UNIVERSIDAD AUTÓNOMA DE MADRID

Departamento de Biología Molecular
Facultad de Ciencias

Analysis and regulation of alternative splicing in the pig infarcted myocardium

Girolamo Giudice

Madrid, 2017

UNIVERSIDAD AUTÓNOMA DE MADRID

Departamento de Biología Molecular
Facultad de Ciencias

Analysis and regulation of alternative splicing in the pig infarcted myocardium

Girolamo Giudice

Madrid, 2017

Departamento de Biología Moleclular

Facultad de Ciencias

Universidad Autónoma de Madrid

Analysis and regulation of alternative splicing in the pig infarcted myocardium

Girolamo Giudice, M.Sc.

Enrique Lara-Pezzi, Ph.D.

Carlos Torroja Fungairiño, Ph.D.

Fundación Centro Nacional

de Investigaciones Cardiovasculares

Acknowledgement

It is difficult to summarize in a few lines my sincere gratitude towards people who have helped me during these years. First, I would like to thank my supervisor Dr. Enrique Lara-Pezzi, for giving me the opportunity to join his team and participate into the CardioNext project. For me, Dr. Enrique Lara-Pezzi represents the true personification of an ideal supervisor, he provided me with a huge amount of help, support, encouragement and advice and his door was always open also during holidays.

Many thanks go to all members, old and new, of the Bioinformatics unit at CNIC, who share some time, conversations, and experiences all these years I have spent in Madrid. Among them, I am particularly, grateful to Fàtima and Carlos for their guidance and mentorship. The advices and suggestions I received from them, I am sure will help me beyond the completion of my thesis.

My colleagues of the CardioNext project, among them Wencke and Navratan, who made my coffee break enjoyable.

I am very grateful to my family for their love, constant support, and encouragement to pursue this challenge. Despite the long distances between us, you always managed to make me feel loved.

I would like to thank the most important person whom contribution to this thesis is priceless: Cosima. You deserve this degree as much as I do. You have put up with a lot in the past three years. Thank you so much for your unlimited help and patience

Finally, I would like to thank the Marie-Curie Early Stage-researcher programme to provide financial support through this research fellowship.

Summary

In all eukaryotes, pre-mRNA splicing is a necessary step for protein synthesis. Pre-mRNA splicing is mediated by the spliceosome, a dynamic complex of 100/150 proteins and RNAs, which is assembled de novo at every splicing event. The same pre-mRNA can be spliced in different ways generating different isoforms that potentially regulate gene expression or affect protein structure, function, or localisation. RNA binding proteins, finely regulate this process by recognising short sequences located in the exons or in the neighbour introns that favour the inclusion or the exclusion of an exon in the mature mRNA.

The advent of RNA sequencing has allowed the investigation of the splicing patterns and concurrently the quantification of mRNA abundance underlying a phenotype. Currently, the interpretation of the changes in the splicing patterns and their regulation represents one of the major challenges in biomedical research.

To collect in a unique resource all the regulatory elements and the RNA-binding proteins that were spread across different repositories, we developed a database named ATtRACT. Subsequently, we investigated the impact of alternative splicing in the pig infarcted myocardium. The analysis focused on four different cell types, cardiomyocytes, fibroblasts, endothelial cells, and macrophages at three different time points before and after infarction. We employed two distinct pipelines that allowed us to detect isoform switches after myocardial infarction, and to identify, thanks to the data available in ATtRACT, the RNA binding proteins and associated motifs that potentially regulate the alternative splicing.

The analysis of the biological impact of large gene/protein sets is currently an open issue in bioinformatics. A widely-adopted procedure involves the enrichment analysis. Here, we developed a new functional enrichment analysis method, called MAGNETO. The main aim of MAGNETO was to gain more insight into of the biological processes underlying a phenotype by employing protein-protein interaction networks.

Our results show that ATtRACT represents an invaluable resource for all the researchers involved in the study of RNA binding proteins. ATtRACT adds 192 motifs not available in any other database by retrieving the information buried in the Protein Data Bank. Two hundred and ten genes were detected as alternatively spliced after myocardial infarction in cardiomyocytes, fibroblast, endothelial cells, and macrophages. Forty-one were also detected at the protein level. Moreover, our findings suggest that alternatively spliced isoforms could be of great use to develop novel therapeutic approaches to avoid adverse cardiac remodelling and improve cardiac repair. Finally, our results confirm that MAGNETO allows to gain more overrepresented terms with respect to the standard enrichment analysis and to generate new hypotheses on the phenotype under investigation.

Resumen

En eucariotas, el *splicing* de pre-ARN mensajero (pre-ARNm) es un paso indispensable en la producción de proteínas. Este proceso está mediado por el Spliceosoma, un complejo de unas 100-150 proteínas y ARNs. Un mismo pre-ARNm puede ser procesado de manera diferente generando isoformas alternativas que pueden alterar el nivel de expresión del gen o la estructura de la proteína y por consiguiente su función y/o localización. Las proteínas de unión a ARN regulan de manera muy precisa este proceso mediante el reconocimiento de secuencias de nucleótidos específicas (motivos) presentes a lo largo de la molécula de ARN mensajero que a la postre definirán la presencia de exones e intrones y permitirán modular la inclusión o exclusión de los exones en el ARN maduro.

La llegada de las nuevas tecnologías de secuenciación masiva (NGS) ha permitido un estudio más exhaustivo y cuantitativo de los patrones de *splicing* presentes en una condición fenotípica dada. Actualmente, la interpretación de los cambios en los patrones de *splicing* y la regulación de los mismos es uno de los mayores desafíos en la investigación biomédica.

Para ayudar en el estudio del AS es necesario tener un acceso fácil y exhaustivo a la información disponible sobre las proteínas reguladoras y sus motivos. Actualmente esta información se encuentra repartida in distintas bases de datos con distintas estructuras y de difícil acceso. Por ello he desarrollado ATtRACT, una base de datos con acceso vía web, que recopila y estructura toda esa información en un solo repositorio. Además en ATtRACT he añadido 192 motivos más obtenidos de las estructuras cristalográficas almacenadas en el “Protein Data Bank” (PDB) que no están disponibles en ninguna otra base de datos.

La interpretación del impacto biológico que un grupo de genes, que cambian en una condición dada, es a día de hoy un problema fundamental y sin resolver. Actualmente el método mas empleado para ayudar en esta interpretación en el análisis de enriquecimiento en términos ontológicos. Este análisis es muy simple y sufre de varías carencias como no englobarse en el contexto biológico y no ser eficiente cuando el número de genes es bajo. Por ello he desarrollado MAGNETO, un nuevo método de análisis de enriquecimiento que incorpora la información disponible sobre las interacciones proteína-proteína, las vías de señalización donde están implicadas y el tejido sobre el cual se han obtenido los datos. Nuestros resultados muestran que MAGNETO es capaz de obtener más términos sobrerrepresentados y más específicos que los métodos tradicionales, especialmente cuando el número de genes en la lista es bajo (~50 genes). MAGNETO permite profundizar de manera más precisa y eficiente en los procesos biológicos afectados en una condición fenotípica dada.

A continuación hemos estudiado el impacto del *splicing* alternativo en el infarto de miocardio en cerdo. El análisis se ha focalizado en cuatro tipos celulares, cardiomiocitos, fibroblastos, células endoteliales y macrófagos a tres tiempos (0, 3 y 7 días) después del infarto. He empleado dos métodos de detección de cambios en isoformas de ARNm, a partir de datos de *RNA-seq*, que han permitido identificar 210 genes que cambian las isoformas de ARNm que producen debido al infarto de miocardio y durante su recuperación en los cuatro tipos celulares. Gracias a ATtRACT hemos podido profundizar en los mecanismos que regulan esos cambios identificando las posibles proteínas reguladoras y sus motivos de unión. De esos 210 cambios, 41 han sido detectados también a nivel de proteína. También MAGNETO ha ayudado en la interpretación de los cambios biológicos que sufren los distintos tipos celulares durante el proceso de recuperación del infarto de miocardio mediante el análisis de las 50 proteínas más expresadas en cada tipo celular y a cada tiempo.

Nuestros resultados sugieren que las isoformas alternativas podrían ser una nueva diana terapéutica para prevenir un remodelado cardíaco adverso y mejorar la reparación del miocardio.

Index of contents

Abbreviations

ACVR2B	Activin Type II receptor
APSP	All-pairs shortest path
AS	Alternative splicing
BP	Biological Process
BWT	Burrows-Wheeler transform
CC	Cellular Component
CGM	Coarse-grained module
CHD	Coronary heart disease
CM	Cardiomyocyte
CVD	Cardiovascular disease
DSD	Diffusion state distance
EC	Endothelial cell
ECM	Extracellular matrix
ESE	Exonic splicing enhancer
ESS	Exonic splicing silencer
FACS	Fluorescence-activated cell sorting
FB	Fibroblast
FH2	Formin homology 2 domain
FST	Follistatin
GO	Gene Ontology
GOP	GO profile
HPA	Human Protein Atlas
HTS	High-throughput sequencing
ISE	Intronic splicing enhancer
ISS	Intronic splicing silencer
JSD	Jansen-Shannon distance
KH	K Homology
MAPK	Mitogen-activated protein kinase
MCL	Markov clustering

MCN	Minimal Connecting Network
MCODE	Molecular complex detection
MFu	Molecular Function
MF	Macrophage
MI	Myocardial infarction
NBEA	Network-based enrichment analysis
NGS	Next generation sequencing
NMR	Nuclear Magnetic Resonance
ORA	Over-representation analysis
PDB	Protein Data Bank
PIN	Protein interaction network
PINA	Protein Interaction Network Analysis
PPI	Protein-protein interaction
PPM	Position-specific probability matrix
pre-mRNA	precursor mRNA
PSI-MI	Proteomics Standards Initiative Molecular Interaction
PUM	Pumilio homology domain
PWM	Position weight matrix
RBD	RNA-binding domain
RBP	RNA binding protein
RRM	RNA Recognition motif
RWR	Random walk with restart
SELEX	Systematic Evolution of Ligands by EXponential enrichment
SP	Shortest path
SRE	Splicing regulatory element
UniProt	UniProt Knowledgebase
Y2H	Yeast two-hybrid

Introduction

Myocardial Infarction

Cardiovascular diseases (CVDs) are the most common cause of death and hospitalisation worldwide. It is estimated that 15.6M people died of CVD in 2010, representing 29.6% of global deaths. Only in Europe, despite steady decreases, CVDs are responsible for 45% of all death equating to more than 4 million deaths per year (Townsend et al. 2015). CVDs dramatically reduce life expectancy, quality of life, working years, and mobility. Among CVD, coronary heart diseases (CHDs) are the predominant contributors to CVDs. Myocardial infarction (MI) is the major clinical manifestations of CHD (Wong 2014).

Myocardial infarction: definition and causes

MI is the end result of a process, typically lasting for decades, called atherosclerosis. Atherosclerosis is characterized by the accumulation of cholesterol deposit in the coronary arteries. The disruption of an atherosclerotic plaque results in the release of a blood clot which leads to the blockade of the blood flow through a coronary artery (Woollard & Geissmann 2010). The interruption of the blood flow connected with the reduced functionalities of the heart is termed MI or commonly heart attack. Reperfusion of the blocked artery and restoration of the blood flow reduces the immediate mortality, but it still causes massive cell death due to oxidative damage. All these changes progressively lead to a decline in cardiac function and eventually to the heart failure.

Structure of the normal myocardium

Three are the main cardiac cell types in a healthy myocardium: Cardiomyocytes (CMs) Endothelial cells (ECs) and Fibroblasts (FBs). CMs are the core machinery and are responsible for the cardiac contraction-relaxation cycle. To perform their function, the CMs are aligned in parallel and in contact with each other. The fuel to perform these functions mainly comes from mitochondria which occupy one-third of the whole volume of the CMs (Schaper et al. 1985). FBs are the most abundant cells in an healthy adult mouse ventricles (Pinto et al. 2016). The relative abundance of each resident cardiac cell is still uncertain and depends mainly on the species, age, and gender (Table 1). FBs are responsible for producing the extracellular matrix, and transmitting the mechanical signal to the CMs (Kakkar & Lee 2010). To work properly, CMs require oxygen and nutrients that are provided through the interacting ECs. Heart functions are regulated by the physical interaction and by the complex interplay between CMs, ECs, and FBs.

Table 1: Main cardiac cell types and their relative abundance. Adapted from (Talman & Ruskoaho 2016)

Cell type	**Cardiomyocytes**	**Endothelial cells**	**Fibroblasts**	**Macrophages and other leukocytes**	**Pericytes and other mesenchymal cells**
Percentage	30%	45%	11%	6%	8%

The three phases of MI

The lack of oxygen-rich blood and nutrients due to ischemia leads to massive deaths of the cardiac resident cells. The reperfusion of the blocked artery, meaning the restoration of the blood flow, facilitates the CMs salvage and reduces immediate mortality, but it still causes massive cell death due to oxidative damage. To compensate this massive cell death, the myocardium starts a series of events that can be divided into three overlapping phases (Table 2): (i) an inflammatory phase, (ii) a proliferative phase, and (iii) a maturation phase. During these phases the reparative process takes place and the damaged ischemic tissue is progressively replaced by a fibrotic scar produced by FBs in order to prevent the cardiac rupture.

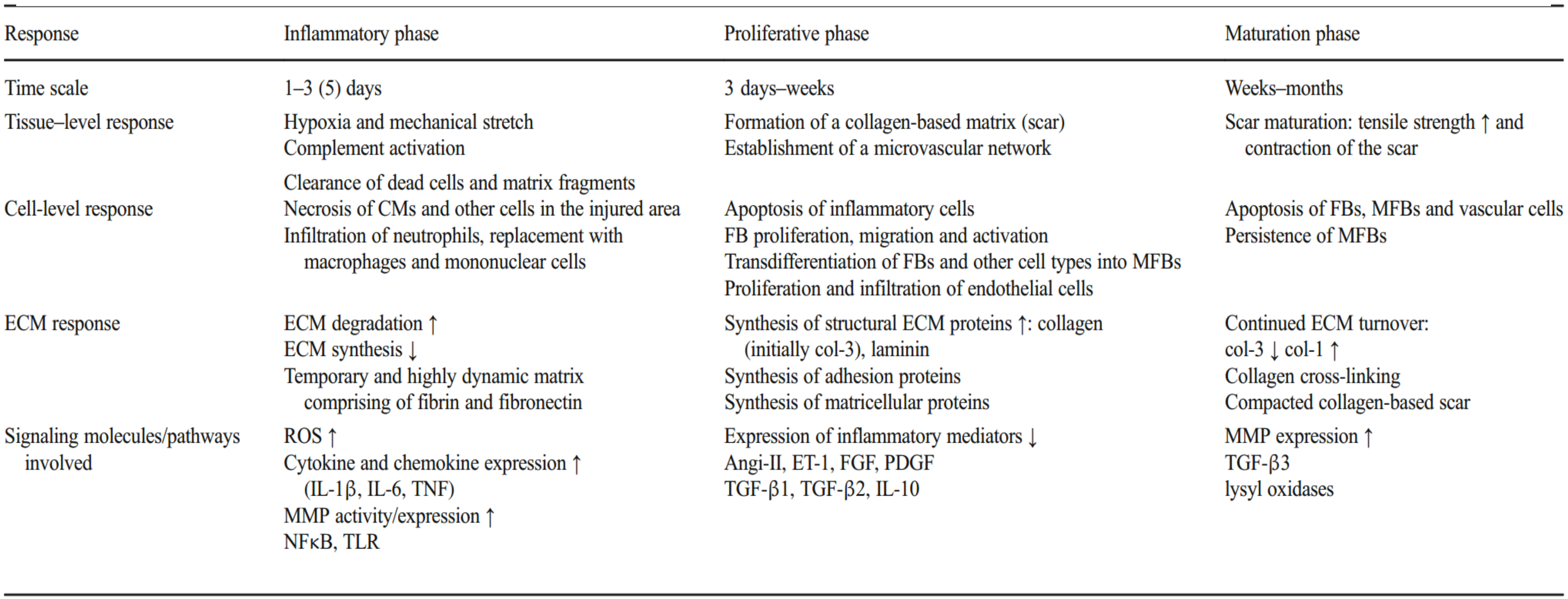
The inflammatory phase

The inflammatory phase lasts for 1-3 days. CMs are less resistant to the lack of oxygen with respect to FBs and ECs. For this reason, it has been proposed that CMs are responsible for triggering the inflammatory reaction in order to preserve the remaining cells and to replace the necrotic CMs. Since ECs, FBs and cardiac mast cells are more resistant to oxidative stress, it has been suggested that these cells are the responsible for receiving the signals from necrotic CMs and activating the innate immune system response (Shinde & Frangogiannis 2014; Frangogiannis 2014). The extracellular matrix (ECM) is degraded by FBs and allows the leukocyte and macrophages (MFs) infiltration into the ischemic tissue. The main role of macrophages and leukocytes is to clear dead cells and ECM debris and allowing the infiltration of new MFs and mononuclear cells and later of FBs for the deposition of the new ECM.

The proliferative phase

The proliferative phase starts approximately 3 days post-MI and can last for weeks. The trigger signals that activate the transition from the inflammatory phase to the proliferative phase are poorly understood. During this phase, the cells recruited and infiltrated in the infarcted tissue during the inflammatory phase are subjected to apoptosis. FBs are then activated to secrete ECM proteins in order to: (i) increase the myocardium tensile strength, (ii) produce the fibrotic scar to prevent the cardiac rupture, (iii) favours the cell migration along the new ECM matrix (van den Borne et al. 2010; Shinde & Frangogiannis 2014).

Table 2: The different phases taking place after myocardial infarction. The inflammatory phase lasts for 1-3 days. During this phase the innate immune response is activated and the extracellular matrix is degraded allowing the leukocyte and macrophage (MF) infiltration into the ischemic tissue. The proliferative phase starts approximately 3 days post-MI and can last for weeks. During this phase the cells recruited and infiltrated in the infarcted tissue during the inflammatory phase are subjected to apoptosis. FBs are then activated to secrete ECM proteins. The ECs proliferate and infiltrate the infarcted area to promote the development of the microvascular network. Then the maturation phase takes over. It lasts for weeks, and the main role is to stabilize the infarcted area. Adapted from (Talman & Ruskoaho 2016)



Response	Inflammatory phase	Proliferative phase	Maturation phase
Time scale	1–3 (5) days	3 days–weeks	Weeks–months
Tissue–level response	Hypoxia and mechanical stretch Complement activation Clearance of dead cells and matrix fragments	Formation of a collagen-based matrix (scar) Establishment of a microvascular network	Scar maturation: tensile strength ↑ and contraction of the scar
Cell-level response	Necrosis of CMs and other cells in the injured area Infiltration of neutrophils, replacement with macrophages and mononuclear cells	Apoptosis of inflammatory cells FB proliferation, migration and activation Transdifferentiation of FBs and other cell types into MFBs Proliferation and infiltration of endothelial cells	Apoptosis of FBs, MFBs and vascular cells Persistence of MFBs
ECM response	ECM degradation ↑ ECM synthesis ↓ Temporary and highly dynamic matrix comprising of fibrin and fibronectin	Synthesis of structural ECM proteins ↑: collagen (initially col-3), laminin Synthesis of adhesion proteins Synthesis of matricellular proteins	Continued ECM turnover: col-3 ↓ col-1 ↑ Collagen cross-linking Compacted collagen-based scar
Signaling molecules/pathways involved	ROS ↑ Cytokine and chemokine expression ↑ (IL-1β, IL-6, TNF) MMP activity/expression ↑ NFκB, TLR	Expression of inflammatory mediators ↓ Angi-II, ET-1, FGF, PDGF TGF-β1, TGF-β2, IL-10	MMP expression ↑ TGF-β3 lysyl oxidases

Additionally, the ECs start to proliferate and infiltrate the infarcted area to promote the development of the microvascular network and support FBs with oxygen and nutrients for proliferation and repair process (Deb & Ubil 2014).

The maturation phase

Afterwards, the maturation phase takes over. The main role of the maturation phase, which lasts for weeks, is to stabilize the infarcted area. During this phase, most of the FBs and ECs activate the apoptotic pathway, the remaining FBs produce type III collagen to progressively substitute the type I collagen in order to increase the stiffness of the scar (Dobaczewski et al. 2010). For a complete review (Talman & Ruskoaho 2016).

Molecular regulation of cellular processes – RNA molecules

In higher eukaryotes, RNA molecules are originated from DNA which is transcribed into a precursor RNA (pre-mRNA). To facilitate maturation of the mRNA, a big molecular complex called spliceosome is assembled. The main task of the spliceosome is to remove the introns in the pre-mRNA and to link the exons together. Splicing is not a rigid process and allows an alternative combination of exons, leading to the generation of different transcripts and proteins from a single gene. RNA binding proteins (RBPs) finely regulate these processes. RBPs define which exons are maintained or skipped, stabilise the mRNA, and export it from the nucleus to the ribosomes.

RNA binding proteins

RNA binding proteins (RBPs) are the key regulators of several cellular processes at mRNA level. Through their binding with RNA, RBPs regulate and control multiple processes including alternative splicing, transport, localization and stability (Glisovic et al. 2008). RBPs directly interact with RNA through particular domains, by recognizing short sequences (6-7 nucleotides long on average) called RNA binding motifs (Lunde et al. 2007). Typically, RBPs recognize thousands of binding sites and potentially regulate thousands of transcripts (Smith & Valcarcel 2000; Barash et al. 2010) with different specificities and affinities. Recent advances in the detection methods and in the solved crystallographic structure of ribonucleoprotein complexes started to shed light on the complex mechanism in which RPBs and RNA are involved in. However, we are currently far from a comprehensive characterization of the complex mechanisms regulating RNA and RBPs interaction and an *in-silico* prediction of RBPs targets and of post-transcriptional gene regulation remains a challenging task. Finally, alterations in RBPs are implicated in many human diseases such as neuropathies, muscular atrophies and cancer (Lukong et al. 2008).

RNA recognition by RNA-binding domains

The RNA-binding domains (RBDs) present in the RBPs are the responsible for the recognition and binding to RNA. The most representative RBDs in eukaryotes are the RNA Recognition motif (RRM), the K Homology (KH) domain, the Pumilio homology domain (PUM) and the Zinc finger domain. RBDs can be found in a single or multiple copies along the RBPs. The majority of RBPs contain multiple copies of the same domain in order to selectively recognize larger and more complex transcripts. The usage of multiple RBDs allows increasing the specificity and the target recognition. RBPs containing only a single copy of RBD usually do not have the strength to interact with RNA in a sequence-specific manner. To increase their specificity, those RBPs containing only one RBD: (i) expand the size of the binding site or (ii) cooperate with other RBPs (Lunde et al. 2007; Jankowsky & Harris 2015a; Ascano et al. 2012).

Single RBDs typically recognize short motifs ranging from 4 to 8 nucleotides. Moreover, the RBDs, usually, recognize degenerate motifs, thus the number of potential interactions and the number of potential binding sites is extremely large (Lunde et al. 2007; Jankowsky & Harris 2015b). The crystallography and NMR structure of RBP-RNA complex highlighted the role of hydrogen bonds and van der Walls forces in the RBP-RNA interaction. Hydrogen bonds, both from the protein backbone and the side chain, contribute to the RNA recognition, while the van der Walls forces enhanced the affinity between RBP and the nucleotides (Jones et al. 2001; Chen & Varani 2013).

Experimental methods for the identification of RBPs

RBP-RNA interactions can be detected through many methods, but essentially can be divided into: (i) low-throughput and (ii) high-throughput techniques. These techniques can be further divided in *in-vivo* and *in-vitro* depending on where the experiments are performed. Each of the methods has its own advantages and drawbacks. For example, in Systematic Evolution of Ligands by EXponential enrichment (SELEX), the binding affinity between RBP and RNA is determined by the alignment of sequences with a strong preference to bind to RBPs. The result is a consensus sequence that reflects the binding affinity between the RBP and the motif. SELEX has clear limitations because each position is not evaluated quantitatively and does not reveal the protein affinity for sub-optimal motifs. In RNAcompete as well as in SELEX there is the chance to identify a false positive target. CLIP-seq experiments cross link the RBP to the RNA using UV radiation. UV cross-linking establishes covalent bonds between proteins and RNA. The main problem is that UV cross-linking binds preferentially certain nucleotides and certain amino acids (Ule et al. 2005). Experimental limitation coupled with the complex interplay between RBP and RNA make the task of determining RBP binding sites a challenging problem.

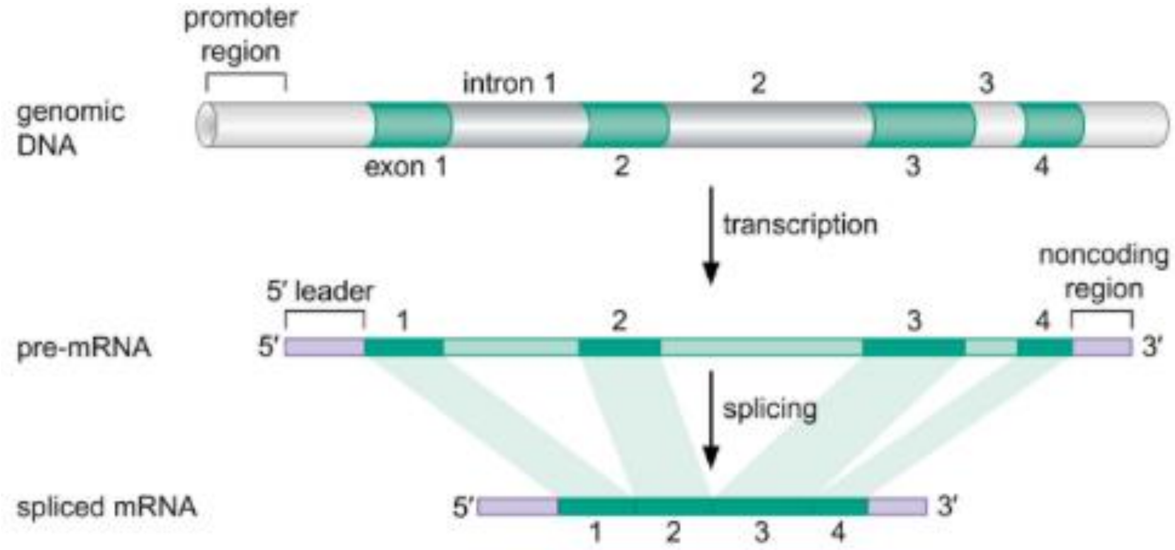


Figure 1: In the canonical splicing the introns are removed and the exons are ligated together.

Basal splicing machinery

Splicing is an eukaryotic regulatory mechanism by which the exons are ligated together and the introns are excised from the precursor-mRNA (pre-mRNA) (Fig 1). Splicing of the precursor-mRNA (pre-mRNAs) is catalysed by a ribonucleoprotein complex called spliceosome that is comprised of more than 100 proteins and five small nuclear RNAs (U1, U2, U4, U5, and U6) (Will & Luhrmann 2011). Four well conserved exonic and intronic signals located within the intron are necessary for the spliceosome assembly (Fig 2): (i) the donor site (5’ at the beginning of the intron), the acceptor site (3’ end of the intron), the branch point and the polypyrimidine tract. The first step in the spliceosome assembly is the recognition of the consensus sequence GTRAGT (where R can be A or G) at 5’ splice site by the U1 snRNP. The branch site and the 3’ splice site are recognized by the U2 snRNP. The U4, U5 and U6 snRNPs are then recruited to act as a link between the U1 and U2. Then, a series of conformational rearrangements allows the bending and the breaking of the intron and exons are joined together (Wahl et al. 2009).

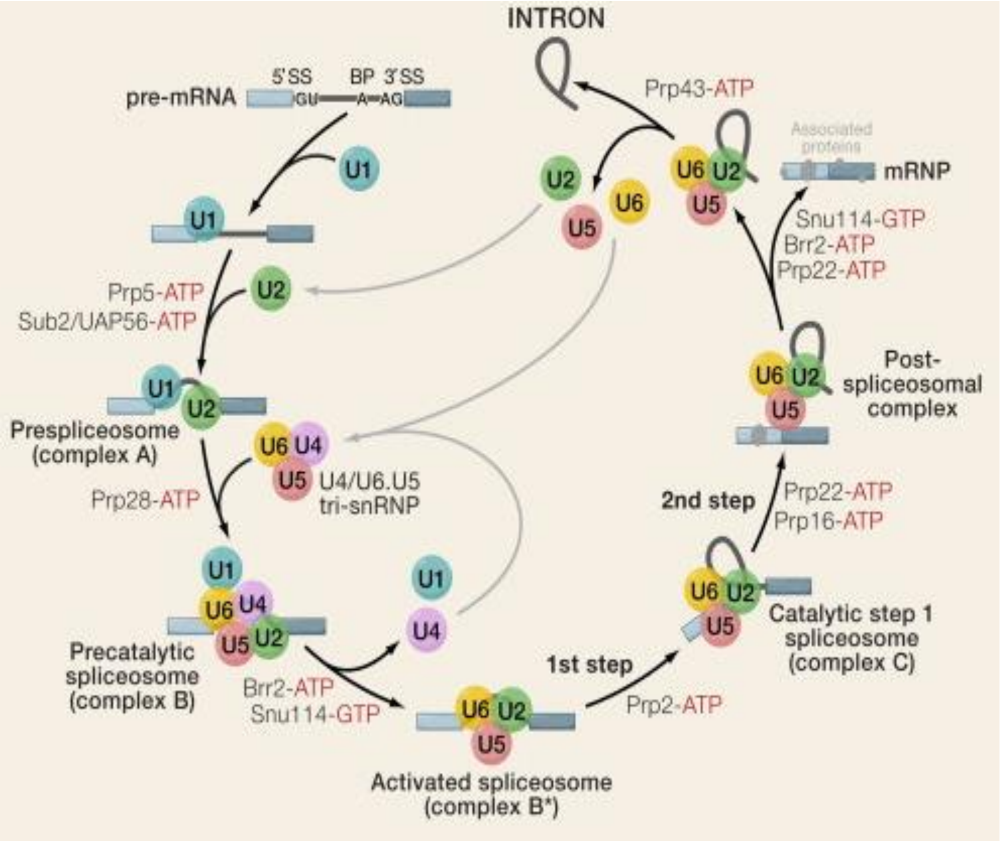


Figure 2: Spliceosome assembly. The stepwise assembly of the spliceosome is shown. The non-snRNP proteins are not shown. Adapted from (Wahl et al. 2009)**.**

Alternative splicing and alternative splicing events

Alternative splicing (AS) refers to the phenomena whereby the pre-mRNA may be spliced in a different combination in order to produce an alternative transcript and, potentially, a protein with a different structure, function or localization.

AS produces different mRNA variants by skipping, shortening and extending exons, retaining introns or via different usages of alternative splice sites. The isoforms represent the possible different forms of mature mRNA that can arise from a single gene.

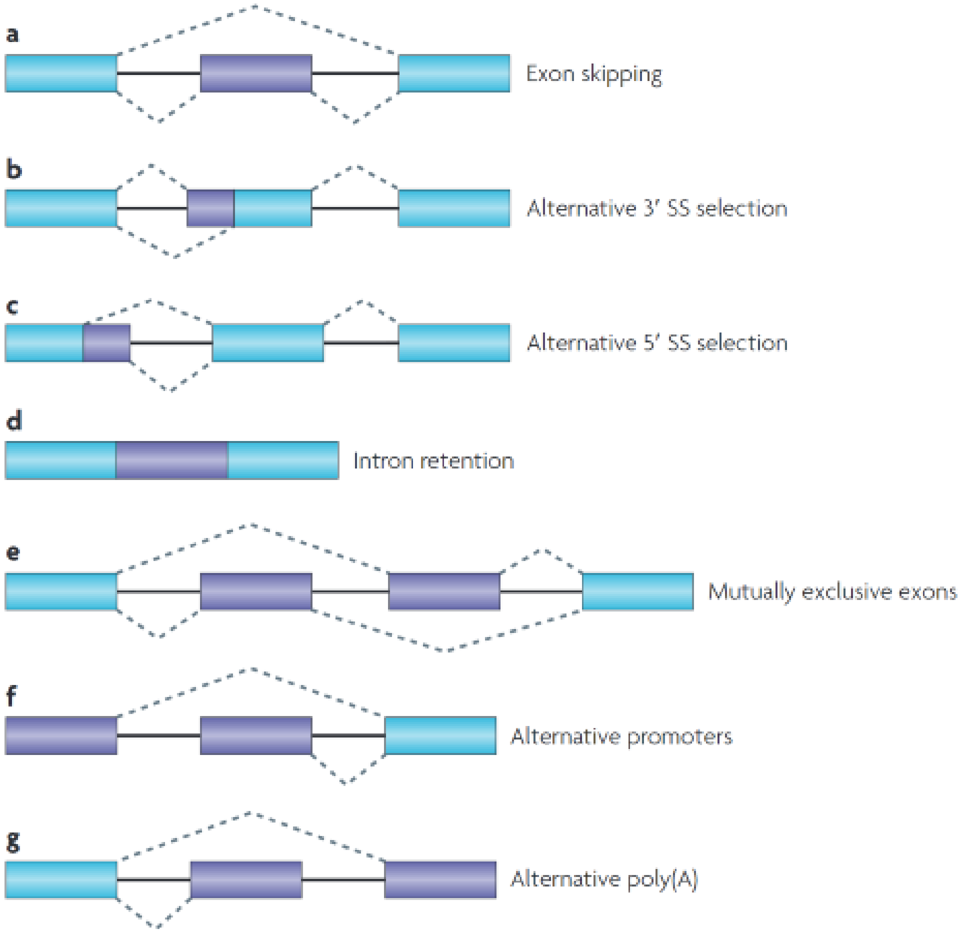


Figure 3: The different AS events. Exon skipping (a), Alternative 3' start site selection (b), Alternative 5' start site selection (c), Intron retention (d), Mutually exclusive splicing (e), Alternative poly-(A) (g). Reproduced from (Keren et al. 2010)**.**

AS events can be classified in different ways:

- Exon skipping (Fig 3a). The exon is removed together with the flanking introns from the pre-mRNA. Exon skipping is the most common AS event in higher eukaryotes.
- Alternative 3' start site selection (Fig 3b). An alternative splice site acceptor is recognized during the spliceosome assembly
- Alternative 5' start site selection (Fig 3c). An alternative splice site donor is recognized during the spliceosome assembly
- Intron retention (Fig 3d). The intron is not spliced out from the pre-mRNA
- Mutually exclusive splicing (Fig 3e). One of the two exons can be included in the pre-mRNA but not both.
- Alternative promoters (Fig 3f). A gene with an alternative promoter allows starting the transcription at a different point

- Alternative poly-(A) (Fig 3g). If present, a different poly-(A) can be selected as an alternative end point for the transcription.

It has been estimated that 95% of human multi-exon genes undergo at least one AS event. AS is the main mechanism responsible for the great disparity in the size of the human genome and proteome, between the 22000 genes and the estimated 100.000 proteins (Pan et al. 2008; McManus & Graveley 2011).

AS is not only the main mechanism responsible for protein diversity but is also responsible for the regulation, localisation, and function of proteins and for controlling gene expression levels via nonsense mediated decay (Wang et al. 2008; Licatalosi & Darnell 2010).

Alternative splicing regulation

The spliceosome complex alone is not sufficient to regulate AS. The recognition of the exon is mediated by additional elements: (i) the *cis-regulatory* elements and (ii) the *trans-regulatory* factors. C*is-regulatory* elements are sequences of approximately 4-8 nucleotides located in the exons or in the neighbour introns that favour the inclusion or the exclusion of an exon in the mature mRNA. These splicing regulatory elements (SREs) can be classified as: (i) exonic splicing enhancers (ESEs) or silencers (ESSs) if they are located in the exonic region and promote the exon inclusion or exclusion respectively, (ii) and as intronic splicing enhancers (ISEs) or silencers (ISSs) if they are located in an intronic region and promote or inhibit usage of adjacent splice sites or exons (Wang & Burge 2008). SREs act as a signal to recruit the *trans-regulatory* factors. T*rans-regulatory* factors are RBPs that recognize the *cis-regulatory* elements and orchestrate the AS events of the transcript. The activity of the SREs depends on the location within the pre-mRNA, the same SRE can function as an enhancer or silencer depending on whether SREs are located in an intronic or in an exonic region (Wang et al. 2008) (Fig. 4). Additionally, the majority of RBPs are able to recognize and bind to degenerate SREs, suggesting flexibility in RBPs recognition. The complex interplay between the *cis-regulatory* sequences and their recognition by the RBPs allows promoting or inhibiting AS.

Given the complexity of the AS regulation and of the spliceosome assembly, it is not surprising that AS is prone to error due to the huge number of factors involved, a disruption of which can lead to a wide range of diseases (Cieply & Carstens 2015; Lara-Pezzi et al. 2013).

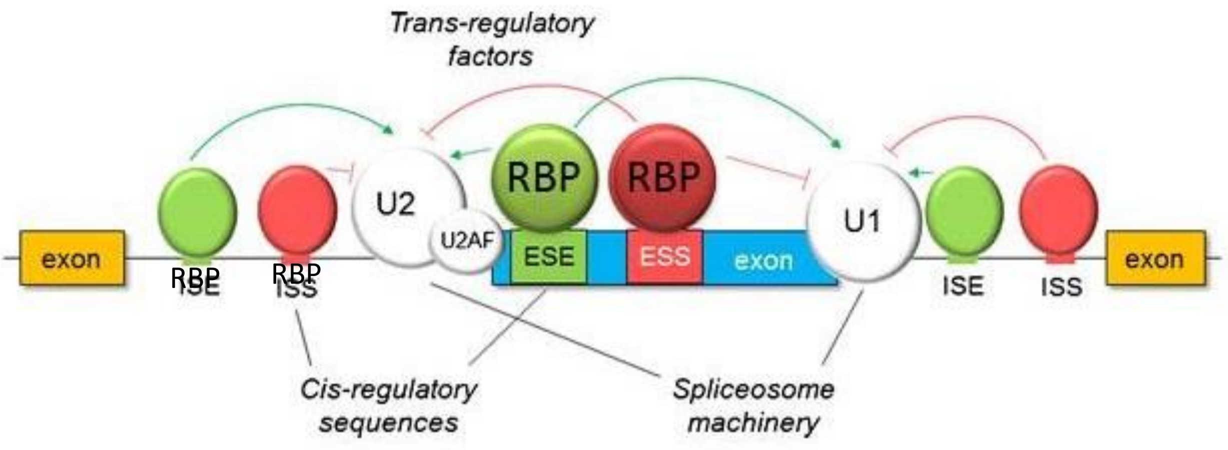


Figure 4: Cis-regulatory elements and trans-regulatory factors regulate alternative splicing. Adapted from (Lara-Pezzi et al. 2013)

High throughput sequencing

After (alternative) splicing is completed, RBPs export the mRNA molecules from the nucleus to the cytoplasm, precisely into the ribosomes, where they are potentially translated into proteins. The mRNA molecules, hence, play a crucial role since they act as an intermediate between the genome and the proteome. Quantifying the abundance of each mRNA is fundamental to fully understand the molecular mechanisms underlying a phenotype, for this reason, the high-throughput sequencing (HTS) methods are, nowadays, widely used. Currently, HTS platforms produce terabytes of biological data in a single run. Manually interpreting the results is unfeasible, thus the natural next step would be to develop a computational approach to provide a better biological interpretation of the generated dataset. This raises three important issues related to:

- the interpretation and management of the information, due to the huge amount of data produced
- the integration of the large list of data with different databases in order to have a broad view and a deeper understanding of the datasets under investigation.
- the development of reliable automatic approaches for the analysis of the data

The typical approach employed to interpret a large genes/proteins list typically involves over-representation analysis (ORA) techniques for identifying candidate genes or gene products (proteins) representative of the underlying biological process (Huang et al. 2009). The major limitation of these tools is that each gene/protein is treated as an isolated entity, therefore the biological information contained in the molecular interaction network underlying the genes/proteins set of interest is usually not considered. Since proteins do not act in isolation, most of their functions and processes are influenced by the neighbouring polypeptide (von Mering et al. 2002). Hence shift the ORA at the interaction network level can help to shed light on new biological processes (Gonzalez & Kann 2012).

Protein interactions

Proteins are fundamental for many biological processes inside and outside the cell, including transport of ions molecules and macromolecules across the membrane, accelerating chemical reactions, transmitting the information from DNA to the RNA, mediating the transduction of signals and much more. Most of these processes are rarely carried out by a single protein but require the interaction of many of them. The interaction occurs when two protein interfaces come in contact with each other and generate a weak intermolecular force, like hydrogen bond, salt bridge, ionic interactions and/or a generic van der Waal's force (Jones & Thornton 1996a).

Types of interaction

It has been observed that proteins involved in the same biological process often tend to interact with each other (von Mering et al. 2002). The analysis of protein-protein interactions (PPIs) is fundamental to understand the molecular mechanisms and functionality of cellular processes. Interactions between two proteins can be classified as physical or logical. Physical interaction occurs when two or more proteins interact physically forming, for example, a stable complex (Jones & Thornton 1996b). Logical interaction or functional association occurs when one or more proteins affect the behaviour of other proteins. Examples of logical interaction are the metabolic pathways or proteins that are part of a complex without being in direct contact (Huynen et al. 2000).

Detection methods

PPIs can be detected through different experimental methods. These methods are classified into two main categories:

- Small scale methods, designed to identify a small number of interactions. These methods usually detect a single binary interaction between two proteins
- High-throughput methods designed to detect large-scale protein interactions; in this case, the methods isolate a set of proteins that act as a complex.

In the 1980's and 1990's, most PPIs were detected via small-scale methods. These methods investigate only a little set of proteins of interest. The classical approach made use of X-ray crystallography to find proteins that were co-crystallized in pairs. Few PPIs were found using this approach because it is very difficult to crystallize a protein, let alone a complex of proteins. In the last decade, the advent of high-throughput methods allowed collecting a huge amount of interaction data. These methods rely on bait-prey interactions. A bait protein is a protein used to "catch" and identify one or more interaction partners (prey proteins). High-throughput methods generate a huge quantity of data in a single run, at the cost of limited accuracy. For example, the high-throughput yeast two-hybrid (Y2H) assays are ∼50% reliable (Sprinzak et al. 2003).

Availability of protein-protein interaction data

PPIs data are collected in several public available databases in the form of a binary interaction. The number of protein interaction data available on each database is variable also within the same species. The variation mainly depends on the level of curation and on the methods adopted by each database to assess the interaction reliability. Some of them, like IntAct (Orchard et al. 2014) (www.intact.com), require two different experiments to assess a single interaction, others like STRING (Szklarczyk et al. 2015) are less restrictive and take advantage of automatic text mining techniques.

Additionally, the format in which the molecular interaction data is stored depends on the database. The lack of an unified standard had delayed the study of PPI and had not permitted the integration of available protein interaction data. Recently, the Proteomics Standards Initiative Molecular Interaction (PSI-MI) standard (Kerrien et al. 2007) has emerged and is being adopted by the majority of the PPI databases. The PSI-MI standard defined a common schema permitting to standardize the molecular interaction data in a common format. Furthermore, adopting the PSI-MI standard allows sharing and synchronizing data between different databases.

Protein-Protein interaction networks

Before the advent of high-throughput techniques, most research efforts were focused on single molecules (e.g. genes). The rapid emergence of high-throughput techniques led to a massive production of PPI data. These techniques allowed to study several biological processes simultaneously and in an integrated manner, leading to what is called 'system biology' (Ideker et al. 2001; Kitano 2002b; Kitano 2002a). Complex networks provide a promising framework for system biology investigations. PPIs can be naturally represented by means of networks or graphs. A protein-

protein interaction network (PIN) can be defined as a complex system of proteins linked by their interaction. In a PIN, the nodes represent the proteins and the edges represent the interactions. PIN are becoming increasingly large and complex, therefore, the interpretation of the functional properties at a large scale is extremely difficult. Hence, there is an increased interest to cluster together the proteins with similar function, since it has been observed that proteins involved in the same pathway interact together (von Mering et al. 2002; Nabieva et al. 2005). The next natural step is to break down the PIN into subnetworks or modules.

A module is a set of interacting proteins and/or metabolites that, in a co-ordinated way, perform a common biological function. In other words, modules are essentially the functional building block of the cells (Hartwell et al. 1999; Alon 2003; Barabasi & Oltvai 2004; Spirin & Mirny 2003; Stuart et al. 2003). Modules can be further distinguished in complexes and functional modules (Spirin & Mirny 2003). A complex is defined as a densely-interconnected subgraph of proteins that interact at the same place and time forming a single multimolecular machine (e.g. spliceosome). A functional module represents a set of proteins that participate in a biological process or in a function, and that bind each other at a different time and place (e.g. MAP signal cascade, phases of cell cycle) (Spirin & Mirny 2003).

Network based enrichment analysis

Standard ORA methods treat each protein as an isolated entity, neglecting the physical interactions between the gene/protein sets of interest. For this reason, several computational models have recently been developed to shift the functional association analysis at the interaction network level. This new class of methods are typically referred to as “network-based enrichment analysis” (NBEA) (Laukens et al. 2015). The NBEA methods allow detecting those biological processes that are not directly inferable from the annotations of the input protein set, and thus not detectable through an ORA. The common objective of all the NBEA methods is to identify modules, the differences lie in the algorithmic techniques used to extract the information from the underlying interaction network.

Several computational approaches have been designed to extract meaningful information from a PIN. These methods can be classified into two main categories:

- Topology based. Algorithms in this category extract the interaction network from the starting set of gene/proteins and then assess how much similar/distant they are with respect to a reference (among them EnrichNet (Glaab et al. 2012), PathNet (Dutta et al. 2012), SANTA (Cornish & Markowetz 2014), JEPETTO (Winterhalter et al. 2014)).

- Module based. Methods in this category extract modules from the PIN and then test if they are involved in a common biological role (among them: PINA (Cowley et al. 2012), FunMod (Natale et al. 2014), NET-GE (Di Lena et al. 2015)).

The most representative and cited methods among the two categories are: EnrichNet and PINA. EnrichNet is a web application that provides a network-based integrative analysis. In EnrichNet, the set of proteins are mapped on a PIN extracted from STRING (Szklarczyk et al. 2015) database. The mapped nodes are used as seeds for a random walk with restart (RWR) procedure. The score obtained from RWR is converted into a distance vector from all the reference pathways. The distance measures the associations between the input protein set and the reference pathways or processes. The output of EnrichNet is a ranking table of pathways or biological processes associated with the distance.

Protein Interaction Network Analysis (PINA) is a web application collecting pre-identified modules detected by the molecular complex detection (MCODE) (Bader & Hogue 2003) algorithm and by the Markov clustering (MCL) (van Dongen & Abreu-Goodger 2012) algorithm. The MCODE algorithm is used to detect small and dense interconnected regions of proteins that should represent complexes of proteins, the MCL algorithm is used to identify large and loosely interconnected regions that should represent a pathway. Each module identified by PINA is annotated through a standard ORA with terms derived from Gene Ontology (Huntley et al. 2015), KEGG (Kanehisa et al. 2012), and PFAM (Finn et al. 2014). PINA maps the input proteins to one or more pre-computed modules and selects those that are overrepresented. The annotation of the overrepresented module is then transferred to the input proteins.

All these methods, present their drawbacks and advantages. Apart from EnrichNet, these methods do not take into consideration the potential tissue-specific expression of the proteins. However, EnrichNet provides only the scores to measure the similarities between the input proteins list and the reference pathways but does not provide information about the statistical significance of the scores. Additionally, apart from NET-GE, the modules extracted with NBEA approaches are not function-specific, meaning that they rely only on statistics or topology of the network to find the interaction partners but do not check whether two interacting proteins participate in similar processes. To our knowledge there is no tool that combines the advantages of topology-based methods and of the modules-based methods.

Objectives

The main objectives of this thesis included the development of the necessary tools and algorithms for the study of post-transcriptional regulation in the context of myocardial infarction in four different cell types (cardiomyocytes, fibroblasts, endothelial cells, and macrophages) at three time points and the development of a new functional enrichment analysis technique. To accomplish these aims, we define the following specific objectives:

1) To develop a central and coherent repository that integrates all the RBP-RNA interactions that are currently sparse through different repositories.
2) To identify the alternatively spliced transcripts after myocardial infarction in the four cell types at different time points
3) To identify the motifs and RBPs that regulate the alternative splicing events in the pig infarcted myocardium
4) To develop a new data-driven functional enrichment analysis method to extract further and more specific biological processes and functions from a genes/proteins list

Materials and Methods

ATtRACT - A daTabase of RNA binding proteins and AssoCiated moTifs

Available databases: direct and indirect sources of information

At the time this work was performed, four were the databases expressly designed for compiling information on RBPs and their binding sites:

- RBPDB (Cook et al. 2011) is a collection of experimentally validated RBPs and associated binding sites linked to a reference in literature
- CISBP-RNA (Ray et al. 2013) contains RBPs and binding sites extracted from in vitro RNAcompete experiments.
- SpliceAid-F (Giulietti et al. 2013) is a database of human splicing factor extracted from literature
- ASD (Stamm et al. 2006) database is no longer maintained. At the present time, the data are available only in a text format file and as a consequence is difficult to search, interrogate and analyse the records.

Among these already available databases, a significant portion of information is available in Protein Data Bank (PDB) (Rose et al. 2015) in the form of co-crystallized structures of protein RNA-complexes. At the time, only RBPDB and SpliceAid-F included a limited number of motifs extracted from RBP-RNA complexes data available in PDB and were attained via literature mining.

ATtRACT: integrating different sources of information together

The data available in RBPDB, CISBP-RNA and SpliceAid-F databases were extracted and integrated to populate ATtRACT. Each entry, in ATtRACT, corresponds to a RBP and to its associated binding sites. Each entry was, furthermore, annotated with the following fields: the official gene name and its synonyms, the gene identifier, the motif associated to the RBP, the experiment performed for the detection of the motif, the PubMed identifier and the domains associated to the RBP according to PFAM or InterPro (Mitchell et al. 2015) annotations. Additionally, the GO terms were integrated in ATtRACT. GO annotations allow identifying the molecular process, the biological function, and the cellular component in which the RBPs are involved and at the same time making unnecessary traversing several sources of information to get the desired information.

Another source of information was the ASD database which was available only as a text file. We integrated the last release of ASD database in order to make the entries again accessible to the public through ATtRACT. Moreover, we updated the data available in ASD data file and completed the information needed to fill in the fields present in ATtRACT but missing in the ASD database. An overview of the data available in attract is shown in Fig 5.

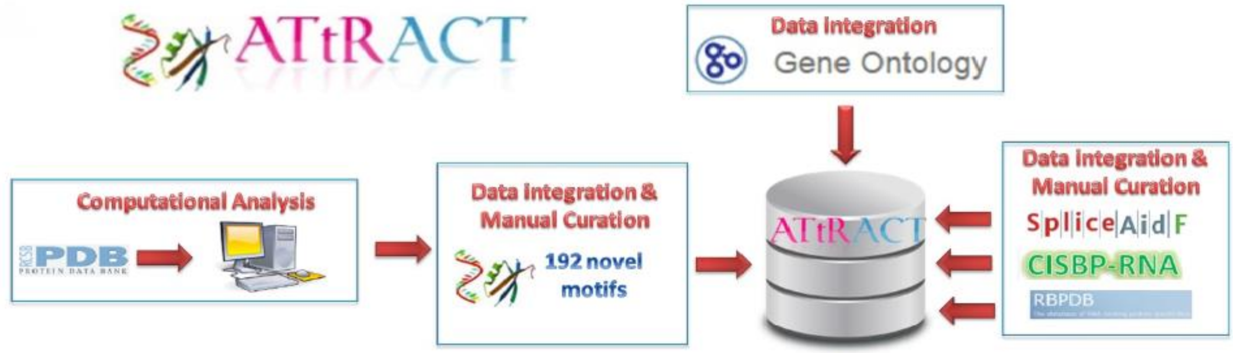


Figure 5: ATtRACT database. (A) Data flows in ATtRACT database. ATtRACT integrates data from SpliceAidF, CISBP-RNA and RBPDB. Moreover, 192 novel motifs, not present in any other database, were extracted from PDB and included in ATtRACT. ATtRACT also integrates Gene Ontology annotation.

