Data analysis for discovering the protein profile dynamics of the human ovarian follicular fluid and BRAF mutated metastatic melanoma tissue.

-

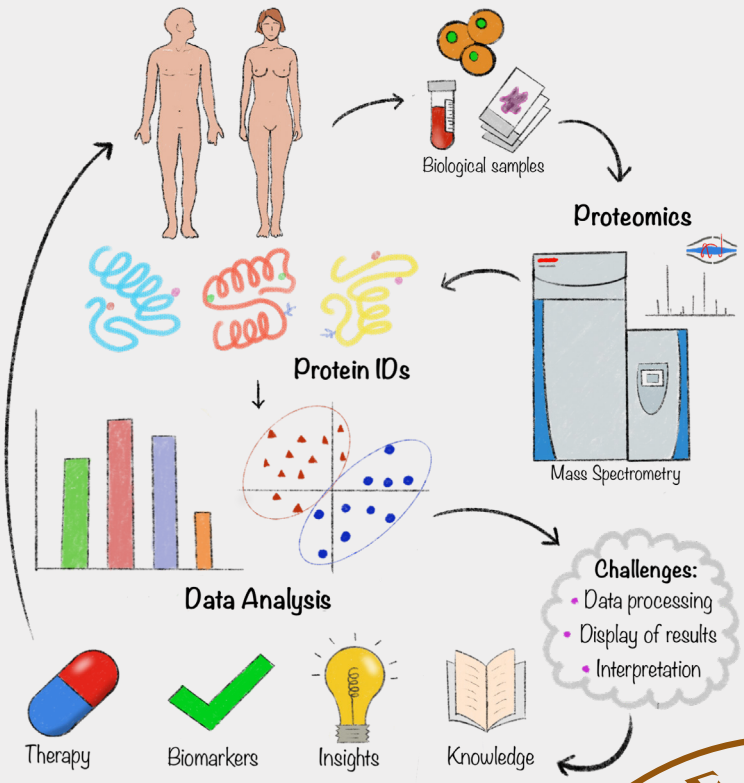
Pla Parada, Indira

2022

Document Version:
Publisher's PDF, also known as Version of record

Data analysis for discovering the protein profile dynamics of the human ovarian follicular fluid and BRAF mutated metastatic melanoma tissue

INDIRA PLÁ PARADA
DEPT. OF TRANSLATIONAL MEDICINE | FACULTY OF MEDICINE | LUND UNIVERSITY



Data analysis for discovering the protein profile dynamics of the human ovarian follicular fluid and BRAF mutated metastatic melanoma tissue

Data analysis for discovering the protein profile dynamics of the human ovarian follicular fluid and BRAF mutated metastatic melanoma tissue

Indira Plá Parada

LUND
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty of Medicine, Lund University, Sweden.
To be defended at Segerfalksalen, BMC A1005, Lund.

Date: 15th of September 2022 at 13:00.

Faculty opponent
Professor, Dr. rer.nat. Marius Ueffing
Director of the Institute for Ophthalmic Research
Co-Chair of the Centre for Ophthalmology at the University Medical Center of Tübingen, Germany

Organization LUND UNIVERSITY	**Document name** DOCTORAL DISSERTATION
	Date of issue
Author(s) Indira Plá Parada	Sponsoring organization

Title and subtitle Data analysis for discovering the protein profile dynamics of the human ovarian follicular fluid and BRAF mutated metastatic melanoma tissue.

Abstract

Proteomics is widely utilized to understand the function of cellular processes at the molecular level. Using liquid chromatography interfaced to mass spectrometry (LC-MS)-based proteomics, thousands of proteins can be identified and quantified in a single experiment and their relationship and interactions can be analyzed. This makes the analysis of high-throughput proteomic data a staple in the escalating field of translational medicine. Our group has been conducting LC-MS-based proteomics experiments to study two complex medical conditions that affect a high rate of the world population, female infertility and malignant melanoma (MM). To study female reproductive disorders our group profiled the protein composition of the ovarian follicular fluid (FF) since it constitutes the microenvironment in which the oocyte develops during antral stages until follicular rupture at ovulation. In addition, it is believed that the FF mirrors what happens at the molecular level in the ovary and plasma due to pathological disorders. In the case of malignant melanoma, we profiled the protein composition of metastatic tumor tissue from patients with BRAF mutation. The large amount of data generated from these experiments involved challenges related to data processing, analysis, and visualization of the results. In this thesis, I performed data analyses to interrogate proteomics data from a bioinformatics, and biostatistical point of view. Using different workflows, analysis and mathematical principles, I combined biological knowledge with bioinformatics and biostatistical approaches to integrate proteomics, clinical, and histopathological data to characterize the protein profiles and obtain relevant biological insights within a disease setting.

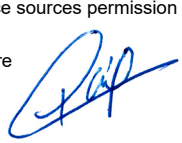
The strategy applied in **paper I,** allowed us to describe progressive proteomic changes occurring in the FF during the ovulation process linked with oocyte maturation, hormone regulation and release of the oocyte. Here, we studied the most detailed temporal ovulatory interval, which included five time points. **Paper II** constituted the first large-scale proteomics characterization of FF extracted from small antral follicles (SAF) (6.1±0.4 mm) in their natural state. Using a multivariate approach, a signature of proteins appeared to play a role in oocyte maturation and meiotic resumption from the early follicular stage. As a follow-up, **paper III** reported evidence of proteomic alterations occurring in the FF of SAF of polycystic ovaries (PCO) for the first time. Alterations were associated with the dysfunction of follicular growth and subsequent oocyte competence usually observed in PCO syndrome. Furthermore, uncharacterized or poorly characterized proteins identified in the FF of unstimulated SAF were assessed. Their functionality during folliculogenesis was described in **paper IV** (manuscript). In **paper V**, data analysis revealed for the first time that the high expression in the MM tumor of the B-raf V600E (mutated) protein could be a significant risk factor for poorer prognosis of patients with stage 3 and 4 of MM. A follow-up of this finding was done on a larger cohort of patients with BRAF mutation, where subgroups of patients with different mortality risks were identified and associated with the activation of different BRAF- related pathways, such as immune response.

This thesis, supported by data-driven results, characterized the protein profile dynamics of human ovarian FF during folliculogenesis (**paper I-IV**) and malignant melanoma tissue of patients with B-raf mutation (paper V). Findings from **paper I** to **IV** may open up new pathways for augmenting or attenuating subsequent oocyte viability in the pre-ovulatory follicle when ready to undergo ovulation, which opens up an opportunity for future advances in reproductive medicine. On the other hand, findings from paper V may enable the eventual delineation of patient responders/non-responders to therapy for malignant melanoma with BRAF mutation.

Key words: proteomics, data analysis, bioinformatics, follicular fluid, malignant melanoma

Classification system and/or index terms (if any)

Supplementary bibliographical information		**Language** English
ISSN and key title 1652-8220		**ISBN** 978-91-8021-271-7
Recipient's notes	**Number of pages 68**	Price
	Security classification	

Signature  Date 2022-08-08

Data analysis for discovering the protein profile dynamics of the human ovarian follicular fluid and BRAF mutated metastatic melanoma tissue

Indira Plá Parada

LUND
UNIVERSITY

Cover drawing by Jimmy Esneider Rodriguez

Faculty of Medicine
Department of Translational Medicine

ISBN 978-91-8021-271-7
ISSN 1652-8220
Lund University, Faculty of Medicine Doctoral Dissertation Series 2022:110

Printed in Sweden by Media-Tryck, Lund University
Lund 2022

To my family and all those who contributed to my personal and professional development

Table of Contents

Lay summary

Proteomics is widely utilized to understand the function of cellular processes at the molecular level. Using liquid chromatography interfaced to mass spectrometry (LC-MS)-based proteomics, thousands of proteins can be identified and quantified in a single experiment and their relationship and interactions can be analyzed. This makes the analysis of high-throughput proteomics data a corner-stone in the escalating field of translational medicine. Our group has been conducting deep-mining LC-MS-based proteomics studies on two complex medical conditions that affect a high rate of the world population, female infertility and malignant melanoma (MM). To study female reproductive disorders, our group profiled the protein composition of the ovarian follicular fluid (FF) since it constitutes the microenvironment in which the oocyte develops during antral stages until follicular rupture at ovulation. In addition, it is believed that the FF mirrors what happens at the molecular level in the ovary and plasma due to pathological disorders. In the case of MM, we profiled the protein composition of metastatic tumor tissue from patients with BRAF mutation. The large amount of data generated from these experiments involves challenges related to data processing, analysis, and visualization of the results. The main challenge in complex disease pathology is the unraveling of the data from experimental outputs. In most cases the answer lies within that biological sample – the challenge is to analyze it and understand the meaning of the data.

In this thesis, I performed data analyses to interrogate proteomics data (high-resolution LC-MS expression data sets) from a bioinformatics and biostatistical point of view. Using different workflows, analyses and mathematical principles, I combined biological knowledge with bioinformatics and biostatistical approaches to integrate proteomics, clinical, and histopathological data in order to obtain new relevant biological insights from protein profiles of ovarian follicular fluids and MM tissues.

The strategy applied in **paper I**, allowed us to describe progressive proteomic changes occurring in the FF during the ovulation process linked with oocyte maturation, hormone regulation and release of the oocyte. Here, we studied the most detailed temporal ovulatory interval, which included five time points. **Paper II** constituted the first large-scale proteomic characterization of FF extracted from small antral follicles (SAF) (6.1±0.4 mm) in their natural state. Using a multivariate approach, a signature of proteins appeared to play a role in oocyte maturation and

oocyte meiotic resumption already from the early follicular stage. As a follow-up, **paper III** reported for the first time evidence of proteomic alterations occurring in the FF of SAF of polycystic ovaries (PCO). Alterations were associated with the dysfunction of follicular growth and subsequent oocyte competence usually observed in PCO syndrome. Furthermore, uncharacterized or poorly characterized proteins identified in the FF of unstimulated SAF were assessed and their functionality during folliculogenesis was described in **paper IV** (manuscript).

In **paper V**, data analysis revealed for the first time that the high expression, in the MM tumor, of the B-raf V600E (mutated) protein could be a significant risk factor for poorer prognosis of patients with stages 3 or 4 of MM. A follow-up of this finding was performed on a larger cohort of patients with BRAF mutation, in which subgroups of patients with different mortality risks were identified and associated with the activation of different BRAF-related pathways, such as the immune response.

Supported by data-driven results, this thesis characterized the protein profile dynamics of human ovarian FF during folliculogenesis (**paper I-IV**) and malignant melanoma tissue of patients with BRAF mutation (**paper V**). Findings from **paper I to IV** may open up new pathways for augmenting or attenuating subsequent oocyte viability in the pre-ovulatory follicle when it is ready to undergo ovulation, which may be of importance to future advances in reproductive medicine. On the other hand, findings from **paper V** may enable the eventual delineation of patient response therapy for MM with BRAF mutation.

Poulärvetenskaplig sammanfattning

Mekanismerna bakom många sjukdomar är ofullständigt kända och för att bättre kunna förebygga, diagnostisera och behandla dem krävs en kartläggning av vad som sker i kroppens olika celler på ett molekylärt plan. En av de viktigaste typerna av molekyler, som finns i alla celler, är proteiner (äggviteämnen) och kartläggningen av proteiner kallas ofta proteomics. Med modern analysteknik, s.k. masspektrometri, kan tusentals proteiner identifieras och det krävs endast mycket små mängder vävnad eller blod. Analysen resulterar i mycket stora datamängder vilka kräver avancerade dataprogram för att bearbeta – ett arbete som brukar kallas bioinformatik.

Vi har studerat två mycket vanliga medicinska problem – infertilitet, en folksjukdom som berör 10-15 procent av barnönskande individer, samt malignt melanom (en form av hudcancer), den femte vanligaste cancerformen i Sverige som under långtid årligen ökat i förekomst (>4000fall/år).

För att bättre förstå processerna som leder fram till ägglossning har vi studerat vätskan i den follikel (vätskeinnehållande blåsa i äggstocken) i vilken ägget utmognar. I första och andra delarbetena kartlade vi förändingar i follikelvätskans sammansättmning i samband med ägglossning – kunskapen kan ligga till grund för såväl behandling av störningar i äggutmognaden, utveckling av nya preventivmedel samt till förbättrade resultat vid provrörsbefruktning. I det tredje delarbetet visar vi att follikelvätskans sammansättning är annorlunda hos kvinnor med cystor i äggstockarna, en vanlig orsak till minskad fertilitet. I fjärde delarbetet identifieras proteiner som tidigare aldrig karakteriserats.

I det femte delarbetet har vi studerat proteininnehållet i malignt melanom och visar att ökad förekomst av en mutation i ett protein kallat B-raf är associerat med en mer aggressiv tumör. Resultatet har stor betydelse för behandlingen av patienter med malignt melanom.

Sammanfattningsvis visar avhandlingens resultat betydelsen av att studera proteiner i patientprover. Många mediciner utövar sin effekt på proteiner varför en kartläggning av dessa är av största vikt för framgångsrik behandling. Dagens analysteknik kan generera mycket stora informationsmängder– information som kan ligga till grund för såväl diagnostik och behandling av olika sjukdomstillstånd som för utveckling av ny, skräddaresydd behandling av t.ex. cancer. En översikt över dessa två projekt kan ses på YouTubes plattform genom att följa länkarna:

European Cancer Moonshot Lund: https://youtu.be/QQVuvB8VMSo

ReproUnion:
https://www.youtube.com/watch?v=NyZNvTIO13c&t=17s&ab_channel=ReproUnion

List of papers included in this thesis

Paper I: Poulsen, L. la C.*; Pla, I. *; Sanchez, A.; Grøndahl, M.L.; Marko-Varga, G.; Yding Andersen, C.; Englund, A.L.M.; Malm, J. **Progressive changes in human follicular fluid composition over the course of ovulation: quantitative proteomic analyses**. *2019, Molecular and Cellular Endocrinology, Vol. 495, Article 110522*. (https://doi.org/10.1016/j.mce.2019.110522)

Paper II: Pla, I. *; Sanchez, A. *; Pors, S.E. *; Pawlowski, K.; Appelqvist, R.; Sahlin, K.B.; Poulsen, L.L.C.; Marko-Varga, G.; Andersen, C.Y.; Malm, J. **Proteome of fluid from human ovarian small antral follicles reveals insights in folliculogenesis and oocyte maturation**. *2020, Human Reproduction, Vol.* 36 (3), *Pag. 756–770 (https://doi.org/10.1093/humrep/deaa335)*

Paper III: Pla, I. *; Sanchez, A.; Pors, S.E.; Kristensen, S.G.; Appelqvist, R.; Sahlin, K.B.; Marko-Varga, G.; Andersen, C.Y.; Malm, J. Proteomic **Alterations in Follicular Fluid of Human Small Antral Follicles Collected from Polycystic Ovaries —A Pilot Study**. *2022*, *Life, Vol. 12 (3), Article 391*. (https://doi.org/10.3390/life12030391)

Paper IV: Pla, I. *, Pawlowski, K., Pors, S.E, Kristensen, S.G., Appelqvist, R., Marko-Varga, G., Andersen, C.Y., Malm, J., Sanchez, A. **Uncharacterized proteins identified in the follicular fluid proteome of small antral follicles reveal associations with folliculogenesis**. *manuscript paper*

Paper V: Betancourt, L.H., Szasz, A.M., Kuras, M., Murillo, J.R., Sugihara, Y.; Pla, I., Horvath, Z., Pawłowski, K., Rezeli, M., Miharada, K., Gil J, Eriksson J., Appelqvist R. Miliotis T., Baldetorp B., Ingvar C., Olsson H., Lundgren L., Horvatovich P., Welinder Ch., Wieslander E., Jeong Kwon H., Malm J., Balazs I., Jönsson J., Fenyö D., Sanchez A, Marko-Varga G. **The hidden story of heterogeneous B-raf V600E mutation quantitative protein expression in metastatic melanoma—Association with clinical outcome and tumor phenotypes**. *2019. Cancers. Vol. 11 (12), Article 1981.* (https://doi.org/10.3390/cancers11121981)

* First authorship

Papers not included in this thesis

During my doctoral studies, I collaborated with other researchers in the frame of the Cancer Moonshot and ReproUnion projects. My primary contribution to the following papers is related to the analysis of proteomics data. (* First authorship)

(1) Betancourt, L. H.; Pawłowski, K.; Eriksson, J.; Szasz, A. M.; Mitra, S.; **Pla, I.**; Welinder, C.; Ekedahl, H.; Broberg, P.; Appelqvist, R.; Yakovleva, M.; Sugihara, Y.; Miharada, K.; Ingvar, C.; Lundgren, L.; Baldetorp, B.; Olsson, H.; Rezeli, M.; Wieslander, E.; Horvatovich, P.; Malm, J.; Jönsson, G.; Marko-Varga, G. Improved Survival Prognostication of Node-Positive Malignant Melanoma Patients Utilizing Shotgun Proteomics Guided by Histopathological Characterization and Genomic Data. *Sci. Rep.* **2019**, *9* (1), 5154. https://doi.org/10.1038/s41598-019-41625-z.

(2) Gil, J.; Betancourt, L. H.; **Pla, I**.; Sanchez, A.; Appelqvist, R.; Miliotis, T.; Kuras, M.; Oskolas, H.; Kim, Y.; Horvath, Z.; Eriksson, J.; Berge, E.; Burestedt, E.; Jönsson, G.; Baldetorp, B.; Ingvar, C.; Olsson, H.; Lundgren, L.; Horvatovich, P.; Murillo, J. R.; Sugihara, Y.; Welinder, C.; Wieslander, E.; Lee, B.; Lindberg, H.; Pawłowski, K.; Kwon, H. J.; Doma, V.; Timar, J.; Karpati, S.; Szasz, A. M.; Németh, I. B.; Nishimura, T.; Corthals, G.; Rezeli, M.; Knudsen, B.; Malm, J.; Marko-Varga, G. ***Clinical Protein Science in Translational Medicine Targeting Malignant Melanoma***. *Cell Biol. Toxicol.* **2019**, *35* (4), 293–332. https://doi.org/10.1007/s10565-019-09468-6.

(3) Sanchez, A.; Kuras, M.; Murillo, J. R.; **Pla, I**.; Pawlowski, K.; Szasz, A. M.; Gil, J.; Nogueira, F. C. S.; Perez-Riverol, Y.; Eriksson, J.; Appelqvist, R.; Miliotis, T.; Kim, Y.; Baldetorp, B.; Ingvar, C.; Olsson, H.; Lundgren, L.; Ekedahl, H.; Horvatovich, P.; Sugihara, Y.; Welinder, C.; Wieslander, E.; Kwon, H. J.; Domont, G. B.; Malm, J.; Rezeli, M.; Betancourt, L. H.; Marko-Varga, G. ***Novel Functional Proteins Coded by the Human Genome Discovered in Metastases of Melanoma Patients***. *Cell Biol. Toxicol.* **2019**, *36* (3), 261–272. https://doi.org/10.1007/s10565-019-09494-4.

