UNIVERSIDAD DE CÓRDOBA

Departamento de Biología Celular, Fisiología e Inmunología

Programa de doctorado en Biomedicina

TESIS DOCTORAL

Functional analysis and modeling of cellular signaling circuits with transcriptomic and proteomic data

---

Análisis funcional y modelado de circuitos de señalización celular con datos transcriptómicos y proteómicos

Memoria presentada para optar al grado de Doctor en Biomedicina por

Martín Garrido Rodríguez-Córdoba

<u>Directores</u>

Marco Antonio Calzado Canale

Joaquín Dopazo Blázquez

Eduardo Muñoz Blanco

Córdoba, Junio de 2021

TITULO: *Functional analysis and modeling of cellular signaling circuits with transcriptomic and proteomic data*

AUTOR: *Martín Garrido Rodríguez-Córdoba*

© Edita: UCOPress. 2021
Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

https://www.uco.es/ucopress/index.php/es/
ucopress@uco.es

TÍTULO DE LA TESIS: Functional analysis and modeling of cellular signaling circuits with transcriptomic and proteomic data

DOCTORANDO: Martín Garrido Rodríguez-Córdoba

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

El documento presentado por el doctorando Martín Garrido Rodríguez-Córdoba, con título “Functional analysis and modeling of cellular signaling circuits with transcriptomic and proteomic data”, corresponde a su trabajo de tesis doctoral realizado en el periodo comprendido entre Septiembre de 2017 y Marzo de 2021. En este trabajo se han cumplido todos los objetivos establecidos al comienzo de la tesis, facilitando el aprendizaje y desarrollo de técnicas computacionales de gran relevancia y utilidad en el campo de la bioinformática aplicada a la investigación biomédica. La ejecución del presente proyecto ha permitido la creación de nuevas metodologías para el modelado de circuitos de señalización celular a partir de datos transcriptómicos y proteómicos, junto a un paquete de software adicional que recopila el estado del arte para la visualización y el análisis estadístico y funcional de los resultados obtenidos. En concreto, el desarrollo de la nueva herramienta para el análisis de la información multi-ómica extraíble de los datos de RNA-Seq, denominada MIGNON, se ha visto reflejado en la publicación de un artículo científico en una revista indexada dentro del primer cuartil de su área. De manera adicional, el paquete de software creado para recopilar el estado del arte en el análisis computacional de datos ómicos se ha usado en múltiples colaboraciones con otros miembros del grupo que han resultado en dos publicaciones en las que el doctorando figura como primer autor, y en otras nueve en las que figura como co-autor, acabando este periodo de formación con un total de doce artículos publicados en revistas científicas. Además de las publicaciones, los resultados obtenidos han sido difundidos a través de su presentación en jornadas y congresos nacionales e internacionales mediante comunicaciones orales y de formato póster.

PUBLICACIONES Y TRABAJOS DERIVADOS DE LA TESIS DOCTORAL

Publicaciones derivadas directamente de la tesis doctoral

- **Garrido-Rodriguez M**, Lopez-Lopez D, Ortuno FM, Peña-Chilet M, Muñoz E, Calzado MA, Dopazo J. A versatile workflow to integrate RNA-seq genomic and transcriptomic data into mechanistic models of signaling pathways. PLoS Comput Biol. 2021 Feb 11;17(2):e1008748. **(Q1)**
- García-Martín A (*), **Garrido-Rodríguez M (*),** Navarrete C, Caprioglio D, Palomares B, DeMesa J, Rollland A, Appendino G, Muñoz E. Cannabinoid derivatives acting as dual PPARγ/CB2 agonists as therapeutic agents for systemic sclerosis. Biochem Pharmacol. 2019 Feb 27;163:321–334. (*) Equal contribution. **(Q1)**
- **Garrido-Rodríguez M**, Ortea I, Calzado MA, Muñoz E, García V. SWATH proteomic profiling of prostate cancer cells identifies NUSAP1 as a potential molecular target for Galiellalactone. J Proteomics. 2019 Feb 20;193:217–229. **(Q2)**

Publicaciones en colaboración durante el desarrollo de la tesis doctoral

- Hartmann O, Reissland M, Maier CR, Fischer T, Prieto-Garcia C, Baluapuri A, Schwarz J, Schmitz W, **Garrido-Rodriguez M**, Pahor N, Davies CC, Bassermann F, Orian A, Wolf E, Schulze A, Calzado MA, Rosenfeldt MT, Diefenbacher ME. Implementation of crispr/cas9 genome editing to generate murine lung cancer models that depict the mutational landscape of human disease. Front Cell Dev Biol. 2021 Mar 2;9:641618. **(Q1)**
- Casares L, García V, **Garrido-Rodríguez M**, Millán E, Collado JA, García-Martín A, Peñarando J, Calzado MA, de la Vega L, Muñoz E. Cannabidiol induces antioxidant pathways in keratinocytes by targeting BACH1. Redox Biol. 2020;28:101321. **(Q1)**
- García-Martín A, **Garrido-Rodríguez M**, Navarrete C, Del Río C, Bellido ML, Appendino G, Calzado MA, Muñoz E. EHP-101, an oral formulation of the cannabidiol aminoquinone VCE-004.8, alleviates bleomycin-induced skin and lung fibrosis. Biochem Pharmacol. 2018 Aug 2;157:304–313. **(Q1)**
- Moreno P, Jiménez-Jiménez C, **Garrido-Rodríguez M**, Calderón-Santiago M, Molina S, Lara-Chica M, Priego-Capote F, Salvatierra Á, Muñoz E, Calzado MA.

Metabolomic profiling of human lung tumor tissues - nucleotide metabolism as a candidate for therapeutic interventions and biomarkers. Mol Oncol. 2018 Sep 13;12(10):1778–1796. **(Q1)**

- Morrugares R, Correa-Sáez A, Moreno R, **Garrido-Rodríguez M**, Muñoz E, de la Vega L, Calzado MA. Phosphorylation-dependent regulation of the NOTCH1 intracellular domain by dual-specificity tyrosine-regulated kinase 2. Cell Mol Life Sci. 2019 Oct 11; **(Q1)**
- Palomares B, Ruiz-Pino F, **Garrido-Rodriguez M**, Eugenia Prados M, Sánchez-Garrido MA, Velasco I, Vazquez MJ, Nadal X, Ferreiro-Vera C, Morrugares R, Appendino G, Calzado MA, Tena-Sempere M, Muñoz E. Tetrahydrocannabinolic acid A (THCA-A) reduces adiposity and prevents metabolic disease caused by diet-induced obesity. Biochem Pharmacol. 2020 Jan;171:113693. **(Q1)**
- Correa-Sáez A, Jiménez-Izquierdo R, **Garrido-Rodríguez M**, Morrugares R, Muñoz E, Calzado MA. Updating dual-specificity tyrosine-phosphorylation-regulated kinase 2 (DYRK2): molecular basis, functions and role in diseases. Cell Mol Life Sci. 2020 May 27; **(Q1)**
- Palomares B, **Garrido-Rodriguez M**, Gonzalo-Consuegra C, Gómez-Cañas M, Saen-Oon S, Soliva R, Collado JA, Fernández-Ruiz J, Morello G, Calzado MA, Appendino G, Muñoz E. Δ9-TETRAHYDROCANNABINOLIC ACID ALLEVIATES COLLAGEN-INDUCED ARTHRITIS: ROLE OF PPARγ AND CB1 RECEPTORS. Br J Pharmacol. 2020 Jun 8; **(Q1)**
- Navarrete C, García-Martin A, **Garrido-Rodríguez M**, Mestre L, Feliú A, Guaza C, Calzado MA, Muñoz E. Effects of EHP-101 on inflammation and remyelination in murine models of Multiple sclerosis. Neurobiol Dis. 2020 Jun 26;104994. **(Q1)**

Comunicaciones presentadas por el doctorando en congresos internacionales

- Using multi-omics to depict the response of human fibroblasts to X-ray radiation. **Poster**. 4th Disease Maps Community Meeting. Sevilla, Spain. 2019-10.

Comunicaciones presentadas por el doctorando en congresos nacionales

- Development of a tool for the bioinformatic analysis of transcriptomic data in the biological big data era. **Poster**. VII University of Córdoba young investigators scientific conference. Córdoba, Spain. 2020-02.

- MGRNA-Seq: A light-weight, portable and scalable RNA-Seq analysis pipeline for biomedical research. **Talk**. X IMIBIC young investigators meeting. Córdoba, Spain. 2019-05.
- SWATH Proteomics Identifies A Marker Profile For Responsiveness To Galiellalactone In Prostate Cancer Cells. **Poster**. VI Proteomic young investigators symposium. Madrid, Spain. 2019-03.
- Integrative analysis of Transcriptomic, Proteomic and Phospho-proteomic data in DNA damage signaling path- ways. **Poster**. XIV Symposium on Bioinformatics. Granada, Spain. 2018-11.
- TRANSCRIPTOMIC AND PROTEOMIC PROFILING ANALYSIS OF HUMAN KERATINOCYTES EXPOSED TO CANNABIDIOL. **Talk**. 19th Annual meeting of the Spanish Society for Cannabinoid Research. Madrid, Spain. 2018-11.
- Bioinformatic Analysis of SILAC-Based Quantitative Proteome and Phosphoproteome for the identification of potential biomarkers for radiodermatitis associated with cancer radiotherapy. **Talk**. 9th IMIBIC young investigators meeting. Córdoba, Spain. 2018-05.

Por todo ello, se autoriza la presentación de la tesis doctoral.

Córdoba, 08 de Junio de 2021

Firma de los directores:

Fdo.: Marco Antonio Calzado Canale

Fdo.: Joaquín Dopazo Blázquez

Fdo.: Eduardo Muñoz Blanco

A mis padres, a mi hermano y a Ana

To my parents, brother, and Ana

“Teach thy tongue to say 'I do not know' and thou shalt progress”

Maimonides (1138–1204)

Table of contents

Abstract

Biomedical research models try to leverage useful insights for the sake of human health. Such models emerge from data, and, in recent years, the advent of omics technologies has revolutionized the way molecular data is generated and processed. Omics technologies allow the systematic evaluation of the molecules in a sample, generating large amounts of data that need to be analyzed and interpreted through computational approaches. Particularly, in the functional analysis step, molecular features are usually combined with prior knowledge to reduce its dimension and increase its interpretability. Across the different functional analysis approaches, those that try to model altered cellular signaling circuits hold the promise to reveal disease mechanisms and to allow the development of effective treatments. Cellular signaling can be traced and modeled from multiple molecular layers, but first proteomics, then transcriptomics and finally genomics offer the closer look to signaling processes. Among them, due to its balance between technology and closeness to the phenotype, the sequencing of RNA molecules (RNA-Seq), has gained great popularity in the past few years.

In the present work, we develop new computational tools to deepen on the analysis and modeling of cellular signaling circuits from transcriptomic and proteomic data. First, we developed MIGNON, a novel workflow for the analysis of RNA-Seq data that can analyze the genomic and transcriptomic information extractable from this technique. MIGNON offers a mechanistic signaling framework to combine the two levels of information, generating an output easy to interpret and link to a given phenotype. Second, we applied and evaluated different computational strategies to infer altered cellular signaling circuits from transcriptomic, proteomic and phosphoproteomic data, highlighting the similarities and differences between them. Finally, we created a basic toolkit to perform the visualization, differential analysis and classic functional analysis of transcriptomic and proteomic data that was applied to create, reinforce, or refute hypotheses in different biomedical research contexts.

Overall, this thesis work aims to improve the functional analysis of transcriptomic and proteomic data, providing the research community with classic and novel methodologies in the form of software packages. In addition to the novelty of MIGNON and of the comparison of strategies to model cellular signaling circuits, we believe that this work can help to widen the bottleneck constituted by the omic data interpretation in biomedical research contexts.

Resumen

Los modelos científicos empleados en la investigación biomédica pretenden generar información útil para mejorar la salud humana. Estos modelos emergen a partir de datos y, en los últimos años, la llegada y el perfeccionamiento de las tecnologías ómicas ha revolucionado la forma en la que los datos moleculares son generados y procesados. Las tecnologías ómicas permiten evaluar sistemáticamente las moléculas en una muestra biológica, generando una gran cantidad de datos que deben ser analizados e interpretados a través de aproximaciones computacionales. Concretamente, en el análisis funcional, los datos ómicos se combinan con información previa, reduciendo su dimensión e incrementando su interpretabilidad. De entre las diferentes estrategias de análisis funcional, aquellas que intentan modelar los circuitos de señalización alterados son especialmente prometedores de cara a revelar los mecanismos subyacentes a procesos patológicos, permitiendo el desarrollo efectivo de nuevos tratamientos. Este modelado se puede llevar a cabo con diferentes capas moleculares, pero, las técnicas proteómicas, transcriptómicas y genómicas, en ese orden, ofrecen la visión más cercana a los procesos de señalización. De entre ellas, debido a su equilibrio entre desarrollo tecnológico y cercanía al fenotipo, la secuenciación de moléculas de ARN (RNA-Seq) ha ganado una gran popularidad en los últimos años.

En este trabajo, hemos desarrollado nuevas herramientas computacionales para profundizar en el análisis y modelado de circuitos de señalización celular a partir de datos transcriptómicos y proteómicos. En primer lugar, desarrollamos MIGNON, un nuevo *workflow* para el análisis de datos de RNA-Seq capaz de analizar la información genómica y transcriptómica que se puede extraer de esta técnica. Además, MIGNON ofrece la posibilidad de combinar los dos niveles de información en el contexto de un modelo de señalización celular, generando un resultado fácil de interpretar y de vincular a determinados fenotipos. En segundo lugar, aplicamos y evaluamos diferentes estrategias computacionales para inferir circuitos de señalización celular a partir de datos transcriptómicos, proteómicos y fosfoproteómicos, explorando la similitudes y diferencias entre ellos. Finalmente, creamos un kit computacional para llevar a cabo la visualización y el análisis estadístico y funcional de datos transcriptómicos y proteómicos. Este kit fue aplicado para crear, reforzar o refutar hipótesis en diferentes contextos de investigación biomédica.

En resumen, el trabajo de esta tesis pretende mejorar el análisis funcional de datos transcriptómicos y proteómicos, proporcionando a la comunidad metodologías novedosas y clásicas en forma de nuevos paquetes de software. Además de la novedad de MIGNON y de la comparación de estrategias para modelar los circuitos de señalización celular, creemos que este trabajo puede ayudar a ensanchar el cuello de botella que representa el análisis e interpretación de datos ómicos en la investigación biomédica.

Abbreviations

List of abbreviations

- Ajulemic Acid: AJA
- ACSN: Atlas of Cancer Signaling Network: ACSN
- Auto gain control: AGC
- Base pairs: bp
- Basic local alignment search tool: BLAST
- Cancer Cell Line Encyclopedia: CCLE
- Clinical Proteomic Tumor Analysis Consortium: CPTAC
- Complementary DNA: cDNA
- Count-per-million: CPM
- Data dependent acquisition system: DDA
- Database of interacting proteins: DIP
- Data-independent acquisition: DIA
- Deoxyribonucleic acid: DNA
- Deoxythymidine triphosphate: dTTP
- Deoxyuridine Triphosphate: dUTP
- EHP-101: EHP
- European Nucleotide Archive: ENA
- False discovery rate: FDR
- Functional Class Scoring: FCS
- Galiellalactone: GL
- Gene Expression Omnibus: GEO
- Gene Ontology: GO
- Gene set enrichment analysis: GSEA
- Generalized linear model: GLM
- Gene-set variation analysis: GSVA
- Genome Analysis Toolkit: GATK
- Genomics of Drug Sensitivity in Cancer cell lines: GDSC
- Guanine and cytosine: GC
- Hierarchical Graph FM index: HGFM
- High-energy collision dissociation: HCD
- High-performance computing: HPC

- Human Protein Atlas: HPA
- Hypertext Markup Language: HTML
- Insertions and deletions: indels
- Integer Linear Programming: ILP
- Kolmogorov-Smirnov: KS
- Kyoto Encyclopedia of Genes and Genomes: KEGG
- Liquid chromatography coupled to mass spectrometry: LC-MS/MS
- Loss of function: LoF
- Mass spectrometry: MS
- Mechanistic integrative analysis of RNA-Seq: MIGNON
- Messenger RNA: mRNA
- Micro-RNA: miRNA
- Mitogen-activated protein kinase: MAPK
- Molecular Signatures Database Hallmarks: MSigDb
- Next generation sequencing: NGS
- Normalized Enrichment Scores: NES
- Over Representation Analysis: ORA
- Pathway Topology: PT
- Polymerase chain reaction: PCR
- Post-translational modifications: PTMs
- Principal Component Analysis: PCA
- Prior Knowledge Network: PKN
- Protein-protein interactions: PPI
- Ribonucleic acid: RNA
- Ribonucleic Acid Sequencing: RNA-Seq
- Ribosomal RNA: rRNA
- Sequential Enrichment from Metal Oxide Affinity Chromatography: SMOAC
- Single nucleotide polymorphisms: SNP
- Stable isotope labeling by amino acids in cell culture: SILAC
- Systems biology markup language: SBML
- Tandem mass tag: TMT
- Transcription factors: TF
- Trimmed Mean of M-values approach: TMM

- Variant Calling Format: VCF
- Variant Effect Predictor: VEP
- Workflow Description Language: WDL

File formats and softwares are named as described in each tool manuscript or online documentation. Genes and proteins are named according to the HUGO Gene Nomenclature Committee (https://www.genenames.org/).

Introduction

1. Models, complexity, omics, and prior knowledge

1.1. Models and biological complexity

In science, modeling can be defined as “the generation of a physical, conceptual, or mathematical representation of a real phenomenon that is difficult to observe directly” (1). Notably, models are everywhere in biomedical research, from cellular or animal models that try to mimic pathological conditions, to the mathematical models that are employed to analyze resulting data. Models enforce a scientific habit of mind, and are crucial to explain natural phenomena from a “militant ignorance” (2). Markedly, “since all models are wrong” were the words chosen by George Box back in 1976 to highlight the limitations of scientific models to capture and explain real world complexity (3). However, although not perfect, such models are our best available tools to try to understand complex problems, like the causes of a disease or a particular drug effectiveness. This is also stated in George Box’s manuscript with the following sentence: “we make tentative assumptions about the real world which we know are false but which we believe may be useful nonetheless”.

Particularly, in biomedical research, scientific models try to simplify biological complexity to derive useful insights for the sake of human health. In general, the sources of such complexity are unknown, although some examples of them include intrinsic and environmental factors, or those related to biological hierarchy. For instance, the biological hierarchy is probably one of the most emphasized aspects when learning biology (4). From molecules to organisms, there is a wide range of levels with different structure, composition, morphology, and distribution. Thus, structural biology groups the entities in molecules, organoids, cells, tissues, organs, and organisms, each of them being a hierarchical level of complexity. This hierarchy is also observed at the central dogma of molecular biology. The deoxyribonucleic acid (DNA) is used as the template to create ribonucleic acid (RNA) molecules, which are translated into proteins that trigger functions such as signaling or metabolism. In essence, models may propagate better or worse through the hierarchical layers, and this is something that should not be forgotten (5).

Another clear example of biological complexity can be found within the combination of intrinsic and extrinsic factors (6). The most compelling evidence is observed in genetics and development. Two cells, carrying the same genetic material, can become cell types with completely different functions, such as a keratinocyte or a neuron, depending on a combination of extrinsic and intrinsic factors. Furthermore, the environment is perhaps one of the aspects that is controlled the most in biological models. In fact, many models are based on the concept of changing certain environmental conditions, leaving the others fixed, to see how a cell, tissue or organism reacts. Normally, such perturbations are modeled when analyzing the results with statistical approaches, as described by Ronald A. Fisher in 1935 (7). However, when such perturbations are not obvious and thus are not considered, models may provide misleading results. A clear example of this is the crosstalk that occurs between signaling pathways in cancer (8), where targeting an individual signaling circuit without considering the rest can be very limiting to a specific use-case.

1.2. "Omics" technologies

We use data for the induction of new models, and over the last 25 years, starting with the automation of DNA microarrays (9), the way data is obtained to study biological systems has progressively changed. Omics technologies have evolved from measuring a small set of molecules to being able to quantify comprehensively and simultaneously hundreds or thousands of molecules. In this sense, the term "omics" is assigned to a group of techniques that aim to study a sample's molecular "-ome", this being the totality of molecules of a specific type. The appearance of such approaches is closely related to the establishment of the term "systems biology", which is now ~ 20 years old (10). Thus, from a biochemical perspective, we can divide the main omics technologies into four groups (11,12):

- **Genomics**, the most mature of the omics fields (12,13), delves into the composition and properties of the DNA molecules in a sample. Moreover, the study of DNA accessibility and modifications constitutes a strongly developed sub-field which is known as epigenomics (14). The output of genomics technologies takes the form of regions, which correspond to coordinates in an organism's genetic code.
- **Transcriptomics** study the transcriptome, which is constituted by the RNA molecules of a sample, including both coding and non-coding RNAs. The approaches employed within this group aim to measure RNA molecules qualitatively and quantitatively (15), and produce an output which is feature based, such as the expression of a given gene, or the presence of a particular micro-RNA (miRNA).

- **Proteomics** aim to deepen into the abundance, structure, modifications, and interactions of all the proteins in a sample. Much like transcriptomics, proteomics approaches produce an output which is feature based and which can have a qualitative or quantitative nature. In particular, since the appearance of approaches to measure protein post-translational modifications (PTMs) in a systematic and effective way (16), the field of proteomics has grown enormously to the point that it has been compared to genomics technologies, although its coverage still behind that provided by genomics and transcriptomics (17).
- **Metabolomics** is the field that attempts to systematically characterize all metabolites in a sample. The metabolome is composed by multiple small molecule types, such as amino acids, lipids, or carbohydrates. Although traditionally employed for biomarker discovery, and given its proximity to the observable phenotype, metabolomics technologies are now considered a powerful tool to elucidate biological mechanisms (18).

In addition to those, there are other techniques that arise from granularity, like translatomics and epigenomics, from studying different organisms, like metagenomics, or from applying several omics to the same set of samples, also known as multi-omics. From a technical point of view, we can group the omics approaches in two broader sets: Techniques that use next generation sequencing (NGS), including genomics and transcriptomics, and techniques that employ mass spectrometry (MS), namely proteomics and metabolomics (19). **Figure 1** provides an overview of the main omics technologies.

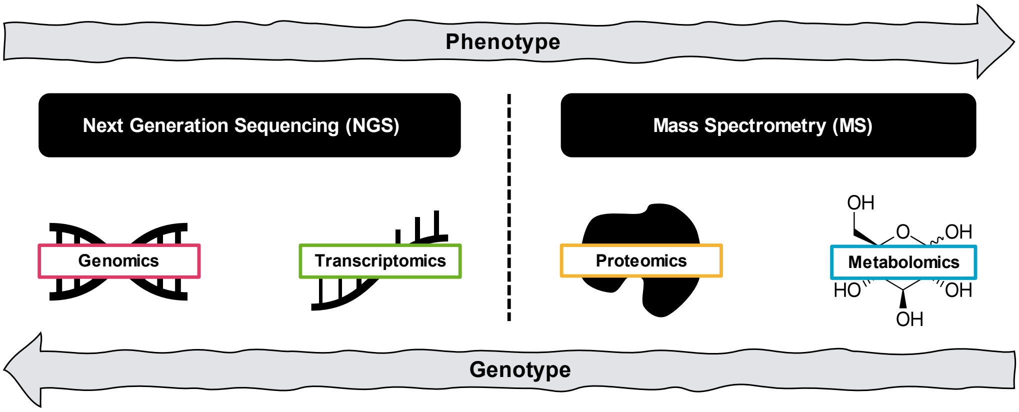

Figure 1. Overview of the different omics approaches. The horizontal axis defines the proximity to the phenotype as defined in Hasin et al. 2017 (13). Boxes indicate the groups of omics approaches from a technical perspective, as defined in Tarazona et al. 2020 (19).

Although each omic technology has its own particularities, all of them generate data in a high-throughput manner. To get an idea of how much information can be obtained using these technologies, while recognizing that actual numbers may vary between assays, it is interesting to take a look to one of the most recent multi-omics studies of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (20). In this study, authors characterized 99 glioblastomas using genomics, transcriptomics, proteomics, and metabolomics. **Table 1** depicts the number of features that were identified from each of the approaches and, as one can see, it can vary from hundreds to thousands of molecules. This dimension of data is something that can no longer be handled using visualization or interpretation techniques that are applied to a few dozen features. Consequently, in recent years, there has been an explosion of computational approaches to handle and analyze omics data.

Omic	Feature	Identified features
Genomics	Somatic mutations	6625
Genomics	Somatic copy number variations	27217
Transcriptomics	Gene expression	45914
Transcriptomics	Circular RNA expression	3671
Transcriptomics	Mature miRNA expression	2883
Proteomics	Protein abundance	10998
Proteomics	Phosphosite abundance	70330
Metabolomics	Lipid abundance	582
Metabolomics	Metabolite abundance	134

Table 1. Number of identified features per omic technique in the glioblastoma study. Data were extracted from the Table S2 of Wang et al. 2021 (20).

1.3. Prior knowledge and databases

Usage of prior knowledge is inherent to the scientific method. Thus, the first step when asking a new question is to perform extensive background research, recapitulating the most recent relevant knowledge in the so-called state-of-the-art. Most of this prior knowledge is stored in the form of unstructured text in scientific articles, which can be retrieved and explored through manual bibliographic review or, more systematically, with text mining approaches that allow natural language processing (21). In recent years, and in part due to the explosion in popularity and availability of omics data, manual retrieval of all knowledge is no longer feasible, as the number of questions that can be answered with such data grows together with its volume. Henceforth, structured biological databases are the perfect complement to a bibliography, as they provide information in a way that is easily usable by computational models (22).

Structured databases make information available in an organized manner, normally through tables or records, providing a quick and easy way of obtaining knowledge. It is important to realize that structured knowledge has been extensively used throughout history, and that a great example of this are non-digital encyclopedias or its digital counterpart, the Wikipedia. In molecular biology, some databases throughout history have revolutionized the way researchers store and use information (23). For example, the initial version of the Protein Data Bank was published in 1977, providing the first computer-based archival file for macromolecular structures, which was distributed in three different geographical locations: Brookhaven (United States), Tokyo (Japan), and Cambridge (Europe) (24). Some years later, in 1990, the appearance of new computational approaches, like the basic local alignment search tool (BLAST) transformed sequence databases into widely used resources (25). A good illustration of this was SWISS-PROT, published in 1997 (26), or the RefSeq database, released in 2000 (27). SWISS-PROT was later integrated into the UniProt archive, which provided a summary of the main characteristics of all known proteins for different organisms, and that was released in 2004 (28). Coupled to those, in the same period, other databases focused on the functional role of genes and proteins, and their interactions, were also published. Such databases contained the knowledge accumulated during decades of research and made it available in a simple and computer-friendly manner. For instance, the first version of the Kyoto Encyclopedia of Genes and Genomes (KEGG) was released in 1995, providing links between sequences and biological functions, and grouping biological entities like genes and proteins into molecular pathways (29). Later, in the year 2000, the Gene Ontology (GO) database was published, providing biologists an unified language for the functional annotation of genes (30). In the same year, the database of interacting proteins (DIP) was published, releasing a comprehensive and integrated tool for browsing information about protein interactions and networks in biological processes (31).

Of these databases, some have matured, some have disappeared, and others have become meta-databases. The latter try to serve as the “Towers of Babel” for knowledge and provide different levels of information that aim to cover different aspects of biology. For example, the current version of KEGG is a multi-level database that covers systems, genomic, chemical and health information, and that due to the recent global pandemic, has decided to make an effort to systematically define interactions between viruses and cellular organisms (32). Another good example of a meta-database is ConsensusPathDB, which gathers information from 32 public resources to integrate interaction networks in Homo Sapiens including binary and complex protein-protein, genetic, metabolic, signaling, and drug-target interactions, as well as biochemical pathways (33). A more recent example of this kind of databases is OmniPath, which contains information about Homo Sapiens transcriptional regulation, protein structure and mechanism, protein-protein interactions (PPI), tissue expression patterns, functional annotations and inter-cellular communication in a single and unified resource (34,35). Overall, there is a long list of databases that group and collect biological information and that are used daily in biomedical research. **Table 2**, adapted from Zou et al. 2015 (36), provides an overview of some of the databases employed for human research.

Name	**URL**	**Brief description**
1000 Genomes	http://www.1000genomes.org	A deep catalog of human genetic variation
ArrayExpress	http://www.ebi.ac.uk/arrayexpress	Database of functional genomics experiments
COSMIC	http://cancer.sanger.ac.uk	Catalog Of Somatic Mutations In Cancer
dbSNP	http://www.ncbi.nlm.nih.gov/snp	Database of single nucleotide polymorphisms
DIP	http://dip.doe-mbi.ucla.edu	Database of Interacting Proteins
DisGeNET	http://www.disgenet.org/web/DisGeNET/v2.1	Gene–disease associations
EGA	http://www.ebi.ac.uk/ega	European Genome–phenome Archive
Ensembl	http://www.ensembl.org	Ensembl genome browser
Expression Atlas	http://www.ebi.ac.uk/gxa	Differential and baseline expression
GENCODE	http://www.gencodegenes.org	Encyclopedia of genes and gene variants
GeneCards	http://www.genecards.org	Integrated database of human genes
GO	http://geneontology.org	Gene ontology
HGNC	http://www.genenames.org	Database of human gene names
HMDB	http://www.hmdb.ca	Human Metabolome Database
Human Protein Atlas	http://www.proteinatlas.org	Tissue-based map of the human proteome

ICGC	http://icgc.org	International Cancer Genome Consortium
JASPAR	http://jaspar.genereg.net	Transcription factor binding profile database
KEGG	http://www.genome.jp/kegg	Kyoto Encyclopedia of Genes and Genomes
KEGG PATHWAY	http://www.genome.jp/kegg/pathway.html	KEGG pathway maps
MINT	http://mint.bio.uniroma2.it/mint	Molecular INTeraction Database
NCBI GEO	http://www.ncbi.nlm.nih.gov/geo	Gene Expression Omnibus
NCBI RefSeq	http://www.ncbi.nlm.nih.gov/refseq	NCBI Reference Sequence Database
OMIM	http://omim.org	Online Mendelian Inheritance in Man
PANTHER	http://www.pantherdb.org	Protein ANalysis THrough Evolutionary Relationships
PDB	http://www.rcsb.org/pdb	Protein Data Bank for 3D structures of macromolecules
Pfam	http://pfam.xfam.org	Database of conserved protein families and domains
PID	http://pid.nci.nih.gov	Pathway Interaction Database
PIR	http://pir.georgetown.edu	Protein Information Resource
PRIDE	http://www.ebi.ac.uk/pride	PRoteomics IDEntifications
PubMed	http://www.ncbi.nlm.nih.gov/pubmed	Database of biomedical literature from MEDLINE
PubMed Central	http://www.ncbi.nlm.nih.gov/pmc	Free full-text literature archive
Reactome	http://www.reactome.org	Curated and peer-reviewed pathway database
TargetScan	http://www.targetscan.org	Predicted miRNA targets in mammals
TCGA	http://cancergenome.nih.gov	The Cancer Genome Atlas
UCSC Genome Browser	http://genome.ucsc.edu	UCSC Genome Browser database
UniProt	http://www.uniprot.org	Universal protein resource

Table 2. List of databases for human research. Adapted from Zou et al. 2015 (36).

1.4. Curated and non-curated information

Given the dimension of biological databases, some of the records are redundant, inconsistent, incomplete, or outdated, an issue that can be addressed by manually reviewing the information, with the support of automatic tools (37,38). This process is known as curation, and provides a quality leap in the delivered information, that represents an extraordinary volume of work (39). Some databases provide curated information, such as UniProt or Reactome (40), a resource for the visualization, interpretation and analysis of pathway knowledge. Such databases make a great effort to enroll domain experts to review and certify the information that they publish, even having public guidelines and tools for such curators. On the other hand, non-curated databases use data or computational approaches to build the served knowledge. Examples of those are NetworKIN (41), that uses annotated kinase-substrate motifs to find new interactions through a sequence similarity search, or the Human Protein Atlas (42), which is designed to serve the results of a group of assays to determine the abundance level of proteins in different tissues and subcellular compartments. Moreover, some databases such as GO or STRING (43), a PPI database, deliver a combination of curated and non-curated knowledge.

Depending on the purpose of the research, scientists may prefer curated or non-curated knowledge. While curated databases provide a solid basis to build new hypotheses, non-curated information may be especially relevant to make new and ground-breaking discoveries. This happens because of the so-called popularity bias, in which researchers try to build knowledge on top of “safe” knowledge that is proven and accepted by the scientific community (44). However, this leads to an observational bias that occurs when people only search for something where it is easiest to look, which is also known as the streetlight effect (44,45). Indeed, some authors have argued that this gene annotation bias in databases is an obstacle for biomedical research, and that part of the unanswered questions would be addressed through data-driven approaches (46). **Figure 2** exemplifies this by showing the top 10 most cited genes from the NCBI Entrez Gene2Pubmed database together with a random selection of genes with a single mention. Almost any biologist would recognize most of the genes at the top of the plot, in contrast to those at the bottom which are largely unfamiliar. Some recent methodologies, such as NicheNet, are starting to include gene popularity measurements within their models to avoid this phenomena (47). The combination of omics technologies and statistical learning seems to be a hopeful tool for entering this dark area of biological knowledge, as some authors have demonstrated at the phosphoproteomic level, a molecular layer that particularly suffers the consequences of this effect (48,49).

