Universidad de Jaén

Escuela de Doctorado

TESIS DOCTORAL

•

APPLIED QUANTITATIVE PROTEOMICS ANALYSIS

PRESENTADA POR:
DAVID OVELLEIRO

DIRIGIDA POR:
MARÍA ÁNGELES PEINADO HERREROS
SANTOS BLANCO RUIZ

JAÉN 30 de OCTUBRE 2019

APPLIED QUANTITATIVE PROTEOMICS ANALYSIS

Chapter 1. Introduction

1.1 Aim of this work

This work carries out an analysis of quantitative proteomics data, using three different data experiments and proteomics approaches:

- In Chapter 3, the early hypoxic response in the cerebral cortex of rats, submitted to two different hypoxic models is evaluated using an isobaric labeling quantification technique (Tandem mass tags or TMT).
- In Chapter 4, proteomic biomarkers of polycystic ovary syndrome (PCOS) in plasma are evaluated using a data independent acquisition proteomics approach (SWATH).
- In Chapter 5, a public data set (iprg2015) is reanalyzed using the label-free proteomics approach by means of different software pipelines, with the objective of setting up an optimized strategy for label-free quantification and also presenting the strengths and limitations of this particular technique.

Additionally, in Chapter 2, the different software and hardware elements used in this work are described, composing a fully functional bioinformatics platform for proteomics analysis. Also, three appendixes have been included at the end of this work, containing methods, pipelines and code used in Chapter 4 (Appendix 1 and 2) and Chapter 5 (Appendix 3).

The analysis of the aforementioned data sets have allowed an exhaustive overview of the present state of the art of quantitative proteomics, both in terms of bioinformatics analysis and biological interpretation of the results obtained. As it will be shown in the next sections, the three techniques chosen here represent the most popular and recent approaches to unravel the complexity of protein functions in living organisms, what has been recently described as “Next-generation proteomics” (1).

1.2 Present situation of proteomics

Proteomics can be defined as “a comprehensive, quantitative description of protein expression and its changes under the influence of biological perturbations such as disease or drug treatment” (2). With a direct relationship with the development of high performance liquid chromatography (HPLC) and mass spectrometry (MS) (3), proteomics has experienced an intense development in recent years (4). In a given sample (or several samples to campare), the expression levels of several thousand of proteins can now be assessed just in a few hours using the approach called “shotgun proteomics” (5).

In contrast to “top-down” proteomics (6), where complete proteins are analyzed, “bottom-up” proteomics approach is based in the previous digestion of complex protein mixtures using enzymes, the most commonly used being trypsin (7). The sample is then transformed from an initially complex mixture of thousands of proteins into an even more complex mixture of peptides. Using the “bottom-up” approach, there are also two possible alternatives: “shotgun” and “targeted” proteomics, where the former tries to identify all peptides present in the sample, while the latter only focuses on certain peptides: the ones mapping to a sub-set of previously chosen proteins. The “shotgun” approach is therefore used in studies where the aim is to analyze the greatest possible number of proteins, while

the “targeted” technique studies preselected lists of proteins of interest (ranging from several dozens to a few hundreds (8)).

At the present moment, “top-down” proteomics is becoming more popular (9), but still faces several important limitations (10): both HPLC and mass spectrometry devices present limitations in terms of accuracy for a proper analysis of complex mixtures of whole proteins (without enzymatic cleavage).

On the other hand, “targeted” proteomics is used for the study of a limited number of proteins contained in a given sample. Using a mass spectrometry technique known as “mass-reaction monitoring” (MRM), the highest levels of accuracy and reproducibility for protein quantification are produced among proteomics techniques (11,12), being also capable of dealing with a high number of samples. All these characteristics make of “targeted” proteomics an excellent approach to the accurate quantification of a limited set of proteins, but not for discovery stages, where a complete overview of the protein levels is desired.

Finally, “shotgun” proteomics (also known as “discovery” proteomics), where thousands of proteins can be identified (and quantified) at once from a given sample, is by far the most popular approach nowadays. Once the sequence of the vast amount of peptides generated after enzymatic cleavage (usually hundreds of thousands) is identified using mass spectrometry and protein databases, the information needs to be integrated back into proteins. This is achieved using different bioinformatics algorithms (13). An overview of the different steps involving a typical “shotgun” experiment is shown in Figure 1.1.

The three methodologies presented in this work (Chapters 3 to 5) correspond to “shotgun” proteomics experiments; thus, the content of this thesis will deal exclusively with this approach.

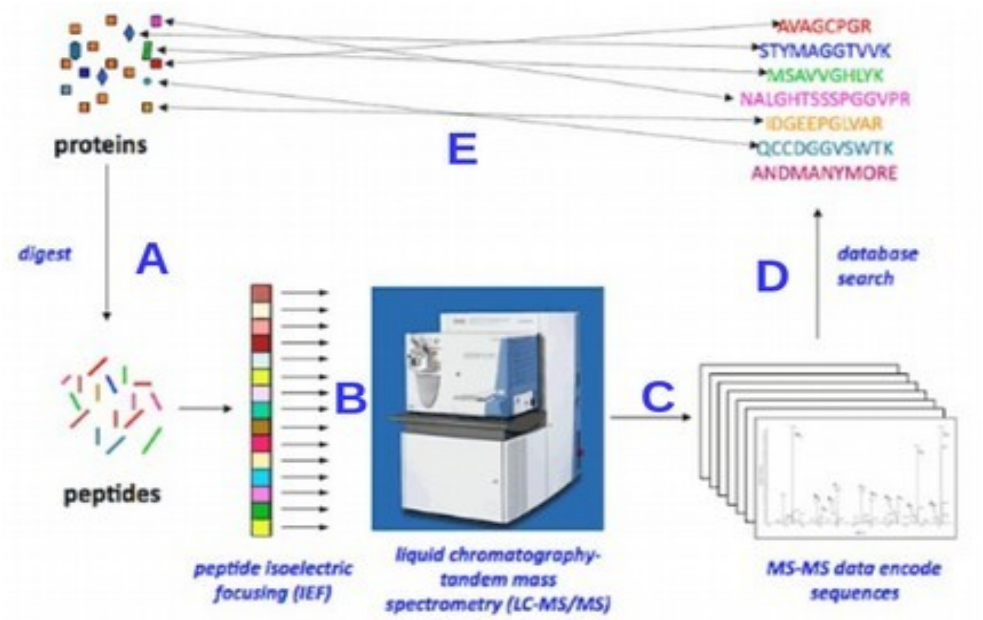


Figure 1.1 Overview of a typical proteomics work-flow, used in “shotgun” proteomics. Proteins are digested (A) using some enzyme (trypsin usually) and processed in a HPLC-MS system (B and C). After a MS-MS analysis, the sequence of the peptides analyzed is identified using protein databases and bioinformatics algorithms (D). The sequenced peptides are then integrated to reflect the sequences of the original proteins, using once more, different bioinformatics algorithms (E).

1.3 Quantitative proteomics strategies

Mass spectrometry based proteomics, at the beginning of the 2000's, provided general information about the proteins contained in certain organisms or tissues; hence, the so called "proteomes" consisted essentially in large listings (from several hundreds up to a few thousands) of proteins (14–16) contained in one sample. Studies providing quantitative information were scarce and used rudimentary approaches based on peptides counts (17), being at this time when several quantification methods started to be developed (18): the analysis of raw peptide signal intensities and also the use of stable (non-radioactive) isotopes labels. Both methods allowed the development of techniques that are widely used these days.

Several classifications and terminologies have been proposed for quantitative proteomics (19–21). In this work, the terms and classification established by Schuber et al. (22) in their 2017 review have been followed, with a few modifications (Figure 1.2).

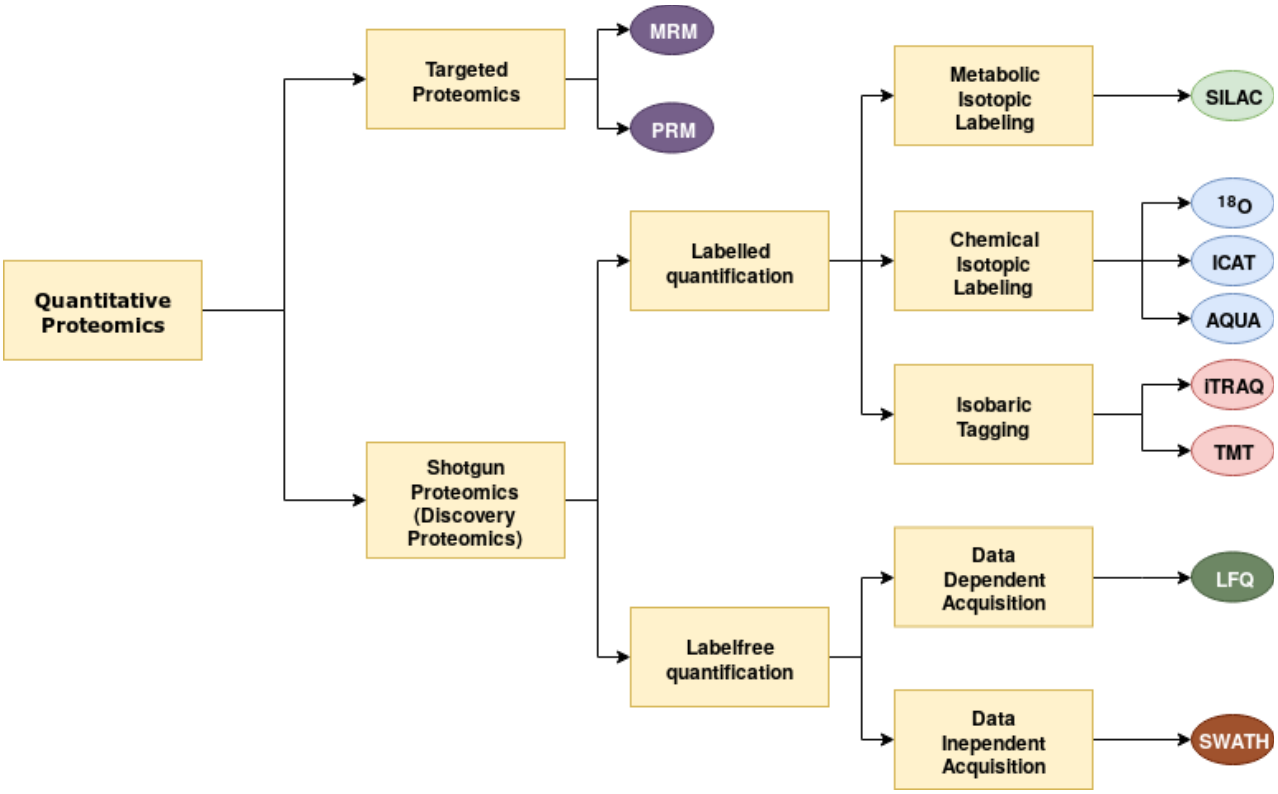


Figure 1.2 Quantitative proteomics approaches (in squares) with several examples (in circles) of the more significant techniques in each case. Targeted proteomics and shotgun (or untargeted) proteomics represent the main subdivision. The untargeted approaches, also known as shotgun proteomics or discovery proteomics, comprise labelled and label-free techniques. Terminology has been adapted from Schuber et al. (22), using throughout this work the term "shotgun" proteomics instead of "discovery" proteomics.

The majority of the quantitative techniques exposed here produce a relative quantification: the amount of a given protein is displayed as a ratio or fold change between two samples or sets of samples. Several strategies allow the adaptation of some of the techniques shown to absolute quantification (23), using peptides labeled with stable isotopes, but the use of relative quantification is still the predominant approach.

Following the summarization used in Figure 1.2, the different quantification techniques in proteomics are:

I. **Targeted proteomics**: a subset of the proteins contained in a particular sample are chosen for analysis before the acquisition. Up to a few hundreds of proteins can be analyzed. The two main techniques are Mass Reaction Monitoring (MRM) (24) and Parallel Reaction Monitoring (PRM) (25), the former being the traditional approach, where a MS^1-MS^2 transition is selected for quantification purposes, while the later is a modern re-implementation, based on high-resolution and high-precision mass spectrometry, where not only one transition is recorded, but all product ions are analyzed. That makes PRM capable of both quantification and identification of peptides at acquisition time.

II. **Shotgun proteomics** (Discovery proteomics): peptides are analyzed without previous knowledge of the proteins present in the sample.

   A) Labelled quantification: a synthetic reagent or "label" is introduced in the sample to produce labelled peptides or proteins.

      1) Metabolic isotopic labeling: cells are grown in culture media supplemented with light and heavy versions of an amino acid, until all amino acids have been replaced by the heavy versions of the amino acid. The most popular method is know as SILAC ("Stable Isotope Labeling by Amino Acids in Cell Culture") (26,27).

      2) Chemical isotopic labeling: chemical reagents are used to derivatize peptides or proteins in one sample. To this category belong, for instance ^{18}O (Proteolytic 18O-labeling (28)), AQUA (Absolute quantification of proteins (29)) and ICAT (Isotope-coded affinity tags (30)).

      3) Isobaric tagging: multiplexed tags with the same total weight ("isobaric") at MS^1 level but with different fragmentation patterns at MS^2 are introduced in the samples. Commercial reagents like TMT (Tandem Mass Tags (31)) and iTRAQ (Isobaric tags for relative and absolute quantitation (32)) are the most used.

   B) Labelfree (label-free) quantification: the quantification is performed without the use of labels added to the samples. The intensity of the precursor ions (peptides at MS^1), the combination of precursor and fragment ions (transitions) intensities or simply peptide counting (spectral counting methods like emPAI (17)) are used for protein quantification.

      1) Data dependent acquisition (DDA) label-free quantification: the precursor ions (peptides) intensities are used as a direct measure of the concentration of proteins in a sample (33). Although not the only DDA label-free method (spectral counting methods are also in this category), the tag "label-free" is commonly used by the proteomics community to refer to this particular approach.

      2) Data independent acquisition (DIA) label-free quantification: the most popular approach nowadays is SWATH quantification (34), where a combination of a DDA library and transitions acquired in intervals of m/z (mass to charge) windows are used for accurate and reproducible protein quantification.

In this work, we have used data obtained using three of the quantitative techniques described above:

- Chapter 3, isobaric tagging using TMT
- Chapter 4, data independent acquisition (DIA) label-free quantification (SWAT)
- Chapter 5, data dependent acquisition (DDA) label-free quantification

As shown in Figure 1.3, the two most popular approaches presently are isobaric tagging and label-free, whereas the emerging SWATH analysis (sometimes referred as Next Generation Proteomics (35)) is gaining ground. These three technologies will be covered in detail in the next chapters.

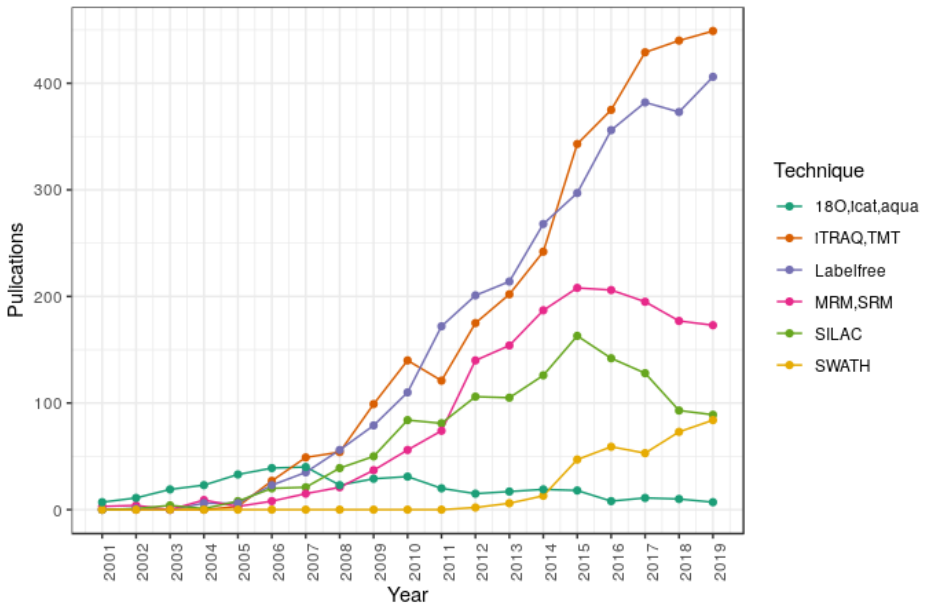


Figure 1.3 Scientific publications related to several terms in quantitative proteomics found in Pubmed between 2001 and 2019 (this last year up to September). Chemical isotopic labelling was searched using the terms 18O, Icat and Aqua. Isobaric labelling was searched by TMT and iTRAQ terms. Labelfree was searched using all the combinations of the term ("label-free", "labelfree" and "label free"), similarly to MRM and SRM for the targeted approaches. Metabolic isotopic labeling was represented only by its most popular approach, SILAC. Finally, SWATH closes the list of the techniques compared.

1.4 References

1. Altelaar AFM, Munoz J, Heck AJR. Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet. 2013 Jan;14(1):35–48.

2. Anderson NL, Anderson NG. Proteome and proteomics: New technologies, new concepts, and new words. Electrophoresis. 1998 Aug;19(11):1853–61.

3. Chen C-H (Winston). Review of a current role of mass spectrometry for proteome research. Anal Chim Acta. 2008 Aug;624(1):16–36.

4. Yates JR, Ruse CI, Nakorchevsky A. Proteomics by mass spectrometry: approaches, advances, and applications. Annu Rev Biomed Eng. 2009 Sep;11:49–79.

5. Wolters DA, Washburn MP, Yates JR. An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. Anal Chem. 2001 Dec;73(23):5683–90.

6. Kelleher NL. Top-down proteomics. Anal Chem. 2004 Jun;76(11):197A–203A.

7. Olsen JV, Ong S-E, Mann M. Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues. Mol Cell Proteomics. 2004 Jun;3(6):608–14.

8. Blackburn K, Goshe MB. Challenges and strategies for targeted phosphorylation site identification and quantification using mass spectrometry analysis. Brief Funct Genomic Proteomic. 2009 Sep;8(2):90–103.

9. Chen B, Brown KA, Lin Z, Ge Y. Top-Down Proteomics: Ready for Prime Time? Anal Chem. 2018 02;90(1):110–27.

10. Catherman AD, Skinner OS, Kelleher NL. Top Down proteomics: Facts and perspectives. Vol. 445. 2014. 683–693 p.

11. Ebhardt HA. Selected reaction monitoring mass spectrometry: a methodology overview. Methods Mol Biol Clifton NJ. 2014;1072:209–22.

12. Elschenbroich S, Kislinger T. Targeted proteomics by selected reaction monitoring mass spectrometry: applications to systems biology and biomarker discovery. Mol BioSyst. 2011;7(2):292–303.

13. Marcotte EM. How do shotgun proteomics algorithms identify proteins? Nat Biotechnol. 2007 Jul;25(7):755–7.

14. Pasini EM, Kirkegaard M, Mortensen P, Lutz HU, Thomas AW, Mann M. In-depth analysis of the membrane and cytosolic proteome of red blood cells. Blood. 2006 Aug 1;108(3):791–801.

15. Mann K. The chicken egg white proteome. Proteomics. 2007 Oct;7(19):3558–68.

16. Shi R, Kumar C, Zougman A, Zhang Y, Podtelejnikov A, Cox J, et al. Analysis of the mouse liver proteome using advanced mass spectrometry. J Proteome Res. 2007 Aug;6(8):2963–72.

17. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, et al. Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. Mol Cell Proteomics. 2005 Sep;4(9):1265–72.

18. Ong S-E, Mann M. Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol. 2005 Oct;1(5):252–262.

19. Vidova V, Spacil Z. A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition. Anal Chim Acta. 2017 Apr 29;964:7–23.

20. Li H, Han J, Pan J, Liu T, Parker CE, Borchers CH. Current trends in quantitative proteomics - an update. J Mass Spectrom JMS. 2017;52(5):319–41.

21. Holman SW, Sims PFG, Eyers CE. The use of selected reaction monitoring in quantitative proteomics. Bioanalysis. 2012 Jul;4(14):1763–86.

22. Schubert OT, Röst HL, Collins BC, Rosenberger G, Aebersold R. Quantitative proteomics: challenges and opportunities in basic and applied research. Nat Protoc. 2017 Jul;12(7):1289–94.

23. Ankney JA, Muneer A, Chen X. Relative and Absolute Quantitation in Mass Spectrometry-Based Proteomics. Annu Rev Anal Chem Palo Alto Calif. 2018 12;11(1):49–77.

24. Yocum AK, Chinnaiyan AM. Current affairs in quantitative targeted proteomics: multiple reaction monitoring-mass spectrometry. Brief Funct Genomic Proteomic. 2009 Mar 1;8(2):145–57.

25. Guerin M, Gonçalves A, Toiron Y, Baudelet E, Pophillat M, Granjeaud S, et al. Development of parallel reaction monitoring (PRM)-based quantitative proteomics applied to HER2-Positive breast cancer. Oncotarget. 2018 Sep 18;9(73):33762–77.

26. Ong S-E, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics MCP. 2002;1(5):376–386.

27. Mann M. Functional and quantitative proteomics using SILAC. Nat Rev Mol Cell Biol. 2006;7(12):952–8.

28. Miyagi M, Rao KCS. Proteolytic18O-labeling strategies for quantitative proteomics. Mass Spectrom Rev. 2007 Jan;26(1):121–36.

29. Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc Natl Acad Sci U S A. 2003 Jun 10;100(12):6940–5.

30. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol. 1999 Oct;17(10):994–9.

31. Thompson A, Schäfer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem. 2003;75(8):1895–1904.

32. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, et al. Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. Mol Cell Proteomics MCP. 2004 Dec;3(12):1154–1169.

33. Wong JWH, Cagney G. An overview of label-free quantitation methods in proteomics by mass spectrometry. Methods Mol Biol Clifton NJ. 2010;604:273–83.

34. Ludwig C, Gillet L, Rosenberger G, Amon S, Collins BC, Aebersold R. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Mol Syst Biol [Internet]. 2018 Aug 13 [cited 2019 Sep 2];14(8). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6088389/

35. Lin Q, Tan HT, Chung MCM. Next Generation Proteomics for Clinical Biomarker Detection Using SWATH-MS. Methods Mol Biol Clifton NJ. 2019;1977:3–15.

Chapter 2. Setting up a Proteomics Analysis Platform

2.1 Abstract

The complete infrastructure needed by a proteomics platform is described in this chapter. Particularly, the tools needed for the production of this thesis are: operative systems and virtual machines, scheduling software, protein identification and quantification tools, statistical and tertiary analysis software and a documentation system. In all cases, free of charge alternatives have been used.

2.2 Introduction

The analysis of millions of mass spectra generated in a quantitative proteomics experiment usually requires an advanced computational infrastructure, both in terms of hardware and software. In many cases, budget and resources limitations play an important role in the ability to perform advanced studies with the available data. In proteomics facilities, the purchase of very expensive equipment is not always backed by a good enough data analysis infrastructure (computers and software), despite the fact that an adequate bioinformatics study is a limiting factor in the quality of the results obtained.

Here, a platform capable of dealing with the analysis of proteomics data is described. A single computer will be used and all software installed will be available at no charge: in some cases "freeware" (with commercial license but freely available) and in most cases, non-commercial and open source (1).

In this chapter, not all the possible proteomics techniques will be covered using the infrastructure described, but many of the most popular proteomics techniques in use these days can be analyzed with the different tools discussed. Moreover, all the software described here will be used, in one way or another, in the different chapters of this thesis.

2.3 General infrastructure

First of all, the computer used has an Intel i7-8700 CPU (3,20 GHz, 6 cores working in 12 threads), with 32 Gb of RAM. The disk storage available sums up to 2 Tb.

The operative system of choice is Ubuntu 18.04 LTS. One of the most popular Linux (2) distributions is Ubuntu, a Debian based distribution that has become sort of a standard in bioinformatics. Two tools distributed with Ubuntu, will be of great importance in this analysis platform:

- R software (3), for statistical analysis and plots generation, alongside with RStudio and Bioconductor (4) and all the necessary modules.
- Mono (5), an open source implementation of Microsoft's .NET Framework, allows running on Linux some Windows native software.

2.3.1 Job scheduling system: Slurm

In scientific analysis, is quite common that long and intensive informatics jobs are running in the same system, being that system a cluster of many computational nodes or a single computer. In proteomics, a single experiment takes several hours to be analyzed even in the most powerful computers. We have used "job (or task) scheduling systems", in order to share the resources of the system in an appropriate manner and queuing jobs waiting for available resources. Typically found in computing clusters, job scheduling systems are also very useful in single computers, offering the capability of assigning certain amount of processors and memory to a given analysis, queuing the jobs that cannot get enough resources. There are many job scheduling systems (Moab-Torque, PBS, LSF,..) but Slurm (6) has become very popular lately, being in use in 60% of the top 500 super-computers in the world (7). Slurm is a free and open-source job scheduler for Linux and Unix-like kernels, and it is easy to install and use . It is a command line application that is usually launched in the form of scripts; those scripts consist in two parts: a header with specifications about number of tasks, processors, memory, paths,.. and a body with the instructions to execute when the job has been allocated resources (Figure 2.1).

A

```
JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
  189 workparti iPRG.201    david PD       0:00      1 (Resources)
  191 workparti PXD00456    david PD       0:00      1 (Priority)
  192 workparti PXD00456    david PD       0:00      1 (Priority)
  183 workparti PXD00181    david  R   10:54:17      1 david-2019
```

B

```
#!/bin/bash
#SBATCH --job-name      = PXD001819         # Job name
#SBATCH --ntasks        = 1                 # Run a single task
#SBATCH --cpus-per-task = 10                # Number of CPU cores (threads) per task
#SBATCH --mem           = 24gb              # Job memory request
#SBATCH --time          = 20:00:00          # Time limit hrs:min:sec
#SBATCH --error         = PXD001819.err     # Error log
#SBATCH --output        = PXD001819.out     # Output log
#SBATCH --workdir       = /mnt/data/doc.chapter4/search.engine04.maxquant/slurm.fold

mono ~/software/10.MaxQuant/bin/MaxQuantCmd.exe \
    /mnt/data/doc.chapter4/search.engine04.maxquant/PXD001819.mqpar.xml
```

Figure 2.1 Slurm schedules jobs and assigns resources to them. (**A**) Command line Slurm output, where four jobs have been scheduled: one of them is running (ST state as "R") and the other three are waiting for resources (ST state as “PD”, with the reason being "Resources" or "Priority"). The working job has been running for almost 11 hours. (**B**) A Slurm script, consisting of a header (in red, with the first character being a "#") with job parameters and a body (in black), that here corresponds to the execution of the MaxQuant software through the Mono framework.

2.3.2 Virtualization: VirtualBox and Docker

Virtualization systems have become increasingly popular: the allow installing an encapsulated operative system (guest) into the main operative system (host), with software installed and working in the guest operative system. The two main reason for using such virtualization systems are:

- Some software applications are very complicated to install and configure, so one option to distribute them is to install the software into a guest operating system and distribute an image of this operating system. In this way, it is possible having a

complete reproducible work-flow: the software will work in the exact same conditions in all the instances of the same virtual machine.

- Many applications are nor compatible with some operating systems. On Linux is not possible running some Windows applications, despite systems like Mono. Using virtualization, it is possible to run a Windows application inside a Windows virtual machine working in a Linux host.

Several virtualization systems are available for free, being the most popular Oracle VirtualBox (8), a proprietary application that can be used for free.

One disadvantage of the virtualization systems is that the host and guest systems have to share the computational resources. To overcome this, a more flexible and efficient system has been developed: the containers. Containers work in a similar way to virtual machines, encapsulating an operative system and its applications, but are smaller and consume less resources. With containers, though, there is no real virtualization: a Windows container can not work on a Linux host. Namely, many Windows applications can be adapted, with a lot of work in some cases, to work in a Linux system: once the work is done, they can be encapsulated in a Linux container and distributed to Linux systems.

The most popular system of containers is Docker (9). It is widely used in all areas of bioinformatics (10), and it is going to be the preferred option here over the use of virtual machines. "Dockerized" applications are shared in a public repository (Dockerhub) in the form of "images", downloaded by Docker and run in form of containers. Docker allows linking directories inside the container with directories in the local machine, allowing in this way directly work on locally located files and folders.

Docker has been used in three different ways throughout this thesis (Figure 2.2):

1) By means of a local console; this way allows a direct access of a program inside a container from the command line: Proteowizard has been used in this way.
2) In interactive mode; this way is very convenient when several programs are to be inspected and used: several applications of the Trans-Proteomic Pipeline have been run inside an interactive console.
3) Running a web server and re-directing the output to a local port; this way allows the local use of complex web applications without installing them: the Trans-Proteomic Pipeline has also been used in this way.

Only in one case a virtual machine has been used in this platform: Skyline can not be adapted (at the moment of writing this document) to work on Linux, being the only option for running Skyline on a Linux system the use of a Windows host system in a virtual machine (inside VirtualBox in this case).

```
# Proteowizard is run from the local console
#   The file spetra.raw is converted to spectra.mzML in the local directory /mnt/data/
docker run -it --rm --memory="10g" -e WINEDEBUG=-all \
    -v /mnt/data/:/data proteowizard/pwiz-skyline-i-agree-to-the-vendor-licenses wine msconvert \
    /data/spetra.raw -o /data

# TPP is run in interactive mode
#   Working inside the container with local directory /mnt/data linked to /data inside the container
docker run --memory="20g" -it -v /mnt/data:/data spctools/tpp /bin/bash

# TPP is accessed in the local web browser
#   The web server is run inside the container, redirecting it locally to port 10401
#   Locally, is accessed using  http://david-2019:10401/tpp/cgi-bin/tpp_gui.pl
docker run --memory="10g" -dit --user=root -p 10401:10401 \
    -v /mnt/data:/data spctools/tpp apache2ctl -DFOREGROUND
```

Figure 2.2 The three ways in which Docker has been used in this work: local console, interactive mode and as web server..

2.4 Proteomics software

2.4.1 Vendor formats converter: Proteowizard

The raw data produced by mass spectrometers is delivered in the form of proprietary formats, most of the times binary and not accessible by third party software. Some of the tools in this platform accept those proprietary formats, but other cannot work with them. In those cases, an intermediary file format, particularly a community open format, is needed to analyze the data generated by the mass spectrometer. Such formats (mzML for raw data) are efficiently produced using an application named Proteowizard (11).

Proteowizard takes the files in the proprietary format and translates them into the mzML format, usable by most of the proteomics analysis pipelines. This software can also perform mass spectra peak integration, necessary for some software to quantify. Due to the fact that vendor software libraries work on Windows systems, Proteowizard needs to be adapted to work on Linux computers using Wine (12), a compatibility layer able to running Windows applications on Linux, BSD and macOS. Adapting Proteowizard for running on Linux systems is not straightforward, so the most convenient way for using this is downloading a Docker image and running locally the corresponding container.

2.4.2 Identification and quantification software

In the several chapters of this thesis, several software applications are going to be used for protein quantification: Maxquant (13), Trans-Proteomic Pipeline , Skyline (14), OpenMS (15), OpenSwath (16) and Proteome Discoverer (17).

The differential characteristics of these three software applications are the following:

- Maxquant is very well suited software for labeled and unlabeled proteomics quantification. It is aimed to Data Dependent Acquisition (DDA). Recently ported to Linux (18) using the Mono framework, can run using a graphical interface or in command line. This last feature makes Maxquant is especially suited for high performance computing (HPC) using Slurm as job scheduler.
- The Trans-Proteomic Pipeline is a mature set of applications that enables working with several search engines, integrating the results into a single search. This

platform has been used in this work for the generation of a spectral library for Swath quantification. It can be directly installed on Linux but is a long and complicated process. A lot more convenient is using its official Docker image, that points its web server to our host Linux system.

- Skyline is the software of reference for targeted proteomics (MRM), but it is also used for DDA and DIA. In this work, has been tested for Swath quantitation. Provides a powerful graphical interface to explore transitions associated to peptide assignments.
- OpenMS: is an open-source software C++ library (+ python bindings) for LC/MS data management and analyses. It has been extensively used in this work in Swath analysis and a one of the pipelines in data-dependent label-free quantification.
- OpenSwath: set of programs aimed to Swath analysis, recently integrated as part of the OpenMS project.
- Proteome Discoverer: integrated platform for identification and quantification of proteins. It is proprietary software and can only be run on Windows. Also, it is limited to data generated by Thermo Fisher Scientific® instruments. The open-source alternatives discussed here provide a complete alternative to its use.

2.4.3 Quantitative and statistical analysis

Quantitation software produces a list of proteins with intensity values associated. In order to organize the information, normalize signals and produce meaningful comparisons between the various phenotypes studied, several solutions have been developed. Some software platforms (like Proteome Discoverer) include this type of analysis as the final part of their pipelines. In the case of Maxquant, a companion application named Perseus (19), reads the output of Maxquant and generates this kind of analysis (and several other features like time-series analysis, cross-omics comparisons and multiple-hypothesis testing). Finally, several other applications have been developed in the Bioconductor project, allowing the import of different quantification pipelines (Proteome Discoverer, Maxquant, Openswath, OpenMS, Skyline among them). Some examples of these Bioconductor packages are MSStats (20), DEqMS (21) and DEP (22). The three of them have been used in this work, testing their performance with data-dependent label-free quantification.

2.5 Tertiary analysis

Tertiary analysis, a term popularized in genomics (23,24), refers to the procedures that allow biological interpretation of the results obtained with proteomics techniques. In this way, the proteins that have been found differentially expressed within samples under study, may be associated with specific biological processes. For elucidating such relationships, several approaches have been developed: gene set enrichment (25), pathway mapping (26), cluster analysis (27), literature annotations (28) and ontologies (29) are among them.

Many of these applications can be used through web interfaces publicly available in internet: some examples are Toppgene (30), Kegg (31), AmiGO (32) or PubMed (33). Another important resource for tertiary analysis, is the Bioconductor (34) project. It does not only allow accessing hundred of proteins, genes and annotation databases, but to

directly use the information stored in those data bases and analyze it, using powerful statistical methods stored in libraries within the R project.

One important consideration can be made here: the use of publicly available data, accessed directly from internet, allows the remote access to huge data bases (some of them with Terabytes of data (35)) without the need of having the information locally available. Only in the case of intensive use of the information, the local install of those databases may be evaluated.

2.6 Reporting and documentation

The last step of any bioinformatics pipeline is reporting the results obtained in some adequate way. Traditionally, documents redacted using word processors (Word, LibreOffice Writer) with the description of the pipelines and spreadsheets (Excel, LibreOffice Calc) containing the results have been employed for this purpose.

More recently, a new paradigm has been introduced by the introduction of Jupyter Notebooks (36) for Python (and its R counterpart with R Notebooks). The system chosen to analyze a significant part of the work exposed in this thesis, using RStudio for the generation of R notebooks, is designed as a complete integrated development environment: the code and the documentation are integrated in the same working environment, as well as code execution and results production. When finished, these notebooks can be exported using different formats, including html for web visualization or pdf documents with high quality.

While long lists of results are still reported as spreadsheets documents for convenience, the rest of information produced in a bioinformatics pipeline can be reported into a single document: the three Appendixes included at the end of this thesis have been completely built using R notebooks.

Another advantage of using R Notebooks is the fact that all the code and algorithms used in some study are shown in a completely transparent way and, additionally, the pipelines generated can be totally reproduced if needed: full reproducibility in bioinformatics pipelines can be easily achieved in this way (37).

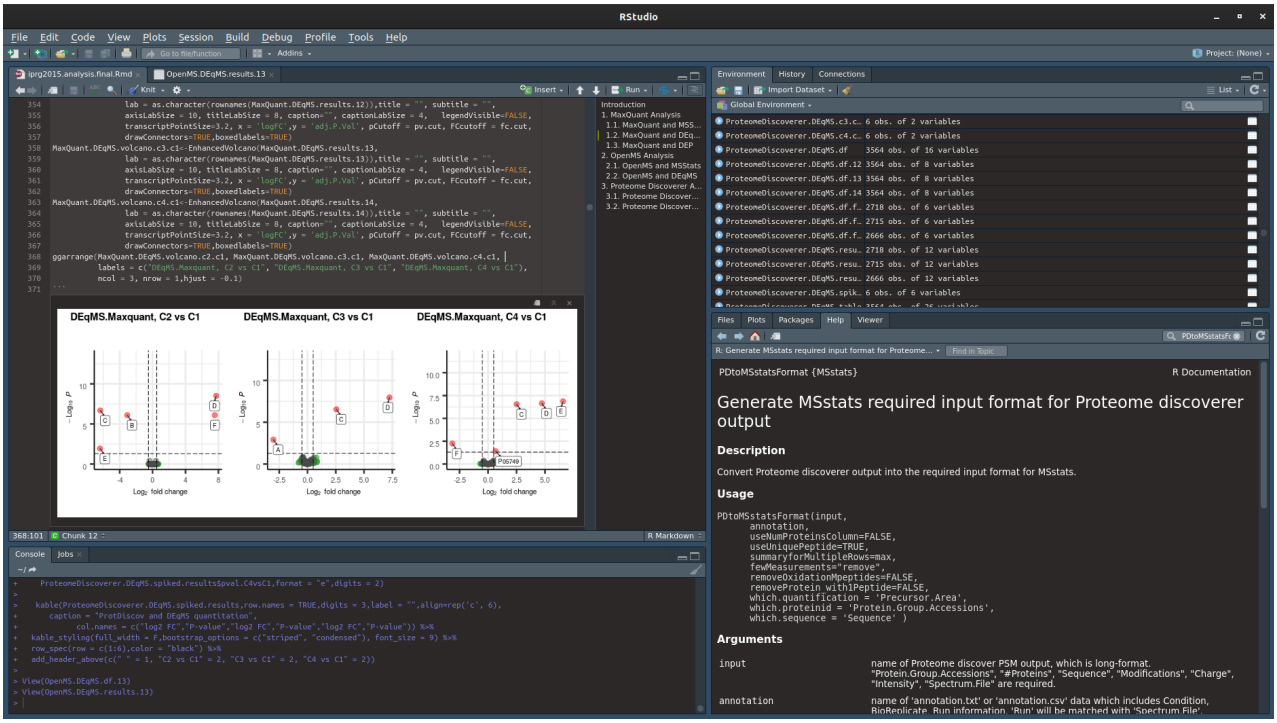


Figure 2.3 Screenshot of RStudio while elaborating an R Notebook used in this work.

2.7 Conclusions

When complex bioinformatics analyses are performed, a complete informatics infrastructure is needed, both in terms if software and hardware. The informatics infrastructure, although usually hidden into the materials and methods sections in the literature, is essential for the development of a correct bioinformatics work . This does not mean that exceptional investments should be done: the complete infrastructure described in this chapter amounts for less than 1,000 euros, hardware and software included. That is achieved thanks to the intensive employment of free of use software, thus reducing costs exponentially. An additional fact that should be taken into account is that this software is in many cases open source, with the advantages in terms of transparency and quality that this represents.

2.8 References

1. Perspectives on Free and Open Source Software [Internet]. [cited 2019 May 18]. Available from: https://dl.acm.org/citation.cfm?id=1051431

2. Möller S, Krabbenhöft HN, Tille A, Paleino D, Williams A, Wolstencroft K, et al. Community-driven computational biology with Debian Linux. BMC Bioinformatics. 2010 Dec 21;11 Suppl 12:S5.

3. R Core Team (2019). R: A language and environment for statistical computing. [Internet]. R Foundation for Statistical Computing, Vienna, Austria.; Available from: https://www.R-project.org/

4. Loraine AE, Blakley IC, Jagadeesan S, Harper J, Miller G, Firon N. Analysis and visualization of RNA-Seq expression data using RStudio, Bioconductor, and Integrated Genome Browser. Methods Mol Biol Clifton NJ. 2015;1284:481–501.

5. Mono project [Internet]. Available from: https://www.mono-project.com/

6. Slurm Workload Manager [Internet]. Available from: https://slurm.schedmd.com/

7. Top500 [Internet]. Available from: https://www.top500.org

8. Oracle VirtualBox [Internet]. Available from: https://www.virtualbox.org/

9. Docker [Internet]. Available from: https://www.docker.com/

10. Di Tommaso P, Palumbo E, Chatzou M, Prieto P, Heuer ML, Notredame C. The impact of Docker containers on the performance of genomic pipelines. PeerJ. 2015;3:e1273.

11. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics. 2008 Nov 1;24(21):2534–6.

12. Wine [Internet]. Available from: https://www.winehq.org/

13. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008 Dec;26(12):1367–72.

14. Keller A, Eng J, Zhang N, Li X, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol. 2005;1:2005.0017.

15. Sturm M, Bertsch A, Gröpl C, Hildebrandt A, Hussong R, Lange E, et al. OpenMS - an open-source software framework for mass spectrometry. BMC Bioinformatics. 2008 Mar 26;9:163.

16. Röst HL, Rosenberger G, Navarro P, Gillet L, Miladinović SM, Schubert OT, et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat Biotechnol. 2014 Mar;32(3):219–23.

17. Colaert N, Barsnes H, Vaudel M, Helsens K, Timmerman E, Sickmann A, et al. Thermo-msf-parser: an open source Java library to parse and visualize Thermo Proteome Discoverer msf files. J Proteome Res. 2011 Aug 5;10(8):3840–3.

18. Sinitcyn P, Tiwary S, Rudolph J, Gutenbrunner P, Wichmann C, Yılmaz Ş, et al. MaxQuant goes Linux. Nat Methods. 2018;15(6):401.

19. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016;13(9):731–40.

20. Choi M, Chang C-Y, Clough T, Broudy D, Killeen T, MacLean B, et al. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinforma Oxf Engl. 2014 Sep 1;30(17):2524–6.

21. Zhu Y. DEqMS: a tool to perform statistical analysis of differential protein expression for quantitative proteomics data. 2019.

22. Zhang X, Smits AH, van Tilburg GB, Ovaa H, Huber W, Vermeulen M. Proteome-wide identification of ubiquitin interactions using UbIA-MS. Nat Protoc. 2018;13(3):530–50.

23. Oliver GR, Hart SN, Klee EW. Bioinformatics for clinical next generation sequencing. Clin Chem. 2015 Jan;61(1):124–35.

24. Moorthie S, Hall A, Wright CF. Informatics and clinical genome sequencing: opening the black box. Genet Med Off J Am Coll Med Genet. 2013 Mar;15(3):165–71.

25. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005 Oct 25;102(43):15545–50.

26. Liu L, Wei J, Ruan J. Pathway Enrichment Analysis with Networks. Genes. 2017 Sep 28;8(10).

27. Chong FS, O'Sullivan MG, Kerry JP, Moloney AP, Methven L, Gordon AW, et al. Understanding consumer liking of beef using hierarchical cluster analysis and external preference mapping. J Sci Food Agric. 2019 Sep 12;

28. Sernadela P, Oliveira JL. A semantic-based workflow for biomedical literature annotation. Database J Biol Databases Curation [Internet]. 2017 Nov 15 [cited 2019 Sep 19];2017. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5691355/

29. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019 Jan 8;47(D1):D330–8.

30. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W305-311.

31. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017 Jan 4;45(Database issue):D353–61.

32. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009 Jan 15;25(2):288–9.

33. PubMed [Internet]. Available from: https://www.ncbi.nlm.nih.gov/pubmed

34. Gatto L, Christoforou A. Using R and Bioconductor for proteomics data analysis. Biochim Biophys Acta BBA - Proteins Proteomics. 2014 Jan;1844(1):42–51.

35. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–15.

36. Mendez KM, Pritchard L, Reinke SN, Broadhurst DI. Toward collaborative open data science in metabolomics using Jupyter Notebooks and cloud computing. Metabolomics Off J Metabolomic Soc. 2019 Sep 14;15(10):125.

37. Kim Y-M, Poline J-B, Dumas G. Experimenting with reproducibility: a case study of robustness in bioinformatics. GigaScience. 2018 01;7(7).

Chapter 3. Isobaric labeling quantification: early hypoxic response in the cerebral cortex

In this chapter, the main steps followed in the analysis of a TMT labeled proteomics experiment are detailed using the data and methods in our study “Comparative proteomic study of early hypoxic response in the cerebral cortex of rats submitted to two different hypoxic models”, issued in the Proteomics - Clinical Applications publication. Although a complete description is going to be given in the Materials and Methods section, the main focus of this chapter is the Bioinformatics analysis and the tools and algorithms used to unravel the biological meaning of the mechanisms at work in the cerebral hypoxia evaluated in the brains of the specimens used in this study.

3.1 Abstract

In “Comparative proteomic study of early hypoxic response in the cerebral cortex of rats submitted to two different hypoxic models”, we analyzed and compared the cortical brain proteomic profiles of two different severity models of cerebral hypoxia in rats (HH: hypobaric hypoxia, HHI: ischemia followed by hypobaric hypoxia) with respect to a control group, in an attempt to describe the alterations of the early molecular hypoxic adaptive response underling each one. The main technology used was Mass Spectrometry and TMT (Tandem mass tags), chemical labels that allow the relative quantification of proteins in complex biological samples.

Altogether, 339 proteins were confidently quantified, 99 of them showing significant variations in the hypoxic conditions with respect to the control. The HHI model presents a global effect of protein down-regulation while HH produces an overall increase of the protein levels. While HH mainly affecting oxidative and energetic metabolism, HHI also interferes with synaptic transmission, neurotransmitter secretion, substantia nigra development and triggers apoptosis through mitochondrial pathway.

3.2 Introduction

Decline or complete deprivation of oxygen flow to brain and posterior re-oxygenation represent a global health issue, as occur after an episode of hypobaric hypoxia or in the cerebral ischemic diseases (1). Given that the decrease or lack of oxygen characterizes all these illnesses, they share several molecular hallmarks: oxidative and nitrosative stresses (2), excitotoxicity (3) or apoptotic and necrotic neuronal death (4). Nevertheless, the available data point out to specific patterns of these molecular responses depending on the multi-factorial aetiology, duration and severity of the hypoxic insult(5). Certainly, these variables define and modulate the type of hypoxic adaptive response as well as the hypoxic damage, although the specific molecular pattern underlying each ones is still scarcely known. In the present work, we propose a quantitative analysis and comparison using isobaric labeling (TMT) of the proteomic profiles of two cerebral hypoxic models of different severity and scope, both simulating brain hypoxic pathologies: high altitude and cerebral ischemic disease.

3.3 Materials and Methods

3.3.1 Sample preparation from animal specimens

Our study has been performed on 15 adult male Wistar rats provided by Harlan Laboratories (Envigo) and weighing 350 g each, kept under standard conditions of light and temperature and allowed *ad libitum* access to food and water. All procedures were performed in accordance with the EU Directive 2010/63/EU (2010), reviewed by the Ethics Committee of the Spanish Council for Scientific Research, approved by the Committee of Bioethics of the University of Jaén (Spain), and comply with the Uniform Requirements for manuscripts submitted to biomedical journals.

Animals were distributed into three different groups (n=5 per group), one for each hypoxic model (HH: hypobaric hypoxia; HHI: ischemia followed by hypobaric hypoxia) and the control group (sham animals under normobaric normoxic conditions).

3.3.2 Hypoxic models

We developed two experimental models of oxygen deprivation with different degree of severity:

1) A gentle model of hypobaric hypoxia (HH) using a slight modification of a previously published procedure by down-regulating the environmental O2 pressure to a final barometric pressure of approximately 300 hPa inside a hypobaric chamber. The rats were placed in the hypobaric chamber in which the air pressure was controlled by means of a continuous vacuum pump and an adjustable inflow valve. The chamber was also provided with a manometer to check the experimental altitude during the process. The conditions, simulating an altitude of 9,144 m (30,000 feet), were maintained for 1 h. Ascent and descent rates were kept below 300 m/min (approximately 1,000 feet/min). After the hypoxic period, the return to normobaric normoxic conditions spanned 30 min.

2) A more severe model of cerebral ischemia followed by hypobaric hypoxia (HHI), which consists of unilateral left common carotid artery occlusion followed by a hypoxic stress for a predetermined time. This model has been successfully applied both to neonatal (6,7) , and adult animals (8,9) and consists on a slight modification of the Levine/Vannucci model. Animals recovered for 2 h after surgery, were submitted to hypobaric hypoxia as previously described. More specifically, rats were anesthetized with ketamine (100 mg/Kg body weight, i.p.) and xylazine (5 mg/Kg body weight, i.p.). Then, we proceeded to the isolation, ligation, and sectioning of left common carotid artery. Animals recovered for 2 h after surgery, and were submitted to hypobaric hypoxia as previously described.

Body temperature was monitored and maintained throughout all the procedures. In both HH and HHI animals were killed intermediately after the hypobaric chamber was opened. Sham animals (controls) were submitted to surgery without vessel sectioning and then kept in the chamber under normobaric normoxic conditions.

3.3.3 Protein extraction

After HH or HHI the left-brain cortices from animals of all experimental groups including controls were extracted and processed according to the following procedure: 0.1 g of the cortices were homogenized with 1.5 mL of extraction buffer pH 8.0 containing 8 M urea, 20 mM dithiothreitol (DTT), 100 mM Tris–HCl, 0.75 mM phenylmethylsulfonyl fluoride (PMSF), and 4% 3-[(3- cholamidopropyl)-dimethylammonio]-1-propane sulfonate (CHAPS). For each

experimental group, there were three replicates of the homogenates, each replicate being made up with a pool of the left-brain cortex from five rats. Proteins were extracted in this buffer for 60 min on ice (the samples were moderately shaken in a vortex every 15 min) and afterwards were centrifuged at 10,000 ×g for 15 min at 4 °C. The protein concentration of the supernatants was measured using the CB-XTM Protein Assay (G-Biosciences, St Louis, USA).

Lessening of detergents from protein extraction buffer was carried out using 100mM triethylammonium bicarbonate (TEAB) by ultrafiltration (millipore 3k) during 30 min at 12500rpm and precipitation (BioRad Protein Sample Cleanup). Isobaric Label Reagent Set (Thermo TMTsixplex™) was performed following the manufacturer's instructions, and followed by desalting (100 mg C18 cartridges, Schalau).

3.3.4 LC/MS/MS analysis

Peptides were scanned and fragmented with the LTQ Orbitrap mass spectrometer (Thermo Fisher Scientific) equipped with a nano UHPLC Ultimate 3000 (Dionex-Thermo Scientifics). Chromatography conditions were: Mobile phase solution A: 0.1% formic acid in ultrapure water; Mobile phase solution B: 80% acetonitrile, 0.1% formic acid. Chromatography gradient was performance in C18 nanocapillary column (Acclaim PepMap C18, 75 um internal diameter, 1.8 um particle size, Dionex-Thermo Scientifics) as follow: 5 min, 4% solution B; 240 min, 4-35% solution B; 10 min, 35-80% B; 10 min, 80% B; 10 min 4% B. The nanoelectrospray voltage was set to 1300 V and the capillary voltage to 50 V at 190 Cº.

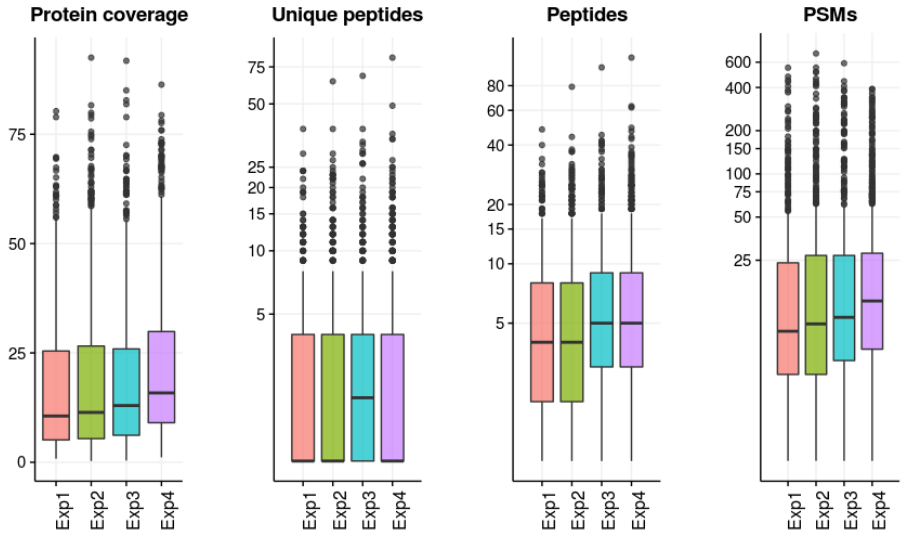
The LTQ Orbitrap was operated in the parallel mode, allowing for the accurate measurement of the precursor survey scan (400–1500 m/z) in the Orbitrap selection, a 30 000 full-width at half-maximum (FWHM) resolution at m/z 400 concurrent with the acquisition of three CID/HCD Data-Dependent MS/MS scans in the LIT and C-Trap for peptide sequence and isotopes quantitation (100–2000 m/z), respectively. HCD Resolution set to at 7500 FWHM at m/z 400. Singly charged ions were excluded. The normalized collision energies used were 40% for HCD and 35% for CID. The maximum injection times for MS and MS/MS were set to 50 ms and 500 ms, respectively. The precursor isolation width was 3 amu and the exclusion mass width was set to 5 ppm. Monoisotopic precursor selection was allowed and singly charged species were excluded. The minimum intensity threshold for MS/MS was 500 counts for the linear ion trap and 1000 counts for the Orbitrap. The Minimum Information About a Proteomics Experiment (MIAPE) (10) for Mass Spectrometry, summarizing all relevant information in this paragraph is shown at Table 3.1.

MIAPE-MS Mass Spectrometry v2.98	
1. General features —1.1 Global descriptors	
Responsible person or role	Santos Blanco (University of Jaen, sblanco@ujaen.es) and Maria Angeles Peinado (University of Jaen, apeinado@ujaen.es)
Instrument manufacturer, model	LTQ Orbitrap mass spectrometer (Thermo Fisher Scientific)
Customisations	HCD cell
2. Ion sources —2.1 Electrospray Ionisation (ESI)	
Supply type	Sprayer fed by ultra performance liquid chromatography
Interface manufacturer, model	Nano UHPLC Ultimate 3000 (Dionex-Thermo Scientifics)
Sprayer type	NSI-2 dynamic
3. Post source component	
3.1 Analyser	Linear trap quadrupole. MS1 survey scans in an Orbitrap and MS2 analysed in a linear trap. Dual conversion dynode detector.
3.2 Activation / dissociation	Acquisition of three CID/HCD Data-Dependent MS/MS scans in the LIT and C-Trap for peptide sequence and isotopes quantification (100–2000 m/z), respectively.
4. Spectrum and peak list generation and annotation	
4.1 Data acquisition	
Software	Xcalibur (Thermo Fischer Scientific) version 2.5.5 SP1
Parameters	Parallel mode: precursor survey scan (400–1500 m/z, 30 000 full-width at half-maximum, FWHM), concurrent with the acquisition of three CID/HCD Data-Dependent MS/MS scans in the LIT and C-Trap.HCD Resolution set to at 7500 FWHM at m/z 400. The normalized collision energies used were 40eV for HCD and 35eV for CID.
4.2 Data analysis	
Software	Conversion from RAW to mzML using ProteoWizard version 3.0.9576 (ProteoWizard Software Foundation). The RAW files where directly loaded and spectra processed in Proteome Discoverer version 1.4.0.288.
Parameters used in the generation of processed spectra	MS1 spectra used as precursor. Precursor masses where selected between 350-5000 Da. Filters: minimum peak count 1, maximum collision energy 100eV, S/N threshold 1.5.
4.3 Resulting data	
Location of source and processed files	The 19 RAW files corresponding to this study are stored in the ProteomeXchange database (PXD004091)
m/z and intensity values	The m/z and intensity values can be accessed at the 19 mzML files stored in the ProteomeXchange database (PXD004091).
MS level	MS2 for CID and HCD
Ion mode	Positive
Precursor m/z and charge	The precursor m/z and intensity values can be accessed at the 19 mzML files stored in the ProteomeXchange database (PXD004091).

Table 3.1 The Minimum Information About a Proteomics Experiment (MIAPE) Mass Spectrometry v2.98 for the analysis performed to the samples in this work.

3.3.5 Protein identification

The protein identification is performed using the software Proteome Discoverer following the *Shotgun proteomics* approach (11). By this methodology, the mass spectra generated by the mass spectrometer are sequenced into a peptide sequence. These sequences, in the order of several thousands, are integrated, using a protein database (12) into one or more proteins. An important limitation to proteomics is that, in many cases, a given peptide can be assigned to more than one protein. This phenomenon, known as the "protein inference problem" (13) is addressed by Proteome Discoverer software (and other software, e.g. Mascot (14)) generating "groups" of proteins with a representative protein reported. Four chromatographic runs have been used to identify and quantify the proteins in this work; a summary of the main data descriptors for each sample is given in Figure 3.1.



	n	Coverage (%)	Unique peptides	Peptides	PSMs
Experiment 1	461	17.6 10.6	3.3 1.0	6.2 4.0	31.2 8.0
Experiment 2	663	18.9 11.4	3.6 1.0	6.4 4.0	33.2 9.0
Experiment 3	630	19.4 13.0	3.7 2.0	7.3 5.0	33.0 10.0
Experiment 4	877	22.0 15.9	3.5 1.0	7.5 5.0	29.7 13.0

Figure 3.1 The four experiments are described here showing the Protein coverage (percent of the residues in each protein sequence that have been identified), Unique peptides (peptides that map exclusively to a given protein), Number of Peptides and PSMs (Peptide-Spectrum Matches, that is, number of spectra that are responsible for the identification of a given protein). The box-plots show the log-scaled distributions of the aforementioned varia-bles in each of the four experiments. The table summarizes the num-ber of proteins (n) and the previous values: their means (black) and standard deviation (light blue).

The main steps involved in the protein identification procedure used in this work are:

1. Loading Raw files (produced by Xcalibur, the acquisition software attached to the mass spectrometer) into Protein Discoverer.

2. Spectra are sequenced (i.e. the signals are translated into sequences of amino acids) using the Sequest (15) search algorithm. This is a database guided sequencing process: tryptic peptides (16) are compared to the spectra provided and proteins are then re-constructed in a final report.
3. The protein database used here is a combined target-decoy database: attached to the original fasta (17) format, an artificial database, consisting in the reversed sequences of the original database is used to asses confidence in the peptide sequencing (18,19). Decoy peptides are non-natural, artificial sequences: when a decoy peptide is identified by the search engine, we can be sure (20) that the corresponding hit is a false positive.
4. Proteins are identified (and quantified) into four separated search processes (one for each original sample). The information is initially stored in a format called Magellan storage file (MSF), a SQLite (21) database file that stores identification and quantification information.
5. The main parameters used in the search describe the precision used in the MS^1 and MS^2 spectra (Precursor Mass Tolerance: 10 ppm, Fragment Mass Tolerance: 0.6 Da), the amino acids that can experiment modifications or dynamic modifications (N-Terminal and K by TMT reagent +229.163 Da, M Oxidation: +15.995 Da), fixed amino acid modifications (Carbamidomethylation of C +57.021 Da) and the peptide lengths and charges that are going to be reported (between 6 and 144 amino acids and +1 to +4).
6. A strict filter is performed over the identified peptides (0.01), using the False discovery rate (22) (FDR) values obtained with the help of the target-decoy database.
7. Results are provided in two different formats: peptide and protein lists, both offering identification and quantification information. Although some information has been extracted from the peptides listings (e.g. Figure 3.2 has been obtained using peptide quantitative information), the main source of data has been the protein identification and quantification reports exported from Protein Discoverer software. The information has been exported as an Excel workbook and converted to tab-delimited files for further manipulation.

A detailed report of the steps, software and parameters used in the protein identification used in this work is offered in Table 3.2. following the MIAPE guidelines for Mass Spectrometry Informatics (23).

MIAPE-MSI Mass Spectrometry Informatics v1.1

1. General features – (a) Global descriptors	
Date stamp	2015-01-01
Responsible person or role	Santos Blanco (University of Jaen, sblanco@ujaen.es) and Maria Angeles Peinado (University of Jaen, apeinado@ujaen.es)
Software name, version and manufacturer	The software used for the identification and quantitation of proteins is Proteome Discoverer (Thermo Fisher Scientific) version 1.4.0.288.
Customisations	The msf files exported from Proteome Discoverer (Thermo Fisher Scientific) where exported to Excel (Microsoft) reports in the four different analysis performed. These four analysis where integrated into one unique report using Perl scripting (integrating the identification and quantification information coming from the four parallel analysis).
Availability of the software	Proteome Discoverer (Thermo Fisher Scientific) can be obtained at www.thermofisher.com. Perl is an open source software that can be obtained at www.perl.org
Location of the files generated	The four msf files (Santos_TMTGroup1.msf , Santos_TMTGroup2.msf , Santos_TMTGroup3.msf and Santos_TMTGroup4.msf , corresponding to the four runs of the mass spectrometry analysis and corresponding bioinformatics analysis) are stored in the ProteomeXchange database (PXD004091).
2. Input data and parameters – (a) Input data	
Description and type of MS data	The data submitted to the Protein Discoverer package were the original RAW files generated by the mass spectrometer.
Availability of MS data	The MS data can be accessed at the 19 mzML files stored in the ProteomeXchange database (PXD004091).
2. Input data and parameters – (b) Input parameters	
Database queried	Uniprot Proteome of Rattus norvegicus (Rat) database, containing 27,820 sequences. Version 2015.01.
Taxonomical restrictions	*Rattus norvegicus* (Rat), Uniprot Organism ID 10116
Description of tool and scoring scheme	Peak lists generated by Proteome Discoverer are analyzed using the vendors proposed method: CID scans are using for peptide identification and HCD scans used for quantitation of the reporter ions generated by fragmentation of the TMT-6plex reagent. Afterwards, the Percolator component of the Proteome Discoverer package has been applied with Maximum Delta Cn: 0.05 and Target FDR (Relaxed): 0.05 and Validation based on: q-Value.
Specified cleavage agent	Trypsin (cleaves after K or R, but not before P) with full cleavage.
Allowed number of missed cleavages	2 missed cleavages allowed.
Permissible amino acids modifications	One static modification set: Carbamidomethyl (+57.021 Da at C), and two dynamic modifications allowed: TMT6plex (+229.163 Da at K or any N-term), Oxidation (+15.995 Da at M), with a maximum of three dynamic modifications per peptide.
Precursor-ion and fragment-ion mass tolerance for tandem MS	Precursor Mass Tolerance: 10 ppm and Fragment Mass Tolerance: 0.6 Da. Both cases using exact mass.
Thresholds; minimum scores for peptides, proteins	Percolator parameters: Max. Delta Cn: 0.05 and Max. Number of Peptides Reported: 10. Proteins with at least 2 high confidence Peptide-Spectrum Matching (PSMs) and one unique peptide allowed.
Any other relevant parameters	Spectrum matching ions: y and b, and neutral losses of a,b,y ions. Min. Peptide Length: 6 Max. Peptide Length: 144. The file “Protocol.Identification.txt “ stored ProteomeXchange database (PXD004091) was exported from the Proteome Discoverer software analysis, summarizing all the steps performed during the identification analysis.
3. The output from the procedure – (a) For identified proteins	
Accession code in the queried database	For each identified protein, the Uniprot accession is provided in Supporting Information 1: Protein and peptide identifications
Protein description	For each identified protein, the Uniprot description is provided in Supporting Information 1: Protein and peptide identifications
Protein scores	For each identified protein, the Protein Discoverer protein score is provided in Supporting Information 1: Protein and peptide identifications, proteins tabs, column T: “Score A(3,6)”
Validation status	The proteins identified by the search engine were accepted without any post-processing.

Number of different peptide sequences (without considering modifications) assigned to the protein.	The number of distinct peptide sequences for each identified protein is provided in Supporting Information 1: Protein and peptide identifications, proteins tabs, column V: "# Peptides A(3,6)".
Percent peptide coverage of protein	The Percent peptide coverage of protein is provided in Supporting Information 1: Protein and peptide identifications, proteins tabs, column C: "Σcoverage".
3. The output from the procedure – (b) For identified peptides	
Sequence	The sequence of each identified peptide is provided in Supporting Information 1: Protein and peptide identifications, peptides tabs, column A: "Sequence".
Peptide scores	The peptide scores of each identified peptide is provided as a q-value and a peptide probability score in Supporting Information 1: Protein and peptide identifications, peptides tabs, columns T and U: "q-Value" and "PEP".
Chemical and post-translational modifications	The chemical modifications of the peptide sequences are provided for each peptide in Supporting Information 1: Protein and peptide identifications, peptides tabs, column F: "Modifications".
Corresponding Spectrum locus	The corresponding spectrum locus for each psm can be obtained from the four msf files stored in the ProteomeXchange database (PXD004091).
Charge assumed for identification and a measurement of peptide mass error	The charge and peptide mass error can both be obtained from the four msf files stored in the ProteomeXchange database (PXD004091).
3. The output from the procedure – (c) Quantitation for selected ions	
Quantitation approach	Quantification of TMT 6plex reporter ions. The peak integration for the quantification used the most confident centroid peak integration approach, with a tolerance of 20 ppm.
Quantity measurement	The quantity measurement, performed by Proteome Discoverer software, is done following the Most Confident Centroid algorithm: Lays a Gaussian curve around the target peak (the tag mass) with a sigma value equal to the mass accuracy or integration window. Then the Gaussian curve normalizes all peaks in the window, and the largest is considered to be the most confident peak.
Data transformation and normalization technique	The quantitative data was normalized by Proteome Discoverer using the protein median ratio.
Number of replicates	Different technical replicates (depending on sample availability) were used, as described in Supporting Information 2: Materials and methods.
Acceptance criteria	A protein quantitation value is only accepted when at least two peptides have been quantified. Also, the coefficient of variation must be less than 30% for each protein. The coefficient of variation is used by Proteome Discoverer as a measure of protein ratio variability, and is calculated as a coefficient-of-variation for log-normal distributed data.
Estimates of uncertainty and the methods for the error analysis	After having used the Coefficient of Variation to limit the error introduced when multiple peptides are used for the quantitation, the use (when possible) of information coming from technical replicates, has been used to modulate (and in a few cases, to eliminate) the quantitative information associated to each protein.
Results from controls	No quantitative controls were used in this study.
4. Interpretation and validation	
Assessment and confidence given to the identification and quantitation	The objective in this study was obtaining quantitative results with the higher possible accuracy. In order to obtain this, in addition to the 30% coefficient of variation threshold in the protein quantitation, the variability of the proteins with respect to the control has been measured following a global standard deviation approach, common to all samples under study. We have used two different thresholds: two standard deviations (equivalent to the 95% of the distribution) and 1.5 (roughly the 87%), and have differeciated clearly these two confidence intervals when interpreting data.
Results of statistical analysis or determination of false positive rate in case of large scale experiments	At the Percolator component level in the Proteome Discoverer application, a Target FDR (Strict): 0.01 and a Target FDR (Relaxed): 0.05 are applied. The subsequent validation is based on the q-Value.
Inclusion/exclusion of the output of the software are provided	The protein and peptide reports were provided as produced by Proteome Discoverer in Supporting Information 1: Protein and peptide identifications. The files " Protocol.quantification1.txt", 1 to 4, stored ProteomeXchange database (PXD004091) were exported from the Proteome Discoverer software analysis, summarizing all the steps performed during the quantitation analysis.

Table 3.2 The Minimum Information About a Proteomics Experiment (MIAPE) Mass Spectrometry

Informatics v2.98 for the analysis performed to the samples in this work.

3.3.6 Protein quantification: TMT for relative protein quantification

The quantification of the proteins present in the samples is performed using Tandem Mass Tags (24), also known as TMT™, which is one of the reagents (alongside with iTRAQ™) known collectively as isobaric mass tags or reporter ion tags. For this analysis we have used the TMT-6plex, that allows labeling of proteins in six different samples. The basic principle is providing a chemical label that can differentiate the origin (one of the six samples labeled) of a protein in a mixture and, quantify the amount of this protein in the original samples. The reagent plays with several isotopes (of C and N) and different distributions in order to obtain a reactive that, when linked to the peptide weights the same in all six peptides (each coming from a different samples) but when the MS^2 fragmentation takes place, a fragment of different weight remains united to the peptide. These MS^2 fragments, known as reporter ions, weight 126, 127, 128, 129, 130 and 131 Daltons. Then, measuring the intensity of each fragment in the common MS^2 spectra, it is possible to infer the relative amount of the original peptides (and by extension, of their corresponding proteins) into the six original samples (Figure 3.2).

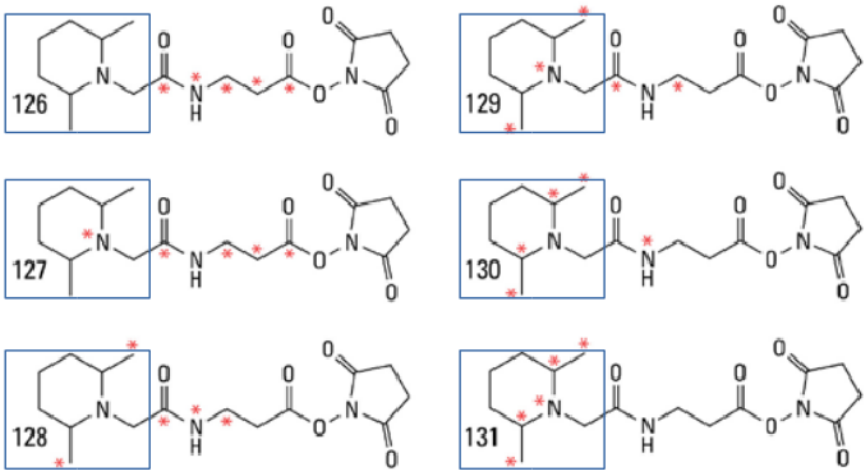


Figure 3.2. Design of six TMT-6plex reagents taken from Product Sheet: TMT Mass Tagging Kits and Reagents. The blue squares delimit the reporter ions and red dots identify the isotopes used in each reagent. The mono-isotopic modification mass, that will be used as a fixed modification by the search engine, is common for all reagents to the sixth decimal: 229.162932 Da. (Figure adapted from Thermo Fisher Scientific™ catalog)

In this work, the relative amount of proteins found in both hypoxic models (HH and HHI) with respect of the sham controls is calculated. In order to do such comparison, the reagents used (the corresponding reporter ion mass) must be identified to perform the comparisons. It is important to note that the “raw” quantitative information obtained here is going to be related to peptides, not proteins. The software for the quantification (Proteome Discoverer) will integrate those peptide ratios (thousands of them) into a list of protein ratios (hundreds).

The isobaric mass tags method has several, well known limitations (25–29):

1. In first place, not all peptides identifying one protein will have a quantitative tag. That means that only peptides identified with high confidence (good enough score) and with a quantitative tag will be used in the quantification. This will drastically

reduce the amount of peptides available for quantification, notably in proteins that were initially identified with a limited number of peptides. The number of identified peptides for a set of randomly picked proteins in the set obtained in this work is shown at Figure 3.3.

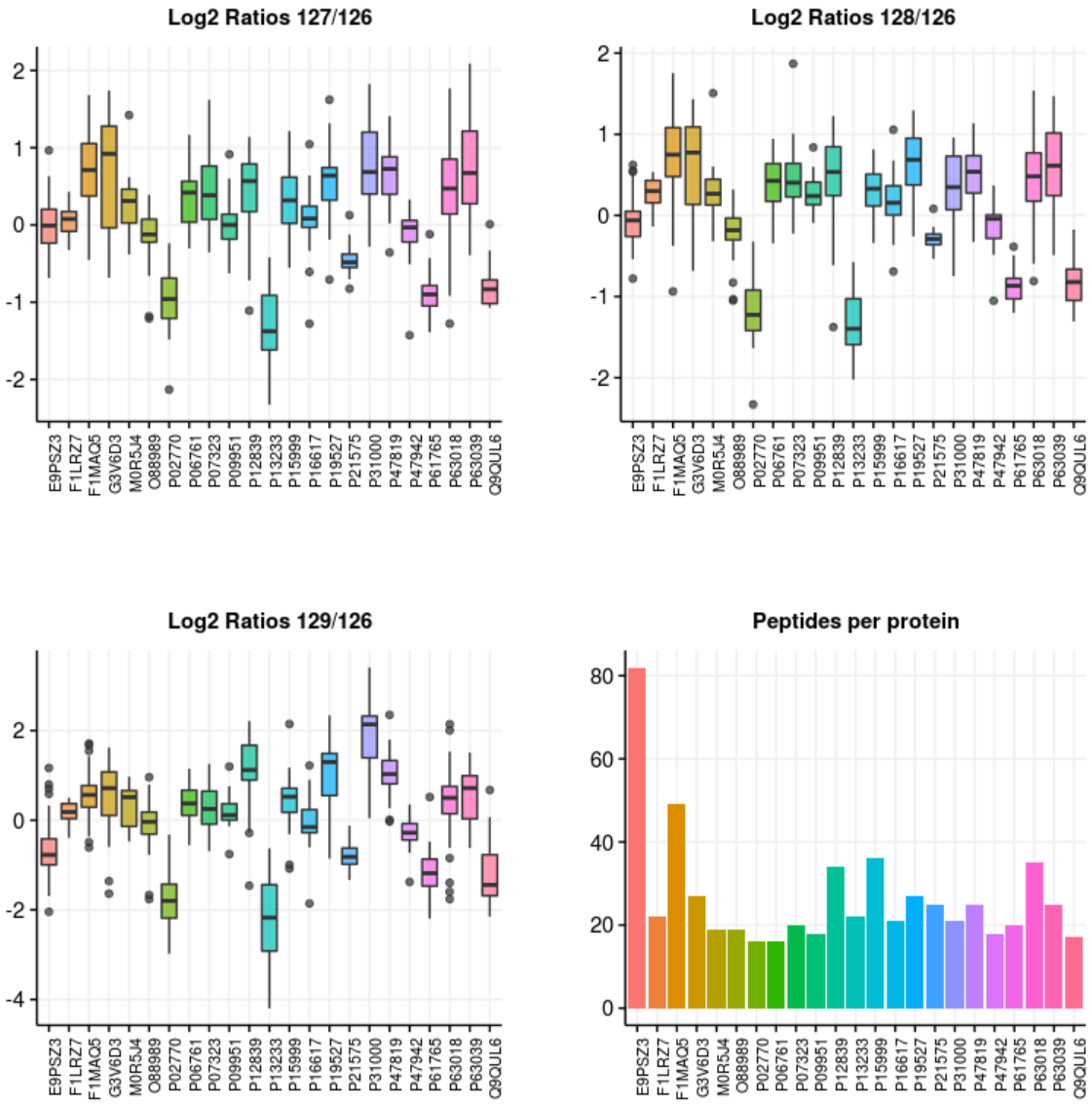


Figure 3.3. Distribution of the Log2 ratios of intensities (three plots with Log2 ratios) and number of peptides (Peptides per protein) from 23 proteins randomly chosen from the set that will be analyzed in this work. In this experiment, the reagent 126 was used for labelling the control.The box-plots show how, typically in this type of analysis, the precision is not as high as would be desirable.

2. The protein inference problem cited before (13) also affects the protein quantification using this approach. If some peptide is found in several proteins, its quantitative information will be a mixture of the original amount of proteins where it proceeds. The ideal approach against this effect would be the exclusive use of

"unique peptides" (those included only in one protein) for quantification. The problem with this approach is that it would reduce, more drastically than the previous point, the population of peptides for quantification and therefore, is not applied.

3. Precision (how close are quantifications of the same protein) and accuracy (how close are measurements of the true amount of that protein) are sometimes quite low. The aforementioned protein inference problem, the inherent instrumental accuracy, false positive peptides, and the protein dynamic range (30) are some of the factors lowering precision and accuracy. With the experimental approach used in this study, the accuracy can not be estimated, and the median value for the replicated measurements of the same protein is adopted as the "true value". Precision is typically estimated using the coefficient of variation (standard deviation relative to the mean) of the ratios obtained with respect to the control (31).

Although the two first points can not be addressed using our experimental approach, the precision can be at least controlled: proteins quantified with a coefficient of variation lower than 30% will be considered as confidently quantified, and the rest will be discarded. This is a conservative approach, limiting greatly the number of proteins that are going to be quantified, but at the same time, increasing robustness and confidence in the results.

3.3.7 Samples and replicates

In the present study, three different samples belonging to each experimental condition are analyzed:

- Control: Pool obtained from samples from five sham rats used as controls in the quantification analysis. The ratios are in all cases calculated with respect to these samples.
- HH: Pool of samples from five rats submitted to hypobaric hypoxia.
- HHI: Pool of samples from five rats submitted to ischemic and hypobaric hypoxia.

These samples, analyzed using mass spectrometry in four labeling groups, produced different files depending on the number of chromatographic runs: several chromatographic replicates were performed when sample availability allowed. Into each of these mass spectrometry analyses, different combinations of samples were performed using the distinct labels of the TMT reagent. The number of technical replicates varies among samples depending, as well, on the sample availability. The next table (Table 3.2) summarizes the composition of the four different groups in which the data is organized, with r1,r2,... meaning different technical replicates. Samples labeled as "not avaliable" (NA) represent samples from a previous experimental design discarded for poor analytical value and lack of proper controls.

	126	127	128	129	130	131
Group 1	**control**	HHI (r1)	HHI (r5)	NA	NA	HH (r3)
Group 2	HH (r2)	NA	NA	HHI (r2)	**control**	HH (r1)
Group 3	NA	NA	NA	**control**	HHI (r3)	NA
Group 4	NA	NA	**control**	HHI (r4)	NA	NA

Table 3.2: TMT labeling of the nine samples used in this study: control (sham individuals), HH and HHI. Inside parenthesis, the number of the technical replicate (r1, r2,...). Samples with gray background were rejected.

The quantitative analysis was performed using the Proteome Discoverer software and the protein reports were exported as four different text files. Inside each of the four groups, the quantification ratios were calculated using the “control” group as reference.

A Perl script has been used to:

- **parse**: data exported from Proteome Discoverer as Protein reports, was converted to tab-delimited text files and parsed by the script
- **classify**: each data coordinate was mapped to the corresponding sample name (e.g. Group2, label 126 was mapped to sample “HH (r2))
- **filter**: only proteins quantified with two or more peptides are used. In addition, quantifications with a coefficient of variation higher than 30% are discarded.
- **integrate** the information, exporting the data to a tab delimited text file.

As a measure of quality for each sample quantification, the global standard deviation (gsd) was used, calculating the standard deviation of the protein ratios for each TMT label. The HHI condition has five technical replicates, and in order to fairly compare the HH and HHI conditions, only the two with best (i.e. lower) gsd - HHI (R2) and HHI (R5) - will be used in the analysis. The gsd values obtained for each sample are shown in Table 3.3.

	sample	gsd
HH	HH (R1)	0.245
	HH (R2)	0.204
HHI	~~HHI (R1)~~	0.217
	HHI (R2)	0.178
	~~HHI (R3)~~	0.406
	~~HHI (R4)~~	0.447
	HHI (R5)	0.119

Table 3.3 Values of global standard deviation (gsd) for each sample. Samples HHI (R1), HHI (R3) and HHI (R4) will not be further used in the analysis. Smaller gsd values mean narrower distributions and therefore, a better chance to correctly distinguish the differentially expressed proteins from the unchanged population.

