Integration of multi-omics data with artificial intelligence

Department of Toxicogenomics, Maastricht University

Integration of multi-omics data with artificial intelligence: studying the toxic effects on the post-transcriptional regulation

Doctoral Thesis Manuscript

Juan Ochoteco Asensio

7-1-2022

Cover design by Juan Ochoteco Asensio
Layout by Fred Feij
Printed by Proefschriftenprinten.nl

Supervisor

Prof. dr. J.C.S. Kleinjans

Co-supervisor

Dr. F. Caiment

Assessment committee

Prof. dr. H.J.M. Smeets, Full Professor of Clinical Genomics
(Department of Toxicogenomics, Maastricht University)

Prof. dr. J. Quackenbush, Professor of Computational Biology and Bioinformatics
(Department of Biostatistics, Harvard University)

Prof. dr. R. Peeters, Full professor of Mathematics in Knowledge Engineering
(Department of Data Science and Knowledge Engineering, Maastricht University)

Dr. S. Hayat, Group Leader of Translational Data Science
(Uniklinik RWTH Aachen)

Dr. L. Eijssen, Assistant Professor
(Department of Bioinformatics, Maastricht University)

The research described in this thesis was conducted at GROW School for Oncology and Reproduction of Maastricht University.

Chapter 1:

General Introduction

Introduction

In the current age of big data, several technologies are helping generate more and more data for the advancement of science. The extraction of data from the cell is highly relevant, as bodily changes are in the end cell changes, especially in the case of toxicology. Therefore, technologies such as transcriptomics and proteomics are very informative on the changes that a cell suffers due to toxicity. The sheer size of that data is manually impossible to analyze, but the computational improvements have not only improved produce that data but eased its analysis too. This data abundance also favors machine learning algorithms, which benefit from an increasing number of data observations. Those algorithms can find trends in the data, build models that predict them when these are missing, or even classify what is generally too complex to understand.

Sequencing RNA

RNA sequencing (RNA-seq) is a high-throughput sequencing technology that allows the identification and quantification of RNA in a cell or tissue. More than a decade old, this technique has become a mainstream and cheap manner to analyze cells under different conditions at a subcellular dimension. The method consists of 3 main steps: library preparation, sequencing, and data analysis (bioinformatics). Library preparation consists of the combination of all input items (library) required for sequencing. Illumina will be used as an example, being one of the most popular sequencing technologies. First, total RNA is extracted. To improve its sensitivity, mRNA (messenger RNA) enrichment or rRNA (ribosomal RNA) depletion are performed. Afterward, this RNA is fragmented, as the sequencing technology is limited in the number of bases it can sequence. It is later retrotranscribed to cDNA (complementary DNA), a much more stable molecule that can be amplified in a polymerase chain reaction (PCR). The following step is the ligation to sequencing adaptors, which will allow the fragments to be amplified, identified, and sequenced. Once the cDNA material has been amplified by PCR, and both library concentration and fragment lengths have been verified (quality control), the library preparation is complete. These PCR-amplified cDNA fragments can be sequenced by detecting and signaling each of the present nucleotides. The way the cDNA is sequenced depends on the technology used. The three more popular technologies are Illumina, Pacific Biosciences, and Oxford Nanopore.

In the **Illumina** workflow, the cDNA molecules are fastened to a flow cell floor, where sequencing is performed by synthesis of the complementary strand of each of those fragments. The PCR is performed in this case through bridge amplification, where the fragment is bound from both ends to the floor, while the reverse strand is being produced. The resulting reverse fragment will also be replicated, leading to

multiple copies of both forward and reverse sequences of the same fragment (clonal amplification). After that, all reverse fragments will be washed out, for the sequencing signal of that cluster of fragments to be equal across all copies. For the sequencing step, a reverse strand will again be synthesized, but this time using fluorescent nucleotides, which when they successfully pair to one of the fragments, a particular fluorescence will be emitted. Combining all fluorescent signals (50-500 base pairs) results in a read sequence. In the simplest case scenario, each amplified fragment is sequenced once from one of its ends (single-end sequencing). That same fragment may be sequenced from both ends (paired-end sequencing), providing a higher coverage of the sequence.

The **Pacific Biosciences** workflow adds sequencing adaptors that circularize the cDNA fragments. These fragments are added to a sequencing chip, which contains as many nanowells as fragments will be sequenced. At the bottom of such nanowells, an immobile polymerase will synthesize a new strand using fluorescent nucleotides (as in Illumina), which will be detected and generate a read (< 50kb).

For the **Oxford Nanopore** library preparation, molecules (aside from the adapter sequence ligation) are attached to a motor protein. When added to the flow cell, the motor protein will dock to a nanopore (embedded into an electrically resistant membrane) and will move a strand of the cDNA molecule through the pore. The passing of an ionic current through the nanopore will be disrupted differently based on the nucleotide going through the pore. This disruption will be measured in a signal trace, which generates reads between 1-10kb.

Different technologies are characterized by different sequencing read lengths. Although the older and most used transcriptomics technologies are short-read sequencing, they have some limitations. Transcripts expressed from the same or homologous genes tend to present different isoform sequences. Because such transcripts share a pronounced proportion of the sequence, the genome mapping of the reads generated from them (identifying the genomic locus of the original transcript) becomes complex.

The most common objective behind RNA-seq is differential expression analysis. Such analysis aims to identify transcripts or genes that are differentially expressed across different conditions in a statistically significant manner. To do so, the analysis starts with the sequencing results in some form of preliminary data, which needs to be pre-processed. The obtained sequenced fragments need to be linked/mapped to their gene of origin. After all fragments have been located in the genome/transcriptome, a quantification informs about the number of amplified transcripts per gene. These raw quantities, though, are not directly comparable across samples due to several factors (such as the total number of fragments per sample), which leads to the need for normalization methods. Having the data pre-processed and normalized, statistical methods are applied to evaluate how significant the expression level differences between

conditions are. Frequently, the significant results are not taken directly at face value using standard thresholds ($p < 0.05$). Due to the high number of genes (~20000) and transcripts, the expected number of false positives under such a threshold is substantially high (1000). Therefore, multiple testing correction, such as the Benjamini–Hochberg one (or False Discovery Rate- FDR), are applied to the probability values.

The relevance of the RNA changes is generally not related to the intrinsic function of those RNAs, but the effect that those changes have when some of those RNAs (messenger RNAs) are translated to proteins. Protein level changes are important at a cellular level because the function to which those are associated will be affected, be it from an excessive generation of their products to a complete lack thereof. The use of transcriptomics as a proxy for the study of proteins, though, is limited[1]. This is partially due to the complex nature of post-transcriptional regulation. Post-transcriptional regulation refers to systems and conditions that enhance or inhibit the translation of transcripts. Examples of this regulation are diverse: from the structure itself of the transcript (%GC content and poly(A) tail), protein regulatory elements (RNA-binding proteins), the efficiency of translation (number of ribosomes simultaneously bound per transcript), or even other transcripts (such as microRNAs).

MicroRNA

Even though messenger RNAs have a key function as intermediaries for protein synthesis, most transcribed RNAs do not code for proteins. Around ~95% of the total expressed RNA consists of non-coding RNAs (ncRNA) essential for translation (ribosomal RNA and transfer RNA)[2]. Other ncRNAs types, instead, inhibit that process: microRNAs[3]. MicroRNAs (miRNAs) are a class of non-coding RNAs of around 22 nucleotides in length. Their main repressive mechanism in mammals relies on their seed region, which is located between the second and seventh nucleotides. This region presents a sequence complementarity specific to one or more protein-coding RNAs (generally in their 3' UTR). Even though this complementarity is per se highly probable to exist by chance due to the short seed sequence, a high number of the miRNA targets are conserved: they are more frequently found than expected by chance. In addition, miRNAs tend to have several simultaneous targets (> 400 conserved targets per miRNA family).

Although the repressive effect by a single miRNA seed is relatively weak, targets that contain conserved sites tend to present on average 4-5 of them for the same or different miRNAs, which leads to cumulative repression. Even so, conservative sites are less frequent than non-conservative ones. MiRNAs do not exert their function in a stand-alone form and require specific processing steps after transcription. The first maturation step occurs in the nucleus, where the pri-miRNA contains at least one hairpin loop structure. Drosha, an RNA-cutting enzyme, will separate the hairpin structure from the rest of the

transcript, leading to the pre-miRNA stage. After its nuclear export into the cytoplasm, RNase Dicer cuts away the loop connecting both strands of the pre-miRNA, getting a miRNA duplex as a product. Even though both strands can be functional, generally only one will be incorporated into the RNA-induced silencing complex (RISC). In it, a protein from the Argonaute family (Ago2) will bind the miRNA, the latter functioning as a guide to find the complementary target. The Argonaute protein will then also bind the target RNA, which will lead to translation inhibition. There are several suggested mechanisms for how this inhibition may occur. One of them is deadenylation of the target (via a deadenylase complex), leading to the 3' polyA end shortening (mRNA destabilization[4]). This shortening will eventually cause a 5' Cap loss, which enables the 5 to 3' mRNA decay. Thus, miRNA differential expression might affect mRNA target levels. In addition, translational repression mechanisms (decreased rate of mRNA translation into proteins within cells) have also been described, such as Cap-40S initiation inhibition, 60S ribosomal unit joining inhibition, elongation inhibition, and ribosome drop-off. Some studies suggest that the influence on translational efficiency is rather secondary, except for early embryonic states. In that stage, a high correlation exists between translational efficiency and poly(A) tail length, thus the shortening of the 3' end does affect their translational process[5]. Other mechanisms, although also leading to a lower protein output, do not affect translation directly. Examples of these are co-translational nascent protein degradation and sequestration in P-bodies.

In a less common mechanism, some miRNAs exhibit a nearly perfect complementarity of their whole sequence to their target, where the latter is cleaved (primary mechanism in plants). This mechanism presents a high resemblance with silencing RNAs (siRNAs). Some evidence suggests that this miRNA cleavage mechanism might be responsible for the cleavage performed in experimental siRNA treatments, as species with defective Argonaute function present minimal response in such experiments[6].

MiRNA post-transcriptional regulation is important both biologically (development, differentiation, cancer, and disease) and analytically. At the biological level, drastic changes in the expression or function of this RNA type can lead to unforeseen consequences to the transcripts it regulates. At a sequencing analysis level, RNA-Seq interpretation does not consider this regulation, thus opening the possibility of false conclusions due to the miRNA function. An extra layer of complexity is added when taking into account factors that regulate miRNAs themselves. Alongside typical transcript turnover, another novel non-coding transcript might be able to decrease miRNA's inhibitory potential: circular RNAs.

Circular RNA

Messenger RNAs require some maturation steps before their translation: splicing (eliminate introns and join exons together), 5' cap addition, and 3' polyadenylation. Some transcripts mature using alternative

splicing mechanisms, which lead to a different structure: circular RNAs (circRNAs). CircRNAs are a result of back-splicing, a splicing event where both 5` and 3` ends join together, thus forming a back-spliced junction, which leads to a circular structure[7]. These transcripts present neither 5' capping nor polyadenylation, thus alternative library preparations, such as ribo-depletion, are necessary for their identification. The circRNA sequence depends strictly on the alternative splicing event, where the resulting transcript may contain one or several exons, and sometimes even retained introns as well.

The most general hypothesized trigger for circularization is based on the introns flanking the circRNA sequence. The relatively long flanking introns have enriched ALU repeats that can base-pair to each other, leading to the circularization (Figure 1: Circularization event).

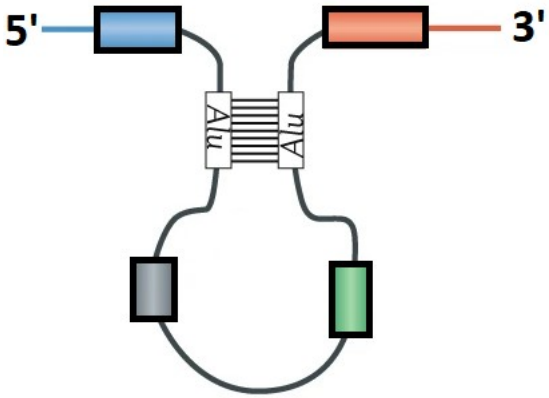


Figure 1: Circularization event

To regulate such circularization, all examples found so far involve RNA-binding proteins (RBP). It has been hypothesized that ADAR proteins (which are RBPs) are involved, as they can mutate the flanking introns[8], weakening and ultimately decreasing the probability of the intron duplex to occur[9,10]. Another mechanism found involves Quaking (QKI), which is also an RBP, but in this case, it favors circularization. It does so when two proteins bind to flanking QKI-binding sites in the linear transcript, and their later dimerization gets both ends of the future BSJ in close contact. In the case of the *Drosophila* muscleblind (mbl) gene[11], the linear transcripts code for the MBL proteins (also an RBP), which in turn stimulate the circularization of its linear transcripts when binding to their specific sites, thus impeding their translation in an autoregulatory or negative feedback system.

As a relatively new discovered type of transcript, several functions have been hypothesized, which are generally related to (post-)transcriptional regulation. Even so, very few examples are known so far that

provide evidence of those functions, which could be classified on whether they induce or inhibit either transcription or translation.

In the group of inhibition, as mentioned before, the formation (in the case of MBL, via protein-binding) of circular RNAs can act as a competing regulator of the linear transcript expression. As generally all of them are transcribed from genes that contain linear isoforms, this effect seems likely. In addition, the ability to function as an RBP sponge could have alternative molecular consequences. Evidence of such a sponging effect is not so straightforward: even though some RBPs have shown a higher cluster density in circRNAs[12], their computational analysis seems to show a lower RBP-binding density[13].

In the inducement function class, a few circular RNAs have been proven to be able to compete as targets for miRNA binding. Some even present copies of the same sequence complementary to a miRNA seed target, allowing a single circular transcript to bind more than one miRNA. When that occurs, circRNAs are said to function as miRNA sponges, adding a new layer of complexity to the post-transcriptional regulation of miRNAs. Again, whether this function is the primordial one is still unclear, as circRNAs are also expressed in organisms that lack miRNA-like repression pathways. In addition, the computational search of circRNAs with enriched seed sites (compared to linear ones) delivered limited additional examples of that function[12,13]. Interestingly, although circRNAs have been, since their discovery, considered a subtype of non-coding RNAs, new evidence via ribosome profiling suggests that they could possess cap-independent protein-coding capabilities[14,15]. It was found that some of them, as some RNA viruses, presented IRES (internal ribosome entry site) activity, which allows eIF3 recruitment (and, thus, translation) independently of the 5'-cap modification[16]. This characteristic could imply that these transcripts could be used as a medium-term source of translation/proteins, as they are more stable (and therefore have a longer half-life) than their linear counterparts; which in turn could help mitigate the current transcriptomics-proteomics gap. Additionally, some circRNAs[17] can regulate their own parental genes' regulation. A subgroup of circRNAs, named exon-intron circRNAs, is located at the nucleus, where they promote the expression of their parental genes by interacting with the U1 small nuclear ribonucleoprotein[18]. Thus, the ability to bind with RBPs might have both transcriptional and post-transcriptional functions. CircRNAs, therefore, might have a role in pre- and post-transcriptional regulation, and because these RNAs can be analyzed via RNA-Seq, their influence on the protein expression level can also be studied.

Proteomics

Proteomics refers to the large-scale study of the "proteome", i.e., the set of all proteins present in a cell, including their expression level, modifications, and interactions. The study of the proteome can be considered more complex than the study of the genome, as the proteome is a consequence of several complex regulatory networks: starting with the genome itself, followed by the intermediary transcriptome, in addition to the post-translational regulation, localization, and possible modifications. Often, instead of proteomics, transcriptomics is used as a proxy to evaluate the molecular changes that could elucidate the phenotypical effects seen at larger systems (on a cellular or tissue level). As already mentioned, RNA levels do not accurately represent either the expression or activity levels of their corresponding proteins, thus the need for proteomics for a more direct approach to the analysis of possible molecular changes[19].

Out of all the different techniques used in proteomics, the most popular one is mass spectrometry (MS), from which other techniques also are derived. Mass spectrometry measures the mass-to-charge ratio of ions. In proteomics, mass spectrometry can be used based on two different methods: top-down or bottom-up proteomics. Top-down proteomics starts with a step of separation and quantitation of the proteins in the sample, followed by MS for identification[20]. In bottom-up proteomics, the sample proteins are instead first enzymatically digested, and then the product peptides are used as input for MS identification[21]. Top-down proteomics is usually used for the identification of a small subset of proteins, while bottom-up has generally a larger sensitivity. In MS, the input peptides are ionized, which can then be split based on their mass-to-charge ratio (m/z). This leads to a mass spectrum, which can be used to identify and quantify proteins. Even so, some technological limitations still exist.

Even though several MS methods have been improving through the years, both reproducibility and repeatability are low across both the same and different MS technologies[22]. Reproducibility, understood as the similarity of results obtained from different technologies, tends to be lower than repeatability, defined as the similarity of results obtained from the same technology, whose wide range comprises from 30-60%[22]. Current proteomics does not generally identify (in a large-scale quantitative manner) post-translational modifications, such as phosphorylation and ubiquitination, which inform of their activity status or their near-future degradation. In addition to that, its sensitivity is greatly limited[23], as the mass spectrometer generally quantifies around 2 to 3 thousand proteins (10-15%), far from the 20,386 manually annotated proteins in humans[24]. Related to sensitivity, dynamic range, referring to the log scale between the most and least abundant proteins, is also one of the metrics aimed to be improved, since the highly expressed proteins tend to completely overwhelm the spectrum to the expense of the low expressed peptides.

In summary, proteomics is one of the technologies that most closely can represent the cell's environment, but the technology itself is limited by its sensitivity. This combination of advantages and disadvantages is curiously the opposite of the ones found in transcriptomics, where the technology allows for high sensitivity while only giving a proxy of the potential protein levels. In addition, considering all the post-transcriptional regulation networks that exist between the two, the ability to take all these considerations simultaneously is manually unpractical. Therefore, the use of computational tools is a must at the aforementioned level of analysis. Among the wide range of bioinformatics approaches, the selection of machine learning algorithms (ML) was preferred due to three main characteristics. First, certain ML models portray exceptional regression capabilities, crucial for the imputation performed in **Chapter 4**. In addition, a selection of ML models is also capable of classification tasks as exceptionally performed in **Chapter 5**. The third characteristic is that most of these models permit the evaluation of the most informative variables/features for the task at hand, which in turn inform the researcher to which degree each feature is related to the desired outcome (as performed in **Chapters 4** and **5**).

Machine learning

Machine learning involves all algorithms that learn or adapt based on the input data[25]. Two of the main groups in which these algorithms can be categorized are supervised and unsupervised. The difference between the two groups relies on whether a target observation is present (supervised[26]) or absent (unsupervised), and therefore their aim changes too: prediction (i.e., whether a compound is toxic) versus grouping (i.e., grouping samples by similarity). Thus, supervised models aim to achieve the highest accuracy or lowest residual values (difference between observed and predicted values).

Supervised machine learning algorithms can also be categorized into two groups based on the nature of the target value: classification and regression. Classification models predict discrete target values (ill or healthy, dog breeds, etc.), while regression ones predict continuous values (height, salary, etc.). Supervised algorithms "learn" from the data by trying to minimize the residual size or error. Building a model to minimize the error in a limited dataset, though, has its limitations. The resulting accuracy can only be attributed to the specific dataset the model has been trained in, usually called the training dataset.

One of the simplest ways to evaluate the accuracy of a model for unseen data is to split the available data into two differing-size datasets (holdout method). The bigger one, which normally consists of around 75 to 80% of the data, is the one used to train the models (training dataset). The complementary small subset is the one used to test the accuracy of the model for unseen data, thus named testing dataset. The model might be tweaked by changing its hyperparameters (parameters that control the learning process) to try to optimize the accuracy of the testing dataset. Again, to avoid bias with such a testing dataset, the dataset

might be subsetted into a third group, the validation set. The latter is used to evaluate the model once the hyperparameters have been optimally set. In this splitting data strategy, there is an inherent potential bias based on which subset is selected as the testing dataset, even when chosen randomly.

For this reason, a more sophisticated method is normally used: *k*-fold cross-validation[27]. In cross-validation, the dataset is also split into k folds, where *k* is the number of folds (subsets) the dataset is divided into. The model is trained using all but one folds (k – 1) and tested on the remaining one (out-of-bag fold). The difference between the previous method and cross-validation is that in the latter, such training and testing are performed with all possible combinations: in each iteration of the loop, one of the k folds will be used as the testing dataset, giving a specific error or metric value. This results in a list of errors/values as large as the number of folds (k errors). Finally, this list is averaged to get a final accuracy metric. Cross-validation, therefore, avoids a possible bias of a specific testing dataset by generating several of them, which gives a more accurate representation of how the model will predict with new input data (unless all the original input data is somehow biased).

Several types of models exist for the prediction based on data learning. Although the number of algorithms keeps increasing at a fast rate, some of the most popular algorithm groups existing in supervised learning can be described based on their prediction strategy:

- **Regression analysis**: statistical methods that try to find the relationship between an independent variable (target to be predicted) and one or more dependent variables (features or variables). The most common one is linear regression[28], in which the optimal solution is found using ordinary least squares. Different subtypes exist based on the nature of such relationship: linear (where the relationship between the dependent and independent variables is fit with a line), polynomial (based on a polynomial relationship of any degree), or logistic regression[29] (used to predict categorical data).

- **Support-vector machines (SVM)**: as a set of methods used commonly for classification (but also for regression), they are defined as a non-probabilistic binary linear classifier[30]. Thus, these methods search to define the optimal line that separates two different classes by maximizing the margin (maximizing the distance between the line and the closest data point from each class). When the separation between two classes is non-linear, SVMs use the kernel trick, in which the data points are transformed to a higher dimension (feature space), where that classification may be performed by the use of a plane (or hyperplane, if the dimensions are above 3).

- **Decision trees**: models that apply decisions on each observation, whose final decision leads to the target value. Each of the decisions (based on the features) are the so-called “branches”, while

each of the final predictions (after all the previous decisions) are named "leaves"[31]. Depending on their target, decision tree models are either classification trees or regression trees. These methods are very popular due to their simplicity and interpretability (the user can straightforwardly visualize how the model reaches a prediction). They are though limited by their tendency to overfit (improve training data metrics at the cost of testing data metrics). Advanced versions such as *random forests* can correct for that, at the cost of their original interpretability.

- **Bayesian networks**: probabilistic models graphed via a directed acyclic graph (DAG)[32]. These models use the probability of random variables, while also taking into account the dependencies and combinations between those probabilities, to predict the probability of an outcome. The outcome with the highest probability is the one predicted to be true. The relationships between the different random variables are graphed using DAGs. In such graphs, the relationship between two vertices (variables) and a single edge (arrow) is understood as the vertex as the end of the edge presents a probability conditioned on the existence of the other vertex, and such a relationship is unidirectional.

- **Artificial neural networks**: groups of models inspired by the biological neurological system, in which neurons are connected via dendrites[33]. In these models, there are at least three layers of neurons or vertices: input, hidden, and output layers. The input neurons contain the input features of the dataset. Each of those observations, before reaching the hidden layer(s), is transformed by the weights and biases. These values are then used in the hidden layer, where a value is extracted based on an activation function chosen beforehand (such as the softplus, rectified linear unit, or sigmoid functions). Again, these values are affected by additional weights before getting to the output layer. In the output layer, the values from the hidden layers are summed together (and corrected by a final bias). Adding additional hidden layers allows fitting data to even more complicated non-linear relationships between inputs and outputs.

- **K-nearest neighbors** (***k*-NN**): in this type of model, which was originally designed for classification, a prediction is made based on how similar the unknown data to predict is to the known training data. Different methods exist to evaluate the neighborhood of a data point, and some do so by clustering, such as Principal Component Analysis or hierarchical clustering. K refers to the number of nearest neighbors to the predicted data point. Those k neighbors are differently classified, and the most popular class in that k-sized population is the one predicted. Regression is also possible, where the prediction results from the average of the k-nearest neighbors values[34].

Aims and outline of the thesis

The aim of this thesis is twofold. First, we want to understand the changes that are a consequence of toxic treatments in the human organism at the molecular level. Namely, the elements of the post-transcriptional regulation responsible for the differences between the transcriptome and proteome at a specific point in time and tissue. Second, and in close relationship to the first aim, we want to further advance the development of data analysis tools that will allow us, and more generally, the experts responsible for risk assessment, to evaluate the effects of any treatment at the molecular level using high-throughput methods such as transcriptomics and proteomics.

Circular RNAs (circRNAs) have recently been shown to be deregulated in different disease scenarios. However, in toxicological conditions, these have still not been well studied. In **Chapter 2**, the expression profiles of several circRNAs affected after cardiotoxicant treatments were analyzed. Was proceeded with an additional validation on 12 differentially expressed circRNAs, and two were selected for further investigation: circCDYL and circGNAS. CircCDYL levels were diminished after anthracycline treatment, which putatively results in de-repression of particular sponged miRNAs. Therefore, the proteins encoded by these miRNA-aimed mRNAs might be lowered. CircGNAS showed a significant cumulative rise across time after being treated with Amiodarone, which was experimentally validated. For both circRNAs, the expression profile of their potential sponged miRNAs and their corresponding mRNA targets were also analyzed. In conclusion, the presence of a regulatory axis involving circCDYL/miR-145-5p/TJP1 was discovered in cardiac cells upon anthracycline treatment, in addition to the validation of the upregulation of circGNAS due to a toxic dose of Amiodarone in different biological replicates, thus indicating the reproducibility of circRNA studies through the use of iPSC-derived cardiomyocytes.

Transcriptomics is at the moment frequently applied as an analytical apparatus to examine the extent of cell expression alterations between two phenotypes or between different conditions. Nevertheless, a substantial section of the significant changes detected in transcriptomics at the gene level is typically not reliably identified at the protein level by proteomics. This weak correlation between the measured transcriptome and proteome is perhaps largely due to post-transcriptional regulation, among which miRNA and circRNA have been suggested to play an important role. Hence, since both miRNA and circRNA are also measured by transcriptomics, a model was proposed to be built in **Chapter 3**. It takes factors related to the post-transcriptional regulation into account to estimate, for each transcript, the fraction of transcripts that would be available for translation. Using a dataset of cells exposed to diverse compounds, the model was evaluated to observe how and whether it was able to improve the correlation between the assessed transcriptome and proteome expression level. The results showed that the model

improved the correlation for a subgroup of genes, possibly due to the regulation of several miRNAs across the genome.

Proteins are often deemed the main biological component in charge of the different functions and structures of a cell. Nevertheless, proteomics, the global study of all proteins which is frequently executed by mass spectrometry, is restricted by its stochastic sampling and can only measure a limited number of proteins per sample. Transcriptomics, which permits an extensive analysis of all expressed transcripts, is regularly used as a surrogate. However, the transcript quantity does not present a high degree of correlation with the subsequent protein quantity, notably due to the existence of several post-transcriptional regulatory mechanisms. In **Chapter 4** is hypothesized that the absent protein values in proteomics could be calculated using machine learning regression techniques, trained with numerous variables extracted from transcriptomics, including previously identified translational regulatory elements such as microRNAs and circular RNAs, among others. After taking into account different machine learning algorithms applied to two different splitting approaches, the random forest algorithm was reported to be able to predict proteins in new samples out of several omics data with good accuracy.

In next-generation transcriptomics, differential expression analysis is established as one of the main approaches to assess the outcomes of two biological conditions on the gene expression of different biological samples. Nevertheless, the current statistical roadmaps offer very mild standardized filters for the selection of differentially expressed genes, leading to a substantial number of false positives. Authors typically incorporate their specific arbitrary thresholds, often reliant on their criteria and prospects of the quantity of differently expressed genes to be attained. This leads to the inclusion of statistically significant genes with expression profiles that, if individually examined, would not be considered biologically relevant to study further. In **Chapter 5**, we focused on developing AutoRel, a machine learning model that incorporates not only the most conventional statistical assessments but also all the complexities that distinguish biologically relevant changes based on manual examinations. AutoRel, which categorizes each evaluated gene into “relevant”, “irrelevant” or “dubious”, informed of the most crucial variables for the selection of relevant genes. The value of the number of replicates on the performance of the model was assessed, through the use of simulated datasets and the biological interpretation of the chosen genes.

In **Chapter 6**, a general discussion of chapters 2 to 5 is presented. First, the discussion begins with a reflection on the dataset used for all chapters, HeCaToS, including its weaknesses and potential improved designs. Following the order of the chapters, the discussion continues with Chapter 2, focusing on the new quantification method used and suggestions for future validation experiments. It continues with the difficulties of formulating the complexity of post-transcriptional regulation found in Chapter 3. It

connects with one of its alternatives, the use of machine learning, which led to the work and later publication of Chapter 4. The latter Chapter is also discussed, including its limitations regarding the use of very stable conditions. The use of machine learning in Chapter 4 goes hand in hand with Chapter 5, where there is an assessment of alternative methods to generate the training datasets. The General Discussion ends with a short conclusion paragraph summarizing the work performed.

Lastly, an impact paragraph describes the potential future uses of the work presented in this manuscript.

Bibliography

1. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between Protein and mRNA Abundance in Yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).

2. Westermann, A. J., Gorski, S. A. & Vogel, J. Dual RNA-seq of pathogen and host. *Nature Reviews Microbiology* vol. 10 618–630 (2012).

3. Baek, D. *et al.* The impact of microRNAs on protein output. *Nature* **455**, 64–71 (2008).

4. Eichhorn, S. W. *et al.* MRNA Destabilization Is the dominant effect of mammalian microRNAs by the time substantial repression ensues. *Mol. Cell* **56**, 104–115 (2014).

5. Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508**, 66–71 (2014).

6. Bartel, D. P. MicroRNAs: Target Recognition and Regulatory Functions. *Cell* vol. 136 215–233 (2009).

7. Barrett, S. P. & Salzman, J. Circular RNAs: Analysis, expression and potential functions. *Dev.* **143**, 1838–1847 (2016).

8. Nishikura, K. Functions and regulation of RNA editing by ADAR deaminases. *Annual Review of Biochemistry* vol. 79 321–349 (2010).

9. Ivanov, A. *et al.* Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. *Cell Rep.* **10**, 170–177 (2015).

10. Rybak-Wolf, A. *et al.* Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Mol. Cell* **58**, 870–885 (2014).

11. Houseley, J. M. *et al.* Noncanonical RNAs from transcripts of the Drosophila muscleblind gene. *J. Hered.* **97**, 253–260 (2006).

12. Guo, J. U., Agarwal, V., Guo, H. & Bartel, D. P. Expanded identification and characterization of mammalian circular RNAs. *Genome Biol.* **15**, 409 (2014).

13. You, X. *et al.* Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. *Nat. Neurosci.* **18**, 603–610 (2015).

14. Pamudurti, N. R. *et al.* Translation of CircRNAs. *Mol. Cell* **66**, 9-21.e7 (2017).

15. Legnini, I. *et al.* Circ-ZNF609 Is a Circular RNA that Can Be Translated and Functions in Myogenesis. *Mol. Cell* **66**, 22-37.e9 (2017).

16. Meyer, K. D. *et al.* 5′ UTR m6A Promotes Cap-Independent Translation. *Cell* **163**, 999–1010 (2015).

17. Hansen, T. B. *et al.* miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. *EMBO J.* **30**, 4414–4422 (2011).

18. Li, Z. *et al.* Exon-intron circular RNAs regulate transcription in the nucleus. *Nat. Struct. Mol. Biol.* **22**, 256–264 (2015).

19. Cox, J. & Mann, M. Is Proteomics the New Genomics? *Cell* vol. 130 395–398 (2007).

20. Kelleher, N. L. Peer Reviewed: Top-Down Proteomics. *Anal. Chem.* **76**, 196 A-203 A (2004).

21. Chait, B. T. Mass spectrometry: Bottom-up or top-down? *Science* vol. 314 65–66 (2006).

22. Tabb, D. L. *et al.* Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **9**, 761–776 (2010).

23. Westont, A. D. & Hood, L. Systems Biology, Proteomics, and the Future of Health Care: Toward Predictive, Preventative, and Personalized Medicine. *Journal of Proteome Research* vol. 3 179–196 (2004).

24. reviewed:yes AND organism:"Homo sapiens (Human) [9606]" in UniProtKB. https://www.uniprot.org/uniprot/?query=*&fil=organism%3A%22Homo+sapiens+%28Human%29+%5B9606%5D%22+AND+reviewed%3Ayes.

25. Tom Michael Mitchell. Machine Learning. *McGraw-Hill Education* 414 https://books.google.nl/books?id=xOGAngEACAAJ&dq=isbn:0070428077&hl=nl&sa=X&redir_esc=y (1997).

26. Brewka, G. Artificial intelligence—a modern approach by Stuart Russell and Peter Norvig, Prentice Hall. Series in Artificial Intelligence, Englewood Cliffs, NJ. *Knowl. Eng. Rev.* **11**, 78–79 (1996).

27. Xu, Y. & Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J. Anal. Test.* **2**, 249–262 (2018).

28. Schneider, A., Hommel, G. & Blettner, M. Lineare regressionsanalyse - Teil 14 der serie zur bewertung wissenschaftlicher publikationen. *Deutsches Arzteblatt* vol. 107 776–782 (2010).

29. Nick, T. G. & Campbell, K. M. Logistic regression. *Methods in molecular biology (Clifton, N.J.)* vol. 404 273–301 (2007).

30. Noble, W. S. What is a support vector machine? *Nature Biotechnology* vol. 24 1565–1567 (2006).

31. Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A. & Brown, S. D. An introduction to decision tree modeling. *J. Chemom.* **18**, 275–285 (2004).

32. Friedman, N., Geiger, D. & Goldszmidt, M. Bayesian Network Classifiers. *Mach. Learn.* **29**, 131–163 (1997).

33. Abbod, M. F., Catto, J. W. F., Linkens, D. A. & Hamdy, F. C. Application of Artificial Intelligence to the Management of Urological Cancer. *Journal of Urology* vol. 178 1150–1156 (2007).

34. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **46**, 175–185 (1992).

Chapter 2:

Effect of cardiotoxicants on circRNAs and their function in post-transcriptional regulation

Effect of cardiotoxicants on circRNAs and their function in post-transcriptional regulation

Juan Ochoteco Asensio[1], Jelmer Faber[1], Jos Kleinjans[1], Twan v.d. Beucken[1], Florian Caiment[1]

[1]Department of Toxicogenomics, School of Oncology and Developmental Biology (GROW), Maastricht University, Maastricht, The Netherlands

(Manuscript under preparation)

1. Abstract

Circular RNAs (circRNAs) have recently been shown to be deregulated in different disease scenarios. However, in toxicological conditions, these have still not been well studied. In this study, we analyzed the expression profiles of several circRNAs affected after cardiotoxicant treatments. We proceeded with additional validation on 12 differentially expressed circRNAs, and we selected two for further investigation: circCDYL and circGNAS. CircCDYL levels were decreased after anthracycline treatment, which hypothetically results in de-repression of specific sponged miRNAs. Consequently, the proteins encoded by these miRNA targeted mRNAs might be reduced. CircGNAS presented a significant cumulative increase across time after being treated with Amiodarone, which we were able to validate experimentally. For both circRNAs, we also analyzed the expression profile of their potential sponged miRNAs and their corresponding mRNA targets. In summary, we discovered the presence of a regulatory axis involving circCDYL/miR-145-5p/TJP1 in cardiac cells upon anthracycline treatment, in addition to validating the upregulation of circGNAS due to a toxic dose of Amiodarone in different biological replicates, thus demonstrating the reproducibility of circRNA studies by using iPSC-derived cardiomyocytes.

2. Introduction

Circular RNAs (circRNAs) are a subtype of non-coding RNA with a circular structure, created by covalent binding of the 3′ and 5′ ends of the RNA molecule[1]. Although circRNAs were firstly discovered in other species several decades ago, their existence and potential functions in humans have been only

recently explored[2]. Around 140,000 potential circular RNAs have been predicted in human cells and deposited in databases such as circBase[3], though many of them tend to be tissue-specific[4]. Interestingly, circRNAs are hypothesized to be stable due to their closed structure, protecting them from exonuclease activity. Although circRNA functions are still poorly understood, one of the most studied functions is their ability to act as miRNA sponges. MiRNAs are another class of non-coding RNAs characterized by a small sequence length (~22 base pairs), whose main function is to bind to messenger RNAs (mRNAs) via base pair complementarity, decreasing the translation efficiency of the latter. In this context, miRNA sponging refers to the capability of some circRNAs to present one or more miRNA binding site, thus sequestering (a.k.a. "sponging") those miRNAs, alleviating the translational inhibition of the coding RNAs.

Recently, a limited number of research articles have been published concerning circular RNAs in the context of cardiac toxic reactions. Although some focus on how these are regulated by RNA-binding proteins[5], most are focused on their miRNA sponging capabilities[6-8]. The upregulation[6,8] or downregulation[7,9] of an individual or multiple circRNAs have been hypothesized to be related to the toxic effects of cardiotoxicant compounds. To this effect, these studies tended to focus on the relationship between the expression of the circRNAs, expression of the miRNAs, and potential toxic effects. Even so, some additional variables are considered to be primordial to understand the regulatory functions of these circRNAs, such as the expression of the gene of origin of these circRNAs and the effect at the proteomics level result of the combination of the different molecular factors.

For the detection or identification of circular RNAs in the RNA-Sequencing data, bioinformatics tools such as CIRI-AS[10] and FUCHS[11] identify back-spliced junctions (BSJs), allowing the closure of the circular structure, which are characteristic of these transcripts. CircExplorer2[12], while including changes to further discover circularized transcripts in both known and *de novo* assembled annotations, also relies on the discovery of BSJs. Two main issues arise from such methodologies: first, the number of circular RNAs found relies on the number of reads that span the BSJ, and because a read spanning a BSJ is a relatively rare event, the probability to find them is highly dependent on the sequencing depth of the experiment. Second, the *de novo* identification of these BSJs and their corresponding circular RNAs leads to either random identification names or naming based on genome location, which in turn also relies on the specific genome version used. This complicates the association of these transcripts with their circular RNA databases.

To study the potential role of circular RNAs in eliciting cardiotoxicity, RNA-Sequencing (RNA-Seq), Small RNA Sequencing, and mass spectrometry were performed on induced Pluripotent Stem Cell-derived cardiac cells after exposure to several known cardiotoxicants. We were able to identify and

quantify circular RNAs with a transcriptome annotation file derived from the Ensembl and circBase databases. After the quantification, a list of circRNAs that were differentially expressed after treatment was obtained for each comparison. Among these, we investigated two specific circRNAs in the lab. First, circCDYL, whose main "sponged" miRNAs were found to be differentially expressed. Following the post-transcriptional network, a fraction of the genes targeted by these miRNAs were also transcriptionally and significantly affected. TJP1, a protein coded by one of the latter genes, showed a time-dependent increase for TJP1 at the protein level. Lastly, we studied circGNAS, for which we were able to experimentally validate its significant increase, and we analyzed the related miRNAs, coding transcript targets, and proteomics.

3. Methods

3.1. Samples

The analyzed data consists of 3D microtissues containing stem-cell-derived cardiomyocytes and fibroblasts in a 4:1 ratio from InSphero. These microtissues were exposed to eight compounds: Fluorouracil (5FU), Amiodarone (AMI), Celecoxib (CEL), Docetaxel (DOC), Mitoxantrone (MXT), Paclitaxel (PTX), Doxorubicin (DOX), and Epirubicin (EPI), plus a fluctuating DMSO control (DF2). The dosing profile was established via the use of the physiologically based pharmacokinetic (PBPK) modeling software PK-Sim, to simulate exposure levels under physiological conditions[13]. For each compound, there were 2 doses: Therapeutic and Toxic. The exposures were done in triplicates, and the data extraction was performed at 8-time points: 0, 2, 8, 24, 72, 168*, 240*, and 336* hours; resulting in 24 data points per dose (except for Doxorubicin and Epirubicin, which did not include a 0h timepoint). * The marked time points were not available for all compounds, as some of them were so toxic that insufficient biological material remained.

3.2. RNA Sequencing (RNA-Seq)

Total RNA from the exposed microtissues was isolated using the Qiagen AllPrep Universal Kit (Cat #80224). Ribo-depletion was achieved by using the Illumina RiboZero Gold kit (Cat #MRZG12324), and the libraries were prepared using the Lexogen SENSE total RNA kit (Cat #009.96). All libraries were then sequenced on an Illumina HiSeq 2000 at 100 bp paired-end at an average sequencing depth of 21 million reads (after Salmon quantification). The adaptors were removed through Trimmomatic version 0.33[2]. We used the following parameters: paired-end, ILLUMINACLIP: TruSeq3-PE.fa:2:30:10, LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, MINLEN:36, HEADCROP:12.

3.3. (non-)coding RNA quantification and analysis

3.3.1. CircRNA prediction with CIRI2

We used CIRI2[14] to predict and quantify potential circular RNAs in our sequencing data. Only circular RNAs with at least 2 reads mapping to the back-splice junction (BSJ) were quantified. The total reads for each circular RNA were obtained by adding both BSJ reads and non-BSJ reads.

3.3.1.1 Differential Expression Analysis with data quantified with CIRI2

Samples that presented a relatively low sequencing depth (less or equal to 25% of the average, ~5 million mapped reads) in comparison with their in-group levels were excluded. We used as a control group for all comparisons both the fluctuating DMSO samples and the timepoint 0h samples available (on each compound experiment). The differential expression analysis was performed by DESeq2 (version '1.24.0')[15]. We considered circular RNAs with a p. adjusted value < 0.05 significant.

3.3.2. CircRNA quantification with Salmon & circBase

The genome version used for all the transcriptomics analyses was the Genome Reference Consortium Human Build 38 (GRCh38.p12). For the circRNA quantification, we used as a reference transcriptome the hg19 circRNAs putative spliced sequence from circBase (Jul 2017 update)[3].

We combined the transcriptomic libraries for both coding (all cDNA) and non-coding (all ncRNA) transcripts from Ensembl (release 92, April 2018 Ensembl Archive)[16] with the library for mature circRNA sequences from circBase into a single library, which we set as the global transcriptome reference to Salmon[17]. We then used Salmon to quantify the RNA-Seq data.

3.3.2.1 Differential Expression Analysis with data quantified with Salmon

We also utilized DeSeq2 as mentioned in 3.3.1.1. Afterward, we filtered the results according to several thresholds. First, we required a p. adjusted value < 0.01 to filter by significance. Second, the median of all samples needed to be non-zero, thus requiring that at least half of all samples expressed such molecule. Third, the first quantile of both control & treatment groups also needed to be non-zero, which requires that at least 75% of the samples in each group present an expression. Fourth, for a group to be considered significantly over-expressed against another, its first quantile needed to be higher than the third quantile of the other group, thus requiring that at least 75% of the first group expression must be higher than the 75% of the other. The previously mentioned filters were inspired by the R-ODAF workflow[18]. Finally, the difference between the 1/7 (14%) quantile and 6/7 (86%) quantile of the control group had to be smaller than its median, therefore filtering out molecules that are highly sensitive to batch effects.

After grouping all the filtered results for all comparisons, we ranked the circular RNAs based on their median expression in decreasing order, and subset the first 100 (the top 100 expressed). The changes were then examined manually, selecting those that expressed both a very significant differential expression and/or a potential time-dependent increase or decrease. That examination resulted in a sub-selection of 10 circular RNAs (circBase IDs): hsa_circ_0010791, hsa_circ_0026129, hsa_circ_0034356, hsa_circ_0055922, hsa_circ_0060999, hsa_circ_0076194, hsa_circ_0078905, hsa_circ_0090448, hsa_circ_0090904, and hsa_circ_0102325. On top of those 10, we also added two circular RNAs known to be expressed in iPSC-derived cardiomyocytes for further validation: circCDYL (differentially expressed in CIRI2 quantification) and circSMARCA5[19].

3.4. Small RNA Sequencing and Quantification

Starting from the same total RNA isolated for the ribo-depleted libraries, an aliquot was size selected and ligated using the TruSeq Small RNA Library Prep Kit (Illumina®). After sequencing on the HiSeq 2500 at an average of 3.6 million reads per sample (after quantification), we quantified the resulting data using miRge2 (last change: 05/06/2018)[3]. miRge2 used the MiRBase database as the reference library (miRBase v22), IsomiRs were not considered for this analysis.

For miRNA data, we also used DeSeq2 as mentioned in 3.3.1.1. We also evaluated as significant those circular RNAs with a p. adjusted value < 0.05.

3.5. MiRNA-gene interactions

To identify which genes were targeted by miRNAs, we decided to source such information from the TargetScan predicted targets (version 7.1)[20,21]. We evaluated the importance of the inhibition based on the 'Cumulative weighted context++ score', which is the value used by default to sort those targets.

3.6. MiRNA-circRNA table

For the prediction of miRNA-circRNA interactions, we used miRanda (version 3.3a, strict condition activated)[22] between human miRNA sequences from miRBase (version 22)[23] and human circRNA sequences from circBase (Jul 2017 update)[3]. This led to 21 million possible interactions between all the molecules, of which around 10 million were unique between both RNA types.

3.7. Experimental validation (qPCR)

3.7.1. Cell culture

Induced Pluripotent Stem Cells (iPSCs, CARIM001A) were maintained on Geltrex (Gibco, Thermo Fisher Scientific) coated 6-well tissue culture plates containing 2 ml of E8 Flex culture (Gibco, Thermo

Fisher Scientific) media. Geltrex was diluted at 1:100 using DMEM:F12 (Gibco, Thermo Fisher Scientific). Plates were incubated with 5% CO2 at 37ºC and cells were passaged when they reached approximately 70% confluency or when spontaneous differentiation was observed in the center of the colonies. For passaging, cells were washed once with magnesium/calcium-free Dulbecco's phosphate-buffered saline (DBPS, Gibco, Thermo Fisher Scientific) and subsequently treated with 50 mM EDTA diluted in magnesium/calcium-free DPBS and incubated at 5% CO2 at 37ºC until colony edges started to lift. To remove and break up colonies into smaller clumps, iPSCs were washed off the plate using 1 ml of RT E8 Flex containing 10 µM of Y-27632 ROCK inhibitor. iPSCs were subsequently plated on GelTrex-coated 6-well tissue culture plates and maintained in 1.5 ml of E8 Flex media supplemented with 10 µM of Y-27632 for the first 24h to aid cell adhesion and survival. After 24h the media was changed to 2 ml of E8 Flex media.

SV40 immortalized human cardiomyocytes (ABMGood) were cultured in RPMI-1640 (Gibco, thermo Fisher Scientific) supplemented with 10% FBS (supplier) and grown on applied extracellular matrix (ABMGood) diluted 1:10 in 20 mM acetic acid. Cells were passaged when reaching approximately 90% confluency. To passage, cells were washed once with DBPS and subsequently incubated with 40 µl/cm2 accutase (Corning) for 5 minutes 5% CO2 at 37ºC.

3.7.2. Cardiomyocyte differentiation

Generation of cardiomyocytes from iPSCs was performed based on a protocol adapted from the one published by Lian et al[24]. When iPSCs reached approximately 70% confluence, cells were dissociated in RT E8 Flex supplemented with 10 µM of Y-27632. Cell clumps were plated on Geltrex-coated 12-well tissue culture plates and maintained in 1 ml of E8 Flex media supplemented with 10 µM of Y-27632 for the first 24h at a density of 140,000 cells per well. After 24h the cell culture media was changed for 1.5 ml of E8 Flex media. Once cells reached approximately 70 percent confluence (usually after 3 days), the culture media was changed for 1.5 ml of RPMI-1640 + GlutaMAX (Gibco, Thermo Fisher Scientific), supplemented with B27 minus insulin (Gibco, Thermo Fisher Scientific) and 10 µM of CHIR99021 (Tocris Bioscience) for 24h. After 24h of exposure, the media was changed to RPMI-1640 + GlutaMAX supplemented with B27 minus insulin (differentiation day 0). On the third day after initial exposure to CHIR99021, a combined media was prepared by collecting 50% of media from each well of the 12-well tissue culture plate and adding 50% of fresh RPMI-1640 + GlutaMAX supplemented with B27 minus insulin. IWP2 (Tocris Bioscience) was added to a final concentration of 5 µM. Left-over media was aspirated from the cells and the combined media was added. 5 days after initial exposure to CHIR99021, the combined media was removed and replaced with fresh RPMI-1640 + GlutaMAX supplemented with B27 minus insulin. On day 7, the media was replaced with fresh RPMI-1640 + GlutaMAX supplemented

with B27 plus insulin (Gibco, Thermo Fisher Scientific) and replaced every two days. On Day 14 a cardiomyocyte selection medium was prepared by adding L-(+)-Lactic acid (Santa Cruz Biotechnology) to RPMI-1640 without glucose (Gibco, Thermo Fisher Scientific) to a final concentration of 4 mM, and 2 ml were added to the cardiomyocytes after washing twice with magnesium/calcium-free DBPS. After three days the media was replaced with fresh RPMI-1640 + GlutaMAX supplemented with B27 plus insulin and replaced every two days until cardiomyocytes were used in experiments.

3.7.3. Validation of Amiodarone Toxic dose effect on circGNAS

Cardiomyocytes were washed three times with RT magnesium/calcium-free DPBS before 1 ml per well of pre-warmed TrypLe (Gibco, Thermo-Fisher Scientific) supplemented with 0.5 U/ml Liberase (Roche) and incubated at 5% CO2 at 37ºC for 5 minutes. Cell clumps were dissociated by repetitive pipetting followed by another 5 minutes of incubation at 5% CO2 at 37ºC. Cell clumps were again dissociated by repetitive pipetting and pelleted at 250 x *g* for 2.5 minutes. The supernatant was discarded and the pellet was redissolved in fresh RPMI-1640 + GlutaMAX supplemented with B27 plus insulin supplemented with 10 µM of Y-27632. Cells were seeded on Geltrex-coated 48-well tissue culture plates and maintained in 1 ml of E8 Flex media supplemented with 10 µM of Y-27632 for the first 24h at a density of 300.000 cells per well. Cells were left over the weekend before Amiodarone treatment commenced.

Amiodarone hydrochloride was aseptically dissolved to a stock concentration of 10 mM in DMSO (Sigma-Aldrich) and stored at –20 ºC until needed. Amiodarone dilutions were prepared using pre-warmed RPMI-1640 + GlutaMAX supplemented with B27 plus insulin and cells were washed once with RT PBS before 500 µL of amiodarone-containing medium was added to the cells.

3.7.4. RNA extraction and cDNA synthesis

After 24h of exposure, the culture medium was discarded and 500 µL of Qiazol (Qiagen) was added to each well to lyse the cells. RNA extraction was carried out according to the manufacturer's protocol and RNA yield and quality were assessed using a Nanodrop 1000 spectrophotometer (Thermo-Fisher Scientific). cDNA was synthesized using the iScript cDNA Synthesis Kit (Bio-Rad Laboratories) according to the manufacturer's protocol.

3.7.5. CircRNA primer design

cDNA FASTA sequences were extracted for circRNAs of interest and primers were designed such that the BSJ (back-spliced junction) would be amplified. This was done so by designing a primer at each end of the sequence, and each of them was directed to the end closest to them. In this manner, only transcripts that contain a BSJ between the start and the end of that sequence will be amplified. Primer sequences can

be found in supplementary table 1. The samples used for validation by qPCR were not the ones used for the bioinformatics analysis.

3.7.6. SYBR Green RT-qPCR Validation

Real-time quantitative PCR (RT-qPCR) reactions were run in triplicates using the CFX Connect Real-Time PCR Detection System (Bio-Rad laboratories) and data were collected with the Bio-Rad CFX manager software (v3.1, Bio-Rad laboratories). cDNA amplification was measured using the IQ Sybr Green Supermix kit (Bio-Rad laboratories) and the abovementioned custom-designed probes. Expression levels were calculated using the 2(-Delta Delta C(T)) method with β-Actin as endogenous control.

3.8. CIRI2 quantification results of validated circRNAs

To investigate whether the validated circRNAs were also quantified by CIRI2, we retrieved the spliced sequence from circBase for each validated circRNA. Moreover, we used BLASTN (Nucleotide-Nucleotide BLAST 2.12.0+)[25] to search the location of those sequences in our genome FASTA file. To filter across possible homolog sequences in different chromosomes, we only retrieved those coordinates that matched the chromosome where the circRNA locus was present (according to circBase). We pooled together all CIRI2 quantification results, searching for predicted circRNAs located in the same loci. In that regard, only predicted circRNAs that matched the chromosome number, circRNA start (locus of the first nucleotide), circRNA end (locus of the last nucleotide), and strand with the BLASTN results were considered both experimentally validated and quantified by CIRI2.

3.9. Additional differential expression analyses

As performed with the miRNA data, we utilized the DESeq2 pipeline with a threshold of p. adjusted value < 0.05 to analyze the potential differential expression of both genes and (linear) transcripts.

3.10. Proteomics analysis

Protein samples were isolated and diluted to a concentration below 0.2M. Subsequently, they were subject to digestion by trypsin and then cleaned up with the usage of Sep-Pak tC18 cartridges (Waters) according to the manufacturer's instructions. The samples were then dried using a vacuum centrifuge and quantified by the usage of an Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific), coupled to a NanoLC-2D HPLC system (Eksigent). The raw MS data processing was performed with Genedata Expressionist software (v.11.0). Noise-reduction and

normalization were applied to the LC-MS peaks, whose properties were later collected (m/z and RT boundaries, m/z and RT center values, intensity). The annotation of the individual MS/MS spectra was obtained using Mascot 2.6. The peak clusters were grouped using protein interference (peptide and protein annotations), and the protein intensities obtainment was performed with the Hi3 method. Proteomics data were normalized using a modified median of medians approach[26].

4. Results

To assess the putative effect of toxicants on circRNAs expression, we quantified relevant elements to the circRNA post-transcriptional regulation (circRNAs, linear RNAs at both transcript and gene levels, microRNAs, and proteins) in both therapeutic and toxic doses of eight compounds (5FU, AMI, CEL, DOC, DOX, EPI, MXT & PTX) across 7 timepoints (2, 8, 24, 72, 168, 240, and 336h) using omics technologies.

To categorize circRNAs in our ribo-depleted data, we used two different methodologies. First, we predicted *in silico* the presence of circRNAs using CIRI2, which predicts and quantifies circular RNAs from the transcriptomics input data based on their back-spliced junction (BSJ). Alternatively, we mapped and quantified all sequenced reads from the ribo-depleted libraries on circRNA sequences extracted from circBase (remapping). We thus describe below the results found in both.

4.1. Differentially expressed circular RNAs

4.1.1. In silico prediction and quantification of circRNAs using CIRI2

We performed a circRNA in silico prediction on all treatment samples and assessed their differential expression using as controls all timepoint 0h samples added to control DMSO samples. Depending on the treatment, this analysis detected between 11 to ~16 thousand unique circRNAs (Table 1). As expected, the number of constitutive circRNAs (circular RNAs quantified in all samples in the comparison) was very low. The number (and proportion) of **D**ifferentially **E**xpressed **C**ircRNA**s** (**DECs**) was the highest in the anthracycline treatments (Doxorubicin and Epirubicin). Even so, the percentage of samples where the DECs were detected was on average between 22 and 10% across all treatments, which was substantially low considering that most DECs were downregulated (higher expression in control samples). In other

words, the majority of differentially expressed circRNAs were not detected in either treatment or control groups for most of the samples.

Table 1: Summary of the CIRI2 and DESeq2 results. 'N quantified circRNA' refers to the number of unique circRNAs that were detected by CIRI2 in any sample involved in the comparison between that treatment and the control. 'N quantified constitutive circRNAs' refers to the number of unique circRNAs that were detected by CIRI2 in all samples involved in the comparison between that treatment and the control. 'N DECs' refers to the number of Differentially Expressed CircRNAs (DECs) in the specific treatment when compared to the control samples (p. adjusted value < 0.05). '% DECs' refers to the percentage of DECs taking the 'N quantified circRNA' as the total/denominator. 'Avg % samples a DEC is quantified' refers to the percentage of samples that detect a circRNA on average when it is differentially expressed (i.e., if the comparison in Epi_The spanned a total of 100 samples, and we analyzed the quantification values for an average DEC, only 18 samples out of the 100 would detect that circular RNA).

Treatment	**N quantified circRNA**	**N quantified constitutive circRNAs**	**N DECs**	**% DECs**	**Avg % samples a DEC is quantified**
Epi_The	13356	1	1135	8.50%	18.37%
Dox_The	11480	1	892	7.77%	19.71%
Dox_Tox	11201	1	764	6.82%	22.04%
Epi_Tox	11713	2	559	4.77%	21.23%
CEL_The	11002	7	391	3.55%	15.61%
MXT_Tox	11553	6	377	3.26%	13.81%
PTX_Tox	11359	6	351	3.09%	14.67%
CEL_Tox	11243	6	303	2.70%	14.10%
PTX_The	11609	8	294	2.53%	12.66%
MXT_The	12256	8	277	2.26%	13.29%
DOC_The	11916	7	248	2.08%	13.94%
DOC_Tox	13595	8	212	1.56%	12.58%
5FU_Tox	14532	10	173	1.19%	10.07%
AMI_The	13596	6	160	1.18%	13.12%
5FU_The	13950	9	150	1.08%	11.01%
AMI_Tox	15695	10	138	0.88%	10.18%

To further investigate this high inconsistency in DEC quantification, we pooled the quantification of all the samples and analyzed the overlap in circRNA detection across all samples. The consequent results (Figure 1) showed that 59% of all detected circRNAs (median = 1) were only quantified in a single sample, 90% were quantified in 8 samples or less, and only one circRNA (circRMRP, or hsa_circ_0001853) was quantified in all 309 samples.

4.1.1.1 circCDYL (hsa_circ_0008285)

Across the differential expression results, we searched for some of the most recently studied circRNAs. Among them, we found circCDYL differentially expressed in 3 anthracycline treatments: the therapeutic (Figure 1) and toxic dose of Doxorubicin, and the therapeutic dose of Epirubicin. All three treatments showed strong down-regulation of circCDYL expression (log2 fold change < -2.5). The sequence predicted by CIRI2 for circCDYL matched perfectly to hsa_circ_0008285[27] from circBase.

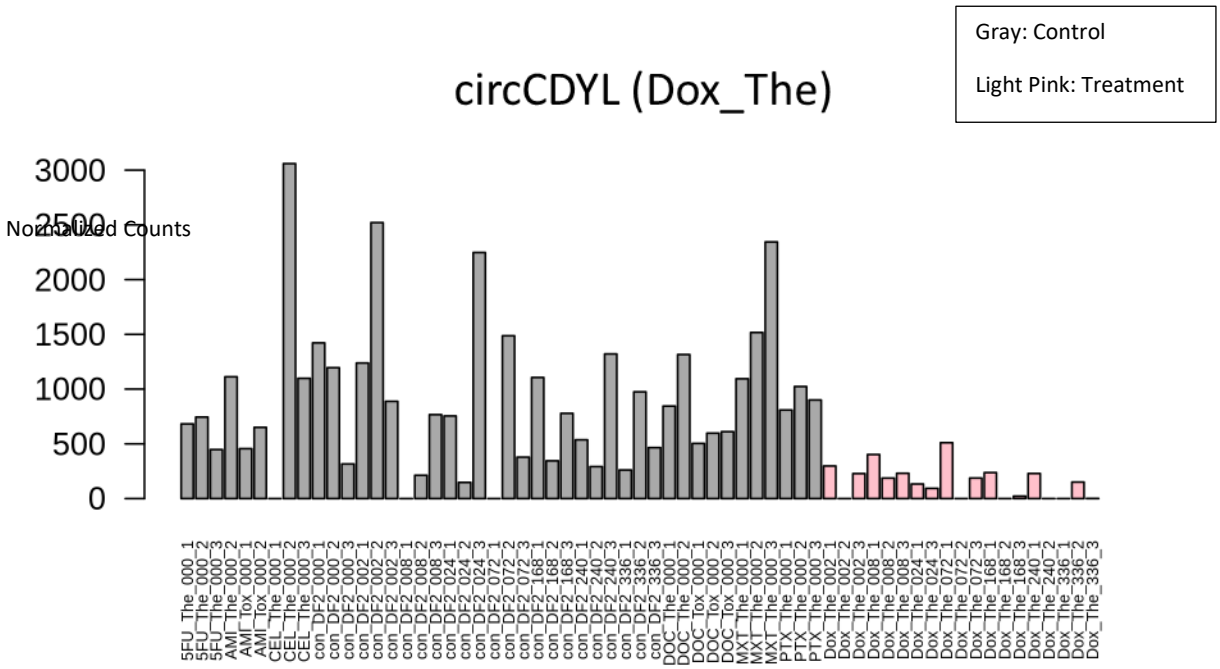


Figure 1: circCDYL (circRNA_ID: 6:4891713\|4892379) quantification for the Control versus the therapeutic dose of Doxorubicin (p.adj. < 0.001) comparison. *Each dark grey bar represents a control sample, while each pink bar represents a treated sample. The quantification shown was the result of the normalization performed by the DESeq2 pipeline. The 'circRNA_ID' is the identification given by CIRI2 based on the genome location. The names in the X-axis refer to the sample names: the first term indicates the treated compound, the second term refers to the dose used, the third term refers to the time-point (in hours), and the fourth term signals the triplicate number.*

To confirm that the change shown was not due to a down-regulation at the gene level, we quantified (using the remapping strategy in **3.3.2**) the expression of the linear CDYL mRNA transcripts. The results (Figure 2) showed no down-regulation for any of the 3 treatments. Instead, the therapeutic dose treatment of both Doxorubicin and Epirubicin presented a significant up-regulation of the linear transcripts. This suggests that the reduced circCDYL levels we observed upon doxorubicin treatment were not due to transcriptional repression of the CDYL gene.

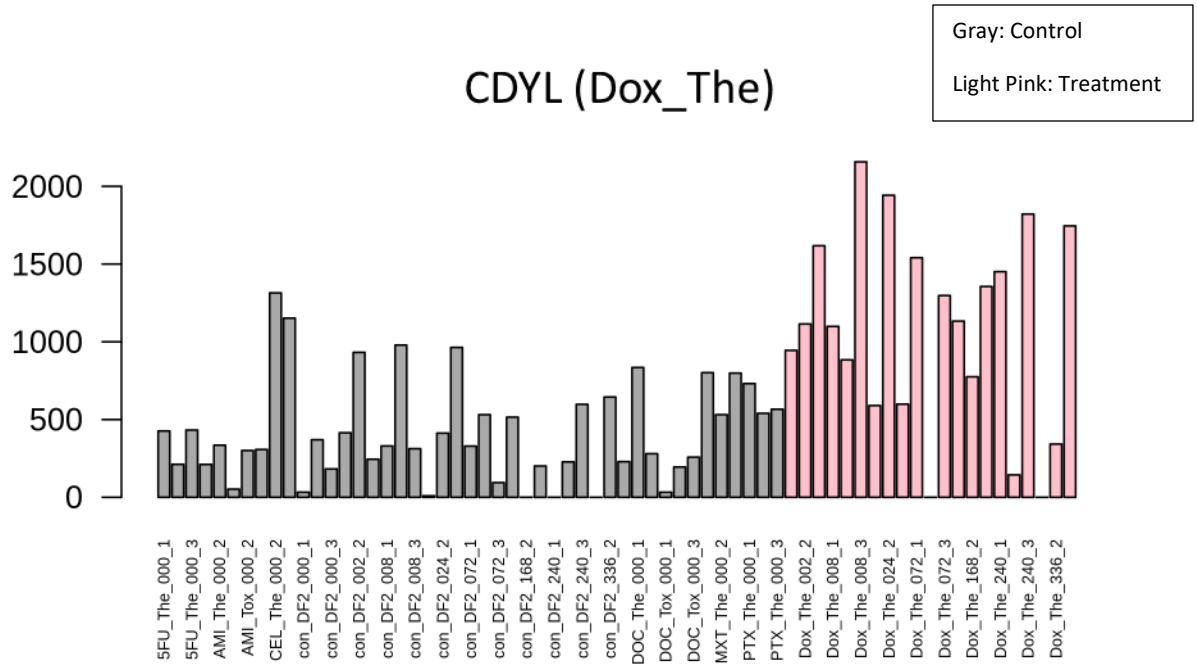


Figure 2: CDYL (ENSG00000153046) quantification by Salmon (circBase) for the Control versus the therapeutic dose of Doxorubicin (p.adj. < 0.01) comparison. *Each dark grey bar represents a control sample, while each pink bar represents a treated sample. The quantification shown was the result of the normalization performed by the DESeq2 pipeline.*

4.1.2. Exhaustive quantification using Salmon & circBase

Since *in silico* prediction is solely based on the detection of BSJ sequences, which is substantially dependent on coverage, we decided to re-quantify all transcripts by remapping against a common reference made of all linear transcripts (both protein and non-coding) and all circRNAs obtained from circBase. The detection range was between 39 and 57 thousand unique quantified circRNAs across all treatments (Table 2). As expected, the number of constitutive circRNAs was higher (576 on average) compared to the *in-silico* prediction (6 on average). The number (and proportion) of DECs was the highest in the anthracycline treatments (Doxorubicin and Epirubicin), similarly to what was observed in Table 1. Even so, the percentage of DECs is higher (12.23% on average) than the one *in silico* prediction (3.33% on average). This might be due to the higher percentage of samples where the DECs were detected (29.07% on average compared to 14.77% on average in CIRI2), where the increased sensitivity might help elucidate the possible changes when comparing the treatment and control groups.

Table 2: Summary of the Salmon + circBase and DESeq2 results. *'N quantified circRNA' refers to the number of unique circRNAs that were detected by Salmon in any sample involved in the comparison between that treatment and the control. 'N quantified constitutive circRNAs' refers to the number of unique circRNAs that were detected by Salmon in all samples involved in the comparison between that treatment and the control. 'N DECs' refers to the number of Differentially Expressed CircRNAs (DECs) in a specific treatment when compared to the control samples (p. adjusted value < 0.05). '% DECs' refers to the percentage of DECs taking the 'N quantified circRNA' as the total/denominator. 'Avg % samples a DEC is quantified' refers to the percentage of samples where a DEC is detected on average.*

Treatment	N quantified circRNA	N quantified constitutive circRNA	N DECs	% DECs	Avg % samples a DEC is quantified
Epi_Tox	39690	94	12438	31.34%	24.62%
Dox_Tox	39323	109	10479	26.65%	36.63%
Dox_The	39645	121	9489	23.93%	35.41%
Epi_The	41529	166	9454	22.76%	35.07%
PTX_Tox	43729	575	6949	15.89%	23.66%
CEL_The	41140	632	4910	11.93%	25.53%
MXT_Tox	43278	713	4662	10.77%	21.59%
DOC_The	39091	683	3207	8.20%	21.04%
AMI_The	40121	645	2615	6.52%	26.41%
5FU_Tox	44673	819	2745	6.14%	27.83%
PTX_The	57661	746	3542	6.14%	33.22%
DOC_Tox	46112	813	2666	5.78%	30.26%
MXT_The	53610	745	2993	5.58%	28.53%
CEL_Tox	54307	725	2689	4.95%	31.02%
AMI_Tox	46755	791	2244	4.80%	24.99%
5FU_The	45670	839	1920	4.20%	39.33%

To further investigate the increased sensitivity found in this method, we pooled the quantification of all the samples and analyzed the overlap in circRNA detection across all samples. The resulting data showed that only 8% of all detected circRNAs were quantified in a single sample, 50% (median) were quantified in 17 samples, and 9 circRNAs were quantified in all samples.

4.1.3. Selection of circRNAs for experimental validation

We selected for experimental validation 10 DECs across all different treatments based on expression, differential expression, and time effect: hsa_circ_0010791 (circCDC42), hsa_circ_0026129 (circTUBA1A), hsa_circ_0034356 (circACTC1), hsa_circ_0055922 (circFHL2), hsa_circ_0060999 (circGNAS), hsa_circ_0076194 (circCDKN1A), hsa_circ_0078905 (intergenic circRNA), hsa_circ_0090448 (circTIMP1), hsa_circ_0090904 (circMSN), and hsa_circ_0102325 (circHIF1A). hsa_circ_0008285 (circCDYL) and hsa_circ_0001445 (circSMARCA5) have been shown to increase

their expression after cardiac differentiation[19] and were thus added for the experimental validation as positive controls.

4.1.3.1 circGNAS (hsa_circ_0060999)

CircGNAS was an example of a highly expressed circRNA with a significant differential expression and time-effect accumulation when treated with a toxic dose of amiodarone. As seen in Figure 5, the circRNA expression increased significantly to levels not found in control conditions. Moreover, the increase was cumulative until 168 hours. At 240 hours (10 days), the expression appeared to return to control levels.

