Master of Science in Omics Data Analysis

Master Thesis

FUNCTIONAL ANNOTATION PIPELINE FOR THE ANALYSIS OF MICROARRAY DATA WITH REGARDS TO SUBCELLULAR PROTEIN LOCALIZATION

by

Maria Noguera Vila-Masana

Supervisor: Lluís Ribas de Pouplana, Gene Translation Laboratory, IRB

Co-Supervisor: Mireia Olivella, Department of Systems Biology, UVIC

Department of Systems Biology

University of Vic – Central University of Catalonia

[September 2017]

ABSTRACT

Subcellular protein localization prediction is a bioinformatics approach (faster and cheaper than experimental procedure) to determine the targeting location of proteins into the cell organelles. Therefore, many programs exist to address this issue, but TPpred is the most up-to-date option regarding mitochondria targeting prediction. On the other hand, the development of a *Drosophila melanogaster* transgenic model to study mitochondria-related diseases in humans has brought up many questions. Some mitochondria mechanisms are still not explored.

The translation pathways present in mitochondria have similarities to the ones present in the cytosol. But they are older, essential in the origins of live and also in the eukaryote cell development from prokaryotes. A new paralog of the seryl tRNA synthetize enzyme that catalyses aminoacylation of $tRNA^{Ser}$ was found in insects. The SLIMP protein, which is not a Seryl tRNA synthetases(SRS) anymore, is now known to interact with SRS2 and LON protease, a fact that can lead to a modulation in both mitochondrial translation and replication, respectively. A microarray of the knock-down of SLIMP expression in S2 *D.melanogaster* cells was performed and their analyzed row data gives rise to the current pipeline reported.

The functional annotation pipeline developed is a new analysis procedure to describe differences in transcriptome considering mitochondria targeting status in a microarray-wide approach. Proportions tests is performed, which in our data did not report a significative increase or depletion of mitochondria targeted transcript products under the SLIMP knockdown condition.

Moreover, the GO enrichment analysis suggested an imbalance in serine metabolism that extends to the selenocysteine biosynthetic pathway. CG1427 protein is behind this enrichment result, taking into account its up-regulation status from microarray and its mitochondria status from the Ttpred prediction. However, since this and other proteins (5%) are theoretical (one analysis step deals with genes having proteins targeted to mitochondria as well as outside of it, due to differences in their sequence from alternative splicing process), experimental evidence should support this results.

TABLE OF CONTENTS

Abbreviations

BP	Biological process
CC	Cellular component
DE	Differently expressed
DNA	Deoxyribonucleic acid
GO	Gene ontology
GRHCRF	Grammatical-Restrained Hidden Conditional Random Fields
GSEA	Gene Set Enrichment Analysis
HMM	Hidden Markov Model
MF	Molecular function
mRNA	Messenger RNA
mtDNA	Mitochondrial DNA
NN	Neural Network
qPCR	Quantitative polymerase chain reaction
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
SLIMP	Seryl tRNA synthetase-like insect mitochondrial protein
SRS	Seryl tRNA synthetases
TFAM	Mitochondrial transcription factor A
tRNA	Transfer RNA
KEGG	Kyoto encyclopedia of genes and genomes

Lists of Tables

Lists of Figures

1.INTRODUCTION:

1.1. Gene translation overview.

A gene is the molecular DNA unit of biological information. DNA and RNA are lineal polymers that encode for genetic information. Proteins are the product of mRNA translation by the ribosomes in the cell. For any existing organism, the transcription of DNA to mRNA and the translation to proteins are the basic steps of the central dogma of molecular biology (Crick, 1970). However, there are many existing regulation steps, some of them being particular for kingdoms or species but also determining differences between individuals.

It is known that prokaryote organisms do not have a membrane-defined nuclei with the genetic information, as eukaryotes do. A second difference is that eukaryotes have a monocistronic gene expression system (one promoter, one gene) while prokaryotes are polycistronic (one promoter, many genes). Because of that, in prokaryotes it is rather used the operon concept: a unit of genomic DNA containing a cluster of genes under the control of a single promoter.

Eukaryotic transcription has a regulation step called alternative splicing. The transcription machinery processes primary mRNA already in cytosol to mature mRNA by eliminating fragments of the transcript (introns) and keeping the rest (exons). Therefore, it allows a genome to be able to code for more proteins than would be expected from its protein-coding genes. Another regulation mechanism is epigenetics.

The genetic code is the set of rules by which information encoded within genetic material (DNA or mRNA sequences) is translated into proteins by living cells. It is said to be degenerated because different codons (triplets of nucleotides in genetic material) may code for the same amino acid (synonymous codons). As a general rule, only 20 different amino acids are coded by cells from the 64 different codon combinations (4^3: four base types and three positions). Evolutionary differences are found throughout the kingdoms of life regarding the genetic code and there is also a codon usage frequency typical from species to species.

This chapter may also briefly explain tRNA molecules. They are short RNA molecules (typically 75 to 95 nucleotides long) that are part of the genes transcripts not coding for proteins. Instead, they fold to achieve a conserved three-dimensional structure consisting of a series of double-stranded stems and single stranded loops (Sprinzl et al., 1998). They go through massive post-transcriptional modifications in order to achieve their role in protein translation.

The translation machinery works in a two-step process. First, amino acids are covalently linked to their cognate tRNA via an aminoacylation reaction catalyzed by a diverse group of proteins, the aminoacyl-tRNA synthetases (aaRS). The aminoacyl-tRNAs are then delivered to the ribosome by elongation factors. Once in the ribosome, the tRNA anticodon is matched to the mRNA codon and the charged tRNA delivers the next residue of a nascent protein chain.

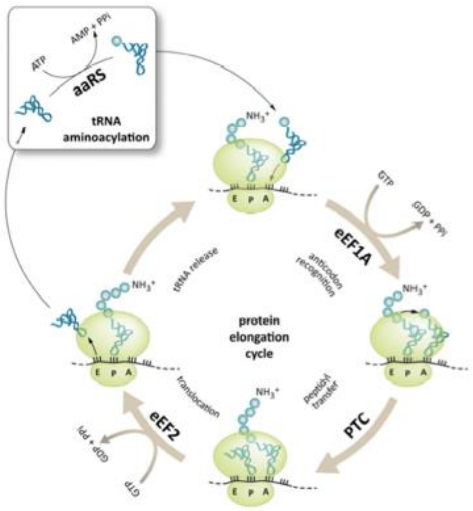


Figure 1. The two phases of protein synthesis. In the first phase, tRNAs are aminoacylated with their cognate amino acid by specific aaRSs. Aminoacylated tRNAs are delivered to the ribosomes to participate in ribosomal translation. The protein elongation cycle is depicted.

1.2. Mitochondria role in translation.

Humans have 3,234.83 Mega-basepairs of DNA per haploid genome, while on their mitochondria, the genome size is of 16,569 base pairs. This is a proportion higher than 99.99% which is maintained, at least, in the animal kingdom. However, mitochondria genes are essential and mutations there can be lethal or rather produce atypical, rare diseases.

Translation occurs in the four known stages as in cytosolic translation: initiation, elongation, termination and recycling. But mitochondria translation differs in many things and is also influenced by many conditions: environmental stress such as oxidative stress, cell division, cell renewal and differentiation, exercise, caloric restriction and low temperature. Moreover, correct mitochondrial biogenesis requires a coordinated regulation of mtDNA replication, mitochondrial fusion and fission processes as well as the synthesis and import of around 1100-1500 proteins encoded by the nuclear genome and synthesized in the cytosol (Pagliarini et al., 2008).

Mitochondrial DNA is composed by two strands, named heavy (H) and light (L) on the basis of their densities in a cesium chloride gradient. Animal mtDNA is a compact circular molecule (ranging in size from 16-20 Kb) typically encoding 13 protein-coding genes involved in oxidative phosphorylation, two rRNA subunits and 22 tRNAs. The non-coding region in the mtDNA, called control region, contains the promoters for mitochondrial transcription and the origin of mtDNA replication (Garesse, 1988). Despite differences in the order of the genes, *Drosophila melanogaster* mitochondrial genome closely resembles that of vertebrates in its overall structure (Figure 2).

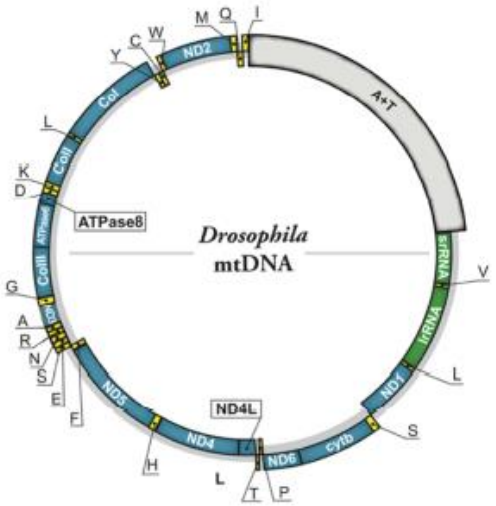


Figure 2. Map of the Drosophila melanogaster mitochondrial DNA. This circular molecule, of 19517 base pairs in length, encodes 13 proteins (in boxes), 22 tRNAs (in a single setter code) and 2 rRNAs (in boxes). The 5 kb control region with a high AT content is indicated in the thicker box and it includes the origins of replication for both strands. Sequence reference NC_001709. Adapted from (Echevarría et al., 2009)

Mitochondria are intracellular organelles organized in complex networks which are dynamic, in part due to fusion and fission processes. As an organelle it contains four distinct compartments:

- *Outer membrane (OM):* phospholipid bilayer. It has translocon protein complexes.
- I*nner membrane (IM):* phospholipid bilayer.
  Inner boundary membrane (IBM): forms contact sites with the OM and interacts with it during the import of proteins.
  Cristae membrane (CM): preferential site of proteins implicated in oxidative phosphorylation (Gilkerson et al., 2003), the iron-sulfur (Fe-S) clusters biogenesis, translation and transport of mitochondrial-encoded proteins.
- *Intermembrane space (IMS):* contains proteins involved in cell physiology, cell death and energy production (Vogel et al., 2006). It has electrochemical potential generated by the OXPHOS process in the CM.
- *Matrix:* contains proteins involved in the expression of mtDNA, tRNA and rRNA, several copies of the mitochondrial genome and a large number of enzymes related to mitochondrial metabolism. This includes tricarboxylic acid cycle, which generates electron carriers (NADH and FADH2), fatty acid β-oxidation pathway to convert long chain fatty acids to Acyl-CoA and metal metabolism (Lill and Mühlenhoff, 2008), like Ca^{2+} homeostasis or heme and Fe-S clusters.