The filtering process applied to proteins data includes:

- initial removal of proteins with less than 2 high confidence peptides,
- removal of quantification tags with a coefficient of variation higher than 30%,
- removal of proteins without data from both of HH and HHI experimental conditions

From an initial set of 1,409 identified-quantified proteins by Protein Discoverer across the two experimental conditions, 1,069 proteins where discarded following the previous criteria, leaving 340 proteins that will be further analyzed.

3.3.8 Protein quantification: threshold for differential expression

In order to define a subset of differentially expressed proteins for each condition under study, a fold-change threshold based on global standard deviation is used. Then, for each of the samples under study, a different global standard deviation is obtained and therefore, a different threshold will be used to delimit the proteins that are differentially expressed in a given condition with respect to the proteins that remain unchanged (or with

non-significant changes).

Given that the distributions generated by quantified proteins usually resemble a normal distribution, with a mean equal (or close) to 1, is a common approach (32–34) to assume that at a distance further than two standard deviations of the mean (considering a two-tailed distribution), only the 5% of proteins with the highest difference will be considered. We have used two different thresholds: two standard deviations (equivalent to the 95% of the distribution) and 1.5 (roughly the 87%). The proteins with ratios bigger to the two gsd threshold will then show an important change in their expression, while the proteins between 1.5 and two gsd units will exhibit a moderate change. Four categories are used:

- “proteins highly under-expressed”, with ratios under 1-(2*gsd)
- “proteins moderately under-expressed”, with ratios between 1-(2*gsd) and 1-(1.5*gsd)
- “proteins moderately over-expressed”, with ratios between 1+(1.5*gsd) and 1+(2*gsd)
- “proteins highly over-expressed”, with ratios higher than 1+(2*gsd)

Proteins with a fold change between -1.5*gsd and 1.5*gsd values will have an unchanged state. The thresholds for each sample are shown at Table 3.4.

	group	TMT label	sample	gsd	proteins highly under-expressed	proteins moderately under-expressed	proteins moderately over-expressed	proteins highly over-expressed
HH	G2	131	HH (R1)	0.245	[0 , 0.51]	[0.51 , 0.632]	[1.368 , 1.49]	[1.49 , +inf]
	G2	126	HH (R2)	0.204	[0 , 0.591]	[0.591 , 0.693]	[1.307 , 1.409]	[1.409 , +inf]
HHI	G2	129	HHI (R2)	0.178	[0 , 0.643]	[0.643 , 0.732]	[1.268 , 1.357]	[1.357 , +inf]
	G1	128	HHI (R5)	0.119	[0 , 0.762]	[0.762 , 0.821]	[1.179 , 1.238]	[1.238 , +inf]

Table 3.4: Values of global standard deviation (gsd) and thresholds used for each sample. Four categories of proteins are used: “highly under-expressed”, “moderately under-expressed”, “moderately over-expressed” and “highly over-expressed”. In some samples, for big gsd values, the threshold will fall under 0 for the under-expressed proteins; in these cases, a NA value is displayed, and no proteins will be taken for that category.

For obtaining the thresholds in the quantification of the different samples, we have followed a similar strategy to the one introduced by Pappin in US HUPO TechTalk, 2010 (32) :*“A simple approach is to use the global statistics derived from any given iTRAQ channel. For example, one can take all ratio measurements in any given iTRAQ channel for all peptides of p>0.05, and calculate mean and standard deviations. If normalized, the mean will be at or close to unity, and standard deviations can be in the range of 0.15-0.5. As a quick filter, one can judge the significance of fold-changes very broadly using this measured value, and fold changes can be ascribed simply in terms of intervals of SD above and below unity (say >2SD up or down)”* .

We have obtained the Quantile-Quantile plots of the samples under analysis, displayed at Figure 3.4, where log2 ratios of the experimentally obtained results are confronted to the values of a normal distribution in the corresponding range . As expected, the central area of the seven distributions resembles quite well a straight line, while the more extreme values, diverge. Those divergent values are precisely the ones that will provide the valuable data that represent the proteins that are under or over-expressed into each experimental condition. Several times throughout this chapter, the log2 of the ratios will be used instead of the ratios themselves. This is done for convenience only when a graphical representation is better displayed centered at 0. All the cutoffs and results will be expressed in terms of ratios, not their logarithms.

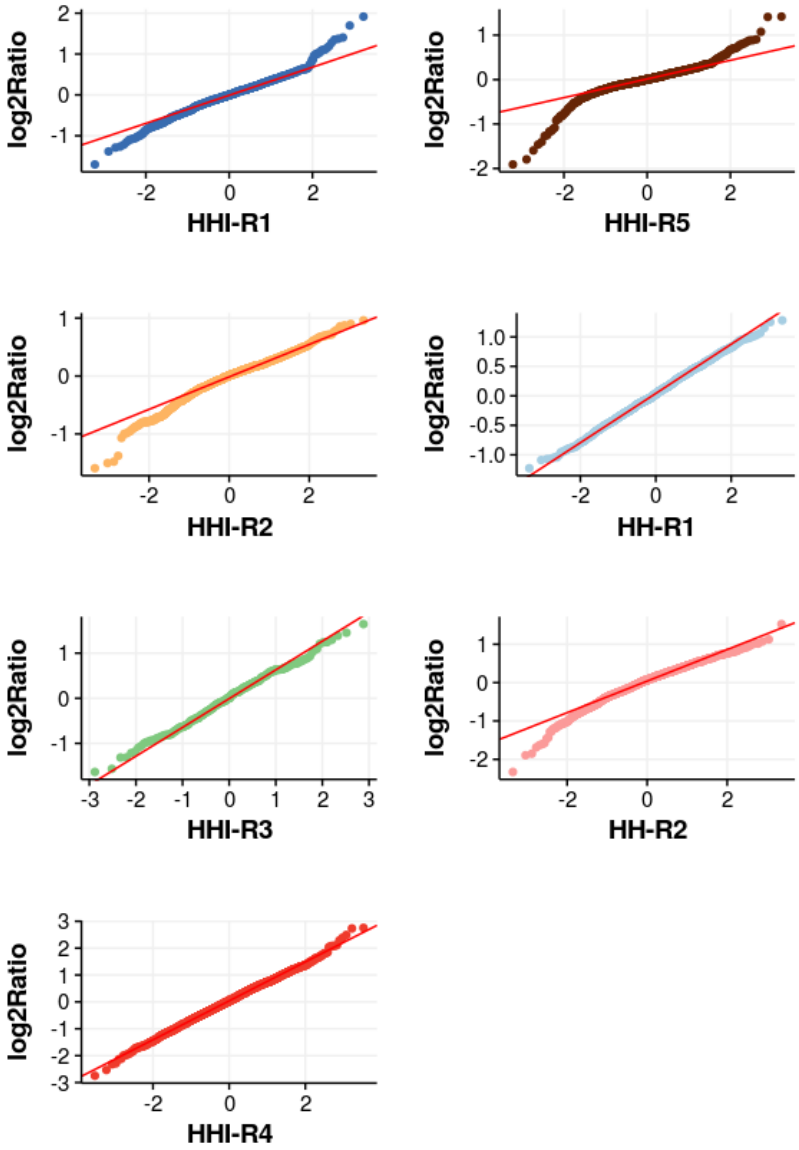


Figure 3.4: Quantile-Quantile plots of the seven technical replicates (two and four respectively) of the HH and HHI samples. The x-axis corresponds to the experimental ratios and the y-axis to the theoretical normal distribution. These plots provide an asses-sment of the goodness of fit to a normal distribution of the experimental values obtained. The thin red line plotted inside each of the graphs crosses the points where the lower and upper quartiles (Q1 and Q3) are located in both distribu-tions. All seven Q-Q plots appear to be fitted quite well to the red line at their central range: thus we can accept that these distri-butions follow normality, at least in their central area. Extreme values, not fitting the line, represent protein ratios that, in addition to the random error in the quantification, represent a significant increase or decrease: those can be interpreted as the proteins that are significantly under or over-expressed in a given sample (HH or HHI) with respect to the sham controls. The exact point to start considering that a protein is actually differentially expressed in each distribution is addressed in Figure 3.5.

In Figure 3.5, it is clear than most of the distributions present a profile close to the typical “bell-shaped” normal distribution, with their maximum close to 1. This is consistent with the Q-Q plots discussed before. Only HHI replicates with the lowest gsd (HHI.R2 and HHI.R5) are used in this analysis. Assuming the cost of the loss of some precious quantitative data from three HHI technical replicates, this approach will ensure a greater level of confidence in the quantification finally reported, as will be demonstrated in the next section.

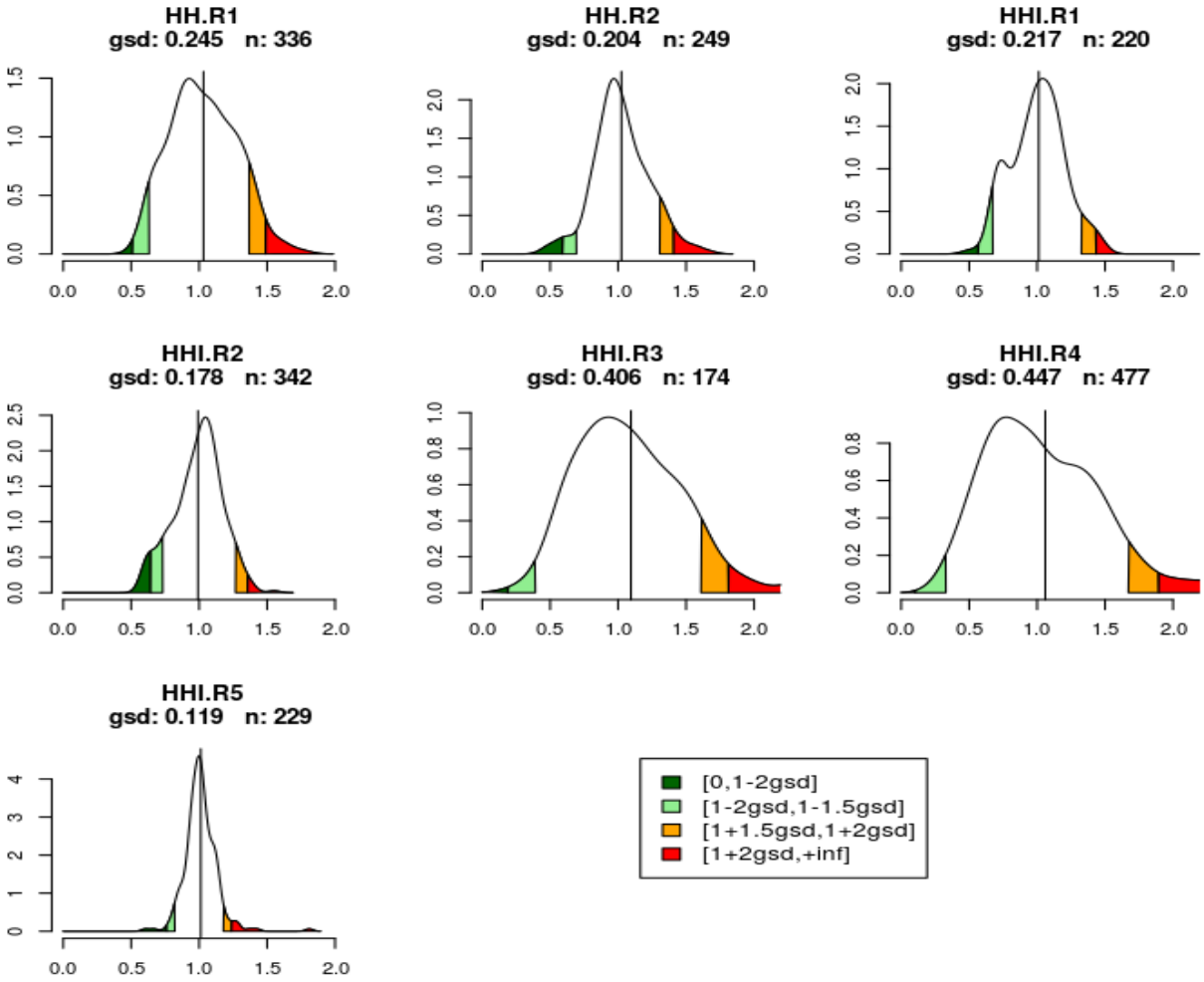


Figure 3.5 The seven plots correspond to the seven protein distributions in the corresponding samples. Narrower distributions (lower global standard deviation, gsd) will represent higher accuracy in the quantification and thus, provide more accurate quantification results. Each distribution has highlighted four categories of protein differential expression: “highly under-expressed” (dark green), “moderately under-expressed” (light green), “moderately over-expressed” (orange) and “highly over-expressed” (red).

3.3.9 Protein quantification: treatment of the technical replicates

Following the experimental design explained above, two technical replicates are available for each experimental condition (HH or HHI). Then, for a given protein in each experimental condition, we will have one or two quantitative determinations. Quantitative information presented by these replicates for a given protein can be then coherent (both replicates showing the over-expression, under-expression or not change at the same time) or not. We also must have present that a protein can appear differentially expressed in one replicate and do not show significant difference (or even show the opposite expression) in the other replicate, this effect being caused by the low precision inherent to the method (27).

To address this, we have used an approach that unifies the quantitative information presented by both replicates, providing a simple and consistent method for merging the information coming from two technical replicates.

We have chosen an approach that deals with experimental evidence. Every ratio associated to an individual measure (one technical replicate) will be translated to a given category depending on the interval of global standard deviation (S) in which is located. These categories of change in an experimental condition with respect to the control are:

- moderate under-expression (between 1.5 and 2 gsd values, -1S)
- no variation (between -1.5 and +1.5 gsd values, NV)
- moderate over-expression (between +1.5 and +2 gsd values, +1S)
- high over-expression (greater than 2 gsd values, +2S)

In Table 3.4, the mapping between a ratio interval and its corresponding category was shown for each sample. Using these categories, we construct a numerical expression that summarizes the behavior of one given protein in the two technical replicates, using the -2S, -1S, NV, +1S and +2S:

- NV value adds 0
- +1S and +2S add +1 and +2, respectively
- -1S and -2S subtract -1 and -2, respectively

Thus, and overall expression value ranging from -4 to +4 is obtained for every protein quantified. A protein with a +4 or -4 is not necessarily more or less expressed than another with +2 or -2: the real meaning of a high variation value is that there is more experimental evidence of the change, not the overall variation.

This can be better explained with two practical examples:

- If we have only one measurement for a protein, coming from one technical replicate, as in [1.779|G2,131], it means that for this protein (Q68FY0, HH experimental condition), only one result has successfully passed the quality threshold (a coefficient of variation inferior to 30%). For the Group 2 of measures (the group this example belongs to), the TMT label 131 presents a set of thresholds as [0,0.51] (proteins highly under-expressed), [0.51,0.632] (proteins moderately under-expressed), [1.368,1.49] (proteins moderately over-expressed) and [1.49,+inf] (proteins highly over-expressed). That gives a +2S (plus 2 global standard deviations) to a ratio of 1.779. The overall variation for this protein will be +2S, meaning that one technical replicate has shown a high over-expression of this protein versus the same protein in the control sample.
- A different measurement, coming from two technical replicates, as in [1.379|G2,126],[1.075|G2,131] (protein P48500, HH experimental condition) presents a more challenging scenario. The first replicate produces a +1S value, the second NV (no variation). The global variation for this sample will be +1S (+1+0=+1). These two replicates are not necessarily contradictory: with a 30% of variation used, both results could actually be over-expressed or show no variation at all.
- Values of variation for all the proteins quantified in this study are found in Table 3.4 in the results section.
- Of the original 1,156 TMT tag quantifications, 153 were removed for presenting a coefficient of variation higher than 30%. Of the remaining 1,003 quantification tags reported in the 340 proteins confidently identified, only in one case (P63041 for condition HH, with -2S|+2S) a contradictory result was obtained: two technical replicates indicating over and under-expression at the same time. That proves, in our opinion, the robustness of this specific approach.

- The work-flow used to generate a confident set of protein quantification is summarized in Figure 3.6, reporting also the number of proteins in each step:
- From an initial set of 1409 proteins quantified by Protein Discoverer, we apply several filters, that leave 339 confidently quantified proteins (only one discarded for presenting contradictory quantitative information).
- In the set of 339 confidently quantified proteins, only 99 present a variations in HH and/or HHI samples with respect to the controls.
- From these 99 proteins that differ in their expression to the controls, only 54 can be included in enriched GO (gene ontology) categories, as will be explained in the next section.

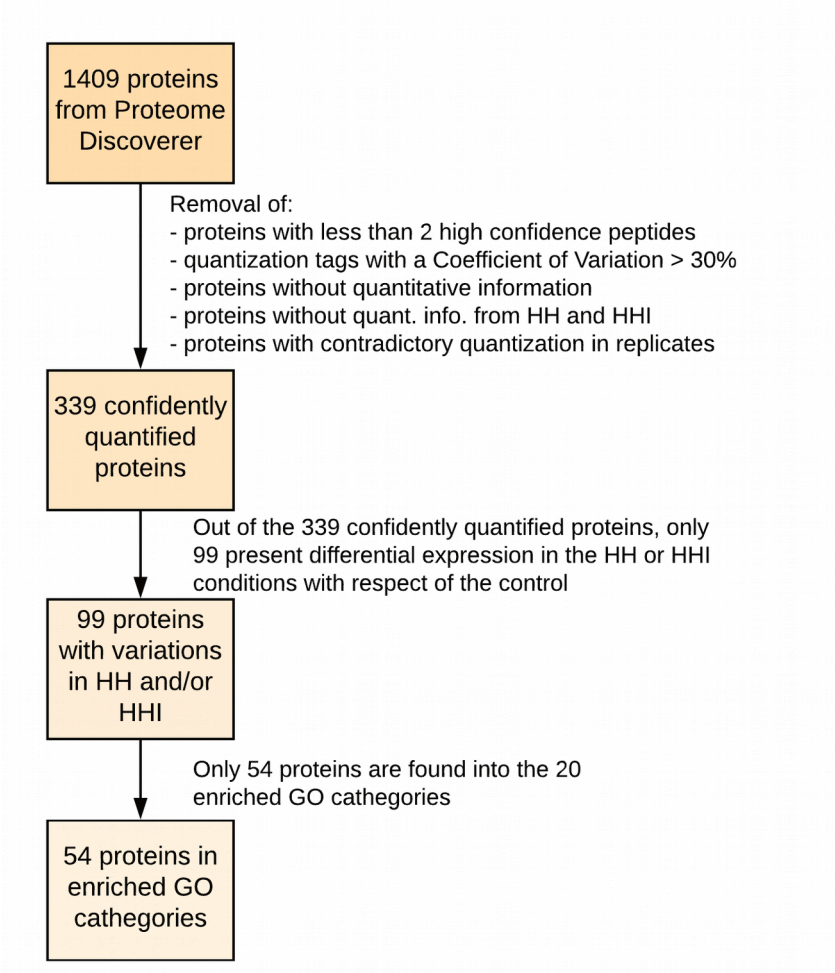


Figure 3.6: Overview of the protein identification and quantification work-flow used in the present analysis.

3.3.10 Gene set enrichment

To help in the interpretation of the results obtained in a proteomics experiment it is very common the use of a technique known as gene set enrichment or functional enrichment analysis (35). This technique uses the functional annotation of the proteins under study to infer which functions are more present in the set. The most used resource to extract functional information of a gene or protein is the Gene Ontology, or GO (36). This ontology is organized into three areas: molecular function (molecular activities of gene products), cellular component (where gene products are active) and biological process (pathways and

larger processes made up of the activities of multiple gene products). In addition to the description provided by these areas of the GO ontology, another linked resource is used: the Gene Ontology Annotation or GOA (37). Inside the GOA for a given organism, each gene (or protein) is associated with one or more terms of the GO. For *Rattus norvegicus*, 20,721 genes are annotated with 461,203 terms (march 2018).

In short, a gene set enrichment makes use of the GOA annotations to provide a list of GO terms that are over-represented. This over-representation points to the fact that some genes or proteins related to a given GO term appear in a higher frequency than expected by chance. To measure this over-representation, a statistical test is used, frequently a binomial (sampling with replacement) or a hypergeometric test (without replacement).

Several software are available at present to generate a gene set enrichment analysis, as desktop software (Ingenuity Pathways Analysis (38), Cytoscape (39) BinGO (40)) or web based application (GOrilla (41) , David (42), Toppgene (43)). We have selected the ClueGO (44) plug-in available in Cytoscape. Unlike other software like BinGO or Gorilla, that assess for over-represented GO terms and reconstruct a hierarchical ontology tree, ClueGO uses kappa statistics (45) to generate a network with GO terms as nodes in a network. To illustrate the way ClueGO displays the information, the graphical output obtained from an example dataset is shown in Figure 3.7.

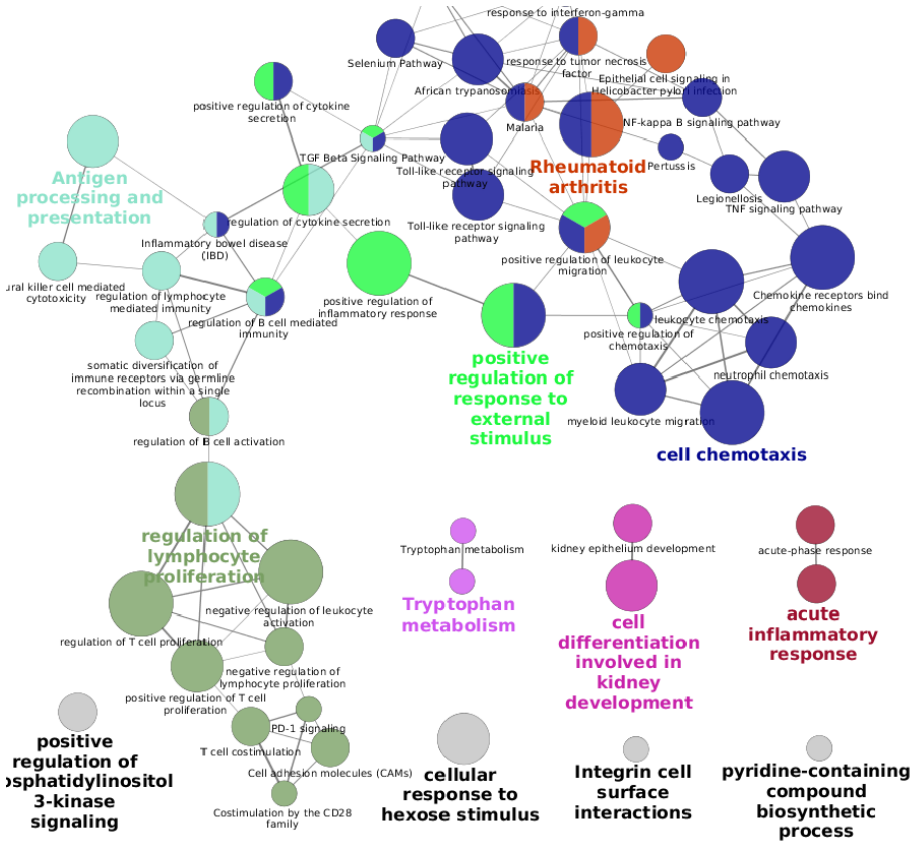


Figure 3.7: Graphical output produced by ClueGO in an example dataset. Each color represents a biological process representative of a set of closely related biological processes. Each node of the network as one or more genes linked (not displayed here). The lines represent interactions between the nodes

ClueGO generates a dynamical network considering the genes of interest (i.e. those over-represented) and integrating GO terms. In this network, genes belonging to two or more functional categories act as links and enriched GO terms (with their corresponding genes)

act as nodes. The closer that the represented genes are allocated in this network, the more likely they are going to interact in one or more biological functions . Additionally, neighboring and higher number of links between biological functions (GO terms) in the diagram will suggest that those functions are going to be more closely related in a biological sense.

3.4 Results

3.4.1 Protein identification and quantification

The results obtained in the proteomics experiment are displayed in Table 3.6, where 339 quantified proteins are shown; a complete description of the fields associated to each protein is provided in Table 3.5.

Protein	Uniprot Accession Number
Gene	Gene symbol associated to the protein
Description	Protein description
num	Total number of quantifications considered in the two conditions evaluated (HH and HHI)
data origin	Sample groups where the data comes from (G1, G2, G3 or G4) and TMT label (from 126 to 131).
variation	Symbols indicating if the ratio translates as high under-expression (-2S), moderate under-expression (-1S), no variation (NV) moderate over-expression (+1S) or high over-expression (+2S) with respect to the control. The table on top shows the mapping for each sample between the ratio value and the correspondent expression.
global var.	The overall expression for the experimental condition considering one or two replicates. A NV value adds 0, and the positive and negative values add or substract their value (e.g. +2S|+2S equals to +4). Only in one case, this method of global variation shows contradictory results: P63041 for HH ratios. The data associated to this protein is shown in the table as red, striked out values, leaving in 339 the set of confidently identified proteins.
symbol	"=" for same expression than the control, "+" for "moderate over-expression", "++" for high over-expression, "-" for moderate under-expression and "++" for high over-expression

Table 3.5 Values of global standard deviation (gsd) and thresholds used for each sample. Four categories of proteins are used: "highly under-expressed", "moderately under-expressed", "moderately over-expressed" and "highly over-expressed". In some samples, for big gsd values, the threshold will fall under 0 for the under-expressed proteins; in these cases, a NA value is displayed, and no proteins will be taken for that category.

Protein	Gene	Description	num	HH data origin	HH variation	HH global var.	HH symbol	HHI data origin	HHI variation	HHI global var.	HHI symbol
A0MY09	Hsp90b1	Endoplasmin	3	[0.892|G2,131],[1.081|G2,126]	NV|NV	0	=	[0.932|G2,129]	NV	0	=
A1L108	Arpc5l	Actin-related protein 2/3 complex sub5-like prot	3	[0.894|G2,126],[1.050|G2,131]	NV|NV	0	=	[1.027|G2,129]	NV	0	=
B0BNE6	Ndufs8	NADH dehyd(Ubiquinone) Fe-S prot8 (Pred), isofCRA_a	4	[0.454|G2,126],[1.219|G2,131]	-2S|NV	-2	- -	[1.172|G1,128],[1.069|G2,129]	NV|NV	0	=
B0K020	Cisd1	CDGSH iron-sulfur domain-containing protein 1	4	[0.814|G2,126],[1.024|G2,131]	NV|NV	0	=	[0.604|G1,128],[0.842|G2,129]	-2S|NV	-2	- -
B2GV73	Arpc3	Actin-related protein 2/3 complex subunit 3	4	[0.949|G2,131],[0.944|G2,126]	NV|NV	0	=	[0.985|G2,129],[1.223|G1,128]	NV|+1S	1	+
B2RYG6	Otub1	Ubiquitin thioesterase OTUB1	3	[1.016|G2,126],[0.694|G2,131]	NV|NV	0	=	[0.755|G2,129]	NV	0	=
B2RYS2	Uqcrb	Cytochrome b-c1 complex subunit 7	4	[1.441|G2,131],[0.848|G2,126]	+1S|NV	1	+	[1.126|G2,129],[1.084|G1,128]	NV|NV	0	=
B2RZ27	Sh3bgrl3	Protein Sh3bgrl3	3	[1.343|G2,131],[1.064|G2,126]	NV|NV	0	=	[1.160|G2,129]	NV	0	=
B2RZD6	Ndufa4	Ndufa4 protein	4	[1.088|G2,126],[1.157|G2,131]	NV|NV	0	=	[0.944|G1,128],[1.021|G2,129]	NV|NV	0	=
B3GNI6	Sept11	Septin-11	4	[1.088|G2,126],[0.986|G2,131]	NV|NV	0	=	[0.865|G1,128],[1.094|G2,129]	NV|NV	0	=
B4F7C2	Tubb4a	Protein Tubb4a	3	[0.908|G2,131]	NV	0	=	[0.979|G2,129],[1.093|G1,128]	NV|NV	0	=
D3ZAF6	Atp5j2	ATP synthase subunit f, mitochondrial	4	[0.721|G2,131],[0.617|G2,126]	NV|-1S	-1	-	[0.746|G2,129],[0.995|G1,128]	NV|NV	0	=
D3ZAY7	Epb41l2	Protein Epb41l2	3	[0.791|G2,126],[1.056|G2,131]	NV|NV	0	=	[1.029|G2,129]	NV	0	=
D3ZCN9	RGD1560073	Protein RGD1560073	3	[0.783|G2,131],[1.279|G2,126]	NV|NV	0	=	[0.843|G2,129]	NV	0	=
D3ZCZ9	LOC100912599	Protein LOC100912599	2	[1.515|G2,126]	+2S	2	+ +	[0.922|G2,129]	NV	0	=
D3ZD09	Cox6b1	Cytochrome c oxidase subunit 6B1	3	[1.411|G2,131]	+1S	1	+	[1.231|G2,129],[1.135|G1,128]	NV|NV	0	=
D3ZDF0	Nptn	Neuroplastin	4	[0.994|G2,126],[0.737|G2,131]	NV|NV	0	=	[0.948|G1,128],[0.806|G2,129]	NV|NV	0	=
D3ZDH8	Sept5	Platelet glycoprotein Ib beta chain	3	[0.879|G2,131],[1.197|G2,126]	NV|NV	0	=	[0.964|G2,129]	NV	0	=
D3ZDU5	Pfn2	Profilin	2	[1.094|G2,131]	NV	0	=	[1.097|G2,129]	NV	0	=
D3ZF13	LOC683884	Acyl carrier protein	3	[1.395|G2,131],[0.808|G2,126]	+1S|NV	1	+	[1.307|G2,129]	+1S	1	+
D3ZG43	Ndufs3	NADH dehyd(Ubiquinone) Fe-S prot8 (Pred), isofCRA_c	3	[0.910|G2,126],[1.002|G2,131]	NV|NV	0	=	[0.963|G2,129]	NV	0	=
D3ZH42	Mov10l1	Protein Mov10l1	3	[1.337|G2,126],[1.053|G2,131]	+1S|NV	1	+	[0.650|G2,129]	-1S	-1	-
D3ZH98		Uncharacterized protein	3	[1.586|G2,126],[1.549|G2,131]	+2S|+2S	4	+ +	[1.377|G2,129]	+2S	2	+ +
D3ZJ08	Hist2h3c2	Histone H3	4	[1.308|G2,126],[0.998|G2,131]	+1S|NV	1	+	[1.174|G1,128],[1.219|G2,129]	NV|NV	0	=
D3ZJF4		Uncharacterized protein (Fragment)	2	[0.988|G2,131]	NV	0	=	[0.999|G2,129]	NV	0	=
D3ZKD9	Mapt	Microtubule-associated protein	2	[1.199|G2,131]	NV	0	=	[1.114|G2,129]	NV	0	=
D3ZNH4	Hist1h2bo	Histone H2B	2	[1.228|G2,131]	NV	0	=	[1.041|G2,129]	NV	0	=
D3ZNI9	Kcnt1	Potassium channel subfamily T member 1	3	[0.982|G2,126],[0.907|G2,131]	NV|NV	0	=	[0.610|G2,129]	-2S	-2	- -
D3ZPP8	Sept3	Neuronal-specific septin-3	3	[0.956|G2,131],[0.945|G2,126]	NV|NV	0	=	[0.927|G2,129]	NV	0	=
D3ZQD3	Ogdhl	2-oxoglutarate dehydrogenase, mitochondrial	3	[1.121|G2,126],[0.902|G2,131]	NV|NV	0	=	[0.909|G2,129]	NV	0	=
D3ZQL7	Tppp	Protein Tppp	4	[1.108|G2,131],[1.039|G2,126]	NV|NV	0	=	[0.962|G2,129],[0.943|G1,128]	NV|NV	0	=
D3ZS58	Ndufa2	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 2	3	[1.085|G2,126],[1.227|G2,131]	NV|NV	0	=	[1.082|G2,129]	NV	0	=
D3ZXP3	H2afx	Histone H2A	3	[1.068|G2,131],[1.014|G2,126]	NV|NV	0	=	[0.933|G2,129]	NV	0	=
D3ZZ51	Prrc2b	Protein Prrc2b (Fragment)	3	[0.869|G2,131],[1.216|G2,126]	NV|NV	0	=	[0.935|G2,129]	NV	0	=
D4A0E2	Napg	Protein Napg	3	[1.024|G2,126],[0.975|G2,131]	NV|NV	0	=	[0.952|G2,129]	NV	0	=
D4A0F5	Sept7	Protein LOC100910754	4	[0.875|G2,131],[1.170|G2,126]	NV|NV	0	=	[0.962|G2,129],[1.043|G1,128]	NV|NV	0	=
D4A0T0	Ndufb10	Protein Ndufb10	2	[1.450|G2,131]	+1S	1	+	[1.108|G2,129]	NV	0	=
D4A133	Atp6v1a	Protein Atp6v1a	4	[0.913|G2,131],[1.119|G2,126]	NV|NV	0	=	[1.038|G2,129],[0.992|G1,128]	NV|NV	0	=
D4A678	Spta1	Protein Spta1	3	[1.394|G2,126],[0.854|G2,131]	+1S|NV	1	+	[1.009|G2,129]	NV	0	=
D4A8H3	Uba6	Protein Uba6	3	[1.282|G2,131],[1.087|G2,126]	NV|NV	0	=	[1.110|G2,129]	NV	0	=
D4AB12		Uncharacterized protein	3	[1.459|G2,126],[1.084|G2,131]	+2S|NV	2	+ +	[1.130|G2,129]	NV	0	=
D4ACQ2	LOC690384	Protein LOC690384	3	[1.370|G2,131],[1.067|G2,126]	+1S|NV	1	+	[1.217|G2,129]	NV	0	=
F1LM47	Sucla2	Succinyl-CoA ligase subunit beta	3	[0.838|G2,131],[1.101|G2,126]	NV|NV	0	=	[0.870|G2,129]	NV	0	=
F1LM82	Hnrnpa2b1	Heterogeneous nuclear ribonucleoproteins A2/B1	4	[1.087|G2,126],[1.236|G2,131]	NV|NV	0	=	[0.989|G1,128],[1.193|G2,129]	NV|NV	0	=
F1LMR7	Dpp6	Dipeptidyl aminopeptidase-like protein 6	2	[1.334|G2,126]	+1S	1	+	[0.804|G2,129]	NV	0	=
F1LMW7	Marcks	Myristoylated alanine-rich C-kinase substrate	3	[0.883|G2,126],[1.515|G2,131]	NV|+2S	2	+ +	[1.380|G2,129]	+2S	2	+ +
F1LNF7	Idh3a	Isocitrate dehyd [NAD] subunit, mitoch.	3	[1.019|G2,131]	NV	0	=	[1.056|G2,129],[1.034|G1,128]	NV|NV	0	=
F1LNN8	Dapk1	Protein Dapk1	3	[1.289|G2,131],[0.862|G2,126]	NV|NV	0	=	[1.160|G2,129]	NV	0	=
F1LPP0	Amph	Protein LOC100910792 (Fragment)	4	[1.260|G2,131],[1.221|G2,126]	NV|NV	0	=	[1.156|G2,129],[0.916|G1,128]	NV|NV	0	=
F1LPS8	Pura	Transcriptional activator protein Pur-alpha	3	[1.151|G2,131],[0.996|G2,126]	NV|NV	0	=	[1.088|G2,129]	NV	0	=
F1LQ05	Sh3gl2	Endophilin-A1 (Fragment)	4	[1.056|G2,131],[0.904|G2,126]	NV|NV	0	=	[1.026|G2,129],[1.003|G1,128]	NV|NV	0	=
F1LQ63	Tnr	Tenascin-R	4	[0.778|G2,131],[0.903|G2,126]	NV|NV	0	=	[0.839|G2,129],[0.903|G1,128]	NV|NV	0	=
F1LQ81	Nsf	Vesicle-fusing ATPase (Fragment)	4	[0.769|G2,131],[1.108|G2,126]	NV|NV	0	=	[0.830|G2,129],[0.944|G1,128]	NV|NV	0	=
F1LQ96	Sncg	Gamma-synuclein	3	[1.598|G2,131],[1.444|G2,126]	+2S|+2S	4	+ +	[1.269|G2,129]	+1S	1	+
F1LRK1	Atp4a	Potassium-transporting ATPase alpha chain 1	3	[0.882|G2,126],[0.566|G2,131]	NV|-1S	-1	-	[0.586|G2,129]	-2S	-2	- -
F1LRZ7	Nefh	Neurofilament heavy polypeptide	3	[0.918|G2,126],[1.135|G2,131]	NV|NV	0	=	[0.949|G2,129]	NV	0	=
F1LTZ6	RGD1559921	Protein RGD1559921 (Fragment)	4	[1.293|G2,131],[0.972|G2,126]	NV|NV	0	=	[1.073|G2,129],[1.031|G1,128]	NV|NV	0	=
F1LUM5	Tubal3	Protein Tubal3	2	[0.651|G2,131]	NV	0	=	[0.795|G2,129]	NV	0	=
F1LUV9	Ncam1	Neural cell adhesion molecule 1 (Fragment)	3	[0.637|G2,131]	NV	0	=	[0.880|G1,128],[0.795|G2,129]	NV|NV	0	=
F1M1D0	Krt79	Protein Krt79	2	[0.894|G2,131]	NV	0	=	[0.748|G2,129]	NV	0	=
F1M208	Piezo2	Protein Piezo2	3	[1.309|G2,131],[1.229|G2,126]	NV|NV	0	=	[1.262|G2,129]	NV	0	=
F1M269		Glyceraldehyde-3-phosphate dehydr (Frag)	3	[1.400|G2,126],[1.084|G2,131]	+1S|NV	1	+	[1.086|G2,129]	NV	0	=
F1M2D3	Vdac1	Uncharacterized protein	4	[1.232|G2,131],[1.128|G2,126]	NV|NV	0	=	[1.163|G2,129],[1.284|G1,128]	NV|+2S	2	+ +
F1M779	Cltc	Clathrin heavy chain	4	[0.626|G2,131],[0.784|G2,126]	-1S|NV	-1	-	[0.637|G2,129],[0.824|G1,128]	-2S|NV	-2	- -
F1M953	Hspa9	Stress-70 protein, mitochondrial	3	[1.138|G2,131]	NV	0	=	[1.092|G2,129],[1.018|G1,128]	NV|NV	0	=
F1M9C1	Leprel1	Prolyl 3-hydroxylase 2 (Fragment)	3	[1.101|G2,126],[1.035|G2,131]	NV|NV	0	=	[0.858|G2,129]	NV	0	=
F1MA36	Sptbn2	Spectrin beta 3	3	[0.669|G2,131],[1.056|G2,126]	NV|NV	0	=	[0.764|G2,129]	NV	0	=
F1MAQ5	Map2	Microtubule-associated protein	4	[1.100|G2,126],[1.223|G2,131]	NV|NV	0	=	[0.981|G1,128],[1.214|G2,129]	NV|NV	0	=
F7EYB9	Omg	Protein Omg	4	[0.675|G2,126],[0.904|G2,131]	-1S|NV	-1	-	[0.855|G1,128],[0.924|G2,129]	NV|NV	0	=
F7FEZ6	Hnrnpa1	Heterogeneous nuclear ribonucleoprotein A1	3	[0.997|G2,126],[0.951|G2,131]	NV|NV	0	=	[0.995|G2,129]	NV	0	=
F7FKI5	Pdha1	Pyruvate dehydrogenase E1 comp.subunit alpha	4	[0.903|G2,131],[0.878|G2,126]	NV|NV	0	=	[0.985|G2,129],[1.011|G1,128]	NV|NV	0	=
F8WG67	Acot7	Acyl-CoA thioesterase 7, isoform CRA_a	3	[0.934|G2,131],[1.029|G2,126]	NV|NV	0	=	[0.983|G2,129]	NV	0	=
G3V6A4	Hnrpd	Heterogeneous nuclear ribonucleop.D, isofCRA_b	4	[1.035|G2,126],[1.162|G2,131]	NV|NV	0	=	[1.157|G1,128],[1.260|G2,129]	NV|NV	0	=
G3V6D3	Atp5b	ATP synthase subunit beta	4	[0.809|G2,126],[1.404|G2,131]	NV|+1S	1	+	[1.118|G1,128],[1.143|G2,129]	NV|NV	0	=
G3V6P2	Dlst	Dihydrolipoamide S-succinyltransferase (E2 comp. of 2-oxo-glutarate complex), isoform CRA_a	3	[1.020|G2,126],[1.239|G2,131]	NV|NV	0	=	[1.047|G2,129]	NV	0	=
G3V6S0	Sptbn1	Protein Sptbn1	4	[1.072|G2,126],[0.698|G2,131]	NV|NV	0	=	[0.860|G1,128],[0.786|G2,129]	NV|NV	0	=
G3V6S8	Srsf6	Serine/arginine-rich splicing factor 6	3	[0.941|G2,131],[1.241|G2,126]	NV|NV	0	=	[1.026|G2,129]	NV	0	=
G3V6X7	Pcsk1n	Proprotein convertase subtilisin/kexin type 1 inhibitor	3	[1.386|G2,131]	+1S	1	+	[1.110|G1,128],[1.246|G2,129]	NV|NV	0	=
G3V6Y6	Pygb	Alpha-1,4 glucan phosphorylase	3	[0.629|G2,131],[0.797|G2,126]	-1S|NV	-1	-	[0.701|G2,129]	-1S	-1	-
G3V721	Wbp2	WW domain binding protein 2, isoform CRA_b	3	[1.098|G2,131],[0.980|G2,126]	NV|NV	0	=	[1.024|G2,129]	NV	0	=
G3V733	Syn2	Synapsin-2	3	[0.884|G2,131]	NV	0	=	[0.982|G2,129],[1.133|G1,128]	NV|NV	0	=
G3V7C6	Tubb4b	RCG45400	3	[0.892|G2,131]	NV	0	=	[1.089|G1,128],[0.971|G2,129]	NV|NV	0	=
G3V7J7	Eif5a2	Eukaryotic translation initiation factor 5A2 (Predicted)	4	[0.949|G2,126],[1.207|G2,131]	NV|NV	0	=	[0.970|G1,128],[1.076|G2,129]	NV|NV	0	=
G3V7L8	Atp6v1e1	ATPase, H+ transport,V1 subunE isoform 1/CRA_a	3	[1.161|G2,131]	NV	0	=	[1.006|G1,128],[1.028|G2,129]	NV|NV	0	=
G3V7Y3	Atp5d	ATP synthase subunit delta, mitochondrial	4	[1.627|G2,131],[1.111|G2,126]	+2S|NV	2	+ +	[1.319|G2,129],[0.971|G1,128]	+1S|NV	1	+
G3V846	Slc1a3	Amino acid transporter	4	[0.618|G2,126],[0.638|G2,131]	-1S|NV	-1	-	[0.928|G1,128],[0.620|G2,129]	NV|-2S	-2	- -
G3V874	Epb41l3	Erythrocyte protein band 4.1-like 3, isoform CRA_b	3	[0.911|G2,126],[1.102|G2,131]	NV|NV	0	=	[1.085|G2,129]	NV	0	=
G3V8C3	Vim	Vimentin	4	[1.234|G2,131],[0.966|G2,126]	NV|NV	0	=	[1.203|G2,129],[0.990|G1,128]	NV|NV	0	=
G3V8K2	Gng3	Guanine nucleotide-binding protein subunit gamma	3	[0.743|G2,131],[1.021|G2,126]	NV|NV	0	=	[0.768|G2,129]	NV	0	=
G3V8Q2	Ina	Alpha-internexin	3	[1.369|G2,131]	+1S	1	+	[0.962|G1,128],[1.056|G2,129]	NV|NV	0	=
G3V936	Cs	Citrate synthase	3	[0.855|G2,131]	NV	0	=	[0.888|G2,129],[1.006|G1,128]	NV|NV	0	=
G3V983	Gstm1	Glutathione S-transferase Mu 1	3	[0.948|G2,131],[1.020|G2,126]	NV|NV	0	=	[0.932|G2,129]	NV	0	=
G3V9B3	Mag	Myelin-associated glycoprotein	3	[0.668|G2,131]	NV	0	=	[0.590|G2,129],[0.841|G1,128]	-2S|NV	-2	- -
G3V9G3	Camk2b	Calcium/calmodulin-dep Pkinase II, beta, isofCRA_a	4	[0.620|G2,131],[1.016|G2,126]	-1S|NV	-1	-	[0.695|G2,129],[0.879|G1,128]	-1S|NV	-1	-
G3V9R8	Hnrnpc	Heterogeneous nuclear ribonucleoprotein C	2	[1.257|G2,131]	NV	0	=	[1.180|G2,129]	NV	0	=
M0R5J4		Uncharacterized protein	4	[1.302|G2,131],[1.289|G2,126]	NV|NV	0	=	[1.125|G2,129],[0.983|G1,128]	NV|NV	0	=
M0R757	LOC100360413	Elongation factor 1-alpha	4	[1.182|G2,126],[0.851|G2,131]	NV|NV	0	=	[0.967|G1,128],[0.854|G2,129]	NV|NV	0	=
M0RAD1	Crym	Ketimine reductase mu-crystallin (Fragment)	3	[1.112|G2,131],[0.842|G2,126]	NV|NV	0	=	[1.091|G2,129]	NV	0	=
M0RAU4		Uncharacterized protein (Fragment)	2	[1.322|G2,131]	NV	0	=	[1.122|G2,129]	NV	0	=
M0RBJ0	Gng2	Guanine nucleotide-binding P.subunit gamma(Frag)	3	[1.148|G2,131],[0.829|G2,126]	NV|NV	0	=	[0.897|G2,129]	NV	0	=
M0RDM4	LOC680322	Histone H2A	3	[1.014|G2,126],[1.082|G2,131]	NV|NV	0	=	[1.002|G2,129]	NV	0	=
O08839	Bin1	Myc box-dependent-interacting protein 1	2	[1.082|G2,131]	NV	0	=	[1.058|G2,129]	NV	0	=
O35244	Prdx6	Peroxiredoxin-6	4	[1.315|G2,131],[0.897|G2,126]	NV|NV	0	=	[1.042|G2,129],[1.069|G1,128]	NV|NV	0	=
O35796	C1qbp	Complement comp. 1Qsubc-bind prot, mitoch.	4	[1.133|G2,131],[0.824|G2,126]	NV|NV	0	=	[1.110|G2,129],[1.104|G1,128]	NV|NV	0	=
O35814	Stip1	Stress-induced-phosphoprotein 1	3	[1.105|G2,131],[1.182|G2,126]	NV|NV	0	=	[1.092|G2,129]	NV	0	=
O88339	Epn1	Epsin-1	2	[1.443|G2,131]	+1S	1	+	[1.124|G2,129]	NV	0	=
O88989	Mdh1	Malate dehydrogenase, cytoplasmic	3	[0.967|G2,131]	NV	0	=	[0.938|G2,129],[1.006|G1,128]	NV|NV	0	=
P00507	Got2	Aspartate aminotransferase, mitochondrial	3	[0.828|G2,131]	NV	0	=	[0.832|G2,129],[0.969|G1,128]	NV|NV	0	=
P01830	Thy1	Thy-1 membrane glycoprotein	4	[0.864|G2,126],[0.809|G2,131]	NV|NV	0	=	[0.930|G1,128],[0.807|G2,129]	NV|NV	0	=
P01946	Hba1	Hemoglobin subunit alpha-1/2	3	[0.969|G2,126],[0.957|G2,131]	NV|NV	0	=	[0.650|G2,129]	-1S	-1	-
P02091	Hbb	Hemoglobin subunit beta-1	2	[0.931|G2,131]	NV	0	=	[0.613|G2,129]	-2S	-2	- -
P02650	Apoe	Apolipoprotein E	3	[0.863|G2,126],[1.140|G2,131]	NV|NV	0	=	[1.077|G2,129]	NV	0	=
P02688	Mbp	Myelin basic protein	3	[1.094|G2,131]	NV	0	=	[0.809|G1,128],[0.736|G2,129]	-1S|NV	-1	-
P02770	Alb	Serum albumin	4	[1.313|G2,126],[0.748|G2,131]	+1S|NV	1	+	[0.990|G1,128],[0.682|G2,129]	NV|-1S	-1	-
P04631	S100b	Protein S100-B	4	[0.885|G2,126],[1.315|G2,131]	NV|NV	0	=	[0.795|G1,128],[1.122|G2,129]	-1S|NV	-1	-
P04636	Mdh2	Malate dehydrogenase, mitochondrial	3	[0.861|G2,131]	NV	0	=	[1.096|G1,128],[0.964|G2,129]	NV|NV	0	=

Protein	Gene	Description	num	HH data origin	HH variation	HH global var.	HH symbol	HHI data origin	HHI variation	HHI global var.	HHI symbol
P04642	Ldha	L-lactate dehydrogenase A chain	4	[0.925|G2,126],[0.736|G2,131]	NV|NV	0	=	[0.974|G1,128],[0.815|G2,129]	NV|NV	0	=
P04692	Tpm1	Tropomyosin alpha-1 chain	2	[1.369|G2,131]	+1S	1	+	[1.278|G2,129]	+1S	1	+
P04785	P4hb	Protein disulfide-isomerase	3	[0.806|G2,126],[0.961|G2,131]	NV|NV	0	=	[1.008|G2,129]	NV	0	=
P04797	Gapdh	Glyceraldehyde-3-phosphate dehydrogenase	4	[1.152|G2,131],[1.240|G2,126]	NV|NV	0	=	[1.132|G2,129],[1.115|G1,128]	NV|NV	0	=
P04904	Gsta3	Glutathione S-transferase alpha-3	2	[1.002|G2,131]	NV	0	=	[0.794|G2,129]	NV	0	=
P04906	Gstp1	Glutathione S-transferase P	4	[0.824|G2,131],[0.790|G2,126]	NV|NV	0	=	[0.804|G2,129],[0.986|G1,128]	NV|NV	0	=
P05065	Aldoa	Fructose-bisphosphate aldolase A	4	[1.330|G2,126],[0.945|G2,131]	+1S|NV	1	+	[1.028|G1,128],[0.891|G2,129]	NV|NV	0	=
P05708	Hk1	Hexokinase-1	4	[0.683|G2,131],[0.796|G2,126]	NV|NV	0	=	[0.693|G2,129],[0.927|G1,128]	-1S|NV	-1	-
P06685	Atp1a1	NA/K-transporting ATPase subunit alpha-1	3	[0.629|G2,131]	-1S	-1	-	[0.963|G1,128],[0.643|G2,129]	NV|-2S	-2	- -
P06686	Atp1a2	Na/K-transporting ATPase subunit alpha-2	3	[0.622|G2,131]	-1S	-1	-	[0.634|G2,129],[0.963|G1,128]	-2S|NV	-2	- -
P06687	Atp1a3	NA/K-transporting ATPase subunit alpha-3	3	[0.606|G2,131]	-1S	-1	-	[0.628|G2,129],[0.936|G1,128]	-2S|NV	-2	- -
P06761	Hspa5	78 kDa glucose-regulated protein	3	[1.349|G2,131]	NV	0	=	[1.015|G1,128],[1.206|G2,129]	NV|NV	0	=
P07171	Calb1	Calbindin	3	[0.855|G2,126],[1.390|G2,131]	NV|+1S	1	+	[1.368|G2,129]	+2S	2	+ +
P07323	Eno2	Gamma-enolase	4	[1.215|G2,126],[1.259|G2,131]	NV|NV	0	=	[0.932|G1,128],[1.066|G2,129]	NV|NV	0	=
P07335	Ckb	Creatine kinase B-type	4	[1.063|G2,131],[1.018|G2,126]	NV|NV	0	=	[1.036|G2,129],[1.042|G1,128]	NV|NV	0	=
P07340	Atp1b1	Sodium/potassium-transporting ATPase subunit beta-1	4	[0.662|G2,131],[0.830|G2,126]	NV|NV	0	=	[0.739|G2,129],[1.006|G1,128]	NV|NV	0	=
P07483	Fabp3	Fatty acid-binding protein, heart	3	[0.922|G2,126],[1.080|G2,131]	NV|NV	0	=	[0.992|G2,129]	NV	0	=
P07825	Syp	Synaptophysin	3	[0.608|G2,131]	-1S	-1	-	[0.931|G1,128],[0.736|G2,129]	NV|NV	0	=
P07895	Sod2	Superoxide dismutase [Mn], mitochondrial	4	[1.197|G2,126],[1.232|G2,131]	NV|NV	0	=	[1.129|G1,128],[1.018|G2,129]	NV|NV	0	=
P07936	Gap43	Neuromodulin	2	[1.624|G2,131]	+2S	2	+ +	[1.277|G2,129]	+1S	1	+
P07943	Akr1b1	Aldose reductase	3	[0.937|G2,131],[0.927|G2,126]	NV|NV	0	=	[0.994|G2,129]	NV	0	=
P08009	Gstm3	Glutathione S-transferase Yb-3	4	[0.949|G2,131],[0.944|G2,126]	NV|NV	0	=	[1.026|G2,129],[1.135|G1,128]	NV|NV	0	=
P08081	Clta	Clathrin light chain A	3	[1.284|G2,131]	NV	0	=	[1.190|G2,129],[1.138|G1,128]	NV|NV	0	=
P08082	Cltb	Clathrin light chain B	4	[1.417|G2,131],[1.210|G2,126]	+1S|NV	1	+	[1.311|G2,129],[1.273|G1,128]	+1S|+2S	3	+ +
P08461	Dlat	Dihydrolipoyllysine-residue acetyltrans.component of pyruvate DHcomplex, mitoch.	3	[1.099|G2,131]	NV	0	=	[1.022|G2,129],[1.051|G1,128]	NV|NV	0	=
P09117	Aldoc	Fructose-bisphosphate aldolase C	3	[1.042|G2,131]	NV	0	=	[1.051|G1,128],[0.926|G2,129]	NV|NV	0	=
P09495	Tpm4	Tropomyosin alpha-4 chain	3	[1.360|G2,131],[1.045|G2,126]	NV|NV	0	=	[1.242|G2,129]	NV	0	=
P09606	Glul	Glutamine synthetase	3	[0.959|G2,131]	NV	0	=	[1.038|G1,128],[0.937|G2,129]	NV|NV	0	=
P09951	Syn1	Synapsin-1	3	[0.978|G2,131]	NV	0	=	[0.931|G2,129],[1.048|G1,128]	NV|NV	0	=
P10111	Ppia	Peptidyl-prolyl cis-trans isomerase A	3	[1.144|G2,131]	NV	0	=	[1.077|G1,128],[1.122|G2,129]	NV|NV	0	=
P10860	Glud1	Glutamate dehydrogenase 1, mitochondrial	4	[1.390|G2,126],[0.933|G2,131]	+1S|NV	1	+	[0.991|G1,128],[1.037|G2,129]	NV|NV	0	=
P10888	Cox4i1	Cytochrome c oxidase sub4 isof1, mitoch	4	[0.862|G2,131],[1.111|G2,126]	NV|NV	0	=	[0.793|G2,129],[1.001|G1,128]	NV|NV	0	=
P10960	Psap	Sulfated glycoprotein 1	3	[1.257|G2,131],[0.842|G2,126]	NV|NV	0	=	[1.138|G2,129]	NV	0	=
P11232	Txn	Thioredoxin	3	[1.274|G2,126],[1.367|G2,131]	NV|NV	0	=	[1.252|G2,129]	NV	0	=
P11275	Camk2a	Ca/calmodulin-dep PK type II subunit alpha	4	[0.994|G2,126],[0.631|G2,131]	NV|-1S	-1	-	[0.928|G1,128],[0.718|G2,129]	NV|-1S	-1	-
P11598	Pdia3	Protein disulfide-isomerase A3	3	[0.966|G2,126],[0.862|G2,131]	NV|NV	0	=	[1.023|G2,129]	NV	0	=
P11951	Cox6c2	Cytochrome c oxidase subunit 6C-2	4	[0.715|G2,131],[1.140|G2,126]	NV|NV	0	=	[0.745|G2,129],[0.904|G1,128]	NV|NV	0	=
P11980	Pkm	Pyruvate kinase PKM	4	[1.194|G2,126],[0.830|G2,131]	NV|NV	0	=	[1.021|G1,128],[0.892|G2,129]	NV|NV	0	=
P12075	Cox5b	Cytochrome c oxidase sub5B, mitoch.	4	[1.152|G2,131],[0.886|G2,126]	NV|NV	0	=	[1.061|G2,129],[0.987|G1,128]	NV|NV	0	=
P12839	Nefm	Neurofilament medium polypeptide	3	[1.364|G2,131]	NV	0	=	[1.053|G2,129],[0.957|G1,128]	NV|NV	0	=
P13221	Got1	Aspartate aminotransferase, cytoplasmic	3	[0.970|G2,131],[0.972|G2,126]	NV|NV	0	=	[1.041|G2,129]	NV	0	=
P13233	Cnp	2',3'-cyclic-nucleotide 3'-phosphodiesterase	4	[0.684|G2,131],[1.190|G2,126]	NV|NV	0	=	[0.599|G2,129],[0.835|G1,128]	-2S|NV	-2	- -
P13383	Ncl	Nucleolin	3	[1.280|G2,126],[1.249|G2,131]	NV|NV	0	=	[1.153|G2,129]	NV	0	=
P13668	Stmn1	Stathmin	3	[1.153|G2,126],[1.200|G2,131]	NV|NV	0	=	[1.182|G2,129]	NV	0	=
P14408	Fh	Fumarate hydratase, mitochondrial	2	[1.309|G2,131]	NV	0	=	[1.113|G2,129]	NV	0	=
P14668	Anxa5	Annexin A5	2	[0.933|G2,131]	NV	0	=	[0.909|G2,129]	NV	0	=
P15205	Map1b	Microtubule-associated protein 1B	4	[1.288|G2,126],[1.297|G2,131]	NV|NV	0	=	[1.001|G1,128],[1.243|G2,129]	NV|NV	0	=
P15999	Atp5a1	ATP synthase subunit alpha, mitochondrial	4	[1.143|G2,131],[0.945|G2,126]	NV|NV	0	=	[1.087|G2,129],[1.082|G1,128]	NV|NV	0	=
P16086	Sptan1	Spectrin alpha chain, non-erythrocytic 1	4	[0.928|G2,131],[1.130|G2,126]	NV|NV	0	=	[1.042|G2,129],[0.943|G1,128]	NV|NV	0	=
P16290	Pgam2	Phosphoglycerate mutase 2	3	[0.983|G2,126],[1.301|G2,131]	NV|NV	0	=	[1.162|G2,129]	NV	0	=
P16617	Pgk1	Phosphoglycerate kinase 1	3	[1.144|G2,131]	NV	0	=	[1.055|G2,129],[0.987|G1,128]	NV|NV	0	=
P17764	Acat1	Acetyl-CoA acetyltransferase, mitochondrial	3	[1.086|G2,131],[1.325|G2,126]	NV|+1S	1	+	[1.119|G2,129]	NV	0	=
P18418	Calr	Calreticulin	3	[1.058|G2,131],[1.128|G2,126]	NV|NV	0	=	[1.080|G2,129]	NV	0	=
P19234	Ndufv2	NADH dehydr.[ubiquinone] flavoprot2, mitoch	4	[0.785|G2,126],[1.132|G2,131]	NV|NV	0	=	[0.944|G1,128],[1.054|G2,129]	NV|NV	0	=
P19511	Atp5f1	ATP synthase F(0) complex subB1, mitoch	4	[0.726|G2,131],[0.897|G2,126]	NV|NV	0	=	[0.774|G2,129],[1.152|G1,128]	NV|NV	0	=
P19527	Nefl	Neurofilament light polypeptide	3	[1.114|G2,126]	NV	0	=	[0.991|G1,128],[1.060|G2,129]	NV|NV	0	=
P19804	Nme2	Nucleoside diphosphate kinase B	3	[1.174|G2,131]	NV	0	=	[1.110|G1,128],[1.108|G2,129]	NV|NV	0	=
P20788	Uqcrfs1	Cytochrome b-c1 complex subRieske, mitoch	2	[0.871|G2,131]	NV	0	=	[0.978|G2,129]	NV	0	=
P21575	Dnm1	Dynamin-1	4	[0.901|G2,131],[0.991|G2,126]	NV|NV	0	=	[0.997|G2,129],[1.047|G1,128]	NV|NV	0	=
P21707	Syt1	Synaptotagmin-1	4	[0.949|G2,126],[0.698|G2,131]	NV|NV	0	=	[0.914|G1,128],[0.702|G2,129]	NV|-1S	-1	-
P22062	Pcmt1	Protein-L-isoaspartate(D-aspartate) O-methyltransferase	3	[0.795|G2,126],[1.023|G2,131]	NV|NV	0	=	[0.973|G2,129]	NV	0	=
P23565	Ina	Alpha-internexin	3	[1.373|G2,131]	+1S	1	+	[0.964|G1,128],[1.063|G2,129]	NV|NV	0	=
P25113	Pgam1	Phosphoglycerate mutase 1	3	[1.007|G2,126],[1.353|G2,131]	NV|NV	0	=	[1.297|G2,129]	+1S	1	+
P26772	Hspe1	10 kDa heat shock protein, mitochondrial	4	[1.464|G2,131],[1.053|G2,126]	+1S|NV	1	+	[1.275|G2,129],[0.951|G1,128]	+1S|NV	1	+
P27139	Ca2	Carbonic anhydrase 2	3	[0.789|G2,131],[0.752|G2,126]	NV|NV	0	=	[0.718|G2,129]	-1S	-1	-
P29419	Atp5i	ATP synthase subunit e, mitochondrial	4	[0.944|G2,126],[1.054|G2,131]	NV|NV	0	=	[1.021|G1,128],[0.992|G2,129]	NV|NV	0	=
P31044	Pebp1	Phosphatidylethanolamine-binding protein 1	3	[1.040|G2,126]	NV	0	=	[1.036|G1,128],[1.174|G2,129]	NV|NV	0	=
P31399	Atp5h	ATP synthase subunit d, mitochondrial	2	[1.452|G2,131]	+1S	1	+	[1.271|G2,129]	+1S	1	+
P31596	Slc1a2	Excitatory amino acid transporter 2	4	[0.676|G2,131],[0.803|G2,126]	NV|NV	0	=	[0.710|G2,129],[1.037|G1,128]	-1S|NV	-1	-
P32551	Uqcrc2	Cytochrome b-c1 complex sub2, mitoch	4	[1.041|G2,131],[0.964|G2,126]	NV|NV	0	=	[1.069|G2,129],[1.128|G1,128]	NV|NV	0	=
P32851	Stx1a	Syntaxin-1A	4	[0.959|G2,126],[0.626|G2,131]	NV|-1S	-1	-	[0.950|G1,128],[0.642|G2,129]	NV|-2S	-2	- -
P34058	Hsp90ab1	Heat shock protein HSP 90-beta	4	[0.842|G2,131],[1.055|G2,126]	NV|NV	0	=	[0.939|G2,129],[0.987|G1,128]	NV|NV	0	=
P34926	Map1a	Microtubule-associated protein 1A	4	[1.308|G2,126],[1.093|G2,131]	+1S|NV	1	+	[1.001|G1,128],[1.081|G2,129]	NV|NV	0	=
P35213	Ywhab	14-3-3 protein beta/alpha	3	[1.039|G2,131]	NV	0	=	[1.059|G2,129],[1.038|G1,128]	NV|NV	0	=
P35332	Hpcal4	Hippocalcin-like protein 4	3	[0.954|G2,126],[1.399|G2,131]	NV|+1S	1	+	[1.133|G2,129]	NV	0	=
P35704	Prdx2	Peroxiredoxin-2	3	[1.246|G2,131]	NV	0	=	[1.083|G2,129],[0.956|G1,128]	NV|NV	0	=
P37805	Tagln3	Transgelin-3	4	[0.959|G2,126],[1.123|G2,131]	NV|NV	0	=	[1.016|G1,128],[1.032|G2,129]	NV|NV	0	=
P39069	Ak1	Adenylate kinase isoenzyme 1	3	[1.041|G2,131]	NV	0	=	[0.891|G1,128],[0.827|G2,129]	NV|NV	0	=
P42123	Ldhb	L-lactate dehydrogenase B chain	4	[1.034|G2,126],[0.883|G2,131]	NV|NV	0	=	[1.001|G1,128],[0.930|G2,129]	NV|NV	0	=
P45592	Cfl1	Cofilin-1	4	[1.256|G2,126],[1.267|G2,131]	NV|NV	0	=	[0.970|G1,128],[1.217|G2,129]	NV|NV	0	=
P47728	Calb2	Calretinin	4	[1.292|G2,131],[0.931|G2,126]	NV|NV	0	=	[1.304|G2,129],[1.014|G1,128]	+1S|NV	1	+
P47819	Gfap	Glial fibrillary acidic protein	4	[1.420|G2,131],[1.107|G2,126]	+1S|NV	1	+	[1.147|G2,129],[1.038|G1,128]	NV|NV	0	=
P47858	Pfkm	ATP-dep.6-phosphofructokinase, muscle	2	[0.673|G2,131]	NV	0	=	[0.738|G2,129]	NV	0	=
P47942	Dpysl2	Dihydropyrimidinase-related protein 2	4	[0.987|G2,126],[1.037|G2,131]	NV|NV	0	=	[1.008|G1,128],[1.012|G2,129]	NV|NV	0	=
P48500	Tpi1	Triosephosphate isomerase	4	[1.379|G2,126],[1.075|G2,131]	+1S|NV	1	+	[1.003|G1,128],[1.059|G2,129]	NV|NV	0	=
P49432	Pdhb	Pyruvate DH E1 component sub.beta, mitoch.	3	[1.192|G2,131]	NV	0	=	[1.065|G2,129],[1.120|G1,128]	NV|NV	0	=
P50399	Gdi2	Rab GDP dissociation inhibitor beta	4	[0.844|G2,131],[1.039|G2,126]	NV|NV	0	=	[0.894|G2,129],[1.117|G1,128]	NV|NV	0	=
P50408	Atp6v1f	V-type proton ATPase subunit F	3	[1.038|G2,131],[0.989|G2,126]	NV|NV	0	=	[0.997|G2,129]	NV	0	=
P50503	St13	Hsc70-interacting protein	2	[0.818|G2,131]	NV	0	=	[0.864|G2,129]	NV	0	=
P50554	Abat	4-aminobutyrate aminotransf., mitoch	3	[0.949|G2,126],[0.628|G2,131]	NV|-1S	-1	-	[0.727|G2,129]	-1S	-1	-
P54311	Gnb1	Guanine nucleot-bind protG(S)(T) sub.b1	4	[0.817|G2,131],[1.316|G2,126]	NV|+1S	1	+	[0.907|G2,129],[1.039|G1,128]	NV|NV	0	=
P54313	Gnb2	Guanine nucleot-bind protG(S)(T) sub.b2	3	[1.316|G2,126],[0.783|G2,131]	+1S|NV	1	+	[0.905|G2,129]	NV	0	=
P56571		ES1 protein homolog, mitochondrial	3	[1.584|G2,126],[0.991|G2,131]	+2S|NV	2	+ +	[1.042|G2,129]	NV	0	=
P59215	Gnao1	Guanine nucleot-bind prot.G(o) sub.alpha	4	[0.697|G2,131],[1.033|G2,126]	NV|NV	0	=	[0.759|G2,129],[0.922|G1,128]	NV|NV	0	=
P60881	Snap25	Synaptosomal-associated protein 25	3	[0.978|G2,131]	NV	0	=	[1.004|G1,128],[0.892|G2,129]	NV|NV	0	=
P61206	Arf3	ADP-ribosylation factor 3	4	[0.780|G2,126],[0.815|G2,131]	NV|NV	0	=	[0.829|G1,128],[0.896|G2,129]	NV|NV	0	=
P61765	Stxbp1	Syntaxin-binding protein 1	3	[0.749|G2,131]	NV	0	=	[1.019|G1,128],[0.785|G2,129]	NV|NV	0	=
P61980	Hnrnpk	Heterogeneous nuclear ribonucleoprotein K	3	[1.201|G2,131],[0.947|G2,126]	NV|NV	0	=	[1.133|G2,129]	NV	0	=
P61983	Ywhag	14-3-3 protein gamma	3	[1.068|G2,131]	NV	0	=	[1.074|G2,129],[1.039|G1,128]	NV|NV	0	=
P62161	Calm1	Calmodulin	2	[1.437|G2,131]	+1S	1	+	[1.189|G2,129]	NV	0	=
P62260	Ywhae	14-3-3 protein epsilon	3	[1.003|G2,131]	NV	0	=	[1.057|G2,129],[1.021|G1,128]	NV|NV	0	=
P62632	Eef1a2	Elongation factor 1-alpha 2	4	[1.246|G2,126],[0.851|G2,131]	NV|NV	0	=	[0.929|G1,128],[0.843|G2,129]	NV|NV	0	=
P62749	Hpcal1	Hippocalcin-like protein 1	3	[1.341|G2,131],[0.940|G2,126]	NV|NV	0	=	[1.159|G2,129]	NV	0	=
P62762	Vsnl1	Visinin-like protein 1	4	[0.879|G2,126],[1.394|G2,131]	NV|+1S	1	+	[0.997|G1,128],[1.198|G2,129]	NV|NV	0	=
P62775	Mtpn	Myotrophin	3	[1.047|G2,126],[1.247|G2,131]	NV|NV	0	=	[1.228|G2,129]	NV	0	=
P62815	Atp6v1b2	V-type proton ATPase subunit B, brain isoform	3	[0.946|G2,131]	NV	0	=	[1.010|G1,128],[1.011|G2,129]	NV|NV	0	=
P62836	Rap1a	Ras-related protein Rap-1A	3	[0.718|G2,126],[0.767|G2,131]	NV|NV	0	=	[0.783|G2,129]	NV	0	=
P62898	Cycs	Cytochrome c, somatic	3	[1.045|G2,126],[1.000|G2,131]	NV|NV	0	=	[1.016|G2,129]	NV	0	=
P62944	Ap2b1	AP-2 complex subunit beta	4	[0.938|G2,126],[0.642|G2,131]	NV|NV	0	=	[1.018|G1,128],[0.682|G2,129]	NV|-1S	-1	-
P62959	Hint1	Histidine triad nucleotide-binding protein 1	4	[1.148|G2,131],[1.013|G2,126]	NV|NV	0	=	[1.103|G2,129],[1.007|G1,128]	NV|NV	0	=
P62963	Pfn1	Profilin-1	2	[1.032|G2,126]	NV	0	=	[0.946|G2,129]	NV	0	=
P63012	Rab3a	Ras-related protein Rab-3A	4	[0.927|G2,126],[0.633|G2,131]	NV|NV	0	=	[1.035|G1,128],[0.776|G2,129]	NV|NV	0	=
P63018	Hspa8	Heat shock cognate 71 kDa protein	3	[1.340|G2,131]	NV	0	=	[1.061|G1,128],[1.223|G2,129]	NV|NV	0	=
P63039	Hspd1	60 kDa heat shock protein, mitochondrial	3	[1.278|G2,131]	NV	0	=	[1.113|G2,129],[1.003|G1,128]	NV|NV	0	=
~~P63041~~	~~Cplx1~~	~~Complexin-1~~	~~3~~	~~[1.516|G2,131],[0.497|G2,126]~~	~~+2S|-2S~~	NA	NA	~~[1.374|G2,129]~~	~~+2S~~	~~2~~	~~+-+~~