ATtRACT: standardization

Inconsistency in nomenclature and annotation is one of the major issues of the last years in life science and big data. The RBPDB, CISBP-RNA, SpliceAid-F and ASD databases, all suffer from ambiguous gene notation. They are not coherent in terms of reference gene names and/or gene identifiers. For example, the Human CELF1 binding protein is named CUGBP1 in SpliceAid-F as well as in RBPDB and, CELF1 in CISBP-RNA. In ATtRACT, the RBPs gene names and identifiers were changed according to UniProt (Anon 2015) official names allowing to reduce the proliferation of the different gene names and duplicated entries. Furthermore, we updated, when available, the gene identifiers according to the last version of Ensembl (Flicek et al. 2014), Xenbase (Karpinka et al. 2014) or European Nucleotide Archive (Leinonen et al. 2011).

Integrating the RBP-RNA interactions in Protein Data Bank data

PDB is a repository of 3D protein structures and molecular complexes. A total of 236 proteins structures of RBP-RNA complexes, excluding the ribosomes and prokaryotes, were available in PDB (release January 2015). The structures available contained at least a RNA-protein complex and were not included in RBPDB and SpliceAid-F databases. The experimental methods used for the detection of protein-RNA contacts were: Nuclear Magnetic Resonance (NMR), electron microscopy and X-ray crystallography with a resolution better than 3.9 Å. To extract information from all these complexes we developed a pipeline to obtain RNA sequence motifs from the structural information stored in PDB experiments. RPBs-RNA complexes are bound together thanks to two types of interactions: the van der Walls forces and the hydrogen bonds (Jones et al. 2001). Since it is not possible to detect the hydrogen bond in the X-ray crystallographic structure, we employed the HBPLUS (McDonald &

Thornton 1994) program to assign the hydrogen bond to the X-ray crystallographic data. We used the NUCPLOT (Luscombe et al. 1997) program to identify amino acids-RNA contacts. We used a distance-based criterion to identify the intermolecular forces involved in the protein-RNA complex structures:

- An hydrogen bond is detected if the distance between the RNA and the protein is less than 3.0 Å.
- A van der Walls interaction takes places if the distance between the RNA and the protein is less than 3.9 Å.

We considered a motif as detected if four or more contiguous nucleotides satisfied the distance criterion defined previously and interacted with any atom of the RBP. We extracted 256 motifs from 110 different RBPs, out of which 192 were novel motifs not available in any other database.

To test the reliability of our method, we performed an ungapped alignment between the motifs extracted from PDB and the motifs belonging to the same RBP and verified by another type of experiment. We aligned 93 motifs out of 256, belonging to 43 RBPs. Forty-eight motifs (51.6%) were perfectly aligned, 27 (29%) differ only in one nucleotide, 4 (4.3%) differ in two nucleotides, the remaining 14 motifs (15.1% of the total) differ in more than two nucleotides. For the motifs that differ for more than two nucleotides we mined the literature and we found evidence that 13 motifs out of 14 were correct even if they were completely different from the motif confirmed with another type of experiment (see supplementary Annex I).

Quality score

RBP-RNA interactions are detected with different experiments and each one with a different level of confidence. For this reason, in ATtRACT a quality score is implemented. The purpose of quality score was to evaluate the binding affinity between RBPs and binding sites. SELEX and RNAcompete experiments identify winner sequences, i.e. sequences with a strong preference to bind to RBPs. The functional motif is assessed through the alignment of winner sequences. The result of the alignment is often represented through IUPAC ambiguous notation (Anon 1986) and shows the most frequent nucleotides found at each position. This representation has clear limitations because each position is not evaluated quantitatively (Staden et al. 1982) since each nucleotide is considered as equally likely. In AEDB and SpliceAid-F databases, the motifs assessed by SELEX or RNAcompete experiments are considered equally likely; therefore, it is not possible to evaluate the binding affinity between RBPs and motifs. To solve this problem, we manually extracted from the literature the winner sequences and aligned them in order to represent the binding preference through a position-specific probability matrix (PPM). If the winner sequences were not annotated, the PPM is generated considering the IUPAC letter encoding for more than one nucleotide as equally likely.

In mathematical terms, the score *S* of a matrix *M* for a motif *m* of length *l* is defined as:

$$S = \prod_{i=1}^{l} P(m_i|M)$$

where $P(m_i|M)$ is the probability of observing the nucleotide *m* in position *l* in the PPM matrix. The quality score represented the probability of observing a given motif within the experiment. Note that each motif in ATtRACT is associated to a PPM. For this reason and according to the definition of quality score, the motifs coming from single sequence experiments, such as UV cross-linking or EMSA, have a quality score equal to 1.0, since it is possible to assess only one motif in the experiment.

Sequence scan

Scanning the sequences searching for potential binding motifs is currently possible only for RBPDB and CISBP-RNA databases. The limitation of both databases lies in in the sequences length that is possible to scan and in the number of motifs available.

The Burrows-Wheeler transform (BWT) algorithm (Burrows & Wheeler 1994) is integrated into ATtRACT. The BWT algorithm allows to scan the sequences searching for those motifs that perfectly match a given sequence or set of sequences. The input file can be a FASTA or multi-FASTA file of 20000 nucleotides maximum. The BWT currently represents one the most efficient algorithms and is widely used by the majority of the first generation NGS algorithms to perfectly align the reads to the transcriptome or to the genome (Li & Durbin 2010; Li & Homer 2010); In a very short time, BWT scans the input sequences and detects the positions of any or a subset of motifs selected by the user, through a perfect match comparison. Additionally, since the number of motifs available in ATtRACT greatly exceeded the ones available in the other databases and because of their small length, it was very easy to find many of those motifs in any input sequence provided. In order to assess the possible biological relevance, we provided a log-odds scores. Three log-odds scores are assigned to each motif of the following species since they are the most represented species in ATtRACT database: *Caenorhabditis Elegans*, *Drosophila Melanogaster*, *Homo Sapiens*, *Mus Musculus*, *Saccharomyces Cerevisiae* and *Xenopus*. To compute the log-odd score, we introduced the concept of genomic functional context. The genomic functional context is a collection of three distinct datasets each one belonging to a species specific genomic region. Each dataset is composed respectively of the sequences of:

- all the exons plus 250 nucleotides upstream and downstream
- all the introns
- all the coding sequences.

The log-odd score is defined as the ratio between the probability of locating the motif in the input sequence and the probability of finding the same motif in any of the genomic functional context of the reference species. Therefore, a log-odd score greater than 0 means that the probability of finding the motif is greater in the input sequence with respect to the corresponding genomic functional context and vice versa if less than 0.

In mathematical terms, the score was calculated in the following way:

Let M = [m_1, m_2, m_3... mn] be the set of the motifs present in the database.

We defined the genomic functional context *GFC = [exon±250, intron, CDS]* as the collection of distinct datasets containing respectively the sequences extracted from all exons plus 250 nucleotides upstream and downstream, all introns and all coding sequences of the reference organism.

$S^{[GFC]}$ = [s_1, s_2 ... s_n] where s_1, s_2 ... s_n were the sequences in the reference organism that represent the genomic functional context. I.e. $S^{[\text{intron\_human}]}$ represented all introns in human

$C^{S[GFC]}$ = [c_{m1},c_{m2},c_{m3},...,c_{mn}] where c_{m1},c_{m2},c_{m3},...,c_{mn} were the occurrences of motifs m_1,m_2,m_3,...,m_n in $S^{[GFC]}$

Let s be an input sequence of length l_s and $m_x \in M$ a motif of length l_m of multiplicity t found in the input sequence. The log odd ratio was defined as:

$$OR_{S[GFC]} = \log_2 \frac{Obs}{Exp_{S\,[GFC]}}$$

where *Obs* was defined as:

$$Obs = \frac{t}{l_s - l_m + 1}$$

$$Exp_{s[GFC]} = \frac{c_{mx}}{\sum_{i=1}^{n}[\,len\left(s_i^{[GFC]}\right) - l_{\,m} + 1]}$$

Where $len\left(s_i^{[\,GFC]}\right)$ was the length of the i^{th} sequence in $S^{[GFC]}$

MEME and Tomtom

ATtRACT integrated a pipeline to discover motifs that occur frequently in a set of sequences and compare them with the ones present in the ATtRACT database. For this aim, MEME (Bailey et al. 2009) and Tomtom (Gupta et al. 2007) programs were integrated into ATtRACT. MEME adopts an extension of the expectation maximization algorithm to produce a statistical model that permits to find a relationship between possibly related unaligned sequences. To evaluate whether a *de novo* motif, enriched in a set of sequences, looks like to any other motif present in the ATtRACT database, the Tomtom algorithm is integrated into the pipeline. A score is assigned by Tomtom to each *de novo* motif, found by MEME, and based on the expected value (E-value). The E-value describes the number of hits one can expect by chance in a database of a particular size. The closer to zero the E-value, the more plausible the match is. The E-Value is strongly correlated with the database's size, for this reason ATtRACT allows to extract a subset of the database according to: (i) the length of the motif, (ii) the experiment assessed, (iii) the organism, and (iv) the domain, in order to fine tune the research and increase the E-Value.

Implementation details

The algorithms in ATtRACT were implemented in python 2.7 (https://www.python.org) and C/C++. To store the information, managing and organizing the database, the SQLite database (http://www.sqlite.org) was adopted. The web2py (http://www.web2py.com) framework and bootstrap (http://getbootstrap.com) framework were used for designing and developing the web interface and for interfacing with the database. Nginx (http://nginx.org) handles the user's request. The plots were implemented using the D3js (http://d3js.org) and highcharts (http://www.highcharts.com) libraries. The tables displaying the results were implemented using the javascript plug-in called DataTables (http://www.datatables.net). The logos were generated through WebLogo3 software (http://weblogo.threeplusone.com)

Analysis and regulation of alternative splicing in the pig infarcted myocardium

RNA extraction

Large white pigs (*Sus scrofa*) were used for the experiments, since they represent a good translational animal model. Myocardial ischemia was induced via balloon occlusion for 45 minutes. The balloon was inserted in the left ascending coronary artery and inflated. Forty-five minutes later the balloon was deflated and the normal circulation flow was restored again. The pigs were sacrificed at day 0 (no infarct), 3 days and 7 days post myocardial infarction by the administration of potassium chloride. From the infarcted tissue area, a piece of tissue was minced and digested with collagenase. The enzyme digestion supernatant was collected in several rounds of digestion, collagenase was inactivated with fetal calf serum and cardiomyocytes (CMs) were decanted by gravity. ECs, FBs and MFs were isolated form the cell supernatant by fluorescence-activated cell sorting (FACS) using specific antibodies for each cell type. For each cell type, total RNA was extracted, quantified and sequenced by the CNIC Genomics unit.

Proteins identification

Protein extracts were obtained by cell lysis. Protein extracts for each time-point were quantified and pooled according to their concentration. Samples were subjected to tryptic digestion and the resulting peptides were labelled for relative quantification (iTRAQ) and separated. The fractionated peptides were analysed by nano-liquid chromatography-tandem mass spectrometry (nanoLC-MS/MS) using a Q-Exactive hybrid quadrupole orbitrap mass spectrometer (Thermo Scientific).

Protein identification was performed by the CNIC Proteomics unit using the SEQUEST HT algorithm integrated in Proteome Discoverer 2.1 (Thermo Scientific). Peptides were identified from MS/MS data using the probability ratio method (Martinez-Bartolome et al. 2008). False discovery rate (FDR) of peptide identifications was calculated by the refined method (Bonzon-Kulichenko et al. 2015, Navarro et al. 2009).

RNA-Seq: how the reads were generated

RNA-Seq is a next generation sequencing (NGS) method that allows sequencing the transcriptome in a sample. RNA-Seq is, currently, widely used to analyse gene expression and alternative splicing. The protocol used for sequencing consisted of the following steps: (i) the poly-adenylated RNA was extracted from each cell types and from each time point; (ii) the RNA was converted into cDNA and fragmented; (iii) the fragments were amplified and the adapters were ligated to the fragments; finally, (iv) the cDNA was sequenced. Samples were sequenced in an Illumina HiSeq 2500, which produced 61 nucleotides long reads.

RNA-Seq analysis: removing the adaptors and low quality reads

The output of the sequencing machine is typically a list of raw sequences data in FastQ format. A quality control check was performed through the FastQC (Wolf, 2013) program to check the quality of the reads. Initially, we got an average of 85,052,976 reads per cell type and time point (Fig 6). The CMs at day 0 showed the fewest reads (51,983,142). Conversely, the best yield was obtained from CMs at day 3 (100,174,650).

We used the Cutadapt (Martin 2011) program to trim the raw sequences data and remove the adapter. Additionally, those reads shorter than 30 nucleotides were removed. After the trimming step, 0.002% of the reads, on average, per all the cell types and all the time points were removed, producing reads of length ranging from 30-61 nucleotides. The FastQC program was run again to check that the qualities of the reads were improved.

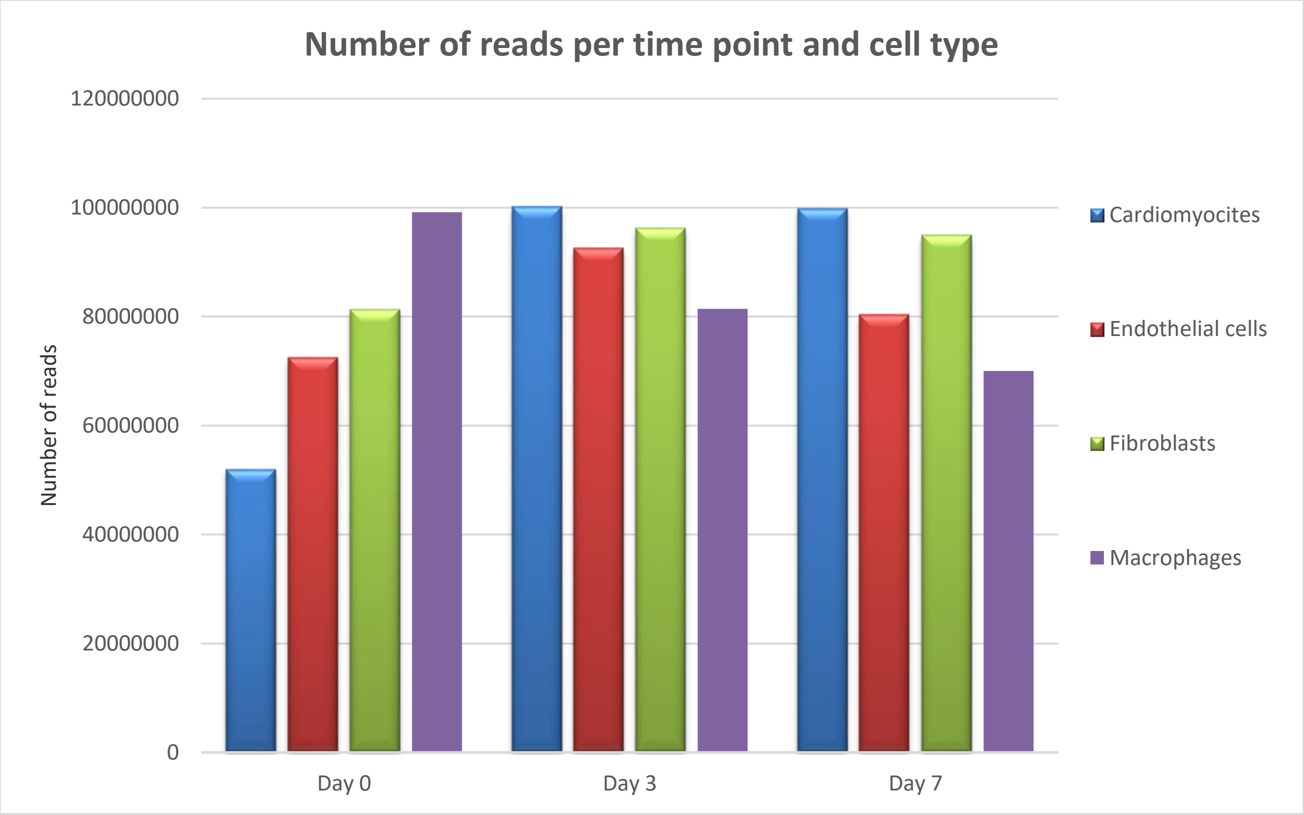


Figure 6: Number of reads per cell type and time point

The trimmed sequences were mapped to the pig reference transcriptome (reference build: *Sus scrofa* 10.2.73 http://www.ncbi.nlm.nih.gov/assembly/304498/). We chose the RSEM (Li & Dewey 2011) algorithm to align the reads, given that recent benchmarks suggest that RSEM outperforms other available tools in the detection of the isoform expression levels (Dapas et al. 2016).

Large scale data analysis

Typically, the output of genome-wide studies is a list of thousands of genes associated with values representing their expression. The biological interpretation of these lists is challenging and led to a great demand for methods to analyse a large amount of data generated in an automatic way.

A widely-used technique for the analysis of the datasets is by means of ORA. Several techniques are available, but all of them try to assess, with a statistical test, whether a gene/protein set is overrepresented in a reference list of known gene/protein sets. Gene Ontology database is widely used as a source of annotated gene/protein sets and is commonly used for the ORA.

Gene ontology database

Ontologies, termed as a set of concepts within a domain, have gained increasing attention in the recent years as tools for representing knowledge. According to Gruber's (Gruber 1993) definition, an ontology is "*an explicit specification of a conceptualization*". Ontologies in bioinformatics play a key role because provide a source of standardized nomenclature terms. The Gene Ontology project concentrates its efforts to provide species-independent descriptions of gene products through a controlled vocabulary and to define the relationships between different terms (GO terms). GO terms are organised in a direct acyclic graph such that the level of specificity of the terms increases while traversing the graph from the root to the leaves. GO vocabulary is also subdivided into three domains:

- Cellular Component (CC) – describes the part of a cell or its extracellular environment in which the gene product is localised
- Molecular Function (MFu) – describes the gene product's elemental activities at the molecular level
- Biological Process (BP) – describes the events in which the gene product is involved

Researchers are often interested to know whether the output of their experiments shares some characteristics with a known set of genes/proteins. ORA method helps investigators to detect genes or proteins annotations significantly enriched within a target set of genes of interest, with respect to a background set.

Fisher's exact test and Bonferroni correction

To proceed with the testing of data, some hypotheses should be made. In a testing problem, two different and complementary hypotheses are formulated, the null hypothesis (H_0) and the alternative hypothesis (H_1).

The purpose of the hypothesis test is to establish if there is some experimental evidence that supports the rejection of the null hypothesis, setting a significance level to a certain significance threshold (α). The p-value measures the strength of the evidence against the null hypothesis in favour of the alternative hypothesis. The comparison of the p-value with α tells if the null hypothesis should be rejected.

Fisher’s exact test is a statistical significance test for categorical data. This non-parametric test is performed to verify if categorical data is compatible with the null hypothesis (H_0). In Fisher’s test, categorical variables are grouped into a contingency table and represent the frequency sample of the variables.

Fisher demonstrated that the probability of observing the sample in a contingency table follows a hypergeometric distribution and in case of true null hypothesis is equal to:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!}$$

Where *a, b, c, d* are the values of the contingency table and *p* is the probability of observing the samples *a, b, c, d* if the null hypothesis is true.

This procedure is valid if and only if the number of the hypotheses tested is equal to one.

When an experiment is performed, usually many hypotheses are considered; hence, adjustment of p-value is needed to reduce the probability of obtaining false positive results when multiple tests are executed on a single data set.

With the Bonferroni correction, the adjusted p-value is equal to:

$$P^I = \frac{P}{k}$$

Where k is the number of tested hypotheses.

Bonferroni’s correction is very stringent; hence, the risk is to throw away real enrichments.

Identification of isoform switches

A gene can give rise to different transcripts, the overall expression of a gene may remain constant between two conditions while the relative abundance of individual transcripts may change resulting in an isoform switching.

To detect the potential isoform switches we developed a pipeline inheriting some of the steps implemented in cufflinks (Trapnell et al. 2010b). The employed pipeline is composed of three main steps:

1) For the same gene and between two conditions, calculate the relative abundance distribution of the transcripts.
2) Calculate the Jensen-Shannon distance of the two transcripts distributions generated previously.

3) Rank the genes according to the Jensen-Shannon distance calculated previously and select only those genes lying over the 95th percentile.

The pipeline selected only those genes showing extreme changes in the abundance percentage of the transcripts. To proper select the transcripts undergoing isoform switching, we further selected the transcripts according to the criteria that are described in the paragraph isoform switching detection and filtering.

The 'IsoPct' field in RSEM output represents the transcript isoform percentage, in other words, the ratio between transcript's abundance and the gene's abundance.

If only one isoform is present, the isoform percentage is set to 100% and we excluded all these transcripts.

An example of how the isoform percentage is calculated is shown in Fig 7.

The Jensen-Shannon distance

The square root of the Jensen-Shannon divergence is a commonly used measure (Trapnell et al. 2010a) to determine which genes were significantly changed (e.g. isoform switches) among two conditions. The JSD is defined as:

$$JSD(P||Q) = \tfrac{1}{2}\, D(P||M) + \tfrac{1}{2}\, D(Q||M)$$

$$where: M = \frac{1}{2}\,(P + Q)$$

and *P* and *Q* represent the isoforms percentage distributions of the transcripts of the same gene in two different conditions. Following the example in Fig 7 P = [0.5, 0.2, 0.3] and Q = [0.06, 0.56, 0.38] and the JSD = 0.46.

The output of the Jansen-Shannon distance (JSD) is a number comprised between 0 and 1 that measures the similarity between the isoform percentage distributions in two conditions. A higher value of the JSD is correlated with a higher probability of observing changes in the transcript abundance percentages between two conditions. In other words, under the null hypotheses of no transcripts abundance percentage change between two conditions, the Jensen-Shannon distance is equal to 0.

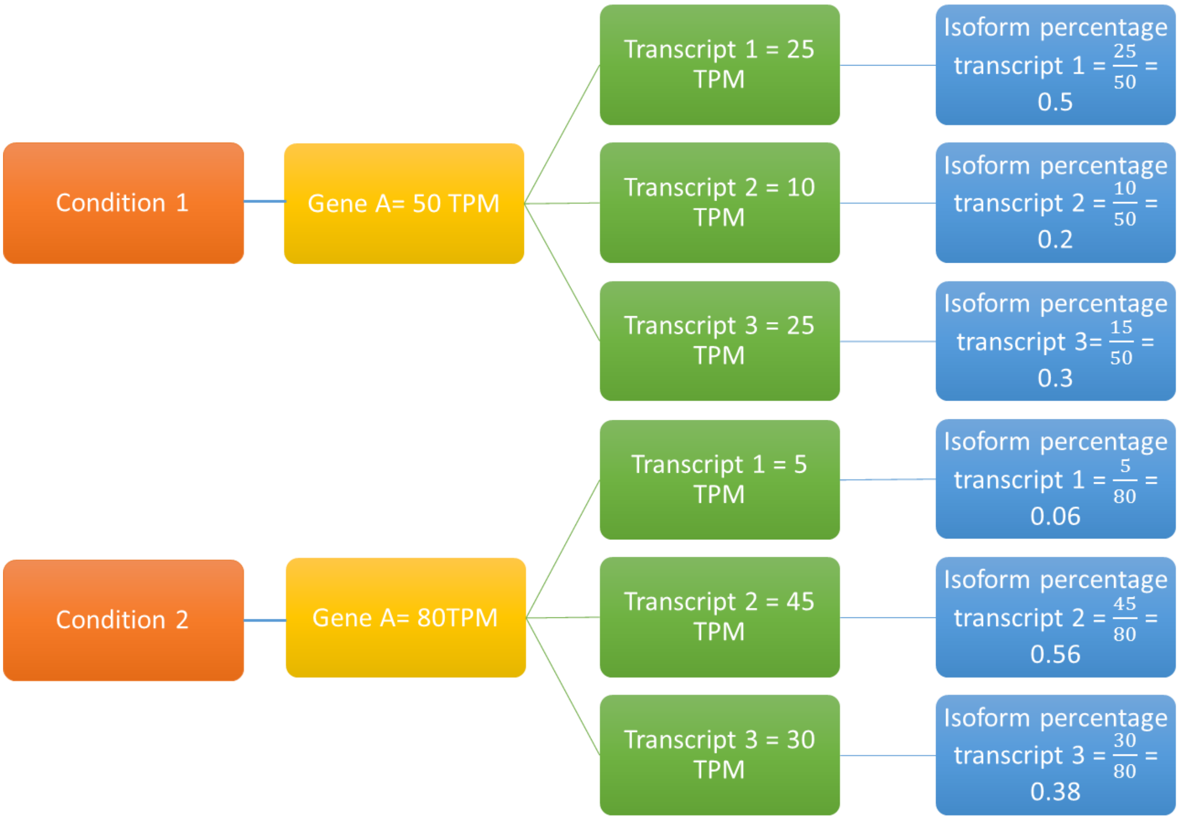

Condition 1
Gene A= 50 TPM
Transcript 1 = 25 TPM
Transcript 2 = 10 TPM
Transcript 3 = 25 TPM
Isoform percentage transcript 1 = $\frac{25}{50}$ = 0.5
Isoform percentage transcript 2 = $\frac{10}{50}$ = 0.2
Isoform percentage transcript 3= $\frac{15}{50}$ = 0.3
Condition 2
Gene A= 80TPM
Transcript 1 = 5 TPM
Transcript 2 = 45 TPM
Transcript 3 = 30 TPM
Isoform percentage transcript 1 = $\frac{5}{80}$ = 0.06
Isoform percentage transcript 2 = $\frac{45}{80}$ = 0.56
Isoform percentage transcript 3 = $\frac{30}{80}$ = 0.38

Figure 7: How the isoform percentage is calculated. The expression of Gene A is equal to 50 TPM in condition 1 and to 80 TPM in condition 2. The TPM are of Gene A are divided across the 3 transcripts. The transcript percentage is the ratio between the gene TPM and the transcript TPM in the two conditions. The values of the isoform percentage represent the isoform percentage distributions employed to calculate the JSD.

Isoform switching detection and filtering

To guarantee a highly dissimilar expression between the transcripts in two conditions, we selected only those genes whose JSD overcame the 95th percentile. To detect an isoform switch, each gene was associated with at least two transcripts. The transcripts were treated pairwise if a gene was transcribed in more than two transcripts.

So far the pipeline acted at the gene level. To properly select the transcripts undergoing an isoform switch the following criteria must be satisfied:

- transcripts whose expression is less than 1 TPM in at least one condition were not considered
- the log fold change of the transcripts must be opposite because we wanted to detect a switch of the isoform in two conditions. We used the following formula:

$$\log\left(\frac{Tr1_{Cond\,2}}{Tr1_{Cond\,1}}\right) > 0 \text{ and } \log\left(\frac{Tr2_{Cond\,2}}{Tr2_{Cond\,1}}\right) < 0 \text{ or vice versa}$$

- Additionally, to avoid the detection of very small changes in isoform switching, the difference in isoform percentage of the same transcript in two conditions must be greater than or equal to 10% in absolute value. In mathematical terms:

$$|(Tr1_{Cond\ 1} - Tr1_{Cond\ 2})| >= 0.1 \text{ and } |(Tr2_{Cond\ 1} - Tr2_{Cond\ 2})| >= 0.1$$

Regulation of alternative splicing and validation at proteomic level

Alternative splicing is regulated by trans-regulatory splicing factors, a subclass of RBPs that recognise short sequences located upstream and downstream of the exon start site. To detect those sequences and the associated splicing factors, we employed a pipeline similar to the one designed in ATtRACT. MEME was adopted to find ungapped motifs that were overrepresented in a set of sequences. Subsequently, Tomtom assessed whether the overrepresented motifs found by MEME were similar to the ones available in ATtRACT.

We run 12 different sessions of MEME one for each possible cell type and contrast. (Fig 8) (4 cell types: CMs, FBs, ECs and MFs, and three possible time courses: day 3 vs day 0, day 7 vs day 0, day 7 vs day 3). The zoops model was employed to assess the overrepresentation of motifs in a set of sequences, the zoops model assumes that each sequence may contain at most one occurrence of each motif. MEME was fed with sequences extracted from a window of 300 nucleotides upstream and downstream (600 nucleotides in total) from all the alternatively spliced exons. Each MEME session was characterised by a different set of sequences in input since different sets of alternatively spliced transcripts were found in each cell type and at each time point.

To better characterise the regulatory elements, we selected as background the sequences extracted from all the alternative spliced exons in the genome. Subsequently, we employed Tomtom to (i) assess whether the motifs found by MEME, with a p-value<0.05, resemble the ones available in ATtRACT, and (ii) to select, among all the possible RBPs, the potential splicing regulatory factor. To validate our findings, we took advantage of the data available in the proteomic profile. The proteomic profile was used to assess whether:

- the transcripts were translated into proteins
- the splicing factor was expressed at the protein level.

It is important to note that the proteomic profile is less accurate than the transcriptomics profile and a large fraction of the proteome is usually not detected (Laukens et al. 2015). Hence, the absence of a protein associated with an alternatively spliced transcript at the proteomic level may be due to (i) the post-transcriptional regulations or (ii) the impossibility to detect the protein.

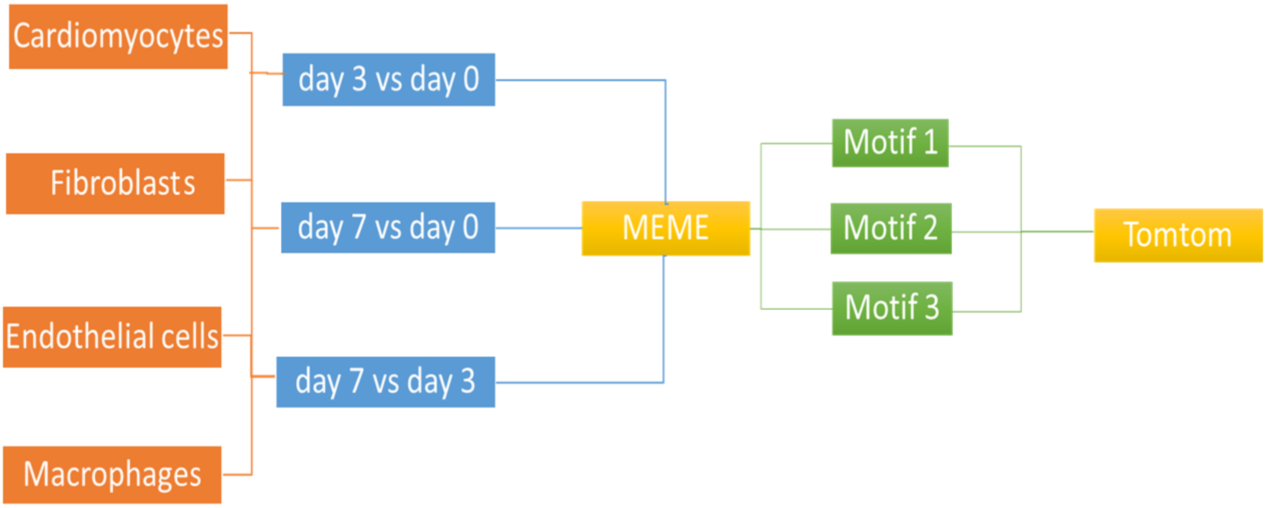


Figure 8: Pipeline adopted to detect the regulatory motifs. Twelve different sessions of MEME one for each possible cell type and contrast (4 cell types: CMs, FBs, ECs and MFs, and three possible time courses: day 3 vs day 0, day 7 vs day 0, day 7 vs day 3) were run. We fed Tomtom with the results obtain by MEME to assess whether the motifs found by MEME resemble the ones available in ATtRACT

Network based enrichment analysis

ORA methods are nowadays widely used to detect a functional profile of a genes/proteins list. We have described above in this section the ORA method (see paragraphs Fisher's exact test and Bonferroni correction). In the following paragraphs, we describe the resources and the algorithms employed to develop a novel ORA method that takes advantage of the PIN. We called this method MAGNETO: augMented functionAl enrichment analysis throuGh proteiN intEracTion netwOrk.

The algorithms designed in MAGNETO make intensive use of four databases to:

- extract the PIN
- identify the top-level pathways
- simplify the network according to the tissue protein expression
- filter the modules according to the top-level pathway annotation.

Databases employed

Despite the recent advances, the PPI detection methods are affected by high false positive rate (Sprinzak et al. 2003), hence it is fundamental that the PPI database contains manually curated interactions and that the procedures for the annotation of the experiments are rigidly controlled.

Currently, there are several databases containing PPI, including STRING (Szklarczyk et al. 2015), BioGRID (Breitkreutz et al. 2008) and IntAct (Orchard et al. 2014). The main difference between them lies in the quality of the annotation. In this respect, IntAct is one of the most stringent. IntAct is an open and freely available database that collects, analyses and stores molecular interaction data. The interactions are manually curated and updated very frequently. Furthermore IntAct is a member of HUPO Proteomics standards initiative (PSI) (Kerrien et al. 2007), and IMEX (Orchard et al. 2012).

Two are the most widely used manually curated databases of pathways, KEGG (Kanehisa et al. 2012) and Reactome (Fabregat et al. 2016). We chose Reactome, because, with respect to KEGG, it includes more proteins. There were 6970 annotated proteins in KEGG (release September 2016) and 10044 in Reactome (version 58).

Recently, several databases have been developed to assess tissue specific protein expression: Human Proteome map (release 29 May 2014) (Kim et al. 2014), ProteomicsDB (release 23 June 2016) (Wilhelm et al. 2014), and Human Protein Atlas (version 16) (HPA) (Uhlén et al. 2015). Amongst them, the HPA database is the most comprehensive one in terms of tissues composition (Table 3).

Table 3: Tissues composition across three different protein expression databases. Human Protein Atlas is the most comprehensive in term of represented tissues.

	Human proteome map	ProteomicsDB	Human Protein Atlas
Tissues	17	33	44

The UniProt-GOA (last release September 2016) (Huntley et al. 2015) database was chosen for the GO annotation. UniProt-GOA is a branch of the most famous Gene Ontology database, which has become the standard de facto for gene ontologies.

Protein-Protein Interaction Data: IntAct Database and generation of PPI networks

IntAct is an open and freely available database whose aim is to collect, analyze and store molecular interaction data. All the interaction data in IntAct derive from manually curated literature and have been annotated with experimental methods, conditions and interacting domains.

Over the years, the IntAct database has grown, with some exception, almost linearly (Fig 9). Nowadays it contains 93,856 proteins and 1,071,798 interactions out of which 61% (653,104) are binary interactions extracted from over 14,346 publications. Here, we present a brief overview of the IntAct database content.

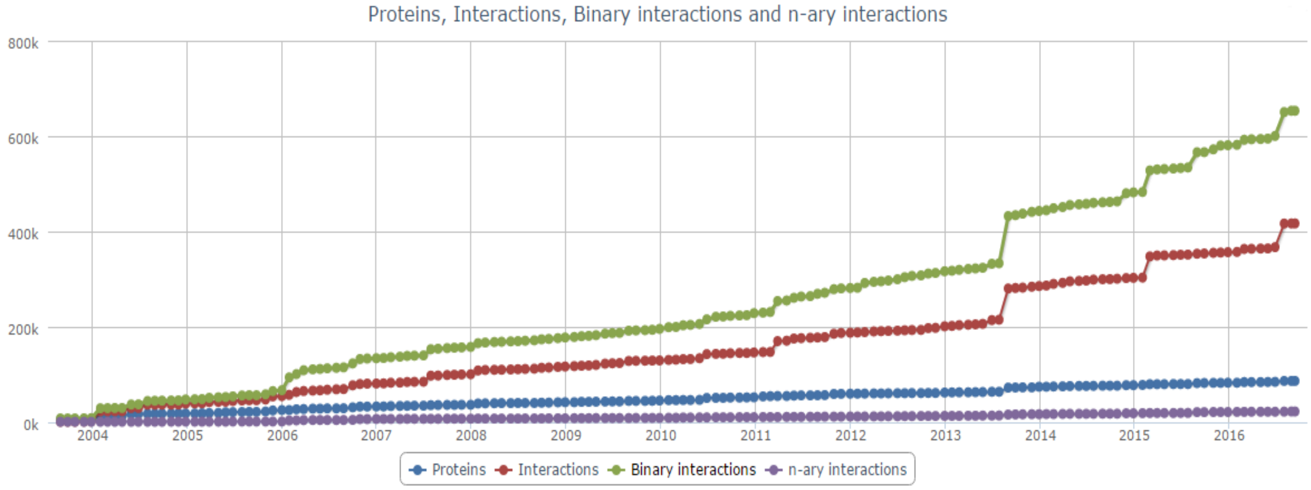


Figure 9: Total number of interactions available in IntAct. The number of interactions increased over time. Adapted from http://www.ebi.ac.uk/intact/about/statistics

The interaction types can be broadly classified into two groups: the physical association (54.9%), and the association (36.4%) (Fig 10).

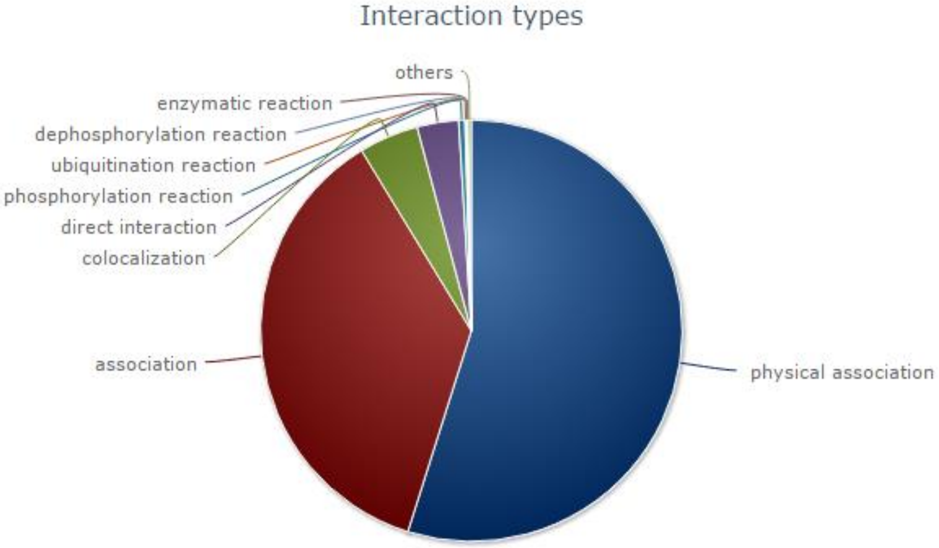


Figure 10: The physical association (54.9%) and the association (36.4%) are the most represented type of interactions in IntAct. Adapted from http://www.ebi.ac.uk/intact/about/statistics

An interaction is classified as *"physical association"* if the method captures only two interactors at a time (yeast two-hybrid) or it involves purified molecules (NMR or X-ray crystallography). The interaction is classified as *"association"* if multiple interactions are captured at the same time (i.e. co-immunoprecipitation, tandem affinity purification). Typically, these methods assess the interaction between a bait protein and an n-ary complex of proteins (prey). A common problem of the association methods regards the classification of the interaction between the bait and the prey since the interactions are annotated in the form of a binary interaction in the database.

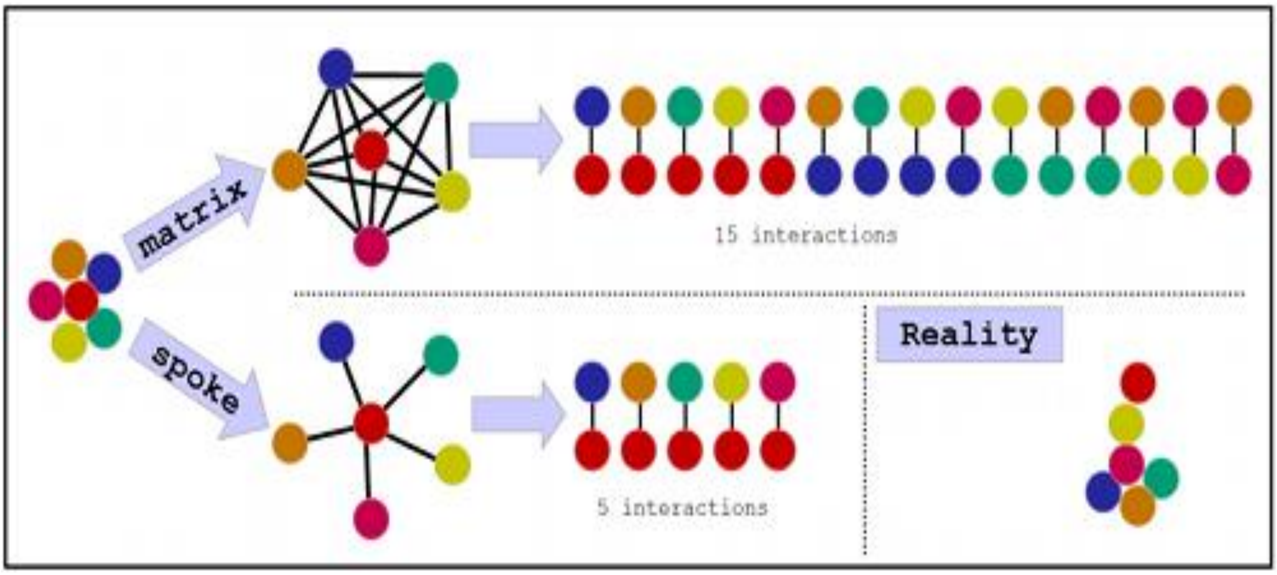


Figure 11: In this example a tandem affinity purification experiment is performed. The red dot is the bait protein and the others are the prey complex. With tne matrix exapansion model the complex is modelled as a complete graph and 15 interactions are detected. With the spoke expansion model only the bait interacts with the n-ary complexes and 5 interactions are detected. But, in reality, it may happen that the red protein has only one interactor, which would be the yellow one. Adapted from (http://www.ebi.ac.uk/intact/main.xhtml)

Two different models have been typically employed (Fig 11):

- In a *Matrix expansion* model, the molecules in the complex are linked together like in a complete graph. If the complex is composed of *N* proteins, it generates $N\times (N-1)$ binary interactions.
- In a *Spoke expansion* model, the binary interaction is identified only between the bait protein and each protein of the n-ary complex. If *N* is the number of proteins in a complex, *N-1* interactions are generated.

IntAct employs the spoke expansion method to avoid the proliferation of false positive interactions. In MAGNETO, two types of networks were extracted from IntAct (release 27th August 2016) (Table 4) and take into account the different level of confidence generated by the detection methods: (i) A high confidence PIN associated only with the low-throughput experiments (only physical association interactions), and a PIN that includes also the high-throughput experiments (physical association and association interactions).

It is evident that high-throughput experiments and spoke models allow to increase the number of edges in the network. This means that considering the IntAct network, a lesser number of nodes need to be traversed to connect two different nodes with respect to the IntAct high confidence network.

Table 4: Nodes and edges distribution of the two networks extracted from IntAct. The spoke model and the high throughput experiments allow increasing the number of edges. An increased number of edges allows decreasing the number of nodes connecting two paths of the network.

	IntAct High Confidence	**IntAct**
Nodes	12672	14802
Edges	60304	102345
Average degree	9.5	13.8

Tissue specific protein expression database: Human Protein Atlas (HPA)

The systematic exploration of the human proteome is fundamental to understand how the expression of the ~20,000 human protein-coding genes affects the corresponding proteins at the tissue level. The HPA database (version 16) contains protein expression data extracted from 44 different human tissues and organs (Fig 12). To derive the protein expression data, the immunohistochemical staining of cell populations in human tissues together with mRNA expression data are taken into consideration. Finally, the HPA database evaluates the expression levels of the proteins in each tissue with four levels: no expression, low, medium and high. In MAGNETO, each protein expression level is associated with a number (no expression=0.01, low=0.33, medium=0.66, high=1). We considered the average of the protein expression if the protein is detected in the same tissue multiple times (i.e. the same protein is detected high in the adipocytes and low in the fibroblasts in liver).

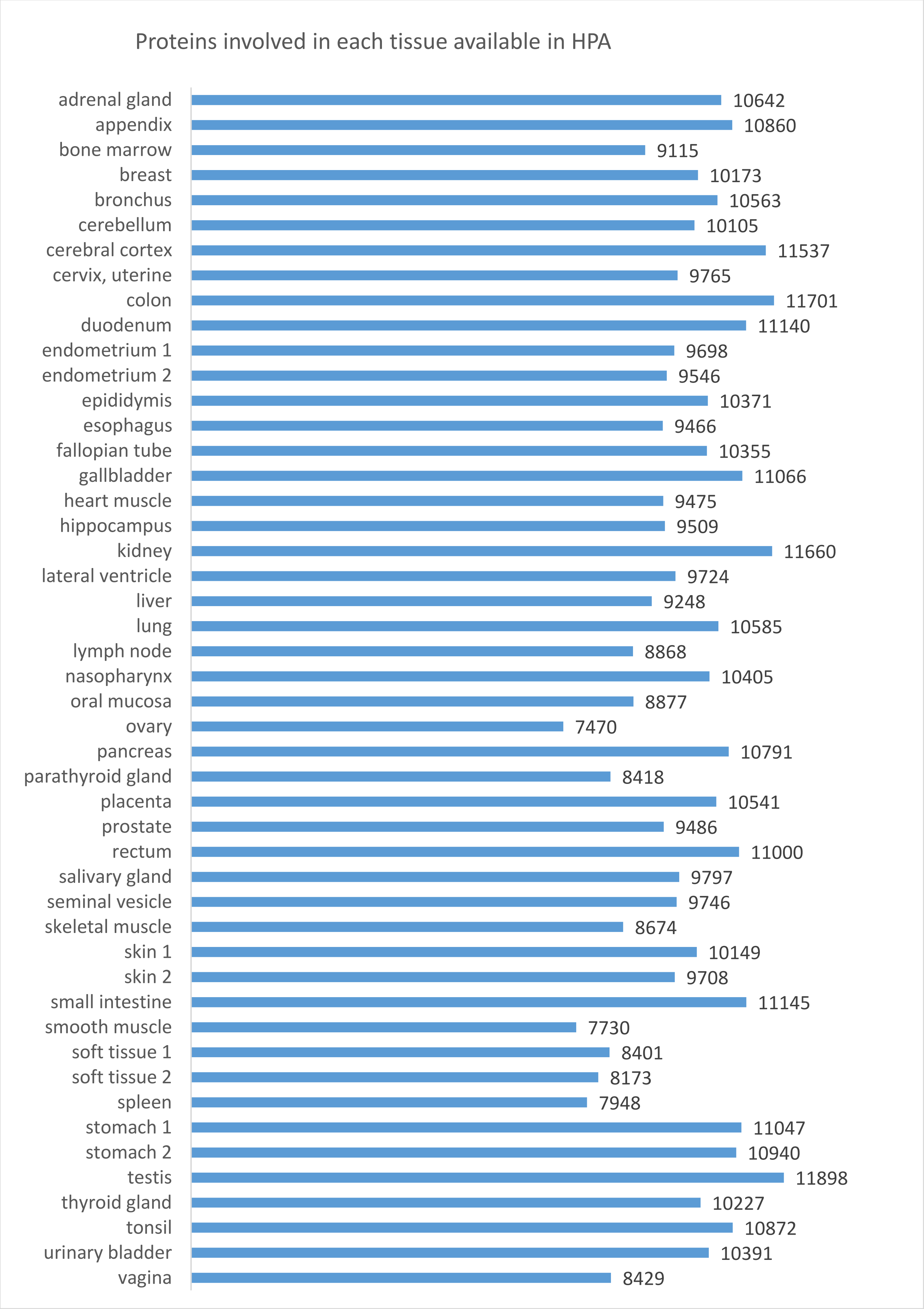


Figure 12: The tissues available in HPA and the number of proteins detected in each tissue**.** The number at the right of the bar represents the number of proteins involved in each tissue.

Pathways data: Reactome database

Reactome is a freely available database of manually curated pathways and reactions. In Reactome, the pathways are organized in a hierarchical way, meaning that moving down the hierarchy more specific biological pathways are found (e.g. signal transduction → signaling by FGFR → signaling by FGFR1). In Reactome, the top-level pathways represent the highest level of the hierarchy and are associated with a general biological process (e.g. hemostasis, gene expression, signal transduction).

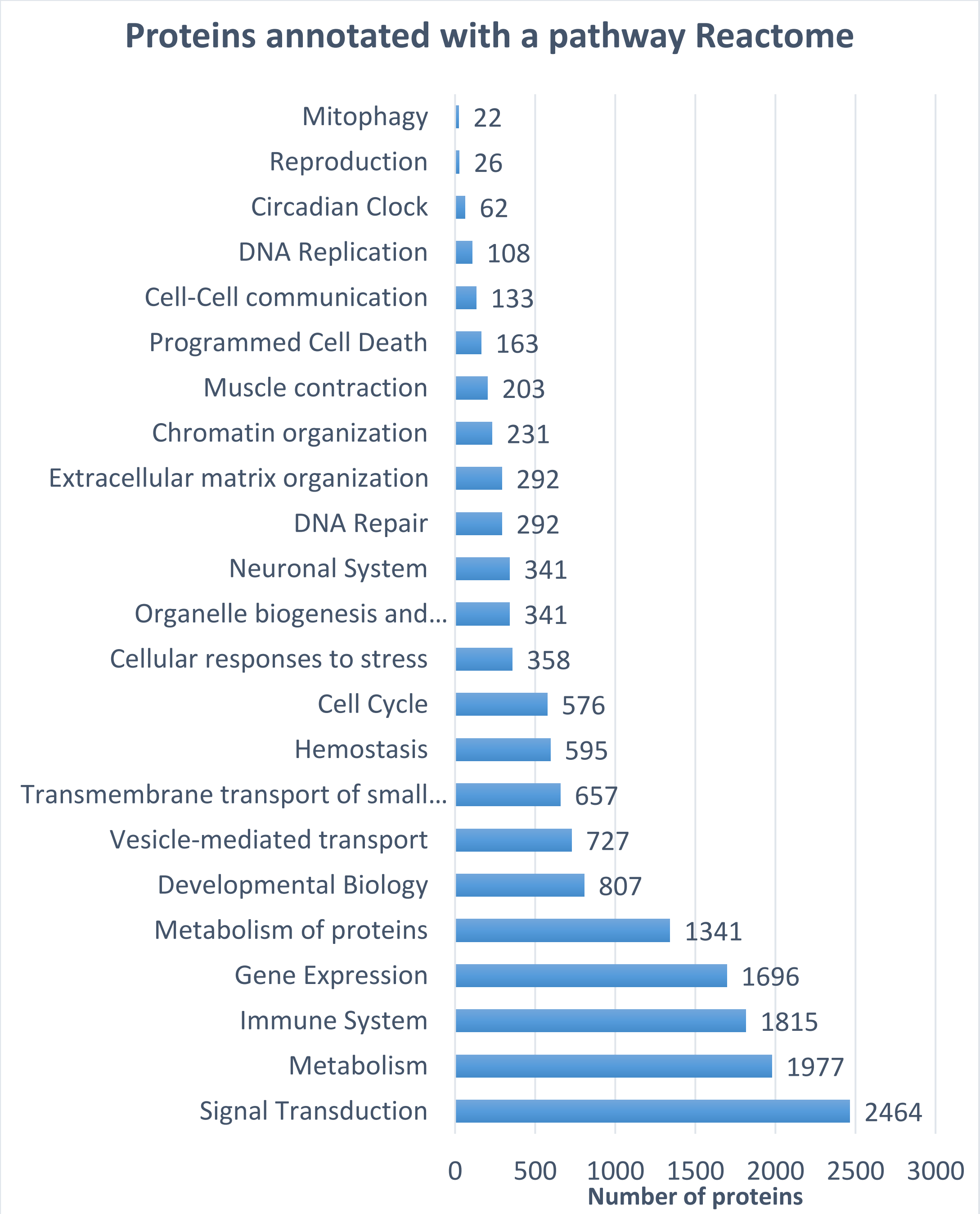


Figure 13: The 23 top-level pathways available in Reactome. The number at the right of the bar represents the number of proteins involved in each of the top-level pathways.

