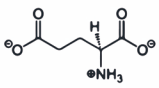


doctoral thesis

Multi-omics Integration for Biomarker Discovery

MARC CLOS GARCIA

2017-2019

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

MULTI-OMICS INTEGRATION FOR BIOMARKER DISCOVERY

MARC CLOS GARCIA

DOCTORAL THESIS

Thesis directors:
Dr. Juan Manuel Falcón Pérez
Dr. Luis Bujanda Fernández de Pierola

Thesis tutor:
Dr. Patricia Aspichueta Celaa

Universidad del País Vasco – Euskal Herriko Unibertsitatea

Center for Cooperative Research in Biosciences (CIC bioGUNE)

Biodonostia Health Research Institute

2017-2019

One by one

Only the good die young

They're only flyin' too close to the sun

Queen (1997) – Eulogy to Freddie Mercury

Acknowledgments

I find your lack of faith disturbing.

A New Hope, 1977.

A JuanMa, por confiar en mi. Por liarme en demasiados proyectos y darme la oportunidad de aprender en tantos campos distintos. Aunque alguna vez me hayas dicho que no hay manera de saturarme, confieso que lo has conseguido. Gracias también por eso, por ayudarme a crecer y entender como funciona el mundo de la ciencia. Por valorar mi trabajo, la autonomía que me has dado, la confianza de dejarme tomar las riendas de vez en cuando y tener mi opinión en cuenta. Gracias por construir un equipo en el que cualquiera querría trabajar.

A Luis, por colaborar en esta oportunidad que me habéis dado. Por dejarme a mi aire pero estando allí cuando era necesario. Por seguir invitándome a las comidas del laboratorio aunque solo haya estado ahí un par de veces al año.

A Naroa, Marina y (por extensión natural ya) Jorge por aparecer en mi vida en el momento y el lugar ideal. Naro, no dejes de bailar por las calles, de ser como quieres ser y de pasar de todos los consejos de quien intente cambiarte. Marina, Jorge, seguid intentando convertir las comidas de los sábados en tradición, pero buscaos alguien más cumplidor que yo. Sois los tres lo más parecido a una familia que he tenido en Bilbo y eso no se paga ni con todo el dinero del mundo, que tampoco os lo iba a pagar siendo catalán. Aun así, tenéis las puertas abiertas a donde sea que vaya siempre que queráis. Os quiero mucho mucho como la trucha al trucho, xafaxarcos.

A Maria, la chica joven aunque-cada-vez-menos-joven del grupo, la eterna estresada. Podría decirte cosas buenas, pero no me pega y lo sabes. Así que lo dejamos en un consejo: respira. Dos veces. Tres veces. Cada vez que entres en el agujero negro que es el despacho del jefe y salgas con esas listas eternas. Take it easy, dale un par de vueltas y disfrutalo, que puedes con todo. Gracias por ser mi reserva de ego y autoestima este año y poco que hemos compartido.

A Rubén, por ser un desastre completo en (casi) todo, aunque no por las horas que he perdido esperandote al quedar contigo. Estoy convencido que si consigues ordenarte la cabeza vas a llegar lejos, así que dale duro.

A Diana&Sebastiaan, la *platform-family*. Gracias por adoptarme en ese zulo como uno más. A ti, Diana, por la complicidad, el entendimiento y las rajadas compartidas. To you, Seb, for the shared opinions about how bullshit everything in science is, including omics. Disfrutad la luz y las vistas ahora que habéis huido del zulo! A Jon, por su paciencia.

A Félix, por enseñarme que los protocolos están bien pero siempre se pueden mejorar. Por ser mi filtro previo a JuanMa. Por tus consejos, lecciones de vida, sinceridad y optimismo particular para hacer todo esto más fácil, divertido y llevadero.

To Charles, for the shared beers, beers and more beers. I hope you'll finally find a way to make exosomes-enriched beer and that I have a chance to taste it! Did I say beer already? Beer.

A Espe, por el soporte moral, la paciencia y la filosofía que me has tratado de inculcar estos años, con más bien poco éxito…

A Jone, por acompañarme en suficientes de mis locuras y muchísimos más cafés y peleas por un cacho de poyata.

A Endika, por ayudarme a desarrollar, probar y demostrar la teoría “trabajando un miércoles salvas la semana”. Eso va a mínimo un journal con IF 25.

A Leti y JL (pronunciado como *jei-el,* ni que fueras Abrams) por ser mi peor pesadilla y mis mejores recursos en la locura compartida con mi jefe de meterme a microbiomear. Sois la peor influencia que he tenido en estos 3 años. Seguid siendolo para otros ahora, que siempre es divertido no dejar piedra sobre piedra.

A Sofi, Maria, Teresa, Marta, Gotxi y el resto del Malu’s Lab. Gracias por hacer que el poco tiempo que he pasado en la poiata haya parecido aún más corto. Fer&David&Jorge, por hacer que las comidas sean el momento de desconexión necesario y por ayudarme a entender la vida propia del bioGUNE. Marujos.

A Koldo, en Biodonostia, por darme acceso a tus recursos y prestarme ayuda siempre.

A la gente de Genómica, Ana&Moni&Co por la paciencia para explicarme el mismo protocolo mil veces y aguantarme las prisas con galletas.

Felix, Mikel, Ibon y el resto de la plataforma de proteómica por su ayuda en este proyecto.

A tu, Anna, per aconseguir que hagi acabat aquesta bogeria de fer una tesi doctoral sense parar (del tot) boig. Saps que els nostres Skypes (cada cop menys habituals) m’han salvat la vida més d’un, dos i tres cops. Que tots els viatges, museus i sopars de postureo que hem imaginat i mai hem fet serveixin d’excusa per veure’ns molt més temps.

Reb, la rubia petirroja hija de Satán, modelo casi-profesional, mi amiga-la-guapa y rompedora de esquemas mentales por excelencia, por ser mi mejor hobby cuando estoy por casa.

Sil&Laia, el duo més boig que conec. D’el-ja-sabeu-perquè a simplement els litres de cervesa que ja no queden al McKiernan’s, Konigs i River.

A l’Eloi, per convertir-se en la meva tradició gironina preferida. No, no tens ni unes braves ni una *Golden Ale* pagades, comunista.

Al recurs barcelo-bisbalenc, de visita anual obligada durant la **segona** millor festa major del món: les Sans, les Soriano i en Roger.

A en Pere, per accedir a retrocedir 5 (¿?) anys en el temps i recuperar la tradició de dissenyar les portades de tesi de mitja facultat de Ciències de la UdG, ara exportada.

I finalment, els més importants. Als meus pares, que m’ho van donar tot. Possiblement necessitaria una altra vida (o tres!) per tenir temps de tornar-vos tot el que heu fet per mi durant tants anys. No sóc capaç de trobar paraules que puguin resumir la gratitud que us dec, permeteu-me deixar-ho en un desig: que la vida us sigui justa i us doni tot el que us mereixeu. Per la meva banda, així intentaré que sigui.

Contents

Publications and congresses

Publications

1) **Clos-Garcia, M.**, Loiziaga-Iriarte, A., Zúñiga-García *et al. Metabolic alterations in urine extracellular vesicles are associated to prostate cancer pathogenesis and progression*. Journal of Extracellular Vesicles, 7:1, 1470442, 2018. **Shared first authorship**.
2) Cubiella, J., **Clos-Garcia, M.** *et al. Targeted UPLC-MS Metabolic Analysis of Human Faeces Reveals Novel Low-Invasive Candidate Markers for Colorectal Cancer.* Cancers, 10:9, 2018. **Shared first authorship.**
3) **Clos-Garcia, M.** *et al. Gut microbiome and serum metabolome analyses identify molecular biomarkers and altered glutamate metabolism in fibromyalgia.* EBioMedicine, 46, 499-511, 2019. **First authorship**.

In preparation

Clos-Garcia, M. *et al. Gut microbiota and fecal targeted metabolomics combination identifies relationships between fecal metabolome and intestinal microbiota and provides with biomarkers for colorectal cancer.* **First authorship.**

Congresses and workshops

1) 1st Euskadi Worskhop on Exosomes (23/03/2017), Derio, Spain.

   Clos-Garcia, M.. *Management of omics data for discovery of predictive biomarkers* – Oral talk.

2) ISEV 2017 Annual Meeting (18-21/05/2017), Toronto, Canada.

   Zuñiga, P., **Clos-Garcia, M.**, Loizaga-Iriarte, A., Sanchez-Mosquera, P., Cortazar, AR, Gónzalez, E., Alonso, C., Pérez-Cormenzana, M., Ugalde Olano, A., Lacasa, I., Castro, A., Royo, P., Unda, M., Carracedo, A., Falcón-Pérez, JM*. *Metabolomics Analysis of Urinary Exosomes reveals novel candidate biomarkers of prostate cancer* – Poster. *Presenter.

3) Bioinformatics Tools to Study Exosomes' Effects (13-15/11/2017), Derio, Spain.

   Clos-Garcia, M. *Combining Cytoscape and STRING database to retrieve protein-protein interactions* – Oral talk and organizer.

4) ISEV 2018 Annual Meeting (02-06/05/2018), Barcelona, Spain.

   Clos-Garcia, M., Loizaga-Iriarte, A., Zuñiga Garcia, P., Sánchez Mosquera, P., Gónzalez, E., Cortazar, AR., Torrano, V., Alonso, C., Pérez Cormenzana, M., Ugalde Olano, A., Castro, A., Royo, F., Unda, M., Carracedo, A., Falcón-Pérez, JM. *Bioinformatics analysis of metabolites present in urinary exosomes identify metabolic pathways altered in prostate cancer* – Poster.

5) 7th International Human Microbiome Consortium Meeting (IHMC) (26-28/06/2018), Killarney, Ireland.

   Clos-Garcia, M., Abecia, L., Lavin, JL., Andrés, N., Fernandez Garcia de Eulate, G., van Liempd, S., Cabrera, D., Valero, A., Errazquin, N., Gomez Vallejo, MC, Govillard, L., Royo, F., Gónzalez, E., Anguita, J., Aransay, AM., Callejo Orcasitas, AM., Maiz, O., Lopez de Munain, A., Falcón-Pérez, JM. *The microbiome composition in fibromyalgia favors the presence of systemic higher levels of the neurotransmitter glutamate* – Poster

6) MIKROBIOGUNE 2018 (11/12/2018), Bilbao, Spain.

   Clos-Garcia, M., Andrés, N., Fernandez Garcia de Eulate, G., Abecia, L., Lavin, JL., van Liempd, S., Cabrera, D., Valero, A., Errazquin, N., Gomez Vallejo, MC, Govillard, L., Royo, F., Gónzalez, E., Anguita, J., Aransay, AM., Callejo Orcasitas, AM., Maiz, O., Lopez de Munain, A., Falcón-Pérez, JM. *Combination of microbiome and metabolomics allows the identification of fibromyalgia patients* – Poster and Oral presentation.

7) Barcelona Debates on Human Microbiome (20-21/06/2019), Barcelona, Spain.
   Clos-Garcia, M., Andrés, N., Fernandez Garcia de Eulate, G., Abecia, L., Lavin, JL., van Liempd, S., Cabrera, D., Valero, A., Errazquin, N., Gomez Vallejo, MC, Govillard, L., Royo, F., Gónzalez, E., Anguita, J., Aransay, AM., Callejo Orcasitas, AM., Maiz, O., Lopez de Munain, A., Falcón-Pérez, JM. *Combination of microbiome and metabolomics allows the identification of fibromyalgia patients* – Poster.

List of Figures

List of Tables

Abbreviations

5-HT	5-hydroxytryptamine
AA	Amino Acids
AC	Acylcarnitines
ACN	Acetonitrile
ACR	American College of Rheumatology
AD	Adenocarcinoma
ADCY5	Adenylate Cyclase 5
ADMA	Asymmetric Dimethylarginine
ALDEx2	ANOVA-Like Differential Expression
ALOX15	Arachidonate 15-Lipoxygenase
ANOVA	Analysis Of Variance
AUC	Area Under the Curve
BA	Bile Acids
BMI	Body Mass Index
BPH	Benign Prostatic Hyperplasia
BST2	Bone Marrow Stromal Cell Antigen 2
C	Control
cAMP	Cyclic adenosine monophosphate
CAMP	Cationic Antimicrobial Peptide
Carb	Carboxylic acids
CCA	Canonical Correlation Analysis
CCM	Derivative carboxylic acids
CD9/CD10/CD63/CD26	Cluster of differentiation 9/10/63/26
CEA	Carcinoembryonic antigen
Cer	Ceramides
CERS4	Ceramide Synthase 4
ChoE	Cholesteryl Esters
CIA	Co-Inertia Analysis
CLR	Centered-Log Ratio
CMH	Monohexosylceramides
CNS	Central Nervous System
COMT	Catechol-O-methyltransferase
CORBATA	CORe microbiome Analysis Tools
COX IV	Cytochrome C oxidase subunit 4 isoform 1
CRC	Colorectal Cancer
CSF	Cerebrospinal Fluid
CYP1A2	Cytochrome P450 1A2
DAPC	Diacylglycerophosphocholine
DAPE	Diacylglycerophosphoethanolamine
DAPI	Diacylglycerophosphoinositol
DAVID	Database for Annotation, Visualization and Integrated Discovery
DG	Diacylglycerols
DHEAS	Dehydroepiandrosterone sulphate
DIABLO	Data Integration Analysis for Biomarker Discovery using Latent cOmponents
DNA	Deoxyribonucleic Acid
E2	Estradiol
EAE	Experimental Autoimmune Encephalomyelitis
EC	Enzyme Code
EDA	Exploratory Data Analysis
ENA	European Nucleotide Archive
ENS	Enteric Nervous System

EVs	Extracellular Vesicles
FAA	Fatty amides
Faith's PD	Faith's Phylogenetic Distance
FFA	Non-sterified fatty acids
FOB	Fecal Occult Blood
GABA	Gamma-Aminobutyric Acid
gadC	Glutamate/Gamma-Aminobutyrate antiporter
GALT	Gut-associated lymphoid tissues
GATM	Glycine Amidinotransferase
GC-MS	Gas Chromatography Mass Spectrometry
GEO	Gene Expression Omnibus
GF	Germ-Free
glnA	Glutamine synthetase A
glsA/glsB	Glutaminase A/B
GO	Gene Onthology
GPI	Glycosylphosphatidylinositol
GRP78	78 kDa glucose-related protein
GWAS	Genome Wide Association Studies
HAllA	Hierarchical All-against-All significance test
HCV	Hepatitis C Virus
HETE	5-hydroxyeicosatetraenoic acid
HMDB	Human Metabolome Database
HMP	The Human Microbiome Project
HPA	Hypothalamic-Pituitary-Adrenal axis
HPLC	High-Pressure Liquid Chromatography
HPV	Human Papilloma Virus
IBD	Intestinal Bowel Disease
ICAM2	Intracellular Adhesion Molecule 2
ICC	Interclass Correlation Coefficient
IL-6/IL-8/IL-9	Interleukin 6/8/9
IMPaLA	Integrated Molecular Pathway Level Analysis
IPA	Ingenuity Pathway Analysis
IQR	Inter Quartile Range
KEGG	Kyoto Encyclopedia of Genes and Genomes
LASSO	Least Absolute Shrinkage and Selection Operator
LCAT	Lecithin-Cholesterol Acyltransferase
LC-MS	Liquid Chromatography Mass Spectrometry
LEfSe	Linear discriminant analysis Effect Size
LPCAT1/LPCAT2	Lysophosphatidylcholine Acyltransferase 1/2
LPPs	Lipid Phosphate Phosphatases
LPS	Lipopolysaccharides
LR	Logistic Regression
MALDI	Matrix-Assisted Laser Desorption Ionization
MAPC	1-monoacylglycerophosphocholine
MAPE	Monoacylglycerophosphoethanolamine
MAPI	Monoacylglycerophosphoinositol
MEMAPC	1-ether, 2-acylglycerophosphocholine
MEMAPE	1-ether, 2-acylglycerophosphoethanolamine
MEPC	1-monoetherglycerophosphocholine
MEPE	1-monoetherglycerophosphoethanolamine
MG	Monoglyceride
mGluR	Metabotropic Glutamate Receptor
MIG	Monokine induced by gamma interferon (CXCL9)
miRNA	Micro RNA
MS	Mass Spectrometry
NMR	Nuclear Magnetic Resonance
nNOS/eNOS/iNOS	Nitric Oxide Synthase neuronal/endothelial/inducible

NO	Nitric Oxide
NPV	Negative Predictive Value
OMIM	Online Mendelian Inheritance in Man
OPLS	Orthogonal Partial Least Squares
OTU	Operational Taxonomic Unit
PAF-16	Platelet Activating Factor
PBS	Phosphate Buffered Saline
PC	Principal Component
PC	Phosphatidylcholine
PCA	Principal Component Analysis
PCa	Prostate Cancer
PCoA	Principal Component Analysis (distance matrices)
PCR	Polymerase Chain Reaction
PCSK9	Proprotein Convertase Subtilisin/Kexin Type 9
PDE4C	Phosphodiesterase 4C
PE	Phosphatidylethanolamine
PERMANOVA	Multivariate ANOVA with permutations
PGE2	Prostaglandin E Receptor 2
PICRUSt/PICRUSt2	Phylogenetic Investigation of Communities by Reconstruction of Unobserved States
PIGK/PIGZ	Phosphatidylinositol Glycan Anchor Biosynthesis Class K/Z
PLD	Phospholipase D
PLS	Partial Least Squares
PLS-DA	Partial Least Squares Discriminant Analysis
PPI	Proton Pump Inhibitors
PPV	Positive Predictive Value
PSA	Polysaccharides
PSA	Prostate-Specific Antigen
PUFA	Polyunsaturated fatty acids
QC	Quality Control
QIIME/QIIME2	Quantitative Insights Into Microbial Ecology
qPCR	Quantitative Polymerase Chain Reaction
rDNA	Ribosomal Deoxyribonucleic Acid
RDP	Ribosomal Database Project
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
rRNA	Ribosomal Ribonucleic Acid
SCFA	Short Chain Fatty Acids
SD	Standard Deviation
SEM	Standard Error of the Mean
SFB	Segmented Filamentous Bacteria
SIAMCAT	Statistical Inference of Associations between Microbial Communities And host phenoTypes
SM	Sphingomyelin
SMPD1/SMPD3	Sphingomyelin Phosphodiesterase 1/3
SOP	Standardized Operating Protocols
sPLS-DA	Sparse Partial Least Squares Discriminant Analysis
SS	Severity Score
SSRI	Selective Serotonin Reuptake Inhibitors
STS	Steroid Sulfatase
SULT2B1/ SULT2A1	Sulfotransferase Family 2B/A Member 1
TAG	Triacylglycerol
TCA	Tricarboxylic Acid Cycle
TCGA	The Cancer Genome Atlas
TG	Triglyceride
T_H	T helper cell
THP	Tamm-Horsfall glycoprotein (uromodulin)

TNF	Tumor Necrosis Factor
ToF	Time of Flight
T_{reg}	Regulatory T cell
Tukey's HSD	Tukey's Honest Significance Difference
UHPLC-MS	Ultra-High Pressure Liquid Chromatography Mass Spectrometry
UPLC	Ultra-Pressure Liquid Chromatography
WGCNA	Weighted Gene Co-expression Network Analysis
WGS	Whole Genome Sequence
WHO	World Health Organization
WPI	Widespread Pain Index

Abstract

Biomarkers constitute all those molecules and substances that are able to measure and describe any change in a biological system. Because of this property, they have been adopted as disease diagnostic and surveillance tools. The number of studies involving new biomarker discoveries has tremendously increased since the development and globalization of high-throughput technologies, such as the omics ones, either performed alone or in combination. The success of these kinds of studies, though, is far from efficient, as much of these newly described biomarker candidates fail in the validation process, being this success ratio calculated to be about 0.1% of proposed biomarkers to end up on clinical standardized practice.

Diseases are considered to be multifactorial, with several combinations of environmental and genomics factors defining a specific phenotype. Omics technologies allow the measurement of thousands of features simultaneously, providing this way with a tool to characterize diseases with great detail. Single omics studies, though, are limited to correlations between one data type, while multi-omics studies may explain interactions between different data types, leading to better-characterized phenotype. Multi-omics integrative studies are still being implemented and, therefore, methods for integration are being developed and discussed. Current methods include dimension reduction analysis and correlation-based methods. Among the widely used omics technologies are genomics, metagenomics, transcriptomics, proteomics and metabolomics analyses. This latter technology is based on the study of small molecules, usually of less than 2,000 Da in size, which are part of all the metabolic pathways. They vary rapidly in front of an environmental change, making them very good candidates for biomarkers. On the other hand, metagenomics studies the structure, composition and metabolic capabilities of the genetic components of a specific community. By metagenomics, we usually refer to microbiome studies, usually restricted to bacterial population, either performed by WGS or 16S gene sequencing.

In this thesis, we combined distinct omics data and biochemical analysis to try to identify new biomarkers in three diseases with elevated socio-economical impact: prostate cancer, colorectal cancer and fibromyalgia. On one hand, we combined metabolomics and transcriptomics for PCa early diagnosis biomarkers identification. In the case of CRC we have combined metabolomics, microbiome and transcriptomics for detection of early biomarkers. Later on, metabolomics, microbiome, cytokines and miRNA profiling for fibromyalgia characterization and diagnostic biomarkers. Finally, focusing on the metabolomics, we developed a tool to facilitate the functional characterization of metabolites identified as potential biomarkers by automatic data retrieval from the most used metabolomics databases.

As result, urine EVs metabolomics analysis revealed higher EVs and different metabolite content in PCa patients when compared to BPH. The combination of metabolomics with publicly available transcriptomics datasets allowed us to provide a functional explanation for the alterations identified. Metabolomics revealed increased levels of sterols and acylcarnitines in PCa patients, while PC family metabolites were decreased. Functional profiling of these alterations mapped metabolite alterations to steroid hormones and cellular energy pathways. Transcriptomics analysis of enzymes related to our altered metabolites showed high concordance with our results, thus providing with a biological context of the described changes and with robustness for our biomarkers, as we used different cohorts for both analyses.

Also, the bioinformatics analysis, in this case of fecal microbiome dataset revealed a reduction of diversity in fibromyalgia patients when compared to healthy individuals. In this case, the functional interpretation by using specific software identify a reduction of bacteria related to neurotransmitter metabolism, which was afterwards confirmed by qPCR analyses. These alterations were concordant with serum metabolomics analysis, which found increased levels of neurotransmitters in patients' blood. Several correlations were found between metabolomics and microbiome datasets, thus confirming the hypothesis of gut microbiota role on the host's metabolome and health. While microbiome and microbiome ability to discriminate between fibromyalgia patients and healthy individuals was not very good, the combination of both the four distinct data types improved the discrimination ability.

For the colorectal cancer study, the fecal metabolomics provided with 18 differentially abundant metabolites between healthy individuals and patients, mostly from ceramides, cholesteryl esters and sphingomyelins metabolite families. A combination of seven of those metabolites was used as a predictive model for the disease, validated later with another cohort. Microbiome study of the CRC patients presented an increased abundance of *Fusobacterium, Bulleidia, Parvimonas, Staphylococcus* and *Gemella.* Lachnospiraceae family bacteria were found to be decreased for CRC individuals. *Adlercreutzia* was found to be increased only in AD patients. The integration of both datasets revealed correlation clusters between altered bacteria and altered metabolite families in CRC patients. The Procrustes analysis revealed similarities between microbiome and metabolomics data, confirming the role of the intestinal microbiome population modulating the fecal metabolomics. Finally, gut microbiota was found to be a better choice to differentiate CRC patients from other sample groups, while metabolomics seemed to discriminate better between healthy individuals and AD patients. The regression model combining both data types performed slightly better than models generated with only one omics data type.

Our work explores the use of multi-omics integration studies to improve the biomarker discovery process, providing options to ensure the robustness of biomarker candidates. We tried to tackle each common challenge multi-omics studies have been discussed to present, trying to solve them in various ways. Finally, we proposed a list of considerations to be followed to ensure good performance of these kinds of projects.

Resumen

Los biomarcadores son todas esas moléculas y sustancias que pueden usarse para medir y describir cualquier alteración en un sistema biológico. Son muy sensibles a cualquier cambio que se produzca en el sistema biológico en el que se miden, presentado las correspondientes modificaciones de forma casi inmediata. De ahí deriva su utilidad como herramientas para el diagnóstico y control de cualquier sistema. Actualmente, los biomarcadores se clasifican en tres categorías en función de su utilidad: biomarcadores predictivos, biomarcadores pronósticos y biomarcadores diagnósticos. Con el desarrollo y globalización de los estudios *high-throughput*, como es el caso de las tecnologías ómicas, el número de estudios relacionados con la búsqueda de nuevos biomarcadores se ha incrementado notablemente. Este incremento se ha dado principalmente en estudios de una sola ómica, aunque recientemente se han empezado a popularizar también los estudios de combinación de distintas ómicas. Desafortunadamente, el éxito de este tipo de estudios dista mucho de ser óptimo, puesto que la mayor parte de los nuevos candidatos a biomarcadores no superan el proceso de validación. Existen distintas razones por las que estos biomarcadores pueden no validarse. La no recolección de datos clínicos y ambientales de calidad, por ejemplo, pueden influir en este proceso, al no poderse discriminar si las diferencias en estos biomarcadores se deben a factores externos o a la propia patología de estudio. La falta de estandarización de los protocolos experimentales durante el procesado de las muestras como de los métodos de análisis bioinformáticos y estadísticos tiene un impacto directo en la reproducibilidad y validación de cualquier biomarcador.

Hoy en día sabemos que la mayor parte de las enfermedades son multifactoriales, cuyo fenotipo depende de una combinación de factores genómicos y ambientales. Las tecnologías ómicas permiten medir miles de variables simultáneamente, siendo así muy útiles para describir cualquier enfermedad. Aun así, los estudios basados en una sola ómica están limitados a las correlaciones que se puedan establecer en el tipo de datos correspondiente mientras que los estudios que combinan distintas ómicas permiten identificar y explicar interacciones entre distintos tipos de datos (moléculas) permitiendo así una mejor caracterización de un fenotipo concreto. Los estudios combinatorios de ómicas están aún en desarrollo, por lo que los métodos analíticos correspondientes están también en desarrollo y en constante discusión. Actualmente, los métodos usados para estos estudios incluyen los análisis de reducción de dimensionalidad de datos y métodos basados en correlaciones.

Entre las tecnologías ómicas más usadas encontramos la genómica, transcriptómica, proteómica, metabolómica y metagenómica. La metabolómica se basa en el estudio de las moléculas pequeñas (<2.000 Da) que forman parte de los miles de rutas metabólicas del organismo. Varían muy rápidamente delante de cualquier cambio ambiental, siendo así una muy buena fuente para la búsqueda de biomarcadores. La metagenómica es el estudio de la estructura, composición y funcionalidad metabólica de los genes de una

comunidad específica. Habitualmente nos referimos a los estudios de microbioma, generalmente restringidos a la población bacteriana, ya sean realizados mediante secuenciación completa del genoma o solo del gen ribosomal 16S. Cada una de estas dos opciones tienes sus ventajas y sus inconvenientes. Así como la secuenciación del gen 16S proporciona datos puramente taxonómicos, los estudios de WGS nos permiten identificar genes individuales, permitiendo así la caracterización funcional de la población microbiana. Aun así, el manejo de datos resultantes de la secuenciación del gen 16S es más sencilla y la resolución taxonómica buena. Además, resulta una tecnología más barata que el WGS y actualmente ya existen herramientas bioinformáticas que nos permiten reconstruir el genoma completa y hacer una aproximación de estudio funcional a partir de estos datos de 16S.

Los estudios de combinación de distintas ómicas son bastante recientes, por lo que metodológicamente aún existen muchas lagunas al respecto. De hecho, en este tipo de estudios se unen los problemas propios de cada ómica que se incluya en el análisis y los problemas derivados de la integración entre ellas. Uno de los inconvenientes más relevantes es el formato de datos en los que cada ómica trabaja, que puede dificultar la integración. Para integrar datos provenientes de distintas ómicas, el primer paso es normalizar adecuadamente cada tipo de datos. Luego podremos aplicar la técnica que consideremos más apropiada, ya sea de reducción de dimensionalidad o basada en correlaciones entre variables. Las técnicas de reducción de dimensionalidad incluyen los análisis de componentes principales, ya sean no supervisadas o supervisadas, los estudios de Procrustes y los análisis de co-inercia. Este tipo de aproximación nos permite comparar la similitud entre dos o más ómicas distintas, aunque dificulta la identificación de biomarcadores derivados de estos estudios. Para facilitar la identificación y caracterización funcional de estos biomarcadores podemos usar las técnicas basadas en correlaciones, que van a identificar que variables de cada ómica están más relacionadas con las variables de otra ómica.

En esta tesis hemos combinado diferentes ómicas para intentar identificar nuevos biomarcadores en tres enfermedades distintas: cáncer de próstata, cáncer colorrectal y fibromialgia. Tanto hemos combinado nuestros propios datos con los datos de acceso libre que se pueden descargar de las bases de datos correspondientes. Para la identificación de nuevos biomarcadores tempranos para cáncer de próstata (PCa) combinamos datos de metabolómica y transcriptómica; para cáncer colorrectal (CRC) hemos combinado metabolómica, microbioma y transcriptómica; y finalmente combinamos metabolómica, microbioma, peptidómica y un cribado de citoquinas y microRNAs para la identificación de biomarcadores para la caracterización y diagnóstico de la fibromialgia. Centrándonos en la metabolómica, hemos desarrollado una herramienta para facilitar la caracterización funcional de los metabolitos identificados como potenciales biomarcadores a partir de la automatización del proceso de selección de información de las bases de datos de metabolómica más comunes.

El PCa es uno de los tipos de cáncer con mayor mortalidad en hombres de los países desarrollados, pero a día de hoy no existen herramientas diagnósticas útiles en los

primeros estadíos de la enfermedad. En este contexto, la metabolómica de las vesículas extracelulares (EV) presentes en la orina podría proveer con biomarcadores capaces de discriminar pacientes con hiperplasia benigna (BPH) de los pacientes con PCa. El uso de EVs de orina puede aliviar las dificultades derivadas del uso de muestras de orina, siendo así una mejor fuente de potenciales biomarcadores. El estudio de la metabolómica de las EVs de orina identificó EVs de mayor tamaño y con un contenido de metabolitos distinto en los pacientes de PCa en comparación de los pacientes de BPH. En este contexto, vimos que el tamaño medio de las poblaciones de EVs era similar a medida que la enfermedad progresaba, pero que en estadíos avanzados aparecía también una subpoblación de EVs de mayor tamaño. La combinación de nuestra metabolómica con los datos de acceso abierto de transcriptómica nos permitió darle una explicación funcional a las alteraciones observadas. La metabolómica identificó niveles elevados de esteroles y acilcarnitinas en los pacientes de PCa, mientras que los metabolitos de la familia de las PC estaban reducidos en estos pacientes. La caracterización funcional de las alteraciones observadas mapeó esos metabolitos a rutas relacionadas con las hormonas esteroides y de obtención de energía. Los EVs de los estadíos avanzados de PCa mostraron además niveles elevados de ceramidas y fosfolípidos, comparado con el primer estadío de PCa. Finalmente, entre los pacientes de los primeros estadíos que presentaban invasión perineural identificamos mayores niveles de esteroles, mientras que también se observó una reducción de AMP cíclico, comparados con los pacientes sin invasión perineural. Los análisis de transcriptómica para las enzimas relacionadas con los metabolitos alterados concordaban con nuestros resultados a la vez que daban un contexto biológico a los cambios descritos. Además, al usar distintas cohortes, la transcriptómica dio robustez a nuestro biomarcadores.

La fibromialgia es una enfermedad de etiopatología desconocida, caracterizada principalmente por un dolor crónico inespecífico. No existe un criterio diagnostico objetivo, siendo el actual basado en una exploración clínica y un formulario que incluye una serie de potenciales síntomas asociados. Aun así, se considera que está estrechamente relacionada con el sistema nervioso central (CNS). En este contexto, nuestra hipótesis de trabajo se basa en el conocido como eje intestino-cerebro, por el cual el microbioma intestinal se comunica con el huésped y regula aspectos físicos y psicológicos del mismo. El análisis del microbioma fecal de los pacientes de fibromialgia mostró una reducción de la diversidad bacteriana en los pacientes y, específicamente, en las bacterias relacionadas con el metabolismo de neurotransmisores, como pudimos validar por qPCR. Estas alteraciones concordaban con los estudios de metabolómica en suero, que identificaron niveles más elevados de neurotransmisores en la sangre de los pacientes de fibromialgia. También identificamos varias correlaciones entre el microbioma fecal y el metaboloma del suero, confirmando así el papel del microbioma en el metaboloma y la salud del huésped. El estudio de las citoquinas y miRNAs identificó algunas alteraciones en diez citoquinas en sangre, mientras que solo un miRNA se encontró estar alterado. Mientras que la capacidad para diferenciar entre pacientes de fibromialgia e individuos sanos de los datos de metabolómica y microbioma por si solos era limitado, la combinación de las 4 ómicas mejoró notablemente esta capacidad

predictiva. Además, la combinación de la metabolómica, miRNAs y citoquinas identificó alteraciones en rutas metabólicas relacionadas con ciertos neurotransmisores como el glutamato y el óxido nítrico (NO) que concordaban además con las alteraciones propuestas a partir del estudio del microbioma.

La relación entre el microbioma y el CRC se ha discutido que sea muy estrecha, llegando a proponer un rol para el microbioma por el cual este es responsable de la acumulación de mutaciones en oncogenes, influyendo así en la aparición y desarrollo del CRC, en la hipótesis llamada *driver-passenger bacteria.* Además, hay que tener en cuenta el papel del microbioma modulando el metaboloma de las heces, puesto que para este proyecto usamos muestras de heces para identificar biomarcadores. Combinando metabolómica y microbioma en muestras de heces, intentamos identificar biomarcadores tempranos útiles para diferenciar entre tres grupos de muestras: controles sanos, pacientes con adenoma y pacientes con CRC. Las muestras analizadas las distribuimos en 3 lotes de muestras, usando los dos primeros para generar un modelo diagnostico con los datos de metabolómica y el tercero para validar el modelo. Para el estudio de microbioma no establecimos ninguna organización particular de las muestras y se trabajó con todas ellas. El estudio de metabolómica identificó 18 metabolitos con diferente abundancia entre los individuos sanos y los pacientes, mayormente ceramidas, ésteres de colesterol y esfingomielinas. Una combinación de 7 metabolitos se usó para generar un modelo predictivo para la enfermedad, que se pudo validar en otra cohorte. Para el estudio del microbioma, combinamos distintas herramientas de análisis, para intentar identificar el máximo número posible de potenciales biomarcadores. El estudio del microbioma para todas las muestras no pudo identificar ninguna diferencia entre los individuos sanos y los diagnosticados con adenoma. Sin embargo, los pacientes con CRC presentaban una mayor abundancia de *Fusobacterium, Bulleidia, Parvimonas, Staphylococcus* y *Gemella.* La familia de bacterias Lachnospiraceae se encontró reducida en los pacientes de CRC. Finalmente, vimos que el género *Adlercreutzia* presentaba mayor abundancia solo para los pacientes de AD. A nivel funcional, se identificaron también alteraciones en la capacidad metabólica de las comunidades microbianas de los pacientes de CRC, con mayor representación de rutas relacionadas con la methanogenesis o la degradación de nitratos, por ejemplo. La integración de los dos tipos de datos mostró grupos de correlaciones entre las bacterias y los metabolitos alterados en los pacientes de CRC. Los análisis de Procrustes identificaron similitudes entre los datos de microbioma y los de metabolómica, confirmando así el papel de la población del microbioma intestinal modulando el metaboloma fecal. Finalmente, vimos que el microbioma intestinal era la mejor opción para diferenciar los pacientes de CRC del resto de grupos incluidos en el estudio, mientras que los datos de metabolómica funcionaron mejor para discriminar entre los individuos sanos y los pacientes de AD. El modelo de regresión generado con la combinación de los dos tipos de datos tuvo una mayor capacidad predictiva que los modelos generados con solo un tipo de ómicas.

Más allá de los tres casos prácticos expuestos y los respectivos biomarcadores identificados, durante el desarrollo de esta tesis también hemos ahondado en aspectos

experimentales y estructurales del microbioma y hemos desarrollado una serie de *scripts* bioinformáticos destinados a facilitar la tarea de descripción, contextualización y caracterización de los metabolitos como biomarcadores. En este sentido, analizamos la especificidad y las diferencias a nivel composicional del microbioma usando distintos sets de *primers*, cada uno dirigido a regiones hipervariables distintas del gen ribosomal 16S. Aun obteniendo un perfil composicional distinto, con especificidad por distintos tipos de bacteria para cada set de *primers*, los análisis comparativos revelaron que ambas aproximaciones proporcionaban un perfil de distribución y diferenciación entre muestras prácticamente idéntico. También analizamos las potenciales diferencias entre distintas poblaciones microbianas dependiendo de la región corporal muestreada. Como se ha descrito previamente, identificamos una composición completamente distinta en función de la región corporal analizada, posiblemente relacionada con los distintos requerimientos funcionales y metabólicos de cada región.

Respecto a la herramienta bioinformática desarrollada, durante el desarrollo de esta tesis nos encontramos con la necesidad de caracterizar funcionalmente los distintos metabolitos que hemos identificado como biomarcadores de alguna de las enfermedades estudiadas. Así, decidimos desarrollar una herramienta que nos permitiese extraer información y caracterizar así los diferentes metabolitos identificados de forma relativamente sencilla y pudiendo trabajar en bloque. Además, decidimos incluir y combinar las dos bases de datos más usadas en el campo, KEGG y HMDB. También quisimos aportar un extra de funcionalidad facilitando la tarea posterior de discusión de resultados y validación experimental de las alteraciones identificadas añadiendo funcionalidades como la búsqueda automática en OMIM, Pubmed, etc. de las enzimas relacionadas con los metabolitos correspondientes. Para facilitar la distribución de la herramienta, decidimos adaptarla a dos de los lenguajes de programación más comunes, R y Python.

Nuestro trabajo explora el uso de los estudios de integración de las tecnologías ómicas para mejorar el proceso de identificación de nuevos biomarcadores, aportando opciones para mejorar la robustez de los candidatos a biomarcador identificados. Mediante la combinación de los tres casos prácticos, hemos intentado afrontar los problemas más habituales de este tipo de estudios combinatorios de distintas formas. Finalmente, hemos propuesto una lista de consideraciones que pueden servir para mejorar el rendimiento de este tipo de proyectos.

Introduction

The Force will be with you. Always.

A New Hope, 1977.

1.1.-Biomarkers

The definition of biomarker was proposed by the World Health Organization (WHO) Task Group on Biomarkers and Risk Assessment: Concepts and Principles in a report from 1989, which summarized a series of meetings dedicated to discussing this term [1]. They adopted the term biomarker to describe the ***"(…) chemicals, metabolites of chemicals, enzymes and other biochemical substances (…) to document the interaction of chemicals with biological systems"***. Although the WHO group were the ones to define what a biomarker was, they were not the first to use the term, as they mention. Technically, the first proposition of a biomarker definition appeared in a US National Academy of Sciences report [2] and in another paper [3]. Notably, in both cases, the term was referred to and proposed in terms of toxicology studies. Combining the three reports, biomarkers were defined within a broad sense, in order to include any measurement that could reflect an interaction between biological systems and potential hazards, which were defined to include either chemical, physical or biological. The use of biomarkers in research has grown exponentially with the development of new high throughput technologies that allow the measurement of a multitude of variables simultaneously such the omics technologies [4]. Behind this growth, there is the need to have direct measurement of potential disease causes and affections, free from recall bias and capable to provide with a biological framework too [5]. Biomarkers have been identified in a range of human tissues and fluids, including blood, brain, cerebrospinal fluid (CSF), urine, feces…

An ideal biomedical diagnostics biomarker should comply with the maximum possible number of the following criteria:

a) It should be present in a minimally invasive source.
b) It should be as much sensitive as possible, in order to be useful for early diagnosis and also be specific, to avoid the potential confounding factors caused by external factors.
c) It should vary fast in response to any change in its biological context, like the disease progression and/or treatment.
d) It should provide a better understanding of the disease's mechanisms, provided that it should be relevant in a biological framework.
e) It should be useful in risk stratification and prognosis.

Biomarkers can be classified into three categories, depending on their usage: predictive, prognostic and diagnostic biomarkers. Predictive biomarkers are used to study responses to therapeutic interventions, such as cancer treatments. Prognostic biomarkers are used to identify the risk of disease progression and/or recurrence, such could be genetic markers for hereditary diseases. Finally, diagnostic biomarkers are those used to identify which patients have a specific disease [5, 6]. Is in this last class of biomarkers that we will focus on the work generated in this thesis dissertation.

When defining a new biomarker, several issues may arise that can complicate the path from the biomarker discovery to its implementation into clinical practice. This has

become especially true with the development of high-throughput omics technologies, which has supposed an increase in the number of potential biomarkers proposed that had failed to be validated and reproduced [7–9]. Nonetheless, it has been calculated that only 0.1% of potential biomarkers have been later implemented in clinical practice [10]. To put this into context, between only 2008 and 2009, the NIH funded with 2.5 billion dollars grants related to biomarker discovery [11]. This lack of success in biomarker discovery studies may be reduced by taking some actions during the discovery process. Indeed, biomarker discovery studies usually involve the following steps: initial discovery in basic studies, validation of potential biomarkers and final clinical implementation.

Usually, biomarkers are discovered and firstly validated in the same study. Thanks to technological developments in omics analyses, thousands of molecules can be measured without previous assumptions, so that hypotheses on biomarkers may be generated after the data analysis, in a data-driven fashion [12]. The cohort generation is an important step and careful planning and clear inclusion and exclusion criteria are needed to avoid potential confounding effects. Either the study is a cohort or a case-control one, metadata is important to correlate potential findings to any external factor that may explain them, such as sex, age, diet, lifestyle, etc. [13]. In case-control studies, matching the potential confounding factors between the cases and the controls will reduce their impact upon the final dataset. To do the validation, ideally, an independent cohort should be used, although is common to use cross-validation methodologies in the same set of patients from the discovery phase [14]. Finally, a detailed report on technical, methodological and bioinformatics protocols is required to ensure reproducibility of the results. This requires also the release of patients' data that may be legally protected, so that it may be difficult depending on the legislation.

Once the cohort is correctly established, the diagnostic biomarker performance is needed to be computed. For this assessment, the number of correctly identified cases and controls is computed, so that we finally have a recount of true positives and true negatives, as summarized in Table 1.

Table 1: *Typical structure of the results summary for a predictive biomarker model.*

	Cohort	
Biomarker result	Control	Case
Positive	False-positive	True positive
Negative	True negative	False-negative

From the table generated, we can calculate the three factors used to characterize the performance of the potential biomarker: sensitivity, specificity and accuracy. Sensitivity is defined by the percentage of correctly identified case-patients, while specificity is the percentage of correctly identified control individuals. Accuracy is the percentage of correctly identified individuals, either cases or controls. Finally, the difference between the true positive predictions and the false positives is known as the positive predictive value (PPV), while the difference between true and false-negative predictions is the negative predictive value (NPV). Sensitivity and specificity can be used to construct a receiver operating characteristic (ROC) curve (FIGURE 1) [15], a plot that depicts the performance of the biomarker, plotting in the X axis the false positive fraction (1-specificity) and in the Y axis the true positive fraction (sensitivity). Once the ROC curve is plotted, the predictive capability of the biomarker may be assessed by computing the Area Under the Curve (AUC) value. A non-informative biomarker will have an AUC value of ≤0.5, while values over 0.5 will represent a predictive biomarker.

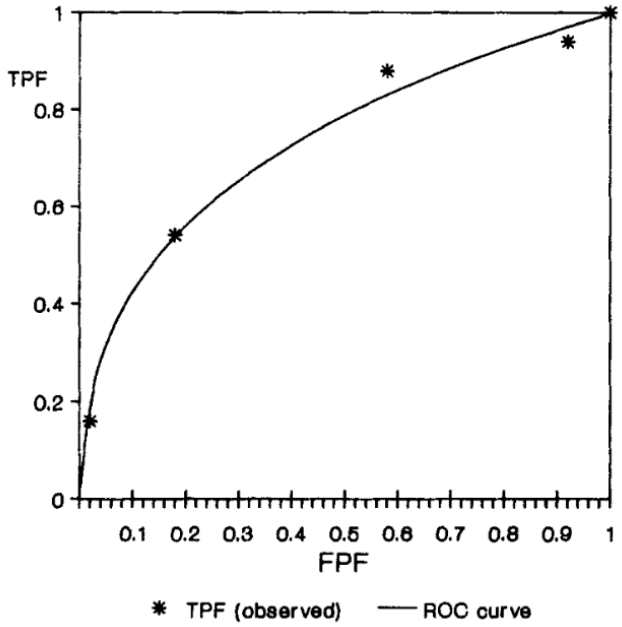


Figure 1: *Typical ROC curve, with False Positive Rate (1-specificity) in the horizontal axis and True Positive Rate (sensitivity) in vertical one.*

1.1.1.-Extracellular Vesicles

One of the most important criteria is the accessibility of the potential biomarkers. This criterion is met by the liquid biopsies, which consist of the analysis of non-solid tissues, such as blood and urine [16]. Liquid biopsies are generally focused on the analysis of circulating tumor cells, free circulant DNA, a set of RNA molecules and extracellular vesicles (EVs) [17].

EVs are microvesicles released by nearly all types of cells which mediate in cell-to-cell communication and signaling [18, 19] (FIGURE 2). EVs contain nucleic acids, proteins and lipids, encapsulating them into a lipid bilayer membrane [19] that protects them from degradation in environments such as blood [20]. While no common protein to all EVs families have been identified, they are commonly enriched in tetraspanins (CD9, CD63,

CD81), major histocompatibility complex (MHC) and other cytosolic proteins, like heat-shock proteins (HSPs) [21]. Relevantly, all these proteins derive from cytoskeleton, cytosol and/or plasma membrane, while no proteins coming from Golgi, endoplasmic reticulum or nucleus have been identified in EVs [22] (FIGURE 2). It has been reported also that tumor cells secrete more EVs that healthy cells [23]. Thus, they represent a valuable source for biomarker discovery process by the analysis of their contents.

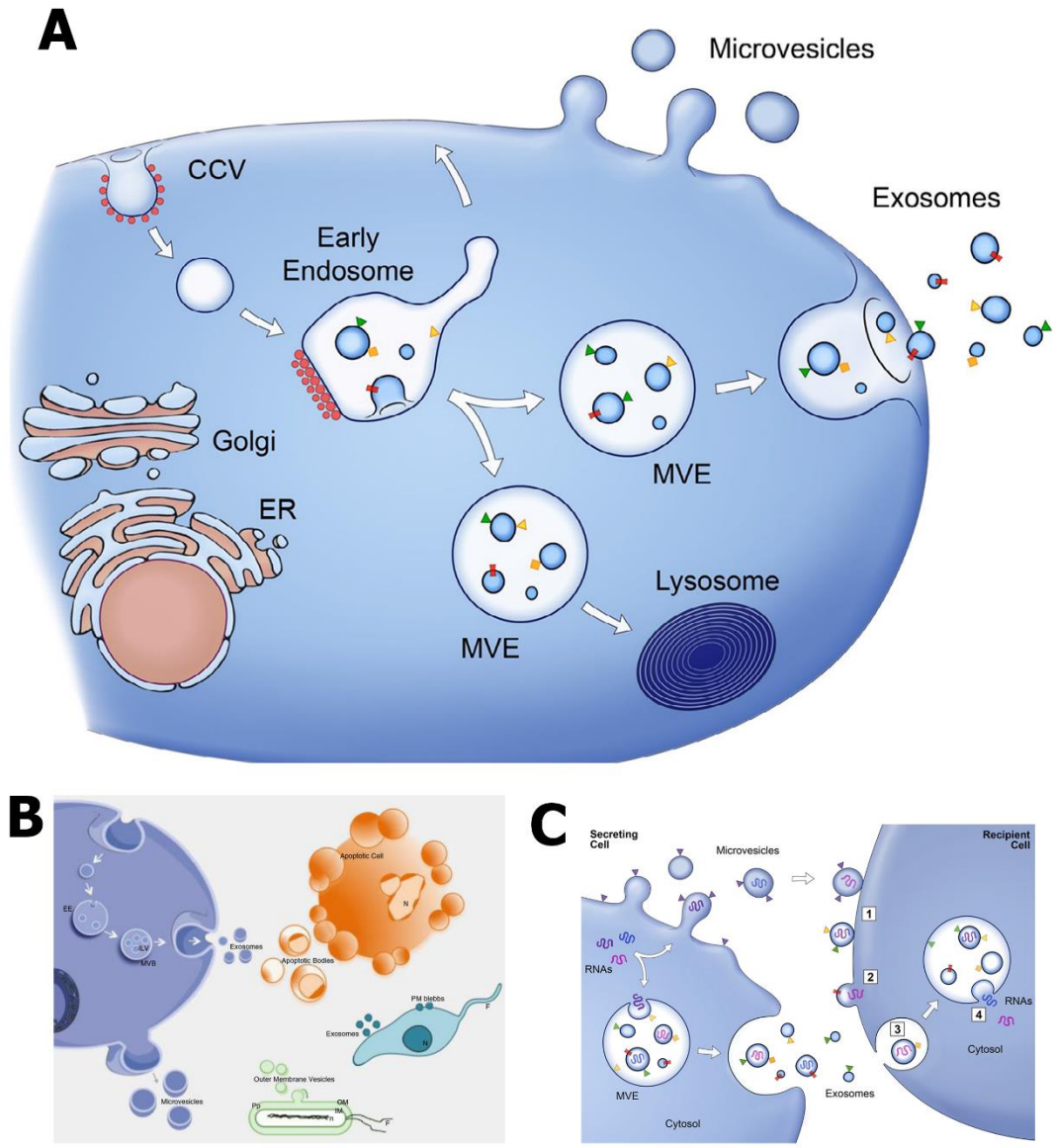


Figure 2: *Biogenesis and release of EVs, adapted from* [19, 24]. *Proposed mechanism of biogenesis by invagination of plasma membrane and release of microvesicles and exosomes (A). Types of EVs depending on the origin and physicochemical characteristics of them, including exosomes, microvesicles, apoptotic bodies and outer membrane vesicles (OMVs) (B). Transference of proteins and RNA between cells mediated by EVs. RNA is internalized in the EVs, while proteins can be membrane associated (triangles) or transmembrane proteins (rectangles) (C).*

1.2.-Data analysis for multi-omics

Multi-omics approach to unravel biomarkers and mechanisms underlying diseases started with the genomics in early 90´s, due to technical development that allowed cost-effective high-throughput analyses to exist. This allowed researchers to map diseases to an enormous number of genetic variants, hence suggesting that, for the majority of diseases, genetics only explained a fraction of their pathology. This fact supposed for the scientific community to assume that diseases, even those clearly regulated by genetic factors, were also dependent upon environmental factors too, something that was not possible to study with genomics, highlighting the requirement of a systemic approach that could integrate multiple factors among genetics. Nowadays, it is assumed that the majority of diseases that affect adults are multifactorial and their phenotype depends on a combination of both genetic and environmental factors [25–27]. This led to the development of new omics technologies and the respective analyses techniques that can provide a holistic understanding of cells and organisms and the progression to disease [28–30]. Technological advances that resulted in a reduction of the cost of omics technologies have also impacted on the advance and acceptance of these multi-omics studies among the scientific community [31–34].

Each omics technique will provide a list of differential features between two populations, depending on the kind of measurement performed. Nowadays, the most common omics technologies are genomics, epigenomics, transcriptomics, proteomics and metabolomics. Microbiome analysis has been recently also included in this family, although it supposes the study of multiple organisms and can be coupled to the other omics, generating metagenomics, metatranscriptomics, etc. fields. Thus, genomics provides with genetic variants, epigenomics provides with epigenome changes, proteomics with protein alterations, etc. All this data, either taken alone or combined, will provide a list of potential biomarkers that can be used both as diagnostics tool and to provide with a biological context, by mapping them into metabolic pathways. This functional characterization of biomarkers can also aid in the therapy designing, targeting the efforts against the identified altered pathways. One omics analysis is limited to correlations with a specific variable of the population, thus reflecting mainly reactive processes instead of causational ones. The combination of multiple omics, however, allows the identification of potential alterations that can lead to disease progression, causative alterations, instead of only reactive ones [35].

Omics technologies can be ordered in a fashion such that they reflect their proximity to the final phenotype, by plotting the importance of genetics vs the importance of the environment, as seen in FIGURE 3.

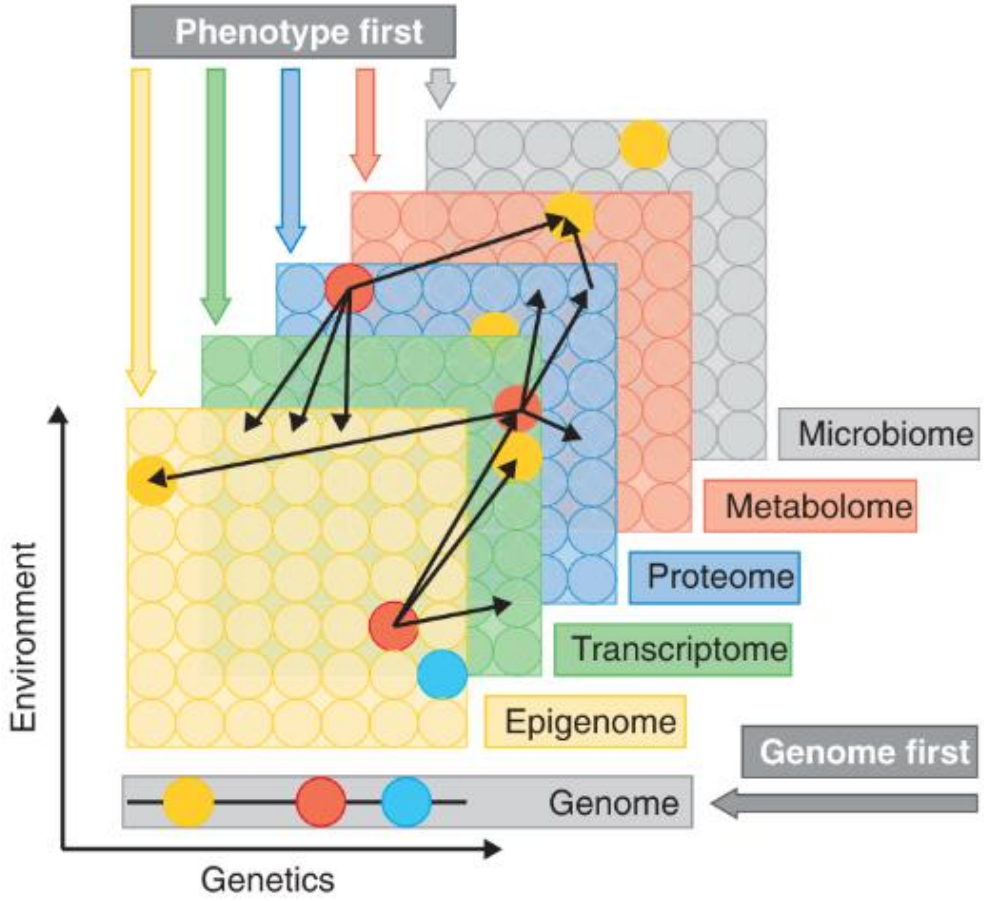


Figure 3: *Organization of omics technologies depending on the impact of each factor. Genome is only represented to be affected by genetics, while the other omics technologies are affected by both, genetics and environment. Arrows represent the potential interactions between different omics features. Circles represent the molecules identified per omics layer, being those colored potential altered ones.*

The integration of distinct omics techniques allows us to understand the flow of information, either starting from genetic alterations to the final phenotype (**genome first approach**) or by starting from the phenotype and trying to reveal which biological alterations may explain it (**phenotype first approach**). The decision on which of these two options should be followed for a multi-omics study will depend highly on the characteristics of the disease studied. Simple diseases may be explained by one or a small subset of gene mutations while more complex ones are usually defined by a combination of factors. Moreover, the phenotype of complex diseases may be the same even when the factors involved are distinct. Thus, the requirement of a multi-omics approach becomes evident in order to be able to elucidate the full complex combination of factors that leads to a specific disease phenotype.

1.2.1.- Omics challenges

The challenge with multi-omics studies starts with the special characteristics of omics data itself. Each omics technique will generate, by default, extremely large, highly variable and noisy datasets [36]. These large datasets are complex and full of redundant non-informative data. Therefore, it may be tricky to assess their relevance and quality.

The solution to this issue will mostly depend on the omics data type and the research community developed standards. Also, because of these specific-omics features, a proper multi-omics study will require the collaboration of multiple researchers, each one specialized in one of the omics involved in the study, in order to ensure that the appropriate standards for each omics measurement and analysis are followed.

Once each omics issues have been resolved, challenges on how to integrate different omics technologies arise. The above discussed fact of the generated datasets complexity implies that most omics will generate qualitative data, instead of a quantitative one. Qualitative data is, by default, hard to reproduce and nearly impossible to compare. Consequently, if only qualitative data is available its integration is hard, when impossible if data is coming from different sources [37–39]. While most of the omics technologies tend to provide qualitative data, this can be solved by the use of standardized operating protocols (SOP) and reference standards that would help in making these studies reproducible. If all these controls are applied and quantitative data is obtained, multi-omics can be performed and even comparisons between distinct laboratories.

Because of the high variability obtained in these high-throughput analyses, metadata collection is also one of the important aspects to consider. The elevated number of variables measured in omics studies makes it feasible that some of them may be affected by environmental factors that need to be controlled and removed [40]. Recurrently, a lot of time and money is invested in collecting molecular data, while no time is destined to metadata collection. This lack of metadata will drastically reduce the quality of the collected molecular omics data, making it non-reproducible and confounding the potential results and biological observations. Hence, it becomes clear the importance of getting data as well annotated as possible, although lack of quality metadata can be overcome by increasing notably the sample number included in the study, which will also increase its cost [41].

Collecting data issues have been discussed already, but a research project does not end there. When analyzing the data generated, other issues will arise. The bioinformatics community is enormous and tends to produce an important number of tools too, each one suited for a specific requirement. Thus, the catalog of tools and software dedicated to multi-omics data integration enormous. Knowing each one of them is just impossible and researchers usually just use what they know how to use or what is in fashion in a specific moment. This could be partially solved with the use of a centralized tool repository that lists all these tools, such as *Bioconductor,* CRAN R packages repository or OMICtools [42]. Benchmarking studies are also a useful tool to help researchers not only to discover new tools but to make better choices in analytical tools selection too. Benchmarking studies use gold-standard data sets to test a set of algorithms and tools in order to identify which are the best ones, by measuring distinct metrics, usually related to the tool performance [43]. This overcrowded tools market for distinct omics technologies analyses contributes to worsen the election of gold standards. One of the most critical for omics analysis is, in fact, the lack of a **gold standard**, either for data nomenclature, processing, and analysis.

At this point, it is also important to consider the usability of the distinct tools. Programs are just a compendium of code lines, which bioinformaticians usually understand and know how to use. To an experimentalist, though, this can be a limitation to choose a tool to use. Hence, it's important to consider the user-friendliness of the developed tool and the operative system compatibility.

Each omics technology will generate data in its own format and even in the same omics technique, distinct machines will export distinct data formats. Remarkably, public databases store data in distinct formats. All this combination of data formats highly limits the use of published data, adding another step in analytical protocols in order to transform all data into the same format. Furthermore, if distinct tools are used, data will need to be transformed to each tool-required format. To solve these issues, the bioinformatics community is starting to move towards open access research and to develop what has been called the FAIR data standards, which stands for Findable, Accessible, Interoperable and Re-usable software and databases development [44].

Finally, the most important issue and probably the most easily solved is the lack of funding for these kinds of projects. Multi-omics studies are expensive, as they make use of expensive equipment and require the involvement of multiple highly-specialized researchers. Even though the technological advancements in the last years have made, the huge number of samples still makes these kinds of studies to be so expensive. Lack of funding for multi-omics studies, apart from the countries on which research investment is not a priority, may be explained in part because of the fail of this field to demonstrate its utility and vindicate itself. While several studies have been already performed, just a few of them have effectively developed new technologies and/or drugs, being instead more basic research oriented, which is less funded (although extremely necessary).

Summarizing, the challenges of multi-omics integration studies include specific-omics processing data issues, the involvement of multiple specialized researchers, requirements for well-designed and annotated studies and lack of standards, both experimental and analytical.

1.2.2.- Integration methodologies

In order to integrate distinct omics technologies, the first-things-first rule is required. Hence, since data comes from different technologies, it is important to firstly normalize (and analyze) individually each dataset. After this step, data will be more or less equilibrated in order to be able to combine and integrate it [45]. Then, both machine learning and statistical tools will be needed in order to integrate distinct omics [46–48].

While the techniques dedicated to integrating distinct omics datasets will vary depending upon which omics we are integrating, there are some common methods that can be used. Among them, probably the most used are the dimension reduction techniques. These techniques take a complex dataset, with usually thousands of variables and generate a new one resulting from the decomposition of the complex dataset into new variables, called components. These components derive from the

combination of distinct variables in such a way that they explain the potential differences between the samples included in the original dataset [49].

1.2.2.1.- Dimension reduction techniques for omics integration

One of the first steps to consider in omics data management is to perform Exploratory Data Analysis (EDA) which consist in summarizing the characteristics of a dataset and allows the identification of potential batch effects and confounding factors [50]. Usually, two approaches exist for the EDA step, knowingly cluster analysis and dimension reduction one. Dimension reduction techniques analyze the whole dataset, thus conserving its global variance, while cluster analysis focuses on individual variable relationships [51]. Hence, cluster analysis loses this information on the relationships between variables, as each variable can only be in one cluster. Because biological phenotypes are often complex and depend on the combination of a set of factors, dimension reduction techniques are more appropriated for omics data analysis, as they retain the potential interactions between variables. They also allow the integration of multiple omics datasets, which in turn helps in discovering global correlation patterns among datasets, so that the findings are more robust against outliers and batch effects.

Usually, omics datasets are presented in a matrix form, where samples are the rows and variables in columns. Because of the high-throughput nature of omics analyses, usually, the number of variables is several times bigger than the number of samples. Dimension reduction tries to identify a set of variables that linearly combined generate a component. Then, each generated component is combined in a new dataset so that this reduced dataset has less components (combinations of variables) than samples. This can be formulated like:

Being **X** the omics dataset a $n \times p$ matrix, where *n* represents the observations and *p* the variables:

$$\mathbf{X} = \left(\mathbf{x_1}, \mathbf{x_2} \ldots, \mathbf{x_p}\right) \quad (1)$$

Dimension reduction will generate a set of new components combining different variables so that the resulting dataset will have fewer components than samples:

$$\mathbf{f} = \mathbf{q_1 x_1} + \mathbf{q_2 x_2} + \cdots + \mathbf{q_p x_p} \quad (2)$$

f is the new variable, which may be called latent variable, component, principal axis, an eigenvector or latent factor, depending on the research field and the researcher. **q** is a *p*-length vector of coefficients in which at least one is not zero, which can be also called loadings. What dimension reduction techniques do is to find a set of **q**'s that maximize the variance of **f**. Depending on the constraint criteria used to achieve this reduction and the optimization protocols, different reduction methods exist. The most commonly used dimension reduction approach is the Principal Components Analysis (PCA) [52].

PCA was first described by Pearson as early as 1901 [53] and in 1933 by Hotelling [54]. Each principal component (PC) of a PCA analysis is a linear combination of original variables in a way that the first PC explains the most variance, meaning that the largest

dispersion of values will occur among this component. Seemingly, second PC will retain the second largest variance and so on [51]. Different PCA types exist for different types of data, which differ in how data is transformed before the decomposition step [55, 56]. PCoA, for example, is indicated for distance matrices and is mostly used for continuous data, being appropriated for microbiome data analysis.

Canonical correlation analysis (CCA) is indicated to the integration of two datasets that, once decomposed in factors, are analyzed in order to identify the variables that most correlate the two datasets. Variations on this approach have been developed in order to improve the performance of the methodology and to adapt it to the new kind of data derived from high-throughput analysis. While CCA cannot be applied to datasets with more variables than samples, new approaches have been proposed that solve this pitfall. Thus, options like sparse CCA [57] or penalized CCA [58] have been applied to the integration of omics datasets. Sparse solutions filter the number of variables, thus simplifying the analytical process and the interpretation of the results.

Another technique widely used is the Partial Least Squares approach (PLS), which does not suffer for these CCA-related constraints. Instead of working with correlations, PLS analyzes the covariance between different components of the datasets, which also makes it more robust against outliers. It has been shown that, for multi-omics studies, sparse PLS works, at least, as well as sparse CCA methods [59].

In addition, Procrustes analysis allows the comparison of the distribution of samples between two PCAs analyses of two datasets [60], by the comparison of its shapes. It superimposes the shape of two PCA objects, moving, scaling and rotating one of them until the difference between shapes is reduced to its minimal value [51].

Finally, co-Inertia Analysis (CIA) is used to combine two different datasets. While it was developed for the integration of ecological measurements, it has been successfully applied to omics integration too [61, 62]. It consists of two steps: (i) perform a dimension reduction technique on each dataset (PCA or similar) and (ii) constrain the projections of the orthogonal axes so that the covariance is maximized [61, 63]. Instead of sparse methods, the CIA does not remove any variable, so that the final result tends to include redundant and non-informative information, hence making more difficult its interpretation [59]. Its finality is to find similarities between two datasets using ordination spaces and being quantitatively measured by a coefficient called RV [51].

1.2.2.2.- Omics – microbiome integration

In general, the integration of microbiome with other omics relies on dimensionality reduction approaches, without being more specific than that, except probably for metabolomics data. Nowadays, omics – microbiome integration studies have been performed with a broad combination of omics, including genomics, epigenomics, proteomics and metabolomics [64]. Strategies for these omics combinations have relied upon correlation measurements, regression approaches and network-based ones, apart

from the dimension reduction ones [64]. In fact, it is not unusual to combine dimension reduction techniques with some of the aforementioned methods.

For the combination of the microbiome and host genomics, it is important to study the heritability of the microbiome, which refers to the part of the microbiome variance explained by genetic variants in the population [65]. This is to ensure that microbiome data is, in fact, hereditary and to remove those variables that do not have this hereditary component. Later, microbiome – genome associations can be tested at individual genes level or the whole genome level. To test individual associations, correlation-based approaches have been previously used, either Spearman's or Pearson's correlations [66, 67]. For full genome associations dimensionality reduction methods are used, such as PCA or PCoA (principal co-ordinate analysis) [67]. The drawback of this approach is that identifying specific correlated variables is less straight forward. With these approaches it has been possible to better describe the impact of diet on microbiome composition, describing correlations between lactase [66] or vitamin D receptor [68] genes and specific bacteria genera. Similar approaches have been followed for the integration of transcriptomics and microbiome, although correlation approaches have been improved to multivariate correlation, like the canonical correlation analysis (CCA) [69]. From these correlation methods, network techniques have also been developed, such as the Weighted Gene Co-expression Network Analysis (WGCNA). WGCNA is a data-driven technique that clusters together those variables (bacterial OTUs and transcripts) most correlated.

Finally, there is the metabolomics – microbiome integration. As mentioned above, while other omics integration with the microbiome is pretty much performed the same way, with metabolomics things are slightly different, as metabolomics datasets need different data pre-processing. Once that is done, the integration may be performed by simple correlations or via dimensionality reduction techniques. Other integration techniques include Procrustes and co-inertia analysis (CIA) that allows the comparison of two PCAs (or PCoAs) from distinct omics datasets. Both techniques are used to study the similarity between two datasets and to establish a correlation value for the whole datasets. While Procrustes is mostly used for these global dataset comparisons, CIA is useful to identify which variables are the most similar between datasets. This approach has been applied, for example, to study the interactions existent between distinct parts of the intestinal mucosa and the microbiome composition [70]. The combination of correlation and network-based methodologies have been used to describe the relationship between microbiome and host's insulin sensitivity [71] or the development and progression of CRC [72, 73].

1.3.- Metabolomics

Metabolomics is one of the latest incorporations to the omics research field [74]. It is dedicated to the study of the metabolome, which was defined by Oliver *et al.* 1998 [75] "***as the complete set of metabolites/low molecular weight compound which is context dependent, varying according to the physiology, development or pathological state of the cell, tissue, organ or organism***". In this metabolites definition, we include all those molecules that are the end-product of any metabolic reaction that occurs in the cell [76]. We consider to be a metabolite those molecules that have a molecular weight <2,000 Da [77]. The complete catalog of all these small molecules is what we call the metabolome that, combined with the omics suffix, refers to metabolomics. Metabolites are biologically active compounds implicated in each biological process of a living cell, from building blocks for macromolecules to energy carriers. Thus, they are part of the thousands of metabolic and biosynthetic pathways required for all the cells and organisms functions [78].

The metabolome is constituted by distinct classes of compounds, from lipids to amino acids, including inorganic species and derivatives of hydrophilic lipids. Thus, unlike genes or proteins, metabolites are harder to study, as they possess an extreme variability in terms of the order and subgroups of the atoms than the 4-letter (genes) or the linear 20-letter (proteins) codes. The huge range of compounds included provides a large variety of chemical and physical properties, such as molecular weight, polarity, solubility and volatility. The number of metabolites included in a determined metabolome depends on the organism studied, ranging from the 600 metabolites of *Saccharomyces cerevisiae* [79] to the 200,000 metabolites of the plant kingdom. While no metabolome size has been proposed for humans, it is accepted that it will be smaller than the plants' metabolome, although more than 4,000 different metabolites have already been annotated [80]. Inside the metabolome, we can differentiate between primary and secondary metabolites. Primary metabolites are those ones that are involved in biological processes, thus making them essential to life. This is the case of amino acids, organic acids, lipids, etc. Secondary metabolites, instead, are those metabolites not essential to cell life, because they have no role in any essential biological process. Hence, they are restricted to a selected set of cells, which synthetize them for specific biological functions. This class of metabolites include xenobiotics, metabolites that come from external sources, such as diet, medication, interactions with microbiota, etc.

Instead of the traditional sequencing approach, useful for genomics, transcriptomics and proteomics, metabolomics requires a *de novo* elucidation of their elemental composition, atoms order and stereochemical orientation. To deal with this complex situation, different approaches have been developed, grouped in two ways: untargeted or targeted metabolomics.

1.3.1.- Targeted metabolomics *vs* untargeted metabolomics

Targeted metabolomics is mainly used for screening purposes. This is an approach that takes into consideration one to few metabolites, usually related between them, in order

to study the effects of a specific alteration. In these kinds of studies, sample preparation protocol should be adapted to the chemical properties of the targeted compounds, to reduce the potential confounding factors and the matrix effect. Matrix effect refers to the events by which metabolite signals may be altered by other metabolites signals [81]. These signals may come from metabolites coming from the sample matrix, may be altered due to degradation of metabolites during processing steps or to instrument influences that may corrupt metabolites quantification [82].

Often, an alteration does not produce a simple, located alteration in a small subset of metabolites. Instead, either because of one metabolite can be found in several biological pathways, the potential interactions established with other metabolites and/or pleiotropic effects, it is plausible that the alteration may impact on other pathways. Therefore, this requires that all metabolites should be measured and considered to describe the role of some biological alteration, which is what we know as **untargeted metabolomics**. The inclusion of the whole metabolome into a study has its own associated complications. Ideally, metabolomics should aim to the inclusion of all metabolites, but this is not always achievable. Therefore, in order to include the maximum possible number of metabolites, different extraction and sample preparation protocols may be needed. The chosen analytical method must have strong enough resolving power maintaining a good sensitivity and selectivity. Ideally, it should be matrix independent too. Because of the nature of the data obtained, a methodology to identify unknown metabolites should be established too. A variant of the untargeted approach is the metabolic fingerprinting, in which full metabolite identification is not needed, as it is used for the classification of samples between two (or more) conditions. Metabolomics datasets are complex by definition. Therefore, appropriated tools are needed for the handling, storing, normalizing and analyzing of the data.

Although technically the first metabolomics studies were performed by gas chromatography mass spectrometry (GC-MS) as early as in the 1970s [83, 84], the later improvements in analytical instruments and the computer advances are the ones responsible for a more broader implementation of metabolomics, enabling the acquisition and interpretation of comprehensive metabolic profiles through multivariate statistical tools [85]. Traditionally, the most used metabolomics technology is Nuclear Magnetic Resonance (NMR) spectroscopy. It offers some advantages, as it is a rapid, high-throughput non-destructive technique with small to no sample preparation. Mass spectrometry (MS), though, provides better performance in sensitivity, something that facilitates the measurement of species with low abundance, an event quite feasible to occur in an untargeted metabolomics study. In contrast, MS needs more sample preparation than NMR and it is usually coupled to another separation technique, like gas or liquid chromatography (GC or LC), in order to better separate the sample metabolites. In this thesis, though, we will work only with MS technology.

1.3.2.- Metabolomics data acquisition and analysis

The typical **metabolomics workflow** is comprised by seven general steps: establish the experimental design (1), ideally within the input from a statistician, sample collection (2)

and preparation (3), running of samples in analytical instruments (4), data processing (5), statistical analysis (6) and pathway and functional analyses (7) [78] (FIGURE 4).

1.3.3.- Experimental design considerations

The experimental design (1) is the most critical step not only for a metabolomics study but for any kind of experimental work. Its design will have a huge impact on the quality of the results obtained. Therefore, well planned studies, with all the potential confounding variables established and controlled will lead to better and easier interpreted results. A good experimental design includes aspects like the correct number of samples, the appropriate classification of each group and a reduced sample variability within groups. It is useful to integrate all subjects related to the study in this step, from the sample collectors and processors to the bioinformaticians or statisticians that will analyze the data.

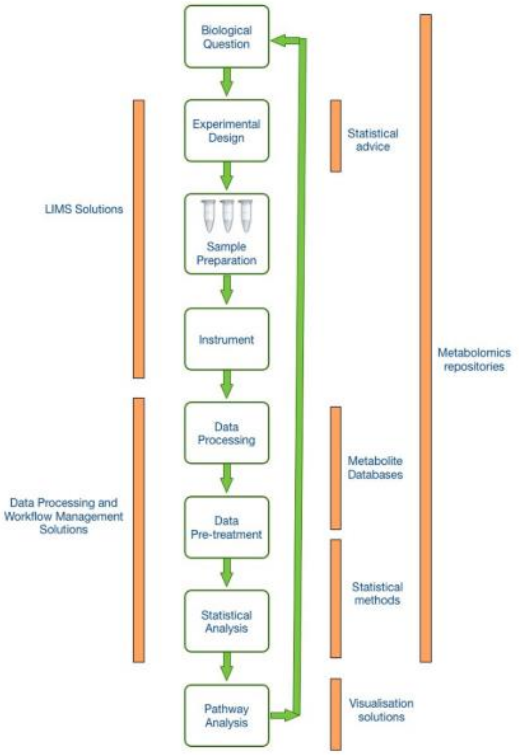


Figure 4: *typical metabolomics workflow, from* [78].

1.3.4.- Sample collection and preparation

Because of the wide variability of physical and chemical properties of different metabolites, there is no unique methodology that can capture the full set of metabolites of a sample. About sample collection, it is important to consider that each biofluid will contain a different set of metabolites, depending on their physicochemical characteristics. Thus, while blood metabolome composition is stable and controlled homeostatically, urine will mostly contain water-soluble metabolites [86]. A well-defined and documented sample collection protocol will also improve the reproducibility of the study. An important consideration to make is also the fact that metabolism is a constant flux, thus stopping any potential metabolic reaction is important in both sample collection, preservation and manipulation in order to avoid any potential loss of metabolites. For liquid samples, freezing to -20ºC or -80ºC is recommended [87] while for solid samples lyophilization is a good option. Conservation at 4ºC may alter the metabolite composition due to the effects of microbial communities present in samples [88]. Metabolite extraction methods from the collected samples have been developed to be effective for specific metabolite families thus each method will also suppose the loss of other metabolite features. Summarizing all potential issues in extraction methods, an ideal extraction step should [89]:

a) Incorporate a preservative so that metabolite composition reflects the original one in the sampling moment.
b) Be as non-selective as possible, in order to incorporate the broader range of metabolites possible. To this aim, a combination of distinct extraction protocols may be considered.
c) Be simple and fast, so that the potential metabolite loss is avoided.
d) Be reproducible.

Extraction methodologies will depend on several factors, including the sample type and the potential aims of the study, that will determine the metabolome coverage required (FIGURE 5). Hence, a targeted metabolomics analysis may need less metabolome coverage so that extraction protocols may be optimized to the metabolite families considered. Untargeted metabolomics, otherwise, is intended to cover as many metabolites as possible, so that they will need a combination of extraction methodologies to cover a wider metabolite range. Even though, some general recommendations can be identified depending on the sample type:

1) Liquid samples: simple, unselective methods are usually applied, like dilutions and solvent precipitations. They enable high coverage and are fast, while if later an LC-MS approach is used ionization suppression events may occur during the electrospray process.
2) Blood samples: the main issue with blood samples, either plasma or serum, is related to the high concentration of proteins. Thus, a protein removal step is required, which can be achieved with ACN or acetone.
3) Solid samples (feces): because of the phase of the sample, an extraction step in order to transfer the metabolites to a liquid phase is needed. Usually, samples come lyophilized, so that extraction is more homogeneous. For the extraction, a combination of solvents with different polarities may be used so that distinct metabolome may be obtained. Commonly, methods that allow the extraction of lipophilic and hydrophilic metabolites are used (chloroform-methanol, chloroform – methanol-water).

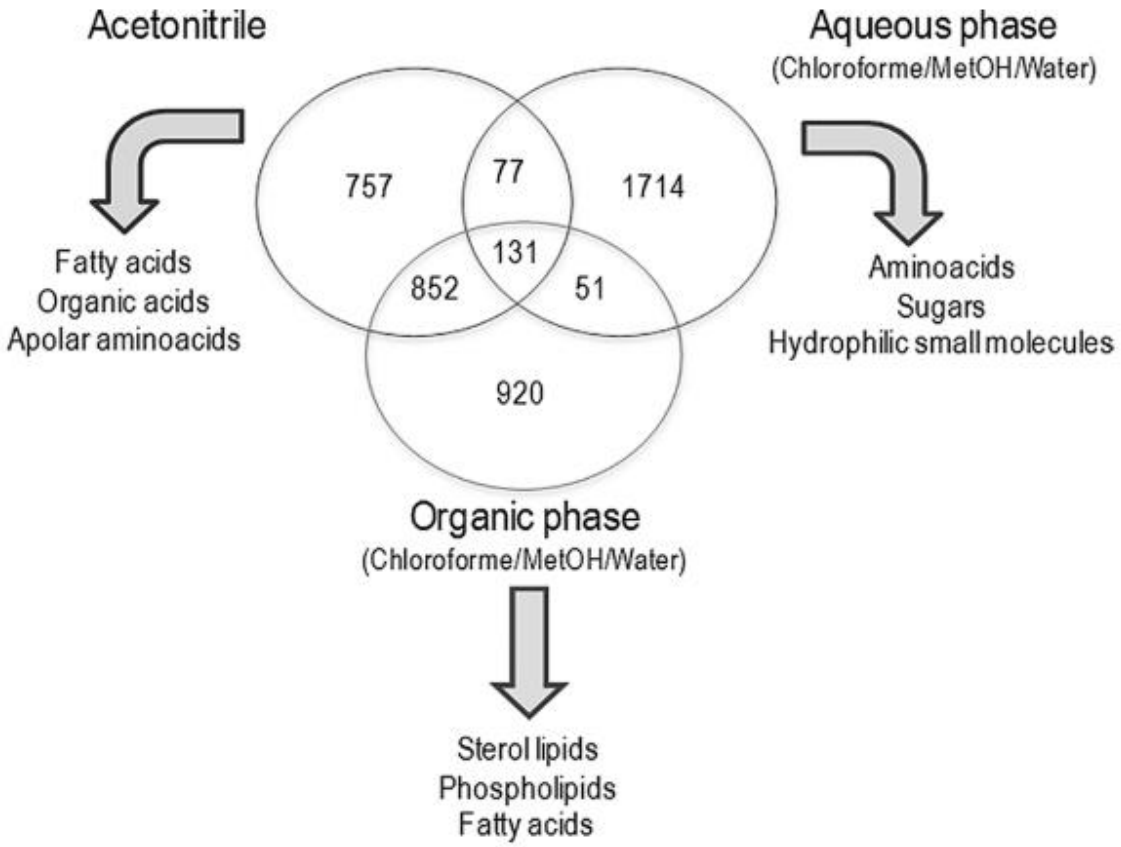


Figure 5: *Metabolite types extracted by different solvents, from* [85].

1.3.5.- Data acquisition

The high diversity of physicochemical features of the whole metabolome does not only affect the sample preparation step, but it impacts also in the data acquisition one. Thus, not one analytical technique will be able to capture the whole range of metabolites in an MS study. Again, a selection of techniques will be needed, biasing this way the output. One selection way is the coupling of separation procedures before the metabolite injection into the MS equipment. Depending on the separation methodology selected, different metabolite types will be conserved. Hence, LC-MS is recommended for the analysis of polar metabolites, while GC-MS works better for small volatile ones.

As El-Aneed *et al.* [90] elegantly defined: "***mass spectrometry relies on the formation of gas-phase ions that can be isolated electrically based on their mass-to-charge ratio (m/z)***". Since its invention in the late 1880s, it was not until 1957 that separation methodologies were coupled to it, being the first the GC [91]. Nowadays, MS is usually combined with other separation techniques, such as high-pressure liquid chromatography (HPLC) and ultra-high pressure liquid chromatography (UHPLC). The mass spectrometer is composed of an ion source, which is the entry point for the ionized metabolites into the equipment, a mass analyzer, which separates the metabolites by *m/z* and the detector. The *m/z* abbreviation indicates the number that results from dividing the mass number (*m*) of an ion by the corresponding charge number (*z*) [92].

1.3.6.-Data processing

Raw data obtained from the mass spectrometer usually are big. For example, for fibromyalgia metabolomics, 200 samples raw data occupied 17GB (approximately), thus making it manual management very difficult. Raw data needs to be converted to a more readable format in order to be able to statistically analyze it. This is usually performed by specialized software, able to deconvolute the full metabolomics spectrum to spectral bins, each one related to a specific *m/z* value.

Basically, this data processing step consists of four parts (FIGURE 4). It starts with the removal of the background signal that may confuse the posterior analyses. Then comes the peak selection, where each *m/z* value is assigned to a peak and its value reported for each sample. Following it comes the peak alignment process, where a correction on the retention time shifts that occurs during data acquisition is performed. This is done to assure that each *m/z* value for a specific molecule appears at the same retention time in each sample. The correction factor applied may be extracted from the use of internal calibration ions, which are injected with each sample, so that the analyst later can identify the corresponding retention time shift in each sample. Finally, missing values analysis should be performed. A missing value proportion threshold is established so that all molecules with more missing values than the threshold are removed from posterior analyses. With the remaining molecules missing value imputation strategies can be studied, being the most common ones using the average and/or the median value of that molecule in the samples or the minimal value divided by a specific factor [93, 94].

1.3.7.- Data analysis

Once data has been properly normalized and pre-treated, it is time to provide the data with a biological context and significance [95]. Usually, metabolomics data analysis is complicated due to several features that are characteristic of high-throughput data: overfitting, because of the much larger variables measured than samples (i); several variables tend to depend one on another (ii); elevated noise levels (iii); the need to separate informative data from the non-informative one (iv) [96]. Metabolomics data can be analyzed in two ways, applying univariate or using multivariate statistics, although a combination of both is recommended [97].

Univariate analysis is referred to as the analysis of just one metabolite at a time. In this approach, more standardized statistics are used to assess the differences between sample groups, such as *t*-test or ANOVA. If univariate statistics are used in an omics study, it is important to perform a multi-test correction to avoid the potential increase on false positives, as it is known that the chance of a comparison to be statistically significant increases proportionally to the tests made [98]. What multi-test correction does is to reduce the significance level so that the chance of getting false positives is reduced, although increasing the possibilities of false negatives.

Multivariate approaches, instead, compare all the variables simultaneously, so that the potential interactions between variables are also considered, considering the correlations and/or covariances between variables [97]. Due to data characteristics,

consisting of a relevant number of variables, usually much larger than samples, multivariate statistics is an appropriated choice. With either PCA [52] and/or PLS-DA [99] methodologies, patterns in the data will arise that may help in explaining specific phenotypes.

As each approach considers different features related to the variables analyzed, it is not strange that each of them provides results that may not agree between them. Reasons for this are various and reviewed in [97], but this event is expectable and nullifies neither multivariate nor univariate results. Differences are explained, mainly, because of the differential characteristics and assumptions of both statistical approaches. Furthermore, the combination of both approaches may result in a better, more holistic view of the results, with an improved fitting in a biological framework.

Either before or after this statistical analysis of the metabolomics data, metabolite identification is needed in order to assign an identity to each metabolomics peak identified. Depending on the kind of metabolomics analysis performed (NMR, GC-MS or LC-MS), the metabolite identification protocol will differ. For LC-MS, obtained spectra are compared against databases such as METLIN [100] or HMDB [101], so that a tentative identification for the metabolite is obtained based on the reported feature mass and the comparison between identified fragmentation patterns and the corresponding entries in such databases. A comparison against commercial standards, analyzed under the same conditions as the samples, is required for complete metabolite identification, as stated by the Metabolomics Standards Initiative working group [102]. During this identification process, issues may arise related to the metabolite nomenclature, especially when non-common metabolites are trying to be identified. Each database uses a particular nomenclature format and no standards have been defined yet. Therefore, for common metabolites databases usually incorporate a synonyms section, in order to facilitate the information retrieval process [103].

Finally, identified metabolites may be used for a more functional-orientated analysis, in order to identify which biological processes may be related to a specific phenotype, using pathway databases such as KEGG [104], WikiPathways [105] and/or Reactome [106].

1.4.- Metagenomics – microbiome analysis

1.4.1.- History of the microbiome studies

It is assumed that the human body contains between 2 to 20 million different bacterial genes, at least 100 times more than human genes (about 20,000 genes) [107–109]. Even though, while the human genome has been widely studied and analyzed, the impact of this huge number of bacterial genes upon the human organism has just started to be studied. Interestingly, while the human genome is fixed at birth and cannot be changed, microbiome composition can be and is modified by a high number of factors [110], including environmental and lifestyle-related factors and diseases.

Technically, the first study of the human's microbiome was performed as early as 1677 by Antonie van Leeuwenhoek, considered the father of microbiology. The inventor of the microscopy, he used it to compare not only his own oral and fecal microbiome but also the microbiome from other people and even the effects of distinct lifestyle factors, such as tobacco and alcohol consumption, upon the microbiome composition [111, 112] (FIGURE 6).