Protein	Gene	Description	num	HH data origin	HH variation	HH global var.	HH symbol	HHI data origin	HHI variation	HHI global var.	HHI symbol
P63086	Mapk1	Mitogen-activated protein kinase 1	4	[1.166|G2,126],[0.815|G2,131]	NV|NV	0	=	[1.007|G1,128],[0.913|G2,129]	NV|NV	0	=
P63100	Ppp3r1	Calcineurin subunit B type 1	3	[0.810|G2,126],[0.795|G2,131]	NV|NV	0	=	[0.949|G2,129]	NV	0	=
P63102	Ywhaz	14-3-3 protein zeta/delta	3	[1.029|G2,131]	NV	0	=	[1.071|G1,128],[1.063|G2,129]	NV|NV	0	=
P63329	Ppp3ca	Serine/threonine-protein phosphatase 2B catalytic subunit alpha isoform	4	[0.703|G2,131],[1.346|G2,126]	NV|+1S	1	+	[0.815|G2,129],[0.959|G1,128]	NV|NV	0	=
P67779	Phb	Prohibitin	4	[0.930|G2,126],[1.292|G2,131]	NV|NV	0	=	[1.000|G1,128],[1.097|G2,129]	NV|NV	0	=
P68035	Actc1	Actin, alpha cardiac muscle 1	2	[0.813|G2,131]	NV	0	=	[0.943|G2,129]	NV	0	=
P68255	Ywhaq	14-3-3 protein theta	3	[0.998|G2,131]	NV	0	=	[1.042|G2,129],[0.999|G1,128]	NV|NV	0	=
P68370	Tuba1a	Tubulin alpha-1A chain	4	[0.783|G2,131],[0.957|G2,126]	NV|NV	0	=	[0.858|G2,129],[1.016|G1,128]	NV|NV	0	=
P68511	Ywhah	14-3-3 protein eta	3	[1.044|G2,131]	NV	0	=	[1.037|G1,128],[1.062|G2,129]	NV|NV	0	=
P69682	Necap1	Adaptin ear-binding coat-associated protein 1	2	[1.192|G2,131]	NV	0	=	[1.134|G2,129]	NV	0	=
P69897	Tubb5	Tubulin beta-5 chain	3	[0.901|G2,131]	NV	0	=	[0.985|G2,129],[1.075|G1,128]	NV|NV	0	=
P70566	Tmod2	Tropomodulin-2	2	[1.105|G2,131]	NV	0	=	[0.995|G2,129]	NV	0	=
P70580	Pgrmc1	Membrane-associated progesterone receptor component 1	3	[0.831|G2,131],[1.170|G2,126]	NV|NV	0	=	[0.897|G2,129]	NV	0	=
P80254	Ddt	D-dopachrome decarboxylase	2	[1.388|G2,131]	+1S	1	+	[1.272|G2,129]	+1S	1	+
P81155	Vdac2	Voltage-dependent anion-selective channel protein 2	4	[0.877|G2,131],[0.903|G2,126]	NV|NV	0	=	[0.887|G2,129],[0.938|G1,128]	NV|NV	0	=
P82995	Hsp90aa1	Heat shock protein HSP 90-alpha	4	[0.825|G2,131],[1.094|G2,126]	NV|NV	0	=	[0.916|G2,129],[0.953|G1,128]	NV|NV	0	=
P84076	Hpca	Neuron-specific calcium-binding protein hippocalcin	4	[1.229|G2,131],[0.966|G2,126]	NV|NV	0	=	[1.145|G2,129],[1.019|G1,128]	NV|NV	0	=
P84087	Cplx2	Complexin-2	4	[0.497|G2,126],[1.327|G2,131]	-2S|NV	-2	- -	[1.148|G1,128],[1.350|G2,129]	NV|+1S	1	+
P84817	Fis1	Mitochondrial fission 1 protein	3	[0.895|G2,131],[0.568|G2,126]	NV|-2S	-2	- -	[0.924|G2,129]	NV	0	=
P85108	Tubb2a	Tubulin beta-2A chain	3	[0.949|G2,131]	NV	0	=	[1.066|G1,128],[1.005|G2,129]	NV|NV	0	=
P85515	Actr1a	Alpha-centractin	3	[0.975|G2,126],[0.779|G2,131]	NV|NV	0	=	[0.932|G2,129]	NV	0	=
P85834	Tufm	Elongation factor Tu, mitochondrial	4	[1.168|G2,126],[0.897|G2,131]	NV|NV	0	=	[0.862|G1,128],[0.847|G2,129]	NV|NV	0	=
P85845	Fscn1	Fascin	4	[1.335|G2,126],[0.867|G2,131]	+1S|NV	1	+	[1.103|G1,128],[0.987|G2,129]	NV|NV	0	=
P85969	Napb	Beta-soluble NSF attachment protein	3	[1.029|G2,131]	NV	0	=	[1.048|G1,128],[0.988|G2,129]	NV|NV	0	=
Q00981	Uchl1	Ubiquitin carboxyl-terminal hydrolase isozyme L1	3	[1.273|G2,131]	NV	0	=	[1.194|G2,129],[1.015|G1,128]	NV|NV	0	=
Q02563	Sv2a	Synaptic vesicle glycoprotein 2A	3	[1.062|G2,126],[0.761|G2,131]	NV|NV	0	=	[0.925|G2,129]	NV	0	=
Q03344	Atpif1	ATPase inhibitor, mitochondrial	4	[1.139|G2,126],[1.438|G2,131]	NV|+1S	1	+	[1.081|G1,128],[1.310|G2,129]	NV|+1S	1	+
Q05140	Snap91	Clathrin coat assembly protein AP180	3	[1.222|G2,126],[0.971|G2,131]	NV|NV	0	=	[1.028|G2,129]	NV	0	=
Q05175	Basp1	Brain acid soluble protein 1	3	[1.779|G2,131]	+2S	2	+ +	[1.551|G2,129],[1.073|G1,128]	+2S|NV	2	+ +
Q05962	Slc25a4	ADP/ATP translocase 1	4	[0.638|G2,131],[0.783|G2,126]	NV|NV	0	=	[0.629|G2,129],[0.940|G1,128]	-2S|NV	-2	- -
Q05982	Nme1	Nucleoside diphosphate kinase A	4	[1.174|G2,131],[0.927|G2,126]	NV|NV	0	=	[1.138|G2,129],[1.109|G1,128]	NV|NV	0	=
Q06645	Atp5g1	ATP synthase F(0) complex subunit C1, mitochondrial	4	[0.818|G2,126],[0.527|G2,131]	NV|-1S	-1	-	[0.918|G1,128],[0.597|G2,129]	NV|-2S	-2	- -
Q06647	Atp5o	ATP synthase subunit O, mitochondrial	4	[0.877|G2,131],[0.903|G2,126]	NV|NV	0	=	[0.859|G2,129],[0.956|G1,128]	NV|NV	0	=
Q09073	Slc25a5	ADP/ATP translocase 2	4	[0.653|G2,131],[0.896|G2,126]	NV|NV	0	=	[0.684|G2,129],[0.949|G1,128]	-1S|NV	-1	-
Q3KR86	Immt	MICOS complex subunit Mic60 (Fragment)	3	[0.911|G2,131],[1.198|G2,126]	NV|NV	0	=	[0.998|G2,129]	NV	0	=
Q3ZB98	Bcas1	Breast carcinoma-amplified sequence 1 homolog (Fragment)	3	[1.712|G2,131],[1.536|G2,126]	+2S|+2S	4	+ +	[1.097|G2,129]	NV	0	=
Q4FZY0	Efhd2	EF-hand domain-containing protein D2	3	[1.203|G2,131],[1.164|G2,126]	NV|NV	0	=	[1.083|G2,129]	NV	0	=
Q4KLX9	Ccdc163	Protein Ccdc163	3	[0.891|G2,131],[0.616|G2,126]	NV|-1S	-1	-	[0.979|G2,129]	NV	0	=
Q4KM73	Cmpk1	UMP-CMP kinase	3	[1.206|G2,131],[0.724|G2,126]	NV|NV	0	=	[0.989|G2,129]	NV	0	=
Q4KMA2	Rad23b	UV excision repair protein RAD23 homolog B	3	[0.826|G2,126],[1.164|G2,131]	NV|NV	0	=	[1.126|G2,129]	NV	0	=
Q4QQV0	Tubb6	Protein Tubb6	2	[0.920|G2,131]	NV	0	=	[1.014|G2,129]	NV	0	=
Q4QRB4	Tubb3	Tubulin beta-3 chain	3	[0.908|G2,131]	NV	0	=	[0.990|G2,129],[1.071|G1,128]	NV|NV	0	=
Q5BJT9	Ckmt1b	Creatine kinase, mitochondrial 1, ubiquitous	4	[1.070|G2,126],[0.845|G2,131]	NV|NV	0	=	[0.982|G1,128],[0.874|G2,129]	NV|NV	0	=
Q5FVI4	Cend1	Cell cycle exit and neuronal differentiation protein 1	3	[1.001|G2,131],[1.174|G2,126]	NV|NV	0	=	[0.889|G2,129]	NV	0	=
Q5M7A7	Cnrip1	CB1 cannabinoid receptor-interacting prot 1	2	[0.835|G2,126]	NV	0	=	[1.018|G2,129]	NV	0	=
Q5M7T6	Atp6v0d1	ATPase, H+ transp, lysos 38kDa, V0subd1	4	[0.958|G2,126],[0.742|G2,131]	NV|NV	0	=	[0.900|G1,128],[0.901|G2,129]	NV|NV	0	=
Q5M9I5	Uqcrh	Cytochrome b-c1 complex subunit 6, mitoch	2	[0.909|G2,126]	NV	0	=	[0.898|G2,129]	NV	0	=
Q5PPN5	Tppp3	Tubulin polymerization-promoting prot.family memb3	3	[1.170|G2,131],[1.163|G2,126]	NV|NV	0	=	[1.029|G2,129]	NV	0	=
Q5PQK2	Fus	Fusion, derived from t(1216) malignant liposarcoma (Human)	2	[1.015|G2,131]	NV	0	=	[1.048|G2,129]	NV	0	=
Q5PQN0	Ncald	Neurocalcin-delta	4	[0.874|G2,126],[1.345|G2,131]	NV|NV	0	=	[0.974|G1,128],[1.158|G2,129]	NV|NV	0	=
Q5RJQ4	Sirt2	NAD-dependent protein deacetylase sirtuin-2	3	[0.977|G2,131]	NV	0	=	[0.919|G1,128],[0.983|G2,129]	NV|NV	0	=
Q5RKJ9	Rab10	RAB10, member RAS oncogene family	4	[0.538|G2,126],[0.537|G2,131]	-2S|-1S	-3	- -	[0.953|G1,128],[0.622|G2,129]	NV|-2S	-2	- -
Q5U318	Pea15	Astrocytic phosphoprotein PEA-15	4	[0.940|G2,126],[1.124|G2,131]	NV|NV	0	=	[0.911|G1,128],[1.119|G2,129]	NV|NV	0	=
Q5XI34	Ppp2r1a	Protein Ppp2r1a	3	[0.910|G2,131],[1.212|G2,126]	NV|NV	0	=	[0.904|G2,129]	NV	0	=
Q5XI73	Arhgdia	Rho GDP-dissociation inhibitor 1	4	[0.982|G2,126],[1.062|G2,131]	NV|NV	0	=	[1.062|G1,128],[1.088|G2,129]	NV|NV	0	=
Q5XIF3	Ndufs4	NADH dehydrogenase [ubiquinone] iron-sulfur protein 4, mitochondrial	3	[1.342|G2,126],[1.251|G2,131]	+1S|NV	1	+	[0.992|G2,129]	NV	0	=
Q5XIF6	Tuba4a	Tubulin alpha-4A chain	3	[0.777|G2,131]	NV	0	=	[0.987|G1,128],[0.830|G2,129]	NV|NV	0	=
Q5XIH7	Phb2	Prohibitin-2	3	[0.925|G2,131],[0.866|G2,126]	NV|NV	0	=	[0.874|G2,129]	NV	0	=
Q62669		Protein Hbb-b1	3	[0.972|G2,131],[1.092|G2,126]	NV|NV	0	=	[0.636|G2,129]	-2S	-2	- -
Q62950	Crmp1	Dihydropyrimidinase-related protein 1	4	[1.092|G2,131],[1.005|G2,126]	NV|NV	0	=	[1.099|G2,129],[1.146|G1,128]	NV|NV	0	=
Q63028	Add1	Alpha-adducin	2	[0.925|G2,131]	NV	0	=	[1.024|G2,129]	NV	0	=
Q63198	Cntn1	Contactin-1	2	[0.836|G2,131]	NV	0	=	[0.930|G2,129]	NV	0	=
Q63228	Gmfb	Glia maturation factor beta	3	[0.883|G2,126],[0.975|G2,131]	NV|NV	0	=	[1.023|G2,129]	NV	0	=
Q63345	Mog	Myelin-oligodendrocyte glycoprotein	3	[0.641|G2,131]	NV	0	=	[0.849|G1,128],[0.581|G2,129]	NV|-2S	-2	- -
Q63429	Ubc	Polyubiquitin-C	2	[0.984|G2,126]	NV	0	=	[1.175|G2,129]	NV	0	=
Q63560	Map6	Microtubule-associated protein 6	2	[1.207|G2,131]	NV	0	=	[1.047|G2,129]	NV	0	=
Q63564	Sv2b	Synaptic vesicle glycoprotein 2B	3	[0.604|G2,126],[0.733|G2,131]	-1S|NV	-1	-	[0.813|G2,129]	NV	0	=
Q63610	Tpm3	Tropomyosin alpha-3 chain	3	[1.064|G2,126],[1.259|G2,131]	NV|NV	0	=	[1.206|G2,129]	NV	0	=
Q63716	Prdx1	Peroxiredoxin-1	3	[0.947|G2,126],[1.133|G2,131]	NV|NV	0	=	[0.965|G2,129]	NV	0	=
Q63754	Sncb	Beta-synuclein	3	[1.449|G2,126],[1.537|G2,131]	+2S|+2S	4	+ +	[1.208|G2,129]	NV	0	=
Q64119	Myl6	Myosin light polypeptide 6	3	[1.114|G2,131],[1.071|G2,126]	NV|NV	0	=	[1.032|G2,129]	NV	0	=
Q66H11	RGD1306195	Protein RGD1306195	3	[1.040|G2,126],[1.034|G2,131]	NV|NV	0	=	[0.922|G2,129]	NV	0	=
Q66HF1	Ndufs1	NADH-ubiquinone oxidoreductase 75 kDa subunit, mitochondrial	3	[0.967|G2,131],[1.005|G2,126]	NV|NV	0	=	[0.958|G2,129]	NV	0	=
Q68FX0	Idh3B	Isocitrate dehydrogenase [NAD] subunit beta, mitochondrial	4	[0.978|G2,131],[1.233|G2,126]	NV|NV	0	=	[1.064|G2,129],[1.087|G1,128]	NV|NV	0	=
Q68FY0	Uqcrc1	Cytochrome b-c1 complex subunit 1, mitochondrial	3	[0.926|G2,131]	NV	0	=	[0.984|G1,128],[1.015|G2,129]	NV|NV	0	=
Q6AXX6	Fam213a	Redox-regulatory protein FAM213A	3	[0.638|G2,131],[0.676|G2,126]	NV|-1S	-1	-	[0.626|G2,129]	-2S	-2	- -
Q6AYH5	Dctn2	Dynactin subunit 2	3	[0.906|G2,126],[1.187|G2,131]	NV|NV	0	=	[1.047|G2,129]	NV	0	=
Q6AZ25	Tpm1	Tropomyosin 1, alpha	3	[1.296|G2,131],[0.989|G2,126]	NV|NV	0	=	[1.210|G2,129]	NV	0	=
Q6P503	Atp6v1d	ATPase, H+ transporting, V1 subunit D, isoform CRA_c	3	[0.870|G2,131],[0.771|G2,126]	NV|NV	0	=	[0.886|G2,129]	NV	0	=
Q6P6R2	Dld	Dihydrolipoyl dehydrogenase, mitochondrial	3	[0.841|G2,131]	NV	0	=	[1.028|G1,128],[0.968|G2,129]	NV|NV	0	=
Q6P6V0	Gpi	Glucose-6-phosphate isomerase	4	[0.932|G2,126],[0.701|G2,131]	NV|NV	0	=	[1.019|G1,128],[0.727|G2,129]	NV|-1S	-1	-
Q6P7Q4	Glo1	Lactoylglutathione lyase	2	[0.869|G2,131]	NV	0	=	[0.759|G2,129]	NV	0	=
Q6PDU7	Atp5l	ATP synthase subunit g, mitochondrial	4	[0.850|G2,131],[0.601|G2,126]	NV|-1S	-1	-	[0.765|G2,129],[1.424|G1,128]	NV|+2S	2	+ +
Q6PEC4	Skp1	S-phase kinase-associated protein 1	3	[0.951|G2,126],[1.137|G2,131]	NV|NV	0	=	[1.104|G2,129]	NV	0	=
Q6QI09	Taf3	ATP synthase subunit gamma, mitochondrial	3	[0.926|G2,131],[0.954|G2,126]	NV|NV	0	=	[1.015|G2,129]	NV	0	=
Q6TXF3	Dbi	Acyl-CoA-binding protein	3	[1.412|G2,131],[0.972|G2,126]	+1S|NV	1	+	[1.277|G2,129]	+1S	1	+
Q71UE8	Nedd8	NEDD8	3	[0.912|G2,126],[1.178|G2,131]	NV|NV	0	=	[1.080|G2,129]	NV	0	=
Q78P75	Dynll2	Dynein light chain 2, cytoplasmic	4	[1.051|G2,126],[0.738|G2,131]	NV|NV	0	=	[1.076|G1,128],[0.822|G2,129]	NV|NV	0	=
Q7M0E3	Dstn	Destrin	3	[1.113|G2,126],[1.351|G2,131]	NV|NV	0	=	[1.239|G2,129]	NV	0	=
Q7M767	Ube2v2	Ubiquitin-conjugating enzyme E2 variant 2	3	[1.267|G2,126],[1.081|G2,131]	NV|NV	0	=	[1.044|G2,129]	NV	0	=
Q7TPK5	Eef1b2l	Ac2-067	3	[1.064|G2,126],[1.115|G2,131]	NV|NV	0	=	[0.906|G2,129]	NV	0	=
Q812E9	Gpm6a	Neuronal membrane glycoprotein M6-a	3	[0.727|G2,126]	NV	0	=	[0.715|G2,129],[0.906|G1,128]	-1S|NV	-1	-
Q8CHN7	Pcp4	Neuron-specific protein PEP-19	3	[0.952|G2,126],[1.273|G2,131]	NV|NV	0	=	[1.094|G2,129]	NV	0	=
Q8R2H0	Atp6v1g2	ATPase, H+ transporting, V1 subunit G isoform 2	3	[1.325|G2,126],[1.328|G2,131]	+1S|NV	1	+	[1.194|G2,129]	NV	0	=
Q8SEZ5		Cytochrome c oxidase subunit 2	2	[0.715|G2,131]	NV	0	=	[0.609|G2,129]	-2S	-2	- -
Q920Q0	Palm	Paralemmin-1	3	[1.194|G2,126],[1.215|G2,131]	NV|NV	0	=	[1.167|G2,129]	NV	0	=
Q99MZ8	Lasp1	LIM and SH3 domain protein 1	3	[1.184|G2,126],[1.080|G2,131]	NV|NV	0	=	[1.117|G2,129]	NV	0	=
Q9EQX9	Ube2n	Ubiquitin-conjugating enzyme E2 N	3	[0.877|G2,126],[1.269|G2,131]	NV|NV	0	=	[1.091|G2,129]	NV	0	=
Q9ER34	Aco2	Aconitate hydratase, mitochondrial	3	[0.843|G2,131]	NV	0	=	[0.861|G2,129],[1.029|G1,128]	NV|NV	0	=
Q9JKB7	Gda	Guanine deaminase	2	[0.918|G2,131]	NV	0	=	[1.110|G2,129]	NV	0	=
Q9R063	Prdx5	Peroxiredoxin-5, mitochondrial	3	[1.327|G2,131],[1.023|G2,126]	NV|NV	0	=	[1.137|G2,129]	NV	0	=

Table 3.6 Complete list of the 339 confidently identified and quantified proteins. P63041 protein is shown in strike-through red: it is rejected for presenting a contradictory result for condition HH (with -2S|+2S).

3.4.2 Gene ontology enrichment

To better understand the biological meaning of the changes in the proteins levels under study, we have performed an enrichment analysis. Using the ClueGO plug-in under the Citoscape software, we have performed the enrichment using the Gene Ontology-Biological Process and GOA annotation, version 08.04.2016_08h58.

The analysis was performed using a two-sided hypergeometric test and the Bonferroni step-down correction for multiple tests. The set of 99 proteins where HH and/or HHI significantly changed with respect to the controls was used. Moreover, the proteins were separated into two different, but with some proteins in common, groups: proteins altered in HH (Cluster#1, 73 proteins), and proteins altered in HHI (Cluster#2 59). Of these, respectively 12 and 10 proteins were not used by the software, because of lacking the protein-to-gene mapping or not being annotated into the GOA database. A total of 54 genes (37 differentially expressed in HH and 36 in HH) were found enriched in 20 different Biological Process ontology terms. The results obtained in the enrichment analysis are shown in Figure 3.8 and Table 3.6. The different enriched functions, 20 in first place, are grouped by ClueGO in a functionally grouped annotation network that reflects the relationships between the terms based on the similarity of their associated genes. The degree of connectivity between terms (and therefore the establishment of a functional group) is calculated using kappa statistics, in a similar way as described in W Huang et. al. (42) .

This analysis offers 20 different biological processes enriched. The criteria here is allowing into the analysis biological processes that are enriched combining the genes for both conditions (HH and HHI), with a threshold that we have selected to be $p<0.001$. After the analysis, we have post-processed the data in two steps:

First, we have calculated the individual enrichment of the two conditions (HH and HHI) for each of the 20 enriched biological processes. To do this, we have used the Bonferroni correction factor for each one of the biological processes (from the ratio between the corrected and the uncorrected Pvalues) and knowing the total amount of annotated genes used in the calculations (16,468), and the genes in each GO term, have used the cumulative hypergeometric distribution to obtain the individual Pvalues and, later, used the Bonferroni step-down correction factor to obtain the corrected Pvalues, that are used in the main text for inferring the differential enrichment of the different biological processes.

NAME	Model	EntrezID	Aliases
Abat	Both	81632	Gabat, beta-AlaAT
Acat1	HH	25014	RATACAL
Aldoa	HH	24189	Aldo1, RNALDOG5
Ap2b1	HHI	140670	
Atp1a1	Both	24211	Nkaa1b
Atp1a2	Both	24212	RATATPA2
Atp1a3	Both	24213	Atpa1a3
Atp5g1	Both	29754	
Atp5h	Both	641434	Atp5jd, Atpq
Atp5j2	HH	690441	
Atp5l	Both	300677	
Atp6v1g2	HH	368044	ATP6G, NG38
Atpif1	Both	25392	Atpi, IF1PA
Basp1	Both	64160	NAP22
Calb1	Both	83839	CaBP28K
Calm2	HH	50663	Cam
Camk2a	Both	25400	PK2CDD, PKCCD
Camk2b	Both	24245	Ck2b
Car2	HHI	54231	Ca2
Cltb	Both	116561	
Cnp	HHI	25275	CNPF, CNPI, Cnp1
COX2	HHI	26198	COII
Cplx2	Both	116657	
Dpp6	HH	29272	DPP VI
Gap43	Both	29423	Basp2
Gfap	HH	24387	
Glud1	HH	24399	Gdh1, Glud,
Gpi	HHI	292804	Amf, Nlk, Pgi
Gpm6a	HHI	306439	M6a
Hk1	HHI	25058	HEXOKIN
Ina	HH	24503	Inexa, Intlaa
LOC688869	HH	688869	Cox6b1
Mag	HHI	29409	
Marcks	Both	25603	KINC, Macs
Mbp	HHI	24547	Mbps
Ndufs4	HH	499529	Aqdq
Omg	HH	450224	
Pgam1	HHI	24642	PGAM-B, Pgmut
Ppp3ca	HH	24674	Calna1
Rab10	Both	50993	Ac1075
S100b	HHI	25742	S100P
Slc1a2	HHI	29482	Eaat2, Glt, Glt-1
Slc1a3	Both	29483	EAAT1, GLAST
Slc25a4	HHI	85333	Ant1
Slc25a5	HHI	25176	Ant2
Sncb	HH	113893	
Stx1a	Both	116470	
Syp	HH	24804	Syp1
Syt1	HHI	25716	P65
Tpi1	HH	24849	Tpi
Uqcrb	HH	362897	Uqcrbl
Vamp2	HHI	24803	RATVAMPB
Vdac1	HHI	83529	
Vsnl1	HH	24877	Nvp1, Ratnvp1

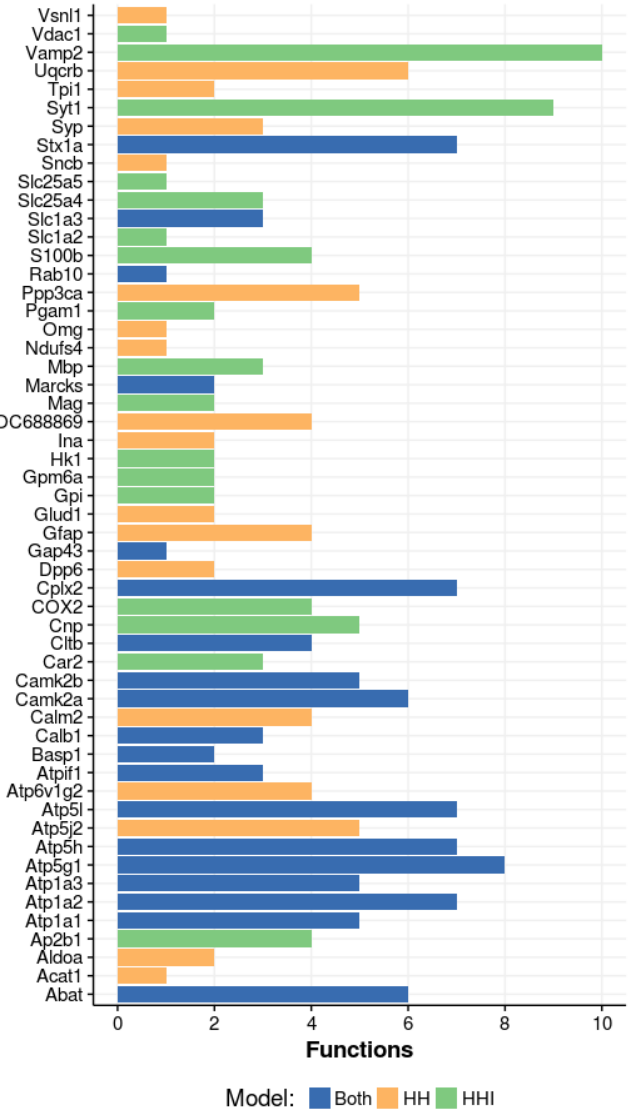


Figure 3.8: The 54 genes found in enriched functions in this analysis. The bar chart shows the number of functions in which each gene is included. A set of 20 GO function has been selected by the software as being enriched in the analysis. Some proteins appear in only one GO biological process (e.g. Acat1) and one (Vamp2) is included in 10. The table details, for each gene, the "cluster origin" (if the gene is differentially expressed in HH and/or HHI), the Entrez Gen ID mapping and the different aliases for each gene. For convenience, the alias Cox6b1 has been used for LOC688869 in the analysis of the enrichment results.

Secondly, we have merged the 20 biological processes into 7 functional groups, in order to better organize and analyze the data. These groups are: Proton, Hydrogen transmembrane and Inorganic cation transport, Brain development, Regulation of mitochondrial membrane permeability, ATP metabolic process, Substantia nigra development, Synaptic transmission, Signal release and Neurotransmitter secretion and Neuron projection morphogenesis. The data obtained in these post-processing steps of the enrichment information can be found in the Table 3.7**.**

GO ID	GOTerm	Nr. Genes	**PValue**	Corr PValue	Associated Genes Found	Genes Cluster HH	Genes Cluster HHI
GO:0006754	ATP biosynthetic process	4	6.3E-05	1.1E-03	Atp5g1, Atp5h, Atp5j2, Atp5l	Atp5g1, Atp5h, Atp5j2, Atp5l	Atp5g1, Atp5h, Atp5l
GO:0007268	synaptic transmission	18	1.4E-10	4.6E-09	Abat, Ap2b1, Calb1, Camk2a, Camk2b, Car2, Cltb, Cplx2, Gfap, Ppp3ca, S100b, Slc1a3, Sncb, Stx1a, Syp, Syt1, Vamp2, Vdac1	Abat, Calb1, Camk2a, Camk2b, Cltb, Cplx2, Gfap, Ppp3ca, Slc1a3, Sncb, Stx1a, Syp	Abat, Ap2b1, Calb1, Camk2a, Camk2b, Car2, Cltb, Cplx2, S100b, Slc1a3, Stx1a, Syt1, Vamp2, Vdac1
GO:0007269	neurotransmitter secretion	7	4.8E-06	1.1E-04	Ap2b1, Camk2a, Cltb, Cplx2, Stx1a, Syt1, Vamp2	Camk2a, Cltb, Cplx2, Stx1a	Ap2b1, Camk2a, Cltb, Cplx2, Stx1a, Syt1, Vamp2
GO:0007420	brain development	13	4.3E-05	7.8E-04	Abat, Acat1, Basp1, Calm2, Cnp, Glud1, Ina, Mag, Marcks, Mbp, Ndufs4, Slc1a2, Syt1	Abat, Acat1, Basp1, Calm2, Glud1, Ina, Marcks, Ndufs4	Abat, Basp1, Cnp, Mag, Marcks, Mbp, Slc1a2, Syt1
GO:0009117	nucleotide metabolic process	14	4.5E-07	1.2E-05	Aldoa, Atp1a2, Atp5g1, Atp5h, Atp5j2, Atp5l, Atpif1, Calm2, Cnp, Gpi, Hk1, Pgam1, Tpi1, Uqcrb	Aldoa, Atp1a2, Atp5g1, Atp5h, Atp5j2, Atp5l, Atpif1, Calm2, Tpi1, Uqcrb	Atp1a2, Atp5g1, Atp5h, Atp5l, Atpif1, Cnp, Gpi, Hk1, Pgam1
GO:0015672	monovalent inorganic cation transport	15	5.5E-08	1.5E-06	Abat, Atp1a1, Atp1a2, Atp1a3, Atp5g1, Atp5h, Atp5j2, Atp5l, Atp6v1g2, COX2, Dpp6, LOC688869, Slc25a4, Uqcrb, Vamp2	Abat, Atp1a1, Atp1a2, Atp1a3, Atp5g1, Atp5h, Atp5j2, Atp5l, Atp6v1g2, Dpp6, LOC688869, Uqcrb	Abat, Atp1a1, Atp1a2, Atp1a3, Atp5g1, Atp5h, Atp5l, COX2, Slc25a4, Vamp2
GO:0015991	ATP hydrolysis coupled proton transport	4	2.3E-05	4.6E-04	Atp1a1, Atp1a2, Atp1a3, Atp5g1	Atp1a1, Atp1a2, Atp1a3, Atp5g1	Atp1a1, Atp1a2, Atp1a3, Atp5g1
GO:0015992	proton transport	12	5.9E-12	2.0E-10	Atp1a1, Atp1a2, Atp1a3, Atp5g1, Atp5h, Atp5j2, Atp5l, Atp6v1g2, COX2, LOC688869, Slc25a4, Uqcrb	Atp1a1, Atp1a2, Atp1a3, Atp5g1, Atp5h, Atp5j2, Atp5l, Atp6v1g2, LOC688869, Uqcrb	Atp1a1, Atp1a2, Atp1a3, Atp5g1, Atp5h, Atp5l, COX2, Slc25a4
GO:0017156	calcium ion regulated exocytosis	5	3.2E-04	4.8E-03	Cplx2, Ppp3ca, Stx1a, Syt1, Vamp2	Cplx2, Ppp3ca, Stx1a	Cplx2, Stx1a, Syt1, Vamp2
GO:0021762	substantia nigra development	6	2.1E-08	6.0E-07	Basp1, Calm2, Cnp, Ina, Mag, Mbp	Basp1, Calm2, Ina	Basp1, Cnp, Mag, Mbp
GO:0023061	signal release	11	7.4E-06	1.5E-04	Abat, Ap2b1, Camk2a, Cltb, Cplx2, Glud1, Ppp3ca, Stx1a, Syt1, Vamp2, Vsnl1	Abat, Camk2a, Cltb, Cplx2, Glud1, Ppp3ca, Stx1a, Vsnl1	Abat, Ap2b1, Camk2a, Cltb, Cplx2, Stx1a, Syt1, Vamp2
GO:0046034	ATP metabolic process	12	1.5E-10	4.7E-09	Aldoa, Atp1a2, Atp5g1, Atp5h, Atp5j2, Atp5l, Atpif1, Gpi, Hk1, Pgam1, Tpi1, Uqcrb	Aldoa, Atp1a2, Atp5g1, Atp5h, Atp5j2, Atp5l, Atpif1, Tpi1, Uqcrb	Atp1a2, Atp5g1, Atp5h, Atp5l, Atpif1, Gpi, Hk1, Pgam1
GO:0046902	regulation of mitochondrial membrane permeability	5	1.9E-06	4.3E-05	Atpif1, Camk2a, Cnp, Slc25a4, Slc25a5	Atpif1, Camk2a	Atpif1, Camk2a, Cnp, Slc25a4, Slc25a5
GO:0048167	regulation of synaptic plasticity	8	1.6E-06	3.8E-05	Calb1, Camk2a, Camk2b, Cplx2, Gfap, S100b, Syp, Vamp2	Calb1, Camk2a, Camk2b, Cplx2, Gfap, Syp	Calb1, Camk2a, Camk2b, Cplx2, S100b, Vamp2
GO:0048489	synaptic vesicle transport	6	2.6E-05	5.0E-04	Ap2b1, Cltb, Cplx2, Stx1a, Syt1, Vamp2	Cltb, Cplx2, Stx1a	Ap2b1, Cltb, Cplx2, Stx1a, Syt1, Vamp2
GO:0048812	neuron projection morphogenesis	10	2.2E-04	3.5E-03	Camk2b, Cnp, Gap43, Gpm6a, Marcks, Mbp, Omg, Ppp3ca, Rab10, Syt1	Camk2b, Gap43, Marcks, Omg, Ppp3ca, Rab10	Camk2b, Cnp, Gap43, Gpm6a, Marcks, Mbp, Rab10, Syt1
GO:0050804	modulation of synaptic transmission	14	7.8E-10	2.4E-08	Abat, Calb1, Camk2a, Camk2b, Car2, Cplx2, Gfap, Ppp3ca, S100b, Slc1a3, Stx1a, Syp, Syt1, Vamp2	Abat, Calb1, Camk2a, Camk2b, Cplx2, Gfap, Ppp3ca, Slc1a3, Stx1a, Syp	Abat, Calb1, Camk2a, Camk2b, Car2, Cplx2, S100b, Slc1a3, Stx1a, Syt1, Vamp2
GO:0050806	positive regulation of synaptic transmission	9	3.2E-08	9.0E-07	Abat, Camk2b, Car2, Gfap, S100b, Slc1a3, Stx1a, Syt1, Vamp2	Abat, Camk2b, Gfap, Slc1a3, Stx1a	Abat, Camk2b, Car2, S100b, Slc1a3, Stx1a, Syt1, Vamp2
GO:0098662	inorganic cation transmembrane transport	14	1.1E-06	2.8E-05	Atp1a1, Atp1a2, Atp1a3, Atp5g1, Atp5h, Atp5l, Atp6v1g2, COX2, Calm2, Dpp6, Gpm6a, LOC688869, Uqcrb, Vamp2	Atp1a1, Atp1a2, Atp1a3, Atp5g1, Atp5h, Atp5l, Atp6v1g2, Calm2, Dpp6, LOC688869, Uqcrb	Atp1a1, Atp1a2, Atp1a3, Atp5g1, Atp5h, Atp5l, COX2, Gpm6a, Vamp2
GO:1902600	hydrogen ion transmembrane transport	10	2.1E-10	6.6E-09	Atp1a1, Atp1a2, Atp1a3, Atp5g1, Atp5h, Atp5l, Atp6v1g2, COX2, LOC688869, Uqcrb	Atp1a1, Atp1a2, Atp1a3, Atp5g1, Atp5h, Atp5l, Atp6v1g2, LOC688869, Uqcrb	Atp1a1, Atp1a2, Atp1a3, Atp5g1, Atp5h, Atp5l, COX2

Table 3.6: The 20 Gene Ontology Biological Processes enriched in this analysis. The different columns refer to: (1) GO ID - unique identifier in the ontology , (2) GO Term - description of the term, (3) Nr. Genes - number of genes in the study included in the biological process, (4) % Associated Genes - ratio of genes in the study associated with the process with respect to all genes associated to that function, (5) Pvalue - probability obtained with a two-sided hypergeometric test, (6) Corr Pvalue – result of the Bonferroni step-down correction for multiple tests, (7) Associated Genes Found – genes found in the biological process under the two conditions under study (HH and HHI), globally, (8) Genes Cluster HH – genes found for HH, (9) Genes Cluster HHI – genes found for HHI.

As shown in Supp. Analysis Table 6, both conditions (HH or HHI) may present the same GO term significantly enriched ($p<0.001$). In this case, the condition with the lowest Pvalue will be chosen as the condition predominantly enriched for this GO term. In the case that both conditions present the same Pvalue, this condition will be labeled as HH/HHI enriched. Each one of the seven possible Functional groups will present also one condition specially enriched; the condition will be chosen attending to the conditions that their individual GO terms present. It is clear that, in almost all functional groups, genes (proteins) that have been expressed differentially with respect to the control in both conditions, will be found at the same time, as it is very usual finding genes included in several GO terms.

GO ID	GO term	Functional group	# genes				Bonferroni correction	HH&HHI Pval	HH&HHI corr Pval	HH Pval	HH Corr Pval	HHI Pval	HHI Corr Pval	GO term - Condition	Functional group - Condition
			GOA	HH&HHI	HH	HHI									
GO:1902600	hydrogen ion transmembrane transport	Proton, Hydrogen transmembrane and Inorganic cation transport	117	10	9	7	31	**2.1E-10**	**6.6E-09**	**4.6E-09**	**1.4E-07**	**1.4E-06**	**4.5E-05**	**HH**	**HH**
GO:0098662	inorganic cation transmembrane transport		613	14	11	9	25	**1.1E-06**	**2.8E-05**	**1.6E-04**	**3.9E-03**	2.6E-03	6.5E-02	**HH**	
GO:0015672	monovalent inorganic cation transport		559	15	12	10	27	**5.5E-08**	**1.5E-06**	**1.3E-05**	**3.6E-04**	**3.3E-04**	**8.9E-03**	**HH**	
GO:0015992	proton transport		149	12	10	8	34	**5.9E-12**	**2.0E-10**	**2.3E-09**	**7.8E-08**	**5.6E-07**	**1.9E-05**	**HH**	
GO:0017156	calcium ion regulated exocytosis	Synaptic transmission, Signal release and Neuro-transmitter secretion	124	5	3	4	15	**3.2E-04**	**4.8E-03**	2.2E-02	3.3E-01	3.0E-03	4.5E-02	**HHI**	**HHI**
GO:0050804	modulation of synaptic transmission		342	14	10	11	30	**7.8E-10**	**2.4E-08**	**5.3E-06**	**1.6E-04**	**6.8E-07**	**2.0E-05**	**HHI**	
GO:0007269	neurotransmitter secretion		140	7	4	7	22	**4.8E-06**	**1.1E-04**	4.6E-03	1.0E-01	**4.8E-06**	**1.1E-04**	**HHI**	
GO:0050806	positive regulation of synaptic transmission		146	9	5	8	28	**3.2E-08**	**9.0E-07**	**6.8E-04**	1.9E-02	**4.8E-07**	**1.4E-05**	**HHI**	
GO:0048167	regulation of synaptic plasticity		171	8	6	6	24	**1.6E-06**	**3.8E-05**	**1.7E-04**	**4.1E-03**	**1.7E-04**	**4.1E-03**	**HHI**	
GO:0023061	signal release		438	11	8	8	21	**7.4E-06**	**1.5E-04**	1.2E-03	2.5E-02	1.2E-03	2.5E-02	**HHI**	
GO:0007268	synaptic transmission		567	18	12	14	33	**1.4E-10**	**4.6E-09**	**1.5E-05**	**5.1E-04**	**4.4E-07**	**1.5E-05**	**HHI**	
GO:0048489	synaptic vesicle transport		122	6	3	6	19	**2.6E-05**	**5.0E-04**	2.1E-02	4.0E-01	**2.6E-05**	**5.0E-04**	**HHI**	
GO:0048812	neuron projection morphogenesis	Neuron projection morphogenesis	532	10	6	8	16	**2.2E-04**	**3.5E-03**	4.2E-02	6.8E-01	3.9E-03	6.3E-02	HHI	HHI
GO:0046902	regulation of mitochondrial membrane permeability	Regulation of mitochondrial membrane permeability	43	5	2	5	23	**1.9E-06**	**4.3E-05**	1.8E-02	4.2E-01	**1.9E-06**	**4.3E-05**	**HHI**	**HHI**
GO:0021762	substantia nigra development	Substantia nigra development	37	6	3	4	29	**2.1E-08**	**6.0E-07**	**7.3E-04**	2.1E-02	**2.9E-05**	**8.3E-04**	**HHI**	**HHI**
GO:0006754	ATP biosynthetic process	ATP metabolic process	45	4	4	3	17	**6.3E-05**	**1.1E-03**	**6.3E-05**	**1.1E-03**	1.3E-03	2.2E-02	**HH**	**HH**
GO:0015991	ATP hydrolysis coupled proton transport		35	4	4	4	20	**2.3E-05**	**4.6E-04**	**2.3E-05**	**4.6E-04**	**2.3E-05**	**4.6E-04**	**HH/HHI**	
GO:0046034	ATP metabolic process		196	12	9	8	32	**1.5E-10**	**4.7E-09**	**4.0E-07**	**1.3E-05**	**4.4E-06**	**1.4E-04**	**HH**	
GO:0009117	nucleotide metabolic process		568	14	10	9	26	**4.5E-07**	**1.2E-05**	**3.7E-04**	**9.7E-03**	1.6E-03	4.0E-02	**HH**	
GO:0007420	brain development	Brain development	735	13	8	8	18	**4.3E-05**	**7.8E-04**	2.4E-02	4.4E-01	2.4E-02	4.4E-01	HH/HHI	HH/HHI

Table 3.7: List of the biological processes enriched (GO ID and GO term), their grouping in Functional groups, the number of genes (#genes) and calculation of the Pvalues (and their Bonferroni corrected counterparts: “corr Pval”) for the genes found in both HH and HHI conditions (HH&HHI), only HH and only HHI. The number and ID of genes can be checked in Supp. Analysis Table 6. In bold, probabilities lower than 0.001. The column “GO term - Condition” selects the condition with the lowest probability for each GO term and the column “Functional group – Condition” choses the most abundant condition inside the Functional group.

3.5 Discussion

Of the 99 proteins expressed differentially in HH and/or HHI conditions: 54 of them belong to at least one of the 20 enriched biological processes found. Both hypoxic models present a similar number of differentially expressed proteins (37 and 36 respectively), but with an overall positive expression in HH (22 over-expressed proteins) and negative in HHI (25 under-expressed proteins) (Figure 3.9). The similar set of processes affected both in HH and HHI points to a common aetiology, while the overall inhibitory nature found in HHI is explained by its greater severity in contrast to HH.

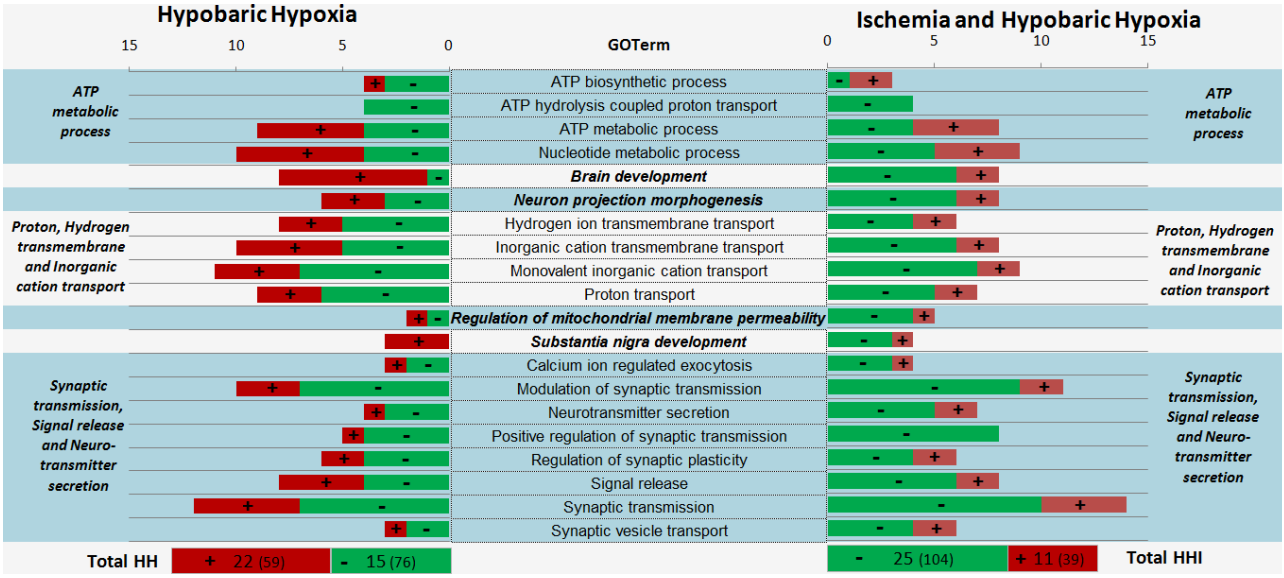


Figure 3.9 The proportion of over/under-expressed proteins in HH (37 proteins) and HHI (36 proteins) is shown for each of the 20 GO biological processes, grouped into seven functional groups (bold). Under the bar chart, the total of over/under expressed proteins (in parentheses the number of times these proteins appear into one biological process), shows a general increase of protein expression in HH (22 protein with increased levels versus 15 decreased) and decrease in HHI (25 decreased versus 11 increased).

The 20 biological processes identified were grouped into seven functional groups attending to the similarity of the processes and genes shared (Figure 3.10):

(i) ATP metabolic process and (ii) Proton, Hydrogen transmembrane and Inorganic cation transport, both showing higher enrichment in HH, present a more down-regulated state in HHI: potassium import across plasma membrane is severely inhibited (Atp1a1, Atp1a3), while calcium exocytosis is also downregulated (Atp1a2 and Vamp2). Furthermore, response to hypoxia (Aldoa) and response to ischemia (HK1) markers show differential expression on their respective conditions.

(iii) Brain development, (iv) Neuron projection morphogenesis and (v) Substantia nigra development present upregulated genes like Gap43, Maks and Basp1, all highly involved in signal transduction pathways, membrane transport and cytoskeletal dynamics . The calmodulin-dependent protein kinases (Camk2a, and Camk2b), that phosphorylate the central bioenergy sensor AMP-activated protein kinase, are downregulated in both HH and HHI; this same tendency is followed by Rab10, a small GTPase acting as regulator of membrane trafficking and fusion also involved in autophagy. Additionally, several proteins related to substantia nigra development (Ina, Calm1, Mbp, Mag and Cnp) show variation in HH and HHI, consistently with previous proteomic studies of changes in Substantia nigra caused by neurodegenerative diseases.

(vi) Synaptic transmission, Signal release and Neuro-transmitter secretion are greatly impaired under HHI, as expected under severe excitotoxic damage; interestingly, the SNARE protein Vamp2, and its regulatory proteins Syt1, both highly involved in glutamate release and neuron damage after ischemic injury, are downregulated but only in HHI.

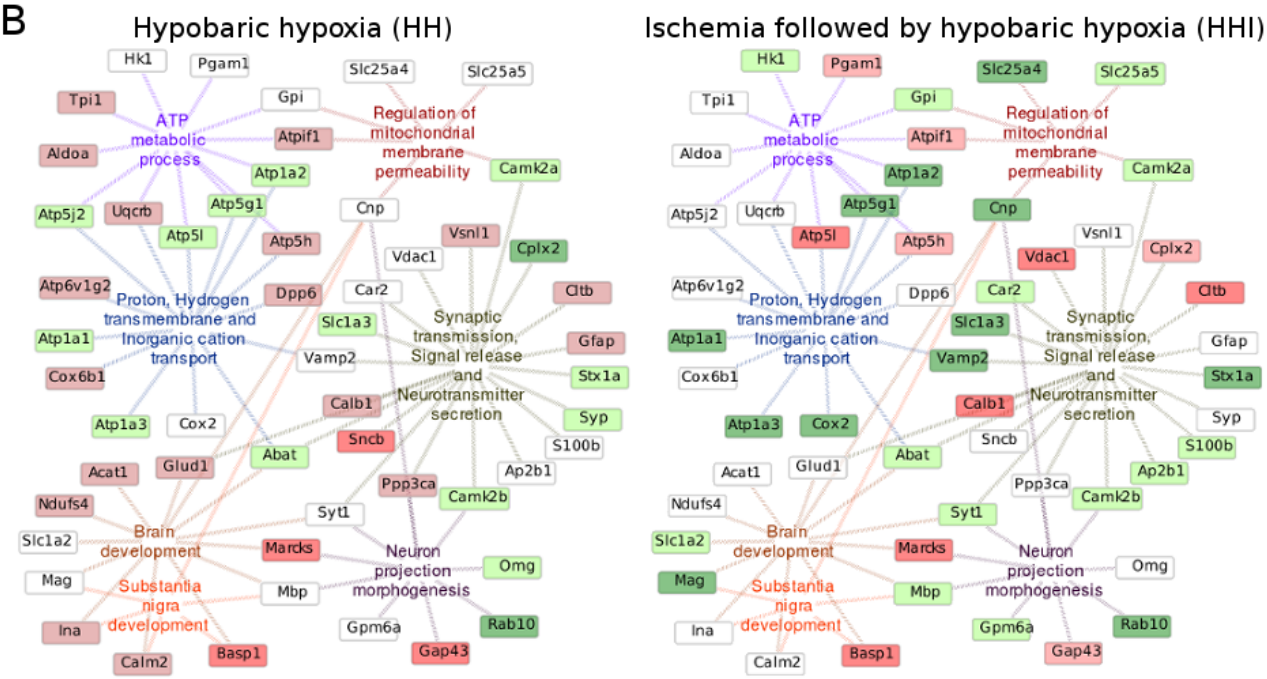
(vii) Regulation of mitochondrial membrane permeability points to the activation of apoptosis through mitochondrial pathways (down-regulation of apoptosis inhibitors Gpi, Slc25a4 Slc25a5 and activation of Atpif1). Components of the mPTP (adenine nucleotide translocator: Slc25a4, Slc25a5 and Vdac1) where also differentially expressed in HH and HHI.

A

GO ID	Functional group	HH Pval	HHI Pval	Condition
GO:0046902	Regulation of mitochondrial membrane permeability	4.2E-01	**4.3E-05**	HHI
GO:0021762	Substantia nigra development	2.1E-02	**8.3E-04**	HHI
GO:0017156	Synaptic transmission, Signal release and Neuro-transmitter secretion	3.3E-01	**4.5E-02**	HHI
GO:0050804		1.6E-04	**2.0E-05**	
GO:0007269		1.0E-01	**1.1E-04**	
GO:0050806		1.9E-02	**1.4E-05**	
GO:0048167		4.1E-03	4.1E-03	
GO:0023061		2.5E-02	2.5E-02	
GO:0007268		5.1E-04	**1.5E-05**	
GO:0048489		4.0E-01	**5.0E-04**	

GO ID	Functional group	HH Pval	HHI Pval	Condition
GO:0048812	Neuron projection morphogenesis	6.8E-01	**6.3E-02**	HHI
GO:0006754	ATP metabolic process	**1.1E-03**	2.2E-02	HH
GO:0015991		4.6E-04	4.6E-04	
GO:0046034		**1.3E-05**	1.4E-04	
GO:0009117		**9.7E-03**	4.0E-02	
GO:1902600	Proton, Hydrogen transmembrane and Inorganic cation transport	**1.4E-07**	4.5E-05	HH
GO:0098662		**3.9E-03**	6.5E-02	
GO:0015672		**3.6E-04**	8.9E-03	
GO:0015992		**7.8E-08**	1.9E-05	
GO:0007420	Brain development	4.4E-01	4.4E-01	HH/HHI

B



C

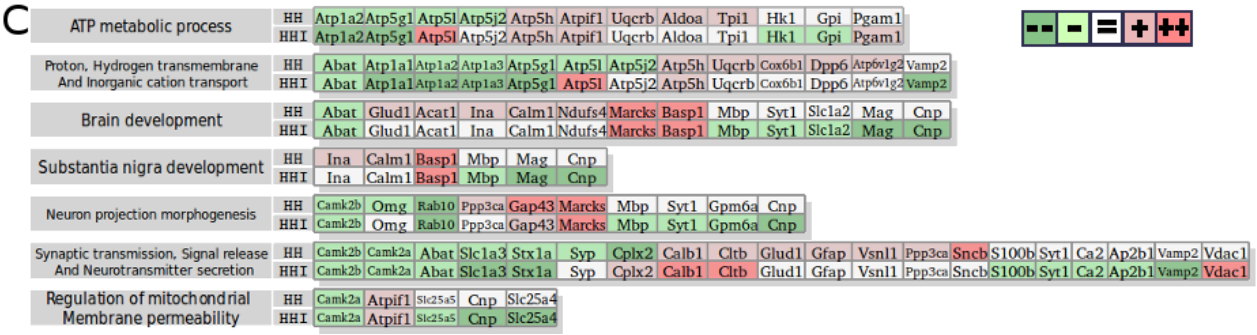


Figure 3.10 Gene enrichment analysis in HH and HHI. (A) Table showing GO terms associated to each functional group, P-values obtained for HH and HHI related genes (in bold the lowest) and condition (HH or HHI) in which the functional group is more enriched. (B) Relationships between functional groups and genes in HH and HHI. Genes are coloured dark and light green for high and moderate evidence of under expression, and dark and light red for high and moderate evidence of over-expression, respectively. (C) For each functional group, using the same legend, the list of genes related to HH and HHI experimental conditions.

3.6 Conclusions

In conclusion, the HHI model presents a global effect of protein downregulation while HH produces an overall increase of the protein levels. With HH mainly affecting oxidative and energetic metabolism, HHI also interferes with synaptic transmission, neurotransmitter secretion, substantia nigra development and triggers apoptosis through mitochondrial pathway.

2.7 References

1. Dugan LL, Choi DW. Hypoxia-Ischemia and Brain infarction. 1999; Available from: https://www.ncbi.nlm.nih.gov/books/NBK28046/

2. Granger DN, Kvietys PR. Reperfusion injury and reactive oxygen species: The evolution of a concept. Redox Biol. 2015 Dec;6:524–551.

3. Aarts MM, Arundine M, Tymianski M. Novel concepts in excitotoxic neurodegeneration after stroke. Expert Rev Mol Med. 2003 Dec;5(30):1–22.

4. Peinado MA, del Moral ML, Esteban FJ, Martínez-Lara E, Siles E, Jiménez A, et al. Aging and neurodegeneration: molecular and cellular bases. Rev Neurol. 2000;31(11):1054–65.

5. Rocha-Ferreira E, Hristova M. Plasticity in the Neonatal Brain following Hypoxic-Ischaemic Injury. Neural Plast. 2016 Mar;2016:1–16.

6. Levine S. Anoxic-Ischemic Encephalopathy in Rats. Am J Pathol. 1960 Sep;36(1):1–17.

7. Rice JE, Vannucci RC, Brierley JB. The influence of immaturity on hypoxic-ischemic brain damage in the rat. Ann Neurol. 1981;9(2):131–141.

8. Basu A, Lazovic J, Krady JK, Mauger DT, Rothstein RP, Smith MB, et al. Interleukin-1 and the interleukin-1 type 1 receptor are essential for the progressive neurodegeneration that ensues subsequent to a mild hypoxic/ischemic injury. J Cereb Blood Flow Metab Off J Int Soc Cereb Blood Flow Metab. 2005;25(1):17–29.

9. Vannucci SJ, Willing LB, Goto S, Alkayed NJ, Brucklacher RM, Wood TL, et al. Experimental stroke in the female diabetic, db/db, mouse. J Cereb Blood Flow Metab Off J Int Soc Cereb Blood Flow Metab. 2001;21(1):52–60.

10. Taylor CF, Paton NW, Lilley KS, Binz P-A, Julian RK, Jones AR, et al. The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol. 2007 Aug;25(8):887–93.

11. Marcotte EM. How do shotgun proteomics algorithms identify proteins? Nat Biotechnol. 2007 Jul;25(7):755–7.

12. Celis JE, Ostergaard M, Jensen NA, Gromova I, Rasmussen HH, Gromov P. Human and mouse proteomic databases: novel resources in the protein universe. FEBS Lett. 1998 Jun 23;430(1–2):64–72.

13. Nesvizhskii AI, Aebersold R. Interpretation of Shotgun Proteomic Data: The Protein Inference Problem. Mol Cell Proteomics. 2005 Oct;4(10):1419–40.

14. Koskinen VR, Emery PA, Creasy DM, Cottrell JS. Hierarchical Clustering of Shotgun Proteomics Data. Mol Cell Proteomics. 2011 Jun 1;10(6):M110.003822.

15. Yates JR, Eng JK, Clauser KR, Burlingame AL. Search of sequence databases with uninterpreted high-energy collision-induced dissociation spectra of peptides. J Am Soc Mass Spectrom. 1996 Nov;7(11):1089–98.

16. Brancia FL, Butt A, Beynon RJ, Hubbard SJ, Gaskell SJ, Oliver SG. A combination of chemical derivatisation and improved bioinformatic tools optimises protein identification for proteomics. Electrophoresis. 2001 Feb;22(3):552–9.

17. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science. 1985 Mar 22;227(4693):1435–41.

18. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 2007 Mar;4(3):207–14.

19. Elias JE, Gygi SP. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. In: Hubbard SJ, Jones AR, editors. Proteome Bioinformatics [Internet]. Totowa, NJ: Humana Press; 2010 [cited 2018 Apr 15]. p. 55–71. Available from: http://link.springer.com/10.1007/978-1-60761-444-9_5

20. Wang G, Wu WW, Zhang Z, Masilamani S, Shen R-F. Decoy Methods for Assessing False Positives and False Discovery Rates in Shotgun Proteomics [Internet]. 2008 [cited 2018 May 13]. Available from: https://pubs.acs.org/doi/abs/10.1021/ac801664q

21. SQLite Home Page [Internet]. [cited 2018 May 10]. Available from: https://www.sqlite.org/index.html

22. Weatherly DB, Atwood JA, Minning TA, Cavola C, Tarleton RL, Orlando R. A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. Mol Cell Proteomics MCP. 2005 Jun;4(6):762–72.

23. Binz P-A, Barkovich R, Beavis RC, Creasy D, Horn DM, Jr RKJ, et al. Guidelines for reporting the use of mass spectrometry informatics in proteomics [Internet]. Nature Biotechnology. 2008 [cited 2018 May 10]. Available from: https://www.nature.com/articles/nbt0808-862

24. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS - Analytical Chemistry (ACS Publications) [Internet]. [cited 2018 May 13]. Available from: https://pubs.acs.org/doi/abs/10.1021/ac0262560

25. Hill EG, Schwacke JH, Comte-Walters S, Slate EH, Oberg AL, Eckel-Passow JE, et al. A Statistical Model for iTRAQ Data Analysis. J Proteome Res. 2008 Aug 1;7(8):3091–101.

26. Herbrich SM, Cole RN, West KP, Schulze K, Yager JD, Groopman JD, et al. Statistical Inference from Multiple iTRAQ Experiments without Using Common Reference Standards. J Proteome Res. 2013 Feb 1;12(2):594–604.

27. Karp NA, Huber W, Sadowski PG, Charles PD, Hester SV, Lilley KS. Addressing Accuracy and Precision Issues in iTRAQ Quantitation. Mol Cell Proteomics. 2010 Sep;9(9):1885–97.

28. Bantscheff M, Boesche M, Eberhard D, Matthieson T, Sweetman G, Kuster B. Robust and Sensitive iTRAQ Quantification on an LTQ Orbitrap Mass Spectrometer. Mol Cell Proteomics. 2008 Sep;7(9):1702–13.

29. Evans C, Noirel J, Ow SY, Salim M, Pereira-Medrano AG, Couto N, et al. An insight into iTRAQ: where do we stand now? Anal Bioanal Chem. 2012 Sep;404(4):1011–27.

30. Zubarev RA. The challenge of the proteome dynamic range and its implications for in-depth proteomics. Proteomics. 2013 Mar;13(5):723–6.

31. Piehowski PD, Petyuk VA, Orton DJ, Xie F, Ramirez-Restrepo M, Engel A, et al. Sources of Technical Variability in Quantitative LC-MS Proteomics: Human Brain Tissue Sample Analysis. J Proteome Res. 2013 May 3;12(5):2128–37.

32. Pappin D. An iTRAQ Primer. http://www.ushupo.org/portals/0/ushupo_techtalk_itraq.pdf. 2010.

33. Emery P, Pappin D. iTRAQ Tips and Tricks ASMS User Meeting. http://www.matrixscience.com/pdf/2010WKSHP2.pdf. 2010.

34. Bauer KM, Watts TN, Buechler S, Hummon AB. Proteomic and functional investigation of the colon cancer relapse-associated genes NOX4 and ITGA3. J Proteome Res. 2014 Nov 7;13(11):4910–8.

35. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005 Oct 25;102(43):15545–50.

36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25–9.

37. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 2017 Jan 4;45(D1):D331–8.

38. Causal analysis approaches in Ingenuity Pathway Analysis [Internet]. [cited 2018 May 16]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3928520/

39. Shannon P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 2003 Nov 1;13(11):2498–504.

40. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics. 2005 Aug 15;21(16):3448–9.

41. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists [Internet]. [cited 2018 May 16]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2644678/

42. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists [Internet]. [cited 2018 May 16]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2375021/

43. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W305-311.

44. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009 Apr 15;25(8):1091–3.

45. Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas. 1960 Apr;20(1):37–46.

Chapter 4. Swath quantification: study of PCOS proteomic biomarkers in plasma

In this chapter, two methods of bioinformatics analysis performed on a proteomics data set for the study of Polycystic Ovary Syndrome (PCOS), using data-independent acquisition mass spectrometry (SWATH), are described. With the results obtained, several bioinformatics and statistical techniques have been used to characterize the phenotypes examined and to assess the different protein levels in them.

4.1 Abstract

The study of the plasma protein levels of twenty women, organized in four phenotypes of lean and obese, diagnosed or not with PCOS, has been performed using data-independent acquisition mass spectrometry. A total of 204 proteins have been quantified. PCOS and obesity present very similar proteomics profiles. Five proteins (FLNA, ADIPOQ, LBP, RBP4 and APOC2) present significant variations between PCOS samples and healthy controls, being RBP4 the most robust marker for PCOS even with interference from obesity. The combination of PCOS and obese phenotypes presents five proteins (ADIPOQ, COLEC11, IGFBP3, SPP2 and IGFALS) down-regulated, as opposed to what happened in lean PCOS subjects.

4.2 Introduction

Polycystic Ovary Syndrome is an hormonal disorder in women of reproductive age, its more evident signs being the presence of cysts in the ovaries, high levels of androgenic hormones and irregular or skipped periods (1–3). PCOS diagnose, based mainly on these three criteria, was systematized in 2003 at the Rotterdam conference on PCOS (4), providing what is known as the “Rotterdam 2003 criteria”. Although recently questioned as insufficient in their prognosis applicability (5), Rotterdam criteria are still the best organized approach to PCOS diagnose. Difficulties in PCOS diagnose have been extensively reported, specially in younger patients (6), and the fundamental basis of PCOS condition is not yet completely understood (7): a set of genetic (PCOM, hyperandrogenemia, insulin resistance, and insulin secretory defects) and environmental factors (prenatal androgen exposure, poor fetal growth and obesity) have been considered. The high number and complexity of the biological pathways involved in PCOS symptomatology (8), add to the challenge of studying its molecular basis with the present knowledge in the area.

In this chapter, a proteomics study of plasma samples of healthy individuals and PCOS patients is performed to unravel the protein signature of PCOS, using also a division of the subjects under study based in their Body mass index (BMI): both lean an obese women have been included in the study under the two categories examined, healthy controls and PCOS patients. The premise here is that the interaction of PCOS with obesity may act as a powerful co-variant and mask the fundamental changes in protein variation among the phenotypes studied.

An introductory section in this chapter (“4.2.1 Phenotypes under study”) will deal with the convenience of dividing the subjects into four different phenotypes (according to the positive or negative diagnose of PCOS and presence or absence of obesity) studying the clinical variables collected for these women. Successfully classifying the subjects into the four different categories described will provide basis for such division.

The technique used for quantitative proteomics is a type of data-independent acquisition known as Swath (9). Bioinformatics analysis of Swath differs substantially from the traditional data-dependent “shotgun” proteomics approach (10,11). The two main platforms for Swath analysis (Skyline (12) and OpenSwath (13)) have been used and further compared with the data generated in this analysis: an overview of the pipelines built and the results obtained is shown in this chapter, alongside with the reasons for the use of OpenSwath, that generated the results that will be discussed here.

And lastly, the proteomics results obtained from the different phenotypes studied are systematically analyzed using different tertiary analysis (14) techniques for giving sense to the different groups of proteins differentially expressed.

4.3 Materials and methods

4.3.1 Phenotypes under study

For this study, plasma samples from a total of 20 females, with ages ranging from 20 to 40 years, are used; ten of them have been diagnosed with PCOS. This cohort, is a subset of a bigger group of study comprising 164 subjects, from whom a comprehensive set of clinical variables has been collected, including biochemical and physiological measures common in clinical practice. In addition of a PCOS diagnose, the subjects have been divided according to their Body mass index (BMI) (15), using a BMI of 30 as a threshold. BMI is calculated dividing weight in kilograms by height in squared meters. Therefore, four different phenotypes will be discussed in this work:

- **HT (healthy-thin)**, five subjects with BMI under 30 and without a PCOS diagnose
- **HO (healthy-obese)**, five subjects with BMI over 30 and without PCOS
- **PT (PCOS-thin)**, five subjects with BMI under 30 and with a PCOS
- **PO (PCOS-obese)**, five subjects with BMI over 30 and with a PCOS

As shown in Figure 1, the BMI of the patients with obesity is in all cases well over 30, whereas the lean patients are also well below that value. In the case of the PT subjects, their BMI values, are in two cases over 25: the relationship of PCOS with obesity (3), has made the task of finding subjects with the pathology and an ideally low BMI difficult.

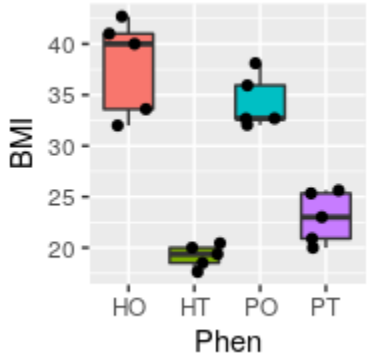


Figure 4.1. Distribution of BMI of the 20 subjects under study grouped by their phenotypes: HT (healthy-thin), HO (healthy-obese), PT (PCOS-thin) and PO (PCOS-obese).

In order to assess the biological validity of the four groups created, previously to the main analysis of this work, a statistical validation of the four groups created is to be performed. The aim of this statistical validation is, using the clinical variables collected, classify the

group of subjects that are going to be analyzed by quantitative proteomics into one of the HT, HO, PT and PO groups. Without embarking into sophisticated machine learning algorithms, the aim of this analysis is to provide some physiological basis to the four groups created: besides the fact that it makes sense to stratify the subjects into these four categories, the clinical variables alone will be able to support this logic.

From the set of 24 initial clinical variables collected (shown in Table 4.1), a subset capable of classifying the subjects into the four groups under study is going to be searched. For this, the complete cohort of 162 subjects is used:

- 20 subjects, analyzed later in this chapter using proteomics, are going to conform a “test subset”, where the variables chosen are tested in their ability to predict the group which they belong.
- The remaining subjects, 142, will conform a “train subset”, where logistic regression is used to extract a model that uses a subset of those variables.

Clinical variables			
hirsutism	menarche	FM	homaindex
cholesterol	LDL	insulin	testosterone
estradiol	thyrotropin	ast	Alt
freeT4	androstenedione	LH.FSF	FSH
HDL	triglycerides	glucose	cortisol
SDHA	hidroxi	LH	prolactina

Table 4.1 The 24 clinical variables collected.

The “train subset” is subjected to binary logistic regression:

- First, a logistic regression is performed taking into consideration two groups: one conformed by HT subjects and the other by the subjects from other groups (i.e. HO, PT and PO together): the binary logistic approach allows only two classes for classification. A model is produced as a linear combination of the variables that best predict this classification with HT versus HO+PT+PO.
- An analysis of variance (ANOVA) is applied on the model, obtaining a P-value that characterizes the performance of each variable in the model.
- Once the model has been built, it is applied over the “train subset”, where each subject is assigned to a group in an interval from 0 to 1, where 0 means complete belonging to the HT group and 1, complete belonging to the second group containing the rest of the groups (HO, PT and PO).
- The performance of the model is tested using a “Receiver operating characteristic” (ROC) curve. The area under the curve is a measure of the performance of the model.
- This process is to be applied likewise with HO versus HT+PT+PO, PT versus HT+HO+PO, PO versus HT+HO+PT and HT+HO versus PT+PO (Healthy versus PCOS).

A detailed report of the approach is provided at "**Appendix 1: Chapter4, Phenotypes inspected**".

The five tables with variables and P-values obtained using ANOVA in each of the five iterations of logistic regressions, are summarized in Table 4.2.

HT ANOVA

Var.	Pr(>Chi)	
waist	3.67e-14	***
FM	4.46e-05	***
testosterone	3.85e-03	**
weight	6.05e-03	**
height	3.58e-03	**
freeT4	2.73e-02	*
LDL	5.62e-04	***
hidroxi	3.44e-02	*
estradiol	1.70e-04	***
hirsutism	2.48e-05	***

HO ANOVA

Var.	Pr(>Chi)	
hip	1.49e-14	***
FM	3.31e-11	***
hirsutism	1.46e-03	**
weight	2.44e-02	*
insulin	7.44e-03	**
height	3.35e-02	*
menarche	7.18e-02	
estradiol	8.06e-02	
glucose	5.21e-02	
HDL	8.85e-02	
waist.hip	6.28e-02	
waist	8.22e-02	
LH	1.23e-01	
homaindex	3.79e-02	*
androstenedione	1.86e-06	***

PT ANOVA

Var.	Pr(>Chi)	
weight	7.45e-10	***
FM	1.32e-07	***
thyrotropin	1.21e-03	**
HDL	6.50e-03	**
hip	1.38e-02	*
LH.FSF	6.81e-02	
prolactina	3.52e-02	*
hirsutism	3.33e-02	*

PO ANOVA

Var.	Pr(>Chi)	
waist.hip	2.46e-09	***
FM	2.67e-07	***
hip	6.22e-04	***
hirsutism	4.68e-05	***
height	8.95e-04	***
LH.FSF	3.04e-02	*
freeT4	3.54e-02	*
thyrotropin	5.51e-02	
ast	1.08e-01	
cholesterol	1.11e-01	
glucose	1.31e-01	

H vs PCOS ANOVA

Var.	Pr(>Chi)	
FM	8.41e-21	***
hirsutism	3.98e-07	***
LH	3.45e-05	***
insulin	1.28e-02	*
height	2.68e-02	*
menarche	2.16e-02	*
homaindex	5.50e-02	
estradiol	1.56e-01	

Table 4.2 Five ANOVA tables obtained from the corresponding linear models obtained by step-wise logistic regression.

In the five iterations, the optimal variables to discern among groups are obtained. From that set of variables, the ones having a P-value inferior to 0.001 are selected: hip, FM, androstenedione, waist, LDL, estradiol, hirsutism, waist.hip, height, weight and LH. And with these variables a Principal Components Analysis (PCA) is implemented. The graphical results are shown in Figure 4.2.

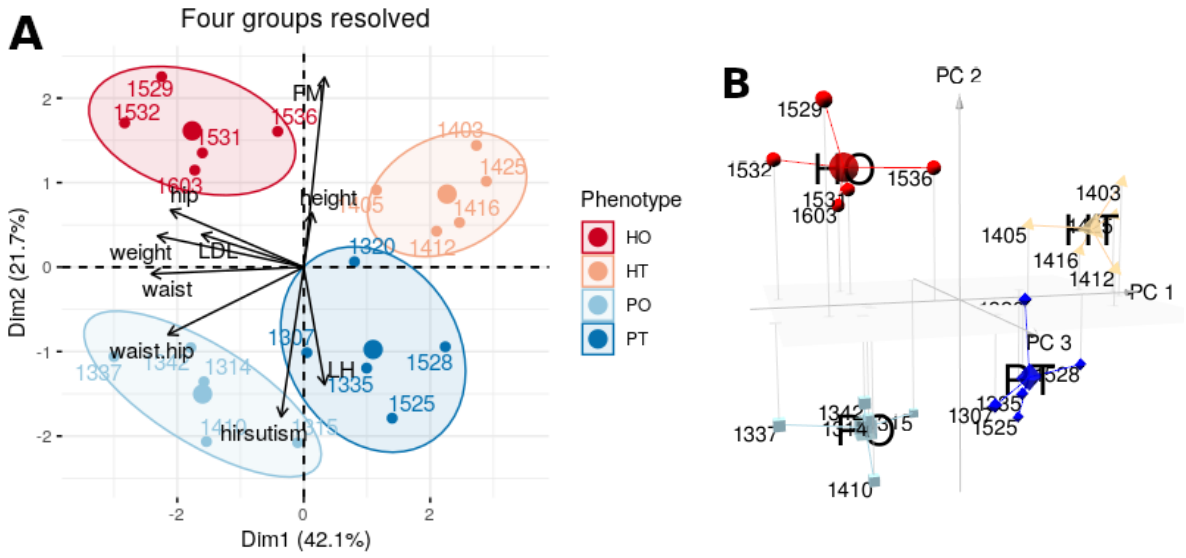


Figure 4.2 (A) PCA biplot (graphical representation of the two principal components) from the PCA) of the eleven selected variables present in the 20 subjects used in the proteomics study. Subjects (identified by numbers as 1337, 1335,...) and variables (waist, FM,...) are projected over the biplot. (B) The three first principal components of the same PCA are used and the subjects

projected in three dimensions.

From observing the two plots displayed in Figure 4.2, it is clear that the four groups have been clearly differentiated using the logistic regression approach described before. More sophisticated methods of machine learning could be applied for this: multinomial logistic regression, neural networks, support vector machines or k-nearest neighbor among them. Furthermore, not previous study of multicollinearity has been used in this approach in order to discard linearly related variables. But this simple approach has allowed to show that classification using a few biochemical and physiological variables is possible for the 20 subjects under study among the four groups created for this work, *QED*.

4.3.2 Mass spectrometry analysis

Plasma samples have been depleted from high abundant albumin and IgG using a column of sepharose based resins (HiTrap Albumin and IgG Depletion, GE Healthcare Life Sciences). Then, non-depleted proteins were concentrated and cleaned by protein precipitation with TCA / acetone and solubilized in 50 µL of 0.2% RapiGest SF (Waters) with 50 mM ammonium bicarbonate. The total protein content was measured using the Qubit Protein Assay Kit (Thermo Fisher Scientific) and 50 µg of protein was subjected to trypsin digestion following a protocol adapted from Vowinckel et al. (16). Retention time reference peptides (iRT peptides, Biognosys) were spiked into each sample.

The proteomics analysis was performed using a TripleTOF 5600+ (Sciex) instrument, using a nano-HPLC NanoSpray III (Sciex) with a sprayer PicoTip Emitter (New Objective) at 2600V spray voltage. The acquisition software used was Analyst (Sciex).

The library used in the study was created analyzing six independent samples created from pools of the 20 samples studied. The acquisition methodology in use consisted in a TOF MS^1 survey scan (350-1250 m/z, 250 ms acquisition time) and a maximum of 65 MS^2 scans (230-1700 m/z, 60 ms acquisition time). Results produced six wiff files used in the next section to build the library.

The 20 subjects under study have been analyzed using a data independent quantitative proteomics approach known as Swath (Sequential Window Acquisition of all Theoretical fragment ions). The Swath (9) technique in the mass spectrometry field refers to an acquisition setup of the mass spectrometer. The objective of the Swath technique is quantifying proteins, generally in a relative quantification mode, where a differential analysis is performed using some type of sample as a reference (e.e. a healthy control), although some examples of absolute quantification can be found (17). Typically, this analysis is performed using the TripleTOF (18) instruments, but it has also been adapted to work with the more advanced models of Orbitrap (19). In this setup (20), ions from a predefined m/z window (called Swath, usually close to 25 m/z) at MS^1 stage (21) are fragmented alongside, producing a highly complex and multiplexed set of MS^2 signals. The mass range where tryptic peptides are expected to be found (400-1200 m/z) is scanned in a 2 to 4 seconds cycle. Then, if using a swath window of 25 m/z and inside a 400-1200 m/z range, 32 “swaths” will be acquired every 2-4 seconds. The element that will allow the interpretation of the highly complex sets of fragmented peptides in each swath is the elaboration, in parallel, of a spectral library acquired in the traditional data dependent acquisition method using the same sample (or pool of samples) under analysis. This spectral library will allow the mapping of the fragmented peptides inside each swath to previously identified peptides. An additional element that allows the unequivocal identification of the peptides is the use of retention time re-alignment techniques using

commercially available reference peptides (22), spiked both in the samples to elaborate the library and the samples to be quantified. Ideally, the instrument and chromatographic column used in the elaboration of the spectral library and the ones used with the samples to quantify should also be the same, although this is not essential.

A total of 50 variable Swath windows were established in this analysis, with a minimum of 9.1 m/z and a maximum of 69.9 m/z, into a mass/charge range going from 399.5 to 1246.9 m/z, suited for detecting the most common peptides produced after protein digestion with trypsin. The use of variable Swath window widths is an approach that ensures that every window contains roughly the same number of precursors after a survey run is analyzed. In the distribution of Swath windows used in this experiment (Figure 4.3), an increase of the widths is observed with higher Swath numbers: that means that a higher number of precursor ions (peptides) is found at a lower m/z range and less peptides are found at higher mass to charge ratios (specially starting at 900 m/z) for a given range. This approach ensures that the same acquisition time is dedicated to each peptide, regardless of the place where is located into the mass to charge scale.

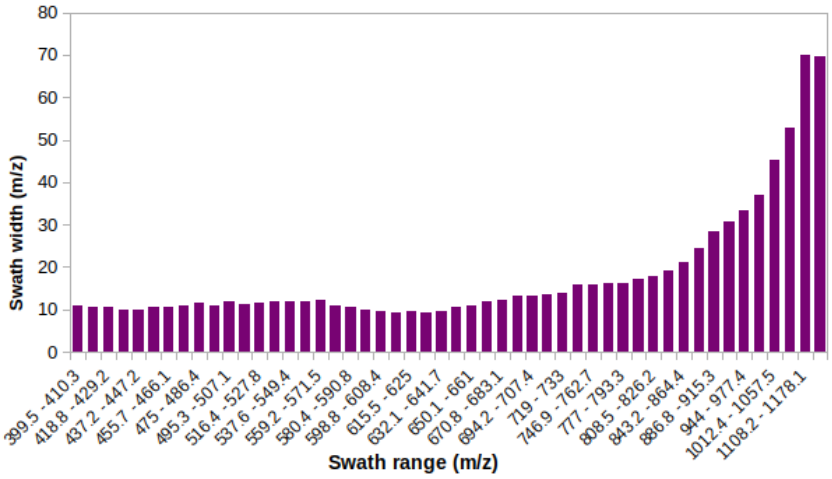


Figure 4.3 Widths of the 50 Swath windows, spanning from 399.5 to 1246.9 m/z.

After the mass spectrometry analysis is done, several files are created:

- Six wiff files acquired in data-dependent mode, corresponding to the libraries.
- 20 wiff files acquired in Swath mode, corresponding to the samples to quantify.

4.3.3 Swath bioinformatics analysis

In contrast with data dependent quantitative approaches, where much proprietary and open source software can be found (23), choices for swath analysis software are more limited. The bioinformatics analysis of a Swath experiment follows a general structure in two steps: generation of the spectral library and swath quantification.

- Elaboration of a spectral library: several samples of the same type that the ones to quantify are analyzed in the mass spectrometer. The bigger the number, the better: these samples are going to be used to identify the peptides that later will be quantified. Typically, a pool of the samples to be analyzed is used for this. One or several runs in the mass spectrometer in data dependent acquisition mode will generate several raw files that will be sequenced using protein database search engines. The identifications will be translated to “transitions”, term originating from

the Selected reaction monitoring mode (24), where only a given ion is selected at MS^1 stage and only another one at MS^2. One or more types of ions will be selected to build this transitions library (with CID or HCD fragmentations, y and b ions, for example).

- After the spectral library is generated, the runs (or samples) generated by the Swath experiments will be analyzed, mapping the transitions stored in the spectral library to the signals obtained from the Swath intensities. The intensities of the mapped ions (mapped to peptides) along retention time are known as extracted ion chromatograms (XIC, Figure 4.4). Several of these XIC will be used (the most intense and most stable alongside samples), using their areas as a measure of the quantity of the peptide in the sample. Finally, like in data dependent acquisition, proteins are constructed using these peptides and a total amount (signal intensity when relative quantification is used) of protein will be calculated for each sample.

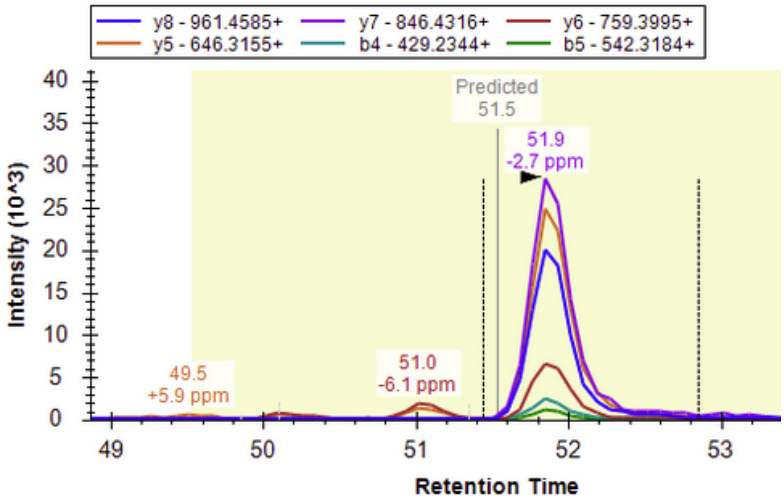


Figure 4.4 Six extracted ion chromatograms (4 y ions and two b ions) generated by one peptide at a given time window (49 to 53 minutes).

The two pipelines more frequently used for Swath analysis to this date are OpenSwath (13) and Skyline (25). Both have been used thoroughly for the analysis of Swath data, being the former pipeline exclusive for Swath analysis and the later used also for Selected reaction monitoring mode for more than ten years. In this work, both have been used to analyze the 20 samples processed and the results obtained have been compared. In both cases, the spectral library has been built using the software Trans-Proteomic Pipeline (26) and two search engines in parallel: Comet (27) and XTandem (28). In the next sections, an overview of the pipelines used with OpenSwath and Skyline will be provided and the reasons that led to choose OpenSwath as the analysis pipeline in this work, explained.

4.3.3.1 OpenSwath analysis

The OpenSwath analysis pipeline uses not only its own tools, but makes extensive use of several other software:

- Proteowizard (29) for conversion of proprietary file formats to open formats.
- Trans-Proteomic pipeline (TPP) for library generation and statistical analysis.
- OpenMS (30) platform for statistical analysis and data conversion and integration.

With the exception of the TPP that can be used through a web interface, the rest of the tools are command line applications, intended for their use in a Linux (31) computer or cluster. In addition, although feasible, the installation process of Proteowizard and the TPP is quite complex: instead, a Docker (32) container has been installed and run in order to use both applications in this pipeline. The method followed in this analysis corresponds to the published material “Building high-quality assay libraries for targeted analysis of SWATH MS data” (33). The complete OpenSwath workflow is included in “**Appendix 2: Chapter4, OpenSwath workflow**”.

An overview of the complete workflow followed with OpenSwath (Figure 4.5) is described in the next points:

- **Library generation**: the analysis of the pools samples produced six raw files from the ABSciex instrument (TTOF) in the form of “wiff” files. These files were converted, using Proteowizard, to the open format mzXML (34). Two parallel searches were performed using Comet and XTandem search engines, with a protein database in fasta format from Uniprot (35) (release April 2019) with 20,417 proteins and one artificial protein containing the Biognosys iRT (22) peptides. Decoy proteins are included too in the database, as reverse sequences of the originals. The search with Comet will produce six pepXML (26) files and also the search with XTandem. Those six files from each search engine will be then combined into two files by the software Interact, producing two other files: one interact.comet.pep.XML and one interact.tandem.pep.XML. Both files will be merged by iProphet to produce one iProphet.combined.pep.XML, with the combined and non-redundant results from Comet and XTandem.

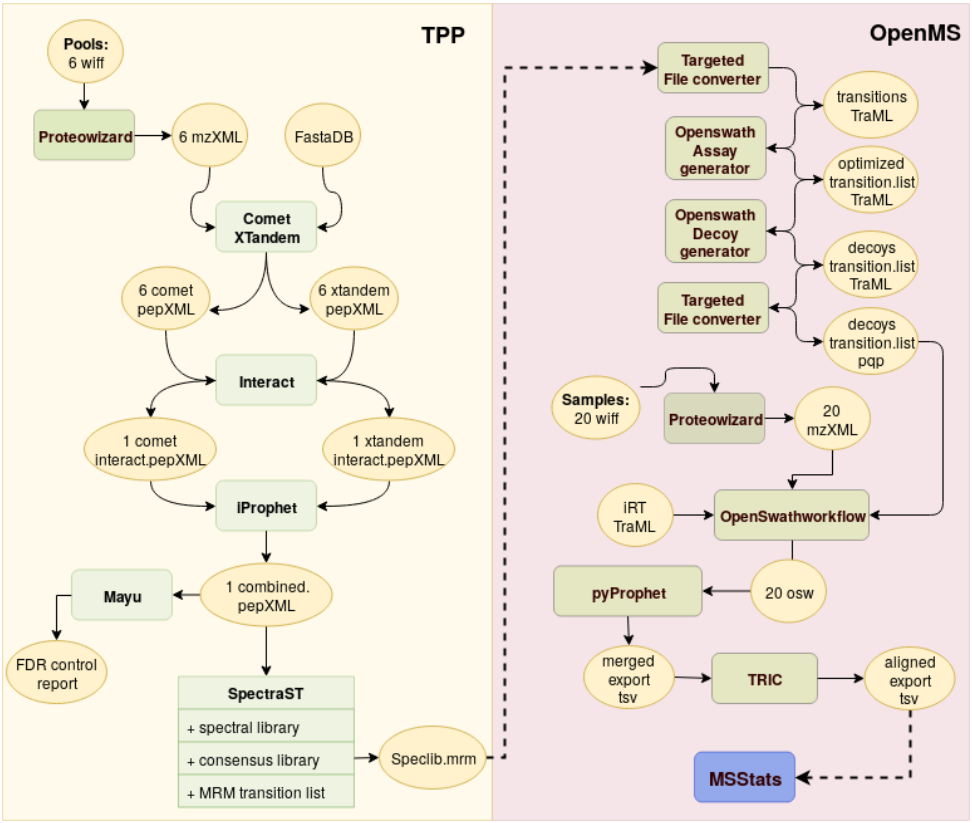


Figure 4.5 Complete workflow with OpenSwath, from library generation to the generation of the file to be used for differential analysis. In yellow circles the files generated and in green squares, the software. All software actions and files generated are enclosed in the Trans-

Proteomic pipeline (TPP) and OpenMS set of tools.

