DOCTORAL THESIS

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

Bioinformatic analysis and deep learning on large-scale human transcriptomic data: studies on aging, Alzheimer's neurodegeneration and cancer

Óscar González Velasco

Ph.D. SUPERVISORS

Javier De Las Rivas Sanz, Ph.D.

José Manuel Sánchez Santos, Ph.D.

Salamanca, Spain.

2021

Dr. Javier De Las Rivas Sanz, con D.N.I. 15949000H, Investigador Científico del Consejo Superior de Investigaciones Científicas (CSIC), director del grupo de Bioinformática y Genómica Funcional en el Centro de Investigación del Cáncer (CiC-IBMCC), y profesor del Programa de Doctorado y del Máster de Biología y Clínica del Cáncer de dicho Centro y de la Universidad de Salamanca (USAL).

Y el **Dr. José Manuel Sánchez Santos**, con D.N.I. 07870414K, Profesor Titular de Universidad del Departamento de Estadística, Facultad de Ciencias de la Universidad de Salamanca (USAL).

Certifican:

Que han dirigido la Tesis Doctoral titulada "***Bioinformatic analysis and deep learning on large-scale human transcriptomic data: studies on aging, Alzheimer's neurodegeneration and cancer***" realizada por D. **Óscar González Velasco**, dentro del Programa de Doctorado *Biociencias: Biología y Clínica del Cáncer y Medicina Traslacional* del Centro de Investigación del Cáncer (CiC-IBMCC, CSIC/USAL).

Y AUTORIZAN

La presentación de la misma, considerando que reúne las condiciones de originalidad y contenidos requeridos para optar al grado de Doctor por la Universidad de Salamanca.

En Salamanca, a 1 de Abril de 2021

Dr. Javier De Las Rivas Sanz
Director

Dr. José Manuel Sánchez Santos
Codirector

Para la realización de esta Tesis Doctoral, el doctorando Óscar González Velasco obtuvo en concurso público una ***ayuda destinada a financiar la Contratación Predoctoral de Personal Investigador***, cofinanciadas por el **Fondo Social Europeo (FSE)** y convocadas por la **Junta de Castilla y León** (ORDEN EDU/310/2015, de 18 de diciembre de 2017).

Además, la investigación de esta Tesis Doctoral ha sido realizada gracias a los fondos proporcionados a varios Proyectos de Investigación competitivos concedidos al **Grupo del Dr. Javier De Las Rivas** (laboratorio 19) en el **Centro de Investigación del Cáncer (CiC-IBMCC, CSIC/USAL)**. En concreto, se pueden citar los Proyectos Nacionales de la AES del Instituto de Salud Carlos III (ISCiii) PI15/00328 y PI18/00591; y el Proyecto Europeo Horizon 2020 *ArrestAD* ref. 737390 (https://cordis.europa.eu/project/id/737390), iniciado en 2017 que terminará en Diciembre de 2021.

Durante el tiempo de trabajo en esta Tesis Doctoral se logró realizar una **estancia de investigación** durante tres meses (de Enero a Abril de 2021) en Heidelberg (Alemania), gracias a la concesión de una beca **EMBO Short-Term Fellowship** (ref. 8927) para trabajar en el laboratorio dirigido por el **Dr. Julio Saez-Rodriguez** (**Full Professor of Biomedical Informatics and Data Analysis**, https://saezlab.org/) del **Institute for Computational Biomedicine** de la **Medical Faculty,** perteneciente a la **Universidad de Heidelberg** y al **Heidelberg University Hospital**. Como se indica, para la realización de dicha estancia se obtuvo de modo competitivo una beca internacional de la EMBO.

Finalmente, con los méritos anteriores y basado en el trabajo de investigación realizado en los últimos años, esta Tesis Doctoral opta a la **Mención de Doctorado Internacional** otorgado por parte de la **Universidad de Salamanca**, y por ello se presenta escrita en **inglés** en su totalidad, adjuntando también un **resumen en castellano**.

A mi padre.

“Now is no time to think of what you do not have.
Think of what you can do with what there is.”

— Ernest Hemingway, The Old Man and the Sea

SECTION INDEX

INDEX OF FIGURES

INDEX OF TABLES

GENERAL OBJECTIVES

This Doctoral Thesis, entitled "***Bioinformatic analysis and deep learning on large-scale human transcriptomic data: studies on aging, Alzheimer's neurodegeneration and cancer***", is focused on two major highly heterogeneous and complex diseases: the first corresponds to the most prevalent neurodegenerative pathology, **Alzheimer's Disease (AD)**, (including a preliminary study on cognitive decline in the human brain due to aging); and the second one corresponds to cancer, focusing in the analysis of **Cancers of Unknown Primary (CUP)** which represent an heterogeneous group of metastatic cancers that have in common that they are all poorly differentiated and whose primary site is not known at the time of diagnosis, and as a result, patient response to treatment and survival are generally poor.

Both groups of diseases share some similarities: they are polygenic and multi-causal, sensitive to multiple external and internal factors; therefore, they are complex in structure with a high degree of interactions and possible regulators. To tackle these problems, we need to develop and apply **robust bioinformatic algorithms and methods** to address the analysis of large cohorts of human omic data derived from patients with these diseases. We postulate that a robust statistical and computational integrative analysis and meta-analysis of large cohorts (thousands of samples) of different types of omic data (mainly transcriptomics, genomics, proteomics, and interactomics) and the search for relationships between entities (i.e., genes and proteins) will allow the identification of clear and robust biomolecular signatures, associated pathways and biological functions that are deregulated in the patients suffering from these **complex diseases**.

Summary of the main OBJECTIVES of this PhD:

Within this thematic context, throughout this Doctoral Thesis, several bioinformatic and statistical methods have been used and developed for the analysis, integration and discovery of biomolecular signatures associated with these diseases. Therefore, this dissertation has been organized into three different chapters arranged to address the three main OBJECTIVES of our scientific work, as follows:

Objective 1: Determination of the transcriptomic signature associated with aging in human brain and its relationship with neurodegeneration and cognitive decline or impairment.

a. Integration of a large collection of **transcriptomic samples** (obtained with

high-density expression microarrays and with RNA-Seq) of **healthy human brain** biopsies, coming from different studies on the **hippocampus, cortex, and cerebellum** regions, and covering distinct ages from childhood to senescence. Design a bioinformatic pipeline and framework to robustly discover **statistically significant patterns of change in expression due to the age factor** within multiple datasets and infer a common pattern throughout the meta-analysis of the results obtained.

b. Development of a **deep learning neural network** method and implementation of an R package to calculate the **biological age (bioage)** per individual based on the obtained transcriptomic signature associated with aging in the human brain. Comparison between the biological age and the chronological age in different individuals.

This Objective, including an introduction to the topic, a detailed description of our results and a specific discussion is presented in ***CHAPTER I: Transcriptomic landscape of aging on the human brain***.

Objective 2: Integrative profiling of transcriptomic data from brain samples of large cohorts of Alzheimer's disease patients and search for new AD biomarkers found in blood.

a. Integration and analysis a large collection of transcriptomic samples (obtained with high-density expression microarrays and with RNA-Seq) of **human brain biopsies from patients with Alzheimer's disease**, mainly from **hippocampus, cortex, and cerebellum** regions. Design of a bioinformatic pipeline and framework to robustly **unravel significant patterns of change in the gene expression due to Alzheimer's disease**, and identification of a common pattern with the meta-analysis of the obtained results, comparing the AD signature with the one generated in the healthy human brain aging analysis.

b. Analyze the transcriptomic profile of new generated **blood samples from AD patients and controls donors**, obtained with microarrays and RNA-Seq, using both **bulk samples and single-cell isolated samples**, to find key gene signatures and deregulation assigned in a significant way to AD.

This Objective, including an introduction to the topic, a detailed description of our results and a specific discussion is presented in **CHAPTER II: Alzheimer's disease gene**

expression signature and new AD biomarkers found in blood samples. All this work has been carried out as part of a collaborative scientific effort with several European research groups within the **European Horizon 2020 Project *ArrestAD*** (https://cordis.europa.eu/project/id/737390), which is still in progress at the time of writing this thesis (until December 2021) and that it has not published its results. For this reason, all the results corresponding to this Chapter are ***Confidential*** and will be kept under temporal embargo.

Objective 3: Construction of a deep learning tool based on large-scale pan-cancer and normal tissues data for the identification and prediction of the primary origin of tumor samples and for the transcriptomic profile of cancer subtypes.

a. Development and implementation of an advanced **deep learning neural network (DLNN)** predictive model with pan-cancer data, aimed at the **diagnosis of cancers of unknown primary site (CUP)** (i.e., metastatic tumors of unknown origin). Integration of thousands transcriptomic data sets including samples from different types of cancer in combination with transcriptomic data from healthy human tissues to train and build this predictor.

b. Development of a biomolecular profiling tool using gene network analysis to assess the prediction of the neural network and provide further insights into the molecular signature of cancer samples, as well as the altered pathways and transcription factor (TF) activities enhanced or deregulated.

This Objective, including an introduction to the topic, a detailed description of our results and a specific discussion is presented in ***CHAPTER III: Pan-cancer deep-learning prediction and profiling tool for tumor samples based on transcriptomic data***.

CHAPTER I

Transcriptomic landscape of aging on the human brain

1 BRIEF SUMMARY

The biological characteristics of human aging that lead to increased disease susceptibility remain poorly understood. Throughout this chapter we present a transcriptomic analysis of the human brain associated with age, derived from a systematic integrative analysis of four independent cohorts of genome-wide expression data from 2,202 brain samples (cortex, hippocampus and cerebellum) of individuals of different ages (from young infants, 5-10 years old, to elderly people, up to 100 years old) categorized in age stages by decades. The study provides a signature of 1 148 genes detected in cortex, 874 genes in hippocampus and 657 genes in cerebellum, that present significant differential expression changes with age using a robust gamma rank correlation as a metric. The signatures show a significant large overlap of 258 genes between cortex and hippocampus, and 63 common genes between the three brain regions. Using these signatures, we performed a functional enrichment analysis and a cell-type fold-change analysis, which provided biological insight about the complex aging signature.