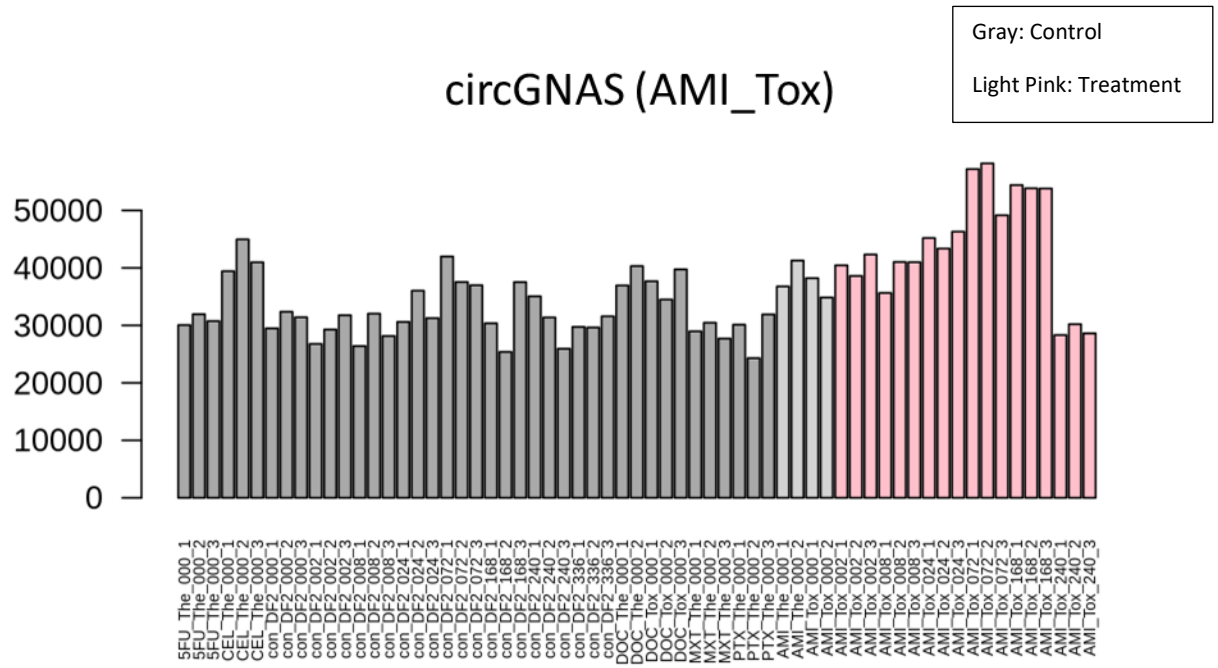
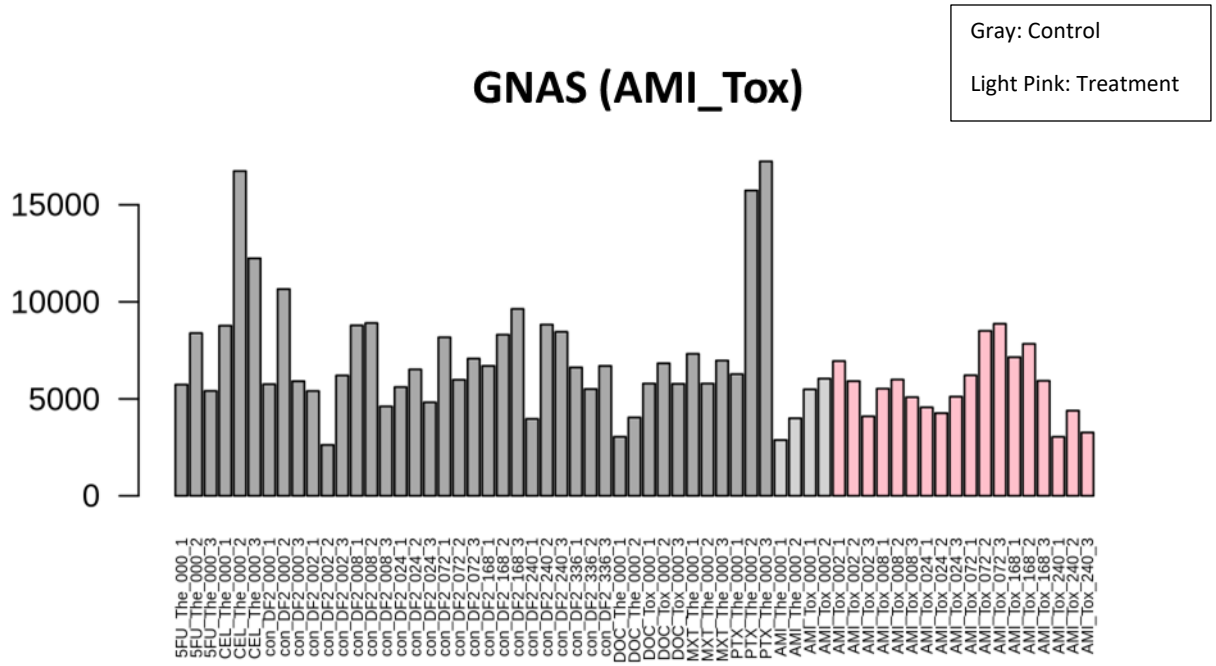


Figure 3: circGNAS (hsa_circ_0060999) quantification for the Control versus the toxic dose of Amiodarone comparison (p.adj. < 0.001). *Each dark grey bar represents a control sample, each light grey bar represents a control sample differentiated in the same conditions as the treatment, and each pink bar represents a treated sample. The quantification shown was the result of the normalization performed by the DESeq2 pipeline. The names in the X-axis refer to the sample names: the first term indicates the treated compound, the second term refers to the dose used, the third term refers to the time-point (in hours), and the fourth term signals the triplicate number.*

As performed for circCDYL, we quantified and analyzed the expression of the linear transcripts originated from the GNAS (Guanine Nucleotide binding protein, Alpha Stimulating activity polypeptide) gene to investigate whether the incremental expression shown was due to an up-regulation at the gene level. The results (Figure 4) showed no significant change for the analyzed treatment (toxic dose of Amiodarone).



366

367 ***Figure 4: GNAS (ENSG00000087460) quantification by Salmon (circBase) for the Control versus the toxic dose of***
368 ***Amiodarone (p.adj. > 0.05) comparison.*** *Each dark grey bar represents a control sample, each light grey bar represents a*
369 *control sample differentiated in the same conditions as the treatment, and each pink bar represents a treated sample. The*
370 *quantification shown was the result of the normalization performed by the DESeq2 pipeline.*

371 ## 4.2. qPCR validation

372 ### 4.2.1. circRNA existence validation

373 To investigate whether the predicted circRNAs exist and are expressed in cardiomyocytes, two batches of
374 cardiomyocytes were derived from iPSCs and checked for expression of our circRNAs of interest. As
375 previously mentioned, circCDYL and circSMARCA5 have already been shown to be expressed in iPSC-
376 derived cardiomyocytes and were therefore included in these experiments as controls. Furthermore, to
377 investigate whether our circRNAs of interest are cardiomyocyte-enriched, their expression was compared
378 to the initial iPSCs level.

379 The results confirmed the expression of three (out of 12) circRNAs existed in myocardial cells after iPSC
380 differentiation (Figure 5A and Figure 5B): circGNAS, circCDYL and circSMARCA5. Aside from the
381 existence of circCDYL and circSMARCA5, our results also confirmed previous results that both
382 circRNA expressions increased along with cardiac differentiation. In addition, we also observed a novel
383 relationship between circGNAS expression and cardiac differentiation. The six transcripts (circTUBA1A,
384 circTIMP1, circHIF1A, circFHL2, circACTC1, and hsa_circ_0078905) that are labelled in both Figures
385 (Figure 5A and Figure 5B) were also quantified, but the results in gel showed several bands, thus the

386 quantification values were a result of several circRNA molecules, which most probably were isoforms of
387 the ones we were studying. In the case of circCDC42, the expression was only detected for the second
388 batch (Figure 5B). CircCDKN1A and circMSN were not validated through qPCR.

389

390 In addition, we also observed a novel relationship between circGNAS expression and cardiac
391 differentiation. For six other circRNAs of interest (circTUBA1A, circTIMP1, circHIF1A, circFHL2,
392 circACTC1, and hsa_circ_0078905) multiple PCR products were formed. this is possibly caused by other
393 circular isoforms but we were not able to rule out aspecific amplification and were therefore excluded
394 from further analysis. Furthermore, circCDC42 was only detected in the second batch of iPSC-CMs.
395 CircCDKN1A and circMSN were not validated though qPCR.

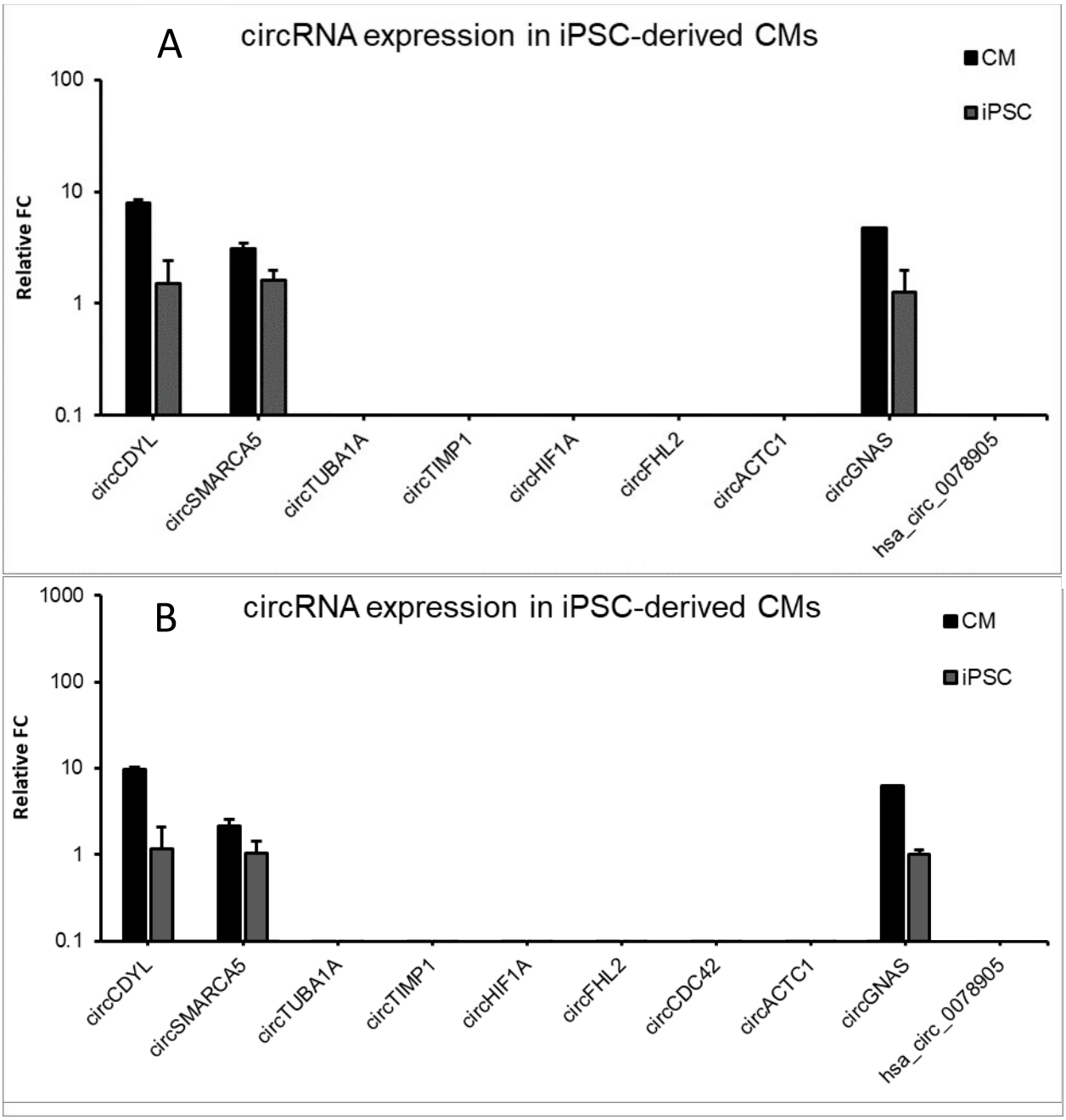


Figure 5: Relative expression of circRNAs by qPCR. A: *9 circRNAs amplified in the first batch.* ***B****: 10 circRNAs (9 + circCDC42) amplified in the second batch. In the vertical axis, Relative FC refers to the relative fold change of expression. In the horizontal axis, each couple of bars is identified by their circular RNA name based on their gene of origin. For the last circRNA, we used the circBase ID due to its intergenic nature. 'CM' stands for CardioMyocytes and 'iPSC' for induced Pluripotent Stem Cells.*

Taking into account the nine circRNAs (and potential isoforms) we experimentally quantified, we evaluated which circRNAs were detected in at least one sample by each of the two methodologies used. This required the cross-reference between gene loci (CIRI2) and circBase IDs (remapping), that is, we

transformed the circRNA names from one methodology to another. We observed that only two circRNAs (circHIF1A and circCDYL) were identified by CIRI2, while all nine validated circRNAs were detected by the remapping strategy.

4.2.2. Validation of Amiodarone effect on circGNAS

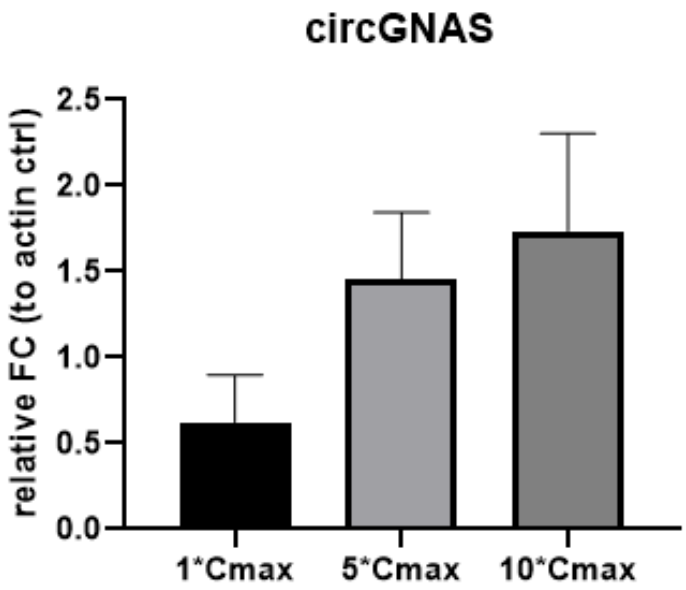


Figure 6: Relative expression of circGNAS upon treatment with three different doses of Amiodarone. *Cmax concentration represents the therapeutically active average plasma maximum concentration values derived from recommended therapeutic doses upon a single-dose administration.*

After observing differential circGNAS expression in our transcriptomics analysis upon amiodarone treatment, experimental validation was carried out. To investigate a possible dynamic dose-response link between circGNAS expression and amiodarone exposure, iPSC-derived cardiomyocytes were exposed to three different Cmax-based concentrations of amiodarone for 24 hours (Figure 6). Cmax concentration represents the therapeutically active average plasma maximum concentration values derived from recommended therapeutic doses upon a single-dose administration. In the case of amiodarone, the selected concentration (0.807 µM) was obtained from a study carried out by Wink et al[28]. The three different concentrations of amiodarone were selected at 1, 5, and 10 times Cmax (0.807, 4.035, and 8.07 µM respectively). After 24 hours of exposure to amiodarone, iPSC-derived cardiomyocytes were harvested for RNA and circGNAS expression was quantified using RT-qPCR (Figure 6). CircGNAS showed a dose-dependent increase expression upon Amiodarone exposure.

4.3. Investigation of the putative function of circRNAs

Next, we wanted to study the ability of circCDYL and circGNAS to act as miRNA sponges. Having access to different omics datasets obtained from the exact same samples, allowed us to robustly investigate the effect of circRNAs on miRNA expression levels and expression of miRNA-target mRNAs themselves as well as their encoded proteins.

First, we investigated which miRNAs could potentially bind to these circRNAs and whether their expression was also affected by the compounds. In addition, we examined the potential mRNA targets of these miRNAs and their expression changes. Finally, we examined potential changes at the proteomics level as a consequence of post-transcriptional dysregulation.

4.4. CircCDYL

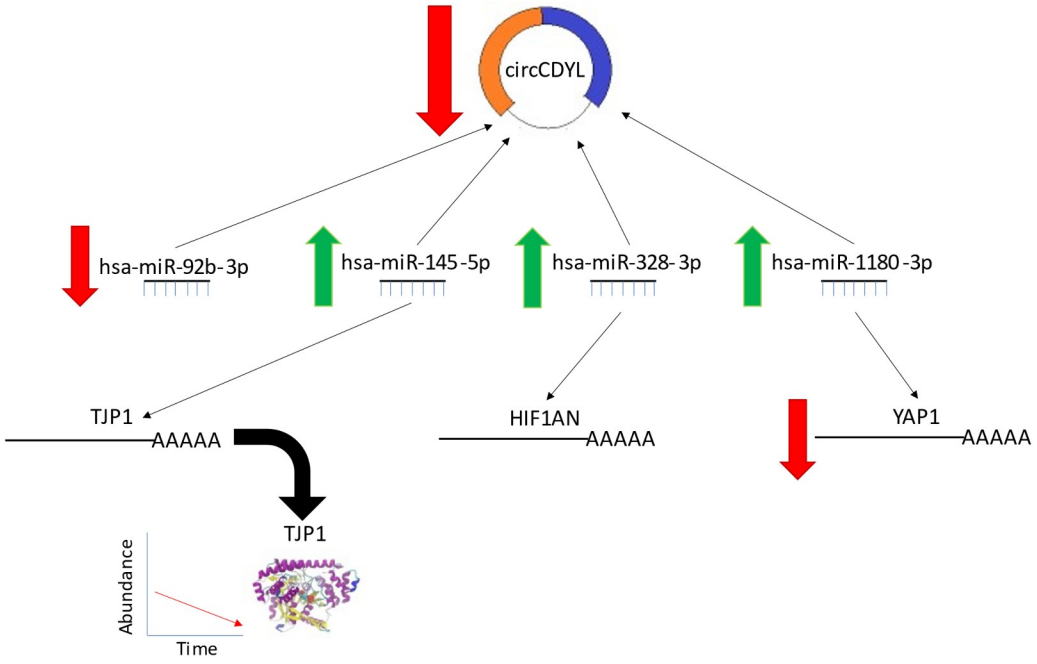


Figure 7: Graphical summary of the expression changes in the post-transcriptional regulation network of circCDYL after the exposure to a therapeutic dose of Doxorubicin. *The red arrows pointing down represent a significant expression decrease of that molecule. The green arrows pointing up represent a significant expression increase of that molecule. The black line arrows represent the binding and/or translational inhibition of miRNAs to their targets. The black bent arrow represents the translation of a transcript to a protein. The red dashed line arrow pointing down represents a non-significant decrease in the expression of a molecule. A red line arrow inside an XY plot represents a time-dependent decrease in the abundance of a molecule.*

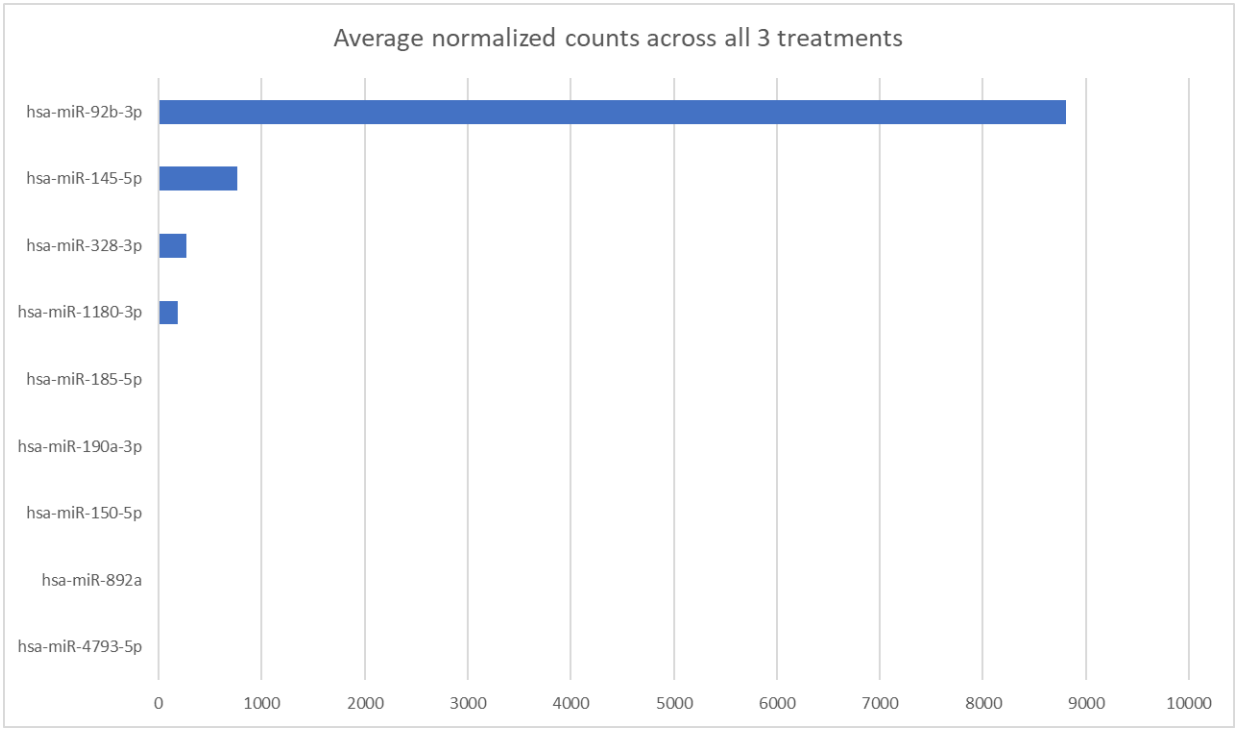
4.4.1. MiRNAs associated with circCDYL

Due to circCDYL being a previously studied circular RNA, we selected miRNAs that had been previously hypothesized to target this circRNA. The selected miRNAs were: hsa-miR-190a-3p[29], hsa-

446 miR-185-5p[30], hsa-miR-4793-5p[31], hsa-miR-150-5p[32], hsa-miR-892a[33] , hsa-miR-328-3p[33], hsa-miR-92b-
447 3p[34], hsa-miR-145-5p[35], and hsa-miR-1180-3p[36].

448 First, we calculated the average expression for each of the miRNAs across the 3 treatments of interest
449 (therapeutic and toxic dose of Doxorubicin, and therapeutic dose of Epirubicin) to filter out miRNAs with
450 low expression, as their influence on their target's translation would be minimal. As a result, we excluded
451 5 miRNAs (Figure 8) with minimal (less than one normalized read) or no expression on average: hsa-
452 miR-185-5p, hsa-miR-190a-3p, hsa-miR-150-5p, hsa-miR-892a, and hsa-miR-4793-5p. Therefore, only
453 the expression of the remaining 4 miRNAs (hsa-miR-92b-3p, hsa-miR-145-5p, hsa-miR-328-3p, and hsa-
454 miR-1180-3p) was analyzed.

455

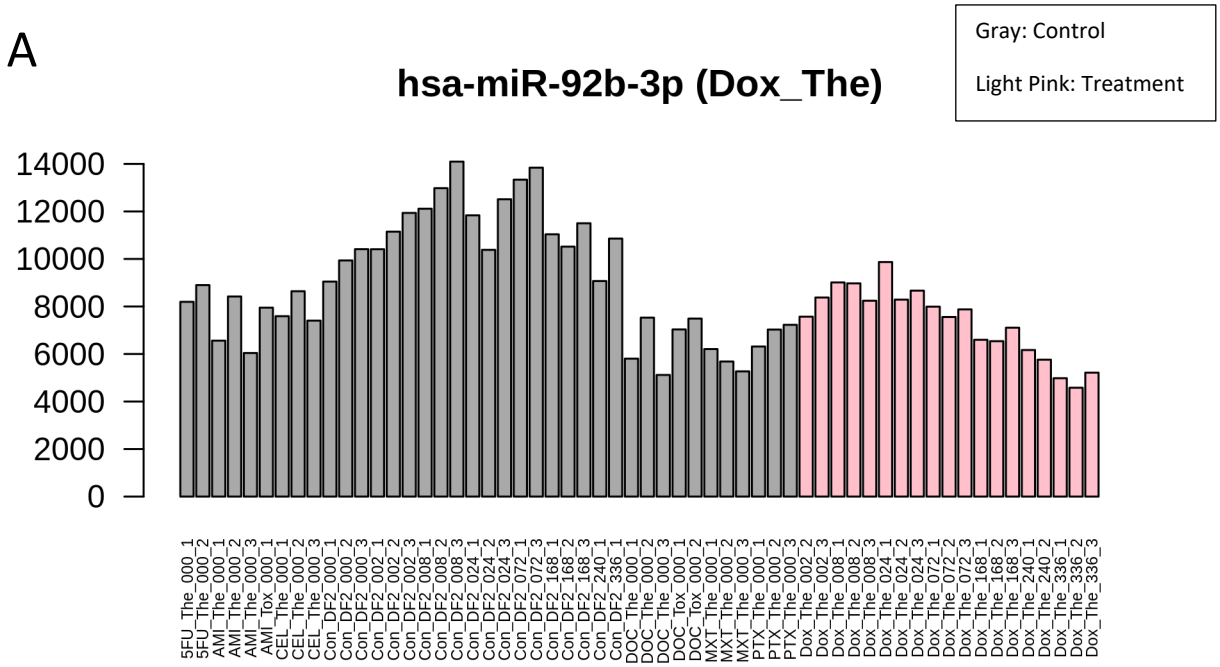


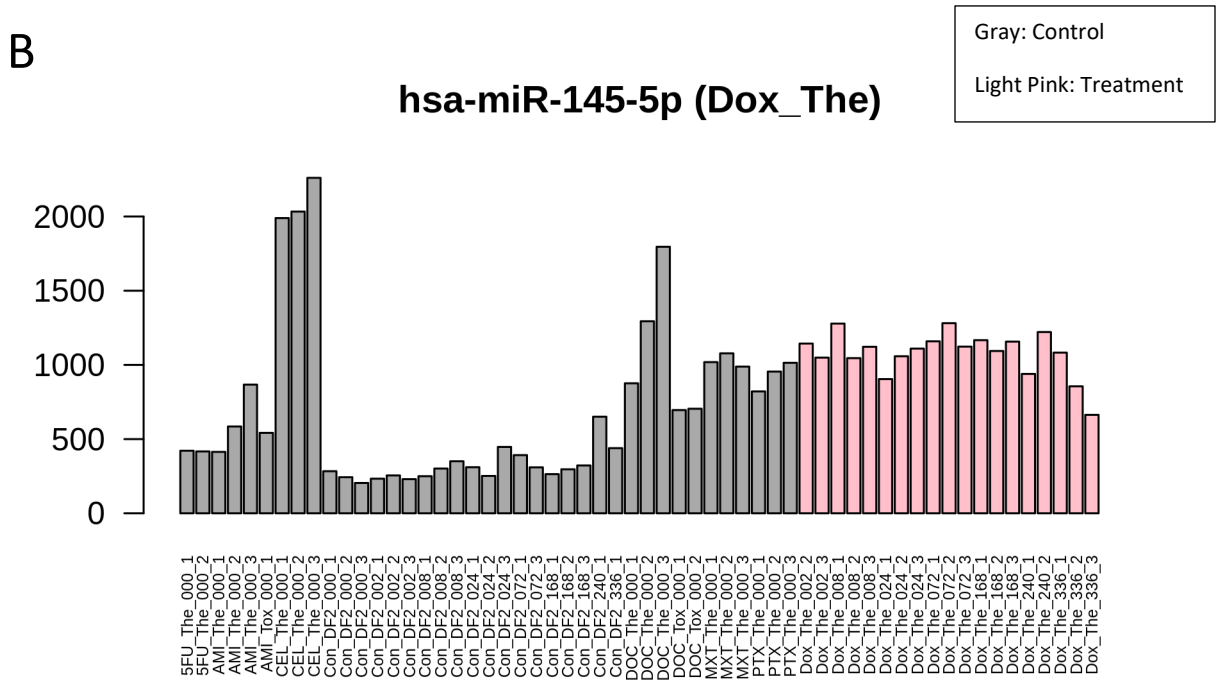
456

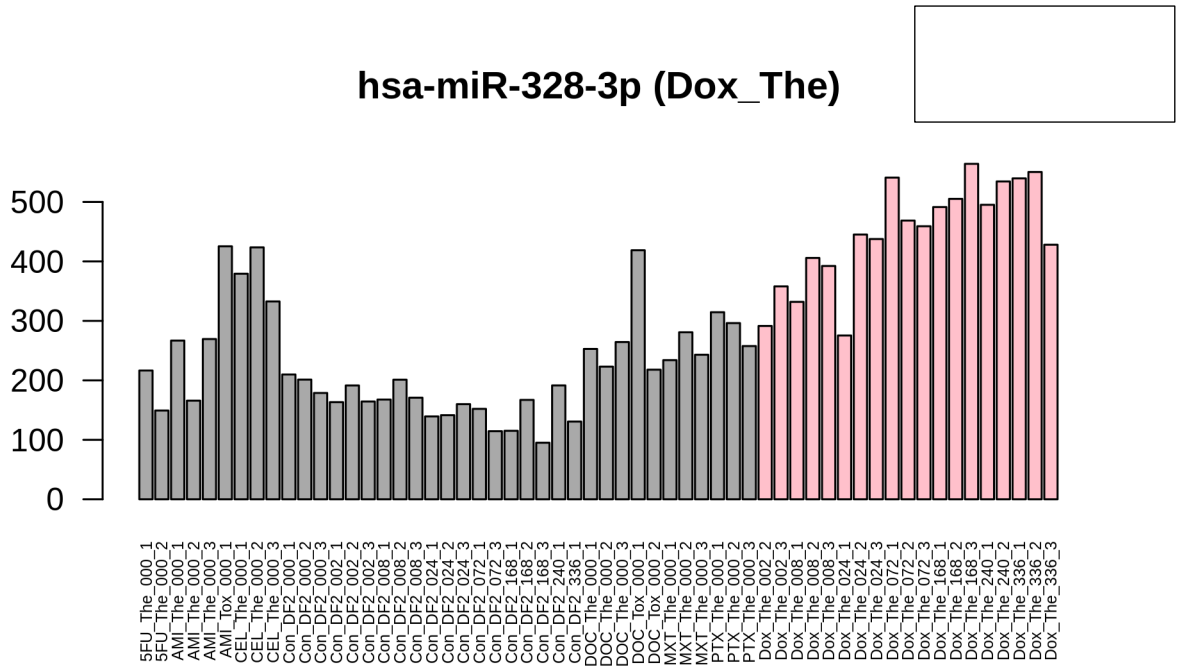
457 ***Figure 8: Bar plot of the average quantified expression of miRNAs associated with circCDYL****. The normalization and resulting*
458 *values were obtained as a result of using the DESeq2 pipeline.*

459 For each miRNA, we obtained the p. adjusted value, while also plotting the normalized expression of all
460 samples involved in the comparison. Hsa-miR-92b-3p (Figure 9A) was the only miRNA that showed a
461 significant decrease as a function of time after exposure to doxorubicin at a therapeutic dose. For the
462 other three miRNAs (Figure 9B, Figure 9C, and Figure 9D), the therapeutic dose of Doxorubicin and
463 Epirubicin always presented a significant up-regulation, while upon the toxic dose of Doxorubicin, only

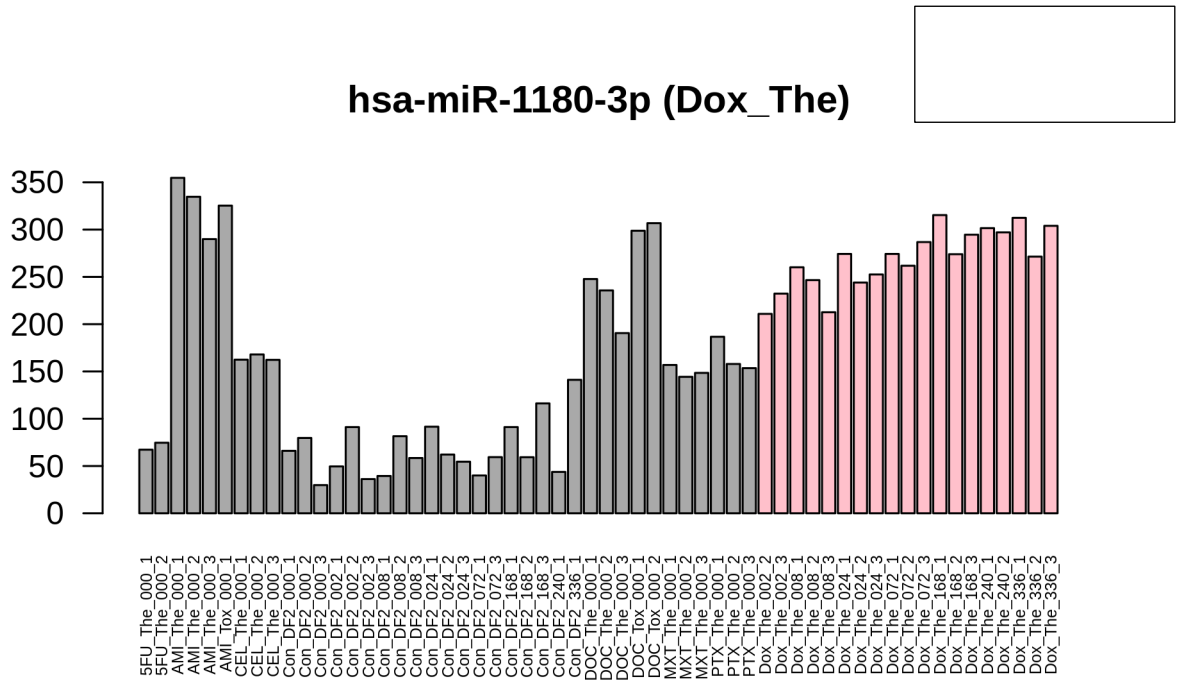
hsa-miR-328-3p was significantly up-regulated in a time-dependent manner (Figure 9C). Interestingly enough, the latter miRNA also showed a time-dependent increase in expression for all three treatments (Supplementary Results).







469



470

471 ***Figure 9: Normalized read counts of four miRNAs by miRge2 for the Control versus the therapeutic dose of Doxorubicin***
472 ***comparison.*** ***A:*** *hsa-miR-92b-3p (p.adj. < 0.01).* ***B:*** *hsa-miR-145-5p (p.adj. < 0.05).* ***C:*** *hsa-miR-328-3p (p.adj. < 0.001).* ***D:***
473 *hsa-miR-1180-3p (p.adj. < 0.001). Each dark grey bar represents a control sample, while each pink bar represents a treated*
474 *sample. The quantification shown was the result of the normalization performed by the DESeq2 pipeline. The title of the plot*
475 *shows the miRNA ID next to the treatment name.*

476

4.4.2. Target transcript expression

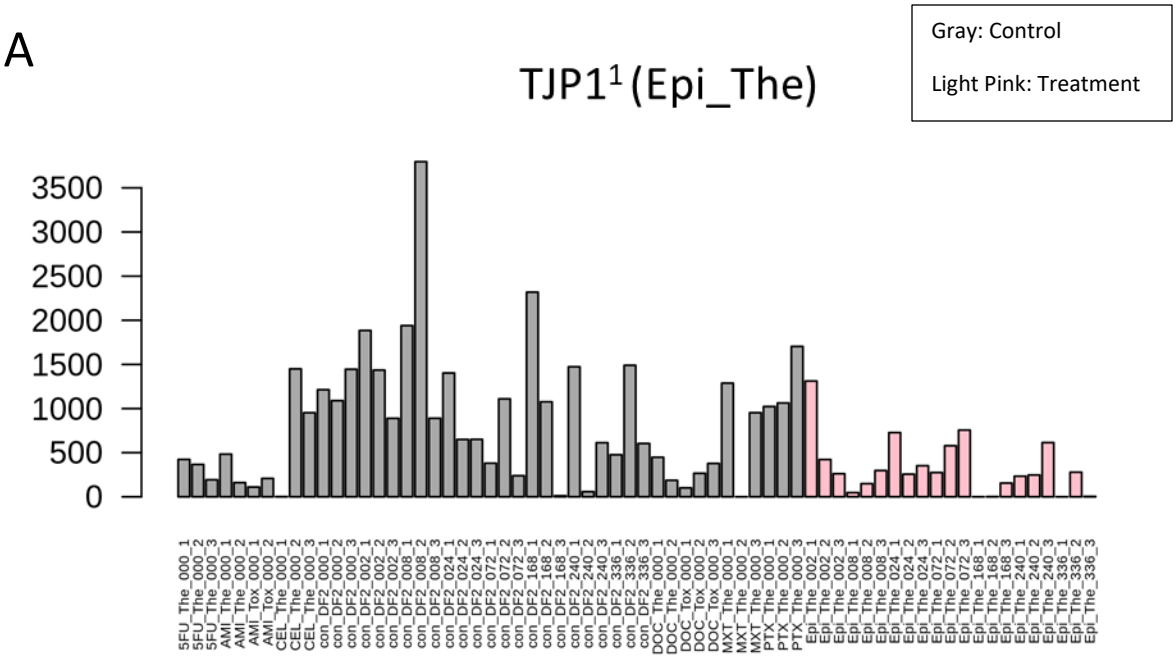
We investigated which targets were associated in earlier studies with the previously mentioned expressed miRNAs. We excluded the targets of hsa-miR-92b-3p, as it was found to be able to degrade circCDYL, and thus the relationship between both entities could not be classified as the classical "miRNA sponging". The differential expression analysis of the target genes (Table 3) showed that all were significantly downregulated for the therapeutic dose of Epirubicin (Figure 10A and Figure 10B), while only YAP1 was significantly downregulated in all 3 treatments (Figure 10D).

Table 3: Overall statistics of known circRNA-regulated genes. *P. adjusted values in bold are significant (p. adj. value < 0.05). [1]: Paralog TJP1 gene (ENSG00000104067). [2]: Paralog TJP1 gene (ENSG00000277401). [3]: Weighted average of both TJP1 gene paralogs (weights were derived from the sum of the three mean expressions for each paralog). L2FC: Log2 Fold Change.*

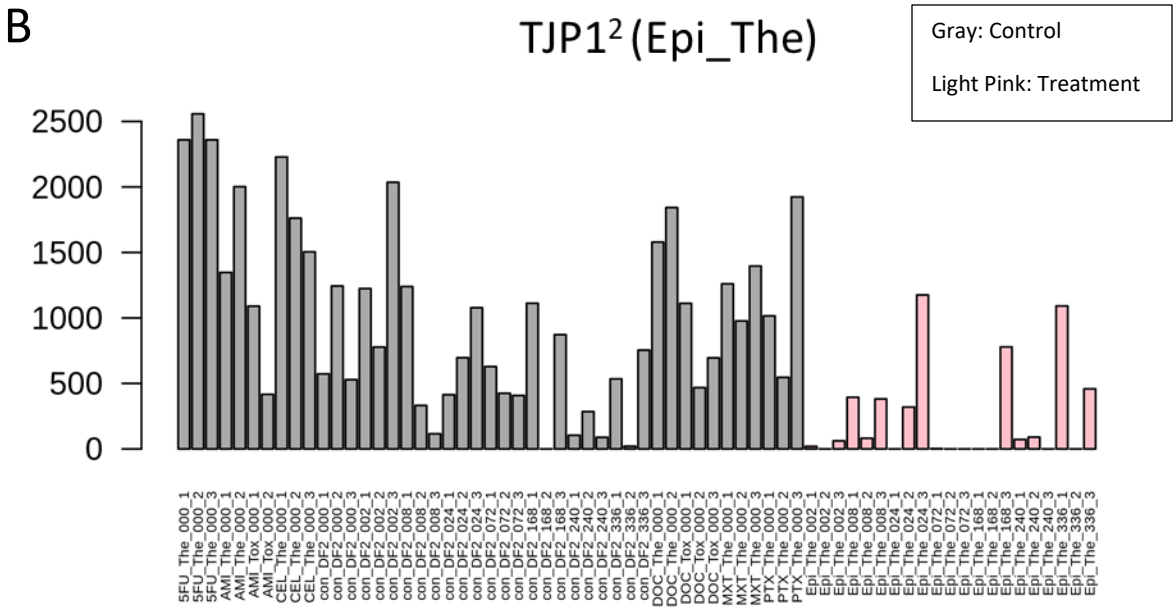
Gene Name	**Mean expression (Dox_The)**	**Mean expression (Dox_Tox)**	**Mean expression (Epi_The)**	**P. adj. value (L2FC) [Dox_The]**	**P. adj. value (L2FC) [Dox_Tox]**	**P. adj. value (L2FC) [Epi_The]**
TJP1[1]	666	689	695	$2.24x10^{-1}$ (-0.81)	$4.53x10^{-1}$ (-0.55)	**$6.86x10^{-3}$ (-1.38)**
TJP1[2]	722	731	770	**$4.72x10^{-2}$ (-1.48)**	**$2.06x10^{-2}$ (-1.39)**	**$4.36x10^{-4}$ (-2.12)**
TJP1[3]	695	711	734	$1.32x10^{-1}$ (-1.16)	$2.28x10^{-1}$ (-0.99)	**$3.52x10^{-3}$ (-1.76)**
HIF1AN	171	151	156	$6.02x10^{-1}$ (-0.50)	$9.23x10^{-2}$ (-1.44)	**$3.04x10^{-3}$ (-2.36)**
YAP1	3331	3511	4020	**$1.36x10^{-10}$ (-1.36)**	**$3.79x10^{-3}$ (-0.90)**	**$1.97x10^{-6}$ (-0.74)**

By sponging miR-145-5p, circCDYL was suggested to have a regulatory effect on the expression of TJP1 in a circCDYL/miR-145-5p/TJP1 axis[35]. From the three treatments, only the samples exposed to a therapeutic dose of Epirubicin showed significant downregulation of TJP1 gene expression in both paralogs (Table 3, Figure 10A and 9B). The other two treatments, though significantly downregulated in the second paralog (ENSG00000277401), were not significant when taking the weighted average effect of both treatments (Table 3).

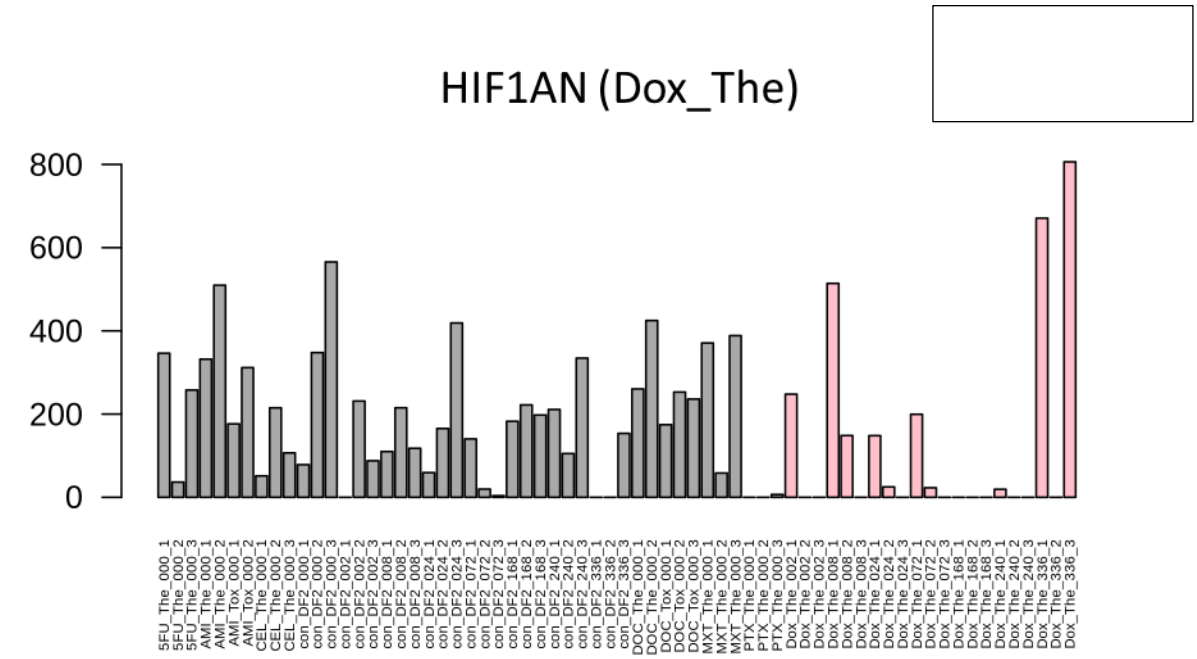
A


TJP1¹ (Epi_The)
Gray: Control
Light Pink: Treatment
3500
3000
2500
2000
1500
1000
500
0

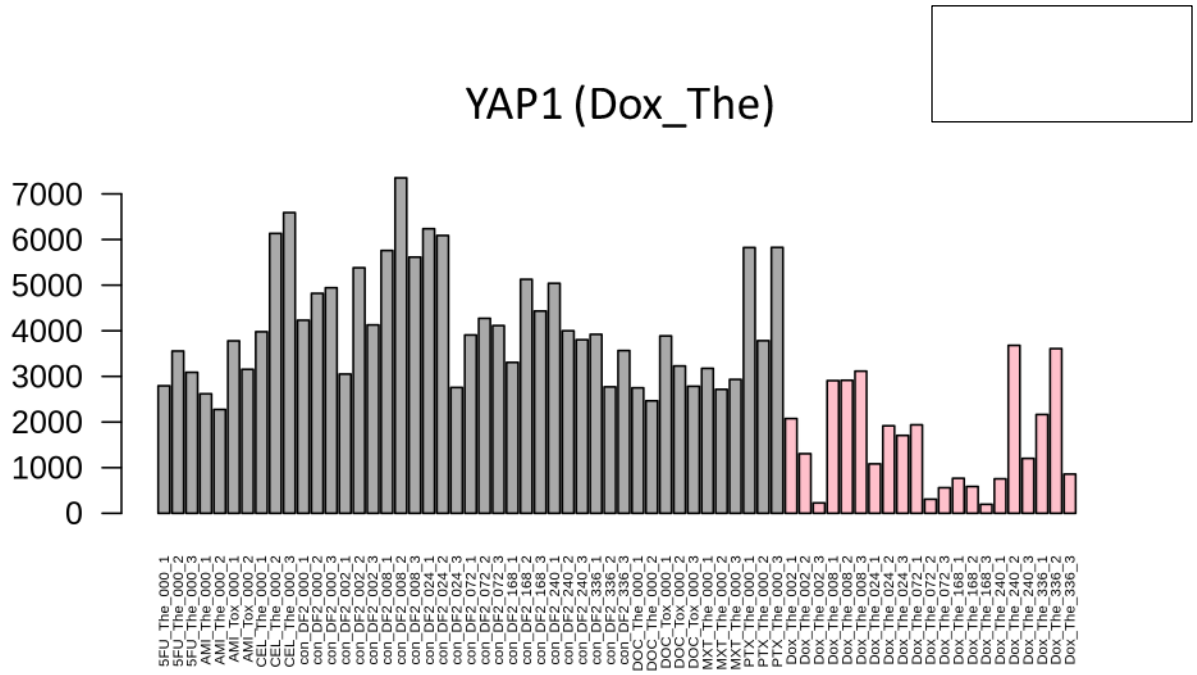
B


TJP1² (Epi_The)
Gray: Control
Light Pink: Treatment
2500
2000
1500
1000
500
0

494



495



496

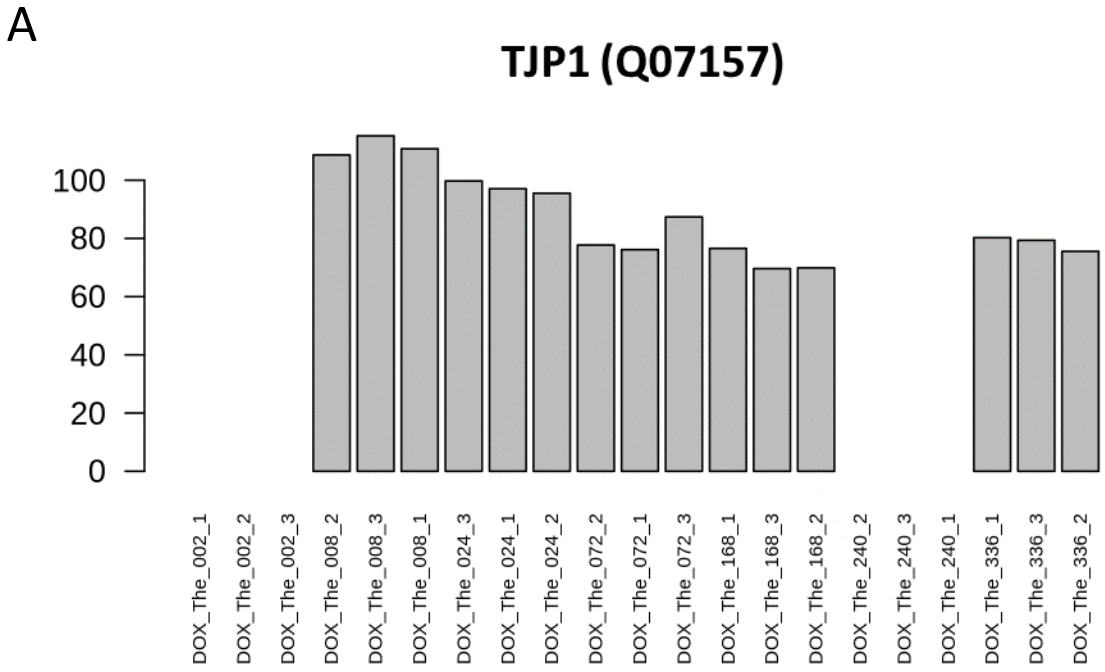
497 ***Figure 10: Gene expression levels after exposure to anthracycline compounds. A&B****: TJP1 gene expression for both paralogs*
498 *after exposure to the therapeutic dose of Epirubicin. Taking the weighted average (Table 3, TJP1[3]) for both paralog genes (1:*
499 *ENSG00000104067 and 2: ENSG00000277401), only the therapeutic dose of Epirubicin was significantly downregulated for*
500 *TJP1.* ***C****: HIF1AN gene expression after exposure to the therapeutic dose of Doxorubicin (p.adj. > 0.05).* ***D****: YAP1 gene*
501 *expression after exposure to the therapeutic dose of Doxorubicin (p.adj. < 0.001). Each dark grey bar represents a control*
502 *sample, while each pink bar represents a treated sample. The quantification shown was the result of the normalization performed*
503 *by the DESeq2 pipeline.*

CircCDYL was suggested to interact with mRNAs encoding hypoxia-inducible factor asparagine hydroxylase (HIF1AN) by acting as the sponge of miR-328-3p[33]. Even though HIF1AN gene expression was not statistically significantly affected by the Doxorubicin treatments (Table 3), we observed a clear down-regulation in all treatments (Figure 10C). The "mild" gene down-regulation, in addition to the up-regulation of miR-328-3p, and down-regulation of circCDYL might result in decreased translation of the target.

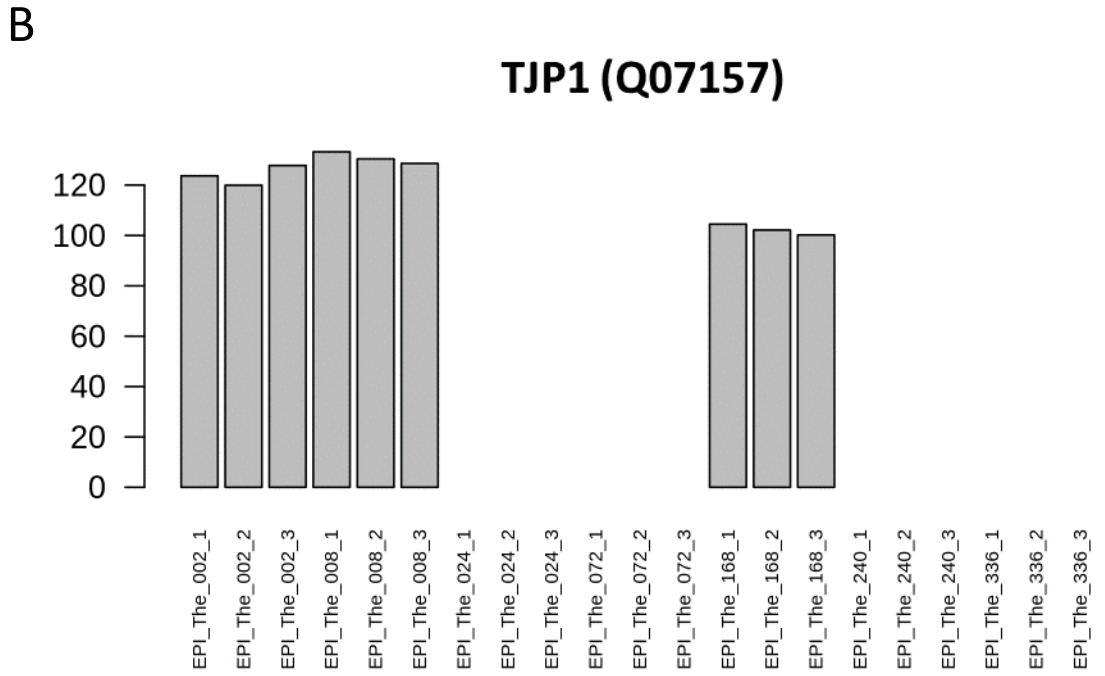
CircCDYL was also suggested to be able to absorb miR-1180, thus alleviating the repression of miR-1180 on YAP[36]. The gene expression of YAP was significantly downregulated for all 3 treatments (Table 3, Figure 10D). Thus, the accumulated effects of the downregulation of circCDYL (Figure 1), the upregulation of miR-1180 for most treatments (Figure 9D), and the downregulation of the gene expression (Figure 10D) should have had a significant effect on YAP protein expression.

4.4.3. Proteomics analysis

Translational inhibition is one of the primary functions of miRNAs, thus we expected to observe the potential post-transcriptional regulation effects at the protein level. Consequently, we analyzed the protein products of the gene targets mentioned above.



520



521

522 ***Figure 11: Proteomics levels of TJP1 (UniProt ID: Q07157). A****: After exposure to the therapeutic dose of Doxorubicin.* ***B****: After*
523 *exposure to the therapeutic dose of Epirubicin.*

524 TJP1 was proposed to be targeted by miR-145-5p, the latter has also been hypothesized to be regulated by
525 circCDYL, leading to a circCDYLßmiR-145-5pàTJP1 regulation. So far, we found that at the gene level,
526 only the therapeutic dose of Epirubicin showed a significant downregulation. When analyzing the

proteomics expression of TJP1-coded proteins, a clear time-dependent decrease was detected in the therapeutic dose of Doxorubicin (Figure 11A), where no significant decrease was found on average for the target gene expression. In the other 2 anthracycline treatments (Doxorubicin toxic dose and Epirubicin therapeutic dose), the protein expression was seldom detected, and only a decrease between 8 and 168h timepoints was identified (Figure 11B).

Due to the stochastic properties and limited sensitivity of mass spectrometry, no other target-coded protein was quantified. In proteomics, this is especially the case for less abundant proteins, which was the expected case for the analyzed proteins if their translation was inhibited.

4.5. CircGNAS

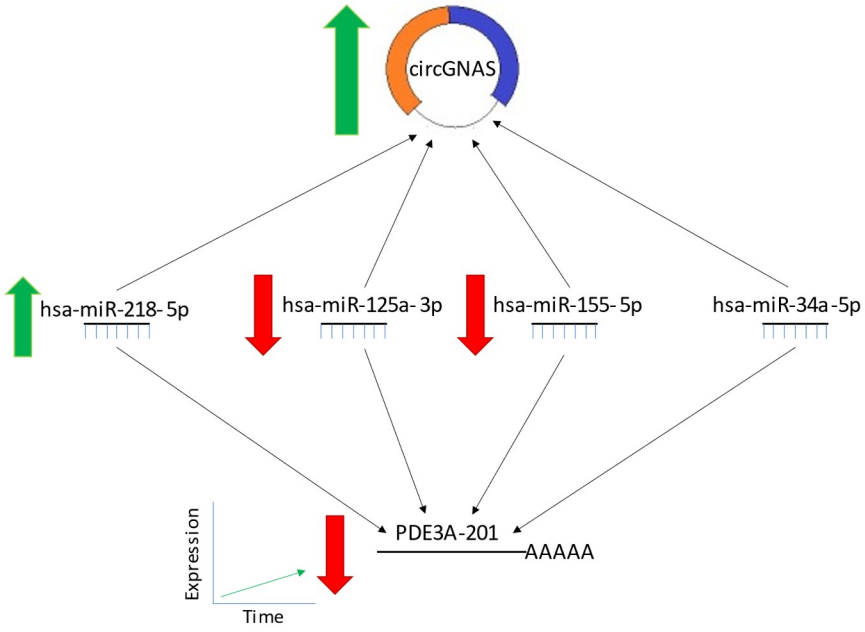


Figure 12: Graphical summary of the expression changes in the post-transcriptional regulation network of circGNAS after Amiodarone treatment with a toxic dose. *The red arrows pointing down represent a significant expression decrease of that molecule. The green arrows pointing up represent a significant expression increase of that molecule. The black line arrows represent the binding and/or translational inhibition of miRNAs to their targets. A green line arrow inside an XY plot represents a time-dependent increase in the abundance of a molecule.*

4.5.1. MiRNAs associated with circGNAS

No research has been published regarding circGNAS so far. Thus, we used miRanda to find miRNAs that could potentially be sponged by circGNAS. Out of the ~ 140 potential miRNAs to bind circGNAS, four of them were the most expressed and presented the highest expression consistency across all samples. Three of them (hsa-miR-218-5p, hsa-miR-125a-3p, and hsa-miR-155-5p) showed differential expression,

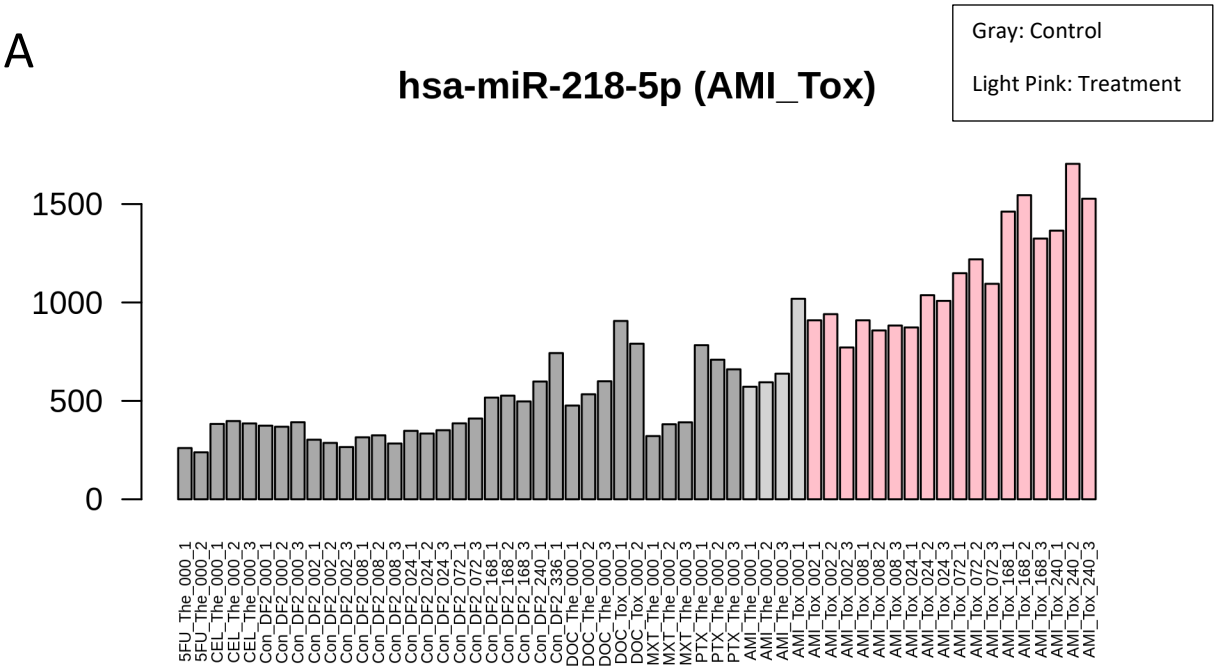
one increasing and two decreasing respectively (Table 4). The remaining miRNA (hsa-miR-34a-5p) did not present differential expression, even though it did present a time-dependent decrease.

Table 4: Overall statistics of expressed miRNAs that were potentially regulated by circGNAS*. P. adjusted values in bold are significant (p. adj. value < 0.05). These statistical results were obtained from DESeq2.*

miRNA Name	**Mean expression (AMI_Tox)**	**Log2 FoldChange (AMI_Tox)**	**P. adj. value (AMI_Tox)**
hsa-miR-218-5p	688	1.25	**$1.16x10^{-17}$**
hsa-miR-125a-3p	44	-0.87	**$3.86x10^{-8}$**
hsa-miR-155-5p	187	-1.26	**$8.13x10^{-4}$**
hsa-miR-34a-5p	132	0.19	$1.42x10^{-1}$

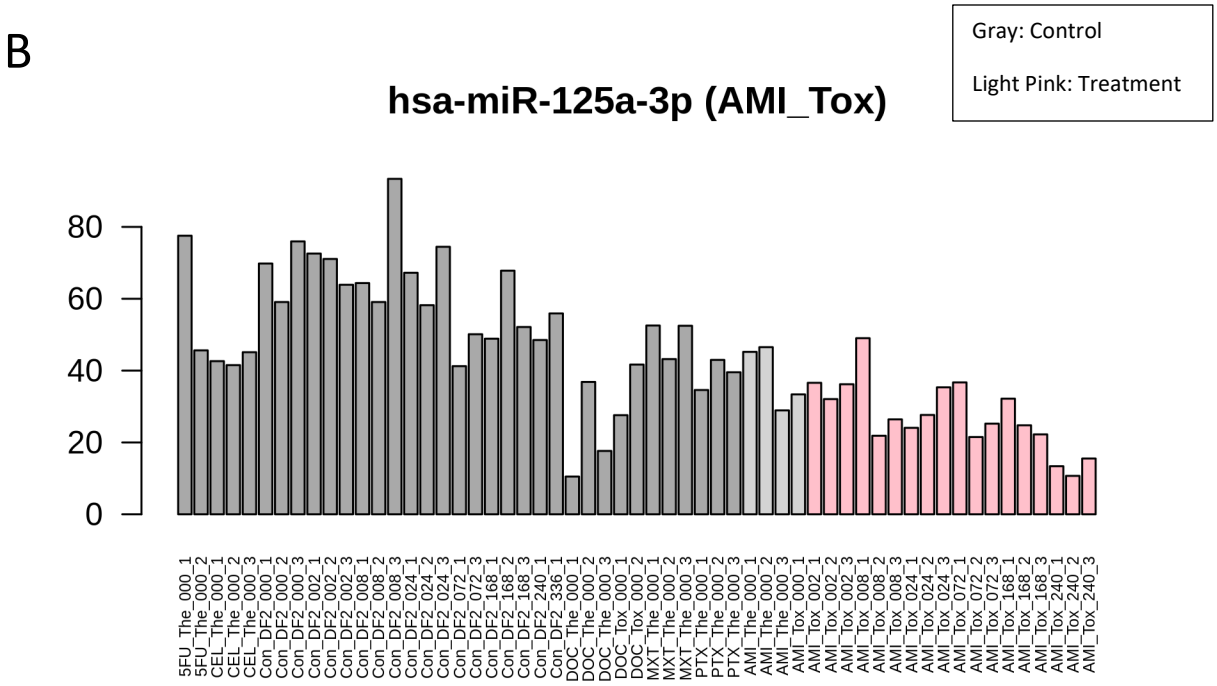
Hsa-miR-218-5p expression showed a significant up-regulation after exposure to the toxic dose of Amiodarone (Figure 13). In addition, a clear time increase was perceived.

A


hsa-miR-218-5p (AMI_Tox)
Gray: Control
Light Pink: Treatment
1500
1000
500
0
5FU_The_000_1
5FU_The_000_2
CEL_The_000_1
CEL_The_000_2
CEL_The_000_3
Con_DF2_000_1
Con_DF2_000_2
Con_DF2_000_3
Con_DF2_002_1
Con_DF2_002_2
Con_DF2_002_3
Con_DF2_008_1
Con_DF2_008_2
Con_DF2_008_3
Con_DF2_024_1
Con_DF2_024_2
Con_DF2_024_3
Con_DF2_072_1
Con_DF2_072_3
Con_DF2_168_1
Con_DF2_168_2
Con_DF2_168_3
Con_DF2_240_1
Con_DF2_336_1
DOC_The_000_1
DOC_The_000_2
DOC_The_000_3
DOC_Tox_000_1
DOC_Tox_000_2
MXT_The_000_1
MXT_The_000_2
MXT_The_000_3
PTX_The_000_1
PTX_The_000_2
PTX_The_000_3
AMI_The_000_1
AMI_The_000_2
AMI_The_000_3
AMI_Tox_000_1
AMI_Tox_002_1
AMI_Tox_002_2
AMI_Tox_002_3
AMI_Tox_008_1
AMI_Tox_008_2
AMI_Tox_008_3
AMI_Tox_024_1
AMI_Tox_024_2
AMI_Tox_024_3
AMI_Tox_072_1
AMI_Tox_072_2
AMI_Tox_072_3
AMI_Tox_168_1
AMI_Tox_168_2
AMI_Tox_168_3
AMI_Tox_240_1
AMI_Tox_240_2
AMI_Tox_240_3

554

B


hsa-miR-125a-3p (AMI_Tox)
Gray: Control
Light Pink: Treatment
80
60
40
20
0
5FU_The_000_1
5FU_The_000_2
CEL_The_000_1
CEL_The_000_2
CEL_The_000_3
Con_DF2_000_1
Con_DF2_000_2
Con_DF2_000_3
Con_DF2_002_1
Con_DF2_002_2
Con_DF2_002_3
Con_DF2_008_1
Con_DF2_008_2
Con_DF2_008_3
Con_DF2_024_1
Con_DF2_024_2
Con_DF2_024_3
Con_DF2_072_1
Con_DF2_072_3
Con_DF2_168_1
Con_DF2_168_2
Con_DF2_168_3
Con_DF2_240_1
Con_DF2_336_1
DOC_The_000_1
DOC_The_000_2
DOC_The_000_3
DOC_Tox_000_1
DOC_Tox_000_2
MXT_The_000_1
MXT_The_000_2
MXT_The_000_3
PTX_The_000_1
PTX_The_000_2
PTX_The_000_3
AMI_The_000_1
AMI_The_000_2
AMI_The_000_3
AMI_Tox_000_1
AMI_Tox_002_1
AMI_Tox_002_2
AMI_Tox_002_3
AMI_Tox_008_1
AMI_Tox_008_2
AMI_Tox_008_3
AMI_Tox_024_1
AMI_Tox_024_2
AMI_Tox_024_3
AMI_Tox_072_1
AMI_Tox_072_2
AMI_Tox_072_3
AMI_Tox_168_1
AMI_Tox_168_2
AMI_Tox_168_3
AMI_Tox_240_1
AMI_Tox_240_2
AMI_Tox_240_3

555

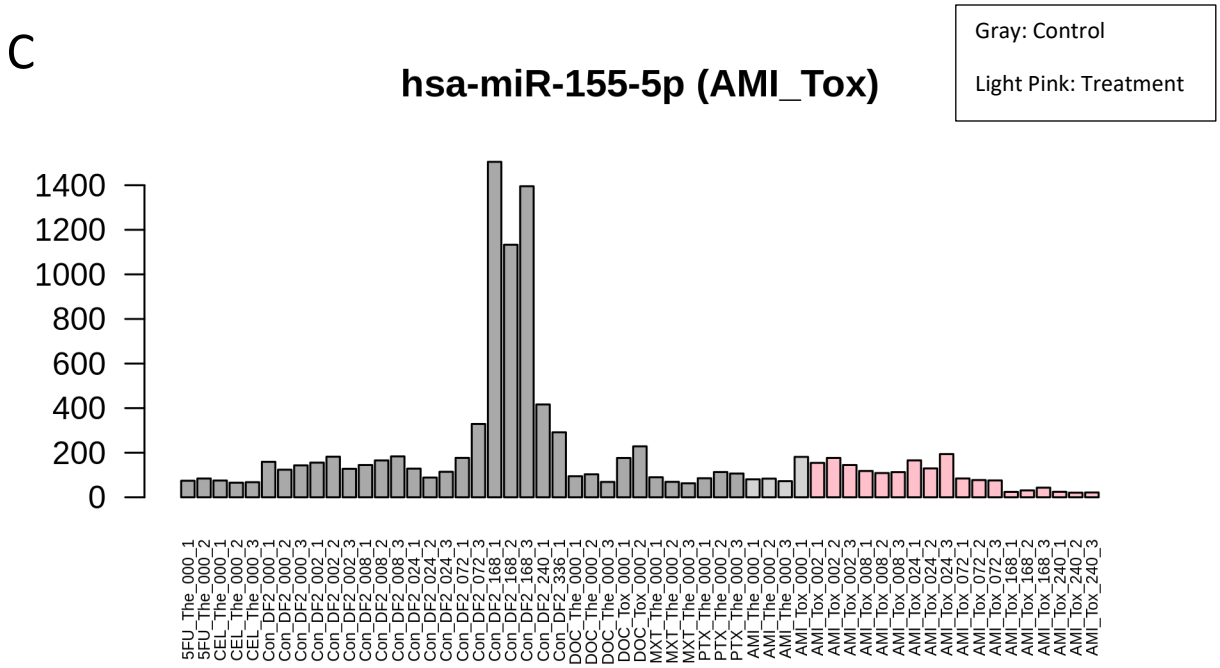


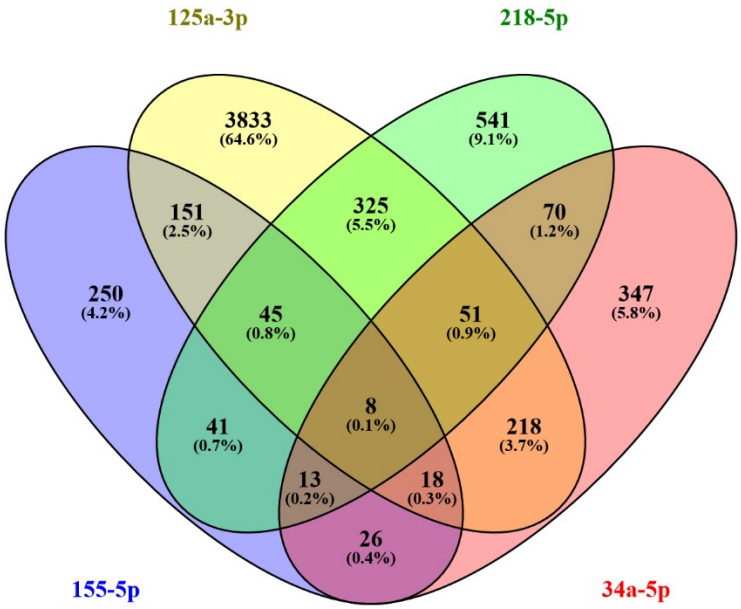
Figure 13: Normalized counts through quantification by miRge2 for the Control versus the toxic dose of Amiodarone comparison. A: *hsa-miR-218-5p (p.adj. < 0.001).* ***B****: hsa-miR-125a-3p (p.adj. < 0.001).* ***C****: hsa-miR-155-5p (p.adj. < 0.001). Each dark grey bar represents a control sample, each light grey bar represents a control sample differentiated in the same conditions as the treatment, and each pink bar represents a treated sample. The quantification shown was the result of the normalization performed by the DESeq2 pipeline. The title of the plot shows the miRNA ID next to the treatment name.*

Hsa-miR-125a-3p also presented a significant differential expression, but as a down-regulation (Figure 13B). For this miRNA, we also observed a time-dependency, where the expression gradually decreased over time.

