Skoltech
Skolkovo Institute of Science and Technology

Ageing transcriptome meta-analysis reveals similarities and differences between key mammalian tissues

Palmer, Daniel et al., *Aging*, vol. 13,3 (2021): 3313-3341

Andrei Stapran, MSc-2, Life Sciences
Daria Kozhevnikova, MSc-1, Life Sciences
Shahzeb Khan, MSc-2, Life Sciences

Outline of the original study:

Meta-analysis on 127 publicly available microarray and RNA-Seq datasets from mice, rats and humans.

1. Performed **linear regression** for each gene inside each dataset

$$Y_{ij} = \beta_{0j} + \beta_{1j} Agei + \epsilon_{ij}$$

2. Selected genes statistically significantly associated with age
3. Meta regression with **Binomial test**
4. Performed **FDR correction** by permutation and sorted the genes by critical P value
5. Identified differentially expressed genes in each set of datasets
6. Identified differentially expressed genes in all datasets combines
7. Enrichment analysis with **David** and **topGO** R tool
8. dN/dS analysis
9. Build **Random Forest ML** models to identify the most important GO terms
10. Run tissue specificity analysis (calculating the **tau index**)

$$\tau = \frac{\sum_{i=1}^{N} (1 - x_i)}{N - 1}$$

Palmer, Daniel et al. “Ageing transcriptome meta-analysis reveals similarities and differences between key mammalian tissues.” *Aging* vol. 13,3 (2021): 3313-3341. doi:10.18632/aging.202648

Results of the original study:

- Overexpression of immune, proteolytic, stress response genes with age; cell response processes activated in muscle and heart
- Underexpression of metabolic, particularly mitochondrial function, and developmental genes
- Genes differentially expressed with age tend to be highly connected and central in protein-protein networks + this is relevant for genes analyzed in brain tissue (these genes might be involved in cognitive ageing)
- Meta-analysis is biased towards broadly expressed genes (across multiple tissues) - however, these genes might indicate ones crucial for ageing

All tissues (Overexpressed)			
GO.ID	**Term**	***p*-value**	**Precision (0.0251)**
GO:2001198	Regulation of dendritic cell differentiation	2.00e-4	0.613
GO:0071276	Cellular response to cadmium ion	2.90e-5	0.571
GO:0071294	Cellular response to zinc ion	3.00e-6	0.452
GO:0006958	Complement activation, classical pathway	1.30e-6	0.276
O:0051043	Regulation of membrane protein ectodomain proteolysis	3.30e-4	0.267

All tissues (Underexpressed)			
GO.ID	**Term**	***p*-value**	**Precision (0.0090)**
GO:0010510	Regulation of acetyl-CoA biosynthetic process from pyruvate	3.20e-6	0.374
GO:0006122	Mitochondrial electron transport, ubiquinol to cytochrome c	1.60e-9	0.356
GO:0006099	Tricarboxylic acid cycle	1.50e-10	0.281
GO:0006107	Oxaloacetate metabolic process	1.20e-4	0.217
GO:0007528	Neuromuscular junction development	3.06e-2	0.204

Weaknesses of the original study & how to fix them

- Cumulative binomial testing was used to filter genes that are differentially expressed across multiple datasets. However, in cases when dataset number is small, statistical significance of such meta-regression testing is small
- Different technologies used - RNAseq and Microarray

ML approach: identifying most important GO terms

Skoltech
Skolkovo Institute of Science and Technology

- Trying to predict whether gene is over- or underexpressed with age based on GO terms
- Run Random Forest (RF) model with random trees (RTs) where nodes are GO terms and leaves are differential expression status
- After training, the model tries to predict the differential expression status of previously unseen gene. When decision path reaches leaf node - the more frequent differential expression status class is selected. The final prediction of expression status stems from voting of all RTs
- Rule-Based Prediction: build several RF models and for each of them several RTs. For each RT and GO term, calculate how many times GO term is used to split the genes into over- and underexpressed categories. Then, calculate how often these splits result in correct prediction - based on this precision is calculated. Then GO terms are ranked based on their precision

Our approach:

Skoltech
Skolkovo Institute of Science and Technology

Conduct linear regression analysis (statsmodels.OLS) for each gene inside each dataset

Filter genes where linear regression slope has p-values < 0.05

Merge datasets into one - if the gene is studied in one dataset but not in the other, the dataset where it is not studied will be omitted for further analysis

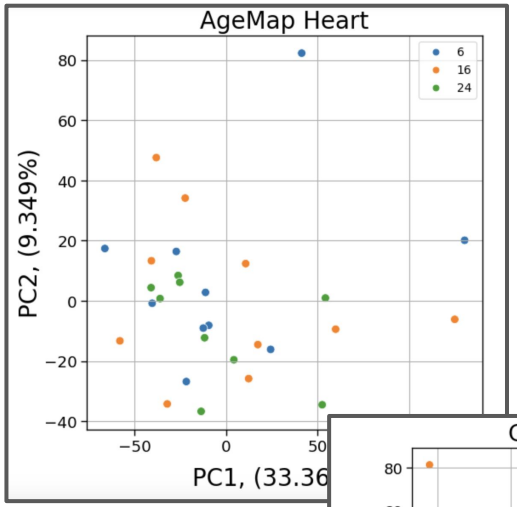
Run PyMare meta-regression analysis on combined dataframe of genes and datasets

