Application of systems biology resources to human diseases : combining transcriptomics data analysis and molecular networks to identify major players

Luis Cristobal Monraz Gomez

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Institute Curie (U900)

Utilisation des ressources de la biologie des systèmes aux maladies humaines : combinaison de l'analyse des données transcriptomiques et les réseaux moléculaires pour identifier les acteurs majeurs.

Application of systems biology resources to human diseases: combining transcriptomics data analysis and molecular networks to identify major players.

Soutenue par

Luis Cristóbal MONRAZ GOMEZ

Le 20 Décembre 2023

Ecole doctorale n° 515

Complexité du vivant

Spécialité

Bioinformatique et biologie des systèmes

Composition du jury :

Nathalie, DOSTATNI Directrice de Recherche, Institute Curie/PSL	*Présidente*
Adrien, ROSSARY Maître de Conférences des Universités, Université Clermont-Auvergne	*Rapporteur*
Marek, OSTASZEWSKI Research Associate, LCSB/University of Luxembourg	*Rapporteur*
Michèle, SABBAH Directrice de Recherche, CRSA/Sorbonne Université	*Examinatrice*
Hédi, SOULA Professeur, Sorbonne Université	*Examinateur*
Inna, KUPERSTEIN Ingénieur de Recherche, Institut Curie/PSL	*Directrice de thèse*
Mathieu, BOISSAN Maître de Conférences des Universités-Praticien Hospitalier, CRSA/Sorbonne Université	*Co-directeur de thèse*

institutCurie | PSL

ACKNOWLEDGEMENTS

This endeavor, that comprised many years, would not have been possible without the advice, help and company of many people that I would like to express my gratitude to.

I would like to thank Inna, for making me part of her team back in 2017. She has trusted in me and included me in different projects, giving me the opportunity to participate in different projects and tasks that have led me to get to know different people, places and ways to do science. I am grateful for the support and interest that she showed when I presented her the idea of the LipoCanPredict project, that led to the application for different grants, and the start of this project, that is still ongoing. I am also grateful for the all the advice, support and encouragement for the completion of this thesis.

I want to thank Mathieu, for accepting being part of the direction of this thesis. I have enjoyed working in your lab and with your team. I am thankful for the great advice, disposition and positivity that you have injected into my projects. I admire your tenacity and I am looking forward to collaborating with you in the future.

My arrival in Paris would not have been possible without Emmanuel Barillot and Andrei Zynoviev. I want to thank you for making me part of the unit and also the sysbio team, the best team to do research. You have created a really pleasing team and unit to be part of, with a great ambience and diverse in scientific interests and backgrounds. Additionally, I would like to thank Laurence, Loredana and Christine for the nice exchanges, scientific or not, and for always have the disposition to have a good laugh. Thank you very much for everything.

After 6+ years of being part of the sysbio team in the U900, I have met many people that has contributed with this work, but also with my scientific efforts. I would like to thank Laura, Gaëlle, Pauline, Maria, Mihaly, Luca, Arnau, Urszula, Nicolas, Om, Jonas, Loïc for the great exchanges, the experiences in conferences, zumba lessons and the nice atmosphere in the team during my first years being part of it until the initiation of my PhD. Also I want to express my gratitude to Jane, Vincent, Marianyela, Saran, Marco, Miguel, Nicolas, Loïc C, Mathieu N, Luca, Altynbek, Alexander, Anne-Claire, Aziz, Jonathan, Lucie, Ellora, Daniel, Martina, Sophia (mi amix), for the great ambience in the team during these last years and months, especially the later. Also, thank you for the amusing, sometimes strangely philosophical, conversations with coffee or some variety of bakery good. I would like to thank all of you for this great time I have passed in the team.

To continue, thank you to all the colleagues in the different U900 teams, without you, this work would not been possible, special thanks to Nicolas S and Pierre for your advice when it came to dealing with the data. Also, thank you very much Caroline, Kati, Emma and Justina for facilitating the administrative tasks but also for sharing and being part of the nice environment in the unit.

I am really grateful with Céline Prunier, for accepting me in her team and invade their installations at CRSA, to perform the needed lab tasks, in order to manipulate my samples. I want to thank Joëlle, Laurence, Sadek, Sabrina, Henry, Tiphaine, Juliane and Robin for the great welcome you made me in your team, as well as to the great feedback and disposition you had when I have presented the results of my project.

The LipoCanPredict consortium, I am indebted with your participation in the planning and execution of this project. Thank you very much Ludger, Anne, Julio, Christian for your disposition and your scientific inputs regarding the project. Also, thank you Charlotte for dealing with the logistics to provide us with the specimens together with Hélène and Andreia. I want to thank Zineb, for the great work she invested in this project during her master internship.

I would like to thank Benjamin, Markus and Andrea, for the great times spent inside and outside the Institute. It has been a pleasure meeting you and to share many thing in common. I want to thank you as well for the many times you have saved me by sharing tips and some supplies.

I thank my friends for your support and encourage throughout all this journey: Jorch, Chema, Octavio, Mayra, Ale, Simon, Shauna, Richard, Ivo, Andres, Ely, Adrian, Xunaxi, Giulia, Zuriel, Mickaël, Luis David, Dan, Juan Carlos, I know the distances or times are difficult to pair, but you have my greatest appreciation.

Finally, I would like to thank to my family members for their unconditional love and support all the time. To my mother, for her tireless support and drive. My siblings Kualziuatl and Jacobo, thank you for being always there for me. Diego, thanks a lot for everything. Jaume, Joan and Raiza, thank you for bringing joy to the family. I am looking forward to seeing all of you after some years.

CONTENTS

List of abbreviations

AF: activity flow

ATF6: activating transcription factor 6

BMI: body mass index

CAA: cancer associated adipocytes

CNV: copy number variation

COVID-19: coronavirus disease 2019

DEG: differentially expressed genes

DNA: deoxyribonucleic acid

EMT: epithelial to mesenchymal transition

ER: endoplasmic reticulum

ER: estrogen receptor (used mostly in chapter 2)

GO: gene ontology

GSEA: gene-set enrichment analysis

GWAS: genome-wide association study

HCoV: human coronavirus

HER2: human epidermal growth factor receptor 2

IRE1: Inositol-requiring enzyme 1

kb: kilobase pairs

KEGG: Kyoto encyclopedia of genes and genomes

Ki-67: Antigen Kiel 67 (Cell proliferation marker)

NGS: next generation sequencing

PCA: principal component analysis

PD: process description

PERK: protein kinase RNA-activated (PKR)-like ER protein kinase

PR: progesterone receptor

RCD: regulated cell death

ROMA: representation of module activity

ROS: reactive oxygen species

SARS: severe acute respiratory syndrome

SARS-CoV-1: severe acute respiratory syndrome coronavirus 1

SARS-CoV-2: severe acute respiratory syndrome coronavirus 2

SASP: senescence-associated secretory phenotype

SBGN: systems biology graphical notation

SIGNOR: the SIGnature Network Open Resource

SNP: single nucleotide polymorphism

UPR: unfolded protein response

VST: variance stabilizing transformation

WHO: world health organization

CHAPTER I. INTRODUCTION

Systems biology is a scientific field that integrates the analysis of large amounts of data to study complex systems that is widely applied in current times. Due to the combination of different sciences and applications of systems biology, there are different approaches that can be applied to answer scientific questions. In biomedical sciences, the application of systems biology approaches has become very popular, especially with the exploitation of de different “omics” types of data. In this thesis, I have explored different systems biology approaches in human diseases, mainly cancer, and used transcriptomics data to understand systems of interest.

A. SYSTEMS BIOLOGY: AN OVERVIEW

A system can be defined as a group of elements that act as parts of a mechanism or an interconnecting network; an intricated whole. From this definition, in systems biology, the goal is to study a given system by looking at all its components simultaneously. Currently, with the exploitation of the different omics data types it has been feasible to acquire knowledge about the biological mechanisms occurring in the system. However, there are different forms to approach a scientific question in this field, generally speaking can be classified as top-down and bottom-up approaches [1,2].

I. TOP-DOWN APPROACHES

The top-down approach, also known as data-driven approach[3], starts from experimental data that is analyzed in order to resolve a biological question or in an exploratory manner that will generate new hypotheses. This type of approaches are integrative and rely in the generation of omics data[4], this data types will be discussed later in this chapter. After the study is designed and performed, the specimens should be prepared for the analysis to be performed, i.e. extraction of nucleic acids, lipids, metabolites, proteins etc. according to the requirements of the study.

After the data is generated, the following step is the processing, to assure the quality and success of the experiment and the fidelity of the results in downstream analyses. For the data analysis, a common approach in the RNA-seq transcriptomics data analysis is the following. The data is normalized so that the values are comparable among the measured samples[5], commonly used methods are log transformations or variance stabilizing transformation[6]. Then data are compared using statistical analyses such as t-tests, ANOVA etc. in order to obtain differentially expressed genes (DEG). This genes usually are determined by fold change differences and p-values, the last are usually adjusted using a multiple testing correction method like false discovery rate (FDR)[7]. With the DEG, a functional analysis can be performed to determine the biological pathways or processes that correspond to the DEG. There are many databases that are used for this type of analysis such as gene ontology (GO)[8], the Kyoto Encyclopedia of Genes and Genomes (KEGG)[9], reactome[10], WikiPathways[11] or the molecular signatures data base (MSigDB)[12]. Then, the information from these pathways can be visualized in the resources that offer that option, such as KEGG and reactome. Also protein-protein interaction networks can be done in software such as Cytoscape[13] or databases as the SIGnaling Network Open Resource (SIGNOR).

With the information obtained from these analyses, then the results can be interpreted according to what is available in the scientific literature. By their integrative nature, these approaches require strong collaborations between experimental biologists, bioinformaticians, biostatisticians, computational biologists, and in the field of human health, epidemiologists, physicians and health care personnel, among others.

II. BOTTOM-UP APPROACHES

The bottom-up approaches, also termed knowledge-based approaches[3], consist in creating detailed models that will be simulated under different conditions. These type of models represent biological knowledge and can be depicted as diagrams, for instance KEGG[9], reactome[10], WikiPathways[11], pathway commons[14] among others. In E. Barillot's group, the atlas of cancer signaling networks (ACSN) has been developed[15]. Alternatively, there are models that are not diagrams, but they describe a biological system by integrating pertinent knowledge and data to perform simulations in order to explore different outcomes[16,17]. These type of models can be represented as a set of equations, like differential equations, or in case of genome scale metabolic models, stoichiometric matrices, like recon[18].

Since these approaches are based on published findings, biocuration is needed, this is the action of extracting the knowledge and formalizing it, according to an established set of guidelines. Then it is represented, for instance as a diagram, and shared to the community [19]. For the coherence of this thesis, the rest of this section will focus on the signaling network representations, as diagrams, however there are overlaps with other bottom-up approaches as well.

Before starting the process of biocuration, it is important to select the most appropriate type of representation to be used. According to the systems biology graphical notation (SBGN)[20], there are three types of representations: (1) process description (PD) diagram, known as bi-partite reaction network graph in chemical kinetics, depicting the biochemical interactions in a network; (2) activity-flow (AF), also referred as regulatory network or influence diagram, represents the flow of information or the interactions between the entities; and (3) entity relationship (ER) diagram, illustrate the relations in which a given entity participates.

When creating a diagram, it is also important to follow stablished standards in order to produce exchangeable comprehensive diagrams. Currently, the SBGN syntax is one of the strongly proposed in the field. This syntax is compatible with various pathway drawing software as well as analytical tools, allowing the representation of cellular compartments and phenotypes, in addition to the biochemical processes. Moreover, to increase compatibility between pathway resources, various formats to exchange information, as BioPAX, SBML, PSI-MI among others, have been recommended[21]; for instance, models in SBML format can be stored in repositories such as BioModels[22,23], facilitating their access and reproducibility. To draw the network, there are some free available software solutions that can be used, such as CellDesigner[24], SBGN-ED[25], PathVisio[26], Newt[27]. An example of these representations, for PD diagrams using the SBGN syntax is in Figure 1.

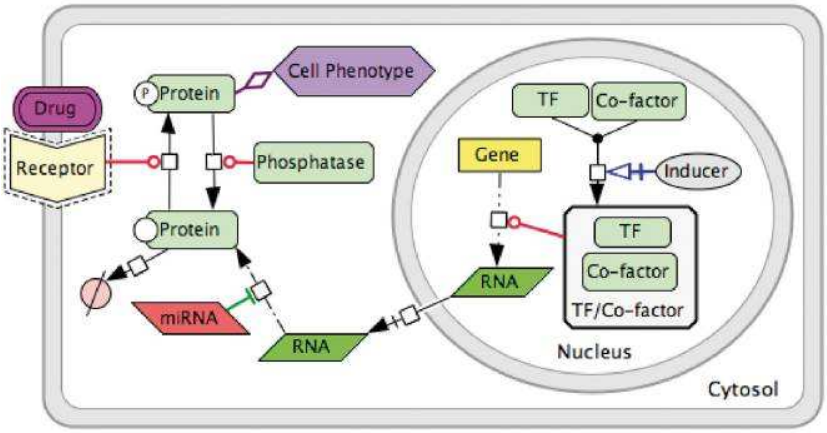


Figure 1. Process description (PD) diagram data model, using CellDesigner and the SBGN syntax.

The produced network diagram can become an issue for visualization when it reaches a high complexity due to a large amount of entities and reactions, this type of diagrams can even reach the dimensions of geographical maps. Hence, some platforms have been developed, such as Pathway projector[28], CellPublisher[29], NaviCell[30,31], MINERVA[32], with the goal of treat these diagrams as geographical maps and integrating the navigation logic and features from Google maps' technology. The adoption of characteristics such as zooming in, scrolling, pinning and callouts in these maps, makes them user friendly. Therefore, in the field, these molecular network diagrams are referred to as maps.

These maps are widely used as databases, source of information or for data visualization. For instance, in ACSN, thanks to the NaviCell technology, it is possible to visualize different omics data types, allowing the user to explore their data in different ways. In the last section of this chapter I will expose some use-cases from these tools. In Figure 2 there is an example of the global environment of ACSN and NaviCell using the angiogenesis map as an example and showing some of the navigation features of the maps.

To conclude this section of this chapter, systems biology offers different approaches useful for research, these approaches are widely used in biology with applications for engineering, environmental sciences, microbiology, food industries, pharmaceutical design, human health among others [1,33,34], and as well as the top-down approaches, they require interdisciplinarity for their successful development.

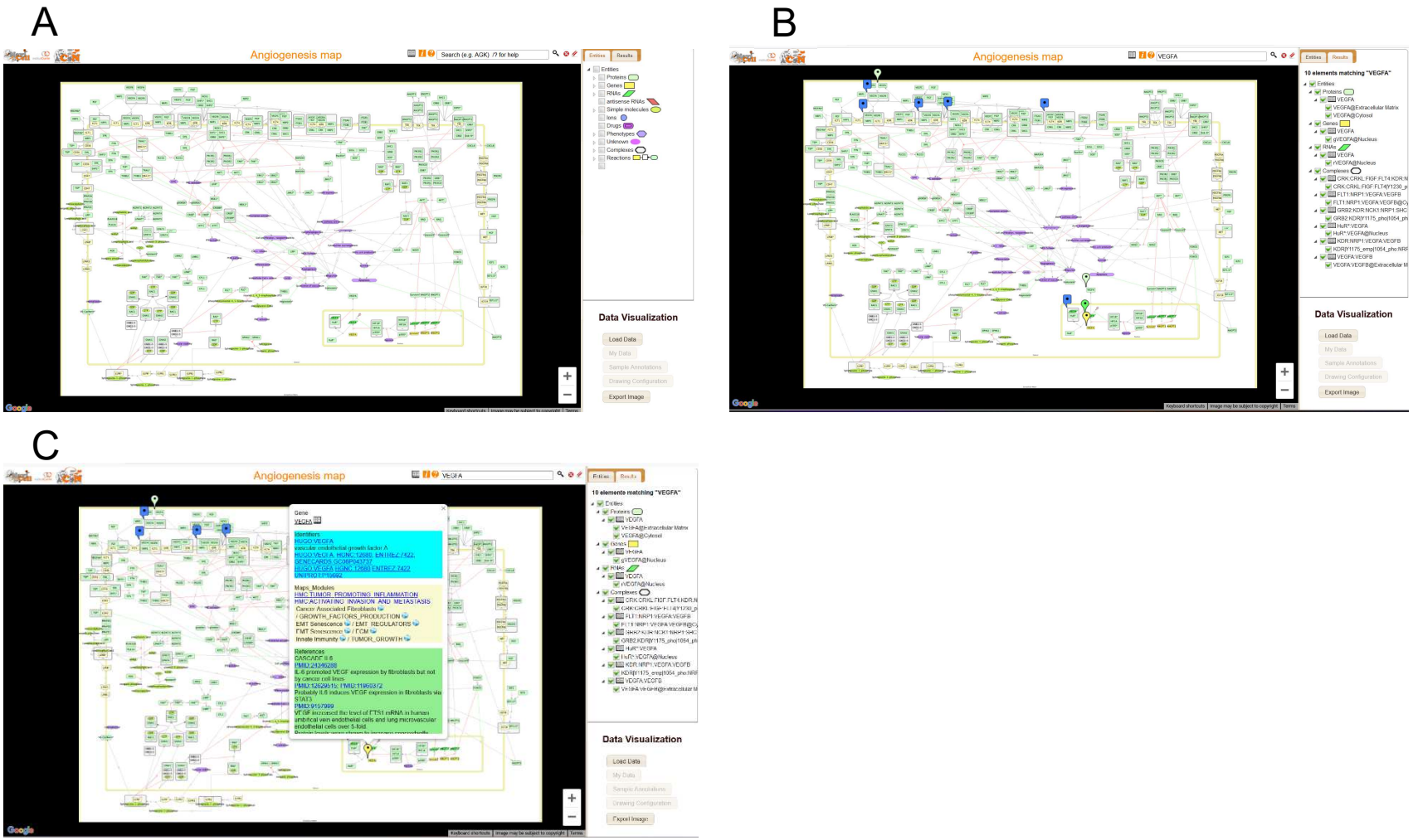


Figure 2. Example of browsing using the Angiogenesis map in ACSN. A**:** Map interface in NaviCell powered Google maps platform with the layout of the map. The interface has the map window, selection panel, data analysis panel and upper panel. B: Example of a query for VEGFA that drops the pins where it is located in the map. C: Callout window that appears when clicking on one of the entities, this call out window displays details such as different identifiers of the entity, where the entity can be found in other ACSN maps as well as the references the curator consulted and added.

B. THE OMICS

The sequencing of the human genome by the Human Genome Project[35], led to the development of new technologies that make possible to measure a vast array of molecules in a biological specimen. These gave rise to the omics disciplines, characterized by measuring biological molecules in a high-throughput manner [4,36]. To date, there are different omics according to the subject of study, in this section, some of them will be described.

I. GENOMICS

Genomics is the study of the genome, which is the complete set of DNA of an organism. The DNA can be complete or partially studied. Examples of studies in genomics include genome-wide association studies (GWAS), in which genetic variants are studied by their loci and are aimed to associate them to a trait, often a disease, usually single nucleotide polymorphism (SNP) chips are used [37,38]. Copy number variation (CNV) refers to a loss or amplification of the number of DNA segments larger than 1 kb, smaller events are known as insertions or deletions (indels), these studies reflect the genomic arrangement in the chromosomes [39,40] the studies of CNV use DNA sequencing, next generation sequencing (NGS) or SNP6 arrays from Affymetrix[41].

II. TRANSCRIPTOMICS

The goal of transcriptomics is to measure the RNA, derived from DNA, to determine gene activation. The transcriptome entails different types of RNA, the messenger RNA (mRNA), the ribosomal RNA (rRNA), transfer RNA (tRNA), microRNA (miRNA), and non-coding RNA (ncRNA). Approaches to study the RNA include the Microarrays which contain probes and the expression is measured by fluorescence. However, with the next generation sequencing (NGS), it has been the most prominent method used for transcriptomics[42,43]. However, the development of technologies has given rise to approaches such as single-cell transcriptomics and spatial transcriptomics. The first looks to analyze the transcriptome of individual cells in a population by separating them and the later to look from a histological-type slide, different places that might correspond to different cell types that can also be annotated by microscopically looking at the slide and determine communication between different cell types[44–46].

III. PROTEOMICS

The proteome is composed by the proteins that are expressed in an organism. An important feature of proteins is that they can be modified by post-translational modifications (PTMs), thus changing their amino acid structure and their three dimensional conformation. These PTMs are important to regulate the protein homeostasis and turnover, and their detection can provide information of the biological system. Proteomics data usually are generated using mass-spectrometry, by getting the mass to charge ratios of the amino acids in their structures [47,48]. Also, the phospho-proteomics approach, that measures phosphorylated proteins, is widely used to investigate the phosphorylation effects in proteins [49,50].

IV. METABOLOMICS/LIPIDOMICS

The metabolomics refers to the quantification of small molecules, referred to as metabolites. These molecules can be seen as indicators of processes occurring in the cell at the level of protein interactions and have been used as biomarkers [51,52]. The common methods to measure the metabolome are combinations of liquid chromatography with mass spectrometry, gas chromatography with mass spectrometry or nuclear magnetic resonance [53]. Lipidomics comprise the study of lipids in depth in different systems, usually measured with mass spectometry [54].

There are other different types of omics, such as epigenomics, studying some interactions like histone modifications or DNA-methylations that can have effects on the gene expression. Also, we have metagenomics, an approach that is aimed to analyse the genotype of the microorganisms in the microbiome of an individual. In the next section of the chapter, we will see some of these omics data types being used in systems biology.

C. EXAMPLES OF SYSTEMS BIOLOGY APPLICATIONS

In this section of the chapter, I present some conducted studies where I have applied systems biology approaches to address some biological questions. The presented work in this section preceded the work in the following chapters of this thesis.

I. IDENTIFICATION OF PLAYERS BETWEEN PRL-3 AND METASTASIS

The phosphatase PRL-3 has been described as a marker of tumor progression, specially metastasis in many cancers, such as gastric carcinoma[55], cervix cancer[56], breast cancer[57] or uveal carcinoma[58]. In this case, a map was created by reviewing literature, depicting the interactions of PRL-3 to different effectors resulting in metastasis. The information about these processes was depicted using CellDesigner[24] following the Systems Biology Graphical Notation (SBGN)[20]. The signaling map is available at: https://acsn.curie.fr/navicell/maps/invasion_motility/master/index.html. The resulting network was then reduced, in order to obtain the key players between PRL-3 and intermediate processes related to metastasis, that were, cell matrix-adhesions, cytoskeleton remodelling, cell cycle and survival, angiogenesis, EMT and motility and invasion. This was done through path analysis on Cytoscape[13], using the plugin BiNoM[59]. This analysis has as a goal to find the shortest path between a source node and a target node in the given network, resulting in different reduced networks. This network reduction approach permits to comprehend the organization principles of the map and to identify the essential paths and the players involved in them.

We depicted PRL-3 relationships to cell cycle, survival and apoptosis, first by RAP1 and downstream activation of G2/M transition. STAT3 activation induces PTEN inhibition which in a similar way with PI3K activation, results in AKT activation with further p53 inhibition resulting in reduced apoptosis. Other mechanisms reducing apoptosis regulated by PRL-3 were the direct induction of p14 as well as PIRH2 induced by EGR1 upstream. PRL-3 induced glycated ULBP2 in order to promote the immune escape of the cancer cells. The inhibition of PTP1B together with the activation of ERK1/2 resulted in the induction of EGF/EGFR that resulted and cell survival and proliferation by activation of downstream STAT and PI3K/AKT/MTOR pathways (Figure 3a).

When investigated the roles of the phosphatase resulting in angiogenesis, we observed that inhibition in the transcription of *IL-4* was directly related to promote angiogenesis, whereas the induction of Src with downstream ERK, and RHOC and RHOA, resulting in downstream VEGF/VEGFR resulted in micro-vessel formation and angiogenesis promotion in cancer cells (Figure 3b).

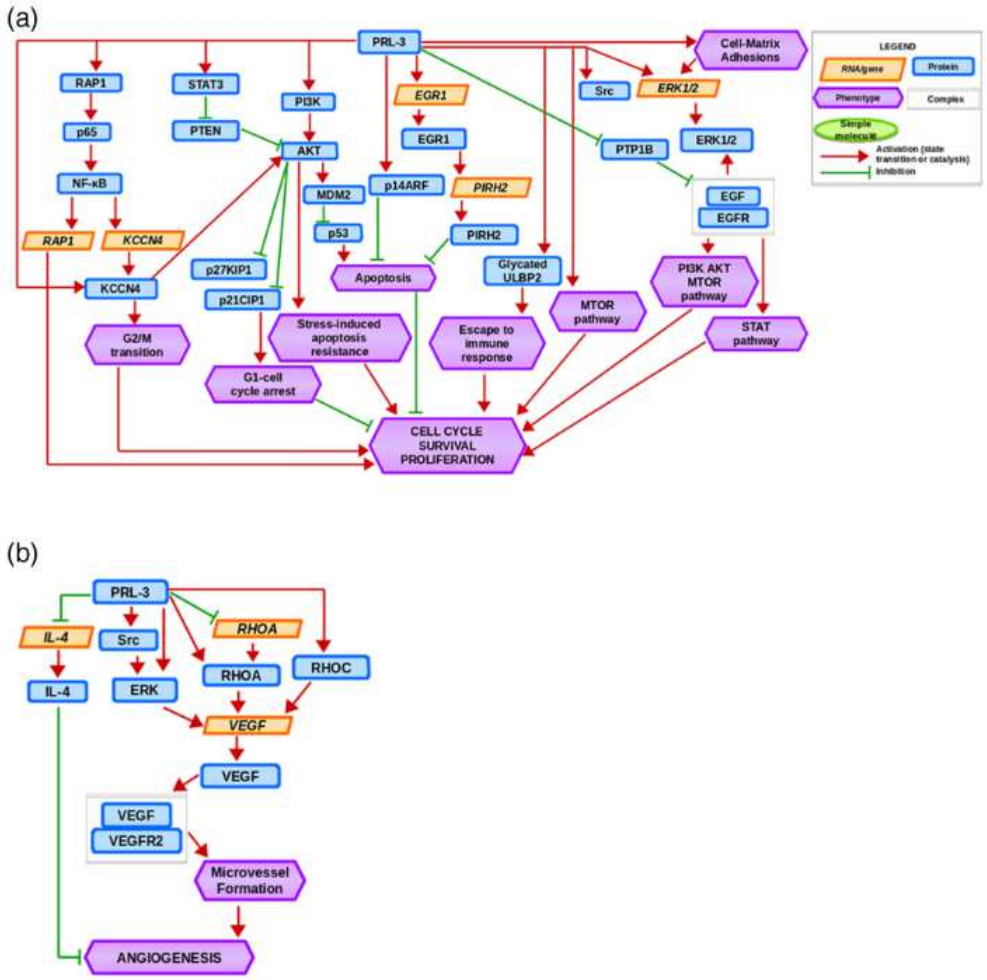


Figure 3. (a) molecular mechanisms induced by PRL-3 leading to cell cycle progression, survival and proliferation in cancer. (b) molecular mechanisms leading to angiogenesis induced by PRL-3 in cancer.

In cell adhesions, the main regulatory mechanism from PRL-3 was in preserving integrins and their recycling. Additionally, activation of focal adhesion kinase (FAK) and Src, important effectors of migration and adhesion. In cytoskeleton remodeling, Ezrin (ERZ) was one of the main induced players. The inhibition of PTP1B and RAC together with induction of RHOA and ROCK result in cytoskeleton remodeling disfunction that leads to a change in cell polarity and further invasive profile (Figure 4).

In figure 5, there are depicted the mechanisms between PRL-3 and EMT and motility as well as invasion, the induction of AKT with further GSK3β inhibition, results in induction of Snail and further mesenchymal markers. For invasion, the degradation of the matrix is essential, and PRL-3 promotes its degradation by the induction of different metalloproteinases. With the presented results, the used approach allowed to organize the information contained in the map into different sub-networks, containing the major players in the interface of PRL-3 activation and metastasis, focusing in the key steps that lead to the latest. The results presented in this section, have been already reviewed and published[60].

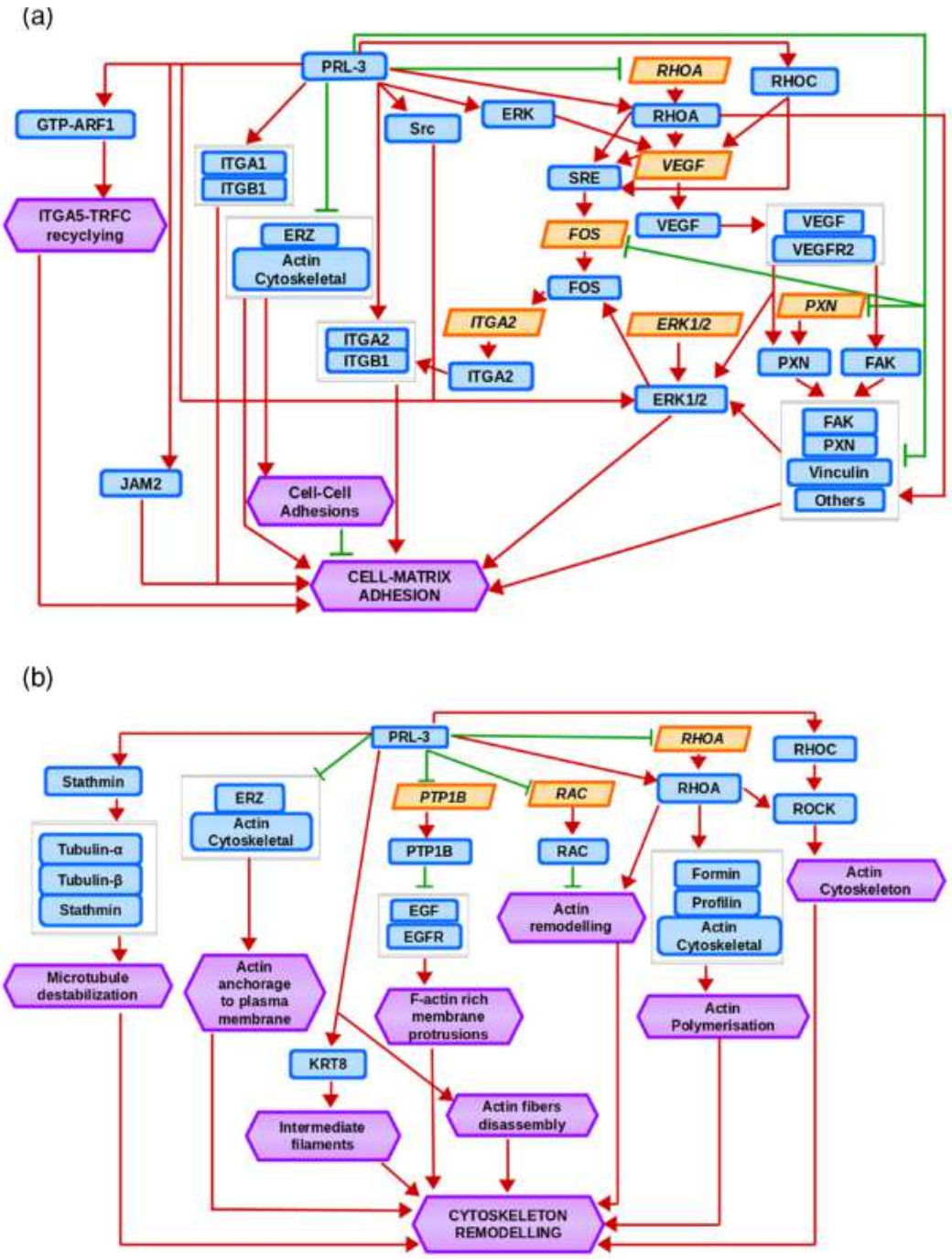


Figure 4. (a) molecular mechanisms of cell matrix adhesion induced by PRL-3 in cancer. (b) molecular mechanism of cytoskeleton remodeling induced by PRL-3 in cancer. for the annotations, please see the legend in **Figure 3.**

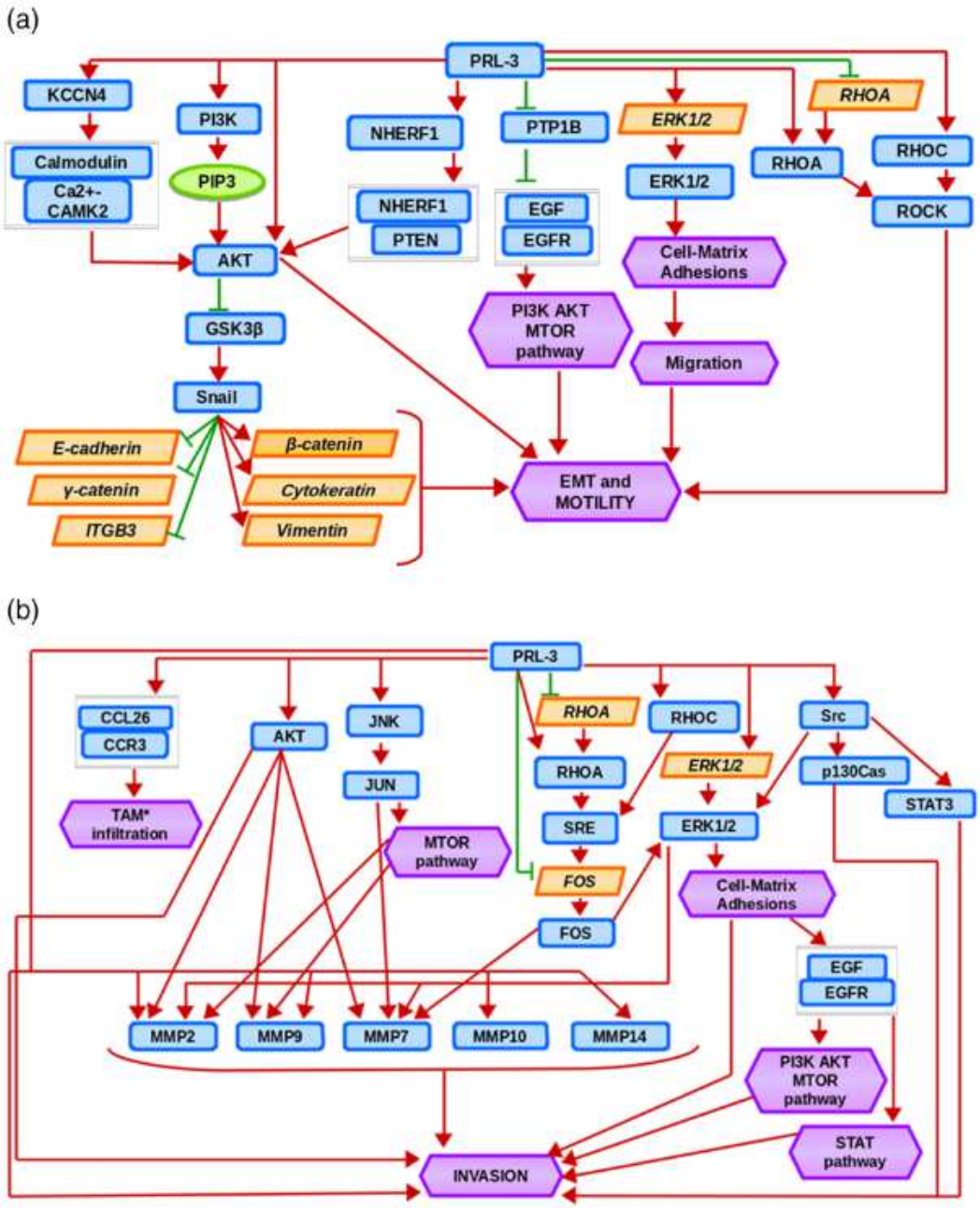


Figure 5. (a) Molecular mechanisms leading to EMT and motility induced by PRL-3 in cancer. (b) molecular mechanisms that promote invasion induced by PRL-3 in cancer. For the annotations, please see the legend to Figure 3.

II. A MAP OF REGULATED CELL DEATH USED TO EXPLORED DISEASE MECHANISMS

Regulated cell death (RCD) comprises different processes that lead to the end of the life of a cell. These processes can be triggered by different mechanisms. In cancer cells, the avoidance of cell death is essential and is a hallmark of this type of cells[61]. A map depicting the different types of RCD was created in ACSN and it was used for visualization of omics data.

Lung cancer and Alzheimer's disease (AD), described to be inverse comorbidities, were compared in the context of this map to explore the differences regarding the RCD processes. AD, as well as other neurogenerative disorders, has as feature an aberrant cell cycle reentry, which leads to cell death[62]. Lung cancer as many other cancer subtypes, on the other hand, are constant dividing cells that try to avoid the activation of RCD[61].

For this study, gene expression datasets from lung cancer, non-small cell lung cancer (NSCLC), and AD samples from the hippocampus, that is a commonly affected area in (GSE36980, GSE48350, GSE5281, GSE19188, GSE19804, GSE33532) were downloaded from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/). For the testing, we grouped the cases of each disease and compared them to their respective controls.

The ROMA[63] method was used to obtain activation scores across the gene-sets. The scores of AD and NSCLC exhibited rather an inverse trend (Figure 6). Regarding RCD types, the modules corresponding to the TRAIL response, Pyroptosis, and Dependence Receptors were more active in AD (Figure 6A). In contrast, in NSCLC, the ligand-receptor modules (TNF response, TRAIL response, and FAS response), as well as some modules of the signaling layer (Pyroptosis and Dependence Receptors), are less active (Figure 6A).

Of note, most modules related to the pyroptosis module appeared to be more active in Alzheimer disease. In addition, in NSCLC several metabolism-related modules (including glucose metabolism, oxidative phosphorylation and the citric acid cycle), as well as ER stress, were more active (Figure 6B,C). This metabolism-related modules activity confirm previous observations[64,65], including on the difference between Alzheimer's disease and lung cancer. Indeed, the integrated comparison of the RCD map across AD and NSCLC is in line with speculations on the inverse comorbidity between both diseases, as well epidemiological studies suggesting that NSCLC occurs less frequently in AD patients than in age-matched individual without AD[66–68].

Following this analysis, the top contributing genes for each disease were identified. For this, the correlation coefficients in all the studies were calculated and those genes that had a correlation coefficient of minimum 0.5 (absolute value), with a significant p-value and that appeared in at least 2 of the 3 data sets were selected. As a result, a list of genes that contribute either positively or negatively to each module was retrieved.

In the case of Alzheimer, that had a great activitation of pyroptosis, it has defined as a CASP1-dependent response to chronic aseptic inflammation[69,70]. Experimental evidence has linked pyroptosis in Alzheimer to the NLRP1[71] inflammasome or NLRP3[72] inflammasome. However, most studies correlating Alzheimer and pyroptosis have been performed in rodent models. Here, we identified *IL18, CASP4*, *GBP2, CASP1*, and *AIM2* as the genes that were contributing most to the pyroptosis module. *IL18* gene is over-expressed in brains of AD patients[73]. CASP4 expression has been hypothesized to mediate inflammatory responses in AD pathology[74]. *CASP1*, together with other genes encoding caspases, is overexpressed in AD patients[75]. AIM2 has been found in mouse models to promote IL1B secretion by neurons, which might also participate in AD pathology. However, GBP2 expression has not yet been evaluated for its potential role in AD pathology. Recently, Saresella and colleagues found that inflammasome components (NLRP1,

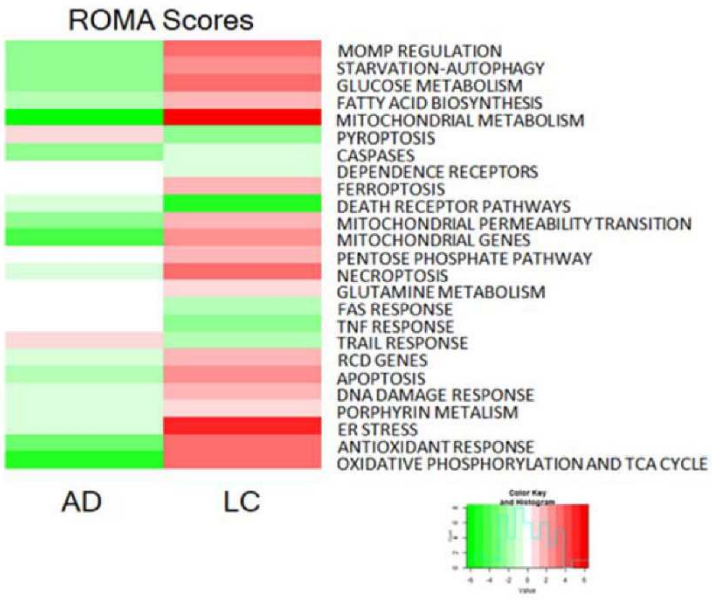
NLRP3, PYCARD, CASP1, CASP5, and CASP8) and downstream effectors (IL1B, IL18) were upregulated in peripheral blood mononuclear cells from patients with moderate and severe AD[76]. All these findings support our results using the RCD map; nevertheless, future research is needed to elucidate if neuroinflammation leads to pyroptosis during AD pathology.

In cancer, ER stress has been identified as an adaptive response that favors either growth or apoptosis[77,78]. In addition, ER stress has also been related to chemotherapy resistance[79]. In our study, we identified the genes *PDIA6, PDIA4, DNAJB11, SEC61G, SEC61A1*, and *CREB3L4* as positively contributing genes and *ITPR1, RYR2, NFKB1*, and *NLRC4*, as negatively contributing genes for the ER stress module. *PDIA6* and *PDIA4* have been demonstrated to be overexpressed in NSCLC biopsies resistant to chemotherapy with cisplatin, and their silencing actually may reverse drug resistance[80]. The *NFKB1* gene has been described to be a key player in the ER stress pathway and cancer survival mechanisms[81,82]. *NLRC4* is downregulated in lung cancer cases[83] . NLRC4 contains a caspase recruitment domain (CARD) through which it can regulate apoptosis via NF-κB signaling pathways, suggesting a possible link between NLRC4 and NFKB1 genes in this module. In contrast, there are no consistent reports on the possible involvement of *DNAJB11, SEC61G, SEC61A*, and *CREB3L4* in NSCLC.

Taken together, these results shown the application of a signaling network map to study disease mechanisms and obtain the major players. In this case, the differences were shown from two diseases that have been described to be inverse comorbidities.

The results of this study have been already published[84]. The presented two examples, depict how systems biology approaches can help to study human diseases. The next chapters describe different approaches that were addressed for different questions and contexts.

A



B

C

Figure 6. Visualization of average ROMA modules activity scores using expression data from Alzheimer disease (AD) hippocampus samples and non-small cell lung cancer (NSCLC) specimens in the RCD map. (a) Heatmap representing ROMA scores for the two diseases (each respective to its normal controls). staining of RCD map with ROMA scores from (b) AD data and (c) NSCLC data. the plotted values correspond to the relative ROMA module score compared to controls (as in a). top contributing genes are represented in their locations on the map, purple positively contributing genes and yellow negatively contributing genes.

D. REFERENCES

1. Shahzad, K. & Loor, J. J. Application of Top-Down and Bottom-up Systems Approaches in Ruminant Physiology and Metabolism. *Curr. Genomics* **13**, 379–94 (2012).

2. Tong, W. Analyzing the Biology on the System Level. *Genomics. Proteomics Bioinformatics* **2**, 6–14 (2004).

3. Bartmanski, B. J., Rocha, M. & Zimmermann-Kogadeeva, M. Recent advances in data- and knowledge-driven approaches to explore primary microbial metabolism. *Current Opinion in Chemical Biology* vol. 75 (2023).

4. Shalhoub, J. *et al.* Systems biology of human atherosclerosis. *Vasc. Endovascular Surg.* **48**, 5–17 (2014).

5. Walach, J., Filzmoser, P. & Hron, K. Data Normalization and Scaling: Consequences for the Analysis in Omics Sciences. *Compr. Anal. Chem.* **82**, 165–196 (2018).

6. Johnson, K. A. & Krishnan, A. Robust normalization and transformation techniques for constructing gene coexpression networks from RNA-seq data. *Genome Biol.* **23**, 1 (2022).

7. Koch, C. M. *et al.* A Beginner's Guide to Analysis of RNA Sequencing Data. *Am. J. Respir. Cell Mol. Biol.* **59**, 145–157 (2018).

8. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

9. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

10. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* (2018) doi:10.1093/nar/gkx1132.

11. Pico, A. R. *et al.* WikiPathways: Pathway Editing for the People. *PLOS Biol.* **6**, e184 (2008).

12. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* (2005) doi:10.1073/pnas.0506580102.

13. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2498–2504 (2003) doi:10.1101/gr.1239303.metabolite.

14. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39**, D685-90 (2011).

15. Kuperstein, I. *et al.* Atlas of Cancer Signalling Network: A systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis* (2015) doi:10.1038/oncsis.2015.19.

16. Olivier, B. G., Swat, M. J. & Moné, M. J. Modeling and simulation tools: From systems biology to systems medicine. *Methods Mol. Biol.* **1386**, 441–463 (2016).

17. Garfinkel, D. Construction of biochemical computer models. *FEBS Lett.* **2**, (1969).

18. Rolfsson, O., Palsson, B. T. & Thiele, I. The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions. *BMC Syst. Biol.* **5**, 1–16 (2011).

19. Dauga, D. Biocuration: a new challenge for the tunicate community. *Genesis* **53**, 132–42 (2015).

20. Novère, N. Le *et al.* The Systems Biology Graphical Notation. *Nature Biotechnology* (2009) doi:10.1038/nbt.1558.

21. Cohen, D., Kuperstein, I., Barillot, E., Zinovyev, A. & Calzone, L. From a biological hypothesis to the construction of a mathematical model. *Methods Mol. Biol.* **1021**, 107–25 (2013).

22. Glont, M. *et al.* BioModels: expanding horizons to include more modelling approaches and formats. *Nucleic Acids Res.* **46**, D1248–D1253 (2018).

23. Malik-Sheriff, R. S. *et al.* BioModels—15 years of sharing computational models in life science. *Nucleic Acids Res.* **48**, D407–D415 (2020).

24. Funahashi, A. *et al.* CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. *Proc. IEEE* **96**, 1254–1265 (2008).

25. Czauderna, T., Klukas, C. & Schreiber, F. Editing, validating and translating of SBGN maps. *Bioinformatics* **26**, 2340–1 (2010).

26. Kutmon, M. *et al.* PathVisio 3: An Extendable Pathway Analysis Toolbox. *PLOS Comput. Biol.* **11**, e1004085 (2015).

27. Balci, H. *et al.* Newt: a comprehensive web-based tool for viewing, constructing and analyzing biological maps. *Bioinformatics* **37**, 1475–1477 (2021).

28. Kono, N. *et al.* Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API. *PLoS One* **4**, (2009).

29. Flórez, L. A., Lammers, C. R., Michna, R. & Stülke, J. CellPublisher: a web platform for the intuitive visualization and sharing of metabolic, signalling and regulatory pathways. *Bioinformatics* **26**, 2997–2999 (2010).

30. Kuperstein, I. *et al.* NaviCell: a web-based environment for navigation, curation and maintenance of large molecular interaction maps. *BMC Syst. Biol.* **7**, 100 (2013).

31. Bonnet, E. *et al.* NaviCell Web Service for network-based data visualization. *Nucleic Acids Res.* (2015) doi:10.1093/nar/gkv450.

32. Gawron, P. *et al.* MINERVA—A platform for visualization and curation of molecular interaction networks. *npj Syst. Biol. Appl.* **2**, 1–6 (2016).

33. Nielsen, J. Systems Biology of Metabolism: A Driver for Developing Personalized and Precision Medicine. *Cell Metab.* **25**, 572–579 (2017).

34. Oulas, A. *et al.* Systems Bioinformatics: Increasing precision of computational diagnostics and therapeutics through network-based approaches. *Brief. Bioinform.* (2017) doi:10.1093/bib/bbx151.

35. Kanavakis, E. & Xaidara, A. The Human Genome Project. *Alcohol Health Res. World* **19**, 190 (1995).

36. Micheel, C. M. *et al.* Omics-Based Clinical Discovery: Science, Technology, and Applications. (2012).

37. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).

38. Evangelou, E. *Genetic Epidemiology. Methods and Protocols. Genetic Epidemiology.* vol. 38 (2018).

39. Pös, O. *et al.* DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomed. J.* **44**, 548 (2021).

40. Shao, X. *et al.* Copy number variation is highly correlated with differential gene expression: A pan-cancer study. *BMC Med. Genet.* **20**, 1–14 (2019).

41. Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**, 676-689.e3 (2018).

42. Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. Transcriptomics technologies. *PLoS Comput. Biol.* **13**, (2017).

43. Dong, Z. C. & Chen, Y. Transcriptomics: advances and approaches. *Sci. China. Life Sci.* **56**, 960–967 (2013).

44. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 1–12 (2017).

45. Ahmed, R. *et al.* Single-Cell RNA Sequencing with Spatial Transcriptomics of Cancer Tissues. *Int. J. Mol. Sci. 2022, Vol. 23, Page 3042* **23**, 3042 (2022).

46. Du, J. *et al.* Advances in spatial transcriptomics and related data analysis strategies. *J. Transl. Med.* **21**, 330 (2023).

47. Bensimon, A., Heck, A. J. R. & Aebersold, R. Mass Spectrometry–Based Proteomics and Network Biology. *https://doi-org.proxy.insermbiblio.inist.fr/10.1146/annurev-biochem-072909-100424* **81**, 379–405 (2012).

48. Ebhardt, H. A., Root, A., Sander, C. & Aebersold, R. Applications of targeted proteomics in systems biology and translational medicine. *Proteomics* **15**, 3193–3208 (2015).

49. Iliuk, A. & Tao, W. A. Quantitative phospho-proteomics based on soluble nanopolymers. *Methods Mol. Biol.* **527**, (2009).

50. Leitner, A. & Lindner, W. Chemical tagging strategies for mass spectrometry-based phospho-proteomics. *Methods Mol. Biol.* **527**, 229–243 (2009).

51. Clish, C. B. Metabolomics: an emerging but powerful tool for precision medicine. *Cold Spring Harb. Mol. Case Stud.* **1**, a000588 (2015).

52. Alseekh, S. *et al.* Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat. Methods* **18**, 747–756 (2021).

53. Nagana Gowda, G. A. & Raftery, D. NMR Based Metabolomics. *Adv. Exp. Med. Biol.* **1280**, 19 (2021).

54. Yang, K. & Han, X. Lipidomics: Techniques, applications, and outcomes related to biomedical sciences. *Trends Biochem. Sci.* **41**, 954 (2016).

55. Miskad, U. A. *et al.* High PRL-3 expression in human gastric cancer is a marker of metastasis and grades of malignancies: An in situ hybridization study. *Virchows Arch.* **450**, 303–310 (2007).

56. Ma, Y. & Li, B. Expression of phosphatase of regenerating liver-3 in squamous cell carcinoma of the cervix. *Med. Oncol.* **28**, 775–80 (2011).

57. Radke, I. *et al.* Expression and prognostic impact of the protein tyrosine phosphatases PRL-1, PRL-2, and PRL-3 in breast cancer. *Br. J. Cancer 2006 953* **95**, 347–354 (2006).

58. Laurent, C. *et al.* High PTP4A3 phosphatase expression correlates with metastatic risk in uveal melanoma patients. *Cancer Res.* **71**, 666–74 (2011).

59. Zinovyev, A., Viara, E., Calzone, L. & Barillot, E. BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks. *Bioinformatics* **24**, 876–877 (2008).

60. Duciel, L., Monraz Gomez, L. C., Kondratova, M., Kuperstein, I. & Saule, S. The Phosphatase PRL-3 Is Involved in Key Steps of Cancer Metastasis. *J. Mol. Biol.* **431**, (2019).

61. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–74 (2011).

62. Modi, P. K., Jaiswal, S. & Sharma, P. Regulation of Neuronal Cell Cycle and Apoptosis by MicroRNA 34a. *Mol. Cell. Biol.* **36**, 84–94 (2015).

63. Martignetti, L., Calzone, L., Bonnet, E., Barillot, E. & Zinovyev, A. ROMA: Representation and quantification of module activity from target expression data. *Front. Genet.* **7**, 1–12 (2016).

64. Hokama, M. *et al.* Altered expression of diabetes-related genes in Alzheimer's disease brains: The Hisayama study. *Cereb. Cortex* **24**, 2476–2488 (2014).

65. Liang, W. S. *et al.* Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. *Proc. Natl. Acad. Sci.* **105**, 4441–4446 (2008).

66. Sánchez-Valle, J. *et al.* A molecular hypothesis to explain direct and inverse co-morbidities between Alzheimer's Disease, Glioblastoma and Lung cancer. *Sci. Rep.* (2017) doi:10.1038/s41598-017-04400-6.

67. Musicco, M. *et al.* Inverse occurrence of cancer and Alzheimer disease: A population-based incidence study. *Neurology* **81**, 322–328 (2013).

68. Ou, S. M. *et al.* Does Alzheimer's disease protect against cancers? A nationwide population-based study. *Neuroepidemiology* **40**, 42–49 (2012).

69. Patel, M. N. *et al.* Inflammasome Priming in Sterile Inflammatory Disease. *Trends Mol. Med.* **23**, 165–180 (2017).

70. Mamik, M. K. & Power, C. Inflammasomes in neurological diseases: Emerging pathogenic and therapeutic concepts. *Brain* **140**, 2273–2285 (2017).

71. Kaushal, V. *et al.* Neuronal NLRP1 inflammasome activation of Caspase-1 coordinately regulates inflammatory interleukin-1-beta production and axonal degeneration-associated Caspase-6 activation. *Cell Death Differ.* **22**, 1676–1686 (2015).

72. Schmid-Burgk, J. L. *et al.* A genome-wide CRISPR (clustered regularly interspaced short palindromic repeats) screen identifies NEK7 as an essential component of NLRP3 inflammasome activation. *J. Biol. Chem.* **291**, 103–109 (2016).

73. Ojala, J. *et al.* Expression of interleukin-18 is increased in the brains of Alzheimer's disease patients. *Neurobiol. Aging* **30**, 198–209 (2009).

74. Kajiwara, Y. *et al.* The human-specific CASP4 gene product contributes to Alzheimer-related synaptic and behavioural deficits. *Hum. Mol. Genet.* **25**, 4315–4327 (2016).

75. P.N., P. *et al.* Caspase gene expression in the brain as a function of the clinical progression of Alzheimer disease. *Arch. Neurol.* **60**, 369–376 (2003).

76. Saresella, M. *et al.* The NLRP3 and NLRP1 inflammasomes are activated in Alzheimer's disease. *Mol. Neurodegener.* **11**, 1–14 (2016).

77. Giampietri, C. *et al.* Cancer microenvironment and endoplasmic reticulum stress response. *Mediators Inflamm.* **2015**, (2015).

78. Avril, T., Vauléon, E. & Chevet, E. Endoplasmic reticulum stress signaling and chemotherapy resistance in solid cancers. *Oncogenesis* **6**, e373 (2017).

79. Salaroglio, I. C. *et al.* PERK induces resistance to cell death elicited by endoplasmic reticulum stress and chemotherapy. *Mol. Cancer* **16**, 91 (2017).

80. Tufo, G. *et al.* The protein disulfide isomerases PDIA4 and PDIA6 mediate resistance to cisplatin-induced cell death in lung adenocarcinoma. *Cell Death Differ.* **21**, 685–695 (2014).

81. Xia, Y., Shen, S. & Verma, I. M. NF- B, an Active Player in Human Cancers. *Cancer Immunol. Res.* **2**, 823–830 (2014).

82. Logue, S. E., Cleary, P., Saveljeva, S. & Samali, A. New directions in ER stress-induced cell death. *Apoptosis* **18**, 537–46 (2013).

83. Välk, K. *et al.* Gene expression profiles of non-small cell lung cancer: Survival prediction and new biomarkers. *Oncology* **79**, 283–292 (2011).

84. Ravel, J. M. *et al.* Comprehensive Map of the Regulated Cell Death Signaling Network: A Powerful Analytical Tool for Studying Diseases. *Cancers (Basel).* **12**, (2020).

CHAPTER II. TRANSCRIPTOMICS PROFILING OF ADIPOSE TISSUES IN BREAST CANCER

A. INTRODUCTION

Breast cancer is the most common type of cancer in females. According to the GLOBOCAN, in 2020, 2.261.419 new cases of breast cancer were reported and 684.996 deaths in the same year were attributed to breast cancer world-wide [1]. Among the risk factors to develop this disease are: Female gender, age, abusive alcohol consumption, family history of breast cancer, exposure to radiation, tobacco smoking, use of postmenopausal hormone therapy and obesity [2].

The female breast could be defined as a cutaneous exocrine gland [3] The breast lies on the anterior thoracic wall between the 2nd and the 6th costal cartilages with the sternum medially and the midaxillary line laterally. The breast consists of skin and subcutaneous tissue, breast parenchyma (ducts and lobules), and supporting stroma [4]. In the breast parenchyma, we can find the glandular tissue, that extends to 10 to 20 lobes. Each lobe is then subdivided into lobule and acini. From each lobule a lactiferous duct leads to the nipple through the lactiferous sinus. The supporting stroma can be divided into fibrous stroma, mainly formed by the suspensory ligaments of Cooper and the fatty stroma, subcutaneous adipose tissue, that gives the main volume to the breast [5].

Breast cancer can develop in any part of the breast. Nevertheless, the most common anatomical sites are the ducts and the lobules, being the first the most common type. Histologically, the breast cancers can be classified regarding if the cells have already invaded the basement membrane or not. Being determined *in situ* for those tumors that have not yet invaded the basal membrane and invasive, that are those that have spread to the basal membrane and cells can be found in another parts of the breast, lymphatic system or even have touched the bloodstream. Since this chapter is focused in the ductal carcinoma, with the aforementioned we can classify them as ductal carcinoma *in situ* (DCIS) or invasive ductal carcinoma (IDC) [6,7].

The invasive ductal carcinomas, usually are subclassified in order to distinguish them and to guide the treatment management. There are many classifications, however, for this chapter we will take four subtypes: Luminal A, Luminal B, Her2-enriched and Triple Negative, that are the most commonly used in the clinical practice. Luminal A tumors are rich in estrogen receptors (ER), correspond to low grade tumors and usually with good prognosis, in addition to ER, progesterone receptor (PR) is highly expressed and the human epidermal growth factor receptor 2 (HER2) is lowly expressed and also they present low Ki67 expression. Luminal B tumors can be positive for ER and PR, at a lower extent than luminal A tumors, and HER2 negative or positive, they also present high Ki67 and are classified as a higher grade than luminal A tumors. HER2-enriched tumors are considered highly aggressive and present a positive expression of the HER2 protein and negative expression of hormone receptors (ER/PR) and with high Ki67. Triple negative tumors are the most aggressive and they are negative for all the receptors ER/PR/HER2 and with high Ki67 [6,8].

The tumor microenvironment (TME) has been widely studied in different cancer types, specially targeting immune infiltrates and fibroblasts. In recent years, the adipose tissue has gained interest, since it has been associated to the initiation and progression of different cancers, like breast cancer. The term Cancer associated adipocytes (CAA) was coined to refer to those adipocytes that are in contact with the tumor cells and have bi-directional interactions [9,10]. It has been observed that this adipocytes can secrete proinflammatory cytokines like IL-6 and Prostaglandin E2, Adipokines, specially leptin, that stimulates the

aromatase expression, lipids that serve as fuel substrate and structural elements for tumor cells [11,12]. The prolonged interaction between adipocytes and tumor cells is believed to transform the mature adipocytes, by de-differentiation, into adipocyte derived fibroblasts [12,13]. A process called "browning" of the white adipocytes has been suggested to be the responsible of the transformation in these cells, and experimental evidence has shown an increase in expression of UCP-1, PGC1α, TOMM20, PLIN1, and HSL that indicate this process to occur [14].

The adipose tissue has been greatly studied in the field of metabolic diseases, especially in obesity. Obesity can be defined as an excess of body weight at the expense of adipose tissue. In order to classify obesity, one of the most used indicators is the body mass index (BMI) the quotient resulting from dividing the body weight in kg between the body surface (height squared in meters). According to the WHO, a person with a BMI ≥ 30 kg/m^2 is considered obese (see Table 1) [15]. It has been observed that obesity can induce a low-grade chronic inflammation through cytokine excretion [16]. This chronic inflammation can lead to recruitment of immune cells, especially macrophages, that result in the "crown-like structures" where macrophages surround the adipocytes and can induce cell death [17]. This low-grade chronic inflammation has been suggested as one of the inducers of different metabolic comorbidities and cancers, including breast cancer [18–20].

BMI	Nutritional status
< 18.5	Underweight
18.5 – 24.9	Normal-weight
25.0 – 29.9	Overweight
30.0 – 34.9	Obesity class I
35.0 – 39.9	Obesity class II
> 40	Obesity class III

Table 1. World health organization (WHO) BMI categories. The classification is meant to be applied on adults over 20 years old.

Taking into account all this information, we came to the hypothesis that the regulators of metabolism and signaling mechanisms in the Tumor Microenvironment components, especially Cancer Associated Adipocytes, provide markers of invasion and metabolic remodeling in breast cancer. In order to explore this hypothesis, we have set a collaborative project in 2019, called LipoCanPredict, where we have a partnership between the Institute Curie and the Saint-Antoine Research Center (CRSA). In this project we aim to explore the CAA role by taking patient biopsies to perform multi-omics (transcriptomics and lipidomics/metabolomics) analyses in the tissues but also we have set co-culture experiments to study more in detail the direct interactions.

In this chapter, I will expose the obtained results from one of the LipoCanPredict branches, the analysis of transcriptomics data from patient biopsies with breast cancer. We have set a cohort of patients, obese or normal-weight, and taken samples from the tumor, as well as adipose tissues, one piece surrounding the tumor and one piece far from the tumor. This cross-sectional approach would give us a snapshot of the transcriptional patterns of the adipose tissues likely CAA signatures.

B. MATERIALS AND METHODS

Study subjects and tissue sampling

Tissue samples were obtained from patients undergoing mastectomy (partial or total) at the Curie Institute Hospital in Paris, France. The studied individuals were selected according to the criteria listed in Table 1. Only the patients that have signed the informed consent approving the use of their biological materials for scientific research were considered for the analysis. The clinical data as well as the tissue specimens from the patients, were obtained in collaboration with the pathology department at the Curie Institute Hospital.

Subsequent to the mastectomy, three different samples were collected as follows: 1) A sample of the tumor, 2) one piece of adipose tissue close to the tumor, no more than one centimeter away from the tumor, which we termed **proximal adipose tissue** and 3) one piece of adipose tissue "distant" from the tumor, at least 5 centimeters away from the tumor, which we termed **distal adipose tissue**. The adipose tissue distances were determined macroscopically. Then the samples were snap-frozen in liquid nitrogen and preserved at -80°C until manipulation.

Inclusion	**Exclusion**	**Elimination**
-Ductal carcinoma -Without neoadjuvant therapy or radiotherapy (surgery first)	-BMI < 18.5 -24.9 < BMI < 30 -Tumor size <10 mm	-Refused to sign the consent.

Table 2. Criteria for the patients in the study

Tissue preparation

The collected frozen samples were ground using porcelain mortar and pestle (Dutscher), until obtaining a powder consistency. Throughout the grinding process, the samples were maintained cold by addition of liquid nitrogen. Once the sample was satisfactory ground, it was divided into tubes and weighted.

RNA extraction and sequencing

To extract the RNA, samples were homogenized using TRIzol reagent (Invitrogen), then the contents were transferred to Phasemaker tubes (Quanta Bio). Then chloroform was added, the samples were incubated and centrifugated. The upper phase was then transferred to a RNeasy minikit column (QIAGEN), afterwards total RNA was extracted according to the manufacturer specifications. The RNA's quality and quantity were measured using nanodrop spectrophotometer (ThermoFisher Scientific) and Bioanalyzer 2000 (Agilent) before sequencing. The extracted RNA samples were sequenced at the CurieCoreTech Next Generation Sequencing (ICGEX) platform, at the Curie Institute, Paris, France. The samples were sequenced using the protocol 3' mRNA-Seq library Prep Kit - FWD-Lexogen. The raw data were processed using the pipeline "rawqc" version 2.2.0 (DOI 10.5281/zenodo.7515638), and the alignment was done using the reference genome GRCh38 (hg38).

Transcriptomics data analysis

All the analysis were performed using R software (v4.2.0). For the gene expression normalization and differential analysis, the packages “DESeq”, “edgeR” and “Limma” were used. For the exploratory analysis, the gene expression matrix was reduced to use the protein coding genes, according to the HUGO Gene Nomenclature Committee (HGNC). The data was normalized using a variance stabilizing transformation (VST). The exploratory analyses were performed using principal component analysis (PCA) and euclidean distances to cluster the samples and measure the gene expression divergence. For Differential Expression analysis, the raw counts were scaled using the trimmed mean of M-values (TMM) normalization method, followed by a voom transformation with a duplicated correlation for fixed effects. Differentially expressed genes had an adjusted p value < 0.05 and an absolute log2 Fold change (log2FC) > 1. The functional analysis was performed taking the differentially expressed genes log2FC using the packages “org.Hs.eg.db", “clusterProfiler" and "AnnotationDbi" with Gene Set Enrichment Analysis (GSEA) to determine enriched pathways using the databases Reactome, KEGG, WikiPathways and the Hallmarks genesets from the molecular signatures database (MSigDB).

CAA mapped pathways

From the previous analyses, according to the literature and to those enriched pathways from the analyses, 21 pathways from WikiPathways and KEGG were selected, the genes were extracted and used to perform further analyses to the transcriptomics data (see Table 16). These pathways were taken as relevant to describe the CAAs.

Representation and Quantification of Module Activity (ROMA) analysis

The ROMA analysis[21] was used to quantify the activity of the different pathways represented in the CAA mapped pathways. The analysis was done using the “rROMA” package in R[22]. Input data was previously normalized using VST. ROMA implements a simplest uni-factor linear model of gene regulation that estimates the expression data of a gene set by its first principal component (PC1). In this algorithm, a random gene set procedure is used to generate a null distribution for the L_1 amount of variance described by the PC1 and computes the p-value by comparing the obtained L_1 to the null distribution. From these calculations, the *overdispersed* modules were identified, corresponding to modules where the amount of variance explained by PC1 calculated for the genes in the module is significantly greater than the variance of a random gene set with the same amount of genes. Top contributing genes were calculated using Pearson’s correlation using the gene expression and the scores provided by the ROMA analysis.

Reproducibility

The files containing the used code for the different analyses, are available in the LipoCanPredict repository on GitHub: https://github.com/sysbio-curie/LipoCanPredict.

C. RESULTS

Population characteristics

We have sequenced 218 samples in total for all tissues. However, the downstream analysis was performed in 171 samples, corresponding to 57 patients where the three tissues were successfully sequenced. The population characteristics can be found in the Table 1.

VARIABLE	OBESE (N = 24)		NORMAL-WEIGHT (N = 33)		TOTAL (N = 57)	
	Mean	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**
AGE (YEARS)	59.75	(±12.60)	58.82	(± 16.96)	59.21	(± 15.16)
BMI (KG/M^2)	33.42	(± 3.13)	22.25	(± 1.74)		
	Frequency	**Proportion**	**Frequency**	**Proportion**	**Frequency**	**Proportion**
MOLECULAR CLASSIFICATION						
INVASIVE	24	100 %	28	85 %	52	91 %
LUMINAL A	14	58 %	13	39 %	27	47 %
LUMINAL B	9	38 %	9	27 %	18	32 %
HER-2 ENRICHED	0	0 %	1	3 %	1	1.8 %
TRIPLE NEGATIVE	1	4.2 %	5	15 %	6	11 %
NON-INVASIVE						
DCIS	0	0 %	5	15 %	5	8.8 %
GRADE (FOR INVASIVE)						
GRADE I	4	17 %	10	36 %	14	27 %
GRADE II	16	67 %	14	50 %	30	58 %
GRADE III	4	17%	4	14 %	8	15 %

Table 3. Characteristics of the studied population. the quantitative variables are described with means and standard deviations (S.D.), whereas the qualitative variables are described as frequencies and proportions, the latest is taken as percentage of the corresponding category, i.e. obese, normal-weight or total.

As we can see in Table 1, the age in both groups appears to be somewhat evenly distributed. The majority of the patients in the cohort correspond to invasive ductal carcinomas (IDC). Frome these, the luminal A and B subtypes are the most predominant, taken together 79% of the cohort. In the invasive subtypes, we can observe that majority of the patients, were classified as to have a grade II. Regarding to the non-invasive ductal carcinomas, we only had 5 patients in the cohort and all of them were Normal-weight in the BMI classification.

Exploratory RNA-seq analyses

For all the individuals, we explored the data as per tissue sampled. For this, a principal component analysis (PCA) has been performed to plot the data regarding the variance of the samples. In this PCA we have observed a tendency of those samples corresponding to tumor tissue and distal adipose tissue to separate from each other, as expected. The samples corresponding to the proximal adipose tissue, show a spreading tendency between the distal adipose tissue and tumor samples (Figure 1). Additionally, we also plotted the PCA to explore the dispersion regarding the BMI categories (Normal weight and obese), BMI and tissue, molecular classification of the tumor, however these variables does not appear to be responsible for the variance in the gene expression in this cohort (Supplementary figures 1-3). Then we wanted to explore the effects of the fact that the samples belong to the same patient could potentially explain the behavior in the PCA For this, we plotted the Euclidean distances of all the samples (Figure 2). In this heatmap we could observe a clustering mainly composed of samples coming from the tumor samples

(from the bottom right corner going up and left). Then, there is a collection of samples clustering together, mostly proximal and distal adipose tissues, with some tumor samples following in this cluster. As expected, some samples coming from the same patient, specifically for adipose tissue samples, were clustering together. For further analyses, we would have to take this into account.

Since this cohort has some samples corresponding to different tumor types (see Table 1), we decided to explore the behavior of the data in a subset of samples. To do so, we took the samples only from the patients that had a tumor classified as Luminal A or Luminal B. Hence, this subset of samples correspond to invasive carcinomas with the better prognosis. This subset (n = 45), is composed by samples of 23 obese and 22 normal-weight patients. After running a PCA, we observed a similar pattern as what we observed with the first analysis (Figure 3). Nevertheless, when plotting the first two dimensions of the PCA, we could observe how the different tissues were dispersed, specially the majority of the distal adipose tissues and the tumors. The proximal adipose tissues appear to be dispersed in between both groups, as we observed with the complete cohort. Additionally, the percentage of variance of the first dimension has dramatically increased after sub setting the samples. As done for the general samples, we also explored for the BMI categories, BMI and tissue, molecular classification of the tumor, though, not striking differences were found (Supplementary figures 4-6).

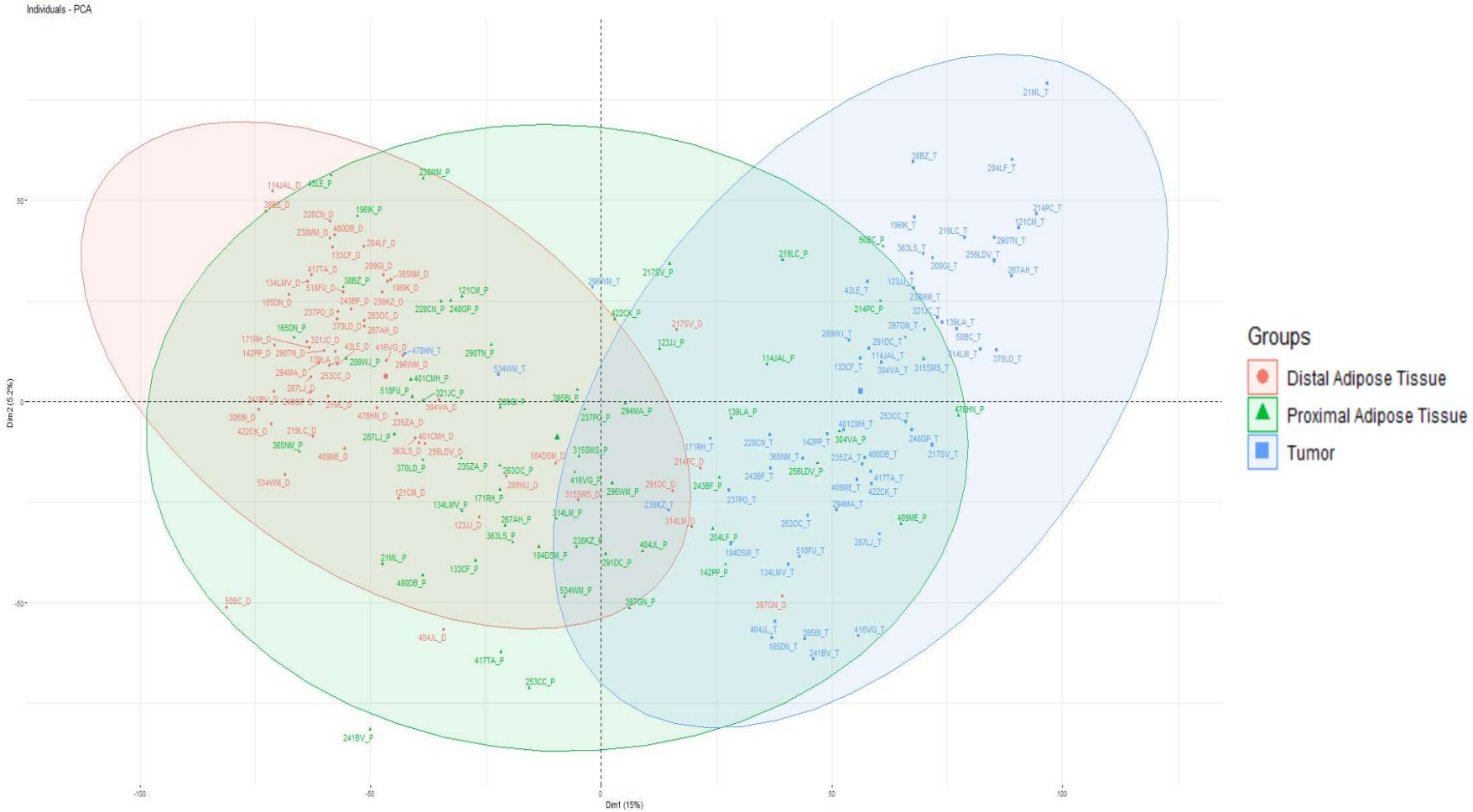


Figure 1. Principal component analysis plot of all the samples for the different tissue types.

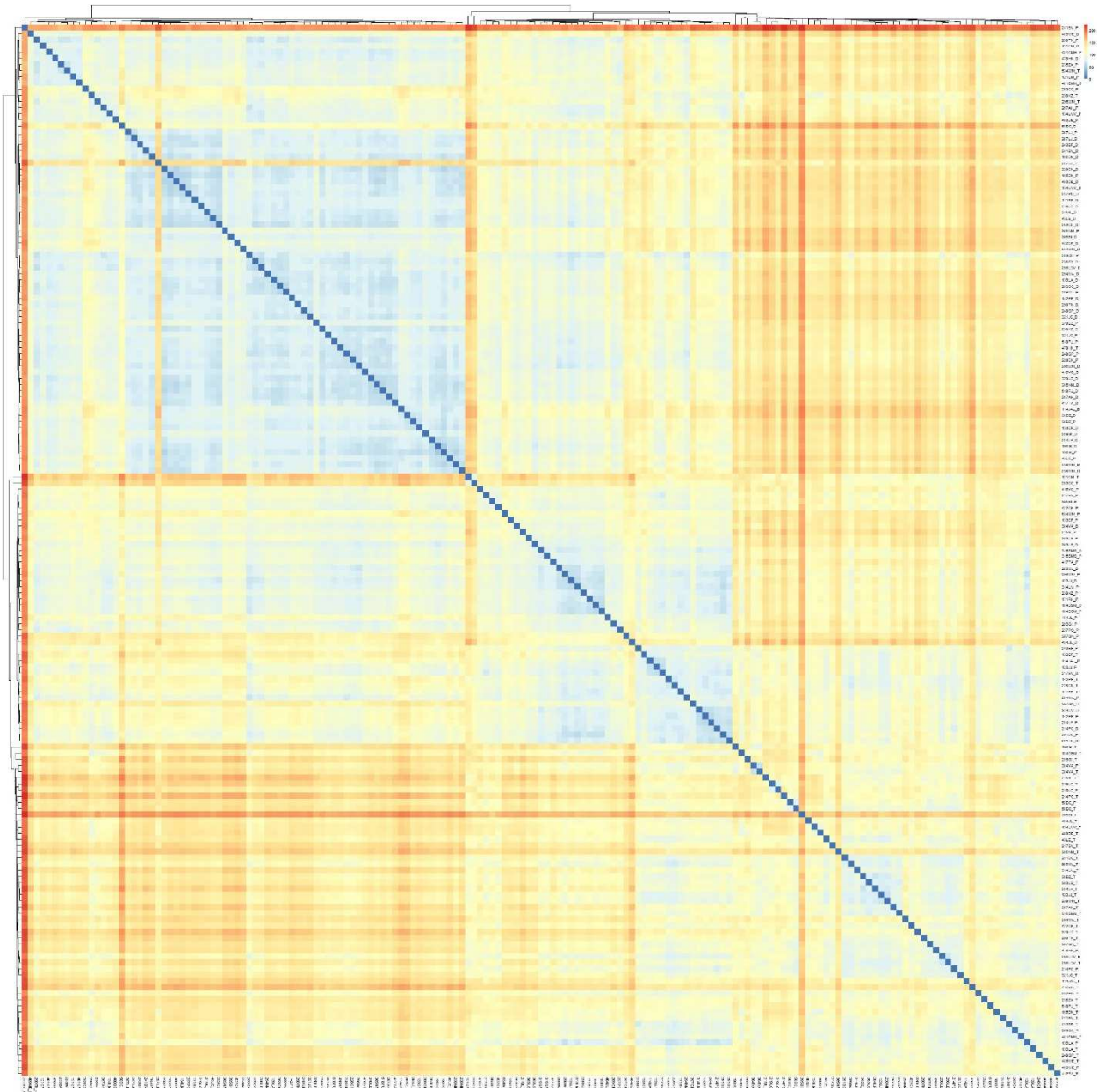


Figure 2. Heatmap corresponding to the Euclidean distances in all the patients for the three tissues.

Figure 3. PCA plot corresponding to the subset of samples (luminal subtypes) for the three different tissue types.

Differential expression and functional analysis

For the differential expression analysis, first we took into account all the samples available in the cohort (n = 57). The comparisons made to obtain the differentially expressed genes (DEG) were: 1) proximal adipose tissue to distal adipose tissue, 2) proximal adipose tissue to tumor tissue and 3) distal adipose tissue to tumor tissue. In the first comparison 1241 DEG were obtained, on the second 2771 DEG were obtained and in the third comparison we obtained 4948 DEG. The overlaps in the three comparisons as well as unique genes per comparison are summarized in the Venn diagram corresponding to Figure 4.

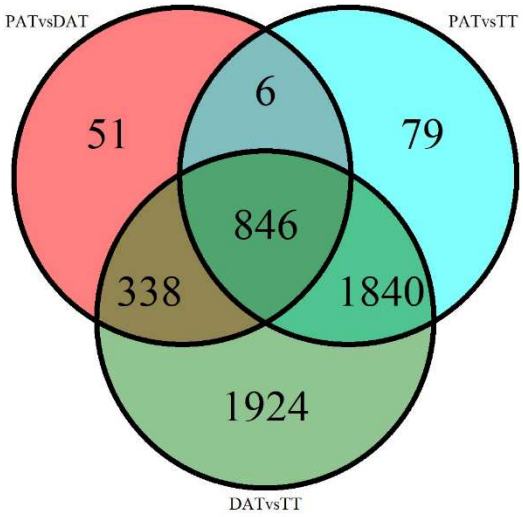


Figure 4. Venn diagram representing the differentially expressed genes across the different comparisons per tissue type. DAT: distal adipose tissue, PAT: proximal adipose tissue, TT tumor tissue

Using the lists of DEG we could then perform the enrichment analyses to see the processes that are related to these genes. Firstly, we have queried for all the comparisons using gene ontology (GO) terms for biological processes (BP). The enriched terms were quite broad and the terms that might be of interest got shadowed in the enriched terms results (data not shown). Then, we decided to perform the enrichment using the following gene sets from the Molecular Signatures Database (MSigDB): Hallmarks, KEGG gene-sets, Reactome gene-sets, and the sets from WikiPathways.

For the proximal adipose tissues as compared to the distal adipose tissues, the pathways that were enriched, from the hallmark gene-sets, were estrogen response for the proximal and the adipogenesis hallmark for the distal adipose tissues (see Table 4). For the KEGG gene sets, only the **PPAR signaling pathway** appear to be enriched in the distal adipose tissue (see Table 5). Then we have the enrichment from Reactome, in which the gene-sets enriched in the proximal adipose tissue seem to be related to cell development whereas in the distal adipose tissue we observed more canonical pathways as **regulation of adipocyte differentiation** or **G alpha (s) signaling events**, related to the Protein Kinase A (see Table 6). When we enriched the gene-sets using WikiPathways, we retrieved very few, the most enriched gene set for proximal adipose tissue was again a pathway related to development, whereas the distal adipose tissue have enriched the **PPAR signaling pathway** as when used the KEGG gene sets and **adipogenesis** as with the hallmarks gene sets (Table 7). The results of the enrichment can indicate the similarities between both tissues but also how the distal adipose tissues have strong enrichment for canonical adipose tissue pathways, as expected.

Gene Set	Enrichment Score	NES	p-value	Adjusted p-value
HALLMARK_ESTROGEN_RESPONSE_EARLY	0.50641563	2.89583977	1.00164812568551e-08	1.35222496967544e-07
HALLMARK_ESTROGEN_RESPONSE_LATE	0.42718786	2.45696096	1.10926328236592e-05	9.98336954129324e-05
HALLMARK_ADIPOGENESIS	-0.8361858	-3.9179996	1.11799596247795e-10	3.01858909869045e-09

Table 4. Enrichment results from the hallmarks gene-sets for the comparison of proximal to distal adipose tissue. Positive values in enrichment score correspond to proximal adipose tissue and negative values to distal adipose tissue. NES: Normalized Enrichment Score.

Gene Set	Enrichment Score	NES	p-value	Adjusted p-value
KEGG_PPAR_SIGNALING_PATHWAY	-0.5385151	-2.2656982	0.00051504	0.01236107

Table 5. Enrichment results from the KEGG gene-sets for the comparison of proximal to distal adipose tissue. Positive values in enrichment score correspond to proximal adipose tissue and negative values to distal adipose tissue. NES: Normalized Enrichment Score.

Gene Set	Enrichment Score	NES	p-value	Adjusted p-value
REACTOME_FORMATION_OF_THE_CORNIFIED_ ENVELOPE	0.564686023	2.342612435	9.90151744518567e-05	0.004554698
REACTOME_KERATINIZATION	0.564686023	2.342612435	9.90151744518567e-05	0.004554698
REACTOME_ESR_MEDIATED_SIGNALING	0.495282491	2.054690349	0.00159285	0.039124398
REACTOME_RHO_GTPASE_CYCLE	0.461659093	2.062440667	0.001984571	0.039124398
REACTOME_SIGNALING_BY_RHO_GTPASES_ MIRO_GTPASES_AND_RHOBTB3	0.346115186	1.974704343	0.001681476	0.039124398
REACTOME_TRANSPORT_OF_SMALL_ MOLECULES	-0.185508003	-1.932895669	0.002863939	0.049402954
REACTOME_G_ALPHA_S_SIGNALLING_EVENTS	-0.473993917	-2.090887184	0.001951802	0.039124398
REACTOME_TRANSCRIPTIONAL_REGULATION_O F_WHITE_ADIPOCYTE_DIFFERENTIATION	-0.746684082	-2.890547816	5.06369805239322e-06	0.00069879

Table 6. Enrichment results from the Reactome gene-sets for the comparison of proximal to distal adipose tissue. Positive values in enrichment score correspond to proximal adipose tissue and negative values to distal adipose tissue. NES: Normalized Enrichment Score.

Gene Set	Enrichment Score	NES	p-value	Adjusted p-value
ECTODERM DIFFERENTIATION-WIKIPATHWAYS_20230610-WP2858-HOMO SAPIENS	0.593444616	2.081134426	0.001787658	0.034561389
PPAR SIGNALING PATHWAY-WIKIPATHWAYS_20230610-WP3942-HOMO SAPIENS	-0.538515088	-2.260688838	0.001196317	0.034561389
ADIPOGENESIS-WIKIPATHWAYS_20230610-WP236-HOMO SAPIENS	-0.542514429	-2.969711233	5.9081957685596e-06	0.000342675

Table 7. Enrichment results from the WikiPathways gene-sets for the comparison of proximal to distal adipose tissue. Positive values in enrichment score correspond to proximal adipose tissue and negative values to distal adipose tissue. NES: Normalized Enrichment Score.

After comparing both adipose tissue types, we also compared the proximal adipose tissue to the tumor tissue samples. We used the same gene-sets to do this comparison. From the hallmarks gene sets, for the proximal adipose tissues we observed an enrichment of classical adipose tissue processes but also others as **TNF signaling** and **hypoxia** which might indicate stress or inflammation in the case of the tumors we observed mostly cell cycle related processes enriched (Table 8). For the KEGG gene sets we observed very similar enriched elements in proximal adipose tissues as in the comparison with hallmarks, with metabolic pathways related to adipose tissue, also the **metabolism of xenobiotics by cytochrome p450**.In the case of the tumor tissue we observed pathways very well known for tumor development, such as **P53 signaling pathway** and **cell cycle** (Table 9). When we did the enrichment using the Reactome gene sets, we obtained 141 pathways enriched (data not shown), most of the pathways were complementary or redundant to others in the results, therefore, the top and bottom 15 pathways are shown

in Table 10. For the proximal adipose tissue the enriched pathways are classical for adipose tissue, however there are also some related to xenobiotic metabolism as previously described, whereas in the tumor tissue we have cell cycle related pathways. In the case of WikiPathways we obtained 45 enriched pathways (see Table 11) from which 32 were enriched in the proximal adipose tissue. From these pathways that many are expected for adipose tissue it is also interesting to see pathways like **VEGFA-VEGFR2 signaling**, **Angiopoietin-like protein 8 regulatory pathways** that are related to angiogenesis and lipid metabolism, as well as the prostaglandin synthesis.

Gene Set	Enrichment Score	NES	p-value	Adjusted p-value
HALLMARK_ADIPOGENESIS	0.592816899	3.439700946	1e-10	1.33333333333333e-09
HALLMARK_FATTY_ACID_METABOLISM	0.525217835	2.625803161	5.93279921124663e-06	3.95519947416442e-05
HALLMARK_HYPOXIA	0.432862173	2.532177823	4.60834062436205e-06	3.68667249948964e-05
HALLMARK_XENOBIOTIC_METABOLISM	0.42518457	2.283951602	0.000193547	0.000967734
HALLMARK_MYOGENESIS	0.422922072	2.536513021	8.0476092171551e-06	4.59863383837434e-05
HALLMARK_BILE_ACID_METABOLISM	0.408208895	1.793094636	0.01385765	0.039593286
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.381988912	2.21641391	0.000235116	0.001044959
HALLMARK_UV_RESPONSE_DN	0.373025681	1.966232667	0.002471064	0.007603275
HALLMARK_KRAS_SIGNALING_UP	0.347814614	2.002296885	0.002148184	0.007160614
HALLMARK_ESTROGEN_RESPONSE_EARLY	-0.296601522	-2.073007256	0.000402302	0.001609209
HALLMARK_SPERMATOGENESIS	-0.472069188	-2.243436024	0.00062437	0.002270435
HALLMARK_MITOTIC_SPINDLE	-0.477908112	-2.762758717	1.63305354124286e-06	1.63305354124286e-05
HALLMARK_E2F_TARGETS	-0.567810993	-3.907446107	1e-10	1.33333333333333e-09
HALLMARK_G2M_CHECKPOINT	-0.576811302	-3.987530016	1e-10	1.33333333333333e-09

Table 8. Enrichment results from the Hallmarks gene-sets for the comparison of proximal adipose tissue to tumor tissue. Positive values in enrichment score correspond to proximal adipose tissue and negative values to tumor tissue. NES: Normalized Enrichment Score.

