performing DEA, as is standard for RNA-seq data (Evans et al., 2018). But when the Wilcoxon test is performed on the same normalized data (following the `edgeR` pipeline for filtering out genes with low counts and using log2-counts per million transformation) as all other evaluated methods, it also appears to exaggerate the FDR – as the other methods. Figure 14 is an amended version of the Figure 2A from Li et al. (2022).

This apparent increase in FDR is thus not imputable to the methods, at least

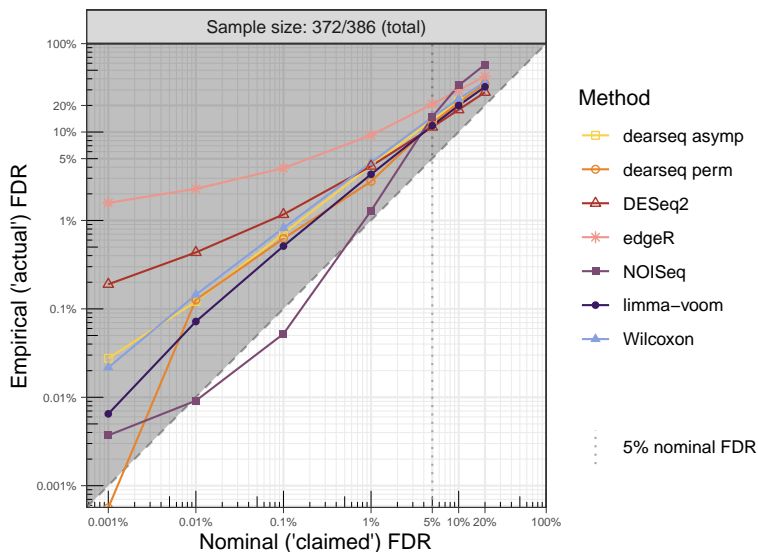


Figure 14: **Empirical FDR control against Nominal FDR level.** Average over 50 semi-synthetic data set generated from the *GTEX Heart atrial appendage VS Heart left ventricle* data. 50% of the truly DE genes are randomly sampled in each semi-synthetic data set (i.e. 2,889 genes) and remains non-permuted as true positives gold-standard. Reproduces Figure 2A from Li et al. (2022) when all methods are applied to first permuted and then normalized full data (372 and 386 samples in each group respectively).

not entirely since normalization obviously had something to do with it. Rather, I noticed it comes from an inappropriate data-generation scheme. In Figure 15 studies the impact of both the sample size and the respective order between the data normalization and the random permutations to generate non-differentially expressed genes on the FDR control, comparing the Wilcoxon test and both asymptotic and permutation tests from `dearseq` (in their discussion, Li et al. advocate for permutation analysis, fortunately `dearseq` already features such a permutation approach which was added to the comparison<sup>4</sup>). In these semi-synthetic simulated data sets, gene expression under  $H_0$  was generated by randomly swapping expression values between samples. However, Li et al. (2022) did not analyze these data directly, but instead normalized them before analysis. The top panel of Figure 15 shows how their permutation scheme leads to an apparent increase in FDR because the expression is no longer generated from  $H_0$  after normalization (e.g.

<sup>4</sup>Of note, when applied to non-normalized data, the heteroskedasticity weights estimated by `dearseq` are subject to caution because observed values are then not comparable across samples.



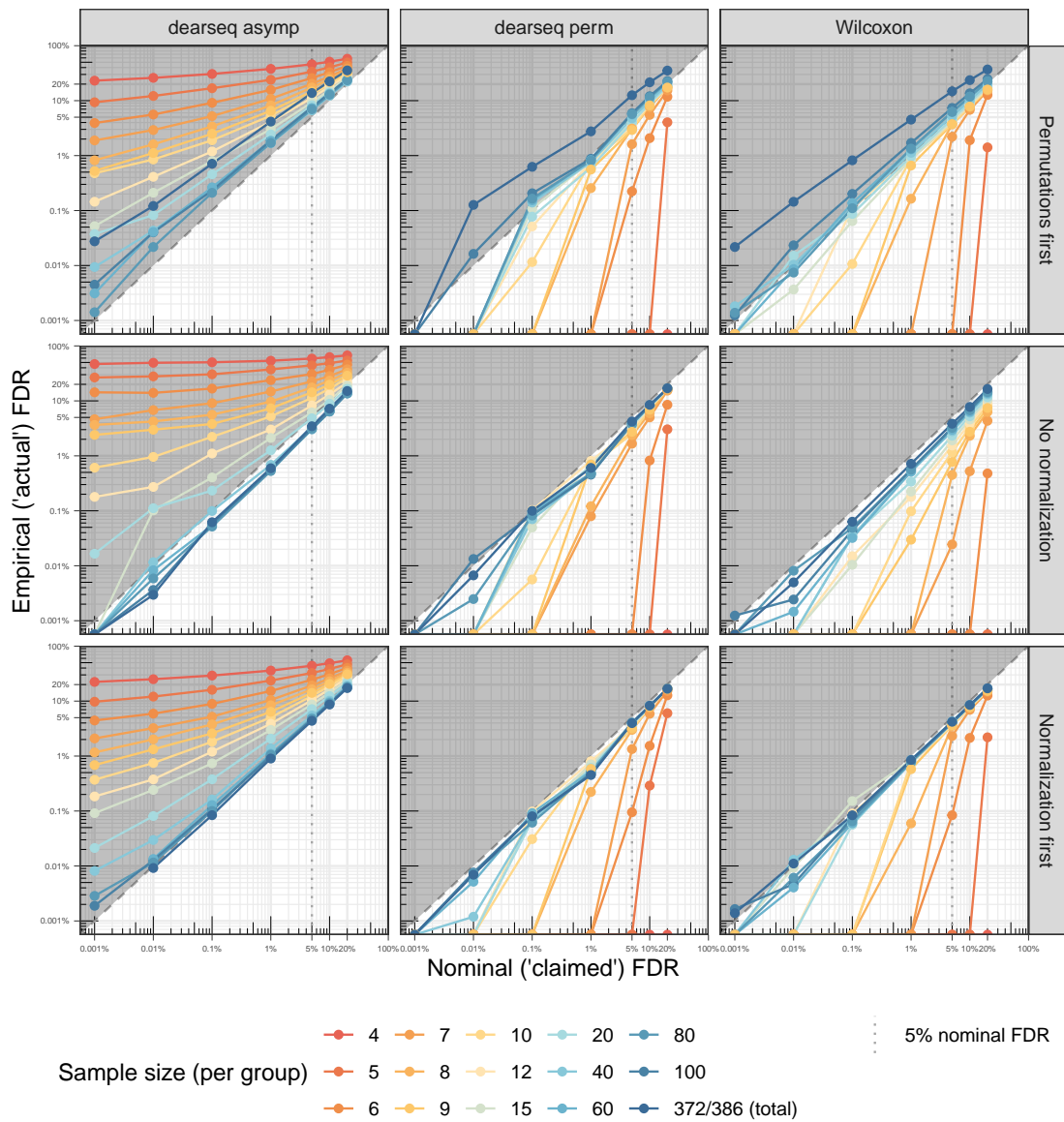


Figure 15: **Empirical FDR control against Nominal FDR level.** Average over 50 semi-synthetic data set generated from the *GTEX Heart atrial appendage VS Heart left ventricle* data. 50% of the truly DE genes are randomly sampled in each semi-synthetic data set (i.e. 2,889 genes) and remain non-permuted as true positives gold-standard. Studies the impact of both the sample size as well as the respective order between the data normalization and the random permutations.

due to a high count being swapped into a sample with a much lower library size, artificially creating a large expression post-normalization). When the data are analyzed without normalization – an approach that would never be used in practice – we show in the middle panel that both `dearseq` and the Wilcoxon test attained the nominal FDR as sample size increased.

The main source of false positives generated in the *permutation first* scheme is likely the difference in library sizes (i.e. the total sum of gene counts in a given sample). Figure 16 displays and characterizes the imbalance of library sizes between the two heart tissues from the *GTEX Heart atrial appendage VS Heart left ventricle* data set used in this example. The confounding between library size difference and the heart tissue (i.e. the condition difference of interest) is noticeable, and can explain the latter bias in the results. Contrary to when all genes are permuted (cf. the analysis presented by Li et al. (2022) in their Figure 1 where neither `dearseq` nor `limma-voom` or `NOISeq` suffer from false positive inflation), when some genes – *a fortiori* DE genes – are left non-permuted, a difference in library size between the two conditions of interest can subsist even after the permutation. In such case, this maintained library size difference will invalidate the normalization. This is exemplified in Figure 17, where this imbalance is mainly conserved in the subset of 5,778 genes that are considered as truly DE by Li et al. (2022) (the intersection of significantly DE genes according to all five methods `DESeq2`, `edgeR`, `NOISeq`, `limma-voom` and Wilcoxon test at a FDR threshold of  $10^{-6}$  on the original data). This also explains results from their supplementary where the higher the proportion of true DE genes, the more false positives are generated by this library size difference remaining after their permutation scheme.

When counts are first normalized, before being permuted under  $H_0$ , we demonstrate that all three tests in Figure 15 adequately controlled the FDR for the full data set. This amended permutation scheme should be preferred as it is fundamental to perform DEA on samples that are normalized to ensure that expression values for a given gene are comparable across samples – in particular to remove the potential effect of library size on the analysis. The null hypothesis of interest is that there is no mean difference between conditions on the data to be analyzed, i.e., the normalized data. These are the data that should be permuted, not the raw expression. Thus the original permutation scheme from Li et al. (2022) is not informative for the desired analysis. Our results indicate that the apparent false positives of `dearseq` in Li et al. (2022) are actually detecting differences in library size. Of note, `dearseq` and Wilcoxon tests both display similarly good performance in Li et al. (2022)’s Figure 1 where their permutation scheme is less problematic as all genes get permuted in that case, whereas for their Figure 2 they introduced a confounding bias from the library size by keeping the top significant genes non-permuted.

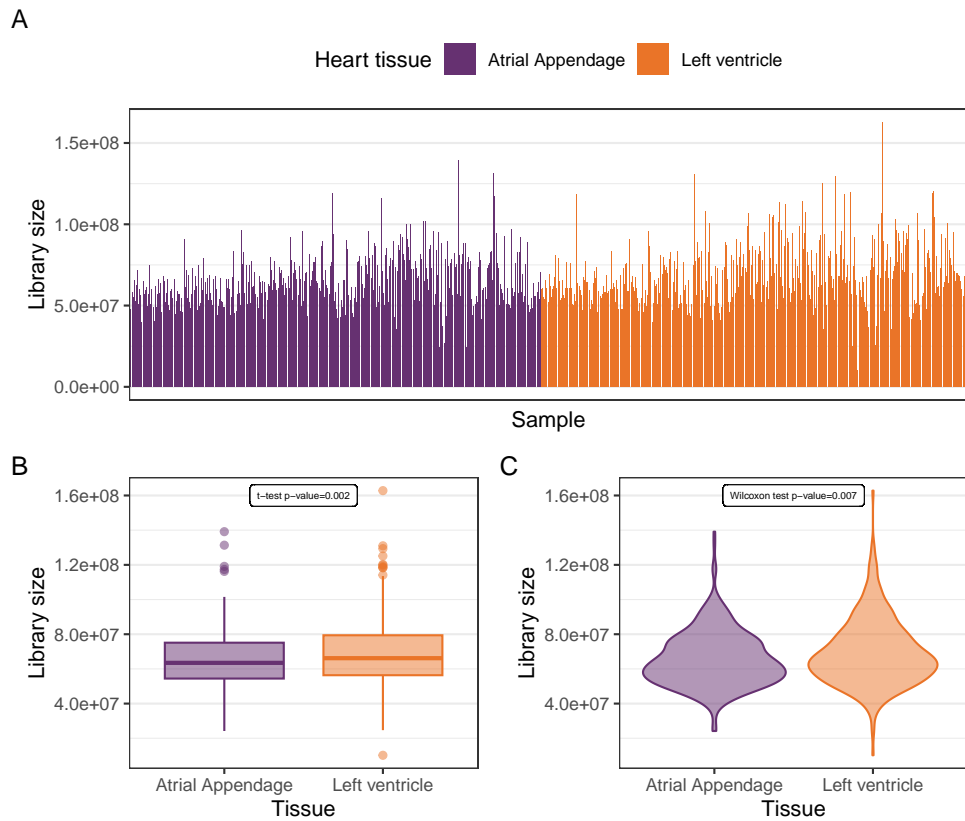


Figure 16: **Library size differences in the *GTEx Heart atrial appendage VS Heart left ventricle* data.** Panel A displays the library sizes of all 758 samples (372 and 386 in the atrial appendage and left ventricle heart tissues respectively). Panel B presents a boxplot highlighting the statistically significant difference with a t-test. Panel C presents a violin plot for a non-parametric comparison with the Wilcoxon test.

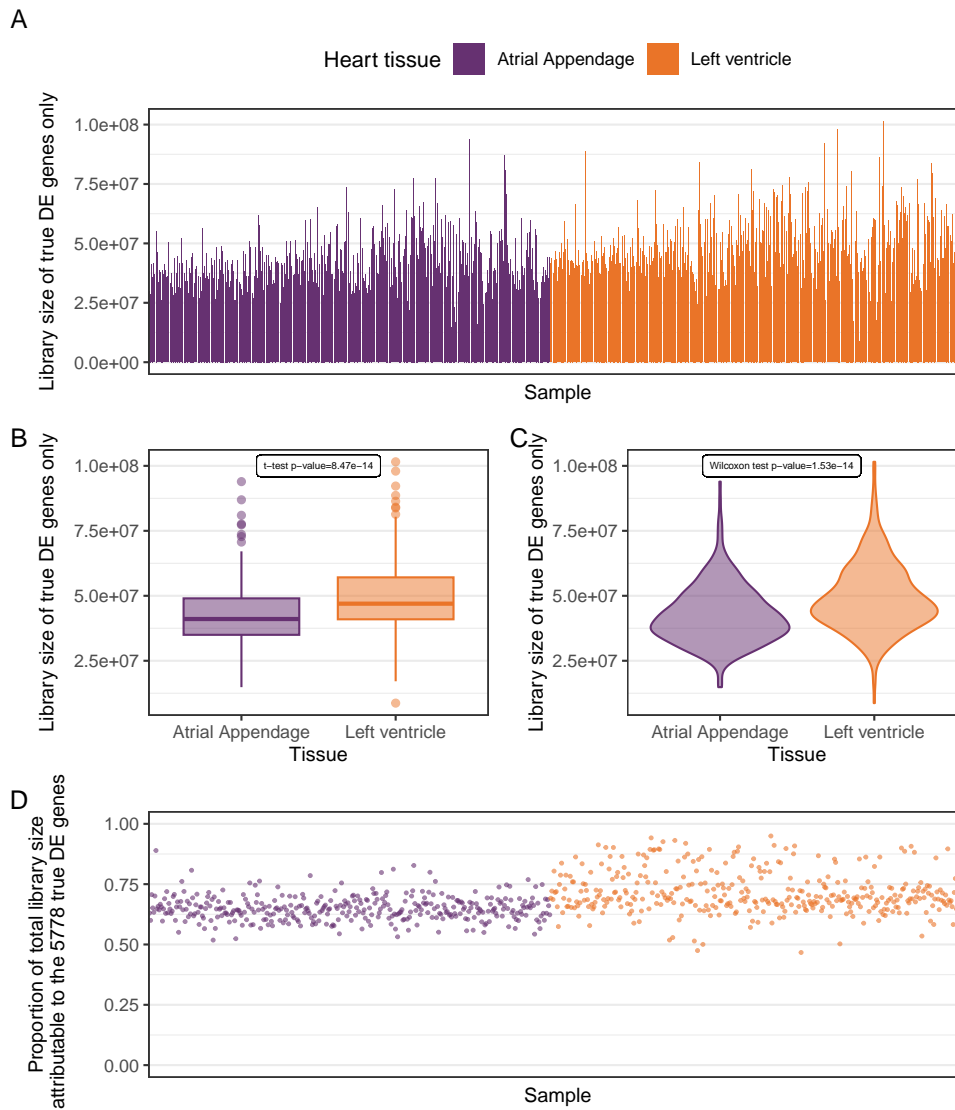


Figure 17: **Library size of true Differentially Expressed (DE) genes only in the *GTEX* Heart atrial appendage VS Heart left ventricle data.** **Panel A** displays the library sizes of all 758 samples (372 and 386 in the atrial appendage and left ventricle heart tissues respectively) when only using the 5,778 true DE genes. **Panel B** presents a boxplot highlighting the statistically significant difference with a t-test. **Panel C** presents a violin plot for a non-parametric comparison with the Wilcoxon test. **Panel D** shows that most of the total library size is accounted for by the subset of the 5,778 true DE genes, and even more so for the left ventricle heart tissue.