Finally, we develop and train a fully connected deep-learning neural network, using stochastic gradient descent as the optimization algorithm and the obtained age-related genes as input variables, to create a R package with the main goal of building a biological age predictor based on the aging signature.

2 INTRODUCTION

The use over the last decade of high performance omic technologies, more specifically next generation sequencing like RNA-Seq, applied to the study of human samples is providing a new vision and understanding of our body at biomolecular level. These achievements are especially notable in the area of transcriptomics, thanks to the accuracy and coverage of genome-wide expression platforms, capable of measuring the complete profile of all human genes in different types of samples and conditions. Additionally, as all these omic data comes publicly available in integrative databases, and the available computational resources has grown exponentially, new methodologies and algorithms need to be designed in order to get new insights into these already existing data, especially in complex diseases where sample groups need to be especially large.

2.1 State of the art in human brain omics

Two recent large-scale efforts on the global transcriptomic profiling of human tissues and cell types, provided a comprehensive original view of the genes that are active in different parts of the human body (Ardlie et al., 2015; Fagerberg et al., 2014; Uhlén et al., 2015). These transcriptome studies included some parts of the cerebrum; however, using a relatively small number of samples of this region of the body, since they were not solely focused on the study of the human brain.

Pioneering studies on the transcriptome of the human brain began by exploring its organization through systematic analysis of gene co-expression relationships (Oldham et al., 2008). Previous efforts yielded biological insight about certain parts of the human brain compared with non-human primate brains (Cáceres et al., 2003); but Oldham et al. produced the first integrated framework describing genome-wide expression in the brain, focusing on three regions: cerebral cortex, caudate nucleus and cerebellum (Oldham et al., 2008). Meanwhile, other relevant research efforts through global transcriptome analysis, explored more than 10 different brain regions (including cerebellum, thalamus, striatum, hippocampus and neocortical areas) (Johnson et al., 2009). However, these analyses primarily focused on the development of the cerebral transcriptome during the prenatal period, since all samples were collected from fetal human brains. The first achievement of a genome-wide expression profile of the whole human brain of adults became a reality in 2012, when Michael Hawrylycz and colleagues, at the Allen Institute for Brain Science, published the first anatomically comprehensive atlas of the adult human brain transcriptome (M. J. Hawrylycz et al., 2012). This work provided a high-resolution map of

gene expression of the human brain, using laser microdissection and microarrays, to assess 900 precise subdivisions in brains from two healthy men. Despite the exhaustive coverage over many regions of the complex anatomy of the brain, this work was only done in two brains of adult men. This study, based on transcriptomic profiles, was later expanded by the same research group, to further investigate the structure and function of the human brain by producing reproducible gene expression patterns across 132 structures (M. Hawrylycz et al., 2015). In this way, they identified 32 anatomically diverse gene expression signatures in the brain, that were recognized consistent and reproducible. Again, a major limitation of this work was the use of only six adult human brains.

All the studies described above focused mainly on the analysis of the human brain in adult individuals or targeting a specific period of life (such as prenatal). However, the demonstrated power of global transcriptome analysis provided an excellent framework to investigate the evolution of the human brain transcriptome throughout the different stages of life, from young individuals to elderly people. Therefore, the study of the effect of age and aging on the human brain became a relevant issue, because it is known that age is a cause of cognitive decline in the elderly and a major risk factor for many neurodegenerative diseases. The first in-depth study that addressed this question using global gene expression profiling was published in Nature in 2004 (Lu et al., 2004). These authors showed that the transcriptional profiling of the human cortex, from individuals ranging from 25 to 105 years of age, defined a set of genes that changed after age 40. This altered gene set included reduced expression of genes involved in synaptic plasticity, vesicular transport, mitochondrial function, and base-excision DNA repair. It also included over expression of some genes related to stress response and antioxidant activation. The work by Lu et al. also found that DNA damage by oxidative stress markedly increased in the promoters of genes with reduced expression in the aged cortex. Thus, DNA damage can reduce the expression of selectively vulnerable genes involved in learning, memory and neuronal survival, initiating a program of brain aging (Lu et al., 2004). However, in this work the majority of the analyses and contrasts were performed considering only two age thresholds (≤42 years for young individuals, and ≥73 years for the elderly); which is a limitation to understand the progression of the brain throughout age.

Later studies on the aging of the human brain have increased the number of age stages to better understand its complexity and dynamics. In this regard, Kang and his colleagues published an outstanding work that presented a spatiotemporal mapping of the transcriptome of the human brain over 15 time-periods: from embryonic and fetal, till infancy, childhood and adulthood (Kang et al., 2011). In this study, the most dramatic differences and changes were detected before birth. Other large-scale transcriptional

studies addressed the human brain aging: focusing on sex differences in gene expression and splicing (Trabzuni et al., 2013); or comparing the transcriptional landscape of brain tissue samples and blood samples, from many individuals with different chronological ages (Hernandez et al., 2012; Peters et al., 2015). Finally, other studies conducted transcriptional analyses to discover signatures of genes up- or down-regulated across multiple regions of the brain, comparing changes in normal human aging versus Alzheimer's disease (Berchtold et al., 2013, 2014).

Considering the described transcriptomic studies, in this work we seek to map the landscape of the human brain gene expression changes throughout aging, to discover the most significant transcriptional alterations that occur in the brain genes with increasing chronological age. To achieve this, we collected and integrated as many valuable sets of transcriptome data produced from normal brain samples as possible, generating a large compendium of comparable and complementary data. In this way, our work provided a collection of 2,202 samples from 3 main brain regions: cortex, hippocampus, and cerebellum. In all cases, the samples came from neurologically healthy individuals, that is, individuals who did not have a neurodegenerative disease or brain pathology. We proceed to the construction of a deep transcriptomic profile of the human brain using individuals of different ages: from children (5-10 years old) to elderly people (80-100 years old) including intermediate age stages. Integrative analyses of all these samples produced robust gene expression signatures associated with aging, which revealed quite a similarity between the cortex and the hippocampus and a more different profile for the cerebellum. The signatures included sets of genes that undergo up- or down-regulation with aging. We analyzed in more detail the gene signature associated with cortex (i.e., the cerebral cortex), because it constitutes the largest part of the cerebrum, includes the grey matter and the largest site of neural integration, and plays a key role in most cognitive functions: memory, attention, perception, thought, language, and consciousness. In this way, the gene signature from cortex was used to produce a biological age predictor by applying a deep-learning algorithm. All these analyses finally provided a broad profile of the human brain transcriptome and a robust identification of brain genes directly related to aging.

3 MATERIAL AND METHODS

3.1 Collection and integration of multiple brain samples in a large compendium

A large exploration of multiple databases and resources was done to obtain a collection of datasets of samples of post-mortem human brains from neurologically healthy individuals with different ages. In this way, we managed to prepare a compendium of more than two thousand samples that had in all cases an adequate phenotypic characterization, plus genome-wide expression data. The sample compendium is described in **Table 1**, which includes datasets of cortex, hippocampus and cerebellum (in a total collection of 2,202 samples). At the same time, information about the age of each individual, donor of samples, was collected. All these data were used for the integrative transcriptomic profiling of the human brain throughout different age stages, from children to the elderly, organizing the samples in similar time intervals, as indicated in **Table 1**. The age stages included approximately similar ages by decades: from children, first decade (D0), to older people, last decades (D7, D8, D9). The number of stages is not the same in each dataset and they were configured flexible with respect to the age but trying to include as similar number of samples in each consecutive stage.

Datasets	Authors / Reference	Data Public	GEO GSE_ID	Transcriptomic Platform Used	Tissue Type	Total number of samples	0	1	2	3	4	5	6	7	8	9	Number of AGE stages
Dataset 1	**Kang et al.** ***Nature (2011)***	**2011**	**GSE25219**	Affymetrix Human Exon 1.0 ST Array	***Cortex*** [1]	**429**	**D0i D0ii** (51) (63)	**D1** (88)	**D23** (99)		**D45** (64)		**D6>** (64)		–	–	**6**
"	"	"	**GSE25219**	Affymetrix Human Exon 1.0 ST Array	***Hippocampus***	**35**	**D0i D0ii** (4) (5)	**D1** (8)	**D23** (7)		**D45** (5)		**D6>** (6)		–	–	**6**
"	"	"	**GSE25219**	Affymetrix Human Exon 1.0 ST Array	***Cerebellum***	**34**	**D0i D0ii** (5) (6)	**D1** (7)	**D23** (8)		**D45** (4)		**D6>** (4)		–	–	**6**
Dataset 2	**Hernandez et al.** ***Neurobiol Dis (2012)***	**2012**	**GSE36192**	Illumina HumanHT-12 V3.0 Expression Beadchip Array	***Cortex*** [2]	**453**	**D0** (13)	**D1** (67)	**D2** (44)	**D3** (46)	**D4** (77)	**D5** (56)	**D6** (34)	**D7** (35)	**D8** (52)	**D9** (29)	**10**
"	"	"	**GSE36192**	Illumina HumanHT-12 V3.0 Expression Beadchip Array	***Cerebellum***	**454**	**D0** (13)	**D1** (67)	**D2** (44)	**D3** (47)	**D4** (78)	**D5** (55)	**D6** (34)	**D7** (35)	**D8** (52)	**D9** (29)	**10**
Dataset 3	**Trabzuni et al.** ***Nat Commun (2013)***	**2013**	**GSE46706**	Affymetrix Human Exon 1.0 ST Array	***Cortex*** [3]	**374**	–	**D12** (31)		**D3** (26)	**D4** (59)	**D5** (77)	**D6** (71)	**D7** (51)	**D89** (59)		**7**
"	"	"	**GSE46706**	Affymetrix Human Exon 1.0 ST Array	***Hippocampus***	**121**	–	**D12** (11)		**D3** (9)	**D4** (20)	**D5** (27)	**D6** (23)	**D7** (17)	**D89** (14)		**7**
"	"	"	**GSE46706**	Affymetrix Human Exon 1.0 ST Array	***Cerebellum***	**130**	–	**D12** (11)		**D3** (9)	**D4** (21)	**D5** (25)	**D6** (24)	**D7** (19)	**D89** (21)		**7**
Dataset 4	**Berchtold et al.** ***Neurobiol Aging (2013)***	**2014**	**GSE48350**	Affymetrix Human Genome U133 Plus 2.0 Array	***Cortex*** [4]	**129**	–	–	**D2** (21)	**D3** (13)	**D4** (25)	**D56** (14)		**D7** (19)	**D8** (17)	**D9** (20)	**7**
"	"	"	**GSE48350**	Affymetrix Human Genome U133 Plus 2.0 Array	***Hippocampus***	**43**	–	–	**D2** (6)	**D3** (4)	**D4** (7)	**D56** (5)		**D7** (5)	**D8** (9)	**D9** (7)	**7**
					TOTAL number of samples	**2202**	**AGE stages in consecutive decades** * (Number of samples in each stage)										