For hsa-miR-155-5p, even though it was significantly down-regulated according to the p. adjusted value, analyzing the values we observed that this significance was mainly due to three triplicate samples with outlier expressions (Figure 13C). In addition, we also observed (as in hsa-miR-125a-3p) a negative cumulative effect across time.

4.5.2. Target transcript and protein expression

To assess which targets might be specifically regulated by circGNAS, we assembled all the potential targets of the aforementioned four miRNAs based on TargetScan (7.1) data. The target overlap across the four miRNAs can be visualized in the Venn Diagram (Figure 14). Eight transcripts were targeted by all 4 of the miRNAs selected.

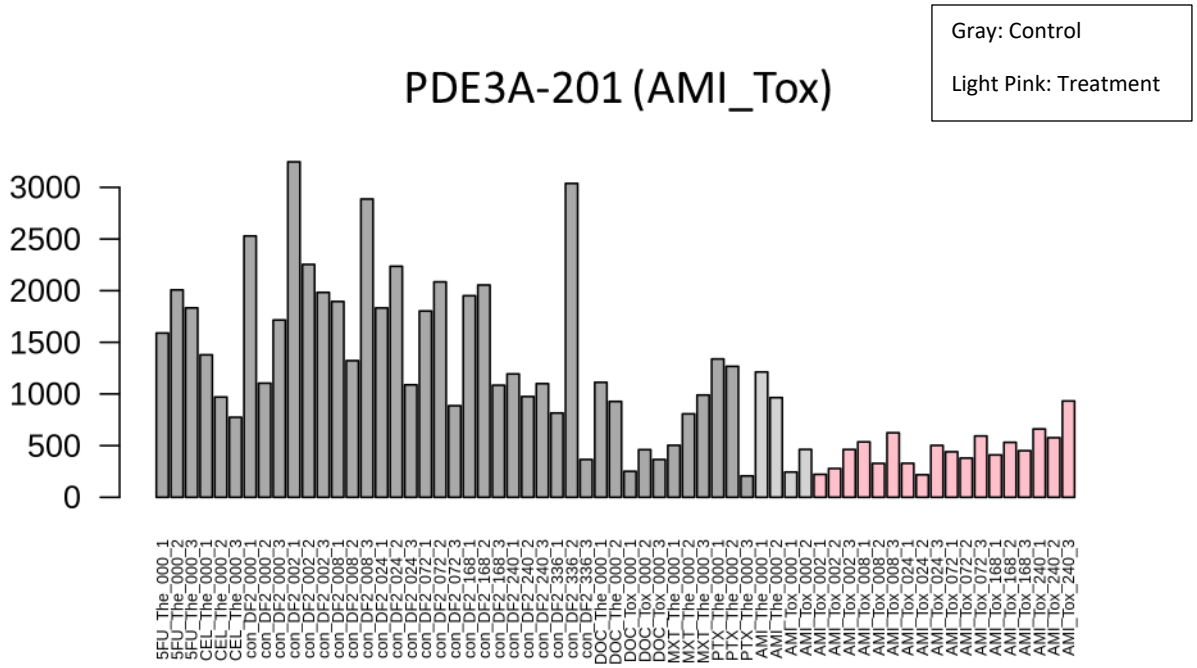


574

575 *Figure 14: Venn Diagram of the four miRNA targets.*

576 Out of those eight transcripts, only three of them were sufficiently expressed: SAMD12-202
577 (ENST00000409003), PURB-201 (ENST00000395699), and PDE3A-201 (ENST00000359062). Two out
578 of the three did significantly differ from the overall control expression: PDE3A-201 (p.adj. < 0.001) and
579 PURB-201 (p. adj. < 0.05). Even so, we observed that the exposure expression of both transcripts was not
580 significantly different from the control technical replicates (AMI_TOX_000_1 and AMI_TOX_000_2).
581 In addition, we were able to note a tendency to slightly increase over time (Figure 15) for the three of
582 them, although the timeframe where this increase occurred varied between them.

583



584

585 ***Figure 15: PDE3A-201 (ENST00000359062) expression in control and toxic dose of Amiodarone treatment samples (p. adj.***
586 ***value < 0.001)***

587 We also analyzed the proteomics values of all 8 proteins coded by those 4 transcripts. None of the
588 proteins were detected in enough samples to be statistically analyzed.

589 5. Discussion

590 We wanted to evaluate whether the expression of any circular RNAs was affected by different
591 cardiotoxicants in an *in vitro* iPSC-derived cardiomyocyte model. We were able to identify and validate
592 the existence of several circular RNAs that showed differential expression after different compound
593 treatments. For circCDYL, we were able to detect a significant decrease in expression in response to
594 several anthracycline treatments. In addition, most miRNAs hypothesized to be sponged by this circRNA
595 were also found to be over-expressed, thus stimulating post-transcriptional regulation of their targets. The
596 targets' mRNA expression was either not affected or significantly decreased, and one of the proteins
597 (TJP1) coded by such target was identified to be decreased across time where no significant decrease was
598 found for the gene expression; thus, potentially linking the decrease to the studied post-transcriptional
599 regulation. Interestingly, the regulatory axis involving circCDYL/miR-145-5p/TJP1 that we discovered in
600 anthracycline treated cardiac cells, has previously been identified in Wilms' Tumor (cancer of the
601 kidneys[35]). A decrease in the abundance of TJP1 may be inducive to an impairment of the atrium-

ventricle electrical conduction, as the knockout of other tight junction proteins such as CAR (Coxsackievirus-adenovirus receptor) are hypothesized to present cross talk effects with gap junctions[37].

For circGNAS, a significant increase in expression was found after exposure to a toxic dose of Amiodarone. Additionally, these results were successfully reproduced using a different iPSC line and cardiomyocytes derived from that iPSC-line, showing the robustness of iPSC-derived cardiomyocytes for studying circRNAs. Several potential microRNAs were investigated as potential sponged candidates, either showing differential expression or not. To understand what potential synergetic effect could be derived from such different behaviors, we analyzed the targets of the microRNAs overlapping across each other. Although the targets that were expressed did not show any differential expression (as miRNAs do not generally degrade mRNAs directly), we found a mild increasing tendency across time for some of them. Unfortunately, none of the abundances of the proteins coded by those targets was quantified by proteomics due to its limited sensitivity.

To quantify circular RNAs, we used CIRI2. CIRI2, a classical circRNA quantification tool, searches for at least two reads spanning the back-spliced junctions to quantify such circular RNA. We realized that the resulting quantification was very inconsistent: most of the quantified circRNAs were only detected in a single sample, while only a single circular RNA was quantified in all samples. The repercussions were also seen in the statistical results, as most differentially expressed circular RNAs presented missing values for several of the samples, making them less reliable. To increase the sensitivity, we developed a novel strategy. We extracted mature circular RNA sequences from circBase and included them in a classical transcriptome comprising linear mRNAs and used this as a reference for Salmon. The quantification results showed a substantial sensitivity increase (compared to CIRI2), where only a minority of circular RNAs (8%) were quantified in a single sample, though the number of constitutively expressed circRNAs was minimally increased (1 versus 9 circRNAs respectively). The sensitivity increase was probably due to the sacrifice of CIRI2's specificity, as Salmon (circBase) does not require the presence of reads that span the back-spliced junctions.

Interestingly enough, substantial experimental evidence validated the existence of circular RNAs quantified by the remapping strategy. This was surprising considering the stochastic quantification of circRNAs from CIRI2, and the hypothesis by some researchers that circRNAs are mostly a product of erroneous splicing. Even more interestingly, most of the circRNAs experimentally validated were not found in our CIRI2 results, certainly for having a lack of sample coverage preventing a consistent identification of back-spliced junctions. Even so, we recognized that using exhaustive mapping for circRNA quantification had its limitations. For this strategy, quantifying a linear RNA and a circular one with the same exon sequence was equivalent, since the circRNA sequences extracted from circBase were

in the traditional 5` to 3` FASTA format and did not span the back-spliced junction as a continuous sequence.

Hence, for a consistent quantification with tools such as CIRI2, a high sequencing depth is required to increase the chance of sequencing reads that span the required back-spliced junction. Pooling together the sequencing reads of all samples might help to increase the chance of identifying existing circular RNAs. Even so, the quantification at the sample level will still be limited, as the *in silico* predictions will be derived independently from each sample.

Further research should investigate the biological impact of the validated circular RNAs. This could be accomplished either by overexpressing such circRNA (as is the case for circGNAS) or selectively repressing their expression (as we see in circCDYL). One of the main aims would be to analyze the regulatory potency of those circRNAs in relation to the translation of the targets involved. In addition, one could then further elucidate whether such an effect might be one of the main drivers of the cascade leading to the toxicity exerted by the compound.

In conclusion, we have identified, quantified, and validated several circular RNAs whose expression has significantly differed after being exposed to certain cardiotoxicants, in particular circCDYL and circGNAS. In addition, we analyzed the differential expression of potentially sponged miRNAs, their correspondent targets, and the proteins coded by those targets, finding for some of them a potential additive effect in different post-transcriptional regulatory events. These circRNAs might help understand the toxic effects/consequences those compounds induce at the molecular level.

6. Supplementary data

6.1. QPCR Primer Sequences

circBase ID	Gene of origin	5'-3'	3'-5'
hsa_circ_0008285	circCDYL	ACAGGCTTAGCTGTTAACGGGA	GTCATAGCCTTTCCACCGAACC
hsa_circ_0010791	CDC42	CCCACCTTCCCAAACCTA	CCCAACAAGCAAGAAAG

		AT	GAG
hsa_circ_0026129	TUBA1A	CAGACCCAAGCTGTCCATTT	CTCCAGCTTGGACTTCTTGC
hsa_circ_0034356	ACTC1	GCCCTGGATTTTGAGAATGA	ATGGACAGGGTCAGTTGGAG
hsa_circ_0055922	FHL2	CGAGTAAGGCACACCCAAAT	GACTCCTGGCTTTTCAGCAAC
hsa_circ_0060999	circGNAS	AAAACAGCAGCAGCAAACAA	CAGCCATCTGTTGTTCCAGA
hsa_circ_0076194	CDKN1A	CAGGGACCACACCCTGTACT	GTTCTGACATGGCGCTTACA
hsa_circ_0078905	None	CCCCTCACACACTTGGTTTT	TACCTGCCCCAGACTGACTT
hsa_circ_0090448	TIMP1	TGTTCCCACTCCCATCTTTC	GCTATCAGCCACAGCAACAA
hsa_circ_0090904	MSN	TTTGGAGGGGTTTATGCTCA	TGGATCATGTCATTGGCAGT
hsa_circ_0102325	HIF1A	CCAAACAGAGCAGGAAAAGG	GGTGAGGGGAGCATTACATC
hsa_circ_0001445	circSMARCA5	CCAAGATGGGCGAAAGTTCAC	AGATTCTGATCCACAAGCCTCC
None	B-Actin	ATGGATGACGATATCGCT	ATGAGGTAGTCTGTCAGGT

6.2. Sanger Primer Sequences

circBase ID	Gene of origin	Sequence	Direction

hsa_circ_0008285	circCDYL	CAGGAAACAGCTATGACCGTCATAGCCTTTCCACCGAACC	3'-5'
hsa_circ_0060999	circGNAS	TGTAAAACGACGGCCAGTAAAACAGCAGCAGCAAACAA	5'-3'
hsa_circ_0001445	circSMARCA5	TGTAAAACGACGGCCAGTCCAAACAGAGCAGGAAAAGG	5'-3'

7. Bibliography

1 Panda, A. C., Grammatikakis, I., Munk, R., Gorospe, M. & Abdelmohsen, K. in *Wiley Interdisciplinary Reviews: RNA* Vol. 8 (Blackwell Publishing Ltd, 2017).

2 Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333-338, doi:10.1038/nature11928 (2013).

3 Glažar, P., Papavasileiou, P. & Rajewsky, N. in *RNA* Vol. 20 1666 (Cold Spring Harbor Laboratory Press, 2014).

4 Maass, P. G. *et al.* in *Journal of Molecular Medicine* Vol. 95 1179-1189 (Springer Verlag, 2017).

5 Gupta, S. K. *et al.* in *Circulation Research* Vol. 122 246-254 (Lippincott Williams and Wilkins, 2018).

6 Wang, X. *et al.* in *Life Sciences* Vol. 265 (Elsevier Inc., 2021).

7 Yang, M. H. *et al.* in *Aging* Vol. 12 2530-2544 (Impact Journals LLC, 2020).

8 Deng, Y., Wang, J., Xie, G., Zeng, X. & Li, H. in *International Journal of Biological Sciences* Vol. 15 2484-2496 (Ivyspring International Publisher, 2019).

9 Zhao, Y. *et al.* in *Journal of Applied Toxicology* (John Wiley and Sons Ltd, 2021).

10 Gao, Y. *et al.* Comprehensive identification of internal structure and alternative splicing events in circular RNAs. *Nat Commun* **7**, 12060, doi:10.1038/ncomms12060 (2016).

11 Metge, F., Czaja-Hasse, L. F., Reinhardt, R. & Dieterich, C. FUCHS-towards full circular RNA characterization using RNAseq. *PeerJ* **5**, e2934, doi:10.7717/peerj.2934 (2017).

12 Zhang, X. O. *et al.* Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res* **26**, 1277-1287, doi:10.1101/gr.202895.115 (2016).

13 Kuepfer, L. *et al.* A model-based assay design to reproduce in vivo patterns of acute drug-induced toxicity. *Arch Toxicol* **92**, 553-555, doi:10.1007/s00204-017-2041-7 (2018).

14 Gao, Y., Zhang, J. & Zhao, F. Circular RNA identification based on multiple seed matching. *Brief Bioinform* **19**, 803-810, doi:10.1093/bib/bbx014 (2018).
15 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
16 Cunningham, F. *et al.* Ensembl 2019. *Nucleic Acids Res* **47**, D745-D751, doi:10.1093/nar/gky1113 (2019).
17 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419, doi:10.1038/nmeth.4197 (2017).
18 Verheijen, M. C. *et al.* R-ODAF: Omics data analysis framework for regulatory application. *Regul Toxicol Pharmacol* **131**, 105143, doi:10.1016/j.yrtph.2022.105143 (2022).
19 Siede, D. *et al.* Identification of circular RNAs with host gene-independent expression in human model systems for cardiac differentiation and disease. *J Mol Cell Cardiol* **109**, 48-56, doi:10.1016/j.yjmcc.2017.06.015 (2017).
20
21 Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**, doi:10.7554/eLife.05005 (2015).
22 John, B. *et al.* Human MicroRNA targets. *PLoS Biol* **2**, e363, doi:10.1371/journal.pbio.0020363 (2004).
23 Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**, D140-144, doi:10.1093/nar/gkj112 (2006).
24 Lian, X. *et al.* Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating Wnt/beta-catenin signaling under fully defined conditions. *Nat Protoc* **8**, 162-175, doi:10.1038/nprot.2012.150 (2013).
25 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402, doi:10.1093/nar/25.17.3389 (1997).
26 Selevsek, N. *et al.* Network integration and modelling of dynamic drug responses at multi-omics levels. *Commun Biol* **3**, 573, doi:10.1038/s42003-020-01302-8 (2020).
27
28 Wink, S., Hiemstra, S. W., Huppelschoten, S., Klip, J. E. & van de Water, B. Dynamic imaging of adaptive stress response pathway activation for prediction of drug induced liver injury. *Arch Toxicol* **92**, 1797-1814, doi:10.1007/s00204-018-2178-z (2018).

29 Wang, S. *et al.* circCDYL Acts as a Tumor Suppressor in Triple Negative Breast Cancer by Sponging miR-190a-3p and Upregulating TP53INP1. *Clin Breast Cancer* **20**, 422-430, doi:10.1016/j.clbc.2020.04.006 (2020).

30 Bian, W. X., Xue, F., Wang, L. Y. & Xing, X. F. Circular RNA CircCDYL Regulates Proliferation and Apoptosis in Non-Small Cell Lung Cancer Cells by Sponging miR-185-5p and Upregulating TNRC6A. *Cancer Manag Res* **13**, 633-642, doi:10.2147/CMAR.S280315 (2021).

31 Zhang, M. *et al.* Circular RNA (circRNA) CDYL Induces Myocardial Regeneration by ceRNA After Myocardial Infarction. *Med Sci Monit* **26**, e923188, doi:10.12659/MSM.923188 (2020).

32 Cui, W., Dai, J., Ma, J. & Gu, H. circCDYL/microRNA-105-5p participates in modulating growth and migration of colon cancer cells. *Gen Physiol Biophys* **38**, 485-495, doi:10.4149/gpb2019037 (2019).

33 Wei, Y. *et al.* A Noncoding Regulatory RNAs Network Driven by Circ-CDYL Acts Specifically in the Early Stages Hepatocellular Carcinoma. *Hepatology* **71**, 130-147, doi:10.1002/hep.30795 (2020).

34 Liang, G. *et al.* MiR-92b-3p Inhibits Proliferation of HER2-Positive Breast Cancer Cell by Targeting circCDYL. *Front Cell Dev Biol* **9**, 707049, doi:10.3389/fcell.2021.707049 (2021).

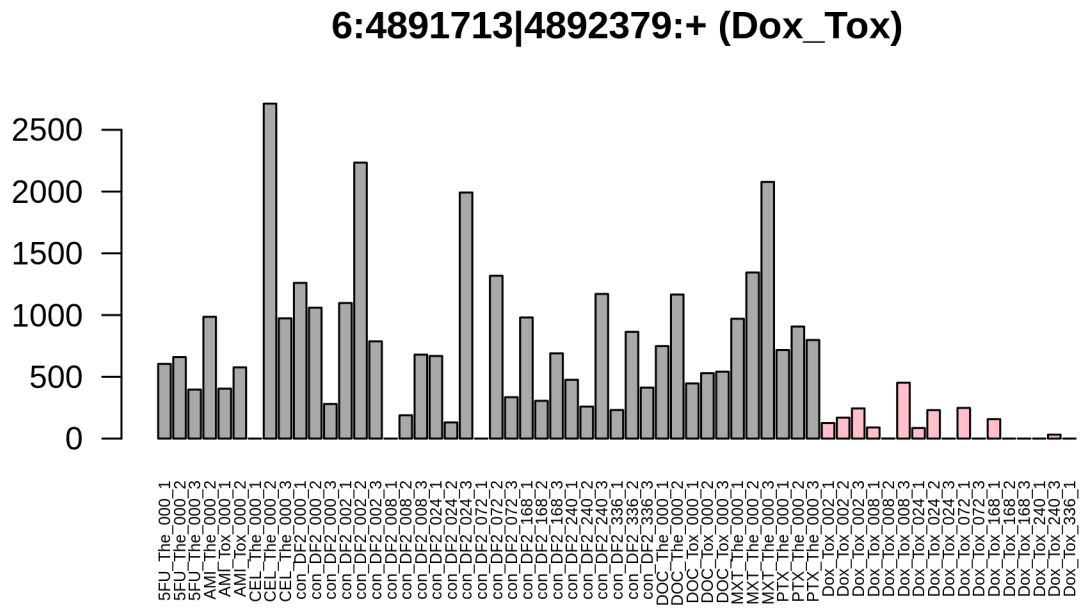
35 Zhou, R. *et al.* CircCDYL Acts as a Tumor Suppressor in Wilms' Tumor by Targeting miR-145-5p. *Front Cell Dev Biol* **9**, 668947, doi:10.3389/fcell.2021.668947 (2021).

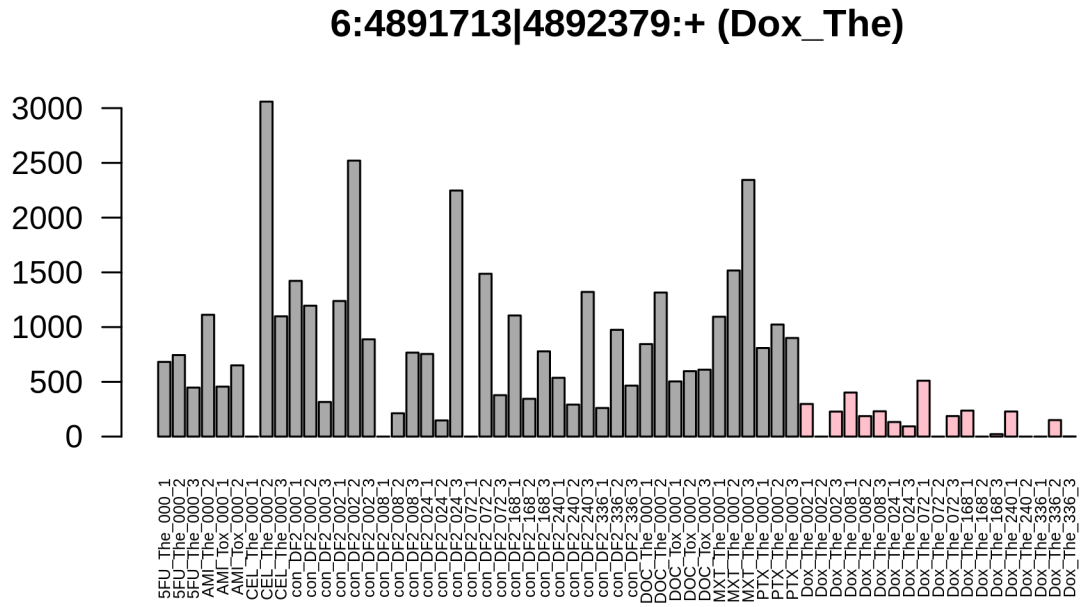
36 Chen, F. *et al.* Circular RNA circ-CDYL sponges miR-1180 to elevate yes-associated protein in multiple myeloma. *Exp Biol Med (Maywood)* **245**, 925-932, doi:10.1177/1535370220918191 (2020).

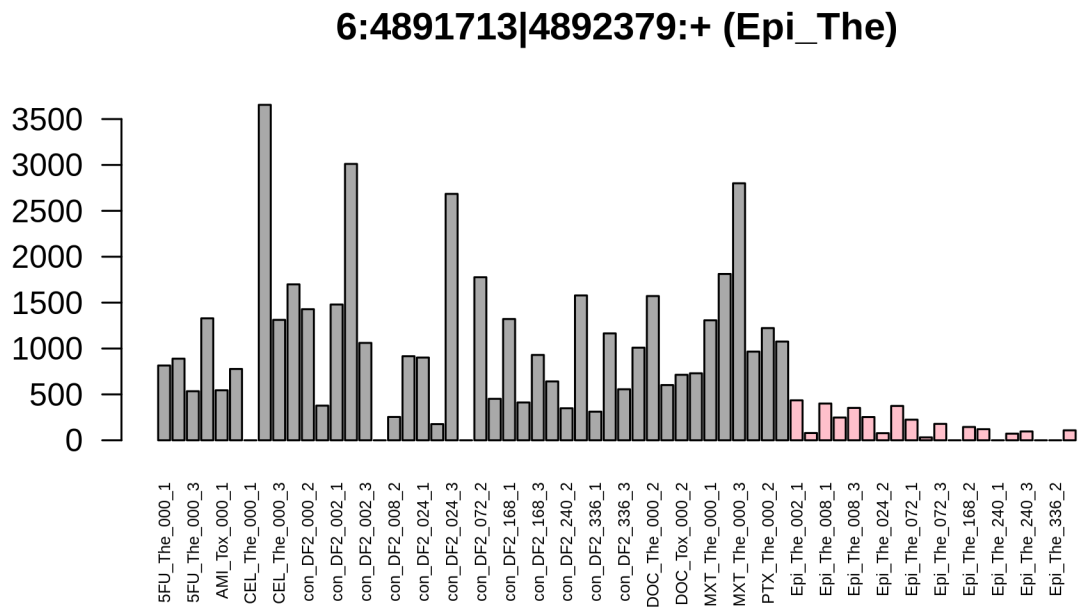
37 Lisewski, U. *et al.* The tight junction protein CAR regulates cardiac conduction and cell-cell communication. *J Exp Med* **205**, 2369-2379, doi:10.1084/jem.20080897 (2008).

Supplementary Results (Chapter 2)

CircCDYL

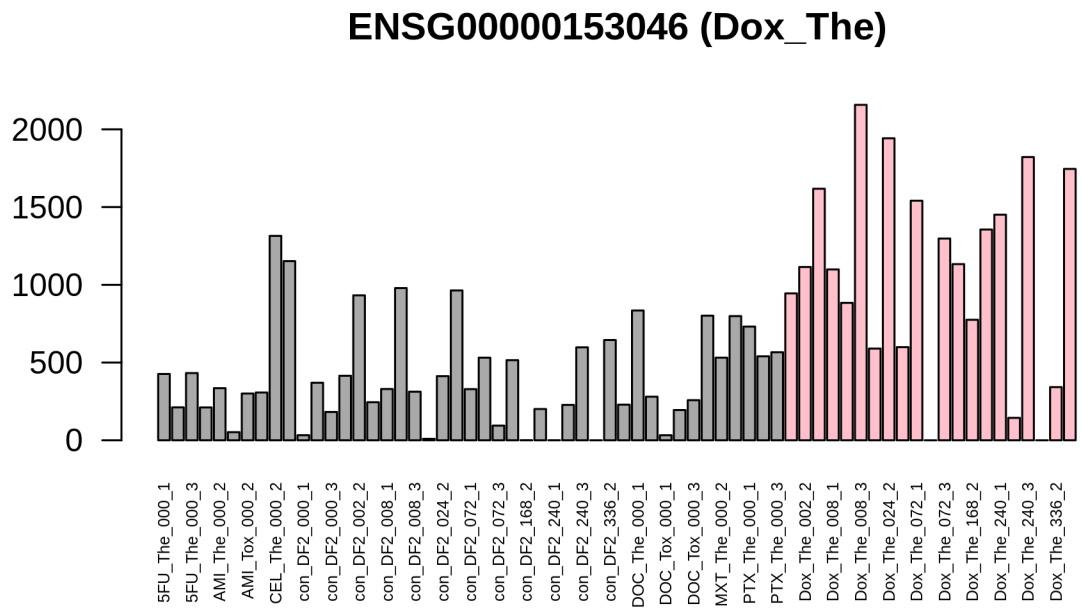




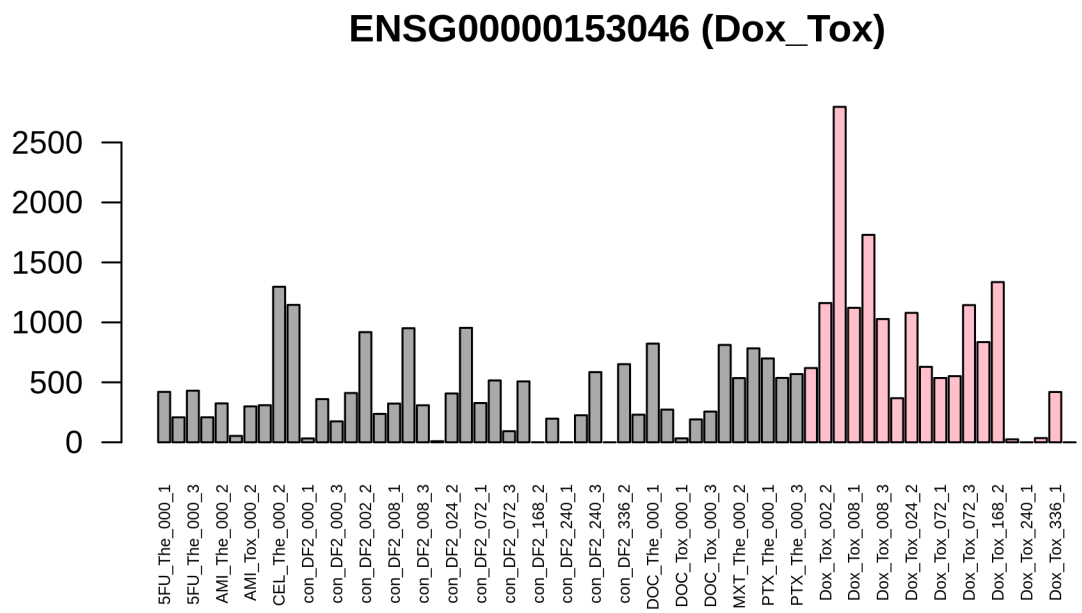

6:4891713|4892379:+ (Epi_The)
3500
3000
2500
2000
1500
1000
500
0
5FU_The_000_1
5FU_The_000_3
AMI_Tox_000_1
CEL_The_000_1
CEL_The_000_3
con_DF2_000_2
con_DF2_002_1
con_DF2_002_3
con_DF2_008_2
con_DF2_024_1
con_DF2_024_3
con_DF2_072_2
con_DF2_168_1
con_DF2_168_3
con_DF2_240_2
con_DF2_336_1
con_DF2_336_3
DOC_The_000_2
DOC_Tox_000_2
MXT_The_000_1
MXT_The_000_3
PTX_The_000_2
Epi_The_002_1
Epi_The_008_1
Epi_The_008_3
Epi_The_024_2
Epi_The_072_1
Epi_The_072_3
Epi_The_168_2
Epi_The_240_1
Epi_The_240_3
Epi_The_336_2

5

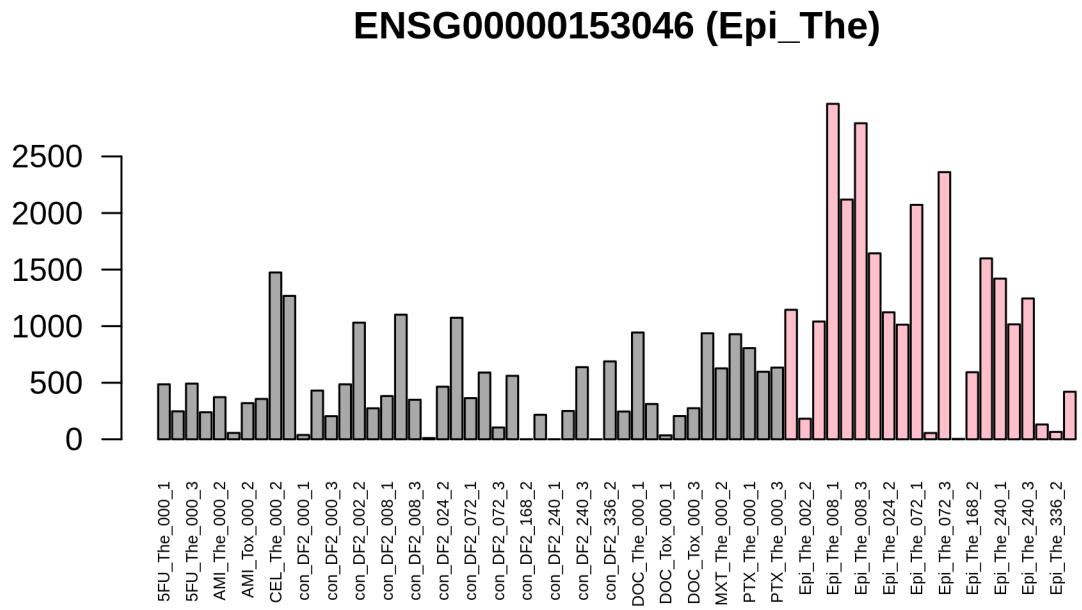
6 CDYL


ENSG00000153046 (Dox_The)
2000
1500
1000
500
0
5FU_The_000_1
5FU_The_000_3
AMI_The_000_2
AMI_Tox_000_2
CEL_The_000_2
con_DF2_000_1
con_DF2_000_3
con_DF2_002_2
con_DF2_008_1
con_DF2_008_3
con_DF2_024_2
con_DF2_072_1
con_DF2_072_3
con_DF2_168_2
con_DF2_240_1
con_DF2_240_3
con_DF2_336_2
DOC_The_000_1
DOC_Tox_000_1
DOC_Tox_000_3
MXT_The_000_2
PTX_The_000_1
PTX_The_000_3
Dox_The_002_2
Dox_The_008_1
Dox_The_008_3
Dox_The_024_2
Dox_The_072_1
Dox_The_072_3
Dox_The_168_2
Dox_The_240_1
Dox_The_240_3
Dox_The_336_2

7

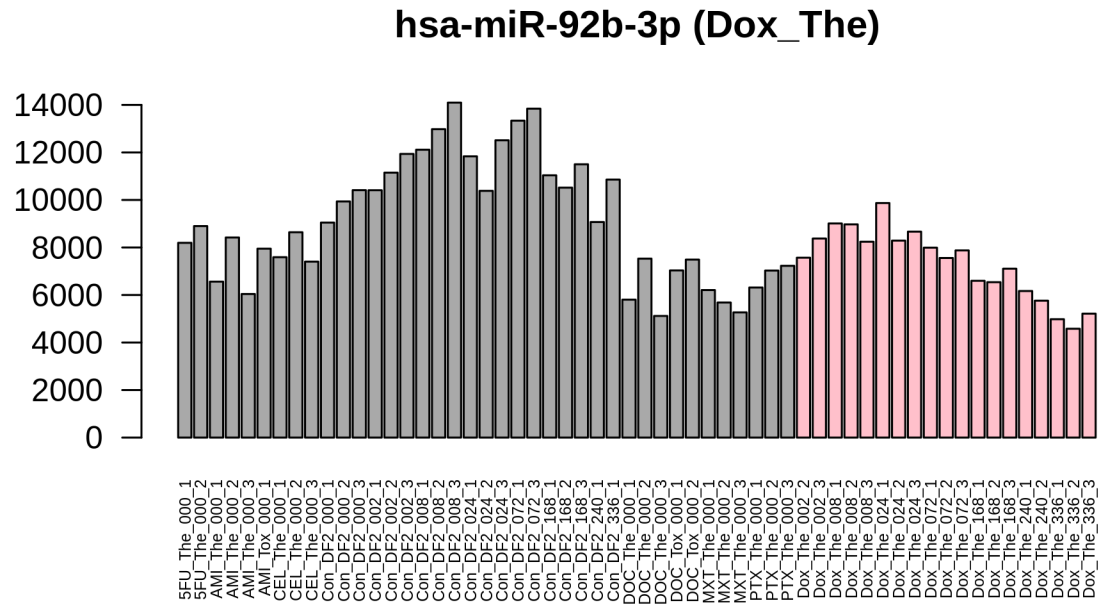

ENSG00000153046 (Dox_Tox)
2500
2000
1500
1000
500
0
5FU_The_000_1
5FU_The_000_3
AMI_The_000_2
AMI_Tox_000_2
CEL_The_000_2
con_DF2_000_1
con_DF2_000_3
con_DF2_002_2
con_DF2_008_1
con_DF2_008_3
con_DF2_024_2
con_DF2_072_1
con_DF2_072_3
con_DF2_168_2
con_DF2_240_1
con_DF2_240_3
con_DF2_336_2
DOC_The_000_1
DOC_Tox_000_1
DOC_Tox_000_3
MXT_The_000_2
PTX_The_000_1
PTX_The_000_3
Dox_Tox_002_2
Dox_Tox_008_1
Dox_Tox_008_3
Dox_Tox_024_2
Dox_Tox_072_1
Dox_Tox_072_3
Dox_Tox_168_2
Dox_Tox_240_1
Dox_Tox_336_1

8

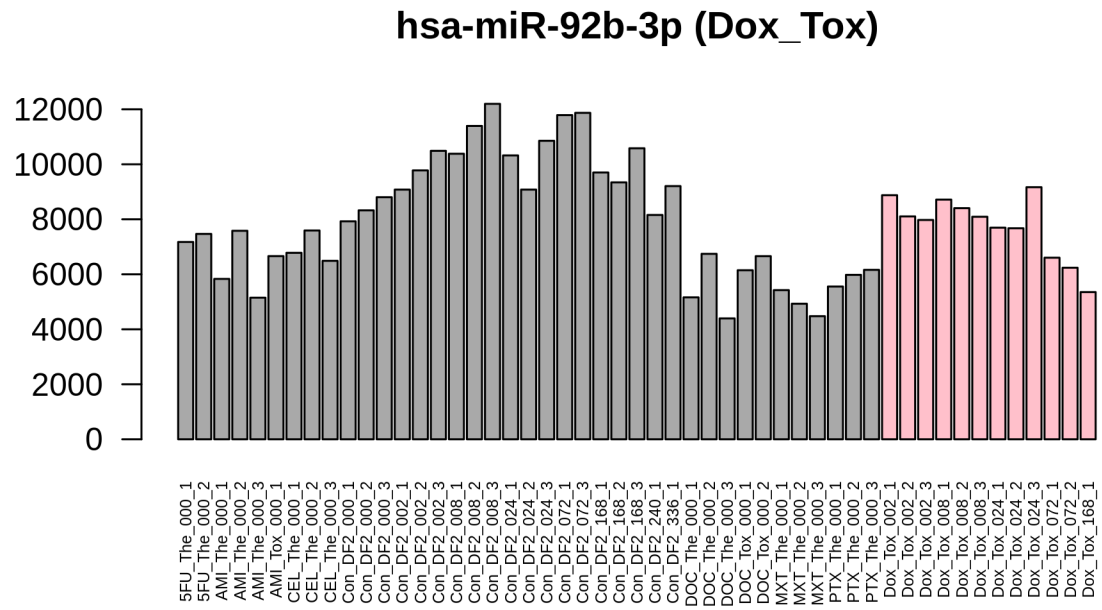

ENSG00000153046 (Epi_The)
2500
2000
1500
1000
500
0
5FU_The_000_1
5FU_The_000_3
AMI_The_000_2
AMI_Tox_000_2
CEL_The_000_2
con_DF2_000_1
con_DF2_000_3
con_DF2_002_2
con_DF2_008_1
con_DF2_008_3
con_DF2_024_2
con_DF2_072_1
con_DF2_072_3
con_DF2_168_2
con_DF2_240_1
con_DF2_240_3
con_DF2_336_2
DOC_The_000_1
DOC_Tox_000_1
DOC_Tox_000_3
MXT_The_000_2
PTX_The_000_1
PTX_The_000_3
Epi_The_002_2
Epi_The_008_1
Epi_The_008_3
Epi_The_024_2
Epi_The_072_1
Epi_The_072_3
Epi_The_168_2
Epi_The_240_1
Epi_The_240_3
Epi_The_336_2

9

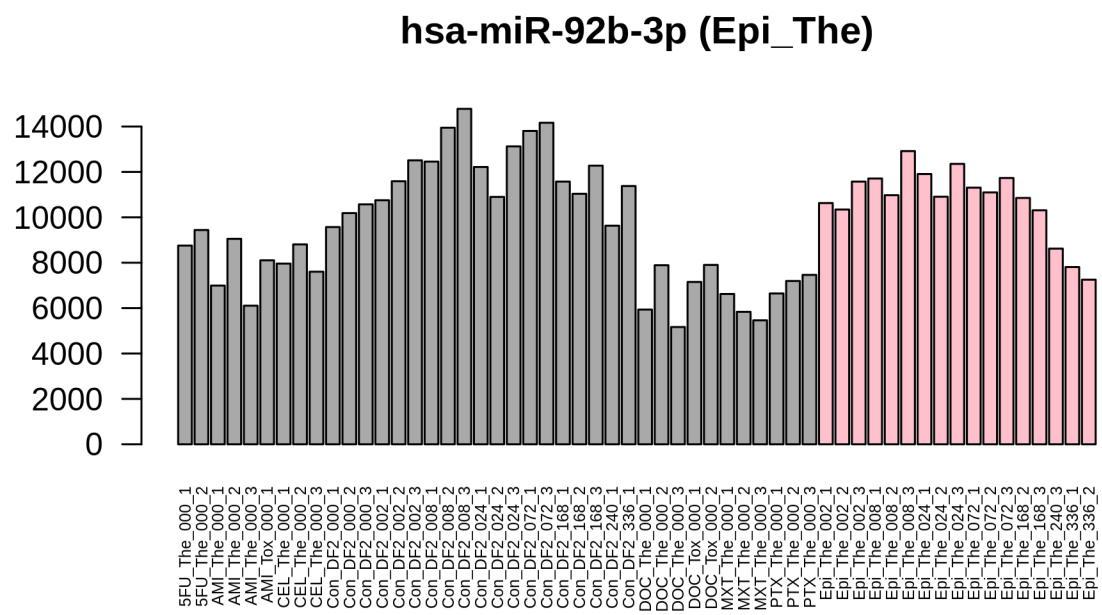
10 hsa-miR-92b-3p



11



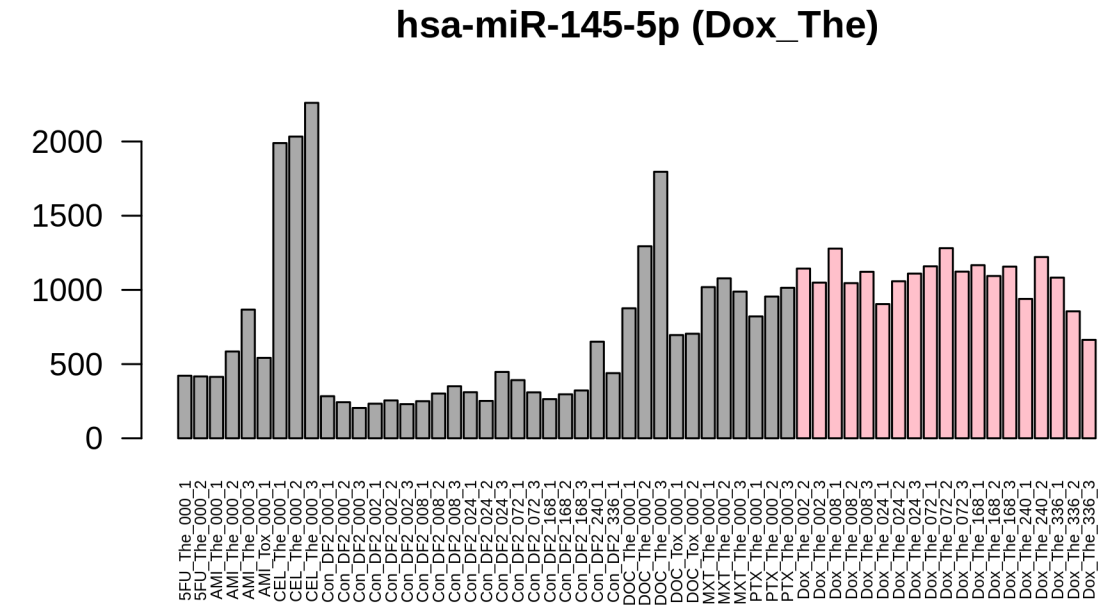
12



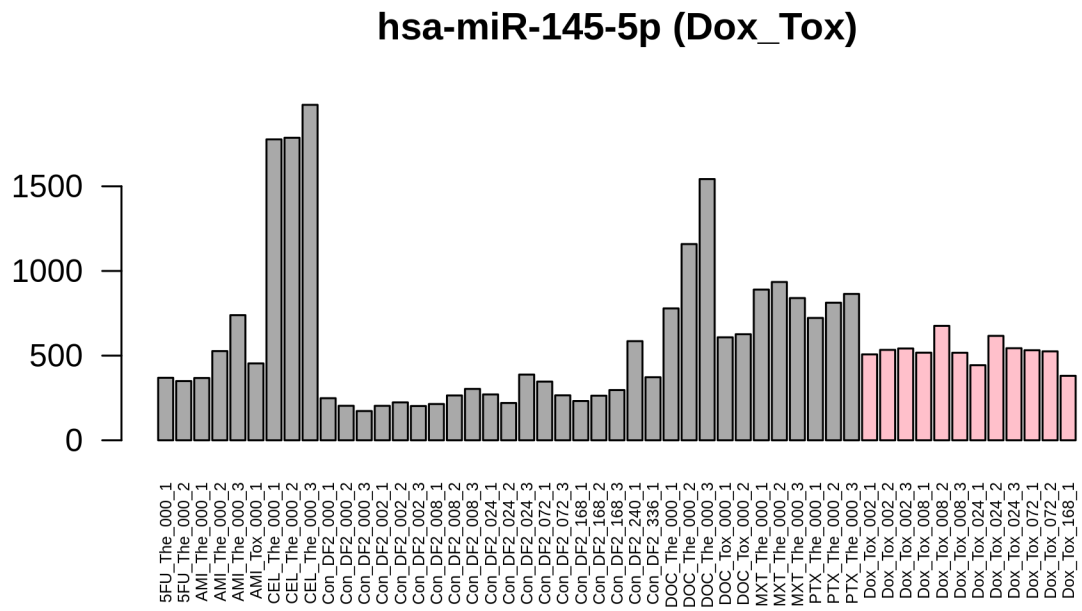
13

14 hsa-miR-145-5p

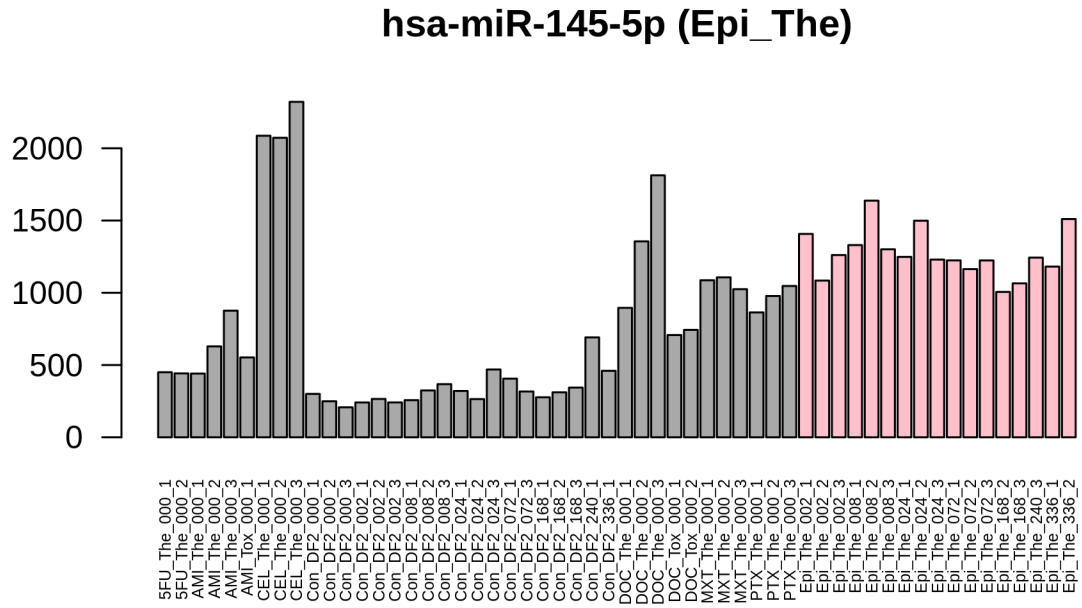
15



16

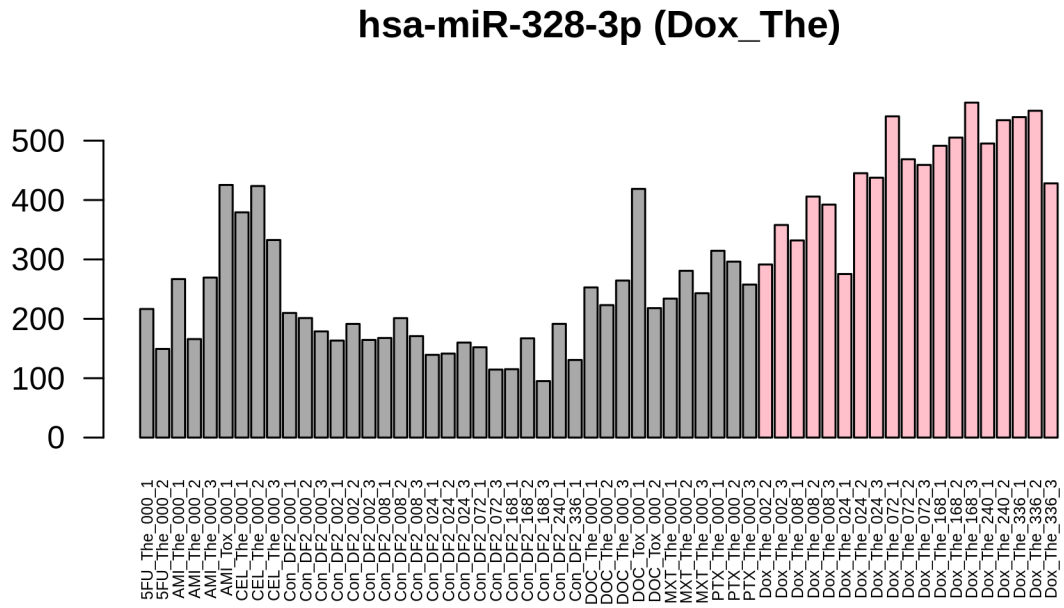

hsa-miR-145-5p (Dox_Tox)
1500
1000
500
0
5FU_The_000_1
5FU_The_000_2
AMI_The_000_1
AMI_The_000_2
AMI_The_000_3
AMI_Tox_000_1
CEL_The_000_1
CEL_The_000_2
CEL_The_000_3
Con_DF2_000_1
Con_DF2_000_2
Con_DF2_000_3
Con_DF2_002_1
Con_DF2_002_2
Con_DF2_002_3
Con_DF2_008_1
Con_DF2_008_2
Con_DF2_008_3
Con_DF2_024_1
Con_DF2_024_2
Con_DF2_024_3
Con_DF2_072_1
Con_DF2_072_3
Con_DF2_168_1
Con_DF2_168_2
Con_DF2_168_3
Con_DF2_240_1
Con_DF2_336_1
DOC_The_000_1
DOC_The_000_2
DOC_The_000_3
DOC_Tox_000_1
DOC_Tox_000_2
MXT_The_000_1
MXT_The_000_2
MXT_The_000_3
PTX_The_000_1
PTX_The_000_2
PTX_The_000_3
Dox_Tox_002_1
Dox_Tox_002_2
Dox_Tox_002_3
Dox_Tox_008_1
Dox_Tox_008_2
Dox_Tox_008_3
Dox_Tox_024_1
Dox_Tox_024_2
Dox_Tox_024_3
Dox_Tox_072_1
Dox_Tox_072_2
Dox_Tox_168_1

17

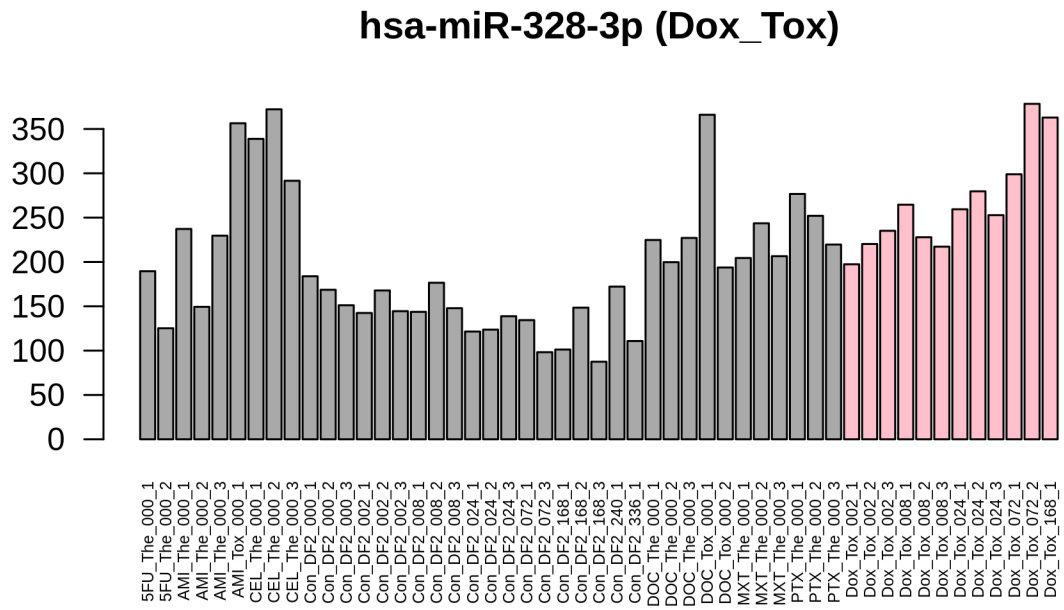

hsa-miR-145-5p (Epi_The)
2000
1500
1000
500
0
5FU_The_000_1
5FU_The_000_2
AMI_The_000_1
AMI_The_000_2
AMI_The_000_3
AMI_Tox_000_1
CEL_The_000_1
CEL_The_000_2
CEL_The_000_3
Con_DF2_000_1
Con_DF2_000_2
Con_DF2_000_3
Con_DF2_002_1
Con_DF2_002_2
Con_DF2_002_3
Con_DF2_008_1
Con_DF2_008_2
Con_DF2_008_3
Con_DF2_024_1
Con_DF2_024_2
Con_DF2_024_3
Con_DF2_072_1
Con_DF2_072_3
Con_DF2_168_1
Con_DF2_168_2
Con_DF2_168_3
Con_DF2_240_1
Con_DF2_336_1
DOC_The_000_1
DOC_The_000_2
DOC_The_000_3
DOC_Tox_000_1
DOC_Tox_000_2
MXT_The_000_1
MXT_The_000_2
MXT_The_000_3
PTX_The_000_1
PTX_The_000_2
PTX_The_000_3
Epi_The_002_1
Epi_The_002_2
Epi_The_002_3
Epi_The_008_1
Epi_The_008_2
Epi_The_008_3
Epi_The_024_1
Epi_The_024_2
Epi_The_024_3
Epi_The_072_1
Epi_The_072_2
Epi_The_072_3
Epi_The_168_2
Epi_The_168_3
Epi_The_240_3
Epi_The_336_1
Epi_The_336_2

18

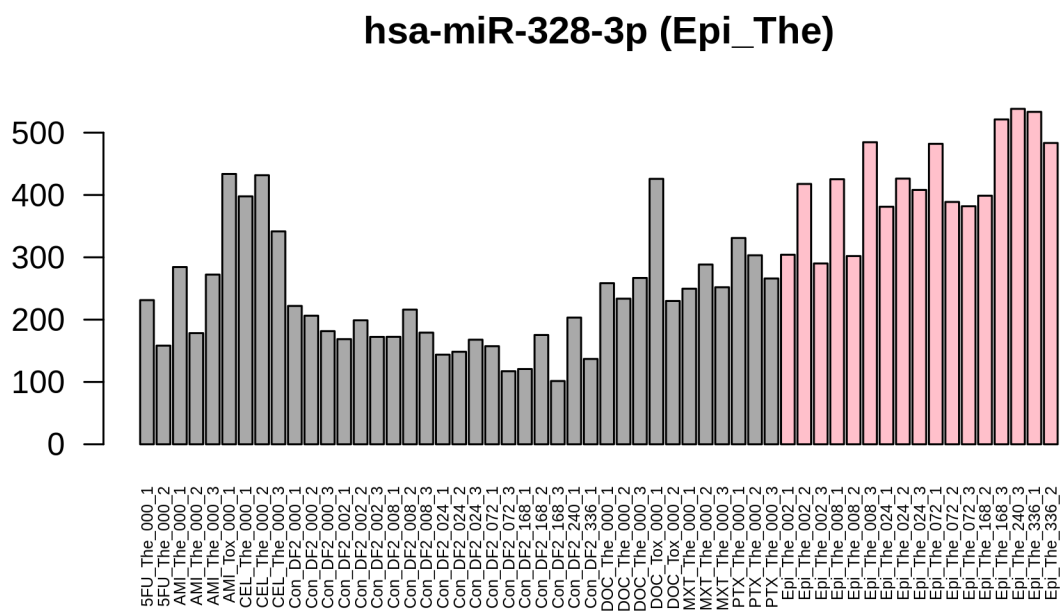
19 hsa-miR-328-3p



20



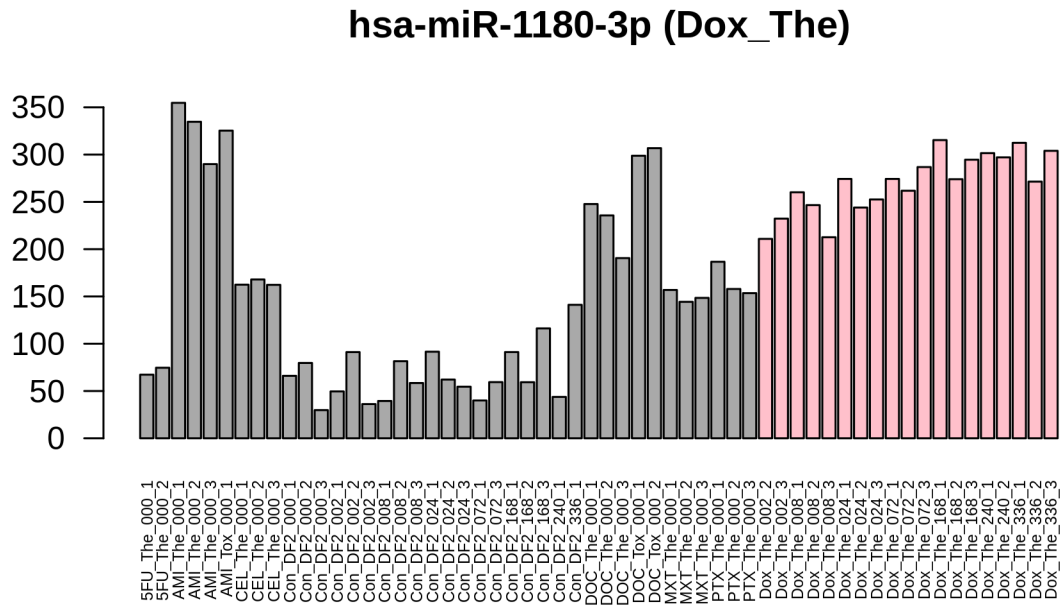
21



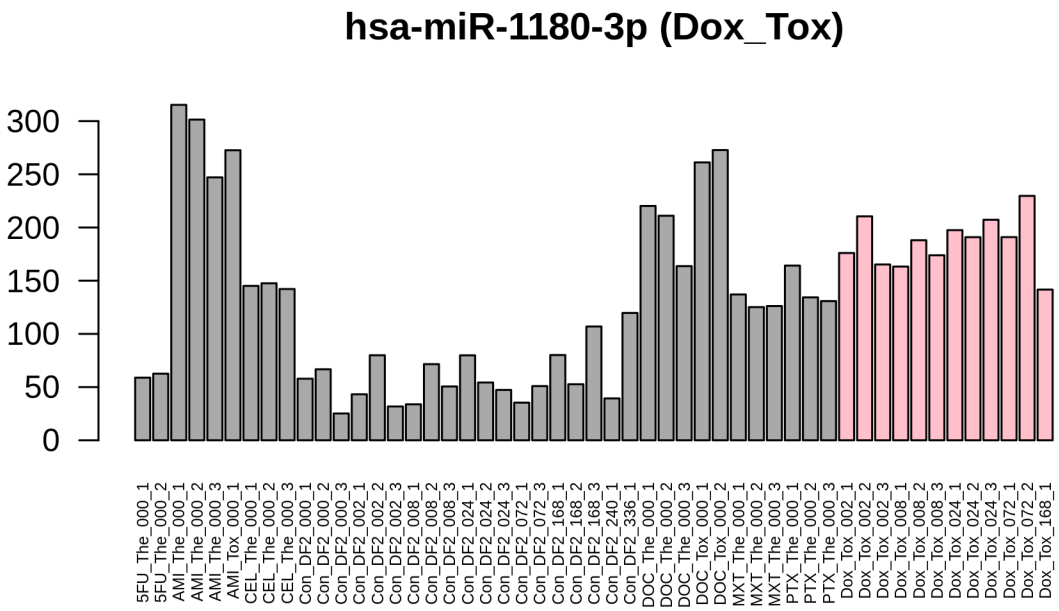
22

23 hsa-miR-1180-3p

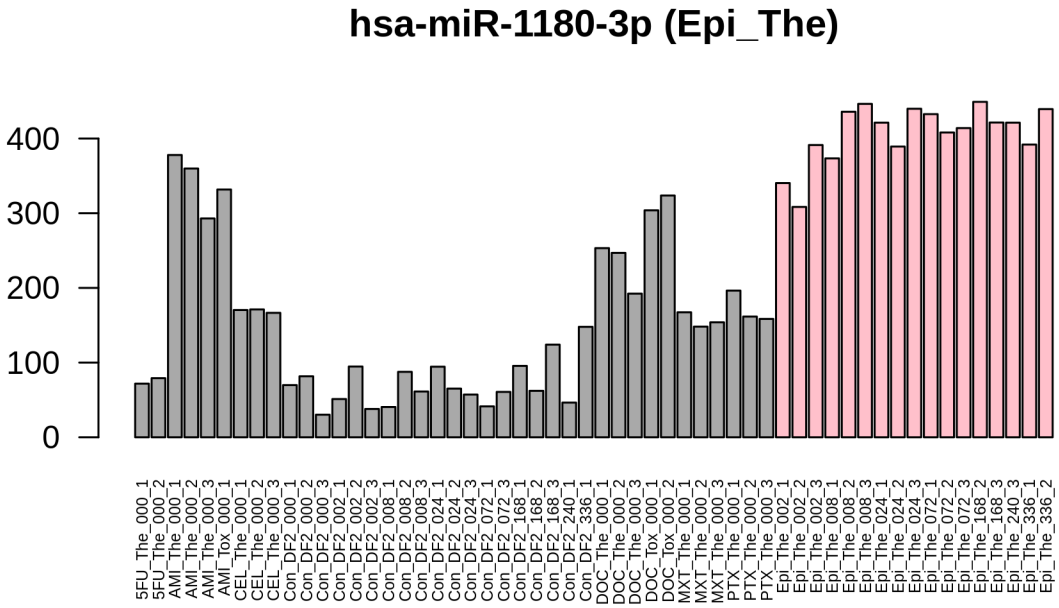
24



25

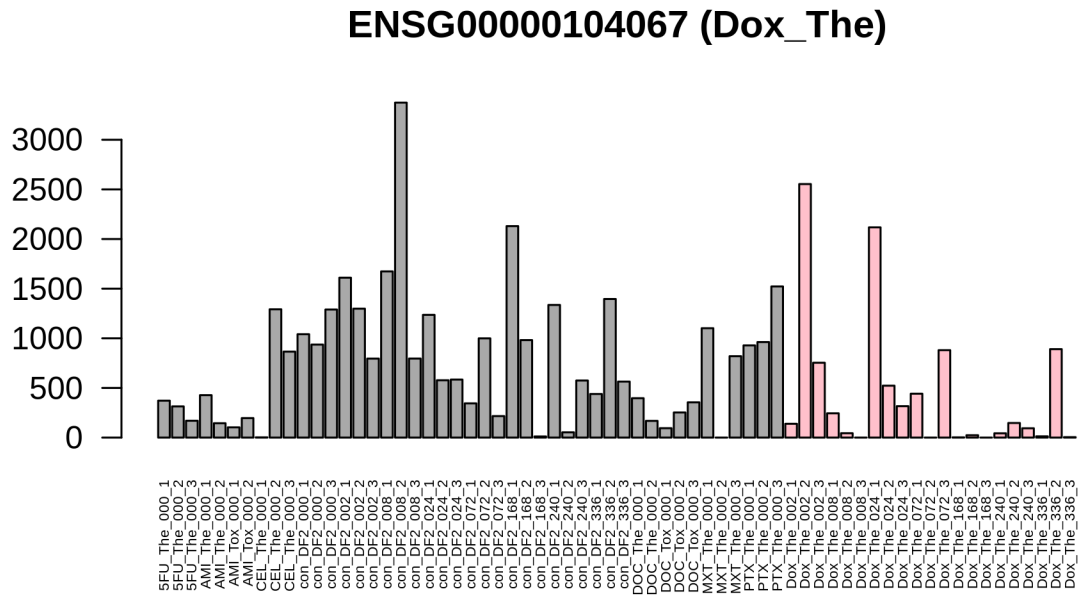

hsa-miR-1180-3p (Dox_Tox)
300
250
200
150
100
50
0
5FU_The_000_1
5FU_The_000_2
AMI_The_000_1
AMI_The_000_2
AMI_The_000_3
AMI_Tox_000_1
CEL_The_000_1
CEL_The_000_2
CEL_The_000_3
Con_DF2_000_1
Con_DF2_000_2
Con_DF2_000_3
Con_DF2_002_1
Con_DF2_002_2
Con_DF2_002_3
Con_DF2_008_1
Con_DF2_008_2
Con_DF2_008_3
Con_DF2_024_1
Con_DF2_024_2
Con_DF2_024_3
Con_DF2_072_1
Con_DF2_072_3
Con_DF2_168_1
Con_DF2_168_2
Con_DF2_168_3
Con_DF2_240_1
Con_DF2_336_1
DOC_The_000_1
DOC_The_000_2
DOC_The_000_3
DOC_Tox_000_1
DOC_Tox_000_2
DOC_Tox_000_3
MXT_The_000_1
MXT_The_000_2
MXT_The_000_3
PTX_The_000_1
PTX_The_000_2
PTX_The_000_3
Dox_Tox_002_1
Dox_Tox_002_2
Dox_Tox_002_3
Dox_Tox_008_1
Dox_Tox_008_2
Dox_Tox_008_3
Dox_Tox_024_1
Dox_Tox_024_2
Dox_Tox_024_3
Dox_Tox_072_1
Dox_Tox_072_2
Dox_Tox_168_1

26

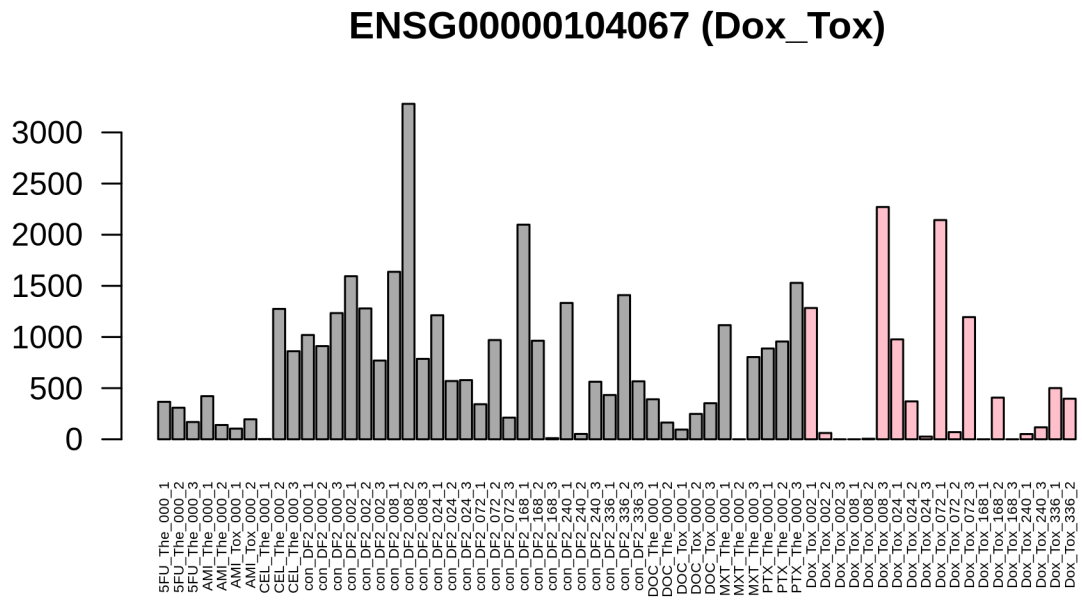

hsa-miR-1180-3p (Epi_The)
400
300
200
100
0
5FU_The_000_1
5FU_The_000_2
AMI_The_000_1
AMI_The_000_2
AMI_The_000_3
AMI_Tox_000_1
CEL_The_000_1
CEL_The_000_2
CEL_The_000_3
Con_DF2_000_1
Con_DF2_000_2
Con_DF2_000_3
Con_DF2_002_1
Con_DF2_002_2
Con_DF2_002_3
Con_DF2_008_1
Con_DF2_008_2
Con_DF2_008_3
Con_DF2_024_1
Con_DF2_024_2
Con_DF2_024_3
Con_DF2_072_1
Con_DF2_072_3
Con_DF2_168_1
Con_DF2_168_2
Con_DF2_168_3
Con_DF2_240_1
Con_DF2_336_1
DOC_The_000_1
DOC_The_000_2
DOC_The_000_3
DOC_Tox_000_1
DOC_Tox_000_2
MXT_The_000_1
MXT_The_000_2
MXT_The_000_3
PTX_The_000_1
PTX_The_000_2
PTX_The_000_3
Epi_The_002_1
Epi_The_002_2
Epi_The_002_3
Epi_The_008_1
Epi_The_008_2
Epi_The_008_3
Epi_The_024_1
Epi_The_024_2
Epi_The_024_3
Epi_The_072_1
Epi_The_072_2
Epi_The_072_3
Epi_The_168_2
Epi_The_168_3
Epi_The_240_3
Epi_The_336_1
Epi_The_336_2