In addition, both `limma-voom` (Law et al., 2014) and `NOISeq` (Tarazona et al., 2015) also controlled the FDR adequately using our amended permutation scheme (note that this simulation is a little harder to operate for `voom-limma`, `edgeR` and `DESeq2` because normalization is baked directly into their implementation without user control). However, `DESeq2` (Love et al., 2014) and `edgeR` (Robinson et al., 2010) still exhibit inflated FDRs in line with previous findings (Agniel and Hejblum, 2017; Gauthier et al., 2020) – see Figure 19. Finally, `dearseq` asymptotic test achieved higher power compared to both `limma-voom` and `NOISeq` (when  $n > 40$  per group – see Figure 18). In conclusion, `dearseq` is capable of handling many experimental designs beyond the simple two conditions comparison setting of the Wilcoxon test, and thus remains a valid and versatile option for DEA of large human population samples.

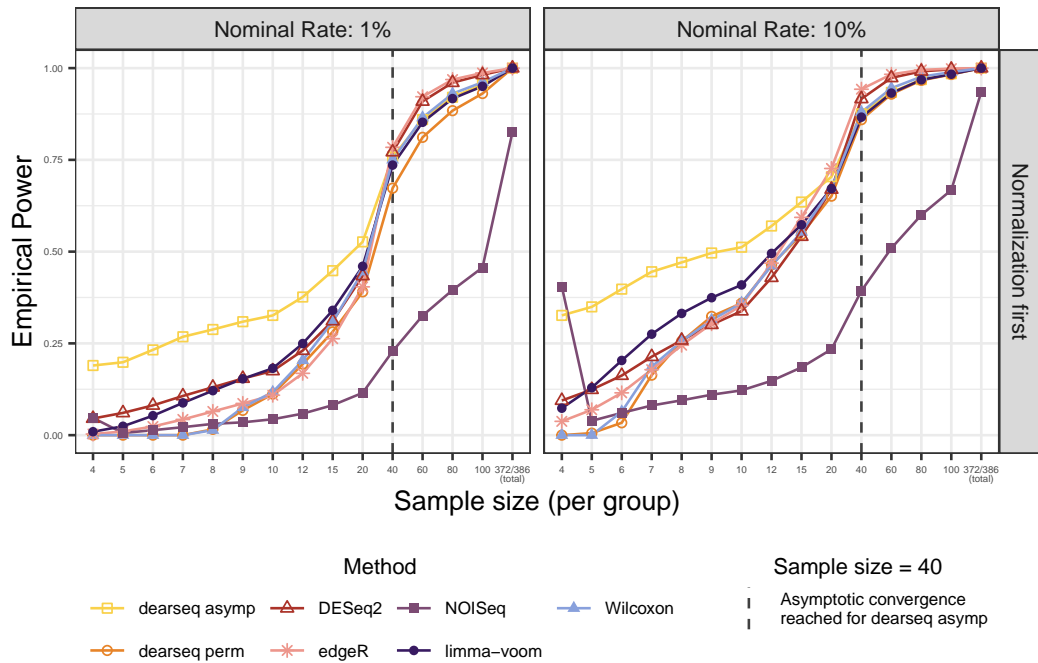


Figure 18: **Empirical statistical power by method.** Settings are identical to Figure 15 with only the amended generation scheme.

## 5 SOFTWARE DISSEMINATION & REPRODUCIBILITY

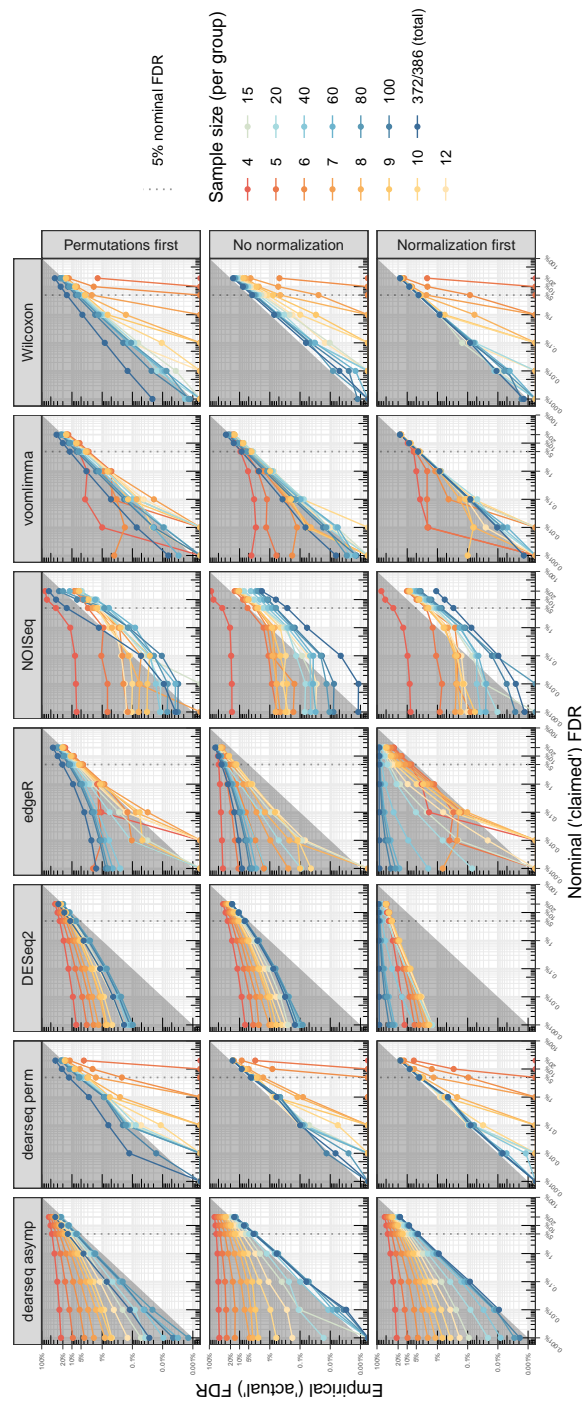


Figure 19: **Empirical FDR control against Nominal FDR level for all method.** Settings are identical to Figure 15

## 5 SOFTWARE DISSEMINATION & REPRODUCIBILITY

## 6 Directions for future research

While my statistical methods can be applied in various areas of clinical epidemiology and biomedical sciences, my primary focus remains centered on driving progress in vaccine clinical development. Through my collaborations with immunologists, and in particular at the VRI, I had the opportunity to work on several vaccines and vaccine candidates, tackling different viruses. Notably, I worked on 3 candidate vaccines against HIV, a therapeutic one in the DALIA trial (Hejblum et al., 2015; Thiébaud et al., 2019), and two prophylactic ones with the LIGHT (Lhomme et al., 2020), ANRS VRI01 (Richert et al., 2022) trials. I also worked on two candidate vaccines against Ebola, that have been since approved and licensed by both the Food and Drug Administration (FDA) in the USA and the European Medicines Agency (EMA) in the European Union : rVSV-ZEBOV (Rechtien et al., 2017), and Ad26.ZEBOV and MVA-BN-Filo (Blengio et al., 2023). Additionally, I studied the immune response to two mRNA vaccines against SARS-CoV-2 (both BNT162b2 and mRNA-1273 in Rinchai et al., 2022). In each of those clinical studies, high-throughput data was generated to deepen our understanding of the immune system and its response to those vaccines. However, those high-throughput have yet to prove they can turn into actionable discoveries for vaccine development. From a methodological standpoint, several bottlenecks need to be resolved before we can leverage these data to their full capacity.

### 6.1 Multi-modal data integration to enhance measurement resolution

While there exists many different high-throughput technologies that each generate different kinds of measurements, the output data are rarely analyzed together in the same study. Integrating several data modalities together, such as transcriptomics and FCM data for instance, can yet provide a broader picture of the immune system. In particular, the generation of FCM data requires the collection of significant amount of blood tubes, an operation that cannot be performed by study participants on their own. On the contrary, gene expression data – and in particular RNA-seq – have demonstrated their capacity to be collected through simple finger-prick (Obermoser et al., 2013b; Rinchai et al., 2022).



The immune system response to an infection or a vaccine can be observed through variations of cell-type proportions in circulating blood. Those variations are also impacting the bulk gene expression in the whole blood, as it is a mixture of cell-type specific expression. However, interpretation is easier at the cell-type level. In addition, their proportion variations can be also be fed into mechanistic modeling and inform understanding of the immune system. Methods for the deconvolution of cell types from bulk gene expression data leverage the relationship between those to infer cell-type proportions, using i) bulk transcriptomic data, and ii) prior biological knowledge in the form of reference signature matrices that map specific gene expression profile to predefined cell-types (Shen-Orr and Gaujoux, 2013).

There have been many approaches proposed to perform such cellular deconvolution of bulk gene expression (Avila Cobos et al., 2018; Hunt et al., 2019; Newman et al., 2019; Chu et al., 2022), and while they have shown promises in some contexts (Sharpe et al., 2018) their practical implementation in vaccine studies falls short of reaching acceptable performance when compared to FCM data in real data in my experience. Multiple factors can contribute to this failure and pave the way for future developments and performance enhancements. First, the reference signature matrices have a tremendous impact on the estimated proportions, yet those are highly dependent on the biological context and not necessarily robust to the technical variations affecting gene expression measurements. Meanwhile, there is currently a lack of a reference signature matrix that would have been generated on a large amount of data, across different platforms, and for the specific context of the immune system in vaccine studies or during an infection. Second, the modeling assumptions underlying all existing approaches suppose that all cell-types constitutive of the mixture are accounted for – an unrealistic assumption in practice. Third, the methods are usually ignorant of the hierarchical relationships between the different cell populations and assume a simple linear mixture.

Finally, if these deconvolution techniques were to reach gold-standard performance, they could replace other means of measuring cellular population variations including FCM. This would unlock the monitoring of many cellular populations at a time resolution never seen before, thanks to gene expression measurements with autonomous finger-prick collection possibly every day – or even every few hours – during a study.

## 6.2 Incorporating prior biological knowledge to enhance results robustness and overcome limited sample sizes

Early phase vaccine trials are characterized by small sample sizes. Yet, as highlighted in my work, they increasingly feature high-dimensional molecular data measurements, to uncover immunogenicity mechanisms and underlying cellular response determining vaccine effect. While those data, such as FCM or scRNA-seq, can be “big” when the statistical units considered are the cell, integrating the results at the individual subject level almost invariably yields limited sample size to a few dozens up to a few hundreds at best. In such settings, many statistical approaches show limitations. To counter-balance the latter, one solution is to incorporate external prior knowledge. Structuring high-dimensional data, for instance by grouping variables into sets derived from biological knowledge, or leveraging repeated observation through longitudinal modeling can spectacularly strengthen estimations (Hejblum et al., 2015; Liquet et al., 2016; Agniel and Hejblum, 2017). This simultaneously reduces the dimension of the estimation problem while multiplying the sample size.

Regarding gene set approaches for data, there remain challenges to summarize the information once significant association has been established. In particular, this is key to visualize and communicate association findings. Beyond this aspect, appropriately summarizing the information from a whole gene set is necessary in order to derive portable prediction signatures from one data set to another. Current approaches focus on crude aggregation of gene-level measure (such as the proportion of significant genes or their average expression) (Rinchai et al., 2022), a measure that might not represent the relevant information detected by advanced DEA methods.

Besides, Bayesian approaches are well suited to integrate external knowledge, thanks to the distribution *a priori* being embedded in their paradigm. While informative priors are relatively straightforward to use in parametric Bayesian models, their specification can become particularly challenging in Bayesian non-parametrics (Kessler et al., 2015; Hejblum et al., 2019). In Hejblum et al. (2019), I have demonstrated the value of non-parametric Bayesian clustering models for single-cell data analysis. However, additional developments are required to adapt those models for count data such as scRNA-seq or Cytometry by Time-Of-Flight (CyTOF) data. In addition, their implementation can be extremely demanding in terms of computation, and do not necessary scale well to the amount of data being generated by current high-throughput technologies.

Finally, the current dimension of gene expression prevents its direct inclusion

into mechanistic modeling of the immune response, such as compartment models for the antibody response (Clairon et al., 2023). One solution yet to integrate this mass of information is to perform dimension reduction, informed by prior biological knowledge. Cellular population estimation through deconvolution actually represents an approach to dramatically reduce the dimension while focusing on variations that are known to be informative for the immune system dynamics. This way, the high-dimension of the original transcriptomic data gets mitigated, and allows estimation given the limited sample size usually available.

### 6.3 Leveraging multiple scales to enhance population generalization

Single-cell technologies like scRNA-seq or FCM provide access to the populational distribution of molecular markers across cells. This adds yet another hierarchical level in multiple sample studies like vaccine trials, in addition to the subject, the condition (e.g. the treatment arm), and the time-point. Integration of these multiple hierarchical scales requires the development of tailored new approaches. For examples, there are no DEA methods for multi-sample scRNA-seq data that can account for those different heterogeneity sources.

Cell-type in scRNA-seq represents yet another intermediary scale that can generate heterogeneity. The current practice to tackle this is to perform data-driven clustering – often in a reduced dimensional space, which can pose its own set of issues (Chari and Pachter, 2023). As discussed in Section 3.3, this requires the development of new testing procedures to preserve Type-I error control through this double use of the data. Current approaches available for such post-clustering inference all need either knowing cell-type (i.e. cluster) specific distribution parameters or that those parameters are constant across clusters, making them inapplicable in practice. Relying on local estimators that would leverage the topology of the data without any clustering assumption represents an interesting avenue to try to resolve this circularity issue.

### 6.4 High-dimensional surrogate construction to enhance clinical relevance

If gene expression can act as a surrogate for immune response, then it could possibly be used to shorten vaccine trials or to quickly measure the effect of vaccination in a population. In light of the recent COVID-19 pandemic, the danger of emerging infectious diseases can hardly be overstated. Because vaccines are


the single most effective intervention against infectious diseases (Pulendran and Ahmed, 2011), efficient vaccine administration is necessary to contain and prevent the most dangerous outbreaks and epidemics. Surrogate markers for vaccine efficacy are mandatory to speed up vaccine development, facilitate licensure, and monitor effectiveness (World Health Organization, 2013). When a vaccine is available, it may not work equally well for all vaccinees (Huttner et al., 2018). A gene expression signature, or combination of important gene expression measurements, predictive of the vaccine response could be instrumental in reducing clinical trial times and developing personalized vaccine regimens, for example, identifying quickly and cheaply which persons did not respond adequately to the initial vaccination and should receive a new dose.