Table 1*: Genome-wide expression datasets used in the integrative transcriptomic profiling of human brain samples of different ages, from children to elderly people. The original raw data series were downloaded from GEO (corresponding to GSE IDs: GSE25219; GSE36192; GSE46706; GSE48350).*

3.2 Gene expression normalization, signal calculation and correlation with age

Each of the 4 datasets collected was treated and analyzed individually, as a strategy to provide independent support and confirmation of the results from each data series. All the data management, analysis and integration were done using R (programming language and environment for statistical computing, https://www.r-project.org/). We also used several software packages and algorithms from Bioconductor (https://www.bioconductor.org/).

Normalization and signal expression calculation were performed using *affy* (Gautier et al., 2004) and *RMA* (Irizarry et al., 2003) algorithms, in the case of the expression datasets obtained with *Affymetrix* platforms (that were high-density oligonucleotide microarrays Human Exon 1.0 and Human Genome U133 Plus 2.0). In the case of the expression datasets obtained with *Illumina* platform, the normalization and signal expression calculation were performed using the R package *beadarray* (Dunning et al., 2007). Assignment of the expression signal from the arrays to human gene entities (i.e. ENSEMBL ID genes, Human Build 38), instead to the probes or probesets of the platforms, was done following the strategy described in (Dai et al., 2005; Risueño et al., 2010).

Further analysis of the signal of the different datasets yielded a strong *batch effect* (a widely extended problem in biostatistics where the main source of signal variability comes from technical conditions and noise, rather than from biological factors)(**Figure 1**), for this reason, and because the sequencing platforms used on the different experiments are different, we opted for a meta-analysis approach in which we combine the results of the different gene expression analysis instead of combining the signal in a large unique dataset altogether.

***Figure 1**: tSNE algorithm applied to three of the brain aging datasets. As it can be clearly seen, the primary source of variation is the dataset itself, thus the batcheffect is the factor driving the clustering.*

Correlation of the expression signatures and trajectories throughout different ages was done using two different but complementary methods: **(i)** *limma* (Ritchie et al., 2015) for differential expression within multiple stages followed by linear regression (i.e., each age stage was contrasted with the immediately following stage, and the gene expression fold-change for every pair of consecutive age stages was calculated, as well as the corresponding p-value); **(ii)** *gamma* correlation (Goodman & Kruskal, 1954), obtained by comparing the expression throughout the age stages and calculating, for each correlation, a p-value based on cross-validation with 1000 iterations. The *Goodman-Kruskal gamma* γ correlation is a measure of ordinal (rank) association between two variables with ranked quantities (Boudt et al., 2012). The range of the gamma statistic is $-1 \leq \gamma \leq 1$ being -1 perfect inverse correlation, 1 perfect correlation and 0 no correlation.

$$\gamma = \frac{C - D}{C + D} \; where: \begin{cases} C \; number \; of \; concordant \; pairs \\ D \; number \; of \; discordant \; pairs \end{cases} \qquad \textbf{Eq. 1}$$

Concordant and discordant pairs were calculated using a *Robust Rank Correlation Coefficient* with the R package *rococo* (Ulrich Bodenhofer et al., 2013): as stated in (U. Bodenhofer & Klawonn, 2008) rank correlations are designed for ordinal data and thus,

they are not ideally suited for measuring rank correlation for numerical data that are perturbed by noise. Consequently, (U. Bodenhofer & Klawonn, 2008) introduces a family of robust rank correlation measures. The idea is to replace the classical ordering of real numbers used in Goodman's and Kruskal's gamma by a fuzzy ordering with smooth transitions — thereby ensuring that the correlation measure is continuous with respect to the data.

Before these analyses, the age of the individuals was divided in age-periods or age-stages by approximate decades (**Table 1**), so that the different consecutive stages included a similar and balanced number of samples. In this way, each of the analyzed datasets included 6 or more stages, from very young ages (corresponding to children: D0 = 1-9 years old, D1 = 10-19 years old) to old ages (i.e. elderly people: D6 = 60-69 years old, D7 = 70-79 years old, D8 = 80-89 years old, etc.) (**Table 1**). Due to the fact that dataset 1 (Kang et al., 2011) included many samples of young people (children), we divided these samples into two (following the strategy used by Kang et al. in their work): stage D0i = corresponding to ages between 1 and 6 years; and stage D0ii = between 6 and 12 years. This was also done to maintain the balance in the number of people included in the age stages of this dataset.

It is important to underline that the categorization of the individual signals in age-stages, carried out in this study, was essential to achieve an accurate capture of gene expression signatures, and their evolution and sustained changes during the life of neurologically healthy individuals. Therefore, the approach provided a robust and consistent gene signature of the brain, which was validated in several independent cohorts and that evolved constantly through the ages, being a signature specifically linked to aging.

In order to delve into the results obtained by the *gamma* statistic, we applied another algorithm to verify whether the main trajectories or trends found with *gamma* could be found with a different method based on co-expression and clustering of genes. In this way, we applied to the genes, included in the signature derived from the *gamma*, a *fuzzy c-means* algorithm (Dembéle & Kastner, 2003), using functions from *mFuzz* (Futschik & Carlisle, 2005) and *e1070* packages. This method discovers groups of genes (i.e., clusters) that follow similar co-expression patterns across the ordinal covariate (the age-stages). Therefore, we used this analysis to corroborate, within each of the 4 datasets, the finding of groups of genes that show a significant trend of upward or downward expression with increasing age. The fuzzy analysis also showed that these two main trends of increase and decrease in expression over time were reproducible and consistent.

3.3 Concurrent functional enrichment analysis of the aging gene signature

After the identification of the aging gene signatures in cortex, hippocampus and cerebellum, we performed a functional enrichment analysis of the largest signature (which was the one derived from cortex) to discover the main biological processes and pathways over-represented or re-pressed (i.e. significantly enriched in this set of genes). This analysis was done using several enrichment tools: *DAVID* (https://david.ncifcrf.gov/) (D. W. Huang et al., 2009) and *GeneTerm Linker* (Aibar et al., 2015; Fontanillo et al., 2011). In particular, the second software tool, *GeneTerm Linker* (http://gtlinker.cnb.csic.es/), performs concurrent functional enrichment analysis in several annotation databases at the same time: the 3 sections of Gene Ontology (GO-BP, GO-MF, GO-CC), KEGG pathways and InterPro motifs and domains. In this way, the tool filters out redundant annotations and provides enriched output data, identifying sets of genes and terms included in metagroups of coherent biological significance (Fontanillo et al., 2011). The resulting metagroups were evaluated measuring the enrichment in specific functional annotations (obtained with enrichment p-values derived from hypergeometric tests) and the coherence of the grouping (obtained with two coefficients: a *Cosine Similarity Coefficient* calculated for the gene-term sets of each metagroup; and a *Silhouette Width Coefficient* that measures how the genes are clustered in the metagroups, taking into account intra-group compactness and inter-groups proximity) (Fontanillo et al., 2011). In our analysis, *GeneTerm Linker* provided the best results since it generates metagroups of genes associated with highly consistent groups of biological functions.

3.4 Finding gene coregulation and transcription factors (TFs) associated to aging

Once the aging gene signature was defined (identified as a set of genes that have a common and consistent expression associated to increasing age), we applied several methods to find whether the promoters and genome regulatory regions of these genes show any significant enrichment in certain transcription factor binding sites (TFBS) and therefore in transcription factors (TFs) known to bind to these TFBS. This analysis was done using iRegulon tool (Janky et al. 2014) [29]. This computational tool allows the identification of transcription factors that are enriched as candidate regulators in a human gene signature (considered as a set of co-expressed genes or a set of genes active and

transcribed in the same conditions). The gene signature is analysed by iRegulon to test if it can be identified as direct target gene set of some specific TFs. iRegulon relies on the analysis of the regulatory sequences around each gene in the gene set to detect enriched TF motifs or ChIP-seq peaks, using databases of nearly 10,000 TF motifs and 1000 ChIP-seq tracks. It associates enriched motifs and tracks with candidate TFs using a Normalized Enrichment Score (NES) and determines the optimal subset of direct target genes (Janky et al. 2014) [29]. The NES score is a z-score indicative of the significance. The default NES cutoff in iRegulon is 3.0, corresponding to an FDR between 3% and 9%.

3.5 Single cell-derived signature to identify cell type age-related changes in brain

The human brain is a tissue of great complexity. In terms of the cell types, it comprises neurons, astrocytes, oligodendrocytes, microglia, etc (which have very different biological functions). We explored our collection of human brain samples, categorized in series of increasing age, to map the gene expression changes along the age to specific cell types present in the brain. To achieve this, we followed the approach provided by (Darmanis et al., 2015), who using single cell RNA sequencing identified a set of molecular markers that allow classifying different cell types of the human brain. In fact, this paper provides a first cellular atlas of the human brain based in cell specific gene-signatures. We used the signatures that they assigned to neurons, astrocytes, and microglia to measure the evolution of these three cell types by decade stages across young ages (from 1 to 39 years old) and across old ages (from 50 to 100 years old).