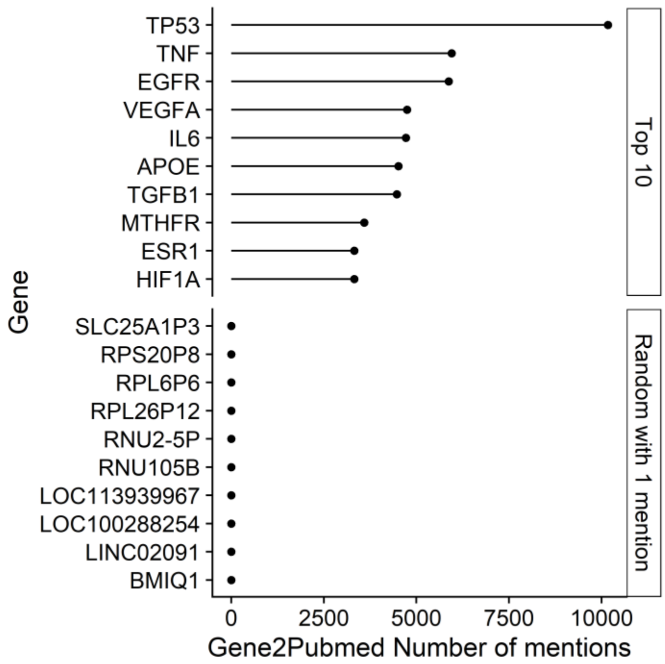

Figure 2. Example of popularity bias from NCBI Entrez Gene2Pubmed database. The upper facet shows the top 10 genes with the higher number of mentions in the database, while the panel at the bottom reflects a random selection of genes with a single mention.

1.5. Functional analysis of omics data

Omics technologies are an important breakthrough in our way of studying biology, given that they allow measuring from hundreds to thousands of molecules in a single assay. However, with this great increase in power comes the challenge of analysis and interpretation, where traditional approaches may no longer be appropriate. For this task, new methodologies are being rapidly developed, causing a drastic growth in the computational biology field in the last two decades, and the emergence of a whole new world of subfields within it. From the different subfields that encompass the analysis of omics data, the functional analysis is the step that translates the results from the molecular level into something interpretable, actionable, and linkable to a given phenotype. Although it does not always involve the use of molecular pathways, this subfield is also known as pathway analysis (50).

Omics data got the ability to generate new knowledge on its own, through inductive unsupervised analyses like sample clustering (51) or correlation based network reconstruction (52). However, extractable information becomes much more powerful when combined with external information, such as sample phenotype or feature functional classification (22). Moreover, this combination is essential to address one of the most important bottlenecks of omics data, namely interpretation. Pathway analysis transforms a large list of molecules into a reduced list of functional features that are easier for us to understand and associate with a given phenotype. Those functional units are sets of related molecular features that are linked to specific biological processes, such as inflammation or cell growth. As with the remaining processing steps for omics data, different statistical approaches are used for the functional analysis. Usually, two main inputs are employed; A list of molecular features of interest extracted from the omics data with or without an associated metric, and prior information that can be defined by knowledge, such as KEGG pathways and GO terms or by data driven approaches, such as the Molecular Signatures Database Hallmarks (MSigDb) (53). In addition, this second data source may also contain prior information about the interactions between the molecular features that can be considered in the analysis. Overall, we can classify the pathway analysis approaches into three terms: Over Representation Analysis (ORA), Functional Class Scoring (FCS) and Pathway Topology Approaches (PT) (50).

To summarize, scientific models are built under certain assumptions in a constant iteration between theory and practice that constitute the advancement of learning. In biomedical research and computational biology, most assumptions are employed to simplify biological complexity, which is caused by factors like hierarchy, and intrinsic and environmental stimuli. Understanding them helps to point out the advantages and limitations of new models during their creation and development. To build such models, scientists use data, and the way data is obtained has changed with the advent of omics technologies. Those technologies are used to perform the systematic evaluation of the molecular state of a sample, allowing the generation, validation or rejection of hypotheses using large datasets. In addition, scientists consider the prior knowledge obtained by others to build new models. This prior knowledge can be retrieved from scientific literature or structured databases and is essential to guide data interpretation and new discoveries. However, extensive usage of prior knowledge may result in a dangerous feedback loop that creates a popularity bias towards events that are easier to observe. Finally, the combination of prior knowledge with omics data in functional analyses transform an initial list of molecules into a reduced list of features that can be associated to a variety of biological processes, and that are easier to interpret than the original dataset.

2. RNA-Seq: The cornerstone

2.1. Why transcriptomics?

From the different omics technologies, transcriptomics is unrivalled in terms of scientific impact in recent years. In the study from Perez-Riverol et. al 2019 (54), authors quantified the impact of each omic level in biomedical research creating custom variables from the number of available datasets, citations, and occurrences of reanalysis, among others. The impact of transcriptomic studies regarding citations and available datasets were impressive when compared to the next most impactful level, genomics. For instance, as of April 2019, transcriptomics accounted for a total of 665,022 citations and 50,699 cited datasets against 8,152 and 3,389 citations and cited datasets respectively for the genomic level (54). As of March 2021, using data from the front page of OmicsDI (55), the number of available datasets is still dominated by transcriptomic technologies, as shown in **Figure 3**.

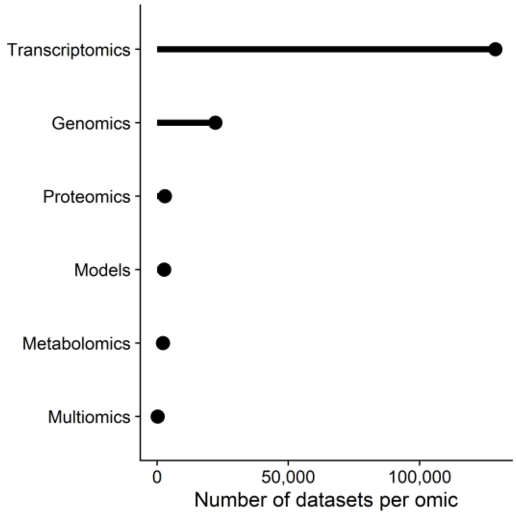

Figure 3. Number of available datasets per omic. Data extracted from OmicsDI front page as of March 2021 (https://www.omicsdi.org/).

But why do transcriptomics studies have so much impact and popularity compared to other omics? Although the reasons behind this are not clear, some assumptions can be made based on the knowledge of its history and application. First, transcriptomics technologies have gone through two major technological breakthroughs in the last thirty years that have boosted its coverage, sensitivity, and specificity for measuring gene expression: Microarrays and NGS (15). Particularly, the application of the sequencing by synthesis (e.g., Illumina) to study the transcriptome, also known as Ribonucleic acid sequencing (RNA-Seq), emerged as one of the most powerful and impactful tools to study biological systems across different fields (56). Second, although those two technological revolutions also happened to genomics, transcriptomics is one step closer to the observable phenotype than genomics, as shown **Figure 1**. Thus, RNA-Seq can provide a closer look to the phenotype while also guaranteeing a high coverage and accuracy of the measured molecular level thanks to its underlying technology. Finally, the maturation of RNA-Seq in recent years has triggered the emergence of new sub-technologies that have allowed us to explore other aspects of RNA, like its structure, translation, or distribution at the single cell level (57).

2.2. Analysis of RNA-Seq data

Among the different RNA-Seq based approaches, bulk RNA short-read sequencing is the most extensively used methodology for gene expression profiling (58). With this technique, total RNA extracted from the samples of interest is processed to select target RNAs. To do so, the two most common enrichment methods are messenger RNA (mRNA) selection through polyadenylated tail capture and ribosomal RNA (rRNA) depletion. The enriched RNA sample is then fragmented, retro-transcribed and amplified to generate a complementary DNA (cDNA) library that is fed to the sequencer. Within this process, an important point to consider is whether the initial DNA strand direction information is kept or not, as this may impact the final gene expression quantification (59). The output of the sequencer consists of files that contain single-end or paired-end reads generated from the cDNA library. Such files contain, on average, between 10 and 100 million sequence reads with a length that varies from 25 to 125 base pairs (bp) (58). In addition, such files are normally the starting point of a wide variety of bioinformatic workflows for the analysis of RNA-Seq data.

The most common initial step is the quality control and trimming of the raw reads that can be performed with tools such as Trimmomatic (60) or fastp (61). The core of these workflows is usually composed of aligners that consider the spliced nature of RNA, such as TopHat2 (62), STAR (63), HISAT2 (64) or Rail-RNA (65), which map reads against a reference genome, or by pseudo-alignment tools as Salmon (66) or kallisto (67), that directly obtain a quantification for the regions of interest, normally genes or exons, using probabilistic models. Additionally, there are pipelines which are intended to be run by the user in local computers or high-performance environments, such as QuickRNASeq (68), or interactively in cloud-based platforms, after uploading raw data to an external service, such as BioJupies (69) or RaNA-Seq (70). Typically, the culmination of the experiment involves differential expression analysis, carried out using different types of statistical models, with packages as edgeR (71), DESeq2 (72) or limma (73). The functional analysis of the processed data is then carried out with different implementations of ORA, FCS and PT methods, such as FatiGO (74), GSEA (75) or subSPIA (76). In addition, the functional analysis can happen before the statistical analysis, but after gene expression normalization, using single sample approaches as GSVA (77), or HiPathia (78).

Despite the existence of different pipelines to perform the aforementioned tasks, most of them present two major drawbacks. First, the genomic information contained in the RNA-Seq reads usually remains unused. However, genomic variants, which may contain crucial information about the functionality and potential activity of the resulting proteins in the different processes where they participate, can be retrieved from such sequences. In this sense, it is well known that RNA-Seq has some limitations for DNA variant calling. There are two main points to consider: (i) lowly expressed genes include lower depth, so variant calling is harder in those regions and (ii) the detection of heterozygous variants can be limited due to allele-specific gene expression (79). Despite these limitations, it has been demonstrated that variants can be called even for low expressed genes in deeper RNA-Seq sequence samples. Moreover, some studies have shown that RNA-Seq variant calling is able to provide a good sensitivity of 99.7%-99.8% in both heterozygous and homozygous variants whereas precision still reaches 97.6% in homozygous but 90% in heterozygous (80). The second major drawback is that conventional functional analysis strategies are mainly descriptive, and very limited in providing insights of the underlying molecular mechanisms that produce the observed phenotypic responses. Recently, a new generation of methods known as mechanistic pathway analyses, within the PT category, are outperforming traditional approaches in both biological explanatory power and interpretability (81).

As can be seen, the analysis of RNA-Seq data usually starts from raw sequences and comprises different steps: Quality control, alignment, quantification, normalization, statistical evaluation, and functional analysis. Current methods, however, present some limitations such as the omission of the genomic information of RNA-Seq data and the inability to use mechanistic functional analysis tools for the interpretation of results.

3. Modeling cellular signaling

3.1. The language of cells

The cells that form up tissues can communicate with other cells and with themselves and to do so, they employ a language called cellular signaling. In this language, an initial mechanical or biochemical stimulus is processed by cells to trigger specific biological functions through the usage of existing cellular components or the synthesis of new ones. Depending on which cell produces and processes the stimulus, signaling can be classified as intracrine, autocrine, juxtracrine, paracrine and endocrine (82). From a theoretical perspective, cell signaling can be conceived as an opened electric circuit, where the initial stimuli generates a message that is transmitted through the circuit until it reaches its target and performs a given function (83). In addition, cellular signaling can be carried out by different types of molecules such as lipids, phospholipids, amino acids, monoamines, proteins, or glycoproteins. Among them, proteins, and their post-translational modifications (PTMs), are major players in signaling and in the majority of cases, they constitute the backbone of the messaging cascades that occur in response to stimuli (84). Particularly, phosphorylation and dephosphorylations, carried out by kinases and phosphatases respectively, play a crucial role in signaling (85). Thus, knowing which exact proteins or PTMs are responsible for the signaling that drives a particular biological process, allows for the conversion of a purely observational result to mechanistic insights, that in turn enable the modification of such process using the proper tools. Moreover, cellular signaling is a process that has been refined throughout our evolution and, when deregulated, can lead to pathological phenotypes such as hyperproliferation in cancer or defective communication in Alzheimer’s disease.

However, modeling the exact signaling cascades that are responsible for a particular process is not a straightforward task. As mentioned in the first section, there are different axes of complexity that prevent the creation of large-scale and accurate cellular signaling models, such as time and space (86). Nevertheless, the efforts of thousands of scientists throughout history have shed light on how cell signaling circuits function, and on their main components. The most common models employed to try to understand signaling are networks. Networks have been used to represent not only the molecular players of signaling, but also their connection with other entities such as drugs, processes and diseases, through nodes and links (87,88). An example of this is found in **Figure 4**, which represents the mitogen-activated protein kinase (MAPK) cascade, one of the most pivotal processes in cell signaling (89). As with the rest of prior knowledge, the nodes and links that compose signaling networks are based upon information that has been refined over the years from experimental data. These nodes and links can be retrieved from generic databases such as KEGG, Reactome, WikiPathways (90) and SIGNOR (91), or from disease specific databases such as the Atlas of Cancer Signaling Network (ACSN) (92).

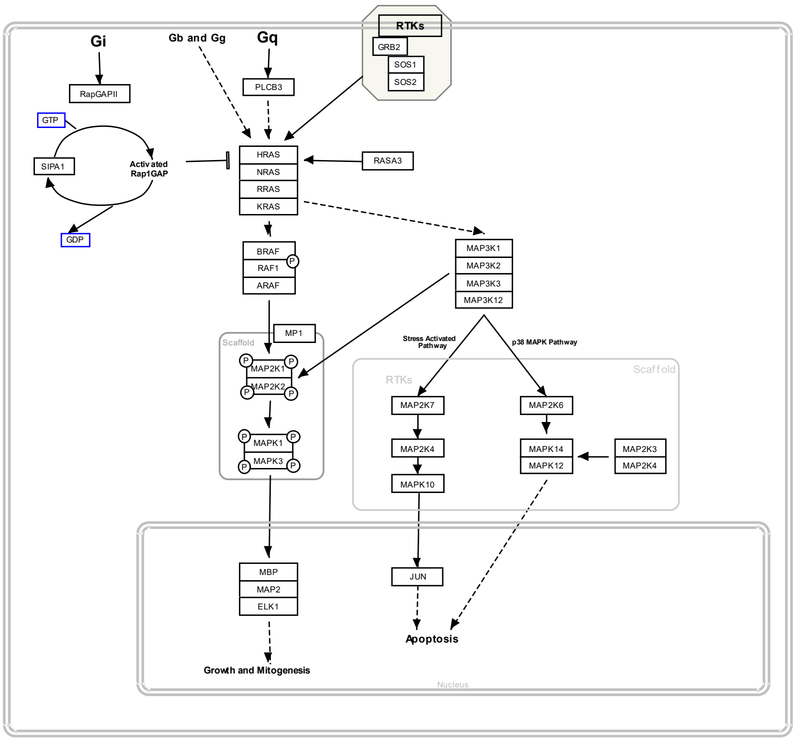

Figure 4. MAPK cascade network. The figure is adapted from the WikiPathways MAPK cascade pathway (https://www.wikipathways.org/index.php/Pathway:WP422).

3.2. Using omics data to infer cellular signaling

Models of cellular signaling aim to obtain a mechanistic or causal view of how the molecules communicate or communicated (in the case of reverse engineering) to orchestrate a given biological process (93). That is, which exact circuits, among all the possibilities, were responsible for the phenotype of interest. In this sense, omics data can provide a systematic evaluation of the biochemical molecules that compose the cellular signaling networks. Of course, such networks become more and more complex depending on the amount of information aiming to be captured, even with the proper data. In fact, models such as the MAPK cascades shown in **Figure 4** are an over-simplification that ignores events like feedback loops (94), protein complexes formation (95) and signaling pathways crosstalk (96). There are dedicated computational formats that aim to capture a higher complexity, such as the systems biology markup language (SBML) (61). Nonetheless, these simpler and intuitive models of cellular signaling have proven to be useful to derive mechanistic insights through different types of computational approaches, like logic modeling (97). In addition, approaches that use prior knowledge have shown overall better results for the inference of causal signaling networks than those that do not (68). **Figure 5** shows a schematic representation on how the omics data can be combined with prior knowledge to investigate cellular signaling. In recent years, many methods have been developed for the mechanistic or causal analysis of cellular signaling circuits from omics data and prior knowledge.

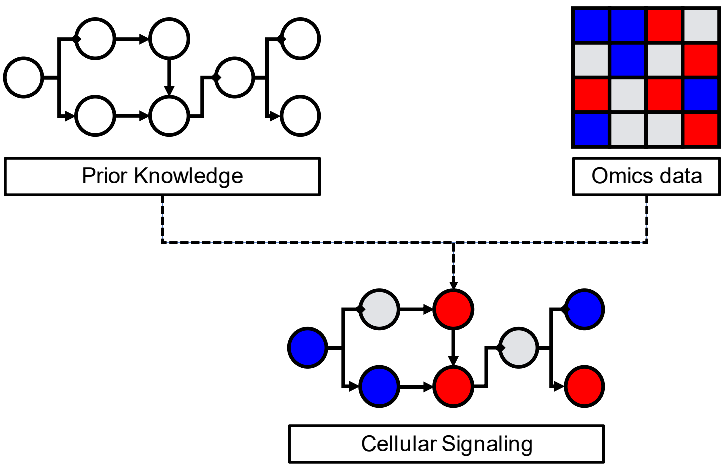

Figure 5. Schematic representation of the combination of omics data with prior knowledge to model cellular signaling.

For example, in 2007, Gao et al. developed TAPPA, which was one of the first functional analysis approaches that used signaling network topology to assess altered pathways from transcriptomics data (98). In this method, gene expression is normalized and used to calculate a pathway connectivity index for each KEGG pathway by multiplying the normalized abundance of each gene pair at all the registered interactions in the pathway. Later, in 2009, Tarca et al. developed a method called SPIA (98). In this method, after assuming a steady state and performing the differential expression analysis, authors calculate two probabilities for each KEGG pathway, one based in an over-representation analysis of differentially expressed genes, and another based on the magnitude of change and connections of genes within the network. Finally, both P values are combined to rank signaling pathways by their probability of being altered.

In 2013, Sebastian-León et al. published Pathiways, a framework for the analysis and visualization of altered signaling pathways, that employs a probabilistic model to calculate a steady state signal transmission status (also in KEGG pathways), intended to be used with transcriptomics data (99). Unlike SPIA, Pathiways do not use differential expression analysis results as input, but a normalized expression matrix. Its main assumption is that gene expression is directly correlated with its protein product and with its ability to transmit the signal through the network. In addition, in this method, authors address an important problem of signaling network modeling which is the multi-protein nodes: “When nodes are composed by more than one protein, we distinguish two situations: alternative proteins or protein complexes. In the first case, the probability of having the node active is taken as the highest probability of any of the proteins in the node, given that they are supposed to be redundant. In the case of protein complex, all the proteins should be present to guarantee the integrity of the complex. Therefore, the probability of having the node active is conditioned to the lowest probability of all the proteins belonging to the complex”. In 2014, Krämer et al. published the algorithms employed by the Ingenuity Pathway Analysis software to perform different types of causal analyses like “Upstream Regulator Analysis”, “Mechanistic Networks”, “Causal Network Analysis”, and “Downstream Effects Analysis” (100). The core of those depends on the results of the “Upstream Regulator Analysis”, which employs a master network extracted from the Ingenuity Knowledge Base to identify nodes responsible for the observed omic profile. However, because the algorithms and knowledge base were developed by a company, there is no open-source implementation that allows their use and application without payment.

In 2015, Terfve et al. developed PHONEMES, a method to create large-scale models of signal propagation from phosphoproteomics at the phosphosite resolution level (101). Starting from a large bipartite background network containing kinase to substrate interactions, and after discretizing the phosphoproteomics data using gaussian mixture models, PHONEMES propose a logic modeling optimization solution that retrieves the signaling network that better matches the data. In a similar way but using transcriptomics Koumakis et al. implemented in 2016 MinePath, a method to evaluate functional sub-pathways in prior knowledge networks delimited by the main KEGG signaling pathways (102). To do so, MinePath first binarizes the gene expression matrix using a supervised entropy-based global discretization approach, and then applies a logic formalism to identify functional sub-pathways for each sample. In a final step, the proportion of functional sub-pathways is compared between the samples of interest using the Fisher's exact test.

Also, in 2016, the developers of Pathiways published HiPathia (78), a method for the mechanistic pathway analysis of transcriptomics data that improved Pathiways in two principal aspects. On the first hand, HiPathia decomposes KEGG signaling networks into sub-circuits from receptors (no input signal) to effectors (no output signal). In this sense, authors showed that increasing the granularity, from entire pathways to circuits, allows capturing more subtle signaling events that are strongly associated with cancer patients' survival. On the other hand, HiPathia core is composed by a new iterative signal propagation engine, based on the Dijkstra algorithm, that allows modeling loops when calculating the signal propagation through sub-circuits: "The advantage if using an iterative method is that the signal becomes steady even in cases of loops in the pathway topology, allowing a more precise estimation of circuit activities". Recently, in 2019, a benchmark compared different mechanistic pathway activity analysis methods and showed that HiPathia performs better in terms of sensitivity and specificity than other similar tools (81).

Bradley et al. published in 2017 the package CausalR, which follows the trend of the Ingenuity Pathway Analysis causal tools and PHONEMES (103), which involves starting from a big prior knowledge network (PKN) and finding a signaling sub-network coherent with omics data. To do so, authors propose to use a directed PKN of public domain, like SIGNOR (103), and discretize the omics input into three categories: Up-regulated nodes, unchanged nodes, and down-regulated nodes. Then, CausalR performs an upstream regulator analysis and uses resulting nodes to reconstruct a network by connecting the master regulators with observed nodes using sequential path-lengths. In 2019, Liu et al. developed CARNIVAL (104), a tool to transform expression footprints into causal signaling networks. The first step of this method consists of inferring transcription factor (TF) activities from gene expression data using an FCS approach with the DoRothEA knowledge base and the VIPER algorithm (105,106). Then, it estimates signaling pathway activities with PROGENy, a regression-based model constructed through the integration of transcriptomics datasets with specific signaling perturbations (107). In a final step, it uses a large PKN based on OmniPath interactions and solves an optimization problem where the TFs are the effector nodes (those at the bottom of the signaling network) and the genes associated with each PROGENy pathway aid to guide the pathway reconstruction through their inclusion in the objective function. In addition, users can provide, or omit, perturbed genes in their experiments, running standard or inverse CARNIVAL, respectively. The optimization is solved efficiently thanks to the usage of an Integer Linear Programming (ILP) implementation, previously developed by Melas et al. in 2015 (108). The same ILP implementation was employed in two recent methods, the first being a new version of PHONEMES, PHONEMES-ILP (109), and the second being COSMOS, a tool for multi-omics data integration that connects signaling and metabolism employing transcriptomics, phosphoproteomics and metabolomics data (110).

Also in 2019, Broaweys et al. published NicheNet, a method that combines prior knowledge with bulk or single-cell transcriptomics data to infer intercellular signaling (47). NicheNet makes extensive usage of prior knowledge to build a model of ligand-target regulatory potential, and then uses data to predict ligand activity from targets or targets activity from ligands. This approach also allows the user to trace back which signaling cascade was the responsible for a given ligand-target association. In addition, NicheNet applies a hub correction factor to address a point often overlooked in protein signaling networks, the fact that a small number of proteins holds the majority of connections (111). **Table 3** summarizes the approaches that were discussed in this section.

Method	Year	Reference
TAPPA	2007	(112)
SPIA	2009	(98)
Pathiways	2013	(99)
IPA Causal Tools	2014	(100)
PHONEMES	2015	(101)
MinePath	2016	(102)

HiPathia	2016	(78)
CausalR	2017	(103)
CARNIVAL	2019	(104)
NicheNet	2019	(47)

Table 3. Methods for signaling network inference from transcriptomics and proteomics data.

In conclusion, cells employ a biochemical language to transmit information and to react to external and internal stimuli known as cellular signaling. It can be carried out by different types of molecules, but proteins and their PTMs have shown a central role within this messaging system. Although understanding cellular signaling is not easy because of its complexity, its modeling provides scientists a mechanistic insight into biology that allows them to understand and potentially intervene in pathogenic biological processes. Those models normally take the form of networks, where nodes and links can represent a wide variety of events and entities. Such networks, that can be obtained from prior knowledge databases, become especially useful in combination with omics data, which measures the status of the components of the network in a high-throughput manner. Overall, the modeling of cellular signaling is a relatively young field, which benefits from employing prior knowledge in combination with the high-throughput capacity of omics techniques. Over the last 15 years, a great number of methods have been published, with each of them proposing a different combination of prior knowledge, omics data and modeling approach.

4. Applications in biomedicine

4.1. Omics technologies in biomedical research

Biomedical research has greatly benefited from the emergence of omics approaches and their ability to study molecules from a systematic point of view (12). Since the appearance of NGS, different consortia have made a great effort to describe the human genome and its variability, like the 1000 genomes project (113) or the United Kingdom's 10K and 100K genomes project (114). In addition to those, other projects have sought to profile the basal molecular state of all the tissues that make up human bodies, such as the Human Protein Atlas (42) or GTEx (115,116). Furthermore, omics technologies have revolutionized the diagnosis, prevention, and treatment of patients through specific studies where such technologies have been the key for new discoveries. For instance, one of the first successful and well-known examples of this was the MammaPrint, a group of 70 genes whose expression, measured through microarrays, was able to detect lymph node negative breast cancer patients with poor prognosis (117).

Of course, the number of individual studies where omics technologies have been used to discover molecular biomarkers or potential targets for the treatment of human diseases is massive. However, what is clear is that there is a reduced group of more prevalent diseases that have received the highest attention and that consequently, account for the majority of omics data generated in the last decade. A clear example of that is cancer, a disease with high molecular heterogeneity and mortality. For example, almost the totality of datasets available in OmicsDI come from cancer (55). This is, in part, caused by big consortia like TCGA and CPTAC, where hundreds of samples are evaluated using multi-omics approaches. However, far from being redundant, each new dataset brings a new molecular perspective to the table that allows for new insights into this disease. A good example of this is the recent proteogenomic study of breast cancer from the CPTAC, where multi-omics data was used to derive a new intermediate subtype between tumors classified as luminal A or B using the PAM50 criteria (118). Likewise, the omics techniques have also shed light into the animal or cellular models of disease. Examples of this are the Cancer Cell Line Encyclopedia (CCLE) (119) or the Genomics of Drug Sensitivity in Cancer cell lines (GDSC) (120). Leaving cancer aside, other resources can also be found, like CADgene (121), a comprehensive database for coronary artery disease genes, or AlzBase (122), an Integrative Database for Gene Dysregulation in Alzheimer's Disease.

In addition to the new insights brought by each individual study, the omics era in biomedical research is also leading to a higher transparency and reproducibility (123). This happens thanks to repositories like GEO (124), ArrayExpress (125), or PRIDE (126), where authors can submit their data before publishing their results. Once published, other researchers can freely access the data and reanalyze it to test their hypotheses. Furthermore, this flow of data allows computational biologists to carry out meta-analyses, where data from different studies of the same disease are shared in a common space and analyzed to extract consistent insights across studies (127).

4.2. Signaling and personalized medicine

Modeling the cell communication system holds the promise of providing a mechanistic perspective of biological phenomena. When it comes to biomedical research, this perspective can be an asset for understanding disease mechanisms on the first hand, and to derive specific treatments or predictions on the second. As an example, Fey et al. published in 2015 a study where authors characterized the networks that regulate cell stress signaling and use them to predict the survival in neuroblastoma patients (128). In another recent study by Eduati et al., authors employed logic modeling of signaling pathways to prioritize and predict patient-specific combination therapies in pancreatic cancer patients (129). Finally, Loucera et al. employed a combination of mechanistic models and machine learning to perform drug repurposing for the treatment of COVID-19 during the 2020 pandemic (130). All of these are successful examples of the application of cellular signaling models in biomedical research and personalized medicine.

Any discussion over personalized medicine implies the usage of genomics (131). Undoubtedly, and partly caused by projects like 1000 Genomes or TCGA, genomics technologies have reached the clinical practice and there are hundreds of companies and facilities devoted to the sequencing of patient samples to assist in the choice of treatment. Some well-known examples of this are the EGFR/KRAS mutations in lung cancer (132) or BRCA1/2 status in breast tumors (133). Nonetheless, the majority of the information retrieved by genomics technologies remains unused, largely because only a small proportion of mutations have been associated with a phenotype in databases like ClinVar (134) or COSMIC (135). For this reason, algorithms like CADD (136) or SIFT (137) aim to infer the functional relevance of genomic information. In this context, the mechanistic models of signaling pathways can be a fundamental aid in the interpretation and discovery of relevant mutations. For instance, Peña et al. showed that a combination of GTEx basal expression data coupled with mechanistic signaling modeling can be used in the clinical interpretation of mutations in diseases like Fanconi anemia and diabetes (138).

To summarize, biomedical research has undergone a revolution with the appearance of omics technologies, that have been employed in a countless number of studies to profile the molecular state of human samples and animal or cellular models of disease. Although the generation of data is biased towards certain diseases, like cancer, omics data in biomedical research are being widely used to gain new insights, not only by data publishers, but by the entire community, through data portals that increase the transparency and reproducibility of research. In addition, modeling cellular signaling is a powerful tool for personalized medicine, and recent studies have shown how to use them directly in clinical setups for tasks like the interpretation of genomics variants or treatment choice.

Aims

Omics technologies have revolutionized biomedical research, providing the ability to systematically evaluate the status of the molecules in biological samples and thus generating a substantial amount of data during recent years. The analysis of such data is composed by several steps, and the use of prior knowledge in combination with functional analysis methods can help to reduce its dimensionality and to increase the interpretability of potential findings. Among the different omics technologies, transcriptomics has gained great popularity in recent years due to a unique combination of underlying technology and proximity to the phenotype. However, current approaches for the analysis of transcriptomics data present some limitations, like omitting the genomic information extractable from RNA-Seq data or using traditional approaches for its functional analysis. In this sense, recent functional analysis methods aim to unravel cell signaling mechanisms from omics profiles, providing researchers a mechanistic or causal insight into the processes that give rise to certain phenotypes. All this background led us to stablish the following objectives for this thesis:

I. To create a new workflow for the analysis of RNA-Seq data with the ability of extracting the genomic information contained on it, and able to integrate it with the transcriptomic data using a mechanistic model of cellular signaling pathways.

II. To apply, evaluate and compare different approaches to model cellular signaling circuits combining the information provided by transcriptomic, proteomic and phosphoproteomic data.

III. To develop a toolkit to perform the analysis of omics data in biomedical research contexts, able to perform pre-processing steps, statistical modeling, and functional analysis of genomics, transcriptomics, proteomics and metabolomics data.

Material and methods

1. Cell culture and stimuli

Primary dermic fibroblasts from 4 different donors were obtained from Tebu-Bio (Le-Perray-en-Yvelines) and pooled together in culture under two initial conditions: With or without stable isotope labeling by amino acids in cell culture (SILAC), referred from now on as “heavy” and “light” respectively. The heavy medium contained the C13 and N15 isotopes. All reagents were obtained from Thermo Fisher Scientific. Then, unlabeled fibroblasts were exposed to two different radiation stimuli using the X-ray irradiator 43855FC-P160 (Faxitron X-Ray LLC):

1. An acute dose of radiation (1 pulse of 2 Gy following lysis after 30 minutes).
2. An accumulative dose (4 pulses of 5 Gy, one each 12h, following lysis after 2h from the last dose).

RNA and protein samples were extracted from lysed cells following standard protocols to obtain 3 samples per condition (Control, acute dose, and accumulative dose of radiation). For proteomic analyses, each light protein extract was pooled together with a heavy protein sample, making a total of 6 control samples and 3 case samples per condition.

2. Genomics and transcriptomics

2.1. Library preparation and sequencing

Library preparation and sequencing of RNA samples was performed at the genomics unit of the Center for Genomic Regulation (CRG), Barcelona, Spain. Before library preparation, the integrity of RNA samples was checked using the Bioanalyzer system (Agilent). Two possible protocols were applied to prepare the cDNA library from total RNA samples. On the first hand, when using the mRNA selection protocol, the TruSeq Stranded mRNA Prep kit (Illumina) was employed. Briefly, this kit uses Oligo-dT beads to capture the poly-A tails of mRNA, and then fragments the captured RNA to synthetize a cDNA library in two steps, keeping the strand information through the replacement of Deoxythymidine triphosphate (dTTP) by Deoxyuridine Triphosphate (dUTP). On the other hand, when the total RNA protocol was applied, the TruSeq Stranded Total RNA Prep kit (Illumina) was used. This kit removes the ribosomal RNA using Ribo-Zero depletion, keeping the mRNA and the non-coding RNA in the sample, and then synthetizes the cDNA library with strand information as previously described for the mRNA Prep kit. Next, cDNA libraries were sequenced using an Illumina HiSeq 2500 and the High-Output v4 mode, generating between 20-60 million of single end reads with 50 base pairs (bp) per sample.