The proteins involved in a top-level pathway participate together in a common biological function, essentially, they represent a coarse-grained module (CGM). Twenty-four top-level pathways are available in Reactome (version V58), of these we excluded the top-level “Disease” pathway because all the proteins involved in this pathway are redundant since they are already included in the other top-level pathways. The histogram in Fig 13 shows the number of proteins involved in any of the 23 top-level pathways.

Graph theory: basic principles

PPIs can be naturally represented by means of graphs (network) in which the nodes represent the proteins and the edges their interactions. In this following paragraphs, we introduce some concepts that will be useful to understand the idea behind of MAGNETO.

Graphs and sub-graphs: some definitions

Graphs are mathematical structures, used to model a great variety of problems. A graph *G = (V, E)* is a pair of sets, where *V* is the set of vertices or nodes $\{V_1, V_2 \ldots V_n\}$ and *E* is a set of edges, such that to each element of *E* correspond a couple of elements of *V*.

Definition: *H = (W, F)* is a sub-graph of *G= (V, E)* if and only if *W ⊆V* and*F⊆E* (Fig 14)

Definition: Given a graph *G = (V, E),* a path $P = (V_1,V_2)$ is a set of distinct nodes $\{V_1, V_2, \ldots, Vn\}$, such that:

$$\exists \left(V_{1,}V_{i+1}\right) \in E \; \forall \, i \; \in \; [1, p-1]$$

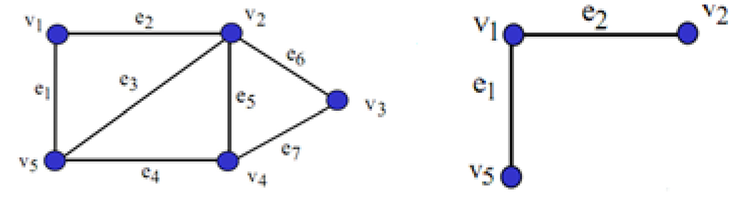


Figure 14: On the left a graph, on the right one of the possible sub graphs extracted from the graph on the left

Shortest path and all pairs shortest Paths

A path is a sequence of nodes that need to be traversed to connect the target and the source node. A path never passes two times over the same edge or node.

In an undirected graph, the shortest path (SP) $d_{i,j}$ between two nodes V_i and V_j represents the minimum distance connecting two nodes. The distance between two nodes is defined as:

$$\begin{cases} \mathrm{d_{i,j}} = \{\min\left(\mathrm{V_i}, V_\mathrm{j}\right) \text{ if the path exist} \\ \qquad\qquad \infty \text{ otherwise} \end{cases}$$

Potentially, there are many paths between two pairs of nodes in a graph. Given a pair of nodes, determining all the possible shortest paths connecting them is called all-pairs shortest path (APSP) problem. Computing the APSP for large graphs like a PIN is a computationally expensive (*$O(n^3)$*) task. To decrease the computational time, in MAGNETO, we implemented a C++ algorithm based on the igraph (http://igraph.org) library, allowing to reduce of several orders of magnitude the time with respect to a python based algorithm.

MAGNETO workflow

The main aim of MAGNETO is to identify modules, meaning sets of interacting proteins involved in a common biological function. To accomplish this task, we adopted a data-driven approach in order to combine the advantages of the topology-based methods and module-based methods.

MAGNETO requires in input a protein list and a tissue chosen from the 44 available in HPA.

The algorithm developed in MAGNETO involves four main phases:

1. A preprocessing phase in which the input proteins are mapped into the 23 top-level pathways available in Reactome.
2. A Connecting phase in which MAGNETO extracts, for each of the top-level pathways, the Minimal Connecting Network (MCN). I.e. the network composed of the union of all shortest paths between each possible pair of proteins.
3. A Simplification phase in which MAGNETO, removes the paths that are barely expressed in the selected tissue.
4. A Maximization phase in which MAGNETO selects the paths whose annotation is highly representative of the CGM.

Definition: The proteins assigned to a CGM are called seed proteins.

Preprocessing phase: mapping the seed proteins to Reactome

During the preprocessing phase, the seed proteins were mapped into the CGMs allowing to define coarse-grained groups of proteins involved in the same biological process. It is important to notice that a seed protein can be involved and assigned to more than one CGM. Since Reactome contains only a small fraction of the human proteome, it could happen that some of the seed proteins remain unmapped. To solve this problem, we started from the consideration that since proteins with similar functions are involved in the same biological process, then proteins sharing similar biological processes should participate in the same CGM. Hence, we extracted a GO profile (GOP) from each of the CGM. The GOP is, essentially, a ranking table representing the probability of finding a BP in the CGM. For example, the first 5 GO terms involved in the vesicle-mediated transport CGM are: Transport, protein transport, vesicle-mediated transport, receptor-mediated endocytosis, and ER to Golgi vesicle-mediated transport. The proteins lacking Reactome annotation are ranked according to the GOP probability. For each unmapped protein, 23 scores are generated, one for each CGM. The score represents the sum of the probabilities of finding the BPs in which the proteins are involved in a CGM. The probabilities are calculated according to the probabilities extracted from each one of the GOP profiles. The unmapped protein is assigned to the CGM by selecting the one with the highest score, in other words, the higher the score is, the higher is the similarity between the biological processes in which the protein is annotated and the biological processes of the CGM.

Connecting phase: computing the Minimal Connecting Network

After all the proteins in the input list were assigned to at least a CGM, the connecting phase takes place. This phase is the most computationally expensive of the whole MAGNETO workflow. For each set of proteins in a CGM, MAGNETO computes the graph constructed by the union of all the shortest paths between any possible pair of nodes involved in a CGM. The resulting network is called minimal connecting interaction network (MCN). The extraction of the MCN involves two main steps.

During the first step, MAGNETO computes all the possible combinations of protein pairs in any of the CGMs. Formally, given a set S of objects (i.e. the seed proteins in a CGM), how many combinations can assume k objects selected from this set. It is a combinatorial problem and the number of possible combination of n elements of length k is given by the formula:

$$\mathrm{C_{n,k}} = \binom{\mathrm{n}}{\mathrm{k}} = \frac{\mathrm{n!}}{\mathrm{k!\,(n-k)!}}$$

Where n is the number of proteins in the CGM, and k is equal to two, corresponding to a protein pair. From this procedure, we exclude the CGMs with only one protein. During the second step, MAGNETO, for each pair of proteins previously extracted, computes all the possible shortest paths.

The MCN permits to understand how the proteins involved in the same biological process are connected to each other in terms of shortest paths. In other words, the MCN allows highlighting the underlying interaction network for each of the CGM. Typically, many MCNs are computed in a single run of MAGNETO, the total number depends on how the seed proteins are assigned to the respective CGM (Fig 15).

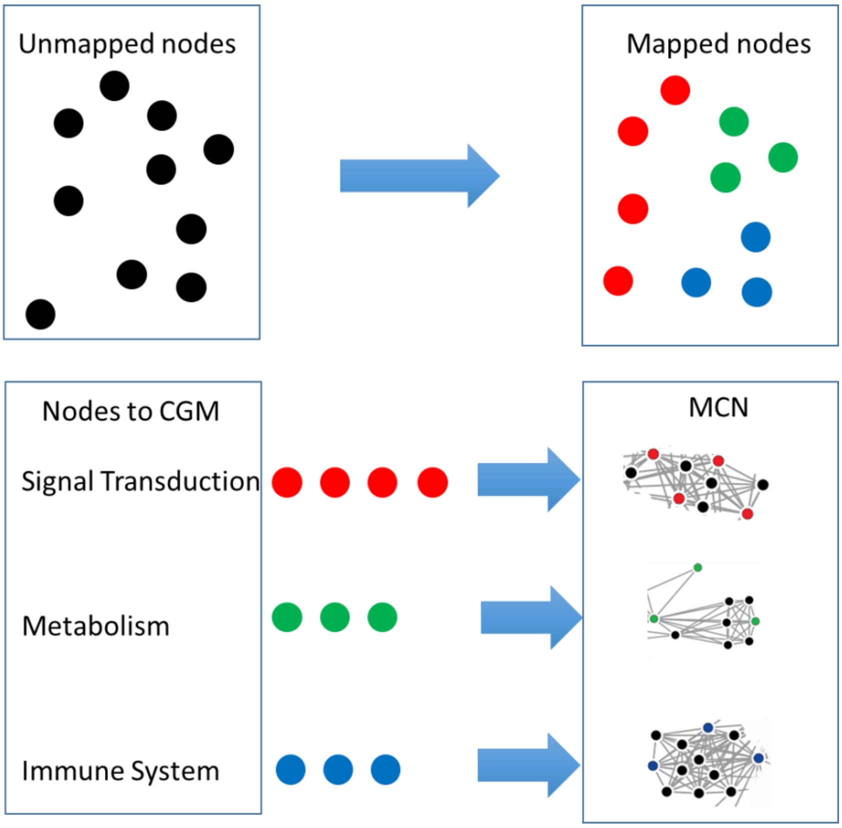


Figure 15: The genes/proteins in input are mapped to one or more CGM. Then for each CGM the MCN is computed by the union of all the shortest paths of the seed nodes.

It is important to note that, typically, there are many shortest paths between two seed proteins. Additionally, excluding the isolated nodes, the obtained MCN is sub-isomorphic to the original graph and constitutes the backbone for the filter phase.

Simplification phase: filter the paths based on tissue expression

Even for medium-small protein sets, the MCNs cover a large fraction of the whole interactome and, therefore, are not representative of the underlying biological processes. To overcome this problem a graph-simplifying (Ruan et al. 2011) procedure is implemented. Simplifying means to eliminate "unimportant" nodes and associated edges from the MCN while preserving its most important characteristics.

In MAGNETO, we applied a graph filtering approach with the difference that instead of removing a single node at a time, all the nodes connecting the source and the target but not the source and the target, are removed in a single run. Initially, the paths that are expressed at low level in the selected tissue are removed.

MAGNETO assigns a value to each node/protein connecting the seed proteins. The node values represent the protein expression extracted from the HPA in the tissue selected by the user. If the protein expression is missing in the selected tissue, MAGNETO assigns a pseudo value of 0.01.

Typically, two seed nodes are connected by many paths, MAGNETO computes the average expression for each path. The vector obtained corresponds to the average expression of all the paths connecting two seed nodes. The mean of the vector corresponds to the threshold assigned by MAGNETO in order to discard those paths that are expressed at a low level in the selected tissue. In other words, during the simplification phase, for each pair of proteins of the input list, those paths that are the most highly expressed in the selected tissue are selected (Fig 16A).

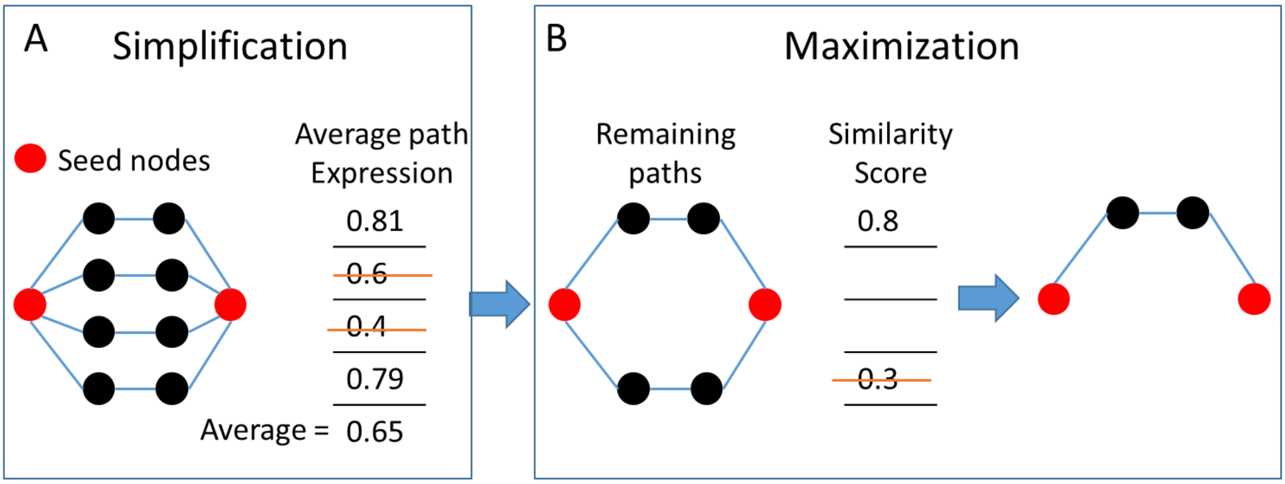


Figure 16: The red nodes are the seed nodes; the black nodes are the nodes involved in the shortest paths. For each path, the average path expression is calculated. The average of the average path expression is the threshold for selecting the paths that are highly expressed in the tissue(A). For the remaining paths, the similarity score is computed. The similarity score represents the likelihood between the biological processes go terms in the paths and in the CGM. The path with the highest rank is selected (B).

Maximization phase: filter the paths based on annotation

The CGM networks obtained previously are further filtered. After the simplification phase, potentially, there are still several paths connecting two seed nodes. The aim of the maximization phase is to rank each path connecting two seed nodes. The rank is given by the sum of the probabilities of finding the BPs, in which the proteins involved in the path are annotated, in the GOP. The selected path is the one that shows the highest rank, i.e. the highest similarity between the annotations of the proteins involved in the path and the GOP of the CGM under investigation (Fig 16B).

From each of the CGM, a subnetwork extracted from the PIN is obtained. The subnetwork is (i) representative of the biological processes that take place in the CGM and (ii) of the tissue-specific interactions that occur between the seed proteins.

Finally, all the nodes extracted from each of the CGM represent the testing set for the ORA.

Databases integrated in MAGNETO

To have a broad view and a deeper understanding of the datasets under investigation, several databases were integrated into MAGNETO. They can be grouped into 5 main categories:

1) Biological pathway databases:
   a) Reactome (Version 58) (Fabregat et al. 2016)
   b) KEGG pathway (Release September 2016) (Kanehisa et al. 2012)
2) Drug databases:
   a) KEGG drugs (Release September 2016) (Kanehisa 2013)
   b) DrugBank (Version 5.0) (Law et al. 2014)
3) Disease databases
   a) OMIM (Release September 2016) (Amberger et al. 2015)
   b) Orphanet (Release September 2016) (Davies 2016)
   c) KEGG disease (Release September 2016) (Kanehisa 2013)
   d) DisGeNet (Release October 2016) (Queralt-Rosinach et al. 2016)
4) Toxin database:
   a) T3DB (Version 2.0) (Lim et al. 2010)
5) Host - Pathogen Interaction database
   a) HPIDB (Version 2.0) (Ammari et al. 2016)

Magneto performance evaluation

One of the main challenges common to all the ORA methods regards the assessment of their performances. The fact that most of the overrepresented terms can be supported by some references in the literature makes the task of evaluating the performance of the ORA methods difficult. To compare MAGNETO with ORA we employed the method proposed by Tarca et al. 2012. Twenty-one datasets (Table 5), associated with a pathology and for which exist a tissue in HPA, were selected from the GEO database (Barrett et al. 2013), each one of these datasets could be associated with a target pathway in KEGG database (i.e. the colorectal cancer is the target pathway from all the GEO dataset studying colorectal cancer). MAGNETO and the standard ORA method were compared in terms of (i) distribution of the -Log(p-values) of the target pathway (the higher the better) and (ii) on the ability of the method to rank the target pathway on the top of the overrepresented terms list (the lower the better). The rank was normalized between 0 and 100 using the formula:

$$rank = \frac{i * 100}{N}$$

where *i* was the rank of the target pathway in the overrepresented list of terms and *N* is equal to 214 representing the total number of pathways in KEGG excluding the metabolism pathways. It could happen that none of the proteins in the input list is associated with the target pathway, therefore a p-value cannot be provided, in this case, the p-value is equal to 1 and the rank is equal to 214.

Table 5: The 21 pathways selected for the benchmark. The first column represents the GEO ID, the second column represents the literature reference (if exists), the third the target pathway and the fourth the associated KEGG id, the fifth column represents the tissue were the experiment were performed and the sixth the tissue selected in MAGNETO

GEOID	Pubmed	Disease/Target pathway	KEGGID	Tissue	Tissue in MAGNETO
GSE1297	14769913	Alzheimer's Disease	hsa05010	Hippocampal CA1	Cerebral cortex
GSE5281	17077275	Alzheimer's Disease	hsa05010	Brain, Entorhinal Cortex	Cerebral cortex
GSE5281	17077275	Alzheimer's Disease	hsa05010	Brain, hippocampus	hippocampus
GSE5281	17077275	Alzheimer's Disease	hsa05010	Brain, Primary visual cortex	Cerebral cortex
GSE20291	15965975	Parkinson's disease	hsa05012	Postmortem brain putamen	Cerebral cortex
GSE4107	17317818	Colorectal Cancer	hsa05210	Mucosa	Colon
GSE8671	18171984	Colorectal Cancer	hsa05210	Colon	Colon
GSE9348	20143136	Colorectal Cancer	hsa05210	Colon	Colon
GSE14762	19252501	Renal Cancer	hsa05211	Kidney	Kidney
GSE781	14641932	Renal Cancer	hsa05211	Kidney	Kidney
GSE15471	19260470	Pancreatic Cancer	hsa05212	Pancreas	Pancreas
GSE16515	19732725	Pancreatic Cancer	hsa05212	Pancreas	Pancreas
GSE19728		Glioma	hsa05214	Brain	Cerebral cortex
GSE21354		Glioma	hsa05214	Brain, Spine	Cerebral cortex
GSE6956	18245496	Prostate Cancer	hsa05215	Prostate	Prostate
GSE6956	18245496	Prostate Cancer	hsa05215	Prostate	Prostate
GSE3467	16365291	Thyroid Cancer	hsa05216	Thyroid	Thyroid gland
GSE3678		Thyroid Cancer	hsa05216	Thyroid	Thyroid gland
GSE18842	20878980	Non-Small Cell Lung Cancer	hsa05223	Lung	Lung
GSE19188	20421987	Non-Small Cell Lung Cancer	hsa05223	Lung	Lung
GSE3585	17045896	Dilated cardiomyopathy	hsa05414	Heart	Heart muscle

The inputs for both MAGNETO and ORA were lists of increased length (10, 50, 100, 200) of differentially expressed genes. The pipeline employed to calculate the differentially expressed genes was the same used by Tarca et al. 2012. Briefly, the datasets were normalized using the RMA algorithm (Irizarry et al. 2003) available with the affy package (Gautier et al. 2004) of Bioconductor (Gentleman et al. 2004). The differential expression between the two sample groups was computed using the limma package (Smyth 2004). The differently expressed genes were selected if the false discovery rate was less than 0.1 and then upregulated genes were ranked according to their log fold change.

Results

The results section is divided into three main subsections. In the first subsection, we present ATtRACT a database collecting the RBPs and associated binding sites. In the second subsection, we present the results regarding the analyses and the regulation of alternative splicing in the pig infarcted myocardium. In the third subsection, we present the results obtained with MAGNETO.

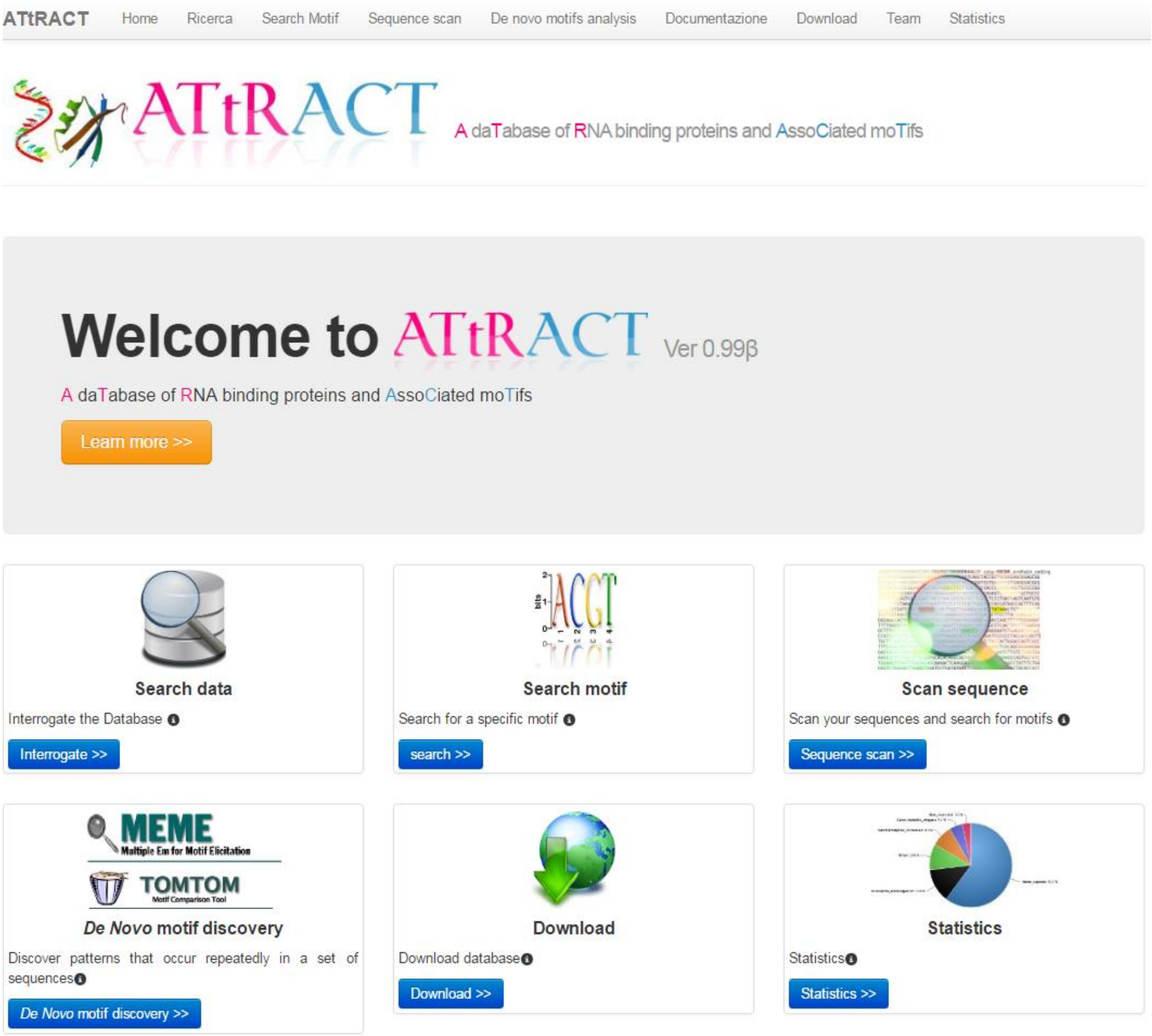
ATtRACT

ATtRACT interface

ATtRACT integrates a user-friendly interface. The interface allows the users to access to six main sections (Fig 17). The first section allows querying the database. The second allows the searches for a specific motif. The third allows scanning one or more RNA sequences searching for RBP binding sites. The fourth allows discovering enriched motifs in a set of related sequences and comparing them with the motifs present in ATtRACT. The fifth allows downloading part of or the entire database. The sixth allows to access to the statistics of the database. Each element of the interface is associated with images, tooltips, and simple explanations to facilitate user experience and usability.

ATtRACT content

ATtRACT contains information on 370 hand-curated and experimentally validated RBPs associated with 1583 consensus motifs out of which 192 were not present in any other database and they account for the 15% of the total content of ATtRACT database.



A daTabase of RNA binding proteins and AssoCiated moTifs

Welcome to ATtRACT Ver 0.99β

A daTabase of RNA binding proteins and AssoCiated moTifs

Search data

Interrogate the Database

Search motif

Search for a specific motif

Scan sequence

Scan your sequences and search for motifs

De Novo motif discovery

Discover patterns that occur repeatedly in a set of sequences

Download

Download database

Statistics

Statistics

Figure 17: Home page and user interface of ATtRACT. The page is divided into 6 main sections. The first section allows querying the database. The second searches for a specific motif. The third scans one or more RNA sequences searching for RBP binding sites. The fourth allows discovering enriched motifs in a set of related sequences and comparing them with the motifs present in ATtRACT. The fifth allows downloading part of or the entire database. The sixth permits to have access to the statistics of the database.

Table 6 shows the percentage of the experimentally validated motifs extracted from each database included in ATtRACT and in parenthesis the percentage that they represent.

Note that the table takes into consideration the binding specificity of the RBP, the experiments, and the organisms, meaning that a motif is considered a distinct entry if: (i) it binds to a different RBP or (ii) it was identified with a different experimental approach or (iii) it was identified in a different organism. E.g. the motif "ACGCGCC" is considered as two distinct entries because it binds either SRSF1 or RBM8A.

Table 6: Total and percentage of experimentally validated consensus motifs extracted from each database. The first row represents the number of motifs from each database included in ATtRACT and the percentage (in parenthesis) that they represent. The second row represents the number of unique motifs, i.e. motifs that are not present in the remaining database. Note that ATtRACT might contain redundant motifs if they were defined based on different methods. From that the discrepancy between the total number of motifs (1583) and those unique (1434).

Database	**CISBP-RNA**	**RBPDB**	**SpliceAid-F**	**AEDB**	**PDB**
Number of consensus motifs	312 (19.7%)	226 (14.3%)	775 (48.9%)	95 (6.0%)	256 (16.2%)
Unique motifs	229 (17.9%)	120 (9.4%)	659 (51.5%)	79 (6.2%)	192 (15.0%)

ATtRACT statistics

ATtRACT contains only motifs whose length varies from 4 to 12 nucleotides. The motifs of 7 nucleotides long are the most represented (Fig 18).

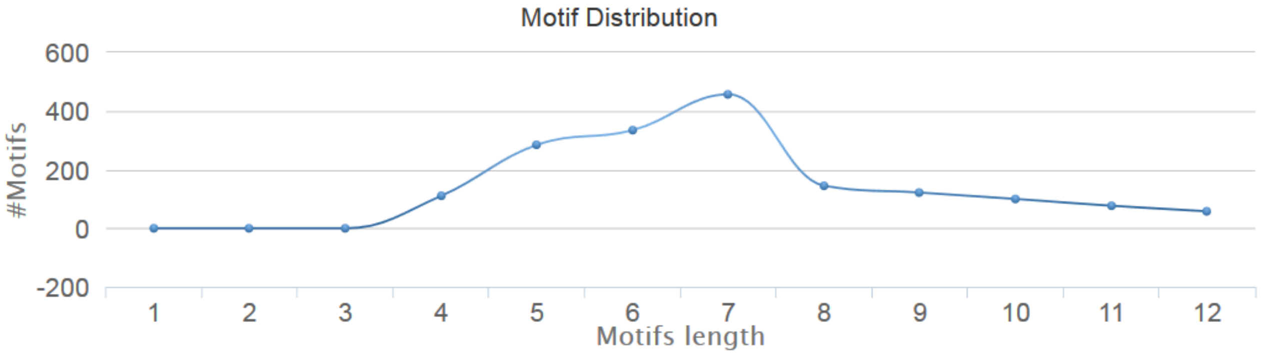


Figure 18: The graph represents the distribution of motifs in ATtRACT. The motifs of 7 nucleotides long are the most represented.

In total, ATtRACT contains motifs from 38 different eukaryotic organisms (Fig 19). The most represented organism, in terms of motifs, is Homo sapiens with 67.2% of the total motifs. Drosophila melanogaster follows with 10.5% of the total motifs.

The most representative experiments in ATtRACT are the SELEX and the UV cross-linking (Fig 20). Notice that this statistic takes into account only the experiment and not the motifs that are generated by the experiments.

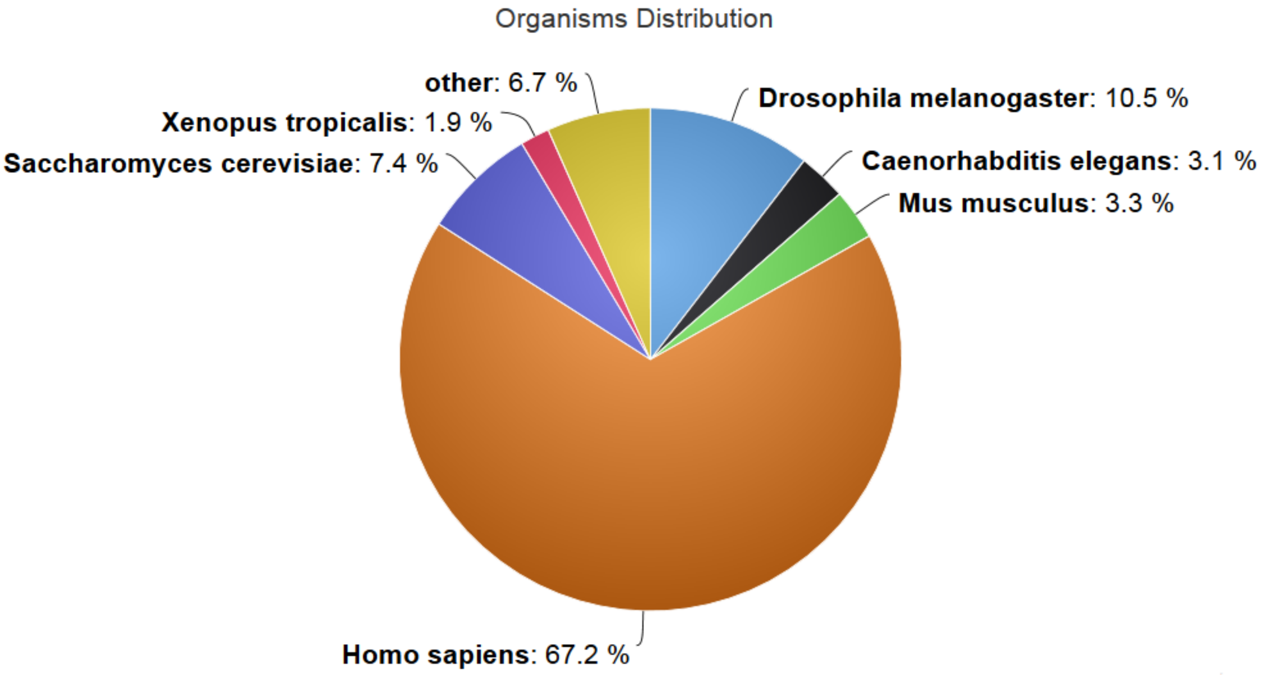


Figure 19: The pie chart represents distribution of organisms in ATtRACT. The human motifs are the most represented with the 67.2% of the whole database

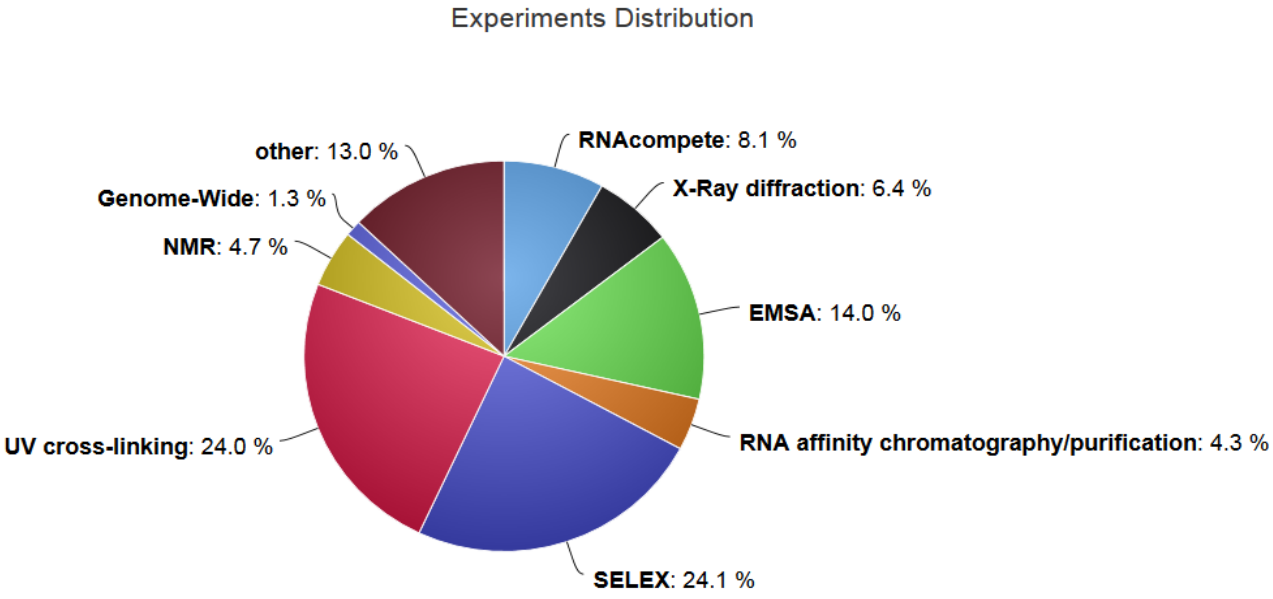


Figure 20: The pie chart represents the distribution of motifs by experimental approach in ATtRACT. The SELEX experiments and the UV cross-linking are the most represented.

The most represented domains are the RRM and the KH. They account for more than the 70% of all the domains in ATtRACT (Fig 21).

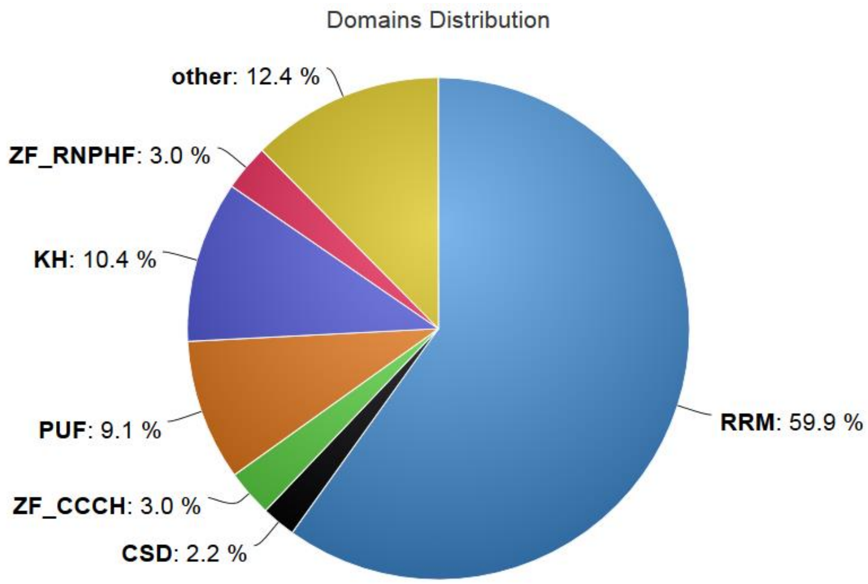
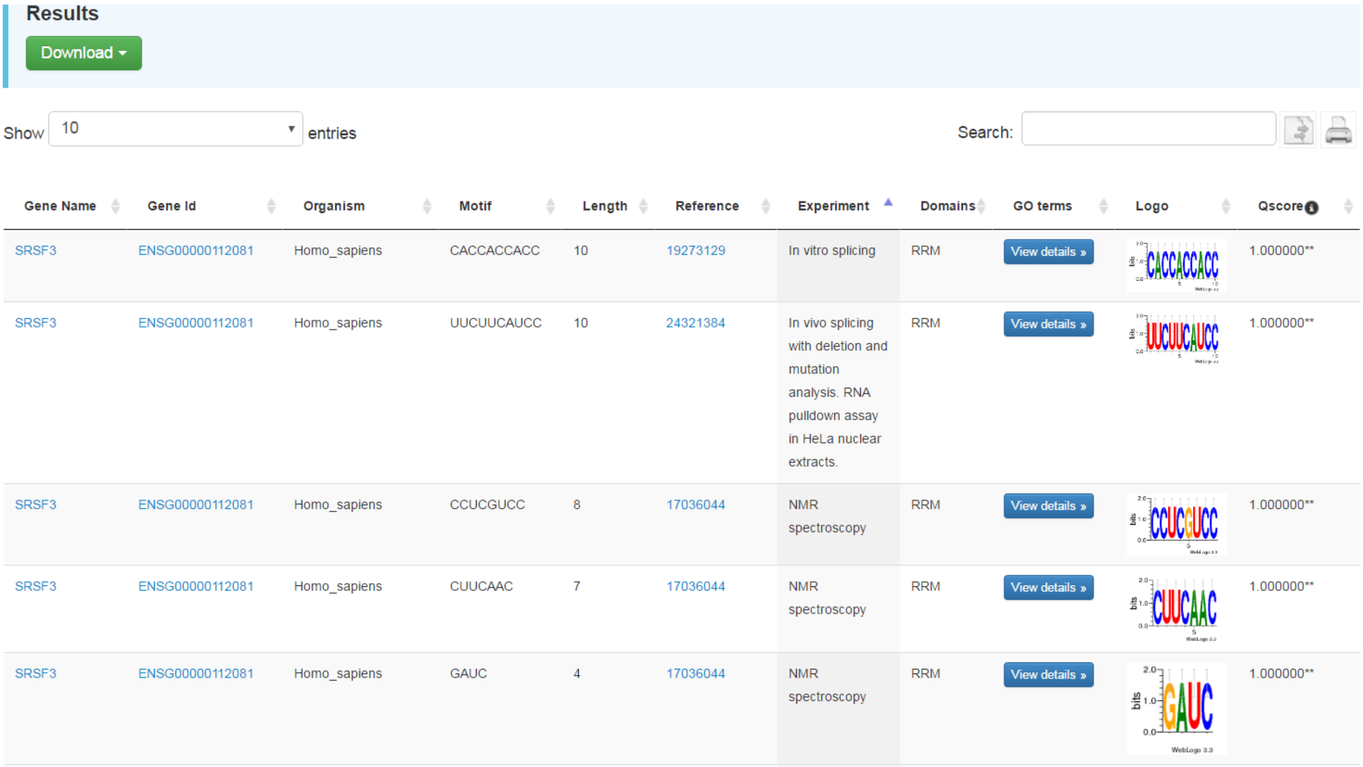


Figure 21: The pie chart represents the domains distribution available in ATtRACT. The RRM and the KH domains are the most represented.

Search for RBPs

Users can interrogate ATtRACT through a broad range of different queries. Users can search information about specific entries of the database simply by typing or choosing one or a combination of the following options: Official Gene name (i.e.: "SRSF4"), Synonyms (i.e. "SFRS4") Gene ID (i.e. "ENSG00000136450"), Minimum or maximum length of motifs, Type of experiments, Organisms and/or Domains. Search criteria can be combined by using combinations of queries. The results are displayed in tabular format (Fig 22). The file, containing the search results, can be downloaded by clicking on the drop-down menu on the top of the page. Two different file formats are available for download: the commas and the tab separated value. The users can further search the entries of the table through a whole text search using the search box. It is also possible to print or copy the results to the clipboard. The tables are sortable simply by clicking on the arrows in the header of the table. The following fields are displayed on the table: gene name, gene identifier, organism, binding sites, PubMed identifier, type of experiment, domain, gene ontology, the sequence logo and quality score. One of the main characteristics of ATtRACT is the integration with the GO database; by clicking the corresponding button in the go terms column, users can investigate the GO terms associated to RBPs. A modal window will show the cellular components, the molecular functions and the biological processes associated to the RBP.



Results

Gene Name	Gene Id	Organism	Motif	Length	Reference	Experiment	Domains	GO terms	Logo	Qscore
SRSF3	ENSG00000112081	Homo_sapiens	CACCACCACC	10	19273129	In vitro splicing	RRM	View details »		1.000000**
SRSF3	ENSG00000112081	Homo_sapiens	UUCUUCAUCC	10	24321384	In vivo splicing with deletion and mutation analysis. RNA pulldown assay in HeLa nuclear extracts.	RRM	View details »		1.000000**
SRSF3	ENSG00000112081	Homo_sapiens	CCUCGUCC	8	17036044	NMR spectroscopy	RRM	View details »		1.000000**
SRSF3	ENSG00000112081	Homo_sapiens	CUUCAAC	7	17036044	NMR spectroscopy	RRM	View details »		1.000000**
SRSF3	ENSG00000112081	Homo_sapiens	GAUC	4	17036044	NMR spectroscopy	RRM	View details »		1.000000**

Figure 22: The typical output of ATtRACT database. The columns represent in order: the gene name, the gene identifier, the organism, the motif, the length of the motif, the PubMed identifier, the experiment, the domain(s), the GO terms, the logo and the Qscore.

A hyperlink is available in the cell table corresponding to the gene name and the gene id. The browser redirects to the UniProt database if the gene name is clicked, otherwise, the browser redirects to the page associated with the Ensembl database if the gene identifier is clicked. In few cases, the redirection, due to the lack of annotation, occurs through other repositories. Moreover, when the experiment is NMR or X-ray we have added a hyperlink to the corresponding PDB entry in the column experiment. Finally clicking on motif logo is possible to download the position-specific probability matrix (PPM).

Search for a specific motif

Users can search for specific motifs by submitting a sequence ranging from 4 to 12 nucleotides. The search module supports IUPAC ambiguous notation. By default, the search module retrieves any motif that contains the sequence in the input. A perfect match search is executed if the motif is enclosed between quotes. The results are displayed as a table and are available for download. The same fields mentioned in the previous paragraph (see Search for RBPs paragraph and Fig 22) are displayed.

Scan sequence interface

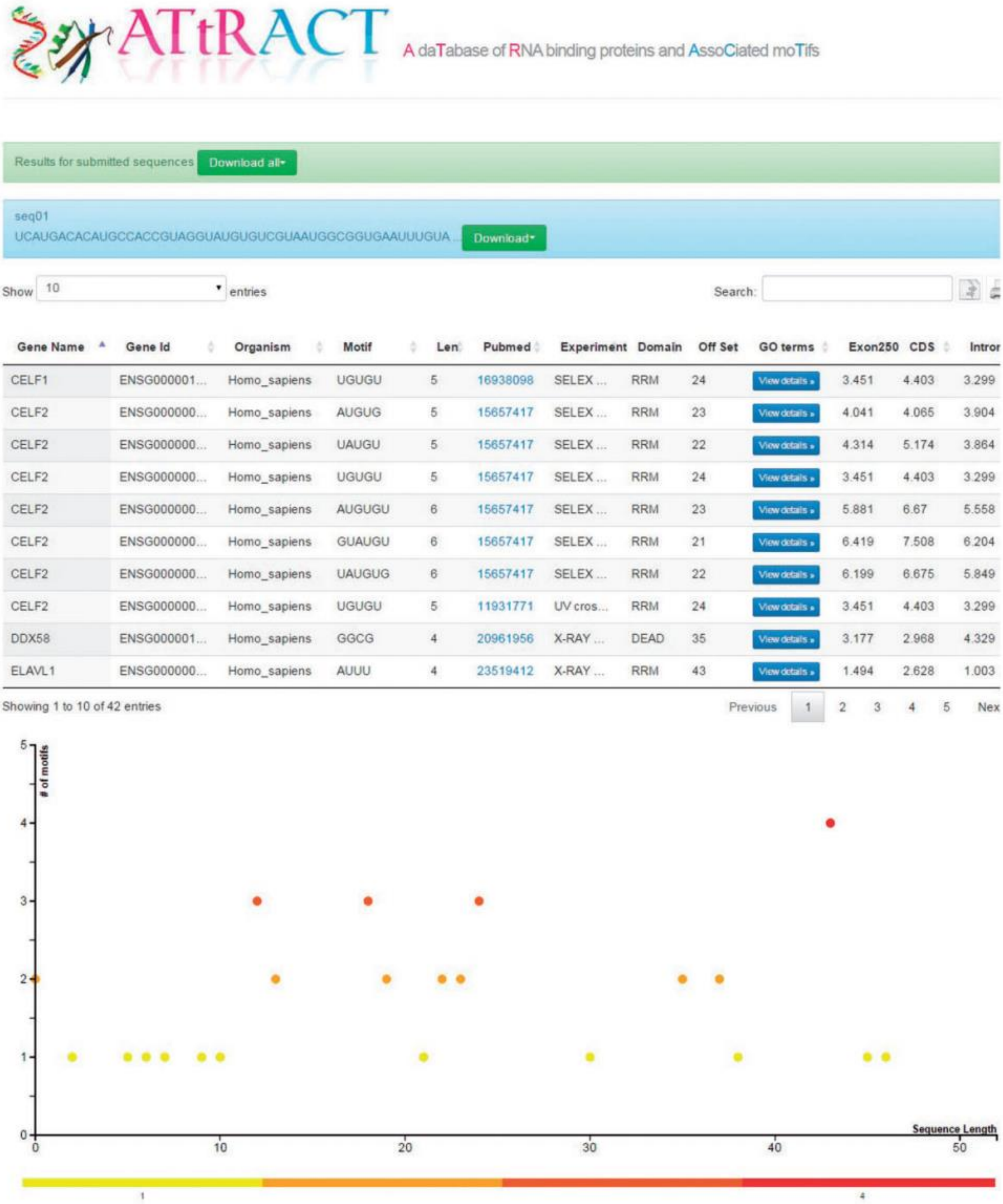
In ATtRACT, the BWT algorithm is implemented to scan the sequences. Users can upload a file containing one or more RNA/DNA sequences in fasta or multi-fasta format and scan the file searching for the presence of motifs. The user can restrict the search by selecting a specific organism and/or motifs of a certain length. The results are provided in tabular format (Fig 23) and are downloadable. With respect to the search RBPs and motifs modules, four more fields are added to the headers of the results tables: (i) the offset (ii) Exon250, (iii) CDS, (iv) intron (please refer to the Materials and Methods section). The offset represents the distance in terms of nucleotides at which it is possible to locate the motif, starting from the beginning of the sequence. In order to better visualize the results, a plot is provided by ATtRACT for each scanned sequence. The x-axis represents the sequence length; each bin represents a nucleotide of the input sequence. The number of motifs is represented in the y-axis. Each point in the graph represents the starting position of a motif across the input sequence. The higher the dot is on the y-axis, the more motifs start in that position. A red-yellow color scale distinguishes the dots. The redder the dots are, the higher the concentration of motifs is. It is possible to zoom-in the graph with the mouse wheel and interact with it by clicking on the dots. A table will show all RBPs, the motifs and the organisms associated with that position.

MEME and Tomtom Interface

ATtRACT integrated a pipeline for *de novo* motif discovery. Users can upload:

- a set of sequences in multi-fasta format (3000 nucleotides maximum). MEME will analyze the sequences to find the *de novo* motifs
- the output of a *de novo* motif analysis done through MEME/MEMERIS (Hiller et al. 2006)
- a position weight matrix (PWM) representing the results of other *de novo* motifs finder such as DRIMust (Leibovich et al. 2013), XXmotifs (Luehr et al. 2012) or cERMIT (Georgiev et al. 2010).

Tomtom is integrated to find those *de novo* motifs that look like the ones present in ATtRACT. If a multi-fasta file is submitted, other parameters must be taken into consideration.



