UNIVERSIDAD
DE GRANADA

PROGRAMA DE DOCTORADO EN BIOMEDICINA (B11.56.1)

International Doctoral Thesis

(Tesis Doctoral con Mención Internacional)

Analysis of Functional Annotations in Regulatory Elements

(Análisis de Anotaciones Funcionales en Elementos Reguladores)

Adrián Garcia Moreno

Directed by

Pedro Carmona Sáez

Centro Pfizer – Universidad de Granada – Junta de Andalucía de Genómica e Investigación Oncológica (GENYO)

Granada, 2023

Editor: Universidad de Granada. Tesis Doctorales
Autor: Adrián García Moreno
ISBN: 978-84-1195-073-2
URI: https://hdl.handle.net/10481/85099

Got up early found something's missing
My only name
No one else sees but I got stuck
And soon forever came

Stopped pushing on for just a second
Then nothing's changed
Who am I this time, where's my name?
Guess it crept away

No one's calling for me at the door
An unpredictable won't bother anymore
And silently gets harder to ignore

I forgot that I might see
So many beautiful things
I forgot that I might need
To find out what life could bring

Beautiful Things - Andain

Agradecimientos

¡Por mi y por todos mis compañeros!

En GENyO. Gracias infinitas Pedro, por todo tu apoyo y por todo lo que me enseñas para saber enfrentarme al mundo de la investigación. Gracias Raul por ser mi mejor pinche, la mitad de esta tesis es tuya. Gracias Jordi (y Eli) por tus inacabables consejos científicos y de vida, eres el sabio del grupo. Gracias Dani por ser una inspiración con tu incansable fuente trabajadora y artística. Gracias Juanan por tu claridad y seriedad ante nuestras tesituras. Gracias Ivan (e Ines) por todos los retiros espirituales que promueves en el grupo. Gracias Samu por tu disposición, haber cogido mi relevo y ayudarme en la etapa final. Gracias a Marina por tu constante "risueñez". Gracias a Jose por tu arrasador paso por el grupo. Gracias a Alberto Ramirez y a Manolo por vuestro apoyo esencial a nivel técnico. Y gracias muchas a quienes en una sola conversación por breve que sea y en cualquier escenario compartimos todos los sentimientos que desata hacer una tesis doctoral hoy en día y que te dan la clave y la motivación para continuar: Iris, Abel, Joan, Navajas, Rita, Silvia, Heavy, Alberto, Alex, Goncho, Amador, Agustin, Ara, Jose, Juanmi, Ismael, Juanra, Laz, Carlos Peris, Felix, Juansan, Jorge, Kala, Olivia, Marisa, Jesus, Joselu, Paula, Ines, Jia Juin, … Y así podría seguir nombrando la gran familia que es GENyO.

En BioInfoGRX y RSG España. Gracias a Luis y a Sara Monzón por haber confiado en mí para continuar estos proyectos y formar equipazos. Sois muy grandes por el esfuerzo altruista e impagable que hacéis para la comunidad. He aprendido muchísimo de vosotros: Marisol, Parra, Ana, Ane, Erika, Mireia, Roberto, Alvaro, Fernando, Tamara, Eva, Zulema, Irene, Carla, Mónica, … Gracias a Oscar Huertas y Jesús por vuestro apoyo con esto y las oportunidades obtenidas.

Thanks to Professor André Gerber and Dr Ibtissam Jabré for welcoming me to the University of Surrey and teaching me a new way of working and for giving me the opportunity to contribute my knowledge to your research.

Gracias Iris, por haberme encontrado otra pasión y haber podido conocer a gente tan maravillosa a la que nos mueve una misma filosofía, trabajar tan duro como divertirse: Ana, Bellonez, Pepe, Agustin, Rogelio, Jesus, Oscar, Cristina, Amalia, Marta, …

Gracias a mis amigos de toda la vida: Andrés y familia por ser un motor de mi pasión bioinformática. A Sergio por tu incesante disponibilidad. A Jota y a Dani por que me molestais de vez en cuando con muchas alegrías.

Gracias Alba, por ser tú, por ser como eres, por que sin ti creo que no estaría ni aquí, ni así de bien.

A mi familia de catetos: mamá, papá, hermano y hermana.

Y gracias a tí también si estás leyendo esto, porque es probable que tenga mucho que agradecerte y yo soy muy malo para los nombres.

Quality Criteria to Apply for the Degree of International PhD" by the University of Granada

This doctoral thesis has been prepared according to the University of Granada requirements to apply for an International PhD.

INTERNATIONAL RESEARCH STAY

During the period September 2022 – December 2022, an international internship was performed at the Faculty of Health and Medical Sciences, University of Surrey department of Microbial Sciences in Guildford, United Kingdom. During this 3 months visit, Prof. André Gerber supervised the project titled "Functional characterization of regulatory networks via multi-omics integration" funded by the European Molecular Biology Organization (EMBO) Scientific Exchange grant.

LANGUAGE OF THE THESIS AND PRESENTATION

This doctoral thesis has been written in English and the conclusions will be defended in English. Following the University of Granada regulation, the abstract and conclusions sections have been written in Spanish too.

Grants and Funding

This doctoral thesis has been possible thanks to the following grants awarded to the candidate:

- Scientific Exchange Grant. European Molecular Biology Organization (EMBO).

In addition, this thesis received support from the following research projects:

- Descubriendo cómo se produce la Transferencia Selectiva de ARN desde Megacariocitos a Plaquetas: un proyecto multidisciplinar. Investigadores Principales: Jorge Cerón Hernández y **Adrian García Moreno**. Plan Propio de Investigación y Transferencia de la Universidad de Granada 2021, Proyectos de Investigación Precompetitivos para Jóvenes Investigadores. Proyectos para estudiantes de doctorado. Reference: PPJIB2021-20. Duration 01/01/2022 - 31/12/2022
- Medicina genómica de precisión: Integración de datos ómicos para descubrimiento de biomarcadores e inferencia de redes. Investigador principal: Pedro Carmona Sáez. Ministerio de ciencia e innovación**,** Proyectos del Plan Nacional 2020. Reference: P20_00335. Duration: 01/09/2021 - 30/11/2024.
- AUTOIMMOMICS: Desarrollo de una plataforma centralizada de datos ómicos en enfermedades autoinmunes para el descubrimiento de nuevos tratamientos y biomarcadores. Investigadores principales: Pedro Carmona Sáez y Marta Alarcón Riquelme. Junta de Andalucía, Proyectos I+D+i del Programa Operativo FEDER 2020. Reference: B-CTS-40-UGR20. Duration: 01/07/2021 - 30/06/2023
- Medicina genómica de precisión en enfermedades autoinmunes: Inferencia de redes de regulación y búsqueda de tratamientos. Investigador principal: Pedro Carmona Sáez. Consejería de Transformación Economía, Industria, Conocimiento y Universidades, Junta de Andalucía, Proyectos I+D+i 2020. Reference: P20_00335. Duration: 04/10/2021 - 30/06/2023

Index

Abstract

The progress in high-throughput techniques, characterised by enhanced measurement accuracy and affordability, has significantly contributed to our improved comprehension of biological systems at the molecular level. This development has propelled the advancement of omics biomedicine research, specially, facing the current challenges that complex diseases present. However, the high heterogeneity of complex diseases stresses the need of a personalised medicine and the integration of the different layers that regulate biological systems. The general purpose of these studies is to identify biomarkers inspecting the crosstalk between the different molecules that govern the genetic information flow. Commonly, the results of omics data investigation yield large lists of candidate biomarkers. Making sense out of these requires bioinformatics methodologies, particularly, the functional annotations enrichment analysis. It applies a statistical test to evaluate the overrepresentation of biological annotations within a list of biomarkers in comparison to a reference background. While it is a well established methodology for genes and proteins there is a notable lack of tools that enable the exploration of functional implications associated with regulatory elements. This thesis's general objective is to address the existing gap contributing to the biomedical scientific community with a functional enrichment tool to analyse regulatory elements.

After carefully reviewing the state-of-the-art enrichment methodologies for miRNAs, we learnt that miRNAs, as well as CpG methylation islands and transcription factors, have a common method that consists of inferring their functional implications through the annotations associated with their target genes. This is because the predominant functional terms databases are dedicated to genes and the annotations of regulatory elements are mainly describing their natural role and not their downstream functional effect on the target genes. In the concrete case of analysing the associated genes of CpGs and miRNAs, the traditional enrichment method which applies a test based on the central hypergeometric distribution over the associated genes produces biassed results towards specific and related functional terms mainly related with cell cycle, regulation processes and cancer. Current tools propose different solutions for the analysis of miRNAs and CpG islands. For instance, to avoid the

traditional approach limitations in miRNAs, direct miRNAs set annotations must be tested which can be obtained either by expert curation or after transforming gene-based annotations to the miRNAs-level. Conversely, a well-established unbiased alternative for CpGs analysis employs the Wallenius noncentral hypergeometric test but, surprisingly, no miRNAs literature hinted about it. Our objective here is focussed on assessing and implementing a novel adaptation of the Wallenius method for the analysis of miRNAs.

The novel method and the evaluation of other known methods for the unbiased functional enrichment analysis of regulatory elements has motivated the development of a new GeneCodis version. To fulfil this objective the new version required a complete reengineering of the application. As a result, GeneCodis 4 offers the latest required methods to perform functional enrichment analysis of lists of genes, proteins, CpGs, miRNAs and transcription factors. The update also provides an improvement of the co-annotation discovery algorithm, an expansion of the annotations and organisms database and new interactive visualisations. It is equally accessible for bioinformatics and bench scientists thanks to its implementation as a webtool with an application programming interface.

Finally, almost no literature studies the enrichment analysis of transcription factors lists. In this context, the authors of the only tool to perform singular enrichment analysis of transcription factors, TFTenricher, appear to have overlooked the biassed enrichment analysis of regulatory elements. This presented an opportunity for us to demonstrate that the varying number of transcription factors per regulated gene contributes to the constant enrichment of signalling pathways, transcription regulation, cell cycle and cancer terms. Finally, we validated the power of the Wallenius approach in the transcription factors context by means of null simulations and two real cases reanalysis.

Resumen

Los avances en las técnicas de alto rendimiento, caracterizadas por una mayor precisión y asequibilidad de las mediciones, han contribuido significativamente a mejorar nuestra comprensión de los sistemas biológicos a nivel molecular. Este desarrollo ha impulsado el avance de la investigación de las ómicas en biomedicina, especialmente, de cara a los retos actuales que plantean las enfermedades complejas. Sin embargo, la gran heterogeneidad de las enfermedades complejas acentúa la necesidad de una medicina personalizada y de la integración de las diferentes capas que regulan los sistemas biológicos. Estos estudios buscan identificar biomarcadores a partir de investigar la relación entre las distintas moléculas que gobiernan el flujo de información genética. Por lo general, los resultados de la investigación de datos ómicos producen grandes listas de biomarcadores candidatos. Para darles sentido se requieren metodologías bioinformáticas, en particular, el análisis de enriquecimiento de anotaciones funcionales. Éste método aplica una prueba estadística para evaluar la sobrerrepresentación de anotaciones biológicas dentro de una lista de biomarcadores en comparación con una referencia. Aunque el análisis de enriquecimiento funcional de genes y proteínas es una metodología establecida, existe una notable carencia de herramientas que permitan explorar las implicaciones funcionales asociadas a elementos reguladores. El objetivo general de esta tesis es abordar el vacío existente contribuyendo a la comunidad científica biomédica con una herramienta de enriquecimiento funcional para analizar listas de elementos reguladores.

Tras revisar detenidamente el estado del arte de las metodologías de enriquecimiento para miARNs aprendemos que tanto estos como las islas CpG de metilación y factores de transcripción, tienen un método común que consiste en inferir sus implicaciones funcionales mediante las anotaciones asociadas a sus genes diana. Esto se debe a que las bases de datos de términos funcionales predominantes están dedicadas a los genes y las anotaciones de los elementos reguladores describen principalmente su papel natural y no su efecto funcional en los genes diana. En el caso concreto del análisis de los genes asociados a CpGs y miARNs, el método tradicional de enriquecimiento que aplica un test basado en la distribución hipergeométrica central sobre los genes asociados produce resultados sesgados hacia términos

funcionales específicos y relacionados principalmente con el ciclo celular, los procesos de regulación y el cáncer. Las herramientas actuales proponen diferentes soluciones para el análisis de miARNs e islas CpG. Por ejemplo, para evitar las limitaciones del enfoque tradicional en miARNs, se deben testar las anotaciones del conjunto de miARNs, que se pueden obtener mediante la curación directa por expertos o tras transformar las anotaciones basadas en genes al nivel de miARNs. Por otro lado, una alternativa no sesgada para el análisis de CpGs emplea la distribución de Wallenius sobre la cual, sorprendentemente, ningún artículo sobre miARNs lo menciona. Nuestro objetivo aquí se centra en la evaluación y aplicación de una nueva adaptación del método de Wallenius para el análisis de miARNs.

El nuevo método y la evaluación de otros conocidos para el análisis de enriquecimiento funcional no sesgado de elementos reguladores ha motivado el desarrollo de una nueva versión de GeneCodis. Para cumplir este objetivo, la nueva versión ha requerido una reingeniería completa de la aplicación. Como resultado, GeneCodis 4 ofrece los últimos métodos necesarios para realizar análisis de enriquecimiento funcional de listas de genes, proteínas, miARNs, CpGs y factores de transcripción. La actualización también proporciona una mejora del algoritmo de descubrimiento de co-anotaciones, una ampliación de la base de datos de anotaciones y organismos y nuevas visualizaciones interactivas. Es igualmente accesible para bioinformáticos y científicos de laboratorio gracias a su implementación como herramienta web con una interfaz de programación de aplicaciones.

Por último, casi ninguna literatura estudia el análisis de enriquecimiento de listas de factores de transcripción. En este contexto, los autores de la única herramienta para realizar análisis de enriquecimiento singular de factores de transcripción, TFTenricher, parecen haber pasado por alto el análisis de enriquecimiento sesgado de elementos reguladores. Esto nos brindó la oportunidad de evaluar y demostrar que el número variable de factores de transcripción por gen regulado contribuye al enriquecimiento constante de términos de vías de señalización, regulación de la transcripción, ciclo celular y cáncer. Por último, hemos validado la potencia del enfoque de Wallenius en el contexto de los factores de transcripción mediante simulaciones nulas y el reanálisis de dos casos reales.

Abbreviations

DNA: deoxyribonucleic acid

RNA: ribonucleic acid

mRNA: messenger RNA

ncRNA: non coding RNA

tRNA: transfer RNA

rRNA: ribosomal RNA

miRNA: microRNA

snRNA: small nuclear RNA

lncRNA: long non coding RNA

TF: transcription factor

cDNA: complementary DNA

RNA-seq: RNA sequencing

scRNA-seq: single-cell RNA-seq

BS-seq: bisulfite sequencing

GO: Gene Ontology

GO BP: Gene Ontology biological process

KEGG: Kyoto Encyclopedia of Genes and Genomes

SEA: singular enrichment analysis

ORA: over-representation analysis

GSEA: gene set enrichment analysis

MEA: modular enrichment analysis

TAM: tool for miRNA set analysis

HMDD: human microRNA disease database

MNDR: mammalian ncRNA-disease repository

API: application programming interface

1. Introduction

1.1. The flow of genetic information

In 1956, Francis Crick delivered lectures aiming to explain the potential pathways that our genetic code could follow in protein synthesis. With limited experimental evidence at the time, Crick made a notable statement: “Once information has got into a protein it can’t get out again” [1]. It was only 14 years later that he further explained what he called, unfortunately or not, Central Dogma of Biology [2]. He meant that a biological system cannot extract information from a protein to replicate it or to obtain the RNA and DNA that encode it, thus determining a clear direction of the genetic information flow. From a point of view purely based on the residue-by-residue transfer of sequential information, Crick’s theory remains true. The replication processes maintain the genetic information of the individual organism during development and reproduction processes. Then the transcription of DNA information into a RNA molecule, known as gene expression, allows the reading of our genetic code during the traduction and generates the proteins necessary for the homeostasis (Figure 1).

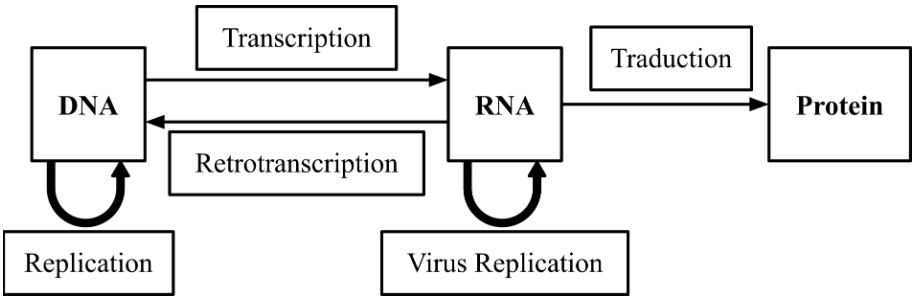


Figure 1. Flow of genetic information according to the central dogma of biology.

Nevertheless, as Crick envisaged, the flow of genetic information is regulated by various actors operating in diverse directions. Actually, genes can be classified into two distinct types based on their ultimate products: coding genes, which give rise to proteins, and non-coding genes, which generate RNAs with diverse functions, some of which remain incompletely understood.

1. INTRODUCTION

The significance of proteins was established prior to their identification as the ultimate products in the flow of genetic information. It is now unequivocally recognized that proteins serve as the primary functional units and play crucial roles in living organisms. They provide structural support, act as biochemical catalysts, function as hormones or enzymes, and their mutation, excess or insufficiency can lead to various diseases, including nervous system disorders, metabolic disturbances, organ failure, and even mortality [3]. Conversely, proteins also serve as targets for therapeutic intervention or serve as therapeutic agents themselves [4,5].

Non-coding genes give rise to non-coding RNAs (ncRNAs), which are RNA molecules that do not undergo translation into proteins. The repertoire of ncRNAs encompasses a wide array of types and functionalities, including transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs), long non-coding RNAs (lncRNAs), and microRNAs (miRNAs).

tRNAs function as carriers of amino acids, facilitating the generation of the protein sequence by serving as the complementary template for mRNA [6]. rRNAs, on the other hand, constitute the structural components of ribosomes and possess catalytic activity in mediating the interaction between tRNAs and mRNA, ultimately leading to the synthesis of protein sequences [7]. snRNAs play a crucial role in pre-mRNA splicing, the process by which gene introns, untranslated regions, are removed and exons, translated regions, are joined together [8]. lncRNAs exert their regulatory influence on gene expression through various mechanisms, including modulation of chromosome structure, transcriptional control, splicing regulation, modulation of mRNA stability and availability, and post-translational modifications [9]. Besides, miRNAs also regulate gene expression but specifically through the silencing of target mRNA [10]. It is important to note that these examples represent only a fraction of the extensive catalogue of ncRNAs known to possess specific functions [11].

The understanding of the flow of genetic information has undoubtedly advanced significantly, revealing its intricate complexity. Merely contemplating the central dogma of biology provides only a constrained glimpse into the multitude of factors that contribute to the functionality of a biological system. The present thesis concentrates on exploring some

well-characterised actors that play pivotal roles in gene regulation, specifically DNA methylation, transcription factor proteins and miRNAs (Figure 2).

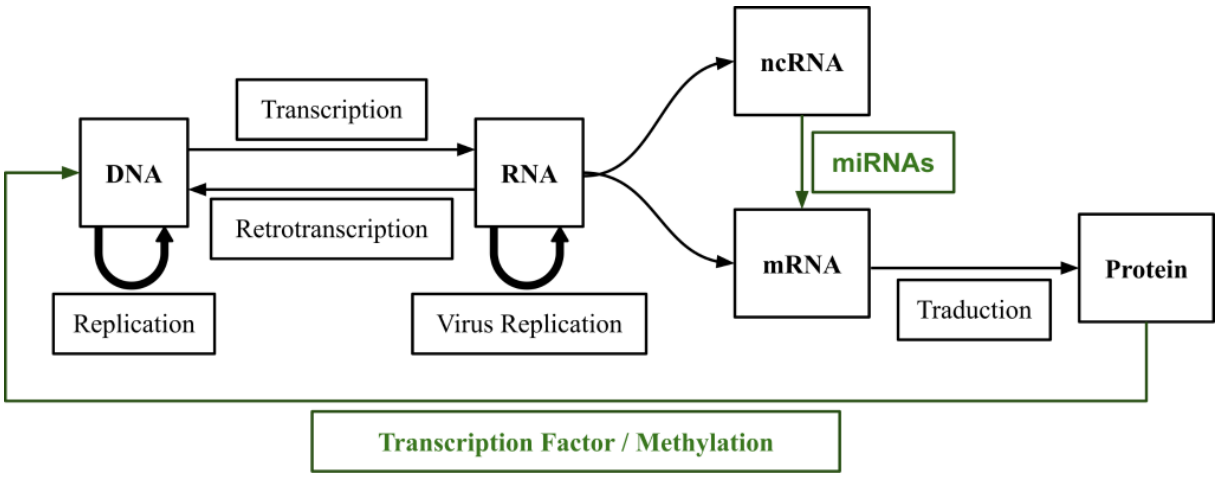


Figure 2. Flow of genetic information according to current knowledge highlighting in green the biological processes evaluated in this doctoral thesis.

1.2. Regulation of gene expression

While there are several mechanisms through which gene expression can be controlled, the scope of this research is limited to the investigation of the following described regulatory factors.

1.2.1. DNA methylation

DNA methylation was first identified as a phenomenon in mammals during the same period when DNA was established as the hereditary material in the 1940s. However, it was not until the 1980s that the involvement of DNA methylation in gene expression regulation was definitively demonstrated [12]. Presently, **DNA methylation** is a well-characterised **epigenetic modification**. Methylation primarily occurs at CpG sites, which are regions in the DNA sequence where a cytosine is followed by a guanosine in the 5' to 3' direction. DNA methylation operates at two levels: maintenance methylation, which preserves the methylation pattern on the newly synthesised daughter strand by copying the pattern from the parental

DNA strand, and de novo methylation, which establishes methylation patterns on previously unmethylated sites [12].

DNA methylation serves as a **repressive mechanism for gene transcription**, particularly for genes with CpG-rich promoters, indicating that DNA methylation patterns are **dynamic and responsive to environmental stimuli** [13]. Consequently, the regulation of DNA methylation removal and reestablishment exhibits substantial variations across different stages of development. Despite the fact that all cells within an organism possess the same DNA sequence, the tissue-specific gene transcription profile is shaped by the distinct methylation patterns [14]. Moreover, DNA methylation plays a pivotal role in genomic imprinting, X-chromosome inactivation, and the suppression of repetitive elements transcription and transposition [15]. Disruptions in the machinery and patterns of DNA methylation have been implicated in a wide range of diseases, spanning from congenital immunodeficiency syndromes, growth phenotypes, and neurodegeneration to haematological cancers [14]. Consequently, the examination of DNA methylation patterns has emerged as a promising avenue for biomarker discovery, disease classification, and potential therapeutic targets in the field of immune-oncology [16].

1.2.2. Transcription factors

The earliest evidence for the presence of DNA sequences responsible for the regulation of gene expression was obtained in the early 1980s through studies on the Drosophila Hsp 70 heat-shock gene [17]. Subsequently, by the end of the same decade, the existence of **transcription factors** (TFs) was firmly established [18].

TFs are **proteins** that possess the ability to **recognize specific DNA sequences**, thereby exerting control over chromatin organisation and transcriptional processes. These factors form a complex regulatory system that **orchestrates the precise expression of the genome**. The sequence characteristics of TFs, including their DNA binding domains, regulatory regions, and physiological roles, are often conserved across metazoans, underscoring the fundamental importance of their regulatory networks. The activities of TFs guide the development and

specialisation of different cell types and are known to exert control over specific pathways, including immune responses. In laboratory settings, TFs can be employed to drive cellular differentiation and can even induce dedifferentiation and trans-differentiation processes [19].

TFs constitute approximately 8% of all human genes and are implicated in a wide range of phenotypes and diseases. Numerous TF-related disorders are associated with neurodevelopmental processes, the immune system and cancer. Studying TF-associated disorders can be challenging due to the highly deleterious nature of mutations affecting TFs, which aligns with their evolutionary conservation [20 21].

1.2.3. MicroRNAs

MicroRNAs (miRNAs) are one class of non-coding RNA that are extensively studied at the moment, primarily owing to their distinctive molecular characteristics and specific functional roles. They were initially discovered and referred to as interfering RNAs in 1993 through investigations conducted in the model organism *Caenorhabditis elegans* [22]. **miRNAs are RNA molecules** of approximately 22 nucleotides conserved in metazoan and plant species [23,24]. Functionally, they act as **post transcriptional regulators silencing** gene expression by **binding to complementary sequences within mRNA transcripts**. This interaction is facilitated by the Argonaute proteins and other components of the RNA-induced Silencing Complex (RISC) [25].

The human genome is estimated to contain approximately 2000 miRNAs [26]. The observed dysregulation of miRNAs in disease states has prompted extensive research into their diagnostic and prognostic potential. Moreover, miRNA-based therapeutics, including miRNA mimics and miRNA inhibitors, have shown promise in preclinical development as novel therapeutic agents [27]. In the context of allergic inflammation, several miRNAs have been identified as important players in diseases such as asthma, atopic dermatitis, allergic rhinitis, and eosinophilic esophagitis [28]. Dysregulated expression of specific miRNAs has been linked to tumour development, progression, and metastasis in various types of cancer. These miRNAs can function as oncogenes or tumour suppressors, influencing critical cellular

processes involved in cancer biology [29].

1.3. Omics revolution in biomedicine

In the field of biological sciences, omics approaches have revolutionised research by employing high-throughput techniques capable of simultaneously measuring and quantifying thousands of properties and entities. However, the development of such techniques was preceded by crucial advancements in the field. To establish a starting point for this revolution, we can trace back to the publication of the **Sanger** sequencing method in 1977 [30]. Another important milestone occurred in 1983 when microarrays were proposed as a method to capture cells [31]. This technology quickly gained momentum and found diverse applications in the years that followed, revolutionising various areas of research. Another groundbreaking development came two years later, the polymerase chain reaction (**PCR**) was presented to change the biochemical field [32]. Furthermore, in 1986, Applied Biosystems (ABI) introduced the first automated DNA sequencer based on Sanger technique, the ABI370. Four years later, the **Human Genome Project** started, although omics approaches were still prohibitive for most laboratories. The omics revolution did not start until 1995 with the development of DNA **microarrays** able to measure gene expression at a very low cost [33]. In 1998, a significant advancement in genome sequencing technology occurred with the development of the first sequencer utilising capillary electrophoresis. This marked the initiation of the first generation in the era of genome sequencing and established itself as the predominant technology in various genome discovery projects. Before the end of the Human Genome Project, the genomes of certain bacteria and metazoans had already been sequenced, it was not until 2003 that the initial version of the human genome was published. The assembly of these reference genomes played a crucial role in enabling subsequent generations of sequencing technologies, leading to a paradigm shift in biomedical sciences and the emergence of the omics revolution.

1.3.1. Genomics

Genomics is the study of the genetic material that codes for an organism or biological system

as a whole, the **genome**, commonly DNA with exceptions in RNA viruses. The genome is constant across all the cells that form an organism and contains the instructions for its homeostasis and to produce copies of itself.

Genomics was born with the first automatic machines that allowed the **DNA sequencing**, DNA-seq, and currently we are in the third generation of this technology. In 2005, the **second generation** of sequencing technologies emerged with Roche at the forefront. This new generation addressed the limitations of the earlier ABI sequencers, which yielded a relatively low number of reads per run. The advent of technologies such as Illumina enabled the simultaneous production of several **billion reads**, a significant increase compared to the previous generation. Notably, these advancements eliminated the need for capillary electrophoresis, making the sequencing process faster and more cost-effective. However, the second-generation technologies are **constrained** by the average **read length**, typically ranging from 75 to 900 paired bases. Despite the substantial increase in the number of reads per run achieved by these technologies, their read lengths remain relatively short. In contrast, the **third generation** of sequencing technologies, represented by platforms like PacBio and Oxford Nanopore [34], has surpassed this limitation. These technologies offer read lengths varying from a thousand to more than ten thousand base pairs. In contrast to short-read technologies, the **long-read technologies** in general have a lower rate of accuracy, which has been overcome with the high-fidelity sequencing (HIFI PacBio) [35].