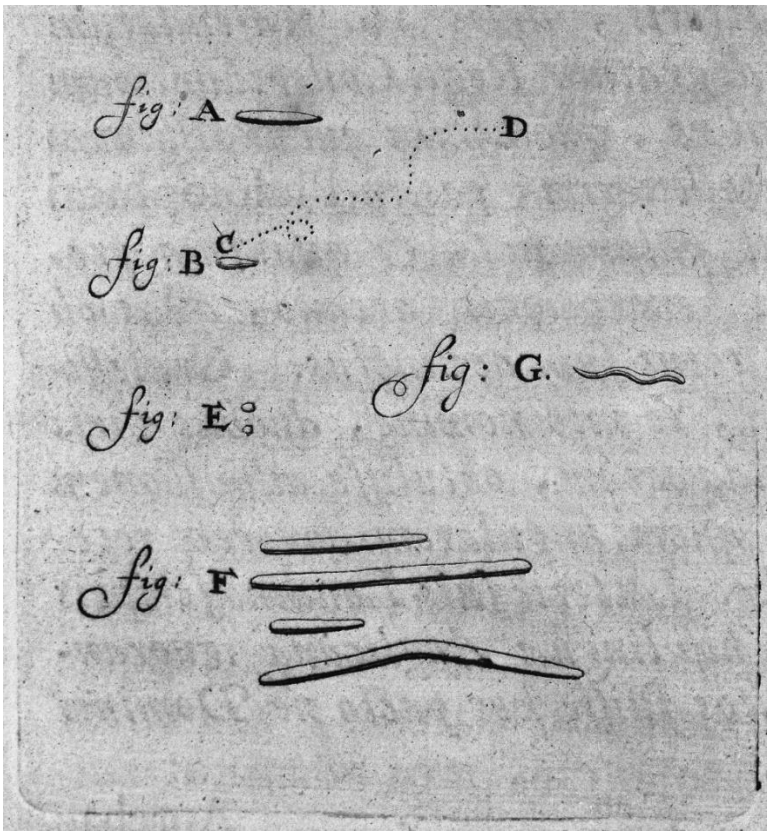


Figure 6: *Oral bacteria as seen by Antonie van Leeuwenhoek, first published in Arcana naturae detecta (Antonie van Leeuwenhoek, 1695). Credit: Wellcome Collection.*

Since then, the microbiome studies have evolved quite significantly. Until the introduction of the 16S rRNA sequencing technique, microbiome studies relied on culture techniques. Even then, it was clear that culture techniques were not able to capture the whole diversity of the human microbiome, assuming that only 20-40% of it

was culturable [113]. The first authors to propose that gene sequences could be appropriate to study evolution, using them as "molecular clocks" to study phylogenies were Pauling and Zuckerkandl [114]. Using their proposal, but focusing on ribosomal RNA sequences, Woese and Fox were the kingdoms division of the life in kingdoms [115], introducing also new domain among the previously existent, the *Archaea* one [116]. They introduced the study of ribosomal RNA, a component of all self-replicating systems, easily isolated and with a sequence that, although it changes with time, it does it slowly, permitting the comparisons between distant species. This became a new standard in the 1980s, mainly with Woese's work demonstrating that phylogenetic relationships of bacteria could be determined by using stable parts from these ribosomal RNA sequences [117, 118]. The adoption of 16S rRNA gene as a gold-standard molecular marker for microbial studies started after that, between 1990-1991, with the first studies being done in environmental microbiology field [119, 120], although the first observations about 16S rRNA sequence conservation were done by Dubnau *et al.* as early as 1960 [121]. In 1999, the first study of human gut microbiome using the 16S rRNA was published [122]. In this study, the authors demonstrated that only 24% of the molecular species identified from the 16S rRNA sequencing corresponded to previously known bacteria. Since then, several studies have been done trying to better characterize the human microbiome and to increase the number of bacteria present in it. Among them, probably the most important study performed to date is the Human Microbiome Project (https://hmpdacc.org).

1.4.2.- The Human Microbiome Project (HMP)

The HMP is an initiative from the National Health Institutes (NIH) established in 2007, in order to study and develop the required tools, protocols and resources to allow the characterization of the human microbiome and its role in health and disease [109]. The HMP is a conglomerate of multiple projects, developed worldwide (including the United States, Europe and Asia). It is not only focused on the effects of microbiome changes in human health but also to determine which factors (either genetics or environmental ones) contribute to the definition of microbiome's composition.

300 healthy individuals were recruited for the HMP study. To determine the structure and function of the healthy human microbiome, 4,788 specimens from 242 of these individuals were analyzed [107]. Two different kinds of data were generated, 16S rRNA sequencing and metagenomics whole genome sequencing (WGS). Different data can be obtained by each of the distinct sequencing options. While 16S rRNA sequencing is restricted to taxonomy analysis, as it just analyzes the 16S gene, WGS can be also used for functional profiling of bacterial communities, because all the bacterial genes are sequenced and analyzed [123]. Therefore, during the HMP study, 16S data was used to analyze the composition of the human microbiome and the potential existence of a core microbiome, while WGS data used to determine its functions.

Resulting from the immense amount of data generated during the project and, specifically, 16S gene reads, two bioinformatics tools, mothur [124] and QIIME [125, 126], were updated in order to incorporate the pipelines developed by HMP Consortium

for the analysis of microbiome data. These pipelines were developed in order to generate a standard data analysis workflow for the microbiome researchers community and to reduce sequencing errors generated by previous tools [127].

1.4.3.- Human microbiota: structure and function

The main results of HMP study (and from other studies like the European MetaHIT) allowed the identification of distinct microbiome niches in the human body, thus identifying different populations of bacteria depending on the sampling location. Notably, the specialization of the bacterial population in each one of these niches was so notable that the microbial communities from the same niche were more similar between individuals that microbial communities from distinct locations of the same individual [107] (FIGURE 7A). To this extent, meta-analyses also showed the importance of this niche-specialization of microbiota populations, showing a preferential clustering of samples by the body site before the clustering by the study [128]. One important parameter in the study of microbiome is named "Microbial diversity" that can be defined as the number and abundance distribution of different bacterial species in a specific environment. By body sites, the most diverse ones where those related to the gastrointestinal tract, including both the oral cavity and the gut and stool. Skin presented with reduced diversity, probably due to modern hygiene habits and finally, the vaginal microbiome was less diverse, mainly dominated by *Lactobacillus* [129]. Gut microbiota is typically composed more than 1,000 bacteria species, although two phyla dominate over the rest, Bacteroides and Firmicutes [107, 108]. Relevantly, it has been demonstrated that stool microbiota composition resembles mostly the colon-specific microbiota, while showing moderate similarities with distal small intestine [130].

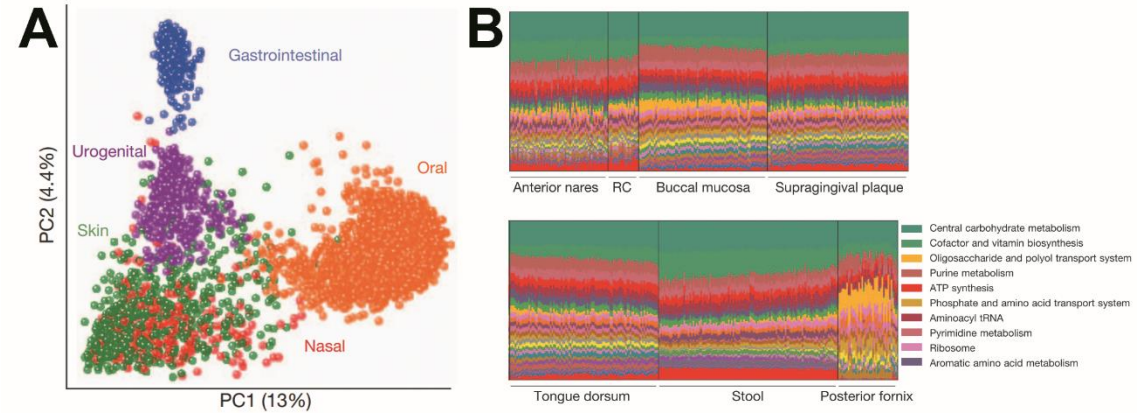


Figure 7: *HMP project results. PCoA plot of HMP data showing samples primary clustering for body area (A) and metabolic pathways abundances in each body area studied. Figure adapted from* [107].

While studies on human microbiota composition have been mostly focused on the bacteria, microbiota extends among the whole tree of life and it is known to include also archaea, viruses and eukaryotes elements too. 16S sequencing is restricted to bacteria identification, although some archaea can be identified too, such has been the case of *Methanobrevibacter* genus, prevalent in the gut [131]. WGS techniques have identified also viral-related genes, which also points towards the existence of a virome, which is suspected to be pretty extensive, with an elevated inter-variability [132, 133], formed

mainly by bacteriophages and with an important impact upon host health [134]. Eukaryotes found in the human body are typically pathogens, although some of them (*Candida, Malassezia* and *Saccharomyces*) are also found in healthy conditions [135]. Relevantly, multicellular eukaryotes such helminths have been a component of the gut microbiota on our evolutionary history [136], although nowadays is removed from Western-cultures gut microbiota.

Studies on microbiome composition have been focused on trying to define what a healthy microbiome is and which bacteria compose it (understanding healthy as the absence of any disease). With the emergence of metagenome and microbiome studies it became evident that the enormous amount of variation in the taxonomical composition of distinct people microbiome would make it impossible to define a common set of bacteria species between individuals [137, 109]. Instead, the idea of a core microbiome turned towards a functional core: an essential set of metabolic functions, common in all humans, provided by bacteria but not necessarily the same bacterial species [138]. The idea of the existence of a core of functionalities was supported by two facts: the broad taxonomical distribution of specific bacterial functions (essential housekeeping functionalities) and/or the enrichment of specific functionalities among bacteria colonizing a specific niche (because of a selective advantage) [139]. Researchers have been able to identify distinct core pathways depending on they were broadly expressed or niche-specific enriched by classifying them accordingly to the taxonomic range. Notably, those pathways related to housekeeping functions (such as biosynthesis of coenzyme A) were broadly distributed among all bacteria, independently of the niche sampled (FIGURE 7B). Pathways that were found to be specific for a niche were related to metabolic functions performed upon specific molecules found in those specific niches. Thus, pathways such as vitamin B12 biosynthesis and short-chain fatty acids (SCFA) production pathways were found to be functional core of gut and oral microbiomes, where bacteria have the required precursors and environmental conditions to perform them [139]. These site-enriched pathways suggest a functional adaptation of niche-specific microbiota to the human body. Notably, microbiota composition changes with time, but those core functional bacterial pathways were temporarily stable too [139].

The most important functional aspects of the gut microbiota includes nutrient metabolism, xenobiotic and drug metabolism, antimicrobial protection, immunomodulation, maintenance of the integrity of both gut barrier and gastrointestinal tract and communication with the Central Nervous System (CNS) [150]. Of this list, the influence of microbiome in nutrient metabolism is one of the most relevant for host health. Gut microbiota obtain their nutrients from carbohydrates derived from the host's digestion. The fermentation of indigestible oligosaccharides by gut microbiota species (*Bacteroides, Roseburia, Bifidobacterium, Enterobacteria* and *Fecalibacterium*) results in the synthesis of short SCFA (butyrate, acetate, propionate) [150], that are an energy source for the host and greatly contributes to its health [151–154].

The healthy microbiota, or healthy functional core, requires also of three characteristics: resistance, resilience and stability; because of the multitude of factors that influence, shape and modify the gut microbiota [137]. Resistance refers to the capability of gut microbiota to resist those perturbations, resilience to the ability to return to the pre-perturbation state (the healthy one) and stability to be more or less constant, either in their composition or their functional capabilities, among time [140]. Factors affecting the gut microbiota include geography [128] and evolutionary events [141], diet [142], host genetics [143], early-life colonization and establishment, including mode of delivery [144], breast-feeding [145–147] and diseases and clinical treatments, such as antibiotics [109, 148, 149].

1.4.4.-The 16S rRNA marker

Three different molecules of RNA are present in the prokaryote ribosome, the 5S, 16S and 23S subunits (S being the units of Svedberg). The 16S rRNA gene is approximately 1,500 base pairs (bp) long, composed by variable and conserved regions (FIGURE 8). It contains enough polymorphisms to provide differentially measurements between different bacteria species.

1) It is present in all bacteria, becoming this way a universal target for bacterial identification.
2) Its function never changed over time, suggesting than random changes in its sequence are a good accurate measure of time, as they must be random changes, not selected ones due to some function alterations.
3) The gene is large enough (≈1,500bp) for informatics purposes, as it can contain statistically relevant sequence information.
4) Approximately, 16S rRNA contains about 50 functional domains, supposing that an introduction of function-altering mutation won't affect importantly the other domains.

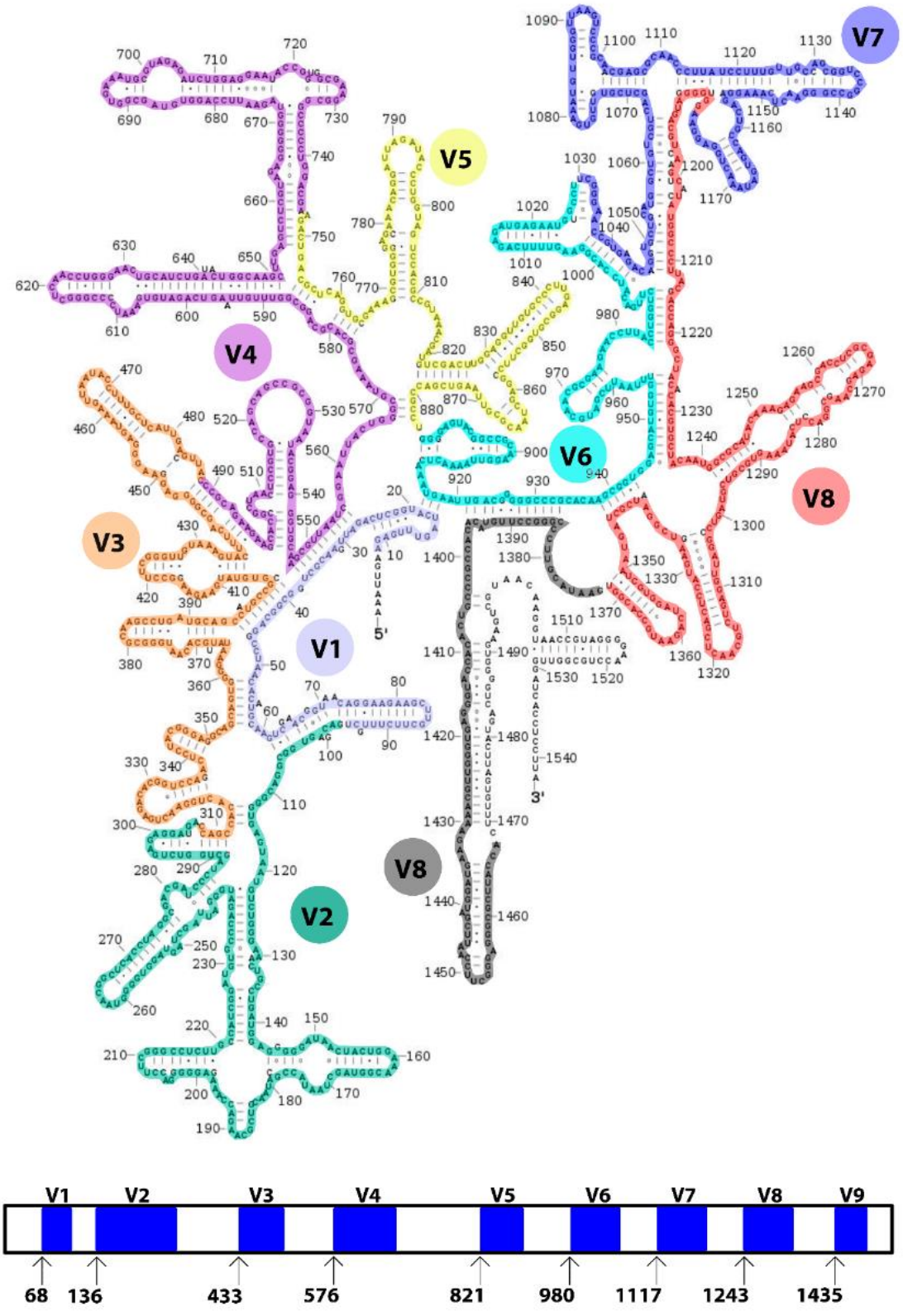


Figure 8: *16S rRNA gene structure, with the different regions colored separately. The base position of each region starting point is indicated in the lower figure.*

Initially, 16S rRNA sequencing was laborious and complex, such that only a small number of laboratories around the world were able to assume its costs [155]. With the invention of the polymerase chain reaction (PCR) in 1986 [156] this situation changed, making it more economic and simple. Depending on the aims of the study, the whole 16S rRNA or just a portion of it can be amplified by PCR and sequenced. This has conducted to the generation of conserved primers targeting conserved regions of the 16S rRNA that allows the amplification for most bacterial species [157, 158]. Interestingly, these universalization of primers allows for the comparison of distinct studies performed by different research groups, as the molecule sequenced is the same [128]. The sequencing of these amplicons and its alignment against referential databases [159] (such as Greengenes [160], SILVA [161] and/or RDP [162]) allows the identification of each bacterial specie present in a sample. This, in turn, supposes also the first pitfall of 16S rRNA amplification studies, as we can only annotate those bacteria already present in a reference database. Another pitfall is that 16S rRNA can be identical between different bacteria from the same genera, even between bacteria from different genera too. This could lead to wrong annotation and identification of bacterial species. To overcome this limitation, functional and biochemical techniques, including bacterial cultures, are needed. Because PCR methodology is quite sensitive to contamination, due to the ubiquitous presence of bacteria in the environment can lead to false assumptions derived from these contaminations. Therefore, to overcome this third pitfall is crucial to incorporate negative controls and blank samples to this kind of taxonomical studies [155]. As 16S rRNA is present only in bacteria and archaea, we lose also the information coming from viruses, protozoa and fungi. It is important to note that they are also a part of the microbiome environment and that can interact with bacteria. Finally, 16S rRNA studied do not comply with Koch's postulates, as they are based just on analyzing the presence of a specific gene, but not in determining if these bacteria are alive and which are their function. Therefore, to determine whether alterations in microbiome are associated to a disease state 16S rRNA sequencing studies are not enough. Relevantly, 16S sequencing provides information about the composition of the microbiome, but not about their genomes and functions [163]. To overcome these issues, whole genome sequencing (WGS) has been developed and applied to microbiome research and Koch's postulates modified accordingly [164].

In order to try to overcome this functionality pitfall from the 16S rRNA sequencing studies, researchers have developed tools that allow the prediction of the functional composition of a specific bacterial community, such as PICRUSt. PICRUSt stands for Phylogenetic Investigation of Communities by Reconstruction of Unobserved States. It takes advantage of data published in the HMP project to, starting from 16S rRNA annotated sequences, predict the whole genome for each one of the bacteria identified, allowing the functional annotation of these predicted bacterial genomes in KEGG database [104]. Its accuracy has been determined to be about 80% although PICRUSt's authors are working on improving it.

1.4.5.-Microbiome data analysis

In brief, a 16S rRNA sequencing study contains four main steps: DNA extraction, PCR amplification, amplicons sequencing and bioinformatics analysis [165].

1) Several protocols for DNA extraction steps exist, most of them being commercially available as extraction kits. What matters most in this step is to consider an extraction protocol appropriated for the sample type (stool, soil, blood, etc.) and to the type of cell lysis needed, as some cell types may resist the common lysis methods. Also relevant, the amount of sample used may be taken into account when selecting the extraction protocol.
2) PCR amplification is, nowadays, performed mostly by sequencing companies already. Therefore, protocols are quite optimized, and the most important consideration here is the selection of 16S primers, which will determine whether the whole 16S is amplified or just some specific regions.
3) Amplicon sequencing is, most often, performed by Next Generation Sequencing (NGS) systems.
4) Finally, the bioinformatics analysis includes everything that happens after the sequencing itself, from the quality control of the reads to the diversity analysis. For these kinds of analyses, several pipelines have been developed, being the most commonly used mothur [124] and QIIME [125]. The following steps have been implemented in, at least, the QIIME tool and can be performed from inside it.
    a. Reads quality control and reads joining (if required). For this step, tools like FLASH [166] and prinseq-lite [167] have been developed.
    b. Removal of chimeras is a necessary step, especially when PCR steps have been performed previously. The most used tools include UCHIME.
    c. Operational Taxonomic Units (OTUs) clustering. After quality controls, reads are clustered into OTUs, which are based on sequence identity (%ID). Depending on the researcher, several thresholds can be used, being the most used one is the 97% similarity threshold. Three types of clustering exist: *de novo*, closed-reference and open-reference.
        i. *De novo* clustering implies that sequences are clustered just by similarity into OTUs, without the use of reference databases.
        ii. Closed-reference OTU picking aligns the reads to a reference database, discarding those sequences that fail to align.
        iii. Open-reference combines both previous approached, starting by close-referencing OTUs and performing *de novo* picking for these sequences that don't align to the reference database.
    d. Diversity analysis. Two kinds of analyses can be performed, α and β diversity analyses [168] In both cases, we have to consider what kind of analysis we want to perform, either qualitative measurements (just considering presence/absence of taxa) or quantitative measurements, if

also taxa abundance is considered. Finally, there is a third consideration to make, whether we consider phylogenetic distance between each pair of taxa studied (divergence-based measures) or we treat each pair of taxa equally (species-based measures).

i. α-diversity measures the diversity between members of a defined population. It allows the comparison of the total intra-diversity between communities. It is used, mostly, to determine whether one community is more or less diverse than other, if it has more or less different bacterial species. Several measures have been developed, both for qualitative (Chao 1 [169], ACE [170], Phylogenetic diversity [171]) and quantitative (Shannon [172], Simpson [173]) approaches.
ii. The β-diversity measure is used to assess the differential distribution of diversity among two or more populations. Different indexes exist, depending on if they are quantitative (weighted) or qualitative (unweighted), including Bray-Curtis, Jaccard, Unifrac, etc. indexes.

e. If the OTU-picking step was performed without the use of a reference database, usually the next step is the taxonomical annotation of the OTUs identified.

Once the OTUs dataset is taxonomically annotated, we can work with either the taxonomical table (at several taxonomic levels) or with the OTUs directly. In both cases, there are two kinds of analyses that can be performed subsequently: population diversity analysis and differential abundances analysis. While the first kind analysis measures how much different and diverse are two or more distinct populations, the second type allows the potential identification of factors that may drive these differences, such a kind of biomarker.

Before being able to perform these analyses, a previous step of normalization is needed. Currently, there is a not a defined, standard protocol for microbiome data analysis steps are to be followed. In fact, the microbiome research community is still debating which kind of data is microbiome data, which kind of statistics should be applied and how normalization of datasets performed. This debate has led to a considerable number of tools developed for microbiome analysis, each one with different assumptions and considerations that may lead to different results and conclusions. While some authors defend that rarefaction approach for dataset normalization is unacceptable because it can suppose the loss of less abundant but still important bacterial features [174], other authors have discussed this affirmation while linking the decision of normalization and rarefaction to the data characteristics and the study aims and assumptions [175]. Other authors have focused on the dataset itself, discussing the traditional approach usually followed of analyzing the normalized OTU counts [176]. They argument, instead, that microbiome datasets should be considered as compositional, so that relative frequencies of bacterial OTUs should be used and different normalization and statistical

approaches applied, because of the limitations introduced by the sequencing step. Comparisons of tools developed by each contenders on this disagreement have been compared, sometimes being more favorable to the non-rarefaction, compositionally-based analysis positions [177] and other times being favorable to the other contender [175].

In any case, after normalization of data, independently of which methodology is selected, traditionally diversity analyses follow. In the case of β-diversity analyses, several indexes exist that measure distinct features, ones incorporating phylogenetic information in the measurements and others just considering global differences. What those indexes do is to measure distances between samples depending on the microbiome composition, so that the final result are distance matrices that are analyzed by multivariate methods [51], including exploratory methods (such as clustering analysis or dimension reduction approaches [49]), interpretative methods and discriminant ones, which have been previously introduced (see section 1.2.2.- Integration methodologies).

1.4.6.- Microbiome impact upon development

Until recently, it was assumed that uterus was a sterile environment and that microbiome was first established after birth [178]. According to this principle, known as *sterile womb paradigm* [179], microbes are acquired both vertically, from the mother, and horizontally, from the environment. However, more recently new research has demonstrated the presence of bacterial genomes upon the uterus, proposing that neither the fetus, placenta nor the amniotic fluid are sterile [180–182]. These postulates have generated controversy on whether the uterus is really sterile or not, with data and opinions going both ways, without reaching a consensus nowadays [183]. In any case, it is known that the host depends highly on the interactions with the microbiome and that the microbiome composition depends highly on the host and the environment [148]. It is known, for example, that microbiome composition is clearly affected, especially during first weeks of life, upon the mode of delivery. Thus, microbiome from babies delivered by C-section resembles the skin microbiome, while vaginally delivered babies' microbiome are dominated by *Lactobacillus, Prevotella* and *Atopobium*, bacteria typically present in the vaginal microbiome community.

Microbiome composition changes during the whole host life (FIGURE 9). Thus, first weeks after birth it starts a diversification process, generating an anaerobe-dominated microbiome [184]. During this time, microbiome is mainly dominated by two phyla, Actinobacteria and Proteobacteria [185]. In this first microbiome community composition, an important modulating factor is the feeding of the babies, existing differences between the own mom's milk, donor's milk and/or formula [145–147]. Microbiome keeps evolving towards a stable composition, with individual-specific temporal patterns [186], that will be maintained during adulthood, only altered by environment factors. Finally, a set of age-related shifts occur in both composition and function of the microbiome. As adult microbiome is mostly composed by bacteria from both Firmicutes and Bacteroidetes phyla, it has been demonstrated that age is

associated to changes on Firmicutes/Bacteroidetes ratio [187]. Notably, elderly people microbiome composition presents a reduction on the capability to perform specific metabolic processes such as SCFA production and amylolysis. This specific microbiome also shows that proteolytic activity is increased [188]. These microbiome metabolic capabilities changes are, in turn, associated to inflammation processes the low grade inflammation event specific of elder individuals [189]. The association between aging, microbiome and specific metabolic processes is, therefore, clear and demonstrated. Seemingly, immune system is developed during the first years of life, as host's microbiome does. Interactions between these two systems will be established, influencing one another's composition [190].

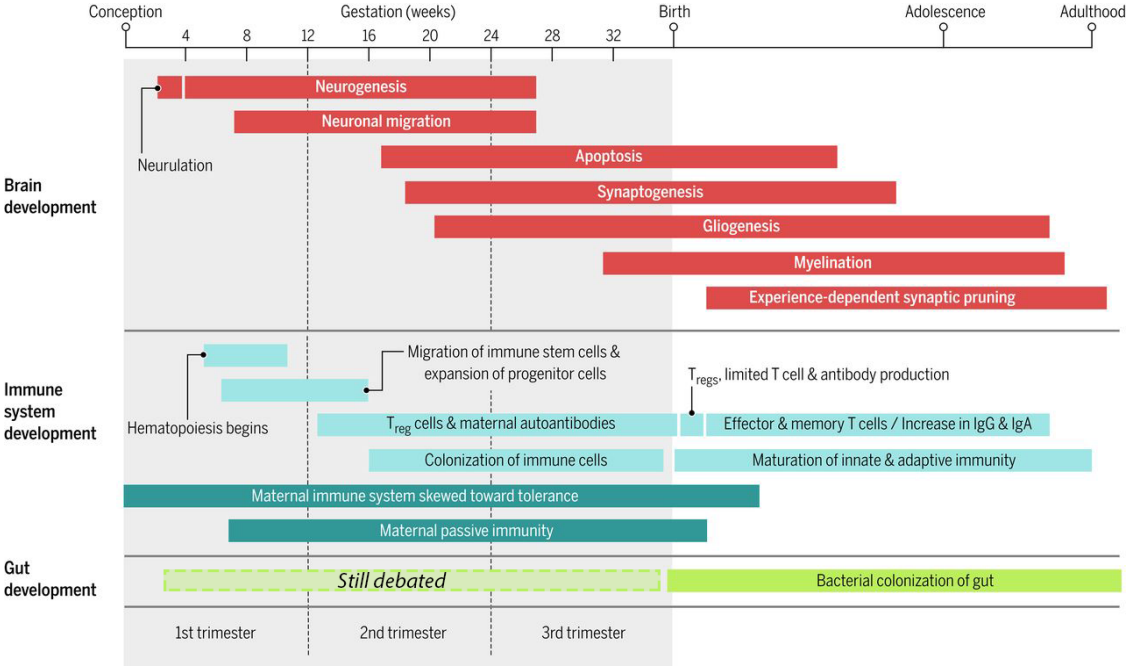


Figure 9: *Timeline of major events occurring in the brain, immune system and gut development from conception to adulthood (adapted from Estes and McAllister, Science (2016)* [191].

1.4.6.1.- Microbiome and immune system

It is accepted that gut microbiota mainly interacts, intensively, with the immune system located in the intestinal mucosa [192]. To this extent, some authors have considered the intestinal mucosa as an immunological niche, an immune-functional organ formed by T cell subpopulation, with the related pro- and anti-inflammatory cytokines, the microbiota and others mediators of inflammation. Innate and adaptive immune systems are believed to play a role in this niche [193]. It's in there were immune and epithelial cells encode receptor molecules for microbial ligands, like capsular polysaccharides (PSA) and lipopolysaccharides (LPS) [194]. Cytokines that are produced during these contacts regulate the differentiation of naïve T cells into regulatory cells (T_{reg}) or a set of helper cells (T_H1, T_H2, T_H17) [195, 196]. Evidence shows that T_H17 and T_{reg} are essential for immune homeostasis and that alteration in their balance can be related to an inadequate microbial gut colonization [197–199] (FIGURE 10).

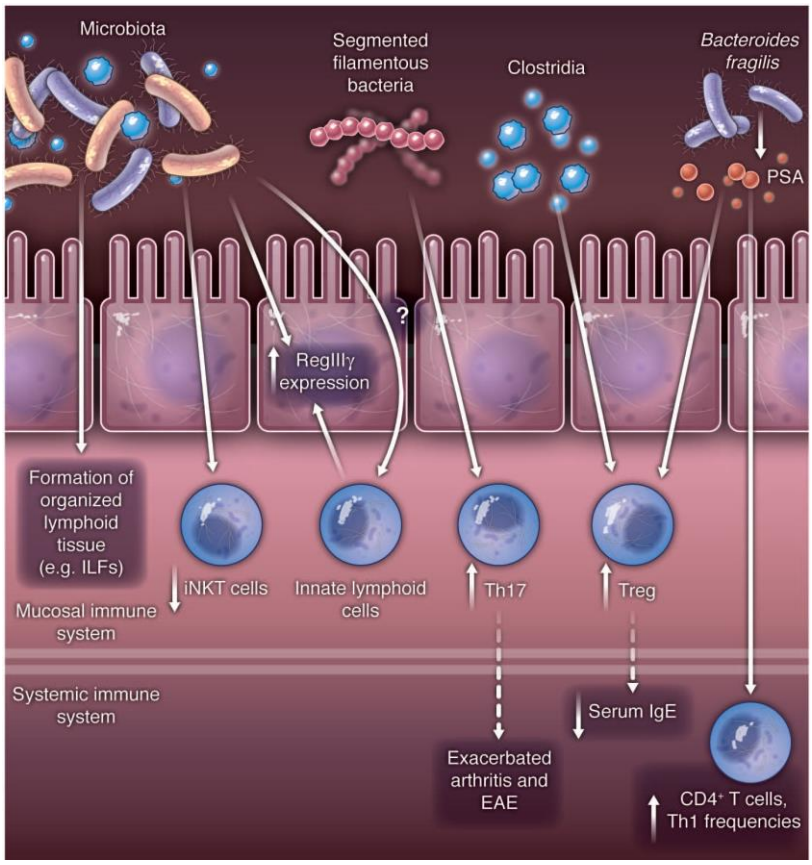


Figure 10: *Summary of the most important manners in which microbiota may shape the host immune system, including mucosal and systemic immunity* [200]**.**

All these effects upon cells and tissues of the immune system have an impact upon its functionality and, especially important, upon the host resistance to pathogen infections [201]. Several mechanisms by which gut microbiota may improve the resistance to pathogens have been described. The two most relevant are the role of microbiota to inhibit the pathogen invasion by competing for the intestinal nutrients [202, 203] and the promoting of the mucosal barrier function, in part with the production of SCFA, which improves the function of the intestinal barrier [150, 151].

1.4.6.2.-The gut-brain axis

The gut-brain axis does not end in the communication during adulthood. Increasing data points towards brain development, function and behavior regulation by gut microbiota [204]. This is due, in part, to the new roles unraveled of molecules related to the peripheral immune system in neurodevelopment, such is the case of cytokines While no clear knowledge exists upon how microbiota influences microglia, it seems that specific bacterial taxa are responsible. GF-associated microglial alterations can be recovered by the administration of SCFA, thus supporting the gut microbiota – microglia association [205].

Notably, microbiota can also modify the integrity of the blood-brain barrier. A reduction of SCFA-producing bacteria supposes an alteration of the tight junction organization, thus increasing the blood-brain barrier permeability [206]. The intrinsic connection gut microbiota – immune system – CNS is clearly demonstrated by the induction of

Experimental Autoimmune Encephalomyelitis (EAE) in GF-mice through the gut colonization by SFB [207], likely via the induction of T_H17, as explained above. This too is supported by the fact that SCFA treatment of these mice reduces EAE and axonal damage by promoting T_{reg} differentiation [208].

Microbiota communicates with the CNS through three main neuroimmune pathways [204] (FIGURE 11):

1) Bacteria-derived and host-derived molecules. The colonization by gut microbiota modulates the host's metabolome, which in turn influences the CNS function by circulating metabolites that can enter the CNS directly affecting neuroactivity [209]. In this particular event, neurotransmitters must be considered, as microbiota can modulate its production and even synthesize them in *de novo* fashion [210]. This microbiota-regulation has been observed for GABA [211], norepinephrine [212], dopamine [212], tryptamine [213] and serotonin, for which is considered that it's mainly produced in the gastrointestinal tract [214]. At this point, is quite relevant to mention the role of GABA as a pain inhibitor [215] and appetite regulator [216].

2) Neuronal signaling. Gut microbiota can also interact with the CNS through the vagus nerve, a nerve that innervates the peripheral organs, such as the gastrointestinal tract. It allows the communication between the CNS and the peripheral organs [217].

3) Neuroendocrine pathways. Through this pathway brain communicates to the gut microbiota, reflecting biochemical changes in the brain to the intestine, modifying its physiology. This kind of communication occurs mainly through the hypothalamic-pituitary-adrenal (HPA) axis [217].

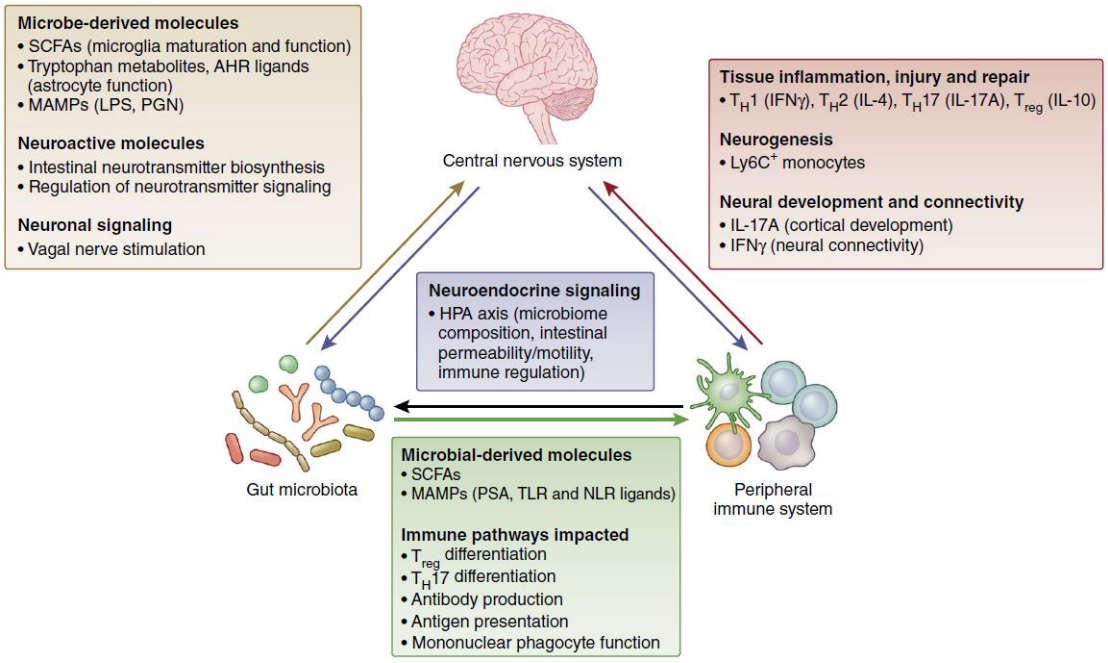


Figure 11: *Summary of the crosswalk pathways between the gut microbiota, the CNS and the immune system* [204].

1.4.6.3.- Gut microbiome and colorectal cancer

Cancer has been associated with a broad of genetic and environmental factors that, in turn, modulates also the gut microbiota. To this extent, it has been calculated that ~20% of human tumors are associated with microorganisms [218]. Notably, the role of bacteria and viruses in cancer development and progression has been known for decades [219], being these especially true for the relationship between cervical cancer and Human Papilloma Virus (HPV) and gastric cancer and *Helicobacter pylori* [220, 221]. This relation has also been proved to be true for colorectal cancer (CRC).

Colorectal cancer is the fourth most prevalent cancer worldwide [222], with more than 600,000 deaths per year. Many of the risk factors are associated with developed countries' lifestyles, which correlates with the global distribution of CRC, being more prevalent in developed countries than in non-developed ones. Its progression follows is described as the “adenoma (AD)-carcinoma sequence” [223] (Fearon and Vogelstein), which stats that accumulated genetic and epigenetic mutations drive epithelial hyperplasia in the colon, which results in CRC. Recently, a new factor has been proposed to act in this AD-CRC transition, the microbiota by the hypothesis known as driver-passenger bacteria. This hypothesis proposes that indigenous bacteria (**driver bacteria**) drive the DNA damage process, accumulating mutations in genes such as APC, CTNNB1, DCC, P53, KRAS and/or MYC. This accumulation of DNA damage will eventually lead to CRC development. This tumorigenesis will modify the environment, favoring the establishment of opportunistic bacteria, the **passenger bacteria**. Passenger bacteria outgrows the driver ones, as they are better fitted to the new tumoral environment and driver bacteria will disappear from the tumoral tissue.

Thesis objectives

Do. Or not do. There is no try.

The Empire Strikes Back, 1980.

Nowadays several high-throughput omics technologies have demonstrated to be a valuable tool to deeply characterize and understand biological processes, what explains their broad application for new biomarkers discovery. In order to increase the efficacy of these technologies, their integration need to be improved it. Thus, in this thesis, we pretended to analyze and choose a standardized pipeline for metabolomics – microbiome integration using current methodologies. The work has been divided in three practical cases, each of them with their specific objectives what guide the bioinformatics approaches. The first project (see results Chapter 2) was focused on early prostate cancer (PCa) biomarkers identification by metabolomics analysis of urinary EVs and its reflection of tissue tumoral metabolism. The second project (results Chapter 4) was a multi-omics study of fibromyalgia disease to identify potential molecular alterations and propose biomarkers, focusing on metabolomics-microbiome integration. Finally, the last practical case included in this thesis (results Chapters 5 and 6) was the integration of metabolomics-microbiome data for the identification of fecal biomarkers for colorectal cancer (CRC) and advanced adenocarcinoma (AD).

The general objectives for this thesis were:

i) To identify, review and summarize the currently available omics-integration tools.
ii) To establish a data analysis pipeline that complies with minimal quality criteria for each omics included in the studies.
iii) To identify and develop bioinformatics tools able to automatically retrieve data from the most common metabolite databases.
iv) To set up and establish a pipeline to analyze and integrate metabolomics and microbiome data.

Specific objectives for the PCa metabolomics project:

Urinary EVs should reflect the biological status of prostate cells, therefore, their content should be useful to discriminate between Benign Prostatic Hyperplasia (BPH) and PCa patients. To test this hypothesis, we proposed the following objectives for the study:

i) To characterize the metabolome of EVs from BPH and PCa patients.
ii) Identify differentially expressed metabolites between these two groups of patients.
iii) To identify differentially expressed metabolites between subgroups of PCa patients (stage 2 vs stage 3 and stage 2 with perineural invasion vs stage 2 without perineural invasion).
iv) To integrate gene expression data obtained from publicly available databases and metabolomics data in order to identify differentially expressed genes.

Specific objectives for the fibromyalgia multi-omics study:

The main objective was the identification of molecular markers that could be useful as diagnostics criteria for fibromyalgia disease, improving the nowadays diagnostics options.

i) To process, analyze and identify potential bacteria strains from microbiome data able differentially present between fibromyalgia patients and control individuals.
ii) To process, analyze and identify potential metabolites differentially present between fibromyalgia patients and healthy individuals.
iii) To analyze data from cytokines and miRNA analyses in order to support the findings in microbiome and metabolomics analyses.
iv) To integrate different types of data (microbiome, metabolomics, cytokines, miRNA) in order to explain potential biological alterations that could explain fibromyalgia pathogenesis.
v) To validate alterations identified by bioinformatics data analysis with experimental methodologies.

Specific objectives for the CRC microbiome-metabolomics integration project:

The first aim and the most important one was to identify changes in both metabolites and microbiota composition in the same fecal sample of CRC and AD patients when compared to healthy controls.

The identified changes were considered as potential biomarkers for CRC and AD. The biological context for the observed alterations was provided by the integration of both data omics. In a more specific way, the objectives of this study were:

i) To characterize the metabolome and metagenome associated with healthy individuals, CRC and advanced AD patients.
ii) To detect and identify differentially expressed metabolites between the three population groups included in the study.
iii) To detect the bacteria species differentially expressed between the three population groups included in the study.
iv) To integrate clinical parameters, metabolomics and metagenomics data in order to identify and describe biological processes related to CRC pathogenesis and progression.
v) To identify and characterize new non-invasive metabolomics-based biomarkers capable of discriminating between healthy individuals and those with AD and/or CRC.
vi) To develop a new diagnostics tool better than the actual ones by statistics methodologies, generating predictive regression models capable of discriminating between healthy, AD and CRC diagnosis better than the actual gold standard.

Results

Never tell me the odds!

The Empire Strikes Back, 1980.

3.1.- Chapter 1. Bioinformatics and data analysis considerations

In the Introduction section, we have already analyzed which are the most challenging aspects of multi-omics integration, specifically from the bioinformatics point of view. In this chapter, I will present which bioinformatics approaches have been used and developed to tackle them, including approaches for individuals omics (metabolomics and microbiome) and for the combination of them.

In order to avoid contributing to the already overcrowded stock of analytical tools, we limited the development of new bioinformatics tools and analytical pipelines to when it was strictly required, favoring the test of several already developed and published tools and pipelines in order to avoid as much as possible contributing to an already overgrowth market of analytical tools, in order to favor the identification of this gold standards.

3.1.1.- Metabolomics

While data cleaning, processing, and analysis did not suppose a relevant challenge because the pipelines are quite established, functional annotation and profiling of metabolomics results needed more work.

Because metabolomics is included in each one of the projects presented in this thesis, we decided to establish a common pipeline for data processing and analysis, with common statistics measurements. The metabolomics data was generated by two ways: targeted metabolomics that was performed in collaboration with OWL Metabolomics and untargeted that was performed in bioGUNE's Metabolomics Platform. Because of the targeted metabolomics characteristics, no peak identification step was needed for these projects. For untargeted metabolomics, though, a first step of potential metabolite annotation and chemical standards validation was needed. To facilitate this step, we decided to focus only on those metabolomics peaks that were differentially expressed between sample groups included in each study.

Either if the data was already processed (with metabolites identified) or not, the first step was to clean the dataset. To this, we identified and removed all the features that presented with more than 30% of missing values, either in all sample groups or in one of them. Only if in one group that features presented less than 30% of missing values and in the other one less than 70% of them we kept the metabolite, considering that this fact could be associated with the biological phenomenon we were studying. Once features with an excessive number of missing values were removed, we proceeded to impute the remaining missing values. To this, of the distinct methods we tested (1/10 minimal value, median, mean and *k*-means nearest neighbor), we decided to use the minimal value of the feature divided by 10, so that the minimal alteration was included in the analysis.

Afterwards, the data was log-normalized and applied custom scripts dedicated to compute the following basic statistics for each sample group in a standard way: mean, winsored mean, median, standard error of the mean (SEM), standard deviation,

coefficient of variance, interquartile range, Kurtosis and Skewness indexes and Shapiro test. Using these measurements fold change and either Student's t-test and/or Wilcoxon signed-rank p-values are computed and the volcano plot is generated to aid in the data interpretation in a visual way.

Once differential metabolites were identified, functional annotation and pathway mapping followed. Is in this step on which more work was performed. Using two publicly available metabolomics databases, KEGG [104] and HMDB [101], we wrote both R and Python scripts that access those databases and extract the required information.

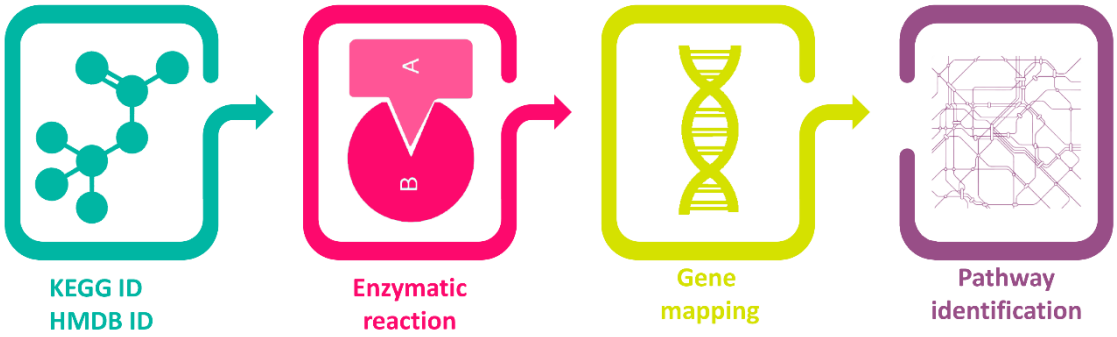


Figure 12: *Developed bioinformatics pipeline. Starting from metabolite ID related enzymes are identified, associated genes retrieved and pathways enrichment performed.*

Because metabolites receive a broad range of names, we decided to focus on the accession codes for both datasets. This allowed us to work with no-ambiguous metabolite identifications and to be able to combine the information on both databases, by translating each accession code to the corresponding one on the other database. Because of the characteristics of each database, complementary information may be obtained this way. Because metabolites are the end-product of biological reactions, our idea was to associate each metabolite to the corresponding enzymes, thus this way we could associate metabolite levels to genes and transcriptomics data. This conversion of metabolites to gene names is what allowed us to perform pathway enrichment analysis by means of tools like PANTHER-db or DAVID functional annotation tool.

KEGG database is structured in several sub-databases related to genomes, biological pathways, diseases, drugs and chemical compounds. The combination of all these data allows the integration of multiple omics into metabolic pathways, being thus a tool for systems biology [224].

Table 2: *The compendium of KEGG databases, with the corresponding contents, updated in 2017 (modified from* [225]).

Category	Database name	Content	KEGG identifier
Systems Information	KEGG PATHWAY	KEGG pathway maps	Map number
	KEGG BRITE	BRITE hierarchies and tables	br/ko number
	KEGG MODULE	KEGG modules	M number
Genomic Information	KEGG ORTHOLOGY	KO groups for functional orthologs	K number
	KEGG GENOME	KEGG organisms (complete genomes) and selected viruses	org code/T number
	KEGG GENES	Gene catalogs ofKEGG organisms, viruses, and addendum category	org:gene
	KEGG SSDB	KEGG SSDB Sequence similarity among GENES entries (computationally generated)	
Chemical Information (KEGG LIGAND)	KEGG COMPOUND	Metabolites and other small molecules	C number
	KEGG GLYCAN	Glycans	G number
	KEGG REACTION	Biochemical reactions	R number
	KEGG RCLASS	Reaction class	RC number
	KEGG ENZYME	Enzyme nomenclature	EC number
Health Information (KEGG MEDICUS)	KEGG DISEASE	Human diseases	H number
	KEGG DRUG	Drugs	D number
	KEGG DGROUP	Drug groups	DG number
	KEGG ENVIRON	Crude drugs and health-related substances	E number

Because of KEGG's architecture, by using metabolite codes we developed an R package dedicated to retrieving general information, enzymes, genes and pathways related to the corresponding metabolite in an automatized fashion.

HMDB database is a multi-purpose bioinformatics database focused on quantitative, analytical and functional annotation of human metabolites [77]. Entries in the HMDB database are structured in the so-called "MetaboCards", each one of them containing more than 90 data fields, including physicochemical and biochemical features of the metabolites, biological and biomedical data and cross-references with other databases [77, 101]. Some of the most relevant entries of MetaboCards are summarized in Table 19.

Our approach with HMDB database included the same data as KEGG's database step and, due to the availability of more data, other important information.

Table 3: *Summary of the MetaboCards included information for HMDB database, adapted from* [77].

Metabolite and medical information	Protein/enzyme information
Common name	Enzyme/protein name
Description	Enzyme/protein synonyms
Synonyms/IUPAC name	Enzyme/protein sequence
Chemical structure	Protein number of residues
Chemical taxonomy	Protein molecular weight
Molecular weight (mono and ave)	Protein pI
SMILES (isomeric and canonical)	Protein gene ontology
KEGG/PubChem/OMIM/MetaGene links	Protein general function
CAS number	Protein specific function
InChi identifier	Protein pathways
Melting point	Protein reactions
Water solubility (predicted and expected)	Protein Pfam domains
State (solid, liquid, gas)	Protein signal sites
pKa or pI	Protein transmembrane regions
LogP or hydrophobicity	Protein metabolic importance
MOL/SDF/PDF text files	Protein/enzyme EC link
MOL/PDB image files	GenBank, SwissProt, PDB ID
NMR spectra (predicted, calculated)	Protein structure data
Location (cell, biofluid, tissue)	Protein cellular location
Concentration (urine, plasma, CSF)	Gene sequence
Associated disorders	GenBank ID
Abnormal concentration (urine, plasma, CSF)	Chromosome location
Metabolic pathways (KEGG, SimCell)	Chromosome locus
Metabolizing enzymes	Protein/enzyme SNP/mutations
	Protein/enzyme references

Combining both approaches, we developed an R package dedicated to the automatic data retrieval of both KEGG and HMDB databases that we called DatR. It had also extra functionalities connected to other databases such as OMIM, Uniprot and/or PubMed. In order to be able to use the full functionalities of the package, the following dependencies were used: *KEGGREST* [226], *XML* [227], *stringi* [228], *rentrez* [229], *GEOquery* [230] and *RISmed* [231]. All the functions developed within this R package were accessible by standard R format. R package was uploaded to GitHub repository (https://github.com/pxtm/DatR).

3.1.1.1.- KEGG database access

KEGG compounds ID follow the following structure: a capital C followed by a combination of 5 numbers, for example, phosphatidylcholine is identified as C00157 in KEGG's database. Because the ID code is unique to each metabolite, we designed the KEGG data retrieval functions to work with these IDs.

In R, with the metabolites identified by their KEGG ID code, we designed mainly three functions, one for metabolite associated enzymes code (EC) retrieval, one for the conversion of these enzymes to their corresponding genes and a final one to map the metabolites to the metabolic pathways in which they are represented.

1.- Metabolites to enzymes

From the KEGG ID metabolite code, the function uses the REST KEGG portal and returns the enzymes that are associated to a specific metabolite. Enzymes are identified through the corresponding EC code, the standard form. The code for the function can be found below:

```
kenz <- function(x){
 KEGG_LINK_BASE <- "http://rest.kegg.jp/link/enzyme/"
 link_REST_url <- paste(KEGG_LINK_BASE, x, sep="")
 link <- readLines(link_REST_url)
 enzs<-sapply(link, function(x){
 strsplit(x, "\t")[[1]][2]
 })
 enzs<-unname(enzs)
 enzs<-sapply(enzs, function(x){
 strsplit(x, "ec:")[[1]][2]
 })
 enzs<-unname(enzs)
 enzs<-enzs[!is.na(enzs)]
}
```

2.- Metabolites to genes

Using a similar approach, this time the function retrieves the genes present in the specific metabolite entry in KEGG. Because of the KEGG's structure, in order to obtain genes from a specific metabolite, we need to retrieve first the enzymes associated with the metabolite and then cross-link them to the corresponding genes database.

```
kgenes <- function(met){
  KEGG_LINK_BASE <- "http://rest.kegg.jp/link/enzyme/"
  link_REST_url <- paste(KEGG_LINK_BASE, met, sep="")
  link <- readLines(link_REST_url)
  enzs<-sapply(link, function(x){
    strsplit(x, "\t")[[1]][2]
  })
  enzs<-unname(enzs)
  enzs<-sapply(enzs, function(y){
    strsplit(y, "ec:")[[1]][2]
  })
  enzs<-unname(enzs)
  enzs<-enzs[!is.na(enzs)]
  kegg_link_gene<-"http://rest.kegg.jp/link/hsa/enzyme"
  total_genes<-readLines(kegg_link_gene)
  genes<-sapply(enzs, function(c){
    total_genes[grep(c, total_genes)]})
  genes<-unlist(sapply(genes, function(z){
    unname(unlist(sapply(z, function(d){
      sapply(d, function(f){strsplit(f, "\t")[[1]][2]})}))))}))
  genes<-unname(unlist(sapply(genes, function(z){
    unname(unlist(sapply(z, function(d){
      sapply(d, function(f){strsplit(f, "hsa:")[[1]][2]})}))))}))
  require(rentrez)
  gene_names<-sapply(genes, function(m){
    sapply(m, function(n){
      entrez_summary('gene', n)
    })
  })
  gene_names<-as.data.frame(gene_names)
```

```
  gene_names<-(unname(unlist(gene_names[2,])))
  return(gene_names)
}
```

2.1.- Enzymes to genes

In case we were starting from the enzymes obtained results, we also included a simplified version of the above function that retrieved the metabolite associated genes but starting from the enzymes list.

```
kenzgen <- function(x){
  kegg_link_gene<-"http://rest.kegg.jp/link/hsa/enzyme"
  total_genes<-readLines(kegg_link_gene)
  genes<-sapply(x, function(c){
    total_genes[grep(c, total_genes)]})
  genes<-unlist(sapply(genes, function(z){
    unname(unlist(sapply(z, function(d){
      sapply(d, function(f){strsplit(f, "\t")[[1]][2]})})))}))
  genes<-unname(unlist(sapply(genes, function(z){
    unname(unlist(sapply(z, function(d){
      sapply(d, function(f){strsplit(f, "hsa:")[[1]][2]})})))})))
  gene_names<-sapply(genes, function(m){
    sapply(m, function(n){
      entrez_summary('gene', n)
    })
  })
  gene_names<-as.data.frame(gene_names)
  gene_names<-(unname(unlist(gene_names[2,])))
}
```

3.- Metabolites mapping to pathways

We also developed a function that maps metabolites to human metabolic pathways and return the code and name of these pathways.

```
kpaths <- function(x){
  temp<-readLines(paste0('http://rest.kegg.jp/link/pathway/', x))
  temp_paths<-sapply(temp, function(x){
    strsplit(x, 'path:')[[1]][2]
  })
  paths<-sapply(temp_paths,
function(y){readLines(paste0('http://rest.kegg.jp/find/pathway/',
y))})
  paths<-sapply(paths, function(z){strsplit(z, '\t')[[1]][2]})
  return(unname(paths))
}
```

4.- Summary of all KEGG data

Finally, in order to facilitate the KEGG's data retrieval step, we wrote a function that wraps up all the above, returning a final list with enzymes, genes and pathways associated to the metabolites.

```
ksumm <- function(x){
  df1<-list(
    list(
      Enzymes=kenz(x),
      Genes=kgenes(x),
      Paths=kpaths(x)))
  return(df1)
}
```

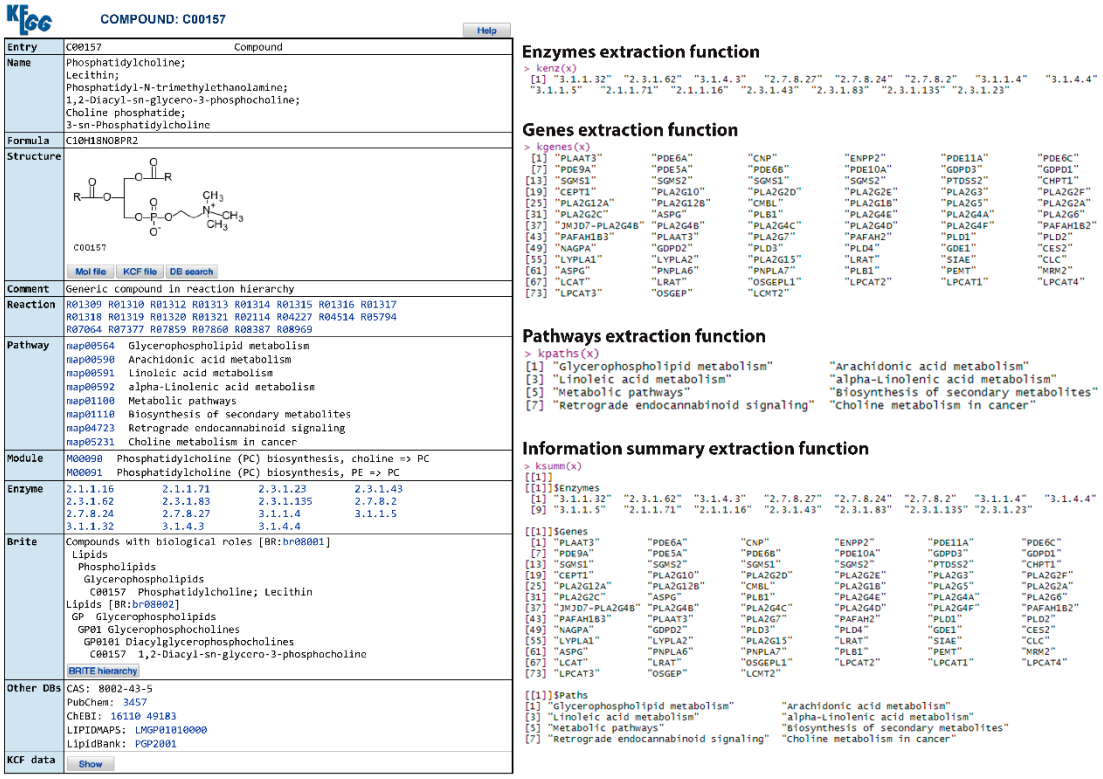


Figure 13: *Comparison of the C00157 metabolite entry in KEGG database (left) and the results obtained using the functions developed and run in the R environment (right) for the same metabolite.*

3.1.1.2.- HMDB database access

For accessing HMDB database, we considered two options. One was accessing each MetaboCard individually, in a sequential form, which was time-consuming. The other was to download the full database and upload it to our own servers or the RStudio environment, which we discarded because of the lack of resources to do that.

1.- **Summary of HMDB data**

This function was designed to retrieve all the available information of interest for our projects: accession code, metabolite name, chemical and SMILEs formulas, related pathways (including metabolic pathways and diseases), the KEGG ID to cross-reference databases and the protein associations (protein accession, name, UniProt ID, gene name and protein type).

```
hsumm <- function (x) {
 temp <- xmlToList(xmlParse(paste0("http://www.hmdb.ca/metabolites/",
                                   x, ".xml")))
 Accession = temp$accession
 Name = temp$name
 Formula = temp$chemical_formul
 Smiles = temp$smile
 Genes <- vector("list", length = length(temp$protein_associations))
 for (i in 1:length(temp$protein_associations)) {
   Genes[i] <- temp$protein_associations[i]$protein$gene_name
 }
 KEGG_ID = temp$kegg_id
```

```
  Paths <- vector("list", length =
   length(temp$biological_properties$pathways))
  if (length(temp$biological_properties$pathways) == 1) {
    Paths = temp$biological_properties$pathways$pathway$name
  }
  else {
    for (a in 1:length(temp$biological_properties$pathways)) {
      Paths[a] <- temp$biological_properties$pathways[a]$pathway$name
    }
  }
  hmdb_data <- list(Acces = Accession, Name = Name, Formula = Formula,
                    SMILES = Smiles, Genes = unlist(Genes), KEGG =
KEGG_ID, Paths = unlist(Paths))
}
```

2.- Databases cross-linking

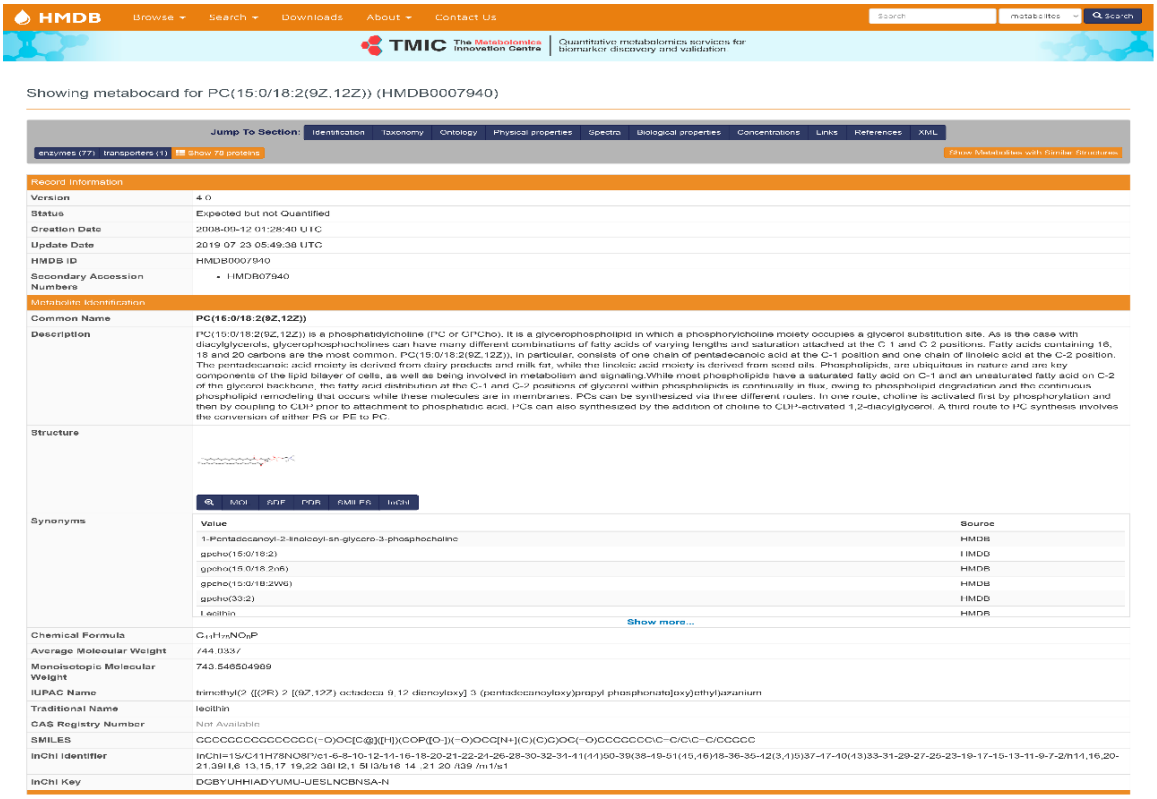
In order to combine HMDB and KEGG databases, we needed one function that could convert between HMDB IDs and KEGG ones. Thus, we developed this function that retrieves the KEGG accession code from an HMDB metabolite entry.

```
hkeggid <- function(x){
  temp<-xmlToList(xmlParse(paste0("http://www.hmdb.ca/metabolites/",
  x,  ".xml")))
  keggid<-temp$kegg_id
  return(keggid)
}
```

3.- Genes retrieval from HMDB metabolite

This was a variant of the first HMDB database related function designed to obtain the gene name of the proteins related to each metabolite so that less time was consumed in comparison with the full function. Because of the specifics of the format of the data, it's a combination of two different functions, one that retrieves the data and the other one that formats it into a readable and easy-exportable way.

```
hgenes <- function(x){
  temp<-xmlToList(xmlParse(paste0("http://www.hmdb.ca/metabolites/",
    x,  ".xml")))
  Genes<-vector('list', length = length(temp$protein_associations))
  for (i in 1:length(temp$protein_associations)){
  Genes[i]<-temp$protein_associations[i]$protein$gene_name
  }
  return(unlist(Genes))
}
```



HMDB summary information function

```
> hsumm (y)
$Acces
[1] "HMDB0007940"

$Name
[1] "PC(15:0/18:2(9Z,12Z))"

$Formula
[1] "C41H78NO8P"

$SMILES
[1] "CCCCCCCCCCCCCCC(=O)OC[C@](COP([O-])(=O)OCC[N+](C)(C)C)OC(=O)CCCCCCC\\C=C/C\\C=C/CCCCC"

$Genes
 [1] "LYPLA1"   "PLA2G15"  "PLA2G5"   "PLA2G2F"  "PLA2G4A"  "PLA2G1B"  "PLA2G12B" "PLA2G10"  "PLA2G2E"  "LYPLA2"   "PLA2G12A"
[12] "PLA2G6"   "LCAT"     "CLC"      "PLA2G2A"  "PLA2G2D"  "PLD2"     "PEMT"     "PLD1"     "PLA2G4C"  "PLA2G3"   "BDH1"
[23] "PCYT1B"   "PCYT1A"   "SGMS2"    "SGMS1"    "CHKA"     "ATP11C"   "ATP11A"   "ATP10A"   "ATP8B1"   "ATP9A"    "ATP10D"
[34] "ATP8A2"   "ATP8A1"   "ATP8B4"   "ATP11B"   "ATP8B3"   "PITPNB"   "ABCB4"    "LRAT"     "PLSCR1"   "PTDSS1"   "PITPNA"
[45] "PLTP"     "CHPT1"    "MOGAT2"   "CEPT1"    "PLCD3"    "LPCAT3"   "PLA2G2C"  "LPCAT1"   "LPCAT2"   "PLD3"     "PLD4"
[56] "PCTP"     "PLSCR2"   "PLSCR3"   "PLSCR4"   "PLSCR5"   "ATP10B"   "ATP8B2"   "PLB1"     "PEBP1"    "PEBP4"    "PITPNM1"
[67] "PITPNM2"  "PITPNM3"  "PLA2G4D"  "PLA2G4E"  "PLA2G4F"  "ARFGAP1"  "PLA2G4B"  "PLD6"     "PNPLA6"   "APOA5"    "SCGB1A1"

$KEGG
[1] "C00157"

$Paths
[1] "Fabry disease"
[2] "Gaucher Disease"
[3] "Globoid Cell Leukodystrophy"
[4] "Krabbe disease"
[5] "Metachromatic Leukodystrophy (MLD)"
[6] "Phosphatidylcholine Biosynthesis PC(15:0/18:2(9Z,12Z))"
[7] "Phosphatidylethanolamine Biosynthesis PE(15:0/18:2(9Z,12Z))"
[8] "Sphingolipid Metabolism"
```

Genes and KEGG ID retrieval function

```
> hgenes(y)
 [1] "LYPLA1"   "PLA2G15"  "PLA2G5"   "PLA2G2F"  "PLA2G4A"  "PLA2G1B"  "PLA2G12B" "PLA2G10"  "PLA2G2E"  "LYPLA2"   "PLA2G12A"
[12] "PLA2G6"   "LCAT"     "CLC"      "PLA2G2A"  "PLA2G2D"  "PLD2"     "PEMT"     "PLD1"     "PLA2G4C"  "PLA2G3"   "BDH1"
[23] "PCYT1B"   "PCYT1A"   "SGMS2"    "SGMS1"    "CHKA"     "ATP11C"   "ATP11A"   "ATP10A"   "ATP8B1"   "ATP9A"    "ATP10D"
[34] "ATP8A2"   "ATP8A1"   "ATP8B4"   "ATP11B"   "ATP8B3"   "PITPNB"   "ABCB4"    "LRAT"     "PLSCR1"   "PTDSS1"   "PITPNA"
[45] "PLTP"     "CHPT1"    "MOGAT2"   "CEPT1"    "PLCD3"    "LPCAT3"   "PLA2G2C"  "LPCAT1"   "LPCAT2"   "PLD3"     "PLD4"
[56] "PCTP"     "PLSCR2"   "PLSCR3"   "PLSCR4"   "PLSCR5"   "ATP10B"   "ATP8B2"   "PLB1"     "PEBP1"    "PEBP4"    "PITPNM1"
[67] "PITPNM2"  "PITPNM3"  "PLA2G4D"  "PLA2G4E"  "PLA2G4F"  "ARFGAP1"  "PLA2G4B"  "PLD6"     "PNPLA6"   "APOA5"    "SCGB1A1"
> hkeggid(y)
[1] "C00157"
>
```

Figure 14: *Comparison of the HMDB0007940 metabolite entry in HMDB database (upper panel) and the results obtained using the functions developed and run in the R environment (lower panel) for the same metabolite. HMDB MetaboCard has been reduced for easier visualization purposes.*

3.1.1.3.- Other functionalities

Apart from the database access functions described, we also included other options to aid in the functional profiling step.

1.- **Accession codes and metabolite names**

Because our center has strong collaborations with OWL Metabolomics Company and they use their personal codes for metabolites, we developed a function that allows us to retrieve the standard name, KEGG and HMDB codes for a specific metabolite. With slight modifications, this function also allowed us to easily convert KEGG IDs to HMDB ones and vice versa.

```
owl_code <- function(x){
  data.frame(
    Name=owl[grep(paste0('^', x, '$'), owl$OWL.Code), 3],
    Alternative=owl[grep(paste0('^', x, '$'), owl$OWL.Code), 4],
    HMDB=owl[grep(paste0('^', x, '$'), owl$OWL.Code), 6],
    KEGG=owl[grep(paste0('^', x, '$'), owl$OWL.Code), 7]
  , stringsAsFactors = F)
}
```

2.- **Gene names**

Genes may be identified by a range of synonyms. This implies that sometimes discrepancies exist between the gene names obtained from KEGG or HMDB pipelines and the gene name used as an identifier in other databases such as the Entrez collection. Thus, we wrote a function that returns all the possible synonym names for a specific gene.

```
galias <- function(x){
  p1<-entrez_search(db='gene', term=paste0(x, '[Gene Name] AND Homo
sapiens[Organism]'))$ids
  if (length(p1)==0)
  {}
  else if (length(p1)==1){
    p2<-entrez_summary(db='gene', id=p1)$otheraliases
  }
  else {
    p2<-sapply(p1, function(z){
      entrez_summary(db='gene', id=z)$otheraliases
    })
  }
  if (isTRUE(p2=="")==TRUE)
  {}
  else {
    return(unlist(strsplit(as.character(p2), ', ')))
  }
}
```

```
> sapply(genes, gallas)
$PLAAT3
 [1] "AdPLA"      "H-REV107"   "H-REV107-1" "HRASLS3"    "HREV107"    "HREV107-1"  "HREV107-3"  "HRSL3"      "PLA2G16"
[10] "PLAAT-3"

$PDE6A
[1] "CGPR-A" "PDEA"   "RP43"

$CNP
[1] "CNP"  "CNP2" "CNP1"

$ENPP2
[1] "ATX"       "ATX-X"     "AUTOTAXIN" "LysoPLD"   "NPP2"      "PD-IALPHA" "PDNP2"

$PDE11A
[1] "PPNAD2"

$PDE6C
[1] "ACHM5" "COD4"  "PDEA2"

$PDE9A
[1] "HSPDE9A2"

$PDE5A
[1] "CGB-PDE" "CN5A"    "PDE5"

$PDE6B
[1] "CSNB3"       "CSNBAD2"     "GMP-PDEbeta" "PDEB"        "RP40"        "rd1"

$PDE10A
[1] "ADSD2"     "HSPDE10A"  "IOLOD"     "LINC00473" "PDE10A19"

$GDPD3
[1] "GDE7"

$GDPD1
[1] "GDE4"

$SGMS1
[1] "MOB"    "MOB1"   "SMS1"   "TMEM23" "hmob33"

$SGMS2
[1] "CDL"  "SMS2"

$SGMS1
[1] "MOB"    "MOB1"   "SMS1"   "TMEM23" "hmob33"

$SGMS2
[1] "CDL"  "SMS2"

$PTDSS2
[1] "PSS2"
```

Figure 15: *Screenshot of the results of the gene aliases function retrieval, using the list of genes obtained from the C00157 KEGG metabolite entry.*

3.- OMIM database access

Online Mendelian Inheritance in Man (OMIM) is the NCBI database that includes information regarding the associations between genes and diseases [232]. Because this bioinformatics tool was designed to retrieve information that could facilitate unraveling the biological context of specific metabolite alterations, we considered this addition to be useful for this aim. Using the list of the genes obtained from any of the aforementioned functions, this function returned a list of all the diseases that have been reported to be associated with each gene.

```
omim <- function(x){
  omim_id<-entrez_search("omim", x)$ids
  if (length(omim_id)==0){
    summary_omim<-c('No entries found for this gene')
  }
  else {
    summary_omim<-sapply(omim_id, FUN=function(x){
      entrez_summary("omim", x)$title
    })
  }
  return(unname(summary_omim))
}
```

```
> sapply(genes, omim)
$PLAAT3
[1] "No entries found for this gene"

$PDE6A
[1] "NEPHRONOPHTHISIS 20; NPHP20"                          "RETINITIS PIGMENTOSA 43; RP43"
[3] "RETINITIS PIGMENTOSA 57; RP57"                        "MACULAR DYSTROPHY, PATTERNED, 2; MDPT2"
[5] "RETINITIS PIGMENTOSA; RP"                             "PHOSPHODIESTERASE 6G, CGMP-SPECIFIC, ROD, GAMMA; PDE6G"
[7] "PHOSPHODIESTERASE 6B, CGMP-SPECIFIC, ROD, BETA; PDE6B"  "PHOSPHODIESTERASE 6A, CGMP-SPECIFIC, ROD, ALPHA; PDE6A"

$CNP
 [1] "SHORT STATURE WITH NONSPECIFIC SKELETAL ABNORMALITIES; SNSK"
 [2] "EPIPHYSEAL CHONDRODYSPLASIA, MIURA TYPE; ECDM"
 [3] "QUAKING, MOUSE, HOMOLOG OF; QKI"
 [4] "TSC22 DOMAIN FAMILY, MEMBER 1; TSC22D1"
 [5] "OLIGODENDROCYTE LINEAGE TRANSCRIPTION FACTOR 2; OLIG2"
 [6] "ACROMESOMELIC DYSPLASIA, MAROTEAUX TYPE; AMDM"
 [7] "NATRIURETIC PEPTIDE PRECURSOR C; NPPC"
 [8] "NATRIURETIC PEPTIDE PRECURSOR B; NPPB"
 [9] "FIBROBLAST GROWTH FACTOR RECEPTOR 3; FGFR3"
[10] "CYCLIC NUCLEOTIDE PHOSPHODIESTERASE; CNP"
[11] "NATRIURETIC PEPTIDE RECEPTOR C; NPR3"
[12] "NATRIURETIC PEPTIDE RECEPTOR 2; NPR2"
[13] "ACHONDROPLASIA; ACH"

$ENPP2
[1] "ECTONUCLEOTIDE PYROPHOSPHATASE/PHOSPHODIESTERASE 3; ENPP3" "ECTONUCLEOTIDE PYROPHOSPHATASE/PHOSPHODIESTERASE 2; ENPP2"

$PDE11A
[1] "PIGMENTED NODULAR ADRENOCORTICAL DISEASE, PRIMARY, 1; PPNAD1"
[2] "PIGMENTED NODULAR ADRENOCORTICAL DISEASE, PRIMARY, 2; PPNAD2"
[3] "PHOSPHODIESTERASE 11A; PDE11A"

$PDE6C
[1] "CONE DYSTROPHY 4; COD4"
[2] "FRAGILE SITE, FOLIC ACID-TYPE, RARE, FRA(10)(q23.3), CANDIDATE GENE 1; FRA10AC1"
[3] "CYCLIC NUCLEOTIDE-GATED CHANNEL, BETA-3; CNGB3"
[4] "PHOSPHODIESTERASE 6C, CGMP-SPECIFIC, CONE, ALPHA-PRIME; PDE6C"
[5] "CYCLIC NUCLEOTIDE-GATED CHANNEL, ALPHA-3; CNGA3"
[6] "ACHROMATOPSIA 2; ACHM2"
```

Figure 16: *Screenshot of the results of the OMIM database entries function retrieval, using the list of the genes obtained from the C00157 KEGG metabolite entry.*

4.- PubMed database access

Following the idea exposed above, accessing the PubMed database was the next logical step. We wanted to develop a function that was able to retrieve the published research papers that comply with a series of criteria defined by the user using the list of the genes obtained from the altered metabolites. The way we designed the function, it retrieved the following information to aid in the comprehension of the identified alterations: title, authors, abstract, journal, volume, year of publication, PubMed ID, DOI and citation.

```
pubmed <- function(x, keywords_pubmed){
  query_search<-EUtilsSummary(paste0(x, '[Gene Name] ',
  keywords_pubmed),type="esearch", db="pubmed")
  summary_pubmed<-as.data.frame(cbind(
    Title=ArticleTitle(EUtilsGet(query_search)),
    Authors=Author(EUtilsGet(query_search)),
    Abstract=AbstractText(EUtilsGet(query_search)),
    Journal=ISOAbbreviation(EUtilsGet(query_search)),
    Volume=Volume(EUtilsGet(query_search)),
    Year=YearPubmed(EUtilsGet(query_search)),
    Article_PubMed_ID=ArticleId(EUtilsGet(query_search)),
    DOI=ELocationID(EUtilsGet(query_search)),
    Citation=RefSource(EUtilsGet(query_search))))
  return(summary_pubmed)
}
```

```
> keywords_pubmed<-c("[Text Word] AND cancer[Title/Abstract] OR colorectal[Title/Abstract] AND colorectal[Text Word]  AND cancer[Text Word] AND Revi
ew[ptyp] AND alteration[Text Word] AND regulation[Text Word]")
> pubmed('PLAAT3', keywords_pubmed)
```

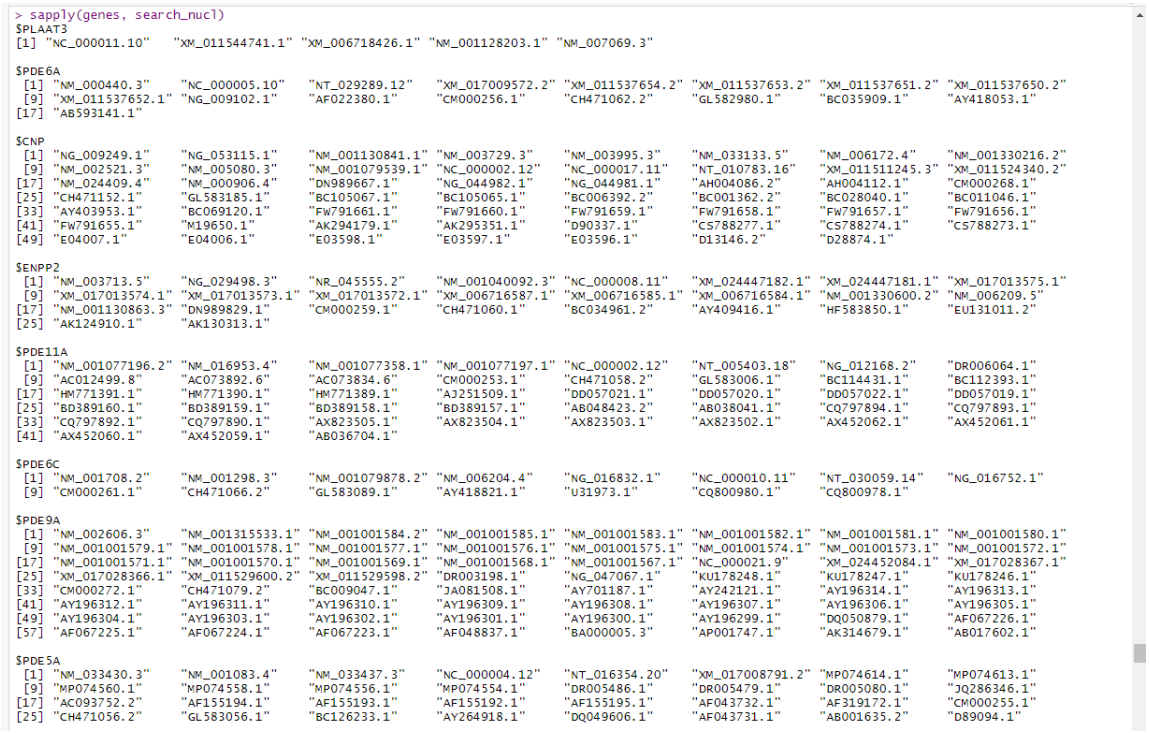
	Title	Authors	Abstract	Journal	Volume	Year	Article_PubMed_ID	DOI
1	The Role of Ubiquitination in Regulating Embryonic ...	*4 variables*	Ubiquitination regulates nearly every aspect of cellular events in eukaryotes...	Int J Mol Sci	20	2019	31151253	E2667
2	LncRNAs with miRNAs in regulation of gastric, liver, ...	*4 variables*	Long noncoding RNA (lncRNA) is a kind of RNAi molecule composed of hundr...	Appl. Microbiol. Biot...	103	2019	31062053	10.1007/s00253-019-(
3	Importance of probiotics in the prevention and treat...	*4 variables*	Colorectal cancer (CRC) remains one of the most common and deadly cance...	J. Cell. Physiol.	234	2019	30912128	10.1002/jcp.28473
4	The Impact of miRNA in Colorectal Cancer Progressi...	*4 variables*	Colorectal cancer (CRC) is one of the most commonly diagnosed malignanci...	Int J Mol Sci	19	2018	30469518	E3711
5	Role of Microbiome in Carcinogenesis Process and E...	*4 variables*	Epigenetic changes during the development of colorectal cancer (CRC) play ...	Methods Mol. Biol.	1856	2018	30178245	10.1007/978-1-4939-8
6	Epigenetic and epitranscriptomic changes in colorec...	*4 variables*	A cancer cell is the final product of a complex mixture of genetic, epigenetic ...	Cancer Lett.	419	2018	29360561	S0304-3835(18)30071
7	Molecular mechanistic pathway of colorectal carcino...	*4 variables*	The colon rectal portion of gastrointestinal tract (GI) is full of microorganism...	Anaerobe	49	2017	29277623	S1075-9964(17)30231
8	Functional significance and therapeutic implication ...	*4 variables*	Accumulative studies revealed that E3 ubiquitin ligases have important roles...	Oncogene	37	2017	28925398	10.1038/onc.2017.31
9	Melatonin for the prevention and treatment of cancer.	*4 variables*	The epidemiological studies have indicated a possible oncostatic property of...	Oncotarget	8	2017	28415828	10.18632/oncotarget.
10	Mutant allele specific imbalance in oncogenes with ...	*4 variables*	Mutant allele specific imbalance (MASI) was initially coined to describe copy ...	Cancer Lett.	384	2016	27725226	S0304-3835(16)30625
11	Genetics and Genetic Biomarkers in Sporadic Colore...	*4 variables*	Sporadic colorectal cancer (CRC) is a somatic genetic disease in which patho...	Gastroenterology	149	2015	26216840	10.1053/j.gastro.2015
12	How the Intricate Interaction among Toll-Like Recept...	*4 variables*	The gut is able to maintain tolerance to microbial and food antigens. The int...	J Immunol Res	2015	2015	26090491	10.1155/2015/489821
13	An update on miRNAs as biological and clinical dete...	*4 variables*	Colorectal carcinogenesis represents a sequential progression of normal colo...	Future Oncol	11	2015	26075447	10.2217/fon.15.83
14	DNA methylation accumulation and its predetermin...	*4 variables*	Aberant DNA methylation is a common epigenomic alteration in carcinogene...	J. Biochem.	156	2014	24962701	10.1093/jb/mvu038
15	Circadian clock circuitry in colorectal cancer.	*4 variables*	Colorectal cancer is the most prevalent among digestive system cancers. Ca...	World J. Gastroenter...	20	2014	24764658	10.3748/wjg.v20.i15.4
16	Functions and Regulation of the PTEN Gene in Color...	*4 variables*	Phosphatase and TENsin homolog deleted on chromosome 10 (PTEN) is a tu...	Front Oncol	3	2014	24475377	10.3389/fonc.2013.00
17	Potential role of probiotics on colorectal cancer prev...	*4 variables*	BACKGROUND: Colorectal cancer represents the most common malignancy ...	BMC Surg	12 Suppl 1	2012	23173670	10.1186/1471-2482-1
18	Regulatory T cells in inflammatory bowel diseases a...	*4 variables*	Regulatory T cells (Tregs)) are key elements in immunological self-tolerance...	World J. Gastroenter...	18	2012	23155308	10.3748/wjg.v18.i40.5
19	Mechanistic aspects of COX-2 expression in colorect...	*4 variables*	The cyclooxygenase-2 (COX-2) enzyme catalyzes the rate-limiting step of pr...	Recent Results Can...	191	2012	22893198	10.1007/978-3-642-30
20	Pharmacological control of autophagy: therapeutic ...	*4 variables*	Autophagy, an intracellular process involved in removing and recycling cellul...	Curr. Pharm. Des.	18	2012	22632751	*NA*
21	Genetics, cytogenetics, and epigenetics of colorecta...	*4 variables*	Most of the colorectal cancer (CRC) cases are sporadic, only 25% of the pati...	J. Biomed. Biotechnol.	2011	2011	21490705	10.1155/2011/792362

Figure 17: *Screenshot of the output of the function for PubMed access for the first gene obtained from C00157 metabolite in KEGG database search, the PLAAT3 gene. Keywords used in the search are included in the image.*

5.- NCBI Nucleotide database access

As we discussed previously, one of the steps of the biomarker discovery pipeline is the validation of the identified potential biomarkers. In our case, because we are working with metabolites and the enzymes that produce them, we thought it would be a good option to include a function that retrieved the transcripts entries related to the identified altered genes. This way, if an experimental validation of altered expression of those genes is required by qPCR, we could speed up the primers design process.

```
search_nucl <- function(x){
  ids<-entrez_search("nucleotide",    term=paste0(x,    "    AND    Homo
    sapiens[porgn]"), retmax=100)$ids
  accession<-sapply(ids, function(x){
    entrez_summary("nucleotide", x)$accessionversion
  })
  return(unname(accession))
}
```



```
> sapply(genes, search_nucl)
$PLAAT3
[1] "NC_000011.10"    "XM_011544741.1" "XM_006718426.1" "NM_001128203.1" "NM_007069.3"

$PDE6A
 [1] "NM_000440.3"    "NC_000005.10"   "NT_029289.12"   "XM_017009572.2" "XM_011537654.2" "XM_011537653.2" "XM_011537651.2" "XM_011537650.2"
 [9] "XM_011537652.1" "NG_009102.1"    "AF022380.1"     "CM000256.1"     "CH471062.2"     "GL582980.1"     "BC035909.1"     "AY418053.1"
[17] "AB593141.1"

$CNP
 [1] "NG_009249.1"    "NG_053115.1"    "NM_001130841.1" "NM_003729.3"    "NM_003995.3"    "NM_033133.5"    "NM_006172.4"    "NM_001330216.2"
 [9] "NM_002521.3"    "NM_005080.3"    "NM_001079539.1" "NC_000002.12"   "NC_000017.11"   "NT_010783.16"   "XM_011511245.3" "XM_011524340.2"
[17] "NM_024409.4"    "NM_000906.4"    "DN989667.1"     "NG_044982.1"    "NG_044981.1"    "AH004086.2"     "AH004112.1"     "CM000268.1"
[25] "CH471152.1"     "GL583185.1"     "BC105067.1"     "BC105065.1"     "BC006392.2"     "BC001362.2"     "BC028040.1"     "BC011046.1"
[33] "AY403953.1"     "BC069120.1"     "FW791661.1"     "FW791660.1"     "FW791659.1"     "FW791658.1"     "FW791657.1"     "FW791656.1"
[41] "FW791655.1"     "M19650.1"       "AK294179.1"     "AK295351.1"     "D90337.1"       "CS788277.1"     "CS788274.1"     "CS788273.1"
[49] "E04007.1"       "E04006.1"       "E03598.1"       "E03597.1"       "E03596.1"       "D13146.2"       "D28874.1"

$ENPP2
 [1] "NM_003713.5"    "NG_029498.3"    "NR_045555.2"    "NM_001040092.3" "NC_000008.11"   "XM_024447182.1" "XM_024447181.1" "XM_017013575.1"
 [9] "XM_017013574.1" "XM_017013573.1" "XM_017013572.1" "XM_006716587.1" "XM_006716585.1" "XM_006716584.1" "NM_001330600.2" "NM_006209.5"
[17] "NM_001130863.3" "DN989829.1"     "CM000259.1"     "CH471060.1"     "BC034961.2"     "AY409416.1"     "HF583850.1"     "EU131011.2"
[25] "AK124910.1"     "AK130313.1"

$PDE11A
 [1] "NM_001077196.2" "NM_016953.4"    "NM_001077358.1" "NM_001077197.1" "NC_000002.12"   "NT_005403.18"   "NG_012168.2"    "DR006064.1"
 [9] "AC012499.8"     "AC073892.6"     "AC073834.6"     "CM000253.1"     "CH471058.2"     "GL583006.1"     "BC114431.1"     "BC112393.1"
[17] "HM771391.1"     "HM771390.1"     "HM771389.1"     "AJ251509.1"     "DD057021.1"     "DD057020.1"     "DD057022.1"     "DD057019.1"
[25] "BD389160.1"     "BD389159.1"     "BD389158.1"     "BD389157.1"     "AB048423.2"     "AB038041.1"     "CQ797894.1"     "CQ797893.1"
[33] "CQ797892.1"     "CQ797890.1"     "AX823505.1"     "AX823504.1"     "AX823503.1"     "AX823502.1"     "AX452062.1"     "AX452061.1"
[41] "AX452060.1"     "AX452059.1"     "AB036704.1"

$PDE6C
 [1] "NM_001708.2"    "NM_001298.3"    "NM_001079878.2" "NM_006204.4"    "NG_016832.1"    "NC_000010.11"   "NT_030059.14"   "NG_016752.1"
 [9] "CM000261.1"     "CH471066.2"     "GL583089.1"     "AY418821.1"     "U31973.1"       "CQ800980.1"     "CQ800978.1"

$PDE9A
 [1] "NM_002606.3"    "NM_001315533.1" "NM_001001584.2" "NM_001001585.1" "NM_001001583.1" "NM_001001582.1" "NM_001001581.1" "NM_001001580.1"
 [9] "NM_001001579.1" "NM_001001578.1" "NM_001001577.1" "NM_001001576.1" "NM_001001575.1" "NM_001001574.1" "NM_001001573.1" "NM_001001572.1"
[17] "NM_001001571.1" "NM_001001570.1" "NM_001001569.1" "NM_001001568.1" "NM_001001567.1" "NC_000021.9"   "XM_024452084.1" "XM_017028367.1"
[25] "XM_017028366.1" "XM_011529600.2" "XM_011529598.2" "DR003198.1"     "NG_047067.1"    "KU178248.1"     "KU178247.1"     "KU178246.1"
[33] "CM000272.1"     "CH471079.2"     "BC009047.1"     "JA081508.1"     "AY701187.1"     "AY242121.1"     "AY196314.1"     "AY196313.1"
[41] "AY196312.1"     "AY196311.1"     "AY196310.1"     "AY196309.1"     "AY196308.1"     "AY196307.1"     "AY196306.1"     "AY196305.1"
[49] "AY196304.1"     "AY196303.1"     "AY196302.1"     "AY196301.1"     "AY196300.1"     "AY196299.1"     "DQ050879.1"     "AF067226.1"
[57] "AF067225.1"     "AF067224.1"     "AF067223.1"     "AF048837.1"     "BA000005.3"     "AP001747.1"     "AK314679.1"     "AB017602.1"

$PDE5A
 [1] "NM_033430.3"    "NM_001083.4"    "NM_033437.3"    "NC_000004.12"   "NT_016354.20"   "XM_017008791.2" "MP074614.1"     "MP074613.1"
 [9] "MP074560.1"     "MP074558.1"     "MP074556.1"     "MP074554.1"     "DR005486.1"     "DR005479.1"     "DR005080.1"     "JQ286346.1"
[17] "AC093752.2"     "AF155194.1"     "AF155193.1"     "AF155192.1"     "AF155195.1"     "AF043732.1"     "AF319172.1"     "CM000255.1"
[25] "CH471056.2"     "GL583056.1"     "BC126233.1"     "AY264918.1"     "DQ049606.1"     "AF043731.1"     "AB001635.2"     "D89094.1"
```

Figure 18: *Screenshot of the accession codes for the Homo Sapiens transcripts identified in Nucleotides database from NCBI data portal for all the genes identified to be related to C00157 metabolite.*

3.1.2.- Microbiome

For microbiome data analysis, the process of getting OTUs from raw reads is in fact pretty well established and standardized, either if using QIIME or mothur [233, 126, 124] pipelines. What follows, though, is more open to discussion. Some debate had arisen related to how microbiome data should be analyzed and which kind of statistics applied.

