Identification of mRNA expression patterns in the human blood thanatotranscriptome: Dead, but not entirely

Laura G. Antiga

KU LEUVEN

upf. Universitat Pompeu Fabra Barcelona

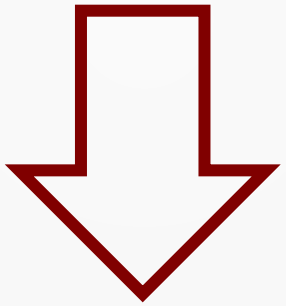
September 2020

BSc in **Genetics**
MSc **Bioinformatics for Health Science** in
Universitat Pompeu Fabra (UPF), Barcelona, Spain

Internship in the **forensic department** of KU Leuven, Belgium
Academic year 2019-2020
Supervisor: Bram Bekaert

PhD Lode Sibbens

Study of cyrcadian rythms with **post-mortem RNA-Seq** data



Collaboration using a dataset meant for a sleep pattern study

Index

1. Introduction

- **Thanathotranscriptomics** is the RNA analysis of post-mortem (PM) tissues.
  - In this project: PM human blood from individuals who donated their body to science

- The **post-mortem interval (PMI)** is the time elapsed from time of death (ToD) of an individual until the discovery of the body.
  - Common methods: *rigor*, *algor* and *livor mortis*

They have become a topic of interest in forensic science due to the essential information they can provide in forensic case investigations.

Background

- Several studies have previously investigated the effect of death on gene transcription, but it has never been conducted with samples of the same individual.
- Different patterns of gene expression after death
- Biological processes are still active up until 48h after death
- Genotype-Tissue Expression (GTEx) Project data to predict PMI

Objectives

First study to perform such transcriptome analysis on blood obtained from a longitudinal sampling procedure on deceased human individuals.

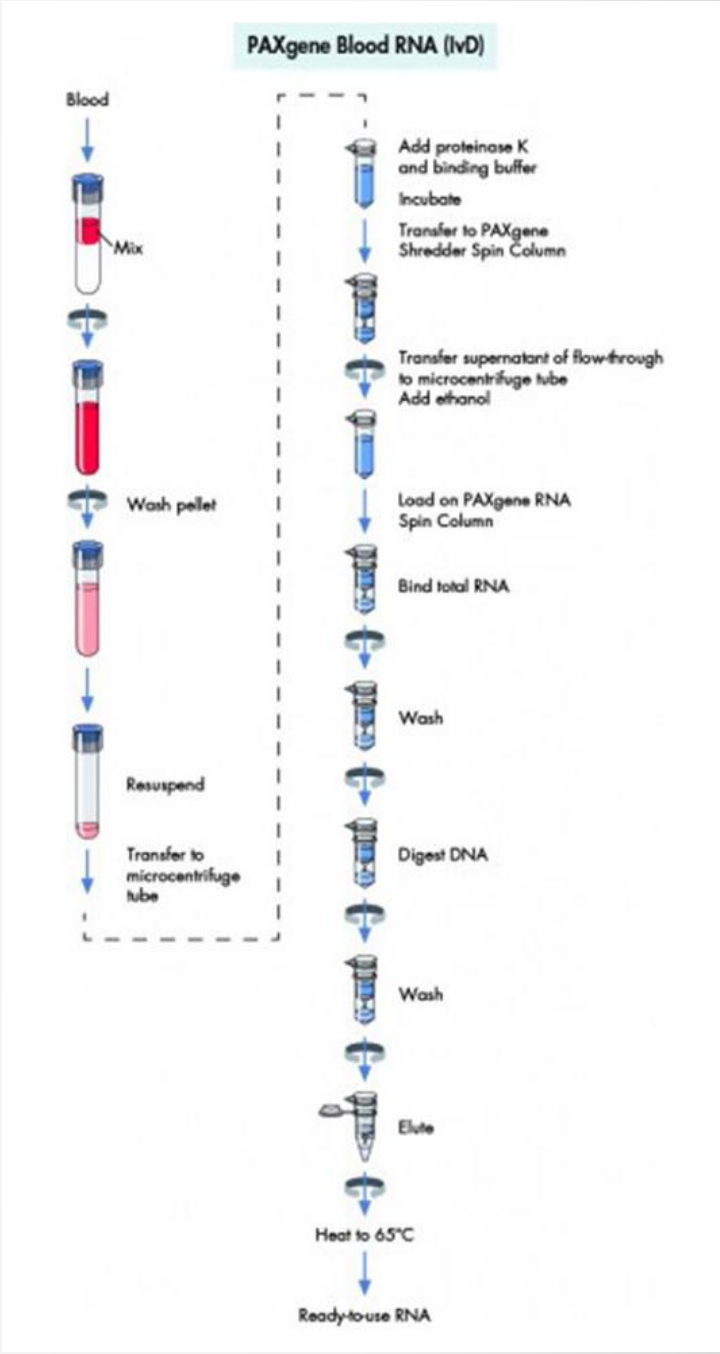
I. Identify **up or downward clusters** or patterns in the human thanatotranscriptome
II. Perform **gene ontology enrichment analysis** on the gene clusters identified in the first objective
III. Develop a **PMI prediction model** based on PM gene expression patterns.

2. Methodology - Sample collection

- 7 deceased individuals with a known time of death who donated their body to science.
- Kept at room temperature (18°C) for the entire sampling procedure.
- Blood samples were taken during a time frame of 24h starting upon arrival to the mortuary and with a 3h interval.
- Blood was alternately collected from the *vena femoralis* (left and right) and *vena subclavian* (left and right).
- Ethical approval for this study was obtained from the Ethical Commission of University Hospital Leuven (case number S58486).