Short-read sequencers are widely employed in genomics studies due to their accessibility and accuracy. Nevertheless, in 2019, the GRCh38.p13 human genome assembly contained approximately 8% of DNA that remained practically unknown. This lack of knowledge was primarily attributed to the complex and repetitive nature of certain genomic regions, which posed challenges for the assembly of short-read contigs. These elusive regions encompass various elements critical to essential biological functions, including pericentromeric and subtelomeric regions, segmental duplications, ampliconic gene arrays, and ribosomal DNA arrays. By means of long-read sequencing, the T2T (telomere to telomere) project could reveal these hidden parts of the human genome and published the T2T-CHM13, the latest reference assembly [36].

Genomics allow us to study, primarily, the mutations in DNA. However, many diseases caused by these mutations do not require omics approaches to be studied. Some of these are discernible at the macromolecular level because they affect the normal structure of the chromosomes, e.g. Down Syndrome. Other disorders are caused by a single gene mutation, called mendelian or monogenic, e.g. cystic fibrosis. Notwithstanding, genomics play a key role in studying complex diseases that are influenced by multiple combinations of mutations occurring in either the same or different genes as is being found in Meniere Disease [37].

1.3.2. Proteomics

Proteomics is the study of the entire complement of proteins produced by an organism or biological system, the **proteome**, as whole. In contraposition with the genome, the proteome is dynamic in a multicellular organism and changes over time. Proteins are the main functional product of the genetic material whose distribution could explain the specialisation of cells and tissues.

Proteins are macromolecules composed of various amino acids, and the specific sequence of these amino acids is encoded by the genetic material. The genetic code serves as a dictionary that assigns specific amino acids to three DNA/RNA molecules known as codons. It is worth mentioning the significant contribution of Marianne Grunberg-Manago's work in Severo Ochoa's laboratory in 1955, where RNA polymerase was isolated and RNA synthesis was achieved purely in vitro. These advancements, along with other key discoveries, paved the way for the subsequent experiments by Niremberg and Khorana, leading to the elucidation of the relationship between codons and the genetic material [38,39]. The genetic code is characterised by its non-overlapping and specific nature, meaning that each codon corresponds to a unique and predetermined amino acid. Furthermore, the genetic code is degenerate or redundant, indicating that multiple codons can code for the same amino acid. Currently, there are 64 codons that encode for the twenty different amino acids present in all taxa, with a few additional variations found in specific organisms [40].

The set of amino acids that constitute a protein is assembled by cellular machinery, and their precise arrangement in a **three-dimensional conformation** is crucial for their functional interactions with various molecules [41]. It is worth noting that for many years, while the amino acid sequences of billions of proteins were inferred through genome sequencing [42], only a small fraction of them, around a hundred thousand, had their three-dimensional structure deciphered [43]. However, recent advancements in computational approaches, exemplified by the milestone development of the **AlphaFold** deep learning algorithm, have revolutionised our ability to accurately predict the three-dimensional structures of proteins, including the entire human proteome [44,45].

Despite these remarkable developments, the experimental characterization of proteins remains a challenging endeavour, primarily due to the wide diversity observed in their dynamic nature, residue modifications, abundance, conformations, molecular sizes, hydrophobicity, and hydrophilicity [46]. Although various methods exist for studying small and specific sets of proteins, these approaches typically do not yield high-throughput measurements [47]. In the field of proteomics, mass spectrometry is the predominant technique, as it enables the identification and quantification of protein mixtures, including their post-translational modifications. Nonetheless, it is often coupled with different protein isolation methods depending on the specific requirements of the study [48].

Proteomics plays a crucial role in clinical research due to the pervasive involvement of proteins in all biological processes. Proteins serve various functions, including enzymatic activity, molecule transport, toxin production, adhesion, invasion, signalling, and receptor interactions. Consequently, proteins play a significant role in the initiation and progression of numerous diseases. By identifying the proteins present in viruses and prokaryotic cells that contribute to infection and disease transmission, researchers can develop targeted vaccines designed to specifically combat these proteins. Additionally, proteomics has facilitated the development of targeted therapies, which involve the use of drugs or other molecules that selectively target disease-associated proteins. These therapies can be more effective and have fewer side effects compared to traditional treatments. Obtaining a comprehensive understanding of the proteome is crucial for unravelling the molecular mechanisms

underlying health and disease [49].

1.3.3. Epigenomics

Epigenomics is the study of all the reversible **DNA modifications** that do **not** alter the DNA **sequence**, the epigenome, as a whole. Along with the genome, the epigenome marks are often maintained from cell to cell and also from progenitors to the next generation.

The epigenome marks include a variety of chemical compounds and proteins, mainly histones, which can attach to DNA and modulate the activation or repression of gene transcription. In essence, the epigenome establishes the catalogue of genes that are accessible for gene expression. These modifications occur naturally during development and tissue differentiation, but they can also be influenced by environmental exposures [50]. Among the various types of epigenetic marks, **DNA methylation** has been extensively studied as a crucial regulator of gene expression. Another important epigenetic mark involves **modifications to histone proteins**, which indirectly influence DNA. These modifications play a critical role in determining the chromatin state, with acetylation often associated with euchromatin and methylation associated with heterochromatin. Euchromatin represents the accessible and readable state of DNA, while heterochromatin represents its counterpart, which is typically less accessible for gene expression [51].

The early attempts to investigate the epigenome started before the Sanger sequencing in 1975, with initial efforts focused on the separation of methylated and unmethylated deoxynucleosides. Few years later, a reversed-phase high performance liquid chromatography technique was developed to quantify 5-methylcytosine, which was further enhanced by incorporating mass spectrometry and thin-layer chromatography. Additional methods include radiolabeling and immunological DNA methylation assays. Through the combination of these techniques and the use of methylation-sensitive restriction enzymes, researchers were able to generate the first drafts of genome-wide methylation profiles. However, a significant breakthrough came with the discovery of sodium bisulfite treatment applied to DNA. This

treatment selectively converts unmethylated cytosine residues to uracil at a much faster rate than methylated cytosines. This key insight was harnessed in DNA sequencing methods, particularly in 1992, whereby unmethylated cytosines are converted to uracil while methylated cytosines remain as cytosines. This **bisulfite sequencing** (BS-seq) approach revolutionised the field by enabling detailed analysis of DNA methylation patterns at single-base resolution [52]. Many of these approaches and variations have been applied in parallel to the improvement in DNA sequencing methods, with the only difference that methylation marks are not conserved during the PCR and necessitate previous treatments to maintain their epigenetic information. BS-seq and, to a lesser extent, DNA methylation arrays, are considered the gold standards for achieving single base resolution in epigenetic sequencing. Nonetheless, these are unable to distinguish among other types of cytosine modifications [53]. Because these marks can play somewhat antagonistic roles in gene regulation, this constraint has probably resulted in a number of inaccurate assumptions [54].
Although there have been advancements in developing specific treatments to detect different cytosine modifications, their implementation is still scarce. Recently, bisulfite-free methodologies have emerged as alternatives to overcome limitations and biases associated with traditional bisulfite sequencing. These new approaches, often coupled with long-read DNA sequencing technologies, aim to improve coverage and reduce experimental biases. However, it is important to note that the accurate determination of cytosine modifications using these methods still heavily relies on computer algorithms for prediction and interpretation. While these approaches show promise, they have not yet surpassed the accuracy of bisulfite sequencing [53].

Being able to measure the epigenomic landscape of health and disease is of great importance. Epigenetic modifications play a pivotal role in activating and silencing genes, thereby establishing tissue- and cell-specific transcriptional programs that significantly impact cellular differentiation and development. For instance, identical pluripotent stem cells can differentiate into various cell types based on specific epigenetic signals. Furthermore, the epigenome is also susceptible to environmental influences, which can lead to phenotypic variations. Monozygotic twins, who share the same genetic makeup, may exhibit differences in physical and behavioural traits due to variations in environmental exposures. Throughout an

individual's lifetime, continuous exposure to various stimuli such as diet, exercise, pollution, and noise can induce specific genetic regulatory programs. Therefore, comprehending the dynamic interplay between epigenetic mechanisms and environmental factors is essential for promoting and maintaining human health. [50].

1.3.4. Transcriptomics

Transcriptomics is the study of the **complete collection of RNA** molecules within a biological system, the **transcriptome**, as a whole. Similar to the proteome and epigenome, the transcriptome is highly specific to cell specialisation. It is worth remembering that these include **coding** RNA which are translated into proteins, and the diverse **non-coding** RNAs some of which have clear regulatory roles in gene expression.

In the 1970s, early techniques were developed to capture mRNA molecules by converting them into **complementary DNA** (cDNA) using reverse transcriptase enzymes. In the following decade, the integration of cDNA synthesis with DNA sequencing led to the creation of the serial analysis of gene expression (SAGE) technique, which allowed the measurement and quantification of known gene expression levels. Around 1995, in conjunction with the emergence of genomics, comprehensive profiling of the transcriptome began with the arrival of the microarrays. A few years later, RNA sequencing (RNA-seq) using second-generation sequencing technologies was introduced, enabling more precise and unbiased transcriptome analysis. Microarrays were the preferred method for transcriptomics until late 2000s given its reduced costs and labour. However, their limitation to a predefined set of transcripts are pushing them to a complete disuse towards the current RNA-seq methods [55].

RNA-seq employs the same sequencing platforms and techniques developed throughout the various generations of genome sequencing. However, it involves additional experimental steps to isolate and capture RNA before converting it into cDNA. These preliminary steps are crucial in determining the specific type of RNA to be sequenced. For instance, mRNA molecules are typically captured using poly-A tail enrichment methods, rRNA can be depleted using taxon-specific probes hybridization, and miRNAs can be isolated based on their size

using gel electrophoresis. In addition to quantifying gene expression by measuring transcript abundance, RNA-seq offers advantages over microarrays by providing insights into transcript variability, such as splice variants. Customization of the RNA-seq experiment is essential to account for the inherent heterogeneity in transcript abundance across different tissues and to achieve sensitivity and accuracy. For instance, rare transcripts require a sufficient number of reads, while long and highly expressed transcripts may need to be normalised during subsequent data analysis to mitigate any potential biases.

Currently, one of the most powerful techniques in transcriptomics is **single-cell RNA-seq**. It applies the same RNA-seq methods previously mentioned, but it enables the identification of gene expression profiles for each individual cell through microfluidics and nucleotide barcoding. Cells are isolated and subsequently distinguished using three distinct sequences: one sequence specific to the cell, a unique molecular identifier for the transcript, and a third sequence that labels each sample. This enables the simultaneous sequencing of multiple samples and the comprehensive analysis of the transcriptome for each captured cell and sample. [56].

The applications of transcriptomics are similar to those of the previously covered omics sciences. In the field of biomedicine, transcriptomics plays a pivotal role in diagnosing and molecularly profiling complex diseases, identifying pathogens, studying environmental responses such as drug responses, and suggesting gene functions through knock-in and knock-out experiments [57].

1.4. Omics bioinformatics methods

The widespread adoption of high-throughput omics technologies has transformed our comprehension of biological systems to the molecular level. Nevertheless, these methods yield vast and intricate datasets, where each omics platform presents distinct challenges in terms of measurements. Consequently, it is essential to employ and create suitable mathematical and computational methodologies for the analysis and integration of these data, thereby enabling the formulation of reliable conclusions.

Bioinformatics methods have the objective of identifying biomarkers for disease diagnosis, prognosis and treatment. These biomarkers can encompass individual genes, extensive gene sets, or other biological molecules. They may or may not be associated with specific biological pathways or functions that can be influenced at various levels of the genetic regulatory network. Subsequently, this section provides a description of some of the common methods in the field which have been directly or indirectly employed in processing the data within the scope of the present thesis.

1.4.1. Differential gene expression, protein abundance and methylation

Traditionally, in omics research, particularly since the advent of microarrays, differential gene expression analysis has been a commonly applied approach. However, this thesis introduction also encompasses differential protein abundance and methylation analysis, as these methods share a similar objective and effectively illustrate the core idea. These methods aim to directly examine **quantitative differences** in the transcriptome, proteome, or epigenome **between two or more groups or phenotypes**, often comparing healthy and diseased states. Basically, these analyses involve contrasting variations in transcript abundance, protein levels, and the frequency of methylated cytosines among the conditions being investigated.

These types of analyses begin with quantifying the reads that align to each gene or identifying and mapping peptides. Normally, the raw counts require a data **normalisation step** because of given biases. For example RNAseq is influenced by various factors, such as transcript length, sequencing depth and library size. Longer transcripts may be overrepresented compared to shorter ones simply because they can span more combinations of short-reads. Additionally, if the total number of reads generated differs among samples, those with higher read counts will generally exhibit higher transcript counts, even if their expression levels are relatively the same. Similar challenges arise with varying numbers of CpGs or peptides that can be identified within a gene or protein. Therefore, normalisation is essential to eliminate these quantification biases and enable accurate comparisons. Utilising the normalised values, the

final step involves calculating the **effect size**, e.g. fold change, and testing the **significance** of differences by computing a **p-value** and a p-value adjusted by **multiple testing correction**. Both measures are necessary to assess the actual impact of the divergence between conditions.

One of the most employed statistical tests is the Student's t-test which assesses whether the means of two groups are equal or not. This test assumes that data follow a normal distribution and that the samples of each condition have equal variance. An adaptation of this test is found in the limma R's package, which generates a linear model that adjusts the variance using an empirical Bayes approach. This method is commonly used in gene expression microarrays and proteomics analyses [58]. In RNA-seq other widespread tools, such as edgeR [59] and DESeq2 [60], test the gene expression using a negative binomial distribution. These tools differ from the methods used to normalise and modelise data estimating the dependence between variance and mean. For proteomics and methylation data analysis edgeR is also presented as an option. In the concrete scenario of the methylation BS-seq data analyses it is not always performed at the gene level, instead, a common method is to merge close CpGs sites into regions in rolling windows across the genome before applying the statistical test. Some of the tests that evaluate differentially methylated regions are based on Fisher's exact test [61,62]. Other methods first normalise the methylation status of neighbouring CpGs and apply an adapted t-test to each CpG site whose results are later combined to define the differentially methylated regions [63].

There is a complete catalogue of methods whose statistical tests are prepared for specific sources and thus are built upon a series of assumptions necessary to acknowledge in order to avoid skewed results. Covering the whole spectrum is not part of this thesis scope, however, it is worth mentioning the existence of benchmarking tests. They consist in assessing the accuracy of a set of methods when analysing the same dataset whose expected results are known. Their principal goal is to facilitate the choosing of the correct methods. Here are cited some current benchmarking applied to differential expression analysis tools for RNA-seq [64], for label-free proteomics methods [65] and for differential methylation analyses for BS-seq [66].

The primary objective of the aforementioned methods is to identify a group of genes or proteins that exhibit dysregulation compared to a reference system, typically a control

phenotype. A difficulty about these methods is to differentiate between causation and causality. The causation of the dysregulation can be a set of clinical variants that have altered the basal levels of gene expression or an external stimuli. The causality of the dysregulation can be demonstrated by the clinical symptoms that are shown in a concrete disease. Regardless, these approaches produce straightforward results that can be directly integrated with further data mining or machine learning methods to develop precision medicine. For example, MyPROSLE is a tool that uses the gene expression values of lupus patients to help the treatment judgement for doctors [67].

1.4.2. Clustering

Clustering is an unsupervised data mining technique. In our context, it used to group samples or biological entities based on their similarities, without the need for prior knowledge or labels. Each cluster represents a set of elements that share a specific pattern, which distinguishes them from elements in other clusters. The objective is to uncover inherent patterns and structures within the data, allowing for further analysis and interpretation.

In the context of this thesis, clustering methods are particularly valuable as they permit the identification of groups of genes or proteins that exhibit similar expression or methylation patterns. Such clusters are likely to be involved in common biological functions or pathways, providing insights into underlying disease mechanisms. By analysing these clusters, it becomes possible to identify potential biomarkers that can aid in the diagnosis and treatment of diseases. Moreover, clustering techniques can also reveal different confounding sources namely, batch effects. Two of the most widespread clustering methods in which many bioinformatics tools are based are the hierarchical and the k-means.

Hierarchical clustering creates small subgroups by recursively dividing the data. It can follow a bottom-up direction (agglomerative), starting from the two most similar genes adding the next closest in each step, or in top-down direction (divisive), dividing all the genes in two different groups that will be splitted again recursively. The hierarchical clustering can use different distance measurements which allows studying both continuous and categorical data.

Although it is a straightforward methodology, the clustering of high dimensional datasets might not produce a consistent granularity. A popular application of the hierarchical clustering in gene expression data is through the weighted correlation network analysis (WGCNA) [68] (Figure 3 A).

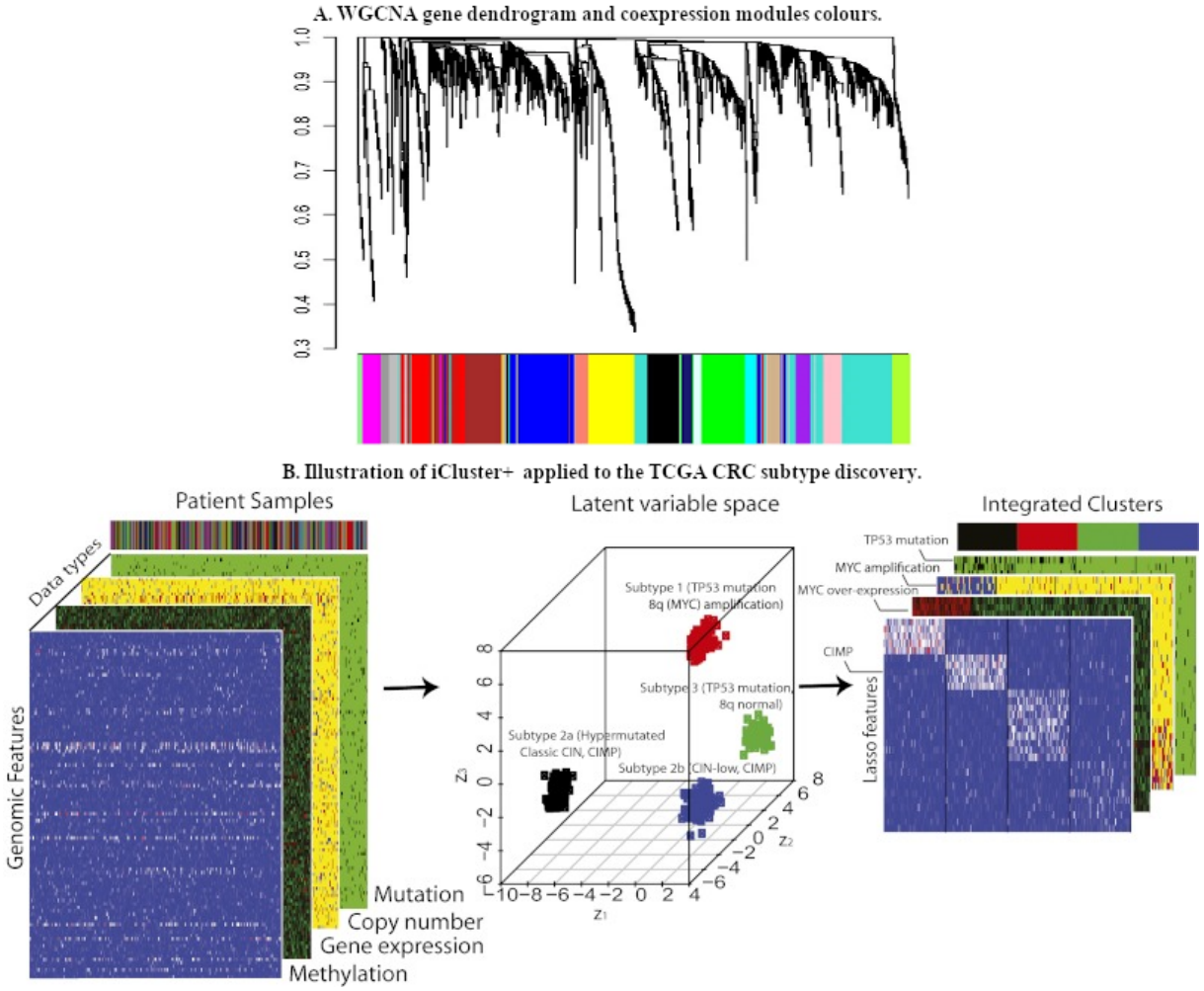


Figure 3. Summary image illustrating the applications of WGCNA (A) and iCluster+ (B), adapted from their original papers.

In the data partitioning clustering methods one of the most extended is the k-means algorithm. This method, unlike hierarchical clustering, requires setting a predefined number in which data is expected to be clustered. This method iteratively divides the data, i.e. genes, in the defined number of clusters and calculates their centroids, when in further permutations the centroids coordinates do not show great variance the algorithm stops. The closest groups of

genes to the centroids define each cluster. The disadvantage is the need of guessing an initial number of groups, however there are different methodologies to propose them. A relevant integration of the k-means clustering is done in the different versions of iCluster, iCluster+ and iClusterBayes [69–71]. They are based on generating different latent variable models to jointly analyse clinical variables and different omics data types (Figure 3 B).

Besides being integrated with different data mining approaches, nowadays, clustering methods are also involved in complete machine learning algorithms, some reviews about this topic can be found here [72–74]. Nevertheless, the overall idea of clustering methods in the omics research is conserved, to obtain groups of candidate genes or proteins that dispense new disease molecular subtypes, possible biomarker selection for diagnosis and therapeutic targeting and, in general sense, also to perform quality controls and generate novel hypotheses based on the subgroups.

1.4.3. Functional enrichment analysis

The analysis of omics datasets using the aforementioned methods often results in extensive lists of potential biomarkers. However, it is crucial to gain insight into the biological processes or pathways affected by these biomarkers in order to validate the experiments and guide further research. Consulting the literature individually for each biomarker is impractical due to the large number of candidates. As a result, this motivated the conception of the functional enrichment analysis. It comprehends a series of widely used methods that statistically associate current knowledge annotations or terms with a given list of biological entities. By employing functional enrichment analysis, researchers can identify the biological functions, pathways, or molecular processes that are significantly enriched within their list of biomarkers. This aids in the interpretation and understanding of the underlying biological implications of the omics data.

The assessment of significant annotations can give us a holistic vision of the action scope of the biomarkers at many levels. Current knowledge annotations are stored in databases that link, mainly genes and proteins, to specific biological functions, phenotypes, localizations or

other molecular species. Thus the creation and maintenance of these databases is essential for the enrichment methods. Majorly, it is the manual expert curation of the scientific literature that backs up the annotations but there are different levels of evidence. Two of the initial and current well established databases, founded in the year 2000, are the Gene Ontology (GO) [75] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [76], both include diverse biological functions, metabolism pathways, and cellular locations where these occur. The number of databases in the field of bioinformatics has increased significantly over the years, providing researchers with a wealth of information and resources for their studies, some of them are covered later in this thesis document.

There are three main types of enrichment analysis that can be classified according to the input list and the annotations to be analysed [77] and are depicted in the following sections.

1.4.3.1. Singular enrichment analysis

The singular enrichment analysis (SEA), was initially proposed in 2003 [78]. This method is also known as over-representation analysis (ORA). Currently, the common SEA approach follows these steps:

1) Define a set of candidate biomarkers. These can be characterised by several methods, for example, because they cluster together or because they have a differential expression when comparing two conditions where biomarkers are selected by a threshold of p-value significance and fold change.
2) Retrieve annotations from a given database and compute the frequencies of each annotation in the input list and the reference list, for instance, all the genes expressed in the studied tissue.
3) Apply a statistical test to evaluate the over-representation of annotations in the input list. A frequently used method is the Fisher's exact test, also commonly named as the central hypergeometric test.
4) Correct p-values for multiple testing.

There are two common limitations of SEA, the overwhelming results that included a large number of enriched annotations and the need to define a set of candidate biomarkers. Despite this, it still is the most widely used method in enrichment analyses because it accepts any type of omics data analysis output and it is implemented in several popular tools such as DAVID, Panther, g:profiler and modenrichR [79–83].

1.4.3.2. Gene set enrichment analysis

The gene set enrichment analysis (GSEA) [84] was developed to overcome some known limitations of SEA. First, the establishment of a significance and/or fold change threshold is arbitrary which can lead to different numbers of candidate biomarkers that can be scarce to capture relevant biological differences or too many in order to reveal a concrete process. Besides, the threshold approach can leave out important biomarkers because it ignores the coordinated effects of the elements that participate in the same pathway [85].

To overcome those limitations, the GSEA is implemented following these steps:

1) Rank all the genes in the experiment according to the difference in expression, for example, fold change. It can also be relative to differences in protein abundance or methylation between two conditions.
2) Calculate an enrichment score for a given annotation commonly based on a normalisation of the Kolmogorov-Smirnov statistic. It expresses if the concentration of genes are overrepresented at the top or bottom of the input ranked list.
3) Finally, it uses a permutation test to calculate an empirical p-value to determine the association of the enrichment score with the given annotation.

This approach is mainly suitable for pair-wise biological studies that produce high-throughout results and permits to establish a ranking criteria. There are currently two main tools to perform this type of analysis, the GSEA standalone tool [85] and a modern alternative in the R's package fgsea [86].

1.4.3.3. Modular enrichment analysis

The modular enrichment analysis (MEA) operates in conjunction with SEA or GSEA, following a similar framework. MEA involves grouping annotations either prior to or following the application of statistical testing. The benefit of this approach is that the integration of related annotations can yield results that offer a comprehensive perspective aligned with the underlying biological data structure.

There are multiple methods available for clustering functional annotations. Several MEA tools make use of the directed acyclic graph structure of the GO, such as topGO [87] and Panther [88,89] before applying SEA. Another tool, ReviGO leverages GO analysis results and produces a visualisation that exploits the semantic similarities of the ontology terms with different measures such as, Resnik and Lin [90]. Finally, different annotation databases can be analysed integratively by using association rules algorithms in tools like GeneCodis [91,92].

MEA is often combined with SEA methods, but it can also apply the statistics from GSEA, thereby sharing the same advantages and limitations as these approaches. Intrinsically, MEA advantage is its ability to decrease redundant outcomes but, in contraposition, genes and terms that have few interrelationships may be underrepresented from the analysis.

Each annotation's functional analysis has its own conveniences. For example, SEA accepts any source of input biomarkers, allowing for the exploration of multiple hypotheses. However, determining an appropriate threshold for fold change and p-value can be challenging which, given its subjectivity, pose a limit to the method replicability. In contrast, GSEA offers higher resolution in such situations, but it requires the use of the entire gene set from the experiment ranked according to a specific criterion that can divide them in two classes, for instance, the fold change positive means overexpressed and the negative underexpressed. Independently of these, MEA can be applied before or after SEA or GSEA, and its main objective is to reduce the numerous enriched annotations into biologically meaningful groups (Figure 4).