27

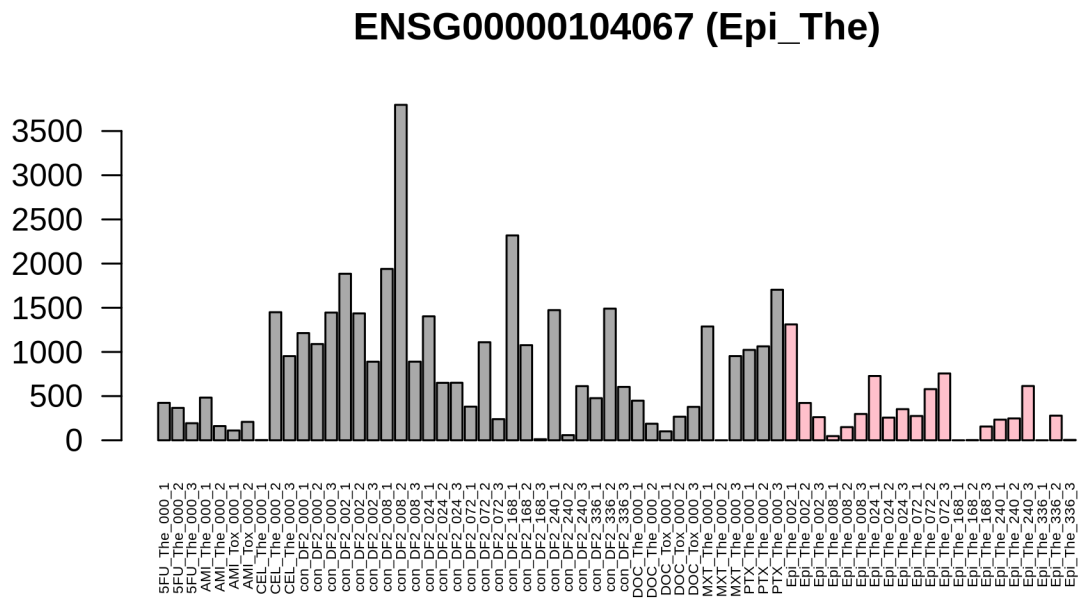
28 TJP1 (ENSG00000104067)



29



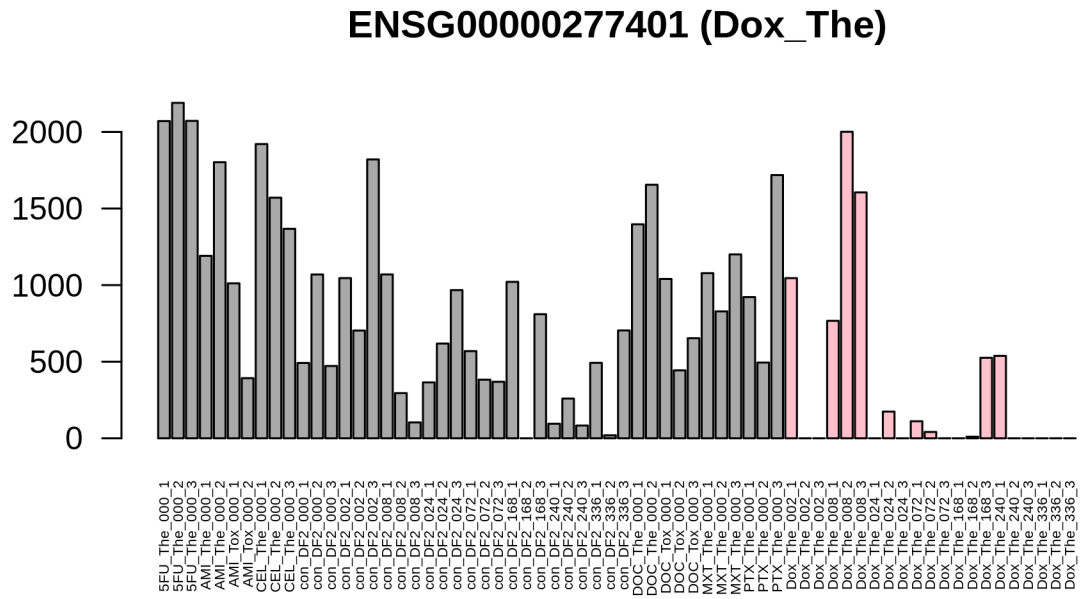
30



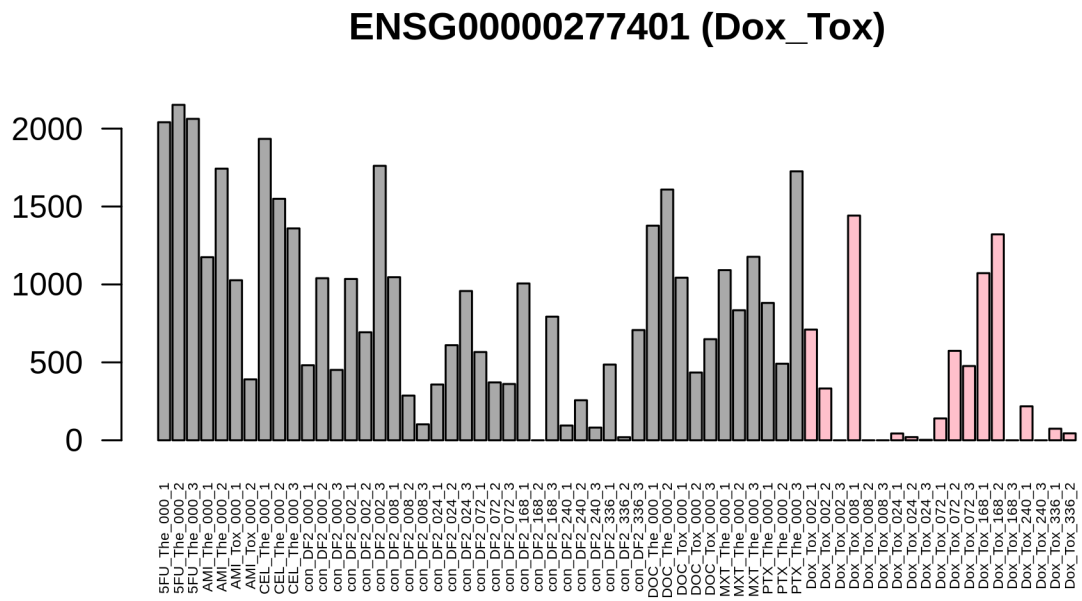
31

32 TJP1 (ENSG00000277401)

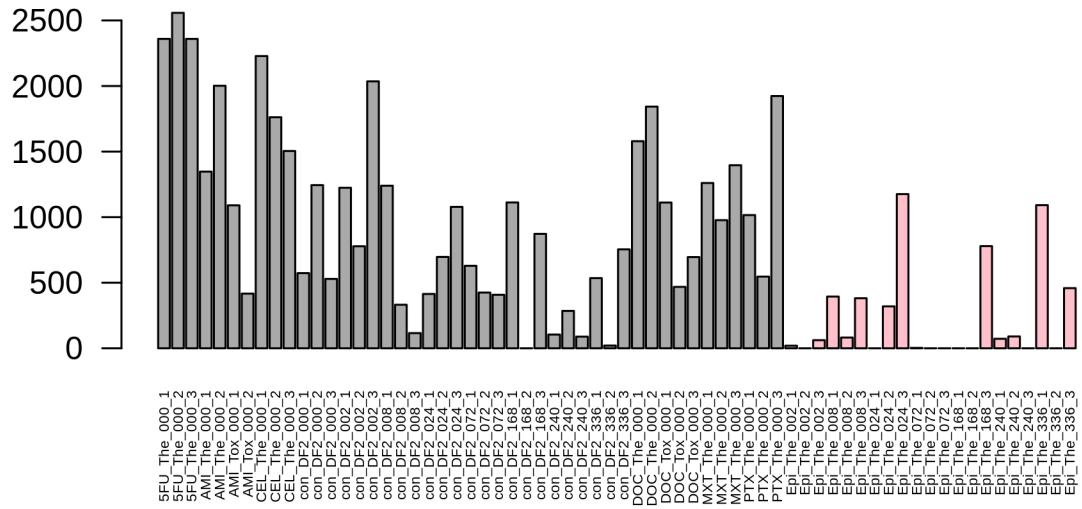
33



34

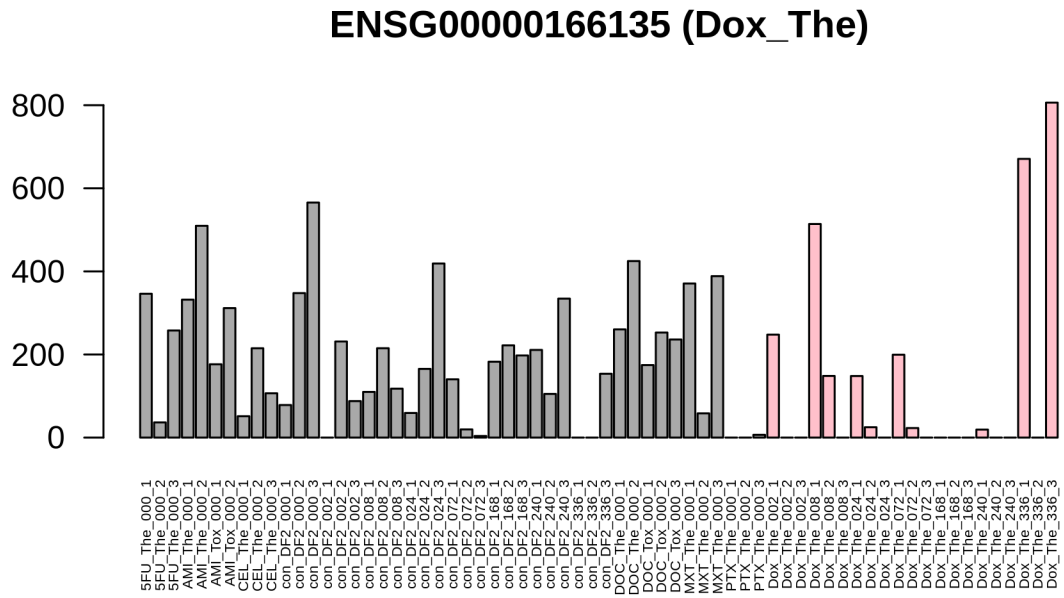

ENSG00000277401 (Dox_Tox)
2000
1500
1000
500
0
5FU_The_000_1
5FU_The_000_2
5FU_The_000_3
AMI_The_000_1
AMI_The_000_2
AMI_Tox_000_1
AMI_Tox_000_2
CEL_The_000_1
CEL_The_000_2
CEL_The_000_3
con_DF2_000_1
con_DF2_000_2
con_DF2_000_3
con_DF2_002_1
con_DF2_002_2
con_DF2_002_3
con_DF2_008_1
con_DF2_008_2
con_DF2_008_3
con_DF2_024_1
con_DF2_024_2
con_DF2_024_3
con_DF2_072_1
con_DF2_072_2
con_DF2_072_3
con_DF2_168_1
con_DF2_168_2
con_DF2_168_3
con_DF2_240_1
con_DF2_240_2
con_DF2_240_3
con_DF2_336_1
con_DF2_336_2
con_DF2_336_3
DOC_The_000_1
DOC_The_000_2
DOC_Tox_000_1
DOC_Tox_000_2
DOC_Tox_000_3
MXT_The_000_1
MXT_The_000_2
MXT_The_000_3
PTX_The_000_1
PTX_The_000_2
PTX_The_000_3
Dox_Tox_002_1
Dox_Tox_002_2
Dox_Tox_002_3
Dox_Tox_008_1
Dox_Tox_008_2
Dox_Tox_008_3
Dox_Tox_024_1
Dox_Tox_024_2
Dox_Tox_024_3
Dox_Tox_072_1
Dox_Tox_072_2
Dox_Tox_072_3
Dox_Tox_168_1
Dox_Tox_168_2
Dox_Tox_168_3
Dox_Tox_240_1
Dox_Tox_240_2
Dox_Tox_240_3
Dox_Tox_336_1
Dox_Tox_336_2
Dox_Tox_336_3

35

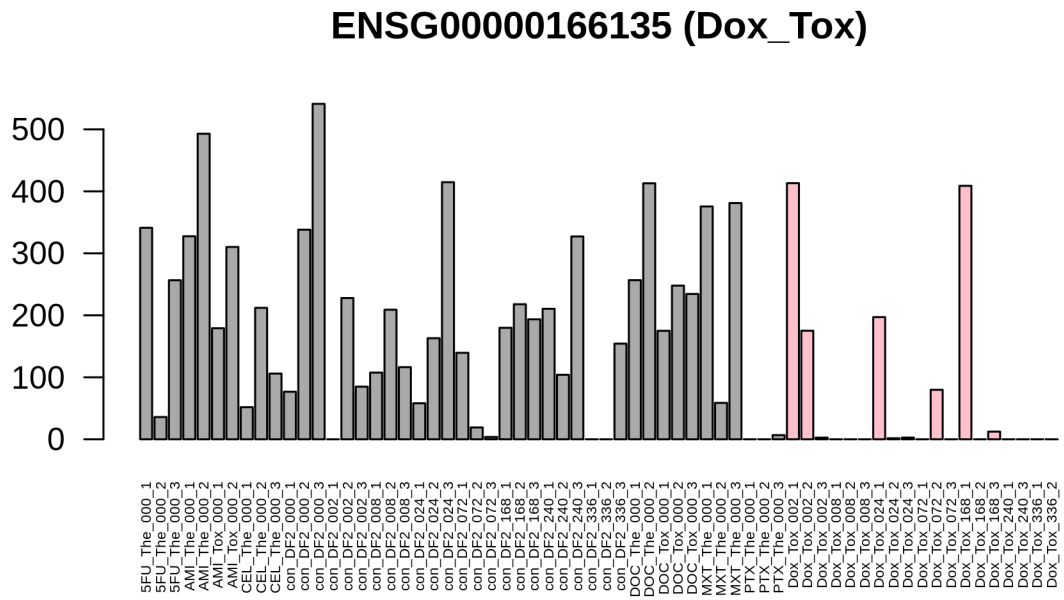

ENSG00000277401 (Epi_The)
2500
2000
1500
1000
500
0
5FU_The_000_1
5FU_The_000_2
5FU_The_000_3
AMI_The_000_1
AMI_The_000_2
AMI_Tox_000_1
AMI_Tox_000_2
CEL_The_000_1
CEL_The_000_2
CEL_The_000_3
con_DF2_000_1
con_DF2_000_2
con_DF2_000_3
con_DF2_002_1
con_DF2_002_2
con_DF2_002_3
con_DF2_008_1
con_DF2_008_2
con_DF2_008_3
con_DF2_024_1
con_DF2_024_2
con_DF2_024_3
con_DF2_072_1
con_DF2_072_2
con_DF2_072_3
con_DF2_168_1
con_DF2_168_2
con_DF2_168_3
con_DF2_240_1
con_DF2_240_2
con_DF2_240_3
con_DF2_336_1
con_DF2_336_2
con_DF2_336_3
DOC_The_000_1
DOC_The_000_2
DOC_Tox_000_1
DOC_Tox_000_2
DOC_Tox_000_3
MXT_The_000_1
MXT_The_000_2
MXT_The_000_3
PTX_The_000_1
PTX_The_000_2
PTX_The_000_3
Epi_The_002_1
Epi_The_002_2
Epi_The_002_3
Epi_The_008_1
Epi_The_008_2
Epi_The_008_3
Epi_The_024_1
Epi_The_024_2
Epi_The_024_3
Epi_The_072_1
Epi_The_072_2
Epi_The_072_3
Epi_The_168_1
Epi_The_168_2
Epi_The_168_3
Epi_The_240_1
Epi_The_240_2
Epi_The_240_3
Epi_The_336_1
Epi_The_336_2
Epi_The_336_3

36

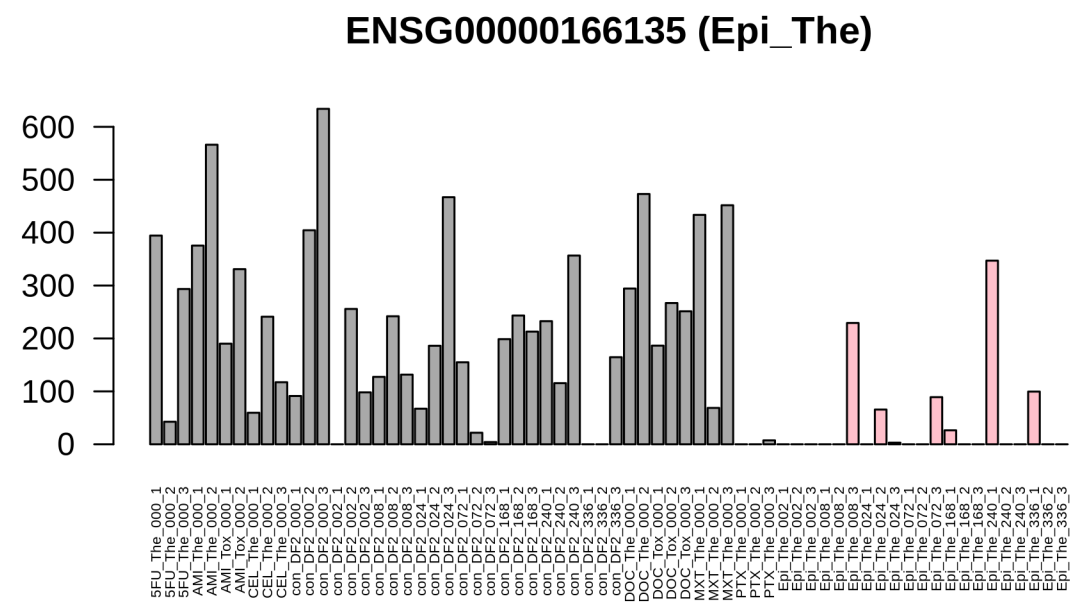
37 HIF1AN



38



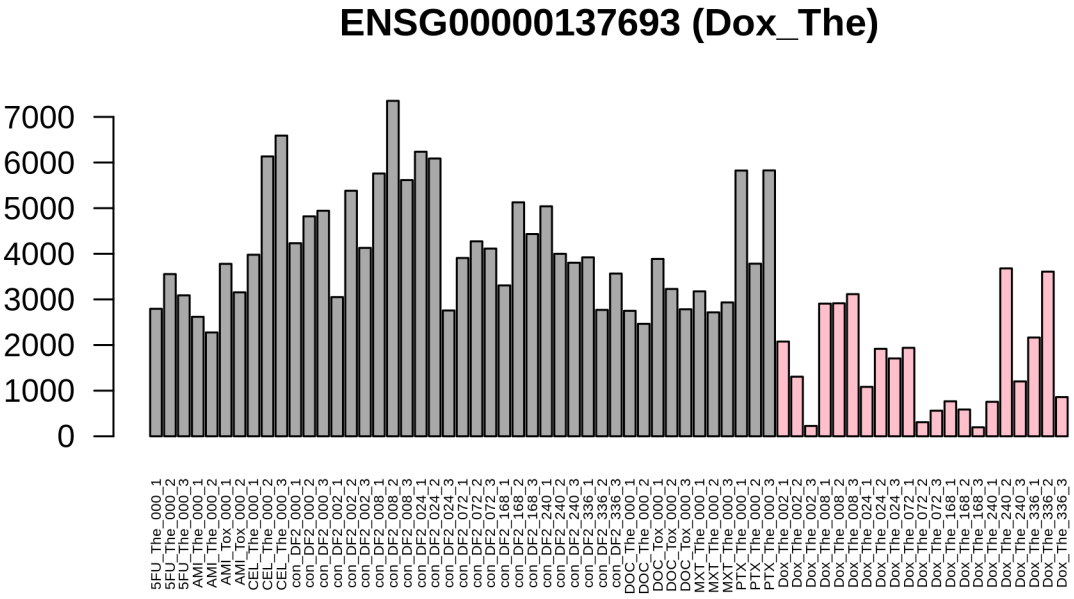
39



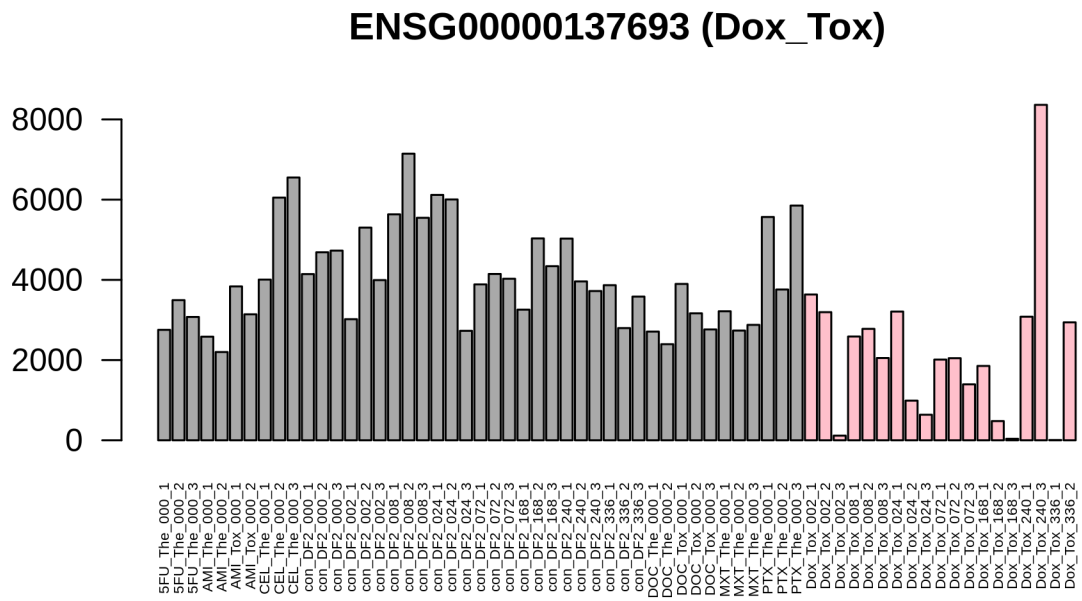
40

41 YAP1

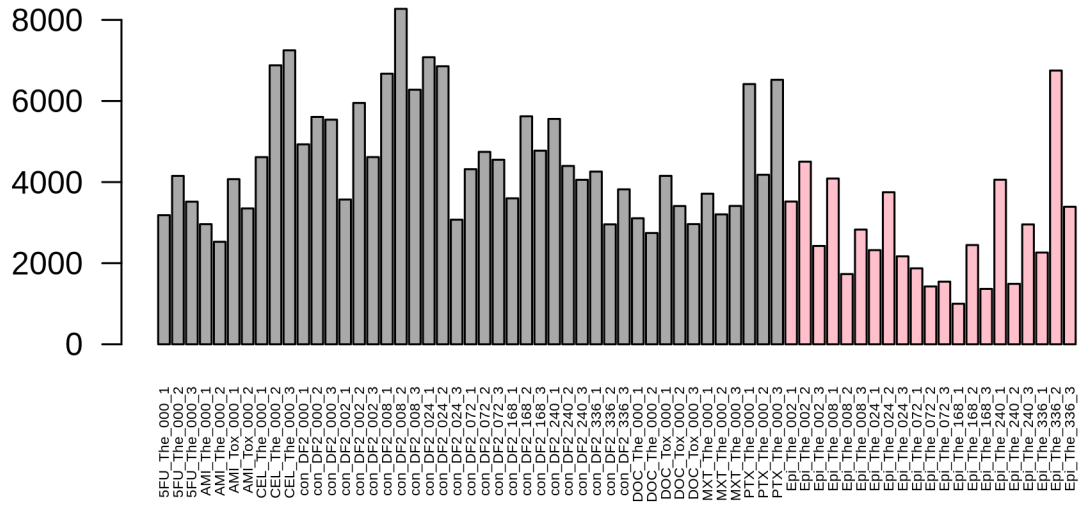
42



43


ENSG00000137693 (Dox_Tox)
8000
6000
4000
2000
0
5FU_The_000_1
5FU_The_000_2
5FU_The_000_3
AMI_The_000_1
AMI_The_000_2
AMI_Tox_000_1
AMI_Tox_000_2
CEL_The_000_1
CEL_The_000_2
CEL_The_000_3
con_DF2_000_1
con_DF2_000_2
con_DF2_000_3
con_DF2_002_1
con_DF2_002_2
con_DF2_002_3
con_DF2_008_1
con_DF2_008_2
con_DF2_008_3
con_DF2_024_1
con_DF2_024_2
con_DF2_024_3
con_DF2_072_1
con_DF2_072_2
con_DF2_072_3
con_DF2_168_1
con_DF2_168_2
con_DF2_168_3
con_DF2_240_1
con_DF2_240_2
con_DF2_240_3
con_DF2_336_1
con_DF2_336_2
con_DF2_336_3
DOC_The_000_1
DOC_The_000_2
DOC_Tox_000_1
DOC_Tox_000_2
DOC_Tox_000_3
MXT_The_000_1
MXT_The_000_2
MXT_The_000_3
PTX_The_000_1
PTX_The_000_2
PTX_The_000_3
Dox_Tox_002_1
Dox_Tox_002_2
Dox_Tox_002_3
Dox_Tox_008_1
Dox_Tox_008_2
Dox_Tox_008_3
Dox_Tox_024_1
Dox_Tox_024_2
Dox_Tox_024_3
Dox_Tox_072_1
Dox_Tox_072_2
Dox_Tox_072_3
Dox_Tox_168_1
Dox_Tox_168_2
Dox_Tox_168_3
Dox_Tox_240_1
Dox_Tox_240_2
Dox_Tox_240_3
Dox_Tox_336_1
Dox_Tox_336_2

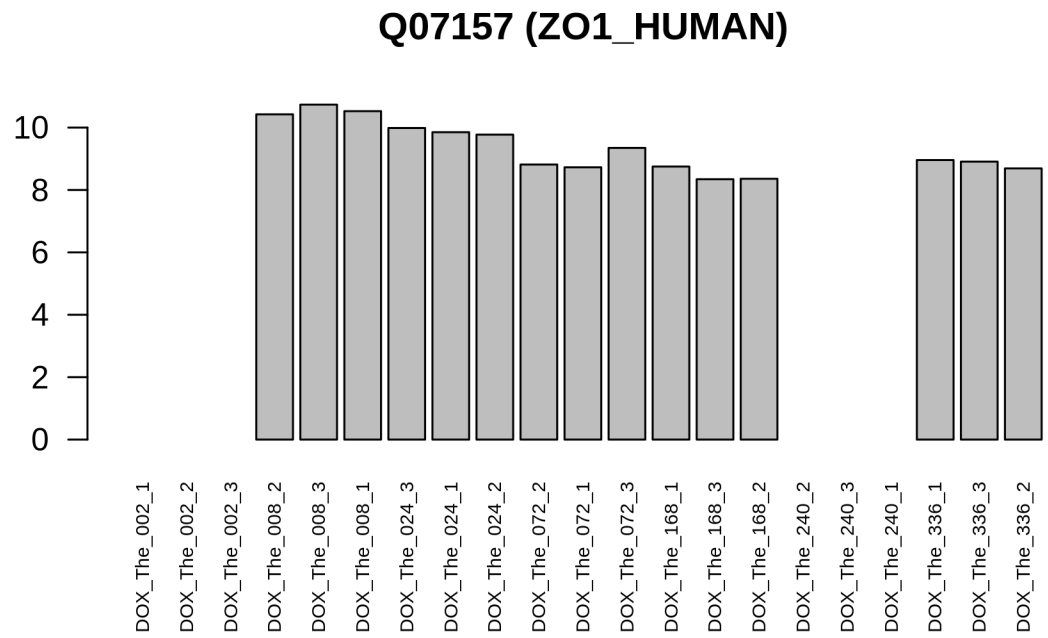
44

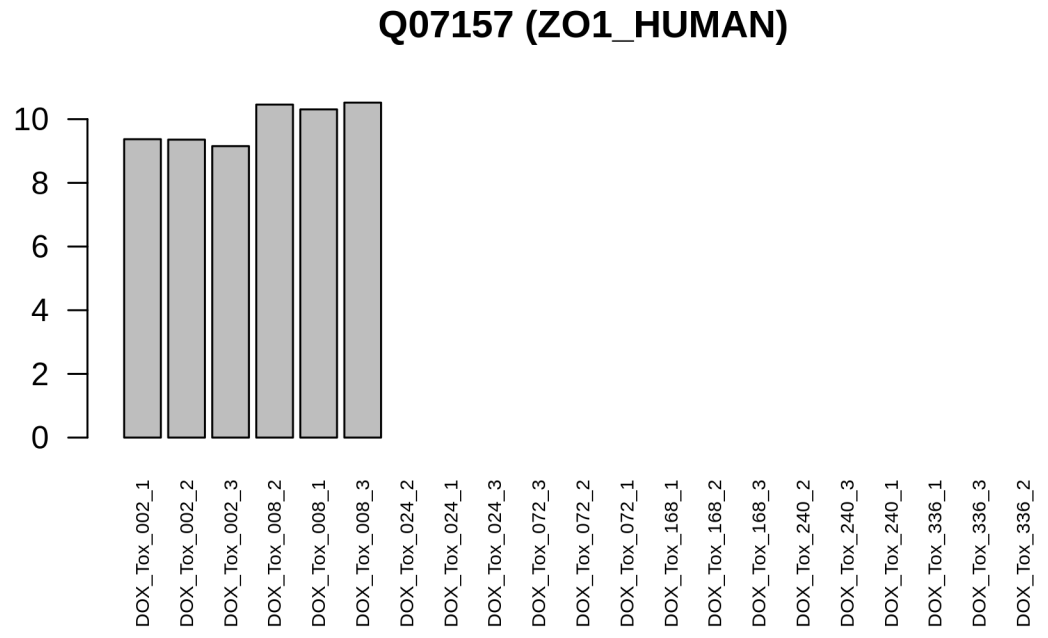

ENSG00000137693 (Epi_The)
8000
6000
4000
2000
0
5FU_The_000_1
5FU_The_000_2
5FU_The_000_3
AMI_The_000_1
AMI_The_000_2
AMI_Tox_000_1
AMI_Tox_000_2
CEL_The_000_1
CEL_The_000_2
CEL_The_000_3
con_DF2_000_1
con_DF2_000_2
con_DF2_000_3
con_DF2_002_1
con_DF2_002_2
con_DF2_002_3
con_DF2_008_1
con_DF2_008_2
con_DF2_008_3
con_DF2_024_1
con_DF2_024_2
con_DF2_024_3
con_DF2_072_1
con_DF2_072_2
con_DF2_072_3
con_DF2_168_1
con_DF2_168_2
con_DF2_168_3
con_DF2_240_1
con_DF2_240_2
con_DF2_240_3
con_DF2_336_1
con_DF2_336_2
con_DF2_336_3
DOC_The_000_1
DOC_The_000_2
DOC_Tox_000_1
DOC_Tox_000_2
DOC_Tox_000_3
MXT_The_000_1
MXT_The_000_2
MXT_The_000_3
PTX_The_000_1
PTX_The_000_2
PTX_The_000_3
Epi_The_002_1
Epi_The_002_2
Epi_The_002_3
Epi_The_008_1
Epi_The_008_2
Epi_The_008_3
Epi_The_024_1
Epi_The_024_2
Epi_The_024_3
Epi_The_072_1
Epi_The_072_2
Epi_The_072_3
Epi_The_168_1
Epi_The_168_2
Epi_The_168_3
Epi_The_240_1
Epi_The_240_2
Epi_The_240_3
Epi_The_336_1
Epi_The_336_2
Epi_The_336_3

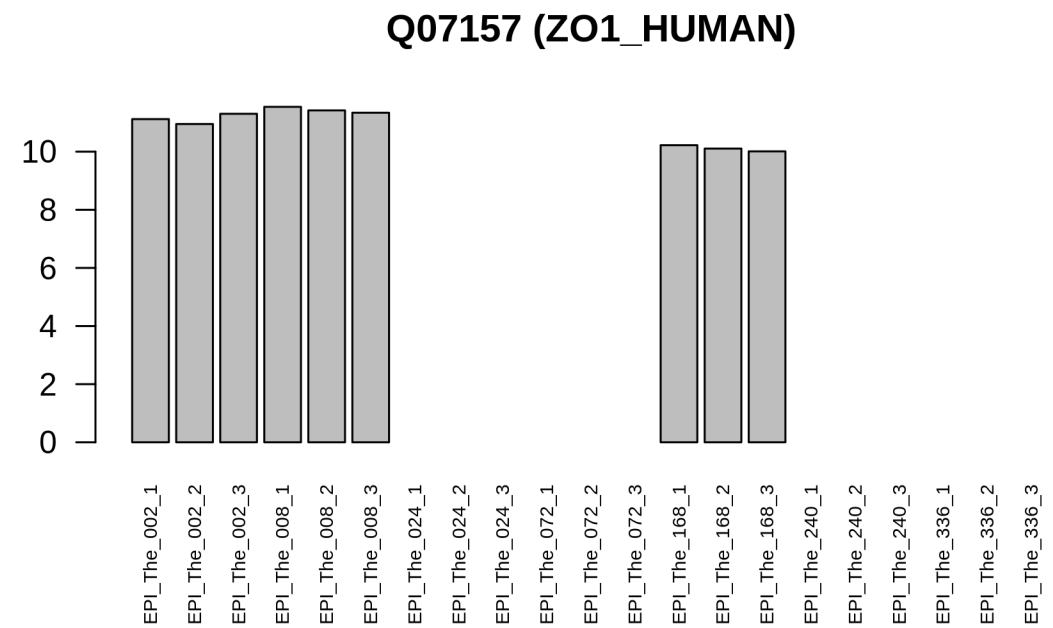
45

TJP1 (Q07157)

Normalized Log2 Values



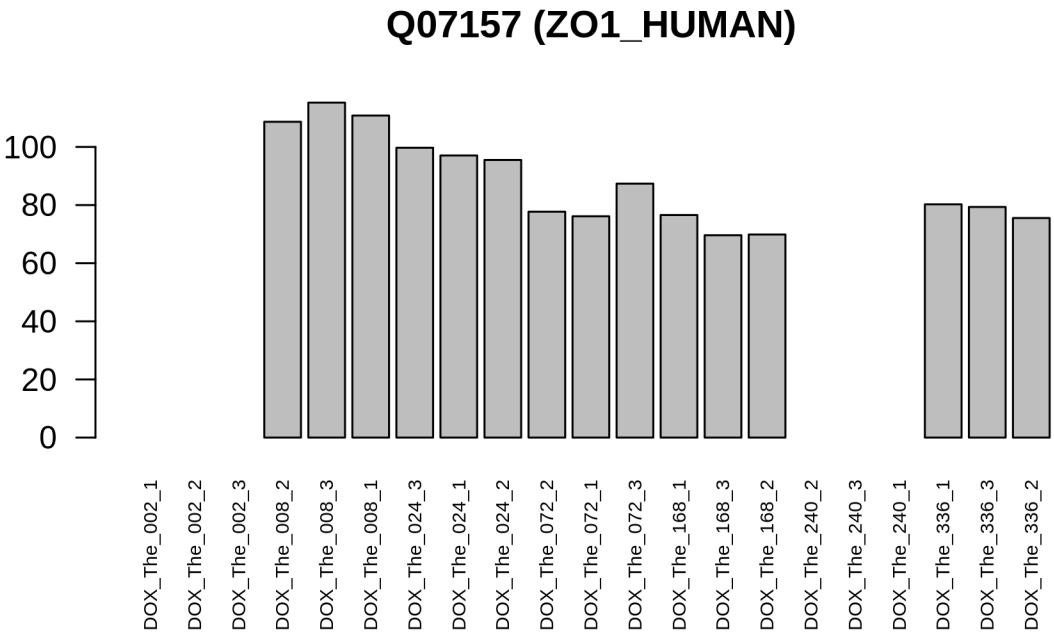




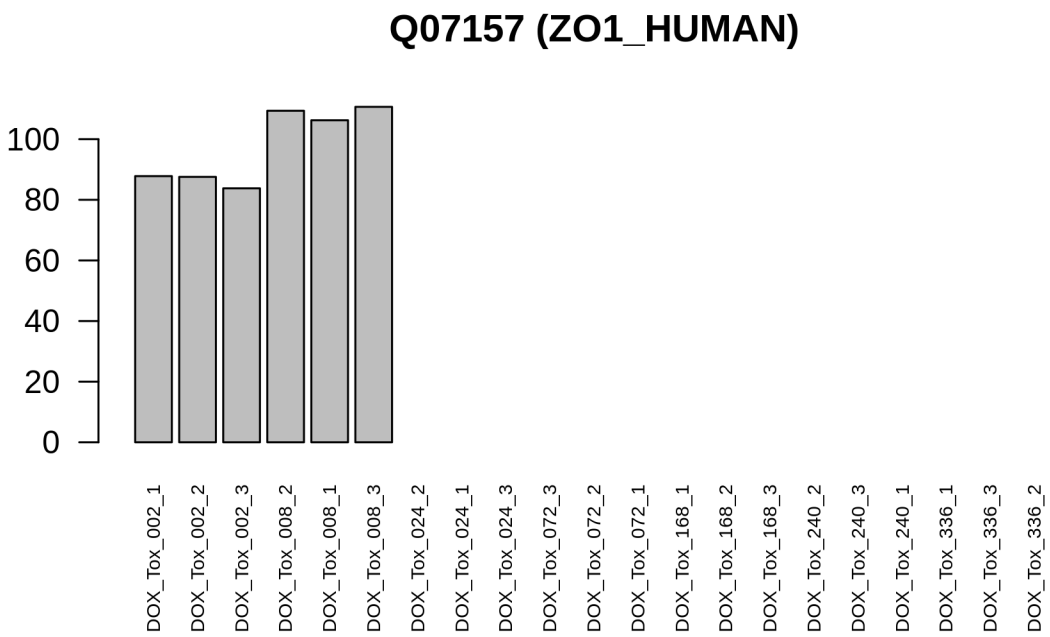
50

51 Normalized Values

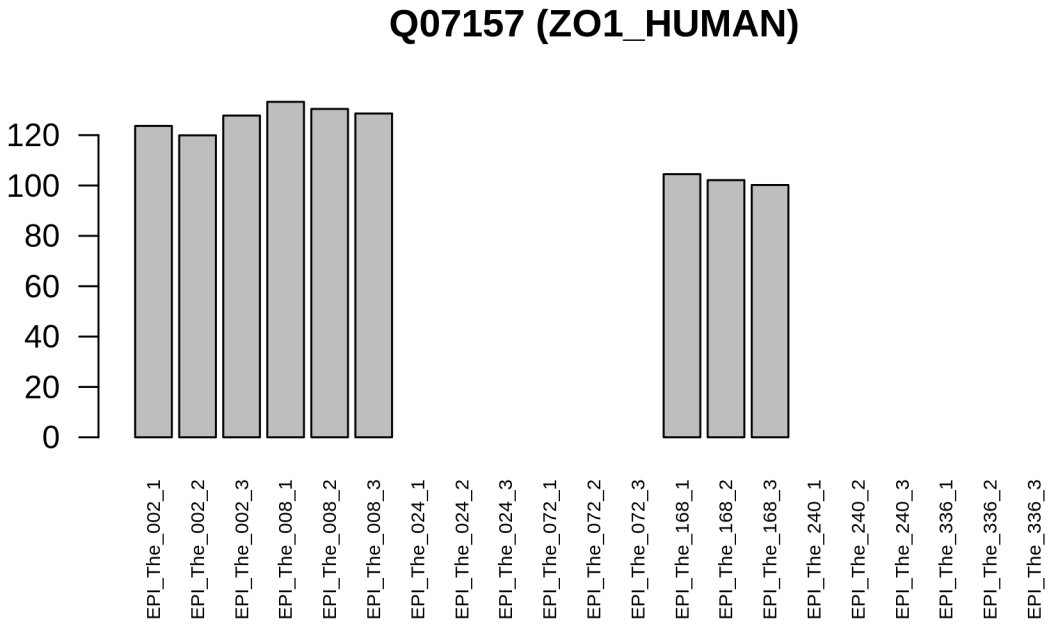
52



53

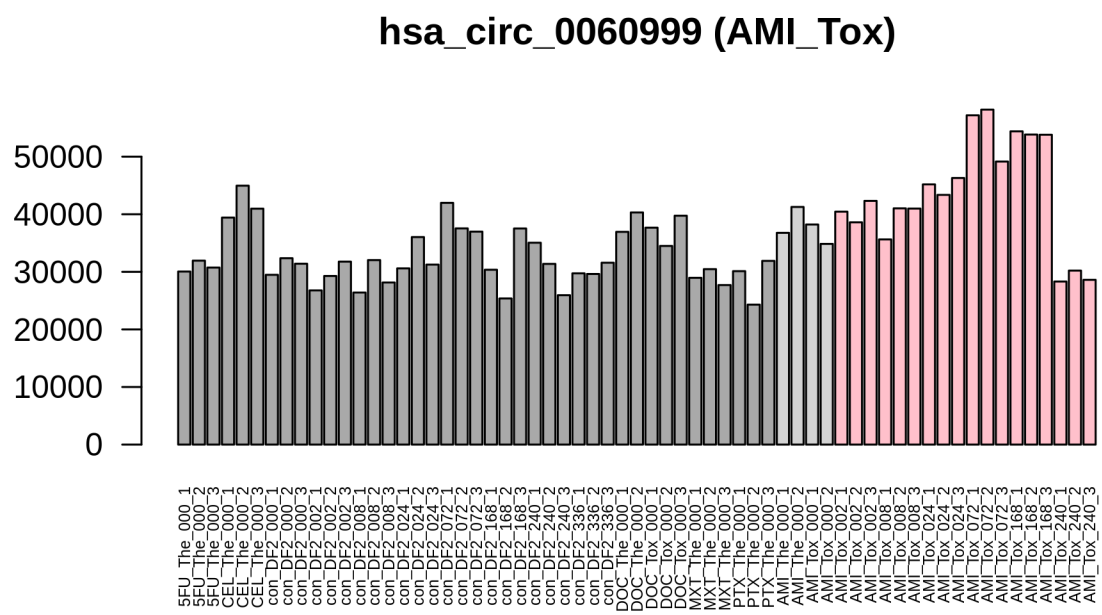

Q07157 (ZO1_HUMAN)
100
80
60
40
20
0
DOX_Tox_002_1
DOX_Tox_002_2
DOX_Tox_002_3
DOX_Tox_008_2
DOX_Tox_008_1
DOX_Tox_008_3
DOX_Tox_024_2
DOX_Tox_024_1
DOX_Tox_024_3
DOX_Tox_072_3
DOX_Tox_072_2
DOX_Tox_072_1
DOX_Tox_168_1
DOX_Tox_168_2
DOX_Tox_168_3
DOX_Tox_240_2
DOX_Tox_240_3
DOX_Tox_240_1
DOX_Tox_336_1
DOX_Tox_336_3
DOX_Tox_336_2

54

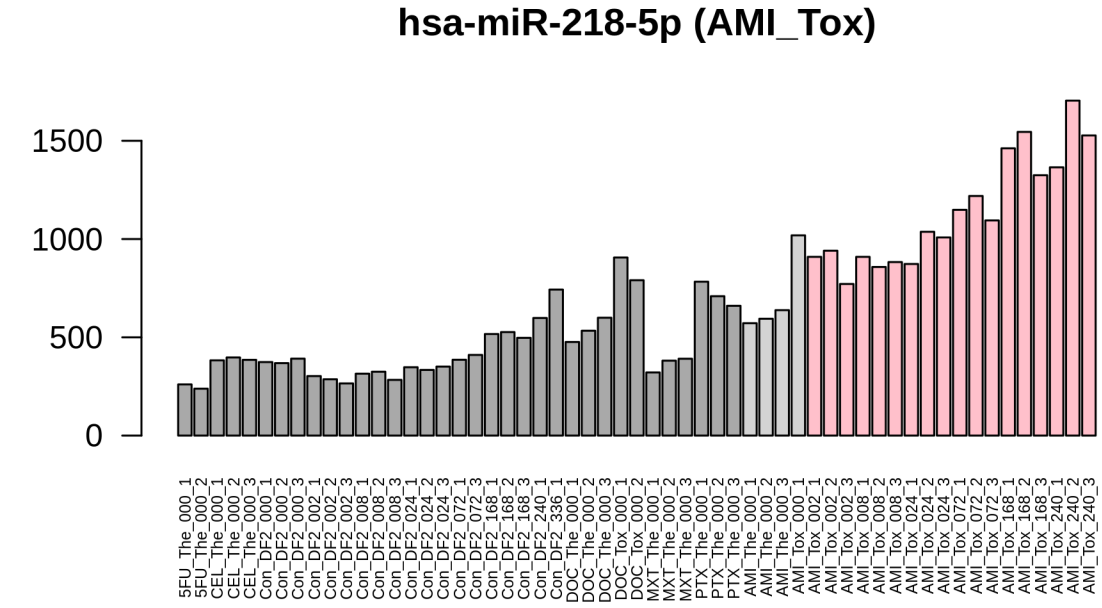

Q07157 (ZO1_HUMAN)
120
100
80
60
40
20
0
EPI_The_002_1
EPI_The_002_2
EPI_The_002_3
EPI_The_008_1
EPI_The_008_2
EPI_The_008_3
EPI_The_024_1
EPI_The_024_2
EPI_The_024_3
EPI_The_072_1
EPI_The_072_2
EPI_The_072_3
EPI_The_168_1
EPI_The_168_2
EPI_The_168_3
EPI_The_240_1
EPI_The_240_2
EPI_The_240_3
EPI_The_336_1
EPI_The_336_2
EPI_The_336_3

55

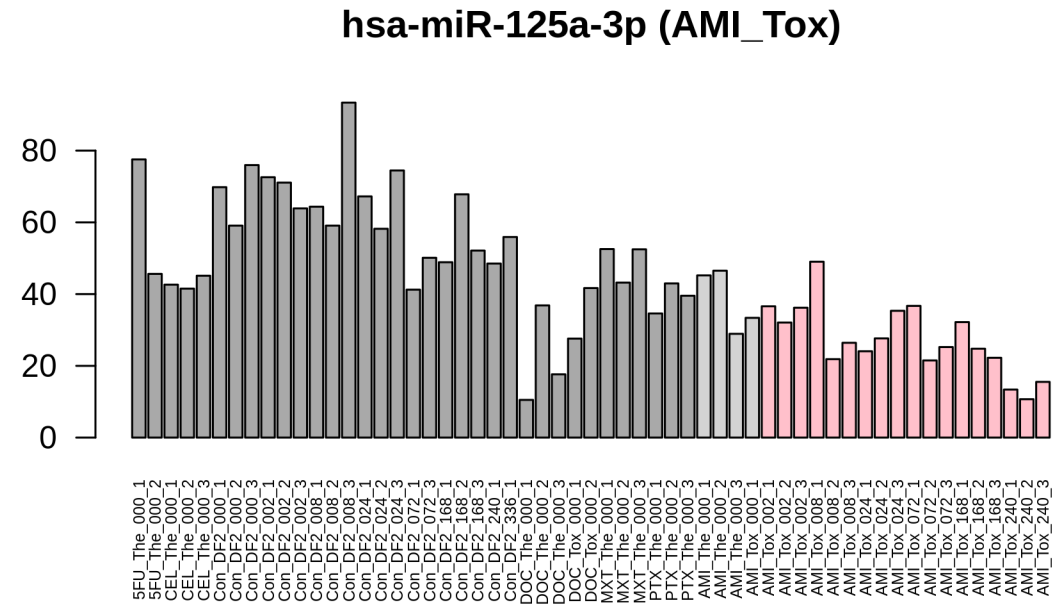
CircGNAS



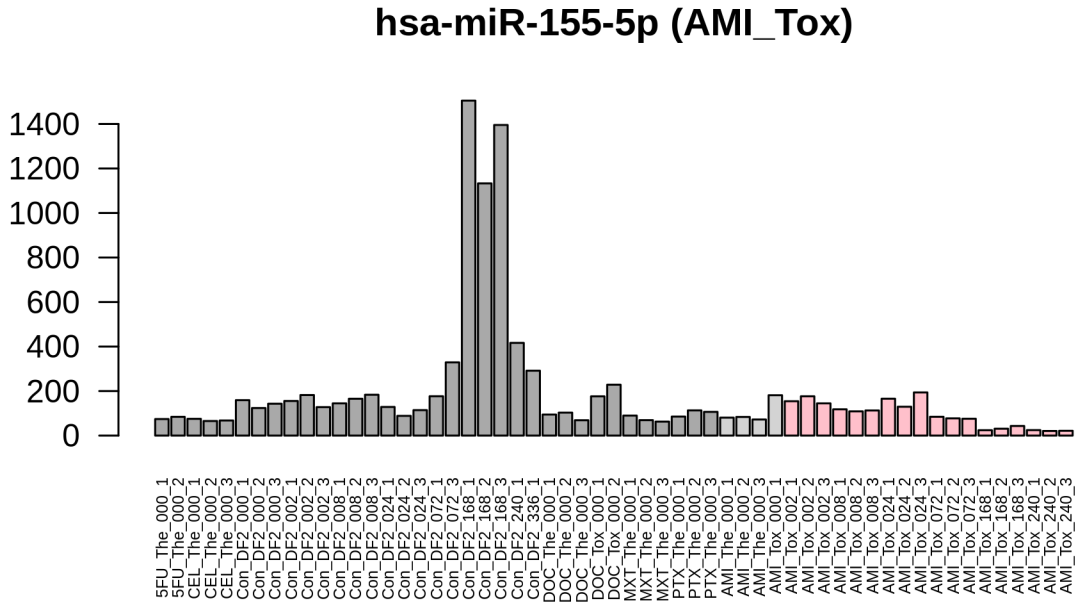
hsa-miR-218-5p



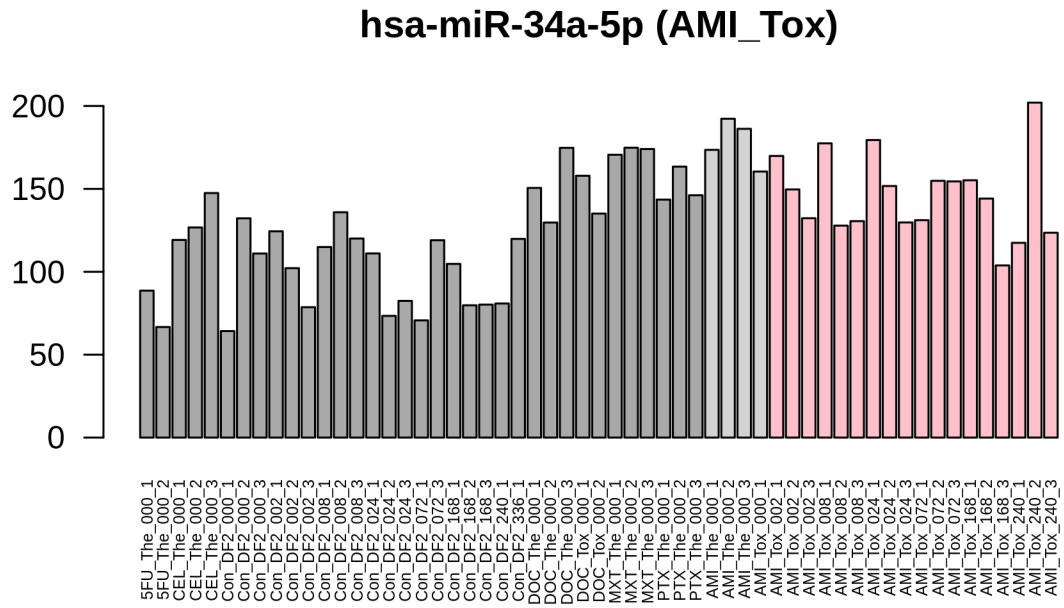
hsa-miR-125a-3p



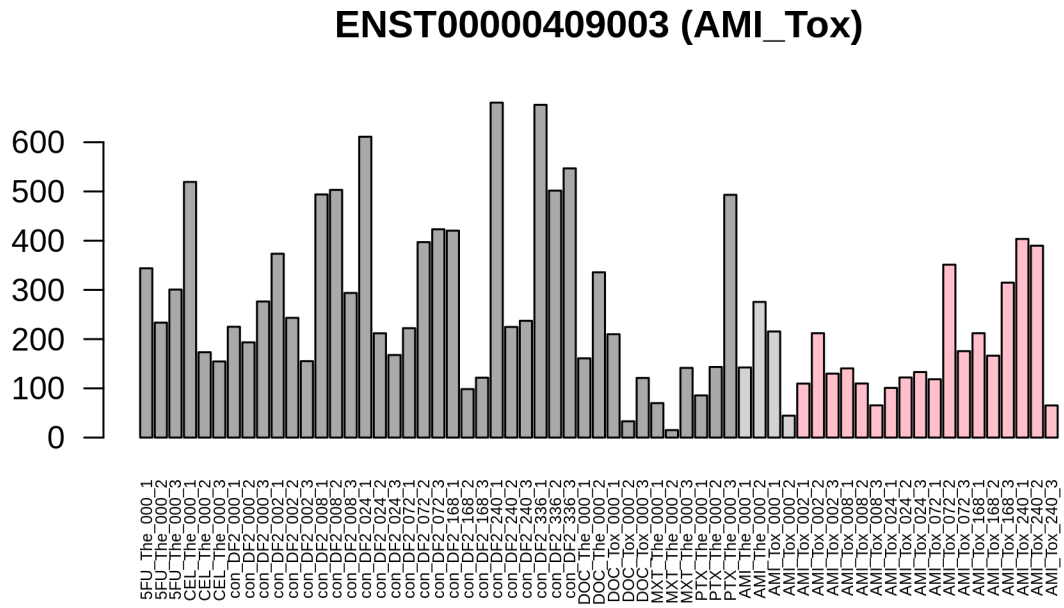
hsa-miR-155-5p



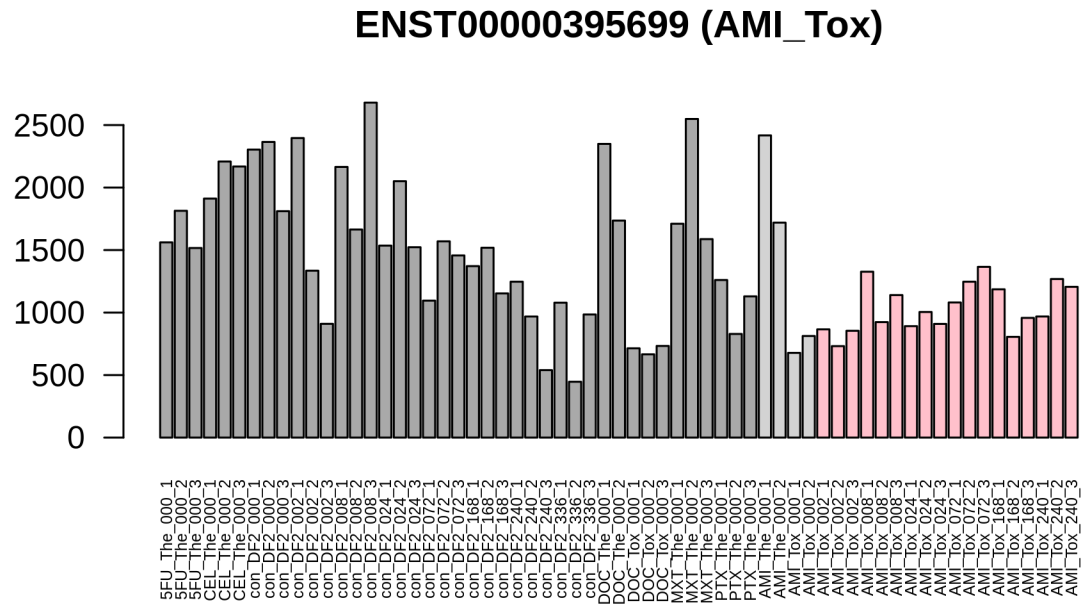
hsa-miR-34a-5p



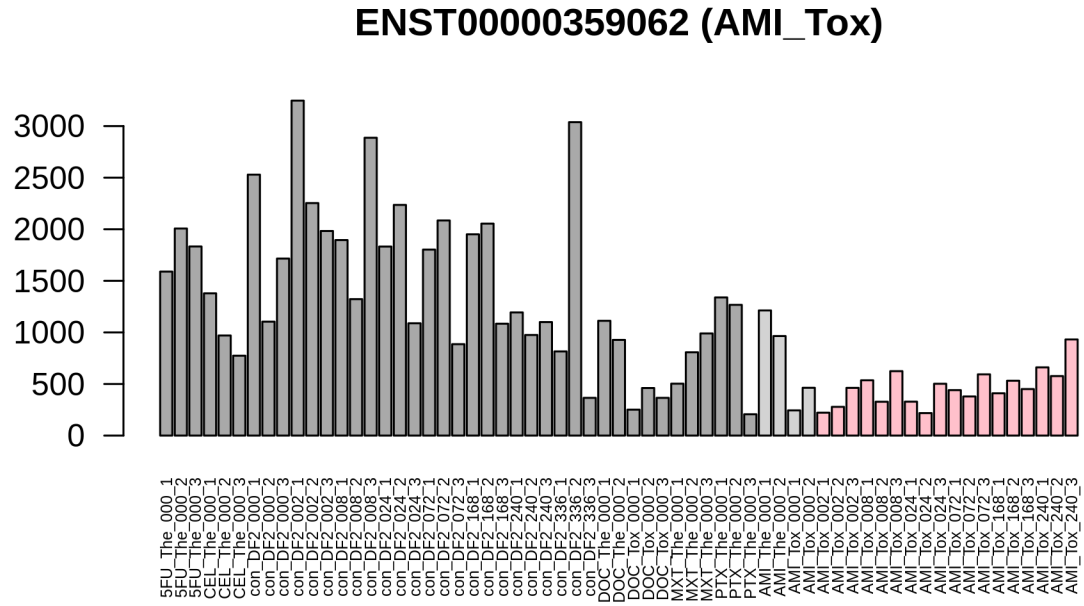
SAMD12-202 (ENST00000409003)



PURB-201 (ENST00000395699)



PDE3A-201 (ENST00000359062)



Chapter 3:

Quantifying the number of translatable transcripts through the use of OMICs involved in post-transcriptional regulation

Quantifying the number of translatable transcripts through the use of OMICs involved in post-transcriptional regulation

Juan Ochoteco Asensio[1], Jos Kleinjans[1], Florian Caiment[1]

[1]Department of Toxicogenomics, School of Oncology and Developmental Biology (GROW), Maastricht University, Maastricht, The Netherlands

(Manuscript preprint available)

Abstract

Transcriptomics is nowadays frequently used as an analytical tool to study the extent of cell expression changes between two phenotypes or between different conditions. However, an important portion of the significant changes observed in transcriptomics at the gene level is usually not consistently detected at the protein level by proteomics. This poor correlation between the measured transcriptome and proteome is probably mainly due to post-transcriptional regulation, among which miRNA and circRNA have been proposed to play an important role. Therefore, since both miRNA and circRNA are also quantified by transcriptomics, we proposed to build a model taking those factors into account to estimate, for each transcript, the fraction of transcripts that would be available for translation. Using a dataset of cells exposed to diverse compounds, we evaluated how our model was able to improve the correlation between the assessed transcriptome and proteome expression levels. The results show that the model improved the correlation for a subset of genes, probably due to the regulation of different miRNAs across the genome.

Introduction

The central dogma of molecular biology states a straightforward flow of information: for a gene to transmit the information it contains, its DNA needs to be transcribed to RNA to produce the desired protein. Each of these biological pools of molecules has its field of study for both their characterization and quantification, which are commonly known as omics. Transcriptomics, for example, refers to the large-scale study of transcripts, and the same applies to proteomics (proteins) or metabolomics (metabolites).

Proteins, being the functional molecules of the cell, are of high interest to be analyzed, as they accurately represent the phenotypic changes of the studied cell or tissue. Unfortunately, the technology' sensitivity and reproducibility behind proteomics, namely mass spectrometry, is currently still limited, especially when the aim is to get an exhaustive protein expression analysis. This is mainly due to the mass spectrometer only being able to measure a fraction of the eluted peptides, hence giving as an output a slight portion of the entire population of proteins[1–3]. To circumvent this issue, an alternative strategy often used is the expression analysis of their precursors: the transcripts. Transcriptomics is indeed usually performed as the surrogate to analyze different disease states or cell conditions. Even so, the fact that transcripts are indeed the cause and origin of proteins is not reflected by a high correlation between both omics[4–8]. This is rather unsurprising, due to the multiplicity and complexity of the factors playing a role in post-transcriptional regulation, in addition to intrinsic variable characteristics of both molecules (such as half-life[9]), and some of those regulatory factors are even transcripts themselves.

One such factor is microRNAs (miRNAs), whose effect on the expression levels of proteins, although mild, has been characterized for more than a decade[10,11]. MiRNAs are short non-coding RNA sequences playing a role in translation regulation[12]. They contain the so-called "seed" region[13], a short sequence with perfect Watson-Crick complementarity with their target (primarily to their 3' UTR region). The binding between the two molecules impairs the translation of the target[14]. Due to the short length of the seed region, a single miRNA often targets several transcripts, and numerous miRNAs can target a single transcript. This inhibitory relationship may differ depending on the level of expression of the miRNA and its target molecules, the number of regions in a single mRNA complementary to the miRNA present, and the combinatorial effect of several miRNAs targeting the same transcript. Taking into account these parameters is necessary to have a better comprehension of the mechanisms behind post-transcriptional regulation.

Another element to be considered in post-transcription regulation is circular RNAs (circRNAs). CircRNAs are transcripts that possess a circular structure due to the covalent binding of their 5' and 3' ends[15,16], hence obtaining a circular structure that bestows them resistance to exonuclease activity[17]. One of their recently discovered functions is described as 'miRNA sponges'[18]. This refers to their ability to contain several copies of short sequences that are complementary to the miRNAs' seed region. Consequently, they also play a role in the regulation of translation machinery by competing against other transcripts for miRNA binding[19]. The number of miRNAs captured by circRNAs influence the number of transcripts available for translation.

CircRNAs are not the only targets competing for miRNA, as other (long) non-coding RNAs (ncRNAs) can also present a seed target for some particular miRNAs[20]. This collection of coding and

non-coding transcripts can be conceptualized as a single group of targets for a shared miRNA, which led them to be generally called competing endogenous RNAs (ceRNAs)[21].

Thus, in the current age of sequencing, accurately associating transcriptomics to the actual phenotype of the cell is a major challenge. Currently, many publications base their biological interpretation on the differentially expressed genes in two groups of samples as transcriptomics is more sensitive and allows the assessment of almost all possible RNA molecules in a single experiment. In this context, we wondered if it would be possible to integrate all available transcriptomics information into a model able to estimate the level of translation of any expressed coding transcript, namely, the fraction of translatable transcripts (TrT). This value, computed for each protein-coding transcript, is an estimation of the number of transcripts that are free to be translated after taking into account the aforementioned post-transcriptional regulation.

To design and assess the model, we used an *in vitro* dataset obtained from human cardiac microtissues exposed to a range of compounds at different doses, which were analyzed with proteomics, RNA-Seq (ribo-depleted libraries), and miRNA-Seq methods. From the RNA sequencing data, we first identified and quantified different transcript biotypes: protein-coding, non-coding, miRNA, and circRNAs. The proposed model was then applied to generate the TrT score from all possible interactions between those molecules. In this manuscript, we analyze the possible benefits of such an approach compared to the state-of-the-art methods for gene expression analysis.

Methods

Samples

The analyzed data consists of 3D microtissues containing stem-cell-derived cardiomyocytes and fibroblasts in a 4:1 ratio obtained from InSphero. These microtissues were exposed to 8 compounds: Fluorouracil (5FU), Amiodarone (AMI), Celecoxib (CEL), Docetaxel (DOC), Doxorubicin (DOX), Epirubicin (EPI), Mitoxantrone (MXT), and Paclitaxel (PTX), in addition to a control group (Untreated/UNTR). The dosing profile was established via the use of the physiologically based pharmacokinetic (PBPK) modeling software PK-Sim, to simulate exposure levels under physiological conditions[22]. For each compound, 2 doses were applied: Therapeutic and Toxic. The exposures were done in triplicates, and the data extraction was performed at 7 time-points: 2, 8, 24, 72, 168, 240, and 336 hours; resulting in 21 data-points per dose. Four samples were excluded due to low sequencing depth: three samples for having too low miRNA sequencing depth

(CEL_Tox_002_3/MXT_Tox_002_3/PTX_Tox_002_3; toxic dose, time-point 2h, 3rd triplicate) and one for its RNA-Seq low read count (UNTR_002_3).

RNA Sequencing (RNA-Seq)

Total RNA from the exposed microtissues was isolated using the Qiagen AllPrep Universal Kit (Cat #80224). Ribo-depletion was achieved by using the Illumina RiboZero Gold kit (Cat #MRZG12324), and the libraries were prepared using the Lexogen SENSE total RNA kit (Cat #009.96). All libraries were then sequenced on an Illumina HiSeq 2000 at 100 bp paired-end at an average coverage range between 20 and 30 million reads. The adaptors were removed through Trimmomatic version 0.33[23]. We used the following parameters: paired-end, ILLUMINACLIP: TruSeq3-PE.fa:2:30:10, LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, MINLEN:36, HEADCROP:12.

Proteomics

Proteins were isolated and diluted to a concentration below 0.2M. The peptides were digested by trypsin, and peptides were cleaned-up using Sep-Pak tC18 cartridges (Waters) according to the manufacturer's instructions. A vacuum centrifuge was used to dry the peptides, before measuring them on an Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific), which was coupled to a NanoLC-2D HPLC system (Eksigent). To process the raw MS data, Genedata Expressionist software (v.11.0) was used. The noise in the LC-MS peaks was reduced and normalized, and afterward, their properties were obtained (m/z and RT boundaries, m/z and RT center values, intensity). To annotate the individual MS/MS spectra, Mascot 2.6 was used. To group peak clusters, protein interference was used (based on peptide and protein annotations), and, using the Hi3 method, protein intensities were computed.

miRNA analysis

Starting from the same total RNA isolated for the ribo-depleted libraries, an aliquot was size selected and ligated using the TruSeq Small RNA Library Prep Kit (Illumina®). After sequencing on the HiSeq 2000 at 3.6 million reads per sample (after quantification), we quantified the resulting data using miRge2 (last change: 05/06/2018)[24]. miRge2 used the MirBase database as the reference library (miRBase v22), bowtie-1.1.1[25] as the mapper, miRge2/sp as the miRge library, human as the species selected, and illumina for the adapter to be removed. The output results were in gff format. Isomirs were not considered for this analysis.

(non-)coding RNA analysis

The genome version used for all the transcriptomics analyses was the Genome Reference Consortium Human Build 38 (GRCh38.p12). For the identification of circRNAs, we performed a *de novo* prediction from our dataset. For this, we concatenated all RNA-Seq forward data (R1) across all 240 Cardiac samples into a single FASTQ file. Afterward, this file was inputted to two recent circRNA prediction software: circExplorer2[26] and CIRI2[27]. For circExplorer2, we first used BWA version 0.7.17[28] to index the genome and align (minimum score to output: 19) the reads to the indexed genome. Afterward, we used circExplorer2 to parse the aligned reads and annotate the circRNAs. For CIRI2, we also required the aligning step via BWA, and then the annotation step through its internal algorithm (-S (single-end), -U 3 -B 13, to set mapping quality thresholds of a junction read and help control False Discovery Rate (FDR)). We then extracted the overlap of identified molecules between both outputs for decreasing the amount of false-positive predictions. The circRNA IDs were comprised of the chromosomal and strand locus of the predicted molecules. Based on this information, we extracted their genetic sequence using BEDTools[29], obtaining a final circRNA transcriptome library.

We downloaded the transcriptomes for both coding (all cDNA) and non-coding (all ncRNA) transcripts from Ensembl (release 96)[30]. We combined these 2 libraries with our predicted circRNAs library into a single library, which we set as the global transcriptome reference for Salmon[31], with which we quantified the RNA-Seq data. Salmon output contained both the number of reads and TPM (Transcript per Million) values for each transcript, the latter of which was used for the analysis. The benefit of using TPM values, instead of raw counts, relies on the inherent normalization performed for both sequencing depth and transcript length. If raw reads would have been used instead, the variability across samples and compounds would have made it impossible to optimize the model consistently. Moreover, since our model computes interactions between molecules of different sizes, using read counts would have greatly biased the output.

CircRNA predictors selection

We selected CIRI2 and circExplorer2 out of 4 possible predictors, which also included find_circ[18] and circRNA_finder[32]. The parameters for find_circ were set as default, and the samtools version was 1.3. For circRNA_finder, it involved 2 steps: running STAR (-c --runThreadN 4 --chimSegmentMin 20 --chimScoreMin 1 --alignIntronMax 500000 --outFilterMismatchNmax 4 --alignTranscriptsPerReadNmax 100000 --twopassMode Basic --outSAMtype BAM SortedByCoordinate --chimOutType SeparateSAMold --outFilterMultimapNmax 2) and the post-processing script (--minLen 100). We formatted all the output equally so that we were able to compare them across tools.

Interaction tables

To evaluate all possible bindings between the molecules, a table containing all possible interactions between miRNAs and any putative competing ceRNAs was generated, each row representing a unique interaction. For miRNA, both identifiers (IDs) and sequences were sourced from miRBase[33]. For all transcripts (both coding and non-coding), we extracted the Ensembl transcript IDs through biomaRt[34]. We obtained the sequences corresponding to these IDs through SAMtools[35], which were searched in the genome FASTA file by using the chromosomal coordinates of the transcripts. For mRNAs, we used only the 3' UTR region. To establish the microRNA target interactions (or MTIs), we used miRanda (version 3.3a)[36] with the -strict parameter to force strict 5' seed pairing. We set the score threshold to 140.0 (corresponding to a full 7 base pair seed with no mismatch). We also used –noenergy, disabling the thermodynamics performance.