3.6 Deep-learning Neural Network method applied to biological age calculation

DLNN is a new-generation machine learning methodology, which mimics the way animal brains operates, that uses multiples layers of connected units or nodes called neurons (a “neuron” is a mathematical function that collects and classifies information according to a specific architecture), each connection acts like a synapse and can send a signal to other neuron, with the final goal of recognizing underlying relationships in a set of data through a process of learning and fitting. Each node is a perceptron and is similar to a multiple linear regression. The perceptron feeds the signal produced by a multiple linear regression into an activation function that may be nonlinear. One of the key points of these

architectures is that Neural networks can adapt to changing input; so, the network generates the best possible result without needing to redesign the output criteria. They became more successful in recent years largely due to the availability of inexpensive, parallel hardware (GPUs, computer clusters) and massive amounts of data.

3.6.1 Defining the model

A deep-learning neural network (DLNN) algorithm was built using R statistical programming language and is available as an R package. We encoded each neuron as a linear function in which each synapse has an input value x_i and an associated weight θ_i plus a unique bias value b corresponding to each neuron as can be seen in **Figure 2**:

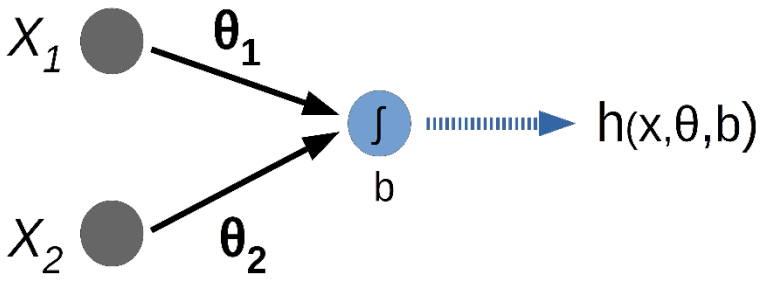


***Figure 2**: Example of a model of a two-layer neural network with two inputs.*

Where: $h(x,\theta,b) = \theta_1 x_1 + \theta_2 x_2 + b$ which can also be written in vector form as: $h(x,\theta,b) = \theta^T x + b$

In order to obtain a non-linear decision boundary and capture more complex patterns on the data, we included an **a**ctivation function that is directly applied after the linear-weighted operation on each neuron. We selected the Parametric Rectified Linear Unit (PReLU) (He et al., 2015) function as our activation for the hidden layers, because is more flexible than ReLU function dealing with negative numbers (PReLU assign a small positive slope for x < 0) and less prone to the vanishing gradient, a common problem of neural networks in which the accumulation of small gradients results in a model that is incapable of learning meaningful insights.

Thus, we construct the new function encoded by each neuron as:

$$f(h(x,\theta,b)) = f(\theta^T x + b), \ \ where\ f(x) = \begin{cases} x, & x \geq 0 \\ ax, & x < 0 \end{cases} \qquad \textit{Eq. 2}$$

$$PReLU = f(x)$$

Note that the **'a'** parameter corresponding to the negative values can be either a

fixed value (this case is known as leaky ReLU if: $0 < a < 1$) or a new parameter that is learned along with the other neural-network parameters (He et al., 2015), we opted for the second as it gives additional flexibility despites being more costly to compute.

When we chain several numbers of layers together each having i neurons (hence *deep learning* when linking a large number of layers) we have a *function of functions* that we can describe as $h(x)$:

$$h^{(1)}(x) = x \quad as\ input\ layer \qquad \textbf{Eq. 3}$$

$$h^{(2)}(x) = f\left(\theta_1^1 x_1 + \theta_2^1 x_2 + \cdots + \theta_i^1 x_i + b\right) = f\left(\left(\theta^{(1)}\right)^T h^{(1)} + b^{(1)}\right)$$

$$h^{(3)}(x) = f\left(\left(\theta^{(2)}\right)^T h^{(2)} + b^{(2)}\right)$$

...

$$h(x) = h^{(L)} = f\left(\left(\theta^{(L-1)}\right)^T h^{(L-1)} + b^{(L-1)}\right)\ where\ L = number\ of\ layers$$

And lastly, we need to define a cost function $J(\theta, b)$ that will calculate the error of the whole model, we will use this error to learn from past observations and fit our parameters. For this purpose, we have chosen the Squared Error function, because as our model will act as a regressor (we want to predict a continuous variable $\hat{y}_i$, e.g.: real numbers) we need a measure of error adequate for regression models, also because of simplicity of computation and the fact that negative and positive errors are weighted equal.

$$J(\theta, b) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2\ where\ \hat{y}_i = h(x) \qquad \textbf{Eq. 4}$$

3.6.2 Stochastic Gradient Descent and Backpropagation algorithms

To minimize the objective function $J(\theta, b)$ described above (and thus, minimize the error) we will use *Stochastic Gradient Descent (SGD)*: an iterative optimization algorithm for finding a *local minimum of a differentiable function*. To find a local minimum of a function using gradient descent, we take steps proportional to the negative of the gradient (the gradient vector can be interpreted as *the direction and rate of fastest increase*, thus the negative gradient vector will point to the direction and rate of *fastest decrease*). Accordingly, we will update the parameters of the artificial neural network as follows:

$$\theta_i = \theta_i - \alpha\Delta\theta_i \; ; for\ any\ given\ weight\ \theta_i \qquad \textbf{Eq. 5}$$

$$b = b - \alpha\Delta b \; ; for\ any\ given\ bias\ b$$

Where α is a *small non-negative scalar* known as the *learning rate (LR)*, this critical parameter is used to fine-tune the gradient optimization: a large α will give aggressive updates whereas a small α will give conservative updates. Ideally, we will want a larger α at the beginning of the learning stage and a small α at the end as we approach a minimum of the objective function. We will adjust automatically the learning rate using the Adagrad algorithm (Duchi et al., 2010), as the negative of the gradient gives the direction of fastest decrease, we can also incorporate knowledge of the geometry of these changes and select an appropriated α learning rate depending on the rate of change.

The last problem in our model is how to find the optimal parameters $\Delta\theta_i$ and Δb. Here is where the *Backpropagation algorithm* comes at handy: the backpropagation algorithm works by computing the *gradient of the loss function with respect to each weight* by making use of the chain rule: computing the gradient one layer at a time (as our objective function $J(\theta, b)$ is composed of function of functions, we use the *chain rule* to compute the partial-derivatives in order to obtain the gradient), iterating backward from the last layer to avoid redundant calculations of intermediate terms in the chain rule. The *chain rule* states that if h and f are differentiable functions, then the chain rule expresses the derivative of their composite $h \circ f$ — the function which maps x to $h\big(f(x)\big)$ — in terms of the derivatives of h and f and the product of functions as follows:

$$(h \circ f)' = (h' \circ f) \cdot f' = h'\big(f(x)\big)f'(x) \qquad \textbf{Eq. 6}$$

If we apply this equation into **Eq. 3**, starting by the output layer in order to calculate the gradient from the objective function (**Eq. 4**) (the output layer in our model is a special case since it does not have an activation function associated) and knowing that:

$$\frac{\partial J(\theta, b)}{\partial x} = 2(\hat{y} - y)\,, where\ \hat{y} = \ h(x) \qquad \textbf{Eq. 7}$$

we obtain that the *delta* δ for the output layer L is:

$$\delta_1^{(L)} = 2\big(h^{(L)} - y\big) \bigodot f'\Big(\big(\theta^{(L-1)}\big)^T h^{(L-1)} + b^{(L-1)}\Big) \qquad \textbf{Eq. 8}$$

Note that part of this equation $\Big(\big(\theta^{(L-1)}\big)^T h_j^{(L-1)} + b_1^{(L-1)}\Big)$ has been computed

already on the feedforward pass (**Eq. 3**), thus by temporarily storing these results we can compute part of the gradient and consequently saving a huge amount of computational time. The δ for the intermediate hidden layers is defined as follow:

$$\delta^{(l)} = \left(\left(\theta^{(l)}\right)^T \delta^{(l+1)}\right) \odot f'\left(\left(\theta^{(l-1)}\right)^T h^{(l-1)} + b^{(l-1)}\right) \qquad \textit{Eq. 9}$$

$$where\ l = (L-1, \dots, 1)$$

Note that $\delta^{(l+1)}$ on **Eq. 9** has been computed on the immediately preceding step (e.g.: the equation term $\delta^{(l+1)}$ corresponds to $\delta_1^{(L)}$ for delta $\delta^{(L-1)}$ on layer L-1), and as already exposed above, the term: $\left(\left(\theta^{(L-1)}\right)^T h_j^{(L-1)} + b_1^{(L-1)}\right)$ has been already computed on the feedforward pass (**Eq. 3**); as *PReLU* function (**Eq. 2**) is applied on all the intermediate hidden layers we just need to apply the derivative of the *PReLU* activation function to the stored values.

$$\Delta\theta^{(l)} = \delta^{(l+1)}\left(h^{(l)}\right)^T \qquad \textit{Eq. 10}$$

$$\Delta b^{(l)} = \delta^{(l+1)}$$

3.6.3 Implementation in R and computational optimization

One of the most challenging problems of AI is the computational cost of running these mathematical models: it needs huge amounts of input data and the complexity of the model grows exponentially with each added layer. With the aim to mitigate this, the R package code built for this thesis underwent a severe optimization. All the model described throughout this section is comprised of matrix and vector operations; R is a scripting programing language known to perform slower than other data science languages when dealing with intensive computing tasks (python and Matlab). But for some specific operations, R make use of BLAS libraries: the BLAS (*Basic Linear Algebra Subprograms*) are routines that provide standard building blocks for performing basic vector and matrix operations, usually coded in C++ and highly optimized.

In recent years, deep learning models have been even furthered improved (specially in image detection and classification problems) thanks to the massive use of *Graphic Processing Units (GPUs)* as the computational instances, GPUs are specially designed for these vector-matrix operations: they are comprised of several hundreds of small computing units with a specialized architecture, although processing at smaller frequencies than the CPU (usually in the range of 100-400 MHz). Using the functions that make use of the

underlaying BLAS library, we can substitute the original BLAS with a custom CUDA BLAS library specially designed for NVIDIA GPU's, in which the routines that provide the vector and matrix operations are executed directly on the Graphics Processing Unit. After a benchmark of the computational resources available in our Bioinformatics and Functional Genomic Laboratory at CICancer, we decided to make use of GPU parallelization.