Because of the sequencing machines' limited capabilities, some authors are pushing the idea of microbiome datasets to be compositional, so those special statistics are needed to interpret this data [176]. Following this idea, the total number of sequenced reads and thus their assignment into OTUs would depend more on the sequencer capability. This means that the total reads count is non-informative in its raw form, as they would not represent the real composition of the microbiome. Instead, these authors suggest that transforming the total counts for each OTU to its relative frequency would recover the real microbiome population composition, including the relationships between different OTUs. The problem with compositional data analysis is that common statistical approaches are not applicable. Thus, new analytical methods that take into consideration the special features of compositional kind of data have been developed. Comparisons between analytical methodologies, though, have not resolved which one is better, usually obtaining opposite results. What it seems, though, is that different methods only differentiate on the false discovery rate, some being stricter (usually the compositional ones) and others less. It has been described that these differences in the false discovery rates may depend on the features of the data, like the libraries' size and the number of samples per group. For our clinical projects (results Chapters 2, 4, 5 and

6), we decided to take into consideration our datasets characteristics and try to analyze which methodology fits better in each case.

Thus, for the fibromyalgia microbiome dataset (Chapter 4), we compared the results obtained with a more standard statistics approach, by using DESeq2 tool [234] with the compositional methods one, by means of ALDEx2 [235, 236]. We saw that common statistics approach (DESeq2) and non-corrected compositional data approach lead to similar results. When multiple testing correction was applied, though, the significance of these results was lost (FIGURE 19). Because we had performed an experimental validation of the difference in fibromyalgia microbiome identified by DESeq2 methodology, we chose to use those results [237]. It has been suggested that a non-even sample number per group may affect ALDEx2 performance [238] and considering that fibromyalgia group had twice the samples as control one we think this may be the explanation for these differences.

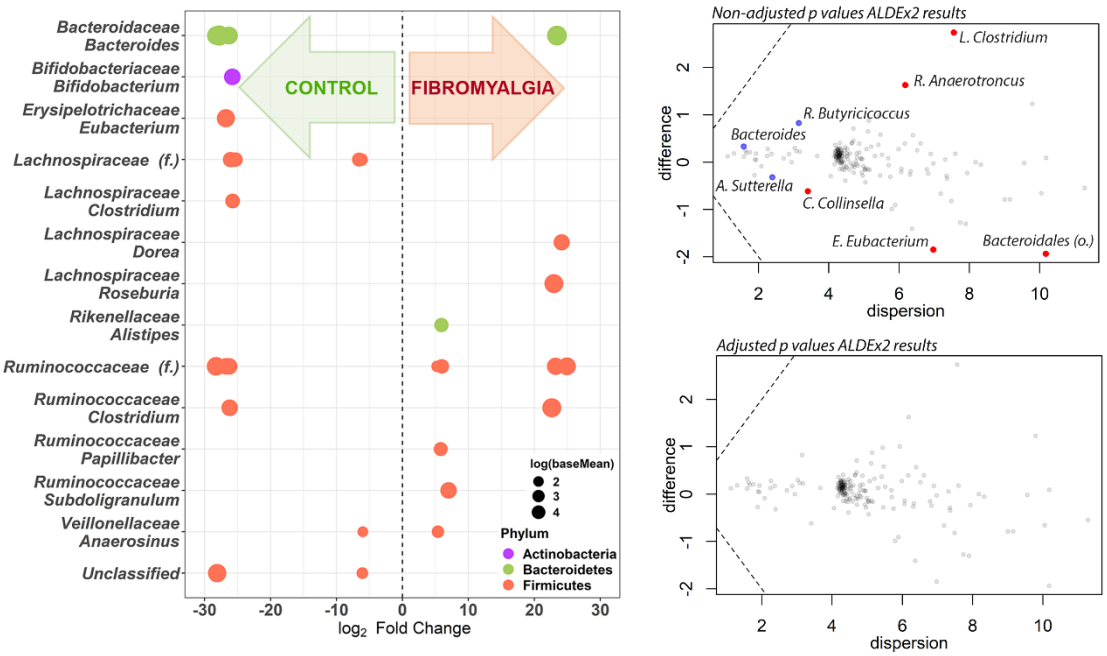


Figure 19: *Differences between distinct microbiome analysis tools at the genus level for fibromyalgia study. To the left, results obtained with DESeq2 methodology as published in* [237]*, upper-right the non-adjusted compositional ALDEx2 results, lower-right the multitest adjusted results for ALDEx2 tool. Grey points represent abundant non-differential features, black points the non-differential rarely abundant features, blue dots the features identified as significantly different by one test (t-test or Wilcoxon) and red ones the significantly different features identified by both tests.*

This comparison among the different tools was also repeated with the data of CRC microbiome project (Chapter 6) in order to determine which analytical tool and pipeline was the most appropriated for the corresponding data characteristics. We combined three tools: ALDEx2, SIAMCAT [239] and LEfSe [240]. While LEfSe identified more bacterial genera to be differentially expressed between sample groups, these

differences were lost when significance was tested by other methodologies. SIAMCAT and ALDEx2 results, though, were consistent (FIGURE 20).

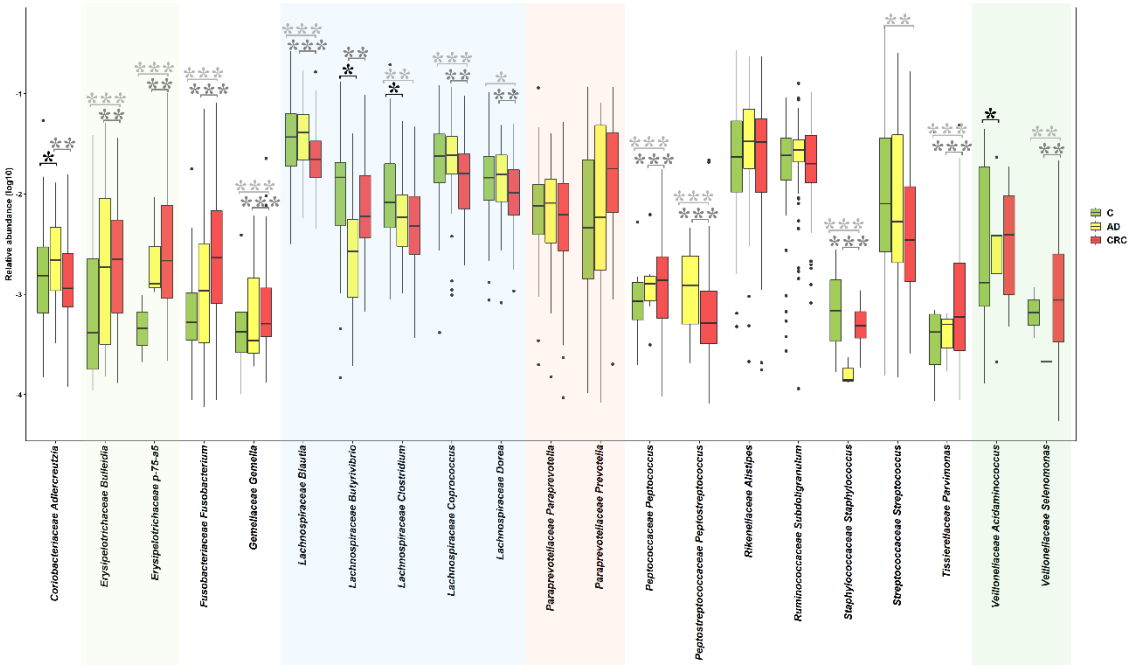


Figure 20: *Relative abundance of genera identified to be differentially abundant between sample groups (C, AD and CRC) by the three mentioned tools. Genera are grouped by family and shaded accordingly. Significance of the pairwise differences are indicated in grayscale as follows: black for C-AD, dark gray for AD-CRC and light gray for C-CRC; * means <.05 significance, **<0.01 and ***<0.001.*

As can be seen in FIGURE 20, the most robust tools were both SIAMCAT and ALDEx2, which implement compositional methodology analytical approaches. Thus, is suggested that results deriving from the application of these two tools may provide with potential biomarkers more prone to be experimentally validated.

Another issue that is currently in discussion is the normalization of microbiome datasets. While until recently rarefaction was a standardized step in the microbiome datasets processing pipeline, nowadays it's been discussed if this could lead to a loss of information [174]. This affirmation, though, has been intensively debated [238], without reaching a consensus. In fact, the most recurrent argument we found in microbiome data analysis is that methodology should be adapted to data characteristics. For our case, we found that no differences were observed when β-diversity indexes were measured on fibromyalgia's microbiome study (FIGURE 21). As can be seen in this figure, the distribution of samples in either non-rarefied and the rarefied data lead to no specific clustering of samples, thus suggesting that rarefaction did not affect the diversity measurements. Accordingly, for this thesis rarefaction procedure was applied to normalize the number of sequences per sample.

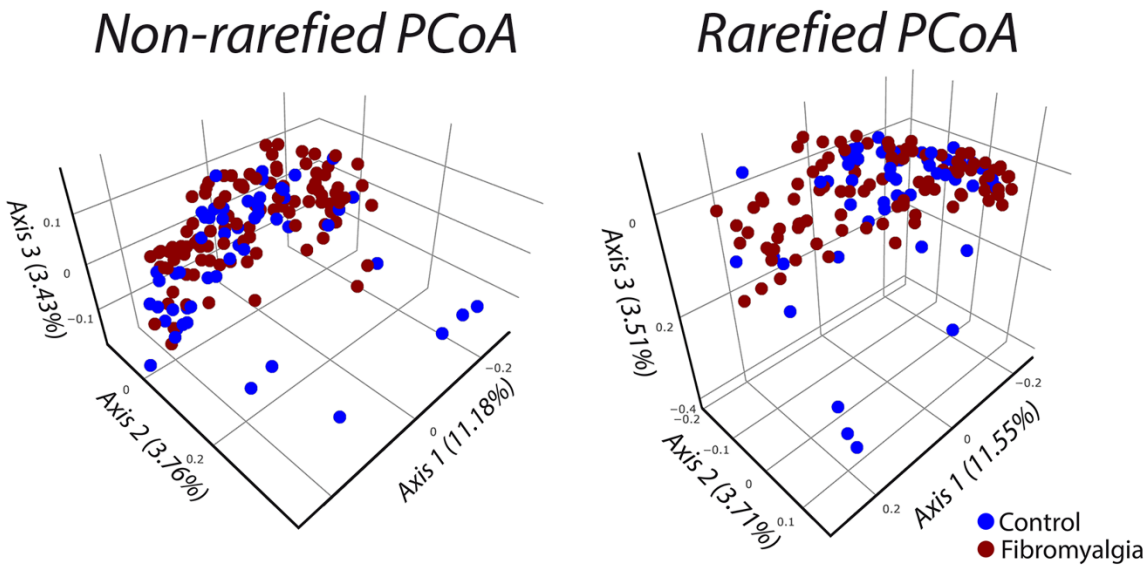


Figure 21: *PCoA plots of non-rarefied and rarefied microbiome datasets for the fibromyalgia project samples. While the distribution of samples seems to change a little bit, no significant differences in diversity of the subpopulations were found.*

3.1.3.- Metabolomics – microbiome integration

We already introduced the current approach options for omics data integration, which in summary rely on dimension reduction techniques and/or the identification of correlations between individual variables of each omics and the between the full omics datasets. While dimension reduction methodologies are useful to easier the interpretation of high-throughput datasets, the identification of potential individual biomarkers from these kinds of analyses is less easy. In contraposition, correlation analysis performed at individual variables level allows the easier identification of potential biomarkers but complicates the identification and interpretation of potential regulatory interactions between different variables, simplifying in excess the representation of a specific phenotype.

Thus, when we designed our analytical and integrative pipeline, we wanted to combine both approaches, in order to better explain and define any potentially interesting result. Among the distinct tools we tested for the dimension reduction approach, mixOmics [241] and especially their team's tool DIABLO [242] was the easier, most user-friendly and less resource consumer option of all.

For the variables correlations analyses, HAllA from Huttenhower's laboratory [243] was found to be an easy to use tool, with clear instructions on how to install and run it. Otherwise, the most standard approach was to calculate Spearman's correlation coefficient between each variable of each omics dataset using R basic tools, such as *cor* and *cor.mtest* R functions, specifying *method = 'Spearman'* in the available options of the functions. In our case, we used both approaches for the distinct project, using HAllA for CRC microbiome-metabolomics project and standard R functions approach for multi-omics fibromyalgia project.

3.1.4.- Final pipeline

As result of the bioinformatics tool combination and use a pipeline is proposed for the independent analysis of metabolomics and microbiome datasets and their integration is summarized in FIGURE 22.

Briefly, raw data is processed and normalized in the first step. Then, each omics data is analyzed separately, using the normalized datasets. For metabolomics, both multivariate and univariate analyses are performed and combined. Functional profiling is performed by means of our ad-hoc scripts, accessing HMDB and KEGG databases.

For microbiome, processing and taxonomical annotation are performed with QIIME2 software. Then, α and β diversity indexes are calculated and statistical differences measured to identify potential differences between the compared sample groups. Differences in specific bacteria are computed with the taxonomically annotated dataset and functional profiling is inferred using specific software.

Both omics datasets are then joined and 3 kinds of multi-omics analytical methods are applied: dimension reduction, using mixOmics package, Procrustes analysis, using basics R functions and correlation analysis, either using R basics functionalities and/or HAllA software. Finally, predictive modeling is performed with the combination of the distinct omics and most discriminating variables identified.

A relation of all the tools, databases and resources used in this thesis can be found on Supplementary Table 1.

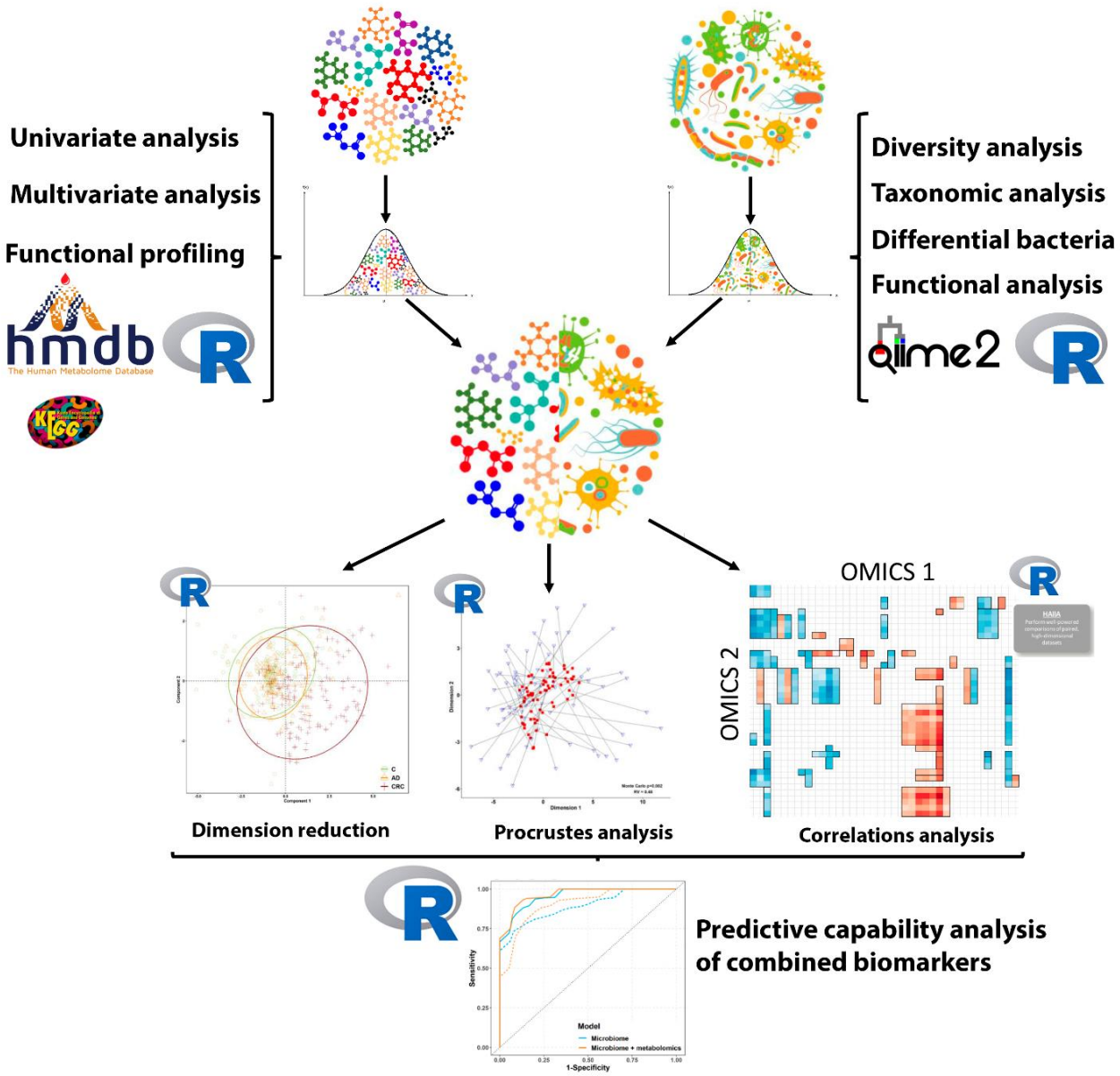


Figure 22: *Proposed analytical pipeline for the processing and integration of metabolomics (left) and microbiome (right) datasets. Analytical approaches used in each step are indicated with text near the appropriated dataset. Tools and databases used are identified by the corresponding logo, including HMDB and KEGG databases, QIIME2, HAllA and R software.*

3.2.- Chapter 2. Prostate Cancer EVs metabolomics

The work related to this case was published as a Research Article, with first shared authorship, in *Journal of Extracellular Vesicles* and can be found as Annex I:

Clos-Garcia, M.*, Loizaga-Iriarte, A.*, Zuñiga-Garcia, P.* *et al.* (2018) 'Metabolic alterations in urine extracellular vesicles are associated to prostate cancer pathogenesis and progression', *Journal of Extracellular Vesicles*. Taylor & Francis, 7(1) [244].

3.2.1.- Introduction

Prostate cancer (PCa) is among the most frequently diagnosed and deadly types of cancer in men in Western countries (http://globocan.iarc.fr). Lack of sensitive and specific diagnostic tools, especially to detect early stages of the disease, and the unknown underlying mechanisms of onset and progression of PCa are the major problems to treat PCa with the highest efficacy. Thus, there is a high demand to discover more sensitive and specific biomarkers to improve PCa diagnosis and prognosis. Nowadays, prostate-specific antigen (PSA) blood screening tests, together with clinical T-stage and Gleason score are the standard tests to discriminate patients with low, intermediate or high risk to suffer PCa [245].

Metabolomics is recognized as the ultimate "omics" discipline with high potential to identify sensitive and specific markers and to understand the mechanisms involved in the development of pathological processes [246]. The recent technological revolution in separation and detection of small molecules, combined with rapid progress in bioinformatics, is making it possible to rapidly measure a large number of metabolites in a small amount of sample [247, 248]. Metabolomics comprises the qualitative and quantitative measurements of the metabolic response to physiological or pathological stimuli. It involves the extraction and measurement of low molecular weight molecules (e.g. amino acids, sugars, bile acids, fatty acids, vitamins, etc.) belonging to different metabolic pathways to generate metabolic profiles of cells, tissues or biofluids [249, 250]. Previous studies have shown the utility of serum metabolite levels as a diagnostic tool for different cancer types [251], and in PCa some metabolites have already been suggested as candidate biomarkers. Increased serum levels of polyunsaturated fatty acids have been associated to reduce risk of PCa, while higher levels of serum testosterone were associated with an increased risk of suffering this malignancy [252]. Other metabolomics approaches have reported alterations of acylcarnitines, glucose, glycerophospholipids (including lysophosphatidylcholines and phosphatidylcholines), amino acids and triglycerides in PCa [253].

Urine samples have been intensely used to identify PCa biomarkers [254], due to its easy availability and handling, and its anatomical proximity to the prostate. As occurs for the serum, there are also several metabolomics studies of urine samples that found alterations in urinary levels of more than 20 metabolites including N-methyl glycine, kynurenine, uracil, glycerol 3-phosphate, dihydroxybutanoic acid, xylonic acid, pyrimidine, ribofuranoside and xylopyranose (reviewed in [255]). These studies have

pointed out that many metabolic pathways may be altered in PCa including glycine synthesis and degradation, and carbohydrate and energy metabolisms. Although all these metabolites need further clinical validation, they support the notion that metabolomics constitutes a suitable technology to identify candidate biomarkers of PCa.

One important drawback of using urine sample for biomarker discovery is that many of their constituents are diluted avoiding to be detected by current technologies. Thus, in order to detect underrepresented molecules, it is still required to concentrate the sample. In this context, cell-secreted extracellular vesicles (EVs) are present in all body fluids, including urine [19], and could provide a concentrated source of molecules. Thus, a deep analysis of the urinary EVs composition could open a window of opportunities to identify more sensitive and specific PCa biomarkers. Inline, a recent lipidomics study performed in these urinary vesicles from healthy and PCa samples reveal up to nine lipid species differentially expressed as to potential PCa biomarkers [256] supporting the existence of metabolic changes in urine EVs from PCa patients.

In the current study, we have compared urinary EVs obtained from PCa and benign prostate hyperplasia (BPH) patients and focused on the analysis of the metabolites that they contain by performing a UHPLC-MS targeted metabolomics analysis. We evaluated the levels of 248 metabolites belonging to different chemical nature including amino acids, nucleosides, vitamins, as well as different lipid species. Among them, 76 metabolites were found significantly altered in PCa compared to BPH. Some of these metabolites were significantly correlated with current markers of PCa (e.g. PSA). Interestingly, dehydroepiandrosterone sulfate was among the most significantly altered metabolites in PCa, supporting the notion that beyond their function as “metabolic machines” [247, 257, 258] EVs could inform about metabolic alterations of cancerous tissue.

3.2.2.- Methods

3.2.2.1.- Patient samples

All urine samples were obtained from the Basque Biobank for research (BIOEF, Basurto University hospital) upon informed consent and with evaluation and approval from the corresponding ethics committee (CEIC code OHEUN11-12 and OHEUN14-14). The clinical classification of the patients is described in Table 2. For each sample, urine (50 ml) was collected by spontaneous micturition, centrifuged at 2,000 × g 10 min, filtered through a 0.22 µm pore membrane and immediately frozen at −80ºC.

3.2.2.2.-Urine extracellular vesicle isolation and characterization

To isolate EVs from urine (average ± SEM; 49.7 ± 0.86 ml), the stored samples were thawed, centrifuged at 10,000 × g for 30 min and the supernatant ultra-centrifuged at 100,000 × g for 75 min. The resulting pellet was washed with an excess of phosphate-saline buffer (PBS), and again ultra-centrifuged at 100,000 × g for 60 min. The final pellet was re-suspended in 150 µL of PBS, aliquot generated and kept at −80°C for further analysis. Protein was determined by Bradford and obtained 32.7 ± 4.6 (mean±SEM)

micrograms on average of total purified protein from the initial urine volume (50 ml). The size distribution of the particles present in the isolated preparations was determined by measuring the Brownian motion using a NanoSight LM10 system equipped with fast video capture and particle-tracking software (Malvern, UK). Pre- and post-acquisition settings were maintained the same for all the samples and each video was analyzed to give the mean, mode, and median vesicle size, as well as an estimate of the particle concentration. Then, an average curve was calculated for each group of patients to be compared among them. Cryoelectron microscopy and Western-blot analysis were performed as described previously [259].

Table 4: *Clinical classification of the samples. In parentheses are indicated the median ± SD of age for each group of samples.*

Disease status	Stage	Perineural invasion	n
Prostate Cancer (PCA) (64±4.41)	Stage 2 (64±4.12)	No (Pn0) (65.5±5.02)	6
		Yes (Pn1) (64±3.47)	10
	Stage 3 (64.5±4.68)	*NA*	15
Benign Hyperplasia (BPH) (70±5.71)	*NA*	*NA*	14

3.2.2.3.-Metabolite extraction and UHPLC-MS analysis

Metabolic profiles of urinary EVs were semi-quantified using four UHPLC-MS-based analytical platforms as previously described [260, 261]. Methanol was first added to urinary EV preparations, and after brief vortex, chloroform was added. Both extraction solvents were spiked with metabolites not detected in unspiked EV extracts: tryptophan-d5(indole-d5), PC(13:0/0:0), FA (19:0), dehydrocholic acid, SM(d18:1/6:0), PE(17:0/ 17:0), PC(19:0/19:0), TAG(13:0/13:0/13:0), Cer(d18:1/17:0), ChoE(12:0), anthranilic acid-(ring-13C6), phe-nylthiohydantoin (PTH)-valine and glycocholic- 2,2,4,4-d4 acid. Samples were incubated at −20°C for 30 min and, after vortex, three different phases were collected. Platform 1 included fatty acyls, bile acids, steroids and lysoglycerophospholipids profiling. Supernatants were collected after centrifugation at 16,000 × g for 15 min, dried, reconstituted in methanol, resuspended for 20 min and centrifuged (16,000 × g for 5 min) before being transferred to vials for UHPLC-MS analysis. Platform 2 included glycerolipids, cholesteryl esters, sphingolipids and glycerophospholipids profiling. Extracts were mixed with water (pH = 9) and after brief vortex mixing, the samples were incubated for 60 min at −20°C. After centrifugation at 16,000 × g for 15 min, the organic phase was collected and the solvent removed. The dried extracts were then reconstituted in acetonitrile/isopropanol (50:50), resuspended for 10 min, centrifuged (16,000xg for 5 min) and transferred to vials for UHPLC-MS analysis. Platform 3 included amino acids profiling; 10 µl aliquots from the extracts prepared for Platform 1 were transferred to microtubes and derivatized for amino acid analysis. Finally, Platform 4 consisted of the analysis of polar metabolites profiling, including central carbon metabolism. Extracts were mixed with chloroform. After brief

vortex mixing, water was added and samples were mixed for 10 min at room temperature. Afterward, the samples were centrifuged at 16,000 × g for 10 min. The supernatants were collected and dried. Extracts were then solubilized in water and after centrifugation, supernatants were transferred to vials for UHPLC-MS analysis.

Chromatographic separation and mass spectrometric detection conditions employed were previously described [260, 261]. The overall quality of the analysis procedure was monitored using six repeat injections of a pooled sample, considered as the quality control sample. For each of the four analytical platforms, randomized sample injections were performed, with QC calibration and validation extracts uniformly interspersed throughout the entire batch run. Generally, the retention time stability was <6 s injection-to-injection variation and the mass accuracy <3 ppm for m/z 400–1200, and <1.2 mDa for m/z 50–400. Details of lipid nomenclature used in this work are provided as supplementary material (https://www.tandfonline.com/doi/full/10.1080/20013078.2018.1470442).

3.2.2.4.- Data processing, statistical and bioinformatics analysis

3.2.2.4.1.- Amount of urine sample and data normalization

A similar volume of urine samples (50 ml) from each patient was employed for obtaining the EV preparations. Then, the complete EV preparations were analyzed by UPLC-MS metabolomics analysis. The peak intensities for each metabolite included in the analysis were normalized to the sum of the peak intensities within each sample. There was no significant correlation (F < Fcrit) between the sum of the peak intensities used for the normalization and the groups being compared in the study.

3.2.2.4.2.- Missing values imputation

First, metabolites that were not detected in at least 70% of the whole set of samples were removed from the analysis. Then, taking the minimal value for each metabolite and dividing it by a factor of 10, missing values were imputed in order to obtain the final data set.

3.2.2.4.3.- Univariate analysis

Three different comparisons were established for the analyses:

- Prostate cancer (PCa) vs benign prostate hyperplasia (BPH).
- PCa pathological stage 3 vs PCa pathological stage 2.
- In the PCa pathological stage 2 group, perineural invasion: Pn1 vs Pn0.

The mean and 90% Winsorized-mean for each metabolite and each group of patients were calculated, as well as, Student's t-test or Wilcoxon signed-rank test, depending on the normality of the data that was assessed using Shapiro-Wilk test. Median, standard error of the mean (SEM), the standard deviation (SD), coefficient of variation and the Interquartile Range (IQR) were also calculated.

Several calculations were performed for the three distinct comparisons. We calculated the F-test of the two variances, the Student's t-test, Wilcoxon signed-rank and Fold Change for each metabolite. To test the discriminatory capacity of each metabolite for each one of the three comparisons we performed Receiver Operating Characteristic

(ROC) analysis, including in the calculations the values of the Area Under the Curve (AUC), sensitivity, specificity, positive predictive value, negative predictive value, Youden index and the optimal cut-off.

For each one of the three pairwise comparisons, we generated box-plots for those metabolites with significant differences between the two groups with adjusted p-values following Bonferroni methodology. Heatmaps indicating log2 value of Fold Change and Bonferroni adjusted p-values were also calculated. Finally, volcano plots were generated with the log2 Fold Change values and Bonferroni adjusted p-values.

All statistical analyses were performed using R software v3.3.2 (R Development Core Team, 2016; http://cran.r-project.org) with *stats, caret, psych* and *OptimalCutpoints* package [262]. Boxplots and volcano plots were generated with *ggplot2* R package. Correlations with clinical parameters such as BMI were done with *cor.test* function in R software, using Spearman's method. Both rho and p-value for each metabolite are reported. We studied the correlation between BMI and metabolite levels with all the samples together and also dividing samples depending on their clinical status.

3.2.2.4.4.- Multivariate analysis

Principal Component Analysis (PCA), Partial Least Squares-Discriminant Analysis (PLS-DA) and Orthogonal Partial Least Squares (OPLS) were performed for each pairwise comparison using SIMCA-P v12.0.1.0 software (Umetrics AB).

3.2.2.4.5.- Metabolites mapping into cellular metabolic pathways and identification of primary enzymes associated with their metabolism

Metabolic pathways were determined with MetScape v3.1.2 application, running under Cytoscape v3.5.0 software, linking them to the KEGG Pathway database (http://www.genome.jp/kegg/pathway.html). Primary enzymes involved in the metabolism of the metabolite of interest, and their corresponding coding genes were retrieved from KEGG (http://www.genome.jp/kegg/compound/) and HMDB (http://www.hmdb.ca/) databases, using dbWalk utility on bioDBnet database searching online utility and specifying "9606" (Homo sapiens) Taxon ID on Organism box (https://biodbnet-abcc.ncifcrf.gov/db/dbWalk.php), with the following paths:

- For KEGG compounds, we started with enzyme EC code:
  EC Number->UniProt Accession->UniProt Entry Name->KEGG Gene ID->Gene ID->Gene Symbol->Gene ID->GenBank Nucleotide Accession.
- For HMDB compounds, we started with the name present on HMDB database:
  HMDB Metabolite->HMDB Enzyme -> UniProt Entry Name->Gene Symbol->Gene ID->GenBank Nucleotide Accession.

For each metabolite included in this step, we reported:

- For KEGG compounds, the related enzymes EC number, UniProt Accession, UniProt Entry Name, KEGG Gene ID, Gene Symbol, GeneID and the GenBank Nucleotide Accession for the corresponding transcripts.

- For HMDB compounds, the HMDB enzyme Gene Symbol, Gene Symbol, Gene ID and GenBank Nucleotide Accession for the corresponding transcripts.

Database normalization: all the datasets used for the data mining analysis were downloaded from GEO or TCGA, and subjected to background correction, log2 transformation and quartile normalization as reported [263, 264]. In the case of using a pre-processed dataset, this normalization was reviewed and corrected if required. For normal vs. PCa comparisons, a two-tailed t-test is performed in order to indicate if the observed differences between the groups are significant. For tumor progression analysis, an ANOVA test was performed in order to evaluate if the observed differences of gene expression levels between the groups were significant. DFS analysis was performed using Taylor and TCGA datasets. In both cases, the patients were stratified by quartiles based on the expression of the gene of interest, Kaplan-Meier Estimator was used in order to estimate the survival function from different groups of patients while a Log-Rank test is calculated to check the significance between the curves. In the case of Taylor dataset, the analysis was performed using the average signal from all the transcripts of a gene.

3.2.3.- Results

Urine samples were collected from patients with BPH (n = 14) and PCa (n = 31) with different pathological characteristics (Table 2). In order to avoid any chemical alteration of the vesicles that could interfere with the metabolomics analysis, we decided to preserve the uromodulin status of the samples by avoiding the use of high-salt concentration or reducing agents. After initial clearing at low centrifugation and ultrafiltration, small EVs (exosomes, small microvesicles and apoptotic blebs) were isolated by differential ultracentrifugation as described in [259]. Cryoelectron microscopy revealed the presence of vesicles in the preparations (FIGURE 23A). Western blot analysis showed that while we could not detect mitochondria (COX IV) or endoplasmic reticulum (GRP78) proteins, we could detect exosomal markers (CD10, CD63, CD9, Flotillin and CD26), and also some uromodulin (UMOD/THP) (FIGURE 23B). As previously, we found high inter-individual variability in the abundance of these proteins [259, 265]. In agreement with previous results [259], physical characterization by NTA analysis of the isolated material revealed significant differences in the size distribution of particles isolated from PCa and BPH samples (FIGURE 23C). Interestingly, the size of the particles increased with the stage of the PCa, thus, the major difference was observed between BPH and PCa stage 3 (FIGURE 23C). A significant higher abundance of particles bigger than 350 nm was observed in samples from PCa stage 3 (FIGURE 23C). The mean concentration of particles per ml for all samples was $8.60 \pm 1.19 \times 10^{10}$ EVs/mL. No differences were found for the concentrations of EVs/mL between different groups (BPH, PCa stage 2 Pn0, stage 2 Pn1 and stage 3).

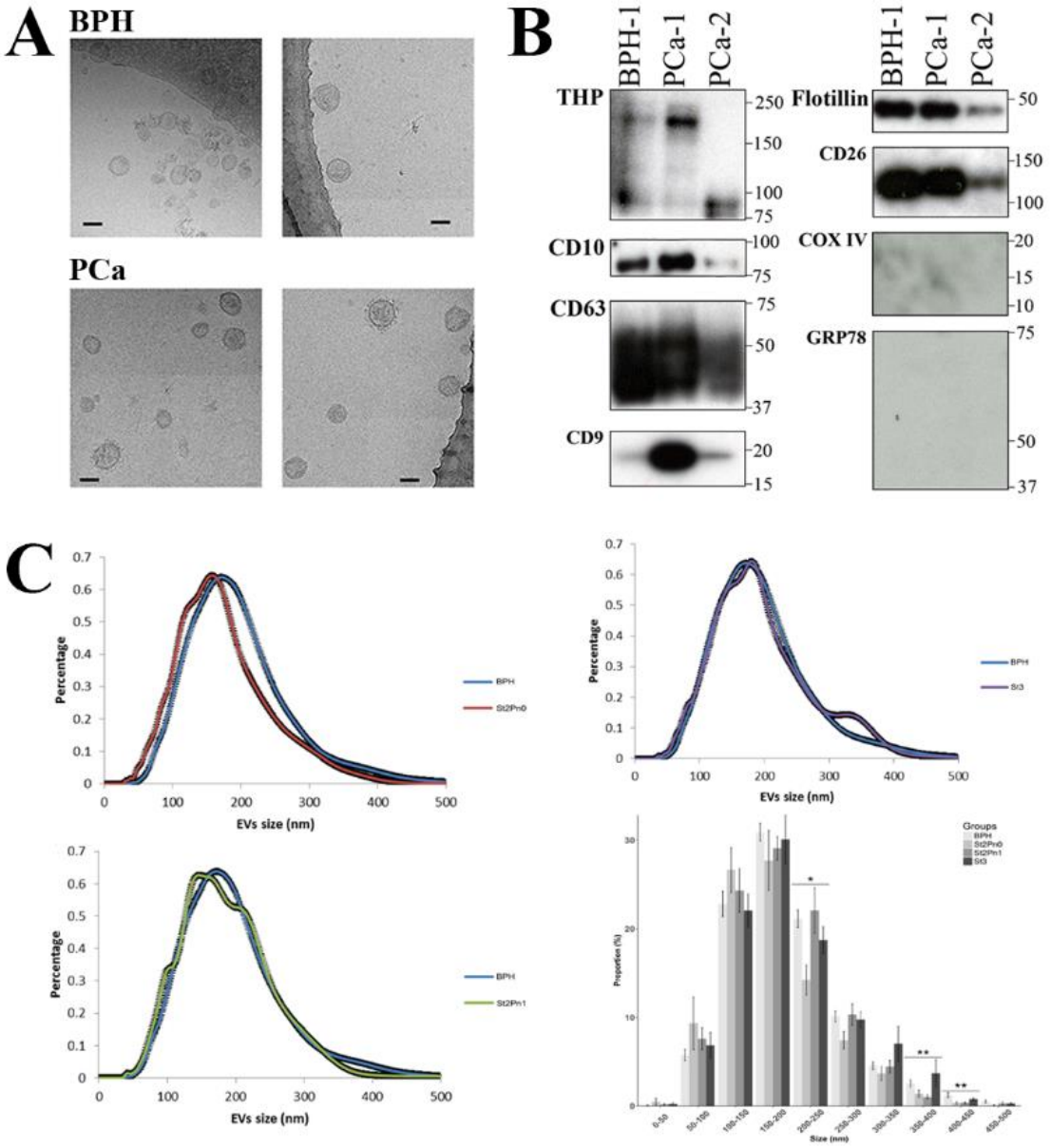


Figure 23: *Isolated EVs biophysical and biochemical characterization. CryoEM pictures of both BPH and PCa sample groups EVs (A); western blot analysis of common EVs markers (CD10, CD63, CD9, CD26, Flotilin), mitochondrial marker (COX IV) and endoplasmic reticulum (GRP78) (B); size distribution and comparison of the EVs population for each PCa subgroup vs BPH (C). All p-values were adjusted by Bonferroni method. Significance levels: * <0.05, ** <0.01 and *** <0.001.*

After this initial characterization, metabolites present in the urinary EV preparations were extracted using different methodologies in order to cover a wide range of molecules with different chemical nature (see *Methods* section). We were able to detect 248 metabolites (https://www.tandfonline.com/doi/full/10.1080/20013078.2018.1470442) including amino acids, vitamins, nucleosides, as well as different lipid species. Considering all the samples, metabolites with more than 70% of missing values were eliminated from the analysis with the exception of PC(14:0/20:4), PC(0:0/20:3) and TG(56:8) because most of the missing values occur mainly in one of the two groups (PCa

or BPH). Afterward, we performed three different statistical analyses comparing BPH and PCa groups, as well as, the association to tumor stage and perineural invasion.

3.2.3.1.- Metabolites differentially altered between BPH and PCa

Univariate analysis revealed that 76 out of 248 metabolites showed statistically significant differences between EVs from PCa and BPH patients. These metabolites were distributed along most chemical families analyzed, although there was a predominance of phosphatidylcholines (PC), fatty acid esters (acylcarnitines) and sterols (FIGURE 24). Whereas a higher abundance of PC was observed in BPH samples, acylcarnitines and sterols were more abundant in PCa samples (FIGURE 24A). In addition, carboxylic acids and glycerolipids were slightly decreased, and vitamins were increased in PCa EVs. The other families of metabolites including amino acids, bile acids, nucleosides, sphingolipids, phosphatidylethanolamines (PE) contained both increased and decreased metabolites (FIGURE 24B). Interestingly, the abundance of ceramides with short carbon number in their acyl chains were increased in PCa samples, while ceramides with long carbon number (>23) in their acyl chains were reduced in PCa EVs. This pattern was not present in other sphingolipids families. In the non-esterified fatty acid family, the abundance of arachidonic acid (20:4n-6) was decreased in PCa samples, while other polyunsaturated fatty acid with shorter carbon chain (16:3n-x) was significantly increased in the PCa group (FIGURE 24).

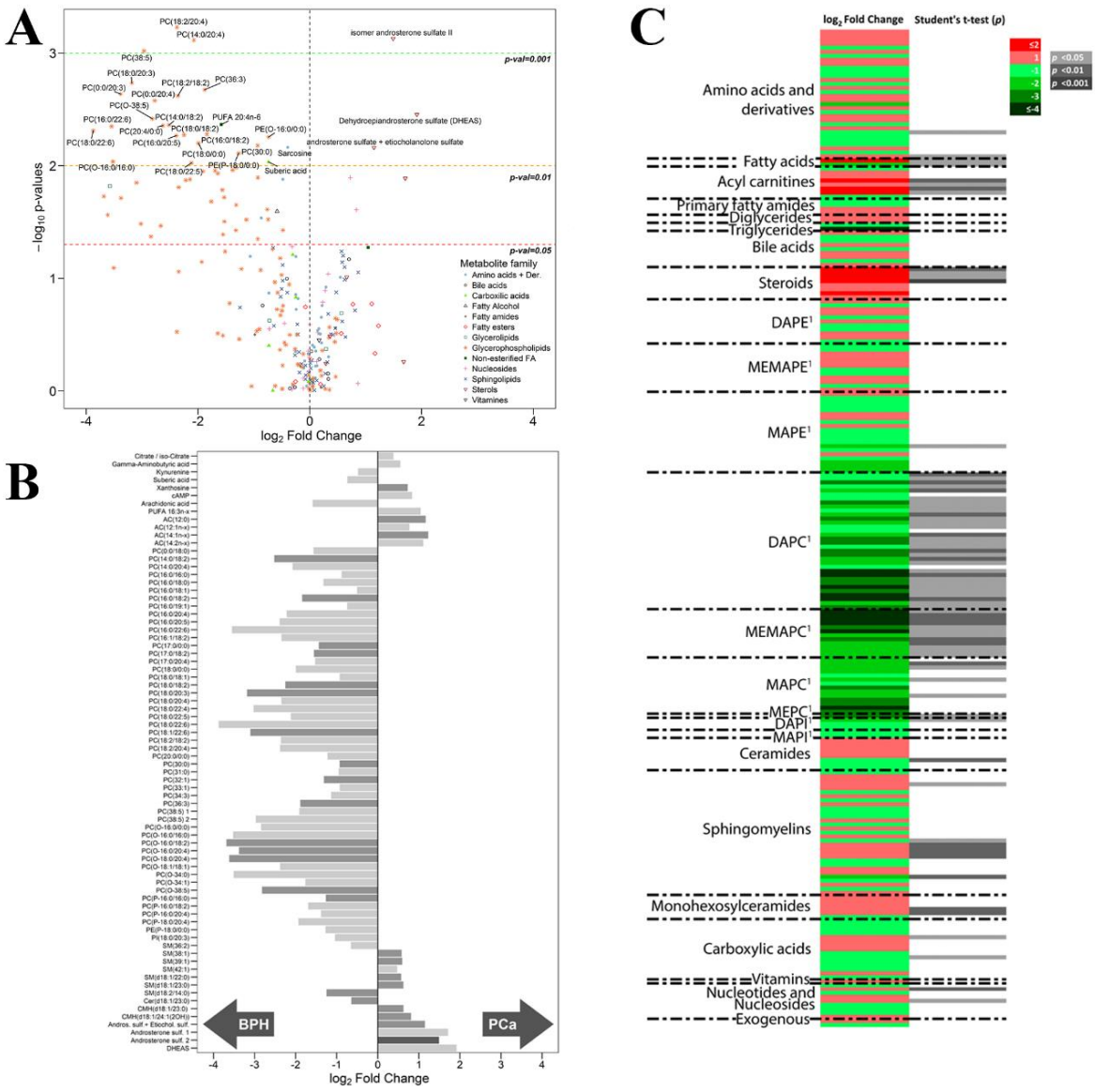


Figure 24: *Univariate differences between PCa (n=31) and BPH (n=14) sample groups. Volcano plot depicting alterations between PCa and BPH metabolites, with PCa increased metabolites in the positive region of the horizontal axis. Points were colored and shaped depending on the metabolite family (A). Relation of the metabolites altered between PCa and BPH sample groups ordered alphabetically. Grayscale indicates the significance value of the difference: light gray <0.05, medium gray <0.01 and dark gray <0.001 (B). Heatmap depicting the full fold-change differences in each metabolite. Red values indicate elevation in PCa and green reduction. P-values are indicated in grayscale (C).*

Multivariate analysis by principal component analysis (PCA) did not show a perfect separation of the two groups, although PCa EV samples tended to aggregate all together, whereas BPH samples were more disperse (FIGURE 25). Statistics of the model indicate a low degree of fit (2n component R2X = 0.49) and also low predictability (2n component Q2X = 0.37). The PCA loadings plot (FIGURE 25) indicated that the differences between PCa and BPH samples were explained mainly by different subfamilies of glycerophospholipids, confirming what was identified with the univariate analysis.

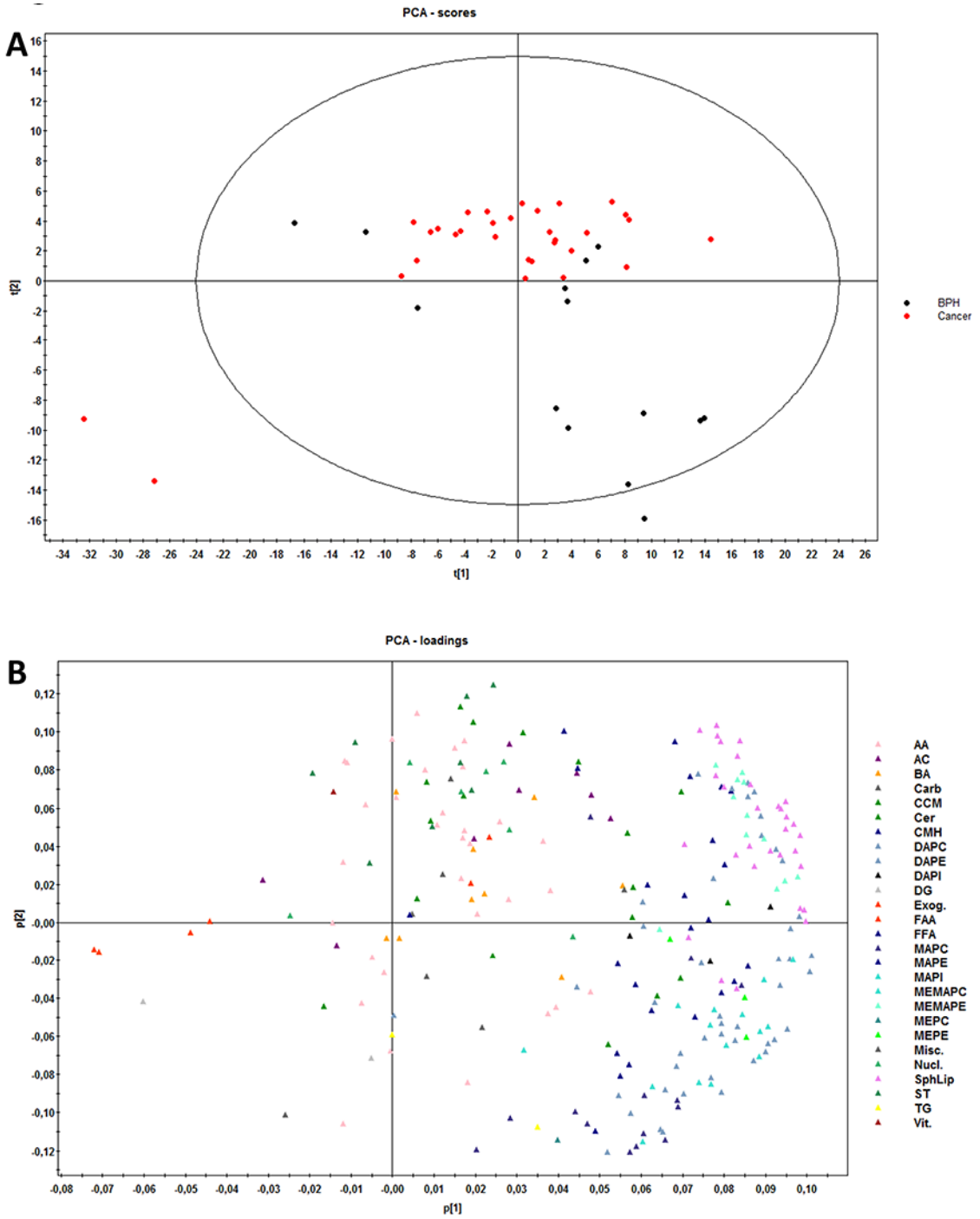


Figure 25: *Multivariate analysis for the comparison PCa (n = 31) vs BHP (n = 14). Dots in score plot (A) have been colored depending on its group (PCa or BHP). Markers in loadings plot (B) have been colored depending on the metabolite family. AA (amino acids), AC (acylcarnitines), BA (bile acids), Carb (carboxylic acids), CCM (derivative carboxylic acids), Cer (ceramides), CMH (monohexosylceramides), DAPC (diacylglycerophosphocholines), DAPE (diacylglycerophosphoethanolamines), DAPI (diacylglycerophosphoinositol), DG (diacylglycerols), Exog. (exogenous), FAA (fatty amides), FFA (non-esterified fatty acids), MAPC (1-monoacylglycerophosphocholine), MAPE (monoacylglycerophosphoethanolamine), MAPI (monoacylglycerophosphoinositol), MEMAPC (1-ether, 2-acylglycerophosphocholine), MEMAPE (1- ether, 2-acylglycerophosphoethanolamine), MEPC (1-monoetherglycerophosphocholine), MEPE (1-monoetherglycerophosphoethanolamine).*

3.2.3.2.- Metabolites differentially altered between PCa stage 2 and stage 3

PCa stage is a pathological sign of disease aggressiveness [245]. In an attempt to identify potential biomarkers to discriminate between different stages of PCa, we performed univariate analysis comparing the PCa stage 2 and stage 3 subgroups. We identified 5 metabolites that showed significant differences between the two groups (FIGURE 26A). These metabolites were three ceramides, Cer(d18:1/16:0), Cer(d18:1/20:0), Cer (d18:1/22:0) one glycerophospholipid PC(30:0) [which is a combination of the isomers PC(16:0/14:0) and PC (14:0/16:0)] and one acylcarnitine, stearoylcarnitine [AC(18:0)]. In addition, we also observed a non-significant trend in other metabolite families. Thus, fatty esters, glycerolipids (both diacylglycerols and triacylglycerols), fatty amides, vitamins and 1-monoetherglycerophosphocholines showed an increase in their abundance in the PCa stage 3 group (https://www.tandfonline.com/doi/full/10.1080/20013078.2018.1470442). In contrast, the levels of most of the metabolites belonging to the sphingolipids family including ceramides, monohexosylceramides and sphingomyelins, as well as fatty alcohols, some glycerophospholipids subgroups and nucleosides were reduced in stage 3. In this comparison, unsupervised multivariate analysis could not achieve any separation between different PCa stages, and although supervised PLS-DA analysis was able to discriminate (R2X 0.47, Q2X 0.07), its loadings plot showed that the major

influence in the separation corresponded to the aforementioned five metabolites (data not shown) detected in the univariate analysis.

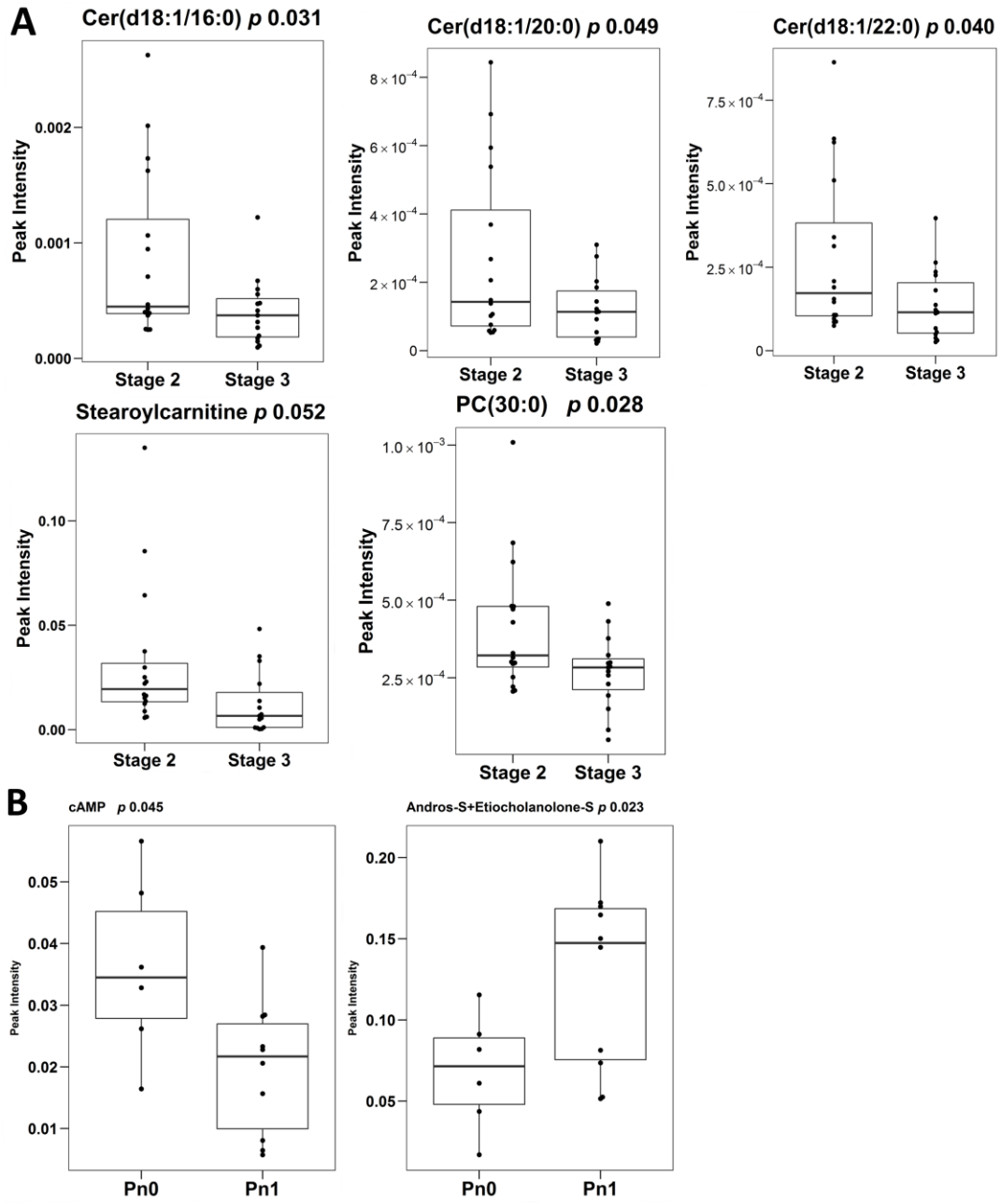


Figure 26: *Boxplots of the differentially expressed metabolites between PCa stages (A) and between presence and absence of perineural invasion (B). Stage 2 n = 16, Stage 3 n = 15, Pn0 = 6, Pn1 n = 10.*

3.2.3.3.- Metabolites differentially altered between PCa stage 2 perineural invasion: Pn1 vs Pn0

Perineural invasion in PCa has been associated with prostate cancer prognosis [266]. Although a limited number of samples were available, we also attempted to identify metabolites tentatively associated to this pathological feature. By univariate analysis, we detected significant lower abundance of cyclic AMP (cAMP) and a higher abundance of the combination of isomers androsterone sulfate and etiocholanolone sulfate in the EV samples obtained from PCa patients with perineural invasion (FIGURE 26B). In addition,

although not significant, three bile acids showed lower levels in samples with perineural invasion (https://www.tandfonline.com/doi/full/10.1080/20013078.2018.1470442). Unsupervised multivariate analysis was not able to separate the two groups of samples, but we could achieve this separation with PLS-DA test (R2X 0.40, Q2X 0.58) *(data not shown)*.

3.2.3.4.- Correlation analysis of metabolic profiling with body mass index (BMI)

When studying circulating metabolites, the systemic metabolic state can be a critical contributing factor that can influence the results of the analysis. Obesogenic diets and obesity impact on biofluid metabolite concentration, and can also have a central effect on tumor tissues [267] by altering their biological features. Therefore, we considered evaluating the changes in urine EV metabolites that were associated with the body mass index (BMI). Samples were divided into three groups, corresponding to their calculated BMI: lean (<25), overweight (>25 and <30) and obese (>30). Taking into account all the samples independently of their BPH or PCa classification, no significant correlation was found between BMI and any of the 248 metabolites analyzed in this study. Afterward, we explored if some metabolites were correlating with BMI inside different groups. In the lean BMI group, some sterol-related metabolites including isomer pregn-5-ene-3,20-diol sulfate and isomer androsterone sulfate showed significant positive correlations with *rho* values of 0.72 and 0.60, respectively (Table 3). On the contrary, diacylglycerol DG(36:3), PC(18:2/00) and triglyceride TG(56:3) showed significant negative correlations with *rho* values of −0.71, −0.69 and −0.67, respectively (Table 3). In the case of the overweight BMI group, a significant positive correlation was found with the exogenous metabolite, hydroxyphenyllactic acid (ρ 0.69). Sphingomyelin SM(43:1) showed a significant negative correlation (ρ −0.67) with BMI values (Table 3). In the obese group, we observed a high degree of correlation of some metabolites with the BMI values. Thus, acylcarnitine AC(8:0) (ρ 0.94) and arginine (ρ 0.85) showed a significant positive correlation, while 13 sphingomyelins, 8 phosphatidylethanolamines and the polyunsaturated fatty acid (16:3n-3) showed negative correlations with *rho* values ranging between −0.95 to −0.78) (Table 3). Finally, we evaluated if any of the metabolites correlated with BMI considering only the PCa group. Inside this group, the highest positive correlations were found for taurocholic acid and dodecanoylcarnitine, AC(12:0), with *rho* values of 0.51 and 0.38, respectively.

Table 5: *Correlation analysis between metabolites and BMI measurements.*

	Metabolite	Class	Correlation (ρ)	*p-value*
Lean	Isomer pregn-5-ene-3,20-diol sulfate	Sterol	0.72	0.003
	Isomer androsterone sulfate	Sterol	0.60	0.011
	Taurodeoxycholic acid	Bile acid	0.59	0.03
	Malate	Carboxylic acid (d)	0.56	0.025
	Arginine	Amino acid	0.53	0.027
	Glycine	Amino acid	-0.56	0.021

	TG(18:1+20:1+18:1)	Glycerolipid	-0.67	0.006
	PC(18:2/0:0)	Glycerophospholipid	-0.69	0.007
	DG(36:3)	Glycerolipid	-0.71	0.008
Overweight	Hydroxyphenyllactic acid	Benzyl alcohol (d)	0.69	0.001
	L-citruline	Amino acid (d)	0.59	0.008
	Vitamin B5	Vitamin	0.52	0.024
	Proline	Amino acid	0.50	0.030
	DG(34:1)	Glycerolipid	0.49	0.035
	4-Pyridoxic acid	Pyridine (d)	0.49	0.036
	PC(O-16:0/20:4)	Glycerophospholipid	0.47	0.042
	PE(18:0/18:1)	Glycerophospholipid	-0.47	0.046
	SM(d18:1/17:0)	Sphinomyelin	-0.48	0.038
	Stearoylcarnitine	Acyl carnitine	-0.49	0.037
	PE(P-18:0/18:1)	Glycerophospholipid	-0.50	0.030
	PE(16:0/18:2)	Glycerophospholipid	-0.51	0.027
	PE(0:0/20:3)	Glycerophospholipid	-0.51	0.027
	PE(P-16:0/18:2)	Glycerophospholipid	-0.53	0.021
	Alpha-Ketoglutarate	Keto-acids (d)	-0.54	0.028
	PE(18:1/18:2)	Glycerophospholipid	-0.55	0.017
	PE(P-18:0/18:2)	Glycerophospholipid	-0.56	0.013
	AC(12:1n-x)	Fatty esters	-0.57	0.014
	PE(18:2/18:2)	Glycerophospholipid	-0.57	0.016
	SM(43:1)	Sphingomyelin	-067	0.003
Obese	L-Octanoylcarnitine	Acyl carnitine	0.94	0.017
	Arginine	Amino acid	0.86	0.024
	PC(O-16:0/18:2)	Glycerophospholipid	0.83	0.058
	Acylcarnitine(8:1n-x)	Acyl carnitine	0.75	0.066
	Deoxycholic acid	Bile acid	0.75	0.066
	PI(18:0/20:4)	Glycerophospholipid	-0.75	0.066
	PE(20:5/16:0)	Glycerophospholipid	-0.75	0.066
	L-Homoserine	Amino acid	-0.75	0.066
	Isoleucine	Amino acid	-0.75	0.066
	SM(43:1)	Sphingomyelin	-0.79	0.048
	SM(d18:1/24:1) + SM(d18:2/24:0)	Sphingomyelin	-0.79	0.048
	SM(d18:1/17:0)	Sphingomyelin	-0.79	0.048
	SM(33:1)	Sphingomyelin	-0.79	0.048
	PE(P-18:0/22:5) + PE(P-20:1/20:4)	Glycerophospholipid	-0.79	0.048
	PUFA (16:3n-x)	Fatty acid	-0.79	0.048
	PE(20:4/18:2)	Glycerophospholipid	-0.79	0.048
	SM(32:1)	Sphingomyelin	-0.82	0.034
	PE(P-16:0/20:4)	Glycerophospholipid	-0.82	0.034
	PE(0:0/22:4)	Glycerophospholipid	-0.82	0.034
	SM(d18:2/22:0)	Sphingomyelin	-0.86	0.024
	SM(d18:1/22:0)	Sphingomyelin	-0.86	0.024
	SM(d18:1/18:0)	Sphingomyelin	-0.86	0.024
	SM(d18:1/16:0)	Sphingomyelin	-0.86	0.024
	PE(18:0/20:4)	Glycerophospholipid	-0.86	0.024
	PE(18:1e/22:4)	Glycerophospholipid	-0.88	0.008
	SM(d18:2/20:0)	Sphingomyelin	-0.89	0.012
	PE(16:0/22:6)	Glycerophospholipid	-0.89	0.012
	SM(42:1)	Sphingomyelin	-0.93	0.007
	SM(d16:1/24:1)	Sphingomyelin	-0.93	0.007
	PE(16:0/20:4)	Glycerophospholipid	-0.93	0.007
	SM(38:1)	Sphingomyelin	-0.96	0.003

3.2.3.5.- Correlation analysis of metabolic profiling with PSA in the PCa group

PSA is the current gold standard non-invasive prognostic marker for PCa while its diagnostic potential remains controversial [268]. We performed a correlation analysis between urinary EV metabolites and the PSA values determined in our cohort of PCa samples. We only observed a significant positive correlation (*rho* value 0.88) of phosphatidylcholine PC(0:0/20:3), and at less extent (*rho* value 0.48) of the primary fatty amide (20:2n-x).

3.2.3.6.- Analysis of enzymes-associated to metabolites differentially expressed between PCa and BPH

We have recently shown that metabolic alterations in PCa are frequently associated with changes in the expression of key enzymes [264]. To better understand the cancer cell-autonomous nature of the metabolite changes observed in urine EVs from PCa patients, we mapped the 76 altered urinary-EV-metabolites into cellular pathways by using MetScape v3.1.2 [269]. We identified several pathways that could be affected in PCa including steroid hormone biosynthesis and metabolism, leukotriene and prostaglandin metabolisms, linoleate and purine metabolisms, glycerophospholipid metabolism, TCA and urea cycle, and tryptophan metabolism. We identified the primary enzymes involved in the metabolism of each of the 76 differentially expressed metabolites between BPH and PCa, by using KEGG or HMDB database (see Methods). A complete list of primary enzymes is supplied as Supplementary Material (https://www.tandfonline.com/doi/full/10.1080/20013078.2018.1470442). Next, we took advantage of publicly available prostate cancer transcriptomes and we queried the expression of the 149 enzymes in PCa. We searched for enzymes which expression changes in PCa would fit the metabolite abundance observed in urine EVs. From these gene list, we identified 7 genes with the expression changes (FIGURE 27) that were concordant with the observed changes in urine EV metabolite abundance among the groups analysed. We found gamma-aminobutyric acid (GABA) increased in PCa urine EVs (FIGURE 24) which was consistent with a reduction in the expression of Glycine Amidinotransferase (GATM- use GABA as substrate for creatine synthesis) (FIGURE 27A). Arachidonic acid abundance was also altered in urine EV samples, being reduced in PCa patients compared with BPH (FIGURE 24). This fatty acid is the product of phospholipase A2 and it is relevant for the synthesis of proinflammatory metabolites by lipooxygenases. Interestingly, we found that the expression of two enzymes (ALOX15 and CYP1A2), that can catabolise arachidonic acid, was increased in PCa tissue (FIGURE 27B,C). Our metabolomics analysis also showed a consistent decrease in phosphatidylcholine. This could be explained by decreased synthesis of the phospholipid or elevated catabolism. When browsing the expression of PC synthesis and degrading enzymes, we found a reduction in the expression of Lysophosphatidylcholine Acyltransferase 2 (LPCAT2) (FIGURE 27D), which transforms lysoPC into PC, and could provide an explanation for the reduction in PC abundance.

Two urine EV metabolites were associated with increased perineural invasion in PCa. On the one hand, we found a decrease in cAMP abundance in EV obtained patients with perineural invasion. The transcriptional analysis revealed changes in the expression of enzymes regulating cAMP synthesis and degradation that were associated with the aggressiveness of

the disease. The expression of adenylate cyclase 5 (ADCY5) was reduced in PCa (FIGURE 27E-G), and a further significant reduction was observed from primary tumors to metastasis. In contrast, the inverse expression pattern (elevation in PCa and further increase from primary tumors to metastasis) was detected in the cAMP degrading enzyme PDE4C (FIGURE 27H-J). In none of these cAMP metabolizing enzymes we could find an association to altered disease-free survival (FIGURE 27G,J).

On the other hand, the steroid biosynthesis-related metabolites were among the most elevated in PCa urine EVs and associated with increased perineural invasion. Interestingly, the three metabolites significantly altered were sulfated steroids in the final steps of androgen synthesis. Whereas these metabolites are found at detectable levels in circulation produced by the adrenal gland, we evaluated whether enzymes regulating their synthesis or degradation could be altered in PCa tissue. Strikingly, we found that the expression steroid sulfatase (STS), which would remove the sulfate group in androsterone sulfate and DHEAS, was decreased in PCa, and this reduction was associated to metastatic disease and reduced disease-free survival in one out of two datasets (FIGURE 27K-M).

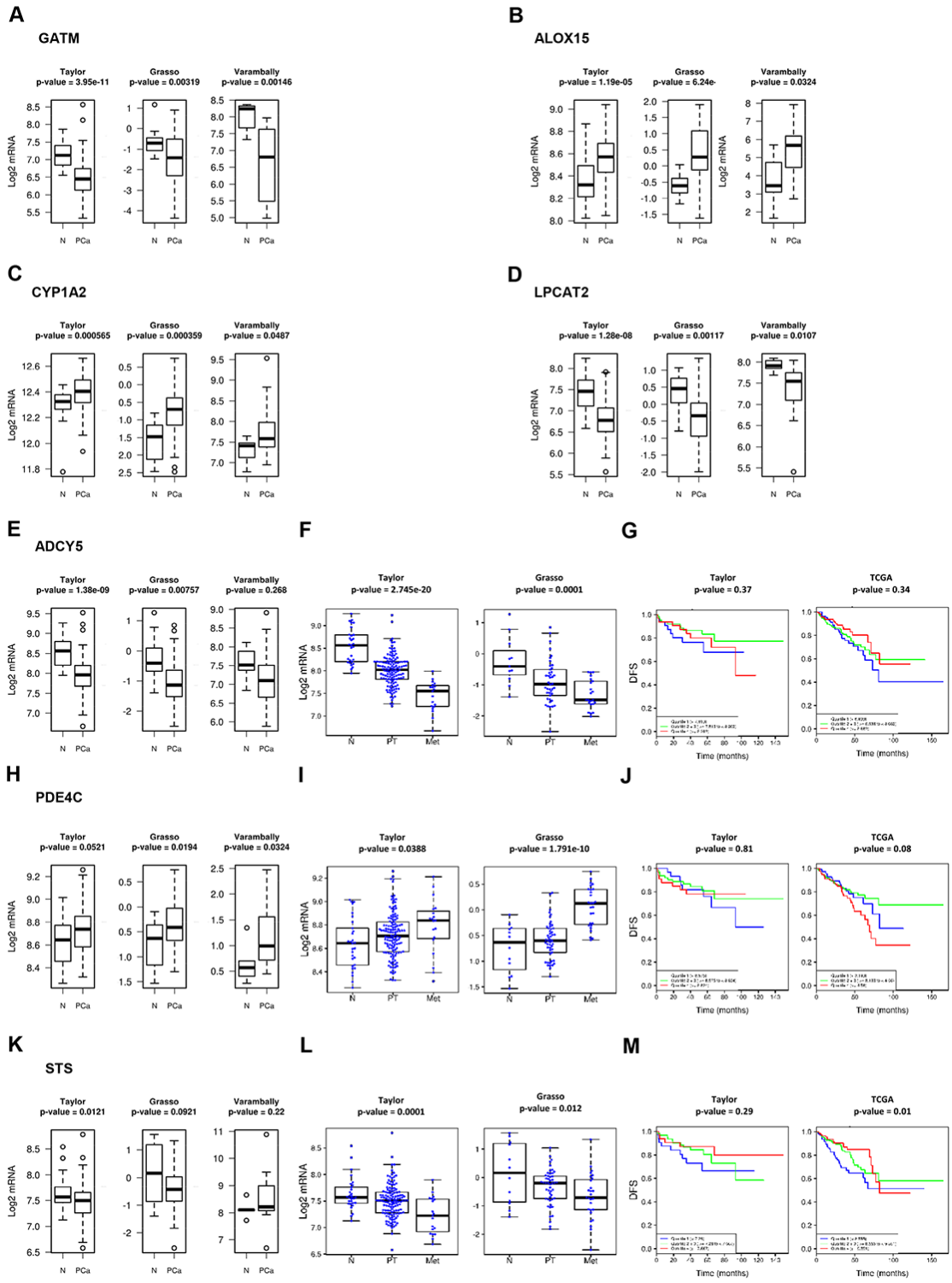


Figure 27: *Gene-enrichment analysis. In-silico transcriptomics analysis of enzymes directly involved in the metabolism of metabolites differentially expressed between PCa and BPH samples.*

3.2.4.- Discussion

EVs are produced by normal and cancerous cells and harbor molecular features of their cells of origin [270]. This encapsulated material can exert biological and metabolic functions

[247, 257, 258, 271], which makes them entities of tremendous interest in cancer biology, both at the level of biomarker discovery and mechanistic. Urine contains EVs from different parts of the urinary tract including kidney and bladder what has awaked great interest to identify biomarkers affecting these organs. In addition, the anatomic proximity of urine to the prostate gland and the already shown presence of tumor cells in the urine sediment [272, 273] support also the development of potential non-invasive diagnoses of PCa using urine-based markers. In agreement with our previous results, we find differences in the size distribution of the urinary EVs between PCa and BPH [259], which we now report to be associated to disease stage (FIGURE 23). Our data show that urine from advanced PCa patients contains a higher proportion of large EVs than BPH patients. Given that our EV isolation procedure (filtration through 0.22 microns, and ultracentrifugation at 10,000×g) removed most of the large EVs from the sample, and enrich in small EVs (mostly exosomes and small microvesicles), this difference could be underestimated in our samples. Importantly, in agreement with our result, it has already been reported that prostate cancer cells release large EVs named oncosomes with a size between 1 and 10 microns [274] that contain a distinct protein cargo [275]. They have also been detected in circulation in models of PCa and shown that their abundance correlates with tumoral progression [274]. Although our studies have been focused on the smaller EVs, it is interesting that we have also observed this size effect.

In a recent targeted lipidomics analysis of urinary EVs from healthy and PCa urine samples [256], the authors analyzed 107 lipid species and found that 9 of them were significantly different between the two groups. Unlike this study, we have focused ours in the comparison between PCa and BPH, in an attempt to provide specific biomarkers to discriminate the two pathological conditions, and contribute to earlier diagnosis, and reduce secondary effects of unnecessary biopsies, so both studies can be considered complementary in terms of sample groups. Both studies are also complementary in the metabolites that they analyze because different metabolite extraction methods and chromatographic procedures were used.

We report changes in the urine EV metabolome at both structural and cargo levels. The composition of the urine EVs analyzed in this study varies in the abundance of phosphatidylcholine species that are major constituents of membranes. In particular, we found reduced abundance of PCs in the EVs from PCa samples, in agreement with previously reported by Puhka and coworkers [276]. This result along with studies reporting increased abundance of PC in PCa tissue [277] could suggest that less PC-containing structures, like membrane vesicles are secreted to the extracellular environment. In addition to the PC content, we found additional metabolites from different chemical nature differentially expressed in EVs from PCa and BPH samples that could be considered candidate biomarkers for PCa including as candidates acylcarnitines, sphingomyelins, and steroids. Although more research is granted, our results indicate that bias in EV size and membrane composition could harbour diagnostic potential in PCa.

Apart from the potential biomarker value of the identified metabolites, they are also valuable to indicate possible metabolic alterations occurring in PCa. We found reduced levels in PCa urine EVs of arachidonic acid, the precursor of eicosanoids and prostaglandins that are important proliferative and inflammatory modulators. Interestingly, it has been also reported that arachidonic acid level is lower in prostatic tissue from PCa patients [278]. In agreement with the reduction of the substrate arachidonic acid in PCa, it has been found that the level of their products (12-and 20-HETE, and PGE2) are higher in the tissue [279, 280] and also in urine [281]. These studies along with many others have already shown that the metabolism of arachidonic acid and their products plays an important role in PCa development, and in fact, represents an important therapeutics target (reviewed in [282]). Importantly, our work suggests that the analysis of this metabolite in EVs isolated from urine samples may be used to evaluate in a non-invasive manner what is occurring in prostatic tissue itself in the context of PCa.

We observed changes in the abundance of metabolites that are carried within the EVs and are a potential cargo in PCa. It is worth mentioning that intermediary metabolites of androgen synthesis were among the most elevated in PCa urine EVs. Moreover, changes in the abundance of these steroids, together with cAMP, were significantly associated with perineural invasion. These results uncover the potential of unbiased urine EV analysis to elucidate novel signaling and metabolic alterations underlying PCa biology. Androgen signaling is among the predominant stimuli supporting PCa growth and the most successful therapeutic approaches have derived from its targeting [283] since prostate tumors frequently remain androgen dependent even at late-stage [284]. We have detected 3beta-hydroxyandors-5-en-17-one-3-sulphate (dehydroepiandrosterone sulphate, DHEAS) in urinary EVs, and its level was significantly elevated in PCa samples. This metabolite, along with estrone sulphate, is one of the main precursors for steroid hormones including androgens. There are many reports showing that steroid-related metabolites and enzymes are important modulators of PCa progression [285]. There are four different genes coding for enzymes that were related to this metabolite: STS, SULT1B1, SULT2B1 and SULT2A1. The fact that urine EVs from PCa patients contain androgen-related metabolites is suggestive of the relevance of this biosynthetic pathway in the disease and the potential role of EVs in providing androgen signalling to neighbour or distal cells. Indeed, expression of STS was reduced in PCa and associated to disease progression, hence providing a feasible explanation for the increase in sulfated steroids. Interestingly, urinary EVs could be used to monitoring androgen metabolism in a non-invasive manner.

Together with the aforementioned metabolites associated with perineural invasion, we also identified molecules that exhibited differential abundance in high-grade tumours. Five metabolites were differentially abundant between pathological stage 2 and stage 3 PCa, and more than half of them were ceramide species. Ceramides are signaling molecules that can regulate various aspects of cancer cell biology, including proliferation, survival and cell death [286]. The selective decrease of ceramides in association with disease aggressiveness provides an exciting perspective of how this family of metabolites could exert cell and non-

cell autonomous functions to limit the progression of PCa. It is worth noting that sarcosine has been proposed also as a PCa biomarker [255]. The urine level of this metabolite was increased in men with metastatic PCa [287]. However, its utility as a potential diagnostic tool is unclear, as its validation as a biomarker has failed in several studies (reviewed in [255, 74]. Interestingly, we have detected sarcosine in urinary EVs, and although not significant ($p = 0.09$), its level was decreased in PCa samples.

Recent molecular and metabolic profiling of PCa also identifies lipid metabolism as a key pathway that undergoes metabolic reprogramming [288, 289]. These changes include an upregulation metabolites involved in de novo lipid biosynthesis [290] and fatty acid β-oxidation [291]. As a consequence, it has been shown the accumulation in the prostatic tissue of acylcarnitines, which are intermediates of fatty acid oxidation [292]. In agreement with this alteration, we found increased levels of acylcarnitines in the urinary EVs from PCa patients. This association of differential levels of carnitines on PCa EVs with a metabolic shifting towards β-oxidation of fatty acids has already been proposed by Puhka and coworkers [276].

In summary, in this work, we report several metabolites associated with urinary EVs, many of them exhibiting differential abundance between BPH and PCa, and mirroring some of the alterations described in PCa.

3.3.- Chapter 3. General considerations on microbiome

3.3.1.- Introduction

As we have previously discussed, several aspects need to be considered when performing microbiome-related projects. In this chapter we will discuss the 16S rDNA region specificity regarding the identified bacteria and data differences.

During the development of this thesis project, two projects were done that involved 16S rDNA sequencing, as we have presented before. Due to technical reasons, two different sequencing services were used, each one specialized in sequencing distinct regions of the 16S gene. Thus, in order to check the consistency of the sequences obtained by each sequencing service, we performed a small comparison in which 4 samples were sequenced twice, first by sequencing the V3-V4 regions and then the V1-V2 regions of the 16S rDNA gene. To this comparison, we used fibromyalgia project samples, choosing 2 control samples and 2 fibromyalgia ones.

3.3.2.-Methods

3.3.2.1.- 16S rDNA region sequencing

Feces samples were delivered to the corresponding hospital by individuals recruited during fibromyalgia project cohort construction. Feces samples were then derived to the Basque Biobank, were DNA extraction was performed using PSP Spin Stool DNA Plus kit (STRATEC Molecular®), following the manufacturer's protocol. Lysis buffer was added to the frozen feces samples to avoid nucleic acids degradation before extraction was performed. Once DNA extraction was performed, samples were aliquoted into 2.5µg of DNA at 100ng/µL concentration aliquots and frozen until sequencing.

3.3.2.1.1.- V3-V4 samples

The 4 samples dedicated to the 16S rDNA distinct sequenced regions comparison aliquots were then split into two parts, so that one was used to sequence the V3-V4 region and the other one for the V1-V2 sequencing. V3-V4 regions sequencing was performed by CIC bioGUNE's genomic platform, in collaboration with FISABIO Sequencing Core Facility. DNA amplicon libraries were generated and sequenced following Illumina Inc's recommendations. V3-V4 surrounding primers pair were selected, leading to a 459bp length amplicons [293]. Amplification reaction methods are detailed in Table 4.

Then, Illumina Inc.'s sequencing adaptors and dual-index barcodes (Nextera XT index kit v2, FC-131-2001) were added to each amplicon and, after PCR purification, libraries were normalized and pooled prior to sequencing. The pool containing indexed amplicons was loaded onto the MiSeq reagent cartridge v3 (MS-102-3003), spiked with 25% PhiX control to improve base calling during sequencing, as recommended by Illumina for amplicon sequencing. Sequencing was conducted using a paired-end, 2x300pb cycle run on an Illumina MiSeq sequencing system. Sequencing was done by FISABIO Sequencing Core Facility, who also performed the quality assessment,

using *prinseq-lite* [167] with the following parameters (min_length: 50, trim_qual_right: 30, trim_qual_type: mean, trim_qual_window: 20), and the sequence joining, with *FLASH* software [166] using default parameters.

Table 6: *Amplification protocol for the V3-V4 16S rDNA region.*

			Volume
Microbial DNA (5 ng/µl)			2.5 µl
Amplicon PCR Forward Primer 1 µM 5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG			5 µl
Amplicon PCR Reverse Primer 1 µM 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC			5 µl
2x KAPA HiFi HotStart ReadyMix (KK2602)			12.5 µl
Total			**25 µl**
PCR cycles protocol			
Step		**Temperature**	**Time**
Denaturation		95ºC	3 min
Annealing	x25 cycles	95ºC	30 sec
		55ºC	30 sec
		72ºC	30 sec
Extension		72ºC	5 min

Joined reads were then uploaded to QIIME2 (v2019.4), demultiplexed and clustered de novo into OTUs at 99% similarity threshold. Common alpha and beta diversity measurements were computed, both phylogenetic and non-phylogenetic ones. Taxonomical annotation was later performed on the OTU table using the GreenGenes database (v13_8).

3.3.2.1.2.- V1-V2 sequencing

Variable regions V1 and V2 of the 16S rRNA gene were amplified using the primer pair 27F-338R in a dual-barcoding approach according to Caporaso *et al*. [294]. DNA was diluted 1:10 prior PCR, and 3 µl of this dilution were finally used for amplification. PCR-products were verified using the electrophoresis in agarose gel. PCR products were normalized using the SequalPrep Normalization Plate Kit (Thermo Fischer Scientific, Waltham, MA, USA), pooled equimolarily and sequenced on the Illumina MiSeq v3 2x300bp (Illumina Inc., San Diego, CA, USA).

Raw reads were then uploaded to QIIME2 (v2019.4), were they were demultiplexed and joined using default configuration. Then they were clustered de novo into OTUs at 99% similarity threshold. Common alpha and beta diversity measurements were computed, both phylogenetic and non-phylogenetic ones. Taxonomical annotation was later performed on the OTU table using the GreenGenes database (v13_8).

3.3.2.1.3.- Comparative analysis

Both OTU tables were joined in QIIME2 and processed again for diversity differences. PCoA was computed upon several diversity indexes. Procrustes Analysis and Mantel's test were also performed for Bray-Curtis, weighted and unweighted UNIFRAC distances to characterize the similarity of samples sequenced by different methods.

3.3.3.-Results

The reads obtained for each sample and each region sequenced, with the number of reads remaining after each quality control step are summarized in Table 7.

Table 7: *Number of reads obtained per sample and 16S regions sequenced. The number of remaining reads after each quality step is indicated and the proportion representing relative to the initial number of reads is indicated between parentheses. V1-V2 regions sequencing are shaded in green, while V3-V4 regions are not shaded.*

	Region	Input	Filtered	Denoised	Non-chimeric
FIBR007F	V3V4	**51,258**	50,476 (98.47%)	49,685 (96.93%)	**23,110 (45.09%)**
	V1V2	**45,850**	39,550 (86.26%)	37,592 (81.99%)	**32,619 (71.14%)**
FIBR011C	V3V4	**42,985**	42,425 (98.70%)	41,636 (96.86%)	**24,127 (56.13%)**
	V1V2	**40,143**	34,451 (85.82%)	33,297 (82.95%)	**28,433 (70.83%)**
FIBR020C	V3V4	**61,493**	60,530 (98.43%)	59,081 (96.08%)	**29,517 (48.00%)**
	V1V2	**68,048**	58,480 (85.94%)	55,608 (81.72%)	**47,340 (69.57%)**
FIBR069F	V3V4	**50,361**	49,688 (98.14%)	48,664 (96.12%)	**25,584 (50.53%)**
	V1V2	**77,557**	66,491 (85.73%)	63,794 (82.25%)	**53,210 (68.61%)**

In general, V1-V2 regions sequencing lead to less chimeric reads, so that the proportion of sequences that passed the quality control was higher than in V3-V4 region sequencing. Thus, more reads per sample of good quality that were later used for posterior analysis were obtained from V1-V2 sequencing than for the V3-V4 16S rDNA region sequencing.

3.3.3.1.1.- V1 – V2

161,602 total reads remained after the quality check, representing 1,313 distinct features. Reads per sample were distributed as follows: a minimum of 28,433 features, a maximum of 53,210 and, on average 40,400.5 reads/sample, with a median of 39,979.5.