Translatable Transcripts (TrT) formulation

We designed a formula (Formula 1) that aimed to estimate the fraction of translatable transcripts based on a basic principle: the number of translatable transcripts is equal to the difference between the total expression of a given protein-coding transcript and its inhibited molecules. The number of inhibited transcripts depends on the number of miRNAs and the probability that these will bind to the target transcript. The probability of a target being targeted by miRNA depends on how abundant the target is in proportion to the number of all the possible targets. For all data analyses, we used R[37,38] as the main programming language.

$$TrT_{transcript} = TPM_{transcript} - miRNA_{factor} * \sum_{i=miRNA}^{N} RPM_i * \frac{TPM_{transcript} * Seeds_{i,transcript}}{\sum_{j=target}^{M} TPM_j * Seeds_{i,j}}$$

Formula 1: TrT formula

TrT = Translatable Transcripts, TPM = Transcripts Per Million, RPM = Reads Per Million, Seeds = the number of seeds present in a specific transcript targeted by a specific miRNA, N = number of miRNAs targeting the transcript, M = number of targets for a specific $miRNA_i$, $miRNA_{factor}$ is a scaling factor, set to 0.1 (TrT model design).

TrT example

For example, consider a protein-coding transcript expressed at 100 TPM targeted by two miRNAs, expressed at 132.5 and 227.5 Reads Per Million (RPM) respectively. Each of them has a single seed or sequence region presented by the target that they can bind to. The first miRNA can also interact

with a single circRNA, expressed at 2 TPM on 3 perfect seed regions, while the second miRNA can bind to an ncRNA with a TPM value of 30 (Equation b).

$$TrT_{transcript} = TPM_{transcript} - miRNA_{factor} * \sum_{i=miRNA}^{N} RPM_i * \frac{TPM_{transcript} * Seeds_{i,transcript}}{\sum_{j=target}^{M} TPM_j * Seeds_{i,j}}$$

Equation a: Entire TrT formula

$$TrT_{transcript} = TPM_{transcript} - miRNA_{factor} * \left(RPM_1 * \frac{100 * 1}{100 * 1 + 2 * 3} + RPM_2 * \frac{100 * 1}{100 * 1 + 30 * 1} \right)$$

Equation b: Calculating the fraction of the transcript of choice as a target for the miRNA

$$TrT_{transcript} = TPM_{transcript} - miRNA_{factor} * (132.5 * 0.94 + 227.5 * 0.77)$$

Equation c: Calculating how many miRNAs will target the transcript of choice

$$TrT_{transcript} = TPM_{transcript} - 0.1 * 300$$

Equation d: Downsizing the miRNA expression to be comparable to transcriptomics

$$TrT_{transcript} = 100 - 30 = 70$$

Equation e: Calculating how many transcripts will be able to be translated

Box 1: TrT example

Those miRNAs can bind to any of the aforementioned ceRNAs (other mRNA targets, circRNAs, or ncRNAs). To know how many miRNAs will bind to each of them, we need to know the probability of each interaction. In our study, we hypothesized that the probability that a miRNA will bind to a molecule is directly proportional to how abundant that molecule is in relation to all possible targets (0.94 and 0.77 in this example, Equation c). Consequently, we estimate that 300 of these miRNA molecules will bind to the target. This number is multiplied by the miRNA factor (Equation d), and we subtract it from the original TPM value of our target, which will be the expected TrT value (Equation e).

Correlation between Transcriptomics and Proteomics

To perform correlation analysis between our transcriptomics and proteomics dataset, we used the untreated samples of the first 4 time-points (2, 8, 24, and 72 hours) for transcriptomics, and all the normalized (Proteomics) time-points for proteomics (2, 8, 24, 72, 168, 240, and 336 hours).

To be able to correlate protein and transcript expressions, we merged protein-coding transcripts by their protein product, and we grouped all values per time-point while maintaining the same order to perform a Pearson correlation at the time-point level. For 168h, we omitted the first triplicate due to a significant batch effect, correlating only the other 2 triplicates with the other 2 triplicates of all other

time-points. We did the same with the RNA-Seq sample of UNTR_002_3, because of the low sequencing depth, as mentioned before. The graphical displays were generated thanks to the 'corrplot' package[39].

Differential Expression Analysis

Proteomics

To obtain differentially expressed proteins (DEPs), we followed the steps performed by Selevsek et al (2020). Briefly, we first log-2 transformed all the proteomic expression values. We then calculated the median for each control sample. Afterward, we calculated the median of all the medians (MoM) and shifted the control samples so that they all shared this median. For every treatment/dose combination, we determined the set of proteins common in both control and treatment samples. This set of proteins was used to determine the MoM on the normalized control samples. We then shifted the distribution of the data by matching their median values to this MoM. Finally, we performed paired t-tests to evaluate the significance between doses, considering as significant (or DEPs) the ones for which the p-values were lower than 0.05.

Transcriptomics

As we aim at evaluating the TrT fraction of the coding transcripts, Salmon mapping output was filtered out for non-coding transcripts. All remaining coding transcript expression values were summed based on their gene of origin. While usually, differential expression analyses are performed via a dedicated sequencing analysis pipeline on raw read count using a pipeline specific for negative binomial distribution (such as DESeq2 or EdgeR), this approach could not be applied here. Indeed, in our case, the values to be compared (TPM and TrT) did not allow such pipelines, either because the input was expected to be un-normalized[40] and/or because other normalization steps were performed instead[41]. We then log-2 transformed both TPM and TrT values, which, like proteomics, also presented originally a negative binomial distribution. We performed t-tests between therapeutic and toxic doses. Due to the high percentage of significant observations and multiple testing, we also applied an FDR/BH p-adjustment. We identified a differentially expressed gene (or DEG) as a gene with a p-adjusted value lower than 0.05

Biological Interpretation with GOrilla

Using as input the lists of genes ranked by increasing p-adjusted value (for each comparison and compound), we ran a gene ontology enrichment analysis with GOrilla[42] using the 'Single ranked list of genes' mode. We studied the GO terms of interest and analyzed some of their genes' expression (both in TPM and TrT). We also explored how the MTIs (using the miRTarBase[43] list) related to such

genes (with at least weak evidence of having a regulatory effect) were expressed in comparison to what TPM and TrT were representing.

Sensitivity, Specificity, and Accuracy

A prediction is normally evaluated by contrasting it to the reality it is trying to model. Such an evaluation focuses on different aspects of the predictor: how many of the positives that have been predicted are real positives (sensitivity), how many of the predicted negatives are indeed negative (specificity); and how many cases are correct out of all cases, independently of whether they are positive or negative (accuracy), making use of binary classification terms. In our case, proteomics was considered the true condition, while transcriptomics was the predicted condition. True or false referred then to the correct or incorrect representation of proteomics by transcriptomics, while positive or negative referred to the presence or absence of differential expression, respectively. For example, if both proteomics and transcriptomics showed differential expression in the same manner, such gene was considered a true positive. If transcriptomics did not present a differential expression while proteomics did, it was considered a false negative. Once these terms were defined, the calculation of such evaluators was performed according to the standard statistical measures of a binary classification test performance.

Results

Test dataset for TrT model assessment

To develop a TrT model, we used a dataset as input derived from a 3D cardiomyocyte culture, from which both transcriptomics (ribo-depleted and small RNA libraries) and proteomics (LC-MS) data were generated. The cell cultures were composed of 8 individual compound treatments, in addition to an untreated control. For every compound, 3 replicates were measured in 7 time-points (2, 8, 24, 72, 168, 240, and 336 hours) for every dose (therapeutic and toxic), resulting in 42 samples per compound. This dataset, of a total of 240 samples, presented the added value to have been generated by the same technician, and all proteomics and transcriptomics samples were generated from the same run, thus reducing an important source of bias. The three different technologies (RNA-Seq, miRNA-seq, and proteomics) were quantified based on state-of-the-art procedures (Methods). While our RNA-Seq data analysis quantified the protein-coding and non-coding transcripts available in the Ensembl database, circular RNA identification required additional steps such as *de novo* prediction.

CircRNA predictors selection

CircRNA prediction programs are known for generating a high level of false positives[44]. To identify the circular RNAs in this transcriptomics dataset, we decided to use an overlapping approach between several tools. For this, we considered the following four tools: CIRI2[27], find_circ[18], circRNA_finder[32], and CIRCexplorer2[26]. When determining the optimal overlap, we interpreted the amount of uniquely predicted identities by a tool as an indicator of the false-positive ratio. We could observe that most of the circRNAs were only predicted by circRNA_finder, which suggested this algorithm has the highest rate of false positives. The other 3 tools showed similar false-positive ratios, so to compare them we focused on the overlap that maximized the number of identified molecules: circExplorer2 and CIRI2. The list of predicted circRNA molecules was then added to the reference transcriptome used to quantify all expressed RNA molecules from the ribo-depleted libraries.

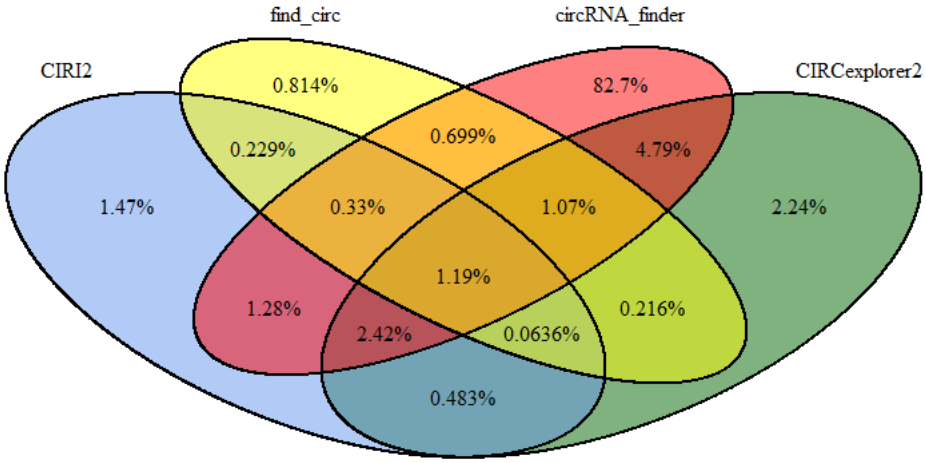


Figure 1: Venn Diagram between 4 circRNA prediction tools output

General Expression Analysis

Such transcriptome was structured in the following manner: 126831 of the identifiers were non-coding RNA (57.37%), 85108 were protein-coding RNAs (38.50%), and 9138 were circRNAs (4.13%) (Figure 2A).

In control conditions, for RNA-Seq, we detected 92969 transcripts expressed in at least 1 sample > 0 reads, with 4815 (5%) of them present in all 240 samples. In microRNA sequencing (miRNA-seq), out of the 1348 quantified molecules in any sample, only 55 miRNAs (4%) were detected constitutively. In proteomics, there were 1392 proteins quantified, of which 362 (26%) were detected in all samples. As expected, proteomics, having one order of magnitude less quantified molecules (both in Total and Constitutively) than RNA-Seq (Figure 2B), was the limiting factor when comparing both. Also, when comparing proteomics and miRNA-seq, although they have similar total

quantification numbers, the difference in their constitutively quantified shows that the former has a higher proportion of highly expressed molecules, while miRNA-seq displays a broader representation of all molecules (in addition to its more volatile/temporary function).

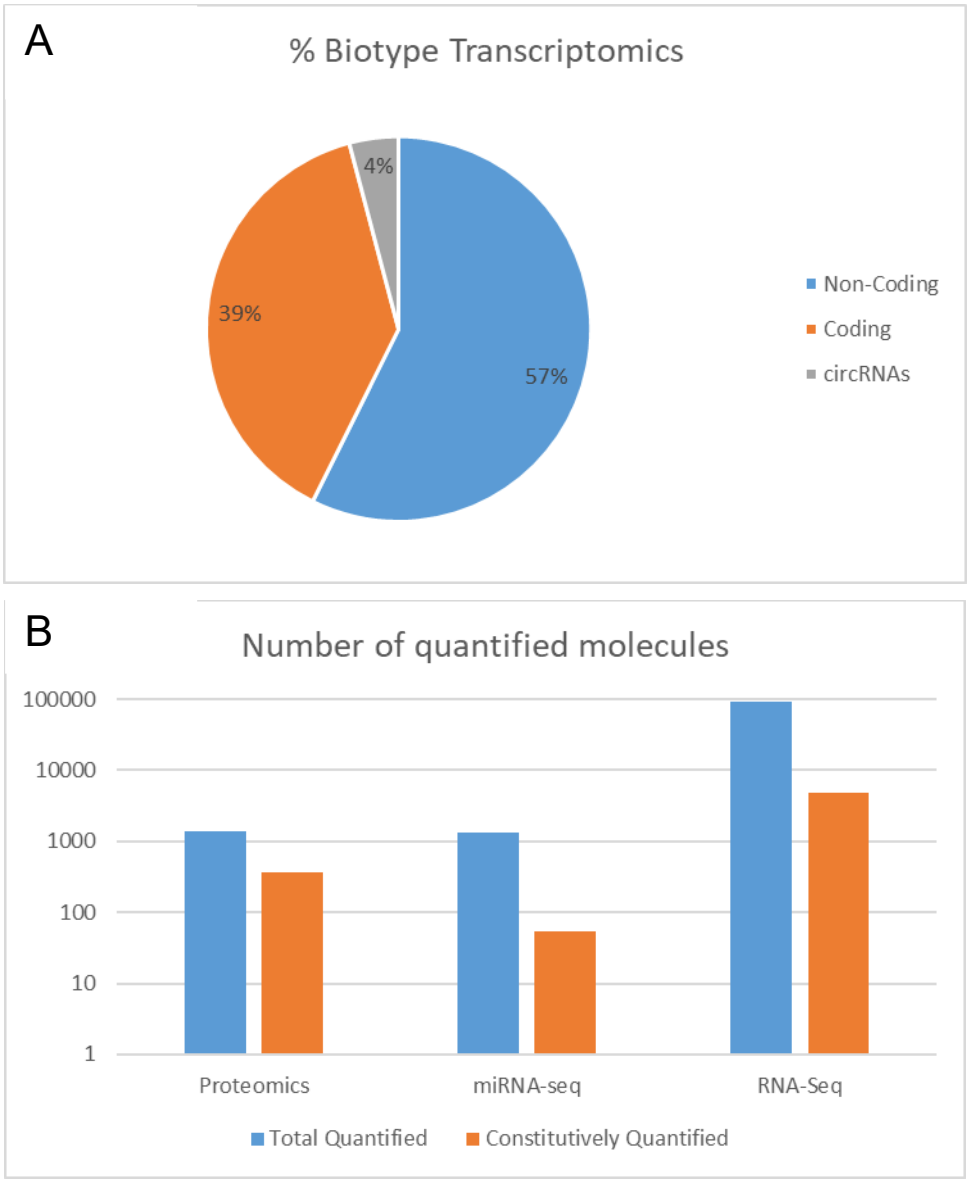


Figure 2: Quantification results across different omics. ***A:*** *Percentage of each biotype present in the Transcriptomics library.* ***B:*** *Number of quantified molecules in the Control condition. Total Quantified refers to the number of molecules that have been at least once quantified by such technology. Constitutively Quantified refers to the number of molecules that have been quantified in all samples by such technology.*

TrT model design

With all individual RNA molecules from the transcriptome characterized and quantified, we investigated the possibility to develop a model to predict the expected translatable fraction (TrT) of any given coding transcript. For this, we hypothesized that taking into account as many post-

transcriptional regulation factors as possible would be fundamental. Therefore, we conceptualized a model based on several key features of the various biological factors available. First, miRNAs decreased a transcript's chance of being translated. An increase in miRNA expression (controlling for all other conditions) would then decrease the number of mRNAs available for translation[45]. Second, the probability that a miRNA transcript interaction (MTI) occurred is directly proportional to the expression level of a target in proportion to the expression level of all possible targets. Lastly, both coding and non-coding RNAs (including circRNAs) that were able to interact with a miRNA inhibiting the coding transcript of interest could work as ceRNA: the higher the expression of a single target, the lower the inhibition for the rest of the targets[46–48]. In other words, we assumed that a given miRNA would be equally distributed among all its possible targets (taking into account the number of individual seeds), and only the fraction available to interact with our gene of interest should be considered to have an inhibiting effect.

The central concept behind this model was that the number of transcripts to be translated was the difference between the total amount of transcripts (in TPM) and the ones that would not be translated due to miRNA inhibition. The number of transcripts that would be inhibited was then based on two main factors: the level of expression of each miRNA that had the mRNA as a target (RPMi), and the probability that each miRNA would interact with the target. We hypothesized that the probability that a miRNA would bind to a target was proportional to how prevalent that target was compared to all possible targets. A target did not always equal a single molecule (a circRNA could present several seed regions/targets). For this reason, we represented a target by its total amount of seed regions for a specific MTI (transcript expression level times its number of seed regions).

Because miRNA libraries present a smaller density than ribo-depleted RNA libraries, the raw number of reads generated for the miRNAs was in general higher than the quantification levels of other transcripts. This difference did not allow us to subtract one from another directly, and miRNA expression should be scaled to a comparable level of the coding RNA transcripts. To know which value should be given for this scaling factor, we first analyzed which factors would be optimal for TrT, that is, in which miRNA had enough power to show a difference between TrT and TPM, but not so powerful that it reduced all expression values to zero. We searched such optimal value by evaluating a range of values between 0 and 1 (0 leading to no miRNA effect at all, and 1 leading to no scaling), and selected the values in which TrT maximized the correction benefit in comparison to TPM (taking proteomics as a reference). The optimal range we observed was between 0.1 and 0.27. Besides, initial investigation of a newly developed sequencing method named Combo-Seq, which allows sequencing both transcripts and miRNAs altogether in a single library preparation, showed us that the proportion observed of miRNA represented around 10% of all transcripts sequenced.

Consequently, we decided to set the miRNA factor to '0.1' (Translatable Transcripts (TrT) formulation)

Correlation between Transcriptomics (TPM/TrT) and Proteomics

Having our method described and finished, we applied our method to all the aforementioned data. Afterward, we aimed to compare it with the state-of-the-art TPM to investigate in which manner TrT could be an improvement. Initially, we wanted to confirm the initial issue (transcriptomics does not accurately represent the proteome), and compare such results with TrT. For that purpose, using the untreated samples, we performed a correlation analysis between proteomics and TPM, followed by its analogous analysis between proteomics and TrT.

The correlation analysis confirmed the low correlation between Transcriptomics (TPM) and proteomics values (Figure 3A), the average correlation of which was 0.39 (± 0.06 SD). There was no specific time interval between both omics that presented an improved correlation. For example, independently of the time difference between both omics, 24h samples (be it proteomics or transcriptomics) showed the highest correlations. Simultaneously, 2h proteomics and transcriptomics samples presented some of the lowest correlations. Even so, the correlation values were not randomly distributed, but substantially influenced by which proteomics time-point they were related to. As an example, time-point 24h presented the highest values regardless of the transcriptomics time-point (avg.: 0.46), while time-point 2 hours presented the lowest (avg.: 0.26). To a lower extent, a similar effect was observed for the transcriptomics time-points. When evaluating TrT at the same level, we saw a similar low correlation with proteomics (Figure 3B), confirming the fact that a representative portion of transcripts (80%) presented equal values between TPM and TrT, due to not presenting any predicted inhibition by any miRNA expressed in our dataset.

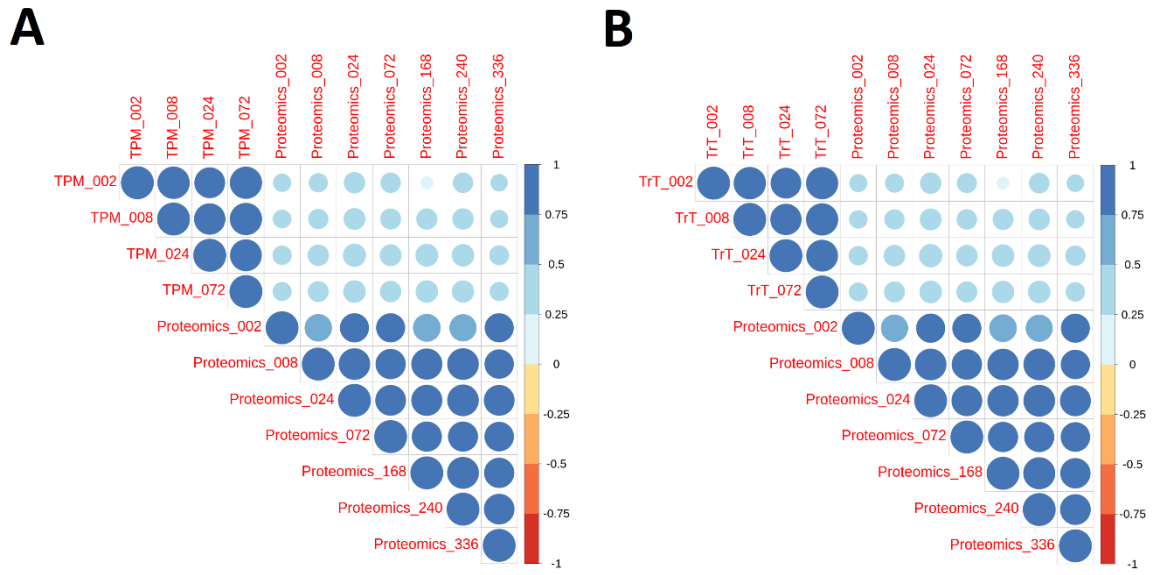


Figure 3: Correlation analysis between Transcriptomics and proteomics. A: *TPM vs Proteomics.* ***B:*** *TrT vs Proteomics. TPM: Transcripts Per Million. TrT: Translatable Transcripts. Timepoint samples are compared across each other (from 2 to 336 hours). The darker and bigger a circle is, the higher the absolute value of correlation it represents. Blue stands for positive correlation, red for negative correlation.*

Differential Expression Analysis

The high similarity between both correlation analyses was not surprising, as we predicted and formulated for miRNA to have generally a small regulatory effect. Even so, we hypothesized that for a portion of the transcripts regulated by miRNA, TrT would be a better proxy for proteomics. To verify this, we ran our formula through all our data. Later, we investigated in which cases traditional transcriptomics was falsely reporting as a proxy by taking proteomics as a reference. The analysis involved the identification of differentially expressed molecules across doses. Two possible scenarios unfolded: there could be a change in transcriptomics not reflected in proteomics, and vice versa: a change in proteomics that was not reflected in transcriptomics. For example, in fluorouracil's (5-FU) exposure, 5.48% of genes presented a change in transcriptomics and not in proteomics, while 4.02% of them did not present a change while proteomics did. Therefore, we focused on the genes for which TrT could correct such cases.

The myosin heavy chain 9 (MYH9) gene was one of them. We could observe a significant increase ($p.value < 0.05$) (Figure 4A) of the protein at a Toxic dose of 5-FU, while for TPM it was not ($p.value > 0.05$) (Figure 4B). TrT, on the contrary, did reflect a significant increase ($p. adjusted < 0.05$) (Figure 4C).

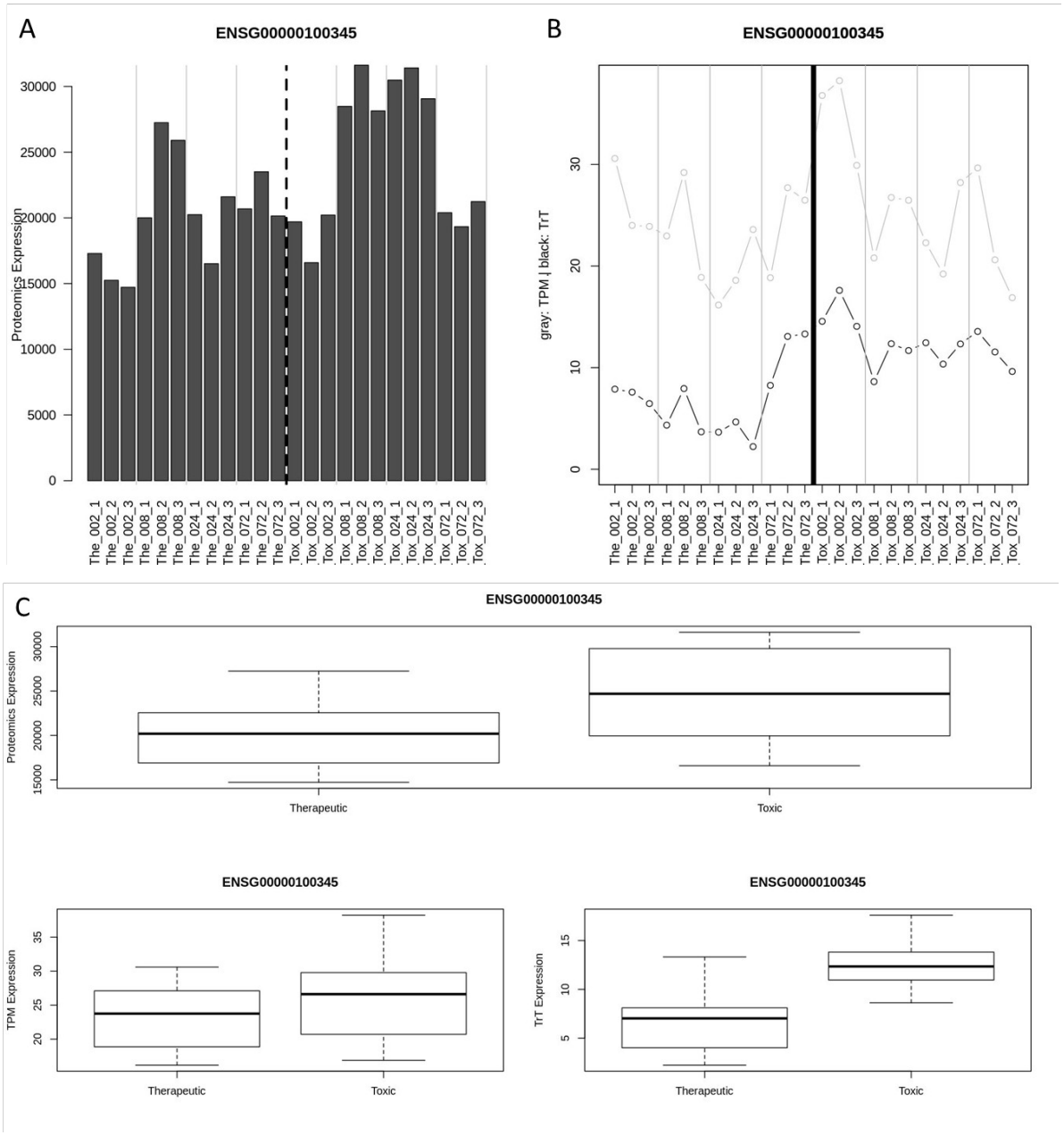


Figure 4: MYH9 expression values for therapeutic and toxic doses of 5-FU. A: *Proteomics expression values.* ***B:*** *TPM and TrT expression values: Gray: TPM, Black: TrT.* ***C****: Boxplots of the expression values for proteomics, TPM, and TrT. TPM: Transcripts Per Million. TrT: Translatable Transcripts.*

We could also find examples of the opposite. Such was the case for TGFBI (transforming growth factor beta-induced). In this case, proteomics did not show any significant change across doses of Docetaxel (Figure 5A), but TPM was significantly increased (p.adjusted < 0.05) in the toxic dose (Figure 5B). TrT, though, was consistent with proteomics, not showing any significant difference between doses (p.value > 0.05) (Figure 5C).

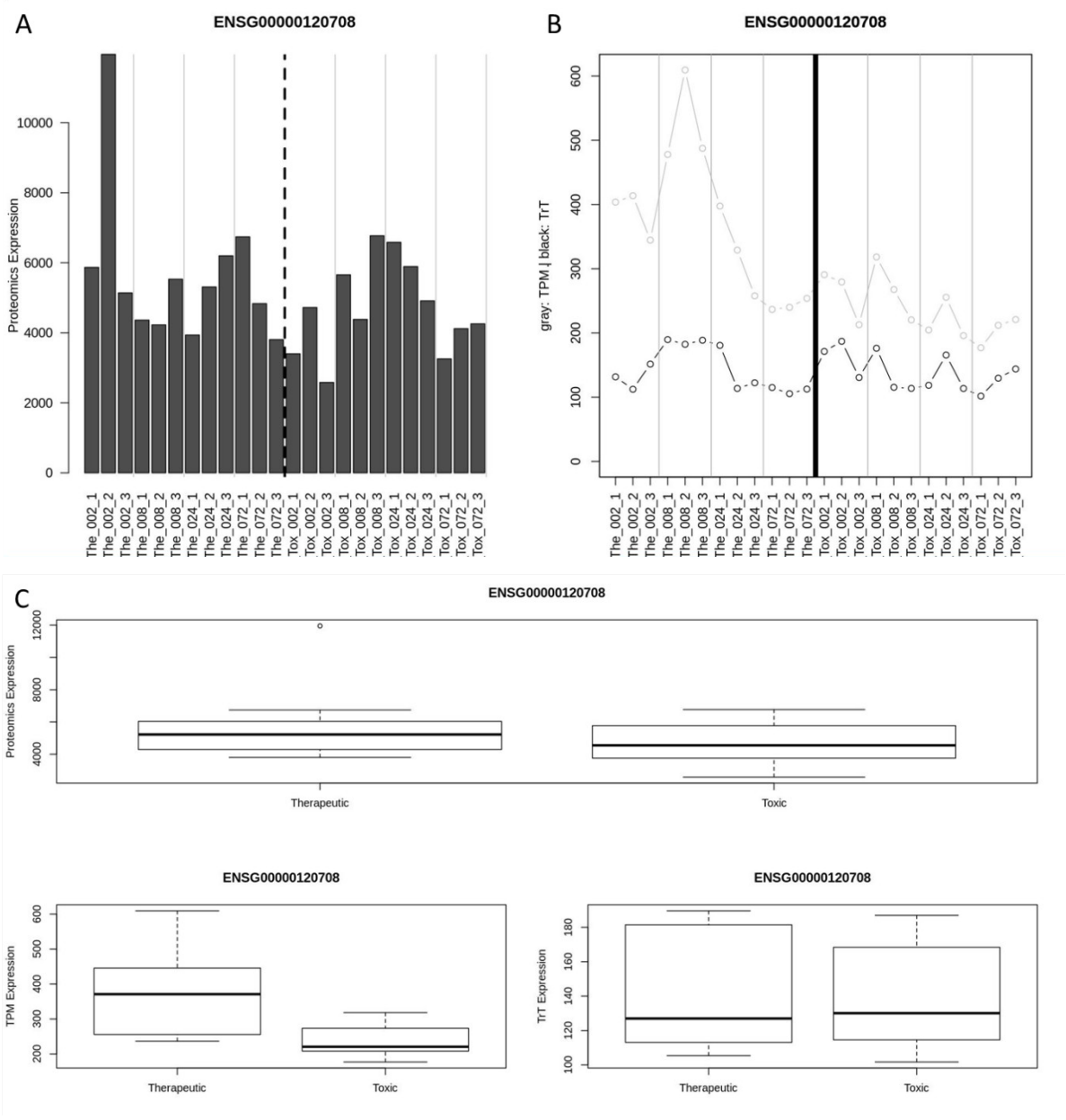


Figure 5: Expression values of TGFBI. ***A:*** *Proteomics expression values.* ***B:*** *TPM and TrT expression values: Gray: TPM, Black: TrT.* ***C:*** *Boxplots of the expression values for Proteomics, TPM, and TrT. TPM: Transcripts Per Million. TrT: Translatable Transcripts.*

Biological Interpretation (GOrilla)

After observing the effects TrT could have, we investigated whether those could also affect the biological interpretation. Specifically, we focused on comparing the enriched gene ontology sets (GO-sets) of both TPM and TrT via GOrilla. The results showed that the number of GO terms was always higher in TPM than in TrT (Figure 6). Such a decrease in GO terms for TrT, though, was not related to a lower number of DEGs from that quantifier. When analyzing the commonalities and differences between TPM and TrT across compounds, we observed 2 different behaviors (Figure 6). For some

compounds (5FU, AMI, EPI & MXT), the number of TrT GO terms shared with TPM was either greater or equal to the exclusive ones, while most of the TPM terms were exclusive. On the other compounds, the GO terms shared between them were rather the minority for both sets.

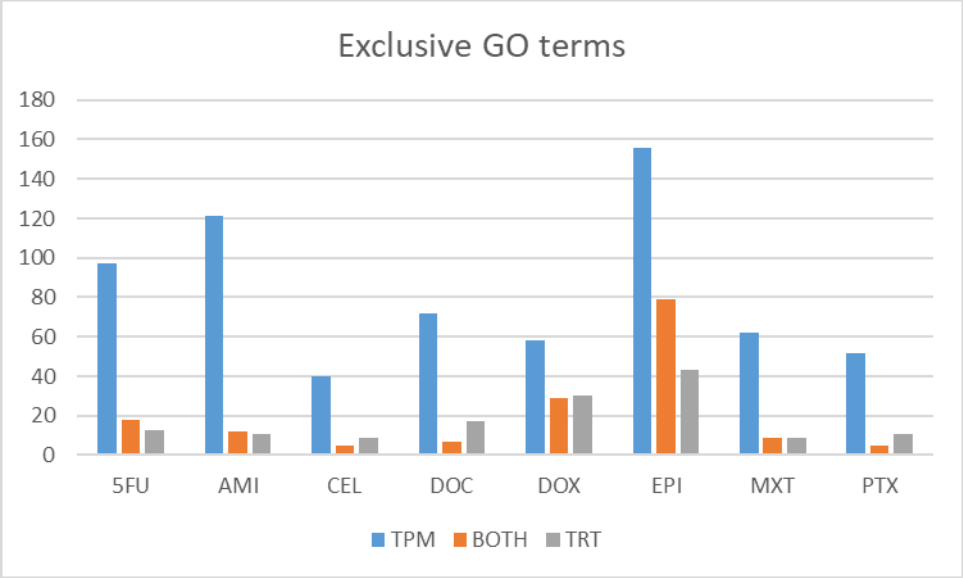


Figure 6: Number of exclusive GO terms for each quantifier (TPM & TRT) and the ones included in both (BOTH). *TPM: Transcripts Per Million. TrT: Translatable Transcripts. 5FU: Fluorouracil. AMI: Amiodarone. CEL: Celecoxib. DOC: Docetaxel. DOX: Doxorubicin. EPI: Epirubicin. MXT: Mitoxantrone. PTX: Paclitaxel.*

An example of how the misclassification of a DEG may affect the biological interpretation of a comparison (with enough genes being misclassified) was the Cardiac-Specific Homeo Box (NKX2-5) gene. According to TPM, the expression of this gene was significantly higher in the UNTR samples when compared to its corresponding samples in the therapeutic dose in 5FU, AMI, and DOC, leading to the enrichment of the Cardiac Muscle Tissue Morphogenesis GO set (of genes). TrT, though, showed no difference between both scenarios (Figure 7A). This meant that the miRNA regulation was predicted to be stronger in the control samples than the treatment ones, to the point where the difference between both conditions was not significant. To contrast that prediction, we searched for all 18 MTIs with either weak or strong evidence of their inhibitory regulation. We observed that the expression of all those miRNAs in the Control samples was greater or equal to the therapeutic ones, confirming the TrT prediction (Figure 7B).

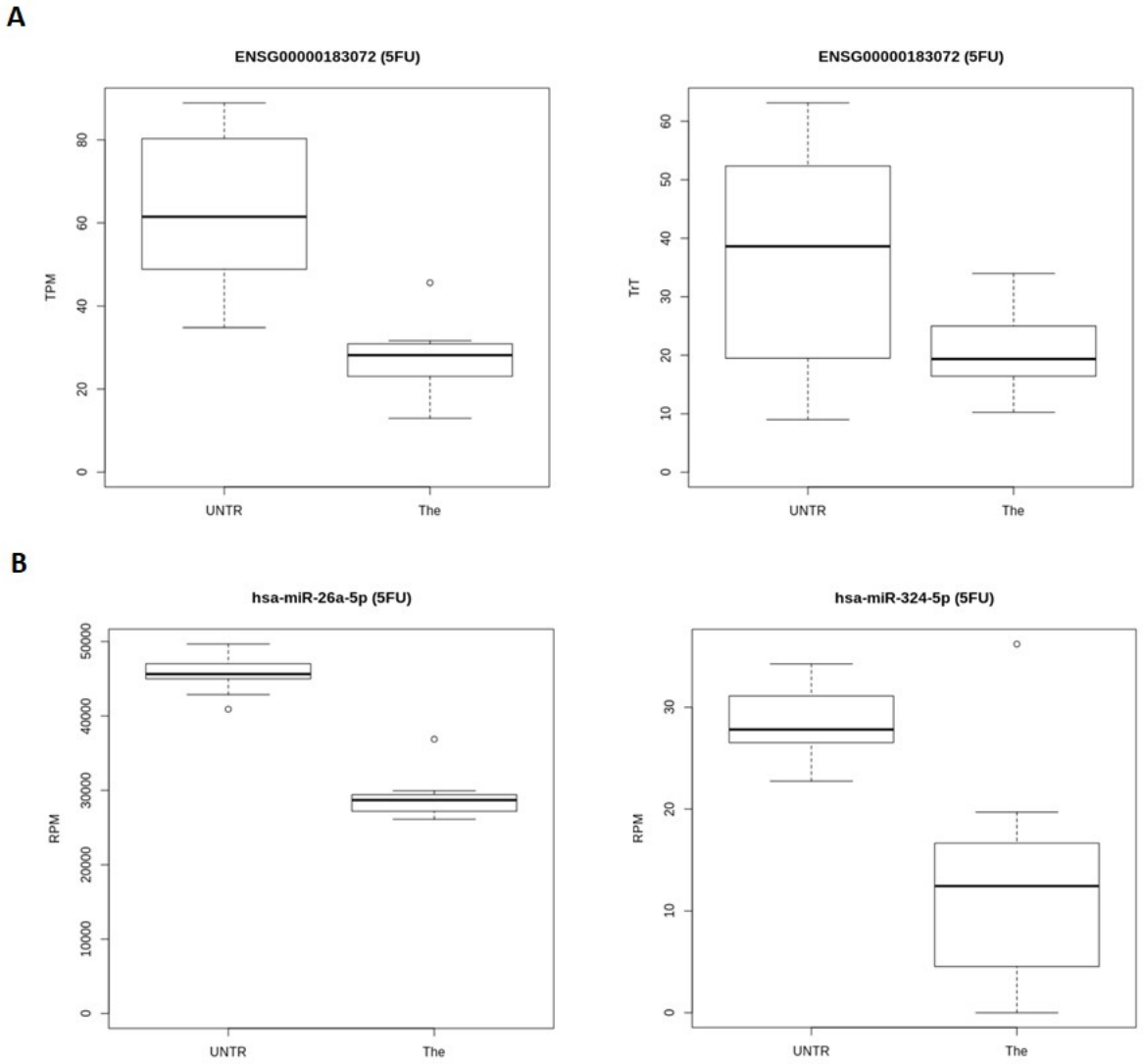


Figure 7: NKX2-5 gene: Expression and Regulation between Untreated (UNTR) and Therapeutic (The) samples treated with Fluorouracil (5FU). A. *TPM (left) and TrT (right) expression represented in boxplots.* ***B.*** *Examples of the expression in Reads Per Million (RPM) of miRNAs that regulate the NKX2-5 gene.*

Sensitivity, Specificity, and Accuracy results in TPM and TRT

So far, we saw that, at least in some cases, TrT could be a better proxy for proteomics than TPM, and that the biological interpretation was simplified when TrT was applied. Afterward, we wanted to assess the global accuracy of TrT and TPM by comparing it to proteomics.

We calculated the accuracy of both transcriptomics values for all compounds in 3 different comparisons: therapeutic versus toxic doses, untreated control versus therapeutic dose, and untreated control versus toxic dose. To evaluate the differential expression of the last 2 comparisons, we used the same methodology as the one used for therapeutic versus toxic doses.

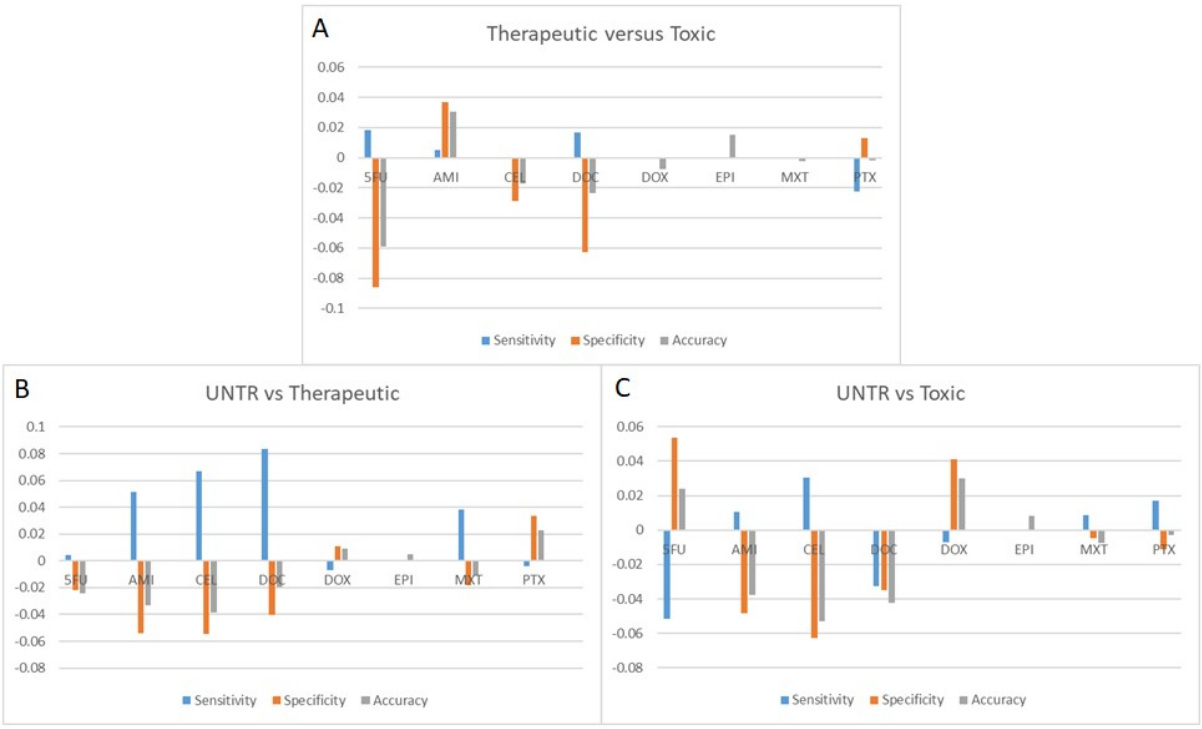


Figure 8: Difference in Sensitivity, Specificity, and Accuracy when changing the quantifier from TPM to TrT, each comparison encompassing 8 compounds: *Fluorouracil (5FU), Amiodarone (AMI), Celecoxib (CEL), Docetaxel (DOC), Doxorubicin (DOX), Epirubicin (EPI), Mitoxantrone (MXT), and Paclitaxel (PTX).* ***A****: Therapeutic versus Toxic doses.* ***B****: Untreated versus Therapeutic dose.* ***C****: Untreated versus Toxic dose.*

We observed an ambiguous effect of TrT in relation to TPM in all 3 comparisons from a global perspective (Figure 8). Individually, though, we observed a compound effect reflected on different accuracy differences for each of the 8 compounds, which could be categorized into 3 groups when analyzing the accuracy differences for each compound. Most of them showed ambiguous results, in which both improvement and decrease in accuracy happened. EPI, instead, showed a continuous improvement across all comparisons. The last group, contrarily (CEL & MXT), showed a continuous decline across all comparisons.

Discussion

The correlation between transcriptomics and proteomics has always tended to be low. This is true both in previous and in the current study, proving transcriptomics as a poor proxy of the phenotypical changes of the cell. Trying to predict the regulatory effect of miRNA (TrT) to improve such proxy (TPM) proved beneficial for a portion of the gene set. This is unsurprising, knowing that not all genes are regulated by miRNAs, and only a small portion of proteins are quantified through the proteomics pipeline. TrT also showed a simplified biological interpretation of the changes across conditions, although its results compared to TPM varied compound-wise.

We did not only take miRNAs into account, though. Other ncRNAs, such as circRNAs (although recent studies have observed some of them as coding), were also quantified. Due to the novelty of circRNAs, in contrast with other OMICs, there was no golden standard database to refer to when seeking their specific IDs nor the way they should be quantified, among other reasons, because the existing databases were seldom updated[18,49] nor online[50].

One of the most recent recommendations[44] in these circumstances was the combined use of at least 2 different circRNA identification software tools. The reason behind this procedure was based on the flawed accuracy of these programs due to their inherent biases. Fortunately, even if this bias differed for every tool, it had been shown that the convergence between the 2 tools offered a much lower proportion of false positives in contrast to their original population of predictions[44].

We decided to pool all samples so that we had the maximum sequencing depth available for the identification process. Otherwise, selecting a random sample would lead to an under-representation of some circRNA molecules, which would not be identified as such, even if they would be in the rest of the samples.

The argumentation behind the selection of CIRI2 and circExplorer2 was twofold. First, their accuracy outcomes in previous reviews portrayed them to be some of the best predictors currently available. Secondly, our comparison of several predictors revealed that both output a substantial proportion of molecules that were also outputted by other predictors (low false-positive rate), without sacrificing sensitivity. This was not the case for circRNA_finder[32], which found several times more molecules than the other predictors with a very low overlap (~7%). find_circ[18], although having a similar proportion of overlap as circExplorer2 and CIRI2, did not present such a good accuracy in previous studies.

The model we applied considered several factors in post-transcriptional regulation. We built it by trying to reach an equilibrium at the complexity level: enough complexity to maximize the representation of the molecular reality of the cell, but without including so many variables that would have made the model either unbuildable or unfeasible to run in practical terms. Our model also reflected how little is still known about post-transcriptional regulation, and recent discoveries are constantly being made public around it[42,51].
This is especially the case with miRNA regulation, where the complementarity and regulatory strength seem to not only be dependent on the seed region but other loci such as the 3' end of miRNAs[42]. These new observations are currently being incorporated into the interaction predictive tools, making probably in the future the more classical tools obsolete.

The genes exemplified as good proxies for proteomics are indeed regulated by miRNAs. MYH9 is a non-muscle myosin chain IIA gene which plays a major role in early mammalian development, while in the adult heart is only expressed in the non-myocyte cells[52]. Several miRNAs have been either positively or negatively associated with the gene's expression[53]. For TGFBI, miRNA regulation has also been shown, to the point of even being related to chemoresistance[54].

Our model was based on miRNA regulation on a 1-on-1 basis: if a miRNA can bind a transcript (after taking into account the endogenous competition), the latter will not be able to be translated. Other recent models prefer to focus on the targeting efficacy (as a function of the affinity between the two molecules) while ignoring the ceRNAs[51]. We acknowledge that both models lack each other's strengths, and a combination of the two would have been optimal.

In addition to that, there could have been other important factors that could have been taken into account. One example could be the translation ratio, i.e., the number of proteins translated per transcript, which depends on the number of ribosomes that may bind to it, and its half-life. Indeed, half-lives, both of the transcript and the protein, make it difficult to transient from one OMICs to another: the increase in proteomics might be by an increase in transcriptomics with a short half-life that cannot be seen, or a long half-life of proteins may delay the decrease in their abundance due to a decrease in translational output.

Proteomics, although being the technology used as a reference to represent the phenotypical changes in the cell, is quite limited in achieving so. This is especially the case due to its lack of sensitivity, where generally only around 1000-3000 proteins are quantified, making the analysis of the proteome quite challenging. In addition to that, the output scale of such quantifications can be quite dramatic when the technology or machinery involved changes.

Our next steps would include trying to breach the gap between both OMICs by, not only including important parameters that can improve the proteomics prediction, but also changing the algorithms involved in the model, namely, using a supervised quantitative machine learning model. That way, the use of new parameters is easier to both implement and evaluate. This is especially relevant for the imputation of proteomics data, prone to missing values due to the stochastic nature of mass spectrometry. Our work concerning this strategy can be found in our published article[57].

We created a formula to model post-transcriptional regulation and its effects on protein expression. We assessed its behavior in both general and specific cases. As expected, due to the variability of each gene's regulation, the formula benefited only a subset of genes. We portrayed the importance of reaching a more accurate description of the cellular changes, and how far we still are from that

objective. Therefore, we must not simply continue with the most popular technology, but try to reach a better approach to the current era of big data.

Supplementary data

Supplementary data can be found in the online preprint version of this manuscript[58].

Bibliography

1. Zhang, Z., Wu, S., Stenoien, D. L. & Paša-Tolić, L. High-throughput proteomics. *Annu. Rev. Anal. Chem. (Palo Alto. Calif).* **7**, 427–54 (2014).
2. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome. *J. Proteome Res.* **2**, 43–50 (2003).
3. Washburn, M. P., Wolters, D. & Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001).
4. Cagney, G. *et al.* Human tissue profiling with multidimensional protein identification technology. *J. Proteome Res.* **4**, 1757–67 (2005).
5. Chen, G. *et al.* Discordant protein and mRNA expression in lung adenocarcinomas. *Mol. Cell. Proteomics* **1**, 304–13 (2002).
6. Lemée, J.-M. *et al.* Integration of transcriptome and proteome profiles in glioblastoma: looking for the missing link. *BMC Mol. Biol.* **19**, 13 (2018).
7. Rogers, S. *et al.* Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics* **24**, 2894–2900 (2008).
8. Dhingra, V., Gupta, M., Andacht, T. & Fu, Z. F. New frontiers in proteomics research: A perspective. *International Journal of Pharmaceutics* vol. 299 1–18 (2005).
9. Belle, A., Tanay, A., Bitincka, L., Shamir, R. & O'Shea, E. K. Quantification of protein half-lives in the budding yeast proteome. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 13004–13009 (2006).
10. Baek, D. *et al.* The impact of microRNAs on protein output. *Nature* **455**, 64–71 (2008).
11. Selbach, M. *et al.* Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58–63 (2008).
12. Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350–355 (2004).
13. Lewis, B. P., Shih, I., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of Mammalian MicroRNA Targets. *Cell* **115**, 787–798 (2003).
14. Lim, L. P. *et al.* Microarray analysis shows that some microRNAs downregulate large numbers of-target mRNAs. *Nature* **433**, 769–773 (2005).
15. Zaphiropoulos, P. G. Exon skipping and circular RNA formation in transcripts of the human

cytochrome P-450 2C18 gene in epidermis and of the rat androgen binding protein gene in testis. *Mol. Cell. Biol.* **17**, 2985–2993 (1997).

16. Chen, L. L. & Yang, L. Regulation of circRNA biogenesis. *RNA Biol.* **12**, 381–388 (2015).
17. Jeck, W. R. *et al.* Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19**, 141–157 (2013).
18. Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–338 (2013).
19. Rong, D. *et al.* An emerging function of circRNA-miRNAs-mRNA axis in human diseases. *Oncotarget* **8**, (2017).
20. Wang, K. *et al.* The long noncoding RNA CHRF regulates cardiac hypertrophy by targeting miR-489. *Circ. Res.* **114**, 1377–88 (2014).
21. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146**, 353–8 (2011).
22. Kuepfer, L. *et al.* A model-based assay design to reproduce in vivo patterns of acute drug-induced toxicity. *Archives of Toxicology* vol. 92 553–555 (2018).
23. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–20 (2014).
24. Baras, A. S. *et al.* miRge - A Multiplexed Method of Processing Small RNA-Seq Data to Determine MicroRNA Entropy. *PLoS One* **10**, e0143066 (2015).
25. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
26. Zhang, X.-O. *et al.* Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res.* **26**, 1277–87 (2016).
27. Gao, Y., Zhang, J. & Zhao, F. Circular RNA identification based on multiple seed matching. *Brief. Bioinform.* **19**, 803–810 (2018).
28. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* vol. 25 1754–1760 https://pubmed.ncbi.nlm.nih.gov/19451168/ (2009).
29. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–2 (2010).
30. ENSEMBL. FTP Download. https://www.ensembl.org/info/data/ftp/index.html (2020).
31. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
32. Westholm, J. O. *et al.* Genome-wide Analysis of Drosophila Circular RNAs Reveals Their Structural and Sequence Properties and Age-Dependent Neural Accumulation. *Cell Rep.* **9**, 1966–1980 (2014).
33. Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34**, D140–D144

(2006).
34. Durinck, S. M. Y. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
35. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
36. John, B. *et al.* Human microRNA targets. *PLoS Biol.* **2**, (2004).
37. R Core Team. R: A Language and Environment for Statistical Computing. (2019).
38. RStudio Team. RStudio: Integrated Development Environment for R. (2019).
39. Wei, T. & Simko, V. R package 'corrplot': Visualization of a Correlation Matrix. (2017).
40. Analyzing RNA-seq data with DESeq2. https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#why-un-normalized-counts.
41. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, 1–9 (2010).
42. Chipman, L. B. & Pasquinelli, A. E. miRNA Targeting: Growing beyond the Seed. *Trends in Genetics* vol. 35 215–222 (2019).
43. Chou, C. H. *et al.* MiRTarBase update 2018: A resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* (2018) doi:10.1093/nar/gkx1067.
44. Hansen, T. B. Improved circRNA identification by combining prediction algorithms. *Front. Cell Dev. Biol.* **6**, 20 (2018).
45. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–33 (2009).
46. Arvey, A., Larsson, E., Sander, C., Leslie, C. S. & Marks, D. S. Target mRNA abundance dilutes microRNA and siRNA activity. *Mol. Syst. Biol.* **6**, 363 (2010).
47. Franco-Zorrilla, J. M. *et al.* Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat. Genet.* **39**, 1033–7 (2007).
48. MS, E., JR, N. & PA, S. MicroRNA Sponges: Competitive Inhibitors of Small RNAs in Mammalian Cells. *Nat. Methods* **4**, (2007).
49. Li, J. H., Liu, S., Zhou, H., Qu, L. H. & Yang, J. H. StarBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* **42**, D92-7 (2014).
50. Liu, Y. C. *et al.* CircNet: A database of circular RNAs derived from transcriptome sequencing data. *Nucleic Acids Res.* **44**, D209–D215 (2016).
51. McGeary, S. E. *et al.* The biochemical basis of microRNA targeting efficacy. *Science (80-.).* **366**, eaav1741 (2019).
52. Ma, X. & Adelstein, R. S. In vivo studies on nonmuscle myosin II expression and function in heart development. *Front. Biosci.* **17**, 545–555 (2012).

53. Yu, M. *et al.* Prognostic impact of MYH9 expression on patients with acute myeloid leukemia. *Oncotarget* **8**, 156–163 (2017).
54. Bissey, P. A. *et al.* Dysregulation of the MiR-449b target TGFBI alters the TGFβ pathway to induce cisplatin resistance in nasopharyngeal carcinoma. *Oncogenesis* **7**, (2018).
55. Nakamoto, M., Jin, P., O'Donnell, W. T. & Warren, S. T. Physiological identification of human transcripts translationally regulated by a specific microRNA. *Hum. Mol. Genet.* **14**, 3813–3821 (2005).
56. Moore, M. J. *et al.* MiRNA-target chimeras reveal miRNA 3′-end pairing as a major determinant of Argonaute target specificity. *Nat. Commun.* **6**, 1–17 (2015).
57. Ochoteco Asensio J, Verheijen M, Caiment F. Predicting missing proteomics values using machine learning: Filling the gap using transcriptomics and other biological features. *Comput Struct Biotechnol J.* 2022 Apr 22;20:2057-2069. doi: 10.1016/j.csbj.2022.04.017. PMID: 35601960; PMCID: PMC9077535.
58. Ochoteco Asensio, J., Kleinjans, J. & Caiment, F. Quantifying the number of translatable transcripts through the use of OMICs involved in post-transcriptional regulation. *bioRxiv*, doi:https://doi.org/10.1101/2022.06.20.496876 (2022).

Chapter 4:

Predicting missing proteomics values using machine learning: Filling the gap using transcriptomics and other biological features

Predicting missing proteomics values using machine learning

Filling the gap using transcriptomics and other biological features

Juan Ochoteco Asensio[1], Marcha Verheijen[1], Florian Caiment[1,*]

1. Department of Toxicogenomics, School of Oncology and Developmental Biology (GROW), Maastricht University, Maastricht, The Netherlands

* Corresponding Author: florian.caiment@maastrichtuniversity.nl

(Manuscript published)

ABSTRACT

Proteins are often considered the main biological element in charge of the different functions and structures of a cell. However, proteomics, the global study of all expressed proteins, often performed by mass spectrometry, is limited by its stochastic sampling and can only quantify a limited amount of protein per sample. Transcriptomics, which allows an exhaustive analysis of all expressed transcripts, is often used as a surrogate. However, the transcript level does not present a high level of correlation with the corresponding protein level, notably due to the existence of several post-transcriptional regulatory mechanisms. In this publication, we hypothesize that the missing protein values in proteomics could be predicted using machine learning regression methods, trained with many features extracted from transcriptomics, including known translational regulatory elements such as microRNAs and circular RNAs. After considering different machine learning algorithms applied on two different splitting strategies, we report that random forest can predict proteins in new samples out of transcriptomics data with good accuracy. The proposed pre-processing and model building scripts can be accessed on GitHub: https://github.com/jochotecoa/ml_proteomics

1 INTRODUCTION

For a cell to react and adapt to any variation of its environment, including for instance the exposure to a foreign compound, a cascade of events leading ultimately to the production of proteins occurs. For that purpose, the cell usually initiates the transcription of its genes (such as transcription factor), and the resulting transcripts containing an open reading frame are translated into proteins. Even though such a schematic view of molecular biology appears straightforward, each of those steps is controlled and affected by a myriad of factors. This complexity led to the development of advanced technologies, named "omics", allowing to deeply study a particular class of biological entity: transcriptomics (characterization and quantification of transcripts), proteomics (proteins), metabolomics (metabolites), etc.

Among those different classes of molecules, proteins are particularly relevant, as their expression level and activity inform profoundly about how the cell is functioning and reacting to its environment, especially when those changes may pose a risk to the integrity and functionality of the whole system, either due to a disease or an infection. To analyze the expression of proteins in different conditions, proteomics (mass spectrometry or MS) is usually applied. Unfortunately, its sensitivity is limited[1-3], and thus only a small subset of proteins (with the highest abundance) can be studied at a time. In addition, the stochastic sampling generates missing identifications across samples, particularly for proteins with an abundance close to the detection limit; even though workflows such as DIA (Data-Independent Acquisition)-MS workflow can increase reproducibility. New technologies are not exempt of these limitations: the latest single-cell proteomics strategies (such as SCoPE2[2]) and newest experimental and computational workflows[3] only obtain ~1000 proteins per cell on average (not including their own limitations[4]), even though their dynamic range allows for the quantification of 3000 distinct proteins.

Proteins are mainly translated from messenger RNAs (mRNAs), which are much easier to analyze. Indeed, while having a shorter half-life than proteins, mRNA transcriptomics has become overwhelming sensitive and cost-efficient over the years with the invention of next-generation sequencing. For these reasons, RNA-Sequencing techniques are usually preferred to statistically study cell changes at the molecular level. However, a given mRNA is not an excellent proxy of its corresponding protein expression level, which is reflected in a very low correlation between transcriptomics and proteomics technologies[5-9]. While the reasons behind this gap can be multiple, the main factors can be categorized into post-transcriptional regulation. By different mechanisms in such regulation, the cell controls the final level of translation of each mRNA into proteins. These factors can

be either determined by the molecules themselves (such as the transcript's or protein's half-life[10]) or by the interaction with external elements.

MicroRNAs (miRNAs), short non-coding transcripts of around 22 nucleotides of length, play an important role in post-transcriptional regulation. They can act as inhibitors of translation[11,12] by base-pairing their seed region[13] (nucleotide 2 to 8) to the target mRNA, usually in their 3` UTR region. While often considered mild individually, the interaction of multiple miRNAs (either the same miRNA or different miRNAs) on the same 3'UTR target can have a significant effect on protein level expression [14,15]. Considering the relatively short length of the seed region, miRNAs can target an average of 200 different targets. Even so, miRNAs are not the only transcripts regulating translation.

Another newly discovered category of RNAs, named circular RNAs (circRNAs), are characterized by their circular form, which is generated by the binding of their 5 and 3` end during splicing (back-splicing)[16,17], forming the so-called back-spliced junction. Due to this particular structure, they are not easily degraded due to the absence of transcript extremities, rendering them immune to exonuclease activity[18]. Several functions have been proposed for these circRNAs, including regulating miRNA activities. It has been demonstrated that circRNAs, which can contain repetitions, could present the same target regions present in miRNA targets, and sometimes several times per molecule. This leads to a target competition[19], where circRNAs bind most miRNAs, which gave to circRNAs the function of 'miRNA sponges'[20]. The post-transcriptional regulation complexity starts to unfold once one realizes that each transcript can be inhibited by several miRNAs, and at the same time, each of those miRNAs can be "sponged" by one or more circRNAs.

The final expression level of a protein results thus from the integration inside the cell of many factors related to transcripts: the level of expression of mRNAs, the number of possible seeds with miRNAs, the expression level of miRNAs, and the expression level (and "sponging" capacity) of circRNAs able to capture these miRNAs. Many other features could also play a role in this final protein expression level. For instance, the GC content of an mRNA has been observed to interfere with the mRNA half-life[21], and thus the total number of proteins formed from a single mRNA. All these RNA elements or characteristics just mentioned could be identified and quantified by transcriptomics with RNA-Sequencing. Since the protein expression level is the most important factor for biological interpretation, and considering the limited sensitivity and stochastic sampling of proteomics in addition to the very low correlation of the mRNA/protein expression level, we considered the possibility of obtaining predicted protein expression levels from the integration of as many possible features available from sev-

eral OMICs data. Although we recognized that methods such as match-between-runs (MBR)[22], DART-ID[23], and IceR[24] have already been developed (and their limitations[25]), including a deep learning approach to extrapolate proteomics values from transcriptomics values[26], none utilized a complex multi-omics strategy to approach in a novel manner the limitations of proteomics.

The amount and complexity of the data render impossible the task of manually integrating all these parameters. Even when inputting such data digitally, it is not straightforward to visualize which is the optimal manner to predict proteomics values. This problem is characteristic of the current big data era, which in turn, has led to the rise of algorithms that use straightforward optimization strategies to rapidly process thousands or millions of observations. Some of those can be categorized as machine learning (ML), which consists of a set of computer algorithms built to automatically improve their prediction with increasing volumes of data[27]. Specifically, the algorithms focused on predicting are part of the supervised learning algorithms, as they require a training phase in which they are exposed to the value to be predicted (target) in conjunction with other variables associated with it (features). Two major classes of machine learning algorithms exist: when predicting categories or labels (qualitative values), algorithms will perform a classification; while when what is predicted are quantitative values, algorithms will perform a regression. The improvement in the accuracy of these models can be evaluated based on how similar the predictions are to the actual observations. The accuracy is only relevant to evaluate with new data (testing dataset), and not with the data used to train the model (training dataset), in order to avoid the generation of a biased model due to overfitting.

In this manuscript, we hypothesized that using machine learning algorithms would allow us to estimate the expression level of the protein not detected by proteomics out of all available data. For the omics data, we made use of an *in vitro* dataset obtained from primary human hepatocytes microtissues which includes 3 omics datasets obtained from the exact same samples batch: RNA-Seq (ribo-depleted libraries), miRNA-Seq (small RNA libraries), and proteomics (mass spectrometry). Both mRNA and circRNAs quantification were extracted from the RNA-Seq data. We thus assessed the accuracy of diverse machine learning predictive models based on different algorithms and data-splitting strategies with the ultimate goal to predict protein expression value from transcriptomics and other mRNA features.

2 METHODS

2.1 DATASET & FEATURES

The description of the biological samples used, in addition to the proteomics and transcriptomics protocols followed to obtain protein and RNA expression values, can be found in the Supplementary Methods.

Proteomics expression values were set as the target to be predicted. We set as features protein properties with nominal values extracted from UniProt that might affect their half-life. The features were the following: protein length (Length), mass (Mass), quantity of each amino acid (Aa_X), organism (Organism), location on which the original gene was encoded (Gene.encoded.by), and the database version of the protein sequence (Version..sequence.). From those, we also derived additional features: linear density (mass divided by length) and proportion of each amino acid based on the protein's length (Aa_X_prop). Finally, we added some irrelevant features (protein sequence version) as negative controls to inform us of the model reliability (based on the importance these features would be given by those models). Concerning protein stability, we included all nine features extracted from the supplementary table: R1-R7, PSI, and SD.

The expression values (in TPM) of protein-related transcripts were added as a feature. Furthermore, we also added diverse transcript properties: strand, transcript length, percentage gene GC content, CDS length, UTR length (or non-CDS length), and proportion of UTR length (UTR length divided by the transcript length). MiRNA expression was also added as a feature, linking it to the transcript targets they could potentially regulate. For this, we used the miRDB's MiRNA Target Interaction (MTI) score in two features in the ML algorithm: one feature with only miRNAs that presented a high probability of targeting such target (`stringent`, score >= 80), and another considering all possible regulations, independently of their score (`all`). CircRNA expression as a feature ('circ') was linked to the proteomics values based on the miRNA sponging effect of the former. We only utilized the expression of those circRNAs that presented more than 7 targeting sites with a specific miRNA. We also added the sponging effect of circRNAs as the feature 'circ_score'.

Transcripts were named based on their Ensembl ID, while proteins were labeled with UniProt IDs. A single UniProt protein could be associated with more than one ENST transcript, potentially with very different features (expression level, transcript length, etc.). Therefore, we needed to summarize the value from all linked transcripts in a single feature. As there was no clear advantage to select a

particular summary method over another, we created a feature for each of those different methods: mean, median, minimum, maximum, sum, and standard deviation. This approach was not only applied to features associated with transcripts coupled to proteins (and their log2-transformed values) but also to the ones associated with miRNAs and circRNAs (and their log2-transformed values as well). Indeed, this problem was also applicable to those molecules (to even a greater extent) when linking them to a single proteomics value: each transcript can be inhibited by several miRNAs and each of those miRNAs can be sponged by several circRNAs. We also extended this strategy to those features that presented a multiplicity of values for a single observation, such as the protein stability data. The combination of all discussed variables led to a total of 196 features.

2.2 PRE-PROCESSING

For both the pre-processing of the data and the construction of the machine learning models, we used the R library 'caret'[28].

2.2.1 Creating dummy variables

Since categorical data (such as gender) cannot be inputted directly into a model, they needed to be transformed into dummy variables. Dummy variables are binary features that indicate the presence (1) or absence (0) of a categorical value. In our data, the dummy variables created were related to strand information (positive (+) or negative (-) strand) and protein version sequence (presence or absence of versions 1 to 7).

2.2.2 Identifying (Near) Zero-Variance and Correlated Predictors

To identify variables with no variance (Zero-Variance or ZV) or insignificant variance (Near Zero-Variance), we used the function 'nzv' (frequency ratio > 95/5, percentage unique < 10%) described in 'caret'[29]. We then discarded those predictors from the dataset. To identify correlated variables (correlation > 0.75), we used the function 'findCorrelation' also from the 'caret' package. We discarded the identified correlated predictors from the dataset. The correlation plot was designed using the 'corrplot' package.

2.2.3 Centering and Scaling

Centering refers to the data transformation where the means of all features are set to a specific value (i.e., 0) while scaling refers to the transformation where the standard deviation is also set to a constant value (i.e., 1). These data transformations avoid a feature importance bias due to value size or scale. No

imputation was performed, but instead, all observations with any missing value were removed from the dataset.

2.2.4 Data splitting and algorithms used

The data split between the training dataset (80% of the whole dataset) and the testing dataset (20% of the whole dataset) was performed based on 2 different strategies: sample names and protein names. For each algorithm used, we performed recursive feature elimination using the 'rfe' function with (10-fold) cross-validation (CV) resampling and the training dataset. After recursive feature elimination, the model with the optimal subset size of variables for each algorithm was selected to predict the testing dataset. As validation, we also used 'rfe' (10-fold cross-validation) for the whole dataset. To split the dataset accordingly, we first generated the 10 folds using the 'groupKFold' function based on the indicated categories (samples and proteins). These folds were used as input in the 'folds' parameter in the 'rfeControl' function.

The algorithms tested were: Boosted Tree ('bstTree'), Random Forest ('rf'), Bagged Model ('bag'), Boosted Tree ('blackboost'), Lasso and Elastic-Net Regularized Generalized Linear Model ('glmnet'), k-Nearest Neighbors ('kknn'), Cubist ('cubist'), and Linear Regression ('lm'). All algorithms were used via 'caret', and thus, the default parameters used by 'caret' were utilized.

2.2.5 Performance based on GO terms

We selected the cardiac dataset, and subselected one sample as testing dataset, while the model training was proceeded with the rest of samples using the 10 features shown in the results. After the training, we predicted the testing dataset with the resulting random forest model, and combined the predictions with the testing observations. We then extracted the GO terms associated for each protein in the testing dataset, which we also combined with the observations and predictions. We discarded GO terms that were categorized in less than 10 proteins. We evaluated the R^2 metrics for each of the groups of proteins associated to each GO term. We ranked the GO term groups from best to worst performing based on R^2. All the code used can be located in the following script on GitHub: 'script/go_terms_analysis/rsquared_on_different_go_terms.R'.

2.2.6 Imputation: a potential use of the random forest model

We also selected the cardiac dataset, but in this case including all proteomics missing values. We subselected all Untreated (UNTR) samples. The training dataset only contained observations with quantified proteomics values, and the 10 features mentioned in the results. We used the random forest algorithm for the training of the model. We then predicted the missing proteomics values using the newly

trained model. We combined the results with the observed data, and sampled proteins with different proportions of missing data. All code run can be found in the following script on GitHub: 'script/imputation/imputing cardiac values.R'.

3 RESULTS

To assess the ability of the regression ML algorithm to estimate the level of proteins, we produced a dataset that presented the added value of having transcriptomics (both ribo-depleted and small RNA libraries) and proteomics (LC/MS), all generated from the exact same sample batches to maximize the interpretability of the interactions. This dataset was composed of a total of 115 *in vitro* samples (61 cardiac and 54 hepatic). The processing of all these samples (Methods) characterized an amount of expressed biological entities summarized in Table 1. The total number of expressed biological entities was 48 266 and 48 715 for the hepatic and cardiac tissues respectively.