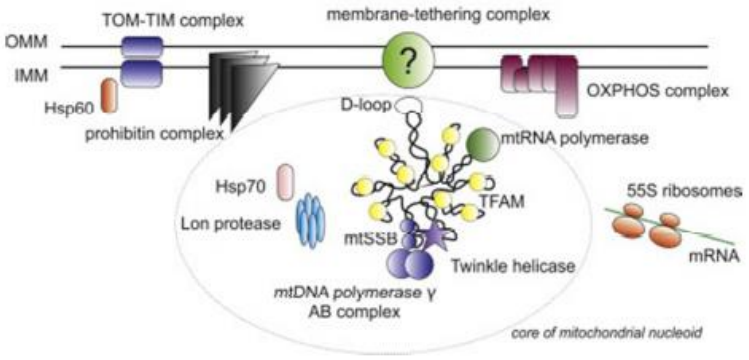


Figure 3. Schematic view of a single human mitochondrial nucleoid. Only one molecule of mtDNA is shown. MtDNA is usually packaged with TFAM. The D-loop is a regulation site for the replication and transcription of mtDNA and is thought to be bound to the inner mitochondrial membrane, probably through a multiprotein complex. Human mitochondrial nucleoids are believed to have a layered structure: mtDNA replication and transcription take place in the core part (circled) while the subsequent RNA processing and translation occurs in the outer part. TFAM, mtRNA polymerase, mtSSB, mitochondrial DNA polymerase γ complex, Twinkle helicase and LON protease are thought to be components of the core of human mitochondrial nucleoids. Adapted from (Bogenhagen et al., 2008).

1.2.1. Drosophila melanogaster as a model.

Drosophila melanogaster has played an important role on modern biology. The main features that make this organism a useful model are the following (Debattisti and Scorrano, 2013):

1) Short life cycle and high level of fertility that allow rapid expansion of populations. The generation time of *Drosophila melanogaster* takes about ten days at 25°C.
2) The fly may be considered multiple model organisms, each with its own specific advantages, defined by developmental stage: the embryo, the larva, the pupa and the adult.
3) The easiness to feed and maintain. Creation of centers of generation and maintenance of stocks
4) Physical mapping of genes on polytene chromosomes.
5) The existence of mutagenic agents to generate large collections of stocks mutants.
6) Complete sequencing and annotation of *Drosophila melanogaster* genome (Adams et al., 2000). It encodes for a little more than 14,000 genes on four chromosomes, only three of which carry the bulk of the genome.
7) Localization of proteins and RNA in tissues and cells: transcriptomics, proteomics, etc.
8) High conservation of genes that control basic developmental processes between *Drosophila melanogaster* and humans. It has been estimated that nearly 75% of disease-related genes in humans have functional orthologues in the fly (Lloyd and Taylor, 2010; Reiter et al., 2001).
9) Numerous genetic tools are available, such as balancers, transgenesis or RNAi technologies.

Drosophila melanogaster is considered close to *Homo sapiens* regarding mitochondrial biogenesis and function. Both species encode the same polypeptides, tRNAs and rRNAs needed for mitochondrial protein synthesis in mitoribosomes. Mechanisms of maintenance and expression of mtDNA, coupling of enzymatic complexes of the mitochondrial respiratory chain and mitochondrial transport are conserved from insects to humans and only minor differences are found. Therefore, to study mitochondria dysfunctions in *Drosophila melanogaster* is expected to elucidate mitochondria disease features in humans [1] (Guitart et al, 2010)

1.2.2. State-of-the-art of the seryl tRNA synthetase-like insect mitochondrial protein (SLIMP) characterization.

Seryl-tRNA synthetases (SRS) are dimeric enzymes responsible for the serylation of $tRNA^{ser}$. On the process to characterize SRS on the *Drosophila melanogaster* proteome, three protein sequences matched the bioinformatics results as human SRS orthologues. One was associated as the cytosolic enzyme (SRS1), the second as its mitochondrial equivalent (SRS2), while the third protein appeared to be an aminoacyl-tRNA synthetase (aaRS) like protein, this is, similar in sequence and structure to SRS but without their biological function. (Guitart et al, 2010).

The aaRS, which is a very old family of proteins believed to appear in conjunction with the first genetic code development (Woese et al., 2000) have suffered many horizontal gene transfers during evolution. This paralogues are documented for both subclasses I and II of the protein family (Dohm et al., 2006), [2]. While some aaRS-like proteins develop signalling functions, others regulate their own expression or regulate glucose metabolism, as examples. A first approach on SLIMP characterization revealed its presence in all available insect genomes, as well as echinoderms and arachnids and also an essential role to fly viability. [1]

Currently, SLIMP is not yet fully characterized but many important features are already known, especially from a thesis devoted to this (D. Picchioni, 2014). It is located inside mitochondria, since it has an N-terminal sequence of 21aa, characterized with mass spectrometry. Molecular modeling of SLIMP structure supports the conservation of the dimeric structure (confirmed by Gel filtration chromatography data) and its N-Terminus coiled coil (common in most SerRS) but a possible loss of its catalytic part. No ortholog of SLIMP is predicted in the mammalian system, and the search for a functional homolog is challenging.

Experimental results showed that SLIMP retains the property to bind mitochondrial $tRNA^{Ser}$ isoacceptors but they do not become aminoacylated. It did not appear to interact with DNA in vitro. However, SLIMP binds to all RNA sequences that are encoded in the mitochondrial genome in vivo as was confirmed by qPCR results on RNAs bound to immunoprecipitated SLIMP. Besides, SLIMP is known to interact with SRS2 and LON protease, a fact that can lead to a modulation in both mitochondrial translation and replication, respectively.

It was shown that SLIMP and SRS2 form a heterodimer and their protein levels are interdependent. It was also proven by *in vitro* aminoacylation assays that SRS2 is only able to perform tRNA aminoacylation activity in the presence of SLIMP, and by *in vivo* pulse chase analysis, that mitochondrial translation is reduced upon SLIMP or SRS2 knockdown.

Moreover, LON is the main mitochondrial protease and it is involved in the degradation of oxidized or misfolded proteins to prevent protein aggregation. LON is responsible for selective degradation of the Mitochondrial Transcription Factor A (TFAM) and it stabilizes the TFAM:mtDNA ratio in *Drosophila melanogaster*. Upon SLIMP knockdown in Schneider 2 (S2) cells, TFAM protein levels are increased, but they are restored back to normal if LON protease is overexpressed. Therefore, SLIMP regulates TFAM protein levels in a LON dependent manner.

Upon SLIMP knockdown, mtDNA copy number is increased, as observed with microscopy, meaning that SLIMP is able to modulate mtDNA replication through the interaction with LON protease. Soon after, *Drosophila* cells accumulate in G2 phase and their proliferation is reduced. Furthermore, the transcription of certain genes involved in the E2F1 pathway is activated, which may lead to an acceleration of the G1/S phase. This has been proven by GSEA results of microarray analysis that will be discussed in the conclusions.

1.3. Microarray experiment of *Drosophila melanogaster* S2 cells with SLIMP knockdown.

The microarray technology is a tool that can measure the expression level of thousands of genes in a particular mRNA sample. This high-throughput expression profiling can be used to compare the level of gene transcription under a control and a condition to study, with the expectations to identify biomarkers, classify diseases, monitor responses to therapy or understand the mechanisms involved in the genesis of disease process (Adi L. Tarca et al, 2006). In our case, the microarray of knockdown of SLIMP to S2 cells in *Drosophila melanogaster* is meant to elucidate mechanisms involved in the condition imposed and quantify transcriptome differences.

The knockdown was possible by using the GAL4 system genetic tool. The methodology and its optimization was described in (Guitart et al, 2010) and used more recently in the Gene Translation Lab at IRB. In the cell culture facility, four cell samples were taken: G1 wild-type, G2 wild-type, G1 knock-down and G2 knock-down. Later, a total of three replicas for each cell sample were analyzed by the Functional Genomics facility. Finally, the Biostatistics/ Bioinformatics facility analyzed the microarray raw data to interpret the intensities by means of fold change and adjusted pvalues. Moreover, a Gene Set enrichment Analysis (GSEA) [3] was conducted so that many information was already available about the experiment.

<table>
<tr><td>G1 WT (3 replicas)</td><td>G2 WT (3 replicas)</td><td>G2WT_G1WT (3 vs. 3)</td><td rowspan="2">G2_G1 (6 vs. 6)</td></tr>
<tr><td>G1 KD (3 replicas)</td><td>G2 KD (3 replicas)</td><td>G2KW_G1KD (3 vs. 3)</td></tr>
<tr><td colspan="2">KD_WT (6 vs. 6)</td><td>-</td><td></td></tr>
<tr><td colspan="3">Interaction group: (KDG1 – KDG2) – (WTG1-WTG2) or (3 vs. 3) vs. (3 vs. 3)</td><td></td></tr>
</table>

Table 1. Samples hybridized in the microarray (grey background) and each comparison analysis. All biological changes in transcriptome due to change of cell cycle phase or the knockdown induction can be addressed. The interaction group appears to be the most powerful comparison for knock-down (KD) condition since it is previously normalized by change of cell phase cycle.

1.4. State-of-the-art of sub-cellular protein position prediction programs: TPpred2.0 for mitochondria.

Sub-cellular protein position prediction programs are bioinformatic tools developed in order to differentiate proteins targeted to a specific organelle. Signaling sequences in proteins can be found through any part of the sequence, even though they are mainly found in the N-terminal part. This sequences are the key to control the import of nuclear encoded proteins into mitochondria, chloroplasts in plants or the nucleus itself, mainly. The signal peptide is typically removed at the destination by a signal peptidase if not cleaved when the native protein is translocated through the outer membrane of the organelle by means of the translocon protein complexes.

Targeting peptides are highly heterogeneous in terms of length (ranging from 10 to 150 residues) and primary sequence (Bruce, 2001; Patron and Waller, 2007). Since the prediction is not trivial, many programs approaching the issue exist:

- *MitoProt (Carlos and Vincens, 1996):* Defines a discriminant function on a pool of 47 physicochemical properties.
- *PredSL (Petsalaki et al., 2006):* Combines Neural Networks (NN), hidden Markov models (HMM) and scoring matrices.
- *TargetP (Emanuelsson et al., 2007):* Based on NN and including ChloroP (1996), iPSORT (2002), Predotar (2004) and PredSL programs.
- *TPpred [4](V. Indio et al., 2013):* Based on Grammatical-Restrained Hidden Conditional Random Fields (GRHCRF) which is a well suited machine learning tool to solve labelling problems (Fariselli et al., 2009, Savogadro et al.,2011).

In addition, WoLF PSORT (P.Horton et al.,2007) is a program that approaches the prediction for all organelles in the cell by k-nearest neighbour classifier. But cross-validation studies within the same publication estimate sensitivity and specificity of around 70%.

For mitochondria and plastids targeting prediction TPpred2.0 is the most robust method nowadays, with a final performance of 96% of accuracy reported (both targeting status and cleavage site) and 0.58 Matthews correlation index. There is a 3.0 version available as web application but for whole proteome or microarray derived data analysis the version 2.0 is meant to be used locally in Linux OS (only 3.0% false-positive rate). Non-redundant data was used to train the program and all experimental evidences of N-terminal sequences for *Homo sapiens* (GRHh37.p5), *Arabidopsis thaliana* (TAIR10) and *Saccharomyces cerevisiae* (EF4) from EnsEMBL were used, thus covering animal, fungi and plant kingdoms.

1.4.1. Goals of the current project.

To code a new pipeline in R in order to use TPpred2.0 predictions from microarray analysed data is proposed. This pipeline is meant to be useful in the current project but also for other projects related to transcript differences in mitochondria.

