U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

2.º CICLO

FCUP
2020

U.PORTO

Multiple Testing in Proteomics

Afonso João Rodrigues dos Santos Castro Videira

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Multiple Testing in Proteomics

Afonso João Rodrigues dos Santos Castro Videira

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Bioinformática e Biologia Computacional

2020

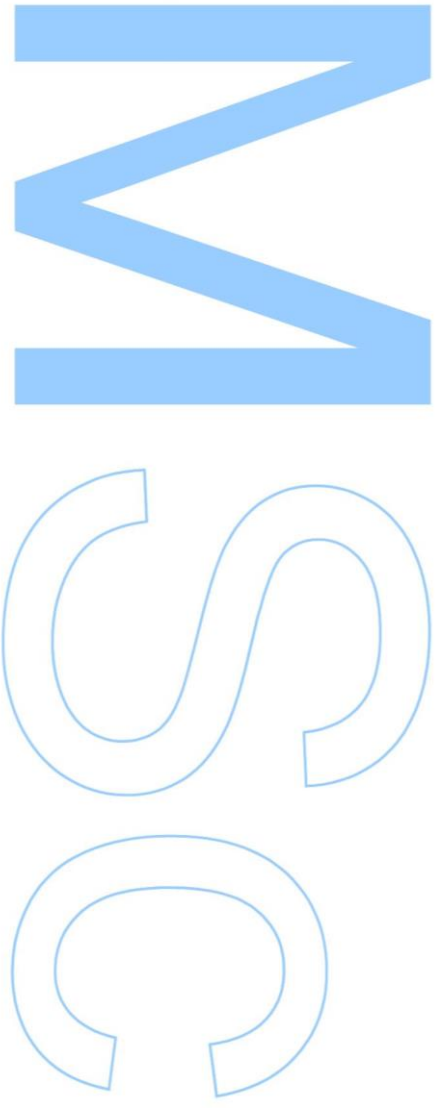
Multiple Testing in Proteomics

Afonso João Rodrigues dos Santos Castro Videira
Mestrado em Bioinformática e Biologia Computacional
Departamento de Biologia | Departamento de Ciência de Computadores
2020

Orientador
Ana Rita Pires Gaio, Professor Auxiliar, Departamento de Matemática, Faculdade de Ciências da Universidade do Porto

Coorientador
Hugo Alexandre de Carvalho Pinheiro Osório, Investigador no i3S / Ipatimup, Professor Afiliado, Faculdade de Medicina da Universidade do Porto

Coorientador
Margarida Maria Araújo Brito, Professor Associado, Departamento de Matemática, Faculdade de Ciências da Universidade do Porto



Todas as correções determinadas pelo júri, e só essas, foram efetuadas.
O Presidente do Júri,

Porto, _____/_____/_____

Abstract

The field of proteomics has seen an ever-increasing development technology wise. Novel high-resolution mass-spectrometry techniques in conjunction with sophisticated computational methods propelled the analysis of both targeted and global proteomes, resulting in a high-throughput of protein data. Given this amount of biological data provided by the technological advancements, powerful statistical tools are needed to handle the downstream analysis and extract information without major deficits.

Hypothesis testing has been a cornerstone in statistical analysis of proteomic data, as it allows to screen for relevant proteins, an important step, for example, when one wants to do biomarkers discovery. However, hypothesis testing faces a challenge in proteomics: the large amount of proteins in the downstream analysis, leading to the performance of multiple tests.

The core problem in multiple hypothesis testing can be traced to type I errors, i.e., testing several hypotheses at the same significance level does not control the error of false discoveries to that significance level. Consequently, this translates into a large amount of false discoveries. To tackle the multiplicity problem with hypothesis testing, two recent methodologies in the field of proteomics were scrutinized. The Binomial sequential goodness of fit metatest (SGoF) and its conservative variation. Through a simulation study, Binomial SGoF showed an overall better power score, when compared to other more widespread adjustment methods, while also keeping the false discovery rate at similar levels.

To ascertain the role of these methodologies in proteomic analysis, their performance in a real case study was evaluated. Results were obtained from a developed R routine. In a first screening, using one-way ANOVA, the two SGoF methods performed very well, with perfect power scores and relatively low false discovery rates. In the second screening, the pair-wise comparisons, their performance in power and false discovery rate deteriorated, when compared to the first screening and to other conventional adjustment methods.

Keywords: Proteomics, Statistics, Bioinformatics, Hypothesis test, Multiple comparison correction methods

Resumo

A área da proteómica tem assistido a um constante aumento no desenvolvimento tecnológico. Novas técnicas de espectrometria de massa de alta resolução em conjunto com métodos computacionais sofisticados, impulsionaram a análise de proteomas, resultando num high-throughput de dados de proteínas. Dada esta quantidade de dados biológicos, fornecidos pelas novas tecnologias, torna-se necessário a utilização de ferramentas estatísticas poderosas, que consigam lidar com a análise de dados e extrair informações sem grandes perdas.

Os testes de hipóteses têm sido um pilar na análise estatística de dados de proteómica. Eles permitem fazer a deteção de proteínas relevantes, um passo importante na descoberta de biomarcadores, por exemplo. Contudo, os testes de hipóteses enfrentam um desafio na proteómica: a elevada quantidade de proteínas na análise de dados, o que leva à necessidade de se fazerem testes múltiplos.

No núcleo do problema dos testes múltiplos de comparações estão os erros do tipo I, i.e., quando testamos várias hipóteses com o mesmo nível de significância, o erro cometido com as falsas rejeições não fica contido para aquele nível de significância. Consequentemente, isto traduz-se numa grande quantidade de falsas descobertas. Para contrariar este problema da multiplicidade, duas metodologias recentes na área da proteómica foram estudadas. O Binomial sequential goodness of fit metatest (SGoF) e a uma versão mais conservativa. Através de um estudo de simulação, o Binomial SGoF mostrou ser, no geral, o método com melhor potência estatística, quando comparado com outras metodologias convencionais, conseguindo manter, no entanto, a taxa de falsas descobertas (FDR) a níveis similares aos outros métodos.

Para averiguar o papel destas metodologias numa análise de proteómica, os seus desempenhos foram avaliados num caso de estudo real. Os resultados foram obtidos por um conjunto de procedimentos desenvolvidos em R. Numa primeira triagem, usando ANOVA de um fator, os dois métodos SGoF mostraram um ótimo desempenho, obtiveram o resultado máximo para a potência estatística e mostraram um FDR relativamente baixo. Na segunda triagem, efetuadas as comparações dois a dois entre grupos, tanto a potência estatística como o FDR dos métodos SGoF se deterioraram, quando comparados aos resultados obtidos na primeira triagem, bem como aos resultados de outros métodos convencionais de ajustamento

Palavras-Chave: Proteómica, Estatística, Bioinformática, Testes de hipóteses, Métodos de correção de testes múltiplos

Agradecimentos

Aos meus orientadores, Professora Rita Gaio, Professora Margarida Brito e Professor Hugo Osório, que me deram um grande apoio e demonstraram enorme paciência. Sem vocês eu não teria conseguido concretizar esta etapa tão importante da minha vida. Estarei para sempre agradecido.

Aos meus amigos e colegas de curso pois, sem a vossa amizade e o vosso companheirismo, estes dois anos não teriam terminado tão rápido. Obrigado por terem estado presentes.

Aos meus pais, Afonso e Eugénia, aos meus irmãos, Nuno, Joana, Martim e Laura, e à minha namorada Lídia, pelo vosso amor e apoio incondicional nesta jornada. Sem a vossa ajuda nada disto teria sido possível. Obrigado por serem uma presença constante na minha vida, obrigado por serem a minha rocha.

Contents

Abstract i
Resumo ii
Agradecimentos iii
Contents iv
List of Figures vi
List of Tables vii
Acronyms ix
Introduction 1
Chapter I – Multiple hypothesis testing 5
Hypothesis testing 5
Errors in hypothesis tests 6
Misinterpretation of hypothesis tests 7
Example of a hypothesis test (one sample t-test) 7
Multiplicity problem 9
Family Wise Error Rate controlling procedures 10
False Discovery Rate 12
Properties of the False Discovery Rate 12
False Discovery Rate controlling procedures 13
The need for a more balancing method in multitest adjustments 14
Chapter II – Sequential Goodness of Fit metatests 15
The mixture model 15
Binomial Sequential Goodness of Fit metatest 15
Binomial SGoF algorithm 18
Conservative SGoF variation for testing the number of effects 19
Simulation study of Binomial SGoF using R programming language 20
Chapter III – Workflow in Proteomics 25
Introduction to Proteomics 25
How do proteins work? 29

Protein analysis techniques 31
Mass spectrometry instrumentation 32
Tandem mass spectrometry (MS/MS) 37
Protein sample preparation 38
Computational methods for MS-based proteomic data 40
Chapter IV – R routine for differential expression analysis and case study 48
Introductory consideration 48
Implementation of procedures in downstream analysis of LFQ proteomics 48
Case Study and Data Acquisition 51
Chapter V – Final Remarks 59
Conclusion 59
Bibliography: 60
Supplemental files 66

List of Figures

Fig 1: Distribution of the test statistic T, with 29 degrees of freedom. Highlighted in grey is the rejection region at a 5% level. 8
Fig 2: Transcription and translation, in highlight is the process of RNA splicing. Adapted from [7, 8]..... 26
Fig 3: Unidirectional flow of information, Central Dogma of Molecular Biology. Adapted from [3]..... 27
Fig 4: Different structures of protein folding. Withdrawn from [2]..... 30
Fig 5: Principle of electro spray ionization. Extracted from [5]..... 33
Fig 6: Illustration of the quadrupole configuration between the source and exit slits. Extracted from [4]..... 35
Fig 7: Transversal cut of an orbitrap along the zr-plane. Ions are moving between the spindle electrode (a) and the outer electrodes (b_1, b_2). The frequency of oscillation of ions along the z-axis can be determined via FT and converted to a m/z. Adapted from [6]. 36
Fig 8: Laboratorial workflow of a proteomic study. Adapted from [1]. 40
Fig 9: Frequency of peptides with different lengths in mouse DB..... 43
Fig 10: Total number of proteins per sample, after applying the initial missing value filter. The horizontal line represents the total number of proteins in the dataset (2620). As this is before missing value imputation, every sample has some of its protein's values missing. 54
Fig 11: Bar plot of the evaluation metrics grouped by the 6 adjustment methods applied to the ANOVA p-values. BH, Bonf, BS, BY, CS and S being respectively Benjamini-Hochberg procedure, Bonferroni correction, Binomial SGoF, Conservative SGoF, ,Benjamini–Yekutieli procedure and Šidák correction. 55
Fig 12: Bar plot of the evaluation metrics grouped by the 6 adjustment methods applied to the contrasts p-values. BH, Bonf, BS, BY, CS and S being respectively Benjamini-Hochberg procedure, Bonferroni correction, Binomial SGoF, Conservative SGoF, ,Benjamini–Yekutieli procedure and Šidák correction. 57

List of Tables

Table 1: Possible outcomes for a single hypothesis test, the keywords positive and negative correspond to rejection and non rejection respectively, as mentioned in the text. Adapted from [29]..... 6

Table 2: Example of proteins and their function in the organism. Adapted from [36]..... 28

Table 3: *Summary of study design, showing 4 samples, each containing 200 ng yeast tryptic digest spiked with the indicated amount of tryptic digest of 6 individual proteins (units in fmols).Adapted from [25].* .....**Erro! Marcador não definido.**

Table 4: Mean percentages of significant cases deteced when the null hypothesis was always true. The simulated null models were tested under t tests. n: sample sizes. S: number of tests. Significant %: Percentage of significant tests at 5% significance. Detect_Bonf %: Percentage of tests corrected by Bonferroni and detected at 5% significance. Detect_BH %: Percentage of tests corrected by Benjamini-Hochberg and detected at 5% significance. Detect_SGoF %: Percentage of tests corrected by Binomial SGoF and detected at 5% significance. Values are avereges through 1000 replicates and their ± stardart deviations..... 66

Table 5: Percentages of significant cases detected at 5% significance (Detect_method), false discovery rate (FDR_method) and Power (Power_methd) after multitest adjustment when the p-values come from families of one-sample t tests where some (% effect) of the alternative hypotheses were true.The alternative hypothesis comes from a N(0.36,1). n: sample size. Prevalence %: Percentage of true alternatives. S: number of tests. Significant %: Percentage of significants before adjustment at 5% significance. Detect_method %: Percentage of significance tests detected after adjustments (Bonferroni, Benjamini-Hochberg, Binomial SGoF) at 5% significance. FDR_method %: Percentage of false discovery rate (Bonferroni, Benjamini-Hochberg and Binomial SGoF). Power_method %: Percentage of statistical power (Bonferroni, Benjamini-Hochberg and Binomial SGoF). Values are avereges through 1000 replicates and their ± stardart deviations. 67

Table 6: Percentages of significant cases detected at 5% significance (Detect_method), false discovery rate (FDR_method) and Power (Power_methd) after multitest adjustment when the p-values come from families of one-sample t tests where some (% effect) of the alternative hypotheses were true.The alternative hypothesis comes from a N(0.7,1). n: sample size. Prevalence %: Percentage of true alternatives. S: number of tests. Significant %: Percentage of significants before adjustment at 5% significance. Detect_method %: Percentage of significance tests detected after adjustments (Bonferroni, Benjamini-Hochberg, Binomial SGoF) at 5% significance. FDR_method %: Percentage of false discovery rate (Bonferroni, Benjamini-Hochberg and Binomial SGoF). Power_method %: Percentage of statistical power

(Bonferroni, Benjamini-Hochberg and Binomial SGoF). Values are avereges through 1000 replicates and their ± stardart deviations. 69
Table 7: Percentages of significant cases detected at 5% significance (Detect_method), false discovery rate (FDR_method) and Power (Power_methd) after multitest adjustment when the p-values come from families of one-sample t tests where some (% effect) of the alternative hypotheses were true.The alternative hypothesis comes from a N(0.96,1). n: sample size. Prevalence %: Percentage of true alternatives. S: number of tests. Significant %: Percentage of significants before adjustment at 5% significance. Detect_method %: Percentage of significance tests detected after adjustments (Bonferroni, Benjamini-Hochberg, Binomial SGoF) at 5% significance. FDR_method %: Percentage of false discovery rate (Bonferroni, Benjamini-Hochberg and Binomial SGoF). Power_method %: Percentage of statistical power (Bonferroni, Benjamini-Hochberg and Binomial SGoF). Values are avereges through 1000 replicates and their ± stardart deviations. 71
Table 8: Confusion matrices (top) and corresponding evaluation matrices (bottom) of the ANOVA (left) and the contrasts (right), both corrected by Binomial SGoF (BS). 73
Table 9: Confusion matrices (top) and corresponding evaluation matrices (bottom) of the ANOVA (left) and the contrasts (right), both corrected by Conservative SGoF (CS). 73
Table 10: Confusion matrices (top) and corresponding evaluation matrices (bottom) of the ANOVA (left) and the contrasts (right), both corrected by Benjamini-Hochberg (BH)..... 74
Table 11: Confusion matrices (top) and corresponding evaluation matrices (bottom) of the ANOVA (left) and the contrasts (right), both corrected by Benjamini–Yekutieli (BY). 74
Table 12: Confusion matrices (top) and corresponding evaluation matrices (bottom) of the ANOVA (left) and the contrasts (right), both corrected by Bonferroni (Bonf). 75
Table 13: Confusion matrices (top) and corresponding evaluation matrices (bottom) of the ANOVA (left) and the contrasts (right), both corrected by Šidák (S). 75

Acronyms

ABRF – Association of biomolecular resource facilities

AM – Accurate mass

ANOVA - Analysis of variance

APCI – Atmospheric pressure chemical ionization

DC – Direct current

DNA – Deoxyribonucleic acid

ELISA – Enzyme-Linked Immunosorbent Assay

ESI – Electrospray ionization

FDR – False discovery rate

FN – False negatives

FP – False positives

FT – Fourier Transformation

FWEC – Family-wise error complete

FWEP – Family-wise error partial

FWER – Family-wise error rate

HCD – High-energy collisions

HPLC – High performance liquid chromatography

HR – High resolution

iBAQ – Intensity-based absolute quantification

iPRG – Proteomic informatics research group

iTRAQ – Isobaric tags for relative and absolute quantitation

LFQ – Label free quantification

MALDI – Matrix-assisted laser desorption/ionization

mRNA – Messenger Ribonucleic acid

MS – Mass spectrometry

PSM – Peptide spectrum match

PTMs – Posttranslational modifications

RF – Radiofrequency

RNA – Ribonucleic acid

SGoF – Sequential goodness of fit

SILAC – Stable isotope labeling using amino acids in cell culture

STP – Simultaneous Test Procedure

TMT – Tandem mass tags

TN – True negatives

TOF – Time-of-flight

TP – True positives

Introduction

Proteomics refers to the study of the full set of proteins of a given organism, quantitative and qualitative wise. As proteins are the effectors of variation in the genome and transcriptome, their detection can provide fundamental information towards the description of a biological system [9-11]. While the genome can be seen as the blueprint information to construct and maintain an organism, the proteome corresponds to the flow of that information within the organism. It shows, in the form of proteins, how genetic information is differentially expressed throughout the cells, tissues, organs and across time [11, 12]. In the context of a genetic disease, even though the gene alteration can give a good description of the illness, it is the protein changes that will define the onset of the disease [13], thus the importance of studying proteins. Therefore, protein study is the next logical step of research, given the central dogma of molecular biology, which states that the genetic information in living organisms flows from their deoxyribonucleic acid (DNA) to their ribonucleic acid (RNA) and finally translates into proteins [10, 14].

As of today, proteomics has seen a giant leap technology wise; high-throughput methods like mass-spectrometry (MS) have propelled the analysis of both targeted and global proteomes, leading to a large scale protein identification [10-12].

From a biological perspective, MS-based proteomics gives three distinct types of experimental results. Firstly, expression proteomics determines the relative and absolute amount of proteins in a sample [12, 13]. Secondly, MS allows for the identification of modifications in proteins, which includes post-translational modifications [12]. Thirdly, MS proteomics is well suited to identify protein interactions [12]. In order to cope with the amount of data provided by MS proteomic experiments, some computational tools have been developed to aid the work of researchers. One of these computational tools commonly used in proteomics laboratories is MaxQuant. It consists of a multitude of algorithms, capable of processing the raw data that come out of liquid chromatography (LC)-MS runs [15]. It supports different labeling techniques, such as MS_1-level and isobaric MS_2-level labeling procedures, as well as free label techniques such as label free quantification (LFQ) [13, 15, 16]. The software also has its own search engine (Andromeda), which allows for peptide identification [15]. When the software finishes analyzing the raw data, it outputs several tables containing qualitative and quantitative information about the proteins found [15].

After the process of data acquisition, a downstream analysis has to be performed, extracting meaningful information about the proteins and their relationship with the study subject (e.g. association between protein expression and a given disease). This type of

analysis is performed mainly by statistical tools, one of them being Perseus, a software developed by the same researchers that built MaxQuant, which integrates several statistical packages that allow researchers to analyse and extract knowledge from data [17].

The motivation of this study lies in the downstream analysis of the data, more precisely, in the multiple hypothesis-testing problem, as it impacts most of the statistical models used for detection of differential expression between conditions (a critical step in a general proteomics study) [18].

When conducting multiple hypotheses tests if the same rejection rule is to be followed for each test, the resulting probability of making at least one Type I error is much larger than the nominal level used for each test. Moreover, this error probability also increases with the number of tests performed [19, 20]. The first developed adjustment methods focused on controlling the family-wise error rate (FWER), which is the probability of making at least one false rejection. The Bonferroni procedure is a FWER controlling method in which the nominal significance of each test must be an equal portion of the total significance [19, 21]. An upgrade to this method was developed by Holm (1979); it follows the same principle as Bonferroni's, but it requires a sequential order of the p-values and their respective hypothesis. The p-values are then compared with a nominal significance portion that sequentially increases, which results in a more powerful method than Bonferroni [19, 22]. Historically, after these FWER methods, a more recent approach to the multiple tests correction focused on controlling the false discovery rate (FDR), which is the expected proportion of rejected hypotheses that have been wrongly rejected. Following this line, Benjamini and Hochberg (1995) introduced an algorithm for the identification of the hypotheses to be rejected while maintaining the FDR equal to or below a predefined level [19, 23]. More recently (2009 and 2011), Uña Alvarez and colleagues came up with two different approaches for controlling Type I errors in multiple testing scenarios, the Binomial sequential goodness of fit metatest (SGoF) and its Conservative counterpart. These are Binomial tests for the expected proportion of rejections under the null hypothesis. Whenever the observed proportion shows a meaningful difference from the expected proportion, this difference reflects the number of hypotheses that can be rejected (sequentially, from the lowest p-value to the highest) [24, 25].

The problem with multiple hypothesis-testing in proteomics can be traced to some challenges when conducting a proteomic study [20].

The first challenge comes from the high peptide/protein throughput by MS experiments. The current mass spectrometers used in proteomics allow for a good peptide transmission, which means that they can analyze most peptides within a sample. If the number of peptides analyzed is large, so it will be the number of proteins, inferred computationally. This high protein throughput will result in many entries/rows in the protein data set [16]. A single global proteomic study might identify and quantify more than 1000 proteins, assuming that the

prevalence of effects in the data is 0%, the expectation is not to find differential expression between groups, but applying a simple hypothesis test at 5% significance level to all comparisons, the expectation is for the test to find 50 differential expressed proteins (100% FDR), a correction to the significance level must be applied in order to control the type I errors (if not corrected, the probability finding a false positive increases exponentially with the number of tests performed) [20].

The second challenge comes from the low prevalence of differential expressed entities and their relatively low effect size, when compared between conditions. Perhaps the key factor determining the usefulness of multiple testing corrections and the final FDR is the percentage of proteins showing a true effect. This percentage plays a similar role to the prevalence of differentially expressed proteins in different proteome conditions. With no true effects, any multiple testing correction method, however strict, will work well, as long as it minimizes the number of false positives. At the other extreme, when all entities have an underlying true effect, multiple testing corrections are detrimental as they reduce the true positives without lowering the FDR. Multiple testing corrections become vital as proteomic data tens to have a low prevalence of effects, many methodologies, especially LFQ requires the samples to be very similar, including those from different conditions [13, 16]. Considering the same data (described in the first challenge), but with a 10% prevalence of effects (100 proteins are differentially expressed), applying a correction method to the tests may exclude some true discoveries and impair further analyses by pathway or biological process enrichment [17, 20]. Regarding the low effect size, some quantification methods, like MS_2-level labeling, might produce ratio compression intensities, these ratio compressions effectively lower the effect size in protein intensities between groups [13, 16, 20]. This leads to higher p-values generated by hypothesis testing because lower differences yield higher p-values. Consequently, any p-value correction will increase the size of the p-values, meaning that fewer proteins pass the corrected thresholds [20].

The third challenge comes from a monetary perspective. As proteomic experiments do not come cheap, the repercussions are reflected in many ways throughout the workflow. One common example is the low sample size of the data, which in statistical terms means a lack of statistical support of the evidence, leading to inflated p-values and lower discoveries. The monetary constraints might not be limited to low sample size, for example, conducting a label-based study is more expensive than label-free, but choosing label-free because of the budget might have consequences in the viability of the data (if the study requires label-based methods) [13, 20].

The fourth and final challenge, facing statistical analysis of proteomic data lies in hypothesis testing. Most hypothesis tests work under the assumption that the data is normally distributed and shows homoscedasticity. However, in many cases, proteomic data does not

have these properties. One example is the situation of ratio compression in MS_2-level labeling. The departure from normality may impact the resulting p-values (coming from the tests where the assumptions are not met), consequently impacting any correction procedure [20].

In spite of the fact that several methodologies have already been established for the statistical analysis of proteomes, the underlying large-scale nature of the proteomics data sets still continue to pose major opportunities and challenges in statistics [20, 21]. These challenges in the workflow of proteomic studies make the development and application of prime methods to correct multiple tests procedures a must have when conducting downstream analysis.

Both Binomial SGoF and Conservative SGoF methods promise to be well-balanced multitest correcting methods. They will complement the already existing analysis tools applied in proteomics, by lowering false discoveries while keeping true ones [24, 25].

This dissertation is laid out in the following way.

In Chapter I, the concept of hypothesis tests is discussed, as well as the problem that arises from performing multiple hypothesis tests and the existing methodologies developed to counteract the multiple hypothesis test problem.