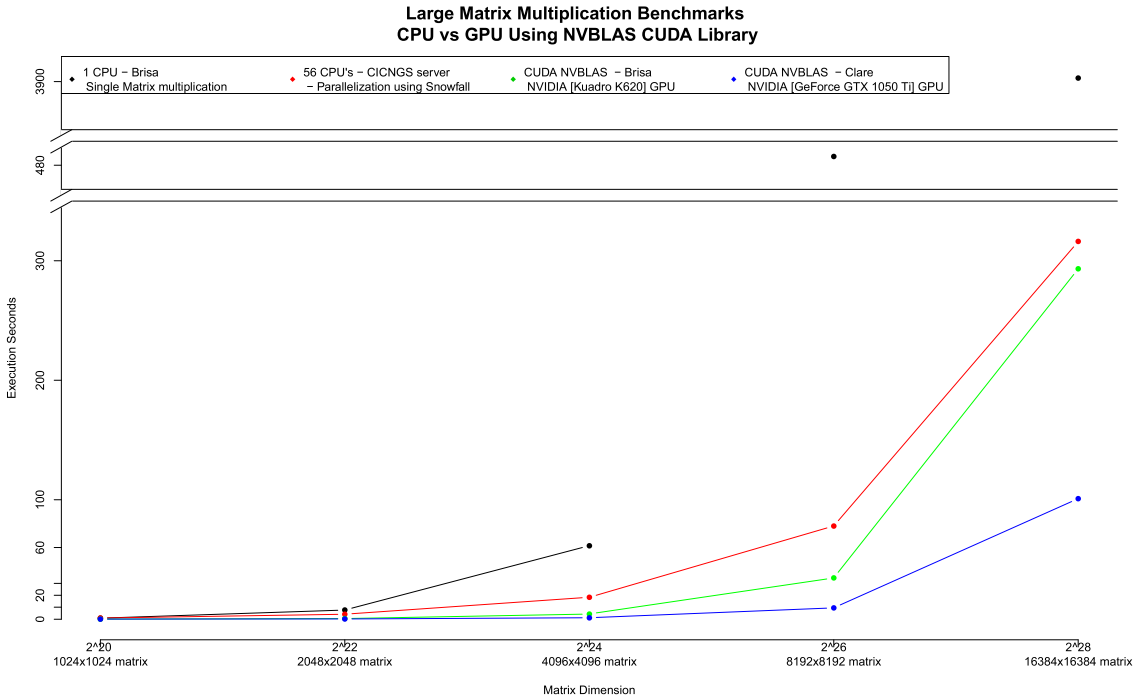


***Figure 3**: Benchmark of a matrix multiplication operation on different hardware architectures available at the home laboratory. Green and blue lines correspond with the two models of GPU. The red line corresponds with the same matrix operation using 56 CPU cores, and black line using just 1 CPU.*

3.6.4 Biological age prediction from transcriptomic data

As stated at the beginning, we applied the aforementioned model to calculate the biological age of a person, based on the data and information provided by the aging gene signature of cortex that we discovered and described in this work.

The training of the Deep Neural Network model took nearly 30 hours. The architecture of the fully connected deep neural network included 19 layers (**Figure** 4).

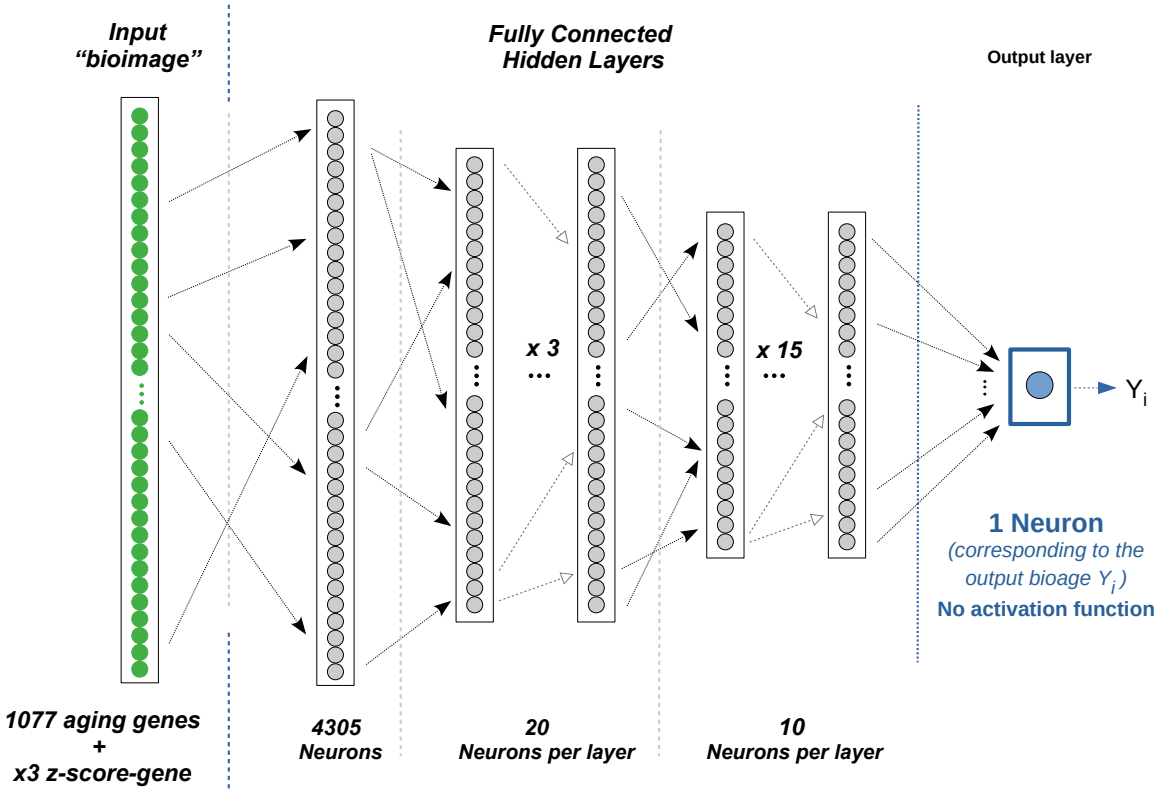


***Figure** 4: Deep Learning Neural Network architecture used in the bio-age predictor model. It is composed of 19 hidden layers and acts as a regression model with a positive continuous real number as output (the bioage).*

The input data used was the gene expression matrix corresponding to 1,077 protein coding genes (i.e. the genes included in the aging gene signature derived from cortex) plus the corresponding chronological age of the person for each sample; using 652 samples (70% of total) for the training of the neural network predictor, and 280 samples (30% of total) for the validation. The total set of samples (652 + 280 = 932), used for the construction of the DLNN, corresponded to the combination all the cortex samples from datasets 1, 3 and 4 (set 1 with 429 samples, set 3 with 374 samples, and set 4 with 129 samples) (**Table 1**). We selected these three datasets because they were produced with similar microarray expression platforms (from *Affymetrix*), and, more important, because they were the ones with the highest coverage over the human transcriptome, mapping to a common set of about twenty thousand human genes (including the 1,077 protein coding genes of the signature).

Before running the neural network method, the input expression data matrix was

scaled, transforming the expression of each sample to a *robust z-score*, using the median and the median absolute deviation (MAD). Besides, three other data matrices were calculated using a different z-score standardization, in which the median was substituted by the expression value in each sample with the three genes that showed the best *gamma* correlation over age (i.e., generating other three gene-centered data matrices, one per gene).

$$robust\ (z-score) = \frac{x - median(x)}{MAD(x)} \qquad \textbf{Eq. 11}$$

$$gene-centered\ (z-score) = \frac{x - Z_k}{MAD(x)} \quad k = 1,2,3$$

where Z_k is the signal of one of the three top aging genes

These four standardized matrices were feed into the neural network as multi-dimensional data derived from the *gamma* rank correlation study. The output of the deep-learning algorithm for any query sample was the prediction of its age as biological age (that is, the predicted age of a person derived from the transcriptomic signature of the brain). Therefore, it is also important to remark that our main goal using the machine learning method was to build a robust regression model predictor of biological age.

Once we constructed the DLNN predictor, we performed an independent analysis to investigate the genes whose expression levels showed the best correlation with the predicted biological age. To do this, we tested how the expression profile of each single gene along the three datasets (1, 3, and 4) correlated with the biological age.

4 RESULTS AND DISCUSSION

4.1 Overview of transcriptomic samples from three regions of the brain using t-SNE

Before entering the study of changes in gene expression with age, we performed an analysis of the whole transcriptomic signal in the samples collected in different datasets. With each of the 4 datasets collected on the compendium (**Table 1**) which included samples from the three regions studied in the brain, we applied the machine learning algorithm t-SNE (t-distributed Stochastic Neighbor Embedding) for the visualization of each dataset (**Figure 5**). The t-SNE is a non-linear dimensionality reduction technique for embedding high-dimensional data in a low-dimensional space (Van Der Maaten & Hinton, 2008). The visualization of the two main dimensions obtained for dataset 1 and dataset 3 shows a clear grouping of the samples according to the 3 regions of the brain, indicating a major separation of the cerebellum versus the cortex and hippocampus. The tSNE over Datasets 2 and 4 also show a clear separation between the two brain regions that the respective datasets include. Moreover, the analysis also suggests that the transcriptome of these two regions, cortex and hippocampus, is considerably similar as shown by the fact that the values of distance between the data points of t-SNE are closer for these regions or even mixed. A similar gene expression profiling and cellular composition of cerebral cortex and hippocampus was already found and reported in the first anatomically comprehensive atlas of the human brain transcriptome (M. J. Hawrylycz et al., 2012).