ATtRACT A daTabase of RNA binding proteins and AssoCiated moTifs

Results for submitted sequences Download all▾

seq01
UCAUGACACAUGCCACCGUAGGUAUGUGUCGUAAUGGCGGUGAAUUUGUA ... Download▾

Show 10 entries

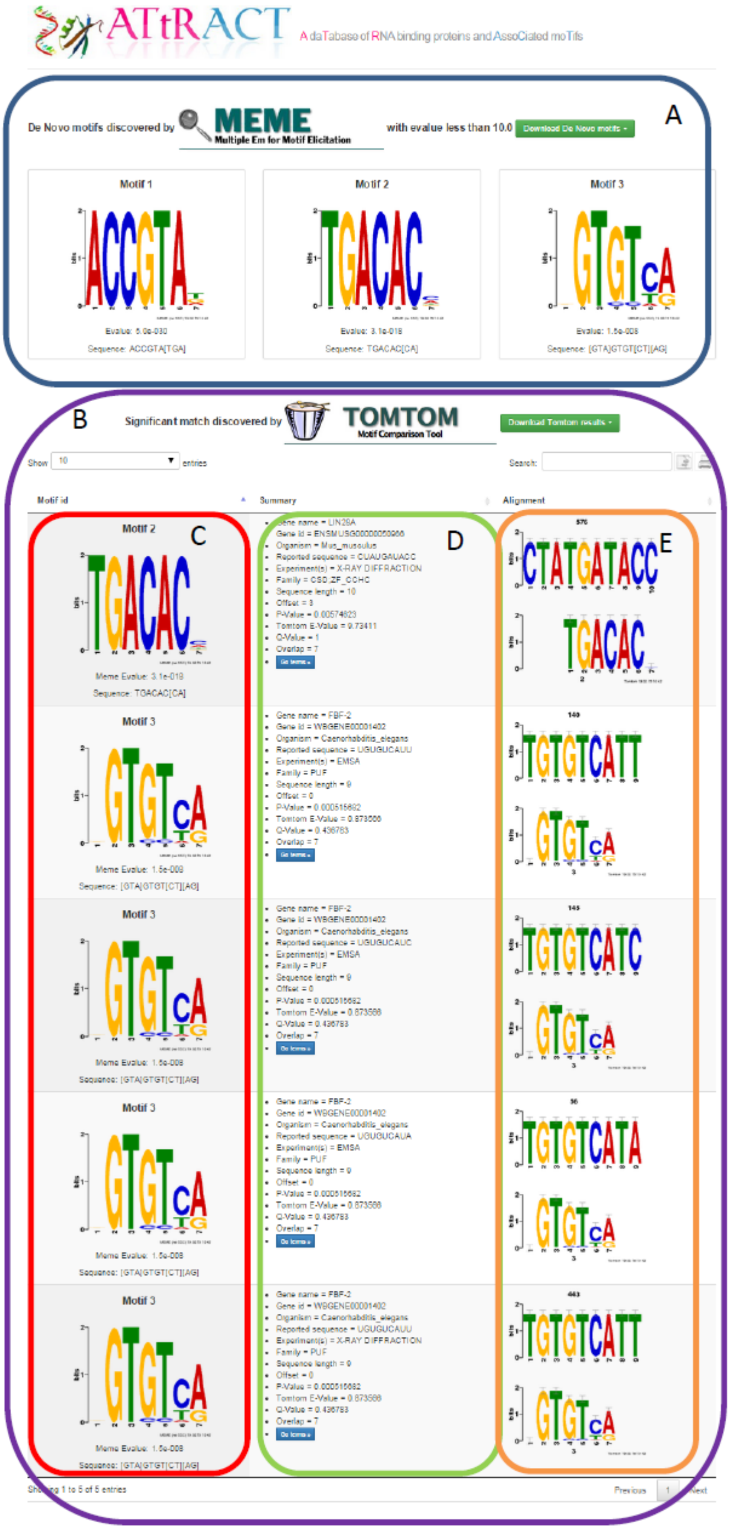
Search:

Gene Name	Gene Id	Organism	Motif	Len	Pubmed	Experiment	Domain	Off Set	GO terms	Exon250	CDS	Intror
CELF1	ENSG000001...	Homo_sapiens	UGUGU	5	16938098	SELEX ...	RRM	24	View details »	3.451	4.403	3.299
CELF2	ENSG000000...	Homo_sapiens	AUGUG	5	15657417	SELEX ...	RRM	23	View details »	4.041	4.065	3.904
CELF2	ENSG000000...	Homo_sapiens	UAUGU	5	15657417	SELEX ...	RRM	22	View details »	4.314	5.174	3.864
CELF2	ENSG000000...	Homo_sapiens	UGUGU	5	15657417	SELEX ...	RRM	24	View details »	3.451	4.403	3.299
CELF2	ENSG000000...	Homo_sapiens	AUGUGU	6	15657417	SELEX ...	RRM	23	View details »	5.881	6.67	5.558
CELF2	ENSG000000...	Homo_sapiens	GUAUGU	6	15657417	SELEX ...	RRM	21	View details »	6.419	7.508	6.204
CELF2	ENSG000000...	Homo_sapiens	UAUGUG	6	15657417	SELEX ...	RRM	22	View details »	6.199	6.675	5.849
CELF2	ENSG000000...	Homo_sapiens	UGUGU	5	11931771	UV cros...	RRM	24	View details »	3.451	4.403	3.299
DDX58	ENSG000001...	Homo_sapiens	GGCG	4	20961956	X-RAY ...	DEAD	35	View details »	3.177	2.968	4.329
ELAVL1	ENSG000000...	Homo_sapiens	AUUU	4	23519412	X-RAY ...	RRM	43	View details »	1.494	2.628	1.003

Showing 1 to 10 of 42 entries

Previous 1 2 3 4 5 Nex

Figure 23: Example of the results page displaying the sequence scan output. On the top of the page the results appear in tabular format. The bottom of the page displays the graph showing the frequency and position of motifs in the input sequence



ATtRACT
A daTabase of RNA binding proteins and AssoCiated moTifs
A
B
C
D
E
Download De Novo motifs ▾
Download Tomtom results ▾
Go terms »
576
140
145
56
443

De Novo motifs discovered by MEME Multiple Em for Motif Elicitation with evalue less than 10.0

Motif 1
Evalue: 5.0e-030
Sequence: ACCGTA[TGA]

Motif 2
Evalue: 3.1e-018
Sequence: TGACAC[CA]

Motif 3
Evalue: 1.5e-008
Sequence: [GTA]GTGT[CT][AG]

Significant match discovered by TOMTOM Motif Comparison Tool

Show 10 ▾ entries
Search:

Motif id | Summary | Alignment

Motif 2
Meme Evalue: 3.1e-018
Sequence: TGACAC[CA]

- Gene name = LIN28A
- Gene id = ENSMUSG00000050966
- Organism = Mus_musculus
- Reported sequence = CUAUGAUACC
- Experiment(s) = X-RAY DIFFRACTION
- Family = CSD;ZF_CCHC
- Sequence length = 10
- Offset = 3
- P-Value = 0.00574623
- Tomtom E-Value = 9.73411
- Q-Value = 1
- Overlap = 7

Motif 3
Meme Evalue: 1.5e-008
Sequence: [GTA]GTGT[CT][AG]

- Gene name = FBF-2
- Gene id = WBGENE00001402
- Organism = Caenorhabditis_elegans
- Reported sequence = UGUGUCAUU
- Experiment(s) = EMSA
- Family = PUF
- Sequence length = 9
- Offset = 0
- P-Value = 0.000515682
- Tomtom E-Value = 0.873588
- Q-Value = 0.436783
- Overlap = 7

Motif 3
Meme Evalue: 1.5e-008
Sequence: [GTA]GTGT[CT][AG]

- Gene name = FBF-2
- Gene id = WBGENE00001402
- Organism = Caenorhabditis_elegans
- Reported sequence = UGUGUCAUC
- Experiment(s) = EMSA
- Family = PUF
- Sequence length = 9
- Offset = 0
- P-Value = 0.000515682
- Tomtom E-Value = 0.873588
- Q-Value = 0.436783
- Overlap = 7

Motif 3
Meme Evalue: 1.5e-008
Sequence: [GTA]GTGT[CT][AG]

- Gene name = FBF-2
- Gene id = WBGENE00001402
- Organism = Caenorhabditis_elegans
- Reported sequence = UGUGUCAUA
- Experiment(s) = EMSA
- Family = PUF
- Sequence length = 9
- Offset = 0
- P-Value = 0.000515682
- Tomtom E-Value = 0.873588
- Q-Value = 0.436783
- Overlap = 7

Motif 3
Meme Evalue: 1.5e-008
Sequence: [GTA]GTGT[CT][AG]

- Gene name = FBF-2
- Gene id = WBGENE00001402
- Organism = Caenorhabditis_elegans
- Reported sequence = UGUGUCAUU
- Experiment(s) = X-RAY DIFFRACTION
- Family = PUF
- Sequence length = 9
- Offset = 0
- P-Value = 0.000515682
- Tomtom E-Value = 0.873588
- Q-Value = 0.436783
- Overlap = 7

Showing 1 to 5 of 5 entries
Previous 1 Next

Figure 24: The typical output of *de novo* motif analysis. A) *De novo* motifs discovered by MEME. B) Tomtom Output. C) Enriched motifs discovered by MEME. D) Brief summary of the characteristics of the RBPs that bind the motif. E) Alignment between the motifs available in ATtRACT and the ones discovered by meme.

The motifs distribution indicates how the occurrences of motifs are distributed along the sequences. Users can choose between three possibilities:

- one motif per sequence (oops model)
- zero or one motif per sequence (zoops model)
- any number of repetitions (anr model) (for further information visit: http://meme.nbcr.net/meme/meme-input.html).

The field named E-value represents the significance threshold for both MEME and Tomtom. Since the E-value is influenced by the dimension of the searching dataset, users can also extract a subset of the ATtRACT database in order to improve the E-Value. For example, it is possible to choose as a search space all the motifs associated to a particular organism or of a particular length, or a combination of both.

The results page (Fig 24) is divided into two sections: at the top of the page, the *de novo* motifs, discovered by MEME/MEMERIS or uploaded by the users through a PWM, are shown, then in the next section the significant matches discovered by Tomtom are displayed. The Tomtom results section is further subdivided into three columns: the *de novo* motifs are represented in the first column; a brief summary of the characteristics of the RBPs is shown in the second column; an alignment figure between the motif present in the ATtRACT database and the *de novo* motif is shown in the third column. Both the results of MEME analysis (if performed through our database) and the significant matches discovered by Tomtom are downloadable.

Analysis and regulation of alternative splicing in the pig infarcted myocardium

Analysis of isoform switching

In order to investigate the changes in alternative splicing after MI, we collected the RNAs of 4 different heart cell types – cardiomyocytes, fibroblasts, endothelial cells and macrophages (CMs, FBs, ECs and MFs) – at 3 different time points (day 0 no infarct, 3 days post MI and 7 days post MI). Transcript abundance was estimated with RSEM by aligning the reads with the pig reference transcriptome.

The aim of our work was to find the genes that showed an isoform switching, meaning transcript pairs that were negatively correlated in terms of expression in two different time points (please see materials and methods section for details). In the pig, 25322 genes are annotated. A little bit more than a half (Fig 25 green bars) were expressed more than 1 TPM in all the cell types and time points.

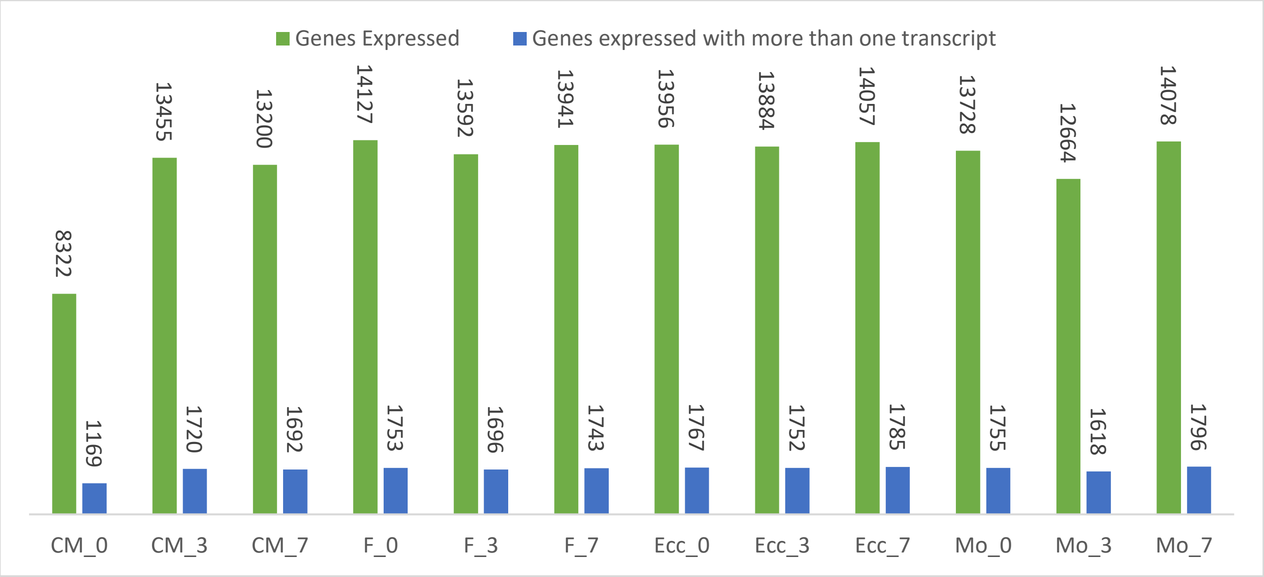


Figure 25: The green bars represent the genes per cell type and time point expressed with more than 1 TPM. The blue bars represent the genes expressed more than 1 TPM and more than one transcript. On average, 87.2% of the genes expressed only one isoform the remaining 12.8% expressed more than one transcript.

An exception was represented by the CMs at day 0 for which only one-third of the annotated genes were expressed more than 1 TPM. On average, 87.2% of the genes expressed a single isoform and more than 1 TPM in each condition, the remaining 12.8% expressed more than one transcript (Fig 25 blue bars).

Globally, we detected 210 genes undergoing an isoform switching in all the cell types and time contrasts, representing only a small part of the potential isoform switching events. Fig 26 shows the quantification of alternatively spliced genes for all the cell types and time points. Alternatively spliced genes are almost equally distributed among all the cell types, in this respect a prevalence of genes undergoing isoform switch is observed in the endothelial cells with respect to the other cell types. With respect to the time contrast, an increased number of alternatively spliced genes was observed in the day0 vs day7.

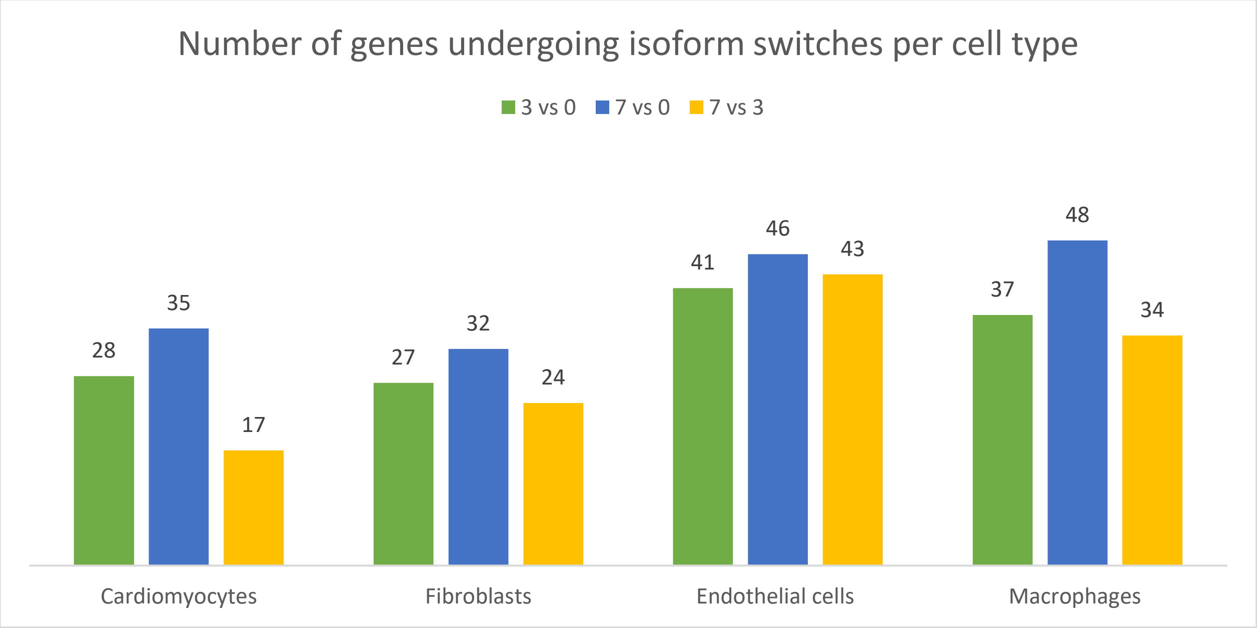


Figure 26: The number of genes undergoing isoform switching in each cell type and time contrast is shown. The numbers on the top of each bar represent the number of genes for which an isoform switch is detected. Each colour represents a different time contrast. Alternatively spliced genes are almost equally distributed among all cell types with a little prevalence of ECs with respect to the other cell types.

Isoform switches in Cardiomyocytes

Globally, 56 unique genes undergoing an isoform switch were identified in CMs in each time contrast (Table 7). For 13 of them (highlighted in red), at least one protein, expressed by the alternatively spliced gene, was detected at the proteomic level in the time contrast.

Table 7: The genes undergoing isoform switch are shown. In total, 56 distinct genes are detected, for 13 of them (highlighted in red) at least a protein is detected at the proteomic level in the time contrast.

Gene	3v0	7v0	7v3	description
ABCA6		X		ATP binding cassette subfamily A member 6
ACTN3			X	actinin, alpha 3
APOD	X			Apolipoprotein D
APOO		X		Apolipoprotein O
ASCC1		X		Sus scrofa activating signal cointegrator 1 complex subunit 1 (ASCC1).
BTK		X	X	Tyrosine-protein kinase BTK
C1orf27			X	chromosome 1 open reading frame 27
CD86	X		X	Sus scrofa CD86 molecule (CD86), mRNA.
CISD3		X		CDGSH iron sulfur domain 3
CMPK2	X			cytidine monophosphate (UMP-CMP) kinase 2, mitochondrial
COMTD1			X	catechol-O-methyltransferase domain containing 1
DIS3L			X	DIS3 like eXsome 3'-5' eXribonuclease

EXC1	X	X		eXcyst complex component 1
FCGR1A		X		High affinity immunoglobulin gamma Fc receptor I
FHOD3			X	formin homology 2 domain containing 3
GGT1		X		Sus scrofa gamma-glutamyltransferase 1 (GGT1), mRNA.
HAGHL	X	X		hydroxyacylglutathione hydrolase-like
HCFC1R1		X		host cell factor C1 regulator 1 (XPO1 dependent)
IFT52	X	X		intraflagellar transport 52
IL17D	X			interleukin 17D
KDELC1	X	X		KDEL (Lys-Asp-Glu-Leu) containing 1
KIAA1549			X	KIAA1549
LARS	X	X		leucyl-tRNA synthetase
LPIN3	X	X		lipin 3
LRPPRC	X	X		leucine rich pentatricopeptide repeat containing
LYRM7			X	LYR motif containing 7
METTL1			X	methyltransferase like 1
MTFMT	X	X		mitochondrial methionyl-tRNA formyltransferase
NDUFAF7	X	X		NADH dehydrogenase (ubiquinone) complex I, assembly factor 7
NELFE	X	X		negative elongation factor complex member E
NIPAL4			X	NIPA-like domain containing 4
NIT1	X	X		nitrilase 1
OSBPL3	X		X	oxysterol binding protein-like 3
PGF	X	X		Placenta growth factor
PLA2G6		X		phospholipase A2, group VI (cytosolic, calcium-independent)
POLR1E	X	X		polymerase (RNA) I polypeptide E, 53kDa
PTRH1		X		peptidyl-tRNA hydrolase 1 homolog
RALGAPB	X	X		Ral GTPase activating protein, beta subunit
RCCD1	X			RCC1 domain containing 1
S100A14			X	S100 calcium binding protein A14
SEMA7A	X	X		semaphorin 7A, GPI membrane anchor (John Milton Hagen blXd group)
STAT5A	X	X		Signal transducer and activator of transcription 5A
TBC1D31			X	TBC1 domain family, member 31
TEX264		X		testis expressed 264
TM4SF18	X	X		transmembrane 4 L six family member 18
TN-X		X	X	Sus scrofa tenascin-X (TN-X), mRNA.
TNFRSF17	X			tumor necrosis factor receptor superfamily, member 17
TNRC6C		X		trinucleotide repeat containing 6C
TRIM44	X	X		tripartite motif containing 44
TRIM52			X	tripartite motif containing 52
UBXN2A	X			UBX domain protein 2A
UXT		X		Protein UXT
ZAP70	X	X		zeta-chain (TCR) associated protein kinase 70kDa
ZNF212		X		zinc finger protein 212
ZNF316	X	X		zinc finger protein 316

For each time contrast, we performed an ORA using the human annotation. All the pig genes orthologues one to one to the human were retrieved and used as testing set. Few GO terms were overrepresented in CMs (Table 8, Table 9, Table 10) reflecting the fact that alternatively spliced genes do not participate in a global process but take part in distinct processes.

Table 8: Enriched GO terms for CM at day 3 vs day 0.

CM3v0	domain	description	genes involved
GO:0045086	BP	positive regulation of interleukin-2 biosynthetic process	CD86, STAT5A

Table 9: Enriched GO terms for CM at day 7 vs day 0

CM7v0	domain	description	genes involved
GO:0047497	BP	mitochondrion transport along microtubule	LRPPRC, UXT
GO:0045579	BP	positive regulation of B cell differentiation	STAT5A, BTK

Table 10: Enriched GO terms for CM at day 7 vs day 3

CM7v3	domain	description	genes involved
GO:0042113	BP	B cell activation	BTK, CD86
GO:0008176	MFu	tRNA (guanine-N7-)-methyltransferase activity	METTL1
GO:0043527	CC	tRNA methyltransferase complex	METTL1

Isoform switches in Fibroblasts

In FBs, 53 distinct genes undergoing isoform switching were identified in each time contrast (Table 11). For 16 of them (highlighted in red) at least one protein, expressed by the alternatively spliced gene, is detected at the proteomic level in the time contrast.

Table 11: The genes undergoing isoform switching in FBs are shown. In total, 56 distinct genes were detected, for 16 of them (highlighted in red) at least one protein is detected in the time point.

Gene	3v0	7v0	7v3	description
ACVR2B			X	Sus scrofa activin A receptor, type IIB (ACVR2B), mRNA.
ADAMTS12		X		ADAM metallopeptidase with thrombospondin type 1 motif, 12
AKIP1		X	X	A-kinase interacting protein 1
ANKRD6		X		ankyrin repeat domain 6
ATAT1	X			Sus scrofa alpha tubulin acetyltransferase 1 (ATAT1), mRNA.
BEST1	X	X		bestrophin 1
C7H6orf12		X	X	Uncharacterized protein
CCNI2	X		X	cyclin I family, member 2
CD46	X	X		CD46 molecule, complement regulatory protein
CD83		X		CD83 molecule
CHM	X		X	choroideremia (Rab escort protein 1)
CISD3		X		CDGSH iron sulfur domain 3
CTLA4			X	Sus scrofa cytotoxic T-lymphocyte-associated protein 4 (CTLA4), mRNA.
CU463271		X		Not Available
DMXL1	X	X		Dmx-like 1
ELP6		X	X	elongator acetyltransferase complex subunit 6
EVI5		X	X	ecotropic viral integration site 5
FANCE	X	X		Fanconi anemia group E protein

FASTKD1	X	X		FAST kinase domains 1
FBX6	X	X		F-box protein 6
GGT1	X	X		Sus scrofa gamma-glutamyltransferase 1 (GGT1), mRNA.
GK	X			glycerol kinase
GPR34	X			G protein-coupled receptor 34
HXK2		X	X	hXk microtubule-tethering protein 2
HSD17B12		X		hydroxysteroid 17-beta dehydrogenase 12
IGF2BP2	X			insulin-like growth factor 2 mRNA binding protein 2
JCHAIN	X		X	joining chain of multimeric IgA and IgM
KIAA1549	X	X		KIAA1549
MAST4	X	X		microtubule associated serine/threonine kinase family member 4
METTL1	X	X		methyltransferase like 1
MTFMT		X	X	mitochondrial methionyl-tRNA formyltransferase
NDUFAF7	X	X		NADH dehydrogenase (ubiquinone) complex I, assembly factor 7
NNAT	X			Sus scrofa neuronatin (NNAT), mRNA.
OSBPL3			X	oxysterol binding protein-like 3
PDZD2	X			PDZ domain containing 2
PNPLA4	X		X	patatin-like phospholipase domain containing 4
RECQL4			X	RecQ protein-like 4
SCML2	X			sex comb on midleg-like 2 (Drosophila)
SLC25A14		X		Brain mitochondrial carrier protein 1
SORCS1		X		sortilin-related VPS10 domain containing receptor 1
STXBP4			X	syntaxin binding protein 4
TBC1D31	X		X	TBC1 domain family, member 31
TFRC	X	X		Sus scrofa transferrin receptor (p90, CD71) (TFRC), mRNA.
THSD7A	X		X	thrombospondin type 1 domain containing 7A
TSHR		X	X	Sus scrofa thyroid stimulating hormone receptor (TSHR), mRNA.
TTC21A		X	X	tetratricopeptide repeat domain 21A
UMPS		X	X	uridine monophosphate synthetase
UPF3B		X		UPF3 regulator of nonsense transcripts homolog B (yeast)
VWF	X		X	Sus scrofa von Willebrand factor (VWF), mRNA.
ZBTB3		X	X	zinc finger and BTB domain containing 3
ZFYVE28	X			zinc finger, FYVE domain containing 28
ZNF77		X	X	zinc finger protein 77
ZNFX1			X	NFX1-type zinc finger-containing protein 1

Interestingly, 11 genes (Table 12) showed a non-significant difference (<10TPM) at the gene expression level between two time points showing that alternative splicing provides additional genetic complexity that is not reflected by total gene expression levels.

Table 12: Genes showing a difference in expression less than 10 TPM between the selected time points

Time contrast	Gene names
FB 7v3	CCNI2, C7H6orf12, MTFMT, PNPLA4, ACVR2B
FB 7v0	CU463271.1, BEST1, CD46, FBXO6, FANCE, NDUFAF7
FB 3v0	FANCE

Also in the case of FBs, few overrepresented GO terms were found (Table 13 and Table 14).

Table 13: Enriched GO terms for FBs at day 3 vs day 0

FB3v0	Domain	Description	Genes involved
GO:0001948	MFu	glycoprotein binding	TFRC, VWF, FBXO6*

Table 14: Enriched GO terms for FBs at day 7 vs day 3

FB7v3	domain	description	genes involved
GO:0071756	CC	pentameric IgM immunoglobulin complex	JCHAIN
GO:0071750	CC	dimeric IgA immunoglobulin complex	JCHAIN

Isoform switches in Endothelial cells

In ECs, 86 distinct genes undergoing isoform switching were identified in each time contrast (Table 15). For 16 of them (highlighted in red) at least one protein, expressed by the alternatively spliced gene, is detected at the proteomic level in the time contrast.

Table 15: The genes undergoing isoform switching in ECs are shown. In total, 86 distinct genes are detected, for 16 of them (highlighted in red) at least one of transcript is detected at the proteomic level in the time contrast.

Gene	**3v0**	**7v0**	**7v3**	**description**
ABHD12B			X	abhydrolase domain containing 12B
ADAMTS12	X		X	ADAM metallopeptidase with thrombospondin type 1 motif, 12
AKAP17A	X			A-kinase anchor protein 17A
ANKRD6		X		ankyrin repeat domain 6
ANKS6		X	X	ankyrin repeat and sterile alpha motif domain containing 6
ATAT1	X	X		Sus scrofa alpha tubulin acetyltransferase 1 (ATAT1), mRNA.
ATG5			X	autophagy protein 5
BTK	X		X	Tyrosine-protein kinase BTK
C1orf27		X		chromosome 1 open reading frame 27
C7H6orf12		X		Uncharacterized protein
CARMIL2		X		Capping protein, Arp2/3 and myosin-I linker protein 2
CC2D1B			X	coiled-coil and C2 domain containing 1B
CDH2	X		X	cadherin 2, type 1, N-cadherin (neuronal)
CEACAM16	X			carcinoembryonic antigen-related cell adhesion molecule 16
CH242-204P3.6	X			Not available
CH242-486P11.2		X		Uncharacterized protein
CLEC5A	X			C-type lectin domain family 5 member A
COQ10A		X		coenzyme Q10A
CSF3	X		X	Sus scrofa colony stimulating factor 3 (granulocyte) (CSF3), mRNA.
DCAF17	X	X		DDB1 and CUL4 associated factor 17
DCLRE1A		X	X	DNA cross-link repair 1A
DHX29		X		DEAH (Asp-Glu-Ala-His) box polypeptide 29
DMXL1	X	X		Dmx-like 1
DNTT		X		DNA nucleotidyleXtransferase

DPPA2		X		developmental pluripotency associated 2
EGFLAM			X	EGF like, fibronectin type III and laminin G domains
ENPP3	X	X		Ectonucleotide pyrophosphatase/phosphodiesterase family member 3
FASTKD1	X	X		FAST kinase domains 1
FCER1A	X	X		Fc fragment of IgE, high affinity I, receptor for; alpha polypeptide
FST			X	Sus scrofa follistatin (FST), mRNA.
GK	X			glycerol kinase
GNB5	X		X	guanine nucleotide binding protein (G protein), beta 5
GPCPD1	X	X		glycerophosphocholine phosphodiesterase 1
HAGHL		X		hydroxyacylglutathione hydrolase-like
HDAC10	X		X	histone deacetylase 10
HP		X		Sus scrofa haptoglobin (HP), mRNA.
ICE2	X		X	interactor of little elongation complex ELL subunit 2
IFT81	X			intraflagellar transport 81
KANSL1L			X	KAT8 regulatory NSL complex subunit 1-like
KIAA1549	X			KIAA1549
LINS		X		lines homolog (Drosophila)
LMLN		X	X	leishmanolysin-like (metallopeptidase M8 family)
LONRF3	X			LON peptidase N-terminal domain and ring finger 3
LRRC36			X	leucine rich repeat containing 36
LYRM7	X	X		LYR motif containing 7
MAGIX		X		MAGI family member, X-linked
METTL1		X		methyltransferase like 1
MME	X	X		Neprilysin
MST1R	X		X	macrophage stimulating 1 receptor
MTCP1			X	mature T-cell proliferation 1
MTFMT	X		X	mitochondrial methionyl-tRNA formyltransferase
MYH7		X	X	Sus scrofa myosin, heavy chain 7, cardiac muscle, beta (MYH7), mRNA.
NCAM1	X	X		neural cell adhesion molecule 1
NCOA1	X			Sus scrofa nuclear receptor coactivator 1 (NCOA1), mRNA.
NNAT		X	X	Sus scrofa neuronatin (NNAT), mRNA.
NOA1		X	X	nitric oxide associated 1
NPHP1		X	X	nephronophthisis 1 (juvenile)
NR6A1	X			Nuclear receptor subfamily 6 group A member 1
OMA1	X		X	OMA1 zinc metallopeptidase
OSBPL3	X			oxysterol binding protein-like 3
OVCA2	X		X	ovarian tumor suppressor candidate 2
PDZD11		X		Sus scrofa PDZ domain containing 11 (PDZD11), mRNA.
POFUT1		X	X	protein O-fucosyltransferase 1
POLR1E	X		X	polymerase (RNA) I polypeptide E, 53kDa
PTGS2	X			Prostaglandin G/H synthase 2
RCCD1	X		X	RCC1 domain containing 1
REPS2		X		RALBP1 associated Eps domain containing 2
RHBDL1		X	X	rhomboid, veinlet-like 1 (Drosophila)
S100A9		X	X	Sus scrofa S100 calcium binding protein A9 (S100A9), mRNA.
S100PBP	X		X	S100P binding protein
SGCA		X		Sus scrofa sarcoglycan, alpha (50kDa dystrophin-associated glycoprotein)

SLA-DRB5			X	Not available
SLBP*	X			stem-lXp binding protein
SLC25A14	X			Brain mitochondrial carrier protein 1
SLC35B1		X	X	solute carrier family 35 member B1
SMPDL3A		X		Acid sphingomyelinase-like phosphodiesterase 3a
SNX21		X	X	sorting nexin family member 21
TCAIM	X		X	T cell activation inhibitor, mitochondrial
THSD7A		X	X	thrombospondin type 1 domain containing 7A
TN-X		X	X	Sus scrofa tenascin-X (TN-X), mRNA.
UMPS	X		X	uridine monophosphate synthetase
ZBTB3		X	X	zinc finger and BTB domain containing 3
ZCCHC18	X	X		zinc finger, CCHC domain containing 18
ZFYVE16		X	X	zinc finger, FYVE domain containing 16
ZMIZ1		X	X	zinc finger, MIZ-type containing 1
ZNF212			X	zinc finger protein 212

Also in ECs, eight genes (Table 16) showed a non-significant difference (<10TPM) at the gene expression level between two time points.

Table 16: Genes showing a difference in expression less than 10 TPM between the selected time points

Time contrast	Gene name
EC 7v3	OMA1, ATG5, CC2D1B, LMLN, KANSL1L
EC7v0	PDZD11
EC3v0	ENPP3, CLEC5A

Few enriched GO terms were found through functional enrichment analysis (Table 17 and Table 18).

Table 17: Enriched GO terms for ECs at day 3 vs day 0

EC3v0	Domain	Description	Genes involved
GO:0019899	MFu	enzyme binding	CDH2, HDAC10, NCOA1, CSF3, NR6A1, MST1R

Table 18: Enriched GO terms for ECs at day 7 vs day 0

EC7v0	Domain	Description	Genes involved
GO:0005911	CC	cell-cell junction	CARMIL2, PDZD11, NPHP1, NCAM1, SGCA

Isoform switches in Macrophages

In MFs, 79 distinct genes undergoing isoform switching were identified in each time contrast (Table 19). For 19 of them (highlighted in red) at least one protein, expressed by the alternatively spliced gene, is detected at the proteomic level in the time contrast.

Table 19: The genes undergoing isoform switching in MFs are shown. In total, 79 distinct genes are detected, for 16 of them (highlighted in red) at least one protein expressed by the gene is detected at the proteomic level in the time contrast.

Gene	3v0	7v0	7v3	description
ADHFE1	X	X		alcohol dehydrogenase, iron containing, 1
AGER		X	X	Sus scrofa advanced glycosylation end product-specific receptor (AGER)

ANO9		X		anoctamin 9
AP4E1	X		X	adaptor-related protein complex 4, epsilon 1 subunit
APOM	X			Apolipoprotein M
ARG2	X			arginase 2
BEST1		X		bestrophin 1
BF		X		Sus scrofa complement factor B (CFB), mRNA.
BMPER		X		BMP binding endothelial regulator
C11orf52			X	chromosome 11 open reading frame 52
C15orf41		X		chromosome 15 open reading frame 41
C15orf52	X		X	chromosome 15 open reading frame 52
CD3D	X		X	Sus scrofa CD3d molecule, delta (CD3-TCR complex) (CD3D), mRNA.
COMTD1	X	X		catechol-O-methyltransferase domain containing 1
CU463271			X	Not available
CXCR2	X			chemokine (C-X-C motif) receptor 2
DCLRE1A		X		DNA cross-link repair 1A
DHX29		X	X	DEAH (Asp-Glu-Ala-His) box polypeptide 29
DZIP1L		X		DAZ interacting zinc finger protein 1-like
ENSSSCG00000001956	X		X	Not available
FAM122C		X		family with sequence similarity 122C
FANCE		X		Fanconi anemia group E protein
FBX6		X		F-box protein 6
FCER1A	X		X	Fc fragment of IgE, high affinity I, receptor for; alpha polypeptide
FCRLB	X		X	Fc receptor-like B
FOXM1		X	X	forkhead box M1
FOXP3		X		forkhead box P3
GNB5	X		X	guanine nucleotide binding protein (G protein), beta 5
HSD17B12	X	X		hydroxysteroid 17-beta dehydrogenase 12
ICE2	X		X	interactor of little elongation complex ELL subunit 2
IL17D			X	interleukin 17D
IL17RE			X	interleukin 17 receptor E
KANSL1L	X	X		KAT8 regulatory NSL complex subunit 1-like
KCP	X			kielin/chordin-like protein
KIAA2018		X	X	KIAA2018
KLRK1	X		X	Sus scrofa killer cell lectin-like receptor subfamily K, member 1 (KLRK1)
LAMA3		X		laminin, alpha 3
LAYN	X	X		layilin
LINS		X		lines homolog (Drosophila)
LRPPRC	X		X	leucine rich pentatricopeptide repeat containing
LYRM7		X	X	LYR motif containing 7
MBTPS2	X	X		membrane-bound transcription factor peptidase, site 2
METTL1	X	X		methyltransferase like 1
MIS18BP1	X			MIS18 binding protein
MTFMT	X	X		mitochondrial methionyl-tRNA formyltransferase
MUC1		X		mucin 1, cell surface associated
MYNN	X			myoneurin
MYOM1		X		myomesin 1

NEMF	X	X		nuclear export mediator factor
OPHN1		X		Oligophrenin-1
PHKA1			X	phosphorylase kinase, alpha 1 (muscle)
PIGA			X	phosphatidylinositol glycan anchor biosynthesis, class A
POLR1E			X	polymerase (RNA) I polypeptide E, 53kDa
PRICKLE3		X	X	prickle homolog 3
RETN	X			Sus scrofa resistin (RETN), mRNA.
RHBDL1		X	X	rhomboid, veinlet-like 1 (Drosophila)
RNASEL			X	2-5A-dependent ribonuclease
S100A14	X	X		S100 calcium binding protein A14
S100A9		X		Sus scrofa S100 calcium binding protein A9 (S100A9), mRNA.
S100PBP		X	X	S100P binding protein
SLA-DRB5		X	X	Not available
SLC25A14	X		X	Brain mitochondrial carrier protein 1
SLC35B4		X	X	solute carrier family 35 member B4
SMARCA1		X		SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 1
SMYD4	X		X	SET and MYND domain containing 4
TCAIM	X		X	T cell activation inhibitor, mitochondrial
TEX264	X	X		testis expressed 264
TMEM27		X	X	Collectrin
TNFRSF17	X			tumor necrosis factor receptor superfamily, member 17
TREML-2		X		Trem-like transcript 2 protein
TSPAN6	X		X	Sus scrofa tetraspanin 6 (TSPAN6), mRNA.
TTC21A		X		tetratricopeptide repeat domain 21A
TTLL7		X		tubulin tyrosine ligase-like family member 7
UMPS		X	X	uridine monophosphate synthetase
UPF3B	X	X		UPF3 regulator of nonsense transcripts homolog B (yeast)
VTN	X	X		Vitronectin
ZBTB3	X			zinc finger and BTB domain containing 3
ZIP4	X	X		Zinc transporter ZIP4
ZNF641		X		zinc finger protein 641

In MFs, 8 genes (Table 20) showed a non-significant difference (<10TPM) at the gene expression level between two time points.

Table 20: Genes showing a difference in expression less than 10 TPM between the selected time points

Time contrast	Gene name
MF 7v3	TMEM27
MF 7v0	ZIP4, S100A14, BEST1, FAM122C, C15orf41, LAYN, TMEM27
MF 3v0	ZIP4, RETN

No enriched GO terms were found in MFs.

Global analysis of alternatively spliced genes

To better understand the effect of MI on alternative splicing, we performed further statistical and structural analyses. Interestingly, we found that a small group of eight genes were alternatively spliced in three cell types without any regards for the time contrasts (Table 21).

Table 21: The table shows the genes alternative spliced in four cell types.

Gene name	Cell type
UMPS	FB, EC, MF
ZBTB3	FB, EC, MF
POLR1E	CM, EC, MF
SLC25A14	FB, EC, MF
OSBPL3	CM, FB, EC
KIAA1549	CM, FB, EC
LYRM7	CM, EC, MF

The information for the role and localisation of these genes is scarce in general. The KIAA1549 gene is localised in the membrane while LYRM7 in the mitochondrion. POLR1E gene and ZBTB3 gene are two transcriptional regulators. The UMPS gene encodes for an enzyme that is involved in the *de novo* uridine 5'-monophosphate biosynthesis. The OSBPL3 gene binds to cholesterol and plays a role in the regulation of cell adhesion and organisation of the actin cytoskeleton (Lehto et al. 2008). The SLC25A14 gene is localised in the mitochondria and is involved in the aerobic respiration.

Subsequently, we investigated the biotype classification of the alternatively spliced transcripts. Based on the transcript biotype annotation available in Ensembl, about 90.2% of transcript isoform pairs are potentially translated into proteins (Fig 27). Among the non-coding transcripts, the processed transcripts, characterized by the absence of an open reading frame, are predominant in EC 3v0, EC 7v0 and FB 7v0 MF 7v3. Nonsense mediated decay transcripts are prevalent mainly in EC 3v0, MF 7v3, MF 3v0, retained intron transcripts are present mainly in EC 7v3, EC 7v0, and FB 3v0.

We next investigated the overlap of alternatively spliced genes among the different time contrasts. CMs show the biggest overlap (20) between the time contrasts 3v0 and 7v0, while only 2 genes were in common between time contrast 3v0 and 7v3 and time contrasts 7v0 and 7v3 (Fig 28a). In other cell types, the overlapped genes were more constant.

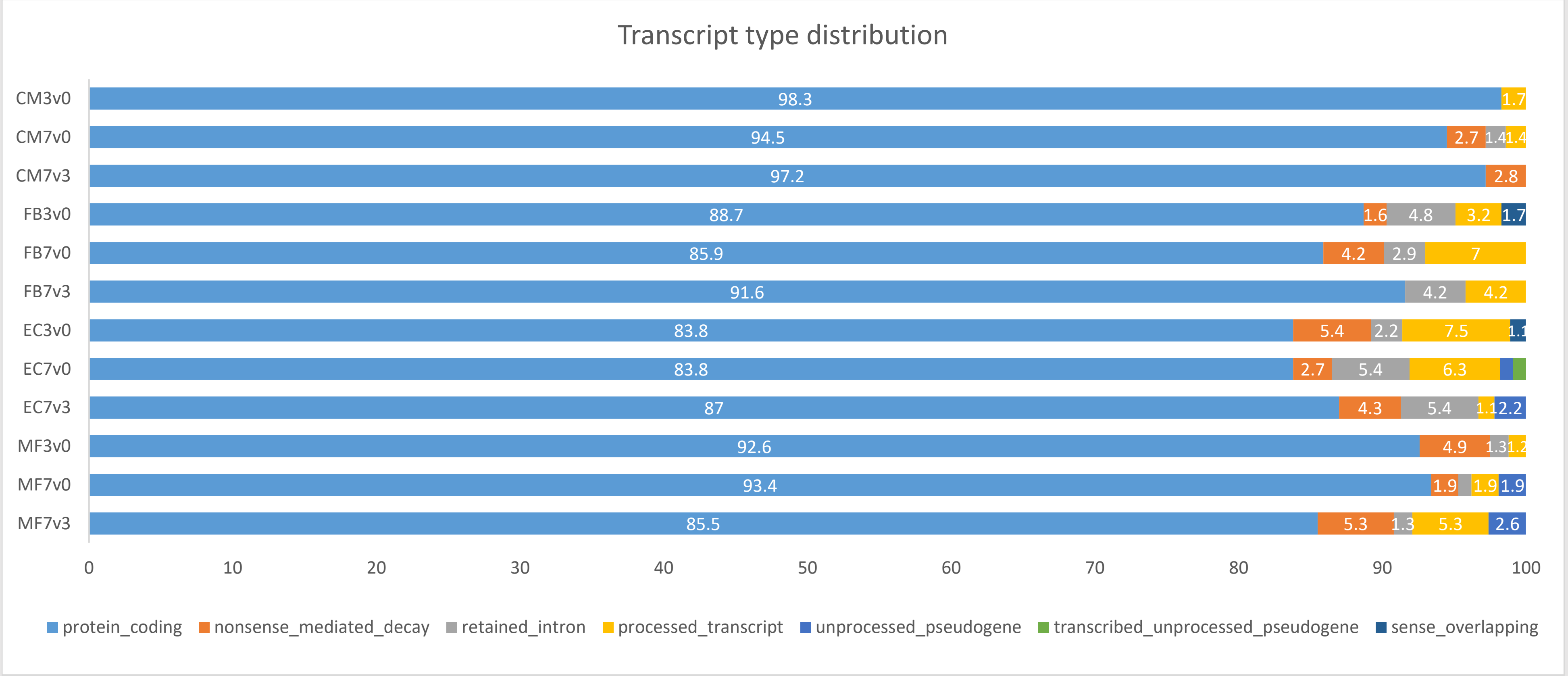


Figure 27: Isoform transcript pairs distributions. Each transcript is classified according to the Ensembl transcript biotype. Most of the transcripts are protein coding genes (90.2% on average). Processed transcript, nonsense mediated decay and retained intron are also present in a significant percentage

In FBs, 12,7, and, 11 genes are shared between time contrasts 3vs0 and 7v0, 3v0 and 7v3, and 7v0 and 7v3 respectively (Fig 28b). In ECs,11,16, and 17 genes are shared between time contrasts 3vs0 and 7v0, 3v0 and 7v3, and 7v0 and 7v3 respectively (Fig 28c). In MFs, 14, 14 and 12 genes are shared between time contrasts 3v0 and 7v0, 3v0 and 7v3, and 7v0 and 7v3 respectively (Fig 28d). These results suggest many of the AS changes observed in the different cell types were maintained over time.

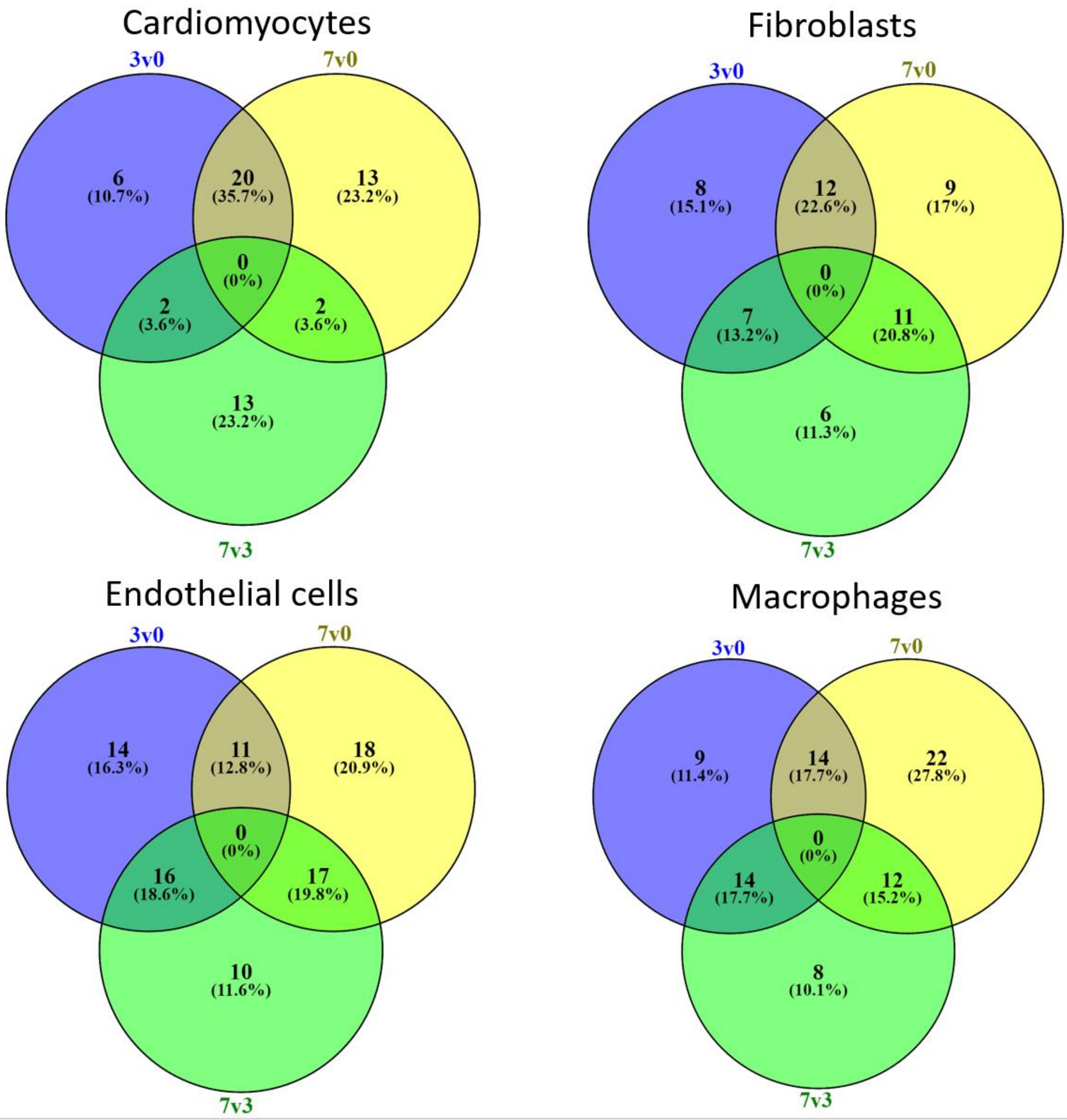


Figure 28: The Venn diagram shows the number of alternative spliced genes across different time points. CMs shared the biggest overlap (20 genes) among time point 3v0 and 7v0 (A). In other cell types, the genes overlap among time points is more stable (B, C,D).

According to Kriventseva et al. 2003, alternative splicing tends to insert or delete or alter protein domains more frequently than expected by chance. To investigate the impact of alternative splicing at the protein structural level, we investigated the consequences that isoform switching potentially had on the gain or loss on the proteins' domains. To characterize the domains annotated in each transcript isoform pairs, we employed the information available in the PFAM database (Version 30.0). Given a transcript switching pair, we defined a domain as inserted if the transcript is upregulated, deleted if downregulated. Notice that, in Table 22, the domains that are in common between the transcript pairs are not shown even though they can be modified by alternative splicing events. Interestingly, alternative splicing of the APOD gene potentially changed the domain by inserting a Lipocalin-like domain and by deleting the Triabin domain. Moreover, in the TNFRSF17 gene, the BCMA, TALL-1 binding domain is inserted in CM3v0 but deleted in MF3v0.

Table 22: Domains altered by alternative splicing events. The columns represent the gene name, the domain inserted or disrupted due to an alternative splicing event and the cell type and time contrast where the domain is being altered.

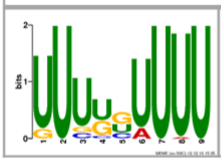
gene name	Inserted domain	Deleted domain	Where
TNFRSF17	BCMA, TALL-1 binding		CM3v0
APOD	Lipocalin-like domain	Triabin	CM3v0
HAGHL	Metallo-beta-lactamase superfamily		CM3v0, CM7v0, EC7v0
NELFE	RNA recognition motif		CM3v0, CM7v0
ZAP70		Protein tyrosine kinase Protein kinase domain	CM7v0
BTK		PH domain Variant SH3 domain Protein kinase domain Protein tyrosine kinase BTK motif SH3 domain	CM7v0, CM7v3, EC7v3
IGF2BP2	RNA recognition motif		FB3v0
TFRC	PA domain Peptidase family M28 Transferrin receptor-like dimerisation domain		FB3v0
ZNF77		Zinc-finger double-stranded RNA-binding	FB3v0, FB7v0
CTLA4	Immunoglobulin V-set domain		FB7v3
NCOA1		Helix-loop-helix DNA-binding domain	EC3v0
LONRF3		Zinc finger, C3HC4 type	EC3v0
NR6A1	Zinc finger, C4 type		EC3v0
ENPP3	Type I phosphodiesterase / nucleotide pyrophosphatase		EC3v0, EC7v0
CDH2	Cadherin cytoplasmic region		EC3v0, EC7v3

ATAT1		GNAT acetyltransferase	EC3v0, EC7v0
NPHP1		SH3 domain Variant SH3 domain	EC7v0, EC7v3
CDH2		Cadherin cytoplasmic region	EC7v3
TNFRSF17		BCMA, TALL-1 binding	MF3v0
ARG2	Arginase family		MF3v0
PRICKLE3	LIM domain PET Domain		MF7v0, MF7v3
MUC1		SEA domain	MF7v0
ANO9		Calcium-activated chloride channel	MF7v0
AGER	Immunoglobulin domain		MF7v0, MF7v3
ADHFE1		Iron-containing alcohol dehydrogenase	MF7v0
SMARCA1		SNF2 family N-terminal domain DEAD/DEAH box helicase Helicase conserved C-terminal domain HAND SLIDE Type III restriction enzyme, res subunit	MF7v0

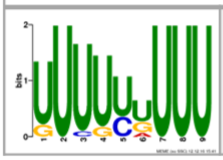
Regulation of Alternative Splicing

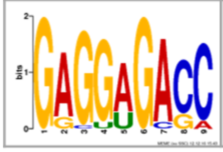
Regulation of alternative splicing is a complicated process in which regulatory motifs are recognized by a splicing factor, a subclass of RBPs, to activate or repress the inclusion of adjacent exons. Our aim was to characterize the regulatory motifs and the splicing factors that regulated alternative splicing during MI. For this reason, we designed a pipeline composed of the combination of MEME and Tomtom (see material and methods for details). To validate our findings, we employed the proteomic profile to assess whether: (i) at least one transcript of the pairs was translated into a protein, and (ii) the splicing factor was detected at the proteomic level. Fig 29 and Fig 30 show the motifs and the splicing factors that potentially regulate alternative splicing in all the cell types. The first column shows the regulatory motifs, the second column the gene(s) alternatively spliced and the third column the splicing factors, the number in parenthesis indicates the splicing factor expression level. A negative number means that the splicing factor is downregulated, and vice versa.

Cardiomyocytes 3v0		

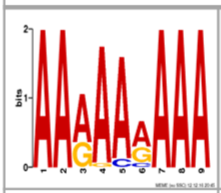
	APOD
	KDELC1
	LARS
	NDUFAF7
		HNRNPD(1.9)
		HNRNPU(1.9)

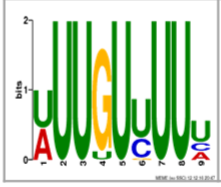
Cardiomyocytes 7v0		

	KDELC1
	NDUFAF7
		HNRNPD(0.4)
		HNRNPU(0.7)

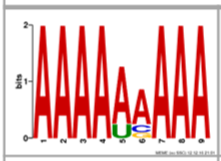
Cardiomyocytes 7v3		
	FHOD3	HNRNPU(-1.5)

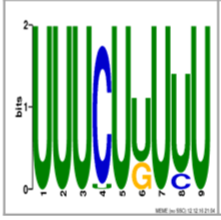
Fibroblasts 3v0		

	CHM
	PDZD2
		HNRNPD(0.4)
		HNRNPM(-0.5)
		HNRNPU(-0.7)
		SRSF6(-1.1)
		SRSF7(0.8)

	MAST4
	NDUFAF7
		CELF1(0.5)
		CELF2(-0.5)
		HNRNPC(-0.8)
		HNRNPD(0.4)
		HNRNPU(-0.7)

Fibroblasts 7v0		

	CISD3
	EVI5
	UMPS
	UPF3B
		HNRNPD(0.2)
		HNRNPU(-1)
		SRSF6(-0.6)
		SRSF7(0.4)

	HSD17B12
	MAST4
		CELF1(0.3)
		CELF2(0.1)
		HNRNPC(-0.7)
		HNRNPD(0.2)
		HNRNPU(-1)
		PTBP2(2)

Fibroblasts 7v3		

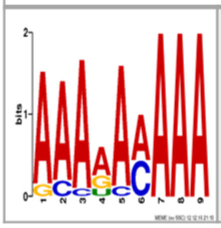
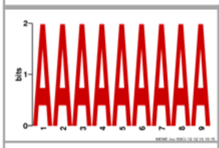
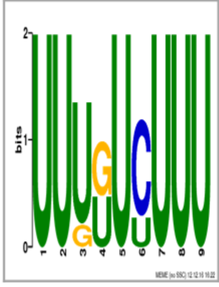
	CHM
	EVI5
	JCHAIN
	TBC1D31
	UMPS
	ZNFX1
		HNRNPD(-0.4)
		HNRNPLL(0.1)
		HNRNPU(-0.4)
		SRSF7(-0.7)

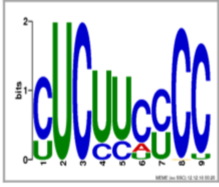
Figure 29: Motifs and splicing factors that potentially regulate alternatively spliced genes in CMs and FBs. The first column shows the motifs, the second row the gene(s) alternatively spliced and the third the splicing factors, the number in parenthesis indicates the protein expression level.