3.3.3.1.2.- V3 – V4

In total, 102,338 reads passed the quality control for this region sequencing, representing 690 distinct features. The median frequency of reads per sample was 24,855.5, with an average of 25.584.5 reads/sample, being the minimum number of reads/sample 23,110 and the maximum 29,517.

3.3.3.1.3.- Joined samples

We measured several α and β-diversity indexes in order to identify whether we could discriminate the samples depending on the rDNA regions sequenced. We saw that independently on which diversity index used, PCoA analysis was able to clearly discriminate the samples depending on the 16S rDNA region sequenced (FIGURE 28).

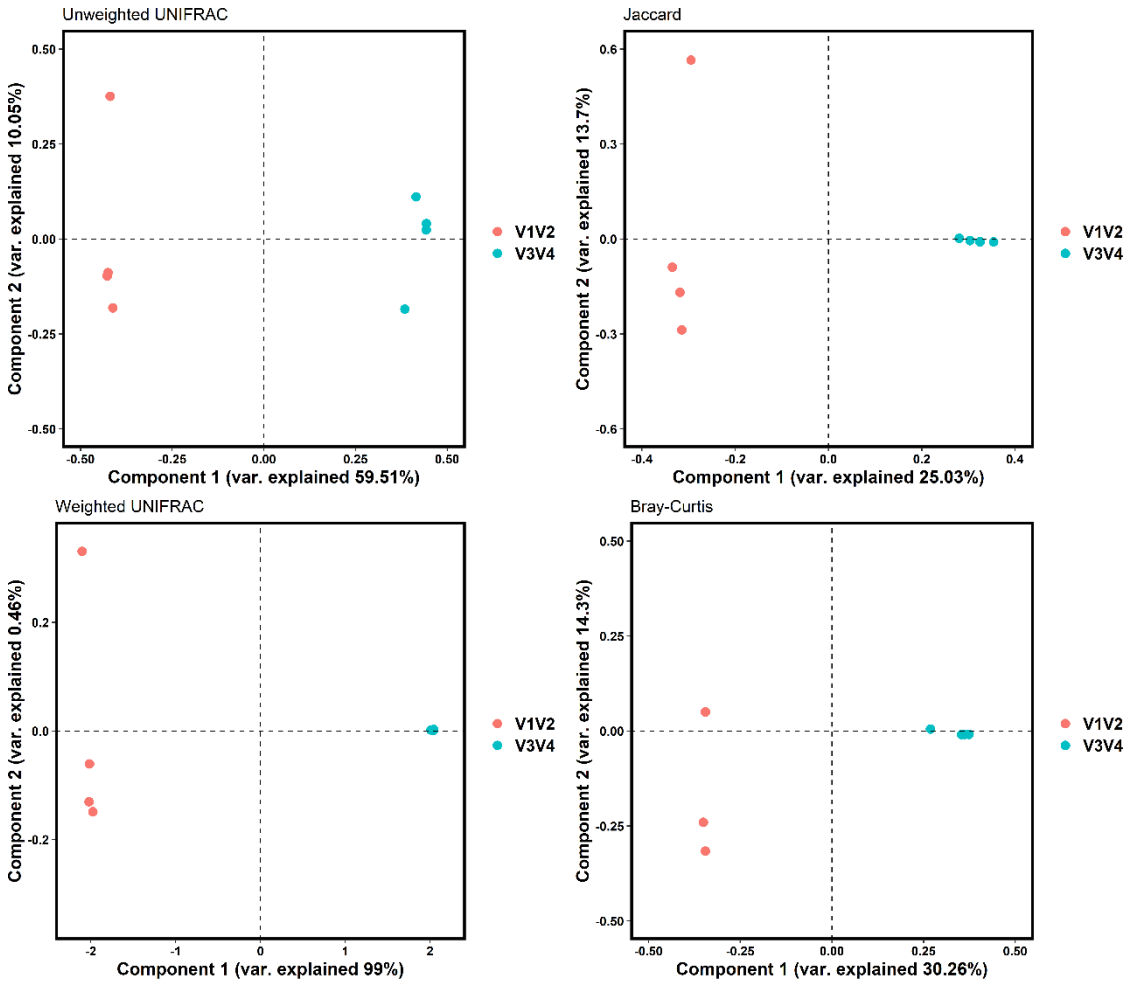


Figure 28: *Beta-diversity measurements of both V1-V2 and V3-V4 amplicons sequencing. Points are colored depending on the region sequenced.*

The taxonomical annotation of both sequencing methods revealed that V3-V4 identified fewer different OTUs than the V1-V2 region sequencing. Consistently, non-weighted alpha-diversity indexes were found to be higher for V1-V2 region sequences. When equilibrated diversity indexes were computed, though, no relevant difference was found between the two regions sequenced (FIGURE 29).

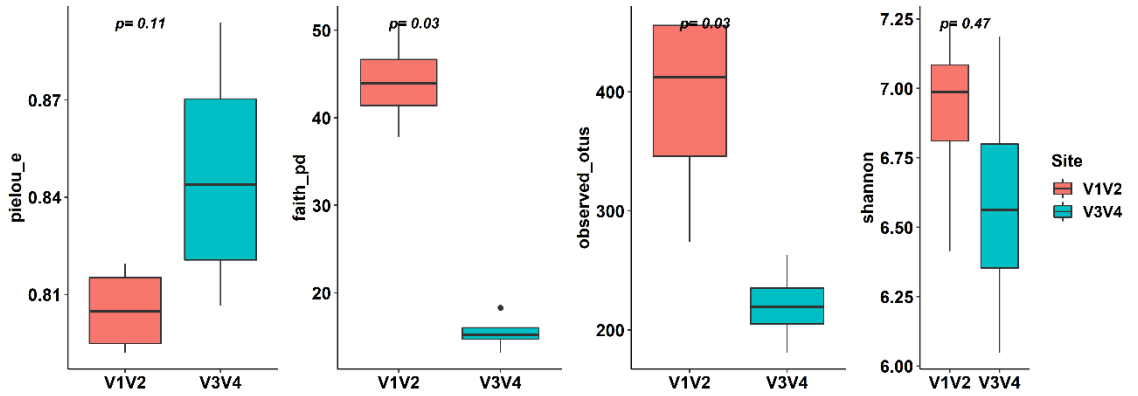


Figure 29: *Alpha-diversity indexes for the V1-V2 and V3-V4 amplicons. Boxplot filling color depends on the 16S rDNA region sequenced.*

Taxonomic analysis revealed that V3-V4 amplicons identified at least one OTU from the Archaea kingdom, while no Archaea sequences were identified for V1-V2 sequencing regions (FIGURE 30).

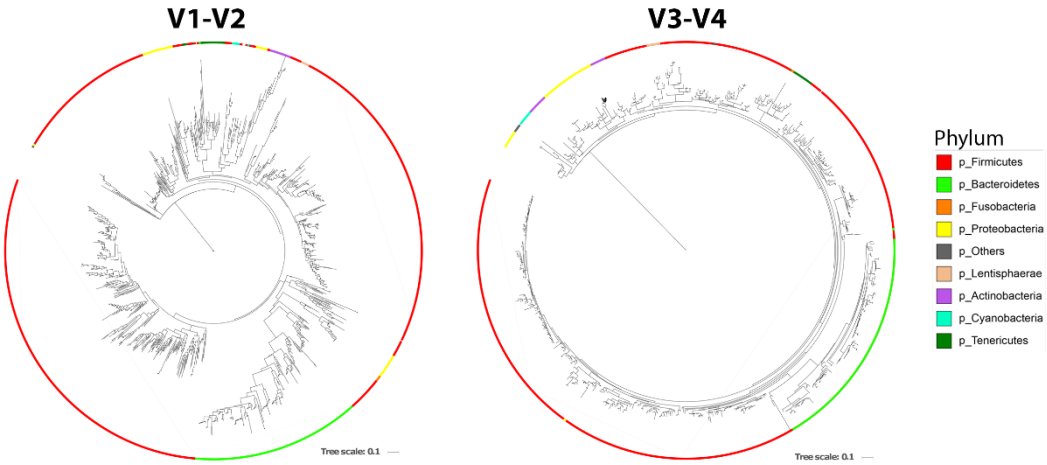


Figure 30: *Rooted phylogenetic trees in circular format display obtained from the quality checked reads for both V1-V2 and V3-V4. Colored ring surrounding each tree indicates the phyla of each branch of the tree. Archaea branch has been removed from the V3-V4 tree for easier visualization reasons.*

As can be seen in FIGURE 30, for both sequenced regions the majority of reads obtained corresponded to the Firmicutes phylum. After that, V3-V4 captured a higher diversity, with more reads mapping into different phyla than V1-V2 regions sequencing. Notably, V3-V4 sequencing regions captured a higher number of OTUs related to both Bacteroidetes, Proteobacteria and Actinobacteria. V1-V2 regions, instead, had more OTUs pertaining to Tenericutes phylum (FIGURE 31). Although these differences in abundance observed, both primers set displayed similar microbiota composition patterns in each sample. At more detailed taxonomical levels (family and genus), more differences were observed, as expected.

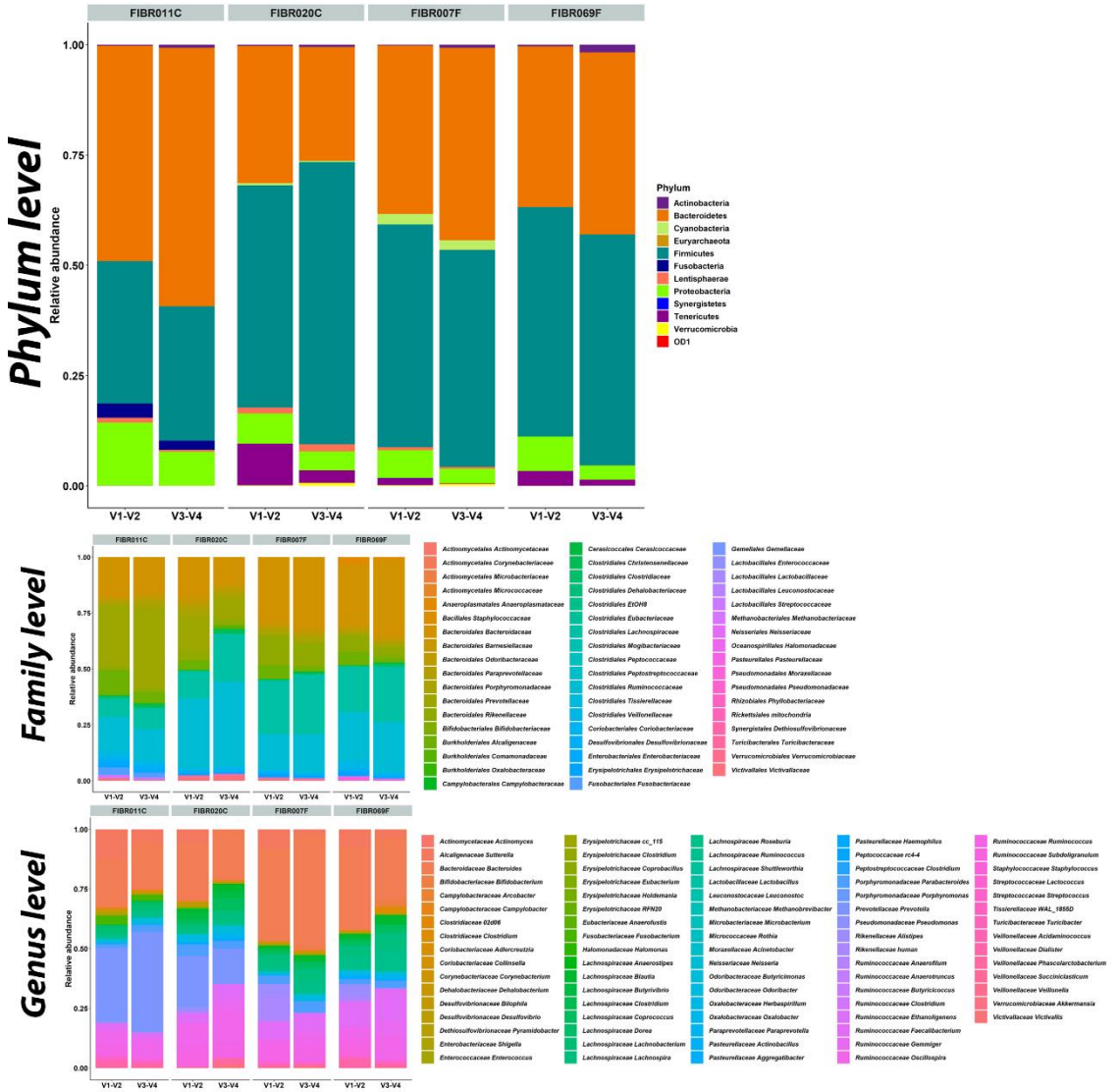


Figure 31: *Relative abundance comparison of each sample per each region sequenced at phylum, family and genus taxonomical levels.*

When the Procrustes analysis was performed, we found that the similarity between the V3V4 and V1V2 distance matrices obtained by three β-diversity indexes (bray-curtis, unweighted and weighted UNIFRAC) was notably higher (FIGURE 32). Specifically, all correlation scores between datasets were over 0.99, while the CIA analysis also showed that, for each index, RV scores were above 0.90. Monte-Carlo calculations on the significance of these similarities showed that weighted UNIFRAC (FIGURE 32C) was the only measurement that resulted to be statistically significant (p-value 0.042). The other two indexes didn't reach the significance threshold, something that may be explained by the reduced number of samples included in the study.

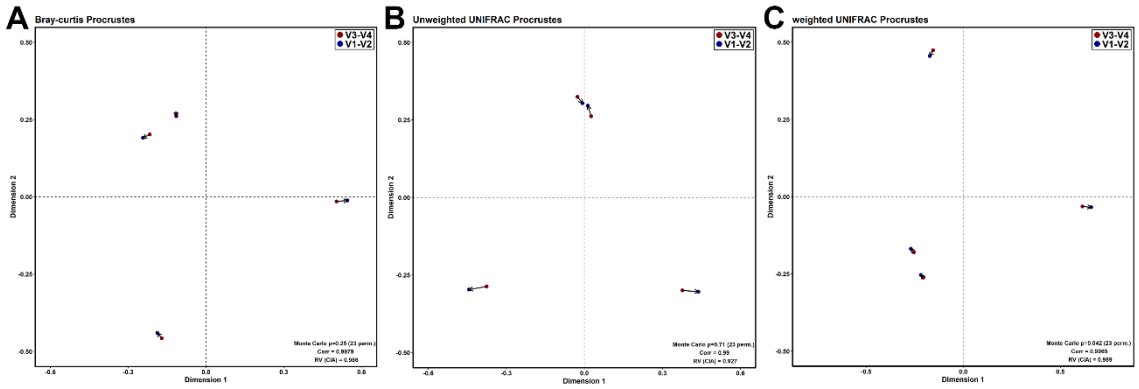


Figure 32: *Procrustes analysis for the comparison of 16S rDNA amplicons. In red, V3-V4 amplicons and in blue the V1-V2. Arrows connect the two amplicons sequenced from the same sample.*

We then performed a compositional data approach, by means of ALDEx2 pipeline, in order to identify which genera were the most different between each sequenced regions (FIGURE 33). We saw that 6 genera were differentially represented between the sequences obtained from V3V4 amplicons and the ones obtained with V1V2 ones: *Ruminococcaceae Gemmiger, Bifidobacteriaceae Bifidobacterium, Lachnospiraceae Coprococcus, Rikenellaceae Alistipes, Staphylococcaceae Staphylococcus* and *Ruminococcaceae Subdoligranulum*.

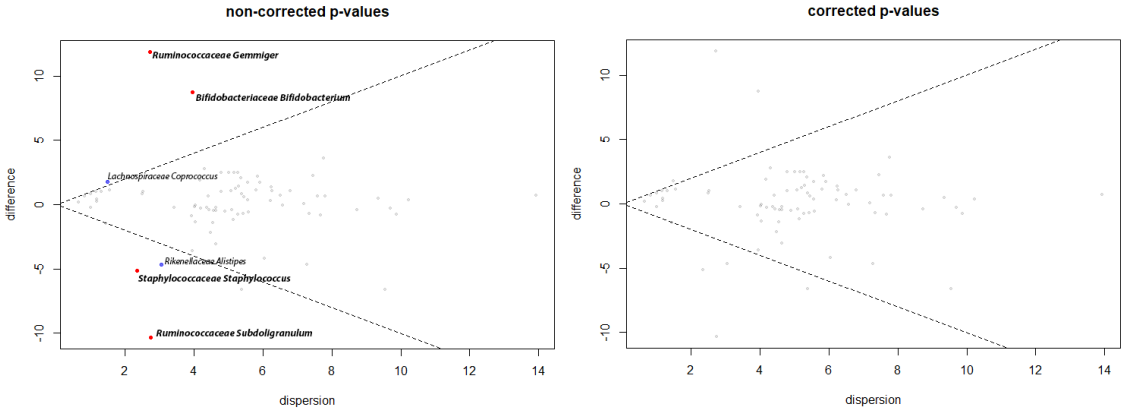


Figure 33: *Compositional data analysis for the genus differences between V3-V4 amplicons and V1-V2 ones. In the left, non-corrected p-values, in the right, corrected ones. Grey points represent abundant non-differential features, black points the non-differential rarely abundant features, blue dots the features identified as significantly different by one test (t-test or Wilcoxon) and red ones the significantly different features identified by both tests.*

3.3.3.1.4.- Functional differences

Finally, we applied PICRUSt2 in order to identify potential differences in the functional capabilities of the distinct microbiota profiles obtained by each 16S rDNA region sequenced. From the pathway abundance data, we identified 34 pathways that were only identified by one of the two sequencing options, 17 in each case (TABLE 8). Interestingly, most of the V3-V4 primers set differential pathways were related to archaeal metabolic functions, thus related to the archaea OTUs amplified by those primers. Methane-related pathways were also identified by V3-V4, indicating thus increased capability for the amplification of anaerobic bacteria.

Table 8: *Amplicon-specific inferred bacterial metabolic pathways from PICRUSt2 tool.*

V1-V2	V3-V4
Creatinine degradation I	methanogenesis from H2 and CO2
Phospholipases	protocatechuate degradation I (meta-cleavage pathway)
Vitamine E biosynthesis (tocopherols)	superpathway of taurine degradation
Sucrose degradation II	superpathway of aerobic toluene degradation
2-nitrobenzoate degradation I	superpathway of bacteriochlorophyll a biosynthesis
2-amino-3-carboxymuconate semialdehyde degradation to 2-oxopentenoate	archaetidylserine and archaetidylethanolamine biosynthesis
L-tryptophan degradation IX	tetrahydromethanopterin biosynthesis
superpathway of CDP-glucose-derived O-antigen building blocks biosynthesis	chorismate biosynthesis II (archaea)
mycolyl-arabinogalactan-peptidoglycan complex biosynthesis	flavin biosynthesis II (archaea)
2-heptyl-3-hydroxy-4(1H)-quinolone biosynthesis	mevalonate pathway II (archaea)
superpathway of quinolone and alkylquinolone biosynthesis	CDP-archaeol biosynthesis
isopropanol biosynthesis	archaetidylinositol biosynthesis
1,5-anhydrofructose degradation	phosphopantothenate biosynthesis III
protein N-glycosylation (bacterial)	7-(3-amino-3-carboxypropyl)-wyosine biosynthesis
ergothioneine biosynthesis I (bacteria)	sucrose biosynthesis III
superpathway of demethylmenaquinol-6 biosynthesis II	isoprene biosynthesis II (engineered)
methanol oxidation to carbon dioxide	sucrose biosynthesis I (from photosynthesis)

3.3.4.- Discussion

The fact that different primers targeting different regions of the 16S rDNA gene have different binding affinity depending on the bacteria has been already reported [295–299]. While firstly the content of GC in each 16S rDNA region was suspected to explain the differences in primers specificity, this has been ruled out nowadays [295]. Instead, primers with different bacterial specificity have been shown to contain mismatches compared to 16S gene sequence for differently abundant bacteria between primers used.

The taxonomical annotation of both sequencing methods revealed that V3-V4 identified less different OTUs than the V1-V2 region sequencing, as expected from the reduced number of sequences that passed the quality controls. It is relevant to note that the V1-V2 primer pair generates an amplicon larger than the V3-V4 standard primer pair. Thus, due to the strict threshold used for OTU clustering (99% similarity), this higher amount of different OTUs observed for V1-V2 may be explained by punctual mutations and may not correspond to different bacterial species completely. Either if the differences are related to the number of quality reads or to the length of the sequenced amplicon, their biological relevance was null, as stated by the high similarity shown by Procrustes analysis. The weighted α-diversity results also supported this idea, seeing that all potential differences were lost.

With the data presented, we can conclude that although each primer combination has a better specificity for specific bacteria they do not alter the biological interpretation of results. Thus, even though β-diversity analysis identifies different microbial

communities depending on the primers used, Procrustes analysis reveals the elevated similarity between each PCoA. Therefore, similar samples clustering should occur independently of the primers used and the same biological conclusions may be obtained. The characterization of these conclusions, though, will depend highly on the primers used. For biomarkers identification, validation and application, multiple 16S rDNA regions, full gene or even whole genome sequencing seem to be more appropriated.

3.4.- Chapter 4. Fibromyalgia multi-omics analysis

The majority of the results presented in this chapter were published in *Ebiomedicine* as a Research Article with first authorship that can be found in Annex II:

Clos-Garcia, M. *et al.* Gut microbiome and serum metabolome analyses identify molecular biomarkers and altered glutamate metabolism in fibromyalgia. EBioMedicine 46, 499–511 (2019).

3.4.1.-Introduction

Fibromyalgia is a complex disease of unknown pathophysiology, for which no specific molecular biomarkers or biochemical alterations have been identified. In 1990, the American College of Rheumatology (ACR) recognised this syndrome as a disease and proposed the Widespread Pain Index (WPI), determined by measuring tenderness on pressure at 18 defined points, as a major diagnostic indicator. In 2010, the ACR introduced the Severity Score (SS), which also takes into account the associated symptoms and their severity [300]. Thus, the diagnosis of fibromyalgia is currently based on subjective pain evaluation and a set of associated signs and symptoms, which are used to assess the severity of the disease.

Even though the fibromyalgia is a complex disease with a multitude of signs and symptoms associated with many organs, the participation of the Central Nervous System (CNS) in its pathogenesis is broadly acknowledged [301]. Some studies have tried to identify molecular signatures that could explain some of the features of fibromyalgia and have provided some potential biomarkers. Several polymorphisms linked to the metabolism and breakdown of neurotransmitters involved in pain modulation have been identified as specific markers of increased risk of fibromyalgia [302]. Such polymorphisms have been found for the serotonin transporter gene 5-HTT [303, 304] and the catechol-O-methyl-transferase (COMT) gene [305, 306]. Some environmental factors, such as viral and bacterial infections [307], e.g. HCV infection [308, 309] and psychological stressors [310], known to produce alterations in the hypothalamic-pituitary-adrenal (HPA) axis, have been associated with this disease. Fibromyalgia is prevalent in individuals with chronic pain attributable to peripheral pain generators, such as rheumatoid arthritis [311]. At the molecular level, glutamate is elevated in the cerebrospinal fluid of fibromyalgia patients [312–314]. A decrease in insular levels of γ-aminobutyric acid (GABA) has also been described [315]. An inflammatory component in the pathogenesis of this disease has also been proposed: certain cells might trigger and perpetuate chronic pain by releasing chemokines and cytokines, such as IL-6 and IL-8, whose levels are elevated in the sera of fibromyalgia patients [316, 317].

The microbiome has a significant role in maintaining health [110]Alterations in the gut microbiome have been linked to a large number of diseases, including intestinal bowel disease (IBD) [318] and metabolic [319] and neurological [320, 217] disorders [150]. The microbiome has been recurrently associated with the CNS [217], indicating the existence of a gut-brain axis [321, 322]. Disturbances in the microbiome might lead, in some cases,

to neural disorders such as depression or autism. Some changes linked to microbial gut dysbiosis, understanding dysbiosis as those differences between healthy individuals and disease-specific patients [323], are also associated with symptoms used to determine the SS2 score in the diagnosis of fibromyalgia. The gut-brain axis has been proposed as a bidirectional communication system between the gastrointestinal tract and the brain, involving both neural and humoral mechanisms (reviewed in Collins, Surette and Bercik, 2012). The intestinal GABA produced by the bacteria from glutamate might affect the behaviour of the host, and it might be involved in anxiety and depression [325, 326, 211, 215]. Alterations in the microbiome composition can escalate the interactions between bacteria and the gut immune system due to the breakage of the intestinal barrier, promoting the release of pro-inflammatory molecules. Such events have been reported in IBD, where a release of IL-2, IL-17, interferon and/or TNFβ has been observed [201]. Interestingly, several pro-inflammatory cytokines can increase the permeability of the blood-brain barrier [322]. The microbiome also has metabolic, immunological and gut-protecting functions in the host. The fermentation of dietary carbohydrates by gut bacteria, for example, results in the production of short-chain fatty acids (SCFAs). These molecules are essential for the maintenance of the integrity of the intestinal barrier [150] and other health-related functions [151], including the correct development and maintenance of the blood-brain barrier [206].

These interactions between the microbiome and other functional systems of the organism have been widely studied. Microbiome data have been scrutinised in conjunction with host's genome, epigenome, transcriptome and metabolome [64]. The integration of different omics data relies mostly on dimension reduction approaches and is not specific to any omics technology, except for the metabolomics data. Correlation, regression and network-based approaches have also been implemented to integrate microbiome data with other omics analyses. As a result, the role of the host genome in regulating microbiome composition has been revealed [327]. A combination of Genome Wide Association Studies (GWAS) and microbiome-GWAS has been applied also to assess the impact of diet on microbiome composition. For example, associations between lactase [66] and variations of vitamin D receptor [68] genes with specific bacteria have been reported. Metabolomics-microbiome integration studies using correlation approaches have shown the effect of microbiome on host's insulin sensitivity [71] and on the development and progression of colorectal cancer [72, 73]. Metabolomics – microbiome integration studies employing a mix of correlation and network methods have obtained a comprehensive profile of the existent interactions between intestinal mucosa and gut microbiome [70]. The authors of these studies have used standard statistical methods but suggested that new, specific methods are needed for omics integration, to take into account the particular omics data characteristics [64].

The aim of this work was to identify potential molecular biomarkers for fibromyalgia diagnosis and characterization, employing different omics technologies: the analysis of microbiome from faeces samples and metabolomics, cytokine and miRNA profiling using serum samples.

3.4.2.- Methods

3.4.2.1.- Cohort recruitment

Individuals included in the study were recruited in two different hospitals in the Basque Country. Both fibromyalgia patients and healthy individuals were given a form with questions concerning several lifestyle variables (diet, smoking, alcohol consumption, physical exercise, other diseases and mood),. Blood samples were obtained from fibromyalgia patients and control individuals. Stool samples were collected from all participants, stored the samples at 4 °C until they could be delivered to the biobank. Blood and stool samples collected in each hospital were then sent to the Basque Biobank. Samples were aliquoted samples and frozen at -80 °C. The hospitals' clinicians (neurologists and rheumatologists) were responsible for the fibromyalgia diagnosis. The following criteria were used:

- Fibromyalgia group: WPI ≥ 7 and SS_T (Severity Score) ≥ 5 or WPI between 3 and 6 and SST ≥ 9. Patients with other diseases with similar symptoms were discarded.
- Control group: healthy individuals without any clinical manifestation of fibromyalgia and/or any other similar disease. To reduce the potential confounding factors associated with lifestyle, they also were age-paired with the patient group and came from the same environment.

All donors signed the informed consent form, and the study was approved by the appropriate ethical committee (CEIC-PI2016037). DNA from faeces was extracted using PSP Spin Stool DNA Plus kit (STRATEC Molecular®), following the manufacturer's protocol. Lysis buffer was added to the frozen samples, to ensure the preservation of nucleic acids. DNA extractions were then aliquoted into samples of 2.5 µg of DNA at the concentration of 100 ng/µL and then frozen until sequencing. All sample processing and distribution were managed by the Basque Biobank. The summary of the collection workflow can be found in FIGURE 34.

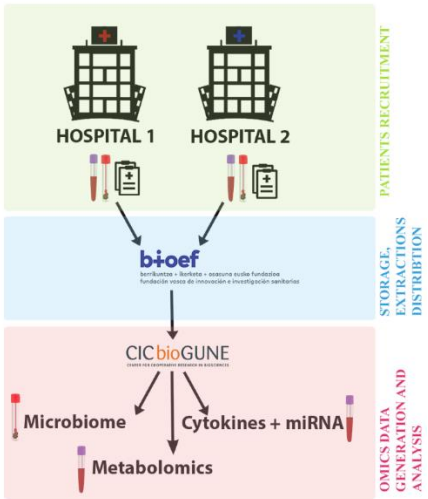


Figure 34: *Fibromyalgia project experimental design workflow, from patient recruitment and sample collection to the arrival of processed samples into the research center and their examination using distinct omics techniques.*

3.4.2.2.- Microbiome

3.4.2.2.1.- V3–V4 16S rDNA sequencing

The amplicon sequencing protocol targeted a fusion fragment containing the V3 and V4 regions (about 459bp) of the 16S genes with the primers designed surrounding conserved regions [293]. The full length primer sequences, using standard IUPAC nucleotide nomenclature, to follow the protocol targeting this fusion region were:

16S Amplicon PCR Forward Primer

5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG

16S Amplicon PCR Reverse Primer

5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC

DNA amplicon libraries were generated following Illumina Inc.'s recommendations. The amplification reactions consisted of:

Table 9: *Amplification reaction mix volumes.*

	Volume
Microbial DNA (5 ng/µl)	2.5 µl
Amplicon PCR Forward Primer 1 µM	5 µl
Amplicon PCR Reverse Primer 1 µM	5 µl
2x KAPA HiFi HotStart ReadyMix (KK2602)	12.5 µl
Total	**25 µl**

And PCR cycling was programmed with an initial denaturation at 95ºC for 3 min, followed by 25 cycles of annealing (95ºC - 30 seconds, 55ºC - 30 seconds, 72ºC - 30 seconds) and an extension at 72ºC for 5 minutes.

Then, lllumina Inc.'s sequencing adaptors and dual-index barcodes (Nextera XT index kit v2, FC-131-2001) were added to each amplicon (see Illumina Inc.'s Protocol for details) and, after PCR purification, libraries were normalized and pooled prior to sequencing. The pool containing indexed amplicons was loaded onto the MiSeq reagent cartridge v3 (MS-102-3003), spiked with 25% PhiX control to improve base calling during sequencing, as recommended by Illumina for amplicon sequencing. Sequencing was conducted using a paired-end, 2x300pb cycle run on an Illumina MiSeq sequencing system.

Sequencing was done by FISABIO Sequencing Core Facility, who also performed the quality assessment, using *prinseq-lite* [167] with the following parameters (min_length: 50, trim_qual_right: 30, trim_qual_type: mean, trim_qual_window: 20), and the sequence joining, with *FLASH* software [166] using default parameters.

3.4.2.2.2.- Microbiome sequences bioinformatics analysis

Joined reads were uploaded to QIIME2 software (v2017.10) [125], specifying the type parameter (SampleData[SequencesWithQualitiy]) and QIIME2 format option for FASTQ data input (SingleEndFastqManifestPhred33). Samples were then clustered *de novo* into OTUs, using the 97% similarity threshold using DADA2 plugin [328]. The resulting OTU table was then rarefied to 12,000 reads per sample, when no increase in diversity was obtained from including more reads. Rarefied table was aligned with mafft plugin [329] and the OTUs phylogenetic tree was then obtained using fasttree plugin [330]. Several alpha and beta diversity indexes were computed with diversity plugin and exported for posterior analysis. Finally, OTUs were annotated with GreenGenes 13_8 database and the resulting table was exported for posterior analysis.

OTU table was then imported into SIMCA-P+ 12.0.1 (Umetrics AB, Umeå, Sweden) in order to compute various multivariate analyses, including PCoA and PLS-DA analyses. OTU table, taxonomy and diversity indexes measurements were imported to R software (R Development Core Team, 2011; http://cran.r-project.org) in order to perform subsequent statistical analysis using *phyloseq* [331], *microbiome* [332] and *DESeq2* [333] R packages. Alpha diversity indexes differences were assessed using Student's t-test for the pairwise comparison (control *vs* fibromyalgia). *p-value* < 0.05 was considered significant. CORBATA [334] approach was used to identify and plot the bacteria corresponding to core microbiome, using the following thresholds: OTUs with a minimum ubiquity of 80% in the respective sample group and minimal abundance of 0.01% on each sample. SIAMCAT [239] was used to assess the potential effects of confounding factors such as sex, hospitals and distinct drug types. Finally, OTUs differential abundance between control and fibromyalgia samples was assessed using *DESeq2* R package [333], considering adjusted *p-value* < 0.05 significant.

3.4.2.2.3.- qPCR validation

From the glutamate cytoplasmatic incorporation and degradation pathways we selected four genes (gadC, glnA, glsA and glsB) to validate our findings related to glutamate and microbiome interaction. The primers were designed using the Primer-BLAST from NCBI website (https://www.ncbi.nlm.nih.gov/tools/primer-blast/), specifying the following five bacterial taxonomies (Bacteroides, Bifidobacterium, Eubacterium, Lachnospiraceae, Ruminococcaceae) and the "nr" database. We selected two pairs of primers for each gene, considering those without human subproducts (Table 10).

Table 10: *Primer pairs for each gene included in the qPCR validation targeting the 5 bacterial families.*

Gene	Direction	Sequence (5'→3')
gadC	Forward	CGGCGCGAATTGCTAAAGTT
	Reverse	TACTACCAGGGTGCCCACTT
glnA	Forward	TGTTTGACGGCTCCTCGATT
	Reverse	GGTTCAAGGATGTCGCAACG
glsA	Forward	TCTGTACGTTAGCCCTTGCG

	Reverse	GCTATGGCCCGGTTATGGAA
	Forward	TCTGGCGAATGTACCAGGTC
	Reverse	GCCCGGTTATGGAAGTTGGT
gadB	Forward	CAGACCTGGGACGACGAAAA
	Reverse	GGGCGAATTTATGCCAGCAG
	Forward	CAAACTGGGGCCGTATGAGT
	Reverse	AGTTTCGGGTGATCGCTGAG

The qPCR reaction was performed as follows:

5uL of SYBR(TM) Select Master Mix from Thermofisher Scientific® (#ref 4472908).

0.4uL of the mix of forward and reverse primers at 10uM.

3.6uL of RNase free water.

1uL of DNA template.

The reaction was runned in a QuantStudio 6 Flex Real-Time PCR System, from Thermofisher Scientific® with the running protocol:

Table 11: *Amplification protocol, including step, time and temperature per each step.*

Stage	Time	Temperature
Hold Stage	2 min	50ºC
	2min	95ºc
PCR Stage (x40 cycles)	15sec	95ºC
	15sec	58ºC
	1min	72ºC
Melt Curve Stage	15sec	95ºC
	1min	60ºC
	15sec	95ºC

3.4.2.3.- Metabolomics

To 40 µL aliquots of human serum, 40 µL of water/0.15% formic acid (FA) was added. Then, the proteins were precipitated by the addition of 120 µL of acetonitrile. To achieve the optimum extraction, after the addition of acetonitrile, the samples were sonicated for 15 minutes and agitated at 1,400 rpm for 30 min (at 4 °C). Next, they were centrifuged at 14,000 rpm for 30 min at 4 °C. The supernatants were transferred to vials. Samples were measured with a UPLC system (Acquity, Waters Inc., Manchester, UK) coupled to a Time of Flight mass spectrometer (ToF MS, SYNAPT G2, Waters Inc.). A 2.1 x 100 mm, 1.7 µm BEH amide column (Waters Inc.), thermostated at 40°C, was used to separate the analytes before entering the MS. Mobile phase solvent A (aqueous phase) consisted of 99.5% water, 0.5% FA and 20 mM ammonium formate while solvent B

(organic phase) consisted of 29.5% water, 70% MeCN, 0.5% FA and 1 mM ammonium formate.

In order to obtain a good separation of the analytes the following gradient was used: from 5% A to 50% A in 2.4 minutes in curved gradient (#8, as defined by Waters), from 50% A to 99.9% A in 0.2 minutes constant at 99.9% A for 1.2 minutes, back to 5% A in 0.2 minutes. The flow rate was 0.250 mL/min and the injection volume was 2 µL. All samples were injected randomly.

The MS was operated in positive electrospray ionization in full scan mode. The cone voltage was 25 V and capillary voltage was 250 V. Source temperature was set to 120 °C and capillary temperature to 450 °C. The flow of the cone and desolvation gas (both nitrogen) were set to 5 L/h and 600 L/h, respectively. A 2 ng/mL leucine-enkephalin solution in water/acetonitrile/formic acid (49.9/50/0.1 %v/v/v) was infused at 10 µL/min and used for a lock mass which was measured every 36 seconds for 0.5 seconds. Spectral peaks were automatically corrected for deviations in the lock mass. Scaled and normalised data were uploaded to R. Principal Component Analysis (PCA) was performed to check whether the differences between sample metabolomes were due to sample origin and to account for the autoclaving process used by one of the hospitals. We excluded the metabolites whose expression differed between the hospitals, to avoid the bias introduced by the sample origin. Metabolomic features with more than 30% of missing values in either hospital were removed from the analysis. Fold changes and *p-values* (adjusted using the Bonferroni method) were computed. Afterwards, differential peaks were selected for further annotation and metabolite identification using the METLIN database [100]. The identification was confirmed using commercial standard injection.

MetScape [335] and Ingenuity Pathway Analysis® were used to map the identified metabolites to corresponding functionalities in humans.

3.4.2.4.- MiRNA & cytokines profiling

Cytokines profiling was performed by Abcam's FirePlex Service Lab (Boston, USA). The cytokine profiling was performed using the FirePlex Human Discovery Cytokine Panel (Abcam, ab227936), allowing for the simultaneous profiling of 70 targets in a single well of sample. Each sample was run in duplicate following the manufacturer's instructions.

In brief, all serum samples were diluted 1:4, adding 12.5µL of samples to 37.5µL of Human Assay Diluent 1X. 150µL of 1X Capture Particles were added to each well of a 96-well plate and filtered. After a single rinse with 175µL 1X Wash Buffer the prepared samples were added to the corresponding wells. The plates were then incubated in the dark overnight at 4°C with 750rpm shaking. After rinsing twice with 175µL of 1X Wash Buffer, 50µL of Detector Antibody Solution were added to each well and incubated in the dark for 1 hour with 750rpm shaking at room temperature. After rinsing twice with 175µL of 1X Wash Buffer, 50µL of 3X Reporting Mix were added to each well and incubated in the dark for 30min with 750rpm shaking at room temperature. After rinsing twice with 175µL of Wash Buffer 1X, 175µL of Run Buffer were added to each well and

the particles were scanned on an EMD Millipore Guava 6HT flow cytometer. The flow cytometer output was analyzed with the FirePlex™ Analysis Workbench Software (http://www.abcam.com/FireflyAnalysisSoftware). Cytokines concentration per sample was interpolated from the standard curve run in duplicate on each plate. Data was log-normalized, and then fold change and Bonferroni adjusted *p-value* were computed to assess the differences between the cytokines profile in the patients.

miRNA profiling was performed by Abcam's FirePlex Service Lab (Boston, USA). The miRNAs were profiled using the FirePlex miRNA Assay Core reagent Kit (Abcam, ab218342) using a custom multiplex panel that was constructed to include 68 miRNA selected from literature revision. Each sample was run in singlicate as previously described [336].

In brief, 20µL of sample was mixed with Digest Buffer + Protease Mix to a final volume of 80µL and was incubated at 60°C for 45min at 750rpm shaking. 35µL of FirePlex Particles were added to each well of a 96-well plate and filtered, including 3 wells for no-sample control. 25µL of Hybridization Buffer was added to each well along with 25µL of sample. In the case of the no-sample controls, water was added instead. The plate was incubated at 37°C for 60min at 750rpm. After rinsing twice with 175µL of Rinse A 1X buffer, 75µL of Labeling Buffer 1X was added to each well and the plate was then incubated at RT for 60min at 750rpm. After two rinses with Rinse B 1X and one of Rinse A 1X, adapted miRNAs were eluted from the particles using 130 µL of 95°C. Particles were then stored in the filter plate at 4°C with 75µL of Rinse A 1X until needed.

30µL of the eluant was added to a clean PCR plate and mixed with 20µL of PCR master mix and underwent 32 cycles of PCR amplification. After removal of Rinse A 1X from the particles stored in the previous step 60µL of Hybridization Buffer were added to each well followed by 20µL of the PCR product. The plate was then incubated at 37°C for 30min at 750rpm. After rinsing twice with Rinse B 1X and once with Rinse A 1X, 75µL of Reporting Buffer 1X was added to each well and then incubated at RT for 15min at 750rpm. After rinsing twice with 175µL of Rinse A 1X, 175µL of Run Buffer were added to each well. The particles were scanned on an EMD Millipore Guava 6HT flow cytometer. Data analysis was performed with the FirePlex™ Analysis Workbench software. Three miRNAs are used for nomalization: hsa-miR-17-5p, hsa-miR-320b, hsa-let-7i-5p were selected using the geNorm algorithm [337]. Data was log-normalized, and then fold change and Bonferroni adjusted *p-value* were computed to assess the differences between the miRNA profile in the patients.

3.4.2.5.- Data integration

3.4.2.5.1.- Microbiome and metabolomics

Spearman's correlation coefficients were computed for relationships between relative abundances of microbiome bacteria with the identified genus and normalised individual metabolomic features. A scaled heatmap was constructed for the correlation matrix, including cladogram classification of the variables, using the default clustering method.

3.4.2.5.2.- Integration of all datasets

We employed the Data Integration Analysis for Biomarker Discovery using Latent cOmponents (DIABLO) implementation in the mixOmics R package [241, 242]. Thirty-six fibromyalgia and 35 control samples were used. Microbiome data was normalised using DESeq2 counts function. The mixOmics block.splsda function, with full weighted design and 10 components, was primarily used to identify the optimal number of components, which was defined in 3 methods using the centroid distance technique. To decide which variables to keep in each component, models with 10, 5, 5 and 5 randomly selected variables were tested for the microbiome, metabolomics, cytokines and miRNAs, respectively. Finally, different model features were obtained and the results were plotted using mixOmics predefined and *ad-hoc* functions. This procedure was followed for both the identified-metabolite dataset and the full dataset of unidentified metabolomics features.

3.4.3.- Results

3.4.3.1.- Clinical samples

One hundred and five confirmed fibromyalgia patients (ACR 2010 modified criteria) [300] and 54 age- and environmentally-paired healthy individuals were recruited. The latter group consisted of individuals who did not present any disease or symptoms related to fibromyalgia and came from the same environment as the fibromyalgia patients. The characteristics of the study cohort are shown in Table 12.

Table 12: *Cohort characteristics. The number of individuals included in each group is given in parentheses. For Age, WPI and SS_T, mean values ± standard deviation are shown.*

	Controls (n = 54)	**Fibromyalgia-diagnosed patients (n = 105)**
Sex	48.15% ♀, 51.85% ♂	69.52% ♀, 30.48% ♂
Age *(years)*	53.5 ± 12.4	52.52 ±10.3
Age at diagnosis *(years)*	*NA*	48.2 ± 11.1
Time since diagnosis *(years)*	*NA*	3.4 ± 6
WPI	*NA*	13.28 ± 3.91
SS_T	*NA*	8.62 ± 2.15
SS1	*NA*	6.6 ± 1.8
SS2	*NA*	2.1 ± 0.4

During WPI evaluation, more than 90% of the patients reported pain in the back, shoulder girdle and abdomen. Neck pain was described by 85% of patients, while upper and lower arm, hip and upper and lower leg pain were reported by 70% of fibromyalgia patients. At least 50% of the patients were affected by jaw and chest pain. The SS_T index is the combination of two sub-indexes, SS_1 (the severity of 3 main symptoms in fibromyalgia: fatigue, sleep quality and cognitive problems) and SS_2 (the list of associated fibromyalgia symptoms). Approximately 90% of patients reported moderate to severe scores for the 3 main symptoms for the SS_1 sub-index in the week preceding the collection of the samples. In the evaluation of associated fibromyalgia symptoms (SS_2), 70.7% of fibromyalgia patients presented at least 4 symptoms from the neurological sphere (muscle pain, fatigue, thinking or memory problems, headache, numbness/tingling, etc.). Among them, 70% used painkillers, while approximately 55% were taking antidepressants and benzodiazepines and approximately 30%, antiepileptic drugs (https://www.ebiomedicine.com/article/S2352-3964(19)30473-6/fulltext). Half of the patients reported some physical exercise and some alcohol consumption, while 23% identified themselves as smokers.

3.4.3.2.- V3-V4 16S rDNA sequencing

We obtained 6,110,564 reads, of which 99.56% passed the quality check. Of the cleaned reads, the 81.91% (4,982,956) were joined. To decide on the number of reads to which the samples should be rarefied; we computed the rarefaction curves for both observed OTUs and Shannon indices (https://www.ebiomedicine.com/article/S2352-3964(19)30473-6/fulltext). After rarefying at 12,000 reads/sample, the median coverage was 96.35 ± 2.33%. Rarefaction step did not reduce diversity (https://www.ebiomedicine.com/article/S2352-3964(19)30473-6/fulltext). Sequencing data was uploaded to ENA under Project Accession code PRJEB27227.

3.4.3.3.- Microbiome

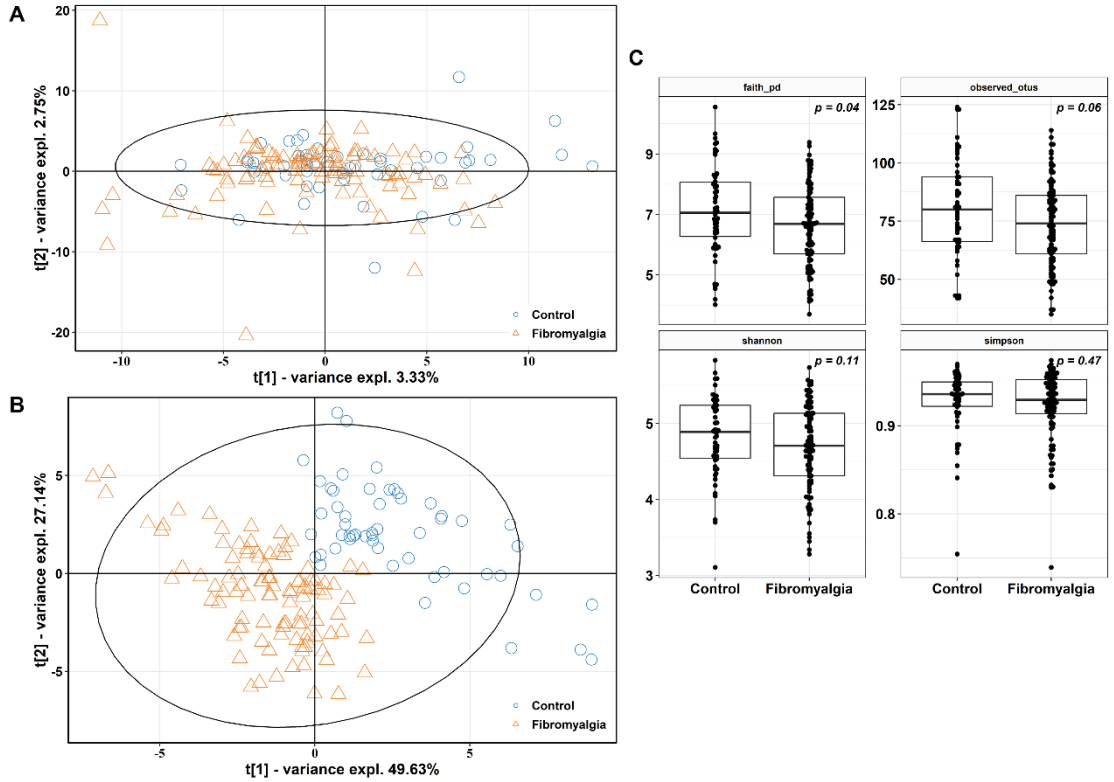


Figure 35: *Microbiome multivariate analysis. (A) Principal Component Analysis (PCoA) of the complete cohort. (B) Supervised Partial Least Squares Discriminant Analysis (PLS-DA) analysis, showing the discrimination between the sample groups. (C) Alpha-diversity indexes for each sample group, showing the adjusted p-value computed using Student's t-test.*

The multivariate unsupervised PCA (FIGURE 35A) did not show any differences between the control and the fibromyalgia samples. The supervised Partial Least Squares Discriminant Analysis (PLS-DA), however, provided two sample groups (FIGURE 35B) (*p-value*, 0.0019). In the specific diversity analysis for 4 alpha-diversity indices (Faith's Phylogenetic Distance, ace, chao1 and observed OTUs) we observed a discrete decrease in bacterial diversity in fibromyalgia patients although only the Faith's PD index showed a statistically significant difference (FIGURE 35C). This reduction in bacterial diversity was also observed in the analysis of the core microbiome at the taxonomic family level. We used CORBATA default parameters (80% ubiquity, 1% abundance) to identify which bacteria families present in both fibromyalgia and control core microbiomes. The two core microbiomes contained the same 4 bacteria families (C. Ruminococcaceae, C. Lachnospiraceae, B. Rikenellaceae and B. Bacteroidaceae). We observed that the control group presented a more diverse bacterial community. The comparison of the two sample groups revealed that Clostridiales Ruminococcaceae was more abundant in the healthy control group than in fibromyalgia patients, although the differences were not statistically significant (FIGURE 36A). After reducing the cut-off to 50% ubiquity, we observed differences between the core microbiomes of the two groups. Specifically, two bacteria families that were absent in the fibromyalgia core microbiome, the Bifidobacteriales Bifidobacteriaceae and the Bacteroidales Prevotella, which were

represented in the control core microbiome (https://www.ebiomedicine.com/article/S2352-3964(19)30473-6/fulltext).

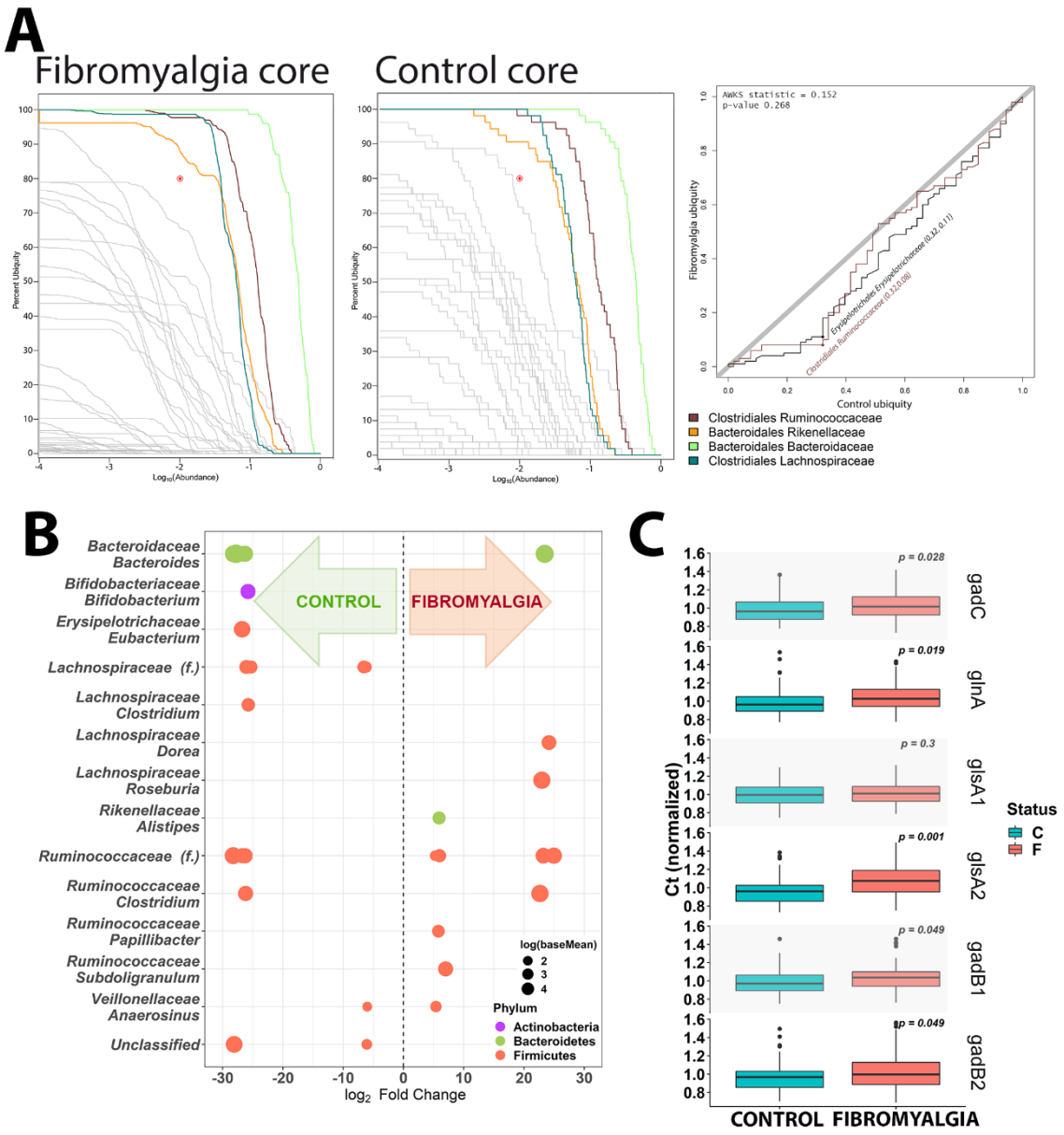


Figure 36: *Core microbiome and genus-discriminant analyses. (A) The composition of core microbiome for each sample group and the comparison of bacterial ubiquity in the two groups. (B) Genera significantly different (adj p N .05) between the control and fibromyalgia samples, obtained using the protocols described in the Methods. Positive log2 fold changes (x-axis) indicate genera with positive fold difference between fibromyalgia and control. Negative log2 fold changes are shown as negative x values. Each point represents a single OTU, coloured by phylum. On the y-axis, the taxonomic genus level is indicated. Size of the points reflects the log-mean abundance of the sequence data. (C) qPCR results for the differential expression of bacterial genes related to glutamate bacterial degradation. Results are indicated in differential Cts count.*

We performed a differential OTU analysis (employing DESeq2) of the core microbiomes in the control and fibromyalgia samples. We identified 32 OTUs distributed among 3 phyla (Actinobacteria, Bacteroidetes and Firmicutes) (FIGURE 36B) whose abundance differed between the two groups, with an adjusted *p-value* of 0.05. In fibromyalgia

patients, the Bacteroidetes and Firmicutes had OTUs both with increased and decreased abundance, and Actinobacteria levels were reduced in this group (FIGURE 36B).

The number of OTUs with the unassigned genus in Bacteroidaceae and Lachnospiraceae families were decreased in fibromyalgia samples; there were also fewer Bifidobacteriaceae and Erysipelotichaceae OTUs in fibromyalgia patients. The Rikenellaceae family showed an increased abundance in fibromyalgia patients (https://www.ebiomedicine.com/article/S2352-3964(19)30473-6/fulltext).

Finally, at the genus level, the abundance of *Bacteroides* OTUs was reduced in fibromyalgia patients, as were *Bifidobacterium, Eubacterium* and *Clostridium* OTUs. However, the abundances of the genera *Dorea, Roseburia* and *Alistipes* were increased in this group (FIGURE 36B).

There were no significant differences between microbiome composition abundances in the two sexes. We did not observe any significant association between drug types (as summarized in https://www.ebiomedicine.com/article/S2352-3964(19)30473-6/fulltext) and the relative microbiome abundance at the genus level *(data not shown)*.

We validated the reduction of the abundance of bacterial species by qPCR technique. For that, we amplified a set of genes dedicated to the glutamate incorporation to bacterial cytoplasm and its transformation to GABA (gadC, glnA, glsA and glsB). We designed specific primers for amplifying genes from 5 bacterial families that we found to be diminished in fibromyalgia patients (Bacteroides, *Bifidobacterium*, *Eubacterium*, Lachnospiraceae and Ruminococcaceae) (FIGURE 36C). We found that the gene encoding the transporter of glutamate into bacterial cytoplasm, represented by gadC, was diminished, as it was also the genes encoding enzymes involved in the transformation of glutamate to L-glutamine (glnA, glsA) and to GABA (gadB) (https://www.ebiomedicine.com/article/S2352-3964(19)30473-6/fulltext), in agreement with the taxonomic analysis of 16S rDNA gene.

3.4.3.4.- Metabolomics

The metabolomics analysis yielded 8543 different metabolic features defined by retention time and mass/charge. One sample was removed due to technical failure. The PCA analysis revealed that the metabolomics profiles differed between hospitals (https://www.ebiomedicine.com/article/S2352-3964(19)30473-6/fulltext). This was expected because of the autoclaving performed in one of the hospitals. Thus, to avoid the bias caused by the chemicals released during the autoclaving procedure, the discriminating hospital features (p = 661), were removed from the study, as well as the features with >30% of missing values. Two hundred and twenty-eight features differed between the fibromyalgia and control groups (FIGURE 37A). Of these 228, only 88 had tentative IDs in the METLIN database. Using MS/MS data and chemical standards, we found that the levels of 7 of these metabolites were significantly altered in the fibromyalgia samples (https://www.ebiomedicine.com/article/S2352-3964(19)30473-6/fulltext): ornithine, L-arginine, Nε-Methyl-L-lysine, L-glutamate, L-glutamine, asymmetric

dimethylarginine (ADMA) and platelet activating factor (PAF-16) (FIGURE 37B). Another metabolic feature among the 228 altered in fibromyalgia was tentatively identified as L-threonine or DL-homoserine (FIGURE 37B). We could not discriminate between these two metabolites as they are structurally similar and have the same molecular mass and fragmentation pattern in LC-MS. We also analyzed the metabolites described in the literature, such as creatinine [338, 339], platelet activating factor [340] and acetylcarnitine [341]. To infer alterations in the biological processes and metabolic and functional pathways associated with the differentially expressed metabolites, we used MetScape [335] and Ingenuity Pathway Analysis® (QIAGEN) (IPA). The analyses showed that cell signaling and inflammatory and hypersensitivity responses were the most relevant biological processes. The most represented metabolic pathways were arginine, nitric oxide (NO) and glutamate metabolism.

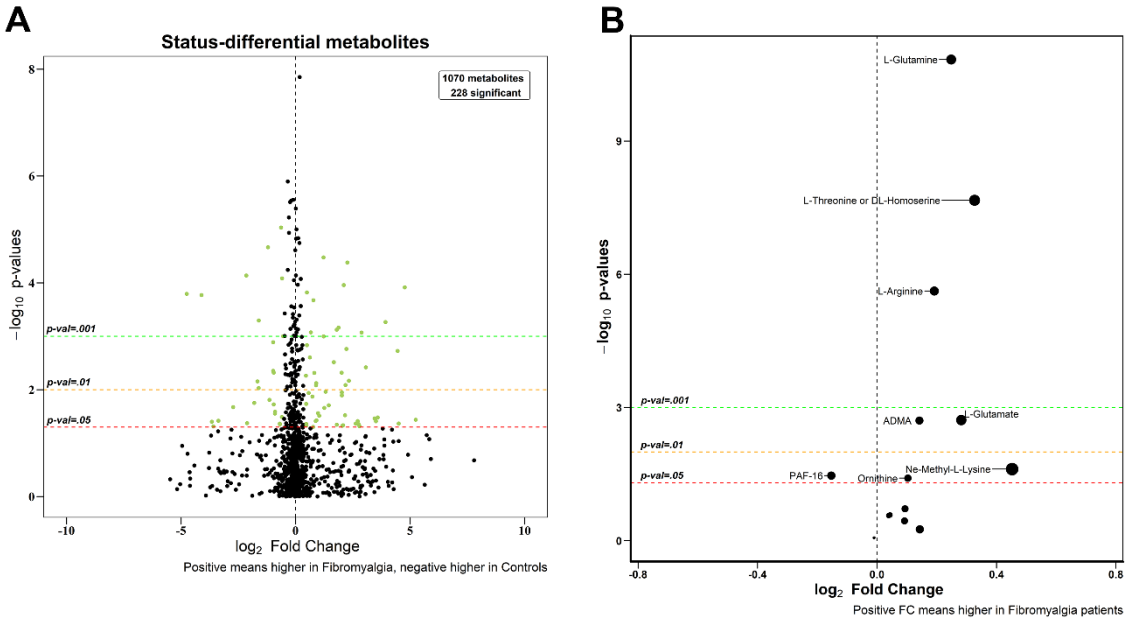


Figure 37: *Univariate metabolomics analysis. (A) Volcano plot of 1070 metabolic features detected in serum samples after background subtraction and removal of the features found in 30% of the data or differing between hospitals. (B) Volcano plot of the identified metabolites. Positive log2 FC indicates increased abundance in fibromyalgia patients. All p-values were adjusted using the Bonferroni method.*

To study the potential dependencies between microbiome composition and the host metabolism and metabolome, we examined the correlations between the two datasets. We computed the Spearman's correlation coefficient for the full set of metabolomics features and microbiome variables. We did not see any clear association patterns between the two complete datasets (https://www.ebiomedicine.com/article/S2352-3964(19)30473-6/fulltext). We also constructed a heatmap of the scaled correlations between the bacteria whose abundance was changed in fibromyalgia and the identified metabolites (FIGURE 38A). Metabolites were grouped into two clusters, depending on the correlations. These were seen mainly with genera *Bifidobacterium* and *Dorea*, which behaved in the opposite manner. The first cluster contained 4 metabolites (3-methyl-L-Lysine, PAF C-16, ADMA, L-Lysine). The second cluster was formed by 8 metabolites (glutamate, L-threonine/DL-homoserine, glutamine, Nε-methyl-L-Lysine, creatinine,

ornithine, arginine and acetylcarnitine), although the metabolite acetylcarnitine behaved differently from the other metabolites in this cluster. *Bifidobacterium*, whose abundance was reduced in fibromyalgia patients, correlated negatively with the first metabolite cluster and positively with the second one. *Dorea*, with increased abundance in fibromyalgia patients, correlated positively with the first metabolite cluster and negatively with the second one.

Finally, we checked, using Virtual Metabolic Human [342] database, whether the different metabolites were produced by the differentially abundant bacteria. We also wanted to study whether they were made by the genera for which we found most correlations (FIGURE 38A). Thus, we limited the search to *Bifidobacterium* and *Dorea* genera. For glutamate, we identified the metabolites upstream and downstream of its production/degradation. For lysine, threonine, homoserine, glutamine, ornithine and arginine (and their modifications), we found that the metabolites themselves, their precursors and degradation products might had been produced by bacteria. No bacterial associations were found for creatinine, PAF C-16, ADMA and acetylcarnitine, consequently suggesting that their origin was exclusively human.

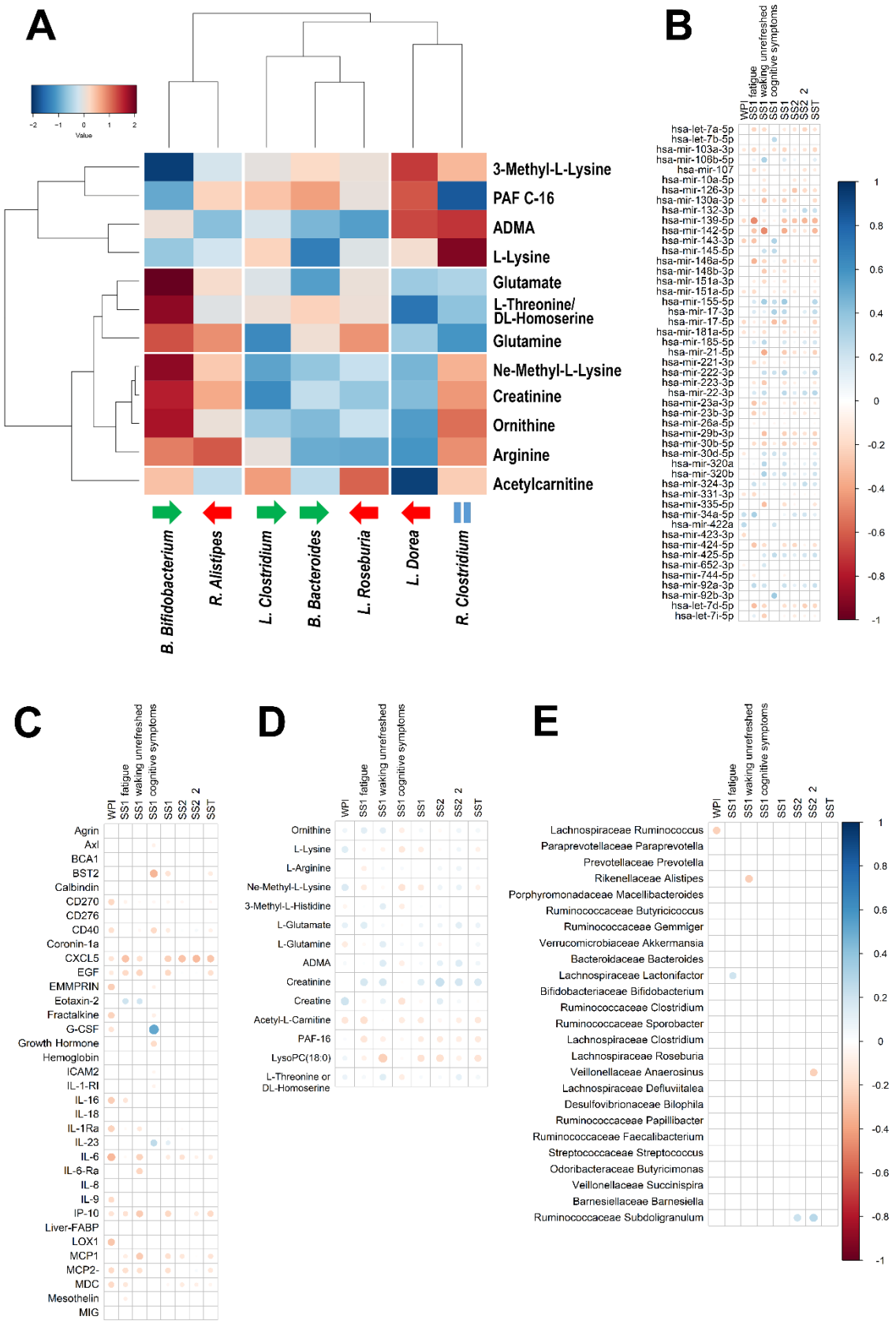


Figure 38: *Heatmap of scaled correlations between the bacteria whose abundance was altered in fibromyalgia and the identified metabolites. The dendrograms were unsupervised. Red arrows mark the bacteria with increased abundance in fibromyalgia, green arrows, with decreased abundance, and "equals" symbol indicates the OTUs with both increased and decreased abundance (A). Omics correlations*

with indexes used in fibromyalgia diagnostics, as defined by ACR 2010 criteria. Only significant correlations (p-value < .05) are coloured. Positive correlations are indicated in red and negative correlations, in blue. Correlations between circulating miRNA levels (B), circulating cytokine levels (C), identified serum metabolites (D) and microbiome composition (at genus level) (E).

Serum factors and miRNA analyses for a subset of samples

A subset of the samples (n = 72; n_C = 36 controls and n_F = 36 fibromyalgia samples) was used to perform multiplex assays for different serum molecules, including miRNAs and cytokines. For the multiplex design, we used 70 molecules and 68 miRNAs that have been associated with fibromyalgia and/or chronic pain. The protein content assays and the miRNAs analyses did not show any differences between the fibromyalgia and the control groups. We observed statistically significant differences for ten serum proteins: PCSK9, mesothelin, BST2 (↑), procalcitonin, Axl, myoglobin, MIG, TNF-alpha, ICAM2 and IL-9 (↓) with fold changes ranging from 0.76 (lower level in patients) for IL-9 to 1.07 for BST2 (https://www.ebiomedicine.com/article/S2352-3964(19)30473-6/fulltext). However, the levels of only one miRNA differed significantly between the fibromyalgia patients and the control group, the hsa-miR-335-5p (https://www.ebiomedicine.com/article/S2352-3964(19)30473-6/fulltext). Predicted target genes were obtained using miRWalk 2.0 database [343]; they were selected if they mapped to at least 8 of the 12 database options. The enrichment of the miRNA targets was performed using ConsensuspathDB [344], selecting the targets with a *p-value* < 0.01. Notably, we identified several pathways related to signaling dedicated to gene regulation processes. The complete results are provided in https://www.ebiomedicine.com/article/S2352-3964(19)30473-6/fulltext.

3.4.3.5.- Correlations between omics data and clinical data

To determine which differences could be associated with the disease, we examined the correlations between different diagnostic indexes obtained for the fibromyalgia patients and the omics data (FIGURE 38B, C, D and E). Notably, miRNA data constituted the omics dataset most correlated with pain indicators (FIGURE 38B), followed by the results of serum protein profiling (FIGURE 38C). Metabolomics also showed a considerable number of correlations with several pain indexes (FIGURE 38D). The microbiome composition (at genus level) (FIGURE 38E) was the omics dataset with the weakest correlation with pain indicators.

We also considered possible effects of medication on the observed differences between the patient and control samples. We checked whether the samples clustered depending on the drug regimen followed. However, we did not find any clusters of samples (neither for serum factors nor for miRNAs) that could be associated with a specific drug or drug combination. We also checked whether any data correlated with distinct drug types; no such correlation was observed (*data not shown*).

3.4.3.6.- Modelisation of microbiome, metabolomics, cytokine and miRNA datasets

We combined the four datasets of the 71 samples ($n_C = 36$, $n_F = 35$) that had all the data. Combining these datasets allowed us to discriminate between the control and fibromyalgia samples when a block sparse PLS-DA model was applied (block sPLS-DA) (FIGURE 39A). The analysis of the individual contribution of each dataset to the differences showed that the most correlated datasets were the microbiome composition and metabolomics data. We also found that the major contributor to the separation of the sample groups was the microbiome dataset, followed by serum metabolomics, proteins and, finally, miRNAs (FIGURE 39B and FIGURE 39C). In this analysis, we used only the metabolomics dataset containing the identified metabolites (n = 14). The sPLS-DA analysis using the whole unidentified metabolomics dataset (n = 1070) showed that using the metabolomics dataset improved the discrimination between the two sample groups, becoming the strongest factor distinguishing the patients from controls (https://www.ebiomedicine.com/article/S2352-3964(19)30473-6/fulltext) although the microbiome showed slightly better predictive ability.

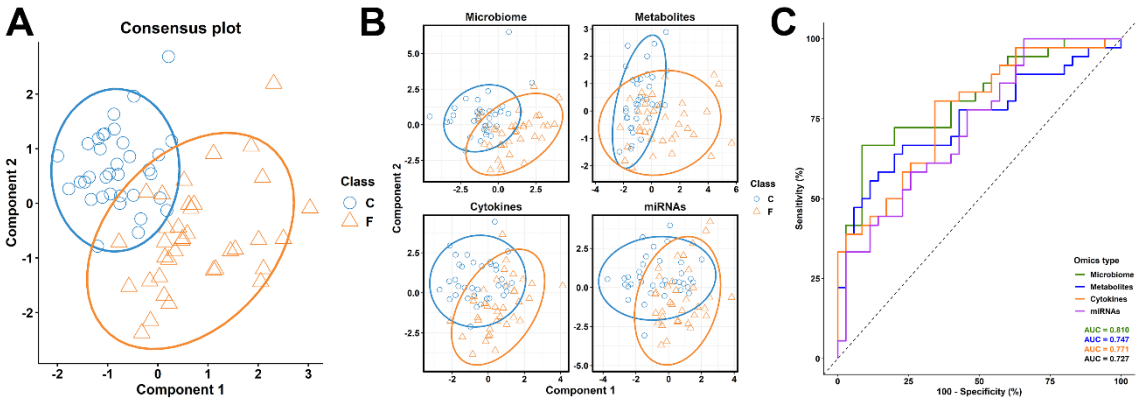


Figure 39: *Multi-omics integration. (A) sPLS-DA consensus plot for the combination of the 4 datasets, showing the nearly complete discrimination of the 71 samples (36 fibromyalgia and 35 control samples). (B) The individual contribution of each dataset to the sPLS-DA final model, in each case showing the score plots for the two first components, indicating the best separation capability for microbiome data, followed by cytokines, metabolomics and miRNAs. (C) ROC curves for each omics dataset, with the Area under the Curve (AUC) values.*

3.4.4.- Discussion

In this study, we applied an omics approach and identified a set of potential molecular markers (Table 13) for the diagnosis of fibromyalgia.

Table 13: *Differences between fibromyalgia and healthy control groups observed using each omics technique (showing alterations in the fibromyalgia patients).*

	Increased (↑)	Decreased (↓)
Microbiome	*Dorea* *Roseburia* *Papillibacter* *Subdoligranulum*	*Bifidobacterium* *Eubacterium* Lachnospiraceae (family) *Clostridium* Firmicutes (phylum)
Metabolomics	L-glutamine L-threonine/DL-homoserine L-arginine ADMA L-glutamate Nε-methyl-L-lysine Ornithine	PAF-16
Cytokines	PCSK9 Mesothelin BST2	Procalcitonin Axl-UFO Myoglobin MIG TNF-alpha ICAM2 IL-9
miRNAs	hsa-miR-335-5p	

The gut microbiome analysis revealed two clusters (FIGURE 35B), one cluster for fibromyalgia patients (modified 2010 ACR diagnostic criteria) and the other for individuals without any clinical manifestation of fibromyalgia. Both core microbiome and alpha-diversity analyses showed a reduction in bacterial diversity in the fibromyalgia group. This is in agreement with the report of reduced microbiota diversity in other pain disorders, such as myalgic encephalomyelitis/chronic fatigue syndrome [345]. Remarkably, our fibromyalgia microbiome analysis showed a reduction in the abundance of several bacterial strains associated with healthy microbiome, such those related to SCFA production (*Bifidobacterium, Eubacterium,* Lachnospiraceae) [150–153], and/or the reduction of Firmicutes phylum OTUs [137, 107, 108], suggesting dysbiosis events in fibromyalgia patients. Due to the current debate upon the dissension on the use of dysbiosis term and its meaning [323], we want to emphasize that with dysbiosis term we refer to those microbiome compositional alterations associated to disease,

either them being causational or consequence of the disease. In these terms, dysbiosis events are also associated with the disruption of the intestinal barrier, which allows the bacteria to interact with the immune system of the host, producing local inflammation [201]. This is supported not only by the large proportion of patients reporting abdominal pain (> 90%) but also by the number of intestinal diseases considered co-morbidities of fibromyalgia. The maintenance of the intestinal barrier is associated with the production of SCFAs, including butyric acid and butyrate [151]. In fibromyalgia, we found a decrease in the abundance of several members of the *Lachnospiraceae* family, the bacteria involved in butyric acid production [346]. Butyrate, the conjugate base of butyric acid, is produced by a small number of bacteria, including several *Eubacterium* species [152], a genus also underrepresented in fibromyalgia patients. The reduction in the diversity of bacteria, especially of those engaged in the production of protective SCFAs, suggests that this process might be implicated in the development of fibromyalgia. Notably, dysbiosis events, in the terms presented here, should be constant among time. Thus, we recognize that multiple time-point data should be studied and that the lack of this data is a limitation of our study. Nevertheless, we would like to highlight that this is a pilot study and that a follow-up study that could reinforce our statements is recommended.

We also found differences between neurotransmitter metabolisms in the patients and control individuals. We detected a significant increase in the serum levels of glutamate in fibromyalgia patients. Moreover, the abundance of bacteria from *Bifidobacterium* and *Lactobacillus* genera (involved in the transformation of glutamate into GABA; [347, 348, 211] was reduced in the fibromyalgia group. This might contribute to the elevated systemic levels of glutamate. The effect of GABA on the gut-brain axis, via the vagus nerve, has been described by several authors [211, 322]. Glutamate affects the development of pain, via glutamatergic synapses [349], and stress can alter the regulation of this pathway [350]. Stress-related events have also been associated with microbiome modifications [211]. The 2010 modified ACR criteria for fibromyalgia diagnosis include several stress-associated symptoms. Whether such elevated systemic levels of glutamate affect the ENS and alter the CNS is still unclear. However, some authors have demonstrated the activation of glutamatergic neurons and glutamate-mediated neurotransmission in the ENS [351–353, 320]. As a result of a reduction in bacterial diversity, the glutamate might enter the host bloodstream after the disruption of the intestinal barrier by the inflammation caused by the dysbiosis. Interestingly, several patients presented with symptoms associated with IBD as fibromyalgia comorbidities (irritable bowel syndrome (46%), pain in abdomen (13%) and in the upper abdomen (45%), diarrhea (20%), etc.). The role of microbiome in IBD pathogenesis has been broadly demonstrated [354, 355], suggesting that dysregulation of intestinal immune system derived from microbiome alterations may lead to disease [356], as demonstrated by patients presenting T-cell responses against commensal bacteria [357]. Specifically, a reduction of Firmicutes phylum bacteria has been recurrently associated with IBD pathogenesis and progression [358, 359], such as that observed for fibromyalgia patients. These common alterations in microbiome composition could

explain thus some of the most frequent comorbidites reported by the patients of our study.

Furthermore, it has been shown that the blood-brain barrier increases its permeability after a decrease in the numbers of SCFA-producing bacteria. This alters the tight junction organization, which can be recovered by colonization with SCFA-producing bacteria and/or by administration of these bacterial metabolites [206]. Cytokines can also modify the blood-brain barrier permeability [360, 361]. Importantly, glutamate levels increase in the cerebrospinal fluid (CSF) of fibromyalgia patients [312]. These data suggest an important role of this neurotransmitter in the pathogenesis of fibromyalgia. How peripheral levels of gut microbiome-derived neurotransmitters can affect the brain function is something still in debate [320], although several mechanisms have been proposed. The alteration of the blood-brain barrier permeability would lead to a modification on the interchange of serum metabolites with the brain. Serum levels of 5-HT have been demonstrated to be altered in germ-free mice [362, 363] and, while 5-HT itself is not known to cross the blood-brain barrier, their precursor levels are able to do it. Microbiome could, instead, alter 5-HT precursor levels, as has been proposed by several authors [364, 325], like tryptophan. This same mechanism has been discussed to be true for other gut microbiome produced neurotransmitters, such as dopamine and GABA [320, 365, 366].

It is essential to keep in mind the relationship between GABAergic pain inhibition and gender as fibromyalgia is 3 times more prevalent in women than in men [367]. Steroid 17β-estradiol (E2) suppresses the GABAergic inhibition in female rats via a sex-specific oestrogen receptor ERα, mGluR and endocannabinoid-dependent mechanism [368]. This suppression requires the activation of mGluR type I receptors by glutamate [369]. Therefore, in the presence of excess glutamate, as observed here in fibromyalgia patients, the pain inhibition by GABA might be suppressed in female patients by this E2-specific regulation. This might partly explain the increased prevalence of fibromyalgia in the female population.

The functional analysis of the metabolomics dataset showed that the most represented pathways were those dedicated to the metabolism of known neurotransmitters, such as glutamate and serine. Both arginine and ornithine levels, related to the widespread pain in fibromyalgia, increased in the sera of fibromyalgia patients. Consistently, IPA analysis identified several pathways related to arginine, such as arginine degradation (I and II) canonical pathways and proline biosynthesis from arginine. These two metabolites are required for the synthesis of nitric oxide (NO) [339]. NO plays an important role in both acute and chronic pain as it is a mediator of nociception [370]. However, NO contributes not only to nociception; it also mediates in analgesia and increases the effect of morphine on pain inhibition [370]. Here, we also observed strengthening of this pathway in fibromyalgia patients (by using IPA). The role of NO in fibromyalgia pathogenesis has been studied but without reaching a consensus [371]. Notably, the levels of iNOS isoform increase in female fibromyalgia sufferers in comparison with healthy controls, while the levels of constitutive isoforms (nNOS and eNOS) do not change [372]. It is

important to remember that our functional profiling was performed using the results obtained from the serum sample analysis. A relevant limitation in this study is, precisely, the metabolomics analysis and, more specifically, the metabolite identification step. We could only identify a small subset of all the metabolic features observed. Thus, the results obtained in this study are constrained by the reduced number of identified metabolites. A better metabolite identification procedure could improve not only the list of potential metabolite biomarkers but also the identification of potentially affected biological pathways and functionalities.

Patients afflicted by chronic pain are likely to participate in many different long-term treatments, which could affect their microbiome composition. Differences in diets and lifestyles will also have some effect. Thus, it is difficult to be certain whether the detected alterations in the microbiota are the cause or consequence of fibromyalgia. Notably, no association between microbiome composition and drug type was found for fibromyalgia patients, although it has been demonstrated that clinical drugs have an impact upon microbiome composition, both antibiotic, non-antibiotic [149] and psychotropic [373] drugs. This lack of associations could be related to the reduced number of patients taking a specific drug family and/or to the interactions between different drugs taken. Proton pump inhibitors (PPI), for example, has been described to have anticommensal activity and was taken by nearly 30% of patients. One study has reported a reduction in Lachnospiraceae and Ruminococcaceae in PPI consumers [374], which is quite consistent with our observations in fibromyalgia patients. Another study could replicate these results, adding also a reduction in *Bifidobacterium* genus in PPI consumers [375]. Both studies also reported a decrease in α-diversity when PPI were taken, consistent with our findings too. Related to psychotropics, it has been reported that they target a similar pattern of bacterial species independently of their chemical similarity, thus suggesting that the anticommensal activity of these drugs may be a part of their mechanism of action instead of a secondary effect [373] We didn't observe any microbiome alteration that could be associated to the ones that have been reported for antidepressant drugs, neither for tricyclic (taken by 12% of patients) nor selective serotonin reuptake inhibitors (SSRI) antidepressants (54% of patients). Regarding the antiepileptic drugs (taken by 29% of patients), it has been shown that neither lithium nor valproate have a significant anticommensal activity, although lithium may increase the relative abundance of Ruminococcaceae and reduce the Bacteroides one while valproate alters the levels of SCFA [376], alterations that we reported to occur in fibromyalgia patients too. Finally, while non antimicrobial activity has been described for morphine [377], opioids (prescribed to 45% of patients) chronic use has been associated with a reduction of Bacteroidaceae, which we also observed in fibromyalgia patients, and Ruminococcaceae [378]. Although no specific associations between specific drugs and microbiome composition were found, probiotics could be useful in the treatment of fibromyalgia as they affect the microbiome composition [326]. Notably, several authors have used this approach to treat chronic fatigue syndrome [379] and one pilot study has examined the effects of probiotics on fibromyalgia patients

[380]. The authors have shown some improvements, mainly in depression symptoms and impulsive behaviour, in comparison with the placebo group [381].

CONCLUSIONS

To the best of our knowledge, this is the first study to report differences between the microbiome composition of fibromyalgia patients and healthy controls. We provided a list of these differences and reported the alterations in the levels of various molecules in the fibromyalgia sufferers, which might be useful as diagnostic biomarkers. We examined the functionality of these molecules and found that the most altered metabolic pathways were related to neurotransmitters, such as glutamate and nitric oxide. We checked possible interactions between the gut microbiome and serum metabolome; our analysis found several individual correlations between the two datasets. We also demonstrated that the combined microbiome and serum metabolome analyses could discriminate between fibromyalgia patients and control individuals. Thus, we report a new set of molecules and bacteria that might improve the diagnosis process, compensating for the current lack of objective biomarkers. Our results should help to shed some new light on the pathogenesis of this disease, provide biomarkers within a biological framework and improve our knowledge of this relatively unknown disease.

Discussion

Just for once, let me look on you with my own eyes.

Return of the Jedi, 1983.

Biomarker discovery is a complicated process, with a low success rate [7–10]. This low percentage of success has worsened with the apparition and generalization of high-throughput technologies, such are the omics ones. The capability to analyze an enormous number of molecules simultaneously has provided with more potential biomarkers that lately fail to be validated, either using validation cohorts and/or by experimental means. Better approaches, technologies and protocols are thus needed in order to improve this success ratio.

This Thesis work stablishes an analytical pipeline that could improve the biomarker discovery effectiveness, by providing more robust candidate molecules. Thus, a combination of distinct omics technologies with bioinformatics tools with the idea to provide a better comprehension of the biological alterations that could explain the alteration of each potential biomarker was studied. It was expected that with a better-defined biological context for the molecules altered the list of potential biomarkers would be more robust, leading to less of them being discarded in the validation process. This hypothesis of binging closer the data to the biological context, has been applied to three practical cases with clinical relevance: the early prostate cancer urinary EVs derived biomarkers; the fibromyalgia multi-omics biomarkers identification; and the colorectal cancer metabolomics-microbiome biomarkers identification, including the advanced adenoma early stage of disease.

For the prostate cancer (Results Chapter 2), metabolomics analysis was performed and the corresponding results combined with publicly available transcriptomics datasets [244]. This combination allowed a better support for metabolomics findings on the altered metabolites, providing with an explanation for that alteration at gene regulation level and with the confirmation of those alteration in other cohorts, although in an indirect way. Another important point of relevance of the project was the choice of the samples to analyze. In this case, urinary EVs were used for the biomarker discovery process. Urine is the most proximal biofluid to prostate [254], thus its metabolome will reflect better the prostatic alterations than other biofluids located further away. The utilization of EVs for biomarker identification could also alleviate the low rates of success. Since the use of urine required the concentration of the sample, EVs represent a concentrated source of molecules [19]. EVs themselves have been shown to be different between early disease stage and later ones, both structurally and in their contents. Since cancer cells seems to release different kind of EVs than normal cells, using EVs as a source for biomarkers may improve the process, identifying more robust candidates because less background molecules may be included in the analysis.

With the fibromyalgia project (Results Chapter 4), a different approach was followed. Since only one cohort was available to test the validity of the identified potential biomarkers, a combination of distinct omics technologies performed upon different sample types of the same individuals was used. Thus, sera metabolome, circulating miRNAs, cytokines and peptides with the fecal microbiome were combined and the interactions, similarities and potential influences between them were studied [237]. Because most of the differences found were related to fecal microbiome and serum

metabolome, the potential role of gut microbiota upon the host's metabolome was inspected. Therefore, one of the criteria applied for ensuring biomarker candidates' robustness was to identify potential roles for altered bacteria upon host's phenotype, by correlating their abundance with metabolite levels. A functional explanation for these correlations was explored, by looking in bacterial metabolomics databases. Finally, the last validation strategy was experimental. Since glutamate related metabolites alterations were identified and these alterations correlated with specific bacterial genes, qPCR analysis was performed on these bacterial genes. This way, a functional confirmation of the difference in taxonomical levels identified by bioinformatics means was obtained.

Finally, a combination of metabolomics and microbiome of fecal samples was explored to identify potential early biomarkers for CRC. While microbiome data acquisition was performed on all samples at once, metabolomics data was acquired and analyzed in two steps. In the first one, the fecal metabolome of 129 individuals was analyzed [449] and those results used to generate a metabolite-composed predictive model, to analyze how good the potential biomarkers could be, dividing the samples in 80% going to model training and 20% to model validation. In order to avoid potential lifestyle and/or population-related confounding factors during the model metabolites selection step, this specific distribution was randomly generated up to 10,000 times, evaluating the model on all of them. Later, the second part of the study, 116 samples, was used as a validation cohort for the model published. CRC, though, is known to be highly associated with alterations on the host's normal gut microbiota [450, 451, 458]. Because of the location of gut microbiota, it has a relevant impact upon fecal metabolome composition too [436, 463]. Therefore, when considering which omics to perform in order to better identify, characterize and explain biomarker candidates, metabolomics and microbiome came to be the most logical options. In fact, these multi-omics analytical methods showed that metabolomics and microbiome revealed higher similarities and interactions between both of them. In a final integration step, the published metabolite-related predictive model was updated to include selected bacterial genera identified to be also differentially abundant.

In summary, in the three projects presented in this thesis work, the best way to use high-throughput omics technologies for biomarker discovery studies has been analyzed, in order to try to convert the raw data into biological context. Several aspects have been specially considered, ranging from the project design to the final analytical tools. All the projects presented have been carefully designed in such a way that would comply with the biomarker discovery checklist. In the case it was not feasible to comply with some of the criteria, strategies were proposed and followed in order to reduce the potential drawbacks of such compliance failure.