Chapter II is dedicated to the discussion of SGoF methods, a novelty in downstream proteomic analysis, and to the comparison of these methods against older ones through a series of simulation studies, with results and discussion.

In Chapter III, the proteomics workflow is presented, with an introduction to this area of research, its importance, and the existing tools that allow researchers to start from a case study and end up with the interesting data.

Chapter IV is dedicated to the presentation of an implemented R routine, containing the novel adjustment methods. An evaluation case study designed by the Proteome Informatics Research Group (iPRG), of the Association of Biomolecular Resource Facilities (ABRF) is also presented and analyzed, using the developed R routine. The results are showed and discussed.

Chapter V is dedicated to the final remarks of the dissertation study as well as some considerations to take into account for future studies with similar scopes.

Chapter I – Multiple hypothesis testing

Hypothesis testing

By itself data has little to no value. It is in its interpretation that interest relies on. We want to be able to provide answers to questions about the data and correctly interpret the results, and statistics provide us with the tools to do just that. Statistical inference provides a level of confidence and likelihood in decision making, thus, it helps researchers to make decisions, when conducting hypothesis tests [26].

In the following section, we will refer only to parametric hypothesis tests.

In statistics, the concept of hypothesis can be thought of as a statement about a population parameter (θ) which can be tested through experimentation [21, 26, 27]. For a given parameter θ in the parameter space Θ, the concept of null and alternative hypothesis is as follows:

- **Null hypothesis (H_0)**, states that the population parameter belongs to a certain subset of the parameter space, i.e., H_0: $\theta \in \Theta_0$, $\Theta_0 \subset \Theta$.
- **Alternative hypothesis (H_1)**, states that the population parameter is in the complementary space of that, specified in the null hypothesis, i.e., H_1: $\theta \in \Theta_0^c$.

The goal of a hypothesis test is to evaluate the veracity of the null hypothesis. In order to do so, first, the area for which $\mathbf{H_0}$ is rejected must be defined. This area, called rejection region, it is set pre experimentation by a probability value **α**, called significance level. This significance level denotes the probability of rejecting the null hypothesis, assuming that $\mathbf{H_0}$ **is true** (the significance level is the rejection region for $\mathbf{H_0}$). To perform an hypothesis test, a **test statistic** must be used [21, 26-28]. This is a random variable with a known distribution under $\mathbf{H_0}$. For the given significance level, and using the distribution, it is possible to calculate the critical value(s), defining the boundaries of the rejection region. After meeting all the test requirements, the hypothesis can now be tested on the sample. Then the test statistic is computed on the sample with the purpose summarizing the data.

A decision can be taken based on the following rule:

- **Value of the test statistic** $\notin$ **rejection region**: There is not enough evidence to reject the null hypothesis, fail to reject $\mathbf{H_0}$.
- **Value of the test statistic** $\in$ **rejection region**: The evidence points to the null hypothesis being false, reject $\mathbf{H_0}$ in favor of $\mathbf{H_1}$.

Another approach for either rejecting or not $\mathbf{H_0}$ consists of using the probability value (**p-value**). The **p-value** is the probability of obtaining a test result at least as extreme as the

value of the test statistic observed in the sample, assuming that the null hypothesis is true. This probability can easily be calculated given the known distribution of the test statistic and can be interpreted directly using the significance level [21, 27].

- **p-value > α:** There is not enough evidence to reject the null hypothesis; fail to reject $\mathbf{H_0}$.
- **p-value ≤ α:** The evidence points to the null hypothesis being false; reject $\mathbf{H_0}$ in favor of $\mathbf{H_1}$.

Errors in hypothesis tests

As already mentioned, the outcome of a statistical hypothesis test is bound to a probabilistic nature, in the sense that the test result may suggest one thing, when in reality the opposite is true [19-21, 24, 27, 28]. In hypothesis testing, there are 4 possible situations:

- **true rejections**, H_0 is false and the test correctly rejects H_0;
- **true non rejections**, H_0 is not false and the test does not reject H_0;
- **false rejections,** H_0 is not false and the test points to a rejection of H_0;
- **false non rejections**, H_0 is false and the test does not reject H_0.

The previous enumeration can be seen in table 1, where the relation between the test results and the real situation is depicted.

Table 1: *Possible outcomes for a single hypothesis test, the keywords positive and negative correspond to rejection and non rejection respectively, as mentioned in the text. Adapted from [29].*

		Reality	
		Positive	Negative
Study finding	Positive	True positive (Power) (1-β)	False positive **Type I Error** (α)
	Negative	False negative **Type II Error** (β)	True negative (1-α)

In this dissertation we will be focusing more on the **Type I errors** and ways to control it; the power and consequently the **Type II errors** may be briefly touched upon.

Misinterpretation of hypothesis tests

The probabilistic nature of hypothesis tests makes the interpretation of the result subject to error.

Rejecting H_0 means that there is enough statistical evidence in the data that the likelihood of the null hypothesis being true is very small accordingly to the significance level [27, 28].

On the other hand, due to the dichotomy of the hypothesis statement, we may be tempted to accept the null hypothesis, but that is **not correct**. Instead it **is correct** to say we have failed to reject the null hypothesis [27, 28].

Example of a hypothesis test (one sample t-test)

Let's assume that a team of scientists wants to test a new drug to see if it has any effect on systolic blood pressure (**sbp**), or no effect at all.

Assume that from previous studies, it is known that the average **sbp** in the population is **127** mmHg and that **sbp** follows a normal distribution.

A sample of 30 participants who have taken the medication presents a mean of **110** mmHg with a standard deviation of **10** mmHg. Does this new drug affect the **sbp**?

1. **Define null and alternative hypotheses.**
   H_0: μ = 127
   H_1: μ ≠127

2. **Define the significance level.**
   α = 0.05

3. **Evaluate the test statistic in the sample.**

   In this case, the sample distribution will follow a t-student distribution:

$$T = \frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)} \sim t(n-1)$$

   $\bar{x}$ is the sample mean, μ_0 is the population mean assuming H_0, s is the sample standard deviation and n is the sample size.

In this particular case, the test statistic follows a t-student distribution with 29 degrees of freedom (**df**).

df = n – 1 = 29

Knowing that:

$\bar{x}$ = 110, μ_0 = 127, s = 10, n = 30

Plugging in every value into the equation of the test statistic, gives **t** = -9.31.

4. State the decision rule (define the rejection region).

Knowing that the test statistic will follow a **t-student distribution with 29 df**, and given the significance level, we can look up the **critical values** for the rejection region, shown in figure1. We are looking for values very different from the population mean, either greater or lower (bilateral test).

Critical value $=^{+}_{-}$ 2.0452

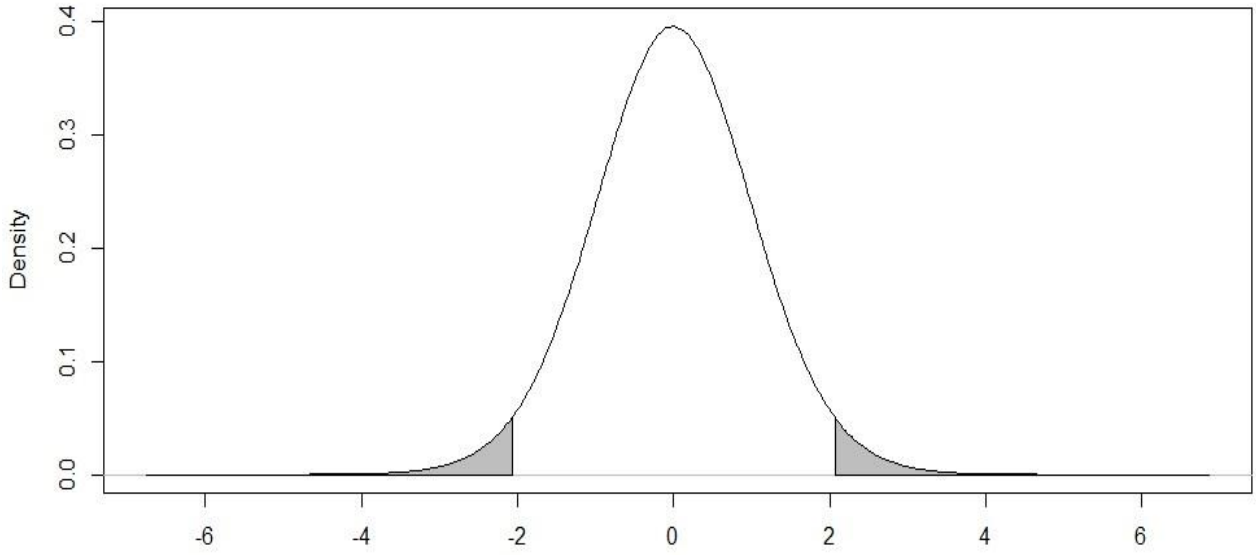


Fig 1: *Distribution of the test statistic T, with 29 degrees of freedom. Highlighted in grey is the rejection region at a 5% level.*

5. Calculate the p-value.

The p-value for the observed test statistic is much lower than **0.0001**.

6. **State the results.**

We previously stated that **Critical value** $= \pm$**2.0452**, which means that **|t|**$\geq$**2.0452** defines the rejection region. So, with an observed **t = -9.31,** which has an absolute value of 9.31 >> 2.0452, the decision is to **reject H_0**, at a 5% significance level. As expected, this is corroborated by a p-value < 0.05.

7. **State the conclusion (what does this mean?).**

The new drug significantly affects the mean systolic blood pressure; in this case, the sample mean suggests that it is lower than 127 mmHg.

Multiplicity problem

As stated in the previous topic, the testing of hypothesis is not without errors. Regarding the **Type I error** in a single test, the probability of incorrectly rejecting the null hypothesis when it is true (False positive) is at most **α**. Even tho in this case the underlying error is being mitigated by the quantity of tests (only one test was performed, so **α** is capped at its designated value), the exact opposite occurs when conducting several independent hypothesis tests, the scalability of this **Type I error** increases tremendously, **α** is no longer maintained at its pre-defined level for that set of hypotheses tested [19, 21, 27, 28].

It is now easy to see the connection between this aspect of randomness of the statistical tests and the way analysis is done in large data sets, as mentioned in the introduction. We arrive at the multiplicity problem, which, in summary, states that the error of wrongly rejecting the null hypothesis increases with the number of hypothesis tests performed [19, 20, 24, 30]. For example, if we conduct one hypothesis test with α = 0.05, assuming that the null hypothesis is true, the probability of making at least one false rejection (in this case there is only one test) is $1 - (1 - 0.05)^1 = 0.05$, in this case the false rejection is contained at α. But, if we conduct 100 hypothesis tests at the same α level, the probability of making at least one false rejection, assuming that all nulls are true, is $1 - (1 - 0.05)^{100} \approx 1$. So, it's almost certain that we will commit a **Type I error,** we failed to contain the error at α.

Concerning the statistical analysis of a data set, and eventually the execution of hypothesis tests, researchers should take into account the following statements, from Young [30], about false positives:

1. With enough testing, false positives will occur.
2. Internal evidence will not contradict a false positive result.
3. Good investigators will come up with a possible explanation.
4. They only happen to the others.

About the third and fourth statement, researchers tend to overestimate their data, in the sense that any conclusions drawn from it must be right. This will lead researchers to formulate an explanation for said results, which is wrong [30].

In order to mitigate the multiplicity problem, one must discuss the intricacies of errors associated to multiple testing. Concerning the control of the **Type I error** when performing multiple hypothesis tests, there is the family-wise error rate (FWER) and the false discovery rate (FDR) [19, 21, 30].

Family Wise Error Rate controlling procedures

We will start by discussing the Family wise Error Rate (FWER), which was first to appear in literature, chronologically speaking. In 1987, Hochberg and Tamhane defined a method, called ***Simultaneous Test Procedure*** (STP), which should be used for any collection of inferences (set of hypotheses to be tested) for which it is meaningful to take into account a combined measure of error (a kind of family error) [30]. Two kinds of family-wise errors were defined:

- **FWEC**, being the probability of rejecting at least one of the hypothesis tested, when the whole set of hypotheses is true.

  $$FWEC = Pr(\text{Reject at least one } H_i \mid \text{all } H_i \text{ are True})$$

- **FWEP,** being the probability of rejecting at least one out of a subset of true hypotheses.

  $$FWEP = Pr(\text{Reject at least one } H_i,\ i = j_1,...,j_n \mid H_{j1},...H_{jn} \text{ are true})$$

The most used definition for FWER in statistics is, the probability of making one or more false discoveries when performing multiple hypotheses tests.

An STP that controls the FWER in the weak sense, guarantees an error rate up to the significance level **α** only when all the null hypotheses are true. An STP controlling the FWER in the strong sense, guarantees an error no bigger than **α** for any configuration of the true and false null hypotheses [30].

Some procedures that control these false discoveries have been developed over the last centuries.

One of the first, and still widely used procedure is the **Bonferroni correction method**, which can be applied whether the tests are independent or not, as stated by the Boole inequality [19, 21]. Given a set of events A_1, A_2, ... A_n, the probability of at least one of them happening is **no greater** than the sum of the marginal probabilities of the events, i.e., $P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$. With that in mind, the Bonferroni procedure states that, in order to reject an individual hypothesis from the full set, their p-value must be less than or equal to a proportion of the global significance level, with respect to the number of tests performed. In other words, given a hypothesis H_i with its p-value p_i, out of a set of n Hypotheses: **reject H_i, if $\boldsymbol{p_i} \leq \frac{\boldsymbol{a}}{\boldsymbol{n}}$** [19, 21, 23].

The **Sidák procedure** is another method like Bonferroni, but only controls the FWER if the tests are independent. Taking into account the notation used in the Bonferroni procedure, each hypothesis H_i of the set is tested at a level α-new = $1 - (1 - \alpha)^{\frac{1}{n}}$. The hypotheses should be rejected if their p-values are **not greater** than α-new [19, 21].

Another similar approach to the multiplicity problem was developed by Holm (1979) [22]; it is a step-up iterative process, as described below:

1. Order the indices 1, …, n by their corresponding p-values, from lowest to highest. Get H_1, …, H_n , and their ordered p-values p_1, … , p_n.
2. Let **R** be the largest **k**, such that $p_k \leq \frac{\alpha}{n + 1 - \boldsymbol{k}}$.
3. Reject the null hypothesis H_1, ..., $H_{\boldsymbol{R}}$. If **R** = 0, then none of the hypotheses are to be rejected.

Later in 1986, Simes extended the Boole inequality and stated that, for an ordered (lowest to greatest) set of p-values, p_1, … , p_n, corresponding to independent continuous tests, and assuming that all hypotheses are true, we have :

$$Prob\left(p_i \geq \frac{i \times a}{n}\right) = 1 - a$$

Where α is the significance level. Using this last inequality, Simes created a simple multiple test method to maintain an error rate at a predefined level [19, 23]. Reject all the null hypotheses H_i in the set, for which:

$$p_i \leq \frac{\boldsymbol{i} \times \boldsymbol{a}}{\boldsymbol{n}}$$

In 1988 Hochberg developed a step-up procedure similar to holm's.

A key aspect to retain about these classical procedures controlling FWER, is that they have substantially less power (false negatives, they miss on true discoveries) than the per comparison procedures (individual tests) of the same level [19, 20].

False Discovery Rate

The False Discovery Rate (FDR) was first described by Benjamini and Hochberg in 1995. It is a less conservative/more powerful (relative to FWER controlling procedures) approach for identification of the non random effects in multiple testing [19, 23, 24].

Consider the outcomes depicted in table 1 but in the multiple testing scenario. Given a set of **n** hypotheses to be tested, a **false positive** and a **true positive** correspond to a **wrongly rejected hypothesis (F)** and a **correctly rejected hypothesis respectively (T)**.

Given a set of hypotheses and a given testing procedure, the number of rejections is a known variable **R**, and by extension of the underlying truthfulness of the hypothesis **R** must be equal to the sum of **F and T, R = F + T** [19, 24]**.**

If the problem is the number of false rejections, as it is in the FWER case, it is logical to look at the problem now, not in the full set of hypotheses tested, but at the set of rejections, more specifically, at the proportion of wrong rejections in relation to the number of rejections $\left(FDR = \frac{F}{R} \right)$.

As mentioned, **F** and **T** are both unobserved random variables, and **FDR** is defined to be the expectation of $\frac{F}{F+T}$, $E\left[\frac{F}{F+T}\right]$ [23, 24].

Properties of the False Discovery Rate

Some relations between FWER and FDR can be established:

- For a given set where all null hypotheses are true, the FDR is equivalent to the FWER; in this particular situation **T = 0** and **F = R**. Thus if **F = 0** then FDR = 0, or, if **F > 0** then FDR = 1, which implies that the probability of having at least one false rejection is equal to FDR [19, 23].
- In the case where the number of **true null hypothesis** is smaller than the **total number of hypotheses** tested, the FDR will be smaller than or equal to the FWER [19, 23]. Here, if **F > 0** then $\frac{F}{R} \leq 1$. Thus, methods that control FWER will also control the FDR.

False Discovery Rate controlling procedures

As mentioned, Benjamini and Hochberg (1995) were the first to come up with this new way of approaching the multiplicity problem. Besides introducing FDR, they proposed a step-up method (all FDR controlling procedures comprise an iterative step over the sorted p-values) for controlling the FDR at a certain significance level α [19, 23].

As already stated, consider testing a set of **n** hypotheses, and let them be sorted by their resulting p-values in a **descending order**, $p_1 \geq p_2, \ldots \geq p_n$. Then, let **k** be the smallest index (**i**) for which $\boldsymbol{p_i} \leq \frac{i \times \alpha}{n}$. Reject all H_i from **k** to **n** (recall that as it is in descending order, these hypotheses will have smaller p-values) [19, 23]. This procedure is valid when the tests are independent and also in various scenarios of dependence, but it is not universally valid.

Benjamini and Yekutieli (2001) described an FDR controlling procedure under arbitrary dependence assumptions, similar to the Benjamini and Hochberg method. It follows a similar arrangement as depicted above, where **k** is the smallest integer such that:

$$p_k \leq \frac{k \times \alpha}{n \times c(n)}$$

Here c(n) = 1 , if the tests are independent or positively correlated; this is the special case where the method is equal to the one described by Benjamini and Hochberg [23, 31].

If the dependencies are random, then c(n) = $\sum_{i=1}^{n} \frac{1}{i}$.

In the case where the tests are negative correlated, c(n) can be approximated by the Euler-Mascheroni constant.

The need for a more balancing method in multitest adjustments

Multitest adjustment methods have gained attention since the appearance of high-dimensional biological data, related to the fast development of the “omic” technologies. These data driven biological studies tend to produce thousands of comparisons at a given time, thus, there is an obvious interest to use multitest adjustment techniques, and even more to know which methods are more appropriate [20, 24, 32].

As it was already stated, FDR controlling procedures tend to be more powerful than FWER controlling methods, which is expected, because any procedure that controls for FWER also controls the FDR, therefore being more stringent with Type I errors. In any case, a transversal feature of any multitest controlling procedure is the constant loss of statistical power, when the number of comparisons tests increases [19-21, 24]. This occurs because the techniques adjust each individual test error rate according to the number of tests performed (the way these methods do this adjustment is what differs between them). The inherent drawback of this adjustment is that the higher the number of hypothesis tests, the lower the chance of detecting even one significant case [24].

In an ideal world, any multitest correction should show a large statistical power and a small FDR under a small number of comparisons, and its statistical power should increase when the number of tests also increases, as it currently occurs in relation to sample size [24].

More recently (2009), a new method to address the multiplicity problem was proposed. The Sequential Goodness of Fit metatest (SGoF).

SGoF was designed to counter the high number of tests and the low power paradigm that afflicts other adjustment techniques.

Chapter II – Sequential Goodness of Fit metatests

The mixture model

When conducting multiple tests it is safe to assume that the n p-values corresponding to each hypothesis tested, $H_1 \ldots H_n$, constitute a random sample from a certain mixture distribution function (df) [25, 33]. More precisely this df is $F(x) = \pi_0 F_0(x) + (1 - \pi_0) F_1(x)$, where π_0 is the proportion of true nulls and F_0 is the df of the p-values from the true null hypotheses, and F_1 corresponds to the unknown df of the p-values obtained from the non true null hypotheses. In simple terms the model sets a statistical independence between the hypothesis tested, and consequently between their p-values [25, 32, 33].

Let p_i be a p-value attached to a certain null hypothesis H_{0i}, for which we do not know the origin of its distribution. The probability of H_{0i} being true is π_0. If we assume that H_{0i} is true, then p_i must be uniformly distributed between 0 and 1. As opposed, for a false null hypothesis, the corresponding p-value must follow the unknown df F_1, as stated by the authors [25].

Binomial Sequential Goodness of Fit metatest

Consider now the previously discussed mixture model and let F_n be the experimental df of the p-values and γ be an initial significance level [25, 32]. Under the complete null hypothesis, i.e., $H_0 = \bigcap_{i=1}^{n} H_{0i}$, the empirical p-values should fit well to a uniform distribution f_0, thus providing an expected proportion for the p-values that fall below the γ level, this being the $\int_0^{\gamma} f_0 dp$ or, in other terms, if F_0 corresponds to the cumulative df of f_0, then the expected proportion is just $F_0(\gamma) = \gamma$. If within F_n there is some proportion of non true nulls, i.e., some of the null hypotheses are false, it is expected that the proportion $F(\gamma)$ will be greater than γ [25].

Given this mixture concept, the SGoF performs a one-sided binomial test for the proportion at level **α** where:

- $H_0(\gamma)$: $F(\gamma) = \gamma$. Given an empirical distribution function of the p-values at level γ, the observed proportion of p-values $\leq \gamma$ is equal to γ.
- $H_1(\gamma)$: $F(\gamma) > \gamma$. Given an empirical distribution function of the p-values at level γ, the observed proportion of p-values $\leq \gamma$ is greater than γ.

In this one-sided test, the rejection region at level **α** is given by a critical value $b_{n,a}(\gamma)$, which represents the minimum significant number (i.e, greater number than the expected, given H_0, plus a buffer zone) of p-values that are lower than the γ threshold. The critical value is defined by $P(Bin(n,\gamma)) \geq b$, this last inequality must be lower than or equal to **α,** thus the respective quantile will be the critical value $b_{n,a}(\gamma)$ [25, 32].

As the number of n (number of p-values) grows, this binomial distribution for the proportion can be approximated by a normal distribution, $Bin(n,\gamma) \approx N(n\gamma, n\gamma(1-\gamma))$. This means that $b_{n,a}(\gamma)$ can be approximated by

$$b_{n,a}(\gamma) \approx n\gamma + \sqrt{n\gamma(1-\gamma)}z_{\alpha}$$

Where $z_a$stands for the (1 - **α**) quantile of the standard normal [25].

Considering the previous approximation, we see that $H_0$is rejected if $nF_n(\gamma) > b_{n,a}(\gamma)$. Using the previous approximation,

$$\frac{nF_n(\gamma)}{n} > \frac{n\gamma + \sqrt{n\gamma(1-\gamma)}\,z_{\alpha}}{n}$$

<=>

$$F_n(\gamma) > \gamma + \sqrt{\frac{\gamma(1-\gamma)}{n}}z_a$$

<=>

$$F_n(\gamma) - \gamma > \sqrt{\frac{\gamma(1-\gamma)}{n}}z_a$$

<=>

$$\frac{F_n(\gamma) - \gamma}{\sqrt{\frac{\gamma(1-\gamma)}{n}}} > z_a$$

This last equation being the asymptotic formulation of a one-sided test for the proportion.

Rejection of H_0 (i.e., rejection of the complete null hypothesis) means that there is at least one false rejection among the n hypotheses tested [24, 25].

The SGoF meta-test figures that the number of non-true null hypotheses is given by the excess number of rejections (significance cases) with respect to the expected number at level γ. This excess number is given by $N_a(\gamma) = nF_n(\gamma) - b_{n,a}(\gamma) + 1$. This means that the

$N_a(\gamma)$ smallest p-values are declared significant by the test. In the special case where only the hypotheses, which p-values $\leq$ **α** are to be considered as candidates for effects, the following correction must be applied $N_a^c(\gamma) = \min[N_a(\gamma), nF_n(a)]$ [25]. In this situation, the p-values will always be less than or equal to **α** without breaking the critical value condition [25].