2. Methodology – RNA preparation and sequencing

- RNA extraction was performed using the PAXgene Blood RNA Kit (IVD) according to manufacturer's instructions.
- RNA quantification was performed using NanoDrop spectrophotometry (Thermo Fisher Scientific) and RNA quality was assessed on the Bioanalyzer 2100 using the RNA 6000 Nano Kit (Agilent).
- cDNA libraries were created using the **QuantSeq 3′ mRNA-Seq** Library Prep Kit FWD for Illumina (Lexogen) including the **Globin Block module** to remove the majority of the globin mRNA transcripts.
- cDNA pools were quantified using the Qubit dsDNA HS assay (Thermo Fisher Scientific).
- Sizing of the pools was performed on the BioAnalyzer 2100 using the High Sensitivity DNA assay (Agilent). Final cDNA libraries were pooled and sequenced on the Illumina NextSeq platform (single read, high output, 75 bp).



QuantSeq 3′ mRNA-Seq

Step 1: Reverse Transcription

4h 30min

The kit uses total RNA as input, hence no prior poly(A) enrichment or rRNA depletion is needed.

Step 1: Reverse Transcription

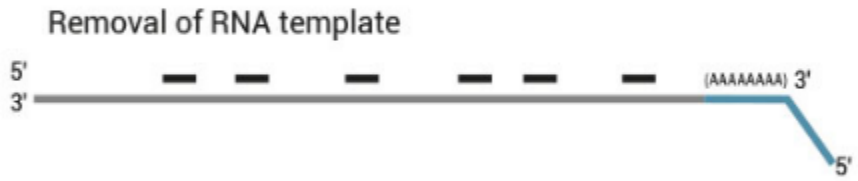
40min

4h 30min

Library generation starts with oligodT priming containing the Illumina-specific Read 2 linker sequence.

Step 2: Removal of RNA

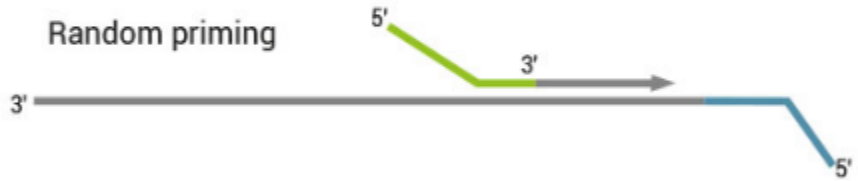
1h — 4h 30min



After first strand synthesis the RNA is removed.

Step 3: Second-Strand Synthesis

2h 30min — 4h 30min



Second strand synthesis is initiated by random priming and a DNA polymerase. The random primer contains the Illumina-specific Read 1 linker sequence. At this step Unique Molecular Identifiers (UMIs) can be introduced by exchanging the Second Strand Synthesis Mix 1 (SS1) from the standard QuantSeq FWD Kit with UMI Second Strand Synthesis Mix (USS).

Step 4: Library Amplification

4h 30min

During the library amplification step sequences required for cluster generation are introduced.

Step 5: Sequencing

NGS reads are generated towards the poly(A) tail and directly correspond to the mRNA sequence. To pinpoint the exact 3' end, longer reads may be required (SR50, SR100, SR150). Although paired-end sequencing is possible, we do not recommend it for QuantSeq FWD. Read 2 would start with the poly(T) stretch, and as a result of sequencing through the homopolymer stretch, the quality of Read 2 would be very low.

2. Methodology – first data processing

- BlueBee platform: **quality control** steps, **trimming** and read **alignment** to the human reference genome (GRCh38) using STAR alignment.
- Raw read counts **were normalized** in BlueBee using the DeSeq2 pipeline using the **median of ratio's** method .
- Only gene transcripts with normalized expression values in >90% of samples were retained for further analysis to eliminate low expressed genes.

For more information about normalization methods visit:

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

2. Bioinformatics Workflow

Clustering assessment
- Hopkins statistic
- Principal Component Analysis (PCA)

Gene expression patterns
- ANOVA-Dunnett's test
- Linear Regression

Biological processes
- Gene Ontology enrichment analysis

PMI prediction model
- Generalised linear model (glmnet)