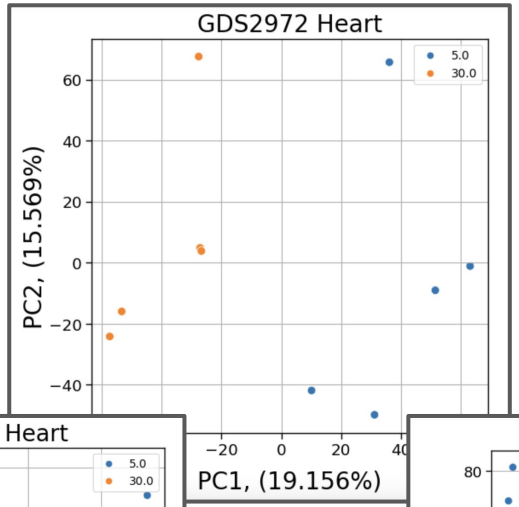
Select genes with adjusted p-values < 0.05 (BH FDR) - diff expressed in at least 2 datasets

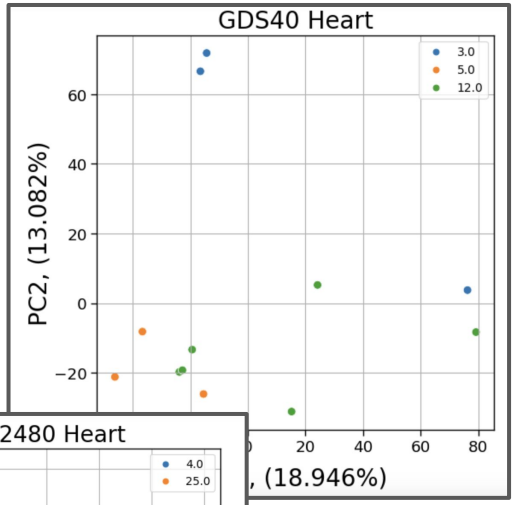
Analyse overlaps in gene lists between tissues & for all tissue combined VS GenAge database

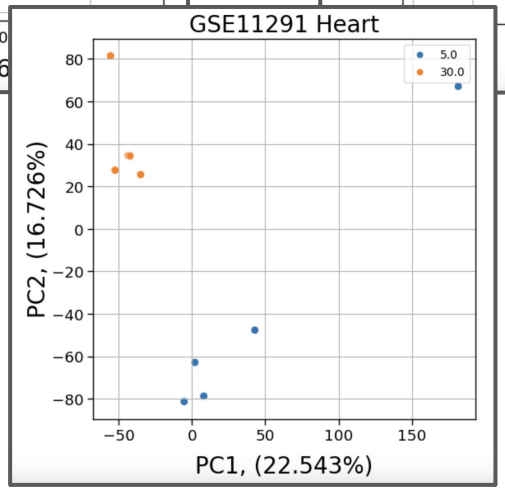
Run binomial cumulative meta-regression testing on combined dataframe of genes and datasets & compare ranks of genes (based on adjusted p-values) between PyMare and binomial approach

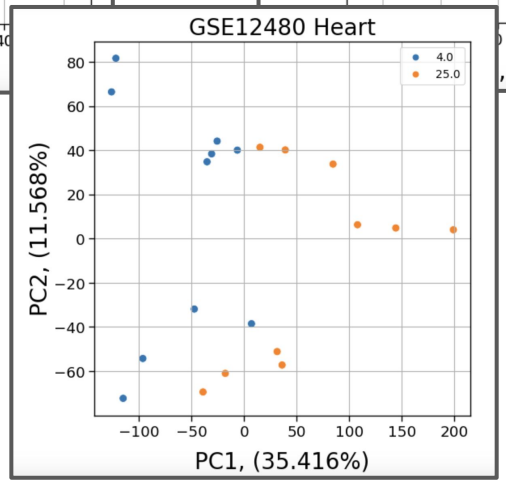
PCA-analysis:



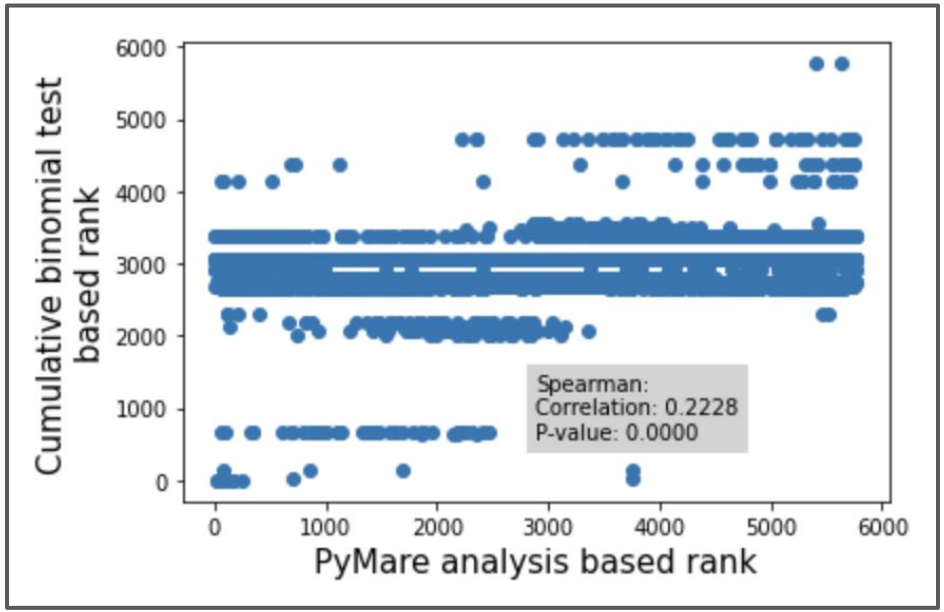




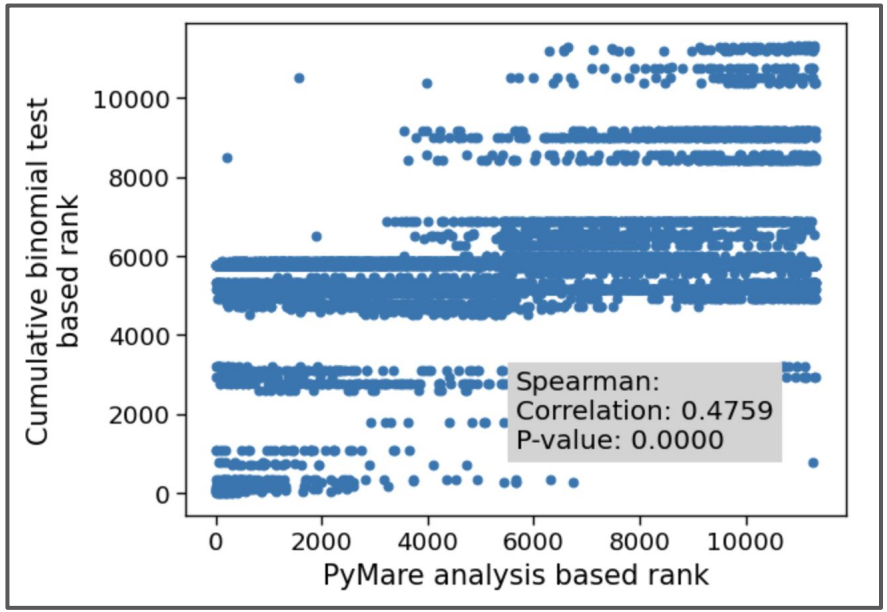




Comparing cumulative binomial test and PyMare meta-regression approaches

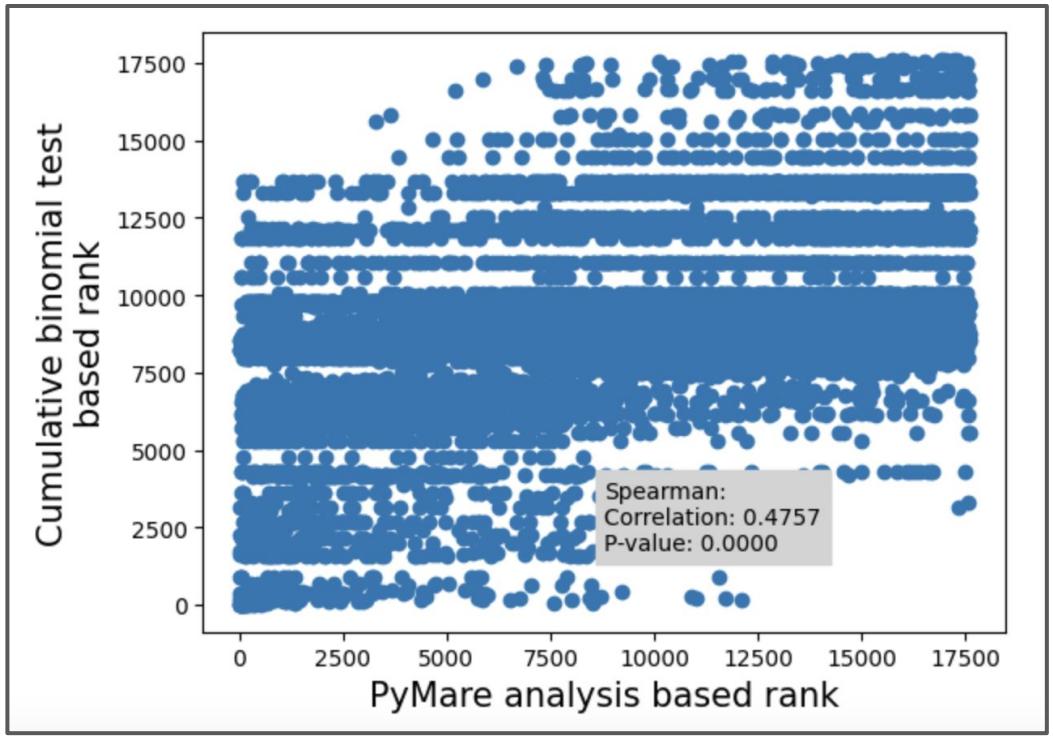


Muscle



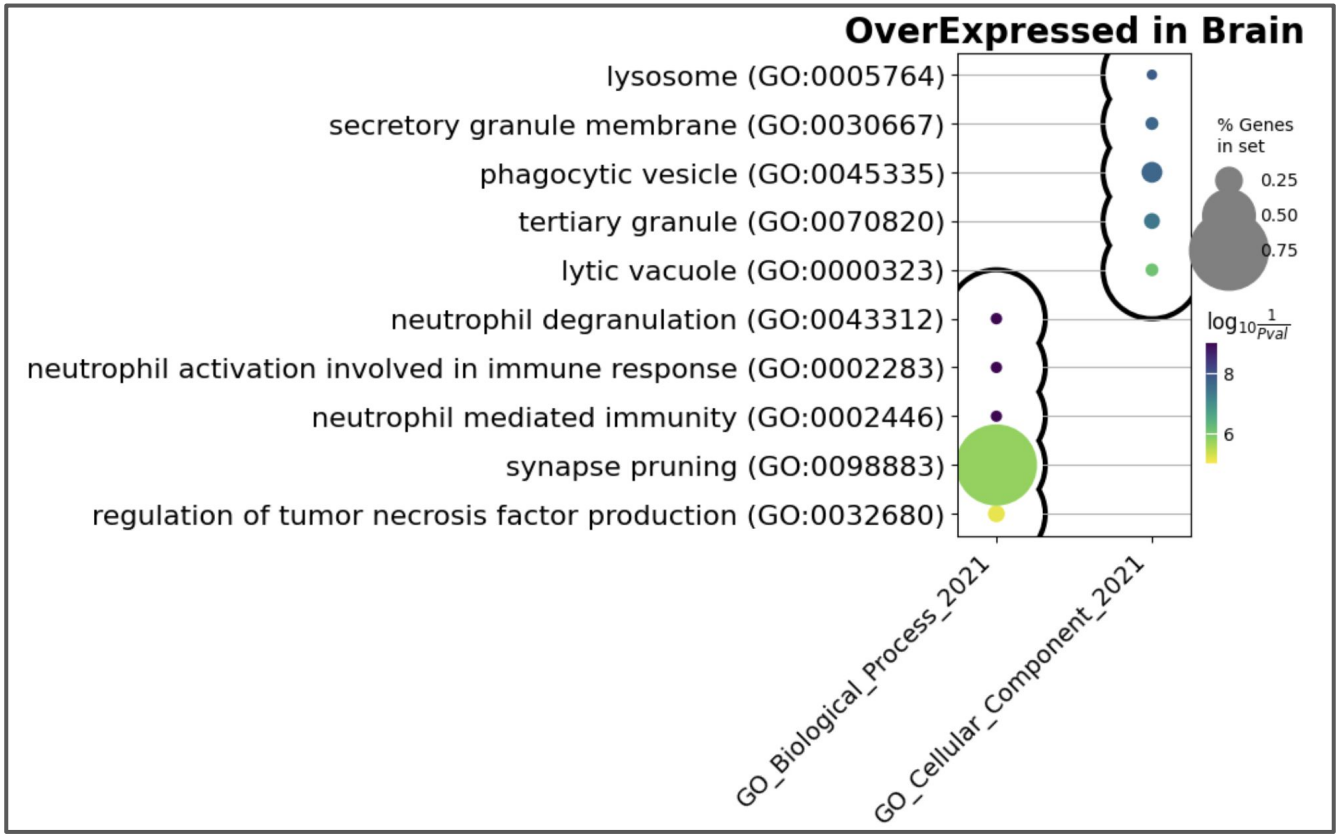