The threshold p-value, denoted by $p^*_{n,a}(\gamma)$, is the p-value for which $N_a(\gamma) = nF_n(p^*_{n,a}(\gamma))$. To determine this p-value we can use the inverse function of the empirical mixture model, which gives the p-value as a function of its proportion. As we know that the threshold proportion (i.e., the proportion of significant cases) is just $F_n(\gamma) - n^{-1}\, b_{n,a}(\gamma) + n^{-1}$. Then the $p^*_{n,a}(\gamma)$ is,

$$p^*_{n,a}(\gamma) = F_n^{-1}\left(F_n(\gamma) - n^{-1} b_{n,a}(\gamma) + n^{-1}\right)$$

Using the approximation for the critical value we have that,

$$p^*_{n,a}(\gamma) \approx F_n^{-1}(F_n(\gamma) - \gamma - \sqrt{\frac{\gamma\,(1-\gamma)}{n}}\,\boldsymbol{z_a} + \frac{1}{n})$$

Since **α** < 1, we have $b_{n,a}(\gamma) \geq 1$ and since F_n^{-1} is a nondecreasing function, the threshold p-value $p^*_{n,a}(\gamma)$ is always less than or equal to γ. This provides that only p-values $\leq \gamma$ may enter the set of hypotheses declared significant by SGoF [32].

It is worth noting that, as n grows to infinity, the threshold p-value $p^*_{n,a}(\gamma)$ approaches to $F_n^{-1}(F_n(\gamma) - \gamma)$. This is because, as we have access to an infinitely large n, our knowledge of the mixture model $F_n(\gamma)$ becomes perfect, and with that the influence of the significance level drops. This property of SGoF directly relates to its power (capacity of making true discoveries) and it is found in the negative term $-\sqrt{\frac{\gamma\,(1-\gamma)}{n}}\,\boldsymbol{z_a}$, which, for a growing value of n, the threshold p-value $p^*_{n,a}(\gamma)$ also grows, leading to a higher count of significant cases [25, 32].

The significance level **α** is controlling the FWER of SGoF meta-test in the weak sense, that is, under the complete null $H_0 = \bigcap_{i=1}^{n} H_{0i}$. With that, SGoF is also controlling the FDR, since under H_0, FWER coincides with FDR [24, 25, 32, 34].

SGoF is based on the interesting principle that there is some sort of connection between the amount $F_n(\gamma) - \gamma$ and the proportion of non true nulls with p-values less than γ, i.e., $P_1(\gamma) = P(H_{0i} = 1, p_i \leq \gamma)$, where H_{0i}= 1 means that H_{0i} is a non true null. Note that $P_1(\gamma)$ is the proportion of the mixture model corresponding to the non true nulls distribution, so $P_1(\gamma) = (1 - \pi_0)F_1(\gamma) = F(\gamma) - \pi_0\gamma \geq F(\gamma) - \gamma$, which means that $F(\gamma) - \gamma$ is a lower bound for $P_1(\gamma)$. Thus, as n grows and since $N_a(\gamma) \approx n(F(\gamma) - \gamma)$, the number of

significant effects detected by SGoF can be considered as a lower bound for the number of true effects below the initial threshold γ [25].

Binomial SGoF algorithm

Consider testing a set of S independent comparisons at a certain significance level γ, with their respective null hypotheses $H_1, H_2 \ldots H_S$ and let $p_1 \leq p_2 \leq \ldots \leq p_S$ be the ordered p-values associated to each test, matching H_i with every P_i . Let K be the observed number of rejections after individually testing all S hypotheses at an γ level. Given that the S null hypotheses are true, the expected number of rejections (in this case false rejections) is given by $E = S * \alpha$.

The SGoF algorithm works as follows [24]:

1. Input: A list of S sorted p-values, from minor to major.
2. Set variable: R = K, to represent the number of p-values below the significance level γ;
3. Loop: Perform a one-sided binomial test at a significance level **α** between the R observed rejections and the Expected rejections.
    a. If the test is significant, count a new significant, then decrease by one the number R of observed rejections and consequently increase the number of non rejections in order to keep S constant. Jump to 3 and repeat.
    b. If the test is not significant: Stop and go to 4.
4. Output the number of new significant findings detected in steps (3, 3.a).
5. Match the output number of significant findings to the number of hypotheses with the lowest p-values, which should be rejected by SGoF.

This algorithm can be improved and the loop removed, by simply calculating the critical value (C). Indeed we have the number of trials S and we have the proportion of success γ, thus the binomial distribution is known, and at a significance level **α** so is the critical value.

The number of new significant findings will just be equal to $K - C + 1$. Then just proceed to step 5 of the algorithm.

Conservative SGoF variation for testing the number of effects

Given the following null and alternative hypotheses:

- $H_0^1(\gamma) : P_1(\gamma) = 0$, meaning that the proportion of non true nulls with p-values less than γ is 0.
- $H_1^1(\gamma) : P_1(\gamma) > 0$, meaning that the proportion of non true nulls with p-values less than γ is greater than 0.

Under the mixture model, $H_0^1(\gamma)$ holds true if $\pi_0 = 1$, in other words, there is a perfect match between the complete null H_0 and $H_0^1(\gamma)$, for every value of γ.

The quantity $P_1(\gamma)$ can be estimated by the empirical value $P_{1,n}(\gamma) = F_n(\gamma) - \pi_{0,n}\gamma$, where, $\pi_{0,n}$ is some consistent estimator of π_0 [25]. Assuming a normal distribution of $P_{1,n}(\gamma)$ for large n (central limit theorem) , $H_0^1(\gamma)$ is to be rejected if the following condition is met for:

$$\frac{P_{1,n}(\gamma)}{\sqrt{Var\left(P_{1,n}(\gamma)\right)}} > z_a$$

Where $\sqrt{Var(P_{1,n}(\gamma))}$ represents the standard deviation of $P_{1,n}(\gamma)$ under $H_0^1(\gamma)$. This modified SGoF method declares as effects the $N_{1,a}(\gamma) = nP_1(\gamma) - n\sqrt{Var(P_{1,n}(\gamma))}z_a + 1$ smallest p-values. This modified SGoF version also controls the FWER in the weak sense at level α [25].

Using a concrete estimator of π_0 [35], namely:

$$\pi_{0,n} = -n^{-1}\sum_{i=1}^{n} log(1-p_i)$$

results in an overestimation of the proportion of true nulls. Under the complete null $H_0^1(\gamma)$ it can be seen that $Var\left(P_{1,n}(\gamma)\right) = \frac{\gamma[1-2(1-\gamma)log(1-\gamma)]}{n}$ [25].

The threshold p-value, $p_{1,n,a}^*(\gamma)$, of this modified SGoF is given by $F_n^{-1}(n^{-1}N_{1,a}(\gamma))$ which is equivalent to the following expression

$$F_n^{-1}(P_{1,n}(\gamma) - \sqrt{Var(P_{1,n}(\gamma))}z_a + n^{-1}) .$$

As the threshold p-value $p_{1,n,a}^*(\gamma)$ is never greater than γ, we arrive at the conclusion that the power of the method increases, because, as both the significance level **α** and the number of tests n increases, $p_{1,n,a}^*(\gamma)$ tends to increase. We note that, as n tends to infinity, $p_{1,n,a}^*(\gamma)$

converges to $F^{-1}(F(\gamma) - \pi_0\gamma)$, which is greater than the p-value threshold convergence of the other original SGoF, which is $F_n^{-1}(F_n(\gamma) - \gamma)$. This results in a more powerful version of SGoF for any given γ and consequently a larger FDR [25].

The threshold p-value of this modified SGoF, when n tends to infinity, is just the point for which the cumulative proportion of p-values equals the proportion of true significant cases that fall below the initial significance level γ [25]. In the finite sample case, the interpretation must take into account the lower limit of significance, level **α**, namely $P_{1,n}(\gamma) - \sqrt{Var(P_{1,n}(\gamma))}z_a + n^{-1}$.

Simulation study of Binomial SGoF using R programming language

1. Simulations and data gathering

We conducted a simulation study to understand how Binomial SGoF differs from other adjustment methods, namely the Benjamini-Hochberg and Bonferroni procedures. As such, several comparisons with different known probabilities of true discoveries were generated. The t-test function generated a set of p-values to be adjusted by the methods. Two different scenarios were assayed and, for each, three different numbers of experiments S (number of tests performed, i.e., number of p-values), were simulated, 100, 1000 and 10000. The scenarios were:

1. The null hypothesis was always true.
2. The alternative hypothesis was true for some of the S tests. Different prevalences of the alternative hypothesis were chosen (5%, 10%, 20% of the total number of S tests). The effect sizes chosen for the alternative hypotheses corresponded to a mean difference of 0.36, 0.7 and 0.97.

The generation of p-values used two set of parameters:

Parameters 1:

n = 5, μ = 0, μ_x = 0, S = 100, prevalence = 0

Parameters 2:

n = 5, μ = 0, μ_x = 0.36, S = 100, prevalence = 0.05

Where n represents the sample size; μ represents the mean value of the normal distribution, from which samples belonging to the null hypothesis are taken; μ_x represents the mean value of the normal distribution, from which samples belonging to the alternative hypothesis are taken; S represents the number of tests performed; prevalence represents the proportion of alternative hypothesis in S.

Algorithm Steps:

Step 1. Load Parameters 1;

Step 2. Draw n values from a N(μ,1);

Step 3. Perform a two-tailed t-test with n-1 degrees of freedom of values in Step 2 against the mean value of 0;

Step 4. Save the p-value of the test performed in step 3 in a vector;

Step 5. Repeat steps 2-4 S times:

Step 5.1. Change μ to μ_x if the iteration number in step 5 is equal to $S \times (1 - prevalence)$;

Remark: The $S \times (1 - prevalence)$ cues the algorithm to introduce the alternative hypotheses p-values into the vector of p-values. When there is a prevalence, the p-values of the effects always appear in the $S \times prevalence$ end of the p-value vector (so we know what are the true significants).

Step 6. Reset μ = 0 and replicate Step 5 1000 times;

Remark: The 1000 replicates of Step 6 purpose is to generate mean and standard deviation values for the metrics.

Step 7. Repeat step 6 for S equal 1000 and 10000;

Step 8. If Parameters 1 are being used:

Step 8.1. Repeat steps 1-7 for n equal to 10 and 20;

Step 8.2. Output results, load Parameters 2 in Step 1 and repeat Steps 2-7.

Remark: Step 8 is skipped when Parameter 2 is loaded. The first iteration of the algorithm until Step 8 simulates the scenario where the null hypothesis is always true.

Step 9. Reset Parameters 2 and repeat Steps 2-7 for prevalences equal to 0.1 and 0.2.

Step 10. Reset Parameters 2 and repeat Steps 2-9 for n equal to 10 and 20;

Step 11. Output results;

Remark: Results from Step 11 correspond to the second scenario, where there is a prevalence of alternative hypotheses. The size of the effect of the alternative hypothesis (0.36, 0.7 and 0.97) correspond to 3 different subsets of the second scenario.

Step 12. Reset to Parameters 2 and repeat Steps 2-11 for μ_x equal to 0.7 and 0.97.

After obtaining the p-values for the different scenarios, and in case of scenario 2, for the different effects of the normal distribution (0.36, 0.7 and 0.97). The p-values were adjusted according to the Binomial SGoF, Benjamini-Hochberg and Bonferroni procedures, using a significance level of 0.05 and a Gamma equal to the significance level in the case of Binomial SGoF.

For the first scenario, the number of rejections on the unadjusted p-values and the adjusted p-values by the 3 methods were taken at a 0.05 significance level, the rejections were averaged by the 1000 replicates and their standard deviations were calculated. The results can be found on the supplemental files in table 4.

For the second scenario, the number of rejections on the unadjusted p-values and the adjusted p-values by the 3 methods were taken at a 0.05 significance level, the rejections were averaged by the 1000 replicates and their standard deviations were calculated. As the alternative hypotheses in this scenario are known, FDR and Power metrics for the 3 adjustment procedures were also calculated and averaged by the 1000 replicates and standard deviations were also calculated for these metrics. The results can be found on the supplemental files in tables 5-7, with effect size of the alternative hypothesis equal to 0.36, 0.7 and 0.97, respectively.

2. Results and discussion

Regarding the first scenario (results in table 4 in the supplemental files), where the null hypothesis is always true:

- The mean percentage of false positives (Significant % column in table 4) obtained in the simulation was close to the theoretical expectation (significance level equal to 5%) in all cases.
- The mean percentages of detection (Detection columns for the 3 methods in table 4) represent false positives (null hypothesis always true). Bonferroni and Benjamini-Hochberg show very similar results regarding type I errors. Binomial SGoF exhibits some similarities with the other 2 methods in some instances (when the number of tests S is 100 and 1000), its detection rate tends to increase in 1 order of magnitude relative to the other 2 methods when S is equal to 10000. As the authors of Binomial SGoF have described, this procedure tends to develop a higher number of rejections when facing a higher number of tests to correct.

Regarding the second scenario (results in tables 5-7 in the supplemental files), with different effects (0.36, 0.7 and 0.97) from the alternative hypothesis:

- The mean percentage of significant results detected at a significance level of 5% shows values higher than 5%, which increase with the increase of the prevalence of alternative hypothesis. This is expected because the more values from an alternative hypothesis, S has, the more likely it is for the p-values to be low, thus passing the 5% threshold detection. The larger the number of samples (n) for each p-value and the greater the effect ($0.36 < 0.7 < 0.97$), the larger the mean percentage of significant results detected, this is because the bigger number of samples and the effect, the more evidence the hypothesis test has to favor the alternative hypothesis, at a fixed significance level (5%).
- The mean FDR percentage (ratio of number of false rejections by total number of rejections) for Bonferroni and Benjamini-Hochberg was similar and presented a down trend, when the number of samples (n) and/or the prevalence of the alternative hypothesis also increased. The Binomial SGoF mean FDR also followed this down trend, when n and/or the prevalence increased, although at a slower pace than the other 2 methods. The size of the effect (0.36, 0.7 and 0.97) also showed a positive impact on the FDR, i.e., as the effect increased (for example table 5 to table 6), the FDR decreased for all the methods when the other parameters were fixed (this is expected because the larger the size of the effect, the more evidence there is to reject the null hypothesis, at a fixed significance level 5%). The number of tests (S) showed mixed results on FDR for the 3 methods. The discrepancy in FDR between Bonferroni or Benjamini-Hochberg and Binomial SGoF can be explained by the fact that under the former there were almost no discoveries, so the margin for false positive ones were also reduced. On the contrary, Binomial SGoF consistently presented more discoveries, especially when facing a large number of tests (S), thus the inflation of its FDR.
- The mean power percentage (ratio of the number of true discoveries by the total number of true cases). Concerning Bonferroni and Benjamini-Hochberg, the procedures exhibited similar results when the effect was 0.36 (table 5) and the number of samples (n) was low. In the other cases, the Benjamini-Hochberg procedure showed a higher statistical power, when compared to Bonferroni's. The statistical power of Binomial SGoF was in the general case higher than the other 2 methods, especially when the number of tests (S) increased. There were some exceptions to this trend in table 7, where the effect was 0.97; Benjamini-Hochberg's statistical power rapidly increased and even surpassed Binomial SGoF's when the number of samples (n) were larger. The down trend in statistical power of Bonferroni and Benjamini-Hochberg when the number of tests (S) increased, was due to their conservative nature when adjusting a large number of tests (for example Bonferroni adjustments increase substantially

with the number of tests, $\frac{significance\ level}{S}$). On the contrary, Binomial SGoF thrived in true discoveries, when facing a large number of tests. As described in [24, 25].

Chapter III – Workflow in Proteomics

Introduction to Proteomics

The term proteomics originates from the word “proteome”, it being the entire collection of proteins encoded by the genome (complete set of genes) in an organism [9, 11]. The main aim is to comprehensively describe the structure, function and interaction of the full set of proteins in an organism, both in temporal (protein content varies across time) and spatial (protein interactions between different locations in the organism) terms [9, 11].

In the majority of organisms, proteins are encoded by DNA and RNA. The general pathway to protein formation is as follows, DNA is transcribed into RNA which is then translated into a protein. As we see, there is a relation between genes (DNA) and proteins, this relation equates to the flow of information inside the cell. It was once thought that this flow of information was interchangeable, in other words, one gene gives origin to one protein [11]. This thought was later disproved mainly through the study of eukaryotic genes.

The study of eukaryotes, which constitute all the complex organisms, brought up the notion of an unidirectional flow of information within the cell [11, 13, 14]. The coding sequences of DNA in eukaryotes are called exons and are interrupted by noncoding extends of nucleotides called introns. Before the transcription into messenger RNA (mRNA) the exons must be spliced, in other words, the introns must be removed from the transcribed sequence of the pre-mRNA and exons must be joined together, depicted in figure 2 is a simple schematic of the splicing process. The different ways the splicing of exons can occur before the protein translation, plus the posttranslational modifications (PTM_S) of proteins, which are chemical modifications (i.e., acetylation, phosphorylation, glycosylation or some association with other biomolecules) after translation is finished, are responsible for the great abundance of proteins. For example, from the more than 20,000 protein-coding genes in the human genome, there are over 100,000 differentially spliced mRNAs, resulting in a proteome consisting of a massive number of proteins, well over 500,000 [10, 11, 14].

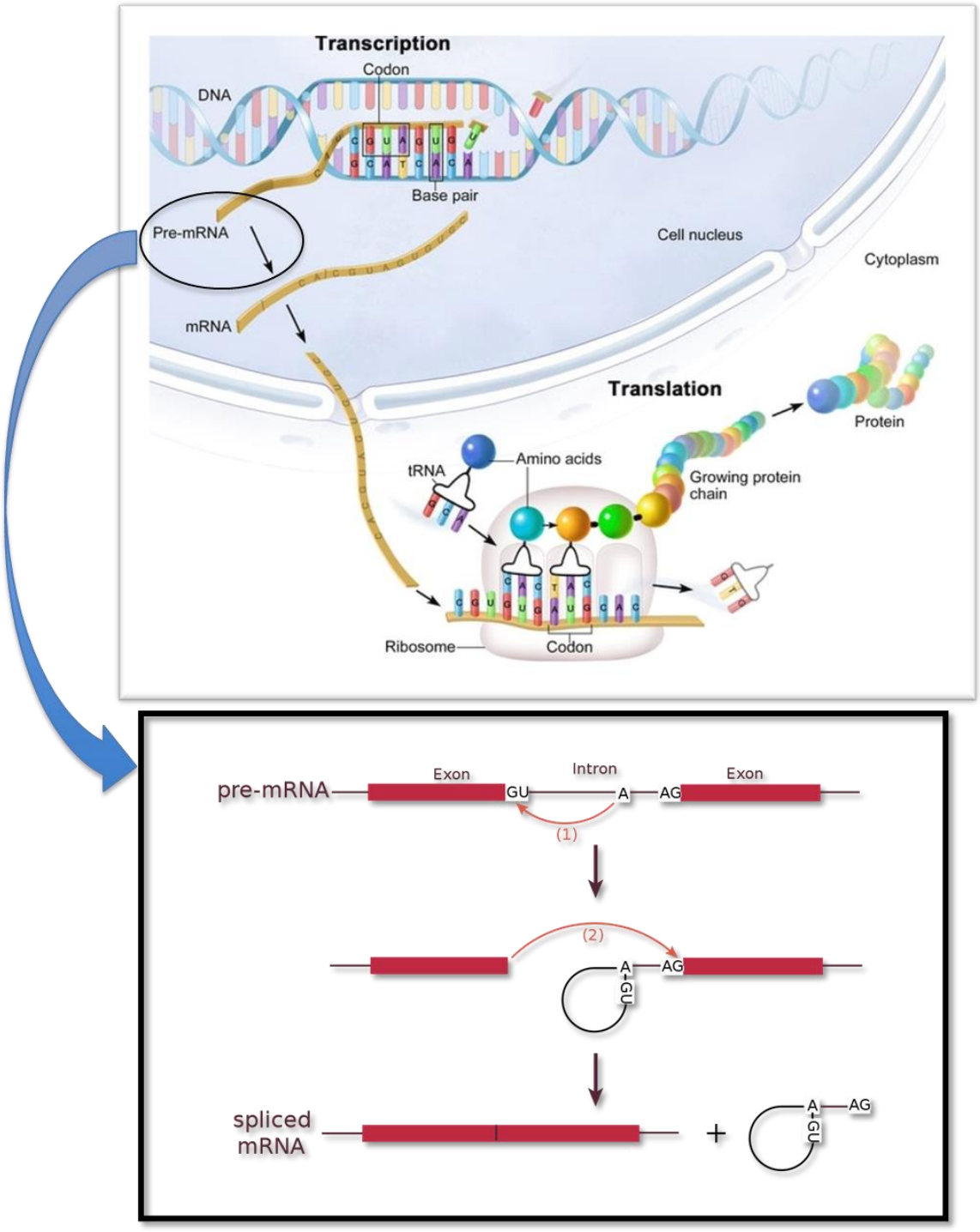


Fig 2: *Transcription and translation, in highlight is the process of RNA splicing. Adapted from [7, 8].*

It became evident that certain genes or DNA segments may code for different proteins or that certain coding sections of a protein in the DNA is spread across huge portions of DNA interrupted only by the noncoding sequences.

As it is stated by the central dogma of molecular biology, depicted in figure 3, the information is transferred from nucleic acid (DNA/RNA) to nucleic acid, or from nucleic acid to protein, making the transference between proteins or from protein to nucleic acid impossible, or at least unlikely [11, 14]. A nice explanation to this concept is to think of nucleic acids as the “brain” of the cell (carries all the information that makes life possible, even the smallest unit is evolved in the greater scheme) and regard proteins as little machines that perform “simple tasks” (it is the sum of their “simple tasks” that make a living being).

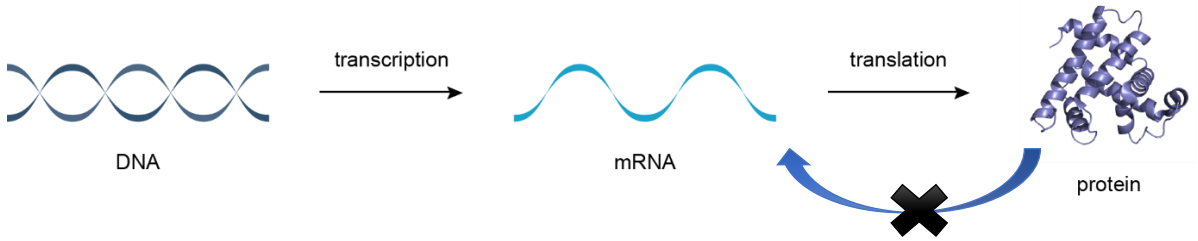


Fig 3: *Unidirectional flow of information, Central Dogma of Molecular Biology. Adapted from [3].*

As the ultimate effectors of variation in the genes, proteins represent the direct linkage to the organism phenotype, in other words, they perform the heavy lifting in order to maintain cellular homeostasis, the tasks they perform range from carrying information inside cells and across the body, to being the building blocks that make up important structures in the organism as it is depicted in table 1 [9, 11, 14]. This is especially evident in the context of a disease. Even though, it is the genetic alterations that predict the likelihood of developing certain disease, it is the phenotypic changes (i.e., protein changes, in this case becomes defective) that define the onset of the disease, becoming an important biomarker and a potential target for a pharmaceutical approach to a treatment [10, 13].

Given this, it becomes clear that proteomics should be the next logical step of study, after/besides genomics, and transcriptomics, it is more complex than the former “omics”. The genome of an organism is somewhat static, in other words, it remains constant in all cell types all the time (it is stable across all stages of the organism development). In contrast, the proteome of an organism is dynamic (it changes), because, as cells differ from tissue to tissue and even among cells from the same type, the characteristic that makes them distinguishable is their protein content, both in terms quantity as well as quality [11]. A change in the proteome must reflect differential activity in genes dependent on the cell type, to express the protein

needed for a particular function. For example, pancreatic cells largely express the insulin gene, which produces the insulin peptide required to regulate the glucose molecules, and with that, maintain an adequate level of sugar in the blood, whereas blood cells, before maturing into erythrocytes and losing the nuclei, predominantly express the hemoglobin gene, which translates into an higher production of the hemoglobin protein, that is responsible for an adequate transportation of oxygen and its delivery throughout the organism, notice table 1 for other instances of protein types as well as their function [11].