Section I | Section II | Section III

Data set overview

Initial number of transcripts: 60,999

Transcripts expressed in, at least, 90% of samples: 10,635

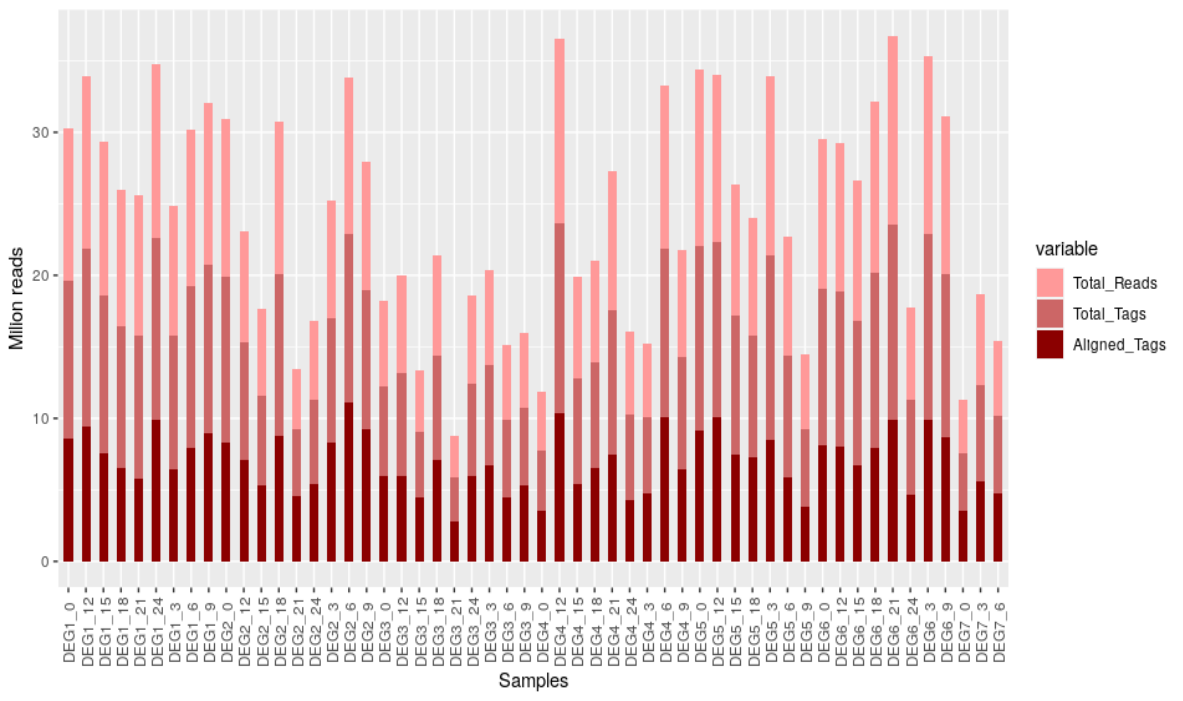


Figure 1. Samples coverage

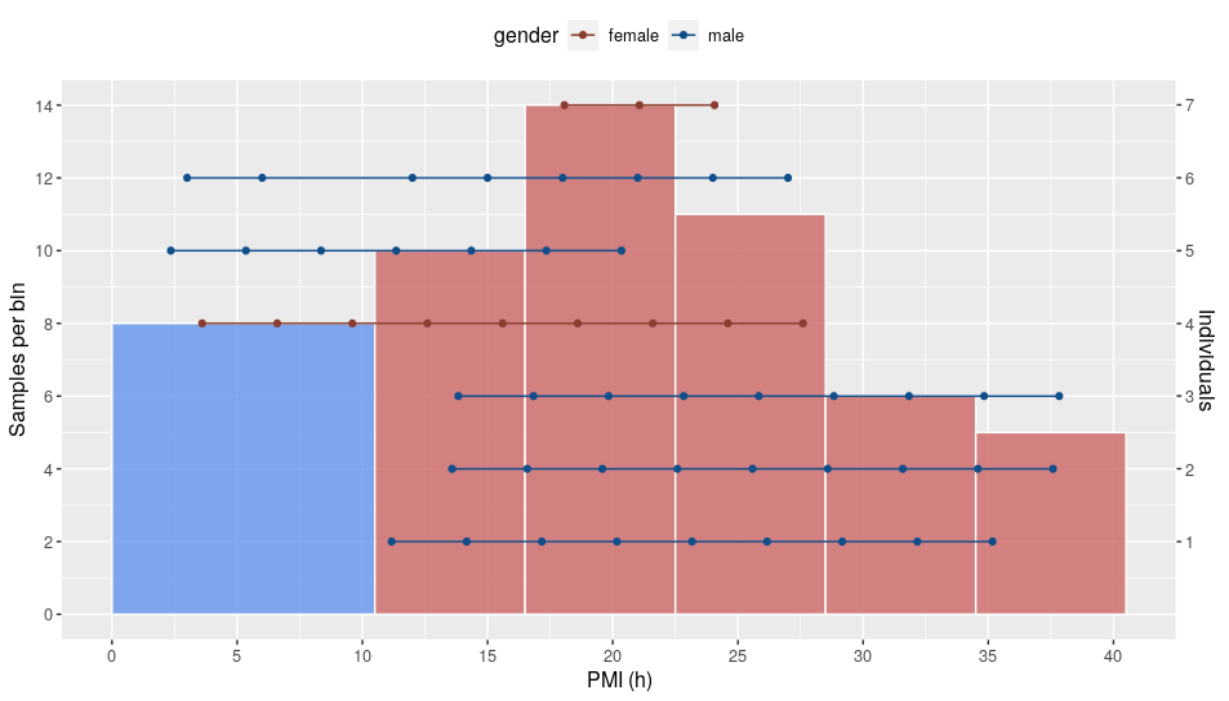
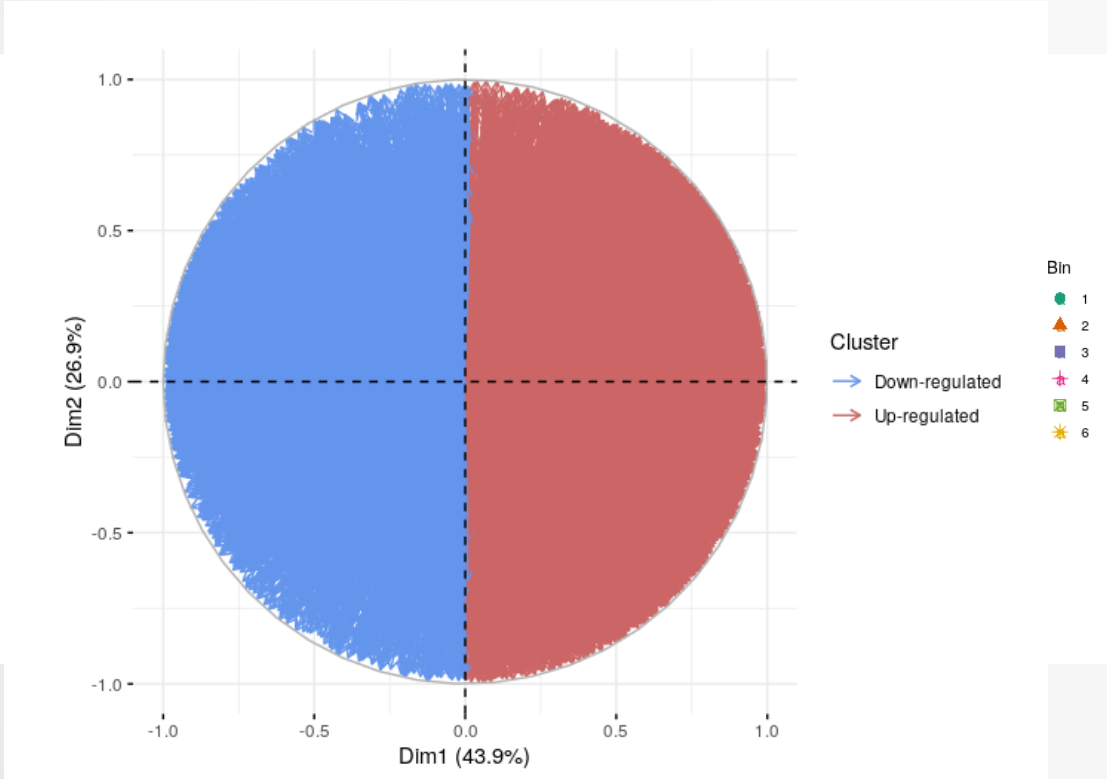


Figure 2. Samples distribution per bin and PMI

Identification of PM gene expression patterns

Clustering assessment

Hopkins statistic: 0.60



transcripts as variables.