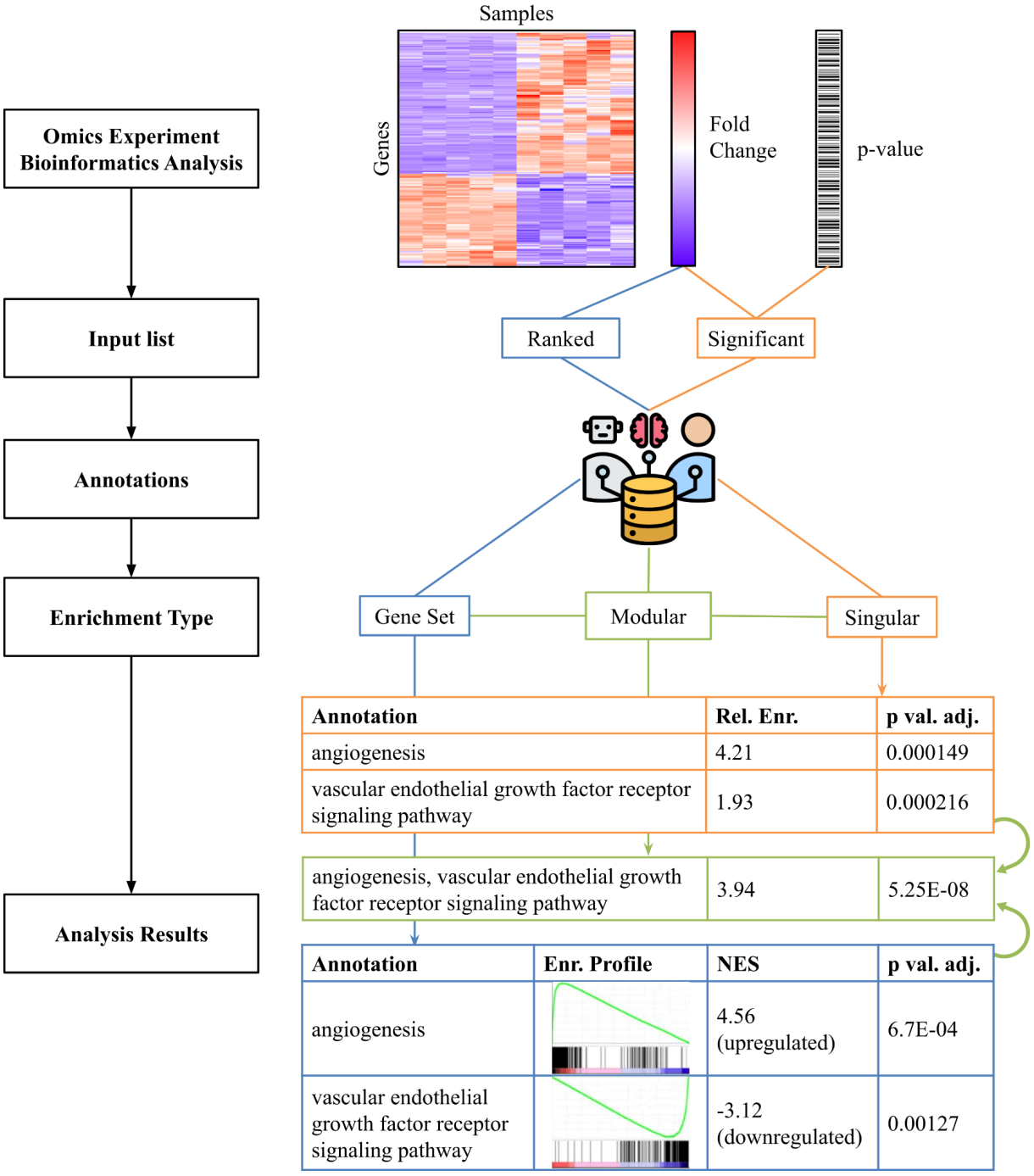


Figure 4. An schematic representation illustrating the differences of the functional enrichment methods. Image contains a draw from Flaticon.com and adapted GSEA plots from [93].

1.4.4. Functional enrichment analysis of regulatory elements

The functional implications of regulatory elements, unlike most genes, are not extensively understood beyond their inherent roles. These elements participate in complex regulation networks where miRNAs and TFs influence one or multiple genes. Conversely, genes themselves can harbour varying numbers of methylation sites, adding to the heterogeneity of these regulatory interactions. Consequently, directly assigning functional annotations to regulatory elements is challenging, however, enrichment analyses have been adopted to investigate the roles of these regulatory actors. The analysis of regulatory elements alone will not produce valuable information because, generally, the annotation databases merely associate them with their basic role, e.g. gene expression regulation. Thus the traditional approach consists in inferring their function by means of their target genes.

The conventional method manifests two challenges that have been discussed in the academic literature. Firstly, there is a lack of dedicated databases that directly annotate the downstream functional effects of regulatory elements. Secondly, the need of using target genes as an indirect measure poses a significant statistical concern in the context of enrichment analysis. This arises from the heterogeneous distribution of regulatory elements associated with a given gene, which violates the assumption of equal selection probability required by the central hypergeometric distribution. These two issues have been specifically observed in the analysis of human methylation sites and miRNAs, wherein the traditional approach has been shown to yield biassed outcomes. The assessment of these biases was conducted by analysing the results derived from numerous random lists of methylation probes and miRNAs.

Initially, it was observed that regardless of the methylation platform used, a clear correlation exists between the probability of a gene appearing as methylated or unmethylated and the number of CpGs within its sequence. As a result, it produces ubiquitous enriched annotations, associated with transcription, development and cell differentiation. These terms have genes annotated possessing a higher number of methylation probes and their enrichment p-value turns strongly correlated with the mean number of probes [94]. CpGs are frequently found near gene promoter regions and their methylation pattern regulates gene expression processes. Since these regions are more likely to be controlled by methylation, it is hypothesised that

genes with a large CpGs population will be linked to the aforementioned processes [95,96].

The evaluation of traditional miRNA SEA in humans, points to skewed results in biological processes that exhibit a close association with cell cycle and cancer biology [97,98]. Such findings could be expected, considering that a significant proportion of identified miRNA targets are implicated in cancer-related pathways. This observation raises the possibility that the observed bias may arise from an inherent imbalance in the studies conducted or from the fundamental roles played by miRNAs themselves in cancer pathogenesis [99].

Currently, various alternative methods have been proposed to overcome the limitations of traditional SEA in methylation sites and miRNAs vary. For methylation data, some researchers adapted the GOseq method [100] focused on the Wallenius statistics. This approach involves fitting a noncentral hypergeometric distribution to evaluate elements that have distinct probabilities of selection. GOseq was developed to mitigate the bias effect in the SEA of genes that are identified as differentially expressed as a function of the transcript length. GOseq tests the annotations weighted by the average transcript length of their genes associated and their differential expression status. In the context of methylation data, this approach weights and tests annotations based on the number of CpG probes associated with each gene. This procedure is implemented in the Bioconductor R package missMethyl [101,102]. Another proposed solutions for SEA and GSEA in methylGSA [102], they make use of the p-values derived from differential methylation analyses applying a meta-analysis approach based on the robust rank aggregation and a logistic regression model. Additionally, the ebGSEA R's package tests ranked genes according to the overall methylation level, by means of all the probes methylation metrics, the M or β values, within each gene [103].

Currently, the prevailing strategy for conducting miRNAs SEA is to use annotations directly associated with them. This can be achieved by converting the annotation gene sets to the set of unique miRNAs that target the annotation genes [97]. Alternatively, dedicated databases focused on miRNA functional annotations can be employed, such as Tool for miRNA Set Analysis (TAM) [104], Human miRNA Disease Database (HMDD) [105] and the Mammalian ncRNA-Disease Repository (MNDR) [106]. TAM includes an enrichment analysis functionality

and its database is also integrated in the enrichment tools miRNet [107] and miEAA [108]. The latter tool also includes gene-based annotations after transforming them to miRNAs sets. This thesis aims to further explore the existing options in miRNAs SEA and propose an alternative approach.

2. Objectives

The current accessibility of sequencing technologies has led to a growing trend in integrating diverse omics data in the study of biological systems. Typically, the analysis of omics data generates lists of candidate biomarkers ranging from dozens to hundreds, resulting in complex outcomes that are challenging to interpret. In order to extract the underlying biological knowledge embedded within these lists, functional enrichment analyses have been developed. While these analyses are well-established for genes and proteins, there is a scarcity of tools specifically designed for the analysis of lists comprising regulatory elements.

The general objective of this thesis is to develop a novel method for the functional annotations analysis of regulatory elements, as well as a complete reengineering of GeneCodis web tool in order to publish this new method. To fulfil this objective, the following research sub-objectives have been proposed in order to effectively attain our goal.

1) Review the **state of the art** in the **functional annotations analysis of regulatory elements.** Focused on assessing and comparing miRNAs databases and tools that perform miRNAs singular enrichment. Further tools able to analyse lists of candidate biomarkers such as methylation CpGs and transcription factors are also explored.

2) Development of a novel method for the functional annotations analysis of regulatory elements. Based on the singular enrichment analysis, the new method must include the statistical framework to avoid the traditional method limitations and be applicable to methylation CpGs, miRNAs and transcription factors.

3) Implementing the new method in a new version of the **Genecodis web tool.** Reengineering of GeneCodis, with a user-friendly web tool developed following an application programming interface to allow its usage programmatically. Update of its annotations database, co-annotations discovery algorithm and visualisation capabilities.

4) Assessment of bias in the singular **enrichment analysis** of **transcription factors** target

genes. Evaluate the traditional approach for singular enrichment analysis of transcription factors target genes via null simulation hypothesis and explore the ability of the previously proposed method to overcome the bias.

3. MicroRNAs Enrichment Analysis

3.1. MicroRNAs databases

MicroRNAs databases typically specialise in three key types of information: their intrinsic nature, their associated targets genes and functional or disease annotations.

One of the main long standing databases is **miRBase** [109]. It is responsible for assigning official miRNA gene names and holds a high-quality encyclopaedia with the genomic sequence covering 271 organisms, totalling 38589 hairpin precursors and 48860 mature forms. Recently, they incorporated functional annotations to 12519 miRNA entries by applying a text-mining approach and further extended with current Gene Ontology terms. This database is the reference in the miRNAs field however Alles et al. discovered that many entries actually report pieces of other small RNAs types [26]. They reanalysed 30000 samples from diverse sources of human small RNA sequencing data and concluded that there are at least 2300 true mature miRNAs of which only 1115 are annotated as such in mirBase, reducing greatly the false positives. An alternative significant resource in the field is **mirGeneDB** [110], which stands out for its focus on including only experimentally validated miRNAs that have undergone a rigorous curation process. This database encompasses a wide range of 75 metazoan organisms, 16670 miRNAs from 1549 families, including different tissue expression matrixes.

The databases for **miRNA targets** are typically created using either prediction algorithms or experimental validation. Namely, miRNAs target prediction algorithms mainly infer the interaction relying on sequence analysis, considering several properties of the interaction between miRNA and mRNA molecules. Diverse reviews assess the different approaches available in a practical sense [111–113]. Nevertheless, in crosslink immunoprecipitation (CLIP) analyses it was found that the miRNA binding events have insignificant functional effects [114,115]. This fact exposes that miRNA - mRNA bindings does not necessarily result exclusively in target gene underexpression, which must be considered when performing miRNAs target

functional analysis. Subsequently, a recommendation is to employ miRNA target gene prediction algorithms that takes into account gene expression data or that filters out genes not expressed in the studied tissue [116]. Important target **prediction tools** are TarPmir [117], TargetScan [114], MirTarget [116] and DIANA microT-CDS [118]. Since the algorithms are trained on different miRNA target interaction features, they lead to different target predictions. Consequently, many authors suggest that combining the results of multiple algorithms improves the target prediction accuracy [119]. They claim that the union set of the targets predicted by different algorithms has greater benefits because it reduces the risk of missing relevant targets or identifying false positives. Additionally, the intersection sets help to increase the confidence of the predicted targets by providing more supporting evidence from different sources [120].

In the context of **experimentally validated miRNA target** interactions there are two principal databases: miRTarBase and DIANA-TarBase. miRTarBase [121] is an extensive and frequently updated database built from manually curating large collections of articles and CLIP-seq data with extra support from several external databases. This produces a database of 37 organisms with 4630 miRNAs that compris millions of interactions with 27172 target genes. It also incorporates nucleotide polymorphisms and disease variants associated with the miRNA binding efficiency, both at the miRNA level and at the target mRNA 3' untranslated region (UTR). Furthermore, it incorporates the expression patterns of miRNAs across various biological samples such as extracellular vesicles, blood, and multiple tissues, encompassing both exosomal miRNAs and tissue-specific miRNAs. The miRTarBase integrative implementation allows to highlight miRNA functions and identification of potential biomarkers. Targets are classified whether the evidence is weak or strong with respect to the experimental technique used in the validation. Strong evidence are reporter assays, western blots, and qRT-PCRs, meanwhile, high throughput techniques are considered weak evidence. TarBase [122], which is a part of the DIANA suite, has undergone eight updates and the latest version was released in 2017. It is a comprehensive repository of over 670000 unique miRNA target pairs that have been manually curated from nearly 1200 research publications and over 350 high-throughput datasets. The targets in TarBase are categorised similarly to miRTarBase, into low- and high-throughput techniques, with some of the most commonly used methods

being reporter assays, Western blotting, qPCR, proteomics, biotin miRNA tagging, RNA sequencing data, and microarrays.

Regarding the **miRNAs functional annotations** databases one of the first developed is miRCancer [123] which links miRNAs with different cancers types, however, its web portal is not accessible at this thesis writing date. Other databases also associate miRNAs with other complex diseases such as the Human miRNA Disease Database [105], which solely features experimentally validated interactions. A broader range of organisms are covered in the Mammal NcRNA-Disease Repository, offering an extensive collection of associations between various ncRNAs and diseases. While most miRNA databases are focused on animals, there are also databases dedicated to plants. For instance, the Plant miRNA Encyclopedia [124] provides a comprehensive and current list of plant miRNAs and it integrates information from other plant databases. Another important database is built for the TAM enrichment tool which includes annotations of the miRNAs biological nature and manually curated functions.

There are diverse databases that also include information of miRNAs although it is not their main topic. Many of these are included in RNAcentral, whose consortium effort aims to generate a complete sequence-based harmonised database of all ncRNAs types from diverse organisms [125].

3.2. Methods for functional enrichment analysis in microRNAs

After conducting an extensive survey of the current tool for miRNAs functional enrichment analysis, this section aims to highlight the different existing approaches. These will be briefly described to provide an overview of the state-of-the-art, emphasising the advantages and limitations of each method, with a specific focus on bias handling.

3.2.1. Target based annotations

In the **traditional approach**, miRNAs are first transformed to the set of genes targeted by them to later apply the enrichment statistics over gene-based databases. As mentioned in the introduction, this is reported to produce **biassed results** in the human miRNAs SEA when the annotations are tested with the central hypergeometric distribution. This approach results in cancer and cell cycle annotations constantly enriched [97,98]. Despite being clearly stated and evidenced, there are plenty of contemporary published studies that ignore it and consider them as useful results [126–129]. This can be explained due to the main interest in gene-based databases whose prominent position offer exhaustive annotation fields and cover several organisms. Additionally, current tools like MIENTURNET [130] and NcPath [127], whose main focus is to display the experimentally validated gene regulation networks of miRNAs and other ncRNAs, offer enrichment analysis functionalities using this traditional approach without further bias control.

Another significant limitation of this approach, and the following target based, is that they depend on the evidence source of the relations between miRNA and targets. These links can be obtained from predictive algorithms or experimentally validated databases. Depending on the research objectives, it is crucial to decide in advance which source best suits the research needs. For instance, using predictive algorithms can provide a broader range of target genes and consequently more functional annotations. On the other hand, relying solely on targets supported by direct empirical evidence can help focus the research on well-established gene regulation mechanisms. The choice between these approaches depends on the specific goals and requirements of the study.

3.2.2. Targets based annotations filtered to tissue specificity

A variation of the traditional approach involves tailoring the annotation datasets to exclusively the set of **target genes expressed in a particular tissue**, as demonstrated by the miTALOS tool [131]. It is conceived towards miRNAs researchers in wet lab settings who aim to identify a set of candidate miRNAs for experiments in specific tissues and cell lines. By

considering tissue-specific target gene expression, this approach partially mitigates the issue of different probabilities of gene selection by eliminating poorly or non-expressed target genes. However, the effectiveness of this approach heavily relies on the availability of high-quality tissue expression data that matches the specific study requirements.

3.2.3. Targets based annotations and empirical sampling

The calculation of **empirical p-values** through the traditional approach is a proposed solution by Bleazard et al. against the miRNAs target based enrichment biassed results [98]. Comparing the overlap of an original miRNAs target gene list associated with a given annotation, they calculate the empirical p-value as the proportion of random miRNAs lists of the original size that produce an equal or greater overlap. The number of random lists is a parameter that directly affects the accuracy of the empirical p-value, especially when numerous annotations are tested and a fine granularity is required to distinguish them. Actually, the minimum considered by Bleazard et al. is one million, which in the analysis of hundreds of miRNAs their implementation might take more than 17 hours in an Intel i7-3820 processor before obtaining an empirical p-value as observed by BUFET [132] tool authors. BUFET is a deeply optimised Python script to perform an empirical sampling like Bleazard et al.

Considering the computational costs of calculating empirical p-values, web tools that anticipate high user demand often reduce, under the proposed minimum, the number of random lists of miRNAs analysed to optimise performance. For example DIANA miRPath [133] article states that it includes the same approach, though it is no longer selectable in their website by the time of this thesis writing, and miRNet only tests a thousand of random lists [134]. Similarly, established tools like miRSystem calculate the empirical p-value as the proportion of p-values obtained with the hypergeometric test lower than a null baseline probability calculated with only a thousand random lists of miRNAs lists which are questionably allowed to be of different sizes.

3.2.4. Transforming target annotations database to miRNAs

A third method is to **transform gene sets annotations** databases **to miRNAs sets**, in order to test each functional term in the new miRNA-based database (Figure 5). This alternative, proposed by Godard and van Eyll [97], ensures not overcounting miRNAs, because each miRNAs it is only represented once in a given annotation whatever the number of its annotated target genes are. They tested the methods with several published miRNAs and found that often results are more specific to the biology under study than the target based approach. Gene-based databases such as KEGG and GO are transformed to be analysed within the miEAA tool. Nonetheless, it was found that this database transformation causes many annotations to be jointly enriched because they share a significant amount of miRNAs. As a solution to this limitation they proposed coupling this method with MEA algorithms.

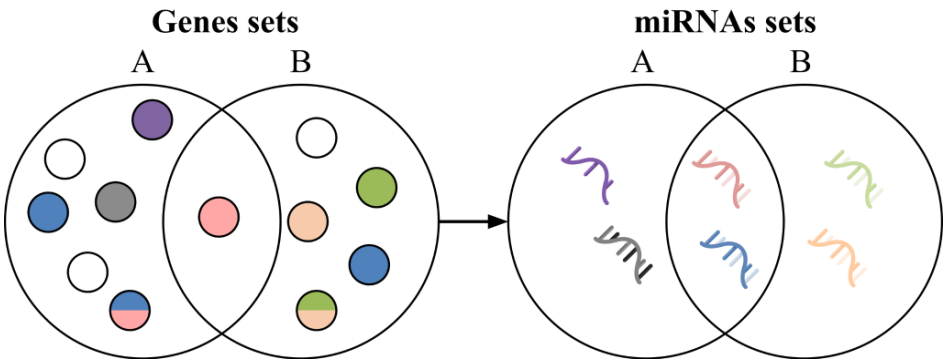


Figure 5. Illustration of the gene-based database (left) transformation to miRNA-based (right). The colours determine the miRNAs gene targeting association.

3.2.5. MicroRNAs based annotations

In this second approach, miRNAs are linked and tested against functional terms directly associated by different **expert curation procedures**. This was first implemented in TAM web tool [104,135]. TAM built an annotation database of the miRNAs families, genome clusters, tissues, diseases and functions by a combination of external sources and manual literature expert curation. TAM annotates around 359 human hairpin miRNAs and 1239 different annotations. They annotate miRNAs following an harmonised vocabulary with the Gene Ontology and the Human Phenotype Ontology [136]. Noteworthy, TAM authors, aware of the

unbalanced study of miRNAs in humans, offer an option to mask cancer terms before testing.

Similar to TAM, the Gene Ontology Consortium has also dedicated efforts to annotate miRNAs. They observed that miRNAs were predominantly associated with annotations related to development regulation and cellular processes. Aiming to address the underrepresentation of miRNAs, the consortium established guidelines for the functional annotation of miRNAs [137] and presently, GO provides over 4400 annotations of approximately 500 miRNAs, hairpins and matures from human, mouse, and rat along with 2400 experimentally validated miRNA target interactions [138].

Other important databases commonly used in miRNA functional enrichment analyses, previously mentioned, are MNDR and HMDD. These, along with direct annotations form TAM and GO are included in the miEAA web tool.

3.4. Summary of reviewed tools

The depicted methods are not exclusive, which means different approaches can be found in the same tool to help cover each research needs. In the following table we surveyed different tools, dedicated to the analysis of miRNAs, highlighting their method and the software implementation.

Tool	Methods	Implementation	Update year
DIANA miRPath	Targets annotations and empirical sampling	Web	2015
miRSystem	Targets annotations and empirical sampling	Web	2016
miTALOS	Targets annotations filtered to tissue specificity	Web	2016
TAM	miRNAs annotations	Web	2018
MIENTURNET	Targets annotations	Web	2019
miRNet	Targets and miRNAs annotations and empirical sampling	Web and R	2020
miEAA	miRNAs annotations and transformed database	Web and Python	2020
BUFET	Targets annotation and empirical sampling	Python and C++	2020
NcPath	Targets annotation	Web	2023

Table 1. MicroRNAs functional enrichment analysis tools.

We wrote a detailed review about miRNAs enrichment analysis tool that can be found here [139].

4. New Regulatory Elements Functional Annotations Analysis Tool (GeneCodis4)

This section is dedicated to the main objectives of this thesis, the development of a **new method** for the functional analysis of regulatory elements and implementing it in a **GeneCodis** new version. GeneCodis is a widely used functional analysis tool that was originally published for modular enrichment analysis [91]. This software has been updated several times, increasing and updating its functionalities to meet the latest needs and usage demands of the research community [140,141]. GeneCodis 4 retains its distinctive capability of conducting modular enrichment analysis and, besides genes and proteins, extends its functionality to analyse regulatory elements such as methylation CpG sites, miRNAs and transcription factors. Taking into consideration the years that have passed since version 3, the new update not only consisted in updating the knowledge database and adapting the novel methods but also a complete **reengineering** of the software implementation together with a migration to a new server.

4.1. Novel method for regulatory elements and statistical methods for functional annotations analysis available

As mentioned in the introduction, the analysis of functional annotations involves the use of statistical tests to identify enriched biological functions in a given list of candidate biomarkers. Since our objective is to integrate the new approach into GeneCodis, we have focused on functional enrichment methods that rely on the Fisher's Exact test, which is commonly employed in various gene enrichment tools. However, during our review of enrichment tools for miRNAs, we have observed that when applying the Fisher's Exact test to the analysis of regulatory elements through their target genes, such as in the case of methylation data and miRNAs, it leads to biassed results towards specific biological processes, particularly those associated with regulation, cell cycle, and cancer processes. This bias arises from the heterogeneous distribution of connections between genes and regulatory elements, causing each target gene to have a different probability of being selected for the

enrichment test. This breaks the assumption of equal sampling probability intrinsic to the Fisher's Exact test. Consequently, genes with a higher number of regulatory factors have a greater probability of being tested, and these target genes are often functionally related, resulting in the pervasive enrichment of specific terms.

Interestingly, during our exploration of unbiased enrichment analysis methods for miRNAs, we discovered that none of the existing tools incorporated the **Wallenius approach**, which has been successfully applied in the analysis of CpG methylation probes in missMethyl [102] R's package publication. While similar strategies were employed to address biassed results in both methylation data and miRNAs associated genes, different authors proposed distinct solutions. It is worth noting that the efficacy of the Wallenius approach in miRNA SEA has not been specifically evaluated in the literature. We did come across an article that mentioned the use of this approach, but the authors did not provide detailed information on its application or justify the choice of this statistical method [142]. The Wallenius noncentral hypergeometric distribution is an adaptation of the central hypergeometric distribution used in Fisher's exact test. However, unlike the standard hypergeometric distribution, the Wallenius distribution takes into account the fact that **items are selected with different probabilities**. Based on our initial hypothesis, we speculate that the bias correction based on the number of CpG probes per gene could be analogous to the correction based on the number of miRNAs and transcription factors that regulate each target gene.

The Wallenius distribution probability mass function, that determines the probability of an annotation being related to our list of candidate biomarkers is calculated as follows [143]. First the functional annotation target gene set is extracted and their frequency in our list and the annotation database is computed. The total number of targets genes regulated by the input regulatory elements is defined as *n,* the targets that are associated with the tested annotation is denoted by *x*, the total number of target genes annotated in the tested functional term is reflected as *N,* the total number of target genes considered in the regulation network, the background set or universe, is expressed as *M*, and finally, ω represents the annotation odd ratio of being present in a random list of target genes. This defines the probability mass function equation of Wallenius' as:

$$p(x, M, n, N, \omega) = \binom{n}{x}\binom{M-n}{N-x}\int_0^1 (1 - t^{\omega/D})^x (1 - t^{1/D})^{N-x} dt \quad (1)$$

Where D is defined as:

$$D = \omega(n - x) + ((M - n) - (N - x)) \quad (2)$$

The calculation of ω is a crucial step in this approach and it varies in the SEA softwares that implements the Wallenius statistics. The first method proposed by GOseq weights each gene with a **power weighting function** by means of a logistic model that is trained on the transcript lengths and predicts the gene weight with a binary vector indicating 1 if the gene is differentially expressed and 0 otherwise. The GOseq method was adapted in missMethyl which trains the model based on the number of CpGs per gene and the binary vector with respect to the gene being considered differentially methylated. Thus, our proposal is to adopt the power weighting function for miRNAs to eventually create a logistic model trained on the number of regulatory elements associated to each target gene and fitting a binary vector, where 1 implies that the gene is targeted by at least one input regulator and 0 if not (Figure 6).

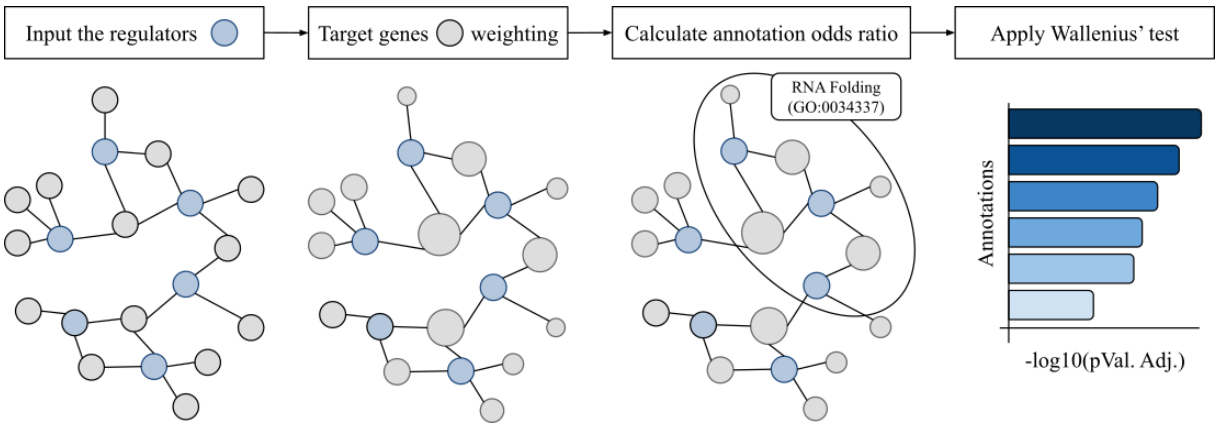


Figure 6. General workflow of the Wallenius approach. First a set of regulatory elements, such as miRNAs are input and their target genes are extracted. The power weighting function is applied to target genes. Next, it is calculated ω, as the fraction between the average weight of the target genes within the functional category i.e. RNA folding, and the average weight of the target genes outside of it.

In addition to the Fisher's Exact test and the novel Wallenius approach, GeneCodis also offers two other distinct approaches for the analysis of miRNAs in its functional enrichment analysis. These approaches involve directly testing miRNAs curated annotations and gene-level annotations transformed into miRNA-based annotations. Later, in GeneCodis we incorporated the Wallenius approach for transcription factors SEA (Section 5). As a result, GeneCodis provides a range of **statistical methods** for the analysis of functional annotations for genes and proteins, CpG probes, miRNAs and TFs:

A. Hypergeometric central test of genes/proteins or target genes annotations
B. Wallenius noncentral hypergeometric test of the associated target genes
C. Hypergeometric central test of miRNAs-based annotations
D. Hypergeometric central test of the transformed gene annotations to miRNAs-based

Our recommendation and default settings apply the hypergeometric test to analyse genes or proteins. The Wallenius noncentral hypergeometric test has been specially implemented to avoid bias in gene selection is set for CpGs and TFs, though is also available for miRNAs. Finally, the miRNAs analysis is done by default with the hypergeometric central test but the databases available are either miRNAs-based or gene-based transformed. Both the obtained Wallenius noncentral hypergeometric and the Hypergeometric central tests p-values are corrected for multiple testing by the Benjamini and Hochberg False Discovery Rate (FDR) method [144].