Gene Set	Enrichment Score	NES	p-value	Adjusted p-value
KEGG_PPAR_SIGNALING_PATHWAY	0.688862657	2.907624729	0.001655629	0.025728988
KEGG_ADIPOCYTOKINE_SIGNALING_PATHWAY	0.681148487	2.634062616	0.001686341	0.025728988
KEGG_METABOLISM_OF_XENOBIOTICS_BY_CYTOCHROME_P450	0.644427308	2.260776745	0.001686341	0.025728988
KEGG_RETINOL_METABOLISM	0.611577047	2.252571206	0.001715266	0.025728988
KEGG_DRUG_METABOLISM_CYTOCHROME_P450	0.550572907	2.027879699	0.003430532	0.032161235
KEGG_GLYCEROLIPID_METABOLISM	0.542253231	1.902330147	0.006745363	0.042158516
KEGG_INSULIN_SIGNALING_PATHWAY	0.541152091	2.307791449	0.001647446	0.025728988
KEGG_CALCIUM_SIGNALING_PATHWAY	0.393439526	1.903940002	0.004754358	0.035657686
KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	0.358127306	2.039867287	0.007407407	0.042735043
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS	-0.359051518	-1.83836394	0.005390836	0.036755697
KEGG_OOCYTE_MEIOSIS	-0.495597084	-2.335781889	0.002531646	0.029844807
KEGG_CELL_CYCLE	-0.549404499	-3.193894061	0.002785515	0.029844807
KEGG_P53_SIGNALING_PATHWAY	-0.568109624	-2.053654474	0.004672897	0.035657686

Table 9. Enrichment results from the KEGG gene-sets for the comparison of proximal adipose tissue to tumor tissue. Positive values in enrichment score correspond to proximal adipose tissue and negative values to tumor tissue. NES: Normalized Enrichment Score.

Gene Set	Enrichment Score	NES	p-value	Adjusted p-value
REACTOME_TRANSCRIPTIONAL_REGULATION_OF_WHITE_ADIPOCYTE_DIFFERENTIATION	0.846577406	3.275101501	1.00E-10	5.37E-09
REACTOME_TRIGLYCERIDE_METABOLISM	0.813835567	2.593382728	8.08E-07	1.31E-05
REACTOME_PLASMA_LIPOPROTEIN_REMODELING	0.745883459	2.376845347	3.47E-05	0.000310805
REACTOME_PLASMA_LIPOPROTEIN_ASSEMBLY_REMODELING_AND_CLEARANCE	0.687389382	2.442628393	3.14E-05	0.000288565
REACTOME_BINDING_AND_UPTAKE_OF_LIGANDS_BY_SCAVENGER_RECEPTORS	0.632579861	2.496768185	1.96E-05	0.000197528
REACTOME_COMPLEMENT_CASCADE	0.606392048	1.932339563	0.005490483	0.015356308
REACTOME_METABOLISM_OF_FAT_SOLUBLE_VITAMINS	0.588246449	2.27571255	0.000521902	0.002100655
REACTOME_SIGNALING_BY_RETINOIC_ACID	0.575522938	2.045112575	0.0019353	0.006294612
REACTOME_RA_BIOSYNTHESIS_PATHWAY	0.566171634	1.893967661	0.008832577	0.022572142
REACTOME_OPIOID_SIGNALLING	0.560964858	1.993380643	0.003357801	0.010104784

REACTOME_DRUG_ADME	0.560535129	2.067879866	0.003465889	0.010333485
REACTOME_INTERACTION_BETWEEN_L1_AND_ANKYRINS	0.519781525	1.811503294	0.016697327	0.039826217
REACTOME_SIGNALING_BY_TYPE_1_INSULIN_LIKE_ GROWTH_FACTOR_1_RECEPTOR_IGF1R	0.517969349	1.805187633	0.01723163	0.04050062
REACTOME_CARGO_RECOGNITION_FOR_CLATHRIN_ MEDIATED_ENDOCYTOSIS	0.511741626	2.083507556	0.001508555	0.005059945
REACTOME_METABOLISM_OF_VITAMINS_AND_ COFACTORS	0.50580945	2.503615272	1.93E-05	0.000197528
REACTOME_SYNTHESIS_OF_DNA	-0.5992411	-2.5914813	1.62E-05	0.00018577
REACTOME_DNA_METHYLATION	-0.608348	-2.5013857	5.78E-05	0.00044348
REACTOME_PRC2_METHYLATES_HISTONES_AND_DNA	-0.6086519	-2.5026353	5.72E-05	0.00044348
REACTOME_S_PHASE	-0.6115992	-2.9890456	1.06E-07	2.60E-06
REACTOME_REGULATION_OF_TP53_ACTIVITY_THROUGH _PHOSPHORYLATION	-0.6143788	-2.1437013	0.00077545	0.0029376
REACTOME_MITOTIC_G2_G2_M_PHASES	-0.6159621	-2.9715651	7.14E-08	2.09E-06
REACTOME_FORMATION_OF_THE_CORNIFIED_ENVELOPE	-0.6169531	-2.536768	3.85E-05	0.00032589
REACTOME_KERATINIZATION	-0.6169531	-2.536768	3.85E-05	0.00032589
REACTOME_ASSEMBLY_OF_THE_ORC_COMPLEX_AT_THE _ORIGIN_OF_REPLICATION	-0.6208597	-2.6256665	2.31E-05	0.00022569
REACTOME_CONDENSATION_OF_PROPHASE_ CHROMOSOMES	-0.6312121	-2.7297433	2.86E-06	4.19E-05
REACTOME_SWITCHING_OF_ORIGINS_TO_A_POST_ REPLICATIVE_STATE	-0.6796124	-2.2934147	0.00041287	0.00174926
REACTOME_CYCLIN_A_B1_B2_ASSOCIATED_EVENTS_ DURING_G2_M_TRANSITION	-0.7274622	-2.4548885	6.79E-05	0.00049715
REACTOME_G0_AND_EARLY_G1	-0.7558855	-2.5508056	1.77E-05	0.00019026
REACTOME_G1_S_SPECIFIC_TRANSCRIPTION	-0.7564335	-2.7156357	8.39E-07	1.31E-05
REACTOME_APC_C_MEDIATED_DEGRADATION_OF_CELL_ CYCLE_PROTEINS	-0.7689318	-2.7605051	3.51E-07	7.06E-06

Table 10. Enrichment results from the Reactome gene-sets for the comparison of proximal adipose tissue to tumor tissue. The enriched pathways in the table correspond to the top and bottom 15 pathways enriched. Positive values in enrichment score correspond to proximal adipose tissue and negative values to tumor tissue. NES: Normalized Enrichment Score.

Gene Set	Enrichment Score	NES	p-value	Adjusted p-value
DIFFERENTIATION OF WHITE AND BROWN ADIPOCYTE-WIKIPATHWAYS_20230610-WP2895-HOMO SAPIENS	0.734251656	2.7638847	4.50E-07	2.12E-05
FAMILIAL PARTIAL LIPODYSTROPHY-WIKIPATHWAYS_20230610-WP5102-HOMO SAPIENS	0.714615012	2.649364848	4.29E-06	0.000115168
TRIACYLGLYCERIDE SYNTHESIS-WIKIPATHWAYS_20230610-WP325-HOMO SAPIENS	0.709865311	2.39281935	3.40E-05	0.000709575
TRANSCRIPTION FACTOR REGULATION IN ADIPOGENESIS-WIKIPATHWAYS_20230610-WP3599-HOMO SAPIENS	0.697071639	2.349694346	4.64E-05	0.000872611
FERROPTOSIS-WIKIPATHWAYS_20230610-WP4313-HOMO SAPIENS	0.689890021	2.235993353	0.00063378	0.00541594
PPAR SIGNALING PATHWAY-WIKIPATHWAYS_20230610-WP3942-HOMO SAPIENS	0.67947524	2.894036361	1.30E-07	8.16E-06
VITAMIN B12 METABOLISM-WIKIPATHWAYS_20230610-WP1533-HOMO SAPIENS	0.665709755	2.310243532	0.000208064	0.002793997
AMP-ACTIVATED PROTEIN KINASE SIGNALING-WIKIPATHWAYS_20230610-WP1403-HOMO SAPIENS	0.656288108	2.338864494	0.000160292	0.00273954
GALANIN RECEPTOR PATHWAY-WIKIPATHWAYS_20230610-WP4970-HOMO SAPIENS	0.651240238	2.110726058	0.0021074	0.014673749
PROSTAGLANDIN SYNTHESIS AND REGULATION-WIKIPATHWAYS_20230610-WP98-HOMO SAPIENS	0.63788216	2.646480391	3.25E-06	0.000101863
VITAMIN A AND CAROTENOID METABOLISM-WIKIPATHWAYS_20230610-WP716-HOMO SAPIENS	0.630156894	2.336244687	0.000191234	0.002793997
FATTY ACID BETA-OXIDATION-WIKIPATHWAYS_20230610-WP143-HOMO SAPIENS	0.606439193	1.965521987	0.007950225	0.035586723
ADIPOGENESIS-WIKIPATHWAYS_20230610-WP236-HOMO SAPIENS	0.606429649	3.123619402	9.56E-10	1.80E-07
NONALCOHOLIC FATTY LIVER DISEASE-WIKIPATHWAYS_20230610-WP4396-HOMO SAPIENS	0.605529896	2.279347691	0.00034805	0.004089591
FOLATE METABOLISM-WIKIPATHWAYS_20230610-WP176-HOMO SAPIENS	0.592172411	2.269238127	0.000541836	0.005078617
FATTY ACID BIOSYNTHESIS-WIKIPATHWAYS_20230610-WP357-HOMO SAPIENS	0.570652174	1.923558657	0.002273418	0.015264379

CHOLESTEROL METABOLISM-WIKIPATHWAYS_20230610-WP5304-HOMO SAPIENS	0.562082398	1.950620694	0.005109489	0.029108603
PHOSPHODIESTERASES IN NEURONAL FUNCTION-WIKIPATHWAYS_20230610-WP4222-HOMO SAPIENS	0.554998687	2.057603029	0.003351783	0.020326945
WHITE FAT CELL DIFFERENTIATION-WIKIPATHWAYS_20230610-WP4149-HOMO SAPIENS	0.545894195	1.945445201	0.005859991	0.030133238
SELENIUM MICRONUTRIENT NETWORK-WIKIPATHWAYS_20230610-WP15-HOMO SAPIENS	0.533073934	2.211646293	0.000406148	0.00428869
LEPTIN SIGNALING PATHWAY-WIKIPATHWAYS_20230610-WP2034-HOMO SAPIENS	0.52923993	1.783965953	0.007388433	0.033878668
GLYCOGEN SYNTHESIS AND DEGRADATION-WIKIPATHWAYS_20230610-WP500-HOMO SAPIENS	0.521924146	1.759305853	0.009183706	0.040152017
THYROID HORMONES PRODUCTION AND PERIPHERAL DOWNSTREAM SIGNALING EFFECTS-WIKIPATHWAYS_20230610-WP4746-HOMO SAPIENS	0.518060238	2.191206709	0.000804729	0.006303708
ANGIOPOIETIN-LIKE PROTEIN 8 REGULATORY PATHWAY-WIKIPATHWAYS_20230610-WP3915-HOMO SAPIENS	0.501403318	2.013111484	0.003161255	0.020139967
THERMOGENESIS-WIKIPATHWAYS_20230610-WP4321-HOMO SAPIENS	0.494319143	2.186605554	0.000410619	0.00428869
SUDDEN INFANT DEATH SYNDROME (SIDS) SUSCEPTIBILITY PATHWAYS-WIKIPATHWAYS_20230610-WP706-HOMO SAPIENS	0.411754013	1.841215487	0.007092433	0.033878668
METAPATHWAY BIOTRANSFORMATION PHASE I AND II-WIKIPATHWAYS_20230610-WP702-HOMO SAPIENS	0.386474185	1.938888752	0.005071074	0.029108603
OREXIN RECEPTOR PATHWAY-WIKIPATHWAYS_20230610-WP5094-HOMO SAPIENS	0.374344865	2.071445366	0.000567292	0.005078617
GPCRS, CLASS A RHODOPSIN-LIKE-WIKIPATHWAYS_20230610-WP455-HOMO SAPIENS	0.347554655	1.876271405	0.005473944	0.030133238
VEGFA-VEGFR2 SIGNALING-WIKIPATHWAYS_20230610-WP3888-HOMO SAPIENS	0.346585192	2.053737664	0.001153445	0.008673909
FATTY ACIDS AND LIPOPROTEINS TRANSPORT IN HEPATOCYTES-WIKIPATHWAYS_20230610-WP5323-HOMO SAPIENS	0.325564915	2.069966742	0.000801853	0.006303708
NUCLEAR RECEPTORS META-PATHWAY-WIKIPATHWAYS_20230610-WP2882-HOMO SAPIENS	0.3224205	1.888405415	0.003213825	0.020139967
CILIOPATHIES-WIKIPATHWAYS_20230610-WP4803-HOMO SAPIENS	-0.389000764	-1.812342236	0.00676039	0.033446141
G1 TO S CELL CYCLE CONTROL-WIKIPATHWAYS_20230610-WP45-HOMO SAPIENS	-0.410768995	-1.760677679	0.010342916	0.044192461
1P36 COPY NUMBER VARIATION SYNDROME-WIKIPATHWAYS_20230610-WP5345-HOMO SAPIENS	-0.456525637	-1.869270767	0.007382819	0.033878668
CHRONIC HYPERGLYCEMIA IMPAIRMENT OF NEURON FUNCTION-WIKIPATHWAYS_20230610-WP5283-HOMO SAPIENS	-0.465958187	-1.907892892	0.005930478	0.030133238
DNA REPAIR PATHWAYS, FULL NETWORK-WIKIPATHWAYS_20230610-WP4946-HOMO SAPIENS	-0.501670532	-1.949334114	0.005630364	0.030133238
CELL CYCLE-WIKIPATHWAYS_20230610-WP179-HOMO SAPIENS	-0.551533289	-2.976984806	7.67E-08	7.21E-06
RETINOBLASTOMA GENE IN CANCER-WIKIPATHWAYS_20230610-WP2446-HOMO SAPIENS	-0.55977502	-2.820460892	6.02E-07	2.27E-05
COHESIN COMPLEX - CORNELIA DE LANGE SYNDROME-WIKIPATHWAYS_20230610-WP5117-HOMO SAPIENS	-0.570164238	-2.090982564	0.001994339	0.014420602
MIRNA REGULATION OF DNA DAMAGE RESPONSE-WIKIPATHWAYS_20230610-WP1530-HOMO SAPIENS	-0.578806432	-2.249059993	0.000549643	0.005078617
GASTRIC CANCER NETWORK 1-WIKIPATHWAYS_20230610-WP2361-HOMO SAPIENS	-0.585937792	-2.335786446	0.000334644	0.004089591
DNA IR-DAMAGE AND CELLULAR RESPONSE VIA ATR-WIKIPATHWAYS_20230610-WP4016-HOMO SAPIENS	-0.617179983	-2.64541636	5.69E-06	0.000133666

Table 11. Enrichment results from the WikiPathways gene-sets for the comparison of proximal adipose tissue to tumor tissue. Positive values in enrichment score correspond to proximal adipose tissue and negative values to tumor tissue. NES: Normalized Enrichment Score.

The last comparison we have made for the enrichments was between distal adipose tissue and tumor. As expected these tissues are completely different biologically speaking. We used the same gene sets as in the previous two comparisons. For the Hallmarks gene sets, we could observe similarities to the enrichment in proximal adipose tissue compared to tumor. The hallmarks **adipogenesis** and **fatty acid metabolism** were the higher scored as expected. In tumors, the cell cycle and proliferation related hallmarks were enriched as well as the hallmarks corresponding to estrogen response (see Table 12). Using the KEGG gene sets we observed the adipose tissue related pathways quite enriched (see Table 13), as well as the pathways observed in tumor tissues previously when compared to proximal adipose tissue. Enrichment using the Reactome gene sets retrieved 153 pathways, as it was the case with the comparison of the proximal adipose tissue to the tumor, there were many redundant or complementary pathways, therefore the top and bottom 15 pathways are shown in Table 14. In this enrichment, we observe the classical pathways enriched in both tissue types, however in distal adipose tissue the pathways related to peroxisome metabolism persist as well as pathways that could be related to immune as the **ADORA2B pathway** that has been related to obesity [PMID: 28104382]. Lastly, when we performed the enrichment using WikiPathways we also observed enrichment similarities with the previous observations with the differentiation of white adipose tissue and triglyceride metabolism pathways in the distal adipose tissue as well as the cell cycle pathways and the genes related to gastric cancer and retinoblastoma in tumor tissues (see Table 15).

Gene Set	Enrichment Score	NES	p-value	Adjusted p-value
HALLMARK_ADIPOGENESIS	0.642512317	4.108894496	1e-10	1.66666666666667e-09
HALLMARK_FATTY_ACID_METABOLISM	0.421698746	2.535472792	1.65004822867683e-06	1.59617568694529e-05
HALLMARK_MYOGENESIS	0.371506259	2.402288265	1.91541082433434e-06	1.59617568694529e-05
HALLMARK_HYPOXIA	0.334010677	2.143456655	4.80495685648771e-05	0.000266942
HALLMARK_OXIDATIVE_PHOSPHORYLATION	0.332313148	1.670624386	0.00933792	0.035915076
HALLMARK_XENOBIOTIC_METABOLISM	0.322107557	2.050620591	0.000203024	0.001015122
HALLMARK_UV_RESPONSE_DN	0.316894608	1.873763005	0.002470706	0.010294607
HALLMARK_MTORC1_SIGNALING	-0.319104039	-1.677869305	0.013552222	0.048400794
HALLMARK_ESTROGEN_RESPONSE_LATE	-0.367929709	-2.247807661	2.62473763874711e-06	1.87481259910508e-05
HALLMARK_ESTROGEN_RESPONSE_EARLY	-0.436049563	-2.642571947	3.69566364758807e-10	4.61957955948509e-09
HALLMARK_SPERMATOGENESIS	-0.440352887	-2.016311981	0.00064113	0.002914226
HALLMARK_MITOTIC_SPINDLE	-0.442576584	-2.341100092	1.5091502566357e-05	9.43218910397314e-05
HALLMARK_E2F_TARGETS	-0.557842856	-3.225321328	1e-10	1.66666666666667e-09
HALLMARK_G2M_CHECKPOINT	-0.571071127	-3.314192373	1e-10	1.66666666666667e-09

Table 12. Enrichment results from the Hallmarks gene-sets for the comparison of distal adipose tissue to tumor tissue. Positive values in enrichment score correspond to distal adipose tissue and negative values to tumor tissue. NES: Normalized Enrichment Score.

Gene Set	Enrichment Score	NES	p-value	Adjusted p-value
KEGG_RETINOL_METABOLISM	0.65516261	2.649380022	9.43610761299079e-06	0.000235903
KEGG_FATTY_ACID_METABOLISM	0.644349767	2.680208721	3.61224947756717e-06	0.00015051
KEGG_PPAR_SIGNALING_PATHWAY	0.613188418	3.054730597	7.19682284863285e-09	8.99602856079106e-07
KEGG_PROPANOATE_METABOLISM	0.608381519	2.282724938	0.000494496	0.007726493
KEGG_PROXIMAL_TUBULE_BICARBONATE_RECLAMATION	0.60716301	2.124578506	0.001694835	0.019259493
KEGG_CITRATE_CYCLE_TCA_CYCLE	0.596006748	2.035670724	0.00280211	0.025018839
KEGG_ADIPOCYTOKINE_SIGNALING_PATHWAY	0.584672846	2.625797366	8.31867045696878e-06	0.000235903
KEGG_METABOLISM_OF_XENOBIOTICS_BY_CYTOCHROME_P450	0.555977305	2.248289422	0.0007642	0.010613889
KEGG_LYSINE_DEGRADATION	0.54775284	1.916691059	0.007857334	0.046769846
KEGG_PYRUVATE_METABOLISM	0.536466177	2.06453508	0.003910853	0.032590439
KEGG_DRUG_METABOLISM_CYTOCHROME_P450	0.485172235	2.018100917	0.002079219	0.021658533
KEGG_PEROXISOME	0.480108743	2.156194328	0.001370093	0.017126165
KEGG_INSULIN_SIGNALING_PATHWAY	0.474799042	2.463539133	2.11987405258143e-05	0.00044164
KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	0.467694896	1.891288507	0.007070362	0.046515538
KEGG_GLYCOLYSIS_GLUCONEOGENESIS	0.434488149	1.906112733	0.004858369	0.037594837

KEGG_GLYCEROLIPID_METABOLISM	0.431879364	1.894667913	0.005413656	0.037594837
KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	0.341803587	2.111401316	4.59704656880506e-05	0.000820901
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS	-0.379984997	-1.773356611	0.007682649	0.046769846
KEGG_OOCYTE_MEIOSIS	-0.385256602	-1.809067653	0.005262589	0.037594837
KEGG_P53_SIGNALING_PATHWAY	-0.489511438	-1.920609418	0.00237246	0.022812119
KEGG_CELL_CYCLE	-0.513108695	-2.560457246	1.86161017568847e-07	1.16350635980529e-05

Table 13. Enrichment results from the KEGG gene-sets for the comparison of distal adipose tissue to tumor tissue. Positive values in enrichment score correspond to distal adipose tissue and negative values to tumor tissue. NES: Normalized Enrichment Score.

Gene Set	**Enrichment Score**	**NES**	**p-value**	**Adjusted p-value**
REACTOME_TRANSCRIPTIONAL_REGULATION_OF_WHITE_ADIPOCYTE_DIFFERENTIATION	0.833861988	3.461247488	1e-10	9.23333333333333e-09
REACTOME_PLASMA_LIPOPROTEIN_REMODELING	0.764944568	2.624039188	1.69852974454741e-06	5.22769710266258e-05
REACTOME_TRIGLYCERIDE_METABOLISM	0.700632145	2.807931625	1.81218604014844e-07	7.17107904458739e-06
REACTOME_GLYCOGEN_METABOLISM	0.684665254	2.210584082	0.000556931	0.005232952
REACTOME_TRIGLYCERIDE_CATABOLISM	0.683381582	2.344248364	9.20912603707018e-05	0.001275464
REACTOME_HEME_SIGNALING	0.655205573	2.115467378	0.001439171	0.010221802
REACTOME_ADORA2B_MEDIATED_ANTI_INFLAMMATORY_CYTOKINES_PRODUCTION	0.607412013	2.083645001	0.001589271	0.010869826
REACTOME_SCAVENGING_BY_CLASS_A_RECEPTORS	0.604590898	2.023405774	0.003398874	0.018643329
REACTOME_PLASMA_LIPOPROTEIN_ASSEMBLY_REMODELING_AND_CLEARANCE	0.592738476	2.528287817	1.2459166579877e-05	0.000246454
REACTOME_FOXO_MEDIATED_TRANSCRIPTION_OF_OXIDATIVE_STRESS_METABOLIC_AND_NEURONAL_GENES	0.590677546	1.907125231	0.006674949	0.030310833
REACTOME_PEROXISOMAL_PROTEIN_IMPORT	0.584166572	2.10707931	0.001048054	0.008538558
REACTOME_SYNTHESIS_OF_BILE_ACIDS_AND_BILE_SALTS	0.567509636	1.8993046	0.008549866	0.03671803
REACTOME_BINDING_AND_UPTAKE_OF_LIGANDS_BY_SCAVENGER_RECEPTORS	0.567473924	2.508501007	5.96228777225587e-06	0.000137629
REACTOME_AQUAPORIN_MEDIATED_TRANSPORT	0.565194059	1.990578131	0.002667021	0.016060103
REACTOME_NR1H2_AND_NR1H3_MEDIATED_SIGNALING	0.554248635	1.952029032	0.003945403	0.0204276
REACTOME_RESOLUTION_OF_D_LOOP_STRUCTURES_THROUGH_SYNTHESIS_DEPENDENT_STRAND_ANNEALING_SDSA	-0.572476	-1.9687342	0.00182051	0.01229953
REACTOME_TRANSCRIPTIONAL_REGULATION_BY_E2F6	-0.5896477	-1.7705319	0.00956479	0.0382198
REACTOME_SWITCHING_OF_ORIGINS_TO_A_POST_REPLICATIVE_STATE	-0.6049253	-2.149732	0.00075514	0.0066404
REACTOME_KERATINIZATION	-0.6126038	-2.662422	1.45982210269644e-07	6.22108803764482e-06
REACTOME_TP53_REGULATES_TRANSCRIPTION_OF_GENES_INVOLVED_IN_G2_CELL_CYCLE_ARREST	-0.6126587	-1.8396269	0.00579483	0.02743878
REACTOME_ORC1_REMOVAL_FROM_CHROMATIN	-0.6166229	-1.9174943	0.00285596	0.01701291
REACTOME_G1_S_SPECIFIC_TRANSCRIPTION	-0.6236579	-2.2839775	0.00011555	0.00156134
REACTOME_FORMATION_OF_THE_CORNIFIED_ENVELOPE	-0.6254944	-2.7054003	7.69545181996601e-08	3.55273359021764e-06
REACTOME_TIGHT_JUNCTION_INTERACTIONS	-0.6365186	-2.0717263	0.00076945	0.00666058
REACTOME_G0_AND_EARLY_G1	-0.6452953	-2.2931953	0.00016226	0.00204303
REACTOME_NUCLEAR_ENVELOPE_BREAKDOWN	-0.6462887	-2.0097452	0.0011159	0.00870717
REACTOME_CYCLIN_A_B1_B2_ASSOCIATED_EVENTS_DURING_G2_M_TRANSITION	-0.6574517	-2.1854321	0.00033616	0.00380061
REACTOME_APC_C_MEDIATED_DEGRADATION_OF_CELL_CYCLE_	-0.6769854	-2.5214651	2.06062314637946e-06	5.70792611547111e-05

PROTEINS				
REACTOME_SUMOYLATION_OF_DNA_REPLICATION_PROTEINS	-0.7322385	-2.2770208	3.20653368521726e-05	0.00050755
REACTOME_POLO_LIKE_KINASE_MEDIATED_EVENTS	-0.74409	-2.3138749	1.7658849801587e-05	0.00030572

Table 14. Enrichment results from the Reactome gene-sets for the comparison of distal adipose tissue to tumor tissue. The enriched pathways in the table correspond to the top and bottom 15 pathways enriched. Positive values in enrichment score correspond to distal adipose tissue and negative values to tumor tissue. NES: Normalized Enrichment Score.

Gene Set	Enrichment Score	NES	p-value	Adjusted p-value
DIFFERENTIATION OF WHITE AND BROWN ADIPOCYTE-WIKIPATHWAYS_20230610-WP2895-HOMO SAPIENS	0.786671607	2.976763975	1.78E-08	1.94E-06
TRANSCRIPTION FACTOR REGULATION IN ADIPOGENESIS-WIKIPATHWAYS_20230610-WP3599-HOMO SAPIENS	0.757651834	2.664958702	3.06E-06	0.000142755
FAMILIAL PARTIAL LIPODYSTROPHY-WIKIPATHWAYS_20230610-WP5102-HOMO SAPIENS	0.729785309	2.790222596	3.39E-07	2.21E-05
OXIDATIVE STRESS RESPONSE-WIKIPATHWAYS_20230610-WP408-HOMO SAPIENS	0.706575169	2.408647923	8.17E-05	0.002054659
TRIACYLGLYCERIDE SYNTHESIS-WIKIPATHWAYS_20230610-WP325-HOMO SAPIENS	0.675896339	2.504851511	3.39E-05	0.00100827
FATTY ACID BIOSYNTHESIS-WIKIPATHWAYS_20230610-WP357-HOMO SAPIENS	0.662408759	2.532618655	2.44E-05	0.000797575
GPCRS, CLASS B SECRETIN-LIKE-WIKIPATHWAYS_20230610-WP334-HOMO SAPIENS	0.654712569	2.138078889	0.001003736	0.013675897
OLIGODENDROCYTE SPECIFICATION AND DIFFERENTIATION, LEADING TO MYELIN COMPONENTS FOR CNS-WIKIPATHWAYS_20230610-WP4304-HOMO SAPIENS	0.632298162	2.064880707	0.001637148	0.019119553
PPAR SIGNALING PATHWAY-WIKIPATHWAYS_20230610-WP3942-HOMO SAPIENS	0.624629506	3.054470634	9.51E-09	1.55E-06
FATTY ACID BETA-OXIDATION-WIKIPATHWAYS_20230610-WP143-HOMO SAPIENS	0.624396776	2.362716302	0.000137255	0.003205882
ADIPOGENESIS-WIKIPATHWAYS_20230610-WP236-HOMO SAPIENS	0.605593524	3.347684577	1.00E-10	3.27E-08
EICOSANOID METABOLISM VIA CYCLOOXYGENASES (COX)-WIKIPATHWAYS_20230610-WP4719-HOMO SAPIENS	0.603119935	2.121409394	0.00163556	0.019119553
AMP-ACTIVATED PROTEIN KINASE SIGNALING-WIKIPATHWAYS_20230610-WP1403-HOMO SAPIENS	0.602155494	2.490378523	4.07E-05	0.001109601
FATTY ACID TRANSPORTERS-WIKIPATHWAYS_20230610-WP5061-HOMO SAPIENS	0.596701822	1.948634608	0.004394768	0.037818139
GLYCOGEN SYNTHESIS AND DEGRADATION-WIKIPATHWAYS_20230610-WP500-HOMO SAPIENS	0.584326205	2.165495347	0.000933539	0.013272494
PROSTAGLANDIN SYNTHESIS AND REGULATION-WIKIPATHWAYS_20230610-WP98-HOMO SAPIENS	0.581578881	2.633542635	1.08E-05	0.00044075
WHITE FAT CELL DIFFERENTIATION-WIKIPATHWAYS_20230610-WP4149-HOMO SAPIENS	0.569130032	2.228969324	0.000348497	0.006511185
STEROL REGULATORY ELEMENT-BINDING PROTEINS (SREBP) SIGNALING-WIKIPATHWAYS_20230610-WP1982-HOMO SAPIENS	0.521864829	1.934015368	0.006036389	0.049347484
PHOSPHODIESTERASES IN NEURONAL FUNCTION-WIKIPATHWAYS_20230610-WP4222-HOMO SAPIENS	0.520716341	2.257516572	0.000683623	0.011177231
FERROPTOSIS-WIKIPATHWAYS_20230610-WP4313-HOMO SAPIENS	0.518823022	2.145734326	0.001407697	0.017704498
VITAMIN A AND CAROTENOID METABOLISM-WIKIPATHWAYS_20230610-WP716-HOMO SAPIENS	0.514982907	2.273141499	0.000330489	0.006511185
VITAMIN B12 METABOLISM-WIKIPATHWAYS_20230610-WP1533-HOMO SAPIENS	0.465606856	1.97031353	0.003256018	0.032264173
THYROID HORMONES PRODUCTION AND PERIPHERAL DOWNSTREAM SIGNALING EFFECTS-WIKIPATHWAYS_20230610-WP4746-HOMO SAPIENS	0.461294335	2.238301299	0.000242854	0.005294211
FOLATE METABOLISM-WIKIPATHWAYS_20230610-WP176-HOMO SAPIENS	0.459934385	2.188485687	0.000718595	0.01118955
NONALCOHOLIC FATTY LIVER DISEASE-WIKIPATHWAYS_20230610-WP4396-HOMO SAPIENS	0.428229354	2.015510428	0.002868071	0.03126197
SELENIUM MICRONUTRIENT NETWORK-WIKIPATHWAYS_20230610-WP15-HOMO SAPIENS	0.402331058	1.910987716	0.003019502	0.031686093
AMINO ACID METABOLISM-WIKIPATHWAYS_20230610-WP3925-HOMO SAPIENS	0.395224217	2.029403521	0.001254715	0.016411674

THERMOGENESIS-WIKIPATHWAYS_20230610-WP4321-HOMO SAPIENS	0.394519136	2.041805358	0.000933313	0.013272494
VITAMIN D RECEPTOR PATHWAY-WIKIPATHWAYS_20230610-WP2877-HOMO SAPIENS	0.303161231	1.774429217	0.004915719	0.041216416
OREXIN RECEPTOR PATHWAY-WIKIPATHWAYS_20230610-WP5094-HOMO SAPIENS	0.299182865	1.814095041	0.004091443	0.037818139
FATTY ACIDS AND LIPOPROTEINS TRANSPORT IN HEPATOCYTES-WIKIPATHWAYS_20230610-WP5323-HOMO SAPIENS	0.280247801	1.943864007	0.000367199	0.006511185
EPITHELIAL TO MESENCHYMAL TRANSITION IN COLORECTAL CANCER-WIKIPATHWAYS_20230610-WP4239-HOMO SAPIENS	-0.338438839	-1.767224035	0.004302066	0.037818139
1P36 COPY NUMBER VARIATION SYNDROME-WIKIPATHWAYS_20230610-WP5345-HOMO SAPIENS	-0.434020802	-1.873040479	0.003875785	0.037275935
DNA IR-DAMAGE AND CELLULAR RESPONSE VIA ATR-WIKIPATHWAYS_20230610-WP4016-HOMO SAPIENS	-0.461470311	-2.081517848	0.000378326	0.006511185
MIRNA REGULATION OF DNA DAMAGE RESPONSE-WIKIPATHWAYS_20230610-WP1530-HOMO SAPIENS	-0.462416112	-1.918370071	0.00310078	0.031686093
DNA DAMAGE RESPONSE-WIKIPATHWAYS_20230610-WP707-HOMO SAPIENS	-0.480857019	-1.974748265	0.004266598	0.037818139
CELL CYCLE-WIKIPATHWAYS_20230610-WP179-HOMO SAPIENS	-0.496276758	-2.445540094	1.79E-06	9.73E-05
COHESIN COMPLEX - CORNELIA DE LANGE SYNDROME-WIKIPATHWAYS_20230610-WP5117-HOMO SAPIENS	-0.567630864	-2.001125467	0.001836597	0.02070922
RETINOBLASTOMA GENE IN CANCER-WIKIPATHWAYS_20230610-WP2446-HOMO SAPIENS	-0.570918243	-2.647537432	7.55E-08	6.17E-06
GASTRIC CANCER NETWORK 1-WIKIPATHWAYS_20230610-WP2361-HOMO SAPIENS	-0.645845879	-2.419836086	1.66E-05	0.000604901

Table 15. Enrichment results from the WikiPathways gene-sets for the comparison of distal adipose tissue to tumor tissue. Positive values in enrichment score correspond to distal adipose tissue and negative values to tumor tissue. NES: Normalized Enrichment Score.

From these enrichment studies, we observed that the two adipose tissue types shared gene expression patterns and enriched for similar pathways when compared to the tumors. The proximal to distal adipose tissue comparison has highlighted some pathways that we detected afterwards when comparing the adipose tissues to the tumors, for example, the estrogen related pathways or the pathways associated to development. This could be due to an artifact in the sampling were the tissues were contaminated with tumor cells.

In addition, we have performed the same analysis taking only into account samples from patients having a tumor of one of the luminal subtypes, and we observed the enrichments per tissue to be very similar to the observations in all the patients for the three tissue types. Furthermore, we have performed the analysis for DEG in obese against the normal-weight patients per tissue type, obtaining very few genes and the GSEA analysis did not provide significant enrichment results. The same comparison was performed in the subset of patients with luminal carcinomas and again, the amount of DEG was very low (data not shown).

Representation of Module Activity (ROMA) analysis

In order to continue exploring the proximal adipose tissue to see if we could get a signature to identify the CAA in the samples, we decided to focus and target pathways that were enriched in the proximal adipose tissue and further continued to analyze them. Therefore, from the pathways observed to be enriched in the GSEA, specially the comparisons involving the proximal adipose tissue, we selected 21 pathways from WikiPathways and KEGG, then extracted the information about the genes that participate in them. This collection, could reflect the adipocyte's behavior as well as the CAA. In Table 16, the 21 pathways selected are listed. Using this map, we could apply then the Representation of Module Activity (ROMA) analysis, to see the over-dispersed gene-sets (activation) in the different groups of samples.

Module	Number of Genes	Module	Number of Genes
Differentiation of white and brown adipocyte	25	White fat cell differentiation	32
Triacylglyceride synthesis	24	Leptin signaling pathway	76
Adipogenesis	135	Angiopoietin-like protein 8 regulatory pathway	132
PPAR signaling pathway	68	Angiogenesis	135
AMP-activated protein kinase signaling	67	Thermogenesis	108
Prostaglandin synthesis and regulation	45	VEGFA-VEGFR2 signaling	432
Vitamin A and carotenoid metabolism	43	Estrogen metabolism	24
Folate metabolism	69	Estrogen receptor pathway	13
Fatty acid biosynthesis	22	Matrix metalloproteinases	30
Cholesterol metabolism	72	Adipokines	67
Fatty acid beta-oxidation	34		

Table 16. Pathways selected to conform the CAA collection.

We have applied the ROMA analysis to compare the three different tissues in the context of the different 21 processes represented as modules. Thus, we have identified the overdispersed modules in the data set. We have found three different clusters in these samples (Figure 5). The first cluster observed in the far left corresponds to a predominant cluster composed in its vast majority for tumor samples. These samples had a higher activity of **leptin signaling pathway**, **Vitamin A and carotenoid metabolism** as well as **cholesterol metabolism** modules. In the modules **matrix metalloproteinases**, **folate metabolism**, **PPAR signaling pathway** and **thermogenesis** these set of samples presented a lower expression. The second cluster was predominantly formed by distal adipose tissue samples with some proximal adipose tissue as well. These samples exhibit an inverse module activity to the tumor samples. Finally, the last cluster was smaller but mostly formed by proximal adipose tissue samples. The module activity in this cluster was mixed to the previously described two clusters. Nonetheless, we observed the intensity of this activities was lower than in the two previous clusters, having some samples more alike the cluster conformed by mostly tumor tissues and other samples resembling the module activity of the cluster of mostly distal adipose tissues.

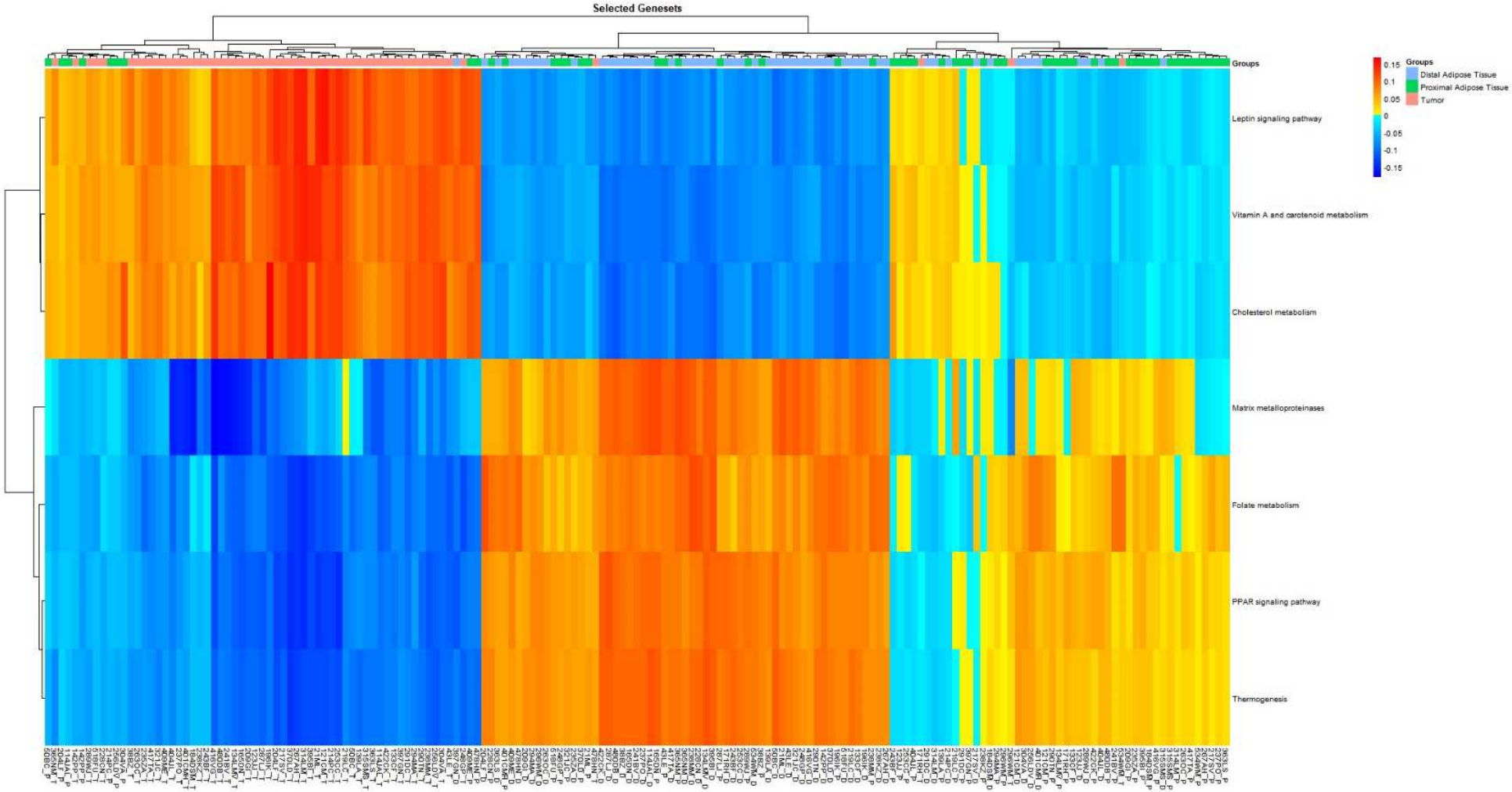


Figure 5. rRoma results corresponding to overdispersed modules in the comparison of the three tissue types. The reddest color indicates higher activation and the bluest lower activation.

With these results, we have compared then the sample scores for these modules across the different modules and groups. The tumor tissues had higher sample scores in the modules **leptin signaling pathway**, **vitamin A and carotenoid metabolism** and **cholesterol metabolism**, as the previous first cluster. The distal adipose tissues exhibited higher scores in the PPAR signaling, Thermogenesis, matrix metalloproteinases and folate metabolism. The proximal adipose tissue appeared to be in the middle of the other two tissues (see Figure 6).

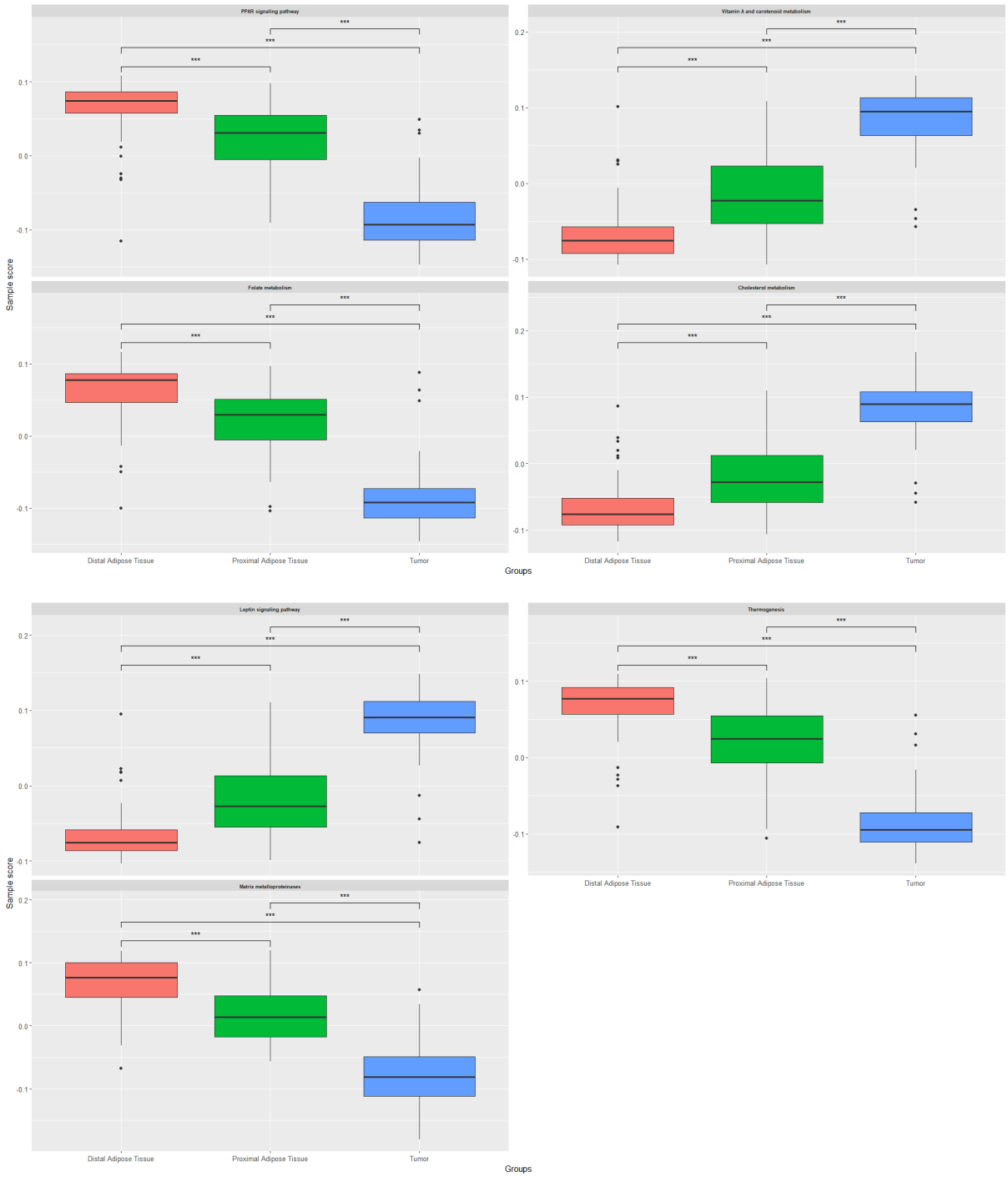


Figure 6. Boxplots showing the distribution for the different overdispersed modules in the three tissues. significance codes: "*" p<0.05, "**" p<0.01, "***" p0.001. Top left corresponds to PPAR signaling pathway module, top right Vitamin A and carotenoid metabolism, second late folate metabolism, second right cholesterol metabolism, third left leptin signaling pathway, bottom right thermogenesis and bottom left matrix metalloproteinases.