In summary, the full set of genes, which constitute the genome, with spatial and temporal stability, must have different gradients of expression, thus having different transcripts and consequently a less stable transcriptome, which results in the production of different proteins. And because each protein controls different and important functions they are the embodiment of cell differentiation and specialization, further allowing the possibility of larger and more complex organisms to exist [9, 11, 14]**.**

Table 2: *Example of proteins and their function in the organism. Adapted from [36].*

Protein Examples	Functions
Amylase, lipase, pepsin, trypsin	Help in digestion of food by catabolizing nutrients into monomeric units
Hemoglobin, albumin	Carry substances in the blood or lymph throughout the body
Actin, tubulin, keratin	Construct different structures, like the cytoskeleton
Insulin, thyroxine	Coordinate the activity of different body systems
Immunoglobulins	Protect the body from foreign pathogens
Actin, myosin	Effect muscle contraction
Legume storage proteins, egg white (albumin)	Provide nourishment in early development of the embryo and the seedling

This protein diversity and functionality is only possible due to their complex biochemical structure.

How do proteins work?

Proteins are made up of smaller units called amino acids, which are small organic molecules that consist of a central carbon atom (alpha) that is linked to an amino group, a carboxyl group, a hydrogen atom and a side chain that is made of variable components (the side chain is what makes amino acids distinguishable). Within a protein, multiple amino acids are linked together by peptide bonds, resulting in a long chain. These bonds are formed by a dehydration reaction, which results in the connection of the amino group of one amino acid to the carboxyl group of the adjacent amino acid and the release of a water molecule. Eventually, all amino acids that constitute a protein link up through this peptide bond and form a linear sequence that represents the primary structure of the protein [37].

Proteins are built from a set of twenty amino acids, each of which has a unique side chain. As it was mentioned, these side chains are what makes amino acids different, because they have different chemistries. Most amino acids have nonpolar or polar side chains, while others have negative or positive charged side chains. It is the chemistry of the side chains that determines the configuration and shape of the final protein. Because the primary structure can have interactions within it, formed between side chains [37, 38]. These bonds are from different types, depending on the amino acid side chain, ionic bonds form between charged amino acids, while hydrogen bonds form from polar amino acids. The rest of the interactions (especially nonpolar) can be described by the weak Van der Waals force, which is the sum of all non-binding forces, both attractive and repulsive [38, 39]. The only known exception to these noncovalent interactions are the disulfide bridges that can form between two cysteines (this type of covalent bond is critically important, when preparing a protein analysis). So, because of the side chain interactions, the sequence and location of amino acids in each protein guides the folds and bends that occur in that protein [38, 39].

The hydrogen bonding between amino groups and carboxyl groups in neighboring regions of the chain sometimes causes certain patterns of folding to happen, for example alpha helices and beta sheets, these patterns make up the secondary structure of a protein [40]. Most proteins contain many sheets and helices, as well as other less common structures. The set of structures and folds in a single linear chain of amino acids constitutes the tertiary structure of a protein [37]. Finally, the assemble of multiple protein molecules, which function as a single protein complex, constitutes the quaternary structure of a protein [41].

The final shape of a newly synthesized protein is typically the most energetically favorable. As they fold, many conformations are tested before reaching their final form. The folded proteins are stabilized by many noncovalent bonds between amino acids, in addition to

the chemical forces coming from the protein environment, which also contribute to their shape and stability [37-41].

In summary, the primary structure of a protein (its amino acid sequence) drives the intra/intermolecular bonding and the folding of itself. Resulting in the unique three-dimensional shape of the protein, as depicted in figure 4.

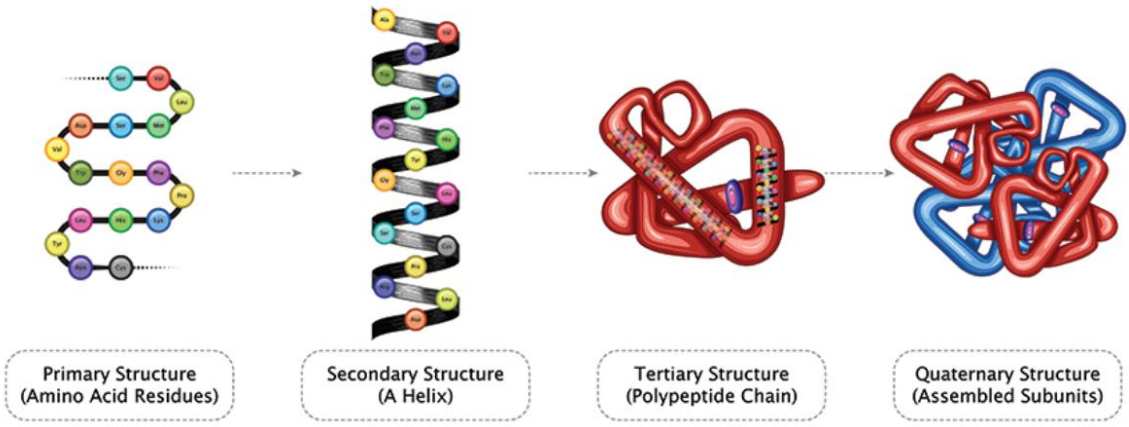


Fig 4: *Different structures of protein folding. Withdrawn from [2].*

It is this three-dimensional shape of the protein that is critical to its functionality, allowing it to interact with its environment, with other molecules and perform other functions [37-41].

All these events of protein assembly and folding vary from protein to protein (protein with different primary structure fold in a different way). As there are twenty unique amino acids that can be used to form proteins, the number of different proteins utilize these unique amino acids increases twenty times each time its length increases by one. So, in “theory” there could be a huge number of different proteins with different functions.

With advances in biochemistry and molecular techniques, protein identification and sequencing quickly disproved that all amino acids configurations were possible. The nature of proteins follow a set of rules in order to work and enable life, that is why they must be observed through experiment. Thus, it is important to discuss what kind of tools are available to conduct proteomic experiments, and how do they work.

Protein analysis techniques

When conducting a protein study, there are multiple methods one can use to analyze a protein sample. In general, proteins may be detected using either immunoassays or mass spectrometry. However, if a complex biological sample is to be analyzed, some sort of biochemical separation must occur before the detection step, as there are too many analytes present in the sample to perform an accurate detection and quantification.

Historically, antibody dependent methods (immunoassays) have formed the foundation of robust and highly sensitive target protein detection [42].

Methods like the enzyme-linked immunosorbent assay (ELISA), which, very briefly, works by immobilizing on a microplate the target protein and then complexed with a specific antibody that is linked to a reporter enzyme measure the reporter activity to determine if the protein was present and in what quantity [42, 43]. Other very widespread immunoassay is the western blot technique, which uses three elements to identify specific proteins from a sample: separation by protein size, solid support transfer and visualization of the target protein by marking it with a proper antibody [44, 45]. For example, immunoblotting (type of western blotting) works by performing SDS-PAGE (electrophoresis, separation by molecular weight) on the sample, the proteins are then transferred to a membrane where they are visualized by probing them with specific antibodies [46].

As for today, advances in mass spectrometry (MS) have propelled this technology to the forefront of both global and targeted analysis of proteomes. Progress in the sensitivity of detection and quantification of proteins have substantially enhanced the breath of applications for MS in proteomics [10, 47]. This range of applications in proteomics spans from: determining the abundance of proteins in different samples/conditions [48]; getting different protein isoforms [49]; determining the protein content and distribution in different body fluids [50]; finding the circadian rhythm of proteins [51]; performing single cell proteomics [52]; finding PTM_S [53]; studying protein interactions [54]; analyzing the structure of proteins [55]; etc. This means that with a single mass spectrometer, instrumentation is no longer the limiting factor when conducting a proteomic study.

MS is a sensitive technique used to detect, identify, and quantify molecules based on their mass to charge ratio (m/z). It was originally developed to measure elemental atomic weights and the abundance of naturally occurring isotopes, the first time it was used in biological sciences was to trace heavy isotopes through biological systems. Recently, MS has been used to sequence peptides, oligonucleotides and analyze nucleotide structure [47, 56].

The development of macromolecule ionization methods, including electrospray ionization (ESI), atmospheric pressure chemical ionization (APCI) and matrix-assisted laser

desorption/ionization (MALDI), propelled the study of protein structures by MS [57-59]. Ionization also allowed researchers to obtain peptide mass “fingerprints”, which could be used to match to proteins and peptides in databases and therefore help to identify unknown targets [59].

The culmination of these technological developments has resulted in methods that can successfully analyze many kinds of samples in different states (solid, liquid, gas) with incredible sensitivity, current mass spectrometers allow the detection of analytes at concentrations of 10^{-18} mol/L [60].

Mass spectrometry instrumentation

The transversal components for all mass spectrometers are:

1. ion source.
2. mass analyzer.
3. ion detector.

The nature of these components varies based on the purpose of the mass spectrometer, what type of data is to be analyzed and the physical properties of the sample. The samples are loaded into the mass spectrometer and then are vaporized and ionized by the ion source, like MALDI or ESI.

1. Ion source

In proteomics, ESI is commonly used because it is a type of soft ionization, it can be performed on solid or liquid samples that are nonvolatile or thermally unstable. This means that ionization of samples such as proteins, peptides and some inorganic molecules can be performed [5, 61]. In principle, ESI applies a high voltage at the tip of the capillary, the electric field generated breaks the liquid flowing out of the capillary into small charged droplets [57, 61]. As the surface to volume ratio of these droplets increases, the solvent evaporates, leading to an increase in charge density on the surface of the droplet. Finally, the droplet completely evaporates into many charged ions (the object of study), allowing the analyte to enter the gas phase in the form of a single charge or multiple charges ion [61]. The previous talking points are well represented in figure 5.

There are some clear advantages to using ESI-MS as an analytical method. One advantage is that this ionization technique is one of the softest ionization methods available, giving it the ability to analyze biological samples that are manly composed by non-covalent interactions in non-volatile solutions. Another advantage is the ability to handle samples with large masses, because the high molecular weight molecules typically carry multiple charges and the distribution of charge states accurately quantifies molecular weight, which in turn provides an accurate information about the molecular mass of the analytes. ESI can also provide multiple ionization modes, in other words, it can charge the compounds with positive charges or negative charges [5, 61].

On the other hand, ESI-MS can struggle to analyze a complex mixture in a sample, this can be mitigated by doing a sample separation (sorting) prior to ESI using for example liquid chromatography (to be discussed). Another disadvantage of ESI is the contamination accumulation inside the capillary needle, these residues tend to build up over usage and the cleaning of the device is no easy task [61].

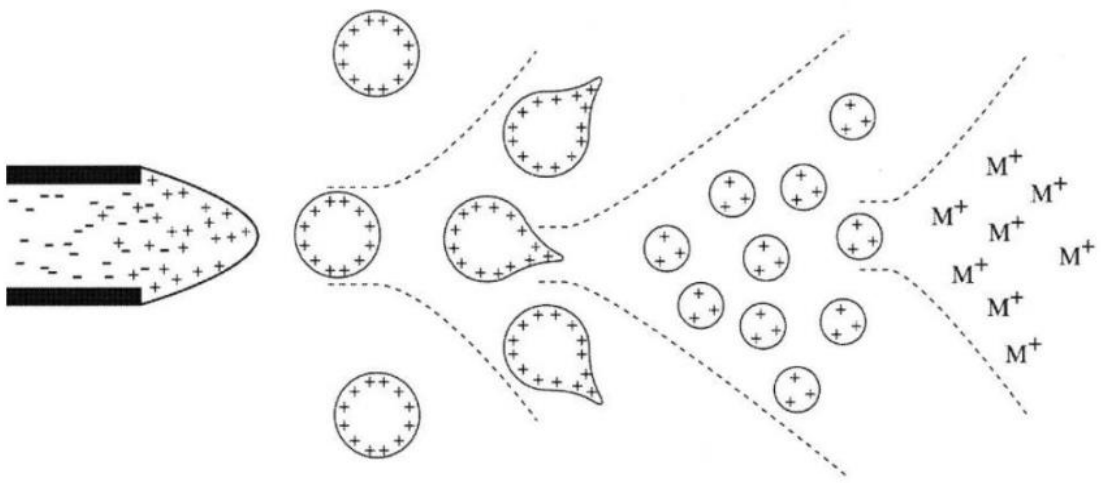


Fig 5: *Principle of electro spray ionization. Extracted from [5].*

After the gas-phase ions have been produced, they are transmitted into the mass analyzer, where they are analyzed according to their mass to charge ratios.

2. Mass analyzer

The core component of MS is the mass analyzer, it is used to determine the mass to charge ratio (m/z), this ratio is used to differentiate between molecular ions that were formed in the previous stage in the ion source. Commonly used mass analyzers include ion traps, time-of-flight (TOF), quadrupoles and orbitraps. The basic principle of the mass analyzer is to sort individual ions by their m/z with the help of electric/magnetic fields.

As the quadrupole and orbitrap are prime instruments and are commonly used in proteomics, it is worth to do an in-depth examination on how these two mass analyzers work.

2.1. Quadrupole

The quadrupole consists of 4 parallel metal rods equally spaced around the central axis (placed in a square configuration). Within the quadrupole assembly a fluctuating field is created by applying radiofrequency (RF) and direct current voltages (DC) to the quadrupole rods, with opposite rods having the same voltage. This field affects the trajectory of ions traveling down the flight path between the rods, based upon their m/z [62]. Each m/z have an optimum RF and DC setting, which creates a stable trajectory through the quadrupole, thus a quadrupole can be programed so that a specific m/z is stable down the length of the quadrupole, resulting in a resonating ion as shown in figure 6. Ions whose m/z is either to large or to small have an unstable fight path, which causes them to strike the rods and be loss, as depicted in figure 6 by the non resonant ion.

Typically, a quadrupole can be operated in one of two ways during a run, static or scanning [62].

In static mode, the RF and DC setting are set to only transmit the ion of interest, with a specific m/z, and all other ions are lost. Thus, operating in static mode provides the higher sensitivity to the ion of interest, because the quadrupole spends the entire time transmitting that specified m/z. This is the most common mode used in targeted quantitative analysis, since is the most sensitive, and additional information within the sample is not needed.

In scanning mode, the RF and DC settings are ramped up over time to create the typical mass spectrum. Ions are only transmitted to the detector when the RF and DC setting is correct to allow stable motion for that m/z. The scan time is the amount of time the quadrupole spends scanning the entire mass range, from the start mass to the end mass. For example, if the quadrupole scans at a 10^4 dalton per second, and the mass range is from 200 to 1200 dalton, the scan time is 0.1 seconds. This is called full scan mode. This mode provides

the most information about all the ions present in the sample. However, it is not as sensitive as doing a static monitoring, because the quadrupole transmits each m/z for only a small percentage of the total scan time [61].

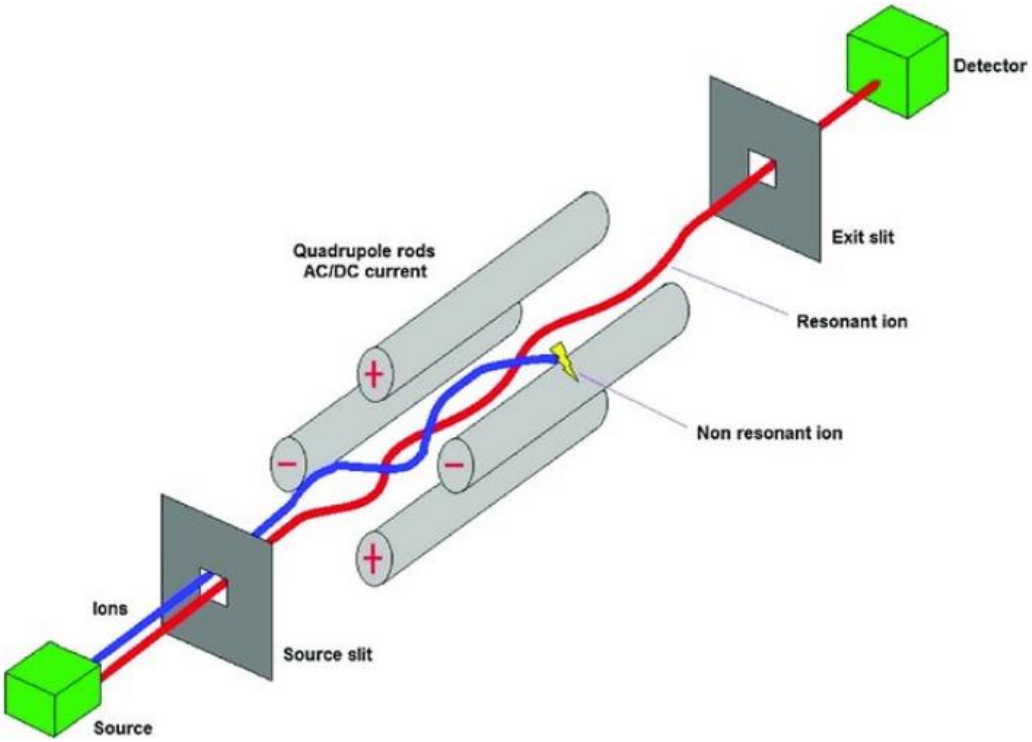


Fig 6: *Illustration of the quadrupole configuration between the source and exit slits. Extracted from [4].*

2.2. Orbitrap

The orbitrap mass analyzer, represented in figure 7, consists essentially of three electrodes. One spindle-shaped central electrode and a pair of bell-shaped outer electrodes, between the inner and outer electrodes there is an empty volume, which is used to accommodate and analyze the ion molecules [6, 63]. With voltage applied between the central and outer electrodes, a radial electric field bends the ion trajectory toward the central electrode while tangential velocity (this velocity is acquired right before entering the orbitrap) creates an opposing centrifugal force, keeping the ions on a rotational path around the spindle [6, 63]. At the same time, the axial electric field caused by the special conical shape of electrodes, pushes ions towards the widest part of the trap, initiating harmonic axial oscillations. The frequency of these harmonic oscillations along the central axis (oscillation along the z-axis in figure 7) depend only on the ion's m/z. The outer electrodes are then used as detector plates

(orbitrap acts as both analyzer and detector) for image current detection of these axial oscillations. The instrument obtains the frequencies of these axial oscillations and therefore the m/z of the ions through Fourier Transformation (FT) [6].

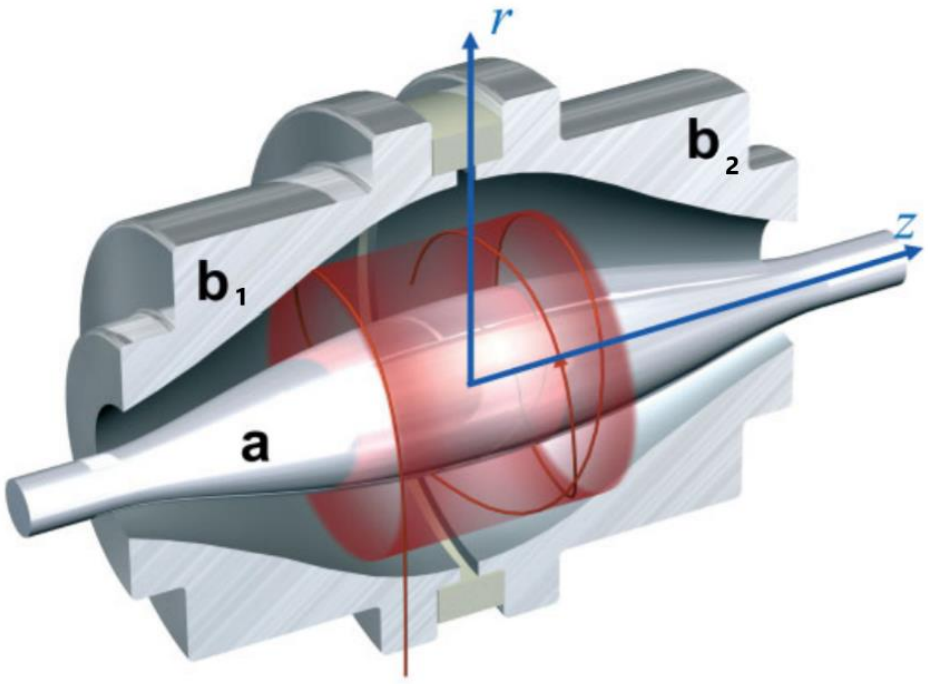


Fig 7: *Transversal cut of an orbitrap along the zr-plane. Ions are moving between the spindle electrode (a) and the outer electrodes (b_1, b_2). The frequency of oscillation of ions along the z-axis can be determined via FT and converted to a m/z. Adapted from [6].*

3. Ion Detector

After the molecular ions pass through the mass analyzer, they are detected and transformed into a usable signal by a detector. Most often, these detectors are electron multipliers (often used with a quadrupole analyzer) or microchannel plates that emit a cascade of electrons when each ion hits the detector plate, due to the small amount of ions leaving the mass analyzer at a given instant, an amplification step is commonly used to obtain a reliable signal. Once the analog signal of the mass-to-charge ratio is recorded, it is then converted to a digital signal and a spectrum representing the data run can be analyzed [64].

Some detectors are made to count ions of a single mass at a time and therefore they detect the arrival of all ions sequentially at one point, while other types of detectors, like photographic plates or image current detectors (the detector present in the orbitrap) can count multiple masses and detect the arrival of all ions simultaneously along a plane [6, 64].

Tandem mass spectrometry (MS/MS)

In order to achieve high performance in ion detection and quantification in proteomics, in other words, to get a high resolution (HR) and an accurate mass (AM) spectrum of the analytes over the MS run, many parallel analysis must be performed inside the mass spectrometer. This is possible when combining several mass analyzers in sequence. As already stated, in proteomics the usual combination is a quadrupole followed by an orbitrap (hybrid mass spectrometer) [65].

In this approach, distinct ions of interest are filtered through the quadrupole, based on their m/z and their relative intensities (precursor/parent ion selection), then, each specific precursor ion is transmitted to a special chamber in which it will suffer dissociation through high-energy collisions (HCD). The resulting fragments are injected into the orbitrap where they undergo HR/AM detection. The analysis of the precursor ion and its fragments is what it is called MS_2 and allows for the identification of the precursor ion (in this case the ionized peptide) when a database is used to match the experimental spectrum with the in-silico spectrum, in the computational analysis [65].

High Performance Liquid Chromatography

When performing targeted or global analysis with a MS instrument some things must be taken into consideration to get reliable results.

One important aspect is the relationship between the overall complexity of the sample and the amount of time the MS instrument can spend analyzing any specific ion. This of course can be managed with some tinkering on the machine parameters. For complex mixtures, one can set relatively low limits to the maximum time the MS instrument can spend on any target, in order to maximize the number of ions analyzed (losing some mass resolution and accuracy in the process). For less complex samples, the instrument can be set to spend more time on each ion, for example, by increasing the injection time of ions into the orbitrap. Of course, this may result in some ions to be missed in the analysis.

In order to mitigate the sample complexity, besides the adjustments that can be made to the MS instrument, a prior separation of the sample must be done. This is where high performance liquid chromatography (HPLC) shines.

In general, a HPLC system contains the following modules: a solvent reservoir, a pump, an injection valve, a column, a detector unit, and a data processing unit [66].

The solvent is delivered by the pump at high pressure and constant speed through the system. The sample is provided to the solvent by the injection valve. The separation principle of HPLC is based on the distribution of the compounds within the sample between a mobile phase (eluent) and a stationary phase (material of the column). Depending on the chemical structure of the sample, the molecules are retarded while passing the stationary phase. The specific intermolecular interactions between the molecules of the sample and the material of the column define their retention time, i.e., the time the molecules take between the start of the process and their moment of detection. Hence, different constituents of a sample are eluted at different times. Thereby, the separation of the sample compounds is achieved. A detection unit, for example an ultraviolet detector, recognizes the analytes after leaving the column. The detected signals are converted and recorded by a data processing system (computer software) and then shown in a chromatogram. After passing the detector unit, the mobile phase can be subjected to additional detector units, in the case of proteomic analysis this will be the mass spectrometer [66].

Protein sample preparation

Now that the important instrumentation required to perform a high resolution and accurate mass proteomic study has been discussed, the next logical step should be to consider the objectives of the study and the instrumentation available, and with that, develop/follow a protocol that enables an unbiased experimentation design. This will ensure a reliable data acquisition and consequently robust results.

This leads us to the next topic, the sample preparation. While the instrumentation used in proteomics is somewhat stable between different studies, the LC-MS instruments stay the same, maybe with some adjustments to their parameters, but they have a broad use in proteomics. What really changes between studies (if their objectives are different) is the way the protein samples are prepared.