Regarding SLIMP characterization the microarray was done in order to address transcript quantification and their functional analysis in a genome-wide level. But the gap in knowledge that the current project wants to cover is the following: *differentiate if proteins are targeted to mitochondria or not and the functional analysis related to each of the new putative subsets*.

To sum up, a generic pipeline will be developed as well as a real example of application will be used during the pipeline development and testing and for the pipeline results output and discussion. To bring up evidences in SLIMP characterization is also expected.

2. PIPELINE OF THE ANALYSIS AND ITS METHODS:

All the pipeline files of the current project can be accessed in the following GitHub link: https://gist.github.com/marianoguera/6c58fe75b893be8cb17d7ff074745660

A single excel (.csv) with the microarray results is meant to go through the pipeline. Specify its name into the very beginning of the first file, in the *read.csv* R function. For more than one excel (ex. KD_WT and interaction as described in Section 1.3), run the pipeline each time. Same new folders and file outputs will be saved automatically so make sure to move your results data once completed one pipeline run.

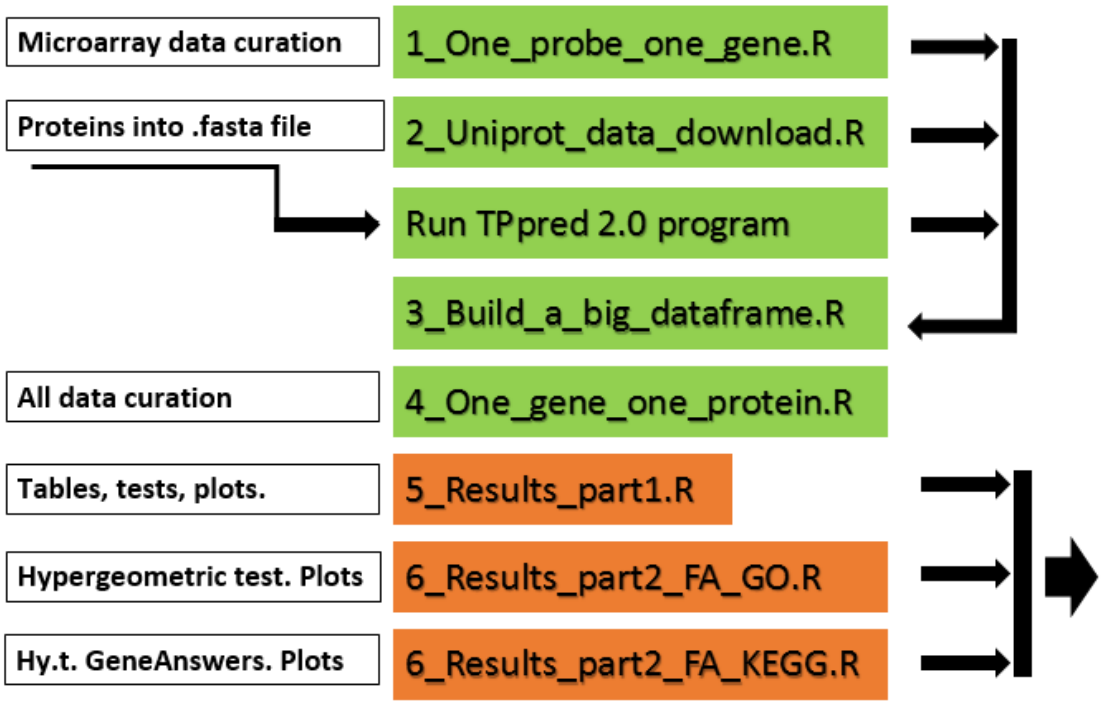


Figure 4. Eight steps of the pipeline: seven R files and the program prediction on third place. The analysis pipeline is depicted. Green steps are from the data preparation. Then, the .csv file output after all data curation is used as input in all the orange steps to obtain biological related information in a ready-to-use knowledge purpose format.

The following chapters in 2.1 and 2.2 sections (green and orange in Figure 4, respectively) address the detailed explanation of each R file content. Then, 3.1 and 3.2 sections (again green and orange in Figure 4, respectively) show the results from each part of the code.

Some R packages need to be installed to run the pipeline. From CRAN: ggplot2 [5], reshape, gridExtra, seqRFLP and readr. From Bioconductor: UniProt.ws [6], AnnotationForge [7], drosophila2.db [8], GSEABase [9], GOstats [10] and GeneAnswers [11].

2.1. Pre-processing steps to get data to work with:

The current section is essential in the establishment of a generic pipeline. It addresses microarray data curation after its analysis from raw data, in terms of improving annotation by direct assignment of gene name and the theoretical protein name to the experimental transcripts.

For a new user it is also relevant to get a first hands-on the data content, and the columns we have for each probe are (from left to right):

Probe.Set.ID	Gene.Symbol	Entrez.Gene	Gene.Title
Chromosomal.Location	global.mean	global.mean	logfc
log.ci.0.025	log.ci.0.975	fc	ci.0.975
ci.0.975	condition1.mean	condition2(basal).mean	t
pval	adj.pval	rej	probDE

Table 2. Column names of the microarray input. Example from the SLIMP knock down microarray. 20 columns are present. Read them from left to right and from up to down. Means are already normalized with logarithms.

From these column names, few are mandatory: Probe.Set.ID, Entrez.Gene, the mean of intensity for each condition, the adjusted pvalue and the fold change. To determine which probes are differentially expressed (DE), the *t* column has the t-test (frequentist statistics) while *probDE* has the Bayesian approach. Overall, a *rej* "rejection" column, summarizes DE probes by -1 (down), 0 and 1 (up) becoming also mandatory to use in the code that follows. In our data, DE was stablished from an absolute fold change value of 1.5 and a false discovery rate (fdr) correction of 0.05. Therefore, for the excel files reported in Section 1.3. , slight differences in adjusted pvalues thresholds are present. See a Volcano Plot representation of this data in Section 3.1.1.1.

2.1.1. Analyzed microarray data filtering: one probe (transcript), one gene.

The goal of the first script is to keep a unique gene annotation for each probe of the microarray. First of all, the probes without gene identification are removed. This can be due to probes designed to check microarray results robustness (common to same microarray type experiments, in our case: Affymetrix, Drosophila_2.na36.annot_20161118) through insertion of probes not present in the current genome (transcriptome) of the organism. This is, to avoid false positives during the experiment or the analysis, some probes are meant to return any identification.

A second issue to address is the alternative splicing process when annotating. Biologically speaking, one gene can produce different mature mRNA. Since we are measuring transcripts, which can be mature or not, when annotating them more than one gene name can be associated with the transcript sequence (intrinsic microarray technical fact). Therefore, the standard process to proceed in further code is the selection of the first identification reported, regardless of how many identifications are present or any other parameter. Just by chance. Besides, a *'firstelement'* function is defined in the pipeline to keep first elements of a list in a column. It is then applied to the Entrez.Gene, Gene.Symbol, Gene.Title and Chromosomal.Location columns data.

Third, there is the case when one gene gets represented by more than one probe. In this situation, we could keep only the first probe but it is better to keep the probe that includes a highest variability (standard deviation) on the mean of their experimental intensities (3 replicas). On doing this, power is lost but we assure the probe is representative for that gene in a uniquely identification. To code this, the function aggregate was used, which allows to consider a condition (max. sd in this case) under the aggregation of Entrez.Genes identifications. Then, to select this data in the initial file, unique identifications should be used and only a *'paste(Entrez.Gene,sd)*' new temporary column was available as coding options to do this. Normally, Probe.Set.ID or Entrez.Gene is used but the first was not available after aggregate function and the second was not unique.

Finally, three pie charts are coded to visualize the filtering total dimensions by number of probes, the proportions under the total data and the number of probes per transcript in the second step before filtering. By using the *table* function and few more code we know if any probe representing the same gene was at the same time up and down regulated. We also infer the thresholds used for DE and save this externally. Yet a Volcano plot comparison for before and after the filtering conditions is coded. All this output is reported in Section 3.1.1.

2.1.2. Uniprot associated data download

Next step in the pipeline is to download the protein sequences associated to the uniquely identified genes for each transcript. The only option to do this is by using Entrez.Genes in UniprotKB. After checking this functionality via website, the package UniProt.ws in R was implemented in the pipeline.

The first step is to ask for *availableUniprotSpecies* and search for *Drosophila melanogaster* name, in our case. Once you know the taxId for the organism, the function UniProt.ws is used to build a Uniprot object variable. *Keytypes* and *columns* functions of this object help customize the data download from the website, which is done by calling the *select* function. As for the current pipeline, column terms in table 3 were queried by standard "ENTREZ_GENE" keytype.

ENTREZ_GENE	UNIPROTKB	PROTEIN-NAMES	FLYBASE
GO-ID	GO	KO	KEGG
SUBCELLULAR-LOCATIONS	SEQUENCE	REVIEWED	

Table 3. Column names of the Uniprot database query. Read from left to right and from up to down.

When using the pipeline more than once for the same microarray type, the query can be saved the first time as an R.Data object. Later on, only the loading to this R.Data is needed and computation time is saved. The code proceeds with a removing of duplicated UNIPROTKB names and those cases (rows) when the protein name or sequence is not annotated.

The curated data with all column data is saved externally as a .csv file in the current working directory. Besides, the .fasta file with each protein name and sequence is also saved externally with the *dataframe2fas* function of seqRFLP package in R.

2.1.3. TPpred2.0 program prediction:

Follow the instructions in: https://tppred2.biocomp.unibo.it/tppred2/default/software to download and install the program locally in your computer.

Once the TPpred2.0 is installed, create a folder called query and another called results in the main folder of the program. Copy the .fasta file created in last chapter into query folder. With the terminal on the main folder of the program, run the following terminal command:

/bin/runTPpred2.sh query/inputTPPRED.fasta results/resultTPPRED.csv

The whole microarray protein sequences prediction can span from one to two days. The output is in the form of Protein ID, Cleavage Site and Score columns, from left to right. Cleavage site reports the length in aminoacids of the N-terminal signaling sequence or a hyphen in case there is no prediction. All scores output are normalized from 0 to 1 with 1 independent of the position predicted.

2.1.4. Get a unique big file for current microarray, Uniprot and predicted data

With the three independent datasets seen above, we have the opportunity to unify them. But the program output needs first a pre-processing step. This is accomplished by reading the output file as lines (*read_lines* function from readr package), then split for any space distance (split = "\\s+") and later select all first, second, and third elements of the list (sapply(t3, function(x) x[1]), sapply(t3, function(x) x[2]), etc), keep each of them as a variable and build them as a data.frame with Protein ID, cleavage site and score as columns instead of a single character row or a list of lists. All scores should indicate 1 as 100% predicted to mitochondria and 0 as 100% out of it. Therefore, we do *1- score* to the predictions outside mitochondria.

The first unification is made with the Uniprot and the predicted data, since they have the same row dimensions and a common unique UniprotKB name. Only correct order has to be assured and this is made as the standard way to code: by putting as row names the unique identifier and then proceed as a selection of one data from the other. This rearranges every identifier and its row from the second data to the order of the first data. Together it is saved as alldata23 variable in the pipeline.