In the results shown on **Figure 5**, the segregation of the age groups in these analyses is not so clear, probably because the whole gene expression data of the samples reveal the variance from many possible factors. Therefore, to find the specific gene factors linked to the age stages, the data should be investigated using more discriminating analytical methods, since the t-SNE only reflects the overall main variance of the transcriptomic data.

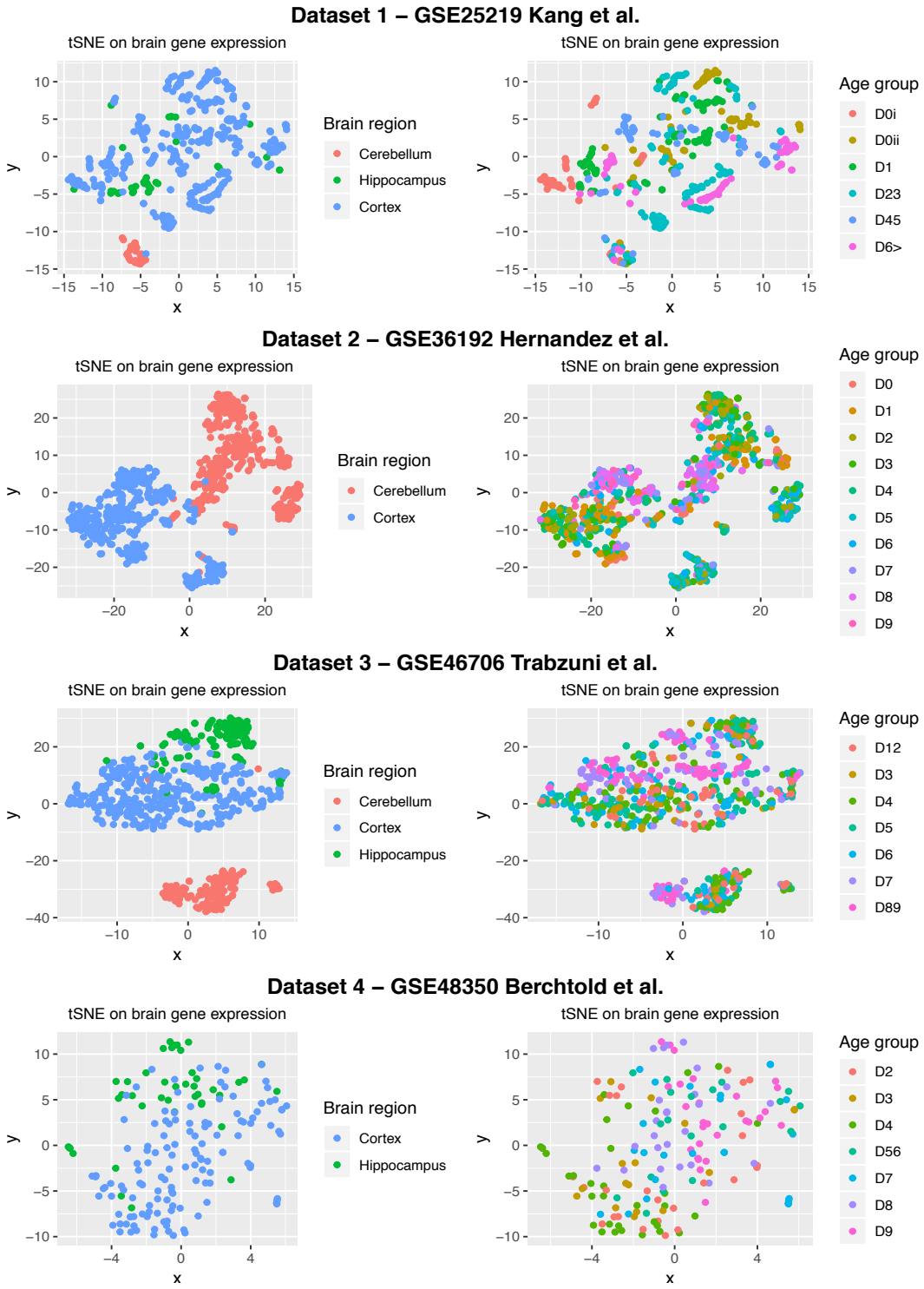


***Figure 5**: tSNE algorithm applied over each of the four brain aging datasets; left column is colored by the brain region where the sample come from, right column is colored by the age group of the individual. Overall results show that spatial difference is stronger than aging difference at transcriptome level (with Cerebellum showing clearly identifiable), however right column figures shows a certain grouping by the age of the individual, specially at older ages.*

4.2 Integrative analysis measuring expression trend of brain genes over age stages

To build a transcriptomic landscape and gene regulatory profile of human brain related to aging, we managed to integrate a large collection of transcriptomic datasets from human post-mortem samples from brains of healthy individuals with different ages, normalized individually per dataset. The compendium includes more than two thousand samples with phenotypic characterization and genome-wide expression data (**Table 1**). The largest number of samples correspond to brain cortex, with 429 samples in dataset 1 (Kang et al., 2011), 453 samples in dataset 2 (Hernandez et al., 2012), 374 samples in dataset 3 (Trabzuni et al., 2013), and 129 samples in dataset 4 (Berchtold et al., 2013). Several algorithms (described in Materials and Methods) were used to identify the expression profile of the genes that follow a trajectory correlated with the increasing age, from young individuals to old people (i.e., with aging). These analyses found genes that present a significant and consistent up-regulation or down-regulation with age stages in the 4 independent datasets. The analyses were performed individually, selecting the samples from each dataset and each region of the brain, and then combining the results of all the datasets for each one of the regions: cortex, hippocampus or cerebellum. In this way, the analyses provided a signature of 1148 genes for cortex, 874 genes for hippocampus and 657 genes for cerebellum. All these genes showed significant gamma rank correlation with age.

Regarding the methodology, it is important to underline that the 4 datasets were treated and analyzed as independent. This provided robust analysis since we reported the agreement of the results in different brain transcriptomic studies. Working in this way, a cross-dataset normalization of the transcriptomic data was not needed. We did not put together or mix the RNA signal from different platforms, labs or studies, and so there was no batch effect that could dramatically affect the outcome of the meta-analysis. Naturally, each independent dataset was internally normalized, so that all samples in each study were comparable.

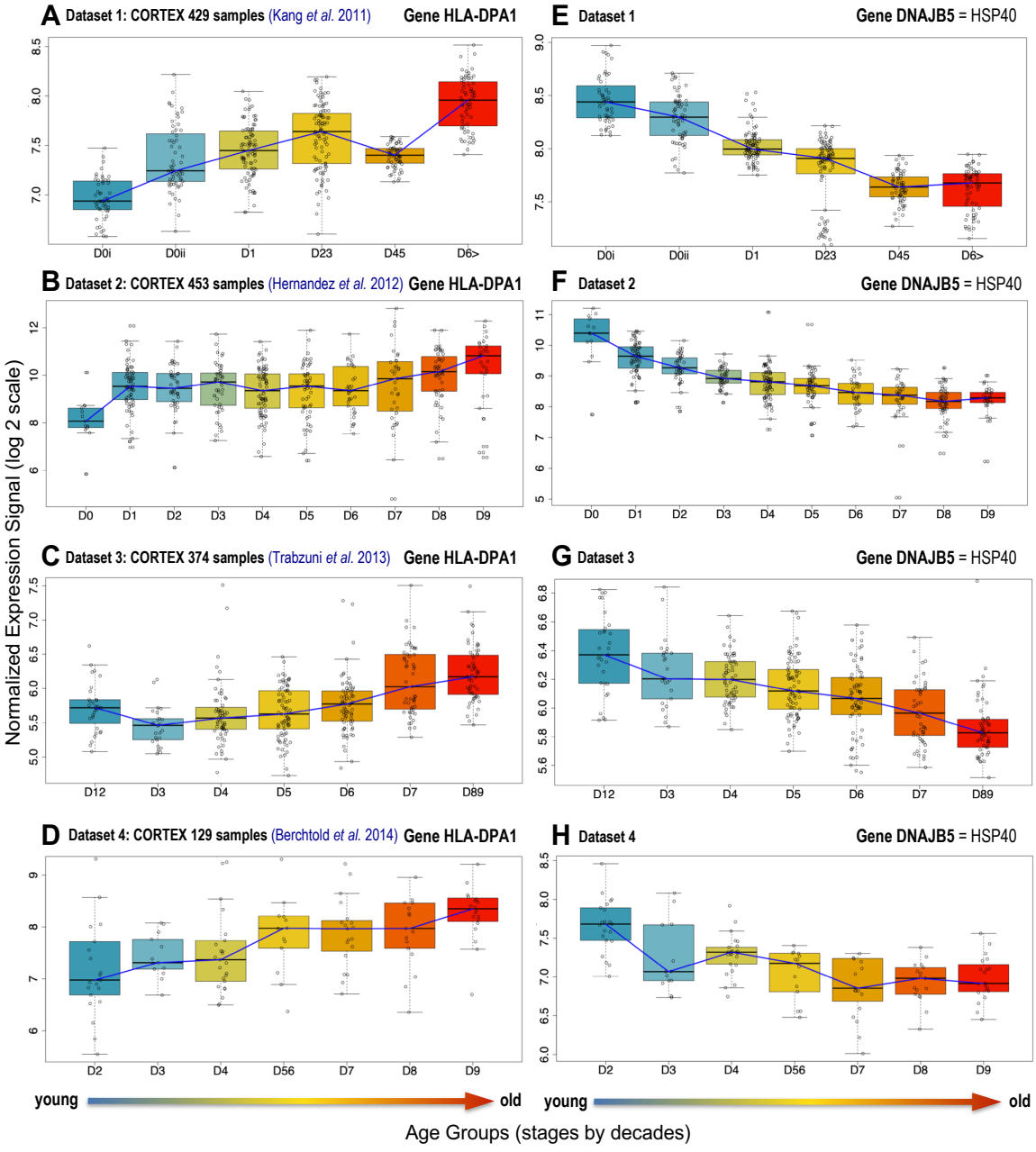


***Figure 6**: Expression profiles throughout age stages in four datasets of human brain cortex: example of 2 genes (HLA-DPA1 and DNAJB5) showing significant up-regulation with aging.*

Another critical point in our methodology is the fact that in the entire study we established age stages by transforming the continuous variable age in a discrete variable. Datasets usually have scarcity of samples corresponding to many age time-points (i.e., in some series we did not have people younger than 20 years old, or older than 70 years old). Therefore, the whole approach of our analysis established age periods, becoming robust and precise at estimating the trends in the expression changes (as it is reflected in **¡Error! No se encuentra el origen de la referencia.**, **Figure 7** and **Figure 8**). Furthermore, the age decades were set up as progressive age ranges or age stages (meaning that they did not need to be exact decades to provide a trend in expression). As a whole, our methodological approach has a clear advantage when using rank statistics (i.e., non-parametric statistics that use ordinal variables), since it does not rely on numbers, but on rankings. In fact, the *gamma* correlation is a non-parametric coefficient that measures the relationship between ordinal variables (**Eq. 1**).