Table 1: Summary table of all quantified biological entities. Total refers to all possible entities to be identified. Expressed (N) refers to the number of entities that were quantified in at least 1 sample. Constitutive (N) refers to the number of entities that were quantified in all samples. Expressed (%) and Constitutive (%) refer to the percentage of (constitutively) expressed entities based on the total number of entities. Constitutive (% Expressed) refers to the percentage of constitutively expressed entities based on the number of expressed entities.

	Tissue	Total	Expressed (N)	Constitutive (N)	Expressed (%)	Constitutive (%)	Constitutive (% Expressed)
Proteomics	Hepatic	1806	1806	283	100.00%	15.67%	15.67%
Proteomics	Cardiac	2217	2217	247	100.00%	11.14%	11.14%
Linear transcripts	Hepatic	211939	135655	894	64.01%	0.42%	0.66%
Linear transcripts	Cardiac	211939	136860	933	64.58%	0.44%	0.68%
MicroRNAs	Hepatic	2744	1561	280	56.89%	10.20%	17.94%

MicroRNAs	Cardiac	2744	1510	250	55.03%	9.11%	16.56%
Circularized transcripts	Hepatic	140317	95106	151	67.78%	0.11%	0.16%
Circularized transcripts	Cardiac	140317	100416	156	71.56%	0.11%	0.16%

To assess the possibility to predict protein expression levels for all genes using ML algorithm, we needed to assemble a list of features, either parametric or categorical. From all the table 1 data, we extracted 12 features focused on the expression level of linear transcripts, 24 features on miRNA expression, and 12 features on circular RNA expression. We added 36 features on transcript characteristics (strand, transcript length, etc.), 48 features on protein characteristics (Protein Mass, Protein Length, etc.), 12 features on MTI (miRNA target interaction) scores, a feature on RNA-Sequencing depth, 6 features on circular scores (number of miRNA binding site per circular RNA), 12 features on circular RNA expression, and 45 features on protein stability. This led to a total of 196 features on the raw dataset. Even so, some of those features might be deemed irrelevant due to their multiplicity and inherent structure. Those features might affect machine learning processes, depending on the algorithms' inherent functionality, by decreasing their accuracy[30]. To avoid their inclusion, we applied several preprocessing filters that removed non-informative features, which are described below.

3.1 Zero- and Near Zero-Variance Variables

Some predictors can have a unique value for all observations (Species: Human), which can make models unstable or decrease their fitness. Those features can be named as Zero-Variance (ZV) variables, and they are generally removed. Similarly, Near Zero-Variance (NZV) variables refer to features that present a value in an overwhelming majority of observations (i.e., genes coded in the nucleic genome vs genes coded in the mitochondrial DNA (Table 2)). These features are generally not helpful in a cost/benefit ratio, as the underrepresented values might have an artificially bigger impact, and these values may not even appear in the subpopulations generated by sub-sampling strategies, generating a ZV variable.

Table 2: Examples of Zero- and Near Zero-Variance variables. The 'Organism' variable contains a single unique value ('Human'), thus this value has no predicting value. The 'Gene encoded by' variable contains 2 possible values, of which 'Nucleus' represents more than 99% of all observations. Even though this variable does indeed have more than a single value, the frequency of its values renders it non-informative.

Protein ID	Organism	Gene encoded by
A – Sample 1	Human	Nucleus
B – Sample 1	Human	Nucleus
C – Sample 1	Human	Nucleus
D – Sample 2	Human	Nucleus
E – Sample 2	Human	Nucleus
F – Sample 2	Human	Nucleus
G – Sample 3	Human	Mitochondrion

Due to both ZV and NZV filters, 44 features were removed from the dataset. Only a few were labeled as ZV, examples of which were 'circ_min' (minimum circular expression) and its log2 transformed version 'circ_min_log2'. Some categories of variables were frequently labeled as NZV: almost all features related to miRNA scores; all maximum, median, and minimum miRNA expressions (non-transformed, log2-transformed, stringent, and all scores); some related to circular scores and some related to circular expression (Supplementary Table 1).

Table 3: Examples of ZV and NZV features with their respective frequency ratios and unique percentages. The metrics for all NZ features were identical, as they only reported a single value (Inf: Infinite). For NZV values, they all presented a frequency ratio above 19 (95/5) and a percentage unique below 10.

Feature name	freqRatio	percentUnique	zeroVar	nzv
circ_min	Inf	0.004	TRUE	TRUE
circ_min_log2	Inf	0.004	TRUE	TRUE
Organism	Inf	0.004	TRUE	TRUE

strand_sd	839.433	0.008	FALSE	TRUE
transcript_length_sd	140.014	3.494	FALSE	TRUE
percentage_gene_gc_content_sd	515.167	0.107	FALSE	TRUE
cds_length_sd	64.757	2.098	FALSE	TRUE
noncds_length_sd	112.344	3.427	FALSE	TRUE
proportion_noncds_length_sd	336.033	3.693	FALSE	TRUE
Gene.encoded.by	1198.048	0.008	FALSE	TRUE

3.2 IDENTIFICATION OF CORRELATED VARIABLES

Having correlated predictors is generally uninformative and sometimes detrimental to build models. For this reason, we removed features that presented a correlation above 0.75. For each pair of correlated features, the feature labeled as 'highly correlated' was the one that presented a higher correlation with the rest of the variables. Having the target inside the dataset would imply that the features that showed a higher correlation with the target would get removed. To avoid this, we removed the target from the dataset before filtering the highly correlated variables. In total, 93 features were removed due to high correlation (Supplementary Table 2). As expected, the abundances of most amino acids were highly correlated to each other, and to the protein mass and length. The same results were not true for the proportion of each amino acid, as they more accurately represent their presence independently of the protein's size. More surprisingly, among all non-filtered features, we observed all possible grouping systems (minimum, mean, median, maximum, standard deviation, and sum of the values they represented), with no clear predominance for any of them, and thus none appeared to present a tendency to be the most informative (i.e., the one with the lowest overall correlation with all features).

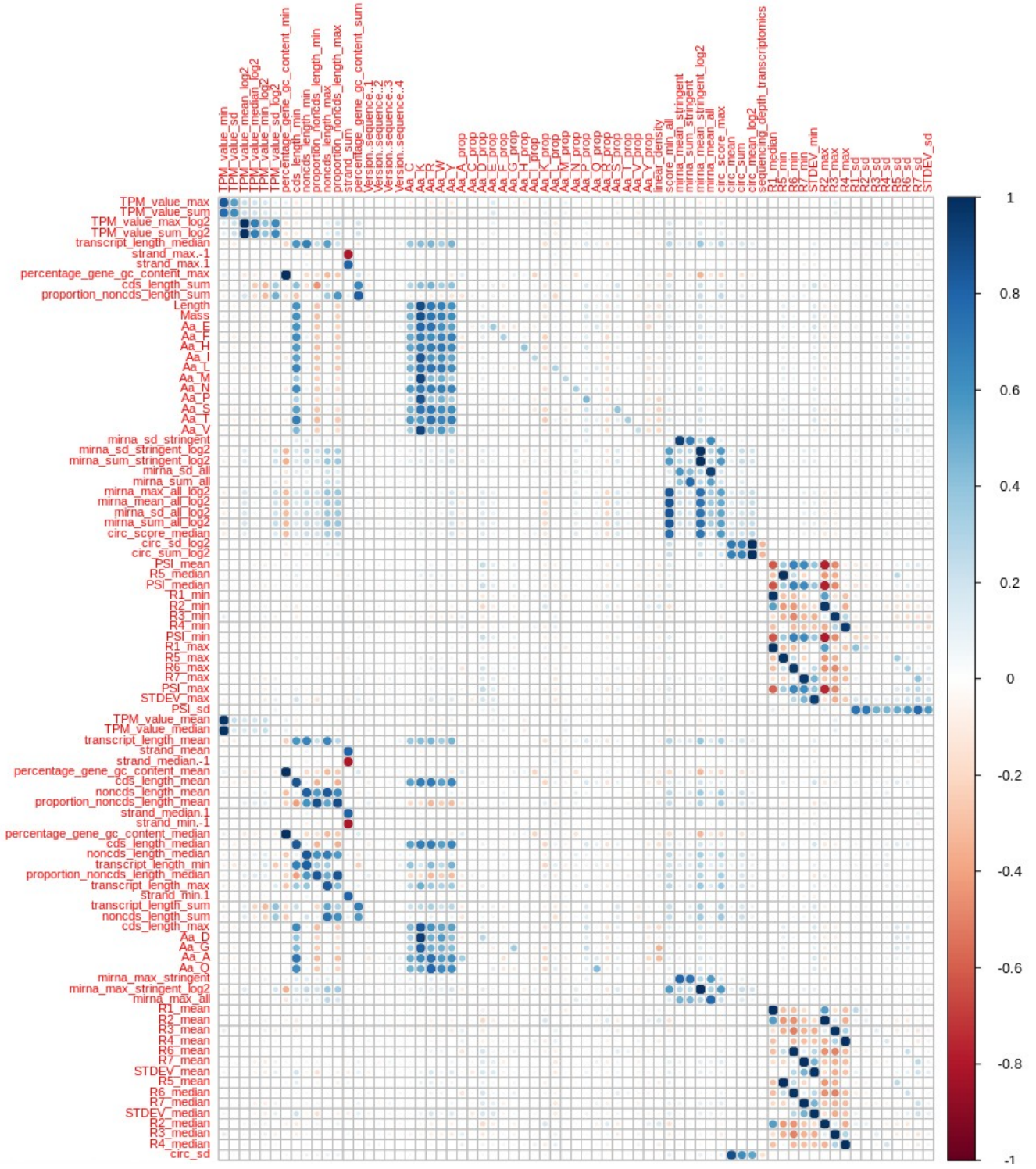
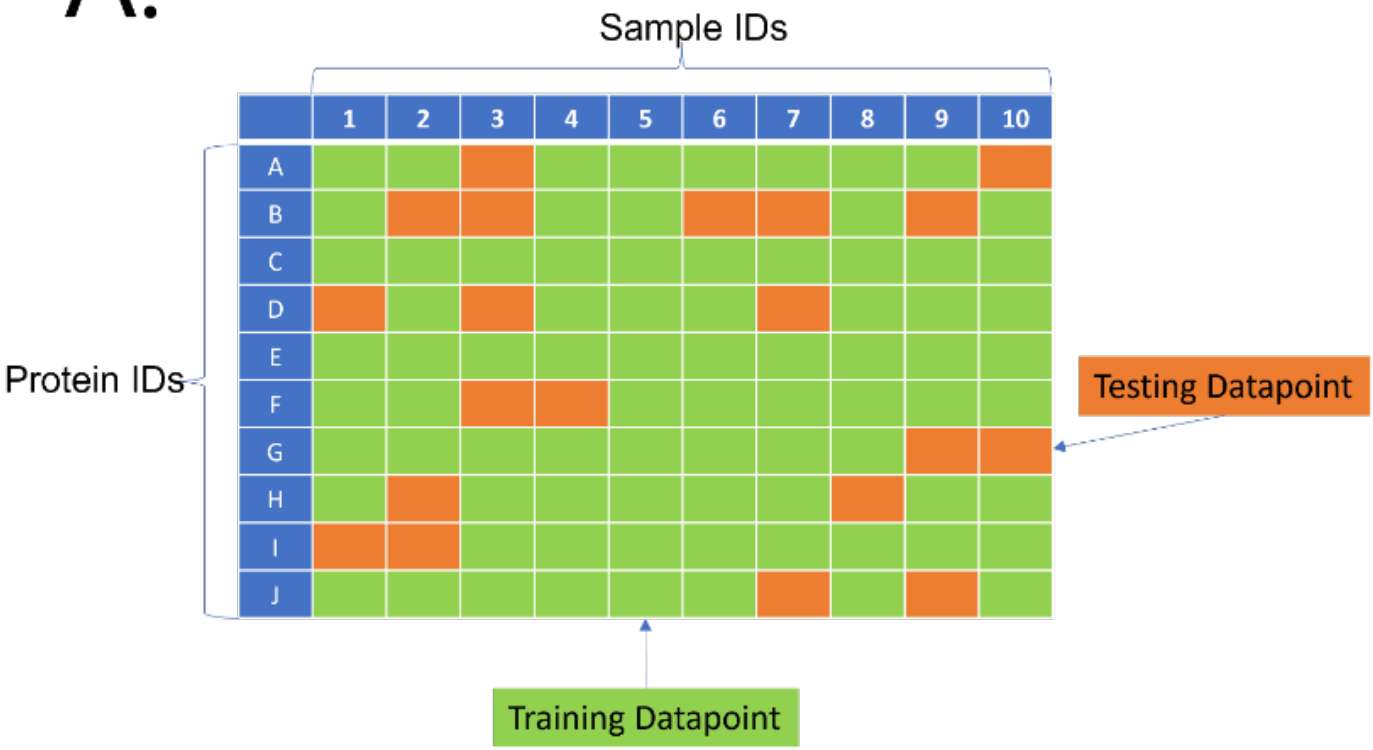


Figure 1: Correlation plot between kept features (horizontal axis) and filtered features (vertical axis). The scale unit on the right side of the figure indicates the correlation values between the features shown based on a range of colors: from dark red (extreme negative correlation) to dark blue (extreme positive correlation), where lighter colors represent a lower absolute correlation value.

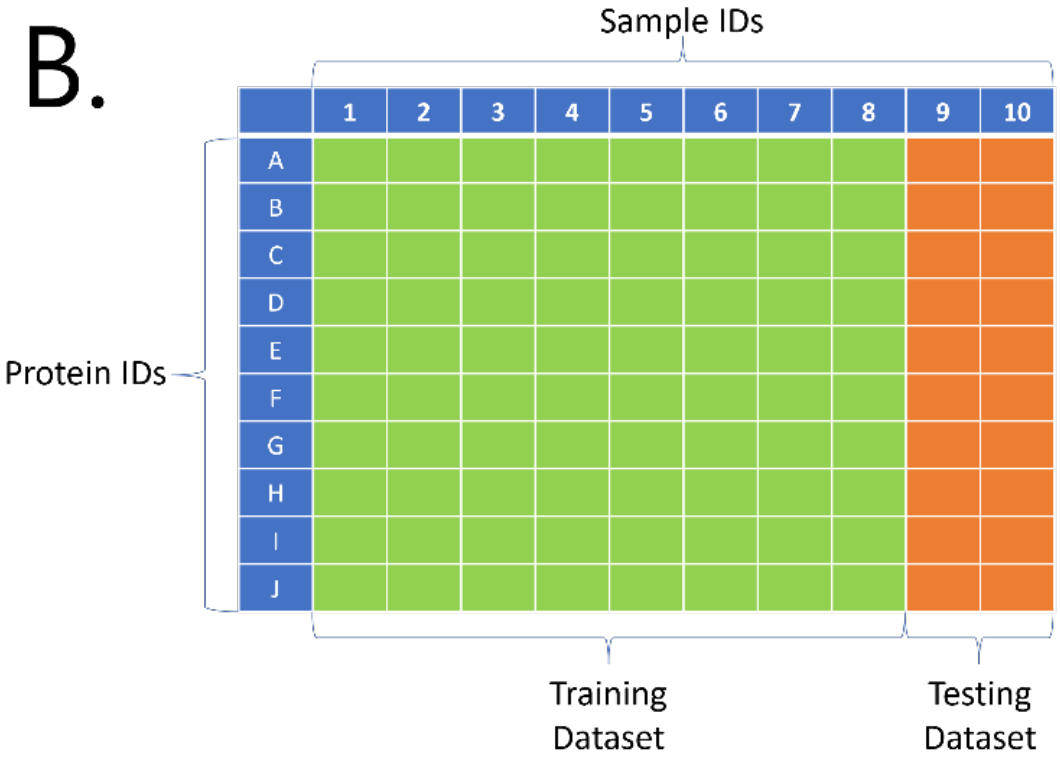
3.3 SPLITTING STRATEGIES

In both hepatic and cardiac datasets, the observations were part of two distinct groups: proteins and samples. Having a random split of our data to form both training and testing datasets would not have enabled us to elucidate the actual accuracy of the model. In a random splitting strategy, the training dataset was highly probable to include most proteins and samples in their observations, rendering the data split futile. Instead, we split (and trained) our models separately in two manners: splitting by sample and splitting by protein (Figure 2). This strategy was applied in the hepatic dataset for both training and testing, and in the cardiac dataset for validation.

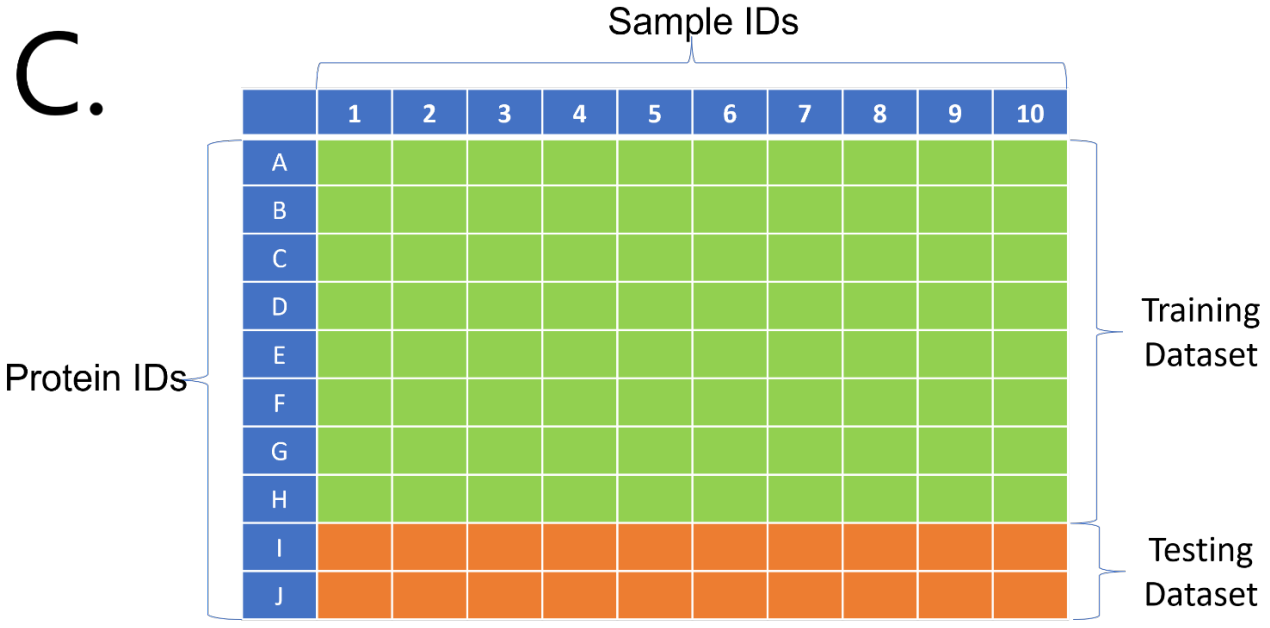
A.


Sample IDs
1
2
3
4
5
6
7
8
9
10
A
B
C
D
E
F
G
H
I
J
Protein IDs
Testing Datapoint
Training Datapoint

288



289



290

291 *Figure 2: Splitting strategies. For all splitting strategies, 80% of the data is used to train the models*
292 *(training dataset), while the other 20% is used for testing the trained models (testing dataset). A. Ran-*
293 *dom splitting strategy, where the algorithm is trained and tested with observations from all proteins*
294 *and samples. B. Sample-splitting strategy: the trained models are tested with 20% of the samples. C.*
295 *Protein-splitting strategy: the trained models are tested with 20% of the proteins.*

3.4 Model training and testing using hepatic sample-split data

When splitting by sample, 80% of hepatic samples were used as the training dataset, while the other 20% was used as the training dataset. For every algorithm, the training dataset was inputted through RFE (cross-validated with 10-fold). Out of all models trained with different subsets of features, the one with the best accuracy was used for the testing step (Figure 3). In terms of root-mean-square error (RMSE, Figure 3A), both k-Nearest Neighbors ('kknn') and Random Forest ('rf') showed the highest accuracies (~ 1.25), the latter having a bigger deviation between training and testing RMSE values. To evaluate these results in a more standard and informative manner, we also analyzed the R squared metrics (Figure 3B). In this figure, we observed that rf and kknn also showed the best performance (R^2 close to 0.7), showing rf better performance in this case.

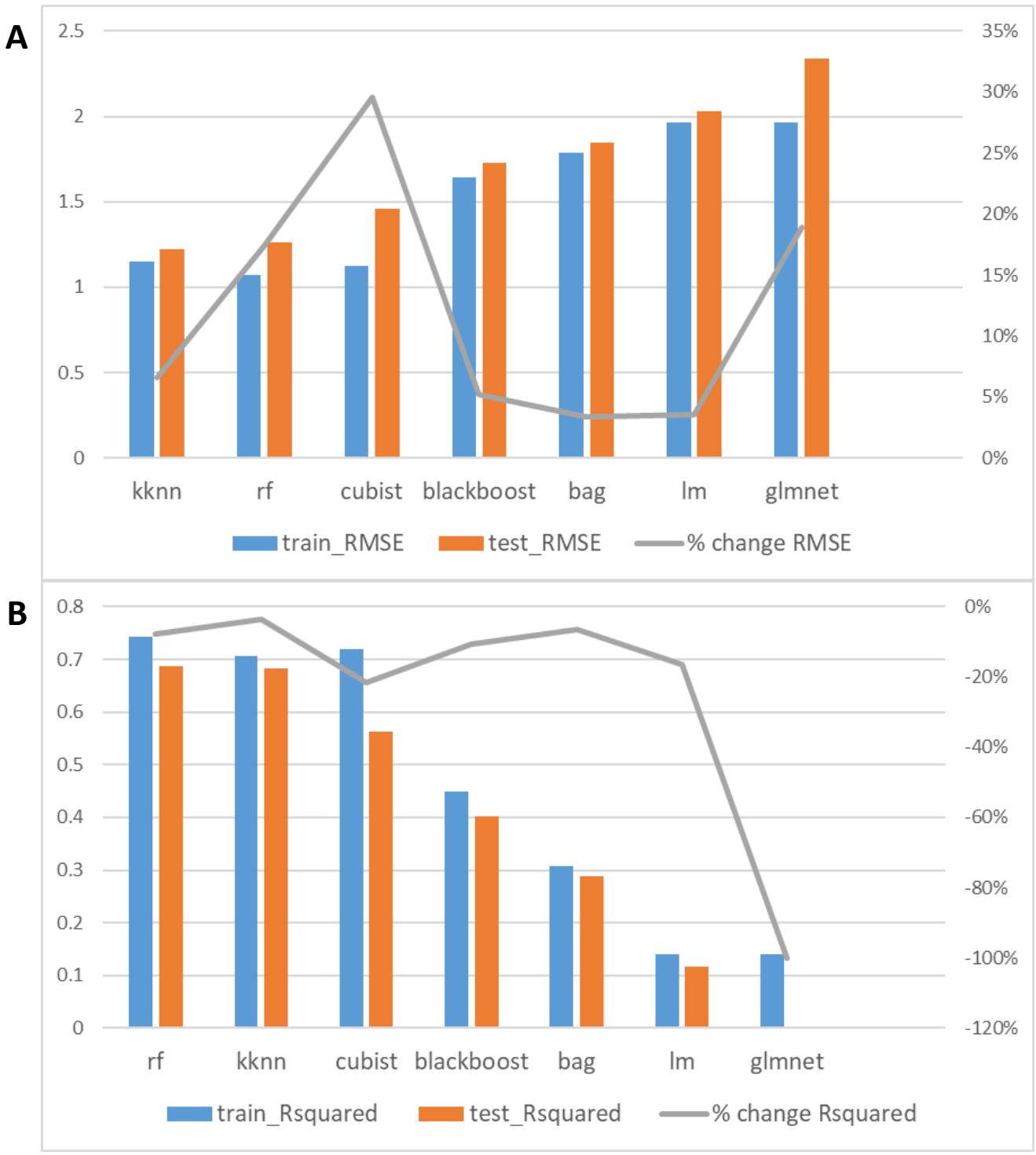


Figure 3: Accuracy results when splitting by sample. A: The blue bar refers to the RMSE value (left vertical axis) after training the model with 80% of the samples, and the orange bar refers to the RMSE value after testing the model with the other 20% of the samples. The gray line refers to the percentual change of RMSE (right vertical axis) between training and testing. B: The blue bar refers to the R^2 value (left vertical axis) after training the model with 80% of the samples, and the orange bar refers to

the R^2 value after testing the model with the other 20% of the samples. The gray line refers to the percentual change of R^2 (right vertical axis) between training and testing.

After validating the aforementioned results by using RFE (10-fold cross-validation) for the whole dataset (Supplementary Figures 1 and 2), we selected random forest ('rf') as the best performing model when splitting by sample. The optimal subset size of features was 51 features, but after close examination of the RFE results (Supplementary Figure 3), we determined that subset sizes above 10 features had a minimal impact on RMSE. The 10 features were selected based on the ranking of feature importance reported by the RFE analysis (Figure 4).

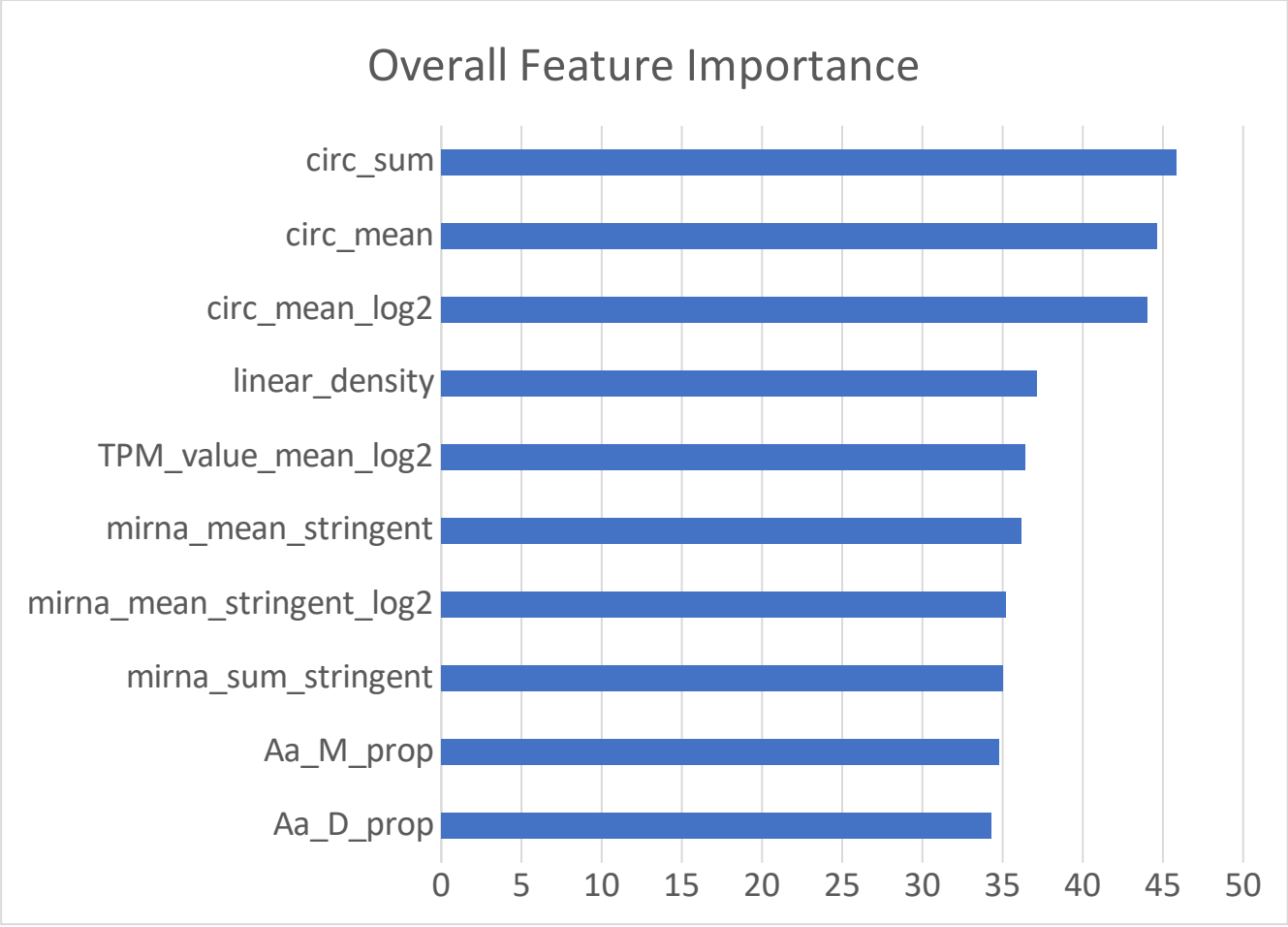


Figure 4: Top 10 features based on Overall importance by RFE when using the rf algorithm. These values represent how important (on average) each feature is to the model, and thus which are the main features used by the model to predict new proteomics values.

3.5 MODEL TRAINING AND TESTING USING HEPATIC PROTEIN-SPLIT DATA

Similar to the splitting by sample strategy, a fifth of all proteins were split to be used as the testing dataset, while the other 4 fifths were used as the training dataset. RFE (10-fold CV) was also performed with similar optimal results as in the training dataset of the sample-splitting strategy (Figure 5). In this case, the best RMSEs in the testing dataset include 'bstTree' and 'rf' (≈2), which almost doubled the error shown when splitting by sample (Figure 5A). To understand how relevant this error increase was, we also evaluated the R-squared values of those values (Figure 5B). We observed that a systematic gap existed between the training and testing steps, leading to minimal R-squared values ($R^2 = 0.15$ for rf).

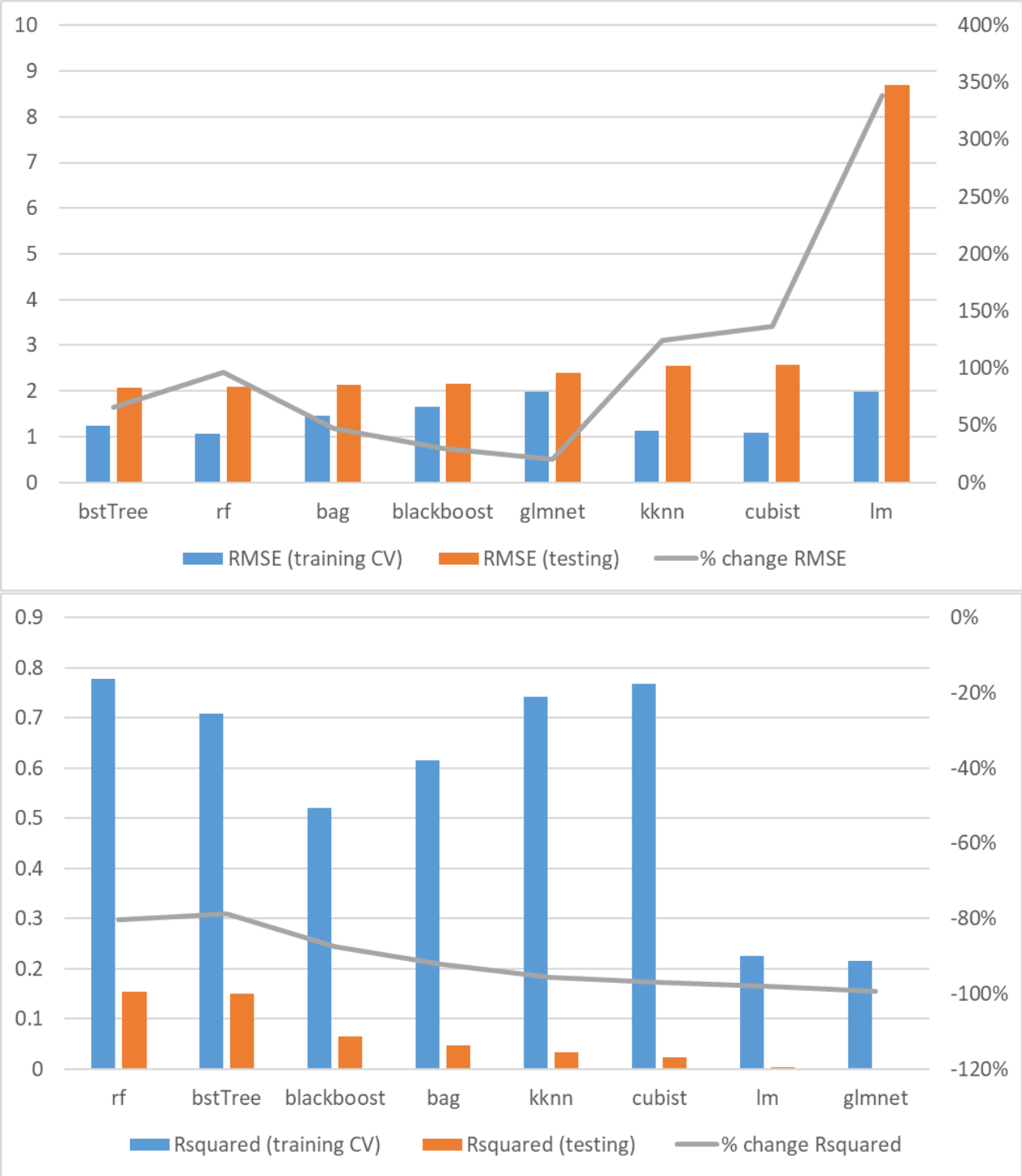


Figure 5: Accuracy results when splitting by protein. A: The blue bar refers to the RMSE value (left vertical axis) after training the model with 80% of the proteins, and the orange bar refers to the RMSE value after testing the model with the other 20% of the proteins. The gray line refers to the percentual change of RMSE (right vertical axis) between training and testing. B: The blue bar refers to the R^2 value (left vertical axis) after training the model with 80% of the proteins, and the orange bar refers to

the R^2 value after testing the model with the other 20% of the proteins. The gray line refers to the percentual change of R^2 (right vertical axis) between training and testing.

For all the results shown above (Figures 3 and 4), we also validated the results using RFE with the whole dataset (no training-testing split), where the folds or splits in the cross-validation step (10-fold) contained exclusively a set of proteins (Supplementary Figures 4 and 5).

3.6 RANDOM FOREST MODEL VALIDATION WITH A CARDIAC SAMPLE-SPLIT DATA

Random forest being the best performing model, we decided to validate its accuracy to predict new samples using a cardiac dataset, which was built in the same manner as the hepatic one. The validation included using the same algorithm (rf) with the same top 10 features (Figure 4), and training and validating it with the cardiac data (27602 observations). The resampling was performed via Cross-Validation (10-fold). Using the cardiac data and the specified model, we validated that the accuracy remained robust across different cell types (RMSE = 1.04, R^2 = 0.75; Supplementary Figure 6).

The only remarkable difference was the feature importance ranking given by the RFE in the hepatic data (Figure 4) compared to the feature importance ranking given by the model itself with the cardiac data (Figure 6). In the latter, linear_density is given the utmost importance, and the importance of the three RNA subtypes relate to how close they are to the protein level: mRNA level, followed by miRNA levels, and finally circRNA levels.

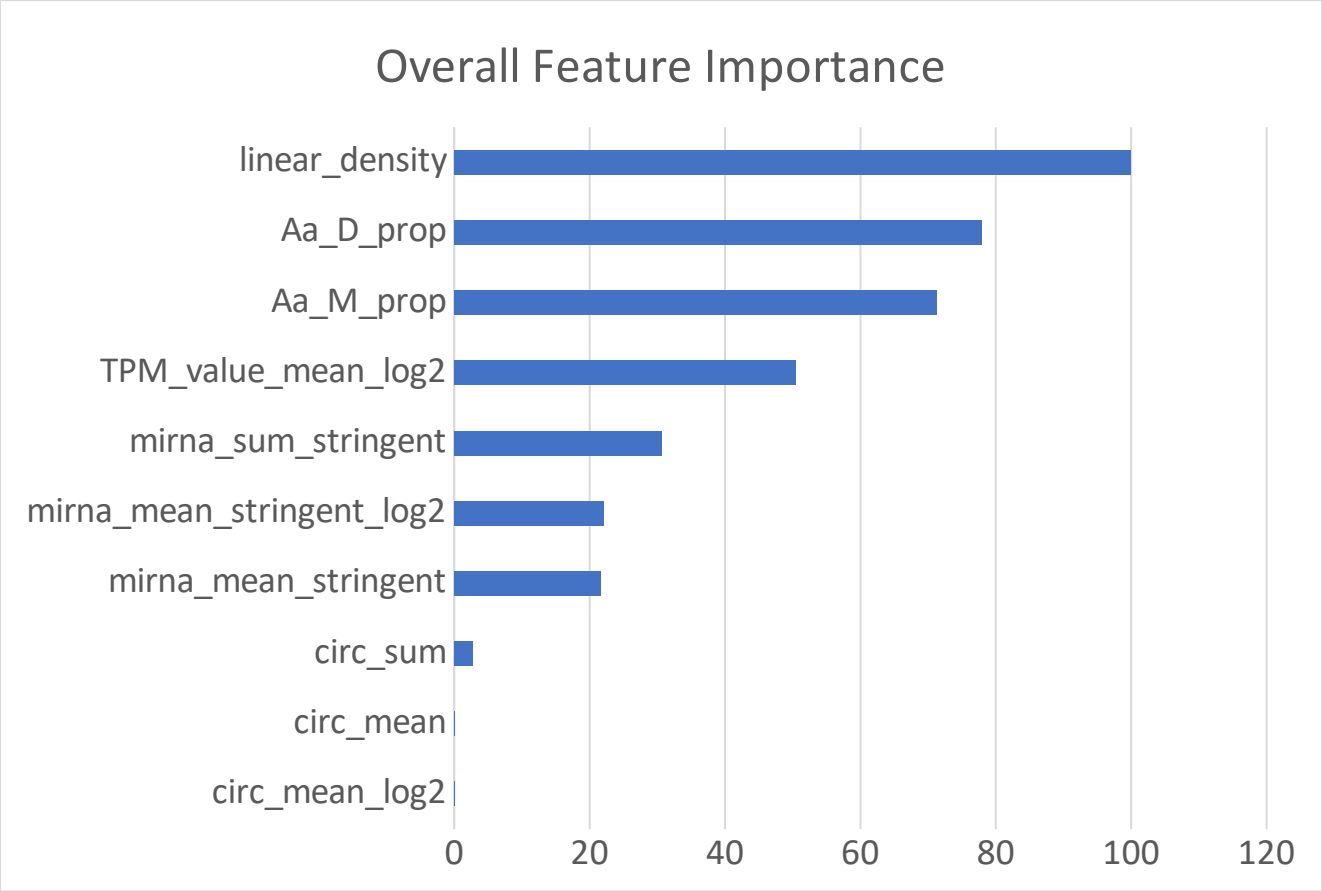


Figure 6: Feature Overall Importance for the rf algorithm when trained in Cardiac data. These values represent how important (on average) each feature is to the model, and thus which are the main features used by the model to predict new proteomics values.

Therefore, the RMSE and R-squared metrics for both cardiac and hepatic models showed that building a random forest model using the aforementioned features allowed to predict with high accuracy full proteomics' samples. Comparing the testing results between sample- and protein-splitting, we observed that the high accuracy was especially due to the prediction of proteins that have already been trained on. Observing the feature importance ranking (Figure 6), we could observe that different biological entities presented a different relevance to the model's accuracy, thus missing some variables will have a minimal effect on the decided outcome.

3.7 Performance based on GO terms

Even though we obtained good substantial results for the prediction of proteomics values at a sample level, these results were an overall representation of all proteomics values, and thus did not inform which protein groups

would be better or worse predicted by our model. For this reason, we decided to stratify the predictions based on GO terms, and then evaluate their R^2 metrics when compared to their counterpart observations. The overall metric for the testing data/sample in this experiment was $R^2 \sim 0.82$. What we observed (Figure 7) is that there were considerable differences in R^2 depending on the GO term the proteins were associated to. While the 6 best-performing GO terms (inflammatory response, magnesium ion binding, mitochondrial nucleoid, unfolded protein binding, ATPase, and negative regulation of cell growth) had near perfect results ($R^2 > 0.9$), the worst performing ones (ligase activity, polysomal ribosome, small ribosomal unit, stress fiber, cell migration, and proteasome complex) showed metrics half the performance shown in the overall results ($R^2 \leq 0.4$).

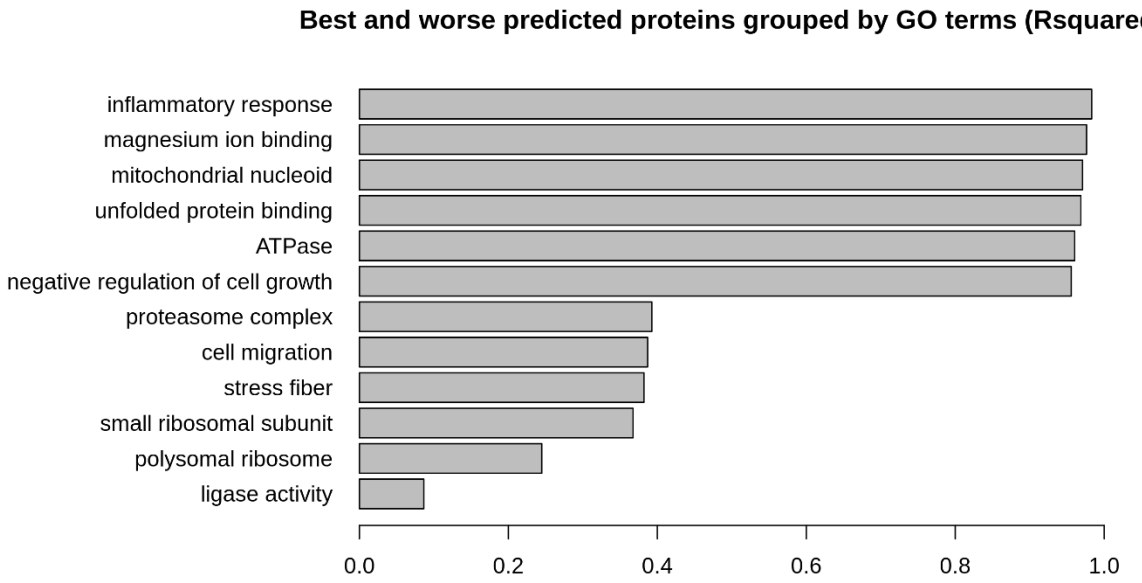


Figure 7: R^2 results categorized in GO terms. The X axis represents the R^2 values, while the bar labels represent the GO terms. Only the 6 best and 6 worst GO terms are depicted.

3.8 IMPUTATION: A POTENTIAL USE OF THE RANDOM FOREST MODEL

As the model showed a promising accuracy for predicting whole replicate samples, we hypothesized that the model could also be used for imputation of missing values for proteins that were at least present in one of the samples of the training data. To showcase a possible example, we trained a random forest model with all the Untreated samples (UNTR) and the corresponding 10 features. The example (Table 4) showed that the proteomics values imputed fitted the range of quantification observed in the quantified values of the same protein, while differing from each other from sample to sample. We also observed that in these samples, values tend to be missing simoultaneously for samples taken at the same time.

Table 4: Imputation of Proteomics Cardiac samples. Every row is identified with a UniProt ID, and represents a protein quantified in at least one of the untreated samples of the cardiac dataset. Each column represents each Untreated (UNTR) sample from the cardiac dataset. On the column names, the first number represents the hour at which the sample was taken (2h, 8h, etc.), while the second identifies the replicate number (002_1 was the first replicate sample taken after 2 hours). The proteins (rows) are sorted by proportion of missing data in a increasing order. Values with a dark green background were quantified by proteomics. Values with a light green background were imputed/predicted by the random forest model.

	002_1	002_2	008_1	008_2	008_3	024_1	024_2	024_3	072_1	072_2	072_3	168_3	240_1	240_2	336_1	336_2	336_3
P22695	15.70	16.09	14.58	14.23	13.91	14.15	14.26	13.83	13.83	13.95	13.89	14.90	14.72	14.35	16.75	16.83	16.94
P51553	12.41	11.03	13.87	13.99	13.76	14.70	14.67	14.82	13.55	13.64	13.34	13.26	13.28	13.45	14.37	14.15	14.27
P62910	15.52	14.49	14.78	14.45	14.68	14.75	14.87	15.04	13.76	13.98	13.67	14.61	15.52	13.81	14.57	14.85	14.49
Q15185	14.43	13.89	14.39	14.28	14.16	14.01	13.99	14.14	14.33	14.17	14.04	11.92	13.48	13.70	13.79	13.95	13.99
P45974	14.66	14.51	11.04	11.26	11.43	13.29	13.24	12.74	14.10	13.16	13.73	13.85	13.07	12.82	13.38	13.65	14.04
P16070	12.72	12.89	13.47	13.55	13.79	13.45	13.22	13.21	13.25	13.44	13.31	12.65	12.92	13.17	12.96	13.28	12.97
P54136	10.93	10.78	11.86	11.91	11.76	12.43	12.57	12.18	13.21	13.13	13.26	12.56	13.23	12.06	11.39	9.76	11.81
O43681	13.08	13.37	13.37	13.77	13.48	13.40	13.31	13.29	12.63	13.06	12.06	12.04	11.83	12.10	11.84	12.14	12.01
P50440	13.22	12.88	13.02	13.05	13.45	12.35	12.45	13.93	14.03	11.90	13.19	13.17	13.16	13.09	13.19	13.18	13.27
P47897	12.93	12.56	13.03	12.90	12.79	12.62	12.62	12.62	12.62	12.62	12.62	12.62	12.62	12.62	12.07	12.29	12.53
Q14141	14.81	14.44	13.49	14.06	13.70	13.48	14.25	13.89	13.83	13.74	14.10	13.41	14.06	14.49	14.17	13.63	13.73
P40763	12.12	11.82	12.29	11.98	11.54	11.47	11.53	11.50	11.13	11.88	12.03	12.18	11.63	11.65	12.24	11.40	11.37
P04844	12.23	14.51	13.74	13.74	13.78	14.13	13.63	13.77	13.94	13.92	13.71	13.75	13.88	13.89	14.02	13.66	14.42
Q9Y6E2	11.52	11.90	12.53	12.32	12.10	12.37	12.22	12.18	12.57	11.89	11.71	12.21	11.04	10.25	12.02	12.17	12.10
P13796	10.98	11.25	11.55	11.37	11.30	11.32	10.97	11.24	10.97	11.25	10.82	11.26	11.25	11.27	11.95	9.84	9.83
Q96KP4	12.09	10.96	11.82	11.97	11.88	11.82	12.05	12.11	11.94	12.51	11.89	12.23	11.95	12.46	12.41	12.55	12.53
Q96RQ3	13.00	12.79	12.98	12.79	12.59	12.98	12.93	12.86	12.59	12.79	12.79	12.24	12.59	12.80	12.93	12.59	12.86

4 DISCUSSION

We wanted to build a machine learning model that tightened the gap between transcriptomics and proteomics, using the former as a predictor of the latter. The results indicate that a random forest model, by using only 10 features, can predict with good accuracy ($R^2 = 0.74$) proteomics values from samples in similar circumstances to the ones where it has been trained on. However, predicting protein expression by training the model on other proteins was highly inefficient ($R^2 = 0.15$).

Interestingly, 7 out of the 10 features used by the model were related to RNA expression (Fisher's Exact Test for Count Data, p-value = 0.0027). Out of these 7, the most important (as expected) was mRNA expression, which is directly linked to translation, and thus, to protein expression. Followed in feature importance came 3 features related to miRNA expression, which is known to inhibit translation to a vast number of coding transcripts. The least important features related to the 3 RNA subtypes referred to circular RNA expression. Circular RNAs have been hypothesized to work as miRNA sponges, and so even though they are involved in post-transcriptional regulation, they have a more indirect effect. It is postulated that most circular RNAs are by-products of faulty splicing[31], and thus

their regulation might just be mainly due to the regulation of their host gene. Even so, their consistent expression would still allow them to have an impact on post-transcriptional regulation.

Linear density (mass of a protein divided by its length) and the proportions of both Aspartic Acid and Methionine were the most important features for the final random forest model. One hypothesis to explain such model behavior was that these three features (and especially linear density) helped to categorize observations protein-wise: an observation with similar values across the three top features could be likely categorized as a similar protein, and thus, also presenting a close expression value. This already made the model highly accurate when trained and tested with similar samples. The other features (related to the current transcript expression level) might have helped to succinctly tune the protein expression already observed in similar proteins during the training step. Another hypothesis, only relevant to linear density, was linked to the proteomics technology itself: linear density was directly linked to protein mass, which is used (along with charge) to identify and quantify protein in mass spectrometry; hence, its relevance as a feature. In addition, having linear density as one of the main features underlines the importance of the training data for our model. A random forest model can only predict values learned beforehand, thus we hypothesize that linear density helps the model to find the most similar protein when predicting. Thus, the use of this model should be to predict proteins that are already quantified in some of the samples, limiting the effect of potential false positives, and therefore also limiting the potential false biological significances created by false positives due to differences that only occur at the transcriptional level.

The observed divergence between the feature importance ranking in RFE and the validation model may be due to how RFE evaluates features while using cross-validation. At the beginning of the process, RFE built 10 different training-testing combinations (based on the 10 folds), and, based on the initial ranking of all the features in each of those combinations, features were removed from least to most important. Each feature was ranked based on the average of all the rankings performed during the feature elimination. In the validation model, instead, the feature importance ranking represented the concrete importance of each variable for that specific model and algorithm.

Considering the relatively high accuracy of the random forest model to impute protein expression from a reduced subset of features, we see an application of this proposed strategy to contribute to compensating for the lack of depth of proteomics. Indeed, since proteomics only allows the analysis of a subset of proteins per sample, with usually only a partial overlap between samples (even at the

replicate level), our model would be able to predict and fill those values, increasing the strength of the statistical analysis of such proteins across treatments.

However, as shown in the GO-term-performance results, the metrics are not uniform for all categories of proteins, and this should be taken into consideration when performing analysis with a specific focus on a certain protein category. This difference may be the result of three different causes: 1/ the correlation of protein abundance with their coding RNA levels may differ across GO categories, 2/ as different GO categories contained an unequal number of proteins, the size of a GO category was inversely proportional to the R^2 metric (a smaller random set of values has a higher chance of obtaining a high R^2, and vice versa), 3/ GO categories with stable protein abundances (and mRNA levels) performed better than otherwise.

An important detail to consider is that drastically different data is generated when utilizing different methods to quantify proteomics intensities: from values that correlate with absolute abundance based on the MS signal of histones (also referred to as the "proteomic ruler" approach[32]), going through intensities inferred based on the ratio of detected peptides (pertaining to each protein) between samples (MaxLFQ[33]), to isobaric proteomics data (TMT/iTRAQ); wherein changes in the peptide intensity from one sample has a ripple effect on the intensities from all the co-isolated samples[34]. In our study, the Hi3 label free method[35] was used to quantify protein intensities, hence values from absolute abundance methodologies are expected to perform similarly. Despite that, isobaric proteomics methods should not be entirely dismissed, as the range of values predicted by a random forest model is highly dependent on the range of the data the model is trained on. The compositional nature of isobaric proteomics experiments results in signals that are highly batch-dependent. Our predictions would not take the batch structure into account, and as a result, a correction would be required. Thus, the inability of random forest models to extrapolate does make them an appealing option for compositional data, but simoultaneously may be a limiting factor for absolute intensity values.

Based on the inefficient accuracy for all models tested in the protein-splitting strategy, we hypothesize that even though we tried to include as much information related to protein expression as possible (transcript expression, transcript properties, protein characteristics, and stability), predicting protein expression anew (without ever training the model with that protein's data) may have required of an even more complete (i.e. RNA binding proteins, long non-coding RNAs, transcript half-life, etc.) or different set of features. For example, a study by Barzine et al[26] showed improved results ($R^2 = 0.51$) extrapolating proteomics values while only using gene expression data, GO terms, and UniProt

keywords. Future research should focus on either including the last two features as features to the dataset, or improving their deep learning model by including our (or other) post-transcriptional features.

In conclusion, after developing different machine learning models to predict proteomics values out of transcriptomics ones, we have achieved to build a random forest model that can predict with significant accuracy the protein expression of a new sample. Building a random forest model with the selected features can thus be used to predict the missing data inherent in proteomics studies, independently of the cell's nature. The code used for the pre-processing of data and the model building process is available on Github (https://github.com/jochotecoa/ml_proteomics)[36].

5 ACKNOWLEDGEMENTS

The current research was funded by the European Union Seventh Framework Programme HeCaToS (FP7/2007-2013) under the [Grant Agreement No. 602156].

6 DATA AVAILABILITY

Data has been submitted to the BioStudies repository (https://www.ebi.ac.uk/biostudies/) and is available under the following accession numbers:

- Hepatic data: S-HECA33, S-HECA34, S-HECA47, S-HECA158, S-HECA457, S-HECA460, S-HECA463
- Cardiac data: S-HECA1, S-HECA9, S-HECA18, S-HECA139, S-HECA447, S-HECA449, S-HECA453

Supplementary data can be found in the online version of this publication.

7 BIBLIOGRAPHY

1 Zhang, Z., Wu, S., Stenoien, D. L. & Pasa-Tolic, L. High-throughput proteomics. *Annu Rev Anal Chem (Palo Alto Calif)* **7**, 427-454, doi:10.1146/annurev-anchem-071213-020216 (2014).

2 Specht, H. *et al.* Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biol* **22**, 50, doi:10.1186/s13059-021-02267-5 (2021).

3 Schoof, E. M. *et al.* Quantitative single-cell proteomics as a tool to characterize cellular hierarchies. *Nat Commun* **12**, 3341, doi:10.1038/s41467-021-23667-y (2021).
4 Cheung, T. K. *et al.* Defining the carrier proteome limit for single-cell proteomics. *Nat Methods* **18**, 76-83, doi:10.1038/s41592-020-01002-5 (2021).
5 Cagney, G. *et al.* Human tissue profiling with multidimensional protein identification technology. *J Proteome Res* **4**, 1757-1767, doi:10.1021/pr0500354 (2005).
6 Chen, G. *et al.* Discordant protein and mRNA expression in lung adenocarcinomas. *Molecular & cellular proteomics : MCP* **1**, 304-313, doi:10.1074/mcp.m200008-mcp200 (2002).
7 Lemée, J.-M. *et al.* Integration of transcriptome and proteome profiles in glioblastoma: looking for the missing link. *BMC molecular biology* **19**, 13, doi:10.1186/s12867-018-0115-6 (2018).
8 Rogers, S. *et al.* Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics* **24**, 2894-2900, doi:10.1093/bioinformatics/btn553 (2008).
9 Dhingra, V., Gupta, M., Andacht, T. & Fu, Z. F. New frontiers in proteomics research: a perspective. *Int J Pharm* **299**, 1-18, doi:10.1016/j.ijpharm.2005.04.010 (2005).
10 Belle, A., Tanay, A., Bitincka, L., Shamir, R. & O'Shea, E. K. Quantification of protein half-lives in the budding yeast proteome. *Proc Natl Acad Sci U S A* **103**, 13004-13009, doi:10.1073/pnas.0605420103 (2006).
11 Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350-355, doi:10.1038/nature02871 (2004).
12 Lim, L. P. *et al.* Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769-773, doi:10.1038/nature03315 (2005).
13 Lewis, B. P., Shih, I.-h., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of Mammalian MicroRNA Targets. *Cell* **115**, 787-798, doi:10.1016/S0092-8674(03)01018-3 (2003).
14 Baek, D. *et al.* The impact of microRNAs on protein output. *Nature* **455**, 64-71, doi:10.1038/nature07242 (2008).
15 Selbach, M. *et al.* Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58-63, doi:10.1038/nature07228 (2008).
16 Zaphiropoulos, P. G. Exon skipping and circular RNA formation in transcripts of the human cytochrome P-450 2C18 gene in epidermis and of the rat androgen binding protein gene in testis. *Mol Cell Biol* **17**, 2985-2993, doi:10.1128/MCB.17.6.2985 (1997).

17 Chen, L. L. & Yang, L. Regulation of circRNA biogenesis. *RNA Biol* **12**, 381-388, doi:10.1080/15476286.2015.1020271 (2015).

18 Jeck, W. R. *et al.* Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19**, 141-157, doi:10.1261/rna.035667.112 (2013).

19 Rong, D. *et al.* An emerging function of circRNA-miRNAs-mRNA axis in human diseases. *Oncotarget* **8**, doi:10.18632/oncotarget.19154 (2017).

20 Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333-338, doi:10.1038/nature11928 (2013).

21 Courel, M. *et al.* GC content shapes mRNA storage and decay in human cells. *Elife* **8**, doi:10.7554/eLife.49708 (2019).

22 Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* **11**, 2301-2319, doi:10.1038/nprot.2016.136 (2016).

23 Chen, A. T., Franks, A. & Slavov, N. DART-ID increases single-cell proteome coverage. *PLoS Comput Biol* **15**, e1007082, doi:10.1371/journal.pcbi.1007082 (2019).

24 Kalxdorf, M., Muller, T., Stegle, O. & Krijgsveld, J. IceR improves proteome coverage and data completeness in global and single-cell proteomics. *Nat Commun* **12**, 4787, doi:10.1038/s41467-021-25077-6 (2021).

25 Lim, M. Y., Paulo, J. A. & Gygi, S. P. Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model. *J Proteome Res* **18**, 4020-4026, doi:10.1021/acs.jproteome.9b00492 (2019).

26 Barzine, M. P. *et al.* Using Deep Learning to Extrapolate Protein Expression Measurements. *Proteomics* **20**, e2000009, doi:10.1002/pmic.202000009 (2020).

27 Mitchell, T. M. *Machine Learning*. (McGraw-Hill, 1997).

28 *CRAN - Package caret*, <https://cran.r-project.org/package=caret> (

29 Kuhn, M. *3 Pre-Processing | The caret Package*, <https://topepo.github.io/caret/pre-processing.html#zero--and-near-zero-variance-predictors> (2019).

30 Butcher, B. & Smith, B. J. Feature Engineering and Selection: A Practical Approach for Predictive Models. *The American Statistician* **74**, 308-309, doi:10.1080/00031305.2020.1790217 (2020).

31 Barrett, S. P. & Salzman, J. Circular RNAs: Analysis, expression and potential functions. *Development (Cambridge)* **143**, 1838-1847, doi:10.1242/dev.128074 (2016).

559 32 Wisniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A "proteomic ruler" for protein copy num-
560 ber and concentration estimation without spike-in standards. *Mol Cell Proteomics* **13**, 3497-
561 3506, doi:10.1074/mcp.M113.037309 (2014).
562 33 Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and
563 maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13**, 2513-2526,
564 doi:10.1074/mcp.M113.031591 (2014).
565 34 O'Brien, J. J. *et al.* Compositional Proteomics: Effects of Spatial Constraints on Protein Quan-
566 tification Utilizing Isobaric Tags. *J Proteome Res* **17**, 590-599, doi:10.1021/
567 acs.jproteome.7b00699 (2018).
568 35 Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P. & Geromanos, S. J. Absolute quantifica-
569 tion of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics* **5**, 144-
570 156, doi:10.1074/mcp.M500230-MCP200 (2006).
571 36 *GitHub - jochotecoa/ml_proteomics*, <https://github.com/jochotecoa/ml_proteomics> (
572

Chapter 5:

AutoRel: Machine Learning model for selecting differentially expressed genes

AutoRel: Machine Learning model for selecting differentially expressed genes.

Juan Ochoteco Asensio[1], Jelmer Faber[1], Twan van den Beucken[1], Marcha Verheijen[1], Florian Caiment[1]

[1]Department of Toxicogenomics, School of Oncology and Developmental Biology (GROW), Maastricht University, Maastricht, The Netherlands

(Manuscript under submission)

Abstract

In next-generation transcriptomics, differential expression analysis has established itself as one of the main strategies to evaluate the effects of two biological conditions on the gene expression of different biological samples. However, the current statistical workflows provide very lenient standardized filters for the selection of differentially expressed genes, leading to an important number of false positives. Authors usually include their arbitrary thresholds, leading to statistically significant genes with expression profiles that, if individually analyzed, would not be considered biological relevant to investigate further. In this study, we focused on the development of a machine learning model, called AutoRel, which encompasses not only the most common statistical evaluations but also all the intricacies that characterize biologically relevant changes based on manual observations. AutoRel classifies each assessed gene into "relevant", "irrelevant" or "dubious". We evaluated the model using both simulated datasets and biological interpretation of the selected genes. GitHub Repository at https://github.com/jochotecoa/AutoRel.git .

Introduction

Next-generation sequencing (or NGS) of RNA (RNA-seq) has brought the capacity to sequence and quantify virtually all RNA molecules within a given sample, being a great advancement from single or few molecule identification methods such as Northern Blot or even a limited set of transcripts in microarray technologies. RNA-seq is especially relevant for contrasting different biological conditions, where the effects at the molecular level can be analyzed in different conditions, such as cells during tumorigenesis or upon exposure to toxic compounds. For the evaluation of these differences, multiple tools have been developed since the dawn of transcriptomics, however, their usage can be mishandled without a proper understanding of the mechanisms behind it. This problem is especially relevant for the selection of differentially expressed genes (DEGs) based solely on statistical results. Using standard

statistical criteria, the selected DEGs between two biological conditions frequently display low read abundance, inconsistent expression among the replicates, or the presence of outliers. Establishing a reference and guideline to avoid such pitfalls is critical. We recently collaborated in the R-ODAF framework[1], which aimed to standardize omics data analyses for regulatory applications, specifically in the toxicogenomics discipline[2]. Selecting a list of genes affected by a compound exposure for regulatory decision-making requires a high level of stringency, as false interpretation could have drastic consequences.

The development of R-ODAF originated primarily from the necessity to implement novel filtering criteria based on the expression profiles of the selected DEGs. The selection process was performed by analyzing the significantly different genes and evaluating their biological relevance by considering the quantification values across all samples and groups. Metrics such as fold change and p-values[3-5] can be misleading (and lead to false positives) if not evaluated in their proper context due to common characteristics such as the presence of outliers or a low read quantification. Systems to prevent the inclusion of statistically significant genes that presented such properties (thus probably not biologically relevant) were set up in the R-ODAF pipeline. However, despite being designed to decrease the frequency of irrelevant genes from the DEG set, the R-ODAF is still based on an adjusted p-value threshold (FDR < 0.01, FDR: False Discovery Rate adjustment), which by definition creates an arbitrary cut-off. Moreover, applying a stringent pipeline essentially generates more false negatives, especially in the typical context of experiments performed in triplicates. A filtering-based method only includes the most frequent characteristics of genes that would be manually discarded, thus frequently requiring a manual selection afterward.

An alternative to the threshold-based selection of DEGs would be to apply a machine learning method trained on gene expression profiles considered “relevant” by experts (relevant defined as “worthy of further investigation” after an analysis of the quantification values and their distribution in each condition). Advancements in artificial intelligence allow practitioners to automatize labeling processes (classification) or predictions of numerical values (regression) through the use of enough data[6-9]. In classification tasks, data is required for the nascent model to differentiate between labels via their different features or characteristics. These methodologies can learn the synchronicities of the labels across numerous examples without the need for explicitly and manually specifying such differences. In addition, the selection of the variables will rely on how informative they are for the classification of each label, and thus by analyzing the predictive model one can also derive which features and to which degree each of them is important.

In this study, we developed and tested machine learning models to automate the biologist-supervised DEG selection process. The best performing model, called AutoRel, informed us of the importance of each of our previously designed norms, while simultaneously providing a novel system to select biologically relevant genes without the expertise needed to evaluate them or based purely on a few arbitrary thresholds. Due to the subjective nature of the notion of what is "biologically relevant", AutoRel outputs not only a "relevant" and "irrelevant" list of genes, but also a class named "dubious" for the gene profiles subject to debate during the training. Ultimately, the AutoRel selected list of genes (named "relevant" list) is compared to conventional methods.

Results

The transcriptomics dataset consisted of 60683 assessed genes in 42 samples: 21 APAP-treated samples and 21 control samples. From this original dataset, we built three datasets based on different comparisons based on the number of replicates: 21 versus 21 (all time points), 9 versus 9 (2h, 8h, 24h), and 3 versus 3 (2h). For this analysis, we considered all samples in each group (treatment or control) as replicates. For each of these datasets, the quantification barplots of individual genes were used to manually label hundreds of genes into two categories: 'irrelevant' or 'relevant', based on whether the gene expression profile was considered worthy of further investigation due to biological differences. Observations that did not suit either of those classes were classified as "dubious". An example of each class is displayed in Figure 1. After labeling genes for all three datasets, we assembled all labeled genes into a new dataset, identified as 'MixR'. Next, we derived features related to the expression levels that might help train a machine learning model to predict the categories of new genes based on the manually annotated data. In total, 160 features were obtained for all datasets (Supplementary Methods). Before the model training, each dataset was pre-processed to remove uninformative or redundant features. We trained and tested all datasets with 11 machine learning algorithms, whose results can be found below.

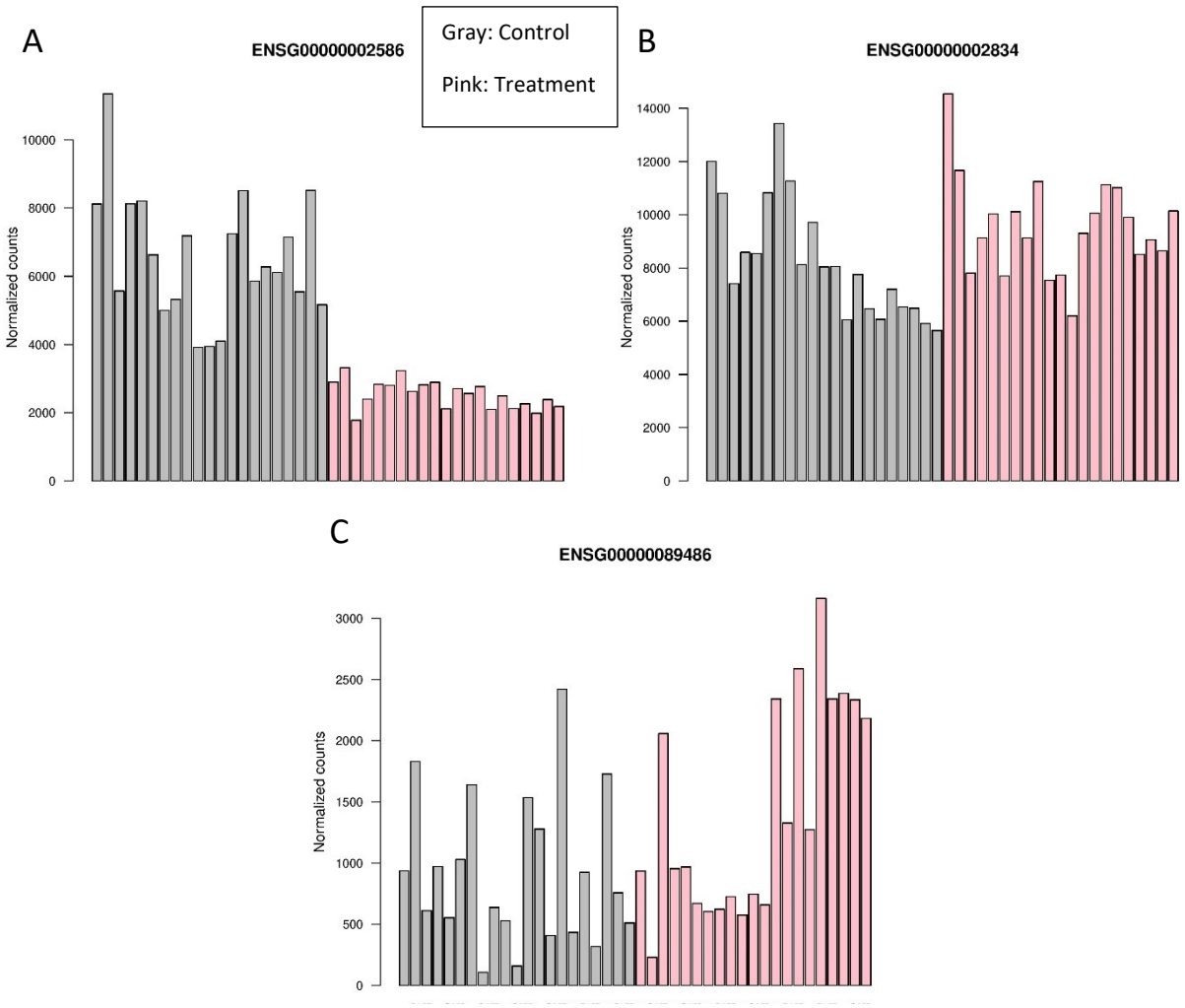


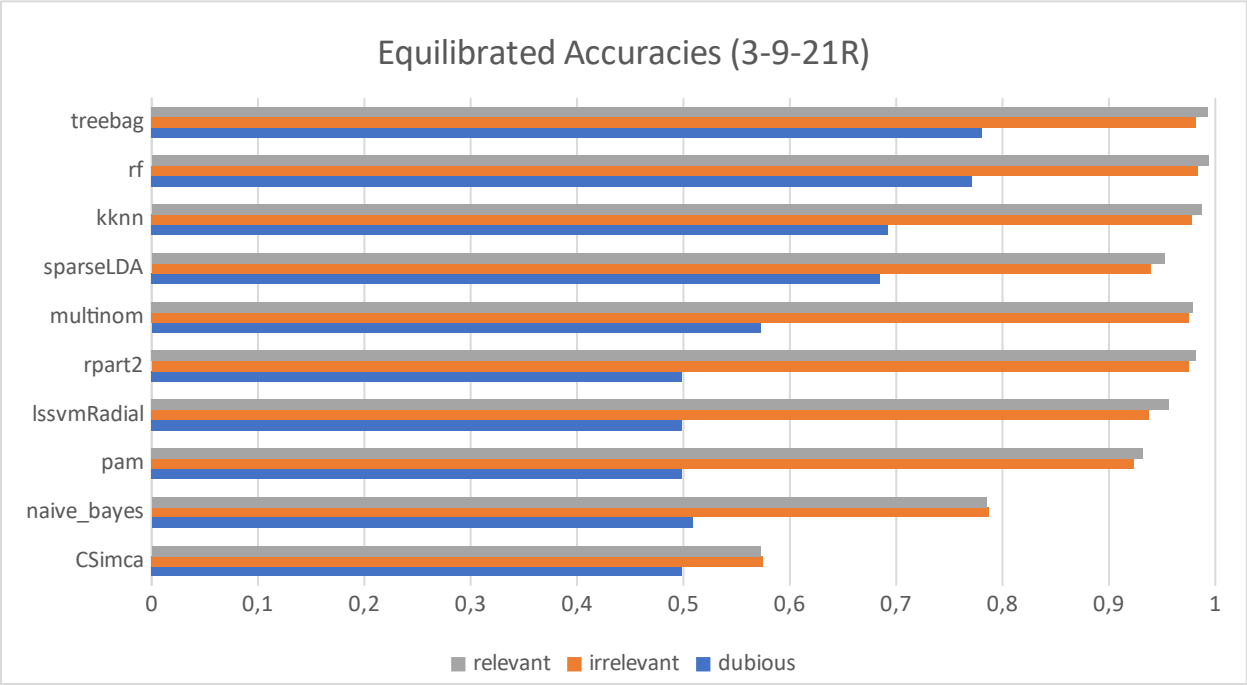
Figure 1: Quantified expression of genes from different classes. ***A:*** *Example of a gene with a relevant biological difference.* ***B****: Example of a gene with an irrelevant biological difference.* ***C****: Example of a gene with a dubious biological difference.*

Selecting the best-performing models in the 3R, 9R, 21R, and MixR datasets

To decide which was the best performing model, we did not utilize accuracy, which is easily biased towards the class with the highest number of observations. Instead, we averaged the sensitivity, specificity, PPV, and NPV for each class, which we named "equilibrated accuracy". We will use the MixR dataset as an example to show the process of the selection of the best model (in each dataset), but all datasets were pre-processed and trained in the same manner, including the same algorithms. The results for the models built in the other three datasets (3R, 9R, and 21R) can be found in the **Supplementary Results**.