Then we explored the most contributing genes per each of the overdispersed modules. To do so, a correlation was computed using the gene expression value and the score of the module. The cut-off value was 20% of the genes in the module to the absolute correlation value. In Figure 7 we can observe a heatmap with the most contributing genes per module. In the **cholesterol metabolism** we had that the genes *CD36, ABCA1, APOB, LRP1, ANGPTL8, CETP, TM7SF2, MGAT1, SORT1, SCARB1,* as positively contributing, and the genes *SQLE, HMGCR, DHCR24,* as to contribute negatively to the module activity. In the module **folate metabolism**, we only obtained positively contributing genes, which were: *SAA1, GPX3, SAA2, SOD3, CAT, HBB, ABCA1, APOB, HBA1, MTHFD1, GPX4, SHMT1*. In the **leptin signaling pathway** we obtained the genes *ESR1, ERBB2, IKBKB* as positive contributors, whereas the genes *PDE3B, FYN, SOS1, FOXO1, CFL2, PTPN11, SOCS3, JAK1, ROCK2, LEPR, STAT5B, PTEN* presented a negative correlation. In the **matrix metalloproteinases** module we had the genes *TIMP4, TIMP3* as positive contributors, whereas the genes *MMP13, MMP3, MMP7, MMP1* were negatively correlated. In the **PPAR signaling pathway** the top contributing genes appeared to correlate negatively to the module activity, these genes were: *PLIN1, FABP4, AQP7, SORBS1, CD36, PPARG, ACSL1, PCK1, ACADL, ANGPTL4, ME1, FABP5, ACOX1.* For the **thermogenesis module**, the genes that contributed positively were: *PLIN1, KLB, PPARG, NPR1, ACSL1, MGLL, PNPLA2, SOS1, CREB5, ADCY5, ADCY4, SLC25A20, ACSL4, RPS6KA2, ADCY6* while the genes *SMARCC1, SMARCA4, CREB3L4, MAPK13, SLC25A29, BMP8A* were negatively contributing. Lastly, in the module **vitamin A and carotenoid metabolism**, all the top contributing genes appeared to do so in a negative manner, these genes were: *RBP4, CD36, RBP7, RETSAT, DHRS3, ALDH1A1, RXRA, CYP26B1*. Additionally, we have also plotted the weights of the genes in the module score as well as their expression in all the samples to interpret better the meaning of the contribution of the genes (Supplementary figures 7-13).

Subsequently, we have analyzed our cohort in the subset of samples from patients having a luminal subtype of cancer (n = 45). We re-ran the analysis of the three tissues, in the luminal samples, where we have found the modules to change their behavior. In Figure 8 we can observe the general pattern of expression for the overdispersed modules. We have obtained six overdispersed modules, that were **matrix metalloproteinases**, **estrogen metabolism**, **folate metabolism**, **cholesterol metabolism**, **vitamin A and carotenoid metabolism** and **thermogenesis**. Strikingly, this six modules appeared to be more active in the cluster conformed mostly by tumors, whereas the cluster conformed by mostly distal adipose tissues had a lower activation. There were two subgroups displayed at the corners of the figure, that were mostly composed of proximal adipose tissue samples, these subgroups shown mixed scores for the overdispersed modules.

Looking closely to these scores, we have seen a similar pattern as with all the samples, where the proximal adipose tissues lie in between the other two tissue types. The modules matrix metalloproteinases, thermogenesis and folate metabolism had a opposite scoring in the previous comparison with all the samples. The modules cholesterol metabolism, vitamin A and carotenoid metabolism, presented a similar overall scoring across the samples with higher scores for the tumor samples. In this analysis, the module **estrogen metabolism** stood out since it was not part of the overdispersed modules in the previous analysis. The modules **leptin signaling pathway** and **PPAR signaling pathway** did not appear to be overdispersed in these samples (see Figure 9).

When computing the top contributing genes, we have obtained the following: in the **cholesterol metabolism**, we have obtained the genes *CD36, LRP1, ABCA1, APOB, ANGPTL8, MGAT1, SORT1, CETP, SCARB1, LBR, TM7SF2* as negatively contributors whereas the genes *SQLE, DHCR24, MSMO1* as positive contributors of the module. In the **estrogen metabolism** module, we have observed the genes *AKR1C3, HSD17B4, CYP3A5* as negative contributing and the gene *HSD17B7* as the only positive contributor. In the module folate metabolism only showed negative contributors which were the genes

SAA1, SOD3, GPX3, SAA2, CAT, HBB, ABCA1. APOB, HBA1, MTHFD1, GPX4, SHMT1. In the module **matrix metalloproteinases**, the genes *TIMP4, TIMP3* appeared as negatively contributing whereas the genes *MMP3, MMP13, MMP1, MMP14* contributed positively. In the module **thermogenesis**, the identified positive contributing genes were *SMARCC1, SMARCA4, CREB3L4, MAPK13, SLC25A29, BMP8A* and the negatively contributing genes were *PLIN1, NP1, KLB, PPARG, MGLL, PNPLQA2, SOS1, CREB5, ADCY4, ADCY5, ACSL4, RPS6KA2, SLC25A20, ACSL5, ADCY6.* Lastly, the module **vitamin A and carotenoid metabolism** presented *CRABP2* as positive contributing gene and the genes *CD36, ALDH1A1, CYP26B1, RBP4, DHRS3, RBP7, RETSAT,* as negative contributors. The plots corresponding to the gene weights and expression in the overdispersed modules are in the Supplementary figures 14-19.

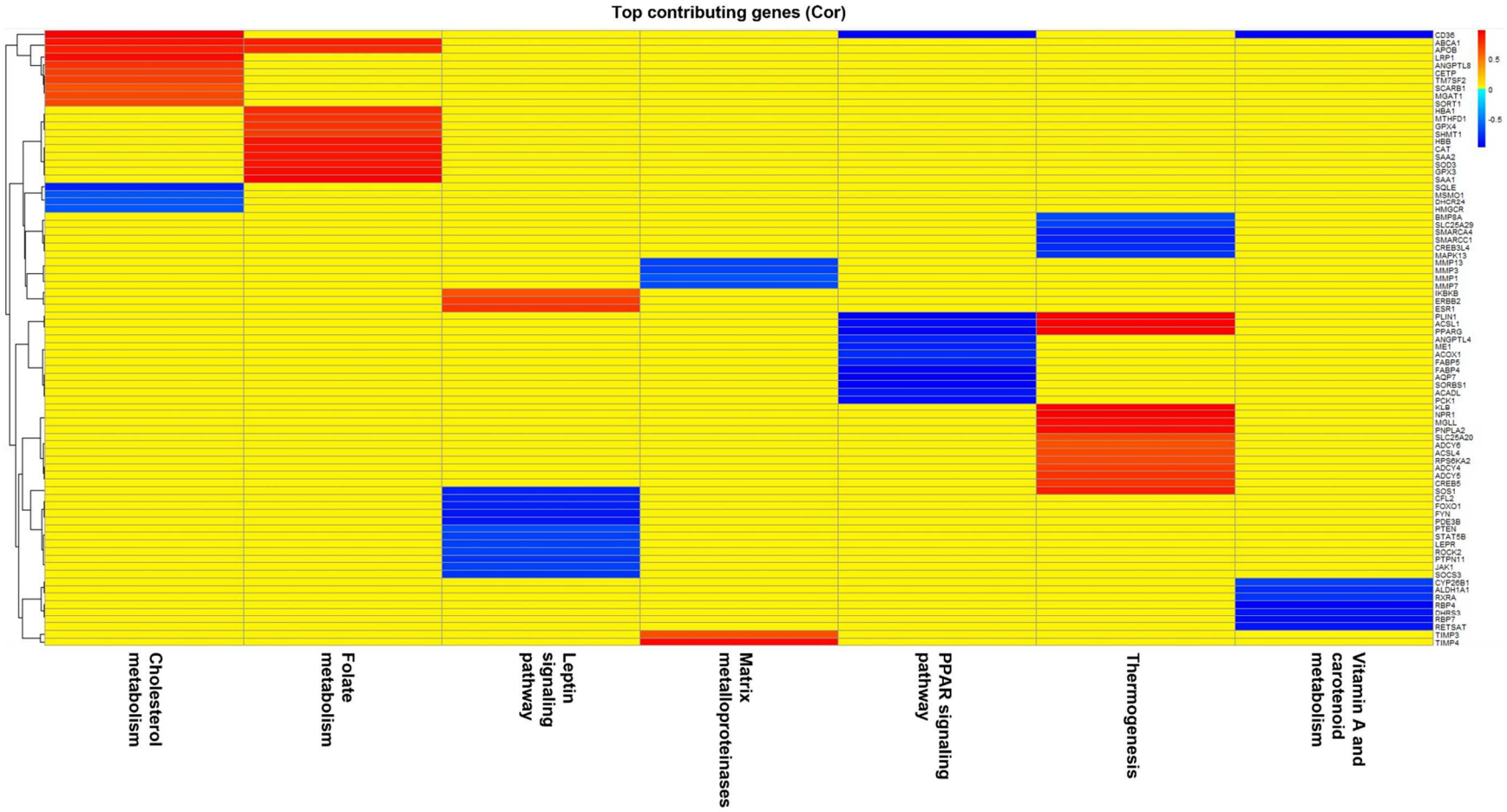


Figure 7. Top contributing genes per module, in the overdispersed modules for the three tissues. Their contribution was calculated as the correlation between the gene expression and the module score.

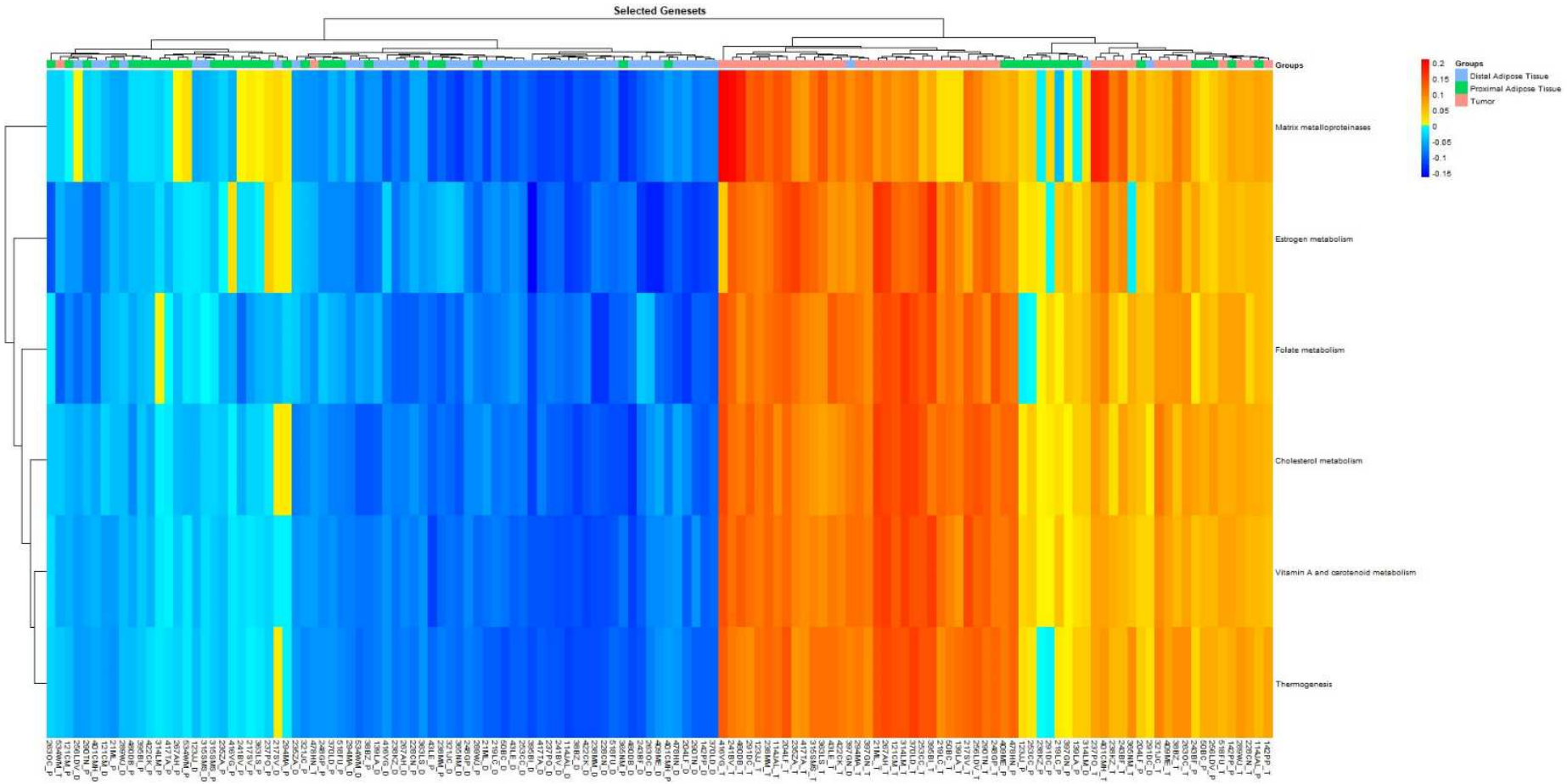


Figure 8. rRoma results corresponding to overdispersed modules in the comparison of the three tissue for the luminal subtypes. The reddest color indicates higher activation and the bluest lower activation.

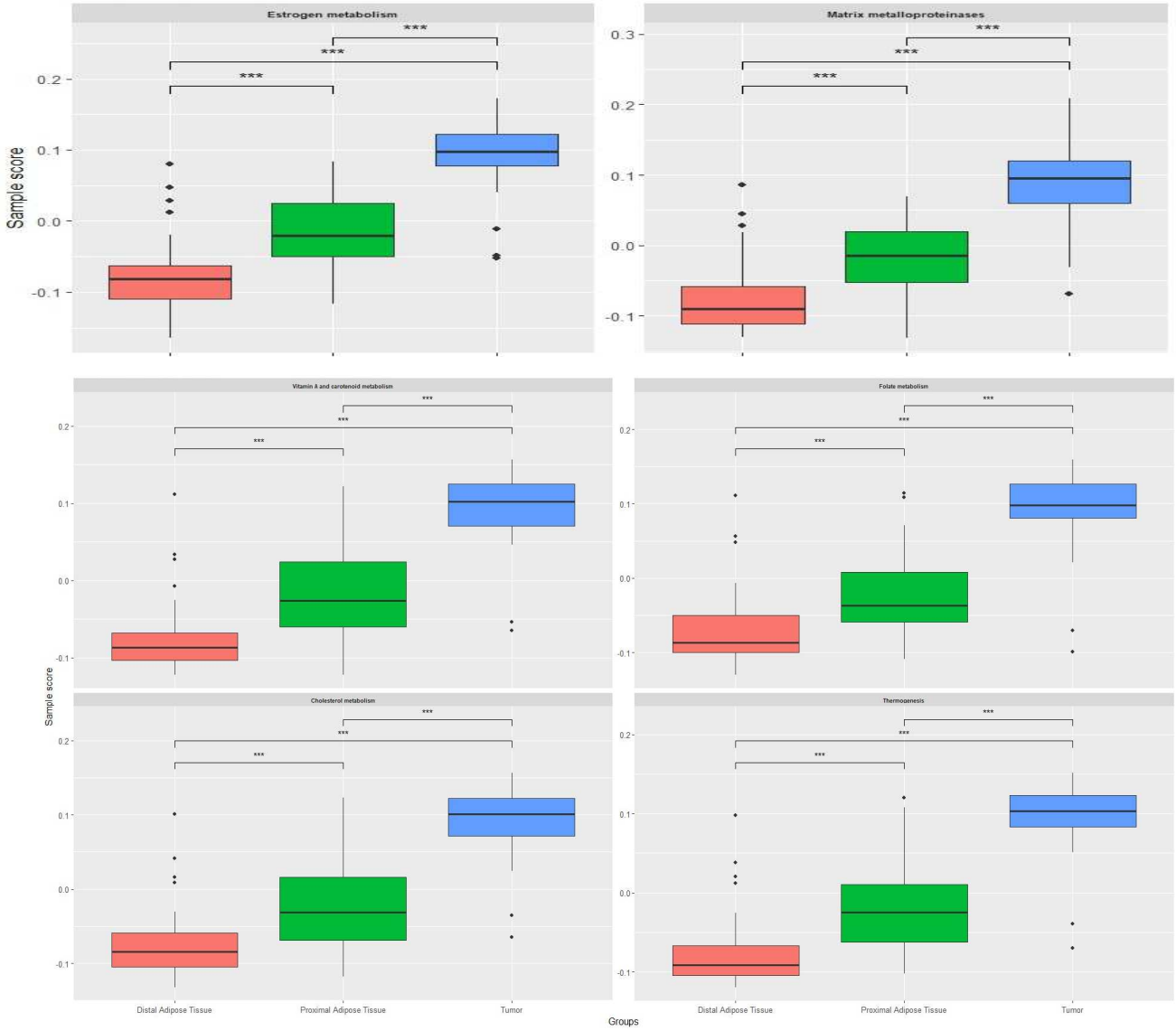


Figure 9. Boxplots showing the distribution for the different overdispersed modules in the three tissues in the luminal patients subset. Significance codes: "*" p<0.05, "**" p<0.01, "***" p0.001. Top left corresponds to Estrogen metabolism module, top right to matrix metalloproteinases, second left Vitamin A and carotenoid metabolism, second right folate metabolism, bottom left cholesterol metabolism and bottom right thermogenesis.

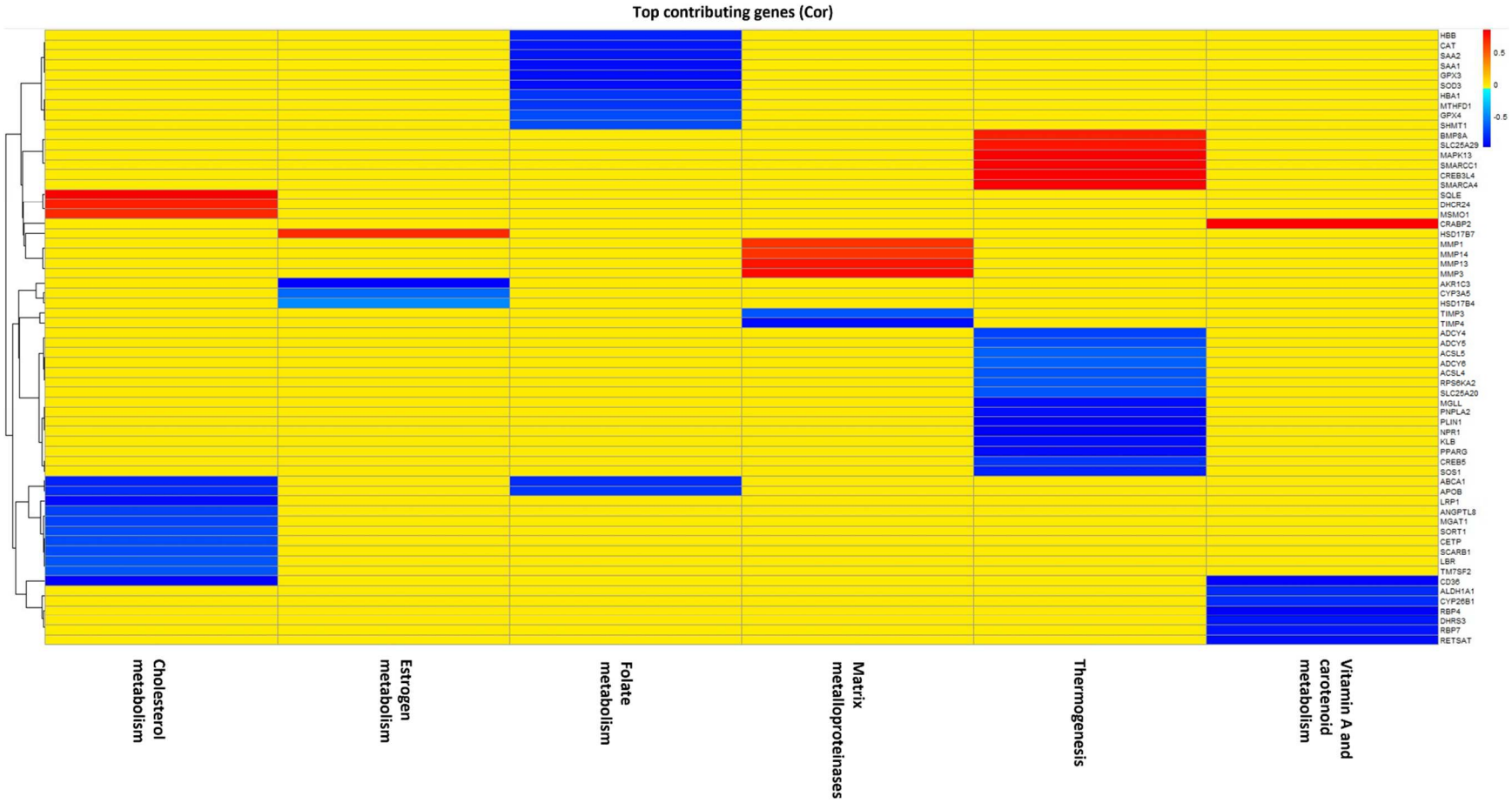
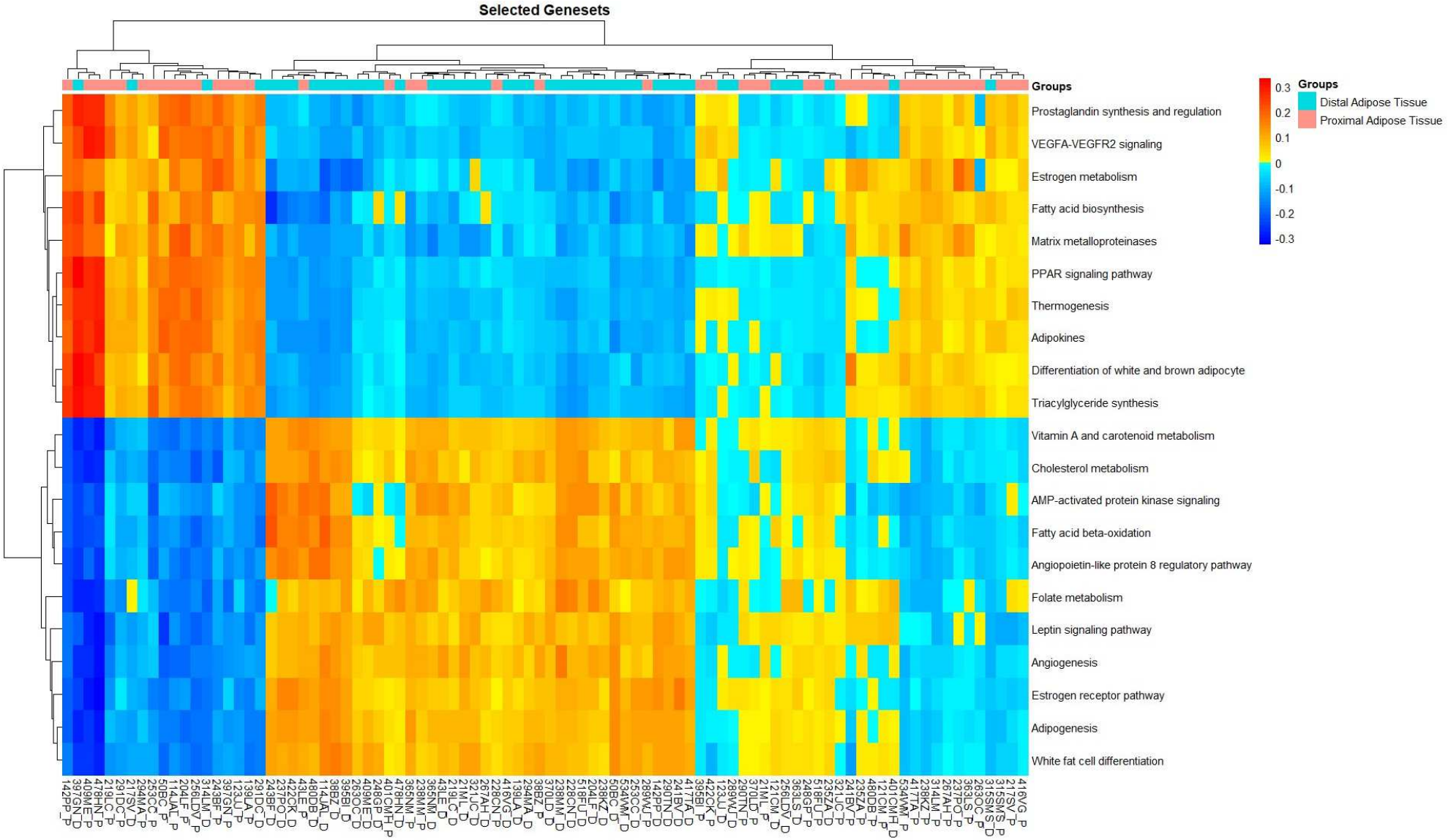


Figure 10. Top contributing genes per module, in the overdispersed modules for the three tissues in the luminal subtypes. Their contribution was calculated as the correlation between the gene expression and the module score.

The results from this analysis combining the three tissues already gave us an insight of the data, especially regarding the behavior of the proximal adipose tissue samples. Next, we wanted to compare only the two types of adipose tissues, in order to see if we could detect signals from CAA using the mapped pathways. After running the analysis, first we looked into all the modules' scores, regardless of the overdispersion, and we saw that there was a solid cluster of proximal adipose tissue samples, that had very extreme scores, then we had a really big cluster that was mostly composed of distal adipose tissue samples which scores seem opposite to the first cluster. Then the third cluster composed of mostly proximal adipose tissues had mixed patterns. This third cluster had some samples of proximal adipose tissue with scores similar to the first cluster but in a milder way (Figure 11).

Afterwards, we have computed the overdispersed modules, that were **vitamin A and carotenoid metabolism**, **cholesterol metabolism**, that had higher scores in distal adipose tissues, **thermogenesis** and **matrix metalloproteinases**, that had higher scores in proximal adipose tissue samples (see Figure 12). Conversely, we have observed that when we had the analyses with the tumor samples, these modules scored higher in this type of samples. When exploring the top contributing genes for these four overdispersed modules, we had for the **cholesterol metabolism** module the genes *CD36, LRP1, ABCA1, TM7SF2, APOB, SORT1, CETPO, MGAT1, DGAT1, TSPO, ANGPTL8* as positively contributing and *HMGCR, SQLE, MSMO1* as negatively contributing, in the **vitamin A and carotenoid metabolism** we had the genes *RBP4, CD36, RETSAT, RXRA, DHRS3, RBP7, CYP26B1,* as positively contributing genes and only *CRABP2* as a negatively contributing gene. For the module **thermogenesis** we had as negatively contributing genes *PLIN1, MGLL, PNPLA2, NPR1, ACSL1, KLB, PPARG, SOS1, PRKACA, ADCY6, SLC25A20, ADCY5, ADCY4* and as positively contributing the genes *SMARCC1, CREB3L4, MAPK13, RPS6KA6, SMARCA4, ZNF516, ACTG1, SLC25A9.* Lastly, the **matrix metalloproteinases** module had as negatively contributing the genes *TIMP4, TIMP3* and as positively contributing the genes *MMP7, MMP16, MMP3*. Additionally, we have also plotted the weights of the genes in the module score as well as their expression in all the samples to interpret better the meaning of the contribution of the genes (Supplementary figures 20-23).


Selected Genesets
Groups
Prostaglandin synthesis and regulation
VEGFA-VEGFR2 signaling
Estrogen metabolism
Fatty acid biosynthesis
Matrix metalloproteinases
PPAR signaling pathway
Thermogenesis
Adipokines
Differentiation of white and brown adipocyte
Triacylglyceride synthesis
Vitamin A and carotenoid metabolism
Cholesterol metabolism
AMP-activated protein kinase signaling
Fatty acid beta-oxidation
Angiopoietin-like protein 8 regulatory pathway
Folate metabolism
Leptin signaling pathway
Angiogenesis
Estrogen receptor pathway
Adipogenesis
White fat cell differentiation
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
Groups
Distal Adipose Tissue
Proximal Adipose Tissue

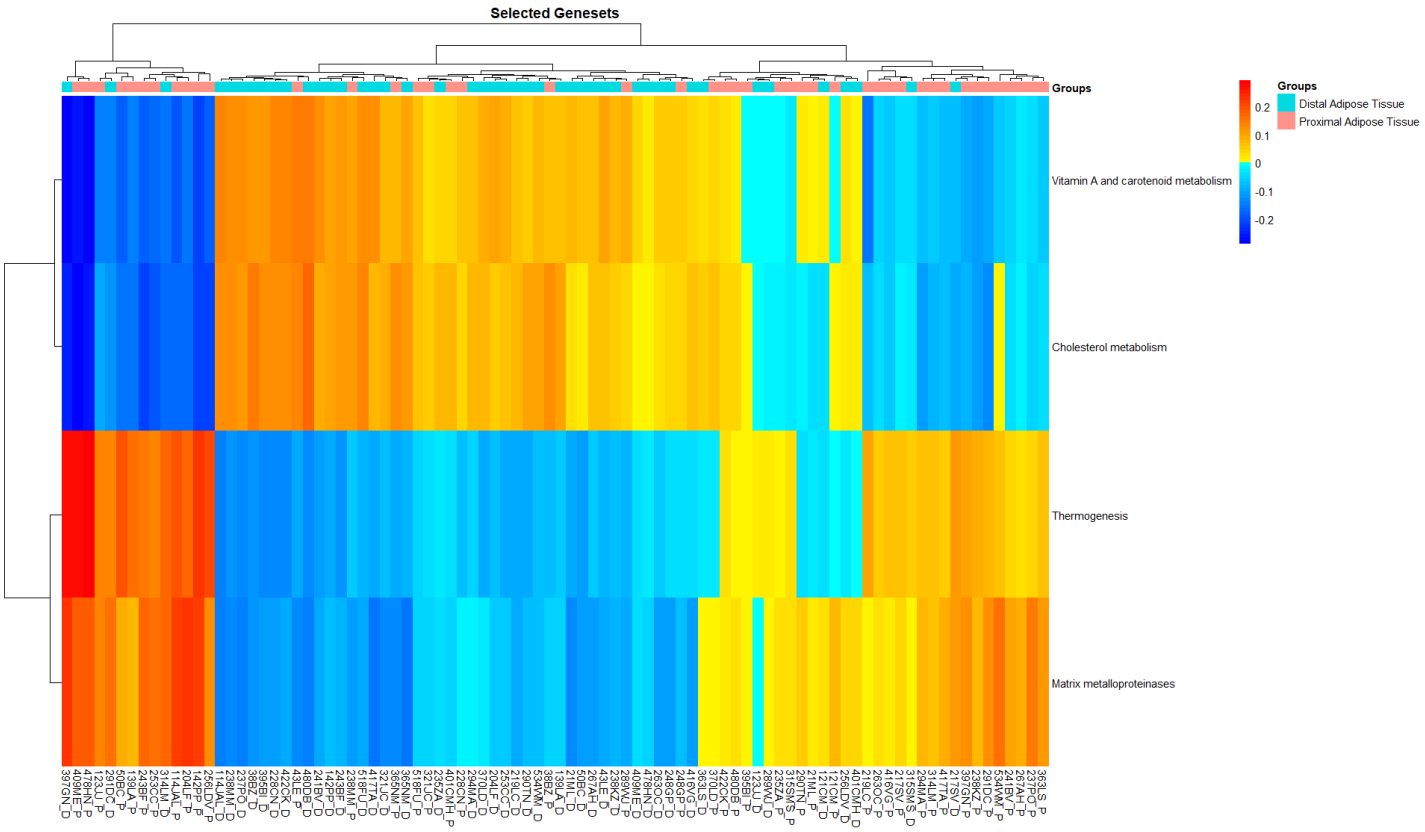


Figure 12. Heatmap of the overdispersed modules between proximal and distal adipose tissue samples.

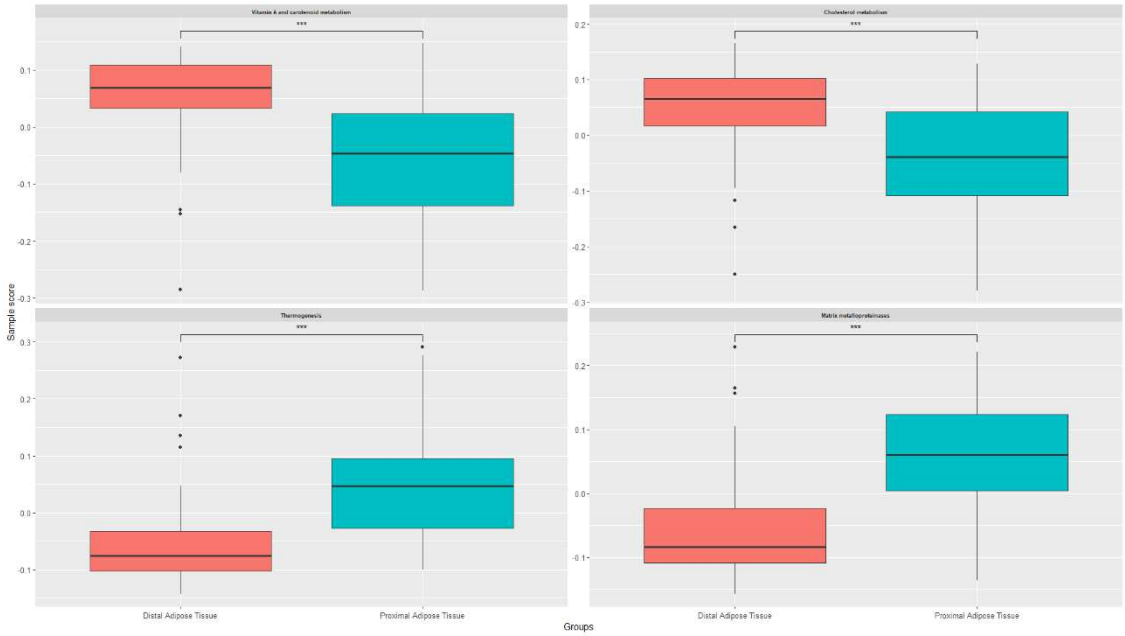


Figure 13. Boxplots showing the sample score distribution in the different overdispersed modules in the distal and proximal adipose tissues. significance codes: "*" p<0.05, "**" p<0.01, "***" p<0.001. Top left boxplots corresponds to vitamin A and carotenoid metabolism module, top right to cholesterol metabolism, bottom left to thermogenesis and bottom right to matrix metalloproteinases.

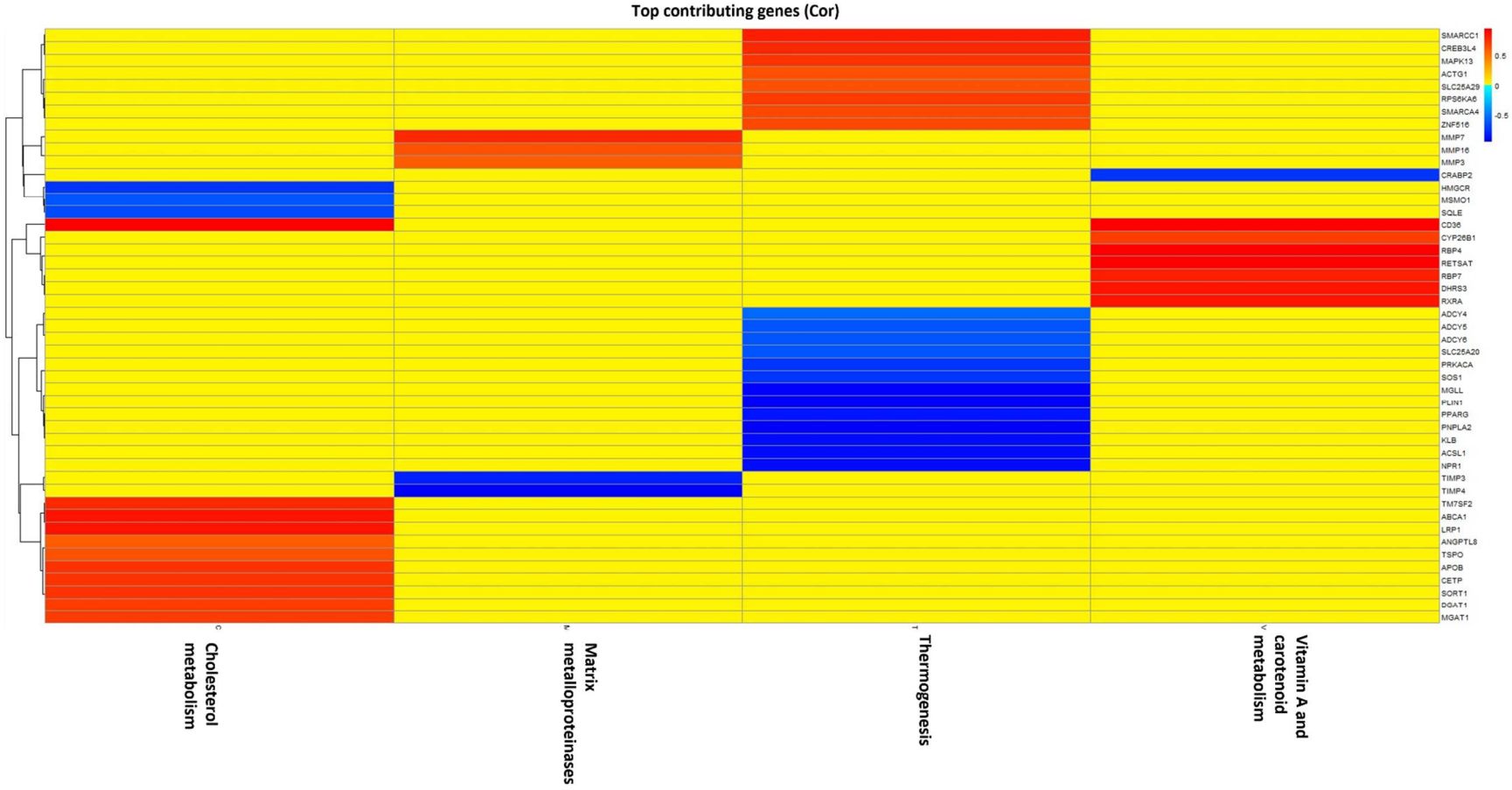
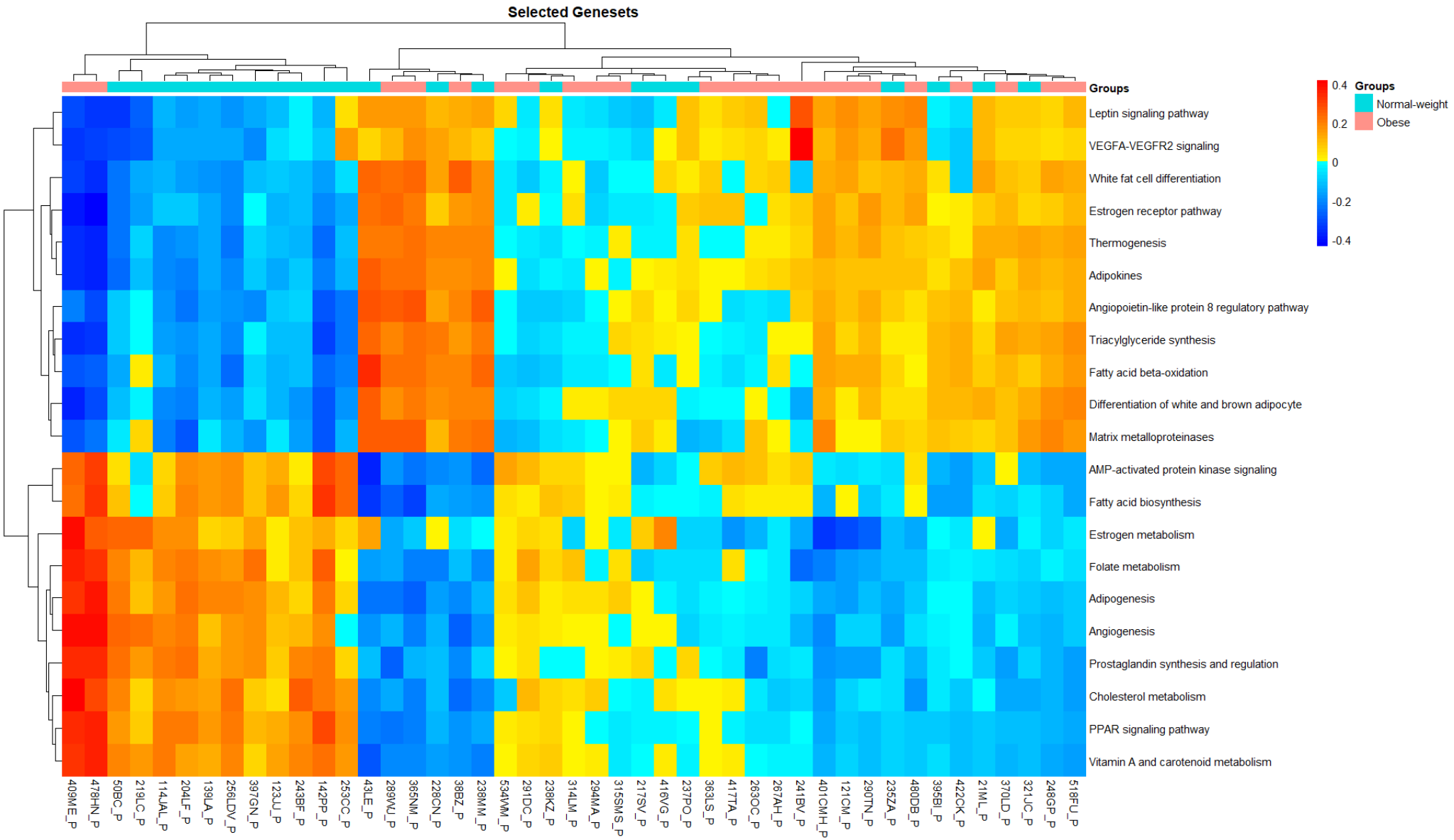


Figure 14. Heatmap of the top contributing genes in the overdispersed modules between proximal and distal adipose tissues.

Furthermore, we interrogated the subset of the proximal adipose tissues, in order to retrieve the mapped pathways comparing the patients by their BMI category. We had 45 samples corresponding to the patients that had a luminal tumor subtype, from them, 23 were Obese and 22 Normal-weight. When exploring the general scores in all the modules, we observed a cluster of patients, mostly corresponding to Normal-weight and then we saw some clusters of mixed patients (Figure 15).

After computing the overdispersed modules, the **vitamin A and carotenoid metabolism** as well as the **cholesterol metabolism** were mostly overdispersed in the cluster with the majority of Normal-weight samples, and the modules Leptin signaling pathway, **Angiopoietin-like protein 8 regulatory pathway** and **thermogenesis** were mostly overdispersed in samples from Obese patients (Figure 16). With this information, we have then explored the sample scores for the BMI categories, observing that the modules **vitamin A and carotenoid metabolism** and **cholesterol metabolism** indeed had a statistically significant difference in the Normal-weight samples compared to the Obese. In the Obese patients the module **thermogenesis** had a higher scores than the Normal-weight patients as well as the Leptin signaling pathway and **Angiopoietin-like protein 8 regulatory pathway**. Nevertheless, only the Thermogenesis module was statistically significant (Figure 17).