2.2. Quality control

The FASTQ format was employed to store the sequencing output, that will be referred to as “raw reads” from now on. The “trimmed reads” term will be used to refer sequence files after removing adapters and low-quality reads. FastQC (v0.11.9) was employed to assess the quality of raw and trimmed reads. A Hypertext Markup Language (HTML) report was created for each sample recapitulating the following information: Number of sequences, sequences flagged as poor quality, sequence length, percentage of guanine and cytosine (GC), per base sequence quality, per tile sequence quality, per sequence quality scores, per base sequence content, per sequence GC content, per base “N” content, sequence length distribution, sequence duplication levels, overrepresented sequences, and adapter content. Similarly, FastQ Screen (v0.13.0) was employed to check cross species contamination in raw reads by performing a multi-genome alignment (139). To do so, a subset of 100000 reads was generated from each FASTQ file and aligned against a set of reference genomes using Bowtie2 (v2.3.5.1) (140). **Table 4** details the reference genome FASTA files that were employed to perform the quality control with FastQ Screen.

Organism	**URL**
Homo Sapiens	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.39_GRCh38.p13/GCF_000001405.39_GRCh38.p13_genomic.fna.gz
Mus Musculus	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/635/GCF_000001635.26_GRCm38.p6/GCF_000001635.26_GRCm38.p6_genomic.fna.gz
Escherichia Coli	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/635/GCF_000001635.26_GRCm38.p6/GCF_000001635.26_GRCm38.p6_genomic.fna.gz
Acholeplasma Laidlawii	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/018/785/GCF_000018785.1_ASM1878v1/GCF_000018785.1_ASM1878v1_genomic.fna.gz
Mycoplasma Arginini	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/547/975/GCF_001547975.1_ASM154797v1/GCF_001547975.1_ASM154797v1_genomic.fna.gz
Mycoplasma Fermentans	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/186/005/GCF_000186005.1_ASM18600v1/GCF_000186005.1_ASM18600v1_genomic.fna.gz

Mycoplasma Hominis	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/085/865/GCF_000085865.1_ASM8586v1/GCF_000085865.1_ASM8586v1_genomic.fna.gz
Mycoplasma Hyorhinis	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/313/635/GCF_000313635.1_ASM31363v1/GCF_000313635.1_ASM31363v1_genomic.fna.gz
Mycoplasma Orale	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/420/105/GCF_000420105.1_ASM42010v1/GCF_000420105.1_ASM42010v1_genomic.fna.gz

Table 4. List of reference genomic FASTA files employed in the FastQ Screen analysis.

2.3. Read filtering and trimming

The trimming of raw reads was performed with Trimmomatic (v0.36) (60) or with fastp (v0.20.0) (61). When using Trimmomatic, the following parameters were employed to remove adapters and low-quality reads: A sliding window of 4 bp with a threshold of a minimum phred quality of 15, a leading and trailing window size of 3 and a minimum read length of 36 bps. For fastp, the following parameters were employed to perform the quality, length, and adapters filtering: A minimum phred quality of 15 to classify a base as qualified, a maximum proportion of unqualified bases of 40% per read, a minimum length of 15 bps per read, and the default strategy of searching for adapters on the first million reads on single end files and at overlapping pairs in paired end files.

2.4. Alignment

Two alternative aligners were used to perform the mapping of trimmed reads against reference genomes. The first option consisted on the usage of STAR v(2.7.0b) (63), a RNA-Seq aligner that uses sequential maximum mappable seed search in uncompressed suffix arrays followed by a seed clustering and stitching procedure. The second option employed to perform the alignment was HISAT2 (v2.1.0) (141), a novel graph-based alignment tool. HISAT2 implements a Hierarchical Graph FM index (HGFM) approach, in which the genome is computationally employed as a graph, where nucleotides are represented as consecutive nodes and single nucleotide polymorphisms (SNP) or insertions and deletions (indels) are represented as alternatives paths within the sequential graph. For both aligners, default parameters were used for the creation of the genome indexes and alignment. The human and mouse genomic FASTA files obtainable from ENSEMBL were used as reference.

Particularly, the employed assemblies were GRCh38 for human and GRCm38 for Mouse, with unmasked repeating regions (**dna.primary_assembly.fa.gz*).

2.5. Variant calling and annotation

The genomic information contained in aligned RNA-Seq reads was retrieved through variant calling using the Genome Analysis Toolkit (GATK) (v4.1.3.0) (142). Hence, a dedicated variant calling sub-workflow was designed following the GATK best practices for variant calling from RNA-Seq data. Like in variant calling from DNA sequencing, the first step in this sub-workflow consisted of marking duplicated reads, which helped to reduce the direct dependency of the depth by gene expression. Additionally, the pipeline also included other steps to deal with RNA-Seq peculiarities for variant calling, like reformatting reads to control the expansion produced by introns. Specifically, reads were split into separate reads when introns were identified inside, thus reducing artifacts in the downstream variant calling. Mapping qualities were also reassigned and adapted to match DNA conventions. Finally, in order to avoid variants called under low evidence, the sub-workflow included a step to filter by depth and only keep those variants found in at least a minimum number of reads. This parameter was set to 5, as recommended in the literature (79). The final output of this sub-workflow is a list of variants in the Variant Calling Format (VCF) for each analyzed sample.

Then, called variants were annotated using the Variant Effect Predictor (VEP) (v99) (143), which transformed the uninformative original VCF into a rich annotated VCF containing genes and transcripts affected by each variant, location and consequence of the variants on protein sequence together with two different scores that summarized the predicted impact of variants in protein stability and functionality: PolyPhen2 (144) and SIFT (137).

2.6. Gene expression quantification

featureCounts (v1.6.4) (145) was employed to extract the number of reads aligned to each genomic region from STAR and HISAT2 alignments. featureCounts was run using default parameters, and thus multi-mapping and multi-overlapping reads were not considered during the quantification. The annotation files (GTF format) corresponding to each genome file employed in the alignment were used to delimit the regions of interest (genes). featureCounts output consists of a table with the number of reads per gene and sample, also known as counts per sample matrix, where genes represent rows and samples represent columns. Alternatively to the alignment, trimmed reads were directly processed with Salmon (v0.13.0) (66), a tool for the quantification of gene expression that uses a quasi-mapping strategy, where reads are not directly mapped to a reference genome, but instead are quasi-mapped to the transcripts in a reference transcriptome, calculating the probability of each read to come from each potential transcript. To do so, the FASTA files corresponding to the cDNA of the previously mentioned genome assemblies were obtained from ENSEMBL and used as the reference for the Salmon pseudo-mapping approach. The count matrix was then generated from Salmon quantifications with the txImport (v1.10.0) R package. The "lengthScaledTPM" option was used to correct the estimated counts by both transcript length and library size.

2.7. Normalization and differential expression analysis

The normalization and differential expression analyses were performed using three alternatives protocols based on DESeq2 (v1.20.0) (72), edgeR (v3.28.0) (71), and limma (v3.46.0) (73) packages. Thus, when using the DESeq2 protocol, low expressed genes were filtered from the matrix by using a minimum cutoff of counts across all samples (by default 15). Then, the filtered count matrix was transformed into a DESeqDataSet object, including experimental groups in the model matrix. Additional variables to consider in the analyses (e.g., sample pairs in paired experimental designs) were also included in the design formula. Next, the standard DESeq2 protocol was applied, modeling the number of counts for each gene with a generalized linear model (GLM) that assumes a binomial negative distribution and estimating the log2 fold changes for each column on the model matrix. The Wald test was used to estimate the significance of the GLM coefficients. For downstream analyses and visualizations, the regularized logarithmic transformation of the count matrix was employed as normalized expression matrix.

On the other hand, when using edgeR, low expressed genes were filtered using the "filter by expression" approach implemented in the package, which was described in Chen et al. 2016 (146). Roughly speaking, this strategy keeps genes that have at least a minimum number of reads in a worthwhile number of samples. More precisely, the filtering keeps genes that have a count-per-million (CPM) value above k in n samples, where k is determined by the *min.counts* argument (by default 10) combined with the sample library sizes and n is determined by the design matrix. Thus, before starting the analysis, a design matrix was created for each analysis, reflecting the experimental groups to which each sample belongs and the additional co-variates to be modeled (as sample pairs in paired experimental designs). Then, normalization factors were calculated using the Trimmed Mean of M-values approach (TMM) (147), eliminating composition biases between libraries. The product of these factors and of the original library sizes defined the effective library size, which replaced the original library size in all downstream analyses. Next, read counts for each gene were modeled using a negative binomial distribution, where the dispersion parameter was moderated using an empirical Bayes approach. Finally, the significance of changes was assessed using a likelihood ratio test. The logarithmic (log2) transformation of the CPMs calculated after applying the TMM factors were used as normalized gene expression values. For limma, the “trend” approach was employed (148). Thus, the edgeR TMM normalized gene expression values were used to fit a linear model and then an empirical Bayes method was employed to squeeze the genewise-wise residual variances towards a global trend. P values calculated in all the methods were adjusted to control the false discovery rate (FDR) using the Benjamini and Hochberg approach (149).

3. Proteomics

3.1. SWATH label-free proteomics

The SWATH label-free proteomics sample processing and MS data analyses were conducted at the proteomics unit of the Maimonides Biomedical Research Institute of Córdoba (IMIBIC), Córdoba, Spain.

3.1.1. Sample preparation for LC-MS analysis

Lysed cells were cleaned to remove contaminants by protein precipitation with TCA/acetone and solubilized in 50 µl of 0.2% RapiGest (Waters) in 50 mM ammonium bicarbonate. Total protein was quantified using Qubit Protein Assay Kit (Thermo Fisher Scientific) and 50 µg proteins from each sample were digested with trypsin. Briefly, protein samples were incubated with 5 mM DTT at 60 ºC for 30 min, and then with 10 mM iodoacetamide at room temperature for 30 min in darkness. Sequencing Grade Modified Trypsin (Promega) was added (ratio 1:40 trypsin:protein) and samples were incubated at 37 °C for 2 h. Afterwards, trypsin was added again (ratio 1:40) and samples were incubated at 37 °C for 15 h. RapiGest was suppressed by precipitation with 0.5% TFA at 37 °C for 1 h and centrifugation. The final volume was adjusted with milliQ water and ACN to a final concentration of 0.5 µg peptide/µL (2.25% ACN and 0.2% TFA), and 1x of the iRT peptides (Biognosis AG) were spiked in each sample.

3.1.2. Creation of the spectral library

To build the spectral library, the peptide solutions were analyzed by a shotgun approach by nano LC-MS/MS. Samples were pooled and 1 µg was separated into a nano-LC system Ekspert nLC400 (Eksigent) using an Acclaim PepMap RSLC C18 column (75 µm × 25 cm, 3 µm, 100 Å) (Thermo Fisher Scientific) at a flow rate of 300 nl/min. Water and ACN, both containing 0.1% formic acid, were used as solvents A and B, respectively. The gradient run consisted of 5% to 35% B for 120 min. Peptides eluted were directly injected into a hybrid quadrupole-TOF mass spectrometer Triple TOF 5600+ (Sciex) operated with a top 65 data dependent acquisition system (DDA) using positive ion mode. A NanoSpray III ESI source (Sciex) was used for the interface between nano-LC and MS, applying a 2600 V. The acquisition mode consisted of a 250 ms survey MS scan from 350 to 1250 m/z, followed by an MS/MS scan from 230 to 1500 m/z (60 ms acquisition time, 350 mDA mass tolerance, rolling collision energy) of the top 65 precursor ions from the survey scan. The fragmented precursors were then added to a dynamic exclusion list for 15 s, excluding any singly charged ions from the MS/ MS analysis. Peptide and protein identifications were performed using Protein Pilot software (v5.0) (Sciex) with a human UniProtKB concatenated target-reverse decoy database, specifying iodoacetamide as cysteine alkylation. The false discovery rate (FDR) was set to 0.01 for peptides and proteins. Then, MS/MS spectra of the identified peptides were used to generate the spectral library for SWATH peak extraction using the add-in for PeakView Software (v2.1) (Sciex) MS/MSALL with SWATH Acquisition MicroApp (v2.0) (Sciex). Peptides with a confidence score above 99% as reported from Protein Pilot databases search were included in this spectral library.

3.1.3. SWATH analysis

Relative protein quantification was performed using SWATH. Each sample (1 µg) was analyzed using the LC-MS equipment and LC gradient described above for building the spectral library but using a SWATH-MS acquisition method. The method consisted of repeating a cycle, which carried out the acquisition of 34 TOF MS/MS scans of overlapping sequential precursor isolation windows (25 m/z isolation width, 1 m/z overlap, high sensitivity mode) covering the 400 to 1250 m/z mass range, with a previous MS scan for each cycle. The accumulation time was 50 ms for the MS scan (from 400 to 1250 m/z) and 100 ms for the product ion scan, thus making a 3.5 s total cycle time. The targeted data extraction of the SWATH runs was performed with PeakView using the MS/MSALL with SWATH Acquisition MicroApp and the spectral library created from the shotgun data. Up to ten peptides per protein and seven fragments per peptide were selected based on signal intensity; any shared and modified peptides were excluded from the extraction. The retention times from the peptides that were selected for each protein were realigned in each run according to iRT peptides (Biognosys AG) spiked in each sample and eluting along the whole-time axis; the extracted ion chromatograms were generated for each selected fragment ion using 15 min windows and 50 ppm widths. PeakView computed a score and an FDR for each assigned peptide using chromatographic and spectra components; only peptides with an FDR of < 1% were used for protein quantitation. The peak areas for peptides were obtained by summing the peak areas of the corresponding fragment ions; protein quantitation was calculated by summing the peak areas of the corresponding peptides. Finally, normalized SWATH protein abundances were obtained using MarkerView (v1.2.1) (Sciex).

3.2. SWATH areas differential abundance analysis

The differential analysis was carried out to compare the protein abundance between the different groups of samples in each study. The effect size of the protein abundance change was estimated using the log transformed mean ratio per comparison. The significance of the change was assessed by performing pair-wise group mean comparisons using a Welch Two Sample’s T-test. Resulting P values were adjusted to control the false discovery rate (FDR) using the Benjamini and Hochberg approach.

3.3. SILAC proteomics

The SILAC proteomics sample processing and MS raw data analyses were carried out at the proteomics unit of the Center for Genomic Regulation (CRG), Barcelona, Spain.

3.3.1. Protein sample preparation

Protein samples (700 µg) were dissolved in 6 M urea with mM ammonium bicarbonate, reduced with dithiothreitol (2100 nmol, 37 ºC, 60 min) and alkylated in the dark with iodoacetamide (4200 nmol, 25 ºC, 30 min). The resulting protein extract was first diluted to 2 M urea with 200 mM ammonium bicarbonate for digestion with endoproteinase LysC (1:100 w:w, 37 ºC, o/n, Wako, cat # 129-02541), and then diluted 2-fold with 200 mM ammonium bicarbonate for trypsin digestion (1:100 w:w, 37 ºC, 8 h, Promega cat # V5113). After digestion, peptide mix was acidified with formic acid and desalted with a Hypersep C18 column (The Nest Group, Inc) prior to LC-MS/MS analysis. 90 µg of each sample was fractionated by strong cation exchange chromatography (3 M, cat # 66889-U), procedure adapted from Rappsilber et al. 2007 (150). The peptides were eluted into six different concentration of Ammonium Acetate (40, 80, 120, 160, 200 and 500 mM). Each fraction was cleaned up with a MicroSpin C18 column (The Nest Group, Inc) prior to LC-MS/MS analysis. In addition, 455 µg of each sample was enriched in phosphopeptides with the High-Select™ TiO_2 Phosphopeptide Enrichment Kit (Thermo Scientific, cat # A32993).

3.3.2. Chromatographic and mass spectrometric analysis

Samples were analyzed using an LTQ-Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific) coupled to an EASY-nLC 1000 (Thermo Fisher Scientific). Peptides were loaded directly onto the analytical column and were separated by reversed-phase chromatography using a 50-cm column with an inner diameter of 75 µm, packed with 2 µm C18 particles spectrometer (Thermo Scientific). Chromatographic gradients started at 95% buffer A and 5% buffer B with a flow rate of 300 nl/min for 5 minutes and gradually increased to 25% buffer B and 78% A in 158 min and then to 40% buffer B and 65% A in 22 min. After each analysis, the column was washed for 10 min with 10% buffer A and 90% buffer B. Buffer A: 0.1% formic acid in water. Buffer B: 0.1% formic acid in acetonitrile.

The mass spectrometer was operated in positive ionization mode with nanospray voltage set at 2.4 kV and source temperature at 275 °C. Ultramark 1621 was used for external calibration of the FT mass analyzer prior the analyses, and an internal calibration was performed using the background polysiloxane ion signal at m/z 445. The acquisition was performed in data-dependent acquisition (DDA) mode and full MS scans with 1 micro scans at resolution of 120,000 were used over a mass range of m/z 350-1500 with detection in the Orbitrap mass analyzer. Auto gain control (AGC) was set to 1E5 and charge state filtering disqualifying singly charged peptides was activated. In each cycle of data-dependent acquisition analysis, following each survey scan, the most intense ions above a threshold ion count of 10000 were selected for fragmentation. The number of selected precursor ions for fragmentation was determined by the “Top Speed” acquisition algorithm and a dynamic exclusion of 60 seconds. Fragment ion spectra were produced via high-energy collision dissociation (HCD) at normalized collision energy of 28% and they were acquired in the ion trap mass analyzer. AGC was set to 1E4, and an isolation window of 1.6 m/z and a maximum injection time of 50 ms were used. All data were acquired with Xcalibur software (v3.0.63).

Digested bovine serum albumin (New england biolabs cat # P8108S) was analyzed between each sample to avoid sample carryover and to assure stability of the instrument and QCloud was used to control instrument longitudinal performance during the project (151).

3.3.3. MS data analysis

Acquired spectra were analyzed using the MaxQuant (152) software suite (v1.6.0.16) and the Andromeda (153) search engine. The data were searched against a Swiss-Prot human database (as in April 2018, 20341 entries) plus a list of common contaminants and all the corresponding decoy entries. For peptide identification a precursor ion mass tolerance of 7 ppm was used for MS1 level, trypsin was chosen as enzyme, and up to three missed cleavages were allowed. The fragment ion mass tolerance was set to 0.5 Da for MS2 spectra. Arg10; Lys8 were used as a label, oxidation of methionine and N-terminal protein acetylation were used as variable modifications whereas carbamidomethylation on cysteines was set as a fixed modification. In phosphorylated samples phosphorylation (STY) also was used as variable modifications. False discovery rate (FDR) in peptide identification was set to a maximum of 5%. The output of the MS data analysis consisted of two matrixes of intensities, one for proteins and other for phosphosites, with light and heavy intensities corresponding to stimuli and control groups, respectively.

3.4. SILAC data normalization and differential abundance analysis

First, proteins and phosphosites with an intensity value of zero in more than 35% of samples on all experimental groups were filtered. By doing so, proteins and phosphosites quantified in at least one experimental condition were retained. Then, missing values (intensity values of zero) were imputed using the sample wise minimum, as recommended in the literature (154). The resulting intensity matrixes were normalized using the Variance Stabilization Normalization approach implemented in the vsn R package (v3.58.0) (155). This normalization method, originally designed for microarray data, has shown consistent results on proteomic data (156). Then, a third matrix was created by averaging the phosphosite abundances for each protein, generating the phosphoprotein abundance matrix. Next, the differential abundance analyses for proteins, phosphosites and phosphoproteins were carried out using the limma package (v3.46.0) (73). To perform such analysis, the design formula included both the experimental condition and sample pairs, those being the “light” and “heavy” SILAC samples that were pooled together. Resulting P values were adjusted to control the false discovery rate (FDR) using the Benjamini and Hochberg approach.

4. Functional analysis and cellular signaling modeling

4.1. Prior knowledge databases

Several databases were employed to perform the functional analysis and modelling of cellular signaling circuits from omics data. The functional categories defined by the Gene Ontology annotations were used to perform ORA and FCS analyses and were obtained from the Bioconductor annotation packages (org.Hs.eg.db for human studies and org.Mm.eg.db for mouse studies) or from the Enrichr web service gene set collection (157). To assess the cancer cell line specificity in the SWATH proteomic study, a gene set collection containing highly expressed genes in each NCI-60 cancer cell line was retrieved from Enrichr. The MSigDb hallmarks were retrieved from the GSEA-MSigDb web site and used to perform FCS analyses (53). An updated version of the DoRothEA regulon collection was employed to perform the transcription factor (TF) activity analysis, employing interactions with a confidence level of A, B or C (105). A collection of kinase-substrate relationships was retrieved from the OmniPath web service, indicating which kinases or phosphatases were responsible for certain phosphorylation and dephosphorylation events, respectively (35). The list of proteins classified as cellular receptors was also retrieved from OmniPath. Signaling networks from the HiPathia R package (v2.6.0) were employed as Prior Knowledge Network (PKN) in the different approaches used to model cellular signaling circuits from transcriptomic and proteomic data. **Table 5** contains a list of the prior knowledge employed and its source URL.

Data	Source URL
Gene Ontology terms (human)	https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html
Gene Ontology terms (mouse)	https://bioconductor.org/packages/release/data/annotation/html/org.Mm.eg.db.html
Gene ontology terms and NCI-60 gene sets	https://maayanlab.cloud/Enrichr/
MSigDb Hallmarks	http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp
DoRothEA TF regulons	https://bioconductor.org/packages/release/data/experiment/html/dorothea.html
Kinase substrate interactions and cellular receptors	https://OmniPathdb.org/
Prior knowledge networks	https://bioconductor.org/packages/release/bioc/html/HiPathia.html

Table 5. Data employed and source URL for prior knowledge.

Particularly, we used a subset of the available networks in HiPathia to model signaling circuits from transcriptomic and proteomic data. Selected pathways are detailed in **Table 6**.

Selected pathway	ID
MAPK signaling pathway	hsa04010
ErbB signaling pathway	hsa04012
Ras signaling pathway	hsa04014
Rap1 signaling pathway	hsa04015
Wnt signaling pathway	hsa04310
Notch signaling pathway	hsa04330
Hedgehog signaling pathway	hsa04340
TGF-beta signaling pathway	hsa04350
Hippo signaling pathway	hsa04390
VEGF signaling pathway	hsa04370
JAK-STAT signaling pathway	hsa04630
NF-kappa B signaling pathway	hsa04064
TNF signaling pathway	hsa04668
HIF-1 signaling pathway	hsa04066
FoxO signaling pathway	hsa04068
Phospholipase D signaling pathway	hsa04072
Sphingolipid signaling pathway	hsa04071
cAMP signaling pathway	hsa04024
cGMP-PKG signaling pathway	hsa04022
PI3K-Akt signaling pathway	hsa04151
AMPK signaling pathway	hsa04152

mTOR signaling pathway	hsa04150
Calcium signaling pathway	hsa04020
Cell cycle	hsa04110
Apoptosis	hsa04210

Table 6. List of selected HiPathia pathways employed in the analysis and modelling of cellular signaling circuits.

4.2. Over representation analysis

ORA was applied in different studies, with the purpose of determining whether a functional category was over-represented in a group of selected features. Those feature groups (normally composed by genes or proteins) were selected using different criteria, like surpassing a statistical cut-off or overlapping other sets of features. For each analysis, a background feature list was prepared, normally composed by the whole proteome for a given organism or by the list of identified features in each assay. Next, a 2x2 contingency table was created for each functional category of interest, summarizing the number of features included or not in the functional category. An example of such contingency table is represented in **Table 7**.

	Selected	Not selected
In gene set	10	1000
Out of gene set	12	3000

Table 7. Example of contingency table employed in the over representation analyses.

The Fisher's exact test was then employed to test the independence of rows and columns in the contingency table with fixed marginals (158). The test was applied using the dedicated function in the stats R package or with the clusterProfiler R package (v3.18.1) (159). Resulting P values were adjusted to control the false discovery rate (FDR) using the Benjamini and Hochberg approach.

4.3. Functional class scoring

The FCS analyses were performed using several approaches. On the first hand, the *fgseaSimple* implementation of the fgsea R package (v1.16.0) was used to perform the preranked FCS analyses (160). Briefly, the input of this approach consisted of a list of features (e.g., genes or proteins) ordered by a metric of interest (e.g., effect size) and a list of functional categories (e.g., groups of functionally related genes and proteins). Next, an enrichment score was calculated for each functional category and its significance was assessed through a permutation scoring system. On the second hand, the single-sample gene-set variation analysis (GSVA) method was employed, as implemented in the GSVA R package (v1.38.2) (77). GSVA takes as input a normalized expression/abundance matrix and the list of functional categories and returns a new matrix of functional features by calculating sample-wise enrichment scores. This approach was applied to the normalized transcriptomic and proteomic matrixes using the functional categories delimited by the signaling pathways detailed in **Table 6**. The resulting GSVA scores were then compared between the sample groups of interest using limma. Finally, the VIPER algorithm was used to estimate the TF and kinase activities from transcriptomic and phosphoproteomic data, respectively (106). Limma moderated T values from genes or phosphosites were employed together with the TF and kinase regulons as the input for the viper R package (v1.24.0). Resulting Normalized Enrichment Scores (NES) were used as the estimate of the TF/Kinase activities. Only kinases and transcription factors with at least 5 substrates in the dataset were considered in this analysis.

4.4. Signal propagation algorithm

The HiPathia R package was employed (v2.6.0) to estimate the activity of the cellular signaling circuits from transcriptomic and proteomic data. HiPathia implements a signal propagation algorithm that uses input matrix values as proxies of potential protein activation values. The inferred protein activity values are then transformed into node activity values using the information on node composition taken from KEGG, which is stored in the HiPathia *metaginfo* object. Briefly, HiPathia first step consists of re-scaling the gene expression matrix to the [0,1] interval. Then, it assigns missing genes a value equal to the median of all values in matrix. Next, the signal intensity across the different receptor (nodes without inputs) to effector (nodes without output) circuits of the pathways is performed by means of an iterative algorithm based on the Dijkstra algorithm (161). The signal value is propagated according to the following recursive rule:

$$S_n = v_n * (1 - \prod_A (1 - S_a)) * \prod_I (1 - S_i)$$

Equation 1. Formula employed to perform the signal intensity propagation in HiPathia, where S_n is the signal intensity for the current node n, v_n is its normalized value, A is the total number of activation signals $S_a$arriving to the current node from activation edges and I is the total number of inhibitory signals S_i arriving to the node from inhibition edges.

Each time the signal value across a node is updated in a recursion and the difference with the previous value is greater than a threshold of 0.000001, all the nodes to which an edge arrives from the current updated node are marked to be updated. This process is repeated up to 100 times. Finally, each circuit is assigned an activity value equal to the signal intensity received by the effector node. Due to the nature of the method, the length of each circuit may influence its signal rank, and so the resulting circuit activity values are normalized by dividing them from the result of running the method with an activity value of 0.5 on all nodes. Thus, using as input a gene expression (or protein abundance) matrix, HiPathia outputs a new matrix containing the predicted activity values for all the signaling circuits in the *metaginfo* object. As a post-processing step, resulting circuits with a network diameter lower than 2 or circuits covered in less than 25% of nodes by at least one gene were filtered to avoid artifacts emerging from the original network decomposition.

To compare the resulting circuit activity values, two different strategies were used. On the first hand, the resulting values were compared between the samples of interest using a Wilcoxon signed rank test. On the other, inferred signaling circuit activities were compared between the samples of interest using limma. Resulting P values were adjusted to control the false discovery rate (FDR) using the Benjamini and Hochberg approach. HiPathia network representations were enhanced using the node betweenness as calculated by the PageRank algorithm (162).

4.5. Multi-omics mean rank

To combine the results obtained from the transcriptomic, proteomic and phosphoproteomic data in the FCS and signal propagation approaches, the results of the differential analysis for each omic layer were ranked in decreasing order using as metric the limma moderated T values. Then, the ranks of each feature in the different omic layers were averaged to derive a unique score indicating the significance and direction of change across the omics levels. This score, derived from the position and not from the significance of the changes, is robust against the differences that arise from the technological background of the different omics approaches.

4.6. Constraint-based signaling network reconstruction

A two-step inference method was employed to reconstruct a signaling network integrating the transcriptomic and proteomic information with CARNIVAL. First, the kinase and TF activities were estimated from transcriptomic and phosphoproteomic data, respectively, using the VIPER algorithm. Then, for each comparison, the multi-omic information was divided in three layers, that tried to resemble the shape of cellular signaling circuits: First, the receptors layer, with protein abundance changes and kinase activity changes included in the OmniPath list of receptors; Second, the intermediate signaling nodes, with protein and kinases values not included in the OmniPath list of receptors; and third, the effectors layer, containing the TF activity changes. For those nodes included in proteomics and inferred kinases, only kinase activity was considered. The **Figure 6** shows a schematic representation of the three layers here described.

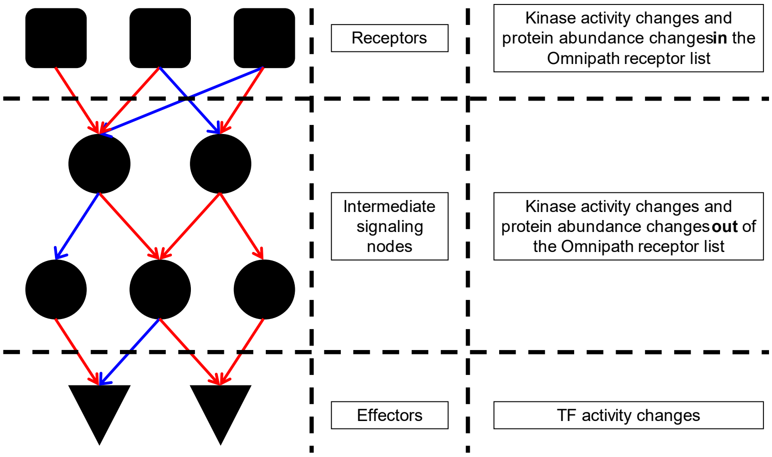

Figure 6. Schematic representation of the constraint-based approach to reconstruct signaling networks from transcriptomic, proteomic and phosphoproteomic data.

To perform the CARNIVAL analysis, all the nodes and interactions contained in the selected signaling pathways detailed in **Table 6** were combined into a shared PKN. Nodes containing more than 1 gene (complexes and multi-gene nodes) were decomposed in their individual components, maintaining all the incoming and outgoing interactions. The constraint-based network reconstruction was then carried out using the CARNIVAL R package (v1.2.0) (104). CARNIVAL includes an Integer Lineal Programming (ILP) implementation that efficiently solves an optimization problem where the objective function tries to minimize the difference between measurements and model predictions while also considering the resulting network size (with parameters named alpha and beta weights, respectively).

The optimizations were run for each comparison of interest using the sign of the receptor values as *inputObj*, the intermediate signaling nodes values scaled to the [-1,1] interval as *weightObj* and the TF activity values as *measObj*. The ILP problems were solved using the IBM CPLEX software (v12.10.0.0) obtained and used through an academic license. CARNIVAL network representations were enhanced using the node betweenness as calculated by the PageRank algorithm (162).

5. Mechanistic Integrative Analysis of RNA-Seq data

A novel bioinformatic workflow for the analysis of RNA-Seq data was developed and named: **M**echanistic **I**nte**G**rative A**N**alysis **O**f R**N**A-Seq (**MIGNON**).

5.1. Workflow implementation

The whole pipeline was developed using the Workflow Description Language (WDL) due to its flexibility, human readability, and easy deployment. WDL makes it straightforward to define complex analysis tasks, chain them together in workflows, and parallelize their execution. It contains three top-level components: Workflows, tasks, and calls. When a workflow is defined, it makes calls to a specific set of tasks. The tasks contain the commands to be executed together with the runtime environment to be used within each task. All the steps of the pipeline were wrapped into WDL tasks that were designed to be executed on independent docker containers. The tools employed to perform each of the steps that compose the pipeline were described in previous sections, so this section only details the assembly of the different components of the workflow.

Starting from raw reads, MIGNON uses **fastp** to perform the quality trimming and filtering of reads. Then, **FastQC** is applied to create a quality report for each trimmed read file. After the quality control step, five execution modes can be selected, starting all of them from the trimmed reads. Each execution mode employs a different combination of "core" tools to perform the alignment or pseudo-alignment of pre-processed reads, as explained in MIGNON's documentation. These core tools are **STAR**, **HISAT2**, **Salmon** and **featureCounts**. The core tools employed by each execution mode and its capabilities are detailed in **Table 8**.

When **Salmon** is employed to quantify gene expression, the **txImport** R package is used to create the count table from salmon intermediate files. The output of the quantification process, which is the counts per sample matrix, is then normalized using the **edgeR** TMM approach, creating the normalized gene expression matrix. All execution modes where an alignment is performed create an intermediate BAM file that is used for the downstream variant calling, enabling the detection of SNPs and indels from aligned RNA-Seq reads. Due to the number of intermediate steps carried out during this process, it was encapsulated on an independent WDL sub-workflow run at the sample level. In this sub-workflow, **GATK** and **VEP** are used to call and annotate the genomic variants, respectively.

Execution mode	Alignment	Quantification	Allows variant calling
"salmon-hisat2"	HISAT2	Salmon	Yes
"salmon-star"	STAR	Salmon	Yes
"hisat2"	HISAT2	featureCounts	Yes
"star"	STAR	featureCounts	Yes
"salmon"	-	Salmon	No

Table 8. MIGNON execution modes.