To continue, the initial microarray data has to be correctly associated to the alldata23, which have different row dimensions. However, since they share Entrez.Gene column (unique for microarray but with repetitions in alldata23), the *merge* function really pays off here. It just takes specifically by Entrez.Gene the information in microarray and writes it as many times needed for the repetitions in alldata23, becoming a data frame of the biggest size (named alldata123 in the pipeline).

A new section in this script is devoted to convert columns of interest (rej and mito_status) in factor variables (*factor* function). This way we will be able to summarize counts in tables much easier in future steps.

Moreover, a last subsection still has to consider the genes that did not encode for any protein but are present in the cell and in the microarray. These are called “NP”, from no protein or no prediction and were annotated automatically with NA “Not Available” when doing the merge step. A matrix of NP is built with the same dimensions as the NA data and then substituted. A final table function checks the ‘rej’ and ‘mito_status’ data to check their factorization is correct.

2.1.5. Big file filtering: one gene (transcript), one protein:

Reached this point, a single representative protein per gene has to be selected and the rest, if present, discarded. This way we are still working at the transcript level, inferring transcript product targeting location from the prediction of the protein associated. The transcript:proteins ratio is not unique, neither trivial to document genome-wide [12].

The code in this script starts by quantifying how many proteins per gene we have from the download query in Section 2.1.2. A detailed bar plot using *ggplot2* package and the summary version with a pie chart is reported in figure 9.

Next step is to count how many mitochondria status (if targeted or not) for proteins by each gene are present. A matrix from a table is obtained. Since we would like to continue subseting this matrix but only data.frames can be subset in R (and just changing to data.frame rearranges columns and rows differently) the solution applied is to save externally this table. Then, it is loaded again as a data.frame (E_file in the pipeline). Therefore, we can proceed doing subsets. One pie chart represents this (See Section 3.1.2) where we have the total of genes with only mitochondria targeted proteins, genes with only no mitochondria predictions, with “NP” or also genes with some proteins targeted into mitochondria and some outside of it.

To select a single protein for every gene in the first three of the fourth cases described above is trivial because the prediction is clear: ‘NP’ are already one gene one (no) protein, and for all mitochondria or all outside of it, the first of the many proteins for each gene are selected, independently of the score reported. When applying to the genes with both positions it was decided to proceed by parts in criteria of a percentage score for the proportions of those. See the table below:

% higher than 0.5	% equal to 0.5	% lower than 0.5
Ex: 3mitochondria, 1 outside	Ex:1mitochondria, 1outside	Ex: 2 outside, 1 mitochondria
Mitochondria selected	Mitochondria selected	Mitochondria selected *

Table 4. Decision criteria for genes with at least one protein targeted to mitochondria and another outside of it. Inner (% equal to 0.5) mitochondria status selection criteria is supported by the fact that overall genes targeted to mitochondria are less than 1% from the whole microarray if only considering genes with one protein at the beginning. See Figure 8A. Besides, the third column was decided to be labelled as mitochondria too instead of NOT because after a first analysis results it was seen that proteins with an important role regarding SLIMP can also be found there. See results and discussion section. Therefore, we change to Figure 8B to Figure 8C proportions.

Finally, all the data after the proper filtering was unified, correct dimensions checked, and saved externally. Few pie charts were coded to show the proportions of this many steps in our example case of *Drosophila melanogaster*, available in Figure 8.

2.2. Results from the current optimal data:

2.2.1. Table of counts by predicted position and differential expression (DE)

The 5_Results_part1.R file has many goals regarding transcriptome quantification and targeting proportions. To start with, the code *addmargins(table(csv$mito_status,csv$rej))* returns the total number of transcript counts by position predicted and DE, respectively. Since it has *addmargins* function it also returns the total sums for columns and rows.

2.2.1.1. Plot TPpred score vs FoldChange (FC).

In order to see if the fold change parameter from the microarray and the predictions score value are related, we code for a scattering plot. The ggplot2 package is used and the capability to define a threshold for two different colors is set by selecting DE and not DE data. Besides, horizontal lines are applied to the 1.5 and -1.5 fold change value in Y-axis (log2fc = 0.0585), while a vertical line is drawn on the 0.5 score in X-axis. Since Y axis only considers fold change and DE is set by fold change and adjusted p value, the DE points appear mainly as expected in the plot but few are not (those without significative pvalue). This is in complete accordance with the table of counts. The *png* function and *dev.off()* are written at the beginning and at the end of the plot code as any other automatically saved image.

2.2.1.2. Proportions test.

The table of counts is saved into a variable. Then, the *prop.test* function from the stats R package is used. It needs an input of numerator values and an input of denominator values from the proportions to test, with more than two relations evaluated if needed. The null hypothesis (H0) is that proportions are not demonstrated to be statistically different while the alternative hypothesis (H1) can be “two.sided”, “less” or “greater”, as determined by the code. The pipeline functions just use two.sided parameter.

Three questions are addressed, therefore, three tests are coded:

1) Are transcripts coding for mitochondria targeted sequences as abundant as the rest?
2) Are up or down mitochondria targeted transcripts expressed differently than the total up or down transcript expression, respectively?
3) Are up-mito transcripts expressed differently than down-mito transcripts regarding total mitochondria expression?

To report this results externally the *capture.output* function is used for the table and for each test variable, with the name of the .txt file created and an *append = TRUE* parameter is set so that this four results are saved consecutively into the file. See results in Section 3.2.1.2.

2.2.1.3. Save subsets of data externally (.csv).

Because we have each subset well defined we save them externally in order to access them later in a much more easy-friendly way. They are: down_mito, down_no_mito, NP.down, NP.up, up_mito and up_no_mito excel files (.csv). Moreover, it is coded to keep only essential columns: Entrez.Gene, Gene.Symbol, UniprotKB, protein.names and GO. The UniprotKB identifier can be used in the Flybase [13] or Entrez.Gene sets can be used in the DAVID (Dennis *et al.*, 2003; Huang *et al.*, 2009) website database for functional annotation. Besides, this files are saved into a specific folder like they are the already described files coded in 2.2.1.

2.2.2. Functional annotation analysis of the subsets:

The code used in the pipeline to answer GO and KEGG enrichment analysis is not the only possible [14, 15]. However, the most up-to-date code approach known has been used in terms of code simplicity and visual-ready results output.

2.2.2.1. Gene ontology (GO) enrichment analysis.

The 6_Results_part2_FA_GO.R file starts with the GSEA object definition, which is a GeneSetCollection of type GOCollection object from drosophila2.db package. This object can be saved externally as a R.data object and if the pipeline is run more than once then the *load(file="gsc.GO.RData")* is only needed.

It then defines the total of genes as: *entrezuniverse = Lkeys(org.Dm.egGO)*, which is the compete genome information. Just after, it codes for a total of twelve gene ontology enrichment results: biological process (BP), molecular function (MF) and cellular component (CC) for the main four subsets. Each of them has its hypergeometric test (probability of k successes in n draws, without replacement, from a finite population of size N that contains exactly K successes). The GOStats R package with the *GSEAGOHyperGParams* and *hyperGTest* function was used. Upon *htmlReport* function all the output into the 0.05 pval cutoff is saved. Besides, for report visualization, a function called my_barplot was brand-new coded so that only ten best terms and their pvalues are shown. All this data is saved automatically in results.part2.FA folder. See results presented in Section 3.2.2.1.

2.2.2.2. Pathway (KEGG) enrichment analysis.

The 6_Results_part2_FA_KEGG.R file could also use the GSEA object system to later code for KEGG enrichment similarly as seen before. However, a new approach with GeneAnswers package is used, which, in turn, could also be used for the GO enrichment part. An advantage of this is that more visually informative results are obtained. But, since the package includes

Cytoscape plug-ins the image opens in an external window and it is not possible to save them automatically while running the file.

The GeneAnswers package works with S4 R objects instead of S3. Because of this, the flyGeneInput variable is saved with Entrez.Gene, pval and fold change values, while flyExpr variable is saved with Entrez.Gene and the two intensity columns from the two conditions in the microarray (up to six if considering replicas instead of means). With this two dataframes the GeneAnswers objects can be built through *geneAnswersBuilder* and *geneAnswersReadable* functions. This last function result allows to check the genes involved in the KEGG enriched pathway (*genesInCategory* slot) and the values associated after the test (enrichmentInfo slot).

With the GeneAnswers object, many functions are available to use in the package. Two are essential in the pipeline, *geneAnswersConceptNet* and *geneAnswersHeatmap* which return an interactive net of KEGG terms (five best) with their genes involved and a table of terms and genes, respectively. The first was coded to be used for the four main categories (up-mito, down-mito, up-no-mito and down-no-mito) while the second was rather used for less specific subsets (allmito, allDE, etc). Only biologically relevant plot result will be reported.

2.3. Integration of the R files to a bash script.

A bash script is published at the same time with the R.files. Therefore, the files can be used individually in the order explained or as a single call from the bash script in the terminal of linux with all R packages previously installed. The command is: source run_all_pipeline.sh

The R files are called with the Rscript command and the program is called as described in Section 2.1.3. To run the pipeline with RStudio with all source option for each file is equally recommended, considering the TPpred2.0 output is already obtained from Linux.

3. RESULTS AND DISCUSSION:

3.1. Concerning pre-processing steps:

3.1.1 One probe (transcript), one gene filtering dimensions.

The following quantification is reported for the Drosophila melanogaster microarray (typical from the microarray type rather than the sample(s) tested):

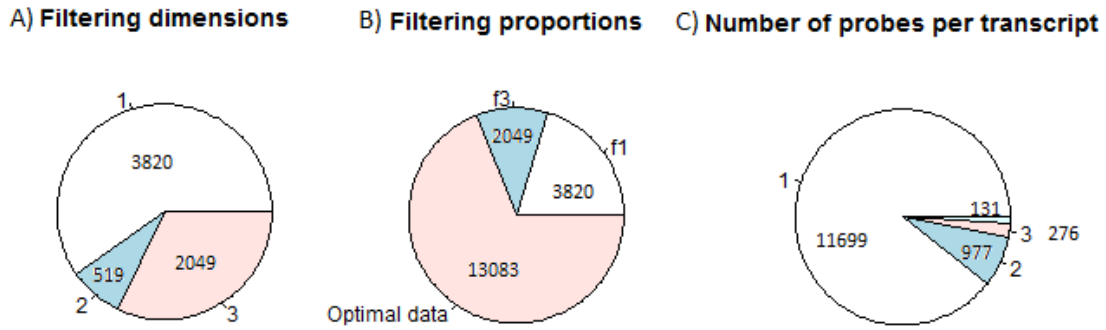


Figure 5. A) Filtering dimensions. 1 are the probes without gene identification, 2 the probes with more than one gene annotated and 3 are the genes represented with more than one probe (more exactly, the probes we remove from this). **B) Filtering proportions**. Optimal data is the unique gene – transcripts that are kept, f1 and f2 are the probes removed from the initial microarray data. **C) Number of probes per transcript**. It details the second filter phenomena. Smallest section stands for 4 or highest number of probes associated for the same transcript.

For interaction microarray no gene was represented with probes up and down regulated at the same time. However, the KD_WT did report *cindr* gene with this paradox.