Endothelial cells 3v0		

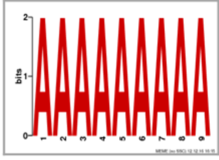
	UMPS	HNRNPD(0.2)
		HNRNPU(-1.4)
		PCBP2(-1.2)

	GPCPD1	CELF1(0.3)
		CELF2(1.5)
		HNRNPC(0.7)
		HNRNPD(0.2)
		HNRNPU(-1.4)
		PCBP2(-1.2)
		PTBP1(-0.3)
		PTBP2(0.8)

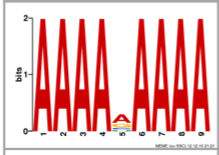
Endothelial cells 7v0		

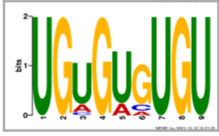
	HP
	POFUT1
	S100A9
	SMPDL3A
	ZMIZ1
		HNRNPU(-0.3)
		PCBP2(-1.2)
		PTBP1(-0.3)
		PTBP2(3.5)
		SRSF6(1.6)

Endothelial cells 7v3		

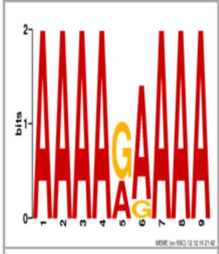
	LRRC36
	UMPS
	ZMIZ1
		HNRNPD(1.3)
		HNRNPU(1.7)
		PCBP2(0.5)

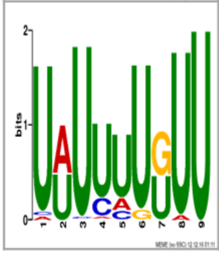
Macrophages 3v0		

	HSD17B12
	TEX264
	UPF3B
		HNRNPD(2.4)
		HNRNPK(1.4)
		HNRNPU(1.8)
		SRSF6(1.1)

	APOM
	COMTD1
	TEX264
		CELF1(-0.1)
		CELF2(0.7)

Macrophages 7v0		

	DHX29
	HSD17B12
	MYOM1
	SMARCA1
	TEX264
	UMPS
	UPF3B
		HNRNPD(2.9)
		HNRNPK(1.6)
		HNRNPU(2.5)
		SRSF6(1.3)

	LAMA3
	NEMF
	METTL1
		CELF1(-0.5)
		CELF2(0.5)
		HNRNPC(0.4)
		HNRNPD(2.9)
		HNRNPK(1.6)
		HNRNPU(2.5)
		PTBP1(-0.2)

Macrophages 7v3		

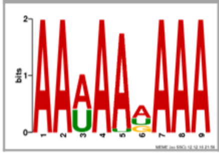
	DHX29
	UMPS
		HNRNPD(0.5)
		HNRNPK(0.3)
		HNRNPU(1.4)
		SRSF6(0.4)

Figure 30: Motifs and splicing factors that potentially regulate alternatively spliced genes in ECs and MFs. The first column shows the motifs, the second row the gene(s) alternatively spliced and the third the splicing factors, the number in parenthesis indicates the protein expression level.

Comparison between MAGNETO and the standard enrichment analysis method

Once we investigated the role of AS in the pig infarcted myocardium, we wanted to assess whether it was possible to gain more insight into the ORA. For this purpose, we developed a new method called MAGNETO, which makes use of information obtained from PINs to improve the ORA results (see materials and methods). To compare in a systematic and unbiased way the effectiveness of MAGNETO with respect to the ORA we employed the approach proposed by Tarca et al. 2012.

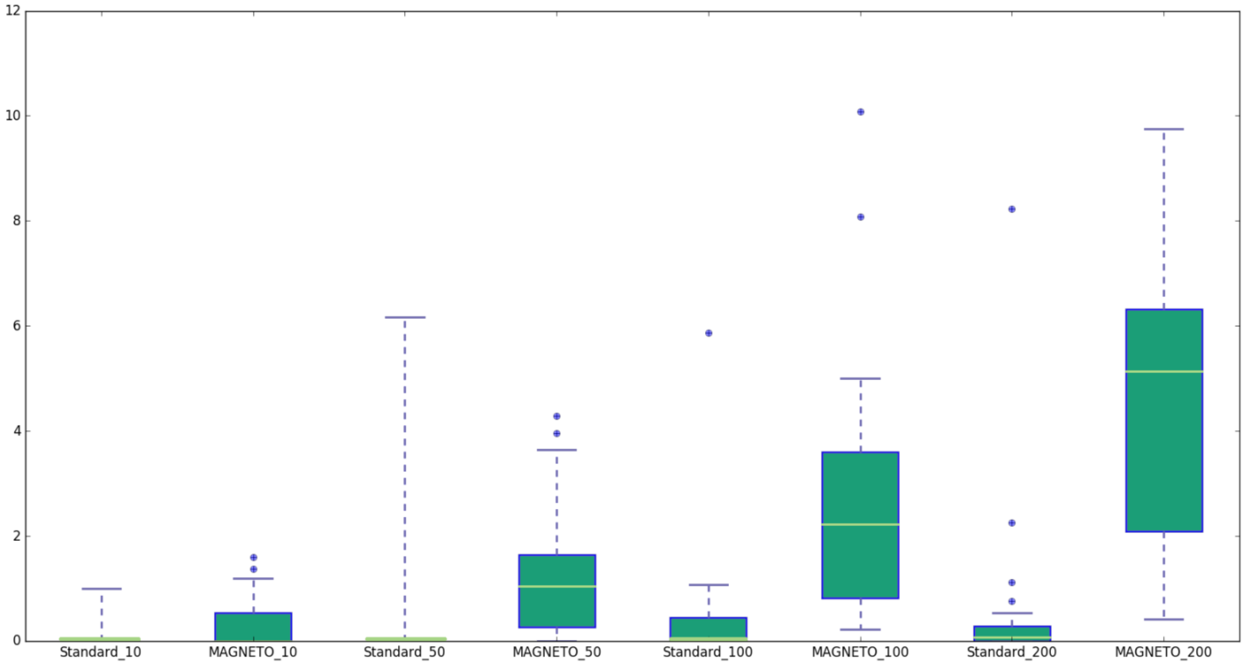


Figure 31: The box plots represent the -Log(p-value) distribution of the 21 datasets. Both MAGNETO and the standard ORA method are shown. The label on the x-axis represents the method (standard and MAGNETO) the number next to the label represents the number of proteins submitted.

We selected 21 datasets of differentially expressed genes from GEO database for which we can assess an association between the phenotype under investigation and a target pathway in KEGG.

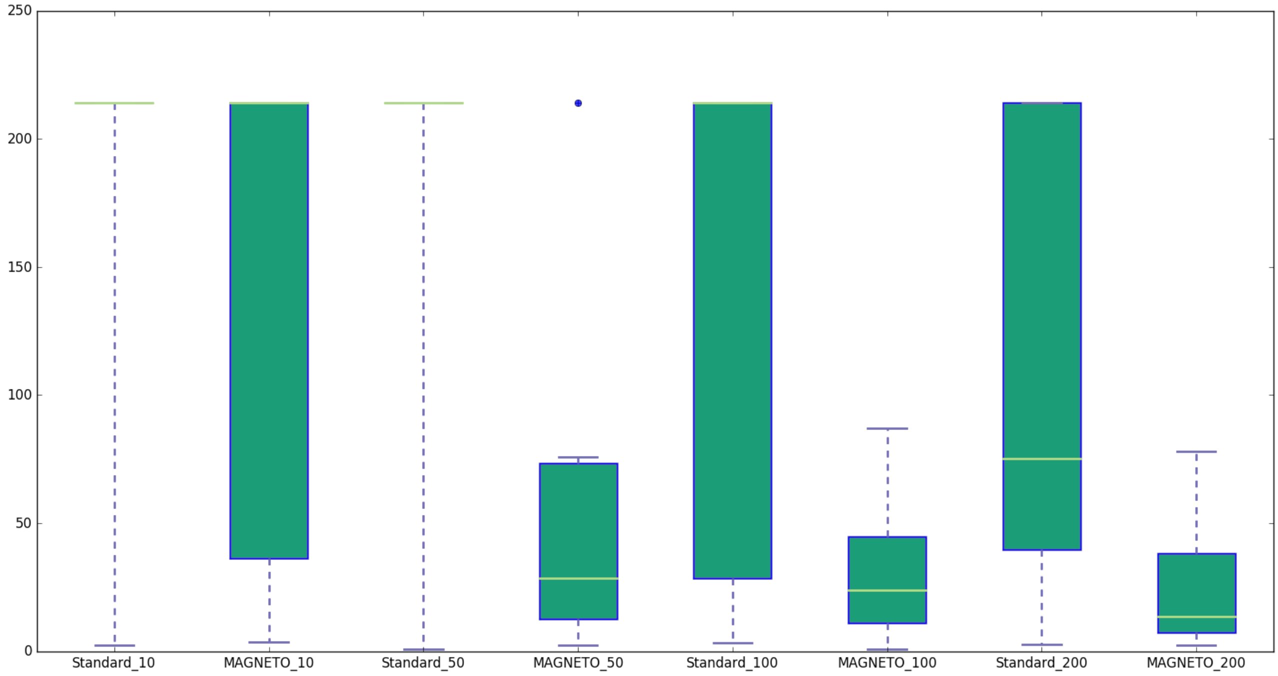


Figure 32: The box plots represent the ranks distribution of the 21 datasets. Both MAGNETO and the standard ORA are shown. The label on the x-axis represents the method (standard or MAGNETO) the number next to the label represents the number of proteins submitted.

We selected lists of upregulated genes of increased length (10, 50, 100, 200) and submitted to MAGNETO and to the standard ORA method. The two methods were compared on the distributions of -Log(p-values) and rank.

The box plots in Fig 31 show the distributions of -Log(p-values) for the 21 datasets of both MAGNETO and the standard ORA method. The box plots in Fig 32 show the ranks distribution for the 21 datasets for both MAGNETO and the standard ORA method. It is evident that MAGNETO outperforms the standard ORA method in terms of p-value and rank distributions when the number of proteins submitted is greater than 50.

Applying MAGNETO to the pig data

MAGNETO currently uses information from the human protein interaction network. To apply MAGNETO to the pig transcriptomic data it was necessary to convert the genes expressed in the pig to the one to one human ortholog. This had also the advantage of employing the more detailed human annotation at the cost of losing some pig genes that were highly expressed. The ORA was performed by selecting the top 50 expressed genes in each cell type and in each time point. We selected a p-value threshold equal to 0.001 Bonferroni corrected. The background was represented by the genes that have at least an annotation in the BP domain. The protein lists employed for the ORA represented the seed nodes for MAGNETO. Additionally, we selected the IntAct network and the heart muscle as tissue since MAGNETO requires a tissue in the input. To demonstrate the advantage of MAGNETO over the ORA method, we compared the BPs identified by MAGNETO to those obtained with the ORA.

Protein distribution among Coarse Grained Modules

The first step of the MAGNETO pipeline was to assign proteins to the CGMs. The CGMs are a gross grained representation of a top-level pathway, meaning a set of proteins involved in a very general biological process (i.e. gene expression, immune system). It is important to notice that proteins can be assigned also to different CGMs, for this reason the sum of the assigned proteins to the CGMs is greater than 50. Fig 33 shows how the proteins are assigned to the CGMs. The x-axis represents the top-level pathways, if the top-level pathway is absent, the proteins were not assigned. The cell types are colour coded and each bar of the same colour corresponds to a different and consecutive time point (i.e. the first blue bar corresponds to the number of proteins in CM at day 0 assigned to the top-level pathway, the second blue bar corresponds to the number of proteins in CM at day 3 assigned to the top-level pathway as so on).

Interestingly, in FBs, about half of the 50 input proteins (yellow bars) are always assigned to "extracellular matrix organization" (25 at day 0, 22 at day 3 and 25 at day 7). FBs are the only cell type showing a constant behaviour in terms of proteins assigned to the top-level pathways regardless of the time points.

Conversely, in CMs (blue bars), the proteins assigned to the top-level pathway fluctuate more and reflect more the impact of the inflammatory phase. A comparison between proteins involved in top-level pathway in day 0 (no infarct) with respect to 3 days post MI shows a decrease of proteins assigned to muscle contraction (10 less) and metabolism (7 less). This decrease is balanced by an increase of proteins assigned to signal transduction (10 more), immune system (14 more), developmental biology (10 more) and extracellular matrix organization (8 more). In comparison with day 3, at day 7, the proteins assigned to muscle contraction show an increase (5 more) while the proteins involved in the immune system are clearly decreased (10 less) coming back to the level of day 0. The proteins involved in metabolism, signal transduction and developmental biology and extracellular matrix remained at the same level of day 3.

In MFs (green bars), the proteins involved in immune system undergo an increase (6 more) at day 3 in comparison with day 0 as well as a slight increase (4 more) is observable in proteins involved in hemostasis but then come back to the same level of day 0.

In ECs (red bars), an increase of proteins involved in developmental biology is observed (6 more) in day 3 and day 7 with respect to day 0. A slight increase (5 more) of proteins involved in the immune system is detected at day 7 in comparison with day 0.

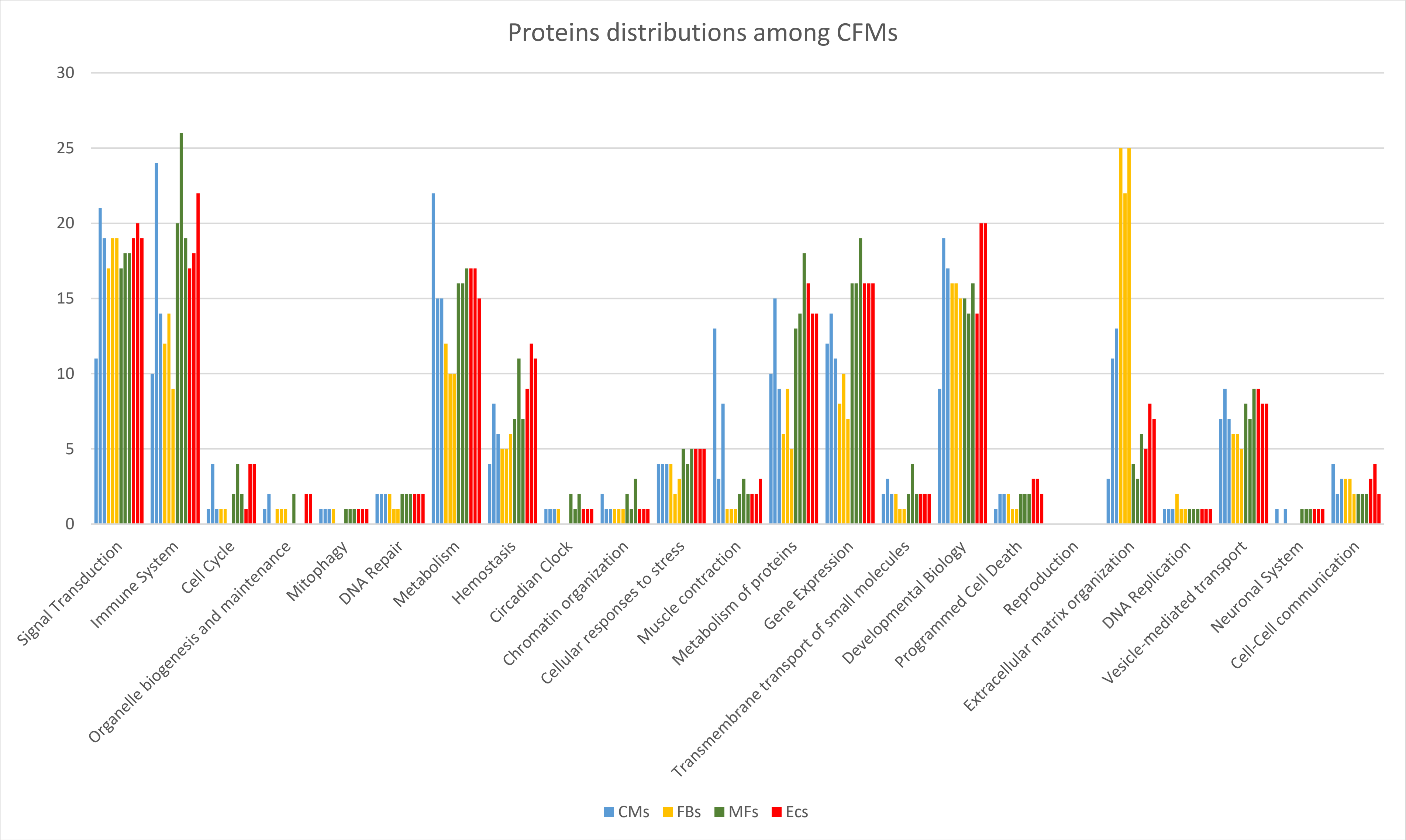


Figure 33: How the proteins are assigned to the CGMs. The cell types are colour coded and each consecutive bar correspond to a different and consecutive time point (i.e. the first blue bar corresponds to the number of proteins in CMs at day 0 assigned to the top-level pathway, the second blue bar corresponds to the number of proteins in CMs at day 3 assigned to the top-level pathway as so on).

MAGNETO: final networks

The network extracted with MAGNETO are highly representative of the interactions that take place in the selected tissue. Fig 34 shows how the number of nodes/proteins needed to connect each couple of seed nodes in any of the CGM networks. On average, the final networks are composed of 261.25 nodes/proteins meaning that about 200 nodes/proteins more were required to connect 50 seed proteins in input.

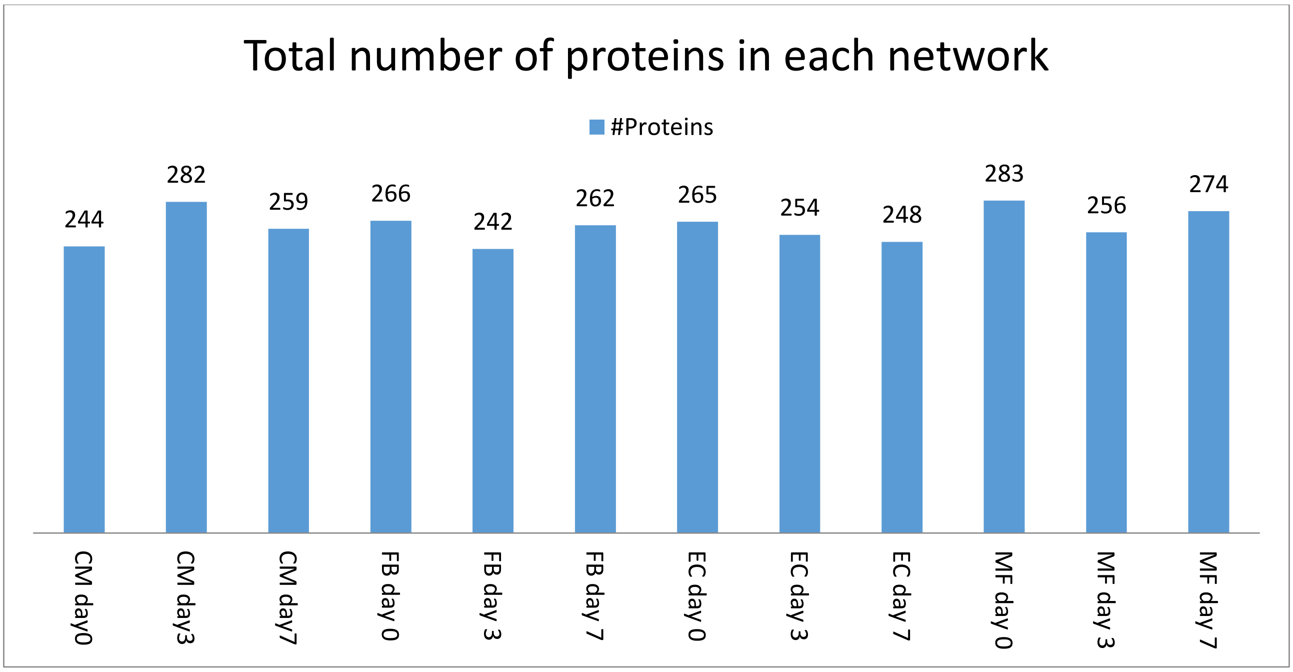
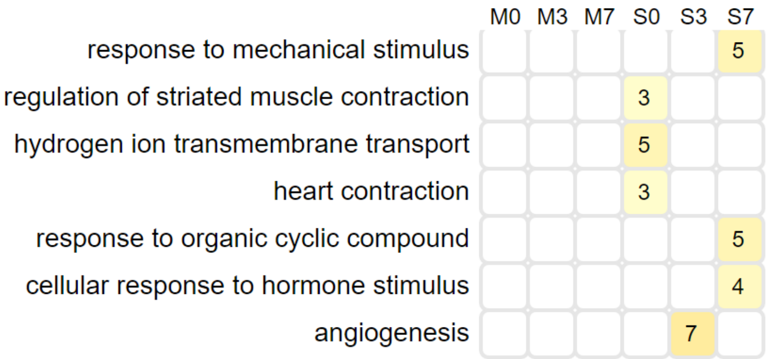


Figure 34: Number of nodes/proteins needed to connect each couple of seed nodes in any of the CGM networks.

The nodes/proteins obtained with MAGNETO represent the testing set for the ORA. The background was represented by the BPs domain of the Gene Ontology database. Even though the BPs were Bonferroni corrected and a stringent p-value (0.001) was chosen, the list of overrepresented BPs was quite large. Visualizing large lists of overrepresented BPs is not a trivial task. The enriched BPs detected by MAGNETO and ORA were visualized through a heatmap. The rows show the name of the enriched BPs, the columns represent the different time points and method used (i.e. M0 means MAGNETO at day 0, S0 means standard at day 0). The squares represent the number of proteins involved in the enriched process, a white square represents no enrichment. Each square adopts a red-yellow colour code. The redder the square is, the higher the number of proteins involved in that BP are. This type of representation helps to figure out which are the BPs in common between the different time points and which one are specific to the cell type at a certain time point.

Enriched processes in Cardiomyocytes

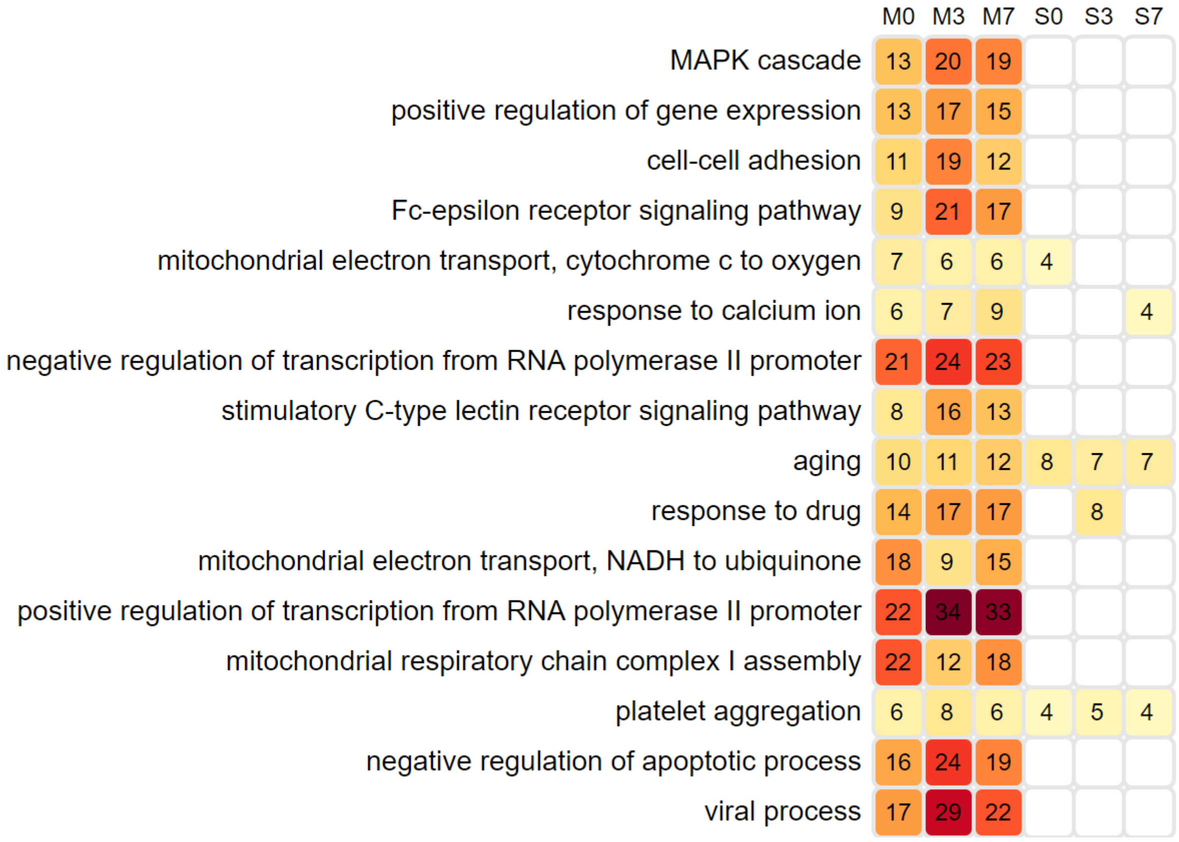
We first analysed the BP terms enriched in the CM gene lists, comparing the performance of MAGNETO to that of ORA. Fig 35 shows the BPs detected in CMs only with ORA but not detected with MAGNETO. MAGNETO did not detect these BPs because it deals with a bigger testing set, therefore, some of the terms may not appear enriched. In CMs, 7 BPs were enriched using the standard ORA. Many of the BPs found by ORA are related to heart and muscle contraction.



	M0	M3	M7	S0	S3	S7
response to mechanical stimulus						5
regulation of striated muscle contraction				3		
hydrogen ion transmembrane transport				5		
heart contraction				3		
response to organic cyclic compound						5
cellular response to hormone stimulus						4
angiogenesis					7	

Figure 35: BPs detected only with the standard ORA but not detected by MAGNETO in CMs

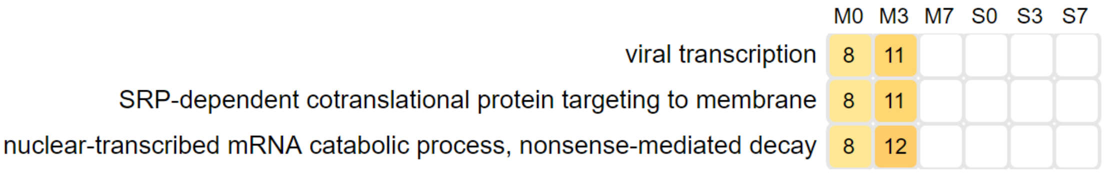
Fig 36 shows the BPs detected by MAGNETO in CMs and commonly enriched at days 0, 3, and 7. MAGNETO detected 16 BPs commonly enriched, five were also detected with the standard ORA but not in all the time points. In comparison with ORA, MAGNETO detected BPs involved in immune response and inflammation (Fc-epsilon receptor signalling pathway, MAPK cascade, stimulatory C-type lectin receptor signalling pathway) and mitochondrial respiration (mitochondrial electron transport, cytochrome c to oxygen, mitochondrial electron transport, NADH to ubiquinone, mitochondrial respiratory chain complex I assembly). Interestingly, the number of proteins involved in immune response and inflammation increased at day 3 and remained stable at day 7, while the proteins involved in mitochondrial respiration decreased at day 3 and increased again at day 7.



	M0	M3	M7	S0	S3	S7
MAPK cascade	13	20	19			
positive regulation of gene expression	13	17	15			
cell-cell adhesion	11	19	12			
Fc-epsilon receptor signaling pathway	9	21	17			
mitochondrial electron transport, cytochrome c to oxygen	7	6	6	4		
response to calcium ion	6	7	9			4
negative regulation of transcription from RNA polymerase II promoter	21	24	23			
stimulatory C-type lectin receptor signaling pathway	8	16	13			
aging	10	11	12	8	7	7
response to drug	14	17	17		8	
mitochondrial electron transport, NADH to ubiquinone	18	9	15			
positive regulation of transcription from RNA polymerase II promoter	22	34	33			
mitochondrial respiratory chain complex I assembly	22	12	18			
platelet aggregation	6	8	6	4	5	4
negative regulation of apoptotic process	16	24	19			
viral process	17	29	22			

Figure 36: BPs detected by MAGNETO and ORA and enriched at day 0, 3 and 7 in CMs

Fig 37 shows the BPs detected by MAGNETO in CMs and commonly enriched at day 0 and 3. MAGNETO detected 5 BPs in CMs commonly enriched between day 0 and 3, none of them was detected with the standard ORA.

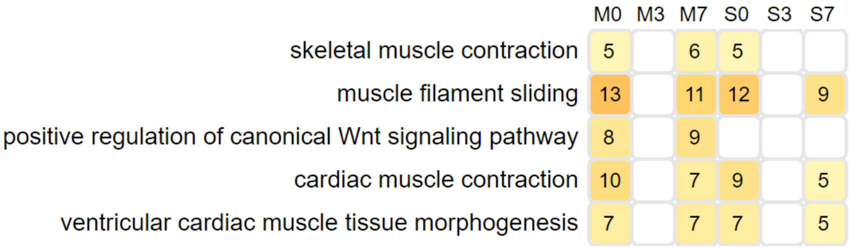


	M0	M3	M7	S0	S3	S7
viral transcription	8	11				
SRP-dependent cotranslational protein targeting to membrane	8	11				
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	8	12				

Figure 37: BPs detected by MAGNETO and commonly enriched at day 0 and 3 in CMs

Fig 38 shows the BPs detected by MAGNETO in CMs and commonly enriched at days 0 and 7. The BPs enriched at time point 0 and 7 showed a very similar profile between MAGNETO and the standard ORA. With respect to the standard ORA, MAGNETO additionally detected the BPs ‘positive regulation of canonical Wnt signalling pathway’ at time point 0 and 7, and ‘skeletal muscle contraction’ at time point 7.

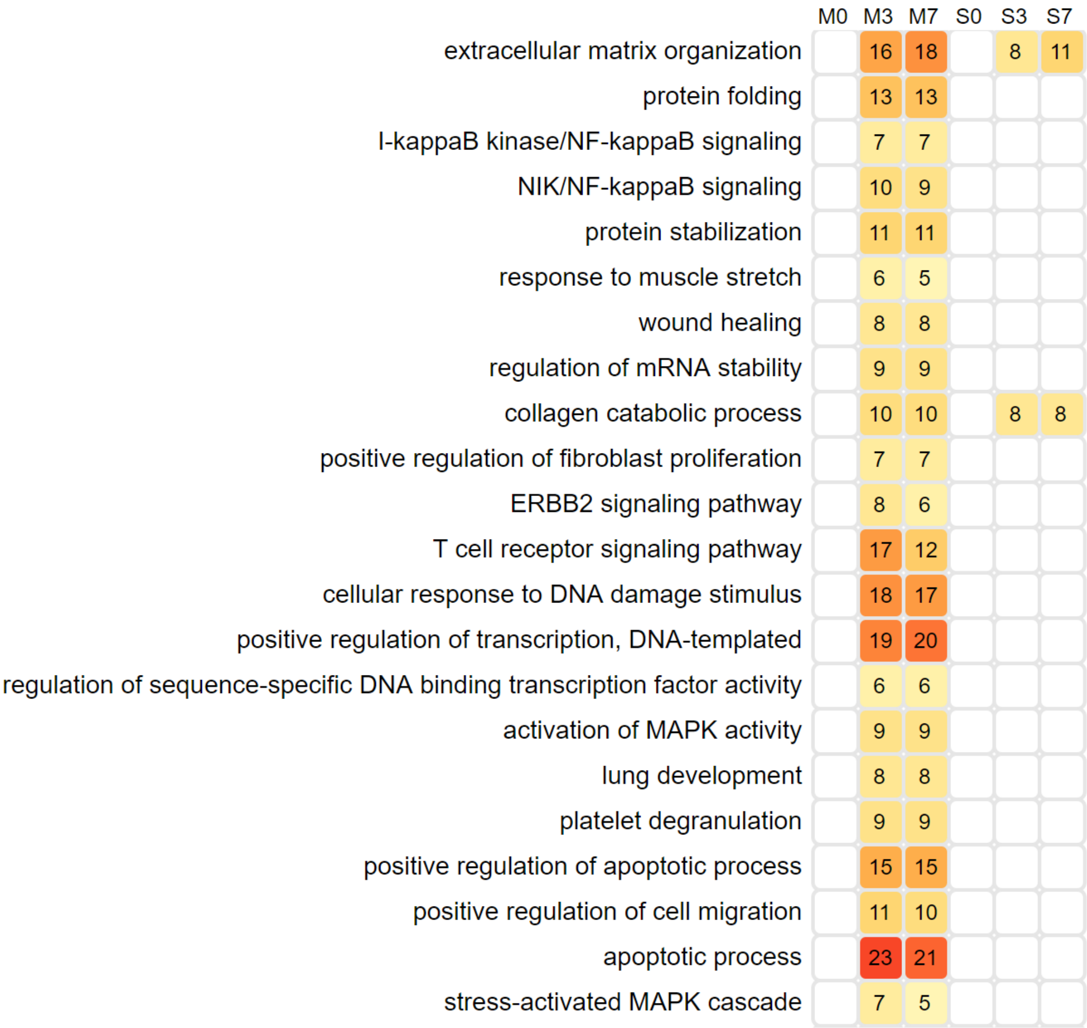
Fig 39 shows the BPs detected by MAGNETO in CMs and commonly enriched at days 3 and 7. MAGNETO detected 22 BPs commonly enriched between day 3 and 7, two of them (extracellular matrix organization and collagen catabolic process) were enriched also with the standard ORA



	M0	M3	M7	S0	S3	S7
skeletal muscle contraction	5		6	5		
muscle filament sliding	13		11	12		9
positive regulation of canonical Wnt signaling pathway	8		9			
cardiac muscle contraction	10		7	9		5
ventricular cardiac muscle tissue morphogenesis	7		7	7		5

Figure 38: BPs detected by MAGNETO and ORA and commonly enriched at day 0 and 7 in CMs

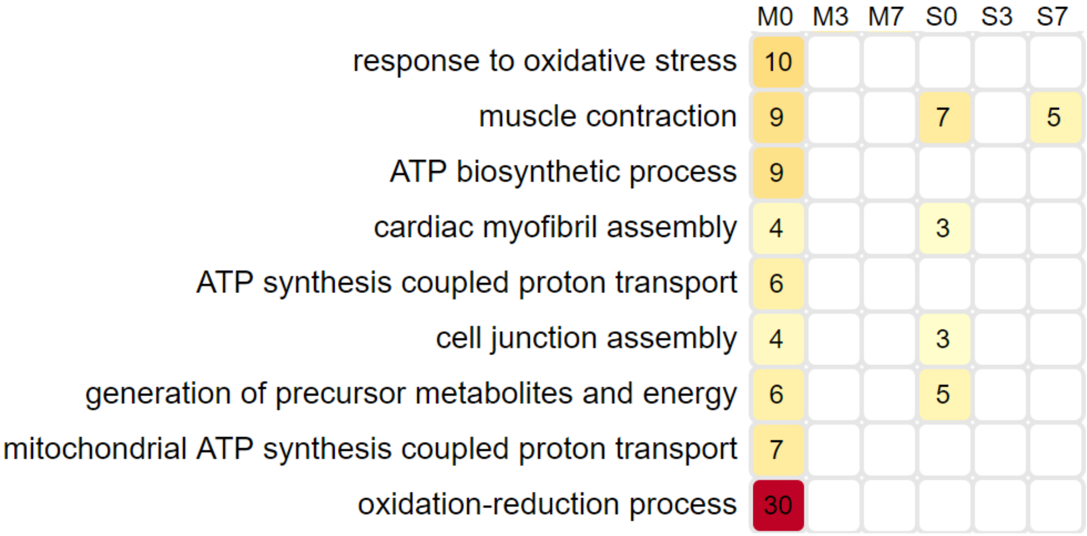
In addition, MAGNETO detected BPs involved in immune response and inflammation (I-kappaB kinase/NF-kappaB signalling, NIK/NF-kappaB signalling, wound healing, response to muscle stretch, stress-activated MAPK cascade, activation of MAPK activity, T cell receptor signalling pathway), proliferation and migration (positive regulation of fibroblast proliferation, positive regulation of cell proliferation) and apoptosis.



	M0	M3	M7	S0	S3	S7
extracellular matrix organization		16	18		8	11
protein folding		13	13			
I-kappaB kinase/NF-kappaB signaling		7	7			
NIK/NF-kappaB signaling		10	9			
protein stabilization		11	11			
response to muscle stretch		6	5			
wound healing		8	8			
regulation of mRNA stability		9	9			
collagen catabolic process		10	10		8	8
positive regulation of fibroblast proliferation		7	7			
ERBB2 signaling pathway		8	6			
T cell receptor signaling pathway		17	12			
cellular response to DNA damage stimulus		18	17			
positive regulation of transcription, DNA-templated		19	20			
regulation of sequence-specific DNA binding transcription factor activity		6	6			
activation of MAPK activity		9	9			
lung development		8	8			
platelet degranulation		9	9			
positive regulation of apoptotic process		15	15			
positive regulation of cell migration		11	10			
apoptotic process		23	21			
stress-activated MAPK cascade		7	5			

Figure 39: BPs detected by MAGNETO and ORA and commonly enriched at day 3 and 7 in CMs

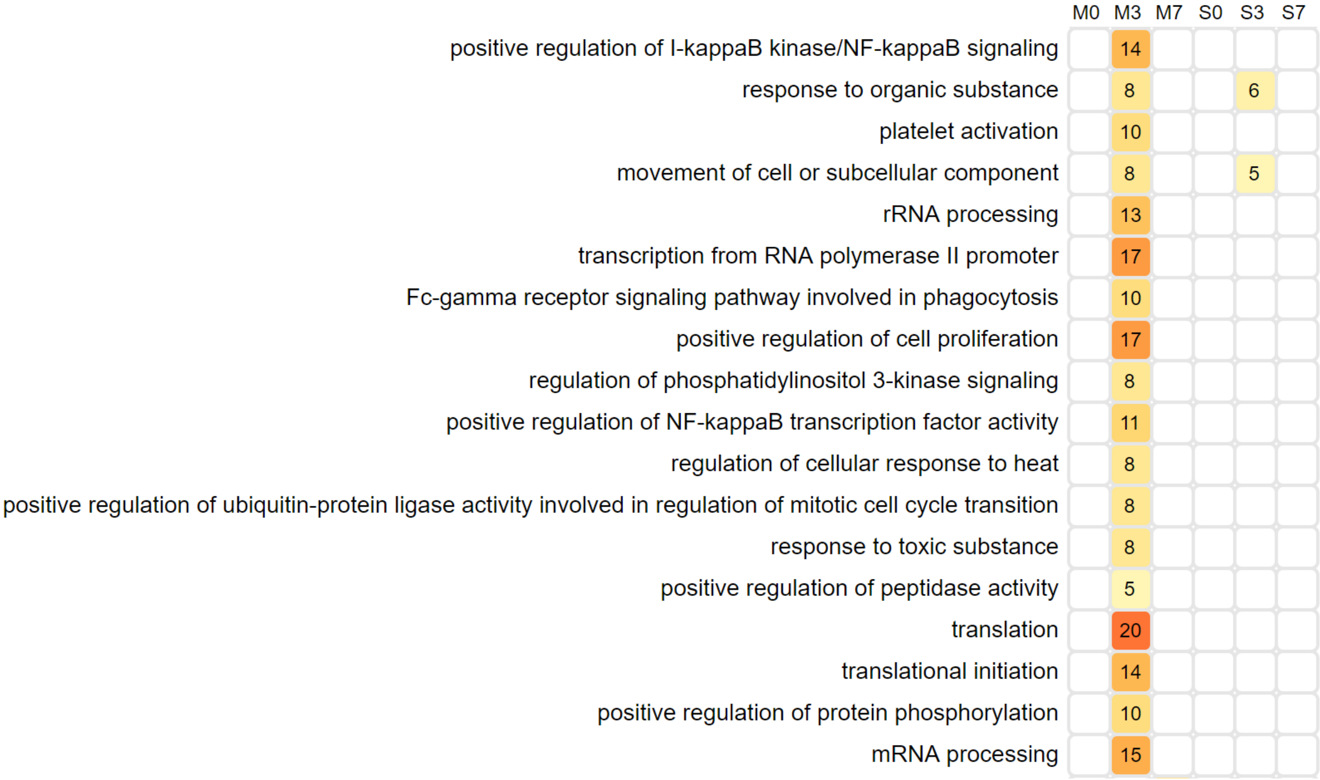
Fig 40 shows the BPs detected by MAGNETO in CMs and commonly enriched at day 0. MAGNETO detected 9 BPs enriched at time point 0, four were also enriched with the standard ORA. The enriched BP “muscle contraction” was detected only with the standard ORA at day 7. In comparison with ORA, MAGNETO detected in addition BPs involved in ATP production typical of the mitochondria.



	M0	M3	M7	S0	S3	S7
response to oxidative stress	10					
muscle contraction	9			7		5
ATP biosynthetic process	9					
cardiac myofibril assembly	4			3		
ATP synthesis coupled proton transport	6					
cell junction assembly	4			3		
generation of precursor metabolites and energy	6			5		
mitochondrial ATP synthesis coupled proton transport	7					
oxidation-reduction process	30					

Figure 40: BPs detected by MAGNETO and ORA and enriched at day 0 in CMs

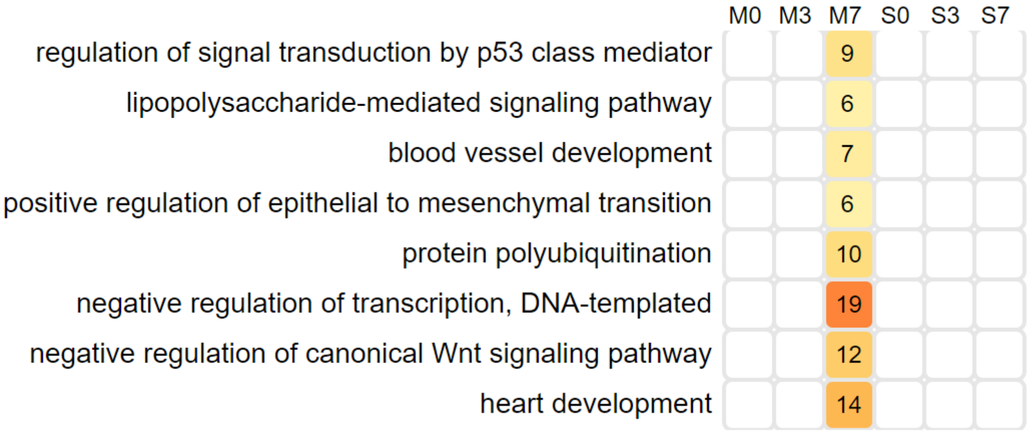
Fig 41 shows the BPs detected by MAGNETO in CMs and enriched at day 3. MAGNETO detected 19 BPs enriched at day 3, two of them (movement of cell or subcellular component and response to organic substance) were also detected by the standard ORA. In comparison with ORA, the BPs detected by MAGNETO at day 3 were related mainly to immune response and phagocytosis (positive regulation of NF-kappaB transcription factor activity, Fc-gamma receptor signaling pathway involved in phagocytosis).



	M0	M3	M7	S0	S3	S7
positive regulation of I-kappaB kinase/NF-kappaB signaling		14				
response to organic substance		8			6	
platelet activation		10				
movement of cell or subcellular component		8			5	
rRNA processing		13				
transcription from RNA polymerase II promoter		17				
Fc-gamma receptor signaling pathway involved in phagocytosis		10				
positive regulation of cell proliferation		17				
regulation of phosphatidylinositol 3-kinase signaling		8				
positive regulation of NF-kappaB transcription factor activity		11				
regulation of cellular response to heat		8				
positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycle transition		8				
response to toxic substance		8				
positive regulation of peptidase activity		5				
translation		20				
translational initiation		14				
positive regulation of protein phosphorylation		10				
mRNA processing		15				

Figure 41: BPs detected by MAGNETO and ORA and enriched at day 3 in CMs

Fig 42 shows the BPs detected by MAGNETO in CMs and commonly enriched at day 7. MAGNETO detected 8 BPs at day 7, none of them was detected with the standard ORA. In comparison with ORA, MAGNETO detected BPs involved mainly in cardiac remodelling (blood vessel development, positive regulation of epithelial to mesenchymal transition, heart development).

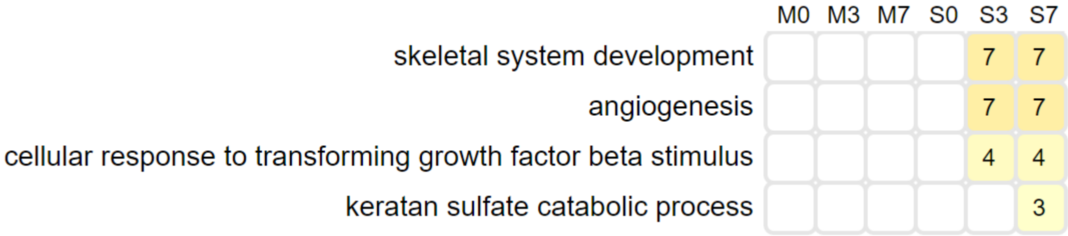


	M0	M3	M7	S0	S3	S7
regulation of signal transduction by p53 class mediator			9			
lipopolysaccharide-mediated signaling pathway			6			
blood vessel development			7			
positive regulation of epithelial to mesenchymal transition			6			
protein polyubiquitination			10			
negative regulation of transcription, DNA-templated			19			
negative regulation of canonical Wnt signaling pathway			12			
heart development			14			

Figure 42: BPs detected by MAGNETO and enriched at day7 in CMs

Enriched processes in Fibroblasts

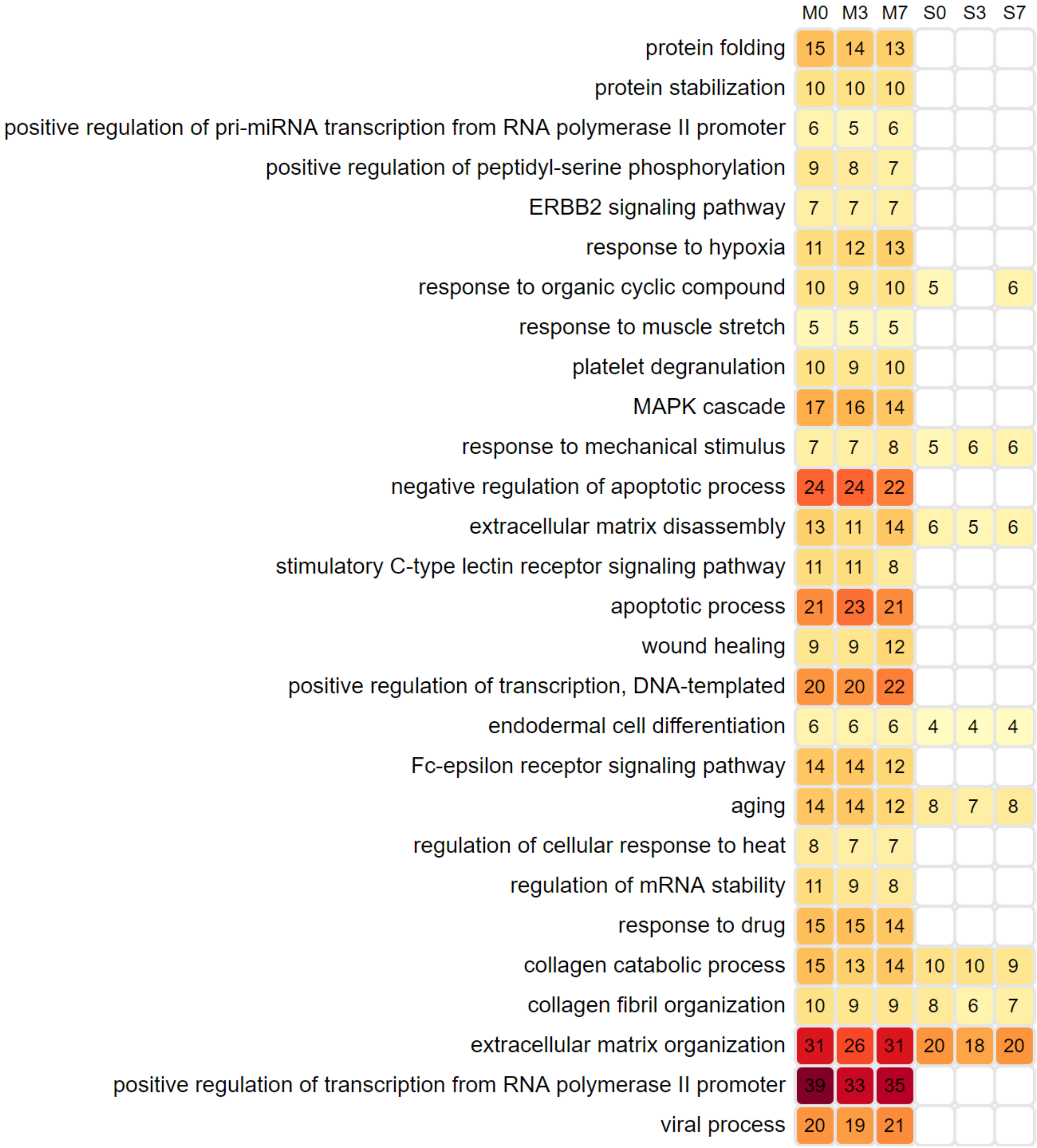
Fig 43 shows the BPs detected in FBs only with ORA but absent in MAGNETO. MAGNETO did not detect these BPs because it deals with a bigger testing set, as a consequence some of the terms may not appear enriched. In FBs, 4 BPs were enriched using the standard ORA.