- **FDR control**: the software Mayu (36) is used for False discovery rate control. Choosing a FDR <5% at the protein level, a value of IP/PPs of 0.63 is found at Mayu's report: this is going to be the cut-off value used later to only use confident proteins. Selecting lower cut-offs like 1% seemed too restrictive while further filters are going to be applied later in the pipeline.
- **SpectraST** (37) software is used to convert, in various steps, the iProphet.combined.pep.xml into a spectral library file in Selected reaction monitoring format (SpecLib_pqp.mrm). In this step, the expected locations of the reference peptides (iRT) are used to align the library. Also, the cut-off at protein level obtained by Mayo is applied.
- **OpenMS**: several steps of file conversion, optimization and generation of decoy transitions are applied here. A file containing the Swath windows definition is used here to align the transitions inside their corresponding window. The final library produced (transitionlist_optimized_decoys.pqp) is now ready to be used by OpenSwath to quantify the proteins in the Swath files.
- **Swath samples**: the 20 samples acquired in Swath mode, and also in wiff format, are converted to mzXML format using Proteowizard.
- **OpenSwathWorkflow** software: the 20 samples in mzXML format, the spectral database in pqp format and a transition list of the reference peptides in TraML (38) format (iRT.TraML) are fed to this software to produce 20 osw files with the transitions mapped to peptides from the spectral database.
- **pyprophet**: this software will merge and apply some strict cut-offs to the peptides and proteins identified (q-value<0.1 at peak group level, q-value<0.05 to peptides and q-value<0.01 to proteins)
- **TRIC** (39): this software performs cross-run alignment; cut-offs 0.05 have been applied both at FDR level and with alignment score. An aligned.export.tsv file is produced, containing a total of 6,645 transition groups (peptides).
- **Normalization** and **Differential expression analysis**: using the Bioconductor (40) package SWATH2stats (41), the data exported by TRIC is transformed to a format that can be used by the MSStats (42) software. Then, using this transformed data and an additional text file mapping the files to the different phenotypes (HT, HO, PT, PO), MSstats processes the data organizing the peptides into the different proteins and samples, normalizes the signals among samples (Figure 4.6) and performs differential analysis (discussed in the section "4.4.1 Differential analysis, general overview"). Once the contaminants (iRT synthetic protein included in the fasta database and four keratines: K2C1_HUMAN, K1C10_HUMAN, K22E_HUMAN, K1C9_HUMAN) have been removed, a total of 204 proteins are quantified.

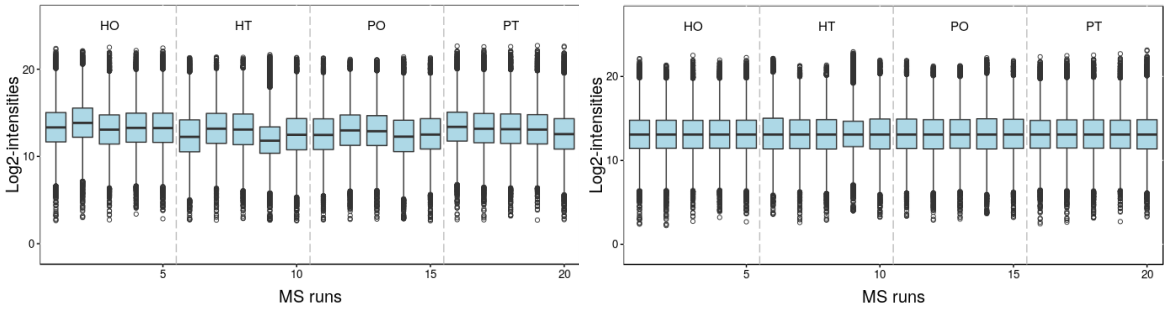


Figure 4.6 Signal normalization by MSStats. Log2 intensities of the 20 files analyzed, grouped by phenotype, prior to normalization (left) and normalized (right).

4.3.3.2 Skyline analysis

The Skyline application is a Windows desktop, open source software that has been used for targeted proteomics for more than ten years. More recently, it has been adapted to Swath analysis. The analysis methodology used here follows the online materials published on the DIA/SWATH Course organized by the Institute of Molecular Systems Biology, ETH Zürich (http://dia-swath-course.ethz.ch/). An overview of the complete workflow followed with Skyline is described in the next points:

- Data-independent acquisition settings are set to a maximum of 2000 m/z.
- The isolation scheme (Swath windows) is imported from one of the Swath wiff files.
- The transition parameters of the eleven iRT peptides (with a total of 66 transitions) are imported using a tab-separated file.
- A spectral library is creating by importing the one created at the OpenSwath pipeline: the file iProphet.combined.pep.XML created by iProphet by merging the results obtained by Comet and XTandem is imported by skyline.
- A protein database in fasta format (without decoy proteins) is imported as the targets to be used in the quantification. Then, decoy proteins are generated inside Skyline.
- Then, the 20 mzXML files (converted by Proteowizard from the raw wiff files) corresponding to the samples to be quantified, are imported.
- Once imported, the files are annotated, that is, the phenotype of each one (HT, HO, PT, PO) is introduced in the workflow. Results can be visualized graphically inside the application (Figure 4.7).

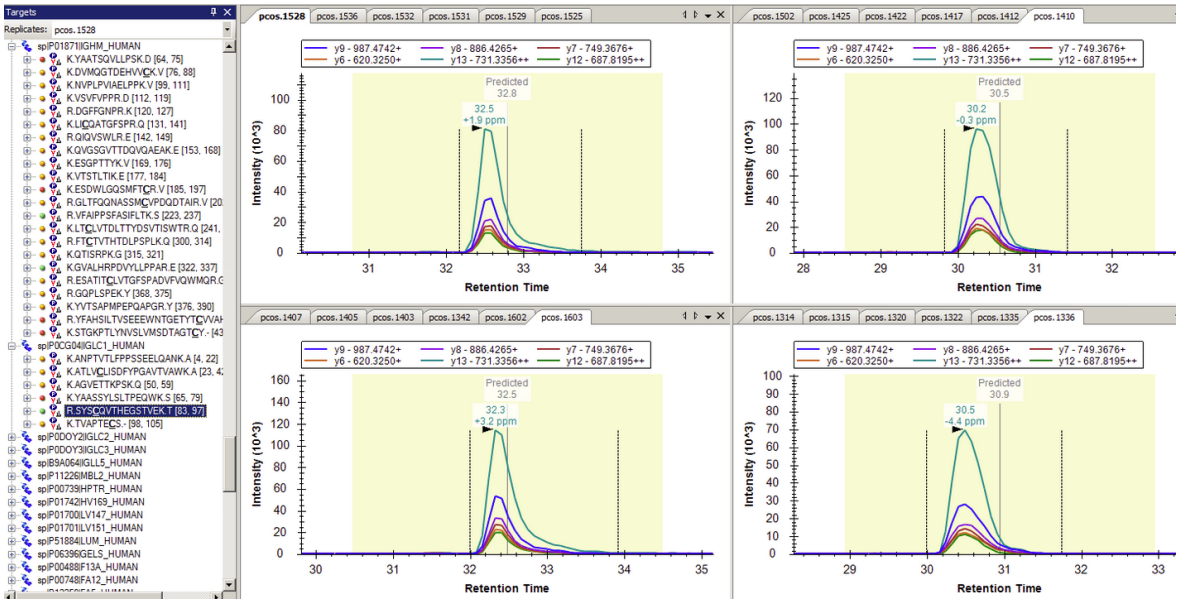


Figure 4.7 Skyline with samples analyzed. On the left, the list of proteins and peptides that have been quantified. On the right, extracted ion chromatograms (XIC) of the transitions corresponding to one peptide selected in the left bar, showing in this example four different samples. The gray bar with the "predicted" caption on each of the four windows represents the predicted retention time for the peptide using the reference (iRT) peptides.

- Using the MSStats plugin that can be directly installed from Skyline, results are exported to a text file. A total of 2,540 transition groups has been exported. The exported file has to be manually edited to remove duplicates (same transition reported twice for the same file).
- Now in the R environment, we import the transitions generated by Skyline using a MSStats function (SkylinetoMSstatsFormat) using a q-value_cut-off of 0.01. Data now is processed by MSStats, normalized (results are visually very similar to the ones generated by the OpenSwath pipeline, shown in Figure 4.5) and a differential expression analysis is performed after including a text file mapping each file to its respective phenotype. Once the contaminants are removed, a total of 209 proteins have been quantified.

4.3.3.3 Skyline/OpenSwath: choosing one approach

As seen in the previous two sections, the two workflows differ greatly in complexity and informatics knowledge required: OpenSwath is far more complex than Skyline, and takes much longer to set up. Moreover, Skyline offers a great graphical interface where users can inspect selected proteins and peptides to get a detailed view on how transitions are quantified. On the other hand, OpenSwath allows a higher control of the procedures done in the pipeline, using a set of filters and adjusts specifically designed for Swath analysis, and this, is reflected in the results obtained: in Figure 4.8, significant proteins (P-values under 0.05) are compared using OpenSwath and Skyline results for four of the possible comparisons. In all cases, OpenSwath gives more proteins as differentially expressed. It is certain that this is not an argument in favor of lack of performance against Skyline, with the samples (and the settings) used in this analysis. But not disposing of reliable standard samples analyzed using Swath, selecting the pipeline that performs better is maybe the best thing to do.

Another, and more objective argument in favor of OpenSwath is the number of peptides quantified by both pipelines: 2,540 transition groups by Skyline and 6,645 transition groups by OpenSwath. The difference is so high that having missed something in the Skyline pipeline appears as a probable explanation. Precisely, this is another argument against Skyline: the amount of information for Swath analysis (published in papers and tutorials) is quite larger in the case of OpenSwath.

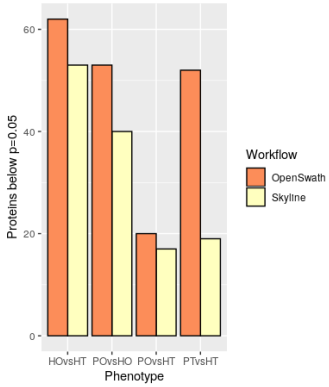


Figure 4.8 Skyline and OpenSwath significant results (P-val<0.05) for HOvsHT, POvsHO, POvsHT and PTvsHT. In all cases, OpenSwath outperforms Skyline.

In Figure 4.9, the correlation plots of Log2 Fold Changes obtained by OpenSwath and Skyline (without P-value filtering) in four different differential analyses are shown. Proteins with higher residuals are marked. In some cases, like CFHR4 for PTvsHT, in Skyline we obtain a log2FC of 1 and -0.5 for OpenSwath. Values of 0 and -2 respectively are obtained for the same protein in HOvsHT. The number of samples with CFHR4 quantified is in both cases almost complete: 19 samples with signal for this protein. The explanation for the differences in log2FC obtained here is that a different number of transitions have been used for the two pipelines. Further analysis of the causes of the differences and tuning the parameters used within Skyline is needed to arrive to a well-founded explanation of these differences.

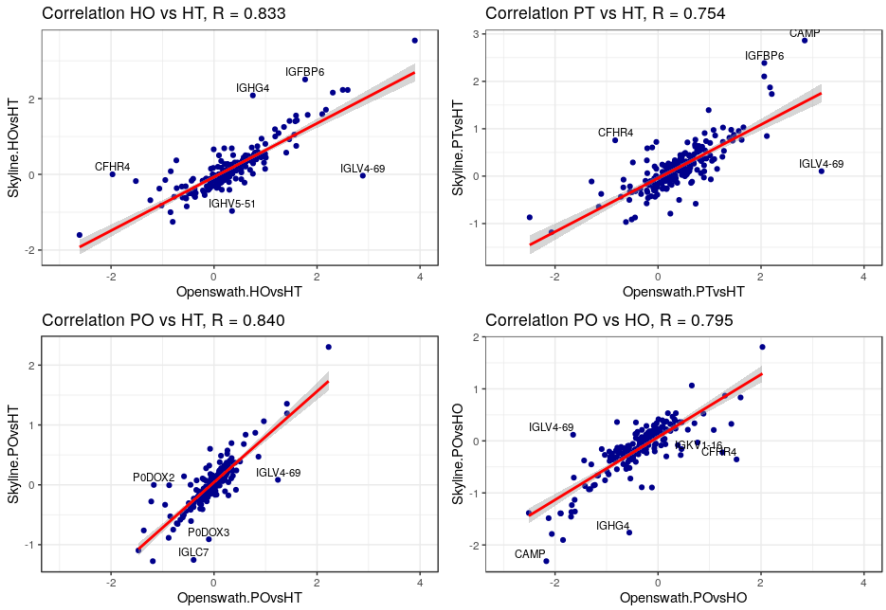


Figure 4.9 Correlation of Log2 Fold Changes obtained by OpenSwath and Skyline (without P-value filtering) in four different differential analyses. Proteins with higher

residuals are highlighted.

The fact that OpenSwath is specifically designed for Swath analysis, while Skyline is not, was the last argument that made us to stop trying parameter optimization with Skyline and proceed with the analysis of the data in this work using OpenSwath.

4.4 Results

4.4.1 Differential analysis, general overview

A study of differential analysis has been realized for different combinations of the samples phenotypes using MSStats (42) Bioconductor package. Eight comparisons have been made (Figure 4.10), keeping the “healthier” state in the denominator (healthy over PCOS, thin over obese):

1. PCOS vs H (diseased versus healthy samples): PT and PO (ten PCOS samples, thin and obese) are compared to HT and HO (ten healthy samples, thin and obese).
2. PCOS vs HT: ten disease (PCOS thin and PCOS obese) samples versus five healthy thin samples (HT).
3. HO vs HT: five healthy obese samples compared to five healthy thin samples.
4. PT vs HT: five PCOS thin samples versus five healthy thin samples.
5. PO vs HT: five PCOS obese samples compared to five healthy thin samples.
6. PO vs HO: five PCOS obese samples compared to five healthy obese samples.
7. PT vs HO: five PCOS thin samples versus five healthy obese samples.
8. PO vs PT: five PCOS obese samples compared to five PCOS thin samples.

A detailed study on each of the comparisons is going to be performed in this section.

For each comparison, MSStats provides an evaluation of the ratio (in log 2 scale) between the groups and a measure of probability, performing a t test for equal means (43), the null hypothesis being that the two groups are actually the same: the lower the probability, the more likely the two groups are different. MSStats also provides an adjusted P-value (44) using Benjamini and Hochberg (45) algorithm. In this work, only the direct P-value, without correction, will be used: the low power (i.e. low number of samples) generally used in proteomics (46) and specifically in this work, makes of adjusted P-value a too harsh filter for the results obtained in the comparisons being made. The distributions obtained from these comparisons are shown in Figure 4.10. The cut-off used in this work for selecting a protein as differentially expressed will a P-value<0.05 and a Log2 fold change (actually a ratio) lower than -0.26 or higher to 0.26. The cut-off value 0.05 for probability is widely used, whereas there is not a universally used cut-off value for fold change:

- some publications use and arbitrary value of fold change: for example, a fold change expressed as ratios (fold change ≥1.5 or ≤0.67, or expressed as FC=1.5 and FC=2/3 for increase and decrease) to make them match to a log2 fold change of ±0.58 (47)
- in other papers, not fold change cut-off is applied at all, although usually this approach is associated to using an adjusted p-value 0.05 as the probability cut-off (48)

- due to the high variability and low abundance of proteins found at plasma (a few hundreds instead of several thousands found at tissue o cultures studies), more relaxed fold change cut-offs have been found in plasma related studies: not cutt-off at all (49), allowing increments as low as the 12% (using p-value<0.05, not adjusted), a 10% ratio increase/decrease (using adjusted p-value) (50).

Then, the cut-off selected for this work will be a p-value of 0.05 and a log2 fold change of 0.26; this represents a compromise between highly strict cut-offs employed in experiments where thousands of proteins are at stake, and the completely permissive approaches where only a P-value is used to assess significant differences between samples. A Log2 fold change of 0.26 represents, roughly, a ±20% minimal expression difference between groups (a +20% increase and -16% decrease) for a given protein.

No missing value imputation (51) has been used in this differential analysis: absent proteins in a given sample will be considered as being below the quantification level.

For each of the eight comparisons, differential expression will be evaluated in to ways:

- A ratio (expressed as a log 2 fold change) obtained from the division of the intensities (mass spectrometry signal) of a group with respect to reference group and a t-test probability of these groups being the same
- Individual intensities obtained from each sample will be used to build a hierarchical cluster, where proteins and groups will group freely according to their intensities; these clusters will also provide visual aid in detecting artifacts. Log-transformed individual intensities obtained from the different samples were scaled and then clustering was performed using the package “Pheatmap” (52) from Bioconductor, using the Ward.D method (dissimilarities are squared and then, Ward's (1963) clustering criterion is applied), both with rows (samples) and with columns (proteins).

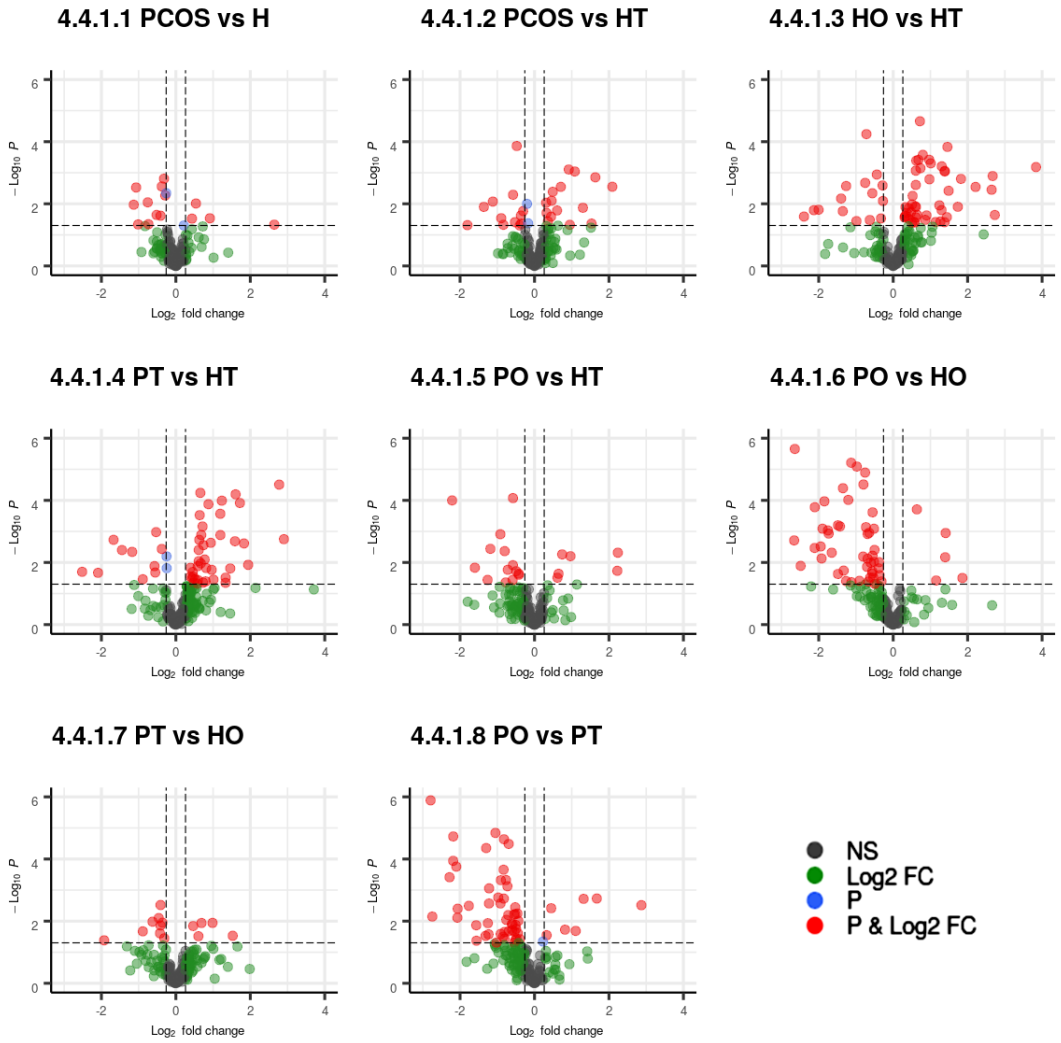


Figure 4.10 Graphical representation of the results obtained by differential expression of the eight comparisons made. Two limits are applied: a log2 Fold change higher of 0.2 or lower than -0.2 and a P-value lower than 0.05. Proteins passing both cut-offs are represented in red. Proteins failing both are represented in black, and for those failing only one, in blue (Fold change) or green (P-value). All eight comparisons comprise 204 proteins, and are here represented using the same scale, for comparison's sake.

The number of samples in which the proteins are found is also an important factor (Figure 4.11). For example, two proteins (IGHG3 and C1QA) appear in nine samples, not being quantified in the other eleven samples. The fact that a sample appears as "quantified" for a given protein depends on several factors, mainly both quality (of the mass spectra obtained) and amount of peptides that compose the protein: thresholds applied at a peptide (or transition) level make possible that, even if a protein is there, is discarded on behalf of the low quality of its constituent peptides. One factor that favors a low peptide score is a very low concentration: the lower the concentration, the less peptides will be correctly detected. In this scenario, as the minimum number of samples for these two proteins is nine, it is then possible that for a given comparison where ten samples are evaluated, those proteins will still be present at nine of the samples. For this reason, proteins will be discarded for this reason (being present in too few samples) only for each comparison, not globally. Only when it becomes clear that for a given comparison (e.g. HO

vs HT) a protein is not found in enough samples for a robust quantification, this protein will be discarded.

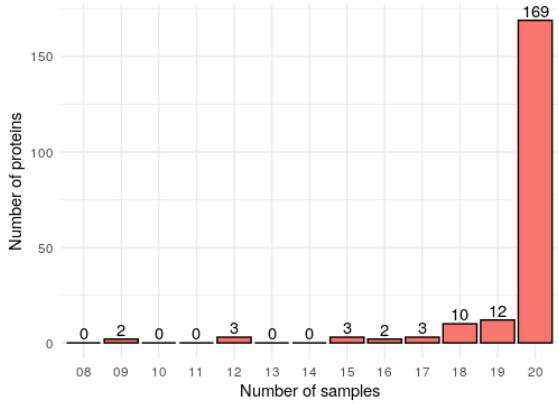


Figure 4.11 Number of proteins found at a given number of samples. The number of proteins that are found in all samples (20 samples) is 169 proteins.

Another possible scenario is the case when, for a given phenotype, one protein is present in all the samples and for the other phenotype (because is below the quantitation limit) the same protein is not detected in some of the samples, presenting in the other samples very low individual concentration levels. In this case, that protein will not be discarded and the ratio (very likely a big one) still considered. This case can be explained as one protein being literally absent in one phenotype and consistently present in the other. Keeping or discarding proteins for a given comparison will be justified in each case.

Finally, for each differential analysis performed, an enrichment analysis will be done in two steps:

1. A differential analysis for Gene Ontology (53) (Biological Process sub-ontology) will be performed using JEPETTO (54), a Cytoscape (55) plugin that performs integrative human gene set analysis, also allowing visualization of interaction networks formed by the enriched terms (Figure 4.12). Using in this work an association threshold equal to 1, a coverage threshold of 0.3 and a triangle threshold of 0.1, the software uses a network-based association score (XD-score) to select significantly enriched GO terms.
2. A second enrichment will be performed used Toppgene (56), using terms of less than 200 genes and a probability density function (instead of cumulative distribution), and showing adjusted P-values using Bonferroni (57) adjustment, more strict than Benjamini & Hochberg and less than Benjamini & Hochberg & Yekutieli (58). The enrichment will be done for all categories, and only those considered of interest for each comparison will be included here.

The results for those enrichments will then be added to each comparison for helping in the interpretation of the groups of proteins found in each of them. It is important to point out that the absence of a GO Biological Process in one comparison does not mean that this process is not taking place: it just means that the number of proteins is not enough to present a significant enrichment. Sometimes, the presence or absence of a single protein will make that one category is present or not for one comparison.

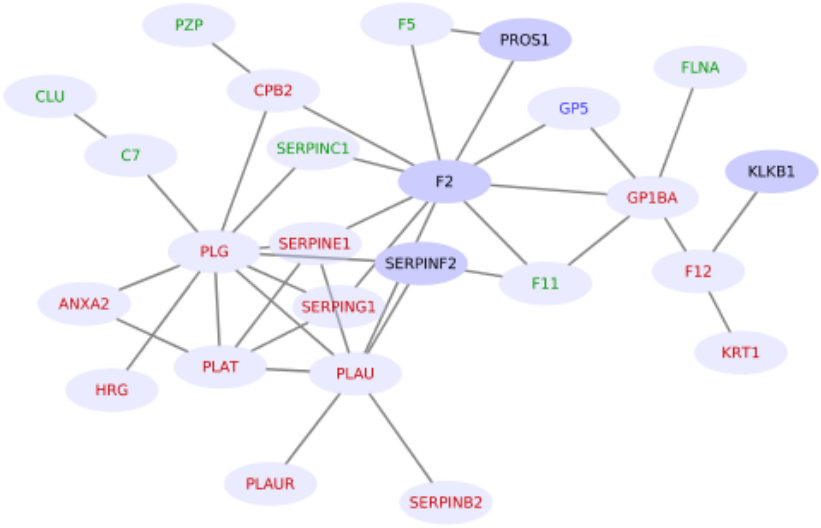


Figure 4.12 JEPETTO example network. The interaction network used is String (59) network. The four highlighted entries (SERPINF2, F2, PROS1 and KLKB1) are proteins in the sample matching the GO category (fibrinolysis in this example). Proteins with text in red are those belonging to the GO category but not in the sample. Those with green text are proteins in the sample but not in the GO category. And the ones in blue(only GP5) are added terms as linking nodes for other interactions.

Then, in the next eight points, two notably interesting comparisons (PCOS vs Healthy and PCOS vs HT) and the six possible comparisons of the four phenotypes studied (HO vs HT, PT vs HT, PO vs HT, PO vs HO, PT vs HO and PO vs PT) are going to be analyzed systematically, using hierarchical clustering and several types of enrichment to get some insight in the kind of changes in protein levels that the different phenotypes undergo.

4.4.1.1 PCOS vs H

Healthy samples (HO and HT) are compared to the ones diagnosed with PCOS (PT and PO), independently if they come from an obese subject or not: **PCOS vs H** comparison, using healthy samples as reference. Here, protein levels in a set of ten samples from healthy subjects are compared to the protein levels in ten samples diagnosed with PCOS. In Figure 4.13, the list of 14 proteins differentially expressed obtained after applying the cut-offs; a hierarchical cluster using the individual intensities has been built. From the cluster, two aspects appear clear:

- The cluster does not produce a perfect grouping of the four phenotypes in display, with a HT phenotype (1425) misplaced among HO.
- The result obtained with protein IGHG3, although appearing with a very high fold change, is likely an artifact. Only nine samples out of the 20 possible have quantified for IGHG3 (five PCOS and four healthy). This protein will be discarded from this comparison: showing a very high expression in one PT sample, while not being detected in the other four, makes that these expression changes surely mean an individual variation, and are not group-related.

Changes in PCOS vs H are mainly negative, with twelve of the proteins under-expressed in PCOS with respect to the healthy samples. Also, in both up and down regulated proteins, changes do not appear to be drastic: after removal of IGHG3, the maximum down-regulation comes from FLNA (-1.13) and the higher up-regulation from APOC2 (0.92).

PCOS vs H

		log2FC	Pval
FLNA	Filamin-A	-1.13	1.07E-02
ADIPOQ	Adiponectin	-1.07	2.98E-03
THBS1	Thrombospondin-1 (Glycoprotein G)	-1.01	4.62E-02
PFN1	Profilin-1	-0.76	8.96E-03
FERMT3	Fermitin family homolog 3	-0.74	4.53E-02
COLEC11	Collectin-11	-0.52	2.28E-02
IGFBP3	Insulin-like growth factor-binding protein 3	-0.41	2.42E-02
C8B	Complement component C8 beta chain	-0.38	2.75E-03
C8G	Complement component C8 gamma chain	-0.32	1.55E-03
C8A	Complement component C8 alpha chain	-0.29	5.39E-03
LBP	Lipopolysaccharide-binding protein (LBP)	0.43	3.00E-02
RBP4	Retinol-binding protein 4	0.54	9.78E-03
APOC2	Apolipoprotein C-II	0.91	2.93E-02
IGHG3	Immunoglobulin heavy constant gamma 3	2.64	4.69E-02

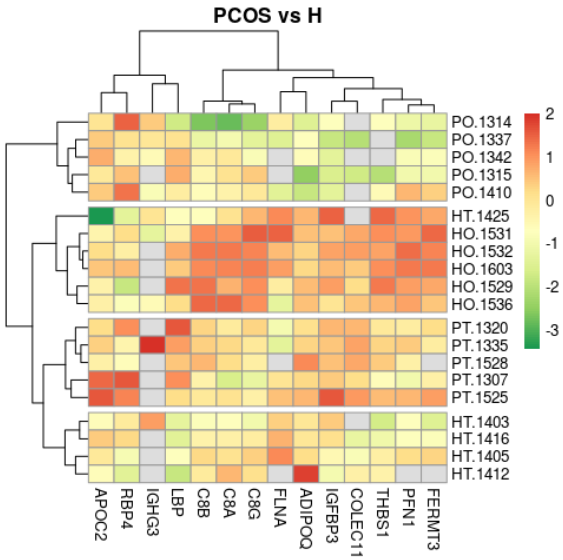


Figure 4.13 Sixteen differentially expressed proteins in PCOS vs Healthy comparison: list with the log2 Fold Changes and P-values (left) and a hierarchical cluster of the individual Log intensities (right), clustering both for samples (rows) and proteins (columns), gray squares representing non-quantified proteins.

Once removed (IGHG3), the resultant volcano plot, with labeled proteins is shown in Figure 4.14. As displayed in the volcano plot, both changes in fold change and levels in probability are not very high. And special caution must be employed with proteins close to P-value (THBS1 and FERMT3) and fold change (C8A, C8B, C8G) thresholds.

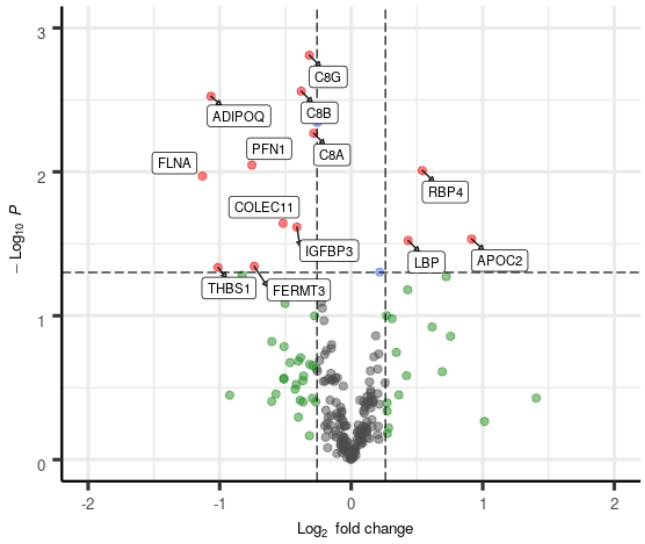


Figure 4.14 Volcano plot of the PCOS vs Healthy comparison. Thirteen differentially expressed proteins in red and labeled. Proteins failing for both fold cut-offs are represented in black, and those failing only one, in blue (Fold change) or green (P-value).

The enrichment study of PCOS vs H is shown at Table 4.3. Both alternative and classical pathways appear represented, while the other terms have been introduced mainly because of interaction of its members with the aforementioned groups. The low P-values here are caused by the low number of differentially expressed proteins (thirteen): only complement activation categories show consistent enrichment values. As it will be seen in

the next comparison, PCOS vs HT, the low number of differentially proteins in this caused by interference from HO samples: while HO and PT share many aspects in common, differences between healthy and PCOS samples, the levels of the former being somewhat diluted by the presence of HO samples.

PCOS vs H

GO Term	XD-score	q-value	n/N
complement activation, alternative pathway	1.199	0.00E+00	4/11
reverse cholesterol transport	0.806	9.02E-01	1/11
complement activation	0.803	1.00E-05	4/22
cytolysis	0.796	0.00E+00	4/16
positive regulation of tumor necrosis factor biosynthetic process	0.788	1.59E-02	2/10
complement activation, classical pathway	0.742	0.00E+00	5/29
positive regulation of intrinsic apoptotic signaling pathway	0.631	9.02E-01	1/14

Term	Description	q-value	n	N	Genes
Pubmed:18492757	High plasma retinol binding protein-4 and low plasma adiponectin concentrations are associated with severity of glucose intolerance in women with previous gestational diabetes mellitus.	5.95E-04	2	2	ADIPOQ, RBP4
Pubmed:17686833	Effect of weight loss on LDL and HDL kinetics in the metabolic syndrome: associations with changes in plasma retinol-binding protein-4 and adiponectin levels.	5.95E-04	2	2	ADIPOQ, RBP4
Pubmed:18445670	Serum levels of retinol-binding protein 4 and adiponectin in women with polycystic ovary syndrome: associations with visceral fat but no evidence for fat mass-independent effects on pathogenesis in this condition.	5.95E-04	2	2	ADIPOQ, RBP4
Pubmed:23864804	Adipokines, insulin-like growth factor binding protein-3 levels, and insulin sensitivity in women with polycystic ovary syndrome.	1.79E-03	2	3	ADIPOQ, IGFBP3
C0030167	Pachymeningitis	1.70E-03	2	3	
C2936179	Obesity, Visceral	6.31E-03	3	44	

Table 4.3 Enriched terms for PCOS vs H. On the left, GO Biological Process terms analyzed with JEPETTO. On the right, Pubmed entries and Disease terms (DisGeNET) from Toppgene.

4.4.1.2 PCOS vs HT

The ten samples diagnosed with PCOS (PT and PO) are compared to the five healthy thin (HT) samples used as reference. In this comparison (**PCOS vs HT**), healthy obese samples are not included, in order to simplify the model and have a true “healthy control” in use.

In Figure 4.15, the list of 27 deferentially expressed proteins and the hierarchical cluster performed, where all samples have been correctly classified using three groups (splitting into three the rows corresponding to the most separate branches of the dendogram). In the dendogram two clear groups of proteins also appear when splitting the two first branches of the dendogram: different color patterns, contributing to an easier visual classification. Surprisingly, proteins within the PT and HT groups appear as the more opposite in the cluster, while proteins in the PO group present intensities that follow the trends in PT group but with less intensity. This fact appears to be contrary to the expectation of finding more severe changes in PO than in PT with respect to healthy controls (HT).

The overall tendency of the proteins differentially expressed in PCOS vs HT is the over-expression, with only IGHM clearly under the -1 log2 fold change (and a P-value just above the cut-off).

PCOS vs HT

		log2FC	Pval
IGHM	Immunoglobulin heavy constant mu	-1.80	4.86E-02
FLNA	Filamin-A	-1.36	1.26E-02
ADIPOQ	Adiponectin	-1.12	8.49E-03
GPX3	Glutathione peroxidase 3	-0.89	2.91E-02
SAA4	Serum amyloid A-4 protein	-0.84	4.74E-02
FBLN1	Fibulin-1 (FIBL-1)	-0.58	5.12E-03
APOD	Apolipoprotein D	-0.53	3.91E-02
C7	Complement component C7	-0.48	1.38E-04
SERPINA1	Alpha-1-antitrypsin	-0.40	4.98E-02
GSN	Gelsolin	-0.36	2.39E-02
A2M	Alpha-2-macroglobulin	-0.36	4.34E-02
F2	Prothrombin (Coagulation factor II)	-0.31	1.72E-02
F11	Coagulation factor XI (FXI)	0.31	9.24E-03
CLU	Clusterin	0.32	1.97E-02
KLKB1	Plasma kallikrein (Fletcher factor)	0.36	3.69E-02
PROS1	Vitamin K-dependent protein S	0.44	2.59E-02
CD14	Monocyte differentiation antigen CD14	0.46	7.99E-03
F5	Coagulation factor V	0.49	4.10E-03
RBP4	Retinol-binding protein 4	0.61	1.64E-02
APCS	Serum amyloid P-component (SAP)	0.71	2.85E-03
LBP	Lipopolysaccharide-binding protein (LBP)	0.92	7.86E-04
PZP	Pregnancy zone protein	0.95	4.73E-02
PRG4	Proteoglycan 4 (Lubricin)	1.08	9.10E-04
APOC2	Apolipoprotein C-II	1.30	1.34E-02
IGLV4-69	Immunoglobulin lambda variable 4-69	1.53	4.33E-02
CAMP	Cathelicidin antimicrobial peptide	1.63	1.41E-03
LPA	Apolipoprotein(a)	2.09	2.82E-03

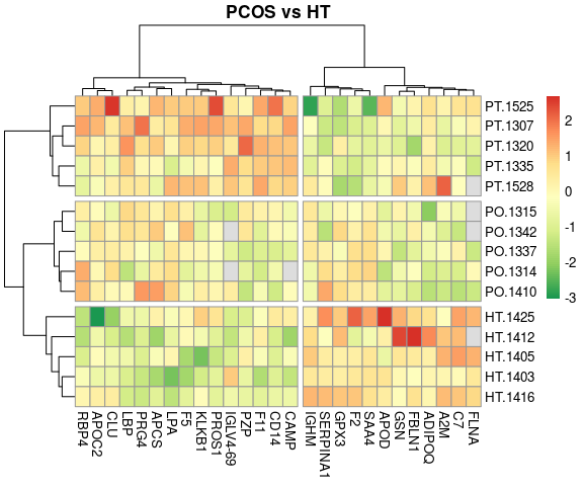


Figure 4.15 27 differentially expressed proteins in PCOS (PT+PO) vs Healthy Thin (HT) comparison: list with the log2 Fold Changes and P-values (left) and a hierarchical cluster of the individual Log intensities (right), clustering both samples (rows) and proteins (columns).

In Figure 4.16 (A), the volcano plot corresponding to PCOS (PT+PO) vs Healthy Thin (HT) proteins has been represented. From the 27 proteins differentially expressed, five are shared with those found as also differentially expressed in the previous section (PCOS vs Healthy (H)) and highlighted in yellow in the plot: FLNA, ADIPOQ, LBP, RBP4 and APOC2.

Interestingly, besides we have taken out the healthy obese (HO) samples, proteins passing the cutoff in the two comparisons, present similar fold changes: in Figure 4.16 (B), fold changes for the two comparisons (the one in this point, PCOS vs HT, and the one in the previous point, PCOS vs H) are shown. The five proteins that are differentially expressed in both (P-value <0.05), are labeled with a star: their fold changes are quite similar, although in the case of PCOS vs HT, the fold changes are slightly more extreme: if the protein is up-regulated in both, PCOS vs HT has a higher fold change than PCOS vs H, and if down-regulated, lower. The fact that introducing HO patients decreases differences within those five proteins, but still preserves the differences, suggests that changes related to the presence of FLNA, ADIPOQ, LBP, RBP4 and APOC2 are also present in both PT and HO, but more intensely in the first case.

The effects of PCOS become clearer in this comparison than in the previous (PCOS vs H), very likely because of PT and HO sharing many variations in common. Here is when the five proteins shared in terms of differential expression with PCOS vs H becomes more important: those proteins resist quite well the “dilution” process (in terms of differential expression) conducted by HO samples.

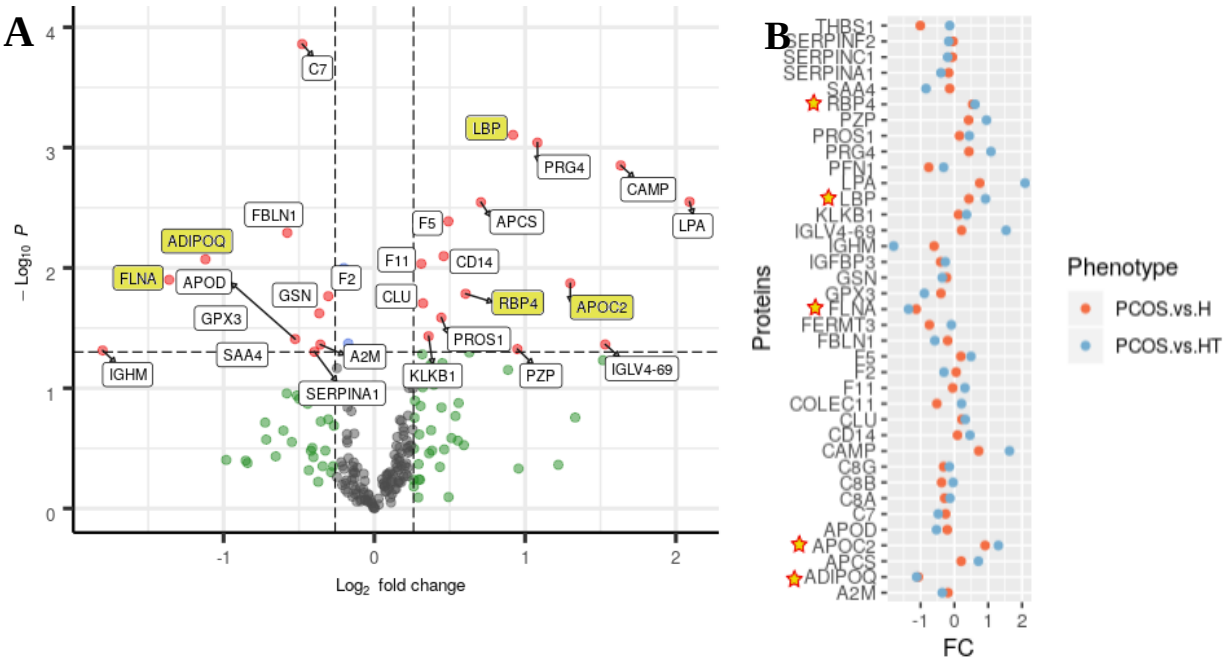


Figure 4.16 (A) Volcano plot of proteins in PCOS (PT+PO) vs Healthy Thin (HT). Labeled those considered differentially expressed. Highlighted in yellow proteins coinciding with differentially expressed in previous section: PCOS vs H. **(B)** Plot representing fold changes for both comparisons (orange PCOS vs H, blue PCOS vs HT) in proteins differentially expressed in at least one them. Proteins labeled with a yellow star are the ones expressed differentially (Pval<0.05) in both comparisons: FLNA, ADIPOQ, LBP, RBP4 and APOC2.

An enrichment study has been performed using the 27 differentially expressed proteins for PCOS vs HT: results are collected in Table 4.4. Enrichment here loses the complement components with two new appearing: coagulation (A2M, SERPINC1, FBLN1, F2, KLKB1 and F11) and acute-phase response (APCS, SERPINC1, SERPINA1, SAA4, LBP, SERPINF2, F2, KLKB1). That does not mean that the complement activation is completely lost here: although not as highly enriched as in the previous comparison, several proteins related are still clearly present here (A2M, APCS, C7, F2, CLU, IGHM and PROS1). Although interactions between coagulation and complement have been described (60), looks like the overall trend in expression changes in PCOS vs HT with respect to PCOS vs H, but the reduced number of proteins in the previous comparison calls for caution in this respect.

PCOS vs HT

GO Term	XD-score	q-value	n/N
negative regulation of fibrinolysis	1.294	3.54E-02	2/10
regulation of blood coagulation	1.283	3.54E-02	2/10
peptidyl-glutamic acid carboxylation	1.283	3.54E-02	2/10
blood coagulation, intrinsic pathway	1.218	2.00E-05	4/17
fibrinolysis	1.155	2.00E-05	4/18
acute-phase response	0.918	0.00E+00	6/31
positive regulation of collagen biosynthetic process	0.911	6.01E-02	2/14
negative regulation of astrocyte differentiation	0.883	8.89E-01	1/10
cytosolic calcium ion homeostasis	0.883	8.89E-01	1/10
positive regulation of blood coagulation	0.883	8.89E-01	1/10
negative regulation of proteolysis	0.733	9.29E-01	1/12

Term	Description	q-value	n	N	Genes
Pubmed:29428584	The preliminary association study of ADIPOQ, RBP4, and BCMO1 variants with polycystic ovary syndrome and with biochemical characteristics in a cohort of Polish women.	1.36E-02	2	3	ADIPOQ, RBP4
Pubmed:12923129	Serum adiponectin levels in women with polycystic ovary syndrome.	1.00E+00	1	1	ADIPOQ
Pubmed:14669168	Polycystic ovary syndrome, the G1691A factor V Leiden mutation, and plasminogen activator inhibitor activity: associations with recurrent pregnancy loss.	4.51E-02	2	5	F2, F5
Pubmed:23415973	Faster thrombin generation in women with polycystic ovary syndrome compared with healthy controls matched for age and body mass index.	1.00E+00	1	1	F2
Pubmed:23571154	Kisspeptin, leptin, and retinol-binding protein 4 in women with polycystic ovary syndrome.	1.00E+00	1	3	RBP4
Pubmed:23685884	Haplotype TGTG from SNP 45T/G and 276G/T of the adiponectin gene contributes to risk of polycystic ovary syndrome.	1.00E+00	1	1	ADIPOQ
Pubmed:25069671	Association between retinol-binding protein 4 and polycystic ovary syndrome: a meta-analysis.	1.00E+00	1	1	RBP4
Pubmed:29428584	The preliminary association study of ADIPOQ, RBP4, and BCMO1 variants with polycystic ovary syndrome and with biochemical characteristics in a cohort of Polish women.	1.36E-02	2	3	ADIPOQ, RBP4
CID000040973	desogestrel	1.58E-13	10	106	
CID000005526	tranexamic acid	3.41E-11	8	64	

Table 4.4 Enriched terms for PCOS vs HT. On the left, GO Biological Process terms analyzed with JEPETTO. On the right, Pubmed entries and Stitch drug database (CID) from Toppgene.

4.4.1.3 HO vs HT

Five samples coming from subjects with a BMI over 30 (HO) are compared to five samples from subjects with a BMI under 30 (HT): **HO vs HT** (Figure 4.17). In this comparison, metabolic effects of overweight are measured, independently of the PCOS pathology studied in this work. A total of 62 proteins pass the cut-off (P-value<0.05, log2 fold change= 0.26).

The hierarchical cluster in Figure 4.17 shows two very well defined groups of proteins: the height of the dendogram clearly shows a group of 15 proteins under-expressed under HO samples and over-expressed for HT samples, while the rest of the proteins present the opposite behavior. While being well classified in the HT group , sample HT.1425 presents a somewhat distinct pattern of expression than the other four HT samples: a more intense over-expression and some of the proteins expected to be under-expressed, actually over-expressed.

In the Figure 4.17 cluster, the protein SOD1 appears over-expressed in HO samples and under expressed in one HT sample, and does not appear at all in the other five HT samples. In this scenario, the protein is not excluded from further analysis because something that could be happening here is that the protein levels are under detection levels in four of the HT samples. Instead of an artifact, maybe some valuable information may be obtained by keeping this protein among the ones studied.

The number of differential proteins obtained in HO vs HT is the highest of all the comparisons done. The expression profile is clearly one of over-expression, with ¾ of the proteins up-regulated in front of ¼ down-regulated.

HO vs HT

		log2FC	Pval
SAA4	Serum amyloid A-4 protein	-1.40	6.77E-03
ORM2	Alpha-1-acid glycoprotein 2	-1.26	2.68E-03
FBLN1	Fibulin-1 (FIBL-1)	-0.75	2.12E-03
F2	Prothrombin (Coagulation factor II)	-0.72	5.72E-05
APOM	Apolipoprotein M	-0.56	4.58E-03
C7	Complement component C7	-0.44	1.14E-03
SERPINC1	Antithrombin-III (ATIII) (Serpin C1)	-0.28	2.57E-03
SERPINF2	Alpha-2-antiplasmin	-0.27	8.12E-03
MASP1	Mannan-binding lectin serine protease 1	0.51	1.04E-02
FN1	Fibronectin	0.52	5.57E-03
F5	Coagulation factor V	0.61	2.47E-03
CFB	Complement factor B	0.61	8.57E-04
CFP	Properdin (Complement factor P)	0.62	4.09E-04
C8B	Complement component C8 beta chain	0.68	3.86E-04
F11	Coagulation factor XI (FXI)	0.72	2.20E-05
CD14	Monocyte differentiation antigen CD14	0.73	7.14E-04
QSOX1	Sulfhydryl oxidase 1	0.80	2.65E-04
HP	Haptoglobin (Zonulin)	0.97	3.83E-04
LBP	Lipopolysaccharide-binding protein (LBP)	0.97	1.62E-03
APCS	Serum amyloid P-component (SAP)	1.01	4.98E-04
CD44	CD44 antigen (CDw44) (Epican)	1.24	1.12E-02
PRG4	Proteoglycan 4 (Lubricin)	1.30	6.22E-04
PIGR	Polymeric immunoglobulin receptor (PIgR)	1.38	9.09E-04
IGFBP6	Insulin-like growth factor-binding protein 6	1.39	8.79E-04
COLEC11	Collectin-11	1.46	1.48E-04
TLN1	Talin-1	1.49	3.80E-03
CAMP	Cathelicidin antimicrobial peptide	1.82	1.58E-03
PF4	Platelet factor 4 (PF-4)	2.21	2.87E-03
IGLV4-69	Immunoglobulin lambda variable 4-69	2.64	3.53E-03
LPA	Apolipoprotein(a)	2.67	1.27E-03
CRP	C-reactive protein	3.84	6.57E-04

		log2FC	Pval
IGHM	Immunoglobulin heavy constant mu	-2.40	2.57E-02
CFHR4	Complement factor H-related protein 4	-2.13	1.60E-02
IGKV1-16	Immunoglobulin kappa variable 1-16	-2.00	1.55E-02
APOC1	Apolipoprotein C-I (Apo-CI)	-1.35	1.73E-02
GPX3	Glutathione peroxidase 3	-0.99	3.60E-02
APOD	Apolipoprotein D	-0.63	3.31E-02
SERPINA7	Thyroxine-binding globulin	-0.33	2.96E-02
C8A	Complement component C8 alpha chain	0.30	2.74E-02
SERPING1	Plasma protease C1 inhibitor	0.32	2.52E-02
C8G	Complement component C8 gamma chain	0.33	1.28E-02
C5	Complement C5	0.35	2.51E-02
C1S	Complement C1s subcomponent	0.37	1.96E-02
BCHE	Cholinesterase	0.42	3.05E-02
SERPIND1	Heparin cofactor 2	0.42	1.37E-02
KLKB1	Plasma kallikrein (Fletcher factor)	0.48	1.86E-02
TGFBI	Transforming GF-beta-induced ig-h3	0.50	4.08E-02
SPP2	Secreted phosphoprotein 24	0.55	3.75E-02
C4BPB	C4b-binding protein beta chain	0.58	4.00E-02
PROS1	Vitamin K-dependent protein S	0.59	1.20E-02
CETP	Cholesteryl ester transfer protein	0.61	1.27E-02
C9	Complement component C9	0.61	2.75E-02
LRG1	Leucine-rich alpha-2-glycoprotein	0.64	2.28E-02
PFN1	Profilin-1	0.86	3.29E-02
SOD1	Superoxide dismutase [Cu-Zn]	0.90	2.36E-02
VCL	Vinculin (Metavinculin)	1.14	3.87E-02
PPBP	Platelet basic protein (PBP)	1.27	3.70E-02
FERMT3	Fermitin family homolog 3	1.30	1.60E-02
YWHAZ	14-3-3 protein zeta/delta	1.38	3.92E-02
ORM1	Alpha-1-acid glycoprotein 1	1.44	2.69E-02
THBS1	Thrombospondin-1 (Glycoprotein G)	1.74	1.25E-02
APOC3	Apolipoprotein C-III	2.73	2.29E-02

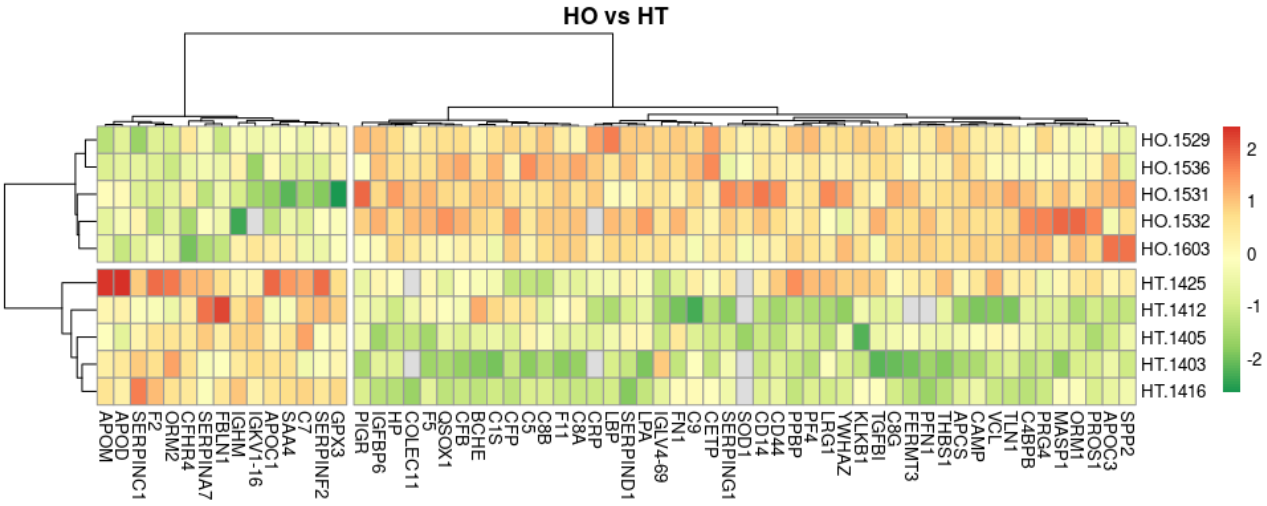


Figure 4.17 62 differentially expressed proteins in healthy obese (HO) vs healthy thin (HT) comparison: list with the log2 Fold Changes and P-values (up) and a hierarchical cluster of the individual Log intensities (bottom), clustering both samples (rows) and proteins (columns).

The enrichment analysis, shown in Table 4.5, presents complement, coagulation and acute phase as the most prominent categories. The high number of differentially expressed proteins makes that the enrichment P-values and the coverage of the GO categories is quite high.

Differences in plasma concentration of proteins have been widely reported in literature: only one example has been included in Table 4.5 with several of the proteins found here. Some pathologies and phenotypes have been included in the enrichment table as well, just to highlight the elevated amount of enriched terms found with the proteins obtained here.

This reflects the fact that this kind of comparison is a well-studied area of knowledge, and that annotation using ontologies and literature is very common.

HO vs HT

GO Term	XD-score	q-value	n/N
complement activation, alternative pathway	2.911	0.00E+00	7/11
peptidyl-glutamic acid carboxylation	1.768	7.58E-02	2/10
negative regulation of fibrinolysis	1.679	1.72E-03	3/10
fibrinolysis	1.641	0.00E+00	5/18
cytolysis	1.635	0.00E+00	6/16
regulation of complement activation	1.553	2.21E-03	3/11
complement activation, classical pathway	1.452	0.00E+00	10/29
complement activation	1.375	0.00E+00	7/22
regulation of blood coagulation	1.268	7.58E-02	2/10
positive regulation of blood coagulation	1.268	7.58E-02	2/10
blood coagulation, intrinsic pathway	1.203	1.90E-04	4/17
acute-phase response	0.935	0.00E+00	7/31
positive regulation of collagen biosynthetic process	0.897	1.33E-01	2/14
negative regulation of astrocyte differentiation	0.868	1.00E+00	1/10
cytosolic calcium ion homeostasis	0.868	1.00E+00	1/10
positive regulation of tumor necrosis factor biosynthetic process	0.768	7.58E-02	2/10

Term	Description	q-value	n	N	Genes
Pubmed:27754964	Investigation of the inflammatory biomarkers of metabolic syndrome in adolescents.	6.97E-05	3	3	APCS, ORM1, PF4
HP:0005339	Abnormality of complement system	2.12E-04	8	109	
HP:0004431	Complement deficiency	2.68E-03	6	69	
CID006450878	Etiocobalamin	2.62E-20	18	177	
CID000040973	desogestrel	1.25E-16	14	106	
CID000005526	tranexamic acid	8.45E-10	9	64	
C0149871	Deep Vein Thrombosis	6.44E-08	9	90	
C0020473	Hyperlipidemia	1.12E-07	11	186	
C0040053	Thrombosis	4.65E-07	7	45	
C0087086	Thrombus	4.65E-07	7	45	
C0242666	Protein S Deficiency	5.88E-07	6	25	

Table 4.5 Enriched terms for HO vs HT. On the left, GO Biological Process terms analyzed with JEPETTO. On the right, Pubmed entries, Disease terms (DisGeNET) and Stitch drug database (CID) from Toppgene.

4.4.1.4 PT vs HT

Five PT samples are compared to five healthy control samples (HT) here. Without the interference (or co-occurrence) introduced by PO and HO, this should the easiest way to approach PCOS interpretation. The list of differentially expressed proteins and the cluster done are shown in Figure 4.18.

PT vs HT

		log2FC	Pval
SAA4	Serum amyloid A-4 protein	-1.67	1.87E-03
GPX3	Glutathione peroxidase 3	-1.45	3.98E-03
ORM2	Alpha-1-acid glycoprotein 2	-1.17	4.56E-03
F2	Prothrombin (Coagulation factor II)	-0.53	1.05E-03
C7	Complement component C7	-0.38	3.67E-03
KLKB1	Plasma kallikrein (Fletcher factor)	0.61	3.95E-03
PLTP	Phospholipid transfer protein	0.62	9.07E-03
F5	Coagulation factor V	0.63	1.87E-03
CFP	Properdin (Complement factor P)	0.64	3.00E-04
F11	Coagulation factor XI (FXI)	0.66	5.74E-05
MASP1	Mannan-binding lectin serine protease 1	0.68	1.30E-03
QSOX1	Sulfhydryl oxidase 1	0.72	6.94E-04
PROS1	Vitamin K-dependent protein S	0.74	2.77E-03
AGT	Angiotensinogen (Serpin A8)	0.77	8.02E-03
CD14	Monocyte differentiation antigen CD14	0.88	1.33E-04
C4BPB	C4b-binding protein beta chain	0.94	2.33E-03
LBP	Lipopolysaccharide-binding protein (LBP)	1.19	2.71E-04
PRG4	Proteoglycan 4 (Lubricin)	1.20	1.31E-03
SPP2	Secreted phosphoprotein 24	1.24	1.03E-04
CD44	CD44 antigen (CDw44) (Epican)	1.59	2.07E-03
COLEC11	Collectin-11	1.61	6.42E-05
IGFBP6	Insulin-like growth factor-binding protein 6	1.72	1.21E-04
PZP	Pregnancy zone protein	1.83	2.44E-03
CAMP	Cathelicidin antimicrobial peptide	2.78	3.12E-05
IGLV4-69	Immunoglobulin lambda variable 4-69	2.90	1.78E-03

		log2FC	Pval
IGHM	Immunoglobulin heavy constant mu	-2.52	2.00E-02
P0DOX3	Immunoglobulin delta heavy chain	-2.09	2.16E-02
IGLV4-60	Immunoglobulin lambda variable 4-60	-0.89	3.47E-02
FBLN1	Fibulin-1 (FIBL-1)	-0.57	1.31E-02
SERPINA1	Alpha-1-antitrypsin	-0.55	2.11E-02
FN1	Fibronectin	0.39	2.88E-02
CLU	Clusterin	0.39	1.49E-02
A1BG	Alpha-1B-glycoprotein	0.44	3.14E-02
BCHE	Cholinesterase	0.45	2.02E-02
F10	Coagulation factor X	0.47	2.82E-02
LUM	Lumican (Keratan sulfate proteoglycan lumican)	0.48	4.48E-02
TGFBI	Transforming GF-beta-induced ig-h3	0.52	3.46E-02
F13B	Coagulation factor XIII B chain	0.58	4.78E-02
APOL1	Apolipoprotein L1	0.59	1.31E-02
RBP4	Retinol-binding protein 4	0.59	3.69E-02
APCS	Serum amyloid P-component (SAP)	0.67	1.07E-02
APOC4	Apolipoprotein C-IV	0.73	4.45E-02
PIGR	Polymeric immunoglobulin receptor (PIgR)	0.80	4.05E-02
APOA1	Apolipoprotein A-I	0.81	1.51E-02
SOD1	Superoxide dismutase [Cu-Zn]	0.96	1.72E-02
TLN1	Talin-1	1.01	3.58E-02
YWHAZ	14-3-3 protein zeta/delta	1.34	4.52E-02
P0DOX7	Immunoglobulin kappa light chain	1.34	3.08E-02
APOC2	Apolipoprotein C-II	1.46	1.56E-02
LPA	Apolipoprotein(a)	1.94	1.19E-02

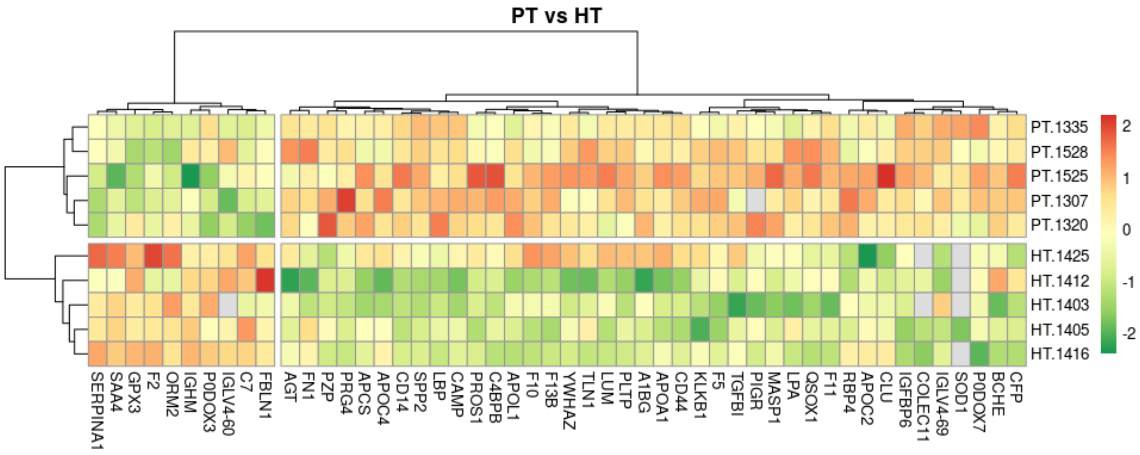


Figure 4.18 50 differentially expressed proteins in PCOS thin (PT) vs healthy thin (HT) comparison: list with the log2 Fold Changes and P-values (up) and a hierarchical cluster of the individual Log intensities (bottom), clustering both samples (rows) and proteins (columns).

From the five proteins (FLNA, ADIPOQ, LBP, RBP4 and APOC2) found differentially expressed in both PCOS vs H and PCOS vs HT:

- LBP (+1.2), RBP4 (+0.6) and APOC2 (+1.5) show very similar levels to those found under PCOS vs HT.
- FLNA (-1.1), although not differentially expressed in this comparison (it has a P-value of 0.053, slightly over the cut-off), shows almost the same log2 fold change that in PCOS vs HT and therefore, should be considered as still being in the previous group.
- With ADIPOQ, the opposite happens: while showing a clear down-regulation in PCOS vs HT (-1.1) here appears with a fold change close to 0 (log2FC=-0.027 and P-value=0.95); while the difference between this comparison (PT vs PO) and PCOS vs HT is the absence here of PO samples, it will be expected finding ADIPOQ highly down-regulated in PO vs HT.

From the cluster, patients appear perfectly classified and two clear groups of proteins are separated by the dendogram. Comparing this cluster to the one obtained in HO vs HT, two points must be considered:

- Similarly to what happened in HO vs HT, the sample HT.1425 is separated from the other four, showing a more over-expressed pattern (although not so clearly).
- The overlap of differentially expressed proteins in HO vs HT and PT vs HT is evident; taking a look at the cluster of ten proteins on the left here, six of the proteins (SAA4, GPX3, F2, ORM2, IGHM and FBLN1) are also found in the left cluster under HO vs HT, in bout cases corresponding to proteins under-expressed in PCOS; many overlapping proteins can also bee found in the right clusters.

An enrichment analysis for PT vs HT is shown in Table 4.6. And again complement, coagulation and acute phase response appear as the main enriched categories.

PT vs HT

GO Term	XD-score	q-value	n/N
peptidyl-glutamic acid carboxylation	2.171	1.05E-03	3/10
regulation of complement activation	1.526	1.34E-03	3/11
fibrinolysis	1.421	1.40E-04	4/18
negative regulation of fibrinolysis	1.282	6.15E-02	2/10
regulation of blood coagulation	1.271	6.15E-02	2/10
blood coagulation, intrinsic pathway	1.206	1.40E-04	4/17
acute-phase response	1.035	0.00E+00	7/31
positive regulation of collagen biosynthetic process	0.899	1.02E-01	2/14
negative regulation of astrocyte differentiation	0.871	8.04E-01	1/10
cytosolic calcium ion homeostasis	0.871	8.04E-01	1/10
positive regulation of blood coagulation	0.871	8.04E-01	1/10
negative regulation of proteolysis	0.721	8.28E-01	1/12
complement activation, alternative pathway	0.708	6.47E-02	2/11
complement activation	0.703	2.90E-04	4/22
positive regulation of cytokine production	0.698	6.47E-02	2/11

Term	Description	q-value	n	N	Genes
HP:0010990	Abnormality of the common coagulation pathway	2.38E-04	5	29	
HP:0005261	Joint hemorrhage	7.82E-04	4	16	
HP:0003645	Prolonged partial thromboplastin time	1.26E-03	5	40	
HP:0010988	Abnormality of the extrinsic pathway	4.42E-03	4	24	
Pubmed:14669168	Polycystic ovary syndrome, the G1691A factor V Leiden mutation, and plasminogen activator inhibitor activity: associations with recurrent pregnancy loss.	2.91E-03	2	5	F2, F5
Pubmed:22341881	Metabolic manifestations of polycystic ovary syndrome in nonobese adolescents: retinol-binding protein 4 and ectopic fat deposition.	2.46E-02	1	1	RBP4
Pubmed:19158194	Retinol-binding protein 4 in polycystic ovary syndrome--association with steroid hormones and response to pioglitazone treatment.	2.46E-02	1	1	RBP4
Pubmed:16275260	The M235T polymorphism of the angiotensinogen gene in women with polycystic ovary syndrome.	2.46E-02	1	1	AGT

Table 4.6 Enriched terms for PT vs HT. On the left, GO Biological Process terms analyzed with JEPETTO. On the right, Pubmed entries and Human Phenotype (HP) from Toppgene.

4.4.1.5 PO vs HT

Five PO samples are evaluated in front of five healthy controls (HT) in this comparison. Being PO a more “extreme” phenotype, presenting both the conditions under study in this work (PCOS and obesity), an exceptionally high number of differentially expressed proteins could be expected. As shown in Figure 4.19, the opposite happens: only 20 proteins differentially expressed appear, representing less than a half of the ones found under PT vs HT.

PO vs HT

		log2FC	Pval
ADIPOQ	Adiponectin	-2.21	1.00E-04
FLNA	Filamin-A	-1.60	1.48E-02
IGHV5-51	Immunoglobulin heavy variable 5-51	-1.26	3.64E-02
COLEC11	Collectin-11	-1.18	3.63E-03
IGFBP3	Insulin-like growth factor-binding protein 3	-0.92	1.23E-03
SPP2	Secreted phosphoprotein 24	-0.80	4.31E-03
CD5L	CD5 antigen-like	-0.77	4.39E-02
APOD	Apolipoprotein D	-0.72	1.72E-02
IGFALS	Insulin-like GF-binding prot.comp. ALS	-0.59	3.66E-02
FBLN1	Fibulin-1 (FIBL-1)	-0.58	1.22E-02
C7	Complement component C7	-0.58	8.49E-05
A2M	Alpha-2-macroglobulin	-0.49	1.90E-02
GSN	Gelsolin	-0.43	2.28E-02
APOM	Apolipoprotein M	-0.42	2.49E-02
RBP4	Retinol-binding protein 4	0.62	3.10E-02
LBP	Lipopolysaccharide-binding protein (LBP)	0.65	2.32E-02
APCS	Serum amyloid P-component (SAP)	0.74	5.50E-03
PRG4	Proteoglycan 4 (Lubricin)	0.97	6.28E-03
CRP	C-reactive protein	2.22	1.86E-02
LPA	Apolipoprotein(a)	2.24	4.85E-03

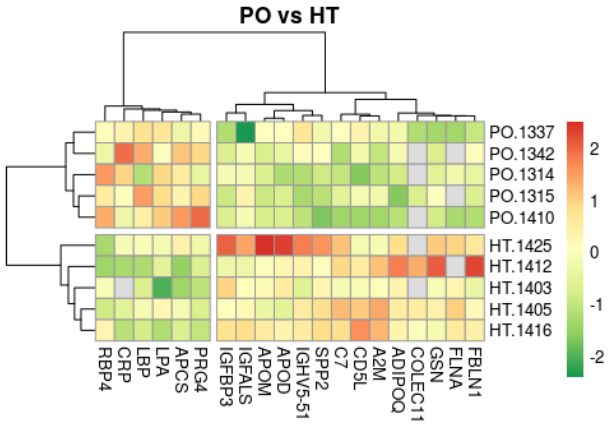


Figure 4.19 20 differentially expressed proteins in PCOS obese (PO) vs healthy thin (HT) comparison: list with the log2 Fold Changes and P-values (up) and a hierarchical cluster of the individual Log intensities (bottom), clustering both samples (rows) and proteins (columns).

Enriched terms using JEPETTO are shown in Table 4.7: only one GO term. Using Toppgene for a Gene Ontology search, using 20 genes as a limit of the groups searched, several terms appear slightly enriched, for example:

- renal protein absorption (GO:0097017): GSN and ADIPOQ (2 out of 3 genes)
- complement activation, lectin pathway (GO:0001867): COLEC11 and A2M (2 out of 3 genes)

The poor enrichment here is obviously related to the limited number of proteins, but even with that, the virtual absence of enriched GO terms should also be explained by some other reason.

PO vs HT

GO Term	XD-score	q-value	n/N
tissue regeneration	0.463	1.54E-01	2/17

Term	Description	q-value	n	N	Genes
Pubmed:18445670	Serum levels of retinol-binding protein 4 and adiponectin in women with polycystic ovary syndrome: associations with visceral fat but no evidence for fat mass-independent effects on pathogenesis in this condition.	4.94E-04	2	2	ADIPOQ, RBP4
Pubmed:18616717	Inflammatory markers and visceral fat are inversely associated with maximal oxygen consumption in women with polycystic ovary syndrome (PCOS).	8.46E-03	1	1	CRP
Pubmed:23864804	Adipokines, insulin-like growth factor binding protein-3 levels, and insulin sensitivity in women with polycystic ovary syndrome.	7.14E-04	2	3	ADIPOQ, IGFBP3

Table 4.7 Enriched terms for PO vs HT. On the left, one GO Biological Process terms analyzed with JEPETTO. On the right, Pubmed entries from Toppgene.

The low level of differential expression could be explained by some sort of *general disorganization* in the protein levels, where only proteins shown in Figure 4.19 would show some coordinated trend. Another possible reason is the emergence of more than one biological mechanism in PO individuals: in contrast with PT, where a list of coherent processes are described by up and under-expression of a set of proteins, in PO several divergent processes in the samples studied would provide *several sub-phenotypes*. The analysis of more samples of the PO phenotype would be necessary for studying such hypotheses. The poor enrichment is compatible with the both explanations: the dispersion of a few differentially expressed proteins among a bunch of ontology categories can be explained by the involvement of many different and distant biological processes (general disorganization) or by the existence of more than one phenotype (sub-phenotypes in PO).

As for the proteins differentially expressed here:

- ADIPOQ is highly down-regulated in PO vs HT, as expected from results previously obtained: being present in PCOS vs HT among differential proteins (log2FC = -1.1), and being absent in PT vs HT, only a very low log2FC here (-2.2) could compensate things.
- FLNA has not been quantified in two of the five HT samples, what would be compatible with the low expression in the other three inside the same phenotype; but the appearance of another sample missed for this protein in PO subjects may suggest some artifact related with this protein.

4.4.1.6 PO vs HO

In this comparison, five PO samples are evaluated in front of five healthy-obese samples as controls (HO). The differentially expressed proteins here should reflect changes originated by PCOS in obese patients. In Figure 4.20 the list of the 53 differential proteins and the

hierarchical cluster are shown. The cluster shows again phenotypes well classified and two subsets of proteins, one under-expressed in HO with only five proteins. The profile of expression here is of global under-expression, that is, proteins generally present higher levels in HO than in PO. Again, this goes against expectations, where a more severe phenotype (PO), would be expected to produce more extreme protein changes.

PO vs HO

		log2FC	Pval
THBS1	Thrombospondin-1 (Glycoprotein G)	-2.66	1.95E-03
COLEC11	Collectin-11	-2.64	2.21E-06
YWHAZ	14-3-3 protein zeta/delta	-2.11	3.41E-03
ADIPOQ	Adiponectin	-2.11	1.66E-04
PPBP	Platelet basic protein (PBP)	-1.95	3.05E-03
FERMT3	Fermitin family homolog 3	-1.90	8.18E-04
IGFBP6	Insulin-like growth factor-binding protein 6	-1.85	1.08E-04
CD44	CD44 antigen (CDw44) (Epican)	-1.75	9.31E-04
TLN1	Talin-1	-1.73	1.18E-03
VCL	Vinculin (Metavinculin)	-1.65	4.86E-03
PFN1	Profilin-1	-1.48	6.37E-04
PIGR	Polymeric immunoglobulin receptor (PIgR)	-1.43	6.92E-04
SPP2	Secreted phosphoprotein 24	-1.35	4.06E-05
IGFBP3	Insulin-like growth factor-binding protein 3	-1.21	9.70E-05
QSOX1	Sulfhydryl oxidase 1	-1.13	6.15E-06
C8B	Complement component C8 beta chain	-0.98	8.16E-06
CFP	Properdin (Complement factor P)	-0.80	3.08E-05
F11	Coagulation factor XI (FXI)	-0.75	1.28E-05
BCHE	Cholinesterase	-0.73	7.30E-04
CD14	Monocyte differentiation antigen CD14	-0.69	1.21E-03
FN1	Fibronectin	-0.65	1.11E-03
C8G	Complement component C8 gamma chain	-0.55	2.43E-04
MASP1	Mannan-binding lectin serine protease 1	-0.55	6.07E-03
C8A	Complement component C8 alpha chain	-0.52	8.20E-04
CFB	Complement factor B	-0.50	3.95E-03
F2	Prothrombin (Coagulation factor II)	0.64	1.95E-04
ORM2	Alpha-1-acid glycoprotein 2	1.41	1.13E-03

		log2FC	Pval
P0DOX2	Immunoglobulin alpha-2 heavy chain	-3.05	4.15E-02
IGLV4-69	Immunoglobulin lambda variable 4-69	-2.48	1.29E-02
PF4	Platelet factor 4 (PF-4)	-1.92	7.44E-03
ORM1	Alpha-1-acid glycoprotein 1	-1.48	2.34E-02
CAMP	Cathelicidin antimicrobial peptide	-1.33	1.83E-02
P0DOX7	Immunoglobulin kappa light chain	-1.26	3.96E-02
IGLV8-61	Immunoglobulin lambda variable 8-61	-1.12	4.33E-02
MST1	Hepatocyte growth factor-like protein	-0.91	3.90E-02
IGFALS	Insulin-like GF-binding prot.comp. ALS	-0.81	6.46E-03
LRG1	Leucine-rich alpha-2-glycoprotein	-0.70	1.38E-02
TGFBI	Transforming GF-beta-induced ig-h3	-0.69	7.59E-03
APOA1	Apolipoprotein A-I	-0.67	4.10E-02
C4BPB	C4b-binding protein beta chain	-0.62	3.05E-02
PLTP	Phospholipid transfer protein	-0.60	1.02E-02
HP	Haptoglobin (Zonulin)	-0.56	1.92E-02
LUM	Lumican (Keratan sulfate proteoglycan lumican)	-0.56	2.27E-02
F13A1	Coagulation factor XIII A chain	-0.53	3.09E-02
C1RL	Complement C1r subcomponent-like protein	-0.51	9.52E-03
GPLD1	Phosphatidylinositol-glycan-spec PPlipase D	-0.49	4.86E-02
C5	Complement C5	-0.39	1.55E-02
SERPING1	Plasma protease C1 inhibitor	-0.38	9.76E-03
HPX	Hemopexin (Beta-1B-glycoprotein)	-0.37	2.55E-02
C1S	Complement C1s subcomponent	-0.31	4.07E-02
APOC1	Apolipoprotein C-I (Apo-CI)	1.15	3.80E-02
SAA4	Serum amyloid A-4 protein	1.40	6.79E-03
CFHR4	Complement factor H-related protein 4	1.86	3.16E-02

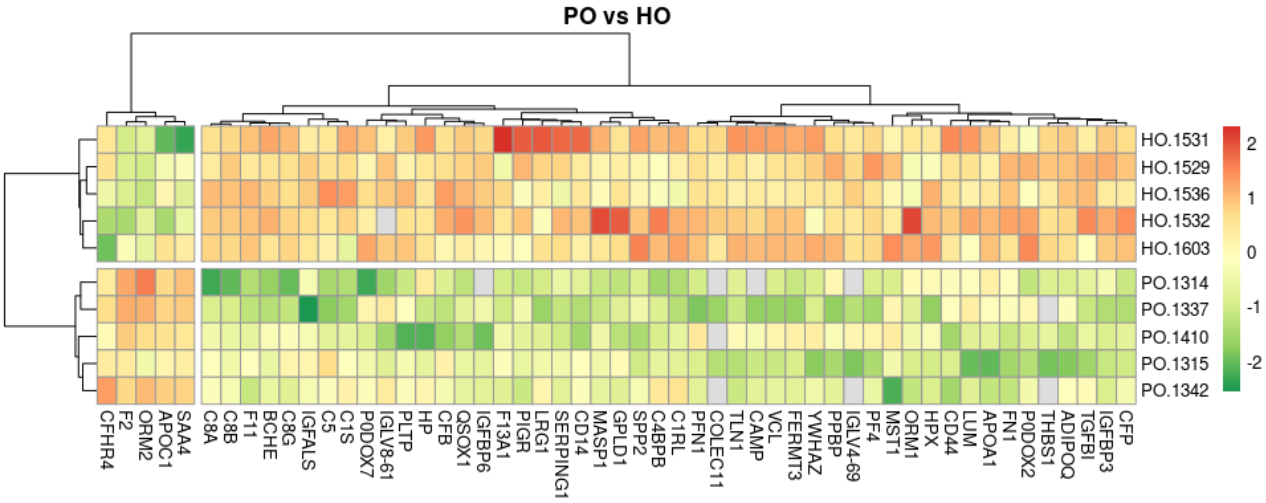


Figure 4.20 53 differentially expressed proteins in PCOS obese (PO) vs healthy obese (HO) comparison: list with the log2 Fold Changes and P-values (up) and a hierarchical cluster of the individual Log intensities (bottom), clustering both samples (rows) and proteins (columns).