Two key challenges arise when attempting to use gene expression in vaccine research: one must establish if and how to use it to measure vaccine effects. First, one must determine if gene expression (measured once or at a few times) captures enough information about the vaccine effect or, specifically, what proportion of the vaccine effect is mediated through gene expression. Available mediation methods for estimating this proportion either rely on restrictive and unverifiable parametric assumptions (Zhou et al., 2020; Zhang et al., 2016; Song et al., 2018; Chén et al., 2018; Zhong et al., 2019; Zhao and Luo, 2022), or are non-parametric (Díaz et al., 2021; Xia and Chan, 2021) but infeasible given the relatively small sample sizes typically available in transcriptomics substudies from vaccine trials. Second, one must determine how to build and use a gene expression signature for estimating vaccine effects in a future study. Even if gene expression mediates all or most of the vaccine effect, a particular gene expression signature is not guaranteed to well capture the effect of the vaccine. Powerful machine learning methods may be used to predict vaccine effects from gene expression, but creating an optimal signature may require more than good prediction when the vaccine effect is not entirely mediated by gene expression (Wang et al., 2020b). Furthermore, using surrogate endpoints (like a gene expression signature) in future studies can lead to bias or over-optimism without proper methods to correct downstream analyses (Wang et al., 2020a).

Available approaches to quantify gene expression mediation are unreliable in small sample sizes and will fail altogether when the dimension of the genes is much larger than the sample size, especially as gene expression data are known to be noisy. Creating an optimal gene expression signature to capture vaccine effects requires new tools. Traditional approaches to creating gene expression signatures have been *ad hoc*, often a collection of genes that were differentially expressed between vaccinated and control groups or before/after vaccination (Querec et al., 2009a; Bucasas et al., 2011; Rehtien et al., 2017). More modern approaches have used machine learning to predict vaccine response (Lee et al., 2016; Gonzalez-Dias

et al., 2020; Cotugno et al., 2020; Richert et al., 2022). However, good prediction is not the only required feature of an optimal signature, and it should also include an adjustment term that additionally encodes part of the relationship between the vaccine and gene expression (Wang et al., 2020b, 2023). These approaches have yet to be extended to the high-dimensional context of gene expression.

## 6.5 Computational efficiency to enhance numerical scalability

Available implementation has become integrative to statistical method development. In statistical genomics, the dominant language currently remains  with platforms like Bioconductor, although more and more Python packages are also being released. One key aspect that applies to all of the research directions outlined above is the importance of computational efficiency for the methods to be developed. As mentioned in Section 6.2, the size and dimension of the data warrants a special attention towards scalability of implementation. For instance, that means that numerical optimization algorithms should be preferred to sampling alternative (e.g. for non-parametric Bayesian models, variational inference is likely to be much faster than Monte Carlo Markov chains). This point is of particular importance to enhance the dissemination of my developments, and to increase their impact in the broader scientific community.

# Curriculum Vitæ

## Research experience

- 2021–present **Faculty Researcher (*Chargé de Recherche*)** in Biostatistics, tenured, Inserm U1219 *Bordeaux Population Health* research center, *SISTM team*, Bordeaux (France).
- 2016–2021 **Associate Professor (*Maître de Conférences*)** in Biostatistics, tenured, ISPED *Bordeaux School of Public Health*, Bordeaux University, Bordeaux (France).
- 2016 **Postdoctoral Research Associate**, Department of Biostatistics, Harvard School of Public Health, Boston (USA).
- 2015–2016 **Postdoctoral Research Fellow**, Department of Biostatistics, Harvard School of Public Health, Boston (USA).
- 2011–2015 **Research Assistant** (Ph.D. student), Inserm U897 *Biostatistics team*, Bordeaux (France).
- Apr.–Sept. 2011 **Research Assistant** (Masters intern), Inserm U897 *Biostatistics team*, Bordeaux (France).  
Development of dynamic statistical models applied to the epidemiology of myocardial infarction.
- May–Jul. 2011 **Statistician Assistant** (Masters intern), *AltraBio* (start-up in biotechnologies), Lyon (France).  
Analysis of transcriptomics data of preclinical trials.

## Education

- 2011–2015 **Ph.D. in Biostatistics**, ISPED *Bordeaux School of Public Health*, Bordeaux University.  
Integrative analysis of high-dimensional data applied to vaccine research.  
Advisors: Rodolphe Thiébaud & Francois Caron

- 2008–2011 **Master of Science (M.Sc.) in Statistics** (*diplôme d'ingénieur*), ENSAI, National School for Statistics and Information Analysis (*École Nationale de la Statistique et de l'Analyse de l'Information*), Rennes (France). Specialization in biostatistics, with high honors.
- 2011 **Master of Science (M.Sc.) in Statistics and Econometrics**, Department of Mathematics, University of Rennes 1, Rennes (France). Dual degree partnership in conjunction with studies at ENSAI (additional education focused on scientific research).
- 2009 **Bachelor of Science (B.Sc.) in Mathematics** (*licence de mathématiques*), Pierre and Marie Curie University – Paris 6 (UPMC), Paris (France). In conjunction with studies at ENSAI (dual curriculum, remote learning).
- 2006–2008 **Post-Secondary Preparatory Classes** (*Classes Préparatoires aux Grandes Écoles – CPGE*), Lycée Hoche, Versailles (France). University-level courses required in preparation for competitive exams into top universities, engineering, and graduate schools (France's *Grandes Écoles*). Major in Mathematics and Physics.
- 2006 **High school diploma**, Lycée Richelieu, Rueil-Malmaison (France). With high honors.

## Teaching experience

- 2019 - present **International Ph.D. course**, Graduate School of Health and Medical Sciences, University of Copenhagen, (Denemark)
- Bayesian methods in biomedical research (graduate class, 3.5 days per year)
- 2018 - present **Ph.D. courses**, Bordeaux University (France)
- R for development & performance (graduate class, 18h per year)
  - Basics for data science using R (graduate class, 12h per year)
- 2019 - present **Master in Public Health**, ISPED, Bordeaux University (France)
- omics data analysis (graduate class, 20h per year)
  - data visualization (undergraduate class, 4h)
- 2021 - present **Master in numerical sciences & bio-health**, École Centrale Nantes (France)
- Statistical learning in high-dimension (graduate class, 2h per year)

## CV

2016 - 2021 **Associate Professor**, Bordeaux University, France

Ph.D. courses:

- Introduction to Bayesian analysis for biometric research (graduate class, 18h per year)

Master in Public Health Data Science & Master in Biostatistics courses:

- likelihood estimation and multivariate regression (graduate class, 30h per year)
- factor methods for multivariate data analysis (graduate class, 30h per year)
- Bayesian analysis and sampling methods (graduate class, 30h per year)
- omics data analysis (graduate class, 20h per year)
- sparse Partial Least Squares methods (graduate class, 7h per year)
- ANOVA regression (graduate class, 7.5h per year)
- hypothesis testing (graduate class, 30h per year)
- advanced R (undergraduate class, 15h per year)

2012 - 2014 **Teaching Assistant**, Bordeaux University, France

Master in Public Health and Master in Biostatistics courses:

- MCMC methods for Bayesian analysis (graduate class, 12h)
- sparse Partial Least Squares methods (graduate class, 5h)
- basic statistics (undergraduate class 16h)
- logistic regression (undergraduate class, 12h)
- R software (undergraduate class 9h)

---

## Scientific supervision

### Postdoctoral researchers

- Laura Villain (2019 – 2021: 100%)
- Hung Van Tran (2019: 50%)

### Ph.D. students

- Ansh Pal (2023 – ... : 50%)
- Arthur Hughes (2023 – ... : 50%)
- Kalidou Ba (2022 – ... : 50%)
- Benjamin Hivert (2020 – ... : 50%)
- Paul Freulon (2019 – 2022: 50%)
- Marine Gauthier (2018 – 2021: 50%)
- Soufiane Ajana (2017 – 2019: 15%)
- Stephanie Chan (2016: 15%)



## Engineers

- Sara Fallet (2023 – . . . : 100%)
- Kalidou Ba (2021 – 2022: 100%)

## Interns

- Rebecca Knowlton (1 month PhD research visit 2023: 100%)
- Arthur Hughes (M2 internship 2023: 100%)
- Maud Perpère (M1 internship 2023: 100%)
- Emma Avisou (M1 internship 2021: 100%)
- Clément Bonnet (M1 internship 2021: 100%)
- Benjamin Hivert (M2 master thesis 2020: 100%)
- Anthony Devaux (M2 master thesis 2019: 100%)
- Aaron Sonabend (2 months PhD research visit 2019: 100%)
- Victor Gasque (M1 internship 2019: 50%)
- Thomas Ferte (M2 master thesis 2019: 100%)
- Marine Gauthier (M2 master thesis 2018: 100%)
- Roxane Coueron (M2 master thesis 2018: 50%)
- Paul Tauzia (M2 master thesis 2017: 50%)
- Chariff Alkassim (M2 master thesis 2015: 50%)
- Damien Chimits (M2 master thesis 2014: 50%)
- Lise Cahuzac (M1 internship 2013: 50%)

## Grants & funding

- 2021-2024 **Principal Investigator** of the Inria associate-team DESTRIER: “DEfining Surrogacy of early Transcriptomics foR vaccInE Response” (32K€ over 3 years)
- 2022-2026 **Work-Package leader** *Réseau de Recherche Impulsion* ”Public Health Data Science”, *Université de Bordeaux*.
- 2023-2027 **Task leader** PEPR *Santé Numérique*, axis “Statistical and AI based Methods for Advanced clinical Trials Challenges in digital Health” (funding 1 PhD student).
- 2023-2027 **Task leader** PEPR *Santé Numérique*, axis “multiScale AI for SingleCell-based precision MEDicine” (funding 50% of 1 PhD student).
- 2016-2024 **Participant** ANRS LabEx Programme “Vaccine Research Institute” (VRI).
- 2020-2024 **Participant** (genomics-statistics referent) in the EU H2020 Framework Programme “IP-cure-B” (*Immune profiling to guide host-directed interventions to cure HBV infections*).

## CV

- 2018-2020 **Principal Investigator** of the Inria associate-team SWAGR: “Statistical Workforce for Advanced Genomics using RNA-seq” (36K€ over 3 years)
- 2019-2021 **Principal Investigator** of the Technology Development Action from Inria Bordeaux Sud-Ouest “VASI” (*Visualization and Analysis Solutions for Immunologists*): 2 year support for a software engineer.
- 2019-2022 **Participant** (computational statistics referent) in the ANR-18-CE36-0004 “DyMES” (*Dynamic Models for Epidemiological Longitudinal Studies of Chronic Diseases*).
- 2018-2020 **Teaching discharge** for research at Inria Bordeaux Sud-Ouest: 96h per year.
- 2017-2020 **Participant** (RNA-seq analysis referent) in the Transcan-2 ERANET “GLIOMA-PRD” (*Multi-parametric analysis of the evolution and progression of low grade glioma*): support for a post-doctoral researcher for 2 years.
- 2016-2019 **Participant** (réfèrent statistique en grande dimension) au *Research and Innovation Programme* n°634479 de EU H2020 EYE-RISK (*Systems medicine for identifying risk factors, molecular mechanisms and therapeutic approaches for age-related macular degeneration*).
- 2016 **Recipient** of a travel grant from the Harvard Program in Quantitative Genomics (PQG) to attend the ENAR conference.
- 2011 **Recipient** of a Ph.D. grant from the EHESP (*École des Hautes Études en Santé Publique*, Rennes, France) – ranked 1<sup>st</sup>.



---









## Patents





- 2021 Invention patent EP20306527/WO2022122959A1 (inventor 1/5<sup>th</sup>)  
Use of cd177 as biomarker of worsening in patients suffering from covid-19
- 2020 Invention patent WO2021058914A1/FR1910515 (inventor 1/7<sup>th</sup>)  
Prediction of the content of omega-3 polyunsaturated fatty acids in the retina by measuring 7 cholesterol ester molecules

---

## Software development & maintenance

- 2023 **citcdf**: an  package for performing Conditional Independence Testing Through Conditional Cumulative Distribution Function Estimation. Available on GitHub . *Co-creator & maintainer*.

- 2022 **CytOpT**: an  package for automatic gating transfer in cytometry data using optimal transport with domain adaptation. Uses Python code.. Available on CRAN, development version on GitHub . *Co-creator & maintainer*.
- 2020 **dearseq**: an  package for Differential Expression Analysis for RNA-seq data through a robust variance component test. Available on , development version on GitHub . *Co-creator & maintainer*.
- 2019 **vici**: an interactive  Shiny application for accurate estimation of vaccine induced cellular immunogenicity with bivariate linear modeling. Available online or locally from the CRAN, development version on GitHub . *Creator & maintainer*.
- 2019 **marqLevAlg**: an  package for (parallelized) optimization of convex multiparametric functions. Available on CRAN, development version on GitHub . *Contributor*.
- 2019 **foodingraph**: an  package for displaying weighted undirected food networks from adjacency matrices. Available on CRAN, development version on GitHub . *Co-creator*.
- 2019 **phenotypr**: an  package for probabilistic phenotyping patients from electronic health records using both diagnosis codes and natural language processed medical notes. Available on CRAN, development version on GitHub . *Creator & maintainer*.
- 2017 **ludic**: an  package for probabilistic record linkage using diagnosis codes. Available on CRAN, development version on GitHub . *Co-creator & maintainer*.
- 2017 **cytometree**: an  package for automatic gating and annotation of flow-cytometry data. Available on CRAN, development version on GitHub . *Co-creator & maintainer*.
- 2017 **sslcov**: an  package for covariance semi-supervised learning. Available on GitHub . *Co-creator*.
- 2016 **tcgsaseq**: an  package for longitudinal RNA-seq data analysis at the gene set level. Available on GitHub . *Co-creator & maintainer*.
- 2017 **kernscr**: an  package for survival analysis by gene sets in presence of competing risks. Available on CRAN, development version on GitHub . *Co-creator & maintainer*.

- 2015 **NPflow**: an  package for clustering of large cell populations with Dirichlet process mixture of skew-Normal and skew-t distributions. Uses C++ code to speed up computation.. Available on CRAN, development version on GitHub . *Co-creator & maintainer.*
- 2014 **TcGSA**: an  package for longitudinal gene-expression data from microarrays at the gene set level. Available on CRAN, development version on GitHub . *Creator & maintainer.*

## Active international research collaborations

**Denis Agniel**, *Rand Corporation, Statistics group*, Santa Monica (CA, USA), Associate Statistician.

**Tianxi Cai**, *Harvard TH Chan School of Public Health, Department of Biostatistics*, Boston (MA, USA), Professor.

**Layla Parast**, *University of Texas at Austin*, Austin (TX, USA), Associate Professor.

## Outreach activities

- 2022-present *Chiche ! 1 Scientifique, 1 Classe* Program by Inria  
1h presentation & open discussion about scientific research with high-school students.
- 2018 Outreach stand "Is there more data in a drop of blood than in my smartphone?" at the 10 year anniversary of Inria Bordeaux Sud-Ouest
- 2012 Poster presentation at the Summer University of Sidaction on longitudinal analysis applied to HIV vaccine research

## Research visits abroad

- 2018-2019 **MRC Biostatistics Unit, Cambridge University**, Cambridge (2×3 weeks) (United-Kingdom)  
invited by Sylvia Richardson, Professor.
- 2018 **Rand Corporation, Statistics group**, Santa Monica (CA, USA) (1 week)  
invited by Denis Agniel, Associate Statistician.
- 2016-2017 **Harvard University, Department of Biostatistics**, Cambridge (2×1 week) (MA, USA)  
invited by Tianxi Cai, Professor.