Brain

Comparing cumulative binomial test and PyMare meta-regression approaches

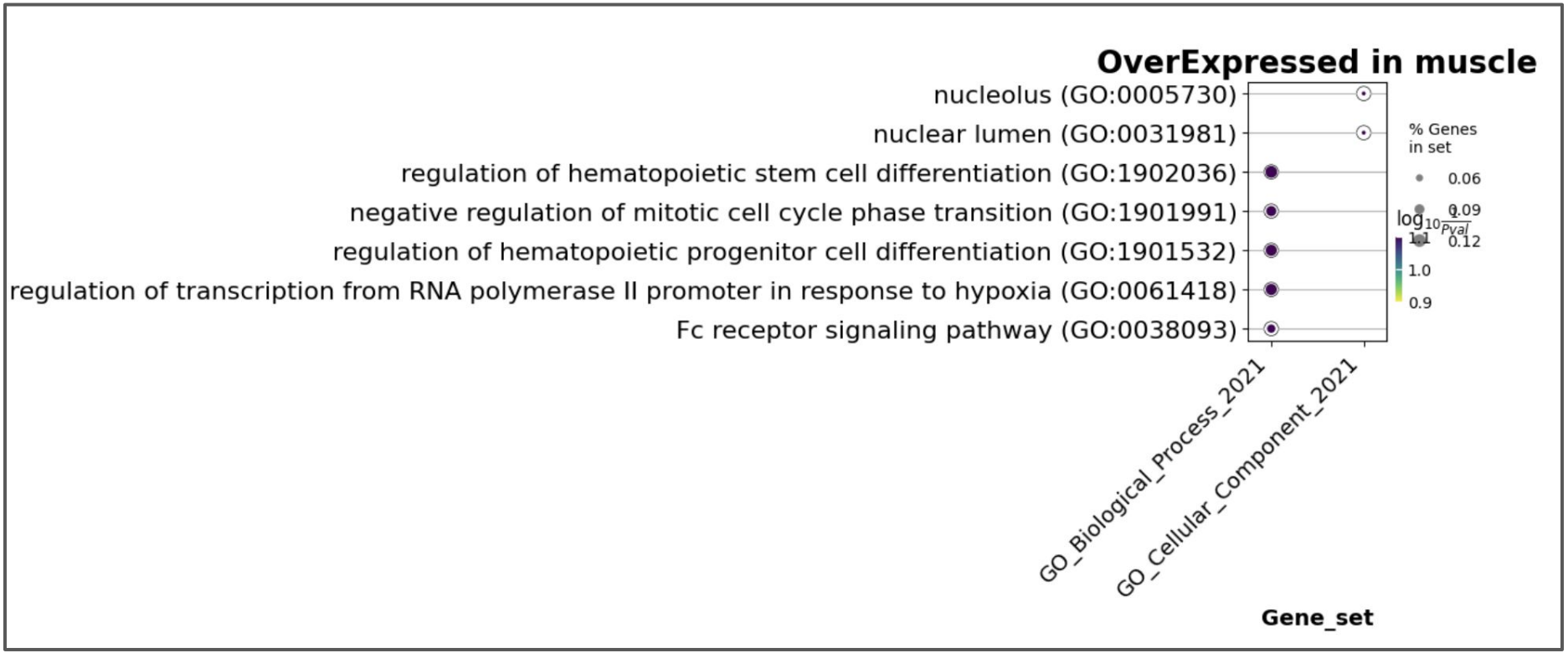


Muscle, Heart, and Brain

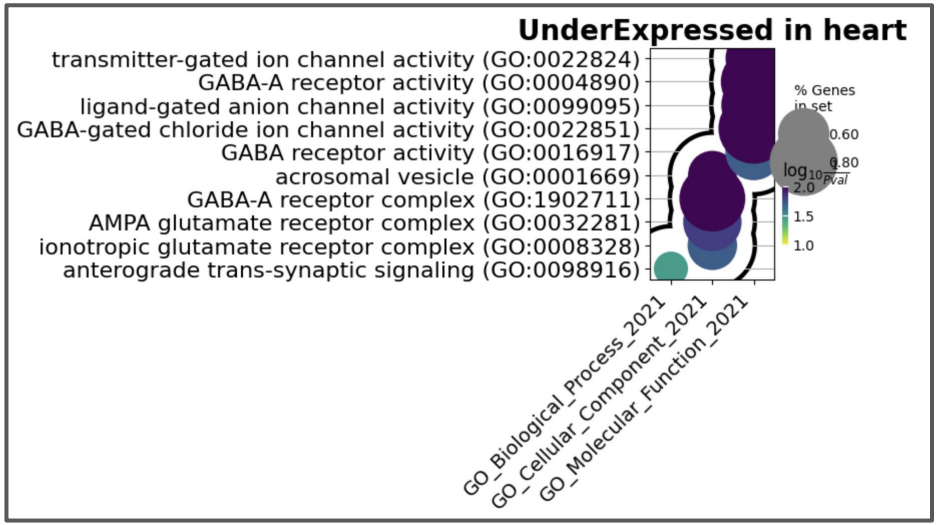
Genes overexpressed with age across multiple datasets ***in brain***