In Table 4.8 the enrichment results are shown: coagulation, acute phase and complement are the main categories, similarly to what happened in PT vs HT and HO vs HT. Proteins related to the innate immune response (MASP1, CD14, SERPING1, C1S, C1RL, C4BPB, C5, C8A, C8B, C8G and CFB) appear here forming a group not seen in previous enrichments.

The consistent way in which PO differentiates from HO brings some light to the low differential expression of proteins in PO vs HT: instead of the co-occurrence of phenotypes

(not likely because PO works as a well defined group here), the hypothesis that proteins were acting in an uncoordinated way in the PO vs HT comparison is strengthen. The conclusion here, with the data available, would be that HT is simply not a good reference to PO for many proteins (but it is for the 20 proteins found differential) and that the complexity of changes taking place into PO phenotypes is better understood using PT as reference.

PO vs HO

GO Term	XD-score	q-value	n/N
positive regulation of blood coagulation	0.780	7.61E-02	2/10
regulation of complement activation	0.756	8.30E-02	2/11
cholesterol efflux	0.730	8.30E-02	2/11
blood coagulation, intrinsic pathway	0.686	8.21E-03	3/17
peptide cross-linking	0.480	8.21E-03	3/17
platelet degranulation	0.461	0.00E+00	10/76
fibrinolysis	0.430	1.89E-01	2/18
positive regulation of reactive oxygen species metabolic process	0.401	2.00E-01	2/19
lipid transport	0.320	6.25E-02	3/36
acute-phase response	0.246	4.33E-02	3/31
platelet activation	0.207	0.00E+00	12/195
negative regulation of angiogenesis	0.180	7.10E-01	2/40
cholesterol metabolic process	0.155	9.03E-01	2/47
innate immune response	0.141	0.00E+00	11/252
cell-matrix adhesion	0.107	1.89E-01	3/64
blood coagulation	0.106	0.00E+00	14/414

Term	Description	q-value	n	N	Genes
HP:0005339	Abnormality of complement system	2.44E-03	6	109	
HP:0004434	C8 deficiency	8.45E-03	2	2	
Pubmed:23864804	Adipokines, insulin-like growth factor binding protein-3 levels, and insulin sensitivity in women with polycystic ovary syndrome.	3.89E-03	2	3	ADIPOQ, IGFBP3
Pubmed:25336505	Serum zonulin is elevated in women with polycystic ovary syndrome and correlates with insulin resistance and severity of anovulation.	2.94E-02	1	1	HP
Pubmed:19368908	Can serum apolipoprotein C-I demonstrate metabolic abnormality early in women with polycystic ovary syndrome?	2.94E-02	1	1	APOC1
Pubmed:23415973	Faster thrombin generation in women with polycystic ovary syndrome compared with healthy controls matched for age and body mass index.	2.94E-02	1	1	F2
Pubmed:28789706	Small leucine-rich proteoglycans (SLRPs) in the endometrium of polycystic ovary syndrome women: a pilot study.	3.86E-02	1	2	LUM
Pubmed:24336678	Increased expression of kindlin 2 in luteinized granulosa cells correlates with androgen receptor level in patients with polycystic ovary syndrome having hyperandrogenemia.	3.86E-02	1	2	FERMT3
Pubmed:22565190	Androgen levels and metabolic parameters are associated with a genetic variant of F13A1 in women with polycystic ovary syndrome.	2.94E-02	1	1	F13A1

Table 4.8 Enriched terms for PO vs HO. On the left, GO Biological Process terms analyzed with JEPETTO (filtered terms due the long list provided). On the right, Pubmed entries and Human Phenotype (HP) from Toppgene.

4.4.1.7 PT vs HO

Five PCOS thin (PT) samples are evaluated in front of five healthy obese samples (HO) in this comparison, illustrating the differences between the simplest models for obesity and PCOS in this work. As previously seen, PT vs HT and HO vs HT are quite similar: protein levels have a lot in common in samples of lean patients diagnosed with PCOS and those without PCOS but with a BMI index higher than 30. For this reason, the low number of differential proteins found here (15 proteins) was in fact expected.

Protein expression levels and the hierarchical cluster are shown in Table 4.21. Protein IGHG3 appears here in only three samples (two as HO control and one as PT); in this case, the protein will discarded for further analysis because, although its levels are consistent in the respective groups, too few samples are analyzed. The hierarchical cluster in Table 4.21 is split in two levels for proteins and also for samples: differences in the dendogram's lengths are quite big between the two first levels for both proteins and samples, suggesting a strong division between groups.

PT vs HO

		log2FC	Pval
CRP	C-reactive protein	-3.03	2.76E-03
P0DOX3	Immunoglobulin delta heavy chain	-1.93	4.15E-02
PFN1	Profilin-1	-0.89	2.13E-02
HP	Haptoglobin (Zonulin)	-0.63	1.04E-02
C8B	Complement component C8 beta chain	-0.46	7.97E-03
FGB	Fibrinogen beta chain	-0.43	2.47E-02
C8G	Complement component C8 gamma chain	-0.41	3.03E-03
CFB	Complement factor B	-0.41	1.38E-02
C8A	Complement component C8 alpha chain	-0.36	1.12E-02
C3	Complement C3	-0.31	3.57E-02
APOM	Apolipoprotein M	0.47	1.44E-02
AGT	Angiotensinogen (Serpin A8)	0.61	3.06E-02
SPP2	Secreted phosphoprotein 24	0.69	1.15E-02
SHBG	Sex hormone-binding globulin	0.99	1.14E-02
MBL2	Mannose-binding protein C	1.53	2.96E-02
IGHG3	Immunoglobulin heavy constant gamma 3	7.08	7.68E-03

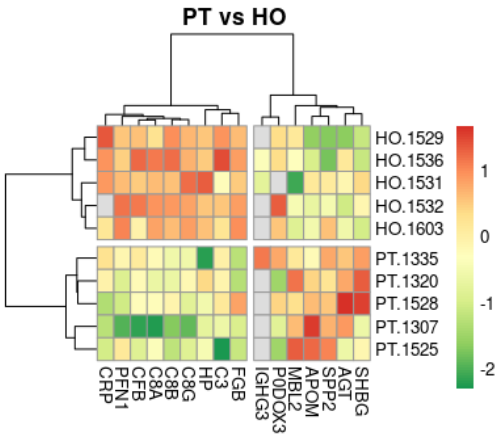


Figure 4.21 15 differentially expressed proteins after discarding IGHG3 in PCOS thin (PT) vs healthy obese (HO) comparison: list with the log2 Fold Changes and P-values (up) and a hierarchical cluster of the individual Log intensities (bottom), clustering both samples (rows) and proteins (columns).

In Table 4.9, the enrichment results are shown. Two groups appear clearly enriched for this comparison: complement activation, alternative and classical pathways. Some Human Phenotype categories, diseases (DisGeNET), drugs (Stich database) and some bibliographical entries are also included.

Proteins related to complement show a general under-expression in PT with respect to HO (FGB, CFB, C3, C8A, C8B and C8G) with the exception of MBL2 that is clearly up-regulated in PT vs HO; the fact that MLB2 is over-expressed here is consistent with the fact that this protein is under-expressed both in HO vs HT and PO vs HT, and over-expressed in PT vs HT, even though in neither of the three cases (HO vs HT, PT vs HT and PO vs HT) the P-value was even close to 0.05. It is only here in PT vs HO (and later in PO vs PT) that MBL2 has significantly different levels from the reference (log2FC=1.53, P-value=2.96E-2).

PT vs HO

GO Term	XD-score	q-value	n/N
complement activation, alternative pathway	1.193	0.00E+00	4/11
complement activation, classical pathway	0.707	0.00E+00	6/29

Term	Description	q-value	n	N	Genes
HP:0100519	Anuria	1.49E-04	3	15	
HP:0004434	C8 deficiency	3.11E-04	2	2	
CID006450878	Etiocobalamin	2.52E-10	8	177	
CID000005470	AC1L1KF5	2.89E-05	5	113	
C0030167	Pachymeningitis	5.95E-07	3	3	
Pubmed:18206145	Increased acylation-stimulating protein, C-reactive protein, and lipid levels in young women with polycystic ovary syndrome.	2.00E+00	2	2	C3, CRP
Pubmed:16275260	The M235T polymorphism of the angiotensinogen gene in women with polycystic ovary syndrome.	1.00E+00	1	1	AGT
Pubmed:23812344	Sex hormone binding globulin, but not testosterone, is associated with the metabolic syndrome in overweight and obese women with polycystic ovary syndrome.	1.00E+00	1	1	SHBG
Pubmed:18616717	Inflammatory markers and visceral fat are inversely associated with maximal oxygen consumption in women with polycystic ovary syndrome (PCOS).	1.00E+00	1	1	CRP
Pubmed:28900795	Haptoglobin levels, but not Hp1-Hp2 polymorphism, are associated with polycystic ovary syndrome.	1.00E+00	1	1	HP
Pubmed:21178921	Sex hormone-binding globulin genetic variation: associations with type 2 diabetes mellitus and polycystic ovary syndrome.	1.00E+00	1	1	SHBG
Pubmed:23257395	Familial aggregation of circulating C-reactive protein in polycystic ovary syndrome.	1.00E+00	1	1	CRP
Pubmed:19440331	Role of haptoglobin in polycystic ovary syndrome (PCOS), obesity and disorders of glucose tolerance in premenopausal women.	1.00E+00	1	1	HP

Table 4.9 Enriched terms for PT vs HO. On the left, GO Biological Process terms analyzed with JEPETTO. On the right, Pubmed entries, Human Phenotype (HP), Disease terms (DisGeNET) and Stitch drug database (CID) from Toppgene.

4.4.1.8 PO vs PT

Finally, five PCOS obese (PO) samples are evaluated in front of five PCOS thin samples (PT) here. This comparison allows to assess differences in protein concentrations for samples diagnosed with PCOS with and without obesity.

Here, 61 proteins are differentially expressed, with an expression profile generally under-expressed (Figure 4.22): the expression levels of proteins is very similar here to the one analyzed with PO vs HO, both in terms of proteins involved and log2FC levels. The hierarchical cluster shows two groups of samples (rows) and proteins (columns) clearly separated. For the same reasons than in previous cases, IGHG3 will be discarded for PO vs PT.

PO vs PT

		log2FC	Pval
COLEC11	Collectin-11	-2.79	1.29E-06
IGLV4-69	Immunoglobulin lambda variable 4-69	-2.75	7.12E-03
CAMP	Cathelicidin antimicrobial peptide	-2.28	3.86E-04
ADIPOQ	Adiponectin	-2.19	1.14E-04
IGFBP6	Insulin-like growth factor-binding protein 6	-2.18	1.88E-05
CD44	CD44 antigen (CDw44) (Epican)	-2.10	1.76E-04
YWHAZ	14-3-3 protein zeta/delta	-2.06	3.99E-03
SPP2	Secreted phosphoprotein 24	-2.04	2.74E-07
PZP	Pregnancy zone protein	-1.76	3.21E-03
IGFBP3	Insulin-like growth factor-binding protein 3	-1.30	4.44E-05
SHBG	Sex hormone-binding globulin	-1.22	2.70E-03
APOA1	Apolipoprotein A-I	-1.22	8.80E-04
QSOX1	Sulfhydryl oxidase 1	-1.05	1.45E-05
C4BPB	C4b-binding protein beta chain	-0.97	1.74E-03
IGFALS	Insulin-like GF-binding prot.comp. ALS	-0.91	2.71E-03
PLTP	Phospholipid transfer protein	-0.90	4.82E-04
CD14	Monocyte differentiation antigen CD14	-0.83	2.21E-04
LUM	Lumican (Keratan sulfate proteoglycan lumican)	-0.82	1.88E-03
CFP	Properdin (Complement factor P)	-0.82	2.31E-05
BCHE	Cholinesterase	-0.77	4.76E-04
MASP1	Mannan-binding lectin serine protease 1	-0.73	7.53E-04
TGFBI	Transforming GF-beta-induced ig-h3	-0.71	6.36E-03
F11	Coagulation factor XI (FXI)	-0.69	3.26E-05
C1RL	Complement C1r subcomponent-like protein	-0.55	6.12E-03
C8B	Complement component C8 beta chain	-0.52	3.58E-03
FN1	Fibronectin	-0.52	5.97E-03
LCAT	Phosphatidylcholine-sterol acyltransferase	-0.46	5.70E-03
F2	Prothrombin (Coagulation factor II)	0.45	3.85E-03
ORM2	Alpha-1-acid glycoprotein 2	1.32	1.92E-03
SAA4	Serum amyloid A-4 protein	1.67	1.87E-03
P0DOX3	Immunoglobulin delta heavy chain	2.87	3.07E-03

		log2FC	Pval
IGHG3	Immunoglobulin heavy constant gamma 3	-5.50	1.46E-02
MBL2	Mannose-binding protein C	-2.07	7.74E-03
P0DOX7	Immunoglobulin kappa light chain	-1.57	1.36E-02
THBS1	Thrombospondin-1 (Glycoprotein G)	-1.56	4.27E-02
IGHV5-51	Immunoglobulin heavy variable 5-51	-1.32	2.99E-02
TLN1	Talin-1	-1.25	1.16E-02
VCL	Vinculin (Metavinculin)	-1.23	2.67E-02
FERMT3	Fermitin family homolog 3	-1.02	4.91E-02
P0DOX5	Immunoglobulin gamma-1 heavy chain	-0.91	2.59E-02
PIGR	Polymeric immunoglobulin receptor (PIgR)	-0.84	3.17E-02
ECM1	Extracellular matrix protein 1	-0.81	2.23E-02
SOD1	Superoxide dismutase [Cu-Zn]	-0.81	3.65E-02
AGT	Angiotensinogen (Serpin A8)	-0.76	9.08E-03
F13B	Coagulation factor XIII B chain	-0.69	2.10E-02
IGF2	Insulin-like growth factor II	-0.61	1.25E-02
GPLD1	Phosphatidylinositol-glycan-spec PPlipase D	-0.60	1.95E-02
PROS1	Vitamin K-dependent protein S	-0.59	1.25E-02
LBP	Lipopolysaccharide-binding protein (LBP)	-0.55	4.89E-02
APOL1	Apolipoprotein L1	-0.54	2.19E-02
KLKB1	Plasma kallikrein (Fletcher factor)	-0.51	1.33E-02
PROC	Vitamin K-dependent protein C	-0.51	3.91E-02
A1BG	Alpha-1B-glycoprotein	-0.50	1.77E-02
CLEC3B	Tetranectin	-0.48	3.69E-02
MASP2	Mannan-binding lectin serine protease 2	-0.45	2.41E-02
HPX	Hemopexin (Beta-1B-glycoprotein)	-0.43	9.84E-03
SERPINA4	Kallistatin (Kallikrein inhibitor)	-0.39	3.80E-02
C3	Complement C3	0.33	2.84E-02
C4B	Complement C4-B	0.82	1.89E-02
GPX3	Glutathione peroxidase 3	1.11	2.06E-02

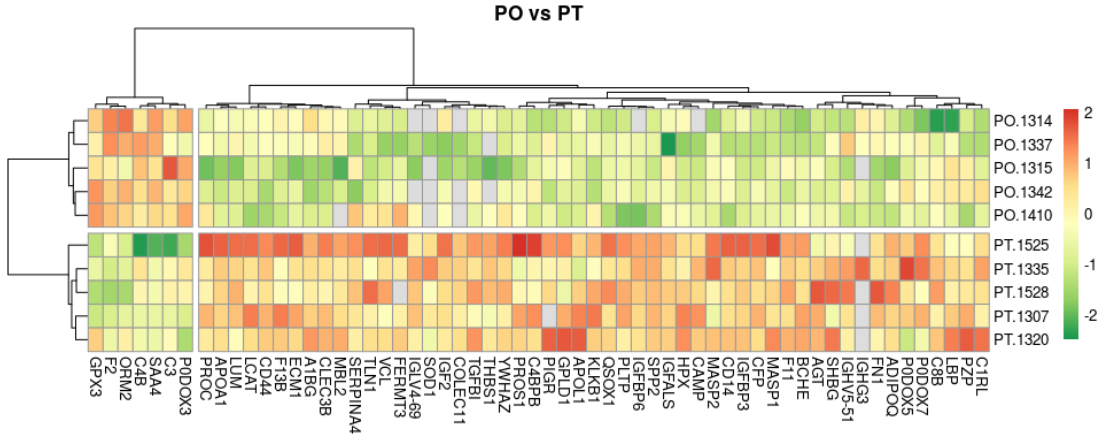


Figure 4.22 61 differentially expressed proteins in PCOS obese (PO) vs PCOS thin (PT) comparison: list with the log2 Fold Changes and P-values (up) and a hierarchical cluster of the individual Log intensities (bottom), clustering both samples (rows) and proteins (columns).

In Table 4.10, the enrichment analysis for this comparison is shown. In the same way that happened in HO vs HT, PT vs HT and PO vs HO, the main categories of Biological Process GO enrichment are coagulation, acute phase and complement (here classical and alternative pathways).

PO vs PT

GO Term	XD-score	q-value	n/N
regulation of complement activation	2.333	4.00E-05	4/11
peptidyl-glutamic acid carboxylation	1.669	2.05E-03	3/10
complement activation, alternative pathway	1.515	2.53E-03	3/11
complement activation	1.287	0.00E+00	6/22
complement activation, classical pathway	0.969	0.00E+00	6/29
fibrinolysis	0.914	8.15E-03	3/18
negative regulation of fibrinolysis	0.769	5.75E-02	2/10
regulation of blood coagulation	0.769	5.75E-02	2/10
positive regulation of blood coagulation	0.769	5.75E-02	2/10
positive regulation of tumor necrosis factor biosynthetic process	0.769	5.75E-02	2/10
positive regulation of phagocytosis	0.734	1.20E-01	2/17
negative regulation of growth of symbiont in host	0.719	6.50E-03	3/16
positive regulation of cytokine production	0.696	6.36E-02	2/11
blood coagulation, intrinsic pathway	0.675	7.30E-03	3/17
acute-phase response	0.614	6.00E-05	5/31
positive regulation vascular endothelial growth factor production	0.612	8.28E-01	1/14
positive regulation of reactive oxygen species metabolic process	0.601	8.86E-03	3/19

Term	Description	q-value	n	N	Genes
Pubmed:23864804	Adipokines, insulin-like growth factor binding protein-3 levels, and insulin sensitivity in women with polycystic ovary syndrome.	3.44E-03	2	3	ADIPOQ, IGFBP3
Pubmed:20200332	Family-based analysis of candidate genes for polycystic ovary syndrome.	3.26E-02	2	27	ADIPOQ, SHBG
Pubmed:29374985	Preptin in women with polycystic ovary syndrome.	3.26E-02	1	1	IGF2
Pubmed:16275260	The M235T polymorphism of the angiotensinogen gene in women with polycystic ovary syndrome.	3.26E-02	1	1	AGT
Pubmed:28294594	Association of kallistatin with carotid intima-media thickness in women with polycystic ovary syndrome.	3.26E-02	1	1	SERPINA 4
Pubmed:21178921	Sex hormone-binding globulin genetic variation: associations with type 2 diabetes mellitus and polycystic ovary syndrome.	3.26E-02	1	1	SHBG
Pubmed:23415973	Faster thrombin generation in women with polycystic ovary syndrome compared with healthy controls matched for age and body mass index.	3.26E-02	1	1	F2
Pubmed:18206145	Increased acylation-stimulating protein, C-reactive protein, and lipid levels in young women with polycystic ovary syndrome.	4.40E-02	1	2	C3
Pubmed:28789706	Small leucine-rich proteoglycans (SLRPs) in the endometrium of polycystic ovary syndrome women: a pilot study.	4.40E-02	1	2	LUM
Pubmed:17154366	Thrombospondin-1 inhibits VEGF levels in the ovary directly by binding and internalization via the low density lipoprotein receptor-related protein-1 (LRP-1).	4.40E-02	1	2	THBS1
Pubmed:19408174	Serum resistin and adiponectin levels in women with polycystic ovary syndrome.	4.40E-02	1	2	ADIPOQ
Pubmed:18192296	The role of sex hormone-binding globulin and androgen receptor gene variants in the development of polycystic ovary syndrome.	4.40E-02	1	2	SHBG
Pubmed:24336678	Increased expression of kindlin 2 in luteinized granulosa cells correlates with androgen receptor level in patients with polycystic ovary syndrome having hyperandrogenemia.	4.40E-02	1	2	FERMT3
MP:0002471	abnormal complement pathway	1.67E-07	6	20	
MP:0005166	decreased susceptibility to injury	1.63E-06	10	169	
MP:0005048	abnormal thrombosis	7.64E-04	7	122	
MP:0005464	abnormal platelet physiology	9.51E-04	7	126	
MP:0004042	decreased susceptibility to kidney reperfusion injury	1.18E-03	4	19	
CID006450878	Etiocobalamin	1.95E-15	15	177	
CID000040973	desogestrel	1.18E-09	10	106	
ctd:C076029	olanzapine	4.57E-09	10	121	
C0019243	Angioedemas, Hereditary	2.82E-05	5	23	
C0085096	Peripheral Vascular Diseases	1.77E-04	6	64	
C0035222	Respiratory Distress Syndrome, Adult	3.18E-04	8	176	
C1704430	Urinary Schistosomiasis	3.37E-04	3	4	
C0040053	Thrombosis	9.65E-04	5	45	

Table 4.10 Enriched terms for PT vs HO. On the left, GO Biological Process terms analyzed with JEPETTO. On the right, Pubmed entries, Human Phenotype (HP), Disease terms (DisGeNET), Mouse Phenotype terms (MP) and Stitch drug database (CID) from Toppgene.

4.4.2 Comparative functional analysis of HO vs HT, PT vs HT and PO vs HT

Individual enrichment analyses performed so far have been useful to characterize the different groups compared. A different approach of GO Biological Process enrichment is applied here. Using ClueGO (61) Cytoscape plugin, the enrichment of the three comparisons with healthy controls (HO vs HT, PT vs HT and PO vs HT) will be done in a coordinated way here.

ClueGO performs its enrichment using two very useful approaches for this situation:

- ClueGO groups highly related GO terms (by the proteins they include), simplifying and reducing the number of enriched categories, producing a subset of “overview” categories that can be used to summarize the biological properties obtained in the enrichment.

- It also performs a comparative enrichment analysis, assigning GO terms to one of the three comparisons analyzed, depending on the number of proteins associated to each of them. Some GO terms will not be associated with any of the three groups, because the relative amount of proteins associated to any of them is not high enough to associate one comparison with the GO term.

The three categories under study here present an intense overlapping of the proteins being differentially expressed on each of them. In Figure 4.23, a graphical representation of this overlap and the main enrichment categories obtained for the three groups.

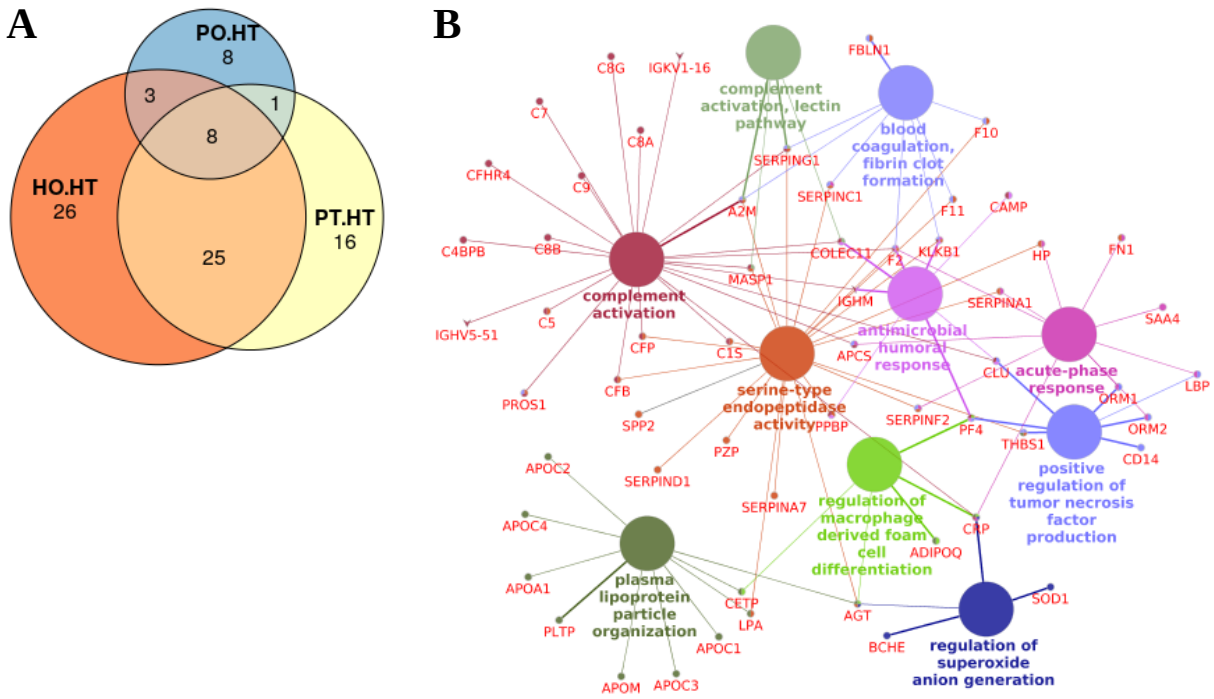


Figure 4.23 (A) Euler-Venn diagram with the numbers of the proteins belonging to HO vs HT, PT vs HT and PO vs HT comparisons. **(B)** ClueGO analysis of the three comparisons produces a set of interconnected GO categories.

Results for ClueGO enrichment are shown in Table 4.11. For a correct interpretation of those results, it is important to keep in mind that a different number of proteins have been introduced for each category: 62 proteins for HO vs HT, 50 for PT vs HT and only 20 for HO vs HT. This fact will make a lot more likely that a given protein from the HO vs HT group will be associated to a GO term than a protein from the PO vs HT comparison, just because the great overlap of proteins between comparisons. Despite this fact, that can not be prevented, the percentage of proteins inside a comparison associated to a given GO category will give an idea on how much related are proteins from a given comparison (for example HO vs HT) to some GO term (for example “complement activation”). From Table 4.11, several conclusions:

- Several terms are found where proteins are evenly represented for the three comparisons: “Complement activation, lectin pathway”, “Regulation of superoxide anion generation” and “Plasma lipoprotein particle organization” are some examples.
- Some terms are related in greater proportion to HO vs HT than the other comparisons. Examples are “positive regulation of tumor necrosis factor

production", "acute-phase response" and the classical and alternative pathways of the Complement activation.

- Only one term is obtained where PT vs HT shows preferential association: "triglyceride homeostasis".
- Not a single GO term has been found preferentially related to PO vs HT .

GOID	GOTerm	GO-Groups	Term Corr-PValue	Group Corr-PValue	% Associated Genes	Nr. Genes	Cluster	%Genes Cluster HOvsHT	%Genes Cluster PTvsHT	%Genes Cluster POvsHT
GO:0001867	**complement activation, lectin pathway**	0	9.50E-06	3.96E-06	25.00	4	None	53.43	35.62	35.62
GO:0010743	**regulation of macrophage derived foam cell diff.**	1	1.53E-05	5.10E-06	11.63	5	None	53.89	17.96	35.93
GO:0032928	**regulation of superoxide anion generation**	3	7.38E-05	1.48E-05	12.50	4	None	53.43	53.43	17.81
GO:0019731	antibacterial humoral response	6	2.82E-03	3.95E-06	5.00	3	None	61.31	61.31	0.00
GO:0097006	regulation of plasma lipoprotein particle levels	8	3.04E-10	1.69E-10	8.80	11	None	41.30	49.56	24.78
GO:0071827	**plasma lipoprotein particle organization**	8	3.56E-11	1.69E-10	13.70	10	None	44.89	53.86	17.95
GO:0034377	plasma lipoprotein particle assembly	8	1.32E-06	1.69E-10	12.50	6	None	46.04	46.04	15.35
GO:0034375	high-density lipoprotein particle remodeling	8	1.33E-10	1.69E-10	31.82	7	None	53.47	40.11	13.37
GO:0032374	regulation of cholesterol transport	8	2.24E-07	1.69E-10	11.48	7	None	42.86	42.86	14.29
GO:0032760	**positive regulation of tumor necrosis factor production**	2	1.19E-05	4.32E-06	6.14	7	**HO vs HT**	69.56	46.37	11.59
GO:0002526	acute inflammatory response	4	2.67E-11	8.55E-12	7.69	13	**HO vs HT**	75.51	50.34	18.88
GO:0006953	**acute-phase response**	4	8.91E-12	8.55E-12	12.36	11	**HO vs HT**	72.77	50.94	21.83
GO:0004867	serine-type endopeptidase inhibitor activity	5	5.79E-09	1.06E-19	8.06	10	**HO vs HT**	60.00	30.00	10.00
GO:0061134	peptidase regulator activity	5	7.18E-12	1.06E-19	5.63	16	**HO vs HT**	65.43	43.62	21.81
GO:0004252	**serine-type endopeptidase activity**	5	1.12E-16	1.06E-19	6.23	21	**HO vs HT**	69.61	43.51	8.70
GO:0004866	endopeptidase inhibitor activity	5	7.19E-11	1.06E-19	6.11	14	**HO vs HT**	64.02	38.41	19.21
GO:0019730	**antimicrobial humoral response**	6	1.40E-05	3.95E-06	4.68	8	**HO vs HT**	68.95	59.10	9.85
GO:0061844	antimicrobial humoral immune resp.med.by.peptide	6	3.53E-04	3.95E-06	4.50	5	**HO vs HT**	77.40	46.44	0.00
GO:0051873	killing by host of symbiont cells	6	4.34E-04	3.95E-06	13.64	3	**HO vs HT**	68.26	45.51	0.00
GO:0006959	humoral immune response	7	4.95E-21	1.55E-21	5.81	27	**HO vs HT**	78.38	39.19	19.60
GO:0006956	**complement activation**	7	1.48E-23	1.55E-21	10.50	23	**HO vs HT**	76.08	38.04	22.83
GO:0006957	complement activation, alternative pathway	7	2.00E-13	1.55E-21	44.44	8	**HO vs HT**	86.72	21.68	10.84
GO:0006958	complement activation, classical pathway	7	1.14E-13	1.55E-21	8.47	15	**HO vs HT**	75.93	29.20	23.36
GO:0030449	regulation of complement activation	7	1.06E-18	1.55E-21	12.88	17	**HO vs HT**	79.73	31.89	10.63
GO:1905953	negative regulation of lipid localization	8	7.46E-05	1.69E-10	8.06	5	**HO vs HT**	71.86	17.96	17.96
GO:0072378	**blood coagulation, fibrin clot formation**	9	1.29E-10	3.28E-09	21.62	8	**HO vs HT**	60.80	50.67	20.27
GO:0007597	blood coagulation, intrinsic pathway	9	2.52E-08	3.28E-09	24.00	6	**HO vs HT**	67.96	40.77	13.59
GO:0061041	regulation of wound healing	9	5.12E-06	3.28E-09	4.52	9	**HO vs HT**	81.14	45.08	9.02
GO:0042730	fibrinolysis	9	4.52E-09	3.28E-09	20.00	7	**HO vs HT**	81.15	46.37	0.00
GO:0051918	negative regulation of fibrinolysis	9	1.11E-04	3.28E-09	23.08	3	**HO vs HT**	79.25	26.42	0.00
GO:0051917	regulation of fibrinolysis	9	5.85E-07	3.28E-09	23.81	5	**HO vs HT**	77.40	46.44	0.00
GO:0031639	plasminogen activation	9	6.80E-05	3.28E-09	13.33	4	**HO vs HT**	77.37	38.69	0.00
GO:0070328	triglyceride homeostasis	8	2.92E-05	1.69E-10	10.00	5	**PT vs HT**	40.00	60.00	0.00

Table 4.11 ClueGO analysis results for HO vs HT, PT vs HT and PO vs HT. Ten different GO groups have been found (0 to 9). GO terms in bold font represent "overview" terms, and can be used to represent the rest of GO terms inside a given group. P-values (Bonferroni corrected) are calculated for each term and for the whole group. Also, a total number of genes and % of associated genes are shown for each GO term. The "Cluster" column assigns each GO term to a given comparison where is more represented (HO vs HT, PT vs HT or None). The three last columns provide percentages of proteins of each comparison inside one GO term.

The obtained results from ClueGO showed that the three groups of proteins compared are mainly related to the same GO terms. That is, the fact that the same set of GO terms is shared among HO, PT and PO indicates that the comparison with more proteins will have more associated enriched terms as is showed in table 4.11. Only in terms like "negative regulation of fibrinolysis", where the ratio for HO and PT was 79%-26%, some real preeminence of HO over PT can be suggested.

Besides the fact that not real differential enrichment has been obtained among the phenotypes studied, ClueGO has provided a useful overview of the GO enriched categories. Therefore, this analysis summarizes the biological processes involved in all experimental groups analyzed, preventing misleading differentiation criteria between phenotypes if only individual comparisons had been done between them.

4.4.3 Pathways analysis of HO vs HT, PT vs HT and PO vs HT

A pathways analysis has been done with protein levels in HO vs HT, PT vs HT and PO vs HT comparisons, using KEGG (62) pathway database. For this study, all proteins and not only those with P-values under 0.05 have been used. In this way, general trends can be observed more easily but, on the other hand, caution must be taken with fold changes that are not backed by a significant P-value. It is also true that proteins showing large fold changes will be frequently associated to significant P-values (the larger the fold change, the more usual), making easier the interpretation of the pathways analysis.

First, an enrichment study has been performed with the 204 proteins expressed using Toppgene (Table 4.12). However, some considerations should be made here:

- A pathway enrichment is very useful in some circumstances, but in this case, some limitations introduce a big bias into the results, the most important being that this study is based on plasma protein levels. That makes obvious that only pathways with a significant involvement in the extracellular space can be conveniently enriched.
- Several diseases appear enriched as can be seen in Table 4.12. The reason, as already seen for several comparisons in the previous sections, is that the immune system is one of the affected biological processes. This immune component is the responsible for the enrichment of pathways related to diseases.
- In some pathways, the same protein appears several times interacting with different elements. Thus, it will be useful to explore pathways that are not enriched or show few proteins mapped to the data set under study.

ID	Name	p-value	q-value Bonferroni	n	N
83073	Complement and coagulation cascades	3.30E-73	2.94E-71	48	79
172846	Staphylococcus aureus infection	6.24E-18	5.56E-16	16	56
101144	Prion diseases	1.21E-11	1.08E-09	10	35
218111	Pertussis	1.90E-10	1.69E-08	12	76
83122	Systemic lupus erythematosus	1.43E-09	1.27E-07	14	133
194384	African trypanosomiasis	7.60E-05	6.76E-03	5	35
167324	Amoebiasis	2.23E-04	1.99E-02	7	96
83042	PPAR signaling pathway	2.24E-03	2.00E-01	5	72
199556	Vitamin digestion and absorption	3.40E-03	3.02E-01	3	24

Table 4.12 Toppgene enrichment for KEGG pathways.

From the enrichment shown in Table 4.12, only “Complement and coagulation cascades” and “PPAR signaling pathway” have some interest in this study.

For a thorough pathway study, instead of an enrichment, a different approach has been followed: using the web application “KEGG Mapper”, the complete list of proteins has been introduced and 111 proteins have been mapped to one or more KEGG pathways. The list of some of the pathways obtained is shown in Table 4.13.

Kegg pathway	Description / (number of proteins)
hsa04610	Complement and coagulation cascades (48)
hsa04979	Cholesterol metabolism (11)
hsa04611	Platelet activation (6)
hsa04080	Neuroactive ligand-receptor interaction (6)
hsa04145	Phagosome (6)
hsa04510	Focal adhesion (6)
hsa04810	Regulation of actin cytoskeleton (6)
hsa03320	PPAR signaling pathway (5)
hsa04151	PI3K-Akt signaling pathway (5)
hsa04918	Thyroid hormone synthesis (4)
hsa04512	ECM-receptor interaction (4)
hsa04015	Rap1 signaling pathway (3)
hsa04977	Vitamin digestion and absorption (3)
hsa04010	MAPK signaling pathway (3)
hsa04072	Phospholipase D signaling pathway (2)
hsa04064	NF-kappa B signaling pathway (2)
hsa04670	Leukocyte transendothelial migration (2)
hsa04216	Ferroptosis (2)
hsa05130	Pathogenic Escherichia coli infection (2)
hsa04062	Chemokine signaling pathway (2)
hsa04620	Toll-like receptor signaling pathway (2)
hsa04640	Hematopoietic cell lineage (2)
hsa04115	p53 signaling pathway (2)
hsa04060	Cytokine-cytokine receptor interaction (2)
hsa04970	Salivary secretion (2)
hsa04514	Cell adhesion molecules (CAMs) (2)
hsa04975	Fat digestion and absorption (2)

Table 4.13 List of KEGG pathways where some of the proteins related to this work are mapped. The number of proteins inside a pathway is shown between parentheses.

After the list of pathways mapped by KEGG has been filtered, discarding members not related to this work (for example, dropping diseases), the Pathview (63) Bioconductor package was used to map protein expression to pathways using the pathway ID's provided by KEGG Mapper (Figure 4.25).

```
library(pathview)
pathways<-c("hsa04610","hsa04979","hsa04611","hsa04080","hsa04145","hsa04510","hsa04810","hsa03320",
            "hsa04620","hsa04640","hsa04933","hsa04115","hsa04060","hsa04970","hsa04514","hsa04975")
for (pathway in pathways) {
  pv.out <- pathview(gene.data = data.to.kegg, pathway.id = pathway,multi.state = TRUE,
                     same.layer = TRUE, match.data = TRUE, kegg.native = TRUE,
                     limit=list(gene=1), bins=list(gene=10), mid ="gray")
}
```

Figure 4.25 Pathview code where sixteen pathways, iterating in a loop using a list, are matched with expression values contained in a text file (imported as a data frame named "data.to.kegg"). The number of bins, 10 in this case, are important to highlight differences between different expression levels. Here, 10 bins are chosen, leaving the center in gray color, which gives an interval of -0.25 to + 0.25 as not quantified, matching the ±0.26 Log2 fold change used as the cutoff used in this work.

In this pathway mapping, the expression values (without considering P-values) of the three comparisons against the HT samples: HO vs HT, PT vs HT and PO vs HT. For every protein found in the list that is provided, the software will split the area of the represented entity into three spaces: the first for HO, the one in the middle for PT and the one on the right for PO. In this way, the variation of protein levels can be compared between the three groups in an easy and convenient way. The interval of fold changes inspected has been -1 to +1:

- Proteins for a given comparison within ±0.26 Log2 fold change will remain gray.
- Proteins with Log2 fold change higher than +0.26 will be increasingly red.

- Proteins with Log2 fold change under -0.26 will be green.
- Over +1 and under -1 Log2 fold change, no changes will be seen: two proteins with a +1 and a +2 Log2 fold change will be seen exactly in the same way.

After automatically generating those pathways with the quantitative information and inspecting the results obtained, two of them have been found specially illustrative for the changes produced by HO, PT and PO with respect to HT:

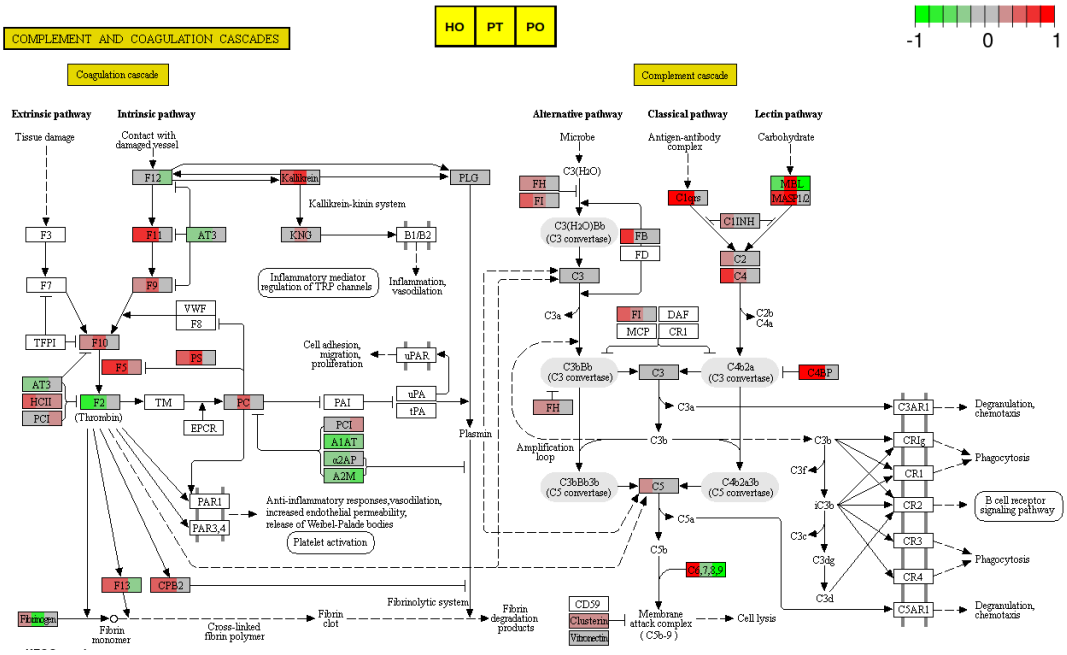


Figure 4.26 Complement and coagulation cascades (hsa04610) KEGG pathway. Proteins are split in three areas, with left corresponding to HO vs HT, middle to PT vs HT and right to PO vs HT. Proteins not quantified are left in white. Proteins under ±0.25 Log2 fold change in gray. Over-expression in red (the higher the fold change, the darker) and under-expression in green.

1. **Complement and coagulation cascades** (Figure 4.26): an overall activation is seen throughout the whole pathway. It is important to observe than in KEGG terminology, the point of an arrow means activation and a a line ending in a flat extreme (like the one joining AT3 and F11 in the coagulation cascade, intrinsic pathway) means inhibition. Some interesting differences can be spotted, but caution should be applied before interpreting the results seen here. For example, in the complement cascade, lectin pathway, MBL is under-expressed both for HO and PO, while over-expressed in PT. This protein corresponds to MBL2 in the lists used in this work (as can be confirmed checking into KEGG webpage). In the three cases, the darkest colors are used, meaning that Log2 fold changes close or higher to ±1 have been obtained. But taking a look into the results obtained, the protein MBL has passed the cut-offs only for PTvsHO and POvsPT: the values seen for this protein here (for HO vs HT, PT vs HT and PO vs HT) cannot be completely trusted, and used only as an orientation on what is happening.
2. **Cholesterol metabolism** (Figure 4.27): in this pathway, almost all the proteins found at the extracellular space are quantified. It makes a lot of sense that the ones exclusively found into the hepatocyte (drawn in the middle), are not mapped: proteins were isolated from plasma. An interesting case arises here: protein ApoC, related to HDL cholesterol, is shown highly over-expressed in the three

comparisons. But the mapping is not simple here: ApoC levels can be evaluated from APOC1, APOC2, APOC3 and APOC4 levels, all of them being among the proteins quantified in this work. All possible combinations of APOC proteins should be inspected before arriving to conclusions.

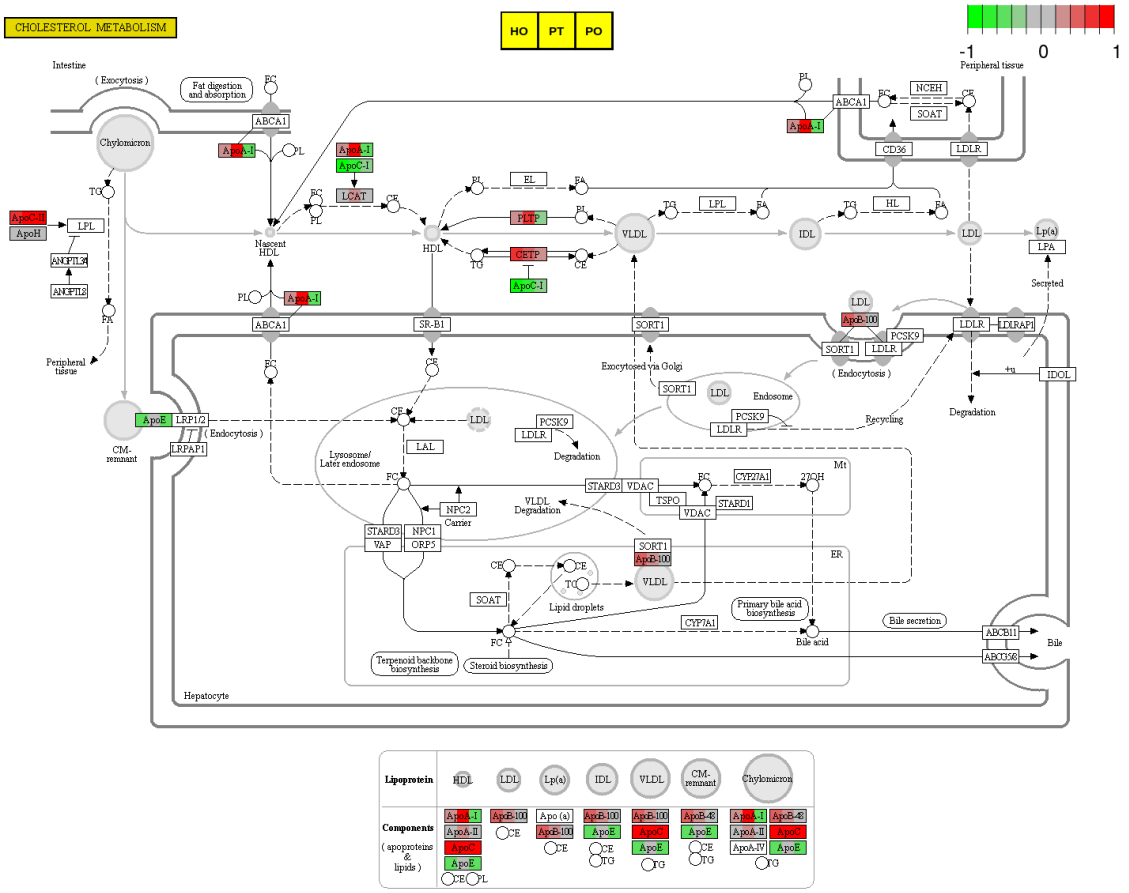


Figure 4.27 Cholesterol metabolism (hsa04979) KEGG pathway. Proteins are split in three areas, with left corresponding to HO vs HT, middle to PT vs HT and right to PO vs HT. Proteins not quantified are left in white. Proteins under ±0.25 Log2 fold change in gray. Over-expression in red (the higher the fold change, the darker) and under-expression in green.

4.5 Discussion

The results shown in the previous sections, both in what has been called “clinical variables” (“4.3.1 Phenotypes under study”) and later with protein levels found in the different phenotypes in the results section (“4.4 Results”) confirm that BMI has a direct impact on the physiological response in subjects with or without a PCOS diagnose. Furthermore, different sets of proteins have been found differentially expressed at a statistically significant level in the four groups created (HT, HO, PT and PO), showing that the generation of these groups makes sense when proteostasis studies are made in PCOS patients. Despite the fact that more phenotypical variables can be in play, the approach used here may be essential to understand changes in protein levels among PCOS patients. It is certain that more variables could be used to stratify patients when studying PCOS proteomics, but a BMI patient stratification has demonstrated its usefulness for classifying both using clinical variables (Figure 4.2) and protein levels (cluster Figure 4.15) with the different phenotypes studied. Obtaining more clinical variables for classification or

establishing a numerical threshold for BMI would require a higher number of samples from each category (PCOS diagnosed and not) than the 20 sample used in this work.

The data independent proteomics (Swath) analysis has shown high stability in measurements, providing in some cases more than 60 differentially expressed proteins out of 204 proteins quantified. Only two proteins, IGHG3 and C1QA, have been detected in less than ten samples, while 169 proteins have been detected in all 20 samples. This has made unnecessary any imputation approach (that is, assigning mathematically obtained values to undetected proteins).

4.5.1 Overall effects of PCOS on protein levels

The way in which protein levels in plasma are changed by the effect of PCOS is summarized in “4.4.1.1 PCOS vs H” and more clearly in “4.4.1.2 PCOS vs HT” section. The reason for using HT samples as reference is that they provide a clearer picture of the changes originated by PCOS, both in terms of more extreme changes (more intense fold changes) and number of proteins differentially expressed (almost double the proteins found in the later): in PCOS vs HT comparison, the interference introduced by the HO samples is not present. As will be demonstrated in the next point, there are many similarities in the changes introduced at protein levels in plasma by both PCOS and obesity.

Despite the fact that in PCOS vs H an interference is introduced by HO samples, this is a very useful comparison because, comparing the proteins with altered levels in PCOS vs H with the ones found in PCOS vs HT, a short list of proteins show consistent levels in both groups. As shown in Figure 4.16, five proteins (FLNA, ADIPOQ, LBP, RBP4 and APOC2) show the same direction in terms of differential expression (up or down-regulation) and similar intensities in both comparisons. That means that, despite the fact that a high BMI samples (HO) introduce similar changes to the ones found in PCOS, these five proteins remain highly indifferent to this effect. This would apparently make those five proteins as strong markers for PCOS in the samples analyzed: actually, low levels of ADIPOQ have been associated to PCOS (64), while high levels have been in the case of LBP (65) and RBP4 (66).

However, several considerations should be made here:

- LBP shows also high over-expression in HO vs HT, therefore making of it a good marker for PCOS only in lean patients.
- With FLNA, not significant results have been found in HO vs HT, but is somewhat under-expressed in that comparison and therefore, could interfere with PCOS diagnose. Something similar happens with APOC2.
- ADIPOQ, as will be seen later, shows an important under-expression for PCOS with obesity, but not for lean PCOS patients, where remains unaltered.
- The only protein that remains as a potential marker for PCOS in all combinations examined is RBP4: this protein has been found differentially over-expressed in PCOS vs HT, PCOS vs H, PT vs HT and PO vs HT with a Log2 fold change close to +0.6 for all these combinations. Furthermore, it also shows similar levels of fold change for PO vs HO and PT vs HO, though with P-values over 0.05.

For assessing the qualities of any of the five mentioned proteins as biomarkers, a big cohort of samples of the same phenotypes should be analyzed specifically for those proteins.

Lastly, from the list of proteins differentially expressed in PCOS vs HT, several proteins (ADIPOQ, RBP4, F2, F5 among them) are well described in literature with respect to their changes in plasma levels of PCOS patients. Other proteins, like CAMP (involved in chronic inflammatory response) or PRG4 (related to immune response) have been less associated with this disease and a lot more with obesity.

4.5.2 Similarities and differences between obesity and PCOS effects

The relationship between PCOS and obesity has been widely described before, because of their high co-occurring rates (3), or by the multiple underlying mechanisms linking the two conditions (67). In this work, the main similarities at protein levels can be inspected from the similarities between HO vs HT and PT vs HT comparisons. Such similarities are shown on Table 4.13, where all the proteins that showed differential expression in those comparisons (P-value<0.05), shown exactly the same general trends (is they are up or down-regulated), and quite similar fold changes. This can only be explained if very similar biological processes are taking effect. This fact is confirmed by the enrichment results shown in Table 4.11, where the high overlapping among HO vs HT and PT vs HT was assessed.

Without analyzing here the causes, it is clear that the consequences (effects at protein level) are very similar both in PT and HO samples. Effects on biological processes such complement activation (68,69), lipoprotein synthesis (70,71), inflammatory response (72,73) and coagulation (74,75) have been described both within PCOS and obesity, so it is not strange that these same processes are affected in this study. But the results shown in Table 4.14 demonstrate how close those effects are in quantitative terms.

		HOvsHT		PTvsHT	
		log2FC	pvalue	log2FC	pvalue
C7	Complement component C7	-0.44	1.14E-03	-0.38	3.67E-03
COLEC11	Collectin-11	1.46	1.48E-04	1.61	6.42E-05
SPP2	Secreted phosphoprotein 24	0.55	3.75E-02	1.24	1.03E-04
LPA	Apolipoprotein(a)	2.67	1.27E-03	1.94	1.19E-02
APCS	Serum amyloid P-component (SAP)	1.01	4.98E-04	0.67	1.07E-02
PRG4	Proteoglycan 4 (Lubricin)	1.30	6.22E-04	1.20	1.31E-03
FBLN1	Fibulin-1 (FIBL-1)	-0.75	2.12E-03	-0.57	1.31E-02
LBP	Lipopolysaccharide-binding protein (LBP)	0.97	1.62E-03	1.19	2.71E-04
F5	Coagulation factor V	0.61	2.47E-03	0.63	1.87E-03
QSOX1	Sulfhydryl oxidase 1	0.80	2.65E-04	0.72	6.94E-04
SERPINC1	Antithrombin-III (ATIII) (Serpin C1)	-0.28	2.57E-03	-0.25	6.36E-03
BCHE	Cholinesterase	0.42	3.05E-02	0.45	2.02E-02
IGFBP6	Insulin-like growth factor-binding protein 6	1.39	8.79E-04	1.72	1.21E-04
CFP	Properdin (Complement factor P)	0.62	4.09E-04	0.64	3.00E-04
YWHAZ	14-3-3 protein zeta/delta	1.38	3.92E-02	1.34	4.52E-02
CD44	CD44 antigen (CDw44) (Epican)	1.24	1.12E-02	1.59	2.07E-03
IGHM	Immunoglobulin heavy constant mu	-2.40	2.57E-02	-2.52	2.00E-02
SERPINF2	Alpha-2-antiplasmin	-0.27	8.12E-03	-0.25	1.53E-02
CAMP	Cathelicidin antimicrobial peptide	1.82	1.58E-03	2.78	3.12E-05
TGFBI	Transforming GF-beta-induced prot.ig-h3	0.50	4.08E-02	0.52	3.46E-02
GPX3	Glutathione peroxidase 3	-0.99	3.60E-02	-1.45	3.98E-03
FN1	Fibronectin	0.52	5.57E-03	0.39	2.88E-02
PROS1	Vitamin K-dependent protein S	0.59	1.20E-02	0.74	2.77E-03
F2	Prothrombin (Coagulation factor II)	-0.72	5.72E-05	-0.53	1.05E-03
KLKB1	Plasma kallikrein (Fletcher factor)	0.48	1.86E-02	0.61	3.95E-03
TLN1	Talin-1	1.49	3.80E-03	1.01	3.58E-02
ORM2	Alpha-1-acid glycoprotein 2	-1.26	2.68E-03	-1.17	4.56E-03
SOD1	Superoxide dismutase [Cu-Zn]	0.90	2.36E-02	0.96	1.72E-02
F11	Coagulation factor XI (FXI)	0.72	2.20E-05	0.66	5.74E-05
MASP1	Mannan-binding lectin serine protease 1	0.51	1.04E-02	0.68	1.30E-03
CD14	Monocyte differentiation antigen CD14	0.73	7.14E-04	0.88	1.33E-04
IGLV4-69	Immunoglobulin lambda variable 4-69	2.64	3.53E-03	2.90	1.78E-03
C4BPB	C4b-binding protein beta chain	0.58	4.00E-02	0.94	2.33E-03
PIGR	Polymeric immunoglobulin receptor (PIgR)	1.38	9.09E-04	0.80	4.05E-02
SAA4	Serum amyloid A-4 protein	-1.40	6.77E-03	-1.67	1.87E-03

Table 4.14 Similarities between HOvsHT and PTvsHT: all proteins differentially expressed in common in HO and PT (35 proteins) present roughly the same expression changes with respect to healthy thin controls (log2 fold change changing in the same direction in all cases).

Differences in protein levels between obesity and PCOS conditions can be easily spotted using PT vs HO comparison (Figure 4.21). Only 15 proteins appear as differentially expressed, half of them showing mild variations (between ±0.5).

Two proteins appear highly under-expressed in PT samples with respect to HO: CRP and P0DOX3. CRP levels have been consistently associated with obesity (76), while in PCOS, it has been shown the opposite: CRP levels are not affected in lean PCOS patients (77). On the other hand, P0DOX3 levels have not been associated with obesity or PCOS (to our knowledge).

Additionally, several proteins appear up-regulated in PT samples with respect to HO: MBL2 (complement activation), SHBG (androgen binding) and SPP2 (coagulation) among others.

4.5.3 Combined effects of PCOS and obesity

Obesity is commonly found in PCOS patients, aggravating many of its reproductive and metabolic symptoms (78). To evaluate the differences between PO and the two closely related HO and PT phenotypes, three comparisons have been used (Table 4.15): the 20 proteins found with a P-value<0.05 for PO vs HT have been aligned with the same proteins found in HO vs HT and PT vs HT, independently of their P-values.

		HOvsHT		PTvsHT		POvsHT	
		log2FC	pvalue	log2FC	pvalue	log2FC	pvalue
ADIPOQ ★	Adiponectin	-0.11	8.06E-01	-0.03	9.50E-01	-2.21	**1.00E-04**
FLNA	Filamin-A	-0.46	3.73E-01	-1.12	5.32E-02	-1.60	**1.48E-02**
IGHV5-51	Immunoglobulin heavy variable 5-51	-0.37	5.11E-01	0.05	9.22E-01	-1.26	**3.64E-02**
COLEC11 ★	Collectin-11	1.46	**1.48E-04**	1.61	**6.42E-05**	-1.18	**3.63E-03**
IGFBP3 ★	Insulin-like growth factor-binding protein 3	0.29	2.35E-01	0.38	1.23E-01	-0.92	**1.23E-03**
SPP2 ★	Secreted phosphoprotein 24	0.55	**3.75E-02**	1.24	**1.03E-04**	-0.80	**4.31E-03**
CD5L	CD5 antigen-like	-0.61	1.02E-01	-0.22	5.35E-01	-0.77	**4.39E-02**
APOD	Apolipoprotein D	-0.63	**3.31E-02**	-0.33	2.35E-01	-0.72	**1.72E-02**
IGFALS ★	Insulin-like growth factor-binding protein complex acid labile subunit (ALS)	0.22	4.08E-01	0.33	2.25E-01	-0.59	**3.66E-02**
FBLN1	Fibulin-1 (FIBL-1)	-0.75	**2.12E-03**	-0.57	**1.31E-02**	-0.58	**1.22E-02**
C7	Complement component C7	-0.44	**1.14E-03**	-0.38	**3.67E-03**	-0.58	**8.49E-05**
A2M	Alpha-2-macroglobulin	-0.34	9.21E-02	-0.22	2.51E-01	-0.49	**1.90E-02**
GSN	Gelsolin	-0.26	1.43E-01	-0.30	8.99E-02	-0.43	**2.28E-02**
APOM	Apolipoprotein M	-0.56	**4.58E-03**	-0.09	5.90E-01	-0.42	**2.49E-02**
RBP4	Retinol-binding protein 4	0.13	6.26E-01	0.59	**3.69E-02**	0.62	**3.10E-02**
LBP	Lipopolysaccharide-binding protein (LBP)	0.97	**1.62E-03**	1.19	**2.71E-04**	0.65	**2.32E-02**
APCS	Serum amyloid P-component (SAP)	1.01	**4.98E-04**	0.67	**1.07E-02**	0.74	**5.50E-03**
PRG4	Proteoglycan 4 (Lubricin)	1.30	**6.22E-04**	1.20	**1.31E-03**	0.97	**6.28E-03**
CRP	C-reactive protein	3.84	**6.57E-04**	0.81	3.49E-01	2.22	**1.86E-02**
LPA	Apolipoprotein(a)	2.67	**1.27E-03**	1.94	**1.19E-02**	2.24	**4.85E-03**

Table 4.15 Levels of proteins differentially expressed in PO vs HT compared to HOvsHT and PTvsHT. Proteins labeled with a star (ADIPOQ, COLEC11, IGFBP3, SPP2 and IGFALS) present completely different levels in PO vs HT (different direction in expression) compared to HO vs HT and PT vs HT. In bold, P-values lower than 0.05.

Five proteins show very different expression levels in PO vs HT compared to the other two phenotypes: ADIPOQ, COLEC11, IGFBP3, SPP2 and IGFALS.

In first place, ADIPOQ shows in PO vs HT a twofold negative change with respect to HO vs HT and PT vs HT. This means that contrary of what has been published (79), where decreased levels of ADIPOQ have been found independent of BMI, in our study the levels of ADIPOQ only decreased for PO subjects, not for PT ones. In the case of RBP4, several studies (80,81) had demonstrated elevated levels of this protein in PCOS patients, independently if they were obese or not; this means that our results correspond to those previously reported, with a nearly exact increment of RBP4 in PT and PO patients.. The fact that ADIPOQ appeared in “4.5.1 Overall effects of PCOS on protein levels” and in Figure 4.15 with a logFC=-1.12 is explained because the high under-expression of ADIPOQ in PO vs HT compensates the null variation of this protein with PT vs HT. In essence, results obtained here contradict BMI-independent under-expression of ADIPOQ found in other works.

LBP, that was introduced previously as one potential marker for PCOS, shows also a high over-expression in HO subjects, what makes it only useful for lean patients.

Additionally, from Table 4.15, several proteins show under-expression under PO and over-expression in HO and PT:

- COLEC11: related to innate immunity (82), apoptosis (83) and embryogenesis (84).
- IGFBP3: with antiproliferative and apoptotic effects (85).
- SPP2: found in association with metabolic disease (86).
- IGFALS: related to the insulin-like growth factor (IGF) system (87).

4.6 Conclusions

Five proteins (FLNA, ADIPOQ, LBP, RBP4 and APOC2) present significant variations between PCOS samples and both H (HT+HO) and HT as controls. Protein RBP4 appears as the most robust marker for PCOS even with interference from obesity.

PCOS and obesity share many traits in common, with at least 35 proteins differentially expressed in both conditions showing virtually the same levels. Proteins related to complement show a general under-expression in PT with respect with HO (FGB, CFB, C3, C8A, C8B and C8G) with the exception of MBL2 that is clearly up-regulated in PT vs HO. Two proteins appear highly under-expressed in PT samples with respect to HO: CRP and P0DOX3, while SHBG (androgen binding) and SPP2 (coagulation) are down-regulated.

Finally, related to the combined effects of PCOS and obesity, five proteins (ADIPOQ, COLEC11, IGFBP3, SPP2 and IGFALS) appear down-regulated, as opposed to what happened in lean PCOS subjects.

4.7 References

1. Mohammad MB, Seghinsara AM. Polycystic Ovary Syndrome (PCOS), Diagnostic Criteria, and AMH. Asian Pac J Cancer Prev. 2017;18(1):17–21.

2. Sirmans SM, Pate KA. Epidemiology, diagnosis, and management of polycystic ovary syndrome. Clin Epidemiol. 2013 Dec 18;6:1–13.

3. Sam S. Obesity and Polycystic Ovary Syndrome. Obes Manag. 2007 Apr;3(2):69–73.

4. Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. Fertil Steril. 2004 Jan;81(1):19–25.

5. Wang R, Mol BWJ. The Rotterdam criteria for polycystic ovary syndrome: evidence-based criteria? Hum Reprod. 2017;32(2):261–4.

6. Agapova SE, Cameo T, Sopher AB, Oberfield SE. Diagnosis and challenges of polycystic ovary syndrome in adolescence. Semin Reprod Med. 2014 May;32(3):194–201.

7. Rosenfield RL, Ehrmann DA. The Pathogenesis of Polycystic Ovary Syndrome (PCOS): The Hypothesis of PCOS as Functional Ovarian Hyperandrogenism Revisited. Endocr Rev. 2016 Oct;37(5):467–520.

8. Mohamed-Hussein Z-A, Harun S. Construction of a polycystic ovarian syndrome (PCOS) pathway based on the interactions of PCOS-related proteins retrieved from bibliomic data. Theor Biol Med Model. 2009 Sep 1;6:18.

9. Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics. 2012 Jun;11(6):O111.016717.

10. Wolters DA, Washburn MP, Yates JR. An automated multidimensional protein identification technology for shotgun proteomics. Anal Chem. 2001 Dec 1;73(23):5683–90.

11. Nesvizhskii AI, Aebersold R. Interpretation of Shotgun Proteomic Data: The Protein Inference Problem. Molecular & Cellular Proteomics. 2005 Oct;4(10):1419–40.

12. Pino LK, Searle BC, Bollinger JG, Nunn B, MacLean B, MacCoss MJ. The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. Mass Spectrom Rev. 2017 Jul 9;

13. Röst HL, Rosenberger G, Navarro P, Gillet L, Miladinović SM, Schubert OT, et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat Biotechnol. 2014 Mar;32(3):219–23.

14. Masseroli M, Kaitoua A, Pinoli P, Ceri S. Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying. Methods. 2016 01;111:3–11.

15. Cresswell J, Fraser R, Bruce C, Egger P, Phillips D, Barker DJP. Relationship between polycystic ovaries, body mass index and insulin resistance. Acta Obstet Gynecol Scand. 2003 Jan;82(1):61–4.

16. Vowinckel J, Capuano F, Campbell K, Deery MJ, Lilley KS, Ralser M. The beauty of being (label)-free: sample preparation methods for SWATH-MS and next-generation targeted proteomics. F1000Res [Internet]. 2014 Apr 7 [cited 2019 Sep 5];2. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3983906/

17. Rice SJ, Liu X, Zhang J, Belani CP. Absolute Quantification of All Identified Plasma Proteins from SWATH Data for Biomarker Discovery. Proteomics. 2019 Feb;19(3):e1800135.

18. Andrews GL, Simons BL, Young JB, Hawkridge AM, Muddiman DC. Performance Characteristics of a New Hybrid Triple Quadrupole Time-of-Flight Tandem Mass Spectrometer. Anal Chem. 2011 Jul 1;83(13):5442–6.

19. Kelstrup CD, Bekker-Jensen DB, Arrey TN, Hogrebe A, Harder A, Olsen JV. Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. J Proteome Res. 2018 05;17(1):727–38.

20. Ludwig C, Gillet L, Rosenberger G, Amon S, Collins BC, Aebersold R. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Mol Syst Biol [Internet]. 2018 Aug 13 [cited 2019 Sep 2];14(8). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6088389/

21. Rardin MJ, Schilling B, Cheng L-Y, MacLean BX, Sorensen DJ, Sahu AK, et al. MS1 Peptide Ion Intensity Chromatograms in MS2 (SWATH) Data Independent Acquisitions. Improving Post Acquisition Analysis of Proteomic Experiments. Mol Cell Proteomics. 2015 Sep;14(9):2405–19.

22. Escher C, Reiter L, MacLean B, Ossola R, Herzog F, Chilton J, et al. Using iRT, a normalized retention time for more targeted measurement of peptides. Proteomics. 2012 Apr;12(8):1111–21.

23. Meyer JG. Fast Proteome Identification and Quantification from Data-Dependent Acquisition–Tandem Mass Spectrometry (DDA MS/MS) Using Free Software Tools. Methods Protoc [Internet]. 2019 Jan 17 [cited 2019 Sep 3];2(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6469856/

24. Ebhardt HA. Selected reaction monitoring mass spectrometry: a methodology overview. Methods Mol Biol. 2014;1072:209–22.

25. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics. 2010 Apr 1;26(7):966–8.

26. Keller A, Eng J, Zhang N, Li X, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol. 2005;1:2005.0017.

27. Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. Proteomics. 2013 Jan;13(1):22–4.

28. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004 Jun 12;20(9):1466–7.

29. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics. 2008 Nov 1;24(21):2534–6.

30. Pfeuffer J, Sachsenberg T, Alka O, Walzer M, Fillbrunn A, Nilse L, et al. OpenMS - A platform for reproducible analysis of mass spectrometry data. J Biotechnol. 2017 Nov 10;261:142–8.

31. Möller S, Krabbenhöft HN, Tille A, Paleino D, Williams A, Wolstencroft K, et al. Community-driven computational biology with Debian Linux. BMC Bioinformatics. 2010 Dec 21;11 Suppl 12:S5.

32. Di Tommaso P, Palumbo E, Chatzou M, Prieto P, Heuer ML, Notredame C. The impact of Docker containers on the performance of genomic pipelines. PeerJ. 2015;3:e1273.

33. Building high-quality assay libraries for targeted analysis of SWATH MS data. - PubMed - NCBI [Internet]. [cited 2019 Sep 3]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25675208

34. Pedrioli PGA, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, et al. A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol. 2004 Nov;22(11):1459–66.

35. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–15.

36. Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, et al. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol Cell Proteomics. 2009 Nov;8(11):2405–17.

37. Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics. 2007 Mar;7(5):655–67.

38. Deutsch EW, Chambers M, Neumann S, Levander F, Binz P-A, Shofstahl J, et al. TraML—A Standard Format for Exchange of Selected Reaction Monitoring Transition Lists. Molecular & Cellular Proteomics. 2012 Apr;11(4):R111.015040.

39. Röst HL, Liu Y, D'Agostino G, Zanella M, Navarro P, Rosenberger G, et al. TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. Nat Methods. 2016;13(9):777–83.

40. pubmeddev, VJ RM and C. Bioconductor: an open source framework for bioinformatics and computational biology. - PubMed - NCBI [Internet]. [cited 2019 Sep 5]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/16939789

41. Blattmann P, Heusel M, Aebersold R. SWATH2stats: An R/Bioconductor Package to Process and Convert Quantitative SWATH-MS Proteomics Data for Downstream Analysis Tools. PLoS ONE. 2016;11(4):e0153160.

42. Choi M, Chang C-Y, Clough T, Broudy D, Killeen T, MacLean B, et al. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics. 2014 Sep 1;30(17):2524–6.

43. Skaik Y. The bread and butter of statistical analysis “t-test”: Uses and misuses. Pakistan Journal of Medical Sciences. 2015 Dec;31(6):1558.

44. Jafari M, Ansari-Pour N. Why, When and How to Adjust Your P Values? Cell Journal (Yakhteh). 2019 Winter;20(4):604.

45. Yoav Benjamini and Yosef 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):289–300.

46. Pascovici D, Handler DCL, Wu JX, Haynes PA. Multiple testing corrections in quantitative proteomics: A useful but blunt tool. Proteomics. 2016;16(18):2448–53.

47. Murillo JR, Goto-Silva L, Sánchez A, Nogueira FCS, Domont GB, Junqueira M. Quantitative proteomic analysis identifies proteins and pathways related to neuronal development in differentiated SH-SY5Y neuroblastoma cells. EuPA Open Proteom. 2017 Sep;16:1–11.

48. Wingo AP, Dammer EB, Breen MS, Logsdon BA, Duong DM, Troncosco JC, et al. Large-scale proteomic analysis of human brain identifies proteins associated with cognitive trajectory in advanced age. Nat Commun. 2019 08;10(1):1619.

49. Geyer PE, Wewer Albrechtsen NJ, Tyanova S, Grassl N, Iepsen EW, Lundgren J, et al. Proteomics reveals the effects of sustained weight loss on the human plasma proteome. Mol Syst Biol. 2016 Dec 22;12(12):901.

50. Oller Moreno S, Cominetti O, Núñez Galindo A, Irincheeva I, Corthésy J, Astrup A, et al. The differential plasma proteome of obese and overweight individuals undergoing a nutritional weight loss and maintenance intervention. Proteomics Clin Appl. 2018;12(1).

51. Berg P, McConnell EW, Hicks LM, Popescu SC, Popescu GV. Evaluation of linear models and missing value imputation for the analysis of peptide-centric proteomics. BMC Bioinformatics. 2019 Mar 14;20(Suppl 2):102.

52. Raivo Kolde. pheatmap: Pretty Heatmaps [Internet]. 2018. Available from: https://cran.r-project.org/web/packages/pheatmap/pheatmap.pdf

53. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25–9.

54. Winterhalter C, Widera P, Krasnogor N. JEPETTO: a Cytoscape plugin for gene set enrichment and topological analysis based on interaction networks. Bioinformatics. 2014 Apr 1;30(7):1029–30.

55. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498–504.

56. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W305-311.

57. Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt. 2014 Sep;34(5):502–8.

58. Chen S-Y, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. J Thorac Dis. 2017 Jun;9(6):1725–9.

59. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019 Jan 8;47(Database issue):D607–13.

60. Oikonomopoulou K, Ricklin D, Ward PA, Lambris JD. Interactions between coagulation and complement—their role in inflammation. Semin Immunopathol. 2012 Jan;34(1):151–65.

61. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009 Apr 15;25(8):1091–3.

62. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017 Jan 4;45(Database issue):D353–61.

63. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics. 2013 Jul 15;29(14):1830–1.

64. Panidis D, Kourtis A, Farmakiotis D, Mouslech T, Rousso D, Koliakos G. Serum adiponectin levels in women with polycystic ovary syndrome. Hum Reprod. 2003 Sep;18(9):1790–6.

65. Zhu Q, Zhou H, Zhang A, Gao R, Yang S, Zhao C, et al. Serum LBP Is Associated with Insulin Resistance in Women with PCOS. PLoS ONE. 2016;11(1):e0145337.

66. Tan BK, Chen J, Lehnert H, Kennedy R, Randeva HS. Raised serum, adipocyte, and adipose tissue retinol-binding protein 4 in overweight women with polycystic ovary syndrome: effects of gonadal and adrenal steroids. J Clin Endocrinol Metab. 2007 Jul;92(7):2764–72.

67. Naderpoor N, Shorakae S, Joham A, Boyle J, De Courten B, Teede HJ. Obesity and polycystic ovary syndrome. Minerva Endocrinol. 2015 Mar;40(1):37–51.

68. Dehdashtihaghighat S, Mehdizadehkashi A, Arbabi A, Pishgahroudsari M, Chaichian S. Assessment of C-reactive Protein and C3 as Inflammatory Markers of Insulin Resistance in Women with Polycystic Ovary Syndrome: A Case-Control Study. J Reprod Infertil. 2013;14(4):197–201.

69. Moreno-Navarrete JM, Fernández-Real JM. The complement system is dysfunctional in metabolic disease: Evidences in plasma and adipose tissue from obese and insulin resistant subjects. Semin Cell Dev Biol. 2019;85:164–72.

70. Kim JJ, Choi YM. Dyslipidemia in women with polycystic ovary syndrome. Obstet Gynecol Sci. 2013 May;56(3):137–42.

71. Magkos F, Mohammed BS, Mittendorfer B. Effect of obesity on the plasma lipoprotein subclass profile in normoglycemic and normolipidemic men and women. Int J Obes (Lond). 2008 Nov;32(11):1655–64.

72. Ellulu MS, Patimah I, Khaza'ai H, Rahmat A, Abed Y. Obesity and inflammation: the linking mechanism and the complications. Arch Med Sci. 2017 Jun;13(4):851–63.

73. Duleba AJ, Dokras A. Is PCOS an inflammatory process? Fertil Steril. 2012 Jan;97(1):7–12.

74. Kornblith L, Howard B, Kunitake R, Redick B, Nelson M, Cohen MJ, et al. Obesity and Clotting: BMI Independently Contributes to Hypercoagulability after Injury. J Trauma Acute Care Surg. 2015 Jan;78(1):30–8.

75. Mannerås-Holm L, Baghaei F, Holm G, Janson PO, Ohlsson C, Lönn M, et al. Coagulation and fibrinolytic disturbances in women with polycystic ovary syndrome. J Clin Endocrinol Metab. 2011 Apr;96(4):1068–76.

76. Aronson D, Bartha P, Zinder O, Kerner A, Markiewicz W, Avizohar O, et al. Obesity is the major determinant of elevated C-reactive protein in subjects with the metabolic syndrome. Int J Obes Relat Metab Disord. 2004 May;28(5):674–9.

77. Oh JY, Lee J-A, Lee H, Oh J-Y, Sung Y-A, Chung H. Serum C-Reactive Protein Levels in Normal-Weight Polycystic Ovary Syndrome. Korean J Intern Med. 2009 Dec;24(4):350–5.

78. Alvarez-Blasco F, Botella-Carretero JI, San Millán JL, Escobar-Morreale HF. Prevalence and characteristics of the polycystic ovary syndrome in overweight and obese women. Arch Intern Med. 2006 Oct 23;166(19):2081–6.

79. Sharifi F, Hajihosseini R, Mazloomi S, Amirmogaddami H, Nazem H. Decreased adiponectin levels in polycystic ovary syndrome, independent of body mass index. Metab Syndr Relat Disord. 2010 Feb;8(1):47–52.

80. Jia J, Bai J, Liu Y, Yin J, Yang P, Yu S, et al. Association between retinol-binding protein 4 and polycystic ovary syndrome: a meta-analysis. Endocr J. 2014;61(10):995–1002.

81. Chan T-F, Tsai Y-C, Chiu P-R, Chen Y-L, Lee C-H, Tsai E-M. Serum retinol-binding protein 4 levels in nonobese women with polycystic ovary syndrome. Fertil Steril. 2010 Feb;93(3):869–73.

82. Henriksen ML, Brandt J, Iyer SSC, Thielens NM, Hansen S. Characterization of the interaction between collectin 11 (CL-11, CL-K1) and nucleic acids. Mol Immunol. 2013 Dec;56(4):757–67.

83. Venkatraman Girija U, Furze CM, Gingras AR, Yoshizaki T, Ohtani K, Marshall JE, et al. Molecular basis of sugar recognition by collectin-K1 and the effects of mutations associated with 3MC syndrome. BMC Biol. 2015 Apr 17;13:27.

84. Rooryck C, Diaz-Font A, Osborn DPS, Chabchoub E, Hernandez-Hernandez V, Shamseldin H, et al. Mutations in lectin complement pathway genes COLEC11 and MASP1 cause 3MC syndrome. Nat Genet. 2011 Mar;43(3):197–203.

85. Ingermann AR, Yang Y-F, Han J, Mikami A, Garza AE, Mohanraj L, et al. Identification of a novel cell death receptor mediating IGFBP-3-induced anti-tumor effects in breast and prostate cancer. J Biol Chem. 2010 Sep 24;285(39):30233–46.

86. te Pas MF, Koopmans S-J, Kruijt L, Boeren S, Smits MA. Changes in Plasma Protein Expression Indicative of Early Diet-induced Metabolic Disease in Male Pigs (Sus scrofa). Comp Med. 2018 Aug;68(4):286–93.

87. Domené HM, Scaglia PA, Martínez AS, Keselman AC, Karabatas LM, Pipman VR, et al. Heterozygous IGFALS gene variants in idiopathic short stature and normal children: impact on height and the IGF system. Horm Res Paediatr. 2013;80(6):413–23.