(4) Kim, Y.; Gil, J.; **Pla, I**.; Sanchez, A.; Betancourt, L. H.; Lee, B.; Appelqvist, R.; Ingvar, C.; Lundgren, L.; Olsson, H.; Baldetorp, B.; Kwon, H. J.; Oskolás, H.; Rezeli, M.; Doma, V.; Kárpáti, S.; Szasz, A. M.; Németh, I. B.; Malm, J.; Marko-Varga, G. ***Protein Expression in Metastatic Melanoma and the Link to Disease Presentation in a Range of Tumor Phenotypes***. *Cancers (Basel).* **2020**, *12* (3), 1–23. https://doi.org/10.3390/cancers12030767.

(5) Velasquez, E.; Szadai, L.; Zhou, Q.; Kim, Y.; **Pla, I.**; Sanchez, A.; Appelqvist, R.; Oskolas, H.; Marko-Varga, M.; Lee, B.; Kwon, H. J.; Malm, J.; Szász, A. M.; Gil, J.; Betancourt, L. H.; Németh, I. B.; Marko-Varga, G. ***A Biobanking Turning-point in the Use of Formalin-fixed, Paraffin Tumor Blocks to Unveil Kinase Signaling in Melanoma***. *Clin. Transl. Med.* **2021**, *11* (8), e466. https://doi.org/10.1002/ctm2.466.

(6) Betancourt, L. H.; Gil, J.; Sanchez, A.; Doma, V.; Kuras, M.; Murillo, J. R.; Velasquez, E.; Çakır, U.; Kim, Y.; Sugihara, Y.; **Pla I.**; Szeitz, B.; Appelqvist, R.; Wieslander, E.; Welinder, C.; Almeida, N. P.; Woldmar, N.; Marko-Varga, M.; Eriksson, J.; Pawłowski, K.; Baldetorp, B.; Ingvar, C.; Olsson, H.; Lundgren, L.; Lindberg, H.; Oskolas, H.; Lee, B.; Berge, E.; Sjögren, M.; Eriksson, C.; Kim, D.; Kwon, H. J.; Knudsen, B.; Rezeli, M.; Malm, J.; Hong, R.; Horvath, P.; Szász, A. M.; Tímár, J.; Kárpáti, S.; Horvatovich, P.; Miliotis, T.; Nishimura, T.; Kato, H.;

Steinfelder, E.; Oppermann, M.; Miller, K.; Florindi, F.; Zhou, Q.; Domont, G. B.; Pizzatti, L.; Nogueira, F. C. S.; Szadai, L.; Németh, I. B.; Ekedahl, H.; Fenyö, D.; Marko-Varga, G. ***The Human Melanoma Proteome Atlas—Complementing the Melanoma Transcriptome.*** *Clin. Transl. Med.* **2021**, *11* (7). https://doi.org/10.1002/CTM2.451.

(7) Betancourt, L. H.; Gil, J.; Kim, Y.; Doma, V.; Çakır, U.; Sanchez, A.; Murillo, J. R.; Kuras, M.; **Pla I.**.; Sugihara, Y.; Appelqvist, R.; Wieslander, E.; Welinder, C.; Velasquez, E.; Almeida, N. P.; Woldmar, N.; Marko-Varga, M.; Pawłowski, K.; Eriksson, J.; Szeitz, B.; Baldetorp, B.; Ingvar, C.; Olsson, H.; Lundgren, L.; Lindberg, H.; Oskolas, H.; Lee, B.; Berge, E.; Sjögren, M.; Eriksson, C.; Kim, D.; Kwon, H. J.; Knudsen, B.; Rezeli, M.; Hong, R.; Horvatovich, P.; Miliotis, T.; Nishimura, T.; Kato, H.; Steinfelder, E.; Oppermann, M.; Miller, K.; Florindi, F.; Zhou, Q.; Domont, G. B.; Pizzatti, L.; Nogueira, F. C. S.; Horvath, P.; Szadai, L.; Tímár, J.; Kárpáti, S.; Szász, A. M.; Malm, J.; Fenyö, D.; Ekedahl, H.; Németh, I. B.; Marko-Varga, G. ***The Human Melanoma Proteome Atlas—Defining the Molecular Pathology.*** *Clin. Transl. Med.* **2021**, *11* (7). https://doi.org/10.1002/ctm2.473.

(8) Almeida, N.; Rodriguez, J.; **Pla I.**.; Perez-Riverol, Y.; Woldmar, N.; Kim, Y.; Oskolas, H.; Betancourt, L.; Valdés, J. G.; Sahlin, K. B.; Pizzatti, L.; Szasz, A. M.; Kárpáti, S.; Appelqvist, R.; Malm, J.; Domont, G. B.; Nogueira, F. C. S.; Marko-Varga, G.; Sanchez, A. ***Mapping the Melanoma Plasma Proteome (MPP) Using Single-Shot Proteomics Interfaced with the WiMT Database***. *Cancers (Basel).* **2021**, *13* (24). https://doi.org/10.3390/CANCERS13246224.

(9) Mendonça, C. F.; Kuras, M.; Nogueira, F. C. S.; **Plá, I**.; Hortobágyi, T.; Csiba, L.; Palkovits, M.; Renner, É.; Döme, P.; Marko-Varga, G.; Domont, G. B.; Rezeli, M. ***Proteomic Signatures of Brain Regions Affected by Tau Pathology in Early and Late Stages of Alzheimer's Disease***. *Neurobiol. Dis.* **2019**, *130*, 104509. https://doi.org/10.1016/j.nbd.2019.104509.

(10) Zhou, Q.; Andersson, R.; Hu, D.; Bauden, M.; Sasor, A.; Bygott, T.; PawŁowski, K.; **Pla, I**.; Marko-Varga, G.; Ansari, D. ***Alpha-1-Acid Glycoprotein 1 Is Upregulated in Pancreatic Ductal Adenocarcinoma and Confers a Poor Prognosis***. *Transl. Res.* **2019**, *212*, 67–79. https://doi.org/10.1016/j.trsl.2019.06.003.

(11) Sahlin, K. B.*; **Pla, I**.*; Sanchez, A.; Pawłowski, K.; Leijonhufvud, I.; Appelqvist, R.; Marko-Varga, G.; Giwercman, A.; Malm, J. ***Short-Term Effect of Pharmacologically Induced Alterations in Testosterone Levels on Common Blood Biomarkers in a Controlled Healthy Human Model***. *Scand. J. Clin. Lab. Invest.* **2019**, *80* (1), 1–7. https://doi.org/10.1080/00365513.2019.1689429.

(12) **Pla, I**.*; Sahlin, K. B.; Pawłowski, K.; Appelqvist, R.; Marko-Varga, G.; Sanchez, A.; Malm, J. A ***Pilot Proteomic Study Reveals Different Protein Profiles Related to Testosterone and Gonadotropin Changes in a Short-Term Controlled Healthy Human Cohort***. *J. Proteomics* **2020**, *220*, 1–3. https://doi.org/10.1016/j.jprot.2020.103768.

(13) Kelemen, O.*; **Pla, I***.; Sanchez, A.; Rezeli, M.; Szasz, A. M.; Malm, J.; Laszlo, V.; Kwon, H. J.; Dome, B.; Marko-Varga, G. Proteomic Analysis Enables Distinction of Early- versus Advanced-stage Lung Adenocarcinomas. *Clin. Transl. Med.* **2020**, *10* (2), 1–18. https://doi.org/10.1002/ctm2.106.

(14) Giwercman, A.; Sahlin, K. B.; **Pla, I**.; Pawlowski, K.; Fehninger, C.; Lundberg Giwercman, Y.; Leijonhufvud, I.; Appelqvist, R.; Marko-Varga, G.; Sanchez, A.; Malm, J. ***Novel Protein Markers of Androgen Activity in Humans: Proteomic Study of Plasma from Young Chemically Castrated Men***. *Elife* **2022**, *11*. https://doi.org/10.7554/ELIFE.74638.

(15) Barbara Sahlin, K.; **Pla, I.**; de Siqueira Guedes, J.; Pawłowski, K.; Appelqvist, R.; Marko-Varga, G.; Domont, G. B.; Nogueira, F. C. S.; Giwercman, A.; Sanchez, A.; Malm, J. ***Short-Term Effect of Induced Alterations in Testosterone Levels on Fasting Plasma Amino Acid Levels in Healthy Young Men***. *Life (Basel, Switzerland)* **2021**, *11* (11). https://doi.org/10.3390/LIFE11111276.

Abbreviations

FF	Follicular fluid
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DNA	Deoxyribonucleic acid
ECM	Extracellular matrix
EMT	Epithelial-to-mesenchymal transition
FDR	false-discovery rate
GC	Granulosa cell
GO	Gene ontology
HPA	Human protein atlas
HPP	Human proteome plasma
HPLC	high-performance liquid chromatography
hSAF	Human small antral follicle
KM	Kaplan-Meier
LASSO	least absolute shrinkage and selection operator
LC-MS	Liquid chromatography -mass spectrometry
MM	Malignant melanoma
mRNA	Messenger ribonucleic acid
MS	Mass spectrometry
OI	Ovulation induction
PCO	Polycystic ovary
PCOS	Polycystic ovary syndrome
PRM	Parallel reaction monitoring
SAF	Small antral follicle
sPLS-DA	Sparse partial least squares– discriminant analysis
TC	Theca cell
WHO	World Health Organization

Introduction

Continuous advances in the fields of life and medical sciences, as well as technology, have resulted in a torrent of data that fueled the emerging growth of data science as an indispensable discipline in the field of translational medicine. Specifically, the so-called 'omics' fields (such as proteomics, genomics, and transcriptomics), are one of the biggest generators of data. These technologies produce high-throughput data, therefore their link to data science is eminent. Data science combines various tools, algorithms, and machine learning principles to discover hidden patterns from raw data. However, in translational medicine, this discipline gains importance when the discoveries are relevant to human diseases, and as a consequence, an improvement in human health can be achieved.

Nowadays, the need to integrate data from different platforms utilized in translational medicine to reach more accurate conclusions about human diseases is increasingly evident. The integration of different omics data has drawn attention as it captures the interconnection between different molecular levels and has proven to be more efficient when trying to understand complex disease pathologies[1–3]. The biggest challenge in the study of complex diseases is the unravelling of the data from experimental outputs. When selecting the sample to evaluate, in most cases the answer often lies within the selected biological sample – the challenge is to analyze it and understand the meaning of the data.

During the course of this thesis, I will present two common and complex medical conditions that are studied in our group. These are: 1) female infertility and 2) malignant melanoma (MM). I will also present different data analysis strategies carried out to characterize the protein profiles of ovarian follicular fluid and MM tissues to obtain relevant biological insights that may benefit these patients.

Medical conditions involved in this thesis

According to the WHO, around 50 % of the cases of infertility, faced by a couple, are due to female reproductive factors. These conclusions stems from data generated from 186 million ever-married women of reproductive age in developing countries (https://www.who.int/). Considering cancer diseases, MM is one of the most aggressive and heterogeneous cancers and is the most frequently mutated tumor type. Specifically, in Sweden, it is the fifth most prevalent type of cancer and more than 4000 cases are diagnosed annually. Its also predicted that the incidence of MM in Sweden is expected to increase by 21%, whereas the mortality is expected to increase 35%, by 2040 [4] (https://gco.iarc.fr/tomorrow). Next, I will be expanding further on these two medical conditions and the biological mechanisms that are central within the progressive developments of these diseases.

Female Infertility

People in general may have heard of infertility, but not many are aware that since 2009, the WHO considers this condition to be a disease. This is "a disease of the reproductive system defined by the failure to achieve a clinical pregnancy after 12 months or more of regular unprotected sexual intercourse"(WHO). Like most other diseases, this condition has a negative impact on the patient's quality of life. Particularly women are at greater risk of having not only physical but also emotional negative consequences. In some cases, they are victims of violence, divorce, depression and anxiety. But what really causes female infertility?

Several medical conditions may lead to infertility. Some have to do with structural problems of the reproductive system, which usually involve the fallopian tubes and/or uterus (https://www.nichd.nih.gov/). Infections located in the reproductive system can also cause infertility. Another abnormality occurs when the cells that normally line the endometrium (uterine cavity) are found outside the uterus. This phenomenon is called endometriosis. The uterus sometimes develops noncancerous growths, called fibroids that may cause infertility depending on their size and location [5]. In addition, some autoimmune diseases may impact fertility. This occurs because the immune system does not recognize endogenous tissue/fluids of the body as normal, and ends up attacking them [6]. However, the most common causes of infertility are related to ovary dysfunction and its relationship with the brain. A failure during the ovulatory process occurs in 40 % of women with infertility issues [7]. This failure may be due to a diminished number of eggs in a woman's ovaries (also called ovarian reserve), or an incapability of the egg to mature. It may also be due to gynecological conditions such as polycystic ovary syndrome (PCOS). PCOS is one of the most common causes of female infertility with a prevalence of ~21% [8]. This condition refers to increased androgen production by the adrenal glands that interfere with the development of ovarian

follicles and the release of eggs during ovulation. To study abnormalities related to ovary function, it is essential to understand the folliculogenesis process and the physiology of the female reproductive axis.

The Female Reproductive Axis

The female reproductive system interacts with a complex network of endocrine-, paracrine- and autocrine-feedback loops that originate in the brain, specifically from the hypothalamus and the pituitary glands [9]. Coordinated communication between the nodes of this network is essential to ensure a proper function of the reproductive system. In this communication, hormones are used by the body to relay signals from one organ to another. During the reproductive age of women, the role of the ovary in this process is to prepare an oocyte (egg) every month for fertilization. Thus, the ovary is responsible for the recruitment, maturation and release of this oocyte, whereas the uterus will be working in parallel to prepare the optimal environment in which the embryo resulting from successful fertilization of the oocyte by the sperm, is going to be implanted to start a pregnancy. These events happen every month as part of a cyclic process in which the hypothalamus and pituitary gland play a key role.

During embryonic development, the gonadotropin-releasing hormone (GnRH) is produced by neurons that migrate from the olfactory area to their primary location within the hypothalamus. During this process, also the pituitary gland develops composed of an anterior and a posterior part. Later, during fetal life, oocytes begin to develop as germ cells (e.g. oogonia) inside a follicle composed of circumferential layers of granulosa- (inner layer) and thecal (outer layer) cells (GCs and TCs). These oocytes are arrested in the prophase of the first meiotic division until after puberty and the initiation of ovulation.

At puberty, women start having a monthly menstrual cycle during their reproductive age. At this point, a women's ovary contains ~300,000 resting primordial follicles (oocytes surrounded by pregranulosa cells) [10], however, only a few of them will be activated each month to start growing until ovulation. Before activation of the follicles at poverty, granulosa cells and theca cells are non-steroidogenic. At puberty, the previously suppressed hypothalamus triggers the pituitary by releasing GnRH which leads to the release of gonadotropins, the follicle-stimulating hormone (FSH) and luteinizing hormone (LH) (Figure 1). The gonadotropins will stimulate the ovary to start the production of steroid hormones. This production takes place in the ovarian follicles (ovarian sacs containing the oocyte and delineated by GCs and TCs). Once the LH binds to its receptors in the follicular TCs, it will stimulate the intracellular conversion of cholesterol to androgens. These androgens are then transported from the TC layer into the GCs to serve as substrates for estrogen production. The binding process between FSH and its receptor (FSH receptor), located in the cellular membrane of the GCs, stimulates the production of aromatase enzymes to convert these androgens into estrogens. Each month, this process begins with a cohort of three

to eleven follicles that start to grow. However, only one becomes the dominant follicle that carries the oocyte's full maturation and ovulation. GCs of the activated follicles upregulate the anti-Müllerian Hormone (AMH) expression to suppress the maturation of the non-activated follicles.

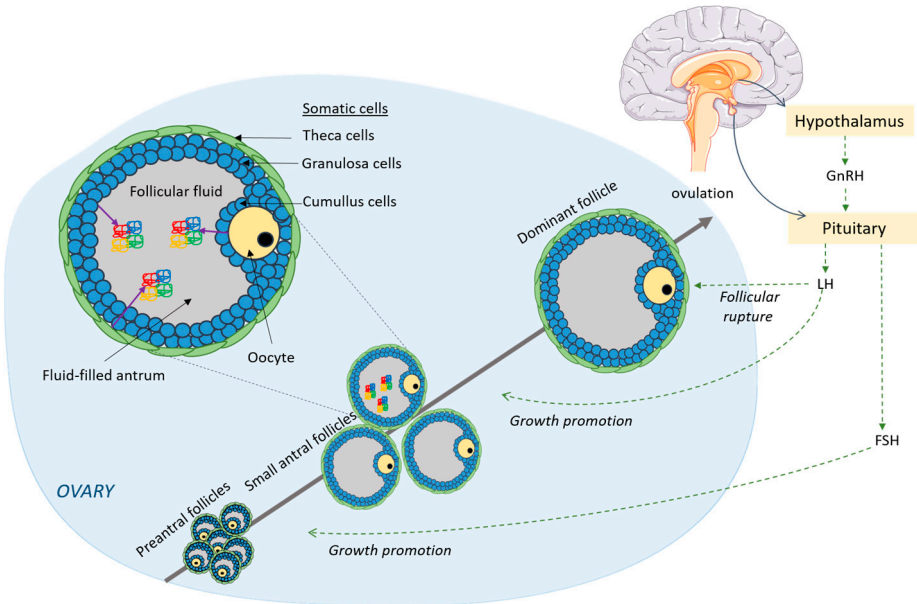


Figure 1. Schematic representation of the Hypothalamic-Pituitary-Gonadal Axis in women. An expanded view of an antral follicle indicates the composition of a typical antral follicle.

Every month, increasing levels of FSH induce the GCs to increase the number of FSH receptors, which results in an exponential increase in the GC's sensitivity to gonadotropins. The follicle that develops the most FSH receptors in response to FSH will become the dominant follicle. The dominant follicle will begin to grow, thereby increasing the number of LH receptors on its TCs. This fact leads to increased production of steroids available for conversion to estrogen. Subsequently, the dominant follicle secretes large amounts of estrogen that support the developing oocyte. Ultimately, LH promotes the follicular rupture and release of the oocyte by upregulating a cascade of proteolytic enzymes and decreasing the gap junction proteins [10].