The process starts with a very complex sample mixture, for example, a cell sample extracted from an organ, from a tissue, plasma, etc. The first step is to lyse the cells present in sample mixture in order to release their protein content [10]. This can be done in two ways:

- Using non-denaturing buffers, for example detergents like triton and NP-40, these will preserve the protein complexes. These buffers should be used if the objective is to do an affinity study, in this case the three-dimensional structure of the protein is important. So, normally, in this kind of studies, the sample is enriched with the bait compound, so

it binds to the protein of interest, and the sample is fractionated in-gel, so the protein of interest can be extracted and further digested.

- Using denaturing buffers, these will break the quaternary structures of the proteins and in turn generate single polypeptide chains that can be digested by proteases. These buffers are normally based on chaotropes (e.g. urea and guanidinium hydrochloride) or detergents like SDS. This denaturation step should be performed if the objective of the study is global proteome or PTM analysis.

Some physical disruption methods like grinding and sonication may be performed prior to the buffer application to allow a good actuation by the buffer [10].

After the solubilization of the proteins by the buffers, the next step is to reduce and alkylate the protein sample, the reduction (e.g., DTT) will break all sulfide bonds (form between cysteines) and the alkylation (iodoacetamide) will prevent them to rebuild and form secondary structures [10]. After this step, comes the digestion of the proteins by proteases. Typically, trypsin is the only one used. Tryptic peptides are easily analyzed by the mass spectrometer, because they are easily ionized in the ion source, they typically form peptide chains with suitable lengths to be detected by MS, and the C-terminal amino acid of the peptides is either arginine or lysine, which makes identification via software more reliable. One may consider using other proteases, but it really depends on the scope of the work.

One very important consideration to have when doing sample preparation is the process of desalting and cleaning. Many reagents and contaminants will interfere with the mass spectrometer analysis, for example detergents form polymers that can be ionized and will compete with the peptides in the mass spectrometer [10]. One solution is to use very low concentrations of detergents and if that is not possible, do an in-gel clean-up process of the sample [10]. For contaminants, it is nearly impossible to avoid them, as they are present everywhere, the best way to avoid them is to keep the material clean and keep the time that the sample is exposed to air to a minimal.

Finally, the digested sample, now a peptide mixture, can be loaded into the HPLC (or other peptide fractionation instrument) so it is further separated based on the properties of the column (as already discussed). The most common type of HPL used in proteomics is reversed phase HPLC, where peptides are sorted according to their polarity, polar peptides elute first followed by less polar ones. HPLC is directly connected to the mass spectrometer and the analysis begins.

What was discussed in these last topics constitutes an in depth look on the materials and methods of a general proteomic study. Depicted in figure 8 is the workflow of the previous discussed topics.

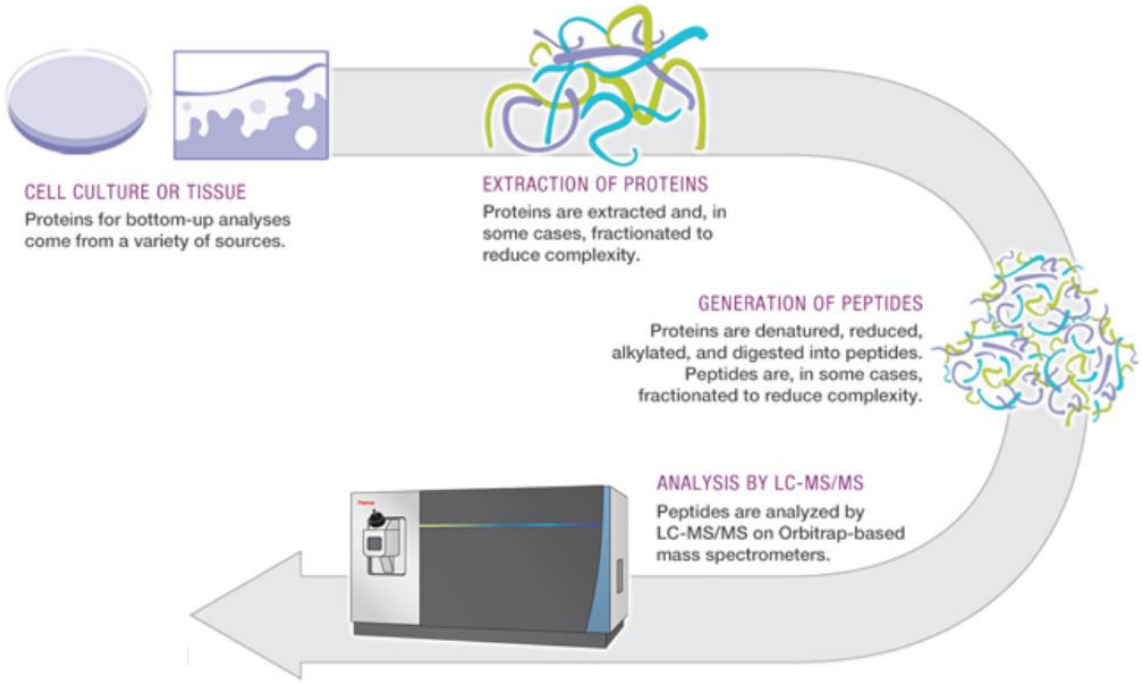


Fig 8: *Laboratorial workflow of a proteomic study. Adapted from [1].*

Computational methods for MS-based proteomic data

The computational methods that are going to be discussed in this topic will be based on bottom-up proteomics, as the previous discussed topics were also based on this assumption (the analysis of peptides and latter assembly into proteins).

The main computational procedures for MS-based proteomic data are:

1. The identification and quantification of peptides, PTM_s and proteins from the spectral data.
2. Analysis and biological interpretation of the data obtained in the previous step.

These points will be discussed in more depth in the following segments. The computational methods described in point 1 are present in the MaxQuant software [67], but can be generalized for other computational tools. Regarding point 2, the analysis tools can be found in the Perseus software (work frame for downstream proteomic analysis) [68], but there are other work frames for downstream analysis which have similar tools.

1.1 Peptide identification

To perform peptide identification, informatic tools utilize the fragmentation spectra (MS_2) obtained by the mass spectrometer. The most common approach is to utilize a database search engine (MaxQuant has an incorporated search engine, Andromeda), the search engine will go through a database of theoretical fragmentations and try to match the theoretical spectra with the empiric spectra [16].

The target database used by the search engine is made up of all known and translated (from DNA/RNA) proteins of an organism. The protein sequences are digested in silico (computationally) into peptides, according to the protease cleavage pattern utilized in the experiment, for example trypsin, which cleaves after a lysine or an arginine. For each in silico peptides, a list of expected fragment masses is calculated, based on the expected bond breakages for the experimental fragmentation technique. Then, for each experimental fragmentation spectrum, the search engine calculates a match score against all theoretical fragmentation spectra within a specified peptide mass tolerance. The best scoring peptide spectrum match (PSM) is taken as a candidate for the identity of the peptide. Since the matching process is made by statistical inference, the best scoring PSM might be a false positive, so it is important to impose a statistical control to the matching process. This is usually done by a double search, one against the target database and the other against a decoy database. The decoy database is generated from the target database, and the usual approach is to take the reverse sequence of peptides. The search against the decoy will provide false positive PSM_s. Comparing the score distributions of target and decoy PSM_s, posterior error probabilities can be calculated, and the FDR can be controlled. Other peptide features besides the matching score can be used to reduce the number of false positives, features such as the length of the peptide and the number of missed cleavages help to increase the confidence of identifications [12, 15, 16].

Other computational technique that can be used to identify peptides from the fragmentation spectra is the de novo peptide sequencing. With this method the peptides are identified using only the information from the experimental spectra and the patterns of fragmentation. Mass differences between certain peak pairs correspond to amino acid masses, as each type of amino acid has its own unique mass (their side chains are different) and if adjacent fragment peaks can be resolved, the whole sequence from N- to C-termini can be continued and the peptide can be sequenced without using a sequence database. Of course, using a database significantly increases detection sensitivity, so hybrid approaches (using a database and doing de novo sequencing) may give interesting results [12, 16].

Regarding PTM_s, once a peptide is identified as having a certain sequence and carrying at least one modification, their position in the sequence might not be localizable with

complete certainty. Given this, a score needs to be calculated that quantifies for each potentially modifiable amino acid in the peptide sequence the certainty of localization. For example, a peptide may contain several phosphorylated amino acids in its sequence, but from the peptide mass (in the experimental spectrum) it is known that it is phosphorylated only once. With this information and looking at the spectral evidence one can derive which amino acid has the highest probability of bearing the modification [16].

1.2 Protein Inference

The assembly of peptides into a list of proteins is a crucial step in the computational work, since the peptides are only a means to an end, as the usual objective is the study of proteins [12, 16].

The connection between proteins and peptides is many-to-many, since after digestion by a protease a protein generates many peptides and a single peptide can also be shared by many different proteins. Besides, using the identified peptides makes proteins that share common sequences indistinguishable from each other, resulting in the need to group those proteins into a redundancy group [12, 16].

The key to correctly infer the protein sequence is to use of peptides that are unique to a protein. One informative aspect to the uniqueness of a given peptide is its length, the longer the peptide is, the more likely it is to be unique. For example, using the *Mus musculus* (Mouse) reference proteome from UniProt (2020_05) [69], and doing an in-silico theoretical digestion of the proteome for all the peptides with lengths four through ten, one can see (figure 9) that the bigger the peptide the less likely it is for it to appear more than once in the proteome, this means that bigger peptides tend to have a 1:1 relationship with proteins.

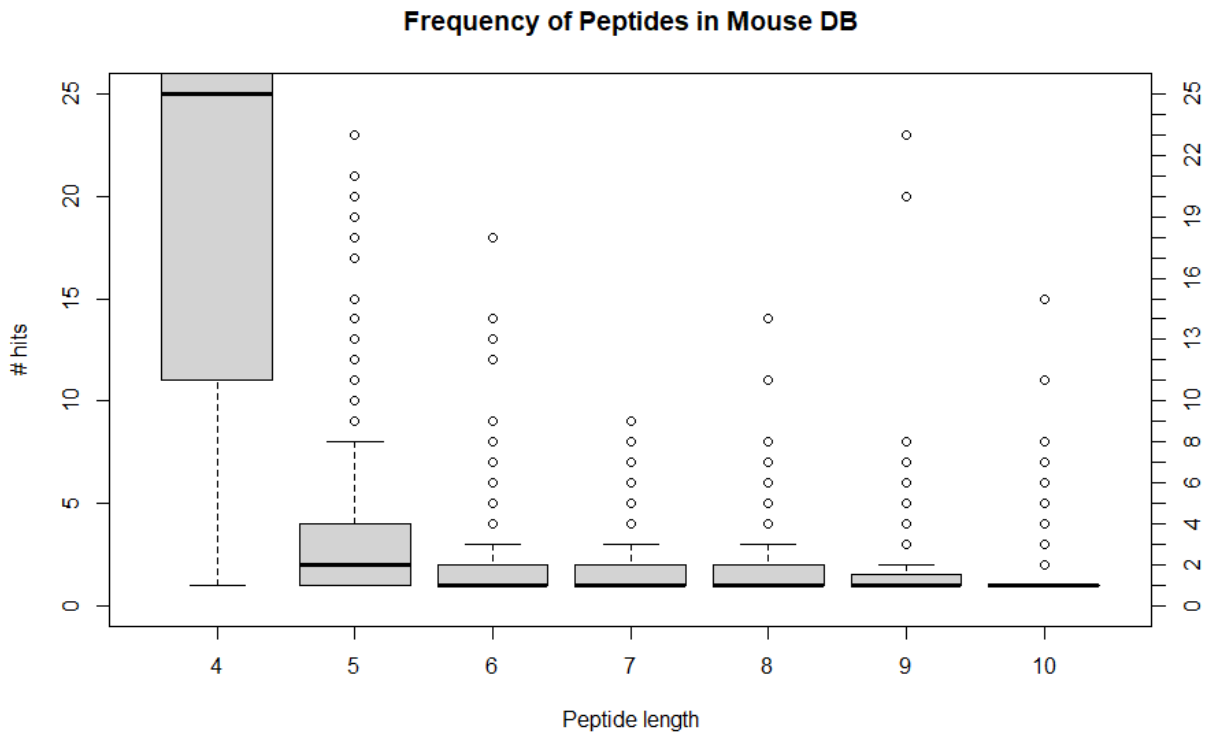


Fig 9: *Frequency of peptides with different lengths in mouse DB.*

The most used algorithms to predict a protein sequence are parsimonious and statistical models [16].

Parsimonious models apply the Occam's razor principle, which in this case states that one must find the smallest set of proteins that can explain the observed peptides in the spectra. This kind of models are usually applied by fast greedy heuristic algorithms to find the smallest set of proteins that explains the data. Statistical models can assemble large amounts of weak peptide identifications to infer the existence of a protein [16].

Using reliable peptide identifications leads to good conclusions about the characteristics of the identified proteins, so in the inferring step, for both models it is worth considering a threshold on peptide identification quality, for example, 1% FDR for PSM_s.

The output of this step should be a list containing groups of proteins. Each group has a set of proteins indistinguishable from each other based on the experimental peptide spectra. Proteins within the same group either have equal sets of identified peptides or the set of peptides for one protein is a proper subset of the set of another protein, leading to the conclusion that one protein does not exist in the sample if the other exists.

It is important to reinforce the concept of error propagation when inferring proteins from peptides. It is not enough to control the FDR at the PSM level, the protein inference methods must also control the FDR so that the number of falsely predicted proteins is contained, and the outcome of the study is relevant and reliable [16].

1.3 Quantification

Performing quantification on MS proteomic data is an indispensable step, as it improves the depth of information, leading to more enriched studies. Protein intensities change through time and in different conditions, it is a continuous process, so, by disregarding their intensities and only consider if a protein is present or absent, information that might be extremely relevant is lost (e.g. finding suitable biomarkers).

Quantitative methods for proteomics can be divided into absolute and relative quantification. In absolute quantification, the main goal is to determine the number of proteins or their concentrations within a sample. In relative quantification, the key is to find a ratio or a relative change of protein concentrations between samples. Since MS is not inherently quantitative, due to the lack of correlation between peptide intensity in a mass spectrum and their abundance in the sample, various label-free and label-based quantification methods have emerged to assist MS in doing both absolute and relative quantification [12, 13, 15, 16].

In label-free quantification (LFQ), the samples go through the standard preparation procedure (as discussed) but they are processed separately, in different MS runs, latter the raw MS data of each sample is quantified together in the computational step. LFQ uses the signal intensities of MS_1 peptide features as input, including the ones identified by matching between runs (it is important that a peptide is present in all samples, in order to be quantified), and outputs the relative protein abundance profiles over the number of samples. This protein intensity profiles are constructed to best fit the protein ratios in all pairwise comparisons between the samples. So, for each pairwise comparison, only peptides that appear in both runs are used, which makes the comparison very precise [12, 13, 15, 16].

Label-based quantification methods tend to be more precise than LFQ, because different samples/conditions are coeluted in the same MS run, reducing experimental errors and nullifying the match between runs procedure. This type of label quantification can be divided into metabolic and chemical labeling, if one decides to conduct a label-based study, one must prepare the sample in a different way from the label-free. The labeling reagents must be incorporated to each sample separately and then the samples can be mixed, the mix is treated like a normal sample that can then be digested and feed into the mass spectrometer, as the proteins were labeled, the mass spectrometer will recognize the different samples in the mixture. Computationally speaking, the distinction between metabolic and chemical labeling is not important. The important distinction is in what MS-level the quantification is done, i.e., if it is done on MS_1 or on MS_2 level [13].

In MS_1-level labeling, the peptide signals correspond to multiple samples, they can be compared, and form multiplexed (from different samples) isotopic patterns in the MS_1 spectra. The protein ratio calculation can be performed along the elution profile (chromatogram)

separately in each MS_1 scan and separately for each isotopic peak. This results in many estimates for the protein ratio, which can be well summarized by the median. When doing MS_1-level quantification, a requantification step is advice, because some isotope patterns might be missing in, especially for peptides that were formed from low-abundant proteins, requantification algorithms allow for a more detailed tracing of these peptides close to the noise level of the data [13, 16]. MS_1-labelling strategies are all based in different isotopic labels. Using heavy isotopic labels versus light isotopic labels allows for the separation of the same peptide in different conditions on the same MS run (they will present different m/z values in MS_1). An example of this type of labeling is the stable isotope labeling using amino acids in cell culture (SILAC). SILAC works by metabolically incorporating stable isotope labeled amino acids into the entire proteome of the study subject. Different cell groups, grown with different weighted labels can have their peptides differentiated on MS1 level of LC-MS/MS [13, 16].

In MS_2-level labeling, the multiplexed signals appear in the fragmentation spectra. The peptides in different samples are labeled with different molecules per sample that have the same mass but that eject different reporter ions when fragmented in MS_2. Comparing to MS_1-level labeling, MS_2-level labeling has the biggest multiplexing capacity, up to 11 samples in a single run can be measured, using the currently available tandem mass tag reagents [16]. On the other hand, the presence of coeluting peptides in the window for fragmentation leads to ratio compression problems, in other words, cofragmentation of different precursor peptides makes the ratio estimates to be off in random ways. There are some computational methods that can reduce ratio compression. For example, ignoring low intensity reporter ions, which are more prone to carry noise from cofragmentation, and giving reporter ions with higher intensity more weight when calculating protein intensities [16]. MS_2-labelling strategies are based on chemical labeling, which is performed after the digestion into peptides (in the sample preparation). In the mass spectrometer, peptides with different labels appear to be similar in MS_1-level, it is only when they enter MS_2 phase that the peptide fragments segregate, according to their labels. Examples of MS_2-level labeling techniques are the Tandem Mass Tags (TMT) and the Isobaric tags for relative and absolute quantitation (iTRAQ, which is similar to TMT) [13, 16].

When trying to decide which quantification method to use for an experiment, one must know the following. It is better to use label-based quantification if the sample size is low and if there are no monetary constraints in the reagents budget. LFQ has the advantage of not being bound to sample size (MS runs are performed separately, so it is limitless, as long as the computing power can scale to that number of samples) and it is way cheaper than its counterparts while also producing good results.

To compare the concentration of different proteins in the same sample, one must calculate the absolute abundance of each protein. For this purpose, there are two different

computational methods, intensity-based absolute quantification (iBAQ) or Top3. iBAQ is derived from the sum of all peptide peaks for a given protein divided by the number of theoretical observed peptides [13, 16]. For Top3, abundance is calculated by comparing the three most intense peptides of a protein to the three most intense peptides of that protein standard [13, 16].

2. Data analysis and interpretation

Once the identification and quantification of proteins is done, a data set is assembled, containing the proteins (or protein groups) as rows, samples/conditions as columns, and the protein intensities or intensity ratios in each cell.

Before jumping to the analysis, interpretation, and translation of the data into significant biological findings, some pre-processing must be performed on the data to make it more interpretable [17]. These data cleansing procedures include:

- Filtering out undesired proteins, such as, proteins that were identified in the contaminants or decoy data bases, proteins that were quantified by a low number of peptides, etc.
- Data normalization, to correct for systematic shifts and to make sure samples are comparable.
- Missing value imputation. The occurrence of missing values is quite common and it is due to: biochemical and analytical constraints (miscleavage, dynamic range, ionization competition, ion suppression, etc.) to bioinformatics mechanisms (peptide misidentification, ambiguous matching of the precursors in the quantitation step, etc.) [70]. Some examples of well performing imputation algorithms are: k Nearest Neighbors (input the average value of the k most similar proteins); Deterministic Minimum Imputation (replacing the missing value by the minimum value observed in the data or in that sample); Probabilistic Minimum Imputation (replacing the missing values with a random value from a Gaussian distribution centered on the minimum observed value and with a variance dependent on the variance of the data) [70].

After all data cleaning procedures have been applied, the protein data is ready to be analyzed and interpreted. So, depending on the objective of the study, many statistical models and learning algorithms can be used to extract knowledge from the data [16, 17]. If the interest is to check for differences between groups or conditions (e.g. differential expression between healthy and disease states), statistical models like the t-test and ANOVA (generalization of t-test for more than 2 conditions) are simple and optimal tools, combining the significantly changing proteins with an enrichment analysis (Fisher’s exact-test), one can obtain a

comprehensive view of the biological roles of the proteins [17]. Principle component analysis and clustering methods, like hierarchical clustering, are also a good choice to visualize expression patterns of groups of proteins along different conditions [16]. Machine learning systems can also be a good approach, e.g., if one wants to build a model to predict health/sickness, based on the protein expressions [16]. The way the data is analyzed and the way it is interpreted depends on the scope of the study.

Chapter IV – R routine for differential expression analysis and case study

Introductory consideration

After the identification and quantification steps done computationally, in this case, by MaxQuant. Proteomics datasets are ready to undergo several analysis procedures, results gathering and consequently knowledge building. For this purpose, the creators of MaxQuant also developed a downstream proteomics analysis tool, called Perseus. As the scope of this dissertation is the differential expression analysis, more specifically, the problem with multiple hypothesis testing, an R routine for differential expression analysis of label-free proteomic data was developed. This R routine executes the same analysis for differential expression as Perseus does, offering at the same time a larger repertoire of missing value imputation methods and p-value adjustment techniques, including Binomial SGoF and Conservative SGoF (methods of interest).

To investigate the ability of the implemented R procedures to detect differential abundant proteins, specifically, to compare the efficiency of the different adjustment methods for the p-values. The developed pipeline was put to the test, against an evaluation case study designed by the Proteome Informatics Research Group (iPRG), of the Association of Biomolecular Resource Facilities (ABRF) back in 2015 [71]. As the intention of the iPRG was to evaluate different computational and statistical tools in differential expression analysis (like the goal of this dissertation) they provided several starting points for the identification and quantification of proteins (raw files, LC-MS/MS spectra, integrated peak intensities and other intermediate results). This being said, in this dissertation work, the raw files (direct outputs of the LC-MS/MS runs) were used, as they are the files required by MaxQuant to perform identification and quantification of proteins.

Implementation of procedures in downstream analysis of LFQ proteomics

The following description is a generalized automated pipeline developed to perform statistical analysis of label-free quantitative proteomics data preprocessed by MaxQuant.

Step 1. To load the “proteinGroups.txt” file into the directory with the user interface script;

Remark: The “proteinGroups.txt” file is one of many files outputted by MaxQuant, for the purpose of statistical analysis it is the most important one, because it contains the quantification for each protein/protein group in each sample.

Step 2. To open the script containing the packages and the functions used by the program and load them into the workspace;

Remark: The libraries that are used by the program are: “SummarizedExpreiment”, “DEP”, ”arrangements”, “sgof”, “EnhancedVolcano”. There are also 5 other scripts with self-made functions that run in the back-end of the user interface script.

Step 3. The console asks for the experimental design in the following order:

a) User input the unique labels corresponding to each condition/group present in the experiment;
b) User input the number of replicates per condition/group;
c) User input the labels as they were set in the experimental design in MaxQuant;

   Remark: This step must have the number of labels equal to the sum of the number of replicates per condition/group.

Remark: Steps 4 through 6 are of the pre-processing nature, which include data filtering, transformation, and missing value handling.

Step 4. The console presents 3 separated tables, which contain proteins/protein groups filtered by 3 different features:

a) Shows table containing information about proteins only identified by site;
b) Shows table containing information about proteins that matched the reverse (decoy) database in the identification step in MaxQuant;
c) Shows table containing information about proteins that matched the contaminant database in the identification set in MaxQuant;
d) User input the filtering combination to discard such proteins based on the previous features;

Step 5. The program performs $\log_2$ transformation of each intensity and presents to the user the distributions of the transformed intensities for each sample for a qualitative assessment;

Step 6. The program guides the user through different procedures to deal with the non-random missing values problem:

a) It eliminates each protein/protein group that does not have at least one condition with a number of missing values lesser than or equal to a threshold given by the user;

Remark: This does not eliminate all missing values but it does significantly trim down random missing values and reduce artificiality presented by imputation methods.

b) It performs missing value imputation based on the method given by the user;

Remark: The user is given a repertoire of imputation methods commonly used in proteomic analysis, such as: Nearest neighbor averaging technique, deterministic minimal value imputation, random draws from a Gaussian distribution centered at a minimum value or at a user specified value.

Remark: Steps 7 and 8 represent the program block responsible for the differential expression analysis. Regarding step 8, the program performs hypothesis tests under the of equality of means assumption (H_0).