In all projects, the best biospecimen for each disease studied was tried to be analyzed. In both cancer-related projects, the most accurate biospecimen was considered to be the most proximal one to the tumor itself. Thus, the selection of urine for PCa and stool for CRC seemed logical. Biospecimen to analyze for the fibromyalgia project was harder,

though. Because of the characteristics of the disease no concrete location of the body can be associated with any main symptomatology. Therefore, a wider approach seemed to be more indicated. This is why different blood fractions and fecal samples were collected from this cohort individuals. Another relevant factor to consider when choosing which biospecimens to collect is what kind of analysis will be performed lately on those samples. For example, metabolomics experimental protocols will depend on the sample analyzed, as will microbiome sequencing ones too. Other molecules won't even be detected depending on the biospecimen used.

An important element when performing either single omics and/or multi-omics studies is the ability to isolate real, phenotype-specific alterations from background noise and alterations due to confounding factors. The large number of variables included in one omics study makes it inevitable to identify data patterns and variable alterations that will be associated with lifestyle, clinical and/or other confounding factors, such as the geographical location of the sample. This issue occurred with fibromyalgia metabolomics, for example, seeing that samples could be completely discriminated depending on the sample's hospital origin. In fact, the influence of confounding factors upon the omics data will depend on the omics itself. Thus, the influence of environmental factors will be higher for microbiome or metabolomics than for genomics. Having bad metadata is, therefore, an important drawback for any omics study. One option to tackle this issue is to increase the number of samples included in the study. This, obviously, will have an important impact upon the costs of the study, but will generate more robust associations and may provide better resolution for the biomarker discovery process. Another option is to try to group common features under one category of the metadata, so that more samples will be included in the same category and statistics may be more robust. This approach was followed for the identification of the potential effects of fibromyalgia patients' drug regimens upon microbiome composition, grouping each individual drug by drug type and indication. This way, patients subgroups with enough sample size were generated, allowing the performance of more robust differential statistical tests. A final option, that could also tackle the small sample size issues, is to perform multi-omics studies instead of single omics one. This way, contrasting and comparison of alterations found by one omics technology with another omics one becomes feasible. Apart from identifying more certain alterations, explained by more than one omics layer, this will also discard some potential confounding factors, because not too many of them are able to affect multiple omics layers at the same time and in the same way.

Related to these confounding factors, the use of different, external validation cohort is highly recommended. Ideally, this validation cohort should be completely independent of the cohort used to generate the corresponding predictive model, including factors such as the geographic location of both cohorts, different sampling time points, sample processing, etc. A biomarker candidate that could be validated in a completely different and independent cohort will have much better options in posterior patenting and commercialization projects. When this completely different validation cohort is not an

option to consider, for whatever reasons, reducing as much as possible the potential effects of sample-related confounding factors should provide highly robust biomarker candidates. For this reason, this strategy in which the cohort was randomly divided into training and testing subpopulations up to 10,000 times was used when selecting which variables should be included in the predictive model. This approach allowed the generation of 10,000 different populations, each one of them with different individuals combination so that all the set of confounding factors that could influence the outcome of the variable selection step were included. This way, a predictive regression model that robust against any factor that may not be related to the phenotype trying to predict could be generated. This approach was shown to be quite effective for the CRC multi-omics project, where in the first part of the project a model for metabolomics data using this strategy was constructed that later was showed to be good enough to be validated within the new cohort of samples introduced in the second part of the project.

Economically, though, multi-omics studies are expensive, so that not all research groups may be able to afford them. Luckily, most research journals require the release of raw data for any publication. This means that there are full datasets being publicly available and that any researcher may use it. It's as easy as downloading and processing one (or more) dataset that has been generated by other authors studying the same phenotype. This was done for PCa project [244] and CRC first part of the project [449], mixing our own metabolomics datasets with transcriptomics obtained from publicly available databases.

One feature that has shown to be important for the validity of biomarkers is their functional characterization what is one of the major aims of this this project. In this aspect, the work in the bioinformatics package to retrieve information from the most commonly used metabolomics databases (KEGG and HMDB) has been useful to aid and accelerate this step. To our knowledge, no app exists that combines as many databases as our approach does, both at physicochemical and functional levels. The most relevant point, though, is that that tool allows the batch search of multiple metabolites (for KEGG and HMDB databases) and/or genes, for the rest of the functions included in the package. This metabolite > enzyme > genes > functionalities path allows the generation of a straightforward analytical pipeline, needing only metabolites codes to perform the complete search.

While dimension reduction techniques have been shown to be useful for distinct datasets integration and comparison, their utility as a biomarker identification strategy is less clear. A reason for that is the dimensionality reduction approach itself because it combines distinct variables into a new one, the principal component, which is responsible for the sample groups differentiation. The identification of individual biomarkers and their potentiality for sample discrimination is, therefore, less straight-forward. Instead, dimension reduction methodologies are useful for identifying alteration patterns that later may be used for biomarkers functional characterization, such as metabolite families and/or specific bacterial phyla. To identify individual biomarkers, the correlation-based approaches were found to be far more informative,

most probably due to the fact that correlations were established between individual variables. Relevantly, the same patterns were identified in both approaches, with the same variables being identified to be more relevant to either the sample groups description and/or prediction. Thus, and seemingly to what happens with metabolomics data analysis [97], the combination of a more global approach as dimension reduction methods could be with a more refined one like the correlations analysis was recommended, as more complete information and better-explained results were obtained from the same set of data. Including a final step in which regression models with the combined datasets were studied was also helpful in order to prove that the identified interactions between distinct omics datasets exist and can be exploited.

Finally, although one of the aims of this thesis was to define a specific, standardized and re-usable pipeline for multi-omics integration studies, the work performed in this thesis suggests that, with the interest of performing the best analysis possible, each pipeline should be adapted to the specifics of the corresponding project. Actually, this can be seen in the three projects presented here, each one with its specific analytical protocol, although some common methods have been applied too. In summary, for a well-performed multi-omics study the following considerations are suggested:

1) Process and analyze each omics separately. The incorporation of specialized researchers in this step will suppose also a better analyzed, good quality data. Test different processing and normalization methodologies to identify which better suits the characteristics of the dataset and/or samples.
2) With the normalized data, perform a range of multi-omics integration, both univariate and multivariate.
3) No approaches should be discarded from the start. Although they may not be as informative as expected, each one is intended to explain specific dataset characteristic or interaction.
4) Biomarkers should be identified or related to more than one omics dataset. This ensures higher robustness of the candidates so that less of them will be lost in validation steps.
5) Experimental validation of bioinformatics results is helpful to confirm or discard some biomarker candidates. This step should always be considered in any study with high dependence on bioinformatics.
6) Multi-omics results can be directly associated with the quality of metadata, as that is the factor that will ensure that differences observed can be (or cannot be) associated with the studied phenotype.

4.1.- Limitations and considerations

In order to tackle the lack of success rate for the biomarker discovery process, several approaches have been presented in the three projects performed in this thesis. The principal limitation of these studies, though, is that two of them have been more prospective projects, without a real validation step using other completely different

cohorts. Even when the identified biomarkers were validated in a different cohort, no significant environmental and lifestyle-related factors were different between cohorts so that we could not rule out the role of confounding factors among the predictability of our biomarkers. Therefore, it's evident that strong international collaborations and consortia are a great tool for biomarker discovery research studies, so that completely different cohorts may be easily accessible for researchers. It is obvious though, that for matters of time this kind of validation with international cohorts was not a feasible objective of the projects presented in this thesis.

It is clear that bioinformatics alone will not resolve the issues presented in this work, even if far better methodological and analytical tools are developed. In fact, the use of only bioinformatics to tackle those challenges was never considered in this thesis. Instead, a list of checkpoints is proposed that, followed during the biomarker discovery process, may lead to a better and more robust list of biomarker candidates. Bioinformatics, thus, should provide a list of candidates, not a definitive and invariable list. Instead, every candidate must be proven to work by experimental methods, laboratory work and clinical trials when that point of the process is reached. Bioinformatics aims, therefore, should be limited to refine and improve the biomarker candidates selection steps, but validating them is a task for another scientific field.

The combination of distinct omics technologies in a single project is not cheap and requires the implication of several researchers, each one specialized in one of those omics. Each omics technologies also have their own equipment, protocols and data formats. If the goal of the multi-omics research field is to achieve a standardization of protocols and data formats, this needs to be tackled from the data acquisition step. Therefore, a joint effort is needed from all the high-throughput research fields to change the current protocols and standards. Finally, the elevated cost of these kinds of technologies makes it hard to globalize and democratize their use. Consequently, if the goal of the omics research field is to become a standard option for biomarkers discovery, new methods and equipment need to be developed that reduces their costs.

Conclusions

Hope.

A New Hope, 1977.

Rogue One, 2016.

- The data mining tool we have developed can aid in the functional description of potential metabolite biomarkers, helping this way to provide a biological context for their alteration, strengthen this way the selection process of robust biomarkers.

- High-throughput tools allow the identification of a large number of potential biomarkers for a range of diseases due to the ability to measure thousands of variables simultaneously.
- A combination of distinct omics technologies has demonstrated to be a useful approach for robust biomarker candidates' identification.
- Analytical methods must be carefully considered for each omics, taking into consideration the data characteristics and structure, which will influence the output of the analytical process.
- The metadata collection process needs to be carefully controlled and supervised, including in the process the opinion of the final data analyst.
- Standardization of experimental protocols and data formats for single omics and a combination of them is still unresolved.

- Urinary EVs are a good source for the biomarker discovery process, due to their easy access and the reduction of sample processing steps, especially for diseases related to the urogenital tract.
- PCa EVs show altered morphology and metabolite content compared to BPH ones.
- Metabolites altered in PCa EVs include PCs, which are related to cellular and extracellular membranes, arachidonic acid, suggesting alterations on inflammatory modulators and metabolites related to steroid hormones.
- Stage 3 PCa patients presented decreased levels of metabolites related to ceramides, acylcarnitines and glycerophospholipids when compared to stage 2 patients.
- Stage 3 PCa patients with perineural invasion were found to have elevated levels of androsterone sulfate + etiocholanolone sulfate and lower levels of cAMP when compared to patients without perineural invasion.
- Metabolomics results for PCa patients were concordant with gene expression alterations identified by independent transcriptomics studies.

- Fibromyalgia patients present microbiota dysbiosis events, such as richness reduction and reduced abundance of SCFA producer bacteria. Furthermore, these alterations couldn’t be associated with any of the clinical metadata.
- Fibromyalgia patients’ microbiome is less abundant on glutamate degradation bacterial enzymes, as identified by 16S sequencing analysis and confirmed by experimental methodologies.

- Fibromyalgia metabolomics analysis revealed alterations on neurotransmitters levels, like glutamate and on metabolites related to NO biosynthesis pathways.
- Microbiome and metabolomics integration for fibromyalgia patients revealed the certain influence of gut microbiome upon serum metabolome, with correlations found for gut bacteria and glutamate levels.
- The combination of the 4 omics datasets was able to identify fibromyalgia patients better than any omics alone.

- Colorectal cancer fecal metabolomics revealed alterations in mainly three metabolite families: cholesteryl esters, ceramides and sphingomyelins.
- A combination of 6 metabolites is predictive enough to be able to identify CRC patients from AD and control individuals, both in the training and validation cohort.
- Microbiome analysis identifies alterations for CRC individuals when compared to C and AD, but is not able to discriminate between the two latter groups.
- CRC patients presented increased abundance of *Bulleidia,* Erysipelotrichaceae (family), *Fusobacterium, Gemella, Butyrivibrio, Peptococcus, Peptostreptococcus, Staphylococcus, Streptococcus, Parvimonas* and *Selenomonas.*
- CRC patients presented decreased abundance of Lachnospiraceae family bacteria.
- AD presents an increased abundance of *Adlercreutzia* when compared to C and CRC individuals
- Specific genera abundance alterations either increases or decreases with disease progression.
- Multi-omics integration of CRC data identified both metabolomics and microbiome datasets to be similar, with strong correlations identified between altered metabolites and bacteria.
- A combination of biomarkers obtained from fecal metabolomics and microbiome generated a better predictive model than each omics separately.

General conclusion:

- The results of the three projects included in this thesis demonstrate the potential and utility of combining distinct omics in order to improve the biomarker discovery process and the identification of altered metabolic pathways that may explain specific diseases pathogenesis.

- La herramienta de *data mining* que hemos desarrollado puede ayudar en la descripción funcional de los potenciales metabolitos biomarcadores, ayudando a proveer con un contexto biológico para su alteración, fortaleciendo así el proceso de selección de biomarcadores robustos.

- Las herramientas de alto rendimiento permiten la identificación de un elevado número de biomarcadores potenciales para múltiples enfermedades a causa de la capacidad para medir miles de variables de forma simultánea.
- La combinación de distintas tecnologías ómicas ha demostrado ser un método útil para la identificación de candidatos robustos a biomarcador.
- Los métodos analíticos deben ser considerados de forma cautelosa para cada tecnología ómica, teniendo en cuenta las características y estructura de los datos, que afectaran sobre el resultado del proceso de análisis.
- El proceso de recolección de los metadatos clínicos tiene que estar muy estrechamente controlado y supervisado y debería incluir en este proceso la opinión del responsable final del análisis de datos.
- La estandarización de los protocolos analíticos y del formato de los datos, tanto para las ómicas individuales como en combinación, no está todavía resuelta.

- Las EVs de orina son una buena fuente para el proceso de identificación de nuevos biomarcadores, por su facilidad de obtención y la reducción de los pasos de procesado de muestras, especialmente para enfermedades relacionadas con el tracto urogenital.
- Las EVs de PCa muestran alteraciones en su morfología y contenido en metabolitos en comparación a las EVs de BPH.
- Los metabolitos alterados en las EVs de PCA incluyen PCs, que se relacionan con las membranas celular y extracelular, el ácido araquidónico, sugiriendo así alteraciones en moduladores de la inflamación y metabolitos relacionados con las hormones esteroides.
- Los pacientes de PCa estadío 3 presentaban niveles reducidos de metabolitos relacionados con las ceramidas, acilcarnitinas y glicerofosfolípidos en comparación a los pacientes en estadío 2.
- Los pacientes de PCa en estadío 3 con invasión perineural tenían niveles elevados de sulfato de androesterona + sulfato de etiocolanolona y niveles reducidos de cAMP en comparación con los pacientes sin invasión.
- Los resultados de metabolómica de los pacientes de PCa concordaban con las alteraciones en la expresión genética identificadas por estudios independientes de transcriptómica.

- Los pacientes de fibromialgia presentan eventos de disbiosis en su microbioma, como la reducción de diversidad y de bacterias productoras de SCFA. Además, estas alteraciones no se pudieron asociar a ningún parámetro clínico.

- El microbioma de los pacientes de fibromialgia presentaba una reducción en la abundancia de las enzimas bacterianas degradadoras de glutamato, como se identificó con el análisis de la secuenciación del gen 16S y se confirmó experimentalmente.
- El análisis de la metabolómica de fibromialgia identificó alteraciones en los niveles de neurotransmisores, como el glutamato y metabolitos relacionados con la ruta de biosíntesis de NO.
- La integración de datos de microbioma y metabolómica de los pacientes identificó una cierta influencia del microbioma intestinal en el metaboloma del suero, con correlaciones identificadas entre las bacterias intestinales y los niveles de glutamato.
- La combinación de las 4 tecnologías usadas permitió identificar los pacientes de fibromialgia mejor que ninguna tecnología por si sola.

- La metabolómica de las heces de los pacientes de cáncer colorrectal identificó alteraciones en principalmente tres familias de metabolitos: esteres de colesterol, cermidas y esfingomielinas.
- Una combinación de 6 metabolitos tiene suficiente capacidad predictiva para diferenciar a los pacientes de CRC de los de AD y los controles sanos, tanto en la cohorte de entrenamiento como la de validación.
- El análisis de microbioma identifica alteraciones para los individuos con cáncer colorrectal comparados con los grupos C y AD, pero no es capaz de discriminar entre estos dos últimos grupos.
- Los pacientes de CRC presentaban una mayor abundancia de *Bulleidia,* Erysipelotrichaceae (familia), *Fusobacterium, Gemella, Butyrivibrio, Peptococcus, Peptostreptococcus, Staphylococcus, Streptococcus, Parvimonas* y *Selenomonas.*
- Los pacientes de CRC mostraban menor abundancia de bacterias de la familia Lachnospiraceae.
- Los individuos con AD presentan mayor abundancia de *Adlercreutzia* en comparación con los individuos de los grupos C y CRC.
- Las alteraciones en abundancia de géneros específicos identificadas o se incrementan o disminuyen con la progresión de na enfermedad.
- La integración de multi-ómicas de los datos de CRC demostraron la existencia de similitudes entre los datos de microbioma y metabolómica, con correlaciones significativas entre metabolitos y bacterias alteradas.
- Una combinación de biomarcadores derivados de la metabolómica fecal y el microbioma permitieron generan un modelo predictivo mejor que los generados con las ómicas por separado.

Conclusión general:

- Los resultados de los tres proyectos que se incluyen en esta tesis demuestran el potencial y utilidad de combinar distintas tecnologías ómicas para mejorar el proceso de identificación de nuevos biomarcadores y la identificación de potenciales alteraciones en rutas metabólicas que puedan explicar la patogénesis de enfermedades concretas.

References

A long time ago in a galaxy far, far away…

A New Hope, 1977.

The Empire Strikes Back, 1980.

Return of the Jedi, 1983.

[1] World Health Organization (WHO) 1993. *Environmental Health Criteria 155 Biomarkers and Risk Assessment : Concepts and Principles*.
[2] National Research Council 1989. *Biologic Markers in Reproductive Toxicology*.
[3] Henderson, B.R.F. et al. 1989. The Use of Biological Markers in Toxicology. *Toxicology*. (1989), 65–82.
[4] Foroutan, B. 2015. Personalized Medicine: A Review with Regard to Biomarkers. *Journal of Bioequivalence & Bioavailability*. 07, 06 (2015), 244–256.
[5] Mayeux, R. 2004. Biomarkers: Potential Uses and Limitations. *NeuroRx, The Journal of the American Society for Experimental NeuroTherapeutics*. 1, 2 (2004), 182–188.
[6] Strimbu, K. and Tavel, J. a 2011. What are Biomarkers? *Curr Opin HIV AIDS*. 5, 6 (2011), 463–466.
[7] Goossens, N. et al. 2015. Cancer biomarker discovery and validation. *Translational Cancer Research*. 4, 3 (2015), 256–269.
[8] Tzoulaki, I. et al. 2011. Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: Meta-epidemiology study. *BMJ*. 343, (2011).
[9] Ioannidis, J.P.A. and Panagiotou, O.A. 2011. Comparison of Effect Sizes Associated With Biomarkers Reported in Highly Cited Individual Articles and in Subsequent Meta-analyses. *JAMA*. 305, 21 (2011).
[10] Poste, G. 2011. Bring on the biomarkers. *Nature*. 469, 7329 (2011), 156–157.
[11] Ptolemy, A.S. and Rifai, N. 2010. What is a biomarker? Research investments and lack of clinical integration necessitate a review of biomarker terminology and validation schema. *Scandinavian Journal of Clinical and Laboratory Investigation*. 70, SUPPL. 242 (2010), 6–14.
[12] Simon, R. 2010. Clinical trials for predictive medicine: New challenges and paradigms. *Clinical Trials*. 7, 5 (2010), 516–524.
[13] Noguera-Julian, M. et al. 2016. Gut Microbiota Linked to Sexual Preference and HIV Infection. *EBioMedicine*. 5, (2016), 135–146.
[14] Duffy, M.J. et al. 2015. Validation of new cancer biomarkers: A position statement from the european group on tumor markers. *Clinical Chemistry*. 61, 6 (2015), 809–820.
[15] Thompson, M.L. and Zucchini, W. 1989. On the statistical analysis of ROC curves. *Statistics in Medicine*. 8, 10 (1989), 1277–1290.
[16] Macías, M. et al. 2018. Liquid Biopsy: From Basic Research to Clinical Practice. *Advances in Clinical Chemistry*. 83, (2018), 73–119.
[17] Poulet, G. et al. 2019. Liquid Biopsy: General Concepts. *Acta Cytologica*. (2019).
[18] Théry, C. et al. 2002. Exosomes: composition, biogenesis and function. *Nature reviews. Immunology*. 2, 8 (2002), 569–579.
[19] Yáñez-Mó, M. et al. 2015. Biological properties of extracellular vesicles and their physiological functions. *Journal of Extracellular Vesicles*. 4, 2015 (2015), 1–60.
[20] Mader, S. and Pantel, K. 2017. Liquid biopsy: Current status and future perspectives. *Oncology Research and Treatment*. 40, 7–8 (2017), 404–408.
[21] Witwer, K.W. et al. 2013. Standardization of sample collection, isolation and analysis methods in extracellular vesicle research. *Journal of Extracellular Vesicles*. 2, 1 (2013).
[22] Théry, C. et al. 2001. Proteomic Analysis of Dendritic Cell-Derived Exosomes: A Secreted Subcellular Compartment Distinct from Apoptotic Vesicles. *The Journal of Immunology*. 166, 12 (2001), 7309–7318.
[23] Szczepanski, M.J. et al. 2011. Blast-derived microvesicles in sera from patients with acute myeloid leukemia suppress natural killer cell function via membrane-associated transforming growth factor-β1. *Haematologica*. 96, 9 (2011), 1302–1309.
[24] Raposo, G. and Stoorvogel, W. 2013. Extracellular vesicles: Exosomes, microvesicles, and friends. *Journal of Cell Biology*. 200, 4 (2013), 373–383.
[25] Hornberg, J.J. et al. 2006. Cancer: A Systems Biology disease. *BioSystems*. 83, 2-3 SPEC. ISS. (2006), 81–90.
[26] Kitano, H. 2004. Cancer as a robust system: Implications for anticancer therapy. *Nature Reviews Cancer*. 4, 3 (2004), 227–235.
[27] Wruck, W. et al. 2015. Multi-omic profiles of human non-alcoholic fatty liver disease tissue highlight heterogenic phenotypes. *Scientific Data*. 2, (2015), 1–10.
[28] Thiele, I. et al. 2013. A community-driven global reconstruction of human metabolism. *Nature Biotechnology*. 31, 5 (2013), 419–425.
[29] Nielsen, J. 2017. Systems Biology of Metabolism. *Annual Review of Biochemistry*. 86, 1 (2017),

245–275.
[30] Cisek, K. et al. 2016. The application of multi-omics and systems biology to identify therapeutic targets in chronic kidney disease. *Nephrology, dialysis, transplantation*. 31, 12 (2016), 2003–2011.
[31] Mardis, E.R. 2008. Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*. 9, 1 (2008), 387–402.
[32] Wang, Z. et al. 2009. RNA-Seq: a revolutionary tool for transcriptomics in Western Equatoria State. *Nature Reviews Genetics*. 10, 1 (2009), 57.
[33] Nassar, A.F. et al. 2017. UPLC–MS for metabolomics: a giant step forward in support of pharmaceutical research. *Drug Discovery Today*. 22, 2 (2017), 463–470.
[34] Beale, D.J. et al. 2018. *Review of recent developments in GC–MS approaches to metabolomics-based research*. Springer US.
[35] Hasin, Y. et al. 2017. Multi-omics approaches to disease. *Genome Biology*. 18, 1 (2017), 83.
[36] Pinu, F.R. et al. 2019. Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. *Metabolites*. 9, 4 (2019), 76.
[37] Fondi, M. and Liò, P. 2015. Multi -omics and metabolic modelling pipelines: Challenges and tools for systems microbiology. *Microbiological Research*. 171, (2015), 52–64.
[38] Dihazi, H. et al. 2018. Integrative omics - from data to biology. *Expert Review of Proteomics*. 15, 6 (2018), 463–466.
[39] Perez-Riverol, Y. et al. 2017. Discovering and linking public omics data sets using the Omics Discovery Index. *Nature Biotechnology*. 35, 5 (2017), 406–409.
[40] Hastings, J. et al. 2019. Multi-Omics and Genome-Scale Modeling Reveal a Metabolic Shift During C. elegans Aging. *Frontiers in Molecular Biosciences*. 6, February (2019), 1–18.
[41] Zeevi, D. et al. 2015. Personalized Nutrition by Prediction of Glycemic Responses. *Cell*. 163, 5 (2015), 1079–1094.
[42] Henry, V.J. et al. 2014. OMICtools: an informative directory for multi-omic data analysis. *Database : the journal of biological databases and curation*. 2014, 13 (2014), 1–5.
[43] Mangul, S. et al. 2019. Systematic benchmarking of omics computational tools. *Nature Communications*. 10, 1 (2019), 1–11.
[44] Wilkinson, M.D. et al. 2016. Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 3, (2016), 1–9.
[45] Misra, B.B. et al. 2018. Integrated omics: tools, advances and future approaches. *Journal of Molecular Endocrinology*. 2016 (2018), R21–R45.
[46] Min, S. et al. 2017. Deep learning in bioinformatics. *Briefings in bioinformatics*. 18, 5 (2017), 851–869.
[47] Libbrecht, M.W. and Noble, W.S. 2015. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*. 16, 6 (2015), 321–32.
[48] Li, Y. et al. 2018. A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics*. 19, 2 (2018), 325–340.
[49] Meng, C. et al. 2016. Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*. 17, 4 (2016), 628–641.
[50] Leek, J.T. et al. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*. 11, 10 (2010), 733–739.
[51] Paliy, O. and Shankar, V. 2016. Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology*. 25, 5 (2016), 1032–1057.
[52] Ringnér, M. 2008. What is principal component analysis? *Nature Biotechnology*. 26, 3 (2008), 303–304.
[53] Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*. 2, (1901), 559–572.
[54] Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*. 24, 7 (1933), 498–520.
[55] Wall, M.E. et al. 2003. Singular Value Decomposition and Principal Component Analysis BT - A Practical Approach to Microarray Data Analysis. *Briefings in Functional Genomics and Proteomics*. D.P. Berrar et al., eds. Springer US. 91–109.
[56] Wouters, L. et al. 2003. Graphical Exploration of Gene Expression Data: A Comparative Study of Three Multivariate Methods. *Biometrics*. 59, (2003), 1131–1139.
[57] Parkhomenko, E. et al. 2009. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*. 8, 1 (2009).

[58] Waaijenborg, S. and Zwinderman, A.H. 2009. Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. *BMC Bioinformatics*. 10, (2009), 315.
[59] Lê Cao, K.A. et al. 2009. Sparse canonical methods for biological data integration: Application to a cross-platform study. *BMC Bioinformatics*. 10, (2009), 1–17.
[60] Gower, J.C. 1971. Statistical methods of comparing different multivariate analyses of the same data. *Mathematics in the archaeological and historical science*. 138–149.
[61] Fagan, A. et al. 2007. A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics*. 7, 13 (2007), 2162–2171.
[62] Culhane, A.C. et al. 2003. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*. 4, 59 (2003).
[63] Dolédec, S. and Cheesel, D. 1994. Co-inertia analysis: an alternative method for studying species–environment relationships. *Freshwater Biology*. 31, 3 (1994), 277–294.
[64] Wang, Q. et al. 2019. Host and microbiome multi-omics integration: applications and methodologies. *Biophysical Reviews*. 11, 1 (2019), 55–65.
[65] Ge, T. et al. 2018. Phenome-wide heritability analysis of the UK Biobank. *PLOS Genetics*. 13, 4 (2018), e1006711.
[66] Bonder, M.J. et al. 2016. The effect of host genetics on the gut microbiome. *Nature Genetics*. 48, (Oct. 2016), 1407.
[67] Igartua, C. et al. 2017. Host genetic variation in mucosal immunity pathways influences the upper airway microbiome. *Microbiome*. 5, 16 (2017).
[68] Wang, J. et al. 2016. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nature Genetics*. 48, (Oct. 2016), 1396.
[69] Schwartz, S. et al. 2012. A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response. *Genome Biology*. 13, r32 (2012).
[70] McHardy, I.H. et al. 2013. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome*. 1, 1 (2013), 1–19.
[71] Pedersen, H.K. et al. 2016. Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature*. 535, 7612 (2016), 376–381.
[72] Weir, T.L. et al. 2013. Stool Microbiome and Metabolome Differences between Colorectal Cancer Patients and Healthy Adults. *PLoS ONE*. 8, 8 (2013).
[73] Nugent, J.L. et al. 2014. Altered Tissue Metabolites Correlate with Microbial Dysbiosis in Colorectal Adenomas. *Journal of Proteome Research*. 13, (2014), 1921–1929.
[74] Monteiro, M.S. et al. 2013. Metabolomics analysis for biomarker discovery: advances and challenges. *Current medicinal chemistry*. 20, 2 (2013), 257–71.
[75] Oliver, S.G. et al. 1998. Systematic functional analysis of the yeast genome. *Trends in Biotechnology*. 16, 9 (1998), 373–378.
[76] Gomase, V.S. et al. 2008. Metabolomics. *Current Drug Metabolism*. 9, (2008), 89–98.
[77] Wishart, D.S. et al. 2007. HMDB: The human metabolome database. *Nucleic Acids Research*. 35, SUPPL. 1 (2007), 521–526.
[78] Dayalan, S. et al. 2018. Metabolome Analysis. *Reference Module in Life Sciences*.
[79] Förster, J. et al. 2003. Genome-Scale Reconstruction of the Saccharomyces cerevisiae Metabolic Network. *Genome Research*. 13, (2003), 244–253.
[80] Psychogios, N. et al. 2011. The human serum metabolome. *PLoS ONE*. 6, 2 (2011).
[81] Redestig, H. et al. 2011. Exploring Matrix Effects and Quantification Performance in Metabolomics Experiments Using Artificial Biological Gradients. *Analytical Chemistry*. (2011), 5645–5651.
[82] Kapoore, R.V. and Vaidyanathan, S. 2016. Towards quantitative mass spectrometry-based metabolomics in microbial and mammalian systems. *Philosophical Transactions of the Royal Society*. 374, 2079 (2016).
[83] Thompson, J.A. and Markey, S.P. 1975. Quantitative Metabolic Profiling of Urinary Organic Acids By Gcms : Comparison of Isolation Methods. *Anal.Chem.* 47, 8 (1975), 1313–1321.
[84] Horning, E.C. and Horning, M.G. 1971. Metabolic profiles: gas-phase methods for analysis of metabolites. *Clinical chemistry*. 17, 8 (1971), 802–9.
[85] Courant, F. et al. 2014. Basics of mass spectrometry based metabolomics. *Proteomics*. 14, (2014), 23969–2388.
[86] Kim, K. et al. 2014. Mealtime, temporal, and daily variability of the human urinary and plasma metabolomes in a tightly controlled environment. *PLoS ONE*. 9, 1 (2014).

[87] Gika, H.G. et al. 2008. Liquid chromatography and ultra-performance liquid chromatography-mass spectrometry fingerprinting of human urine. Sample stability under different handling and storage conditions for metabonomics studies. *Journal of Chromatography A*. 1189, 1–2 (2008), 314–322.
[88] Lauridsen, M. et al. 2007. Human urine as test material in 1H NMR-based metabonomics: Recommendations for sample preparation and storage. *Analytical Chemistry*. 79, 3 (2007), 1181–1186.
[89] Vuckovic, D. 2012. Current trends and challenges in sample preparation for global metabolomics using liquid chromatography-mass spectrometry. *Analytical and Bioanalytical Chemistry*. 403, 6 (2012), 1523–1548.
[90] El-Aneed, A. et al. 2009. Mass spectrometry, review of the basics: Electrospray, MALDI, and commonly used mass analyzers. *Applied Spectroscopy Reviews*. 44, 3 (2009), 210–230.
[91] Holmes, J.C. and Morrell, F.A. 1956. Oscillographic Mass Spectrometric Monitoring of Gas Chromatography. *Applied Spectroscopy*. (1956).
[92] Todd, J.F.J. 1991. Recommendations for nomenclature and symbolism for mass spectroscopy (including an appendix of terms used in vacuum technology). *Pure and Applied Chemistry*. 63, 10 (1991), 1541–1566.
[93] Armitage, E.G. et al. 2015. Missing value imputation strategies for metabolomics data. *Electrophoresis*. 36, 24 (2015), 3050–3060.
[94] Hrydziuszko, O. and Viant, M.R. 2012. Missing values in mass spectrometry based metabolomics: An undervalued step in the data processing pipeline. *Metabolomics*. 8, (2012), 161–174.
[95] Burgess, K. et al. 2014. Chapter 10 - Metabolomics. S.B.T.-H. of P. and S.M. Padmanabhan, ed. Academic Press. 181–205.
[96] Hendriks, M.M.W.B. et al. 2011. Data-processing strategies for metabolomics studies. *TrAC - Trends in Analytical Chemistry*. 30, 10 (2011), 1685–1698.
[97] Saccenti, E. et al. 2014. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*. 10, 3 (2014), 361–374.
[98] Broadhurst, D.I. and Kell, D.B. 2006. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*. 2, 4 (2006), 171–196.
[99] Gromski, P.S. et al. 2015. A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta*. 879, (2015), 10–23.
[100] Guijas, C. et al. 2018. METLIN: A Technology Platform for Identifying Knowns and Unknowns. *Analytical Chemistry*. 90, 5 (2018), 3156–3164.
[101] Wishart, D.S. et al. 2018. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Research*. 46, D1 (2018), D608–D617.
[102] Sumner, L.W. et al. 2007. Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics*. 3, 3 (2007), 211–221.
[103] J. Carroll, A. 2012. Online Metabolomics Databases and Pipelines. *Metabolomics*. U. Roessner, ed. 47–72.
[104] Kanehisa, M. and Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. 28, 1 (2000), 27–30.
[105] Slenter, D.N. et al. 2018. WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*. 46, D1 (2018), D661–D667.
[106] Fabregat, A. et al. 2018. The Reactome Pathway Knowledgebase. *Nucleic acids research*. 46, D1 (2018), D649–D655.
[107] Human Microbiome Project Consortium, T. et al. 2012. Structure, function and diversity of the healthy human microbiome. *Nature*. 486, 7402 (2012), 207–214.
[108] Qin, J. et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. 464, 7285 (2010), 59–65.
[109] Turnbaugh, P.J. et al. 2007. The Human Microbiome Project. *Nature*. 449, 7164 (2007), 804–810.
[110] Knight, R. et al. 2017. The Microbiome and Human Biology. *Annual Review of Genomics and Human Genetics*. 183, March (2017), 65–86.
[111] Lane, N. 2015. The unseen world : reflections on Leeuwenhoek (1677) “Concerning little animals.” *Philosophical Transactions of the Royal Society of London*. 370, 20140344 (2015).
[112] Leewenhoeck, A. 1684. An abstract of a letter from Mr. Anthony Leevvenhoeck at Delft, dated Sep. 17. 1683. Containing some microscopical observations, about animals in the scurf of the

teeth, the substance call'd worms in the nose, the cuticula consisting of scales. *Philosophical Transactions of the Royal Society of London*. 14, 159 (1684), 568–574.
[113] Langendijk, P.S. et al. 1995. Quantitative fluorescence in situ hybridization of Bifidobacterium spp. with genus-specific 16S rRNA-targeted probes and its application in fecal samples. *Applied and Environmental Microbiology*. 61, 8 (1995), 3069–3075.
[114] Zuckerkandl, E. and Pauling, L. 1965. Molecules as documents of history. *Journal of Theoretical Biology*. 8, (1965), 357–366.
[115] Woese C, F.G. 1977. Phylogenetic structure of the prokaryotic domain. *Proceedings of the National Academy of Sciences*. 74, 11 (1977), 5088–5090.
[116] Ramazzotti, M. and Bacci, G. 2017. *16S rRNA-Based Taxonomy Profiling in the Metagenomics Era*. Elsevier Inc.
[117] Woese, C.R. et al. 1985. A Phylogenetic Definition of the Major Eubacterial Taxa. *Systematic and Applied Microbiology*. 6, 2 (1985), 143–151.
[118] Woese, C.R. 1987. Bacterial Evolution. *Microbiological reviews*. 51, 2 (1987), 221–271.
[119] Giovannoni, S.J. et al. 1990. Genetic diversity in Sargasso Sea bacterioplankton. *Nature*. 345, 6270 (1990), 60–63.
[120] G. Weisburg, W. et al. 1991. 16S Ribosomal DNA Amplification for Phylogenetic Study. *Journal of Bacteriology*. 173, 2 (1991), 697–703.
[121] Dubnau, D. et al. 1965. Gene conservation in Bacillus species. I. Conserved genetic and nucleic acid base sequence homologies. *Proceedings of the National Academy of Sciences of the United States of America*. 54, 2 (Aug. 1965), 491–498.
[122] Suau, A. et al. 1999. Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Applied and environmental microbiology*. 65, 11 (1999), 4799–807.
[123] Jovel, J. et al. 2016. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology*. 7, APR (2016), 1–17.
[124] Schloss, P.D. et al. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*. 75, 23 (2009), 7537–7541.
[125] Caporaso, J.G. et al. 2010. QIIME allows analysis of high- throughput community sequencing data. *Nature Methods*. 7, 5 (2010), 335–336.
[126] Bolyen, E. et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME2. *Nature Biotechnology*. 37, August (2019), 848–857.
[127] Gevers, D. et al. 2012. Bioinformatics for the Human Microbiome Project. *PLoS Computational Biology*. 8, 11 (2012).
[128] Lozupone, C. et al. 2013. Meta-analysis studies of the human microbiota. *Genome Research*. 23, (2013), 1704–1714.
[129] Morgan, X.C. et al. 2013. Biodiversity and functional genomics in the human microbiome. *Trends in Genetics*. 29, 1 (2013), 51–58.
[130] Yasuda, K. et al. 2015. Biogeography of the intestinal mucosal and lumenal microbiome in the rhesus macaque. *Cell Host and Microbe*. 17, 3 (2015), 385–391.
[131] Horz, H.-P. 2015. Archaeal Lineages within the Human Microbiome: Absent, Rare or Elusive? *Life*. 5, 2 (2015), 1333–1345.
[132] Minot, S. et al. 2013. Rapid evolution of the human gut virome. *Proceedings of the National Academy of Sciences*. 110, 30 (2013), 12450–12455.
[133] Reyes, A. et al. 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*. 466, 7304 (2010), 334–338.
[134] Virgin, H.W. 2014. The virome in mammalian physiology and disease. *Cell*. 157, 1 (2014), 142–150.
[135] Underhill, D.M. and Iliev, I.D. 2014. The mycobiota: Interactions between commensal fungi and the host immune system. *Nature Reviews Immunology*. 14, 6 (2014), 405–416.
[136] Loke, P. and Lim, Y.A.L. 2015. Helminths and the microbiota: parts of the hygiene hypothesis. *Parasite Immunology*. 37, 6 (2015), 314–323.
[137] Lloyd-Price, J. et al. 2016. The healthy human microbiome. *Genome Medicine*. 8, 1 (2016), 1–11.
[138] Shafquat, A. et al. 2014. Functional and phylogenetic assembly of microbial communities in the human microbiome. *Trends in Microbiology*. 22, 5 (2014), 261–266.
[139] Lloyd-Price, J. et al. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature*. (2017).

[140] Bäckhed, F. et al. 2012. Defining a healthy human gut microbiome: Current concepts, future directions, and clinical applications. *Cell Host and Microbe*. 12, 5 (2012), 611–622.
[141] Davenport, E.R. et al. 2017. The human microbiome in evolution. *BMC Biology*. 15, 1 (2017), 1–12.
[142] De Filippo, C. et al. 2010. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences*. 107, 33 (2010), 14691–14696.
[143] Carmody, R.N. et al. 2015. Diet dominates host genotype in shaping the murine gut microbiota. *Cell Host & Microbe*. 17, 1 (2015), 72–84.
[144] Rautava, S. et al. 2012. Microbial contact during pregnancy, intestinal colonization and human disease. *Nature Reviews Gastroenterology and Hepatology*. 9, 10 (2012), 565–576.
[145] Alcántara, C. et al. 2018. Preterm Gut Microbiome Depending on Feeding Type: Significance of Donor Human Milk. *Frontiers in Microbiology*. 9, June (2018), 1–10.
[146] Boix-Amorós, A. et al. 2016. Relationship between milk microbiota, bacterial load, macronutrients, and human cells during lactation. *Frontiers in Microbiology*. 7, APR (2016), 1–9.
[147] Pannaraj, P.S. et al. 2017. Association between breast milk bacterial communities and establishment and development of the infant gut microbiome. *JAMA Pediatrics*. 171, 7 (2017), 647–654.
[148] Nicholson, J.K. et al. 2012. Host-gut microbiota metabolic interactions. *Science*. 336, 6086 (2012), 1262–1267.
[149] Maier, L. et al. 2018. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature*. (2018).
[150] Jandhyala, S.M. et al. 2015. Role of the normal gut microbiota. *World Journal of Gastroenterology*. 21, 29 (2015), 8836–8847.
[151] Ríos-Covián, D. et al. 2016. Intestinal short chain fatty acids and their link with diet and human health. *Frontiers in Microbiology*. 7, FEB (2016), 1–9.
[152] Morrison, D.J. and Preston, T. 2016. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes*. 7, 3 (2016), 189–200.
[153] Tan, J. et al. 2014. *The Role of Short-Chain Fatty Acids in Health and Disease*. Elsevier Inc.
[154] Bourassa, M.W. et al. 2016. Butyrate, neuroepigenetics and the gut microbiome: Can a high fiber diet improve brain health? *Neuroscience Letters*. 625, (2016), 56–63.
[155] Y. Han, X. 2000. Bacterial Identification Based on 16S Ribosomal RNA Gene Sequence Analysis. *Methodology*. (2000), 323–332.
[156] Horn, G. et al. 2012. Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harbor Symposia on Quantitative Biology*. 51, 0 (2012), 263–273.
[157] Greisen, K. et al. 1994. PCR Primers and Probes for the 16S rRNA Gene of Most Species of Pathogenic Bacteria , Including Bacteria Found in Cerebrospinal Fluid. *Journal of Clinical Microbiology*. 32, 2 (1994), 335–351.
[158] Han, X.Y. et al. 2002. Rapid and Accurate Identification of Mycobacteria by Sequencing Hypervariable Regions of the 16S Ribosomal RNA Gene. *American Journal of Clinical Pathology*. 118, (2002), 796–801.
[159] Mardanov, A. V. et al. 2017. *Metagenomics: A Paradigm Shift in Microbiology*. Elsevier Inc.
[160] McDonald, D. et al. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME Journal*. 6, 3 (2012), 610–618.
[161] Quast, C. et al. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*. 41, D1 (2013), 590–596.
[162] Brown, C.T. et al. 2013. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*. 42, D1 (2013), D633–D642.
[163] Morgan, X.C. and Huttenhower, C. 2014. Meta'omic analytic techniques for studying the intestinal microbiome. *Gastroenterology*. 146, 6 (2014), 1437-1448.e1.
[164] Fredricks, D.N. and Relman, D.A. 1996. Sequence-Based Identification of Microbial Pathogens : a Reconsideration of Koch's Postulates. *Clinical Microbiology Reviews*. 9, 1 (1996), 18–33.
[165] Goodrich, J.K. et al. 2014. Conducting a microbiome study. *Cell*. 158, 2 (2014), 250–262.
[166] Magoč, T. and Salzberg, S.L. 2011. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. 27, 21 (2011), 2957–2963.
[167] Schmieder, R. and Edwards, R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. 27, 6 (2011), 863–864.

[168] Lozupone, C.A. and Knight, R. 2008. Species divergence and the measurement of microbial diversity. *FEMS Microbiology Reviews*. 32, 4 (2008), 557–578.
[169] Chao, A. 1984. Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics*. 11, 4 (1984), 265–270.
[170] Chazdon, R.L. et al. 1998. Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of NE Costa Rica. *Forest Biodiversity Research, Monitoring and Modeling*. F. Dallmeier and J.A. Comiskey, eds. UNESCO Paris and The Parthenon Publishing Group.
[171] Faith, D.P. 1992. Conservation evaluation and phylogentic diversity. *Biological Conservation*. 61, (1992), 1–10.
[172] Shannon, C. and Weaver, W. 1964. *The Mathematical Theory of Communication*. The University of Illinois Press, Urbana, IL.
[173] Simpson, E.H. 1949. Measurement of Diversity. *Nature*. 163, (1949), 688.
[174] McMurdie, P.J. and Holmes, S. 2014. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Computational Biology*. 10, 4 (2014).
[175] Weiss, S. et al. 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*. 5, 1 (2017), 1–18.
[176] Gloor, G.B. et al. 2017. Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*.
[177] Quinn, T.P. et al. 2018. Benchmarking differential expression analysis tools for RNA-Seq: Normalization-based vs. log-ratio transformation-based methods. *BMC Bioinformatics*. 19, 1 (2018), 1–15.
[178] Perez-muñoz, M.E. et al. 2017. A critical assessment of the “sterile womb” and “in utero colonization” hypotheses : implications for research on the pioneer infant microbiome. *Microbiome*. 5, 48 (2017), 1–19.
[179] Funkhouser, L.J. and Bordenstein, S.R. 2013. Mom Knows Best : The Universality of Maternal Microbial Transmission. *PLoS Biology*. 11, 8 (2013), 1–9.
[180] Collado, M.C. et al. 2016. Human gut colonisation may be initiated in utero by distinct microbial communities in the placenta and amniotic fluid. *Nature Publishing Group*. March (2016), 1–13.
[181] Blaser, M.J. and Dominguez-Bello, M.G. 2016. The Human Microbiome before Birth. *Cell Host and Microbe*. 20, 5 (2016), 558–560.
[182] Aagaard, K. et al. 2014. The Placenta Harbors a Unique Microbiome. *Science Translational Medicine*. 65, (2014).
[183] Dominguez-bello, M.G. et al. 2019. Role of the microbiome in human development. *Gut*. 0, (2019), 1–7.
[184] Mitsou, E.K. et al. 2008. Fecal microflora of Greek healthy neonates. *Anaerobe*. 14, (2008), 94–101.
[185] Thursby, E. and Juge, N. 2017. Introduction to the human gut microbiota. *Biochemical Journal*. 474, 11 (2017), 1823–1836.
[186] Palmer, C. et al. 2007. Development of the human infant intestinal microbiota. *PLoS Biology*. 5, 7 (2007), 1556–1573.
[187] Mariat, D. et al. 2009. The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age. *BCM Microbiology*. 6, (2009), 1–6.
[188] Woodmansey, E.J. et al. 2004. Comparison of Compositions and Metabolic Activities of Fecal Microbiotas in Comparison of Compositions and Metabolic Activities of Fecal Microbiotas in Young Adults and in Antibiotic-Treated and Non-Antibiotic-Treated Elderly Subjects. *Applied and Environmental Microbiology*. 70, 10 (2004), 6113–6122.
[189] Biagi, E. et al. 2013. Ageing and gut microbes: Perspectives for health maintenance and longevity. *Pharmacological Research*. 69, 1 (2013), 11–20.
[190] Simon, A.K. et al. 2015. Evolution of the immune system in humans from infancy to old age. *Proceedings of the Royal Society B*. 282, 1821 (2015), 1–12.
[191] Estes, M.L. and McAllister, A.K. 2016. Maternal immune activation: Implications for neuropsychiatric disorders. *Science*. 353, 6301 (2016), 772–777.
[192] Cianci, R. et al. 2018. The Microbiota and Immune System Crosstalk in Health and Disease. *Mediators of Inflammation*. (2018).
[193] Pagliari, D. et al. 2015. The Interactions between Innate Immunity and Microbiota in Gastrointestinal Diseases. *Journal of Immunology Research*. 2015, (2015), 1–3.

[194] Sjögren, Y.M. et al. 2009. Altered early infant gut microbiota in children developing allergy up to 5 years of age. *Clinical and Experimental Allergy*. 39, 4 (2009), 518–526.
[195] Platt, A.M. and Mowat, A.M.I. 2008. Mucosal macrophages and the regulation of immune responses in the intestine. *Immunology Letters*. 119, 1–2 (2008), 22–31.
[196] Romagnani, S. 2006. Regulation of the T cell response. *Clinical & Experimental Allergy*. 36, (2006), 1357–1366.
[197] Rook, G.A.W. and Brunet, L.R. 2005. Microbes, immunoregulation, and the gut. *Gut*. 54, 3 (2005), 317–320.
[198] Yazdanbakhsh, M. et al. 2002. Immunology: Allergy, parasites, and the hygiene hypothesis. *Science*. 296, 5567 (2002), 490–494.
[199] Wills-Karp, M. et al. 2001. The germless theory of allergic disease: revisiting the hygiene hypothesis. *Nature Reviews Immunology*. 1, October (2001), 1–7.
[200] Hooper, L. V. et al. 2012. Interactions Between the Microbiota and the Immune System. *Science*. 336, (2012), 1268–1273.
[201] Kamada, N. et al. 2013. Role of the gut microbiota in immunity and inflammatory disease. *Nature Reviews Immunology*. 13, 5 (2013), 321–335.
[202] Momose, Y. et al. 2008. Competition for proline between indigenous Escherichia coli and E. coli O157:H7 in gnotobiotic mice associated with infant intestinal microbiota and its contribution to the colonization resistance against E. coli O157:H7. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*. 94, 2 (2008), 165–171.
[203] Leatham, M.P. et al. 2009. Precolonized Human Commensal Escherichia coli Strains Serve as a Barrier to E. coli O157:H7 Growth in the Streptomycin-Treated Mouse Intestine. *Infection and Immunity*. 77, 7 (2009), 2876–2886.
[204] Fung, T.C. et al. 2017. Interactions between the microbiota, immune and nervous systems in health and disease. *Nature Neuroscience*. 20, 2 (2017), 145–155.
[205] Borre, Y.E. et al. 2014. Microbiota and neurodevelopmental windows: implications for brain disorders. *Trends in molecular medicine*. 20, 9 (2014), 509–518.
[206] Braniste, V. et al. 2014. The gut microbiota influences blood-brain barrier permeability in mice. *Science Translational Medicine*. 6, 263 (2014).
[207] Lee, Y.K. et al. 2010. Proinflammatory T-cell responses to gut microbiota promote experimental autoimmune encephalomyelitis. *Proceedings of the National Academy of Sciences*. 108, Supplement_1 (2010), 4615–4622.
[208] Haghikia, A. et al. 2015. Dietary Fatty Acids Directly Impact Central Nervous System Autoimmunity via the Small Intestine. *Immunity*. 43, 4 (2015), 817–829.
[209] Rothhammer, V. et al. 2016. Type I interferons and microbial metabolites of tryptophan modulate astrocyte activity and CNS inflammation via the aryl hydrocarbon receptor. *Nature Medicine*. 22, 6 (2016), 586–597.
[210] Lyte, M. 2013. Microbial Endocrinology in the Microbiome-Gut-Brain Axis: How Bacterial Production and Utilization of Neurochemicals Influence Behavior. *PLoS Pathogens*. 9, 11 (2013), e1003726.
[211] Bravo, J.A. et al. 2011. Ingestion of Lactobacillus strain regulates emotional behavior and central GABA receptor expression in a mouse via the vagus nerve. *Proceedings of the National Academy of Sciences*. 108, 38 (2011), 16050–16055.
[212] Asano, Y. et al. 2012. Critical role of gut microbiota in the production of biologically active, free catecholamines in the gut lumen of mice. *American Journal of Physiology-Gastrointestinal and Liver Physiology*. 303, 11 (2012), G1288–G1295.
[213] Williams, B.B. et al. 2014. Discovery and characterization of gut microbiota decarboxylases that can produce the neurotransmitter tryptamine. *Cell Host and Microbe*. 16, 4 (2014), 495–503.
[214] Gershon, M.D. and Tack, J. 2007. The Serotonin Signaling System: From Basic Understanding To Drug Development for Functional GI Disorders. *Gastroenterology*. 132, 1 (2007), 397–414.
[215] Mazzoli, R. and Pessione, E. 2016. The neuro-endocrinological role of microbial glutamate and GABA signaling. *Frontiers in Microbiology*. 7, NOV (2016), 1–17.
[216] Delgado, T.C. 2013. Glutamate and GABA in appetite regulation. *Frontiers in Endocrinology*. 4, AUG (2013), 1–8.
[217] Sharon, G. et al. 2016. The Central Nervous System and the Gut Microbiome. *Cell*. 167, 4 (2016), 915–932.
[218] De Martel, C. et al. 2012. Global burden of cancers attributable to infections in 2008: A review and

synthetic analysis. *The Lancet Oncology*. 13, 6 (2012), 607–615.
[219] Hagland, H.R. and Søreide, K. 2015. Cellular metabolism in colorectal carcinogenesis: Influence of lifestyle, gut microbiome and metabolic pathways. *Cancer Letters*. 356, 2 (2015), 273–280.
[220] Hausen, H. Zur 2009. The search for infectious causes of human cancers: Where and why (Nobel Lecture). *Angewandte Chemie - International Edition*. 48, 32 (2009), 5798–5808.
[221] Warren, J.R. 2006. Helicobacter: The ease and difficulty of a new discovery (Nobel lecture). *ChemMedChem*. 1, 7 (2006), 672–685.
[222] Siegel, R.L. et al. 2019. Cancer statistics, 2019. *CA: a cancer journal for clinicians*. 69, 1 (2019), 7–34.
[223] Fearon, E.R. and Vogelstein, B. 1990. A Genetic Model for Colorectal Tumorigenesis. *Cell*. 61, (1990), 759–767.
[224] Kanehisa, M. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*. 34, 90001 (2006), D354–D357.
[225] Kanehisa, M. et al. 2017. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*. 45, D1 (2017), D353–D361.
[226] Tenenbaum, D. 2019. R Package "KEGGREST."
[227] Lang, D.T. and The CRAN Team 2019. R Package "XML."
[228] Gagolewski, M. et al. 2019. R Package "stringi."
[229] Xml, I. et al. 2019. R Package "rentrez."
[230] Davis, S. 2019. R Package "GEOquery."
[231] Kovalchik, S. 2017. R Package "RISmed": Download content from NCBI databases.
[232] Amberger, J.S. et al. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders. *Nucleic Acids Research*. 43, D1 (2015), D789–D798.
[233] Caporaso, J.G. et al. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME Journal*. 6, 8 (2012), 1621–1624.
[234] Love, M.I. et al. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. 15, 12 (2014), 550.
[235] Fernandes, A.D. et al. 2013. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. *PLoS ONE*. 8, 7 (2013).
[236] Fernandes et al. 2014. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*. 2, 15 (2014), 1–13.
[237] Clos-Garcia, M. et al. 2019. Gut microbiome and serum metabolome analyses identify molecular biomarkers and altered glutamate metabolism in fibromyalgia. *EBioMedicine*. 46, (2019), 499–511.
[238] Weiss, S. et al. 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*. (2017).
[239] Zych, K. et al. 2018. SIAMCAT: Statistical Inference of Associations between Microbial Communities And host phenoTypes. (2018).
[240] Segata, N. et al. 2011. Metagenomic biomarker discovery and explanation. *Genome Biology*. 12, 6 (2011).
[241] Rohart, F. et al. 2017. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*. 13, 11 (2017), e1005752.
[242] Singh, A. et al. 2018. DIABLO: from multi-omics assays to biomarker discovery, an integrative approach. *bioRxiv*. (Jan. 2018).
[243] High-sensitivity pattern discovery in large multi'omic datasets: *http://huttenhower.sph.harvard.edu/halla*. Accessed: 2019-07-07.
[244] Clos-Garcia, M. et al. 2018. Metabolic alterations in urine extracellular vesicles are associated to prostate cancer pathogenesis and progression. *Journal of Extracellular Vesicles*. 7, 1470442 (2018).
[245] Freedland, S.J. 2011. Screening, risk assessment, and the approach to therapy in patients with prostate cancer. *Cancer*. 117, 6 (2011), 1123–1135.
[246] Fernie, A.R. et al. 2004. Metabolite profiling: from diganostics to systems biology. *Nature Reviews Molecular Cell Biology*. 5, September (2004), 1–7.
[247] Royo, F. et al. 2017. Hepatocyte-secreted extracellular vesicles modify blood metabolome and endothelial function by an arginase-dependent mechanism. *Scientific Reports*. 7, January (2017), 1–15.

[248] Gonzalez, E. et al. 2012. Serum UPLC-MS/MS metabolic profiling in an experimental model for acute-liver injury reveals potential biomarkers for hepatotoxicity. *Metabolomics*. 8, 6 (2012), 997–1011.
[249] Alonso, C. et al. 2017. Metabolomic Identification of Subtypes of Nonalcoholic Steatohepatitis. *Gastroenterology*. 152, 6 (2017), 1449-1461.e7.
[250] Holmes, E. et al. 2015. The promise of metabolic phenotyping in gastroenterology and hepatology. *Nature Reviews Gastroenterology &Amp; Hepatology*. 12, (Jul. 2015), 458.
[251] Griffin, J.L. and Shockcor, J.P. 2004. Metabolic profiles of cancer cells. *Nature Reviews Cancer*. 4, 7 (2004), 551–561.
[252] Pentyala, S. et al. 2016. Prostate cancer markers: An update (Review). *Biomedical Reports*. (2016), 263–268.
[253] Giskeødegård, G.F. et al. 2015. Metabolic markers in blood can separate prostate cancer from benign prostatic hyperplasia. *British Journal of Cancer*. 113, (2015), 1712–1719.
[254] Di Meo, A. et al. 2017. Liquid biopsy: A step forward towards precision medicine in urologic malignancies. *Molecular Cancer*. 16, 1 (2017), 1–14.
[255] Lima, A.R. et al. 2016. Biomarker discovery in human prostate cancer: An update in metabolomics studies. *Translational Oncology*. 9, 4 (2016), 357–370.
[256] Skotland, T. et al. 2017. Molecular lipid species in urinary exosomes as potential prostate cancer biomarkers. *European Journal of Cancer*. 70, (2017), 122–132.
[257] Iraci, N. et al. 2017. Extracellular vesicles are independent metabolic units with asparaginase activity. *Nature Chemical Biology*. 13, 9 (2017), 951–955.
[258] Royo, F. et al. 2017. Metabolically active extracellular vesicles released from hepatocytes under drug-induced liver-damaging conditions modify serum metabolome and might affect different pathophysiological processes. *European Journal of Pharmaceutical Sciences*. 98, (2017), 51–57.
[259] Royo, F. et al. 2016. Transcriptomic profiling of urine extracellular vesicles reveals alterations of CDH3 in prostate cancer. *Oncotarget*. 7, 6 (2016), 6835–6846.
[260] Barr, J. et al. 2012. Obesity-dependent metabolic signatures associated with nonalcoholic fatty liver disease progression. *Journal of Proteome Research*. 11, 4 (2012), 2521–2532.
[261] Martínez-Uña, M. et al. 2013. Excess S-adenosylmethionine reroutes phosphatidylethanolamine towards phosphatidylcholine and triglyceride synthesis. *Hepatology*. 58, 4 (2013), 1296–1305.
[262] López-Ratón, M. et al. 2014. OptimalCutpoints : An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. *Journal of Statistical Software*. 61, 8 (2014), 1–36.
[263] Zabala-Letona, A. et al. 2017. MTORC1-dependent AMD1 regulation sustains polyamine metabolism in prostate cancer. *Nature*. 547, 7661 (2017), 109–113.
[264] Torrano, V. et al. 2016. The metabolic co-regulator PGC1α suppresses prostate cancer metastasis. *Nature Cell Biology*. 18, 6 (2016), 645–656.
[265] Royo, F. et al. 2016. Different EV enrichment methods suitable for clinical settings yield different subpopulations of urinary extracellular vesicles from human samples. *Journal of Extracellular Vesicles*. 5, 1 (2016).
[266] Dell'Atti, L. 2016. Prognostic significance of perineural invasion in patients who underwent radical prostatectomy for localized prostate cancer. *Journal of B.U.ON.* 21, 5 (2016), 1219–1223.
[267] Rodriguez, C. et al. 2007. Body Mass Index, Weight Change, and Risk of Prostate Cancer in the Cancer Prevention Study II Nutrition Cohort. *Cancer Epidemiology Biomarkers & Prevention*. 16, 1 (Jan. 2007), 63 LP – 69.
[268] J., L.D. et al. 2017. Recent Changes in Prostate Cancer Screening Practices and Epidemiology. *Journal of Urology*. 198, 6 (Dec. 2017), 1230–1240.
[269] Gao, J. et al. 2010. Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. *Bioinformatics (Oxford, England)*. 26, 7 (Apr. 2010), 971–973.
[270] Torrano, V. et al. 2016. Vesicle-MaNiA: extracellular vesicles in liquid biopsy and cancer. *Current opinion in pharmacology*. 29, (Aug. 2016), 47–53.
[271] Peinado, H. et al. 2012. Melanoma exosomes educate bone marrow progenitor cells toward a pro-metastatic phenotype through MET. *Nature medicine*. 18, 6 (Jun. 2012), 883–891.
[272] Fujita, K. et al. 2009. Specific detection of prostate cancer cells in urine by multiplex immunofluorescence cytology. *Human pathology*. 40, 7 (Jul. 2009), 924–933.
[273] Siebren, D. et al. 2014. Prostate Cancer Biomarker Profiles in Urinary Sediments and Exosomes. *Journal of Urology*. 191, 4 (Apr. 2014), 1132–1138.

[274] Di Vizio, D. et al. 2012. Large oncosomes in human prostate cancer tissues and in the circulation of mice with metastatic disease. *The American journal of pathology*. 181, 5 (Nov. 2012), 1573–1584.
[275] Minciacchi, V.R. et al. 2015. Large oncosomes contain distinct protein cargo and represent a separate functional class of tumor-derived extracellular vesicles. *Oncotarget*. 6, 13 (2015), 11327–41.
[276] Puhka, M. et al. 2017. Metabolomic Profiling of Extracellular Vesicles and Alternative Normalization Methods Reveal Enriched Metabolites and Strategies to Study Prostate Cancer-Related Changes. *Theranostics*. 7, 16 (Aug. 2017), 3824–3841.
[277] Li, J. et al. 2016. Integration of lipidomics and transcriptomics unravels aberrant lipid metabolism and defines cholesteryl oleate as potential biomarker of prostate cancer. *Scientific reports*. 6, (Feb. 2016), 20984.
[278] Faas, F.H. et al. 2003. Decreased prostatic arachidonic acid in human prostatic carcinoma. *BJU International*. 92, 6 (Oct. 2003), 551–554.
[279] Yang, P. et al. 2012. Arachidonic acid metabolism in human prostate cancer. *International journal of oncology*. 41, 4 (Aug. 2012), 1495–1503.
[280] Chaudry, A.A. et al. 1994. Arachidonic acid metabolism in benign and malignant prostatic tissue in vitro: Effects of fatty acids and cyclooxygenase inhibitors. *International Journal of Cancer*. 57, 2
[281] Nithipatikom, K. et al. 2006. Elevated 12- and 20-hydroxyeicosatetraenoic acid in urine of patients with prostatic diseases. *Cancer letters*. 233, 2 (2006), 219—225.
[282] Nithipatikom, K. and Campbell, W.B. 2008. Roles of Eicosanoids in Prostate Cancer. *Future lipidology*. 3, 4 (Aug. 2008), 453–467.
[283] Rodrigues, D.N. et al. 2017. The molecular underpinnings of prostate cancer: impacts on management and pathology practice. *The Journal of Pathology*. 241, 2 (Jan. 2017), 173–182.
[284] de Bono, J.S. et al. 2011. Abiraterone and Increased Survival in Metastatic Prostate Cancer. *New England Journal of Medicine*. 364, 21 (May 2011), 1995–2005.
[285] Capper, C.P. et al. 2016. The Metabolism, Analysis, and Targeting of Steroid Hormones in Breast and Prostate Cancer. *Hormones & cancer*. 7, 3 (Jun. 2016), 149–164.
[286] Saddoughi, S.A. and Ogretmen, B. 2013. Chapter Two - Diverse Functions of Ceramide in Cancer Cell Death and Proliferation. *The Role of Sphingolipids in Cancer Development and Therapy*. J.S.B.T.-A. in C.R. Norris, ed. Academic Press. 37–58.
[287] Sreekumar, A. et al. 2009. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*. 457, 7231 (2009), 910–914.
[288] Zadra, G. et al. 2013. The fat side of prostate cancer. *Biochimica et biophysica acta*. 1831, 10 (Oct. 2013), 1518–1532.
[289] Deep, G. and Schlaepfer, I.R. 2016. Aberrant Lipid Metabolism Promotes Prostate Cancer: Role in Cell Survival under Hypoxia and Extracellular Vesicles Biogenesis. *International journal of molecular sciences*. 17, 7 (Jul. 2016), 1061.
[290] Rysman, E. et al. 2010. De novo Lipogenesis Protects Cancer Cells from Free Radicals and Chemotherapeutics by Promoting Membrane Lipid Saturation. *Cancer Research*. 70, 20 (Oct. 2010), 8117 LP – 8126.
[291] Carracedo, A. et al. 2013. Cancer metabolism: fatty acid oxidation in the limelight. *Nature reviews. Cancer*. 13, 4 (Apr. 2013), 227–232.
[292] Al-Bakheit, A. et al. 2016. Accumulation of Palmitoylcarnitine and Its Effect on Pro-Inflammatory Pathways and Calcium Influx in Prostate Cancer. *The Prostate*. 76, 14 (Oct. 2016), 1326–1337.
[293] Klindworth, A. et al. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*. 41, 1 (2013), 1–11.
[294] Caporaso, J.G. et al. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME Journal*. 6, 8 (2012), 1621–1624.
[295] Ward, D. V. et al. 2012. Evaluation of 16s rDNA-based community profiling for human microbiome research. *PLoS ONE*. 7, 6 (2012).
[296] Bukin, Y.S. et al. 2019. The effect of 16s rRNA region choice on bacterial community metabarcoding results. *Scientific Data*. 6, (2019), 1–14.
[297] Zhang, J. et al. 2018. Evaluation of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. *Science of the Total Environment*. 618, (2018), 1254–1267.
[298] Claesson, M.J. et al. 2010. Comparison of two next-generation sequencing technologies for

resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research*^{o}. 38, 22 (2010).
[299] Chen, Z. et al. 2019. Impact of Preservation Method and 16S rRNA Hypervariable Region on Gut Microbiota Profiling. *mSystems*. 4, e00271-18 (2019).
[300] Wolfe, F. et al. 2010. The American College of Rheumatology preliminary diagnostic criteria for fibromyalgia and measurement of symptom severity. *Arthritis Care and Research*. 62, 5 (2010), 600–610.
[301] Häuser, W. et al. 2015. Fibromyalgia. *Nature Reviews Disease Primers*. August (2015), 15022.
[302] Ablin, J.N. et al. 2006. Mechanisms of Disease: Genetics of fibromyalgia. *Nature Clinical Practice Rheumatology*. 2, 12 (2006), 671–678.
[303] Offenbaecher, M. et al. 1999. POSSIBLE ASSOCIATION OF FIBROMYALGIA WITH A POLYMORPHISM IN THE SEROTONIN TRANSPORTER GENE REGULATORY REGION. *Arthritis & Rheumatism*. 42, 11 (1999), 2482–2488.
[304] Cohen, H. et al. 2002. Confirmation of an association between fibromyalgia and serotonin transporter promoter region (5-HTTLPR) polymorphism, and relationship to anxiety-related personality traits. *Arthritis and Rheumatism*. 46, 3 (2002), 845–847.
[305] Zubieta, J.K. et al. 2003. COMT val158 genotype affects μ-opioid neurotransmitter responses to a pain stressor. *Science*. 299, 5610 (2003), 1240–1243.
[306] Gürsoy, S. et al. 2003. Significance of catechol-O-methyltransferase gene polymorphism in fibromyalgia syndrome. *Rheumatology International*. 23, (2003), 104–107.
[307] Buskila, D. et al. 2008. Etiology of fibromyalgia: The possible role of infection and vaccination. *Autoimmunity Reviews*. 8, 1 (2008), 41–43.
[308] Buskila, D. et al. 1997. Fibromyalgia in Hepatitis C Virus Infection. *Archives of Internal Medicine*. 157, (1997), 2497.
[309] Rivera, J. et al. 1997. Fibromyalgia-associated hepatitis C virus infection. *British Journal of Rheumatology*. 36, (1997), 981–985.
[310] Häuser, W. et al. 2011. Emotional, physical, and sexual abuse in fibromyalgia syndrome: A systematic review with meta-analysis. *Arthritis Care & Research*. 63, 6 (2011), 808–820.
[311] Ablin, J. et al. 2013. Frequency of axial spondyloarthropathy among patients suffering from fibromyalgia. A magnetic resonance imaging study applying the assessment of spondylo-arthritis international society classification criteria. *Arthritis Rheum. Abstr.* 65, (2013), 128.
[312] Sarchielli, P. et al. 2007. Sensitization, glutamate, and the link between migraine and fibromyalgia. *Current Pain and Headache Reports*. 11, 5 (2007), 343–351.
[313] Peres, M.F.P. et al. 2004. Cerebrospinal fluid glutamate levels in chronic migraine. *Cephalalgia*. 24, 9 (2004), 735–739.
[314] Gallai, V. et al. 2003. Glutamate and nitric oxide pathway in chronic daily headache: Evidence from cerebrospinal fluid. *Cephalalgia*. 23, 3 (2003), 166–174.
[315] Foerster, B.R. et al. 2012. Reduced insular γ-aminobutyric acid in fibromyalgia. *Arthritis and Rheumatism*. 64, 2 (2012), 579–583.
[316] Milligan, E.D. and Watkins, L.R. 2009. Pathological and protective roles of glia in chronic pain. *Nature Reviews Neuroscience*. 10, 1 (2009), 23–36.
[317] Uçeyler, N. et al. 2011. Systematic review with meta-analysis: cytokines in fibromyalgia syndrome. *Bmc Musculoskelet Di*. 12, 1 (2011), 245.
[318] Kennedy, P.J. et al. 2014. Irritable bowel syndrome: A microbiome-gut-brain axis disorder? *World Journal of Gastroenterology*. 20, 39 (2014), 14105–14125.
[319] Karlsson, F. et al. 2013. Assessing the human gut microbiota in metabolic diseases. *Diabetes*. 62, 10 (2013), 3341–3349.
[320] Sampson, T.R. and Mazmanian, S.K. 2015. Control of brain development, function, and behavior by the microbiome. *Cell Host and Microbe*. 17, 5 (2015), 565–576.
[321] Forsythe, P. et al. 2010. Mood and gut feelings. *Brain, Behavior, and Immunity*. 24, 1 (2010), 9–16.
[322] Cryan, J.F. and O'Mahony, S.M. 2011. The microbiome-gut-brain axis: From bowel to behavior. *Neurogastroenterology and Motility*. 23, 3 (2011), 187–192.
[323] Hooks, K.B. and O'Malley, M.A. 2017. Dysbiosis and Its Discontents. *mBio*. 8, 5 (2017), 1–11.
[324] Collins, S.M. et al. 2012. The interplay between the intestinal microbiota and the brain. *Nature Reviews Microbiology*. 10, 11 (2012), 735–742.
[325] Sharon, G. et al. 2014. Specialized metabolites from the microbiome in health and disease. *Cell*

Metabolism. 20, 5 (2014), 719–730.
[326] Hemarajata, P. and Versalovic, J. 2013. Effects of probiotics on gut microbiota: Mechanisms of intestinal immunomodulation and neuromodulation. *Therapeutic Advances in Gastroenterology*. 6, 1 (2013), 39–51.
[327] Goodrich, J.K. et al. 2016. Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host and Microbe*. 19, 5 (2016), 731–743.
[328] Callahan, B.J. et al. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*. 13, 7 (2016), 581–583.
[329] Katoh, K. and Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*. 30, 4 (2013), 772–780.
[330] Price, M.N. et al. 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE*. 5, 3 (2010).
[331] McMurdie, P.J. and Holmes, S. 2013. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE*. 8, 4 (2013).
[332] Lahti, L. and Shetty, S. 2018. microbiome R package. (2018).
[333] Love, M.I. et al. 2014. *DESeq2 package__Differential analysis of count data*.
[334] Li, K. et al. 2013. Analyses of the Stability and Core Taxonomic Memberships of the Human Microbiome. *PLoS ONE*. 8, 5 (2013).
[335] Karnovsky, A. et al. 2012. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics*. 28, 3 (2012), 373–380.
[336] Tackett, M.R. and Diwan, I. 2017. Using FirePlex™ Particle Technology for Multiplex MicroRNA Profiling Without RNA Purification. *Methods in molecular biology (Clifton, N.J.)*. 1654, (2017), 209—219.
[337] Vandesompele, J. et al. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology*. 3, 7 (Jun. 2002), research0034.1.
[338] Malatji, B.G. et al. 2017. A diagnostic biomarker profile for fibromyalgia syndrome based on an NMR metabolomics study of selected patients and controls. *BMC Neurology*. 17, 1 (2017), 1–15.
[339] Hadrévi, J. et al. 2015. Systemic differences in serum metabolome: A cross sectional comparison of women with localised and widespread pain and controls. *Scientific Reports*. 5, March (2015), 1–13.
[340] Caboni, P. et al. 2014. Metabolomics analysis and modeling suggest a lysophosphocholines-PAF receptor interaction in fibromyalgia. *PLoS ONE*. 9, 9 (2014), 1–8.
[341] Freidin, M.B. et al. 2018. Metabolomic markers of fatigue: Association between circulating metabolome and fatigue in women with chronic widespread pain. *Biochimica et Biophysica Acta - Molecular Basis of Disease*. 1864, 2 (2018), 601–606.
[342] Noronha, A. et al. 2019. The Virtual Metabolic Human database: Integrating human and gut microbiome metabolism with nutrition and disease. *Nucleic Acids Research*. 47, D1 (2019), D614–D624.
[343] Dweep, H. and Gretz, N. 2015. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nature Methods*. 12, 8 (2015), 697–697.
[344] Kamburov, A. et al. 2013. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Research*. 41, 793–800 (2013), \.
[345] Giloteaux, L. et al. 2016. Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome. *Microbiome*. 4, 1 (2016), 30.
[346] Meehan, C.J. and Beiko, R.G. 2014. A phylogenomic view of ecological specialization in the lachnospiraceae, a family of digestive tract-associated bacteria. *Genome Biology and Evolution*. 6, 3 (2014), 703–713.
[347] Barrett, E. et al. 2012. γ-Aminobutyric acid production by culturable bacteria from the human intestine. *Journal of Applied Microbiology*. 113, 2 (2012), 411–417.
[348] Yunes, R.A. et al. 2016. GABA production and structure of gadB/gadC genes in Lactobacillus and Bifidobacterium strains from human microbiota. *Anaerobe*. 42, (2016), 197–204.
[349] Osikowicz, M. et al. 2013. The glutamatergic system as a target for neuropathic pain relief. *Experimental Physiology*. 98, 2 (2013), 372–384.
[350] Popoli, M. et al. 2012. The stressed synapse: The impact of stress and glucocorticoids on glutamate

transmission. *Nature Reviews Neuroscience*. 13, 1 (2012), 22–37.
[351] Liu, M.T. et al. 1997. Glutamatergic enteric neurons. *The Journal of neuroscience : the official journal of the Society for Neuroscience*. 17, 12 (1997), 4764–4784.
[352] Kirchgessner, A.L. et al. 1997. Excitotoxicity in the enteric nervous system. *Journal of Neuroscience*. 17, 22 (1997), 8804–8816.
[353] Chen, W.P. and Kirchgessner, A.L. 2002. Activation of group II mGlu receptors inhibits voltage-gated Ca2+ currents in myenteric neurons. *Am J Physiol Gastrointest Liver Physiol*. 283, 6 (2002), G1282-9.
[354] Sartor, R.B. 2008. Microbial Influences in Inflammatory Bowel Diseases. *Gastroenterology*. 134, 2 (2008), 577–594.
[355] Frank, D.N. et al. 2007. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proceedings of the National Academy of Sciences of the United States of America. *PNAS*. 104, 34 (2007), 13780–13785.
[356] Strober, W. et al. 2007. The fundamental basis of inflammatory bowel disease. *Journal of Clinical Investigation*. 117, 3 (2007), 514–21.
[357] Pirzer, U. et al. 1991. Reactivity of infiltrating T lymphocytes with microbial antigens in Crohn's disease. *The Lancet*. 338, 8777 (1991), 1238–1239.
[358] Frank, D.N. et al. 2011. Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. *Inflammatory Bowel Diseases*. 17, 1 (2011), 179–184.
[359] Morgan, X.C. et al. 2012. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology*. 13, 9 (2012), R79.
[360] Wong, D. et al. 2004. Cytokines, nitric oxide, and cGMP modulate the permeability of an in vitro model of the human blood-brain barrier. *Experimental Neurology*. 190, 2 (2004), 446–455.
[361] Boveri, M. et al. 2006. Highly purified lipoteichoic acid from gram-positive bacteria induces in vitro blood-brain barrier disruption through glia activation: Role of pro-inflammatory cytokines and nitric oxide. *Neuroscience*. 137, 4 (2006), 1193–1209.
[362] Wikoff, W.R. et al. 2009. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proceedings of the National Academy of Sciences*. 106, 10 (2009), 3698–3703.
[363] Yano, J.M. et al. 2015. Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis. *Cell*. 161, 2 (2015), 264–276.
[364] O'Mahony, S.M. et al. 2015. Serotonin, tryptophan metabolism and the brain-gut-microbiome axis. *Behavioural Brain Research*. 277, (2015), 32–48.
[365] Matsumoto, M. et al. 2012. Impact of intestinal microbiota on intestinal luminal metabolome. *Scientific Reports*. 2, (2012), 1–10.
[366] Velagapudi, V.R. et al. 2009. The gut microbiota modulates host energy and lipid metabolism in mice. *Journal of Lipid Research*. 51, 5 (2009), 1101–1112.
[367] Queiroz, L.P. aul. 2013. Worldwide epidemiology of fibromyalgia. *Current pain and headache reports*. 17, 8 (2013), 356.
[368] Tabatadze, N. et al. 2015. Sex Differences in Molecular Signaling at Inhibitory Synapses in the Hippocampus. *Journal of Neuroscience*. 35, 32 (2015), 11252–11265.
[369] Huang, G.Z. and Woolley, C.S. 2012. Estradiol Acutely Suppresses Inhibition in the Hippocampus through a Sex-Specific Endocannabinoid and mGluR-Dependent Mechanism. *Neuron*. 74, 5 (2012), 801–808.
[370] Cury, Y. et al. 2011. Pain and analgesia: The dual effect of nitric oxide in the nociceptive system. *Nitric Oxide - Biology and Chemistry*. 25, 3 (2011), 243–254.
[371] Pernambuco, A.P. et al. 2016. Involvement of Oxidative Stress and Nitric Oxide in Fibromyalgia Pathophysiology : A Relationship to be Elucidated. *Fibromyalgia : Open Access*. 1, 1 (2016), 1–7.
[372] McIver, K.L. et al. 2006. NO-mediated alterations in skeletal muscle nutritive blood flow and lactate metabolism in fibromyalgia. *Pain*. 120, 1–2 (2006), 161–169.
[373] Cussotto, S. et al. 2019. Psychotropics and the Microbiome: a Chamber of Secrets…. *Psychopharmacology*. (2019).
[374] Jackson, M.A. et al. 2016. Proton pump inhibitors alter the composition of the gut microbiota. *Gut*. 65, 5 (2016), 749–756.
[375] Imhann, F. et al. 2016. Proton pump inhibitors affect the gut microbiome. *Gut*. 65, 5 (2016), 740–748.

[376] Cussotto, S. et al. 2018. Differential effects of psychotropic drugs on microbiome composition and gastrointestinal function. *Psychopharmacology*. (2018).
[377] Rosenberg, P.H. and Renkonen, O. V. 1985. Antimicrobial Activity of Bupivacaine and Morphine. *Anesthesiology*. 62, (1985), 178–179.
[378] Acharya, C. et al. 2017. Chronic opioid use is associated with altered gut microbiota and predicts readmissions in patients with cirrhosis. *Alimentary Pharmacology and Therapeutics*. 45, 2 (2017), 319–331.
[379] Roman, P. et al. 2018. Are probiotic treatments useful on fibromyalgia syndrome or chronic fatigue syndrome patients? A systematic review. *Beneficial Microbes*. 9, 4 (2018), 603–611.
[380] Roman, P. et al. 2017. Probiotics for fibromyalgia: study design for a pilot double-blind, randomized controlled tria. *Nutrición Hospitalaria*. 34, 5 (2017), 1246–1251.
[381] Roman, P. et al. 2018. A Pilot Randomized Controlled Trial to Explore Cognitive and Emotional Effects of Probiotics in Fibromyalgia. *Scientific Reports*. 8, 1 (2018), 1–9.
[382] Ferlay, J. et al. 2015. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*. 136, 5 (2015), E359–E386.
[383] Vogelstein, B. et al. 2013. Cancer Genome Landscapes. *Science*. 339, 6127 (2013), 1546–1558.
[384] Zauber, A.G. et al. 2012. Colonoscopic Polypectomy and Long-Term Prevention of Colorectal-Cancer Deaths. *The New England Journal of Medicine*. 366, (2012).
[385] Quintero, E. et al. 2015. Colonoscopy versus Fecal Immunochemical Testing in Colorectal-Cancer Screening. *New England Journal of Medicine*. 366, 8 (2015), 697–706.
[386] Lindholm, E. et al. 2008. Survival benefit in a randomized clinical trial of faecal occult blood screening for colorectal cancer. *British Journal of Surgery*. 95, 8 (2008), 1029–1036.
[387] Faivre, J. et al. 2004. Reduction in colorectal cancer mortality by fecal occult blood screening in a French controlled study. *Gastroenterology*. 126, 7 (2004), 1674–1680.
[388] Atkin, W.S. et al. 2010. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. *The Lancet*. 375, 9726 (2010), 1624–1633.
[389] Segnan, N. et al. 2005. Randomized trial of different screening strategies for colorectal cancer: Patient response and detection rates. *Journal of the National Cancer Institute*. 97, 5 (2005), 347–357.
[390] Imperiale, T.F. et al. 2014. Multitarget Stool DNA Testing for Colorectal-Cancer Screening. *New England Journal of Medicine*. 370, 14 (2014), 1287–1297.
[391] Levin, B. et al. 2008. Screening and Surveillance for the Early Detection of Colorectal Cancer and Adenomatous Polyps, 2008: A Joint Guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *CA: A Cancer Journal for Clinicians*. 58, 3 (2008), 130–160.
[392] Regula, J. et al. 2006. Colonoscopy Screening for Detection of Advanced Neoplasia. *New England Journal of Medicine*. 355, 18 (2006), 1863–1872.
[393] Bujanda, L. et al. 2007. Low adherence to colonoscopy in the screening of first-degree relatives of patients with colorectal cancer. *Gut*. 56, 12 (2007), 1714–1718.
[394] Puente Gutiérrez, J.J. et al. 2011. Effectiveness of a colonoscopic screening programme in first-degree relatives of patients with colorectal cancer. *Colorectal Disease*. 13, 6 (2011), 145–153.
[395] Mansouri, D. et al. 2015. Temporal trends in mode, site and stage of presentation with the introduction of colorectal cancer screening: A decade of experience from the West of Scotland. *British Journal of Cancer*. 113, 3 (2015), 556–561.
[396] Cubiella, J. et al. 2017. The fecal hemoglobin concentration, age and sex test score: Development and external validation of a simple prediction tool for colorectal cancer detection in symptomatic patients. *International Journal of Cancer*. 140, 10 (2017), 2201–2211.
[397] Cubiella, J. et al. 2016. Development and external validation of a faecal immunochemical test-based prediction model for colorectal cancer detection in symptomatic patients. *BMC medicine*. 14, 1 (2016), 128.
[398] Westwood, M. et al. 2017. Faecal immunochemical tests (FIT) can help to rule out colorectal cancer in patients presenting in primary care with lower abdominal symptoms: A systematic review conducted to inform new NICE DG30 diagnostic guidance. *BMC Medicine*. 15, 1 (2017), 1–17.
[399] Chen, C. et al. 2007. LC-MS-based metabolomics in drug metabolism. *Drug Metabolism Reviews*. 39, 2–3 (2007), 581–597.
[400] Clarke, C.J. and Haselden, J.N. 2008. Metabolic Profiling as a Tool for Understanding Mechanisms

of Toxicity. *Toxicologic Pathology*. 36, 1 (2008), 140–147.
[401] Nicholson, J.K. and Wilson, I.D. 2003. Understanding "global" systems biology: Metabonomics and the continuum of metabolism. *Nature Reviews Drug Discovery*. 2, 8 (2003), 668–676.
[402] Nordström, A. et al. 2006. Non-linear Data Alignment for UPLC-MS and HPLC-MS based Metabolomics: Application to Endogenous and Exogenous Metabolites in Human Serum. *Anal Chem*. 15, 78 (2006), 3289–3295.
[403] Nováková, L. et al. 2006. Advantages of ultra performance liquid chromatography over high-performance liquid chromatography: Comparison of different analytical approaches during analysis of diclofenac gel. *Journal of Separation Science*. 29, 16 (2006), 2433–2443.
[404] Zhang, F. et al. 2017. Metabolomics for biomarker discovery in the diagnosis, prognosis, survival and recurrence of colorectal cancer: a systematic review. *Oncotarget*. 8, 21 (2017), 35460–35472.
[405] Cross, A.J. et al. 2014. A prospective study of serum metabolites and colorectal cancer risk. *Cancer*. 120, 19 (2014), 3049–3057.
[406] Ikeda, A. et al. 2012. Serum metabolomics as a novel diagnostic approach for gastrointestinal cancer. *Biomedical Chromatography*. 26, 5 (2012), 548–558.
[407] Leichtle, A.B. et al. 2012. Serum amino acid profiles and their alterations in colorectal cancer. *Metabolomics*. 8, 4 (2012), 643–653.
[408] Li, F. et al. 2013. Lipid profiling for early diagnosis and progression of colorectal cancer using direct-infusion electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Rapid Communications in Mass Spectrometry*. 27, 1 (2013), 24–34.
[409] Nishiumi, S. et al. 2012. A novel serum metabolomics-based diagnostic approach for colorectal cancer. *PLoS ONE*. 7, 7 (2012), 1–10.
[410] Ma, Y. et al. 2012. An integrated proteomics and metabolomics approach for defining oncofetal biomarkers in the colorectal cancer. *Annals of Surgery*. 255, 4 (2012), 720–730.
[411] Ritchie, S. a et al. 2010. Reduced levels of hydroxylated, polyunsaturated ultra long-chain fatty acids in the serum of colorectal cancer patients: implications for early screening and detection. *BMC medicine*. 8, (2010), 13.
[412] Tan, B. et al. 2013. Metabonomics Identi fi es Serum Metabolite Markers of Colorectal Cancer. (2013).
[413] Zhu, J. et al. 2014. Colorectal cancer detection using targeted serum metabolic profiling. *Journal of Proteome Research*. 13, 9 (2014), 4120–4130.
[414] Manna, S.K. et al. 2014. Biomarkers of coordinate metabolic reprogramming in colorectal tumors in mice and humans. *Gastroenterology*. 146, 5 (2014), 1313–1324.
[415] Mirnezami, R. et al. 2014. Rapid diagnosis and staging of colorectal cancer via high-resolution magic angle spinning nuclear magnetic resonance (HR-MAS NMR) spectroscopy of intact tissue biopsies. *Annals of Surgery*. 259, 6 (2014), 1138–1149.
[416] Wang, H. et al. 2013. ^{1}H NMR-based metabolic profiling of human rectal cancer tissue. *Molecular Cancer*. 12, 1 (2013), 121.
[417] Silva, C.L. et al. 2011. Investigation of urinary volatile organic metabolites as potential cancer biomarkers by solid-phase microextraction in combination with gas chromatography-mass spectrometry. *British Journal of Cancer*. 105, 12 (2011), 1894–1904.
[418] Lin, Y. et al. 2016. NMR-based fecal metabolomics fingerprinting as predictors of earlier diagnosis in patients with colorectal cancer. *Oncotarget*. 7, 20 (2016), 29454–29464.
[419] Irrazábal, T. et al. 2014. The multifaceted role of the intestinal microbiota in colon cancer. *Molecular Cell*. 54, 2 (2014), 309–320.
[420] Gao, Z. et al. 2015. Microbiota disbiosis is associated with colorectal cancer. *Frontiers in Microbiology*. 6, FEB (2015), 1–9.
[421] Yan, G. et al. 2016. Lipidome in colorectal cancer. *Oncotarget*. 7, 22 (2016), 33429–33439.
[422] Martínez-Arranz, I. et al. 2015. Enhancing metabolomics research through data mining. *Journal of Proteomics*. 127, (2015), 275–288.
[423] Cock, P.J.A. et al. 2009. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 25, 11 (2009), 1422–1423.
[424] Cokelaer, T. et al. 2013. BioServices: A common Python package to access biological Web Services programmatically. *Bioinformatics*. 29, 24 (2013), 3241–3242.
[425] Szklarczyk, D. et al. 2017. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic acids research*. 45, D1 (2017), D362–D368.
[426] Valcz, G. et al. 2014. Myofibroblast-derived SFRP1 as potential inhibitor of colorectal carcinoma

field effect. *PLoS ONE*. 9, 11 (2014), 18–20.
[427] Luo, W. and Brouwer, C. 2013. Pathview: An R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*. 29, 14 (2013), 1830–1831.
[428] Phua, L.C. et al. 2013. Global gas chromatography/time-of-flight mass spectrometry (GC/TOFMS)-based metabonomic profiling of lyophilized human feces. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*. 937, (2013), 103–113.
[429] Saric, J. et al. 2008. Species variation in the fecal metabolome gives insight into differential gastrointestinal function. *Journal of Proteome Research*. 7, 1 (2008), 352–360.
[430] Zheng, X. et al. 2011. The Footprints of Gut Microbial-Mammalian Co-Metabolism. *Journal of Proteome Research*. 10, December (2011), 5512–5522.
[431] Jump, R.L.P. et al. 2014. Metabolomics analysis identifies intestinal microbiota-derived biomarkers of colonization resistance in clindamycin-treated mice. *PLoS ONE*. 9, 7 (2014).
[432] Martin, F.J. et al. 2010. Dietary Modulation of Gut Functional Ecology Studied by Fecal Metabonomics Francois-Pierre. *Journal of Proteome Research*. (2010), 5284–5295.
[433] Chow, J. et al. 2014. Fecal metabolomics of healthy breast-fed versus formula-fed infants before and during in vitro batch culture fermentation. *Journal of Proteome Research*. 13, 5 (2014), 2534–2542.
[434] Stella, C. et al. 2006. Susceptibility of human metabolic phenotypes to dietary modulation. *Journal of Proteome Research*. 5, 10 (2006), 2780–2788.
[435] Xu, W. et al. 2017. Development of High-Performance Chemical Isotope Labeling LC-MS for Profiling the Human Fecal Metabolome. *Analytical Chemistry*. 89, 12 (2017), 6758–6765.
[436] Zhao, Y. et al. 2013. Gut Microbiota Composition Modifies Fecal Metabolic Profiles in Mice. *Journal of Proteome Research*. (2013).
[437] Gao, X. et al. 2009. Metabolite analysis of human fecal water by gas chromatography/mass spectrometry with ethyl chloroformate derivatization. *Analytical Biochemistry*. 393, 2 (2009), 163–175.
[438] Gao, X. et al. 2010. Development of a quantitative metabolomic approach to study clinical human fecal water metabolome based on trimethylsilylation derivatization and GC/MS analysis. *Analytical Chemistry*. 82, 15 (2010), 6447–6456.
[439] Poroyko, V. et al. 2011. Diet creates metabolic niches in the “inmature gut” that shape microbial communities. *Nutricion Hospitalaria*. 26, 6 (2011), 1283–1295.
[440] Ponnusamy, K. et al. 2011. Microbial community and metabolomic comparison of irritable bowel syndrome faeces. *Journal of Medical Microbiology*. 60, 6 (2011), 817–827.
[441] Sciorra, V.A. and Morris, A.J. 2002. Roles for lipid phosphate phosphatases in regulation of cellular signaling. *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids*. 1582, 1–3 (2002), 45–51.
[442] Tang, X. et al. 2015. Lipid phosphate phosphatases and their roles in mammalian physiology and pathology. *Journal of Lipid Research*. 56, 11 (2015), 2048–2060.
[443] Weaver, G.A. et al. 1988. Short chain fatty acid distribution of enema samples from a sigmoidoscopy population:an association of high acetate and low butyrate ratios with adenomatous polyps and colon cancer. *Gut*. 29, (1988), 1539–1543.
[444] Stewart, B.W. and Wild, C.P. 2014. World Cancer Report 2014. *World Health Organization*. (2014), 1–2.
[445] Lasry, A. et al. 2016. Inflammatory networks underlying colorectal cancer. *Nature Immunology*. 17, 3 (2016), 230–240.
[446] Cross, A.J. et al. 2010. A large prospective study of meat consumption and colorectal cancer risk: An investigation of potential mechanisms underlying this association. *Cancer Research*. 70, 6 (2010), 2406–2414.
[447] 2007. *Food, Nutrition, Physical Activity, and the Prevention of Cancer*.
[448] Bénard, F. et al. 2018. Systematic review of colorectal cancer screening guidelines for average-risk adults: Summarizing the current global recommendations. *World Journal of Gastroenterology*. 24, 1 (2018), 124–138.
[449] Cubiella, J. et al. 2018. Targeted UPLC-MS Metabolic Analysis of Human Faeces Reveals Novel Low-Invasive Candidate Markers for Colorectal Cancer. *Cancers*. 10, 9 (2018), 300.
[450] Sobhani, I. et al. 2013. Microbial dysbiosis and colon carcinogenesis: Could colon cancer be considered a bacteria-related disease? *Therapeutic Advances in Gastroenterology*. 6, 3 (2013), 215–229.