4.3 Transcriptomic profiles of brain cortex and hippocampus with age

Following the methodologies described above, we performed a complete analysis of the expression trajectories throughout the age stages of each gene, in each of the 4 datasets in the 3 brain regions studied. With this approach, we obtained profiles (as the ones presented in **¡Error! No se encuentra el origen de la referencia.**), that are independent for each dataset and each brain region. The most significant genes followed two major profiles: one that steadily increases expression over the progressive stages of life, and another in which the expression decreases across age stages. One of the up-regulated genes with most significant correlation in the cortex transcriptome data is HLA-DPA1 (**¡Error! No se encuentra el origen de la referencia.**A,B,C,D), which corresponds to a protein member of the major histocompatibility complex (MHC class II, DP alpha 1) and plays a central role in the immune system by presenting antigen peptides derived from extracellular proteins. The observation that HLA proteins increase in brain with age had been reported by (Darmanis et al., 2015) who observed the expression of major histocompatibility complex genes in a subset of adult human neurons, but not in fetal neurons. It is expected that molecules expressed in antigen presenting cells increase levels in elderly people, since their brain immunological system has been exposed to many more antigenic elements. Indeed, our transcriptomic profiling also revealed that several other HLA genes were up-regulated with age: HLA-DMB, HLA-DPB1, HLA-DPB2 and HLA-DRA. With respect to genes down-regulated with age, we present in **Figure 6** (E, F, G, H) the

profile in the 4 datasets of the gene that showed the most significant down-regulation: the heat shock protein DNAJB5 (member of the HSP40 family). Heat shock proteins (HSPs)

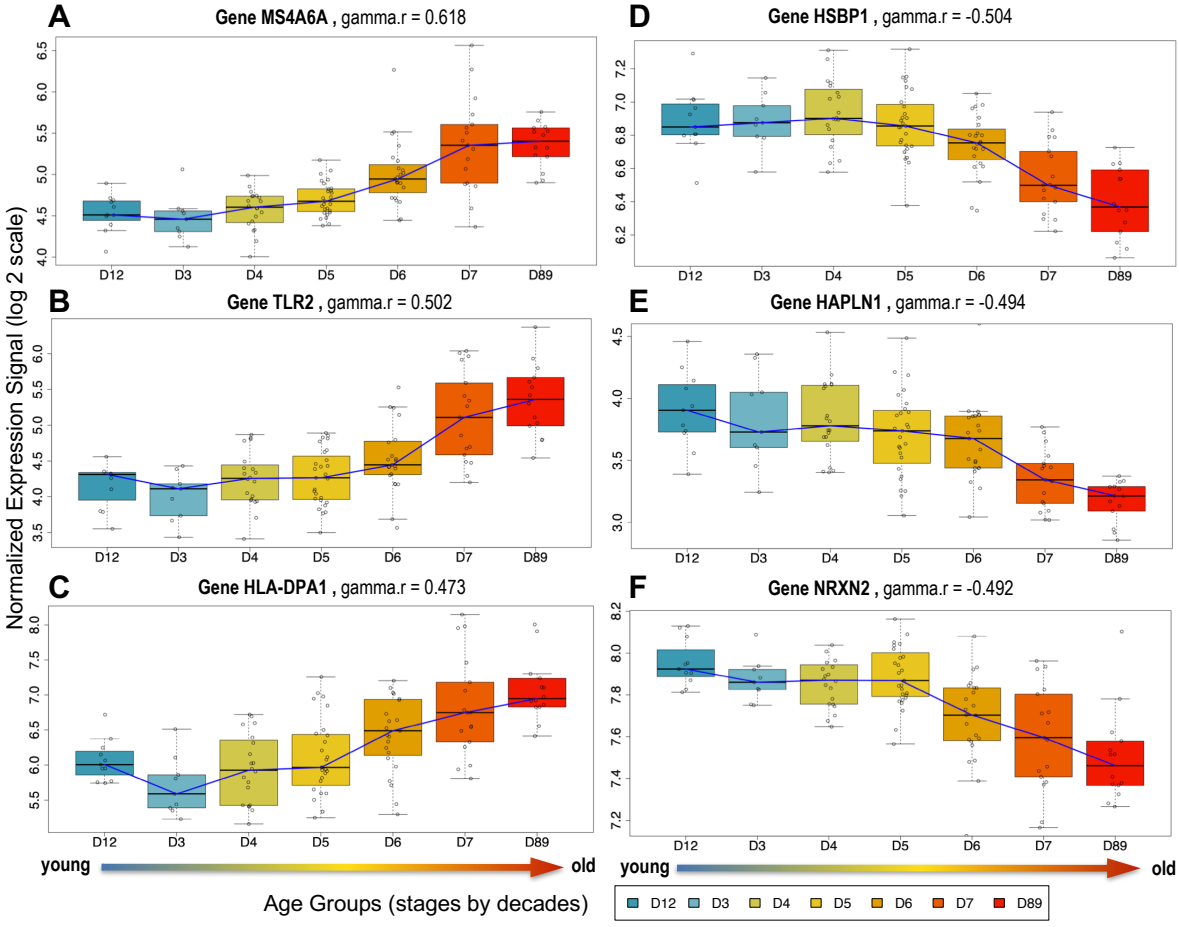


***Figure 7*: Gene expression profiles throughout aging in human hippocampus.** *Gene expression profiles of six genes across age stages (from young to elderly individuals) in 121 human hippocampus samples (from dataset 3, Trabzuni et al. 2013). Plots **A**, **B** and **C** correspond to 3 genes up-regulated: MS4A6A, TLR2 and HLA-DPA1. Plots **D**, **E** and **F** correspond to 3 genes down-regulated: HSBP1, HAPLN1 and NRXN2. The gamma r coefficient of each gene is indicated on the label.*

are molecular chaperones that protect the proteome by folding denature polypeptides and promoting the degradation of severely damaged proteins. These molecules are regulated in mammalian and non-mammalian cells, responding to stimuli that enhance longevity and become impaired during aging (Calderwood et al., 2009). As a general result, the trends observed for these two genes (HLA-DPA1 showing overexpression with age, and DNAJB5 showing repression with age) were consistent across the 4 datasets analyzed (**Figure 7**). This emphasizes the coherence and replicability of the results achieved by our method with independent datasets.

The dataset 2 (Hernandez et al., 2012) was produced with a transcriptomic platform (*Illumina HumanHT-12 V3.0 Expression Beadchip* arrays) that included a smaller number of human genes (approximately 12,000 human UniGene IDs). For this reason, a direct comparison or combination with the other 3 datasets was not possible. These were performed with *Affymetrix* high-density microarrays (*HG U133 Plus 2.0* and *Human Exon 1.0 ST*), and included a common set of approximately 20,000 human ENSG IDs. Therefore, only the gamma correlations and p-values of these 3 datasets were mathematically combined, and dataset 2 was used to corroborate the results for the genes measured in common with the other platforms. We generated combined adjusted p-values using the Stouffer's Z-score method (a powerful method for combining probabilities in meta-analysis (Zaykin, 2011)), merging the p-values obtained within each of the 3 comparable datasets (1, 3 and 4 in **Table 1**).

The combination of gamma correlations and p-values provided a robust evaluation of each gene and allowed the selection of the best features to produce the gene signatures associated with the datasets of cortex, hippocampus and cerebellum. The genes in each signature were ranked by significance, based on the combined p-values (setting up a threshold <0.01), and selecting only the genes that maintained the same sign of the gamma correlation (positive or negative) across datasets. Following these criteria, we proceed to the selection of the genes included in the aging signatures. The signature from cortex included 1148 genes (456 up-regulated and 692 down-regulated). The signature from hippocampus included 874 genes (546 up-regulated and 328 down-regulated). The signature from cerebellum included 657 genes (323 up-regulated and 334 down-regulated).

To further test the existence of a consistent profile of brain genes that are up-regulated with aging and another profile of genes that are down-regulated with aging, we applied a different methodology, separated from the gamma correlation study. This has been described in Materials and Methods, and the results of this alternative analysis are shown in **Figure 8**. A pattern of up-regulation is found in the four datasets and also a

pattern of down-regulation, confirming the results obtained with the application of the gamma correlation to each dataset.