Genes overexpressed with age across multiple datasets ***in muscle tissue***

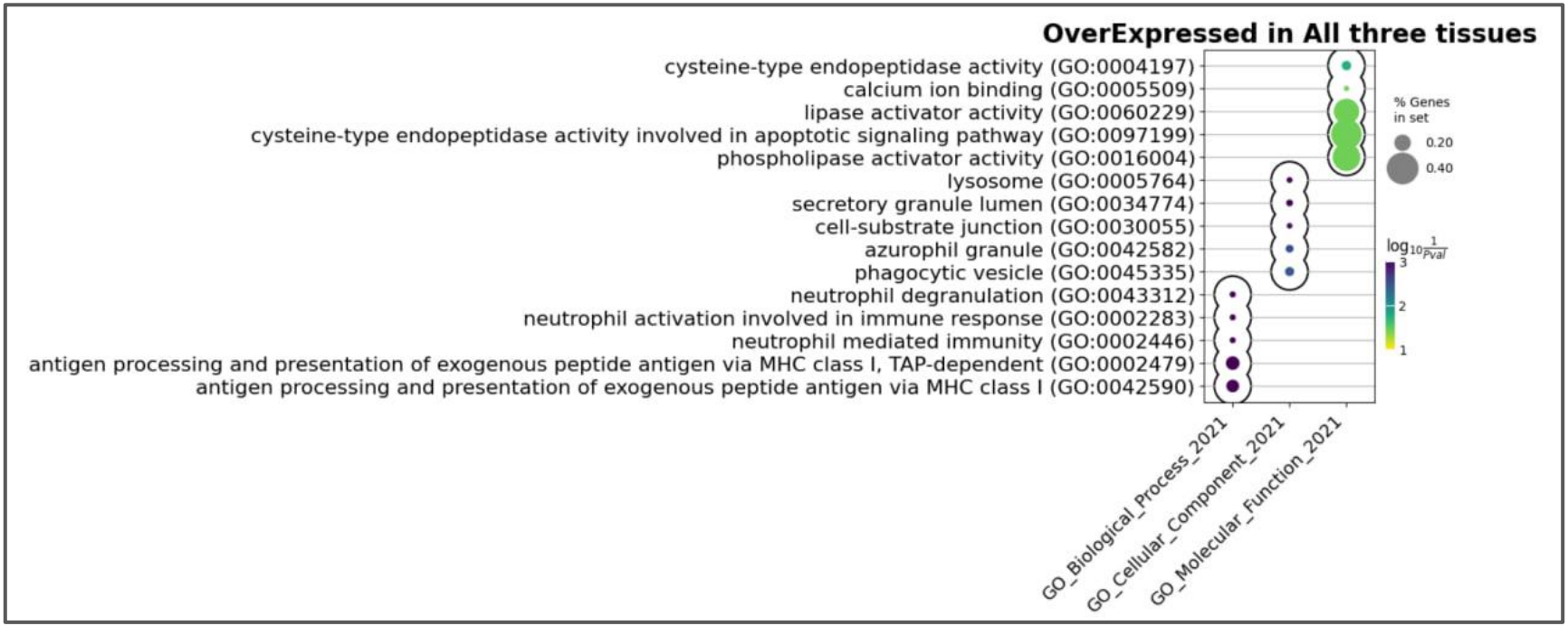


Genes differentially expressed with age across multiple datasets ***in heart***

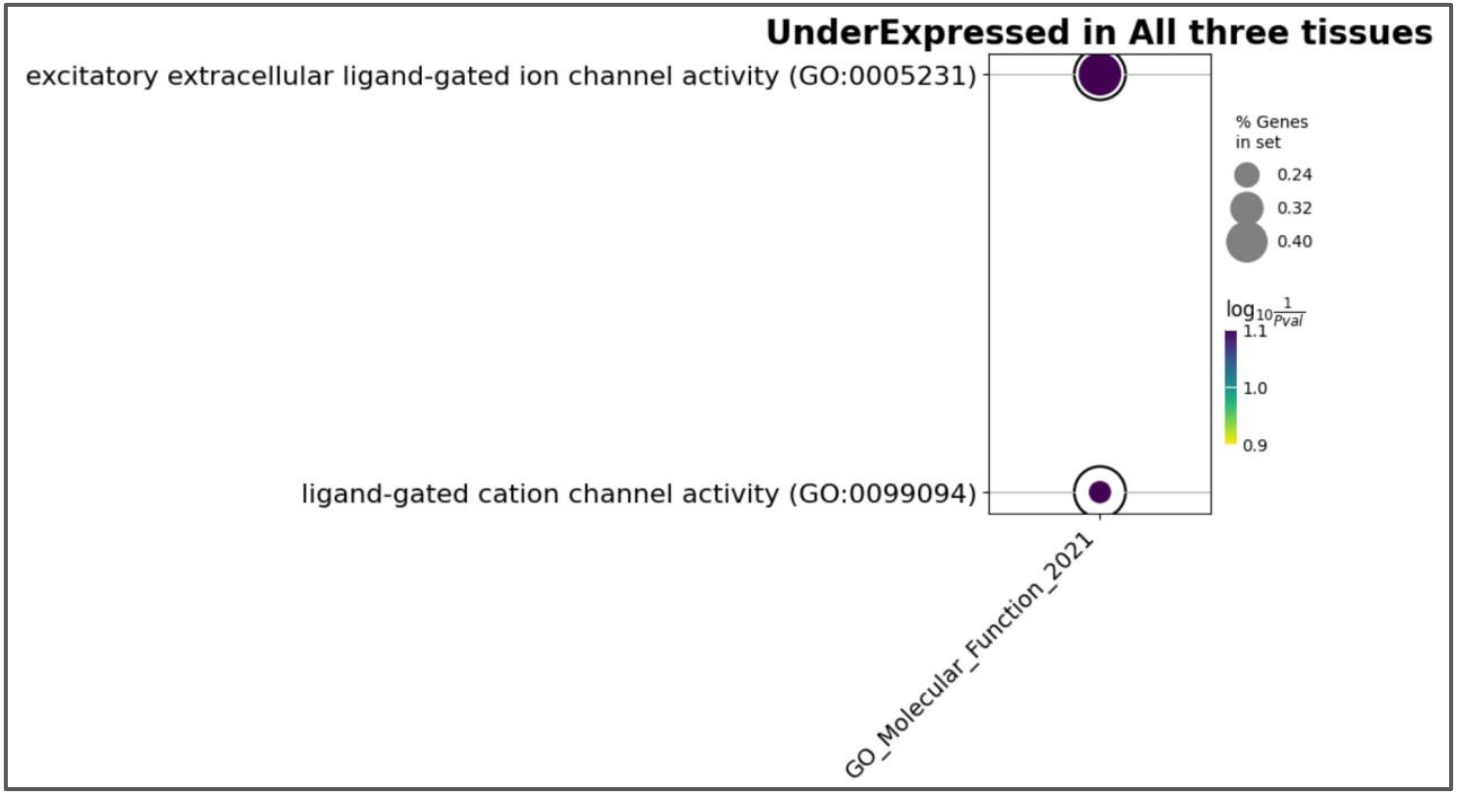


Skoltech