All the GeneCodis analysis options can be configured in the advanced settings in the website, where it is also possible to activate the comparative analysis of the intersecting and exclusive sets of two input lists. It is also possible to define the **scope of the universe** or provide a customised background reference. It should be noted that setting up the universe has significant implications when conducting SEA [145]. Here, the common issue can be illustrated by recognising that the epigenome, transcriptome and proteome of a given cell type or tissue is highly specific. In experimental designs that focus on a specific system and compare two conditions, the functional enrichment analysis of the resulting biomarkers is likely to yield

annotations that are limited to the biology of the studied system. For example, if a cell culture is treated with specific inhibitor substances, the enrichment analysis of dysregulated genes is likely to produce terms specific to the inhibited pathway. This highlights the importance of establishing a custom background gene list to ensure that the functional analysis provides novel insights. A custom universe should encompass all the genes that could be robustly measured in the experiment and are relevant to the biological system, based on its transcriptome expression profile (Figure 7). This consideration is important for the case of regulatory elements as remarked by miTALOS tool.

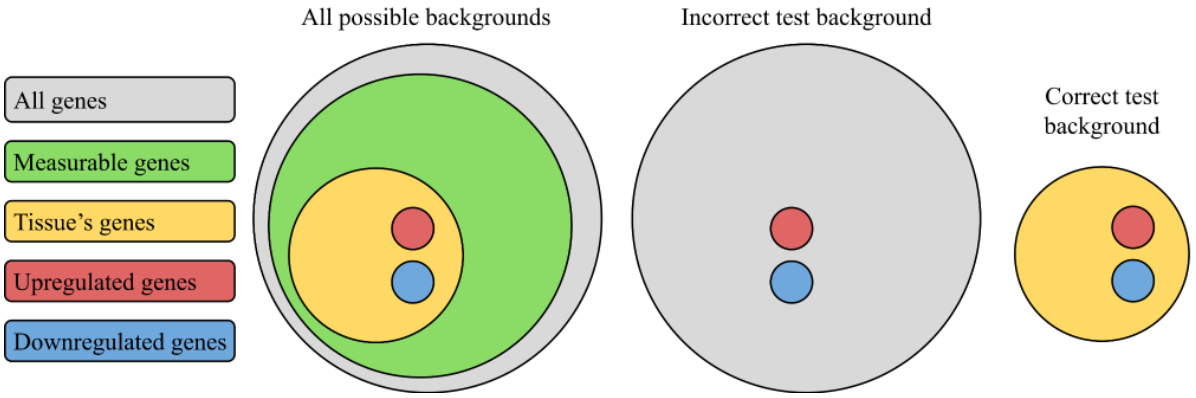


Figure 7. Graphical representation of the adequate background universe setting.

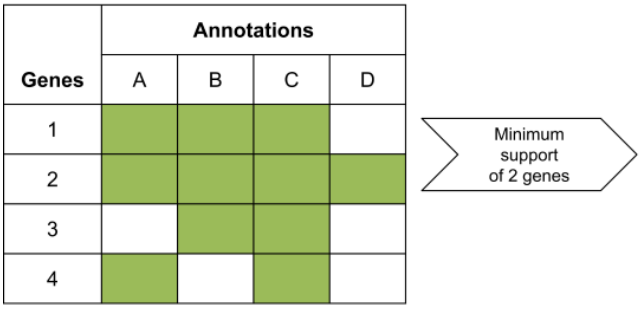
Nevertheless, in GeneCodis, the default setting for the universe is defined as the set of genes, miRNAs, or target genes that have at least one annotation in each selected database. Alternatively, there is an option to use the entire set of genes present in the database as the universe scope, although this is less recommended.

Another metric available in GeneCodis is the relative enrichment score which measures the proportion between the number of genes found of an annotation in the input divided by the size of the input and how many genes are under that annotation divided by all the genes in the database. High numbers imply a better relative enrichment.

4.2. Co-annotation algorithm

In previous versions, GeneCodis employed the apriori algorithm to identify closed co-occurring biological annotations. However, in the latest update, a more advanced approach based on the Frequent Pattern algorithms has been implemented. These algorithms leverage tree structures, enabling a more efficient discovery of concurrent biological annotations [146].

GeneCodis incorporates two distinct **Frequent Pattern algorithms** in its analysis process, FPgrowth and FPmax. These algorithms serve the purpose of identifying **co-annotations**, although they differ in the types of co-annotations they report. Specifically, **FPgrowth** focuses on finding closed co-annotations, which are individual or combinations of annotations that share a minimum number of input elements, the minimum co-annotation support. The **FPmax** algorithm implementation provides a novelty to the GeneCodis tool because it identifies maximal closed co-annotations, that are a superset of other closed co-annotations. In other words, maximal closed co-annotations represent the largest possible combinations of annotations within the dataset (Figure 8).



Co-annotation	Genes	N. Genes	Type
A, B, C	1, 2	2	Maximal
A, B	1, 2	2	Closed
A, C	1, 2, 4	3	Closed
C, B	1, 2, 3	3	Closed
A	1, 2, 4	3	Closed
B	1, 2, 3	3	Closed
C	1, 2, 3, 4	4	Closed
D	2	1	Infrequent

Figure 8. Example of a co-annotation analysis to illustrate the difference between frequent closed co-annotations and maximal closed co-annotations. Note that, in order to be a maximal co-annotation, all its annotation subcombinations must be closed.

While FPmax is generally faster and yields non-redundant results, it may be more challenging to interpret when numerous terms are co-annotated. In such cases, the output from FPmax

may be too complex to extract meaningful insights. Nonetheless, the availability of both algorithms within GeneCodis provides users with flexibility and options to tailor their analysis approach to their specific needs.

In GeneCodis the default algorithm is the FPgrowth approach. Moreover, unless fewer are introduced, both MEA algorithms need annotations to share at least three of the input elements to be considered a co-annotation. It is important to note that increasing the number of annotation sources and input elements while reducing the minimum support, could result in a significant growth in the number of co-annotations discovered. Subsequently, the complexity of the results widens along with the computation time required to generate and assess them. To ensure **computational efficiency**, GeneCodis employs certain strategies. Firstly, the co-annotation analysis is limited to input lists of up to a thousand elements and a maximum of two different sources of annotations can be used per analysis. Additionally, If the analysis exceeds a time threshold of 5 minutes, the minimum number of input elements required for a co-annotation is increased by one. This helps reduce calculation time and minimise the inclusion of noisy co-annotations. Conversely, if no co-annotation is found, which can occur if the user sets a high minimum support, the minimum support value is automatically decreased by one. These restrictions and helper functionalities ensure that the co-annotation discovery process remains manageable.

4.3. Databases included

GeneCodis 4 upholds information of genes and proteins for **14 different model species and 20 different annotation resources**. MicroRNAs are available only for 5 of these, transcription factors from *Mus musculus* and *Homo sapiens,* and CpG sites only for human.

4.3.1. Input data

GeneCodis 4 utilises a comprehensive gene catalogue that includes a combination of the most widely used gene databases, namely **NCBI** and **Ensembl**. To achieve this, the latest genome

annotation files of Ensembl, in GTF (Gene Transfer Format), are procured for each organism and subsequently aligned with the NCBI gene repository to ensure consistency and accuracy. Furthermore, GeneCodis offers various nomenclatures to cater to different user preferences, including HUGO official gene symbols and Uniprot, along with all the synonyms found in NCBI annotation. This enhances the flexibility and usability of the platform, allowing users to conduct their gene analysis in a more personalised and efficient manner. Next, miRNAs identifiers are matched to the **miRBase** [109] nomenclature and their target genes are extracted from **miRTarBase** [147] limit to only the links which are evidenced by techniques such as western blot, qRT-PCR and reporter assay. Likewise, to ensure a comprehensive and reliable SEA of TF target genes, GeneCodis incorporates data from **DoRothEA** [148], specifically, from the three highest confidence scoring criteria, A, B, and C. These TF-gene regulons have been incorporated by meticulous literature review experts, which additionally are supported by at least two curated databases and ChIP-seq interactions plus an extra level of evidence. Human genome CpGs included are from the Illumina's **MethylationEPIC** microarray.

4.3.2. Annotations sources

Twenty different collections grouped into four annotation categories can be found in GeneCodis 4. The categories are: functional databases, which define biological processes and pathways, regulatory annotations include the regulons of miRTarBase and DoRothEA and miRNAs specific databases, drugs related with genes and finally, a set of phenotypes, principally diseases, clinical signs and symptoms.

In the **functional** category the following databases are gathered: KEGG Pathways [149], Gene Ontology divided in its three main categories: Biological Process, Molecular Function and Cellular Component [150]; Panther Pathways [89], Mouse Genome Informatics database [151], BioPlanet [152], Reactome [153] and WikiPathways [154].

The **regulatory** category, apart from the two aforementioned curated interactomes, encompass a specialised subcategory that covers miRNAs-based annotations: TAM, HMDD and MNDR. Originally, all annotations from TAM and HMDD exclusively point to precursor miRNAs,

which is done on purpose by their authors during the curation process, in contraposition to MNDR. Concretely, TAM transforms mature identifiers of miRNAs to precursors before applying the enrichment analysis. In GeneCodis we decided to facilitate the analysis of miRNAs that TAM and HMDD annotations are also associated with both precursors and corresponding mature identifiers.

Next, gene and **chemical** pairs from the Pharmacogenomics Knowledgebase (PharmGKB) [155], the Comparative Toxicogenomics Database (CTD) [156] and the Library of Integrated Network-Based Cellular Signatures (LINCS) [157] are added to the GeneCodis 4 database.

Lastly included are gene to **phenotype** relations from the Online Mendelian Inheritance in Man (OMIM) [158] catalogue, the Human Phenotype Ontology (HPO) [136] and DisGeNET repository [159].

4.4. Web development and API implementation

Web applications, including many software tools, consist of two fundamental components: the front-end responsible for displaying and facilitating user interaction, and the back-end responsible for accessing the databases and processing information. It is essential to maintain a clear separation between these layers. In web tools, the client-side bears the burden of visualisations, making it crucial to avoid overloading it. By optimising the distribution of processing tasks between the front-end and back-end, the performance and user experience of the web tool can be enhanced.

GeneCodis 4's back-end is developed using **Python** 3.8 and the **Flask** microframework, which enables the creation of a simple yet robust **application programming interface** (API). The API is deployed with the Gunicorn library which assists handling quick requests and responses and additionally sends the core analysis to a job scheduling system called Slurm [160]. It permits creating a job queue in case the server demand is too high and liberates the API endpoints to receive constant analyses. Nevertheless, to avoid any abusive use of the tool that might overload the server we incorporated the Flask limiter extension to our application and

the number of analyses per user is limited to ten per minute. The web page is sent to the client-side using an NGINX server, which also acts as a front-end reverse proxy for the API. This software implementation and our collaboration with Dr. Hackenberg allows a direct GeneCodis query from the **SRNAToolBox-sRNAbench** [161]. SRNAToolBox is a valuable resource to perform sRNA expression profiling on next-generation sequencing data which can yield a set of candidate miRNAs that can be functionally described using GeneCodis4. This integrated approach provides a convenient and streamlined process for studying the regulatory role of miRNAs in various biological systems.

The database is built with PostgreSQL 12, a powerful open-source database management system accessed using the psycopg2 Python module. To implement the frequent pattern algorithms, GeneCodis4 employs the MLxtend library [162], while the probability weighting function is obtained using pyGAM [163]. The statistics methods are built-in functions of the scipy [164], pandas [165], numpy [166] and statsmodels libraries.

GeneCodis4's front-end is built using plain HTML, JavaScript, and CSS. To render HTML pages, GeneCodis4 uses EJS, a powerful JavaScript templating language. The interactive table within the results report is displayed using the DataTables plugin, a highly customizable jQuery library that enables the manipulation of HTML tables. The plots generated depend on two powerful JavaScript libraries, D3.js and jQuery. GeneCodis4's CSS mainly derives from the open source Bulma framework, a responsive and modern CSS framework widely used in web development.

GeneCodis4 is deployed on an Ubuntu Server 18 operating system, running on a computer with 252 GB of RAM memory and an Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz microprocessor.

GeneCodis uses only **open source** technologies thus, inherently, itself is also open source and available in GitHub (https://github.com/GENyO-BioInformatics/GeneCodis). The repository includes two wrappers in R and Python to exploit the API, whose complete documentation is available in Postman (https://documenter.getpostman.com/view/21704552/2s7YfPfuPN).

4.5. Results report

Results page is divided in two columns, a navigation menu across the different annotation databases that the user indicates and the results report sections besides it. Some screenshots are displayed in the final section of this chapter for better understanding.

The first results consist in a quality control of the input query list and the universe list, if provided, to capture unrecognised input elements. It is done at two different levels, a general control over the available input in GeneCodis and a second one to notice input elements that are not in the investigated database. Often, input elements might contain strange characters that could cause the software not to identify them in our database. If an opted annotation has cero introduced genes, miRNAs or targets, it will be also noted here. Then the number query entities in the annotation database, the number that are not present and the total number of genes or miRNAs in the universe is also shown. In case of applying MEA, in this section the co-annotation minimum support applied is also informed.

GeneCodis generates an interactive table per annotation analysed in its results report, allowing users to explore up to the top 100 enriched terms in the website, displayed without any p-value cutoff. On top of that, the complete result table is available to download for more in-depth examination. In case that the user wants to download all the annotations resulting tables at once, a dedicated button is available at the results navigation menu.

Two downloadable visualisations are generated, namely a network diagram and a bar chart. In both, the user has the flexibility to determine the number of top enriched annotation terms to be displayed. The network diagram illustrates the relationships between the top enriched annotations and their associated genes or miRNAs. Annotations that share a larger number of genes tend to cluster together, thus aiding in the intuitive comprehension of the roles of the analysed elements. The size of each annotation node within the network is proportional to the adjusted p-value in -log10 scale, providing a visual indicator of the statistical significance. The network also allows to hide or show the gene nodes and the annotation label to generate a more comprehensive picture. The bars chart provides a more classical alternative view of the

enriched annotations, displaying their significance scores as a function of the bar width. The number of genes in the input that cause the enrichment is reflected in the colour intensity of both figures.

Finally, if miRNAs are analysed via the Wallenius approach a table that links them with their targets annotated in the selected databases is shown.

4.6. GeneCodis 4 analysis of arrhythmia miRNAs

The primary objective of the GeneCodis 4 update was to enable the analysis of miRNAs, which are the only entities that can be functionally characterised using all four implemented methods in GeneCodis. In this section, we will adopt a **tutorial-like** approach to analyse a list of miRNAs associated with arrhythmia. We will apply each method while discussing the advantages and limitations of each approach. The background universe used for all the analyses will be the default option, which is the annotated universe scope. The candidate miRNAs (miR-1, miR-133, miR-208a, miR-212, miR-328) are obtained from table one of a conducted literature revision that investigates cardiac excitability miRNAs [167] (Table 2).

Review miRNAs	Hairpin miRNAs	Mature miRNAs
miR-1	hsa-mir-1-1 hsa-mir-1-2	hsa-miR-1-3p hsa-miR-1-5p
miR-133	hsa-mir-133a-1 hsa-mir-133a-2 hsa-mir-133b	hsa-miR-133a-3p hsa-miR-133a-5p hsa-miR-133b
miR-208a	hsa-mir-208a	hsa-miR-208a-3p hsa-miR-208a-5p
miR-212	hsa-mir-212	hsa-miR-212-3p hsa-miR-212-5p
miR-328	hsa-mir-328	hsa-miR-328-3p hsa-miR-328-5p

Table 2. GeneCodis 4 use case input miRNAs.

As such, the miRNAs in the sourcing review are not written following a strict gene nomenclature, instead they are referred to the id root. Thus they must be converted to a proper

identification convention, for example the miRBase annotation. By means of manual search in miRBase prepending the human tag ("hsa-") we are able to extract the correct ids of the miRNAs.

Additionally to explore the different methods results, this use case also pretends to illustrate the different biological information fields available in GeneCodis. For that, in the forthcoming analyses one of each annotation category is selected: for the functional category GO biological process, in the regulatory the miRNAs-based MNDR, the chemicals from PharmGKB and finally the phenotypes of DisGeNET. Both hairpins and mature miRNAs are introduced as input because gene-based databases, such as GO, are annotating miRNAs in hairpin nomenclature and because MNDR distinguishes between precursors and mature miRNAs.

The first analysis is the **hypergeometric test of miRNAs direct annotations** using the **gene-based databases transformed** to miRNAs-level and the miRNAs **expert curated annotations** database. This is the default setting in GeneCodis when indicating miRNAs as input type. The base analysis settings are *Homo sapiens* and genes/proteins as input type. Changing it to miRNAs will cause the miRNAs-based annotations to become selectable and at the left bottom appear an extra functionality of GeneCodis, the miRNAs converter. This functionality allows the user to add or to convert the mature or precursor ids of the input miRNAs. Just above the orange "Launch Analysis" bottom the analysis can be named and provide an email to be written when the results are ready (Figure 9). The general aspect of the website front end might vary across different browsers and operating systems, precisely, this and the following screenshots are taken with Mozilla Firefox 112 in Ubuntu 20.

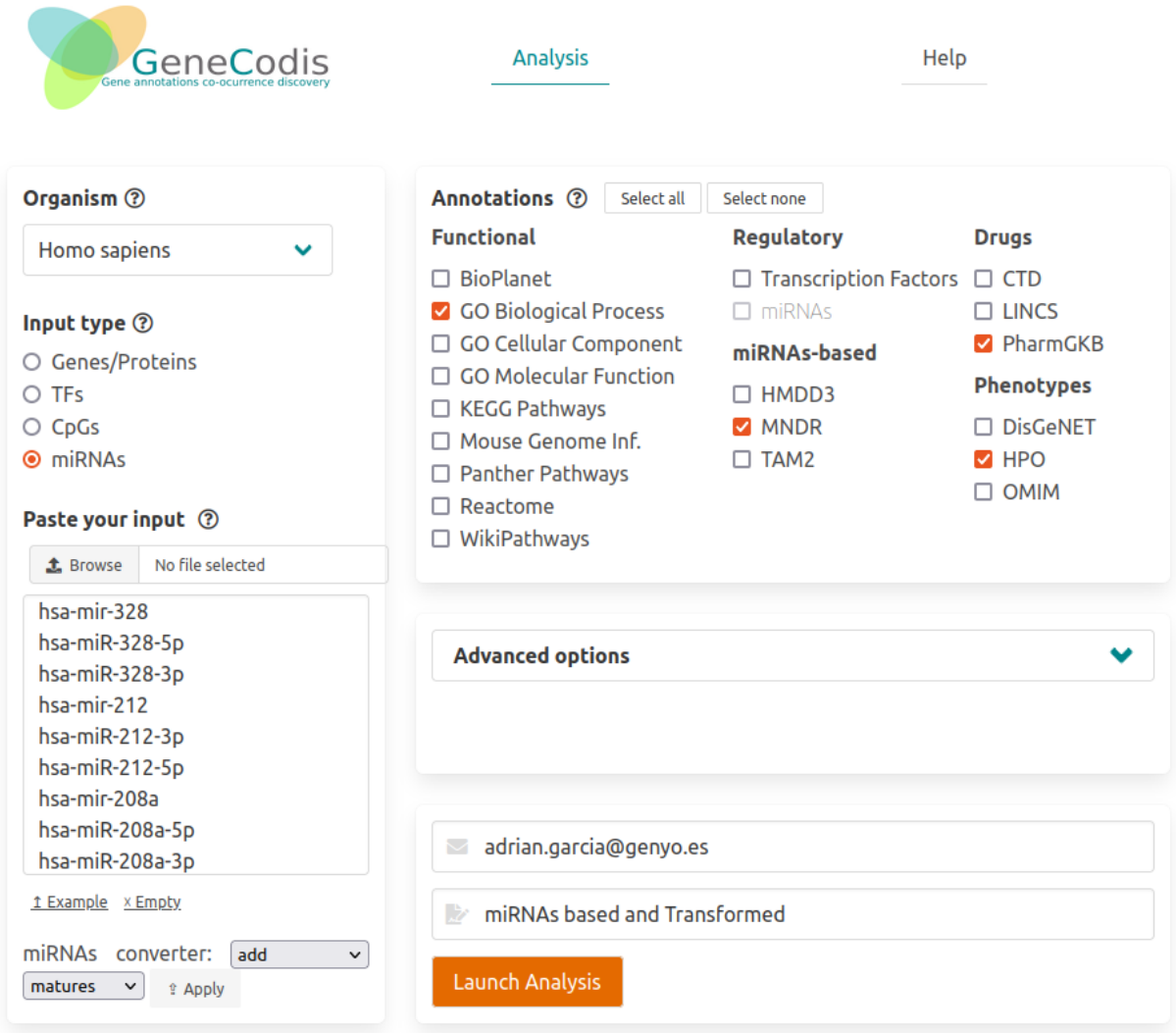


Figure 9. Screenshot of GeneCodis landing page with the form filled ready to analyse the set of miRNAs against the transformed to miRNAs and miRNAs-based databases.

As previously mentioned, the resulting page is divided in two columns. The left one contains two navigation menus, the top includes the inputs database results tabs that can be selected for displaying, the second helps you to navigate through the shown results report, which is in the right column, and contains different options to customise the gene-annotation network and the bars chart (Figure 10).

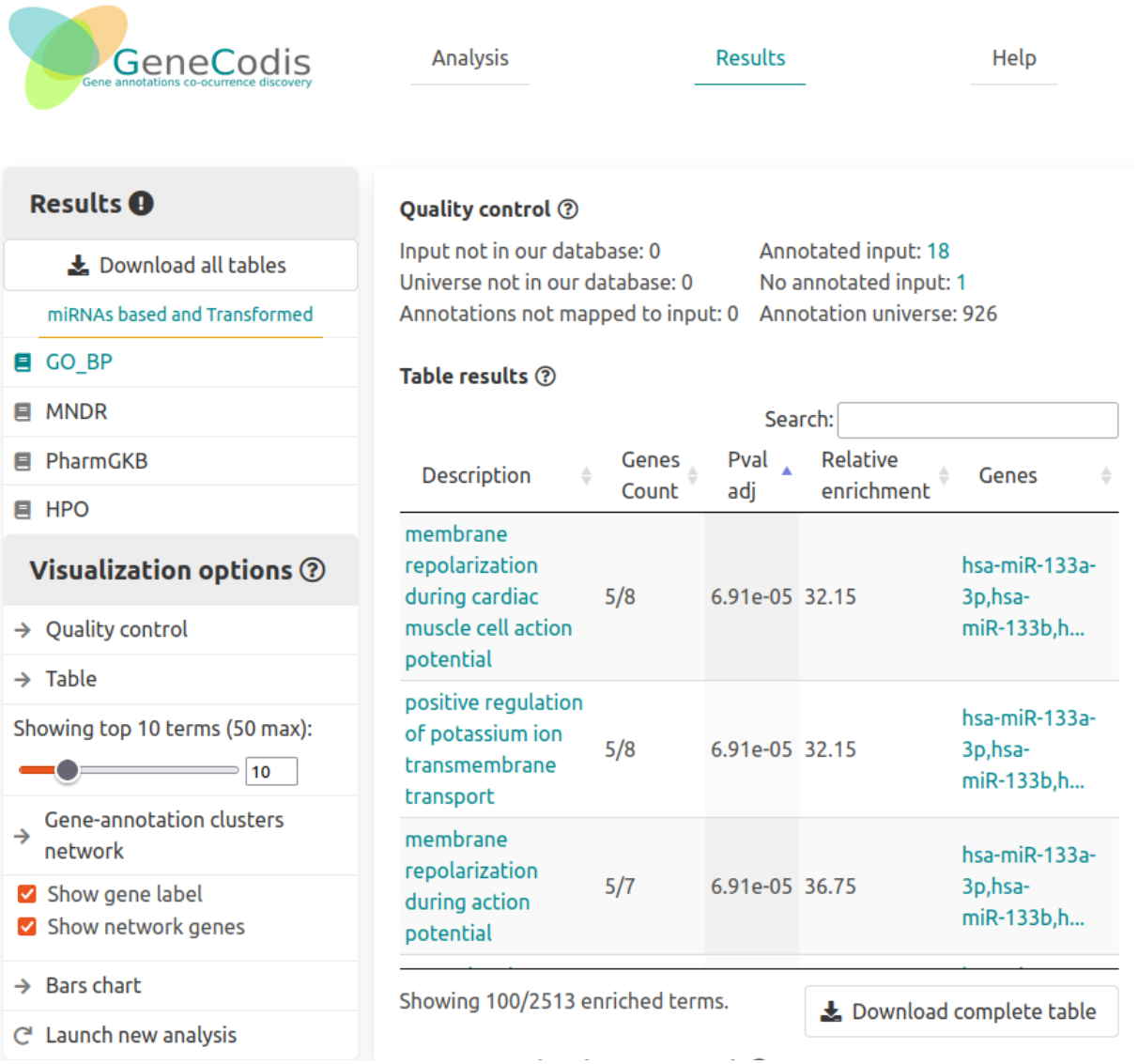


Figure 10. Screenshot of GeneCodis GO biological process results page resulted from the analysis of arrhythmia-related miRNAs against the databases transformed to miRNAs and miRNAs-based.

Further details of the results report are explained in section 4.4. In this way, at the moment we can use these results to focus on understanding the results visualisations. For example, we could select the top 15 enriched terms from GO biological process results, centre into the gene-annotation clusters network and click some annotation nodes (blue) to display their label and associated miRNAs (Figure 11).

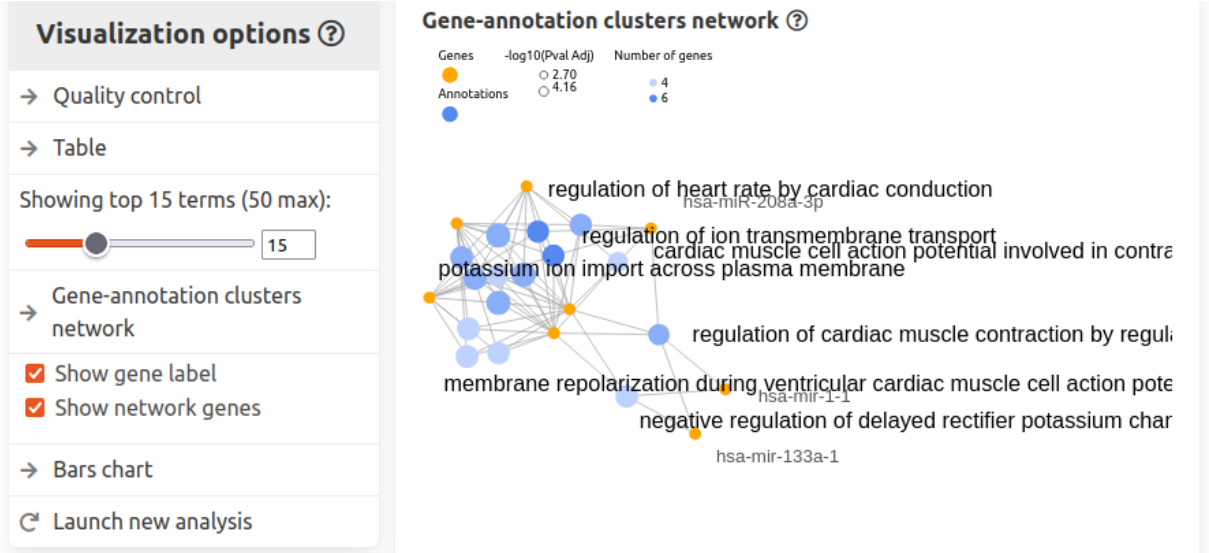


Figure 11. Screenshot of the gene-annotation clusters network with the top 15 enriched terms from GO biological process results, with the genes and their labels displaying activated. Results from an arrhythmia-related miRNAs against the databases transformed to miRNAs and miRNA-based.

Another option is to hide the genes and their labels, which automatically displays the selected top annotations labels (Figure 12).