3.1.1.1. Volcano Plot comparison for before and after the filters.

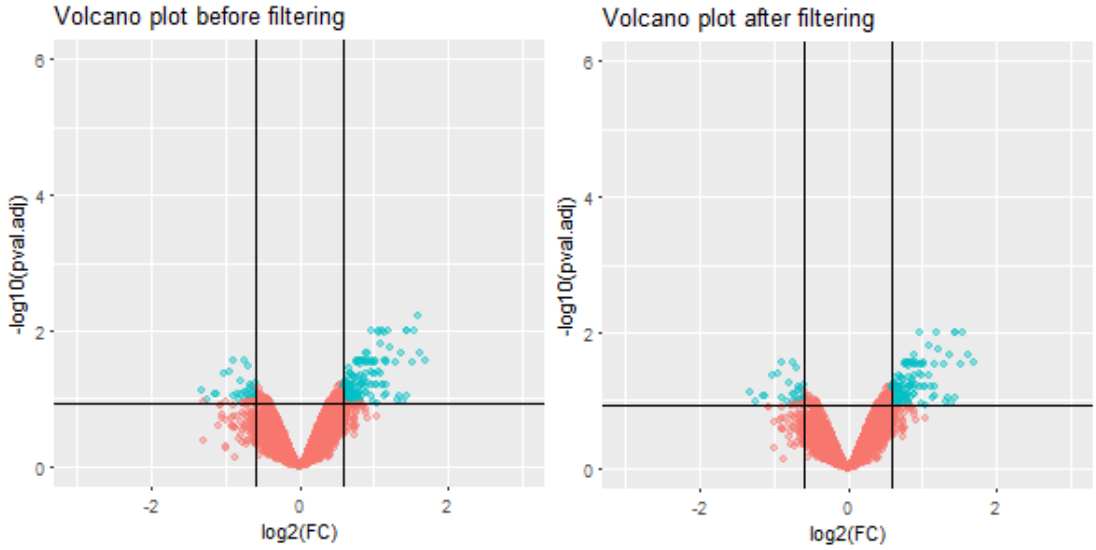


Figure 6. Interaction group data. Left) VP before filtering. Right) VP after filtering. The trend of differential expression enrichment in up-regulation is maintained. Remember interaction group points out differences in knockdown versus wild type after of G1 vs G2 normalization.

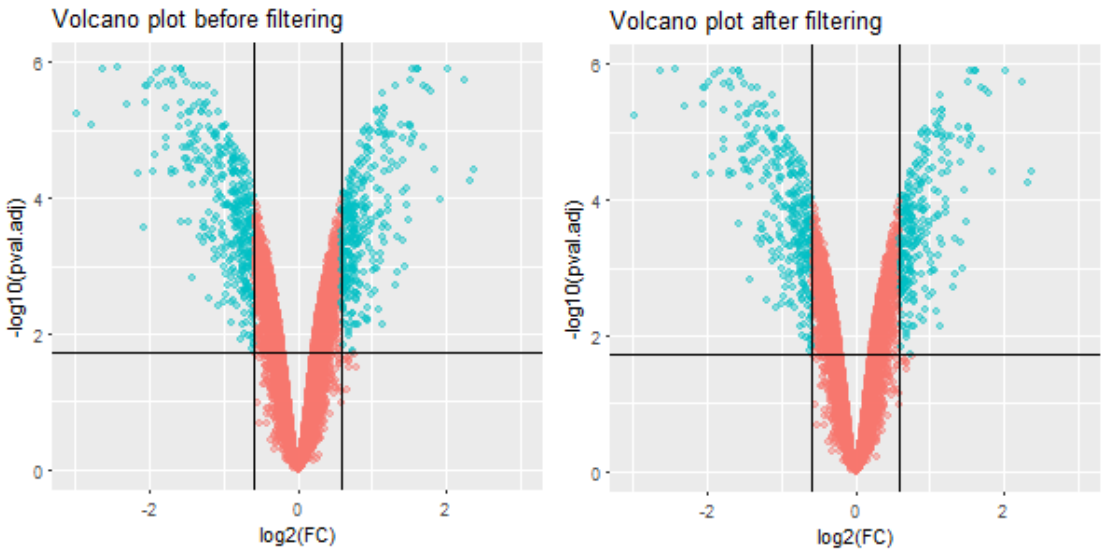


Figure 7. KD_WT data. Left) VP before filtering. Right) VP after filtering. The trend is maintained. However, we see down regulation is a bit stronger than up regulation. Remember KD_WT addresses the condition with no previous normalization for G2 vs G1.

3.1.2 One gene (transcript), one protein filtering dimensions.

Only conclusions from transcript proportions are expected in the next sections. However, we have to previously report the curating process used to associate the subcellular location regarding mitochondria for the transcripts, through their protein. Because of that, a ratio 1:1 is being used.

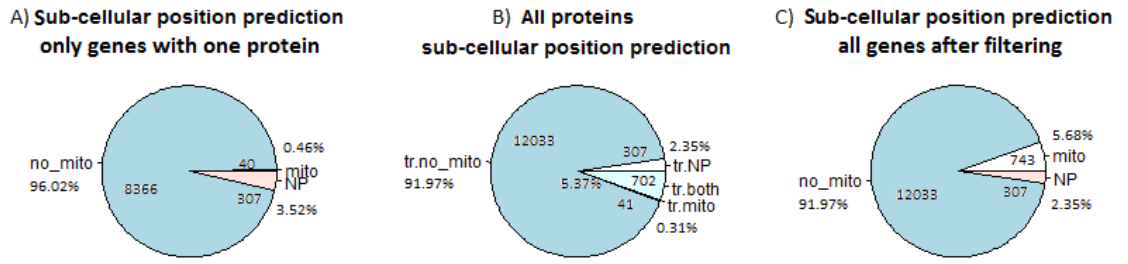


Figure 8. A) Sub-cellular position prediction only for genes with one protein. Most are targeted outside of mitochondria, while only 0.46% is targeted inside of it. **B) All proteins sub-cellular position prediction.** Since no criteria has been applied yet here, genes with proteins targeted both to mitochondria or out of it are present in the both variable (5.37%). **C) Sub-cellular position prediction for all genes after filtering.** This is the final distribution. The 'both' labelling in B has been unified to mitochondria targeted. 5.68% of data is targeted to mitochondria while 91.97% is targeted outside of it. See figure 9 and 10 for discussion of 'both' cases.

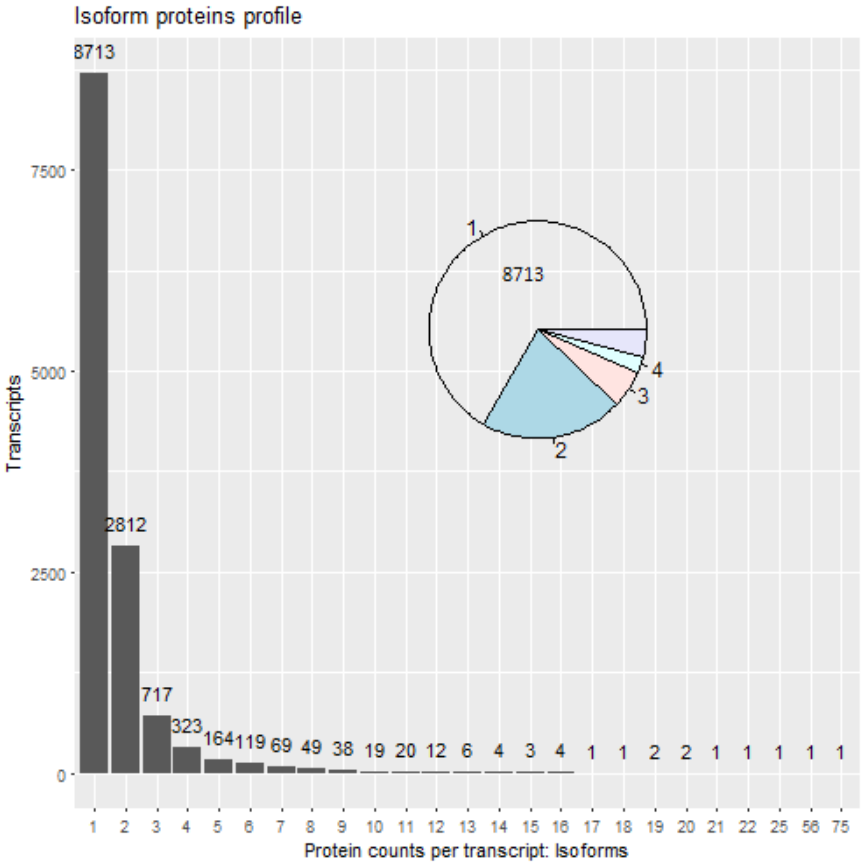


Figure 9. Isoform proteins profile. To have this data in our pipeline is intrinsic from the Uniprot data download of proteins by their Entrez.Gene. It is later incorporated in the prediction program, in the big data file and finally filtered accordingly. As a detail, the 75 value in the bar plot corresponds to a gene coding for membrane-channel protein(s).

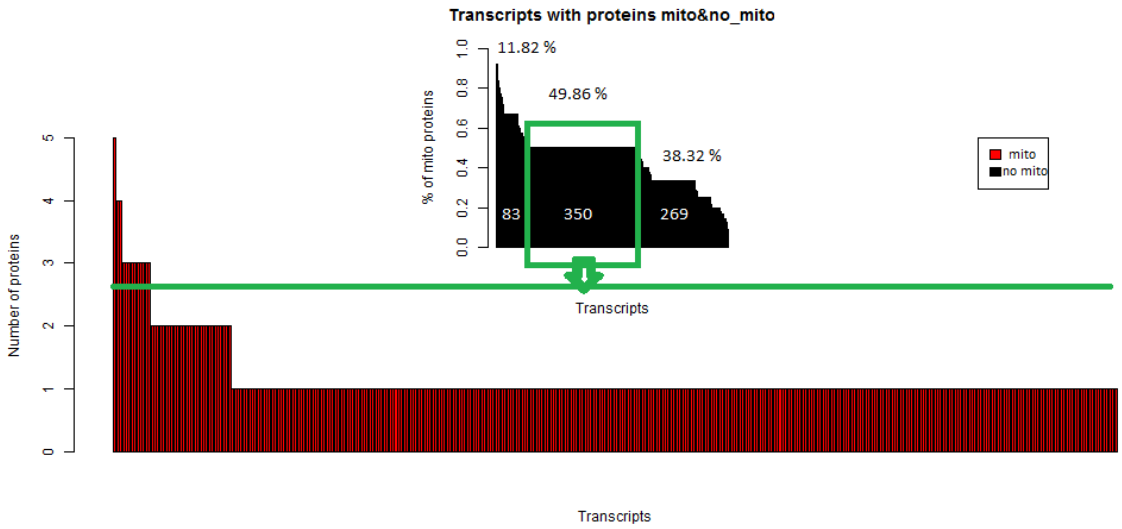


Figure 10. Profile of the genes with proteins targeted to mitochondria and outside of it. It is the same data as the 5.37% visualized in figure 8 B. In addition, we see that most of the genes under this condition have a proportion equal to 0.5. Moreover, this green selected data is mostly due to a single protein of each status for the same gene

3.2. Concerning biological questions:

3.2.1. Table of counts by predicted position and differential expression (DE)

INTERACTION	down	not_DE	up	Sum	KD_WT	down	not_DE	up	Sum
mito	2	731	10	743	mito	25	699	19	743
no_mito	24	11869	140	12033	no_mito	308	11472	253	12033
NP	0	305	2	307	NP	5	297	5	307
Sum	26	12905	152	13083	Sum	338	12468	277	13083
G2WT_G1WT	down	not_DE	up	Sum	G2KD_G1KD	down	not_DE	up	Sum
mito	22	699	22	743	mito	6	721	16	743
no_mito	413	11354	266	12033	no_mito	110	11687	236	12033
NP	4	303	0	307	NP	1	304	2	307
Sum	439	12356	288	13083	Sum	117	12712	254	13083

Table 5. Table of counts for interaction group, KD_WT, G2WT_G1WT and G2KD_G1KD. It is seen that the total number of probes are the same but they report different numbers in the small subsets.