The 'treebag' algorithm showed the best performance out of 10 tested algorithms in total (Figure 2A), closely followed by 'rf' (both algorithms also performed the best in the 3R and 21R datasets). Even with

the biggest training dataset, four algorithms ('rpart2', 'lssvmRadial', 'pam', and 'CSimca') were not able to predict 'dubious'-class observations.



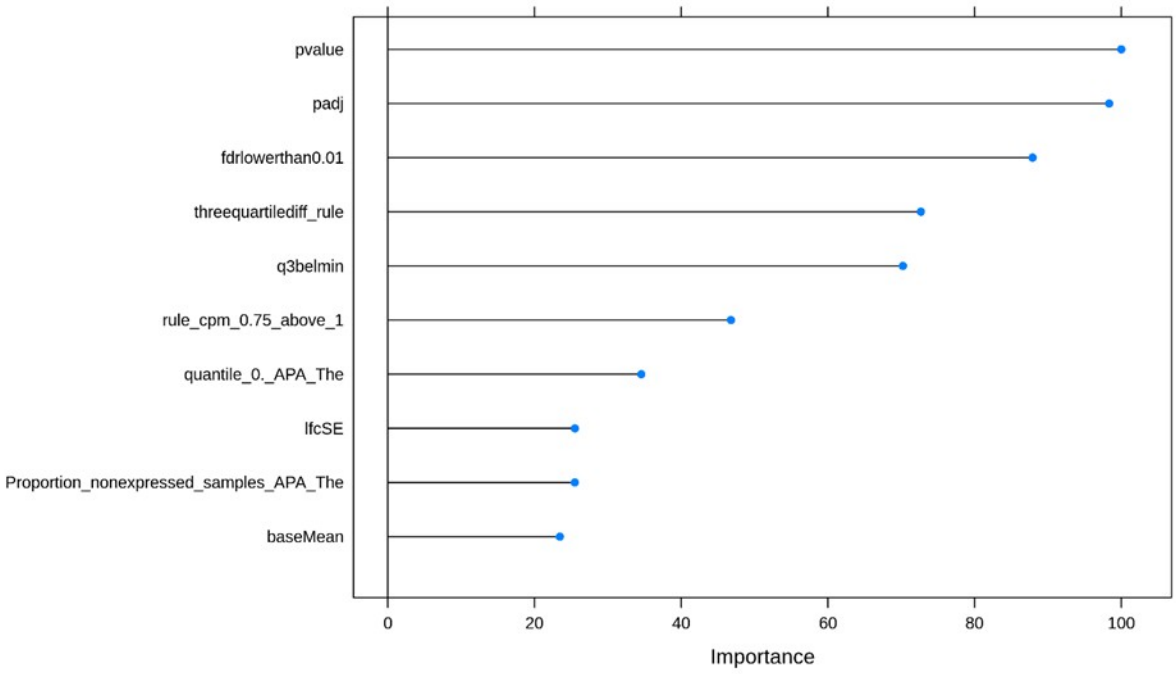


Figure 2: Selection and evaluation of the best-performing model using MixR as the training dataset*. A:* *Equilibrated accuracies for the models built with 11 machine learning algorithms using the MixR dataset. The models were sorted by the*

*average of equilibrated accuracies in descending order. **B**: Variable importance for the 10 most important features sorted in decreasing order for the 'treebag' model trained with the combined dataset.*

For the variable importance of the model (Figure 2B), 'pvalue' and 'padj' became again (as in the 'treebag_3R' model) the most important features. Similarly, variables featured as the most important for the best-performing models in differing training datasets were also found in the top 10 features for the combined dataset: 'fdrlowerthan0.01', 'threequartilediff_rule', 'q3belmin', 'rule_cpm_0.75_above_1', 'lfcSE', and 'baseMean'. Curiously, two additional features were also part of the top 10 that were not observed before: the minimum expression value for the treatment group ('quantile_0._APA_The'), and the proportion of non-expressed samples in the treatment group ('Proportion_nonexpressed_samples_APA_The').

Testing the best-performing models with all the testing datasets

Even though the four best-performing models were trained and tested with their characteristic datasets, we hypothesized that these models might present different predicting capabilities for datasets not similar to the ones they were trained for. To test the hypothesis, we tested each of the best-performing models with each of the corresponding test datasets and calculated the Cohen's Kappa Coefficient.

***Table 1: Kappa score for all best-performing models**. Each row represented one of the best-performing models, labeled by the name of the algorithm and its training dataset. Each column, in turn, represented the testing dataset utilized to obtain the Kappa score.*

		Testing datasets				
		3R	**9R**	**21R**	**MixR**	**Average score**
Best model in each dataset	**treebag (3R)**	0.970	0.894	0.485	0.849	0.800
	sparseLDA (9R)	0.614	0.918	0.316	0.609	0.614
	rf (21R)	0.538	0.919	0.915	0.658	0.757
	treebag (MixR)	0.974	0.955	0.912	0.946	0.947
	Average score	0.774	0.922	0.657	0.766	

As seen in the results (Table 1), substantial differences were shown when training and testing with different datasets. The best performing model for predicting any contrast independently of the number of replicates was the treebag (MixR) model, which we referred to as 'AutoRel'. Remarkably, the latter model presented a higher Kappa Score when testing the 3R and 9R datasets than the models trained with those datasets. In turn, the testing dataset with the highest average Kappa Score (the testing dataset best predicted by all models) was '9R'. The quality of the dataset did not appear to influence directly both

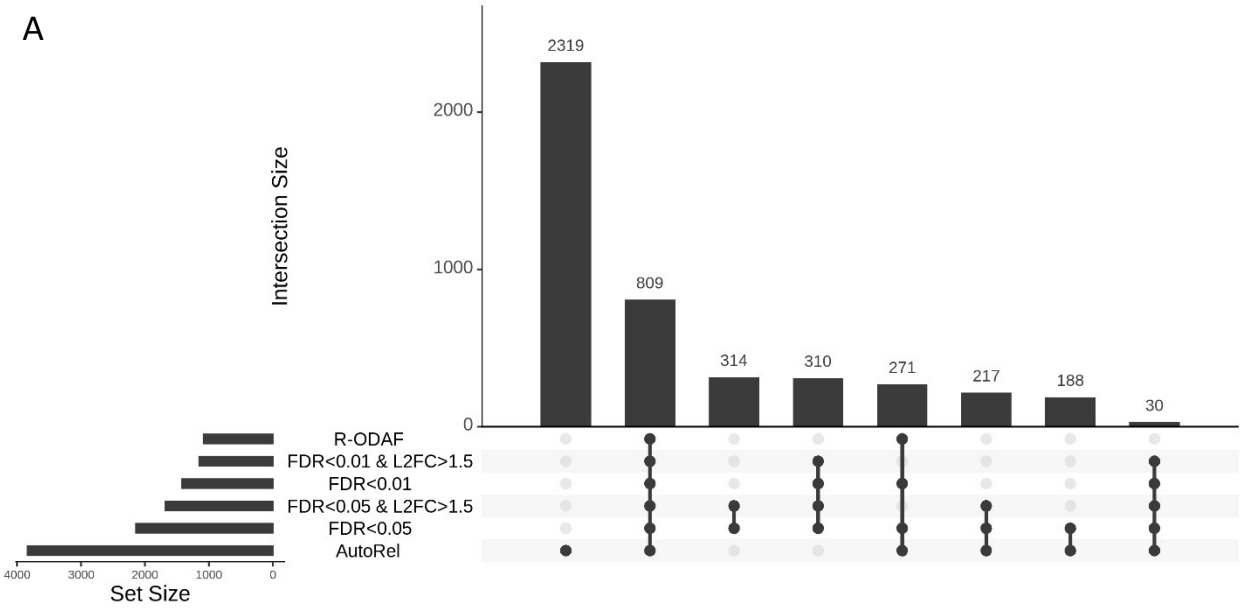
training and testing steps, as the average training and average testing scores presented a Pearson correlation of -0.521.

Model comparison

To further investigate how the 'treebag_MixR' model classified genes as 'relevant', we decided to compare the predictions of this model with several methods that could be used to select genes of interest in a transcriptomics study: FDR (p.adjusted value) < 0.05, FDR < 0.05 in combination with a log2 fold change > 1.5, FDR < 0.01, FDR < 0.01 in combination with a log2 fold change > 1.5, and R-ODAF. The statistical values were obtained using DESeq2. The input used to compare the different methods were the full 3R, 9R, and 21R datasets.

For the first comparison, we made use of the 3R dataset for which results showed the highest difference between 'relevant' and significant genes (Figure 3A). In addition to the 'relevant' set size being 1.8 times bigger than the significant set size, only 39.4% of the 'relevant' genes were significant (FDR < 0.05), and of these, 69.7% presented an L2FC > 1.5. Although only 28.1% of the 'relevant' genes were detected by R-ODAF, 100% of the R-ODAF genes were 'relevant'.

A


Intersection Size
2319
809
314
310
271
217
188
30
2000
1000
0
R-ODAF
FDR<0.01 & L2FC>1.5
FDR<0.01
FDR<0.05 & L2FC>1.5
FDR<0.05
AutoRel
4000
3000
2000
1000
0
Set Size

144

B

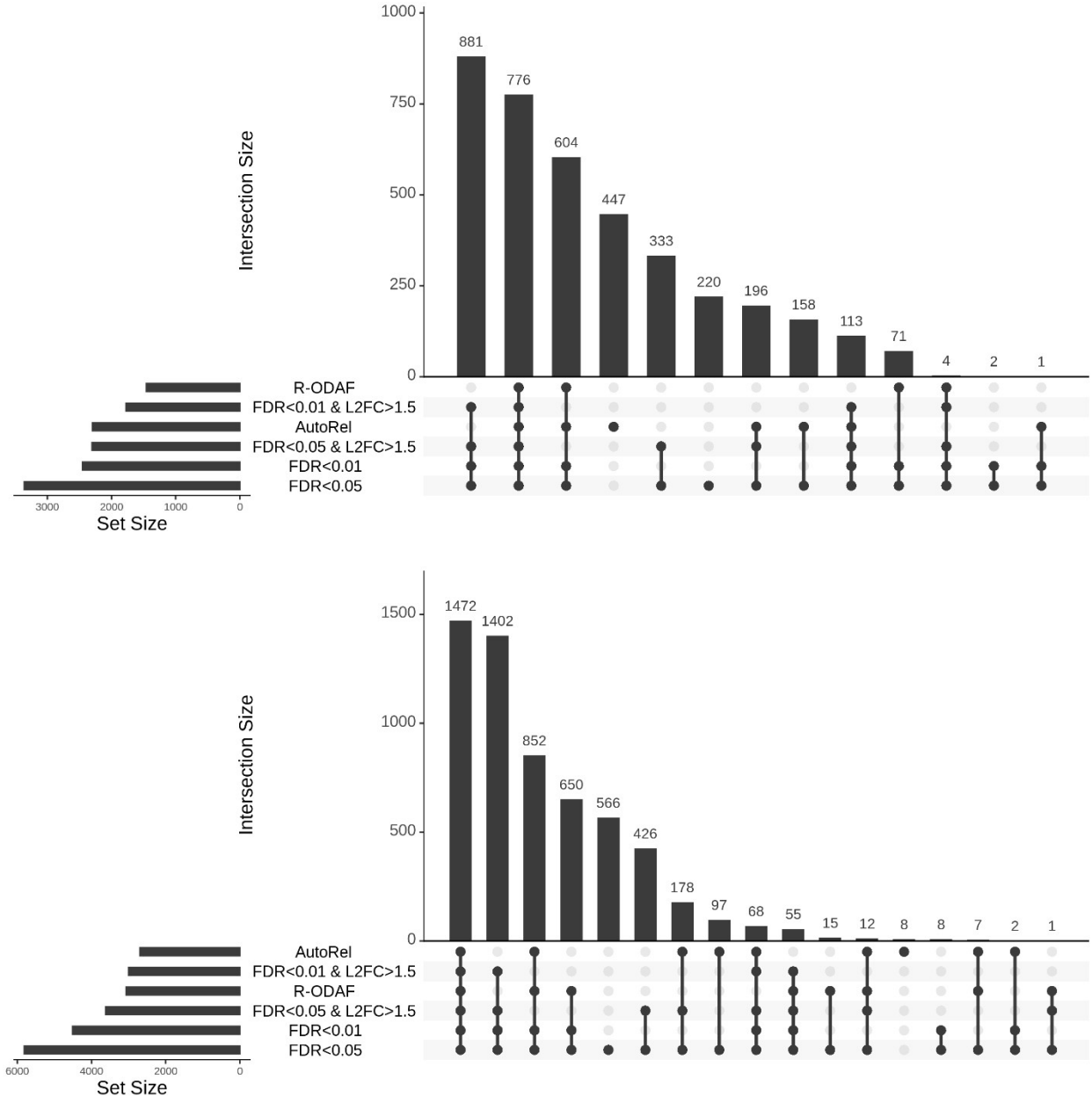


Figure 3: UpSet plot of the intersections between the evaluated methods using the diverse datasets. ***A:*** *3R dataset.* ***B:*** *9R dataset.* ***C:*** *21R dataset. Set Size referred to the total number of genes positively labeled or predicted by any of the methods*

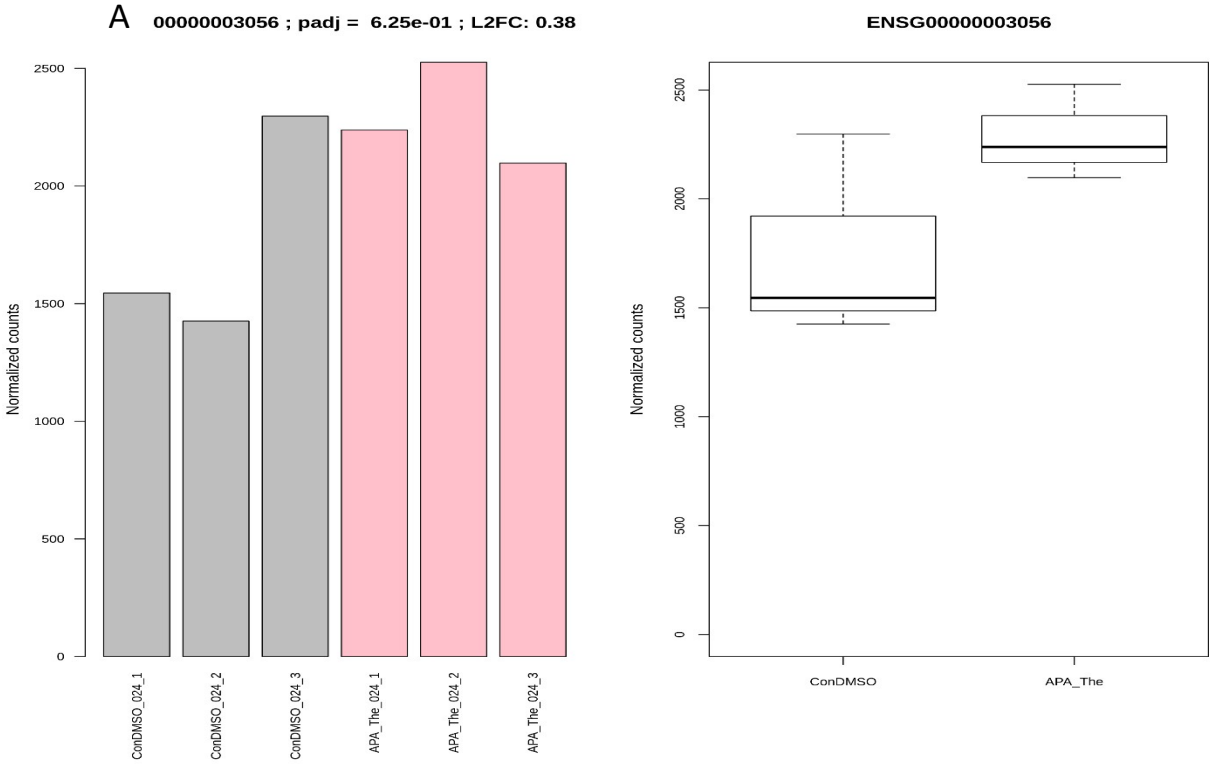
For the 9R dataset (Figure 3B), the 'relevant' class was no more the smallest group by Set Size. The impact of significance for the 'relevant' genes was decreased: 80.7% of them were significant (FDR < 0.05). The 9R results, when compared with the 21R results, showed that the 'relevant' class was less reliant on p-values the fewer replicates were used for the contrast. The decreased reliance on statistical values also decreased the proportion of R-ODAF genes in the 'relevant' set (60.3% of 'relevant' genes detected by R-ODAF), although simultaneously almost all genes in the R-ODAF set were also 'relevant'

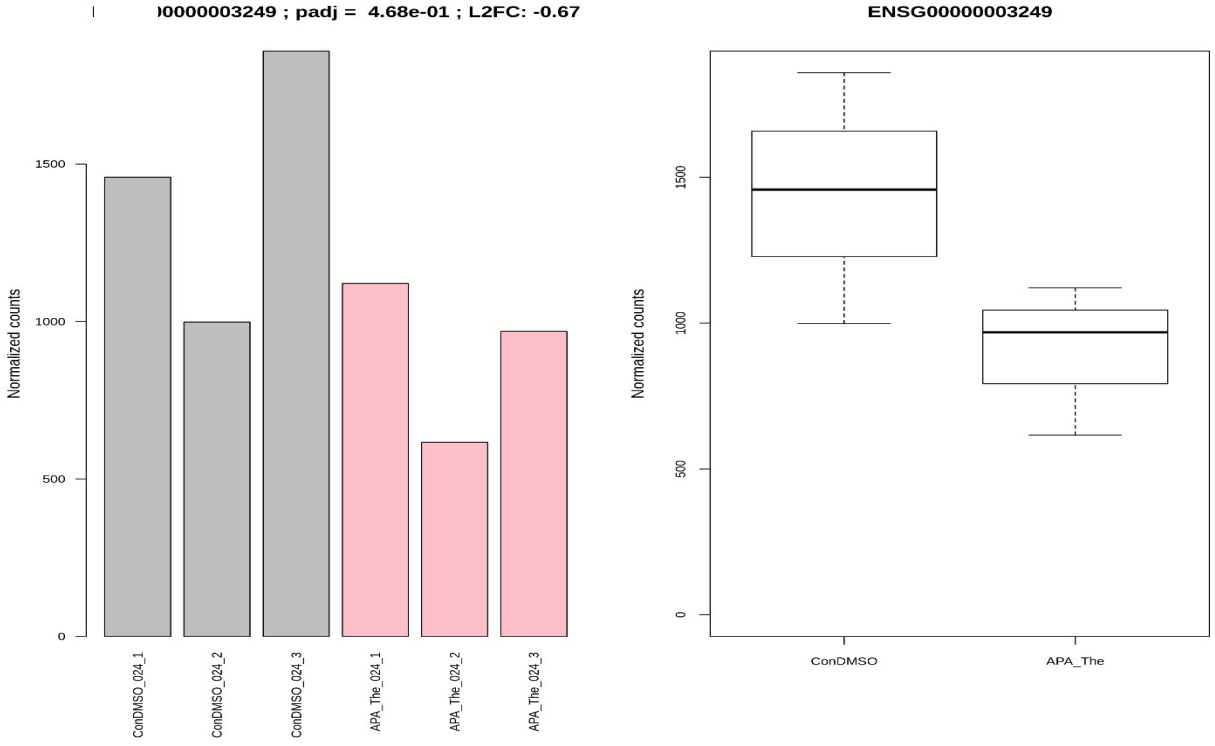
(94.8% of the R-ODAF set of genes were also ‘relevant’). 72.6% of the significant (p.adj < 0.05) ‘relevant’ genes presented an L2FC > 1.5. For the significant genes (FDR < 0.05), 68.1% of them were also ‘relevant’.

Lastly, we performed the same analysis using the 21R dataset as input (Figure 3C). For this number of replicates, the genes labeled as ‘relevant’ appeared to be a more restricted subset of genes compared to any other significant set of genes (as seen in the Set Size). The ‘relevant’ genes were mostly significant (99.7% of them with an FDR < 0.05), and slightly biased to have a high L2FC (64.3% of the significant ‘relevant’ genes presented an L2FC > 1.5). The ‘relevant’ genes also presented a high overlap with R-ODAF (87.0% of them detected by R-ODAF), but R-ODAF identified in total more genes (76.3% of the latter were ‘relevant’). For the significant genes (FDR < 0.05), 46.15% of them were also ‘relevant’, showing that for this number of replicates, significance was a necessary condition for ‘relevant’ genes, but not sufficient.

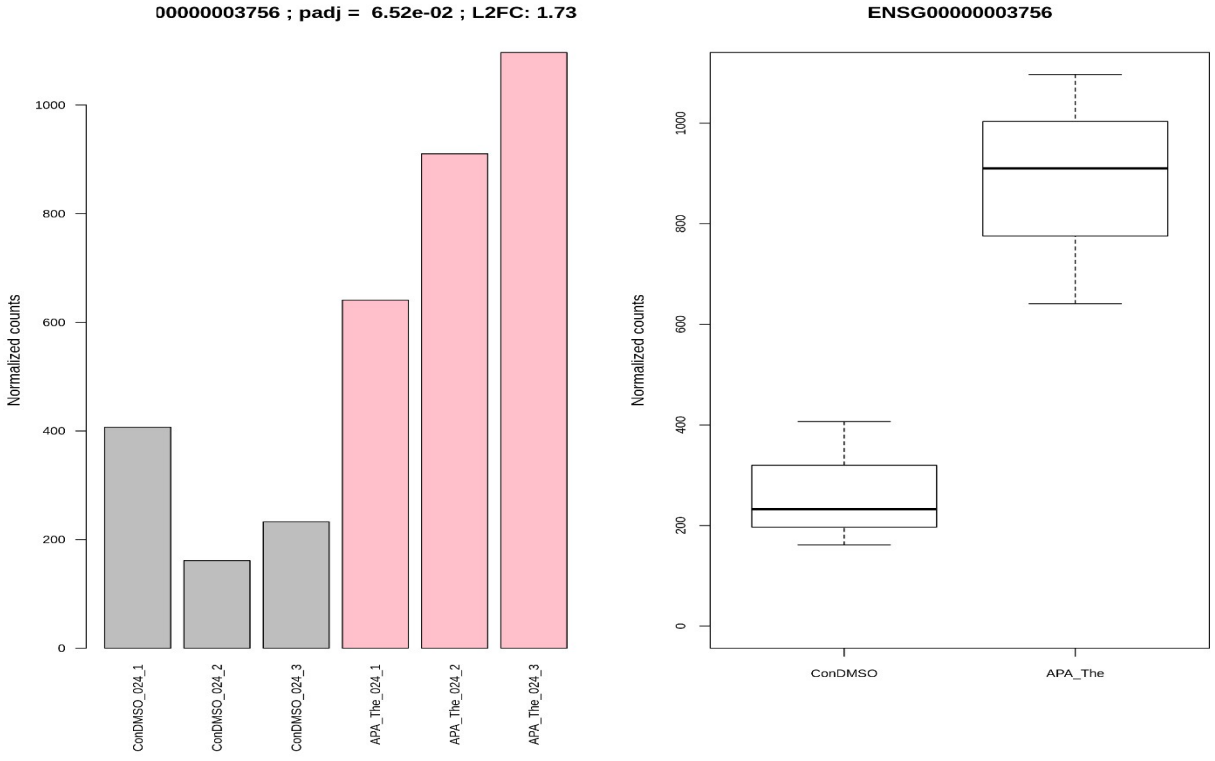
 Analysis of exclusively ‘relevant’ genes

The results of the model in the 3R dataset showed that the majority of the 'relevant' genes were not statistically significant using an FDR threshold of 0.05. We thus further analyzed these gene expressions to visualize how a gene expression change could be simultaneously relevant and not significant. For this reason, we extracted the normalized reads of the first three genes from the "exclusively relevant" set (M6PR, DBNDD1, and RBM5), and generated a barplot and boxplot for each of them (Figure 4A, Figure 4B, and Figure 4C).




)0000003249 ; padj = 4.68e-01 ; L2FC: -0.67
Normalized counts
0
500
1000
1500
ConDMSO_024_1
ConDMSO_024_2
ConDMSO_024_3
APA_The_024_1
APA_The_024_2
APA_The_024_3
ENSG00000003249
Normalized counts
0
500
1000
1500
ConDMSO
APA_The

175


00000003756 ; padj = 6.52e-02 ; L2FC: 1.73
Normalized counts
0
200
400
600
800
1000
ConDMSO_024_1
ConDMSO_024_2
ConDMSO_024_3
APA_The_024_1
APA_The_024_2
APA_The_024_3
ENSG00000003756
Normalized counts
0
200
400
600
800
1000
ConDMSO
APA_The

176

Figure 4: Barplot and boxplot of the normalized count expression of exclusively relevant genes. *A: AM6PR (ENSG00000003056, FDR = 6.25E-01. L2FC = 0.38).* ***B:*** *DBNDD1 (ENSG00000003249, FDR = 4.68E-01, L2FC = -0.67).* ***C:*** *RBM5 (ENSG00000003756, FDR = 6.52E-02, L2FC = 1.73). In the barplot (left half), the gray columns represented the gene expression in the control samples, while the pink columns represented the gene expression in the treated samples. The boxplot (right half) represented the same values while summarizing the values for each condition in a boxplot.*

M6PR gene expression (Figure 4A) showed that, in a triplicate setting, a single sample (ConDMSO_024_3) could affect substantially the statistical values, while still being potentially relevant for further investigation. For DBNDD1 (Figure 4B), a pattern could be seen for both groups triplicate-wise, while expressing in different scales, the low number of replicates makes these two conditions not significantly different.

Lastly, the expression levels of RBM5 (Figure 4C) showed a clear difference between both conditions, including a high log2 fold change (1.73). Even so, the use of a standard 5% FDR threshold identified the expression levels in the treatment group as not significantly different from the control condition.

Effect of relevant genes on biological interpretation

The AutoRel model detected 3839, 2293, and 2686 relevant genes in the 3R, 9R, and 21R datasets respectively. To evaluate the effect on the biological interpretation of the genes labeled by the model, we decided to use the relevant gene set derived from the 21R dataset. As a contrast gene set, we selected the significant (p. adj. < 0.05) genes from the same dataset, with a set size of 5811 genes.

The analysis showed that the GO terms in the three different ontologies (Biological Process, Biological Function, and Biological Component) were simplified in the relevant genes in comparison to the significant ones (Figure 5). In the Biological Process ontology, GO terms exclusive to the significant gene set were very variable and unrelated to hepatocytes exposed to APAP. Some examples were spermatogenesis (p-value = 4.51E-5), male gamete generation (p-value = 6.04E-5), regulation of neuron differentiation (p-value = 1.77E-4), and oocyte development (p-value = 2.54E-4). Out of the exclusive GO terms in the relevant gene set, the most significant were response to stimulus (p-value = 9.51E-5) and biological_process (p-value = 5.02E-4). The other three GO terms were related to coagulation (negative regulation of coagulation, negative regulation of blood coagulation, and negative regulation of hemostasis), which has been hypothesized to be enhanced via p62 in APAP-induced liver injury[29].

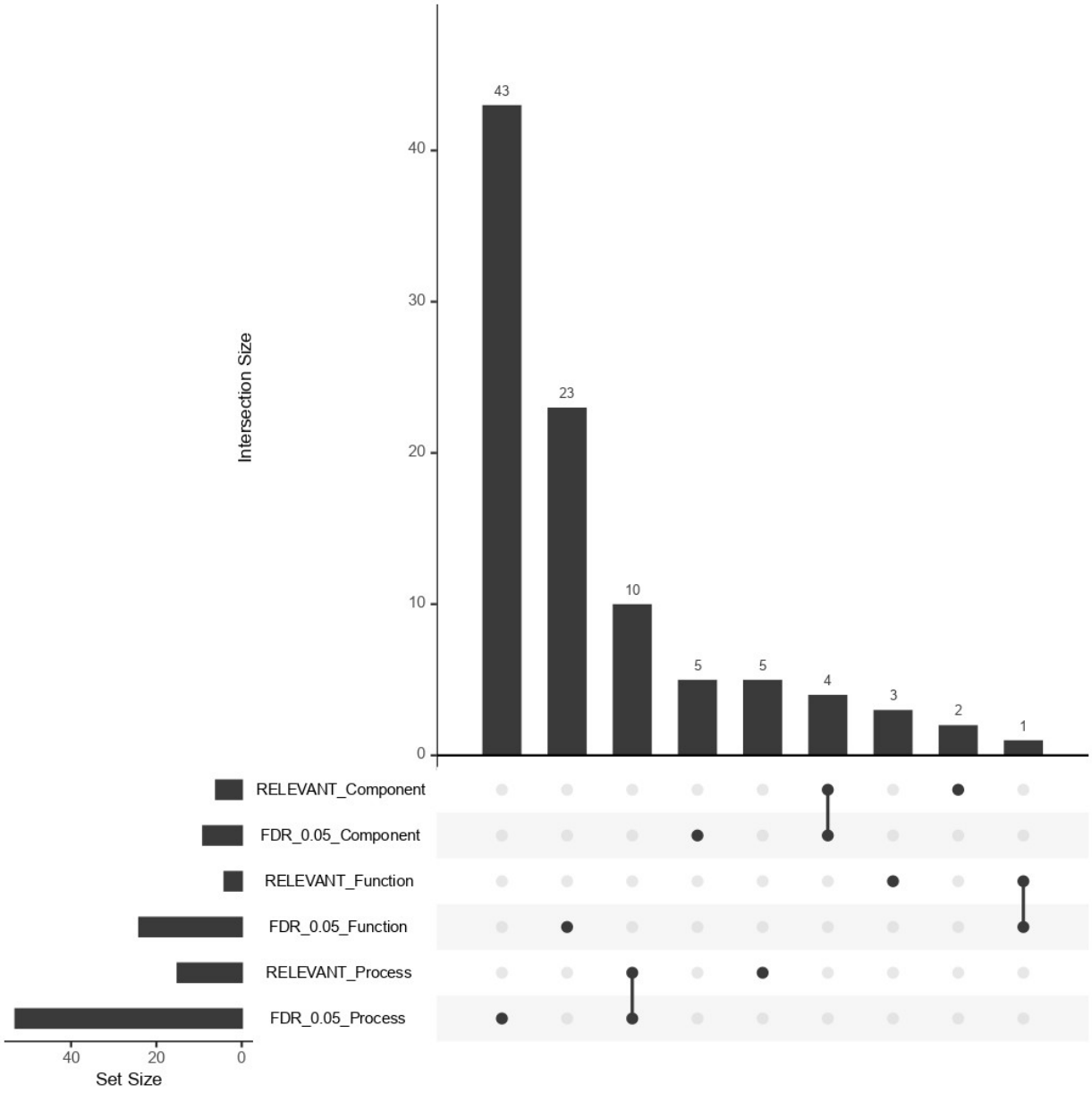


Figure 5: UpSet plot of the different enriched GO terms between significant (FDR 0.05) and relevant genes. *Set Size shows the number of enriched GO terms for a specific method (FDR < 0.05 or RELEVANT) in a concrete ontology (Process, Function, or Component). The Intersection Size shows how many enriched GO terms were exclusive for that method (single dot) or shared with the equivalent method for the same ontology (two dots connected by a line).*

For the Biological Function ontology, the only shared GO term between both methods was anion binding. The significant gene set showed a diverse range of terms related to the catalytic activity (phosphotransferase activity, kinase activity, sterol 14-demethylase activity) and molecule binding (peptide hormone receptor binding, CD4 receptor binding, drug binding), while the relevant set was limited to cytoskeletal protein binding. Lastly, the Biological Component ontology showed, despite the

lower number of GO terms in the relevant set, the highest overlap of the three ontology classes. Interestingly, the significant set presented 'neuron part' as an enriched GO term, but not the relevant set.

Dataset simulation

To evaluate the predictive value of the model, we generated simulated datasets with different numbers of replicates: 50 (50R), 21 (21R), 9 (9R), and 3 (3R). For every number of replicates setting, 100 iterations were performed to avoid skewed results. As predictors, we selected significant genes (FDR < 0.05), relevant genes, and a combination of relevant and dubious genes. We compared the predicted results with the original population difference and evaluated it using several parameters: Accuracy, Kappa, Sensitivity (or Recall), Specificity, PPV (or Precision), NPV, and Balanced Accuracy.

In the 50R setting (Supplementary Figure 7), relevant genes presented a similar accuracy to significant ones. Kappa and sensitivity showed the biggest decrease, the latter of which (in combination with a slightly increased specificity) led to a decrease in the balanced accuracy. PPVs were instead increased, in combination with a slight decrease in NPVs. Including the dubious genes (Supplementary Figure 8) decreased the differences with the significant genes among all metrics. In the 21R setting (Supplementary Figures 9 and 10), highly similar results to the 50R setting were found (for both relevant and dubious+relevant groups), which suggested that similar results might be interpolated in the range between 21 and 50 replicates.

In the 9R setting (Supplementary Figure 11), the differences across metrics generally decreased. Accuracy continued to be similar between relevant and significant genes. Kappa was also still decreased for relevant genes, but the difference was smaller. Sensitivity was also slightly decreased, but in combination with an almost unchanged specificity, led to an almost equally balanced accuracy between both groups. The PPV was also mostly decreased, while the NPV was overwhelmingly unaffected. Including dubious genes increased sensitivity but at the cost of PPV (Supplementary Figure 12).

In the 3R setting (Supplementary Figure 13), we found the biggest differences across groups. Accuracy was slightly decreased, and kappa, although also decreased, showed a much bigger variance depending on the dataset/iteration. Sensitivity was strongly increased, while specificity was slightly decreased, thus balanced accuracy was only marginally decreased. The increase in sensitivity was at the cost of PPV, which strongly decreased. Adding dubious genes mostly decreased most metrics: accuracy, kappa, specificity, PPV, and balanced accuracy (Supplementary Figure 14).

Discussion

As an alternative to the typical selection of differentially expressed genes using a statistical threshold, we chose to build a model that would select genes with biologically relevant differences across treatments trained on the criteria selection we generally use for further validation. For this, the model was trained with our decision-making based on the characteristics observed in the expression of hundreds of genes. Training a 'treebag' algorithm with observations from diverse datasets resulted in the optimal model to classify 'relevant', 'dubious', and 'irrelevant' genes.

Among the best-performing algorithms trained on distinct training datasets (3R, 9R, 21R, and MixR), the model that showed the best overall Kappa accuracy was the one trained with the MixR dataset using the 'treebag' algorithm (which we named 'AutoRel'). This was consistent with our suppositions, as the training dataset contained the highest number of observations in addition to the most variance among them.

Analyzing the variable importance of the AutoRel model allowed us to evaluate what factors were critical for selecting genes as potentially relevant in a biological setting. As expected, P-values and FDR values were vital predictors for the model, as extreme changes statistically tended to refute the null hypothesis. Secondly, the 'threequartilerule' feature (**Supplementary Methods**), which refers to gene changes where there are three quartile differences across groups (such as the minimum of one group presenting a value higher than the 3rd quartile of the other group), also appeared highly important in AutoRel. This might be explained by the fact that these genes, while being substantially different across groups, were more informative than more extreme rules (such as 'fourquartilediff'), as the former difference occurred more frequently than the latter.

Even though statistical values were critical for the AutoRel model, the overlap between relevant and significant genes decreased drastically as the number of replicates involved decreased (, Figure 3). This was especially the case with the dataset with the lowest number of replicates (3R, Figure 3A). This might have been due to the higher impact of every sample for the whole contrast. As seen in Figure 4, the expression of a single replicate (the third control triplicate) resulted in a high p-value, independently of a clear difference between most samples in both groups. This was also an example of how using L2FC as a statistical threshold might have discarded this gene: the average of the control group was highly affected by a single outlier, decreasing the L2FC value. In addition, M6PR, with a relatively high gene expression, was modestly increased in relative terms to the control average, even though (discarding the outlier), there was at least a 500 normalized count difference between both groups. Thus, we observed that AutoRel selected relevant genes based on the differences between the majority of the replicates from both groups,

presenting resilience to the presence of potential outliers, in contrast to the use of statistical p-values or fold changes.

Additionally, the effects of manually selecting biologically relevant genes through the developed model were also perceivable during the biological interpretation phase. When evaluating the differences between significant (FDR < 0.05) and relevant genes, we observed that in all the three main GO categories (Process, Function, and Component) there was a reduction in the number of enriched GO terms in the relevant set. The shrinkage did not appear to be stochastic, as some GO terms completely unrelated to the hepatic toxicity of APAP disappeared (spermatogenesis, oocyte development), while other terms, such as negative regulation of coagulation, were exclusive to the relevant set. Previous research has found an inducement of coagulation in APAP-induced liver injury[29], thus compensatory mechanisms might be in effect in therapeutic doses.

A potential improvement of the model could include previously selected differentially expressed genes, especially if those were later validated (or not) *in vitro*. In addition, providing an online tool for quick assessments of random quantification barplots to increase the size of the training dataset and help diversify the selection criteria outside of our department would be beneficial. This would of course require strict criteria to recruit external experts in the training, to avoid the addition of noisy data to the dataset.

In conclusion, we generated a machine learning model named AutoRel that can artificially simulate the manual selection often required for further research on a select number of differentially expressed genes based on our understanding of the biology and sequencing processes. AutoRel was more stringent than standard FDR thresholds in experiments with a high number of replicates, decreasing the number of false positives. For a low number of replicates, the effects were inverted, where the number of relevant genes was superior to the number of significant ones, decreasing the number of false negatives in a low informative experiment. The model (and the code used to generate it) is publicly available on GitHub: https://github.com/jochotecoa/AutoRel.git .

Bibliography

1 Verheijen, M. C. *et al.* R-ODAF: Omics data analysis framework for regulatory application. *Regul Toxicol Pharmacol* **131**, 105143, doi:10.1016/j.yrtph.2022.105143 (2022).

2 Lovett, R. A. Toxicogenomics. Toxicologists brace for genomics revolution. *Science* **289**, 536-537, doi:10.1126/science.289.5479.536 (2000).

3 Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. Moving to a World Beyond "p < 0.05". *The American Statistician* **73**, 1-19, doi:10.1080/00031305.2019.1583913 (2019).

4 Hurlbert, S. H., Levine, R. A. & Utts, J. Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires. *The American Statistician* **73**, 352-357, doi:10.1080/00031305.2018.1543616 (2019).

5 McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. Abandon Statistical Significance. *The American Statistician* **73**, 235-245, doi:10.1080/00031305.2018.1527253 (2019).

6 Ochoteco Asensio, J., Verheijen, M. & Caiment, F. Predicting missing proteomics values using machine learning: Filling the gap using transcriptomics and other biological features. *Comput Struct Biotechnol J* **20**, 2057-2069, doi:10.1016/j.csbj.2022.04.017 (2022).

7 Weis, C. *et al.* Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. *Nat Med* **28**, 164-174, doi:10.1038/s41591-021-01619-9 (2022).

8 Zhang, J. *et al.* Rapid Antibiotic Resistance Serial Prediction in Staphylococcus aureus Based on Large-Scale MALDI-TOF Data by Applying XGBoost in Multi-Label Learning. *Front Microbiol* **13**, 853775, doi:10.3389/fmicb.2022.853775 (2022).

9 Jabal, M. S. *et al.* Interpretable Machine Learning Modeling for Ischemic Stroke Outcome Prediction. *Front Neurol* **13**, 884693, doi:10.3389/fneur.2022.884693 (2022).

10 Kuepfer, L. *et al.* Applied Concepts in PBPK Modeling: How to Build a PBPK/PD Model. *CPT Pharmacometrics Syst Pharmacol* **5**, 516-531, doi:10.1002/psp4.12134 (2016).

11 Team, R. C. *R: A Language and Environment for Statistical Computing*, <https://www.R-project.org/> (2021).

12 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

13 Max Kuhn [aut, c., Jed Wing [ctb], Steve Weston [ctb], Andre Williams [ctb], Chris Keefer [ctb], Allan Engelhardt [ctb], Tony Cooper [ctb], Zachary Mayer [ctb], Brenton Kenkel [ctb], R Core Team [ctb], Michael Benesty [ctb], Reynald Lescarbeau [ctb], Andrew Ziem [ctb], Luca Scrucca [ctb], Yuan Tang [ctb], Can Candan [ctb], Tyler Hunt [ctb]. *caret: Classification and Regression Training*, <https://github.com/topepo/caret/> (2019).

14 Todorov, V. *rrcov: Scalable Robust Estimators with High Breakdown Point*, <https://github.com/valentint/rrcov> (2021).

15 Klaus Schliep [aut, c., Klaus Hechenbichler [aut], Antoine Lizee [ctb]. *kknn: Weighted k-Nearest Neighbors*, <https://github.com/KlausVigo/kknn> (2016).

16 Zeileis, A. K. a. A. S. a. K. H. a. A. kernlab -- An {S4} Package for Kernel Methods in {R}. *Journal of Statistical Software* **11**, 1--20, doi:10.18637/jss.v011.i09 (2004).

17 Ripley, W. N. V. a. B. D. *Modern Applied Statistics with S*. Fourth edn, (Springer, 2002).

18 Majka, M. *naivebayes: High Performance Implementation of the Naive Bayes Algorithm in R*, <https://CRAN.R-project.org/package=naivebayes> (2019).

19 Wurm, M. J., Rathouz, P. J. & Hanlon, B. M. Regularized Ordinal Regression and the ordinalNet R Package. *J Stat Softw* **99**, doi:10.18637/jss.v099.i06 (2021).

20 T. Hastie, R. T., Balasubramanian Narasimhan, Gil Chu. *pamr: Pam: Prediction Analysis for Microarrays*, <https://cran.r-project.org/package=pamr> (2019).

21 Wiener, A. L. a. M. Classification and Regression by randomForest. *R News* **2**, 18-22 (2002).

22 Terry Therneau [aut], B. A. a., cre], Brian Ripley [trl] (producer of the initial R port, maintainer 1999-2017). *rpart: Recursive Partitioning and Regression Trees*, <https://github.com/bethatkinson/rpart> (2018).

23 Line Clemmensen, c. b. M. K. *sparseLDA: Sparse Discriminant Analysis*, <http://www.imm.dtu.dk/~lhc> (2016).

24 Andrea Peters [aut], T. H. a., cre], Brian D. Ripley [ctb], Terry Therneau [ctb], Beth Atkinson [ctb]. ipred: Improved Predictors. (2021).

25 Kuhn, M. *8 Models Clustered by Tag Similarity*, <https://topepo.github.io/caret/models-clustered-by-tag-similarity.html> (2014).

26 Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48, doi:10.1186/1471-2105-10-48 (2009).

27 Assefa, A. T., Vandesompele, J. & Thas, O. SPsimSeq: semi-parametric simulation of bulk and single-cell RNA-sequencing data. *Bioinformatics* **36**, 3276-3278, doi:10.1093/bioinformatics/btaa105 (2020).

28 Zhang, W. *et al.* Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol* **16**, 133, doi:10.1186/s13059-015-0694-1 (2015).

29 Qian, H. *et al.* Dual roles of p62/SQSTM1 in the injury and recovery phases of acetaminophen-induced liver injury in mice. *Acta Pharm Sin B* **11**, 3791-3805, doi:10.1016/j.apsb.2021.11.010 (2021).

Methods

Dataset composition

The original dataset used for this manuscript originated from a 3D liver cell culture made of spheroids by InSphero Inc. 42 samples were taken, 21 per group: control DMSO and acetaminophen (APAP) exposure. Those 21 samples were the accumulation of samples in triplicate at 7 different time points: 2, 8, 24, 72, 168, 240, and 336 hours. The applied dose was selected based on a PBPK profile, whose design is described in Kuepfer et al[10]. Total RNA was extracted and ribo-depleted. The libraries were sequenced on an Illumina Hiseq 2000, where the average sequencing depth was 27.8 million reads. The nomenclature of the samples was as follows: "Exposure"_" Timepoint"_" Replicate", where exposure could be either the control group ("ConDMSO") or the treatment ("APAP"), followed by any of the seven time points (002 – 336), and ending with any of the triplicates (1 – 3).

This dataset was used as input to the DESeq2 pipeline in three different configurations: 3 samples (triplicates from a single time point: 24 hours) from each group contrasted against each other, 9 samples (triplicates from 3-time points: 2, 8, and 24 hours) from each group, and 21 samples (triplicates from 7-time points: 2, 8, 24, 72, 168, 240, and 336 hours) from each group. All programming scripts that pertain to the analyses of this manuscript were written in the R language (major version 3, minor version 6.3)[11]. For all 3 comparisons, we used the R package 'DESeq2' (version 1.26.0)[12] to obtain both the quantitative and statistical features that would be used as input for the model. The quantitative features were obtained from the normalized counts. The normalization was performed using DESeq2's median of ratios. The features used for our models are listed and described in the 'Supplementary Methods' document.

Model training

The targets to be predicted were annotated manually for the training dataset. The three labels assigned were 'irrelevant', 'dubious', and 'relevant'. The 'irrelevant' label was assigned when a gene was not estimated to present a biological effect, as a consequence of (but not limited to) both groups being almost identical, their distributions being so variable that no clear change can be identified, or not enough reads are sequenced for most samples for that gene. 'Relevant' genes were those where a substantial change could be detected for most samples in both groups, so that a researcher would consider them for further research for their biological effect, such as the distributions of both groups spanning a different range of high quantification levels. Observations that did not classify for either of the previously described classes were labeled as 'Dubious'. To evaluate each gene in each comparison, we plotted the normalized counts for both groups using both barplots and boxplots. For each comparison, we labeled a random set of observations to include a diverse set of distribution changes for the model to learn from. Thus, after each comparison, we obtained a distinct dataset, which we named in an increasing number of replicates: datasets 3R, 9R, and 21R.

Pre-processing and modeling

After forming the complete dataset, it led to a total of 160 features. Aside from the 3 datasets derived from the 3 different comparisons (3, 9, and 21 replicates), we generated a new dataset that included the observations of all 3 datasets (dataset named MixR). For all four datasets, we performed several pre-processing steps to filter out redundant or uninformative features that might either worsen or slow down the model training process. For the first 2 steps, we used an under-sampled version of the dataset. Under-sampling reduces the size of all groups to the size of the smallest one. In the first step, we used a (near) zero-variance filter using the 'nearZeroVar' function from the caret[13] library. Without under-sampling, a feature with a common value that was predominant in the most common labels would have presented a reduced variance, and thus would have been more prone to be discarded due to over-representation of a target label, and not due to low variance. The second step involved the removal of features highly correlated to each other (correlation > 0.99). The last step of pre-processing involved the exclusion of features with a linear dependency on other features, using caret's 'findLinearCombos' function.

After the pre-processing of the dataset, we built different models using 11 different algorithms available to use via the 'caret' package: 'CSimca' (SIMCA[14]), 'kknn' (k-Nearest Neighbors[15]), 'lssvmRadial' (Least Squares Support Vector Machine with Radial Basis Function Kernel[16]), 'multinom' (Penalized Multinomial Regression[17]), 'naive_bayes' (Naïve Bayes[18]), 'ordinalNet' (Penalized Ordinal Regression[19]), 'pam' (Nearest Shrunken Centroids[20]), 'rf' (Random Forest[21]), 'rpart2' (CART[22]), 'sparseLDA' (Sparse Linear Discriminant Analysis[23]), and 'treebag' (Bagged CART[24]). We selected these algorithms based on maximal dissimilarity sampling starting from the 'rf' algorithm[25].

Model evaluation

To evaluate the models, we assessed the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for the 3 labels (for each model). To summarize those 4 metrics, we averaged their values at the label level, resulting in 3 metrics per model. We named each of these values 'equilibrated accuracies'. Next, to select the best model per comparison, the mean of these equilibrated accuracies (average of the previous 3 values) was computed; and the model with the highest value was selected. Even so, we verified in all cases that the metrics were not biased for labels that were relatively less important (i.e. a model with high performance for the 'dubious' label but low for the 'relevant' label). Concretely, we first selected models that presented a near-maximum equilibrated accuracy for the 'relevant' label. A near maximum value was defined as a value with at least 98.5% of the maximum value (i.e. if the maximum value was 1, values between 0.985 and 1 were considered near maximum). Out of the selected models, we further selected those with near-maximum equilibrated accuracies of the 'irrelevant' label. And lastly, we selected models with near-maximum equilibrated accuracies of the 'dubious' label. If several models were selected in the last step, the model with the highest average equilibrated accuracy was selected.

To calculate the variable importance for the models, we used the 'varImp' function from the 'caret' package. We used the default parameters to calculate and plot the variables' importance. The importance values were scaled between 0 and 100. This resulted in different calculations depending on the algorithm. For example, 'rf' and 'treebag' presented their variable importance calculation, while algorithms like 'sparseLDA' used ROC (Receiver Operating Characteristic) curve variable importance, where the variables were sorted by maximum importance across the classes.

Model comparison

To compare the models' accuracies across different datasets, for each model trained with a specific dataset, we tested each of those datasets: 3R, 9R, 21R, and MixR datasets. In case the same dataset would be used for training and testing, we split the full dataset into training and testing datasets to prevent an overfit. The accuracy metric used was Cohen's Kappa Coefficient, as it takes into account class imbalance while providing a straightforward metric.

Method comparison

We utilized the same initial datasets to compare the different methods to detect differential expression. The p-values, FDR values, and log2 fold-change values were all obtained from the results table from DESeq2 for each comparison. We use the '&' symbol to indicate that the conditions on both sides of the symbol need to be true. The methods were as follows: FDR < 0.05, FDR < 0.05 & |log2 Fold Change value| > 1.5, FDR < 0.01, FDR < 0.01 & |log2 Fold Change value| > 1.5, R-ODAF, and 'Relevant' labelled genes by our model.

Enrichment analysis

We used GORILLA[26] for the Gene Ontology enrichment analysis. The advanced parameters included a 'P-value threshold' of 10^-3, and fast mode was enabled. The data used were derived from the 21R dataset. The FDR 0.05 gene set contained the significantly differentially expressed genes based on an FDR threshold of 0.05, which contained 5811 genes. The RELEVANT gene set contained the genes labeled as 'relevant' by our 'AutoRel' model, which contained 2686 genes.

Simulated dataset

The simulated datasets were generated using the SPsimSeq R package[27]. The Zhang bulk RNA-seq data[28] was used as a reference for the data generation process. Four comparisons were made using a different number of replicates: 50 (as performed in the package demonstration), 21, 9, and 3. The total number of genes was 3000, and the proportion of Differentially Expressed Genes was 10% (300 genes), with at least a 0.5 log-fold-change in the source data. Each of the 4 comparisons was performed 100-fold, generating a new/random dataset for each iteration. We applied the DESeq2 pipeline and used the consequent results as input for AutoRel's prediction process.

Supplementary Methods (Chapter 5)

Model features

The features used in the machine learning model are derived from the normalized counts. The mean, standard deviation (SD), and variance of all normalized counts per gene were added as features. 40 features were added (20 per group), each of them representing a quantile from 0% to 100% every 5% (i.e. quantile_55%_control). 20 features were related to the normalized values: all replicates (per group) were divided into 10 subgroups based on the order of their imputation. The median of each subgroup was obtained and saved as a feature (i.e. out of 21 samples, '1st subset median' would be the median between the 1st and 2nd replicates; '2nd subset median' would represent the median between the 3rd and 4th replicates, etc.). In addition, four features of N (absolute number) and proportion (N / number of samples) of zero values per group (i.e. N_nonexpressed_samples_treatment) were added. Next, 12 features for outlier features (N and proportion per group for 3 groups of outliers: total, mild, and extreme) were included. For each of the aforementioned features, an additional feature was added, where the feature of the treatment was divided by the same feature for the control group, whose feature name would start with 'foldchange' (foldchange_N_mild_outliers).

Quartile rules: all potential differences between the quartiles of the 2 groups were described with different features. Four degrees of changes were recorded as features depending on the shift between both group distributions: mild changes with only a 25% difference (i.e. the 50% quantile of the treatment group above the 75% quantile of the control group), 50% difference (i.e. 50% quantile of the treatment group above the 100% quantile of the control group), 75% difference (i.e. 25% quantile of the treatment group above the 100% quantile of the control group), and 100% difference (i.e. 0% quantile of the treatment group above the 100% quantile of the control group). The names for the rules were, respectively: `onequartilediff_rule', 'twoquartilediff_rule', 'threequartilediff_rule', and 'fourquartilediff_rule'. In addition, all possible differences were added as features (i.e. 'q1belmin' stood for 25% quantile of the control group below 0% quantile of the treatment group). The last feature regarding quartiles was named 'quartilediff_score'. The score represented an overall description of all potential changes between the two distributions. Any quartile rule where the treatment group values were above the control ones contributed +1 to the score. Any quartile rule where the treatment group values were below the control ones contributed -1 to the score. This resulted in a score that ranged between -10 and +10. For example, if the quantile 50% control group was below the quantile 0% treatment group, this resulted in two rules being true: 'q2belmin' and 'q1belmin'; thus the resulting score would be +2. If simultaneously the 75% quantile control group was above the 100% quantile treatment group, a negative rule would also apply ('q3abomax'), and thus the score would result in +1.

Aside from the third quartile rule, which inspired the aforementioned rules as features, we also incorporated additional rules inspired by the R-ODAF method[1]. The "expression consistency" rule uses values transformed to counts per million (CPM), and for an observation to be true, 75% of the values on either the control or treatment groups need to be above 1 CPM. A "spurious spike" rule was designed to detect single samples with most of either group's reads, where the threshold limit for the proportion of reads depends on the number of replicates: $Spike\,threshold = 1.4\,N^{-0.66}$, where N is the number of replicates of the group).

Another source of features for the dataset was the result table from each DESeq analysis. The result table included base means across samples ('baseMean'), standard errors ('lfcSE'), log2 fold changes ('log2FoldChange'), p-values ('pvalue'), adjusted p-values ('padj'), and test statistics ('stat'). An additional feature was added from R-ODAF, where an FDR threshold limit of 0.01 was established ('fdrlowerthan0.01'). All NA values from the result table for 'log2FoldChange' and 'stat' were transformed to zero, while all NA values for 'lfcSE', 'pvalue', and 'padj' were set to one.

Supplementary Results (Chapter 5)

Dataset 3R: 3 replicates per group

The 'relevant' class was generally the best performing across the three classes, followed closely by 'irrelevant' (Figure 1). The 'dubious' class was generally the worst-performing class, where a clear difference could be seen for all models but one ('CSimca'). We selected 'treebag' as the best-performing model on average.

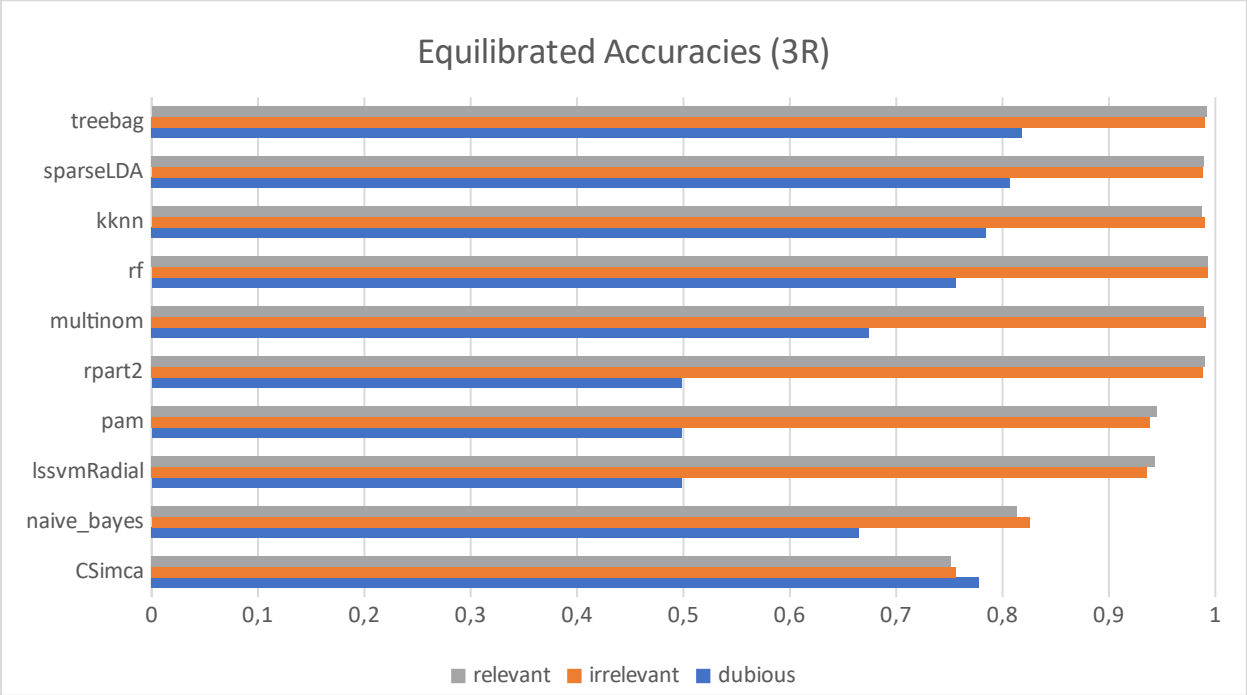


Figure 1: Equilibrated accuracies for the models built with 11 machine learning algorithms using the 3R dataset. The models were sorted by the average of equilibrated accuracies in descending order.

To further investigate the best-performing model, we calculated the variable importance of the 'treebag' model ('treebag_3R') to evaluate which features were crucial for a gene to be classified. Only the 20 most important features were plotted (Figure 2). Seven features distinguished as predominantly important, from most to least importance: 'pvalue', 'padj', 'rule_cpm_0.75_above_1', 'fdrlowerthan0.01', 'lfcSE', 'threequartilediff_rule', and 'twoquartilediff_rule'.

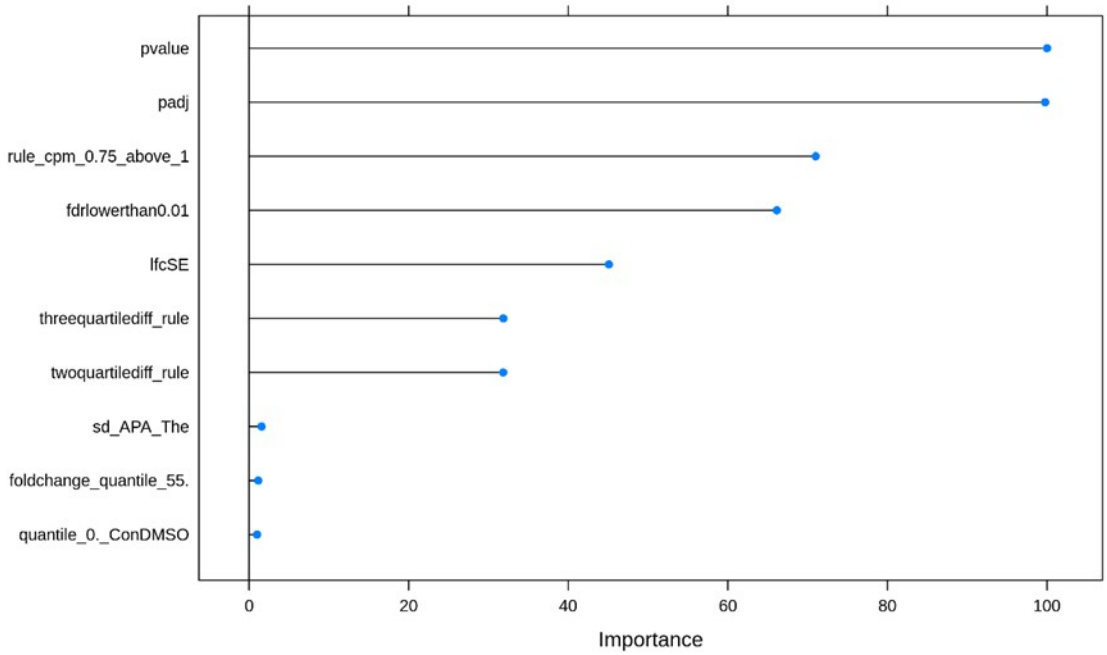


Figure 2: Variable importance for the 10 most important features sorted in decreasing order for the 'treebag' model trained with the 3R dataset.

The first two relate to the statistical significance, which is confirmed to be of primordial importance in a high number of observations even with only three replicates. The CPM rule described whether a gene was considered expressed in at least 75% of the samples of either group. This rule was essential in our decision-making during the labeling process, and thus it was also reflected in the model's variable importance. 'fdrlowerthan0.01' was able to label the most extreme cases of statistical significance. 'lfcSE' (Log Fold Change Standard Error) reflected the confidence of the log2 Fold Change, while the feature that represented the latter 'log2FoldChange' was situated in the lower half of the variable importance ranking. The last two features of the most important seven were related to the quartile differences between groups: these "rules" showed how different both groups were by comparing their quartiles.

Dataset 9R: 9 replicates per group

Even though the 'treebag' algorithm also performed correctly in the contrast with nine replicates per group (Figure 3), the best performing model was 'sparseLDA'. 'multinom' and 'treebag' presented a higher equilibrated accuracy for the 'relevant' class, but the difference was minimal. On the other hand, the equilibrated accuracy for the 'dubious' class was substantially higher for sparseLDA, leading to a higher mean equilibrated accuracy.

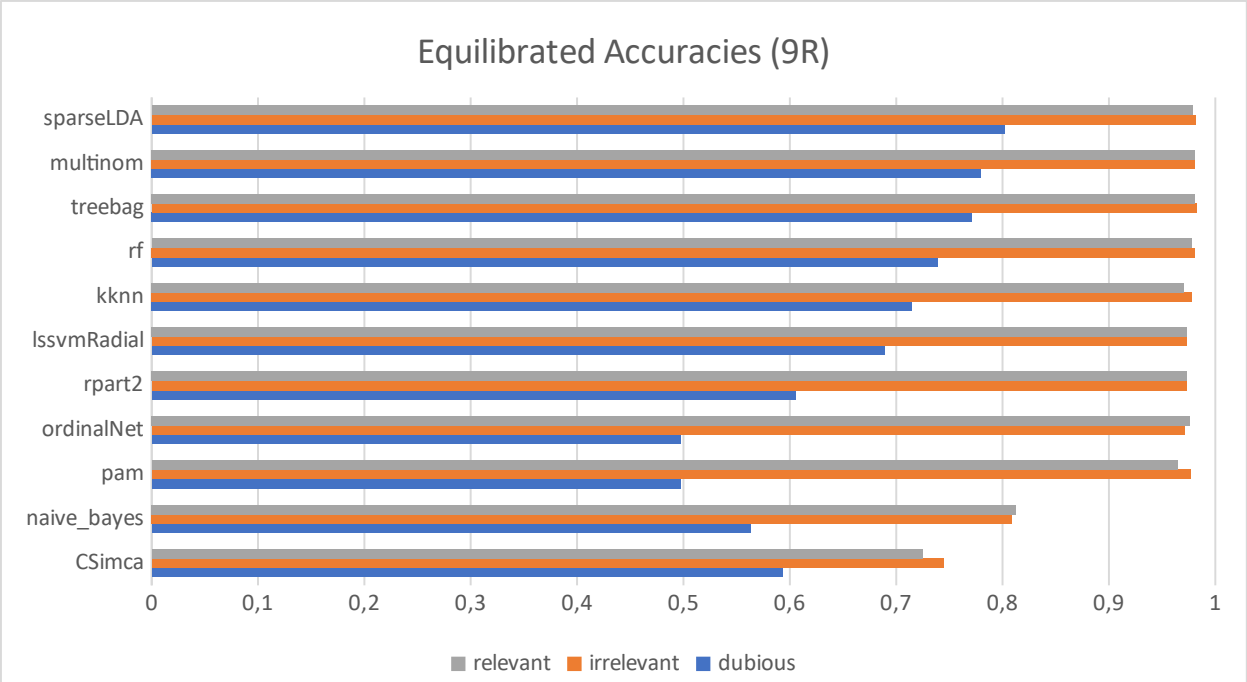


Figure 3: Equilibrated accuracies for the models built with 11 machine learning algorithms using the 9R dataset. The models were sorted by the average of equilibrated accuracies in descending order.

The variable importance calculation was obtained via the use of the ROC curve (Figure 4). As there were variable importances for each class, the variables were sorted by maximum importance across the classes. Focusing on the 20 most important variables, we were able to detect some similarities with the most important ones from the previous contrast. All the seven most important features for the 'treebag_3R' model were also present: the quartile difference rules ('twoquartilediff_rule', 'onequartilediff'), the statistical features ('padj', 'pvalue', 'fdrlowerthan0.01', and 'lfcSE'), and the CPM rule ('rule_cpm_0.75_above_1'). Other features related to those groups were also detected ('threequartilediff_rule' and 'baseMean'). In addition, we observed important features related to the treatment group ('X8th_subset_median_APA_The', 'var_APA_The', 'sd_APA_The', 'Proportion_nonexpressed_samples_APA_The'), and up-regulation ('q3belq2', 'q2belq1', 'q1belmin', 'q3belq1', 'q2belmin', and 'q3belmin').

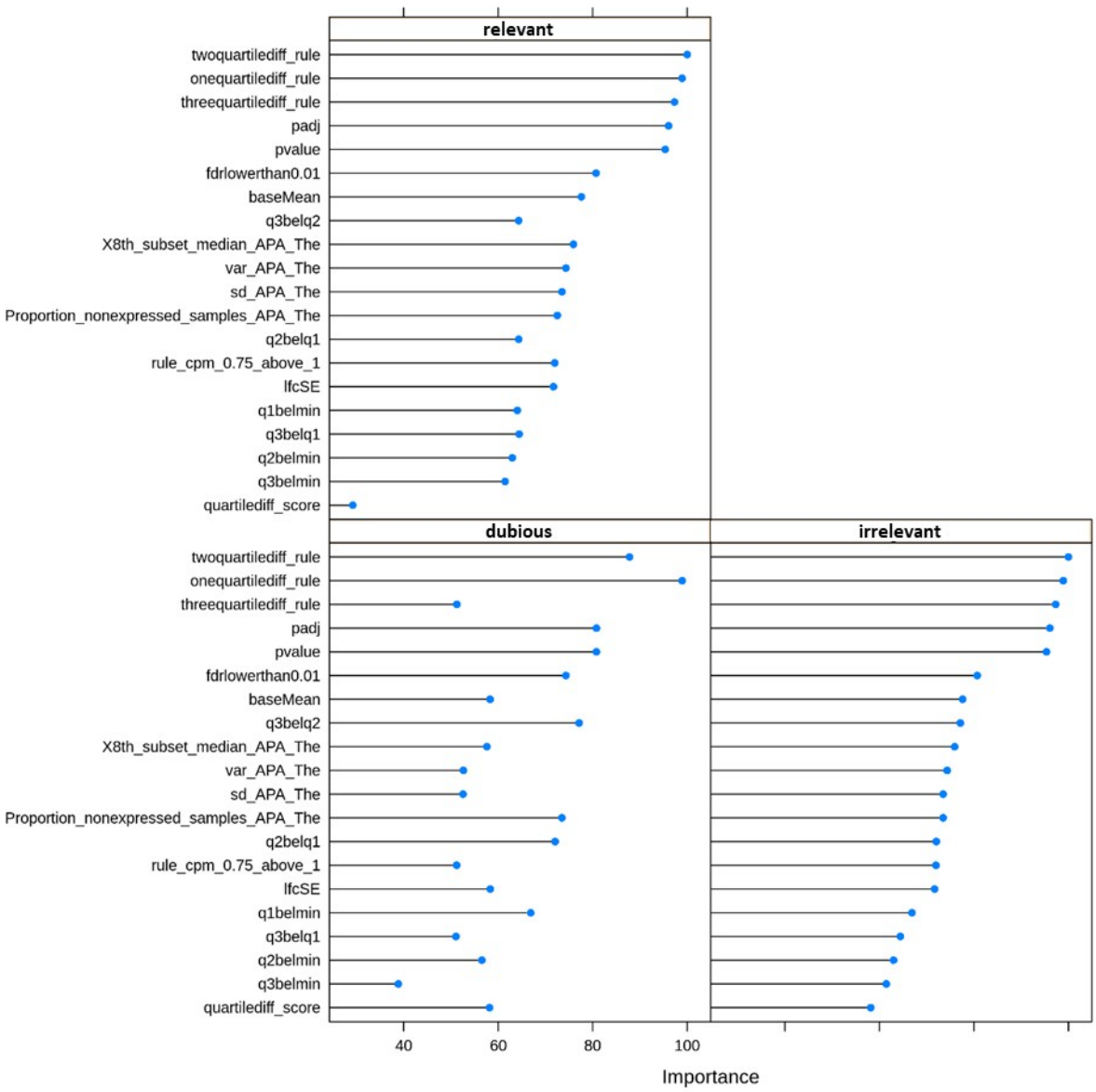


Figure 4: Variable importance for the 20 most important features for the 'sparseLDA' model trained with the 9R dataset. The features were sorted in a decreasing order based on the maximum importance value across the three classes.