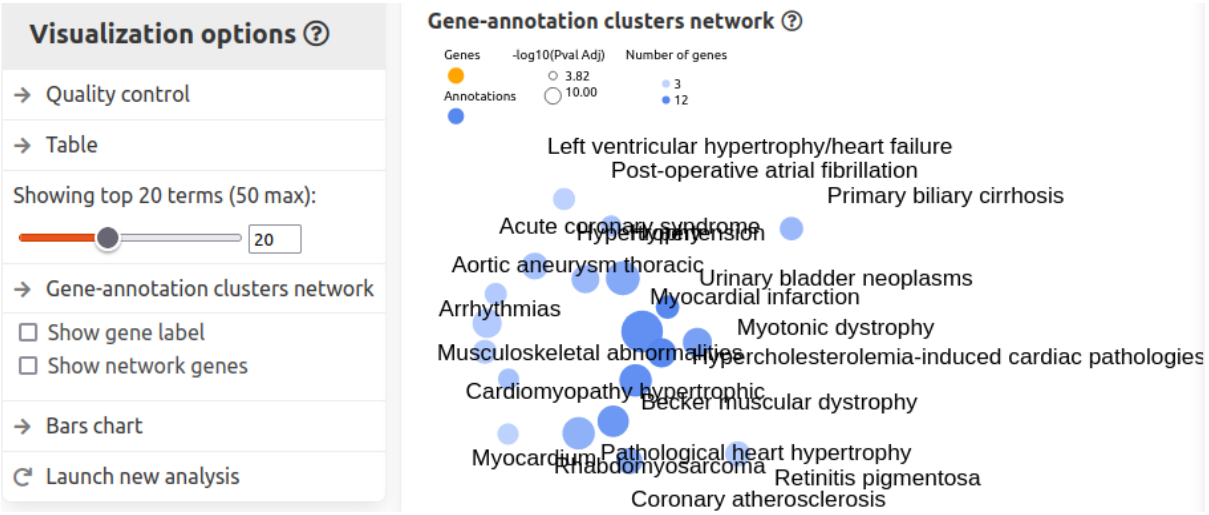


Figure 12. Screenshot of the gene-annotation clusters network with the top 20 enriched terms from MNDR results, with the genes and their labels displaying deactivated. Results from an arrhythmia-related miRNAs against the databases transformed to miRNAs and miRNA-based.

Or in case that the resulting network is too tangled we can jump directly to the simple but

effective bars chart (Figure 13).

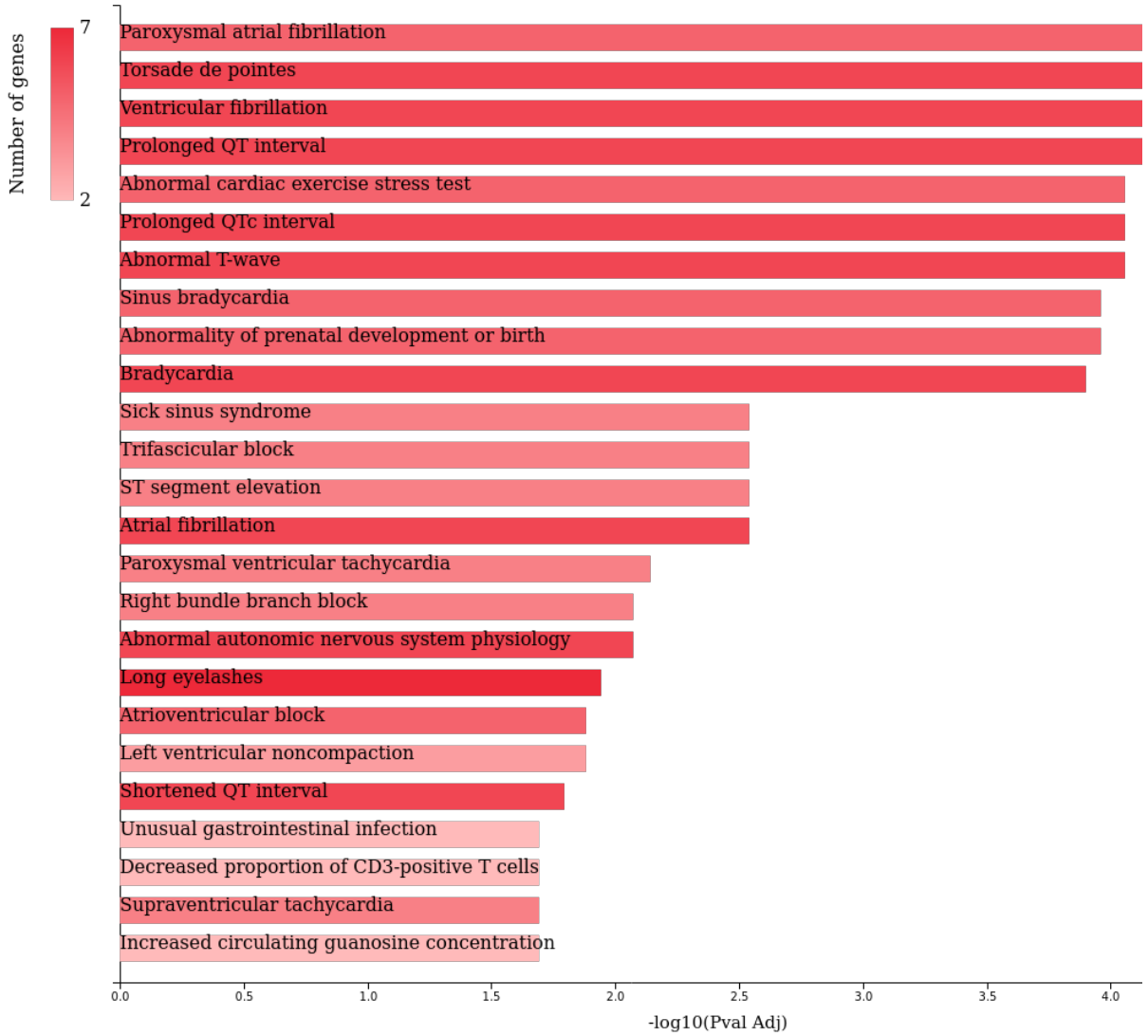


Figure 13. Bars plot downloaded from GeneCodis with the top 25 enriched terms from the HPO results. Results from an arrhythmia-related miRNAs against the databases transformed to miRNAs and miRNA-based.

All the results displayed in the previous figures include two miRNAs SEA methods from GeneCodis, first the gene annotations transformed to miRNAs, GO and HPO, and the direct miRNA-based annotations from MNDR. Given that the idea is to assess the different approaches available in GeneCodis, before going into details about these results, the following figures explain how to launch the remaining two analyses in GeneCodis.

The next analysis is the **Wallenius approach,** described in section 3.4. We could launch a new analysis or, more straightforward, click in the "Analysis" tab at the top navigation bar to have the input and all the previous options already prepared. To this extent we would only need to change the enrichment stats to "Wallenius" in the advanced options of GeneCodis and launch the analysis. Those are the different options to tune the GeneCodis SEA and perform any of the different approaches explained in section 4.3 (Figure 14).

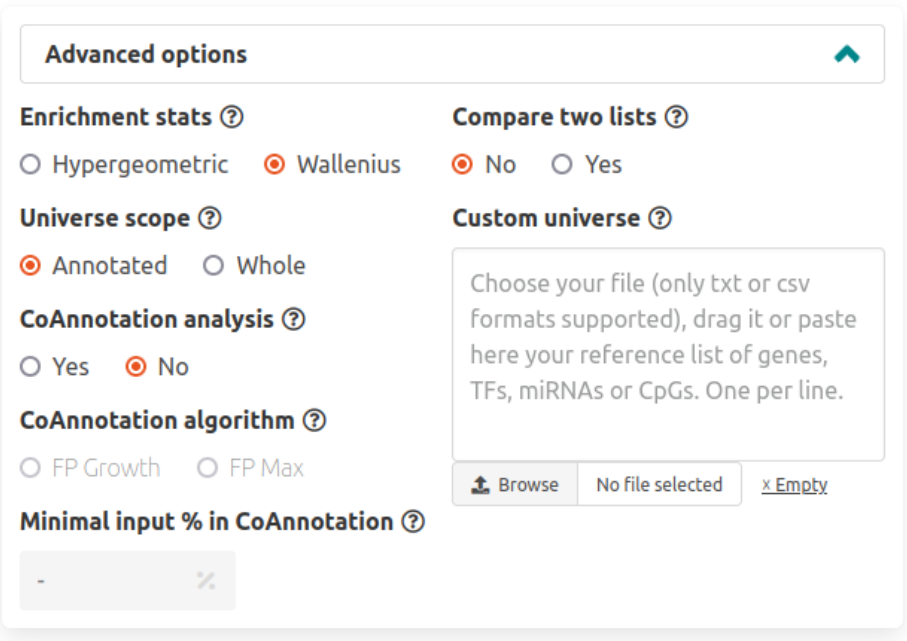


Figure 14. Screenshot of the GeneCodis 4 advanced settings configuring the Wallenius approach for the miRNAs SEA.

Now miRNAs SEA is done by means of the target genes weighted. An unique feature of the Wallenius approach in GeneCodis is the construction of a table that links the miRNAs and its targets annotated in the selected databases. It is linked at the bottom of the left navigation panel in the previous. Additionally, it can be observed that the numbers obtained at the "Quality control" section are higher compared to the results obtained with the previous approaches (Figure 15). Concretely, with the Wallenius approach the universe scoped set to the annotation considers only the genes that are a target of miRNAs in miRTarbase and the miRNAs directly included in gene-based resources. This is noted to create awareness about the importance of the input and the annotation universe count, which it will be discussed when comparing the different approaches.

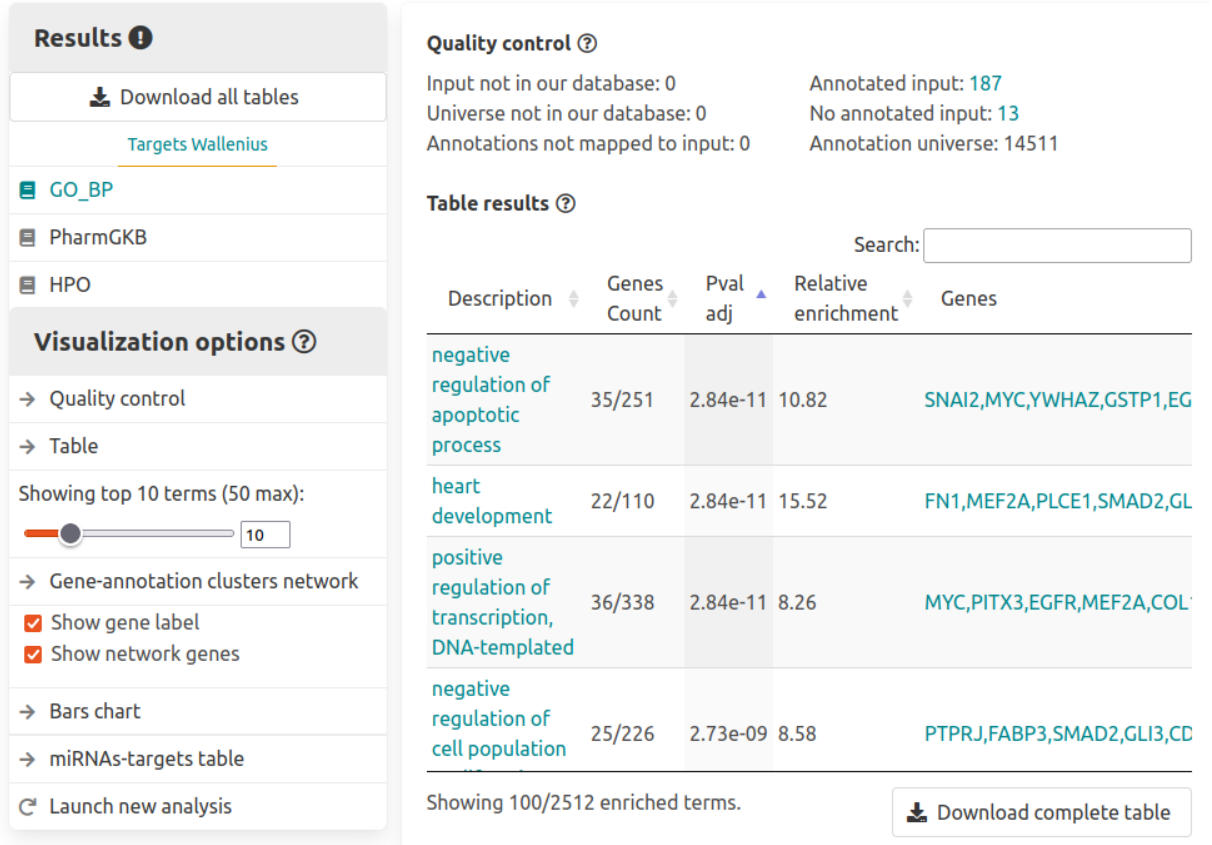


Figure 15. Screenshot of GeneCodis GO biological process results page resulted from the analysis of arrhythmia-related miRNAs using the Wallenius approach.

The final analysis is the **traditional approach**, testing the annotations with the central hypergeometric distribution using the target genes. Thus, the first step is to extract all the target genes available, which can be done by clicking the "Annotated input" number for each database. For example, clicking the "187" here produces a plain text table with the recognised genes in gene symbol nomenclature, their description and in synonyms the gene name of your input (Figure 16).

Official Symbol	Description	Synonyms
ABCB1	ATP binding cassette subfamily B member 1	ABCB1
ABCG2	ATP binding cassette subfamily G member 2 (Junior blood group)	ABCG2
ACHE	acetylcholinesterase (Cartwright blood group)	ACHE
ADAR	adenosine deaminase RNA specific	ADAR
ADCY1	adenylate cyclase 1	ADCY1
AFTPH	aftiphilin	AFTPH
AGO1	argonaute RISC component 1	AGO1
AKT1	AKT serine/threonine kinase 1	AKT1
ANXA2	annexin A2	ANXA2
API5	apoptosis inhibitor 5	API5
ARPC5	actin related protein 2/3 complex subunit 5	ARPC5

Figure 16. Screenshot of GeneCodis genic information obtained for some of the targets of arrhythmia-related miRNAs annotated in GO biological process.

Once all target genes have been collected we can proceed to analyse them by clicking "Launch new analysis". This will return to the home page with the default configuration of the GeneCodis form, genes/proteins as input type and the hypergeometric test, what we are aiming for. After copying and pasting the target genes and selecting the proposed databases the analysis query can be launched.

With all the results available we can compare the different approaches and database results. The common procedure to assess the results of an enrichment analysis is to observe the **top significant annotations**. In our study case these are expected to recover terms that were employed to generate conclusions of the sourcing miRNAs original paper. Chiefly, the top results ought to reflect properties associated with the **cardiac system excitability**, such as automaticity, conduction, as well as regulation of Ca^{2+} and other ion channels, repolarization, and spatial divergence. Besides, other essential but general properties, such as apoptosis and fibrosis, might be recovered due to their correlation with the pathological and physiological changes in the cardiac system.

The following **bubble plots**, one per database, gather the top 20 enriched terms obtained with the different approaches available sorted by significance. In them the p-value is reflected as the size of the bubble, and in yellow to red gradient the growing relative enrichment. In the Y axis are the annotation terms and in the X axis the approach applied (Figures 17 to 21).

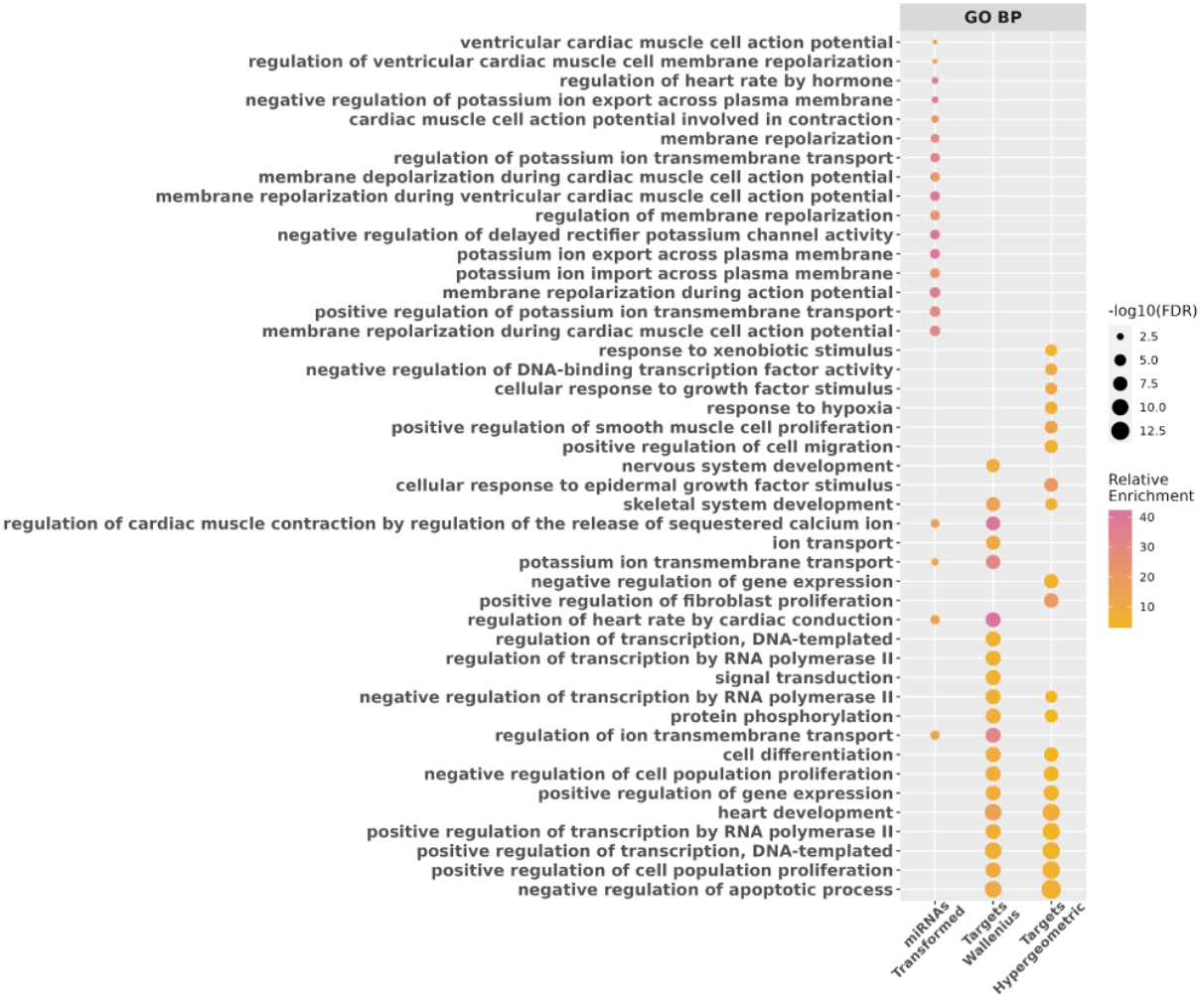


Figure 17. GO biological process GeneCodis results obtained with approaches: the database transformed and the Wallenius and standard hypergeometric tests applied to the targets.

The complete top twenty terms obtained with the transformed GO database (Figure 17) are accurate with the anticipated results. Specifically, these are annotations that point to cardiac muscle, membrane repolarization and ion channels processes. Four of these are shared exclusively with the Wallenius approach, i.e. “regulation of heart rate by cardiac conduction” or “regulation of potassium ion transmembrane transport”. Worth noting is the higher relative enrichment score compared to other terms in the targets based approaches. Although, uniquely these methods find enriched processes related to apoptosis or heart development, in general and especially with the hypergeometric approach, most terms are nonspecific and are

also noninformative such as those related to the miRNAs basic function: regulation of transcription and gene expression.

Figure 18. HPO GeneCodis results obtained with approaches: the database transformed and the Wallenius and standard hypergeometric tests applied to the targets.

Once again, the transformation of the HPO provides the best results, although this time, it shares more specific annotations with the miRNAs target hypergeometric test (Figure 18). These are cardiac pathology phenotypes, such as: “Atrial fibrillation”, “Sinus bradycardia”,

"Abnormal cardiac exercise stress test", "Torsade de pointes". Common grounds of the three strategies are "Ventricular fibrillation", "Abnormal T-wave" and "Prolonged QTc interval", which stand out within the Wallenius results given their higher relative enrichment. Other hypergeometric results highlight "Syncope" shared with Wallenius, and "Shortened QT interval" exclusive of it. The rest are general terms i.e. inheritance or referred to abnormal body clinical signs i.e. "Wide nasal bridge".

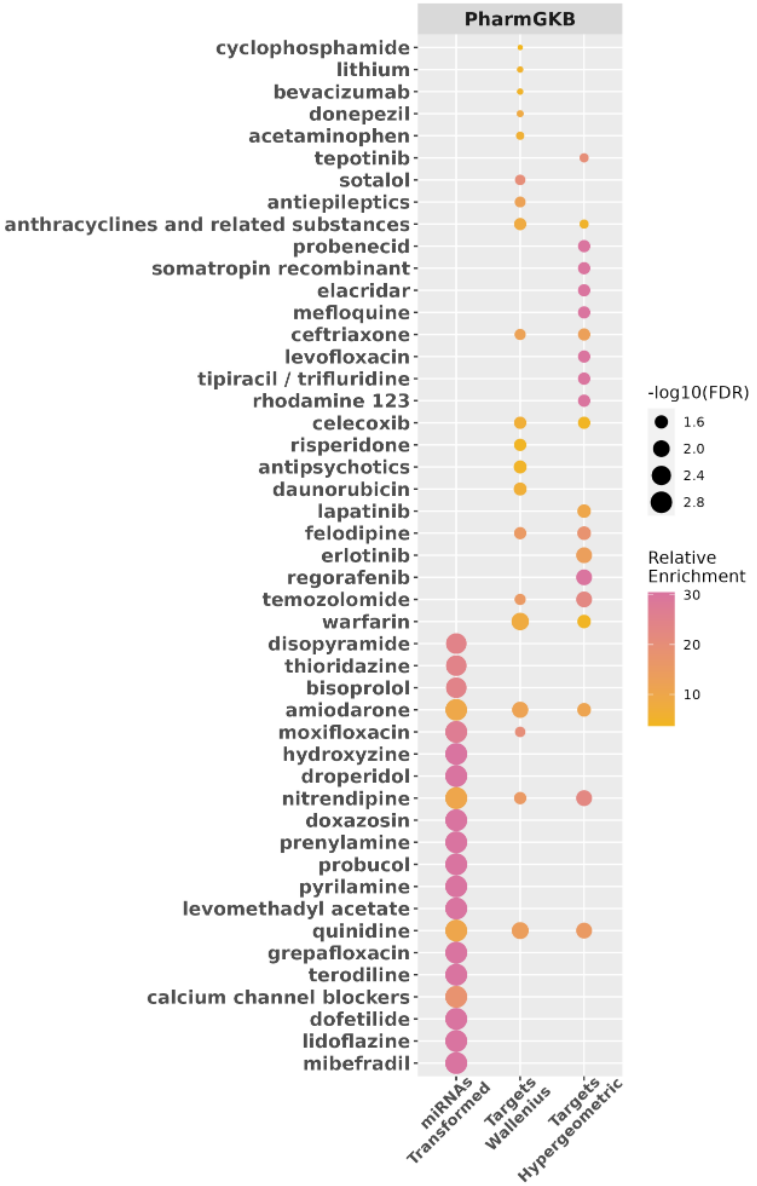


Figure 19. PharmGKB GeneCodis results obtained with approaches: the database transformed and the Wallenius and standard hypergeometric tests applied to the targets.

For the case of PharmGKB (Figure 19) the top enriched terms share antiarrhythmic drugs across the three methods, like the "quinidine" and "amiodarone" along with Ca^{2+} ion channel blocker "nitrendipine". "Felodipine" is also a calcium channel blocker but is only missed from the top 20 results testing the database transformed to miRNAs. This procedure, nonetheless, finds another antiarrhythmic, "dofetilide", beside the broad annotation "calcium channel blockers" and its derivational drug "mibefradil". Other interesting medicines could be the "lidoflazine" as an HERG K+ channels blocker, and the antihypertensive "doxazosin". Pharmaceutical drugs like "grepafloxacin", "moxifloxacin", "prenylamine" or "levomethadyl acetate", could provoke cardiac afflictions, like long QT syndrome and torsades de pointes. All these terms are closely related to arrhythmias and are found mainly in the miRNAs transformed database. Finally, though a bit distant from our study case, "warfarin" is an anticoagulant drug, found only in the top 20 results of the target- based approaches.

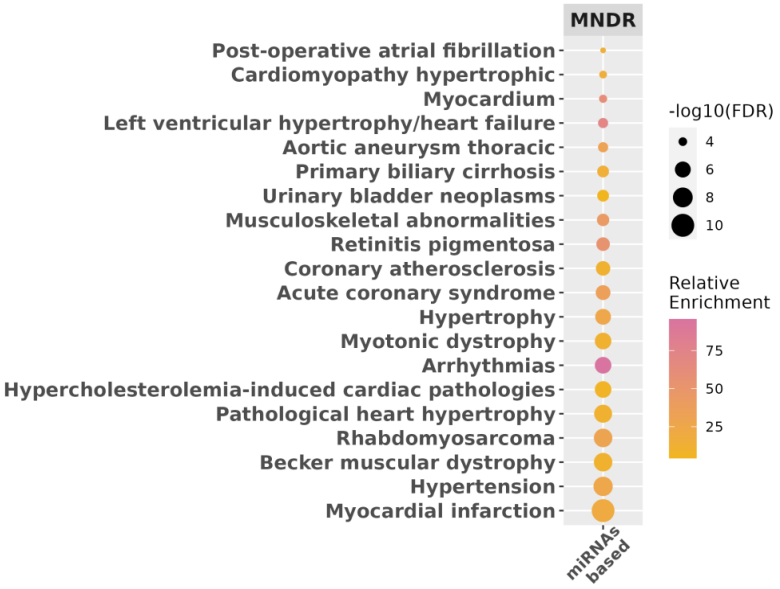


Figure 20. GeneCodis results obtained with MNDR, a miRNA-based database.

Finally, the MNDR database is built for miRNAs and its analysis is here the only approach completely free from limitations. Its top results yield sixteen heart diseases within the top 20, the four remaining are unrelated to arrhythmia terms: "Primary biliary cirrhosis", "Urinary bladder neoplasms", "Musculoskeletal abnormality" and "Retinitis pigmentosa" (Figure 20).

In conclusion, the results obtained agree with some of the methodologies reviewed in the previous section. For example, the transformation of the gene-based databases to miRNAs annotations provide the most phenotype accurate results. However the reduced background universe to the miRNAs level can produce very stringent statistics. Furthermore, the analysis via the target genes can provide deep insight within the regulatory network context. Although common biassed or irrelevant terms appear as significant, the p-value in combination with a high relative enrichment score makes the Wallenius approach to recover phenotype specific biological signals.

5. Bias Assessment In Transcription Factors Enrichment Analysis

Throughout the review of the current miRNAs enrichment tools and the development of GeneCodis we realised that the enrichment analysis for TFs is much less explored compared to the methylation or miRNAs field. As far as we are aware, there are scarce tools for conducting TFs functional annotations enrichment analysis. Among the few published methods, the database TRRUST (Transcriptional Regulatory Relationships Unravelled by Sentence-based Text mining) [168] provides information on TF target gene interactions via natural language processing algorithms later scrutinised via manual curation. Its latest version contains 8444 TF target gene pairs for 800 TFs in humans and 6552 for 828 TFs in mouse. Accordingly, TRRUST applies the Fisher's exact test to each individual TF set of target genes testing the annotations from the Disease Ontology [169], KEGG pathways and the GO biological process category. Another tool, CistromeGO [170], is a web-based tool that evaluates ChIP-Seq peaks and expression data to rank the target genes of TFs and assesses GO terms and KEGG pathways using a GSEA approach with the minimum hypergeometric test. Lastly, **TFTenricher** [171] is a Python toolkit that converts a list of TFs to their target genes and discover the overrepresentation of gene sets annotated in KEGG, GO, Reactome, the GWAS catalogue, and also allowing user-uploaded custom gene sets implementing the Fisher's exact test.