Follicular Fluid

During this lengthy period of follicular development, the avascular follicle increases from approximately 45 µm to about 20 mm in diameter. This process involves several developmental checkpoint stages, one of which is the development of the follicular fluid (FF)-filled antrum that begins to form when the human ovarian follicles reach a

diameter of approximately 250 µm [11]. The FF is comprised of secretions from the oocyte, GCs (including cumulus and mural GCs), TCs and transudates from circulation that are filtered through the basal membrane (Figure 1). TC secretions are diffused across the basal membrane surrounding the follicle. The basal membrane acts as a molecular filter and proteins with a relatively high molecular weight can only penetrate the FF to a limited extent [11]. For instance, FF does not coagulate due to low concentrations of the high molecular weight fibrinogen. The composition of FF is highly-variable and associated with the follicles' developmental stage. In particular, FF reflects GC activity, which is also highly variable and strongly dependent on gonadotropins, other hormones and growth factors. For example, the TGF-β growth factor anti-Müllerian hormone (AMH) is present at very high concentrations in small antral follicles (SAF) with a peak in follicular content around a diameter (size) of 8 mm [12]. Conversely, sex steroids such as estradiol and progesterone accumulate at very high concentrations in the pre-ovulatory follicles, in orders of magnitude higher than in small antral follicles [12].

The FF and GCs constitute the microenvironment in which oocytes develop. In particular, FF affects the development of immature oocytes. Consequently, the protein composition of GC and oocyte together with the FF have attracted considerable interest and several proteomics studies have been conducted [13–17]. Figure 2 shows a 'hypothetical' representation of the normal and abnormal ovarian function and possible studies to be performed. The proteomics study of the FF from different follicular stages and specific medical conditions (such as PCOS) could give new insights into ovarian function that may be relevant for the advancement in reproductive medicine.

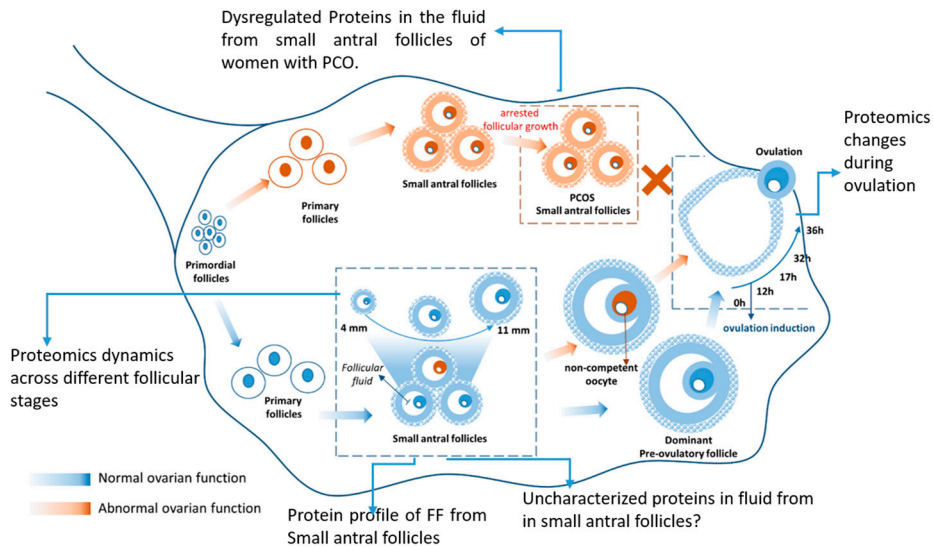


Figure 2. Hypothetical representation of a normal and abnormal ovarian function and possibles studies to carry out.

Malignant Melanoma

Malignant melanoma is the most aggressive and lethal form of all skin cancers [18]. It develops from melanocytes which are the cells that produce pigmentation (melanin). Melanocytes are derived from the neural crest and colonized in the skin, the eyes, and several tissues during the development of the human body [18]. Most melanoma tumors arise in the skin, however, they may also arise from mucosal surfaces or eyes. The cause of this disease may be attributed to a multitude of risk factors, where ultraviolet (UV) exposure is one of the most common, together with genetics. Increased exposure to UV light (e.g, sun light) plus a genetic susceptibility may induce the accumulation of genetic mutations in melanocytes. These mutations may activate oncogenes that inactivate tumor suppressors genes and interrupt DNA repair. Ultimately leading to an induction of melanocytes to proliferate, blood vessel growth, tumor invasion, and diminished immune response [19]. Clinical subtypes of this tumor are classified according to cancer's location, morphology, color and pathophysiology. Melanoma classification according to WHO includes superficial spreading melanoma (the most common), nodular melanoma, lentigo MM and acral lentiginous melanoma.

The optimal treatment for MM patients depends primarily on the stage of the disease. According to the TNM classification (globally recognized standard classification of malignant tumors), MM can be classified into five major stages (with subcategories). These stages are assigned based on the thickness of the tumor, whether there are metastases in nearby lymph nodes and whether there are distant metastases. Stage 0 (also called *in situ*) is the less aggressive form. During this stage, the cancer is confined to the epidermis, the outermost skin layer. Stage IV is the most aggressive form and indicates that the disease is spread and there are metastases in lymph nodes and other parts of the body. Therapy options includes surgery as the front-line therapy to remove the tumor. Unlike other cancer types, chemotherapy and radiation are rarely used due to their poor efficacy and side effects [20]. The second frontline therapies for MM are immunotherapy and targeted therapy. Immunotherapy is mainly used for advanced MM; specifically, the PD-1 antibody has generated promising results in patients with metastatic disease. This antibody inhibits the PD-1 immune checkpoint protein, a well-known regulator of the immune system stimulated by cancer cells to evade attacks from the immune system [21]. Based on the same principle, cytotoxic T-lymphocyte antigen 4 (CTLA-4) antibodies are used for tumors with low PD1 expression. In addition, tyrosine kinase inhibitors are used to counter the BRAF V600E mutation that many patients harbor. Immunotherapy and BRAF inhibitors are commonly used in combination for optimal patient outcomes.

Genetically defined subgroups have also been described [18], presenting different mutational profiles, e.g., BRAF, RAS, NF1, as well as triple wild-type (WT). The first two (the most common) are oncogenes within the mitogen-activated protein

kinase (MAPK) pathways. No universal mutation for all cutaneous melanomas has been identified. However, oncogene and/or tumor suppressor mutations often result in the MAPK pathway's constitutive activation [22]. In addition, among the most common pathways affecting MM is the phosphoinositol 3-kinase (PI3K) pathway, along with the Wnt signaling pathway. An intratumoral variation has also been observed in MM tissues. In a previous study performed with the most extensive melanoma data repository [23] collected from 500 tumor samples, we observed an interplay between stromal and tumoral cells in primary tumors. In addition, a diversity of clonal evolutionary pathways of metastatic tumors have been previously reported [23]. The field of digital pathology is making rapid progress using artificial intelligence (AI) to characterize MM tumors at the single-cell level to associate individual cells characteristic to a specific phenotype [24]. Our group was one of the first to outline digital pathology and AI for melanoma patients [23].

In general, MM is considered one of the most heterogeneous diseases. Variations in clinical symptoms, appearance, morphology and molecular profile of the tumor makes it difficult to perform an accurate diagnosis of the disease. Consequently, it makes the therapy decision-making hard. [25]. While patients diagnosed at an early stage may be cured through surgical excision, those who develop metastasis, rapidly progress to regional lymph nodes (stage 3) and distal organs (stage 4). This evolvement accelerates the tumor progression and reduces survival time, often to less than 1-year [26]. Figure 3 illustrates the heterogeneity that may be observed in patients with BRAF mutated tumors. Some patients develop a homogenous tumor (Figure 3A), while others present a more heterogeneous and dispersed B-raf expression (Figure 3B).

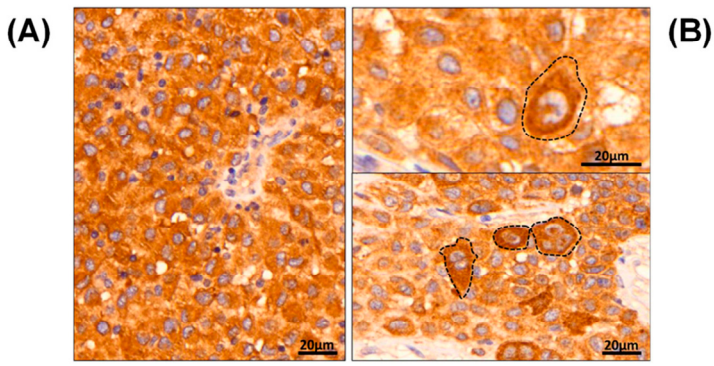


Figure 3. Immunohistochemical images of mutated B-raf V600E, displayed from two patients with MM. (A) a patient with homogeneous B-raf expression, (B) two IHC images generated from two different areas of the same tumor with heterogeneous and dispersed B-raf expression, highlighted by brown colorimetric reaction.

Patients with BRAF mutation

Approximately 40-50 % of MM patients harbor the BRAF mutation [27] and about 90% of these are BRAF V600E [28]. B-raf is a serine/threonine protein kinase encoded by the BRAF gene on chromosome 7q34 that activates the MAP kinase/ERK-signaling pathway to induce cell growth and proliferation. The mutation is located to codon 600 and consists of a single nucleotide mutation resulting in the substitution of glutamic acid for valine (V600E). Current therapies aim to reduce both the activity and subsequent development of metastases by targeting BRAF V600E; and above all, lowering of the tumor burden of the patient.

BRAF inhibitors (e.g., Vemurafenib, Dabrafenib) combined with MEK inhibitors (Trametinib) constitute the effective therapy to counter BRAF V600E mutation in patients with MM. However, after a period of successful treatment, most patients develop resistance, and/or get a relapse, which induces an accelerated progression of the disease. The reactivation of the MAPK pathways and ERK1/2 activity have been suggested as possible causes of this resistance. Although there are several studies on the potentially underlying factors causing resistance to BRAF inhibitors, many clinical questions require alternative research approaches to address the molecular mechanisms resulting in metastasis development and treatment-resistant melanoma. Even within a genetically defined subgroup of patients (such as BRAFmut patients), heterogeneity still exists. Clone-specific expression diagnosis can be made by digital pathology, whereby the heterogeneity patterns is defined. Proteomics is a highly promising research field that can be applied in order to generate new insights into the microenvironment of metastatic tumors with BRAF mutation. In particular, this can be achieved using a minimal amount of sample at single clone levels.

Studying diseases at the molecular level

What happens in the cell?

Our body is made up of functional systems and organs, which in turn are made up of cells that have the power to control our health status. What happens at the cellular level to control our health status is intricate and beautiful. The cells continuously replicate themselves not only to ensure the growth of tissues that comprise the human body, but also to ensure procreation. Specifically, germ cells (precursors of gametes, either eggs or sperm) undergo a nuclear division process called meiosis. During this process, gametes are produced in male and female gonads. Involving two rounds of nuclear division and an ordered series of events, meiosis produces four cells that are genetically different. Ultimately, during a process of fertilization, cells produced by meiosis from a male and female will fuse to create the zygote (a

cell that contains the genetic information of a new individual organism). On the other hand, somatic cells (non-sex cell) replicate themselves through mitosis. Unlike meiosis, during mitosis, the genetic material (DNA) in a cell is duplicated and divided equally to generate two identical daughter cells.

Both during replication and when other cellular functions are carried out, the protein molecules are key players. Proteins can be regarded as cellular workhorses, responsible for many of the various cell actions needed for function. The cell carries in its nucleus, the DNA that stores our phenotypical information and instructions to produce proteins and RNA molecules. The DNA distributes this information among chromosomes which are made up of segments of DNA, called genes. These genes are the unit of heredity but they only come to life after being translated to proteins.

Proteins are the gene's products and are considered as the main functional actors in the human body. They perform a broad range of functionalities including biochemical reactions, signaling, transport, and structural support. For a gene to be translated into a protein, it first has to be transcribed into a molecule called RNA (ribonucleic acid). Briefly, transcription and translation are the two processes that transform a sequence of nucleotides (genes) from DNA into a sequence of amino acids to build a protein (Figure 4). During transcription (carried out in the nucleus), a strand of DNA is used as a template to build a similar molecule of RNA (transcripts), which in turn will constitute the link between the DNA and the production of proteins. During the translation process carried out in the cytoplasm, the transcript is converted into an amino acid sequence which will constitute the building blocks of functional proteins.

Proteins do not function in isolation, they must interact with each other and with other types of molecules (e.g DNA, RNA) in order to mediate not only cellular processes but also metabolic- and signaling- processes [29]. Some of them remain in the cell after being synthesized to perform intracellular functions, whereas others are transported out of the cell to carry out different extracellular functions.

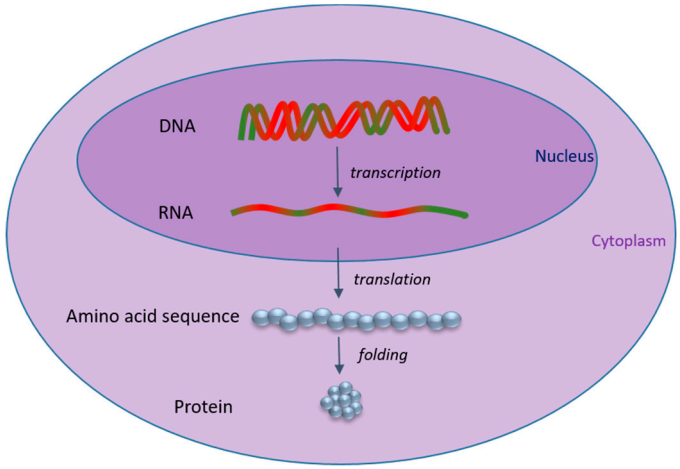


Figure 4. Protein syntesis. The transition from DNA to protein within the cell.

Due to the important role that proteins have in the human body, the way they act or interact also influences the mechanisms leading to diseases. This is the main reason why an increasing number of proteomic studies are being carried out in the field of translational medicine. Most of them aim to illustrate the relationship between protein expression and clinical phenotypes.

Clinical Proteomics

Clinical proteomics has an enormous impact and potential in generating knowledge on the molecular mechanisms of diseases. This field studies the proteome of cells and other clinical specimens such as human tissues and body fluids. In general, proteomics deals with the large-scale determination of gene and cellular function directly at the protein level [30]. A valuable characteristic of this discipline is the capacity it has to evaluate hundreds to thousands of proteins simultaneously. At the core of proteomics is the technique LC-MS, which provides a sensitive analysis of complex mixtures of proteins and reveals the qualitative and quantitative status of the molecular profile in a given clinical sample.

Mass spectrometry-based proteomics

MS-based proteomics is a technology established to interpret the information encoded in genomes. It is used to study protein-protein interactions, do mapping of numerous organelles, perform quantitative protein profiling of diverse species and specimens and detect post-translational modifications (PTM) [30]. MS is today's method of choice for the analysis of complex protein samples. This analytical

technique detects the presence and abundance of ionized peptides (or other biomolecules such as metabolites, lipids and proteins), by measuring their mass-to-charge ratio (m/z) [30,31]. This process begins with a step of sample preparation in which the biological specimens at hand are biochemically enriched by extracting the proteins. The overall goal of sample preparation is to obtain a mixture of peptides, thus, the extracted solubilized proteins are enzymatically digested with a sequence-specific protease; in most cases with trypsin alone or in combination with other enzymes. The complex peptide mixture is then injected into a sensitive LC-MS system with high resolving power to separate and ultimately quantify the peptides.

High-Resolution Liquid Chromatography Separation

Proteomic studies usually utilizes a chromatographic separation of peptides generated from the protein content of samples, rather than proteins separation. The reason being, that proteins are more difficult to separate and analyze, than peptides. This also holds true for the MS part of the data generation.

The principle of chromatographic separation of a peptide's mixture, is based on the interaction of the solutes with the solid support (stationary phase) and the mobile phase[32]. The remarkable volume of peptides generated in proteomic experiments outperforms direct mass spectral analyses. The high-performance liquid chromatography approach in proteomics aims to fractionate peptide mixtures to allow and maximize peptide identification and quantification by mass spectrometry. The analysis constitutes a challenge, even for the most advanced chromatographic separation equipment and mass spectrometers. The simplest and most direct way to combine the system is using a single chromatography method (online reverse phase column coupled to MS). However, the peptide mixture is still very complex and multidimensional liquid chromatography (MDLC) analysis are often necessary to increase the overall peak capacity, resolution and consequently the proteome coverage by mass spectrometry [33,34]. MLDC is the process of separating peptides using two or more physical properties with different chromatographic separation schemes (columns or dimensions). The separation is considered orthogonal when these methods are independent [35]. The most commonly exploited physical properties and their associated chromatography methods are mass (size-exclusion chromatography (SEC)), charge (ion exchange chromatography), hydrophilicity (normal chromatography), hydrophobicity (reverse phase chromatography), and biological interaction (affinity chromatography) [33,34,36,37]. The most efficient approach to generate separation is nano-chromatography, and its elution into the MS. Thus, it has to be emphasized that there is no separation technology available today that manages to separate and isolate hundreds of thousands of peptides in for instance a tumor sample. This is still a limiting factor for clinical proteomics.

Mass spectrometry

The analytical instrument of this technology is the mass spectrometer, which simultaneously holds a powerful capacity for peptide separation. This instrument consists of an ion source, a mass analyzer and a detector that registers the number of ions at each m/z value. The most common way mass spectrometers are used to analyze complex samples is by integrating high-performance liquid chromatography (HPLC) with an ion source called electrospray ionization (ESI) [38]. This interfaced system is commonly referred to as LC-MS. Peptide mixtures are commonly separated based on their hydrophobicity and charge using the HPLC system. Here, the eluent is introduced into the MS and volatilized and ionized using ESI. Although other ionization methods exist, such as matrix-assisted laser desorption/ionization (MALDI), ESI is often the preferred method for analyzing complex mixtures of proteins. ESI converts peptides from the liquid phase to gaseous ions by pumping, at high voltage, the liquid containing the peptides through a micrometer-sized orifice. This induces the disintegration of the liquid, leaving peptide ions in the gas phase (John Fenn received the Nobel Prize in 2002 for this discovery) [31]. Depending on the hydrophobicity and charge of the peptides, they will diffuse through the chromatography column at different velocities. The time a peptide is retained in the column before being eluted into the mass spectrometer is called the retention time (RT).