The workflow can be executed in personal computers or in high-performance computing (HPC) environments, both locally or in cloud-based services with **cromwell**, a Java based software that control and interpret WDL, using a JSON file as input. To run MIGNON, three dependencies are required: **Java** (v1.8.0), cromwell, and an engine able to run the containerized software (such as **Docker** or **Singularity**). The list of docker containers employed by MIGNON can be found in **Table 9.**

Software	Version	Docker container URL
fastp	v0.20.0	quay.io/biocontainers/fastp:0.20.0
fastqc	v0.11.5	biocontainers/fastqc:v0.11.5_cv4
samtools	v1.9	quay.io/biocontainers/samtools:1.9
HISAT2	v2.1.0	quay.io/biocontainers/hisat2:2.1.0
STAR	v.2.7.2b	quay.io/biocontainers/star:2.7.2b
salmon	v.0.13.0	quay.io/biocontainers/salmon:0.13.0
GATK	v4.1.3.0	broadinstitute/gatk:4.1.3.0
picard	v2.20.7	broadinstitute/picard:2.20.7
VeP	v99	ensemblorg/ensembl-vep:release_99.1
txImport	v1.10.0	quay.io/biocontainers/bioconductor-tximport:1.10.0
edgeR	v3.28.0	quay.io/biocontainers/bioconductor-edger:3.28.0
HiPathia	v2.2.0	quay.io/biocontainers/bioconductor-HiPathia:2.2.0

Table 9. List of docker containers employed by MIGNON.

5.2. Integrative mechanistic signaling pathway activity analysis

The HiPathia signaling propagation model is employed to perform the functional analysis in MIGNON, either using transcriptomic data alone, or integrating it with the genomic data in the so called integrative mechanistic analysis. Since the model is mechanistic, it allows to infer the effect of an intervention (e.g., a knock-out) on the resulting signaling (and functional) profile, a concept that can easily be assimilated to a loss of function (LoF). In practical terms, MIGNON considers that a gene harbors a LoF if it presents at least one genomic variant with a SIFT score < 0.05 and a PolyPhen score > 0.95 (default values). When this occurs, an *in-silico* knock-down is simulated by multiplying the scaled normalized expression values of affected gene-sample pairs by 0.01 before applying HiPathia, like described in Peña et al. 2019 (138). Hence, resulting inferred signaling circuit activity values integrate both types of data (genomic and transcriptomic) into a unique functional quantitative value.

5.3. MIGNON performance evaluation and proof-of-concept

To assess MIGNON performance and resource consumption, the workflow was executed over 6 different human datasets, comprising a total of 42 samples. The test runs were performed in a HPC environment, using cromwell (v47) and singularity (v3.5). To test resource consumption in response to parallelization, 6 different CPU configurations were employed on tasks allowing multi-threading: 1, 2, 4, 8, 16 and 24 threads. The memory and time consumed by each call was retrieved using the *sacct* command from the Slurm Workload Manager. **Table 10** details the datasets employed to perform the performance analysis and their accession in the European Nucleotide Archive (ENA).

Accession	URL	Library layout	Samples
PRJEB34009	https://www.ebi.ac.uk/ena/data/view/PRJEB34009	Single end	6
PRJEB18808	https://www.ebi.ac.uk/ena/data/view/PRJEB18808	Single end	6
PRJEB35799	https://www.ebi.ac.uk/ena/data/view/PRJEB35799	Single end	6
PRJEB27811	https://www.ebi.ac.uk/ena/data/view/PRJEB27811	Paired end	9
PRJEB22671	https://www.ebi.ac.uk/ena/data/view/PRJEB22671	Paired end	9
PRJEB30062	https://www.ebi.ac.uk/ena/data/view/PRJEB30062	Paired end	6

Table 10. List of datasets employed to perform the MIGNON performance evaluation.

To evaluate how the proposed strategy affects the predicted signaling circuit activities, two different runs of MIGNON were carried out over 462 unrelated human lymphoblastoid cell line samples from the 1000 genomes sample collection, corresponding to the CEU, FIN, GBR, TSI and YRI populations (163). In the reference run, only transcriptomic information (raw) was used, while for the case run the integrative analysis strategy was applied. Raw RNA-Seq reads were retrieved from the ENA project with accession **PRJEB3366**.

6. Data and code availability

Raw mass spectrometry proteomic data employed at the prostate cancer study were deposited in the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier **PXD010993**. Raw mass spectrometry data employed in the radiated fibroblast study have also been deposited in the PRIDE repository under the accession **PXD025656.** Raw RNA-Seq data for the systemic sclerosis study have been deposited in the Gene Expression Omnibus (GEO) under the accession **GSE121659.** RNA-Seq data on the radiated fibroblast study have been deposited in ArrayExpress and can be accessed using the following identifier **E-MTAB-10456**.

The functions employed to perform the basic differential expression/abundance analysis of omics data, as well as the classic functional analysis approaches (ORA and FCS) were unified in an R package that can be obtained from the following repository:

https://github.com/martingarridorc/biokit

MIGNON WDL code together with a bash script ready to perform a dry run of the workflow can be accessed in the following repository:

https://github.com/babelomics/MIGNON/

The following link leads to a detailed version of the MIGNON documentation, which includes examples, how to prepare the inputs and the format of the outputs:

https://babelomics.github.io/MIGNON/

All the analyses carried out to model signaling circuits from transcriptomic and proteomic data were formatted as a *workflowr* webpage, that also includes a link to the source code, and which can be accessed in the following link:

https://martingarridorc.github.io/skinxcare_workflow/

Results

1. MIGNON

At the beginning of this thesis work, we developed a basic workflow for the analysis of RNA-Seq data. First, because it was needed for the main dataset employed during the thesis, and second, because the analysis of this type of data was a bottleneck in projects carried out within both research groups. Hence, we reviewed and assembled a basic workflow containing the state-of-the-art tools for the analysis of RNA-Seq data, that included steps like alignment of reads, quantification, and differential expression analysis. However, during the review process, we noticed that most of the modern available workflows were not employing the genomic information extractable from RNA-Seq data, nor were using current approaches for the functional analysis of the transcriptomic information. Because of this, we developed a new method to carry out the **M**echanistic **I**nte**G**rative A**N**alysis **O**f R**N**A-Seq data, which we named **MIGNON**.

1.1. A novel approach for the analysis of RNA-Seq data

By doing so, we created the first publicly available workflow not only able to extract the genomic and transcriptomic information from RNA-Seq data, but also capable of integrating both levels within a mechanistic signaling framework. MIGNON covers the whole process from raw reads using state-of-the-art tools and produces two key matrixes, one containing the normalized gene expression, and other containing the gene/sample combinations that harbor loss of function (LoF) mutations. Then, MIGNON combines these two matrixes by using an *in-silico* knock-down approach, where the gene expression of gene/sample pairs affected by a LoF mutation is multiplied by a knock-down factor (0.01 by default). On a final step, MIGNON employs the HiPathia model to perform the mechanistic pathway analysis. The propagation algorithm implemented in HiPathia stops the signal propagation in perturbed nodes and thus summarizes both the genomic and transcriptomic information in a single and mechanistic output. This output, which consists of a signaling circuit activity matrix, can be further analyzed using comparative approaches or supervised and unsupervised classification methods. In addition, we implemented MIGNON using portable and modular design that emerged from the combined usage of the Workflow Description Language (WDL) and software containers. **Figure 7** represents the chain of tasks and tools that we employed to build MIGNON.

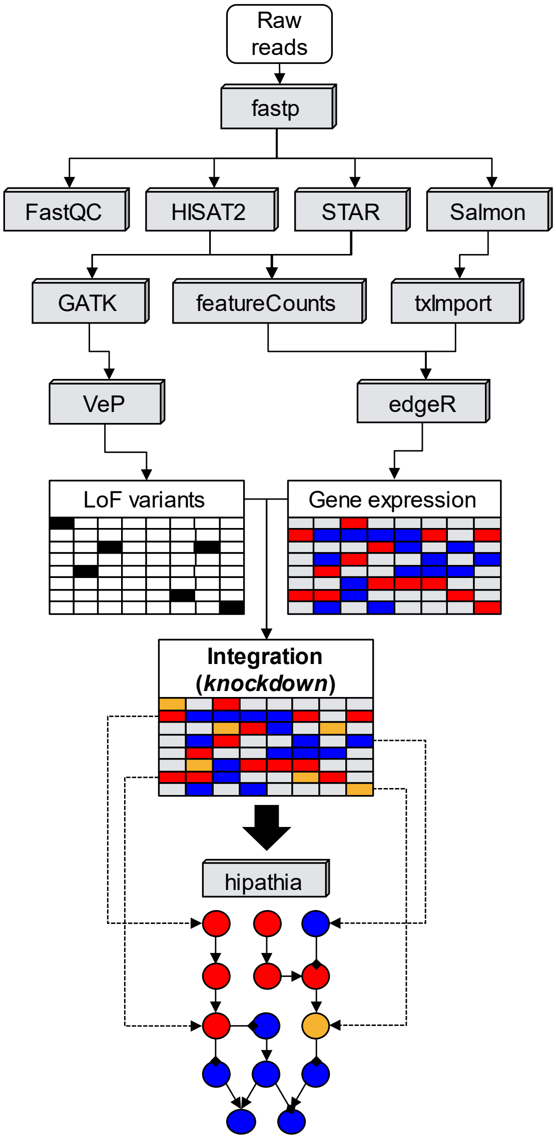

Figure 7. Schematic representation of MIGNON. Each grey box represents a piece of containerized software. The figure depicts the analysis steps from raw reads to the two matrixes that are combined: the loss of function (LoF) variants matrix (left) and the normalized gene expression matrix (right).

To evaluate how the proposed strategy affects the inferred signaling circuit activities, we carried out two different runs of MIGNON over 462 unrelated human lymphoblastoid cell line samples from the 1000 genomes sample collection, corresponding to the CEU, FIN, GBR, TSI and YRI populations. In the reference run, only transcriptomic information (raw) was used, while in the case run, the integrative knock-down strategy was applied. **Figure 8 A** and **B** depicts how the knock-down due to LoF mutations interrupted the transduction of the signal in three circuit/sample pairs. Moreover, **Figure 8 C** shows that the overall predicted signaling circuit activities were significantly lower (paired Wilcoxon signed-rank test P value < 2.2×10^{-16}) when the genomic information was integrated in the model. This example shows how the use of transcriptomic data alone produced an incomplete picture of the real signaling activity and proves the usefulness of the proposed multi-omic data integration.

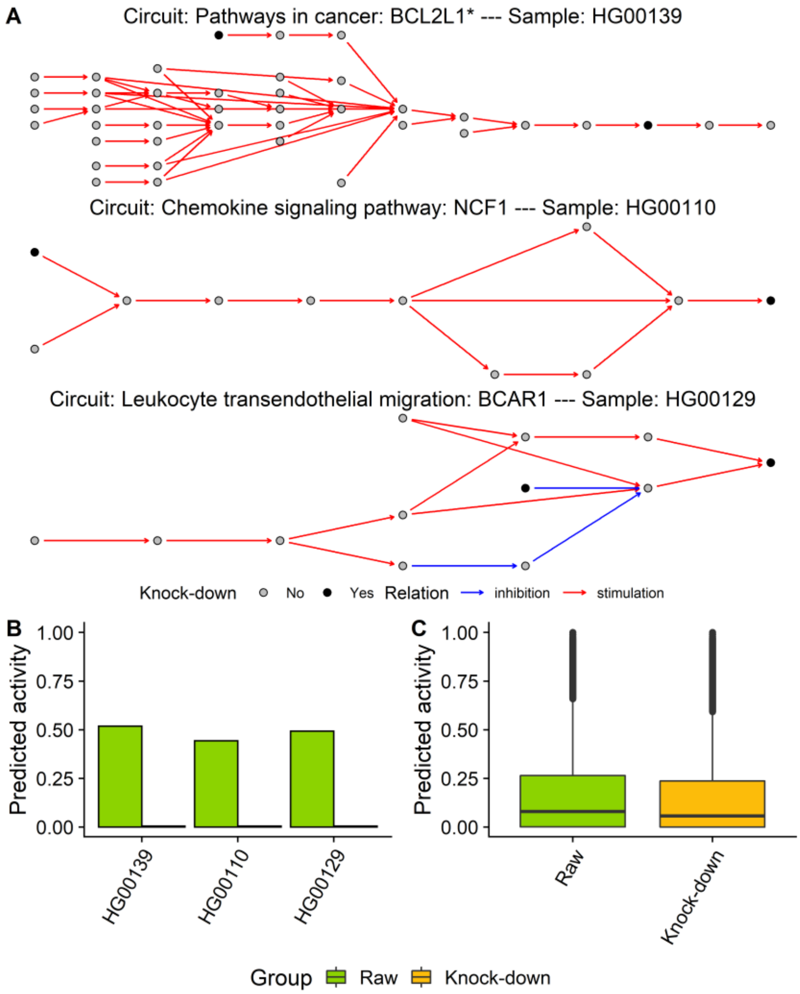

Figure 8. *In-silico* knock-downs effect on predicted signaling circuit activities. **A)** Network representation of three signaling circuits that contain genes with LoF variants for three subjects from the 1000 genomes cohort. Node color indicates whether a gene contained on in it has a LoF variant (black) or not (grey). Red and blue arrows indicate stimulations and inhibitions, respectively. **B)** Predicted signaling activity for the three circuit/sample pairs on the on the sub-figure A. Color represents signaling circuit activity with and without considering the genomic information (raw and knock-down, respectively). **C)** Box plots showing the whole predicted signaling circuit activities with and without the genomic information for the 1000 genomes cohort (paired Wilcoxon signed-rank test P value < 2.2×10^{-16}).

1.2. Performance evaluation

Next, we decided to perform a small benchmark to measure MIGNON's computational requirements in terms of memory and time. Thus, we applied it to 6 different datasets, comprising a total of 42 samples, and using 6 different CPU configurations on tools that allowed multi-threading. This analysis revealed that the most limiting steps of the entire pipeline are the alignment steps (HISAT2 and STAR) and the *MarkDuplicates* and *HaplotypeCaller* steps of the GATK sub-workflow. **Figure 9** summarizes the time and memory consumption of the tools that allow multi-threading using 6 different CPU configurations. While HISAT2 is slower than STAR, the second one makes a more intensive usage of available memory. Therefore, we decided to keep both aligners in MIGNON since this tradeoff should be considered if planning to deploy the workflow in cloud-computing based environments or, contrarily, in limited memory computing environments.

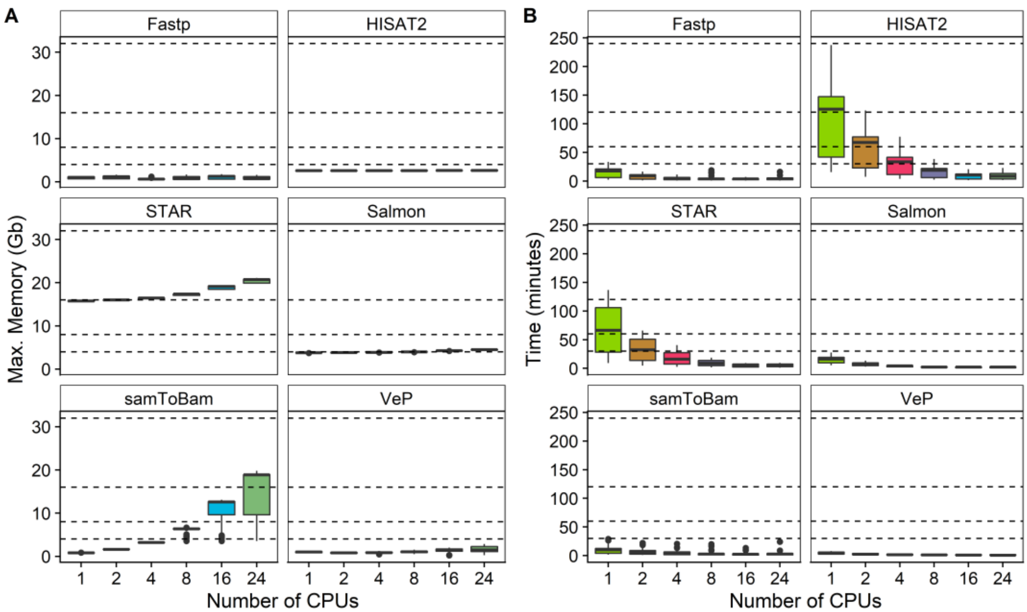

Figure 9. MIGNON performance analysis results for multi-thread tasks. **A)** Memory consumption by task. Each boxplot represents the maximum memory consumption in gigabytes (Y axis) for each CPU configuration (X axis) and task (facets). Dashed lines indicate the following memory limits: 4, 8, 16 and 32 gigabytes. **B)** Elapsed time by task. Each boxplot represents the elapsed time (Y axis) for each CPU configuration (X axis) and task (facets). Dashed lines indicate the following time points: 30, 60, 120 and 240 minutes.

In addition, **Figure 10** shows the time and memory consumption of the different steps that make up the variant calling sub-workflow. In this process, *MarkDuplicates* displays the highest memory consumption and *HaplotypeCaller* shows the longest runtime. Overall, the different tasks carried out by the workflow show a maximum memory usage under 32 gigabytes, which makes the pipeline deployable under most computational environments.

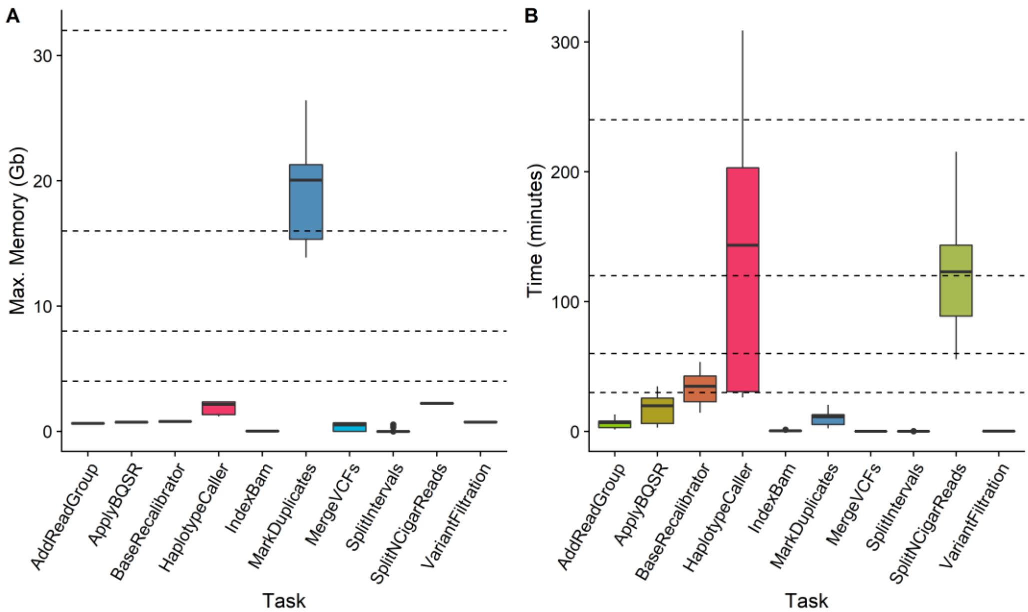

Figure 10. GATK sub-workflow performance analysis results. **A)** Memory consumption by task. Each boxplot represents the maximum memory consumption in gigabytes (Y axis) for each task (X axis). Dashed lines indicate the following memory limits: 4, 8, 16 and 32 gigabytes. **B)** Elapsed time by task. Each boxplot represents the elapsed time (Y axis) for each task (X axis). Dashed lines indicate the following time points: 30, 60, 120 and 240 minutes.

2. Functional modeling of cellular signaling circuits with transcriptomic and proteomic data

The core of this thesis is focused on creating and evaluating computational approaches to perform the functional analysis of omics data, in order to extract actionable insights about cellular signaling processes. In MIGNON, we integrate the multi-omics data in the molecular space, mainly because of the binary nature of the genomic information (harboring or not a deleterious mutation). However, transcriptomics and proteomics offer a quantitative perspective of the main players in cellular signaling, which are proteins and their activity. Hence, as shown in the introduction, a wide variety of methods have been developed to model signaling circuits from transcriptomic, proteomic and phosphoproteomic data. Most of those try to infer altered signaling circuits through the combination of omics data and prior knowledge. Moreover, some methods also provide frameworks flexible enough to integrate different omic layers during the analysis. Nonetheless, due to the recentness of this field within computational biology, there are no studies where the functional outputs from those levels are directly compared or integrated in any manner.

For this reason, we adapted and applied three different methodologies to infer altered signaling circuits from transcriptomic, proteomic and phosphoproteomic data. We contrasted the functional outputs from the different omics layers and proposed new strategies to combine them. Out of the functional analysis step, we employed a fixed framework in terms of prior knowledge and statistical approaches, to make the solutions as comparable to each other as possible. The three selected methods aimed to represent three different manners of modeling altered signaling circuits from the combination of omics data and prior knowledge:

- First, standing for more traditional approaches, we employed GSVA, a functional class scoring (FCS) approach that does not consider the signaling network topology, and that uses a Kolmogorov-Smirnov (KS) like random walk statistic to summarize the molecular data into pathway scores.
- Second, representing signal propagation approaches, we selected HiPathia, a method that first digests signaling pathways into receptor-to-effector circuits and then maps molecular data to network nodes to calculate each circuit signaling activity. To do so, it employs a signal propagation algorithm able to deal with complexes and loops.
- Finally, exemplifying approaches that first extract “master regulators” from omics data and then uses them to guide the reconstruction of altered signaling networks, we adapted CARNIVAL. First, we extracted transcription factor (TF) and kinase activities from transcriptomic and phosphoproteomic data, respectively. Then, we used altered regulators together with the proteomic data to reconstruct a coherent signaling network through an optimization step.

2.1. The SkinXCare dataset

To test the proposed approaches, we used the main dataset from the SkinXCare project. SkinXCare is a public-private initiative to develop an effective treatment for radiodermatitis, a side effect of radiotherapy mainly caused by the response of the fibroblasts in the dermal layer of the skin. To study the underlying signaling mechanisms of this response, we generated a multi-omic dataset to depict the molecular reaction of human fibroblasts to X-rays. This dataset was composed by the transcriptomic, proteomic and phosphoproteomic profile of primary human fibroblasts exposed to two different doses of radiation: an acute radiation dose, and an accumulative radiation dose. **Figure 11** shows a graphical abstract of the conditions, approaches, and next steps to take in the project.

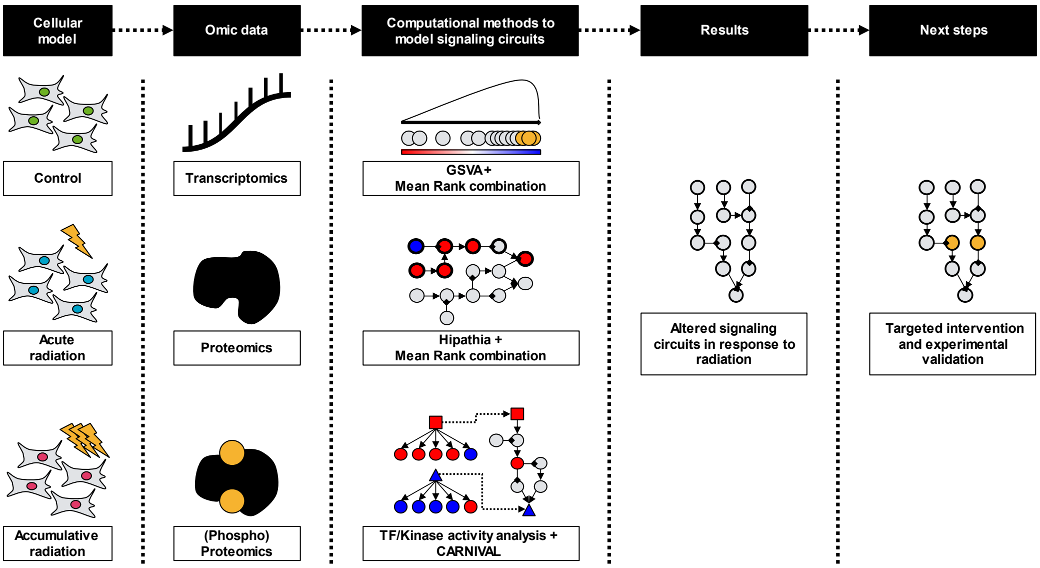

Figure 11. Graphical abstract of the initial biological material, omic data, computational methods, and next steps to take in the project.

To evaluate the changes occurring at the transcriptomic level, we employed total RNA-Seq with ribosomal RNA depletion, generating between 20 and 30 million single-end reads of 50 bp for each sample. Then, raw RNA-Seq reads were processed with MIGNON using the "hisat2" mode, employing edgeR to filter low expressed genes and to normalize the count matrix. For proteomics, we employed a SILAC labeling approach, culturing control samples with medium containing C13 and N15 isotopes ("heavy"). Before the mass spectrometry (MS) analysis, a portion of the fragmented protein extracts was enriched in phosphopeptides using Sequential Enrichment from Metal Oxide Affinity Chromatography (SMOAC) with TiO2. Next, both protein and phosphoprotein samples were analyzed through liquid chromatography coupled to MS (LC-MS/MS). Raw MS data were processed using the MaxQuant suite, and then we processed the intensity tables filtering non quantified features, imputing missing values, and applying the VSN normalization method. The **Figure 12 A** shows the final number of genes, proteins, phosphosites and phosphoproteins that were quantified in this analysis. In addition, the **Figure 12 B** represents the overlap between genes, proteins, and phosphoproteins as a Venn diagram. To generate the phosphoprotein abundance matrix, we averaged the abundance of all quantified phosphosites for each protein. As expected, many more features were quantified in the transcriptomic analysis than in the proteomics analyses.

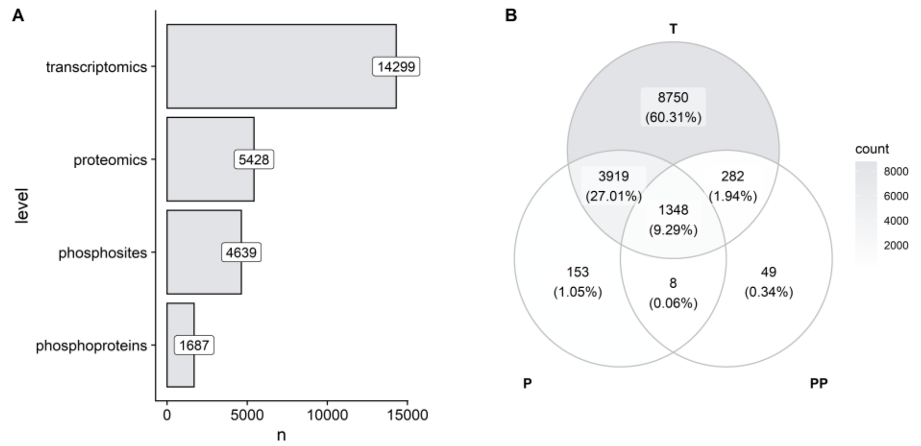

Figure 12. SkinXCare dataset overview. **A)** Bar plot representing the number of quantified features per omic level. **B)** Venn diagram showing the overlap between quantified genes, proteins, and phosphoproteins. Color indicates the number of features in each intersection. T = Transcriptomics, P = Proteomics, PP = Phosphoproteins.

2.2. Molecular moderated T values correlation

Next, we performed the differential expression/abundance analysis to evaluate the changes at the gene/protein level in response to both doses of X-ray radiation. To do so, we employed linear models as implemented in the limma package. **Figure 13** represents the results of this analysis in the form of volcano plots for each molecular level and comparison. Notably, both the significance and magnitude of the changes observed at the transcriptomic level were higher than in other molecular levels. Overall, the proportion of differential features per molecular layer comprised between 5 and 15% of all the quantified features. Then, to systematically investigate the interrelationship of the changes between the different molecular layers, we explored the moderated T value correlation among comparisons. By doing so, we summarized in a single metric (the Pearson's correlation coefficient) the agreement between the significance and direction of altered features in response to radiation. This analysis was performed for the common space of 1348 features that overlapped between genes, proteins, and phosphoproteins.

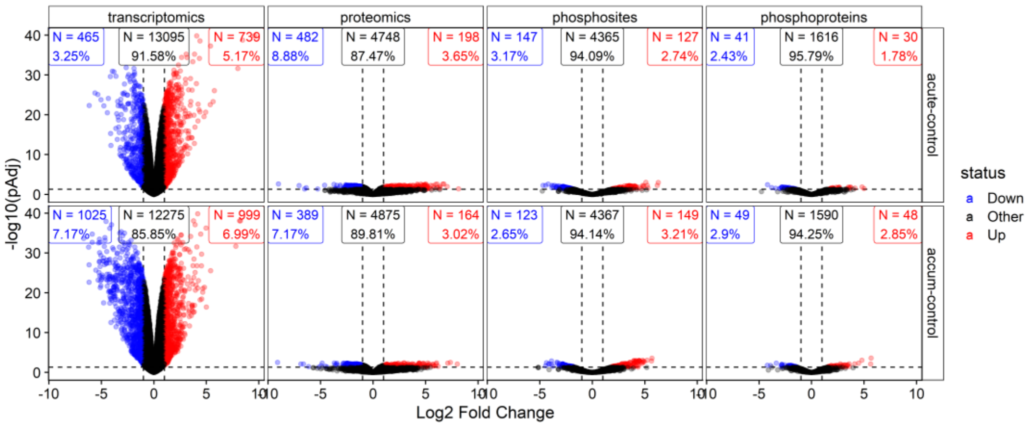

Figure 13. Volcano plots showing the differential expression/abundance analysis results at the gene/protein level. Points represent molecular features, and its color indicates whether a feature surpassed a cutoff of adjusted P value < 0.05 and log2 fold change higher or lower than 1 and -1 (as red and blue, respectively). Horizontal and vertical facets split the volcano plots by comparison and molecular level, respectively. The labels in the top of each facet indicate the number of features that surpassed (or not) the previously mentioned cutoff.

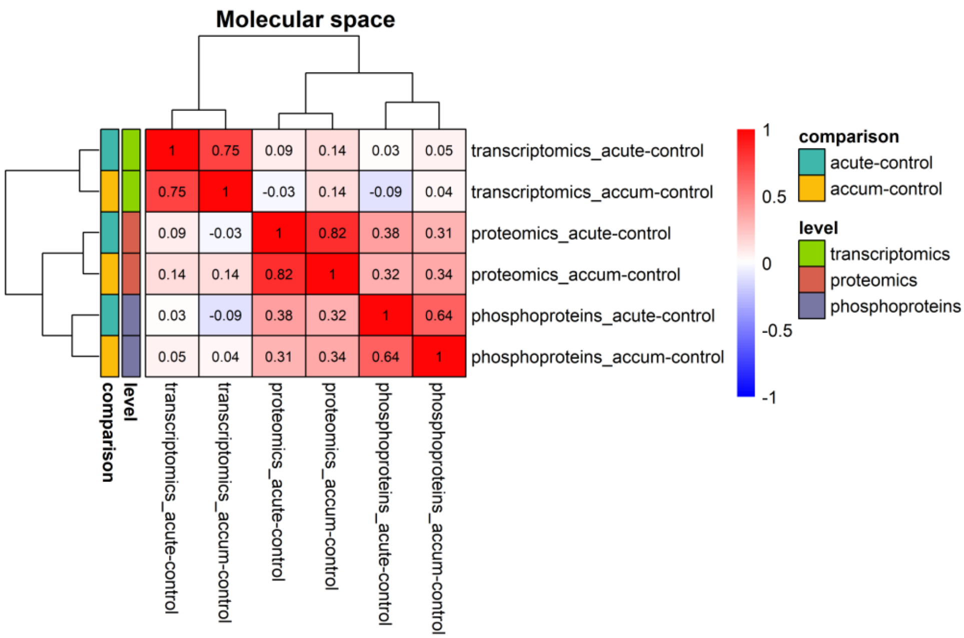

Figure 14. Moderated T value correlation matrix. Heatmap color indicates the Pearson's correlation coefficient for each pair of T value vectors. The left side colors annotation shows the comparison and omic level that produced each T value vector. The top and left side dendrograms indicate the hierarchical clustering results, using the complete linkage method with the Euclidean distance as metric.

In addition, we evaluated the similarities between levels and comparisons through the hierarchical clustering of the resulting correlation matrix. **Figure 14** shows the correlation matrix and clustering results in the form of a heatmap. Markedly, the T value correlations were dominated by assay rather than by radiation stimuli. The correlation coefficients between comparisons in the same omic levels ranged from 0.63 to 0.82, indicating that most changes induced by the acute radiation stimuli were kept in the accumulative radiation stimuli. In terms of omic levels, we observed that proteomics and phosphoproteomics had a higher correlation between them with respect to transcriptomics, with correlation coefficients between 0.31 and 0.64. Moreover, the correlation coefficients for all the pairs involving the transcriptomic level and other omic levels ranged from -0.09 to 0.14. Finally, **Figure 15** depicts the origin of the correlation matrix using a pairs plot, where the scatter plots, kernel density estimations and correlation coefficients can be traced back for each pair of T value vectors.