Chapter 5. Data Dependent Acquisition and Label Free Quantification: iprg2015 reanalysis

5.1 Abstract

The aim of this work is to find the most suitable tools and parameters for the analysis of data-dependent, label free quantitative proteomics data. It is with that goal in mind that the reanalysis of the data set provided by the iprg2015 study has been performed. Three quantitative pipelines (Proteome Discoverer, MaxQuant and OpenMS) and three statistical R packages (MSStats, DEqMS and DEP) have been evaluated. The use of MaxQuant and MSstats, using a P-value of 0.05 and a Log2 fold change of ±1 as thresholds has demonstrated to be the most robust approach when dealing with complex protein mixtures. Label free quantification accuracy has been evaluated, with overall correct results but with higher errors when measuring the more extreme ratios: accuracy is lower for measures involving very high or very low ratios. Also, imputation of censored values has been explored: the “accelerated failure time” model for imputation has been chosen as the most robust approach, albeit the use of various imputation strategies may prevent the emergence of artifacts.

5.2 Introduction

In the year 2015, the Proteome Informatics Research Group (iPRG) of the Association of Biomolecular Resource Facilities (ABRF) released the proteomics analysis of four samples of a tryptic digest of *Saccharomyces cerevisiae*, spiked with six proteins at known concentrations. Technical triplicates of the four samples were analyzed following a data-dependent label-free quantification approach, using a a Thermo Scientific Q-Exactive mass spectrometer. The spiked proteins presented different concentrations for each sample, only known by the organizers of the project. Then, sixty anonymous volunteers from around the world, used their bioinformatics pipelines to analyze the samples, and the results were summarized and commented in the publication “ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of differentially abundant proteins in label-free quantitative LC-MS/MS experiments” (1). The raw data and the protein database was made available at the ProteomeXchange (2) platform under the identifier PXD010981.

The characteristics of the iprg2015 data set are summarized here:

- Four different samples, all containing the same amount of a tryptic protein digest of *Saccharomyces cerevisiae* culture (200 ng), were spiked with different concentrations of six proteins (Table 5.1).
- The identity of the proteins was known at the moment of this reanalysis: chicken ovoalbumin (P01012, OVAL_CHICK), horse myoglobin (P68082, MYG_HORSE), rabbit Glycogen phosphorylase, muscle form (P00489, PYGM_RABIT), Beta-galactosidase from Escherichia coli, strain K12 (P00722, BGAL_ECOLI), bovine serum albumin (P02769, ALBU_BOVIN) and bovine Carbonic anhydrase 2 (P00921, CAH2_BOVIN). For convenience, these six proteins have been respectively labeled with letters A to F, in the same way that was done in the original study.

- The four samples are named here as C1 to C4. The comparisons made throughout this chapter will be only of C2, C3 and C4 with respect to C1 (C2 vs C1, C3 vs C1 and C4 vs C1). More combinations could have been used, but for the purpose of this work and to limit the complexity and extent of the results presented, only those three comparisons will be made. Due to the fact that the amount of *S. cerevisiae* is constant, the proteins coming from this background protein digest should ideally have a Log2 fold change equal to zero. Any protein from the yeast digest that shows differential expression in some of the comparisons must be accounted as an artifact or associated to some issue in the bioinformatics analysis pipeline.
- On the other hand, the six spiked proteins (A to F) must show, in almost all cases, some significant difference in the three comparisons studied. Because the relative concentrations (ratios or fold changes) can be obtained as the quotient of the spikes concentrations in the samples compared, theoretical Log2 fold changes have been calculated and shown in Table 5.1.

	Samples				Theoretical Log2 fold changes		
	C1	**C2**	**C3**	**C4**	**C2 vs C1**	**C3 vs C1**	**C4 vs C1**
A	65	55	15	2	-0.24	-2.12	-5.02
B	55	15	2	65	-1.87	-4.78	0.24
C	15	2	65	55	-2.91	2.12	1.87
D	2	65	55	15	5.02	4.78	2.91
E	11	0.6	10	500	-4.20	-0.14	5.51
F	10	500	11	0.6	5.64	0.14	-4.06

Table 5.1 Each sample contained 200 ng yeast tryptic digest spiked with the indicated amounts (in fmols) of tryptic digest of six individual proteins. The A to F labels correspond to P01012 (OVAL_CHICK), P68082 (MYG_HORSE), P00489 (PYGM_RABIT), P00722 (BGAL_ECOLI), P02769 (ALBU_BOVIN) and P00921 (CAH2_BOVIN). The theoretical ratios (Log2 fold changes) have been calculated using the known amounts of the spiked proteins.

Throughout this chapter, data are going to be analyzed using different analysis pipelines and the results will be expressed both in a numerical and a graphical way. For the graphical representation of those results, volcano plots (3) are going to be used. In Figure 5.1, three volcano plots, reproduced from the original publication (1) are shown: the dots labeled as A to F letters represent the spiked proteins and the unlabeled dots, the proteins coming from the yeast protein digest.

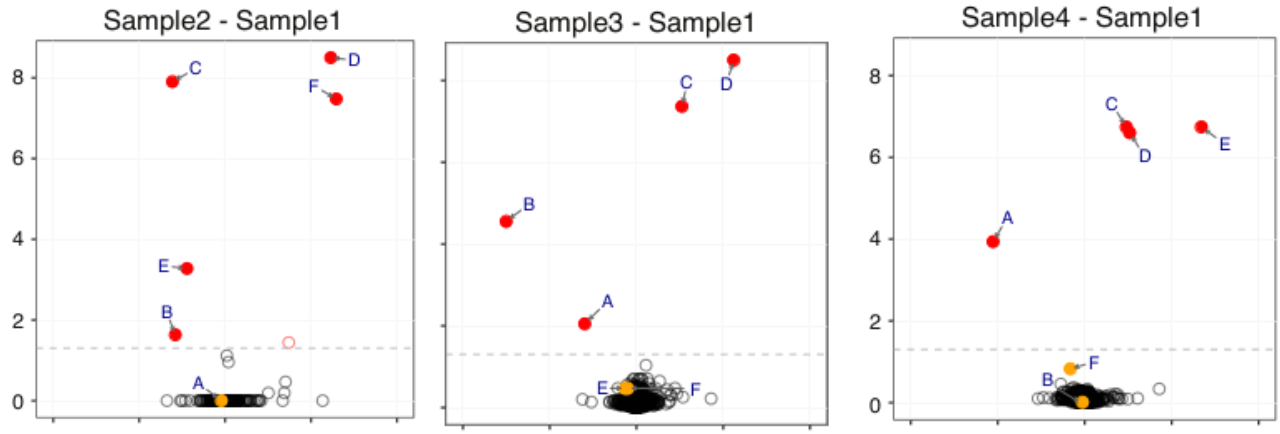


Figure 5.1 Results, in the form of volcano plots, for the C2 vs C1, C3 vs C1 and C4 vs C1 comparisons, taken from the original publication (1) by Choi M. et al (2017) , Figure 4. The x axis represents the Log2

fold change (ranging from -10 to +10) and the y axis the -Log10 adjusted P-Values. The cutoff (adjusted P-value of 0.05) used in the original publication is drawn using a dashed line.

5.3 Materials and Methods

The analysis of the data from the iprg2015 study will be performed using three different identification and quantification pipelines (MaxQuant (4,5), OpenMS (6,7) and Proteome Discoverer (8)) and three Bioconductor (9) packages for the statistical analysis of quantitative proteomics data (MSStats (10), DEqMS (11) and DEP (12)). The aim of this work is not to compare the performance of those pipelines and software packages, but to find a combination of software, parameters and procedures that allows a confident quantification of data-dependent label free proteomics experiments. It is possible that using other parameters than those applied here would have allowed to improve the results obtained with some of the software packages tested; but following the different documentation to the best of our knowledge has produced the results that will be shown here.

5.3.1 Identification and quantification software

The data sets generated by the iprg2015 project have been analyzed using the raw files as the starting point for the three pipelines: MaxQuant, Proteome Discoverer and OpenMS. The protein database used for identification in this work is the one provided by the project itself, where spiked proteins were disguised using deceptive S.cerevisiae protein accessions (P44015, P55752, P44374, P44983, P44683 and P55249 for A to F respectively). A decoy database (using reversed sequences) was employed by the three pipelines in order to calculate the False Discovery Rate (FDR) at peptide and protein levels. Also, Cysteine Carbamidomethylation as fixed modification and Oxidation of Methionine as variable modification have been used. Digestion using trypsin (KR^P) was employed. The description of the three pipelines is shown in the next sections.

5.3.1.1 MaxQuant

MaxQuant is described as “a quantitative proteomics software package designed for analyzing large mass-spectrometric data sets”. It supports the analysis of several labeled techinques (iTRAQ and TMT among them) and label-free quantification. A companion application, named Perseus (13), is sometimes used in combination with MaxQuant for statistical analysis of the quantitative data in some pipelines. Perseus has not been employed in this work because it is designed for its use in combination with only MaxQuant results.

MaxQuant provides a graphical interface for the elaboration of a parameters text file, where all the information needed in the analysis is supplied (Figure 5.2). The MaxQuant graphical interface is also capable of launching the application.

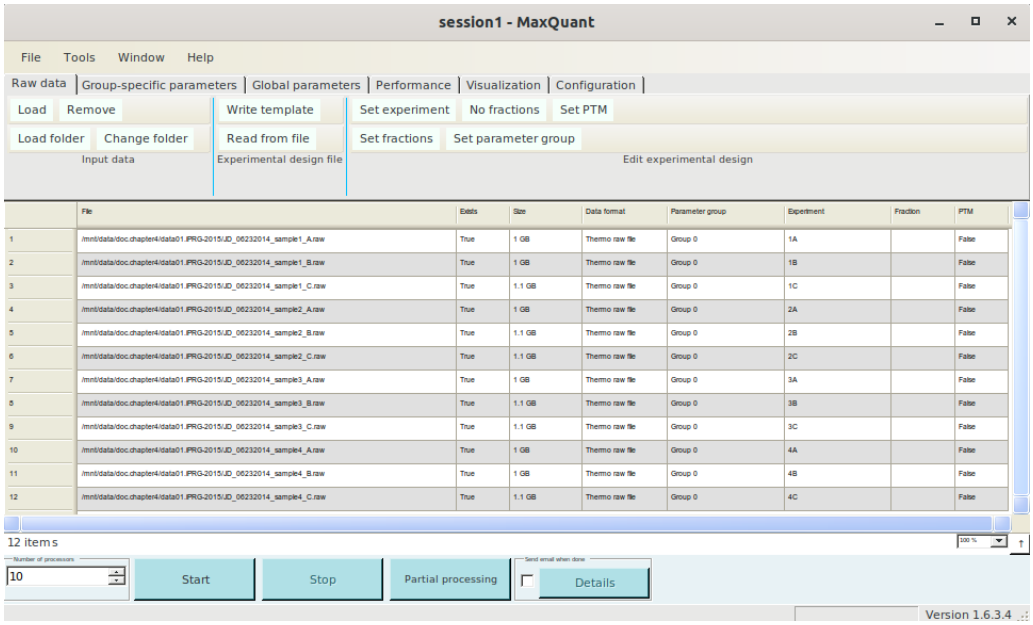


Figure 5.2 Graphical interface to MaxQuant. Raw data menu is shown, with the list of raw files to be analyzed in this work.

Once the parameters have been introduced by using the graphical interface, the MaxQuant pipeline can be launched using the command line by an application called MaxQuantCmd.exe. This application can be used in Linux systems, through the Mono (14) framework, and easily combined with the Slurm (15) workload manager; it is precisely using this setup in which MaxQuant has been used in this work for identification and quantification of proteins. Using MaxQuant or any other pipeline, take several hours, and automating and scheduling this process is very convenient. Parameters used by MaxQuant include, for the identification, a peptide-spectrum matching (PSM) and protein FDR both of 0.01. The instrument setup selected has been an Orbitrap, with a “main search tolerance” at MS^1 level of 4.5 ppm. The rest of the parameters of importance (like nature and number of the peptides used in the quantification) will be especified later with the statistical software used.

After the quantification is performed, several files with the information produced by MaxQuant are generated into a folder named “combined/txt”; two of these files, proteinGroups.txt (with a summary listing the proteins and the quantification information) and evidence.txt (with a detailed list of the peptides quantified), will be used by the statistical packages used later.

5.3.1.2 OpenMS

The OpenMS project comprises several software tools for the analysis (both for protein identification and quantification) of mass spectrometry data. It can be used by means of the command line or using the Knime (16) analytics platform. This platform helps building scientific workflows and executing them into the workspace. Those workflows consist of nodes that perform both general procedures (like generating a text file) or very specific (like performing proteomics identification) and are arranged answering for the specific needs of the analysis. The Knime platform can be run using Linux, Windows and Mac platforms

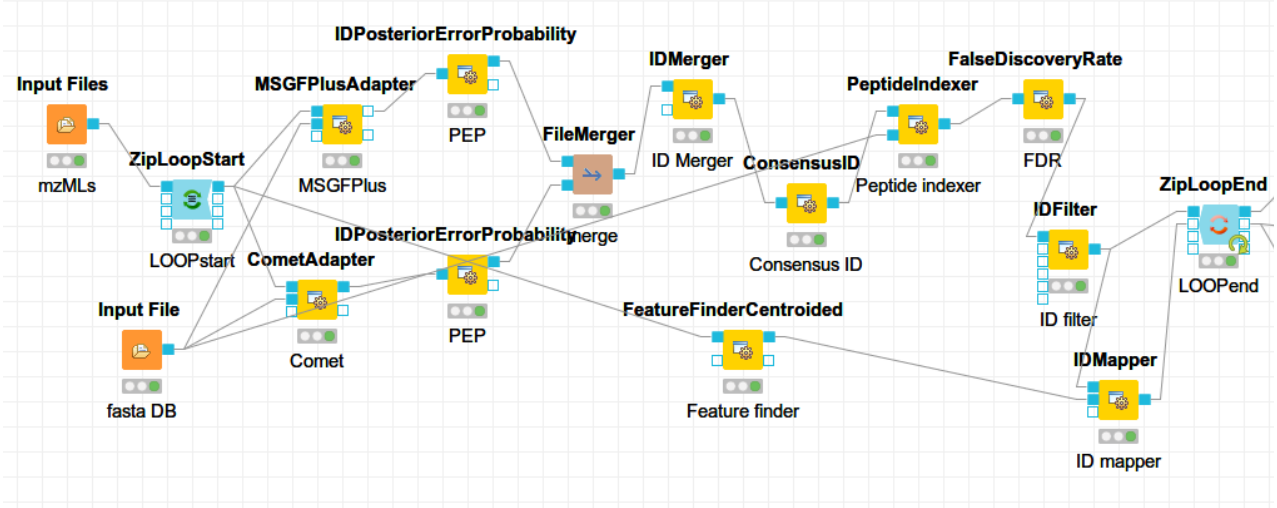


Figure 5.3 First part of the OpenMS workflow for label-free quantification. Raw files are used as input and a loop cycles over them (in blue, ZipLoopStart and ZipLoopEnd), executing two protein identification steps (Comet and MSGF-Plus) and merging the probabilities produced by them for each peptide. Alongside to the peptide identification, the workflow calculates the areas of the centroided peaks (FeatureFinder) and maps the identification and quantitation information.

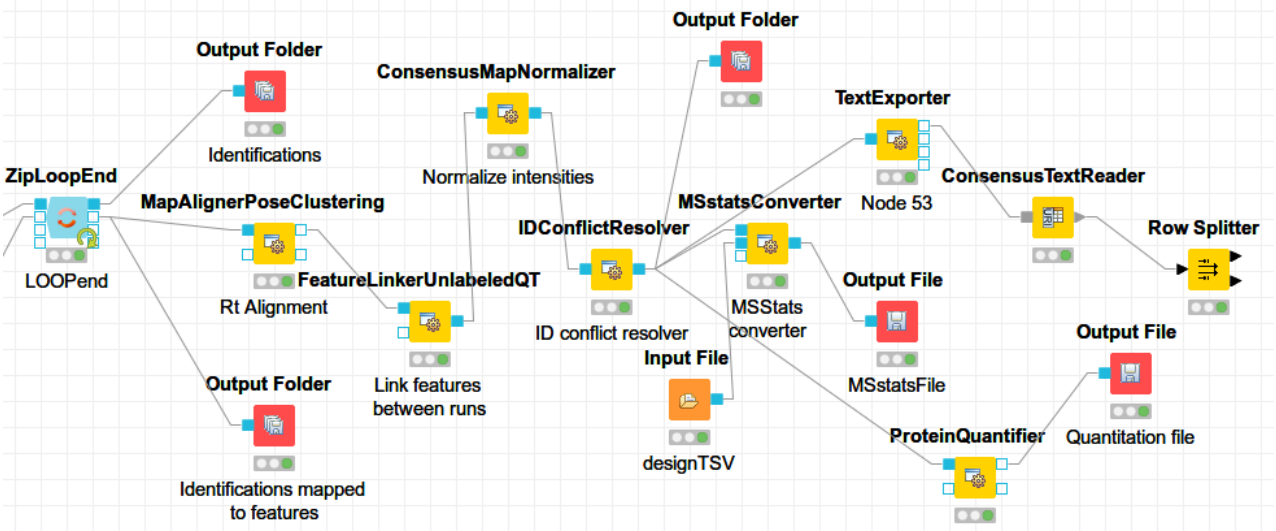


Figure 5.4 Second part of the OpenMS workflow for label-free quantification. Features (quantified peptides) are aligned using retention times (Rt Alignment). Then, the information from different files is merged (FeatureLinker), normalized (ConsensusMapNormalizer) and several text files are exported with the identifications and quantitative information. Ultimately, two text files are produced: a generic Quantitation File produced by OpenMS and a MSStats-compatible file to be used by the statistical packages used later.

The OpenMS analysis platform was directly downloaded using the Knime application. Then, following the documentation available, a workflow to identify and quantify the data provided by the iprg2015 study was built. An overview of the OpenMS quantitation pipeline that was created for this work is provided in two parts:

- In Figure 5.3, the initial part of the workflow is shown, with raw files processed in a loop where peptides are identified and quantified.
- In Figure 5.4, the final part of the workflow, where identification and quantification information is integrated and a text file, with the quantified data, is produced.

As the main filter used, and FDR value of 0.05 for both protein and peptide identification have been used. After the pipeline has finished, two text files with the results are generated: a generic Quantitation File produced by OpenMS and a MSStats-compatible file to be used by the statistical packages used later.

5.3.1.3 Proteome Discoverer

Proteome Discoverer is a proprietary software produced by the Thermo Fisher Scientific company. The main limitation of Protein Discoverer is that it is designed to work only with instruments produced by the same company. Another limitation is that only works on Windows operative systems. On the other hand, it has a very complete graphical interface and some capabilities of batch processing for the files that are analyzed.

The software works by designing pipelines with nodes executing the different steps. The analysis workflow is divided in two elements by the application:

- A “Child step” (or “Processing step”, Figure 5.5) that performs the identification of proteins (using Sequest HT (17) and Percolator (18)) and a feature detection (at MS1 level) for protein quantification.
- A “Consensus step” (or “Integration step”, Figure 5.6) that integrates the results obtained from different files, performing the quantification using the features detected in the “Processing step”.

The parameters used by the Proteome Discoverer pipeline are the ones provided by a predefined configuration for the same instrument used in this analysis: Proteome Discoverer allows selecting a default pipeline for label-free quantification using the Q-Exactive mass spectrometer.

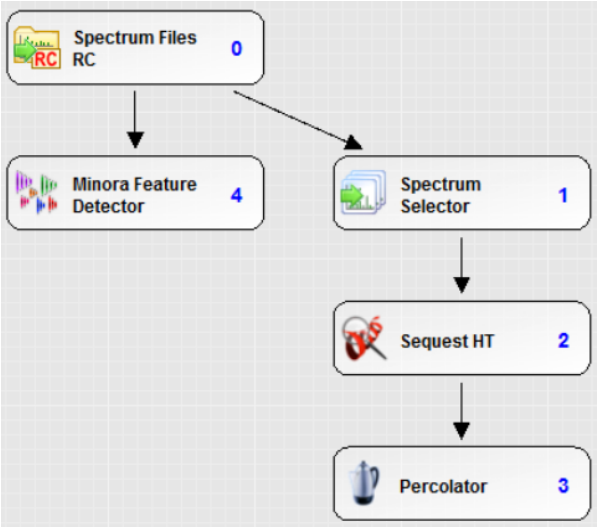


Figure 5.5 Child Step (Processing step) in Proteome Discoverer.

Proteome Discoverer shows the results in the graphical interface, using an advanced interface that allows inspecting spectra, proteins, quantification information... It also allows exporting the results produced in several formats: Excel files, some proteomics XML standards and also tabulated text files. The file format that will allow the integration of Proteome Discoverer with the statistical packages used later, is the file containing the identification and quantification information organized as PSM (Peptide Spectrum Matches). This is a text file that will be imported by the R packages used later.

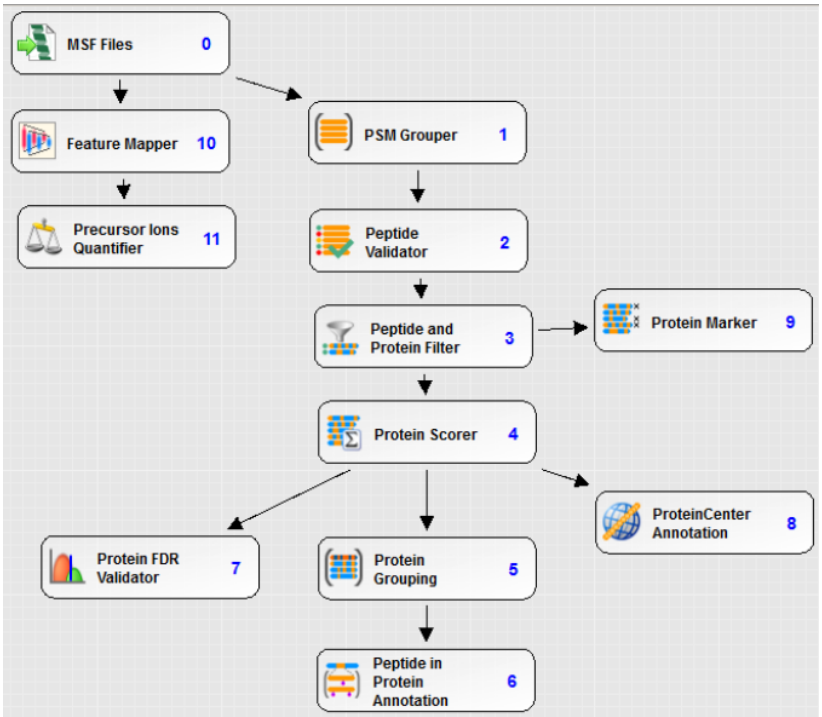


Figure 5.6 Consensus Step (Integration step) in Proteome Discoverer.

5.3.2 Statistical protein quantification software

Three different R packages included in the Bioconductor project are used in the statistical analysis, integration of results and filtering of the quantitative data produced here: MSStats, DEqMS and DEP. These packages are described in the next points.

5.3.2.1 MSStats

MSStats is a mature Bioconductor package (version 3.14.1 used here), that is used both in data dependent and data independent (Swath) acquisition setups. With direct bindings to the most used pipelines in quantitative proteomics, it is also very well documented and widely used. The MSStats package includes:

- several functions to import data from the different pipelines supported,
- a processing step, where the quantitative information is integrated, normalized and filtered using some advanced features and,
- a final step where the different conditions under study (phenotypes, times in time-series experiments,...) are defined and compared.

MSStats allows different methods of normalization (median equalization, quantile normalization or based on standard proteins). It also performs quantification by using several combinations of features (using all peptides, the three most intense,...). Different censoring methods are also supported. One interesting functionality (only available with MaxQuant data) is the possibility of discarding peptides containing methionine, as will be discussed in the section corresponding to the MaxQuant and MSStats combination.

In addition to the quantitative information related directly to proteins, MSStats also requires the use of the features identified at the PSM level.

5.3.2.2 DEqMS

The DEqMS package is a more recent alternative for statistical analysis of label-free data (1.0.1 version used here). It supports the two main kinds of data-dependent acquisition: label-free quantization ans isobaric labeling (TMT and iTRAQ). It is build on top of the limma (19) package, a widely used solution used in transcriptomics for more than fifteen years. It is very flexible while importing data produced by different software pipelines, allowing the direct import of MaxQuant data and also is capable of working with either PSM or protein tables. Different ways of importing and analyzing the quantitative data are described in its documentation.

Options for normalization are less advanced in this package compared to the other two: only medianSummary, medianSweeping and medpolishSummary functions are available.

The workflow for DEqMS is by far the most complex of the three statistical packages used here: although a detailed description of several common quantitative pipelines are described in the documentation, processing the data with this software involves many steps that can be tuned in multiple combinations.

5.3.2.3 DEP

DEP, the last of the three statistical packages used in this work, is a completely new option: at the moment of writing this document, a dedicated publication has not been released yet. It is designed to work with only two quantitative pipelines (MaxQuant for label free and IsobarQuant for isobaric labelling), making special emphasis on visualization. It provides many integrated plots for quality control and results evaluation. It only supports one method of normalization (variance stabilizing transformation), but on the other hand, provides many alternatives for data imputation and graphical tools to explore the type of missing values predominant in the data analyzed.

The use of this package is very straightforward, providing a set of functions to be used sequentially. Although DEP performs worse than DEqMS and MSStats (as will be shown later), it presents a very interesting set of tools and, with further development, can become a very strong tool in quantitative proteomics analyses.

5.3.3 Visualization software: Enhanced Volcano

The results obtained with the combination of identification and quantification pipelines (MaxQuant, Proteome Discoverer and OpenMS) with statistical analysis packages (MSStats, DEqMS and DEP) are summarized in this work using two approaches:

- First, lists of Log2 fold changes and adjusted P-values will be displayed as tables for the three comparisons and the six protein spikes used.
- Secondly, volcano plots for every comparison will also be provided.

The use of volcano plots with expression data in general (transcriptomics and proteomics) is very common and convenient: in addition to supply a graphical representation of the results, it provides a very good glimpse of the distribution that the quantified data follows for every analysis. The cutoffs used in the comparisons made (corrected P-value<0.05 and a Log2 fold change of ±0.5 will be used in most cases) are represented using dotted lines, and different colors will be used for points surpassing the different thresholds.

The software used for drawing the volcano plots is EnhancedVolcano (20), that allows to easily set thresholds, labelling specific proteins and scaling axis in the most convenient way for visualization.

5.4 Results

In the reanalysis of the iprg2015 data, three quantitation pipelines have been used (MaxQuant, OpenMS and Proteome Discoverer) in combination with three statistical analysis packages (MSStats, DEqMS and DEP). In each case, different approaches have been taken:

- MSStats incorporates direct methods for importing the data produced by MaxQuant, Proteome Discoverer and OpenMS, something that has facilitated the combined use of MSStats with the three quantitation pipelines used in this work.
- In the case of DEqMS, the package has been designed to directly import MaxQuant data, and has been adapted here, using a Perl script created specifically for this purpose in this work, to be used also with OpenMS and Proteome Discoverer, providing acceptable results, specially in its use with Proteome Discoverer.
- In the case of DEP, it has only used in combination with MaxQuant: several attempts, using the same kind of script that was designed for DEqMS, produced sub-optimal results that have not been included in this chapter.

In all combinations, the same threshold has been applied: a corrected P-value<0.05 and a Log2 fold change of ±0.5. The spiked proteins, represented with letters A to F (Table 5.1) will be inspected in each case, and the same will be done with the eventual apparition of non-spiked proteins. Those non-spiked proteins, that will represent false positives in the quantification procedures, will be represented with their respective Uniprot Accession Number (21) (e.g. P54000 or Q08773). The fold changes obtained for the spiked proteins will be expected to be similar to the theoretical ones shown in Table 5.1, while non-spiked proteins are not expected to pass the threshold used in the subsequent comparisons.

The different workflows are discussed below and the code that generated the results is included in “**Appendix 3: Chapter5, iprg2015 Reanalysis**”.

5.4.1 MaxQuant and MSStats

The combination of MaxQuant and MSStats, in a first approach, has produced the results shown in Figure 2.17. In the import of the MaxQuant data by MSStats, only unique peptides (those mapping uniquely one protein) have been used, and proteins quantified only using one peptide have been removed. Data processing has used “equalized medians” as the normalization approach and Tukey's median polish as the summarization approach. A total of 2,519, 2,515 and 2,509 proteins have been quantified for C2 vs C1, C3 vs C1 and C4 vs C1 comparisons.

In C2 vs C1, four of the expected spikes (B, C, D and E) appear, three for C3 vs C1 (A, C and D) and four spikes (A,C,D and E) and two false positives (P54000 and P08525) show up in C4 vs C1.

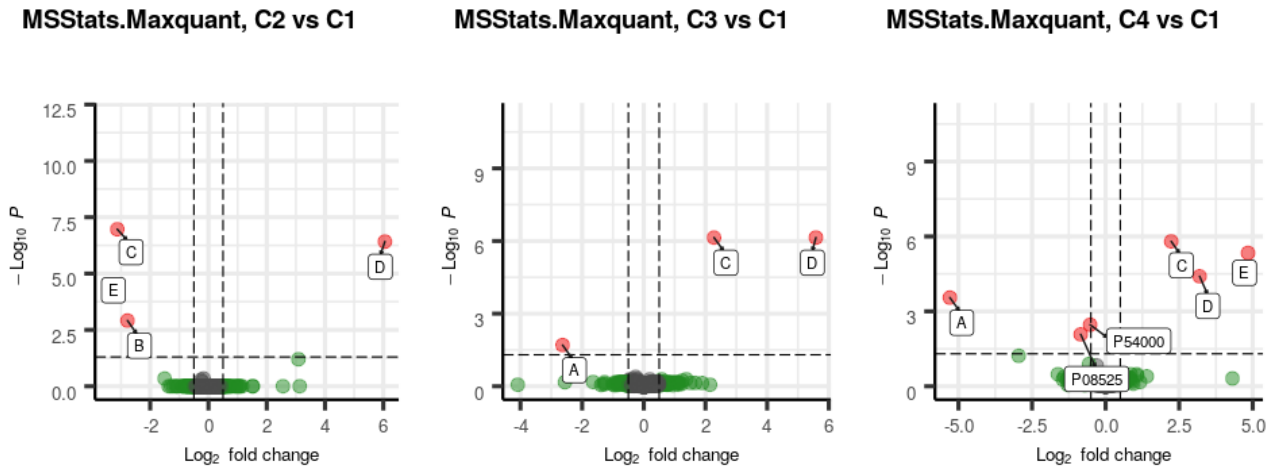


Figure 5.7 MaxQuant and MSStats results. Peptides containing methionine have been used. A total of 2519, 2515 and 2509 proteins have been respectively quantified for C2 vs C1, C3 vs C1 and C4 vs C1 comparisons.

Using the same parameters than before, but removing peptides containing methionine in the import of MaxQuant data, has produced the results shown in Figure 5.8. In this case, for C2 vs C1, five of the expected spikes (B, C, D, E and F) appear, three for C3 vs C1 (A, C and D) and four spikes (A,C,D,E and F) for C4 vs C1. None of the non-spiked proteins have appeared as differentially found in any of the comparisons.

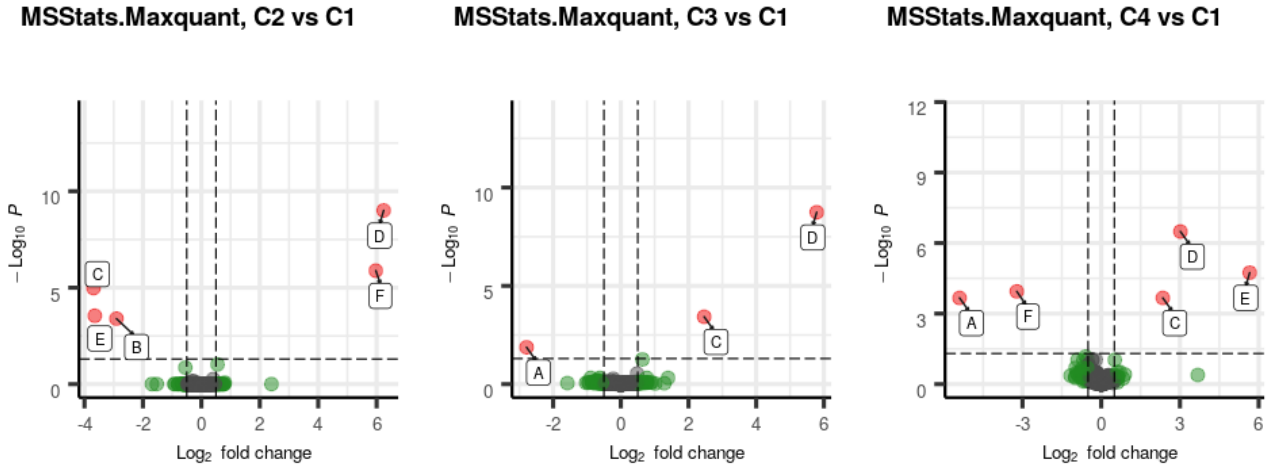


Figure 5.8 MaxQuant and MSStats results with peptides containing methionine removed: 2365, 2360 and 2335 proteins have bee quantified for C2 vs C1, C3 vs C1 and C4 vs C1 comparisons, respectively.

The removal of methionine is justified by the oxidation process that this amino acid presents with high frequency in the proteomics experiments (22). It is common that methionine sulfoxides appear as an artifact during the processing of samples in proteomics experiments. The accidental appearance of oxidized methionines, in different protein locations for the different samples produced, may introduce a serious bias in quantification if this is not corrected: since the same peptide can be found as oxidized for some percentage (for example, 80%) in some replicate and the corresponding peptide in another replicate can appear differently oxidized (for example 50%), only a 20% of the original peptide (not oxidized) will be considered in the first case and 50% in the second: this effect will produce an important underestimation of this peptide in the first replicate and an artifact will be introduced in the final quantification for the whole protein. This effect will

be more important if a higher number of methionine residues are present in the tryptic digest.

Comparing the results shown in figure 5.7 to the one in Figure 5.8, the removal of peptides containing methionine has improved the results (with two more spikes detected and the removal of non-spiked proteins), while losing a relatively low amount of the proteins quantified (close to the 6% in the three comparisons).

MaxQuant and MSStats quantitation

	C2 vs C1		C3 vs C1		C4 vs C1	
	log2 FC	P-value	log2 FC	P-value	log2 FC	P-value
A	-0.53	9.94e-01	-2.79	1.34e-02	-5.40	2.16e-04
B	-2.91	4.03e-04	0.00	0.00e+00	-0.25	5.09e-01
C	-3.69	1.05e-05	2.46	3.74e-04	2.35	2.16e-04
D	6.24	9.88e-10	5.79	1.80e-09	3.02	3.24e-07
E	-3.65	2.85e-04	-0.60	8.21e-01	5.66	1.82e-05
F	5.97	1.31e-06	-0.37	8.41e-01	-3.22	1.14e-04

Table 5.2 MaxQuant and MSStats results for the spiked proteins, using removal of peptides containing methionine. Spike B for C3 vs C1 comparison is not detectable.

Results shown in Table 5.2 show fold changes resembling to the ones expected by calculating the theoretical fold changes (Table 5.1). The only issue found with the MaxQuant and MSStats analysis is that the spiked protein B is missing completely for the C3 vs C1 comparison. This artifact will be discussed later in this chapter.

5.4.2 MaxQuant and DEqMS

The results obtained for MaxQuant and DEqMS are shown in Figure 5.9 and Table 5.3. The analysis has followed the process described in the DEqMS documentation, where MaxQuant output file “proteinGroups.txt” was used.

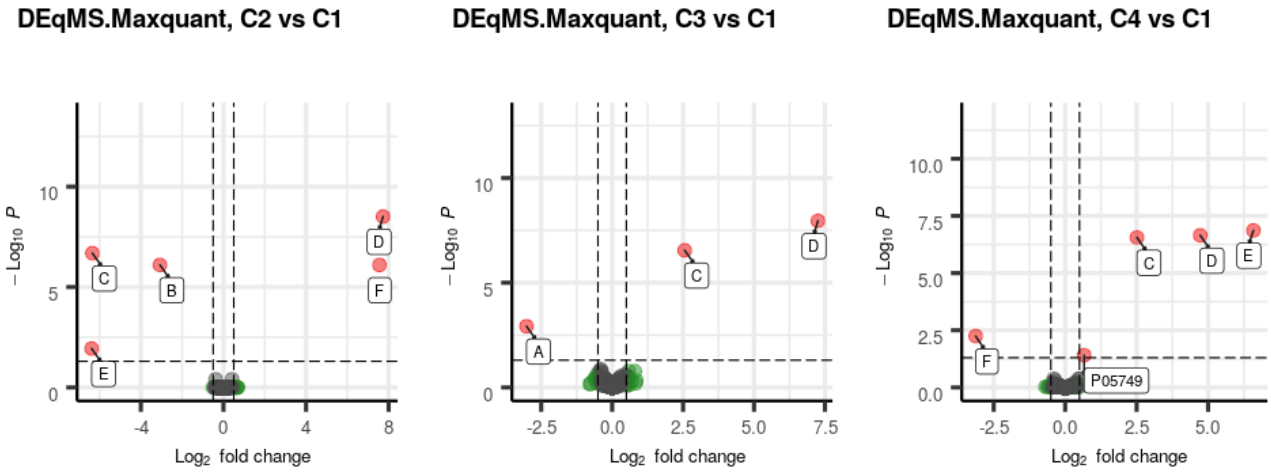


Figure 5.9 MaxQuant amd DEqMS results. A total of 1966, 1911 and 1911 proteins have been quantified for C2 vs C1, C3 vs C1 and C4 vs C1, respectively.

Roughly the same results that were shown with MaxQuant and MSStats are obtained here. The number of quantified proteins is lower here (an average of 1930 proteins here, while

2353 were quantified with MSStats), and one non-spiked protein, P05749, appears for C4 vs C1, while it is certain that with very low values of P-value and fold change.

MaxQuant and DEqMS quantitation

	C2 vs C1		C3 vs C1		C4 vs C1	
	log2 FC	P-value	log2 FC	P-value	log2 FC	P-value
A	-0.38	3.93e-01	-3.02	1.20e-03	0.00	0.00e+00
B	-3.08	7.98e-07	0.00	0.00e+00	-0.26	8.22e-01
C	-6.35	2.07e-07	2.56	2.84e-07	2.51	2.77e-07
D	7.73	3.09e-09	7.25	1.11e-08	4.73	2.23e-07
E	-6.38	1.17e-02	-0.52	6.40e-01	6.59	1.37e-07
F	7.55	7.98e-07	-0.19	8.78e-01	-3.13	5.71e-03

Table 5.3 MaxQuant and DEqMS results for the spiked proteins. Spiked proteins B and A are not quantified for C3 vs C1 and C4 vs C1, respectively.

Two spiked proteins have not passed the threshold here: B in C3 vs C1 and A in C4 vs C1 comparisons. The case of B in C3 vs C1 will be discussed later in the chapter, while the case of A in C4 vs C4 is different. Having this spike a very low concentration (presenting a theoretical Log2 fold change of -5.02) and with only two peptides detected in the three replicates of C4 (as will be seen later in Figure 5.15), this protein has been filtered by the DEqMS algorithm for this specific comparison, due to a low number of confidently identified peptides.

5.4.3 MaxQuant and DEP

The results for the MaxQuant and DEP combinations are shown in Figure 5.10 and Table 5.4. DEP data is filtered for proteins with missing values for at least one condition and then normalized using a variance stabilizing transformation function. Then, data is imputed following the “MinProb” approach.

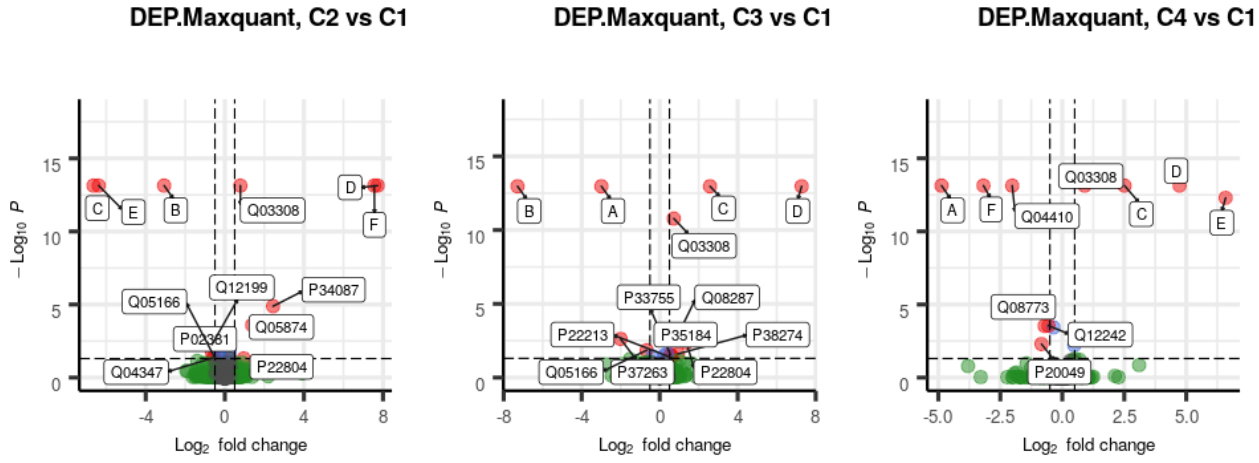


Figure 5.10 MaxQuant and DEP results. A total of 2019 proteins have been quantified for all comparisons.

From the distributions shown in Figure 5.10 it is clear that a considerable amount of false positive (non-spiked) proteins appear as significantly differential in the groups compared.

Further optimization of the parameters employed and more stringent filters could be tried to remove the non-spiked proteins appearing here. On the other hand, all spiked proteins appear correctly quantified: experimental and theoretical fold changes are close enough.

One factor that makes the DEP package an interesting choice, is that it approaches imputation (as will be discussed later in this chapter) in a very rigorous way, offering several algorithms to deal with missing values.

MaxQuant and DEP quantitation

	C2 vs C1		C3 vs C1		C4 vs C1	
	log2 FC	P-value	log2 FC	P-value	log2 FC	P-value
A	-0.38	4.98e-01	-3.00	2.07e-13	-5.86	2.13e-13
B	-3.07	1.20e-13	-8.13	2.07e-13	-0.26	9.52e-01
C	-6.93	1.20e-13	2.58	2.80e-08	2.51	4.07e-06
D	7.73	1.20e-13	7.27	2.07e-13	4.72	2.13e-13
E	-6.38	1.20e-13	-0.50	9.26e-01	6.58	5.12e-13
F	7.56	1.20e-13	-0.17	9.45e-01	-3.64	4.43e-13

Table 5.4 MaxQuant and DEP results for the spiked proteins. In all cases, an estimated P-value and fold change is reported for the spiked proteins.

5.4.4 OpenMS and MSStats

Data obtained with OpenMS have been processed using MSStats here, and results shown in Figure 5.11 and Table5.5. MSStats parameters correspond to the ones used with the MaxQuant data, with the difference that a MSStats compatible file has been created directly by the OpenMS workflow and no importing function has been needed here. Peptides containing methionine have not been removed in this analysis.

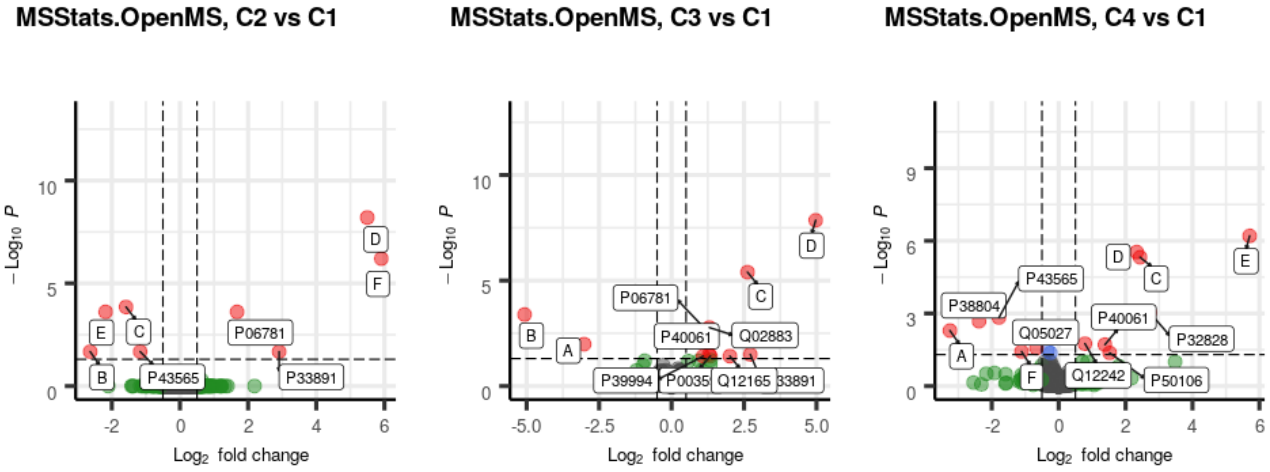


Figure 5.11 OpenMS and MSStats results. A total of 2557, 2558 and 2555 proteins have been respectively quantified for C2 vs C1, C3 vs C1 and C4 vs C1 comparisons.

From Figure 5.11, several non-spiked proteins appear for the three comparisons. It is also certain that all the spiked proteins are present as well (the ones that are expected to show up by their concentration values). The high amount of proteins quantified (averaging more that 2550 proteins per comparison) suggests that room is left for more restrictive filters to be applied on the quantified proteins. For example, removal of methionine containing

peptides could produce an improvement of the results obtained here; the problem here is that MSStats does not allow this option for OpenMS data.

The accuracy shown in the spiked proteins quantification (Table5.5) and the fact that the OpenMS pipeline includes many steps that can be optimized makes of the results obtained here a first approach to the full capabilities of this pipeline.

OpenMS and MSStats quantitation

	C2 vs C1		C3 vs C1		C4 vs C1	
	log2 FC	P-value	log2 FC	P-value	log2 FC	P-value
A	-0.27	9.99e-01	-3.02	1.04e-02	-3.25	5.12e-03
B	-2.63	2.12e-02	-5.07	4.08e-04	-0.23	8.51e-01
C	-1.58	1.42e-04	2.62	4.01e-06	2.43	4.80e-06
D	5.50	6.29e-09	4.97	1.40e-08	2.33	2.92e-06
E	-2.18	2.45e-04	-0.46	5.21e-01	5.71	6.26e-07
F	5.91	6.28e-07	0.10	9.14e-01	-1.11	3.75e-02

Table 5.5 OpenMS and MSStats results for the spiked proteins.

5.4.5 OpenMS and DEqMS

DEqMS does not support, in a straightforward way, importing results from OpenMS. Nevertheless, the results produced by OpenMS have been adapted to be imported by this statistical package writing and executing a Perl script. The results obtained are shown in Figure 5.12 and Table 5.6.

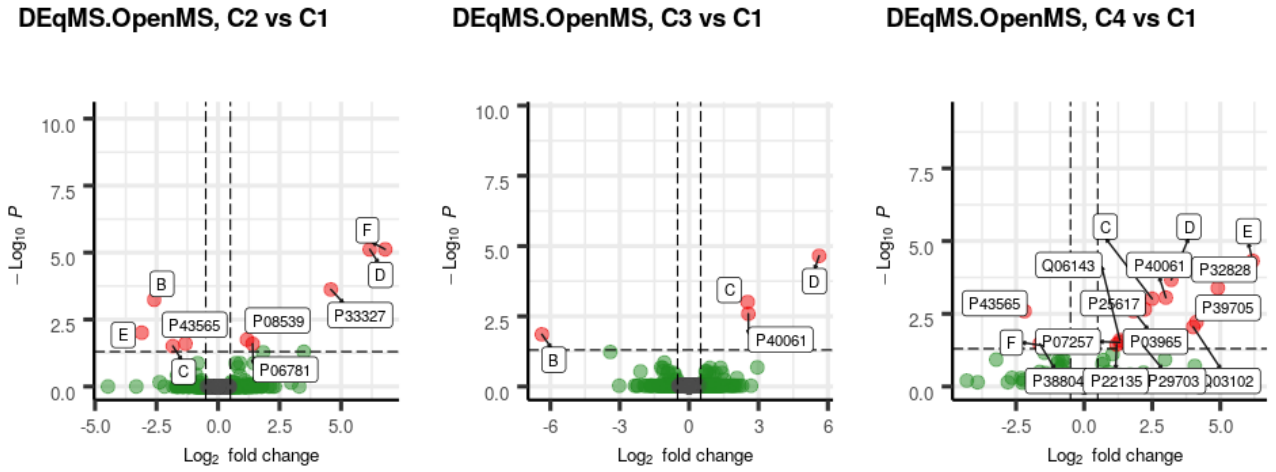


Figure 5.12 OpenMS and DEqMS results. 2527, 2529 and 2521 proteins have been respectively quantified for C2 vs C1, C3 vs C1 and C4 vs C1 comparisons.

In the same way that happened with the OpenMS and MSStats combined analysis, several non-spiked proteins appear in the three comparisons. Also, several spiked proteins disappear from the volcano plots, although for example, spike A is just below the threshold for C3 vs C1 (with an adjusted P-value of 5.9E-2 and a Log2Fold change of -3.4). The same conclusions can be extracted here: OpenMS conforms a highly complex pipeline and much room is left for improving parameters in the different steps that conform it.

As shown in Table 5.6, all spikes are reported, with values that are quite close to the theoretical fold changes shown in Table 5.1.

OpenMS and DEqMS quantitation

	C2 vs C1		C3 vs C1		C4 vs C1	
	log2 FC	P-value	log2 FC	P-value	log2 FC	P-value
A	0.10	1.00e+00	-3.41	5.93e-02	-4.31	6.29e-01
B	-2.60	5.73e-04	-6.37	1.40e-02	-0.06	8.95e-01
C	-1.84	3.10e-02	2.53	9.85e-04	2.50	9.63e-04
D	6.16	7.57e-06	5.62	2.24e-05	3.20	2.16e-04
E	-3.10	9.85e-03	-0.31	9.41e-01	6.19	4.80e-05
F	6.79	7.57e-06	-0.45	9.41e-01	-1.50	3.94e-02

Table 5.6 OpenMS and DEqMS results obtained for the spiked proteins.

5.4.6 Proteome Discoverer and MSStats

The results obtained after processing Proteome Discoverer results with MSStats are shown in Figure 5.13 and Table 5.7. The first thing that can be observed is the high amount of quantified proteins: this is actually the combination that includes more proteins reported. The parameters used by MSStats are very similar to those used with MaxQuant results, the only difference being that the import feature in MSStats for Proteome Discoverer does not include the removal of methionine containing peptides: it only allows the removal of peptides containing oxidized methionine residues; this option has not been used here, because some tests showed that no improvement was obtained with it.

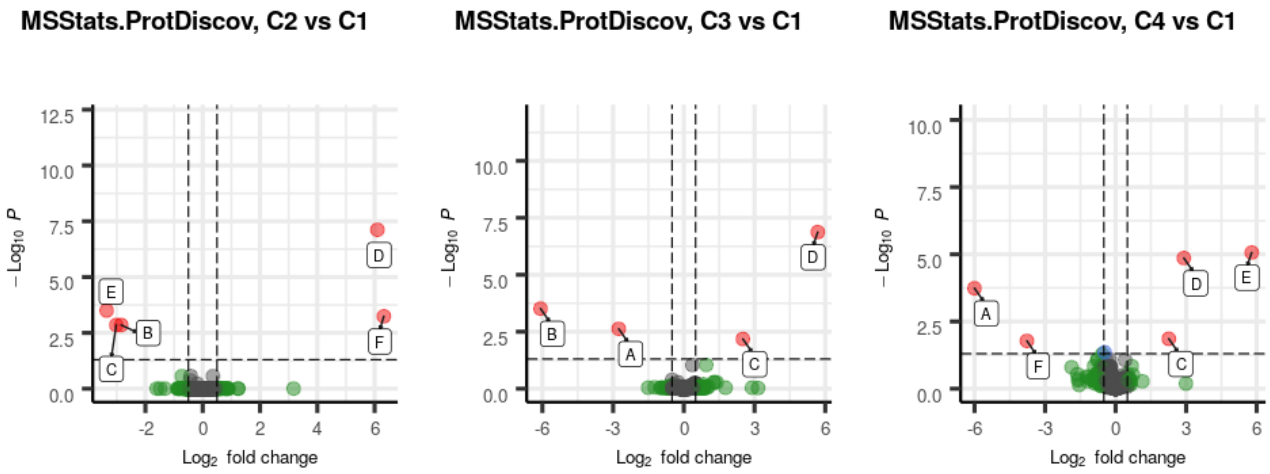


Figure 5.13 Proteome Discoverer and MSStats results. A total of 2854, 2852 and 2824 proteins proteins have been quantified for C2 vs C1, C3 vs C1 and C4 vs C1, respectively.

From Figure 5.13, it is clear that the results provided by Proteome Discoverer and MSStats are close to perfect: all the spikes that could be reported for the three comparisons are present in the volcano plots, and not a single non-spiked protein can be shown.

In Table 5.7, all spikes have been detected by this combination of software, even spikes with very low concentration levels (like B in C3 vs C1 and A in C4 vs C1 comparisons) have been properly quantified here.

Alongside with the MaxQuant and MSStats combination, this one will be selected for further discussion, as presenting the, *a priori*, most consistent results among all the different possibilities evaluated.

ProtDiscov and MSStats quantitation

	C2 vs C1		C3 vs C1		C4 vs C1	
	log2 FC	P-value	log2 FC	P-value	log2 FC	P-value
A	-0.39	9.94e-01	-2.76	2.37e-03	-6.00	1.83e-04
B	-2.85	1.41e-03	-6.08	3.07e-04	-0.14	7.30e-01
C	-3.02	1.41e-03	2.51	6.63e-03	2.26	1.39e-02
D	6.10	7.64e-08	5.68	1.34e-07	2.90	1.38e-05
E	-3.35	3.17e-04	-0.60	8.55e-01	5.79	8.59e-06
F	6.32	5.74e-04	-0.63	9.36e-01	-3.77	1.66e-02

Table 5.7 Proteome Discoverer and MSStats results obtained for the spiked proteins.

5.4.7 Proteome Discoverer and DEqMS

The results obtained with the last of the software combinations evaluated, Proteome Discoverer and DEqMS, are shown in Figure 5.14 and Table 5.8. As was the case with OpenMS and DEqMS, a Perl script has been used to make possible the import and analysis of the data produced by Proteome Discoverer by DEqMS.

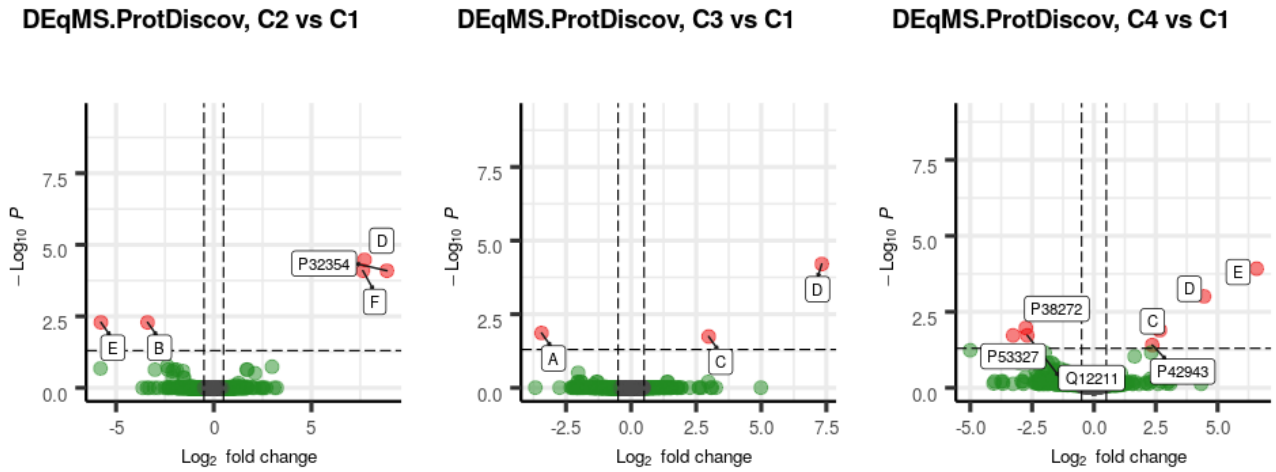


Figure 5.14 Proteome Discoverer and DEqMS results. A total of 2718, 2715 and 2666 proteins were quantified for C2 vs C1, C3 vs C1 and C4 vs C1, respectively.

Inspecting the volcano plots in figure 5.14, one non-spiked protein appear in C2 vs C1 and four in C4 vs C1. Also, similarly to what happened with MaxQuant and DEqMS, the B spike is missing from C3 vs C1 and A and F spikes from C4 vs C1 (although F is dismissed by presenting an adjusted P -value of 5.81E-2, just below the cutoff).

ProtDiscov and DEqMS quantitation

	C2 vs C1		C3 vs C1		C4 vs C1	
	log2 FC	P-value	log2 FC	P-value	log2 FC	P-value
A	-0.31	9.99e-01	-3.44	1.38e-02	0.00	0.00e+00
B	-3.39	5.17e-03	0.00	0.00e+00	-0.08	9.10e-01
C	-5.80	2.10e-01	2.98	1.81e-02	2.67	1.28e-02
D	7.72	3.35e-05	7.33	6.21e-05	4.46	9.83e-04
E	-5.78	5.17e-03	-0.58	9.85e-01	6.58	1.21e-04
F	7.63	8.07e-05	-0.48	9.85e-01	-5.00	5.81e-02

Table 5.8 Proteome Discoverer and DEqMS results obtained for the spiked proteins.

5.5 Discussion

From the results shown in the previous points, it is clear that the three pipelines tested (Proteome Discoverer, MaxQuant and OpenMS combined with MSStats, DEqMS and DEP) perform quite well in terms of proteins quantified and correct detection of the spikes (A to F letters) in terms of up and down-regulation of proteins levels. Some of them have not been detected in some combinations, but it is necessary to highlight here the challenging nature of the test: some of the spiked proteins show ratios of 50 to 1 in the upper region, while others present 1 to 30 in the lower zone. To evaluate the performance of the different pipelines, the parameters that will be prioritized will be the proper detection of the spikes (using adjusted P-value<0.05 and log2 fold change above/below ±0.5 cutoffs) and the absence of proteins not among the spiked (identified using Uniprot AC, e.g. P32354). The more spiked proteins and the less not-spiked in the significant area, the better.

Using this criteria, two pipelines out-perform the rest: MaxQuant and Proteome Discoverer combined with MSStats. It is important to insist that the different applications are not being compared here: choosing a more optimal set of analysis parameters, would surely have improved OpenMS, DEqMS and DEP results. But following the different documentation available, the results obtained are those that have been shown. In this way, in the next points, the results obtained with MaxQuant and Proteome Discoverer used in combination with MSStats are going to be discussed. The number of total proteins quantified and the ones that have been obtained as being significantly different (using two different thresholds) are shown in Table 5.9. From that table, the first thing to highlight is that the number of proteins quantified by Proteome Discoverer is roughly a 20% higher than using MaxQuant. This can be explained by more restrictive settings used by MaxQuant (the removal of peptides containing Methionine with MaxQuant is one example). From Figure 5.13 (Proteome Discoverer) and Figure 5.8 (MaxQuant), all spikes have been correctly quantified (with the only exception of B for C3 vs C1 in MaxQuant, that will be discussed below), and not “false positives” have been found with any of them: not proteins appear as differential using the Corrected P-value<0.05 and a Log2 fold change of ±0.5 used in all cases.

	MaxQuant MSStats			Proteome Discoverer MSStats		
	Total	**Sign.1**	**Sign.2**	**Total**	**Sign.1**	**Sign.2**
C2 vs C1	2365	22	5	2854	24	7
C3 vs C1	2360	43	4	2852	34	9
C4 vs C1	2335	66	7	2824	66	9

Table 5.9 Proteins quantified using MaxQuant and Proteome Discoverer combined with MSStats. The three comparisons are shown (C2, C3 and C4 samples compared to C1) and the total number of proteins quantified is shown. Also, the number of proteins detected as differentially significant is shown, using two thresholds: for **Sign.1**, a P-value<0.05 and a Log2 fold change of ±0.5, for **Sign.2**, a P-value<0.05 and a Log2 fold change of ±1. In both cases, the P-values are not adjusted.

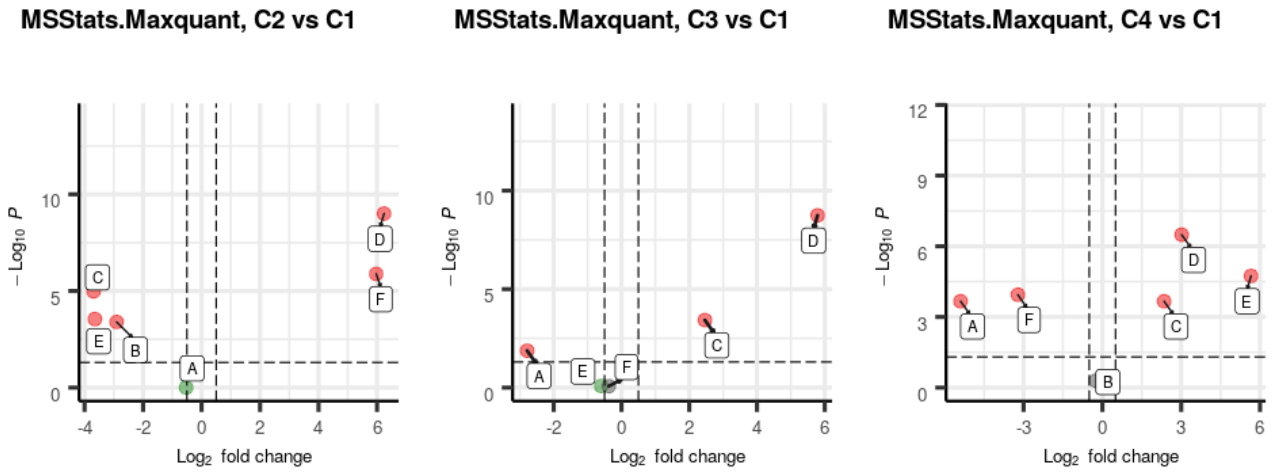
Both Proteome Discoverer and MaxQuant perform similarly in terms of results (with more relaxed cutoffs, as will be seen later, MaxQuant performs better), but one important difference among them must be emphasized here: Proteome Discoverer is a software that works only with results obtained using a Thermo Fisher Scientific instruments (like the one used to analyze the samples used in this work) and therefore, can not be used with data coming from other instrument manufacturers. Also, Proteome Discoverer is a proprietary, not free of charge software, while MaxQuant is a freeware (5) application.

As an illustration of the spikes (letters A to F) detection by the search engines, in Figure 5.15, the performance of MaxQuant and MSStats is shown, using three volcano plots where all non-spiked proteins have been removed and one table showing the molecular weight of each spiked protein in KDaltons, with their concentration in each sample and the number of peptides obtained for each sample.

From Figure 5.15, the spike B under C3 vs C1 appears with a theoretical Log2 Fold Change value of -4.78, corresponding to 2 fmols of protein B in sample C3 and 55 fmols of protein B in sample C1. The value of 2 fmols in the volume used for injection, although can be close to the detection limit of the instrument used, has proved to be detectable by using OpenMS and Proteome Discoverer, both providing values for B in this comparison. The cutoffs used with MaxQuant, apparently more severe for this protein than with the other two software tools, have eliminated this protein in the MaxQuant results from C2 replicates: the protein is simply not there. That can be explained by the low scores that peptides at low concentrations are given: in this case, all peptides obtained for protein B at C3 sample have been removed by MaxQuant (for C1, C2 and C4 samples and average of 11, 5 and 10 unique peptides have been found, respectively, as shown in Figure 5.15).

From observing the other spiked proteins in Table 5.15, the number of peptides detected for quantification, can be correlated with two effects:

- Any given peptide, at a very low concentration in some sample, will be detected with lower scores by the fact that its associated mass spectrum will be of lower quality (23). This fact makes that for a less concentrated protein, fewer peptides will be identified with enough reliability and therefore, will be lost for quantification.
- Secondly, the length of the protein studied will directly produce a higher amount of different tryptic peptides. The number of those peptides, combined with the concentration of the protein in a given sample, will generate a higher or lower amount of peptide available for quantification.



	MW	Spikes (fmol)				Sample C1			Sample C2			Sample C3			Sample C4		
	KDa	C1	C2	C3	C4	1A	1B	1C	C2A	C2B	C2C	C3A	C3B	C3C	C4A	C4B	C4C
A	45	65	55	15	2	9	7	5	8	6	7	1	3	2	1	0	1
B	17	55	15	2	65	12	10	11	4	6	6	0	0	0	10	11	10
C	97	15	2	65	55	33	36	34	2	3	1	62	66	59	59	61	57
D	116	2	65	55	15	8	10	10	57	53	50	53	51	49	28	30	30
E	66	11	0.6	10	500	27	25	26	2	2	8	20	21	21	85	85	84
F	29	10	500	11	0.6	5	5	4	25	22	21	5	5	5	2	2	1

Figure 5.15 Above, spiked proteins (letters A to F) shown as a significantly different in the three conditions under study (C2 vs C1, C3 vs C1 and C4 vs C1) using three volcano plots where all non-spiked proteins have been removed. Below, a table showing the molecular weight of each spiked protein in KDaltons, their concentration in each sample (in femto mols) and the number of unique-mapping peptides obtained for each sample.

From the results obtained using MaxQuant and MSStats one fact can be underlined: using more relaxed cutoffs with MaxQuant would surely have allowed the quantification of B spike under C3 vs C1 comparison, but at the same time, would have surely generated a higher amount of proteins false quantified as significantly different between samples.

5.5.1 Cutoffs for differential expression

The cutoff used in the previous sections has been, in all cases, an adjusted P-value of 0.05 and a Log2 fold change of ±0.5 (amounting to a positive ratio of 1.41 and a negative of 0.71). That approach, used in the original publication where the data analyzed here was made publicly available (1), has been questioned in the literature as being “too restrictive” or simply “blunt” (24). The use of corrected P-values as threshold, a practice frequently used in genomics (25), has demonstrated to be too restrictive in proteomics , mainly because the lack of power in proteomics measurements caused by a low number of replicates combined with ratio compression; ratio compression refers to lower fold changes obtained in proteomics with respect to the ones obtained in transcriptomics. Other alternatives have been suggested instead of the use of adjusted p-values; among them, the control of false positives at the peptide level, something that has been carried on along this study.

Having said this, at the end, some threshold has to be chosen, in order to identify proteins that are expressed in different ways among different phenotypes. It is precisely the data obtained in this work the kind of material that can effectively be used to delimit a threshold that is not arbitrary: evaluating the number of detected spikes and the number of non-spiked proteins, using a given threshold, provides a clear picture of what is going to

be obtained when complex biological material is under study. Analyzing several times the same dilution to be studied (technical replicates), the precision of the proteomics analysis can be assessed and also, a threshold to be used with real comparisons can be obtained.

Some considerations must be observed, though: the cutoff obtained, after delimiting the non-spiked proteins into optimal intervals of P-value and fold change, should not be extrapolated to other experimental set-ups (concentration ranges used, type of sample, chromatography column, solvents,…) or instrument used (peptide ionization and detection will change drastically using different instruments).

In Figure 5.16, a new cutoff has been applied on the proteins quantified by MaxQuant and Proteome Discoverer in combination with MSStats: a P-value (not corrected) of 0.05 and a fold change of ±1. As the six volcano plots demonstrate (the three conditions compared by the two pipelines), these cutoffs are quite effective in separating the non-spiked proteins from the spiked ones. Only three proteins are incorrectly classified as differential using MaxQuant while 12 are using Proteome Discoverer. It is important to appreciate that the fold change has been increased from ±0.5 to ±1: using a ±0.5 fold change cutoff with a non-corrected P-value would have generated many proteins wrongly characterized as changing in a significant way.

While the use of the more strict “adjusted P-value” threshold is very useful for highlighting proteins that have been quantified with stronger confidence, it is clear that many proteins that were actually differentially expressed will be lost by using this filter.

Finally, using this new threshold is evident that MaxQuant performs better in terms of detecting spiked proteins while leaving out background proteins. One of the reasons for this is that more strict cutoffs have been used with the default configuration employed by MaxQuant. On the other hand, it is important to note that about 20% more proteins have been quantified by using Proteome Discoverer: this comparison is not completely fair.

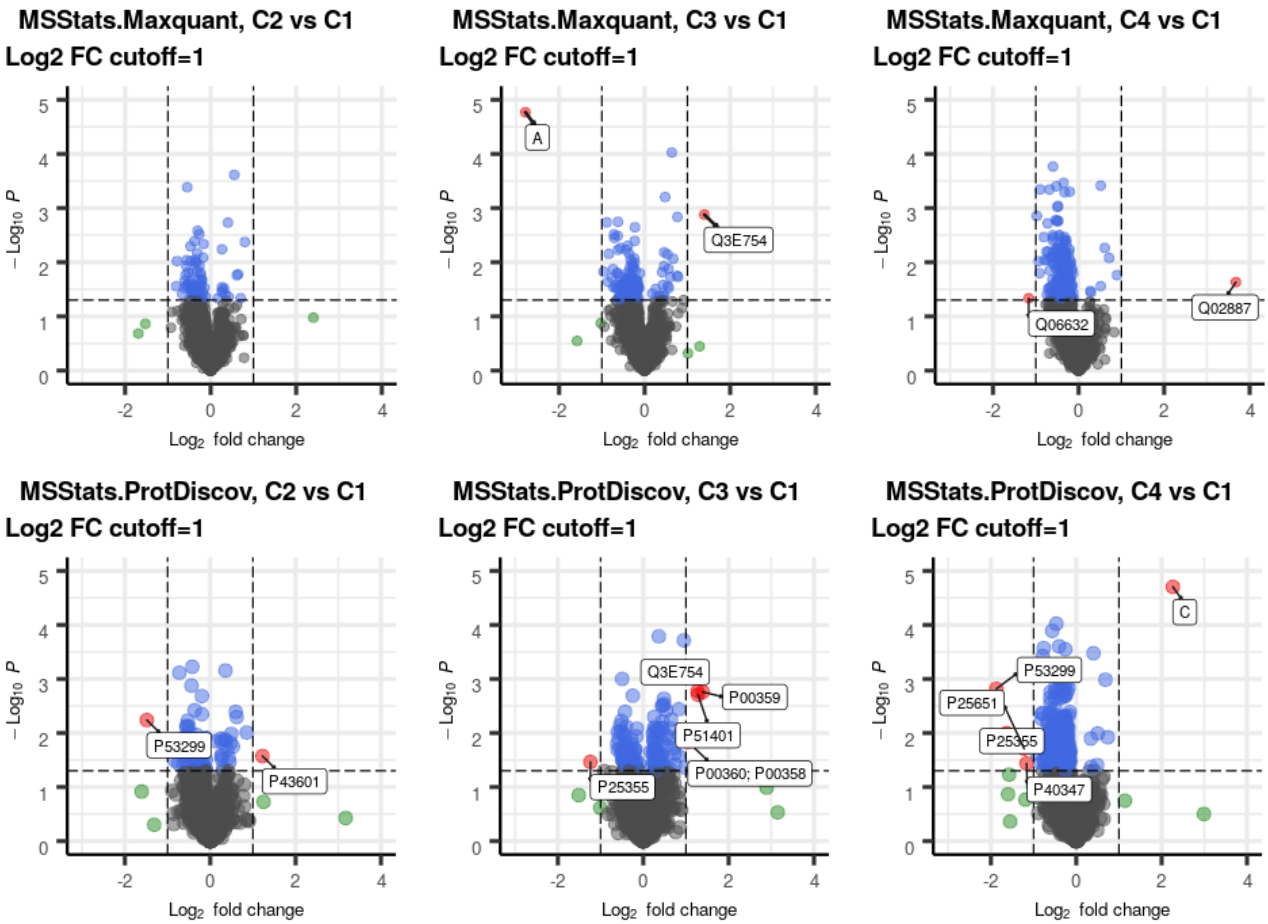


Figure 5.16 Volcano plots of the proteins quantified using MaxQuant (above) and Proteome Discoverer (below) for the three conditions studied (C2 vs C1, C3 vs C1 and C4 vs C1) using a P-value (not corrected) of 0.05 and a Log2 fold change of ±1. The range of the volcano plots have been zoomed to better appreciate the protein distributions, leaving most of the spiked proteins outside the inspected range. In red, proteins passing both thresholds, in green proteins passing only the fold change threshold, in blue proteins passing the P-value threshold and in black proteins passing none.

5.5.2 Quantification accuracy

The differences between the real values of the spiked proteins (the ratios obtained from known concentrations) and the experimental ratios obtained (using MaxQuant and Proteome discoverer) are evaluated here. The results are shown in Table 5.10. In all cases, the true nature of the differential expression has been detected: over and under expression are perfectly characterized in all cases. Only in the case of the detection of the spiked protein B for C3 vs C1 using MaxQuant, an issue has appeared. It is an interesting fact that the same ratio, generated by Proteome Discoverer, is four times lower than the true ratio (0.01 in front of 0.04), pointing to a difficult detection of the peptides from B into the C3 sample.

	Theoretical ratios			MaxQuant and MSStats			ProtDiscov and MSStats		
	C2 vs C1	C3 vs C1	C4 vs C1	C2 vs C1	C3 vs C1	C4 vs C1	C2 vs C1	C3 vs C1	C4 vs C1
A	0.85	0.23	0.03	0.69	0.14	0.02	0.76	0.15	0.02
B	0.27	0.04	1.18	0.13	-	0.84	0.14	0.01	0.91
C	0.13	4.33	3.67	0.08	5.50	5.10	0.12	5.70	4.79
D	32.50	27.50	7.50	75.58	55.33	8.11	68.59	51.27	7.46
E	0.05	0.91	45.45	0.08	0.66	50.56	0.10	0.66	55.33
F	50.00	1.10	0.06	62.68	0.77	0.11	79.89	0.65	0.07
				0.869	0.998	0.999	0.974	0.998	0.999
					R^2			R^2	

Table 5.10 Ratios for the three comparisons (C2 vs C1, C3 vs C1 and C4 vs C1) of the six spiked proteins (A to F) are shown here in three tables: first, the theoretical values calculated from the known concentrations; then, the values obtained using MaxQuant and Proteome Discoverer. Three squared correlation coefficients (R^2, one for each comparison) are shown at the bottom for each of the two software pipelines.

As the correlation coefficients shown in Table 5.10 show, the two software pipelines have performed quite well, although some serious bias can be observed for the more concentrated proteins in both cases. The accuracy of the measurements, though, is far from ideal: accuracy found in the relative quantifications performed here is over 20% in most cases. Said this, the detection of the overall tendencies and an approximation to the true intensities of changes is achieved in a remarkable way by the two software pipelines used here.

5.5.3 Censored values and imputation

Censored values, in statistics, correspond to those values that are unknown for being above of a given point (right censoring), in a given interval (interval censoring) or below some point (left censoring) (26). Imputation, is defined as the mechanism that deals with censored values assigning numerical magnitudes to them. In proteomics, two scenarios have been described (27):

- Missing Completely At Random (MCAR): the propagation of minor errors and random effects (peptides out of the retention time window in quantitation or below the cutoffs in the identification steps) generate a non-quantified peptide. This is a sub-type of the Missing At Random (MAR) case, and both can be treated equally in the proteomics landscape (28). Typically in this scenario, for a given protein, some peptides are measured and others are not in one condition, while different peptides can then be measured in a different condition. Imputation here is required in order to generate more accurate results for differential expression.
- Missing Not at Random (MNAR): missing values respond to a clear cause. In proteomics, MNAR values are obtained because of a left censoring process: if all peptides from a given protein are below the detection level, this protein will not be quantified for a given phenotype or condition under study. For other conditions, the concentrations can be well above the quantification threshold and therefore, some way to deal with this scenario must be found if ratios (or fold changes) are going calculated.

The two types of missing values (MCAR and MNAR) are going to be discussed using examples that have appeared in the analysis performed throughout this chapter.

5.5.3.1 Example 1: dealing with MCAR values

In order to detect some case where MCAR values are important in the differential expression under study, the MaxQuant with MSStats pipeline has been run again twice: in one case using imputation and in the other, not using it. The results obtained are shown in Figure 5.17: when not using imputation at all, several proteins appear as diferentially expressed in C3 vs C1 comparison (P38850) and C4 vs C1 comparison (P46675, P45000 and P38850). As those proteins are expected to have the same concentrations in the four groups, it is clear that some error has been introduced when not using imputation.

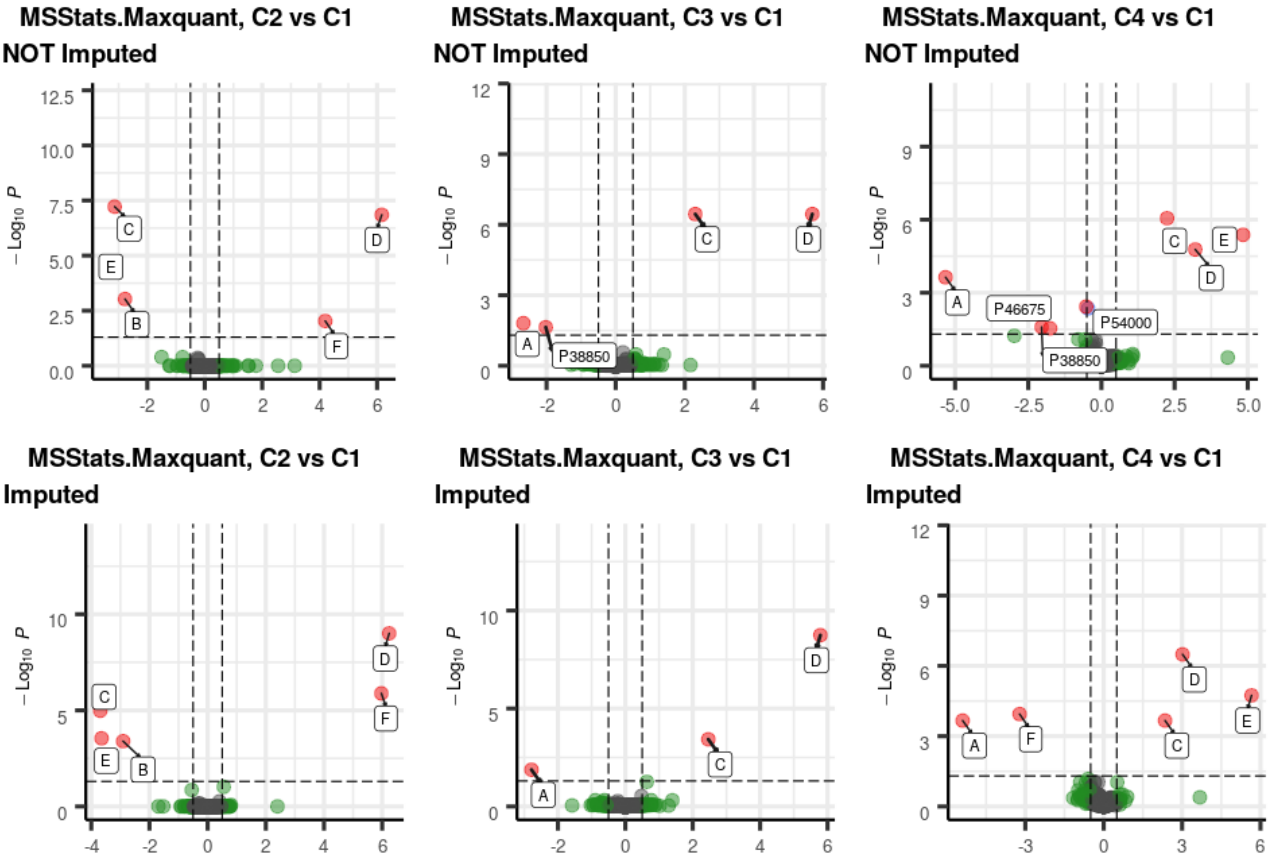


Figure 5.17 Results obtained using MaxQuant and MSStats pipeline not using imputation (above) and using “Accelerated failure model” imputation (below): several proteins appear as deferentially expressed in the C3 vs C1 comparison (P38850) and in the C4 vs C1 comparison (P46675, P45000 and P38850) in the “NOT Imputed” model. Also, F protein spike disappears from C4 vs C1 if imputation is not used.