When obtaining the top contributing genes we have observed that in the module **Angiopoietin-like protein 8 regulatory pathway**, we had the genes *MLXIPL, RXRA, SLC2A4, RHOQ, PCK1, RAPGEF1, SCD, MAPK10, ANGPTL8, TRIP10, MAPK3, MAPK11, SOS1, FASN, NR1H3, FOXO1, IRS2, AKT2, MAP3K9* positively contributing, whereas the genes *MAP3K1, RPS6KA5, RPS6KA6, PIK3C2G, MAP2K6, MAPK13, MAP4K5*, where negatively contributing. In the module **cholesterol metabolism** the genes *CD36, ABCA1, LRP1, MGAT1, CETP, APOE, TM7SF2, APOB, ANGPTL8,* were negatively correlated, whereas the genes *HMGCR, MSMO1, SQLE, DHCR24, HMGCS1,* were positively contributing to the module scoring. In the **leptin signaling pathway** the genes *ANGPTL8, ESR1, ERBB2, PRKAA2, SRC, CCND1, IRS1* contributed negatively and the genes *FYN, SCOS3, ROCK2, CFL2, SOS1, PTPN11, PDE3B, FOXO1, STAT5B,* contributed positively. In the **thermogenesis** module, the genes *PLIN1, MGLL, NPR1, PNPLA2, ACSL1, KLB, PPARG, SOS1, PRKACA, ACSL5, RPS6KA2, ADCY4* were positively contributing whereas the genes *SMARCC1, CREB3L4, RPS6KA6, MAPK13, SMARCE1, FRS2, SMARCA4, SLC25A29, KDM1A* were found to be negatively contributing. Lastly, in the **vitamin A and carotenoid metabolism**, the genes *CD36, RBP4, RETSAT, RXR4, RBP7, DHRS3, CYP26B1* were negatively contributing and the gene *CRABP2* was positively contributing. As in the previous comparisons, we have also plotted the weights of the genes in the module score as well as their expression in all the samples to interpret better the meaning of the contribution of the genes (Supplementary figures 24-28). From this analysis, it appears that the differences regarding the BMI are minimal. Also it could be due to the sample size of the analysis (n = 45).


Selected Genesets
Groups
Leptin signaling pathway
VEGFA-VEGFR2 signaling
White fat cell differentiation
Estrogen receptor pathway
Thermogenesis
Adipokines
Angiopoietin-like protein 8 regulatory pathway
Triacylglyceride synthesis
Fatty acid beta-oxidation
Differentiation of white and brown adipocyte
Matrix metalloproteinases
AMP-activated protein kinase signaling
Fatty acid biosynthesis
Estrogen metabolism
Folate metabolism
Adipogenesis
Angiogenesis
Prostaglandin synthesis and regulation
Cholesterol metabolism
PPAR signaling pathway
Vitamin A and carotenoid metabolism
0.4
0.2
0
-0.2
-0.4
Groups
Normal-weight
Obese
409ME_P
478HN_P
50BC_P
219LC_P
114JAL_P
204LF_P
139LA_P
256LDV_P
397GN_P
123JJ_P
243BF_P
142PP_P
253CC_P
43LE_P
289WJ_P
365NM_P
228CN_P
38BZ_P
238MM_P
534VM_P
291DC_P
238KZ_P
314LM_P
294MA_P
315SMS_P
217SV_P
416VG_P
237PO_P
363LS_P
417TA_P
263OC_P
267AH_P
241BV_P
401CMH_P
121CM_P
290TN_P
235ZA_P
480DB_P
395BI_P
422CK_P
21ML_P
370LD_P
321JC_P
248GP_P
518FU_P

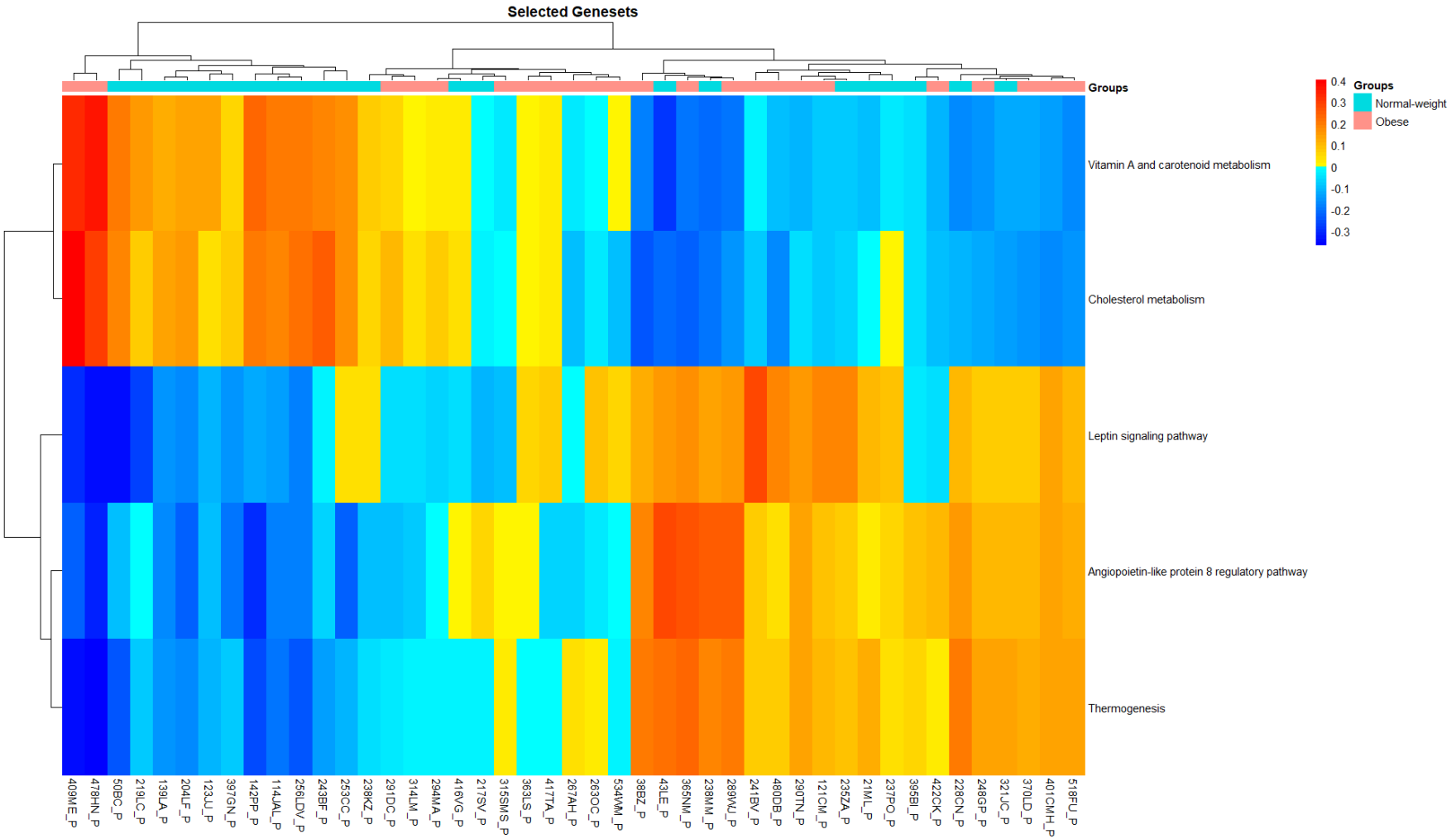


Figure 16. Heatmap of the sample scores in the overdispersed modules in proximal adipose tissues, grouped by their BMI category.

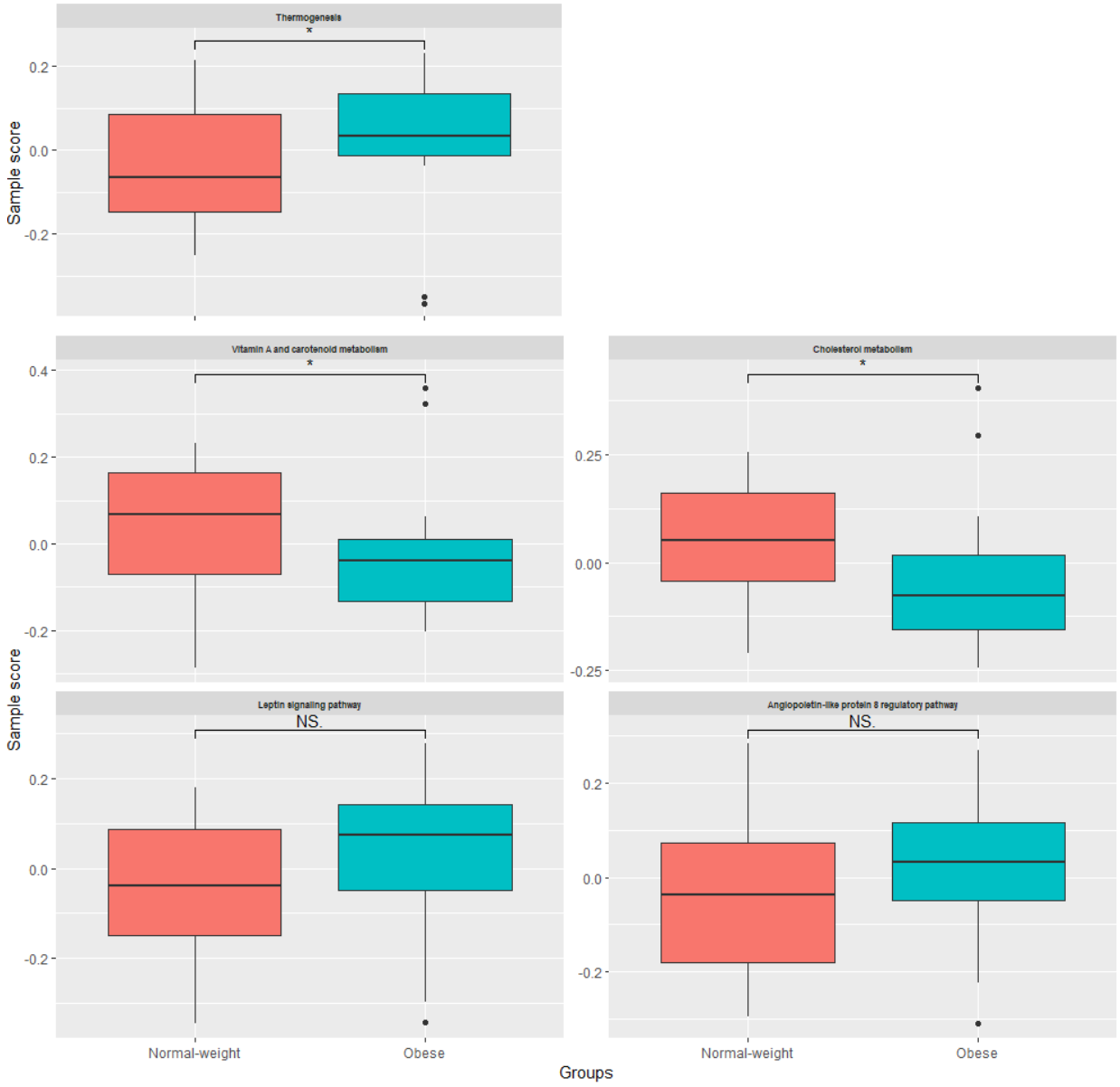


Figure 17. Boxplots depicting the sample scores distributions in the overdispersed modules obtained from the proximal adipose tissue samples when comparing by the BMI categories. Top left boxplots correspond to thermogenesis module, second left to vitamin A and carotenoid metabolism, top right cholesterol metabolism, bottom right to leptin signaling pathway and bottom right to angiopoietin-like protein 8 regulatory pathway.

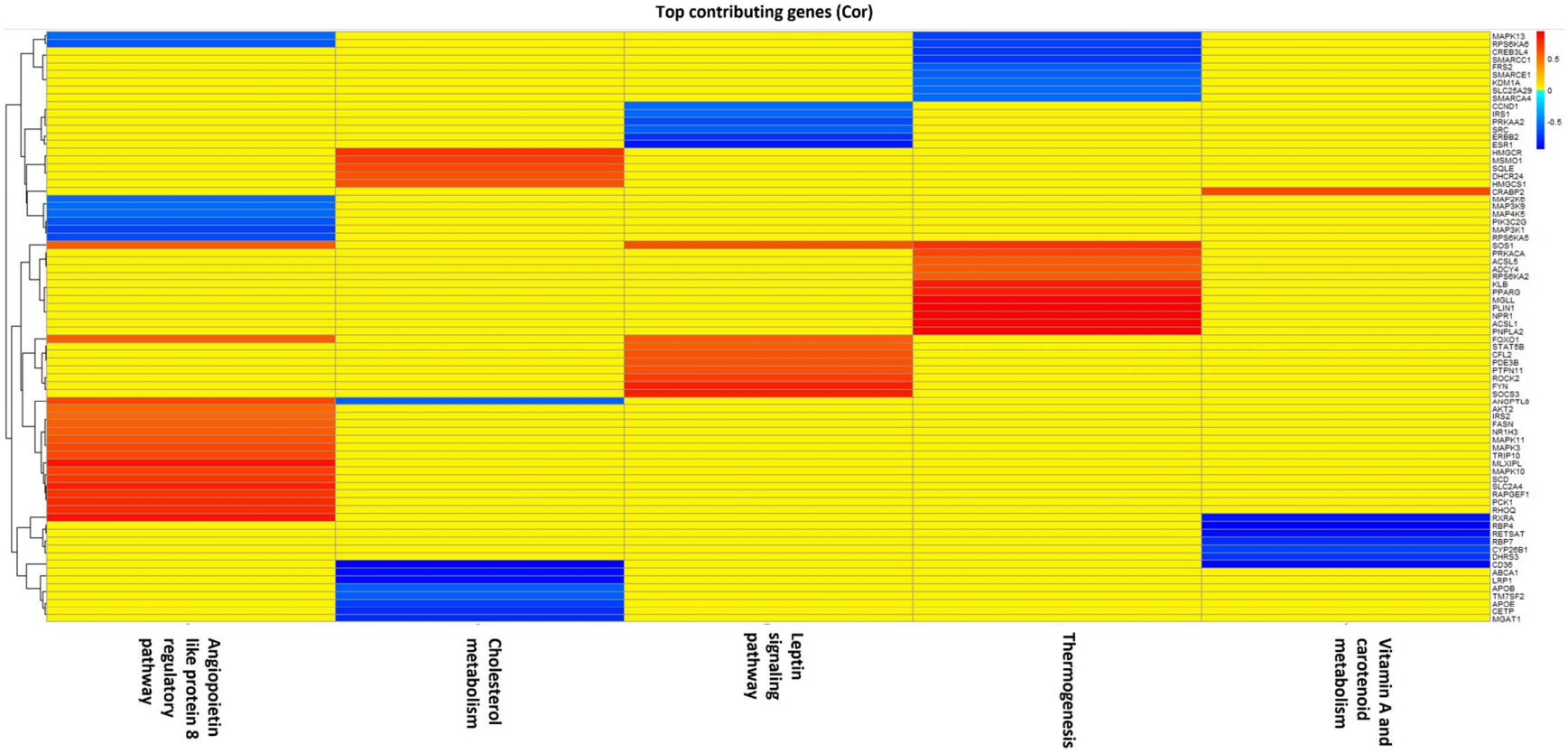


Figure 18. Heatmap of the top contributing genes in the overdispersed modules in the proximal adipose tissues when comparing the BMI categories.

D. DISCUSSION

Here we have presented an integrative study that entailed the experimental design, biological specimen's collection and processing as well as the data analysis to study the gene expression profiles of mammary adipose tissue and to identify signatures of cancer associated adipocytes in the context of ductal carcinomas. The general cohort has more than 100 patients, however, for the preparation of this thesis, we could gather 57 patients with all the tissues of interest sequenced. From this cohort, predominated the patients that had an invasive carcinoma corresponding to the luminal subtypes.

The exploratory analyses have shown the overlap of the proximal adipose tissue samples between tumor and distal adipose tissue samples. Additionally, to consider, there were cases where samples coming from the same patient presented similar gene expression and consequently tended to cluster together. This factor was important to take into account for the differential expression analyses. In our cohort, we did not observe notable differences when considering BMI or the combination of BMI and tissue type. Then we consider to only analyze the luminal subtypes, due to the fact that these subtypes shared many similarities. In this case, the samples seemed to be more localized in groups, specially the tumor and the distal adipose tissue samples.

From the differential expression analysis, when we compared both adipose tissues, we obtained the lower amount of DEG as expected. From the functional analysis, comparing proximal and distal adipose tissue types we observed that the proximal adipose tissue enriched for estrogen signaling pathways as well as pathways involved in cell cycle and related to epithelial function. However, the comparison against the tumor samples revealed these processes to be more enriched in tumor samples whereas the proximal adipose tissues had enriched pathways related to adipose tissue functions. Therefore, these proximal adipose tissue samples, from the enriched pathways, appeared to have functionalities related to tumor but preserving their adipose tissue characteristics. To notice, since the sampling was done macroscopically we can infer that tumor cells could be present in these adipose tissue.

The goal of this study was to capture the CAA signatures from tissue specimens at the transcriptome level. Hence, the results from functional analysis were taken and refined, summarizing to a collection of 21 pathways that might be relevant. CAA as part of the TME in certain cancers, have gained popularity in the last years [23]. We have used adipose tissue in the proximity of the tumor as a proxy to explore the CAA gene expression. In our results from the ROMA analysis, the pathways thermogenesis and matrix metalloproteinases appeared to be more active when compared to adipose tissues. The matrix metalloproteinases have been associated to be induced in the tumor microenvironment cells by the tumor cells[24]. In our study, we have found the *MMP7, MMP16,* and *MMP3* to be positively expressed in proximal adipose tissues. It has been described that leptin, a major adipokine, can induce the production of metalloproteinases, especially *MMP2* and *MMP9* in breast cancer cells[25]. *MMP3* and *MMP9* have been found to be overexpressed in breast cancer cells upon co-culture with adipocytes[26]. MMP11 has been also associated to be expressed in adipocytes co-cultured with cancer cells[27]. *MMP7* expression has been to be expressed in cells termed adipocyte derived fibroblasts[13] .Taken together, however their expression has been seen in tumors, and in the tumors their expression was higher. Taken together, this results shown an increase activity in matrix metalloproteinases as a pathway, and suggests that their players should be further investigated.

The thermogenesis module that appeared to be more active in the proximal adipose tissues, is related to non-shivering heat production, this process has been associated to brown adipose tissue, that is characterized by having higher number of mitochondria and smaller lipid droplets[28]. In cancer, white adipocyte dedifferentiation has been described to occur and to favor the tumor progression, during this

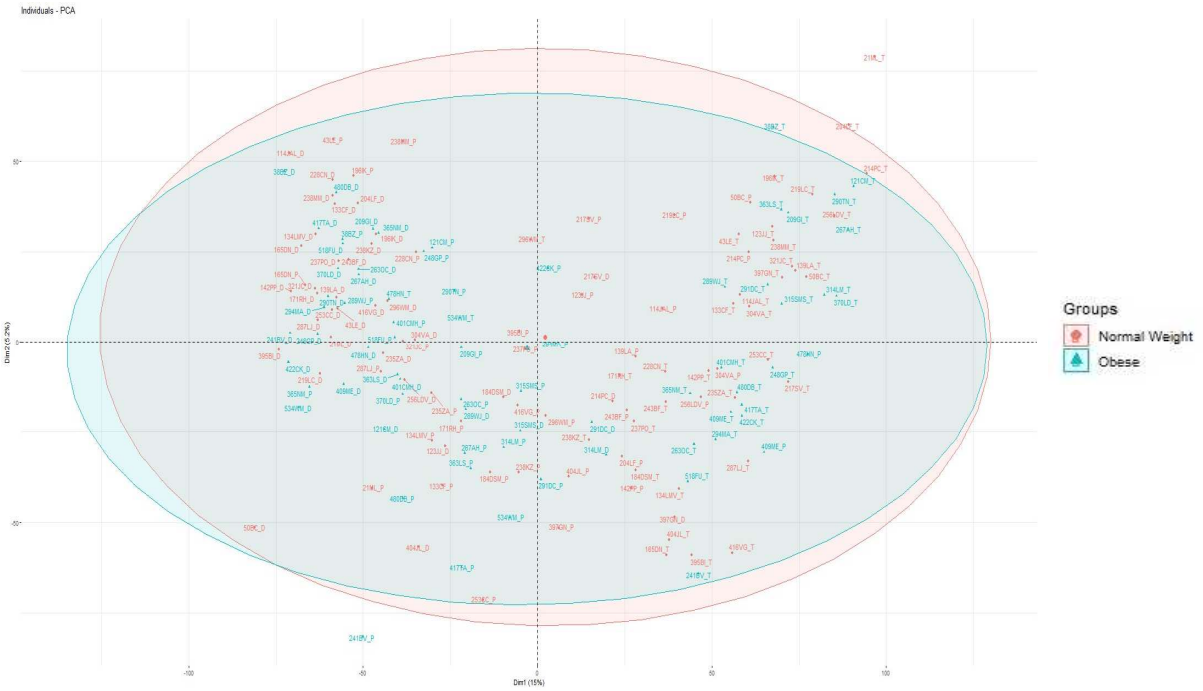
process, the adipose cells exhibit a “browning” effect, this is adopting the brown adipose tissue phenotype, while providing lipids to tumor cells and loss of their white adipocyte features [29–31]. From the genes in this module, we found the genes *PLIN1* and *PPARG* to be negatively contributing, which are highly expressed in white adipocytes[32].

The module adipokines, seemed to not be overdispersed in any of the performed comparisons. Nevertheless, it is worth to mention that molecules such as leptin, IL-6, ATX or TNF-α, have been related to cancer initiation and progression[31,33,34].

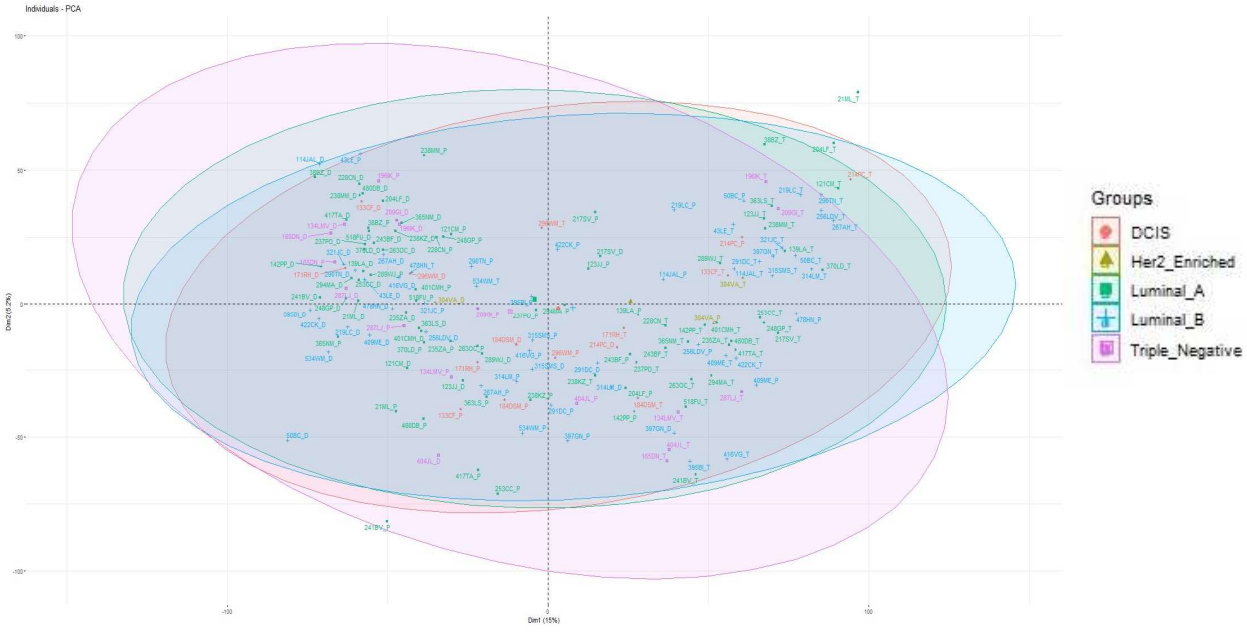
This study has some limitations, the sampling of the samples, that has been done macroscopically, could result in contamination of different cell types in the proximal adipose tissue samples. Alternatives to study the role of the CAA in tumor cells, could be to perform co-cultures with primary adipocytes from patients, and then extracting the pure adipocytes samples RNA and do the gene expression analysis. Additionally, other approaches, such as spatial transcriptomics, with proper annotation of the cells, could help to have a better picture of the CAA gene expression profiles. These two approaches, are already ongoing in the context of the LipoCanPredict project, and we expect to extend the presented findings.

In this study, we had described the gene expression profile of adipose tissue in the proximity of the tumor. In bulk, it appeared to have conserved adipose functions but also to interact with the tumor. The identified modules, metalloproteinases and thermogenesis, need to be validated to uncover the biology of the CAA.

E. SUPPLEMENTARY FIGURES



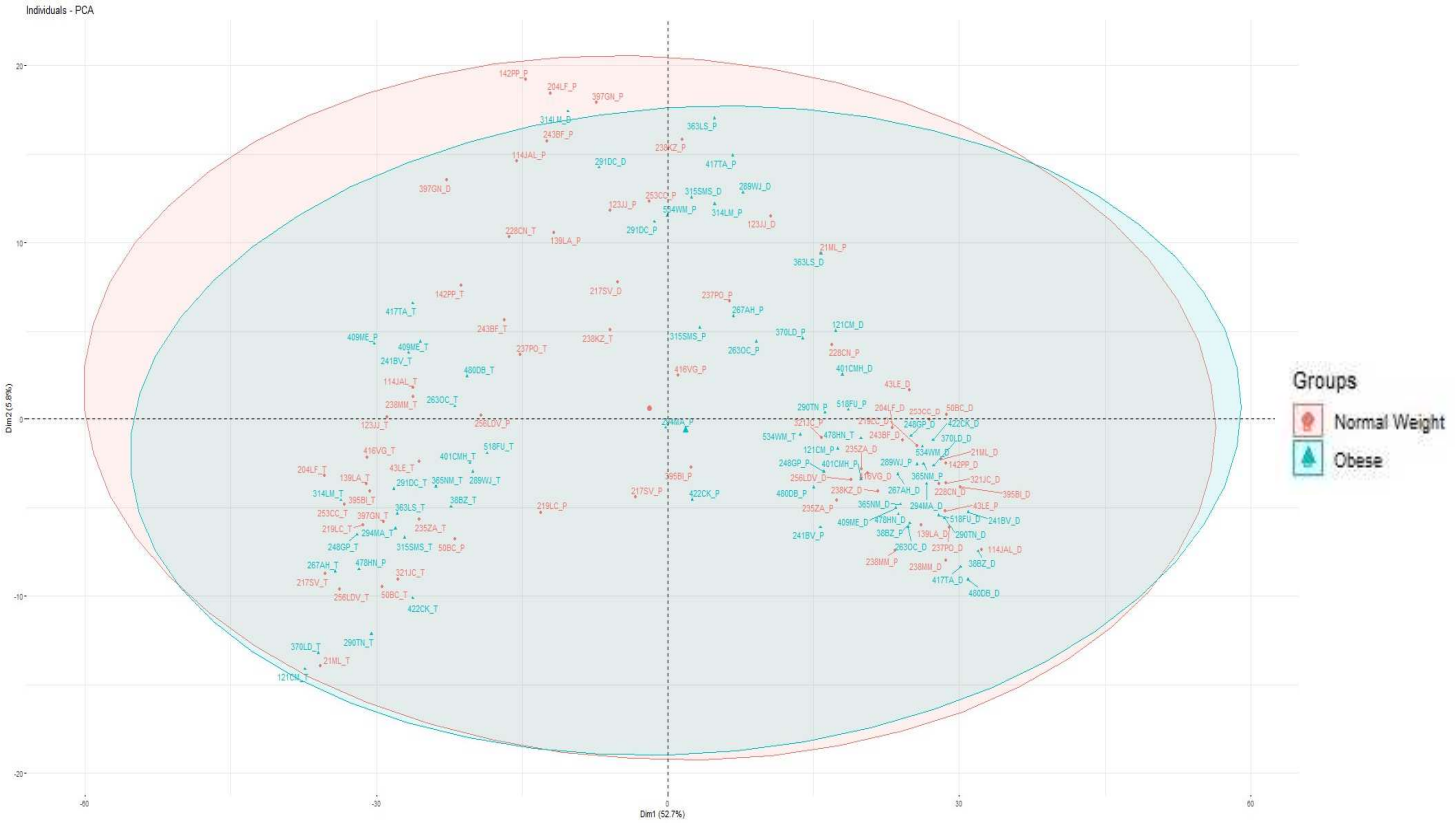
Supplementary figure 1. PCA of all the samples coloring the BMI classifications.



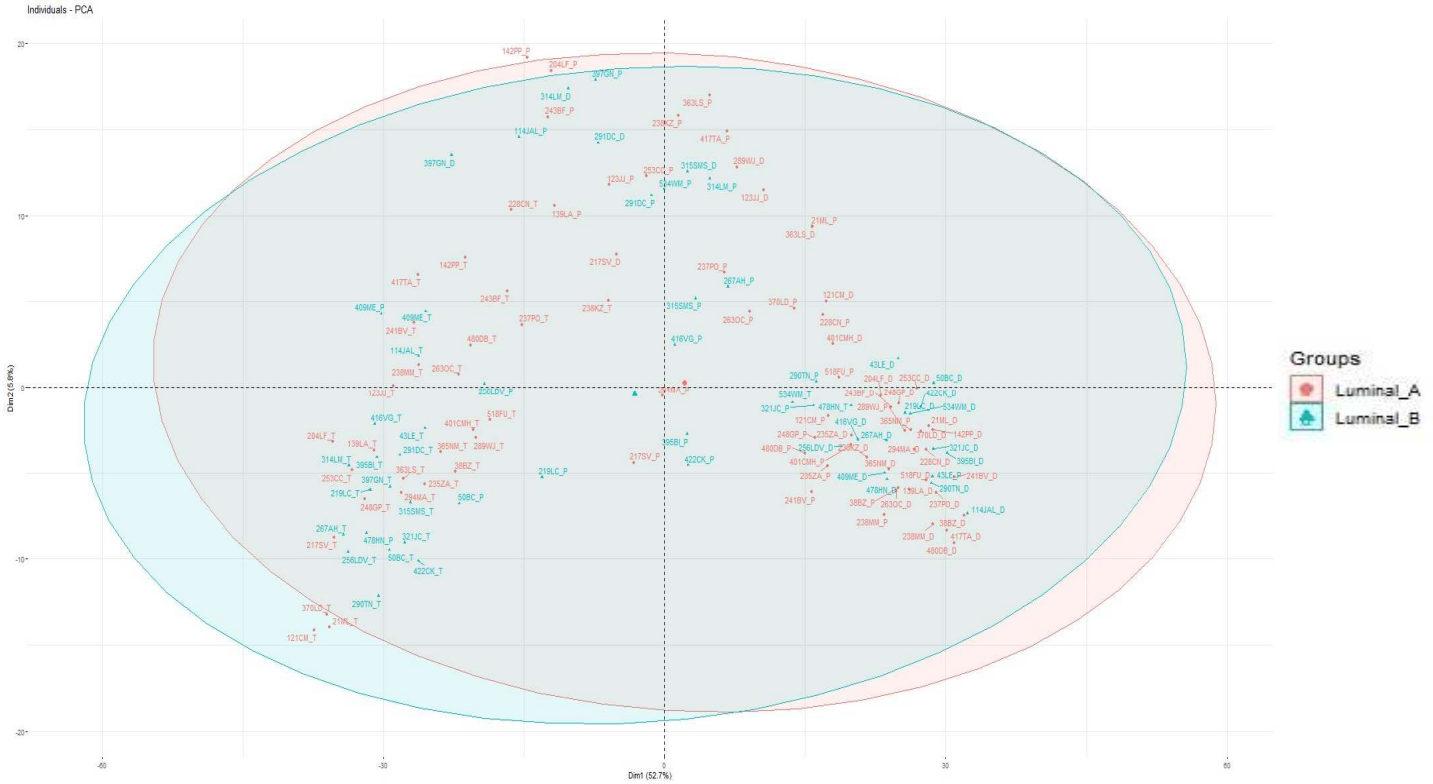
Supplementary figure 2. PCA featuring the molecular subtypes for all the samples**.**


Individuals - PCA
Dim1 (15%)
Dim2 (5.2%)
Groups
Distal Adipose Tissue Normal-weight
Distal Adipose Tissue Obese
Proximal Adipose Tissue Normal-weigh
Proximal Adipose Tissue Obese
Tumor Normal-weight
Tumor Obese

Supplementary figure 3. PCA plot emphasizing the tissue type and the patients' BMI category for all the samples.

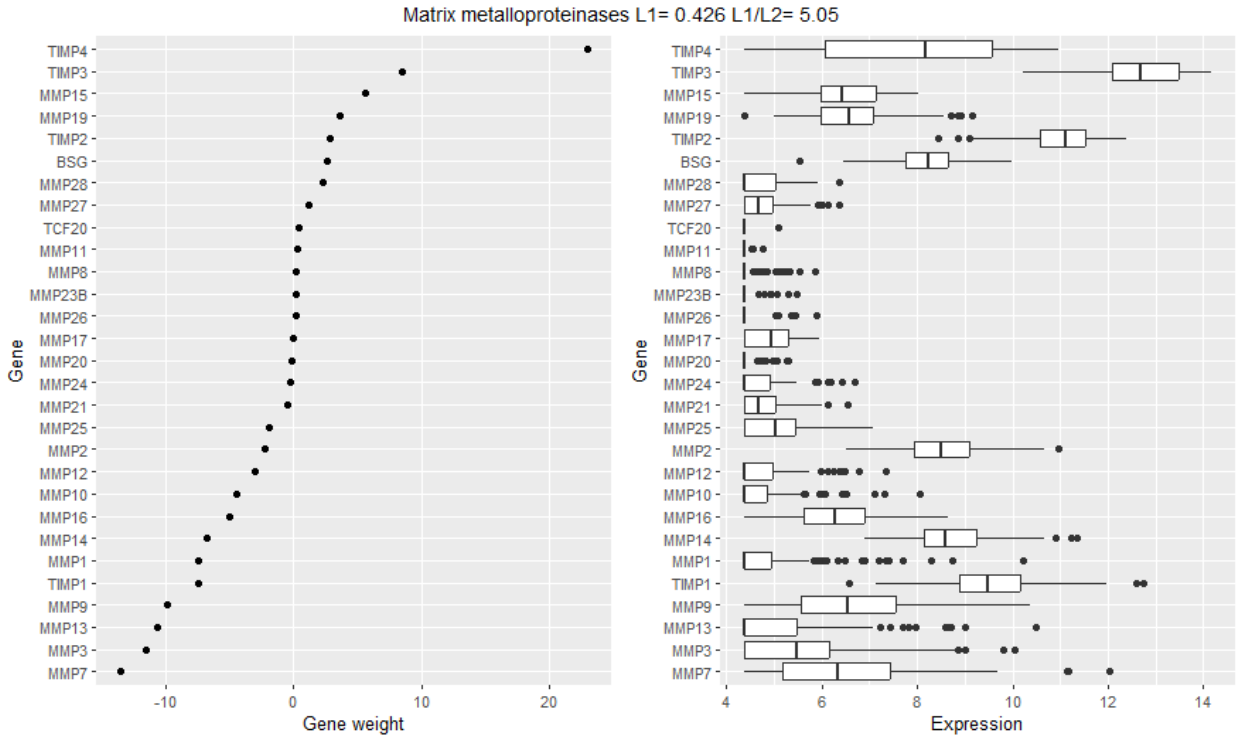


Supplementary figure 4. PCA of the samples from the patients' with tumors corresponding to luminal subtypes, coloring the BMI classifications.

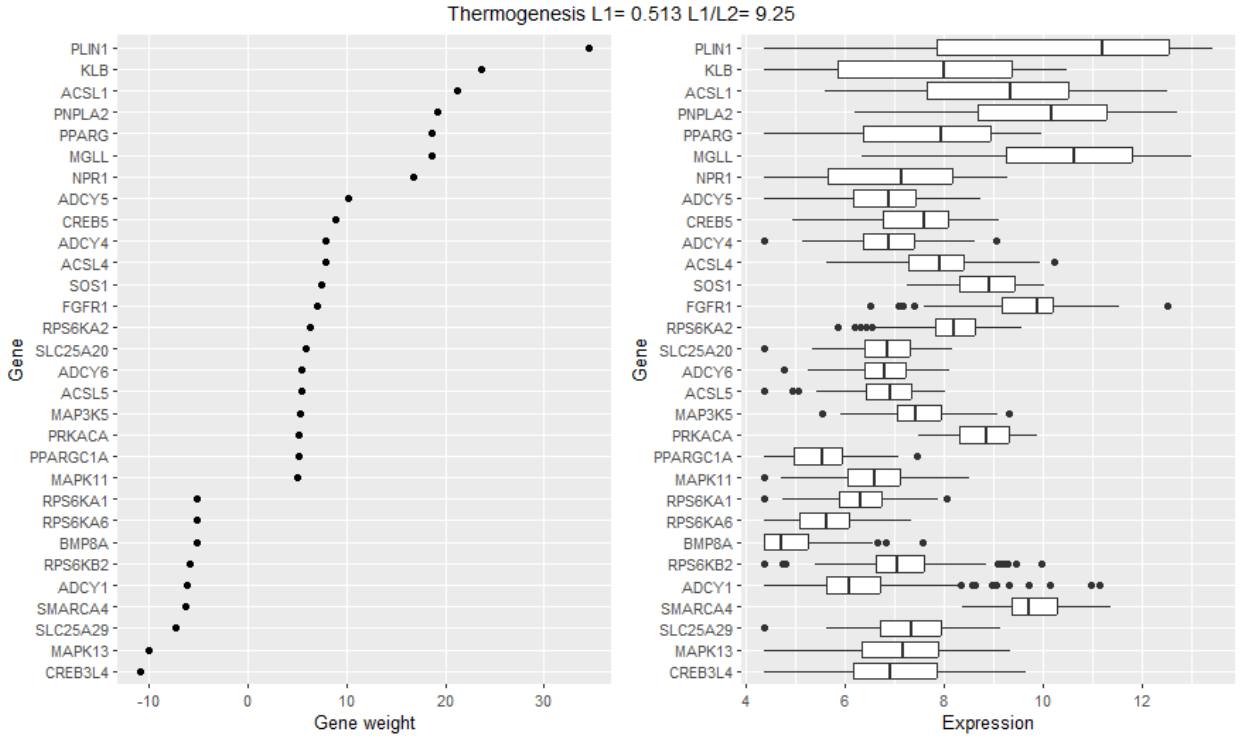


Supplementary figure 5. PCA plot of the samples from luminal patients displaying the class of luminal subtypes.

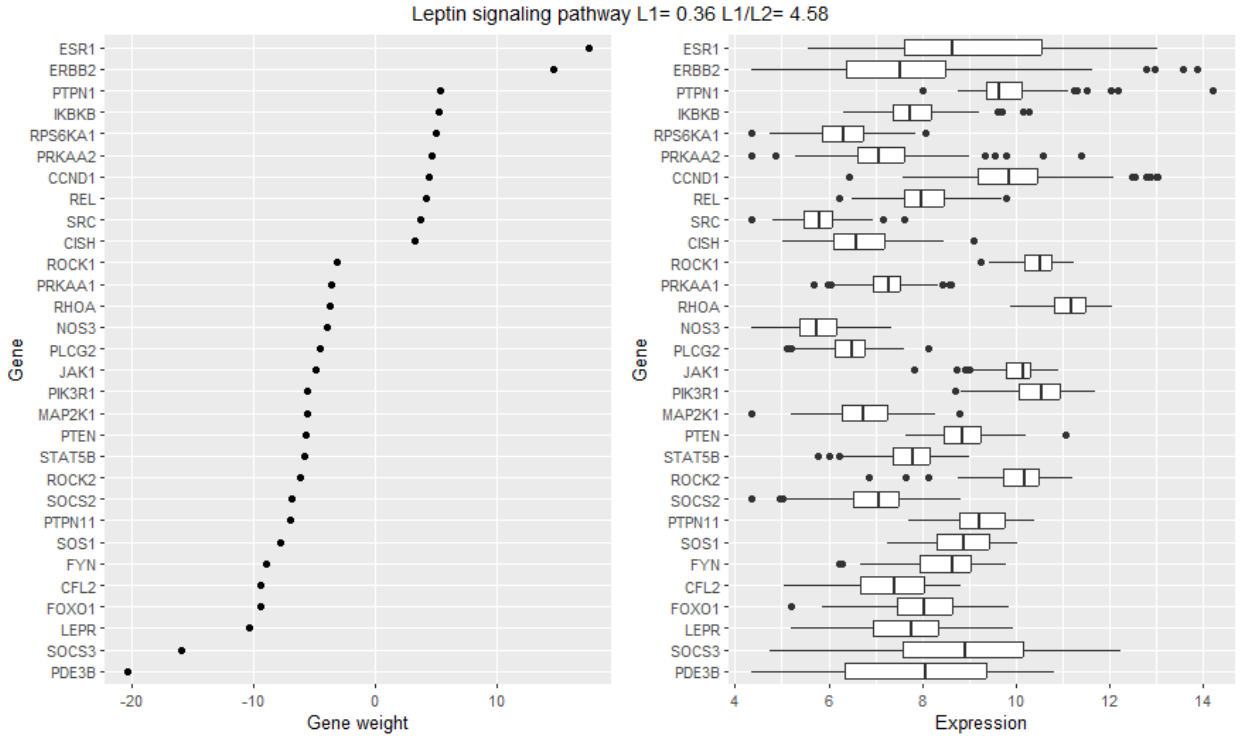
Supplementary figure 6. PCA plot emphasizing the tissue type and THE BMI classification for the luminal subset of patients.



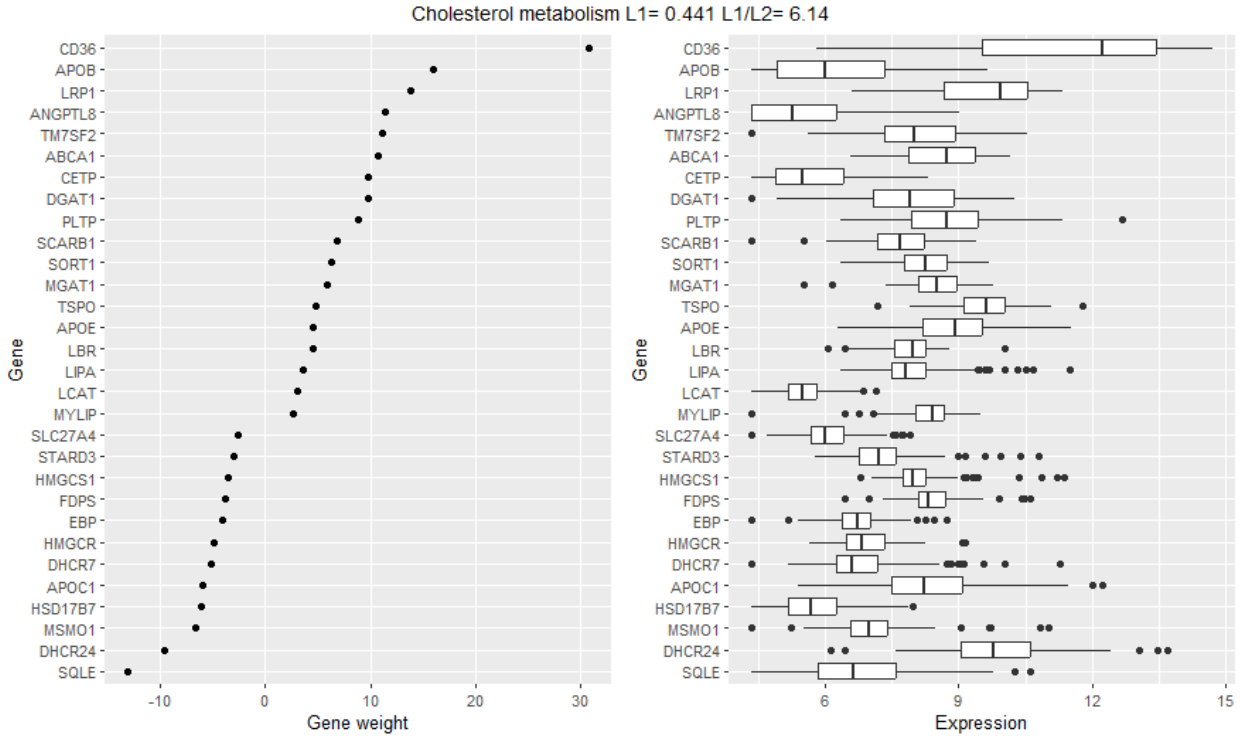
Supplementary figure 7. Gene weights and expression of the 30 most representative genes for the module matrix metalloproteinases for the three different tissue types.