The mass analyzer of the spectrometer aims to separate ions by modulating their trajectories in electrical fields. The most common are the Quadrupoles, usually combined with a time-of-flight (TOF)-, and an Orbitrap- analyzer. The first one separates ions using an oscillating electrical field between four cylindrical rods in a parallel arrangement. After an induced acceleration, the TOF mass analyzer separates the ions based on the velocities they reach and their subsequent arrival times at the detector. In the case of the Orbitrap mass analyzer, it distinguishes ions based on their oscillation frequencies. Ions are tangentially injected and then trapped in the Orbitrap, moving along the length axis of a central metal spindle [31]. A quantitative readout of the strength of individual ion packets is then achieved after transforming (Fourier transformation) the so-called ‘image current’ induced by the rapidly oscillating ions into a frequency domain.

Finally, the MS instruments sequence peptides by using precursor ions, firstly isolated by the quadrupole, and subsequently fragmented through collision with inert gases to break them apart at the lowest energy bonds. This process generates the MS/MS spectrum that shows the amino acid sequence of the peptides after sequencing.

Large-scale MS-data acquisition and quantification

Different methodologies are used for data acquisition. They can be divided into two main categories: targeted proteomics and discovery proteomics. Targeted proteomics focuses on hypothesis-driven methods that use a predefined set of

peptides; in this case, the protein(s) of interest is/are known. The two most common strategies for targeted proteomics are based on approaches where the triple quadrupole mass analyzer is operated in multiple reactions monitoring (MRM) mode [39], or in parallel reaction monitoring (PRM) mode [40–42]. Both methods simultaneously perform relative or absolute quantitative detection of multiple target proteins in complex biological samples. Peptides of interest are selected based on their precursor ion mass in the first quadrupole (Q1). Differences between these two methods lie in the second step for protein detection. MRM detects specific product ions after the fragmentation of the precursor ions, and qualitatively quantifies the protein through the one-to-one correspondence between the precursor and product ions. PRM detects all the product ions through a high-resolution detector after fragmenting the precursor ions. Overall, both methods are widely used; however, PRM has been described as more accurate than MRM, with much less ion interference [43,44].

Discovery proteomics comprises approaches not limited to the large-scale identification of a predefined set of proteins (in this case, the proteins of interest are unknown). The data acquisition can be dependent or independent. The data-dependent acquisition (DDA) is based on user-defined rulers (e.g m/z, charge, intensity, and cross-section) followed by the mass spectrometer which selects as many peptides as possible for acquiring MS/MS spectra. This method has a stochastic nature since during the selection of the peptides there are more peptides than analysis time, resulting in the generation of missing values. On the other hand, data-independent acquisition (DIA) is a relatively new approach that promises to solve this limitation. Unlike DDA, DIA fragments every single peptide in a sample. This leads to very complex MS/MS spectra and demands high-performance instruments.

Discovery proteomics commonly relies on the relative quantification of peptides. The strategies divides into two classes referred to as label-free quantification (LFQ), and label-based quantification. The former type implies spectral counting and ion intensity-based quantification, and the label-based strategies such as tandem mass tag (TMT) include metabolic, enzymatic, or chemical labeling strategies[45]. LFQ is experimentally the most straightforward and cost-efficient of the two methods. The major advantage of TMT is the ability to evaluate multiple samples within a single LC–MS/MS run; nowadays even up to 16 samples [46,47].

Data analysis for MS-based proteomics

The output generated by the mass spectrometer will include MS1 and MS2 data. This information is later processed using software that contains a search engine (such as SEQUEST or X!Tandem, FragPipe [48], Andromeda [49], among others) responsible for matching MS/MS spectra to peptide sequences stored in empirical or spectral databases. Here, proteins are inferred using algorithms to assemble the

peptides back to proteins and finally, the quantification at the peptide or protein levels is obtained. To indicate the matching quality between experimental- and theoretical- spectra of peptides in the database search, a score is calculated for each peptides spectrum match (PSM). Several techniques have been created to validate the search results and assigns a FDR for a given threshold. The most common method is based on decoy searching. It reverses all the protein sequences in a protein database and appends the reversed 'decoy' proteins to the target proteins. Then, an estimation based on the frequency of matches to decoy proteins allows one to estimate the specificity of the search [50]. This strategy is based on the premise that the decoy sequences correspond with possible incorrect search results generated by the search engines.

The final output from all these processes is a matrix with a list of proteins and their corresponding abundances per sample, that is filtered using false-discovery rate (FDR) cut-offs. This matrix is used to address the biological questions in the study at hand.

Data analysis to address biological questions

The analysis of the output from MS-based proteomics experiments may involve drawing conclusions from the derived protein list's nature, protein abundance analysis, or comparisons with other studies [50]. No standard workflows exist for the analysis of data generated from proteomics experiments. The proteomics data analysis can be very specific to a particular research area. However, three major steps are typically followed when analyzing high-throughput proteomics data (Figure 5). These are 1) data pre-processing, 2) statistical analysis to select relevant proteins, and 3) functional enrichment analysis [51]. Data pre-processing includes considerations on handling missing values and data normalization to obtain a matrix with comparable and reliable data for downstream analysis.

When performing downstream analysis, one can either generate a new hypothesis out of existing data (data-driven hypothesis), or produce new data for an existing hypothesis (hypothesis-driven analysis). Although both approaches are widely used in clinical proteomics, the first one has emerged as crucial when working with large-scale data (such as LC-MS data). In this context, correlation replaces causation, and complex statistical algorithms are capable of finding patterns not visible to the naked eye. Depending on each case, different methods are used for featuring extraction (i.e., selection of relevant proteins for further analysis). These methods are classified into several categories depending on their mechanisms and purpose. For example, algorithms based on supervised learning (such as sPLS-DA) can be used to uncover relationships between proteins and clinical outcomes or phenotypes. Whereas unsupervised algorithms (such as PCA and hierarchical clustering) uncover naturally occurring patterns or grouping in the data [52].

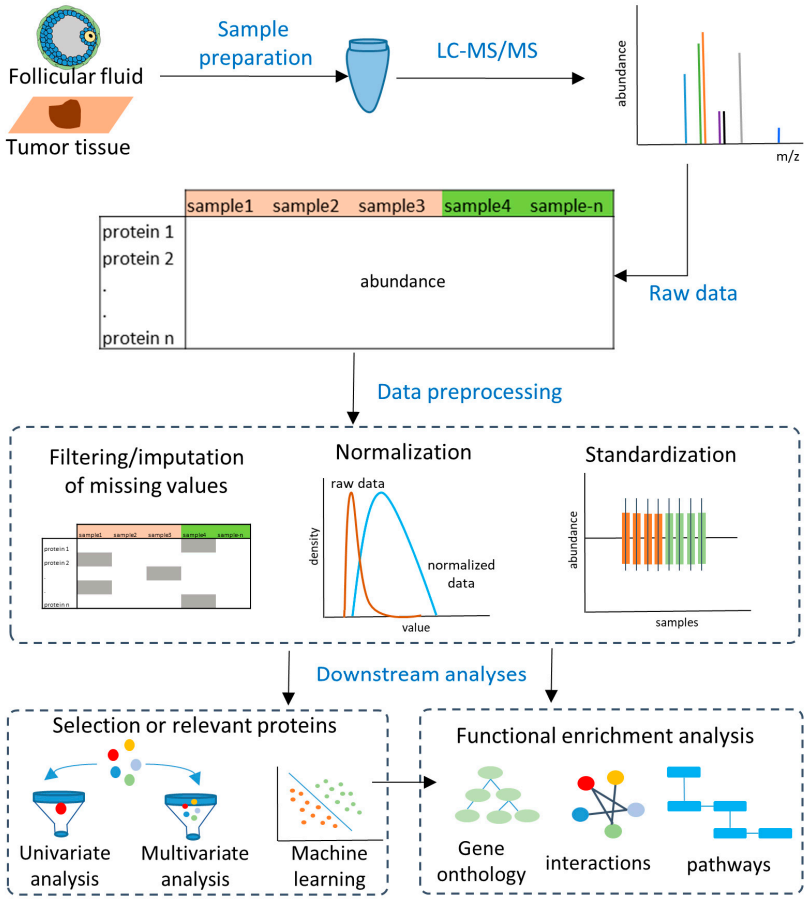


Figure 5. General workflow of bioinformatics analysis in MS-based proteomics

On the other hand, statistical methods can also be classified as univariate or multivariate. The first one, including popular tests such as ANOVA and Student t-Test, analyzes proteins individually. In proteomics data analysis, these previous tests should be followed by tests aimed at controlling the FDR generated for multiple testing. However, the FDR method must be carefully selected because some aim to control the false positive rate and others the false negative rate. On the other hand, multivariate analysis (such as sPLS-DA) considers the expression of all proteins in a matrix simultaneously. These methods are computationally more demanding but also more resistant to errors during protein selection. Other tests based on empirical Bayes approaches are used for the detection of differentially expressed proteins. For example, LIMMA (linear models for microarray data) R package was implemented to account for the realistic distribution of biological variance. This approach was introduced to analyze gene expression data and later

expanded into the field of proteomics. Although this approach can achieve better results, it seems susceptible to missing values commonly obtained in proteomics experiments, such as those based on label-free quantification [53].

The resulting outcomes obtained from the statistical tests are then analyzed based on the functional classification of the proteins and their previously studied relationship. Several bioinformatics tools are available today (such as DAVID [54], FunRich[55], STRING[56], PANTHER [57]) to perform functional enrichment and over-representation analyses that permit information on cellular, molecular functions, biological processes, and pathways in which our proteins of interest are involved. Other algorithms (implemented in GSEA, PSEA, and Perseu tools) are based on computational methods that determine whether an *a-priori-defined* set of genes/proteins statistically shows significant differences between two biological states[58–61]. Overall, many of these analyses are implemented as packages in R free-platform. Two examples are the Bioconductor software [62] for examining and comprehending high-throughput data generated by wet lab experiments and the mixOmics package [63], which perform, among other things, multi-omic integration.

In general, the effectiveness of a data-driven system goes beyond a measure of performance, such as an AUC or a p-value. A data-driven approach is practical when it produces actionable outputs for suitable patients at the right time. For instance, when the output is able to predict relevant information that can help a clinician decide the most effective treatment for a particular patient as soon as a diagnosis is made, we are in the presence of an effective data-driven system. Furthermore, when evaluating the clinical implementation of such a system, it is important to know whether it has been tested in an experimental setting and whether it has shown a meaningful impact in a population similar to the one for which it is being considered. The latest cancer treatments involves an oncological team consideration, where molecular expression data such as proteogenomics data are taken into consideration, for therapy decision-making.

Aims of the thesis

The overall aim of this thesis was to interrogate proteomics data from a bioinformatics and biostatistical point of view to discover insights into the protein profile dynamics of the human ovarian follicular fluid and BRAF mutated metastatic melanoma tissue.

Using different workflows, analyses, and mathematical principles, this thesis aimed to combine biological knowledge with bioinformatics and biostatistical approaches to integrate proteomics, clinical, and histopathological data to expand on biological insights.

The specific aims of this thesis were to:

I. Map dynamic changes in FF protein composition during ovulation in humans using a sensitive quantitative proteomic approach to develop a detailed understanding of the annotated proteins and pathways;

II. Create a detailed fingerprint of proteins present in FF from hSAF to identify from the early follicular stage, candidate proteins that support follicular growth and development;

III. Identify proteomic alterations in the FF of unstimulated hSAF from polycystic ovaries;

IV. Identify folliculogenesis-related functionalities of uncharacterized or poorly characterized proteins identified in the FF of hSAF; and

V. Discover and define the association between protein expression, clinical outcome and tumor phenotypes in MM patients with BRAF mutation.

Material and Methods

Data origin and subjects

Data analyzed in this thesis originated from research projects where FF and MM cancer tumor tissues have been analyzed by MS-based proteomics to get insight into the female ovarian function and MM metastatic tumors. Raw files obtained from label-free (**paper I-IV**) and tandem mass tag (TMT) (**paper V**) quantification were processed using search engine the SEQUEST HT serch engine integrated into Proteome Discovery (Thermo Scientific, San Jose, CA, USA) software. The output data was subjected to data analysis.

To map FF dynamic changes during ovulation (**paper I**), a cohort of 25 women (age: 18-35 years) was selected. These women were previously hormonally stimulated in connection with IVF-ICSI treatment. Using vaginal ultrasound-guided puncture, a FF sample was collected from a large antral follicle (>14 mm of diameter) from each woman. In addition, descriptive clinical parameters and results from hormonal analyses were collected.

To investigate fingerprints of proteins present in FF from SAF (size 6.1±0.4 mm), a first study (**paper II**) included a cohort of 31 unstimulated women (age: ~28 years) undergoing unilateral ovariectomy for fertility preservation whose ovary had a macroscopically normal appearance. Only women with diseases unrelated to the ovary were considered. Proteomics data was obtained in three blocks. Firstly, FF samples from 15 of the 31 women were pooled and evaluated. Next, ten other samples were individually evaluated for protein verification, and ultimately, 13 FF from 6 women were evaluated.

Furthermore, we evaluated FF from SAF (size 4.6–9.8 mm) of polycystic ovaries (**paper III**). Proteomic data from this study was obtained from ten women (5 with PCO and 5 with non-PCO, age: 17-33 years) who donated in total 20 FF samples.

The folliculogenesis-related functionalities of uncharacterized or poorly characterized proteins identified in SAF (**paper IV**) were assessed by analysing FF samples collected from different stages of ovarian antral follicles. Samples were extracted from 50 women (only one sample per woman). FF from small antral follicles (diameter: 5-13 mm) was extracted from 30 women (non-hormone stimulated) undergoing unilateral ovariectomy for fertility preservation. Only patients with diseases unrelated to the ovary were included, and in all cases, the

ovary had a macroscopically normal appearance. FF from large antral follicles (diameter: >15 mm) were extracted from 20 women undergoing assisted reproductive therapy (IVF- or ICSI). Ten out of 20 women were non-hormone stimulated; they were only treated with an ovulation trigger (hCG) to induce ovulation.

To get insight into the protein profile of MM metastatic tumors (**paper V**), 56 patients (24-89 years old, mean = 64, 40 men, 16 women) diagnosed with metastatic MM were evaluated. Only two received targeted B-raf treatment with Vemurafenib. The overall survival was 2.9 ± 3.5 (0.1–17.4) years. The majority of the studied metastatic tissues were from the lymph nodes (82%), while the remainder were cutaneous, subcutaneous and visceral. Four patients younger than 40 years of age at diagnosis were excluded from the analysis since, as described in several studies[64,65], these young patients presented an imbalance towards a much higher overall survival.

Ethical approvals

Informed consent was obtained from all participants included in this thesis. The studies were approved by:
Paper I - The Danish Data Protection Agency and the Scientific Ethical Committee of Region Zealand, Denmark (SJ-530).
Paper II-III-IV - The Ethics committee of the municipalities of Copenhagen and Frederiksberg (H-2-2011-044).
Paper V - The Regional Ethical Committee at Lund University, Southern Sweden (DNR 191/2007, 101/2013 and 2015/266, 2015/618).

Data analysis

Data pre-processing

Output from Proteome Discoverer software (raw data) was filtered based on missing values to work with proteins quantified in at least 70% of the samples of at least one condition. Considering the low number of samples and quantified proteins in **paper I**, missing values were replaced following two approaches for imputing missing values. The first approach imputed values similar to the measured data. This approach assumed that missing values were at random as a result of ion suppression or from the stochastic nature of the DDA methods. The second approach considered that missing values were not at random. It assumed that values were missing due to low abundance of ions (below the instrument's detection limit). In this case, missing values were replaced by simulating low signals, meaning that the new values were biased toward the lower part of the normal distribution of the measured data.

Data pre-processing included a normalization step in which protein intensities were log2 transformed and centred (i.e., standardized) across samples to perform subsequent parametric statistical testing as needed. From **paper I to IV** where label-free quantification was performed, the data standardization consisted of subtracting the sample median from each intensity value, the median of its sample. In the case of **paper V**, a TMT (labeled) approach for protein quantification was applied. To enable comparison across the entire sample set, relative protein abundances were calculated as the ratio between the protein intensity in the sample and the intensity of the protein in the reference.

Selection of relevant proteins

Univariate and multivariate tests were used to select relevant proteins (**Table 1**). p-values from the univariate tests were adjusted to control the FDR generated from multiple testing (i.e., multiple proteins tested or multiple pair-wise comparisons). FDR < 0.05 were accepted. Specifically, in **paper I**, the FDR was controlled using Tukey HSD post-hoc test [66] to control the family error rate provoked by multiple pair-wise comparisons. Tukey HSD is suitable when the sample sizes for each group are equal. The remaining univariate tests were followed by an FDR control based on the Benjamini-Hochberg method [67]. This test calculates the expected proportion of false discoveries amongst the rejected hypotheses (i.e., the proportion of proteins falsely reported as significantly different between groups). Proteins included in the functional enrichment analysis performed in **paper V** were selected based on significance levels of 1% (i.e., p-values <0.01).

Multivariate analyses performed in **paper I** and **paper V** were unsupervised. These allowed us to corroborate the discriminative nature of the differentially expressed proteins detected from the univariate analysis. The multivariate analyses carried out in **papers II** and **III** were supervised (sPLS-DA) to detect discriminative proteins

using methods based on feature selection (LASSO penalization). sPLS-DA is a multivariate analysis that classifies the samples by performing a multivariate regression using the protein expression matrix as predictors and the sampling origin or phenotype as the response. To carry out this analysis, I used the mixOmics R package[63], which includes a LASSO penalization method to select the most informative predictors (e.g., proteins) responsible for discriminating samples.

Table 1. Statistical tests for discovering relevant proteins. *Tests followed by a false discovery rate test.

	Univariate analysis	Multivariate analysis
Paper I	ANCOVA* Tukey HSD post hoc*	PCA Hierarchical clustering
Paper II	Pearson correlation* paired Student t-Test*	sPLS-DA (mixOmics) LASSO penalization Hierarchical clustering
Paper III	Student t-Test*	sPLS-DA (mixOmics) LASSO penalization Hierarchical clustering
Paper IV	Pearson correlation*	-
Paper V	Kaplan–Meier ROC curve Student t-Test	PCA Hierarchical clustering

Functional enrichment analysis

Table 2 shows the different bioinformatics tools used to interpret the results from the statistical analyses. Except for IPA-QIAGEN software, the tools are free (non-commercial) to use by the scientific community. FunRich [55] was mainly used for GO annotations and enrichment analysis to identify altered pathways in paper II. DAVID tool was used to perform functional annotation clustering. January 2022). In this type of analysis, proteins with similar GO/pathway terms are most likely involved in similar biological mechanisms [54,68]. R package clusterProfiler [69,70] was utilized in paper IV to perform GO analysis and significant pathways were assessed by applying the 1D annotation enrichment algorithm proposed by Cox & Mann [60] and available in Perseus platform [71].