The wild type vs. knock down comparison without cell cycle normalization (green) reports a total of 44 up-mito and down-mito transcripts. This is similar to the total 44 transcripts for G2WT_G1WT comparison (blue), which are the highest values. Since G2KD_G1KD reports only 26 total up and down transcripts, we conclude that the knock down condition decreases the number of transcripts targeted to mitochondria. This is confirmed by the interaction group (yellow) as we see that only a total of 12 transcripts are differentially expressed targeting mitochondria.

Comparing which are the transcripts mito-up reported in the interaction group comparison and the two previously calculated comparisons (blue and purple), we can infer some conclusions. A set of genes are present in this last two but not the interaction: pdm3, dnr1, ATP8B, alph, brp and Mmp1. Therefore, this genes are involved in the cell cycle change and their function is maintained. On the other hand, four genes (Asph, sm, trol and l(2)03659) are present in the mito-up G2KD_G1KD, not in the wild-type comparison and again in the interaction comparison. Therefore, they are likely involved in the up-regulated response targeting mitochondria for the SLIMP knockdown condition in S2 cells. But they are not the unique genes because the six remaining genes in interaction comparison are: CG4896, row, CG1427, plx, Lerp and CG42613.

Moving to the analogous mito-down comparison we also infer some conclusions. The set of genes maintained down-regulated in the cell cycle change are: CG12895, Dyb, CG1427, ea, CG34355 and crb. On the other hand, there appear to be no genes associated to both G2KD_G1KD and interaction comparison for the thresholds set. However, from the probe intensities procedure we do have the two genes, which are: CG42855 and CG43125. Their proteins appear to be uncharacterized in Uniprot (uploading content status in FlyBase database but serine-type endopeptidase activity GO term is annotated for the last gene.

Regarding the genes not coding for proteins there is also something to show. First, the four down-regulated genes in G2WT_G1WT (CR42254, CG7730, U4-U6-60K and CG7441) do not appear in the G2KD_G1KD (only CG3902), and none appears for interaction comparison. Then, no gene appears in G2WT_G1WT, the CG8677 and CR45567 genes appear for G2KD_G1KD and CG8677 and CG7730 are present in interaction group as up-regulated. For this last two genes, the CG7730 function is not known but for CG8677, Flybase search says "encodes a protein that dimerizes with Iswi to form a chromatin remodeling factor RSF (remodeling and spacing factor). RSF is involved in silent chromatin formation via His2Av replacement. [Date last reviewed: 2016-12-01]. "

Because of this, SLIMP knockdown is susceptible to induce less transcripts targeted to mitochondria, from the nuclei genome, (both up and down regulated) in comparison to wild type conditions through a silent chromatin formation induced by CG8677 gene.

Considering only the four main subsets for each comparison we see the following proportions:

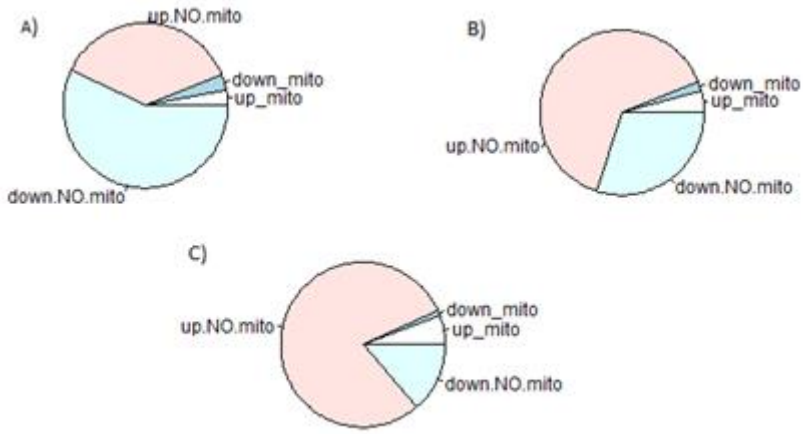


Figure 11. Four main subsets proportions by DE and position. A) G2WT_G1WT B) G2KD_G1KD C) Interaction. The proportion of up regulated transcripts outside mitochondria increases as well as the up regulated transcripts inside mitochondria for the interaction data. However, the total number of transcripts of these pie charts are not the same. Numbers reported in table 5 should not be forgotten.

Because of figure 11 we can say that SLIMP knockdown in S2 cells induces a higher response for up-regulated transcripts than for down-regulated ones, independently of the subcellular protein position predicted.

3.2.1.1. Plot TPpred score vs FoldChange (FC).

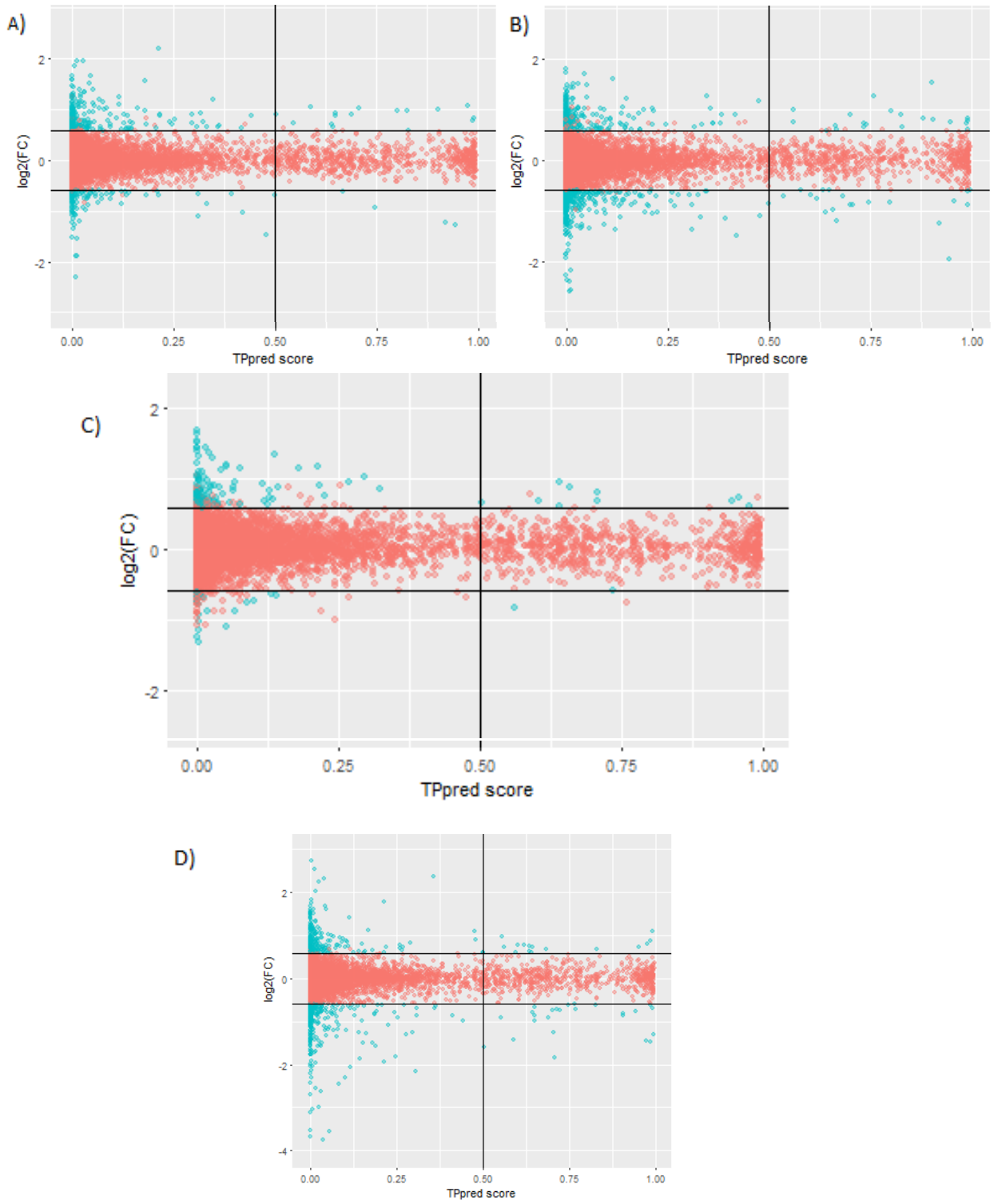


Figure 12. Scattering plot for Ttpred score vs. log(fc). A) G2WT_G1WT B) G2KD_G1KD C) Interaction D) KD_WT. Greenish blue color points in the upper part are up regulated transcripts while the ones from below correspond to down regulated transcripts. Red color is for not DE. Besides, the 0.5 score separates the mitochondria targeted transcripts (right) and the outside of mitochondria (left).

3.2.1.2. Proportions test.

It cannot be said that mitochondria up or down transcripts to their up and down totals are different in proportion to the total mitochondria transcripts regarding all transcripts for any of the G2WT_G1WT (p-value = 0.30), G2KD_G1KD (p-value = 0.88), Interaction (p-value = 0.81) or KD_WT (p-value = 0.29) comparisons.

It can neither be said that outside mitochondria up or down transcripts to their up and down totals are different in proportion to the total outside mitochondria transcripts regarding all transcripts for any of the G2WT_G1WT (p-value = 0.27), G2KD_G1KD (p-value = 0.62), Interaction (p-value = 0.99) or KD_WT (p-value = 0.79) comparisons. Since the first test did not appear significative for any comparison, this was also expected to do not be significative (very few NP).

Regarding the third proportions test, only interaction was significative (p-value = 0.04), meaning that up mito transcripts (10) are more expressed than down mito transcripts (2) from the total mitochondria expression (743 transcripts). Moreover, G2KD_G1KD is almost significative (p-value = 0.053) while G2WT_G1WT (p-value = 1) and KD_WT (p-value = 0.44) are clearly not.