Figure 4. PCA of the significant transcripts according to ANOVA-Dunnett's test, before power (10,635)

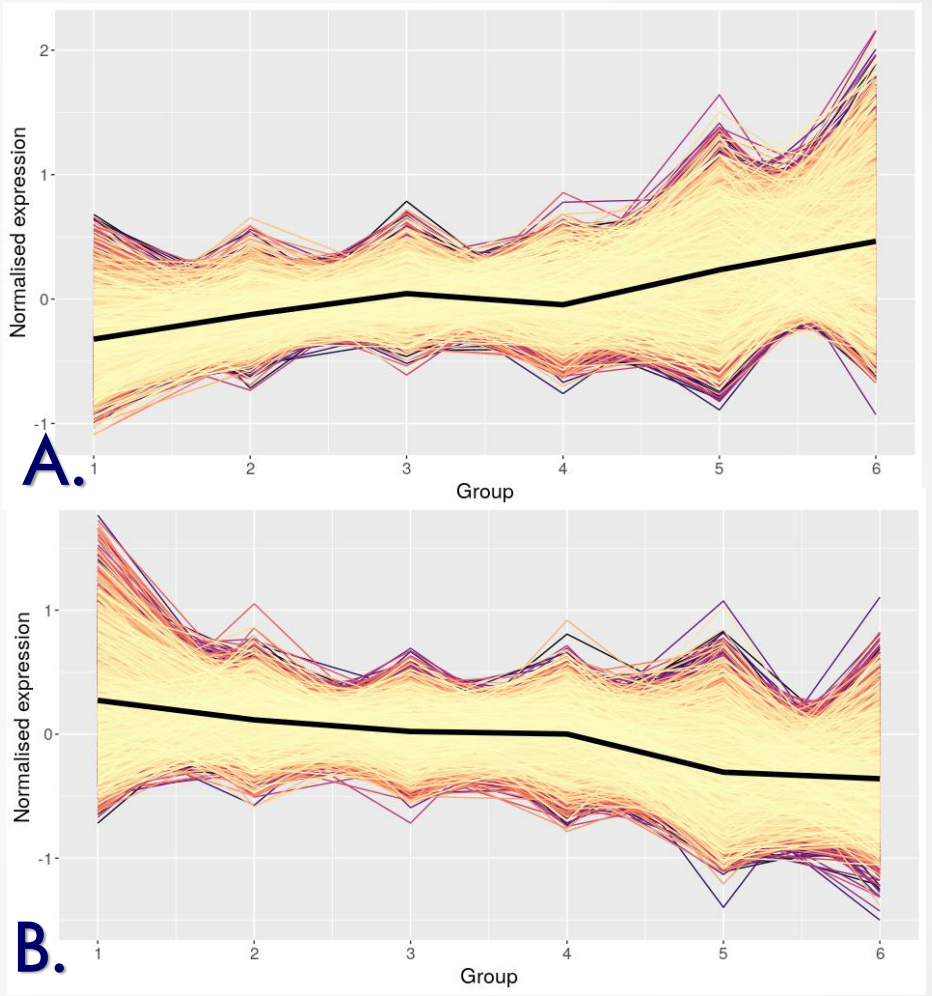


Figure 5. Mean expression, per bin, of all transcripts (10,635) in the two clusters found with PCA: up-regulated (A) and down-regulated (B)

ANOVA-Dunnett's test

Table 1. Significant transcripts according to ANOVA-Dunnett's test, before and after filtering with power analysis.

	Up-regulated	Down-regulated	Total
ANOVA-Dunnett	644	414	1058
After power	254	58	312

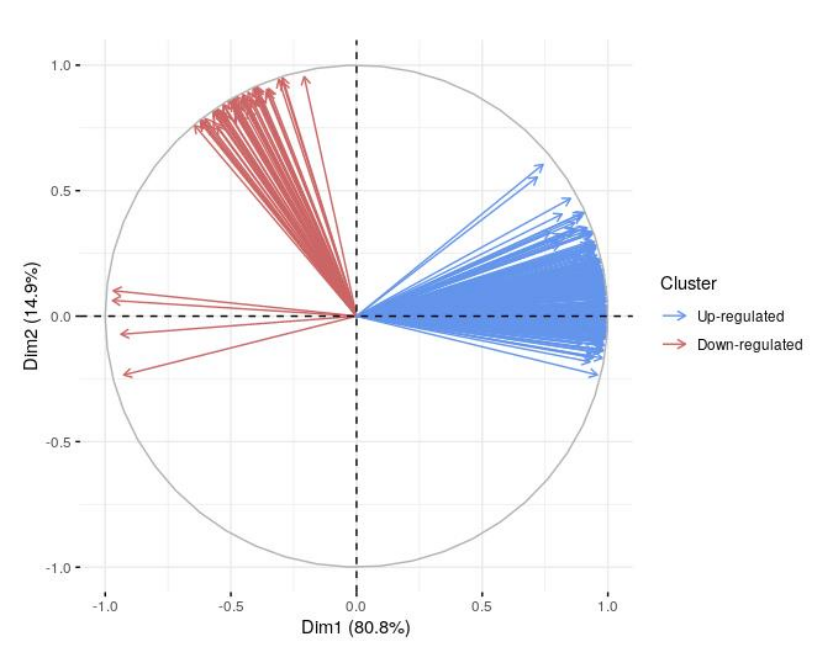


Figure 6. PCA of the significant transcripts according to ANOVA-Dunnett's test, after power (312) as variables.