The three mentioned tools infer the downstream functional implications of TF indirectly via the targets. This fact exposes a complete lack of databases that annotate such effects in the regulation networks and that the current functions attributed to TFs majorly highlight their gene expression regulation role. Even though these are true annotations, little to no information can be subtracted from their regulation effects performing SEA of TFs as genes. Adding up the fact that each TF regulates different numbers of genes, the analysis of the TFs targets with the traditional SEA approach might also cause biassed results. However, TRRUST explores the functions of each TF individually, thus no target gene is overcounted. On the other hand CistromeGO, analyses target genes revealed by an experimental technique, not transforming TFs to targets, besides, it applies GSEA. It is only the recently published

TFTenricher tool that actually does SEA of TF lists in the traditional sense.

In the next section, the TFs target enrichment analysis based on the Fisher's exact test is evaluated for biassed results together with the ability of the Wallenius non central hypergeometric approach to reduce the false positives. The analysis looks for biassed terms in the functional databases of KEGG, GO biological process, WikiPathways and Reactome.

5.1. Source of TF target genes

The first step in this part of the research is to determine the collection of TF and its target gene links, a set of TF regulons.

Unlike the miRNAs, TF gene binding prediction algorithms are less extended, probably because of the complexity in the interaction DNA - protein. Still there are different approaches available that make use of current knowledge of protein interactions and Chip-Seq data analysis to suggest transcription factor binding sites and candidate TF target genes [172–174]. Despite this, several databases sustain experimentally screened and computationally predicted associations between TF and target gene. In our proposed analysis, the regulons of **DoRothEA**[148] are selected because it consolidates human TFs regulons from various well established sources into a single resource further expanding it with analysis of gene expression data. The database includes four different levels of evidence that backs each different source. The first and most trusted source is a set of twelve public databases whose TF target gene links come from manual literature curation. Next, it includes the results of ReMap [175], a high resolution analysis of manifold DNA-binding experiments. The third level is the prediction of TF binding sites (TFBS) on gene promoters from HOCOMOCO [176] and JASPAR [177]. At last, the previous databases are expanded by applying ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) [178] to The Cancer Genome Atlas [179] and normal human tissues from GTEx [180] to infer novel regulons. Finally, depending on the distribution of those evidences across all the TF target pairs, DoRothEA classifies them in five levels, A to E, in decreasing available proof.

Which level to use depends on the demands of the research, in our context, we want to evaluate the functional enrichment analysis of TFs target genes thus it is normally desired an equilibrium between experimental evidence and annotation coverage. To address the DoRothEA evidence level selection, we observe the coverage of functional terms of human TFs and TF regulons (Figure 21).

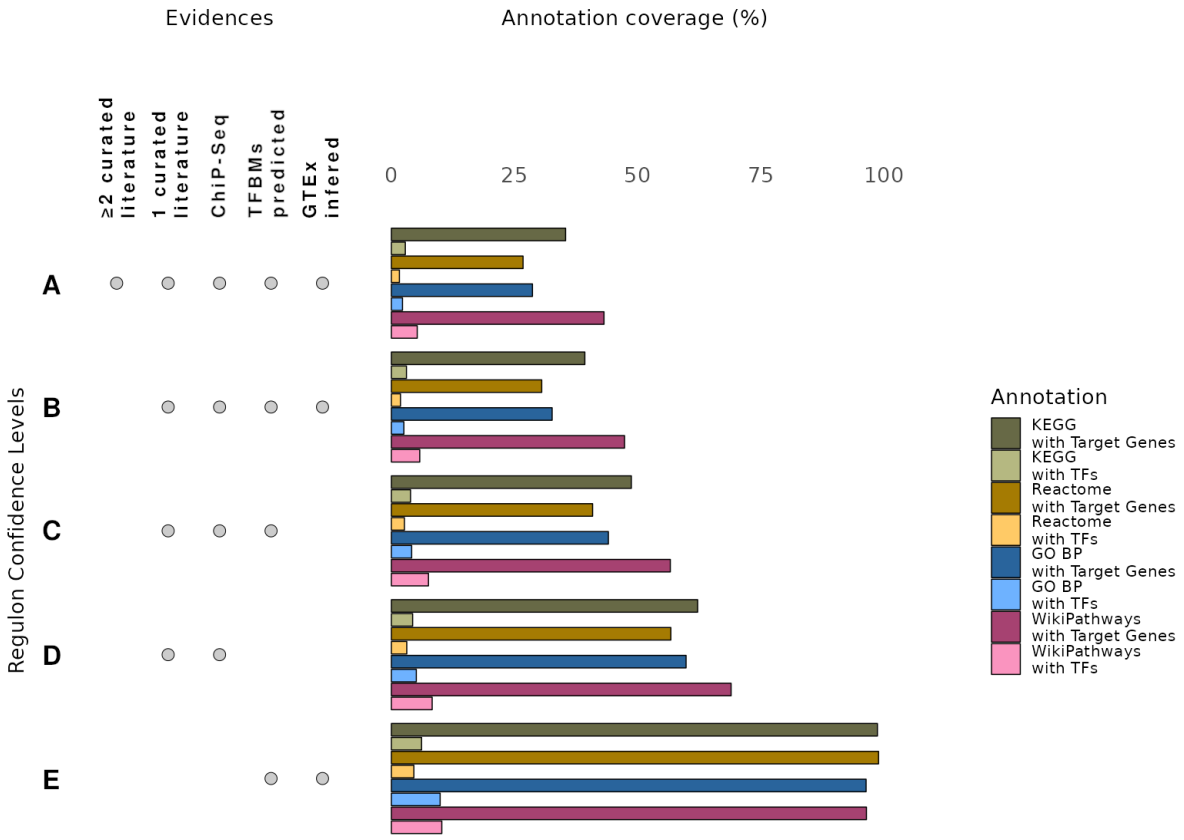


Figure 21. Evidence distribution of the 5 DoRothEA levels of confidence and their annotation coverage in different databases. It shows the type of evidence and the annotation coverage, in percentage, that each TF regulon's confidence level has using only the TFs or the target genes for each database.

The human TFs in all DoRothEA levels represent only a small percentage of annotations across all functional databases. However, when analysing the coverage within TF regulons, a growing variation is observed. The manually curated regulons in DoRothEA A, and B cover a range of 26% for Reactome in DoRothEA A to 47% for WikiPathways in DoRothEA B.

DoRothEA C, which includes TF regulons evidenced by two proofs, TF binding sites and curated resources or ChiP-seq data, shows an annotation coverage between 40-57%. On the other hand, DoRothEA levels D and E, which contain TF regulons with a single source of evidence and those computationally inferred, capture 56-69% and close to 100% of the terms respectively. Based on these observations, it is reasonable to use the TF regulons from DoRothEA C, which have the maximal coverage and sufficient experimental support for the TFs targets mapping.

5.2. Assessment based on null simulations tests

To explore the possibility of biassed results in TFs targets SEA with the approach of TFTenricher we are going to follow the same approach used for the SEA of miRNAs and CpGs associated genes. The method is a **permutation-based** approach, which involves conducting multiple simulations of **SEA using randomly selected** input **TFs** lists of the same size without replacement. For each simulation, the proportion of significant terms (p-value *adjusted* < 0.05) is calculated by dividing the number of times each term appears as significant by the total number of repetitions. The simulations are repeated 1000 times, using input lists of different sizes of (2, 3, 4, 5, 10, 20, and 30) TFs selected at random.

In order to test three different hypotheses, three **different SEA approaches** simulations are explored:

1) SEA of **TF genes** does not provide useful information (Section 5.2.1.).
2) **Fisher's exact** SEA **TF target genes** produce biassed results (Section 5.2.2.).
3) **Wallenius's** SEA of **TF target genes** reduces the overrepresentation of skewed results (Section 5.2.3.),

The Wallenius's method here proposed follows the same alternative rationale applied in the miRNAs target genes analysis, **see section 4.1**. The exclusive difference is that the gene power weighting function is modelled with the number of TFs linked to each target and, equally, it is fitted with a binary vector, where 1 implies a gene being targeted by at least one input TF and 0 otherwise.

5.2.1. SEA of TF genes

As previously mentioned, TFs are annotated with specific terms of functions closely related to gene expression and transcription processes. Subsequently, it is expected that the SEA of TFs lists will not provide novel information about their functional implications in the regulatory network. To test this hypothesis we apply the first proposed null simulation, where the annotations are tested against randomly generated lists of TFs. Not surprisingly, in the 1000 TF lists tested, **concrete terms are found to be always significant** (p-value < 0.05). Precisely, GO biological process category terms related to the regulation of transcription appeared in 100% of the simulations, which also happens in Reactome close to that frequency too. Likewise, the KEGG pathways of transcription and cancer related are the most constant, >90% of the time, while WikiPathways shows ubiquitous terms associated with adipogenesis. These biassed results can be observed to be directly correlated with the size of the TFs list where functional annotations appear more often significantly enriched and lower p-values (Figure 22).

5. BIAS ASSESSMENT IN TRANSCRIPTION FACTORS ENRICHMENT ANALYSIS

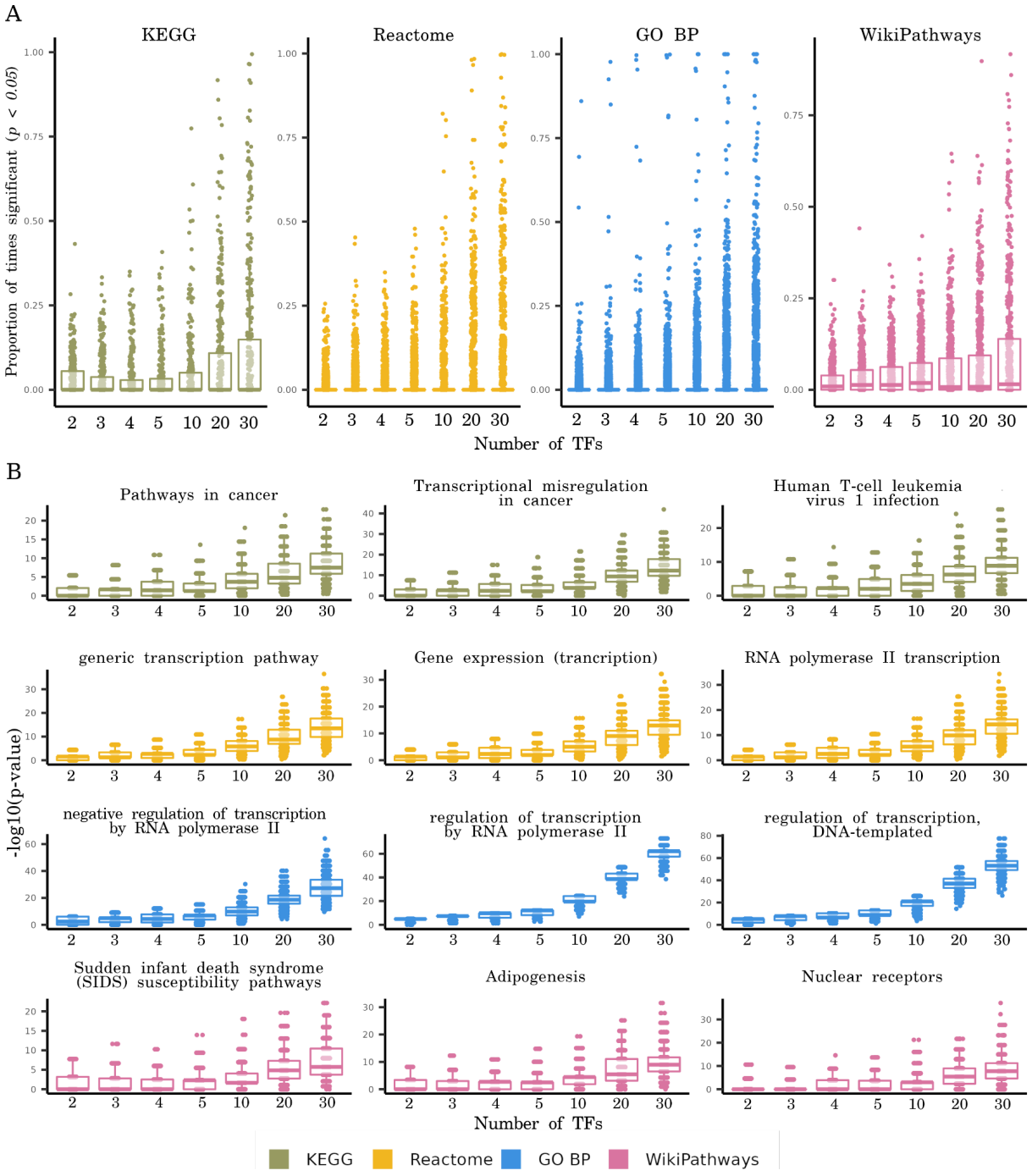


Figure 22. Null simulations results from Fisher's exact SEA of the annotations of TF genes with 1000 random TF lists of different sizes. A) Boxplot that reflects the proportion of times that each term results significant ($p < 0.05$). B) Boxplot of the -log10 of the p-values for the top three more frequent terms for each database.

5.2.3. Fisher's exact and Wallenius's TF target genes SEA

The previous results prove the need of performing TFs SEA via their target genes in order to find the biological functions that they regulate. Studying the distribution of TFs and targets across the DoRothEA C regulon might hint us the different probability each gene has of being selected with random lists of TFs (Figure 23).

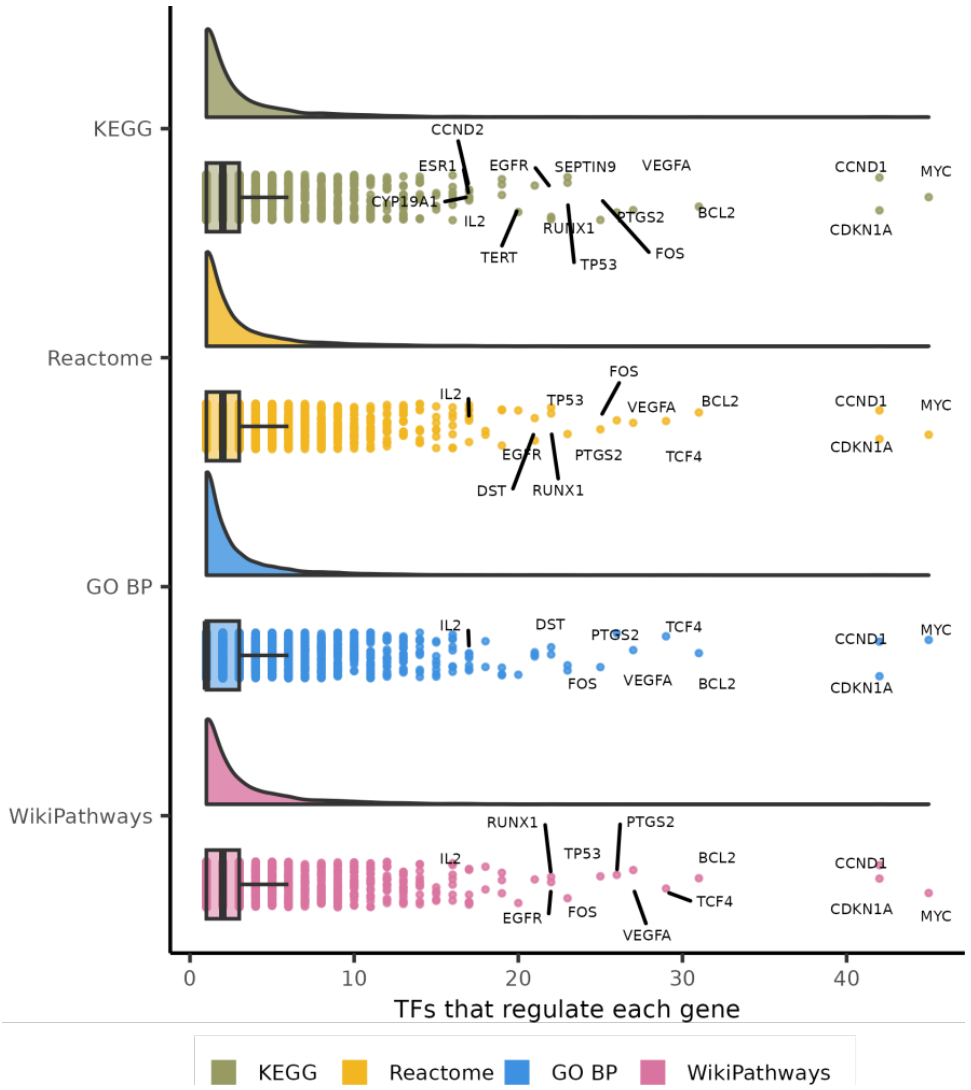


Figure 23. This raincloud plot represents the number of TFs that are regulating each gene according to DoRothEA C regulons in the context of each annotation database. Most genes are regulated by a few TFs as shown at the left distribution while only a few genes are regulated by many TFs differentially.

5. BIAS ASSESSMENT IN TRANSCRIPTION FACTORS ENRICHMENT ANALYSIS

Clearly there are a few genes with an exceptional number of TFs associated, particularly MYC, CCDN1 and CDKN1A have more than forty with enough experimental evidence in DoRothEA. Meanwhile, only one TF targets around 46-53% of the annotated genes. The uneven distribution of TFs targeting specific genes breaks the hypergeometric distribution assumption and highlights a potential bias in the traditional SEA of TFs regulons. Some genes are more likely to appear in the list of targets when transforming a list of TFs. Furthermore, if we extract the annotations of the twenty target genes with the greatest number of TFs associated it can be observed that they are referred to similar biological processes (Figure 24).

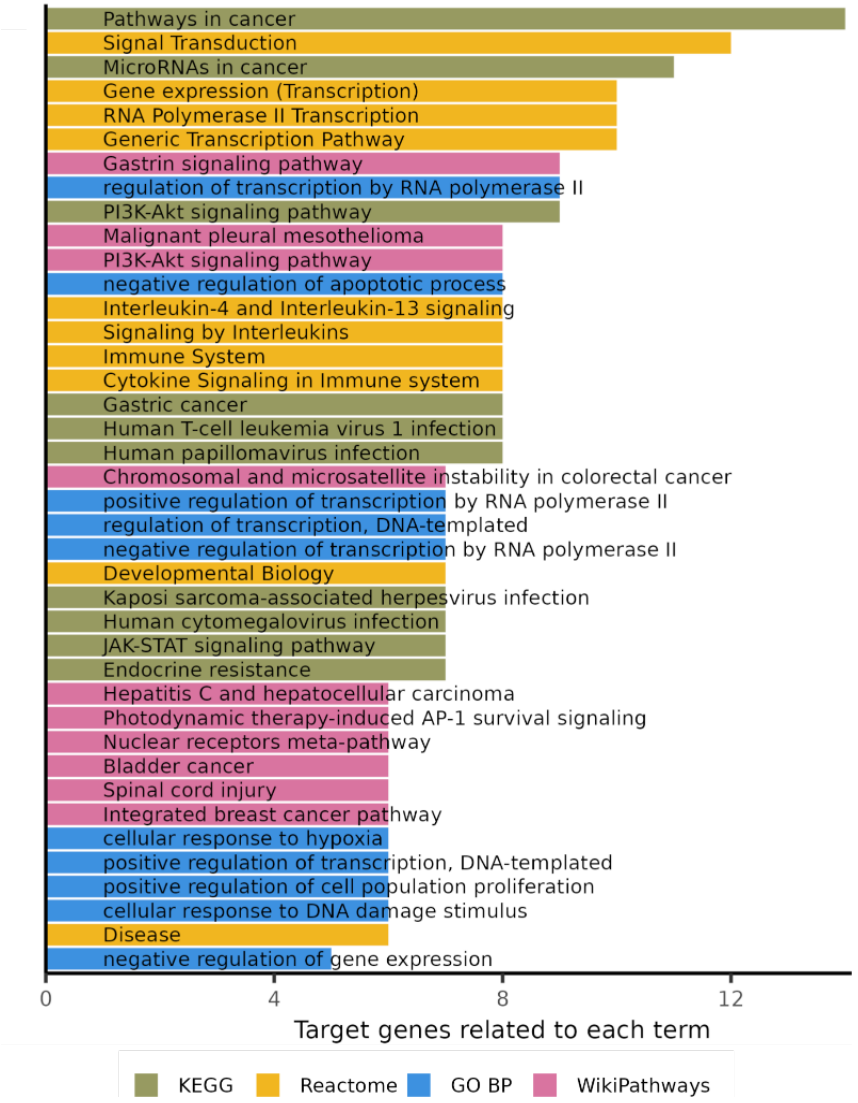


Figure 24. Annotations count using the 20 genes regulated by more TFs. As it can be observed, the overrepresented terms share similar functions.

In KEGG, more than ten genes with the highest number of TFs are associated with terms such as "Pathways in cancer" and "microRNAs in cancer," while Reactome mainly reports transcription terms. Similarly, in GO BP and WikiPathways, the most frequent annotations are "regulation of transcription by RNA polymerase II" and "Gastrin signaling pathway". These observations imply that certain functions may appear differentially enriched in TFs regulons using randomly generated lists of target genes

Finally, observing the results of the **TFs targets SEA null simulations** applying both the Fisher's exact test and the Wallenius's approach we check the effects of the heterogeneous distribution between TFs and target genes and how it can be reduced (Figure 25).

The **Fisher's** exact test traditional SEA approach null simulations report as significant (p-value < 0.05) recurrently the same terms shown in the previous figure, above 90% of the times. In the context of KEGG, the terms related to cancer are ubiquitously found, whereas in Reactome, the most frequent terms are associated with general signalling pathways. On the other hand, GO BP typically returns transcription-related terms, and WikiPathways reveals terms related to both cancer and cell signalling. Here it is also observed that larger sets of TFs are more likely to yield higher significant frequencies and thus a biassed report of terms.

The **Wallenius'** TFs targets SEA null simulations exhibit a reduction in the frequency of significant terms (p-value < 0.05) compared to the proportions obtained from the Fisher's exact test method. Generally, with the Wallenius test, most terms appear by chance less than 25% of the time, except for those with larger TFs lists, although not by far. Notable differences are observed in Wallenius, particularly in the p-values obtained for the top frequent significant terms with Fisher's exact test. It is noteworthy that unlike Fisher's exact test, Wallenius' test moderates the significant frequency and p-values independently of the number of TFs in the simulation, albeit it has a similar trend.

5. BIAS ASSESSMENT IN TRANSCRIPTION FACTORS ENRICHMENT ANALYSIS

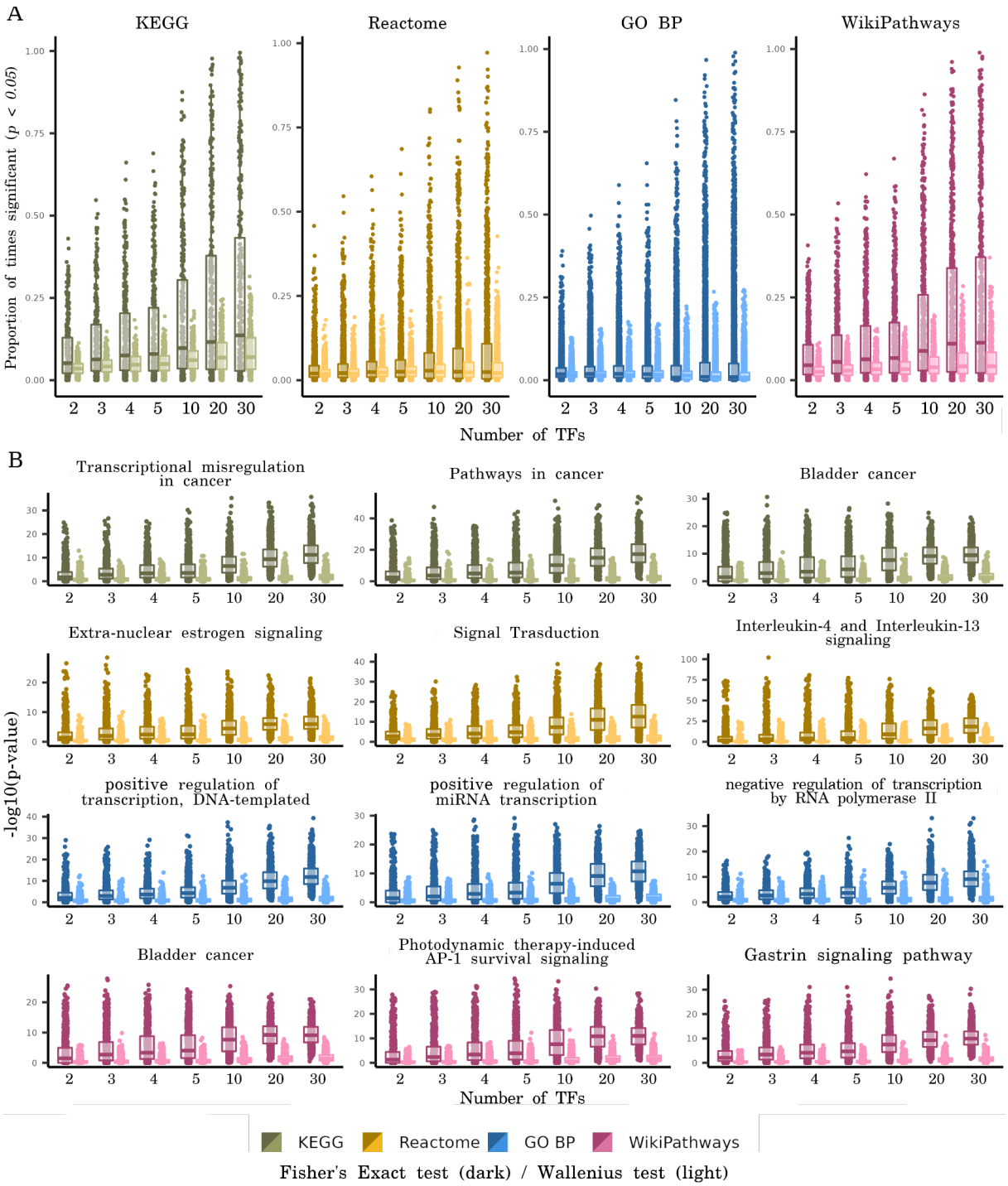


Figure 25. Results of TFs regulons SEA null simulations with Fisher's exact and Wallenius' tests. A) Each point, in cero to one scale, represents the number of times that each annotation resulted significant ($p < 0.05$) after analysing a thousand random TFs lists. B) In these plots is shown the -log(p-value) distribution of the three terms with the highest proportion of times significant.

5.3. Comparative Analysis of the Standard and the Bias-Correction Approaches

To demonstrate the efficacy of our proposed alternative approach in comparison to the standard analysis, we reanalyzed two TFs lists derived from actual cases. The first set consists of 14 TFs that were identified to show differential activity in SLE patients compared to healthy individuals. These TFs biomarkers are a results of a TF activity inference study conducted using two independent expression datasets of patients with Systemic Lupus Erythematosus (SLE): a paediatric dataset comprising 158 patients and 46 healthy controls, and an adult dataset comprising 301 patients and 20 healthy controls [181]. The second TFs list collects 70 TFs reported as significant biomarkers across multiple cancer types. This was concluded from a pan-cancer study performed using RNA-seq gene expression data from 1,056 cancer cell lines and 9,250 primary tumours [182].

After the reanalysis, the top 15 enriched terms obtained by each approach can be compared to understand the different and common conclusions that both methods reach. Precisely, calculating their respective rank differences (Fisher's term rank - Wallenius' term rank) we can observe that small differences generally point to similar top ranking positions. Then, negative differences indicate biassed terms that are reduced in significance by the Wallenius approach. Oppositely, high positive differences are indicative of the recovery of biological terms that are concealed by the biassed terms obtained with the Fisher's exact test.

The results of the lupus TFs SEA manifest that both Fisher's exact and Wallenius' tests, in all the databases, return similar enriched terms related to immune processes or more specifically autoimmune diseases, virus infection and interferon pathways. Other terms also closely related with the previous terms are found in top positions only by the Wallenius' approach, specially noted in the WikiPathways results with the higher positive difference of ranking. Interestingly, other terms whose ranking difference results in high numbers are essentially more specific terms and are commonly related across the different databases. Here, it is important to understand that this scenario is also a result of Wallenius undermining the non SLE related terms, i.e. cancer and cell cycle terms. Finally, the annotations with little or

negative rank difference are mainly associated with genes that are regulated, in average, by more TFs, which is reflected with a darker colouring of the bars (Figure 26).