Table 2. Bioinformatics tools utilized for functional enrichment analysis

	Gene ontology	Functional annotation clustering	Protein network	Pathways
Paper I	FunRich	-	STRING	-
Paper II	FunRich	DAVID	IPA (QIAGEN)	FunRich
Paper III	FunRich	DAVID	-	1D annotation enrichment (Cox & Mann)
Paper IV	clusterProfiler	-	IPA (QIAGEN)	-
Paper V	IPA (QIAGEN)	-	IPA (QIAGEN)	IPA (QIAGEN)

Data analysis performed in paper V

Data analysis workflow for **paper V,** and its extended analysis is shown in Figure 6**.** From a cohort of 49 patients with BRAF mutation (transcript), the mutated protein B-raf V600E was quantified in 16 patients.

Association between B-raf V600E expression and patient survival

The receiver operator characteristic (ROC) curve was used in **paper V** to detect a cut-off point based on the ability of the expression of B-raf V600E protein to discriminate between patients with less than and more than three years of survival. This variable was used to generate a KM curve which showed an association between high expression of the B-raf V600E mutated protein and significantly reduced overall survival.

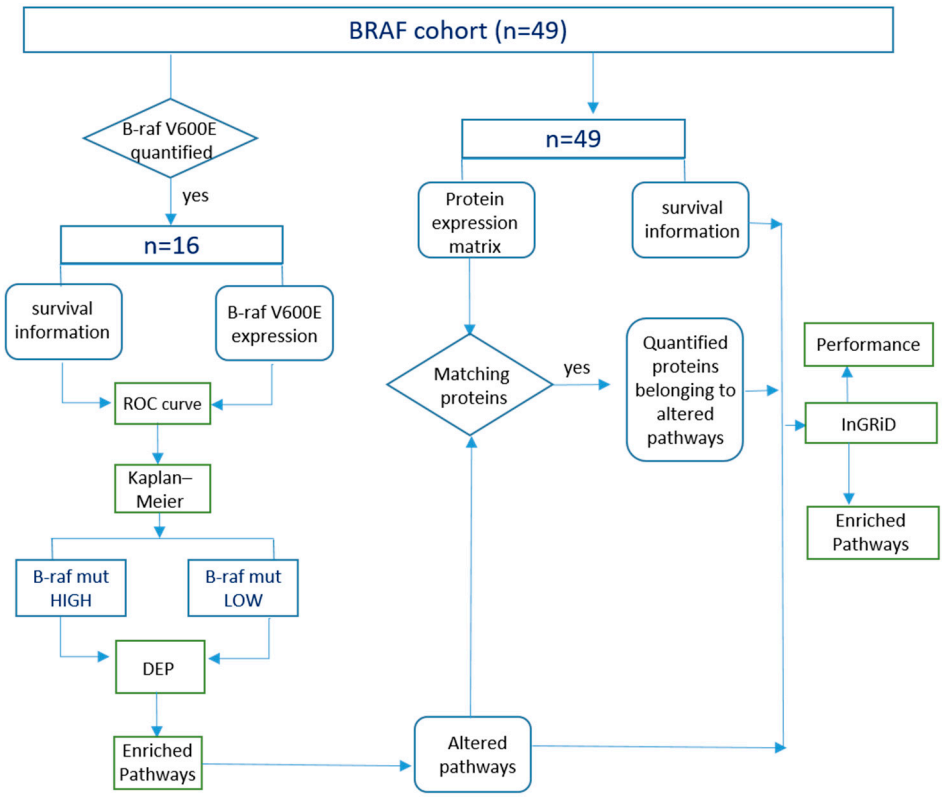


Figure 6. Workflow to determine the association between B-raf V600E expression and patient survival. Green rectangles indicate statistical test

Identification of mortality risk subgroups of BRAF V600E mutated patients

Subgroups of patients with different mortality risk rates were identified using an R package called ‘InGRiD’ [72]. This package provides a pathway-guided identification of patient subgroups based on protein expression while utilizing patient survival information as the outcome variable. Proteins belonging to pathways that emerged as altered between the two groups of patients with different levels of BRAF mutation were selected, and their expression was used in the analysis. The survival information was the patient survival time from sample collection to death or censoring. All default parameters of 'InGRiD' were kept.

Results and Discussion

The methodology for proteomic data analysis approached in this thesis was applied to two different fields of translational medicine: women's reproductive disorders and malignant melanoma. For each field, different cohorts of patients were analyzed to characterize the proteomic profile of the ovarian follicular fluid and the tumor tissue of patients with BRAF mutation to obtain new biological insights that help to alleviate these medical conditions.

In this section, I present a general workflow established for proteomic data analysis and a summary of the results presented in the scientific articles included in this thesis. A complete presentation of the results can be found in the original papers attached at the end of this thesis. In addition, I present a brief discussion of the generated results.

General workflow for data analysis

Clinical proteomics research is based on data generated from analysis of patient samples and other types of clinical information (e.g., age, sex, disease stage). In some cases, the clinic can provide relevant data collected from clinical chemistry analyses and/or histopathology, the latter being common in cancer research. The integration of these different types of information is crucial in order to obtain relevant and meaningful results.

In this thesis, a global adaptable workflow was established to perform the clinical proteomic data analysis (Figure 7). The workflow is composed of three major parts that were adapted to each study (i.e., paper) presented in this thesis. The first part (blue color) is performed prior to the real data analysis. Here we meet with all parties involved in the project to be aware of how much data we can collect from the clinic, what is the clinical/biological problem and/or hypothesis and which patient cohort will be included in the study.

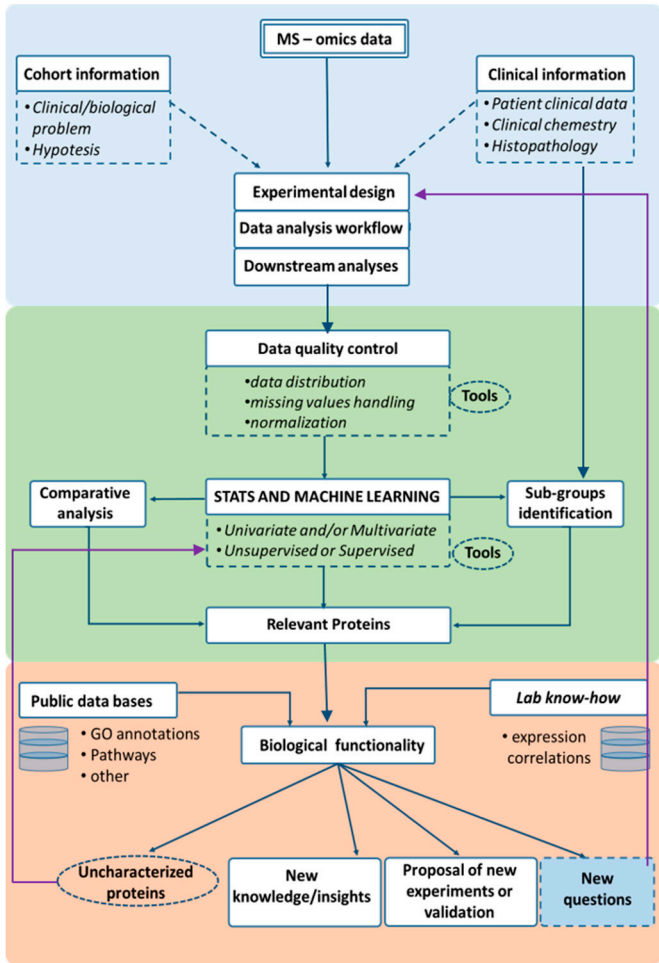


Figure 7. General workflow for clinical proteomic data analysis

This first part generates an experimental design and a data analysis workflow adapted for the specific study. The second part (green color) of the overall workflow starts with data pre-processing, where quality control and data cleaning are performed. Relevant features (e.g., proteins) or patterns are detected by applying biostatistical tests and/or machine learning machine techniques in the second part. This part could include the combination of several statistical strategies and tools. The third part of the workflow aims to investigate the biological functionalities of the detected relevant features or patterns. This part may include the combination of several bioinformatic analyses procedures using public databases and/or laboratory know-how. We proposed four major outputs derived from this workflow. The data analysis could provide 1) the discovery of new knowledge/insights, 2) the proposal of new experiments, both for discovery and/or validation of the results, 3) new

questions derived from unexpected results, or 4) the identification of features with unknown functions which will require a deeper analysis to obtain biological insights.

Throughout the workflow, several challenges may emerge. For example, choosing the most appropriate statistical test to perform a comparative analysis can be challenging. Sometimes we have to decide beforehand whether the data should be transformed or not to a specific probabilistic distribution (e.g., normal distribution), which in turn will affect the decision on which test to choose. On the other hand, the generated data may not exactly fit into the designed pipeline. This is the case with some pipelines published as packages on R or Python platforms, that sometimes are tested on data generated from genomic experiments. In order to apply the proposed scientific methodology to your data (e.g., proteomic data), one needs to adapt the pipeline to consider a different type of data. The challenge here is not only the adaptation, but making sure that your changes reveal reliable results. At this point, the researcher not only needs computational skills to adapt the code and handle the data, there is also a need to understand the nature of the data, but also the analytical technique from which it was generated, and the problem that needs to be overcome. Another difficulty, not often discussed but very important, is the representation of the results. How to represent high-throughput data is cumbersome. The results should be illustrated in a figure that is easy to understand not only by researchers working in the field of molecular biology but also by non-expert collaborators, like clinicians, collaborators from other disciplines, and sponsors.

Hereafter, I will be presenting new biological insights, which were obtained after applying different workflows to the analysis that combine biological information with bioinformatics and biostatistical solutions.

The proteome of the ovarian follicular fluid

Proteomic changes during ovulation

Ovulation constitutes the final step of folliculogenesis and is crucial in order to ensure women's reproductive health. In an attempt to investigate the protein profile dynamics of the ovarian FF during folliculogenesis, we first focused on what happens during the last stage of ovulation (**Paper I**). The study was designed (Figure 8) to assess the FF protein profile dynamics at five different time points (before ovulation induction (OI) and at 12h, 17h, 32h and 36h respectively after OI).

A comparative analysis performed beforehand with clinical and endocrine parameters (**paper I**) concluded that the size of the follicles aspired at 32h was significantly larger (adjusted p-value < 0.05, Kruskal-Wallis test followed by

Bonferroni *post hoc* test). Therefore, 'follicular size' was included in the ANCOVA test as a covariable.

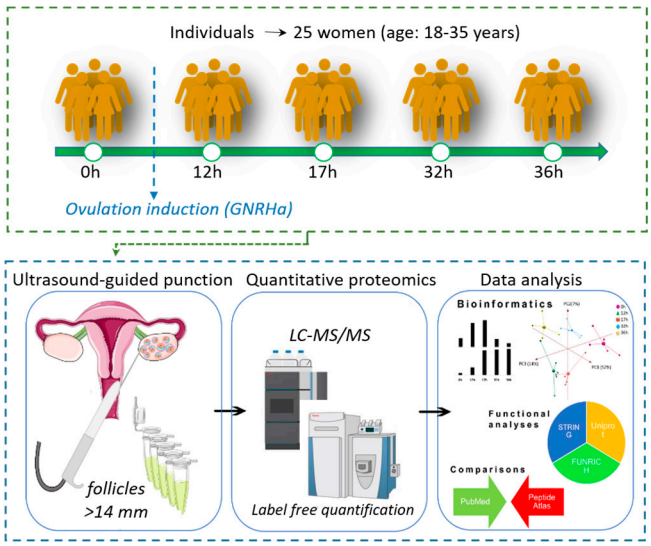


Figure 8. Study design to assess the protein profile dynamics of the ovarian FF during ovulation

Quantitative large-scale proteomics allowed the identification of 400 proteins of which 342 were detected at all time points. As determined by an univariate (ANCOVA followed by post hoc Tukey HSD) analysis, 24 proteins changed significantly (FDR less than 5%) at least at one time point during ovulation. Out of these 24 proteins, 15 were most likely considered to be secreted by follicle cells since they were only present in the FF when compared to a high-confidence human proteome plasma (HPP) library constructed from 91 LC-MS/MS datasets [73]. A second approach based on missing values highlighted 16 additional proteins as turned on or shut off at least at one time point during ovulation (referred to as 'on-off proteins'). Multivariate analyses based on PCA and unsupervised hierarchical clustering revealed a distinguished difference throughout the time course, observing that the major changes occurred immediately after OI (Figure 9a). The heat map and dendrogram generated by the hierarchical clustering manifested different patterns of protein expression across the studied time points.

According to the gene ontology categorization, the altered proteins, which most likely represent the contribution of GC and TC to FF, are mainly involved in biological processes such as cell growth and/or maintenance and cell communication (Figure 9a). The 15 proteins in FF (superior in FF compared to plasma) were mainly matrix stabilizers, inflammatory factors and growth factors. The combination of bioinformatics tools for protein-protein interaction and functional enrichment analyses revealed an enrichment of extracellular proteins

with molecular functions such as protein binding, receptor binding and growth factor activity (Figure 9b). In addition, these were proteins involved in biological processes related to inflammatory immune functions, secretion and extracellular structural organization. Furthermore, based on known and predicted interactions [74], the altered proteins conformed to a network where a protease system (conformed by metalloproteinase and plasminogen) interacts with inflammatory factors and the insulin-like growth factors system.

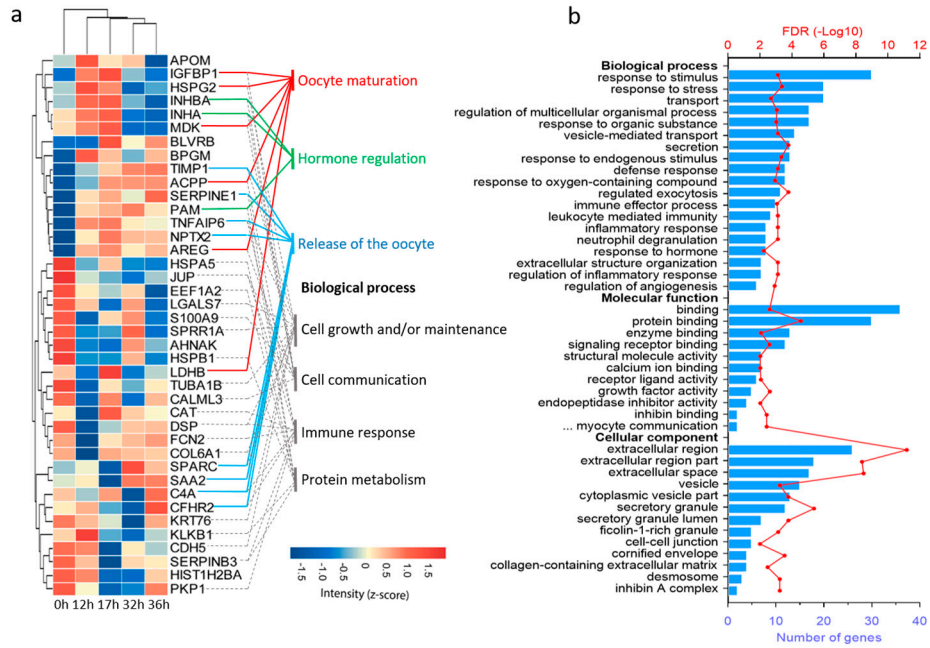


Figure 9. Hierarchical clustering of altered proteins during ovulation and their functionality. **a)** Heatmap and dendrogram of 40 altered proteins (24 differentially expressed and 16 on-off proteins). The right panel shows a summary of the functionality of some dysregulated proteins. This figure combines results shown in Figure 2b, Figure 4 and Table 4 of **paper I**. **b)** Functional enrichment analysis performed with the altered proteins. Only the top GO terms are displayed. This figure illustrates the results shown in Table 3 of paper I.

Overall, we were able to conclude that some proteins were associated to well-known processes that occur during ovulation (Figure 9a). The first one was 'oocyte maturation' involving proteins that increased their expression right after OI, reaching a significant peak mostly at 17 hours (AREG, IGFBP1, MDK, HSPG2, LDHB, ACPP). AREG transduces the LH signal from GC to the oocyte, which then resumes meiosis [75]. IGFBP1 may be secreted by GC [76,77] to dampen the effect of IGF proteins acting in the FF to drive ovulatory-related changes (e.g., steroidogenesis). MDK and HSPG2 proteins were suggested to have a common purpose. The growth factor binding properties of HSPG2 may play a role during ovulation by retaining the oocyte maturation effect of MDK inside the follicle. Interestingly, MDK expression decreased after 17h.

The second process we observed was 'follicle hormone regulation' involving proteins inhibin alpha chain (INHA), inhibin beta A chain (INHBA), which peaked at 12h-17h. It also included the peptidyl-glycine alpha-amidating monooxygenase (PAM) peaking at 17h-32h. Although the midovulatory peak (12h-17h) of INBA and INHBA has not been reported previously, we hypothesized that it might serve to terminate the FSH secretion, to induce oocyte maturation or to briefly intensify the androgen production. Lastly, the 'release of the oocyte' process was linked to proteins which mostly peaked (increased) at 36h (SERPINE1, TIM1, SPARC (32h-36h), NPTX2, C4A, CFHR2, SAA2) and protein TNFAIP6 peaking at 12h-17h. TNFAIP6 is a matrix stabilizer protein that, together with NPTX2 seems to be involved in cumulus expansion [78]. The remained proteins were linked to inflammatory-like processes involved in the follicle rupture and coagulation.

Follicular fluid from small antral follicles. Proteomic characterization

Historically, the proteomic study of ovarian FF has been conducted on FF from pre-ovulatory large antral follicles. This is because the most accessible way to obtain these samples is during assisted reproductive procedures. At this follicular stage, the FF contains a high number of plasma constituents transferred through the follicular basal membrane (due to a follicular expansion), which attenuates the detection of low abundant proteins and, therefore, the number of identifications fluctuates around hundreds of proteins. The second proteomics study (**paper II**) presented in this thesis was thought following the hypothesis that by using mass spectrometry-based proteomics, a wider range of functional proteins could be detected in FF from small antral follicles compared to large follicles. As a consequence, in **paper II** we reported the first large-scale proteomic characterization of the FF from human SAF (hSAF) (size 6.1±0.4 mm).

The characterization was based on 2461 identified proteins, out of which 1108 were found for the first time in human FF (Figure 10a). This constituted the largest number of proteins reported to date in human FF. 94% of the proteins were previously found at the transcript level in GC [14] and 39% were identified at the protein level in oocytes [16], which leads us to presume that those were possibly secreted from GCs and the oocyte.