- 2013-2014 **University of Oxford, Department of Statistics**, Oxford  
(3×1 week) (United-Kingdom)  
invited by François Caron, Research Fellow.
- 2012 **Benaroya Research Institute**, Chaussabel Laboratory, Seattle  
(1 month) (WA, USA)  
invited by Damien Chaussabel, Director of Systems Immunology.
- 2011 **Baylor Institute for Immunology Research**, Dallas (TX, USA).  
(1 month)

### Scientific evaluation

- 2023 **Reviewer for the 2023 MESSIDORE project call from Inserm IReSP “Méthodologie des ESSais cliniques Innovants, Dispositifs, Outils et Recherches Exploitant les données de santé et biobanques”**
- 2023 **Member of the Scientific Committee for the CNC23 9<sup>th</sup> Channel Network Conference of the International Biometric Society 2023**
- 2021 **Member of the PhD defense committee of Shaima Belhechmi, *Université Paris-Saclay***
- 2021 **Reviewer for the ANRT, (*Association Nationale de la Recherche Technologique*)**
- 2021 **Member of the Scientific Committee for the 42<sup>nd</sup> ISCB conference**
- 2021 **Member of the Pharm. D. defense committee of Blandine Malbos, *Université d’Angers***
- 2019 **Invited member of the PhD defense committee of Soufiane Ajana, *Université de Bordeaux***

### Reviewer for international peer-reviewed scientific journals

*Annals of Applied Statistics, Bayesian Analysis, BioData Mining, Bioinformatics, Biometrics, Cell Reports Methods, Cancer Reports, Computational Statistics Data Analysis, Journal of Open Source Software, Journal of Statistical Computation and Simulation, PLOS Computational Biology, Scientific Reports, STAT, Statistics in Medicine, Statistical Applications in Genetics and Molecular Biology, WIREs Applications in Genetics and Molecular Biology*

## Academic responsibilities

- 2023–present **Member of the Organizing Committee for the next annual conference “*Journées de Statistique*” of the French Statistical Society (SFdS)**
- 2021–present **French Biometric Society correspondant to the Channel Network region of the International Biometrics Society**
- 2019–present **Member of the Bureau of the French Biometric Society (*Société Française de Biométrie*) – webmaster**
- 2019 **Co-organizer of the Bordeaux Statistics Seminar series (quarterly)**
- 2017–present **Organizer of the Public Health Department Biostatistics Seminar series (biweekly)**
- 2018 **Co-organizer of the workshop in honor of Daniel Comenges’ 70<sup>th</sup> birthday**
- 2012–2014 **Founder of the ISPED Ph.D. students (weekly) seminar**
- 2009–2010 **President (formerly Secretary General) of the ENSAI Business Networking Forum**  
Responsible for organizing the yearly networking event between companies and ENSAI students
- 2009 **Vice President of the ENSAI Student Council**  
Organize and coordinate associative activities and social life at the school

## Selected communications

▷ **Oral communications:** (\* indicates invited talks)

- Mexico 2022\* Hejblum B, Parast L, Agniel D, Transcriptomics: a potential early surrogate for vaccine response ?, *BIRS-CMO 22w5184*, Oaxaca.
- Latvia 2022 Hejblum B, Gauthier M, Ba K, Thiébaud R, Agniel D, Distribution-free complex hypothesis testing for single-cell RNA-seq differential expression analysis, *31<sup>st</sup> International Biometric Conference*, Riga.
- France 2022\* Hejblum B, Machine learning approaches for the analysis of bulk and single-cell RNA-seq data, *4<sup>th</sup> GenMed workshop on Medical Genomics*, Paris.
- Germany 2022\* Hejblum B, Teaching Bayesian statistics during a pandemic, *German Association for Medical Informatics, Biometry and Epidemiology (GMDS) Teaching & Didactics workshop*, Saarbrücken.

- France 2021\* Prague M, Collin A, Wittkop L, Dutartre D, Clairon Q, Moireau P, Thiébaud R, Hejblum B, Leveraging random effects to estimate the impact of NPIs on epidemic dynamics across French regions, *8<sup>th</sup> Channel Network Conference of the International Biometric Society*, Paris.
- France 2021\* Hejblum B, Clustering of flow cytometry data using non parametric Bayesian modeling, *Séminaire LMBA*, Vannes.
- France 2021\* Hejblum B, Distribution-free complex hypothesis testing for single-cell RNA-seq differential expression analysis, *Statistical Methods for Post-Genomic Data (SMPGD) – 2021*, online.
- France 2020 Hejblum B, Gauthier M, Thiébaud R, Agniel D, A variance component score test for flexible RNA-Seq data differential analysis, *Statistical Methods for Post-Genomic Data (SMPGD) – 2020*, Paris.
- France 2019\* Hejblum B, Montani I, Leffondré K, Diallo G, Mouglin F, Pariente A, Richert L, Thiessard F, Joly P, Alioum A, Tzourio C, Thiébaud R, Enseigner la science des données en santé publique, *Colloque Francophone International sur l'Enseignement de la Statistique (CFIES)*, Strasbourg.
- France 2019\* Hejblum B, Gauthier M, Thiébaud R, Agniel D, Controlling Type-I error in RNA-seq differential analyses through a variance component score test with an application to tuberculosis infection, *Séminaire de l'équipe de Statistique de l'Institut de Recherche MATHématique de Rennes (IRMAR)*, Rennes.
- UK 2019\* Hejblum B, Kirk PDW, Scaling up nonparametric Bayesian clustering with MCMC for big data applications, *12<sup>th</sup> International Conference of the ERCIM WG on Computational and Methodological Statistics*, Londres.
- Taiwan 2019\* Hejblum B, Gauthier M, Thiébaud R, Agniel D, A variance component score test applied to RNA-Seq differential analysis, *3<sup>rd</sup> EcoSta Conference*, Taichung.
- France 2019 Hejblum B, Lhomme E, Thiébaud R, Richert L, VICI: a Shiny app for accurate estimation of Vaccine Induced Cellular Immunogenicity with bivariate modeling, *UseR! 2019*, Toulouse.
- France 2018\* Hejblum B, Gauthier M, Thiébaud R, Agniel D, Controlling type-I error and false discoveries in RNA-seq differential analyses through a variance component score test, *Bioinfo-Biostat GenoToul Annual Day*, Toulouse.

## CV

- Spain 2018 Hejblum B, Agniel D, A variance component score test for RNA-seq differential analysis in vaccine trials, *29<sup>th</sup> International Biometric Conference*, Barcelona.
- UK 2017\* Hejblum, Alkhassim, Gottardo, Caron, Thiébaud, Dirichlet Process Mixtures of Multivariate Skew t-distributions for Unsupervised Clustering of Cell Populations from Flow-Cytometry Data, *BSU invited Seminar*, Cambridge.
- Spain 2017 Hejblum B, Agniel D, Type I error and false discovery rate control in RNA-seq differential analyses through a variance component score test, *38<sup>th</sup> Annual Conference of the International Society for Clinical Biostatistics*, Vigo.
- USA 2016 Hejblum B, Agniel D, Time-course Gene Set Analysis of longitudinal RNA-seq data, *ENAR 2016 Spring Meeting*, Austin (TX).
- Italy 2014 Hejblum B, Caron F, Thiébaud R, Bayesian analysis of time-course flow cytometry data with Dirichlet process mixture modeling, *27<sup>th</sup> International Biometric Conference*, Florence.
- France 2014 Hejblum B, Genuer R, Thiébaud R, Variable selection in high-dimensional dataset: comparison of sPLS with other approaches in an HIV vaccine trial, *8<sup>th</sup> International Conference on Partial Least Squares and Related Methods*, Paris.
- France 2014\* Hejblum B, Caron F, Thiébaud R, Bayesian nonparametric modeling of flow cytometry data with Dirichlet process mixtures, *Ph.D. students working group of the LSTA (Laboratoire de Statistique Théorique et Appliquée) in Paris 6 University*, Paris.
- Spain 2013 Thiébaud R, Hejblum B, Skinner J, Montes M, Chêne G, Palucka K, Banchereau J, Lévy Y, Integrative Analysis of Responses to Dendritic-Cell Vaccination Identifies Signatures Correlated with Control of HIV Replication: The DALIA Trial, *AIDS Vaccine 2013, AIDS Research and Human Retroviruses*, Barcelone.
- Norway 2012 Hejblum B, Skinner J, Thiébaud R, Application of Gene Set Analysis of Time-Course gene expression in a HIV vaccine trial, *33<sup>rd</sup> Annual Conference of the International Society for Clinical Biostatistics*, Bergen.
- ▷ [Written communications](#)
- USA 2015 Hejblum B, Cai T, Weber G, Probabilistic Patient Linkage Algorithms for PIC-SURE, *BD2K all Hands Meeting 2015*, Bethesda (MD).



UK 2014 Hejblum B, Caron F, Thiébaud R, Hierarchical Analysis of Time-Course Flow Cytometry Data with Dirichlet Process Mixture Modeling, *Medical Research Council Conference on Biostatistics in celebration of the MRC Biostatistics Unit's centenary year*, Cambridge.

## List of scientific publications

### Published articles

#### 2024

- A1. Hejblum BP, Ba K, Thiébaud R, and Agniel D. Neglecting normalization impact in semi-synthetic RNA-Seq data simulation generates artificial false positives. *Genome Biology*, in press, 2024.
- A2. Hivert B, Agniel D, Thiébaud R, and Hejblum BP. Post-clustering difference testing: valid inference and practical considerations with applications to ecological and biological data. *Computational Statistics & Data Analysis*, 107916, 2024. DOI: [10.1016/j.csda.2023.107916](https://doi.org/10.1016/j.csda.2023.107916).

#### 2023

- A3. Freulon P, Bigot J, and Hejblum BP. CytOpT: Optimal Transport with Domain Adaptation for Interpreting Flow Cytometry data. *Annals of Applied Statistics*, 17(2): 1086–1104, 2023. DOI: [10.1214/22-AOAS1660](https://doi.org/10.1214/22-AOAS1660).
- A4. Agniel D, Hejblum BP, Thiébaud R, and Parast L. Doubly-robust evaluation of high-dimensional surrogate markers. *Biostatistics*, 24(4): 985–999, 2023. DOI: [10.1093/biostatistics/kxac020](https://doi.org/10.1093/biostatistics/kxac020).
- A5. Collin A, Hejblum BP, Vignals C, Lehot L, Thiébaud R, Moireau P, and Prague M. Using population based Kalman estimator to model COVID-19 epidemic in France: estimating the effects of non-pharmaceutical interventions on the dynamics of epidemic. *The International Journal of Biostatistics*, in press, 2023. DOI: [10.1515/ijb-2022-0087](https://doi.org/10.1515/ijb-2022-0087).
- A6. Colas C, Hejblum B, Rouillon S, Thiébaud R, Oudeyer PY, Moulin-Frier C, and Prague M. Epidemioptim: A toolbox for the optimization of control policies in epidemiological models. *Journal of Artificial Intelligence Research*, 71: 479–519, 2021. DOI: [10.1613/jair.1.12588](https://doi.org/10.1613/jair.1.12588).
- A7. Thiébaud R, Hejblum B, Mouglin F, Tzourio C, and Richert L. Chatgpt and beyond with artificial intelligence (ai) in health: Lessons to be learned. *Joint Bone Spine*, 90(5): 105607, 2023. DOI: [10.1016/j.jbspin.2023.105607](https://doi.org/10.1016/j.jbspin.2023.105607).

- A8. Blengio F, Hocini H, Richert L, Lefebvre C, Durand M, Hejblum B, Tisserand P, McLean C, Luhn K, Thiebaut R, and Lévy Y. Identification of early gene expression profiles associated with long-lasting antibody responses to the Ebola vaccine Ad26. ZEBOV/MVA-BN-Filo. *Cell Reports*, 42(9): 113101, 2023. DOI: [10.1016/j.celrep.2023.113101](https://doi.org/10.1016/j.celrep.2023.113101).
- A9. Vignals C, Hejblum BP, and Prague M. Modéliser la covid-19: de la population à l'individu. *Interstices*, 2023. URL <https://interstices.info/modeliser-la-covid-19-de-la-population-a-lindividu/>.

## 2022

- A10. Ferté T, Jouhet V, Greffier R, Hejblum BP, and Thiébaut R. The benefit of augmenting open data with clinical data-warehouse EHR for forecasting SARS-CoV-2 hospitalizations in Bordeaux area, France. *JAMIA open*, ooac086, 2022. DOI: [10.1093/jamiaopen/ooac086](https://doi.org/10.1093/jamiaopen/ooac086).
- A11. Richert L, Lelièvre JD, Lacabaratz C, Hardel L, Hocini H, Wiedemann A, Lucht F, Poizot-Martin I, Bauduin C, Diallo A, Rieux V, Durand M, Hejblum BP, Launay O, Thiébaut R, Lévy Y, and on behalf of the ANRS VRI01 Study group . T-cell immunogenicity, gene expression profile and safety of four heterologous prime-boost combinations of hiv vaccine candidates in healthy volunteers - results of the randomized multi-arm phase i/ii anrs vri01 trial. *Journal of Immunology*, 208(12): 2663–2674, 2022. DOI: [10.4049/jimmunol.2101076](https://doi.org/10.4049/jimmunol.2101076).
- A12. Rinchai D, Deola S, Zoppoli G, Ahamed Kabeer BS, Taleb S, Pavlovski I, Maacha S, Gentilcore G, Toufiq M, Mathew L, Liu L, Vempalli FR, Mubarak G, Lorenz S, Sivieri I, Cirmena G, Dentone C, Cuccarolo P, Giacobbe D, Baldi F, Garbarino A, Cigolini B, Cremonesi P, Bedognetti M, Ballestrero A, Bassetti M, Hejblum BP, Augustine T, Van Panhuys N, Thiébaut R, Branco R, Chew T, Shojaei M, Short K, Feng C, PREDICT-19 consortium , Zughailer SM, De Maria A, Tang B, Ait Hssain A, Bedognetti D, Grivel JC, and Chaussabel D. High-temporal resolution profiling reveals distinct immune trajectories following the first and second doses of COVID-19 mRNA vaccines. *Science Advances*, 8(45): eabp9961, 2022. DOI: [10.1126/sciadv.abp9961](https://doi.org/10.1126/sciadv.abp9961).