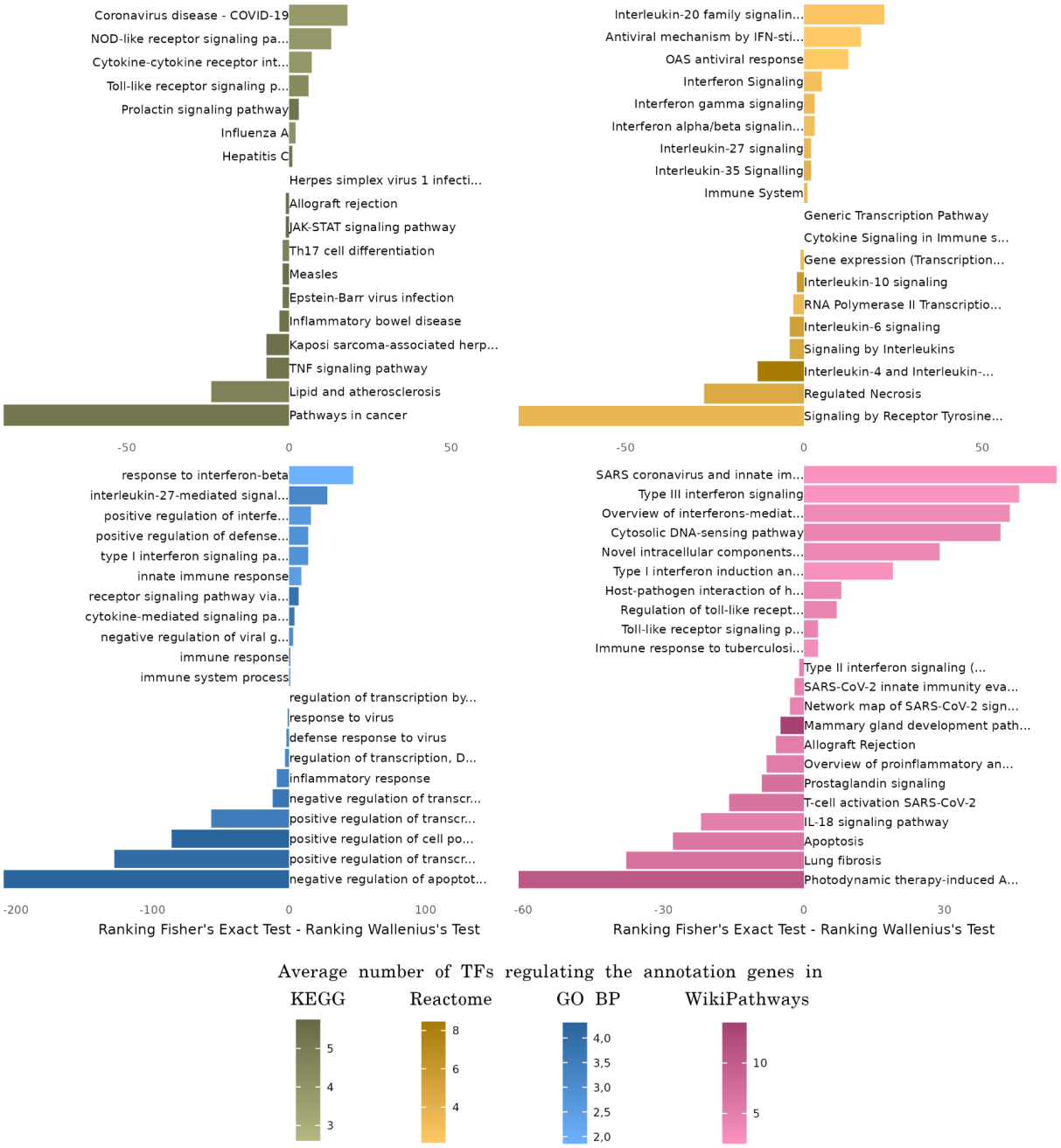


Figure 26. Difference of ranking between the top 15 enriched terms obtained with Fisher's exact and Wallenius' tests in the SLE TFs targets SEA. Bars are coloured with the average number of TFs associated to the genes in the term.

The selection of the second study case related to cancer aims to demonstrate that the

Wallenius approach does not dismiss terms that are biassed with the traditional approach. By looking into the results, a similar fashion to the previous case is shown where the top enriched annotations of each database by both approaches are primarily associated with the study's context, such as cell cycle and cancer (Figure 27).

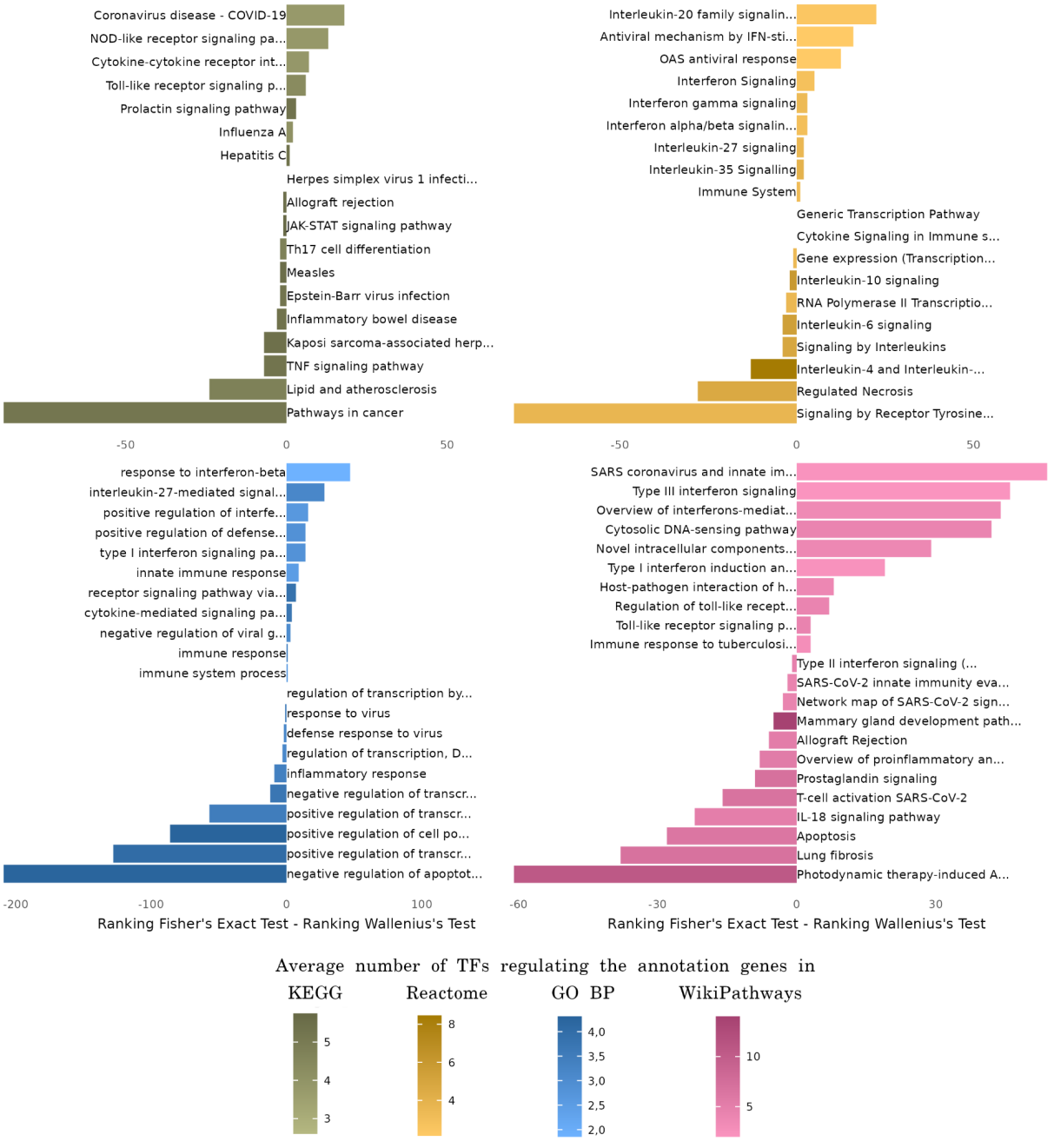


Figure 27. Difference of ranking of the top 15 enriched terms obtained with Fisher's exact and Wallenius' tests in the pan cancer TFs targets SEA. Bars are coloured with the average number of TFs associated to the genes in the term.

Some of these top enriched terms exhibit significant negative differences in the previous use case. This result illustrates that the Wallenius' approach does not penalise biassed annotations if the input is specific to the phenotype under investigation. Furthermore, the terms recovered in the top positions by the Wallenius approach are highly consistent across different databases. Again, the smaller differences in ranking between both approaches can be attributed to a large number of TFs associated which increases the chance of biassed terms occurrence.

In conclusion, this results evinces the existence of biassed terms in the TFs targets SEA using the traditional approach based on the hypergeometric analysis and the capacity of the Wallenius approach as an alternative method which has been also implemented in the GeneCodis 4 application.

7. Discussion

The analysis of functional annotations is a way to extract current biological knowledge from lists of candidate biomarkers. There are different approaches in functional enrichment analyses, but this thesis is focused on the GeneCodis methods [141]. GeneCodis is a web tool that performs MEA through the discovery of co-annotations which along with individual annotations are tested with the SEA Fisher's exact test. The alternative method is the GSEA which, unlike the SEA, requires the whole set of measured elements ranked by a given metric.

SEA are straightforward methods daily applied for analysing genes or proteins and a complete catalogue of well-known tools exist, particularly, DAVID [79] and Enrichr [183]. Contrarily, the tools available for the analysis of regulatory elements, such as methylation data, miRNAs and, specially in TFs, are rare and their methodology varies. Most of these perform the functional characterisation of these regulators via their associated genes. However, it was described that the Fisher's exact test SEA of CpGs and miRNAs associated genes reports ubiquitously, as top significant, specific functional annotations, such as transcription regulation, cell differentiation and development in methylation [94] and cell cycle and cancer in miRNAs [97]. The issue arises because the different number of CpGs and miRNAs distributed per gene breaks the basic statistical assumption of Fisher's exact test, the equal probability of item selection of the central hypergeometric distribution. A similar situation could be accounted for in the analysis of TFs target lists. Surprisingly, current tools for the enrichment analysis of miRNAs and TFs do not consider this methodological limitation and still test the annotations via Fisher's exact test, such as the MIENTURNET [130] or NcPath [127] and TFTenricher [171] respectively.

Before this research work, the bias solution was only proposed for methylation CpGs and miRNAs. Namely, the SEA of CpGs associated genes is an adaptation of the GOseq methodology which is based on the Wallenius noncentral hypergeometric distribution test [94]. This approach is currently the gold standard for methylation SEA done at the CpG probe level, but it is implemented only as R's package in missMethyl [102]. On the other hand the miRNAs SEA approaches to overcome the bias differ. For instance, the calculation of

empirical p-values was one of the first approaches implemented in tools like miRSystem [184] or DIANA miRPath [133]. Nevertheless, the empirical sampling in the unbiased miRNAs SEA is reported by Bleazard et al. [98] to be very stringent and it requires millions of permutations in order to obtain enough p-value granularity which complicates its integration in web tools due to computational costs. Web tools like miRNet [134] only permutes a thousand lists and, currently, only BUFET [132] generates a proper null distribution with a million of permutations thanks to deeply optimised Python's script. The miTALOS [131] tool adjusts the background universe to a concrete tissue gene expression thus removing target genes that are not expressed, this reduces the bias effect partially and its useful for only those research studies that share the same tissue. Other well-established tools implement, the inclusion of miRNAs curated annotations and the transformation of gene-based annotations to miRNA-based using experimentally validated targets. TAM [104] builds its own curated miRNAs annotations database which is incorporated along with other direct annotations in tools like miEAA [108] or miRNet. miEAA additionally incorporates the gene-based annotations of KEGG and GO transformed by means of the curated targets of miRTarBase [121] database.

Other methods that tackle the issue of biassed results like methylGSA [185] or ebGSEA [103] were not explored in our research because they are based on the GSEA approach and we aimed for a generalist approach. These methods require analysis in conjunction with an associated metric, such as p-values. In contrast, the Wallenius-based methods, like the Fisher's exact test, allow for testing various types of data sources without the need for additional metrics.

Currently, GeneCodis 4 [92] implements known methods for the functional enrichment analysis of CpGs such as the Wallenius approach but it is the only web tool that offers it. GeneCodis also enables established techniques for the analysis of miRNAs, like the transformation of the gene based annotation databases to miRNAs sets using only targets confirmed experimentally from miRTarBase [121], as it is done by miEAA. This type of transformed miRNAs annotations can often produce many significant results due to the same set of miRNAs. A solution proposed by Godard et al. [97] to avoid redundant results is based on clustering annotations based on the miRNAs shared which is inherent in the GeneCodis MEA algorithm that generates co-annotations via the Frequent pattern algorithm [162]. Moreover, GeneCodis allows

the use of curated annotations from the TAM database, HMDD [105] and MNDR [106] which directly attribute miRNAs to functions and diseases. An additional way of analysing miRNAs targets functional annotations in GeneCodis requires the user to provide a customised background of the genes that are expressed in the studied tissue, following the miTALOS strategy. Although this might seem trivial, many web tools commonly utilised do not offer this possibility such as modEnrichr [83], g:profiler [82] or Panther [88]. However the novelty is that we studied an adaptation of the Wallenius approach for miRNAs whose suitability of could be confirmed with the real study cases analyses of arrhythmias related miRNAs. In the use cases we compared different approaches available for miRNAs SEA available in GeneCodis in order to illustrate the advantages and limitations of each approach when trying to decipher biological implications of candidate miRNAs lists. We realised that there are scarce tools for the SEA of TFs, and that the only one existing analyse them via its regulated genes ignoring a possible bias similar to CpGs and miRNAs due to the intrinsic nature of the regulation networks, namely, the heterogeneous distribution of regulators pointing to each gene. Thus following the same rationale that proved the bias in CpGs and miRNAs we confirmed the presence of biassed annotations that point to cell cycle, transcription regulation, signalling pathways and cancer. We finally proved the Wallenius approach as a solution for the unbiased enrichment of TFs target genes.

During the use cases we can envision some of the limits of this thesis results. For instance, it can be argued that the Wallenius approach is not immune to biassed results for miRNAs and TFs. Nevertheless, this approach should be the preferential choice when the downstream effects of regulatory elements want to be discovered through their target genes or simply because it is the only unbiased approach available as it happens in GeneCodis for TFs and CpGs. It has been observed that the combination of the Wallenius approach and a high relative enrichment score could provide a better prioritisation to select the top enriched terms, though a more in depth investigation is required. In the specific case of miRNAs it was found that the databases of miRNAS directly annotated are mainly focused on human disease which might limit the possibility to discover novel mechanisms and the proposal of new research hypotheses. Additionally, the transformation of gene-based annotations, and the Wallenius method, depends on the sourcing database that collects the miRNAs targets associated, thus

results will differ depending on the target evidence background choice. This dependence of the approaches offers a new window of research to discover associations between the different experimental evidence levels of the regulation links and their biological functions. A future line of work of this thesis would be to perform a benchmark analysis of the singular enrichment methods applied to regulatory elements.

Our results provide the inclusion of different methodologies in GeneCodis 4 for the unbiased analysis of functional annotations in regulatory elements. This along with the MEA algorithm, makes GeneCodis a unique web tool for the SEA and MEA for the functional characterisation of genes, proteins, miRNAs, TFs and CpGs. Nowadays, most of the functional annotations analysis tools implement only SEA or GSEA, then in the case of the analysis of regulatory elements they include a single method to handle the bias and normally are prepared for a single type of biological entity, and some of them require computational skills to be used. Thereupon, GeneCodis4 is a unique type of web application that allows different enrichment analysis strategies such as SEA and MEA, and abides for the biological implications discovery behind large lists of genes, proteins, miRNAs, TFs and CpGs. GeneCodis is prepared for both bioinformatics, given that it is open source and developed as an API, and for bench scientists, as an intuitive web tool with interactive plots. GeneCodis is ideal for the investigators that are planning to combine different types of omics data and jointly analyse different knowledge databases focused on providing the state-of-the-art in the regulatory elements functional annotations analysis.

8. Conclusions

In this doctoral thesis, a novel approach for the functional annotations analysis of gene expression regulatory elements has been developed and it has been included in a new version of GeneCodis. During this research the final conclusions are:

1. Although the functional annotation analysis is a method that has been in use for many years, tools have mainly focused on gene and protein lists, and little has been done in the field of regulatory elements, partly because their annotation databases have not been developed as much and high-throughput analysis has been refined later in these fields.

2. A novel method for the unbiased enrichment of miRNAs has been proposed based on an adaptation of the Wallenius approach applied to the analysis of genome-wide methylation data.

3. The new method is implemented in GeneCodis 4, a pioneering web application for singular unbiased enrichment analysis of regulatory elements such as CpGs, miRNAs and transcription factors. In addition, it includes the traditional method for gene and protein lists, which is also applied to study direct functional annotations of miRNAs, obtained directly through expert curation or by transforming gene annotations at the miRNA level. In addition, its annotation database, co-annotation discovery algorithm and visualisation capabilities have been improved. It has been built as an API to ensure its versatile integration into an easy-to-use web tool and into the programmatic workflows of bioinformatics scientists.

4. The assessment of transcription factor singular enrichment analysis using Fisher's exact test yields biassed annotations, similar to what has been observed with miRNAs and CpGs. Our proposed solution utilising the Wallenius distribution test effectively mitigates these biases.

9. Conclusiones

En esta tesis doctoral se ha desarrollado un novedoso método para el análisis de anotaciones funcionales de elementos reguladores de la expresión génica, que se ha incluido en una nueva versión de GeneCodis. Durante esta investigación las conclusiones finales son:

1. Aunque el análisis de anotaciones funcionales es un método que lleva utilizándose muchos años, las herramientas se han centrado principalmente en listas de genes y proteínas, y se ha hecho poco en el campo de los elementos reguladores, en parte porque sus bases de datos de anotaciones no se han desarrollado tanto y su análisis de alto rendimiento se ha perfeccionado más tarde en estos campos.

2. Se ha propuesto un método novedoso para el enriquecimiento no sesgado de miARNs basado en una adaptación del enfoque de la distribución de Wallenius aplicado en el análisis de datos de metilación.

3. El nuevo método está implementado en GeneCodis 4, una aplicación web pionera para el análisis de enriquecimiento singular no sesgado de elementos reguladores como CpGs, miRNAs y factores de transcripción. Además, incluye el método tradicional para listas de genes y proteínas, que también se aplica para estudiar anotaciones funcionales directas de miARNs, obtenidas directamente a través de la curación por expertos o mediante la transformación de anotaciones de genes a nivel de miARNs. Además, se han mejorado su base de datos de anotaciones, su algoritmo de descubrimiento de co-anotaciones y sus capacidades de visualización. Ha sido construido como una API para asegurar su integración versátil en una herramienta web fácil de usar y en los flujos de trabajo programáticos de los científicos bioinformáticos.

4. La evaluación del análisis de enriquecimiento singular de factores de transcripción mediante la prueba exacta de Fisher arroja anotaciones sesgadas, de forma similar a lo observado con los miARNs y los CpGs. La solución que proponemos utilizando la prueba de distribución de Wallenius mitiga eficazmente estos sesgos.

10. References

1. Cobb M. 60 years ago, Francis Crick changed the logic of biology. *PLOS Biol*. 2017;15(9):e2003243. doi:10.1371/journal.pbio.2003243

2. Crick F. Central dogma of molecular biology. *Nature*. 1970;227(5258):561-563. doi:10.1038/227561a0

3. LaPelusa A, Kaushik R. Physiology, Proteins. In: *StatPearls*. StatPearls Publishing; 2022. Accessed February 1, 2023. http://www.ncbi.nlm.nih.gov/books/NBK555990/

4. Ha J, Park H, Park J, Park SB. Recent advances in identifying protein targets in drug discovery. *Cell Chem Biol*. 2021;28(3):394-423. doi:10.1016/j.chembiol.2020.12.001

5. Benedini L. Advanced Protein Drugs and Formulations. *Curr Protein Pept Sci*. 2022;23(1):2-5. doi:10.2174/1389203722666211210115040

6. Santos FB, Del-Bem LE. The Evolution of tRNA Copy Number and Repertoire in Cellular Life. *Genes*. 2022;14(1):27. doi:10.3390/genes14010027

7. Sloan KE, Warda AS, Sharma S, Entian KD, Lafontaine DLJ, Bohnsack MT. Tuning the ribosome: The influence of rRNA modification on eukaryotic ribosome biogenesis and function. *RNA Biol*. 2017;14(9):1138-1152. doi:10.1080/15476286.2016.1259781

8. Morais P, Adachi H, Yu YT. Spliceosomal snRNA Epitranscriptomics. *Front Genet*. 2021;12:652129. doi:10.3389/fgene.2021.652129

9. Fernandes JCR, Acuña SM, Aoki JI, Floeter-Winter LM, Muxel SM. Long Non-Coding RNAs in the Regulation of Gene Expression: Physiology and Disease. *Non-Coding RNA*. 2019;5(1):17. doi:10.3390/ncrna5010017

10. Green D, Dalmay T, Chapman T. Microguards and micromessengers of the genome. *Heredity*. 2016;116(2):125-134. doi:10.1038/hdy.2015.84

11. Santosh B, Varshney A, Yadava PK. Non-coding RNAs: biological functions and applications: NON-CODING RNAS: POWER AND PROMISES. *Cell Biochem Funct*. 2015;33(1):14-22. doi:10.1002/cbf.3079

12. Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology*. 2013;38(1):23-38. doi:10.1038/npp.2012.112

13. Mattei AL, Bailly N, Meissner A. DNA methylation: a historical perspective. *Trends Genet*. 2022;38(7):676-707. doi:10.1016/j.tig.2022.03.010

14. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol*. 2019;20(10):590-607. doi:10.1038/s41580-019-0159-6

15. Al Aboud NM, Tupper C, Jialal I. Genetics, Epigenetic Mechanism. In: *StatPearls*.

StatPearls Publishing; 2022. Accessed February 2, 2023. http://www.ncbi.nlm.nih.gov/books/NBK532999/

16. Jones PA, Ohtani H, Chakravarthy A, De Carvalho DD. Epigenetic therapy in immune-oncology. *Nat Rev Cancer*. 2019;19(3):151-161. doi:10.1038/s41568-019-0109-9

17. Pelham HRB. A regulatory upstream promoter element in the Drosophila Hsp 70 heat-shock gene. *Cell*. 1982;30(2):517-528. doi:10.1016/0092-8674(82)90249-5

18. Mitchell PJ, Tjian R. Transcriptional Regulation in Mammalian Cells by Sequence-Specific DNA Binding Proteins. *Science*. 1989;245(4916):371-378. doi:10.1126/science.2667136

19. Lambert SA, Jolma A, Campitelli LF, et al. The Human Transcription Factors. *Cell*. 2018;172(4):650-665. doi:10.1016/j.cell.2018.01.029

20. Bejerano G, Pheasant M, Makunin I, et al. Ultraconserved elements in the human genome. *Science*. 2004;304(5675):1321-1325. doi:10.1126/science.1098119

21. Barrera LA, Vedenko A, Kurland JV, et al. Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science*. 2016;351(6280):1450-1454. doi:10.1126/science.aad2257

22. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*. 1993;75(5):843-854. doi:10.1016/0092-8674(93)90529-Y

23. Bartel DP. Metazoan MicroRNAs. *Cell*. 2018;173(1):20-51. doi:10.1016/j.cell.2018.03.006

24. Gramzow L, Theißen G. Plant miRNA Conservation and Evolution. *Methods Mol Biol Clifton NJ*. 2019;1932:41-50. doi:10.1007/978-1-4939-9042-9_3

25. Groot M, Lee H. Sorting Mechanisms for MicroRNAs into Extracellular Vesicles and Their Associated Diseases. *Cells*. 2020;9(4):1044. doi:10.3390/cells9041044

26. Alles J, Fehlmann T, Fischer U, et al. An estimate of the total number of true human miRNAs. *Nucleic Acids Res*. 2019;47(7):3353-3364. doi:10.1093/nar/gkz097

27. Santulli G, ed. *MicroRNA: Medical Evidence: From Molecular Biology to Clinical Practice*. Vol 888. Springer International Publishing; 2015. doi:10.1007/978-3-319-22671-2

28. Dissanayake E, Inoue Y. MicroRNAs in Allergic Disease. *Curr Allergy Asthma Rep*. 2016;16(9):67. doi:10.1007/s11882-016-0648-z

29. Rupaimoole R, Slack FJ. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat Rev Drug Discov*. 2017;16(3):203-222. doi:10.1038/nrd.2016.246

30. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977;74(12):5463-5467. doi:10.1073/pnas.74.12.5463

31. Chang TW. Binding of cells to matrixes of distinct antibodies coated on solid surface. *J Immunol Methods*. 1983;65(1-2):217-223. doi:10.1016/0022-1759(83)90318-6

32. Mullis KB. The unusual origin of the polymerase chain reaction. *Sci Am*. 1990;262(4):56-61, 64-65. doi:10.1038/scientificamerican0490-56

33. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995;270(5235):467-470. doi:10.1126/science.270.5235.467

34. Kchouk M, Gibrat JF, Elloumi M. Generations of Sequencing Technologies: From First to Next Generation. *Biol Med*. 2017;09(03). doi:10.4172/0974-8369.1000395

35. Wenger AM, Peluso P, Rowell WJ, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*. 2019;37(10):1155-1162. doi:10.1038/s41587-019-0217-9

36. Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. *Science*. 2022;376(6588):44-53. doi:10.1126/science.abj6987

37. Escalera-Balsera A, Roman-Naranjo P, Lopez-Escamez JA. Systematic Review of Sequencing Studies and Gene Expression Profiling in Familial Meniere Disease. *Genes*. 2020;11(12):1414. doi:10.3390/genes11121414

38. Nirenberg M, Leder P, Bernfield M, et al. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proc Natl Acad Sci U S A*. 1965;53(5):1161-1168. doi:10.1073/pnas.53.5.1161

39. Yanofsky C. Establishing the triplet nature of the genetic code. *Cell*. 2007;128(5):815-818. doi:10.1016/j.cell.2007.02.029

40. Rother M, Krzycki JA. Selenocysteine, pyrrolysine, and the unique energy metabolism of methanogenic archaea. *Archaea Vanc BC*. 2010;2010:453642. doi:10.1155/2010/453642

41. Litwack G. Protein Biosynthesis. In: *Human Biochemistry*. Elsevier; 2022:357-375. doi:10.1016/B978-0-323-85718-5.00015-7

42. Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat Methods*. 2019;16(7):603-606. doi:10.1038/s41592-019-0437-4

43. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res*. 2019;47(D1):D520-D528. doi:10.1093/nar/gky949

44. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with

AlphaFold. *Nature*. 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2

45. David A, Islam S, Tankhilevich E, Sternberg MJE. The AlphaFold Database of Protein Structures: A Biologist's Guide. *J Mol Biol*. 2022;434(2):167336. doi:10.1016/j.jmb.2021.167336

46. Pandey A, Mann M. Proteomics to study genes and genomes. *Nature*. 2000;405(6788):837-846. doi:10.1038/35015709

47. Aslam B, Basit M, Nisar MA, Khurshid M, Rasool MH. Proteomics: Technologies and Their Applications. *J Chromatogr Sci*. 2017;55(2):182-196. doi:10.1093/chromsci/bmw167

48. Cui M, Cheng C, Zhang L. High-throughput proteomics: a methodological mini-review. *Lab Invest*. 2022;102(11):1170-1181. doi:10.1038/s41374-022-00830-7

49. Zubair M, Wang J, Yu Y, et al. Proteomics approaches: A review regarding an importance of proteome analyses in understanding the pathogens and diseases. *Front Vet Sci*. 2022;9:1079359. doi:10.3389/fvets.2022.1079359