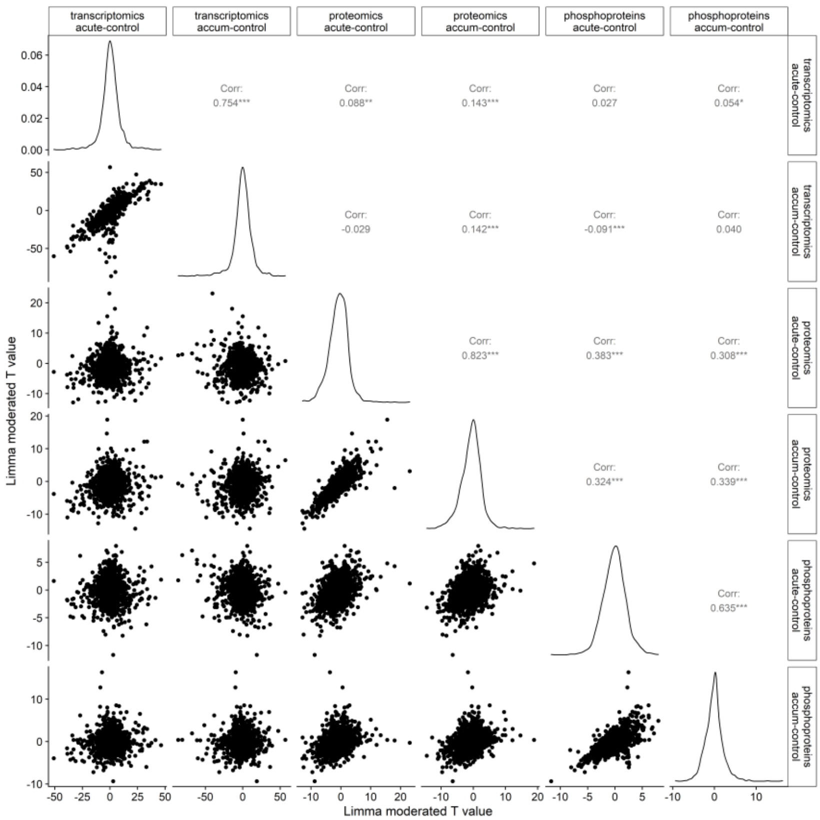

Figure 15. Moderated T values pairs plot. In the lower left section, a scatter plot is shown for each pair of T value vectors, while the Pearson's correlation coefficient is shown in the upper right corner. For correlation

coefficients, asterisks indicate the significance of the correlation: * ($P < 0.05$), ** ($P < 0.01$), and *** ($P < 0.001$). The diagonal plots represent the kernel density estimation plots for each T value vector.

In a nutshell, this first analysis allowed us to analyze the changes that occurred in the different molecular layers in response to radiation. Additionally, it also helped us to evaluate the basal agreement between such layers, exploring the correlation of the limma moderated T values in the space of common features. As expected, the transcriptomic layer quantified the highest number of features showed the most pronounced changes. On the other hand, the correlation analysis revealed that the interrelationship between T values was dominated by assay rather than by radiation stimuli, and that the overall correlation between omic layers is not very strong (with a maximum value of 0.34).

2.3. Perturbed signaling pathways using FCS

We employed different methods to infer altered signaling circuits from SkinXCare data, integrating the information from the different omics layers in a functional space. To do so, we employed the prior knowledge contained in a subset of available pathways in the HiPathia R package, originally derived from KEGG. This prior knowledge summarizes not only which genes/proteins belong to each pathway, but also the connections that occur between them in the form of nodes and links. Thus, in the first approach, we applied the GSVA algorithm in the normalized transcriptomic, proteomic, and phosphoproteomic matrixes. We ran GSVA using default parameters and a minimum gene set size threshold of 5 genes. By doing so, we transformed the original molecular space, which was different across omic levels, in a common feature space composed by the functional categories defined by the prior pathway collection.

Then, we performed the downstream comparison of the resulting normalized enrichment scores (NES) between samples as previously done with the molecular measurements. **Figure 16** depicts the results of the differential enrichment analysis as volcano plots. As it can be seen, the new feature space contained far fewer variables than the original one. In terms of significance of the changes, we observed the same pattern that occurred at the molecular level, with transcriptomics being the most altered omic layer. In addition, because 3 pathways were discarded using the set size threshold in the phosphoprotein layer, the final common pathway space was composed by 22 functional categories instead of 25.

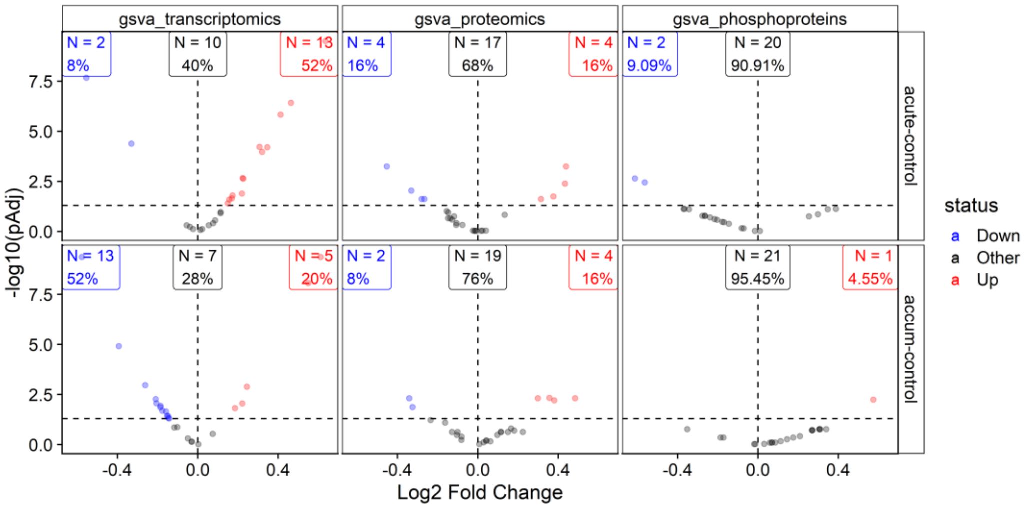

Figure 16. Volcano plots showing the differential enrichment analysis results at the pathway level. Points represent pathways, and its color indicates whether a pathway surpassed a cutoff of adjusted P value < 0.05, with color reflecting the direction of change (as red and blue for up and down, respectively). Horizontal and vertical facets split the volcano plots by comparison and molecular level, respectively. The labels in the top of each facet indicate the number of features that surpassed (or not) the previously mentioned cutoff.

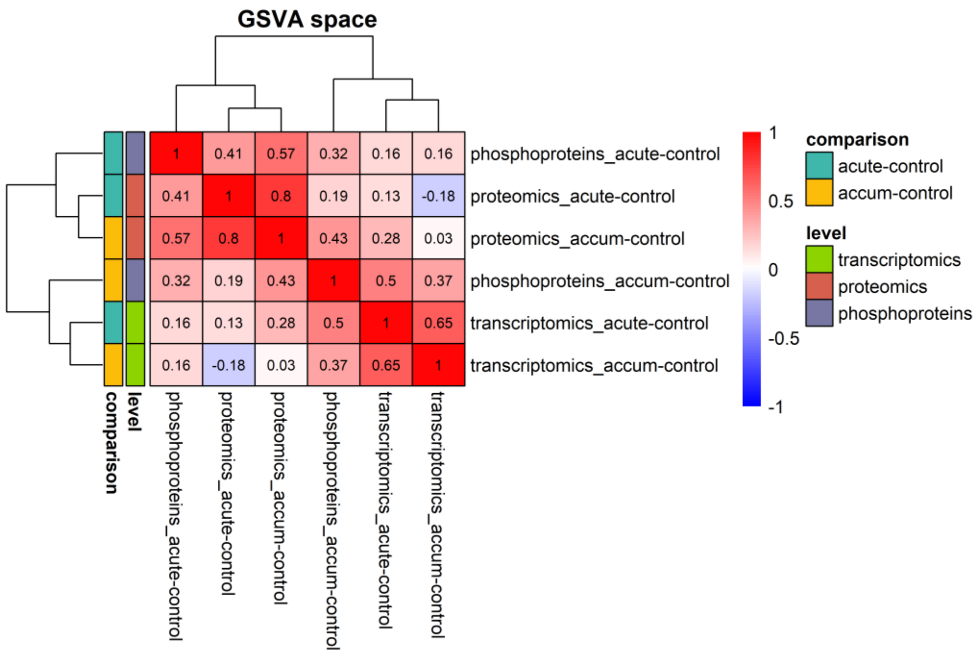

Figure 17. Moderated T value correlation matrix in the GSVA space. Heatmap color indicates the Pearson's correlation coefficient for each given pair of T value vectors. The left side colors annotation shows the comparison and omic level that produced each T value vector. The top and left side dendrograms indicate the hierarchical clustering results, using the complete linkage method with the Euclidean distance as metric.

We also explored the correlation between the moderated T values that emerged from this analysis to identify similarities and differences between molecular levels and comparisons. **Figure 17** shows the resulting T value correlation matrix. Interestingly, with respect to the previous analysis, the phosphoproteomic level was divided according to the stimulus. This was mainly caused by the correlation coefficient between the phosphoprotein vectors, which decreased from 0.63 to 0.32. On the other hand, the similarity between the changes defined by the transcriptomic and proteomic vectors was conserved in this new functional space. Finally, the correlation coefficients between transcriptomics and proteomics levels were more pronounced in this analysis, widening the previous range of -0.09 to 0.15 to a new range of -0.19 to 0.35.

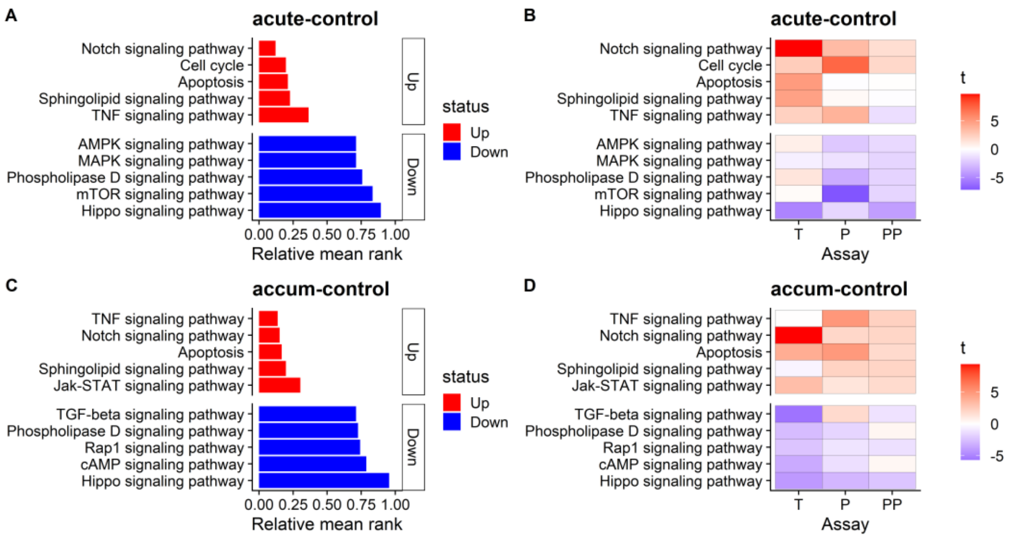

Figure 18. GSVA mean rank multi-omic combination result. Left side plots represent the relative mean rank across the three molecular levels for the top 5 up and down regulated pathways. Right side plots represent the individual limma moderated T values on each molecular level as a heatmap (T = Transcriptomics, P = Proteomics, PP = Phosphoproteins). Subfigures **A** and **B** depicts the results for the acute radiation vs control comparison and subfigures **C** and **D** for the accumulative radiation vs control comparison.

After that, we derived a single score to rank the most perturbed pathways according to the three molecular levels. To do so and being aware of the magnitude and significance differences between assays, we decided to employ a rank-based approach. Thus, we ranked the pathways on each molecular level by the limma moderated T value, placing the most up-regulated pathways at the top of the list and leaving the most down-regulated ones at the bottom. Finally, we used the mean rank across omic levels to sort the perturbed pathways. **Figure 18 A** and **C** panels show this score in a relative space for the two comparisons performed, showing the top 5 up and down regulated features. In this relative space, a score of 0 means that a pathway is the most up-regulated feature of the list in all the omics levels, and a score of 1, the same but for down-regulations. Additionally, **Figure 18 B** and **D** depicts the origin of this score, representing the moderated limma T value on each individual level and comparison for each selected pathway.

As it can be seen, the top up-regulated pathways in response to the acute and accumulative doses of radiation were the "Notch signaling pathway" and "TNF signaling pathway", respectively. On the other hand, the "Hippo signaling pathway" was the top down-regulated feature in both comparisons, showing a consistent direction of change. Additionally, other pathways like "Apoptosis" or "Sphingolipid signaling pathway" also appeared in this list for both stimuli, although the direction of change was not as consistent across omics as it was for previous pathways. Finally, the rest of pathways diverged between comparisons, indicating a stimulus-specific regulation of cellular signaling in the GSVA approach.

2.4. Perturbed signaling circuits using signal propagation

The second strategy that we employed to infer perturbed signaling circuits was based in the signal propagation algorithm implemented in HiPathia. On it, the nodes and links that form each signaling pathway are decomposed into receptor to effector circuits, and a propagation algorithm is employed to explore how an initial signal of 1 is propagated along the different nodes of the networks, using molecular measurements as proxies of the node signaling activities. Thus, we applied HiPathia over the normalized expression/abundance matrixes as previously done in the FCS approach, transforming the original molecular space into a new space of receptors-to-effector signaling circuits. As the pathways were composed by 1666 genes, missing genes were assigned a value equivalent to each matrix's median value. The proportion of estimated missing values was 0.01% for transcriptomics, 17.18% for proteomics and 46.95% for phosphoproteins. Additionally, circuits with a network diameter < 2 or with less than 25% of nodes covered by at least one molecular feature assay were filtered to avoid artifacts emerging from the network decomposition process and uncovered circuits. This filtering resulted in 423 circuits in transcriptomics, 377 circuits in proteomics and 159 circuits for phosphoproteomics. After that, we performed the differential analysis of the inferred circuit activity values as previously done for the molecular level and FCS approach. **Figure 19** depicts the results of the differential signaling activity analysis as volcano plots.

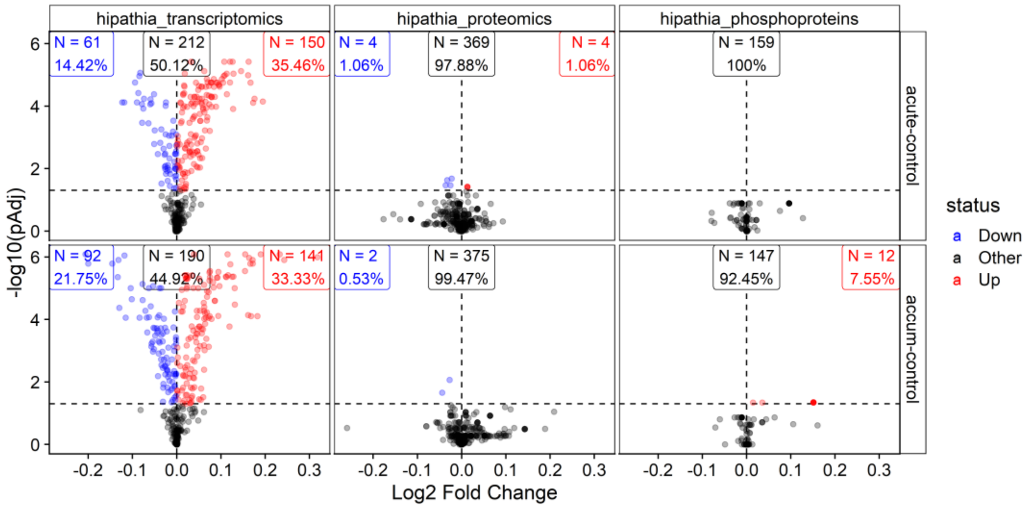

Figure 19. Volcano plots showing the differential activity analysis results at the circuit level. Points represent circuits, and its color indicates whether a circuit surpassed a cutoff of adjusted P value < 0.05, with color reflecting the direction of change (as red and blue for up and down, respectively). Horizontal and vertical facets split the volcano plots by comparison and molecular level, respectively. The labels in the top of each facet indicate the number of features that surpassed (or not) the previously mentioned cutoff.

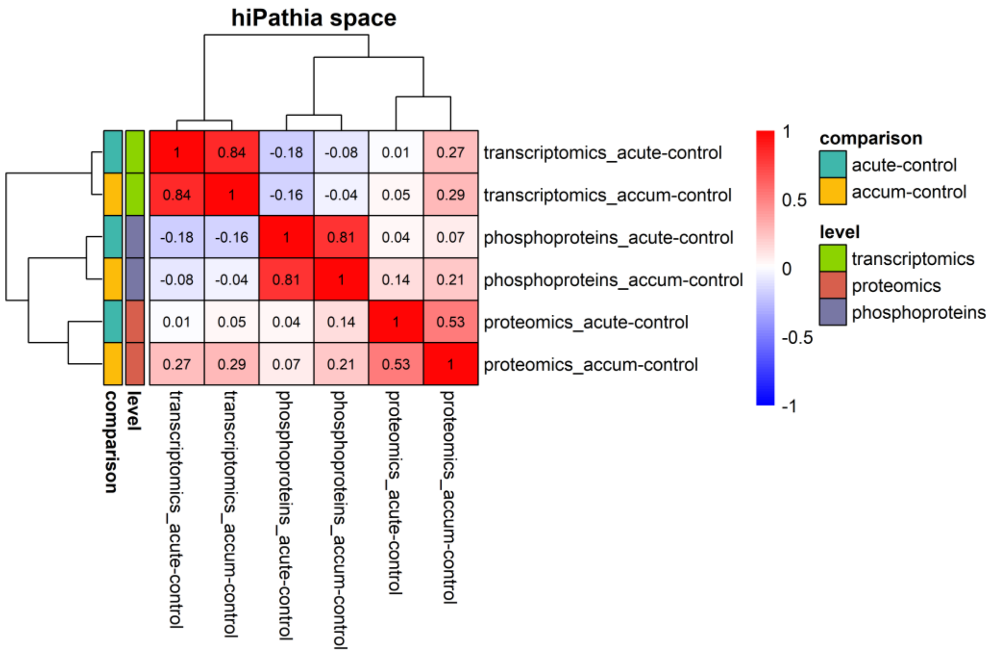

Figure 20. Moderated T value correlation matrix in the HiPathia signaling circuits space. Heatmap cells color indicates the Pearson's correlation coefficient for each given pair of T value vectors. The left side colors

annotation shows the comparison and omic level that produced each T value vector. The top and left side dendrograms indicate the hierarchical clustering results, using the complete linkage method with the Euclidean distance as metric.

Then, we evaluated the similarities between the changes across molecular levels by correlating the moderated T values. As it can be seen in **Figure 20**, the correlation coefficients that emerged from the differential activity analysis were again dominated by molecular assay rather than by comparison. In this sense, we found the strongest correlation coefficients between stimuli at the transcriptomic and phosphoproteomic levels (0.84 and 0.81, respectively), while this coefficient was lower for proteomic data (0.53). Next, we used the rank-based approach to aggregate the multi-omic results in a single score. **Figure 21 A** and **C** panels show the resulting score in the relative space for the two comparisons, showing the top 5 up and down regulated circuits, while **B** and **D** panels depicts the origin of this score, representing the moderated limma T value on each individual level and comparison per circuit.

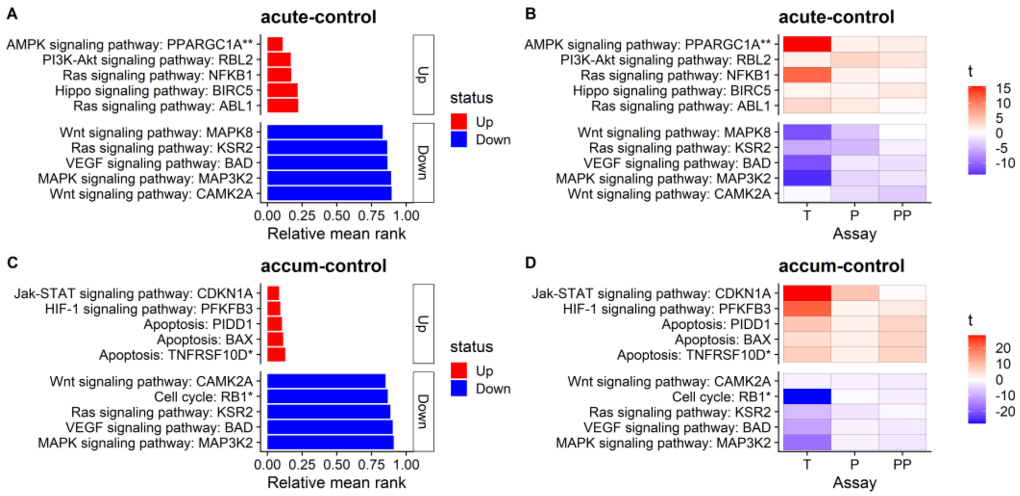

Figure 21. HiPathia mean rank multi-omic combination result. Left side plots represent the relative mean rank across the three molecular levels for the top 5 up and down regulated circuits. Right side plots represent the individual limma moderated T values on each molecular level as a heatmap (T = Transcriptomics, P = Proteomics, PP = Phosphoproteins). Subfigures **A** and **B** depicts the results for the acute radiation vs control comparison and subfigures **C** and **D** for the accumulative radiation vs control comparison.

We found very different up-regulated circuits between stimuli, belonging the top altered circuits to the “AMPK signaling pathway”, “PI3K-Akt signaling pathway”, “Ras signaling pathway” and “Hippo signaling pathway” in the acute radiation dose comparison and to the “Jak-STAT signaling pathway”, “HIF-1 signaling pathway” and “Apoptosis” pathway in the accumulative radiation dose comparison. On the other hand, we found a higher overlap between circuits in the top down-regulated circuits set, where circuits belonging to the “MAPK signaling pathway”, “VEGF signaling pathway” and “Ras signaling pathways” were shared between stimuli. In addition, to have an overview of the top altered circuits, we explored the topology of the following circuits: “AMPK signaling pathway: PPARGC1A**”, “Jak-STAT signaling pathway: CDKN1A”, “Wnt signaling pathway: CAMK2A” and “MAPK signaling pathway: MAP3K2”. A network representation of those circuits can be found in **Figure 22**, where nodes are labeled as in the original HiPathia networks, links represent stimulations and inhibitions, and node size indicates node betweenness calculated by the PageRank algorithm, which was employed to aid visual interpretation of resulting networks.

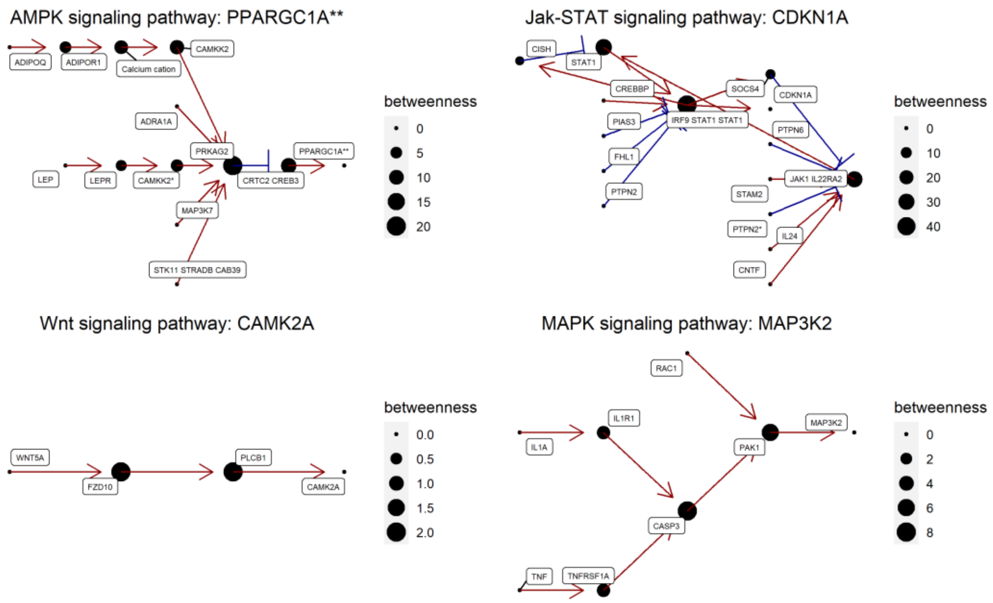

Figure 22. Network representation of top altered HiPathia circuits. Edge color represents stimulations and inhibitions in red and blue, respectively. Node size indicates each node betweenness calculated by the PageRank algorithm.

2.5. Upstream regulators analysis

The last method that we used to infer perturbed signaling circuits in response to radiation was a constraint-based network reconstruction approach. To apply it, we first inferred the activity of upstream transcription factors (TF) and kinases from transcriptomic and phosphoproteomic data, respectively. To do so, we employed regulator-target interactions obtained from prior knowledge and the VIPER algorithm as implemented in the viper R package. The TF-target interactions were retrieved from the

DoRothEA collection, including interactions with a confidence level of A, B or C. The kinase-substrate interactions with phosphosite resolution were retrieved from the OmniPath database. Next, we employed the limma moderated T value vectors obtained at each comparison for transcriptomics and phosphosites data as the input of the VIPER algorithm.

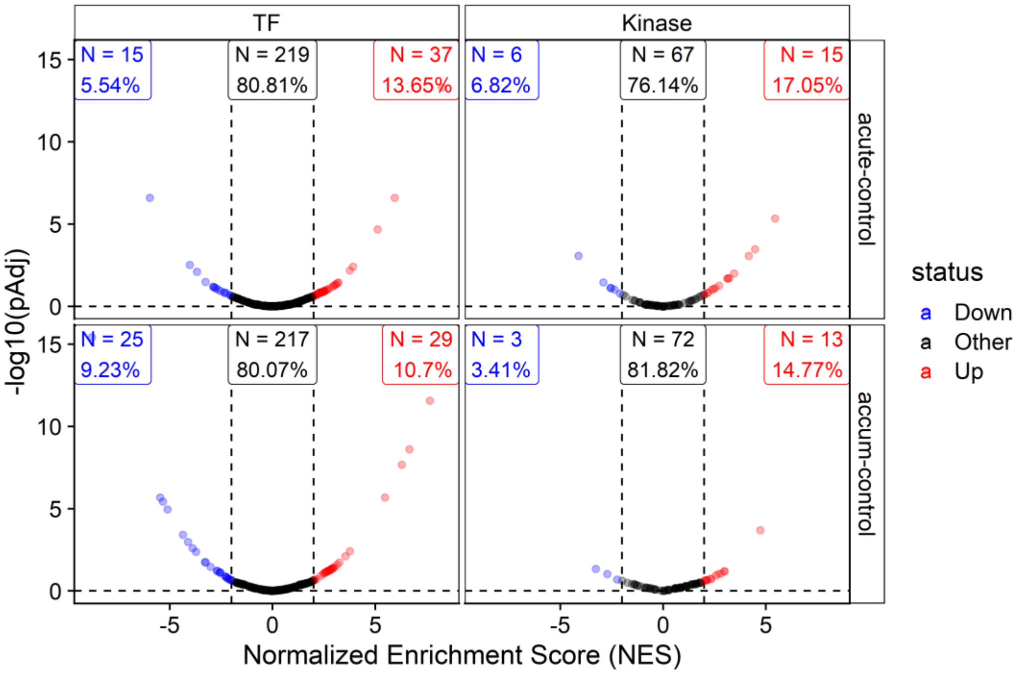

Figure 23. Upstream regulator analysis results volcano plots. The X axis represents the VIPER normalized enrichment score (NES) and the Y axis indicates the significance of the activity change for each regulator. Red and blue points indicate regulators that surpassed a cutoff of a NES higher or lower than 2 and -2, respectively. Horizontal and vertical facets split the volcano plots by comparison and regulator type, respectively. The labels in the top of each facet indicate the number of features that surpassed (or not) the previously mentioned cutoff.

This resulted in the activity estimation for 271 TFs and 88 kinases for each radiation stimuli. Then, we assessed the significance of the resulting normalized enrichment scores (NES) calculated by VIPER using the null models included in the viper R package, which were created using a parametric approach equivalent to shuffle the input features. **Figure 23** shows the results of this upstream regulator analysis in the form of volcano plots for TFs and kinases. Red and blue points indicate regulators that showed an absolute NES > 2. As it can be observed, the magnitude and significance of the inferred activity changes were higher for the TFs than for the kinases. In addition, we depicted the NES for the top altered regulators in both comparisons, which are represented in **Figure 24**. As it can be seen, we found ATM and STAT2 as the top up-regulated kinase and TF in both radiation stimuli, respectively. On the other hand, we found CSNK2A2 as the top down-regulated kinase in both comparisons, while the down-regulated TFs varied between conditions. In this sense, while in the acute-control comparison we observed TFs like ZNF263, MYC or MNT, in the accumulative-control comparison we found TFs from the E2F family (E2F4, E2F1, E2F2). Finally, others up-regulated TFs were also shared between conditions, like RELA and NFKB1. The rest of regulators diverged between stimuli.

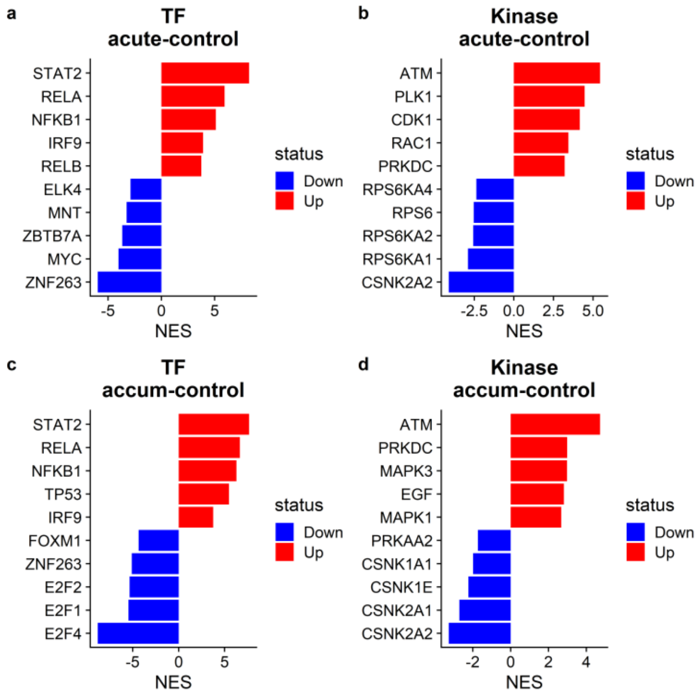

Figure 24. Top altered regulators. The bar plot reflects the top 5 altered regulators, by NES, per regulator type, comparison, and direction of change. Subfigures **A** and **C** depicts the top altered TFs, while subfigures **B** and **D** show the kinases.

2.6. Constraint-based network reconstruction

Once we performed the upstream regulators analysis, we used the resulting inferred activity values (NES) coupled to the proteomic data to perform a constraint-based signaling network reconstruction with CARNIVAL. Hence, we used an absolute NES cutoff of 2 to filter upstream regulators, and an adjusted P value of 0.05 to select proteins in the differential abundance analysis. Next, we divided those features in the three categories of CARNIVAL:

- Inputs: Formed by altered kinases and proteins that are cellular receptors.
- Weights: Formed by the rest of kinases and proteins.
- Measurements: Formed by altered TFs.

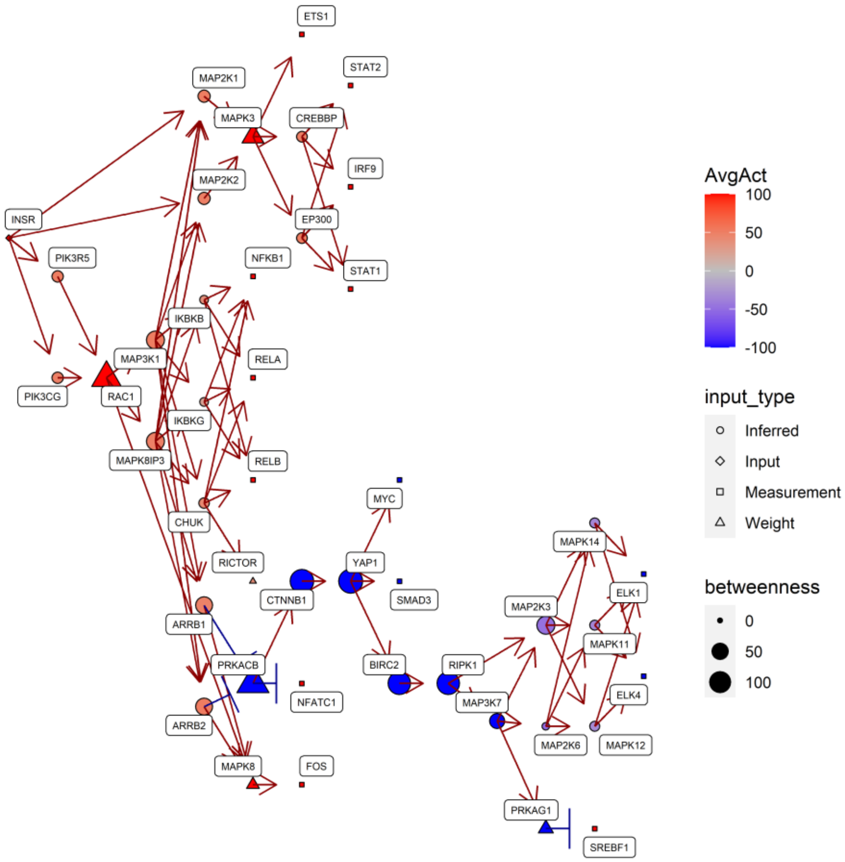

Figure 25. Signaling network obtained using the constraints defined by the altered features in response to an acute dose of radiation. Node colors represent the average activity in the pool of 100 solutions, node shape represent its type and node size indicate the betweenness score obtained from the PageRank algorithm.