	M0	M3	M7	S0	S3	S7
skeletal system development					7	7
angiogenesis					7	7
cellular response to transforming growth factor beta stimulus					4	4
keratan sulfate catabolic process						3

Figure 43: BPs detected only with the standard ORA but not detected by MAGNETO in FBs

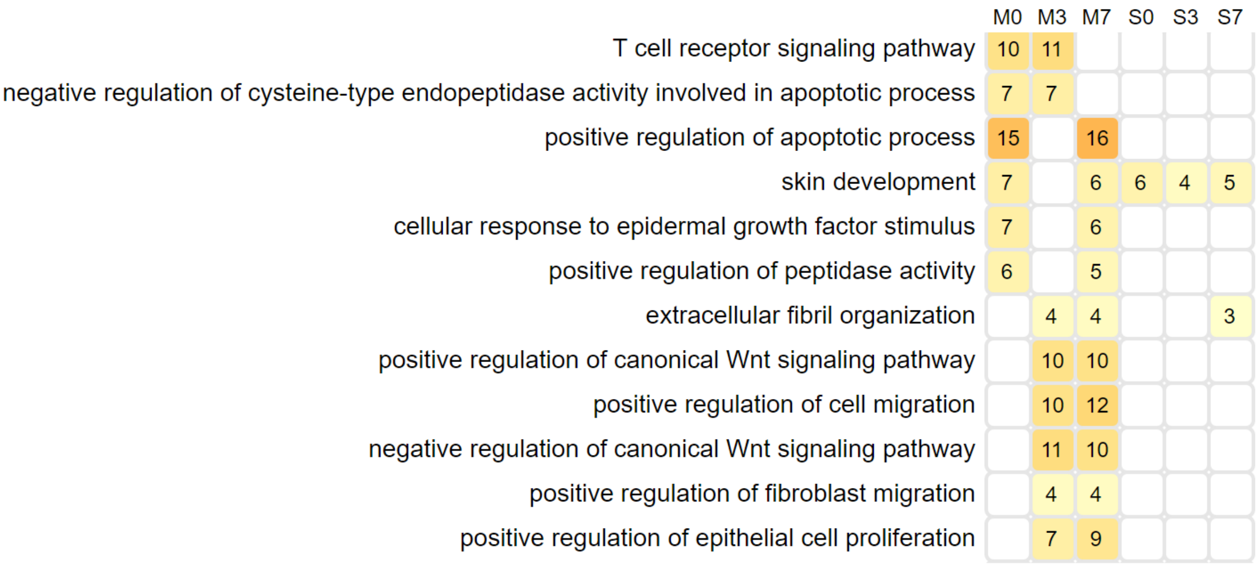
Fig 44 shows the BPs detected by MAGNETO in FBs and commonly enriched at days 0, 3, and 7. MAGNETO detected 28 BPs commonly enriched in all the time points in FBs, 8 of them were detected also with the standard ORA, both methods mainly enriched BPs related to extracellular matrix. MAGNETO additionally detected BPs related to immune response and inflammation (MAPK cascade, stimulatory C-type lectin receptor signalling pathway, wound healing, Fc-epsilon receptor signalling pathway, response to muscle stretch, response to hypoxia). We did not observe a fluctuation in terms of proteins assigned to any of the BPs commonly enriched, unlike in CMs. This is consistent with the constant behaviour in terms of proteins assigned to the CGMs (please see paragraph Protein distribution among Coarse Grained Modules)



	M0	M3	M7	S0	S3	S7
protein folding	15	14	13			
protein stabilization	10	10	10			
positive regulation of pri-miRNA transcription from RNA polymerase II promoter	6	5	6			
positive regulation of peptidyl-serine phosphorylation	9	8	7			
ERBB2 signaling pathway	7	7	7			
response to hypoxia	11	12	13			
response to organic cyclic compound	10	9	10	5		6
response to muscle stretch	5	5	5			
platelet degranulation	10	9	10			
MAPK cascade	17	16	14			
response to mechanical stimulus	7	7	8	5	6	6
negative regulation of apoptotic process	24	24	22			
extracellular matrix disassembly	13	11	14	6	5	6
stimulatory C-type lectin receptor signaling pathway	11	11	8			
apoptotic process	21	23	21			
wound healing	9	9	12			
positive regulation of transcription, DNA-templated	20	20	22			
endodermal cell differentiation	6	6	6	4	4	4
Fc-epsilon receptor signaling pathway	14	14	12			
aging	14	14	12	8	7	8
regulation of cellular response to heat	8	7	7			
regulation of mRNA stability	11	9	8			
response to drug	15	15	14			
collagen catabolic process	15	13	14	10	10	9
collagen fibril organization	10	9	9	8	6	7
extracellular matrix organization	31	26	31	20	18	20
positive regulation of transcription from RNA polymerase II promoter	39	33	35			
viral process	20	19	21			

Figure 44: BPs detected by MAGNETO and ORA and enriched at days 0, 3 and 7 in FBs

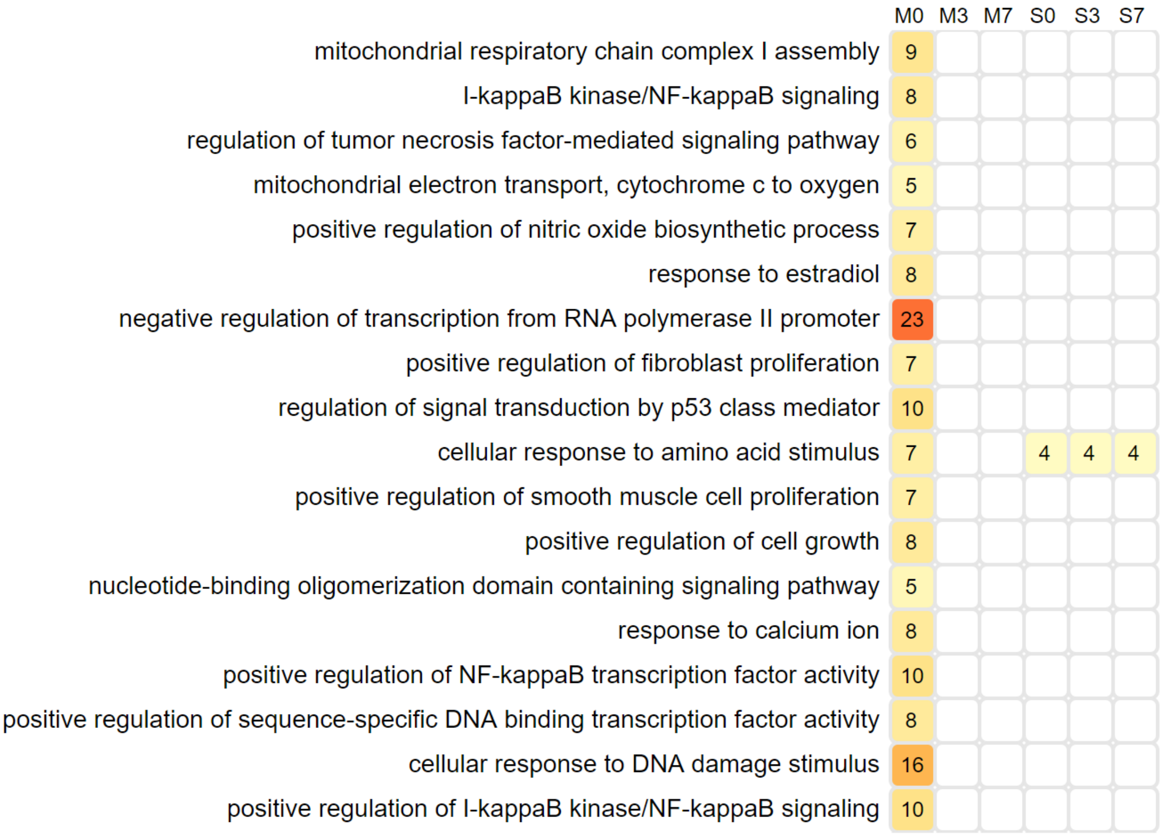
Fig 45 shows the BPs detected by MAGNETO in FBs and commonly enriched at days 0 and 3, days 0 and 7 and days 3 and 7. MAGNETO detected 2 BPs commonly enriched at day 0 and 3, four BPs were commonly enriched at days 0 and 7, of these the "skin development" was detected also with the standard ORA and at day 3. MAGNETO detected 6 BPs commonly enriched at days 3 and 7, of these, the BP "extracellular fibril organization" was detected with ORA but only at day 7. Interestingly, in comparison with ORA, the BPs detected with MAGNETO and commonly enriched were related to fibroblast and cell migration and epithelial cell proliferation.



	M0	M3	M7	S0	S3	S7
T cell receptor signaling pathway	10	11				
negative regulation of cysteine-type endopeptidase activity involved in apoptotic process	7	7				
positive regulation of apoptotic process	15		16			
skin development	7		6	6	4	5
cellular response to epidermal growth factor stimulus	7		6			
positive regulation of peptidase activity	6		5			
extracellular fibril organization		4	4			3
positive regulation of canonical Wnt signaling pathway		10	10			
positive regulation of cell migration		10	12			
negative regulation of canonical Wnt signaling pathway		11	10			
positive regulation of fibroblast migration		4	4			
positive regulation of epithelial cell proliferation		7	9			

Figure 45: BPs detected by MAGNETO and ORA and commonly enriched at days 0 and 3, at days 0 and 7, and at days 3 and 7 in FBs

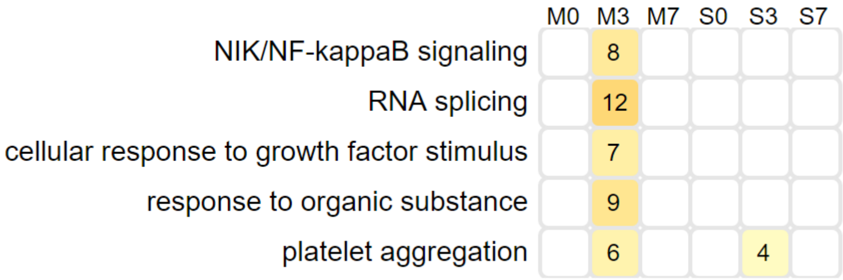
Fig 46 shows the BPs detected by MAGNETO and ORA in FBs and enriched at day 0. MAGNETO detected 18 BPs enriched only at day 0. The "cellular response to amino acid stimulus" was detected also by the standard ORA and at day 3 and day 7. In comparison with ORA, MAGNETO added enriched BPs related to immune response, fibroblast proliferation and the tumor necrosis factor signalling.



	M0	M3	M7	S0	S3	S7
mitochondrial respiratory chain complex I assembly	9					
I-kappaB kinase/NF-kappaB signaling	8					
regulation of tumor necrosis factor-mediated signaling pathway	6					
mitochondrial electron transport, cytochrome c to oxygen	5					
positive regulation of nitric oxide biosynthetic process	7					
response to estradiol	8					
negative regulation of transcription from RNA polymerase II promoter	23					
positive regulation of fibroblast proliferation	7					
regulation of signal transduction by p53 class mediator	10					
cellular response to amino acid stimulus	7			4	4	4
positive regulation of smooth muscle cell proliferation	7					
positive regulation of cell growth	8					
nucleotide-binding oligomerization domain containing signaling pathway	5					
response to calcium ion	8					
positive regulation of NF-kappaB transcription factor activity	10					
positive regulation of sequence-specific DNA binding transcription factor activity	8					
cellular response to DNA damage stimulus	16					
positive regulation of I-kappaB kinase/NF-kappaB signaling	10					

Figure 46: BPs detected by MAGNETO and ORA enriched at day 0 in FBs

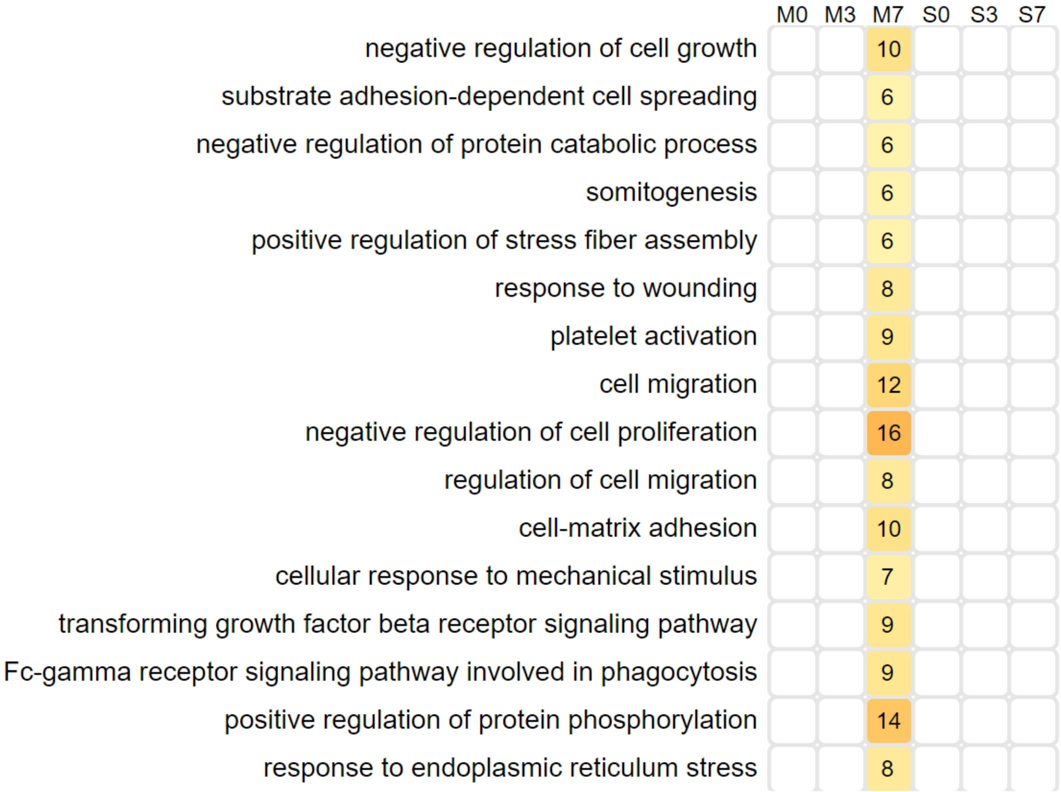
Fig 47 shows the BPs detected by MAGNETO and ORA in FBs and commonly enriched at day 3. MAGNETO detected 5 enriched BPs at day 3. The BP “platelet aggregation” was detected also with the standard ORA at the same day. In comparison with ORA, MAGNETO added BPs related to immune response (NIK/NF-kappaB signalling) and platelet aggregation.



	M0	M3	M7	S0	S3	S7
NIK/NF-kappaB signaling		8				
RNA splicing		12				
cellular response to growth factor stimulus		7				
response to organic substance		9				
platelet aggregation		6			4	

Figure 47: BPs detected by MAGNETO and ORA and enriched at day 3 in FBs

Fig 48 shows the BPs detected by MAGNETO in FBs and commonly enriched at day 7. MAGNETO detected 16 BPs enriched at day 7, none of them was detected with ORA. The enriched BPs detected by MAGNETO, but not with ORA, were mainly related to cell adhesion (cell-matrix adhesion, substrate adhesion-dependent cell spreading), proliferation, migration, and signalling (transforming growth factor beta receptor signaling pathway, Fc-gamma receptor signaling pathway involved in phagocytosis).

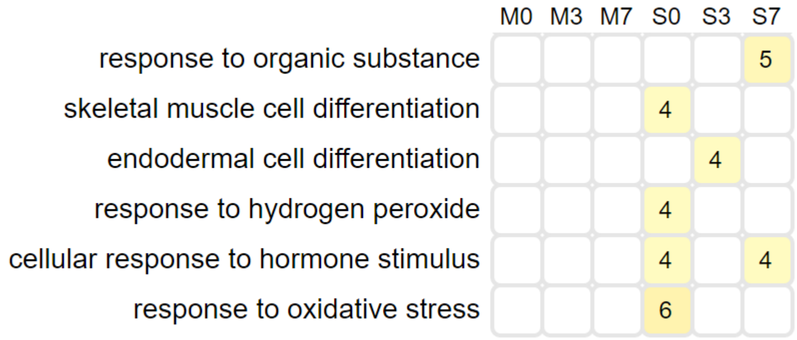


	M0	M3	M7	S0	S3	S7
negative regulation of cell growth			10			
substrate adhesion-dependent cell spreading			6			
negative regulation of protein catabolic process			6			
somitogenesis			6			
positive regulation of stress fiber assembly			6			
response to wounding			8			
platelet activation			9			
cell migration			12			
negative regulation of cell proliferation			16			
regulation of cell migration			8			
cell-matrix adhesion			10			
cellular response to mechanical stimulus			7			
transforming growth factor beta receptor signaling pathway			9			
Fc-gamma receptor signaling pathway involved in phagocytosis			9			
positive regulation of protein phosphorylation			14			
response to endoplasmic reticulum stress			8			

Figure 48: BPs detected by MAGNETO and enriched at day 7 in FBs

Enriched processes in Endothelial cells

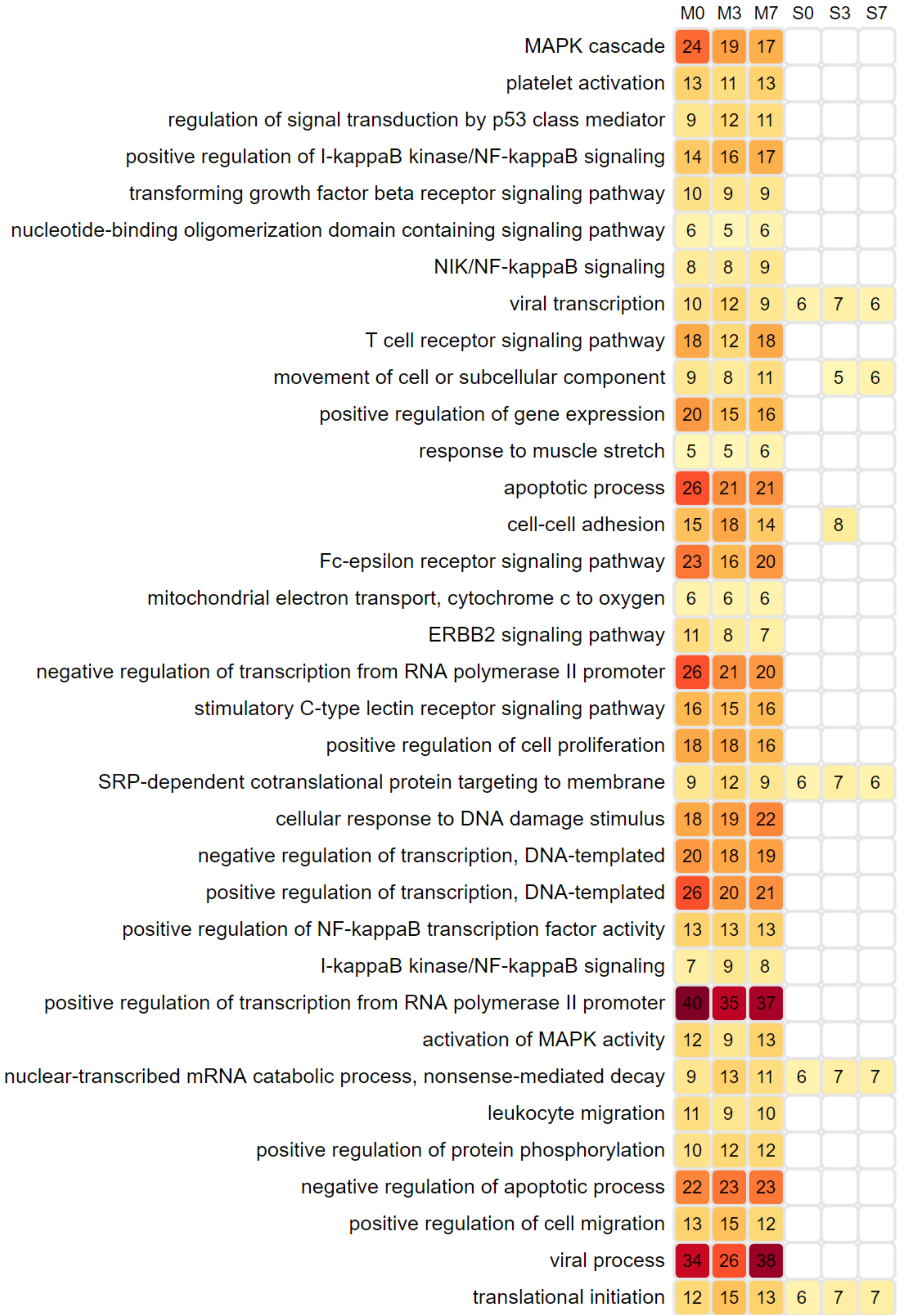
Fig 49 shows the BPs detected with ORA but not with MAGNETO in ECs. Considering any of the days, ORA detected 6 BPs, two of them were related to cell differentiation even though in two different days.



	M0	M3	M7	S0	S3	S7
response to organic substance						5
skeletal muscle cell differentiation				4		
endodermal cell differentiation					4	
response to hydrogen peroxide				4		
cellular response to hormone stimulus				4		4
response to oxidative stress				6		

Figure 49: BPs detected only with ORA method but not detected by MAGNETO in ECs

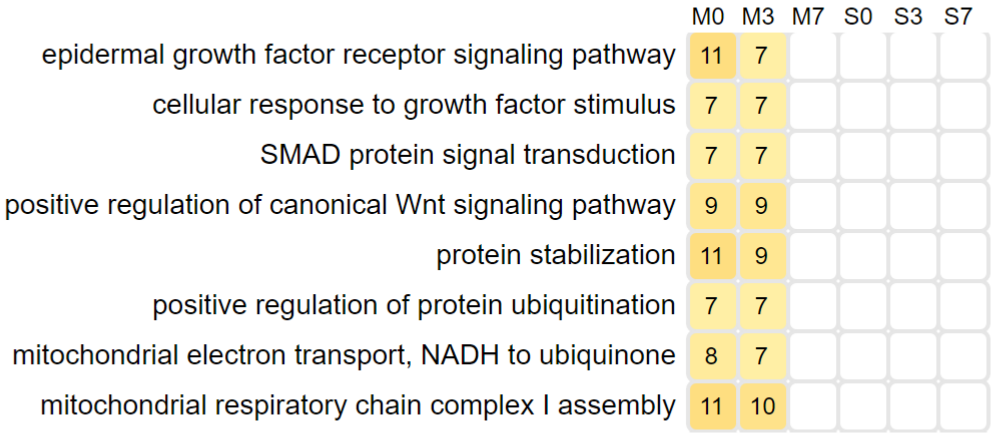
Fig 50 shows the BPs detected by MAGNETO in ECs and commonly enriched at days 0, 3, and 7. MAGNETO detected 35 BPs enriched in all the days in ECs, 6 of them were detected also with the ORA method, but two of them (cell-cell adhesion and movement of cell or subcellular component) were not enriched in all the days. As in all the other cell types, most of the BPs detected with MAGNETO in ECs were related to the immune response, additionally MAGNETO detected the BP “leukocyte migration”. Moreover, several BPs detected only with MAGNETO are associated to signalling (transforming growth factor beta receptor signalling pathway, Fc-epsilon receptor signalling pathway). Like in the CMs, with respect to day 0, the proteins involved in several BPs (MAPK cascade, Apoptotic process, positive regulation of gene expression, positive regulation of transcription, DNA-templated, negative regulation of transcription from RNA polymerase II promoter) decreased at day 3 and remained almost constant at day 7. The proteins involved in BPs related to immune response (T cell receptor signalling pathway and Fc-epsilon receptor signalling pathway) decreased at day 3 and then increased at day 7, reaching almost the same number of day 0.



	M0	M3	M7	S0	S3	S7
MAPK cascade	24	19	17			
platelet activation	13	11	13			
regulation of signal transduction by p53 class mediator	9	12	11			
positive regulation of I-kappaB kinase/NF-kappaB signaling	14	16	17			
transforming growth factor beta receptor signaling pathway	10	9	9			
nucleotide-binding oligomerization domain containing signaling pathway	6	5	6			
NIK/NF-kappaB signaling	8	8	9			
viral transcription	10	12	9	6	7	6
T cell receptor signaling pathway	18	12	18			
movement of cell or subcellular component	9	8	11		5	6
positive regulation of gene expression	20	15	16			
response to muscle stretch	5	5	6			
apoptotic process	26	21	21			
cell-cell adhesion	15	18	14		8	
Fc-epsilon receptor signaling pathway	23	16	20			
mitochondrial electron transport, cytochrome c to oxygen	6	6	6			
ERBB2 signaling pathway	11	8	7			
negative regulation of transcription from RNA polymerase II promoter	26	21	20			
stimulatory C-type lectin receptor signaling pathway	16	15	16			
positive regulation of cell proliferation	18	18	16			
SRP-dependent cotranslational protein targeting to membrane	9	12	9	6	7	6
cellular response to DNA damage stimulus	18	19	22			
negative regulation of transcription, DNA-templated	20	18	19			
positive regulation of transcription, DNA-templated	26	20	21			
positive regulation of NF-kappaB transcription factor activity	13	13	13			
I-kappaB kinase/NF-kappaB signaling	7	9	8			
positive regulation of transcription from RNA polymerase II promoter	40	35	37			
activation of MAPK activity	12	9	13			
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	9	13	11	6	7	7
leukocyte migration	11	9	10			
positive regulation of protein phosphorylation	10	12	12			
negative regulation of apoptotic process	22	23	23			
positive regulation of cell migration	13	15	12			
viral process	34	26	38			
translational initiation	12	15	13	6	7	7

Figure 50: BPs detected by MAGNETO and ORA and enriched at days 0, 3 and 7 in ECs

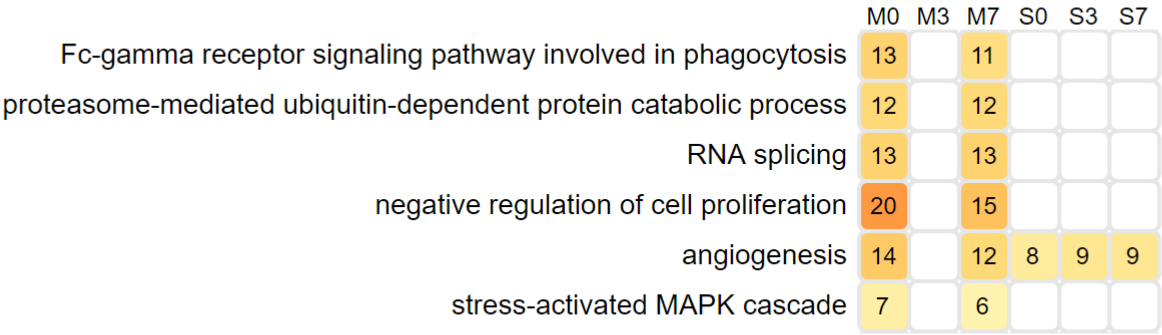
Fig 51 shows the BPs detected by MAGNETO in ECs and commonly enriched at days 0 and 3. MAGNETO detected 8 BPs commonly enriched at days 0 and 3, none of them was detected with ORA. Two BPs detected with MAGNETO were related to growth factor (epidermal growth factor receptor signalling pathway and cellular response to growth factor stimulus), and two to mitochondrial respiration.



	M0	M3	M7	S0	S3	S7
epidermal growth factor receptor signaling pathway	11	7				
cellular response to growth factor stimulus	7	7				
SMAD protein signal transduction	7	7				
positive regulation of canonical Wnt signaling pathway	9	9				
protein stabilization	11	9				
positive regulation of protein ubiquitination	7	7				
mitochondrial electron transport, NADH to ubiquinone	8	7				
mitochondrial respiratory chain complex I assembly	11	10				

Figure 51: BPs detected by MAGNETO commonly enriched at day 0 and 3 in ECs

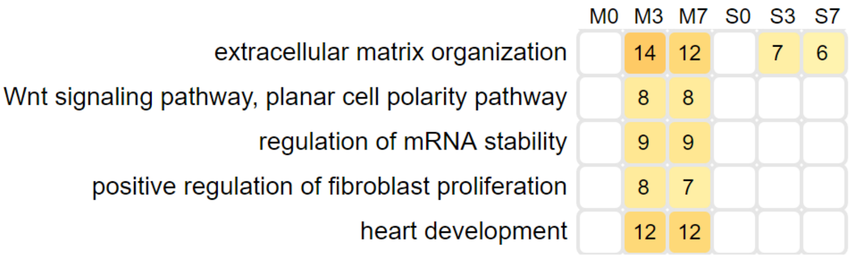
Fig 52 shows the BPs detected by MAGNETO and ORA in ECs and commonly enriched at day 0 and 7. MAGNETO detected 6 BPs commonly enriched at days 0 and 7, the BP “angiogenesis” was detected also by the standard ORA and also at day 3.



	M0	M3	M7	S0	S3	S7
Fc-gamma receptor signaling pathway involved in phagocytosis	13		11			
proteasome-mediated ubiquitin-dependent protein catabolic process	12		12			
RNA splicing	13		13			
negative regulation of cell proliferation	20		15			
angiogenesis	14		12	8	9	9
stress-activated MAPK cascade	7		6			

Figure 52: BPs detected by MAGNETO and ORA commonly enriched at day 0 and 7 in ECs

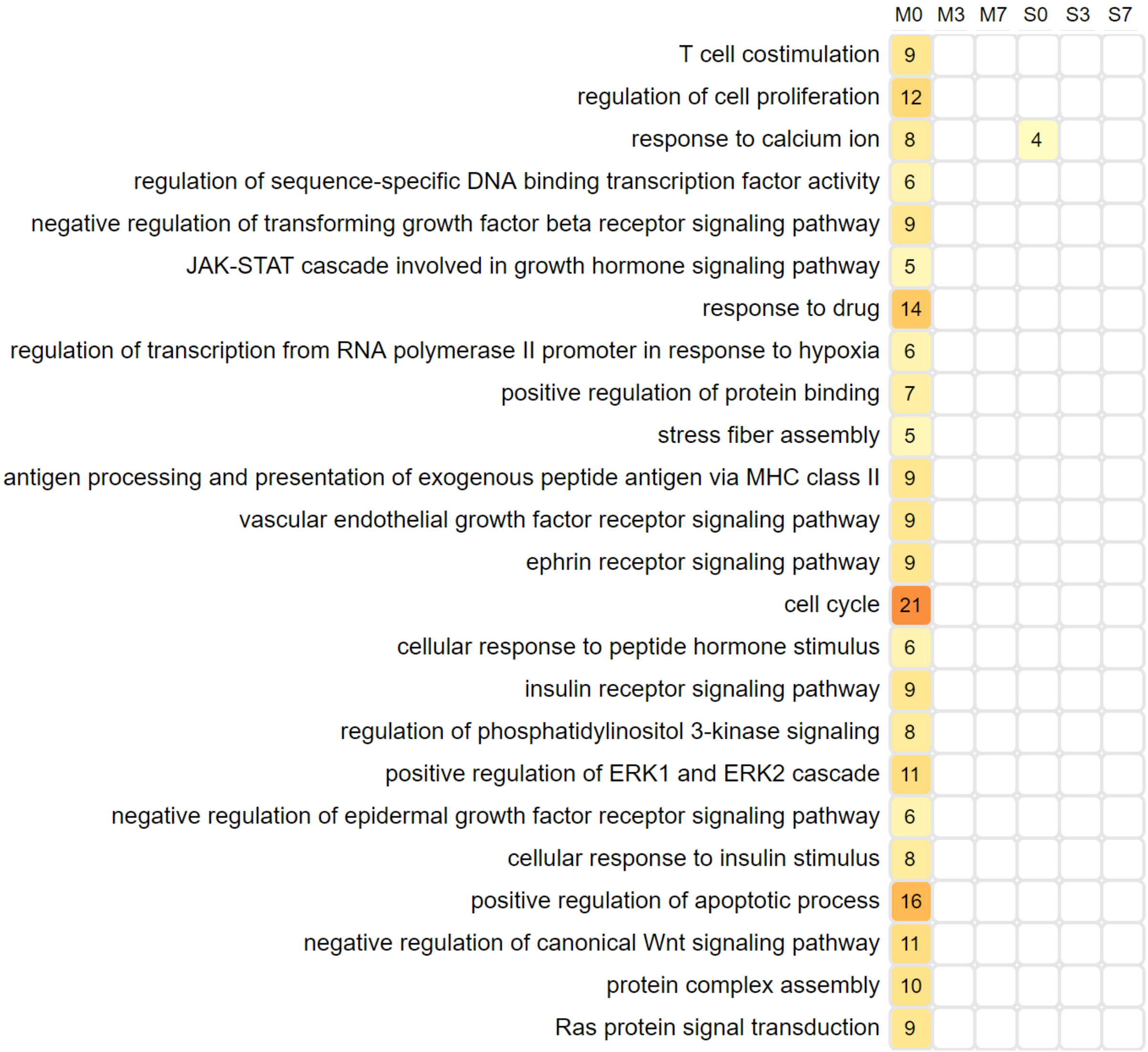
Fig 53 shows the BPs detected by MAGNETO in ECs and commonly enriched at day 3 and 7. MAGNETO detected 5 BPs commonly enriched at days 3 and 7, the BP “extracellular matrix organization” was detected also by the standard ORA at the same days. In comparison with ORA, MAGENTO detected two BPs related to cardiac remodelling (heart development and positive regulation of fibroblast proliferation).



	M0	M3	M7	S0	S3	S7
extracellular matrix organization		14	12		7	6
Wnt signaling pathway, planar cell polarity pathway		8	8			
regulation of mRNA stability		9	9			
positive regulation of fibroblast proliferation		8	7			
heart development		12	12			

Figure 53: BPs detected by MAGNETO and ORA commonly enriched at day 3 and 7 in ECs

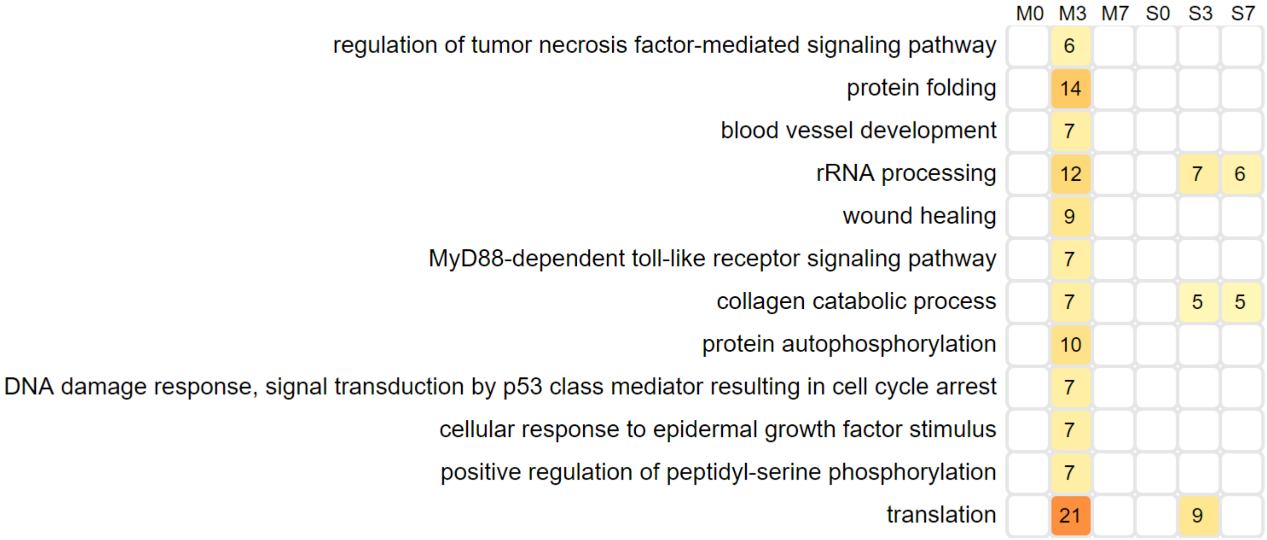
Fig 54 shows the BPs detected by MAGNETO and ORA in ECs and enriched at day 0. MAGNETO detected 24 enriched BPs in ECs at day 0. Of these, only the BP “response to calcium ion” was detected with the standard ORA. Additionally, MAGNETO detected BPs related to signalling pathway (negative regulation of transforming growth factor beta receptor signaling pathway, JAK-STAT cascade involved in growth hormone signaling pathway), and angiogenesis.



	M0	M3	M7	S0	S3	S7
T cell costimulation	9					
regulation of cell proliferation	12					
response to calcium ion	8			4		
regulation of sequence-specific DNA binding transcription factor activity	6					
negative regulation of transforming growth factor beta receptor signaling pathway	9					
JAK-STAT cascade involved in growth hormone signaling pathway	5					
response to drug	14					
regulation of transcription from RNA polymerase II promoter in response to hypoxia	6					
positive regulation of protein binding	7					
stress fiber assembly	5					
antigen processing and presentation of exogenous peptide antigen via MHC class II	9					
vascular endothelial growth factor receptor signaling pathway	9					
ephrin receptor signaling pathway	9					
cell cycle	21					
cellular response to peptide hormone stimulus	6					
insulin receptor signaling pathway	9					
regulation of phosphatidylinositol 3-kinase signaling	8					
positive regulation of ERK1 and ERK2 cascade	11					
negative regulation of epidermal growth factor receptor signaling pathway	6					
cellular response to insulin stimulus	8					
positive regulation of apoptotic process	16					
negative regulation of canonical Wnt signaling pathway	11					
protein complex assembly	10					
Ras protein signal transduction	9					

Figure 54: BPs detected by MAGNETO and ORA and enriched at day 0 in ECs

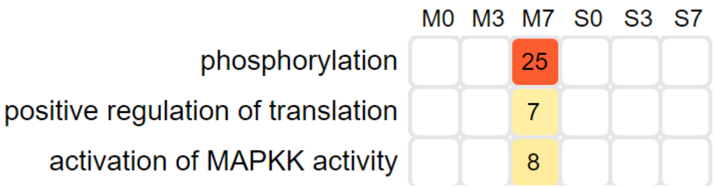
Fig 55 shows the BPs detected by MAGNETO in ECs and enriched at day 3. MAGNETO detected 12 enriched BPs in ECs at day 3. Of these, three of them were enriched also with the standard ORA, two of them (rRNA processing, collagen catabolic process) were enriched also at day 7. MAGNETO additionally detected BPs related to signalling (regulation of tumor necrosis factor-mediated signalling pathway, MyD88-dependent toll-like receptor signaling pathway), healing (wound healing and blood vessel development) and the collagen catabolic process.



	M0	M3	M7	S0	S3	S7
regulation of tumor necrosis factor-mediated signaling pathway		6				
protein folding		14				
blood vessel development		7				
rRNA processing		12			7	6
wound healing		9				
MyD88-dependent toll-like receptor signaling pathway		7				
collagen catabolic process		7			5	5
protein autophosphorylation		10				
DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest		7				
cellular response to epidermal growth factor stimulus		7				
positive regulation of peptidyl-serine phosphorylation		7				
translation		21			9	

Figure 55: BPs detected by MAGNETO and ORA commonly enriched at day 3 in ECs

Fig 56 shows the BPs detected by MAGNETO in ECs and commonly enriched at day 7. MAGNETO detected 3 BPs enriched in ECs at day 7, none of them were detected by the standard ORA.

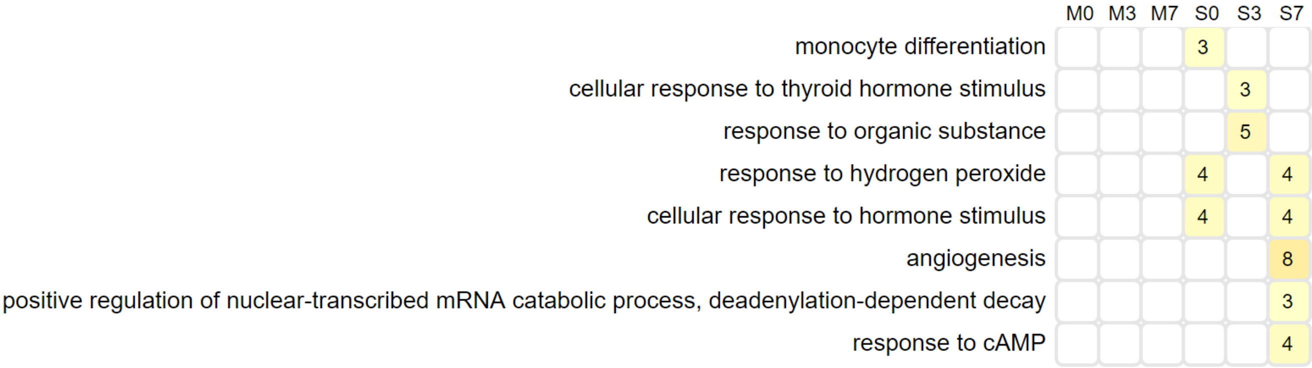


	M0	M3	M7	S0	S3	S7
phosphorylation			25			
positive regulation of translation			7			
activation of MAPKK activity			8			

Figure 56: BPs detected by MAGNETO and enriched at day 7 in ECs

Enriched processes in Macrophages

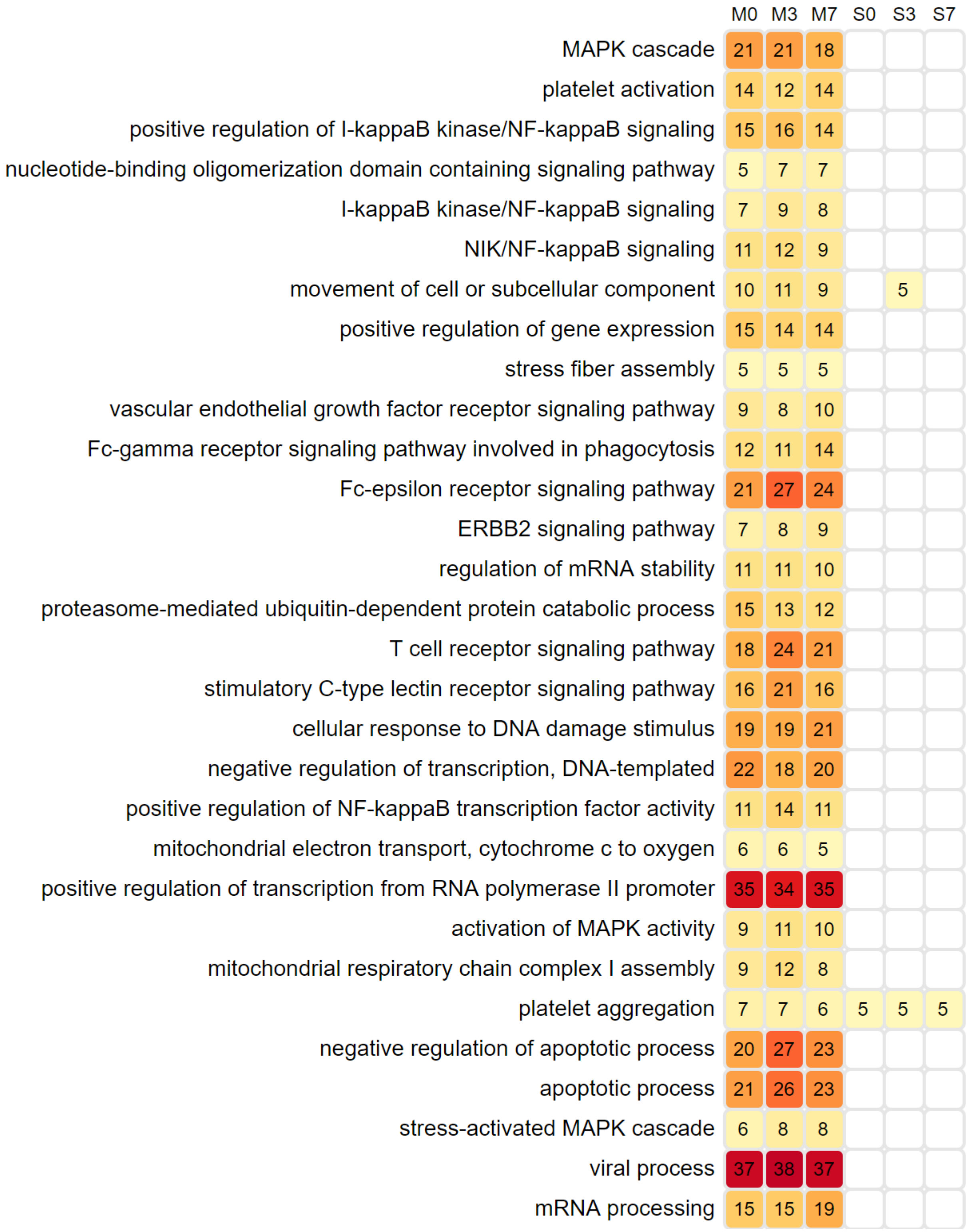
Without any regards for the day, ORA detected 8 BPs in MFs. These BPs were not enriched with MAGNETO.



	M0	M3	M7	S0	S3	S7
monocyte differentiation				3		
cellular response to thyroid hormone stimulus					3	
response to organic substance					5	
response to hydrogen peroxide				4		4
cellular response to hormone stimulus				4		4
angiogenesis						8
positive regulation of nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay						3
response to cAMP						4

Figure 57: BPs detected only with the standard ORA but not detected by MAGNETO in MFs

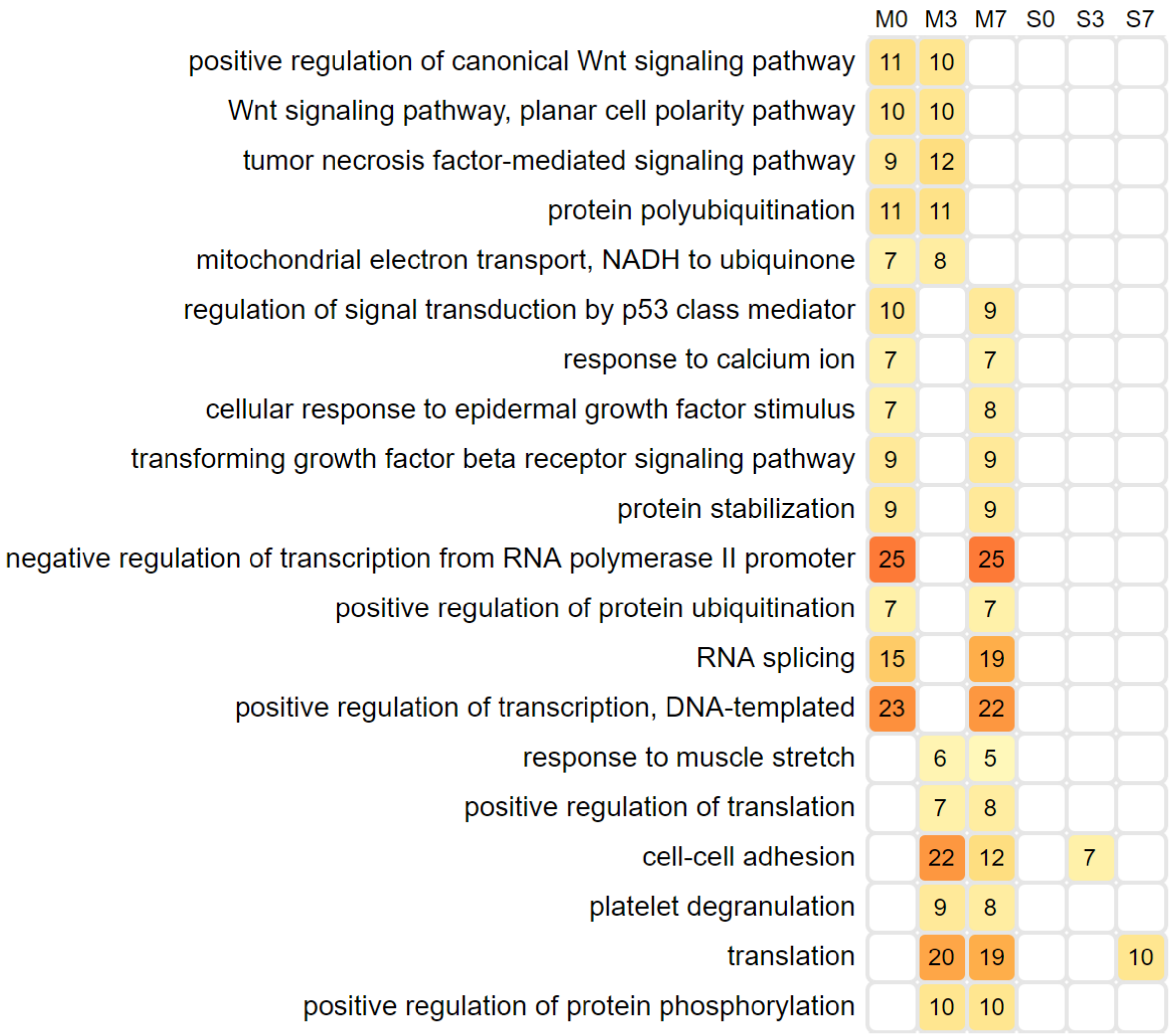
Fig 58 shows the BPs detected by MAGNETO in MFs and commonly enriched at day 0, 3, and 7. MAGNETO detected 30 BPs commonly enriched in all the days in MFs, the standard ORA shared 2 enriched BPs but one of them (movement of cell or subcellular component) was enriched only at day 3. Like in all the other cell types, also in MFs, most of the BPs detected with MAGNETO were related to immune response. Most of the proteins involved in BPs related to the immune response and apoptosis, (Fc-epsilon receptor signalling pathway, T cell receptor signalling pathway, stimulatory C-type lectin receptor signalling pathway, negative regulation of apoptotic process, apoptotic process) were increased at day 3 and slightly decreased at day 7.



	M0	M3	M7	S0	S3	S7
MAPK cascade	21	21	18			
platelet activation	14	12	14			
positive regulation of I-kappaB kinase/NF-kappaB signaling	15	16	14			
nucleotide-binding oligomerization domain containing signaling pathway	5	7	7			
I-kappaB kinase/NF-kappaB signaling	7	9	8			
NIK/NF-kappaB signaling	11	12	9			
movement of cell or subcellular component	10	11	9		5	
positive regulation of gene expression	15	14	14			
stress fiber assembly	5	5	5			
vascular endothelial growth factor receptor signaling pathway	9	8	10			
Fc-gamma receptor signaling pathway involved in phagocytosis	12	11	14			
Fc-epsilon receptor signaling pathway	21	27	24			
ERBB2 signaling pathway	7	8	9			
regulation of mRNA stability	11	11	10			
proteasome-mediated ubiquitin-dependent protein catabolic process	15	13	12			
T cell receptor signaling pathway	18	24	21			
stimulatory C-type lectin receptor signaling pathway	16	21	16			
cellular response to DNA damage stimulus	19	19	21			
negative regulation of transcription, DNA-templated	22	18	20			
positive regulation of NF-kappaB transcription factor activity	11	14	11			
mitochondrial electron transport, cytochrome c to oxygen	6	6	5			
positive regulation of transcription from RNA polymerase II promoter	35	34	35			
activation of MAPK activity	9	11	10			
mitochondrial respiratory chain complex I assembly	9	12	8			
platelet aggregation	7	7	6	5	5	5
negative regulation of apoptotic process	20	27	23			
apoptotic process	21	26	23			
stress-activated MAPK cascade	6	8	8			
viral process	37	38	37			
mRNA processing	15	15	19			

Figure 58: BPs detected by MAGNETO and ORA and enriched at day 0, 3 and 7 in MFs

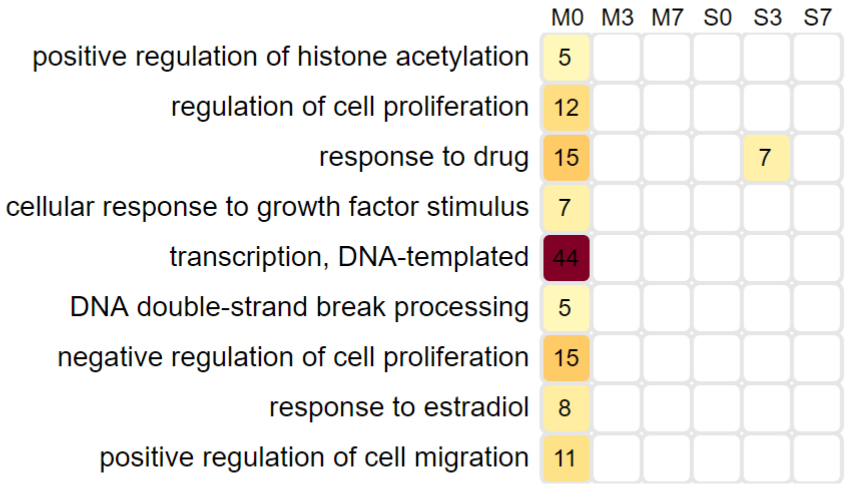
Fig 59 shows the BPs detected by MAGNETO in MFs and commonly enriched at day 0 and 3, at day 0 and 7, and at day 3 and 7. MAGNETO detected 5 BPs commonly enriched at day 0 and 3. Two of them were involved in the Wnt signalling pathway, one in the tumor necrosis factor signalling pathway. MAGNETO detected 9 BPs commonly enriched between day 0 and 7. MAGNETO detected 6 BPs commonly enriched at day 3 and 7, of these, the BP "cell-cell adhesion" showed a discrepancy in the number of proteins involved at day 3 with respect to day 7. The BP “translation” was also detected but in a different day.