According to the GO classification, 19% were extracellular proteins, while 38% were cytosolic and nuclear proteins. The presence of intracellular proteins in the follicular fluid may be due to cellular apoptosis of follicular cells, which usually occurs during follicular development [79,80]. A high number of proteins with catalytic activity (36%) were identified, such as hydrolases (16%), transferases (5%), oxidoreductase (4%), and enzyme modulators (11%). On the other hand, we found that 43% of the proteins had a binding molecular function. These proteins included nucleic acid binding proteins (12%), signaling molecules (8%), receptors (4%), and calcium-binding proteins.

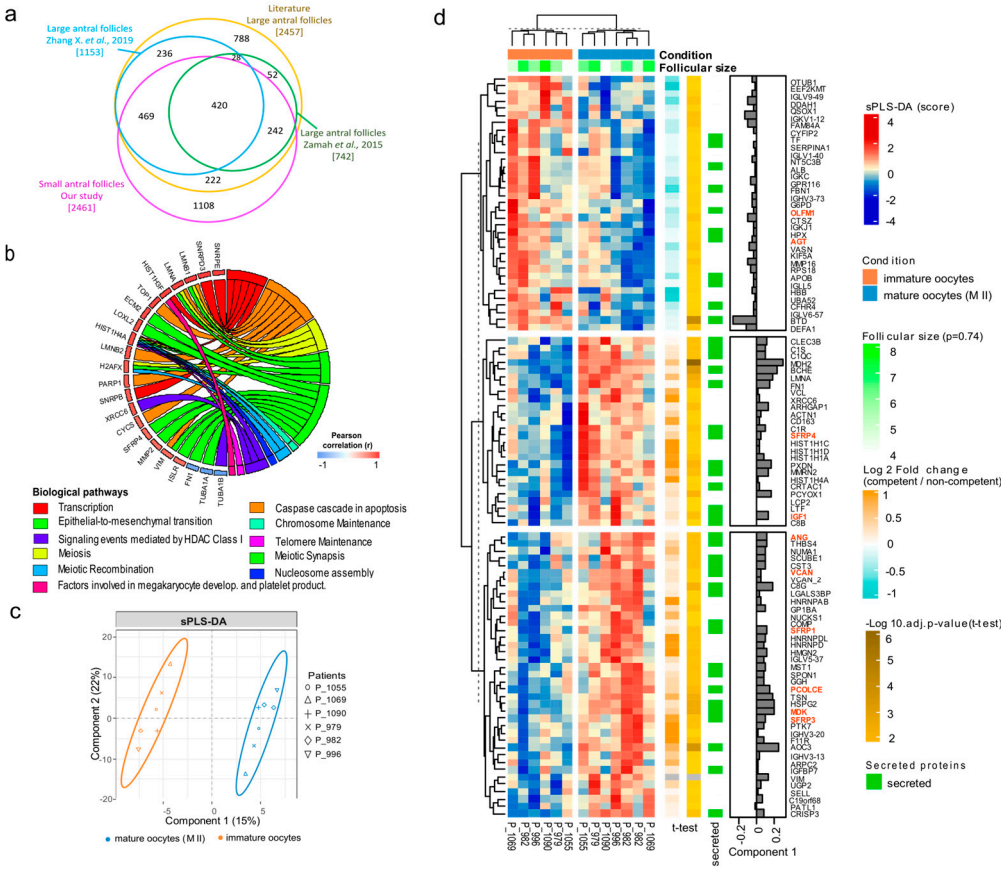


Figure 10. Proteins identified in human follicular fluid (FF) studies. a) Venn diagram comparing the number of proteins identified in our study and proteins previously identified in FF. The yellow circle ('Literature') denotes proteins identified in previous FF studies (up to 2020). The circles in light blue and green colour, represent the two proteomics studies that had previously identified the largest number of proteins in FF using mass spectrometry. Magenta color represents the number of proteins identified in FF from unstimulated small antral follicles (paper II). b) Biological pathways significantly enriched (BH method: adjusted P-value <0.05) by proteins correlated positively or negatively with MDK and VIM. **c)** Sparse partial squares discriminant analysis performed with 750 proteins quantified in 13 paired FF samples extracted from small antral follicles coming from six women. The analysis discriminated between FF surrounding oocytes capable of achieving metaphase II (MII) after IVM (n=7, blue) and FF surrounding oocytes unable to mature (n=6, orange) after in vitro maturation. **d)** Top 100 proteins that contributed to Component 1 of sPLS-DA to discriminate between FF samples. Positive and negative sPLS-DA scores mean that the protein is up- and down-regulated in FF surrounding oocytes capable of reaching M2, respectively. The bar chart indicates each protein's contribution in Component 1 (sPLS-DA) to discriminate between groups.

The workflow designed to identify high abundant proteins possibly more accessible by MS in FF from SAF as compared to the high abundant proteome of FF from large follicles highlighted a list of 24 non-plasma proteins, of which four were secreted proteins; AMH, MDK, HTRA1 and LOXL2. This comparison also confirmed the superiority (i.e., more abundant) of candidate proteins AREG, TNFAIP6, SERPINE1, and ACPP (ovulatory-related proteins) in large antral follicles. The presence of the 24 proteins in FF from SAF was verified by data-dependent

acquisition (DDA) and/or parallel reaction monitoring (PRM) in 23 different samples collected from 16 women. Specifically, AMH, MDK, and vimentin (VIM) were grouped (according to the functional annotation clustering) in a cluster of proteins involved in ovarian follicle development.

A further analysis based on protein expression correlations and functional enrichment analysis suggested that specifically MDK and the protein pair MDK/VIM might play a fundamental role in follicle development and oocyte maturation already from early antral follicles. Expression correlation is known as an indication of a functional association between genes and proteins [81]. Proteins positively correlated ($r \geq |0.7|$, adjusted p-value < 0.02) to MDK and VIM (evaluated in a subgroup of 10 individual FF samples) belong to gene regulation pathways such as transcription, chromosome maintenance, and meiosis and may act as part of the chromosomal organization of GC that supports the progression of follicular growth and maturation (Figure 10b). In addition, these proteins enriched the epithelial-to-mesenchymal transition (EMT) pathway, which is well known to play a crucial role during folliculogenesis (Kim et al. 2014).

The association of MDK and VIM with subsequent oocyte maturation was further confirmed in an additional cohort of six women. From each woman, we collected at least two FF samples: One FF sample from a follicle containing an oocyte that was capable to mature to metaphase II (MII) after IVM, and a FF sample from another follicle containing an oocyte incapable of maturing to MII. After a discriminative multivariate analysis based on the sPLS-DA method, a total of 100 proteins, including MDK, VIM and IGF1, were dysregulated in FF from hSAF surrounding oocytes capable of maturing (Figure 10c,d). A functional annotation clustering made with these proteins grouped nine secreted proteins (SFRP1, SFRP4, FRZB, MDK, AGT, PCOLCE, ANG, OLFM1, VCAN) in a cluster of EMT up-regulated processes such as development, growth factors and Wnt signal. Furthermore, these cell-secreted proteins correlated with proteins involved in transcription, signaling by NOTCH and EMT pathways.

Altogether, these results provide evidence that a broader range of functional proteins can be found in FF from SAF. The experimental design combined with the methodology applied for data analysis revealed that changes at the protein level occur already in FF from small antral follicles related to subsequent oocyte maturation. We demonstrated that the ability of the enclosed oocyte to sustain meiotic resumption could be predicted from SAF.

Follicular fluid proteomic alterations linked to polycystic ovaries. Small antral follicles

A polycystic ovary (PCO) is characterized by having antral follicles that arrest at a size of 3 to 11 mm in diameter. This fact affects the selection of the dominant follicle and therefore the subsequent ovulatory process. In the previous paper (paper II), we demonstrated that the ability of the oocyte to sustain meiotic resumption can be detected at a non-selected stage of the follicle (i.e., small antral follicles). This led us to a follow-up where we aimed to detect possible proteomic alterations occurring in the FF of unstimulated SAF from PCO associated with a disruption of folliculogenesis.

Paper III constitutes a pilot study (20 FF samples from 10 women were evaluated) that describes the first large-scale proteomics study performed in FF of SAF (4.6-9.8 mm), collected from unstimulated PCO. Here, we performed a multivariate analysis based on the sPLS-DA method to discriminate the protein profile of FF samples (n=10) collected from five PCOs from the protein profile of FF samples (n=10) collected from five normal ovaries (i.e., non-PCO). The analysis was performed based on the log2 protein expression of 850 proteins quantified in at least 70% of the samples of one of the two conditions. The LASSO penalization included in the sPLS-DA method for feature selection detected 115 dysregulated proteins. Furthermore, unsupervised clustering of the samples confirmed the discriminative nature of these proteins.

The bioinformatics strategy carried out to investigate the functionality of the dysregulated proteins in FF from PCO samples included a GO overrepresentation analysis, a pathway enrichment analysis (1D annotation enrichment analysis) and functional annotation clustering of the secreted proteins. Results obtained from the comparative analysis showed alterations at the protein level in the FF of PCO related to the immune and inflammatory systems, extracellular matrix (ECM) receptor interaction, collagens-containing the extracellular matrix, regulation of signaling, response to oxidative stress and growth factor/hormone activities. Significantly dysregulated proteins involved in these processes are depicted in Figure 11.

We suggested that an altered cell signaling and communication is present in SAF's FF of PCO that interrupts the crosstalk among paracrine follicular cells. On the other hand, increased immune and inflammatory processes were observed in PCO samples. This may be related to an activation of the pro-inflammatory nuclear factor-kappa B (NF-kB) signaling pathway mediated by inflammatory cytokines derived from the peripheral circulation that enter into the follicles through the ovarian circulation system to activate the NF-kB factor in FF [82,83]. In addition, an increased GC inflammatory cascade provokes mitochondrial damage [84], which exacerbates the generation of reactive oxygen species (ROS)-induced oxidative stress and, thereby, leads to a reduction of cell proliferation, ultimately affecting the growth and development of oocytes. Specifically, secreted proteins SFRP1, THBS4,

and C1QC significantly decreased their expression in PCO FF. This downregulation was associated with impaired future oocyte competence since these proteins were found in **paper II** as down-regulated in SAF's FF surrounding oocytes incapable of maturing to MII.

Overall, the data analysis carried out in this study allowed us to identify proteomics alterations occurring in the FF of PCO hSAF that may be related to the dysfunction of follicular growth and subsequent oocyte competence, finally affecting the selection of the dominant follicle.

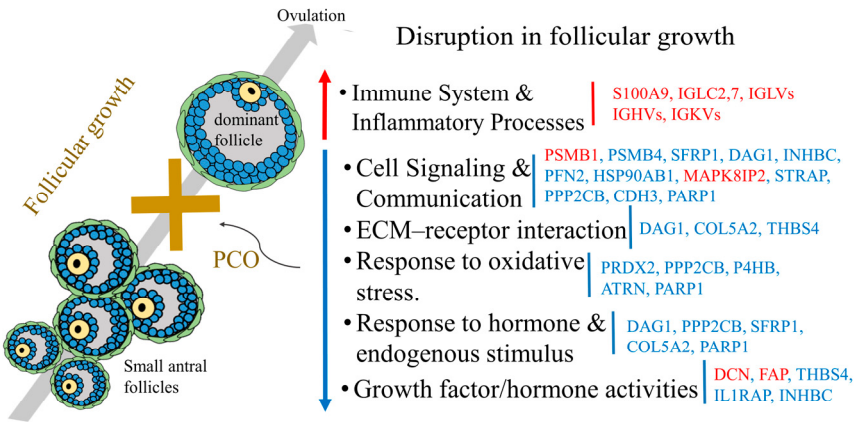


Figure 11. Follicular fluid proteomics alterations are associated with a disruption in follicular growth of polycystic ovaries (PCO). Red and blue colors in arrows and letters represent upregulation and downregulation in PCO respectively.

Folliculogenesis-related uncharacterized proteins.

Disruptions in the folliculogenesis process may lead to medical conditions related to female infertility. The dynamic molecular changes that occur in the follicular fluid (FF) must be well coordinated to ensure correct folliculogenesis. Almost nothing is known about the function of uncharacterized/hypothetical proteins acting in the FF of antral follicles. In this study, we aim to discover folliculogenesis-associated uncharacterized or poorly characterized proteins present in the FF of antral follicles. Figure 12, shows the study design. Using mass spectrometry (MS)-based proteomics, We evaluated FF samples collected from different sizes (5-13 mm) of unstimulated small antral follicles of 30 women and two pools of FF samples collected from large antral follicles (>15 mm) of 10 hormonal stimulated women or 10 unstimulated women which were compared with a pool of 15 FF from unstimulated SAF.

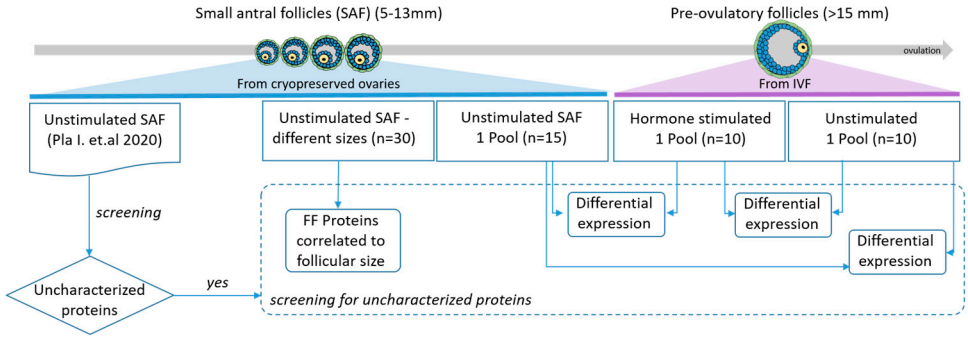


Figure 12. Study design. Uncharacterized or poorly characterized proteins were screened from a previous proteomics study performed in follicular fluid (FF) from unstimulated small antral follicles (SAF). Their behaviour in different follicular stages was evaluated after performing mass spectrometry (MS)-based proteomics using FF from three new different cohorts.

A total of 1641 proteins were identified and 107 were significantly correlated to follicular size. After screening for proteins with no functional domains/regions assigned or unknown functions and proteins without or with very generic GO annotations, fourteen proteins uniquely identified in our previous study (**paper II**) were selected as functionally-uncharacterized or poorly-characterized. Six additional proteins were also selected, taking their novelty in FF into account despite being functionally characterized up to a certain point. The six proteins are adhesion G-protein coupled receptor G6 (ADGRG6), adhesion G-protein coupled receptor G2 (ADGRG2), adhesion G-protein coupled receptor F5 (ADGRF5), extracellular serine/threonine-protein kinase (FAM20C), putative kinase FAM198B and sushi repeat-containing protein SRPX (SRPX). Some of these proteins were associated with follicle-stimulating hormone and follicular growth (Figure 13**)** and fourteen (70%) out of 20 (uncharacterized or poorly characterized proteins) were verified by PRM in ten individual samples.

To dig into the functionality of these proteins, the global proteome detected in each cohort was used to perform bioinformatics and statistical analyses. A total of 107 proteins were found to be significantly correlated to follicular size (adj.$p<0.05$, $|r| > 0.5$). Twenty-eight proteins were positively correlated to follicular size (e.i. increased their expression while the follicular size was higher), and 79 proteins were negatively correlated to follicular size. Enrichment of the collagen-containing extracellular matrix was observed. The ECM provides structural support to the follicle, maintains cellular organization and connectivity, and provides biochemical signals that promote follicle development and maturation [85]. Specifically, the ECM collagen content decreases as the follicle develops [86,87]. Three uncharacterized proteins appeared to correlate to follicular growth (FAM3C, AN32B, STIP1). Four of these proteins decreased their expression in large follicles (ANP32A, ANP32B, STIP1, GPR116), behaving like AMH and INHBB; three increased together with INHBA and INHA in FF from large follicles (SRPX,

FAM3C, C1orf123). Taken altogether, the selected, poorly-characterized proteins detected in the current follicular proteomic study provide an opportunity to delineate novel markers to dissect the status of follicular regulatory processes.

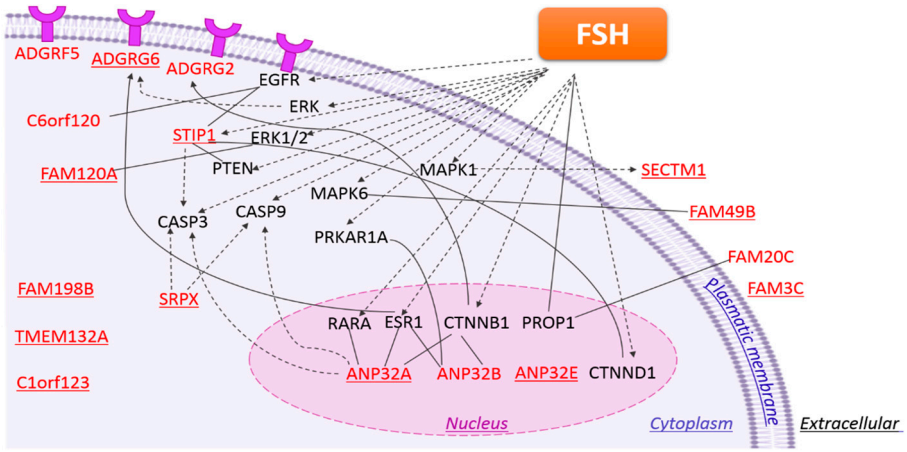


Figure 13. Functional relationships involving the selected uncharacterized/poorly-characterized FF proteins (in red) and follicle-stimulating hormone (FSH), based on IPA analysis. Included are proteins that link FSH to this protein set. Subcellular locations are indicated. Underlined proteins were validated by PRM. Continuous lines indicate a direct interaction, while dashed lines represent indirect relationships and lines without arrows indicate binding events only.

※

Until here, I have summarized the results of the study where proteomic data analysis was applied to get insights into the ovarian follicular fluid. From here, I will be presenting the results of the proteomic data analysis applied in the field of skin cancer, Malignant Melanoma.