To illustrate why those proteins have been wrongly labeled as differentially expressed, the protein P38850 for the C3 vs C1 comparison has been inspected. In Table 5.11, the results obtained for this protein with the MaxQuant-MSStats pipeline, both using or not using imputation are shown: without imputation, protein P38850 is not detected in C2 vs C1, but obtains a significant difference (adjusted P-values below 0.05) and intense fold changes (four times less concentrated) with C3 vs C1 and C4 vs C1 comparisons. That means that protein P38850 is completely missing from C2 with or without imputation (resembling a case of censored value of the MNAR type, where this protein would be below the detection limit) and detected four times less concentrated in C3 and C4 with respect to C1.

Protein P38850 Not Imputed

Comparison	log2FC	SE	adj.pvalue	issue	Missing Percentage	Imputation Percentage
C2vsC1	-Inf	NA	0.00E+00	OCM	75.00%	0.00%
C3vsC1	-2.023	0.053	2.30E-02	none	58.33%	0.00%
C4vsC1	-2.044	0.067	2.57E-02	none	66.67%	0.00%

Protein P38850 Imputed

Comparison	log2FC	SE	adj.pvalue	issue	Missing Percentage	Imputation Percentage
C2vsC1	-Inf	NA	0.00E+00	OCM	75.00%	25.00%
C3vsC1	0.024	0.042	9.42E-01	none	58.33%	41.67%
C4vsC1	0.007	0.053	9.62E-01	none	66.67%	33.33%

Table 5.11 Values obtained from the differential analysis with MaxQuant and MSStats for protein P38850 in the three comparisons (C2 vs C1, C3 vs C1 and C4 vs C1) are shown, not using imputation (above) and using “Accelerated failure model” imputation (below). Column SE refers to “Standard error” and column “issue” highlights if an issue has appeared in the quantification: “none” or “One Condition Missing (OCM)”.

Further inspecting this protein, only two peptides are used in P38850 quantification (Table 5.12): each replicate makes use of only those two peptides and, for some samples, only two (in C3) and even one (in C4) peptides are used for quantification. The fact that in sample C1 only one peptide is found (DCQVYISK) and in samples C3 and C4 the other one is found (CINLVNDIPGGVDTIGSVLK), makes this protein prone to quantification errors if imputation is not used at all. It is important to remark here that the intensity values obtained for these two peptides are just below the median (about 1E+07) and well above of the minimum intensities (about 1.5E+06) obtained from all the peptides analyzed in the study: that means that, being low, the concentration of these peptides is well above the limit of detection in the samples, placing these censored values inside the Missing Completely At Random category.

Sample	Replicate	Peptide	Intensity
C1	1	CINLVNDIPGGVDTIGSVLK_2	NA
	1	DCQVYISK_2	1.08E+07
	2	CINLVNDIPGGVDTIGSVLK_2	NA
	2	DCQVYISK_2	1.15E+07
	3	CINLVNDIPGGVDTIGSVLK_2	NA
	3	DCQVYISK_2	9.84E+06
C2	1	CINLVNDIPGGVDTIGSVLK_2	NA
	1	DCQVYISK_2	NA
	2	CINLVNDIPGGVDTIGSVLK_2	NA
	2	DCQVYISK_2	NA
	3	CINLVNDIPGGVDTIGSVLK_2	NA
	3	DCQVYISK_2	NA
C3	1	CINLVNDIPGGVDTIGSVLK_2	NA
	1	DCQVYISK_2	NA
	2	CINLVNDIPGGVDTIGSVLK_2	3.00E+06
	2	DCQVYISK_2	NA
	3	CINLVNDIPGGVDTIGSVLK_2	2.70E+06
	3	DCQVYISK_2	NA
C4	1	CINLVNDIPGGVDTIGSVLK_2	NA
	1	DCQVYISK_2	NA
	2	CINLVNDIPGGVDTIGSVLK_2	NA
	2	DCQVYISK_2	NA
	3	CINLVNDIPGGVDTIGSVLK_2	2.59E+06
	3	DCQVYISK_2	NA

Table 5.12 Peptide intensities of the two peptides quantified for protein P38850 (NA values for not quantified peptides). Only two different peptides have been quantified (CINLVNDIPGGVDTIGSVLK and DCQVYISK, both with charge 2), none for sample C2.

Fold change and probability values obtained for P38850 using imputation (Table 5.11), with Accelerated Failure Time model (29,30), reflect more accurately the true concentrations of

the protein and, more importantly, prevent the onset of artifacts like the one described here.

5.5.3.2 Example 2: dealing with MNAR values

When all peptides for a given protein are under the limit of detection, not a single peptide will be associated to the protein and therefore, the corresponding ratio will not be calculated and reported by the software in use. It is important to note that all peptides associated to a certain protein for a given experimental condition (e.g. phenotype) must act in a coordinated way: if some peptides or features show concentrations far from the detection limit, this will be a case of MCAR censoring, not MNAR.

The value obtained for the spike B under the C3 vs C1 comparison and using MaxQuant in combination with MSStats and DEP, will allow the inspection of the two different approaches used when a Missing Not at Random value is found in a proteomics experiment:

- The approach followed by the software DEP when imputation is used (Figure 5.10) is assigning a value close to 0 (using some minimal value across features), and very similar (or equal) between replicates. That will generate a set of proteins where a maximum P-value is reached (in Figure 5.10, that corresponds to a corrected P-value of 1.09E-13 for C3 vs C1) and for B, an extreme Log2 fold change of -7.190.
- The approach followed by MSStats (Figure 5.7), even when using imputation (left censoring assumed) is not providing any results associated to this protein for the C3 vs C1.

Not reporting a given protein, like MSStats does, is quite problematic: if one protein is perfectly quantified in a given experimental condition and disappears in another condition, this can be the result of some important biological process, that will go unaccounted for if not reported in some way. On the other hand, the imputation approach used by DEP generates several proteins that are simply artifacts, not being spiked proteins and detected as deferentially expressed.

In the case exposed here, it is clear that the B spike is present in both samples, because it is correctly detected by, for example, Proteome Discoverer. The only reason that makes spike B going undetected in C3 vs C1 using MaxQuant is that none of the peptides have passed the cutoffs established by the software.

Using the MSStats imputation approach, given the quality of the results provided, seems the more sensible thing to do; but this will mean that, in some cases, some of the proteins experimenting dramatic under-expression in a biological context can go unnoticed.

5.5.3.3 Coexistence of MCAR and MNAR values: a global strategy

One of the causes that makes imputation a serious issue in label free proteomics is the fact that both MCAR and MNAR censored values coexist (31). Several mathematical tools are used to deal with both kinds of missing values but, unfortunately, the are only well suited to work with one kind at a time.

The “accelerated failure time” method used by MSStats in this work, assumes that all missing values are produced by left censoring (32). In Figure 5.13, the effect and the extent of imputation among the values obtained from the protein spikes under C2 vs C1 is shown.

	True ratios	WITHOUT imputation C2 vs C1			WITH imputation C2 vs C1			
Spike-in	C2vsC1	log2 FC	adj Pvalue	Missing %	log2 FC	adj Pvalue	Missing %	Imputation %
A	-0.24	-0.48	9.98E-01	2.38%	-0.53	9.94E-01	2.38%	2.38%
B	-1.87	-2.78	9.37E-04	37.96%	-2.91	4.03E-04	37.96%	37.96%
C	-2.91	-3.14	6.04E-08	77.59%	-3.69	1.05E-05	77.59%	77.59%
D	5.02	6.16	1.41E-07	47.98%	6.24	9.88E-10	47.98%	47.98%
E	-4.20	-3.43	4.28E-05	86.07%	-3.65	2.85E-04	86.07%	86.07%
F	5.64	4.19	9.29E-03	40.38%	5.97	1.31E-06	40.38%	40.38%

Table 5.13 Log2 Fold changes and corrected P-values obtained using MaxQuant and MSStats without (left) and with (right) imputation. The percentage of missing values (peptides) for every protein into the three possible replicates of each sample is provided under Missing%. When imputation is used, all missing values are imputed.

From Table 5.13, several observations can be made:

- Imputation here, does not greatly affect the fold changes obtained with respect to the ones obtained without imputation: in some cases, the values obtained resemble more to the “true ratios” and in some cases, less.
- For some proteins, imputation values as high as the 86% are achieved. That does not mean that 86% of peptides are undetected: one peptide can be detected many times in the case of an abundant protein, in several retention times and with different charge states: this percentage refers to “features”, not peptides. Because three replicates of the same sample are used in this quantification, it is very common that for one sample a peptide is detected in a given retention time for a replicate and not detected in the other two. That said, for the proteins evaluated in the table, quite good levels of accuracy have been achieved even with the high levels of missing values obtained. In Figure 5.15, where the number of peptides for each condition is shown, one possible explanation for this effect emerges: in the case of protein E, where the highest amount of missing values has been found (86.07%), more that 25 unique peptides are found in replicates from sample C1, while only 2 unique peptides are found for two of the replicates from sample C2: in every case, when an unmatched peptide is found among samples, a missing value will appear.
- The number of total and unique peptides used for quantification for every protein are both important variables that can be used to assess the confidence for the concentration levels given for a certain protein under one comparison. This is possible when using MSStats, but not in a straightforward way.

At the moment of writing this work, a complete solution for the treatment of censored values in label free proteomics is not yet available (27,33–35). One possible strategy, after all that has been discussed in this chapter, would follow the next steps:

- MSStats, with “Accelerated failure model” imputation will be used as the primary resource for obtaining quantitation values. Using P-value=0.05 and Log2 fold change ±1 as cutoffs.
- A second analysis, using DEP with a MNAR imputation approach (e.g. “MinProb”) and highly restrictive cutoffs (Adjusted P-value=0.05, Log2 fold change=±1) would

allow to highlight proteins that could have been removed from the report obtained in the previous point.

- Inspection of the peptides associated to the quantified proteins, both in terms of number and identification quality, will provide basis for inclusion or rejection of a given protein from the final quantified set.
- Additionally, several imputation approaches can be tried with DEP, in order to retrieve more information about the way that missing data behaves into the data set under study.

5.6 Conclusions

The reanalysis of the iPRG2015 data set has allowed the adoption of a set of criteria that can be used as a guideline to perform label free quantitative analyses, defining cutoffs and providing certain levels of confidence on the results obtained. It is certain that the values obtained here can only be consistently applied if the same experimental setup is used: the complexity of the samples, the instrumentation employed and experimental protocols followed should resemble the ones used to generate the data used here if the same values are going to be employed. Nevertheless, the preparation of samples resembling the ones that are used by the iprg2015 study and their analysis with the instrumentation employed by a proteomics facility, should be something completely feasible, both in terms of work and economic burden.

The use of MaxQuant and MSStats as the primary software pipeline seems clear from the results obtained. Ease of use in an automated way (MaxQuant and MSStats can easily be used in conjunction with Slurm in a Linux environment), and the consistency of the results obtained make of this an ideal combination for a proteomics facility to use. Other software packages and pipelines can be used as well in conjunction of the aforementioned.

It is also clear that the use of MaxQuant and MSStats provides a robust quantification pipeline, offering an effective detection of proteins differentially expressed in samples. However, the accuracy on the ratios obtained with the different software packages used, particularly with very low or very high ratios, is far from ideal.

The use of a P-value of 0.05 and Log2 fold changes ±1 as cutoff values, with the MaxQuant and MSStats combination, provides a reliable filtering system where almost all proteins labeled as significantly detected can be relied on.

Finally, as a strategy for dealing with imputation of missing values in label free quantitation, the use of MSStats and an “Accelerated failure model”, although far from being perfect, has proved to be the most accurate. Additionally, the use of different imputation strategies with DEP may help to detect protein ratios that would go unaccounted for if only MSStats is employed.

5.7 References

1. Choi M, Eren-Dogu ZF, Colangelo C, Cottrell J, Hoopmann MR, Kapp EA, et al. ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of Differentially Abundant Proteins in Label-Free Quantitative LC-MS/MS Experiments. J Proteome Res. 2017 03;16(2):945–57.

2. Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Ríos D, et al. ProteomeXchange provides globally co-ordinated proteomics data submission and dissemination. Nat Biotechnol. 2014 Mar;32(3):223–6.

3. Li W. Volcano plots in analyzing differential expressions with mRNA microarrays. J Bioinform Comput Biol. 2012 Dec;10(6):1231003.

4. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008 Dec;26(12):1367–72.

5. Sinitcyn P, Tiwary S, Rudolph J, Gutenbrunner P, Wichmann C, Yılmaz Ş, et al. MaxQuant goes Linux. Nat Methods. 2018;15(6):401.

6. Pfeuffer J, Sachsenberg T, Alka O, Walzer M, Fillbrunn A, Nilse L, et al. OpenMS - A platform for reproducible analysis of mass spectrometry data. J Biotechnol. 2017 Nov 10;261:142–8.

7. Sturm M, Bertsch A, Gröpl C, Hildebrandt A, Hussong R, Lange E, et al. OpenMS - an open-source software framework for mass spectrometry. BMC Bioinformatics. 2008 Mar 26;9:163.

8. Sweet SMM, Jones AW, Cunningham DL, Heath JK, Creese AJ, Cooper HJ. Database Search Strategies for Proteomic Data Sets Generated by Electron Capture Dissociation Mass Spectrometry. Journal of Proteome Research. 2009 Dec 4;8(12):5475–84.

9. pubmeddev, VJ RM and C. Bioconductor: an open source framework for bioinformatics and computational biology. - PubMed - NCBI [Internet]. [cited 2019 Sep 5]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/16939789

10. Choi M, Chang C-Y, Clough T, Broudy D, Killeen T, MacLean B, et al. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics. 2014 Sep 1;30(17):2524–6.

11. Zhu Y. DEqMS: a tool to perform statistical analysis of differential protein expression for quantitative proteomics data. 2019.

12. Zhang X, Smits AH, van Tilburg GB, Ovaa H, Huber W, Vermeulen M. Proteome-wide identification of ubiquitin interactions using UbIA-MS. Nat Protoc. 2018;13(3):530–50.

13. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016;13(9):731–40.

14. Mono project [Internet]. Available from: https://www.mono-project.com/

15. Slurm Workload Manager [Internet]. Available from: https://slurm.schedmd.com/

16. Warr WA. Scientific workflow systems: Pipeline Pilot and KNIME. J Comput Aided Mol Des. 2012 Jul;26(7):801–4.

17. Tabb DL. The SEQUEST Family Tree. J Am Soc Mass Spectrom. 2015 Nov;26(11):1814–9.

18. The M, MacCoss MJ, Noble WS, Käll L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. J Am Soc Mass Spectrom. 2016;27(11):1719–27.

19. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:Article3.

20. Blighe K. EnhancedVolcano: publication-ready volcano plots with enhanced colouring and labeling [Internet]. Available from: https://github.com/kevinblighe/EnhancedVolcano

21. The Universal Protein Resource (UniProt). Nucleic Acids Res. 2008 Jan;36(Database issue):D190–5.

22. Ghesquière B, Gevaert K. Proteomics methods to study methionine oxidation. Mass Spectrom Rev. 2014 Apr;33(2):147–56.

23. Urban PL. Quantitative mass spectrometry: an overview. Philos Trans A Math Phys Eng Sci [Internet]. 2016 Oct 28 [cited 2019 Sep 17];374(2079). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5031646/

24. Pascovici D, Handler DCL, Wu JX, Haynes PA. Multiple testing corrections in quantitative proteomics: A useful but blunt tool. Proteomics. 2016;16(18):2448–53.

25. Lystig TC. Adjusted P values for genome-wide scans. Genetics. 2003 Aug;164(4):1683–7.

26. Lagakos SW. General right censoring and its impact on the analysis of survival data. Biometrics. 1979 Mar;35(1):139–56.

27. Lazar C, Gatto L, Ferro M, Bruley C, Burger T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. J Proteome Res. 2016 Apr 1;15(4):1116–25.

28. Karpievitch YV, Dabney AR, Smith RD. Normalization and missing value imputation for label-free LC-MS analysis. BMC Bioinformatics. 2012;13 Suppl 16:S5.

29. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. Br J Cancer. 2003 Aug 4;89(3):431–6.

30. Hougaard P. Fundamentals of survival data. Biometrics. 1999 Mar;55(1):13–22.

31. Webb-Robertson B-JM, Wiberg HK, Matzke MM, Brown JN, Wang J, McDermott JE, et al. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. J Proteome Res. 2015 May 1;14(5):1993–2001.

32. Taylor SL, Leiserowitz GS, Kim K. Accounting for Undetected Compounds in Statistical Analyses of Mass Spectrometry ‘Omic Studies. Stat Appl Genet Mol Biol. 2013 Dec 1;12(6):703–22.

33. O’Brien JJ, Gunawardena HP, Paulo JA, Chen X, Ibrahim JG, Gygi SP, et al. The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. Ann Appl Stat. 2018 Dec;12(4):2075–95.

34. Li Q, Fisher K, Meng W, Fang B, Welsh E, Haura EB, et al. GMSimpute: a generalized two-step Lasso approach to impute missing values in label-free mass spectrum analysis. Bioinformatics. 2019 Jun 14;

35. Tang J, Zhang Y, Fu J, Wang Y, Li Y, Yang Q, et al. Computational Advances in the Label-free Quantification of Cancer Proteomics Data. Curr Pharm Des. 2018;24(32):3842–58.

Appendix 1: Chapter4, Phenotypes inspected

Contents

Appendix 1: Chapter4, Phenotypes inspected

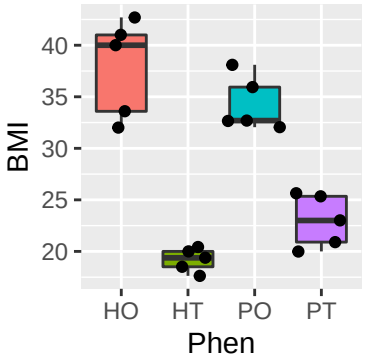
Dividing subjects into four classes

The 20 subjects analyzed using proteomics have been extracted from a bigger cohort (162 individuals). The criteria for generating the four groups, HT-HO-PT-PO is their diagnose of PCOS (P from PCOS, H from Healthy) and their Body mass index (O from obese, with BMI higher than 30, and T from thin, with BMI lower than 30). Here, using clinical variables collected for the 20 patients and the bigger cohort, the four groups will be studied using Principal components analysis (PCA) and Logistic regression.

Some of the procedures followed in this analysis have been inspired from the "Handbook of Biological Statistics" (John McDonald) and its "R Companion" (Salvatore S. Mangiafico).

First, an exploration of the four groups divided according to their BMI is performed. A boxplot for each group is built and each subject represented as a black dot.

```
clinical.data.df<-read.csv(file = "20patients.clinical.csv",header = TRUE)
bmi.boxplot <- ggplot(clinical.data.df, aes(x = Phen, y = BMI, fill=Phen)) +
  geom_boxplot() + theme(legend.position = "none")
bmi.boxplot <- bmi.boxplot + geom_jitter()
bmi.boxplot
```



Principal components analysis

A PCA is performed to assess the grouping of patients using the clinical variables. Twenty-four clinical variables are considered: hirsutism, menarche, FM, homaindex, HDL, triglycerides, glucose, cholesterol, LDL, insulin, testosterone, SDHA, hidroxi, LH, FSH, estradiol, thyrotropin, ast, Alt, LH.FSF, prolactina, cortisol, freeT4 and androstenedione. The PCA summary shows that 80% of variability is reached at PC6.

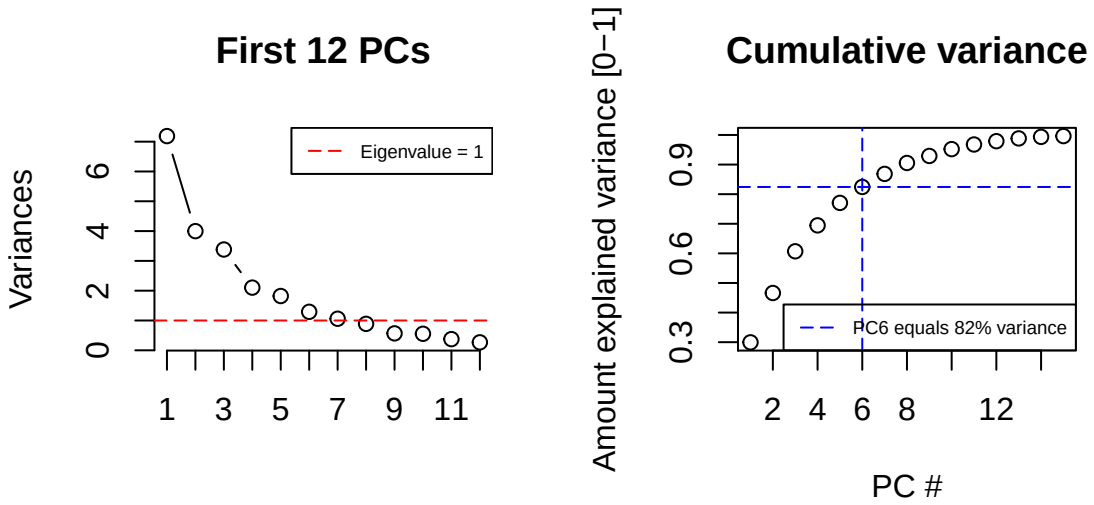
```
numeric.data<-as.data.frame(clinical.data.df[c(10:33)])
rownames(numeric.data)<-clinical.data.df$Patient
patients.clinical.data.pr <- prcomp(numeric.data, center = TRUE, scale = TRUE)
summary(patients.clinical.data.pr)
```

Two graphs below:

- the first 7 PCs are over the eigenvlaue=1, remaining the higher (until 12 in this graph) below. That means that they do not further add significative information

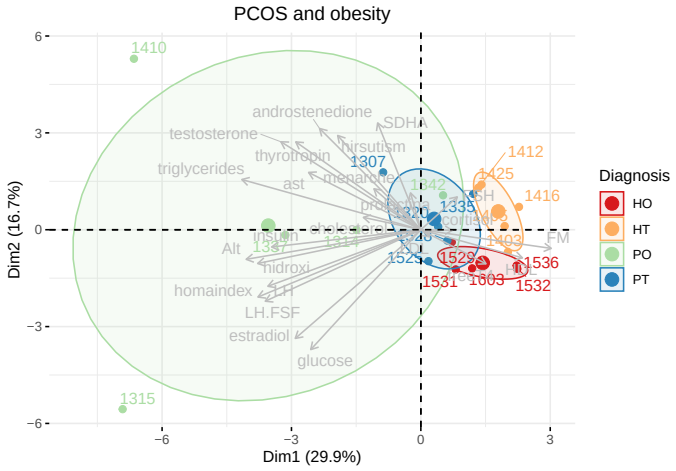
- a graphical display of the fact that 82% of variability is reached at PC6

```
plot(patients.clinical.data.pr, type = "l", npcs = 12, main = "First 12 PCs")
abline(h = 1, col="red", lty=5)
legend("topright", legend=c("Eigenvalue = 1"), col=c("red"), lty=5, cex=0.6)

cumpro <- cumsum(patients.clinical.data.pr$sdev^2 / sum(patients.clinical.data.pr$sdev^2))
plot(cumpro[0:15], xlab = "PC #", ylab = "Amount explained variance [0-1]", main = "Cumulative variance")
abline(v = 6, col="blue", lty=5)
abline(h = 0.8243, col="blue", lty=5)
legend("bottomright", legend=c("PC6 equals 82% variance"),col=c("blue"), lty=5, cex=0.6)
```



A PCA using the clinical data of the samples used in the proteomics study (20 subjects) is done. As displayed, the four groups, with the 24 variables in use, highly overlap, specially the PO group (green). HT is the only group that appears separated from the other three.

```
fviz_pca_biplot(patients.clinical.data.pr, pointshape = 21,
             fill.ind = clinical.data.df$Phen, col.ind = clinical.data.df$Phen, col.var = "grey",
             palette = "Spectral", addEllipses = TRUE, pointsize = 2, ellipse.level=0.55,
             repel = TRUE,legend.title = "Diagnosis") +
  ggtitle("PCOS and obesity") + theme(plot.title = element_text(hjust = 0.5))
```



Logistic regression

In order to reduce the number of variables used in the previous PCA plot (twenty-four), a binary logistic regression is going to be used here. The aim is finding a sub-set of variables that improve the grouping of subjects according to their phenotype. For this, five step-wise logistic regressions are performed, using the "step" function, that selects models to minimize AIC. No correlation study is performed with the variables used in the models. Each logistic regression takes into account one of the five different scenarios:

- Logistic reg: HT. HT is considered one group (0), and HO ,PT and PO a different single group (1)
- Logistic reg: HO. HO is considered one group (0), and HT ,PT and PO a different single group (1)
- Logistic reg: PT. PT is considered one group (0), and HT ,HO and PO a different single group (1)
- Logistic reg: PO. PO is considered one group (0), and HT ,PT and HO a different single group (1)
- Logistic reg: H vs D. Healthy subjects (HT and HO) are considered one group (0), and diseased (PT and PO) another group (1)

For each of the logistic regressions, a set of variables will be obtained as significants, and as suggested by the function "step", will be directly used (without checking correlations or biological significance) to build a definitive model. This model will be checked using ANOVA and a ROC curve with the predicted vs. real results.

The complete list of clinical data over 162 subjects is used here. The subjects not included in the proteomic study are ging to be used as TRAIN data (142), and the 20 used in proteomics analysis, will be the TEST data.

```
complete.clinical.data <- read.csv("162patients.clinical.csv", header = T)
complete.clinical.data.df<-complete.clinical.data[,c(5:33)]
complete.clinical.data.df$Phen <-complete.clinical.data$Phen
complete.clinical.data.df$Patient <-complete.clinical.data$Patient
```

Binary Logistic regression: HT

Obtaining the variables that better differentiate between HT and the rest of the groups is the purpose of the logistic regression and the subsequent ANOVA analysis. Clinical data from the big cohort (162 subjects) is used. The phenotypic group is changed from HT to 0, and the other three (HO, PT and PO) to 1. Once the data is organized, a null (empty model, the starting point of the step function) and full (using all variables) models are obtained. Then, the step function minimizes the AIC (Akaike information criterion) combining the different variables fed to the algorithm. A final model is produced as a linear combination of a subset of the original variables.

```
data.used.HT<-as.matrix(complete.clinical.data.df)
data.used.HT[data.used.HT=="HT"]<-0
data.used.HT[data.used.HT=="PT"]<-1
data.used.HT[data.used.HT=="PO"]<-1
data.used.HT[data.used.HT=="HO"]<-1
data.used.HT<-as.data.frame(data.used.HT)
rownames(data.used.HT)<-NULL
data.used.HT[] <- lapply(data.used.HT, function(x) {
    if(is.factor(x)) as.numeric(as.character(x)) else x
})
train.HT <- data.used.HT[!data.used.HT$Patient %in% clinical.data.df$Patient,]
test.HT <- data.used.HT[data.used.HT$Patient %in% clinical.data.df$Patient,]
test.HT.list.patients<-test.HT$Patient
train.HT<-subset(train.HT, select=-c(Patient))
test.HT<-subset(test.HT, select=-c(Patient))

model.null.HT = glm(Phen ~ 1, data=train.HT, family = binomial(link="logit"))
model.full.HT = glm(Phen ~ ., data=train.HT, family = binomial(link="logit")
                    )
step(model.null.HT, scope = list(upper=model.full.HT), direction="both", test="Chisq", data=train.HT)
```

We build the final model using the significant variables. The model in this case is a linear combination of 10 variables (with a positive or negative sign) and an intercept. In this case, ann increase of waist, testosterone, weight, estradiol and hirsutism will favor belonging to any or some of the HO, PO or PT groups, while an incrrease of FM, height, freeT$, LDL and hidroxi will favor belonging to the HT group.

```
model.final.HT = glm(formula = Phen ~ waist + FM + testosterone + weight + height +
    freeT4 + LDL + hidroxi + estradiol + hirsutism, family = binomial(link = "logit"),  data = train.HT)
model.final.HT
```

Table 1: HT ANOVA

Var.	Pr(>Chi)	
waist	3.67e-14	***
FM	4.46e-05	***
testosterone	3.85e-03	**
weight	6.05e-03	**
height	3.58e-03	**
freeT4	2.73e-02	*
LDL	5.62e-04	***
hidroxi	3.44e-02	*
estradiol	1.70e-04	***
hirsutism	2.48e-05	***

```
##
## Call:  glm(formula = Phen ~ waist + FM + testosterone + weight + height +
##     freeT4 + LDL + hidroxi + estradiol + hirsutism, family = binomial(link = "logit"),
##     data = train.HT)
##
## Coefficients:
##  (Intercept)         waist            FM  testosterone        weight
##     4141.087        17.004      -778.323      2972.392        32.472
##       height        freeT4           LDL       hidroxi     estradiol
##    -4260.005       -61.352        -3.136      -750.536        15.582
##    hirsutism
##       17.578
##
## Degrees of Freedom: 141 Total (i.e. Null);  131 Residual
## Null Deviance:       151.5
## Residual Deviance: 1.091e-06     AIC: 22
```

An ANOVA test is performed with the variables selected for the final model for HT. Variables that more accurately contribute to the model will have lower Probability. Here, waist, FM, LDL, estradiol and hirsutism are the best ones, with a P-value lower than 0.001 (three stars).

```
best1<-names(model.final.HT$coefficients)[-1]
best2<-as.numeric(anova(model.final.HT, test="Chisq")$`Pr(>Chi)`[-1])
best.coeff.HT<-data.frame(best1,best2)
best.coeff.HT<-best.coeff.HT[best.coeff.HT$best2<0.001,]
colnames(best.coeff.HT)<-c("Variable","Prob.")

anova.HT<-anova(model.final.HT, test="Chisq")
stars.HT <-with (anova.HT,ifelse(`Pr(>Chi)`< 0.001,"***",
      ifelse(`Pr(>Chi)`< 0.01,"**",ifelse(`Pr(>Chi)`< 0.05,"*",""))))
anova.HT.df<-data.frame(rownames(anova.HT),
      as.character(formatC(anova.HT[[5]],format = "e", digits = 2)),stars.HT)
colnames(anova.HT.df)<-c("Var.","Pr(>Chi)","")
rownames(anova.HT.df)<-NULL
anova.HT.df<-anova.HT.df[-1,]
kable(anova.HT.df,row.names = FALSE,digits = 3,label = "",caption = "HT ANOVA") %>%
  kable_styling(full_width = F,
  bootstrap_options = c("striped", "condensed"), font_size = 9)
```

A ROC curve is drawn comparing the predicted and real values. In this case, a 97% of the aurea under the curve is obtained. Inspecting the table test.HT, one prediction mistake is found: Patient 1320, actually PT, is classified as HT.

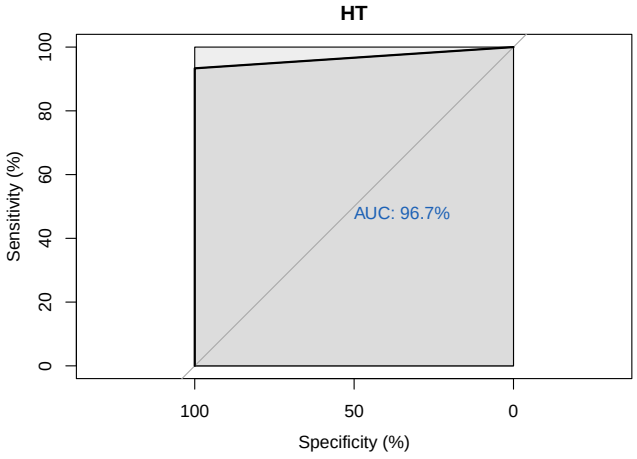
```
fitted.results.HT <- predict(model.final.HT,test.HT,type="response")
test.HT$predy<-predict(model.final.HT,test.HT,type="response")
test.HT$patient<-test.HT.list.patients
roc<-roc(test.HT$Phen, as.numeric(test.HT$predy), percent = TRUE, print.auc=TRUE, print.auc.col = "#1c61b6",
               auc.polygon = TRUE, max.auc.polygon = TRUE, main = "HT", plot=TRUE)
```

Table 2: HO ANOVA

Var.	Pr(>Chi)	
hip	1.49e-14	***
FM	3.31e-11	***
hirsutism	1.46e-03	**
weight	2.44e-02	*
insulin	7.44e-03	**
height	3.35e-02	*
menarche	7.18e-02	
estradiol	8.06e-02	
glucose	5.21e-02	
HDL	8.85e-02	
waist.hip	6.28e-02	
waist	8.22e-02	
LH	1.23e-01	
homaindex	3.79e-02	*
androstenedione	1.86e-06	***

Table 3: PT ANOVA

Var.	Pr(>Chi)	
weight	7.45e-10	***
FM	1.32e-07	***
thyrotropin	1.21e-03	**
HDL	6.50e-03	**
hip	1.38e-02	*
LH.FSF	6.81e-02	
prolactina	3.52e-02	*
hirsutism	3.33e-02	*



Binary Logistic regression: HO, PT, PO and H vs PCOS

Same process than in "Binary Logistic regression: HT" is performed here. For HO, PO and H vs PCOS, a 100% AUC is obtained in the ROC curves. Only for PT , a 91% of AUC is obtained. Inspecting test.PT we realize that two patients are predicted as inter-class: PO-1315 (0.3) PT-1320 (0.5), and one is wrongly classified: PT-1307, classified as no-PT when she actually is PT. Variables and P-value for these comparisons are shown at "Table 2: HO ANOVA", "Table 3: PT ANOVA", "Table 4: PO ANOVA" and "Table 5: H vs PCOS ANOVA".

PCA using variables with P<0.001 for each comparison

A new PCA is performed using only the variables showing a Pval<0.001 at the ANOVA test on any of the five models obtained. The list of variables is composed by: hip, FM, androstenedione, waist, LDL, estradiol, hirsutism, waist.hip, height, weight and

Table 4: PO ANOVA

Var.	Pr(>Chi)	
waist.hip	2.46e-09	***
FM	2.67e-07	***
hip	6.22e-04	***
hirsutism	4.68e-05	***
height	8.95e-04	***
LH.FSF	3.04e-02	*
freeT4	3.54e-02	*
thyrotropin	5.51e-02	
ast	1.08e-01	
cholesterol	1.11e-01	
glucose	1.31e-01	

Table 5: H vs PCOS ANOVA

Var.	Pr(>Chi)	
FM	8.41e-21	***
hirsutism	3.98e-07	***
LH	3.45e-05	***
insulin	1.28e-02	*
height	2.68e-02	*
menarche	2.16e-02	*
homaindex	5.50e-02	
estradiol	1.56e-01	

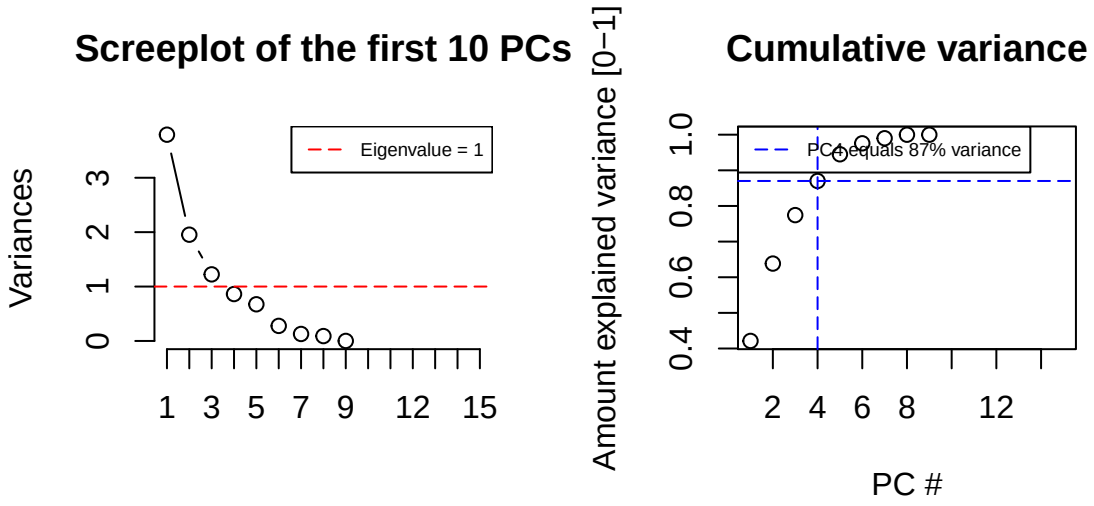
LH. Interestingly, if one of "hip", "waist" or "waist.hip" (a priori redundant) is removed, the PCS significantly worsens. So, the eleven variables remain.

```
patients.clinical.data.filtered<-clinical.data.df
#three.stars<-c(three.stars.HO,three.stars.HT,three.stars.PO,three.stars.PT)
three.stars<-c(as.character(best.coeff.HO$Variable), as.character(best.coeff.HT$Variable),
               as.character(best.coeff.PO$Variable), as.character(best.coeff.PT$Variable),
               as.character(best.coeff.HvsPcos$Variable))
three.stars<-unique(three.stars)
three.stars<-c("Phen","Patient",three.stars)
patients.clinical.data.filtered<-patients.clinical.data.filtered[,
                              colnames(patients.clinical.data.filtered) %in% three.stars]
rownames(patients.clinical.data.filtered)<-patients.clinical.data.filtered$Patient
numeric.data.filtered<-as.data.frame(patients.clinical.data.filtered[c(3:11)])
```

```
patients.clinical.data.pr.filtered <- prcomp(numeric.data.filtered, center = TRUE, scale = TRUE)
summary(patients.clinical.data.pr.filtered)
```

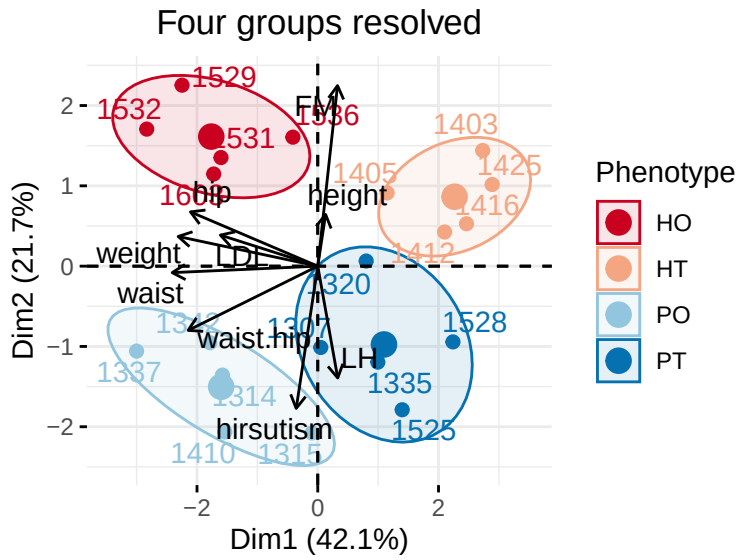
Only the three first principal components appear to be useful and PC4 equals 87% variance

```
screeplot(patients.clinical.data.pr.filtered, type = "l", npcs = 15, main = "Screeplot of the first 10 PCs")
abline(h = 1, col="red", lty=5)
legend("topright", legend=c("Eigenvalue = 1"), col=c("red"), lty=5, cex=0.6)

cumpro.filtered <- cumsum(patients.clinical.data.pr.filtered$sdev^2
                     / sum(patients.clinical.data.pr.filtered$sdev^2))
plot(cumpro.filtered[0:15], xlab = "PC #", ylab = "Amount explained variance [0-1]",
     main = "Cumulative variance")
abline(v = 4, col="blue", lty=5)
abline(h = 0.87013, col="blue", lty=5)
legend("topleft", legend=c("PC4 equals 87% variance"), col=c("blue"), lty=5, cex=0.6)
```



The biplot obtained from the PCA shows the groups of five subjects clearly separated. The second component, the vertical axis, with healthy subjects on top and diseased below, has FM, LH and hirsutism as the predominant variables.

```
library("factoextra")
fviz_pca_biplot(patients.clinical.data.pr.filtered, pointshape = 21,
             pointsize = 2, fill.ind = patients.clinical.data.filtered$Phen,
             col.ind = patients.clinical.data.filtered$Phen, palette = "RdBu", addEllipses = TRUE,
             label = "all", col.var = "black", repel = TRUE, ellipse.level=0.7, legend.title = "Phenotype") +
  ggtitle("Four groups resolved") + theme(plot.title = element_text(hjust = 0.5))
```



For inspecting the first three components of the PCA, gathering 77% of variablity, a library called “pca3d” is used. Because creates a graphical device is not included here, but the 3D display is included in the main text though.

```
library(pca3d)
pca3d(patients.clinical.data.pr.filtered, group=patients.clinical.data.filtered$Phen,
      radius = 2,fancy = TRUE, col = c("red","red","red","red","red","burlywood1",
          "burlywood1","burlywood1","burlywood1","burlywood1", "lightblue","lightblue",
          "lightblue","lightblue","lightblue","blue","blue","blue","blue","blue"))
```

Appendix 2: Chapter4, OpenSwath workflow

Contents

Appendix 2: Chapter4, OpenSwath workflow

Files conversion and Comet and Xtandem searches

First, a conversion of all wiff raw files (library and Swath files) to mzXML format is performed with Proteowizard, using a Docker container. When all files are converted, searches of files that will build the library are performed with the web interface of TPP (from a TPP Docker container redirecting the TPP web server to 10401 port in the local machine). The search engines used are Comet and XTandem. The search will generate two pepXML files (one from each search engines) for each wiff file. Those pepXML will be merged, generating two files: interact.tandem.pep.xml and interact.comet.pep.xml files.

```
# Conversion of wiff files to mzXML

docker run -it --rm -e WINEDEBUG=-all \
 -v /mnt/data2/swath:/data proteowizard/pwiz-skyline-i-agree-to-the-vendor-licenses wine \
 msconvert --mzXML --filter "peakPicking true 1-" /data/*.wiff
```

Building the spectral library

Then, a Trans-Proteomic pipeline (TPP) Docker image is run in interactive mode, to generate the spectra library. Several programs are going to be used sequentially : xinteract, iProphet, Mayu and spectraST. After spectraST, we will go outside the Docker image, and will use a local installation of OpenMS, executing TargetedFileConverter and OpenSwathAssayGenerator. The spectral library generated (transitionlist_optimized_decoys.pqp) will be used in the next point to quantify the Swath files.

```
# Interactive console to TPP
docker run --memory="20g" -it -v /mnt/data2/swath:/data spctools/tpp /bin/bash

# Searches were performed with the web interface of TPP, producing
#interact.tandem.pep.xml and interact.comet.pep.xml files

### PeptideProphet on Comet and Tandem pep.xml. Then, iProphet joining
#     interact.comet.pep.xml and interact.tandem.pep.xml
#     -OARPdl parameter not used => scoring with Rt (R param) caused problems with comet
#Run PeptideProphet on Comet and XTandem
xinteract -dDECOY_ -OAPdl -Ninteract.tandem.pep.xml \
                  /data/pools/*.pep.xml &>> PCOS_log1.txt
xinteract -dDECOY_ -OAPdl -Ninteract.comet.pep.xml \
                  /data/pools/*.pep.xml &>> PCOS_log2.txt

# Run iProphet
InterProphetParser DECOY=DECOY_ interact.comet.pep.xml interact.tandem.pep.xml \
    iProphet.combined.pep.xml &>> PCOS_log3.txt

# Mayu : 2019-08-06_10.01.14_main_1.07.xlsx with 0.01 and 0.05, we select IP/PPs
#   value corresponding to <5% protein FDR (0.630895)
perl /usr/local/tpp/bin//Mayu.pl -A iProphet.combined.pep.xml \
      -C uniprot.SP.human.apr2019.irt.DECOY.fasta -E DECOY_   \
      -G 0.01 -H 101 -I 0 &>> PCOS_log4.txt

#Combination of search results to a spectral library
# irtkit.txt file used from DIA2018 tutorial
# generation of a spectral library using SpectraS.
# mzXML files need to be avaliable under the location where iProphet file is
# Careful about -c_IRT.. argument! : exactly this spelling
```

```
# The iRT.txt file MUST have exactly the format used
spectrast -cNSpecLib -cICID-QTOF -cf'Protein!~DECOY_' -cP0.630895 \
  -algorithm:retentionTimeInterpretation -c_IRT../data/iRT.txt \
  -c_IRR iProphet.combined.pep.xml

# generate a consensus library by running the following command:
spectrast -cNSpecLib_cons -cICID-QTOF -cAC SpecLib.splib

#generate a SpectraST MRM transition list:
spectrast -cNSpecLib_pqp -cICID-QTOF -cM SpecLib_cons.splib

#### Now outsite TPP !
cd /mnt/data2/pcos.proteomics/analysis.pcos.swath

# Peptide-query parameter library generation and conversion of the SpectraST MRM to TraML
TargetedFileConverter -in SpecLib_pqp.mrm -out transitionlist.TraML &>> PCOS_log8.txt

#Generate target assays
# isolation.windows.txt obtained from one of the wiffs with Skyline
OpenSwathAssayGenerator -in transitionlist.TraML -out transitionlist_optimized.TraML \
-swath_windows_file isolation.windows.txt &>> PCOS_log9.txt

#Append decoy transitions to the spectral library:
OpenSwathDecoyGenerator -in transitionlist_optimized.TraML \
-out transitionlist_optimized_decoys.TraML -method shuffle &>> PCOS_log9.txt

#Convert the library to the pqp format for the further OpenSWATH analysis
TargetedFileConverter -in transitionlist_optimized_decoys.TraML \
  -out transitionlist_optimized_decoys.pqp &>> PCOS_log11.txt

# Finally, to inspect the library, we will convert it back to the tsv format.
TargetedFileConverter -in transitionlist_optimized.TraML \
-out transitionlist_optimized.tsv &>> PCOS_log12.txt
```

Quantification of Swath files

Swath files in mzXML format, are quantified and processed using OpenSwathWorkflow, pyprophet and TRIC (feature_alignment.py). The aligned quantitation data generated (in "tab separated values" format) is called aligned.export.tsv here. This file will be converted to a MSStats compatible format.

```
##### OpenSwath
# hroest_DIA_iRT.TraML file from DIA2018 course
for file in $(ls /mnt/data2/pcos.proteomics/swath/swaths.mzxml/*.mzXML)
do
OpenSwathWorkflow -in ${file} -tr transitionlist_optimized_decoys.pqp \
-tr_irt hroest_DIA_iRT.TraML -batchSize 1000 -min_upper_edge_dist 1 \
-Scoring:stop_report_after_feature 5 -out_osw $(basename ${file}).osw \
-threads 10 &>> PCOS_log13.txt
done

### Merge osw files with PyProphet ans analyze them
pyprophet merge --out=training.osw --subsample_ratio=0.33 pcos*.osw
pyprophet merge --out=merged.osw --subsample_ratio=1 pcos*.osw
pyprophet score --in=training.osw --level=ms2
pyprophet score --in=merged.osw --level=ms2 --apply_weights=training.osw
pyprophet peptide --in=merged.osw --context=run-specific peptide \
    --in=merged.osw --context=experiment-wide \
peptide --in=merged.osw --context=global
pyprophet export --in=merged.osw --out=merged_export.tsv --format=legacy_merged \
    --max_rs_peakgroup_qvalue 0.1 --max_global_peptide_qvalue 0.05 \
    --max_global_protein_qvalue 0.01
pyprophet export --in=merged.osw --format=score_plots

### TRIC software
feature_alignment.py --in merged_export.tsv --out aligned.export.tsv --method LocalMST \
  --realign_method lowess --max_rt_diff 60 --mst:useRTCorrection True \
```

```
  --mst:Stdev_multiplier 3.0 --target_fdr -1 --fdr_cutoff 0.05 \
  --max_fdr_quality -1 --alignment_score 0.05
```

MSstats format conversion

The SWATH2stats Bioconductor library is used to import the quantified data (aligned.export.tsv) into a data structure compatible with MSStats. Several steps are followed using the documentation to filter the data and generate a data frame that can be used by MSStats.

```
library(SWATH2stats)
library(MSstats)

data.openswsath <- read.delim2('aligned.export.tsv',
    dec='.',
    sep='\t',
    header=TRUE,
    stringsAsFactors = FALSE)
#Due to the new format of our data, we need to adjust some column names in order to
#be recognized in the following procedure.
names(data.openswsath)[names(data.openswsath) == "FullUniModPeptideName"] <-
  "FullPeptideName"
names(data.openswsath)[names(data.openswsath) == "aggr_fragment_annotation"] <-
  "aggr_Fragment_Annotation"
names(data.openswsath)[names(data.openswsath) == "aggr_peak_area"] <-
  "aggr_Peak_Area"

dim(data.openswsath)
#Before we apply this filter on the actual data, we reduce the number of columns to the
#ones we really need for the further analysis.
data.openswsath.reduced <- reduce_OpenSWATH_output(data.openswsath)
#View(data.reduced)

#We also want to get rid of iRT peptides that are still in the data, we only needed them
#for RT calibration. Additionally we will filter out all non-proteotypic peptides.

data.openswsath.reduced <- data.openswsath.reduced[grep("iRT",
                                data.openswsath.reduced$ProteinName,invert = TRUE),]
data.openswsath.reduced <- data.openswsath.reduced[grep(";",
                                data.openswsath.reduced$ProteinName, invert = TRUE),]

annotation.file <- read.delim(file = 'analysis.PCOS.openswath.annotation.txt',
                              sep='\t',header=TRUE)
data.openswsath.annotated <- sample_annotation(data.openswsath.reduced, annotation.file)

count_analytes(data.openswsath.annotated)
# We can now filter our data set and to make sure that we just use complete observations.
data.openswsath.filtered <- filter_mscore_condition( data.openswsath.annotated,
                                                     mscore=0.01, n.replica=5)
count_analytes(data.openswsath.filtered)
# To feed the data into MSstats (or mapDIA) we need to split the transition groups
#       into single transitions.
data.openswsath.transition <- disaggregate(data.openswsath.filtered)
# columns are renamed to match the requirements for MSstats
MSstats.openswath.input <- convert4MSstats(data.openswsath.transition)
```

MSstats data processing

The MSStats function “dataProcess” will process the Swath data, assigning transitions to peptides (and protein) quantification. It will also perform data normalization between samples, using the “equalizeMedians” method.

```
#MSstats analysis
data.openswath.processed.normalized <- dataProcess(MSstats.openswath.input,
                              normalization = "equalizeMedians",
                              featureSubset="all",
                              summaryMethod = "TMP", censoredInt = "NA",
                              cutoffCensored = "minFeature", MBimpute = FALSE)
```

MSstats group comparison

The different comparisons to be made are designed here. Eight comparisons are built: "HOvsHT","PTvsHT","POvsHT", "POvsHO", "PTvsHO", "POvsPT","PCOS.vs.HT","PCOS.vs.H". Before those comparisons are calculated, we remove the keratines found in the samples (ProcessedData and RunlevelData levels in the normalized data structure). The result for each comparison will be accessed using the names provided to each of them (e.g. "HOvsHT").

```
# 3.2. Group Comparison
levels(data.openswath.processed.normalized$ProcessedData$GROUP_ORIGINAL)
# We get the order of the different phenotypes: "HO" "HT" "PO" "PT"
# Then, each comparison is built as a matrix. If four groups are used, a 0.5 is used,
#  with a negative sign if the group is used as the reference
comparison1<-matrix(c(1,-1,0,0),nrow=1)
comparison2<-matrix(c(0,-1,0,1),nrow=1)
comparison3<-matrix(c(0,-1,1,0),nrow=1)
comparison4<-matrix(c(-1,0,1,0),nrow=1)
comparison5<-matrix(c(-1,0,0,1),nrow=1)
comparison6<-matrix(c(0,0,1,-1),nrow=1)
comparison7<-matrix(c(0,-1,0.5,0.5),nrow=1)
comparison8<-matrix(c(-0.5,-0.5,0.5,0.5),nrow=1)
comparison <- rbind(comparison1, comparison2, comparison3, comparison4,comparison5,
                    comparison6,comparison7,comparison8)
row.names(comparison)<-c("HOvsHT","PTvsHT","POvsHT", "POvsHO", "PTvsHO",
                         "POvsPT","PCOS.vs.HT","PCOS.vs.H")

#Removal of keratines
keratines<-c("sp|P04264|K2C1_HUMAN","sp|P13645|K1C10_HUMAN",
             "sp|P35908|K22E_HUMAN","sp|P35527|K1C9_HUMAN")
data.openswath.processed.normalized$ProcessedData<-
  data.openswath.processed.normalized$ProcessedData[
    !data.openswath.processed.normalized$ProcessedData$PROTEIN %in% keratines,]
data.openswath.processed.normalized$RunlevelData<-
  data.openswath.processed.normalized$RunlevelData[
    !data.openswath.processed.normalized$RunlevelData$Protein %in% keratines,]

result.GroupComparison.openswath <- groupComparison(
  contrast.matrix = comparison,
  data = data.openswath.processed.normalized)
```

Appendix 3: Chapter5, iprg2015 Reanalysis

Contents

Appendix 3: Chapter5, iprg2015 Re-analysis

The iprg2015 dataset consists in four samples of known composition, each containing 200 ng of tryptic digests from S. cerevisiae cultures. Each sample has been independently spiked with different quantities of six individual protein digests, as shown in the table below (from "ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of differentially abundant proteins in label-free quantitative LC-MS/MS experiments"). Raw data has been downloaded from ProteomeXchange (PXD010981): http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD010981

The four samples generated in the iprg2015 experiment are called here C1, C2, C3 and C4. The "differential expression" has been calculated using C1 as reference in all cases studied here. Although more combinations could be made, we think that this approach is sufficient: the Log 2 fold changes obtained for the spiked proteins range from 5.6 for F protein in C2 with respect to C1 (almost 50 times more concentrated) to -5.02 for protein A in C4 with respect to C1 (more than 30 times less concentrated). In "Table 1: theoretical Log2 FC", the theoretical Log2 Fold change values of the spiked in proteins (using sample C1 as reference) are shown.

In this document, three different quantification pipelines are used (MaxQuant, OpenMS and Proteome Discoverer) and three Bioconductor packages for differential expression analysis used (MSStats, DEqMS and DEP). In the case of OpenMS and Proteome Discoverer pipelines, the DEP approach has not been used, because it is not designed for them and attempts to adapt the data in a convenient way produced aberrant results. The R code used with them and the results obtained are shown here. The chunks of code that are redundant, are hidden for convenience.

1. MaxQuant Analysis

Information generated by MaxQuant will be used for normalization and statistical analysis. Two files from the "combined/txt" directory produced by MaxQuant will be used: "proteinGroups.txt" and "evidence.txt". The number of target and decoy PSM hits is shown.

```
MaxQuant.proteinGroups <- read.table("./data/iprg2015.MaxQuant.proteinGroups.txt",sep = "\t", header = TRUE)
MaxQuant.evidence <- read.table("./data/iprg2015.MaxQuant.evidence.txt", sep = "\t", header = TRUE)
decoy.hits<-nrow(MaxQuant.proteinGroups[MaxQuant.proteinGroups$Reverse=="+",])
target.hits<-nrow(MaxQuant.proteinGroups[!MaxQuant.proteinGroups$Reverse=="+",])
cat(decoy.hits," decoy / ",target.hits," target")
```

				Samples			
	Name	Origin	Molecular Weight	1	2	3	4
A	Ovalbumin	Chicken Egg White	45KD	65	55	15	2
B	Myoglobin	Equine Heart	17KD	55	15	2	65
C	Phosphorylase b	Rabbit Muscle	97KD	15	2	65	55
D	Beta-Galactosidase	Escherichia Coli	116KD	2	65	55	15
E	Bovine Serum Albumin	Bovine Serum	66KD	11	0.6	10	500
F	Carbonic Anhydrase	Bovine Erythrocytes	29KD	10	500	11	0.6

Figure 1: Spiked proteins

Table 1: Spikes, theoretical Log2 FC

		C2 vs C1	C3 vs C1	C4 vs C1
		log2 FC	log2 FC	log2 FC
A	P01012 (OVAL_CHICK)	-0.24	-2.12	-5.02
B	P68082 (MYG_HORSE)	-1.87	-4.78	0.24
C	P00489 (PYGM_RABIT)	-2.91	2.12	1.87
D	P00722 (BGAL_ECOLI)	5.02	4.78	2.91
E	P02769 (ALBU_BOVIN)	-4.20	-0.14	5.51
F	P00921 (CAH2_BOVIN)	5.64	0.14	-4.06

```
## 39  decoy /  3268  target
```

1.1. MaxQuant and MSStats

Data generated by MaxQuant is converted to a data frame using the MSstats function MaxQtoMSstatsFormat. Three arguments are needed: an annotation file with the information of groups to be formated, and two data frames produced by reading the evidence and proteinGroups text files. Function dataProcess is used to produce a list of data frames with all the quantitative information needed br MSstats to generate the differential analysis.

```
MSStats.annot <- read.csv("./data/iprg2015.MSStats.design.csv", header = TRUE)
MaxQtoMSstatsFormat.data <- MaxQtoMSstatsFormat(evidence=MaxQuant.evidence, annotation=MSStats.annot,
                                        proteinGroups=MaxQuant.proteinGroups,
                                        useUniquePeptide=TRUE,
                                        removeMpeptides=TRUE,
                                        fewMeasurements='remove', removeProtein_with1Peptide=TRUE)

MaxQuant.MSstats.processed.quant <- dataProcess(MaxQtoMSstatsFormat.data, logTrans=2,
                                            normalization='equalizeMedians', fillIncompleteRows=TRUE,
                                            featureSubset="all", summaryMethod="TMP",
                                            cutoffCensored="minFeature",
                                            censoredInt="NA",remove50missing=FALSE,MBimpute=TRUE,
                                            maxQuantileforCensored=0.999)
```

The four conditions are compared using condition 1 (Sample 1 in Figure 1) as reference using groupComparison function. A matrix is built with the three comparisons done (“C2vsC1”,“C3vsC1”,“C4vsC1”).

```
comparison1<-matrix(c(-1,1,0,0),nrow=1)
comparison2<-matrix(c(-1,0,1,0),nrow=1)
comparison3<-matrix(c(-1,0,0,1),nrow=1)
comparison <- rbind(comparison1, comparison2, comparison3)
row.names(comparison)<-c("C2vsC1","C3vsC1","C4vsC1")
MaxQuant.MSstats.Comparisons <-groupComparison(contrast.matrix=comparison,
                                           data=MaxQuant.MSstats.processed.quant)
```

Deceptive proteins (the six spike-ins showed in figure 1: P44015, P55752, P44374, P44983, P44683, P55249) are replaced by letters A to F, and the three comparisons are subsetted. Then, proteins having an NA value as pvalue (due to compare one group without values) are removed.

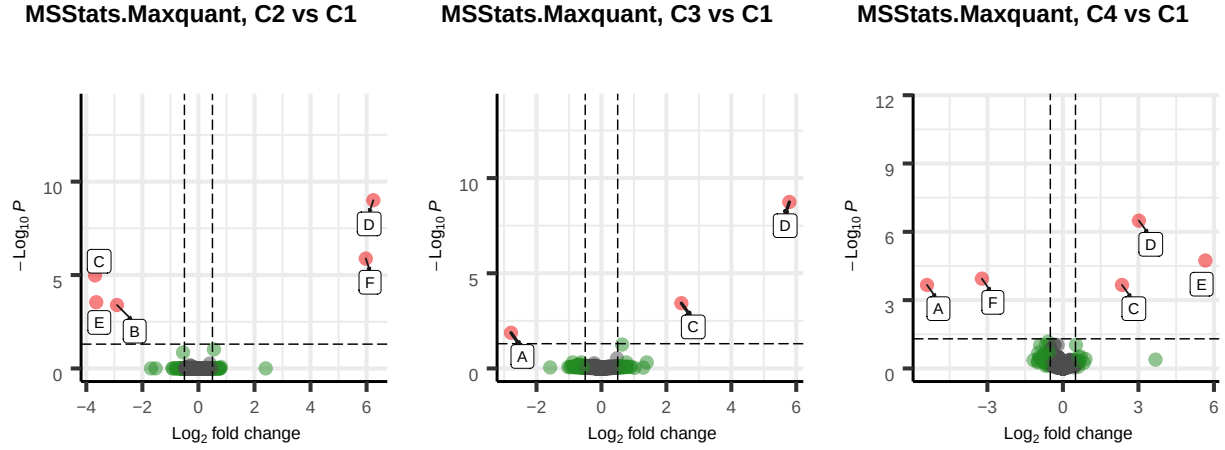
```
MaxQuant.MSstats.results<-MaxQuant.MSstats.Comparisons$ComparisonResult

MaxQuant.MSstats.results$Protein<-as.character(MaxQuant.MSstats.results$Protein)
MaxQuant.MSstats.results$Protein<-str_split_fixed(MaxQuant.MSstats.results$Protein, "\\|", 3)[,2]
MaxQuant.MSstats.results$Protein[MaxQuant.MSstats.results$Protein=='P44015']<-"A"
MaxQuant.MSstats.results$Protein[MaxQuant.MSstats.results$Protein=='P55752']<-"B"
MaxQuant.MSstats.results$Protein[MaxQuant.MSstats.results$Protein=='P44374']<-"C"
MaxQuant.MSstats.results$Protein[MaxQuant.MSstats.results$Protein=='P44983']<-"D"
MaxQuant.MSstats.results$Protein[MaxQuant.MSstats.results$Protein=='P44683']<-"E"
MaxQuant.MSstats.results$Protein[MaxQuant.MSstats.results$Protein=='P55249']<-"F"

MaxQuant.MSstats.results.c2.c1<-subset (MaxQuant.MSstats.results,MaxQuant.MSstats.results$Label=='C2vsC1')
MaxQuant.MSstats.results.c3.c1<-subset (MaxQuant.MSstats.results,MaxQuant.MSstats.results$Label=='C3vsC1')
```

```
MaxQuant.MSstats.results.c4.c1<-subset (MaxQuant.MSstats.results,MaxQuant.MSstats.results$Label=='C4vsC1')

MaxQuant.MSstats.results.c2.c1.clean<-MaxQuant.MSstats.results.c2.c1 %>% drop_na(pvalue)
MaxQuant.MSstats.results.c3.c1.clean<-MaxQuant.MSstats.results.c3.c1 %>% drop_na(pvalue)
MaxQuant.MSstats.results.c4.c1.clean<-MaxQuant.MSstats.results.c4.c1 %>% drop_na(pvalue)

C2.C1.prots<-nrow(MaxQuant.MSstats.results.c2.c1.clean)
C2.C1.prots.removed<- nrow(MaxQuant.MSstats.results.c2.c1)-nrow(MaxQuant.MSstats.results.c2.c1.clean)
C3.C1.prots<-nrow(MaxQuant.MSstats.results.c3.c1.clean)
C3.C1.prots.removed<- nrow(MaxQuant.MSstats.results.c3.c1)-nrow(MaxQuant.MSstats.results.c3.c1.clean)
C4.C1.prots<-nrow(MaxQuant.MSstats.results.c4.c1.clean)
C4.C1.prots.removed<- nrow(MaxQuant.MSstats.results.c4.c1)-nrow(MaxQuant.MSstats.results.c4.c1.clean)

cat (" C2vsC1 comp.: ",C2.C1.prots, "prots  (after ",C2.C1.prots.removed," NA removed)","\n",
"C3vsC1 comp.: ",C3.C1.prots, "prots  (after ",C3.C1.prots.removed," NA removed)","\n",
"C4vsC1 comp.: ",C4.C1.prots, "prots  (after ",C4.C1.prots.removed," NA removed)")
```

```
##  C2vsC1 comp.:  2365 prots  (after  10  NA removed)
##  C3vsC1 comp.:  2360 prots  (after  15  NA removed)
##  C4vsC1 comp.:  2352 prots  (after  23  NA removed)
```

The differential abundance of proteins observed in the different comparisons is compared using volcano plots.

```
pv.cut<-0.05
fc.cut<-0.5
MaxQuant.MSstats.volcano.c2.c1<-EnhancedVolcano(MaxQuant.MSstats.results.c2.c1.clean,
                lab = as.character(MaxQuant.MSstats.results.c2.c1.clean$Protein),title = "", subtitle = "",
                axisLabSize = 10, titleLabSize = 8, caption="",
                captionLabSize = 4,   legendVisible=FALSE,
                transcriptPointSize=3.2,
                x = 'log2FC',y = 'adj.pvalue', pCutoff = pv.cut, FCcutoff = fc.cut,
                drawConnectors=TRUE,boxedlabels=TRUE)
MaxQuant.MSstats.volcano.c3.c1<-EnhancedVolcano(MaxQuant.MSstats.results.c3.c1.clean,widthConnectors = 0.9,
                lab = as.character(MaxQuant.MSstats.results.c3.c1.clean$Protein),title = "", subtitle = "",
                axisLabSize = 10, titleLabSize = 8, caption="", captionLabSize = 4,   legendVisible=FALSE,
                transcriptPointSize=3.2, x = 'log2FC',y = 'adj.pvalue', pCutoff = pv.cut, FCcutoff = fc.cut,
                drawConnectors=TRUE,boxedlabels=TRUE)
MaxQuant.MSstats.volcano.c4.c1<-EnhancedVolcano(MaxQuant.MSstats.results.c4.c1.clean,
                lab = as.character(MaxQuant.MSstats.results.c4.c1.clean$Protein),title = "", subtitle = "",
                axisLabSize = 10, titleLabSize = 8, caption="", captionLabSize = 4,   legendVisible=FALSE,
                transcriptPointSize=3.2, x = 'log2FC',y = 'adj.pvalue', pCutoff = pv.cut, FCcutoff = fc.cut,
                drawConnectors=TRUE,boxedlabels=TRUE)

ggarrange(MaxQuant.MSstats.volcano.c2.c1, MaxQuant.MSstats.volcano.c3.c1, MaxQuant.MSstats.volcano.c4.c1,
        labels = c("MSStats.Maxquant, C2 vs C1","MSStats.Maxquant, C3 vs C1","MSStats.Maxquant, C4 vs C1"),
        ncol = 3, nrow = 1,hjust = -0.1)
```



```
MaxQuant.MSstats.sign.c2.c1<-nrow(rbind(
  subset(MaxQuant.MSstats.results.c2.c1.clean,log2FC< -fc.cut & pvalue<pv.cut),
  subset(MaxQuant.MSstats.results.c2.c1.clean,log2FC>  fc.cut & pvalue<pv.cut)))
MaxQuant.MSstats.sign.c3.c1<-nrow(rbind(
  subset(MaxQuant.MSstats.results.c3.c1.clean,log2FC< -fc.cut & pvalue<pv.cut),
  subset(MaxQuant.MSstats.results.c3.c1.clean,log2FC>  fc.cut & pvalue<pv.cut)))
MaxQuant.MSstats.sign.c4.c1<-nrow(rbind(
  subset(MaxQuant.MSstats.results.c4.c1.clean,log2FC< -fc.cut & pvalue<pv.cut),
  subset(MaxQuant.MSstats.results.c4.c1.clean,log2FC>  fc.cut & pvalue<pv.cut)))

cat ("Number of significant proteins using P-value<0.05 (not adjusted P-value):\n C2vsC1 sign. proteins: ",
     MaxQuant.MSstats.sign.c2.c1," Total proteins: ",nrow(MaxQuant.MSstats.results.c2.c1.clean),"\n",
      "C3vsC1 sign. proteins: ",MaxQuant.MSstats.sign.c3.c1,
      " Total proteins: ",nrow(MaxQuant.MSstats.results.c3.c1.clean),"\n",
     "C4vsC1 sign. proteins: ",MaxQuant.MSstats.sign.c4.c1,
     " Total proteins: ",nrow(MaxQuant.MSstats.results.c4.c1.clean))
```

```
## Number of significant proteins using P-value<0.05 (not adjusted P-value):
##  C2vsC1 sign. proteins:  22  Total proteins:  2365
##  C3vsC1 sign. proteins:  43  Total proteins:  2360
##  C4vsC1 sign. proteins:  66  Total proteins:  2352
```

We extract the fold changes and P-values from the spiked proteins for each condition, to generate a table with Fold Changes and adjusted P-values obtained for each comparison (C2vsC1, C3vsC1 and C4vsC1) using Maxquant pipeline and MSStats: "Table 2: MaxQuant and MSStats quantitation".

```
# We filter only results for spikes.
# C2 vs C1
MaxQuant.MSstats.c2.c1.spiked.results <- subset(MaxQuant.MSstats.results.c2.c1.clean[,
                                      c("Protein","log2FC","adj.pvalue")],
                                            Protein %in% c("A","B","C","D","E","F"))
MaxQuant.MSstats.c2.c1.spiked.results <- MaxQuant.MSstats.c2.c1.spiked.results[
                                      order(MaxQuant.MSstats.c2.c1.spiked.results$Protein),]
# C3 vs C1
MaxQuant.MSstats.c3.c1.spiked.results <- subset(MaxQuant.MSstats.results.c3.c1.clean[,
                                                  c("Protein","log2FC","adj.pvalue")],
                                      Protein %in% c("A","B","C","D","E","F"))
# B is not quantified in C3.C1, but we need to add it,
#     to produce a 6 rows data frame like C2.C1 and C4.C1
MaxQuant.MSstats.c3.c1.spiked.results <- rbind(MaxQuant.MSstats.c3.c1.spiked.results,c("B",0,0))
MaxQuant.MSstats.c3.c1.spiked.results <- MaxQuant.MSstats.c3.c1.spiked.results[
                                  order(MaxQuant.MSstats.c3.c1.spiked.results$Protein),]
# C4 vs C1
MaxQuant.MSstats.c4.c1.spiked.results <- subset(MaxQuant.MSstats.results.c4.c1.clean[,
                                                  c("Protein","log2FC","adj.pvalue")],
                                         Protein %in% c("A","B","C","D","E","F"))
MaxQuant.MSstats.c4.c1.spiked.results <- MaxQuant.MSstats.c4.c1.spiked.results[
                                  order(MaxQuant.MSstats.c4.c1.spiked.results$Protein),]

# And join the four sorted data frames
rownames(MaxQuant.MSstats.c2.c1.spiked.results) <- c()
rownames(MaxQuant.MSstats.c3.c1.spiked.results) <- c()
rownames(MaxQuant.MSstats.c4.c1.spiked.results) <- c()
MaxQuant.MSstats.c2.c1.spiked.results <- MaxQuant.MSstats.c2.c1.spiked.results[,2:3]
MaxQuant.MSstats.c3.c1.spiked.results <- MaxQuant.MSstats.c3.c1.spiked.results[,2:3]
MaxQuant.MSstats.c4.c1.spiked.results <- MaxQuant.MSstats.c4.c1.spiked.results[,2:3]
colnames(MaxQuant.MSstats.c2.c1.spiked.results) <- c("FC.C2vsC1","pval.C2vsC1")
colnames(MaxQuant.MSstats.c3.c1.spiked.results) <- c("FC.C3vsC1","pval.C3vsC1")
colnames(MaxQuant.MSstats.c4.c1.spiked.results) <- c("FC.C4vsC1","pval.C4vsC1")
MaxQuant.MSstats.spiked.results <- cbind(MaxQuant.MSstats.c2.c1.spiked.results,
                                         MaxQuant.MSstats.c3.c1.spiked.results,
                                         MaxQuant.MSstats.c4.c1.spiked.results)
MaxQuant.MSstats.spiked.results <- as.data.frame(sapply( MaxQuant.MSstats.spiked.results, as.numeric ))
rownames(MaxQuant.MSstats.spiked.results) <-c ("A","B","C","D","E","F")
MaxQuant.MSstats.spiked.results$FC.C2vsC1<-formatC(MaxQuant.MSstats.spiked.results$FC.C2vsC1,
MaxQuant.MSstats.spiked.results$FC.C3vsC1<-formatC(MaxQuant.MSstats.spiked.results$FC.C3vsC1,
```

Table 2: MaxQuant and MSStats quantitation

	C2 vs C1		C3 vs C1		C4 vs C1	
	log2 FC	P-value	log2 FC	P-value	log2 FC	P-value
A	-0.53	9.94e-01	-2.79	1.34e-02	-5.40	2.16e-04
B	-2.91	4.03e-04	0.00	0.00e+00	-0.25	5.09e-01
C	-3.69	1.05e-05	2.46	3.74e-04	2.35	2.16e-04
D	6.24	9.88e-10	5.79	1.80e-09	3.02	3.24e-07
E	-3.65	2.85e-04	-0.60	8.21e-01	5.66	1.82e-05
F	5.97	1.31e-06	-0.37	8.41e-01	-3.22	1.14e-04

```
                                                format = "f",digits = 2)
MaxQuant.MSstats.spiked.results$FC.C4vsC1<-formatC(MaxQuant.MSstats.spiked.results$FC.C4vsC1,
                                                format = "f",digits = 2)
MaxQuant.MSstats.spiked.results$pval.C2vsC1<-formatC(MaxQuant.MSstats.spiked.results$pval.C2vsC1,
                                               format = "e",digits = 2)
MaxQuant.MSstats.spiked.results$pval.C3vsC1<-formatC(MaxQuant.MSstats.spiked.results$pval.C3vsC1,
                                               format = "e",digits = 2)
MaxQuant.MSstats.spiked.results$pval.C4vsC1<-formatC(MaxQuant.MSstats.spiked.results$pval.C4vsC1,
                                               format = "e",digits = 2)
kable(MaxQuant.MSstats.spiked.results,row.names = TRUE,digits = 3,label = "",align=rep('c', 6),
      caption = "MaxQuant and MSStats quantitation",
      col.names = c("log2 FC","P-value","log2 FC","P-value","log2 FC","P-value")) %>%
  kable_styling(full_width = F,bootstrap_options = c("striped", "condensed"), font_size = 9) %>%
  row_spec(row = c(1:6),color = "black") %>%
  add_header_above(c(" " = 1, "C2 vs C1" = 2, "C3 vs C1" = 2, "C4 vs C1" = 2))
```

1.2. MaxQuant and DEqMS

DEqMS requires several elements, extracted from the ProteinGroups MaxQuant file export to work properly:

- "LFQ intensity 1A" to "LFQ intensity 4C" values,
- "Razor + unique peptides 1A" to "Razor + unique peptides 4C"

```
# We remove reverse proteins from the list
MaxQuant.proteinGroups.DEqMS<-MaxQuant.proteinGroups[!MaxQuant.proteinGroups$Protein.IDs %like% "REV__", ]
# Only proteins uniquely mapped
MaxQuant.proteinGroups.DEqMS<-MaxQuant.proteinGroups.DEqMS[MaxQuant.proteinGroups.DEqMS$Number.of.proteins == 1, ]
#Only proteins with more than one peptide
MaxQuant.proteinGroups.DEqMS<-MaxQuant.proteinGroups.DEqMS[
                                    !as.numeric(MaxQuant.proteinGroups.DEqMS$Peptide.counts..unique.) <2, ]
# We obtain the Uniprot AC and replace the skiked proteins
MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs<-str_split_fixed(
                                    MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs, "\\|", 3)[,2]
MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs[MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs=='P44015']<-"A"
MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs[MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs=='P55752']<-"B"
MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs[MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs=='P44374']<-"C"
MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs[MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs=='P44983']<-"D"
MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs[MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs=='P44683']<-"E"
MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs[MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs=='P55249']<-"F"
# We extract columns of Label free quantitation intensities from the previously loaded MaxQuant.proteinGroups
# (iprg2015.MaxQuant.proteinGroups.txt). To select the proper columns, we execute colnames
# (MaxQuant.proteinGroups) and choose columns from "LFQ.intensity.1A" to "LFQ.intensity.4C", that correspond
# to columns 80:91. This, of curse,  depends on the number of samples.
MaxQuant.DEqMS.df = MaxQuant.proteinGroups.DEqMS[,80:91]
MaxQuant.DEqMS.df[MaxQuant.DEqMS.df==0] <- NA
# Rownames are added using the "Majority.protein.IDs" column
rownames(MaxQuant.DEqMS.df) = MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs
# Number of NA is counted for each sample group (4 conditions) and columns are created accordingly.
MaxQuant.DEqMS.df$na_count_1 = apply(MaxQuant.DEqMS.df,1,function(x) sum(is.na(x[1:3])))
MaxQuant.DEqMS.df$na_count_2 = apply(MaxQuant.DEqMS.df,1,function(x) sum(is.na(x[4:6])))
MaxQuant.DEqMS.df$na_count_3 = apply(MaxQuant.DEqMS.df,1,function(x) sum(is.na(x[7:9])))
MaxQuant.DEqMS.df$na_count_4 = apply(MaxQuant.DEqMS.df,1,function(x) sum(is.na(x[10:12])))
```