## 2021

- A13. Lin L and Hejblum BP. Bayesian mixture models for cytometry data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13: e1535, 2021. DOI: [10.1002/wics.1535](https://doi.org/10.1002/wics.1535).
- A14. Ferte T, Cossin S, Schaeverbeke T, Barnetche T, Jouhet V, and Hejblum BP. Automatic phenotyping of electronic health record: Phevis algorithm. *Journal of Biomedical Informatics*, 117: 103746, 2021. DOI: [10.1016/j.jbi.2021.103746](https://doi.org/10.1016/j.jbi.2021.103746).
- A15. Philipps V, Hejblum BP, Prague M, Commenges D, and Proust-Lima C. Robust and efficient optimization using a marquardt-levenberg algorithm with r package marqlevalg. *The R Journal*, 13(2): 365–379, 2021. DOI: [10.32614/RJ-2021-089](https://doi.org/10.32614/RJ-2021-089).
- A16. Zhang HG\*, Hejblum BP\*, Weber G, Palmer N, Churchill S, Szolovits P, Murphy S, Liao K, Kohane I, and Cai T. Atlas: An automated association test using probabilistically linked health records with application to genetic studies. *Journal of the American Medical Informatics Association*, 28(12): 2582–2592, 2021. DOI: [10.1093/jamia/ocab187](https://doi.org/10.1093/jamia/ocab187).
- A17. Lévy Y, Wiedemann A\*, Hejblum BP\*, Durand M, Lefebvre C, Surénaud M, Lacabaratz C, Perreau M, Foucat E, Déchenaud M, Tisserand P, Blengio F, Hivert B, Gauthier M, Cervantes-Gonzalez M, Bachelet D, Laouénan C, Bouadma L, Timsit JF, Yazdanpanah Y, Pantaleo G, Hocini H\*, and Thiébaud R\*. Cd177, a specific marker of neutrophil activation, is associated with coronavirus disease 2019 severity and death. *iScience*, 24(7): 102711, 2021. DOI: [10.1016/j.isci.2021.102711](https://doi.org/10.1016/j.isci.2021.102711).
- A18. Ajana S, Cougnard-Grégoire A, Colijn J, Merle BM, Verzijden T, de Jong P, Hofman A, EYE-RISK Consortium, Vingerling J, Hejblum BP, Korobelnik JF, Meester-Smoor M, Jacqmin-Gadda H, Klaver C, and Delcourt C. Predicting progression to advanced age-related macular degeneration from clinical, genetic and lifestyle factors using machine learning. *Ophthalmology*, 128(4): 587–597, 2021. DOI: [10.1016/j.ophtha.2020.08.031](https://doi.org/10.1016/j.ophtha.2020.08.031).
- A19. Lefèvre-Arbogast S, Hejblum BP, Helmer C, Klose C, Manach C, Low DY, Urpi-Sarda M, Andres-Lacueva C, González-Domínguez R, Aigner L, Altendorfer B, Lucassen PJ, Ruigrok SR, De Lucia C, Du Preez A, Proust-Lima C, Thuret S, Korosi A, and Samieri C. Early signature in the blood lipidome associated with subsequent cognitive decline in the elderly: A case-control analysis nested within the three-city cohort study. *EBioMedicine*, 64: 103216, 2021. DOI: [10.1016/j.ebiom.2021.103216](https://doi.org/10.1016/j.ebiom.2021.103216).

- A20. Acar N, Merle BMJ, Ajana S, He Z, Grégoire S, Hejblum BP, Martine L, Buaud B, Bron AM, Creuzot-Garcher CP, Korobelnik JF, Berdeaux O, Jacqmin-Gadda H, Bretillon L, Delcourt C, and for the Biomarkers of Lipid Status And metabolism in Retinal ageing (BLISAR) Study Group . Predicting the retinal content in omega-3 fatty acids for age-related macular-degeneration. *Clinical and Translational Medicine*, 11(7): e404, 2021. DOI: [10.1002/ctm2.404](https://doi.org/10.1002/ctm2.404).

## 2020

- A21. Gauthier M, Agniel D, Thiébaud R, and Hejblum BP. dearseq: a variance component score test for rna-seq differential analysis that effectively controls the false discovery rate. *NAR Genomics and Bioinformatics*, 2(4): lqaa093, 2020. DOI: [10.1093/nargab/lqaa093](https://doi.org/10.1093/nargab/lqaa093).
- A22. Lhomme E, Hejblum BP, Lacabaratz C, Wiedemann A, Lelièvre JD, Lévy Y, Thiébaud R, and Richert L. Analyzing cellular immunogenicity in vaccine clinical trials: a new statistical method including non-specific responses for accurate estimation of vaccine effect. *Journal of Immunological Methods*, 477: 112711, 2020. DOI: [10.1016/j.jim.2019.112711](https://doi.org/10.1016/j.jim.2019.112711).
- A23. Chan SF, Hejblum BP, Chakraborty A, and Cai T. Semi-supervised estimation of covariance with application to phenome-wide association studies with electronic medical records data. *Statistical Methods in Medical Research*, 29: 455–465, 2020. DOI: [10.1177/0962280219837676](https://doi.org/10.1177/0962280219837676).
- A24. Wiedemann A, Foucat E, Hocini H, Lefebvre C, Hejblum BP, Durand M, Krüger M, Keita AK, Ayouba A, Mély S, Fernandez JC, Touré A, Fourati S, Lévy-Marchal C, Raoul H, Delaporte E, Koivogui L, Thiébaud R, Lacabaratz C, Lévy Y, and PostEboGui Study Group . Long-lasting severe immune dysfunction in ebola virus disease survivors. *Nature Communications*, 11: 3730, 2020. DOI: [10.1038/s41467-020-17489-7](https://doi.org/10.1038/s41467-020-17489-7).
- A25. Bouadma L, Wiedemann A, Patrier J, Surenaud M, Wicky PH, Foucat E, Diehl JL, Hejblum BP, Sinnah F, de Montmollin E, Lacabaratz C, Thiébaud R, Timsit JF, and Lévy Y. Immune alterations during sars-cov-2-related acute respiratory distress syndrome. *Journal of Clinical Immunology*, 40: 1082–1092, 2020. DOI: [10.1007/s10875-020-00839-x](https://doi.org/10.1007/s10875-020-00839-x).

## 2019

- A26. Hejblum BP, Weber GM, Liao KP, Palmer NP, Churchill S, Shadick NA, Szolovits P, Murphy SN, Kohane IS, and Cai T. Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes. *Scientific Data*, 6: 180298, 2018b. DOI: [10.1038/sdata.2018.298](https://doi.org/10.1038/sdata.2018.298).
- A27. Hejblum BP, Alkhassim C, Gottardo R, Caron F, and Thiébaud R. Sequential dirichlet process mixture of skew t-distributions for model-based clustering of flow cytometry data. *Annals of Applied Statistics*, 13(1): 638–660, 2019. DOI: [10.1214/18-AOAS1209](https://doi.org/10.1214/18-AOAS1209).
- A28. Thiébaud R, Hejblum BP, Hocini H, Bonhabau H, Skinner J, Montes M, Lacabaratz C, Richert L, Palucka K, Banchereau J, and Levy Y. Gene expression signatures associated with immune and virological responses to therapeutic vaccination with dendritic cells in hiv-infected individuals. *Frontiers in Immunology*, 10: 874, 2019. DOI: [10.3389/fimmu.2019.00874](https://doi.org/10.3389/fimmu.2019.00874).
- A29. Ajana S, Niyazi A, Bretillon L, Hejblum BP, Jacquemin-Gadda H, and Cécile D. Benefits of dimension reduction in penalized regression methods for high dimensional grouped data: a case study in low sample size. *Bioinformatics*, 35: 3628–3634, 2019. DOI: [10.1093/bioinformatics/btz135](https://doi.org/10.1093/bioinformatics/btz135).
- A30. Low DY, Lefèvre-Arbogast S, González-Domínguez R, Urpi-Sarda M, Micheau P, Petera M, Centeno D, Durand S, Estelle P, Korosi A, Lucassen PJ, Aigner L, Proust-Lima C, Hejblum BP, Helmer C, Andres-Lacueva C, Thuret S, Samieri C, and Manach C. Diet-related metabolites associated with cognitive decline revealed by untargeted metabolomics in a prospective cohort. *Molecular Nutrition & Food Research*, 63: 1900177, 2019. DOI: [10.1002/mnfr.201900177](https://doi.org/10.1002/mnfr.201900177).

## 2018

- A31. Hejblum BP, Cui J, Lahey LJ, Cagan A, Sparks JA, Sokolove J, Cai T, and Liao KP. Association between anti-citrullinated fibrinogen antibodies and coronary artery disease in rheumatoid arthritis. *Arthritis Care & Research*, 70: 1113–1117, 2018a. DOI: [10.1002/acr.23444](https://doi.org/10.1002/acr.23444).
- A32. Commenges D, Alkhassim C, Gottardo R, Hejblum BP, and Thiébaud R. cytometree: a binary tree algorithm for automatic gating in cytometry analysis. *Cytometry: Part A*, 93(11): 1132–1140, 2018. DOI: [10.1002/cyto.a.23601](https://doi.org/10.1002/cyto.a.23601).

- A33. Neykov M, Hejblum BP, and Sinnott JA. Kernel machine score test for pathway analysis in the presence of semi-competing risks. *Statistical Methods in Medical Research*, 27(4): 1099–1114, 2018. DOI: [10.1177/0962280216653427](https://doi.org/10.1177/0962280216653427).
- A34. Sinnott JA, Cai F, Yu S, Hejblum BP, Hong C, Kohane IS, and Liao KP. PheProb: probabilistic phenotyping using diagnosis codes to improve power for genetic association studies. *Journal of the American Medical Informatics Association*, 25(10): 1359–1365, 2018. DOI: [10.1093/jamia/ocy056](https://doi.org/10.1093/jamia/ocy056).
- A35. Lefèvre-Arbogast S, Gaudout D, Bensalem J, Letenneur L, Dartigues JF, Hejblum BP, Féart C, Delcourt C, and Samieri C. Pattern of polyphenol intake and the long-term risk of dementia in older persons. *Neurology*, 2018. DOI: [10.1212/WNL.0000000000005607](https://doi.org/10.1212/WNL.0000000000005607).

## 2017

- A36. Agniel D and Hejblum BP. Variance component score test for time-course gene set analysis of longitudinal RNA-seq data. *Biostatistics*, 18(4): 589–604, 2017. DOI: [10.1093/biostatistics/kxx005](https://doi.org/10.1093/biostatistics/kxx005).
- A37. Rechten A, Richert L, Lorenzo H, Martrus G, Hejblum B, Dahlke C, Kasona R, Zinser M, Stubbe H, Matschl U, Lohse A, Krähling V, Eickmann M, Becker S, Agnandji ST, Krishna S, Kremsner PG, Brosnahan JS, Bejon P, Njuguna P, Addo MM, Becker S, Krähling V, Siegrist CA, Huttner A, Kieny MP, Moorthy V, Fast P, Savarese B, Lapujade O, Thiébaud R, Altfeld M, and Addo M. Systems Vaccinology Identifies an Early Innate Immune Signature as a Correlate of Antibody Responses to the Ebola Vaccine rVSV-ZEBOV. *Cell Reports*, 20(9): 2251–2261, 2017. DOI: [10.1016/j.celrep.2017.08.023](https://doi.org/10.1016/j.celrep.2017.08.023).
- A38. Liao KP, Sparks JA, Hejblum BP, Kuo I, Cui J, Lahey LJ, Cagan A, Gainer VS, Liu W, Cai TT, Sokolove J, and Cai T. Phenome-wide association study of autoantibodies to citrullinated and noncitrullinated epitopes in rheumatoid arthritis. *Arthritis & Rheumatology*, 69(4): 742–749, 2017. DOI: [10.1002/art.39974](https://doi.org/10.1002/art.39974).

CV

## 2016

- A39. Liquet B, De Micheaux PL, Hejblum BP, and Thiébaud R. Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics*, 32(1): 35–42, 2016. DOI: [10.1093/bioinformatics/btv535](https://doi.org/10.1093/bioinformatics/btv535).

## 2015

- A40. Hejblum BP, Skinner J, and Thiébaud R. Time-Course Gene Set Analysis for Longitudinal Gene Expression Data. *PLOS Computational Biology*, 11(6): e1004310, 2015. DOI: [10.1371/journal.pcbi.1004310](https://doi.org/10.1371/journal.pcbi.1004310). URL <http://dx.plos.org/10.1371/journal.pcbi.1004310>.

## 2014

- A41. Furman D\*, Hejblum BP\*, Simon N, Jovic V, Dekker CL, Thiébaud R, Tibshirani RJ, and Davis MM. Systems analysis of sex differences reveals an immunosuppressive role for testosterone in the response to influenza vaccination. *Proceedings of the National Academy of Sciences*, 111(2): 869–874, 2014. DOI: [10.1073/pnas.1321060111](https://doi.org/10.1073/pnas.1321060111).
- A42. Thiébaud R, Hejblum BP, and Richert L. The analysis of "Big Data" in clinical research. *Revue d'Épidémiologie et de Santé Publique*, 62(1): 1–4, 2014. DOI: [10.1016/j.respe.2013.12.021](https://doi.org/10.1016/j.respe.2013.12.021).

## 2013

- A43. Commenges D and Hejblum BP. Evidence synthesis through a degradation model applied to myocardial infarction. *Lifetime Data Analysis*, 19(1): 1–18, 2013. DOI: [10.1007/s10985-012-9227-3](https://doi.org/10.1007/s10985-012-9227-3).

## Books

- B1. Desquilbet L, Granger S, Hejblum BP, Legrand A, Pernot P, and Rougier N. *Vers une recherche reproductible : Faire évoluer ses pratiques*. Urfist de Bordeaux, 2019. ISBN 979-10-97595-05-0. URL <https://rr-france.github.io/bookrr/>.



## Preprints

- P1. Bigot J, Freulon P, Hejblum BP, and Leclaire A. On the potential benefits of entropic regularization for smoothing wasserstein estimators. *arXiv*, 2210.06934, 2022. DOI: [10.48550/arXiv.2210.06934](https://doi.org/10.48550/arXiv.2210.06934).
- P2. Gauthier M, Agniel D, Thiébaud R, and Hejblum BP. Distribution-free complex hypothesis testing for single-cell rna-seq differential expression analysis. *bioRxiv*, 2021.05.21.445165, 2021. DOI: [10.1101/2021.05.21.445165](https://doi.org/10.1101/2021.05.21.445165).
- P3. Villain L, Ferté T, Thiébaud R, and Hejblum BP. Gene set analysis for time-to-event outcome with the generalized berk-jones statistic. *bioRxiv*, 2021.09.07.459329, 2021. DOI: [10.1101/2021.09.07.459329](https://doi.org/10.1101/2021.09.07.459329).
- P4. Hejblum BP, Kunzmann K, Lavagnini E, Hutchinson A, Robertson DS, Jones SC, and Eckes-Shephard AH. Realistic and robust reproducible research for biostatistics. *Preprints*, 2020060002, 2020. DOI: [10.20944/preprints202006.0002.v1](https://doi.org/10.20944/preprints202006.0002.v1).