Dataset 21R: 21 replicates per group

When analyzing the results for the equilibrated accuracies of this contrast (Figure 5), we noted that most models presented an equilibrated accuracy for the 'dubious' class of approximately 0.5. This was due to the lack of prediction of this class for these models, even though the training dataset included

observations from that class. The best performing model was thus random forest ('rf'), with high metrics for 'relevant' and 'irrelevant' classes, but especially for the 'dubious' class, being the only model above 0.6.

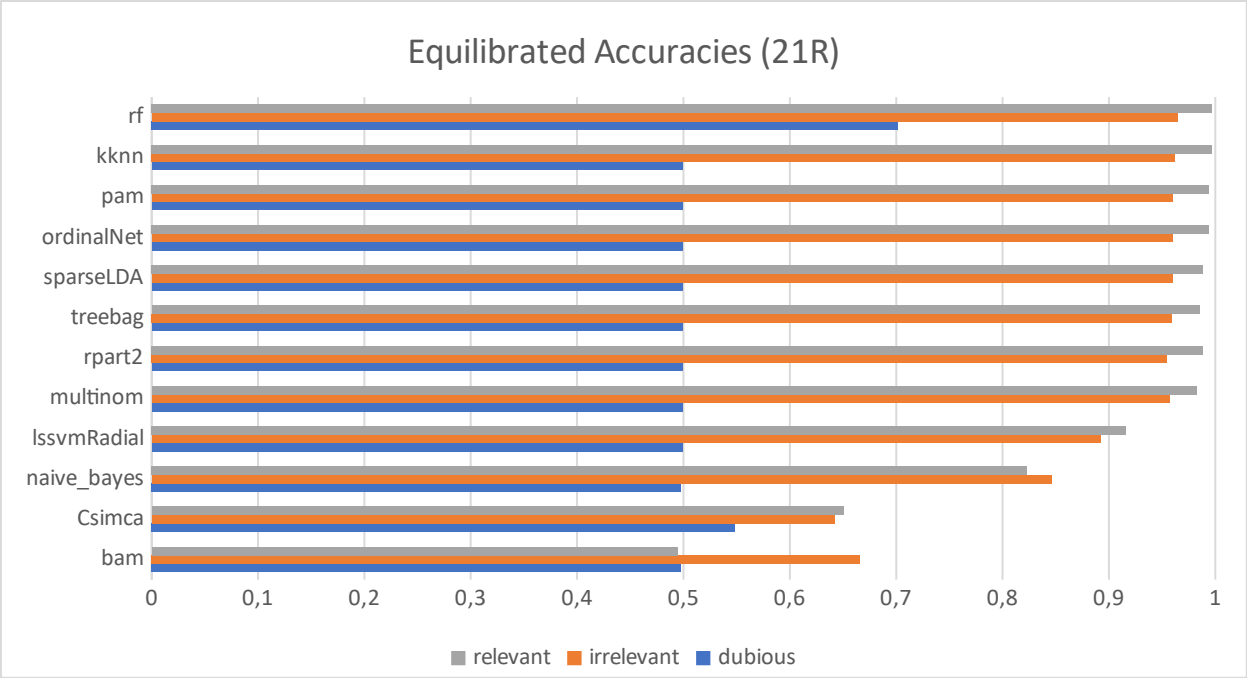


Figure 5: Equilibrated accuracies for the models built with 11 machine learning algorithms using the 21R dataset. The models were sorted by the average of equilibrated accuracies in descending order.

As performed for the previous contrasts, we evaluated the most important variables for the 'rf' model (Figure 6). The 'threequartilediff_rule' was predominantly the most important feature, having an importance value 400% higher than the second most important feature. Again, we observed similarities with the previous datasets: quartile difference features ('threequartilediff_rule', 'quartilediff_score', and 'maxbelq1'), statistical features ('pvalue', 'padj', and 'lfcSE'), and the CPM rule ('rule_cpm_0.75_above_1').

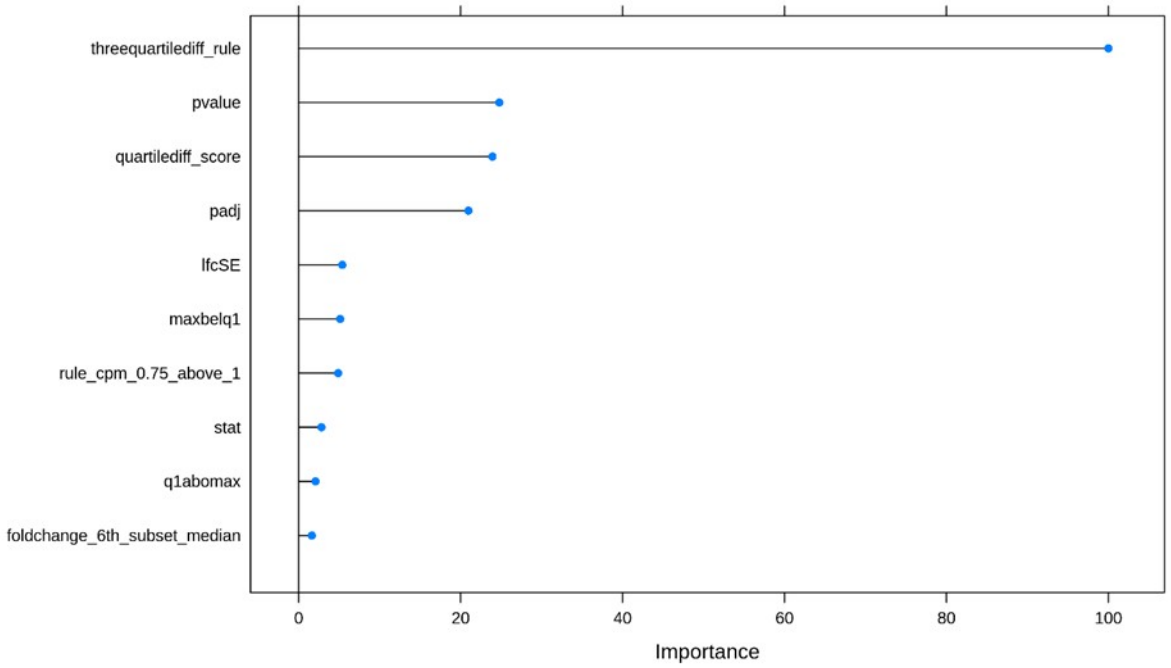


Figure 6: Variable importance for the 10 most important features sorted in decreasing order for the 'rf' model trained with the 21R dataset.

50 replicates

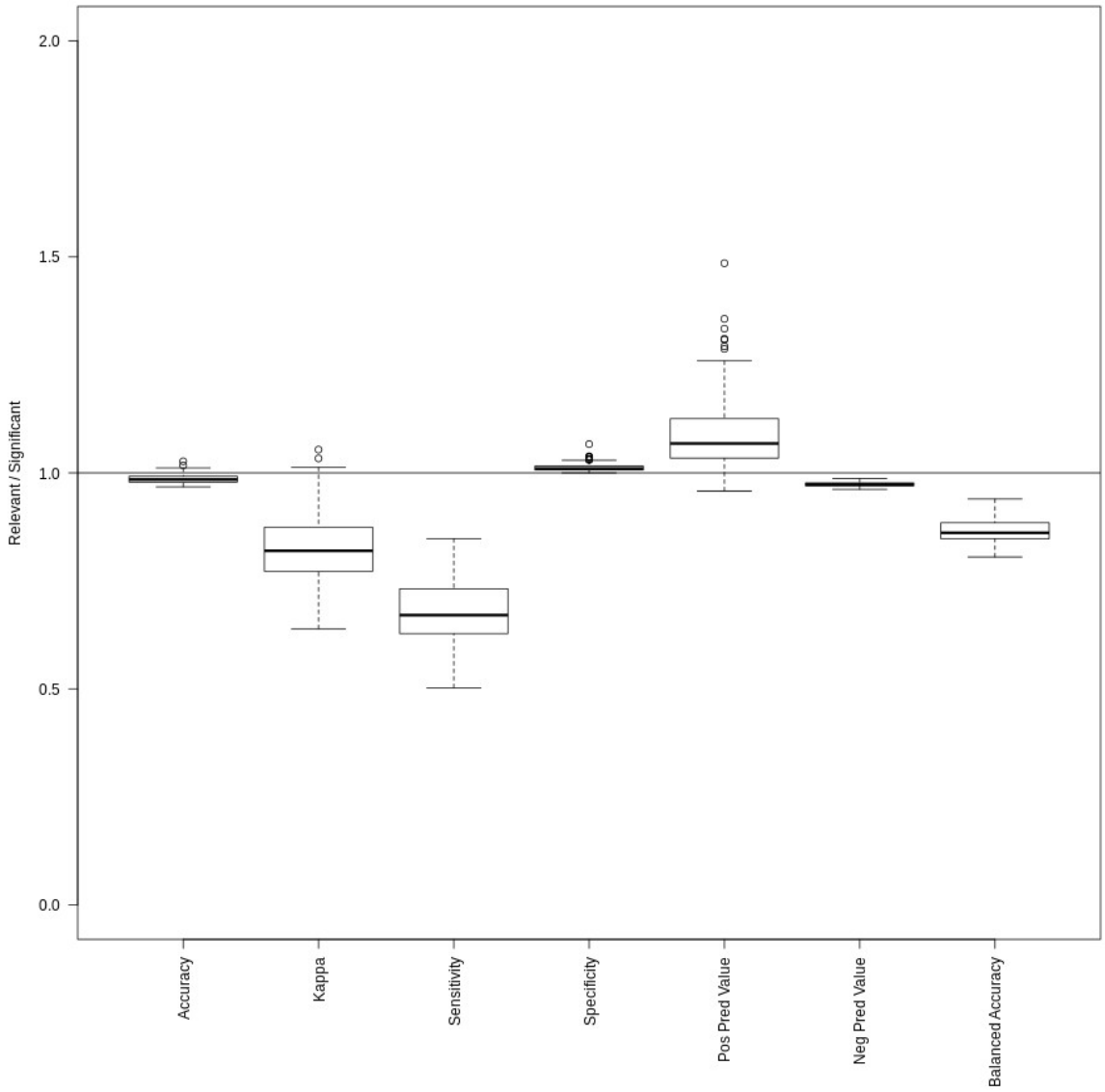


Figure 7: Boxplot of the ratio of several metrics between relevant and significant genes with 50 replicates per group. For every simulated dataset, a metric is derived for both relevant and significant genes compared to the population/original difference. The result of each metric from the relevant list of genes is divided by the result of each metric from the significant list of genes. If they perform equally, the value will be 1. If the relevant list of genes performs better, the value of the ratio will be above 1. If the relevant list of genes performs worse than the significant list of genes, the value of the ratio will be below 1. A vertical line at 1 visualizes where the equilibrium is situated.

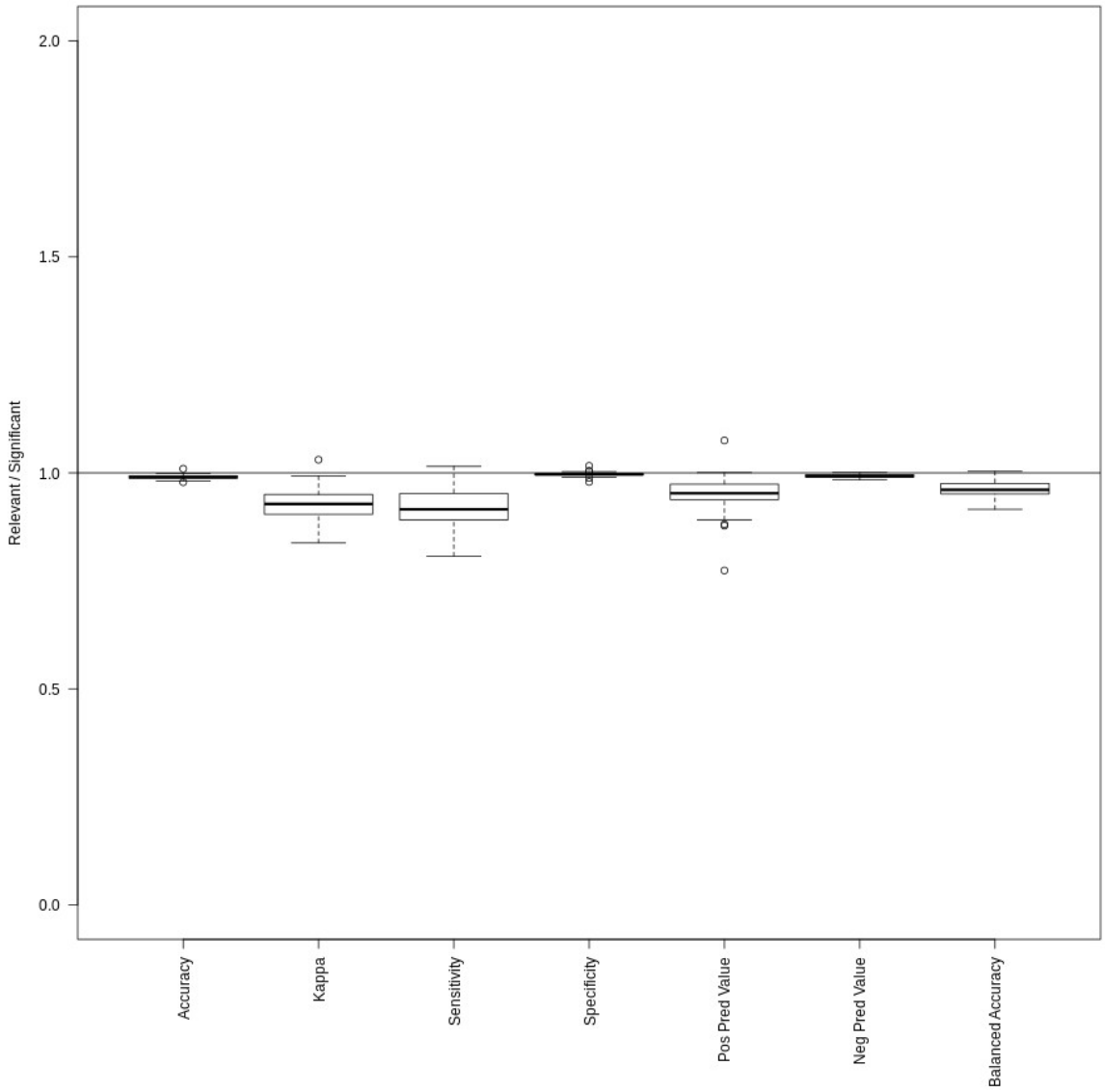


Figure 8: Boxplot of the ratio of several metrics between relevant-dubious and significant genes with 50 replicates per group. For every simulated dataset, a metric is derived for both relevant and significant genes compared to the population/original difference. The result of each metric from the relevant list of genes is divided by the result of each metric from the significant list of genes. If they perform equally, the value will be 1. If the relevant list of genes performs better, the value of the ratio will be above 1. If the relevant list of genes performs worse than the significant list of genes, the value of the ratio will be below 1. A vertical line at 1 visualizes where the equilibrium is situated.

21 replicates

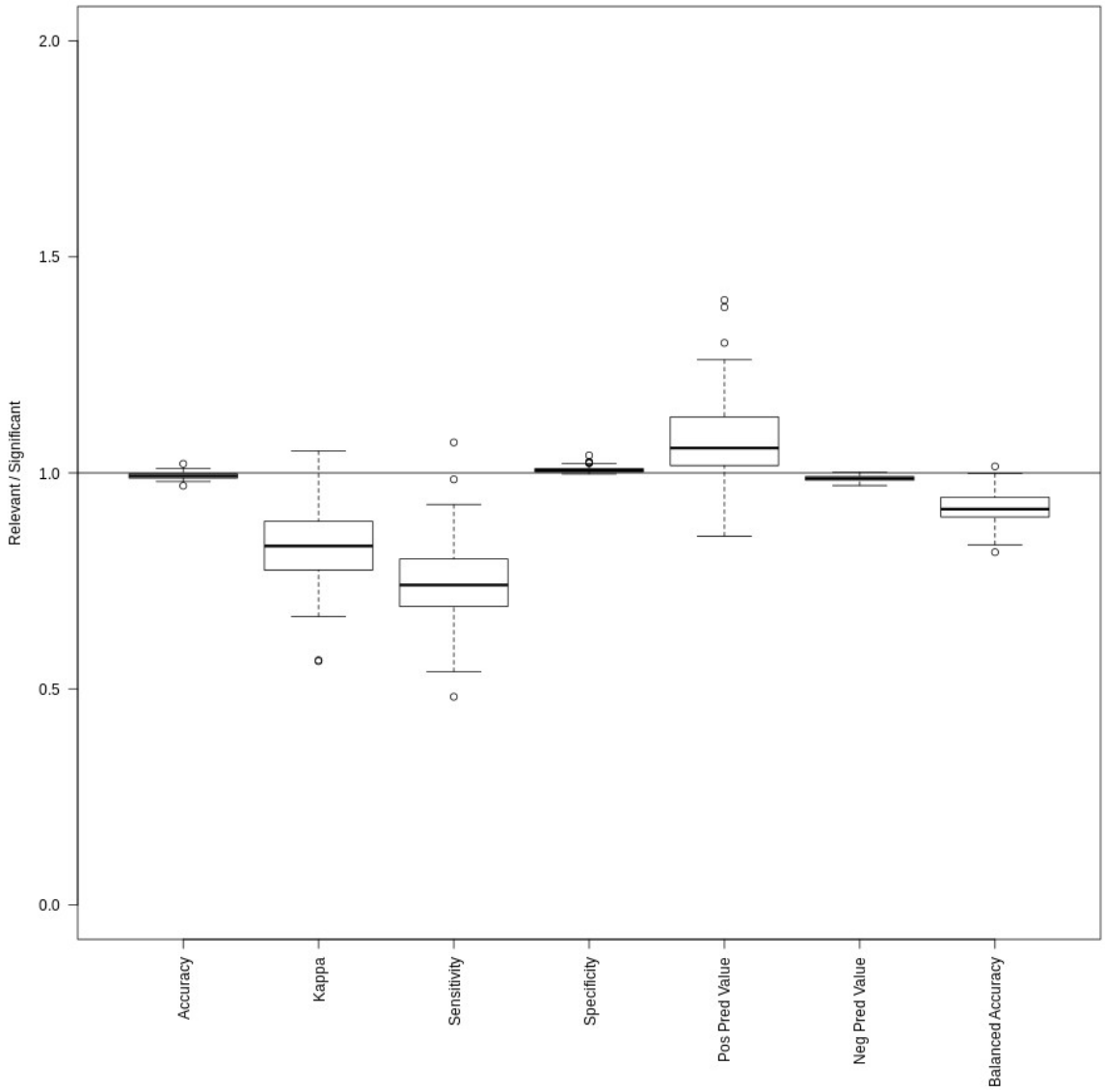


Figure 9: Boxplot of the ratio of several metrics between relevant and significant genes with 21 replicates per group. For every simulated dataset, a metric is derived for both relevant and significant genes compared to the population/original difference. The result of each metric from the relevant list of genes is divided by the result of each metric from the significant list of genes. If they perform equally, the value will be 1. If the relevant list of genes performs better, the value of the ratio will be above 1. If the relevant list of genes performs worse than the significant list of genes, the value of the ratio will be below 1. A vertical line at 1 visualizes where the equilibrium is situated.

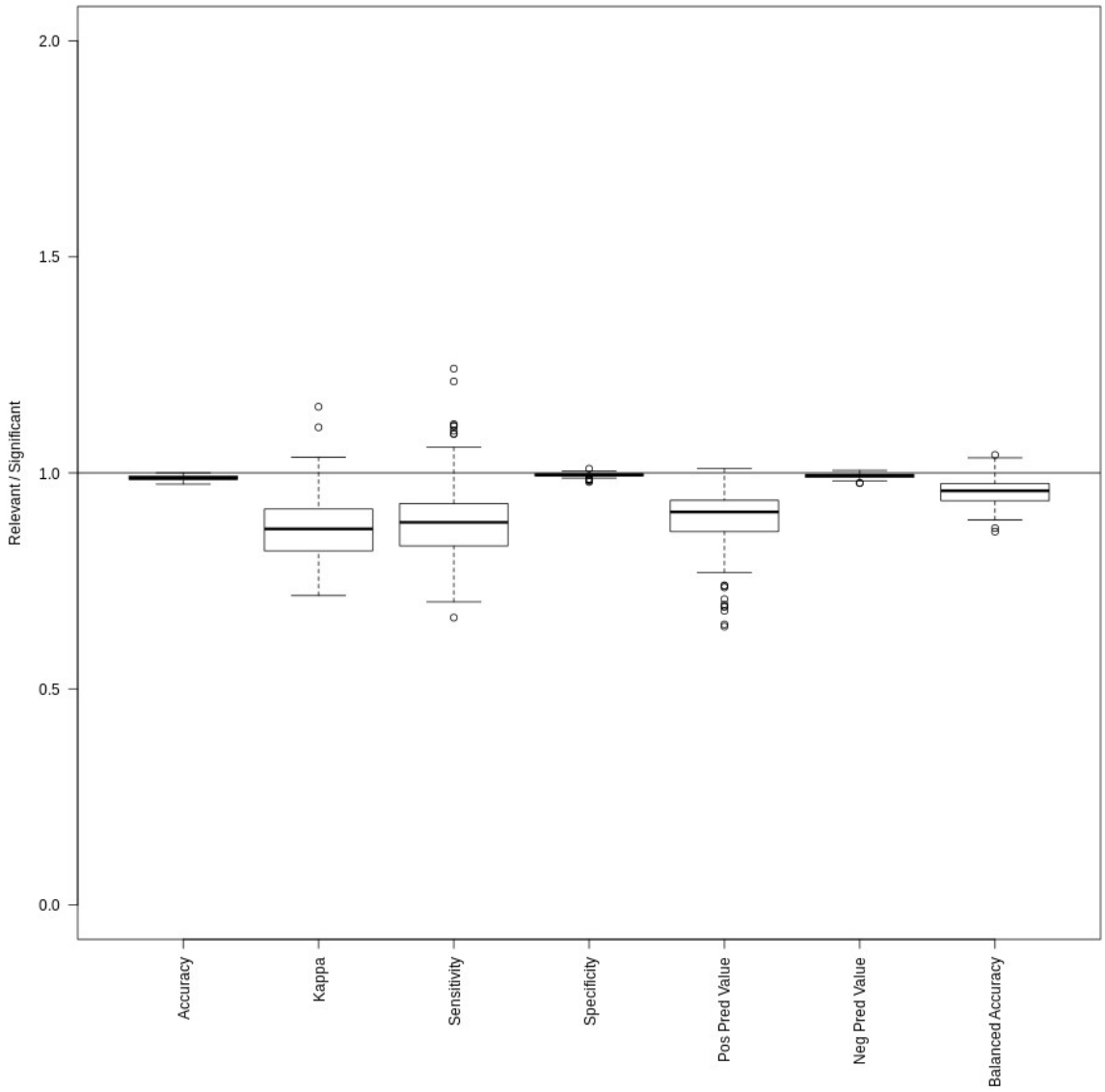


Figure 10: Boxplot of the ratio of several metrics between relevant-dubious and significant genes with 21 replicates per group. For every simulated dataset, a metric is derived for both relevant and significant genes compared to the population/original difference. The result of each metric from the relevant list of genes is divided by the result of each metric from the significant list of genes. If they perform equally, the value will be 1. If the relevant list of genes performs better, the value of the ratio will be above 1. If the relevant list of genes performs worse than the significant list of genes, the value of the ratio will be below 1. A vertical line at 1 visualizes where the equilibrium is situated.

9 replicates

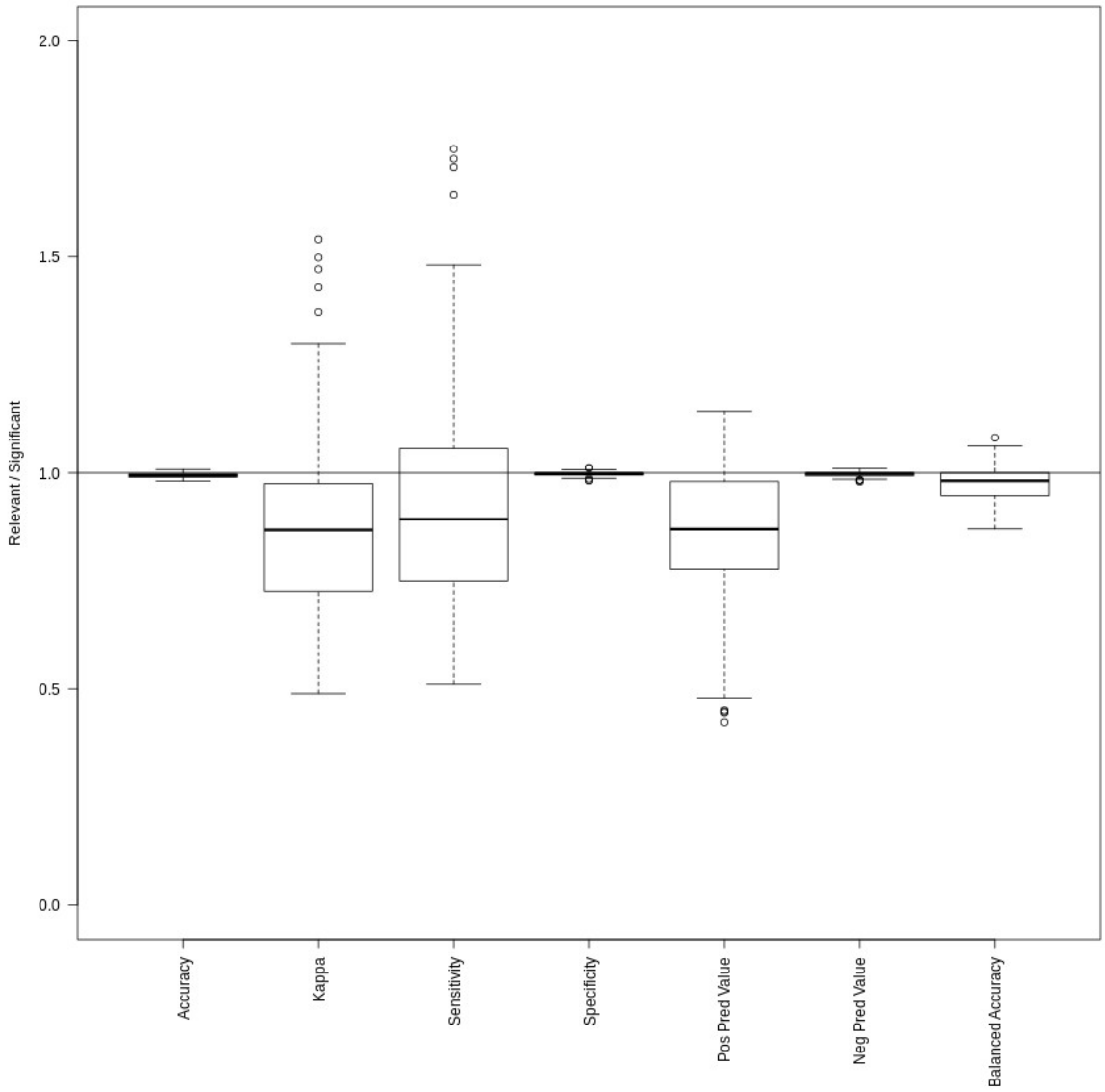


Figure 11: Boxplot of the ratio of several metrics between relevant and significant genes with 9 replicates per group. For every simulated dataset, a metric is derived for both relevant and significant genes compared to the population/original difference. The result of each metric from the relevant list of genes is divided by the result of each metric from the significant list of genes. If they perform equally, the value will be 1. If the relevant list of genes performs better, the value of the ratio will be above 1. If the relevant list of genes performs worse than the significant list of genes, the value of the ratio will be below 1. A vertical line at 1 visualizes where the equilibrium is situated.

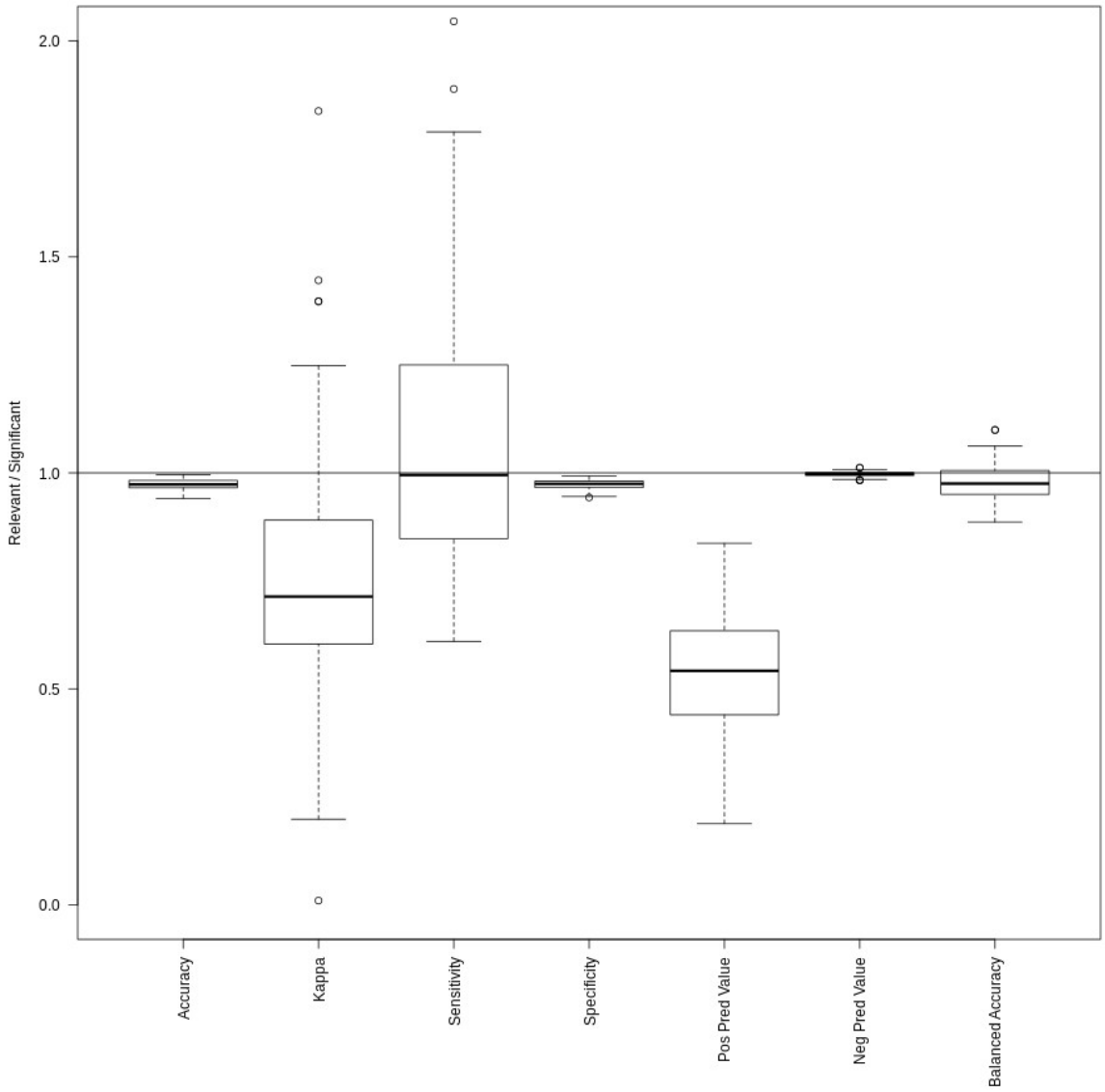


Figure 12: Boxplot of the ratio of several metrics between relevant-dubious and significant genes with 9 replicates per group. For every simulated dataset, a metric is derived for both relevant and significant genes compared to the population/original difference. The result of each metric from the relevant list of genes is divided by the result of each metric from the significant list of genes. If they perform equally, the value will be 1. If the relevant list of genes performs better, the value of the ratio will be above 1. If the relevant list of genes performs worse than the significant list of genes, the value of the ratio will be below 1. A vertical line at 1 visualizes where the equilibrium is situated.

3 replicates

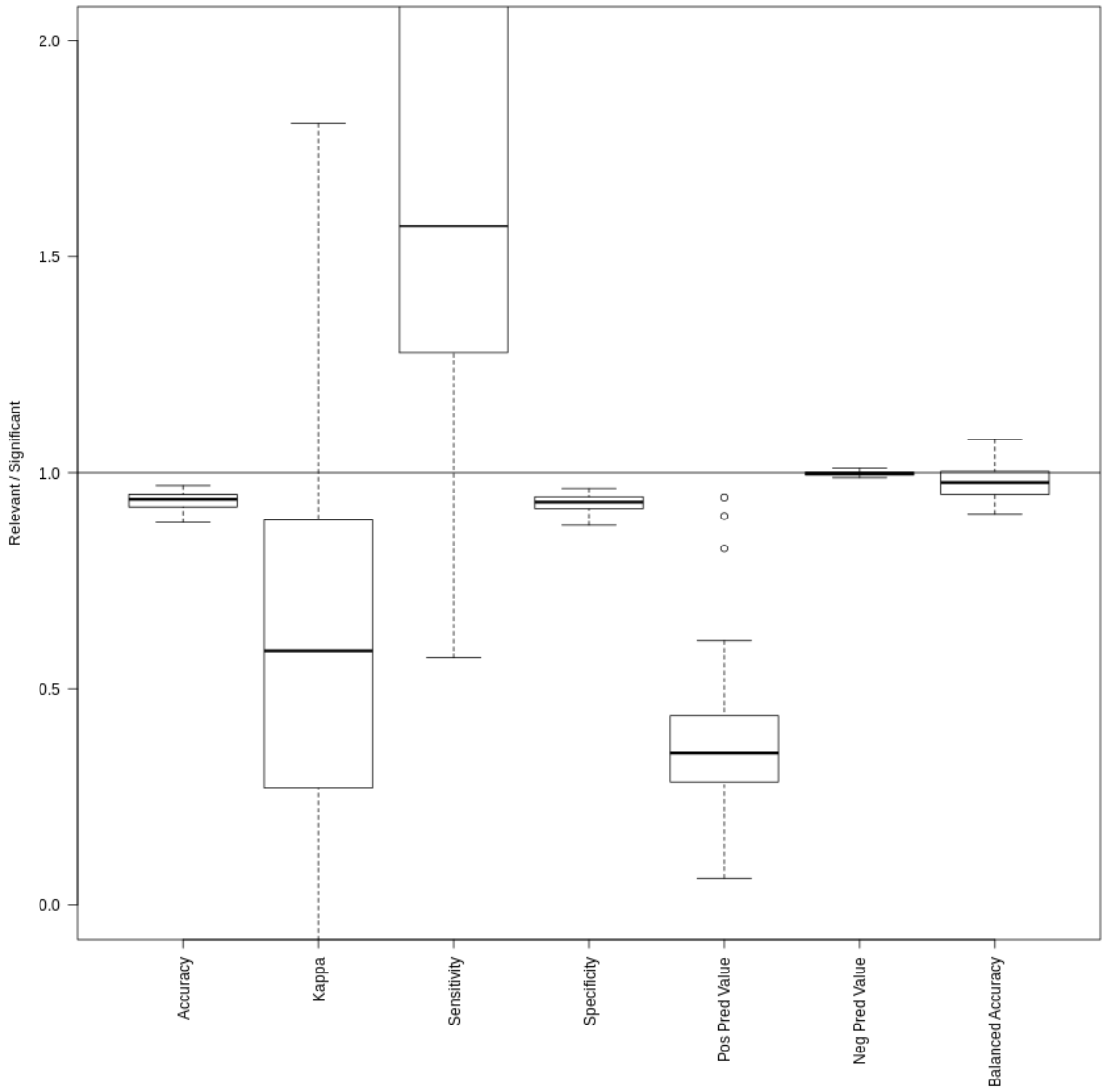
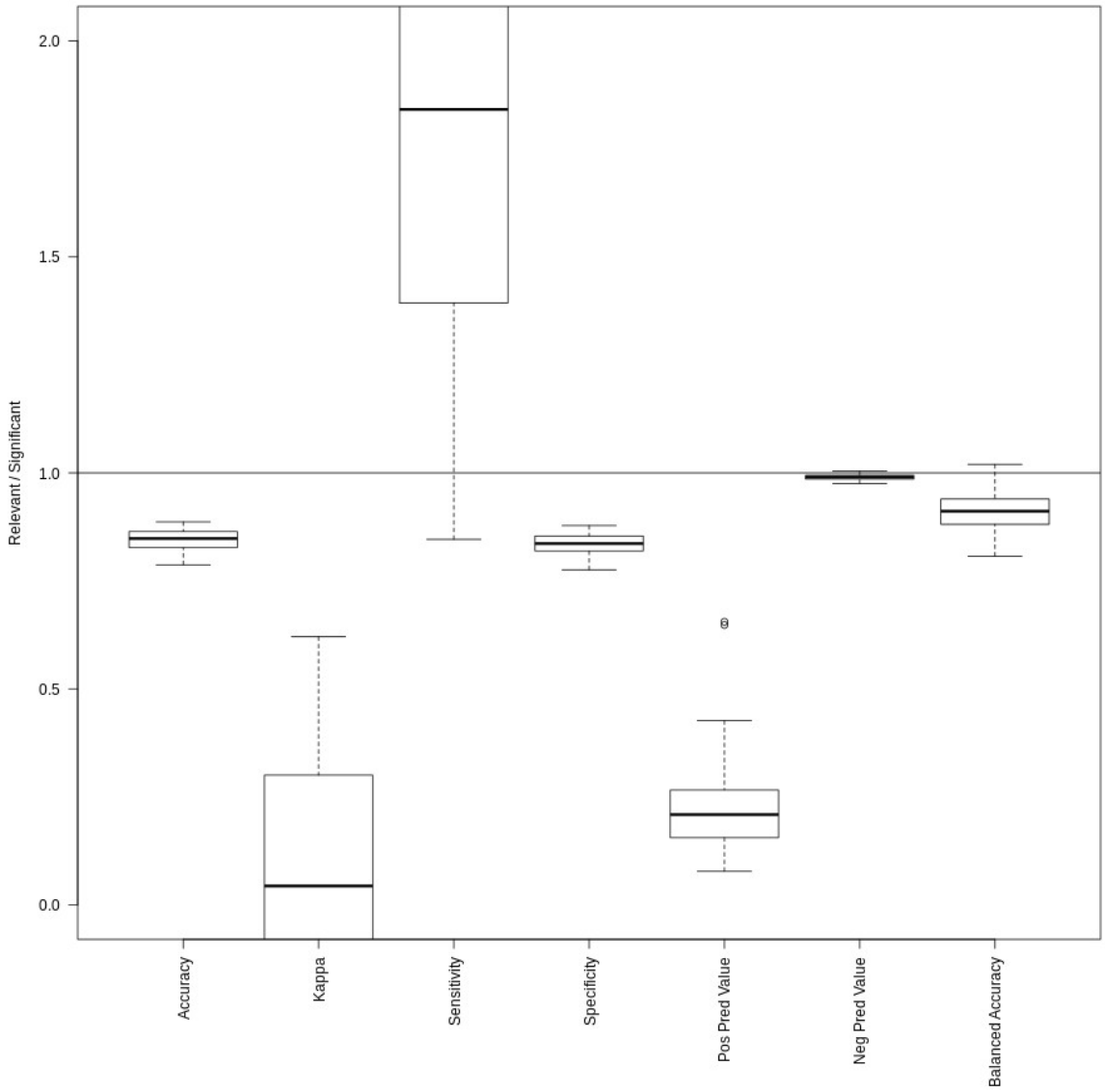


Figure 13: Boxplot of the ratio of several metrics between relevant and significant genes with 3 replicates per group. For every simulated dataset, a metric is derived for both relevant and significant genes compared to the population/original difference. The result of each metric from the relevant list of genes is divided by the result of each metric from the significant list of genes. If they perform equally, the value will be 1. If the relevant list of genes performs better, the value of the ratio will be above 1. If the relevant list of genes performs worse than the significant list of genes, the value of the ratio will be below 1. A vertical line at 1 visualizes where the equilibrium is situated.



126

127 *Figure 14: Boxplot of the ratio of several metrics between relevant-dubious and significant genes with 3 replicates per group.*
128 *For every simulated dataset, a metric is derived for both relevant and significant genes compared to the population/original*
129 *difference. The result of each metric from the relevant list of genes is divided by the result of each metric from the significant list*
130 *of genes. If they perform equally, the value will be 1. If the relevant list of genes performs better, the value of the ratio will be*
131 *above 1. If the relevant list of genes performs worse than the significant list of genes, the value of the ratio will be below 1. A*
132 *vertical line at 1 visualizes where the equilibrium is situated.*

Chapter 6:

General Discussion and Summary

Chapter 6: General Discussion and Summary

OMICs data has become the new tool to study all kinds of areas related to molecular biology: from cancer and its prognosis[1] to microbiology discovery[2]. Its use has become so universal that the data generation process has led to an enormous library of molecular data. Even though the size of data is crucial for research, its management and analysis are at least as important. In the case of this thesis, the data used originated in the HeCaToS project. This project spanned tens of compounds, with different doses, timepoints, and cell models. Even so, the abundance of data was not free of important limitations. The number of replicates was minimal (three). In addition, these were technical replicates, i.e., repeated simultaneously and originating from the same batch of cells. A great limitation of such design is that the results obtained are at best indicative of potential changes, but cannot be used as evidence per se. The full investment in the data generation process, while helpful, undermines its own results by not being able to have access to further resources to validate the results. A different approach might be to increase the number of both technical and biological replicates, even at the cost of decreasing the number of conditions tested. The danger of the latter approach is that, while being more scientifically sound, it might hinder its financing, as it may be evaluated at face value to be less ambitious than other broader-scoped projects.

As evidence of our intention to be biologically sound and more data skeptical, one can observe that in **Chapter 2** we applied *in vitro* techniques to validate the existence of the circular RNAs studied. Ideally, we would have not only validated their existence but ratified their effect at the toxicological level by either overexpressing or knocking down each of the transcripts described in the post-transcriptional regulation axes. This is especially relevant for the circRNAs themselves, which, even though they have been frequently studied since their hypothesized function as miRNA sponges[3], some bioinformatic analyses showed that this is rather the exception than the rule[4,5]. In addition, artificially modifying the expression levels of the studied circRNAs would help us discern whether those changes are simply a result or also a mechanism of the toxicity exerted by those compounds. Other modifications would have introduced more specificity in our novel method to detect and quantify circRNAs. For example, one of the main weaknesses of the remapping strategy is that circRNAs are not mapped based on their most characteristic property: the back-spliced junction (BSJ). This is because the sequences introduced to be mapped were structured in the classical or genomic 5' to 3' form. To this effect, no read that mapped the BSJ would be assigned to their circRNA sequence. One solution would be to truncate or split the original sequences in half and join the 3' end of the second half to the 5' end of the first one. In this manner, the characteristical BSJ would lay in the middle of the sequence. The reads mapping to the BSJ would not only increase the read count for each circRNA but would help to more correctly assess their expression.

This is due to the algorithms used by quantification software, which assign shared reads (reads that map to more than one location) by looking at the proportion of non-shared reads (reads that uniquely map to a genomic locus). This change in the transcriptome file would still allow quantifying circRNAs more broadly than CIRI2[6], as circRNAs without a BSJ read would still be identified, but simultaneously improve the quantification step for the circRNAs whose BSJ reads were sequenced.

As seen in **Chapter 2**, the integration of all the multi-omics is not a straightforward process. Even the study of two circular RNAs leads to a complex network of interactions that affect the translation of messenger RNAs, upon which the whole structure and function of the cell are based. For this reason, we proposed in **Chapter 3** to formulate an equation that would help quantify the number of transcripts available for translation. Even though from the toxicological view the dataset used (HeCaToS) involved certain complexities due to the various factors to consider (time, dose, or batch effects), it included different omics data extracted simultaneously from the same samples, thus a special design for a big dataset that is truly beneficial when aiming for a new crossomics (integration of multiple omics data) strategy. In addition, the fact that the RNA-Sequencing data was ribo-depleted (instead of using poly (A) capture) permitted the quantification of circRNAs, thus expanding the level of RNA types analyzed. Including important factors that affect the levels of free transcripts can become endless, not only for the calculations necessary, but also due to data that is not accessible at that moment in time (such as epigenetics or translatomics), and the different number of factors that affect each gene: some will be limited by transcriptional factors while others might be mostly regulated by histone acetylation. Thus, even to the best of our efforts, the results showed that the prediction of transcript availability was much more complex than a formulation derived from our current knowledge of the main drivers of post-transcriptional regulation.

This issue brought about our published work[7] encapsulated in **Chapter 4**, where we addressed the issue from a different perspective. First, we selected the proteomics expression as the target to be predicted. Second, we switched our method from human formulation to machine learning. These changes were sounder for several reasons. First, having proteomics data allowed us to directly evaluate the accuracy of our research. Second, machine learning is mainly limited by the data it uses. So, if we fed the data in the most comprehensive way possible, it would output the most optimized prediction, as was the case. Even so, one would acknowledge that the information available is still far from completely describing the molecular landscape of a cell. And so, even though the metrics showed a good prediction accuracy in terms of R^2, it is fair to wonder how much of the model accuracy is due to being trained in untreated conditions: if the model has been trained with the same protein on a different sample, outputting the same value that was trained on is a safe bet which would also lead to a potentially good correlation score. In

any case, if the conditions were met, the ideal situation to generate the best model would be to make a model for each protein (to be predicted), where the variables encapsulate all other molecules. In this manner, all potential effects would be considered, from miRNAs to transcription factors, and thus if such an abundant source of data was available, one could in principle predict any molecule's quantity based on all other variables. Of course, such a model would only be useful if one would also have access to the huge number of variables that trained the model, and thus it would probably be more of a proof of concept of whether we can completely map and understand the molecular biology of the cell. In addition, such a model would allow us to modify the abundance of specific molecules in order to study the effects on all the components of the cell. A model with these characteristics is, in the author's opinion, still not possible nowadays due to technological limitations, and so one would caution in favor of skepticism whenever projects aim to develop a model of similar characteristics to predict all toxic effects inside a cell.

This is though not a leeway as a scientist to presume that using machine learning is per se a universal method to solve current scientific problems as long as there is enough data. As an example, just feeding a model with all published transcriptomics data and assumed results derived from that will mostly result, in the author's opinion, in a big model with very stochastic results. This is because the human component in the decisions taken, be it due to different methods, thresholds, or prejudices used, will ultimately lead to different conclusions, and these decisions would all be mixed into a single model. For this reason, even though one of the main limitations of **Chapter 5** is that most of the observations fed into the AutoRel were labeled by the author of this thesis, it also ensures a more replicable model of the reasoning behind those decisions, that is, that the model can more accurately label observations in concordance to how it has been trained. As mentioned in the chapter, this would represent a further step in the direction of automatizing manual filtering steps in RNA-Sequencing analyses, as was the case for our collaboration in R-ODAF[8]. In any case, it can be argued that different strategies could have been applied to represent more extensively the expert decision-making. The most direct manner would have been to introduce all possible interpretations from the maximal number of scientists in a single model. One could also instead have built a model per researcher, and a global algorithm would be used to assign a label based on the majority vote from all those distinct models. A different strategy could have been to only label as relevant genes that have been previously proven to be differentially expressed via several distinct confirmation experiments, or at least train these models more extensively by using simulation datasets, where the differentially expressed genes (DEGs) are known beforehand. A limitation of the use of simulated datasets, though, is that it selects its DEGs by fold change alone, which is a limited metric to use. It is based on the comparison of means, which are easily tilted by the presence of outliers. In addition, simulated datasets do not take into account low quantification values, and thus a difference of a few reads

is considered a DEG, as small values have a higher chance at random to present a higher fold change value. In practice, this can easily be the result of an irrelevant small variation of expression.

It is of interest to note the use of previous machine learning approaches to evaluate differentially expressed genes. For example, LASSO regression analysis is usually utilized after a differential expression analysis to further filter the number of DEGs[9]. The use of random forest has been recently suggested for the selection of important genes in microarray data[10]. The theory is based on the hypothesis that if two conditions (control vs treatment) are able to be separated by the data available (gene expression data), an artificial intelligence model should be able to differentiate between them. In addition, the commonplace existence of variable importance values would help identify which variables (genes) are the most responsible for the differences/splitting between the two conditions. We consider the evaluation of several classification algorithms (including LASSO and random forest) of great interest to discern to which extent these algorithms can correctly identify such genes, by the use of validated gene expression results.

Summary

Novel bioinformatics approaches have been applied to elucidate the function of circular RNAs in the mechanism of action of several cardiotoxicants through their importance in the post-transcriptional regulation. We further our investigation of this regulation via a formulation to predict the quantity of coding transcripts that are available for translation. The complexity of this task makes necessary the introduction of machine learning strategies, which help us predict and impute proteomics values in untreated samples. Finally, we make use of the artificial intelligence methods to classify genes according to their biological relevance based on expert labeling.

Bibliography

1 Lin, E. *et al.* Integrative Analysis of the Genomic and Immune Microenvironment Characteristics Associated With Clear Cell Renal Cell Carcinoma Progression: Implications for Prognosis and Immunotherapy. *Front Immunol* **13**, 830220, doi:10.3389/fimmu.2022.830220 (2022).

2 Grujcic, V., Taylor, G. T. & Foster, R. A. One Cell at a Time: Advances in Single-Cell Methods and Instrumentation for Discovery in Aquatic Microbiology. *Front Microbiol* **13**, 881018, doi:10.3389/fmicb.2022.881018 (2022).

3 Ma, C. *et al.* circRNA CDR1as Promotes Pulmonary Artery Smooth Muscle Cell Calcification by Upregulating CAMK2D and CNN3 via Sponging miR-7-5p. *Mol Ther Nucleic Acids* **22**, 530-541, doi:10.1016/j.omtn.2020.09.018 (2020).

4 Guo, J. U., Agarwal, V., Guo, H. & Bartel, D. P. Expanded identification and characterization of mammalian circular RNAs. *Genome Biol* **15**, 409, doi:10.1186/s13059-014-0409-z (2014).

5 You, X. *et al.* Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. *Nat Neurosci* **18**, 603-610, doi:10.1038/nn.3975 (2015).

6 Gao, Y., Zhang, J. & Zhao, F. Circular RNA identification based on multiple seed matching. *Brief Bioinform* **19**, 803-810, doi:10.1093/bib/bbx014 (2018).

7 Ochoteco Asensio, J., Verheijen, M. & Caiment, F. Predicting missing proteomics values using machine learning: Filling the gap using transcriptomics and other biological features. *Comput Struct Biotechnol J* **20**, 2057-2069, doi:10.1016/j.csbj.2022.04.017 (2022).

8 Verheijen, M. C. *et al.* R-ODAF: Omics data analysis framework for regulatory application. *Regul Toxicol Pharmacol* **131**, 105143, doi:10.1016/j.yrtph.2022.105143 (2022).

9 Miduo Tan, G. H., Jingjing Chen, Jiansheng Yi, Xi Liu, Ni Liao, Yi Hu, Wei Zhou, Qiong Guo. Construction and validation of an eight pyroptosis-related lncRNA risk model for breast cancer. *American Journal of Translational Research* **14(5)**, 2779-2800 (2022).

10 Hediger, S., Michel, L. & Naf, J. On the use of random forest for two-sample testing. *Comput Stat Data An* **170**, doi:ARTN 107435

10.1016/j.csda.2022.107435 (2022).

Impact

Scientific and Social Impact

Currently, regulatory agencies rely on the use of animals for testing drugs for several reasons. First, the most accurate way of testing drugs would be testing them in humans, which is not ethically possible. Second, modeling everything that happens in our bodies in a Petri dish is nowadays unfeasible: the complexity of a human body requires 30 trillion cells to keep functioning. In addition, each of those cells also involves its level of complexity: even a simple yeast cell contains 42 million proteins[1], which does not include other essential molecules like sugars and fatty acids. Even so, recent incentives (such as the regulation (EC) No 1223/2009 of the European parliament[2]) have pushed the scientific community to search for alternative testing methods without the use of animals. Specifically, the area of Toxicology is affected, as it studies the potentially toxic effects of drugs both before and after being released on the market. A popular method for studying the effects a compound can have in humans is by testing them *in vitro*, that is, by exposing human cells. Doing so helps to narrow the bridge between what is being tested (human cells outside the body) and the actual goal of the study (human cells inside the body) when compared to animal testing, where the model and the end goal belong to different species. Nowadays, the development of induced pluripotent stem cell (iPSC) technology allows reverting any human skin cell to another tissue type (such as cardiac cells that contain the same DNA as the donor of the skin cells), without the need for surgery or invasive biopsies. Although some drugs may kill cells by simply destroying the membrane that encapsulates the cell, most of them disturb the cell in more subtle ways. One of these ways is the deregulation of the number of proteins synthesized by a cell. Proteins are essential molecules, as they perform most of the cell functions. For this reason, disturbing or blocking their production can lead to the disruption or death of a cell and/or the ones that depend on it. The processes that lead to the making of proteins involve mainly RNAs, molecules that work as messengers from DNA to proteins. Therefore, in Toxicogenomics, studying how specific treatments can affect these molecules can help understand their mechanisms of toxicity.

In **Chapter 2**, we assessed how a recently discovered class of RNA, called circular RNAs (circRNAs), are disturbed in heart cells by known toxicants. These circRNAs have been hypothesized to regulate microRNAs (miRNAs) by letting the latter bind to the former. When messenger RNAs (mRNAs) are not bound to miRNAs, they can provide the instructions to produce proteins. When miRNAs are occupied binding circRNAs, they are not able to bind to mRNAs. For this reason, by assessing changes in the number of circRNAs, miRNAs, mRNAs, and proteins; we helped better understand how these compounds (that are still in use) are being toxic in the human heart without the use of animal testing. Going further into the thesis, we realized that knowing how many proteins are in a cell is crucial to

understanding how a drug (or disease, or any other perturbation) affects a cell. Unfortunately, the technology used for doing it, mass spectrometry, does not measure all proteins in a cell. Instead, researchers tend to use transcriptomics, which exhaustively measures the quantity of RNAs. Nevertheless, the number of RNAs and the number of proteins do not always perfectly correlate with each other.

In **Chapter 3**, we designed an equation to estimate how many RNAs are available to produce proteins. We did that by counting the total number of mRNAs and subtracting the ones that will be affected by miRNAs, but only those miRNAs that are not binding to other RNAs (like another mRNA or circRNA). Nonetheless, the formula, which is focused mainly on RNA molecules, demonstrated an added value for only a subset of proteins. As a result, in **Chapter 4**, we went a step further. We built a large dataset with RNAs and their corresponding proteins and trained a machine learning model to predict the latter. Machine learning algorithms "learn" how to predict values by looking at how other similar values behave. Using our data, our model predicted well the increases and decreases of proteins, which can help others to predict how many proteins there are in a sample of cells based on how many there are in similar ones.

As mentioned before, in the area of Toxicogenomics it is of great interest to study the changes happening in a cell. Transcriptomics is exceptionally good at counting how many RNAs there are, consequently using this technology helps us understand which molecules change in quantity due to a specific cause. The statistical tools used to detect changes, though, do not work without fault. For this reason, experts in this technology can manually detect these errors. On the flip side, this manual curation is pretty time-intensive when taking into account the number of genes to be evaluated, and requires specialist knowledge to do so. That is why, in **Chapter 5**, we again trained a machine learning model. In this case, we taught the model to recognize the profile of genes that are typically of interest to the researcher. We built several of them with different characteristics and selected the best one, which we named 'AutoRel'. Even though AutoRel was not flawless, it showed improvements by removing genes that were not of interest.

In an era of increasing societal pressure against animal testing, added to the inherent shortcomings of animal assays, regulatory agencies need to reevaluate their historical procedure of risk assessment. With the rapid development of methodologies that allow the analysis of the complete set of biological entities in a cell exposed to any substance, regulators will need both a better understanding of all the complex interactions behind molecular biology and powerful data analysis tools to integrate them. The work of this thesis contributes to this necessary transition toward a next-generation risk assessment. This is achieved by the discovery of new changes that happen when a toxic compound affects a cell, predicting protein measures that are usually unknown with artificial intelligence, and filtering results in an automated way to

have a better understanding of the changes that occur in a cell. In aggregate, this contributes to assess more accurately how toxic a drug is, making the use of treatments more safe and reliable.

Bibliography

1 Ho, B., Baryshnikova, A. & Brown, G. W. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Syst* **6**, 192-205 e193, doi:10.1016/j.cels.2017.12.004 (2018).

2 *Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products (Text with EEA relevance)*, <http://data.europa.eu/eli/reg/2009/1223/oj> (2009).

Juan Ochoteco Asensio was born on the 29th of March of 1994 in Barcelona, Spain. In 2012, he started his Bachelor in Biochemistry at the University of Navarra. During his bachelor's, he took an internship in preanalytical and analytical phases in a clinical laboratory at the Biochemistry Lab of the Hospital Clinic de Barcelona. In addition, he attended a basic proteomics course at the Centre of Applied Medical Research in Pamplona (University Clinic of Navarra). For the bachelor's thesis, he worked on the analysis of the astroglia and microglia activation in the aging brain for the study of neurodegenerative diseases at the Centre of Applied Medical Research in Pamplona (University Clinic of Navarra).

After finishing his BSc degree in 2016, he started his Master's degree in Bioinformatics at the Faculty of Biosciences of the Autonomous University of Barcelona. For his master's thesis, he worked on a project that focused on re-annotating the genome of *Drosophila buzzatii*.

In March 2018, he started working as a PhD student at the Toxicogenomics department of Maastricht University under the supervision of Prof. Dr. Jos Kleinjans and Dr. Florian Caiment. His PhD project took advantage of the big datasets generat ed in the HeCaToS project, the latter of which aimed at developing integrative in silico tools for predicting human liver and heart toxicity. The results achieved throughout these four years are presented in this thesis.

List of papers

Published

- Verheijen MC, Meier MJ, Asensio JO, et al. R-ODAF: **Omics data analysis framework for regulatory application**. *Regul Toxicol Pharmacol*. 2022;131:105143. doi:10.1016/j.yrtph.2022.105143
- Ochoteco Asensio J, Verheijen M, Caiment F. **Predicting missing proteomics values using machine learning: Filling the gap using transcriptomics and other biological features.** *Comput Struct Biotechnol J*. 2022;20:2057-2069. Published 2022 Apr 22. doi:10.1016/j.csbj.2022.04.017

In submission process

- **AutoRel: Machine Learning model for selecting differentially expressed genes.** Juan Ochoteco Asensio, Jelmer Faber, Twan van den Beucken, Marcha Verheijen, Florian Caiment.

Preprints

- **Quantifying the number of translatable transcripts through the use of OMICs involved in post-transcriptional regulation.** Juan Ochoteco Asensio, Jos Kleinjans, Florian Caiment. bioRxiv 2022.06.20.496876; doi: https://doi.org/10.1101/2022.06.20.496876

Other contributions

- Evaluation of transcriptomic quantification tools – **Bioinformatics & Systems Biology Conference 2019**
- P11-06 Unraveling novel gene networks in anthracycline-induced cardiotoxicity. JG Faber, JO Asensio, F Caiment, T van den Beucken. Toxicology Letters 368, S164. 2022
- SOC-IV-09 Investigating the role of nuclear receptors in valproic acid-induced liver steatosis. K Guo, J Faber, JO Asensio, F Caiment, T van den Beucken. Toxicology Letters 368, S54-S55. 2022

Acknowledgments

Let's be honest. Most people (or at least I would) will first come to this section to see if their names are here. Indeed, this is one of the lessons you learn in academia: we may like science, but that curiosity is also present in the human affairs. What is this section, you may ask? This is where I acknowledge that all these pages would not have been possible without so many people that have brought me to this moment, and for that I would like to thank them.

First, I would like to thank **Florian Caiment**. Thanks for teaching me so many valuable lessons about how to do research, which involves creativity (which we both share) and deadlines. I am grateful to have met a rigorous scientist in so many aspects, which has helped me being open and critical about the work we do. Interestingly enough, you also have fun while trying to make the most out of your work. All while not having sacrificed everything for your work. That sounds a pretty good future if I ever get there. And of course, it was so fun to always learn new French idioms with you, such as "Couper l'herbe sous le pied de quelqu'un".

Prof. Dr. Jos Kleinjans, thank you for giving me the opportunity to be a part of the last year of the HeCaToS project and providing the funding for my doctoral contract. Ever since your departure from the department, I have missed your productive motivation to move forward in delivering results.

I would like to acknowledge all the members of my assessment committee: **Prof. dr. H.J.M. Smeets, Prof. dr. J. Quackenbush, Prof. dr. R. Peeters, Dr. S. Hayat**, and **Dr. L. Eijssen.**

I also want to thank **Dr. T. van den Beucken** for his abundant help and contribution in the discussion of my work, giving always a realist criticism from a biological perspective, so necessary in the Bioinformatics field. Needless to say, I need to add also **Dr. M. Verheijen** here, as she has been there from the start helping me getting to understand the HeCaToS project, while also giving a friendly environment to the department. You literally celebrated my birthday on my second day of work! Even if our party attendance ratios are completely opposite. **Dr. D. Jennen** for your suggestions and hall conversations. **Dr. J. Briede** for your educational role. **Dr. J. Krauskopf** for your bioinformatic tips and bits. **Christa Graus** for so much help in everything needed in my day to day work. For our conversations at the office. Ik wens u nog success verder. Misschien zie ik je in Catalonië. **Duncan** for being the wisest of all of us, and make us wonder whether we are actually worth a doctoral degree. **Rosella**, la postdoc italiana più forte, grazie per la tua energia e immediatezza. Quei due sono sicuramente una combinazione potente. Saluta il ragazzo di Nazaret. E, come sempre, Forza Italia! **Margot**, ik heb je voor een korte tijd ontmoet, maar ik zie dat je heel energetisch bent, en dat is heel goed voor deze afdeling. Bedankt voor al onze gesprekken.

I would also like to thank my paranymphs. First, **Yannick Schrooders**. What can I say? I cannot imagine how my junior years could have been without you being part of the Gossip Guys group. It would have been then difficult to think how many things we would have done together, from you coming to my place in Barcelona, to me marrying you soon. In any case, you know I will there to give some of my energy for the Spirit Bomb. Weest altijd jezelf, en doe je ding! Mem**, Antoni Vallbona Garcia**, que

daixonare de tu? M'has vist amb xubec i fins i tot mesclar ous amb cargols. A vegades et sembla que cerc na Maria per sa cuina, i vols que a la defensa vagi gat perque mira que en soc de poma! De totes maneres, gracies per estar amb mi tot aquest temps. La nostra amistat demostra que persones molt diferents poden viure juntes, i esper que pugui saber de tu fins al final en aquest mon terrenal. Porta't bé, i un dels millors consells: digues la veritat (o com a minim no menteixis), i estima (o com a minim no odiïs).

I would like to thank **Jarno** for letting me have his place in the office. Thank you **Evelyn** for being so direct with everyone, to the point of leaving the office because of me. Thank you **Rajinder** for offering your house for parties (Raji shots!) and cooking such delicious foods (that I was able to try), in addition to let us know that cricket is a thing. Thanks to **Tim**, who helped me so much in the beginning of the doctorate with programming, and for sharing so many memories together. Hopefully your hockey skills are unmatched, and you keep reading as much as you do. **Manon Marjoleen Jacqueline**, what a fun time was to have you around in the office! So much energy and feminine taste in our otherwise male environment. And so much coffee and food of mine in your table! **Marta**, maggari, sei l'ultima collega che avevo in ufficio prima di essere estradato in isolamento e reclusione. Spero che le tue piante crescano alte e sane, e apprezzo così tanto le nostre lunghe conversazioni su così tante cose importanti. Mantieni intatta la tua curiosità e cerca sempre la verità! **Sean**, wat een gast hoor! Zo open om over alles te praten (en grappen makken). Ik wil jou danken voor al de tijd die je hebt in onze vriendschaap geinvesteerd, en ik wardeer je openheid. Je bent echt slim (maar wees nederig). De expert in belegging en motors! Gedraag je, en ik hoop dat we nog meer kunnen elkaar leren kennen in de toekomst. Thanks to all the interns that I have also shared a room with. Unfortunately, my memory is really bad, so thanks to Irene, Wendy, the Destiny 2 guy, and so many others that were fascinated with my madness. Also I remember now **Nkoli**, the tallest Brittish woman I had ever met, and almost as unhinged as me.

Now, time to thank other (ex)PhD students. **Heloïse**, enchanté de vous rencontrer dame française! Et maintenant maman apparemment. **Nhan**, the expert judo and colleague of HeCaToS. We started and will finish (God willing) the same year. Thanks for all the help throughout this time. Thanks for the Vietnamese transparent tacos and tiếng mèo kêu. Bạn là một người phụ nữ mạnh mẽ. Tiếp tục theo cách đó! **Daniela**, from the Spanish region of Portugal. Esta linda senhora é na verdade uma PhD agora! Espero que você seja o melhor na Suíça e continue rindo desse seu jeito característico. E lembre-se, o Amor e a Verdade prevalecerão! E eles realmente existem! E como você sempre disse, tudo ficará bem no final. **Jelmer**, je bent echt lief. Bedankt voor al je samenwerk, zonder jou zou ik deze thesis niet beeindigt. We zijn niet altijd eens met elkaar, maar nog heb ik echt genieten van deze tijd met jou. Wees altijd grappig and raar zoals altijd, met jou dansen in de gang en expert racen in Mario Kart. **Nikki**, de meest briljant Nederlans meisje die nog een PhD doet in ons afdeling. Bedankt dat je de meest normale van ons bent, en tegelijkertijd ook zo grappig en leuk het is om met jou te zijn. Zeg aan **Mickey** dat ik nog hem herinner (en dat hij raketten schoon maakt). Wijchen!. **Nicolaj**, der neue Deutsche in der Stadt. Schön, Sie auch kennenzulernen, und hoffentlich können Sie in einem so interessanten Büro überleben. **Somaieh**, جوان ترین دختر در TGX. مواظب خودت باش من برای مادران باشد که این جوان ها را تا پایان دکترا شاد کنید. احترام زیادی قائلم. **Julia (& Frank)**, thanks for everything! So many deep conversations, your sweetness, your acrobatic skills, your work ethic, your wedding invitation, your heart blessings, your wine/knife selling, and so many other things. Thanks for your yeehaw time. **Qingfeng, Zhiling, Yueqin, Kaidi**, and **Na**:向所有中国女孩致敬。不幸的是，我总是很难记住你所有的名字。但是感谢中国的糖果！ **Milena,** dziękuję za informacje o sytuacji bezrobocia w Holandii. **Brian, n**imekujua kwa muda mfupi tu,

lakini ninathamini tabasamu lako na ukimya wako, ingawa kama Mhispania ningekuhimiza kuzungumza zaidi! **Saad**, میں بھی آپ کو صرف تھوڑی دیر کے لیے جانتا ہوں، لیکن کسی بھی بایو انفارمیٹیشن کا ہمیشہ خیرمقدم ہوتا ہے۔ امید ہے کہ آپ اس ملک کو برفانی سویڈن کے مقابلے میں کم از کم تھوڑا گرم پائیں گے۔.

Salutations to other PhD's I have met from other departments: **Grzegorz** (jedyny inny katolicki dr. uczennica, która była na przyjęciach i mszy świętej), **Jana** (das klügste Mädchen, das fortschrittliche Einzelzell-Transkriptomik verwendet, um Menschen mit der Schönheit (ihrer Handlungen) zu beeindrucken), **Niccolo** (sei pazzo come me, goditi la vita negli Stati Uniti e continua a essere il ragazzo più sicuro di sé che abbia mai incontrato finora), **Delia** (arriba la gente del acueducto, eres genial y muy divertida, acuerdate de pórtarte bien y deja a la gente con un poco de sangre querida, espero que mis plaquetas te traten bien), **Julia** (che energica ragazza italiana, spero che tu possa continuare a scalare pareti sia fisicamente che nella vita)**,** and **Celia** (bona damisela de Barcelona, espero que el teu temps a Maastricht te'n recordi encara que sigui una mica a Barna).

Thanks to the Twitch streamers **JPStudyTime** and **Sarayoposita** for their digital company in my working from home time.

Gracias a mi familia por estar siempre ahí. Sé que siempre podré volver a casa a descansar y disfrutar de la buena vida hasta que me sienta inútil y tenga que trabajar de nuevo. Para que quede patente, ahí van los nombres: **Juan Ignacio, María Montserrat, Marcos, María, Isabel, Luis, Beatriz, Ignacio, Sofía, y los tres hermanos que nos esperan en el cielo.** Os conozco desde hace ya tiempo, así que son demasiadas anécdotas y de todas formas no hace falta que os coma las orejas.

Bedankt **Melanie** om zo liefdevol met mij altijd te zijn. Je herinnert me van de liefde van God: bemminen zoals ik ben.