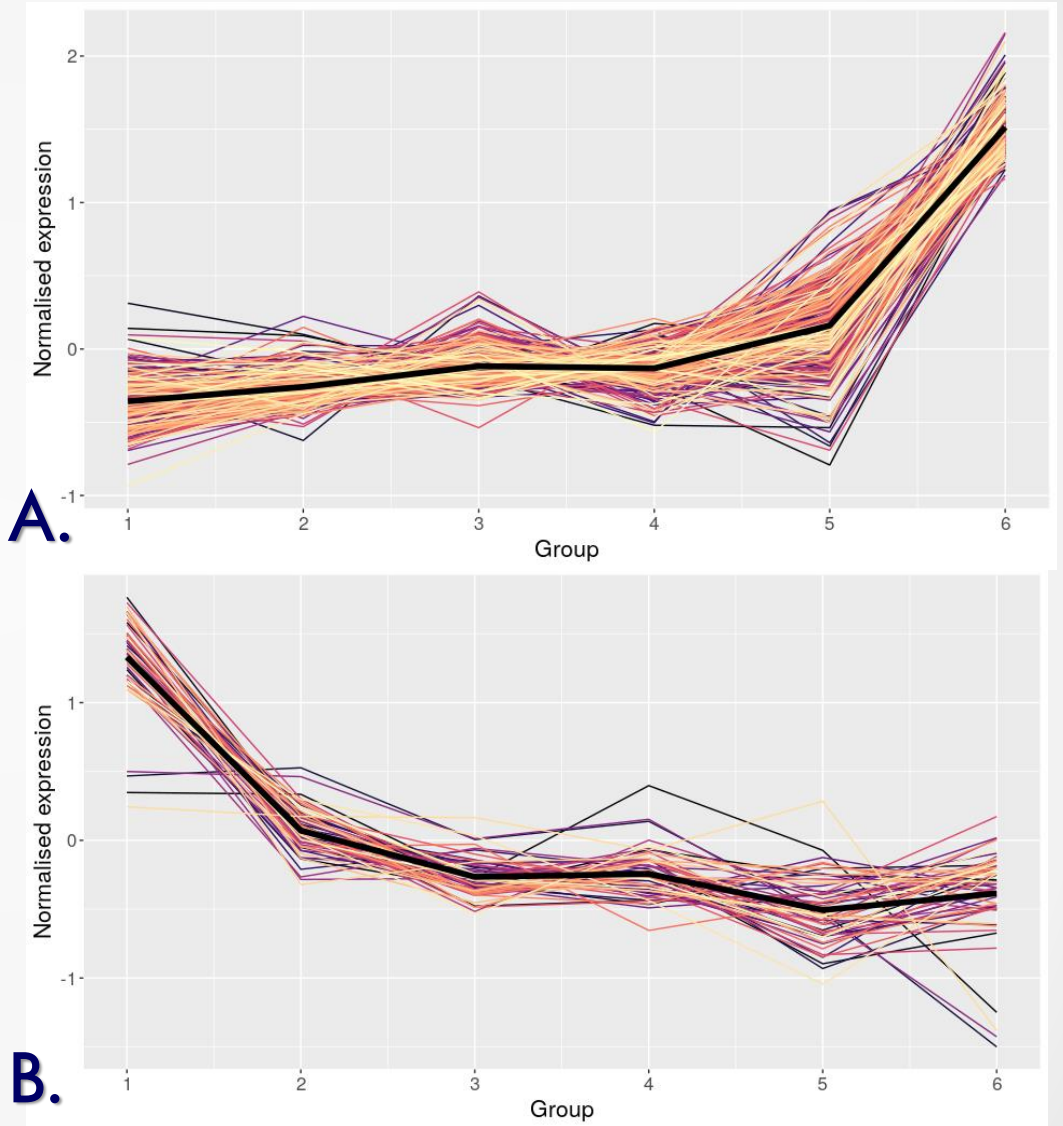


Figure 7. Mean expression, per bin, of ANOVA significant transcripts, after power (312) in the two clusters found with PCA: up-regulated (A. 254 transcripts) and down-regulated (B. 58 transcripts)

Linear regression analysis

Table 2. Significant transcripts according to regression analysis, before and after filtering with power analysis.

	Up-regulated	Down-regulated	Total
Regression	1416	1263	2679
After power	499	465	964