Skolkovo Institute of Science and Technology

Genes differentially overexpressed with age across multiple datasets ***in all three tissues***

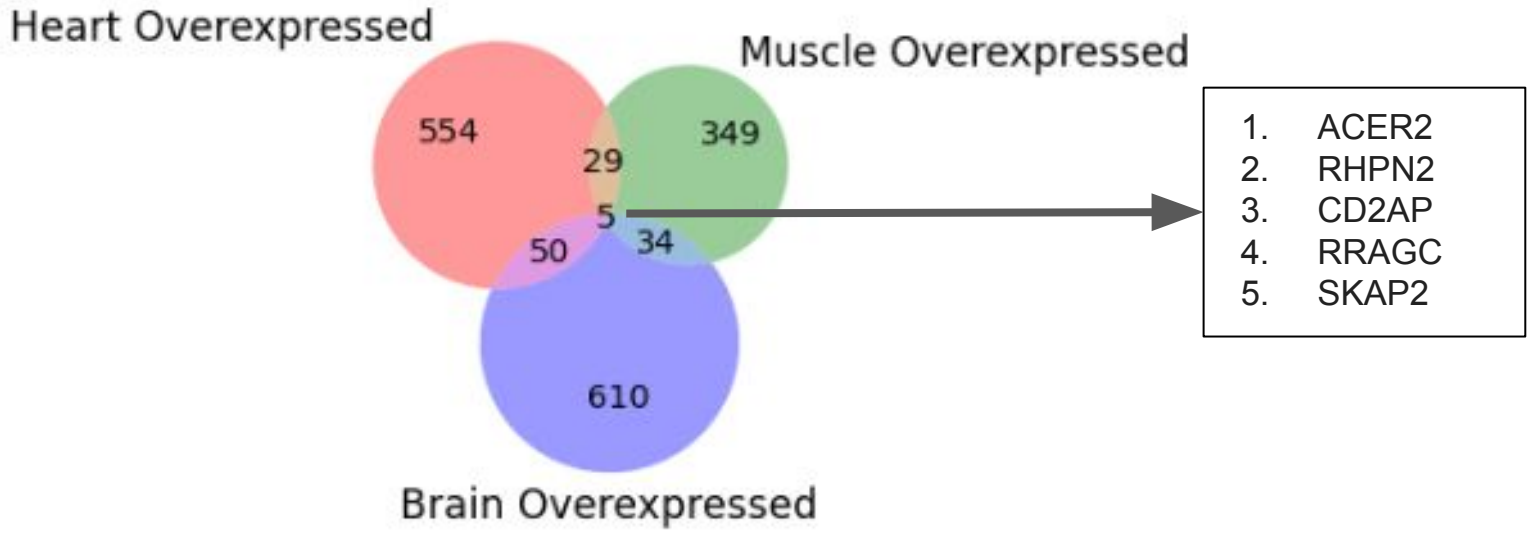


Genes differentially underexpressed with age across multiple datasets ***in all three tissues***