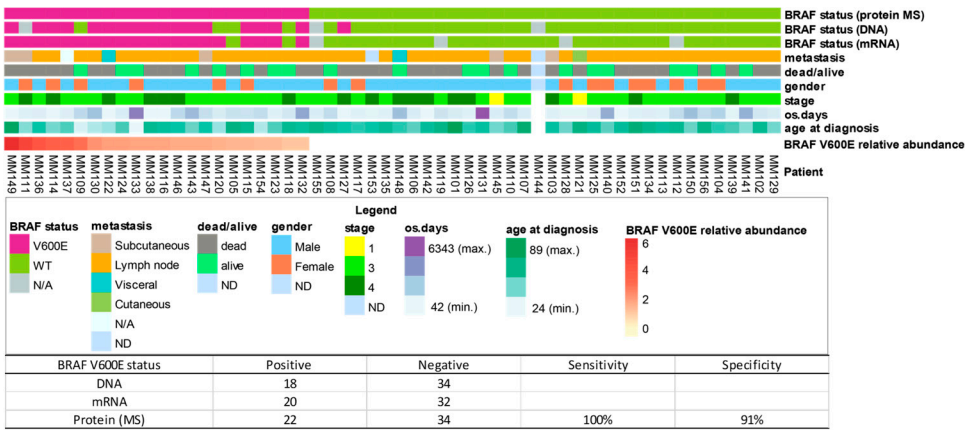
※

Malignant melanoma – Patients with BRAF mutation

Malignant Melanoma is the most frequently mutated tumor type. Around 50% of the cases harbor activating BRAF mutations [27]. In these cases, a targeted inhibitor is given as treatment; however, these patients develop therapy resistance after a while. This indicates that much remains to be understood about what happens at the molecular level in this type of patient.

We performed TMT quantitative proteomics to study the BRAF mutational status of 56 MM metastatic tumor samples (**paper V**). More than 12000 proteins were quantified, including the mutated B-raf V600E protein, which was quantified for the first time in 22 MM tumor samples. The proteomic BRAF mutational status was confirmed by genomics in 50 samples previously analyzed at DNA and mRNA levels. Specifically, 20 of the 22 B-raf V600E positive tumors found by proteomics were confirmed by genomics (Figure 14), while the negative cases were all confirmed. This indicated that the applied MS strategy could truly detect the BRAF V600E mutational status with a sensitivity of 100 % and a specificity of 91%.

Interestingly, the relative abundance of the B-raf V600E protein revealed a high degree of variability (CV=57%) across the samples (Figure 14, red color). This led us to analyze the correlation between different expression levels of the B-raf V600E protein and patient survival.



BRAF V600E status	Positive	Negative	Sensitivity	Specificity
DNA	18	34		
mRNA	20	32		
Protein (MS)	22	34	100%	91%

Figure 14. Patient clinical data and BRAF status for metastatic melanoma. Heat map representation of patient clinical data and BRAF determination by genomic and proteomics techniques.

Association between B-raf V600E expression, immune system and patient survival.

Based on the relative expression of the protein B-raf V600E and the overall survival information of sixteen patients, we did set up a data analysis strategy to understand the effect of this protein on patient survival.

Patient's metastatic tumors were categorized as having low and high B-raf V600E mutation. This was done by applying a ROC curve to detect a cut-off point based on the ability of the expression of B-raf V600E protein to discriminate between patients with less than, and more than three years of survival. This variable was used to generate a KM curve which showed an association between high expression of the B-raf V600E mutated protein and significantly reduced survival (Figure 15a). The histological images of mutation-positive metastatic melanoma samples confirmed the differences between these two groups of tumors by observing different morphological patterns (Figure 15b). The global differential proteome between patients with low and high B-raf V600E expression was assessed and 697 differential expressed proteins emerged (Student t-test, $p<0.01$) (Figure 15c). Overall, high expression of the B-raf V600E protein was associated with more aggressive tumor progression and shorter survival. After a functional enrichment analysis, we associated this event with the downregulation of proteins involved in immune response pathways, while proteins belonging to pathways associated with a proliferative nature of the tumor were upregulated.

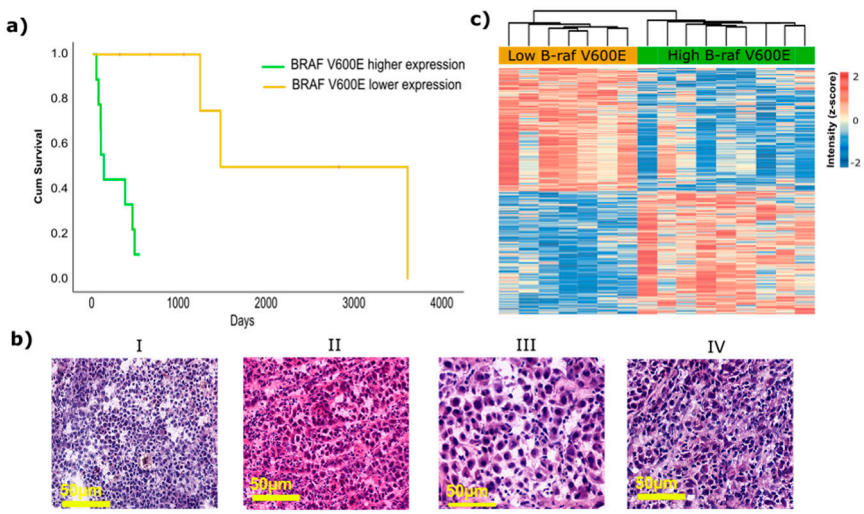


Figure 15. B-raf V600E expression correlated with patient survival and tumor phenotype. a) Overall survival (OS) of malignant melanoma patients according to B-raf V600E mutation levels (log-rank p = 0.001, Breslow p = 0.002 and Tarone Ware p = 0.001). b) Histological images of mutation-positive metastatic melanoma samples: (I and II) tumors with high expression of the B-raf V600E mutated protein; and tumor III and IV are tumors with low expression of the B-raf V600E mutated protein. For all images the magnification and scale were 10x and 50 µm, respectively. c) Hierarchical clustering and heat map of 697 differentially expressed proteins between the two groups of mutation-positive metastatic melanomas.

Furthermore, we extended the analysis to find subgroups with different mortality risks within a cohort of 49 patients with BRAF mutation (transcript). Considering that the B-raf V600E protein variant was not quantified in 33 out of 49 tumors, we could not repeat the same strategy. Instead, we used the list of pathways that emerged as actives in the previous analysis and performed a pathway-guided identification of patient subgroups based on the expression of the proteins that belong to these pathways while utilizing patient survival information as the outcome variable. As a result, patients were classified into low, medium and high mortality risk groups (median survival was 5.1, 2.3 and 0.5 years, respectively) (Figure 16a). Notably, 8 out of 9 of the patients with high levels of B-raf V600E mutated protein were in the high and medium risk groups (Figure 16b). In contrast, most of the patients with low levels of B-raf V600E mutated protein had a low mortality risk (4 out of 5). This indicated that the mortality risk classification might be strongly associated with the expression level of the B-raf V600E protein variant. In other words, patients with high levels of B-raf V600E seem to have a higher risk of mortality. Overall, proteins involved in the classification of subgroups of patients were mostly enriched in pathways related to metabolism, immune system and signal transduction (Figure 16c). Specifically, the observation of the immune system pathways in this extension of the study corroborated that an imbalance of the immune system in MM patients with BRAF mutation may induce an aggressive tumor progression and shorter survival. Although proteins that enriched these pathways were not only downregulated proteins, as observed in the previous study, but also key regulating proteins such as proteasome activator complex appeared to decrease in the group of high mortality risk. This complex is the main degradation system for oxidatively damaged proteins and low expression of these proteins has been associated with short survival of cancer patients (www.proteinatlas.org).

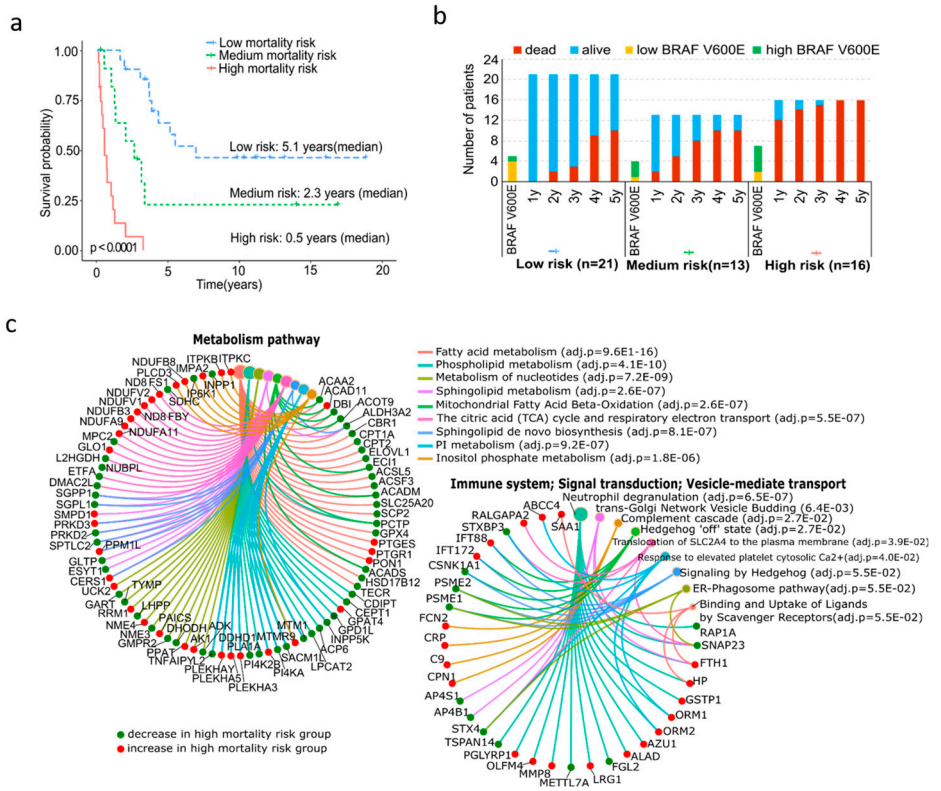


Figure 16. Identification of subgroups of patients with BRAF mutation. **a)** Kaplan-Meier curves of the molecularly-defined subgroups of BRAF mutated patients. Survival probabilities color subgroups: red (high risk of mortality, n=16), green (medium risk, n=12) and blue (low risk, n=21). Median survival times for the three groups is shown. **b)** Distribution of BRAF patients dead or alive after 1 to 5 years from sample collection. On x-axis, patient distribution is stratified according to mortality risk classification. The number of patients in whom the levels of mutated B-raf V600E protein could be quantified is indicated in yellow (low) and green (high). **c)** Proteins from the most significant pathways enriched in mortality risk subgroups of patients with BRAF mutation.(Figure also published on thesis ISBN: 978-91-7895-919-8 (pdf)).

Conclusions and Future Perspectives

Data science has emerged as an indispensable discipline in most fields of life science research, especially in the field of translational medicine. The integration of proteomics and clinical data carried out in this thesis allowed us to obtain new information that may be of importance for future advances in reproductive medicine and an eventual delineation of patient responders/non-responders to therapy for malignant melanoma with BRAF mutation.

Results from **paper I** confirmed that dynamic changes occur in the protein composition of the ovarian FF during ovulation. Changes occur right after ovulation induction from 12 h until 36 h to impact processes such as oocyte maturation, follicle hormonal regulation and release of the oocyte. Proteins involved in these processes could be screened after ovulation induction in women undergoing IVF therapy to predict a successful outcome of the treatment. However, there is a long way to go to get to this point. First, changes in these proteins during ovulation must be assessed in a larger cohort of patients and further analyses should be done to validate the results in the clinic. Considering the difficulty of obtaining homogeneous groups of patients, the data analysis strategy should be designed so that models can be adjusted/corrected to avoid the effect of confounding variables.

Paper II-IV demonstrated that by doing MS, a larger number of functional proteins can be identified in FF from SAF compared to large follicles. FF from large follicles has a high dynamic range, which attenuates the ability to detect low abundant functional proteins. The study described in **paper II** revealed that significant differences exist in the FF from SAF, predicting the ability of the enclosed oocyte to sustain meiotic resumption. If this can be confirmed in further studies, it demonstrates that the viability of the oocyte is determined early on in follicular development and may open up new pathways for augmenting or attenuating subsequent oocyte viability in the pre-ovulatory follicle when ready to undergo ovulation. The direct impact of protein midkine (MDK) on oocytes *in-vitro* maturation is being tested in the clinic. Preliminary results confirming an increase in the rate of oocyte maturation by adding MDK to the culture medium have already been patented by our group. The final goal is to augment the number of good-quality oocytes available for IVF treatment.

Paper III revealed that proteomics alterations related to PCOS occur already at the non-selected stage of the follicle to affect the follicular growth and subsequent

oocyte competence. This study proved that signatures of proteins indicative of women's health status can be identified in the FF of SAF. Following this concept, we are currently testing FF from SAF to discriminate between ovarian cancer patients and other types of cancer.

The study described in **paper IV** revealed a novel association between FSH and uncharacterized or poorly characterized proteins. Furthermore, this study captured the proteomic dynamics of the FF associated with follicular growth during the middle follicular phase. In order to have a broader picture of the system involved in folliculogenesis, we planned a joint proteomics evaluation of FF, GC and the oocyte collected from the same follicle.

Paper V provided evidence that a higher expression level of the B-raf V600E mutated protein is associated with a more aggressive tumor progression in MM patients with BRAF mutation. This event may be linked to an imbalance in the immune system of these patients. The analysis of the B-raf V600E protein, and associated proteins, together with a histopathological characterization of the tumor isolated from MM patients with BRAF mutation, may enable the delineation of patients who will respond or not to the treatment. Furthermore, a subsequent treatment combining BRAF inhibitors with immune therapy could increase the survival rate of those patients.

Many others studies are being carried out in our group to get insights into MM disease. They include several ‘omics’ strategies such as proteogenomics, transcriptomics and metabolomics. The most efficient, albeit challenging, way to study MM-related biological systems is to integrate the data generated from these platforms. This may lead, for instance, to the discovery of new biomarkers or drug target proteins.

Acknowledgments

I want to thank those who accompanied me during my doctoral studies. Special thanks to my supervisors and co-supervisor, Johan Malm, György Marko-Varga and Marcell Szász as well as Aniel Sánchez. You guided me through this journey and fueled my passion for science.

Johan and György, I will never forget that you welcomed me into this group of excellent scientists and helped and supported me until the very last minute.

Aniel, life will not be enough for me to thank you for your unconditional support and your patience in my moments of stress.

Thanks to Krzysztof Pawlowski and Péter Horvatovich for your kindness and wisdom in sharing your knowledge of bioinformatics with me. Thanks to Marcel and Viktória, who welcomed me when I went to Hungary and showed me the daily life in the clinic. Since then, my commitment to science and patients is bigger.

Thanks to the mass spectrometry team, Aniel, Barbara, Lázaro, Jimmy, Melinda, Jeovanis, Erika, Magdalena, Natalia, Kim and Nicole. Without your incredible experimental work, none of this would have been possible. I want to thank you also for the moments of joy we spent together and for celebrating each published paper with me.

Thank you very much, Jimmy, for the freehand artwork you draw for my thesis cover.

Thanks to Jonatan and Bea for the expertise, you brought regarding bioinformatics to the group.

Thanks to Zsolt Horvath, my IT savior. Thanks for your patience and knowledge in solving critical IT problems.

Thanks to Roger, Henriett, Henrik, Boram and Tillie for always being when I needed you. Your support regarding practical things related to the project and biobank has been crucial to focus on my research studies.

Thanks to the European Cancer Moonshot project and especially to the Szeged team. István, Leticia, and Agnes, thanks for sharing all your knowledge on malignant melanoma with me.

Thanks to ReproUnion and particularly to Professors Claus Y. Andersen and Aleksander Giwercman, Susanne Pors, Liv la Cour Poulsen, Stine Gry Kristensen and Yvonne Lundberg. Thanks to all of you for introducing me to the field of Reproductive medicine and teaching me all I know about it. Working with you has been an absolute pleasure.

Thanks to my dear Cariocas friends Gilberto, Fabio and Solange. Thanks for being who you are, for your advice, for your love for science and for being present every time I need you.

Many thanks to my family for being there, and support me unconditionally. You are my inspiration to keep going in times of weakness.

Many thanks also to my friends all around the world. Thanks for being in touch and worry about my progress day by day.

References

1. Das, T.; Andrieux, G.; Ahmed, M.; Chakraborty, S. Integration of Online Omics-Data Resources for Cancer Research. *Front. Genet.* **2020**, *11*.
2. Subramanian, I.; Verma, S.; Kumar, S.; Jere, A.; Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform. Biol. Insights* **2020**, *14*.
3. Marttinen, M.; Paananen, J.; Neme, A.; Mitra, V.; Takalo, M.; Natunen, T.; Paldanius, K.M.A.; Mäkinen, P.; Bremang, M.; Kurki, M.I.; et al. A multiomic approach to characterize the temporal sequence in Alzheimer's disease-related pathology. *Neurobiol. Dis.* **2019**, *124*, 454–468.
4. Ferlay J, Laversanne M, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, B.F. Global Cancer Observatory (2020): Cancer Tomorrow. Lyon, France: International Agency for Research on Cancer.
5. Medicine, A.S. for R. Fibroids and Fertility. *Am. Soc. Reprod. Med.* **2015**.
6. Nothnick, W.B. Treating endometriosis as an autoimmune disease. *Fertil. Steril.* **2001**, *76*, 223–231.
7. Jose-Miller, A.B.; Boyden, J.W.; A Frey, K. Infertility . *Am. Fam. Physician* **2007**, *75*, 849–856.
8. Deswal, R.; Narwal, V.; Dang, A.; SPundir, C.; Pundir, C.S. The Prevalence of Polycystic Ovary Syndrome: A Brief Systematic Review. **2020**.
9. Bates, G.W.; Bowling, M. Physiology of the female reproductive axis. *Periodontol. 2000* **2013**, *61*, 89–102.
10. Atwood, C.S.; Vadakkadath Meethal, S. The spatiotemporal hormonal orchestration of human folliculogenesis, early embryogenesis and blastocyst implantation. *Mol. Cell. Endocrinol.* **2016**, *430*, 33–48.
11. Rodgers, R.J.; Irving-Rodgers, H.F. Formation of the ovarian follicular antrum and follicular fluid. *Biol. Reprod.* **2010**, *82*, 1021–1029.
12. Jeppesen, J. V.; Anderson, R.A.; Kelsey, T.W.; Christiansen, S.L.; Kristensen, S.G.; Jayaprakasan, K.; Raine-Fenning, N.; Campbell, B.K.; Yding Andersen, C. Which follicles make the most anti-Mü llerian hormone in humans? Evidence for an abrupt decline in AMH production at the time of follicle selection. *Mol. Hum. Reprod.* **2013**, *19*, 519–527.
13. Hou, S.; Hao, Q.; Zhu, Z.; Xu, D.; Liu, W.; Lyu, L.; Li, P. Unraveling proteome changes and potential regulatory proteins of bovine follicular Granulosa cells by mass spectrometry and multi-omics analysis. **2019**, *17*, 1–11.
14. Kõks, S.; Velthut, A.; Sarapik, A.; Altmäe, S.; Reinmaa, E.; Schalkwyk, L.C.; Fernandes, C.; Lad, H.V.; Soomets, U.; Jaakma, Ü.; et al. The differential

transcriptome and ontology profiles of floating and cumulus granulosa cells in stimulated human antral follicles. *MHR Basic Sci. Reprod. Med.* **2010**, *16*, 229–240.