	M0	M3	M7	S0	S3	S7
positive regulation of canonical Wnt signaling pathway	11	10				
Wnt signaling pathway, planar cell polarity pathway	10	10				
tumor necrosis factor-mediated signaling pathway	9	12				
protein polyubiquitination	11	11				
mitochondrial electron transport, NADH to ubiquinone	7	8				
regulation of signal transduction by p53 class mediator	10		9			
response to calcium ion	7		7			
cellular response to epidermal growth factor stimulus	7		8			
transforming growth factor beta receptor signaling pathway	9		9			
protein stabilization	9		9			
negative regulation of transcription from RNA polymerase II promoter	25		25			
positive regulation of protein ubiquitination	7		7			
RNA splicing	15		19			
positive regulation of transcription, DNA-templated	23		22			
response to muscle stretch		6	5			
positive regulation of translation		7	8			
cell-cell adhesion		22	12		7	
platelet degranulation		9	8			
translation		20	19			10
positive regulation of protein phosphorylation		10	10			

Figure 59: BPs detected by MAGNETO and ORA and commonly enriched at day 0 and 3, day 0 and 7, day 3 and 7 in MFs

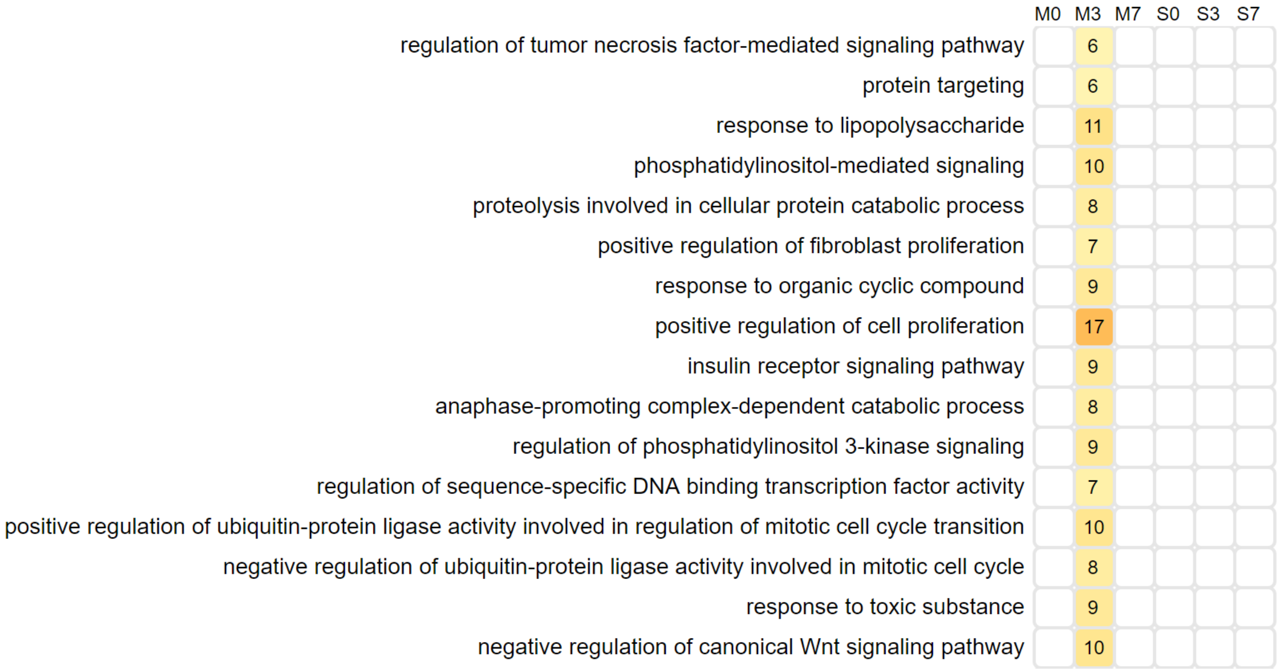
Fig 60 shows the BPs detected by MAGNETO and ORA in MFs and enriched at day 0. MAGNETO detected 9 BPs enriched in MFs at day 0, of these only the BP “response to drug” was enriched with the standard ORA. Additionally, MAGNETO detected BPs related to cell proliferation and migration.



	M0	M3	M7	S0	S3	S7
positive regulation of histone acetylation	5					
regulation of cell proliferation	12					
response to drug	15				7	
cellular response to growth factor stimulus	7					
transcription, DNA-templated	44					
DNA double-strand break processing	5					
negative regulation of cell proliferation	15					
response to estradiol	8					
positive regulation of cell migration	11					

Figure 60: BPs detected by MAGNETO and ORA and enriched at day 0 in MFs

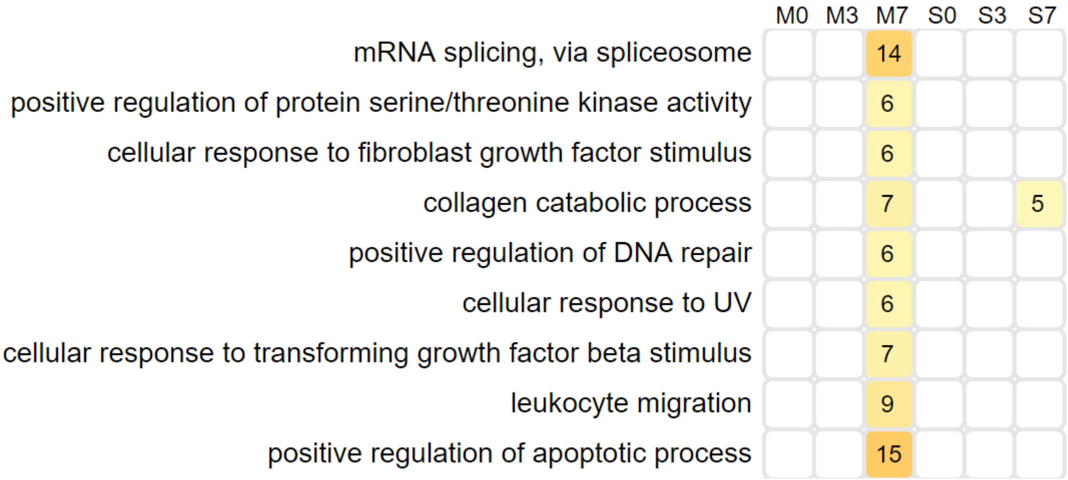
Fig 61 shows the BPs detected by MAGNETO in MFs and commonly enriched at day 3. MAGNETO detected 16 BPs enriched in MFs at day 3, none of them was detected with the standard ORA. The BPs detected by MAGNETO were mainly involved in the immune response (regulation of tumor necrosis factor-mediated signalling pathway, response to lipopolysaccharide, insulin receptor signalling pathway) and cell proliferation.



	M0	M3	M7	S0	S3	S7
regulation of tumor necrosis factor-mediated signaling pathway		6				
protein targeting		6				
response to lipopolysaccharide		11				
phosphatidylinositol-mediated signaling		10				
proteolysis involved in cellular protein catabolic process		8				
positive regulation of fibroblast proliferation		7				
response to organic cyclic compound		9				
positive regulation of cell proliferation		17				
insulin receptor signaling pathway		9				
anaphase-promoting complex-dependent catabolic process		8				
regulation of phosphatidylinositol 3-kinase signaling		9				
regulation of sequence-specific DNA binding transcription factor activity		7				
positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycle transition		10				
negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle		8				
response to toxic substance		9				
negative regulation of canonical Wnt signaling pathway		10				

Figure 61: BPs detected by MAGNETO and enriched at day 3 in MFs

Fig 62 shows the BPs detected by MAGNETO and ORA in MFs and commonly enriched at day 7. MAGNETO detected 9 BPs enriched in MFs at day 7, of these only the "collagen catabolic process" was commonly enriched with the standard ORA. Two BPs detected by MAGNETO were related to the cellular response to growth factor (cellular response to fibroblast growth factor stimulus and cellular response to transforming growth factor beta stimulus) and leukocyte migration.



	M0	M3	M7	S0	S3	S7
mRNA splicing, via spliceosome			14			
positive regulation of protein serine/threonine kinase activity			6			
cellular response to fibroblast growth factor stimulus			6			
collagen catabolic process			7			5
positive regulation of DNA repair			6			
cellular response to UV			6			
cellular response to transforming growth factor beta stimulus			7			
leukocyte migration			9			
positive regulation of apoptotic process			15			

Figure 62: BPs detected by MAGNETO and ORA and enriched at day 7 in MFs

Discussion

The work presented in this thesis can be divided into three main part. In the first part, we characterized the RBPs and the associated motifs by developing ATtRACT. In the second part, we detected the AS events after MI and thanks to the data available in ATtRACT we determined the splicing factors and the motifs that potentially regulate AS after MI. In the third part, we developed MAGNETO that permitted to depict a global picture of the BPs occurring before and after MI in each cell type. In this respect, MAGNETO, is an invaluable resource to gain more insight into the molecular processes and to discover new ones.

ATtRACT: - A daTabase of RNA binding proteins and AssoCiated moTifs

RBPs play a fundamental role in almost every cellular process. The complex and orchestrated interplay between RBPs and RNA controls and regulates multiple steps of RNA metabolism such as alternative splicing, polyadenylation, mRNA export, translation, stability and degradation (Glisovic et al. 2008). Knowledge the RBPs specificity and affinity with the associated binding motifs is essential to understand the transcriptional, post-transcriptional and the regulatory mechanisms in which RBPs are involved. However, the available information on RBPs and their motifs is currently limited, incomplete, and sometimes outdated. Despite advances in recent years, the lack of a unified, updated, and non-redundant repository that collects the information on RBPs and their associated binding sites not only delays the study of RBP function itself but also precludes the study of RNA processing, localization, and regulation in a global manner.

To address this gap, we developed ATtRACT: A database of RNA-binding proteins and associated motifs. ATtRACT was developed to be a 'one-stop shop' for the researchers that are involved in the study of RBPs.

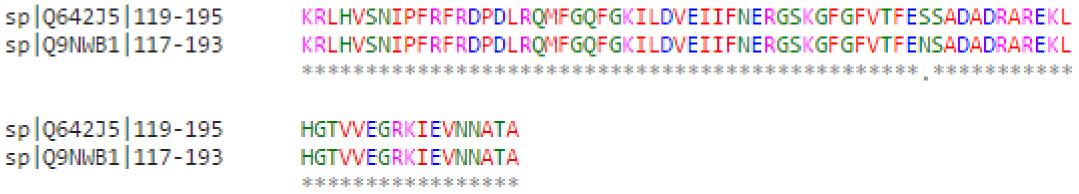
ATtRACT represents a unique resource containing manually curated and experimentally validated RNA binding proteins and their associated motifs. In comparison to other similar databases, ATtRACT adds 192 motifs not available in any other database from 110 different RBPs by retrieving the information buried in the PDB database. To our knowledge, ATtRACT is the largest and most updated collection of experimentally validated RBPs and associated binding sites. For this reason, it represents an invaluable resource to improve our understanding of RBP-RNA interactions and how they are regulated.

In comparison with the available databases, only CISBP-RNA includes more motifs than ATtRACT. This is because CISBP-RNA includes also motifs inferred by homology. The strategy adopted by

CISBP-RNA is to align the RBDs of different species. If two RBDs show a sequence similarity greater than 70%, it is likely that they share similar RNA binding sites. In ATtRACT, the inferred motifs are not included because only experimentally validated motifs are contained. Furthermore, a sequence similarity greater than 70% is not a condition sufficient to determine that the binding site will be identical also in other species. For example: RBFOX1 in humans binds to “UGCAUG” while RBFOX1 in zebrafish binds to “GCAUG”, even if the domains are almost 100% identical (Fig 63). In Auweter et al. 2006 the authors claim that:

1. U_1, G_2 and C_3 are necessary to form a hydrophobic cage around phenylalanine.

- U_1 and C_3 form one hydrogen bond
- The substitution of U_1 by either A or C leads to a loss of free binding energy (ΔΔG) of 4.0 and 4.5 kJ/mol, respectively.



```
sp|Q642J5|119-195      KRLHVSNIPFRFRDPDLRQMFGQFGKILDVEIIFNERGSKGFGFVTFESSADADRAREKL
sp|Q9NWB1|117-193      KRLHVSNIPFRFRDPDLRQMFGQFGKILDVEIIFNERGSKGFGFVTFENSADADRAREKL
                       ************************************************.***********

sp|Q642J5|119-195      HGTVVEGRKIEVNNATA
sp|Q9NWB1|117-193      HGTVVEGRKIEVNNATA
                       *****************
```

Figure 63: Alignment between the RBFOX1 domains in zebrafish (up) and in human (down**).** Identical and similar amino acids are indicated with stars and dots, respectively. The colours represent the residues according to their physicochemical properties (Red : Small (small+ hydrophobic (incl.aromatic -Y)), Blue : Acidic, Magenta: Basic - H, Green : Hydroxyl + sulfhydryl + amine + G, Grey: Unusual amino/imino acids etc)

The binding specificity of “UGCAUG” for RBFOX1 in humans is also confirmed by other papers (Wang & Burge 2008; Yeo et al. 2007; Castle et al. 2008; Ponthier et al. 2006). For this reason, we decided to leave out the inferred binding motifs.

ATtRACT was designed to be highly integrated with other databases and tools. To integrate other domains of knowledge, the other available databases adopt a cross-referencing technique forcing the users to traverse from a database to another to get the desired information. In ATtRACT, the information and tools are integrated into the database. For example, the GO database is integrated and users can explore the GO terms in which the RBPs are involved in. Additionally, to our knowledge, an RBP database integrating a tool to identify *de novo* motifs in a set of sequences was not available. The users were obliged to upload their files from one application to another. Moreover, to our knowledge, in Tomtom, the only available dataset of motifs is derived from CISBP-RNA, which we believe is not sufficient alone to capture the variability of RNA binding sites (please see Table 6).

ATtRACT was also designed to be highly standardized. In this respect, the inconsistency in nomenclature remains a major issue and a source of confusion: i) the investigators lose time in converting alternative gene name into the official one, ii) the nomenclature in a database must not be ambiguous by definition iii) the entries of a database are usually the starting points for further analyses (i.e. the intersection between two or more sources of information are partially captured only because the gene names are different) , iv) life sciences are interdisciplinary fields whose actors are not only biologists but also computer scientists, bioinformaticians and statisticians whose knowledge of the alternative gene names could be limited v) the search in PubMed can lead to incomplete results because the genes are annotated with different names. For all these reasons, in ATtRACT, we renamed the gene names according to the UniProt standard.

Even though with ATtRACT, many databases have been integrated, more work needs to be done to better characterize the RBPs and their binding sites and more data needs to be integrated. For example, a huge source of information, neglected in ATtRACT and in all the other databases, is the recently published POSTAR database (Hu et al. 2016). POSTAR collects in a unique resource a large amount of CLIP-seq experiments (including HITS-CLIP, PAR-CLIP and iCLIP). We are planning to integrate these large amounts of data in the next release of ATtRACT 2.0. Additionally, ATtRACT, thanks to the presence of experimentally validated data, could be a precious resource for developing new and more accurate machine learning methods to predict the RBP binding sites. Recently, new machine learning methods have been developed to predict the sequence specificities of RNA-binding proteins. Among them DeepBind (Alipanahi et al. 2015) is considered one of the best. DeepBind employs a convolutional neural network and predicts RNA binding sequences of 207 distinct RBPs. DeepBind was trained only with data available in CISBP-RNA, neglecting the information available in other databases, RNA secondary structure and the physicochemical properties of the amino acids in the RBD. To tackle all the aforementioned shortcomings, we plan to take advantage of the large amount of data available in ATtRACT and implement a “deep learning” method capable of predict RBP-RNA interactions better than the other state-of-the-art methods.

The -omics data

Cardiovascular diseases, including diseases of the heart and the blood vessels such as myocardial infarction, cardiac arrhythmias, and stroke, are the world's top cause of death and cause one-third death globally. Genome-wide and proteomics profile are nowadays necessary to fully understand the molecular mechanisms that regulate the response to myocardial infarction in a global manner. The exploration of these mechanisms has considerable limitations in humans. In this regard, animal models of myocardial infarction play a critical role in biomedical research. However, the evolutionary gap between humans and small animals like rats or mice is a strong disadvantage for translational research. In this respect, the pig model for myocardial infarction gained massively in importance during the last years mainly because of many anatomical and physiological similarities with the human heart (Swindle et al. 2012). To our knowledge, this is the first study that successfully isolated CMs, FBs, ECs, and MFs from the pig infarcted area and that used transcriptomics and proteomics analyses on these cells in order to unveil the molecular mechanisms that regulate the response to infarction.

Alternative splicing after MI

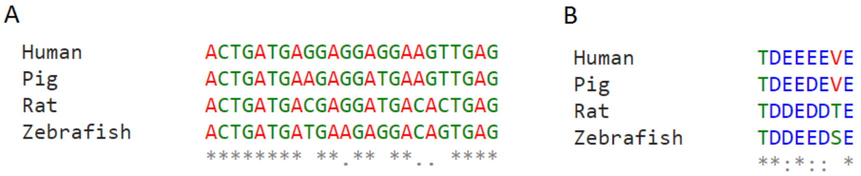
AS is one of the main mechanism responsible for giving rise to the final set of transcripts present in a cell. Even though our knowledge of the molecular mechanisms that underlie post-infarction heart remodelling has increased over years, it remains still incomplete and mainly relates to gene expression data and does not consider the effect of AS changes after the myocardial infarction. A query in PubMed database with the words: “alternative splicing” and “myocardial infarction” produces only 41 manuscripts, meaning that very little is known about alternative splicing and its relationship with heart diseases. The current research regarding the role of AS in the context of heart failure is limited to the study of individual gene isoforms (Kong et al. 2010; Qiu et al. 2008; Woolard et al. 2009; Felkin et al. 2011). In this context, RNA-Seq analysis has greatly facilitated the study of alternative splicing, but the algorithms to detect the isoform switch remain a challenge for the computational biologists. The lack of guidelines and common practice in combination with the proliferation of several algorithms for the detection of AS events make it more difficult to establish an effective solution. We employed a pipeline that allowed us to detect, on a global scale, pairs of isoform switching transcripts after MI in 4 cell types at 3 different time points. We identified 210 distinct genes undergoing AS in all the conditions, 41 of them were also detected at the protein level. It is important to notice that proteins must be expressed in a sufficient quantity to be identified. Hence the undetectability, at the protein level, of an alternatively spliced transcript does not mean that it is not translated.

The main challenge concerned the interpretation of the impact and functions of alternatively spliced genes. An ORA of AS genes did not provide meaningful information suggesting that AS genes regulate different processes but do not participate in a common process. The interpretation of the results remains a big issue. In this case, we considered that a literature mining of the most interesting and well annotated genes was the best approach even though time-consuming.

Alternative splicing in Cardiomyocytes

We detected 53 AS genes, 28 of them detected in CMs at day3v0, 35 at day7v0 and 17 at day7v3. Of these genes, 20 were in common between day3v0 and day7v0 suggesting that the same genes continued to be alternatively spliced also on day 7 (Fig 28A). The enriched BPs in any time contrast 'positive regulation of interleukin-2 biosynthetic process', 'positive regulation of B cell differentiation' and 'B cell activation' suggest that AS in CMs may play a role in the activation of the immune response.

A literature analysis revealed that, in CMs, FHOD3, UXT and LRPPRC genes were the ones for which at least one article regarding their AS has been published. The FHOD3 gene was alternatively spliced in time contrast 7v3. The FHOD3 gene is mainly involved in the assembly and maintenance of cardiac myofibrils. Structurally, the gene is a member of the formin subfamily and contains multiple domains including the formin homology 2 domain (FH2). Iskratsch and collegues (Iskratsch et al. 2010; Iskratsch et al. 2013) identified in CMs a novel striated muscle–specific micro-exon in FHOD3 gene, which inserts as exon 26 in humans and as exon 25 in mice. This alternative exon is highly conserved between Zebrafish and Human. To assess whether the micro-exon is conserved also in pig, we performed a multiple sequence alignment of the nucleotide (Fig 64A) and amino acid sequences (Fig 64B).



```
Human        ACTGATGAGGAGGAGGAAGTTGAG
Pig          ACTGATGAAGAGGATGAAGTTGAG
Rat          ACTGATGACGAGGATGACACTGAG
Zebrafish    ACTGATGATGAAGAGGACAGTGAG
             ******** **.** **.. ****
```

```
Human        TDEEEEVE
Pig          TDEEDEVE
Rat          TDDEDDTE
Zebrafish    TDDEEDSE
             **:*:: *
```

Figure 64: Multiple sequence alignment of conserved sequences of alternative spliced micro-exon in four species (Human, Pig, Rat and, Zebrafish). Identical and similar amino acids are indicated with stars and dots, respectively. The colours of the amino acid alignment (Fig 64B) represent the residues according to their physicochemical properties (Red : Small (small+ hydrophobic (incl.aromatic -Y)), Blue : Acidic, Magenta: Basic - H, Green : Hydroxyl + sulfhydryl + amine + G, Grey: Unusual amino/imino acids etc) A) Multiple sequence alignment of nucleotide sequences. B) Multiple sequence alignment of translated region.

The alignment showed a very high similarity in both nucleotide and amino acid sequences. The micro-exon encodes for eight additional amino acids at the C-terminal end of the FH2 domain. According to Iskratsch and collegues, this insertion introduces a phosphorylation site in the FH2 domain which, if phosphorylated, alters its interaction with sequestosome 1 (SQSTM1). Thus, this new isoform appears to be required for the myofibril maintenance. The authors also investigated the role of FHOD3 in cardiomyopathy, including myocardial ischemia, and detected a downregulation of the isoform containing the exon with a concomitant up-regulation of the exon–lacking isoform. We observed an analogous behaviour in our data (Fig 65). Notice that we did not detect the isoform switch at time contrast 3v0 because our pipeline did not consider the expression of the isoform lacking the exon relevant, since it did not exceed the threshold of 1 TPM.

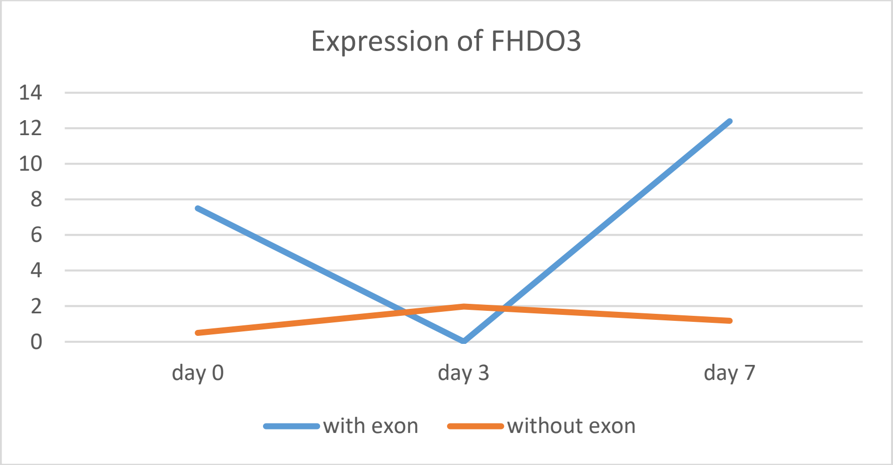


Figure 65: The expression profile of the FHOD3 gene is shown. The x-axis represents the time points, the y-axis represents the expression in terms of TPMs. A downregulation of isoforms containing the exon (blue line) with a concomitant up-regulation of the exon–lacking isoforms (orange line) can be observed.

It has been proposed by Larochelle 2016; X. Yang et al. 2016 that alternative splicing can affect the way in which the proteins interact each other. Alternatively spliced isoforms tend to behave like distinct proteins rather than minor variants. In this context, two alternatively spliced genes, UXT and LRPPRC, the last one detected also at proteins level, interact with each other. According to Moss et al. 2007, at higher expression levels, UXT interacts with LRPPRC leading to a progressive aggregation of mitochondria and cell death. AS of one of the genes can potentially disrupt this interaction and may lead to a pro-survival phenotype in CMs.

Alternative splicing in Fibroblasts

We detected 53 AS genes in FBs, 27 of them were detected at day3v0, 32 at day7v0 and 24 at day7v3. The enriched BPs did not provide any meaningful information about the role of alternative splicing in FBs. A literature mining did not show any relevant paper associated to the AS genes. An analysis of the BPs revealed that the HSD17B12 and ADAMTS12 genes are involved in the extracellular matrix organisation and cell-matrix adhesion respectively. ADAM8 is involved in extracellular matrix degradation (Choi et al. 2001). The role of AS in FBs requires further investigations.

Alternative splicing in Endothelial Cells

We detected 86 AS genes in ECs, 41 of them were detected at day3v0, 46 at day7v0 and 43 at day7v3. The enrichment analysis of the ECs at time contrast 7v0 showed that 5 genes (CARMIL2, PDZD11, NPHP1, NCAM1, SGCA) were localised in cell-cell junction suggesting that these genes may play a role in maintaining the integrity of the endothelium and in leukocyte extravasation.

A literature analysis revealed that, in ECs, MST1R, FST and SLBP genes were the ones for which there was information available about their alternative splicing. MST1R (also known as RON) gene is a tyrosine kinase receptor for the Macrophage-stimulating protein. It is involved in cell growth and protection from apoptosis, cell dissociation, motility, and matrix invasion. ΔRON is an active isoform generated through the skipping of exon 11. This skipping provokes a change in the protein structure, as a consequence of which the ΔRon is always active also in the absence of its ligand, permitting to increase the cell motility (Collesi et al. 1996; Ghigna et al. 2005).

One of the main roles of Follistatin (FST) is the inhibition of the myostatin activity. Myostatin is a transforming growth factor-β family member that negatively regulates skeletal muscle growth. Follistatin (FST) is alternatively spliced at time contrast 7v3 (Fig 66). The FST gene generates two isoforms, a long and a short one. The short isoform is much more efficient in the inhibition of the myostatin pathway compared to the long one (Hashimoto et al. 1997; Lee & McPherron 2001).

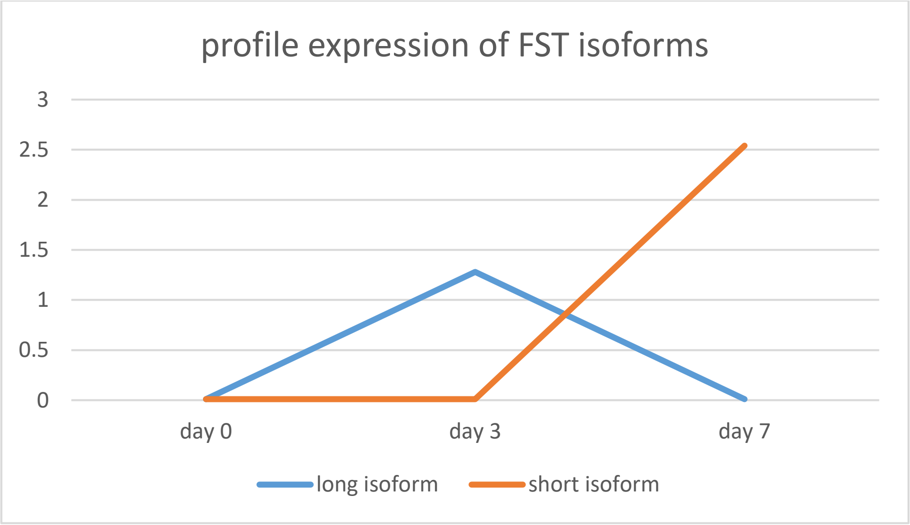
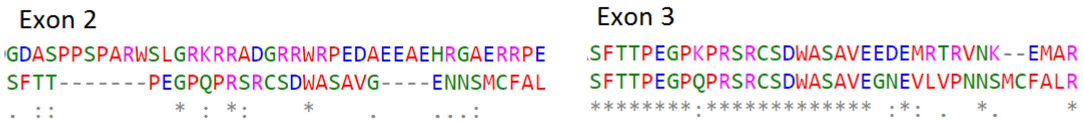


Figure 66: The expression profile of the FST gene isoforms. The x-axis represents the days, the y-axis represents the expression in terms of TPMs. The isoform switch occurred at the time contrast 7v3. The short isoform (orange line) increased its expression at day 7 suggesting a potential role in promoting muscle growth

This suggests that the short isoform of FST gene may play a role in promoting muscle growth by inhibiting the myostatin pathway. Additionally, it has been shown that adeno-associated virus-mediated delivery of FST allows an increased muscle size and strength with reduced fibrosis in a muscular dystrophy mouse model (Rodino-Klapac et al. 2009). This suggests a possible gene therapy treatment for MI. Interestingly, the Activin Type II receptor (ACVR2B) gene is alternatively spliced in FBs at the same time contrast. It has been shown that ACVR2B competes with FST for the binding of Myostatin (Donaldson et al. 1999) suggesting an additional way to regulate the myostatin pathway through AS .

The RNA-binding protein SLBP regulates the expression of all histone genes. SLBP is essential for histone mRNA 3′ end formation and is involved in the translation of histone mRNA. It also plays a role in the control of histone mRNA stability (Rattray & Muller 2012; Marzluff et al. 2008). The human SLBP gene generates, under replication stress conditions, two alternatively spliced variants lacking exon 3 and exon 2, respectively. According to Rattray et al. 2013, the isoform lacking exon 2 is localised in the nucleus and the cytoplasm, making the SLBP gene partially unable to participate in histone mRNA regulation, whereas the isoform lacking exon 3 is localised in the nucleus. The same authors suggest that exon 3 is important for the binding of proteins involved in translation initiation, suggesting that alternative splicing of the SLBP gene is a way to control the histone gene expression. To assess whether these findings were valid also in pig, we performed a sequence alignment of the coding sequences of exon 2 and exon 3 between the pig gene and the human ortholog (Fig 67).



```
Exon 2
GDASPPSPARWSLGRKRRADGRRWRPEDAEEAEHRGAERRPE
SFTT-------PEGPQPRSRCSDWASAVG----ENNSMCFAL
. ::         * : *:    *    .    ...:

Exon 3
SFTTPEGPKPRSRCSDWASAVEEDEMRTRVNK--EMAR
SFTTPEGPQPRSRCSDWASAVEGNEVLVPNNSMCFALR
********:************* :*: .  *.     *
```

Figure 67: Alignment of the coding region of Exon 2 and Exon 3 of the SLBP gene. Identical and similar amino acids are indicated with stars and dots, respectively. The colours represent the residues according to their physicochemical properties (Red : Small (small+ hydrophobic (incl.aromatic -Y)), Blue : Acidic, Magenta: Basic - H, Green : Hydroxyl + sulfhydryl + amine + G, Grey: Unusual amino/imino acids etc).

The alignment showed a high similarity for exon 3, but a poor similarity for exon 2, suggesting that the findings might not be valid also in pig.

Alternative splicing in Macrophages

We detected 79 AS genes in MFs, 37 of them were detected at day3v0, 48 at day7v0 and 34 at day7v3. The standard ORA did not show any enriched GO term in any of the three domains of gene ontology.

A literature analysis revealed that in MFs, AGER, and MUC1 were the main genes for which at least an article regarding AS had been published. The Receptor for Advanced Glycation End-products AGER (also known as RAGE) is a transmembrane protein expressed in the heart in CMs, ECs, FBs, and MFs (Ramasamy & Schmidt 2012). Lu et al. 2010 showed that RAGE and its ligands were upregulated in the heart during ischemia/reperfusion and that oxidative stress was enhanced, Tsoporis et al. 2012 showed that S100B, via AGER ligation, favours FBs proliferation contributing to scar formation in infarcted myocardium, additionally antagonizing AGER allowed to reduce the infarcted area and to improve cardiac function. The AGER mRNA was alternatively spliced in time contrast 7v0 and 7v3. Alternative splicing of AGER mRNA in the pig alters the immunoglobulin-like domain that is necessary for the binding of S100B ligand (Leclerc et al. 2007). These isoforms may therefore serve as a binding regulator for the S100B ligand (Hudson et al. 2008).

MUC1 mRNA was alternatively spliced in time contrast 7v0. MUC1 is a transmembrane protein whose main feature is the variable number of tandem repeats (VNTR) region forming the extracellular domain. The VNTR region can extend 300–500 nm above the cell surface playing a fundamental role in cell-cell and cell-matrix interactions (Fukuda 2002; Hilkens et al. 1992). Alternative splicing may regulate the extension of the VNTR region. In the pig, the VNTR is encoded by exon 1, we observed an expression of the long VNTR isoform mainly in day 0 and the short VNTR isoform in day 7. It has been suggested by Zhang et al. 2013 and Carson 2008 that the short isoform preserves important signalling functions or interfere with the long isoform by competing for ligand binding.

Analyses of the domains

Changing the sequence content of a gene product is one of the main characteristics of AS. In this respect, AS provides a new layer of complexity since it affects the protein structure and function. AS tends to alter the protein structure by inserting, deleting, and altering a proteins' domain. The most interesting effects of AS on the proteins' domain were found in the APOD and TNFRSF17 genes. AS of the APOD gene in CM at day 3v0 potentially changed the domain from a triabin domain to a lipocalin-like domain. For TNFRSF17, the BCMA, TALL-1 binding domain is inserted in CM at day 3v0 and deleted in MFs at the same time contrast. Recent studies suggested that APOD gene is highly upregulated during heart failure (Wei et al. 2008). Additionally, it has been shown that APOD has cardioprotective effects on CMs in vitro in mouse models of MI with ischemia/reperfusion (I/R) injury (Tsukamoto et al. 2013). The same authors performed an adenovirus-mediated administration of APOD gene and detected a significant decrease in the infarct size suggesting a potential gene therapy approach. The triabin domain structurally looks similar to the lipocalin domain, they differ only in the direction of β-strands and in the general conformation of the β-barrel. Triabin interacts with thrombin, the principal promoter of blood clotting and inhibits vasoconstriction, platelet aggregation and coagulation (Fuentes-Prior et al. 1997; Hernández-Vargas et al. 2016). The APOD containing the lipocalin domain can bind to a number of ligands which may be protective including, progesterone, and arachidonic acid but with low affinity (Rassart et al. 2000). Tsukamoto et al. 2013 suggested, in contrast, that APOD might be cardioprotective because it is an antioxidant.

TNFRSF17 is a member of the TNF-receptor superfamily and in combination with TNF-related cytokines promotes cell death or cell proliferation and differentiation (Smith et al. 1994). Overexpression of the TNFRSF17, like in other members of the family, activates the mitogen-activated protein kinase (MAPK) pathway and NF-κB, but, unlike the other members of the family, TNFRSF17 lacks of the "death domain", suggesting that TNFRSF17 is involved in cell survival and proliferation (Hatzoglou et al. 2000). Hence, the insertion of the BCMA, TALL-1 binding domain in TNFRSF17 gene indicates that AS could play a role in the signals involved in CMs survival. However, the effect of the deletion of the same domain at the same time contrast in MFs remains unclear.

Regulation of alternative splicing

The lack of standard computational methods in combination with the complex regulatory layer generated by the RBPs-RNA interaction makes it difficult to assess the biological relevance of our *in-silico* analysis, for this reason further experiments are needed. Here we developed a pipeline which makes use of MEME to find motifs overrepresented in the alternatively spliced exons and Tomtom to assess whether these motifs look like any of the ones available in ATtRACT. The main drawback of the pipeline is that it finds master regulators, meaning a set of RBPs regulating the majority alternatively spliced exons. This happens because the zoops model implemented in MEME assumes that each sequence may contain at most one occurrence of each motif. Therefore, to be overrepresented, a motif must be present in most of the sequences. Hence, it is not surprising that the overrepresented splicing factors hnRNPU and hnRNPD are the ones capable of binding RNA also non-specifically. This may be due to the fact that both splicing factors carry a RGG box region. The RGG box region confers the splicing factor the ability to bind to poly(A), poly(U), poly(G) and poly(C) with an intermediate affinity (Fackelmayer et al. 1994; Kiledjian & Dreyfuss 1992). However, the role of hnRNPU in the regulation of alternative splicing remains an open question. The group of Rappsilber et al. 2002 identified the hnRNPU as a core component of the spliceosome, the group of Zhou et al. 2002, on the contrary, did not find the protein. More recently has been suggested by Xiao et al. 2012 that hnRNPU is a global regulator of alternative splicing because regulates U2 snRNP maturation. Additionally, Ye et al. 2015 showed that hnRNPU is necessary for normal pre-mRNA splicing and postnatal heart development and function. These findings may be in agreement with previous results showing that developmental mechanisms are reactivated following injury (Bergmann et al. 2015; Senyo et al. 2013). Additional experiments are necessary to validate our *in-silico* analyses and the role of hnRNPD. Moreover, the computational methods for detecting the regulatory motifs need to be improved to gain a complete understanding of the alternative splicing regulatory mechanisms.

Network-based enrichment analysis with MAGNETO: performance evaluation

ORA is a common approach to assess whether there is a statistically significant overlap between a gene/protein set of interest and a database of known gene/protein sets representing cellular processes or pathways. Since proteins rarely act in isolation, the main limitation of the ORA method is that the molecular interaction network in a gene/protein set of interest is not considered. To solve this problem, here we developed a new network-based enrichment analysis method called MAGNETO.

MAGNETO is a data-driven method whose main goal is to gain more insight into the molecular mechanisms of the phenotype under investigation by combining the advantages of topology-based methods and modules-based methods. MAGNETO initially splits the starting list of proteins/genes in modules. Each module represents a general biological process in accordance with the Reactome classification. Subsequently, MAGNETO exploits the information contained within the protein interaction network and in the expression of the proteins at the tissue level to expand the input list and detects additional functional associations not directly inferable with the standard ORA. Typically, MAGNETO yields a big list of novel functional associations, most of which are coherent with the phenotype and the literature references. Others represent potentially new BPs associated with the phenotype that never were investigated before. Therefore, MAGNETO could be of great use to formulate new hypotheses associated with a pool of proteins related to a phenotype.

The main challenge, common to all the ORA methods, is to evaluate their performances. In MAGNETO we adopted the approach proposed by Tarca et al. 2012. The advantage of this method is that it eliminates any human bias in the interpretation of the results at the cost of having a dataset associated with a target pathway (for details please see the materials and methods section). A disadvantage of the target pathway method is that it focuses only on a single pathway and neglects those BPs and pathways that may play a fundamental role in the phenotype.

We evaluated MAGNETO on 21 real datasets. The evaluation was focused on the ability of MAGNETO to identify the target pathway as significant and with a rank as high as possible. MAGNETO outperformed the standard ORA when more than 50 genes/proteins were considered. It was also evident that MAGNETO allows to detect the target pathway with just a few proteins in the input, suggesting that it may be useful to investigate the biological impact of small lists of genes/proteins (i.e. less than 50). Additional benchmarks are needed to compare MAGNETO with other NBEA methods (i.e. Enrichnet, NET-GE, PINA) and gene set enrichment analysis methods (Subramanian et al. 2005).

Large scale data analysis of the pig infarcted myocardium

In a healthy myocardium, the most abundant cells are CMs, FBs, and ECs (Pinto et al. 2016). Subsequently to myocardial ischemia, these proportions are subject to variation and also MFs play a crucial role in the inflammation and healing of the myocardium (Frangogiannis 2014; Christia & Frangogiannis 2013). In this respect, ORA methods can provide an idea of the molecular mechanisms characterising the response to MI and to depict the role of each cell type.

The current interest in the pig as a large animal model has raised the problem of the limited pig gene and protein annotations, efforts to provide a complete annotation of the pig genome and protein functions are still ongoing (Groenen 2016). Frequently, the annotations are transferred via sequence similarity. It is commonly accepted that proteins sharing more than 60% of sequence identity are also likely to share similar functions. Many of the annotated pig proteins have been inferred via sequence similarity methods and lack any experimental evidence. For this reason, we employed the human annotations to perform the ORA and converted the pig genes to one to one orthologue human genes. To compare MAGNETO and the standard ORA method, we employed the top 50 expressed genes in each cell type and time point. A term was considered enriched if the p-value (Bonferroni corrected) was less than 0.001

Analysis of overrepresented biological processes in all the cell types

MAGNETO detected 178 unique BPs in total in all the cell types. In contrast, the standard ORA detected 48 unique BPs in total in all the cell types. It is important to note that MAGNETO increases the testing set at the risk that some BPs, overrepresented with the standard ORA, may not appear enriched in MAGNETO.

Cardiac repair after MI is an orchestrated event driven, mainly, by CMs, FBs, ECs and, MFs with sometimes overlapping roles (Talman & Ruskoaho 2016). Cardiac injury activates innate immune mechanisms initiating an inflammatory reaction. Therefore, it is not surprising that most of the enriched BPs following MI detected by MAGNETO and common in all the cell types and time points are related to the immune response and apoptosis (positive regulation of I-kappaB kinase/NF-kappaB signalling, I-kappaB kinase/NF-kappaB signalling, positive regulation of NF-kappaB transcription factor activity, NIK/NF-kappaB signalling, NF-kappaB signalling, stimulatory C-type lectin receptor signalling pathway, T cell receptor signalling pathway, apoptosis). Many of these BPs have been experimentally validated (Cai et al. 2016, Lin & Knowlton 2014, Ghigo et al. 2014), suggesting that the results obtained with MAGNETO are biologically relevant.

In CMs at day 0 most of the BPs, both in MAGNETO and in the standard ORA, are connected to muscle contraction and generation of energy. MAGNETO detected additional BPs associated with the inflammatory phase of MI (wound healing, platelet activation and platelet degranulation, Fc-gamma receptor signalling pathway involved in phagocytosis) and with the generation of the scar (extracellular matrix organization, collagen catabolic process, positive regulation of cell proliferation, positive regulation of fibroblast proliferation). CMs at day 7 are characterised by the enrichment of BPs related to muscle contraction and cardiac remodelling detected by both MAGNETO and the standard ORA approach, suggesting that proteins involved in muscle contraction are reappearing in the infarcted area after the acute phase. Additionally, other enriched BPs revealed signals involved in cardiac regeneration and scar formation (heart development, blood vessel development, positive regulation of epithelial to mesenchymal transition, ERBB2 signalling pathway). Some of these BPs have been experimentally validated (Zhou & Pu 2011; Ozcelik et al. 2002).

in FBs, the BPs enriched, both with MAGNETO and the standard ORA approach, in all the time points reflect the inflammatory and proliferative phases of the cell type with BPs typical of cardiac regeneration, extracellular matrix degradation and synthesis (extracellular matrix organization, extracellular matrix disassembly, collagen catabolic process, collagen fibril organization). Additionally, MAGNETO detected BPs at day 3 and/or at day 7 involved in cell migration (positive regulation of fibroblast migration, cell migration, regulation of cell migration, positive regulation of cell migration), extracellular matrix organization (extracellular fibril organization, cell-matrix adhesion, substrate adhesion-dependent cell spreading and positive regulation of stress fiber assembly) and recruitment of other cells employed as a source of fibroblasts (positive regulation of epithelial cell proliferation). This BP has been also experimentally validated (Hinz et al. 2007). Additionally, it has been shown that the epidermal growth factor induced heart fibrosis and proliferation of cardiac FBs (Lian et al. 2012) and the enriched BP “cellular response to epidermal growth factor stimulus” detected by MAGNETO seems to confirm this hypothesis. Finally, the BPs “transforming growth factor beta receptor signalling pathway” enriched at day 7 may regulate a wide range of cellular responses critical to cardiac repair. Also in FBs, we found several BPs experimentally validated, for a review of these BPs see Bujak & Frangogiannis 2007 .

In ECs, the BP “angiogenesis” is overrepresented in all the time points with the standard ORA, in MAGNETO the same process is enriched only in day 0 and 7 but not in day 3. Additionally, MAGNETO enriches BPs related to ECs at day 0 (vascular endothelial growth factor receptor signalling pathway, ephrin receptor signalling pathway). These two pathways have been also experimentally validated (Gale & Yancopoulos 1999). The BPs enriched at day 3 and/or 7 confirm the role of ECs in immune response and regeneration (cell-cell adhesion, leukocyte migration, wound healing, blood vessel development transforming growth factor beta receptor signalling pathway, SMAD protein signal transduction, regulation of tumor necrosis factor-mediated signalling pathway, MhD88-dependent toll-like receptor signalling pathway, platelet activation, heart development, extracellular matrix organization, collagen catabolic process). Many of these BPs have been experimentally validated (Muller 2003, Sumpio et al. 2002, Saini et al. 2005, Euler 2015, Y. Yang et al. 2016, Michiels 2003, Guarda et al. 1993)

Most of the BPs detected in MFs by MAGNETO are correlated with the immune response. Additionally, the BPs detected by MAGNETO are consistent with the inflammatory response and remodelling of the heart at day 3 (regulation of tumor necrosis factor-mediated signalling pathway, tumor necrosis factor-mediated signalling pathway, response to lipopolysaccharide, proteolysis involved in cellular protein catabolic process, insulin receptor signalling pathway, positive regulation of fibroblast proliferation, response to muscle stretch, cell-cell-adhesion, positive regulation of cell proliferation) and at day 7 (cellular response to fibroblast growth factor stimulus, cellular response to transforming growth factor beta stimulus, transforming growth factor beta receptor signalling pathway, leukocyte migration, collagen catabolic process, cell-cell-adhesion, response to muscle stretch). Some of these BPs have been experimentally validated (Chen & Frangogiannis 2016; Fujiu et al. 2014; Frangogiannis et al. 2002).

These few examples demonstrate that MAGNETO could be of great use to identify BPs and consequently also the proteins involved in a complex phenotype like the response to MI, for which the available knowledge is limited. Additionally, MAGNETO is of great use for the investigators to formulate new hypotheses, that need, eventually, to be tested in the laboratory.

MAGNETO employs a data-driven approach, meaning that as the annotations of GO, Reactome and HPA improve, also the assignments of proteins to the modules and the detection of BPs will improve. Moreover, in this study we only showed the gene ontology domain related to BPs, but, in MAGNETO, other databases are integrated to give additional layers of knowledge and to better evaluate the protein list of interest in the context of pathways, diseases, viruses, toxins, and drugs.

Finally, we believe that there is space for an improvement of MAGNETO's algorithm. For example, instead of using the shortest path to connect two seed nodes, a more effective way could be the employment of the diffusion state distance (DSD) metric that is able to quantify topological similarity in a PIN in a more fine-grained way (Cao et al. 2014).

In summary, MAGNETO represents a way of generating new knowledge from the available data, at no cost, allowing to propose new hypotheses and unveil new biological processes associated with a given phenotype.

Conclusions

1. ATtRACT collects in a unique resource all the experimentally validated RNA binding proteins and associated motifs that were previously spread across several databases. Compared to existing databases, ATtRACT adds 192 motifs not available in any other database by retrieving the information buried in the Protein Data Bank. Currently, ATtRACT is the largest and most updated collection of experimentally validated RNA binding proteins and associated binding sites.

2. In comparison with the heart of small animals, the pig heart shares many anatomical and physiological similarities with the human heart. However, the use of pig as animal model has raised the problem of its functional annotation. The use of another species annotations through the identification of the gene orthologs helps to interpret the results when the species of interest is poorly annotated.

3. Two hundred and ten genes were detected as alternatively spliced after myocardial infarction in cardiomyocytes, fibroblasts, endothelial cells, and macrophages. Forty-one were also detected at the protein level. These results indicate that either the technology sensitivity limits the number of alternative splicing events that can be detected in the infarcted heart or that very few AS events take place in response to a MI.

4. A functional enrichment analysis of the alternatively spliced genes following myocardial infarction retrieved only few enriched terms supported by only few genes, suggesting that the identified genes did not participate in a common biological process. A literature analysis of these genes suggested that the alternative Apolipoprotein D and Follistatin isoforms were the ones that may have the strongest biological impact after myocardial infarction.

5. The functional enrichment analysis is greatly improved when the protein interaction network is taken into account. In this respect, the benchmarks showed that MAGNETO outperforms the standard enrichment analysis from 50 proteins onwards. MAGNETO improves the ranks and the p-values of target pathways. Therefore, MAGNETO represents a precious resource to gain more biological knowledge also when a list of just a few genes/proteins is available.

6. Applied to the top 50 genes orthologs one to one to human expressed in the pig infarcted myocardium, MAGNETO provided more overrepresented terms. MAGNETO detected 178 unique biological processes in total in all the cell types. In contrast, the standard enrichment analysis detected 48 unique biological processes in all the cell types.

7. The biological processes detected in the pig infarcted myocardium with MAGNETO but not with the standard enrichment analysis method showed that all the cell types analysed, in response to MI, activate the innate immune mechanisms initiating an inflammatory reaction and at the same time deliver signals to promote cardiac remodelling and fibroblast proliferation.

Conclusiones

1. ATtRACT recopila en un único repositorio todas las proteínas de unión a ARN y sus motivos asociados que estaban esparcidos por diversas bases de datos. Comparado con las demás bases de datos, ATtRACT añade 192 motivos más obtenidos mediante un algoritmo diseñado para extraer los motivos de los datos cristalográficos presentes en "Protein Data Bank". Actualmente, ATtRACT es la colección más amplia y actualizada de proteínas de unión a ARN y sus motivos de interacción validados experimentalmente.

2. En comparación con el corazón de animales pequeños, el corazón del cerdo comparte muchas similitudes anatómicas y fisiológicas con el corazón humano. Sin embargo, el uso del cerdo como animal modelo trae consigo el inconveniente de la deficiente anotación funcional de sus genes. El uso de las anotaciones de otras especies mediante la identificación de los genes ortólogos ayuda en la interpretación de los resultados cuando la especie de interés está poco anotada.

3. Se han detectado en total doscientos diez genes con eventos de *splicing* alternativo después del infarto de miocardio en cardiomiocitos, fibroblastos, células endoteliales y macrófagos. Cuarenta y uno de esos eventos también se han detectado a nivel de proteína. Estos resultados indican que, o bien la sensibilidad de detección de la tecnología no es óptima, o bien el *splicing* alternativo no es un mecanismo relevante en la generación de cambios en respuesta al infarto de miocardio.

4. El análisis de enriquecimiento funcional de estos eventos de *splicing* alternativo en respuesta al infarto de miocardio muestra pocos términos enriquecidos, y estos están a su vez soportados por muy pocos genes, sugiriendo que no hay ningún proceso biológico cuya regulación se base en el *splicing* alternativo como mecanismo principal. Un análisis basado en al literatura nos muestra que las isoformas alternativas de Apolipoproteina D y Folistatina son las que más impacto biológico podrían tener en el infarto de miocardio.

5. El análisis de enriquecimiento funcional mejora drásticamente cuando se incorporan las redes de interacción proteína-proteína. A este respecto, los análisis de rendimiento muestran que MAGNETO supera a los métodos estándar de enriquecimiento funcional a partir de un número de 50 proteínas en adelante. MAGNETO mejora la clasificación y p-

valores de las vías diana de este análisis. Por consiguiente, MAGNETO es una herramienta valiosa para mejorar el conocimiento biológico de un estado fenotípico dado cuando el número de genes disponible para su interpretación es reducido.

6. Cuando usamos los ortólogos de humano de los 50 genes más expresados en el corazón de cerdo infartado, MAGNETO obtiene más términos sobrerrepresentados. MAGNETO detecta 178 procesos biológicos en total entre todos los tipos celulares estudiados. Sin embargo, el análisis de enriquecimiento funcional estándar sólo detectó 48 procesos biológicos.

7. Los procesos biológicos detectados con MAGNETO, y no con los métodos estándar, muestran que en todos los tipos celulares analizados, en respuesta al infarto de miocardio, se disparan los mecanismos para activar la respuesta inmune innata iniciando una reacción inflamatoria y al mismo tiempo liberan señales que promueven el remodelado cardíaco y la proliferación de fibroblastos.