Then, we ran CARNIVAL for each radiation stimuli using the CPLEX software to resolve the ILP formulation of the optimization problem, that seeks the most parsimonious solution with respect to the inputs within the PKN. For this analysis, the PKN was composed by all genes and connections that formed the HiPathia pathways object, except for the transcriptional regulation interactions that were considered in the TF activity analysis. We applied CARNIVAL with the following parameters: An alpha weight of 1, a beta weight of 0.2, an allowed number of solutions to be generated of 10000, an allowed number of solutions to be kept in the final solution pool of 100, and a time limit of 30 minutes. CPLEX solved the optimization problems with a gap value of 0.08% for the acute-control problem and of 0.97% for the accumulative-control comparison. **Figure 25** and **Figure 26** show the obtained consensus signaling networks for features derived from the acute and accumulative radiation stimuli, respectively.

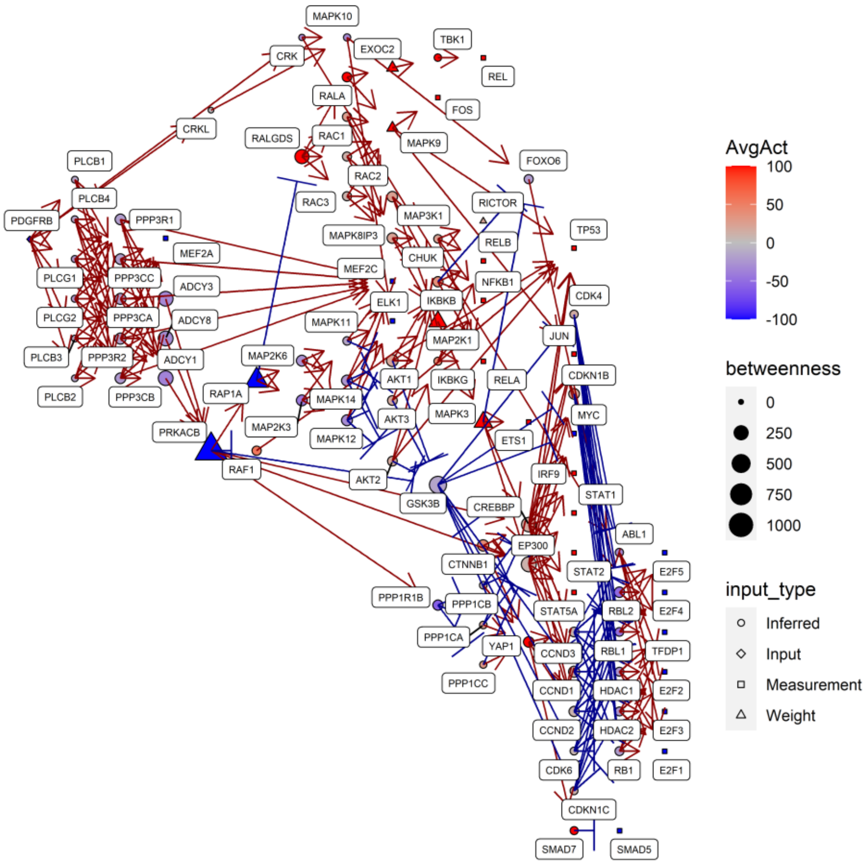

Figure 26. Signaling network obtained using the constraints defined by the altered features in response to an accumulative dose of radiation. Node colors represent the average activity in the pool of 100 solutions, node shape represent its type and node size indicate the betweenness score obtained from the PageRank algorithm.

2.7. Comparison of solutions

In a final step, we performed a systematic comparison of the inferred signaling networks obtained from each method. Thus, and given the "ranked" nature of GSVA and HiPathia combined scores, we used those scores to divide the obtained pathways or circuits in ten folds and extracted the HUGO symbols for genes/proteins included in each fold. For CARNIVAL networks, all nodes included in the final solutions were used.

Next, we computed an overlap coefficient between gene vectors for each pair of methods using the Jaccard index. This similarity coefficient has a value between 0 and 1, receiving a value of zero if the compared sets have no elements in common, and a value of 1 if the compared sets contain exactly the same elements. **Figure 27** represent the resulting Jaccard index (Y-axis) for genes included in the solutions on each step in the relative mean rank (X-axis).

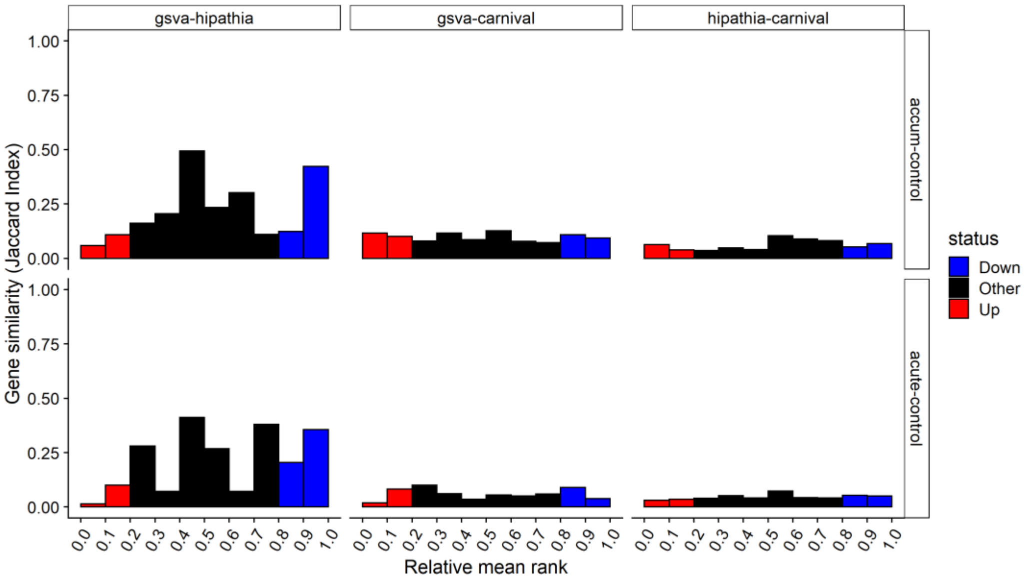

Figure 27. Stepwise comparison of resulting networks. Y-axis represent the Jaccard index between each pair of vectors containing the genes of altered signaling networks for each method comparison (vertical facets) and fold in the relative mean rank (X-axis). Horizontal facets divide the plot by radiation stimuli. Sections where up and down regulated features are most likely to be found are highlighted in red and blue, respectively.

Overall, we found a greater similarity between GSVA and HiPathia with respect to the similarity of both methods with CARNIVAL. While the Jaccard index between GSVA and HiPathia reached a value of 0.5 for some sections, the maximum value for the GSVA-CARNIVAL and HiPathia-CARNIVAL comparisons was of 0.12. Markedly, for the GSVA-HiPathia comparison, the similarity between methods was higher in the region of down-regulated features than in the sections belonging to up-regulated features.

3. A computational toolkit for biomedical research

During this thesis work, we developed a computational toolkit to perform the statistical and functional analysis of omics data in different biomedical research contexts. In addition, we automatized several tasks to retrieve data and prior knowledge from public databases, which helped to create, reinforce, or refute each project's specific hypotheses. The basic capabilities of the toolkit, implemented as an R package, are depicted in **Figure 28**. This toolkit made significant contributions in a wide variety of research lines, and the following sections detail examples of its application and use. In all projects, the computational analysis was accompanied by an experimental validation of the most important insights retrieved by the toolkit, highlighting the potential of the combined computational and experimental approaches.

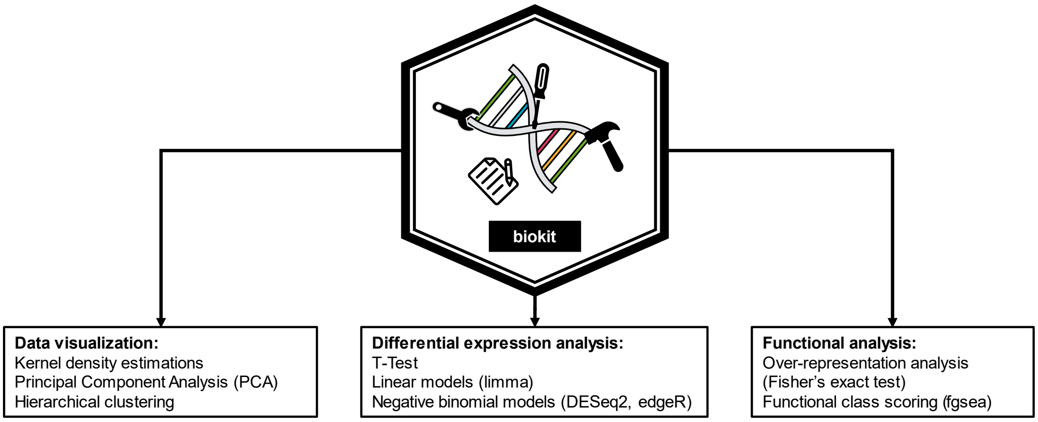

Figure 28. Biokit package capabilities. Each box indicates the analyses that can be carried out with the biokit package within the following three categories: "Data visualization", "Statistical analysis" and "Functional analysis".

3.1. Pre-processing and analysis of transcriptomic data

A very early version of MIGNON and of the toolkit were used to transform bulk RNA-Seq raw data into biological insights across different projects. In all of them, we performed the quantification of gene expression, normalization, statistical modeling, and functional analysis of transcriptomic data from cellular and animal models of human disease. As an example, we will show the results for one of those projects, focused on profiling the effect of two different treatments for scleroderma, an autoimmune pathology that affects skin through chronic inflammation and fibrosis. In this project, we employed a mouse model based in bleomycin (BLM) treatment, a drug that induces proinflammatory and fibrogenic cytokines in the skin, mimicking what occurs in human scleroderma patients. In combination with the bleomycin, we applied two different treatments based in two cannabinoids: Ajulemic Acid (AJA) and EHP-101 (EHP). Next, skin total RNA was isolated and processed using RNA-Seq in two rounds of sequencing. Thus, we obtained the transcriptomic profile for the skin of 12 mice under 4 conditions: Untreated (control), BLM, BLM + AJA and BLM + EHP (N=3 per group).

In the first step of the computational analysis, we processed the raw RNA-Seq data by performing the trimming and quality control of raw reads. Then, trimmed reads were aligned against the mouse reference genome and reads overlapping genomic features were quantified to obtain the count matrix. The resulting count matrix was filtered to remove genes without at least 15 counts across samples and normalized using DESeq2 regularized logarithmic transformation, regressing out the batch effect corresponding to the two different rounds of sequencing. Then, we applied three types of analysis to explore, compare and interpret the obtained transcriptomic profiles. First, we used a principal component analysis (PCA) to evaluate the distribution of samples in a reduced dimensional space accounting for most of the variability in the dataset. **Figure 29** shows the results of this analysis, placing samples in a bidimensional space formed by the two first principal components, that explained 65.8% of the dataset's variance. As it can be seen, samples clustered forming 4 groups, that corresponded to the experimental conditions under study.

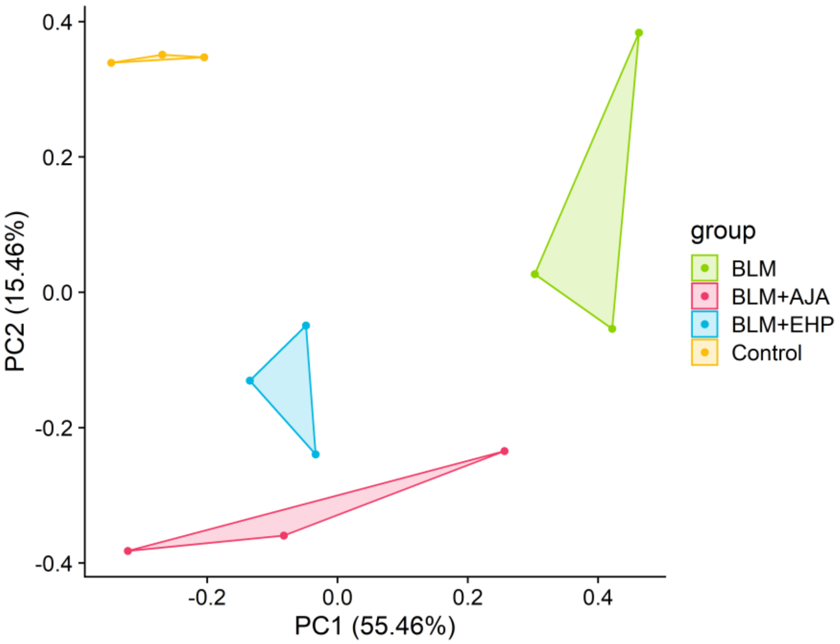

Figure 29. Principal Component Analysis (PCA) results. Each point represents a sample in the bidimensional space created by the two first principal components. Point color indicates the experimental group to which each sample belongs. The percentage showed in the X and Y axes indicate the proportion of the variance explained by each principal component.

Next, we decided to evaluate the transcriptomic changes produced by each of the treatments in mice skin. Thus, we used DESeq2 negative binomial models to perform a differential expression analysis carrying out the following comparisons: BLM treated skin versus non-treated skin (control), BLM+AJA versus BLM, and BLM+EHP versus BLM. By doing so, we evaluated the changes induced by the treatment that creates the damage and the action of the treatments that try to reduce it. The results of this analysis are represented as volcano plots in the **Figure 30**. Markedly, the changes induced by the BLM treatment in non-treated skin were the more pronounced in terms of magnitude and significance. A total of 3155 (17.26%) genes surpassed the cutoff of adjusted P < 0.05 and absolute log2 fold change > 1. On the other hand, the changes induced by both cannabinoids in BLM treated skin were less pronounced. A total of 920 (6.35%) and 564 (3.21%) genes surpassed the aforementioned cutoff for EHP and AJA treatments, respectively.

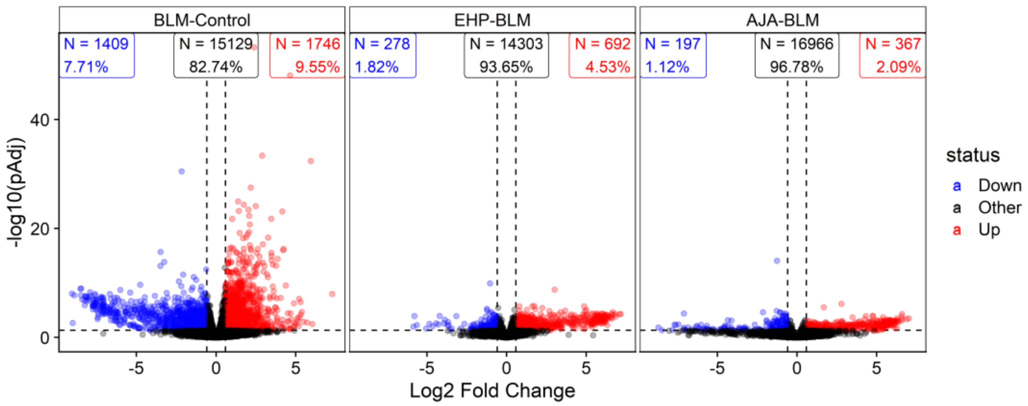

Figure 30. Differential expression analysis results as volcano plots. The X axis indicate the magnitude of the gene expression change as the log2 fold change, while the Y axis represents the significance of the change as the -log10 transformed adjusted P value. Each point represents a gene, and its color indicates whether a gene surpassed or not a cutoff of adjusted P value < 0.05 and a log2 fold change lower or higher than 1 as blue (Down) or red (Up), respectively. The facets split the volcano plots by comparison and the labels at the top of each facet indicate the number of genes that surpassed the cutoffs.

Finally, to evaluate the results in a functional context, and to guide their interpretation, we performed the functional analysis of the differential expression results using a functional class scoring (FCS) approach. Hence, for each comparison, we ranked the genes using the log2 fold change and performed a gene set enrichment analysis (GSEA) using MSigDb hallmarks as functional categories and the gene permutation approach implemented in the fgsea R package as algorithm. **Figure 31** shows the hallmarks that were enriched (adjusted P < 0.01) in the BLM versus Control comparison and in at least one of the other two comparisons. This analysis revealed that 7 of the 10 hallmarks that were enriched in one direction in BLM challenged mice followed the opposite trend after treatment with either EHP or AJA. Among them, the inflammatory response and myogenesis hallmarks, related to the pathogenesis of the disease, showed a similar pattern for both treatments, something that was subsequently confirmed through histological studies. The three hallmarks that did not follow a common trend between treatments consisted of functional categories related to hypoxia, interferon alpha and interferon gamma response.

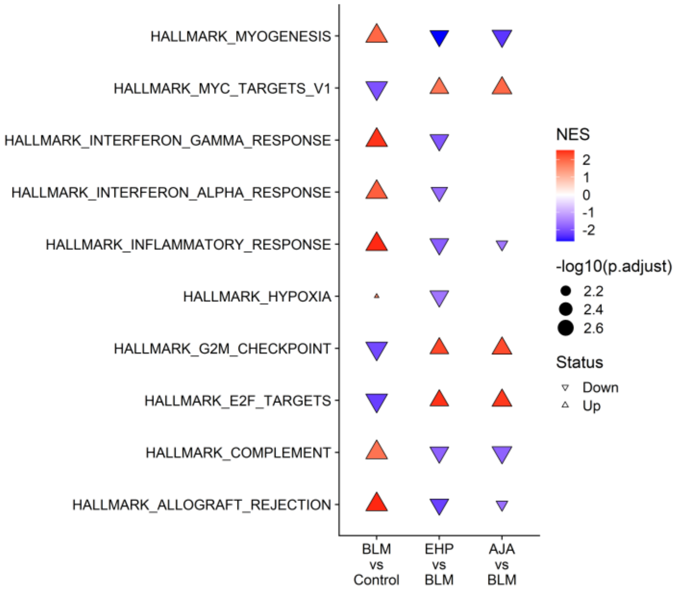

Figure 31. Functional analysis results. For each comparison of interest (X axis), the enriched MSigDb hallmarks (Y axis, Adjusted P < 0.01) are represented as points with arrow shape. Points color represents the normalized enrichment score (NES) and its size indicates the significance of the enrichment as the -log10 transformed adjusted P value.

Overall, the analysis developed in this project is an example of how to apply basic data visualization and statistical modeling techniques to evaluate the similarities and differences between novel treatments for scleroderma, using a combination of animal models and RNA-Seq. This project was followed by further experimental validations and was used as the seed for the development of a new scleroderma treatment, based in the EHP formulation. This treatment is currently in the clinical trial phase and is showing promising results for the treatment of patients suffering from this disease. On the other hand, the exemplified strategy was applied to 4 different research lines during the thesis, that resulted in 5 co-authored publications.

3.2. Analysis of proteomic data

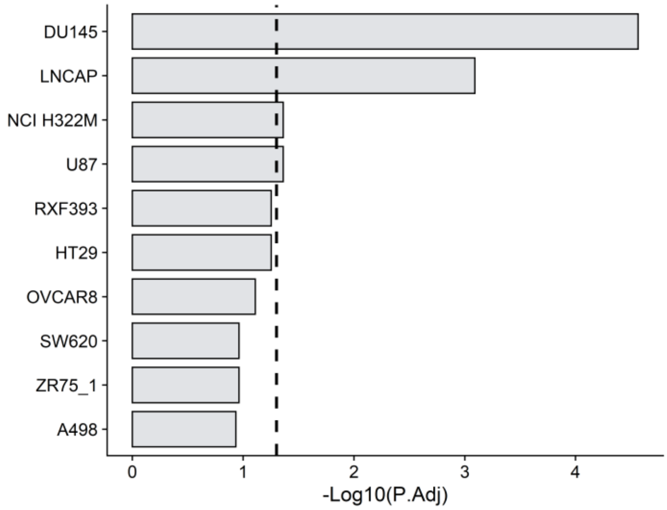

Figure 32. NCI-60 over-representation analysis results. The bar plot shows the top 10 enriched NCI-60 cell lines in the Y axis, while the X axis indicates the significance of the enrichments as the -log10 transformed adjusted Fisher's exact test P value. The dashed line. Dashed line indicates a cutoff of adjusted P = 0.05.

In a very similar way to that described above for transcriptomic data, our toolkit was used for the analysis of proteomic data generated using the SWATH label-free method in two different research lines. We will exemplify the analyses carried out for those projects showing the results of one of them in which the computational analysis made a particularly significant contribution. In this project, the main objective was to identify the mechanism of action of a potential treatment for prostate cancer, a fungal metabolite called Galiellalactone (GL). Thus, to investigate the drug's mechanism of action, we employed two different cell lines: LNCaP and DU145, representing androgen sensitive and insensitive prostate cancer cells, respectively. In the first experiment, we applied the GL treatment to both cell lines GL (10 µM) during 24h and profiled their proteome using the SWATH label-free method. In this analysis, we identified 2558 proteins and quantified a total of 1863. As a quality control step, and to check the specificity of the proteomic data, we performed an over-representation analysis of all the quantified proteins against a gene set library containing highly expressed genes in the NCI-60 cancer cell line collection. As expected, we found DU145 and LNCaP at the top of the resulting list (**Figure 32**).

Then, to explore the dataset variability in a reduced dimensional space as done before for transcriptomic data, we performed a principal component analysis (PCA).

Figure 33 shows the distribution of samples in a bidimensional space formed by the two first principal components, which explained the 54.28% of the variance in the dataset. Surprisingly, we found three groups of samples. On the one hand, the greatest difference between samples was marked by cell type (distance the PC1 axis), indicating that this is the most relevant factor in terms of variance. On the other hand, while we could observe two sample groups for DU145 cells, indicating whether cells were or not treated with GL (PC2 axis), this separation was not as clear for LNCaP samples. This led us to conclude that the data from DU145 cells seem to reflect the consequences of the perturbation caused by GL-based treatment better than LNCaP cells.

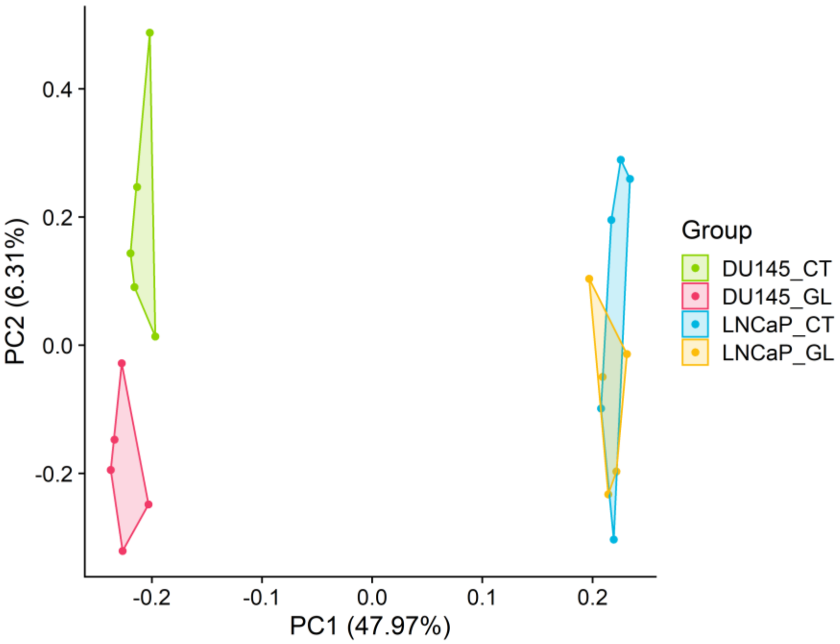

Figure 33. Principal Component Analysis (PCA) results. Each point represents a sample in the bidimensional space created by the two first principal components. Point color indicates the experimental group to which each sample belongs. The percentage showed in the X and Y axes indicate the proportion of the variance explained by each principal component.

Next, we evaluated the differences between the experimental conditions at the protein level by performing a differential abundance analysis. To do so, we employed a T-Test to compare the normalized SWATH area means between experimental groups. **Figure 34** depicts the results of this analysis as volcano plots. As expected, we found the most prominent differences between the proteomic profiles of DU145 and LNCaP cells. In this comparison, we found a 49.95% of proteins with an adjusted $P < 0.05$. On the other hand, the changes in response to the treatment followed the trend previously observed in the principal component analysis. While in DU145 cells, the GL treatment mobilized an 8.38% of proteins (P < 0.05), on LNCaP cells the treatment did not generate many changes (1.39% of proteins with P < 0.05).

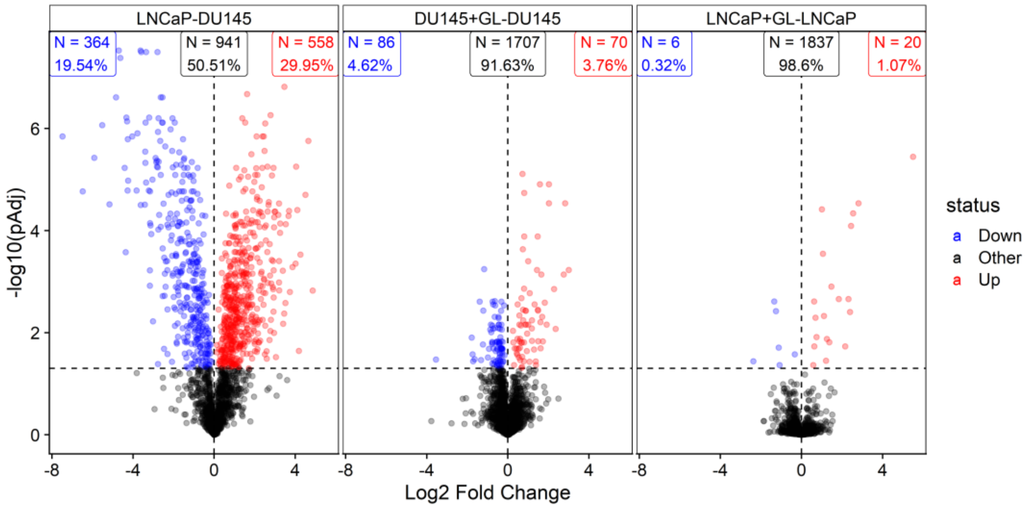

Figure 34. Differential abundance analysis results as volcano plots. The X axis indicate the magnitude of the protein abundance change as the log2 fold change, while the Y axis represents the significance of the change as the -log10 transformed adjusted P value. Each point represents a gene, and its color indicates whether a gene surpassed or not a cutoff of adjusted P value < 0.05 and a log2 fold change lower or higher than 0 as blue (Down) or red (Up), respectively. The facets split the volcano plots by comparison and the labels at the top of each facet indicate the number of proteins that surpassed the cutoffs.

With these observations, we decided to validate this differential response between cell types using cell viability assays, which showed that DU145 cells undergo cell cycle arrest after treatment, but not LNCaP cells. In addition, the experimental assays demonstrated that DU145 cells became resistant to treatment when confluence was reached. Hence, we decided to investigate potential targets of GL by performing a second proteomic analysis where DU145 cells with and without confluence status were analyzed using the same SWATH proteomic approach. In this second analysis, we quantified 2011 proteins, and 990 (49.22%) showed an adjusted P value < 0.05 when we compared the SWATH areas between confluent and non-confluent cells. On a final analysis, we searched potential GL targets between the proteins that were significantly altered by the GL treatment in DU145 cells and that show significant changes in the cell line and confluence comparisons. This analysis revealed 40 shared proteins between the three comparisons, as represented in the Venn diagram shown at **Figure 35** and in the heatmap presented in **Figure 36**.

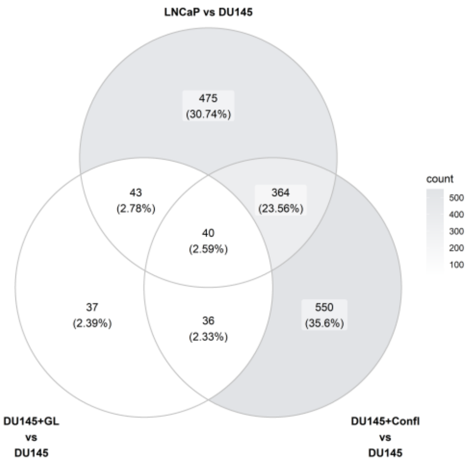

Figure 35. Venn diagram showing the overlap between differentially altered proteins (T-Test adjusted P value < 0.05) in the three comparisons of interest.

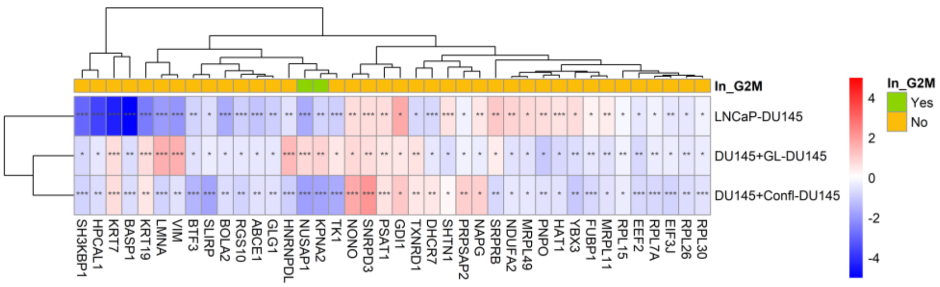

Figure 36. Heatmap depicting the differential abundance analysis results for the 40 proteins overlapping the three comparisons of interest (columns). Cell colors indicate the log2 transformed fold change, and cell labels the significance of the differential abundance analysis for each comparison (rows): * P < 0.05, ** P < 0.01, *** P < 0.001. Column annotation indicates which proteins are included in the MSigDb hallmark "G2M checkpoint".

As we had previous evidence that GL treatment could affect the transition from the G2 to M phases of the cell cycle, we decided to evaluate the proteins from the resulting list that were included within the MSigDb "G2M checkpoint" functional category. From the 40 proteins, 2 were part of this hallmark: NUSAP1 and KPNA2. A final experiment, where NUSAP1 was silenced, revealed that this protein could be responsible for the GL treatment responsiveness in DU145 cells, stablishing it as a potential drug target.

3.3. Analysis of metabolomic data

Although initially designed for transcriptomic and proteomic data, the toolkit that we developed contains generic-purpose tools that are applicable to other types of omics data. As an example, here we show its application to a dataset that we generated to study the metabolomic properties of lung cancer. This dataset is, to our knowledge, the largest metabolomic dataset generated from solid tissues in lung cancer. It covered two different types of tumors: Adenocarcinomas and squamous cell carcinomas. For each tumor sample analyzed, we also obtained the metabolomic profile of paired normal adjacent tissue. The total dataset was composed by 136 lung tissue samples and a total of 851 metabolites were identified. After quality control, 439 and 798 metabolites were quantified for adenocarcinoma and squamous cell carcinoma samples, respectively. In the first analysis, we visualized the entire dataset using two heatmaps, represented in **Figure 37**. The hierarchical clustering of the normalized metabolite abundance values revealed an almost perfect segregation of tumor and normal samples (except for 2 samples in adenocarcinoma and 1 sample in squamous cell carcinoma).

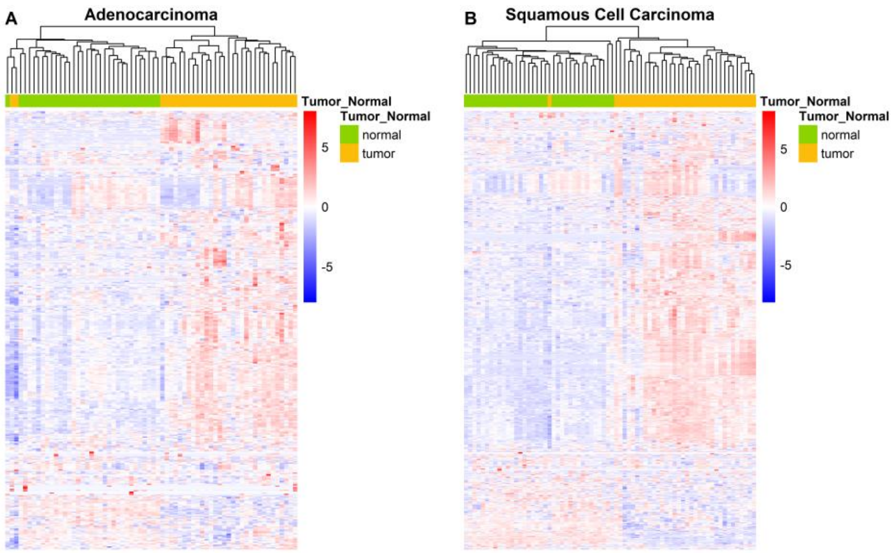

Figure 37. Metabolomic dataset overview and hierarchical clustering results. For each cancer type (A, B), the whole dataset is represented as a heatmap. Cell colors represent the scaled normalized metabolite abundance. The column annotation indicates which samples were generated from tumoral tissue or normal adjacent tissue. The top dendrogram represents the results of the hierarchical clustering of samples, using the average-linkage method with Pearson's correlation as metric.

Next, we performed a differential analysis using a paired T-Test to compare the metabolite abundance between tumor and normal samples. This analysis revealed that a great proportion of metabolites were altered between paired samples for both tumor types, as shown in **Figure 38**. The resulting list of differential metabolites was then manually scrutinized grouping metabolites per metabolic pathway.