3.2.2. Functional annotation analysis of the subsets:

3.2.2.1. Gene ontology (GO) enrichment analysis.

Results from interaction comparison are mainly reported because they address our final biological goal. However, see an example in the figure above of GO terms only considering the change of phase by both wild-type and knock-down condition:

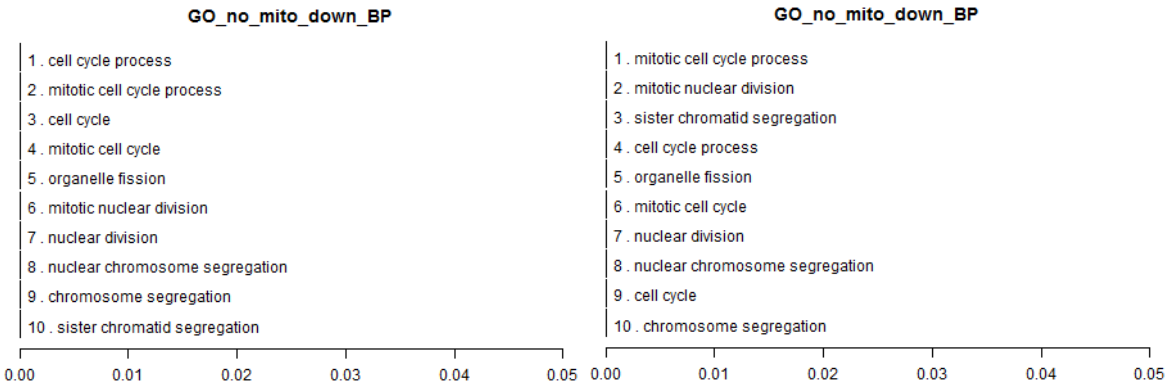


Figure 13. Horizontal bar plots of the ten most significative GO terms for Biological Process in no-mito down genes, for G2WT_G1WT (left) and G2KD_G1KD (right). G2 phase cells have DNA division down-regulated in comparison to G1/S phase cells. This fact is seen in the GO terms, which are more significative to this fact than to the possible change induced by SLIMP knockdown. Therefore, normalization to the change of cell phase cycle has been important.

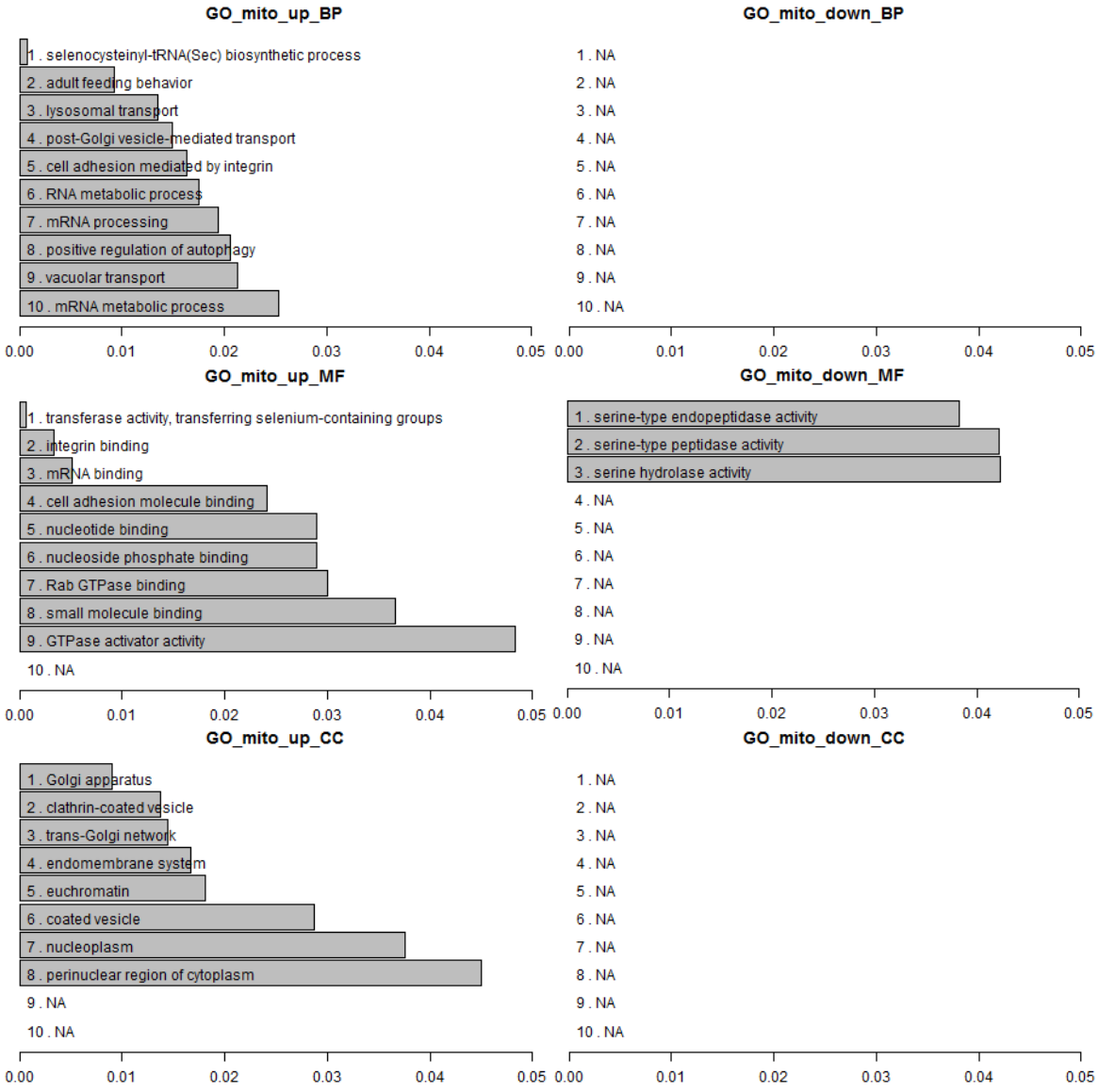


Figure 14. Horizontal bar plots of the ten most significative GO terms (BP, MF, CC) for <u>mitochondria</u> up (left) and down (right) subsets of genes for interaction data. NA are 'not available' terms, whereas there is less than ten GO terms enriched after the test or because the gene set was too small to test. But GO terms give clues to the existing biological mechanism after knockdown of SLIMP in S2 cells.

It is seen that the ten genes mito-up produce an enrichment in the selenocysteinylt-RNA(Sec) biosynthetic and an enrichment in selenium transferase activity (due to CG1427). To recover the tRNA aminoacylation activity after the SLIMP knock down condition can seem bizarre at a first sight. As far as for the two genes mito–down, the serine type activity is reported, as expected.

Regarding transcription, mRNA biological process is enriched for mitochondria genes as well as few related molecular functions: mRNA binding, nucleotide binding or small molecule binding.

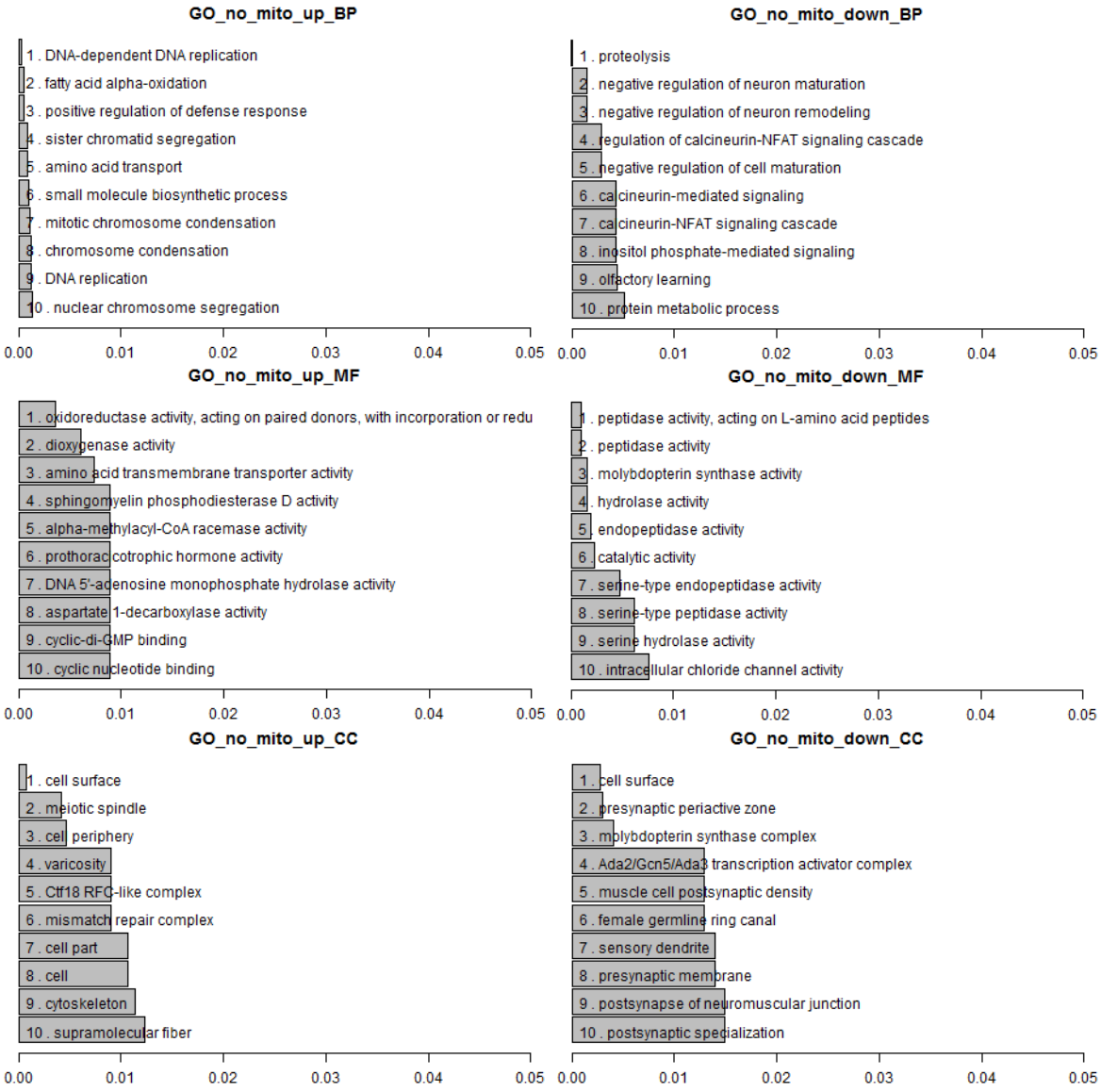


Figure 15. Horizontal bar plots of the ten most significative GO terms (BP, MF, CC) for <u>outside mitochondria</u> up (left) and down (right) subsets of genes for interaction data. Gene Ontology terms are indicators of the existing biological mechanism after knockdown of SLIMP in S2 cells.

From the 24 genes no-mito down-regulated I highlight the enrichment of negative-regulation of cell maturation term, which is in accordance to the arrest at G2 phase for those cells with the knock down of SLIMP, seen in cell culture. Then we see that peptidase activity is enriched as well as the serine-type peptidase activity (again, but outside mitochondria), meaning that less number of proteins present in the cell will be degraded.

From the 140 genes no-mito up-regulated I see terms associated to chromosome condensation, DNA replication and chromatid segregation. This can suggest that maybe part of the G2 cell samples taken were not really in the G2 phase but still in the S phase, which maybe could not be fully completed due to SLIMP knockdown. To confirm this, three biological replicas of the same experiment should be performed (we are working with one biological experiment having three technical replicas (.CEL files) for each sample, See table1).