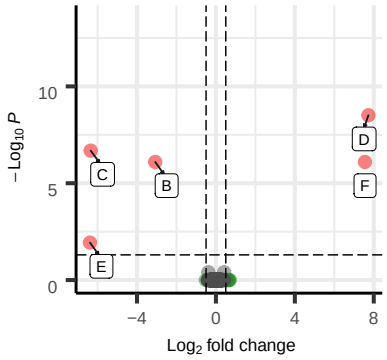
In the same way we did with MSStats, the three comparisons made here are C2 vs C1, C3 vs C1 and C4 vs C1.

```
# Filter protein table. DEqMS requires a minimum of two values for each group. This needs to be done
# separately because NAs will differ between groups.
MaxQuant.DEqMS.df.12<-MaxQuant.DEqMS.df[,c(1:6,13:14)]
MaxQuant.DEqMS.df.13<-MaxQuant.DEqMS.df[,c(1:3,7:9,13,15)]
MaxQuant.DEqMS.df.14<-MaxQuant.DEqMS.df[,c(1:3,10:12,13,16)]
MaxQuant.DEqMS.df.filter.12 = MaxQuant.DEqMS.df.12[MaxQuant.DEqMS.df.12$na_count_1<2 &
                                                   MaxQuant.DEqMS.df.12$na_count_2<2, 1:6]
MaxQuant.DEqMS.df.filter.13 = MaxQuant.DEqMS.df.13[MaxQuant.DEqMS.df.13$na_count_1<2 &
                                                   MaxQuant.DEqMS.df.13$na_count_3<2, 1:6]
MaxQuant.DEqMS.df.filter.14 = MaxQuant.DEqMS.df.14[MaxQuant.DEqMS.df.14$na_count_1<2 &
                                                   MaxQuant.DEqMS.df.14$na_count_4<2, 1:6]
# A data frame of unique peptide minimum count per protein (Unique + Razor) is made for the groups involved
# in the three comparisons: 23:28 for C2 vs C1, 23:25 and 29:31 for C3 vs C1 and 23:25 and 32:34 for C4 vs C1
pep.count.table.12 = data.frame(count = rowMins(as.matrix(MaxQuant.proteinGroups.DEqMS[,23:28])),
                                row.names = MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs)
pep.count.table.13 = data.frame(count = rowMins(as.matrix(MaxQuant.proteinGroups.DEqMS[,c(23:25,29:31)])),
                                row.names = MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs)
pep.count.table.14 = data.frame(count = rowMins(as.matrix(MaxQuant.proteinGroups.DEqMS[,c(23:25,32:34)])),
                                row.names = MaxQuant.proteinGroups.DEqMS$Majority.protein.IDs)
# As the DEqMS software documentation states, a minimum peptide count of some proteins can be 0 and
# adding a pseudocount 1 to all proteins is needed
pep.count.table.12$count = pep.count.table.12$count+1
pep.count.table.13$count = pep.count.table.13$count+1
pep.count.table.14$count = pep.count.table.14$count+1
#Finally, the DEqMS analysis on LFQ data
protein.matrix.12 = log2(as.matrix(MaxQuant.DEqMS.df.filter.12))
protein.matrix.13 = log2(as.matrix(MaxQuant.DEqMS.df.filter.13))
protein.matrix.14 = log2(as.matrix(MaxQuant.DEqMS.df.filter.14))
class.12 = as.factor(c("1","1","1","2","2","2"))
class.13 = as.factor(c("1","1","1","2","2","2"))
class.14 = as.factor(c("1","1","1","2","2","2"))
# Fitting without intercept
design.12 = model.matrix(~0+class.12)
design.13 = model.matrix(~0+class.13)
design.14 = model.matrix(~0+class.14)
fit1.12 = lmFit(protein.matrix.12,design = design.12)
fit1.13 = lmFit(protein.matrix.13,design = design.13)
fit1.14 = lmFit(protein.matrix.14,design = design.14)
# Here we need to check the colnames for the class factors
cont.12 <- makeContrasts(class.122-class.121, levels = design.12) # The reference goes on the right
cont.13 <- makeContrasts(class.132-class.131, levels = design.13)
cont.14 <- makeContrasts(class.142-class.141, levels = design.14)
fit2.12 = contrasts.fit(fit1.12,contrasts = cont.12)
fit2.13 = contrasts.fit(fit1.13,contrasts = cont.13)
fit2.14 = contrasts.fit(fit1.14,contrasts = cont.14)
fit3.12 <- eBayes(fit2.12)
fit3.13 <- eBayes(fit2.13)
fit3.14 <- eBayes(fit2.14)
fit3.12$count = pep.count.table.12[rownames(fit3.12$coefficients),"count"]
fit3.13$count = pep.count.table.13[rownames(fit3.13$coefficients),"count"]
fit3.14$count = pep.count.table.14[rownames(fit3.14$coefficients),"count"]
fit4.12 = spectraCounteBayes(fit3.12)
fit4.13 = spectraCounteBayes(fit3.13)
fit4.14 = spectraCounteBayes(fit3.14)
MaxQuant.DEqMS.results.12 = outputResult(fit4.12,coef_col = 1)
MaxQuant.DEqMS.results.13 = outputResult(fit4.13,coef_col = 1)
MaxQuant.DEqMS.results.14 = outputResult(fit4.14,coef_col = 1)
```

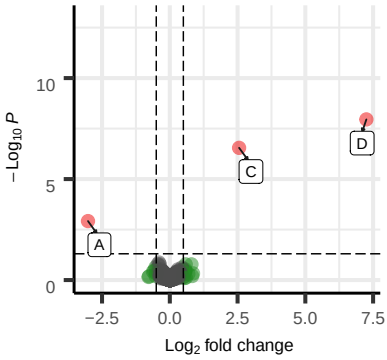
Table 3: MaxQuant and DEqMS quantitation

	C2 vs C1		C3 vs C1		C4 vs C1	
	log2 FC	P-value	log2 FC	P-value	log2 FC	P-value
A	-0.38	3.93e-01	-3.02	1.20e-03	0.00	0.00e+00
B	-3.08	7.98e-07	0.00	0.00e+00	-0.26	8.22e-01
C	-6.35	2.07e-07	2.56	2.84e-07	2.51	2.77e-07
D	7.73	3.09e-09	7.25	1.11e-08	4.73	2.23e-07
E	-6.38	1.17e-02	-0.52	6.40e-01	6.59	1.37e-07
F	7.55	7.98e-07	-0.19	8.78e-01	-3.13	5.71e-03

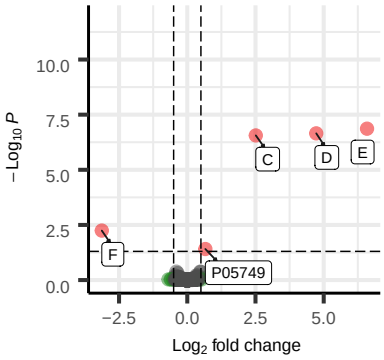
DEqMS.Maxquant, C2 vs C1



DEqMS.Maxquant, C3 vs C1



DEqMS.Maxquant, C4 vs C1




```
MaxQuant.DEqMS.sign.c2.c1<-nrow(rbind(
  subset(MaxQuant.DEqMS.results.12,logFC< -fc.cut & P.Value<pv.cut),
  subset(MaxQuant.DEqMS.results.12,logFC>  fc.cut & P.Value<pv.cut)))
MaxQuant.DEqMS.sign.c3.c1<-nrow(rbind(
  subset(MaxQuant.DEqMS.results.13,logFC< -fc.cut & P.Value<pv.cut),
  subset(MaxQuant.DEqMS.results.13,logFC>  fc.cut & P.Value<pv.cut)))
MaxQuant.DEqMS.sign.c4.c1<-nrow(rbind(
  subset(MaxQuant.DEqMS.results.14,logFC< -fc.cut & P.Value<pv.cut),
  subset(MaxQuant.DEqMS.results.14,logFC>  fc.cut & P.Value<pv.cut)))
cat ("Number of significant proteins using P-value<0.05 (not adjusted P-value):\n C2 vs C1 sign. proteins: ",
     MaxQuant.DEqMS.sign.c2.c1,"  Total proteins: ",nrow(MaxQuant.DEqMS.results.12),"\n",
     "C3 vs C1 sign. proteins: ",MaxQuant.DEqMS.sign.c3.c1,
     " Total proteins: ",nrow(MaxQuant.DEqMS.results.14),"\n",
     "C4 vs C1 sign. proteins: ",MaxQuant.DEqMS.sign.c4.c1,
    "  Total proteins: ",nrow(MaxQuant.DEqMS.results.14))
```

```
## Number of significant proteins using P-value<0.05 (not adjusted P-value):
##  C2 vs C1 sign. proteins:  7   Total proteins:  1966
##  C3 vs C1 sign. proteins:  18  Total proteins:  1911
##  C4 vs C1 sign. proteins:  8   Total proteins:  1911
```

1.3. MaxQuant and DEP

Elements needed by DEP are:

- MaxQuant.proteinGroups: iprg2015.MaxQuant.proteinGroups.txt
- DEP.annot: experimental design for iprg2015. It is a csv file with “label”, “condition” and “replicate” as columns.

```
DEP.annot<-read.table("./data/iprg2015.DEP.design.csv",sep=",",header = TRUE)
DEP.annot$label<-as.character(DEP.annot$label)
# We remove reverse proteins from the list
MaxQuant.proteinGroups.DEP<-MaxQuant.proteinGroups[!MaxQuant.proteinGroups$Protein.IDs %like% "REV__", ]
```

```
# Only proteins uniquely mapped
MaxQuant.proteinGroups.DEP<-MaxQuant.proteinGroups.DEP[MaxQuant.proteinGroups.DEP$Number.of.proteins == 1, ]
#Only proteins with more than one peptide
MaxQuant.proteinGroups.DEP<-MaxQuant.proteinGroups.DEP[!as.numeric(
                                                MaxQuant.proteinGroups.DEP$Peptide.counts..unique.) < 2, ]
MaxQuant.data.DEP<-import_MaxQuant(MaxQuant.proteinGroups.DEP, DEP.annot,
                #filter = c("Reverse","Potential.contaminant"),
                intensities = "LFQ", names = "Protein.IDs",
                ids = "Protein.IDs", delim = ";")
```

Data is filtered for proteins with missing values for at least one condition (thr=0) and then normalized using the function normalize_vsn, where variance stabilizing transformation is performed using the vsn-package.

```
# Filter
MaxQuant.data.DEP_fil <- filter_missval(MaxQuant.data.DEP, thr = 0)
#Normalize
MaxQuant.data.DEP_norm <- normalize_vsn(MaxQuant.data.DEP_fil)
```

Data is imputed following MinProb approach. Many other approaches are avaliable: “bpca”, “knn”, “QRILC”, “MLE”, “MinDet”, “MinProb”, “man”, “min”, “zero”, “mixed” or “nbavg”. Here, missing not at random (MNAR) censored values was assumed.

```
# Imputation not used
MaxQuant.data_imp <- impute(MaxQuant.data.DEP_norm, fun = "MinProb")
# test_diff performs a differential enrichment test based on protein-wise linear models and empirical Bayes
# statistics using limma.
# The control used (Cond1) is defined as the condition introduced in DEP.annot file
MaxQuant.data_diff <- test_diff(MaxQuant.data_imp, type = "control", control = "Cond1")
# add_rejections marks significant proteins based on defined cutoffs.
MaxQuant.dep <- add_rejections(MaxQuant.data_diff, alpha = 0.05, lfc = log2(1.5))
# Generate a results table, where the three possible comparisons (C2 vs C1, C3 vs C1 and C4 vs C1) are done.
MaxQuant.dep.data_results <- get_results(MaxQuant.dep)
MaxQuant.dep.data_results$ID<-str_split_fixed(MaxQuant.dep.data_results$ID, "\\|", 3)[,2]
# Spiked proteins are converted to letters A to F
MaxQuant.dep.data_results$ID[MaxQuant.dep.data_results$ID=='P44015']<-"A"
MaxQuant.dep.data_results$ID[MaxQuant.dep.data_results$ID=='P55752']<-"B"
MaxQuant.dep.data_results$ID[MaxQuant.dep.data_results$ID=='P44374']<-"C"
MaxQuant.dep.data_results$ID[MaxQuant.dep.data_results$ID=='P44983']<-"D"
MaxQuant.dep.data_results$ID[MaxQuant.dep.data_results$ID=='P44683']<-"E"
MaxQuant.dep.data_results$ID[MaxQuant.dep.data_results$ID=='P55249']<-"F"
# Number of significant proteins in the three comparisons. This gives the proteins found significant in at
# least one comparison: here we obtain 98 proteins. For the individual comparisons we obtain 26, 72 and 32
# for C2vsC1, C3vsC1 and C4vsC1.
MaxQuant.dep.number.sign<-as.numeric(MaxQuant.dep.data_results %>% filter(significant) %>% nrow())
```

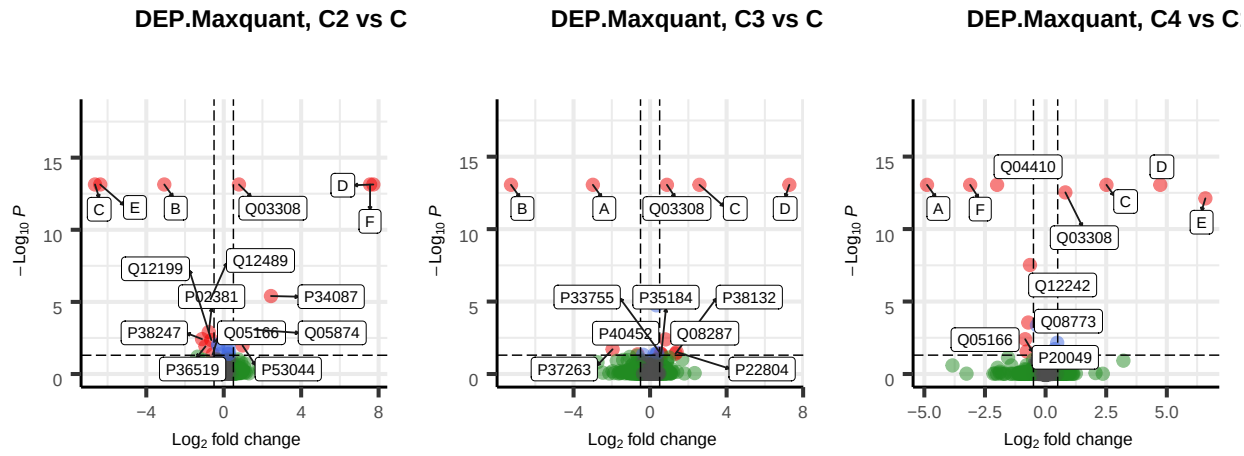


Table 4: MaxQuant and DEP quantitation

	C2 vs C1		C3 vs C1		C4 vs C1	
	log2 FC	P-value	log2 FC	P-value	log2 FC	P-value
A	-0.38	6.62e-01	-2.99	8.71e-14	-4.89	8.89e-14
B	-3.07	7.26e-14	-7.27	8.71e-14	-0.25	7.47e-01
C	-6.65	7.26e-14	2.58	8.71e-14	2.51	8.89e-14
D	7.73	7.26e-14	7.28	8.71e-14	4.73	8.89e-14
E	-6.38	7.26e-14	-0.50	9.47e-01	6.59	7.62e-13
F	7.56	7.26e-14	-0.16	9.47e-01	-3.11	8.89e-14

```
## Number of significant proteins using P-value<0.05 (not adjusted P-value):
##  C2vsC1 sign. proteins:  34    Total proteins:  2019
##  C3vsC1 sign. proteins:  46   Total proteins:  2019
##  C4vsC1 sign. proteins:  41   Total proteins:  2019
```

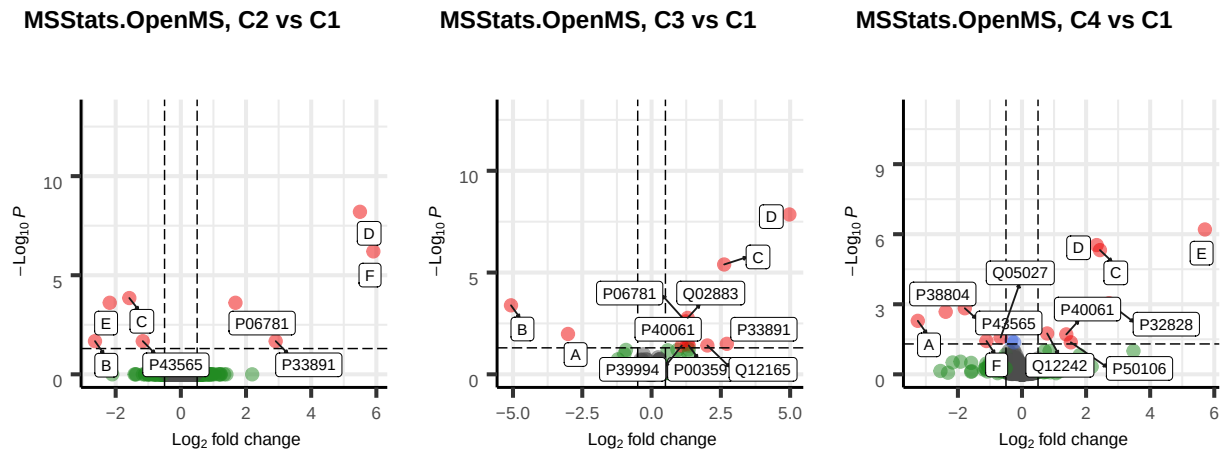
2. OpenMS Analysis

The results obtained with OpenMS are going to be analyzed using MSStats and DEqMS. OpenMS automatically generates a text file that can be directly imported by MSStats. For DEaMS, a Perl script has been made to generate a text file (combining the data generated by the MSStats-ready file) that is compatible with what DEqMS expects. In the case of DEP, writing a different script was attempted to adapt information required by DEP from OpenMS export, but the results were not satisfactory.

2.1. OpenMS and MSStats

OpenMS generates (by using an exporter in the worflow) a file that can be directly used by MSStats (iprg2015.OpenMS.Norm.MSstats.no1pep.csv). With the exception of the file import by MSStats, code is hidden here because is quite simiar to what was done before using MSStats for MaxQuant data.

```
OpenMS.MSStats.Formated.results<-read.csv("./data/iprg2015.OpenMS.Norm.MSstats.no1pep.csv")
OpenMS.MSstats.processed.quant <- dataProcess(OpenMS.MSStats.Formated.results, logTrans=2,
                                normalization='equalizeMedians', fillIncompleteRows=TRUE, featureSubset="all",
                                summaryMethod="TMP", cutoffCensored="minFeature", censoredInt="NA",
                                remove50missing=FALSE, MBimpute=TRUE, maxQuantileforCensored=0.999)
```




```
## Number of significant proteins using P-value<0.05 (not adjusted P-value):
##  C2vsC1 sign. proteins:  26   Total proteins:  2557
##  C3vsC1 sign. proteins:  41   Total proteins:  2558
##  C4vsC1 sign. proteins:  52   Total proteins:  2555
```

Table 5: OpenMS and MSStats quantitation

	C2 vs C1		C3 vs C1		C4 vs C1	
	log2 FC	P-value	log2 FC	P-value	log2 FC	P-value
A	-0.27	9.99e-01	-3.02	1.04e-02	-3.25	5.12e-03
B	-2.63	2.12e-02	-5.07	4.08e-04	-0.23	8.51e-01
C	-1.58	1.42e-04	2.62	4.01e-06	2.43	4.80e-06
D	5.50	6.29e-09	4.97	1.40e-08	2.33	2.92e-06
E	-2.18	2.45e-04	-0.46	5.21e-01	5.71	6.26e-07
F	5.91	6.28e-07	0.10	9.14e-01	-1.11	3.75e-02

2.2. OpenMS and DEqMS

We use the output from OpenMS that was intended for use with MSstats and transform it to the format needed by DEqMS. Is important to note here that the OpenMS output provides unique or proteotypic peptides, not razor peptides.In order to perform that conversion, we have used a Perl script that reorganizes the information and counts the unique peptides found for each condition and protein.

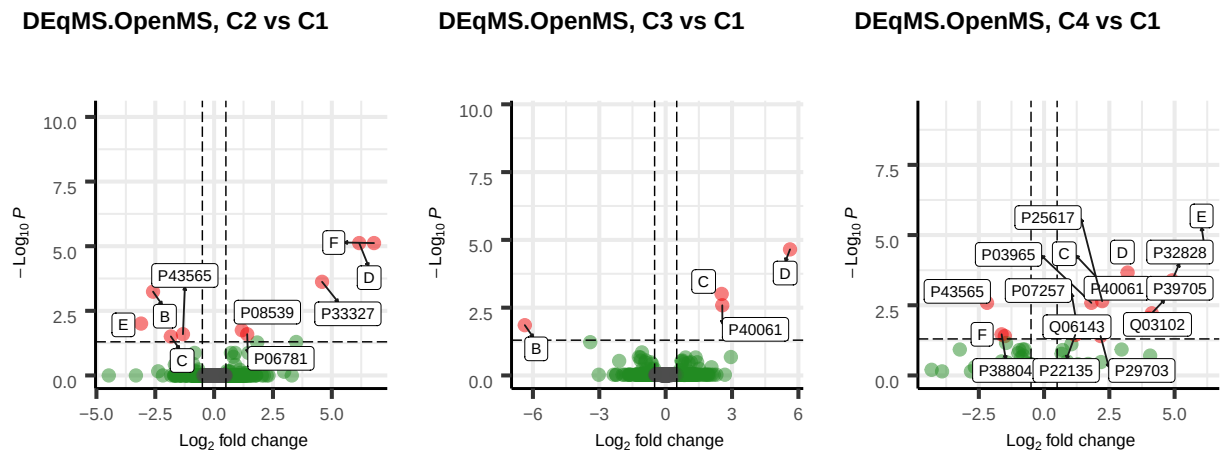
```
##################################      Perl script       ##################################
open my $openMSExportData,"<","./data/iprg2015.OpenMS.Norm.MSstats.no1pep.csv";
my %protein_data;
my $number_of_runs=12;
# We read the information storing it in a hash. The number of runs is introduced manually, although it
# could be inferred directly from the data.
while(<$openMSExportData>){
    next if /^ProteinName/;
    chomp;
    my ($ProteinName,$PeptideSequence,$PrecursorCharge,$FragmentIon,$ProductCharge,$IsotopeLabelType,
      $Condition,$BioReplicate,$Run,$Intensity)= split /\,/,$_;
    if(defined $protein_data{$ProteinName}{$Run}{TotalIntensitiy}){
        $protein_data{$ProteinName}{$Run}{TotalIntensitiy}=
                          $Intensity+$protein_data{$ProteinName}{$Run}{TotalIntensitiy};
        $protein_data{$ProteinName}{$Run}{Peptides}+=1;
        }
    else{
        $protein_data{$ProteinName}{$Run}{TotalIntensitiy}=$Intensity;
        $protein_data{$ProteinName}{$Run}{Peptides}=1;
        }
    }
# Then, we write that information in a tabular format
open OUT,">","./data/iprg2015.OpenMS.DEqMS.txt";
print OUT "Protein\t";
for my $i(1..$number_of_runs){
    print OUT "Run$i\t";
    }
for my $i(1..$number_of_runs){
    print OUT "Unique$i\t";
    }
print OUT "\n";
foreach my $prot(sort keys %protein_data){
    print OUT $prot,"\t";
    for my $i(1..$number_of_runs){
        defined $protein_data{$prot}{$i}{TotalIntensitiy} ?
              printf OUT ("%d",$protein_data{$prot}{$i}{TotalIntensitiy}) : print OUT 0;
        print  OUT "\t";
        }
    for my $i(1..$number_of_runs){
     defined $protein_data{$prot}{$i}{Peptides} ? print OUT $protein_data{$prot}{$i}{Peptides} : print OUT 0;
     print  OUT "\t";
    }
    print OUT "\n";
    }
##################################      end of Perl script       ##################################
```

The Perl script generates a file named “iprg2015.OpenMS.DEqMS.txt” that is imported by DEqMS.

Table 6: OpenMS and DEqMS quantitation

	C2 vs C1		C3 vs C1		C4 vs C1	
	log2 FC	P-value	log2 FC	P-value	log2 FC	P-value
A	0.10	1.00e+00	-3.41	5.93e-02	-4.31	6.29e-01
B	-2.60	5.73e-04	-6.37	1.40e-02	-0.06	8.95e-01
C	-1.84	3.10e-02	2.53	9.85e-04	2.50	9.63e-04
D	6.16	7.57e-06	5.62	2.24e-05	3.20	2.16e-04
E	-3.10	9.85e-03	-0.31	9.41e-01	6.19	4.80e-05
F	6.79	7.57e-06	-0.45	9.41e-01	-1.50	3.94e-02

```
OpenMS.DEqMS.table<-read.table("./data/iprg2015.OpenMS.DEqMS.txt",header = TRUE)
# We extract columns of Label free quantitation intensities from the tabular data file we have created
# (iprg2015.OpenMS.DEqMS.txt).
# Then, we select the proper columns, organized by the previous Perl script in a convenient way.
OpenMS.DEqMS.df = OpenMS.DEqMS.table[,2:13]
OpenMS.DEqMS.df[OpenMS.DEqMS.df==0] <- NA
# Rownames are added using the "Majority.protein.IDs" column
rownames(OpenMS.DEqMS.df) = OpenMS.DEqMS.table$Protein
# Number of NA is counted for each sample group (4 conditions) and columns are created accordingly.
OpenMS.DEqMS.df$na_count_1 = apply(OpenMS.DEqMS.df,1,function(x) sum(is.na(x[1:3])))
OpenMS.DEqMS.df$na_count_2 = apply(OpenMS.DEqMS.df,1,function(x) sum(is.na(x[4:6])))
OpenMS.DEqMS.df$na_count_3 = apply(OpenMS.DEqMS.df,1,function(x) sum(is.na(x[7:9])))
OpenMS.DEqMS.df$na_count_4 = apply(OpenMS.DEqMS.df,1,function(x) sum(is.na(x[10:12])))
```




```
## Number of significant proteins using P-value<0.05 (not adjusted P-value):
##  C2 vs C1 sign. proteins:  75    Total proteins:  2527
##  C3 vs C1 sign. proteins:  82    Total proteins:  2529
##  C4 vs C1 sign. proteins:  116   Total proteins:  2521
```

We extract the fold changes and P-values from the spiked proteins for each condition.

3. Proteome Discoverer Analysis

Data generated by Proteome Discoverer is analyzed here. Both MSStats and DEqMS are used, the later using an strategy equivalent the one used with OpenMS: a Perl script generated a text file that can be properly used by DEqMS using the data generated by Proteome Discoverer.

3.1. Proteome Discoverer and MSStats

Using the PSMs file exported from Proteome Discoverer (iprg2015.ProteomeDiscoverer_PSMs.txt), data is imported by MSStats. The rest of the workflow is equivalent to previous MSStats data analysis, and therefore, hidden.

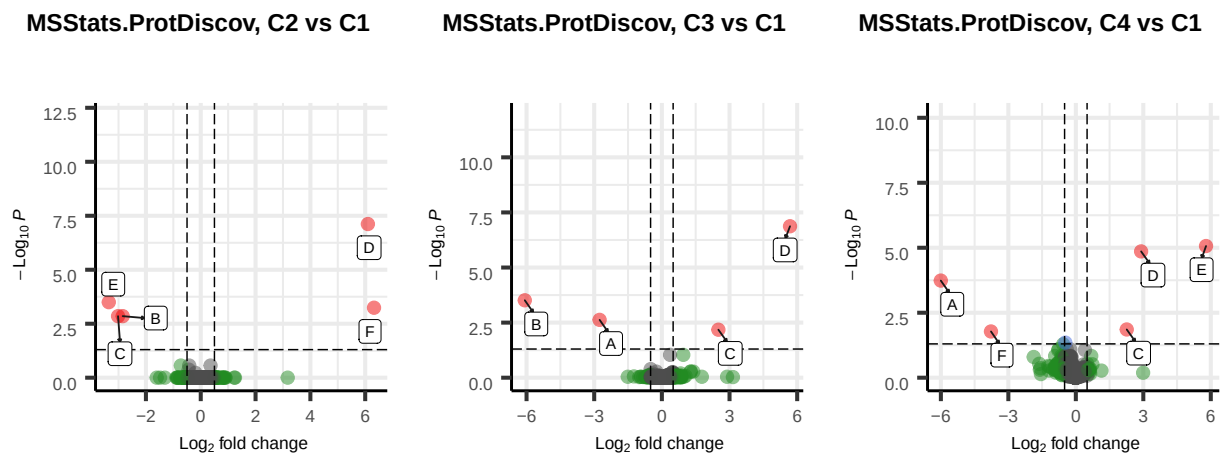
Table 7: ProtDiscov and MSStats quantitation

	C2 vs C1		C3 vs C1		C4 vs C1	
	log2 FC	P-value	log2 FC	P-value	log2 FC	P-value
A	-0.39	9.94e-01	-2.76	2.37e-03	-6.00	1.83e-04
B	-2.85	1.41e-03	-6.08	3.07e-04	-0.14	7.30e-01
C	-3.02	1.41e-03	2.51	6.63e-03	2.26	1.39e-02
D	6.10	7.64e-08	5.68	1.34e-07	2.90	1.38e-05
E	-3.35	3.17e-04	-0.60	8.55e-01	5.79	8.59e-06
F	6.32	5.74e-04	-0.63	9.36e-01	-3.77	1.66e-02

```
ProteomeDiscoverer.annot <- read.csv("./data/iprg2015.MSStats.ProteomeDiscoverer.design.csv", header = TRUE)
ProteomeDiscoverer.psms <- read.table("./data/iprg2015.ProteomeDiscoverer_PSMs.txt", sep = "\t", header = TRUE)

ProteomeDiscoverertoMSstatsFormat.data <- PDtoMSstatsFormat(ProteomeDiscoverer.psms,
                        annotation=ProteomeDiscoverer.annot,which.quantification ="Precursor.Abundance",
                        which.proteinid = 'Protein.Accessions',which.sequence = 'Annotated.Sequence',
                        useUniquePeptide=TRUE,fewMeasurements=TRUE,
                        removeProtein_with1Peptide=TRUE)
```

The four conditions are compared using condition 1 (Sample 1 in Figure 1) as reference using groupComparison function. A matrix is built with the three comparisons done ("C2−C1","C3−C1","C4−C1").

```
##  C2vsC1 comp.:  2854 prots  (after  115  NA removed)
##  C3vsC1 comp.:  2852 prots  (after  117  NA removed)
##  C4vsC1 comp.:  2824 prots  (after  145  NA removed)
```




```
## Number of significant proteins using P-value<0.05 (not adjusted P-value):
##  C2vsC1 sign. proteins:  24   Total proteins:  2854
##  C3vsC1 sign. proteins:  34   Total proteins:  2852
##  C4vsC1 sign. proteins:  66   Total proteins:  2824
```

3.2. Proteome Discoverer and DEqMS

With DEqMS and Proteome Discoverer, a Perl script has been used, in the same way that in the previous section with Open MS.

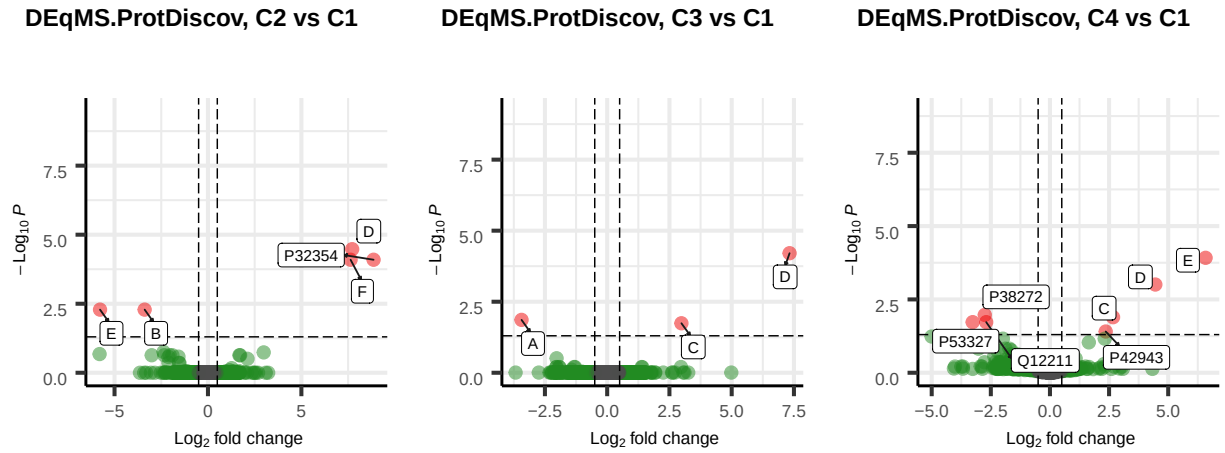
```
#################################      Perl script      ###################################
my %protein_data;
my $number_of_runs=12;
my @runs=qw /F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12/;
open IN, "<","./data/iprg2015.ProteomeDiscoverer_PSMs.txt";
```

```
while(<IN>){
    next if /^\"Checked/;
    chomp;
    s/\"//g;
    my $Annotated_Sequence=(split /\t/)[4];
    my $num_proteins=(split /\t/)[7];
    next if $num_proteins>1;
    my $protein=(split /\t/)[9];
    my $charge=(split /\t/)[11];
    my $peptide=$Annotated_Sequence."_".$charge;
    my $group=(split /\t/)[28];
    my $abundance=(split /\t/)[34];

    if(defined $protein_data{$protein}{$group}{TotalIntensitiy}){
        next unless $abundance;
        $protein_data{$protein}{$group}{TotalIntensitiy}=
                    $abundance+$protein_data{$protein}{$group}{TotalIntensitiy};
        $protein_data{$protein}{$group}{Peptides}+=1;
        }
    else{
        next unless $abundance;
        $protein_data{$protein}{$group}{TotalIntensitiy}=$abundance;
        $protein_data{$protein}{$group}{Peptides}=1;
        }
    }
open OUT,">","./data/iprg2015.ProteinDiscoverer.DEqMS.txt";
print OUT "Protein\t";
print OUT "$_\t" foreach @runs;
print OUT "Unique $_\t" foreach @runs;
print OUT "\n";
foreach my $prot(sort keys %protein_data){
    print OUT $prot,"\t";
    foreach my $i(@runs){
        defined $protein_data{$prot}{$i}{TotalIntensitiy} ?
                  printf OUT ("%d",$protein_data{$prot}{$i}{TotalIntensitiy}) : print OUT 0;
        print  OUT "\t";
        }
    foreach my $i(@runs){
        defined $protein_data{$prot}{$i}{Peptides} ? print OUT $protein_data{$prot}{$i}{Peptides} : print OUT 0;
        print  OUT "\t";
    }
    print OUT "\n";
    }
##################################      end of Perl script       ##################################
```

```
ProteomeDiscoverer.DEqMS.table<-read.table(
                                "./data/iprg2015.ProteinDiscoverer.DEqMS.txt",header = TRUE,sep = "\t")
# We extract columns of Label free quantitation intensities from the tabular data file
# we have created (iprg2015.OpenMS.DEqMS.txt).

# Then, we select the proper columns, organized by the previous Perl script in a convenient way.
ProteomeDiscoverer.DEqMS.df = ProteomeDiscoverer.DEqMS.table[,2:13]
ProteomeDiscoverer.DEqMS.df[ProteomeDiscoverer.DEqMS.df==0] <- NA
# Rownames are added using the "Majority.protein.IDs" column
rownames(ProteomeDiscoverer.DEqMS.df) = ProteomeDiscoverer.DEqMS.table$Protein
# Number of NA is counted for each sample group (4 conditions) and columns are created accordingly.
ProteomeDiscoverer.DEqMS.df$na_count_1 = apply(ProteomeDiscoverer.DEqMS.df,1,function(x) sum(is.na(x[1:3])))
ProteomeDiscoverer.DEqMS.df$na_count_2 = apply(ProteomeDiscoverer.DEqMS.df,1,function(x) sum(is.na(x[4:6])))
ProteomeDiscoverer.DEqMS.df$na_count_3 = apply(ProteomeDiscoverer.DEqMS.df,1,function(x) sum(is.na(x[7:9])))
ProteomeDiscoverer.DEqMS.df$na_count_4 = apply(ProteomeDiscoverer.DEqMS.df,1,function(x) sum(is.na(x[10:12])))
```

Table 8: ProtDiscov and DEqMS quantitation

	C2 vs C1		C3 vs C1		C4 vs C1	
	log2 FC	P-value	log2 FC	P-value	log2 FC	P-value
A	-0.31	9.99e-01	-3.44	1.38e-02	0.00	0.00e+00
B	-3.39	5.17e-03	0.00	0.00e+00	-0.08	9.10e-01
C	-5.80	2.10e-01	2.98	1.81e-02	2.67	1.28e-02
D	7.72	3.35e-05	7.33	6.21e-05	4.46	9.83e-04
E	-5.78	5.17e-03	-0.58	9.85e-01	6.58	1.21e-04
F	7.63	8.07e-05	-0.48	9.85e-01	-5.00	5.81e-02




```
## Number of significant proteins using P-value<0.05 (not adjusted P-value):
##  C2 vs C1 sign. proteins:  97   Total proteins:   2718
##  C3 vs C1 sign. proteins:  103   total proteins:  2715
##  C4 vs C1 sign. proteins:  210   Total proteins:  2666
```

Summary of the thesis in Spanish

La presente tesis describe el análisis bioinformático de las principales técnicas sobre cuantificación "*label-free*" utilizadas en proteómica. En los diferentes capítulos se abordan diversas aproximaciones a la cuantificación de proteínas aplicadas sobre estudios concretos; concretamente:

- Capítulo 1: Incluye una introducción a los aspectos actuales de la cuantificación en proteómica, ofreciendo un visión general del campo y poniendo en contexto el trabajo elaborado en esta tesis.

- Capítulo 2: En este capítulo se ofrece una visión general de los diferentes elementos que componen una infraestructura de análisis de datos cuantitativos, generados en experimentos de proteómica. Se describen elementos de *hardware* y de *software*, explicando cómo interaccionan para conseguir un grado elevado de automatización.

- Capítulo 3: Se centra en la siguiente temática: "Cuantificación mediante marcaje isobárico: repuesta hipóxica temprana en la corteza cerebral". Concretamente en la aproximación realizada en este capítulo, se aborda mediante el uso de un reactivo de marcaje isotópico (TMT), la medición de las variaciones de los niveles de proteínas en la corteza cerebral de ratas, en dos estados de hipoxia con diferente severidad: hipoxia hipobárica (HH) e hipoxia hipobárica con isquemia (HHI). El modelo HHI presenta un perfil global de inhibición, mientras HH presenta un incremento generalizado de los niveles proteicos. Mientras en HH se afecta principalmente el metabolismo oxidativo y energético, en HHI se observa también interferencias en la transmisión sináptica, secreción de neurotransmisores, desarrollo de substancia nigra y activación de la apoptosis mediante la vía mitocondrial.

- Capítulo 4: Incluye una aproximación que se centra en: "Cuantificación SWATH: estudio de marcadores proteicos en plasma". En este caso se utiliza una técnica de Adquisición Independiente de Datos (DIA) llamada SWATH para cuantificar proteínas en plasma de pacientes con Ovario Poliquístico (PCOS) y obesidad con respecto a pacientes sanas. Cinco proteínas (FLNA, ADIPOQ, LBP, RBP4 y APOC2) presentan variaciones significativas en pacientes con PCOS, con RBP4 como el marcador más robusto incluso en pacientes con obesidad. Obesidad y PCOS presentan muchas características en común, con al menos 35 proteínas diferencialmente expresadas y en ambos casos con niveles similares.

- Capítulo 5: Este capítulo trata sobre la "Adquisición Dependiente de Datos y cuantificación sin marcaje: re-análisis del estudio iprg2015 ". Concretamente se usan los datos públicos generados en el estudio iprg2015 para analizar diferentes tipos de software y técnicas de análisis en cuantificación sin marcaje (*label-free*) en proteómica. La combinación de los software MaxQuant y MSStats ha sido escogida como la más conveniente en términos de resultados alcanzados, capacidad de automatización y aplicabilidad a datos obtenidos mediante espectrómetros de masas de diferentes fabricantes. También se han evaluado estrategias de filtrado de resultados, mediante el establecimiento de valores de corte donde un *P-value* de 0.05 y un *ratio* de ±1 han sido seleccionados como valores de referencia para posteriores estudios. Por último, se han estudiado diferentes aproximaciones a la imputación de valores no informados ("*missing values*"). Dicho estudio arroja la conclusión de que el uso del modelo "*Accelerated failure*" representa la aproximación más conveniente, demostrándose además que la combinación de diferentes modelos resulta de gran utilidad.

Publication: Comparative proteomic study of early hypoxic response in the cerebral cortex of rats submitted to two different hypoxic models

DATASET BRIEF

Comparative proteomic study of early hypoxic response in the cerebral cortex of rats submitted to two different hypoxic models

David Ovelleiro[1], Santos Blanco[2]*, Raquel Hernández[2] and María Ángeles Peinado[2]*

[1] Area of Bioinformatics, Instituto Maimónides de Investigación Biomédica de Córdoba, Jaén, Spain
[2] Area of Cellular Biology, Department of Experimental Biology, University of Jaén, Jaén, Spain

Received: April 7, 2017
Revised: June 21, 2017
Accepted: June 30, 2017

Purpose: The present study analyses and compares the cortical brain proteomic profiles of two different models of cerebral hypoxic insult in rats (HH: hypobaric hypoxia and HHI: ischemia followed by hypobaric hypoxia) in an attempt to describe the alterations of the early molecular hypoxic adaptive response underlying each one.
Experimental Design: A quantitative proteomic profile of left-brain cortices of rats under HH, HHI, and control conditions was determined using isobaric labeling (Tandem Mass Tag™) on the protein extracts from pools of five individuals. Data are available via ProteomeXchange with identifier PXD004091.
Results: Altogether, 339 proteins were confidently quantified, 99 of them showing significant variations in the hypoxic conditions with respect to the control. The HHI model presents a global effect of protein downregulation while HH produces an overall increase of the protein levels. While HH mainly affecting oxidative and energetic metabolism, HHI also interferes with synaptic transmission, neurotransmitter secretion, *substantia nigra* development, and triggers apoptosis through mitochondrial pathway.
Conclusions and Clinical Relevance: The findings obtained show an overview of protein alterations under two hypoxic models of different aetiology and provide a basis for more detailed studies in order to unravel new specific mechanisms and therapies for hypoxic pathologies.

Keywords:
Hypoxia / Ischemia / Proteomics / TMT

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

Decline or complete deprivation of oxygen flow to brain and posterior reoxygenation represent a global health issue, as occur after an episode of hypobaric hypoxia or in the cerebral ischemic diseases [1]. Given that decrease or lack of oxygen characterizes all these illnesses, they share several molecular hallmarks: oxidative and nitrosative stresses [2], excitotoxicity [3] or apoptotic and necrotic neuronal death [4]. Nevertheless, the available data point out to specific patterns of these molecular responses depending on the multifactorial aetiology, duration, and severity of the hypoxic insult [5]. Certainly, these variables define and modulate the type of hypoxic adaptive response as well as the hypoxic damage, although the specific molecular pattern underlying each ones is still scarcely known. In the present work, we propose a quantitative analysis and comparison using isobaric labeling (TMT, Tandem Mass Tag™) of the proteomic profiles of two cerebral hypoxic models of different aetiology and scope, both simulating brain hypoxic pathologies: high altitude and cerebral ischemic disease.

Our study has been performed on 15 adult male Wistar rats provided by Harlan Laboratories (Envigo) and weighing

Correspondence: María Ángeles Peinado, Department of Experimental Biology, University of Jaén, Campus Las Lagunillas s/n, 23071 Jaén, Spain
E-mail: apeinado@ujaen.es

Abbreviations: **g.s.d**, global standard deviation; **GO**, Gene Ontology; **HH**, Hypobaric hypoxia; **HHI**, Ischemia followed by hypobaric hypoxia; **TMT**, Tandem Mass Tag™

*These authors have contributed equally to this work.

Colour Online: See the article online to view Figs. 1 and 2 in colour.

350 g each, kept under standard conditions of light and temperature and allowed ad libitum access to food and water, all procedures performed in accordance with the EU animal welfare directive 2010/63/EU. Animals were distributed into three different groups ($n = 5$ per group) depending on the hypoxic model: hypobaric hypoxia (HH), ischemia followed by hypobaric hypoxia (HHI) and a sham control group without any ischemic or hypoxic manipulations. The first one was submitted to a model of hypobaric hypoxia (HH) using a slight modification of a previously published procedure [6] by downregulating the environmental O_2 pressure to a final barometric pressure of approximately 300 hPa inside a hypobaric chamber. The rats were placed in the hypobaric chamber in which the air pressure was controlled by means of a continuous vacuum pump and an adjustable inflow valve. The conditions, simulating an altitude of 9144 m, were maintained for 1 h (temperature and humidity conditions being 23±1°C and 60–70%, respectively). Ascent and descent rates were kept below 300 m/min and the return to normobaric normoxic conditions spanned 30 min. The second group was submitted to a model of cerebral ischemia followed by hypobaric hypoxia (HHI), which consists of unilateral left common carotid artery occlusion followed by a hypoxic stress for a predetermined time, consisting of a slight modification of the Levine/Vannucci model [7]. Animals, recovered for 2 h after surgery, were then exposed to hypobaric hypoxia as previously described. Sham control animals were submitted to surgery without vessel sectioning and then kept in the chamber under normobaric normoxic conditions. In all cases, the survival rate was 100% and animals were anesthetized with ketamine (100 mg/Kg body weight, i.p.) and xylazine (5 mg/Kg body weight, i.p.) and killed after the hypobaric chamber was opened. Body temperature was monitored and maintained throughout all procedures.

The left-brain cortices from animals of all experimental groups were extracted and processed according to the following procedure: 0.1 g of the cortices were homogenized with 1.5 mL of extraction buffer pH 8.0 containing 8 M urea, 20 mM dithiothreitol (DTT), 100 mM Tris–HCl, 0.75 mM phenylmethylsulfonyl fluoride (PMSF), and 4% 3-[(3-cholamidopropyl)-dimethylammonio]-1-propane sulfonate (CHAPS). Proteins were extracted in this buffer for 60 min on ice. Every 15 min, the samples were moderately shaken in a vortex and afterwards were centrifuged at 10 000 × *g* for 15 min at 4°C. The protein concentration of supernatants was measured using the CB-XTM Protein Assay (G-Biosciences, St Louis, USA). Lessening of detergents from protein extraction buffer was carried out using 100 mM triethylammonium bicarbonate (TEAB) by ultrafiltration (millipore 3 k) during 30 min at 12 500 rpm and precipitation (BioRad Protein Sample Cleanup). Isobaric Label Reagent Set (Thermo TMTsixplex™) was performed following the manufacturer's instructions, and followed by desalting (100 mg C18 cartridges, Schalau). Experiments and analysis were performed in blind manner.

The different experimental conditions were then distributed into four different Tandem mass Tag™ (TMTsixplex™) labeled samples, which were analysed using a LTQ Orbitrap mass spectrometer (Thermo Fisher Scientific). Briefly, peptides were analyzed with the Orbitrap mass spectrometer equipped with a nano UHPLC Ultimate 3000 (Dionex-Thermo Scientifics). Chromatography conditions were: mobile phase solution A: 0.1% formic acid in ultrapure water; mobile phase solution B: 80% acetonitrile, 0.1% formic acid, in a C18 nanocapillary column (Acclaim PepMap C18, 75 um internal diameter, 1.8 um particle size, Dionex-Thermo Scientifics) as follow: 5 min, 4% solution B; 240 min, 4–35% solution B; 10 min, 35–80% B; 10 min, 80% B; 10 min 4% B. Nanoelectrospray voltage was set to 1300 V and capillary voltage to 50 V at 190C°. The LTQ Orbitrap was operated in parallel mode, allowing for the accurate measurement of the precursor survey scan (400–1500 *m/z*) in the Orbitrap selection, a 30 000 full-width at half-maximum (FWHM) resolution at *m/z* 400 concurrent with the acquisition of three CID/HCD Data-Dependent MS/MS scans in the LIT and C-Trap for peptide sequence and isotopes quantitation (100–2000 *m/z*), respectively. HCD Resolution set to 7500 FWHM at *m/z* 400. Singly charged ions were excluded. The normalized collision energies were 40% for HCD and 35% for CID. The maximum injection times for MS and MS/MS were set to 50 and 500 ms, respectively. The precursor isolation width was 3 amu and the exclusion mass width was set to 5 ppm. Monoisotopic precursor selection was allowed and singly charged species were excluded. A more extensive description of the experimental procedures and a MIAPE [8] compliant report are found in Supporting Information Methods.

The MS proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [9] partner repository with the dataset identifier PXD004091 (Username: reviewer12476@ebi.ac.uk, Password: tAAJD9QV). Data were analyzed afterwards using Proteome Discoverer (Thermo Fisher Scientific) and searched against a Uniprot Proteome of *Rattus norvegicus* database, containing 27 820 sequences (version 2015.01), resulting in the initial identification of 1409 proteins. Of this initial set of proteins, only 339 were further used in this study as confidently quantified. For a protein to be considered so, it had to present at least two identified peptides with FDR<5%, present quantitative information into the three groups of the analysis (HH, HHI, and control) and its quantitation tags with a coefficient of variation inferior to 30%. The complete list of 339 quantified proteins can be consulted in Supporting Information Data. Of these proteins, only 99 showed differential expression with respect to the control in HH, HHI, or both conditions (Table 1). The variability of a given protein is reported as the amount of positive or negative variation evidence: taking the global standard deviation of the ratios distribution (g.s.d.) as threshold, proteins over/under expressed more than two units of g.s.d. in the same technical replicate, or between 1.5 and two

Table 1. List of 99 rat proteins expressed differentially in HH and HHI with Uniprot accession, gene symbol, description, ratios, and overall evidence of variation: "++" high, "+" moderate, "−−" high, and "−" moderate evidence of over and underexpression, respectively, and " = " for unchanged expression

Protein	Gene	Description	HH data	ΔHH	HHI data	ΔHHI
G3V6Y6	Pygb	α-1,4 glucan phosphorylase	0.63 ± 0.2 0.80 ± 0.2	−	0.70 ± 0.1	−
G3V9G3	Camk2b	Ca/calmodulin-dep. PK II, β, isof.CRA_a	1.02 ± 0.3 0.62 ± 0.2	−	0.69 ± 0.1 0.88 ± 0.1	−
P11275	Camk2a	Ca/calmodulin-dep.PK type II sub.α	0.99 ± 0.2 0.63 ± 0.1	−	0.93 ± 0.1 0.72 ± 0.1	−
P50554	Abat	4-aminobutyrate aminotransf. Mitoch	0.95 ± 0.1 0.63 ± 0.0	−	0.73 ± 0.1	−
F1LRK1	Atp4a	K-transporting ATPase α chain 1	0.57 ± 0.1 0.88 ± 0.1	−	0.59 ± 0.1	−−
F1M779	Cltc	Clathrin heavy chain	0.78 ± 0.2 0.63 ± 0.1	−	0.64 ± 0.1 0.82 ± 0.1	−−
G3V846	Slc1a3	Amino acid transporter	0.62 ± 0.0 0.64 ± 0.1	−	0.93 ± 0.2 0.62 ± 0.0	−−
P06685	Atp1a1	Na/K-transporting ATPase sub. α-1	0.63 ± 0.2	−	0.64 ± 0.1 0.96 ± 0.1	−−
P06686	Atp1a2	Na/K-transporting ATPase sub. α-2	0.62 ± 0.1	−	0.63 ± 0.1 0.96 ± 0.1	−−
P06687	Atp1a3	Na/K-transporting ATPase sub. α-3	0.61 ± 0.1	−	0.94 ± 0.1 0.63 ± 0.1	−−
P32851	Stx1a	Syntaxin-1A	0.96 ± 0.0 0.63 ± 0.1	−	0.95 ± 0.2 0.64 ± 0.1	−−
Q06645	Atp5g1	ATP synthase F(0) complex sub.C1,mitoch	0.53 ± 0.0 0.82 ± 0.0	−	0.60 ± 0.0 0.92 ± 0.1	−−
Q6AXX6	Fam213a	Redox-regulatory protein FAM213A	0.68 ± 0.1 0.64 ± 0.0	−	0.63 ± 0.0	−−
Q6PDU7	Atp5l	ATP synthase sub. g, mitochondrial	0.85 ± 0.1 0.60 ± 0.1	−	1.42 ± 0.1 0.77 ± 0.1	++
D3ZAF6	Atp5j2	ATP synthase sub. f, mitochondrial	0.62 ± 0.1 0.72 ± 0.0	−	0.99 ± 0.0 0.75 ± 0.1	=
F7EYB9	Omg	Protein Omg	0.90 ± 0.0 0.68 ± 0.1	−	0.85 ± 0.1 0.92 ± 0.0	=
P07825	Syp	Synaptophysin	0.61 ± 0.0	−	0.93 ± 0.1 0.74 ± 0.1	=
Q4KLX9	Ccdc163	Protein Ccdc163	0.62 ± 0.1 0.89 ± 0.0	−	0.98 ± 0.0	=
Q63564	Sv2b	Synaptic vesicle glycoprotein 2B	0.60 ± 0.1 0.73 ± 0.2	−	0.81 ± 0.1	=
Q5RKJ9	Rab10	RAB10, member RAS oncogene family	0.54 ± 0.0 0.54 ± 0.0	−−	0.62 ± 0.0 0.95 ± 0.2	−−
P84087	Cplx2	Complexin-2	1.33 ± 0.4 0.50 ± 0.1	−−	1.15 ± 0.3 1.35 ± 0.3	+
B0BNE6	Ndufs8	NADH-DH(Ubiq.)Fe-S prot8 (Pred),isoCRA_a	0.45 ± 0.1 1.22 ± 0.1	−−	1.17 ± 0.2 1.07 ± 0.1	=
P84817	Fis1	Mitochondrial fission 1 protein	0.90 ± 0.2 0.57 ± 0.0	−−	0.92 ± 0.1	=
D3ZH42	Mov10l1	Protein Mov10l1	1.34 ± 0.1 1.05 ± 0.0	+	0.65 ± 0.0	−
P02770	Alb	Serum albumin	0.75 ± 0.1 1.31 ± 0.3	+	0.68 ± 0.1 0.99 ± 0.1	−
D3ZF13	LOC683884	Acyl carrier protein	0.81 ± 0.0 1.40 ± 0.1	+	1.31 ± 0.1	+
P04692	Tpm1	Tropomyosin α-1 chain	1.37 ± 0.2	+	1.28 ± 0.2	+
P26772	Hspe1	10 kDa heat shock protein, mitochondrial	1.46 ± 0.4 1.05 ± 0.3	+	0.95 ± 0.1 1.27 ± 0.3	+
P31399	Atp5h	ATP synthase sub. d, mitochondrial	1.45 ± 0.3	+	1.27 ± 0.3	+

(*Continued*)

Table 1. Continued

Protein	Gene	Description	HH data	ΔHH	HHI data	ΔHHI
P80254	Ddt	D-dopachrome decarboxylase	1.39 ± 0.2	+	1.27 ± 0.1	+
Q03344	Atpif1	ATPase inhibitor, mitochondrial	1.44 ± 0.2	+	1.31 ± 0.2	+
			1.14 ± 0.1		1.08 ± 0.1	
Q6TXF3	Dbi	Acyl-CoA-binding protein	0.97 ± 0.1	+	1.28 ± 0.2	+
			1.41 ± 0.1			
P07171	Calb1	Calbindin	1.39 ± 0.1	+	1.37 ± 0.2	++
			0.85 ± 0.1			
P08082	Cltb	Clathrin light chain B	1.21 ± 0.3	+	1.31 ± 0.1	++
			1.42 ± 0.2		1.27 ± 0.0	
B2RYS2	Uqcrb	Cytochrome b-c1 complex sub. 7	1.44 ± 0.1	+	1.13 ± 0.1	=
			0.85 ± 0.2		1.08 ± 0.2	
D3ZD09	Cox6b1	Cytochrome c oxidase sub. 6B1	1.41 ± 0.3	+	1.23 ± 0.2	=
					1.14 ± 0.2	
D3ZJ08	Hist2h3c2	Histone H3	1.00 ± 0.1	+	1.17 ± 0.0	=
			1.31 ± 0.2		1.22 ± 0.2	
D4A0T0	Ndufb10	Protein Ndufb10	1.45 ± 0.0	+	1.11±0.1	=
D4A678	Spta1	Protein Spta1	0.85 ± 0.1	+	1.01±0.0	=
			1.39 ± 0.1			
D4ACQ2	LOC690384	Protein LOC690384	1.37 ± 0.0	+	1.22±0.1	=
			1.07 ± 0.1			
F1LMR7	Dpp6	Dipeptidyl aminopeptidase-like p6	1.33 ± 0.2	+	0.80±0.0	=
F1M269	NA	Glyceraldehyde-3-phosphate DH Frag.	1.08 ± 0.1	+	1.09±0.1	=
			1.40 ± 0.2			
G3V6D3	Atp5b	ATP synthase sub. β	1.40 ± 0.4	+	1.14 ± 0.2	=
			0.81 ± 0.2		1.12 ± 0.2	
G3V6 × 7	Pcsk1n	Proprot. convertase subtilisin/kexin T1 inhib	1.39 ± 0.4	+	1.11 ± 0.1	=
					1.25 ± 0.3	
G3V8Q2	Ina	α-internexin	1.37 ± 0.3	+	0.96 ± 0.1	=
					1.06 ± 0.2	
O88339	Epn1	Epsin-1	1.44 ± 0.1	+	1.12±0.1	=
P05065	Aldoa	Fructose-bisphosphate aldolase A	0.94 ± 0.1	+	0.89 ± 0.1	=
			1.33 ± 0.4		1.03 ± 0.1	
P10860	Glud1	Glutamate DH 1, mitochondrial	0.93 ± 0.1	+	1.04 ± 0.1	=
			1.39 ± 0.4		0.99 ± 0.1	
P17764	Acat1	Acetyl-CoA acetyltrans. Mitoch.	1.09 ± 0.0	+	1.12±0.0	=
			1.32 ± 0.1			
P23565	Ina	α-internexin	1.37 ± 0.4	+	0.96 ± 0.1	=
					1.06 ± 0.2	
P34926	Map1a	Microtubule-associated protein 1A	1.09 ± 0.2	+	1.08 ± 0.2	=
			1.31 ± 0.3		1.00 ± 0.2	
P35332	Hpcal4	Hippocalcin-like prot. 4	1.40 ± 0.3	+	1.13±0.2	=
			0.95 ± 0.2			
P47819	Gfap	Glial fibrillary acidic protein	1.11 ± 0.2	+	1.15 ± 0.2	=
			1.42 ± 0.4		1.04 ± 0.2	
P48500	Tpi1	Triosephosphate isomerase	1.07 ± 0.1	+	1.06 ± 0.1	=
			1.38 ± 0.3		1.00 ± 0.1	
P54311	Gnb1	Guanine nucl-bind prot G(I)/G(S)/G(T) sub. β-1	0.82 ± 0.1	+	1.04 ± 0.1	=
			1.32 ± 0.1		0.91 ± 0.2	
P54313	Gnb2	Guanine nucl-bind prot G(I)/G(S)/G(T) sub. β-2	0.78 ± 0.1	+	0.91±0.0	=
			1.32 ± 0.1			
P62161	Calm1	Calmodulin	1.44 ± 0.4	+	1.19±0.2	=
P62762	Vsnl1	Visinin-like protein 1	1.39 ± 0.3	+	1.00 ± 0.2	=
			0.88 ± 0.2		1.20 ± 0.1	
P63329	Ppp3ca	Ser/Thr-prot Pase 2B catalytic sub. α isof	0.70 ± 0.1	+	0.81 ± 0.2	=
			1.35 ± 0.3		0.96 ± 0.1	
P85845	Fscn1	Fascin	1.33 ± 0.4	+	1.10 ± 0.2	=
			0.87 ± 0.1		0.99 ± 0.3	

(*Continued*)

Table 1. Continued

Protein	Gene	Description	HH data	ΔHH	HHI data	ΔHHI
Q5XIF3	Ndufs4	NADH DH [ubiquinone] Fe-S prot4, mitoch	1.25 ± 0.2 1.34 ± 0.4	+	0.99±0.3	=
Q8R2H0	Atp6v1g2	ATPase, H+ transporting, V1 sub. G isoform 2	1.32 ± 0.4 1.33 ± 0.2	+	1.19±0.1	=
F1LQ96	Sncg	Gamma-synuclein	1.44 ± 0.1 1.60 ± 0.1	++	1.27±0.1	+
G3V7Y3	Atp5d	ATP synthase sub. delta, mitochondrial	1.63 ± 0.2 1.11 ± 0.1	++	0.97 ± 0.1 1.32 ± 0.1	+
P07936	Gap43	Neuromodulin	1.62 ± 0.3	++	1.28±0.4	+
D3ZH98	NA	Uncharacterized protein	1.55 ± 0.2 1.59 ± 0.5	++	1.38±0.2	++
F1LMW7	Marcks	Myristoylated Ala-rich C-kinase substrate	1.52 ± 0.1 0.88 ± 0.2	++	1.38 ± 0.1	++
Q05175	Basp1	Brain acid soluble protein 1	1.78 ± 0.3	++	1.55 ± 0.2 1.07 ± 0.0	++
D3ZCZ9	LOC100912599	Protein LOC100912599	1.52 ± 0.2	++	0.92 ± 0.0	=
D4AB12	NA	Uncharacterized protein	1.46 ± 0.2 1.08 ± 0.1	++	1.13 ± 0.1	=
P56571	NA	ES1 protein homolog, mitochondrial	1.58 ± 0.4 0.99 ± 0.2	++	1.04 ± 0.0	=
Q3ZB98	Bcas1	Breast carcinoma-ampl.seq1.hom(Frag)	1.71 ± 0.2 1.54 ± 0.4	++	1.10 ± 0.2	=
Q63754	Sncb	β-synuclein	1.54 ± 0.2 1.45 ± 0.3	++	1.21 ± 0.1	=
P01946	Hba1	Hemoglobin sub. α-1/2	0.96 ± 0.1 0.97 ± 0.2	=	0.65 ± 0.2	-
P02688	Mbp	Myelin basic protein	1.09 ± 0.1	=	0.74 ± 0.1 0.81 ± 0.1	-
P04631	S100b	Protein S100-B	1.31 ± 0.4 0.89 ± 0.2	=	1.12 ± 0.1 0.80 ± 0.0	-
P05708	Hk1	Hexokinase-1	0.80 ± 0.1 0.68 ± 0.1	=	0.93 ± 0.1 0.69 ± 0.1	-
P21707	Syt1	Synaptotagmin-1	0.95 ± 0.2 0.70 ± 0.1	=	0.91 ± 0.1 0.70 ± 0.1	-
P27139	Ca2	Carbonic anhydrase 2	0.79 ± 0.1 0.75 ± 0.1	=	0.72 ± 0.1	-
P31596	Slc1a2	Excitatory amino acid transporter 2	0.80 ± 0.2 0.68 ± 0.1	=	1.04 ± 0.3 0.71 ± 0.2	-
P62944	Ap2b1	AP-2 complex sub. β	0.64 ± 0.2 0.94 ± 0.0	=	1.02 ± 0.1 0.68 ± 0.0	-
Q09073	Slc25a5	ADP/ATP translocase 2	0.90 ± 0.2 0.65 ± 0.1	=	0.68 ± 0.1 0.95 ± 0.1	-
Q6P6V0	Gpi	Glucose-6-phosphate isomerase	0.93 ± 0.2 0.70 ± 0.1	=	0.73 ± 0.1 1.02 ± 0.0	-
Q812E9	Gpm6a	Neuronal membrane glycoprotein M6-a	0.73 ± 0.1	=	0.71 ± 0.2 0.91 ± 0.0	-
B0K020	Cisd1	CDGSH Fe-S domain-cont. Prot1	1.02 ± 0.1 0.81 ± 0.2	=	0.84 ± 0.1 0.60 ± 0.1	- -
D3ZNI9	Kcnt1	K channel subfamily T member 1	0.98 ± 0.3 0.91 ± 0.1	=	0.61 ± 0.1	- -
G3V9B3	Mag	Myelin-associated glycoprotein	0.67 ± 0.0	=	0.59 ± 0.0 0.84 ± 0.1	- -
P02091	Hbb	Hemoglobin sub. β-1	0.93 ± 0.1	=	0.61 ± 0.1	- -
P13233	Cnp	2′,3′-cyclic-nucleotide 3′-Pdiesterase	1.19 ± 0.3 0.68 ± 0.1	=	0.60 ± 0.1 0.83 ± 0.2	- -
Q05962	Slc25a4	ADP/ATP translocase 1	0.78 ± 0.2 0.64 ± 0.2	=	0.63 ± 0.1 0.94 ± 0.1	- -
Q62669	NA	Protein Hbb-b1	1.09 ± 0.2 0.97 ± 0.2	=	0.64 ± 0.1	- -

(*Continued*)

Table 1. Continued

Protein	Gene	Description	HH data	ΔHH	HHI data	ΔHHI
Q63345	Mog	Myelin-oligodendrocyte glycoprotein	0.64 ± 0.1	=	0.58 ± 0.1 0.85 ± 0.1	- -
Q8SEZ5	NA	Cytochrome c oxidase sub. 2	0.71 ± 0.0	=	0.61 ± 0.1	- -
R9PY00	Vamp2	Vesicle-assoc membr. prot2 (Frag)	0.94 ± 0.2	=	0.76 ± 0.0	- -
B2GV73	Arpc3	Actin-related protein 2/3 complex sub. 3	0.94 ± 0.0 0.95 ± 0.1	=	0.98 ± 0.0 1.22 ± 0.2	+
P25113	Pgam1	Phosphoglycerate mutase 1	1.01 ± 0.2 1.35 ± 0.2	=	1.30 ± 0.4	+
P47728	Calb2	Calretinin	1.29 ± 0.2 0.93 ± 0.1	=	1.30 ± 0.2 1.01 ± 0.2	+
F1M2D3	Vdac1	Uncharacterized protein	1.23 ± 0.2 1.13 ± 0.3	=	1.28 ± 0.2 1.16 ± 0.2	+ +
Q9Z2L0	Vdac1	Voltage-dep anion-sel. channel prot1	1.19 ± 0.2 1.07 ± 0.3	=	1.28 ± 0.2 1.12 ± 0.2	+ +

g.s.d. in two technical replicates, are considered as highly over/under expressed, while variation between 1.5 and two g.s.d in a single technical replicate is considered as moderate evidence of over/under expression (Supporting Information Analysis).

Using Cytoscape ClueGO plug-in [10], we performed a Gene Ontology (GO) enrichment analysis (two-sided hypergeometric test and Bonferroni step-down correction) of the 99 proteins expressed differentially in HH and/or HHI conditions: 54 of them belong to at least one of the 20 enriched biological processes found. Both hypoxic models present a similar number of differentially expressed proteins (37 and 36, respectively), but with an overall positive expression in HH (22 overexpressed proteins) and negative in HHI (25 underexpressed proteins) (Fig. 1). The similar set of affected processes, both in HH and HHI, points to similar pathologic patterns [1], while the overall inhibitory nature found in HHI is explained by its greater severity in contrast to HH [5].

The 20 biological processes identified were grouped into seven functional groups attending to the similarity of the processes and genes shared (Fig. 2):

(i) ATP metabolic process and (ii) Proton, Hydrogen transmembrane, and Inorganic cation transport, both showing higher enrichment in HH, present a more downregulated state in HHI: potassium import across plasma membrane is severely inhibited (Atp1a1, Atp1a3), while calcium

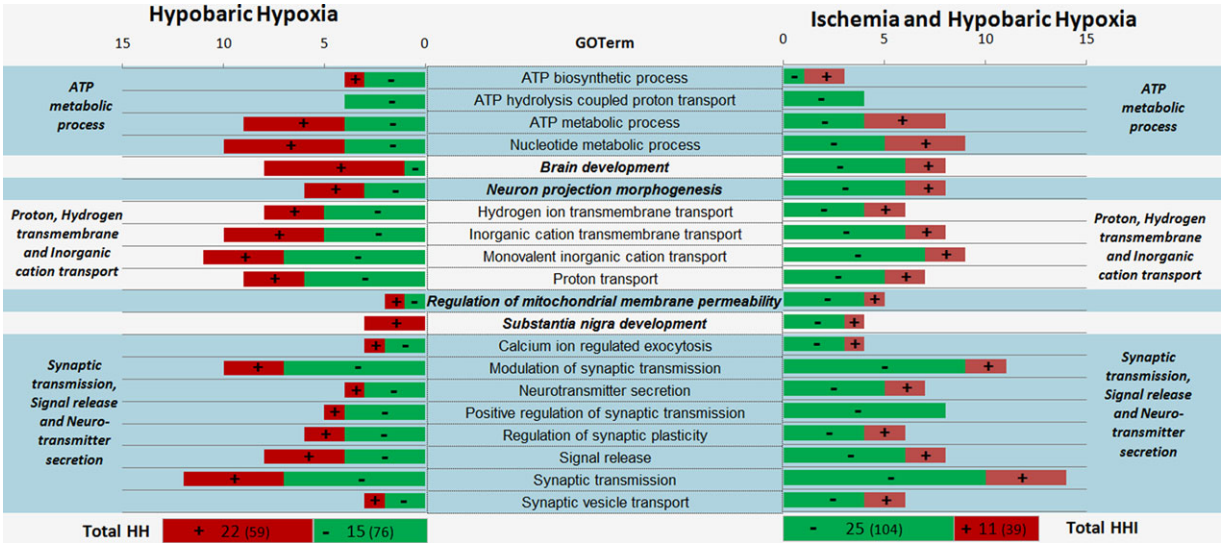


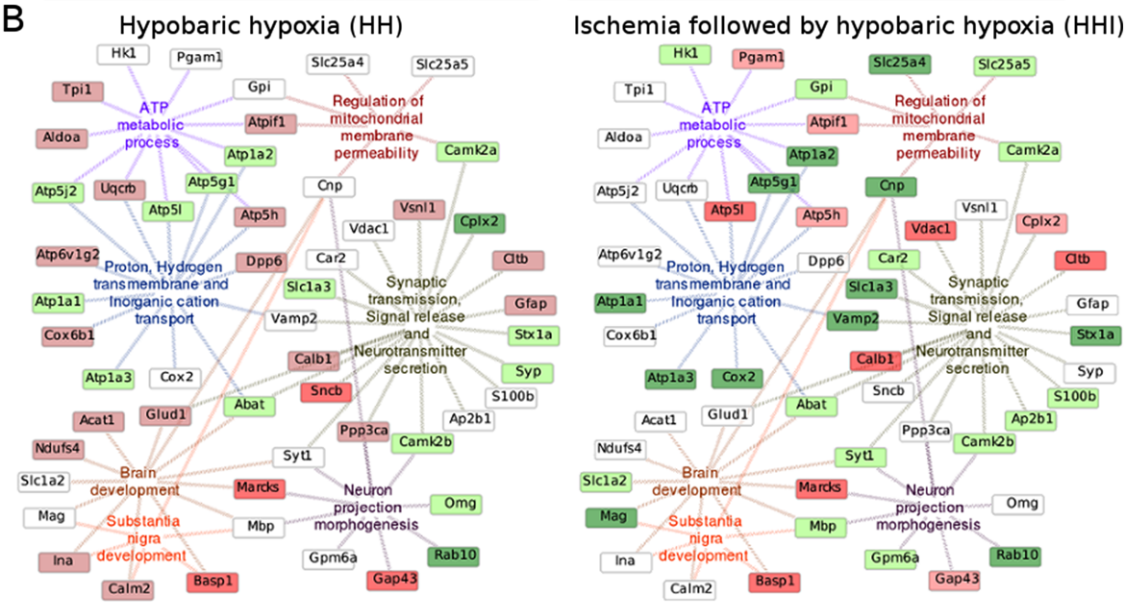
Figure 1. The proportion of over/underexpressed proteins in HH (37 proteins) and HHI (36 proteins) is shown for each of the 20 GO biological processes, grouped into seven functional groups (bold). Under the bar chart, the total of over/underexpressed proteins (in parentheses the number of times these proteins appear into one biological process), shows a general increase of protein expression in HH (22 protein with increased levels versus 15 decreased) and decrease in HHI (25 decreased versus 11 increased).

A

GO ID	Functional group	HH Pval	HHI Pval	Condition
GO:0046902	Regulation of mitochondrial membrane permeability	4.2E-01	**4.3E-05**	HHI
GO:0021762	Substantia nigra development	2.1E-02	**8.3E-04**	HHI
GO:0017156	Synaptic transmission, Signal release and Neuro-transmitter secretion	3.3E-01	**4.5E-02**	HHI
GO:0050804		1.6E-04	**2.0E-05**	
GO:0007269		1.0E-01	**1.1E-04**	
GO:0050806		1.9E-02	**1.4E-05**	
GO:0048167		4.1E-03	4.1E-03	
GO:0023061		2.5E-02	2.5E-02	
GO:0007268		5.1E-04	**1.5E-05**	
GO:0048489		4.0E-01	**5.0E-04**	

GO ID	Functional group	HH Pval	HHI Pval	Condition
GO:0048812	Neuron projection morphogenesis	6.8E-01	**6.3E-02**	HHI
GO:0006754	ATP metabolic process	**1.1E-03**	2.2E-02	HH
GO:0015991		4.6E-04	4.6E-04	
GO:0046034		**1.3E-05**	1.4E-04	
GO:0009117		**9.7E-03**	4.0E-02	
GO:1902600	Proton, Hydrogen transmembrane and Inorganic cation transport	**1.4E-07**	4.5E-05	HH
GO:0098662		**3.9E-03**	6.5E-02	
GO:0015672		**3.6E-04**	8.9E-03	
GO:0015992		**7.8E-08**	1.9E-05	
GO:0007420	Brain development	4.4E-01	4.4E-01	HH/HHI

B



C

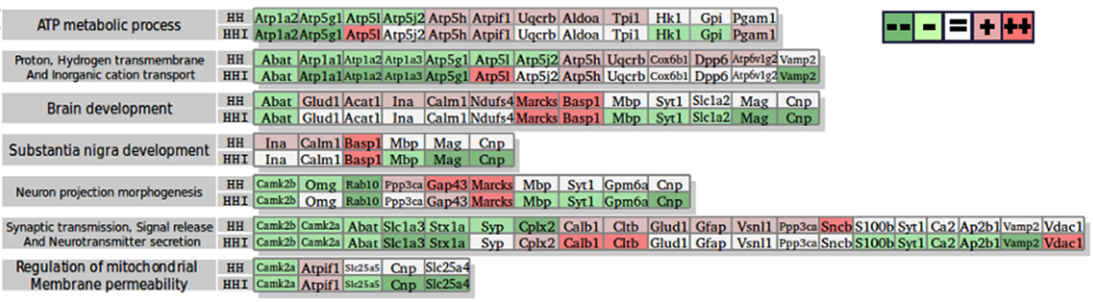


Figure 2. Gene enrichment analysis in HH and HHI. (A) Table showing GO terms associated to each functional group, *p*-values obtained for HH and HHI related genes (in bold the lowest) and condition (HH or HHI) in which the functional group is more enriched. (B) Relationships between functional groups and genes in HH and HHI. Genes are coloured dark and light green for high and moderate evidence of under expression, and dark and light red for high and moderate evidence of overexpression, respectively. (C) For each functional group, using the same legend, the list of genes related to HH and HHI experimental conditions.

exocytosis is also downregulated (Atp1a2 and Vamp2). Furthermore, response to hypoxia (Aldoa) and response to ischemia (HK1) markers [11] show differential expression on their respective conditions.

(iii) Brain development, (iv) Neuron projection morphogenesis, and (v) *Substantia nigra* development present upregulated genes—Gap43, Maks, and Basp1—all highly involved in signal transduction pathways, membrane transport and cytoskeletal dynamics [12]. The calmodulin-dependent protein kinases (Camk2a and Camk2b), that phosphorylate the central bioenergy sensor AMP-activated protein kinase [13], are downregulated in both HH and HHI; this same tendency is followed by Rab10, a small GTPase acting as regulator of membrane trafficking and fusion also involved in autophagy [14]. Additionally, several proteins related to substantia nigra development (Ina, Calm1, Mbp, Mag, and Cnp) show variation in HH and HHI, consistently with previous proteomic studies of changes in *Substantia nigra* caused by neurodegenerative diseases [15].

(vi) Synaptic transmission, Signal release, and Neurotransmitter secretion are greatly impaired under HHI, as expected under severe excitotoxic damage; interestingly,

the SNARE protein Vamp2, and its regulatory proteins Syt1, both highly involved in glutamate release and neuron damage after ischemic injury, are downregulated but only in HHI [16].

(vii) Regulation of mitochondrial membrane permeability points to the activation of apoptosis through mitochondrial pathways (downregulation of apoptosis inhibitors Gpi, Slc25a4 Slc25a5, and activation of Atpif1). Components of the mPTP (adenine nucleotide translocator: Slc25a4, Slc25a5, and Vdac1) [17] where also differentially expressed in HH and HHI.

In conclusion, HHI model presents a global effect of protein downregulation while HH produces an overall increase of the protein levels. With HH mainly affecting oxidative and energetic metabolism, HHI also interferes with synaptic transmission, neurotransmitter secretion, *substantia nigra* development and triggers apoptosis through mitochondrial pathway.

This study was supported by the Spanish Ministry of Science and Innovation (SAF2008-03938).

The authors have declared no conflict of interest.

2 References

[1] Dugan, L. L., Choi, D. W., Hypoxia-ischemia and brain infarction. In *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*. 6th edition (Eds.: Siegel, G. J., Agranoff, B. W., Albers, R. W., Fisher, S. K., Uhler, M. D.). Lippincott Williams and Wilkins, Philadelphia 1999.

[2] Granger, D. N., Kvietys, P. R., Reperfusion injury and reactive oxygen species: the evolution of a concept. *Redox Biol.* 2015, *6*, 524–551.

[3] Aarts, M. M., Arundine, M., Tymianski, M., Novel concepts in excitotoxic neurodegeneration after stroke. *Expert Rev. Mol. Med.* 2003, *5*, 1–22.

[4] Peinado, M. A., del Moral, M. L., Esteban, F. J., Martínez-Lara, E. et al., Aging and neurodegeneration: molecular and cellular bases. *Rev. Neurol.* 2000, *31*, 1054–1065.

[5] Rocha-Ferreira, E., Hristova, M., Plasticity in the neonatal brain following hypoxic-ischaemic injury. *Neural Plast.* 2016, *2016*, 4901014.

[6] Hernández, R., Blanco, S., Peragón, J., Pedrosa, J. Á., Peinado, M. Á. et al., Hypobaric hypoxia and reoxygenation induce proteomic profile changes in the rat brain cortex. *Neuromol. Med.* 2013, *15*, 82–94.

[7] Adhami, F., Liao, G., Morzov, Y. M. ,Schloemer, A. et al., Cerebral ischemia-hypoxia induces intravascular coagulation and autophagy. *Am. J. Pathol.* 2006, *169*, 566–583.

[8] Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P. A. et al., The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* 2007, *25*, 887–893.

[9] Vizcaíno, J. A., Côté, R. G., Csordas, A., Dianes, J. A. et al., The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* 2013, *41*(Database issue), D1063–D1069.

[10] Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P. et al., ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinforma. Oxf. Engl.* 2009, *25*, 1091–1093.

[11] Pasdois, P., Parker, J. E., Griffiths, E. J., Halestrap, A. P. et al., The role of oxidized cytochrome c in regulating mitochondrial reactive oxygen species production and its perturbation in ischaemia. *Biochem J.* 2011, *436*, 493–505.

[12] Sandal, M., Paltrinieri, D., Carloni, P., Musiani, F., Giorgetti, A. et al., Structure/function relationships of phospholipases C Beta. *Curr. Protein Pept. Sci.* 2013, *14*, 650–657.

[13] Khatri, N., Man, H.-Y., Synaptic activity and bioenergy homeostasis: implications in brain trauma and neurodegenerative diseases. *Front Neurol.* 2013, *4*, 199.

[14] Szatmári, Z., Sass, M., The autophagic roles of Rab small GTPases and their upstream regulators. *Autophagy* 2014, *10*, 1154–1166.

[15] Chen, S., Lu, F. F., Seeman, P., Liu, F. et al., Quantitative proteomic analysis of human substantia nigra in Alzheimer's disease, Huntington's disease and multiple sclerosis. *Neurochem. Res.* 2012, *37*, 2805–2813.

[16] Wang, Z., Wei, X., Liu, K., Yang, F. et al., NOX2 deficiency ameliorates cerebral injury through reduction of complexin II-mediated glutamate excitotoxicity in experimental stroke. *Free Radic. Biol. Med.* 2013, *65*, 942–951.

[17] Tanno, M., Miura, T., Adenine nucleotide translocator, a mitochondrial carrier protein, and fate of cardiomyocytes after ischaemia/reperfusion. *Cardiovasc. Res.* 2008, *80*, 1–2.