Step 7. The program helps the user to do a quality assessment of homoscedasticity between conditions/groups by plotting the pairwise protein variance ratios;

Remark: If the pairwise boxplot of variance ratios significantly differs from 1 the variances should be considered unequal. This will relax the equal variance assumption of the following comparison tests.

Step 8. The program performs differential expression analysis on the data:

a) It asks the user for the significance level;

b) It detects the experiment set up and assess the best comparison configuration:

b.1) If the number of conditions/groups is equal to 2:

b.1.1) The program performs a two-sample t-test for each protein/protein group and it corrects the p-values using 6 (*) different methods;

b.2) If the number of conditions/groups is greater than 2:

b.2.1) The program performs one-way ANOVA for each protein/protein group and adjusts the p-values using 6 (*) different methods;

b.2.2) For each adjustment method in b.2.1) the program rejects the proteins/protein groups which p-values fall below the one set by the user in a).

b.2.3) The contrasts of each rejection are calculated based on two-sample t-test and the p-values are adjusted using the same method that adjusted the p-values of ANOVA in b.2.1).

(*) ***Remark:*** The adjustment methods implemented in the program are: Binomial SGoF, Conservative SGoF, Benjamini-Hochberg procedure, Benjamini–Yekutieli procedure, Bonferroni correction and Šidák correction

Remark: Steps 9 and 10 are the visualization of results from the differential expression analysis.

Step 9. For each adjustment method (combination of methods in case of ANOVA) used and for each contrast the program presents a plot of the $-\log_{10}$(p-value) in function of the fold change;

Step 10. In case of more than 2 conditions/groups, the program will output 6 (one for each combination of adjustments) tables. Each table contains the identification of the contrast in each row, and the respective proteins/protein groups deemed significative in the ANOVA step with their fold change and p-value.

Case Study and Data Acquisition

1. Experimental design

The study conducted by iPRG was based on four artificially made samples of known composition. Each sample containing a constant 200 ng background of tryptic digests of *Saccharomyces cerevisiae* (yeast). Each one of the four samples was separately spiked with different quantities of six individual protein digests. All proteins were reduced and alkylated with iodoacetamide before being digested by trypsin. Table 3 shows the concentrations of the spiked-in proteins in each sample.

Erro! Ligação inválida.2. LC-MS/MS set up

Each one of the four samples were analyzed in triplicated with LC-MS/MS, for a total of twelve runs (performed in random order).

The samples were separated using a Thermo Scientific Easy-nLC 1000 system (LC) with a 110-minute linear gradient of 0-40% acetonitrile in 0.1% formic acid at 250 nL/min, directly connected to a Thermo Scientific Q-Exactive mass spectrometer. The data was acquired in data-dependent mode, with each MS survey scan followed by 10 MS/MS HCD scans. Both MS_1 and MS_2 data were acquired in profile mode in Orbitrap, with a resolution of 70000 and 17500 for MS_2. The MS_1 scan range was 300-1650 m/z.

3. Protein Identification and Quantification

The raw data gathered from the LC-MS/MS runs was analyzed using MaxQuant v1.6.17.0.

The FASTA file containing the search database for the background yeast proteins and the six artificial spiked-in proteins, was loaded into the global parameters of MaxQuant, the decoy database was set to reverse, with the purpose to find false identified proteins.

The match between runs option was enabled, as it is required to perform label-free quantification using the LFQ algorithm of MaxQuant.

The label-free quantification tab in the group-specific parameters of MaxQuant was enabled, as well as the normalization of the sampled data (normalization of the data is required to perform unbiased comparisons between samples and between conditions).

Finally, the raw data was loaded into MaxQuant and the experimental set up was set to 12 independent LC-MS/MS runs with no fractionation. Every other parameter of that version of MaxQuant was left in the default setting.

After finishing the data processing, the software outputted the desired file, namely, “proteinGroups.txt”, this file contains the necessary information to perform downstream differential expression analysis.

4. Differential expression analysis using the R routine

Using the developed R routine, described in this chapter, differential expression analysis was performed on the outputted “proteinGroups.txt” file.

Going through the R pipeline, the experimental design was set, as 4 different conditions with 3 replicates each.

Following this, a qualitative assessment of the undesirable proteins was made, through the usage of 3 categorical columns:

1. Only identified by site, which contains information of proteins matched due to single amino acid modification;
2. Reverse, which contains false protein identifications;
3. Potential contaminants, which contains information about proteins that might not be native of the sample;

The information in point 1 and 2 was not relevant for the differential expression, so these undesirable proteins were removed. However, a careful oversight of the potential

contaminants showed that one of the artificial spiked-in proteins (Bovine serum albumin) was considered a contaminant by the MaxQuant search engine. The action here was to not discard the potential contaminants proteins.

The following step was to deal with the missing values, as already discussed, missing values in LC-MS/MS experiments are non-random, they occur because of the detection threshold of the mass spectrometer. This being said, first a maximum number threshold of 0 was set, to discard proteins that were not identified in all replicates of at least one condition, this will mitigate artificiality when performing imputation of missing values. After applying this threshold, from the initial 3269 proteins only 2620 were kept, as shown in figure 10, after the first missing values removal, samples were left with few missing values.

The remaining missing values were replaced using the MinDet imputation method, which replaces missing values by the minimum value observed, this being the q-th quantile (q = 0.01) of the observed values in that sample.

The qualitative assessment of the variances homoscedasticity was skipped, as the number of replicates per condition is low (3 replicates per condition), the equal variance assumption was relaxed and subsequent hypothesis tests were performed in accordance.

Steps followed in the differential expression analysis:

1. The significance level was set to 0.05;
2. The Gamma parameter was set to be equal to the significance level (Gamma is used by Binomial SGoF and Conservative SGoF, as the expected number of rejections, the authors recommend to use the same value as the significance level);
3. Differences between the conditions means were investigated by one-way ANOVA for each protein. The adjustment on the p-values adjustment was applied as described in the R routine;
4. All pair-wise comparisons were considered for the significant results identified in step 3, using two-sample t-test and the same p-value adjustment;
5. Steps 3 and 4 were repeated for each adjustment method considered;

Once in this situation the proteins that are differentially expressed are known, confusion matrices and evaluation matrices were then considered.

For each adjustment method, 2 confusion matrices were constructed, one for the adjustment of ANOVA p-values, and the other for the adjustment of the contrasts. Each one of these matrices has its own evaluation table, with performance metrics. These tables can be found in the supplemental files, tables 8-13.

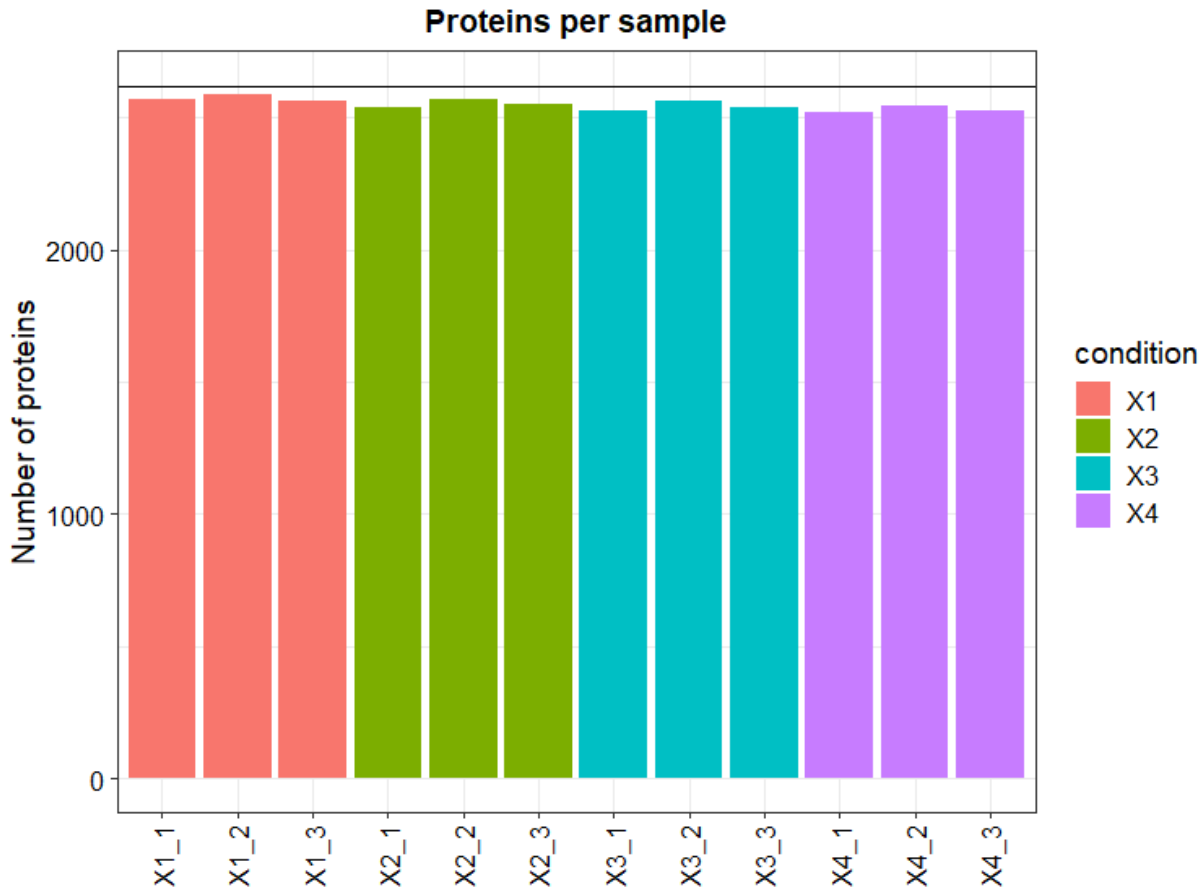


Fig 10: *Total number of proteins per sample, after applying the initial missing value filter. The horizontal line represents the total number of proteins in the dataset (2620). As this is before missing value imputation, every sample has some of its protein's values missing.*

5. Results and Discussion

The evaluation metrics chosen to compare the adjustment methods were the following:

1. $Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$
2. $Precision = \frac{TP}{TP + FP}$
3. $Sensitivity/Power = \frac{TP}{TP + FN}$
4. $Specificity = \frac{TN}{TN + FP}$
5. $FDR = \frac{FP}{TP + FP}$

In these 5 metrics, TP represents the true positive findings, i.e., the intersection of the real positives and the study positives, TN represents the true negative, i.e., the intersection of the real negatives and the study negatives, FP represents the false positives, i.e., the intersection of the real negatives and the study positives and FN represents the false negatives, i.e., the intersection of the real positives and the study negatives.

From the one-way ANOVA results (left side of tables 8-13 in the supplemental files and figure 11 below):

- Accuracy and specificity were similar in all methods, above 99.4%.
- Regarding precision (1 – FDR), Bonferroni and Šidák correction methods showed the highest values, 44.4% for both, and Binomial SGoF presented the worse precision score at 30.0%, Conservative SGoF showed the second worst precision score, at 35.3%.
- In terms of sensitivity/power, the Benjamini-Hochberg procedure, Binomial SGoF and Conservative SGoF found all differentially expressed proteins, having a sensitivity/power score of 100.0%. The other methods identified 4 out of the 6 right proteins, obtaining a sensitivity/power score of 66.7%.

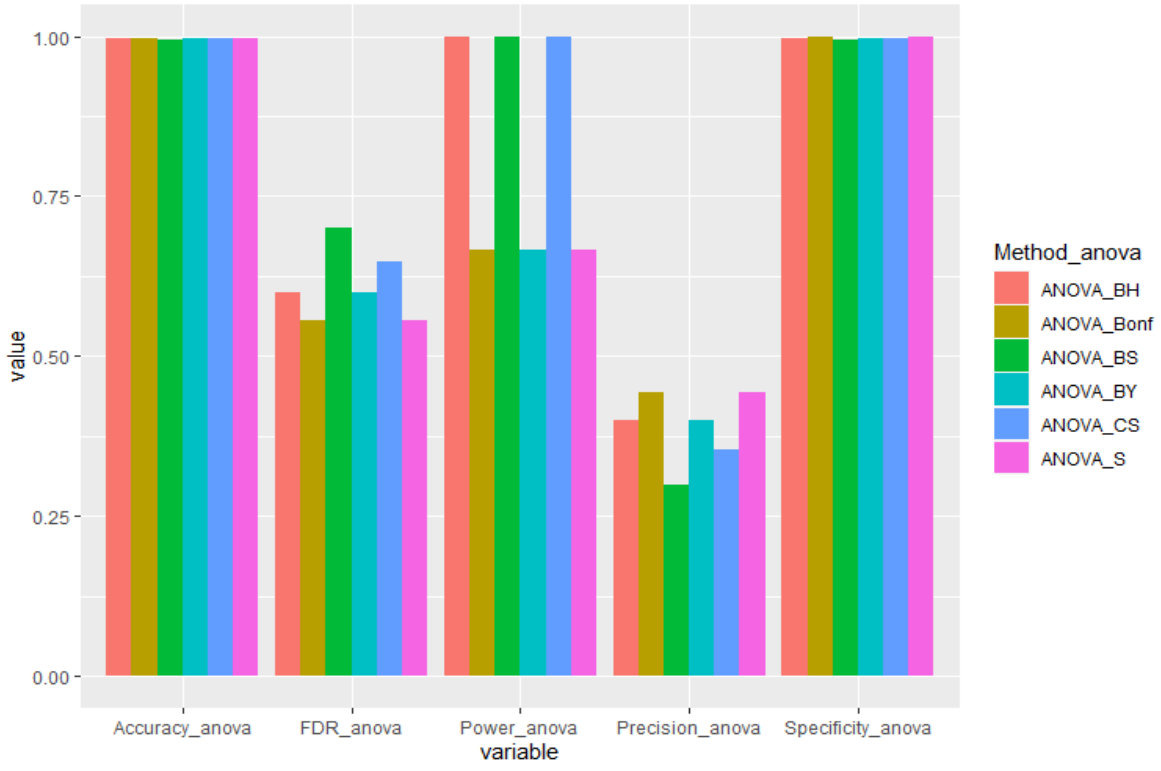


Fig 11: *Bar plot of the evaluation metrics grouped by the 6 adjustment methods applied to the ANOVA p-values. BH, Bonf, BS, BY, CS and S being respectively Benjamini-Hochberg procedure, Bonferroni correction, Binomial SGoF, Conservative SGoF, ,Benjamini–Yekutieli procedure and Šidák correction.*

The previous results show that there are many more real negative cases in the data, i.e., proteins that are not differentially expressed, than true positive cases, i.e., proteins that are differentially expressed. This can be verified through the specificity, precision, and accuracy metrics. All adjustment methods show almost perfect specificity scores, with some decimal points of percentage of difference between them. Which points to a large difference between the number of true negatives and false positives. On the contrary, precision was low across all methods, toping at 14% difference between Bonferroni/Šidák and Binomial SGoF. This points to a number of false positives greater than the number of true positives, possibly in the same order of magnitude. This information, plus the fact that accuracy shows the same results as specificity across all methods, is an indicator that the bulk of the data is composed by real negative cases. In fact, the total number of real negative cases is 2614, while there are only 6 real positive cases.

Another observation to be made is the strength these methods have in controlling the type I errors (1 – specificity), which is in compliance with the fact that the methods were developed to correct type I errors.

The inflated values of FDR are due to the low number of true positives; in fact false positives (which happen due to the large number of negative cases) greatly increase this ratio.

Regarding sensitivity/power, the results show perfect scores for Benjamini-Hochberg, Binomial SGoF and Conservative SGoF procedures. With the exception of the power result for Benjamini–Yekutieli's, the other results are in accordance with the way these methods work (Benjamini-Hochberg, Binomial SGoF and Conservative SGoF are more able to find true positives in their corrections than Bonferroni and Šidák). Regarding the differences in power between these methods, they are due to the low number of real positive cases, only 6, so missing just 1 out of 6 translates to a drop of 16.7% in power.

For these results concerning one-way ANOVA, all methods show good overall performance, with special considerations for Benjamini-Hochberg, Binomial SGoF and Conservative SGoF methods, which stand out in the sensitivity/power category with perfect scores.

From the pair-wise comparisons (right side of tables 8-13 in the supplemental files and figure 12 below):

- Regarding the accuracy metric, all methods showed similar results. Benjamini-Hochberg procedure was the best, with a score of 73.3%, and Bonferroni method was the worst, with 68.5%.

- Inspecting precision (1 – FDR), the Šidák procedure exhibited the best score, 66.7%, while Binomial SGoF and Conservative SGoF showed the worse performances, 49.1% and 55.0% respectively.
- In terms of specificity, the Bonferroni, Šidák and Conservative SGoF showed the best scores, 73.3% for both Bonferroni and Šidák and 72.7% for Conservative SGoF. The Benjamini-Hochberg procedure had the worst specificity, 64.8%.
- Regarding sensitivity/power, Benjamini-Hochberg exhibited the highest score, 86.1%; among the 36 differentially expressed contrasts (the 6 spiked-in proteins in each one of the 6 pair-wise comparisons), it deemed 31 as being differentially expressed. Benjamini–Yekutieli and Binomial SGoF methods showed similar sensitivity/power results, at 75.0% and 72.2% respectively out of 24 and 36 differentially expressed contrasts respectively. Bonferroni and Šidák procedures exhibited similar power results, at 62.5% and 66.7% respectively out of 24 differential expressed contrasts. Conservative SGoF method presented the worst sensitivity/power score, at 61.1% out of 36 differential expressed contrasts.

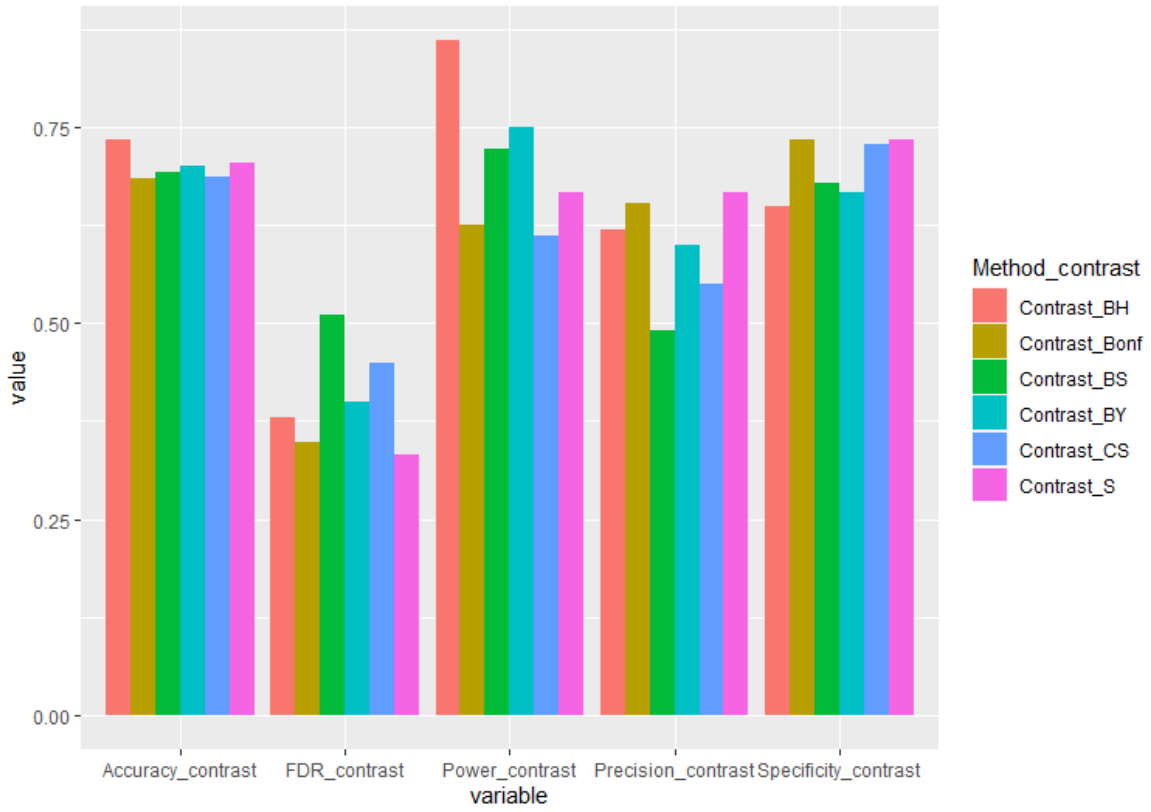


Fig 12: *Bar plot of the evaluation metrics grouped by the 6 adjustment methods applied to the contrasts p-values. BH, Bonf, BS, BY, CS and S being respectively Benjamini-Hochberg procedure, Bonferroni correction, Binomial SGoF, Conservative SGoF, ,Benjamini–Yekutieli procedure and Šidák correction.*

The previous results exhibit a more balancing distribution of true positives and true negatives. This is evident from the similar results obtained for specificity, precision, and accuracy. For example, Benjamini-Hochberg method got very similar results for these metrics, indicating similar true positive results and true negative results; in fact, this method discovered 31 pair-wise comparisons with differentially expressed proteins and 35 without differential expression. On the contrary, Binomial SGoF and Conservative SGoF procedures got lower precision scores, compared to other methods, and compared with the other metrics, because of the false positives carried over in the one-way ANOVA analysis, thus increasing the number of real negative cases in the pair-wise comparisons; in fact, the SGoF methods were the only methods that had to correct more than 100 pair-wise comparisons of proteins.

Regarding just specificity and precision, Bonferroni and Šidák procedures got the best scores, as expected, because these metrics benefit methods leading to few false positives, which Bonferroni and Šidák methods were tailored to do (to correct type I errors). Conservative SGoF showed a large specificity and a much lower precision, because it has more real negative cases than Bonferroni and Šidák; while keeping specificity at the same value as these two methods, the higher absolute value of false positives drastically lowers its precision. The same happens for Binomial SGoF.

As for sensitivity/power, the best performing method was Benjamini-Hochberg with a 31 out of 36 true discovery rate. Benjamini–Yekutieli and Binomial SGoF appeared to be the second and third best methods for this metric with very similar scores, with Benjamini–Yekutieli discovering 18 out of 24 diferentially expressed proteins and Binomial SGoF 26 out of 36 (similarly to Bonferroni and Šidák, Benjamini–Yekutieli only had 66.7% power at one-way ANOVA, so in the pair-wise comparison the method was working with less real positive cases). Bonferroni and Šidák, as expected, were the least effective at discovering the true positives out of the real positive cases. The surprise in this metric was the Conservative SGoF, that got the lowest power score despite having perfect power score in one-way ANOVA.

For these results, concerning the pair-wise comparisons, the method that stud out in a positive note was Benjamini-Hochberg's. It detained a very high power score, while keeping the FDR low, which is something to look for in an adjustment method. Whereas in the results of one-way ANOVA, Benjamini-Hochberg, Binomial SGoF and Conservative SGoF exhibited very similar values across all metrics and none stud out, in these results of the pair-wise comparisons Benjamini-Hochberg kept his performance very high. On the contrary, both SGoF methods took a hard hit performance wise. This detrimental in performance of both SGoF methods was probably due to the low number of tests, as described in [24, 25]. Both SGoF techniques excel when correcting a high number of tests (in one-way ANOVA they corrected a total of 2620 tests, while in the pair-wise the number dropped to the hundreds).

Chapter V – Final Remarks

Conclusion

This dissertation focused on the multiple hypothesis testing problem and on how it affects the differential expression analysis of proteomics data sets. In an ideal world, any multitest correction should show a large statistical power and a small FDR under a small number of comparisons, and its statistical power should increase when the number of tests also increases.

The studied SGoF methodologies exhibited, through the simulation studies and the real case study, well balanced results in terms of FDR and statistical power. The procedures were able to keep FDR relatively low, without sacrificing much of their power. An important aspect of these methods is their scalability with the number of tests performed, while other adjustment methodologies tend to get more conservative, thus indirectly reducing their power, as more comparison tests are performed, the SGoF methods were able to maintain their performance characteristics, and even to increase their statistical power. This last aspect makes them more than suitable to be utilized in differential expression analysis of proteomes.

The developed R routine, used to analyze the case study, was designed to perform differential expression analysis on any LFQ data set outputted by MaxQuant, which makes this tool a suitable complement to other software's used in proteomic laboratories, that focus their studies on biomarkers discovery from LFQ data.