50. Costantino S, Paneni F. The Epigenome in Atherosclerosis. In: von Eckardstein A, Binder CJ, eds. *Prevention and Treatment of Atherosclerosis*. Vol 270. Handbook of Experimental Pharmacology. Springer International Publishing; 2020:511-535. doi:10.1007/164_2020_422

51. Singh D, Nishi K, Khambata K, Balasinor NH. Introduction to epigenetics: basic concepts and advancements in the field. In: *Epigenetics and Reproductive Health*. Elsevier; 2020:xxv-xliv. doi:10.1016/B978-0-12-819753-0.02001-8

52. Harrison A, Parle-McDermott A. DNA methylation: a timeline of methods and applications. *Front Genet*. 2011;2:74. doi:10.3389/fgene.2011.00074

53. Searle B, Müller M, Carell T, Kellett A. Third‐Generation Sequencing of Epigenetic DNA. *Angew Chem Int Ed*. Published online January 25, 2023. doi:10.1002/anie.202215704

54. Mellén M, Ayata P, Heintz N. 5-hydroxymethylcytosine accumulation in postmitotic neurons results in functional demethylation of expressed genes. *Proc Natl Acad Sci*. 2017;114(37). doi:10.1073/pnas.1708044114

55. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. *PLOS Comput Biol*. 2017;13(5):e1005457. doi:10.1371/journal.pcbi.1005457

56. Chambers DC, Carew AM, Lukowski SW, Powell JE. Transcriptomics and single‐cell RNA‐sequencing. *Respirology*. 2019;24(1):29-36. doi:10.1111/resp.13412

57. Kulkarni A, Anderson AG, Merullo DP, Konopka G. Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Curr Opin Biotechnol*. 2019;58:129-136. doi:10.1016/j.copbio.2019.03.001

10. REFERENCES

58. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3:Article3. doi:10.2202/1544-6115.1027

59. Robinson MD, McCarthy DJ, Smyth GK. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140. doi:10.1093/bioinformatics/btp616

60. 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. doi:10.1186/s13059-014-0550-8

61. Lister R, Pelizzola M, Dowen RH, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*. 2009;462(7271):315-322. doi:10.1038/nature08514

62. Dolzhenko E, Smith AD. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics*. 2014;15(1):215. doi:10.1186/1471-2105-15-215

63. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol*. 2012;13(10):R83. doi:10.1186/gb-2012-13-10-r83

64. Baik B, Yoon S, Nam D. Benchmarking RNA-seq differential expression analysis methods using spike-in and simulation data. *PloS One*. 2020;15(4):e0232271. doi:10.1371/journal.pone.0232271

65. Dowell JA, Wright LJ, Armstrong EA, Denu JM. Benchmarking Quantitative Performance in Label-Free Proteomics. *ACS Omega*. 2021;6(4):2494-2504. doi:10.1021/acsomega.0c04030

66. Piao Y, Xu W, Park KH, Ryu KH, Xiang R. Comprehensive Evaluation of Differential Methylation Analysis Methods for Bisulfite Sequencing Data. *Int J Environ Res Public Health*. 2021;18(15):7975. doi:10.3390/ijerph18157975

67. Toro-Domínguez D, Martorell-Marugán J, Martinez-Bueno M, et al. Scoring personalized molecular portraits identify Systemic Lupus Erythematosus subtypes and predict individualized drug responses, symptomatology and disease progression. *Brief Bioinform*. 2022;23(5):bbac332. doi:10.1093/bib/bbac332

68. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9(1):559. doi:10.1186/1471-2105-9-559

69. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*. 2009;25(22):2906-2912. doi:10.1093/bioinformatics/btp543

70. Mo Q, Wang S, Seshan VE, et al. Pattern discovery and cancer gene identification in

integrated cancer genomic data. *Proc Natl Acad Sci U S A*. 2013;110(11):4245-4250. doi:10.1073/pnas.1208949110

71. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*. 2018;19(1):71-86. doi:10.1093/biostatistics/kxx017

72. Chauvel C, Novoloaca A, Veyre P, Reynier F, Becker J. Evaluation of integrative clustering methods for the analysis of multi-omics data. *Brief Bioinform*. 2020;21(2):541-552. doi:10.1093/bib/bbz015

73. Alashwal H, El Halaby M, Crouse JJ, Abdalla A, Moustafa AA. The Application of Unsupervised Clustering Methods to Alzheimer's Disease. *Front Comput Neurosci*. 2019;13:31. doi:10.3389/fncom.2019.00031

74. Lan K, Wang D tong, Fong S, Liu L sheng, Wong KKL, Dey N. A Survey of Data Mining and Deep Learning in Bioinformatics. *J Med Syst*. 2018;42(8):139. doi:10.1007/s10916-018-1003-9

75. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res*. 2019;47(D1):D330-D338. doi:10.1093/nar/gky1055

76. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res*. 2019;47(D1):D590-D595. doi:10.1093/nar/gky962

77. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009;37(1):1-13. doi:10.1093/nar/gkn923

78. Drăghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA. Global functional profiling of gene expression☆☆This work was funded in part by a Sun Microsystems grant awarded to S.D., NIH Grant HD36512 to S.A.K., a Wayne State University SOM Dean's Post-Doctoral Fellowship, and an NICHD Contraception and Infertility Loan to G.C.O. Support from the WSU MCBI mode is gratefully appreciated. *Genomics*. 2003;81(2):98-104. doi:10.1016/S0888-7543(02)00021-6

79. Sherman BT, Hao M, Qiu J, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*. 2022;50(W1):W216-W221. doi:10.1093/nar/gkac194

80. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57. doi:10.1038/nprot.2008.211

81. Mi H, Ebert D, Muruganujan A, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res*. 2021;49(D1):D394-D403. doi:10.1093/nar/gkaa1106

82. Raudvere U, Kolberg L, Kuzmin I, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res*. 2019;47(W1):W191-W198. doi:10.1093/nar/gkz369

83. Kuleshov MV, Diaz JEL, Flamholz ZN, et al. modEnrichr: a suite of gene set enrichment analysis tools for model organisms. *Nucleic Acids Res*. 2019;47(W1):W183-W190. doi:10.1093/nar/gkz347

84. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003;34(3):267-273. doi:10.1038/ng1180

85. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci*. 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102

86. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. *Fast Gene Set Enrichment Analysis*. Bioinformatics; 2016. doi:10.1101/060012

87. Adrian Alexa JR. topGO. Published online 2017. doi:10.18129/B9.BIOC.TOPGO

88. Mi H, Muruganujan A, Huang X, et al. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc*. 2019;14(3):703-721. doi:10.1038/s41596-019-0128-8

89. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L, Mi H. PANTHER : Making genome‐scale phylogenetics accessible to all. *Protein Sci*. 2022;31(1):8-22. doi:10.1002/pro.4218

90. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. Gibas C, ed. *PLoS ONE*. 2011;6(7):e21800. doi:10.1371/journal.pone.0021800

91. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol*. 2007;8(1):R3. doi:10.1186/gb-2007-8-1-r3

92. Garcia-Moreno A, López-Domínguez R, Villatoro-García JA, et al. Functional Enrichment Analysis of Regulatory Elements. *Biomedicines*. 2022;10(3):590. doi:10.3390/biomedicines10030590

93. Yang Y, Chen L, Gu J, et al. Recurrently deregulated lncRNAs in hepatocellular carcinoma. *Nat Commun*. 2017;8(1):14421. doi:10.1038/ncomms14421

94. Geeleher P, Hartnett L, Egan LJ, Golden A, Raja Ali RA, Seoighe C. Gene-set analysis is severely biased when applied to genome-wide methylation data. *Bioinforma Oxf Engl*. 2013;29(15):1851-1857. doi:10.1093/bioinformatics/btt311

95. Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes Dev*. 2011;25(10):1010-1022. doi:10.1101/gad.2037511

96. Paweł K, Maria Małgorzata S. CpG Island Methylator Phenotype—A Hope for the Future or a Road to Nowhere? *Int J Mol Sci*. 2022;23(2):830. doi:10.3390/ijms23020830

97. Godard P, van Eyll J. Pathway analysis from lists of microRNAs: common pitfalls and alternative strategy. *Nucleic Acids Res*. 2015;43(7):3490-3497. doi:10.1093/nar/gkv249

98. Bleazard T, Lamb JA, Griffiths-Jones S. Bias in microRNA functional enrichment analysis. *Bioinforma Oxf Engl*. 2015;31(10):1592-1598. doi:10.1093/bioinformatics/btv023

99. Mork S, Pletscher-Frankild S, Palleja Caro A, Gorodkin J, Jensen LJ. Protein-driven inference of miRNA-disease associations. *Bioinformatics*. 2014;30(3):392-397. doi:10.1093/bioinformatics/btt677

100. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol*. 2010;11(2):R14. doi:10.1186/gb-2010-11-2-r14

101. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics*. 2016;32(2):286-288. doi:10.1093/bioinformatics/btv560

102. Maksimovic J, Oshlack A, Phipson B. Gene set enrichment analysis for genome-wide DNA methylation data. *Genome Biol*. 2021;22(1):173. doi:10.1186/s13059-021-02388-x

103. Dong D, Tian Y, Zheng SC, Teschendorff AE. ebGSEA: an improved Gene Set Enrichment Analysis method for Epigenome-Wide-Association Studies. Birol I, ed. *Bioinformatics*. 2019;35(18):3514-3516. doi:10.1093/bioinformatics/btz073

104. Li J, Han X, Wan Y, et al. TAM 2.0: tool for MicroRNA set analysis. *Nucleic Acids Res*. 2018;46(W1):W180-W185. doi:10.1093/nar/gky509

105. Huang Z, Shi J, Gao Y, et al. HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res*. 2019;47(D1):D1013-D1017. doi:10.1093/nar/gky1010

106. Ning L, Cui T, Zheng B, et al. MNDR v3.0: mammal ncRNA-disease repository with increased coverage and annotation. *Nucleic Acids Res*. 2021;49(D1):D160-D164. doi:10.1093/nar/gkaa707

107. Fan Y, Siklenka K, Arora SK, Ribeiro P, Kimmins S, Xia J. miRNet - dissecting miRNA-target interactions and functional associations through network-based visual analysis. *Nucleic Acids Res*. 2016;44(W1):W135-141. doi:10.1093/nar/gkw288

108. Kern F, Fehlmann T, Solomon J, et al. miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. *Nucleic Acids Res*. 2020;48(W1):W521-W528. doi:10.1093/nar/gkaa309

109. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res*. 2019;47(D1):D155-D162. doi:10.1093/nar/gky1141

110. Fromm B, Høye E, Domanska D, et al. MirGeneDB 2.1: toward a complete sampling of all major animal phyla. *Nucleic Acids Res*. 2022;50(D1):D204-D210. doi:10.1093/nar/gkab1101

111. Riffo-Campos Á, Riquelme I, Brebi-Mieville P. Tools for Sequence-Based miRNA Target Prediction: What to Choose? *Int J Mol Sci*. 2016;17(12):1987. doi:10.3390/ijms17121987

112. Peterson SM, Thompson JA, Ufkin ML, Sathyanarayana P, Liaw L, Congdon CB. Common features of microRNA target prediction tools. *Front Genet*. 2014;5. doi:10.3389/fgene.2014.00023

113. M. Witkos T, Koscianska E, J. Krzyzosiak W. Practical Aspects of microRNA Target Prediction. *Curr Mol Med*. 2011;11(2):93-109. doi:10.2174/156652411794859250

114. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*. 2015;4:e05005. doi:10.7554/eLife.05005

115. Wang X. Composition of seed sequence is a major determinant of microRNA targeting patterns. *Bioinformatics*. 2014;30(10):1377-1383. doi:10.1093/bioinformatics/btu045

116. Liu W, Wang X. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome Biol*. 2019;20(1):18. doi:10.1186/s13059-019-1629-z

117. Ding J, Li X, Hu H. TarPmiR: a new approach for microRNA target site prediction. *Bioinformatics*. 2016;32(18):2768-2775. doi:10.1093/bioinformatics/btw318

118. Paraskevopoulou MD, Georgakilas G, Kostoulas N, et al. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res*. 2013;41(W1):W169-W173. doi:10.1093/nar/gkt393

119. Le TD, Zhang J, Liu L, Li J. Ensemble Methods for MiRNA Target Prediction from Expression Data. Mari B, ed. *PLOS ONE*. 2015;10(6):e0131627. doi:10.1371/journal.pone.0131627

120. Quillet A, Saad C, Ferry G, et al. Improving Bioinformatics Prediction of microRNA Targets by Ranks Aggregation. *Front Genet*. 2020;10:1330. doi:10.3389/fgene.2019.01330

121. Huang HY, Lin YCD, Cui S, et al. miRTarBase update 2022: an informative resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res*. 2022;50(D1):D222-D230. doi:10.1093/nar/gkab1079

122. Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, et al. DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. *Nucleic Acids Res*. 2018;46(D1):D239-D245. doi:10.1093/nar/gkx1141

123. Xie B, Ding Q, Han H, Wu D. miRCancer: a microRNA-cancer association database constructed by text mining on literature. *Bioinforma Oxf Engl*. 2013;29(5):638-644.

doi:10.1093/bioinformatics/btt014

124. Guo Z, Kuang Z, Zhao Y, et al. PmiREN2.0: from data annotation to functional exploration of plant microRNAs. *Nucleic Acids Res*. 2022;50(D1):D1475-D1482. doi:10.1093/nar/gkab811

125. RNAcentral Consortium. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res*. 2021;49(D1):D212-D220. doi:10.1093/nar/gkaa921

126. De Palma FDE, Carbonnier V, Salvatore F, Kroemer G, Pol JG, Maiuri MC. Systematic Investigation of the Diagnostic and Prognostic Impact of LINC01087 in Human Cancers. *Cancers*. 2022;14(23):5980. doi:10.3390/cancers14235980

127. Li Z, Zhang Y, Fang J, et al. NcPath: a novel platform for visualization and enrichment analysis of human non-coding RNA and KEGG signaling pathways. Kendziorski C, ed. *Bioinformatics*. 2023;39(1):btac812. doi:10.1093/bioinformatics/btac812

128. Yang J, Li C, Chi S, Wei H, Du W, Hu Q. Upregulation of microRNA‐762 suppresses the expression of GIPC3 in systemic lupus erythematosus and neuropsychiatric systemic lupus erythematosus. *Immun Inflamm Dis*. 2022;10(11). doi:10.1002/iid3.719

129. Dweep H, Showe LC, Kossenkov AV. Functional Annotation of MicroRNAs Using Existing Resources. In: Allmer J, Yousef M, eds. *MiRNomics*. Vol 2257. Methods in Molecular Biology. Springer US; 2022:57-77. doi:10.1007/978-1-0716-1170-8_3

130. Licursi V, Conte F, Fiscon G, Paci P. MIENTURNET: an interactive web tool for microRNA-target enrichment and network-based analysis. *BMC Bioinformatics*. 2019;20(1):545. doi:10.1186/s12859-019-3105-x

131. Preusse M, Theis FJ, Mueller NS. miTALOS v2: Analyzing Tissue Specific microRNA Function. *PloS One*. 2016;11(3):e0151771. doi:10.1371/journal.pone.0151771

132. Zagganas K, Vergoulis T, Paraskevopoulou MD, Vlachos IS, Skiadopoulos S, Dalamagas T. BUFET: boosting the unbiased miRNA functional enrichment analysis using bitsets. *BMC Bioinformatics*. 2017;18(1):399. doi:10.1186/s12859-017-1812-8

133. Vlachos IS, Zagganas K, Paraskevopoulou MD, et al. DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res*. 2015;43(W1):W460-466. doi:10.1093/nar/gkv403

134. Chang L, Zhou G, Soufan O, Xia J. miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. *Nucleic Acids Res*. 2020;48(W1):W244-W251. doi:10.1093/nar/gkaa467

135. Lu M, Shi B, Wang J, Cao Q, Cui Q. TAM: A method for enrichment and depletion analysis of a microRNA category in a list of microRNAs. *BMC Bioinformatics*. 2010;11(1):419. doi:10.1186/1471-2105-11-419

136. Köhler S, Carmody L, Vasilevsky N, et al. Expansion of the Human Phenotype Ontology

(HPO) knowledge base and resources. *Nucleic Acids Res*. 2019;47(D1):D1018-D1027. doi:10.1093/nar/gky1105

137. Huntley RP, Sitnikov D, Orlic-Milacic M, et al. Guidelines for the functional annotation of microRNAs using the Gene Ontology. *RNA*. 2016;22(5):667-676. doi:10.1261/rna.055301.115

138. Huntley RP, Kramarz B, Sawford T, et al. Expanding the horizons of microRNA bioinformatics. *RNA*. 2018;24(8):1005-1017. doi:10.1261/rna.065565.118

139. Garcia-Moreno A, Carmona-Saez P. Computational Methods and Software Tools for Functional Analysis of miRNA Data. *Biomolecules*. 2020;10(9). doi:10.3390/biom10091252

140. Nogales-Cadenas R, Carmona-Saez P, Vazquez M, et al. GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Res*. 2009;37(Web Server issue):W317-322. doi:10.1093/nar/gkp416

141. Tabas-Madrid D, Nogales-Cadenas R, Pascual-Montano A. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic Acids Res*. 2012;40(Web Server issue):W478-483. doi:10.1093/nar/gks402

142. Wu F, Luo K, Yan Z, et al. Analysis of miRNAs and their target genes in five Melilotus albus NILs with different coumarin content. *Sci Rep*. 2018;8(1):14138. doi:10.1038/s41598-018-32153-3

143. Fog A. Calculation Methods for Wallenius' Noncentral Hypergeometric Distribution. *Commun Stat - Simul Comput*. 2008;37(2):258-273. doi:10.1080/03610910701790269

144. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B Methodol*. 1995;57(1):289-300. Accessed January 18, 2021. http://www.jstor.org/stable/2346101

145. Timmons JA, Szkop KJ, Gallagher IJ. Multiple sources of bias confound functional enrichment analysis of global -omics data. *Genome Biol*. 2015;16(1):186. doi:10.1186/s13059-015-0761-7

146. Agrawal, Rakesh, Ramakrishnan Srikant. Fast algorithms for mining association rules. *Proc 20th Int Conf Very Large Data Bases VLDB*. 1994;1215.

147. Huang HY, Lin YCD, Li J, et al. miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res*. 2020;48(D1):D148-D154. doi:10.1093/nar/gkz896

148. 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;29(8):1363-1375. doi:10.1101/gr.240663.118

149. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27-30. doi:10.1093/nar/28.1.27

150. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25-29. doi:10.1038/75556

151. Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE, Mouse Genome Database Group. Mouse Genome Database (MGD) 2019. *Nucleic Acids Res*. 2019;47(D1):D801-D806. doi:10.1093/nar/gky1056

152. Huang R, Grishagin I, Wang Y, et al. The NCATS BioPlanet – An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and Chemical Genomics. *Front Pharmacol*. 2019;10. doi:10.3389/fphar.2019.00445

153. Jassal B, Matthews L, Viteri G, et al. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2020;48(D1):D498-D503. doi:10.1093/nar/gkz1031

154. Martens M, Ammar A, Riutta A, et al. WikiPathways: connecting communities. *Nucleic Acids Res*. 2021;49(D1):D613-D621. doi:10.1093/nar/gkaa1024

155. Thorn CF, Klein TE, Altman RB. PharmGKB: the Pharmacogenomics Knowledge Base. *Methods Mol Biol Clifton NJ*. 2013;1015:311-320. doi:10.1007/978-1-62703-435-7_20

156. Davis AP, Grondin CJ, Johnson RJ, et al. Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Res*. 2021;49(D1):D1138-D1143. doi:10.1093/nar/gkaa891

157. Stathias V, Turner J, Koleti A, et al. LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. *Nucleic Acids Res*. 2020;48(D1):D431-D439. doi:10.1093/nar/gkz1023

158. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. 2015;43(D1):D789-D798. doi:10.1093/nar/gku1205

159. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res*. 2020;48(D1):D845-D855. doi:10.1093/nar/gkz1021

160. Yoo AB, Jette MA, Grondona M. SLURM: Simple Linux Utility for Resource Management. In: Feitelson D, Rudolph L, Schwiegelshohn U, eds. *Job Scheduling Strategies for Parallel Processing*. Vol 2862. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2003:44-60. doi:10.1007/10968987_3

161. Aparicio-Puerta E, Gómez-Martín C, Giannoukakos S, et al. sRNAbench and sRNAtoolbox 2022 update: accurate miRNA and sncRNA profiling for model and non-model organisms. *Nucleic Acids Res*. 2022;50(W1):W710-W717. doi:10.1093/nar/gkac363

162. Raschka S. MLxtend: Providing machine learning and data science utilities and

extensions to Python's scientific computing stack. *J Open Source Softw*. 2018;3(24):638. doi:10.21105/joss.00638

163. Servén daniel, Brummitt C, Abedi H, hlink. dswah/pyGAM: v0.8.0. Published online October 2018. doi:10.5281/zenodo.1476122

164. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261-272. doi:10.1038/s41592-019-0686-2

165. Team TPD. pandas-dev/pandas: Pandas. Published online March 16, 2023. doi:10.5281/ZENODO.3509134

166. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357-362. doi:10.1038/s41586-020-2649-2

167. Kim GH. MicroRNA Regulation of Cardiac Conduction and Arrhythmias. *Transl Res J Lab Clin Med*. 2013;161(5):381-392. doi:10.1016/j.trsl.2012.12.004

168. Han H, Cho JW, Lee S, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res*. 2018;46(D1):D380-D386. doi:10.1093/nar/gkx1013

169. Schriml LM, Munro JB, Schor M, et al. The Human Disease Ontology 2022 update. *Nucleic Acids Res*. 2022;50(D1):D1255-D1261. doi:10.1093/nar/gkab1063

170. Li S, Wan C, Zheng R, et al. Cistrome-GO: a web server for functional enrichment analysis of transcription factor ChIP-seq peaks. *Nucleic Acids Res*. 2019;47(W1):W206-W211. doi:10.1093/nar/gkz332

171. Magnusson R, Lubovac-Pilav Z. TFTenricher: a python toolbox for annotation enrichment analysis of transcription factor target genes. *BMC Bioinformatics*. 2021;22(1):440. doi:10.1186/s12859-021-04357-4

172. Jayaram N, Usvyat D, R. Martin AC. Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*. 2016;17(1):547. doi:10.1186/s12859-016-1298-9

173. Essebier A, Lamprecht M, Piper M, Bodén M. Bioinformatics approaches to predict target genes from transcription factor binding data. *Methods*. 2017;131:111-119. doi:10.1016/j.ymeth.2017.09.001

174. Lim H, Xie L. Target Gene Prediction of Transcription Factor Using a New Neighborhood-regularized Tri-factorization One-class Collaborative Filtering Algorithm. *ACM-BCB ACM Conf Bioinforma Comput Biol Biomed ACM Conf Bioinforma Comput Biol Biomed*. 2018;2018:1-10. doi:10.1145/3233547.3233551

175. Hammal F, de Langen P, Bergon A, Lopez F, Ballester B. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res*. 2022;50(D1):D316-D325. doi:10.1093/nar/gkab996

176. Kulakovskiy IV, Vorontsov IE, Yevshin IS, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res*. 2018;46(D1):D252-D259. doi:10.1093/nar/gkx1106

177. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2022;50(D1):D165-D173. doi:10.1093/nar/gkab1113

178. Margolin AA, Nemenman I, Basso K, et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*. 2006;7(S1):S7. doi:10.1186/1471-2105-7-S1-S7

179. Alvarez MJ, Shen Y, Giorgi FM, et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet*. 2016;48(8):838-847. doi:10.1038/ng.3593

180. Carithers LJ, Ardlie K, Barcus M, et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation Biobanking*. 2015;13(5):311-319. doi:10.1089/bio.2015.0032

181. Lopez-Dominguez R, Toro-Dominguez D, Martorell-Marugan J, et al. Transcription Factor Activity Inference in Systemic Lupus Erythematosus. *Life*. 2021;11(4):299. doi:10.3390/life11040299

182. Garcia-Alonso L, Iorio F, Matchan A, et al. Transcription Factor Activities Enhance Markers of Drug Sensitivity in Cancer. *Cancer Res*. 2018;78(3):769-780. doi:10.1158/0008-5472.CAN-17-1679

183. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. 2016;44(W1):W90-W97. doi:10.1093/nar/gkw377

184. Lu TP, Lee CY, Tsai MH, et al. miRSystem: an integrated system for characterizing enriched functions and pathways of microRNA targets. *PloS One*. 2012;7(8):e42390. doi:10.1371/journal.pone.0042390

185. Ren X, Kuan PF. methylGSA: a Bioconductor package and Shiny app for DNA methylation data length bias adjustment in gene set testing. *Bioinforma Oxf Engl*. 2019;35(11):1958-1959. doi:10.1093/bioinformatics/bty892

11. Scientific Production

This epigraph is dedicated to show a relation of the doctoral candidate's scientific production during the time of this thesis research, including original author works and research collaborations out of the scope of the thesis.

11.1. Articles with thesis results

1. **Garcia-Moreno, A.**; Carmona-Saez, P. Computational Methods and Software Tools for Functional Analysis of miRNA Data. Biomolecules 2020, 10, 1252. https://doi.org/10.3390/biom10091252
2. **Garcia-Moreno, A.**; López-Domínguez, R.; Villatoro-García, J.A.; Ramirez-Mena, A.; Aparicio-Puerta, E.; Hackenberg, M.; Pascual-Montano, A.; Carmona-Saez, P. Functional Enrichment Analysis of Regulatory Elements. Biomedicines 2022, 10, 590. https://doi.org/10.3390/biomedicines10030590
3. **Garcia-Moreno, A.**; López-Domínguez, R.; Villatoro-García, J.A.; Carmona-Saez, P.; Transcription factors targets enrichment analysis: assessment of biassed results and an alternative approach. In preparation.

11.2. Co-authorship in collaborations

1. Rivero, I.; Gorines-Cordero, G.J.; Rubio-Rodríguez, L.A.; Soler-Sáez, I.; Perpiñá-Clérigues, C.; **García-Moreno, A**.; Monzón, S.; Hernández-Beeftink, T. ISCB RSG-Spain and Highlights from the VIII Spanish Student Symposium in Bioinformatics and Computational Biology in 2021; Scientific Communication and Education, 2022; https://doi.org/10.1101/2022.08.19.504447
2. Martorell-Marugán J,; Villatoro-García JA,; **García-Moreno A,;** López-Domínguez R,; Requena F,; Merelo JJ,; Lacasaña M,; de Dios Luna J,; Díaz-Mochón JJ,; Lorente JA,; Carmona-Sáez P. DatAC: A visual analytics platform to explore climate and air quality indicators associated with the COVID-19 pandemic in Spain. Sci Total Environ. 2021 Jan 1;750:141424. https://doi.org/10.1016/j.scitotenv.2020.141424
3. Lopez-Dominguez, R.; Toro-Dominguez, D.; Martorell-Marugan, J.; **Garcia-Moreno,**

A.; Holland, C.H.; Saez-Rodriguez, J.; Goldman, D.; Petri, M.A.; Alarcon-Riquelme, M.E.; Carmona-Saez, P. Transcription Factor Activity Inference in Systemic Lupus Erythematosus. Life 2021, 11, 299. https://doi.org/10.3390/life11040299

4. Martorell-Marugán, J.; López-Domínguez, R.; **García-Moreno, A.**; Toro-Domínguez, D.; Villatoro-García, J.A.; Barturen, G.; Martín-Gómez, A.; Troule, K.; Gómez-López, G.; Al-Shahrour, F.; et al. A Comprehensive Database for Integrated Analysis of Omics Data in Autoimmune Diseases. BMC Bioinformatics 2021, 22, 343, https://doi.org/10.1186/s12859-021-04268-4
5. Aparicio-Puerta, E.; Gómez-Martín, C.; Giannoukakos, S.; Medina, J.M.; Scheepbouwer, C.; **García-Moreno, A.**; Carmona-Saez, P.; Fromm, B.; Pegtel, M.; Keller, A.; et al. SRNAbench and SRNAtoolbox 2022 Update: Accurate MiRNA and SncRNA Profiling for Model and Non-Model Organisms. Nucleic Acids Research 2022, 50, W710–W717, https://doi.org/10.1093/nar/gkac363.