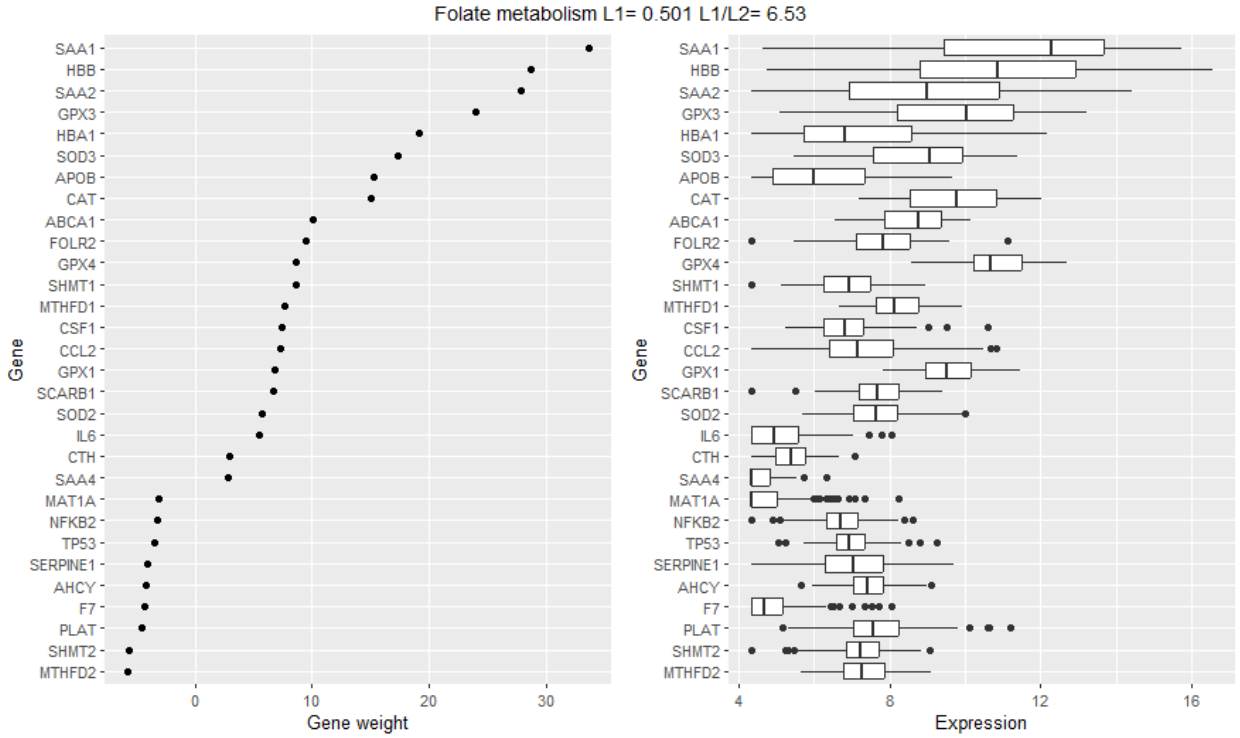
Supplementary figure 8. Gene weights and expression of the 30 most representative genes for the module Thermogenesis for the three different tissue types.



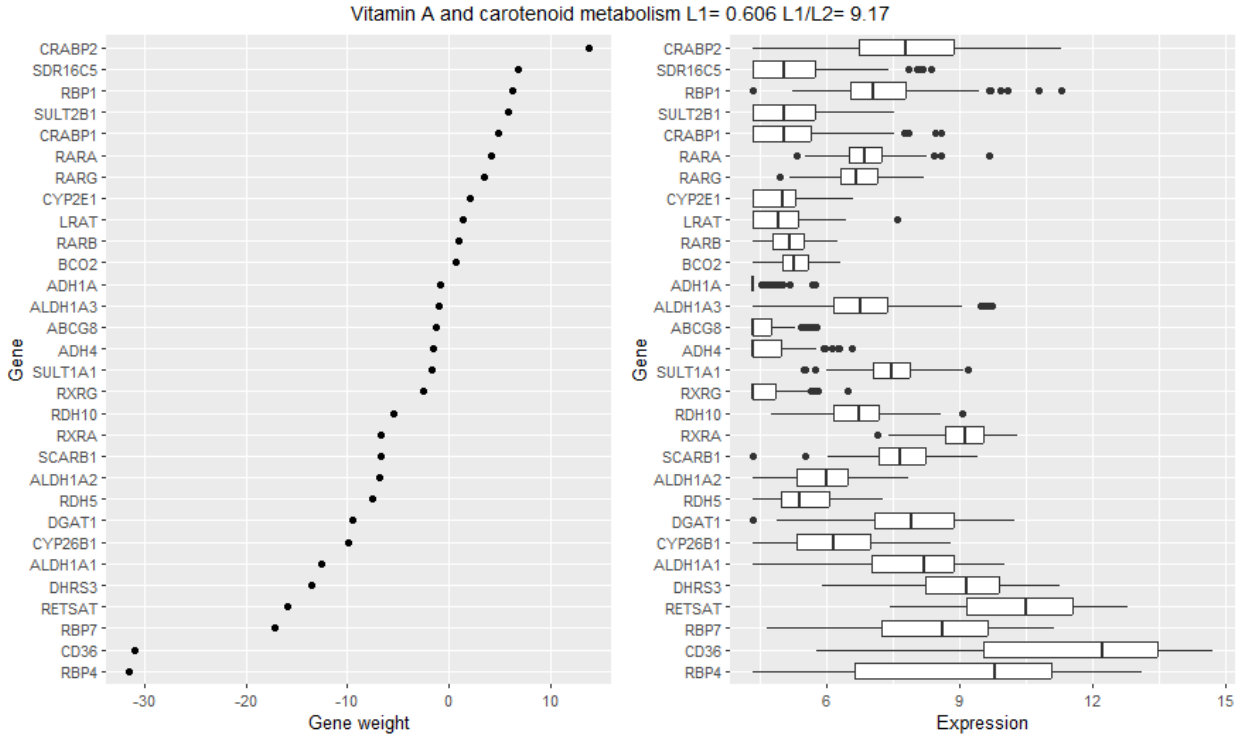
Supplementary figure 9. Gene weights and expression of the 30 most representative genes for the module Leptin signaling pathway for the three different tissue types.



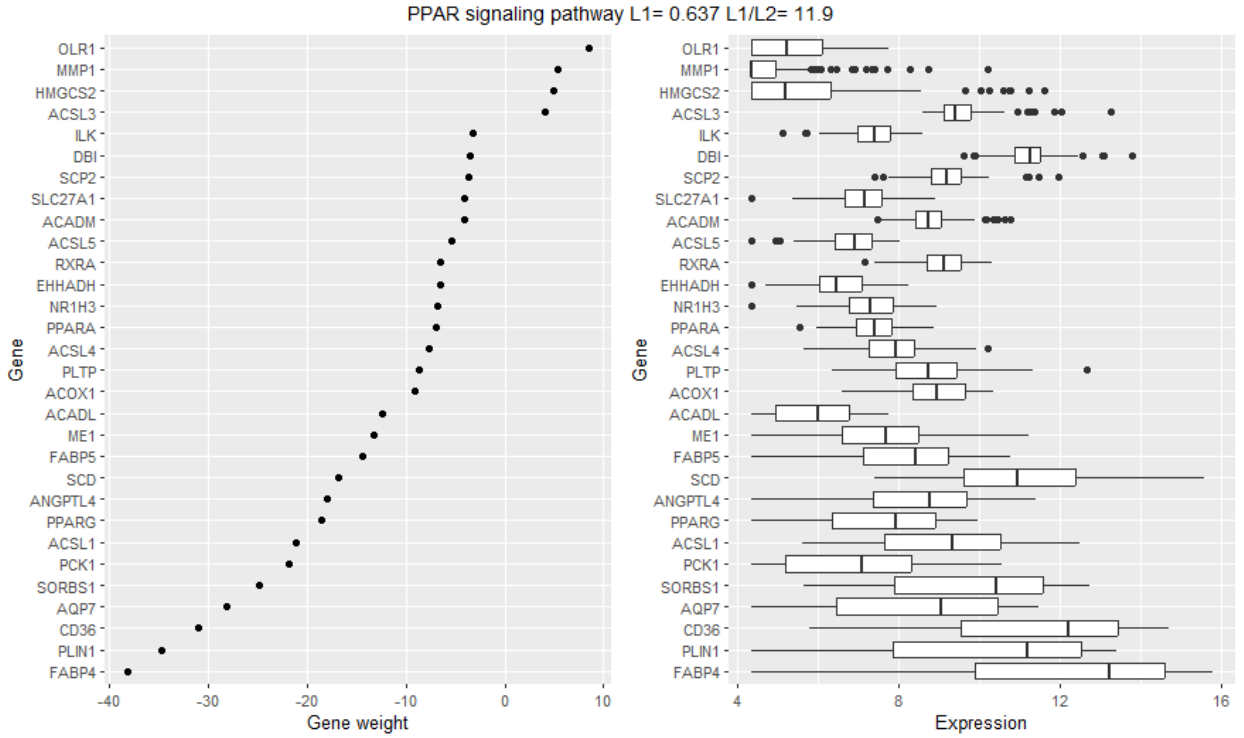
Supplementary figure 10. Gene weights and expression of the 30 most representative genes for the module Cholesterol metabolism for the three different tissue types.



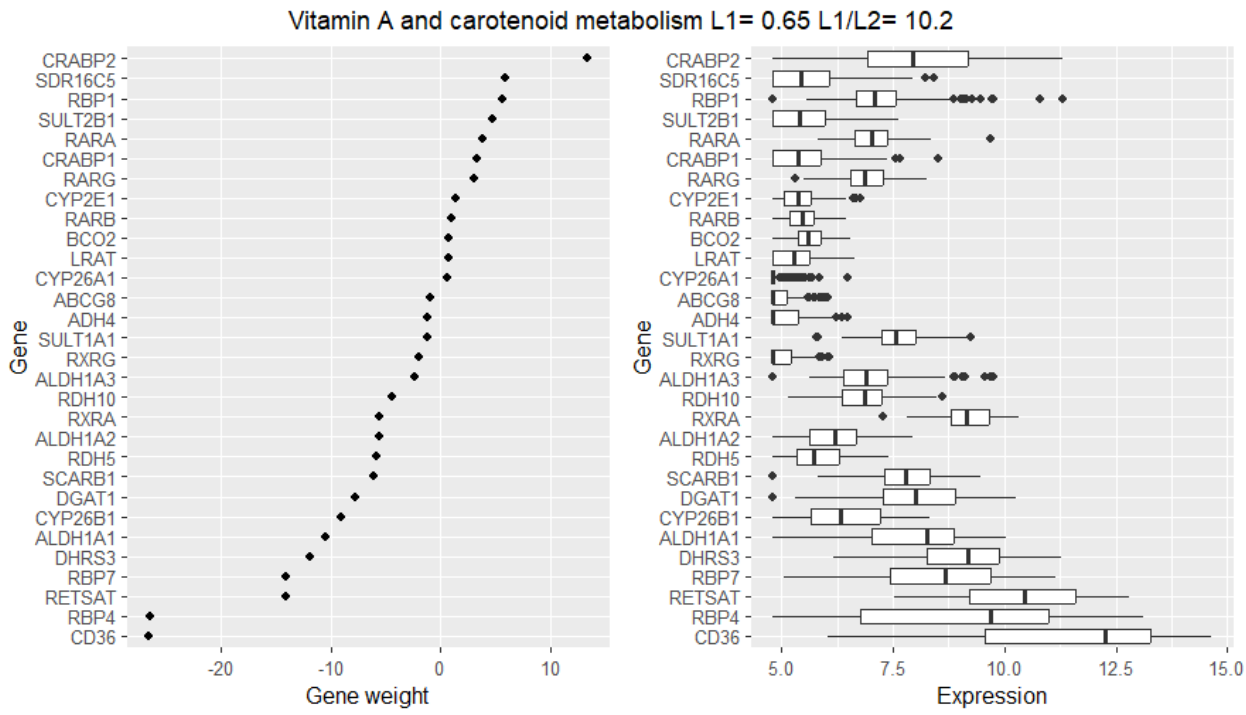
Supplementary figure 11. Gene weights and expression of the 30 most representative genes for the module Folate metabolism for the three different tissue types.



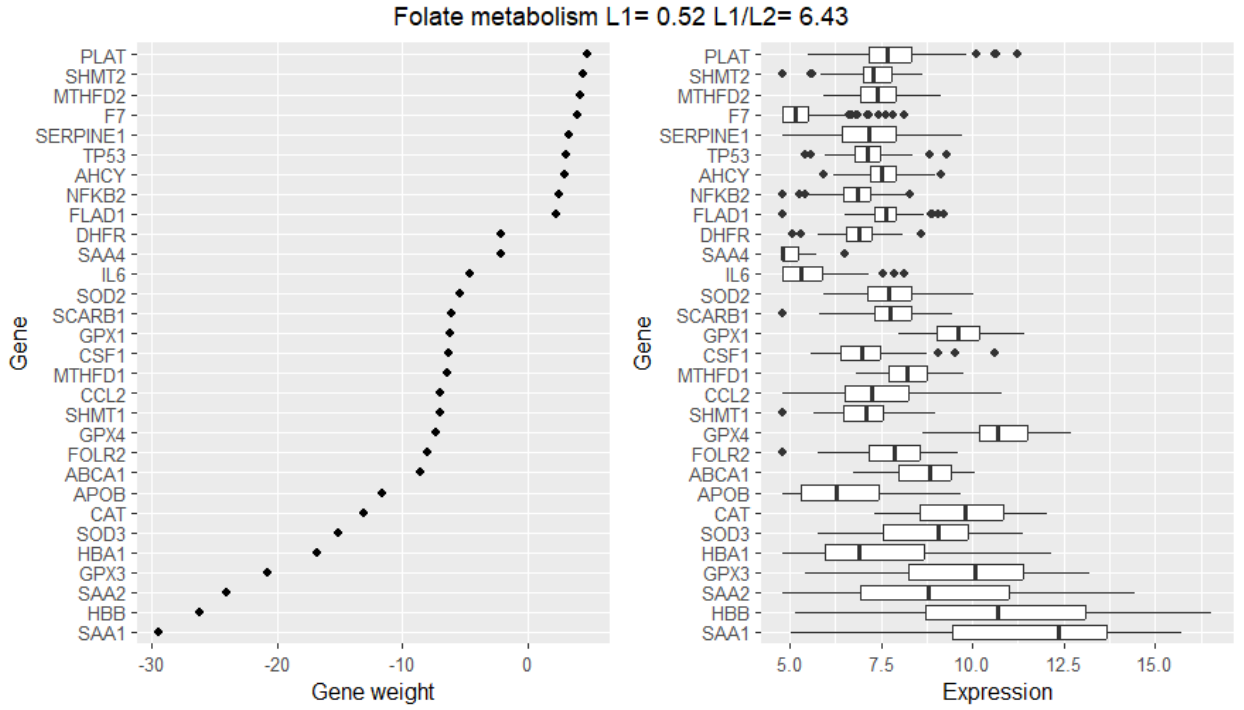
Supplementary figure 12. Gene weights and expression of the 30 most representative genes for the module Vitamin A and carotenoid metabolism for the three different tissue types.



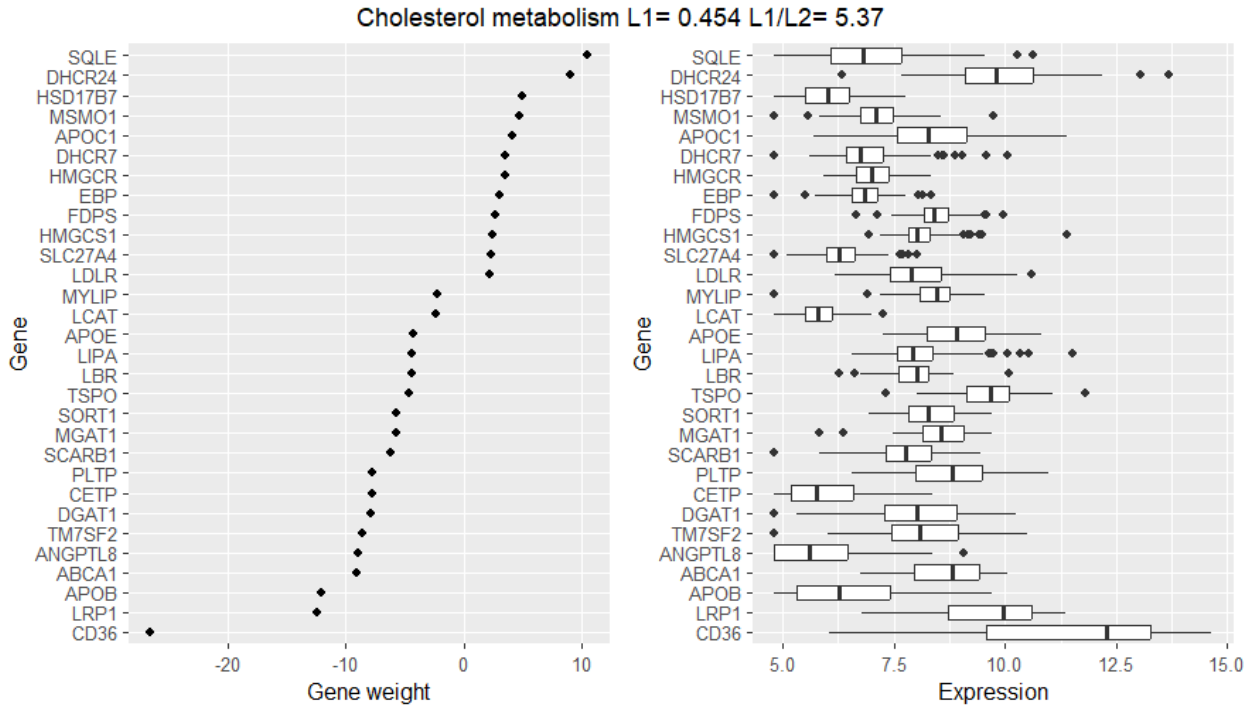
Supplementary figure 13. Gene weights and expression of the 30 most representative genes for the module PPAR signaling pathway for the three different tissue types.



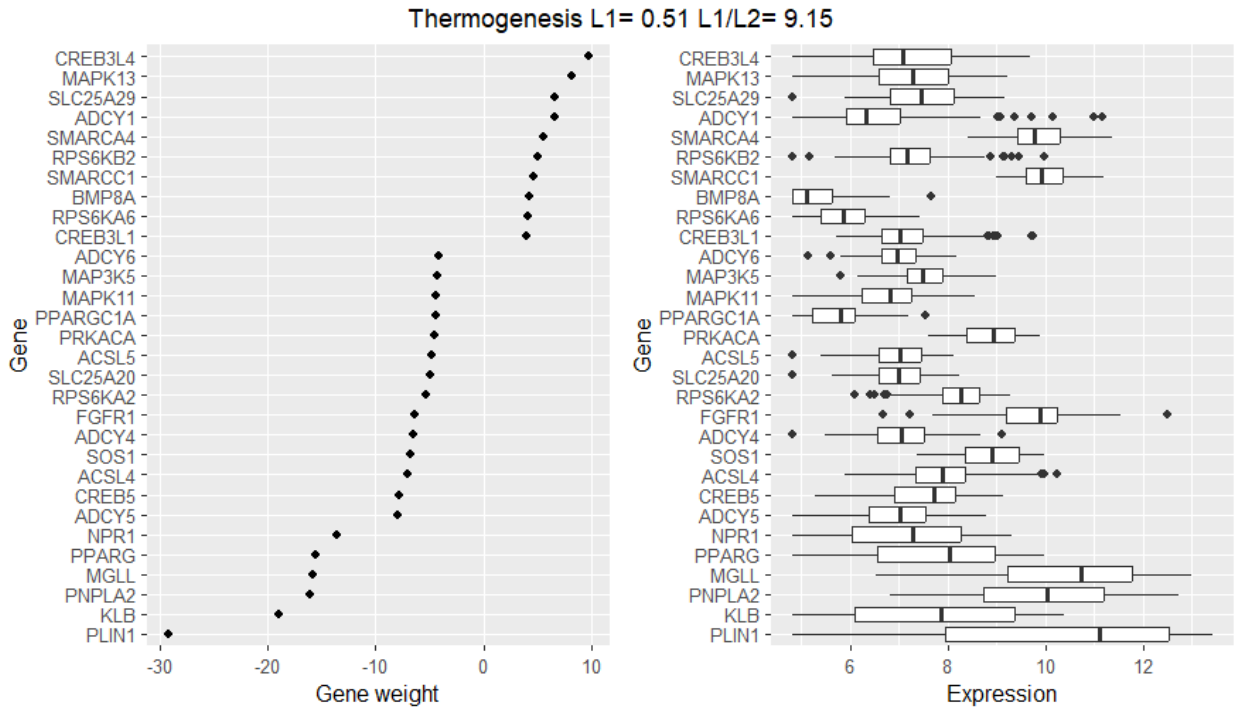
Supplementary figure 14. Gene weights and expression of the 30 most representative genes for the module Vitamin A and carotenoid metabolism for the three different tissue types in patients with a luminal subtype.



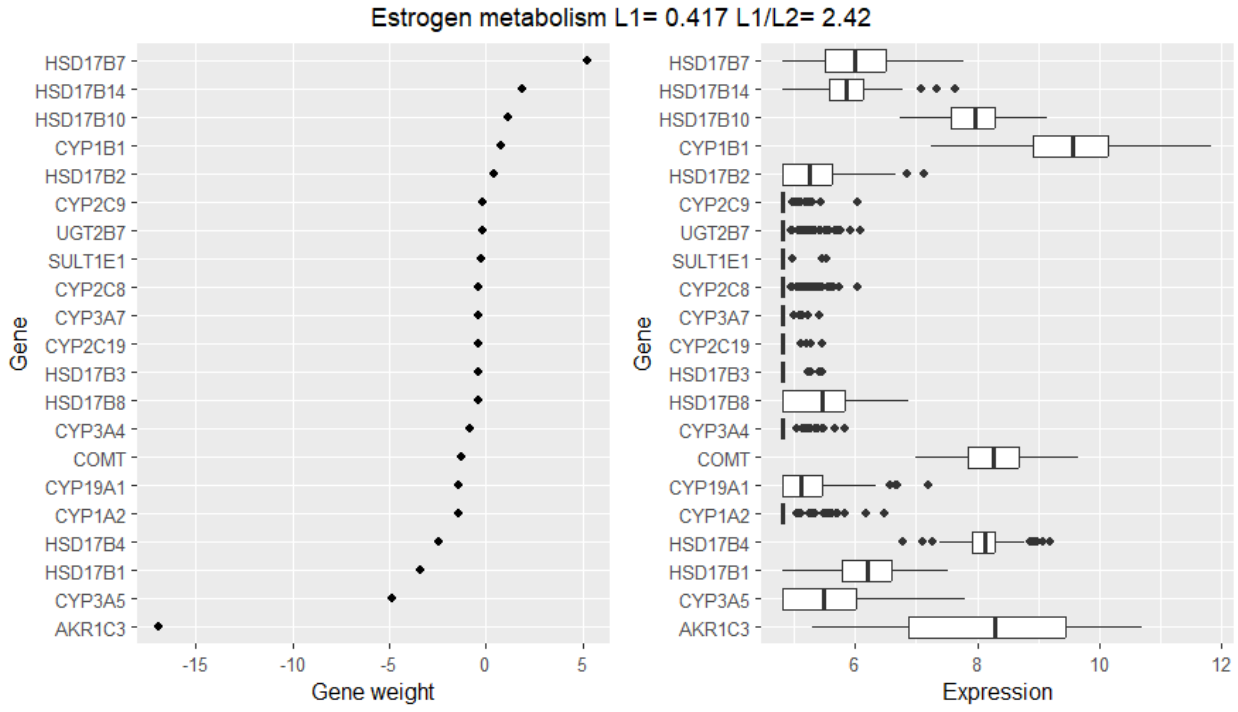
Supplementary figure 15. Gene weights and expression of the 30 most representative genes for the module Folate metabolism for the three different tissue types in patients with a luminal subtype.



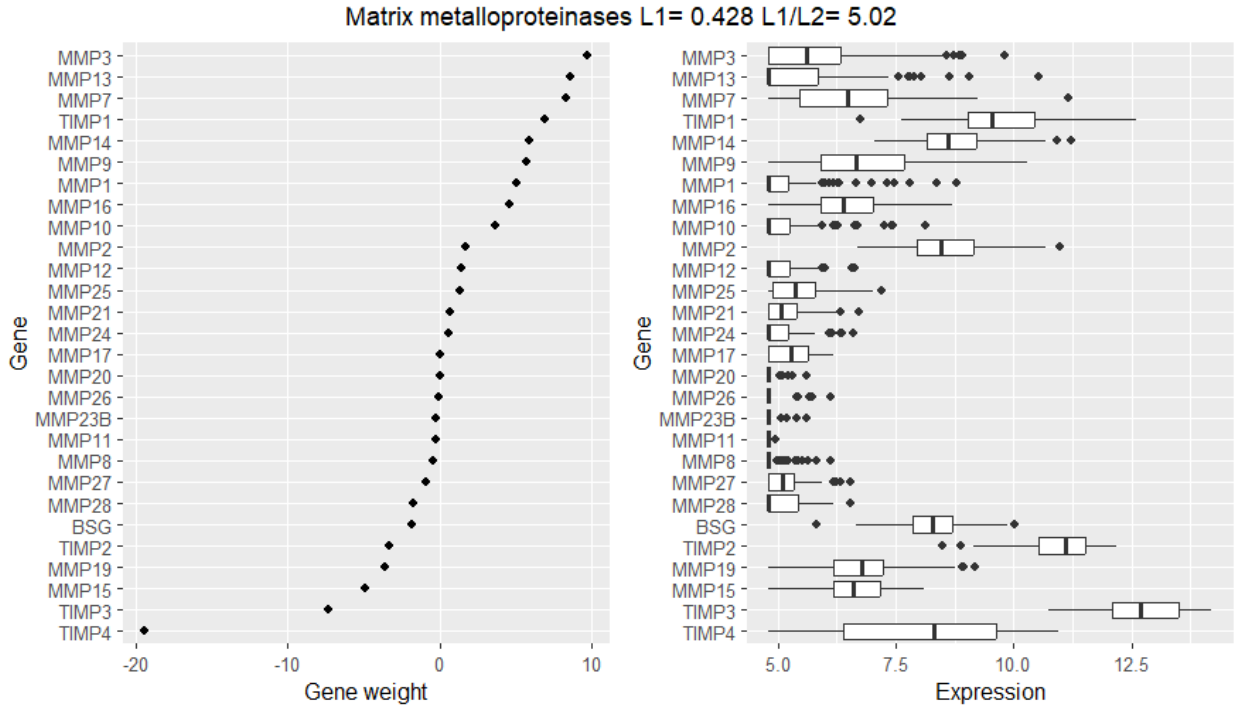
Supplementary figure 16. Gene weights and expression of the 30 most representative genes for the module Cholesterol metabolism for the three different tissue types in patients with a luminal subtype.



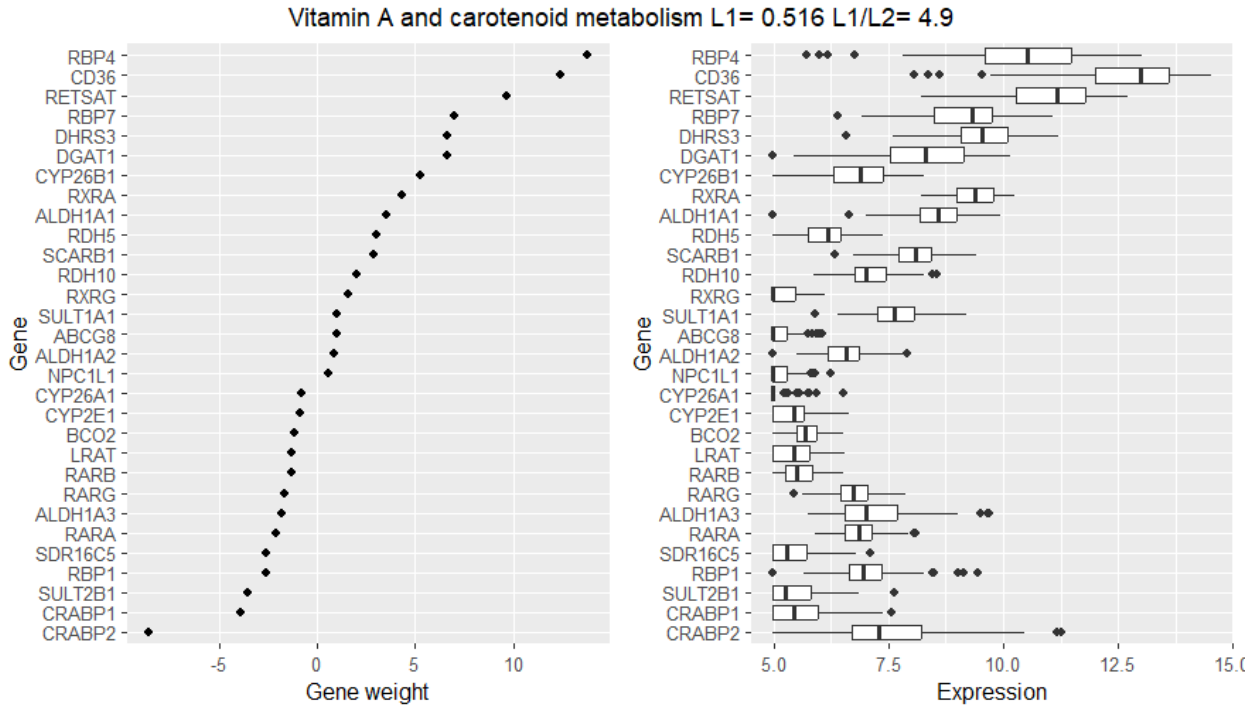
Supplementary figure 17. Gene weights and expression of the 30 most representative genes for the module Thermogenesis for the three different tissue types in patients with a luminal subtype.



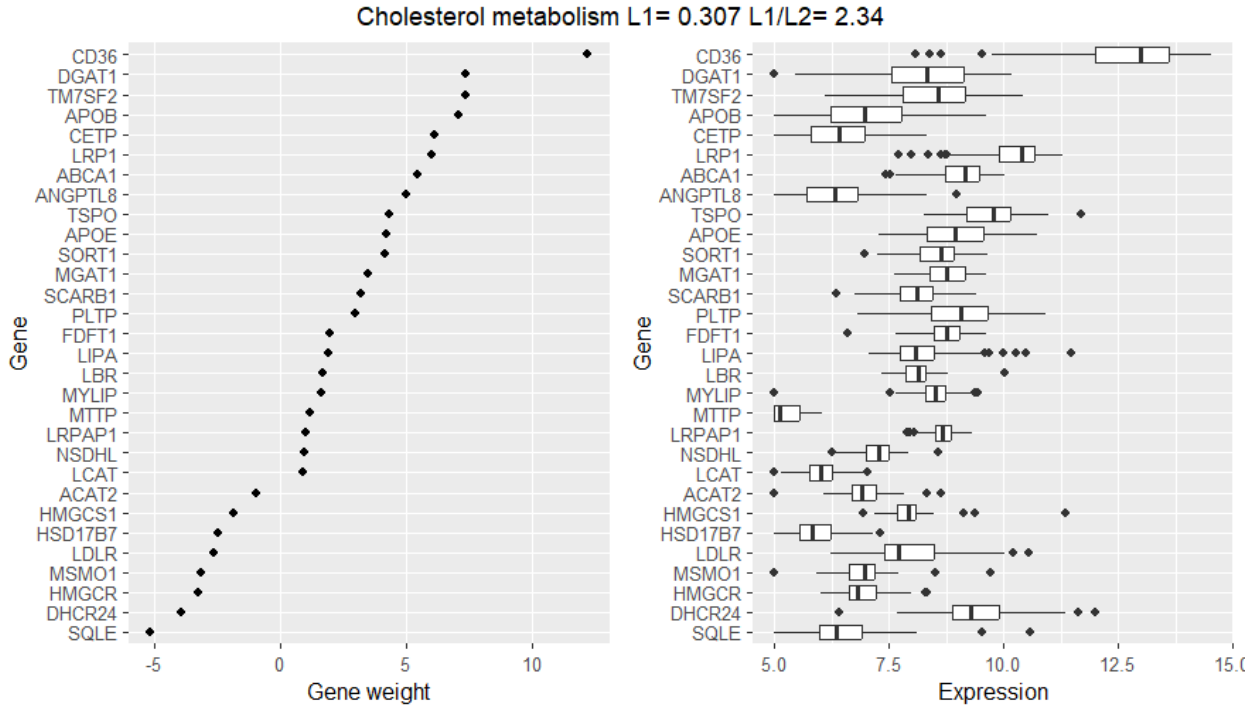
Supplementary figure 18. Gene weights and expression of the 30 most representative genes for the module Estrogen metabolism for the three different tissue types in patients with a luminal subtype.



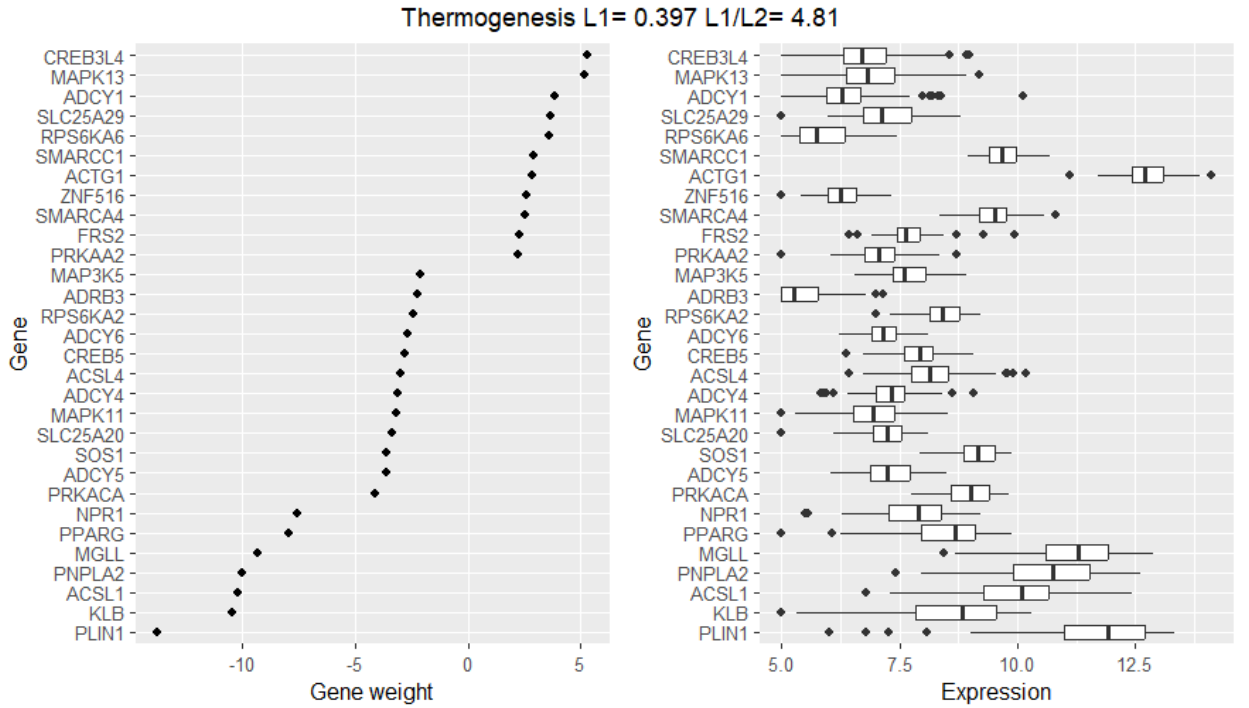
Supplementary figure 19. Gene weights and expression of the 30 most representative genes for the module Matrix metalloproteinases pathway for the three different tissue types in patients with a luminal subtype.



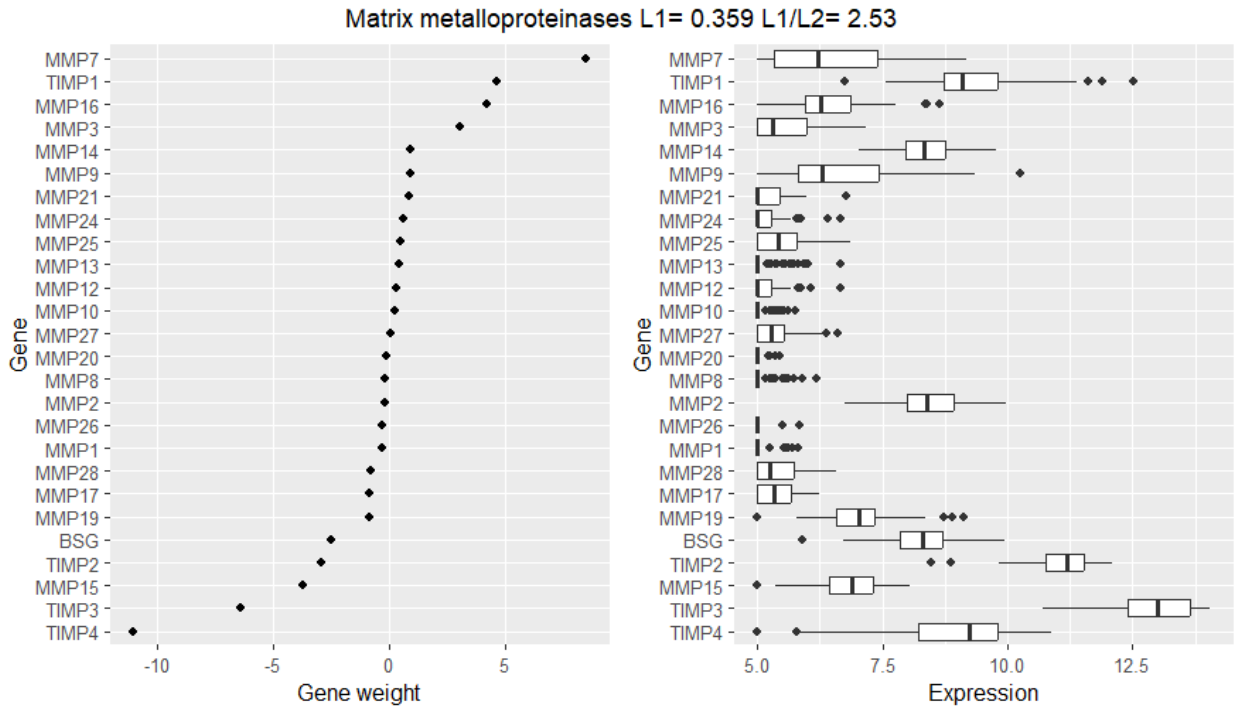
Supplementary figure 20. Gene weights and expression of the 30 most representative genes for the module Vitamin A and carotenoid metabolism for the proximal and distal adipose tissues.



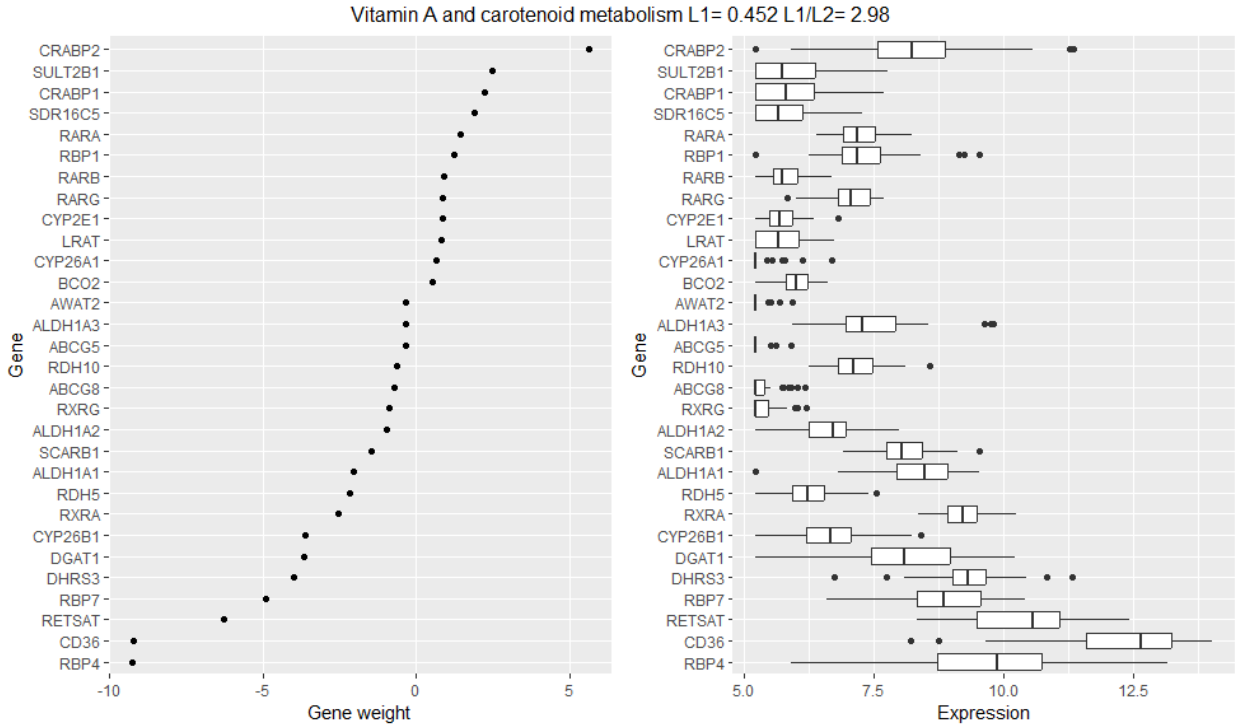
Supplementary figure 21 Gene weights and expression of the 30 most representative genes for the module Cholesterol metabolism for the proximal and distal adipose tissues.



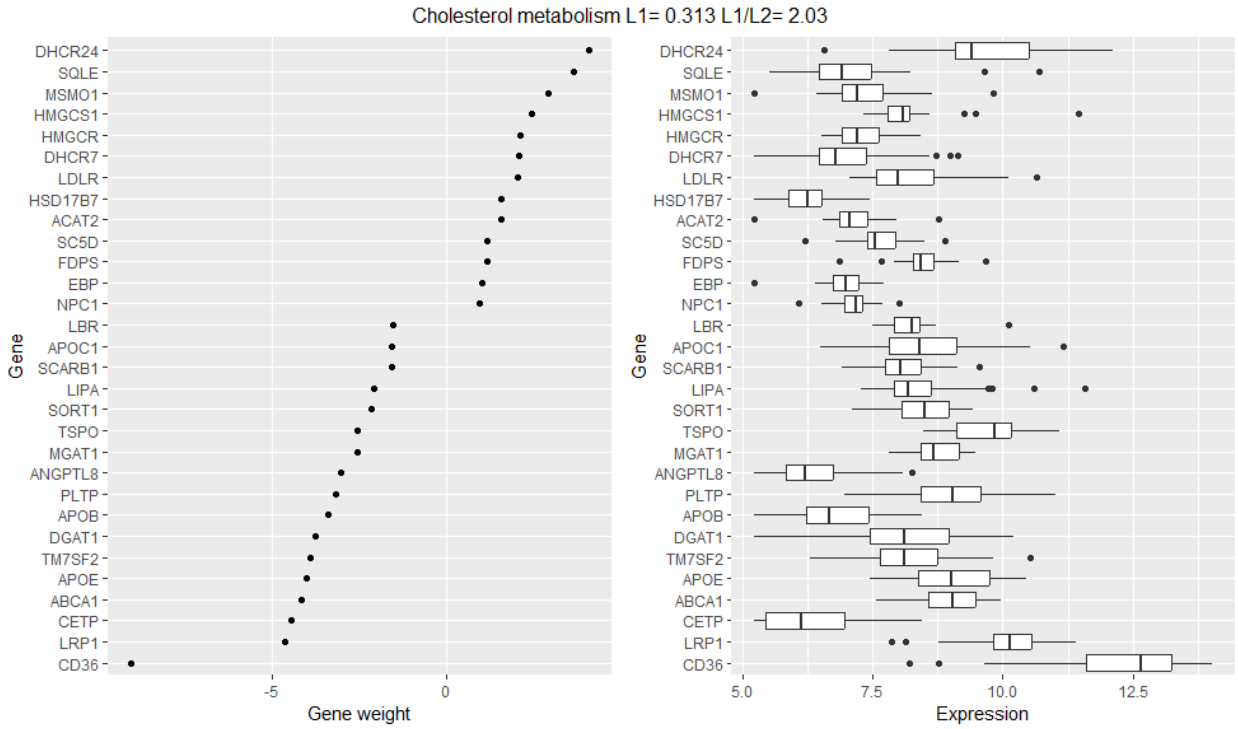
Supplementary figure 22. Gene weights and expression of the 30 most representative genes for the module Thermogenesis for the proximal and distal adipose tissues.