3.2.2.2. Pathway (KEGG) enrichment analysis.

Because the number of transcripts targeted to mitochondria that are up and down regulated for interaction group (12), are very few to produce KEGG output and build the heatmap-like plot, the equivalent for KD_WT(44 genes, no change phase normalization) result is showed:

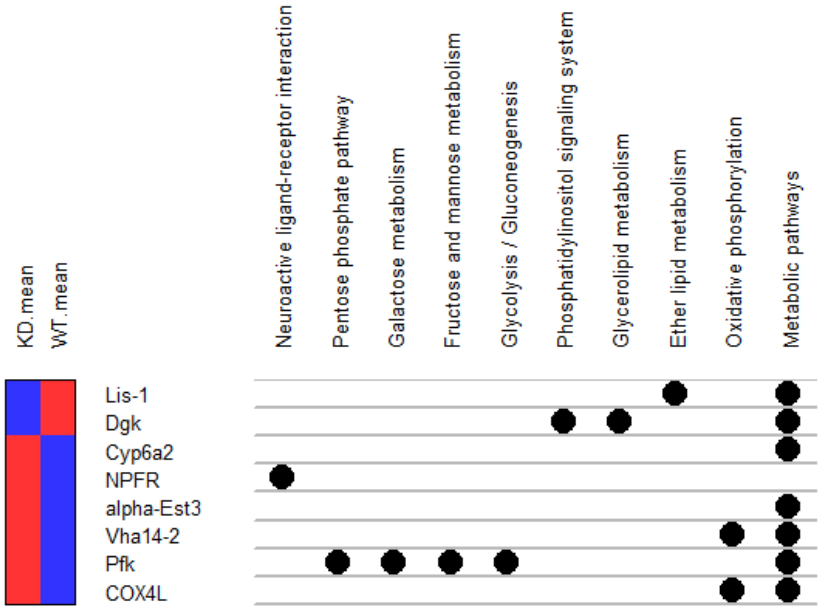
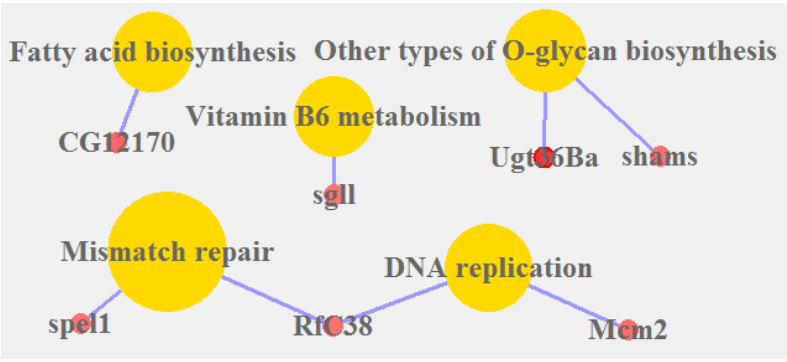


Figure 16. GeneAnswersHeatmap for the KEGG enrichment results in up and down mitochondria targeted gene products from KD_WT. The ten most enriched pathways are present in the X axis by clustering gene order, from the genes in Y. A total of 44 genes were given as input. Red bar means less expression than blue (downregulation of the pathway).

So far, the oxidative phosphorylation pathway is known to be lost and here it appears as expected. Besides, we see that the SLIMP knockdown condition produces a downregulation to many important glucidic pathways (Pentose phosphate, galactose, fructose and mannose and glycolysis/gluconeogenesis) which in turn could provoke the upregulation of glycerolipid and/or ether lipid metabolism. Overall, it could all be a consequence of the lack of ATP production.

Regarding the set of genes from interaction data classified as no mitochondrial, up and down regulated, we have:



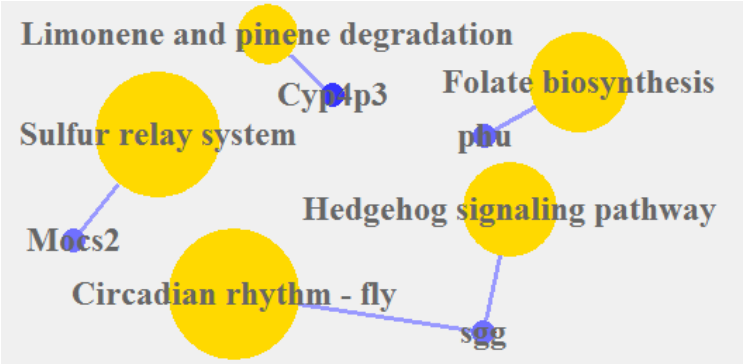


Figure 17. GeneAnswersConceptNet plots for interaction gene sets: no mito up (upper plot, red genes) and no mito down (plot below, blue genes). The five most significant pathway names enriched appear for each plot. The yellow circle size is as bigger as significative is the pvalue of the KEGG enrichment.

From this results we can say that upon genes outside mitochondria that are up-regulated, the DNA replication pathway is enriched. Meanwhile, the fatty acid biosynthesis pathway and other type of O-glycan biosynthesis are enriched too, so a trend to liphidic or other types of O-glycan biosynthesis that are not the current is required. From the pathways enriched by down-regulated genes outside mitochondria, no relations can be inferred.

4. CONCLUSIONS:

The pipeline in R to use TPpred2.0 protein sub-cellular position predictions together with microarray data is reported. Therefore, transcriptome differences studied in microarray data can also be studied in terms of mitochondria targeting status. To adapt this code for a different organism than *Drosophila melanogaster*, only few code should be modified in the Uniprot download R file and in the functional analysis code. Besides, all code used is written in the most easy-to-follow manner for users, with a tidy style and comments thought the lines.

The SLIMP knockdown microarray (Affimetrix *Drosophila* 2) data in S2 cells used in the current project lead to many results. Regarding targeting proportions, we can say that 5.68% (743) of the transcripts are targeted to mitochondria, 2.35% (307) are transcripts not coding for proteins and 91.97% (12033) are transcripts that translate to proteins that do not enter mitochondria. Then, the up and down regulated transcripts for each position were reported in Table 5. For sample interaction, 12 genes DE were defined as mitochondria targeted (10 up, 2 down) and 162 genes DE were defined as NOT mitochondria targeted (140 up, 24 down).

Focusing on SLIMP-involved cell mechanisms, we can say that its knockdown in S2 cells does not lead to significant changes in number of mitochondria-targeted transcripts (up or down) in comparison to all mitochondria targeted transcripts. It neither happens for transcripts with proteins not entering mitochondria for any of the microarray comparisons. Moreover, from the proportions in Figure 11 we conclude that SLIMP knock-down leads to a bigger response in up-regulation than in the number of down-regulated transcripts, for both inside and outside mitochondria transcripts. This trend is not present in G2WT_G1WT or KD_WT (comparisons that consider change of cell phase cycle).

However, we must focus on the analysis step where, from having proportions reported in Figure 8B (5.37%, 702transripts, with proteins from same gene targeting both mitochondria and NOT) we incorporate them as transcripts coding for proteins targeted to mitochondria (from 0.31% to 5.68%). This way we can infer proportions in a microarray-wide approach but we are losing the biologic information associated to genes coding for proteins targeted to mitochondria or NOT depending on the alternative splicing process of the transcripts. Examples of this are protein matrix metalloproteinase (Mmp1) (37949 gene: 2 proteins NOT, 4 mitochondrial) or CG1427 protein (40681 gene: 1 mitochondria, 2 proteins NOT). Nevertheless, such cases will always require additional experimental validation to identify which of the potential transcripts are differentially expressed.

Finally, the GO terms analysis reported peptidase as well as serine-peptidase activity for down-regulated NOT mitochondria targeted genes. It also reported enrichment in Selenocysteinyl tRNA(Sec) biosynthetic process as well as selenium transferase activity due to CG1427* (an up-regulated mitochondria targeted protein). This results is interesting due to the functional interaction between SLIMP and mitochondrial seryl-tRNA synthetase, a central component of this pathway that, consistent with previous data, is not found changed in this microarray data analysis. However, the increase in expression of CG1427 suggests that the depletion of SLIMP generates an imbalance in serine metabolism that extends to the selenocysteine biosynthetic pathway.

4.1. Subset functional annotation analysis vs. GSEA

The GSEA is a genome-wide analysis from all the microarray data as a single set of genes. The transcripts are ordered by fold change, and the enrichment score for functional analysis terms is calculated. In addition, the GSEA associated to the current microarray data reports Hallmark (GO terms universe simplified) and Slim (all three categories mixed) GO categories as well as the three main ones: BP, MF and CC. To use MSigDB data is required in doing Hallmark. Besides, gene homology from fly to human is required and a GO enrichment function in R allowing for a hallmark GeneSetCollection is required as well. (None is the GOStats used). Because of that, a direct comparison of GSEA results vs. pipeline results at a GO hallmark level is not reported. The E2F1 pathway was not present in the pipeline results.

REFERENCES:

[1] Guitart, Tanit, et al. "New aminoacyl-tRNA synthetase-like protein in insecta with an essential mitochondrial function." Journal of Biological Chemistry 285.49 (2010): 38157-38166.

[2] Debard, Sylvain, et al. "Nonconventional localizations of cytosolic aminoacyl-tRNA synthetases in yeast and human cells." Methods 113 (2017): 91-104.

[3] Subramanian, Aravind, et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proceedings of the National Academy of Sciences 102.43 (2005): 15545-15550.

[4] Indio, Valentina, et al. "The prediction of organelle-targeting peptides in eukaryotic proteins with Grammatical-Restrained Hidden Conditional Random Fields." *Bioinformatics* 29.8 (2013): 981-988.

[5] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

[6] Carlson M (2017). UniProt.ws: R Interface to UniProt Web Services. R package version 2.16.0.

[7] Carlson M and Pages H (2017). AnnotationForge: Code for Building Annotation Database Packages. R package version 1.18.2.

[8] Carlson M (2016). drosophila2.db: Affymetrix Drosophila Genome 2.0 Array annotation data (chip drosophila2). R package version 3.2.3.

[9] Morgan M, Falcon S and Gentleman R (2017). GSEABase: Gene set enrichment data structures and methods. R package version 1.38.1

[10] Falcon S and Gentleman R (2007). “Using GOstats to test gene lists for GO term association.” Bioinformatics, 23(2), pp. 257-8.

[11] Huang L, Feng G, Du P, Xia T, Wang X, Jing, Wen, Kibbe W and Lin S (2014). GeneAnswers: Integrated Interpretation of Genes. R package version 2.18.0.

[12] Vogel, Christine, and Edward M. Marcotte. "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses." *Nature reviews. Genetics* 13.4 (2012): 227.

[13] FlyBase Consortium. "The FlyBase database of the Drosophila genome projects and community literature." Nucleic acids research 31.1 (2003): 172-175

[14] Feng, Gang, et al. "A collection of bioconductor methods to visualize gene-list annotations." BMC research notes 3.1 (2010): 10.

[15] Khatri, Purvesh, and Sorin Drăghici. "Ontological analysis of gene expression data: current tools, limitations, and open problems." Bioinformatics 21.18 (2005): 3587-3595.