[451] Tjalsma, H. et al. 2012. A bacterial driver-passenger model for colorectal cancer: Beyond the usual suspects. *Nature Reviews Microbiology*. 10, 8 (2012), 575–582.
[452] Dove, W.F. et al. 1997. Intestinal neoplasia in the Apc(Min) mouse: Independence from the microbial and natural killer (beige locus) status. *Cancer Research*. 57, 5 (1997), 812–814.
[453] Sellon, R.K. et al. 1998. Resident enteric bacteria are necessary for development of spontaneous colitis and immune system activation in interleukin-10-deficient mice. *Infection and Immunity*. 66, 11 (1998), 5224–5231.
[454] Uronis, J.M. et al. 2009. Modulation of the intestinal microbiota alters colitis-associated colorectal cancer susceptibility. *PLoS ONE*. 4, 6 (2009).
[455] Vogelstein, B. and Kinzler, K.W. 1993. The multistep nature of cancer. *Trends in Genetics*. 9, 4 (1993), 138–141.
[456] Kostic, A.D. et al. 2013. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor immune microenvironment. *Cell Host and Microbe*. 14, 2 (2013), 207–215.
[457] Rubinstein, M.R. et al. 2013. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. *Cell Host and Microbe2*. 14, 2 (2013), 195–206.
[458] Zeller, G. et al. 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular systems biology*. 10, 11 (2014), 766.
[459] Shah, M.S. et al. 2018. Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. *Gut*. 67, 5 (2018), 882–891.
[460] Feng, Q. et al. 2015. Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nature Communications*. 6, (2015).
[461] Peng, H. et al. 2014. Co-occurrence of driver and passenger bacteria in human colorectal cancer. *Gut Pathogens*. 6, 1 (2014), 26.
[462] Sinha, R. et al. 2016. Fecal Microbiota, Fecal Metabolome, and Colorectal Cancer Interrelations. *Plos One*. 11, 3 (2016), e0152126.
[463] Yachida, S. et al. 2019. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nature Medicine*. 25, 6 (2019), 968–976.
[464] Mayo, R. et al. 2018. Metabolomic-based noninvasive serum test to diagnose nonalcoholic steatohepatitis: Results from discovery and validation cohorts. *Hepatology Communications*. 2, 7 (2018), 807–820.
[465] Cavill, R. et al. 2011. Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. *PLoS Computational Biology*. 7, 3 (2011).
[466] Mi, H. et al. 2019. PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*. 47, D1 (2019), D419–D426.
[467] Picart-Armada, S. et al. 2018. FELLA: An R package to enrich metabolomics data. *BMC Bioinformatics*. 19, 1 (2018), 1–9.
[468] Bindea, G. et al. 2009. ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*. 25, 8 (2009), 1091–1093.
[469] Shannon, P. et al. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*. 13, 22 (2003), 6.
[470] psych: Procedures for Personality and Psychological Research, Version = 1.8.12.: 2018. *https://cran.r-project.org/package=psych*.
[471] R package "corrplot": Visualization of a Correlation Matrix (Version 0.84): 2017. .
[472] Edgar, R.C. et al. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*. 27, 16 (2011), 2194–2200.
[473] Oksanen, J. et al. 2018. Vegan: community ecology package. *R Package Version 2. 4-6*. (2018).
[474] Sing, T. et al. 2009. ROCR: Visualizing the performance of scoring classifiers. *R package version*. 1, (2009), 4.
[475] Douglas, G.M. et al. 2019. PICRUSt2: An improved and extensible approach for metagenome inference. *bioRxiv*. (Jan. 2019), 672295.
[476] Castellarin, M. et al. 2012. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Research*. 22, (2012), 299–306.
[477] Farshidfar, F. et al. 2016. A validated metabolomic signature for colorectal cancer: Exploration of the clinical value of metabolomics. *British Journal of Cancer*. 115, 7 (2016), 848–857.
[478] Louis, P. et al. 2014. The gut microbiota, bacterial metabolites and colorectal cancer. *Nature*

Reviews Microbiology. 12, 10 (2014), 661–672.
[479] Wang, T. et al. 2012. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *The ISME Journal*. 6, 2 (2012), 320–329.
[480] McCoy, A.N. et al. 2013. Fusobacterium Is Associated with Colorectal Adenomas. *PLoS ONE*. 8, 1 (2013).
[481] Warren, R.L. et al. 2013. Co-occurrence of anaerobic bacteria in colorectal carcinomas. *Microbiome*. 1, 1 (2013), 1–12.
[482] Bullman, S. et al. 2017. Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. *Science*. 5240, November (2017), eaal5240.
[483] Vogtmann, E. et al. 2016. Colorectal cancer and the human gut microbiome: Reproducibility with whole-genome shotgun sequencing. *PLoS ONE*. 11, 5 (2016), 1–13.
[484] Gagnière, J. et al. 2016. Gut microbiota imbalance and colorectal cancer. *World Journal of Gastroenterology*. 22, 2 (2016), 501–518.
[485] Flemer, B. et al. 2017. Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut*. 66, 4 (2017), 633–643.
[486] Han, Y.W. et al. 2000. Interactions between periodontal bacteria and human oral epithelial cells: Fusobacterium nucleatum adheres to and invades epithelial cells. *Infection and Immunity*. 68, 6 (2000), 3140–3146.
[487] Weiss, E.I. et al. 2000. Attachment of Fusobacterium nucleatum PK1594 to mammalian cells and its coaggregation with periodontopathogenic bacteria are mediated by the same galactose-binding adhesin. *Oral Microbiology and Immunology*. 15, 6 (2000), 371–377.
[488] Krisanaprakornkit, S. et al. 2000. Inducible expression of human β-defensin 2 by Fusobacterium nucleatum in oral epithelial cells: Multiple signaling pathways and role of commensal bacteria in innate immunity and the epithelial barrier. *Infection and Immunity*. 68, 5 (2000), 2907–2915.
[489] Flanagan, L. et al. 2014. Fusobacterium nucleatum associates with stages of colorectal neoplasia development, colorectal cancer and disease outcome. *European Journal of Clinical Microbiology and Infectious Diseases*. 33, 8 (2014), 1381–1390.
[490] Ito, M. et al. 2015. Association of Fusobacterium nucleatum with clinical and molecular features in colorectal serrated pathway. *International Journal of Cancer*. 137, 6 (2015), 1258–1268.
[491] Mima, K. et al. 2016. Fusobacterium nucleatum in colorectal carcinoma tissue and patient prognosis. *Gut*. 65, 12 (2016), 1973–1980.
[492] Dinh, D.M. et al. 2015. Intestinal Microbiota, microbial translocation, and systemic inflammation in chronic HIV infection. *Journal of Infectious Diseases*. 211, 1 (2015), 19–27.
[493] Fleissner, C.K. et al. 2010. Absence of intestinal microbiota does not protect mice from diet-induced obesity. *British Journal of Nutrition*. 104, 6 (2010), 919–929.
[494] Martínez, I. et al. 2009. Diet-induced metabolic improvements in a hamster model of hypercholesterolemia are strongly linked to alterations of the gut microbiota. *Applied and Environmental Microbiology*. 75, 12 (2009), 4175–4184.
[495] Clavel, T. et al. 2014. The Family Coriobacteriaceae. *The Prokaryotes: Actinobacteria*. 1–1061.
[496] Marchesi, J.R. et al. 2011. Towards the human colorectal cancer microbiome. *PLoS ONE*. 6, 5 (2011).
[497] Maruo, T. et al. 2008. Adlercreutzia equolifaciens gen. nov., sp. nov., an equol-producing bacterium isolated from human faeces, and emended description of the genus Eggerthella. *International Journal of Systematic and Evolutionary Microbiology*. 58, 5 (2008), 1221–1227.
[498] Zheng, W. et al. 2019. Compositional and functional differences in human gut microbiome with respect to equol production and its association with blood lipid level: A cross-sectional study. *Gut Pathogens*. 11, 1 (2019), 1–9.
[499] Murphy, N. et al. 2018. A prospective evaluation of plasma polyphenol levels and colon cancer risk. *International Journal of Cancer*. 143, 7 (2018), 1620–1631.
[500] Han, S. et al. 2018. Role of intestinal flora in colorectal cancer from the metabolite perspective: A systematic review. *Cancer Management and Research*. 10, (2018), 199–206.
[501] Buitenwerf, E. et al. 2017. Cholesterol delivery to the adrenal glands estimated by adrenal venous sampling: An in vivo model to determine the contribution of circulating lipoproteins to steroidogenesis in humans. *Journal of Clinical Lipidology*. 11, 3 (2017), 733–738.
[502] Farhana, L. et al. 2016. Bile acid: A potential inducer of colon cancer stem cells. *Stem Cell Research and Therapy*. 7, 1 (2016), 1–10.
[503] Ajouz, H. et al. 2014. Secondary bile acids: An underrecognized cause of colon cancer. *World*

Journal of Surgical Oncology. 12, 1 (2014), 1–5.
[504] Smith, P.M. et al. 2013. The Microbial Metabolites, Short-Chain Fatty Acids, Regulate Colonic Treg Cell Homeostasis. *Science*. 341, 6145 (2013), 569–573.
[505] Chang, P. V. et al. 2014. The microbial metabolite butyrate regulates intestinal macrophage function via histone deacetylase inhibition. *Proceedings of the National Academy of Sciences*. 111, 6 (2014), 2247–2252.
[506] Murphy, E.C. and Frick, I.M. 2013. Gram-positive anaerobic cocci - commensals and opportunistic pathogens. *FEMS Microbiology Reviews*. 37, 4 (2013), 520–553.
[507] Roccarina, D. et al. 2010. The role of methane in intestinal diseases. *American Journal of Gastroenterology*. 105, 6 (2010), 1250–1256.
[508] Scanlan, P.D. et al. 2008. Human methanogen diversity and incidence in healthy and diseased colonic groups using mcrA gene analysis. *BMC Microbiology*. 8, (2008), 1–8.
[509] Ishaq, S.L. et al. 2016. The Pathology of Methanogenic Archaea in Human Gastrointestinal Tract Disease. *The Gut Microbiome - Implications for Human Disease*.
[510] Abell, G.C.J. et al. 2006. Methanogenic archaea in adult human faecal samples are inversely related to butyrate concentration. *Microbial Ecology in Health and Disease*. 18, 3–4 (2006), 154–160.
[511] Wu, X. et al. 2018. Effects of the intestinal microbial metabolite butyrate on the development of colorectal cancer. *Journal of Cancer*. 9, 14 (2018), 2510–2517.

Supplementary information

Supplementary Table 1: *Summary of the tools, resources and methods used in this Thesis project. The table includes the resource name, it's accessibility, a brief description and the link to the tool.*

Application	*Type*	*Access*	*Description*	*Reference*
ClueGO	Cytoscape plugin	Free (license)	Groups genes in clusters depending on their functional annotation.	http://apps.cytoscape.org/apps/cluego
MetScape	Cytoscape plugin	Free	Visualization, interpretation and pathway enrichment of metabolomics data.	http://apps.cytoscape.org/apps/metscape
GEO	Database	Free	Database of transcriptomics studies.	https://www.ncbi.nlm.nih.gov/geo/
GreenGenes	Database	Free	16S rRNA gene database, useful for taxonomic annotation.	https://greengenes.secondgenome.com
HMDB	Database	Free	Database for human metabolites, including functional, experimental, clinical and physicochemical features.	http://www.hmdb.ca
KEGG	Database	Free (limited)	Collection of databases including genomes, metabolome, biological pathways and diseases, etc.	https://www.genome.jp/kegg/
METLIN	Database	Free	Repository of experimental mass spectrometry data, used for metabolite identification from fragmentation peaks.	https://metlin.scripps.edu/landing_page.php?pgcontent=mainPage
Nucleotide	Database	Free	Database of sequences from published articles.	https://www.ncbi.nlm.nih.gov/nucleotide/
OMIM	Database	Free	Database of known phenotype-genotype associations, focused on human genetic disorders.	https://www.omim.org

Application	*Type*	*Access*	*Description*	*Reference*
PubMed	Database	Free	Database of published articles.	https://www.ncbi.nlm.nih.gov/pubmed/
STRING	Database	Free	Protein-protein interaction database, including experimental, computationally predicted and literature mined ones.	https://string-db.org
UniProt	Database	Free	Database of proteins, including functional, clinical and physicochemical characteristics.	https://www.uniprot.org
Virtual Metabolic Human	Database	Free	Combined database of manually curated metabolic reconstructions, both human and bacterial.	https://www.vmh.life
Python	Programming language	Free	Programming language for general purposes.	https://www.python.org
R	Programming language	Free	Programming language and environment especially developed for statistical computing and graphical output.	https://cran.r-project.org
Biopython	Python module	Free	Set of python tools dedicated to biological computation, including access and interaction with biological databases.	https://biopython.org
Bioservices	Python module	Free	Python extension dedicated to provide access to several databases.	https://pypi.org/project/bioservices/
NumPy	Python module	Free	Python extension for working with large multi-dimensional arrays and matrices.	https://numpy.org

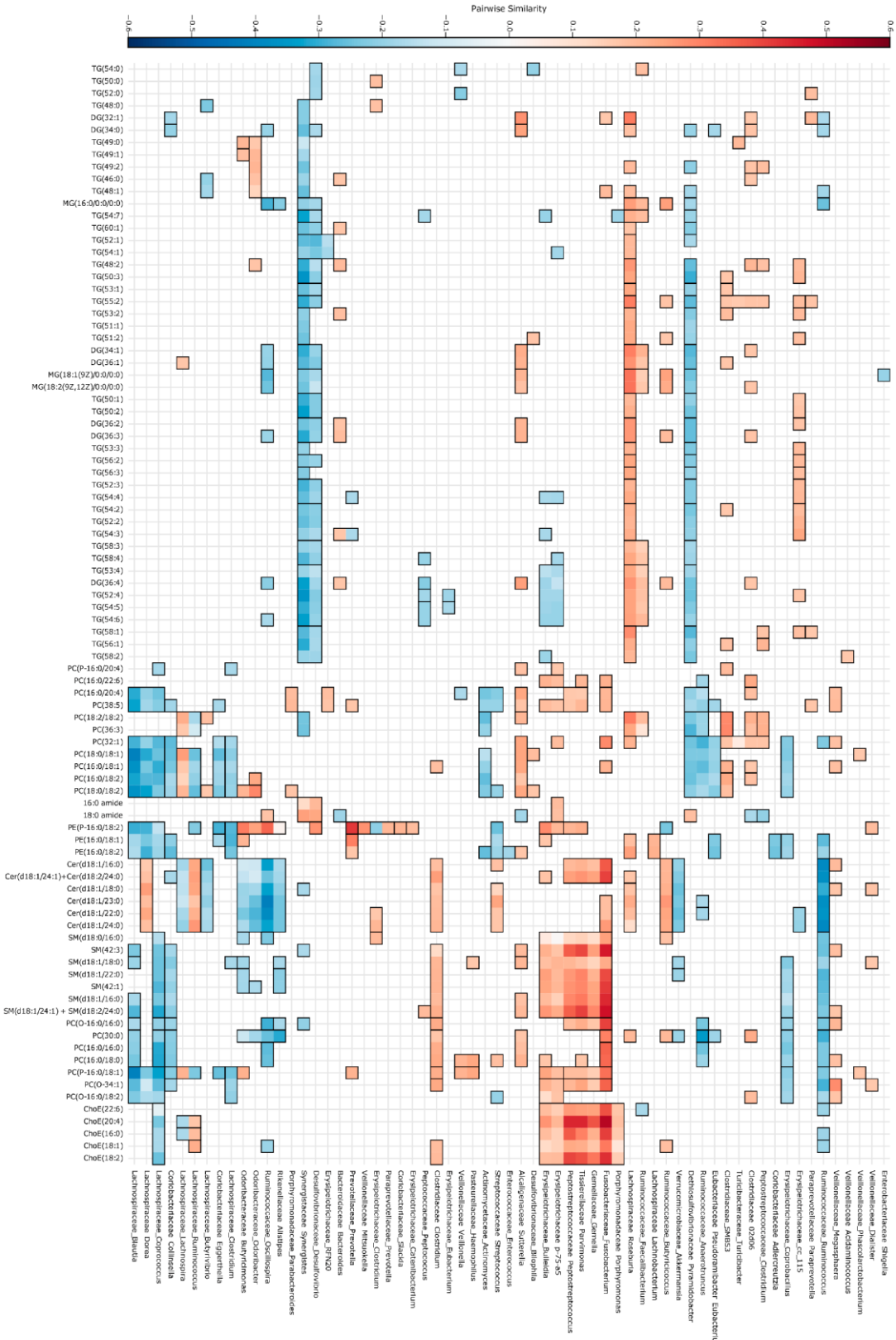
Application	*Type*	*Access*	*Description*	*Reference*
pandas	Python module	Free	Python extension that offers easy-to-use data structures and analysis tools.	https://pandas.pydata.org
DADA2	QIIME2 plugin	Free	Processing tool for Illumina demultiplexed reads, generates the final OTU table.	http://benjjneb.github.io/dada2/index.html
ALDEx2	R package	Free	Analytical tool for microbiome compositional data analysis.	https://bioconductor.org/packages/release/bioc/h
corrplot	R package	Free	Offers both functionalities and graphical options for correlation matrices analysis.	https://cran.r-project.org/web/packages/corrplot/index.html
DatR	R package	Free	Offers a set of functions to programmatically access biological databases, mainly KEGG and HMDB to aid in functional metabolomics analysis.	https://github.com/pxtm/DatR
DESeq2	R package	Free	Tool for differential analysis for count data.	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
FactoMineR	R package	Free	Extends the graphical utilities of factoextra.	http://factominer.free.fr
FELLA	R package	Free	Tool for metabolomics data enrichment against KEGG database, useful for distinct organisms enrichment.	https://bioconductor.org/packages/release/bioc/html/FELLA.html

Application	*Type*	*Access*	*Description*	*Reference*
ggplot2	R package	Free	Extends the R' s graphical options.	https://ggplot2.tidyverse.org
ggpubr	R package	Free	Extension of ggplot2 package, useful for significance analysis and display on graphs.	https://cran.r-project.org/web/packages/ggpubr/index.html
glmnet	R package	Free	LASSO and logistic regression modelling tool.	https://cran.r-project.org/web/packages/glmnet/index.html
made4	R package	Free	Includes tools for multivariate data analysis and extension of heatplot functions.	https://bioconductor.org/packages/release/bioc/html/made4.html
microbiome	R package	Free	Extension of phyloseq, designed for microbiome common analysis.	https://microbiome.github.io
mixOmics	R package	Free	Tools for multivariate data analysis and omics integration studies, focused on variable selection.	http://mixomics.org
pathview	R package	Free	Allows the download and modification of user' s selected KEGG pathways.	https://bioconductor.org/packages/release/bioc/html/pathview.html
phyloseq	R package	Free	Allows the microbiome data upload, transformation, annotation and analysis in the R environment.	https://joey711.github.io/phyloseq/

Application	*Type*	*Access*	*Description*	*Reference*
ROCR	R package	Free	Tool for evaluating and representing the performance of a classifier model.	https://rocr.bioinf.mpi-sb.mpg.de
SIAMCAT	R package	Free	R-based pipeline for microbiome-phenotype association studies and biomarker identification.	https://bioconductor.org/packages/release/bioc/html/SIAMCAT.html
vegan	R package	Free	Includes functionalities for diversity measurements and ordination methods.	https://cran.r-project.org/web/packages/vegan/index.html
plotly	R package / Python module	Free (require registration)	Allows the generation and sharing of interactive plots.	https://plot.ly
CORBATA	Software	Free	Pipeline for the identification and characterization of bacteria belonging to the core microbiome of different populations.	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3646044/
Cytoscape	Software	Free	Software for visualizing and integrating complex networks.	https://cytoscape.org
IPA®	Software	License	Enrichment tool for multiple omics data types, including up and downstream analysis, pathway identification using manually curated databases.	https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/
PICRUSt2	Software	Free	Tool for the reconstruction and abundance prediction of bacterial whole genomes from 16S gene marker data.	https://github.com/picrust/picrust2

Application	*Type*	*Access*	*Description*	*Reference*
QIIME2	Software	Free	Pipeline for the processing and basical analysis of microbiome data, extensible by correspondant plugins.	https://qiime2.org
SIMCA-P	Software	License	Software for multivariate statistics analysis, including unsupervised (PCA, HC) and supervised (PLS-DA, OPLS-DA) analyses.	https://umetrics.com/products/simca
Adonis	Statistical method	*NA*	Adonis is the MANOVA methodology used in ecology studies to explain communities with environmental variables, based on permutations of distance matrices.	
PCA	Statistical method	*NA*	Principal Components Analysis is a data dimension reduction that allows the reduction of the complexity derived from the omics data type special features. Instead, it combines the dataset variables generating a small set of principal components that collect all the complexity of the dataset in few variables, making the data easier to analyse.	
PCoA	Statistical method	*NA*	The Principal Coordinates Analysis is a variation of PCA that allows the analysis of distance matrices, expanding this way the capabilities of PCA.	
PERMANOVA	Statistical method	*NA*	Permutational Analysis of Variance is a non-parametric multivariate test used to compare groups of samples and test the corresponding null hypothesis.	
Procrustes	Statistical method	*NA*	Procrustes analysis determines the similarity between two datasets by comparing, transforming and matching the resulting shapes of the configuration of points resulting from the corresponding datasets graphical visualisation.	
Cancertool	Webtool	Free	Gene expression analysis tool from public datasets.	http://web.bioinformatics.cicbiogune.es/CANCERTOOL/

Application	*Type*	*Access*	*Description*	*Reference*
DAVID	Webtool	Free	Contains a set of functional annotation tools for large lists of genes.	https://david.ncifcrf.gov
IMPaLa	Webtool	Free	Webtool for the enrichment analysis of both gene expression and metabolites data.	http://impala.molgen.mpg.de
iTOL	Webtool	Free	Webtool for the visualization, annotation and modification of phylogenetic trees.	https://itol.embl.de
LEfSe	Webtool	Free	Webtool and pipeline to identify features that mostly contributes to differentiation between groups applying common statistical significance tests.	https://bitbucket.org/biobakery/biobakery/wiki/lefse
MetaQuery	Webtool	Free	Webtool for the annotation and abundance analysis of specific genes and bacteria in >2,000 human gut metagenome studies.	http://metaquery.docpollard.org
PANTHER-db	Webtool	Free	Tool for the enrichment analysis of genes and proteins, with options for molecular function, biological process, cellular location and pathways analyses.	http://www.pantherdb.org



Supplementary Figure 1: *Complete correlations plot between genera annotated microbiome dataset and metabolomics dataset, as obtained from HAllA tool.*

Annexes

JOURNAL OF EXTRACELLULAR VESICLES
2018, VOL. 7, 1470442
https://doi.org/10.1080/20013078.2018.1470442

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Metabolic alterations in urine extracellular vesicles are associated to prostate cancer pathogenesis and progression

Marc Clos-Garcia[a], Ana Loizaga-Iriarte[b,c], Patricia Zuñiga-Garcia[a], Pilar Sánchez-Mosquera[a], Ana Rosa Cortazar[a,c], Esperanza González[a], Verónica Torrano[a,c], Cristina Alonso[d], Miriam Pérez-Cormenzana[d], Aitziber Ugalde-Olano[c,e], Isabel Lacasa-Viscasillas[c,e], Azucena Castro[d], Felix Royo[a,f], Miguel Unda[b,c], Arkaitz Carracedo[a,c,g,h] and Juan M. Falcón-Pérez[a,f,g]

[a]CIC bioGUNE, Bizkaia Technology Park, Derio, Spain; [b]Department of Urology, Basurto University Hospital, Bilbao, Spain; [c]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC); [d]OWL Metabolomics, Bizkaia Technology Park, Derio, Spain; [e]Department of Pathology, Basurto University Hospital, Bilbao, Spain; [f]Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD); [g]Ikerbasque, Basque foundation for science, Bilbao, Spain; [h]Biochemistry and Molecular Biology Department, University of the Basque Country (UPV/EHU), Bilbao, Spain

ABSTRACT

Urine contains extracellular vesicles (EVs) that concentrate molecules and protect them from degradation. Thus, isolation and characterisation of urinary EVs could increase the efficiency of biomarker discovery. We have previously identified proteins and RNAs with differential abundance in urinary EVs from prostate cancer (PCa) patients compared to benign prostate hyperplasia (BPH). Here, we focused on the analysis of the metabolites contained in urinary EVs collected from patients with PCa and BPH. Targeted metabolomics analysis of EVs was performed by ultra-high-performance liquid chromatography–mass spectrometry. The correlation between metabolites and clinical parameters was studied, and metabolites with differential abundance in PCa urinary EVs were detected and mapped into cellular pathways. We detected 248 metabolites belonging to different chemical families including amino acids and various lipid species. Among these metabolites, 76 exhibited significant differential abundance between PCa and BPH. Interestingly, urine EVs recapitulated many of the metabolic alterations reported in PCa, including phosphathidylcholines, acyl carnitines, citrate and kynurenine. Importantly, we found elevated levels of the steroid hormone, 3beta-hydroxyandros-5-en-17-one-3-sulphate (dehydroepiandrosterone sulphate) in PCa urinary EVs, in line with the potential elevation of androgen synthesis in this type of cancer. This work supports urinary EVs as a non-invasive source to infer metabolic changes in PCa.

ARTICLE HISTORY
Received 15 January 2018
Accepted 17 April 2018

KEYWORDS
Prostate; urine; exosomes; metabolomics; metabolism; biomarkers

Introduction

Prostate cancer (PCa) is among the most frequently diagnosed and deadly types of cancer in men in Western countries (http://globocan.iarc.fr). Lack of sensitive and specific diagnostic tools, especially to detect early stages of the disease, and the unknown underlying mechanisms of onset and progression of PCa are the major problems to treat PCa with the highest efficacy. Thus, there is a high demand to discover more sensitive and specific biomarkers to improve PCa diagnosis and prognosis. Nowadays, prostate-specific antigen (PSA) blood screening tests, together with clinical T-stage and Gleason score are the standard tests to discriminate patients with low, intermediate or high risk to suffer PCa [1].

Metabolomics is recognised as the ultimate "omics" discipline with high potential to identify sensitive and specific markers, and to understand the mechanisms involved in the development of pathological processes [2]. The recent technological revolution in separation and detection of small molecules, combined with rapid progress in bioinformatics, is making possible to rapidly measure a large number of metabolites in a small amount of sample [3,4]. Metabolomics comprises the qualitative and quantitative measurement of the metabolic response to physiological or pathological stimuli. It involves the extraction and measurement of low molecular weight molecules (e.g. amino acids, sugars, bile acids, fatty acids, vitamins, etc.) belonging to different metabolic pathways to generate metabolic profiles of cells, tissues or biofluids [5,6]. Previous

CONTACT Arkaitz Carracedo acarracedo@cicbiogune.es; Juan M. Falcón-Pérez jfalcon@cicbiogune.es
Marc Clos-Garcia, Ana Loizaga-Iriarte and Patricia Zuñiga-Garcia contributed equally to the work
Supplemental data for this article can be accessed here.

studies have shown the utility of serum metabolite levels as a diagnostic tool for different cancer types [7], and in PCa some metabolites have already been suggested as candidate biomarkers. Increased serum levels of polyunsaturated fatty acids have been associated to reduce risk of PCa, while higher levels of serum testosterone were associated with an increased risk of suffering this malignancy [8]. Other metabolomics approaches have reported alterations of acyl carnitines, glucose, glycerophospholipids (including lysophosphatidylcholines and phosphatidylcholines), amino acids and triglycerides in PCa [9].

Urine samples have been intensely used to identify PCa biomarkers [10], due to its easy availability and handling, and its anatomical proximity to the prostate. As occurs for the serum, there are also several metabolomics studies of urine samples that found alterations in urinary levels of more than 20 metabolites including N-methyl glycine, kynurenine, uracil, glycerol 3-phosphate, dihydroxybutanoic acid, xylonic acid, pyrimidine, ribofuranoside and xylopyranose (reviewed in [11]). These studies have pointed out that many metabolic pathways may be altered in PCa including glycine synthesis and degradation, and carbohydrate and energy metabolisms. Although all these metabolites need further clinical validation, they support the notion that metabolomics constitutes a suitable technology to identify candidate biomarkers of PCa.

One important drawback of using urine sample for biomarker discovery is that many of their constituents are diluted avoiding to be detected by current technologies. Thus, in order to detect underrepresented molecules, it is still required to concentrate the sample. In this context, cell-secreted extracellular vesicles (EVs) are present in all body fluids, including urine [12], and could provide a concentrated source of molecules. Thus, a deep analysis of the urinary EVs composition could open a window of opportunities to identify more sensitive and specific PCa biomarkers. In line, a recent lipidomics study performed in these urinary vesicles from healthy and PCa samples reveal up to nine lipid species differentially expressed as potential PCa biomarkers [13] supporting the existence of metabolic changes in urine EVs from PCa patients.

In the current study, we have compared urinary EVs obtained from PCa and benign prostate hyperplasia (BPH) patients, and focused on the analysis of the metabolites that they contain by performing an UHPLC-MS targeted metabolomics analysis. We evaluated the levels of 248 metabolites belonging to different chemical nature including amino acids, nucleosides, vitamins, as well as different lipid species. Among them, 76 metabolites were found significantly altered in PCa compared to BPH. Some of these metabolites were significantly correlated with current markers of PCa (e.g. PSA). Interestingly, dehydroepiandrosterone sulphate was among the most significantly altered metabolites in PCa, supporting the notion that beyond their function as "metabolic machines" [4,14,15], EVs could inform about metabolic alterations of cancerous tissue.

Materials and methods

Patient samples

All urine samples were obtained from the Basque Biobank for research (BIOEF, Basurto University hospital) upon informed consent and with evaluation and approval from the corresponding ethics committee (CEIC code OHEUN11-12 and OHEUN14-14). Clinical classification of the patients is described in Table 1. For each sample, urine (50 ml) was collected by spontaneous micturition, centrifuged at 2,000 × g 10 min, filtered through a 0.22 µm-pore membrane and immediately frozen at −80ºC.

Urine extracellular vesicle isolation and characterisation

To isolate EVs from urine (average ± SEM; 49.7 ± 0.86 ml), the stored samples were thawed, centrifuged at 10,000 × g for 30 min and the supernatant ultra-centrifuged at 100,000 × g for 75 min. The resulting pellet was washed with an excess of phosphate-saline buffer (PBS), and again ultra-centrifuged at 100,000 × g for 60 min. Final pellet was re-suspended in 150 µL of PBS, aliquot generated and kept at −80ºC for further analysis. Protein was determined by Bradford and obtained 32.7 ± 4.6 (mean±SEM) micrograms on average of total purified protein from the initial

Table 1. Clinical classification of the samples.

Disease status	Stage	Perineural invasion	*n*
Prostate cancer (PCA) (64 ± 4.41)	Stage 2 (64 ± 4.12)	No (Pn0) (65.5 ± 5.02)	6
		Yes (Pn1) (64 ± 3.47)	10
	Stage 3 (64.5 ± 4.68)	*NA*	15
Benign hyperplasia (BPH) (70 ± 5.71)	*NA*	*NA*	14

In parentheses are indicated the median ± SD of age for each group of samples.

urine volume (50 ml). Size distribution of the particles present in the isolated preparations was determined by measuring the Brownian motion using a NanoSight LM10 system equipped with a fast video capture and particle-tracking software (Malvern, UK). Pre- and post-acquisition settings were maintained the same for all the samples and each video was analysed to give the mean, mode, and median vesicle size, as well as an estimate of the particle concentration. Then, an average curve was calculated for each group of patients to be compared among them. Cryo-electron microscopy and Western-blot analysis were performed as describe previously [16].

Metabolite extraction and UHPLC-MS analysis

Metabolic profiles of urinary EVs were semi-quantified using four UHPLC-MS based analytical platforms as previously described [17,18]. Methanol was first added to urinary EV preparations, and after brief vortex, chloroform was added. Both extraction solvents were spiked with metabolites not detected in unspiked EV extracts: tryptophan-d5(indole-d5), PC(13:0/0:0), FA (19:0), dehydrocholic acid, SM(d18:1/6:0), PE(17:0/17:0), PC(19:0/19:0), TAG(13:0/13:0/13:0), Cer(d18:1/17:0), ChoE(12:0), anthranilic acid-(ring-13C6), phenylthiohydantoin (PTH)-valine and glycocholic-2,2,4,4-d4 acid. Samples were incubated at −20°C for 30 min and, after vortex, three different phases were collected. Platform 1 included fatty acyls, bile acids, steroids and lysoglycerophospholipids profiling. Supernatants were collected after centrifugation at 16,000 × g for 15 min, dried, reconstituted in methanol, resuspended for 20 min and centrifuged (16,000 × g for 5 min) before being transferred to vials for UHPLC-MS analysis. Platform 2 included glycerolipids, cholesteryl esters, sphingolipids and glycerophospholipids profiling. Extracts were mixed with water (pH = 9) and after brief vortex mixing, the samples were incubated for 60 min at −20°C. After centrifugation at 16,000 × g for 15 min, the organic phase was collected and the solvent removed. The dried extracts were then reconstituted in acetronitrile/isopropanol (50:50), resuspended for 10 min, centrifuged (16,000 × g for 5 min) and transferred to vials for UHPLC-MS analysis. Platform 3 included amino acids profiling; 10 µl aliquots from the extracts prepared for Platform 1 were transferred to microtubes and derivatised for amino acid analysis. Finally, Platform 4 consisted in the analysis of polar metabolites profiling, including central carbon metabolism. Extracts were mixed with chloroform. After brief vortex mixing, water was added and samples were mixed for 10 min at room temperature. Afterwards, samples were centrifuged at 16,000 × g for 10 min. The supernatants were collected and dried. Extracts were then solubilised in water and after centrifugation, supernatants were transferred to vials for UHPLC-MS analysis.

Chromatographic separation and mass spectrometric detection conditions employed were previously described [17,18]. The overall quality of the analysis procedure was monitored using six repeat injections of a pooled sample, considered as the quality control sample. For each of the four analytical platforms, randomised sample injections were performed, with QC calibration and validation extracts uniformly interspersed throughout the entire batch run. Generally, the retention time stability was <6 s injection-to-injection variation and the mass accuracy <3 ppm for m/z 400–1200, and <1.2 mDa for m/z 50–400. Details of lipid nomenclature used in this work is provided as supplementary material.

Data processing, statistical and bioinformatics analyses

Amount of urine sample and data normalisation

A similar volume of urine sample (50 ml) from each patient was employed for obtaining the EV preparations. Then, the complete EV preparations were analysed by UPLC-MS metabolomics analysis. The peak intensities for each metabolite included in the analysis were normalised to the sum of the peak intensities within each sample. There was no significant correlation (F < Fcrit) between the sum of the peak intensities used for the normalisation and the groups being compared in the study.

Missing values imputation

First, metabolites that were not detected in at least 70% of the whole set of samples were removed from the analysis. Then, taking the minimal value for each metabolite and dividing it by a factor of 10, missing values were imputed in order to obtain the final data set.

Univariate analysis

Three different comparisons were established for the analyses:

- Prostate cancer (PCa) *vs* benign prostate hyperplasia (BPH).
- PCa pathological stage 3 *vs* PCa pathological stage 2.
- In the PCa pathological stage 2 group, perineural invasion: Pn1 *vs* Pn0.

The mean and 90% Winsorized-mean for each metabolite and each group of patients were calculated, as well as, Student's *t*-test or Wilcoxon signed-rank test, depending on the normality of the data that was assessed using Shapiro-Wilk test. Median, standard error of the mean (SEM), the standard deviation (SD), coefficient of variation and the Interquartile Range (IQR) were also calculated.

Several calculations were performed for the three distinct comparisons. We calculated the F-test of the two variances, the Student's *t*-test, Wilcoxon signed-rank and Fold Change for each metabolite. To test the discriminatory capacity of each metabolite for each one of the three comparisons we performed Receiver Operating Characteristic (ROC) analysis, including in the calculations the values of the Area Under the Curve (AUC), sensitivity, specificity, positive predictive value, negative predictive value, Youden index and the optimal cut-off.

For each one of the three pairwise comparisons, we generated box-plots for those metabolites with significant differences between the two groups with adjusted *p*-values following Bonferroni methodology. Heatmaps indicating $\log_2$ value of Fold Change and Bonferroni adjusted *p*-values were also calculated. Finally, volcano plots were generated with the $\log_2$ Fold Change values and Bonferroni adjusted *p*-values.

All statistical analyses were performed using R software v3.3.2 (R Development Core Team, 2016; http://cran.r-project.org) with stats, caret, psych and OptimalCutpoints package [19]. Boxplots and volcano plots were generated with ggplot2 R package. Correlations with clinical parameters such as BMI were done with cor.test function in R software, using Spearman's method. Both *rho* and *p*-value for each metabolite are reported. We studied the correlation of BMI and metabolite levels with all the samples together and also dividing samples depending on their clinical status.

Multivariate analysis

Principal Component Analysis (PCA), Partial Least Squares-Discriminant Analysis (PLS-DA) and Orthogonal Partial Least Squares (OPLS) were performed for each pairwise comparison using SIMCA-P v12.0.1.0 software (Umetrics *AB*).

Metabolites mapping into cellular metabolic pathways and identification of primary enzymes associated with their metabolism

Metabolic pathways were determined with MetScape v3.1.2 application, running under Cytoscape v3.5.0 software, linking them to KEGG Pathway database (http://www.genome.jp/kegg/pathway.html). Primary enzymes involved in the metabolism of the metabolite of interest, and their corresponding coding genes were retrieved from KEGG (*http://www.genome.jp/kegg/compound/*) and HMDB (*http://www.hmdb.ca/*) databases, using dbWalk utility on bioDBnet database searching online utility and specifying "9606" (Homo sapiens) Taxon ID on Organism box (https://biodbnet-abcc.ncifcrf.gov/db/dbWalk.php), with the following paths:

- For KEGG compounds, we started with enzyme EC code:

EC Number->UniProt Accession->UniProt Entry Name->KEGG Gene ID->Gene ID->Gene Symbol->Gene ID->GenBank Nucleotide Accession.

- For HMDB compounds, we started with the name present on HMDB database:

HMDB Metabolite->HMDB Enzyme -> UniProt Entry Name->Gene Symbol->Gene ID->GenBank Nucleotide Accession.

For each metabolite included in this step, we reported:

- For KEGG compounds, the related enzymes EC number, UniProt Accession, UniProt Entry Name, KEGG Gene ID, Gene Symbol, GeneID and the GenBank Nucleotide Accession for the corresponding transcripts.
- For HMDB compounds, the HMDB enzyme Gene Symbol, Gene Symbol, Gene ID and GenBank Nucleotide Accession for the corresponding transcripts.

Database normalisation: all the datasets used for the data mining analysis were downloaded from GEO or TCGA, and subjected to background correction, $\log_2$ transformation and quartile normalisation as reported [20,21]. In the case of using a pre-processed dataset, this normalisation was reviewed and corrected if required. For normal *vs.* PCa comparisons, a two-tailed *t*-test is performed in order to indicate if the observed differences between the groups are significant. For tumour progression analysis, an ANOVA test was performed in order to evaluate if the observed differences of gene expression levels between the groups were significant. DFS analysis was performed using Taylor and TCGA datasets. In both cases, the patients were stratified by quartiles based on the expression of the

gene of interest, Kaplan-Meier Estimator was used in order to estimate the survival function from different groups of patients while a Log-Rank test is calculated to check the significance between the curves. In the case of Taylor dataset, the analysis was performed using the average signal from all the transcripts of a gene.

Results

Urine samples were collected from patients with BPH (n = 14) and PCa (n = 31) with different pathological characteristics (Table 1). In order to avoid any chemical alteration of the vesicles that could interfere with the metabolomics analysis, we decided to preserve uromodulin status of the samples by avoiding the use of high-salt concentration or reducing agents. After initial clearing at low centrifugation and ultrafiltration, small EVs (exosomes, small microvesicles and apoptotic blebs) were isolated by differential ultracentrifugation as described in [16]. Cryo-electron microscopy revealed the presence of vesicles in the preparations (Supplementary Figure 1A). Western-blot analysis showed that while we could not detect mitochondria (COX IV) or endoplasmic reticulum (GRP78) proteins, we could detect exosomal markers (CD10, CD63, CD9, Flotillin and CD26), and also some uromodulin (THP) (Supplementary Figure 1B). As previously, we found a high inter-individual variability in the abundance of these proteins [16,22]. In agreement with previous results [16], physical characterisation by NTA analysis of the isolated material revealed significant differences in the size distribution of particles isolated from PCa and BPH samples (Figure 1). Interestingly, the size of the particles increased with the stage of the PCa, thus, the major difference was observed between BPH and PCa stage 3 (Figure 1). A significant higher abundance of particles bigger than 350 nm were observed in samples from PCa stage 3 (Figure 1(d)). The mean concentration of particles per mL for all samples was $8.60 \pm 1.19 \times 10^{10}$ EVs/mL. No differences were found for the concentrations of EVs/mL between different groups (BPH, PCa stage 2 Pn0, stage 2 Pn1 and stage 3).

After this initial characterisation, metabolites present in the urinary EV preparations were extracted using different methodologies in order to cover a wide range of molecules with different chemical nature (see *Methods* section). We were able to detect 248 metabolites (Supplementary Table 1) including amino acids, vitamins, nucleosides, as well as different lipid species. Considering all the samples, metabolites with more than 70% of missing values were eliminated from the analysis with the exception of PC(14:0/20:4), PC (0:0/20:3) and TG(56:8) because most of missing values occur mainly in one of the two groups (PCa or BPH). Afterwards, we performed three different statistical analysis comparing BPH and PCa groups, as well as, the association to tumour stage and perineural invasion.

Metabolites differentially altered between BPH and pca

Univariate analysis revealed that 76 out of 248 metabolites showed statistically significant differences between EVs from PCa and BPH patients. These metabolites were distributed along most chemical families analysed, although there was a predominance of phosphatidylcolines (PC), fatty acid esters (acyl carnitines) and sterols (Figures 2, 3 and Supplementary Figure 2). Whereas higher abundance of PC was observed in BPH samples, acyl carnitines and sterols were more abundant in PCa samples (Figures 2 and 3). In addition, carboxylic acids and glycerolipids were slightly decreased, and vitamins were increased in PCa EVs. The other families of metabolites including amino acids, bile acids, nucleosides, sphingolipids, phosphatidylethanolamines (PE) contained both increased and decreased metabolites (Figure 3). Interestingly, the abundance of ceramides with short carbon number in their acyl chains were increased in PCa samples, while ceramides with long carbon number (>23) in their acyl chains were reduced in PCa EVs. This pattern was not present in other sphingolipids families. In the non-esterified fatty acid family, the abundance of arachidonic acid (20:4n-6) was decreased in PCa samples, while other polyunsaturated fatty acid with shorter carbon chain (16:3n-x) was significantly increased in the PCa group (Figure 3).

Multivariate analysis by principal component analysis (PCA) did not show a perfect separation of the two groups, although PCa EV samples tended to aggregate all together, whereas BPH samples were more disperse (Figure 4(a)). Statistics of the model indicate low degree of fit (2n component R2X = 0.49) and also low predictability (2n component Q2X = 0.37). The PCA loadings plot (Figure 4(b)) indicated that the differences between PCa and BPH samples were explained mainly by different subfamilies of glycerophospholipids, confirming what was identified with the univariate analysis.

Metabolites differentially altered between PCa stage 2 and stage 3

PCa stage is a pathological sign of disease aggressiveness [1]. In an attempt to identify potential biomarkers

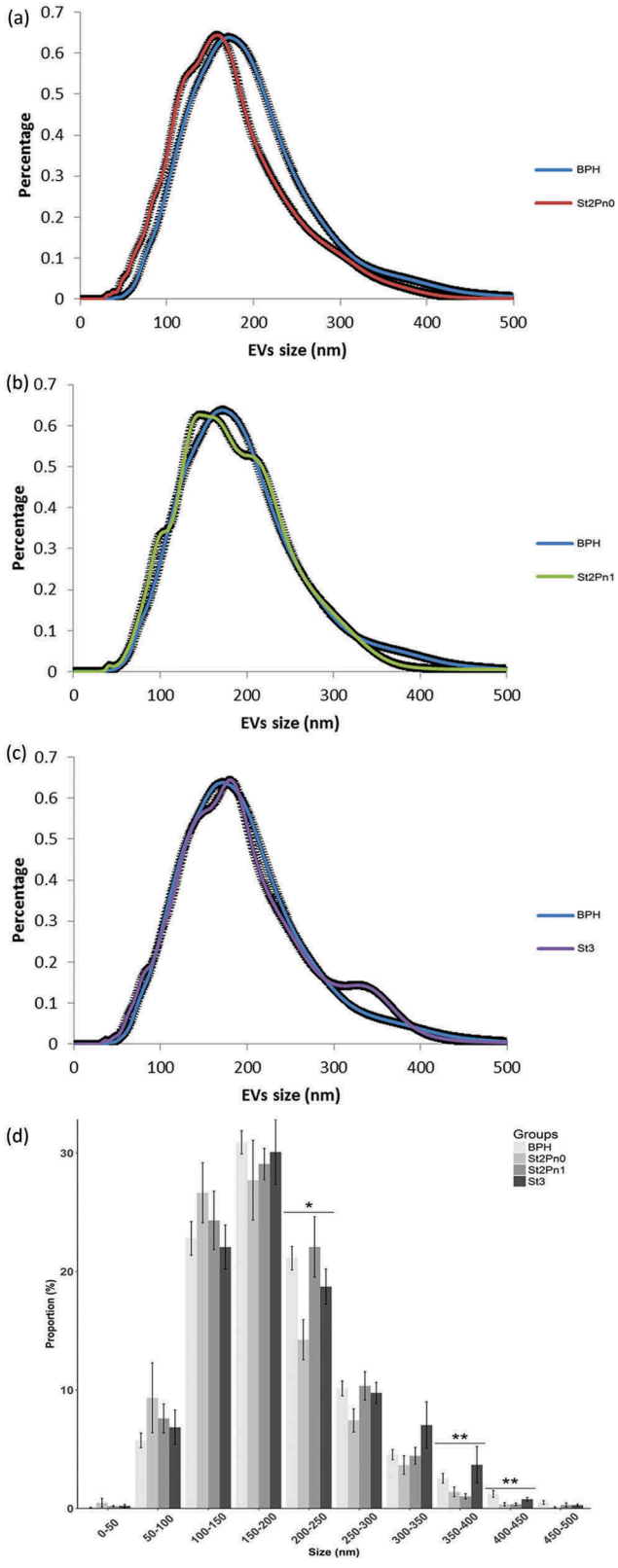


Figure 1. Size distribution of urinary EVs isolated from the BPH and PCa groups.

Pairwise comparison BPH vs PCa stage 2 without perineural invasion (a), pairwise comparison BPH vs PCa stage 2 with perineural invasion (b) and pairwise comparison BPH vs PCa stage 3 (c). Size distribution of the particles isolated from each patient, including SEM error bars (d). Number of samples: BPH (n = 14), Stg.2 Pn0 (n = 6), Stg.2 Pn1 (n = 10) and Stg.3 (n = 13), all of them analysed in duplicate. Kruskal-Wallis Rank Sum test was applied to study the significance of the sizes distribution differences (*p < 0.05 and **p < 0.01).

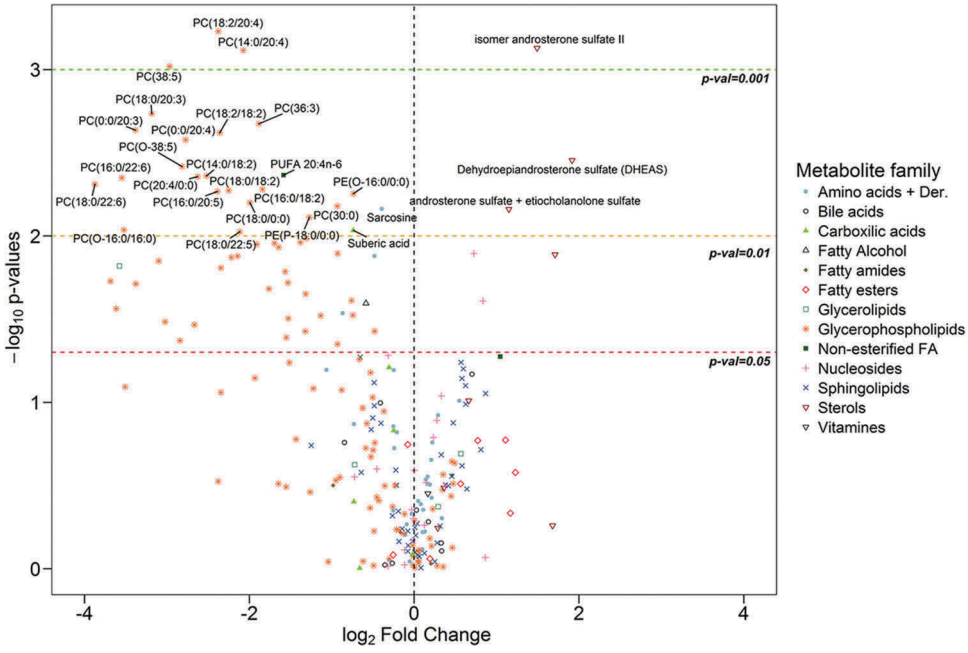


Figure 2. Volcano plot for BPH (n = 14) *vs* PCa (n = 31).

Positive fold change indicates an increase of the metabolite in PCa, while a negative value indicates that the levels are reduced in PCa. Dots shape and colour depend on metabolite families.

to discriminate between different stages of PCa, we performed univariate analysis comparing the PCa stage 2 and stage 3 subgroups. We identified 5 metabolites that showed significant differences between the two groups (Figure 5(a)). These metabolites were three ceramides, Cer(d18:1/16:0), Cer(d18:1/20:0), Cer (d18:1/22:0) one glycerophospholipid PC(30:0) [which is a combination of the isomers PC(16:0/14:0) and PC (14:0/16:0)] and one acyl carnitine, stearoylcarnitine [AC(18:0)]. In addition, we also observed a non-significant trend in other metabolite families. Thus, fatty esters, glycerolipids (both diacylglycerols and triacylglycerols), fatty amides, vitamins and 1-monoetherglycerophosphocholines showed an increase in their abundance in the PCa stage 3 group (Supplementary Table 1). In contrast, the levels of most of the metabolites belonging to the sphingolipids family including ceramides, monohexosylceramides and sphingomyelins, as well as fatty alcohols, some glycerophospholipids subgroups and nucleosides were reduced in stage 3. In this comparison, unsupervised multivariate analysis could not achieve any separation between different PCa stages, and although supervised PLS-DA analysis was able to discriminate (R2X 0.47, Q2X 0.07), its loadings plot showed that the major influence in the separation corresponded to the aforementioned five metabolites (*data not shown*) detected in the univariate analysis.

Metabolites differentially altered between PCa stage 2 perineural invasion: Pn1 *vs* Pn0

Perineural invasion in PCa has been associated to prostate cancer prognosis [23]. Although a limited number of samples were available, we also attempted to identify metabolites tentatively associated to this pathological feature. By univariate analysis, we detected significant lower abundance of cyclic AMP (cAMP) and higher abundance of the combination of isomers androsterone sulphate and etiocholanolone sulphate in the EV samples obtained from PCa patients with perineural invasion (Figure 5(b)). In addition, although not significant, three bile acids showed lower levels in samples with perineural invasion (Supplementary Table 1). Unsupervised multivariate analysis was not able to separate the two groups of samples, but we could achieve this separation with PLS-DA test (R2X 0.40, Q2X 0.58) (*data not shown*).

Correlation analysis of metabolic profiling with body mass index (BMI)

When studying circulating metabolites, the systemic metabolic state can be a critical contributing factor that can influence the results of the analysis. Obesogenic diets and obesity impact on biofluid metabolite concentration, and can also have a central effect on tumour tissues [24] by altering their biological

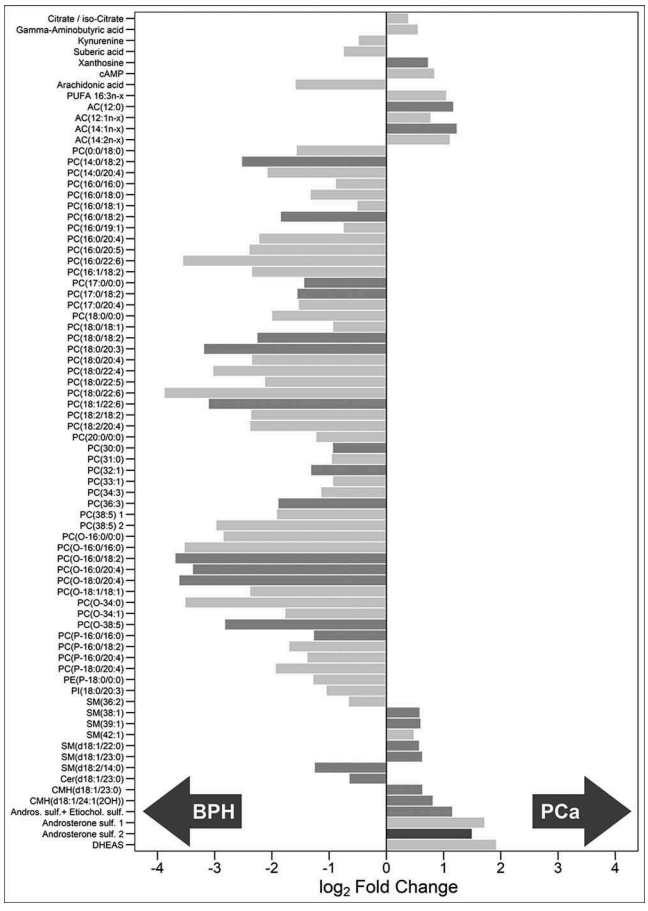


Figure 3. Metabolites associated to urinary EVs differentially expressed between BPH (n = 14) and PCa samples (n = 31).

Bars have been coloured depending on the significance of the differences, being lighter gray for the p-values between 0.05 and 0.01, medium gray for p-values between 0.01 and 0.001 and darker gray for p-values lower than 0.001.

features. Therefore, we considered evaluating the changes in urine EV metabolites that were associated to the body mass index (BMI). Samples were divided into three groups, corresponding to their calculated BMI: lean (<25), overweight (>25 and <30) and obese (>30). Taking into account all the samples independently of their BPH or PCa classification, no significant correlation was found between BMI and any of the 248 metabolites analysed in this study. Afterwards, we explored if some metabolites were correlating with BMI inside different groups. In the lean BMI group, some sterol-related metabolites including isomer pregn-5-ene-3,20-diol sulphate and isomer androsterone sulphate showed significant positive correlations with *rho* values of 0.72 and 0.60, respectively (Table 2). On the contrary, diacylglycerol DG(36:3), PC(18:2/00) and triglyceride TG(56:3) showed significant negative correlation with *rho* values of −0.71, −0.69 and −0.67, respectively (Table 2). In the case of the overweight BMI group, significant positive correlation was found with the exogenous metabolite, hydroxyphenyllactic acid (ρ 0.69). Sphingomyelin SM(43:1) showed significant negative correlation (ρ −0.67) with BMI values (Table 2). In the obese group, we observed a high degree of correlation of some metabolites with the BMI values. Thus, acyl carnitine AC(8:0) (ρ 0.94) and arginine (ρ 0.85) showed significant positive correlation, while 13 sphingomyelins, 8 phosphatidylethanolamines and the polyunsaturated fatty acid (16:3n-3) showed negative correlations with *rho* values ranging between −0.95 to −0.78) (Table 2). Finally, we evaluated if any of the metabolites correlated with BMI considering only the PCa group. Inside this group the highest positive correlations were found for taurocholic acid and dodecanoylcarnitine, AC(12:0), with *rho* values of 0.51 and 0.38, respectively.

Correlation analysis of metabolic profiling with PSA in the PCa group

PSA is the current gold standard non-invasive prognostic marker for PCa while its diagnostic potential remains controversial [25]. We performed a correlation analysis between urinary EV metabolites and the PSA values determined in our cohort of PCa samples. We only observed a significant positive correlation (*rho* value 0.88) of phosphatidylcoline PC(0:0/20:3), and at less extent (*rho* value 0.48) of the primary fatty amide (20:2n-x).

Analysis of enzymes-associated to metabolites differentially expressed between PCa and BPH

We have recently shown that metabolic alterations in PCa are frequently associated to changes in the expression of key enzymes [21]. To better understand the cancer cell autonomous nature of the metabolite changes observed in urine EVs from PCa patients, we mapped the 76 altered urinary-EV-metabolites into cellular pathways by using MetScape v3.1.2 [26]. We identified several pathways that could be affected in PCa including steroid hormone biosynthesis and metabolism, leukotriene and prostaglandin metabolisms, linoleate and purine metabolisms, glycerophospholipid metabolism, TCA and urea cycle, and tryptophan metabolism. We identified the primary enzymes involved in the metabolism of each of the 76 differentially expressed metabolites between BPH and PCa, by using KEGG or HMDB database (see *Methods*). A complete list of primary enzymes is supplied as Supplementary Table 2. Next, we took advantage of publicly available prostate cancer transcriptomes and

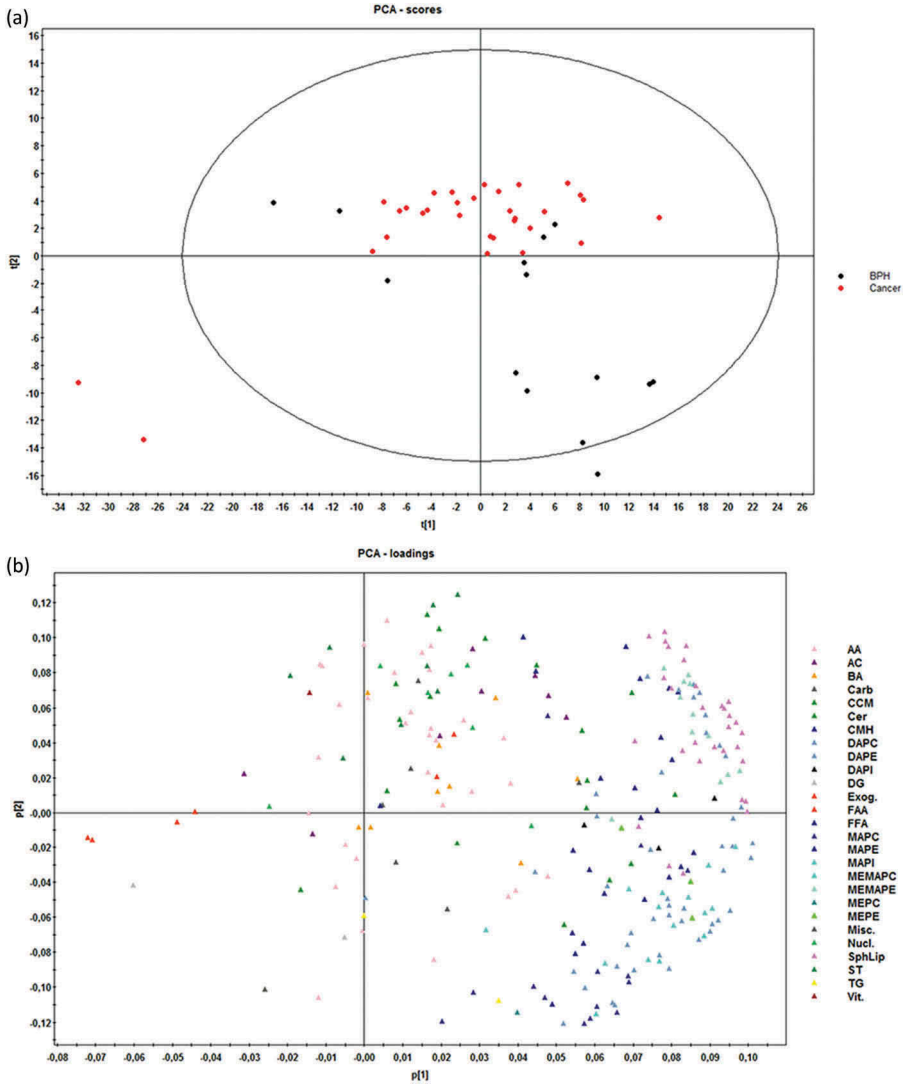


Figure 4. Score (a) and loadings (b) plots of PCA model for the comparison PCa (n = 31) *vs* BHP (n = 14).

Dots in score plot (A) have been coloured depending on its group (PCa or BHP). Markers in loadings plot (B) have been coloured depending on metabolite family. AA (amino acids), AC (acyl carnitines), BA (bile acids), Carb (carboxylic acids), CCM (derivative carboxilic acids), Cer (ceramides), CMH (monohexosylceramides), DAPC (diacylglycerophosphocholines), DAPE (diacylglycerophosphoethanolamines), DAPI (diacylglycerophosphoinositol), DG (diacylglycerols), Exog. (exogenous), FAA (fatty amides), FFA (non-esterified fatty acids), MAPC (1-monoacylglycerophosphocholine), MAPE (monoacylglycerophosphoethanolamine), MAPI (monoacylglycerophosphoinositol), MEMAPC (1-ether, 2-acylglycerophosphocholine), MEMAPE (1-ether, 2-acylglycerophosphoethanolamine), MEPC (1-monoetherglycerophosphocholine), MEPE (1-monoetherglycerophosphoethanolamine). See more details of the nomenclature in supplemental material.

we queried the expression of the 149 enzymes in PCa. We searched for enzymes which expression changes in PCa would fit the metabolite abundance observed in urine EVs. From these gene list, we identified 7 genes with the expression changes (Figure 6) that were concordant with the observed changes in urine EV metabolite abundance among the groups analysed. We found gamma-aminobutyric acid (GABA) increased in PCa urine EVs (Figure 3) which was consistent with a reduction in the expression of Glycine Amidinotransferase (GATM- use GABA as substrate for creatine synthesis) (Figure 6(a)). Arachidonic acid abundance was also altered in urine EV samples, being reduced in PCa patients compared with BPH (Figure 3). This fatty acid is the product of phospholipase A2 and it is relevant for the synthesis of pro-inflammatory metabolites by lipooxygenases. Interestingly, we found that the expression of two enzymes (ALOX15 and CYP1A2), that can catabolise arachidonic acid, was increased in PCa tissue (Figure 6 (b,c). Our metabolomics analysis also showed a consistent decrease in phosphatidylcholine. This could be explained by decreased synthesis of the phospholipid or elevated catabolism. When browsing the expression of PC synthesis and degrading enzymes, we found a reduction in the expression of Lysophosphatidylcholine

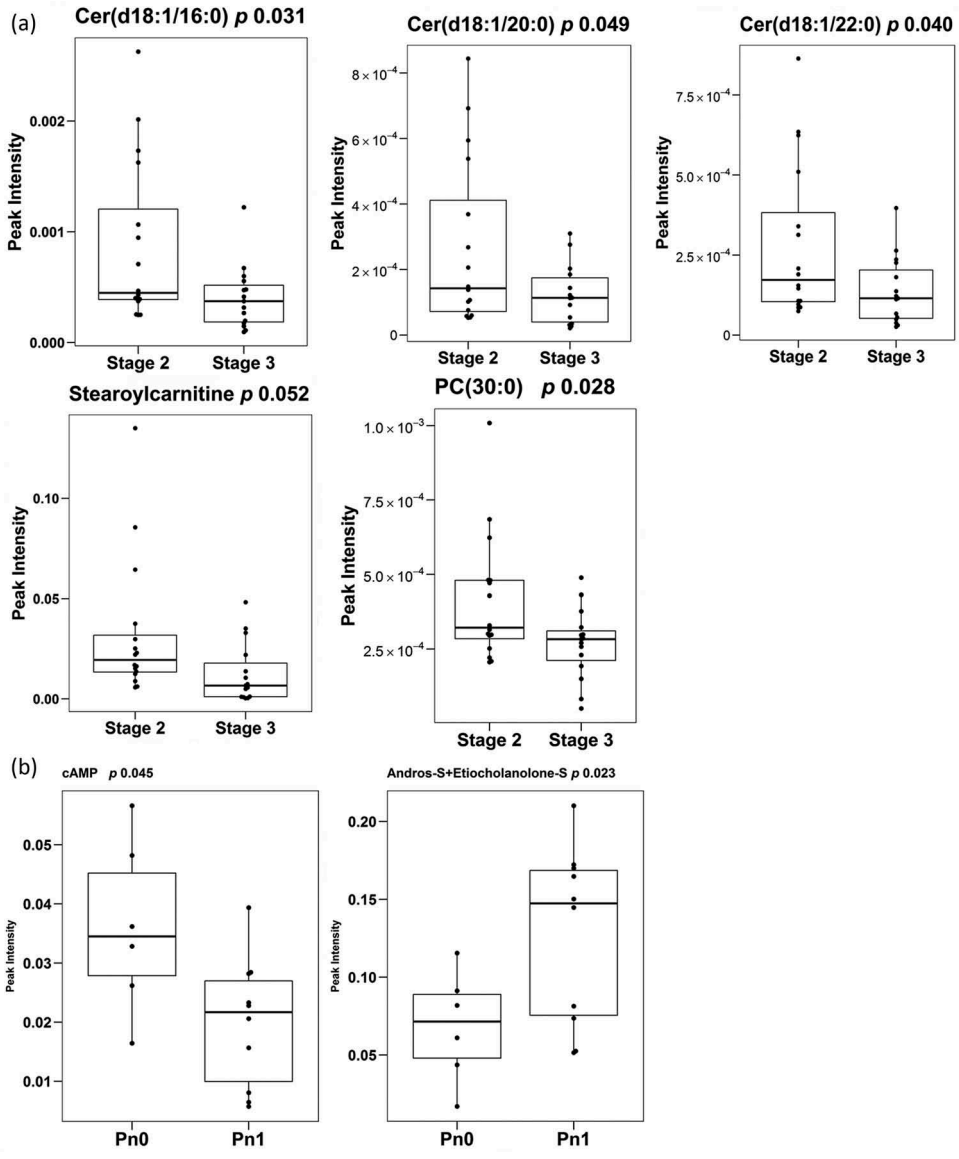


Figure 5. Differentially-expressed metabolites.

Box-plots of differentially expressed metabolites between PCa stages (A) (stage 2 $n = 16$ and stage 3 $n = 15$) and of differentially expressed metabolites between the presence and absence of perineural invasion (B) (Pn0 $n = 6$ and Pn1 $n = 10$). Significance is indicated next to metabolite name.

Acyltransferase 2 (LPCAT2) (Figure 6(d)), which transforms lysoPC into PC, and could provide an explanation for the reduction in PC abundance.

Two urine EV metabolites were associated to increased perineural invasion in PCa. On the one hand, we found a decrease in cAMP abundance in EV obtained patients with perineural invasion. The transcriptional analysis revealed changes in the expression of enzymes regulating cAMP synthesis and degradation that were associated to the aggressiveness of the disease. The expression of adenylate cyclase 5 (ADCY5) was reduced in PCa (Figure 6(e–g))), and a further significant reduction was observed from primary tumours to metastasis. In contrast, the inverse expression pattern (elevation in PCa and further increase from primary tumours to metastasis) was detected in the cAMP degrading enzyme PDE4C (Figure 6(h–j)). In none of these cAMP metabolising enzymes we could find an association to altered disease-free survival (Figure 6(g,j)).

On the other hand, the steroid biosynthesis-related metabolites were among the most elevated in PCa urine EVs, and associated with increased perineural invasion. Interestingly, the three metabolites significantly altered were sulphated steroids in the final steps of androgen synthesis. Whereas these metabolites are found at detectable levels in circulation produced by the adrenal gland, we evaluated whether enzymes regulating their synthesis or

Table 2. Correlation analysis of metabolites and BMI.

	Metabolite	Class	Correlation (*ρ*)	*p-value*
Lean	Isomer pregn-5-ene-3,20-diol sulphate	Sterol	0.72	0.003
	Isomer androsterone sulphate	Sterol	0.60	0.011
	Taurodeoxycholic acid	Bile acid	0.59	0.03
	Malate	Carboxylic acid (d)	0.56	0.025
	Arginine	Amino acid	0.53	0.027
	Glycine	Amino acid	−0.56	0.021
	TG(18:1 + 20:1 + 18:1)	Glycerolipid	−0.67	0.006
	PC(18:2/0:0)	Glycerophospholipid	−0.69	0.007
	DG(36:3)	Glycerolipid	−0.71	0.008
Overweight	Hydroxyphenyllactic acid	Benzyl alcohol (d)	0.69	0.001
	L-citruline	Amino acid (d)	0.59	0.008
	Vitamin B5	Vitamin	0.52	0.024
	Proline	Amino acid	0.50	0.030
	DG(34:1)	Glycerolipid	0.49	0.035
	4-Pyridoxic acid	Pyridine (d)	0.49	0.036
	PC(O-16:0/20:4)	Glycerophospholipid	0.47	0.042
	PE(18:0/18:1)	Glycerophospholipid	−0.47	0.046
	SM(d18:1/17:0)	Sphinomyelin	−0.48	0.038
	Stearoylcarnitine	Acyl caritine	−0.49	0.037
	PE(P-18:0/18:1)	Glycerophospholipid	−0.50	0.030
	PE(16:0/18:2)	Glycerophospholipid	−0.51	0.027
	PE(0:0/20:3)	Glycerophospholipid	−0.51	0.027
	PE(P-16:0/18:2)	Glycerophospholipid	−0.53	0.021
	Alpha-Ketoglutarate	Keto-acids (d)	−0.54	0.028
	PE(18:1/18:2)	Glycerophospholipid	−0.55	0.017
	PE(P-18:0/18:2)	Glycerophospholipid	−0.56	0.013
	AC(12:1n-x)	Fatty esters	−0.57	0.014
	PE(18:2/18:2)	Glycerophospholipid	−0.57	0.016
	SM(43:1)	Sphingomyelin	−067	0.003
Obese	L-Octanoylcarnitine	Acyl carnitine	0.94	0.017
	Arginine	Amino acid	0.86	0.024
	PC(O-16:0/18:2)	Glycerophospholipid	0.83	0.058
	Acylcarnitine(8:1n-x)	Acyl carnitine	0.75	0.066
	Deoxycholic acid	Bile acid	0.75	0.066
	PI(18:0/20:4)	Glycerophospholipid	−0.75	0.066
	PE(20:5/16:0)	Glycerophospholipid	−0.75	0.066
	L-Homoserine	Amino acid	−0.75	0.066
	Isoleucine	Amino acid	−0.75	0.066
	SM(43:1)	Sphingomyelin	−0.79	0.048
	SM(d18:1/24:1) + SM(d18:2/24:0)	Sphingomyelin	−0.79	0.048
	SM(d18:1/17:0)	Sphingomyelin	−0.79	0.048
	SM(33:1)	Sphingomyelin	−0.79	0.048
	PE(P-18:0/22:5) + PE(P-20:1/20:4)	Glycerophospholipid	−0.79	0.048
	PUFA (16:3n-x)	Fatty acid	−0.79	0.048
	PE(20:4/18:2)	Glycerophospholipid	−0.79	0.048
	SM(32:1)	Sphingomyelin	−0.82	0.034
	PE(P-16:0/20:4)	Glycerophospholipid	−0.82	0.034
	PE(0:0/22:4)	Glycerophospholipid	−0.82	0.034
	SM(d18:2/22:0)	Sphingomyelin	−0.86	0.024
	SM(d18:1/22:0)	Sphingomyelin	−0.86	0.024
	SM(d18:1/18:0)	Sphingomyelin	−0.86	0.024
	SM(d18:1/16:0)	Sphingomyelin	−0.86	0.024
	PE(18:0/20:4)	Glycerophospholipid	−0.86	0.024
	PE(18:1e/22:4)	Glycerophospholipid	−0.88	0.008
	SM(d18:2/20:0)	Sphingomyelin	−0.89	0.012
	PE(16:0/22:6)	Glycerophospholipid	−0.89	0.012
	SM(42:1)	Sphingomyelin	−0.93	0.007
	SM(d16:1/24:1)	Sphingomyelin	−0.93	0.007
	PE(16:0/20:4)	Glycerophospholipid	−0.93	0.007
	SM(38:1)	Sphingomyelin	−0.96	0.003

degradation could be altered in PCa tissue. Strikingly, we found that the expression steroid sulfatase (STS), which would remove the sulphate group in androsterone sulphate and DHEAS, was decreased in PCa, and this reduction was associated to metastatic disease and reduced disease-free survival in one out of two datasets (Figure 6(k–m).

Discussion

EVs are produced by normal and cancerous cells and harbour molecular features of their cells of origin [27]. This encapsulated material can exert biological and metabolic functions [4,14,15,28], which makes them entities of tremendous interest in cancer biology, both at the level of

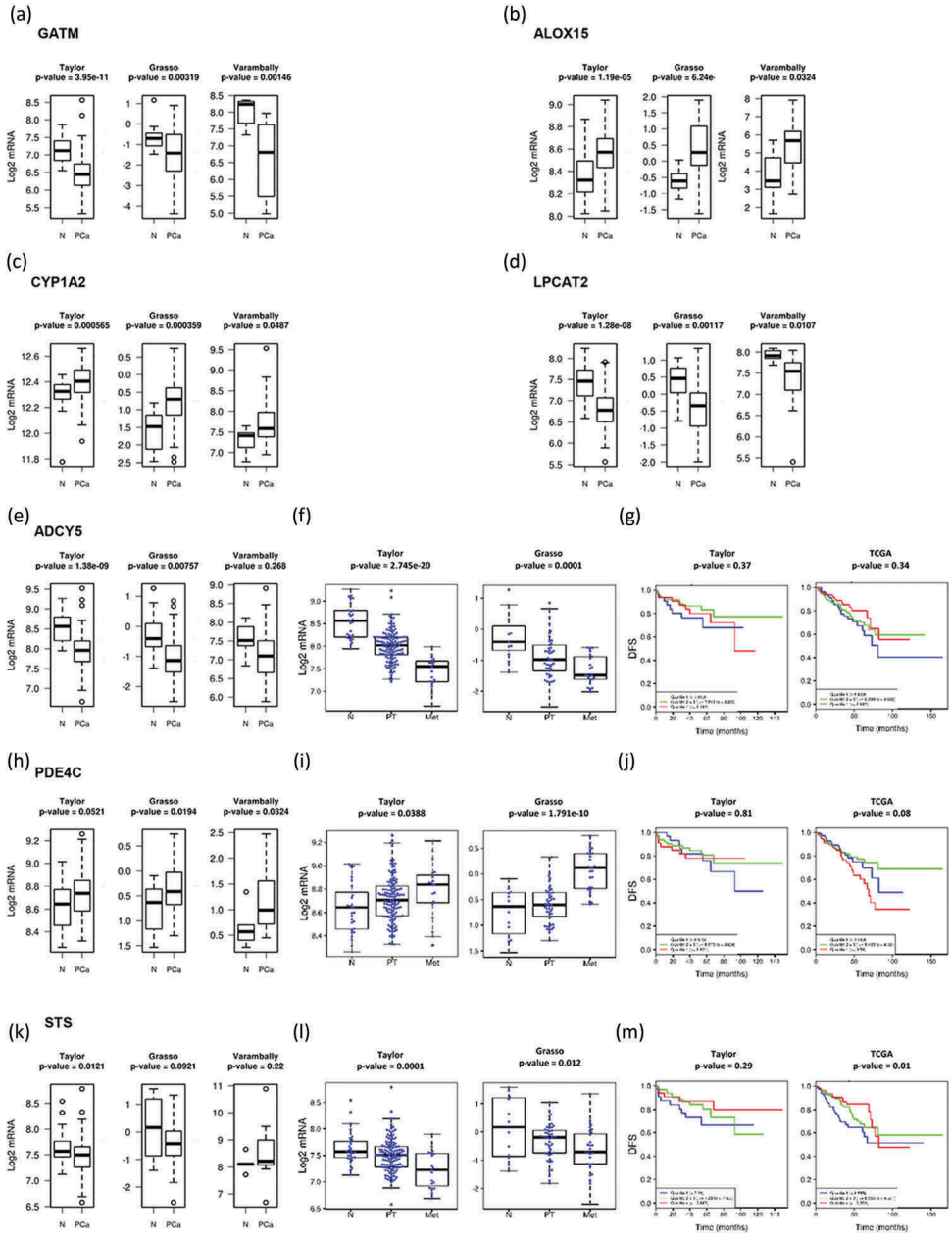


Figure 6. Gene-enrichment analysis.

In-silico transcriptomics analysis of enzymes directly involved in the metabolism of metabolites differentially expresed between PCa and BPH samples.

biomarker discovery and mechanistic. Urine contains EVs from different parts of the urinary track including kidney and bladder what has awaked great interest to identify biomarkers affecting these organs. In addition, the anatomic proximity of urine to the prostate gland and the already shown presence of tumour cells in the urine sediment [29,30] support also the development of potential non-invasive diagnoses of PCa using urine-based markers. In agreement with our previous results, we find differences in the size distribution of the urinary EVs between PCa and BPH [16], which we now report to be associated to disease stage (Figure 1). Our data show that urine from advanced PCa patients contains a higher proportion of large EVs than BPH patients. Given that our EV isolation procedure (filtration through 0.22 microns, and ultracentrifugation at 10,000 × g) removed most of the large EVs from the sample, and enrich in small EVs (mostly exosomes and small microvesicles), this difference could be underestimated in our samples. Importantly, in agreement with our result, it has already been reported that prostate cancer cells release large EVs named oncosomes with a size between 1 and 10 microns [31] that contain a distinct

protein cargo [32]. They have also been detected in circulation in models of PCa and shown that their abundance correlate with tumoral progression [31]. Although, our studies have been focused in the smaller EVs, it is interesting that we have also observed this size effect.

In a recent targeted lipidomics analysis of urinary EVs from healthy and PCa urine samples [13], the authors analysed 107 lipid species and found that 9 of them were significantly different between the two groups. Unlike this study, we have focused ours in the comparison between PCa and BPH, in an attempt to provide specific biomarkers to discriminate the two pathological conditions, and contribute to earlier diagnosis, and reduce secondary effects of unnecessary biopsies, so both studies can be considered complementary in terms of sample groups. Both studies are also complementary in the metabolites that they analyse because different metabolite extraction method and chromatographic procedures were used.

We report changes in the urine EV metabolome at both structural and cargo levels. The composition of the urine EVs analysed in this study varies in the abundance of phosphatidylcholine species that are major constituents of membranes. In particular we found reduced abundance of PCs in the EVs from PCa samples, in agreement with previously reported by Puhka and coworkers [33]. This result along with studies reporting increased abundance of PC in PCa tissue [34] could suggest that less PC-containing structures, like membrane vesicles are secreted to the extracellular environment. In addition, to the PC content, we found additional metabolites from different chemical nature differentially expressed in EVs from PCa and BPH samples that could be considered candidate biomarker for PCa including as candidate acyl carnitines, sphingomyelins, and steroids. Although more research is granted, our results indicate that a bias in EV size and membrane composition could harbour diagnostic potential in PCa.

Apart of the potential biomarker value of the identified metabolites, they are also valuable to indicate possible metabolic alterations occurring in PCa. We found reduced levels in PCa urine EVs of arachidonic acid, the precursor of eicosanoids and prostaglandins that are important proliferative and inflammatory modulators. Interestingly, it has been also reported that arachidonic acid level is lower in prostatic tissue from PCa patients [35]. In agreement with the reduction of the substrate arachidonic acid in PCa, it has been found that the level of their products (12-and 20-HETE, and PGE2) are higher in the tissue [36,37] and also in urine [38]. These studies along with many others have already shown that the metabolism of arachidonic acid and their products plays an important role in PCa development, and in fact, represents an important therapeutics target (reviewed in [39]). Importantly, our work suggests that the analysis of this metabolite in EVs isolated from urine samples may be used to evaluate in a non-invasive manner what is occurring in prostatic tissue itself in the context of PCa.

We observed changes in the abundance of metabolites that are carried within the EVs and are a potential cargo in PCa. It is worth mentioning that intermediary metabolites of androgen synthesis were among the most elevated in PCa urine EVs. Moreover, changes in the abundance of these steroids, together with cAMP, were significantly associated to perineural invasion. These results uncover the potential of unbiased urine EV analysis to elucidate novel signalling and metabolic alterations underlying PCa biology. Androgen signalling is among the predominant stimuli supporting PCa growth and the most successful therapeutic approaches have derived from its targeting [40], since prostate tumours frequently remain androgen dependent even at late-stage [41]. We have detected 3beta-hydroxyandors-5-en-17-one-3-sulphate (dehydroepiandrosterone sulphate, DHEAS) in urinary EVs, and its level was significantly elevated in PCa samples. This metabolite, along with estrone sulphate, is one of the main precursor for steroid hormones including androgens. There are many reports showing that steroid-related metabolites and enzymes are important modulators of PCa progression [42]. There are four different genes coding for enzymes that were related to this metabolite: STS, SULT1B1, SULT2B1 and SULT2A1. The fact that urine EVs from PCa patients contain androgen-related metabolites is suggestive of the relevance of this biosynthetic pathway in the disease and the potential role of EVs in providing androgen signalling to neighbour or distal cells. Indeed, expression of STS was reduced in PCa and associated to disease progression, hence providing a feasible explanation for the increase in sulfated steroids. Interestingly, urinary EVs could be used to monitoring androgen metabolism in a non-invasive manner.

Together with the aforementioned metabolites associated to perineural invasion, we also identified molecules that exhibited differential abundance in high grade tumours. Five metabolites were differentially abundant between pathological stage 2 and stage 3 PCa, and more than half of them were ceramide species. Ceramides are signalling molecules that can regulate various aspects of cancer cell biology, including proliferation, survival and cell death [43]. The selective decrease of ceramides in association with disease

aggressiveness provides an exciting perspective of how this family of metabolites could exert cell and non-cell autonomous functions to limit the progression of PCa.

It is worth noting that sarcosine has been proposed also as a PCa biomarker [11]. The urine level of this metabolite was increased in men with metastatic PCa [44]. However, its utility as a potential diagnostic tool is unclear, as its validation as a biomarker has failed in several studies (reviewed in [11,45]). Interestingly, we have detected sarcosine in urinary EVs, and although not significant ($p = 0.09$), its level was decreased in PCa samples.