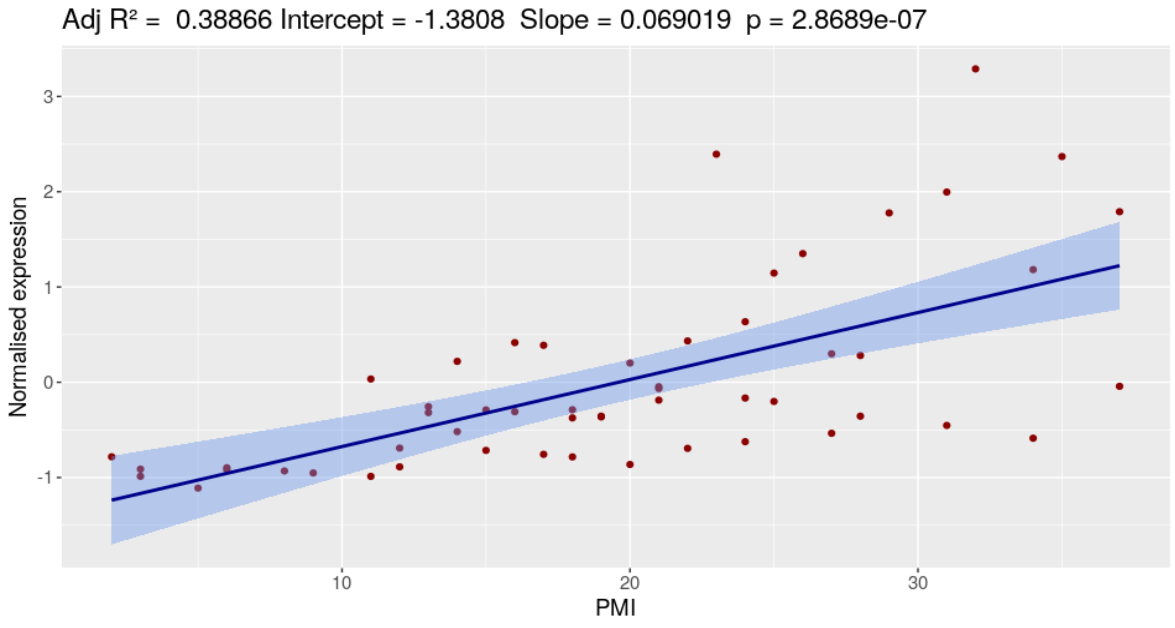


Figure 8. Example of linear regression analysis, for transcript ENSG00000184260.

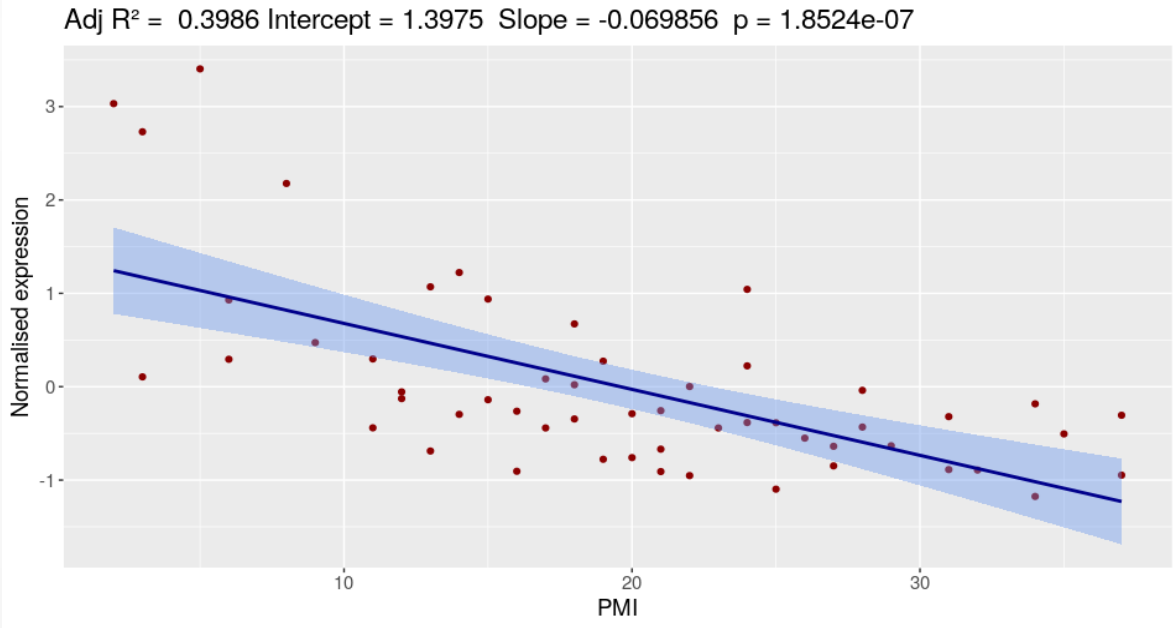


Figure 9. Example of linear regression analysis, for transcript ENSG00000171860.

Methods comparison

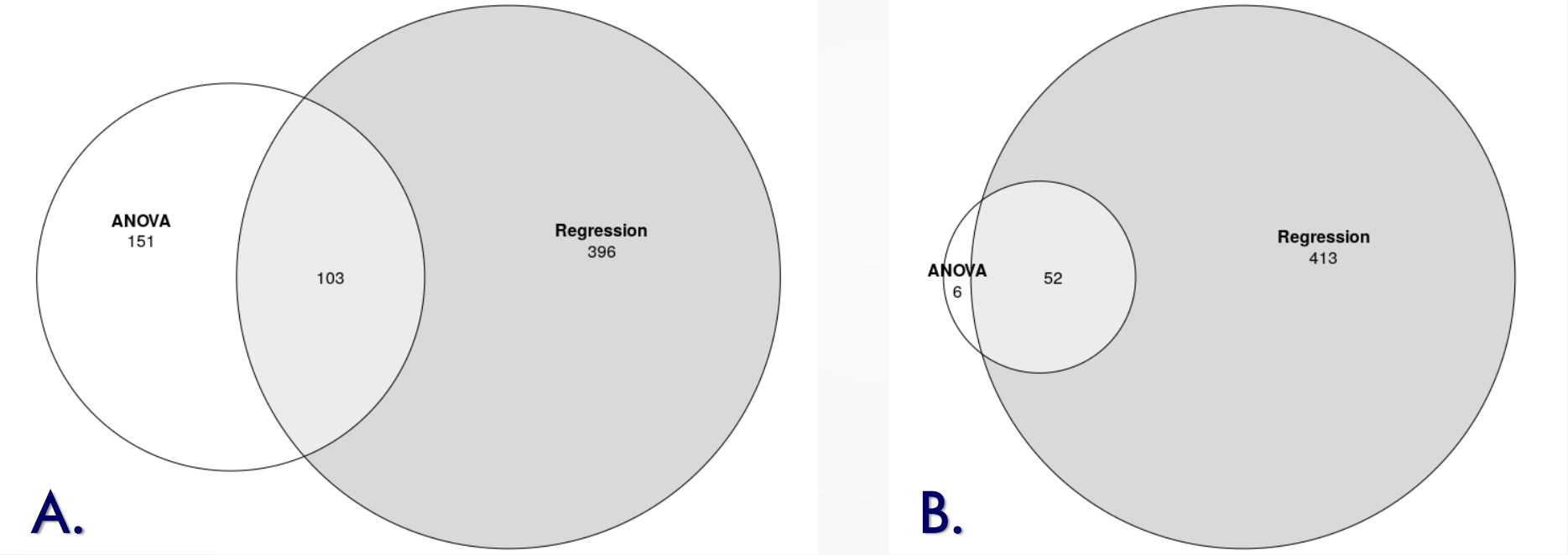


Figure 10. Venn's diagram of the common significant transcripts for ANOVA-Dunnett and regression analysis, after power. A. shows the results of the up-regulated cluster (650) and B. shows the results for the down-regulated cluster (471).