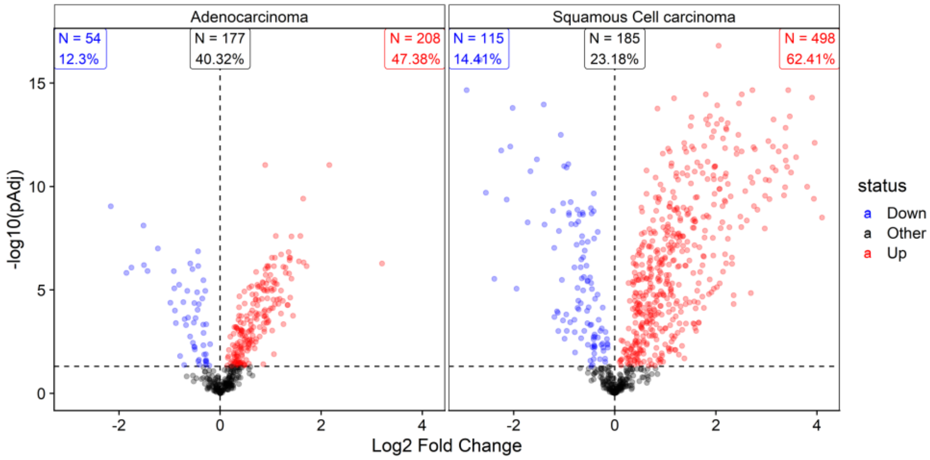

Figure 38. Differential abundance analysis results as volcano plots. The X axis indicate the magnitude of the metabolite abundance change as the log2 fold change, while the Y axis represents the significance of the change as the -log10 transformed adjusted P value. Each point represents a gene, and its color indicates whether a gene surpassed or not a cutoff of adjusted P value < 0.05 and a log2 fold change lower or higher than 0 as blue (Down) or red (Up), respectively. The facets split the volcano plots by cancer type and the labels at the top of each facet indicate the number of metabolites that surpassed the cutoffs.

3.4. Analysis of prior knowledge

In addition to the toolkit previously described, that was mainly employed to analyze new omic datasets, we also automatized a series of tasks to retrieve prior knowledge from biological databases. These tools were used to support and complement experimental research lines and here we show an example of its use and application. We employed one of the tools in a research project focused on exploring the interaction between two proteins: DYRK2 and NOTCH1. The main hypothesis in this research line was that DYRK2 had the capacity of phosphorylating NOTCH1, inducing its degradation. Thus, to elucidate the possible cancer types where this post-translational regulation could be taking place, we processed the data from the Human Protein Atlas (HPA) database. This database contains discrete measurements of the abundance of almost all proteins in different healthy and pathological tissues, based on immunohistochemical staining assays. Hence, by analyzing HPA data, we created a potential list of cancer types where this regulation may occur. We sorted the list by the abundance difference between the two proteins. To calculate this, every staining level was assigned to a number (Not detected: 1, Low: 2, Medium: 3 and High: 4) and multiplied by the number of patients for each tissue and protein. Then, the absolute mean differences were calculated for every tumor tissue. **Figure 39** shows the results we obtained after pre-processing and evaluating HPA data. Overall, we discovered that this type of regulation was more likely to occur in cervical cancer, colorectal cancer, or pancreatic cancer, and less likely in other cancer types like thyroid cancer, prostate cancer or head and neck cancer.

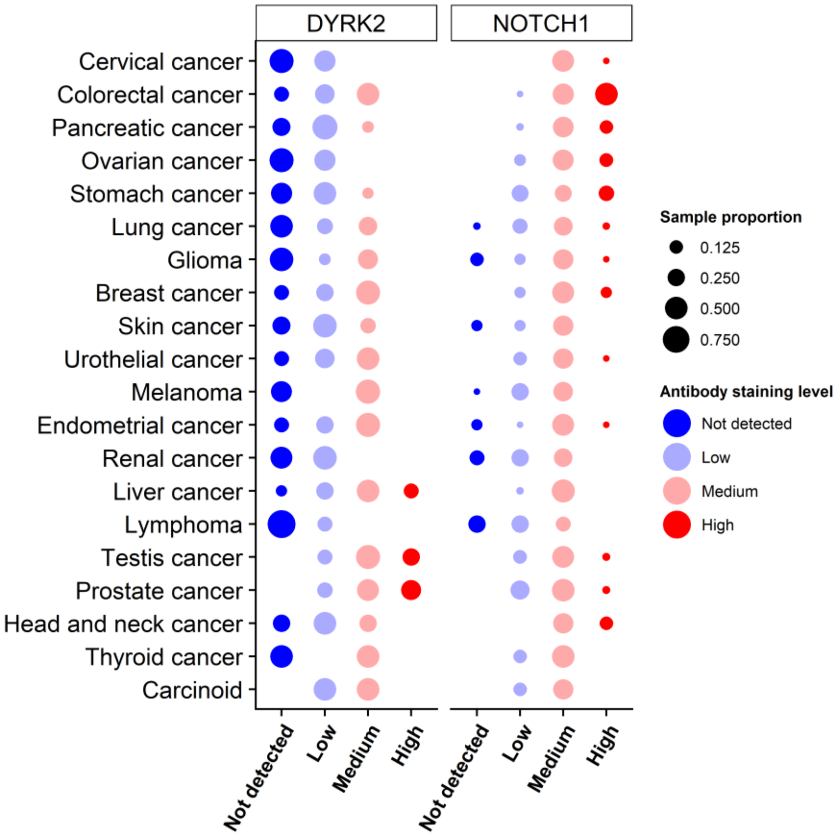

Figure 39. DYRK2 and NOTCH1 protein abundance in tumor tissues obtained from HPA. Column and circle color show the antibody stain level observed in tumor tissues. The point size indicates the number of patients showing a particular expression level relative to the total number of patients. The tumor tissues were sorted based on the abundance score differences between proteins.

Discussion

1. MIGNON

One of the main products of this thesis work is MIGNON (**M**echanistic **I**nte**G**rative A**N**alysis **O**f R**N**A-Seq data), a versatile workflow to integrate RNA-seq genomic and transcriptomic data into mechanistic models of signaling pathways. The pipeline covers the whole analysis from RNA-Seq raw reads to an output that integrates both genomic and transcriptomic data, and which is easy to interpret and link to a given phenotype. To compare MIGNON against other similar tools, we firstly performed a comprehensive review of available workflows for RNA-Seq data processing. We only considered those published from 2015 onwards and able to use raw read files (*.fastq*) as input data. Nine workflows fulfilled these criteria: QuickRNASeq (68), SEPIA (164), Recount2 (165), RNACocktail (166), ARCHS4 (167), GREIN (168), VaP (80), DEWE (169) and RaNA-Seq (70). **Table 11** summarizes their number of citations extracted from Google Scholar (as of April 2020), year of publication and implementation type. In addition, **Table 12** lists the outputs produced by each of them, hence comparing their functionalities.

Workflow	URL	Google scholar citations	Year	Implementation
QuickRNASeq	https://sourceforge.net/projects/quickrnaseq/	26	2016	Shell, Perl, and R scripts
SePIA	http://anduril.org/sepia	25	2016	Anduril workflow
Recount2	https://jhubiostatistics.shinyapps.io/recount/	154	2017	Shiny app and R package
RNACocktail	https://bioinform.github.io/rnacocktail/	71	2017	Python scripts
BioJupies (ARCHS4)	https://amp.pharm.mssm.edu/biojupies/	143	2018	Web service
GREIN	https://shiny.ilincs.org/grein	9	2019	Shiny app and R package
VaP	https://modupeore.github.io/VAP/	1	2019	Perl scripts
DEWE	http://www.sing-group.org/dewe/	3	2019	Java app
RaNa-Seq	https://ranaseq.eu/	0	2019	Web service
MIGNON	https://github.com/babelomics/MIGNON	-	2021	WDL workflow

Table 11. Summary of available workflows for RNA-Seq data analysis.

While a comparison at the level of computational resources does not make much sense, since it will depend on the tools that make up each pipeline, a comparison of their outputs does. Henceforth, the first noticeable aspect is that there are three outputs that most workflows produce, which are the normalized gene expression values, the differential expression results and the transcriptomic-based functional analysis output. In addition, although some of them can carry out genomic variant calling (QuickRNASeq, SEPIA, RNACocktail and VaP), none of them provides a way to integrate the called variants with the gene expression data as MIGNON does. Among the workflows, only SEPIA provides an option for the functional analysis of both omic results, where transcriptomic and genomic data are treated independently. While the real usage level of these workflows is always difficult to estimate, Google Scholar citations can provide an approximate measurement of the relative impacts in terms of scientific document quotations. According to these observations, SEPIA displays a modest 6% of use among the available workflows. Conversely, Recount2 (36%), ARCHS4 (33%) and RNACocktail (16%) together account for 85% of the citations. Among these, only one (ARCHS4) provides functional analysis, by conventional ORA. Thus, a workflow capable of not only extracting transcriptomic and genomic information from RNA-seq reads, but also of integrating them in a mechanistic model of signaling pathways seems to be a good step forward.

Workflow	**NGE**	**DGE**	**TFR**	**GV**	**AGV**	**IGF**
QuickRNASeq	Yes	-	-	Yes	-	-
SePIA	Yes	Yes	Yes	Yes	Yes	-
Recount2	Yes	-	-	-	-	-
RNACocktail	Yes	Yes	-	Yes	-	-
ARCHS4	Yes	Yes	Yes	-	-	-
GREIN	Yes	Yes	-	-	-	-
VaP	-	-	-	Yes	Yes	-
DEWE	Yes	Yes	Yes	-	-	-
RaNa-Seq	Yes	Yes	Yes	-	-	-
MIGNON	Yes	Yes	Yes	Yes	Yes	Yes

Table 12. Comparison of workflow functionalities. Each column represents a different type of output that can be produced by each workflow. NGE = Normalized Gene Expression, DGE = Differential Gene Expression, TFR = Transcriptomic-based functional results, GV = Genomic Variants, AGV = Annotated Genomic Variants, IGF = Integrated Functional Results.

Although it is true that MIGNON provides a novel way to analyze RNA-Seq data, it also makes some assumptions that need to be carefully drawn to highlight the limitations of the proposed methodology. First, MIGNON uses alignments to detect and retrieve genomic variants like Single Nucleotide Polymorphisms (SNP) and Insertions/Deletions (indels). Calling genomic variants from RNA-Seq data is not a new idea, since it appeared together with the first versions of the technique. For example, in

2008, Cloonan et al. described one of the first RNA-Seq protocols to profile stem cell transcriptomes where a dedicated SNP discovery pipeline was used to retrieve genomic information (170). More recent examples include the analysis of bovine pituitary gland (171), or pig hypothalamus and liver (172). Moreover, in 2020, Tang et al. used the genomic information extractable from single-cell RNA-Seq to confirm the mutations detected through single-cell DNA-Seq in skin melanocytes (173). The main biological assumption behind this is that the genomic variants detected in the RNA molecules sequences were generated from the original template DNA during the transcription process.

There are 3 steps where this previous assumption may fail: (i) The RNA polymerase may error during the transcription process, inserting a variant at the transcriptomic level that does not occur at the genomic level, (ii) in RNA-Seq, additional errors can be introduced by the reverse transcriptase that is used to generate the initial set of cDNA molecules from RNA samples, or (iii), in the polymerase chain reaction (PCR) used during the amplification process. Hence, while it has been estimated that the RNA and DNA polymerases employed in (i) and (iii) produce one error for each 300,000 bases synthetized (0.0000033%), the reverse transcriptase used in (ii) has been shown to be notoriously error-prone and expected to make one error every 10,000 to 30,000 bases (0.0001% to 0.000033%) (174,175). It is important to consider this set of biological error rates when interpreting MIGNON's results, which cannot be avoided through computational strategies.

It is also worth highlighting the limitations of the computational side of variant calling which, unlike biological errors, we can control with software modifications. In this sense, the main challenges to overcome include handling splice junctions, variant detection in low-expressed regions, and managing of duplicated reads (176,177). Hence, different engines have been employed to perform the variant calling from aligned RNA-Seq reads. Most of these engines were originally developed for DNA-Seq, but have been adapted to the peculiarities of RNA-Seq. Sahraeian et al. 2017 compared the accuracy of two packages to carry out variant calling from RNA-Seq data using the NIST high-confidence calls as gold standard (166,178). In this work, authors apply the GATK HaplotypeCaller and SAMtools mpileup softwares, in combination with different aligners, showing that GATK performs consistently in terms of precision and sensitivity. Particularly, authors showed that 93 to 97% of GATK true positive calls had genotypes consistent with the NIST HC predictions. Moreover, in the VaP manuscript, authors show how the use of GATK can provide a good sensitivity of 99.7%-99.8% in both heterozygous and homozygous variants while precision reaches 97.6% and 90% in homozygous and heterozygous variants respectively (7). Consequently, we chose to use the GATK workflow adapted for RNA-Seq to perform the variant calling from aligned reads, which has demonstrated robustness in the few benchmark studies that have been conducted.

We also make several assumptions in the functional analysis step proposed to integrate the genomic and transcriptomic data. First, we use VEP to annotate the called variants and assume that those overlapping gene regions and receiving a SIFT score < 0.05 and a PolyPhen2 score > 0.95, trigger deleterious mutations in the corresponding gene products. Both SIFT and PolyPhen2 are algorithms designed to predict the impact of genetic variants over resulting proteins, which showed consistent performances if sufficiently strict cutoffs are used (137,144). Because the predictions are not experimentally validated, we considered alternative approaches to annotate variants during MIGNON's development, like using the curated information from databases like ClinVar (134), or restricting the workflow to specific applications using disease specific information, like those retrieved by COSMIC (135). However, we decided to design MIGNON using these broad-spectrum tools, and to implement a modular design that allows one to adapt the workflow to the specific needs of each project. In addition, while it is true that we cannot experimentally validate all variants annotated by MIGNON, we believe that applying a combination of methodologies and a strict cutoff reduces the proportion of false positives. However, a large benchmark is needed to choose how to score the variants in the proposed approach.

Following variant annotation and filtering, we employ the HiPathia algorithm to calculate signal propagation over cellular signaling circuits. The main assumptions made in this step are derived from the HiPathia model itself and from the proposed *in-silico* knockdown approach. HiPathia implements an algorithm that employs normalized expression measurements as proxies for the resistance of the nodes to transmit the signal through the circuits. Therefore, highly expressed genes allow a higher signal transmission while the less expressed genes produce the opposite effect. Due to an obvious lack of a ground truth or gold standard in the functional analysis of omics data, we cannot determine whether the algorithm used is the most appropriate with respect to a biological reality. Thus, we can only evaluate the methods from what we “expect” to obtain given a biological scenario and our prior knowledge. In this sense, a benchmark carried out in 2018 covering 9 mechanistic methods concluded that HiPathia has the higher specificity and sensitivity in the task of recovering cancer-specific altered signaling circuits (81).

We make use of the mechanistic nature of the HiPathia method to perform an *in-silico* knockdown of genes that trigger loss-of-function mutations, with the aim of identifying why signal propagation is prevented by these events, as demonstrated in our analysis of the 1000 genomes RNA-Seq data. This approach was employed before in two studies: First, Peña et al. demonstrated in 2019 the usefulness of the HiPathia knockdown approach to interpret complex genetic variation using GTEx transcriptomic data in two case studies, one for Fanconi anemia and other for type 2 diabetes (138). Second, Falco et al. 2020 showed how the proposed strategy can be used to predict drug response in glioblastoma at the single cell resolution (179). Overall, although not perfect, we believe that the integrative phase of MIGNON is based upon methodology that is well supported by several studies, demonstrating its usefulness in different biological scenarios.

Finally, from a computational perspective, and being aware that we had to create a pipeline easy to maintain and modify, we implemented MIGNON using a combination of a workflow language and containerized software. Thus, the modular design of the pipeline makes it easy to replace any tool for another one providing it matches the input/output schema used. Users can easily replace tools in the pipeline by making small changes to the MIGNON WDL code, as explained in the documentation.

In conclusion, we believe that MIGNON is a tool that provides an unprecedented functionality for the analysis of RNA-Seq data: the integrated use of genomic and transcriptomic information within a framework based on mechanistic models of cellular signaling circuits. Although its applicability still depends on benchmarks that should be carried out, we believe that this tool can find its full potential in the field of personalized medicine. MIGNON raises the possibility of not only extracting genomic information from an RNA sample, but also of providing mechanistic insights that may translate to the clinic in the form of novel treatments and biomarkers. This tool is intended to widen the bottleneck that currently exists between the generation and analysis of RNA-Seq data in clinical setups, and thus, the next logical step in this project is to perform a proof of concept applying MIGNON in conjunction with the usual clinical practice in a small cohort of patients. Given the heterogeneous nature of the disease, the field of oncology is postulated as a good niche to demonstrate the capabilities of MIGNON, due to its ability to interpret putative driver mutations along with gene expression in the context of signaling activity.

2. Functional modeling of cellular signaling circuits with transcriptomic and proteomic data

Another large block of this thesis work is composed by the results obtained from modeling altered cellular signaling circuits with transcriptomic, proteomic and phosphoproteomic data. To do so, we employed the SkinXCare dataset, a small-scale cellular model that aims to reflect the response of skin fibroblasts to X-ray radiation. Hence, before talking about the applied methodologies, it is important to be aware about the molecular regulation layers that this dataset attempts to capture. SkinXCare data tries to obtain a systematic view of fibroblast's molecular response to two different radiation stimuli, using RNA-Seq for capturing transcriptome level alterations, and SILAC coupled to LC-MS/MS to capture changes that occur at the proteomic and phosphoproteomic levels. Because of the SILAC design of the proteomic experiment, the number of samples analyzed differed between the transcriptomic and the proteomic assays, although the analyzed experimental conditions remained the same (control, acute radiation, and accumulative radiation). Overall, the basal hypothesis in this project is that profiling those molecular levels with state-of-the-art techniques captures most of the cellular signaling events that occur in response to radiation.

For this reason, we compared and integrated the outputs obtained from each molecular level in the different applied strategies. However, even by focusing on levels with a pivotal role in the central dogma of biology, we are still subject to a great complexity that was not directly addressed within our analysis. For instance, some examples of processes that affect cellular signaling and that we did not consider include RNA splicing, transport, modification and degradation, microRNA regulation, or protein folding, localization, ubiquitylation and degradation (180). In our models, we try to find an agreement between the molecular layers either directly or in a different feature space composed by inferred signaling features. In this sense, we are also prisoners of the differences between RNAs, proteins and phosphoproteins, which prevent a perfect correlation between levels, and which is observable in the T value correlations found at the molecular layer. Such differences may appear depending on whether the studied conditions reflect a steady or transitional state, on the availability of ribosomes, or even if we look at different cellular compartments (180,181). This was exemplified in a recent proteomic study from the GTEx consortium, where 201 samples from 32 different tissue types were profiled using tandem mass tag (TMT) based proteomics, quantifying the relative protein levels of over 12,000 genes (182). In this study, the average RNA to protein abundance Spearman correlation was of 0.46, with an interquartile range of 0.24–0.65. In another recent benchmark, performed within the framework of a DREAM challenge with CPTAC data, authors tried to predict protein and phosphoprotein abundances from a combination of DNA, RNA and protein measurements (183). In this study, best scoring teams obtained a Pearson's correlation coefficients of 0.51 and 0.53 when predicting protein abundance, and of 0.42 and 0.33 when predicting phosphoprotein abundance (for Breast and Ovarian cancer, respectively). Both studies are clear examples that, even using leading edge technology, there is a set of latent factors that prevent a perfect agreement between gene expression, protein abundance, and the post-translational modifications that encompass cellular signaling.

Being aware of those limitations, we tried to infer altered signaling circuits in response to radiation using different methods and the available omic profiles. These methods recapitulate the abundance of RNAs, proteins and phosphosites in a basal state (non-radiated fibroblasts) and in two perturbed states (radiated fibroblasts). Whether the perturbed conditions correspond to transitional or steady states, from a signaling perspective, is something that remains unknown. In addition, as done in several CPTAC studies (20,118,184), we also created a phosphoprotein matrix by averaging phosphosite data. Next, we chose three different methods to model altered signaling circuits which, in our view, were good representatives of the different types of approaches that exist for the functional analysis of signaling circuits from omic data and prior knowledge. We chose GSVA (77), HiPathia (78) and CARNIVAL (104) as representatives of FCS, signal propagation and network reconstruction approaches respectively. It is important to realize that, although some benchmarks exist (81,185), and each tool's manuscript includes a comparison with similar methodologies, this work is, to our knowledge, the first attempt to compare functional approaches that differ substantially both in the way prior knowledge and omic data are used.

As explained in the introduction, the functional analysis of omic data can help to reduce its dimensionality and increase its biological interpretability. GSVA achieves both by using a per-sample FCS approach that employs a Kolmogorov-Smirnov (KS) like random walk statistic. Thus, the main benefit of employing GSVA, or in general, any FCS or ORA approach, is that the large original molecular space is transformed into a smaller space that is easier to interpret and associate to a given phenotype. While those traditional approaches can be very useful if the functional categories are adapted to each case of use, they present some weaknesses in the way they use the prior knowledge. Firstly, they treat all the molecular components of the functional categories in the same manner, or in other words, with the same weight. Consequently, molecular players with central or peripheric roles within the signaling networks contribute equally to the resulting score. Secondly, the redundancy between functional categories is not considered, thus benefiting molecules that appear repeatedly through signaling pathways. A study by Cantini et al. in 2018 showed how this functional redundancy can affect the informative value of gene sets within the context of cancer (186). Thirdly, granularity of signaling is not considered, and several studies have shown how, even in the same pathway, two different circuits can lead to opposite biological functions (76,78,102). In other words, the granularity at which the prior knowledge is employed is too shallow and can produce confounding results. Finally, with this being an important aspect of cellular signaling, connections between the molecular components of each signaling networks are not considered in the model. The other two approaches employed to infer signaling circuits consider those issues from different modeling perspectives that also provide a mechanistic framework to trace back the main components of the perturbed networks.

HiPathia employs a network propagation-based methodology to model altered signaling circuits from molecular data and prior knowledge networks. First, the granularity of signaling is considered by decomposing the initial signaling networks into receptor-to-effector circuits. Those circuits aim to represent the way we understand signaling, where the initial reception of a stimulus by nodes on the top of the hierarchy, is propagated through the backbone of the networks to the final effectors, which trigger the biological functions. Then, HiPathia calculates an activity for each node of the network considering the molecular measurements of the features contained on them, summarizing complex nodes and multi-gene nodes in a single value. Finally, HiPathia applies a signal propagation algorithm to explore how an initial theoretical signal is propagated through the circuits, and hence addressing one of the major points not considered by GSVA: the connection between the molecular components of each network. In addition, the signal propagation algorithm also deals with loops, resulting in a more realistic approach to model cellular signaling, where both positive and negative regulatory loops are common elements (94). However, HiPathia also make assumptions that may fail. By mapping the molecular measurements to nodes, HiPathia assumes that such values are good proxies of node signaling activities. However, contrary to previous studies, we found that this is better controlled in our setup through the inclusion of the proteomic and phosphoproteomic levels, which are closer to the final protein activities. Finally, HiPathia does not consider the redundancy between signaling circuits and hence, also suffers from the problem of the same signaling events being evaluated in multiple circuits. The main consequence of this is that HiPathia cannot make “causal” statements, but instead “mechanistic” ones, which allow for a “tracing back” mechanism that leads to the obtained result without declaring it as the sole factor responsible for the observed phenotype.

In the third approach, we adapted the current version of CARNIVAL, and its multi-omic version, called COSMOS (110), to our particular scenario. CARNIVAL uses an approach in which node activity is not taken from molecular measurements directly, but rather from the analysis of upstream regulators performed through FCS. Hence, we inferred TF activities from transcriptomic data, and kinase activities from phosphoproteomic data, using the latter in a different “resolution” than in GSVA and HiPathia (phosphosites instead of phosphoproteins). We included the proteomic layer to complement the regulator activities, but only if no data was found for a particular receptor or intermediate signaling node. In a second step and using a simplified and non-redundant version of the PKN, CARNIVAL solves an optimization problem approached through integer linear programming (ILP) that uses node activities as constraints when reconstructing the network. This implementation, originally developed by Melas et al. in 2015 (108), includes specific parameters to control network size and mismatches between constraints and model predictions, considering the sign of interactions. The main benefit of this approach is that it creates a reduced signaling network that is coherent with the inferred regulatory layers and that, theoretically, is “causal” for the observed phenotype, given that connections are not considered in a redundant manner. However, like the other described methods, there are some problems that CARNIVAL fails to solve. Firstly, protein complexes are simplified into their individual components before the analysis and thus, not included in the model in any manner. Secondly, the ILP algorithm used to solve the optimization problem makes it infeasible to consider feedback loops, failing to model this aspect of cellular signaling. Finally, the inclusion of the network size in the objective function results in networks sometimes too small and parsimonious for the number of signaling events that may, theoretically, occur in response to perturbation. All the discussed features for each method are summarized in **Table 13**.

Feature	**GSVA**	**HiPathia**	**CARNIVAL**
Reduced dimensionality and increased interpretability	Yes	Yes	Yes
Topological information is considered	-	Yes	Yes
Receptor-to-effector circuits	-	Yes	Yes
Complexes are considered	-	Yes	-
Feedback loops are considered	-	Yes	-
Non-redundant evaluation of interactions	-	-	Yes
Upstream regulator analysis	-	-	Yes

Table 13. Feature comparison for methods employed to model signaling circuits from omic data and prior knowledge.

We used limma to compare the measurements between sample groups in the original molecular space and in the functional spaces derived from GSVA and HiPathia (73). We chose this for several reasons: firstly, the linear models implemented in limma offer great flexibility that allowed us to compare samples between different conditions considering the experimental design of each omic assay. Secondly, the empirical bayes approach used to estimate the variance considers not only the row-wise variation, but also global trends, resulting in increased effective degrees of freedom that are very convenient in small sample size scenarios, like the one in the SkinXCare dataset. Thirdly, while limma is the state-of-the-art for the analysis of transcriptomic and proteomic data, previous studies also employed it to assess the difference between the scores estimated by GSVA and HiPathia (77,179). Finally, the choice of a common statistical model allowed us to study how the resulting T value vectors correlate with each other, eliminating the possible variability that could arise from using dedicated statistical models for each assay. This was the same principle that guided us to employ a fixed prior knowledge background for the three methods. In this sense, we employed a reduced version of the HiPathia pathways, created from the nodes and interactions in

KEGG that in our view, reflects most of the canonical cellular signaling processes. This filtering allowed us to remove pathways like “Pancreatic cancer”, “Cocaine addiction” or “Dopaminergic synapse”. It is true that by doing so, we fell into the trap of the “streetlight effect”, and that one of the aspects that remains to be explored is how the results would change as we include more and more information with less support in our case study (187).

With this setup, we quantified the similarity of the molecular changes that occurred in response to radiation using the Pearson’s correlation between moderated T values. Our results indicated that the similarity of changes was dominated by omic level rather than by stimuli, finding a closer correlation between proteins and phosphoproteins and a lower correlation of the transcriptomic level with the other two. As mentioned above, we expected this difference between transcriptomic and proteomic changes, given the wide variety of factors that can explain it (181). While we observed a similar pattern at the HiPathia level, the GSVA scores showed an overall higher agreement between levels. Whether this is caused by the prior knowledge information added in the model, or by the fact that measurements are “averaged” by GSVA before the differential analysis, is something that remains unclear. In addition, another aspect that remains unexplored in our analysis is whether the pathway and circuit scores that we obtained were mainly driven by protein-protein interactions or by genetic regulatory events (188).

We employed a consensus rank approach to integrate the multi-omic information obtained from GSVA and HiPathia because of the significance and magnitude differences that we found between the transcriptomic-based scores and those that were generated from proteomic data. A very similar approach was used by Väremo et al. in 2013 to aggregate the results from different enrichment methods into a single ranking score (189). In our study, the underlying assumption behind this consensus approach was that the molecular components of signaling pathways and circuits more affected by radiation would show pronounced changes at all the quantified levels. In a final step, we compared the solutions obtained from each method. In an ideal setup, we would have a ground truth to score each of the solutions. However, as mentioned in the introduction, a ground truth does not exist in cellular signaling at the moment (46,48,187). There are some phenomena that we may expect more than others, but without a guarantee that these encompass the entire response to radiation.

Thus, we decided to compare the obtained signaling networks instead of scoring them. As shown in the results of the comparison, the HiPathia and GSVA approaches showed an overall higher degree of similarity between them than with the solutions exerted by CARNIVAL. This can be explained by the fact that GSVA and HiPathia estimate perturbed signaling in a pre-delimited space, where the prior knowledge is divided by pathways or receptor-to-effector signaling circuits. In comparison, the CARNIVAL starting point consists of all the nodes and interactions contained in the prior knowledge. In addition, in CARNIVAL, two out of the three levels are not used through direct mapping to nodes, but instead are employed as surrogates of the activity of master regulators (TFs and kinases).

On the biological side, we can review whether the most prominent altered circuits and nodes for each of the models are consistent with previous literature. In this sense, we expect to see signaling circuits mainly related to the DNA damage response caused by X-rays (190). Starting with GSVA results, we found the “Notch signaling pathway” as the top up-regulated pathway in response to the acute dose of radiation. Notch signaling is a process known to modulate the DNA damage response, competing with FOXO3a for ATM binding (191,192). Moreover, in a manuscript from 2019, authors showed that NOTCH signaling promotes the survival of irradiated basal airway stem cells in response to X-ray radiation (193). The most down-regulated GSVA feature in both stimuli was “Hippo signaling pathway”. Phosphorylation of YAP1, one of the main players in this pathway, is considered a critical step in the activation of proapoptotic genes in response to DNA damage (194,195). Hence, its down-regulation may be an indicative of the fibroblasts’ resistance to trigger pro-apoptotic mechanisms.

Within the HiPathia top altered circuits, we found a circuit of the “AMPK signaling pathway” leading to the activation of PPARGC1A as the top up-regulated circuit in response to the acute dose of radiation. Several studies has demonstrated that this effector, which is a PPARG cofactor, is associated with protective role against DNA damage and telomere malfunction in different studies (196,197). In contrast, we found a sub-circuit of the “MAPK signaling pathway” leading to the MAP3K2 kinase as the top down-regulated feature in the same stimulus. While we could not find a direct association of this effector with DNA damage response, we did find a study that described its ability to regulate the Hippo pathway, one of the results previously obtained in the GSVA approach (198). We also discovered different up-regulated circuits in response to the accumulative radiation dose, which reflected a higher "stimulus-specificity" of the method in this direction of change. We found a circuit of the "Jak-STAT signaling pathway" leading to CDKN1A as the top up-regulated feature. This effector, also known as p21, is a cell-cycle inhibitor and has been shown to regulate cell cycle, apoptosis and gene transcription in response to DNA damage (199). Thus, this up-regulation may be an indicator of cell cycle arrest that becomes more prominent in response to the accumulation of radiation doses.

In the final approach, carried out with CARNIVAL, the TF analysis revealed a conserved signature of up-regulated TFs, with STAT2 being the top up-regulated feature in response to both doses of radiation. In a study from 2013, STAT2 was studied as one of the effectors of IFNβ mediated signaling to promote resistance to DNA damage, which could explain the obtained result (200). The down-regulated TFs in response to radiation varied between stimuli. While we could not find a clear connection between the studied phenotype and the top down-regulated TF in response to the acute radiation dose (ZNF263), we found a clear explanation for those that reduced their activity in response to the accumulative dose of radiation. In particular, the top down-regulated TFs were composed of E2F family members, which are known to control cell cycle progression to help DNA repair (201). This down-regulation could be an indicator of the mechanisms used by the fibroblasts to stop the cell-cycle in an attempt to repair the damage caused by radiation. Regarding kinase analysis, we found more consistent results between stimuli. Markedly, the top up-regulated kinase was ATM in both radiation doses, probably the best known player of the response to DNA damage (202). We found CSNK2A2 as the top down-regulated kinase in response to both stimuli. This kinase has been associated with a wide variety of biological processes, and its inhibition has been linked to reduced survival in cancer cells in response to DNA damage therapies (203). In the final step, we used the regulators' activity changes coupled to the proteomic data to derive a single coherent signaling network solving an integer linear programming (ILP) optimization problem. This resulted in two networks, one per radiation stimuli, that contained not only the regulators here discussed, but also other proteins related with the results of the other two methods, like the YAP1 protein.

In conclusion, while the FCS approaches are a popular choice to reduce omic data dimensionality and increase its interpretability, they lack a mechanistic component to trace back and identify the most important molecular players in the altered pathways. In this sense, both HiPathia and CARNIVAL use the topological information of signaling networks to model altered circuits, thus considering the interactions between genes or proteins. The analysis of resulting networks with topological measurements, like the PageRank node betweenness score (162), provides a powerful framework to select molecular features that can drive the modeled signaling response. Hence, possible future directions for this work could include perturbing those "central" nodes of the cellular signaling that occur in response to radiation, to explore whether the different modeling strategies allow us to really modulate the response of fibroblasts to radiation.