Recent molecular and metabolic profiling of PCa also identifies lipid metabolism as a key pathway that undergoes metabolic reprogramming [46,47]. These changes include an upregulation metabolites involved in *de novo* lipid biosynthesis [48] and fatty acid β-oxidation [49]. As consequence, it has been shown the accumulation in the prostatic tissue of acyl carnitines, which are intermediates of fatty acid oxidation [50]. In agreement with this alteration, we found increased levels of acyl carnitines in the urinary EVs from PCa patients. This association of differential levels of carnitines on PCa EVs with a metabolic shifting towards β-oxidation of fatty acids has already been proposed by Puhka and coworkers [33].

In summary, in this work we report several metabolites associated to urinary EVs, many of them exhibiting differential abundance between BPH and PCa, and mirroring some of the alterations described in PCa.

Acknowledgments

We are thankful to the Basque Biobank for research (BIOEF), for the acquisition, maintenance and distribution of urine samples, and Dr. Sebastiaan van Liempd for is technical assistance in the treatment of the metabolomics data.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work of JF-P is supported by ISCIII [PI12/01604], Spanish Ministry of Economy and Competitiveness MINECO [SAF2015-66312] and GAP1 Movember Foundation. The work of A.C. is supported by the department of education of the Basque Government [IKERTALDE IT1106-16], the BBVA foundation, the MINECO [SAF2016-79381-R (FEDER/EU)]; European Research Council [Starting Grant 336343, PoC 754627]. The participation of A.C., A.R.C and V.T. as part of CIBERONC was co-funded with FEDER funds. V.T. is founded by Fundación Vasca de Innovación e Investigación Sanitarias, BIOEF [BIO15/CA/052], the AECC J.P. Bizkaia and the Basque Department of Health [2016111109]. We thank MINECO for the REDIEX (Spanish Excellence Network in Exosomes) and the Severo Ochoa Excellence Accreditation [SEV-2016-0644].

ORCID

Marc Clos-Garcia http://orcid.org/0000-0002-0208-1372
Pilar Sánchez-Mosquera http://orcid.org/0000-0003-2194-234X

References

[1]. Freedland SJ. Screening, risk assessment, and the approach to therapy in patients with prostate cancer. Cancer. 2011;117:1123–1135.
[2]. Fernie A, Trethewey R, Krotzky A, et al. Metabolite profiling: from diagnostics to systems biology. Nat Rev Mol Cell Biol. 2004;5:763–769.
[3]. Gonzalez E, Van Liempd S, Conde-Vancells J, et al. Serum UPLC-MS/MS metabolic profiling in an experimental model for acute-liver. Metabolomics. 2012;8:997–1011.
[4]. Royo F, Moreno L, Mleczko J, et al. Hepatocyte-secreted extracellular vesicles modify blood metabolome and endothelial function by an arginase-dependent mechanism. Sci Rep. 2017;7:42798.
[5]. Alonso C, Fernandez-Ramos D, Varela-Rey M, et al. Metabolomic identification of subtypes of nonalcoholic steatohepatitis. Gastroenterology. 2017;152:1449–61 e7.
[6]. Holmes E, Wijeyesekera A, Taylor-Robinson SD, et al. The promise of metabolic phenotyping in gastroenterology and hepatology. Nat Rev Gastroenterol Hepatol. 2015;12:458–471.
[7]. Griffin JL, Shockcor JP. Metabolic profiles of cancer cells. Nat Rev Cancer. 2004;4:551–561.
[8]. Pentyala S, Whyard T, Pentyala S, et al. Prostate cancer markers: an update (Review). Biomed Rep. 2016;263–268. DOI:10.3892/br.2016.586
[9]. Giskeødegård GF, Hansen AF, Bertilsson H, et al. Metabolic markers in blood can separate prostate cancer from benign prostatic hyperplasia. Br J Cancer. 2015;113:1712–1719.
[10]. Di Meo A, Bartlett J, Cheng Y, et al. Liquid biopsy: a step forward towards precision medicine in urologic malignancies. Mol Cancer. 2017;16:80.
[11]. Lima AR, Bastos Mde L, Carvalho M, et al. Biomarker discovery in human prostate cancer: an update in metabolomics studies. Transl Oncol. 2016;9:357–370.
[12]. Yanez-Mo M, Siljander PR, Andreu Z, et al. Biological properties of extracellular vesicles and their physiological functions. J Extracell Vesicles. 2015;4:27066.
[13]. Skotland T, Ekroos K, Kauhanen D, et al. Molecular lipid species in urinary exosomes as potential prostate cancer biomarkers. Eur J Cancer. 2017;70:122–132.
[14]. Iraci N, Gaude E, Leonardi T, et al. Extracellular vesicles are independent metabolic units with asparaginase activity. Nat Chem Biol. 2017;13:951–955.
[15]. Royo F, Palomo L, Mleczko J, et al. Metabolically active extracellular vesicles released from hepatocytes under

drug-induced liver-damaging conditions modify serum metabolome and might affect different pathophysiological processes. Eur J Pharm Sci. 2017;98:51–57.
[16]. Royo F, Zuñiga-Garcia P, Torrano V, et al. Transcriptomic profiling of urine extracellular vesicles reveals alterations of CDH3 in prostate cancer. Oncotarget. 2016;7:6835–6846.
[17]. Barr J, Caballeria J, Martinez-Arranz I, et al. Obesity-dependent metabolic signatures associated with nonalcoholic fatty liver disease progression. J Proteome Res. 2012;11:2521–2532.
[18]. Martinez-Una M, Varela-Rey M, Cano A, et al. Excess S-adenosylmethionine reroutes phosphatidylethanolamine towards phosphatidylcholine and triglyceride synthesis. Hepatology. 2013;58:1296–1305.
[19]. López-Ratón M, Rodríguez-Álvarez MX, Suárez CC, et al. OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. J Stat Softw. 2014;61:1–36.
[20]. Zabala-Letona A, Arruabarrena-Aristorena A, Martin-Martin N, et al. mTORC1-dependent AMD1 regulation sustains polyamine metabolism in prostate cancer. Nature. 2017;547:109–113.
[21]. Torrano V, Valcarcel-Jimenez L, Cortazar AR, et al. The metabolic co-regulator PGC1alpha suppresses prostate cancer metastasis. Nat Cell Biol. 2016;18:645–656.
[22]. Royo F, Zuniga-Garcia P, Sanchez-Mosquera P, et al. Different EV enrichment methods suitable for clinical settings yield different subpopulations of urinary extracellular vesicles from human samples. J Extracell Vesicles. 2016;5:29497.
[23]. Dell'Atti L. Prognostic significance of perineural invasion in patients who underwent radical prostatectomy for localized prostate cancer. J Buon. 2016;21:1219–1223.
[24]. Rodriguez C, Freedland SJ, Deka A, et al. Body mass index, weight change, and risk of prostate cancer in the cancer prevention study ii nutrition cohort. Cancer Epidemiol Biomarkers Prev. 2007;16:63–69.
[25]. Lee DJ, Mallin K, Graves AJ, et al. Recent changes in prostate cancer screening practices and epidemiology. J Urol. 2017;198:1230–1240.
[26]. Gao J, Tarcea VG, Karnovsky A, et al. Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. Bioinformatics. 2010;26:971–973.
[27]. Torrano V, Royo F, Peinado H, et al. Vesicle-MaNiA: extracellular vesicles in liquid biopsy and cancer. Curr Opin Pharmacol. 2016;29:47–53.
[28]. Peinado H, Aleckovic M, Lavotshkin S, et al. Melanoma exosomes educate bone marrow progenitor cells toward a pro-metastatic phenotype through MET. Nat Med. 2012;18:883–891.
[29]. Fujita K, Pavlovich CP, Netto GJ, et al. Specific detection of prostate cancer cells in urine by multiplex immunofluorescence cytology. Hum Pathol. 2009;40:924–933.
[30]. Dijkstra S, Birker IL, Smit FP, et al. Prostate cancer biomarker profiles in urinary sediments and exosomes. J Urol. 2014;191:1132–1138.
[31]. Di Vizio D, Morello M, Dudley AC, et al. Large oncosomes in human prostate cancer tissues and in the circulation of mice with metastatic disease. Am J Pathol. 2012;181:1573–1584.
[32]. Minciacchi VR, You S, Spinelli C, et al. Large oncosomes contain distinct protein cargo and represent a separate functional class of tumor-derived extracellular vesicles. Oncotarget. 2015;6:11327–11341.
[33]. Puhka M, Takatalo M, Nordberg ME, et al. Metabolomic profiling of extracellular vesicles and alternative normalization methods reveal enriched metabolites and strategies to study prostate cancer-related changes. Theranostics. 2017;7:3824–3841.
[34]. Li J, Ren S, Piao HL, et al. Integration of lipidomics and transcriptomics unravels aberrant lipid metabolism and defines cholesteryl oleate as potential biomarker of prostate cancer. Sci Rep. 2016;6:20984.
[35]. Faas FH, Dang AQ, White J, et al. Decreased prostatic arachidonic acid in human prostatic carcinoma. BJU Int. 2003;92:551–554.
[36]. Yang P, Cartwright CA, Li J, et al. Arachidonic acid metabolism in human prostate cancer. Int J Oncol. 2012;41:1495–1503.
[37]. Chaudry AA, Wahle KW, McClinton S, et al. Arachidonic acid metabolism in benign and malignant prostatic tissue in vitro: effects of fatty acids and cyclooxygenase inhibitors. Int J Cancer. 1994;57:176–180.
[38]. Nithipatikom K, Isbell MA, See WA, et al. Elevated 12- and 20-hydroxyeicosatetraenoic acid in urine of patients with prostatic diseases. Cancer Lett. 2006;233:219–225.
[39]. Nithipatikom K, Campbell WB. Roles of eicosanoids in prostate cancer. Future Lipidol. 2008;3:453–467.
[40]. Rodrigues DN, Boysen G, Sumanasuriya S, et al. The molecular underpinnings of prostate cancer: impacts on management and pathology practice. J Pathol. 2017;241:173–182.
[41]. De Bono JS, Logothetis CJ, Molina A, et al. Abiraterone and increased survival in metastatic prostate cancer. N Engl J Med. 2011;364:1995–2005.
[42]. Capper CP, Rae JM, Auchus RJ. The metabolism, analysis, and targeting of steroid hormones in breast and prostate cancer. Horm Cancer. 2016;7:149–164.
[43]. Saddoughi SA, Ogretmen B. Diverse functions of ceramide in cancer cell death and proliferation. Adv Cancer Res. 2013;117:37–58.
[44]. Sreekumar A, Poisson LM, Rajendiran TM, et al. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature. 2009;457:910–914.
[45]. Monteiro MS, Carvalho M, Bastos ML, et al. Metabolomics analysis for biomarker discovery: advances and challenges. Curr Med Chem. 2013;20:257–271.
[46]. Zadra G, Photopoulos C, Loda M. The fat side of prostate cancer. Biochim Biophys Acta. 2013;1831:1518–1532.
[47]. Deep G, Schlaepfer IR. Aberrant lipid metabolism promotes prostate cancer: role in cell survival under hypoxia and extracellular vesicles biogenesis. Int J Mol Sci. 2016;17. DOI:10.3390/ijms17071061
[48]. Rysman E, Brusselmans K, Scheys K, et al. De novo lipogenesis protects cancer cells from free radicals and chemotherapeutics by promoting membrane lipid saturation. Cancer Res. 2010;70:8117–8126.
[49]. Carracedo A, Cantley LC, Pandolfi PP. Cancer metabolism: fatty acid oxidation in the limelight. Nat Rev Cancer. 2013;13:227–232.
[50]. Al-Bakheit A, Traka M, Saha S, et al. Accumulation of palmitoylcarnitine and its effect on pro-inflammatory pathways and calcium influx in prostate cancer. Prostate. 2016;76:1326–1337.

cancers

MDPI

Article

Targeted UPLC-MS Metabolic Analysis of Human Faeces Reveals Novel Low-Invasive Candidate Markers for Colorectal Cancer

Joaquin Cubiella [1,†,*], Marc Clos-Garcia [2,3,†], Cristina Alonso [4], Ibon Martinez-Arranz [4], Miriam Perez-Cormenzana [4], Ziortza Barrenetxea [5], Jesus Berganza [5], Isabel Rodríguez-Llopis [5], Mauro D'Amato [6,7], Luis Bujanda [3,*], Marta Diaz-Ondina [1] and Juan M. Falcón-Pérez [2,7,8,*]

1 Department of Gastroenterology, Complexo Hospitalario Universitario de Ourense, Instituto de Investigación Biomédica Ourense-Vigo-Pontevedra, 32005 Ourense, Spain; m.ondina@hotmail.com
2 Exosomes Laboratory, CIC bioGUNE, CIBERehd, Bizkaia Technology Park, Derio, 48160 Bizkaia, Spain; mclos.biodonostia@cicbiogune.es
3 Department of Gastroenterology, Hospital Donostia/Instituto Biodonostia, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Universidad del País Vasco (UPV/EHU), 20014 San Sebastián, Spain
4 OWL Metabolomics, Bizkaia Technology Park, Derio, 48160 Bizkaia, Spain; calonso@owlmetabolomics.com (C.A.); imartinez@owlmetabolomics.com (I.M.-A.); mperez@owlmetabolomics.com (M.P.-C.)
5 GAIKER-IK4 Technology Centre, Ed. 202, 48170 Zamudio, Spain; barrenetxea@gaiker.es (Z.B.); berganza@gaiker.es (J.B.); rodriguez@gaiker.es (I.R.-L.)
6 Gastrointestinal Genetics Unit, Biodonostia HRI, 20014 San Sebastián, Spain; mauro.damato.mda@gmail.com
7 IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain
8 Metabolomics Platform, CIC bioGUNE, CIBERehd, Bizkaia Technology Park, Derio, 48160 Bizkaia, Spain
* Correspondence: Joaquin.Cubiella.Fernandez@sergas.es (J.C.); luis.bujandafernandezdepierola@osakidetza.eus (L.B.); jfalcon@cicbiogune.es (J.M.F.-P.)
† Joaquin Cubiella and Marc Clos-Garcia shared first authorship.

Received: 11 July 2018; Accepted: 28 August 2018; Published: 1 September 2018

Abstract: Low invasive tests with high sensitivity for colorectal cancer and advanced precancerous lesions will increase adherence rates, and improve clinical outcomes. We have performed an ultra-performance liquid chromatography/time-of-flight mass spectrometry (UPLC-(TOF) MS)-based metabolomics study to identify faecal biomarkers for the detection of patients with advanced neoplasia. A cohort of 80 patients with advanced neoplasia (40 advanced adenomas and 40 colorectal cancers) and 49 healthy subjects were analysed in the study. We evaluated the faecal levels of 105 metabolites including glycerolipids, glycerophospholipids, sterol lipids and sphingolipids. We found 18 metabolites that were significantly altered in patients with advanced neoplasia compared to controls. The combinations of seven metabolites including ChoE(18:1), ChoE(18:2), ChoE(20:4), PE(16:0/18:1), SM(d18:1/23:0), SM(42:3) and TG(54:1), discriminated advanced neoplasia patients from healthy controls. These seven metabolites were employed to construct a predictive model that provides an area under the curve (AUC) median value of 0.821. The inclusion of faecal haemoglobin concentration in the metabolomics signature improved the predictive model to an AUC of 0.885. In silico gene expression analysis of tumour tissue supports our results and puts the differentially expressed metabolites into biological context, showing that glycerolipids and sphingolipids metabolism and GPI-anchor biosynthesis pathways may play a role in tumour progression.

Keywords: colorectal cancer; metabolomics; faecal samples; biomarkers

1. Introduction

Colorectal cancer (CRC) is the second leading cause of cancer death in developed countries [1]. Although knowledge of the genetic- and diet-associated mechanisms involved in CRC establishment and progression is rapidly increasing [2], still the best prognosis is obtained when malignancy is detected early. CRC screening, which detects both precancerous polyps and CRC, can reduce both colorectal cancer incidence and mortality [3–7]. Through screening, the incidence of colorectal cancer can be reduced by 30% with a mortality reduction of 50% depending on the screening modality and the participation rates [7,8]. These data clearly support the strategy to have efficient and sensitive screening methods. Screening tests available include detecting haemoglobin or DNA mutations/alterations in feces [4,9], radiologic or endoscopic (flexible sigmoidoscopy, colonoscopy, and computed tomographic colonography) methods [10]. Each test has its own advantages, has demonstrated to be cost-effective, and has associated limitations and risks [10]. Although colonoscopy is considered the most accurate test for early detection and prevention of colorectal cancer [11], its applicability is limited due to the secondary effects associated with it (mild and severe), the low adherence in average and familial-risk populations and the limited resources available [12,13].

On the other hand, most of CRC are still diagnosed in symptomatic patients, even when CRC screening programs are established [14]. In this regard, symptoms and symptom-based prediction models have a limited accuracy for CRC detection in this population. CRC diagnostic biomarkers, such as faecal haemoglobin, can improve the diagnostic process either alone or within prediction models [15–17]. For all those reasons, the development of non-invasive methods to detect CRC either in asymptomatic and symptomatic patients is an area of interest for patients, clinicians and healthcare providers.

Metabolomics is the omics technology dedicated to the measurement of small molecules (<2000 Da) that are present in a biological system. Major advances and new development of analytical instruments, together with the implementation of bioinformatics tools for robust data analysis allows simultaneous measurement and analysis of a huge number of metabolites from a biological system [18–21]. In consequence, metabolomics has become one of the main technologies for biomarker identification and for unraveling pathophysiological mechanisms in many diseases, including cancer. The development of ultra-performance liquid chromatography (UPLC) has improved both resolution and sensitivity of metabolomics analysis. It has also allowed the rapid separation of metabolites when compared to conventional LC methods [22,23]. Notably, several metabolomics studies have been performed aiming to identify new CRC biomarkers, as reviewed by Zhang et al. [24]. For diagnostics purpose, several studies exist, although the majority of them have been performed on serum samples [25–33], tissue [34–36] and urine [37]. To our knowledge, only one study was found that studied metabolomics differences directly in human feces samples, like our project design, using NMR-based metabolomics [38]. Metabolomics study of faeces may be more effective in detecting novel colon cancer makers than other approaches because faeces are in close proximity to the colorectal mucosa and are a product of interactions between dietary components and the microbiota. This latter is affected by and seems to play an important role in the progression of colon cancer [39,40]. Existent literature has identified several metabolites, some being consistently altered in CRC individuals and others being increased in some studies and decreased in other ones [24]. These studies have allowed the identification of several altered metabolic pathways, including carbohydrate and amino acid metabolisms, and lipid-related metabolic pathways. Significantly, most of the studies found differences in metabolites of the tricarboxilic acid (TCA) cycle. Also, importantly, alterations on short-chain fatty acids (SCFAs) levels were found for feces-metabolomics study, which clearly indicates a role for the CRC-specific microbiota composition [38]. Lipid metabolism is an important pathway of cellular energy metabolism and its alteration has been related to CRC development and progression. Alterations on metabolic pathways for the eight distinct pathways of lipid metabolism, including corresponding genes and lipid-specific cell receptors, have been reviewed by Yan et al. 2016 [41].

In this study, we evaluate by UPLC-MS the levels of 105 metabolites in lyophilized faeces from a cohort of 129 samples including patients with advanced adenoma or colon carcinoma and healthy individuals. After applying univariate analysis, we found significant changes between healthy individuals and advanced neoplasia patients in 18 metabolites including sphingomyelins, ceramides, glycerophospholipis and cholesteryl esters. A combined analysis of ChoE(18:1), ChoE(18:2), ChoE(20:4), PE (16:0/18:1), SM(d18:1/23:0), SM(42:3) and TG(54:1) provides an AUC value of 0.821. This work supports the usefulness of metabolomics to develop low invasive diagnostic tools for colon cancer population screenings.

2. Results

For the study, we have analysed faecal samples collected from 49 healthy, 40 CRC patients and 40 AD patients (see Materials and Methods for more details). On these samples, we have performed a metabolomics profiling using the UPLC-MS approach as described in Materials and Methods. There is no single method to analyse the entire set of metabolites of a biological sample, mainly due to the wide concentration range of the metabolites joined to their extensive chemical diversity. For this study, we have employed an UPLC-MS method (Supplementary Figure S1) capable of detecting consistently the 105 identified metabolites listed in Supplementary Table S1, that includes fatty acyls, glycerolipids, glycerophospholipids, sterol lipids and sphingolipids.

2.1. Multivariate Analysis

First, we analysed the metabolomic profiling of the 105 metabolites by unsupervised principal component analysis (PCA). We did not find any clustering of samples according to their classification as cases (AD and CRC) and controls (C), as seen on the score plot in Figure 1; neither, did if each group (AD, CRC and C) was compared separately each other (Supplementary Figure S2).

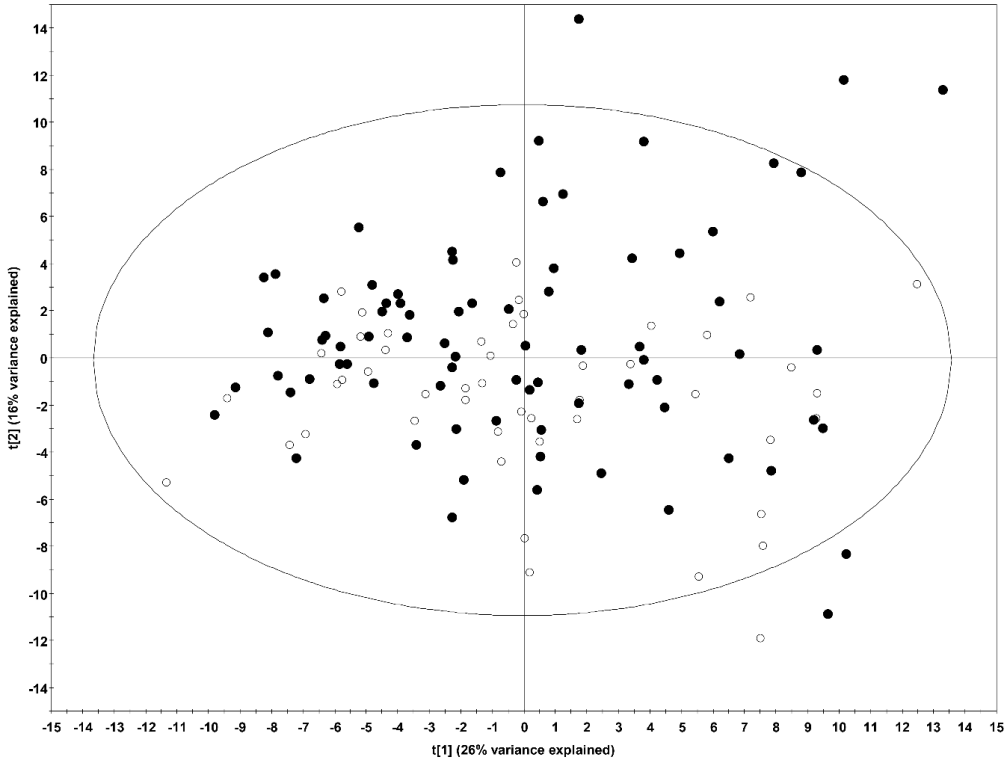


Figure 1. PCA scores plot of healthy individuals and patients with advanced neoplasia. (t[1]: R2X = 0.26 and Q2 = 0.22, t[2]: R2X = 0.16 and Q2 = 0.18): CRC and AD patients (n = 80), filled circles; healthy individuals (n = 49), open circles.

Neither the application of orthogonal (partial least squares) projections to latent structures (OPLS) or multivariate analysis was suitable for obtaining a separation between the groups of samples (data not shown). This lack of discrimination between groups through multivariate analysis highlights the expected high heterogeneity that exists between individuals.

2.2. Univariate Analysis

As it is complementary to the multivariate analysis, we have applied a univariate approach that has been shown to be an alternative for metabolomics data sets with elevated heterogeneity [26]. The comparison of the 105 metabolites between cases (AD plus CRC) versus control (C) samples, showed significant (adjusted p-value < 0.05) difference of the fold change for 18 of them as can be observed in the Volcano plot (Figure 2A). Differences were mostly seen in sphingolipid family (SM and Cer, but not CMH), but also included ChoE, PC, PE and TG metabolites. The most altered metabolite was Cer(42:3), and all metabolites were higher in the case group, except for two of them, Cer(d18:1/16:0) and TG(54:1), which were lower than the control group (Figure 2A). Other highly altered metabolites (log2 fold change < 1) were Cer(d18:1/24:1) + Cer(d18:2/24:0), PE(16:0/18:1), PE(16:0/18:2) and TG(54:1) (Figure 2A).

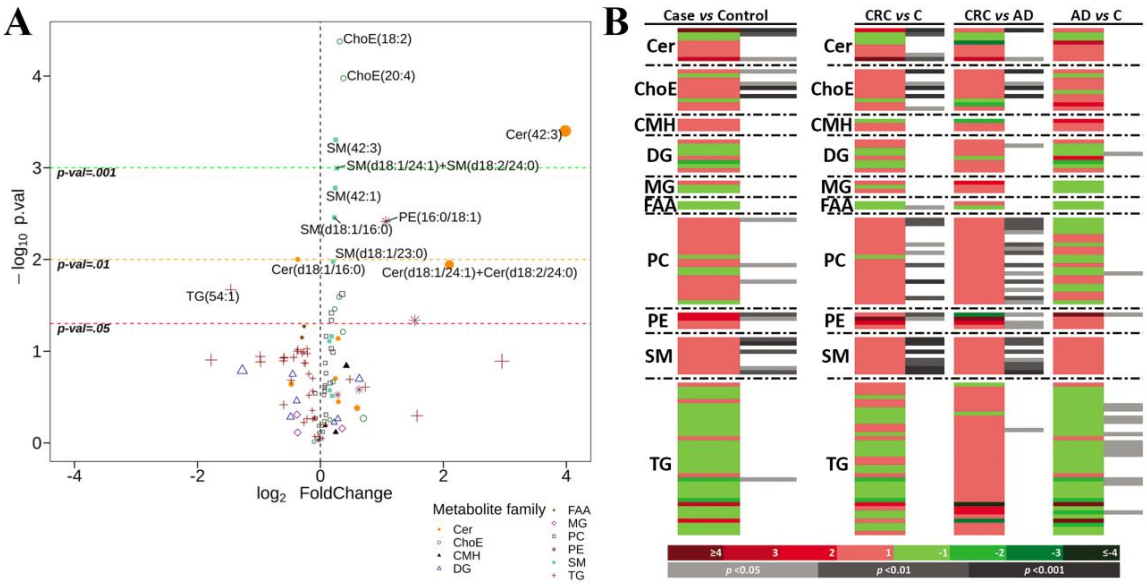


Figure 2. Volcano plot representation of metabolic changes in stools from control, CRC and AD sample groups. [log10 (p-value) vs. log2 (fold-change)] for the comparison between healthy individuals and patients with advanced neoplasia (CRC and AD). The shape and colour of the points indicates metabolite family, while the size is determined by the absolute value of the log2 Fold Change (**A**). Heatmap of metabolites altered in stools from control, CRC and AD sample groups (**B**).

Paired comparisons of sample groups revealed significant differences for some metabolic classes between CRC and AD, and also between CRC and C individuals (Table 1). Stool samples of patients with CRC had higher levels than AD or C samples of PC and also ChoE and SM metabolite classes. TG family showed the maximum differences when AD was compared to C samples, with alterations in 12 metabolites of the family; it was lower in AD than C. Actually, most of the differences between AD and C groups were found in this metabolite family, with only one metabolite altered for DG, PC and PE families. CMH and MG families did not show any difference in any comparison.

Ceramides, ChoE, PC and SM metabolite families were consistently increased in cancer samples. Only TG metabolites showed a specific trend for AD samples, being decreased with respect to the control samples, but showing no differences when comparing C versus CRC samples. Only PE family was consistently increased in both CRC and AD samples when compared to C group.

Table 1. Alteration in metabolic classes. Number of metabolites per metabolic classes differentially expressed in cases vs. control (C), CRC vs. AD, and CRC vs. control. Arrows indicate if metabolites are higher (↑), or lower (↓) in the Case, CRC or AD, depending on the comparison. In parentheses, the number of metabolites analyzed for each family is indicated.

	Case vs. Control	C vs. CRC	C vs. AD	AD vs. CRC
Cer (8)	2↑ 1↓	3↑ 1↓	0	2↑
ChoE (10)	4↑	5↑	0	4↑
CMH (3)	0	0	0	0
DG (8)	0	0	1↓	1↑
MG (3)	0	0	0	0
FAA (2)	0	1↓	0	0
PC (21)	3↑	7↑	1↓	13↑
PE (4)	2↑	2↑	1↑	3↑
SM (9)	5↑	7↑	0	7↑
TG (37)	1↓	0	12↓	1↑

The analysis of the individual metabolites also showed a difference between sample groups (Figure 2B). The heatmaps display the fold change of the 105 metabolites included in the analysis and their significances according to the Student's t-test for the comparisons performed between CRC and C, CRC and AD and between AD and C groups. In the comparison of case (AD plus CRC) versus C groups, significant metabolites were found mainly in Cer, ChoE, PE and SM families. While the ceramide family included both increased and decreased metabolites; only increased levels of metabolites belonging to ChoE, PC, PE and SM families were found in the case group.

The comparisons of CRC versus C, and CRC versus AD groups also revealed significant alteration of the levels of metabolites belonging to Cer, ChoE, PE and SM families, but in this case also the abundance of many metabolites belonging to the PC family were significantly altered. Most of the metabolites of these families were elevated in the CRC group in both comparisons. All these changes were not observed when comparing the AD and control groups indicating that those metabolites were mostly altered in the CRC group. Interestingly, a significant down-regulation of metabolites belonging to the TG family was observed mainly in the AD group (Figure 2B).

We also performed ANOVA test to detect significant differences in the metabolic profile between the three groups studies (C vs. AD vs. CRC). As a result, 29 differentially expressed metabolites belonging to Cer, ChoE, PC, PE and SM classes were found to be statistically significant in agreement with the previous paired analysis (Supplementary Table S2). Also, in concordance with the previous analysis, TG altered metabolites showed a specific pattern, being decreased in the AD group.

2.2.1. Predictive Models

In order to construct prediction models for cases (CRC and AD), the cohort was randomly separated in the training set containing 80% of the samples, and the validation set containing the remaining 20% of samples. To avoid possible bias derived from the data separation, we applied a bootstrap method, generating 10,000 different combinations of both training and validation datasets. By applying general linear models to the training set, we were able to find seven metabolites that when combined provide an AUC value of 0.821 (sensitivity 0.833 and specificity 0.800) (Figure 3). The metabolites were ChoE(18:1), ChoE(18:2), ChoE(20:4), PE(16:0/18:1), SM(d18:1/23:0), SM(42:3) and TG(54:1) and the model was:

$$\begin{aligned} Y = -5.308 - 1.92 \times \text{ChoE(18:2)} + 3.087 \times \text{ChoE(18:1)} - 1.564 \times \text{ChoE(20:4)} - 1.025 \times \text{PE(16:0/18:1)} \\ - 0.289 \times \text{SM(d18:1/23:0)} - 0.678 \times \text{SM(42:3)} + 0.386 \times \text{TG(54:1)} \end{aligned}$$

We computed also the potential effects of age and sex upon the performance of our model. We were able to slightly increase the predictive ability of the model when adding the age (AUC = 0.838),

sex (AUC = 0.837) and the combination of both (AUC = 0.848) features to the model (Figure 3C). When combining our metabolite model with faecal occult blood (FOB) parameter we were able to increase the AUC value up to 0.885.

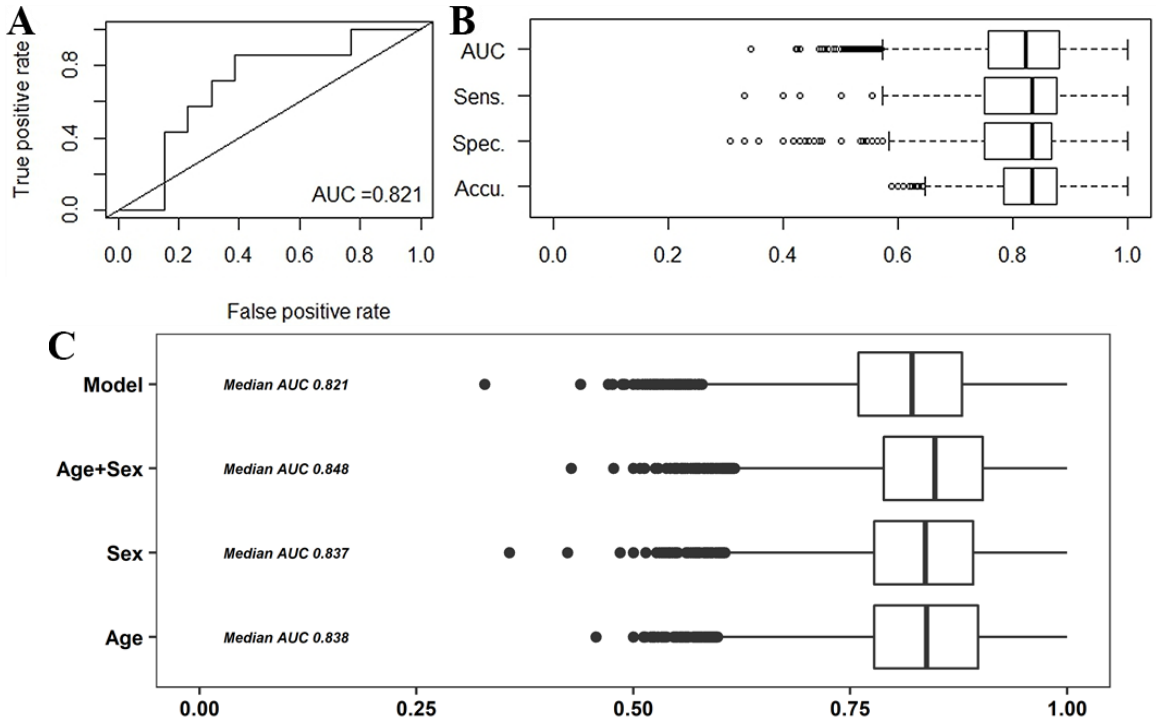


Figure 3. ROC curve of the predictive model constructed with the seven specified metabolites, including the value of the median AUC (**A**). Distribution of the model's features (AUC, sensitivity, specificity and accuracy) obtained from the 10,000 iterations done (**B**). Distribution of AUC measurements for the combination of our model with age, sex and the age + sex combination (**C**).

2.2.2. Correlation of the Metabolites with Clinical Parameters

A number of clinical parameters were available for the 129 samples analysed in this study including age, gender, FOB test (cut-off 100 ng/mL), carcinoembryonic antigen (CEA) test and COLONPREDICT index. COLONPREDICT is a CRC prediction model that takes into account demographic, symptoms, laboratory and anorectal examination results applicable both in primary and secondary healthcare units [16]. Thus, we evaluated if any of the 105 metabolites analysed in faecal samples correlated with any of the clinical parameters (Supplementary Table S3). There was not strong correlation with age, neither with CEA nor COLONPREDICT or gender, and there were only minor correlations with some clinical data as follows. Several TG metabolites correlated inversely with age data. Also, some metabolites belonging to the DG family correlate with age data, in the same direction as the TG metabolites. COLONPREDICT test showed the highest degree of correlation with metabolites of different families including CMH, PC, ChoE, PE, and SM. Although only slightly, ChoE(18:2) correlated directly with the FOB parameter (Supplementary Table S3).

We also studied how clinical parameters classified samples between the three groups (C, CRC and AD) and between two groups (C and Case) (Supplementary Figure S3). Both ANOVA test for the classification into three groups (Table 2) and Tukey's HSD test for the classification into two groups (Table 2) showed that COLONPREDICT was the best index to discriminate between samples, followed by FOB. We could see that gender had nearly no differences upon the discrimination between groups, compared to all other clinical parameters. It is important to note that no clinical parameter was able to significantly differentiate between C and AD sample groups.

Table 2. Differences between sample classification of several clinical parameters, either for the groups comparison (C, AD and CRC) and for the pairwise comparison (Control vs. Case). ANOVA test has been used for the study of differences between the three groups classification (C, AD and CRC) and Tukey's HSD test was used to analyse pairwise classifications (C vs. AD, C vs. CRC and AD vs. CRC). Tukey's HSD column depicts those pairwise combinations (of the three tested combinations) that showed to be significantly different. Avg. stands for average.

C, AD and CRC	Avg_C	Avg_{AD}	Avg_{CRC}	*p*-Value	Tukey's HSD
Gender	35.4% men	56.4% men	60% men	0.042	*NA*
Age	62.52	68.64	73.50	0.0003	CRC vs. C
FOB *	0	49	873	1.6×10^{-9}	CRC vs. C CRC vs. AD
CEA	1.90	1.72	14.85	0.00546	CRC vs. C CRC vs. AD
COLONPREDICT	0.048	0.104	0.470	$<2 \times 10^{-16}$	CRC vs. C CRC vs. AD
Control vs. Case	$Avg_{CONTROL}$	Avg_{CASE}	*p*-Value		
Gender	35.4% men	58.3% men	0.013		
Age	62.52	71.10	0.00083		
FOB *	0	336	7.09×10^{-8}		
CEA	1.900	8.367	0.0036		
COLONPREDICT	0.0477	0.289	1.231×10^{-10}		

* For FOB index, median values are given instead of mean, due to the non-normal distribution of the measurements.

2.2.3. Gene Expression Analysis of Enzymes Involved in the Metabolism of Altered Metabolites

Metabolites that were differentially expressed between case and control samples (Figure 2A), and with a KEGG or HMDB code already defined, were employed to identify possible metabolic pathways altered in colorectal cancer. By using the differentially expressed metabolites, we could in-silico identify 211 gene-encoding proteins that mainly clustered in three different metabolic pathways (Figure 4A). The identified pathways were glycerophospholipids metabolism, sphingolipids metabolism and the glycosylphosphatidylinositol (GPI)-anchor biosynthesis pathway suggesting that these pathways could be altered in colorectal cancer (Supplementary Figure S4). We analysed the expression levels of these gene-encoding proteins in the available gene-expression dataset of biopsies of colorectal cancer and normal mucosae of the colon [42]. We have observed that 15 of them showed a significantly different fold change between control and cancer (case) samples (Figure 4B). We have also observed a downregulation of CERS4, SMPD1 and SMPD3 (Figure 4B), which are responsible for the transformation of sphingosines and sphingomyelins to ceramides. We also observed downregulation of genes that encoded enzymes that catalyse the degradation of phosphocholine into choline metabolite, mainly from the phospholipase D (PLD) family: PLB1, PLD1, PNPLA7, PLA2G12B, PLA2G4C (Figure 4B). Furthermore, there was a significant downregulation of the genes PIGK and PIGZ, which encode enzymes involved in GPI-anchor biosynthesis. In addition, an upregulation of the genes LPCAT1 and LCAT (Figure 4B) that encode enzymes involved in the synthesis of phosphatidylcholine and cholesteryl esters, respectively, was also observed. Together, all these alterations on genes involved in lipid metabolism of the tumoral tissue support the lipid changes detected in the faecal samples.

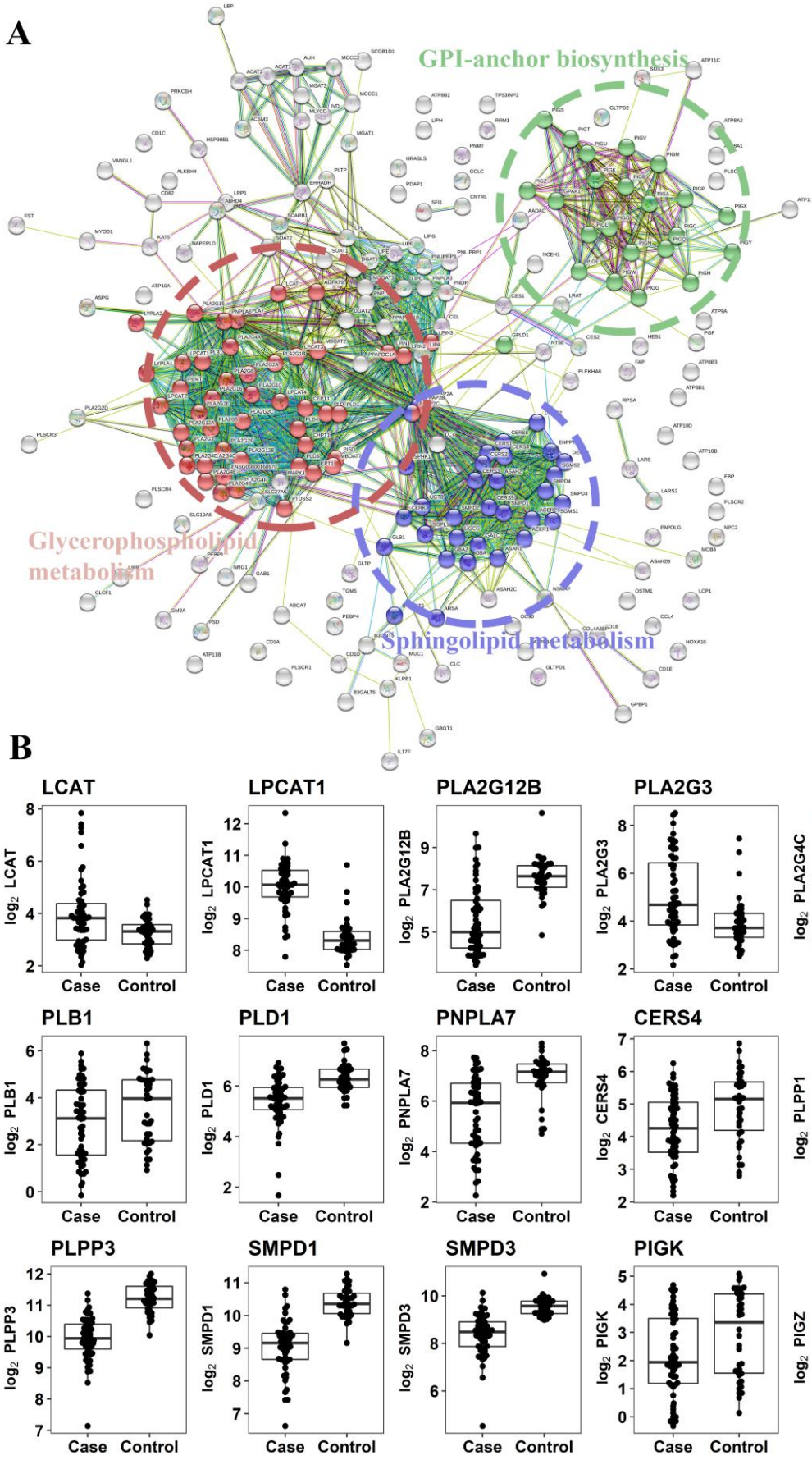


Figure 4. Gene networks of enzymes related with metabolism of stool CRC-altered lipids. Three major

pathways could be observed: Sphingolipid and glycerophospholipid metabolisms, and GPI-anchor biosynthesis (**A**). Gene expression in silico analysis of CRC tumoral tissue. The expression of gene-encoding enzymes involved in the metabolism of stool-altered lipids was analysed in publicly available GEO dataset GSE37364 that compared tumoral versus healthy tissue of the same individual. All displayed genes were highly significant (p-value < 0.001) except PLPP1 (p-value = 0.05) and PIGK (p-value = 0.02) (**B**).

3. Discussion

CRC screening with faecal occult blood (FOB) test has demonstrated efficacy in randomized trials. Nonetheless, the low sensitivity for advanced neoplasia of the test suggests the need for more accurate alternative diagnostic tests. In the present study, we have performed an UPLC-based targeted metabolomics analysis of stool to detect candidate endogenous metabolites suitable for the assessment of colon cancer using minimally invasive techniques. Metabolomic study of faeces can be more effective, because faeces are in close proximity to the colorectal mucosa. To date, metabolomics analyses of faecal samples have mostly been restricted to experimental studies in animal and small cross-sectional studies in humans [42–52]. While GC/MS-based metabolic profiling of faecal water has been reported [53–55], there exists only limited studies on the profiling and identification of metabolites within the complete faecal material; notably, lyophilized human faeces where its metabotype was confirmed to be more comprehensive than faecal water [47]. Previously, Ponnusamy et al. [56] profiled whole faeces from irritable bowel syndrome using GC/MS and identified several metabolites as candidate biomarkers for the disease. In the current work, a semi-quantitative analysis of 105 metabolites reveals significant differences in the faecal composition of cancer samples in the following lipids: PC(16:0/16:0), PC(32:1), PC(O-16:0/16:0), PE(16:0/18:1), PE(16:0/18:2), SM(d18:1/16:0), SM(d18:1/23:0), SM(d18:2/24:1) + SM(d18:1/24:0), SM(42:1), Cer(d18:1/16:0), Cer(d18:1/24:1) + Cer(d18:2/24:0), Cer(42:1), SM(42:3), ChoE(16:0), ChoE(18:1), ChoE(18:2), ChoE(20:4), TG(54:1). These lipid alterations detected in stools were supported by the gene expression profile observed in tumoral tissues showing deregulation of enzymes involved in glycerophospholipids and the glycosphingolipids metabolisms (Figure 4B). Some of the genes were of special interest as they serve as union nexuses of different metabolic pathways. Thus, PLPP1 and PLPP3 genes encoded lipid phosphate phosphatases (LPPs) with broad substrate specificity that dephosphorylate lipid substrates including phosphatidic acid, lysophosphatidic acid, ceramide 1-phosphate, sphingosine 1-phosphate, and diacylglycerol pyrophosphate [57]. One of their enzymatic reactions is the conversion of phosphatidic acid to diacylglycerol which is a central lipid for glycerophospholipids, triacylglycerols and sphingolipid metabolisms. In consequence, they modulate different signalling pathways and generate building blocks for lipid metabolism-regulating physiological and pathological processes including vascular function and tumor progression [58]. These also indicate that the altered metabolism of the tumour could be detected in stools, and consequently be detected in a non-invasive manner.

In our study, the most significant lipids altered in stool were cholesteryl esters, particularly ChoE(18:2) and ChoE(20:4) that were increased in CRC samples. This was in agreement with the fact that acetate—a short chain fatty acid—which is the precursor molecule for endogenous cholesterol production, has been reported to be elevated in CRC [59]. In addition, our in silico analysis of the gene expression profile of tumoral tissue reported by Valcz et al. [42] shows increased tumoral levels of the gene encoding the enzyme phosphatidylcholine-sterol acyltransferase responsible for the cholesteryl ester synthesis. Together, the data suggest that the levels of cholesteryl esters in stools can be a suitable non-invasive measurement to detect and follow up colorectal cancer. Based on the cholesteryl esters ChoE(18:1), ChoE(18:2) and ChoE(20:4), and complemented by PE(16:0/18:1), SM(d18:1/23:0), SM(42:3) and TG(54:1), we have built a robust stool metabolomic signature with an AUC value of 0.821 (sensitivity 0.833 and specificity 0.800). In our set of samples, the AUC of the FOB was 0.744, showing that our model of 7-metabolites performed better than the FOB in the detection of CCR.

Interestingly, the combination of FOB with our 7-metabolites of our metabolomics model increases the discriminating ability as judged by the AUC value that passed from 0.821 to 0.885.

It is important to highlight that one of the strengths of our study includes careful processing and preservation of the faecal specimens, and our quantification of within-subject intraclass correlation coefficient (ICC), from which we could estimate statistical power with our cutting-edge faecal metabolomics platform. Our platform has high sensitivity and technical reproducibility, but it has limited ability to detect some volatile and larger molecules.

Our study's major limitations are its small size and cross-sectional, hospital-based case–control design. It provided no assessment of temporality and could only detect strong associations with CRC. Also, the fact that this is a targeted metabolomics obviously biases the results towards lipid species, which is also an important limitation. As we mentioned in the introduction section, lipid alterations have been previously associated with CRC development and progression [41]. We considered, therefore, that our panel of metabolites would be sufficient to find potential CRC biomarkers. Also, keeping in mind the diagnostics aim of this study, we decided to use targeted metabolomics because it's cheaper than an untargeted one, making it a more affordable option. Targeted metabolomics allows an easier interpretation of results and, therefore, an easier translation to clinical practice, which we also considered to be an important point for the diagnostics purpose. As no restriction on diet was provided to the participants in the study, another limitation is the lack of control for potential diet-confounding factors. Nevertheless, we believe this potential diet's effects to be minimal, as all participants came from two Spanish regions that share the same dietary patterns. We did not specifically control for age, sex, tumour position and staging for this study, which constitutes another important limitation. The decision of not to control for those factors was done taking into account the sample size, not big enough to generate sufficiently big subgroups to obtain statistically robust data. In order to minimize those variables effects, we incorporated the 10,000 iterations through random subsetting of the population for the modelization step, thus generating 10,000 different populations, covering a huge range of different composition trains and test subpopulations that could reduce the potential bias towards some of the mentioned factors. Another strength of our study is the comparison against the FOB test and other clinical parameters. For every one of these comparisons, our model composed by the 7-metabolites performed better than the clinical parameters alone. Also, the integration of gene expression data in the study supports the identification of differentially expressed metabolites and puts them into context, providing some insights on how and why the levels are different between healthy controls and cancer patients.

4. Materials and Methods

4.1. Chemicals

HPLC-MS grade solvents were purchase from Sigma Aldrich (St. Louis, MO, USA). Reference metabolite standard compounds were obtained from Sigma Aldrich, Larodan Fine Chemicals (Malmö, Sweden) and Avanti Polar Lipids (Alabaster, AL, USA).

4.2. Clinical Samples and Study Population

The samples were collected during COLONPREDICT study, a multicentre, cross-sectional, blinded study of diagnostic tests aimed to create and validate a CRC prediction index in symptomatic patients based on available biomarkers, clinical and demographical data [16]. The study was approved by the Clinical Research Ethics Committee of Galicia (Code 2011/038). As the samples were collected from the COLONPREDICT study, the population selection characteristics were the same of that study. The cohort consisted of consecutive patients with gastrointestinal symptoms referred for colonoscopy from primary and secondary health care to Complexo Hospitalario Universitario de Ourense, Spain. Exclusion criteria for the COLONPREDICT study were: age under 18, pregnancy, asymptomatic individuals undergoing colonoscopy for CRC screening, patients with previous history

of colonic disease, patients requiring hospital admission, patients whose symptoms had ceased within 3 months of evaluation, and patients who declined to participate after reading the informed consent form. Patients self-collected a faecal sample from one bowel movement without specific diet or medication restrictions the week before a colonoscopy was performed at home and delivered to the hospital. The faecal sample was brought to the laboratory in less than 4 hours, split in aliquots and immediately frozen at −80 °C. We selected samples from 40 patients with advanced adenoma-AD- (≥10 mm, villous histology, high-grade dysplasia), 40with CRC and 49 with a normal colonoscopy. The characteristics of the patients differed with respect to age (CRC = 73.1 ± 10.6 years, AD = 68.8 ± 44.6 years, normal = 61.5 ± 14.4 years; $p < 0.001$) and sex (CRC = 60.0% male, AD = 59.1% male, normal = 27.5% male; $p = 0.004$). The CRC were located in the rectum (32.5%), colon distal to splenic flexure (45%) and proximal to splenic flexure (22.5%). The tumour stage at diagnosis was: I (24.2%), II (30.3%), III (30.3%) and IV (15.2%).

4.3. Sample Preparation and UPLC®-MS Metabolomics Analysis

A UPLC−time-of-flight (TOF)-MS-based platform was used to analyze chloroform/methanol extracts, including glycerolipids, cholesteryl esters, sphingolipids, primary fatty amides and glycerophospholipids among the identified ion features. The metabolite extraction procedure was as follows. Stools were lyophilized during 3 days by using the instrument Telstar LyoQuest −85. Afterward, 15 milligrams of lyophilized stool samples were mixed with 45 µL sodium chloride (50 mM) and 450 µL chloroform/methanol (30:1) in 1.5 mL microtubes at room temperature. The extraction solvent was spiked with compounds not detected in unspiked human stool samples [SM(d18:1/16:0), PE(17:0/17:0), PC(19:0/19:0), TAG(13:0/13:0/13:0), Cer(d18:1/17:0) and ChoE(12:0)]. After brief vortex mixing, the samples were incubated for 1 hour at −20 °C. After centrifugation at 16,000 × g for 15 min, 35 µL of the lower organic phase was collected and the solvent was removed. The dried extracts were then reconstituted in 1000 µL acetronitrile/isopropanol (1:1), centrifuged (16,000 × g for 5 min), and transferred to vials for UPLC®-MS analysis on an Acquity-Xevo G2 QTof system (Waters Corp., Milford, MA, USA). Samples were randomly divided into three batches, which contained a maximum of 78 samples. Chromatographic method and mass spectrometric detection conditions were described by Barr et al. [60]. Of the different platforms described, the one corresponding to ours was Platform 3.

4.4. Data Pre-Processing

Data pre-processing was processed using the TargetLynx application manager for MassLynx 4.1 (Waters Corp). A total of 105 UPLC-MS features were analysed, all of them identified prior to the analysis. Peak detection and noise reduction were performed as previously described [61,62]. Intra- and inter-batch normalization process was based on multiple internal standards and the pool calibration samples approach described by Martinez-Arranz et al. [62].

4.5. Data Analysis

The biomarker assessment in this study was organized in sequential and consecutive phases for discovery and biological validation. Firstly, 133 metabolites including glycerolipids, glycerophospholipids, sterol lipids and sphingolipids were selected as candidate biomarkers for initial analysis faeces samples from advanced neoplasia cases, colorectal cancer and cancer-free controls (Discovery Phase). Secondly, the potential clinical use of the most promising validated candidates was tested in faeces samples from colon cancer cases, a small set of adenomas, and cancer-free controls. Reported STARD guidelines have been the basis for defining our protocol.

Metabolites with less than 70% of the values present were removed from the analysis (remaining 105 metabolites into the analysis). Remaining missing values were imputed metabolite by metabolite, taking the minimal value for the metabolite and dividing it by 10. Data was then normalized with the log10 transformation. Univariate statistical analyses were also performed calculating group percentage changes and the analysis of variance (ANOVA) for the comparison among the different groups: CRC,

AD and control (C). Student's t-test *p*-values were calculated for the comparison between cases (AD and CRC) and C groups, as well as for the comparisons CRC and C, CRC and AD and between AD and C groups. Multivariate analyses were also performed, including both Principal Component Analysis (PCA) and Partial Least Squares Discriminant Analysis one (PLS-DA). ANOVA tests and Tukey's HSD tests were also calculated for several clinical parameters (FOB, sex, age, CEA and COLONPREDICT test) to determine its effectiveness to classify our samples into categories (CRC, AD and C or Case-Control). All *p*-values were adjusted with Bonferroni methodology unless otherwise stated.

A logistic regression (LR) was performed to identify a predictive signature capable of distinguishing between cases and control groups. LR is a commonly used technique for data classification. We first analysed the correlations between metabolites, establishing a cut-off at ρ 0.75. For each pair of correlated metabolites, we removed the one that separated the worst out of the two groups. A forward stepwise method was selected as variable selection approach, where the analysis started with an empty model and variables were added one at a time as long as these additions are worthy, by measuring the Area Under the Curve (AUC) value. This process finished when no more variables could be added. All samples were randomly divided into estimation (80% of all subjects; n = 101) and validation (20% of all subjects; n = 26) groups, both cohorts having an equal proportional representation of individuals belonging to cases and control groups. Ten-thousand iterations of both subsetting into estimation and validation groups and model constructing were generated, to avoid population-based biases. Receiver operating characteristic (ROC) curve analysis was used to assess its discriminatory power. Overall diagnostic accuracy for a given two-class comparison was given by the area under the ROC curve (AUC) with its associated standard error. Sensitivity, specificity and accuracy values were calculated.

All calculations were performed using statistical software package R v.3.1.1 (R Development Core Team, 2011; http://cran.r-project.org) with caret, caTools and receiver operating characteristic R (ROCR) packages to produce ROC curves and AUC estimate; MASS package was used to generate the LR. Additionally, SIMCA-P+ 12.0.1 (Umetrics AB, Umeå, Sweden) was used for PCA and PLS-DA multivariate data analysis.

Retrieval of genes and enzymes related with differentially expressed metabolites found in the study was done with custom Python scripts, which takes advantage of the published Python packages Biopython [63] and bioservices [64], which were used to access both HMDB and KEGG databases. These custom scripts retrieve information on the metabolite entries on both HMDB and KEGG databases regarding the enzymes involved in the metabolism of cited metabolites, as in which pathways are they present. We identified gene-encoding proteins involved in the metabolism of the seven metabolites of the predictive model, and we uploaded those genes to the STRING database [65], in order to identify the interaction between them, any potential clusterization and possible affected metabolic pathways. Genetic expression was obtained from publicly available GEO databaset GSE37364 [42]. The datasets were uploaded to R and the expression of selected genes was plotted into boxplots. Mapping of both metabolites and genes into metabolic pathways was done with pathview package [66] and custom R scripts.

5. Conclusions

This study highlights the power of UPLC-MS-based metabolomics approach in the discovery of novel non-invasive markers for colorectal cancer. With this study, we identified alterations in two main metabolic pathways, the glycerophospholipids and glycosphingolipids metabolisms. We found 18 metabolites differentially expressed between case samples (CRC + AD) and healthy controls, being mainly increased in case ones. We also showed how a discrimination model based only on metabolite species was able to differentiate between case (CRC+AD) samples and healthy ones and is better than those used nowadays, based in several clinical parameters like FOB, CEA, etc. The model generated included these metabolites: ChoE(18:1), ChoE(18:2), ChoE(20:4), PE (16:0/18:1), SM(d18:1/23:0), SM(42:3) and TG(54:1). Finally, we showed how the integration of different omics technologies

might be useful for supporting findings of one of them and to gain insights on how to explain the results obtained.

Supplementary Materials: The following are available online at http://www.mdpi.com/2072-6694/10/9/300/s1, Supplementary Figure S1, Workflow of the UPLC-MS-based targeted metabolomic profiling. Supplementary Figure S2, Multivariate analysis of paired group. (A) CRC vs. AD: R2X = 0.29 and Q2 = 0.24 t[2]: R2X = 0.19 and Q2 = 0.22). Black CRC, grey AD. (B) CRC vs. control: R2X = 0.30 and Q2 = 0.25, t[2]: R2X = 0.19 and Q2 = 0.24). Black CRC, white healthy. (C) AD vs. control: R2X = 0.28 and Q2 = 0.24, t[2]: R2X = 0.15 and Q2 = 0.15. Grey AD, white healthy. Supplementary Figure S3, Boxplot representation of the clinical parameters distribution on the distinct groups of samples (C, AD and CRC). Supplementary Figure S4, Mapping of altered genes and metabolites into the three metabolic pathways identified: sphingolipid metabolism (A), glycerophospholipid metabolism (B) and glycosylphosphatidylinositol (GPI)-anchor biosynthesis (C). Genes detected are coloured in a range green-red, depending on the Fold Change and metabolites in a range blue-yellow. Supplementary Table S1, List of the 105 metabolites analysed in the study. Supplementary Table S2, Metabolites differentially expressed between control, AD and CRC groups (ANOVA test). Supplementary Table S3, Clinical correlations between metabolites included in the study and the following parameters: FOB, Sex, Age, COLONPREDICT and CEA. Supplementary Table S4, Number of missing values obtained for each metabolite.

Author Contributions: Conceptualization, J.M.F.-P., J.C., L.B. and M.C.G.; Methodology, M.C.G.; Software, M.C.G. and I.M.-A.; Validation, M.C.G., C.A., I.M.-A. and J.M.F.-P.; Formal Analysis, M.C.G.; Investigation, C.A., I.M.-A., M.P.-C., Z.B., J.B., I.R.-L., M.D., M.D.-O.; Resources, J.M.F.-P., C.A.; Data Curation, M.C.G.; Writing-Original Draft Preparation, M.C.G., J.M.F.-P.; Writing-Review & Editing, J.M.F.-P., M.C.G., J.C., L.B.; Visualization, M.C.G.; Supervision, J.M.F.-P.; Project Administration, J.M.F.-P., L.B., J.C.; Funding Acquisition, J.M.F.-P., L.B., J.C.

Funding: Centro de Investigación Biomédica en Red en el Área temática de Enfermedades Hepáticas y Digestivas (CIBERehd) is funded by the Institute of Health Carlos III. This work has been supported by Instituto de Salud Carlos III (PI12/01604 to JMF-P) and BG2016-INVESTIGACION COLABORATIVA EN MEDICINA DE PRECISION Y BIOMARCADORES (Ref. KK-2016/00026) funded by Basque Government. All of them co-financed by ERDF (FEDER) Funds from the European Commission, "A way of making Europe".

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Abbreviations

AUC	Area Under the Curve
CEA	Carcinoembryonic Antigen
Cer	Ceramides
ChoE	Cholesteryl esters
CMH	Monohexosylceramides
DAG	Diacylglycerides
DAPC	Diacylglycerophosphocholines
FAA	Fatty acid amides (Primary Fatty Amides)
FOB	Faecal Occult Blood
MAG	Monoacylglycerides
MEMAPC	1-ether, 2-acylglycerophosphocholines
PC	Phosphatidylcholines
PCA	Principal Component Analysis
PE	Phosphatidylethanolamines
PI	Phosphatidylinositols
PLS-DA	Partial Least Square Discriminant Analysis
SM	Sphingomyelins
TAG	Triacylglycerides
UPLC®-MS	Ultra performance liquid chromatography-mass spectrometry

References

1. Ferlay, J.; Soerjomataram, I.; Dikshit, R.; Eser, S.; Mathers, C.; Rebelo, M.; Parkin, D.M.; Forman, D.; Bray, F. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **2015**, *136*, E359–E386. [CrossRef] [PubMed]
2. Vogelstein, B.; Papadopoulos, N.; Velculescu, V.E.; Zhou, S.; Diaz, L.A., Jr.; Kinzler, K.W. Cancer Genome Landscapes. *Science* **2013**, *339*, 1546–1558. [CrossRef] [PubMed]
3. Zauber, A.G.; Winawer, S.J.; O'Brien, M.J.; Lansdrop-Vogelaar, I.; van Ballegooijen, M.; Hankey, B.F.; Shi, W.; Bond, J.H.; Schapiro, M.; Panish, J.F.; et al. Colonoscopic Polypectomy and Long-Term Prevention of Colorectal-Cancer Deaths. *N. Engl. J. Med.* **2012**, *366*. [CrossRef] [PubMed]
4. Quintero, E.; Castells, A.; Bujanda, L.; Cubiella, J.; Salas, D.; Lanas, Á.; Andreu, M.; Hernández, C.; Jover, R.; Montalvo, I.; et al. Colonoscopy versus Fecal Immunochemical Testing in Colorectal-Cancer Screening. *N. Engl. J. Med.* **2015**, *366*, 697–706. [CrossRef] [PubMed]
5. Lindholm, E.; Brevinge, H.; Haglind, E. Survival benefit in a randomized clinical trial of faecal occult blood screening for colorectal cancer. *Br. J. Surg.* **2008**, *95*, 1029–1036. [CrossRef] [PubMed]
6. Faivre, J.; Dancourt, V.; Lejeune, C.; Tazi, M.A.; Lamour, J.; Gerard, D.; Dassonville, F.; Bonithon-Kopp, C. Reduction in colorectal cancer mortality by fecal occult blood screening in a French controlled study. *Gastroenterology* **2004**, *126*, 1674–1680. [CrossRef] [PubMed]
7. Atkin, W.S.; Edwards, R.; Kralj-Hans, I.; Wooldrage, K.; Hart, A.R.; Northover, J.M.; Parkin, D.M.; Wardle, J.; Duffy, S.W.; Cuzick, J. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: A multicentre randomised controlled trial. *Lancet* **2010**, *375*, 1624–1633. [CrossRef]
8. Segnan, N.; Senore, C.; Andreoni, B.; Arrigoni, A.; Bisanti, L.; Cardelli, A.; Castiglione, G.; Crosta, C.; DiPlacido, R.; Ferrari, A.; et al. Randomized trial of different screening strategies for colorectal cancer: Patient response and detection rates. *J. Natl. Cancer Inst.* **2005**, *97*, 347–357. [CrossRef] [PubMed]
9. Imperiale, T.F.; Ransohoff, D.F.; Itzkowitz, S.H.; Levin, T.R.; Lavin, P.; Lidgard, G.P.; Ahlquist, D.A.; Berger, B.M. Multitarget Stool DNA Testing for Colorectal-Cancer Screening. *N. Engl. J. Med.* **2014**, *370*, 1287–1297. [CrossRef] [PubMed]
10. Levin, B.; Lieberman, D.A.; McFarland, B.; Smith, R.A.; Brooks, D.; Andrews, K.S.; Dash, C.; Giardiello, F.M.; Glick, S.; Levin, T.R.; et al. Screening and Surveillance for the Early Detection of Colorectal Cancer and Adenomatous Polyps, 2008: A Joint Guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *CA Cancer J. Clin.* **2008**, *58*, 130–160. [CrossRef] [PubMed]
11. Regula, J.; Rupinski, M.; Kraszewska, E.; Polkowski, M.; Pachlewski, J.; Orlowska, J.; Nowacki, M.P.; Butruk, E. Colonoscopy Screening for Detection of Advanced Neoplasia. *N. Engl. J. Med.* **2006**, *355*, 1863–1872. [CrossRef] [PubMed]
12. Bujanda, L.; Sarasqueta, C.; Zubiaurre, L.; Cosme, A.; Muñoz, C.; Sánchez, A.; Martín, C.; Tito, L.; Piñol, V.; Castells, A.; et al. Low adherence to colonoscopy in the screening of first-degree relatives of patients with colorectal cancer. *Gut* **2007**, *56*, 1714–1718. [CrossRef] [PubMed]
13. Puente Gutiérrez, J.J.; Marín Moreno, M.A.; Domínguez Jiménez, J.L.; Bernal Blanco, E.; Díaz Iglesias, J.M. Effectiveness of a colonoscopic screening programme in first-degree relatives of patients with colorectal cancer. *Color. Dis.* **2011**, *13*, 145–153. [CrossRef] [PubMed]
14. Mansouri, D.; McMillan, D.C.; Crearie, C.; Morrison, D.S.; Crighton, E.M.; Horgan, P.G. Temporal trends in mode, site and stage of presentation with the introduction of colorectal cancer screening: A decade of experience from the West of Scotland. *Br. J. Cancer* **2015**, *113*, 556–561. [CrossRef] [PubMed]
15. Cubiella, J.; Digby, J.; Rodríguez-Alonso, L.; Vega, P.; Salve, M.; Díaz-Ondina, M.; Strachan, J.A.; Mowat, C.; McDonald, P.J.; Carey, F.A.; et al. The fecal hemoglobin concentration, age and sex test score: Development and external validation of a simple prediction tool for colorectal cancer detection in symptomatic patients. *Int. J. Cancer* **2017**, *140*, 2201–2211. [CrossRef] [PubMed]
16. Cubiella, J.; Vega, P.; Salve, M.; Díaz-Ondina, M.; Alves, M.T.; Quintero, E.; Álvarez-Sánchez, V.; Fernández-Bañares, F.; Boadas, J.; Campo, R.; et al. COLONPREDICT study investigators Development and external validation of a faecal immunochemical test-based prediction model for colorectal cancer detection in symptomatic patients. *BMC Med.* **2016**, *14*, 128. [CrossRef] [PubMed]

17. Westwood, M.; Lang, S.; Armstrong, N.; van Turenhout, S.; Cubiella, J.; Stirk, L.; Ramos, I.C.; Luyendijk, M.; Zaim, R.; Kleijnen, J.; et al. Faecal immunochemical tests (FIT) can help to rule out colorectal cancer in patients presenting in primary care with lower abdominal symptoms: A systematic review conducted to inform new NICE DG30 diagnostic guidance. *BMC Med.* **2017**, *15*, 1–17. [CrossRef] [PubMed]
18. Chen, C.; Gonzalez, F.J.; Idle, J.R. LC-MS-based metabolomics in drug metabolism. *Drug Metab. Rev.* **2007**, *39*, 581–597. [CrossRef] [PubMed]
19. Clarke, C.J.; Haselden, J.N. Metabolic Profiling as a Tool for Understanding Mechanisms of Toxicity. *Toxicol. Pathol.* **2008**, *36*, 140–147. [CrossRef] [PubMed]
20. Fernie, A.R.; Trethewey, R.N.; Krotzky, A.J.; Willmitzer, L. Metabolite profiling: from diganostics to systems biology. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 1–7. [CrossRef] [PubMed]
21. Nicholson, J.K.; Wilson, I.D. Understanding "global" systems biology: Metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discov.* **2003**, *2*, 668–676. [CrossRef] [PubMed]
22. Nordström, A.; O'Maille, G.; Qin, C.; Siuzdak, G. Non-linear Data Alignment for UPLC-MS and HPLC-MS based Metabolomics: Application to Endogenous and Exogenous Metabolites in Human Serum. *Anal Chem* **2006**, *15*, 3289–3295. [CrossRef] [PubMed]
23. Nováková, L.; Solichová, D.; Solich, P. Advantages of ultra performance liquid chromatography over high-performance liquid chromatography: Comparison of different analytical approaches during analysis of diclofenac gel. *J. Sep. Sci.* **2006**, *29*, 2433–2443. [CrossRef] [PubMed]
24. Zhang, F.; Zhang, Y.; Zhao, W.; Deng, K.; Wang, Z.; Yang, C.; Ma, L.; Openkova, M.S.; Hou, Y.; Li, K. Metabolomics for biomarker discovery in the diagnosis, prognosis, survival and recurrence of colorectal cancer: a systematic review. *Oncotarget* **2017**, *8*, 35460–35472. [CrossRef] [PubMed]
25. Cross, A.J.; Moore, S.C.; Boca, S.; Huang, W.-Y.; Xiong, X.; Stolzenberg-Solomon, R.; Sinha, R.; Sampson, J.N. A prospective study of serum metabolites and colorectal cancer risk. *Cancer* **2014**, *120*, 3049–3057. [CrossRef] [PubMed]
26. Ikeda, A.; Nishiumi, S.; Shinohara, M.; Yoshie, T.; Hatano, N.; Okuno, T.; Bamba, T.; Fukusaki, E.; Takenawa, T.; Azuma, T.; et al. Serum metabolomics as a novel diagnostic approach for gastrointestinal cancer. *Biomed. Chromatogr.* **2012**, *26*, 548–558. [CrossRef] [PubMed]
27. Leichtle, A.B.; Nuoffer, J.M.; Ceglarek, U.; Kase, J.; Conrad, T.; Witzigmann, H.; Thiery, J.; Fiedler, G.M. Serum amino acid profiles and their alterations in colorectal cancer. *Metabolomics* **2012**, *8*, 643–653. [CrossRef] [PubMed]
28. Li, F.; Qin, X.; Chen, H.; Qiu, L.; Guo, Y.; Liu, H.; Chen, G.; Song, G.; Wang, X.; Li, F.; et al. Lipid profiling for early diagnosis and progression of colorectal cancer using direct-infusion electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Rapid Commun. Mass Spectrom.* **2013**, *27*, 24–34. [CrossRef] [PubMed]
29. Nishiumi, S.; Kobayashi, T.; Ikeda, A.; Yoshie, T.; Kibi, M.; Izumi, Y.; Okuno, T.; Hayashi, N.; Kawano, S.; Takenawa, T.; et al. A novel serum metabolomics-based diagnostic approach for colorectal cancer. *PLoS ONE* **2012**, *7*, 1–10. [CrossRef] [PubMed]
30. Ma, Y.; Zhang, P.; Wang, F.; Liu, W.; Yang, J.; Qin, H. An integrated proteomics and metabolomics approach for defining oncofetal biomarkers in the colorectal cancer. *Ann. Surg.* **2012**, *255*, 720–730. [CrossRef] [PubMed]
31. Ritchie, S.A.; Ahiahonu, P.W.K.; Jayasinghe, D.; Heath, D.; Liu, J.; Lu, Y.; Jin, W.; Kavianpour, A.; Yamazaki, Y.; Khan, A.M.; et al. Reduced levels of hydroxylated, polyunsaturated ultra long-chain fatty acids in the serum of colorectal cancer patients: implications for early screening and detection. *BMC Med.* **2010**, *8*, 13. [CrossRef] [PubMed]
32. Tan, B.; Qiu, Y.; Zou, X.; Chen, T.; Xie, G.; Cheng, Y.; Dong, T.; Zhao, L.; Feng, B.; Hu, X.; et al. Metabonomics Identi fi es Serum Metabolite Markers of Colorectal Cancer. *J. Proteome Res.* **2013**, *12*, 3000–3009. [CrossRef] [PubMed]
33. Zhu, J.; Djukovic, D.; Deng, L.; Gu, H.; Himmati, F.; Chiorean, E.G.; Raftery, D. Colorectal cancer detection using targeted serum metabolic profiling. *J. Proteome Res.* **2014**, *13*, 4120–4130. [CrossRef] [PubMed]
34. Manna, S.K.; Tanaka, N.; Krausz, K.W.; Haznadar, M.; Xue, X.; Matsubara, T.; Bowman, E.D.; Fearon, E.R.; Harris, C.C.; Shah, Y.M.; et al. Biomarkers of coordinate metabolic reprogramming in colorectal tumors in mice and humans. *Gastroenterology* **2014**, *146*, 1313–1324. [CrossRef] [PubMed]

35. Mirnezami, R.; Jiménez, B.; Li, J.V.; Kinross, J.M.; Veselkov, K.; Goldin, R.D.; Holmes, E.; Nicholson, J.K.; Darzi, A. Rapid diagnosis and staging of colorectal cancer via high-resolution magic angle spinning nuclear magnetic resonance (HR-MAS NMR) spectroscopy of intact tissue biopsies. *Ann. Surg.* **2014**, *259*, 1138–1149. [CrossRef] [PubMed]
36. Wang, H.; Wang, L.; Zhang, H.; Deng, P.; Chen, J.; Zhou, B.; Hu, J.; Zou, J.; Lu, W.; Xiang, P.; et al. ^{1}H NMR-based metabolic profiling of human rectal cancer tissue. *Mol. Cancer* **2013**, *12*, 121. [CrossRef] [PubMed]
37. Silva, C.L.; Passos, M.; Cmara, J.S. Investigation of urinary volatile organic metabolites as potential cancer biomarkers by solid-phase microextraction in combination with gas chromatography-mass spectrometry. *Br. J. Cancer* **2011**, *105*, 1894–1904. [CrossRef] [PubMed]
38. Lin, Y.; Ma, C.; Liu, C.; Wang, Z.; Yang, J.; Liu, X.; Shen, Z.; Wu, R. NMR-based fecal metabolomics fingerprinting as predictors of earlier diagnosis in patients with colorectal cancer. *Oncotarget* **2016**, *7*, 29454–29464. [CrossRef] [PubMed]
39. Irrazábal, T.; Belcheva, A.; Girardin, S.E.; Martin, A.; Philpott, D.J. The multifaceted role of the intestinal microbiota in colon cancer. *Mol. Cell* **2014**, *54*, 309–320. [CrossRef] [PubMed]
40. Gao, Z.; Guo, B.; Gao, R.; Zhu, Q.; Qin, H. Microbiota disbiosis is associated with colorectal cancer. *Front. Microbiol.* **2015**, *6*, 1–9. [CrossRef] [PubMed]
41. Yan, G.; Li, L.; Zhu, B.; Li, Y. Lipidome in colorectal cancer. *Oncotarget* **2016**, *7*, 33429–33439. [CrossRef] [PubMed]
42. Valcz, G.; Patai, Á.V.; Kalmár, A.; Péterfia, B.; Furi, I.; Wichmann, B.; Muzes, G.; Sipos, F.; Krenács, T.; Mihály, E.; et al. Myofibroblast-derived SFRP1 as potential inhibitor of colorectal carcinoma field effect. *PLoS ONE* **2014**, *9*, 18–20. [CrossRef] [PubMed]
43. Chow, J.; Panasevich, M.R.; Alexander, D.; Vester Boler, B.M.; Rossoni Serao, M.C.; Faber, T.A.; Bauer, L.L.; Fahey, G.C. Fecal metabolomics of healthy breast-fed versus formula-fed infants before and during in vitro batch culture fermentation. *J. Proteome Res.* **2014**, *13*, 2534–2542. [CrossRef] [PubMed]
44. Zheng, X.; Xie, G.; Zhao, A.; Zhao, L.; Yao, C.; Chiu, N.H.L.; Zhou, Z.; Bao, Y.; Jia, W.; Nicholson, J.K.; et al. The Footprints of Gut Microbial-Mammalian Co-Metabolism. *J. Proteome Res.* **2011**, *10*, 5512–5522. [CrossRef] [PubMed]
45. Jump, R.L.P.; Polinkovsky, A.; Hurless, K.; Sitzlar, B.; Eckart, K.; Tomas, M.; Deshpande, A.; Nerandzic, M.M.; Donskey, C.J. Metabolomics analysis identifies intestinal microbiota-derived biomarkers of colonization resistance in clindamycin-treated mice. *PLoS ONE* **2014**, *9*. [CrossRef] [PubMed]
46. Martin, F.J.; Sprenger, N.; Montoliu, I.; Rezzi, S.; Kochhar, S.; Nicholson, J.K. Dietary Modulation of Gut Functional Ecology Studied by Fecal Metabonomics Francois-Pierre. *J. Proteome Res.* **2010**, *9*, 5284–5295. [CrossRef] [PubMed]
47. Phua, L.C.; Koh, P.K.; Cheah, P.Y.; Ho, H.K.; Chan, E.C.Y. Global gas chromatography/time-of-flight mass spectrometry (GC/TOFMS)-based metabonomic profiling of lyophilized human feces. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2013**, *937*, 103–113. [CrossRef] [PubMed]
48. Saric, J.; Wang, Y.; Li, J.; Coen, M.; Utzinger, J.; Marchesi, J.R.; Keiser, J.; Veselkov, K.; Lindon, J.C.; Nicholson, J.K.; et al. Species variation in the fecal metabolome gives insight into differential gastrointestinal function. *J. Proteome Res.* **2008**, *7*, 352–360. [CrossRef] [PubMed]
49. Stella, C.; Beckwith-Hall, B.; Cloarec, O.; Holmes, E.; Lindon, J.C.; Powell, J.; Van Der Ouderaa, F.; Bingham, S.; Cross, A.J.; Nicholson, J.K. Susceptibility of human metabolic phenotypes to dietary modulation. *J. Proteome Res.* **2006**, *5*, 2780–2788. [CrossRef] [PubMed]
50. Weir, T.L.; Manter, D.K.; Sheflin, A.M.; Barnett, B.A.; Heuberger, A.L.; Ryan, E.P. Stool Microbiome and Metabolome Differences between Colorectal Cancer Patients and Healthy Adults. *PLoS ONE* **2013**, *8*. [CrossRef] [PubMed]
51. Xu, W.; Chen, D.; Wang, N.; Zhang, T.; Zhou, R.; Huan, T.; Lu, Y.; Su, X.; Xie, Q.; Li, L.; et al. Development of High-Performance Chemical Isotope Labeling LC-MS for Profiling the Human Fecal Metabolome. *Anal. Chem.* **2017**, *89*, 6758–6765. [CrossRef] [PubMed]
52. Zhao, Y.; Wu, J.; Li, J.V.; Zhou, N.; Tang, H.; Wang, Y. Gut Microbiota Composition Modifies Fecal Metabolic Profiles in Mice. *J. Proteome Res.* **2013**, *12*. [CrossRef] [PubMed]

53. Gao, X.; Pujos-Guillot, E.; Martin, J.F.; Galan, P.; Juste, C.; Jia, W.; Sebedio, J.L. Metabolite analysis of human fecal water by gas chromatography/mass spectrometry with ethyl chloroformate derivatization. *Anal. Biochem.* **2009**, *393*, 163–175. [CrossRef] [PubMed]
54. Gao, X.; Pujos-Guillot, E.; Sébédio, J.L. Development of a quantitative metabolomic approach to study clinical human fecal water metabolome based on trimethylsilylation derivatization and GC/MS analysis. *Anal. Chem.* **2010**, *82*, 6447–6456. [CrossRef] [PubMed]
55. Poroyko, V.; Morowitz, M.; Bell, T.; Ulanov, A.; Wang, M.; Donovan, S.; Bao, N.; Gu, S.; Hong, L.; Alverdy, J.C.; et al. Diet creates metabolic niches in the "inmature gut" that shape microbial communities. *Nutr. Hosp.* **2011**, *26*, 1283–1295. [CrossRef] [PubMed]
56. Ponnusamy, K.; Choi, J.N.; Kim, J.; Lee, S.Y.; Lee, C.H. Microbial community and metabolomic comparison of irritable bowel syndrome faeces. *J. Med. Microbiol.* **2011**, *60*, 817–827. [CrossRef] [PubMed]
57. Sciorra, V.A.; Morris, A.J. Roles for lipid phosphate phosphatases in regulation of cellular signaling. *Biochim. Biophys. Acta-Mol. Cell Biol. Lipids* **2002**, *1582*, 45–51. [CrossRef]
58. Tang, X.; Benesch, M.G.K.; Brindley, D.N. Lipid phosphate phosphatases and their roles in mammalian physiology and pathology. *J. Lipid Res.* **2015**, *56*, 2048–2060. [CrossRef] [PubMed]
59. Weaver, G.A.; Krause, J.A.; Miller, T.L.; Wolin, M.J. Short chain fatty acid distribution of enema samples from a sigmoidoscopy population:an association of high acetate and low butyrate ratios with adenomatous polyps and colon cancer. *Gut* **1988**, *29*, 1539–1543. [CrossRef] [PubMed]
60. Barr, J.; Caballería, J.; Martínez-Arranz, I.; Domínguez-Díez, A.; Alonso, C.; Muntané, J.; Pérez-Cormenzana, M.; García-Monzón, C.; Mayo, R.; Martín-Duce, A.; et al. Obesity-dependent metabolic signatures associated with nonalcoholic fatty liver disease progression. *J. Proteome Res.* **2012**, *11*, 2521–2532. [CrossRef] [PubMed]
61. Saccenti, E.; Hoefsloot, H.C.J.; Smilde, A.K.; Westerhuis, J.A.; Hendriks, M.M.W.B. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics* **2014**, *10*, 361–374. [CrossRef]
62. Martínez-Arranz, I.; Mayo, R.; Pérez-Cormenzana, M.; Mincholé, I.; Salazar, L.; Alonso, C.; Mato, J.M. Enhancing metabolomics research through data mining. *J. Proteomics* **2015**, *127*, 275–288. [CrossRef] [PubMed]
63. Cock, P.J.A.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423. [CrossRef] [PubMed]
64. Cokelaer, T.; Pultz, D.; Harder, L.M.; Serra-Musach, J.; Saez-Rodriguez, J.; Valencia, A. BioServices: A common Python package to access biological Web Services programmatically. *Bioinformatics* **2013**, *29*, 3241–3242. [CrossRef] [PubMed]
65. Szklarczyk, D.; Morris, J.H.; Cook, H.; Kuhn, M.; Wyder, S.; Simonovic, M.; Santos, A.; Doncheva, N.T.; Roth, A.; Bork, P.; et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **2017**, *45*, D362–D368. [CrossRef] [PubMed]
66. Luo, W.; Brouwer, C. Pathview: An R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **2013**, *29*, 1830–1831. [CrossRef] [PubMed]

EBioMedicine 46 (2019) 499–511

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

EBioMedicine
Published by THE LANCET

Gut microbiome and serum metabolome analyses identify molecular biomarkers and altered glutamate metabolism in fibromyalgia

Marc Clos-Garcia [a,b], Naiara Andrés-Marin [c], Gorka Fernández-Eulate [c,d], Leticia Abecia [e], José L. Lavín [f], Sebastiaan van Liempd [g], Diana Cabrera [g], Félix Royo [a], Alejandro Valero [h], Nerea Errazquin [i], María Cristina Gómez Vega [j], Leila Govillard [k], Michael R. Tackett [l], Genesis Tejada [l], Esperanza Gónzalez [a], Juan Anguita [e,m], Luis Bujanda [b], Ana María Callejo Orcasitas [j], Ana M. Aransay [n], Olga Maíz [h], Adolfo López de Munain [c,d,o,p], Juan Manuel Falcón-Pérez [a,g,m,*]

[a] Exosomes Laboratory, CIC bioGUNE, CIBERehd, Bizkaia Technology Park, Derio, Spain
[b] Department of Gastroenterology, Instituto Biodonostia, Universidad del País Vasco (UPV/EHU), CIBERehd (Centro de investigación en red de enfermedades hepáticas y digestiva) San Sebastian, Spain
[c] Department of Neurology, Donostia University Hospital, San Sebastian, Spain
[d] Neuroscience Area, Biodonostia Health Research Institute, San Sebastian, Spain
[e] Macrophage and Tick Vaccine Laboratory, CIC bioGUNE, Bizkaia Technology Park, Derio, Spain
[f] Bioinformatics Unit, CIC bioGUNE, Bizkaia Technology Park, Derio, Spain
[g] Metabolomics Platform, CIC bioGUNE, CIBERehd, Bizkaia Technology Park, Derio, Spain
[h] Department of Rheumatology, Donostia University Hospital, San Sebastian, Spain
[i] Department of Rheumatology, Gipuzcoa Policlinic, San Sebastian, Spain
[j] Department of Anesthesiology and Pain Unit, Hospital Galdakao-Usansolo, Bizkaia, Spain
[k] Deusto University, San Sebastian, Spain
[l] Abcam, Cambridge, MA, United States
[m] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[n] Genome Analysis Platform, CIC bioGUNE, CIBERehd, Bizkaia Technology Park, Derio, Spain
[o] Network Center for Biomedical Research in Neurodegenerative Diseases (CIBERNED), Spain
[p] Department of Neurosciences, University of Basque Country UPV/EHU, San Sebastian, Spain

ARTICLE INFO

Article history:
Received 12 March 2019
Received in revised form 24 June 2019
Accepted 10 July 2019
Available online 18 July 2019

Keywords:
Fibromyalgia
Gut microbiota
Pain
Metabolomics
Cytokines
miRNAs
Omics integration

ABSTRACT

Background: Fibromyalgia is a complex, relatively unknown disease characterised by chronic, widespread musculoskeletal pain. The gut-brain axis connects the gut microbiome with the brain through the enteric nervous system (ENS); its disruption has been associated with psychiatric and gastrointestinal disorders. To gain an insight into the pathogenesis of fibromyalgia and identify diagnostic biomarkers, we combined different omics techniques to analyse microbiome and serum composition.

Methods: We collected faeces and blood samples to study the microbiome, the serum metabolome and circulating cytokines and miRNAs from a cohort of 105 fibromyalgia patients and 54 age- and environment-matched healthy individuals. We sequenced the V3 and V4 regions of the 16S rDNA gene from faeces samples. UPLC-MS metabolomics and custom multiplex cytokine and miRNA analysis (FirePlex™ technology) were used to examine sera samples. Finally, we combined the different data types to search for potential biomarkers.

Results: We found that the diversity of bacteria is reduced in fibromyalgia patients. The abundance of the *Bifidobacterium* and *Eubacterium* genera (bacteria participating in the metabolism of neurotransmitters in the host) in these patients was significantly reduced. The serum metabolome analysis revealed altered levels of glutamate and serine, suggesting changes in neurotransmitter metabolism. The combined serum metabolomics and gut microbiome datasets showed a certain degree of correlation, reflecting the effect of the microbiome on metabolic activity. We also examined the microbiome and serum metabolites, cytokines and miRNAs as potential sources of molecular biomarkers of fibromyalgia.