II. GO enrichment analysis

Up-regulated cluster sub-patterns

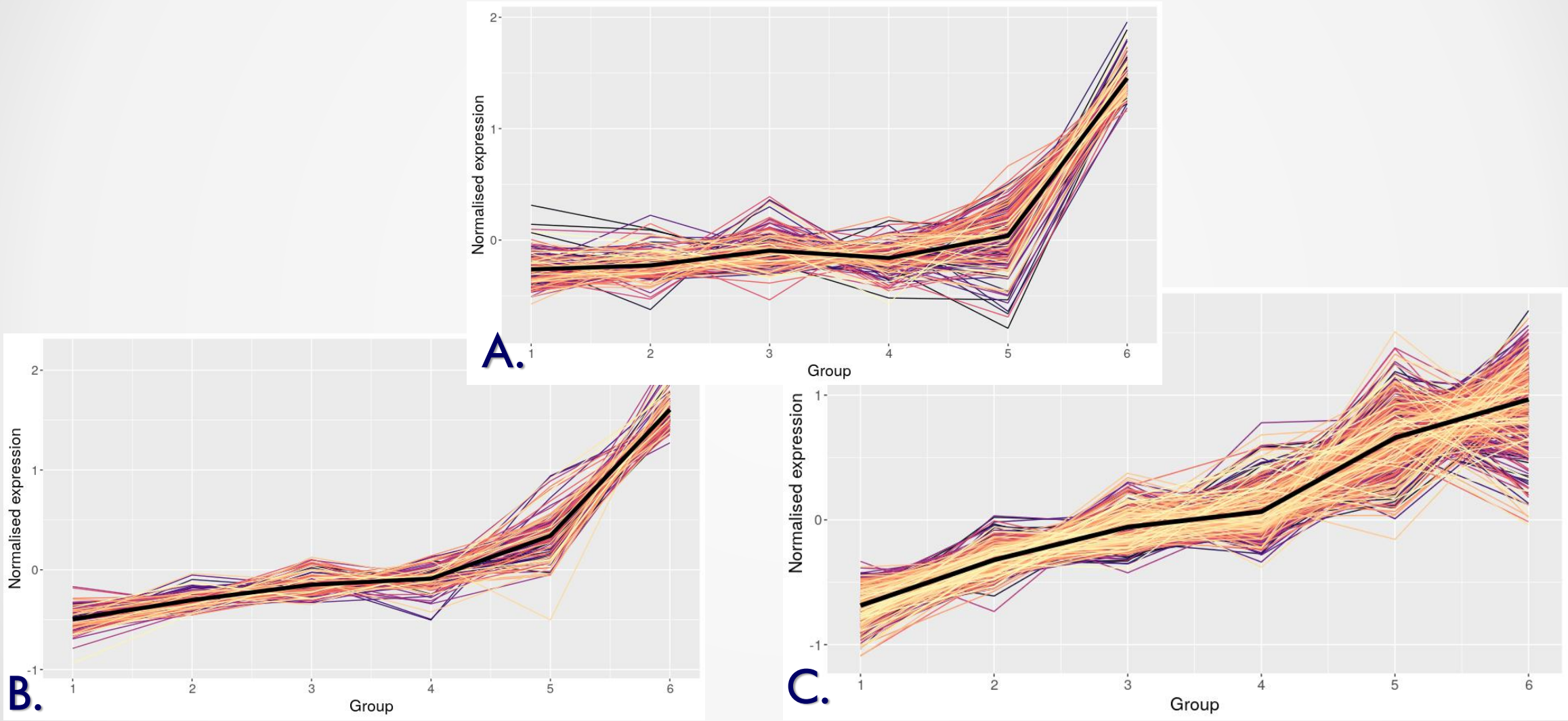


Figure 11. Mean expression, per bin, of significant transcripts, after power, found with A. only ANOVA-Dunnett (151), B. both methods (103) and C. only linear regression (396) in the up-regulated cluster

Up-regulated cluster

ANOVA-Dunnett's test

- 254 transcripts
- **Blood processes** (e.g. hemoglobin metabolic process and oxygen transport)
- **Cell localization processes** (e.g. mitochondrial transport along microtubule and the establishment of mitochondrion localization)

- 151 transcripts
- **Blood processes** (e.g. hemoglobin metabolic process and oxygen transport)
- **Heart embryonic development** (e.g. heart looping and embryonic heart tube development)

Linear regression analysis

- 499 transcripts
- Most pathways found were related to stress responses
- One transport process involved (SRP-dependent cotranslational protein targeting to membrane)

Down-regulated cluster

ANOVA-Dunnett's test

- 58 transcripts
- **Cellular response to stimulus** (e.g. dendritic cell chemotaxis, chemokine-mediated signaling pathway, cellular defense response and cellular response to mechanical stimulus)
- **Cellular locomotion** (e.g. dendritic cell chemotaxis and regulation of mononuclear cell migration)

Linear regression analysis

- 465 transcripts
- **Immune system processes** (e.g. neutrophil homeostasis and natural killer cell activation)
- **Response to stimulus** (e.g. negative regulation of platelet aggregation, dendritic cell chemotaxis and positive regulation of double-strand break repair via homologous recombination)

III. PMI prediction model

PMI prediction model

- Generalised linear model (**glmnet**)
- **10-repeat 5-fold cross-validation** (not enough samples to do training and test sets)
- **RMSE** as validation parameter

Table 3. Mean of the lowest RMSE, for each repetition and every alpha tested in the model. Standard deviation, best lambda and number of transcripts included in the model are also included.