# Bibliography

- Ackermann M and Strimmer K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1): 47, 2009. 25
- Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, Sweeney TE, Gyang E and Shah NH. Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association*, 23(6): 1166–1173, 2016. 44
- Aghaeepour N, Finak G, Hoos H, Mosmann TR, Brinkman RR, Gottardo R and Scheuermann RH. Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, 10(3): 228–38, 2013. 14, 16, 27, 30, 36
- Agniel D and Hejblum BP. Variance component score test for time-course gene set analysis of longitudinal RNA-seq data. *Biostatistics*, 18(4): 589–604, 2017. 16, 20, 21, 25, 68, 73
- Agniel D, Hejblum BP, Thiébaud R and Parast L. Doubly-robust evaluation of high-dimensional surrogate markers. *Biostatistics*, 24(4): 985–999, 2023. 48, 51
- Anders S and Huber W. Differential expression analysis for sequence count data. *Genome Biology*, 11(10): R106, 2010. 20
- Assefa AT, De Paepe K, Everaert C, Mestdagh P, Thas O and Vandesompele J. Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA-sequencing data. *Genome Biology*, 19(1): 96, 2018. 16, 21
- Avila Cobos F, Vandesompele J, Mestdagh P and De Preter K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*, 34(11): 1969–1979, 2018. 72
- Azzalini A, Browne RP, Genton MG and McNicholas PD. On nomenclature for, and the relative merits of, two formulations of skew distributions. *Statistics and Probability Letters*, 110: 201–206, 2016. 28

## BIBLIOGRAPHY

- BD Biosciences–US. BD FACSymphony flow cytometer, 2019. <http://www.bdbiosciences.com/us/instruments/research/cell-analyzers/bd-facsymphony/m/6022968/overview>. 13
- Benjamini Y and Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57: 289–300, 1995. 11, 16, 21, 22, 24
- Blengio F, Hocini H, Richert L, Lefebvre C, Durand M, Hejblum B, Tisserand P, McLean C, Luhn K, Thiebaut R and Lévy Y. Identification of early gene expression profiles associated with long-lasting antibody responses to the Ebola vaccine Ad26. ZEBOV/MVA-BN-Filo. *Cell Reports*, 42(9): 113101, 2023. 71
- Boulesteix AL, Hoffmann S, Charlton A and Seibold H. A Replication Crisis in Methodological Research? *Significance*, 17(5): 18–21, 2020. 58, 62
- Bouveyron C, Celeux G, Murphy TB and Raftery AE. *Model-Based Clustering and Classification for Data Science: With Applications in R*, volume 50. Cambridge University Press, 2019. 28
- Brusic V, Gottardo R, Kleinstein SH, Davis MM and HIPC steering committee. Computational resources for high-dimensional immune analysis from the human immunology project consortium. *Nature Biotechnology*, 32(2): 146–148, 2014. 36
- Bucasas KL, Franco LM, Shaw CA, Bray MS, Wells JM, Niño D, Arden N, Quarles JM, Couch RB and Belmont JW. Early patterns of gene expression correlate with the humoral immune response to influenza vaccination in humans. *The Journal of infectious diseases*, 203(7): 921–929, 2011. 75
- Bühlmann P and van de Geer SA. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, 2011. 9
- Burden CJ, Qureshi SE and Wilson SR. Error estimates for the analysis of differential expression from RNA-seq count data. *PeerJ*, 2: e576, 2014. 62
- Burnham KP and Anderson DR. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2): 261–304, 2004. 29
- Chari T and Pachter L. The specious art of single-cell genomics. *PLOS Computational Biology*, 19(8): e1011288, 2023. 74
- Chén OY, Crainiceanu C, Ogburn EL, Caffo BS, Wager TD and Lindquist MA. High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, 19(2): 121–136, 2018. 75

## BIBLIOGRAPHY

- Chen T and Lumley T. Numerical evaluation of methods approximating the distribution of a large quadratic form in normal variables. *Computational Statistics & Data Analysis*, 139: 75–81, 2019. [22](#)
- Chen YT and Witten DM. Selective inference for k-means clustering. *Journal of Machine Learning Research*, 24(152): 1–41, 2023. [35](#)
- Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C and Newey W. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5): 261–265, 2017. [51](#)
- Chu T, Wang Z, Pe’er D and Danko CG. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nature Cancer*, 3(4): 505–517, 2022. [72](#)
- Clairon Q, Prague M, Planas D, Bruel T, Hocqueloux L, Prazuck T, Schwartz O, Thiébaud R and Guedj J. Modeling the kinetics of the neutralizing antibody response against SARS-CoV-2 variants after several administrations of Bnt162b2. *PLOS Computational Biology*, 19(8): e1011282, 2023. [74](#)
- Collin A, Hejblum BP, Vignals C, Lehot L, Thiébaud R, Moireau P and Prague M. Using Population Based Kalman Estimator to Model COVID-19 Epidemic in France: Estimating the Effects of Non-Pharmaceutical Interventions on the Dynamics of Epidemic. *The International Journal of Biostatistics*, page in press, 2023. [47](#)
- Commenges D, Sayyareh A, Letenneur L, Guedj J and Bar-Hen A. Estimating a difference of kullback–leibler risks using a normalized difference of aic. *The Annals of Applied Statistics*, 2(3): 1123–1142, 2008. [30](#)
- Commenges D, Alkassim C, Gottardo R, Hejblum BP and Thiébaud R. cytometree: a binary tree algorithm for automatic gating in cytometry analysis. *Cytometry: Part A*, 93(11): 1132–1140, 2018. [30](#), [38](#)
- Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X and Mortazavi A. A survey of best practices for RNA-seq data analysis. *Genome Biol*, 17(1): 13–13, 2016. [11](#)
- Corey L, Gilbert PB, Tomaras GD, Haynes BF, Pantaleo G and Fauci AS. Immune correlates of vaccine protection against hiv-1 acquisition. *Science translational medicine*, 7(310): 310rv7, 2015. [13](#)
- Cossarizza A, Chang HD, Radbruch A, Abrignani S, Addo R, Akdis M, Andrä I, Andreatta F, Annunziato F, Arranz E, Bacher P, Bari S, Barnaba V,

## BIBLIOGRAPHY

- Barros-Martins J, Baumjohann D, Beccaria CG, Bernardo D, Boardman DA, Borger J, Böttcher C, Brockmann L, Burns M, Busch DH, Cameron G, Cammarata I, Cassotta A, Chang Y, Chirido FG, Christakou E, Čičin Šain L, Cook L, Corbett AJ, Cornelis R, Cosmi L, Davey MS, De Biasi S, De Simone G, del Zotto G, Delacher M, Di Rosa F, Di Santo J, Diefenbach A, Dong J, Dörner T, Dress RJ, Dutertre CA, Eckle SBG, Eede P, Evrard M, Falk CS, Feuerer M, Fillatreau S, Fiz-Lopez A, Follo M, Foulds GA, Fröbel J, Gagliani N, Galletti G, Gangaev A, Garbi N, Garrote JA, Geginat J, Gherardin NA, Gibellini L, Ginhoux F, Godfrey DI, Gruarin P, Haftmann C, Hansmann L, Harpur CM, Hayward AC, Heine G, Hernández DC, Herrmann M, Hoelsken O, Huang Q, Huber S, Huber JE, Huehn J, Hundemer M, Hwang WYK, Iannacone M, Iverson SM, Jäck HM, Jani PK, Keller B, Kessler N, Ketelaars S, Knop L, Knopf J, Koay HF, Kobow K, Kriegsmann K, Kristyanto H, Krueger A, Kuehne JF, Kunze-Schumacher H, Kvistborg P, Kwok I, Latorre D, Lenz D, Levings MK, Lino AC, Liotta F, Long HM, Lugli E, MacDonald KN, Maggi L, Maini MK, Mair F, Manta C, Manz RA, Mashreghi MF, Mazzoni A, McCluskey J, Mei HE, Melchers F, Melzer S, Mielenz D, Monin L, Moretta L, Multhoff G, Muñoz LE, Muñoz-Ruiz M, Muscate F, Natalini A, Neumann K, Ng LG, Niedobitek A, Niemz J, Almeida LN, Notarbartolo S, Ostendorf L, Pallett LJ, Patel AA, Percin GI, Peruzzi G, Pinti M, Pockley AG, Pracht K, Prinz I, Pujol-Autonell I, Pulvirenti N, Quatrini L, Quinn KM, Radbruch H, Rhys H, Rodrigo MB, Romagnani C, Saggau C, Sakaguchi S, Sallusto F, Sanderink L, Sandrock I, Schauer C, Scheffold A, Scherer HU, Schiemann M, Schildberg FA, Schober K, Schoen J, Schuh W, Schüler T, Schulz AR, Schulz S, Schulze J, Simonetti S, Singh J, Sitnik KM, Stark R, Starossom S, Stehle C, Szelinski F, Tan L, Tarnok A, Tornack J, Tree TIM, van Beek JJP, van de Veen W, van Gisbergen K, Vasco C, Verheyden NA, von Borstel A, Ward-Hartstonge KA, Warnatz K, Waskow C, Wiedemann A, Wilharm A, Wing J, Wirz O, Wittner J, Yang JHM and Yang J. Guidelines for the use of flow cytometry and cell sorting in immunological studies (third edition). *European Journal of Immunology*, 51(12): 2708–3145, 2021. [13](#)
- Cotugno N, Santilli V, Pascucci GR, Manno EC, De Armas L, Pallikkuth S, Deodati A, Amodio D, Zangari P, Zicari S et al. Artificial intelligence applied to in vitro gene expression testing (iviget) to predict trivalent inactivated influenza vaccine immunogenicity in hiv infected children. *Frontiers in immunology*, 11: 2270, 2020. [76](#)
- Cui S, Ji T, Li J, Cheng J and Qiu J. What if we ignore the random effects when analyzing RNA-seq data in a multifactor experiment. *Statistical Applications in Genetics and Molecular Biology*, 15(2): 87–105, 2016. [25](#)

## BIBLIOGRAPHY

- Darrah PA, Patel DT, De Luca PM, Lindsay RW, Davey DF, Flynn BJ, Hoff ST, Andersen P, Reed SG, Morris SL and Roederer M. Multifunctional th1 cells define a correlate of vaccine-mediated protection against leishmania major. *Nature medicine*, 13(7): 843, 2007. 13
- De Rosa SC, Herzenberg LA, Herzenberg LA and Roederer M. 11-color, 13-parameter flow cytometry: identification of human naive T-cells by phenotype, function, and T-cell receptor diversity. *Nature medicine*, 7(2): 245, 2001. 12
- Desquilbet L, Granger S, Hejblum BP, Legrand A, Pernot P and Rougier N. *Vers une recherche reproductible : Faire évoluer ses pratiques*. Urfist de Bordeaux, 2019. <https://rr-france.github.io/bookrr/>. 55
- Díaz I, Hejazi NS, Rudolph KE and van Der Laan MJ. Nonparametric efficient causal mediation with intermediate confounders. *Biometrika*, 108(3): 627–641, 2021. 75
- Eberwine J, Sul JY, Bartfai T and Kim J. The promise of single-cell sequencing. *Nature methods*, 11(1): 25–27, 2014. 12
- Etard JF, Sow MS, Leroy S, Touré A, Taverne B, Keita AK, Msellati P, Magassouba N, Baize S, Raoul H, Izard S, Kpamou C, March L, Savane I, Barry M, Delaporte E, Ayouba A, Baize S, Bangoura K, Barry A, Barry M, Cissé M, Cissé M, Delaporte E, Delfraissy JF, Delmas C, Desclaux A, Diallo SB, Diallo MS, Diallo MS, Étard JF, Etienne C, Faye O, Fofana I, Granouillac B, Hébert EH, Izard S, Kassé D, Keita AK, Keita S, Koivogui L, Kpamou C, Lacarabartz C, Leroy S, Marchal CL, Levy Y, Magassouba N, March L, Mendiboure V, Msellati P, Niane H, Peeters M, Pers YM, Raoul H, Sacko SL, Savané I, Sow MS, Taverne B, Touré A, Traoré FA, Traoré F, Youla Y and Yazdanpanah Y. Multidisciplinary assessment of post-Ebola sequelae in Guinea (Postebogui): An observational cohort study. *The Lancet Infectious Diseases*, 17(5): 545–552, 2017. 52
- European Commission. General Data Protection Regulation (GDPR), 2018. <https://gdpr-info.eu/>. 57
- Evans C, Hardin J and Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, 19(5): 776–792, 2018. 62
- Everitt BS and Hothorn T. *A handbook of statistical analyses using R*. Chapman & Hall, Boca Raton, FL, 2006. 28

## BIBLIOGRAPHY

- Ferte T, Cossin S, Schaefferbeke T, Barnetche T, Jouhet V and Hejblum BP. Automatic phenotyping of electronic health record: Phevis algorithm. *Journal of Biomedical Informatics*, 117: 103746, 2021. [45](#)
- Ferté T, Jouhet V, Greffier R, Hejblum BP and Thiébaud R. The benefit of augmenting open data with clinical data-warehouse EHR for forecasting SARS-CoV-2 hospitalizations in Bordeaux area, France. *JAMIA open*, page ooac086, 2022. [46](#)
- Finak G, Langweiler M, Jaimes M, Malek M, Taghiyar J, Korin Y, Raddassi K, Devine L, Obermoser G, Pekalski ML, Pontikos N, Diaz A, Heck S, Villanova F, Terrazzini N, Kern F, Qian Y, Stanton R, Wang K, Brandes A, Ramey J, Aghaeepour N, Mosmann T, Scheuermann RH, Reed E, Palucka K, Pascual V, Blomberg BB, Nestle F, Nussenblatt RB, Brinkman RR, Gottardo R, Maecker H and McCoy JP. Standardizing Flow Cytometry Immunophenotyping Analysis from the Human ImmunoPhenotyping Consortium. *Scientific Reports*, 6: 20686, 2016. [36](#)
- Freedman LS, Graubard BI and Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in medicine*, 11(2): 167–178, 1992. [48](#)
- Freulon P, Bigot J and Hejblum BP. CytOpT: Optimal Transport with Domain Adaptation for Interpreting Flow Cytometry data. *Annals of Applied Statistics*, 17(2): 1086–1104, 2023. [32](#), [39](#)
- Fulwyler MJ. Electronic separation of biological cells by volume. *Science*, 150 (3698): 910–911, 1965. [13](#)
- Gao LL, Bien J and Witten DM. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, 119(545): 332–342, 2024. [34](#), [35](#)
- Gauthier M, Agniel D, Thiébaud R and Hejblum BP. dearseq: a variance component score test for rna-seq differential analysis that effectively controls the false discovery rate. *NAR Genomics and Bioinformatics*, 2(4): lqaa093, 2020. [16](#), [20](#), [21](#), [62](#), [68](#)
- Gauthier M, Agniel D, Thiébaud R and Hejblum BP. Distribution-free complex hypothesis testing for single-cell rna-seq differential expression analysis. *bioRxiv*, page 2021.05.21.445165, 2021. [20](#), [23](#)
- Germain PL, Vitriolo A, Adamo A, Laise P, Das V and Testa G. RNAon-theBENCH: Computational and empirical resources for benchmarking RNAseq

## BIBLIOGRAPHY

- quantification and differential expression methods. *Nucleic Acids Research*, 44(11): 5054–5067, 2016. [16](#), [21](#)
- Giraud C. *Introduction to High-Dimensional Statistics*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press, second edition edition, 2021. [9](#)
- Goeman JJ, van de Geer SA and van Houwelingen HC. Testing against a high dimensional alternative. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 68: 477–493, 2006. [22](#)
- Goeman JJ and Bühlmann P. Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics*, 23(8): 980–987, 2007. [25](#)
- González-Delgado J, Cortés J and Neuvial P. Post-clustering Inference under Dependency. *arXiv*, page 2310.11822, 2023. [35](#)
- Gonzalez-Dias P, Lee EK, Sorgi S, de Lima DS, Urbanski AH, Silveira EL and Nakaya HI. Methods for predicting vaccine immunogenicity and reactogenicity. *Human Vaccines & Immunotherapeutics*, 16(2): 269–276, 2020. [75](#)
- Guo J and Geng Z. Collapsibility of logistic regression coefficients. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 263–267, 1995. [51](#)
- Halpern Y, Horng S, Choi Y and Sontag D. Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association*, 23(4): 731–740, 2016. [44](#)
- Hartigan JA and Hartigan PM. The dip test of unimodality. *Annals of statistics*, 13(1): 70–84, 1985. [35](#)
- Hejblum BP, Skinner J and Thiébaud R. Time-Course Gene Set Analysis for Longitudinal Gene Expression Data. *PLOS Computational Biology*, 11(6): e1004310, 2015. [25](#), [71](#), [73](#)
- Hejblum BP, Alkhassim C, Gottardo R, Caron F and Thiébaud R. Sequential dirichlet process mixture of skew t-distributions for model-based clustering of flow cytometry data. *Annals of Applied Statistics*, 13(1): 638–660, 2019. [27](#), [29](#), [32](#), [73](#)
- Henel G and Schmitz JL. Basic theory and clinical applications of flow cytometry. *Laboratory Medicine*, 38(7): 428–436, 2007. [28](#)
- Hennig C, Meila M, Murtagh F and Rocci R. *Handbook of cluster analysis*. CRC Press, 2015. [28](#)