3. A computational toolkit for biomedical research

In contrast to the strategies developed to model signaling circuits from multi-omics data, the toolkit that we implemented and applied in different biomedical research contexts was not focused on creating new methods for the analysis of omics data. Instead, we summarized different mathematical and statistical approaches widely employed in the analysis of omics data, placing the focus on its application rather than on its methodological development. By doing so, we created a single unified toolkit intended to increase the reproducibility of omics data analysis, given the lack of consensus amongst scientist about which methods to apply (204). We discovered how even the simplest methods, from a mathematical perspective, are extremely powerful when applied in the appropriated contexts. An example of this is the over-representation analysis that was used in the SWATH prostate cancer proteomic study as a quality control step. As explained previously, while it is true that proteomic technologies are currently developing by leaps and bounds, they are still behind genomic and transcriptomic technologies in terms of total proteome coverage. Thus, in the SWATH proteomic study, only 1863 proteins were quantified, comprising around a ~10% of all the proteins annotated in Swiss-Prot (205). To certify that the quantified proteome was representative of the studied cell lines, we used an over-representation analysis against a background formed by the most expressed genes across the NCI-60 cancer cell lines (206). The method used in the core of this approach, which is the Fisher's exact test, was published in the fifth version of "Statistical Methods for Research Workers" in 1934, and is a test for independence as opposed to association in 2x2 contingency tables (158). Although more complex approaches exist to perform such analyses, we believe that our method was the most appropriate in this context, revealing that the two cell lines under study were the top enriched features in this analysis.

The same principle applies to the visualization, clustering, differential analysis and overlap exploration methodologies that we used across the different biomedical research projects where omics data were analyzed. The methods included in the toolkit helped to address questions such as:

- Which proteins are responsible for the effect of the drug?
- Does the treatment produce a systematic reversal of the pathological phenotype in the target tissue?
- Is our data grouped according to their source in an unsupervised manner?

In addition to the omic data generated in each project and analyzed with our toolkit, we showed how the usage of prior knowledge stored in biological databases can be used to create, reinforce, or refute each project's specific hypotheses. Overall, we believe that the software tool developed can address common tasks in the analysis of omics data within biomedical research contexts, and help to the unification of computational analyses in this field.

4. Towards new data, tools, and insights

In this thesis work, we summarized, developed, and applied computational approaches to perform the functional analysis and modeling of cellular signaling circuits from different types of omic data. While the biokit employed state of the art techniques to answer questions in different biomedical research contexts, MIGNON and the proposed strategies to model signaling from multi-omic data aimed to obtain a mechanistic perspective of the observed phenotypes. In this sense, as mentioned previously, the output of our methods could not be properly scored using a ground truth, a problem practically inherent to biology, where the latent space of unknown phenomena compromises all the extracted functional conclusions.

However, new multi-omic technologies hold the promise of resolving the molecular state of samples with spatial and temporal resolution, and at the single-cell level (207,208). The biomedical scientific community is moving towards an exciting era, where for the first time we will be able to answer an almost infinite number of questions with the data generated by these new techniques. However, to be able to answer these questions, we will need to adapt to a more high-throughput focused mindset, where computational approaches will be a cornerstone of research in almost any area of knowledge. The work of this thesis is our contribution to this new era, where the proposed methodologies are intended to open the door to new discoveries and help researchers to break down the cause or mechanism of pathological phenotypes through the use of bioinformatics. With these techniques, our intention is to extract more accurate insights from available data, with the final aim of improving patients' lives through translational biomedical research.

Conclusions

1. MIGNON is the first available workflow able to extract and integrate the genomic and transcriptomic information from RNA-Seq data using a mechanistic model of signaling pathways.

2. MIGNON makes use of state-of-the-art methods to perform the initial processing of raw reads and uses an *in-silico* knockdown strategy to stop the signal propagation in genes affected by loss-of-function mutations.

3. Although further experimental validation of the proposed approach is needed, MIGNON has an enormous potential of application in personalized medicine, especially in the analysis of cancer transcriptomes, given its ability to interpret putative driver mutations along with gene expression in the context of signaling activity.

5. In the framework of the SkinXCare dataset, there is a poor agreement between the changes occurring at transcriptomic, proteomic and phosphoproteomic level. In the molecular space, the across-level correlations are dominated by assay rather than by stimuli, and the correlations between proteomic levels are higher than with the transcriptomic changes.

6. In the functional space provided by the GSVA analysis, the correlation between levels is still dominated by assay except for the phosphoproteomic level. On the other hand, for the HiPathia functional space, the same pattern observed at the molecular level is conserved.

7. The lack of a ground truth in the field of functional analysis makes it impossible to score the inferred signaling networks. However, its comparison revealed a better agreement between GSVA and HiPathia than between CARNIVAL and the other two approaches. This can be caused by the fact that GSVA and HiPathia estimate perturbed signaling in a pre-delimited space.

8. A toolkit for the basic computational analysis of omic data in biomedical research was developed and applied in different contexts, showing its potential to create, reinforce or refute each project’s specific hypotheses when combined with experimental validations.

Bibliography

1. Rogers K. Scientific modeling. Encyclopedia Britannica [Internet]. 2012 May 21 [cited 2021 Apr 6]; Available from: https://www.britannica.com/science/scientific-modeling

2. Epstein JM. Why Model? Journal of Artificial Societies and Social Simulation. 2008;11(4):12.

3. Box GEP. Science and Statistics. J Am Stat Assoc. 1976 Dec;71(356):791–799.

4. Valentine JW, May CL. Hierarchies in biology and paleontology. Paleobiology. 1996;22(1):23–33.

5. Derbal Y. On modeling of living organisms using hierarchical coarse-graining abstractions of knowledge. J Biol Syst. 2013 Mar;21(01):1350008.

6. Toni T, Tidor B. Combined model of intrinsic and extrinsic variability for computational network design with application to synthetic biology. PLoS Comput Biol. 2013 Mar 28;9(3):e1002960.

7. Fisher RA. The Design of Experiments. Edinburgh: Oliver and Boyd; 1935.

8. Prahallad A, Bernards R. Opportunities and challenges provided by crosstalk between signalling pathways in cancer. Oncogene. 2016 Mar 3;35(9):1073–1079.

9. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995 Oct 20;270(5235):467–470.

10. Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet. 2001;2:343–372.

11. Horgan RP, Kenny LC. Omic‘ technologies: genomics, transcriptomics, proteomics and metabolomics. The Obstetrician & Gynaecologist. 2011 Jul;13(3):189–195.

12. Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinformatics. 2018 Mar 1;19(2):286–302.

13. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017 May 5;18(1):83.

14. Wang KC, Chang HY. Epigenomics: technologies and applications. Circ Res. 2018 Apr 27;122(9):1191–1199.

15. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017 May 18;13(5):e1005457.

16. Mann M, Jensen ON. Proteomic analysis of post-translational modifications. Nat Biotechnol. 2003 Mar;21(3):255–261.

17. Cox J, Mann M. Is proteomics the new genomics? Cell. 2007 Aug 10;130(3):395–398.

18. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. Nat Rev Mol Cell Biol. 2016 Mar 16;17(7):451–459.

19. Tarazona S, Balzano-Nogueira L, Gómez-Cabrero D, Schmidt A, Imhof A, Hankemeier T, et al. Harmonization of quality metrics and power calculation in multi-omic studies. Nat Commun. 2020 Jun 18;11(1):3092.

20. Wang L-B, Karpova A, Gritsenko MA, Kyle JE, Cao S, Li Y, et al. Proteogenomic and metabolomic characterization of human glioblastoma. Cancer Cell. 2021 Apr 12;39(4):509–528.e20.

21. Westergaard D, Stærfeldt H-H, Tønsberg C, Jensen LJ, Brunak S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. PLoS Comput Biol. 2018 Feb 15;14(2):e1005962.

22. Reshetova P, Smilde AK, van Kampen AHC, Westerhuis JA. Use of prior knowledge for the analysis of high-throughput transcriptomics and metabolomics data. BMC Syst Biol. 2014 Mar 13;8 Suppl 2:S2.

23. Imker HJ. 25 years of molecular biology databases: A study of proliferation, impact, and maintenance. Front Res Metr Anal. 2018 May 29;3.

24. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol. 1977 May 25;112(3):535–542.

25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–410.

26. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database: its relevance to human molecular medical research. J Mol Med. 1997 May;75(5):312–316.

27. Maglott DR, Katz KS, Sicotte H, Pruitt KD. Ncbi's locuslink and refseq. Nucleic Acids Res. 2000 Jan 1;28(1):126–128.

28. Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R. UniProt archive. Bioinformatics. 2004 Nov 22;20(17):3236–3237.

29. Kanehisa M. Toward Pathway Engineering: A New Database of Genetic and Molecular Pathways. Science & Technology Japan. 1996;59:34–38.

30. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25–29.

31. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: the database of interacting proteins. Nucleic Acids Res. 2000 Jan 1;28(1):289–291.

32. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021 Jan 8;49(D1):D545–D551.

33. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res. 2013 Jan;41(Database issue):D793–800.

34. Türei D, Valdeolivas A, Gul L, Palacio-Escat N, Klein M, Ivanova O, et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Mol Syst Biol. 2021;17(3):e9923.

35. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods. 2016 Nov 29;13(12):966–967.

36. Zou D, Ma L, Yu J, Zhang Z. Biological databases for human research. Genomics Proteomics Bioinformatics. 2015 Feb 21;13(1):55–63.

37. Bouadjenek MR, Verspoor K, Zobel J. Literature consistency of bioinformatics sequence databases is effective for assessing record quality. Database (Oxford). 2017 Jan 1;2017(1).

38. Bursteinas B, Britto R, Bely B, Auchincloss A, Rivoire C, Redaschi N, et al. Minimizing proteome redundancy in the UniProt Knowledgebase. Database (Oxford). 2016 Dec 26;2016.

39. Chen Q, Britto R, Erill I, Jeffery CJ, Liberzon A, Magrane M, et al. Quality matters: biocuration experts on the impact of duplication and other data quality issues in biological databases. Genomics Proteomics Bioinformatics. 2020 Apr;18(2):91–103.

40. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2020 Jan 8;48(D1):D498–D503.

41. Linding R, Jensen LJ, Ostheimer GJ, van Vugt MATM, Jørgensen C, Miron IM, et al. Systematic discovery of in vivo phosphorylation networks. Cell. 2007 Jun 29;129(7):1415–1426.

42. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015 Jan 23;347(6220):1260419.

43. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019 Jan 8;47(D1):D607–D613.

44. Battaglia M, Atkinson MA. The streetlight effect in type 1 diabetes. Diabetes. 2015 Apr;64(4):1081–1090.

45. Evans BJ. The streetlight effect: regulating genomics where the light is. J Law Med Ethics. 2020;48(1):105–118.

46. Haynes WA, Tomczak A, Khatri P. Gene annotation bias impedes biomedical research. Sci Rep. 2018 Jan 22;8(1):1362.

47. Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. Nat Methods. 2020;17(2):159–162.

48. Needham EJ, Parker BL, Burykin T, James DE, Humphrey SJ. Illuminating the dark phosphoproteome. Sci Signal. 2019 Jan 22;12(565).

49. Ochoa D, Jarnuczak AF, Viéitez C, Gehre M, Soucheray M, Mateus A, et al. The functional landscape of the human phosphoproteome. Nat Biotechnol. 2020;38(3):365–373.

50. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012 Feb 23;8(2):e1002375.

51. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 2019;20(5):273–282.

52. Care MA, Westhead DR, Tooze RM. Parsimonious Gene Correlation Network Analysis (PGCNA): a tool to define modular gene co-expression for refined molecular stratification in cancer. NPJ Syst Biol Appl. 2019 Apr 11;5:13.

53. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015 Dec 23;1(6):417–425.

54. Perez-Riverol Y, Zorin A, Dass G, Vu M-T, Xu P, Glont M, et al. Quantifying the impact of public omics data. Nat Commun. 2019 Aug 5;10(1):3512.

55. Perez-Riverol Y, Bai M, da Veiga Leprevost F, Squizzato S, Park YM, Haug K, et al. Discovering and linking public omics data sets using the Omics Discovery Index. Nat Biotechnol. 2017 May 9;35(5):406–409.

56. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57–63.

57. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019 Jul 24;20(11):631–656.

58. Van den Berge K, Hembach KM, Soneson C, Tiberi S, Clement L, Love MI, et al. RNA sequencing data: hitchhiker’s guide to expression analysis. Annu Rev Biomed Data Sci. 2019 Jul 22;2(1).

59. Zhao S, Zhang Y, Gordon W, Quan J, Xi H, Du S, et al. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. BMC Genomics. 2015 Sep 3;16:675.

60. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–2120.

61. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018 Sep 1;34(17):i884–i890.

62. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013 Apr 25;14(4):R36.

63. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.

64. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015 Apr;12(4):357–360.

65. Nellore A, Collado-Torres L, Jaffe AE, Alquicira-Hernández J, Wilks C, Pritt J, et al. Rail-RNA: scalable analysis of RNA-seq splicing and coverage. Bioinformatics. 2017 Dec 15;33(24):4033–4040.

66. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017 Apr;14(4):417–419.

67. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016 Apr 4;34(5):525–527.

68. Zhao S, Xi L, Quan J, Xi H, Zhang Y, von Schack D, et al. QuickRNASeq lifts large-scale RNA-seq data analyses to the next level of automation and interactive visualization. BMC Genomics. 2016 Jan 8;17:39.

69. Torre D, Lachmann A, Ma’ayan A. BioJupies: Automated Generation of Interactive Notebooks for RNA-Seq Data Analysis in the Cloud. Cell Syst. 2018 Nov 28;7(5):556–561.e3.

70. Prieto C, Barrios D. RaNA-Seq: Interactive RNA-Seq analysis from FASTQ files to functional analysis. Bioinformatics. 2019 Nov 15;

71. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–140.

72. Love MI, Huber W, Anders S. Moderated estimation of fold change and' ' dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

73. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47.

74. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics. 2004 Mar 1;20(4):578–580.

75. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005 Oct 25;102(43):15545–15550.

76. Li X, Shen L, Shang X, Liu W. Subpathway Analysis based on Signaling-Pathway Impact Analysis of Signaling Pathway. PLoS One. 2015 Jul 24;10(7):e0132813.

77. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013 Jan 16;14:7.

78. Hidalgo MR, Cubuk C, Amadoz A, Salavert F, Carbonell-Caballero J, Dopazo J. High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget. 2017 Jan 17;8(3):5160–5178.

79. Brouard J-S, Schenkel F, Marete A, Bissonnette N. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. J Anim Sci Biotechnol. 2019 Jun 21;10:44.

80. Adetunji MO, Lamont SJ, Abasht B, Schmidt CJ. Variant analysis pipeline for accurate detection of genomic variants from transcriptome sequencing data. PLoS One. 2019 Sep 23;14(9):e0216838.

81. Amadoz A, Hidalgo MR, Çubuk C, Carbonell-Caballero J, Dopazo J. A comparison of mechanistic signaling pathway activity analysis methods. Brief Bioinformatics. 2019 Sep 27;20(5):1655–1668.

82. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular Biology of the Cell. 4th ed. 2002.

83. Janes KA, Chandran PL, Ford RM, Lazzara MJ, Papin JA, Peirce SM, et al. An engineering design approach to systems biology. Integr Biol (Camb). 2017 Jul 17;9(7):574–583.

84. Deribe YL, Pawson T, Dikic I. Post-translational modifications in signal integration. Nat Struct Mol Biol. 2010 Jun;17(6):666–672.

85. Ardito F, Giuliani M, Perrone D, Troiano G, Lo Muzio L. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). Int J Mol Med. 2017 Aug;40(2):271–280.

86. Kholodenko BN. Cell-signalling dynamics in time and space. Nat Rev Mol Cell Biol. 2006 Mar;7(3):165–176.

87. Cheng F, Kovács IA, Barabási A-L. Network-based prediction of drug combinations. Nat Commun. 2019 Mar 13;10(1):1197.

88. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004 Feb;5(2):101–113.

89. Plotnikov A, Zehorai E, Procaccia S, Seger R. The MAPK cascades: signaling components, nuclear roles and mechanisms of nuclear translocation. Biochim Biophys Acta. 2011 Sep;1813(9):1619–1633.

90. Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, et al. WikiPathways: connecting communities. Nucleic Acids Res. 2021 Jan 8;49(D1):D613–D621.

91. Licata L, Lo Surdo P, Iannuccelli M, Palma A, Micarelli E, Perfetto L, et al. SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update. Nucleic Acids Res. 2020 Jan 8;48(D1):D504–D510.

92. Kuperstein I, Bonnet E, Nguyen HA, Cohen D, Viara E, Grieco L, et al. Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. Oncogenesis. 2015 Jul 20;4:e160.

93. Hill SM, Heiser LM, Cokelaer T, Unger M, Nesser NK, Carlin DE, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. Nat Methods. 2016 Apr;13(4):310–318.

94. Brandman O, Meyer T. Feedback loops shape cellular signals in space and time. Science. 2008 Oct 17;322(5900):390–395.

95. Hlavacek WS, Faeder JR, Blinov ML, Perelson AS, Goldstein B. The complexity of complexes in signal transduction. Biotechnol Bioeng. 2003 Dec 30;84(7):783–794.

96. Nishi H, Demir E, Panchenko AR. Crosstalk between signaling pathways provided by single and multiple protein phosphorylation sites. J Mol Biol. 2015 Jan 30;427(2):511–520.

97. Morris MK, Saez-Rodriguez J, Sorger PK, Lauffenburger DA. Logic-based models for the analysis of cell signaling networks. Biochemistry. 2010 Apr 20;49(15):3216–3224.

98. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim J-S, et al. A novel signaling pathway impact analysis. Bioinformatics. 2009 Jan 1;25(1):75–82.

99. Sebastián-León P, Carbonell J, Salavert F, Sanchez R, Medina I, Dopazo J. Inferring the functional effect of gene expression changes in signaling pathways. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W213–7.

100. Krämer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2014 Feb 15;30(4):523–530.

101. Terfve CDA, Wilkes EH, Casado P, Cutillas PR, Saez-Rodriguez J. Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. Nat Commun. 2015 Sep 10;6:8033.

102. Koumakis L, Kanterakis A, Kartsaki E, Chatzimina M, Zervakis M, Tsiknakis M, et al. MinePath: Mining for Phenotype Differential Sub-paths in Molecular Pathways. PLoS Comput Biol. 2016 Nov 10;12(11):e1005187.

103. Bradley G, Barrett SJ. CausalR: extracting mechanistic sense from genome scale data. Bioinformatics. 2017 Nov 15;33(22):3670–3672.

104. Liu A, Trairatphisan P, Gjerga E, Didangelos A, Barratt J, Saez-Rodriguez J. From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. NPJ Syst Biol Appl. 2019 Nov 11;5:40.

105. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019 Jul 24;29(8):1363–1375.

106. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat Genet. 2016 Jun 20;48(8):838–847.

107. Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S, et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun. 2018 Jan 2;9(1):20.

108. Melas IN, Sakellaropoulos T, Iorio F, Alexopoulos LG, Loh W-Y, Lauffenburger DA, et al. Identification of drug-specific pathways based on gene expression data: application to drug induced lung injury. Integr Biol (Camb). 2015 Aug;7(8):904–920.

109. Gjerga E, Dugourd A, Tobalina L, Sousa A, Saez-Rodriguez J. PHONEMeS: Efficient Modeling of Signaling Networks Derived from Large-Scale Mass Spectrometry Data. J Proteome Res. 2021 Mar 8;

110. Dugourd A, Kuppe C, Sciacovelli M, Gjerga E, Gabor A, Emdal KB, et al. Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. Mol Syst Biol. 17(1):e9730.

111. Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001 May 3;411(6833):41–42.

112. Gao S, Wang X. TAPPA: topological analysis of pathway phenotype association. Bioinformatics. 2007 Nov 15;23(22):3100–3102.

113. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74.

114. The United Kingdom’s 10K genome project. Nat Methods. 2015 Nov;12(11):1010–1010.

115. Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. Biopreserv Biobank. 2015 Oct;13(5):307–308.

116. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020 Sep 11;369(6509):1318–1330.

117. van ’t Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002 Jan 31;415(6871):530–536.

118. Krug K, Jaehnig EJ, Satpathy S, Blumenberg L, Karpova A, Anurag M, et al. Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. Cell. 2020 Nov 25;183(5):1436–1456.e31.

119. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012 Mar 28;483(7391):603–607.

120. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A landscape of pharmacogenomic interactions in cancer. Cell. 2016 Jul 28;166(3):740–754.

121. Liu H, Liu W, Liao Y, Cheng L, Liu Q, Ren X, et al. CADgene: a comprehensive database for coronary artery disease genes. Nucleic Acids Res. 2011 Jan;39(Database issue):D991–6.

122. Bai Z, Han G, Xie B, Wang J, Song F, Peng X, et al. Alzbase: an integrative database for gene dysregulation in alzheimer’s disease. Mol Neurobiol. 2016 Jan;53(1):310–319.

123. Kolker E, Özdemir V, Martens L, Hancock W, Anderson G, Anderson N, et al. Toward more transparent and reproducible omics studies through a common metadata checklist and data publications. OMICS. 2014 Jan;18(1):10–14.

124. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013 Jan;41(Database issue):D991–5.

125. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, et al. ArrayExpress update - from bulk to single-cell expression data. Nucleic Acids Res. 2019 Jan 8;47(D1):D711–D715.

126. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 2019 Jan 8;47(D1):D442–D450.

127. Caldas J, Vinga S. Global meta-analysis of transcriptomics studies. PLoS One. 2014 Feb 26;9(2):e89318.

128. Fey D, Halasz M, Dreidax D, Kennedy SP, Hastings JF, Rauch N, et al. Signaling pathway models as biomarkers: Patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. Sci Signal. 2015 Dec 22;8(408):ra130.

129. Eduati F, Jaaks P, Wappler J, Cramer T, Merten CA, Garnett MJ, et al. Patient-specific logic models of signaling pathways from screenings on cancer biopsies to prioritize personalized combination therapies. Mol Syst Biol. 2020;16(2):e8664.

130. Loucera C, Esteban-Medina M, Rian K, Falco MM, Dopazo J, Peña-Chilet M. Drug repurposing for COVID-19 using machine learning and mechanistic models of signal transduction circuits related to SARS-CoV-2 infection. Signal Transduct Target Ther. 2020 Dec 11;5(1):290.

131. Ginsburg GS, Willard HF. Genomic and personalized medicine: foundations and applications. Transl Res. 2009 Dec;154(6):277–287.

132. Eberhard DA, Johnson BE, Amler LC, Goddard AD, Heldens SL, Herbst RS, et al. Mutations in the epidermal growth factor receptor and in KRAS are predictive and prognostic indicators in patients with non-small-cell lung cancer treated with chemotherapy alone and in combination with erlotinib. J Clin Oncol. 2005 Sep 1;23(25):5900–5909.

133. Tung NM, Garber JE. BRCA1/2 testing: therapeutic implications for breast cancer management. Br J Cancer. 2018 Jun 5;119(2):141–152.

134. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014 Jan;42(Database issue):D980–5.

135. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2019 Jan 8;47(D1):D941–D947.

136. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019 Jan 8;47(D1):D886–D894.

137. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812–3814.

138. Peña-Chilet M, Esteban-Medina M, Falco MM, Rian K, Hidalgo MR, Loucera C, et al. Using mechanistic models for the clinical interpretation of complex genomic variation. Sci Rep. 2019 Dec 12;9(1):18937.

139. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. [version 2; peer review: 4 approved]. F1000Res. 2018 Aug 24;7:1338.

140. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357–359.

141. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019 Aug 2;37(8):907–915.

142. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–1303.

143. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol. 2016 Jun 6;17(1):122.

144. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–249.

145. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014 Apr 1;30(7):923–930.

146. Chen Y, Lun ATL, Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. [version 2; peer review: 5 approved]. F1000Res. 2016 Jun 20;5:1438.

147. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010 Mar 2;11(3):R25.

148. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014 Feb 3;15(2):R29.

149. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995 Jan;57(1):289–300.

150. Rappsilber J, Mann M, Ishihama Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc. 2007;2(8):1896–1906.

151. Chiva C, Olivella R, Borràs E, Espadas G, Pastor O, Solé A, et al. QCloud: A cloud-based quality control system for mass spectrometry-based proteomics laboratories. PLoS One. 2018 Jan 11;13(1):e0189209.

152. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008 Dec;26(12):1367–1372.

153. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011 Apr 1;10(4):1794–1805.

154. Liu M, Dongre A. Proper imputation of missing values in proteomics datasets for differential expression analysis. Brief Bioinformatics. 2020 Jun 10;

155. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. 2002;18 Suppl 1:S96–104.

156. Välikangas T, Suomi T, Elo LL. A systematic evaluation of normalization methods in quantitative label-free proteomics. Brief Bioinformatics. 2018 Jan 1;19(1):1–11.

157. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013 Apr 15;14:128.

158. Fisher RA. Statistical Methods for Research Workers. Oliver and Boyd, editor. 1934.

159. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012 May;16(5):284–287.

160. Sergushichev A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. BioRxiv. 2016 Jun 20;

161. Dijkstra EW. A note on two problems in connexion with graphs. Numer Math. 1959 Dec;1(1):269–271.

162. Page L, Brin S, Motwani R, Winograd T. The PageRank Citation Ranking Bringing Order to the Web. Stanford InfoLab; 1999 Nov. Report No.: 1999-66.

163. Lappalainen T, Sammeth M, Friedländer MR, t Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013 Sep 26;501(7468):506–511.

164. Icay K, Chen P, Cervera A, Rantanen V, Lehtonen R, Hautaniemi S. SePIA: RNA and small RNA sequence processing, integration, and analysis. BioData Min. 2016 May 20;9:20.

165. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al. Reproducible RNA-seq analysis using recount2. Nat Biotechnol. 2017 Apr 11;35(4):319–321.

166. Sahraeian SME, Mohiyuddin M, Sebra R, Tilgner H, Afshar PT, Au KF, et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nat Commun. 2017 Jul 5;8(1):59.

167. Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, et al. Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun. 2018 Apr 10;9(1):1366.

168. Mahi NA, Najafabadi MF, Pilarczyk M, Kouril M, Medvedovic M. GREIN: An Interactive Web Platform for Re-analyzing GEO RNA-seq Data. Sci Rep. 2019 May 20;9(1):7580.

169. López-Fernández H, Blanco-Míguez A, Fdez-Riverola F, Sánchez B, Lourenço A. DEWE: A novel tool for executing differential expression RNA-Seq workflows in biomedical research. Comput Biol Med. 2019 Feb 27;107:197–205.

170. Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods. 2008 Jul;5(7):613–619.

171. Pareek CS, Smoczyński R, Kadarmideen HN, Dziuba P, Błaszczyk P, Sikora M, et al. Single Nucleotide Polymorphism Discovery in Bovine Pituitary Gland Using RNA-Seq Technology. PLoS One. 2016 Sep 8;11(9):e0161370.

172. Martínez-Montes AM, Fernández A, Pérez-Montarelo D, Alves E, Benítez RM, Nuñez Y, et al. Using RNA-Seq SNP data to reveal potential causal mutations related to pig production traits and RNA editing. Anim Genet. 2017 Apr;48(2):151–165.

173. Tang J, Fewings E, Chang D, Zeng H, Liu S, Jorapur A, et al. The genomic landscapes of individual melanocytes from human skin. Nature. 2020 Oct 7;586(7830):600–605.

174. Gout J-F, Li W, Fritsch C, Li A, Haroon S, Singh L, et al. The landscape of transcription errors in eukaryotic cells. Sci Adv. 2017 Oct 20;3(10):e1701484.

175. Potapov V, Ong JL. Examining sources of error in PCR by single-molecule sequencing. PLoS One. 2017 Jan 6;12(1):e0169774.

176. Quinn EM, Cormican P, Kenny EM, Hill M, Anney R, Gill M, et al. Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. PLoS One. 2013 Mar 26;8(3):e58815.

177. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. Am J Hum Genet. 2013 Oct 3;93(4):641–651.

178. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data. 2016 Jun 7;3:160025.

179. Falco MM, Peña-Chilet M, Loucera C, Hidalgo MR, Dopazo J. Mechanistic models of signaling pathways deconvolute the glioblastoma single-cell functional landscape. NAR Cancer. 2020 Jun 1;2(2).

180. Buccitelli C, Selbach M. mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet. 2020 Jul 24;21(10):630–644.

181. Liu Y, Beyer A, Aebersold R. On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell. 2016 Apr 21;165(3):535–550.

182. Jiang L, Wang M, Lin S, Jian R, Li X, Chan J, et al. A quantitative proteome map of the human body. Cell. 2020 Oct 1;183(1):269–283.e19.

183. Yang M, Petralia F, Li Z, Li H, Ma W, Song X, et al. Community Assessment of the Predictability of Cancer Protein and Phosphoprotein Levels from Genomics and Transcriptomics. Cell Syst. 2020 Aug 26;11(2):186–195.e9.

184. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016 Jun 2;534(7605):55–62.

185. Nguyen T-M, Shafi A, Nguyen T, Draghici S. Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biol. 2019 Oct 9;20(1):203.

186. Cantini L, Calzone L, Martignetti L, Rydenfelt M, Blüthgen N, Barillot E, et al. Classification of gene signatures for their information value and functional redundancy. NPJ Syst Biol Appl. 2018;4:2.

187. Edwards AM, Isserlin R, Bader GD, Frye SV, Willson TM, Yu FH. Too many roads not taken. Nature. 2011 Feb 10;470(7333):163–165.

188. Szalai B, Saez-Rodriguez J. Why do pathway methods work better than they should? FEBS Lett. 2020 Dec 14;594(24):4189–4200.

189. Väremo L, Nielsen J, Nookaew I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res. 2013 Apr;41(8):4378–4391.

190. Parplys AC, Petermann E, Petersen C, Dikomey E, Borgmann K. DNA damage by X-rays and their impact on replication processes. Radiother Oncol. 2012 Mar;102(3):466–471.

191. Adamowicz M, Vermezovic J, d Adda di Fagagna F. NOTCH1 Inhibits Activation of ATM by Impairing the Formation of an ATM-FOXO3a-KAT5/Tip60 Complex. Cell Rep. 2016 Aug 23;16(8):2068–2076.

192. Vermezovic J, Adamowicz M, Santarpia L, Rustighi A, Forcato M, Lucano C, et al. Notch is a direct negative regulator of the DNA-damage response. Nat Struct Mol Biol. 2015 May;22(5):417–424.

193. Giuranno L, Wansleeben C, Iannone R, Arathoon L, Hounjet J, Groot AJ, et al. NOTCH signaling promotes the survival of irradiated basal airway stem cells. Am J Physiol Lung Cell Mol Physiol. 2019 Sep 1;317(3):L414–L423.

194. Levy D, Adamovich Y, Reuven N, Shaul Y. Yap1 phosphorylation by c-Abl is a critical step in selective activation of proapoptotic genes in response to DNA damage. Mol Cell. 2008 Feb 15;29(3):350–361.

195. Cottini F, Hideshima T, Xu C, Sattler M, Dori M, Agnelli L, et al. Rescue of Hippo coactivator YAP1 triggers DNA damage-induced apoptosis in hematological cancers. Nat Med. 2014 Jun;20(6):599–606.

196. Xiong S, Patrushev N, Forouzandeh F, Hilenski L, Alexander RW. PGC-1α Modulates Telomere Function and DNA Damage in Protecting against Aging-Related Chronic Diseases. Cell Rep. 2015 Sep 1;12(9):1391–1399.

197. Lai C-Q, Tucker KL, Parnell LD, Adiconis X, García-Bailo B, Griffith J, et al. PPARGC1A variation associated with DNA damage, diabetes, and cardiovascular diseases: the Boston Puerto Rican Health Study. Diabetes. 2008 Apr;57(4):809–816.

198. Lu J, Hu Z, Deng Y, Wu Q, Wu M, Song H. MEKK2 and MEKK3 orchestrate multiple signals to regulate Hippo pathway. J Biol Chem. 2021 Feb 8;100400.

199. Cazzalini O, Scovassi AI, Savio M, Stivala LA, Prosperi E. Multiple roles of the cell cycle inhibitor p21(CDKN1A) in the DNA damage response. Mutat Res. 2010 Jun;704(1-3):12–20.

200. Cheon H, Holvey-Bates EG, Schoggins JW, Forster S, Hertzog P, Imanaka N, et al. IFNβ-dependent increases in STAT1, STAT2, and IRF9 mediate resistance to viruses and DNA damage. EMBO J. 2013 Oct 16;32(20):2751–2763.

201. Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, et al. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. Genes Dev. 2002 Jan 15;16(2):245–256.

202. Bakkenist CJ, Kastan MB. DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. Nature. 2003 Jan 30;421(6922):499–506.

203. Rabalski AJ, Gyenis L, Litchfield DW. Molecular pathways: emergence of protein kinase CK2 (CSNK2) as a potential target to inhibit survival and DNA damage response and repair pathways in cancer cells. Clin Cancer Res. 2016 Jun 15;22(12):2840–2847.

204. Papin JA, Mac Gabhann F, Sauro HM, Nickerson D, Rampadarath A. Improving reproducibility in computational biology research. PLoS Comput Biol. 2020 May 19;16(5):e1007881.

205. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, et al. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. Methods Mol Biol. 2016;1374:23–54.

206. Weinstein JN. Spotlight on molecular profiling: "Integromic" analysis of the NCI-60 cancer cell lines. Mol Cancer Ther. 2006 Nov 6;5(11):2601–2605.

207. Hu Y, An Q, Sheu K, Trejo B, Fan S, Guo Y. Single Cell Multi-Omics Technology: Methodology and Application. Front Cell Dev Biol. 2018 Apr 20;6:28.

208. Tang L. Multiomics sequencing goes spatial. Nat Methods. 18(1):31.