15. Poulsen, L.C.; Bøtkjær, J.A.; Østrup, O.; Petersen, K.B.; Yding Andersen, C.; Grøndahl, M.L.; Englund, A.L.M. Two waves of transcriptomic changes in periovulatory human granulosa cells. *Hum. Reprod.* **2020**, *35*, 1230–1245.
16. Virant-Klun, I.; Leicht, S.; Hughes, C.; Krijgsveld, J. Identification of Maturation-Specific Proteins by Single-Cell Proteomics of Human Oocytes. *Mol. Cell. Proteomics* **2016**, *15*, 2616–2627.
17. Zamah, A.M.; Hassis, M.E.; Albertolle, M.E.; Williams, K.E. Proteomic analysis of human follicular fluid from fertile women. *Clin. Proteomics* **2015**, *12*, 1–12.
18. Dimitriou, F.; Krattinger, R.; Ramelyte, E.; Barysch, M.J.; Micaletto, S.; Dummer, R.; Goldinger, S.M. *The World of Melanoma: Epidemiologic, Genetic, and Anatomic Differences of Melanoma Across the Globe*; 2018; Vol. 20, p. 87;.
19. Thompson, J.F.; Scolyer, R.A.; Kefford, R.F. Cutaneous melanoma. *Lancet (London, England)* **2005**, *365*, 687–701.
20. Mattia, G.; Puglisi, R.; Ascione, B.; Malorni, W.; Carè, A.; Matarrese, P. Cell death-based treatments of melanoma:conventional treatments and new therapeutic strategies. *Cell Death Dis.* **2018**, *9*.
21. Pardoll, D.M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **2012**, *12*, 252–264.
22. Sullivan, R.J.; Flaherty, K. MAP kinase signaling and inhibition in melanoma. *Oncogene* **2013**, *32*, 2373–2379.
23. Betancourt, L.H.; Gil, J.; Kim, Y.; Doma, V.; Çakır, U.; Sanchez, A.; Murillo, J.R.; Kuras, M.; Parada, I.P.; Sugihara, Y.; et al. The human melanoma proteome atlas—Defining the molecular pathology. *Clin. Transl. Med.* **2021**, *11*.
24. Piccinini, F.; Balassa, T.; Szkalisity, A.; Molnar, C.; Paavolainen, L.; Kujala, K.; Buzas, K.; Sarazova, M.; Pietiainen, V.; Kutay, U.; et al. Advanced Cell Classifier: User-Friendly Machine-Learning-Based Software for Discovering Phenotypes in High-Content Imaging Data. *Cell Syst.* **2017**, *4*, 651-655.e5.
25. Marcell Szasz, A.; Malm, J.; Rezeli, M.; Sugihara, Y.; Betancourt, L.H.; Rivas, D.; Gyorffy, B.; Marko-Varga, G. Challenging the heterogeneity of disease presentation in malignant melanoma---impact on patient treatment. *Cell Biol. Toxicol.* **2018**.
26. Eton, O.; Legha, S.S.; Bedikian, A.Y.; Lee, J.J.; Buzaid, A.C.; Hodges, C.; Ring, S.E.; Papadopoulos, N.E.; Plager, C.; East, M.J.; et al. Sequential biochemotherapy versus chemotherapy for metastatic melanoma: results from a phase III randomized trial. *J. Clin. Oncol.* **2002**, *20*, 2045–2052.
27. Griewank, K.G. Biomarkers in melanoma. *Scand. J. Clin. Lab. Invest. Suppl.* **2016**, *245*, S104–S112.
28. Ascierto, P.A.; Kirkwood, J.M.; Grob, J.J.; Simeone, E.; Grimaldi, A.M.; Maio, M.; Palmieri, G.; Testori, A.; Marincola, F.M.; Mozzillo, N. The role of BRAF V600 mutation in melanoma. *J. Transl. Med.* **2012**, *10*, 1–9.
29. Gonzalez, M.W.; Kann, M.G. Chapter 4: Protein interactions and disease. *PLoS Comput. Biol.* **2012**, *8*.

30. Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422*, 198–207.

31. Sinha, A.; Mann, M. A beginner's guide to mass spectrometry–based proteomics. *Biochem. (Lond).* **2020**, *42*, 64–69.

32. Issaq, H.J.; Chan, K.C.; Blonder, J.; Ye, X.; Veenstra, T.D. Separation, detection and quantitation of peptides by liquid chromatography and capillary electrochromatography. *J. Chromatogr. A* **2009**, *1216*, 1825–1837.

33. Stoll, D.R.; Carr, P.W. Two-Dimensional Liquid Chromatography: A State of the Art Tutorial. *Anal. Chem.* **2017**, *89*, 519–531.

34. Betancourt, L.H.; De Bock, P.J.; Staes, A.; Timmerman, E.; Perez-Riverol, Y.; Sanchez, A.; Besada, V.; Gonzalez, L.J.; Vandekerckhove, J.; Gevaert, K. SCX charge state selective separation of tryptic peptides combined with 2D-RP-HPLC allows for detailed proteome mapping. *J. Proteomics* **2013**, *91*, 164–171.

35. Giddings, J.C. Concepts and comparisons in multidimensional separation. *J. High Resolut. Chromatogr.* **1987**, *10*, 319–323.

36. Oh, J.-Y.; Barrett-Connor, E.; Wedick, N.M.; Wingard, D.L.; Rancho Bernardo Study No Title. **2002**, *25*, 55–60.

37. Shiio, Y.; Aebersold, R. Quantitative proteome analysis using isotope-coded affinity tags and mass spectrometry. *Nat. Protoc.* **2006**, *1*, 139–145.

38. Liu, Q.; Cobb, J.S.; Johnson, J.L.; Wang, Q.; Agar, J.N. Performance Comparisons of Nano-LC Systems, Electrospray Sources and LC–MS-MS Platforms. *J. Chromatogr. Sci.* **2014**, *52*, 120–127.

39. RA, Y.; CG, E. Triple quadrupole mass spectrometry for direct mixture analysis and structure elucidation. *Anal. Chem.* **1979**, *51*.

40. Gallien, S.; Duriez, E.; Crone, C.; Kellmann, M.; Moehring, T.; Domon, B. Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. *Mol. Cell. Proteomics* **2012**, *11*, 1709–1723.

41. Peterson, A.C.; Russell, J.D.; Bailey, D.J.; Westphall, M.S.; Coon, J.J. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol. Cell. Proteomics* **2012**.

42. Lesur, A.; Ancheva, L.; Kim, Y.J.; Berchem, G.; van Oostrum, J.; Domon, B. Screening protein isoforms predictive for cancer using immunoaffinity capture and fast LC-MS in PRM mode. *Proteomics. Clin. Appl.* **2015**, *9*, 695–705.

43. Andrews, G.L.; Simons, B.L.; Young, J.B.; Hawkridge, A.M.; Muddiman, D.C. Performance characteristics of a new hybrid quadrupole time-of-flight tandem mass spectrometer (TripleTOF 5600). *Anal. Chem.* **2011**, *83*, 5442–5446.

44. Michalski, A.; Damoc, E.; Hauschild, J.P.; Lange, O.; Wieghaus, A.; Makarov, A.; Nagaraj, N.; Cox, J.; Mann, M.; Horning, S. Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **2011**, *10*.

45. Megger, D.A.; Pott, L.L.; Ahrens, M.; Padden, J.; Bracht, T.; Kuhlmann, K.; Eisenacher, M.; Meyer, H.E.; Sitek, B. Comparison of label-free and label-based strategies for proteome analysis of hepatoma cell lines. *Biochim. Biophys. Acta* **2014**, *1844*, 967–976.

46. Li, J.; Van Vranken, J.G.; Pontano Vaites, L.; Schweppe, D.K.; Huttlin, E.L.; Etienne, C.; Nandhikonda, P.; Viner, R.; Robitaille, A.M.; Thompson, A.H.; et al. TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat. Methods 2020 174* **2020**, *17*, 399–404.

47. Thompson, A.; Wölmer, N.; Koncarevic, S.; Selzer, S.; Böhm, G.; Legner, H.; Schmid, P.; Kienle, S.; Penning, P.; Höhle, C.; et al. TMTpro: Design, Synthesis, and Initial Evaluation of a Proline-Based Isobaric 16-Plex Tandem Mass Tag Reagent Set. *Anal. Chem.* **2019**, *91*, 15941–15950.

48. Kong, A.T.; Leprevost, F. V.; Avtonomov, D.M.; Mellacheruvu, D.; Nesvizhskii, A.I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **2017**, *14*, 513–520.

49. Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc. 2016 1112* **2016**, *11*, 2301–2319.

50. Deutsch, E.W.; Lam, H.; Aebersold, R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol. Genomics* **2008**, *33*, 18–25.

51. Chen, C.; Hou, J.; Tanner, J.J.; Cheng, J. Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. *Int. J. Mol. Sci.* **2020**, *21*.

52. Sanchez-Pinto, L.N.; Luo, Y.; Churpek, M.M. Big Data and Data Science in Critical Care. *Chest* **2018**, *154*, 1239–1248.

53. Kammers, K.; Cole, R.N.; Tiengwe, C.; Ruczinski, I. Detecting significant changes in protein abundance. *EuPA Open Proteomics* **2015**, *7*, 11–19.

54. Huang, D.W.; Sherman, B.T.; Tan, Q.; Collins, J.R.; Alvord, W.G.; Roayaei, J.; Stephens, R.; Baseler, M.W.; Lane, H.C.; Lempicki, R.A. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* **2007**, *8*, R183.

55. Pathan, M.; Keerthikumar, S.; Ang, C.-S.; Gangoda, L.; Quek, C.Y.J.; Williamson, N.A.; Mouradov, D.; Sieber, O.M.; Simpson, R.J.; Salim, A.; et al. FunRich: An open access standalone functional enrichment and interaction network analysis tool. *Proteomics* **2015**, *15*, 2597–2601.

56. Franceschini, A.; Szklarczyk, D.; Frankild, S.; Kuhn, M.; Simonovic, M.; Roth, A.; Lin, J.; Minguez, P.; Bork, P.; Von Mering, C.; et al. STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **2013**, *41*, 808–815.

57. Mi, H.; Ebert, D.; Muruganujan, A.; Mills, C.; Albou, L.P.; Mushayamaha, T.; Thomas, P.D. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* **2021**, *49*, D394–D403.

58. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 15545–15550.

59. Lavallée-Adam, M.; Rauniyar, N.; McClatchy, D.B.; Yates, J.R. PSEA-quant: A protein set enrichment analysis on label-free and label-based protein quantification data. *J. Proteome Res.* **2014**, *13*, 5496–5509.

60. Cox, J.; Mann, M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* **2012**, *13 Suppl 1*, S12.

61. Lualdi, M.; Fasano, M. Statistical analysis of proteomics data: A review on feature selection. *J. Proteomics* **2019**, *198*, 18–26.

62. Hahne, F.; Huber, W.; Gentleman, R.; Falcon, S. *Bioconductor Case Studies (Use R!)*; 2008; ISBN 0387772391.

63. Rohart, F.; Gautier, B.; Singh, A.; Lê Cao, K.-A.A. mixOmics: An R package for 'omics feature selection and multiple data integration. **2017**, *13*, e1005752.

64. Ilmonen, S.; Asko-Seljavaara, S.; Kariniemi, A.L.; Jeskanen, L.; Pyrhönen, S.; Muhonen, T. Prognosis of primary melanoma. *Scand. J. Surg.* **2002**, *91*, 166–171.

65. Plym, A.; Ullenhag, G.J.; Breivald, M.; Lambe, M.; Berglund, A. Clinical characteristics, management and survival in young adults diagnosed with malignant melanoma: A population-based cohort study. *http://dx.doi.org/10.3109/0284186X.2013.854928* **2014**, *53*, 688–696.

66. TUKEY, J.W. Comparing individual means in the analysis of variance. *Biometrics* **1949**, *5*, 99–114.

67. Benjaminit, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300.

68. Dennis, G.; Sherman, B.T.; Hosack, D.A.; Yang, J.; Gao, W.; Lane, H.C.; Lempicki, R.A. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **2003**, *4*, R60.

69. Wu, T.; Hu, E.; Xu, S.; Chen, M.; Guo, P.; Dai, Z.; Feng, T.; Zhou, L.; Tang, W.; Zhan, L.; et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innov.* **2021**, *2*, 100141.

70. Yu, G.; Wang, L.G.; Han, Y.; He, Q.Y. ClusterProfiler: An R package for comparing biological themes among gene clusters. *Omi. A J. Integr. Biol.* **2012**, *16*, 284–287.

71. Tyanova, S.; Cox, J. Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. In; Humana Press, New York, NY, 2018; pp. 133–148.

72. Wei, W.; Sun, Z.; da Silveira, W.A.; Yu, Z.; Lawson, A.; Hardiman, G.; Kelemen, L.E.; Chung, D. Semi-supervised identification of cancer subgroups using survival outcomes and overlapping grouping information. *Stat. Methods Med. Res.* **2019**, *28*, 2137–2149.

73. Farrah, T.; Deutsch, E.W.; Omenn, G.S.; Campbell, D.S.; Sun, Z.; Bletz, J.A.; Mallick, P.; Katz, J.E.; Malmström, J.; Ossola, R.; et al. A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol. Cell. Proteomics* **2011**.

74. Szklarczyk, D.; Morris, J.H.; Cook, H.; Kuhn, M.; Wyder, S.; Simonovic, M.; Santos, A.; Doncheva, N.T.; Roth, A.; Bork, P.; et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **2017**, *45*, D362–D368.

75. Park, J.-Y.; Su, Y.-Q.; Ariga, M.; Law, E.; Jin, S.-L.C.; Conti, M. EGF-like growth factors as mediators of LH action in the ovulatory follicle. *Science* **2004**, *303*, 682–4.

76. Jalkanen, J.; Suikkari, A.M.; Koistinen, R.; Butzow, R.; Ritvos, O.; Seppala, M.; Ranta, T. Regulation of Insulin-Like Growth Factor-Binding Protein-1 Production in Human Granulosa-Luteal Cells. *J. Clin. Endocrinol. Metab.* **1989**, *69*, 1174–1179.

77. Kwon, H.; Choi, D.H.; Bae, J.H.; Kim, J.H.; Kim, Y.S. mRNA expression pattern of insulin-like growth factor components of granulosa cells and cumulus cells in women with and without polycystic ovary syndrome according to oocyte maturity. *Fertil. Steril.* **2010**, *94*, 2417–2420.

78. Baranova, N.S.; Inforzato, A.; Briggs, D.C.; Tilakaratna, V.; Enghild, J.J.; Thakar, D.; Milner, C.M.; Day, A.J.; Richter, R.P. Incorporation of pentraxin 3 into hyaluronan matrices is tightly regulated and promotes matrix cross-linking. *J. Biol. Chem.* **2014**, *289*, 30481–30498.

79. Ambekar, A.S.; Nirujogi, R.S.; Srikanth, S.M.; Chavan, S.; Kelkar, D.S.; Hinduja, I.; Zaveri, K.; Prasad, T.S.K.S.K.; Harsha, H.C.C.; Pandey, A.; et al. Proteomic analysis of human follicular fluid: A new perspective towards understanding folliculogenesis. *J. Proteomics* **2013**, *87*, 68–77.

80. Markström, E.; Svensson, E.C.; Shao, R.; Svanberg, B.; Billig, H. Survival factors regulating ovarian apoptosis -- dependence on follicle differentiation. *Reproduction* **2002**, *123*, 23–30.

81. Pita-Juárez, Y.; Altschuler, G.; Kariotis, S.; Wei, W.; Koler, K.; Green, C.; Tanzi, R.E.; Hide, W. The Pathway Coexpression Network: Revealing pathway relationships. *PLoS Comput. Biol.* **2018**, *14*, e1006042.

82. González, F.; Rote, N.S.; Minium, J.; Kirwan, J.P. Increased activation of nuclear factor kappaB triggers inflammation and insulin resistance in polycystic ovary syndrome. *J. Clin. Endocrinol. Metab.* **2006**, *91*, 1508–1512.

83. Liu, Y.; Liu, H.; Li, Z.; Fan, H.; Yan, X.; Liu, X.; Xuan, J.; Feng, D.; Wei, X. The Release of Peripheral Immune Inflammatory Cytokines Promote an Inflammatory Cascade in PCOS Patients via Altering the Follicular Microenvironment. *Front. Immunol.* **2021**, *12*.

84. González, F.; Rote, N.S.; Minium, J.; Kirwan, J.P. Reactive oxygen species-induced oxidative stress in the development of insulin resistance and hyperandrogenism in polycystic ovary syndrome. *J. Clin. Endocrinol. Metab.* **2006**, *91*, 336–340.

85. Berkholtz, C.B.; Lai, B.E.; Woodruff, T.K.; Shea, L.D. Distribution of extracellular matrix proteins type I collagen, type IV collagen, fibronectin, and laminin in mouse folliculogenesis. *Histochem. Cell Biol.* **2006**, *126*, 583–592.

86. Irving-Rodgers, H.F.; Rodgers, R.J. Extracellular matrix of the developing ovarian follicle. *Semin. Reprod. Med.* **2006**, *24*, 195–203.

87. Rodgers, R.J.; Irving-Rodgers, H.F.; Van Wezel, I.L. Extracellular matrix in ovarian follicles. *Mol. Cell. Endocrinol.* **2000**, *163*, 73–79.

Printed by Media-Tryck, Lund 2022 NORDIC SWAN ECOLABEL 3041 0903

Indira Plá Parada is an informatics engineer who has a background in bioinformatics and data analysis. In 2015, she got a master's degree in Biotechnology, and in 2016 continued her career as a data scientist in the fields of proteogenomics associated with cancer and reproductive medicine. Fields in which she pursued her doctoral studies.

LUND UNIVERSITY

FACULTY OF MEDICINE

Department of Translational Medicine

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2022:110
ISBN 978-91-8021-271-7
ISSN 1652-8220