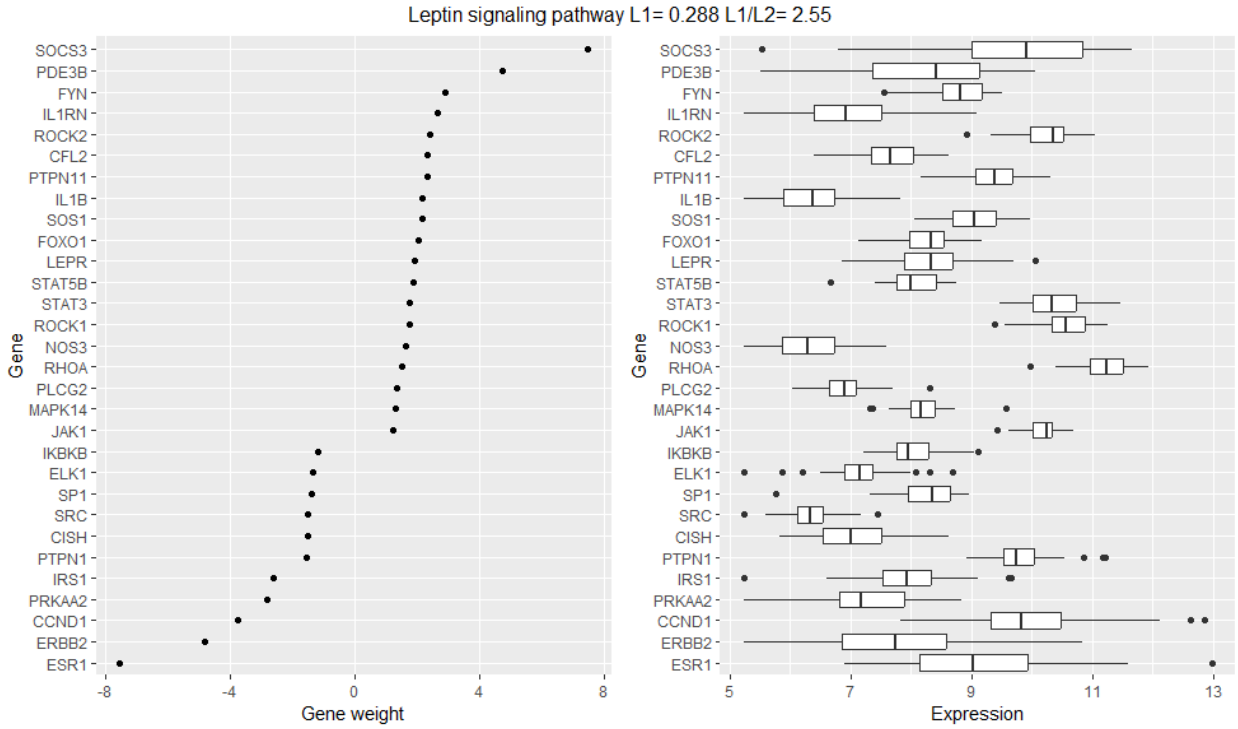
Supplementary figure 23. Gene weights and expression of the 30 most representative genes for the module Matrix metalloproteinases for the proximal and distal adipose tissues.



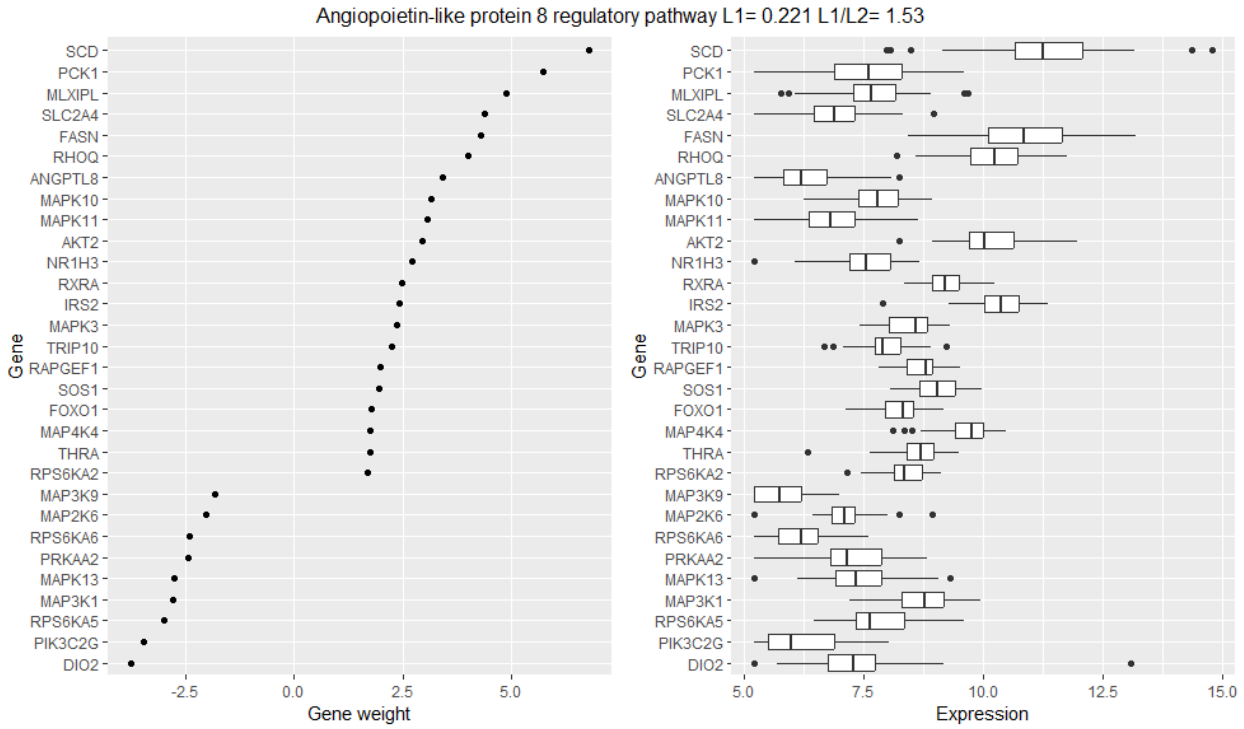
Supplementary figure 24. Gene weights and expression of the 30 most representative genes for the module Vitamin A and carotenoid metabolism for the proximal adipose tissues when compared by BMI categories.



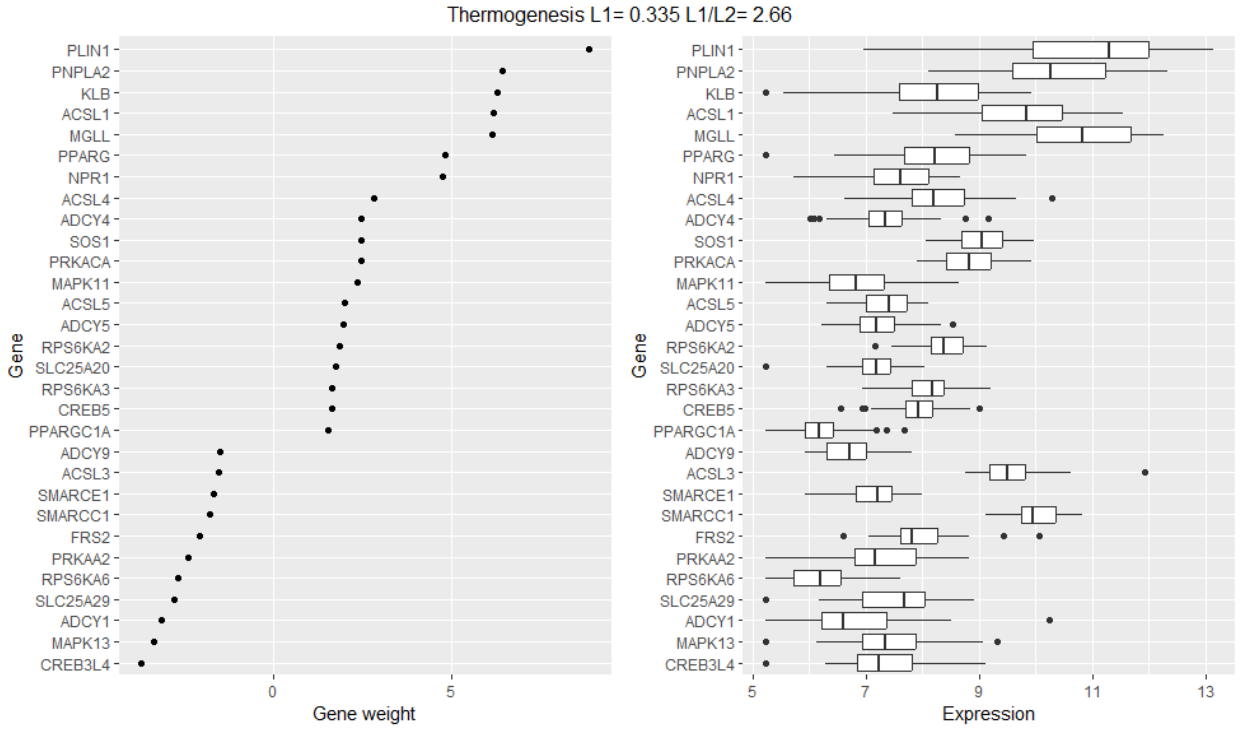
Supplementary figure 25. Gene weights and expression of the 30 most representative genes for the module Cholesterol metabolism for the proximal adipose tissues when compared by BMI categories.



Supplementary figure 26. Gene weights and expression of the 30 most representative genes for the module Leptin signaling pathway for the proximal adipose tissues when compared by BMI categories.



Supplementary figure 27. Gene weights and expression of the 30 most representative genes for the module Angiopoietin-like protein 8 regulatory pathway for the proximal adipose tissues when compared by BMI categories.



Supplementary figure 28. Gene weights and expression of the 30 most representative genes for the module Thermogenesis for the proximal adipose tissues when compared by BMI categories.

F. REFERENCES

1. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).

2. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, B. F. No Title. *(2020). Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer.* https://gco.iarc.fr/today.

3. Jesinger, R. A. Breast anatomy for the interventionalist. *Tech. Vasc. Interv. Radiol.* **17**, 3–9 (2014).

4. McGhee, D. E. & Steele, J. R. Breast biomechanics: What do we really know? *Physiology* **35**, 144–156 (2020).

5. Khan, Y. S. & Sajjad, H. *Anatomy, Thorax, Mammary Gland. StatPearls* (StatPearls Publishing, 2023).

6. Harbeck, N. *et al. Breast cancer. Nature Reviews Disease Primers* vol. 5 (2019).

7. Chamalidou, C. *et al.* Survival patterns of invasive lobular and invasive ductal breast cancer in a large population-based cohort with two decades of follow up. *The Breast* **59**, 294–300 (2021).

8. Orrantia-Borunda, E., Anchondo-Nuñez, P., Acuña-Aguilar, L. E., Gómez-Valles, F. O. & Ramírez-Valdespino, C. A. Subtypes of Breast Cancer. in *Breast Cancer* 31–42 (Exon Publications, 2022). doi:10.36255/exon-publications-breast-cancer-subtypes.

9. Dirat, B. *et al.* Cancer-associated adipocytes exhibit an activated phenotype and contribute to breast cancer invasion. *Cancer Res.* **71**, 2455–2465 (2011).

10. Wang, Y.-Y. *et al.* Adipose tissue and breast epithelial cells: A dangerous dynamic duo in breast cancer. *Cancer Lett.* **324**, 142–151 (2012).

11. Brown, K. A. & Scherer, P. E. Update on Adipose Tissue and Cancer. *Endocr. Rev.* **44**, 961–974 (2023).

12. Bouche, C. & Quail, D. F. Fueling the Tumor Microenvironment with Cancer-Associated Adipocytes. *Cancer Res.* **83**, 1170–1172 (2023).

13. Bochet, L. *et al.* Adipocyte-Derived Fibroblasts Promote Tumor Progression and Contribute to the Desmoplastic Reaction in Breast Cancer. *Cancer Res.* **73**, 5657–5668 (2013).

14. Pagnotta, P. *et al.* Peritumoral adipose tissue promotes lipolysis and white adipocytes browning by paracrine action. *Front. Endocrinol. (Lausanne).* **14**, 1144016 (2023).

15. A healthy lifestyle - WHO recommendations. https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations.

16. Hildebrandt, X., Ibrahim, M. & Peltzer, N. Cell death and inflammation during obesity: “Know my methods, WAT(son)”. *Cell Death Differ.* **30**, 279–292 (2023).

17. Ouchi, N., Parker, J. L., Lugus, J. J. & Walsh, K. Adipokines in inflammation and metabolic disease. *Nat. Rev. Immunol.* **11**, 85–97 (2011).

18. Harborg, S. *et al.* New Horizons: Epidemiology of Obesity, Diabetes Mellitus, and Cancer Prognosis. *J. Clin. Endocrinol. Metab.* (2023) doi:10.1210/clinem/dgad450.

19. Pietrzyk, L., Torres, A., Maciejewski, R. & Torres, K. Obesity and Obese-related Chronic Low-grade Inflammation in Promotion of Colorectal Cancer Development. *Asian Pac. J. Cancer Prev.* **16**, 4161–8 (2015).

20. Kawai, T., Autieri, M. V. & Scalia, R. Adipose tissue inflammation and metabolic dysfunction in obesity. *Am. J. Physiol. - Cell Physiol.* **320**, C375–C391 (2021).

21. Martignetti, L., Calzone, L., Bonnet, E., Barillot, E. & Zinovyev, A. ROMA: Representation and quantification of module activity from target expression data. *Front. Genet.* **7**, 1–12 (2016).

22. Najm, M. *et al.* Representation and quantification Of Module Activity from omics data with rROMA. *bioRxiv* 2022.10.24.513448 (2023) doi:10.1101/2022.10.24.513448.

23. Dirat, B. *et al.* Cancer-associated adipocytes exhibit an activated phenotype and contribute to breast cancer invasion. *Cancer Res.* (2011) doi:10.1158/0008-5472.CAN-10-3323.

24. Gonzalez-Avila, G., Sommer, B., García-Hernández, A. A. & Ramos, C. Matrix metalloproteinases' role in tumor microenvironment. *Adv. Exp. Med. Biol.* **1245**, 97–131 (2020).

25. Juárez-Cruz, J. C. *et al.* Leptin induces cell migration and invasion in a FAK-Src-dependent manner in breast cancer cells. *Endocr. Connect.* **8**, 1539 (2019).

26. Jafari, N. *et al.* Adipocyte-derived exosomes may promote breast cancer progression in type 2 diabetes. *Sci. Signal.* **14**, eabj2807 (2021).

27. Dirat, B. *et al.* Cancer-Associated Adipocytes Exhibit an Activated Phenotype and Contribute to Breast Cancer Invasion. *Cancer Res.* **71**, 2455–2465 (2011).

28. Fenzl, A. & Kiefer, F. W. Brown adipose tissue and thermogenesis. *Horm. Mol. Biol. Clin. Investig.* **19**, 25–37 (2014).

29. Álvarez-artime, A., García-soler, B., Sainz, R. M. & Mayo, J. C. Emerging Roles for Browning of White Adipose Tissue in Prostate Cancer Malignant Behaviour. *Int. J. Mol. Sci.* **22**, (2021).

30. Song, T. & Kuang, S. Adipocyte dedifferentiation in health and diseases. *Clin. Sci. (Lond).* **133**, 2107 (2019).

31. Choi, J., Cha, Y. J. & Koo, J. S. Adipocyte biology in breast cancer: From silent bystander to active facilitator. *Prog. Lipid Res.* **69**, 11–20 (2018).

32. Dalen, K. T. *et al.* Adipose Tissue Expression of the Lipid Droplet–Associating Proteins S3-12 and Perilipin Is Controlled by Peroxisome Proliferator–Activated Receptor-γ. *Diabetes* **53**, 1243–1252 (2004).

33. Booth, A., Magnuson, A., Fouts, J. & Foster, M. Adipose tissue, obesity and adipokines: role in cancer promotion. *Horm. Mol. Biol. Clin. Investig.* **21**, 57–74 (2015).

34. Bochet, L. *et al.* Adipocyte-derived fibroblasts promote tumor progression and contribute to the desmoplastic reaction in breast cancer. *Cancer Res.* **73**, 5657–5668 (2013).

CHAPTER III. EXPLOITATION OF A SIGNALING NETWORK TO STUDY EMT AND SENESCENCE

A. INTRODUCTION

The vast majority of cancer deaths are due to Metastasis. This process can be summarized as invasion, that happens at the primary local tumor site, intravasation, when it can enter the lymphatic ducts and the bloodstream, extravasation and colonization of another site [1]. For metastasis to occur, in epithelial tumor cells, must undergo an epithelial-mesenchymal transition (EMT), that provides the necessary modifications for these cells to invade and migrate distant tissues, the activation of this program is crucial for metastasis to happen [2].

Cellular senescence is a stress response, characterized by cell cycle arrest, mostly in the G1 phase[3]. Many insults can trigger senescence, such as oxidative stress, that emerges caused by the excessive production of reactive oxygen species (ROS) [4]. Oxidative stress has also been associated with tumor progression affecting the stroma and the extracellular matrix (ECM) [5]. Senescent cells have been described to actively communicate with neighboring cells through molecules that have been called senescence-associated secretory phenotype (SASP) [6]. The SASP is composed of multiple molecules, mostly involved in immune functions, matrix remodeling and angiogenesis and it has been controversial if it exerts a role as anti-tumor or pro-tumor agent [6,7]. It has been also observed that senescent cells might induce, through SASP, specially IL6 and IL8, EMT even in non-aggressive epithelial tumor cells [8].

While evidence indicates a role for senescent stromal cells, in particular senescent fibroblasts and macrophages, participating in cancer progression, the probability that tumor cells themselves could be senescent also, has been disregarded since they have been considered to be more a defense mechanism against cancer and not a contributor. It remains to be studied if tumor cells that go through EMT could also be senescent.

In this chapter, I present the work of a collaborative project with M. Boissan. In this work, there was a computational studies part in which I contributed and is the body of this chapter, but also a very interesting fundamental biology part. This work started from observations in experimental data, using a cell model that had silenced a metastasis suppressor gene: *NME1*. NME1 belongs to the nucleoside diphosphate kinases (NDPK) family of enzymes, which catalyze the synthesis of nucleoside triphosphates, mostly GTP, from corresponding nucleoside diphosphates and ATP[9]. As previously stated, NME1 has been recognized as a metastasis suppressor gene since it has been observed to be lowly expressed in metastatic primary tumors[10,11], however, when its expression has been upregulated in multiple metastatic cell-lines, their metastatic potential has been reduced[12,13]. Previous experiments using a knocked-down cell lines for NME1 have shown that senescence, as well as invasion markers, were highly expressed in the cell lines with the NME1 knock-down (Figure 1).

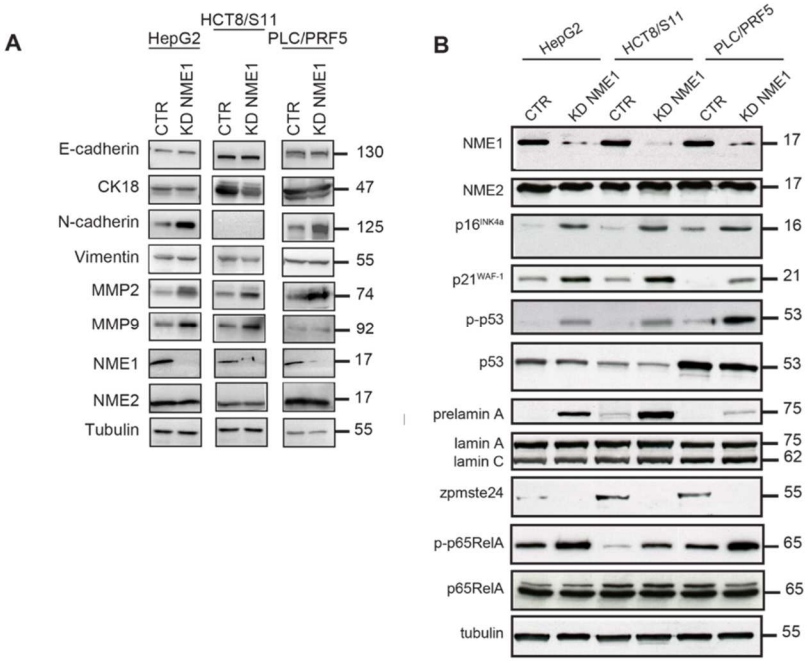


Figure 1. Loss of NME1 metastasis suppressor induces EMT and features of senescence in tumor cells. A) representative western blots from three independent experiments of HepG2 (CTR, KD NME1), HCT8/S11 (CTR, KD NME1) and PLC/PRF5 (CTR, KD NME1). Cell lysates revealed with antibodies directed against E-cadherin, Cytokeratin 18 (ck18), N-cadherin, Vimentin, MMP2, MMP9, NME1 and NME2. Tubulin was used as a loading control. molecular weights are in KDa. N-cadherin was undetectable in HCT8/S11 cells. B) representative western blots from three independent experiments of HepG2 (CTR, KD NME1), HCT8/s11 (CTR, KD NME1), and PLC/PRF5 (CTR, KD NME1), cell lysates revealed with antibodies directed against NME1, NME2, $p16^{ink4a}$, $p21^{waf-1}$, phospho-p53, p53, prelamin A, Lamin A/C, zmpste24, phospho-p65RelA and p65RelA. Tubulin was used as a loading control. molecular weights are in KDa. (Courtesy of M. Boissan)

To contribute with a systems biology approach, I enlarged a previously constructed network describing interactions resulting in EMT, senescence, cell polarity, extracellular matrix among others, that has been already published[14]. Firstly, I used a path analysis to reduce the network and explore the key players between NME1, senescence and EMT. Then using the information contained in the network, i.e. genes and processes (modules), I could investigate in a colorectal cancer gene expression dataset, the progression through the four stages via the activation of the different modules. From these analyses, it was possible to study how both processes could be related, as well as to pinpoint key players that can trigger them.

B. MATERIALS AND METHODS

EMT-senescence signaling network map construction and availability

The EMT and senescence signaling network map was created as an update of an already published EMT map [14]. The update of the canonical molecular mechanisms implicated in the EMT map and extension towards the senescence molecular mechanisms were done through literature curation and further knowledge formalization [15]. The map was build using standards indicated in: https://github.com/sysbio-curie/NaviCell/blob/master/map_construction_procedures/NaviCellMapperAdminGuide.pdf. The map has been constructed using the commonly accepted systems biology graphical notation (SBGN) syntax in the molecular diagrams editor software CellDesigner [16]. The different entities composing the map were assigned common identifiers converted into links to the corresponding entity descriptions in the HGNC, UniProt, Entrez, GeneCards, Reactome, and KEGG databases. The corresponding literature references, that support the map contents, are provided. The EMT-Senescence signalling network map is available online: https://acsn.curie.fr/navicell/maps/emt_senescence/master/index.html.

Map reduction to study players between NME1 and EMT and Senescence

The EMT and senescence map was used to obtain the shortest paths between the protein NME1 and the EMT and Senescence phenomena, with the aim to obtain the major players in such interactions. This was done using the Cytoscape software v2.8 and the BiNoM plugin using the Path Analysis functionality.

Gene expression data analysis and visualization using the EMT-senescence signaling network map

In order to estimate the involvement of different signaling pathways and molecular modules into the disease, we have performed a pathway enrichment study on the colorectal cancer gene expression data from TCGA (https://www.cbioportal.org), comparing groups of patients with different clinical characteristics. We have used the 'Representation and Quantification of Module Activity from Target Expression Data (ROMA)' software [17] to quantify the activity of gene sets included into the signaling pathways and molecular modules (named module activity from now and on). The analysis was executed using the R packages "rROMA" and "rROMADash" (available at: https://github.com/sysbio-curie/rRoma). To obtain the top contributing genes, ROMA calculates Pearson correlation coefficients for the expression of the gene across the data set and the ROMA score of the module. The visualization of the module activity scores was performed using Cytoscape v2.8 [18] and the BiNoM plugin [19], through the option "Stain CellDesigner Map". This TCGA dataset was used as well to calculate the mean expression between stages I and IV of certain genes of interest in some modules. One tailed T-tests using mean expression between the patients classified as stage I and stage IV using most representative genes for the modules: EMT regulators, senescence, lysosome-endosome and mitochondrial oxidative stress were performed. P-values were considered significant when <0.05.

C. RESULTS

We represented the genes involved in the form of a network by using the Systems Biology Graphical Notation syntax [20] in agreement with the Systems Biology Markup Language [21], which allowed us to perform further computational analysis. This structured signaling network map was based on data from >1400 publications, it covers 1636 biochemical reactions and represents 1067 gene and protein molecules. The signaling network map was separated into eleven modules representing structures or processes important in the EMT and senescence: adherens junctions, gap junctions, tight junctions, desmosomes, cell–matrix adhesions, cytoskeleton polarity, extracellular matrix, EMT regulators, lysosome-endosome, mitochondria oxidative stress and senescence. In this map some updates were done as NF-κB canonical and non-canonical dimers inducing SASP or p21 participation in the cell cycle.

From the reduction using the shortest paths we could observe that NME1 exerts an action in clathrin dependent endosomal pathway that preserves the adherens junctions preventing EMT. By inhibiting TIAM1 and activating ZMPSTE24 the cytoskeleton remodeling is interrupted so that the EMT is barred. The inhibition of NF-κB and HIF-1α leads to a reduced stimulation of transcription factors such as TWIST1, SNAI1, SNAI2, ZEB1 that block E-Cadherin, resulting in restraint EMT. By decreasing the mitochondrial ROS the downstream activation of NF-κB and SASP formation, as well as the activation of players such as MYC and MAPK14 that results in prevention of Senescence. Also we can see that the downstream hindrance of proteins such as p21CIP1, p16INK4A and downstream signaling prevent Senescence (Figure 2).

In order to have a global picture of these two processes, and to see if they are related, we proceeded to analyze a publicly available data-set from patients with colorectal cancer. We calculated the activity scores of the signaling pathways and modules by using the ROMA method in its implementation in R (see Material and Methods). ROMA maximizes the variance of the first principal component among the genes in a given set of genes (i.e. a module), to produce a score per module indicating the activity of that module. These activity scores were visualized in the context of the signaling network map at four stages (I, II, III and IV) of colorectal tumorigenesis by using the BiNoM plugin in Cytoscape (Figure 3).

Visual inspection of the signaling network maps and pair-wise comparisons of modules, using the obtained ROMA scores, found that the EMT regulators and senescence modules were both more active in stages III and IV than in stages I and II, suggesting that the genes and signaling pathways represented by these modules might be related to each other. Like the EMT regulators and senescence modules, the mitochondria oxidative stress module was more active in stages III and IV than in stages I and II. The activities of the senescence and mitochondria oxidative stress modules were indistinguishable from each other throughout. Three main pathways of oxidative stress-induced cell senescence are known[22]. One is the DNA damage response pathway, through which oxidative stress causes DNA damage that activates p53 and up-regulates p21 expression to cause senescence. The second is the NF-κB pathway, in which intracellular accumulation of reactive oxygen species activates the IκB kinase, which phosphorylates IκB to activate NF-κB and stimulate transcription of the genes associated with the senescence-associated secretory phenotype. The third is the p38 MAPK pathway, which is activated by reactive oxygen species and up-regulates expression of the senescence marker p16INK4a. Strikingly, the activity of the lysosome-endosome module also increased progressively up to stage IV. During senescence, lysosomes increase in number, pH and size resulting in increased SA-βGal activity [23], consistent with this observed increase in module activity through tumorigenesis.

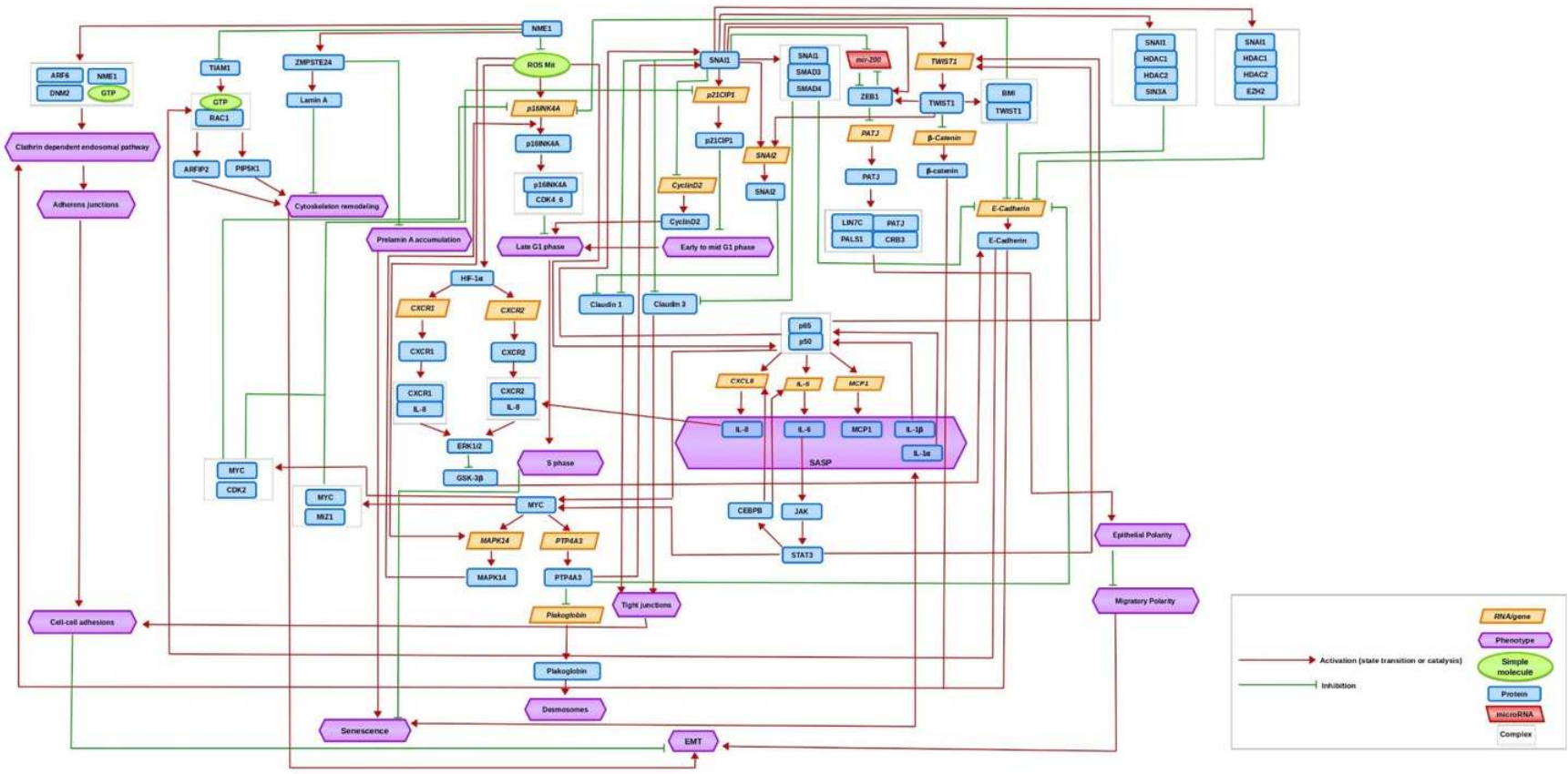


Figure 2. Diagram of the downstream players from nme1 that participate in senescence and EMT mechanisms. Some of the NME1 roles as metastatic suppressor. NME1 produces GTP that is used for some proteins involved in the clathrin dependent pathway. Also it participates by preventing the cytoskeleton remodeling, via induction of ZMPSTE24 and inhibition tiam1 but also Gelsolin (GSN, not shown). from this diagram we can infer that a major role of NME1 would be as inhibitor of ROS formation which are involved in the activity of different proteins that participate in senescence and EMT.

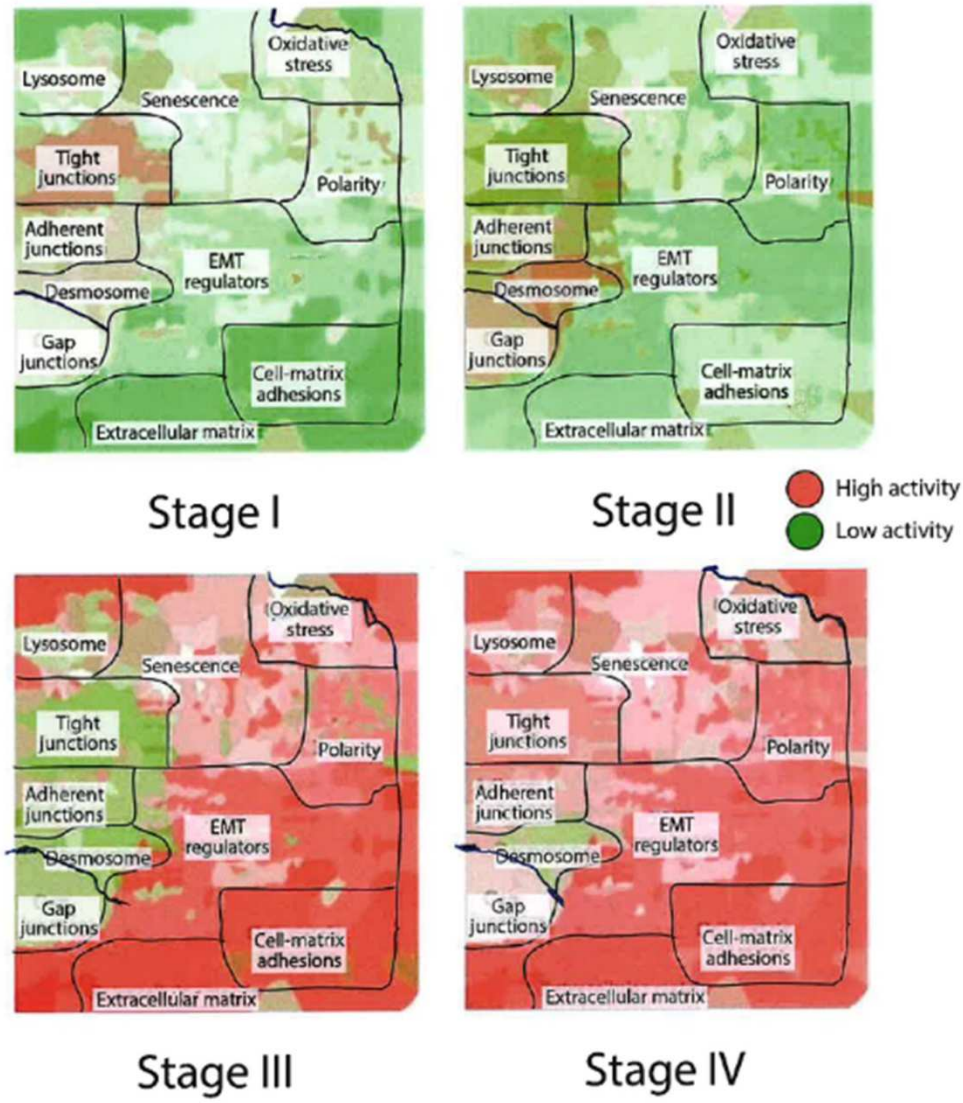


Figure 3. Shift of EMT and senescence module activity among the four stages in colorectal cancer patients. the obtained ROMA scores were used to visualize the module activity using the map information. in the first stages we observe a low activity of genes related to modules senescence, EMT regulators and mitochondria oxidative stress. notwithstanding, in the later stages (III and IV), there is an indication of higher activity in these modules.

The activities of the extracellular matrix and cell–matrix adhesions modules were strongly activated in stages III and IV when compared to stages I and II, consistent with remodeling of the extracellular matrix during cancer invasion. The activity of the cytoskeleton polarity module was also activated in stages III and IV when compared with stages I and II, consistent with switching from a non-migrating baso-apical polarized epithelial phenotype to a polarized migrating/invading mesenchymal phenotype during EMT and cancer invasion. Also in stage IV, we saw a slight increase in the activities of the gap junctions, adherens junctions and tight junctions modules when compared to stage III, but no change in the desmosomes module, whose activity was very low. All of these modules are related to cell–cell adhesion.

We correlated the scores for module activation with individual gene expression in order to extract the gene contribution per module . Briefly, we found the genes *MSN* and *ZEB2* contributed positively to the activity of the EMT regulators module whereas *CRB3* and *MAPK13* contributed negatively. In the senescence module, *HIF1A* and *IL6ST* contributed positively whereas *CDK4* and *NME1* contributed strongly negatively (Tables 1-4).

To identify any differences between the stages I and IV regarding the expression of key genes in the EMT regulators, senescence, mitochondria oxidative stress and lysosome-endosome modules, we computed the mean gene expression of those genes and compared them. In the EMT regulators module, we saw an increase in mean gene expression in stage IV including the pro-EMT genes *CDH2*, *VIM* and *MMP14*, and in genes encoding transcription factors that drive the EMT, especially *SNAI1*, *TWIST1* and *ZEB1*, when compared to stage I. In the senescence module, we observed increased expression in stage IV of genes that participate in the senescence-associated secretory phenotype: *IL1A*, *IL1B*, *IL6* and *IL8*. Also in the senescence module, *CDKN2A*, which encodes the senescence marker p16INK4a, was more highly expressed in stage IV than in stage I, indicating that some cells may arrest their cell cycle as the disease progresses, consistent with senescence. In the mitochondria oxidative stress module, we observed increased expression of the pro-oxidant genes *MPO*, *NOX4*, *NOX5* in stage IV and decreased expression of *SDHA* and *SDHB*, which encode subunits of complex II of the mitochondrial respiratory chain. Genes In the lysosome–endosome module, *CLN5*, *CTSF*, *GALNS*, *IDS* and *LAMP1* tended to be more highly expressed in stage IV, all of which encode proteins contained in lysosomes. Notably, *LAMP1* was highly expressed; it encodes LAMP1, the main marker of lysosomes. These data together indicate that senescence is related to EMT (Supplementary Tables 1-4).

MODULE EMT REGULATORS			
POSITIVE CONTRIBUTION		**NEGATIVE CONTRIBUTION**	
Gene	**Correlation**	**Gene**	**Correlation**
MSN	0.9101	CRB3	-0.4298
ZEB2	0.9033	GNG5	-0.3736
LAMC1	0.8809	MYB	-0.3662
ARHGAP31	0.8801	MAPK13	-0.3643
LAMA4	0.8774	MYL5	-0.3225

Table 1. Top contributing genes from the ROMA analysis in the module EMT regulators

MODULE SENESCENCE			
POSITIVE CONTRIBUTION		**NEGATIVE CONTRIBUTION**	
Gene	**Correlation**	**Gene**	**Correlation**
HIF1A	0.8906	CDK4	-0.5641
ROCK1	0.6734	NME1	-0.5357
IL6ST	0.6425	BAX	-0.3567
SNAI2	0.6235	AKT2	-0.3498
NOTCH2	0.5925	RCE1	-0.3315

Table 2. Top contributing genes from the ROMA analysis in the module senescence

MODULE LYSOSOME-ENDOSOME			
POSITIVE CONTRIBUTION		**NEGATIVE CONTRIBUTION**	
Gene	**Correlation**	**Gene**	**Correlation**
NOTCH1	0.9885	UBE2D2	-0.4482
NOTCH3	0.4551	UBE2D1	-0.4297
NOTCH4	0.3255	PSEN1	-0.3825
NOTCH2	0.2028	CTNND1	-0.102
DLL1	0.1901		

Table 3. Top contributing genes from the ROMA analysis in the module lysosome-endosome.

MODULE MITOCHONDRIA OXIDATIVE STRESS			
POSITIVE CONTRIBUTION		**NEGATIVE CONTRIBUTION**	
Gene	**Correlation**	**Gene**	**Correlation**
PIK3C2A	0.6401	MAP2K2	-0.8246
PIK3CA	0.6101	MAPK3	-0.6443
PIK3CB	0.5821	AKT2	-0.6432
MAPK1	0.5096	HRAS	-0.5946
EIF4E	0.4815	MKNK2	-0.5884

Table 4. Top contributing genes from the ROMA analysis in the module mitochondria oxidative stress.

D. DISCUSSION

We have presented a systems biology approach using different tools to investigate the relationship between EMT and Senescence. First we have showed the reduced network of the NME1 connections in EMT and Senescence from a larger signalling network. The reduction of mitochondrial ROS, and the subsequent oxidative stress, by NME1, prevents the occurring of the EMT and Senescence. It has been observed that reduced ROS, from targeting the mitochondrial complex I resulted in less EMT in colorectal cancer [24]. Also in breast cancer cells, it has been observed that reduced ROS also result in reduction of EMT markers [25]. Mitochondrial oxidative stress, induced by an excessive amount of ROS, can also result in Senescence [26]. The downstream induction of NF-κB resulted in the induction of both cell programs. This transcription factor has been widely studied for inducing both programs [7,27] but also as a target, for instance NF-κB targeting has resulted in reduced metastases in colorectal cancer in murine models [28]. From this obtained results, we can then conclude that Mitochondrial ROS production and NF-κB activation, that can be activated by different proteins, are key in both processes.

The activation of different programs, through the progression of the disease, has been shown exploring transcriptomics data from patients with colorectal cancer. From this, we saw an activation of Senescence in the most advanced stages. It has been proposed that tumor cells as well as cells from the TME, like CAF, can induce EMT if they are in a senescent state [29,30], specifically by signaling of the SASP molecules. Taken together the results presented, from the systems biology application, we could conclude that these two processes are strongly linked.

From this collaborative project, there were experiments using senolytic agents to treat cells. They have used NME1 depleted cells as well as aerotactic EMT-positive MCF10A cells. Both presented features of cell senescence and they became sensitive to senolysis. The treatments with senolytics its being reviewed lately in combination with other treatments having promising results in cancer but also in the context of other diseases [31–33].

Finally, in this study we can see that applying systems biology tools, like these molecular networks, result in a comprehensive overview of the studied system. We have shown the crosstalk of different proteins that have common effectors between EMT and senescence. In the future, it could be interesting to apply this type of approach to study the inverse process, the Mesenchymal-epithelial transition that has an important role in the establishment of metastasis.

E. SUPPLEMENTARY TABLES

	STAGE I (n = 74)		STAGE IV (n = 61)		
Gene	Mean	S.D.	Mean	S.D.	T-test (1 tailed)
CLN5	612.7559	211.0769	818.8901	511.5455	0.002122
CTSF	252.3945	187.9283	356.8917	253.6308	0.004377
GALNS	426.6519	201.8659	527.625	248.038	0.005927
GLB1	2531.1788	735.9242	2432.854	876.5544	0.2436685
IDS	1834.864	705.3488	2497.207	1873.816	0.005442
LAMP1	8114.744	3698.366	9895.907	4132.787	0.00506
NEU1	2371.771	1243.812	2867.703	1450.249	0.018597
PLA2G15	497.0409	164.748	566.8375	219.7297	0.021339
SLC11A1	116.7673	130.1934	177.2104	235.6541	0.038371
SMPD1	525.4288	311.6779	670.5865	426.9186	0.014491

Supplementary table 1. Mean expression and standard deviation (S.D.) of genes related to lysosome

	STAGE I (n = 74)		STAGE IV (n = 61)		
Gene	Mean	S.D.	Mean	S.D.	T-Test (1 tailed)
CDH1	14069.13	4527.126	14178.64	5116.672	0.448265
CDH2	27.29264	42.85729	65.18369	99.31203	0.003455
KRT18	30410.46	16224.48	33602.91	19817.77	0.157359
MKI67	4426.73	2232.476	3798.055	1996.179	0.043346
MMP14	5299.444	3531.589	7097.619	6190.865	0.023458
SNAI1	127.2326	69.574	199.1302	147.7709	0.000385
SNAI2	124.6267	90.475	194.2126	310.096	0.047388
TWIST1	45.13247	40.24358	97.01585	130.069	0.001877
TWIST2	7.392951	7.690907	8.616046	9.308008	0.206676
VIM	5986.839	4004.615	10247.35	20599.3	0.058302
ZEB1	268.6577	150.6874	331.8404	216.2262	0.028255
ZEB2	308.2077	204.1187	356.9129	359.3997	0.174669

Supplementary table 2. Mean expression and standard deviation (S.D.) of genes related to EMT.