## BIBLIOGRAPHY

- Hicks SC, Townes FW, Teng M and Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4): 562–578, 2018. [23](#)
- Hivert B, Agniel D, Thiébaud R and Hejblum BP. Post-clustering difference testing: valid inference and practical considerations with applications to ecological and biological data. *Computational Statistics & Data Analysis*, page 107916, 2024. [34](#), [35](#), [40](#), [41](#)
- Hripcsak G and Albers DJ. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1): 117–121, 2013. [44](#)
- Hu Y, Gao L et al. Detection of deregulated modules using deregulatory linked path. *PloS one*, 8(7): e70412, 2013. [25](#)
- Huang YT and Lin X. Gene set analysis using variance component tests. *BMC Bioinformatics*, 14(1): 210–210, 2013. [19](#)
- Hunt GJ, Freytag S, Bahlo M and Gagnon-Bartsch JA. Dtangle: Accurate and robust cell type deconvolution. *Bioinformatics*, 35(12): 2093–2099, 2019. [72](#)
- Huttner A, Agnandji ST, Combescure C, Fernandes JF, Bache EB, Kabwende L, Ndungu FM, Brosnahan J, Monath TP, Lemaître B et al. Determinants of antibody persistence across doses and continents after single-dose rVSV-zebov vaccination for ebola virus disease: an observational cohort study. *The Lancet Infectious Diseases*, 18(7): 738–748, 2018. [75](#)
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822): 860–921, 2001. [9](#)
- Jin H, Wan YW and Liu Z. Comprehensive evaluation of RNA-seq quantification methods for linearity. *BMC Bioinformatics*, 18(S4): 117, 2017. [12](#)
- Joffe MM and Greene T. Related causal frameworks for surrogate outcomes. *Biometrics*, 65(2): 530–538, 2009. [48](#), [49](#)
- Johnson WE, Li C and Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1): 118–27, 2007. [61](#)
- Kalogeratos A and Likas A. Dip-means: an incremental clustering method for estimating the number of clusters. *Advances in neural information processing systems*, 25: 2393–2401, 2012. [28](#)

## BIBLIOGRAPHY

- Kennedy RB and Poland GA. The Top Five “Game Changers” in Vaccinology: Toward Rational and Directed Vaccine Development. *OMICS: A Journal of Integrative Biology*, 15(9): 533–537, 2011. [47](#)
- Kessler DC, Hoff PD and Dunson DB. Marginally Specified Priors for Non-Parametric Bayesian Estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(1): 35–58, 2015. [73](#)
- Kimes PK, Liu Y, Neil Hayes D and Marron JS. Statistical significance for hierarchical clustering. *Biometrics*, 73(3): 811–821, 2017. [28](#)
- Kriegeskorte N, Simmons WK, Bellgowan PS and Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5): 535–540, 2009. [34](#)
- Kuonen D. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, 86(4): 929–935, 1999. [22](#)
- Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerwinkler N, Mahfouz A et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1): 1–35, 2020. [12](#), [33](#)
- Law CW, Chen Y, Shi W and Smyth GK. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2): R29–R29, 2014. [12](#), [20](#), [62](#), [68](#)
- Lee EK, Nakaya HI, Yuan F, Querec TD, Burel G, Pietz FH, Benecke BA and Puelandran B. Machine learning for predicting vaccine immunogenicity. *INFORMS Journal on Applied Analytics*, 46(5): 368–390, 2016. [75](#)
- Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K and Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10): 733–739, 2010. [11](#), [59](#)
- Lhomme E, Hejblum BP, Lacabaratz C, Wiedemann A, Lelièvre JD, Lévy Y, Thiébaud R and Richert L. Analyzing cellular immunogenicity in vaccine clinical trials: a new statistical method including non-specific responses for accurate estimation of vaccine effect. *Journal of Immunological Methods*, 477: 112711, 2020. [71](#)
- Li C and Fan X. On nonparametric conditional independence tests for continuous variables. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(3): e1489, 2020. [23](#)

## BIBLIOGRAPHY

- Li S, Roupahel N, Duraisingham S, Romero-Steiner S, Presnell S, Davis C, Schmidt DS, Johnson SE, Milton A, Rajam G, Kasturi S, Carlone GM, Quinn C, Chaussabel D, Palucka AK, Mulligan MJ, Ahmed R, Stephens DS, Nakaya HI and Pulendran B. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nature Immunology*, 15(2): 195–204, 2014. [52](#)
- Li Y and Ge X. Processed datasets for differential expression analysis on population-level RNA-seq data (Version v4) [Data set]. *Zenodo*, 2022. DOI: 10.5281/zenodo.6326786. [62](#)
- Li Y, Ge X, Peng F, Li W and Li JJ. Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome Biology*, 23(1): 79, 2022. [16](#), [21](#), [59](#), [62](#), [63](#), [65](#)
- Liao KP, Sparks JA, Hejblum BP, Kuo I, Cui J, Lahey LJ, Cagan A, Gainer VS, Liu W, Cai TT, Sokolove J and Cai T. Phenome-wide association study of autoantibodies to citrullinated and noncitrullinated epitopes in rheumatoid arthritis. *Arthritis & Rheumatology*, 69(4): 742–749, 2017. [44](#)
- Lin L and Hejblum BP. Bayesian mixture models for cytometry data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13: e1535, 2021. [27](#), [29](#), [32](#), [37](#)
- Lin L, Finak G, Ushey K, Seshadri C, Hawn TR, Frahm N, Scriba TJ, Mahomed H, Hanekom W, Bart PA, Pantaleo G, Tomaras GD, Rerks-Ngarm S, Kaewkungwal J, Nitayaphan S, Pitisuttithum P, Michael NL, Kim JH, Robb ML, O’Connell RJ, Karasavvas N, Gilbert P, De Rosa SC, McElrath MJ and Gottrardo R. COMPASS identifies T-cell subsets correlated with clinical outcomes. *Nature Biotechnology*, 33(6): 610–616, 2015. [13](#)
- Lin X. Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2): 309–326, 1997. [19](#)
- Liquet B, De Micheaux PL, Hejblum BP and Thiébaud R. Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics*, 32(1): 35–42, 2016. [73](#)
- Liu Y, Hayes DN, Nobel A and Marron JS. Statistical significance of clustering for high-dimension, low-sample size data. *Journal of the American Statistical Association*, 103(483): 1281–1293, 2008. [28](#)

## BIBLIOGRAPHY

- Love MI, Huber W and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 2014. [12](#), [20](#), [62](#), [68](#)
- Lowe R, Shirley N, Bleackley M, Dolan S and Shafee T. Transcriptomics technologies. *PLOS Computational Biology*, 13(5): e1005457, 2017. [10](#)
- Lumley T. Analysis of complex survey samples. *Journal of Statistical Software*, 9(1): 1–19, 2004. R package version 2.2. [22](#)
- Maecker HT and McCoy JP. A model for harmonizing flow cytometry in clinical trials. *Nature immunology*, 11(11): 975–978, 2010. [28](#), [32](#)
- Marioni JC, Mason CE, Mane SM, Stephens M and Gilad Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9): 1509–1517, 2008. [20](#)
- Marwick B, Boettiger C and Mullen L. Packaging Data Analytical Work Reproducibly Using R (and Friends). *The American Statistician*, 72(1): 80–88, 2018. [58](#)
- Mazzoni G, Kogelman LJA, Suravajhala P and Kadarmideen HN. Systems Genetics of Complex Diseases Using RNA-Sequencing Methods. *International Journal of Bioscience, Biochemistry and Bioinformatics*, 5(4): 264–279, 2015. [16](#), [21](#)
- Morris TP, White IR and Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11): 2074–2102, 2019. [62](#)
- Neufeld A, Gao LL, Popp J, Battle A and Witten DM. Inference after latent variable estimation for single-cell RNA sequencing data. *Biostatistics*, 25(1): 270–287, 2022a. [21](#), [35](#)
- Neufeld AC, Gao LL and Witten DM. Tree-values: Selective inference for regression trees. *Journal of Machine Learning Research*, 23(305): 1–43, 2022b. [35](#)
- Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS, Luca BA, Steiner D, Diehn M and Alizadeh AA. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, 37(7): 773–782, 2019. [72](#)
- Nowicka M, Krieg C, Weber LM, Hartmann FJ, Guglietta S, Becher B, Levesque MP and Robinson MD. CyTOF workflow: Differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Research*, 6: 748, 2017. [13](#)

## BIBLIOGRAPHY

- Oberg AL, McKinney BA, Schaid DJ, Pankratz VS, Kennedy RB and Poland GA. Lessons learned in the analysis of high-dimensional data in vaccinomics. *Vaccine*, 33(40): 5262–5270, 2015. [9](#), [47](#)
- Obermoser G, Presnell S, Domico K, Xu H, Wang Y, Anguiano E, Thompson-Snipes L, Ranganathan R, Zeitner B, Bjork A, Anderson D, Speake C, Ruchaud E, Skinner J, Alsina L, Sharma M, Dutartre H, Cepika A, Israelsson E, Nguyen P, Nguyen QA, Harrod aC, Zurawski SM, Pascual V, Ueno H, Nepom GT, Quinn C, Blankenship D, Palucka K, Banchereau J and Chaussabel D. Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines. *Immunity*, 38(4): 831–844, 2013a. [47](#)
- Obermoser G, Presnell S, Domico K, Xu H, Wang Y, Anguiano E, Thompson-Snipes L, Ranganathan R, Zeitner B, Bjork A, Anderson D, Speake C, Ruchaud E, Skinner J, Alsina L, Sharma M, Dutartre H, Cepika A, Israelsson E, Nguyen P, Nguyen QA, Harrod aC, Zurawski SM, Pascual V, Ueno H, Nepom GT, Quinn C, Blankenship D, Palucka K, Banchereau J and Chaussabel D. Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines. *Immunity*, 38(4): 831–844, 2013b. [59](#), [71](#)
- OECD. *Making Open Science a Reality*. Number 25 in OECD Science, Technology and Industry Policy Papers. OECD Publishing, Paris, 2015. DOI: [10.1787/5jrs2f963zs1-en](https://doi.org/10.1787/5jrs2f963zs1-en). [57](#)
- Ozier-Lafontaine A, Fourneaux C, Durif G, Vallot C, Gandrillon O, Giraud S, Michel B and Picard F. Kernel-Based Testing for Single-Cell Differential Analysis. *arXiv*, page 2307.08509, 2023. [12](#)
- Parast L, McDermott MM and Tian L. Robust estimation of the proportion of treatment effect explained by surrogate marker information. *Statistics in medicine*, 35(10): 1637–1653, 2016. [48](#)
- Perfetto SP, Chattopadhyay PK and Roederer M. Seventeen-colour flow cytometry: unravelling the immune system. *Nature Reviews Immunology*, 4(8): 648, 2004. [12](#), [14](#)
- Phipson B and Smyth GK. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1), 2010. [22](#)
- Plotkin SA and Gilbert PB. Nomenclature for immune correlates of protection after vaccination. *Clinical Infectious Diseases*, 54(11): 1615–1617, 2012. [49](#)

## BIBLIOGRAPHY

- Poland GA and Oberg AL. Vaccinomics and bioinformatics: Accelerants for the next golden age of vaccinology. *Vaccine*, 28(20): 3509–3510, 2010. 9
- Poland GA, Ovsyannikova IG and Jacobson RM. Personalized vaccines: The emerging field of vaccinomics. *Expert Opinion on Biological Therapy*, 8(11): 1659–1667, 2008. 9
- Poland GA, Kennedy RB and Ovsyannikova IG. Vaccinomics and Personalized Vaccinology: Is Science Leading Us Toward a New Path of Directed Vaccine Development and Discovery? *PLoS Pathogens*, 7(12): e1002344, 2011. 9
- Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4): 431–440, 1989. 48
- Price BL, Gilbert PB and van der Laan MJ. Estimation of the optimal surrogate based on a randomized trial. *Biometrics*, 74(4): 1271–1281, 2018. 49
- Public Law 104-191. Health Insurance Portability and Accountability Act (HIPAA) of 1996, 1996. 57
- Pulendran B and Ahmed R. Immunological mechanisms of vaccination. *Nature Immunology*, 131(6): 509–517, 2011. 75
- Querec TD, Akondy RS, Lee EK, Cao W, Nakaya HI, Teuwen D, Pirani A, Gernert K, Deng J, Marzolf B, Kennedy K, Wu H, Bennouna S, Oluoch H, Miller J, Vencio RZ, Mulligan M, Aderem A, Ahmed R and Pulendran B. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nature Immunology*, 10(1): 116–125, 2009a. 75
- Querec TD, Akondy RS, Lee EK, Cao W, Nakaya HI, Teuwen D, Pirani A, Gernert K, Deng J, Marzolf B, Kennedy K, Wu H, Bennouna S, Oluoch H, Miller J, Vencio RZ, Mulligan M, Aderem A, Ahmed R and Pulendran B. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nature Immunology*, 10(1): 116–125, 2009b. 47
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 1997. 17
- Rahmatallah Y, Emmert-Streib F and Glazko G. Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline. *Briefings in Bioinformatics*, 17(3): 393–407, 2016. 25
- Rao S, Ghosh D, Asturias EJ and Weinberg A. What can we learn about influenza infection and vaccination from transcriptomics? *Human Vaccines & Immunotherapeutics*, 0(0): 1–9, 2019. 47

## BIBLIOGRAPHY

- Rechtien A, Richert L, Lorenzo H, Martrus G, Hejblum B, Dahlke C, Kasonta R, Zinser M, Stubbe H, Matschl U, Lohse A, Krähling V, Eickmann M, Becker S, Agnandji ST, Krishna S, Kreamsner PG, Brosnahan JS, Bejon P, Njuguna P, Addo MM, Becker S, Krähling V, Siegrist CA, Huttner A, Kiény MP, Moorthy V, Fast P, Savarese B, Lapujade O, Thiébaut R, Altfeld M and Addo M. Systems Vaccinology Identifies an Early Innate Immune Signature as a Correlate of Antibody Responses to the Ebola Vaccine rVSV-ZEBOV. *Cell Reports*, 20(9): 2251–2261, 2017. [52](#), [71](#), [75](#)
- Redko I, Courty N, Flamary R and Tuia D. Optimal transport for multi-source domain adaptation under target shift. In Chaudhuri K and Sugiyama M, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 849–858, 2019. [33](#)
- Richert L, Lelièvre JD, Lacabaratz C, Hardel L, Hocini H, Wiedemann A, Lucht F, Poizot-Martin I, Bauduin C, Diallo A, Rieux V, Durand M, Hejblum BP, Lounay O, Thiébaut R, Lévy Y and on behalf of the ANRS VRI01 Study group. T-cell immunogenicity, gene expression profile and safety of four heterologous prime-boost combinations of hiv vaccine candidates in healthy volunteers - results of the randomized multi-arm phase i/ii anrs vri01 trial. *Journal of Immunology*, 208(12): 2663–2674, 2022. [71](#), [76](#)
- Rigaill G, Balzergue S, Brunaud V, Blondet E, Rau A, Rogier O, Caius J, Maugis-Rabusseau C, Soubigou-Taconnat L, Aubourg S, Lurin C, Martin-Magniette ML and Delannoy E. Synthetic data sets for the identification of key ingredients for RNA-seq differential analysis. *Briefings in bioinformatics*, bbw092: 1–12, 2016. [16](#), [21](#)
- Rinchai D, Deola S, Zoppoli G, Ahamed Kabeer BS, Taleb S, Pavlovski I, Maacha S, Gentilcore G, Toufiq M, Mathew L, Liu L, Vempalli FR, Mubarak G, Lorenz S, Sivieri I, Cirmena G, Dentone C, Cuccarolo P, Giacobbe D, Baldi F, Garbarino A, Cigolini B, Cremonesi P, Bedognetti M, Ballestrero A, Bassetti M, Hejblum BP, Augustine T, Van Panhuys N, Thiébaut R, Branco R, Chew T, Shojaei M, Short K, Feng C, PREDICT-19 consortium, Zughaiier SM, De Maria A, Tang B, Ait Hssain A, Bedognetti D, Grivel JC and Chaussabel D. High-temporal resolution profiling reveals distinct immune trajectories following the first and second doses of COVID-19 mRNA vaccines. *Science Advances*, 8(45): eabp9961, 2022. [59](#), [71](#), [73](#)
- Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A and Smyth GK. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23(20): 2700–2707, 2007. [61](#)