Genes overexpressed in ***brain, heart, and muscle tissues***

- 699 genes overexpressed in brain
- 638 genes overexpressed in heart
- 417 genes overexpressed in muscle



Genes overexpressed in ***brain, heart, and muscle tissues***

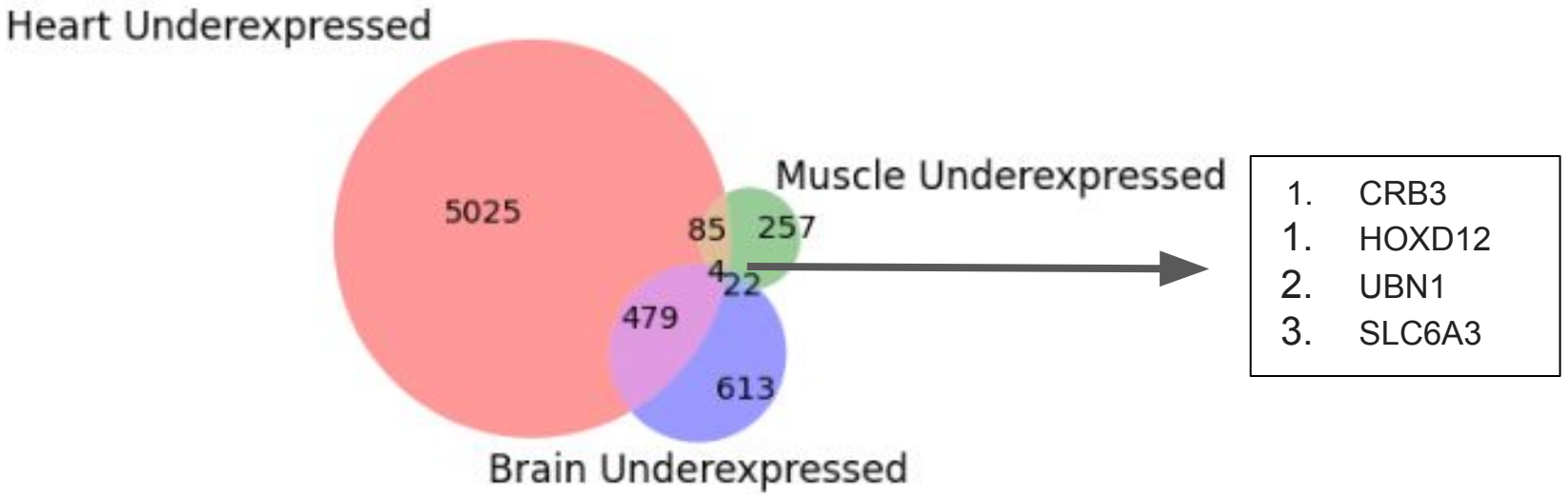
- **ACER2:** may lead to changes in the metabolism of ceramide and other lipid molecules
- **RHPN2**: may affect cell adhesion, migration, and signaling, which could impact various physiological and pathological processes
- **CD2AP**: may disrupt the structure and function of the glomerular filtration barrier in the kidney, leading to proteinuria, inflammation, and progressive kidney damage.
- **RRAGC**: may affect the activity of the mTORC1 signaling pathway, which plays a key role in regulating cell growth, autophagy, and metabolism
- **SKAP2**: Differential expression of SKAP2 may influence T-cell activation and migration, which are essential for immune surveillance and defense against infections and cancer

Genes underexpressed in ***brain, heart, and muscle tissues***

- **CRB3**: encodes a protein that is important for maintaining the structure and function of epithelial cells
- **HOXD12**: member of the homeobox gene family that is involved in regulating embryonic development and differentiation
- **UBN1**: involved in chromatin remodeling, which is important for regulating gene expression.
- **SLC6A3**: encodes a dopamine transporter protein that is involved in the regulation of dopamine signaling in the brain

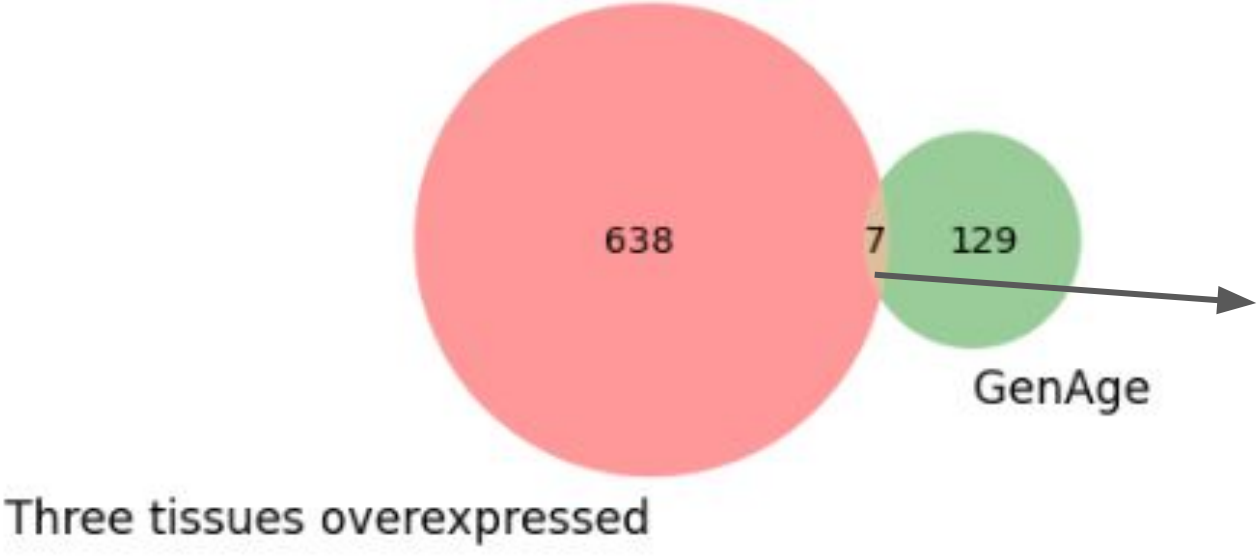
Genes underexpressed in ***brain, heart, and muscle tissues***