* Corresponding author at: Exosomes Laboratory, CIC bioGUNE, CIBERehd, Bizkaia Technology Park, Derio, Spain
E-mail addresses: mclos.biodonostia@cicbiogune.es (M. Clos-Garcia), gorka.fernandezgarciadeeulate@osakidetza.net (G. Fernández-Eulate), labecia@cicbiogune.es (L. Abecia), jllavin@cicbiogune.es (J.L. Lavín), smvanliempd@cicbiogune.es (S. van Liempd), dcabrera@cicbiogune.es (D. Cabrera), froyo.ciberehd@cicbiogune.es (F. Royo), jesusalejandro.valerojaimes@osakidetza.eus (A. Valero), nerea.errazquinaguirre@osakidetza.eus (N. Errazquin), leila.govillard@deusto.es (L. Govillard), michael.tackett@abcam.com (M.R. Tackett), genesis.tejada@abcam.com (G. Tejada), egonzalez@cicbiogune.es (E. Gónzalez), janguita@cicbiogune.es (J. Anguita), luis.bujandafernandezdepierola@osakidetza.eus (L. Bujanda), mariacristina.gomezvega@osakidetza.eus (A.M.C. Orcasitas), anamariafrancisca.callejoorcasitas@osakidetza.eus, amaransay@cicbiogune.es (A.M. Aransay), olga.maizalonso@osakidetza.eus (O. Maíz), ADOLFOJOSE.LOPEZDEMUNAINARREGUI@osakidetza.eus (A. López de Munain), jfalcon@cicbiogune.es (J.M. Falcón-Pérez).

https://doi.org/10.1016/j.ebiom.2019.07.031

Conclusions: Our results show that the microbiome analysis provides more significant biomarkers than the other techniques employed in the work. Gut microbiome analysis combined with serum metabolomics can shed new light onto the pathogenesis of fibromyalgia. We provide a list of bacteria whose abundance changes in this disease and propose several molecules as potential biomarkers that can be used to evaluate the current diagnostic criteria.

Research in context

Evidence before this study

Fibromyalgia is a complex disease with chronic pain as its primary symptom. To date, no molecular biomarkers exist for it, leaving its diagnosis up to subjective questionnaires. Several alterations in fibromyalgia patients have pointed towards the central nervous system as the origin of this pathology. The gut microbiome can influence the CNS through the gut-brain axis.

Added value of this study

Employing microbiome and metabolomics analysis along with cytokine and miRNA profiling we identified several alterations between healthy controls and fibromyalgia patients that could be used as potential biomarkers. We also studied how the microbiome and metabolomics datasets correlated with each other to elucidate the role of microbiome alterations in host metabolism.

Implications of all available evidence

Taken together, this study provides candidate molecular biomarkers for fibromyalgia, and supports an alteration of neurotransmitter levels in fibromyalgia patients.

1. Background

Fibromyalgia is a complex disease of unknown pathophysiology, for which no specific molecular biomarkers or biochemical alterations have been identified. In 1990, the American College of Rheumatology (ACR) recognised this syndrome as a disease and proposed the Widespread Pain Index (WPI), determined by measuring tenderness on pressure at 18 defined points, as a major diagnostic indicator. In 2010, the ACR introduced the Severity Score (SS), which also takes into account the associated symptoms and their severity [102]. Thus, the diagnosis of fibromyalgia is currently based on subjective pain evaluation and a set of associated signs and symptoms, which are used to assess the severity of the disease.

Even though the fibromyalgia is a complex disease with a multitude of signs and symptoms associated with many organs, the participation of the Central Nervous System (CNS) in its pathogenesis is broadly acknowledged [33]. Some studies have tried to identify molecular signatures that could explain some of the features of fibromyalgia and have provided some potential biomarkers. Several polymorphisms linked to the metabolism and breakdown of neurotransmitters involved in pain modulation have been identified as specific markers of increased risk of fibromyalgia [2]. Such polymorphisms have been found for the serotonin transporter gene 5-HTT [14,68] and the catechol-*O*-methyltransferase (COMT) gene [30,106]. Some environmental factors, such as viral and bacterial infections [10], e.g. HCV infection [9,78] and psychological stressors [32], known to produce alterations in the hypothalamic-pituitary-adrenal (HPA) axis, have been associated with this disease. Fibromyalgia is prevalent in individuals with chronic pain attributable to peripheral pain generators, such as rheumatoid arthritis [1]. At the molecular level, glutamate is elevated in the cerebrospinal fluid of fibromyalgia patients [26,71,85]. A decrease in insular levels of γ-aminobutyric acid (GABA) has also been described [21]. An inflammatory component in the pathogenesis of this disease has also been proposed: certain cells might trigger and perpetuate chronic pain by releasing chemokines and cytokines, such as IL-6 and IL-8, whose levels are elevated in the sera of fibromyalgia patients [62,95].

The microbiome has a significant role in maintaining health [37,47]. Alterations in the gut microbiome have been linked to a large number of diseases, including intestinal bowel disease (IBD) [45] and metabolic [43] and neurological [84,89] disorders [40]. The microbiome has been recurrently associated with the CNS [89], indicating the existence of a gut-brain axis [16,22]. Disturbances in the microbiome might lead, in some cases, to neural disorders such as depression or autism. Some changes linked to microbial gut dysbiosis, understanding dysbiosis as those differences between healthy individuals and disease-specific patients [35], are also associated with symptoms used to determine the SS_2 score in the diagnosis of fibromyalgia. The gut-brain axis has been proposed as a bidirectional communication system between the gastrointestinal tract and the brain, involving both neural and humoral mechanisms (reviewed in [15]). The intestinal GABA produced by the bacteria from glutamate might affect the behaviour of the host, and it might be involved in anxiety and depression [8,34,57,88]. Alterations in the microbiome composition can escalate the interactions between bacteria and the gut immune system due to the breakage of the intestinal barrier, promoting the release of pro-inflammatory molecules. Such events have been reported in IBD, where a release of IL-2, IL-17, interferon and/or TNFβ has been observed [41]. Interestingly, several pro-inflammatory cytokines can increase the permeability of the blood-brain barrier [16]. The microbiome also has metabolic, immunological and gut-protecting functions in the host. The fermentation of dietary carbohydrates by gut bacteria, for example, results in the production of short-chain fatty acids (SCFAs). These molecules are essential for the maintenance of the integrity of intestinal barrier [40] and other health-related functions [77], including the correct development and maintenance of the blood-brain barrier [7].

These interactions between the microbiome and other functional systems of the organism has been widely studied. Microbiome data have been scrutinised in conjunction with host's genome, epigenome, transcriptome and metabolome [99]. The integration of different omics data relies mostly on dimension reduction approaches and is not specific to any omics technology, except for the metabolomics data. Correlation, regression and network-based approaches have also been implemented to integrate microbiome data with other omics analyses. As a result, the role of the of the host genome in regulating microbiome composition has been revealed [28]. Combination of Genome Wide Association Studies (GWAS) and microbiome-GWAS has been applied also to assess the impact of diet on microbiome composition. For example, associations between lactase [5] and variations of vitamin D receptor [98] genes with specific bacteria have been reported. Metabolomics-microbiome integration studies using correlation approaches have shown the effect of microbiome on host's insulin sensitivity [70] and on the development and progression of colorectal cancer [66,100]. Metabolomics – microbiome integration studies

employing a mix of correlation and network methods have obtained a comprehensive profile of the existent interactions between intestinal mucosa and gut microbiome [58]. The authors of these studies have used standard statistical methods but suggested that new, specific methods are needed for omics integration, to take into account the particular omics data characteristics [99].

The aim of this work was to identify potential molecular biomarkers for fibromyalgia diagnosis and characterisation, employing different omics technologies: the analysis of microbiome from faeces samples and metabolomics, cytokine and miRNA profiling using serum samples.

2. Methods

2.1. Cohort recruitment

Individuals included in the study were recruited in two different hospitals in the Basque Country. Both fibromyalgia patients and healthy individuals were given a form with questions concerning several lifestyle variables (diet, smoking, alcohol consumption, physical exercise, other diseases and mood),. Blood samples were obtained from fibromyalgia patients and control individuals. Stool samples were collected from all participants, stored the samples at 4 °C until they could be delivered to the biobank. Blood and stool samples collected in each hospital were then sent to the Basque Biobank. Samples were aliquoted samples and frozen at −80 °C. The hospitals clinicians (neurologists and rheumatologists) were responsible for the fibromyalgia diagnosis. The following criteria were used:

- Fibromyalgia group: WPI ≥ 7 and SS_T (Severity Score) ≥ 5 or WPI between 3 and 6 and SST ≥ 9. Patients with other diseases with similar symptoms were discarded.
- Control group: healthy individuals without any clinical manifestation of fibromyalgia and/or any other similar disease. To reduce the potential confounding factors associated with lifestyle, they also were age-paired with the patient group and came from the same environment.

All donors signed the informed consent form, and the study was approved by the appropriate ethical committee (CEIC-PI2016037). DNA from faeces was extracted using PSP Spin Stool DNA Plus kit (STRATEC Molecular®), following the manufacturer's protocol. Lysis buffer was added to the frozen samples, to ensure the preservation of nucleic acids. DNA extractions were then aliquoted into samples of 2.5 μg of DNA at the concentration of 100 ng/μL and then frozen until sequencing. All sample processing and distribution were managed by the Basque Biobank. The summary of the collection workflow can be found in Fig. 1.

2.2. V3–V4 16S rDNA sequencing

DNA amplicon libraries were generated and sequencing performed following the recommendations of Illumina Inc. Sequencing was conducted at the FISABIO Sequencing Core Facility, as were the quality assessment using *prinseq-lite* [87] and the sequence joining, employing *FLASH* software [53] with default parameters. The complete protocol can be found in the Supplementary Methods file.

2.3. Microbiome sequences bioinformatics analysis

QIIME2 package (v. 2017.10) [12] was employed to perform the Operational Taxonomic Units (OTU) clustering and identification, using de novo methodology at 97% similarity threshold. Diversity analysis was performed, and the OTUs were annotated using GreenGenes 13_8 database. The OTU table was exported to SIMCA-9+ 12.0.1 (Umetrics AB, Umeå, Sweden) to perform multivariate analysis and to R programme (R Development Core Team [108]; http://cran.r-project.org) to conduct the statistical analysis using *phyloseq* [60], *microbiome* [48] and *DESeq2*

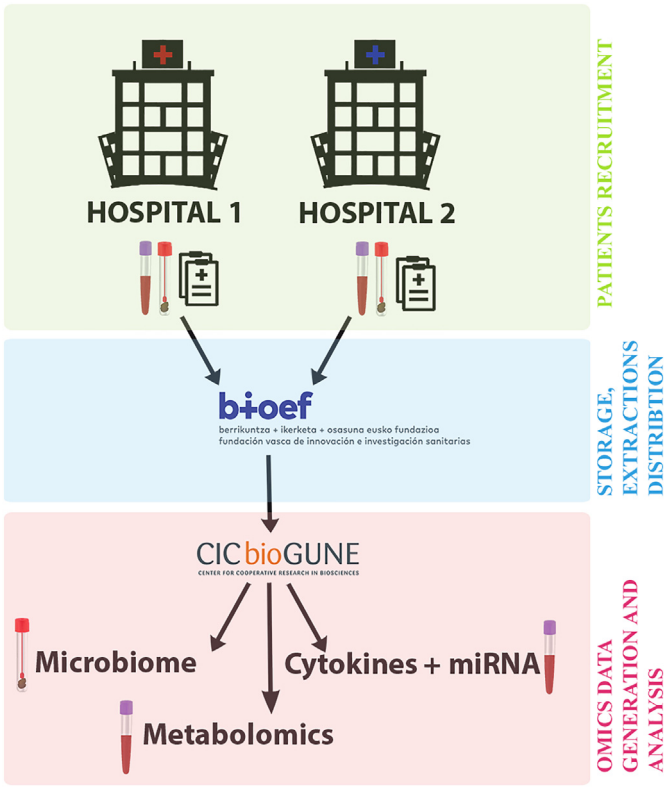


Fig. 1. Experimental design workflow, from patient recruitment and sample collection to the arrival of processed samples into the research centre and their examination using distinct omics techniques.

[52] R packages. CORBATA [49] approach was used to identify and plot the bacteria in the core microbiome. SIAMCAT tool [107] was used to assess the potential effects of confounding factors such as sex, different hospitals and distinct drug types. The adjusted *p-value* < .05 was considered statistically significant unless stated otherwise. The complete protocol can be found in the Supplementary Methods file.

2.4. qPCR validation

From the glutamate cytoplasmatic incorporation and degradation pathways we selected four genes (gadC, glnA, glsA and glsB) to validate our findings related to glutamate and microbiome interaction. We designed specific primer pairs with Primer-BLAST from NCBI webtool (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) indicating specificity for five bacterial families: Bacteroides, Bifidobacterium, Eubacterium, Lachnospiraceae and Ruminococcaceae. Complete procotol can be found in the Supplementary Methods file.

2.5. Metabolomics methodology

To 40 μL aliquots of human serum, 40 μL of water/0.15% formic acid (FA) was added. Then, the proteins were precipitated by addition of 120 μL of acetonitrile. To achieve the optimum extraction, after the addition of acetonitrile, the samples were sonicated for 15 min and agitated at 1400 rpm for 30 min (at 4 °C). Next, they were centrifuged at 14,000 rpm for 30 min at 4 °C. The supernatants were transferred to vials. Samples were examined using a UPLC system (Acquity, Waters Inc., Manchester, UK) coupled to a time-of-flight mass spectrometer (ToF MS, SYNAPT G2, Waters Inc.). A 2.1 × 100 mm, 1.7-μm BEH amide column (Waters Inc.), kept at 40 °C, was used to separate the analytes before the MS. The MS was operated in positive electrospray ionisation full scan mode. Spectral peaks were automatically corrected

for deviations in the lock mass. The complete specifications can be found in the Supplementary Methods.

Scaled and normalised data were uploaded to R. Principal Component Analysis (PCA) was performed to check whether the differences between sample metabolomes were due to sample origin and to account for the autoclaving process used by one of the hospitals. We excluded the metabolites whose expression differed between the hospitals, to avoid the bias introduced by the sample origin. Metabolomic features with >30% of missing values in either hospital were removed from the analysis. Fold changes and *p-values* (adjusted using the Bonferroni method) were computed. Afterwards, differential peaks were selected for further annotation and metabolite identification using the METLIN database [29]. The identification was confirmed using commercial standard injection.

MetScape [44] and Ingenuity Pathway Analysis® were used to map the identified metabolites to corresponding functionalities in humans.

2.6. Cytokine and miRNA profiling

The cytokine and miRNA profiling was performed by Abcam FirePlex Service (Boston, USA). The cytokine analysis was conducted using the FirePlex Human Discovery Cytokine Panel (Abcam, ab227936), allowing simultaneous profiling of 70 targets in a single well. Each sample was analysed in duplicate, following the manufacturer's instructions. The flow cytometer output was analysed using the FirePlex™ Analysis Workbench software (http://www.abcam.com/FireflyAnalysisSoftware). Cytokine concentration in a sample was interpolated from the standard curve obtained in duplicate for each plate. The data was log-normalised, and then the fold changes and Bonferroni-adjusted *p-values* were computed to assess the differences between the cytokine profiles.

The miRNAs were profiled using the FirePlex miRNA Assay Core Reagent Kit (Abcam, ab218342) employing a custom multiplex panel with 68 miRNAs selected on the basis of literature review. Each sample was run in singlicate, as previously described [93]. Data analysis was performed using the FirePlex™ Analysis Workbench software. Three miRNAs used for normalisation, hsa-miR-17-5p, hsa-miR-320b and hsa-let-7i-5p, were selected employing the geNorm algorithm [96]. The data was log-normalised, and then the fold changes and Bonferroni-adjusted *p-values* were computed to evaluate the differences between the miRNA profiles.

2.7. Omics integration

2.7.1. Microbiome and metabolomics

Spearman's correlation coefficients were computed for relationships between relative abundances of microbiome bacteria with the identified genus and normalised individual metabolomic features. A scaled heatmap was constructed for the correlation matrix, including cladogram classification of the variables, using the default clustering method.

2.7.2. Integration of all datasets

We employed the Data Integration Analysis for Biomarker Discovery (DIABLO) using Latent cOmponents implementation in the mixOmics R package [79,90]. Thirty-six fibromyalgia and 35 control samples were used. Microbiome data was normalised using DESeq2 counts function. The mixOmics block.splsda function, with full weighted design and 10 components, was primarily used to identify the optimal number of components, which was defined in 3 methods using the centroid distance technique. To decide which variables to keep in each component, models with 10, 5, 5 and 5 randomly selected variables were tested for the microbiome, metabolomics, cytokines and miRNAs, respectively. Finally, different model features were obtained and the results were plotted using mixOmics predefined and ad-hoc functions. This procedure was followed for both the identified-metabolite dataset and the full dataset of unidentified metabolomics features.

Table 1
Cohort characteristics. The number of individuals included in each group is given in parentheses. For Age, WPI and SS_T, mean values ± standard deviation are shown.

	Controls ($n = 54$)	Fibromyalgia-diagnosed patients ($n = 105$)
Sex	48.15% ♀, 51.85% ♂	69.52% ♀, 30.48% ♂
Age *(years)*	53.5 ± 12.4	52.52 ± 10.3
Age at diagnosis *(years)*	*NA*	48.2 ± 11.1
Time since diagnosis *(years)*	*NA*	3.4 ± 6
WPI	*NA*	13.28 ± 3.91
SS_T	*NA*	8.62 ± 2.15
SS1	*NA*	6.6 ± 1.8
SS2	*NA*	2.1 ± 0.4

3. Results

3.1. Clinical samples

One hundred and five confirmed fibromyalgia patients (ACR 2010 modified criteria) [102] and 54 age- and environmentally-paired healthy individuals were recruited. The latter group consisted of individuals who did not present any disease or symptoms related to fibromyalgia and came from the same environment as the fibromyalgia patients. The characteristics of the study cohort are shown in Table 1.

During WPI evaluation, >90% of the patients reported pain in the back, shoulder girdle and the abdomen. Neck pain was described by 85% of patients, while upper and lower arm, hip and upper and lower leg pain were reported by 70% of fibromyalgia patients. At least 50% of the patients were affected by jaw and chest pain. The SS_T index is the combination of two sub-indexes, SS_1 (the severity of 3 main symptoms in fibromyalgia: fatigue, sleep quality and cognitive problems) and SS_2 (the list of associated fibromyalgia symptoms). Approximately 90% of patients reported moderate to severe scores for the 3 main symptoms for the SS_1 sub-index in the week preceding the collection of the samples. In the evaluation of associated fibromyalgia symptoms (SS_2), 70.7% of fibromyalgia patients presented at least 4 symptoms from the neurological sphere (muscle pain, fatigue, thinking or memory problems, headache, numbness/tingling, etc.). Among them, 70% used painkillers, while approximately 55% were taking antidepressants and benzodiazepines and approximately 30%, antiepileptic drugs (SUPPLEMENTARY TABLE S1). Half of the patients reported some physical exercise and some alcohol consumption, while 23% identified themselves as smokers.

3.2. V3 + V4 16S rDNA sequencing

We obtained 6,110,564 reads, of which 99.56% passed the quality check. Of the cleaned reads, the 81.91% (4,982,956) were joined. To decide on the number of reads to which the samples should be rarefied; we computed the rarefaction curves for both observed OTUs and Shannon indices (Supplementary Fig. S1A). After rarefying at 12,000 reads/sample, the median coverage was 96.35 ± 2.33%. Rarefaction step did not reduce diversity (Supplementary Fig. S1B). Sequencing data was uploaded to ENA under Project Accession code PRJEB27227.

3.3. Microbiome analysis

The multivariate unsupervised PCA (Fig. 2A) did not show any differences between the control and the fibromyalgia samples. The supervised Partial Least Squares Discriminant Analysis (PLS-DA), however, provided two sample groups (Fig. 2B) (*p-value*, 0.0019). In the specific diversity analysis for 4 alpha-diversity indices (Faith's Phylogenetic Distance, ace, chao1 and observed OTUs) we observed a discrete decrease in bacterial diversity in fibromyalgia patients although only the Faith's PD index showed a statistically significant difference (Fig. 2C). This

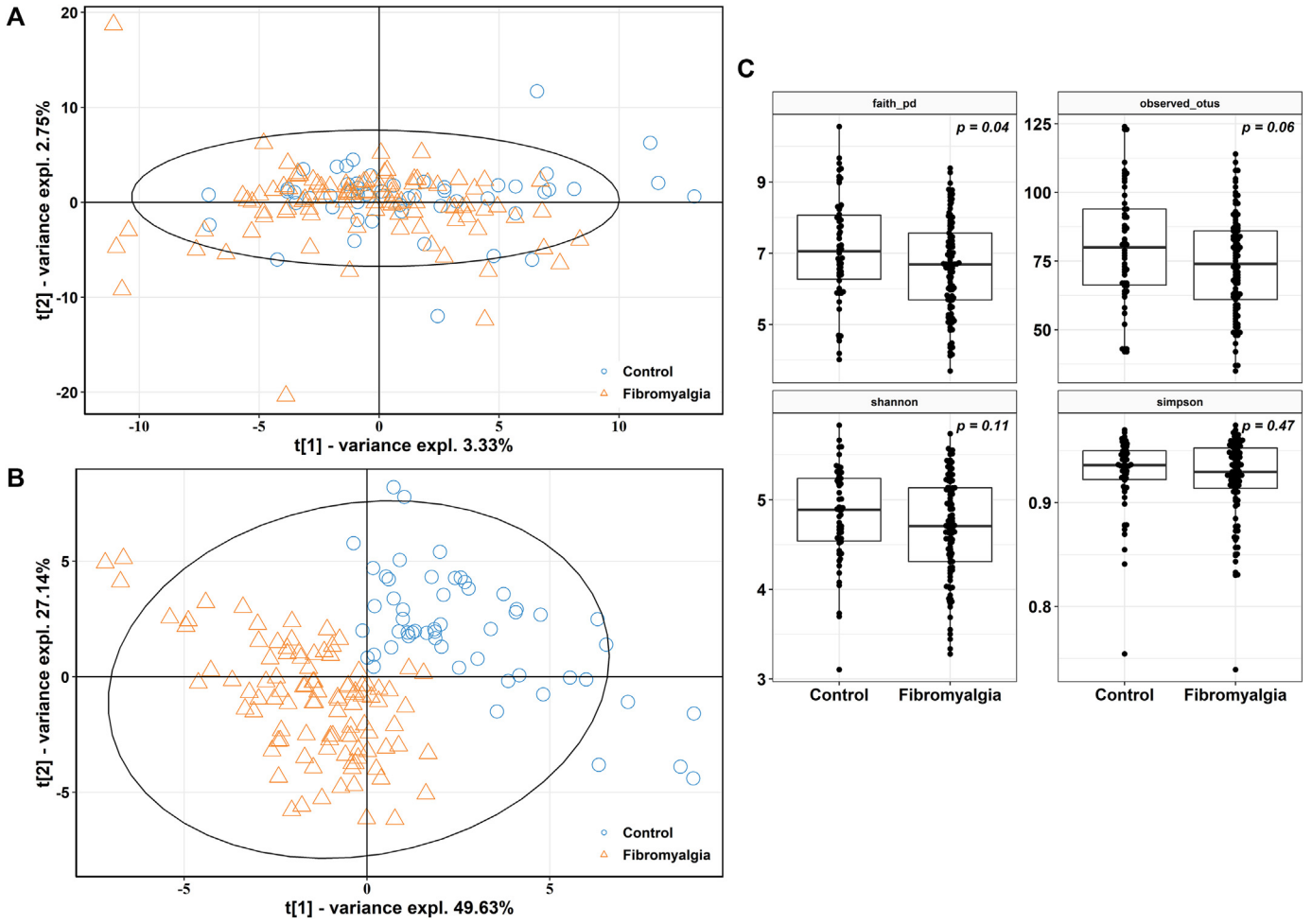


Fig. 2. Microbiome multivariate analysis. (A) Principal Component Analysis (PCoA) of the complete cohort. (B) Supervised Partial Least Squares Discriminant Analysis (PLS-DA) analysis, showing the discrimination between the sample groups. (C) Alpha-diversity indexes for each sample group, showing the adjusted *p-value* computed using Student's *t*-test.

reduction in bacterial diversity was also observed in the analysis of the core microbiome at the taxonomic family level. We used CORBATA default parameters (80% ubiquity, 1% abundance) to identify which bacteria families present in both fibromyalgia and control core microbiomes. The two core microbiomes contained the same 4 bacteria families (C. Ruminococcaceae, C. Lachnospiraceae, B. Rikenellaceae and B. Bacteroidaceae). We observed that the control group presented a more diverse bacterial community. The comparison of the two sample groups revealed that Clostridiales Ruminococcaceae was more abundant in the healthy control group than in fibromyalgia patients, although the differences were not statistically significant (Fig. 3A). After reducing the cut-off to 50% ubiquity, we observed differences between the core microbiomes of the two groups. Specifically, two bacteria families that were absent in the fibromyalgia core microbiome, the Bifidobacteriales Bifidobacteriaceae and the Bacteroidales Prevotella, which were represented in the control core microbiome (Supplementary Fig. S2A).

We performed a differential OTU analysis (employing DESeq2) of the core microbiomes in the control and fibromyalgia samples. We identified 32 OTUs distributed among 3 phyla (Actinobacteria, Bacteroidetes and Firmicutes) (Fig. 3B) whose abundance differed between the two groups, with an adjusted *p-value* of 0.05. In fibromyalgia patients, the Bacteroidetes and Firmicutes had OTUs both with increased and decreased abundance, and Actinobacteria levels were reduced in this group (Fig. 3B).

The number of OTUs with the unassigned genus in Bacteroidaceae and Lachnospiraceae families were decreased in fibromyalgia samples; there were also fewer Bifidobacteriaceae and Erysipelotichaceae OTUs in fibromyalgia patients. The Rikenellaceae family showed an increased abundance in fibromyalgia patients (Supplementary Table S2).

Finally, at the genus level, the abundance of *Bacteroides* OTUs was reduced in fibromyalgia patients, as were *Bifidobacterium*, *Eubacterium* and *Clostridium* OTUs. However, the abundances of the genera *Dorea*, *Roseburia* and *Alistipes* were increased in this group (Fig. 3B).

There were no significant differences between microbiome composition abundances in the two sexes. We did not observe any significant association between drug types (as summarized in Supplementary Table S1) and the relative microbiome abundance at the genus level (*data not shown*).

We validated the reduction of the abundance of bacterial species by qPCR technique. For that, we amplified a set of genes dedicated to the glutamate incorporation to bacterial cytoplasm and its transformation to GABA (gadC, glnA, glsA and glsB). We designed specific primers for amplifying genes from 5 bacterial families that we found to be diminished in fibromyalgia patients (Bacteroides, Bifidobacterium, Eubacterium, Lachnospiraceae and Ruminococcaceae) (Fig. 3C). We found that the gene encoding the transporter of glutamate into bacterial cytoplasm, represented by gadC, was diminished, as it was also the genes encoding enzymes involved in the transformation of glutamate to L-glutamine (glnA, glsA) and to GABA (gadB) (Supplementary Fig. S2B), in agreement with the taxonomic analysis of 16S rDNA gene.

3.4. Metabolomics analysis

The metabolomics analysis yielded 8543 different metabolic features defined by retention time and mass/charge. One sample was removed

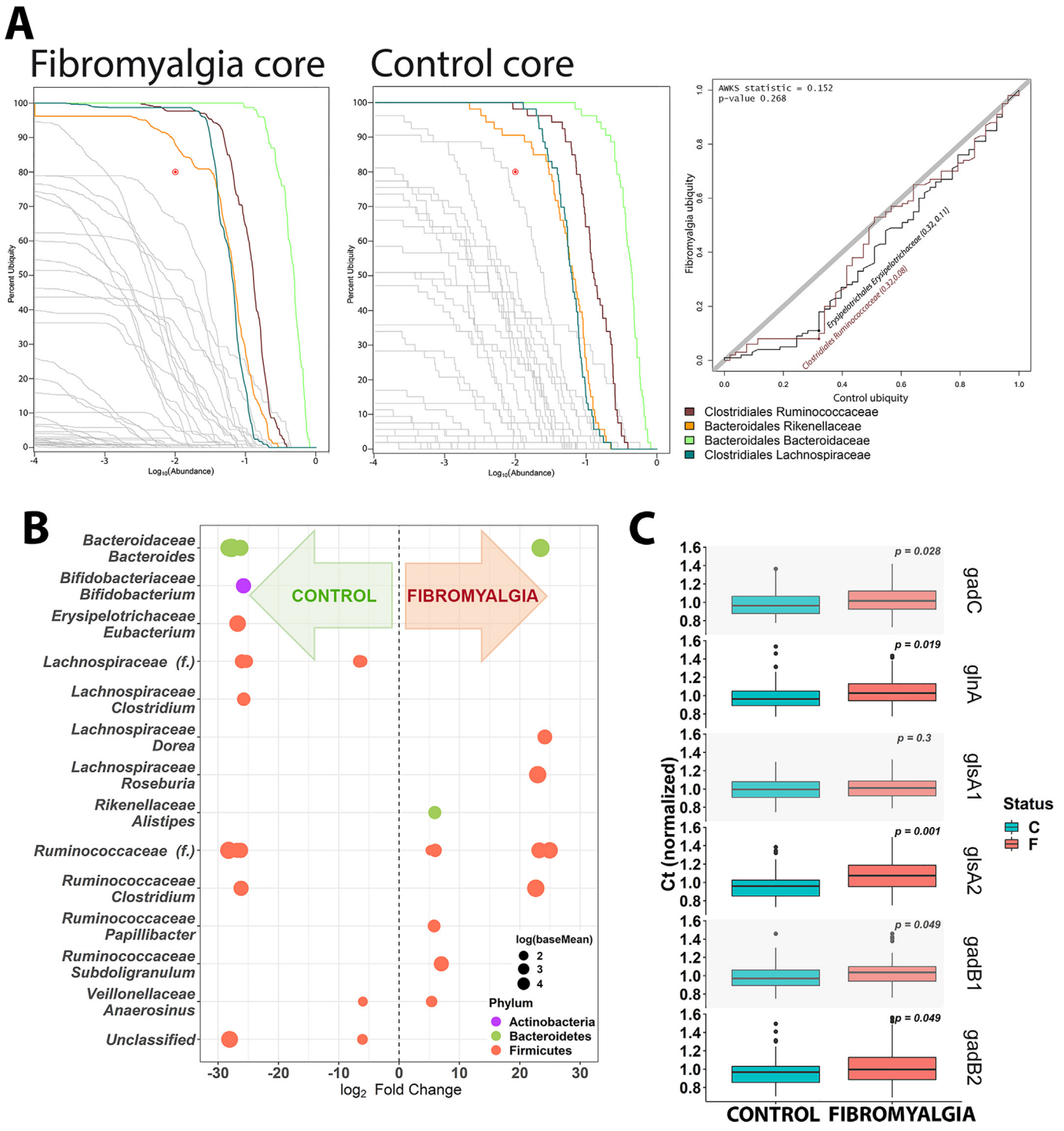


Fig. 3. Core microbiome and genus-discriminant analyses. (A) The composition of core microbiome for each sample group and the comparison of bacterial ubiquity in the two groups. (B) Genera significantly different (*adj p* > .05) between the control and fibromyalgia samples, obtained using the protocols described in the Methods. Positive $\log_2$ fold changes (x-axis) indicate genera with positive fold difference between fibromyalgia and control. Negative log2 fold changes are shown as negative x values. Each point represents a single OTU, coloured by phylum. On the y-axis, the taxonomic genus level is indicated. Size of the points reflect the log-mean abundance of the sequence data. (C) qPCR results for the differential expression of bacterial genes related to glutamate bacterial degradation. Results are indicated in differential Cts count.

due to technical failure. The PCA analysis revealed that the metabolomics profiles differed between hospitals (Supplementary Fig. 3). This was expected because of the autoclaving performed in one of the hospitals. Thus, to avoid the bias caused by the chemicals released during the autoclaving procedure, the discriminating hospital features ($p = 661$), were removed from the study, as well as the features with >30% of missing values. Two hundred and twenty-eight features differed between the fibromyalgia and control groups (Fig. 4A). Of these 228, only 88 had tentative IDs in the METLIN database. Using MS/MS data and chemical standards, we found that the levels of 7 of these metabolites were significantly altered in the fibromyalgia samples (Supplementary Table S3): ornithine, L-arginine, Nε-Methyl-L-lysine, L-glutamate, L-glutamine, asymmetric dimethylarginine (ADMA) and platelet activating factor (PAF-16) (Fig. 4B). Another metabolic feature among the 228

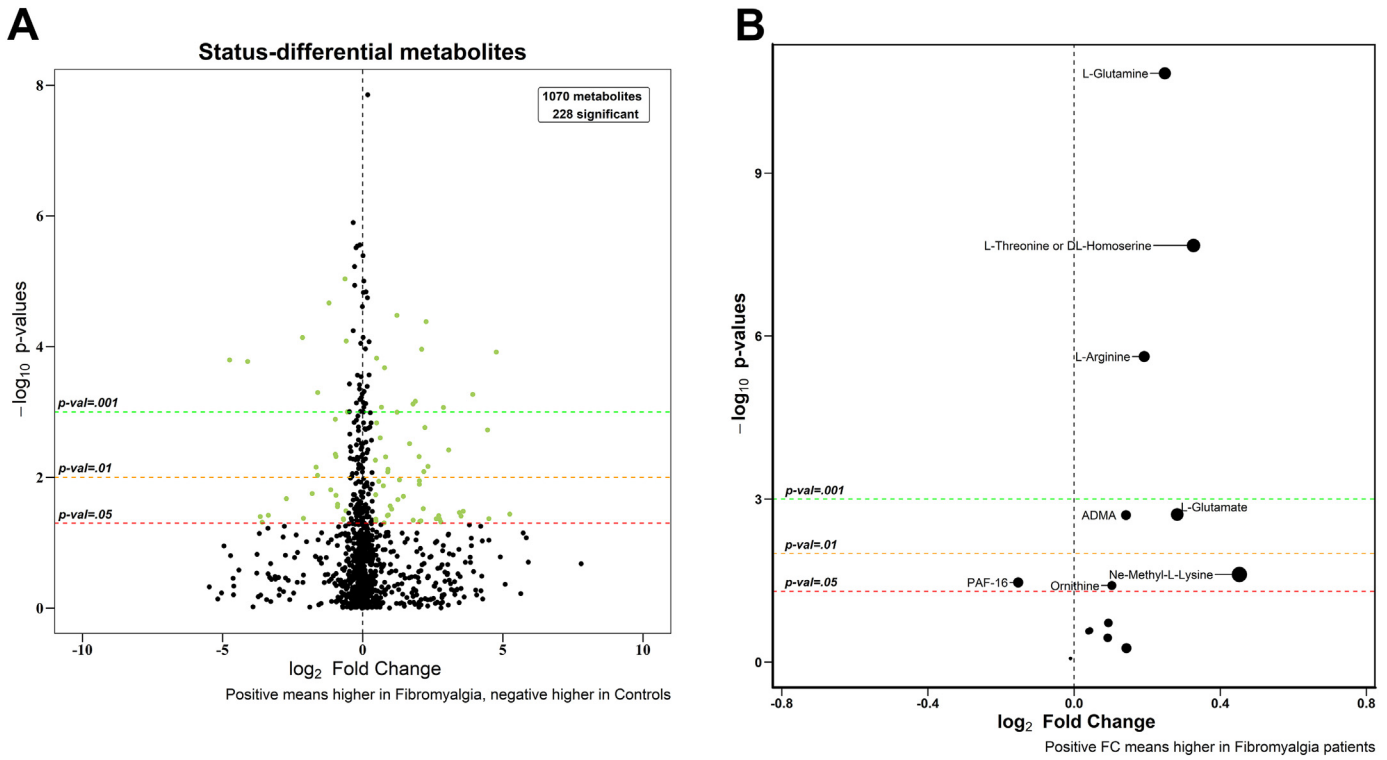


Fig. 4. Univariate metabolomics analysis. (A) Volcano plot of 1070 metabolic features detected in serum samples after background subtraction and removal of the features found in <30% of the data or differing between hospitals. (B) Volcano plot of the identified metabolites. Positive $\log_2$ FC indicates increased abundance in fibromyalgia patients. All *p-values* were adjusted using the Bonferroni method.

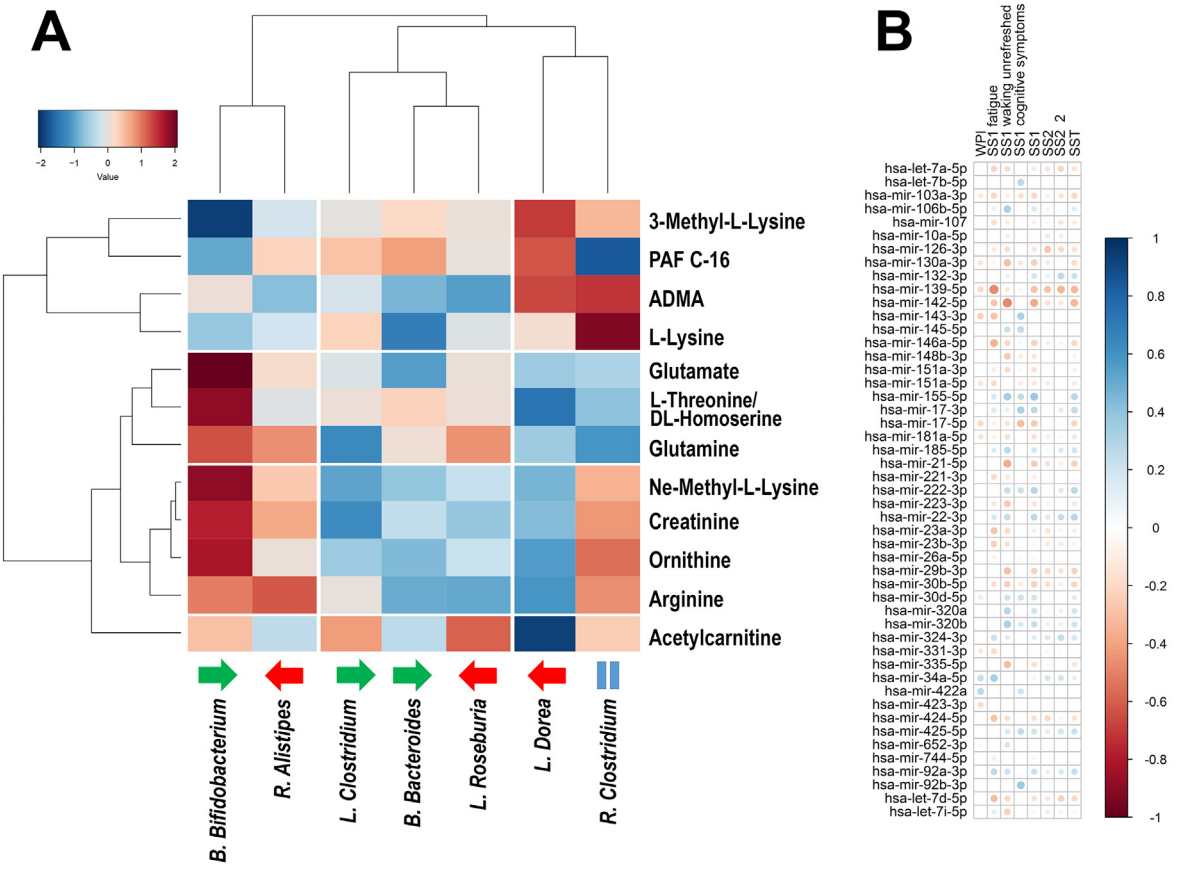
altered in fibromyalgia was tentatively identified as L-threonine or DL-homoserine (Fig. 4B). We could not discriminate between these two metabolites as they are structurally similar and have the same molecular mass and fragmentation pattern in LC-MS. We also analysed the metabolites described in the literature, such as creatinine [31,55], platelet activating factor [11] and acetylcarnitine [25]. To infer alterations in the biological processes and metabolic and functional pathways associated with the differentially expressed metabolites, we used MetScape [44] and Ingenuity Pathway Analysis® (QIAGEN) (IPA). The analyses showed that cell signalling and inflammatory and hypersensitivity responses were the most relevant biological processes. The most represented metabolic pathways were arginine, nitric oxide (NO) and glutamate metabolism.

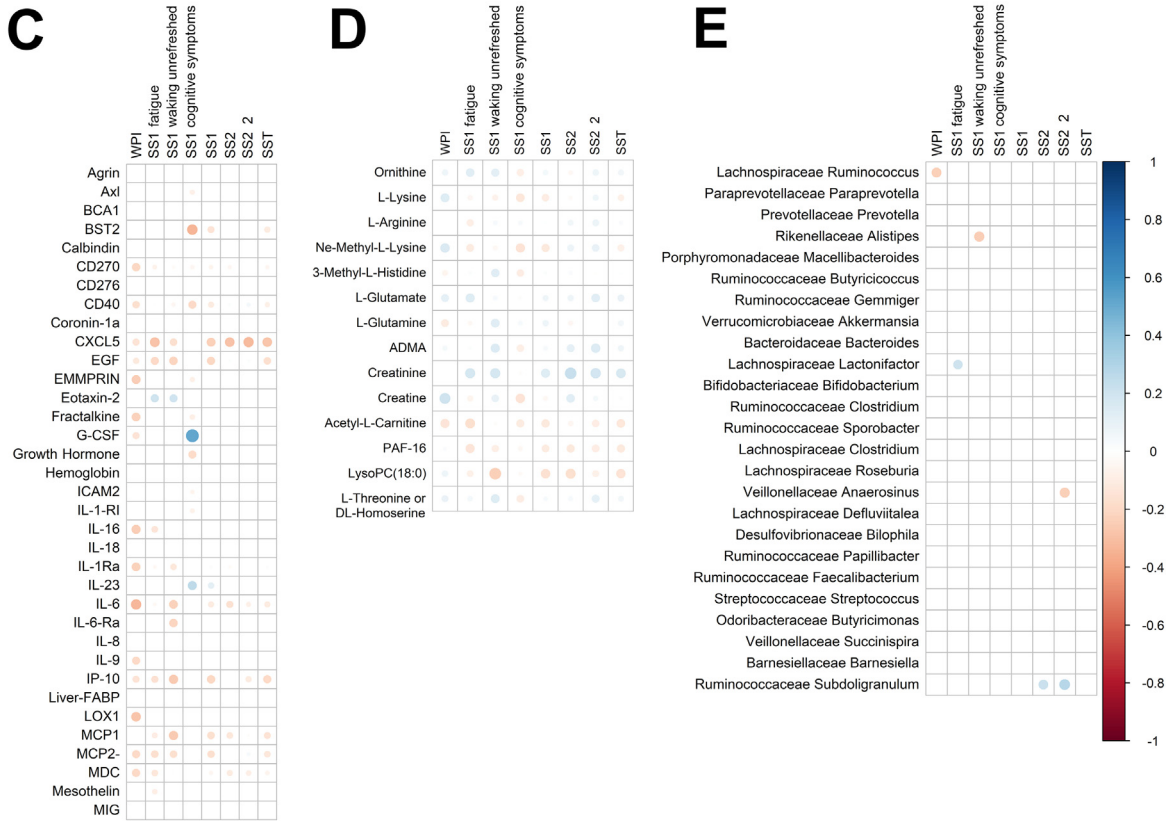
To study the potential dependencies between microbiome composition and the host metabolism and metabolome, we examined the correlations between the two datasets. We computed the Spearman's correlation coefficient for the full set of metabolomics features and microbiome variables. We did not see any clear association patterns between the two complete datasets (Supplementary Fig. S4). We also constructed a heatmap of the scaled correlations between the bacteria whose abundance was changed in fibromyalgia and the identified metabolites (Fig. 5A). Metabolites were grouped into two clusters, depending on the correlations. These were seen mainly with for genera *Bifidobacterium* and *Dorea*, which behaved in an opposite manner. The first cluster contained 4 metabolites (3-methyl-L-Lysine, PAF C-16, ADMA, L-Lysine). The second cluster was formed by 8 metabolites (glutamate, L-threonine/DL-homoserine, glutamine, Nε-methyl-L-Lysine, creatinine, ornithine, arginine and acetylcarnitine), although the metabolite acetylcarnitine behaved differently from the other metabolites in this cluster. *Bifidobacterium*, whose abundance was reduced in fibromyalgia patients, correlated negatively with the first metabolite cluster and positively with the second one. *Dorea*, with increased abundance in fibromyalgia patients, correlated positively with the first metabolite cluster and negatively with the second one.

Finally, we checked, using Virtual Metabolic Human [65] database, whether the different metabolites were produced by the differentially abundant bacteria. We also wanted to study whether they were made by the genera for which we found most correlations (Fig. 5A). Thus, we limited the search to *Bifidobacterium* and *Dorea* genera. For glutamate, we identified the metabolites upstream and downstream of its production/degradation. For lysine, threonine, homoserine, glutamine, ornithine and arginine (and their modifications), we found that the metabolites themselves, their precursors and degradation products might had been produced by bacteria. No bacterial associations were found for creatinine, PAF C-16, ADMA and acetylcarnitine, consequently suggesting that their origin was exclusively human.

3.5. Serum factors and miRNA analysis for a subset of samples

A subset of the samples ($n = 72$; $n_C = 36$ controls and $n_F = 36$ fibromyalgia samples) was used to perform multiplex assays for different serum molecules, including miRNAs and cytokines. For the multiplex design, we used 70 molecules and 68 miRNAs that have been associated with fibromyalgia and/or chronic pain. The protein content assays and the miRNAs analysis did not show any differences between the fibromyalgia and the control groups. We observed statistically significant differences for ten serum proteins: PCSK9, mesothelin, BST2 (↑), procalcitonin, Axl, myoglobin, MIG, TNF-alpha, ICAM2 and IL-9 (↓) with fold changes ranging from 0.76 (lower level in patients) for IL-9 to 1.07 for BST2 (Supplementary Fig. S5A). However, the levels of only one miRNA differed significantly between the fibromyalgia patients and the control group, the hsa-miR-335-5p (Supplementary Fig. S5B). Predicted target genes were obtained using miRWalk 2.0 database [20]; they were selected if they mapped to at least 8 of the 12 database options. The enrichment of the miRNA targets was performed using ConsensuspathDB [42], selecting the targets with a *p-value* < .01. Notably, we identified several pathways related to signalling dedicated to


A
Value
3-Methyl-L-Lysine
PAF C-16
ADMA
L-Lysine
Glutamate
L-Threonine/ DL-Homoserine
Glutamine
Ne-Methyl-L-Lysine
Creatinine
Ornithine
Arginine
Acetylcarnitine
B. Bifidobacterium
R. Alistipes
L. Clostridium
B. Bacteroides
L. Roseburia
L. Dorea
R. Clostridium
B
WPI
SS1 fatigue
SS1 waking unrefreshed
SS1 cognitive symptoms
SS1
SS2
SS2 2
SST


C
D
E
WPI
SS1 fatigue
SS1 waking unrefreshed
SS1 cognitive symptoms
SS1
SS2
SS2 2
SST
Agrin
Axl
BCA1
BST2
Calbindin
CD270
CD276
CD40
Coronin-1a
CXCL5
EGF
EMMPRIN
Eotaxin-2
Fractalkine
G-CSF
Growth Hormone
Hemoglobin
ICAM2
IL-1-RI
IL-16
IL-18
IL-1Ra
IL-23
IL-6
IL-6-Ra
IL-8
IL-9
IP-10
Liver-FABP
LOX1
MCP1
MCP2-
MDC
Mesothelin
MIG
Ornithine
L-Lysine
L-Arginine
Ne-Methyl-L-Lysine
3-Methyl-L-Histidine
L-Glutamate
L-Glutamine
ADMA
Creatinine
Creatine
Acetyl-L-Carnitine
PAF-16
LysoPC(18:0)
L-Threonine or DL-Homoserine
Lachnospiraceae Ruminococcus
Paraprevotellaceae Paraprevotella
Prevotellaceae Prevotella
Rikenellaceae Alistipes
Porphyromonadaceae Macellibacteroides
Ruminococcaceae Butyricicoccus
Ruminococcaceae Gemmiger
Verrucomicrobiaceae Akkermansia
Bacteroidaceae Bacteroides
Lachnospiraceae Lactonifactor
Bifidobacteriaceae Bifidobacterium
Ruminococcaceae Clostridium
Ruminococcaceae Sporobacter
Lachnospiraceae Clostridium
Lachnospiraceae Roseburia
Veillonellaceae Anaerosinus
Lachnospiraceae Defluviitalea
Desulfovibrionaceae Bilophila
Ruminococcaceae Papillibacter
Ruminococcaceae Faecalibacterium
Streptococcaceae Streptococcus
Odoribacteraceae Butyricimonas
Veillonellaceae Succinispira
Barnesiellaceae Barnesiella
Ruminococcaceae Subdoligranulum

gene regulation processes. The complete results are provided in SUPPLEMENTARY TABLE S4.

3.6. Correlations between omics data and clinical data

To determine which differences could be associated with the disease, we examined the correlations between different diagnostic indexes obtained for the fibromyalgia patients and the omics data (Fig. 5B, C, D and E). Notably, miRNA data constituted the omics dataset most correlated with pain indicators (Fig. 5B), followed by the results of serum protein profiling (Fig. 5C). Metabolomics also showed a considerable number of correlations with several pain indexes (Fig. 5D). The microbiome composition (at the genus level) (Fig. 5E) was the omics dataset with the weakest correlation with pain indicators.

We also considered possible effects of medication on the observed differences between the patient and control samples. We checked whether the samples clustered depending on the drug regimen followed. However, we did not find any clusters of samples (neither for serum factors nor for miRNAs) that could be associated with a specific drug or drug combination. We also checked whether any data correlated with distinct drug types; no such correlation was observed (data not shown).

3.7. Modelisation of microbiome, metabolomics, cytokine and miRNA datasets

We combined the four datasets of the 71 samples ($n_C = 36$, $n_F = 35$) that had all the data. This allowed us to discriminate between the control and fibromyalgia samples when a block sparse PLS-DA model was applied (block sPLS-DA) (Fig. 6A). The analysis of the individual contribution of each dataset to the differences showed that the most correlated datasets were the microbiome composition and the metabolomics data. We also found that the major contributor to the separation of the sample groups was the microbiome dataset, followed by serum metabolomics, proteins and, finally, miRNAs (Fig. 6B and C). In this analysis, we used only the metabolomics dataset containing the identified metabolites ($n = 14$). The sPLS-DA analysis using the whole unidentified metabolomics dataset ($n = 1070$) showed that using the metabolomics dataset improved the discrimination between the two sample groups, becoming the strongest factor distinguishing the patients from controls (Supplementary Fig. S6) although the microbiome showed slightly better predictive ability.

4. Discussion

In this study, we applied an omics approach and identified a set of potential molecular markers (Table 2) for the diagnosis of fibromyalgia.

The gut microbiome analysis revealed two clusters (Fig. 2B), one cluster for fibromyalgia patients (modified 2010 ACR diagnostic criteria) and the other for individuals without any clinical manifestation of fibromyalgia. Both core microbiome and alpha-diversity analyses showed a reduction in the bacterial diversity in the fibromyalgia group. This is in agreement with the report of reduced microbiota diversity in other pain disorders, such as myalgic encephalomyelitis/chronic fatigue syndrome [27]. Interestingly, our fibromyalgia microbiome analysis showed a reduction in the abundance of several bacterial strains associated with healthy microbiome, such as those linked to SCFA production (*Bifidobacterium, Eubacterium and* Lachnospiraceae) [40,64,77,94], and/or to the reduction in Firmicutes phylum OTUs ([75]; Human Microbiome Project Consortium et al., 2012; [51]), suggesting dysbiosis events in fibromyalgia patients. Currently, there is no consensus on the use of the term "dysbiosis" or its meaning [35]. Thus, we would like to clarify that we refer to alterations in microbiome composition linked to disease (either causing the disease or appearing as its consequence). Dysbiosis events are also associated with the disruption of the intestinal barrier; this increases the interactions of bacteria with the immune system of the host, producing local inflammation [41]. This is supported not only by the large proportion of patients reporting abdominal pain (>90%) but also by the number of intestinal diseases considered co-morbidities of fibromyalgia. The maintenance of the intestinal barrier is associated with the production of SCFAs, including butyric acid and butyrate [77]. In fibromyalgia, we found a decrease in the abundance of several members of the *Lachnospiraceae* family, the bacteria involved in butyric acid production [61]. Butyrate, the conjugate base of butyric acid, is produced by a small number of bacteria, including several *Eubacterium* species [64], a genus also underrepresented in fibromyalgia patients. The reduction in the diversity of bacteria, especially of those engaged in the production of protective SCFAs, suggests that this process might be implicated in the development of fibromyalgia. If this is the case, the dysbiosis events, as understood here, should be persistent. Thus, we recognise that multiple time-point data should be acquired and studied; lack of this data is a limitation of our study. We would like to emphasise that this is a pilot study and that a follow-up analysis, which might reinforce our findings, is recommended.

We also found differences between neurotransmitter metabolisms in the patients and control individuals. We detected a significant increase in the serum levels of glutamate in fibromyalgia patients. Moreover, the abundance of bacteria from *Bifidobacterium* and *Lactobacillus* genera (involved in the transformation of glutamate into GABA; [4,8,105] was reduced in the fibromyalgia group. This might contribute to the elevated systemic levels of glutamate. The effect of GABA on the gut-brain axis, via the vagus nerve, has been described by several authors [8,16]. Glutamate affects the development of pain, via glutamatergic synapses [69], and stress can alter the regulation of this pathway [74]. Stress-related events have also been associated with microbiome modifications [8]. The 2010 modified ACR criteria for fibromyalgia diagnosis include several stress-associated symptoms. Whether such elevated systemic levels of glutamate affect the ENS and alter the CNS is still unclear. However, some authors have demonstrated the activation of glutamatergic neurons and glutamate-mediated neurotransmission in the ENS [13,46,50,84]. As a result of a reduction in bacterial diversity, the glutamate might enter the host bloodstream after the disruption of the intestinal barrier by the inflammation caused by the dysbiosis. Interestingly, several patients presented with symptoms associated with IBD as fibromyalgia co-morbidities (irritable bowel syndrome (46%), abdominal pain (13%) and the pain in the upper abdomen (45%), diarrhoea (20%), etc.). The role of microbiome in IBD pathogenesis has been broadly demonstrated [23,86]; a dysregulation of intestinal immune system caused by microbiome alterations may lead to disease [91], as demonstrated by patients presenting T-cell responses against commensal bacteria [73]. Specifically, a reduction in the abundance of Firmicutes phylum bacteria (observed in fibromyalgia patients) has been recurrently associated with IBD pathogenesis and progression [24,63]. These common alterations in microbiome composition could explain some of the most frequent co-morbidites reported by the patients in our study.

Furthermore, it has been shown that the blood-brain barrier increases its permeability after a decrease in the numbers of SCFA-producing bacteria. This alters the tight junction organisation, which

Fig. 5. Heatmap of scaled correlations between the bacteria whose abundance was altered in fibromyalgia and the identified metabolites. The dendrograms were unsupervised. Red arrows mark the bacteria with increased abundance in fibromyalgia, green arrows, with decreased abundance, and "equals" symbol indicates the OTUs with both increased and decreased abundance (A). Omics correlations with indexes used in fibromyalgia diagnostics, as defined by ACR 2010 criteria. Only significant correlations (*p-value* < .05) are coloured. Positive correlations are indicated in red and negative correlations, in blue. Correlations between circulating miRNA levels (B), circulating cytokine levels (C), identified serum metabolites (D) and microbiome composition (at genus level) (E).

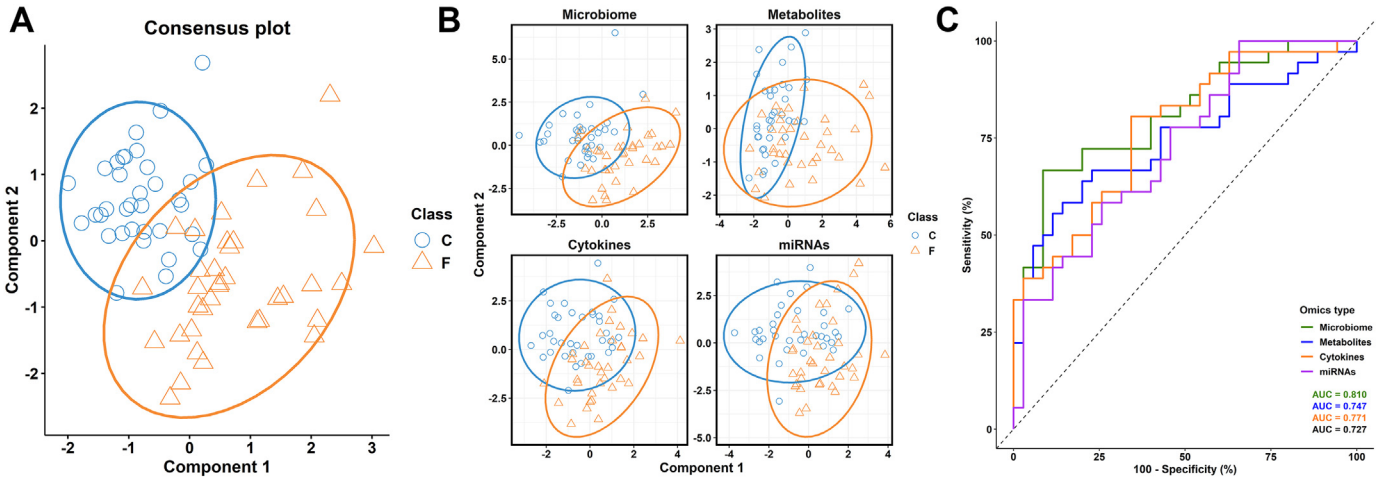


Fig. 6. Multi-omics integration. (A) sPLS-DA consensus plot for the combination of the 4 datasets, showing the nearly complete discrimination of the 71 samples (36 fibromyalgia and 35 control samples). (B) The individual contribution of each dataset to the sPLS-DA final model, in each case showing the score plots for the two first components, indicating the best separation capability for microbiome data, followed by cytokines, metabolomics and miRNAs. (C) ROC curves for each omics dataset, with the Area under the Curve (AUC) values.

can be recovered by colonisation with SCFA-producing bacteria and/or by the administration of these bacterial metabolites [7]. Cytokines can also modify the blood-brain barrier permeability [6,103]. Importantly, glutamate levels increase in the cerebrospinal fluid (CSF) of fibromyalgia patients [85]. These data suggest an important role of this neurotransmitter in the pathogenesis of fibromyalgia. The manner in which the peripheral levels of gut microbiome derived neurotransmitters can affect the brain function is still under debate [84], although several mechanisms have been proposed. Alterations in the blood-brain barrier permeability could modify the interchange of serum metabolites with the brain. Serum levels of 5-HTs are altered in germ-free mice [101,104]. Even though 5-HT itself is not known to cross the blood-brain barrier, its precursor can. The microbiome might alter the 5-HT precursor (e.g. tryptophan) levels, as has been proposed by several authors [67,88]. The same mechanism has been suggested for other gut microbiome neurotransmitters, such as dopamine and GABA [56,84,97].

It is essential to keep in mind the relationship between GABAergic pain inhibition and gender as fibromyalgia is 3 times more prevalent in women than in men [76]. Steroid 17β-estradiol (E2) suppresses the GABAergic inhibition in female rats via a sex-specific oestrogen receptor ERα, mGluR and endocannabinoid-dependent mechanism [92]. This suppression requires the activation of mGluR type I receptors by glutamate [36]. Therefore, in the presence of excess glutamate, as observed here in fibromyalgia patients, the pain inhibition by GABA might be suppressed in female patients by this E2-specific regulation. This might partly explain the increased prevalence of fibromyalgia in the female population.

Table 2
Differences between fibromyalgia and healthy control groups observed using each omics technique (showing alterations in the fibromyalgia patients).

	Increased (↑)	Decreased (↓)
Microbiome	*Dorea* *Roseburia* *Papillibacter* *Subdoligranulum*	*Bifidobacterium* *Eubacterium* Lachnospiraceae (family) *Clostridium* Firmicutes (phylum)
Metabolomics	L-glutamine L-threonine/DL-homoserine L-arginine ADMA L-glutamate Nε-methyl-L-lysine Ornithine	PAF-16
Cytokines	PCSK9 Mesothelin BST2	Procalcitonin Axl-UFO Myoglobin MIG TNF-alpha ICAM2 IL-9
miRNAs	hsa-miR-335-5p	

The functional analysis of the metabolomics dataset showed that the most represented pathways were those dedicated to the metabolism of known neurotransmitters, such as glutamate and serine. Both arginine and ornithine levels, related to the widespread pain in fibromyalgia, increased in the sera of fibromyalgia patients. Consistently, IPA analysis identified several pathways related to arginine, such as arginine degradation (I and II) canonical pathways and proline biosynthesis from arginine. These two metabolites are required for the synthesis of nitric oxide (NO) [31]. NO plays an important role in both acute and chronic pain as it is a mediator of nociception [17]. However, NO contributes not only to nociception; it also mediates in analgesia and increases the effect of morphine on pain inhibition [17]. Here, we also observed a strengthening of this pathway in fibromyalgia patients (by using IPA). The role of NO in fibromyalgia pathogenesis has been studied but without reaching a consensus [72]. Notably, the levels of iNOS isoform increase in female fibromyalgia sufferers in comparison with healthy controls, while the levels of constitutive isoforms (nNOS and eNOS) do not change [59]. It is important to remember that our functional profiling was performed using the results obtained from the serum sample analysis. One of the limitations of this study is the metabolomics analysis, and specifically, the metabolite identification step. We could only identify a small subset of all the metabolic features observed. Thus, the results obtained here are constrained by the relatively small number of identified metabolites. An improved metabolite identification procedure could not only expand the list of potential metabolite biomarkers but also advance the identification of potentially affected biological pathways and functionalities.

Patients afflicted by chronic pain are likely to participate in many different long-term treatments, which could affect their microbiome composition. Differences in diets and lifestyles will also have some effect. Thus, it is difficult to be certain whether the detected alterations in the microbiota are the cause or consequence of fibromyalgia. No association between microbiome composition and drug type was found for fibromyalgia patients. However, it has been demonstrated that clinical drugs have an impact upon microbiome composition; this seems to be true for antibiotic, non-antibiotic [54] and psychotropic [19] drugs. The lack of associations shown here could have been caused by the small number of patients taking medication from a specific drug family and/or by the interactions between different drugs prescribed. Proton

pump inhibitors (PPI), for example, have an anticommensal activity and were taken by nearly 30% of the patients. One study has reported a reduction in the abundance of Lachnospiraceae and Ruminococcaceae in PPI consumers [39], which is consistent with our observations for fibromyalgia patients. Another study obtained similar results and considered in its analysis the decrease in the abundance of *Bifidobacterium* genus in PPI consumers [38]. Both studies have reported a decrease in α-diversity after PPI administration, which is also consistent with our findings. It has been reported that psychotropics target a similar pattern of bacterial species irrespective of the degree of their chemical similarity. This suggests that the anticommensal activity of these drugs is a part of their mechanism of action rather than a secondary effect [19].

We did not observe any microbiome alterations that could be associated with antidepressant drugs, either for the tricyclic antidepressants (taken by 12% of patients) or for the selective serotonin reuptake inhibitors (SSRI), 54% of patients). The antiepileptic drugs (here taken by 29% of patients), such as lithium or valproate, do not have a significant anticommensal activity. However, lithium may increase the relative abundance of Ruminococcaceae and reduce the abundance of Bacteroides, while valproate alters the levels of SCFA [18]; there were also alterations found in fibromyalgia patients. Finally, while no antimicrobial activity has been reported for morphine [83], chronic use of opioids (prescribed to 45% of patients) has been associated with a reduction in Bacteroidaceae (which we also observed in fibromyalgia patients) and Ruminococcaceae [3]. Even though no associations between specific drugs and microbiome composition was found, probiotics could be useful in the treatment of fibromyalgia as they affect the microbiome composition [34]. Notably, several authors have used this approach to treat the chronic fatigue syndrome [82] and one pilot study has examined the effects of probiotics on fibromyalgia patients [80]. The authors have shown some improvements, mainly in depression symptoms and impulsive behaviour, in comparison with the placebo group [81].

5. Conclusions

To the best of our knowledge, this is the first study to report differences between the microbiome composition of fibromyalgia patients and healthy controls. We provided a list of these differences and reported the alterations in the levels of various molecules in the fibromyalgia sufferers, which might be useful as diagnostic biomarkers. We examined the functionality of these molecules and found that the most altered metabolic pathways were related to neurotransmitters, such as glutamate and nitric oxide. We checked possible interactions between the gut microbiome and serum metabolome; our analysis found several individual correlations between the two datasets. We also demonstrated that the combined microbiome and serum metabolome analyses could discriminate between the fibromyalgia patients and control individuals. Thus, we report a new set of molecules and bacteria that might improve the diagnosis process, compensating for the current lack of objective biomarkers. Our results should help to shed some new light on the pathogenesis of this disease, provide biomarkers within a biological framework and improve our knowledge of this relatively unknown disease.

Ethics approval and consent to participate

All donors signed the informed consent form, and the study was approved by the ethical committee (CEIC-PI2016037).

Availability of data and materials

Sequencing data was submitted to the ENA repository, under the Project Accession code PRJEB27227.

We used free, open-access software, except for SIMCA-P+ 12.0.1 (Umetrics AB, Umeå, Sweden) and Ingenuity Pathway Analysis® from QIAGEN, which require a license. Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Juan M. Falcón-Pérez (jfalcon@cicbiogune.es). Sharing of patient data may be restricted due to anonymity considerations.

Funding

This project was funded by the Basque Government's Health Department (ref. 2015111149 (2016-2018)).

Author contributions

Experimental design: Marc Clos-Garcia, Naiara Andrés, Gorka Fernández-Eulate, Olga Maíz, Adolfo López de Munain and Juan Manuel Falcón-Pérez.

Patient recruitment and sample collection: Naiara Andrés, Gorka Fernández-Eulate, Alejandro Valero, Nerea Errazquin, María Cristina Gómez Vallejo, Leila Govillard, Ana María Callejo Orcasitas, Olga Maíz and Adolfo López de Munain.

Sample processing and data analysis: Marc Clos-Garcia, Leticia Abecia, José L. Lavín, Sebastiaan van Liempd, Diana Cabrera, Félix Royo, Esperanza Gónzalez, Ana M. Aransay, Michael R Tackett and Genesis Tejada.

Manuscript writing: Marc Clos-Garcia, Leticia Abecia, José L. Lavín, Naiara Andrés, Gorka Fernández-Eulate and Juan Manuel Falcón-Pérez.

Discussion and review of the manuscript: Marc Clos-Garcia, Leticia Abecia, José L. Lavín, Juan Anguita, Luis Bujanda, Ana María Callejo Orcasitas, Ana M. Aransay, Olga Maíz, Adolfo López de Munain and Juan Manuel Falcón-Pérez.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ebiom.2019.07.031.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgements

We are thankful to the Basque Biobank for Research (BIOEF) for the acquisition, maintenance and distribution of faeces, blood and DNA samples and to Bizi Bide association (Asociación Guipuzcoana de Fibromialgia y Síndrome de Fatiga Crónica) for their help in patient recruitment.

References

[1] Ablin J, et al. Frequency of axial spondyloarthropathy among patients suffering from fibromyalgia. A magnetic resonance imaging study applying the assessment of spondylo-arthritis international society classification criteria. Arthritis Rheum Abstr 2013;65:128. https://doi.org/10.1007/s11606-010-1338-5.
[2] Ablin JN, Cohen H, Buskila D. Mechanisms of disease: genetics of fibromyalgia. Nat Clin Pract Rheumatol 2006;2(12):671–8. https://doi.org/10.1038/ncprheum0349.
[3] Acharya C, et al. Chronic opioid use is associated with altered gut microbiota and predicts readmissions in patients with cirrhosis. Aliment Pharmacol Ther 2017;45(2):319–31. https://doi.org/10.1111/apt.13858.
[4] Barrett E, et al. γ-Aminobutyric acid production by culturable bacteria from the human intestine. J Appl Microbiol 2012;113(2):411–7. https://doi.org/10.1111/j.1365-2672.2012.05344.x.
[5] Bonder MJ, et al. The effect of host genetics on the gut microbiome. Nat Genet 2016;48:1407 Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. Available at: https://doi.org/10.1038/ng.3663.
[6] Boveri M, et al. Highly purified lipoteichoic acid from gram-positive bacteria induces in vitro blood-brain barrier disruption through glia activation: role of pro-inflammatory cytokines and nitric oxide. Neuroscience 2006;137(4):1193–209. https://doi.org/10.1016/j.neuroscience.2005.10.011.
[7] Braniste V, et al. The gut microbiota influences blood-brain barrier permeability in mice. Sci Transl Med 2014;6(263). https://doi.org/10.1126/scitranslmed.3009759.
[8] Bravo JA, et al. Ingestion of Lactobacillus strain regulates emotional behavior and central GABA receptor expression in a mouse via the vagus nerve. Proc Natl Acad Sci 2011;108(38):16050–5. https://doi.org/10.1073/pnas.1102999108.
[9] Buskila D, et al. Fibromyalgia in hepatitis C virus infection. Arch Intern Med 1997;157:2497. https://doi.org/10.1001/archinte.1997.00440420129014.

[10] Buskila D, Atzeni F, Sarzi-Puttini P. Etiology of fibromyalgia: the possible role of infection and vaccination. Autoimmun Rev 2008;8(1):41–3. https://doi.org/10.1016/j.autrev.2008.07.023.

[11] Caboni P, et al. Metabolomics analysis and modeling suggest a lysophosphocholines-PAF receptor interaction in fibromyalgia. PLoS One 2014;9(9):1–8. https://doi.org/10.1371/journal.pone.0107626.

[12] Caporaso JG, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010;7(5):335–6. https://doi.org/10.1038/nmeth0510-335 Nature Publishing Group.

[13] Chen WP, Kirchgessner AL. Activation of group II mGlu receptors inhibits voltage-gated Ca2+ currents in myenteric neurons. Am J Physiol Gastrointest Liver Physiol 2002;283(6):G1282–9. https://doi.org/10.1152/ajpgi.00216.2002.

[14] Cohen H, et al. Confirmation of an association between fibromyalgia and serotonin transporter promoter region (5-HTTLPR) polymorphism, and relationship to anxiety-related personality traits. Arthritis Rheum 2002;46(3):845–7. https://doi.org/10.1002/art.10103.

[15] Collins SM, Surette M, Bercik P. The interplay between the intestinal microbiota and the brain. Nat Rev Microbiol 2012;10(11):735–42. https://doi.org/10.1038/nrmicro2876 Nature Publishing Group.

[16] Cryan JF, O'Mahony SM. The microbiome-gut-brain axis: from bowel to behavior. Neurogastroenterol Motil 2011;23(3):187–92. https://doi.org/10.1111/j.1365-2982.2010.01664.x.

[17] Cury Y, et al. Pain and analgesia: the dual effect of nitric oxide in the nociceptive system. Nitric Oxide 2011;25(3):243–54. https://doi.org/10.1016/j.niox.2011.06.004 Elsevier Inc.

[18] Cussotto S, et al. Differential effects of psychotropic drugs on microbiome composition and gastrointestinal function. Psychopharmacology 2018. https://doi.org/10.1007/s00213-018-5006-5.

[19] Cussotto S, et al. Psychotropics and the microbiome: a chamber of secrets…. Psychopharmacology 2019. https://doi.org/10.1007/s00213-019-5185-8.

[20] Dweep H, Gretz N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. Nat Methods 2015;12(8):697. https://doi.org/10.1038/nmeth.3485.

[21] Foerster BR, et al. Reduced insular γ-aminobutyric acid in fibromyalgia. Arthritis Rheum 2012;64(2):579–83. https://doi.org/10.1002/art.33339.

[22] Forsythe P, et al. Mood and gut feelings. Brain Behav Immun 2010;24(1):9–16. https://doi.org/10.1016/j.bbi.2009.05.058 Elsevier Inc.

[23] Frank DN, et al. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proceedings of the National Academy of Sciences of the United States of America. PNAS 2007;104(34):13780–5 (www.pnas.org%02cgi%02doi%0210.1073%02pnas.0706625104).

[24] Frank DN, et al. Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. Inflamm Bowel Dis 2011;17(1):179–84. https://doi.org/10.1002/ibd.21339.

[25] Freidin MB, et al. Metabolomic markers of fatigue: association between circulating metabolome and fatigue in women with chronic widespread pain'. Biochim Biophys Acta 2018;1864(2):601–6. https://doi.org/10.1016/j.bbadis.2017.11.025 Elsevier.

[26] Gallai V, et al. Glutamate and nitric oxide pathway in chronic daily headache: evidence from cerebrospinal fluid. Cephalalgia 2003;23(3):166–74. https://doi.org/10.1046/j.1468-2982.2003.00552.x.

[27] Giloteaux L, et al. Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome. Microbiome 2016;4(1):30. https://doi.org/10.1186/s40168-016-0171-4.

[28] Goodrich JK, et al. Genetic determinants of the gut microbiome in UK twins. Cell Host Microbe 2016;19(5):731–43. https://doi.org/10.1016/j.chom.2016.04.017 Elsevier Inc.

[29] Guijas C, et al. METLIN: a technology platform for identifying knowns and unknowns. Anal Chem 2018;90(5):3156–64. https://doi.org/10.1021/acs.analchem.7b04424.

[30] Gürsoy S, et al. Significance of catechol-O-methyltransferase gene polymorphism in fibromyalgia syndrome. Rheumatol Int 2003;23:104–7. https://doi.org/10.1007/s00296-002-0260-5.

[31] Hadrévi J, et al. Systemic differences in serum metabolome: a cross sectional comparison of women with localised and widespread pain and controls. Sci Rep 2015;5(March):1–13. https://doi.org/10.1038/srep15925 Nature Publishing Group.

[32] Häuser W, et al. Emotional, physical, and sexual abuse in fibromyalgia syndrome: a systematic review with meta-analysis. Arthritis Care Res 2011;63(6):808–20. https://doi.org/10.1002/acr.20328.

[33] Häuser W, et al. Fibromyalgia'. Nat Rev Dis Primers 2015(August):15022. https://doi.org/10.1038/nrdp.2015.22.

[34] Hemarajata P, Versalovic J. Effects of probiotics on gut microbiota: mechanisms of intestinal immunomodulation and neuromodulation. Therap Adv Gastroenterol 2013;6(1):39–51. https://doi.org/10.1177/1756283X12459294.

[35] Hooks KB, O'Malley MA. Dysbiosis and its discontents. mBio 2017;8(5):1–11. https://doi.org/10.1128/mbio.01492-17.

[36] Huang GZ, Woolley CS. Estradiol acutely suppresses inhibition in the Hippocampus through a sex-specific endocannabinoid and mGluR-dependent mechanism. Neuron 2012;74(5):801–8. https://doi.org/10.1016/j.neuron.2012.03.035 Elsevier Inc.

[37] Human Microbiome Project Consortium, T. et al. Structure, function and diversity of the healthy human microbiome. Nature 2012;486(7402):207–14. https://doi.org/10.1038/nature11234 Nature Publishing Group.

[38] Imhann F, et al. Proton pump inhibitors affect the gut microbiome. Gut 2016;65(5):740–8. https://doi.org/10.1136/gutjnl-2015-310376.

[39] Jackson MA, et al. Proton pump inhibitors alter the composition of the gut microbiota. Gut 2016;65(5):749–56. https://doi.org/10.1136/gutjnl-2015-310861.

[40] Jandhyala SM, et al. Role of the normal gut microbiota. World J Gastroenterol 2015;21(29):8836–47. https://doi.org/10.3748/wjg.v21.i29.8787.

[41] Kamada N, et al. Role of the gut microbiota in immunity and inflammatory disease. Nat Rev Immunol 2013;13(5):321–35. https://doi.org/10.1038/nri3430.

[42] Kamburov A, et al. The ConsensusPathDB interaction database: 2013 update'. Nucleic Acids Res 2013;41:793–800. https://doi.org/10.1093/nar/gks1055.

[43] Karlsson F, et al. Assessing the human gut microbiota in metabolic diseases. Diabetes 2013;62(10):3341–9. https://doi.org/10.2337/db13-0844.

[44] Karnovsky A, et al. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. Bioinformatics 2012;28(3):373–80. https://doi.org/10.1093/bioinformatics/btr661.

[45] Kennedy PJ, et al. Irritable bowel syndrome: a microbiome-gut-brain axis disorder? World J Gastroenterol 2014;20(39):14105–25. https://doi.org/10.3748/wjg.v20.i39.14105.

[46] Kirchgessner AL, Liu MT, Alcantara F. Excitotoxicity in the enteric nervous system. J Neurosci 1997;17(22):8804–16. https://doi.org/10.1016/S0896-6273(02)00778-X.

[47] Knight R, et al. The microbiome and human biology. Annu Rev Genomics Hum Genet 2017;183(March):65–86. https://doi.org/10.1146/annurev-genom-083115.

[48] Lahti L, Shetty S. microbiome R Package. Available at: http://microbiome.ogithub.io; 2018.

[49] Li K, Bihan M, Methé BA. Analyses of the stability and core taxonomic memberships of the human microbiome. PLoS One 2013;8(5). https://doi.org/10.1371/journal.pone.0063139.

[50] Liu MT, et al. Glutamatergic enteric neurons. J Neurosci 1997;17(12):4764–84. https://doi.org/10.1523/JNEUROSCI.17-12-04764.1997.

[51] Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Genome Med 2016;8(1):1–11. https://doi.org/10.1186/s13073-016-0307-y.

[52] Love MI, Anders S, Huber W. DESeq2 package__Differential Analysis of Count Data; 2014. https://doi.org/10.1101/002832.

[53] Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 2011;27(21):2957–63. https://doi.org/10.1093/bioinformatics/btr507.

[54] Maier L, et al. Extensive impact of non-antibiotic drugs on human gut bacteria. Nature 2018. https://doi.org/10.1038/nature25979.

[55] Malatji BG, et al. A diagnostic biomarker profile for fibromyalgia syndrome based on an NMR metabolomics study of selected patients and controls. BMC Neurol 2017;17(1):1–15. https://doi.org/10.1186/s12883-017-0863-9.

[56] Matsumoto M, et al. Impact of intestinal microbiota on intestinal luminal metabolome. Sci Rep 2012;2:1–10. https://doi.org/10.1038/srep00233.

[57] Mazzoli R, Pessione E. The neuro-endocrinological role of microbial glutamate and GABA signaling. Front Microbiol 2016;7(NOV):1–17. https://doi.org/10.3389/fmicb.2016.01934.

[58] McHardy IH, et al. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. Microbiome 2013;1(1):1–19. https://doi.org/10.1186/2049-2618-1-17.

[59] McIver KL, et al. NO-mediated alterations in skeletal muscle nutritive blood flow and lactate metabolism in fibromyalgia. Pain 2006;120(1–2):161–9. https://doi.org/10.1016/j.pain.2005.10.032.

[60] McMurdie PJ, Holmes S. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 2013;8(4). https://doi.org/10.1371/journal.pone.0061217.

[61] Meehan CJ, Beiko RG. A phylogenomic view of ecological specialization in the lachnospiraceae, a family of digestive tract-associated bacteria. Genome Biol Evol 2014;6(3):703–13. https://doi.org/10.1093/gbe/evu050.

[62] Milligan ED, Watkins LR. Pathological and protective roles of glia in chronic pain. Nat Rev Neurosci 2009;10(1):23–36. https://doi.org/10.1038/nrn2533.

[63] Morgan XC, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol 2012;13(9):R79. https://doi.org/10.1186/gb-2012-13-9-r79.

[64] Morrison DJ, Preston T. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. Gut Microbes 2016;7(3):189–200. https://doi.org/10.1080/19490976.2015.1134082 Taylor & Francis.

[65] Noronha A, et al. The virtual metabolic human database: integrating human and gut microbiome metabolism with nutrition and disease. Nucleic Acids Res 2019;47(D1):D614–24. https://doi.org/10.1093/nar/gky992.

[66] Nugent JL, et al. Altered tissue metabolites correlate with microbial dysbiosis in colorectal adenomas. J Proteome Res 2014;13:1921–9. https://doi.org/10.1021/pr4009783.

[67] O'Mahony SM, et al. Serotonin, tryptophan metabolism and the brain-gut-microbiome axis. Behav Brain Res 2015;277:32–48. https://doi.org/10.1016/j.bbr.2014.07.027 Elsevier B.V.

[68] Offenbaecher M, et al. Possible association of fibromyalgia with a polymorphism in the serotonin transporter gene regulatory region. Arthritis Rheum 1999;42(11):2482–8. https://doi.org/10.1126/science.274.5292.1527.

[69] Osikowicz M, Mika J, Przewlocka B. The glutamatergic system as a target for neuropathic pain relief. Exp Physiol 2013;98(2):372–84. https://doi.org/10.1113/expphysiol.2012.069922.

[70] Pedersen HK, et al. Human gut microbes impact host serum metabolome and insulin sensitivity. Nature 2016;535(7612):376–81. https://doi.org/10.1038/nature18646 Nature Publishing Group.

[71] Peres MFP, et al. Cerebrospinal fluid glutamate levels in chronic migraine. Cephalalgia 2004;24(9):735–9. https://doi.org/10.1111/j.1468-2982.2004.00750.x.

[72] Pernambuco AP, et al. Involvement of oxidative stress and nitric oxide in fibromyalgia pathophysiology: a relationship to be elucidated. Fibromyalgia 2016;1(1):1–7. https://doi.org/10.4172/foa.1000105.

[73] Pirzer U, et al. Reactivity of infiltrating T lymphocytes with microbial antigens in Crohn's disease. Lancet 1991;338(8777):1238–9. https://doi.org/10.1016/0140-6736(91)92104-A.
[74] Popoli M, et al. The stressed synapse: the impact of stress and glucocorticoids on glutamate transmission. Nat Rev Neurosci 2012;13(1):22–37. https://doi.org/10.1038/nrn3138 Nature Publishing Group.
[75] Qin J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010;464(7285):59–65. https://doi.org/10.1038/nature08821.
[76] Queiroz LP. Worldwide epidemiology of fibromyalgia. Curr Pain Headache Rep 2013;17(8):356. https://doi.org/10.1007/s11916-013-0356-5.
[77] Ríos-Covián D, et al. Intestinal short chain fatty acids and their link with diet and human health. Front Microbiol 2016;7(FEB):1–9. https://doi.org/10.3389/fmicb.2016.00185.
[78] Rivera J, et al. Fibromyalgia-associated hepatitis C virus infection. Br J Rheumatol 1997;36:981–5. https://doi.org/10.1093/rheumatology/36.9.981.
[79] Rohart F, et al. mixOmics: an R package for 'omics feature selection and multiple data integration'. PLoS Comput Biol 2017;13(11):e1005752. https://doi.org/10.1371/journal.pcbi.1005752.
[80] Roman P, et al. Probiotics for fibromyalgia: study design for a pilot double-blind, randomized controlled tria. Nutr Hosp 2017;34(5):1246–51. https://doi.org/10.3305/nh.2013.28.sup4.6783.
[81] Roman P, Estévez AF, et al. A pilot randomized controlled trial to explore cognitive and emotional effects of probiotics in fibromyalgia. Sci Rep 2018;8(1):1–9. https://doi.org/10.1038/s41598-018-29388-5.
[82] Roman P, Carrillo-Trabalón F, et al. Are probiotic treatments useful on fibromyalgia syndrome or chronic fatigue syndrome patients? A systematic review. Benefic Microbes 2018;9(4):603–11. https://doi.org/10.3920/BM2017.0125.
[83] Rosenberg PH, Renkonen OV. Antimicrobial activity of bupivacaine and morphine. Anesthesiology 1985;62:178–9.
[84] Sampson TR, Mazmanian SK. Control of brain development, function, and behavior by the microbiome. Cell Host Microbe 2015;17(5):565–76. https://doi.org/10.1016/j.chom.2015.04.011.
[85] Sarchielli P, et al. Sensitization, glutamate, and the link between migraine and fibromyalgia. Curr Pain Headache Rep 2007;11(5):343–51. https://doi.org/10.1007/s11916-007-0216-2.
[86] Sartor RB. Microbial influences in inflammatory bowel diseases. Gastroenterology 2008;134(2):577–94. https://doi.org/10.1053/j.gastro.2007.11.059.
[87] Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics 2011;27(6):863–4. https://doi.org/10.1093/bioinformatics/btr026.
[88] Sharon G, et al. Specialized metabolites from the microbiome in health and disease. Cell Metab 2014;20(5):719–30. https://doi.org/10.1016/j.cmet.2014.10.016 Elsevier Inc.
[89] Sharon G, et al. The central nervous system and the gut microbiome. Cell 2016;167(4):915–32. https://doi.org/10.1016/j.cell.2016.10.027 Elsevier Inc.
[90] Singh A, et al. DIABLO: from multi-omics assays to biomarker discovery, an integrative approach. bioRxiv 2018. https://doi.org/10.1101/067611.
[91] Strober W, Fuss I, Mannon P. The fundamental basis of inflammatory bowel disease. J Clin Investig 2007;117(3):514–21. https://doi.org/10.1172/JCI30587.514.
[92] Tabatadze N, et al. Sex differences in molecular signaling at inhibitory synapses in the Hippocampus. J Neurosci 2015;35(32):11252–65. https://doi.org/10.1523/JNEUROSCI.1067-15.2015.
[93] Tackett MR, Diwan I. Using FirePlex™ particle technology for multiplex MicroRNA profiling without RNA purification. Methods Mol Biol 2017;1654:209–19. https://doi.org/10.1007/978-1-4939-7231-9_14.
[94] Tan J, et al. The role of short-chain fatty acids in health and disease. Advances in Immunology. 1st ed. Elsevier Inc.; 2014. https://doi.org/10.1016/B978-0-12-800100-4.00003-9.
[95] Uçeyler N, Häuser W, Sommer C. Systematic review with meta-analysis: cytokines in fibromyalgia syndrome. Bmc Musculoskelet Di 2011;12(1):245. https://doi.org/10.1186/1471-2474-12-245.
[96] Vandesompele J, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 2002;3(7). https://doi.org/10.1186/gb-2002-3-7-research0034 p. research0034.1.
[97] Velagapudi VR, et al. The gut microbiota modulates host energy and lipid metabolism in mice. J Lipid Res 2009;51(5):1101–12. https://doi.org/10.1194/jlr.m002774.
[98] Wang J, et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. Nat Genet 2016;48:1396 Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. Available at: https://doi.org/10.1038/ng.3695.
[99] Wang Q, et al. Host and microbiome multi-omics integration: applications and methodologies. Biophys Rev 2019;11(1):55–65. https://doi.org/10.1007/s12551-018-0491-7.
[100] Weir TL, et al. Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. PLoS One 2013;8(8). https://doi.org/10.1371/journal.pone.0070803.
[101] Wikoff WR, et al. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. Proc Natl Acad Sci 2009;106(10):3698–703. https://doi.org/10.1073/pnas.0812874106.
[102] Wolfe F, et al. The American College of Rheumatology preliminary diagnostic criteria for fibromyalgia and measurement of symptom severity. Arthritis Care Res 2010;62(5):600–10. https://doi.org/10.1002/acr.20140.
[103] Wong D, Dorovini-Zis K, Vincent SR. Cytokines, nitric oxide, and cGMP modulate the permeability of an in vitro model of the human blood-brain barrier. Exp Neurol 2004;190(2):446–55. https://doi.org/10.1016/j.expneurol.2004.08.008.
[104] Yano JM, et al. Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis. Cell 2015;161(2):264–76. https://doi.org/10.1016/j.cell.2015.02.047 Elsevier.
[105] Yunes RA, et al. GABA production and structure of gadB/gadC genes in Lactobacillus and Bifidobacterium strains from human microbiota. Anaerobe 2016;42:197–204. https://doi.org/10.1016/j.anaerobe.2016.10.011 Elsevier Ltd.
[106] Zubieta JK, et al. COMT val158 genotype affects μ-opioid neurotransmitter responses to a pain stressor. Science 2003;299(5610):1240–3. https://doi.org/10.1126/science.1078546.
[107] Zych K, et al. SIAMCAT: Statistical Inference of Associations Between Microbial Communities and host phenoTypes; 2018.
[108] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2019https://www.R-project.org/.

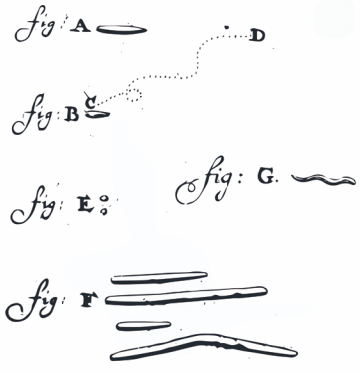


fig: A
D
fig: B
C
fig: E.
fig: G.
fig: F