Alpha	RMSE	sd	N° of factors	Lambda
0	5.03	0.28	10635	63.57
0.1	5.43	0.29	110	4.09
0.2	5.15	0.32	79	0.37
0.3	5.04	0.31	65	0.26
0.4	4.88	0.44	36	0.67
0.5	5.55	0.31	33	0.16
0.6	6.11	0.46	27	1.14
0.7	5.34	0.24	25	0.25
0.8	5.97	0.26	21	0.71
0.9	5.78	0.13	20	1.45
1	5.60	0.30	19	0.21

PMI prediction model

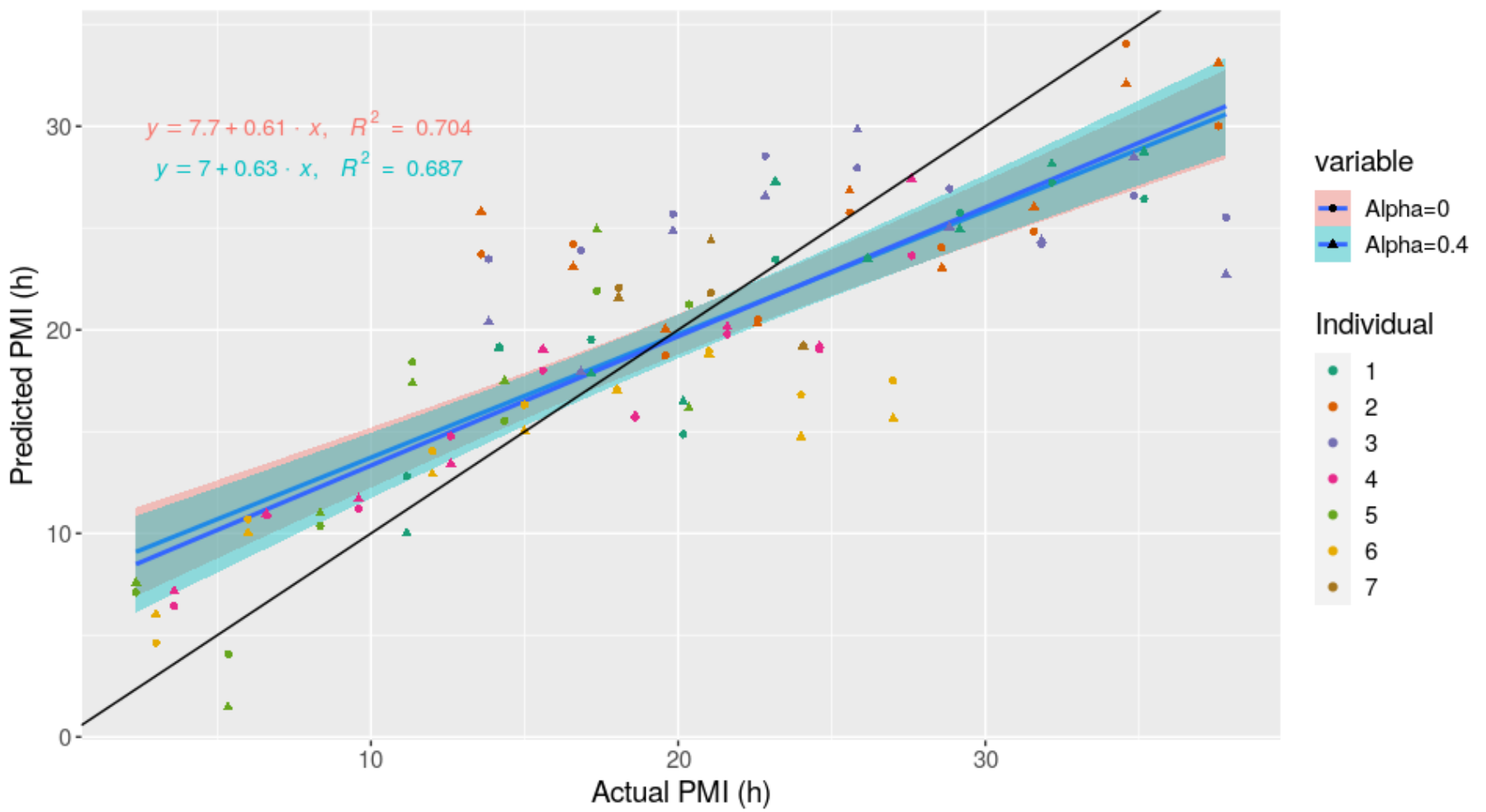


Figure 12. Actual PMI vs. mean of the predicted value for the 10 repetitions with the two models that returned the lowest RMSE (alpha = 0 and alpha = 0.4). The regression of both models is also shown.

3. Conclusion

- There is, indeed, differential expression of gene transcripts after death
- Two clear main behaviours of up and down-regulated clusters
- More than one pattern of gene expression
- GO enrichment analysis supports previous findings
- PMI prediction model optimised for our data set
- Lack of samples was the main limitation (ANOVA and PMI prediction model)
- Further studies are required with larger data sets

Collaborators:

- Lode Sibbens: Conceptualization, Methodology, Sample Collection, Transcriptome Experiments, Funding Acquisition
- Yasmina Abakkouy: Methodology, Analysis
- Ronny Decorte: Commenting on Manuscript
- Wouter Van Den Bogaert: Methodology, Commenting on Manuscript
- Wim Van de Voorde: Commenting on Manuscript, Funding Acquisition
- Bram Bekaert: Conceptualization, Methodology, Writing, Funding Acquisition

Thank you for your attention

Any questions?

KU LEUVEN

upf. Universitat Pompeu Fabra Barcelona

September 2020