- 1351 genes underexpressed in brain
- 5604 genes underexpressed in heart
- 368 genes underexpressed in heart



Skoltech
Skolkovo Institute of Science and Technology

GenAge database intersection with genes overexpressed in ***brain, heart, and muscle tissues***

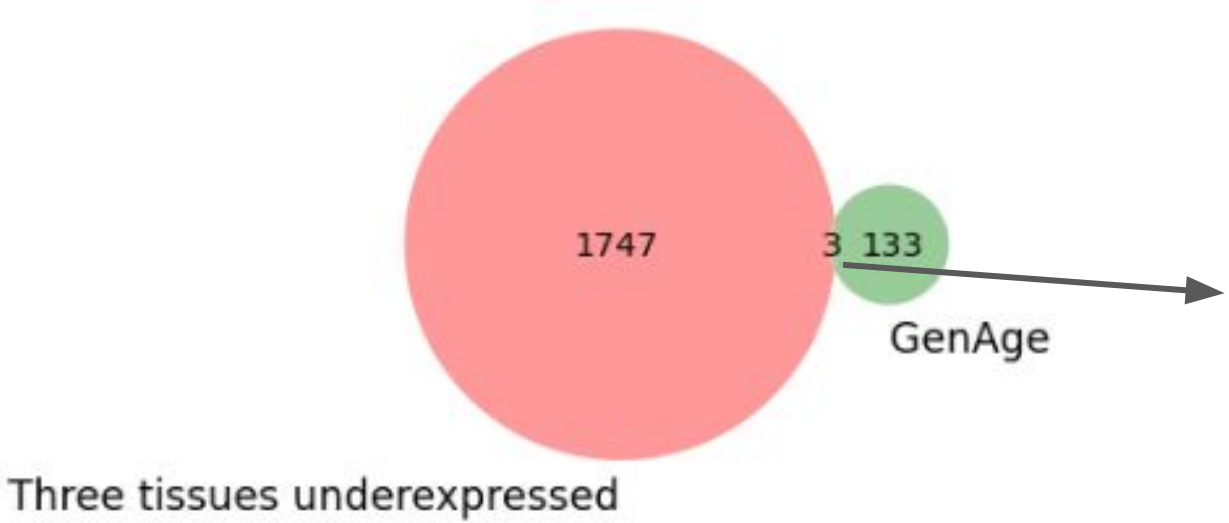


1. EFEMP1
2. TRP53BP1
3. PAWR
4. CDKN1A
5. GRN
6. GPX4
7. HNRNPD

GenAge database intersection with genes overexpressed in ***brain, heart, and muscle tissues***

- **EFEMP1**: encodes a protein that is involved in cell adhesion and signaling
- **TRP53BP1**: This gene is involved in the DNA damage response and plays a role in maintaining genomic stability
- **PAWR**: involved in the regulation of apoptosis (programmed cell death) and has been implicated in various physiological and pathological processes, including cancer and neurodegeneration
- **CDKN1A**: This gene encodes a protein called p21, which is a key regulator of cell cycle progression and DNA damage response
- **GRN**: encodes a protein called progranulin, which is involved in various physiological processes, including inflammation, wound healing, and neuronal development
- **GPX4**: encodes an enzyme called glutathione peroxidase 4, which plays a critical role in protecting cells from oxidative stress
- **HNRNPD**: This gene encodes a protein called heterogeneous nuclear ribonucleoprotein D, which is involved in RNA processing and gene expression regulation

GenAge database intersection with genes underexpressed in ***brain, heart, and muscle tissues***



1. POU1F1
2. FOXM1
3. NUDT1

GenAge database intersection with genes underexpressed in ***brain, heart, and muscle tissues***

- **POU1F1** (Pit-1): a transcription factor that plays a critical role in the development and function of the pituitary gland
- **FOXM1**: a transcription factor that regulates the expression of genes involved in cell cycle progression, DNA replication, and repair
- **NUDT1** (nucleoside diphosphate-linked moiety X motif 1): an enzyme involved in nucleotide metabolism and DNA damage response

Conclusions:

- PyMare approach allowed identifying genes differentially expressed across multiple datasets, perhaps, more specific than binomial test used originally
- Genes upregulated with age across three tissues studied were mainly associated with immunity and inflammation (for muscle tissue - also with cell differentiation, and for heart with cell signalling)
- Genes downregulated with age were not as clearly associated with particular GO category (although for heart metabolic function was enriched)
- Therefore, some of the results we have found are consistent with original findings

Additional slides:

Widely studied genes linked to aging

Gene Alteration	**life span effect**
Pou1f1 knockout	Increased by 40% ↑
GPX4 knockout	Increased by 7% ↑
GRN knockout	Decreased ↓
CDKN1A (p21) knockout	p21 in mice with dysfunctional telomeres prolonged lifespan ↑
Efemp1 knockout	~20% decrease in average lifespan ↓

https://genomics.senescence.info/genes/index.html