This dissertation also addressed other foundations for future studies that intend to address the statistical challenges in proteomics, such as: the lack of replicates, which translates to a strict statistical outcome; the low prevalence of relevant proteins that truly represent certain conditions and their low relative expression intensity, resulting in a statistical struggle to find such differences. It is important to establish new methodologies for statistical analysis, bearing in mind these limitations, which, although briefly discussed in this study, were not the main focus. To better understand what can and can´t be done in downstream statistical analysis, one must evaluate every step of the proteomics' workflow.

Bibliography:

1. *Proteomics workflow*. 01/09/2020]; Available from: https://planetorbitrap.com/bottom-up-proteomics#tab:overview.
2. *Protein analysis techniques*. [cited 2020 08/08]; Available from: https://www.atascientific.com.au/3-protein-analysis-techniques/.
3. *Transcription, translation and replication*. [cited 2020 2/08]; Available from: https://www.atdbio.com/content/14/Transcription-Translation-and-Replication.
4. Santoiemma, G., *Recent methodologies for studying the soil organic matter.* Applied Soil Ecology, 2018. **123**: p. 546-550.
5. Ho, C.S., et al., *Electrospray ionisation mass spectrometry: principles and clinical applications.* The Clinical Biochemist Reviews, 2003. **24**(1): p. 3.
6. Scigelova, M. and A. Makarov, *Orbitrap mass analyzer–overview and applications in proteomics.* Proteomics, 2006. **6**(S2): p. 16-21.
7. *RNA Splicing*. [cited 2020 3/08]; Available from: https://en.wikipedia.org/wiki/RNA_splicing.
8. *Transcription And Translation*. [cited 2020 1/08]; Available from: https://visualsonline.cancer.gov/details.cfm?imageid=11683.
9. Anderson, N.L. and N.G. Anderson, *Proteome and proteomics: new technologies, new concepts, and new words.* Electrophoresis, 1998. **19**(11): p. 1853-1861.
10. Hughes, C.S., et al., *Single-pot, solid-phase-enhanced sample preparation for proteomics experiments.* Nature protocols, 2019. **14**(1): p. 68-85.
11. Mishra, N.C., *Introduction to proteomics: principles and applications*. Vol. 148. 2011: John Wiley & Sons.
12. Cox, J. and M. Mann, *Quantitative, high-resolution proteomics for data-driven systems biology.* Annual review of biochemistry, 2011. **80**: p. 273-299.
13. Ankney, J.A., A. Muneer, and X. Chen, *Relative and absolute quantitation in mass spectrometry–based proteomics.* Annual Review of Analytical Chemistry, 2018. **11**: p. 49-77.
14. Nirenberg, M., *Deciphering the genetic code.* Office of NIH History, 2010.
15. Tyanova, S., T. Temu, and J. Cox, *The MaxQuant computational platform for mass spectrometry-based shotgun proteomics.* Nature protocols, 2016. **11**(12): p. 2301.
16. Sinitcyn, P., J.D. Rudolph, and J. Cox, *Computational methods for understanding mass spectrometry–based shotgun proteomics data.* Annual Review of Biomedical Data Science, 2018. **1**: p. 207-234.

17. Tyanova, S. and J. Cox, *Perseus: a bioinformatics platform for integrative analysis of proteomics data in cancer research*, in *Cancer systems biology*. 2018, Springer. p. 133-148.
18. Diz, A.P., A. Carvajal-Rodríguez, and D.O. Skibinski, *Multiple hypothesis testing in proteomics: a strategy for experimental work.* Molecular & Cellular Proteomics, 2011. **10**(3).
19. Austin, S.R., I. Dialsingh, and N. Altman, *Multiple hypothesis testing: A review.* J Indian Soc Agric Stat, 2014. **68**(2): p. 303-14.
20. Pascovici, D., et al., *Multiple testing corrections in quantitative proteomics: A useful but blunt tool.* Proteomics, 2016. **16**(18): p. 2448-2453.
21. Ewens, W.J. and G.R. Grant, *Statistical methods in bioinformatics: an introduction*. 2006: Springer Science & Business Media.
22. Holm, S., *A simple sequentially rejective multiple test procedure.* Scandinavian journal of statistics, 1979: p. 65-70.
23. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal statistical society: series B (Methodological), 1995. **57**(1): p. 289-300.
24. Carvajal-Rodríguez, A., J. de Uña-Alvarez, and E. Rolán-Alvarez, *A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests.* BMC bioinformatics, 2009. **10**(1): p. 209.
25. de Uña-Alvarez, J., *On the statistical properties of SGoF multitesting method.* Statistical Applications in Genetics and Molecular Biology, 2011. **10**(1).
26. Urdan, T.C., *Statistics in plain English*. 2016: Taylor & Francis.
27. Brownlee, J., *Statistical methods for machine learning: Discover how to transform data into knowledge with Python*. 2018: Machine Learning Mastery.
28. Casella, G. and R.L. Berger, *Statistical inference*. Vol. 2. 2002: Duxbury Pacific Grove, CA.
29. Banerjee, A., et al., *Hypothesis testing, type I and type II errors.* Industrial psychiatry journal, 2009. **18**(2): p. 127.
30. Westfall, P.H. and S.S. Young, *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Vol. 279. 1993: John Wiley & Sons.
31. Benjamini, Y. and D. Yekutieli, *The control of the false discovery rate in multiple testing under dependency.* Annals of statistics, 2001: p. 1165-1188.
32. Castro-Conde, I. and J. de Uña-Álvarez, *Adjusted p-values for SGoF multiple test procedure.* Biometrical Journal, 2015. **57**(1): p. 108-122.
33. Brown, B.W. and K. Russell, *Methods correcting for multiple testing: operating characteristics.* Statistics in medicine, 1997. **16**(22): p. 2511-2528.

34. Castro-Conde, I. and J. de Uña-Álvarez, *Power, FDR and conservativeness of BB-SGoF method.* Computational Statistics, 2015. **30**(4): p. 1143-1161.
35. Dalmasso, C., P. Broët, and T. Moreau, *A simple procedure for estimating the false discovery rate.* Bioinformatics, 2005. **21**(5): p. 660-668.
36. *Function of Proteins*. [cited 2020 1/08]; Available from: https://courses.lumenlearning.com/wm-biology1/chapter/reading-function-of-proteins/.
37. Haurowitz, D.E.K.a.F. *Protein*. 15/08/2020]; Available from: https://www.britannica.com/science/protein.
38. Spassov, V.Z., L. Yan, and P.K. Flook, *The dominant role of side-chain backbone interactions in structural realization of amino acid code. ChiRotor: A side-chain prediction algorithm based on side-chain backbone interactions.* Protein Science, 2007. **16**(3): p. 494-506.
39. Sung, S.S., *Peptide folding driven by V an der W aals interactions.* Protein Science, 2015. **24**(9): p. 1383-1388.
40. *Protein Secondary Structure: α-Helices and β-Sheets*. 17/08/2020]; Available from: https://proteinstructures.com/Structure/Structure/secondary-sructure.html.
41. Ouellette, R.J. and J.D. Rawn, *Principles of organic chemistry*. 2015: Academic Press.
42. Lequin, R.M., *Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA).* Clinical chemistry, 2005. **51**(12): p. 2415-2418.
43. Wiederschain, G.Y., *The ELISA guidebook*. 2009, Springer Nature BV.
44. Mahmood, T. and P.-C. Yang, *Western blot: technique, theory, and trouble shooting.* North American journal of medical sciences, 2012. **4**(9): p. 429.
45. Kurien, B.T. and R.H. Scofield, *Western blotting.* Methods, 2006. **38**(4): p. 283-293.
46. Gallagher, S., et al., *Immunoblotting and immunodetection.* Current protocols in molecular biology, 2008. **83**(1): p. 10.8. 1-10.8. 28.
47. Finehout, E.J. and K.H. Lee, *An introduction to mass spectrometry applications in biological research.* Biochemistry and molecular biology Education, 2004. **32**(2): p. 93-100.
48. Tyanova, S., et al., *Proteomic maps of breast cancer subtypes.* Nature communications, 2016. **7**(1): p. 1-11.
49. Liu, Y., et al., *Impact of alternative splicing on the human proteome.* Cell reports, 2017. **20**(5): p. 1229-1241.
50. Geyer, P.E., et al., *Plasma proteome profiling to assess human health and disease.* Cell systems, 2016. **2**(3): p. 185-195.

51. Robles, M.S., S.J. Humphrey, and M. Mann, *Phosphorylation is a central mechanism for circadian control of metabolism and physiology.* Cell metabolism, 2017. **25**(1): p. 118-127.
52. Zhu, Y., et al., *Nanodroplet processing platform for deep and quantitative proteome profiling of 10–100 mammalian cells.* Nature communications, 2018. **9**(1): p. 1-10.
53. Barthélemy, N.R., et al., *Tau phosphorylation rates measured by mass spectrometry differ in the intracellular brain vs. extracellular cerebrospinal fluid compartments and are differentially affected by Alzheimer's disease.* Frontiers in aging neuroscience, 2019. **11**: p. 121.
54. Hein, M.Y., et al., *A human interactome in three quantitative dimensions organized by stoichiometries and abundances.* Cell, 2015. **163**(3): p. 712-723.
55. Schneider, M., A. Belsom, and J. Rappsilber, *Protein tertiary structure by crosslinking/mass spectrometry.* Trends in biochemical sciences, 2018. **43**(3): p. 157-169.
56. Mallick, P. and B. Kuster, *Proteomics: a pragmatic perspective.* Nature biotechnology, 2010. **28**(7): p. 695.
57. Chowdhury, S., V. Katta, and B. Chait, *Electrospray ionization mass spectrometric peptide mapping: A rapid, sensitive technique for protein structure analysis.* Biochemical and biophysical research communications, 1990. **167**(2): p. 686-692.
58. Kulyk, D.S., et al., *Direct mass spectrometry analysis of complex mixtures by nanoelectrospray with simultaneous atmospheric pressure chemical ionization and electrophoretic separation capabilities.* Analytical chemistry, 2019. **91**(18): p. 11562-11568.
59. Thiede, B., et al., *Peptide mass fingerprinting.* Methods, 2005. **35**(3): p. 237-247.
60. Forsgard, N., M. Salehpour, and G. Possnert, *Accelerator mass spectrometry in the attomolar concentration range for 14 C-labeled biologically active compounds in complex matrixes.* Journal of Analytical Atomic Spectrometry, 2010. **25**(1): p. 74-78.
61. *Electrospray Ionization Mass Spectrometry*. 20/08/2020]; Available from: https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Supplemental_Modules_(Analytical_Chemistry)/Instrumental_Analysis/Mass_Spectrometry/Mass_Spectrometers_(Instrumentation)/Electrospray_Ionization_Mass_Spectrometry.
62. Ubelaker, D., *Encyclopedia of forensic sciences*. 2013, Academic Press Waltham (MA). p. 603-608.
63. *Orbitrap mass spectrometers offer the highest levels of accuracy and precision*. 23/08/2020]; Available from: https://www.thermofisher.com/pt/en/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-systems/orbitrap-lc-ms.html.

64. *Mass Spec*. 23/08/2020]; Available from: https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Supplemental_Modules_(Analytical_Chemistry)/Instrumental_Analysis/Mass_Spectrometry/Mass_Spec.
65. *Overview of Mass Spectrometry for Protein Analysis*. 24/08/2020]; Available from: https://www.thermofisher.com/pt/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-mass-spectrometry.html.
66. *Principle of HPLC*. 27/08/2020]; Available from: https://www.knauer.net/en/Systems-Solutions/Analytical-HPLC-UHPLC/HPLC-Basics---principles-and-parameters.
67. *MaxQuant*. 09/08/2020]; Available from: http://coxdocs.org/doku.php?id=maxquant:start.
68. *Perseus*. 09/08/2020]; Available from: http://coxdocs.org/doku.php?id=perseus:start.
69. *Proteomes - Mus musculus (Mouse)*. 22/05/2020]; Available from: https://ebi14.uniprot.org/proteomes/UP000000589.
70. Lazar, C., et al., *Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies.* Journal of proteome research, 2016. **15**(4): p. 1116-1125.
71. Choi, M., et al., *ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of Differentially Abundant Proteins in Label-Free Quantitative LC–MS/MS Experiments.* Journal of proteome research, 2017. **16**(2): p. 945-957.

Supplemental files

***Table 3**: Mean percentages of significant cases deteced when the null hypothesis was always true. The simulated null models were tested under t tests. n: sample sizes. S: number of tests. Significant %: Percentage of significant tests at 5% significance. Detect_Bonf %: Percentage of tests corrected by Bonferroni and detected at 5% significance. Detect_BH %: Percentage of tests corrected by Benjamini-Hochberg and detected at 5% significance. Detect_SGoF %: Percentage of tests corrected by Binomial SGoF and detected at 5% significance. Values are avereges through 1000 replicates and their ± stardart deviations.*

n	S	Significant %	Detect_Bonf %	Detect_BH %	Detect_SGoF %
5	100	5.103 ± 2.2279	0.057 ± 0.2362	0.063 ± 0.2629	0.063 ± 0.3939
5	1000	4.9717 ± 0.6725	0.0047 ± 0.0216	0.0052 ± 0.0248	0.0085 ± 0.0681
5	10000	5.0156 ± 0.2153	4e-04 ± 0.0019	4e-04 ± 0.0023	0.0049 ± 0.0284
10	100	4.989 ± 2.2081	0.048 ± 0.2139	0.055 ± 0.245	0.053 ± 0.3071
10	1000	5.0355 ± 0.6782	0.004 ± 0.0201	0.0046 ± 0.0237	0.0153 ± 0.0896
10	10000	4.9919 ± 0.2228	5e-04 ± 0.0024	6e-04 ± 0.0026	0.0045 ± 0.026
20	100	5.062 ± 2.1389	0.044 ± 0.2052	0.049 ± 0.2338	0.054 ± 0.351
20	1000	5.0275 ± 0.7039	0.0054 ± 0.0235	0.0063 ± 0.0281	0.0167 ± 0.0958
20	10000	5.0014 ± 0.2186	5e-04 ± 0.0022	5e-04 ± 0.0024	0.0053 ± 0.0311

Table 4: *Percentages of significant cases detected at 5% significance (Detect_method), false discovery rate (FDR_method) and Power (Power_methd) after multitest adjustment when the p-values come from families of one-sample t tests where some (% effect) of the alternative hypotheses were true.The alternative hypothesis comes from a N(0.36,1). n: sample size. Prevalence %: Percentage of true alternatives. S: number of tests. Significant %: Percentage of significants before adjustment at 5% significance. Detect_method %: Percentage of significance tests detected after adjustments (Bonferroni, Benjamini-Hochberg, Binomial SGoF) at 5% significance. FDR_method %: Percentage of false discovery rate (Bonferroni, Benjamini-Hochberg and Binomial SGoF). Power_method %: Percentage of statistical power (Bonferroni, Benjamini-Hochberg and Binomial SGoF). Values are avereges through 1000 replicates and their ± stardart deviations.*

n	Prevalence %	S	Significant %	Detect_Bonf %	FDR_Bonf %	Detect_BH %	FDR_BH %	Detect_SGoF %	FDR_SGoF %	Power_Bonf %	Power_BH %	Power_SGoF %
5	5	100	5.264 ± 2.2502	0.053 ± 0.2372	84 ± 37.0328	0.059 ± 0.2676	85.2941 ± 35.063	0.068 ± 0.3944	87.963 ± 29.7105	0.0016 ± 0.0178	0.0016 ± 0.0178	0.0012 ± 0.0155
5	5	1000	5.2354 ± 0.7008	0.0051 ± 0.0229	88.7755 ± 31.0666	0.0053 ± 0.0245	88.7755 ± 31.0666	0.0325 ± 0.1309	91.6407 ± 14.0543	1e-04 ± 0.0015	1e-04 ± 0.0015	7e-04 ± 0.0052
5	5	10000	5.2312 ± 0.2252	5e-04 ± 0.0023	90.3846 ± 29.7678	6e-04 ± 0.0026	90.3846 ± 29.7678	0.0385 ± 0.0866	89.7329 ± 14.4042	0 ± 1e-04	0 ± 1e-04	8e-04 ± 0.0022
5	10	100	5.421 ± 2.3823	0.067 ± 0.258	83.0769 ± 37.7874	0.086 ± 0.3671	83.2108 ± 36.3239	0.101 ± 0.5187	85.8654 ± 30.8777	0.0011 ± 0.0104	0.0014 ± 0.0126	0.0013 ± 0.0122
5	10	1000	5.4756 ± 0.7149	0.005 ± 0.0223	79.5918 ± 40.7206	0.0054 ± 0.0247	80 ± 40.4061	0.0589 ± 0.2062	83.0309 ± 24.4361	1e-04 ± 0.001	1e-04 ± 0.001	0.001 ± 0.0047
5	10	10000	5.4687 ± 0.218	6e-04 ± 0.0024	61.8182 ± 49.031	7e-04 ± 0.0028	63.5593 ± 48.0934	0.152 ± 0.161	79.0162 ± 12.8818	0 ± 1e-04	0 ± 2e-04	0.0031 ± 0.0035
5	20	100	5.828 ± 2.352	0.066 ± 0.2524	60.7692 ± 48.8079	0.069 ± 0.2689	61.5385 ± 48.2257	0.12 ± 0.4918	59.2254 ± 45.0633	0.0013 ± 0.008	0.0014 ± 0.0084	0.0023 ± 0.0126
5	20	1000	5.9738 ± 0.7438	0.0055 ± 0.0237	66.0377 ± 47.8113	0.0063 ± 0.0274	66.3636 ± 46.2026	0.201 ± 0.3623	62.4284 ± 28.6703	1e-04 ± 7e-04	1e-04 ± 7e-04	0.0037 ± 0.0073
5	20	10000	5.9383 ± 0.2253	7e-04 ± 0.0027	52.8986 ± 49.9147	8e-04 ± 0.003	53.8095 ± 49.1614	0.5786 ± 0.2244	64.0158 ± 7.5027	0 ± 1e-04	0 ± 1e-04	0.0103 ± 0.0043
10	5	100	5.633 ± 2.3837	0.079 ± 0.2736	71.1538 ± 45.2405	0.091 ± 0.3175	71.9512 ± 44.5121	0.105 ± 0.516	71.7687 ± 37.743	0.0046 ± 0.03	0.005 ± 0.0325	0.0054 ± 0.0359
10	5	1000	5.6117 ± 0.7297	0.0065 ± 0.0266	73.3333 ± 43.6343	0.0087 ± 0.0382	69.3122 ± 42.2228	0.0874 ± 0.235	74.0855 ± 28.3552	3e-04 ± 0.0026	5e-04 ± 0.0033	0.0042 ± 0.0136
10	5	10000	5.6203 ± 0.22	8e-04 ± 0.0029	61.6034 ± 48.6538	9e-04 ± 0.0033	62.3984 ± 47.4891	0.2724 ± 0.1997	73.8207 ± 12.5689	1e-04 ± 3e-04	1e-04 ± 4e-04	0.0136 ± 0.0105
10	10	100	6.285 ± 2.4256	0.08 ± 0.2787	48.7179 ± 50.3071	0.091 ± 0.3238	50.4115 ± 48.4464	0.19 ± 0.6545	64.8867 ± 39.2198	0.004 ± 0.0196	0.0044 ± 0.021	0.0066 ± 0.0286

10	10	1000	6.2766 ± 0.7227	0.01 ± 0.0332	49.0842 ± 48.8992	0.012 ± 0.0405	46.6667 ± 46.9898	0.3255 ± 0.4579	60.8263 ± 26.1992	5e-04 ± 0.0022	6e-04 ± 0.0027	0.0124 ± 0.019
10	10	10000	6.2468 ± 0.2363	8e-04 ± 0.0029	48.7179 ± 48.9993	0.0011 ± 0.0039	49.0196 ± 47.3751	0.8868 ± 0.2363	61.0902 ± 6.0309	0 ± 2e-04	1e-04 ± 3e-04	0.0342 ± 0.0094
10	20	100	7.377 ± 2.5078	0.121 ± 0.3442	30 ± 45.306	0.145 ± 0.4474	32.1083 ± 43.8379	0.413 ± 1.0254	43.1914 ± 40.7541	0.0042 ± 0.0142	0.0049 ± 0.0167	0.0111 ± 0.0317
10	20	1000	7.4765 ± 0.8073	0.0116 ± 0.0336	31.5315 ± 46.6749	0.0157 ± 0.0491	31.5826 ± 43.9605	1.2943 ± 0.7721	42.196 ± 17.3784	4e-04 ± 0.0014	5e-04 ± 0.0018	0.0366 ± 0.0224
10	20	10000	7.4887 ± 0.2602	0.0015 ± 0.0038	28.8732 ± 45.0862	0.0019 ± 0.0051	28.5794 ± 42.6049	2.1288 ± 0.2602	44.6229 ± 3.4636	1e-04 ± 2e-04	1e-04 ± 2e-04	0.0589 ± 0.0075
20	5	100	6.387 ± 2.2831	0.144 ± 0.3787	40.6173 ± 48.5532	0.171 ± 0.4669	43.6451 ± 46.5127	0.164 ± 0.5979	53.4722 ± 45.4295	0.0172 ± 0.0589	0.0194 ± 0.0656	0.014 ± 0.0636
20	5	1000	6.4206 ± 0.7414	0.0203 ± 0.0452	19.0346 ± 38.3239	0.0271 ± 0.0634	22.9058 ± 37.9382	0.4183 ± 0.5109	48.5855 ± 26.742	0.0033 ± 0.0081	0.0041 ± 0.0106	0.0402 ± 0.0492
20	5	10000	6.4098 ± 0.2443	0.0028 ± 0.0053	16.0468 ± 35.6502	0.0041 ± 0.0089	18.8524 ± 34.4554	1.0498 ± 0.2443	53.5476 ± 5.45	5e-04 ± 0.001	7e-04 ± 0.0015	0.0964 ± 0.0205
20	10	100	7.935 ± 2.5427	0.205 ± 0.4617	15.93 ± 35.9536	0.257 ± 0.6126	16.3298 ± 34.1579	0.566 ± 1.2054	37.0655 ± 37.4583	0.0174 ± 0.0429	0.0211 ± 0.0524	0.0331 ± 0.0747
20	10	1000	7.8556 ± 0.7893	0.0345 ± 0.0594	13.6207 ± 32.9463	0.0564 ± 0.1077	14.9805 ± 30.5856	1.6593 ± 0.7804	35.7308 ± 13.9916	0.003 ± 0.0056	0.0047 ± 0.0093	0.1033 ± 0.0471
20	10	10000	7.8302 ± 0.2593	0.0048 ± 0.0066	8.5043 ± 25.1738	0.0109 ± 0.0173	9.7961 ± 22.3205	2.4702 ± 0.2593	41.0596 ± 3.2337	4e-04 ± 6e-04	0.001 ± 0.0015	0.1453 ± 0.0147
20	20	100	10.738 ± 3.0263	0.388 ± 0.6098	6.6363 ± 23.6928	0.579 ± 0.932	7.9917 ± 22.6938	2.231 ± 2.4015	20.9751 ± 25.4218	0.0181 ± 0.0295	0.0264 ± 0.0428	0.0839 ± 0.0891
20	20	1000	10.7135 ± 0.9379	0.0633 ± 0.0793	6.1748 ± 21.8608	0.1509 ± 0.205	7.9155 ± 19.6644	4.5135 ± 0.9379	25.8318 ± 6.6427	0.003 ± 0.0038	0.0069 ± 0.0094	0.1664 ± 0.0335
20	20	10000	10.6662 ± 0.2905	0.009 ± 0.0093	4.2497 ± 18.5264	0.0398 ± 0.0476	5.724 ± 14.5125	5.3062 ± 0.2905	27.2802 ± 2.0315	4e-04 ± 5e-04	0.0019 ± 0.0022	0.1928 ± 0.0102