	STAGE I (n = 74)		STAGE IV (n = 61)		
Gene	Mean	S.D.	Mean	S.D.	T-Test (1 tailed)
CCL2	247.4911	221.1236	267.6956	247.6051	0.310735
CDKN1A	3451.997	2366.18	3271.928	2263.136	0.326477
CDKN2A	66.48832	70.59072	143.0202	201.9057	0.003084
IL1A	42.62975	50.89968	57.65319	107.4554	0.159377
IL1B	444.8082	423.7025	1005.227	3727.429	0.123652
IL6	94.24523	222.542	180.3197	623.1458	0.154096
IL8	1997.696	2808.659	3845.274	9744.384	0.078267
MKI67	4426.73	2232.476	3798.055	1996.179	0.043346
RCE1	534.5684	140.4531	540.4699	116.7111	0.395083

RELA	2266.317	456.3211	2400.623	474.9218	0.049258
ZMPSTE24	1855.751	610.9636	1945.991	917.2856	0.256187

Supplementary table 3. Mean expression and standard deviation (S.D.) of genes related to Senescence.

	STAGE I (n = 74)		STAGE IV (n = 61)		
Gene	Mean	S.D.	Mean	S.D.	T-Test (1 tailed)
CAT	2107.433	795.6023	2129.124	758.3643	0.435874
MPO	3.64546	3.457329	6.867098	7.06524	0.000821
NDUFA6	1132.865	459.0848	1011.248	379.7705	0.047215
NDUFA9	2585.076	1095.751	2264.513	992.3116	0.038543
NDUFS1	1739.007	526.9671	1580.16	392.4974	0.023519
NDUFS6	1240.129	427.4574	1401.002	567.2077	0.035121
NDUFV2	1524.589	731.1335	1248.692	580.5896	0.007931
NOX1	4556.063	4016.479	4437.768	4051.885	0.43284
NOX3	0.005731	0.049301	0.014189	0.077704	0.231565
NOX4	31.15232	49.10578	54.47589	61.82838	0.009254
NOX5	0.564588	1.098879	1.339283	3.458093	0.048599
SDHA	3279.756	1244.83	2825.529	1207.833	0.016923
SDHB	1892.078	482.6642	1628.689	515.9744	0.001448
SDHC	2192.25	617.6359	2238.32	780.5585	0.354411
SDHD	2535.223	830.9403	2359.993	857.7037	0.116557
SOD1	4133.309	1869.557	3778.076	1797.311	0.131906
SOD2	5214.081	2555.736	4941.316	3776.505	0.315904
TXNRD1	2629.202	758.8631	2745.945	885.686	0.209046
TXNRD2	462.5513	178.6488	442.9827	155.2277	0.248607
UQCR10	1423.37	567.8754	1248.83	495.3033	0.029361
UQCR11	1869.928	712.4046	1888.611	702.8836	0.439413
UQCRB	3116.407	1244.37	3220.984	1391.338	0.324702
UQCRC1	6138.42	2781.906	5199.232	2327.253	0.017265
UQCRC2	5380.364	1438.112	5103.237	1421.73	0.132129
UQCRFS1	2307.227	758.6101	2041.226	611.4379	0.012861
UQCRH	2604.092	1059.385	2468.571	1390.717	0.266316
UQCRQ	2815.721	1303.159	2897.487	1396.979	0.363901

Supplementary table 4. Mean expression and standard deviation (S.D.) of genes related to Oxidative Stress.

F. REFERENCES

1. Steeg, P. S. Tumor metastasis: mechanistic insights and clinical challenges. *Nat. Med. 2006 128* **12**, 895–904 (2006).

2. Mittal, V. Epithelial Mesenchymal Transition in Tumor Metastasis. *https://doi.org/10.1146/annurev-pathol-020117-043854* **13**, 395–412 (2018).

3. Pezone, A. *et al.* Inflammation and DNA damage: cause, effect or both. *Nat. Rev. Rheumatol.* **19**, 200–211 (2023).

4. Davalli, P., Mitic, T., Caporali, A., Lauriola, A. & D'Arca, D. ROS, Cell Senescence, and Novel Molecular Mechanisms in Aging and Age-Related Diseases. *Oxid. Med. Cell. Longev.* **2016**, (2016).

5. Nikitovic, D., Corsini, E., Kouretas, D., Tsatsakis, A. & Tzanakakis, G. ROS-major mediators of extracellular matrix remodeling during tumor progression. *Food Chem. Toxicol.* **61**, 178–186 (2013).

6. Faget, D. V., Ren, Q. & Stewart, S. A. Unmasking senescence: context-dependent effects of SASP in cancer. *Nat. Rev. Cancer 2019 198* **19**, 439–453 (2019).

7. Wei, W. & Ji, S. Cellular senescence: Molecular mechanisms and pathogenicity. *J. Cell. Physiol.* **233**, 9121–9135 (2018).

8. Laberge, R. M., Awad, P., Campisi, J. & Desprez, P. Y. Epithelial-Mesenchymal Transition Induced by Senescent Fibroblasts. *Cancer Microenviron.* **5**, 39 (2012).

9. Lascu, I. & Gonin, P. The catalytic mechanism of nucleoside diphosphate kinases. *J. Bioenerg. Biomembr.* **32**, 237–246 (2000).

10. Terasaki-Fukuzawa, Y. *et al.* Decreased nm23 expression, but not Ki-67 labeling index, is significantly correlated with lymph node metastasis of breast invasive ductal carcinoma. *Int. J. Mol. Med.* **9**, 25–29 (2002).

11. Guan-Zhen, Y. *et al.* Reduced protein expression of metastasis-related genes (nm23, KISS1, KAI1 and p53) in lymph node and liver metastases of gastric cancer. *Int. J. Exp. Pathol.* **88**, 175–183 (2007).

12. Baba, H. *et al.* Two Isotypes of Murine nm23/Nucleoside Diphosphate Kinase, nm23-M1 and nm23-M2, Are Involved in Metastatic Suppression of a Murine Melanoma Line. *Cancer Res.* **55**, 1977–1981 (1995).

13. Parhar, R. S. *et al.* Effects of cytokine-mediated modulation of nm23 expression on the invasion and metastatic behavior of B16F10 melanoma cells. *Int. J. cancer* **60**, 204–210 (1995).

14. Kuperstein, I. *et al.* Atlas of Cancer Signalling Network: A systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis* (2015) doi:10.1038/oncsis.2015.19.

15. Kondratova, M., Sompairac, N., Barillot, E., Zinovyev, A. & Kuperstein, I. Signalling maps in cancer research: construction and data analysis. *Database (Oxford).* **2018**, 1–15 (2018).

16. Funahashi, A. *et al.* CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. *Proc. IEEE* **96**, 1254–1265 (2008).

17. Martignetti, L., Calzone, L., Bonnet, E., Barillot, E. & Zinovyev, A. ROMA: Representation and quantification of module activity from target expression data. *Front. Genet.* **7**, 1–12 (2016).

18. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2498–2504 (2003) doi:10.1101/gr.1239303.metabolite.

19. Zinovyev, A., Viara, E., Calzone, L. & Barillot, E. BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks. *Bioinformatics* **24**, 876–877 (2008).

20. Novère, N. Le *et al.* The Systems Biology Graphical Notation. *Nature Biotechnology* (2009) doi:10.1038/nbt.1558.

21. Hucka, M. *et al.* The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics* (2003) doi:10.1093/bioinformatics/btg015.

22. Liguori, I. *et al.* Oxidative stress, aging, and diseases. *Clin. Interv. Aging* **13**, 757–772 (2018).

23. Herranz, N. & Gil, J. Mechanisms and functions of cellular senescence. *J. Clin. Invest.* **128**, 1238–1246 (2018).

24. Bastin, J. *et al.* Downregulation of mitochondrial complex I induces ROS production in colorectal cancer subtypes that differently controls migration. *J. Transl. Med.* **21**, 1–16 (2023).

25. Monti, E., Mancini, A., Marras, E. & Gariboldi, M. B. Targeting Mitochondrial ROS Production to Reverse the Epithelial-Mesenchymal Transition in Breast Cancer Cells. *Curr. Issues Mol. Biol.* **44**, 5277–5293 (2022).

26. Mas-Bargues, C. Mitochondria pleiotropism in stem cell senescence: Mechanisms and therapeutic approaches. *Free Radic. Biol. Med.* **208**, 657–671 (2023).

27. Oh, A. *et al.* NF-κB signaling in neoplastic transition from epithelial to mesenchymal phenotype. *Cell Commun. Signal. 2023 211* **21**, 1–17 (2023).

28. de Carvalho, T. G. *et al.* Inhibition of murine colorectal cancer metastasis by targeting M2-TAM through STAT3/NF-kB/AKT signaling using macrophage 1-derived extracellular vesicles loaded with oxaliplatin, retinoic acid, and Libidibia ferrea. *Biomed. Pharmacother.* **168**, (2023).

29. Chambers, C. R., Ritchie, S., Pereira, B. A. & Timpson, P. Overcoming the senescence-associated secretory phenotype (SASP): a complex mechanism of resistance in the treatment of cancer. *Mol. Oncol.* **15**, 3242–3255 (2021).

30. Li, H. *et al.* Senescent Fibroblasts Generate a CAF Phenotype through the Stat3 Pathway. *Genes (Basel).* **13**, (2022).

31. Kirkland, J. L. & Tchkonia, T. Senolytic drugs: from discovery to translation. *J. Intern. Med.* **288**, 518–536 (2020).

32. Wang, K. *et al.* KDM4C-mediated senescence defense is a targetable vulnerability in gastric cancer harboring TP53 mutations. *Clin. Epigenetics* **15**, (2023).

33. Zhu, Y. *et al.* Identification of a novel senolytic agent, navitoclax, targeting the Bcl-2 family of anti-apoptotic factors. *Aging Cell* **15**, 428–435 (2016).

CHAPTER IV: KNOWLEDGE FORMALIZATION TO TACKLE AN EMERGENT PANDEMIC

A. INTRODUCTION

The coronavirus disease 2019 (COVID-19) was an emergent and rapidly spreading disease that had begun in late 2019. This disease which is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) had 446,260 confirmed cases and caused 36,847 deaths worldwide at the end of March of 2020 [1].

Due to the rapid increase of the epidemic across the world, that led to a major confinement worldwide, it resulted to be a topic of interest for researchers. Hence, as a part of disease maps (https://disease-maps.org) consortium[2,3], a large-scale community dedicated to the biocuration, construction and exploitation of molecular networks represented as maps in different diseases, we have gathered together to construct the COVID-19 Disease map (https://disease-maps.org/covid-19/). This map, is composed of many diagrams and has been used to analyze omics data, and has been already published[4]. However, for clarity of this chapter, I will focus it on my main contribution to this effort, i.e. the construction and refinement of the endoplasmic reticulum (ER) stress map.

The endoplasmic reticulum, is a vast organelle with many functions for cell homeostasis, such as Ca^{2+} storage, synthesis and folding of proteins as well as carbohydrate and lipid metabolism. Many conditions, such as oxidative stress, altered Ca^{2+} homeostasis, fails on protein folding can cause the accumulation of unfolded or misfolded proteins in the ER, leading to the stress of this organelle. The ER has many pathways to resolve this stress, however, when it fails to restore its function can trigger cell apoptosis [5,6].

The expression of some human coronavirus (HCoV) proteins during infection, specially the S glycoprotein, might induce the activation of the ER stress in the host's cells [7]. The unfolded protein response (UPR) pathways are key to assure the ER homeostasis, these pathways are activated by the protein kinase RNA-activated (PKR)-like ER protein kinase (PERK), inositol-requiring enzyme 1 (IRE1), and activating transcription factor 6 (ATF6) [8]. During SARS-Cov-1 infection, it has been proved the activation of PERK[9], IRE1[10] and, ATF6 pathways[11].

IRE1 is the most evolutionary conserved UPR protein. IRE1 by a luminal domain and a cytoplasmic effector domain. It has kinase and endoribonuclease activity. In absence of stress it is maintained as a monomer and bound to the chaperone BiP towards the lumen. When there is an accumulation of unfolded proteins, IRE1 is activated by detachment of BiP and subsequent oligomerization and autophosphorylation. Then it exerts ribonuclease activity on *XBP1* mRNA that results in the translation of the transcription factor XBP1 in its active form that induces the transcription of response proteins. IRE1 can also degrade mRNAs and miRNAs to decrease the quantity of proteins incoming the ER, this process is called IRE1-dependent decay (RIDD). Additionally, when phosphorylated, it can bind to TRAF2 to promote c-Jun N-terminal kinase (JNK) resulting in activation of cell death programs [12–14].

PERK, Similar to IRE1, also is a monomeric protein that as a cytoplasmic domain and a luminal domain that is also bound to BiP. Upon unfolded protein accumulation, BiP dissociates from PERK, leading to dimerization and autophosphorylation, making PERK activated. PERK then phosphorylates the α subunit of the eukaryotic initiation factor 2 (eIF2α), reducing thus eIF2-GTP resulting in a decreased translation as well as activation of stress proteins such as ATF4 and CHOP[8,13,15].

ATF6 is a transmembrane protein in the form of monomers and dimers with intra- and inter-molecular disulfide bonds. These are bound to the chaperon BiP. Upon an insult, BiP dissociates and the disulfides are reduced by protein disulfide isomerases. This leads to monomeric forms of ATF6 that traffic to the Golgi apparatus, where they are proteolytically processed, releasing the cytosolic active part of ATF6, a bZIP transcription factor, that localizes to the nucleus and induces the expression of genes related to ER quality control [8,13,16].

This map depicts the activation of the main UPR actors (ATF6, IRE1 and PERK) upon unfolded protein accumulation, and their role of this response to Ca2+ release into the cytoplasm as well as the activation of molecules that can lead to apoptosis and cell death.

B. MATERIALS AND METHODS

Obtaining related key players from a larger network

To construct the ER Stress map, in the context of SARS-CoV-2 infection, we have started from an already existing map to gather the key elements relevant for the infection. This larger map was the regulated cell death (RCD) map [17], which is available online: https://acsn.curie.fr/navicell/maps/rcd/master/index.html. In this map, we find the module ER stress under the stress response layer. We have centered the focus in the three main players of the UPR: ATF6, IRE1 and PERK. Then by using Cytoscape v2.8 [18] and the BiNoM plugin [19], we could extract a subnetwork with the entities related to these key players.

Literature curation and map annotation

The previously obtained network was manipulated using the CellDesigner[20] software. Entities and reactions were annotated using the MIRIAM registry[21]. The curation rules were established to have a homogeneous process in the community. A summary of the biocuration guidelines is available in Appendix 1. After having the reduced network, we refined it by reviewing the literature regarding similar diseases, such as SARS-CoV infection, which caused an outbreak in 2002[22] or HCoV in general.

C. RESULTS

The ER stress map consisted of 121 species from which 64 corresponded to proteins, 9 genes, 9 RNAs, 2 simple molecules, 19 complexes, 15 phenotypes, 3 ions, 83 reactions and is based on 23 scientific papers (between reviews and original papers). Inside the map, there were 7 compartments, comprising the endoplasmic reticulum, the nucleus, the Golgi apparatus, the mitochondria, the mitochondria inner membrane, the cytoplasm and the autophagosome Figure 1.

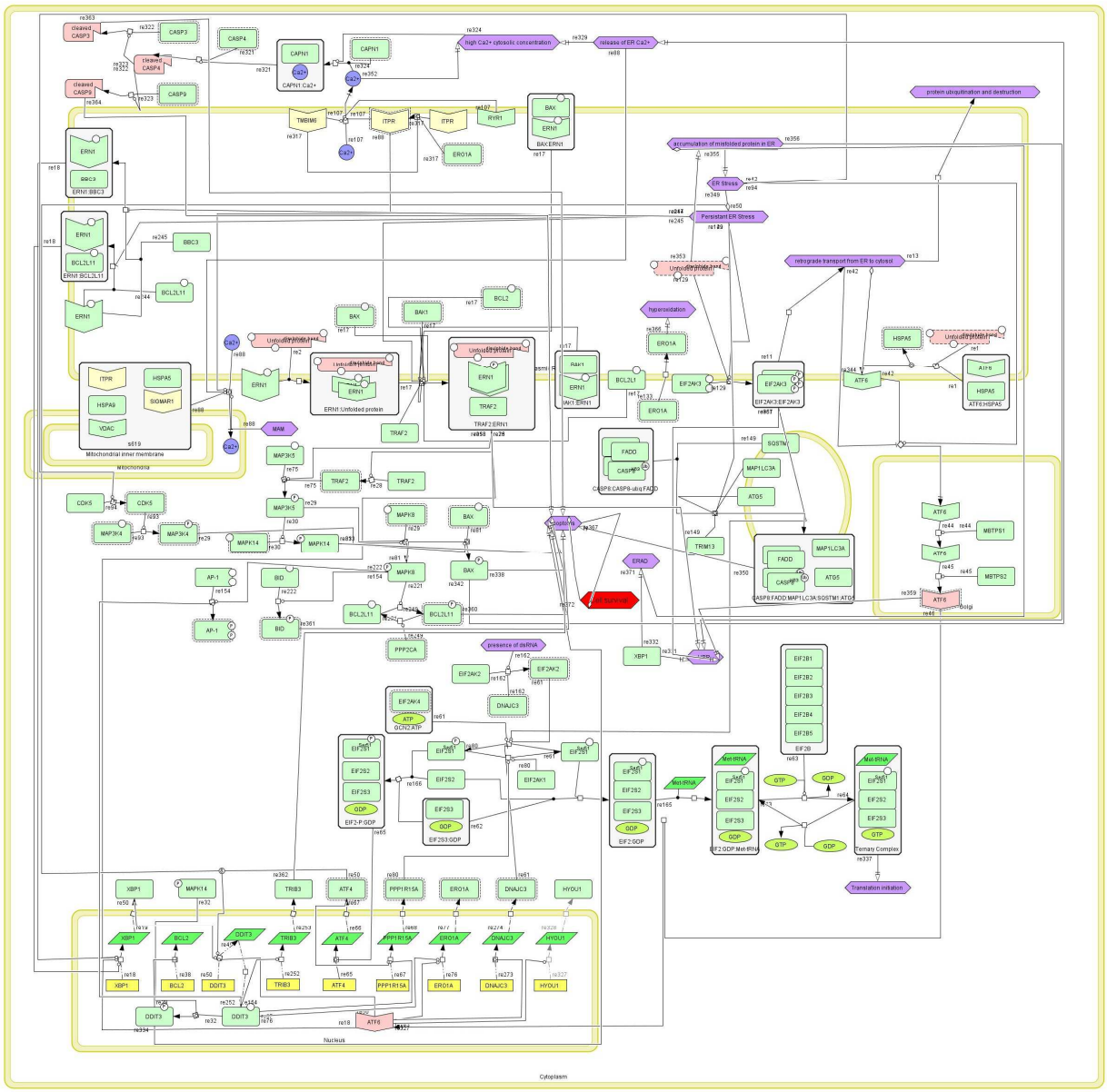


Figure 1. Overall view of the ER-stress map. This map was constructed based on the main players of the UPR: ATF6, IRE1 and PERK.

The ER stress map depicts in a simplified manner the main actors of the UPR response and their downstream effects. This map was then enriched with the mitochondria-associated ER membrane (MAM) and calcium homeostasis (collaboration with Barbara Brauner, Helmholtz Zentrum, Germany). As a result, this extension have connected the ER Stress and the Mitochondria in a more integrative manner. The top level view is in Supplementary figure 1.

This map was integrated to 20 other manually curated maps to assemble the COVID-19 map (available at: covid19map.elixir-luxembourg.org). The maps integrated 4 groups which were the virus replication cycle, with the attachment and entry, the transcription, translation and replication as well as assembly and release. The viral subversion of the host defense, where ER stress was integrated with

apoptosis and autophagy. The integrative stress response that had the renin-angiotensin system, coagulopathy and the Innate Immune Response: with PAMP signaling, Induction of interferons and the cytokine storm and altered host metabolism, the top level view of the COVID-19 map is in Figure 2.

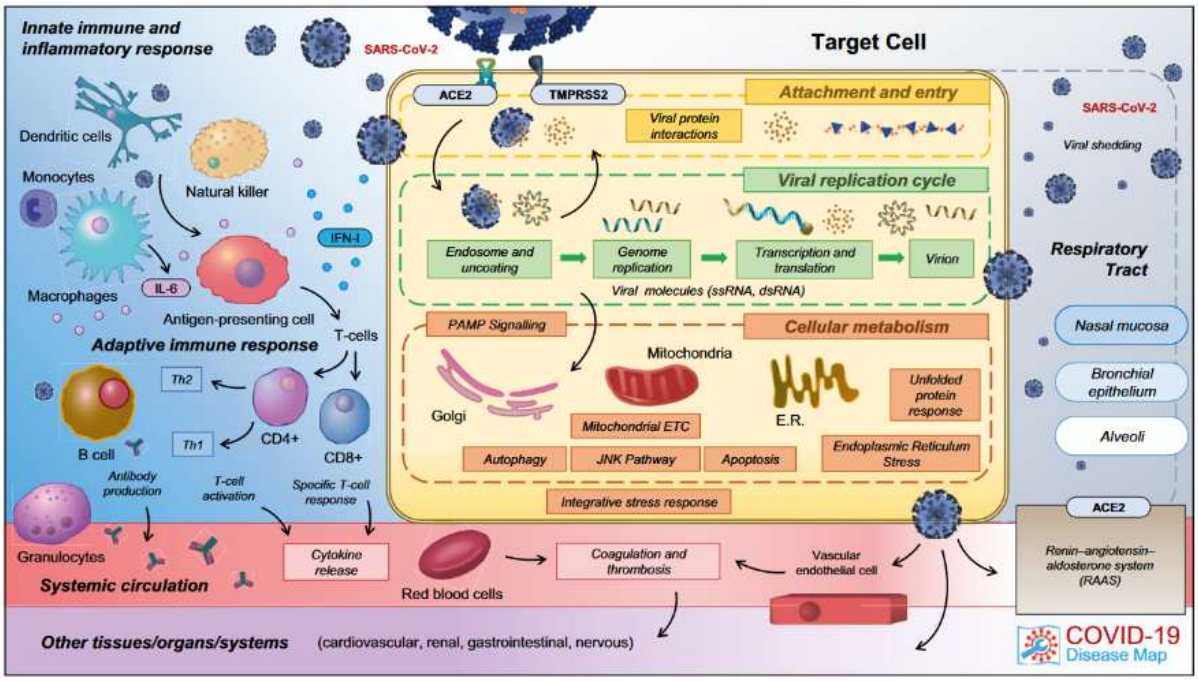


Figure 2. Top level view and contents of the COVID-19 map. This integrative map has covered different aspects of the disease.

The construction of these maps has helped to analyze some omics data that have been reported as case studies[4]. For instance, there was a case study using RNA-Seq data from nasopharyngeal swabs[23] from cases of COVID-19 and controls, were they focused on the apoptosis pathway (from the apoptosis map). They observed an overall downregulation of CASP3 and CASP7 sub pathways and inhibition of the circuit that ended in CASP3 probably due to the downregulation of AKT1 and BAD with the downstream inhibition of BAX. Although the BAX downstream genes appeared up-regulated, the signal arriving at them was reduced due to the effect of the previous nodes. Although CASP8 was up-regulated, the aggregated effect of the individual node activities resulted in the inhibition of CASP7. In fact, inflammatory response via CASP8 has been reported to occur with SARS-CoV-2 infection[24], and the caspase-induced apoptosis together with the ripoptosome/caspase-8 have been considered as a pro-inflammatory checkpoint[25], which in turn, could trigger the activation of other pathways that result in the progression of the disease (Supplementary figure 2).

D. DISCUSSION

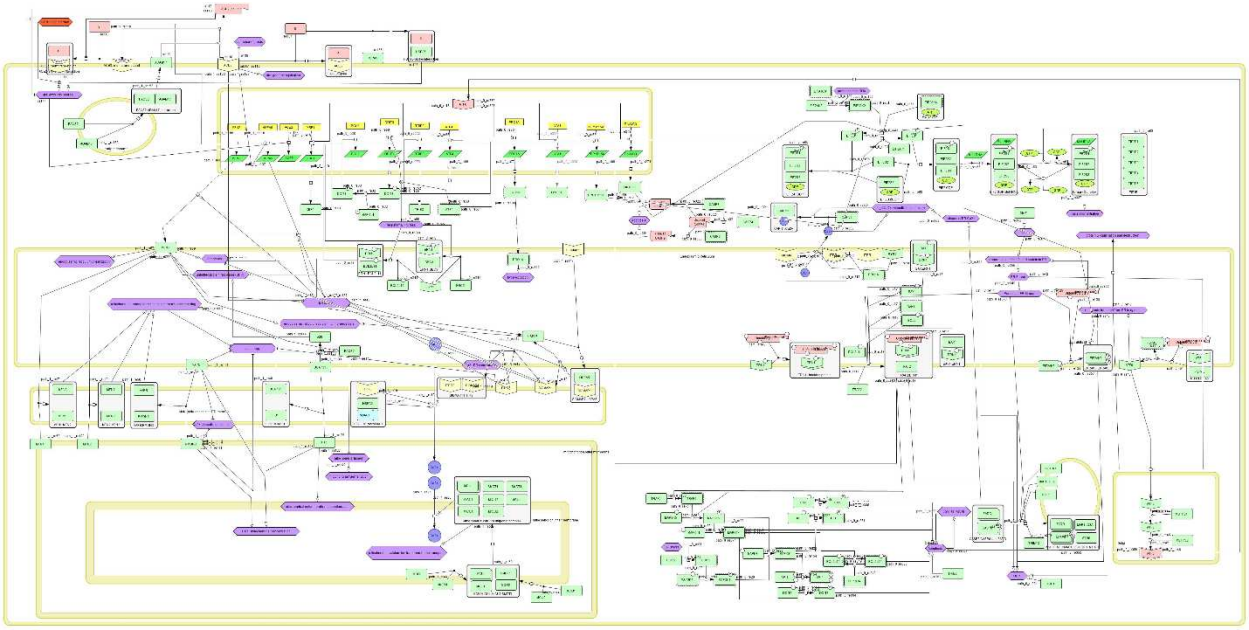
In this chapter I have presented a work of biocuration applied to a viral respiratory disease that was part of a large, international, community effort. This effort gathered scientists from different fields, like biology, bioinformatics, computer science, medical sciences among others. This allowed to develop an integrative, comprehensive and powerful tool that was the COVID-19 map.

My contribution working in the ER stress map involved the revision and curation of scientific literature that could be relevant for SARS-CoV-2 infection and the progression to COVID-19. The resulting ER stress map, ensembled with the other curated maps, resulted in a comprehensive tool depicting the key

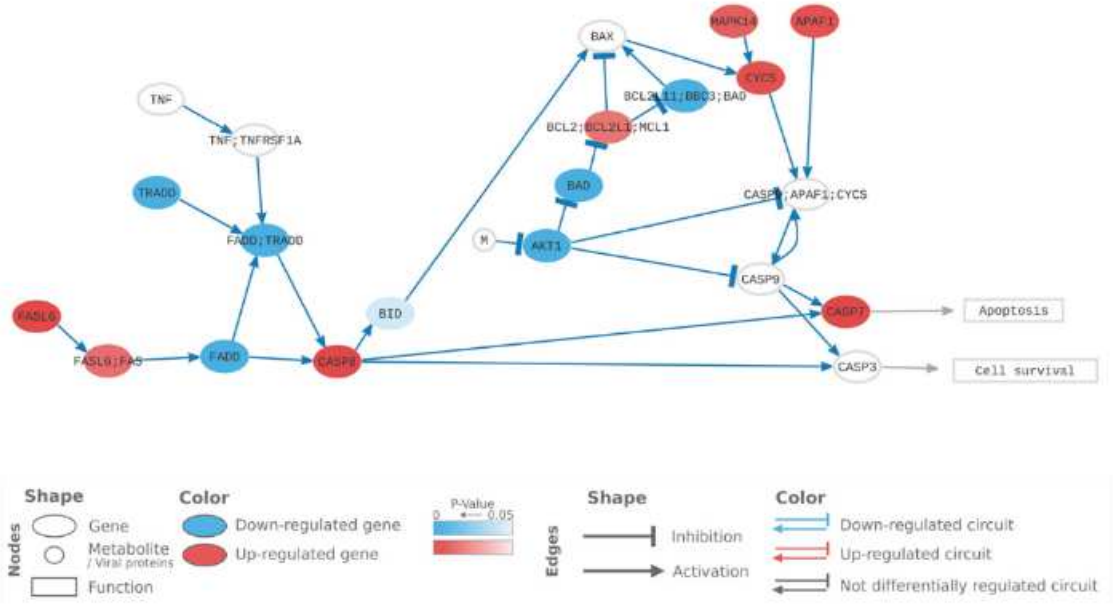
mechanisms and molecular relationships taking place during the disease. Nevertheless, the information contained in this type of tools, can also serve as a starting point to study other diseases, in the case of my map, the core was the three components of UPR, that can take place in any cell undergoing ER stress. Using transcriptomics data analysis, it was presented the different activation levels of the apoptosis pathway. The results could give explanations of interactions on pathway elements, that can give rise to new hypotheses.

Overall, I have expose an application of biocuration to construct a tool that could be applied to a disease. This tools have the properties that can be human and computer readable, so that they can be applied for other *in silico* approaches. Currently, the community effort is ongoing, however, the focus on COVID-19 has decreased. Noteworthy, all the teams participating in the disease maps community are specialized in a disease or related diseases, in the case of E. Barillot's team, it is cancer.

E. SUPPLEMENTARY FIGURES



Supplementary figure 1. Top level view of the ER stress map after the integration of the mitochondrial-associated ER membrane.



Supplementary figure 2. Representation of the different activation levels for the apoptosis pathways in nasopharyngeal swabs.

F. REFERENCES

1. WHO Coronavirus (COVID-19) Dashboard | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data. https://data.who.int/dashboards/covid19/cases.

2. Ostaszewski, M. *et al.* Community-driven roadmap for integrated disease maps. *Brief. Bioinform.* **20**, 659–670 (2019).

3. Mazein, A. *et al.* Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. *NPJ Syst. Biol. Appl.* **4**, (2018).

4. Ostaszewski, M. *et al.* COVID19 Disease Map, a computational knowledge repository of virus–host interaction mechanisms. *Mol. Syst. Biol.* **17**, e10387 (2021).

5. Oakes, S. A. & Papa, F. R. The role of endoplasmic reticulum stress in human pathology. *Annu. Rev. Pathol.* **10**, 173–94 (2015).

6. Senft, D. & Ronai, Z. A. UPR, autophagy, and mitochondria crosstalk underlies the ER stress response. *Trends Biochem. Sci.* **40**, 141–148 (2015).

7. Fukushi, M. *et al.* Monitoring of S Protein Maturation in the Endoplasmic Reticulum by Calnexin Is Important for the Infectivity of Severe Acute Respiratory Syndrome Coronavirus. *J. Virol.* **86**, 11745–11753 (2012).

8. Fung, T. S. & Liu, D. X. Human Coronavirus: Host-Pathogen Interaction. *Annu. Rev. Microbiol.* **73**, 529–557 (2019).

9. Krähling, V., Stein, D. A., Spiegel, M., Weber, F. & Mühlberger, E. Severe Acute Respiratory Syndrome Coronavirus Triggers Apoptosis via Protein Kinase R but Is Resistant to Its Antiviral Activity. *J. Virol.* **83**, 2298–2309 (2009).

10. DeDiego, M. L. *et al.* Severe acute respiratory syndrome coronavirus envelope protein regulates cell stress response and apoptosis. *PLoS Pathog.* **7**, (2011).

11. Sung, S. C., Chao, C. Y., Jeng, K. S., Yang, J. Y. & Lai, M. M. C. The 8ab protein of SARS-CoV is a luminal ER membrane-associated protein and induces the activation of ATF6. *Virology* **387**, 402–413 (2009).

12. Cabral-Miranda, F. *et al.* Unfolded protein response IRE1/XBP1 signaling is required for healthy mammalian brain aging. *EMBO J.* **41**, e111952 (2022).

13. Wiseman, R. L., Mesgarzadeh, J. S. & Hendershot, L. M. Reshaping endoplasmic reticulum quality control through the unfolded protein response. *Mol. Cell* **82**, 1477–1491 (2022).

14. Park, S. M., Kang, T. Il & So, J. S. Roles of XBP1s in Transcriptional Regulation of Target Genes. *Biomedicines* **9**, (2021).

15. Donnelly, N., Gorman, A. M., Gupta, S. & Samali, A. The eIF2α kinases: their structures and functions. *Cell. Mol. Life Sci. 2013 7019* **70**, 3493–3511 (2013).

16. Chen, X., Shen, J. & Prywes, R. The luminal domain of ATF6 senses endoplasmic reticulum (ER) stress and causes translocation of ATF6 from the ER to the Golgi. *J. Biol. Chem.* **277**, 13045–13052 (2002).

17. Ravel, J. M. *et al.* Comprehensive Map of the Regulated Cell Death Signaling Network: A Powerful Analytical Tool for Studying Diseases. *Cancers (Basel).* **12**, (2020).

18. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2498–2504 (2003) doi:10.1101/gr.1239303.metabolite.

19. Zinovyev, A., Viara, E., Calzone, L. & Barillot, E. BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks. *Bioinformatics* **24**, 876–877 (2008).

20. Funahashi, A. *et al.* CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. *Proc. IEEE* **96**, 1254–1265 (2008).

21. Juty, N., Le Novere, N. & Laibe, C. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res.* **40**, (2012).

22. Hui, D. S., Azhar, E. I., Memish, Z. A. & Zumla, A. Human Coronavirus Infections—Severe Acute Respiratory Syndrome (SARS), Middle East Respiratory Syndrome (MERS), and SARS-CoV-2. *Encycl. Respir. Med.* **4**, 146 (2022).

23. Lieberman, N. A. P. *et al.* In vivo antiviral host transcriptional response to SARS-CoV-2 by viral load, sex, and age. *PLoS Biol.* **18**, (2020).

24. Li, S. *et al.* SARS-CoV-2 triggers inflammatory responses and cell death through caspase-8 activation. *Signal Transduct. Target. Ther.* **5**, 235 (2020).

25. Chauhan, D. *et al.* BAX/BAK-Induced Apoptosis Results in Caspase-8-Dependent IL-1β Maturation in Macrophages. *Cell Rep.* **25**, 2354-2368.e5 (2018).

CONCLUDING REMARKS AND PERSPECTIVES

In this manuscript, I have described the fundamentals of systems biology and omics sciences, as well as some examples where these approaches were used to investigate different questions regarding biological processes in human diseases.

I presented an integrative study that was performed in collaboration with different partners at the Insitut Curie and the Centre de Recherche Saint-Antoine. This study represented a series of challenges for all the partners participating on it. By using the transcriptomics data from diverse tissue samples, I could identify some likely mechanisms to be indicators of presence of the cancer associated adipocytes. The found mechanisms, thermogenesis and matrix metalloproteinases, shall be further investigated and validated. In this project it is envisaged to perform lipidomics/metabolomics in a vast array of samples, in order to add another molecular layer and have a better approximation of the underlying biology. Additionally, spatial transcriptomics analyses are planned for a set of samples, in order to profile the adipocytes that are in close contact with tumor cells. Nevertheless, in another branch of this large collaborative project, cell co-cultures using primary adipocytes from patients and cancer cell lines are being carried out. There are some indicators that suggest that the cancer cells co-cultured with adipocytes shown increased expression of invasion markers. In this co-cultured cells, it is planned to perform RNA-Seq, so that we can compare the results that were obtained in the bulk data. In conclusion, for this part of the thesis, the used approach gave us an insight in the mechanisms that could characterize the cancer associated adipocytes. A paper regarding these findings is being prepared.

Then, I described a project that was done in collaboration with M. Boissan, where we investigated the relationship between EMT and cellular senescence. My contribution to this project, was to depict the players of NME1, metastatic suppressor, in both processes. I have found different players of interest that can relate both processes. Then I analyzed some colorectal cancer data, corresponding to different stages, to see the differences in terms of gene-sets from an already existing signaling network diagram of EMT. The results indicated that in later stages of the disease, key genes of senescence and EMT were overexpressed in the stage IV when compared to stage I. On the other hand, the biological experiments that have been performed suggest a strong relationship between both processes, indicating a senolytic like phenotype in cells with positive markers of EMT. My contributions to this project, together with the experimental findings are already in the form of a manuscript that soon will be submitted.

Lastly, as a complementary chapter, I described my contribution to a project that occurred at the beginning of the PhD, regarding the knowledge formalization and description of the endoplasmic reticulum stress in cells infected with SARS-CoV2, this network diagram has helped efforts for many groups applying other systems biology approaches, such as modelling. The integrated network has been mentioned in a series of publications.

To conclude, systems biology offers many opportunities to address different biological questions, the exploitation of transcriptomics, can provide valuable information regarding the characteristics of certain cells, the relationships between two processes, among others that result in connections to wire the molecular networks that occur during the human diseases.

RÉSUMÉ

Les systèmes biologiques sont des structures complexes avec des interactions complexes entre leurs composants. Grâce à la combinaison de différents domaines scientifiques, il est désormais possible d'étudier ces systèmes et de répondre à différentes questions qui ont des applications différentes. Dans cette thèse, j'ai exploré des outils et des approches utilisés en biologie des systèmes afin de trouver des acteurs moléculaires ainsi que des mécanismes importants dans les réseaux moléculaires des systèmes biologiques. J'ai intégré des techniques d'analyse de données transcriptomiques. J'ai également utilisé des approches de formalisation des connaissances afin de construire ou d'étendre des réseaux moléculaires descriptifs existants dans différentes maladies.

J'ai principalement étudié le rôle du tissu adipeux dans le cancer du sein. Le tissu adipeux constitue une partie fondamentale et importante de l'anatomie du sein. Il a été suggéré que ce tissu adipeux, principalement composé d'adipocytes blancs, interagit avec les cellules cancéreuses au front invasif de la tumeur, favorisant ainsi la progression tumorale. Ces cellules ont été appelées "Adipocytes associés au cancer (CAA, en anglais)". Il a été émis l'hypothèse selon laquelle l'interaction entre CAA et cellules tumorales serait amplifiée en cas d'obésité. Ainsi, une cohorte de patientes atteintes d'un carcinome canalaire mammaire et classées comme obèses ou normo-pondérales a été constituée. J'ai analysé des échantillons de tissu adipeux de ces patients, proches (proximaux) ou éloignés (distal) de la tumeur, au niveau du transcriptome. Les deux types de tissus présentaient des motifs d'expression génique similaires. Cependant, avec l'analyse d'enrichissement, les échantillons proximaux présentaient des voies de signalisation des œstrogènes enrichies et des voies liées à l'épithélium par rapport aux échantillons distaux. Par rapport aux échantillons de tumeurs, les échantillons proximaux montraient principalement des voies menant à la fonction du tissu adipeux, telles que l'adipogenèse, le métabolisme des acides gras, la signalisation de PPAR entre autres. J'ai appliqué l'analyse ROMA pour déterminer l'activation des voies d'intérêt à partir des résultats d'enrichissement, et nous avons constaté que la thermogenèse et les métalloprotéinases matricielles étaient plus actives dans les tissus adipeux proximaux. Les gènes *MMP7*, *MMP16, MMP3, SMARCC1, CREB3L4, MAPK13, RPS6KA6, SMARCA4, ZNF516, ACTG1, SLC25A9* sont apparus comme contributeurs majeurs.

Les réseaux moléculaires peuvent être représentés sous forme de diagrammes. Les informations contenues dans ces réseaux peuvent servir à exploiter l'analyse des données transcriptomiques. Auparavant, l'Atlas du réseau de signalisation du cancer avait été constitué. Cette ressource est composée de processus biologiques pour le développement et la progression du cancer sous la forme de cartes. J'ai utilisé l'une des cartes, la sénescence cellulaire et la transition épithélio-mésenchymateuse (EMT, en anglais), pour explorer le rôle du prototype gène suppresseur de métastase, *NME1* (appelé auparavant *NM23-H1*) dans ces processus. J'ai enrichi la carte avec les fonctions de la protéine NME1 et utilisé les informations pour compiler les acteurs impliqués dans la sénescence cellulaire et l'EMT. Certains acteurs intéressants liés aux deux processus ont été identifiés, comme NF-κB, montrant que la sénescence a une relation avec l'EMT. Ensuite, j'ai utilisé des données transcriptomiques provenant de patients atteints d'un cancer colorectal pour observer l'activité des différents modules du réseau afin d'observer la progression à travers les différents stades de la maladie.

Finalement, en raison de l'épidémie de COVID-19, j'ai participé à un effort où nous avons construit une carte de l'interaction hôte-virus, la carte COVID-19. Ma contribution s'est concentrée sur la construction du réseau représentant le stress du réticulum endoplasmique.

MOTS CLÉS

Biologie des systèmes, transcriptomique, maladies, réseaux moléculaires

ABSTRACT

Biological systems are complex structures with multiple interactions between their components. Thanks to the combination of fields such as mathematics, computational science, biology, physiology etc. it is now possible to study these systems and answer different questions that have different applications, like in human health. In this thesis I have explored some tools and approaches used in systems biology in order to find molecular players as well as mechanisms that are important in the molecular networks for the biological systems. For this thesis, I have integrated data analysis techniques to transcriptomics data in different diseases. Also, I have used knowledge formalization approaches in order to construct or extend existing descriptive molecular networks in different diseases.

I have studied the role of adipose tissue in breast cancer. The adipose tissue constitutes a fundamental and large part of the breast anatomy. Mammary adipocytes have been hypothesized to interact with cancer cells at the invasive front of the tumor, supporting the progression of the disease. These adipocytes have been termed “Cancer Associated Adipocytes (CAA)”. The interaction of these CAA and the progression of the disease have been suggested to be worse in obese patients. Therefore, to have an insight on the mechanism , a cohort of patients that had ductal breast carcinoma and that are considered as obese or normal-weight was created. I have analyzed adipose tissue samples of these patients, that were either close (proximal) or far (distal) from the tumor, at the transcriptome level. Both tissue types showed similar gene expression patterns. However, with the enrichment analysis, proximal samples had enriched estrogen signaling pathways, and pathways related to epithelium when compared to distal samples. When compared to tumor samples, proximal showed mostly pathways to their adipose tissue function, as adipogenesis, fatty acid metabolism PPAR signaling among others. We applied ROMA analysis to determine activation of pathways of interest from the enrichment results, and we found thermogenesis and matrix metalloproteinases to be more active in the proximal adipose tissues. The genes *MMP7, MMP16, MMP3, SMARCC1, CREB3L4, MAPK13, RPS6KA6, SMARCA4, ZNF516, ACTG1, SLC25A9* appeared as major contributors.

Molecular networks can be depicted as diagrams in order to facilitate their exploration and visualization. The information contained in these networks may serve to exploit the analysis of transcriptomics data using techniques such as gene-set enrichment analysis. Previously, the Atlas of Cancer Signaling Network was assembled. This resource is composed of known biological processes that are relevant for cancer development and progression in the form of maps depicting molecular interactions. I have used one of the maps, cellular senescence and Epithelial to Mesenchymal Transition (EMT), to explore the role of prototypic metastasis suppressor gene *NME1* (previously called *NM23-H1*) in these processes. I had enriched the map with functions of the protein and also used the information to compile the players that are involved in cellular senescence and EMT. Some interesting players that are related were identified to both processes, like NF-κB, showing that senescence has a relationship with EMT. Then, I used transcriptomics data from colorectal cancer patients to observe the activity of the different modules in the network to observe the progression through the different stages of the disease.

Lastly, due to the COVID-19 epidemic, I have participated in a multi-research groups’ effort where we constructed a map of the host-virus interaction, the COVID-19 map. My contribution was focused on building the network representing the endoplasmic reticulum stress.

KEYWORDS

Systems biology, transcriptomics, diseases, molecular networks