## BIBLIOGRAPHY

- Robinson MD, McCarthy DJ and Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1): 139–140, 2010. [12](#), [20](#), [62](#), [68](#)
- Rocke DM, Ruan L, Zhang Y, Gossett JJ, Durbin-Johnson B and Aviran S. Excess false positive rates in methods for differential gene expression analysis using rna-seq data. *bioRxiv*, page 020784, 2015. [16](#), [21](#), [62](#)
- Roederer M, Brenchley JM, Betts MR and De Rosa SC. Flow cytometric analysis of vaccine responses: how many colors are enough? *Clinical immunology*, 110(3): 199–205, 2004. [14](#)
- Rozenendaal R, Hendriks J, Van Effelterre T, Spiessens B, Dekking L, Solfrosi L, Czapska-Casey D, Bockstal V, Stoop J, Splinter D, Janssen S, Baelen BV, Verbruggen N, Serroyen J, Dekeyster E, Volkmann A, Wollmann Y, Carrion R, Giavedoni LD, Robinson C, Leyssen M, Douoguih M, Luhn K, Pau MG, Sadoff J, Vandebosch A, Schuitemaker H, Zahn R and Callendret B. Nonhuman primate to human immunobridging to infer the protective effect of an Ebola virus vaccine candidate. *npj Vaccines*, 5(1): 112, 2020. [52](#)
- Saeyns Y, Van Gassen S and Lambrecht BN. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, 16(7): 449–463, 2016. [30](#)
- Saxena A, Dagur PK and Biancotto A. Multiparametric flow cytometry analysis of naïve, memory, and effector t cells. In *Immunophenotyping*, pages 129–140. Springer, 2019. [40](#)
- Seshadri C, Lin L, Scriba TJ, Peterson G, Freidrich D, Frahm N, DeRosa SC, Moody DB, Prandi J, Gilleron M and Mahomed H. T-cell responses against mycobacterial lipids and proteins are poorly correlated in south african adolescents. *The Journal of Immunology*, 195(10): 4595–4603, 2015. [13](#)
- Shapiro HM. *Practical flow cytometry*. John Wiley & Sons, 2005. [13](#)
- Sharpe C, Davis J, Mason K, Tam C, Ritchie D and Koldej R. Comparison of gene expression and flow cytometry for immune profiling in chronic lymphocytic leukaemia. *Journal of Immunological Methods*, 463: 97–104, 2018. [72](#)
- Shen-Orr SS and Gaujoux R. Computational deconvolution: Extracting cell type-specific information from heterogeneous samples. *Current Opinion in Immunology*, 25(5): 571–578, 2013. [72](#)



- Shi W, Oshlack A and Smyth GK. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Research*, 38(22): e204–e204, 2010. 11
- Siddiqui S and Livák F. Principles of advanced flow cytometry: A practical guide. In Bosselut R and Vacchio MS, editors, *T-Cell Development: Methods and Protocols*, pages 89–114. Springer US, 2023. 13
- Siffer A, Fouque PA, Termier A and Largouët C. Are your data gathered? In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2210–2218, 2018. 28
- Sinnott JA, Cai F, Yu S, Hejblum BP, Hong C, Kohane IS and Liao KP. PheProb: probabilistic phenotyping using diagnosis codes to improve power for genetic association studies. *Journal of the American Medical Informatics Association*, 25(10): 1359–1365, 2018. 44
- Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. 11
- Soneson C and Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4): 255–261, 2018. 12, 22
- Song Y, Zhou X, Zhang M, Zhao W, Liu Y, Kardia SL, Roux AVD, Needham BL, Smith JA and Mukherjee B. Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *Biometrics*, 2018. 75
- Srivastava A, Malik L, Sarkar H, Zakeri M, Almodaresi F, Sonesson C, Love MI, Kingsford C and Patro R. Alignment and mapping methodology influence transcript abundance estimation. *Genome Biology*, 21(1): 239, 2020. 12
- Stein D, O’Connor D, Blohmke C, Sadarangani M and Pollard A. Gene expression profiles are different in venous and capillary blood: Implications for vaccine studies. *Vaccine*, 34(44): 5306–5313, 2016. 59, 61
- Steinbach M, Ertöz L and Kumar V. The challenges of clustering high dimensional data. In *New directions in statistical physics*, pages 273–309. Springer, 2004. 28, 59
- Stubington MJ, Rozenblatt-Rosen O, Regev A and Teichmann SA. Single-cell transcriptomics to explore the immune system in health and disease. *Science*, 358(6359): 58–63, 2017. 12

## BIBLIOGRAPHY

- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES and Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43): 15545–15550, 2005. 25
- Sullivan NJ, Martin JE, Graham BS and Nabel GJ. Correlates of protective immunity for ebola vaccines: implications for regulatory approval by the animal rule. *Nature Reviews Microbiology*, 7(5): 393–400, 2009. 52
- Svensson V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, 38(2): 147–150, 2020. 23
- Tarazona S, Furió-Tarí P, Turrà D, Pietro AD, Nueda MJ, Ferrer A and Conesa A. Data quality aware analysis of differential expression in rna-seq with noiseq r/bioc package. *Nucleic Acids Research*, 43(21): e140–e140, 2015. 62, 68
- Taylor JM, Wang Y and Thiébaud R. Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics*, 61(4): 1102–1111, 2005. 48
- The Turing Way Community. *The Turing Way: A handbook for reproducible, ethical and collaborative research*. Zenodo, 2022. DOI: [10.5281/zenodo.3233853](https://doi.org/10.5281/zenodo.3233853). <https://book.the-turing-way.org/>. 56
- The Turing Way Community and Scriberia. Illustrations from the turing way book dashes, 2019. DOI: [10.5281/zenodo.3332808](https://doi.org/10.5281/zenodo.3332808). 56
- Thiébaud R, Hejblum BP and Richert L. The analysis of "Big Data" in clinical research. *Revue d'Épidémiologie et de Santé Publique*, 62(1): 1–4, 2014. 9
- Thiébaud R, Hejblum BP, Hocini H, Bonnabau H, Skinner J, Montes M, Lacabaratz C, Richert L, Palucka K, Banchereau J and Levy Y. Gene expression signatures associated with immune and virological responses to therapeutic vaccination with dendritic cells in hiv-infected individuals. *Frontiers in Immunology*, 10: 874, 2019. 71
- Townes FW, Hicks SC, Aryee MJ and Irizarry RA. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, 20(1): 295, 2019. 23
- Van der Laan MJ, Polley EC and Hubbard AE. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007. 51

## BIBLIOGRAPHY

- Van Mechelen I, Boulesteix AL, Dangl R, Dean N, Hennig C, Leisch F, Steinley D and Warrens MJ. A white paper on good research practices in benchmarking: The case of cluster analysis. *WIREs Data Mining and Knowledge Discovery*, 13(6): e1511, 2023. 62
- VanderWeele TJ. Surrogate measures and consistent surrogates. *Biometrics*, 69(3): 561–565, 2013. 49
- Vandewalle P, Kovacevic J and Vetterli M. Reproducible research in signal processing. *IEEE Signal Processing Magazine*, 26(3): 37–47, 2009. 55, 57
- Wagholikar KB, Estiri H, Murphy M and Murphy SN. Polar labeling: Silver standard algorithm for training disease classifiers. *Bioinformatics*, 36(10): 3200–3206, 2020. 44
- Wang S, McCormick TH and Leek JT. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48): 30266–30275, 2020a. 75
- Wang T, Li B, Nelson CE and Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC bioinformatics*, 20(1): 40, 2019. 12
- Wang X, Parast L, Tian L and Cai T. Model-free approach to quantifying the proportion of treatment effect explained by a surrogate marker. *Biometrika*, 107(1): 107–122, 2020b. 75, 76
- Wang X, Parast L, Han L, Tian L and Cai T. Robust approach to combining multiple markers to improve surrogacy. *Biometrics*, 79(2): 788–798, 2023. 76
- Wasserman L. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer-Verlag, New York, 2006. 20
- Weber LM, Saelens W, Cannoodt R, Sonesson C, Hapfelmeier A, Gardner PP, Boulesteix AL, Saeys Y and Robinson MD. Essential guidelines for computational method benchmarking. *Genome Biology*, 20(1): 125, 2019. 61
- Weiner J, Domaszewska T, Donkor S, Kaufmann SHE, Hill PC and Sutherland JS. Changes in transcript, metabolite and antibody reactivity during the early protective immune response in humans to Mycobacterium tuberculosis infection. *Clinical Infectious Diseases*, page ciz785, 2019. 47
- Wiedemann A, Foucat E, Hocini H, Lefebvre C, Hejblum BP, Durand M, Krüger M, Keita AK, Ayoub A, Mély S, Fernandez JC, Touré A, Fourati S,

## BIBLIOGRAPHY

- Lévy-Marchal C, Raoul H, Delaporte E, Koivogui L, Thiébaud R, Lacabaratz C, Lévy Y and PostEboGui Study Group. Long-lasting severe immune dysfunction in ebola virus disease survivors. *Nature Communications*, 11: 3730, 2020. 52
- World Health Organization. Correlates of vaccine-induced protection: methods and implications. Technical report, World Health Organization, 2013. 75
- Xia F and Chan KCG. Identification, semiparametric efficiency, and quadruply robust estimation in mediation analysis with treatment-induced confounding. *Journal of the American Statistical Association*, just-accepted: 1–28, 2021. 75
- Xu R and Wunsch D. *Clustering*, volume 10. John Wiley & Sons, 2008. 28
- Yi H, Plotkin A and Stanley N. Benchmarking differential abundance methods for finding condition-specific prototypical cells in multi-sample single-cell datasets. *Genome Biology*, 25(1): 9, 2024. 12
- Yu S, Ma Y, Gronsbell J, Cai T, Ananthakrishnan AN, Gainer VS, Churchill SE, Szolovits P, Murphy SN, Kohane IS, Liao KP and Cai T. Enabling phenotypic big data with PheNorm. *Journal of the American Medical Informatics Association*, 25(1): 54–60, 2018. 44
- Zhang H, Zheng Y, Zhang Z, Gao T, Joyce B, Yoon G, Zhang W, Schwartz J, Just A, Colicino E et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20): 3150–3154, 2016. 75
- Zhang HG, Hejblum BP, Weber G, Palmer N, Churchill S, Szolovits P, Murphy S, Liao K, Kohane I and Cai T. Atlas: An automated association test using probabilistically linked health records with application to genetic studies. *Journal of the American Medical Informatics Association*, 28(12): 2582–2592, 2021. 44
- Zhao Y and Luo X. Pathway Lasso: Pathway estimation and selection with high-dimensional mediators. *Statistics and Its Interface*, 15(1): 39–50, 2022. 75
- Zhong W, Spracklen CN, Mohlke KL, Zheng X, Fine J and Li Y. Multi-snp mediation intersection-union test. *Bioinformatics*, 35(22): 4724–4729, 2019. 75
- Zhou RR, Wang L and Zhao SD. Estimation and inference for the indirect effect in high-dimensional linear mediation models. *Biometrika*, 107(3): 573–589, 2020. 75
- Zou H and Hastie T. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2): 301–320, 2005. 45

## BIBLIOGRAPHY

# Glossary & acronyms

## Glossary

**Bioconductor:** The Bioconductor project focuses on the analysis of genomic data and hosts many R packages <https://bioconductor.org/>. 17, 58, 59, 76

**CRAN:** The Comprehensive R Archive Network <https://cran.r-project.org/>. 17, 58, 59

**EM algorithm:** Expectation-Maximization algorithm. 30

**read:** an RNA sequence of various length (usually 100 or 150 bases long) read by a sequencer. 11

**RNA-seq:** “Next-generation” sequencing of RNA (ribonucleic acid) molecules from a biological sample to quantify gene expression, i.e. its transcriptome. 9–12, 15–17, 20–22, 59, 62, 71, 73

**scRNA-seq:** single-cell RNA-seq, that provide gene expression measurement at the cell resolution, as opposed to bulk RNA-seq. 9, 12, 13, 16, 17, 21–23, 26, 33, 73, 74

## Acronyms

**AIC:** Akaike Information Criteria 30, 31

**CCDF:** conditional cumulative distribution function 23, 24

**CDF:** cumulative distribution function 23

**CIT:** Conditional Independence Test 23

**CLT:** Central Limit Theorem 22

- COVID-19:** Coronavirus disease 2019 4, 45
- CyTOF:** Cytometry by Time-Of-Flight 13, 31, 73
- DE:** Differentially Expressed 20, 21, 63–65, 67
- DEA:** Differential Expression Analysis 10–12, 16, 17, 19–23, 61, 62, 65, 68, 73, 74
- DNA:** Desoxyribonucleic Acid 9, 11, 47
- EHR:** Electronic Health Records 17, 43–47
- FCM:** Flow Cytometry 13–17, 27–33, 36–39, 71–74
- FDR:** False Discovery rate 16, 21, 22, 62, 63, 65, 68
- FlowCAP:** *Flow Cytometry: Critical Assessment of Population Identification Methods* project 31, 36
- GDPR:** General Data Protection Regulation 57
- GSA:** Gene Set Analysis 25
- HIPAA:** Health Insurance Portability and Accountability Act 57
- HIPC:** Human Immunology Project Consortium 31, 36–38
- i.i.d.:** independent and identically distributed 48
- ICD-10:** 10<sup>th</sup> revision of International Classification of Diseases published by the World Health Organisation initially in 1999 44, 45
- mRNA:** messenger Ribonucleic Acid 11, 47
- NLP:** Natural Language Processing 45
- OECD:** Organisation for Economic Co-operation and Development 57
- OLS:** Ordinary Least Squares 22
- OT:** Optimal Transport 32, 33
- PBMC:** Peripheral Blood Mononuclear Cells 36

## GLOSSARY & ACRONYMS

**PCA:** Principal Component Analysis 61

**PTE:** Proportion of Treatment effect Explained 48, 50, 51, 54

**RNA:** Ribonucleic Acid 9, 11, 59, 71

**RT-PCR:** Reverse Transcriptase Polymerase Chain Reaction 45, 47

**SARS-CoV-2:** Severe Acute Respiratory Syndrome Corona Virus 2 45, 47

**SNP:** Single Nucleotide Polymorphism 9

**VRI:** Vaccine Research Institute 15, 71