Bibliography

Alipanahi, B., Delong, A., Weirauch, M.T. & Frey, B.J., 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotech*, 33(8), pp.831–838. Available at: http://dx.doi.org/10.1038/nbt.3300.

Alon, U., 2003. Biological networks: the tinkerer as an engineer. *Science (New York, N.Y.)*, 301(5641), pp.1866–1867.

Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F. & Hamosh, A., 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic acids research*, 43(Database issue), pp.D789-98.

Ammari, M.G., Gresham, C.R., McCarthy, F.M. & Nanduri, B., 2016. HPIDB 2.0: a curated database for host-pathogen interactions. *Database : the journal of biological databases and curation*, 2016.

Anon, 1986. Nomenclature Committee for the International Union of Biochemistry (NC-IUB). Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. *Molecular biology and evolution*, 3(2), pp.99–108.

Anon, 2015. UniProt: a hub for protein information. *Nucleic acids research*, 43(Database issue), pp.D204-12.

Ascano, M., Hafner, M., Cekan, P., Gerstberger, S. & Tuschl, T., 2012. Identification of RNA-protein interaction networks using PAR-CLIP. *Wiley interdisciplinary reviews. RNA*, 3(2), pp.159–177.

Auweter, S.D., Fasan, R., Reymond, L., Underwood, J.G., Black, D.L., Pitsch, S. & Allain, F.H.-T., 2006. Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. *The EMBO journal*, 25(1), pp.163–173.

Bader, G.D. & Hogue, C.W. V, 2003. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4, p.2.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. & Noble, W.S., 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, 37(Web Server issue), pp.W202-8.

Barabasi, A.-L. & Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. *Nature reviews. Genetics*, 5(2), pp.101–113.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J. & Frey, B.J., 2010.

Deciphering the splicing code. *Nature*, 465(7294), pp.53–59.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S. & Soboleva, A., 2013. NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids research*, 41(Database issue), pp.D991-5.

Bergmann, O., Zdunek, S., Felker, A., Salehpour, M., Alkass, K., Bernard, S., Sjostrom, S.L., Szewczykowska, M., Jackowska, T., Dos Remedios, C., Malm, T., Andra, M., Jashari, R., Nyengaard, J.R., Possnert, G., Jovinge, S., Druid, H. & Frisen, J., 2015. Dynamics of Cell Generation and Turnover in the Human Heart. *Cell*, 161(7), pp.1566–1575.

Bonzon-Kulichenko E, Garcia-Marques F, Trevisan-Herraz M, Vazquez J., 2015. Revisiting peptide identification by high-accuracy mass spectrometry: problems associated with the use of narrow mass precursor windows. J Proteome Res 14:700-710 doi:10.1021/pr5007284

van den Borne, S.W.M., Diez, J., Blankesteijn, W.M., Verjans, J., Hofstra, L. & Narula, J., 2010. Myocardial remodeling after infarction: the role of myofibroblasts. *Nature reviews. Cardiology*, 7(1), pp.30–37.

Breitkreutz, B.-J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D.H., Bahler, J., Wood, V., Dolinski, K. & Tyers, M., 2008. The BioGRID Interaction Database: 2008 update. *Nucleic acids research*, 36(Database issue), pp.D637-40.

Bujak, M. & Frangogiannis, N.G., 2007. The role of TGF-β Signaling in Myocardial Infarction and Cardiac Remodeling. *Cardiovascular research*, 74(2), pp.184–195.

Burrows, M. & Wheeler, D.J., 1994. *A block-sorting lossless data compression algorithm.*, Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.6774.

Cai, W.-F., Liu, G.-S., Wang, L., Paul, C., Wen, Z.-L. & Wang, Y., 2016. Repair Injured Heart by Regulating Cardiac Regenerative Signals. *Stem Cells International*, 2016.

Cao, M., Pietras, C.M., Feng, X., Doroschak, K.J., Schaffner, T., Park, J., Zhang, H., Cowen, L.J. & Hescott, B.J., 2014. New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics* , 30(12), pp.i219–i227. Available at: http://bioinformatics.oxfordjournals.org/content/30/12/i219.abstract.

Carson, D.D., 2008. The cytoplasmic tail of MUC1: a very busy place. *Science signaling*, 1(27), p.pe35.

Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A. V, Kalsotra, A., Cooper, T.A. & Johnson, J.M., 2008. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nature genetics*, 40(12), pp.1416–1425.

Chen, B. & Frangogiannis, N.G., 2016. Macrophages in the Remodeling Failing Heart. *Circulation research*, 119(7), pp.776–778.

Chen, Y. & Varani, G., 2013. Engineering RNA-binding proteins for biology. *The FEBS journal*, 280(16), pp.3734–3754.

Choi, S.J., Han, J.H. & Roodman, G.D., 2001. ADAM8: a novel osteoclast stimulating factor. *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research*, 16(5), p.814—822. Available at: http://dx.doi.org/10.1359/jbmr.2001.16.5.814.

Christia, P. & Frangogiannis, N.G., 2013. Targeting inflammatory pathways in myocardial infarction. *European journal of clinical investigation*, 43(9), pp.986–995.

Cieply, B. & Carstens, R.P., 2015. Functional roles of alternative splicing factors in human disease. *Wiley interdisciplinary reviews. RNA*, 6(3), pp.311–326.

Collesi, C., Santoro, M.M., Gaudino, G. & Comoglio, P.M., 1996. A splicing variant of the RON transcript induces constitutive tyrosine kinase activity and an invasive phenotype. *Molecular and cellular biology*, 16(10), pp.5518–5526.

Cook, K.B., Kazan, H., Zuberi, K., Morris, Q. & Hughes, T.R., 2011. RBPDB: a database of RNA-binding specificities. *Nucleic acids research*, 39(Database issue), pp.D301-8.

Cornish, A.J. & Markowetz, F., 2014. SANTA: quantifying the functional content of molecular networks. *PLoS computational biology*, 10(9), p.e1003808.

Cowley, M.J., Pinese, M., Kassahn, K.S., Waddell, N., Pearson, J. V, Grimmond, S.M., Biankin, A. V, Hautaniemi, S. & Wu, J., 2012. PINA v2.0: mining interactome modules. *Nucleic acids research*, 40(Database issue), pp.D862-5.

Dapas, M., Kandpal, M., Bi, Y. & Davuluri, R. V, 2016. Comparative evaluation of isoform-level gene expression estimation algorithms for RNA-seq and exon-array platforms. *Briefings in Bioinformatics* . Available at:

http://bib.oxfordjournals.org/content/early/2016/02/26/bib.bbw016.abstract.

Davies, W., 2016. Insights into rare diseases from social media surveys. *Orphanet journal of rare diseases*, 11(1), p.151.

Deb, A. & Ubil, E., 2014. Cardiac fibroblast in development and wound healing. *Journal of molecular and cellular cardiology*, 70, pp.47–55.

Dobaczewski, M., Gonzalez-Quesada, C. & Frangogiannis, N.G., 2010. The extracellular matrix as a modulator of the inflammatory and reparative response following myocardial infarction. *Journal of molecular and cellular cardiology*, 48(3), pp.504–511.

Donaldson, C.J., Vaughan, J.M., Corrigan, A.Z., Fischer, W.H. & Vale, W.W., 1999. Activin and inhibin binding to the soluble extracellular domain of activin receptor II. *Endocrinology*, 140(4), pp.1760–1766.

van Dongen, S. & Abreu-Goodger, C., 2012. Using MCL to extract clusters from networks. *Methods in molecular biology (Clifton, N.J.)*, 804, pp.281–295.

Dutta, B., Wallqvist, A. & Reifman, J., 2012. PathNet: a tool for pathway analysis using topological information. *Source code for biology and medicine*, 7(1), p.10.

Euler, G., 2015. Good and bad sides of TGFβ-signaling in myocardial infarction. *Frontiers in Physiology*, 6.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., Matthews, L., May, B., Milacic, M., Rothfels, K., Shamovsky, V., Webber, M., Weiser, J., Williams, M., Wu, G., Stein, L., Hermjakob, H. & D'Eustachio, P., 2016. The Reactome pathway Knowledgebase. *Nucleic acids research*, 44(D1), pp.D481-7.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J. & Punta, M., 2014. Pfam: the protein families database. *Nucleic acids research*, 42(Database issue), pp.D222-30.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Searle, S.M.J., et al., 2014. Ensembl 2014. *Nucleic acids research*, 42(Database issue), pp.D749-55.

Frangogiannis, N.G., 2014. The inflammatory response in myocardial injury, repair, and remodelling. *Nature reviews. Cardiology*, 11(5), pp.255–265.

Frangogiannis, N.G., Smith, C.W. & Entman, M.L., 2002. The inflammatory response in myocardial infarction. *Cardiovascular research*, 53(1), pp.31–47.

Fuentes-Prior, P., Noeske-Jungblut, C., Donner, P., Schleuning, W.-D., Huber, R. & Bode, W., 1997. Structure of the thrombin complex with triabin, a lipocalin-like exosite-binding inhibitor derived from a triatomine bug. *Proceedings of the National Academy of Sciences* , 94(22), pp.11845–11850. Available at: http://www.pnas.org/content/94/22/11845.abstract.

Fujiu, K., Wang, J. & Nagai, R., 2014. Cardioprotective function of cardiac macrophages. *Cardiovascular research*, 102(2), pp.232–239.

Fukuda, M., 2002. Roles of mucin-type O-glycans in cell adhesion. *Biochimica et biophysica acta*, 1573(3), pp.394–405.

Gale, N.W. & Yancopoulos, G.D., 1999. Growth factors acting via endothelial cell-specific receptor tyrosine kinases: VEGFs, angiopoietins, and ephrins in vascular development. *Genes & development*, 13(9), pp.1055–1066.

Gautier, L., Cope, L., Bolstad, B.M. & Irizarry, R.A., 2004. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics (Oxford, England)*, 20(3), pp.307–315.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H. & Zhang, J., 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10), p.R80.

Georgiev, S., Boyle, A.P., Jayasurya, K., Ding, X., Mukherjee, S. & Ohler, U., 2010. Evidence-ranked motif identification. *Genome biology*, 11(2), p.R19.

Ghigna, C., Giordano, S., Shen, H., Benvenuto, F., Castiglioni, F., Comoglio, P.M., Green, M.R., Riva, S. & Biamonti, G., 2005. Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene. *Molecular cell*, 20(6), pp.881–890.

Ghigo, A., Franco, I., Morello, F. & Hirsch, E., 2014. Myocyte signalling in leucocyte recruitment to the heart. *Cardiovascular Research*, 102(2), p.270 LP-280. Available at: http://cardiovascres.oxfordjournals.org/content/102/2/270.abstract.

Giulietti, M., Piva, F., D’Antonio, M., De Meo, P.D.O., Paoletti, D., Castrignanò, T., D’Erchia, A.M., Picardi, E., Zambelli, F., Principato, G., Pavesi, G. & Pesole, G., 2013. SpliceAid-F: A database of human splicing factors and their RNA-binding sites. *Nucleic Acids Research*, 41(D1).

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R. & Valencia, A., 2012. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics (Oxford, England)*, 28(18), pp.i451–i457.

Glisovic, T., Bachorik, J.L., Yong, J. & Dreyfuss, G., 2008. RNA-binding proteins and post-transcriptional gene regulation. *FEBS letters*, 582(14), pp.1977–1986.

Gonzalez, M.W. & Kann, M.G., 2012. Chapter 4: Protein Interactions and Disease. *PLOS Computational Biology*, 8(12), pp.1–11. Available at: http://dx.doi.org/10.1371%2Fjournal.pcbi.1002819.

Groenen, M.A.M., 2016. A decade of pig genome sequencing: a window on pig domestication and evolution. *Genetics, selection, evolution : GSE*, 48, p.23.

Gruber, T.R., 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2), pp.199–220. Available at: http://www.sciencedirect.com/science/article/pii/S1042814383710083.

Guarda, E., Myers, P.R., Brilla, C.G., Tyagi, S.C. & Weber, K.T., 1993. Endothelial cell induced modulation of cardiac fibroblast collagen metabolism. *Cardiovascular research*, 27(6), pp.1004–1008.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. & Noble, W.S., 2007. Quantifying similarity between motifs. *Genome biology*, 8(2), p.R24.

Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W., 1999. From molecular to modular cell biology. *Nature*, 402(6761 Suppl), pp.C47-52.

Hashimoto, O., Nakamura, T., Shoji, H., Shimasaki, S., Hayashi, Y. & Sugino, H., 1997. A novel role of follistatin, an activin-binding protein, in the inhibition of activin action in rat pituitary cells. Endocytotic degradation of activin and its acceleration by follistatin associated with cell-surface heparan sulfate. *The Journal of biological chemistry*, 272(21), pp.13835–13842.

Hatzoglou, A., Roussel, J., Bourgeade, M.-F., Rogier, E., Madry, C., Inoue, J., Devergne, O. & Tsapis, A., 2000. TNF Receptor Family Member BCMA (B Cell Maturation) Associates with TNF Receptor-Associated Factor (TRAF) 1, TRAF2, and TRAF3 and Activates NF-κB, Elk-1, c-Jun N-Terminal Kinase, and p38 Mitogen-Activated Protein Kinase. *The Journal of Immunology* , 165(3), pp.1322–1330. Available at: http://www.jimmunol.org/content/165/3/1322.abstract.

Hernández-Vargas, M.J., Santibáñez-López, C.E. & Corzo, G., 2016. An Insight into the Triabin Protein Family of American Hematophagous Reduviids: Functional, Structural and Phylogenetic Analysis G. F. King, ed. *Toxins*, 8(2).

Hilkens, J., Ligtenberg, M.J., Vos, H.L. & Litvinov, S. V, 1992. Cell membrane-associated mucins and their adhesion-modulating property. *Trends in biochemical sciences*, 17(9), pp.359–363.

Hiller, M., Pudimat, R., Busch, A. & Backofen, R., 2006. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic acids research*, 34(17), p.e117.

Hinz, B., Phan, S.H., Thannickal, V.J., Galli, A., Bochaton-Piallat, M.-L. & Gabbiani, G., 2007. The myofibroblast: one function, multiple origins. *The American journal of pathology*, 170(6), pp.1807–1816.

Hu, B., Yang, Y.-C.T., Huang, Y., Zhu, Y. & Lu, Z.J., 2016. POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Research* . Available at: http://nar.oxfordjournals.org/content/early/2016/10/05/nar.gkw888.abstract.

Huang, D.W., Sherman, B.T. & Lempicki, R.A., 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1), pp.1–13.

Hudson, B.I., Carter, A.M., Harja, E., Kalea, A.Z., Arriero, M., Yang, H., Grant, P.J. & Schmidt, A.M., 2008. Identification, classification, and expression of RAGE gene splice variants. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 22(5), pp.1572–1580.

Huntley, R.P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M.J. & O'Donovan, C., 2015. The GOA database: gene Ontology annotation updates for 2015. *Nucleic acids research*, 43(Database issue), pp.D1057-63.

Huynen, M., Snel, B., Lathe, W. 3rd & Bork, P., 2000. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome research*, 10(8), pp.1204–1210.

Ideker, T., Galitski, T. & Hood, L., 2001. A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, 2, pp.343–372.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. & Speed, T.P., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, 4(2), pp.249–264.

Iskratsch, T., Lange, S., Dwyer, J., Kho, A.L., dos Remedios, C. & Ehler, E., 2010. Formin follows function: a muscle-specific isoform of FHOD3 is regulated by CK2 phosphorylation and promotes myofibril maintenance. *The Journal of cell biology*, 191(6), pp.1159–1172.

Iskratsch, T., Reijntjes, S., Dwyer, J., Toselli, P., Degano, I.R., Dominguez, I. & Ehler, E., 2013. Two distinct phosphorylation events govern the function of muscle FHOD3. *Cellular and molecular life sciences : CMLS*, 70(5), pp.893–908.

Jankowsky, E. & Harris, M.E., 2015a. Specificity and nonspecificity in RNA-protein interactions. *Nature reviews. Molecular cell biology*, 16(9), pp.533–544.

Jankowsky, E. & Harris, M.E., 2015b. Specificity and nonspecificity in RNA-protein interactions. *Nature reviews. Molecular cell biology*, 16(9), pp.533–544.

Jones, S., Daley, D.T.A., Luscombe, N.M., Berman, H.M. & Thornton, J.M., 2001. Protein–RNA interactions: a structural analysis. *Nucleic Acids Research*, 29(4), pp.943–954. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC29619/.

Jones, S. & Thornton, J.M., 1996a. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 93(1), pp.13–20.

Jones, S. & Thornton, J.M., 1996b. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 93(1), pp.13–20.

Kakkar, R. & Lee, R.T., 2010. Intramyocardial fibroblast myocyte communication. *Circulation research*, 106(1), pp.47–57.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M., 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(Database issue), pp.D109-14.

Kanehisa, M., 2013. Molecular network analysis of diseases and drugs in KEGG. *Methods in molecular biology (Clifton, N.J.)*, 939, pp.263–275.

Karpinka, J.B., Fortriede, J.D., Burns, K.A., James-Zorn, C., Ponferrada, V.G., Lee, J., Karimi, K., Zorn, A.M. & Vize, P.D., 2014. Xenbase, the Xenopus model organism database; new virtualized system, data types and genomes. *Nucleic acids research*.

Keren, H., Lev-Maor, G. & Ast, G., 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nature reviews. Genetics*, 11(5), pp.345–355. Available at: http://dx.doi.org/10.1038/nrg2776.

Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A.F., Vinod, N., Bader, G.D., Xenarios, I., Wojcik, J., Sherman, D., Tyers, M., Salama, J.J., Moore, S., Ceol, A., Chatr-Aryamontri, A., Oesterheld, M., Stumpflen, V., Salwinski, L., Nerothin, J., Cerami, E., Cusick, M.E., Vidal, M., Gilson, M., Armstrong, J., Woollard, P., Hogue, C., Eisenberg, D., Cesareni, G., Apweiler, R. & Hermjakob, H., 2007. Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions. *BMC biology*, 5, p.44.

Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Pandey, A., et al., 2014. A draft map of the human proteome. *Nature*, 509(7502), pp.575–581.

Kitano, H., 2002a. Computational systems biology. *Nature*, 420(6912), pp.206–210.

Kitano, H., 2002b. Systems biology: a brief overview. *Science (New York, N.Y.)*, 295(5560), pp.1662–1664.

Kriventseva, E. V, Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M.S. & Sunyaev, S., 2003. Increase of functional diversity by alternative splicing. *Trends in genetics : TIG*, 19(3), pp.124–128.

Lara-Pezzi, E., Gomez-Salinero, J., Gatto, A. & Garcia-Pavia, P., 2013. The alternative heart: impact of alternative splicing in heart disease. *Journal of cardiovascular translational research*, 6(6), pp.945–955.

Larochelle, S., 2016. Systems biology: Protein isoforms: more than meets the eye. *Nat Meth*, 13(4), p.291. Available at: http://dx.doi.org/10.1038/nmeth.3828.

Laukens, K., Naulaerts, S. & Berghe, W. Vanden, 2015. Bioinformatics approaches for the functional interpretation of protein lists: from ontology term enrichment to network analysis. *Proteomics*, 15(5–6), pp.981–996.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z.T., Han, B., Zhou, Y. & Wishart, D.S., 2014. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids* *research*, 42(Database issue), pp.D1091-7.

Leclerc, E., Fritz, G., Weibel, M., Heizmann, C.W. & Galichet, A., 2007. S100B and S100A6 differentially modulate cell survival by interacting with distinct RAGE (receptor for advanced glycation end products) immunoglobulin domains. *The Journal of biological chemistry*, 282(43), pp.31317–31331.

Lee, S.J. & McPherron, A.C., 2001. Regulation of myostatin activity and muscle growth. *Proceedings of the National Academy of Sciences of the United States of America*, 98(16), pp.9306–9311.

Lehto, M., Mayranpaa, M.I., Pellinen, T., Ihalmo, P., Lehtonen, S., Kovanen, P.T., Groop, P.-H., Ivaska, J. & Olkkonen, V.M., 2008. The R-Ras interaction partner ORP3 regulates cell adhesion. *Journal of cell science*, 121(Pt 5), pp.695–705.

Leibovich, L., Paz, I., Yakhini, Z. & Mandel-Gutfreund, Y., 2013. DRIMust: a web server for discovering rank imbalanced motifs using suffix trees. *Nucleic acids research*, 41(Web Server issue), pp.W174-9.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., Zalunin, V. & Cochrane, G., 2011. The European Nucleotide Archive. *Nucleic Acids Research* , 39(suppl 1), pp.D28–D31. Available at: http://nar.oxfordjournals.org/content/39/suppl_1/D28.abstract.

Di Lena, P., Martelli, P.L., Fariselli, P. & Casadio, R., 2015. NET-GE: a novel NETwork-based Gene Enrichment for detecting biological processes associated to Mendelian diseases. *BMC genomics*, 16 Suppl 8, p.S6.

Li, B. & Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1), pp.1–16. Available at: http://dx.doi.org/10.1186/1471-2105-12-323.

Li, H. & Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5), pp.589–595.

Li, H. & Homer, N., 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5), pp.473–483.

Lian, H., Ma, Y., Feng, J., Dong, W., Yang, Q., Lu, D. & Zhang, L., 2012. Heparin-binding EGF-like growth factor induces heart interstitial fibrosis via an Akt/mTor/p70s6k pathway. *PloS one*, 7(9), p.e44946.

Licatalosi, D.D. & Darnell, R.B., 2010. RNA processing and its regulation: global insights into biological networks. *Nature reviews. Genetics*, 11(1), pp.75–87.

Lim, E., Pon, A., Djoumbou, Y., Knox, C., Shrivastava, S., Guo, A.C., Neveu, V. & Wishart, D.S., 2010. T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic acids research*, 38(Database issue), pp.D781-6.

Lin, L. & Knowlton, A.A., 2014. Innate immunity and cardiomyocytes in ischemic heart disease. *Life sciences*, 100(1), pp.1–8.

Lu, L., Zhang, Q., Xu, Y., Zhu, Z., Geng, L., Wang, L., Jin, C., Chen, Q., Schmidt, A.M. & Shen, W., 2010. Intra-coronary administration of soluble receptor for advanced glycation end-products attenuates cardiac remodeling with decreased myocardial transforming growth factor-beta1 expression and fibrosis in minipigs with ischemia-reperfusion injury. *Chinese medical journal*, 123(5), pp.594–598.

Luehr, S., Hartmann, H. & Soding, J., 2012. The XXmotif web server for eXhaustive, weight matriX-based motif discovery in nucleotide sequences. *Nucleic acids research*, 40(Web Server issue), pp.W104-9.

Lukong, K.E., Chang, K., Khandjian, E.W. & Richard, S., 2008. RNA-binding proteins in human genetic disease. *Trends in genetics : TIG*, 24(8), pp.416–425.

Lunde, B.M., Moore, C. & Varani, G., 2007. RNA-binding proteins: modular design for efficient function. *Nature reviews. Molecular cell biology*, 8(6), pp.479–490.

Luscombe, N.M., Laskowski, R.A. & Thornton, J.M., 1997. NUCPLOT: a program to generate schematic diagrams of protein-nucleic acid interactions. *Nucleic acids research*, 25(24), pp.4940–4945.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal; Vol 17, No 1: Next Generation Sequencing Data Analysis*. Available at: http://journal.embnet.org/index.php/embnetjournal/article/view/200.

Martinez-Bartolome S, Navarro P, Martin-Maroto F, Lopez-Ferrer D, Ramos-Fernandez A, Villar M, Garcia-Ruiz JP, Vazquez J., 2008 Properties of average score distributions of SEQUEST: the probability ratio method. Mol Cell Proteomics pp.1135-1145.

Marzluff, W.F., Wagner, E.J. & Duronio, R.J., 2008. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nature reviews. Genetics*, 9(11), pp.843–854.

McDonald, I.K. & Thornton, J.M., 1994. Satisfying hydrogen bonding potential in proteins. *Journal of molecular biology*, 238(5), pp.777–793.

McManus, C.J. & Graveley, B.R., 2011. RNA structure and the mechanisms of alternative splicing. *Current opinion in genetics & development*, 21(4), pp.373–379.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. & Bork, P., 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887), pp.399–403.

Michiels, C., 2003. Endothelial cell functions. *Journal of cellular physiology*, 196(3), pp.430–443.

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., Sangrador-Vegas, A., Scheremetjew, M., Rato, C., Yong, S.-Y., Bateman, A., Punta, M., Attwood, T.K., Sigrist, C.J.A., Redaschi, N., Rivoire, C., Xenarios, I., Kahn, D., Guyot, D., Bork, P., Letunic, I., Gough, J., Oates, M., Haft, D., Huang, H., Natale, D.A., Wu, C.H., Orengo, C., Sillitoe, I., Mi, H., Thomas, P.D. & Finn, R.D., 2015. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research* , 43(D1), pp.D213–D221. Available at: http://nar.oxfordjournals.org/content/43/D1/D213.abstract.

Moss, T.N., Vo, A., McKeehan, W.L. & Liu, L., 2007. UXT (Ubiquitously Expressed Transcript) causes mitochondrial aggregation. *In vitro cellular & developmental biology. Animal*, 43(3–4), pp.139–146.

Muller, W.A., 2003. Leukocyte-endothelial-cell interactions in leukocyte transmigration and the inflammatory response. *Trends in immunology*, 24(6), pp.327–334.

Nabieva, E., Jim, K., Agarwal, A., Chazelle, B. & Singh, M., 2005. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics (Oxford, England)*, 21 Suppl 1, pp.i302-10.

Natale, M., Benso, A., Di Carlo, S. & Ficarra, E., 2014. FunMod: a Cytoscape plugin for identifying functional modules in undirected protein-protein networks. *Genomics, proteomics & bioinformatics*, 12(4), pp.178–186.

Navarro P, Vazquez J., 2009. A refined method to calculate false discovery rates for peptide identification using decoy databases. J Proteome Res 8(1) pp.792-1796.

Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., Bridge, A., Briganti, L., Brinkman, F.S.L., Cesareni, G., Chatr-aryamontri, A., Chautard, E., Chen, C., Dumousseau, M., Goll, J., Hancock, R.E.W., Hannick, L.I., Jurisica, I., Khadake, J., Lynn, D.J., Mahadevan, U., Perfetto, L., Raghunath, A., Ricard-Blum, S., Roechert, B., Salwinski, L., Stumpflen, V., Tyers, M., Uetz, P., Xenarios, I. & Hermjakob, H., 2012. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature methods*, 9(4), pp.345–350.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R.C., Meldal, B., Melidoni, A.N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., van Roey, K., Cesareni, G. & Hermjakob, H., 2014. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42(Database issue), pp.D358-63.

Ozcelik, C., Erdmann, B., Pilz, B., Wettschureck, N., Britsch, S., Hubner, N., Chien, K.R., Birchmeier, C. & Garratt, A.N., 2002. Conditional mutation of the ErbB2 (HER2) receptor in cardiomyocytes leads to dilated cardiomyopathy. *Proceedings of the National Academy of Sciences of the United States of America*, 99(13), pp.8880–8885.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J., 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12), pp.1413–1415.

Pinto, A.R., Ilinykh, A., Ivey, M.J., Kuwabara, J.T., D'Antoni, M.L., Debuque, R., Chandran, A., Wang, L., Arora, K., Rosenthal, N.A. & Tallquist, M.D., 2016. Revisiting Cardiac Cellular Composition. *Circulation research*, 118(3), pp.400–409.

Ponthier, J.L., Schluepen, C., Chen, W., Lersch, R.A., Gee, S.L., Hou, V.C., Lo, A.J., Short, S.A., Chasis, J.A., Winkelmann, J.C. & Conboy, J.G., 2006. Fox-2 splicing factor binds to a conserved intron motif to promote inclusion of protein 4.1R alternative exon 16. *The Journal of biological* *chemistry*, 281(18), pp.12468–12474.

Queralt-Rosinach, N., Pinero, J., Bravo, A., Sanz, F. & Furlong, L.I., 2016. DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases. *Bioinformatics (Oxford, England)*, 32(14), pp.2236–2238.

Ramasamy, R. & Schmidt, A.M., 2012. Receptor for Advanced Glycation End Products (RAGE) and Implications for the Pathophysiology of Heart Failure. *Current heart failure reports*, 9(2), pp.107–116.

Rappsilber, J., Ryder, U., Lamond, A.I. & Mann, M., 2002. Large-scale proteomic analysis of the human spliceosome. *Genome research*, 12(8), pp.1231–1245.

Rassart, E., Bedirian, A., Do Carmo, S., Guinard, O., Sirois, J., Terrisse, L. & Milne, R., 2000. Apolipoprotein D. *Biochimica et biophysica acta*, 1482(1–2), pp.185–198.

Rattray, A.M.J. & Muller, B., 2012. The control of histone gene expression. *Biochemical Society transactions*, 40(4), pp.880–885.

Rattray, A.M.J., Nicholson, P. & Muller, B., 2013. Replication stress-induced alternative mRNA splicing alters properties of the histone RNA-binding protein HBP/SLBP: a key factor in the control of histone gene expression. *Bioscience reports*, 33(5).

Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., Na, H., Irimia, M., Matzat, L.H., Dale, R.K., Smith, S.A., Yarosh, C.A., Kelly, S.M., Nabet, B., Mecenas, D., Li, W., Laishram, R.S., Qiao, M., Lipshitz, H.D., Piano, F., Corbett, A.H., Carstens, R.P., Frey, B.J., Anderson, R.A., Lynch, K.W., Penalva, L.O.F., Lei, E.P., Fraser, A.G., Blencowe, B.J., Morris, Q.D. & Hughes, T.R., 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457), pp.172–177.

Rodino-Klapac L.R., Haidet, A.M., Kota, J., Handy, C., Kaspar, B.K. & Mendell, J.R., 2009. Inibition of myostatin with emphasis on follistatin as a therapy for muscle disease. *Muscle & nerve*, 39(3), pp.283–296.

Rose, P.W., Prlic, A., Bi, C., Bluhm, W.F., Christie, C.H., Dutta, S., Green, R.K., Goodsell, D.S., Westbrook, J.D., Woo, J., Young, J., Zardecki, C., Berman, H.M., Bourne, P.E. & Burley, S.K., 2015. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic acids research*, 43(Database issue), pp.D345-56.

Ruan, N., Jin, R. & Huang, Y., 2011. Distance Preserving Graph Simplification. *2011 IEEE 11th International Conference on Data Mining*, pp.1200–1205.

Saini, H.K., Xu, Y.-J., Zhang, M., Liu, P.P., Kirshenbaum, L.A. & Dhalla, N.S., 2005. Role of tumour necrosis factor-alpha and other cytokines in ischemia-reperfusion-induced injury in the heart. *Experimental & Clinical Cardiology*, 10(4), pp.213–222.

Schaper, J., Meiser, E. & Stammler, G., 1985. Ultrastructural morphometric analysis of myocardium from dogs, rats, hamsters, mice, and from human hearts. *Circulation research*, 56(3), pp.377–391.

Senyo, S.E., Steinhauser, M.L., Pizzimenti, C.L., Yang, V.K., Cai, L., Wang, M., Wu, T.-D., Guerquin-Kern, J.-L., Lechene, C.P. & Lee, R.T., 2013. Mammalian heart renewal by pre-existing cardiomyocytes. *Nature*, 493(7432), pp.433–436. Available at: http://dx.doi.org/10.1038/nature11682.

Shinde, A. V & Frangogiannis, N.G., 2014. Fibroblasts in myocardial infarction: a role in inflammation and repair. *Journal of molecular and cellular cardiology*, 70, pp.74–82.

Smith, C.A., Farrah, T. & Goodwin, R.G., 1994. The TNF receptor superfamily of cellular and viral proteins: activation, costimulation, and death. *Cell*, 76(6), pp.959–962.

Smith, C.W. & Valcarcel, J., 2000. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends in biochemical sciences*, 25(8), pp.381–388.

Smyth, G.K., 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3, p.Article3.

Spirin, V. & Mirny, L.A., 2003. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21), pp.12123–12128.

Sprinzak, E., Sattath, S. & Margalit, H., 2003. How reliable are experimental protein-protein interaction data? *Journal of molecular biology*, 327(5), pp.919–923.

Staden, R., Staden, R., Road, H. & Road, H., 1982. Nucleic Acids Research. *Nucleic Acids Research*, 10, pp.2951–2961.

Stamm, S., Riethoven, J.-J., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Tang, Y., Barbosa-

Morais, N.L. & Thanaraj, T.A., 2006. ASD: a bioinformatics resource on alternative splicing. *Nucleic acids research*, 34(Database issue), pp.D46-55.

Stuart, J.M., Segal, E., Koller, D. & Kim, S.K., 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, N.Y.)*, 302(5643), pp.249–255.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. & Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), pp.15545–15550.

Sumpio, B.E., Riley, J.T. & Dardik, A., 2002. Cells in focus: endothelial cell. *The international journal of biochemistry & cell biology*, 34(12), pp.1508–1512.

Swindle, M.M., Makin, A., Herron, A.J., Clubb, F.J.J. & Frazier, K.S., 2012. Swine as models in biomedical research and toxicology testing. *Veterinary pathology*, 49(2), pp.344–356.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., Kuhn, M., Bork, P., Jensen, L.J. & von Mering, C., 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(Database issue), pp.D447-52.

Talman, V. & Ruskoaho, H., 2016. Cardiac fibrosis in myocardial infarction-from repair and remodeling to regeneration. *Cell and tissue research*.

Tarca, A.L., Draghici, S., Bhatti, G. & Romero, R., 2012. Down-weighting overlapping genes improves gene set analysis. *BMC bioinformatics*, 13, p.136.

Townsend, N., Nichols, M., Scarborough, P. & Rayner, M., 2015. Cardiovascular disease in Europe--epidemiological update 2015. *European heart journal*, 36(40), pp.2696–2705.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. & Pachter, L., 2010a. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5), pp.511–515. Available at: http://www.nature.com/nbt/journal/v28/n5/full/nbt.1621.html\nhttp://www.nature.com/nbt/journal/v28/n5/pdf/nbt.1621.pdf.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. & Pachter, L., 2010b. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5), pp.511–515.

Tsoporis, J.N., Izhar, S., Proteau, G., Slaughter, G. & Parker, T.G., 2012. S100B-RAGE dependent VEGF secretion by cardiac myocytes induces myofibroblast proliferation. *Journal of molecular and cellular cardiology*, 52(2), pp.464–473.

Tsukamoto, K., Mani, D.R., Shi, J., Zhang, S., Haagensen, D.E., Otsuka, F., Guan, J., Smith, J.D., Weng, W., Liao, R., Kolodgie, F.D., Virmani, R. & Krieger, M., 2013. Identification of apolipoprotein D as a cardioprotective gene using a mouse model of lethal atherosclerotic coronary artery disease. *Proceedings of the National Academy of Sciences of the United States of America*, 110(42), pp.17023–17028.

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C.A.-K., Odeberg, J., Djureinovic, D., Takanen, J.O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J.M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J. & Pontén, F., 2015. Tissue-based map of the human proteome. *Science*, 347(6220). Available at: http://science.sciencemag.org/content/347/6220/1260419.abstract.

Ule, J., Jensen, K., Mele, A. & Darnell, R.B., 2005. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods (San Diego, Calif.)*, 37(4), pp.376–386.

Wahl, M.C., Will, C.L. & Luhrmann, R., 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell*, 136(4), pp.701–718.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P. & Burge, C.B., 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221), pp.470–476.

Wang, Z. & Burge, C.B., 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA (New York, N.Y.)*, 14(5), pp.802–813.

Wei, Y.-J., Huang, Y.-X., Zhang, X.-L., Li, J., Huang, J., Zhang, H. & Hu, S.-S., 2008. Apolipoprotein D as a novel marker in human end-stage heart failure: a preliminary study. *Biomarkers : biochemical indicators of exposure, response, and susceptibility to chemicals*, 13(5), pp.535–

548.

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.-H., Bantscheff, M., Gerstmair, A., Faerber, F. & Kuster, B., 2014. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502), pp.582–587.

Will, C.L. & Luhrmann, R., 2011. Spliceosome structure and function. *Cold Spring Harbor perspectives in biology*, 3(7).

Winterhalter, C., Widera, P. & Krasnogor, N., 2014. JEPETTO: a Cytoscape plugin for gene set enrichment and topological analysis based on interaction networks. *Bioinformatics (Oxford, England)*, 30(7), pp.1029–1030.

Wong, N.D., 2014. Epidemiological studies of CHD and the evolution of preventive cardiology. *Nature reviews. Cardiology*, 11(5), pp.276–289.

Woollard, K.J. & Geissmann, F., 2010. Monocytes in atherosclerosis: subsets and functions. *Nature reviews. Cardiology*, 7(2), pp.77–86.

Xiao, R., Tang, P., Yang, B., Huang, J., Zhou, Y., Shao, C., Li, H., Sun, H., Zhang, Y. & Fu, X.-D., 2012. Nuclear Matrix Factor hnRNP U/SAF-A Exerts a Global Control of Alternative Splicing by Regulating U2 snRNP Maturation. *Molecular Cell*, 45(5), pp.656–668.

Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G.M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y.A., Murray, R.R., Spirohn, K., Begg, B.E., Duran-Frigola, M., MacWilliams, A., Pevzner, S.J., Zhong, Q., Trigg, S.A., Tam, S., Ghamsari, L., Sahni, N., Yi, S., Rodriguez, M.D., Balcha, D., Tan, G., Costanzo, M., Andrews, B., Boone, C., Zhou, X.J., Salehi-Ashtiani, K., Charloteaux, B., Chen, A.A., Calderwood, M.A., Aloy, P., Roth, F.P., Hill, D.E., Iakoucheva, L.M., Xia, Y. & Vidal, M., 2016. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*, 164(4), pp.805–817. Available at: http://dx.doi.org/10.1016/j.cell.2016.01.029.

Yang, Y., Lv, J., Jiang, S., Ma, Z., Wang, D., Hu, W., Deng, C., Fan, C., Di, S., Sun, Y. & Yi, W., 2016. The emerging role of Toll-like receptor 4 in myocardial inflammation. *Cell Death & Disease*, 7(5), p.e2234-.

Ye, J., Beetz, N., O'Keeffe, S., Tapia, J.C., Macpherson, L., Chen, W. V, Bassel-Duby, R., Olson, E.N. &

Maniatis, T., 2015. hnRNP U protein is required for normal pre-mRNA splicing and postnatal heart development and function. *Proceedings of the National Academy of Sciences of the United States of America*, 112(23), pp.E3020-9.

Yeo, G.W., Xu, X., Liang, T.Y., Muotri, A.R., Carson, C.T., Coufal, N.G. & Gage, F.H., 2007. Alternative splicing events identified in human embryonic stem cells and neural progenitors. *PLoS computational biology*, 3(10), pp.1951–1967.

Zhang, L., Vlad, A., Milcarek, C. & Finn, O.J., 2013. Human Mucin MUC1 RNA Undergoes Different Types of Alternative Splicing Resulting in Multiple Isoforms. *Cancer immunology, immunotherapy : CII*, 62(3), pp.423–435.

Zhou, B. & Pu, W.T., 2011. Epicardial epithelial-to-mesenchymal transition in injured heart. *Journal of Cellular and Molecular Medicine*, 15(12), pp.2781–2783.

Zhou, Z., Licklider, L.J., Gygi, S.P. & Reed, R., 2002. Comprehensive proteomic analysis of the human spliceosome. *Nature*, 419(6903), pp.182–185.

ANNEX I

Supplementary

```
>ENSG00000149187
 UUGUG
||||||
UGUGUG
distance: 1
>ENSG00000149187
UUGUU
|||||
UGUGU
distance: 2
>ENSG00000149187
UGUU
|||||
UGUGU
distance: 1
>ENSG00000149187
UGUGUG
||||||
UGUGUG
distance: 0
>ENSG00000149187
UGUUU
|||||
UGUGU
distance: 1
>ENSG00000161547
UCCAGU
|||||||
UCGAGAU
distance: 2
>ENSG00000161547
UGGAGU
|||||||
UGGAGAU
distance: 1
>ENSG00000011304
UCUCU
|||||
UCUCU
distance: 0
>ENSG00000011304
CUCUCU
||||||
CUCUCU
distance: 0

>ENSG00000011304
CUCU
||||
CUCU
distance: 0
>ENSG00000197111
CCCUAA
||||||||
CCCUUAAA
distance: 1
```

```
>ENSG00000197111
CCCU
||||||
CCAUUC
distance: 1
>FBGN0264270 see Hennig,J., Militti,C., Popowicz,G.M., Wang,I.,
Sonntag,M., Geerlof,A., Gabel,F., Gebauer,F. and Sattler,M. (2014)
Structural basis for the assembly of the Sxl-Unr translation regulatory
complex. Nature, 515, 287–290. (see figure 2f )
GUGA
|||||||
UUUUUUU
distance: 4
>FBGN0264270
UGUUUUUUUU
||||||||
UUUUUUUU
distance: 1
>FBGN0264270 see Hennig,J., Militti,C., Popowicz,G.M., Wang,I.,
Sonntag,M., Geerlof,A., Gabel,F., Gebauer,F. and Sattler,M. (2014)
Structural basis for the assembly of the Sxl-Unr translation regulatory
complex. Nature, 515, 287–290.
GCACG
|||||||
UUUUUUU
distance: 5
>FBGN0264270
UUUUUUUGAGCA
|||||||
UUUUUUU
distance: 0
>ENSG00000063244
UUUUUUU
|||||||
UUUUUUC
distance: 1
>ENSG00000063244
UUUUU
|||||||
UUUUUCC
distance: 0
>ENSG00000234414
 ACAAGAC
|||||
CACAA
distance: 0
>ENSG00000131914
AGGAGAU
|||||||
AGGAGAA
distance: 1
```

```
>ENSG00000138385
     UUUU
|||||||||
UGCUGUUUU
distance: 0
>ENSG00000138385
    AUUU
|||||||||
UGCUGUUUU
distance: 1
>ENSG00000138385
UGCUG
|||||||||
UGCUGUUUU
distance: 0
>WBGENE00001595
  CUAAC
||||||||||||
AUCUACUCAUAU
distance: 2
>WBGENE00001595
  CUACUCAUAU
||||||||||||
AUCUACUCAUAU
distance: 0
>ENSG00000078328
UGCAUGU
||||||
UGCAUG
distance: 0
>YOR359W see Aviv,T., Lin,Z., Ben-Ari,G., Smibert,C.A. and Sicheri,F.
(2006) Sequence-specific recognition of RNA hairpins by the SAM domain of
Vts1p. Nat. Struct. Mol. Biol., 13, 168–176. (figure 3a)
UCUUUGA
|||||||
GCUGGUG
distance: 4
>YOR359W
 CUGGCA
|||||||
GCUGGCC
distance: 1
>XB-GENE-6252591 see Lee,B.M., Xu,J., Clarkson,B.K., Martinez
Yamout,M.A., Dyson,H.J., Case,D.A., Gottesfeld,J.M. and Wright,P.E.
(2006) Induced Fit and 'Lock and Key' Recognition of 5 S {RNA} by
Zinc Fingers of Transcription Factor {IIIA}. J. Mol. Biol., 357, 275–291.
(figure 1b left)
CCUGGUUAG
|||||||
GGGUGGG
distance: 5
>XB-GENE-6252591 see Lee,B.M., Xu,J., Clarkson,B.K., Martinez
Yamout,M.A., Dyson,H.J., Case,D.A., Gottesfeld,J.M. and Wright,P.E.
(2006) Induced Fit and 'Lock and Key' Recognition of 5 S {RNA} by
Zinc Fingers of Transcription Factor {IIIA}. J. Mol. Biol., 357, 275–291.
(figure 1b right)
```

```
CCAUAC
|||||||
GGGUGGG
distance: 6
>ENSG00000147274
 UCAAA
||||||
AUCAAA
distance: 0
>ENSG00000134644
UGUAAUAUU
||||||||
UGUAAAUA
distance: 1
>ENSG00000134644 see Gupta,Y.K., Nair,D.T., Wharton,R.P. and
Aggarwal,A.K. (2008) Structures of Human Pumilio with Noncognate {RNAs}
Reveal Molecular Mechanisms for Binding Promiscuity. Structure, 16, 549-
557.
UUUAAUGUU
||||||||
UGUAAAUA
distance: 4
>ENSG00000134644
UGUAUAUA
||||||||
UGUAUAUA
distance: 0
>ENSG00000134644
UGUAAAUA
||||||||
UGUAAAUA
distance: 0
>ENSG00000134644
UGUACAUA
||||||||
UGUACAUA
distance: 0
>ENSG00000134644
UGUACAUC
||||||||
UGUACAUA
distance: 1
>ENSG00000134644
UGUAGAUA
||||||||
UGUAAAUA
distance: 1
>ENSG00000134644
UGUCCAG
||||||||
UGUACAUA
distance: 2
>ENSG00000134644
UGUACAU
||||||||
UGUACAUA
```

```
distance: 0
>ENSG00000134644
UGUAUAU
||||||||
UGUAUAUA
distance: 0
>ENSG00000066044
AUUUU
|||||
AUUUA
distance: 1
>ENSG00000066044
UUUU
|||||
UUUUU
distance: 0
>ENSG00000066044
AUUU
|||||
AUUUA
distance: 0
>ENSG00000112531
ACUAAC
||||||
ACUAAC
distance: 0
>YGL014W
UGUAU
||||||||
UGUAUAUA
distance: 0
>YGL014W
UGUAUAUUA
|||||||||
UGUAUAUUA
distance: 0
>YGL014W
UGUAUAUA
||||||||
UGUAUAUA
distance: 0
>ENSG00000070756
AAAAAAAA
|||||||
AAAAAAA
distance: 0
>ENSG00000070756
AAAAAAA
|||||||
AAAAAAA
distance: 0
>ENSG00000077312
AUUGCACC
|||||||
AUUGCAC
distance: 0
```

```
>ENSG00000077312
AUUGCAC
|||||||
AUUGCAC
distance: 0
>ENSG00000102081 no reference
GCUGC
|||||||
GGACAGG
distance: 4
>YGL044C see Leeper,T.C., Qu,X., Lu,C., Moore,C. and Varani,G. (2010)
Novel Protein-Protein Contacts Facilitate mRNA 3'-Processing Signal
Recognition by Rna15 and Hrp1. J. Mol. Biol., 401, 334-349.
AAUAAU
||||||
UGUUGU
distance: 4
>YGL044C see Leeper,T.C., Qu,X., Lu,C., Moore,C. and Varani,G. (2010)
Novel Protein-Protein Contacts Facilitate mRNA 3'-Processing Signal
Recognition by Rna15 and Hrp1. J. Mol. Biol., 401, 334-349.
UAUAUAUAA
||||||
UGUUGU
distance: 5
>ENSG00000162374
 UAUUUAUUUA
||||||||||
UUAUUUAUUU
distance: 1
>ENSG00000162374
AUUU
|||||||
UUUUUUU
distance: 1
>ENSG00000113742
CUUUA
||||||
UUUUUU
distance: 2
>ENSG00000055917
UGUACAUC
||||||||
UGUACAUA
distance: 1
>ENSG00000055917
UGUAGAUA
||||||||
UGUAGAUA
distance: 0
>ENSG00000055917
UGUAAAUA
||||||||
UGUAAAUA
distance: 0
>YLL013C
UGUAUAUA
```

```
||||||||||
CAUGUAUAUA
distance: 0
>YLL013C
UGUAAAUA
||||||||||
CAUGUAAAUA
distance: 0
>YOL123W
UAUAUAU
|||||||
UAUAUAA
distance: 1
>ENSG00000139910
UCACC
||||||
AUCACC
distance: 0
>ENSG00000139910
 CAGUCAC
||||||||
UCAGUCAC
distance: 1
>WBGENE00011279 see Kuwasako,K., Takahashi,M., Unzai,S., Tsuda,K.,
Yoshikawa,S., He,F., Kobayashi,N., Guntert,P., Shirouzu,M., Ito,T., et
al. (2014) RBFOX and SUP-12 sandwich a G base to cooperatively regulate
tissue-specific splicing. Nat. Struct. Mol. Biol., 21, 778–786. (results
and figure 1a)
 GUGUGC
|||||||
AGCAUGC
distance: 3
>WBGENE00011279
UGCAUGG
|||||||
UGCAUGA
distance: 1
>ENSG00000104967 see Lewis,H.A., Musunuru,K., Jensen,K.B., Edo,C.,
Chen,H., Darnell,R.B. and Burley,S.K. (2000) Sequence-Specific {RNA}
Binding by a Nova {KH} Domain: Implications for Paraneoplastic Disease
and the Fragile X Syndrome. Cell, 100, 323–332. (figure 2 complex 1)
CCUAGAUCACC
||||||
AACACC
distance: 5
>ENSG00000104967
GAUCACC
||||||
AUCACC
distance: 0
>ENSG00000152518
UUAUUUAUU
|||||||||
UUAUUUAUU
distance: 0
>ENSG00000120948
```

```
GUGAAUGA
||||||
GAAUGA
distance: 0
>WBGENE00001402
UGUGUUAUC
|||||||||
UGUGUUAUC
distance: 0
>WBGENE00001402
UGUGCCUUA
|||||||||
UGUGCCAUA
distance: 1
>WBGENE00001402
UGUAA
|||||||||
UGUAAAAUC
distance: 0
>WBGENE00001402
UGUACCAUA
|||||||||
UGUACCAUA
distance: 0
>WBGENE00001402 see Qiu,C., Kershner,A., Wang,Y., Holley,C.P.,
Wilinski,D., Keles,S., Kimble,J., Wickens,M. and Hall,T.M.T. (2012)
Divergence of Pumilio/fem-3 mRNA binding factor (PUF) protein specificity
through variations in an RNA-binding pocket. J. Biol. Chem., 287, 6949-
6957. (see paragraph An Upstream C Is Required for Tight Binding by FBF)
CAUGUGC
|||||||||
UGUGUCAUC
distance: 5
>WBGENE00001402 see Wang,Y., Opperman,L., Wickens,M. and Hall,T.M.T.
(2009) Structural basis for specific recognition of multiple mRNA targets
by a PUF regulatory protein. Proc. Natl. Acad. Sci. U. S. A., 106, 20186-
20191. (figure 2a)
CUGUGC
|||||||||
UGUGCCAUA
distance: 1
>WBGENE00001402 see Wang,Y., Opperman,L., Wickens,M. and Hall,T.M.T.
(2009) Structural basis for specific recognition of multiple mRNA targets
by a PUF regulatory protein. Proc. Natl. Acad. Sci. U. S. A., 106, 20186-
20191. (figure 2b)
 AUAC
|||||||||
UGUAAAAUC
distance: 3
>WBGENE00001402
UGUGUCAUU
|||||||||
UGUGUCAUU
distance: 0
>WBGENE00001402
UGUGC
```

```
|||||||||
UGUGCCAUA
distance: 0
>WBGENE00001402
UGUACUAUA
|||||||||
UGUACUAUA
distance: 0
>ENSG00000048740
UGUU
|||||
AUGUU
distance: 0
>ENSG00000168066
AUACUAACAA
|||||||||||
UAUACUAACAA
distance: 0
>ENSMUSG00000003410
AUUUAUUUU
|||||||
UUUUUUU
distance: 1
>ENSG00000136527
AGAA
|||||
AAGAA
distance: 0
>ENSG00000136527
AGAAC
||||||
AAGAAC
distance: 0
>ENSG00000136450
UGAAGGAC
||||||||
AGAAGGAC
distance: 1
>WBGENE00006321
 GUGUGC
|||||||
AGUGUGA
distance: 1
>WBGENE00006321
GUGUG
|||||||
AGUGUGA
distance: 0
>YLR116W
AUACUAAC
|||||||
UACUAAC
distance: 0
>YLR116W
UACUAACA
|||||||
UACUAAC
distance: 0
>YLR116W
UACUAAC
|||||||
UACUAAC
distance: 0
```