***Table 5**: Percentages of significant cases detected at 5% significance (Detect_method), false discovery rate (FDR_method) and Power (Power_methd) after multitest adjustment when the p-values come from families of one-sample t tests where some (% effect) of the alternative hypotheses were true.The alternative hypothesis comes from a N(0.7,1). n: sample size. Prevalence %: Percentage of true alternatives. S: number of tests. Significant %: Percentage of significants before adjustment at 5% significance. Detect_method %: Percentage of significance tests detected after adjustments (Bonferroni, Benjamini-Hochberg, Binomial SGoF) at 5% significance. FDR_method %: Percentage of false discovery rate (Bonferroni, Benjamini-Hochberg and Binomial SGoF). Power_method %: Percentage of statistical power (Bonferroni, Benjamini-Hochberg and Binomial SGoF). Values are avereges through 1000 replicates and their ± stardart deviations.*

n	Prevalence %	S	Significant %	Detect_Bonf %	FDR_Bonf %	Detect_BH %	FDR_BH %	Detect_SGoF %	FDR_SGoF %	Power_Bonf %	Power_BH %	Power_SGoF %
5	5	100	6.019 ± 2.3178	0.071 ± 0.2608	62.8571 ± 48.6675	0.081 ± 0.3075	62.5 ± 48.752	0.135 ± 0.5376	79.7186 ± 35.8006	0.0052 ± 0.0318	0.0056 ± 0.0342	0.0052 ± 0.0365
5	5	1000	5.8946 ± 0.743	0.0077 ± 0.0278	68.9189 ± 45.8577	0.0087 ± 0.0322	71.2719 ± 43.3418	0.1729 ± 0.3454	75.6132 ± 24.4068	5e-04 ± 0.0032	5e-04 ± 0.0035	0.0084 ± 0.0199
5	5	10000	5.889 ± 0.2266	9e-04 ± 0.0029	66.6667 ± 47.4236	0.001 ± 0.0036	67.8161 ± 45.8055	0.5293 ± 0.2257	73.6297 ± 7.617	1e-04 ± 3e-04	1e-04 ± 4e-04	0.0277 ± 0.013
5	10	100	6.905 ± 2.4888	0.088 ± 0.3072	59.8765 ± 48.3605	0.111 ± 0.401	59.7701 ± 47.3661	0.297 ± 0.8897	66.4336 ± 38.0085	0.0035 ± 0.0189	0.0044 ± 0.0228	0.0094 ± 0.0351
5	10	1000	6.7575 ± 0.771 5	0.0078 ± 0.0279	58.6667 ± 48.8885	0.0102 ± 0.0371	58.8353 ± 46.3679	0.6579 ± 0.6278	57.9361 ± 23.0561	3e-04 ± 0.0018	4e-04 ± 0.0025	0.0269 ± 0.0277
5	10	10000	6.7834 ± 0.2434	9e-04 ± 0.003	61.3095 ± 48.6883	0.0011 ± 0.0039	62.9213 ± 46.3131	1.4234 ± 0.2434	59.3983 ± 4.1051	0 ± 2e-04	0 ± 2e-04	0.0577 ± 0.0109
5	20	100	8.341 ± 2.6037	0.108 ± 0.3353	32.5 ± 46.2618	0.157 ± 0.4966	35.2339 ± 43.1952	0.738 ± 1.3665	38.9552 ± 38.9469	0.0036 ± 0.0135	0.005 ± 0.0177	0.0225 ± 0.0459
5	20	1000	8.5417 ± 0.8559	0.0116 ± 0.0339	32.7273 ± 46.1531	0.0145 ± 0.0431	33.9031 ± 44.3088	2.343 ± 0.8518	39.9576 ± 11.2784	4e-04 ± 0.0014	5e-04 ± 0.0017	0.0696 ± 0.0268
5	20	10000	8.5448 ± 0.2625	0.0013 ± 0.0035	33.0709 ± 46.8112	0.0018 ± 0.0048	31.4286 ± 43.4377	3.1848 ± 0.2625	41.1616 ± 2.7511	0 ± 1e-04	1e-04 ± 2e-04	0.0936 ± 0.0081
10	5	100	7.349 ± 2.3918	0.205 ± 0.4418	24.3386 ± 42.0892	0.262 ± 0.5914	25 ± 40.1265	0.383 ± 0.9483	45.2604 ± 40.4532	0.031 ± 0.0772	0.0386 ± 0.0963	0.0392 ± 0.1077
10	5	1000	7.2508 ± 0.7634	0.0254 ± 0.0516	17.4208 ± 36.3337	0.0365 ± 0.0832	18.1583 ± 33.5329	1.078 ± 0.7131	38.4708 ± 17.5103	0.0042 ± 0.0091	0.0058 ± 0.0135	0.1258 ± 0.079
10	5	10000	7.2832 ± 0.2455	0.0032 ± 0.0056	14.9697 ± 34.4572	0.006 ± 0.0119	16.1666 ± 30.8833	1.9232 ± 0.2455	44.4314 ± 3.8376	6e-04 ± 0.0011	0.001 ± 0.002	0.2128 ± 0.0236
10	10	100	9.538 ± 2.6171	0.32 ± 0.5421	12.6316 ± 31.929	0.491 ± 0.8802	13.75 ± 30.1616	1.327 ± 1.7513	26.3145 ± 30.7473	0.0279 ± 0.0507	0.0424 ± 0.0797	0.0944 ± 0.1276

10	10	1000	9.5873 ± 0.8386	0.0457 ± 0.07	7.8404 ± 25.311	0.0955 ± 0.1646	10.1182 ± 23.3214	3.3873 ± 0.8386	29.0678 ± 8.0884	0.0042 ± 0.0068	0.0086 ± 0.015	0.2371 ± 0.0525
10	10	10000	9.5657 ± 0.2636	0.006 ± 0.0078	7.7154 ± 24.4386	0.0196 ± 0.029	7.4467 ± 18.2147	4.2057 ± 0.2636	32.2371 ± 2.3805	6e-04 ± 8e-04	0.0018 ± 0.0027	0.2848 ± 0.0171
10	20	100	14.062 ± 2.9283	0.605 ± 0.7304	6.9002 ± 22.9897	1.148 ± 1.5403	8.6623 ± 20.5473	5.107 ± 2.8322	15.1951 ± 16.7978	0.028 ± 0.035	0.0516 ± 0.0702	0.211 ± 0.1146
10	20	1000	14.1243 ± 0.928	0.0904 ± 0.0931	5.3439 ± 19.6427	0.3759 ± 0.4164	5.8522 ± 13.2087	7.9243 ± 0.928	19.2586 ± 4.5392	0.0043 ± 0.0045	0.0176 ± 0.0194	0.3192 ± 0.0356
10	20	10000	14.1322 ± 0.3009	0.0116 ± 0.0106	4.0445 ± 17.596	0.1323 ± 0.1213	4.5444 ± 8.2328	8.7722 ± 0.3009	20.3467 ± 1.3606	6e-04 ± 5e-04	0.0063 ± 0.0057	0.3493 ± 0.0118
20	5	100	9.027 ± 2.2448	1.052 ± 0.9067	4.4899 ± 17.8785	1.613 ± 1.4004	6.5854 ± 17.2655	0.887 ± 1.385	15.9107 ± 26.164	0.2006 ± 0.1771	0.2938 ± 0.2487	0.1384 ± 0.2067
20	5	1000	8.963 ± 0.7252	0.3151 ± 0.1766	1.1065 ± 6.6886	1.1824 ± 0.543	4.4623 ± 6.4893	2.763 ± 0.7252	14.7661 ± 8.3517	0.0623 ± 0.0351	0.2245 ± 0.1014	0.4628 ± 0.0986
20	5	10000	8.9678 ± 0.2207	0.0703 ± 0.0266	0.4929 ± 2.7255	1.1328 ± 0.1833	4.6964 ± 2.0219	3.6078 ± 0.2207	20.6834 ± 2.8147	0.014 ± 0.0053	0.2157 ± 0.034	0.5714 ± 0.0248
20	10	100	12.974 ± 2.3814	2.099 ± 1.3026	2.0549 ± 9.4673	4.134 ± 2.3811	4.8544 ± 10.0967	3.994 ± 2.3412	8.1788 ± 13.9591	0.205 ± 0.1285	0.3881 ± 0.2184	0.3557 ± 0.1942
20	10	1000	12.928 ± 0.7597	0.6134 ± 0.2381	0.8061 ± 4.0183	3.7375 ± 0.8037	4.3957 ± 3.5969	6.728 ± 0.7597	11.9553 ± 4.6115	0.0609 ± 0.0238	0.3566 ± 0.0744	0.5903 ± 0.0542
20	10	10000	12.9271 ± 0.2427	0.1413 ± 0.0381	0.3279 ± 1.5073	3.7254 ± 0.2586	4.5081 ± 1.0534	7.5671 ± 0.2427	14.7787 ± 1.4951	0.0141 ± 0.0038	0.3557 ± 0.024	0.6446 ± 0.016
20	20	100	20.891 ± 2.453	4.088 ± 1.8509	0.8189 ± 4.4954	10.737 ± 3.419	4.0554 ± 6.1436	11.891 ± 2.453	5.593 ± 6.7515	0.2024 ± 0.0916	0.5124 ± 0.1582	0.558 ± 0.106
20	20	1000	20.8676 ± 0.783	1.2078 ± 0.3403	0.3243 ± 1.6549	10.626 ± 1.1116	4.1082 ± 1.9711	14.6676 ± 0.783	8.2883 ± 2.4174	0.0602 ± 0.017	0.5092 ± 0.0515	0.6721 ± 0.0305
20	20	10000	20.8656 ± 0.2568	0.2796 ± 0.0511	0.1639 ± 0.7674	10.6238 ± 0.3449	4.0271 ± 0.6	15.5056 ± 0.2568	9.3259 ± 0.8203	0.014 ± 0.0026	0.5098 ± 0.016	0.7029 ± 0.0096

Table 6: *Percentages of significant cases detected at 5% significance (Detect_method), false discovery rate (FDR_method) and Power (Power_methd) after multitest adjustment when the p-values come from families of one-sample t tests where some (% effect) of the alternative hypotheses were true.The alternative hypothesis comes from a N(0.96,1). n: sample size. Prevalence %: Percentage of true alternatives. S: number of tests. Significant %: Percentage of significants before adjustment at 5% significance. Detect_method %: Percentage of significance tests detected after adjustments (Bonferroni, Benjamini-Hochberg, Binomial SGoF) at 5% significance. FDR_method %: Percentage of false discovery rate (Bonferroni, Benjamini-Hochberg and Binomial SGoF). Power_method %: Percentage of statistical power (Bonferroni, Benjamini-Hochberg and Binomial SGoF). Values are avereges through 1000 replicates and their ± stardart deviations.*

n	Prevalence %	S	Significant %	Detect_Bonf %	FDR_Bonf %	Detect_BH %	FDR_BH %	Detect_SGoF %	FDR_SGoF %	Power_Bonf %	Power_BH %	Power_SGoF %
5	5	100	6.6 ± 2.2854	0.115 ± 0.3315	47.2973 ± 49.4689	0.149 ± 0.4765	48.6232 ± 47.2428	0.199 ± 0.6811	65.8 ± 38.7034	0.0122 ± 0.0487	0.0142 ± 0.0544	0.0124 ± 0.0567
5	5	1000	6.6632 ± 0.7543	0.0095 ± 0.0316	53.4091 ± 49.5934	0.0106 ± 0.0362	54.0741 ± 48.1435	0.5865 ± 0.5896	59.9693 ± 23.6837	9e-04 ± 0.0044	0.001 ± 0.0048	0.0454 ± 0.0489
5	5	10000	6.6403 ± 0.2432	9e-04 ± 0.003	54.2169 ± 49.5128	0.0011 ± 0.0038	53.7879 ± 48.1276	1.2803 ± 0.2432	60.2228 ± 4.4808	1e-04 ± 4e-04	1e-04 ± 5e-04	0.1013 ± 0.0198
5	10	100	8.265 ± 2.6049	0.128 ± 0.3518	39.3443 ± 48.6298	0.166 ± 0.474	39.5674 ± 45.9741	0.708 ± 1.3633	42.0009 ± 38.4654	0.0078 ± 0.0279	0.0099 ± 0.0343	0.0381 ± 0.0804
5	10	1000	8.3053 ± 0.8221	0.0121 ± 0.0338	26.9231 ± 44.3042	0.0165 ± 0.0484	26.378 ± 42.0192	2.1056 ± 0.8213	42.7313 ± 12.2634	9e-04 ± 0.003	0.0013 ± 0.0042	0.119 ± 0.0483
5	10	10000	8.309 ± 0.2662	0.0014 ± 0.0037	27.9487 ± 43.9841	0.0017 ± 0.0048	27.3551 ± 41.9394	2.949 ± 0.2662	44.7885 ± 2.9031	1e-04 ± 3e-04	1e-04 ± 4e-04	0.1626 ± 0.0153
5	20	100	11.68 ± 2.9162	0.202 ± 0.451	18.7845 ± 37.7221	0.333 ± 0.7904	19.9257 ± 33.8481	2.946 ± 2.5262	24.814 ± 25.8487	0.0082 ± 0.02	0.0131 ± 0.0321	0.1074 ± 0.0938
5	20	1000	11.7037 ± 0.941	0.0206 ± 0.0447	24.515 ± 42.0565	0.031 ± 0.0741	24.9083 ± 39.217	5.5037 ± 0.941	28.0573 ± 6.0022	8e-04 ± 0.0019	0.0012 ± 0.0032	0.1975 ± 0.0349
5	20	10000	11.6332 ± 0.2924	0.002 ± 0.0045	19.8276 ± 38.6996	0.0034 ± 0.0085	18.531 ± 33.9739	6.2732 ± 0.2924	28.5868 ± 1.7662	1e-04 ± 2e-04	1e-04 ± 4e-04	0.2239 ± 0.0105
10	5	100	8.6 ± 2.3003	0.506 ± 0.7003	6.2913 ± 21.9608	0.737 ± 1.0578	8.2132 ± 21.3599	0.717 ± 1.2941	24.6687 ± 31.8818	0.0942 ± 0.1335	0.1316 ± 0.1868	0.0992 ± 0.1758
10	5	1000	8.6572 ± 0.7329	0.0837 ± 0.0913	5.1406 ± 18.2911	0.2451 ± 0.2853	6.1023 ± 14.0145	2.4572 ± 0.7329	22.8219 ± 9.8144	0.0157 ± 0.0175	0.0451 ± 0.0519	0.3711 ± 0.0943
10	5	10000	8.6357 ± 0.2258	0.0125 ± 0.0113	3.7176 ± 15.9799	0.1058 ± 0.0914	5.3424 ± 9.0577	3.2757 ± 0.2258	28.2741 ± 2.8601	0.0024 ± 0.0022	0.0199 ± 0.0171	0.4691 ± 0.026
10	10	100	12.122 ± 2.4093	0.962 ± 0.9452	3.9208 ± 17.2302	1.906 ± 1.8581	4.9371 ± 14.7961	3.2 ± 2.2723	11.4439 ± 17.0514	0.0923 ± 0.0922	0.1787 ± 0.1705	0.2718 ± 0.1829
10	10	1000	12.3333 ± 0.7842	0.1691 ± 0.1242	2.7751 ± 13.4965	1.1524 ± 0.7451	5.1267 ± 7.8924	6.1333 ± 0.7842	17.3998 ± 5.2104	0.0165 ± 0.0124	0.1085 ± 0.0693	0.5044 ± 0.054

10	10	10000	12.2924 ± 0.2419	0.0245 ± 0.0163	1.946 ± 11.2212	0.9293 ± 0.2848	4.5545 ± 2.2875	6.9324 ± 0.2419	19.7633 ± 1.6064	0.0024 ± 0.0016	0.0885 ± 0.0268	0.556 ± 0.0159
10	20	100	19.553 ± 2.749	1.885 ± 1.2639	1.9525 ± 10.138	5.842 ± 3.62	3.9532 ± 7.836	10.553 ± 2.749	8.2264 ± 8.6859	0.0922 ± 0.0621	0.2774 ± 0.1685	0.4791 ± 0.1135
10	20	1000	19.5509 ± 0.8675	0.3404 ± 0.1831	1.1638 ± 7.0068	5.3462 ± 1.4029	3.9108 ± 2.6807	13.3509 ± 0.8675	11.0958 ± 2.8283	0.0168 ± 0.0091	0.2564 ± 0.066	0.5929 ± 0.0341
10	20	10000	19.5665 ± 0.268	0.0487 ± 0.0216	0.8612 ± 4.5364	5.2702 ± 0.4394	3.9885 ± 0.8676	14.2065 ± 0.268	12.1616 ± 0.8823	0.0024 ± 0.0011	0.253 ± 0.0208	0.6239 ± 0.0106
20	5	100	9.702 ± 2.0551	2.876 ± 1.118	1.3682 ± 6.9831	4.122 ± 1.2493	5.1747 ± 9.9993	1.172 ± 1.5423	3.4212 ± 11.3594	0.566 ± 0.2195	0.7716 ± 0.2146	0.22 ± 0.2768
20	5	1000	9.6566 ± 0.6767	1.3831 ± 0.3064	0.2361 ± 1.3107	3.9862 ± 0.4064	4.7114 ± 3.3455	3.4566 ± 0.6767	3.6359 ± 3.9038	0.2759 ± 0.0611	0.7585 ± 0.0703	0.6627 ± 0.1126
20	5	10000	9.6702 ± 0.2107	0.4737 ± 0.0663	0.1032 ± 0.4691	3.9815 ± 0.137	4.7439 ± 1.0618	4.3102 ± 0.2107	6.683 ± 1.7949	0.0946 ± 0.0132	0.7584 ± 0.0236	0.8038 ± 0.0281
20	10	100	14.333 ± 2.0475	5.791 ± 1.5581	0.9677 ± 4.0308	8.851 ± 1.5541	4.3975 ± 6.6027	5.334 ± 2.0447	1.6367 ± 5.1442	0.5732 ± 0.1548	0.8412 ± 0.1301	0.5205 ± 0.1874
20	10	1000	14.3732 ± 0.6908	2.7255 ± 0.4517	0.1736 ± 0.863	8.8981 ± 0.5034	4.592 ± 2.2222	8.1732 ± 0.6908	3.3128 ± 2.4224	0.2721 ± 0.0452	0.8484 ± 0.0421	0.7891 ± 0.055
20	10	10000	14.3359 ± 0.2109	0.9461 ± 0.0915	0.0372 ± 0.1935	8.867 ± 0.1531	4.4913 ± 0.6921	8.9759 ± 0.2109	4.8493 ± 0.9585	0.0946 ± 0.0091	0.8468 ± 0.0129	0.8539 ± 0.0145
20	20	100	23.655 ± 1.9626	11.424 ± 2.203	0.3471 ± 1.7273	19.073 ± 1.696	3.9305 ± 4.4045	14.655 ± 1.9626	1.2723 ± 2.9431	0.5692 ± 0.1102	0.9141 ± 0.0679	0.7223 ± 0.0895
20	20	1000	23.6697 ± 0.6401	5.4235 ± 0.6282	0.0508 ± 0.3071	19.0111 ± 0.5383	3.9976 ± 1.4612	17.4697 ± 0.6401	2.2956 ± 1.322	0.271 ± 0.0314	0.9123 ± 0.0213	0.8532 ± 0.0262
20	20	10000	23.6807 ± 0.1919	1.893 ± 0.1325	0.0152 ± 0.0884	18.9934 ± 0.1807	3.998 ± 0.4473	18.3207 ± 0.1919	3.0597 ± 0.4508	0.0946 ± 0.0066	0.9117 ± 0.0069	0.888 ± 0.0072

Table 7: *Confusion matrices (top) and corresponding evaluation matrices (bottom) of the ANOVA (left) and the contrasts (right), both corrected by Binomial SGoF (BS).*

ANOVA_BS_BS		
Column1	**POSITIVE**	**NEGATIVE**
Rejected	6	14
Not_rejected	0	2600

Contrasts_BS_BS		
Column1	**POSITIVE**	**NEGATIVE**
Rejected	26	27
Not_rejected	10	57

ANOVA_BS_BS_eval_metrics		
Column1	**Metric**	**Value**
1	Accuracy	0.995
2	Precision	0.300
3	Sensitivity_Power	1.000
4	Specificity	0.995
5	FDR	0.700

Contrasts_BS_BS_eval_metrics		
Column1	**Metric**	**Value**
1	Accuracy	0.692
2	Precision	0.491
3	Sensitivity_Power	0.722
4	Specificity	0.679
5	FDR	0.509

Table 8: *Confusion matrices (top) and corresponding evaluation matrices (bottom) of the ANOVA (left) and the contrasts (right), both corrected by Conservative SGoF (CS)..*

ANOVA_CS_CS		
Column1	**POSITIVE**	**NEGATIVE**
Rejected	6	11
Not_rejected	0	2603

Contrasts_CS_CS		
Column1	**POSITIVE**	**NEGATIVE**
Rejected	22	18
Not_rejected	14	48

ANOVA_CS_CS_eval_metrics		
Column1	**Metric**	**Value**
1	Accuracy	0.996
2	Precision	0.353
3	Sensitivity_Power	1.000
4	Specificity	0.996
5	FDR	0.647

Contrasts_CS_CS_eval_metrics		
Column1	**Metric**	**Value**
1	Accuracy	0.686
2	Precision	0.550
3	Sensitivity_Power	0.611
4	Specificity	0.727
5	FDR	0.450

Table 9: *Confusion matrices (top) and corresponding evaluation matrices (bottom) of the ANOVA (left) and the contrasts (right), both corrected by Benjamini-Hochberg (BH)..*

ANOVA_BH_BH		
Column1	**POSITIVE**	**NEGATIVE**
Rejected	6	9
Not_rejected	0	2605

Contrasts_BH_BH		
Column1	**POSITIVE**	**NEGATIVE**
Rejected	31	19
Not_rejected	5	35

ANOVA_BH_BH_eval_metrics		
Column1	**Metric**	**Value**
1	Accuracy	0.997
2	Precision	0.400
3	Sensitivity_Power	1.000
4	Specificity	0.997
5	FDR	0.600

Contrasts_BH_BH_eval_metrics		
Column1	**Metric**	**Value**
1	Accuracy	0.733
2	Precision	0.620
3	Sensitivity_Power	0.861
4	Specificity	0.648
5	FDR	0.380

Table 10: *Confusion matrices (top) and corresponding evaluation matrices (bottom) of the ANOVA (left) and the contrasts (right), both corrected by Benjamini–Yekutieli (BY).*

ANOVA_BY_BY		
Column1	**POSITIVE**	**NEGATIVE**
Rejected	4	6
Not_rejected	2	2608

Contrasts_BY_BY		
Column1	**POSITIVE**	**NEGATIVE**
Rejected	18	12
Not_rejected	6	24

ANOVA_BY_BY_eval_metrics		
Column1	**Metric**	**Value**
1	Accuracy	0.997
2	Precision	0.400
3	Sensitivity_Power	0.667
4	Specificity	0.998
5	FDR	0.600

Contrasts_BY_BY_eval_metrics		
Column1	**Metric**	**Value**
1	Accuracy	0.700
2	Precision	0.600
3	Sensitivity_Power	0.750
4	Specificity	0.667
5	FDR	0.400

Table 11: *Confusion matrices (top) and corresponding evaluation matrices (bottom) of the ANOVA (left) and the contrasts (right), both corrected by Bonferroni (Bonf).*

ANOVA_Bonf_Bonf		
Column1	**POSITIVE**	**NEGATIVE**
Rejected	4	5
Not_rejected	2	2609

Contrasts_Bonf_Bonf		
Column1	**POSITIVE**	**NEGATIVE**
Rejected	15	8
Not_rejected	9	22

ANOVA_Bonf_Bonf_eval_metrics		
Column1	**Metric**	**Value**
1	Accuracy	0.997
2	Precision	0.444
3	Sensitivity_Power	0.667
4	Specificity	0.998
5	FDR	0.556

Contrasts_Bonf_Bonf_eval_metrics		
Column1	**Metric**	**Value**
1	Accuracy	0.685
2	Precision	0.652
3	Sensitivity_Power	0.625
4	Specificity	0.733
5	FDR	0.348

Table 12: *Confusion matrices (top) and corresponding evaluation matrices (bottom) of the ANOVA (left) and the contrasts (right), both corrected by Šidák (S).*

ANOVA_S_S		
Column1	**POSITIVE**	**NEGATIVE**
Rejected	4	5
Not_rejected	2	2609

Contrasts_S_S		
Column1	**POSITIVE**	**NEGATIVE**
Rejected	16	8
Not_rejected	8	22

ANOVA_S_S_eval_metrics		
Column1	**Metric**	**Value**
1	Accuracy	0.997
2	Precision	0.444
3	Sensitivity_Power	0.667
4	Specificity	0.998
5	FDR	0.556

Contrasts_S_S_eval_metrics		
Column1	**Metric**	**Value**
1	Accuracy	0.704
2	Precision	0.667
3	Sensitivity_Power	0.667
4	Specificity	0.733
5	FDR	0.333