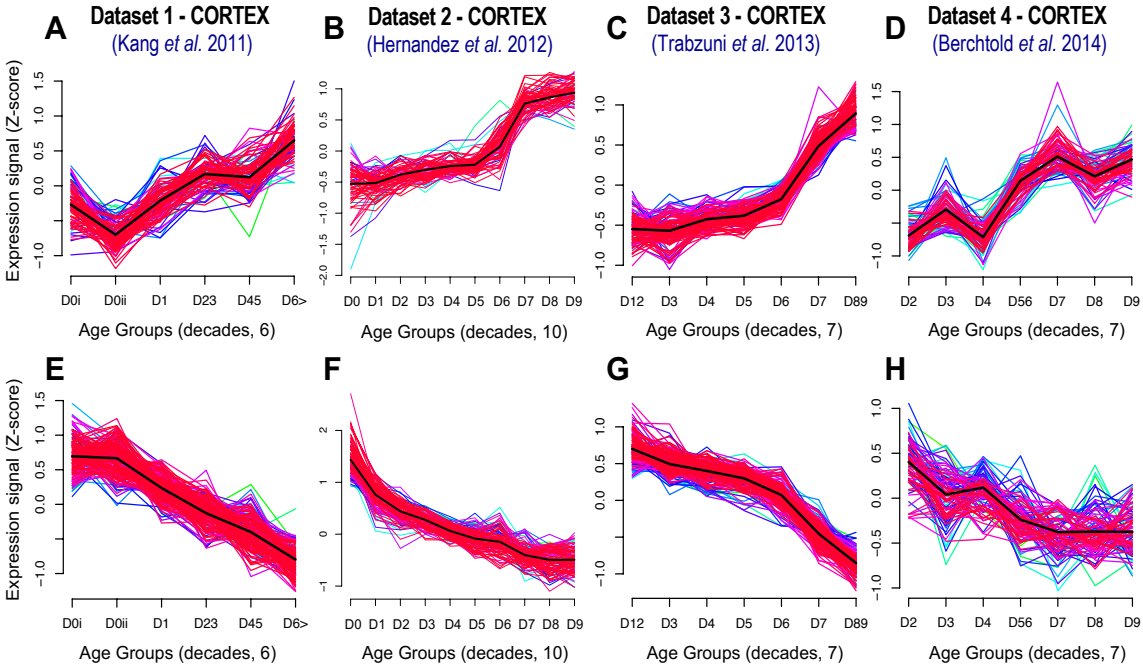


Figure 8: ***Main coexpression trends with aging detected in human cerebral cortex.*** *Identification of the most significant coexpression clusters of genes found in the human brain cortex samples of four different independent datasets. The datasets are described in* ***Table 1****. The clusters were obtained using the fuzzy c-means algorithm and show a significant correlation along age, presenting two main trends: up-regulation with age (plots* ***A****,* ***B****,* ***C*** *and* ***D****) and down-regulation with age (plots* ***E****,* ***F****,* ***G*** *and* ***H****). The samples of each independent dataset were grouped by decades with increasing age.*

Finally, it is worth to indicate that the gene signature found associated with aging in cortex includes more down-regulated genes (692) than up-regulated genes (456), revealing that the gene or genomic effect of aging, at least in the brain cortex, seems more tilted towards a loss-of-function than to a gain-of-function. However, this cannot be taken as general observation because changes in the hippocampus do not follow this trend. To avoid an excessive focus only on the cerebral cortex, we performed a similar analysis of the data obtained from the hippocampus and the cerebellum. **Figure 7** presents the expression profiles produced for 6 genes from hippocampus that showed a high gamma correlation with age stages (using the hippocampus dataset with the largest number of samples: 121) (Trabzuni et al., 2013). Again, the analysis of the expression profiles throughout 7 age stages showed some genes with a clear over-expression trend, from young to old ages, and other group of genes that showed significant repression with age. The genes involved in these two opposite trajectories are similar to those found in the cerebral cortex: showing, for example, the increase of an HLA gene product (HLA-DPA1) and the decrease of a Heat shock protein (HSBP1).

4.4 Presence of genes related to heparan sulfate proteoglycan biology

Interestingly, among the altered genes, several are related to heparan sulfate proteoglycan (HSPG) biology. Examples of these are full and partial time HSPG core proteins CD44, SDC4 and neurexins (Mitsou et al., 2017); the HS biosynthetic glycosyltransferase EXTL3 that starts HS chain biosynthesis; and several HSPG binding proteins including: CAMK2, EGR1, EPHB2 and EPHB3, IGF1R, LPL and LRP2. In brain, HSPGs play an important role in synaptic stability, development of specific synaptic connectivity patterns important for neural circuit function, and in axon guidance (Condomitti & De Wit, 2018). To date, about 800 genes in the whole human genome have been reported to be involved in HS biology, either by binding HSPGs or because they are involved in their biosynthesis (Ori et al., 2011). The observed alteration of HSPG biology-related genes in this study, particularly those associated with functions of neurons, synapses, and neuronal receptors, agrees with the potential involvement of these complex sugars in aging (Huynh et al., 2012). One particular enzyme involved in heparan sulfate homeostasis, specific of brain and neurons, called HS3ST2, have been associated with neurodegeneration (Sepulveda-Diaz et al., 2015). However, this work showed a clear link of this enzyme with Alzheimer's disease-related Tau pathology, more than with aging.

4.5 Aging signatures and regulatory profile derived from TFBS and TF enrichment

In order to identify Transcription Factors (TFs) involved in the regulation of the gene signatures that we found in the brain aging, we performed a genomic analysis of the significant enrichment in Transcription Factors Binding Sites (TFBS) in the set of genes included in the aging signatures of cortex, hippocampus and cerebellum. Before looking for the candidate regulators of these gene signatures, we performed a global comparison to show their overlap. **Figure 9A** presents a comparative analysis of the 3 aging signatures, based on the use of proportional *Venn* diagrams, that reveals a larger intersection between cortex and hippocampus and a clear separation of cerebellum. In fact, 73% of the genes in the cerebellum signature (480 / 657) were only detected in this region of the brain. This result agrees with the distribution of samples observed in the t-SNE analysis presented in **Figure 5**.

The overlap between cortex and hippocampus allowed the identification of a common gene signature that included 258 genes. **Figure 9B** presents a schematic network view of these 258 genes, separated by gamma correlation in a group of 129 up-regulated genes (i.e., genes with positive gamma, marked with blue border in the figure), and another group of 129 down-regulated genes (with negative gamma, marked with red border in the figure). We used this common signature to search for enrichment in TFBS and associated TF regulators as indicated in Methods. The search on the promoters of these genes provided the identification of 4 TFs with significant enrichment score, that regulate 67 genes (of the 129 over-expressed with age) and 91 genes re-pressed with age. This sub-set of genes (67+91=158) are marked in purple color in **Figure 9B**, and represent the genes that can be directly regulated by the 4 TFs identified: FOSL1,2; MEF2A,D; PDX1 and RFX5. As it can be seen, the identification of a specific TF is not unique because two paralogous genes (like MEF2A and MEF2D) bind to the same TFBS and therefore are both candidate regulators of the gene signature tested. The results also indicated that FOSL1,2 and RFX5 had more links with up-regulated genes, and MEF2A,D and PDX1 had more links with down-regulated genes.

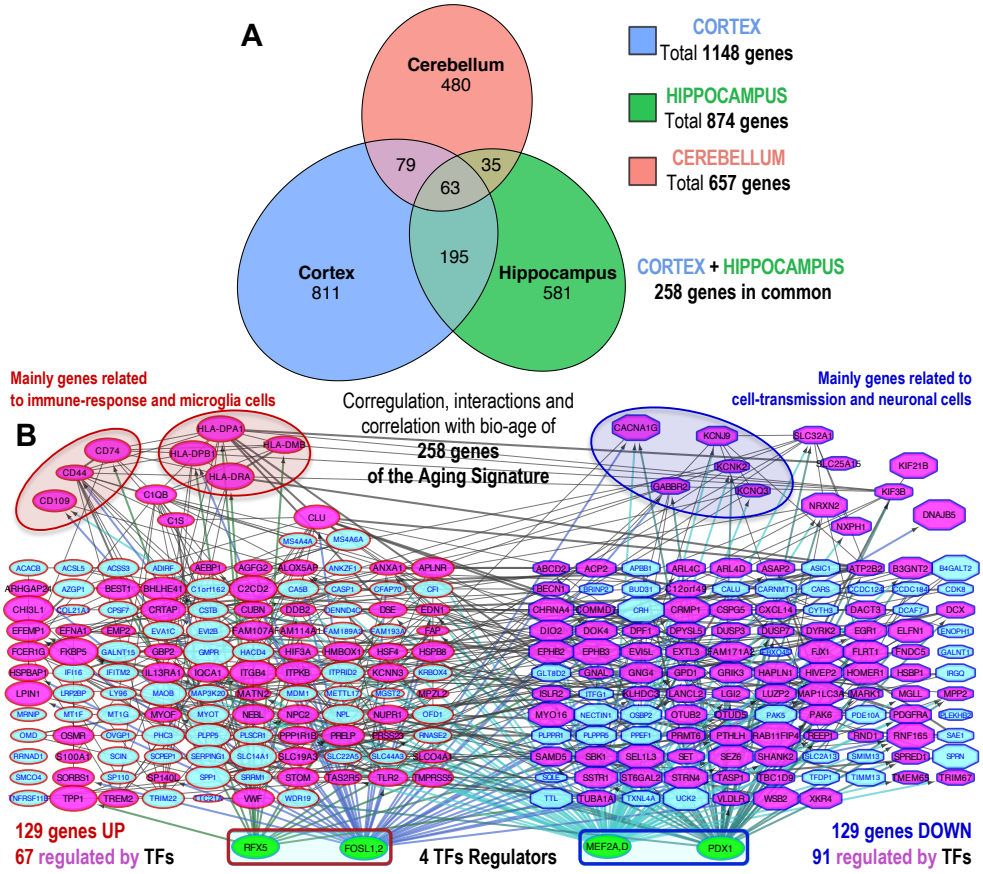


***Figure 9**: Overlap of brain aging signatures and regulatory network derived. (A) Superposition of the genes that were obtained in the three brain aging signatures: obtained for cortex (1148 genes), for hippocampus (874 genes), and for cerebellum (657 genes). The figure is produced using proportional Venn diagrams, and the number of genes in each intersection are indicated, as well as the number of genes included in the non-overlapping regions. A set of 258 genes were found in common in cortex and hippocampus. (B) Network presenting the most significant TF regulators of these 258 genes (FOSL1,2 and RFX5 for up-regulation, and MEF2A,D and PDX1 for down-regulation); as well as, the interactions between the genes, and the correlation that each of these genes showed with the biological-age (bio-age, calculated with the DLNN predictor). The size of each gene nodes is proportional to the absolute value of their correlation with bio-age. Up-regulated genes have a red border color and down-regulated genes a dark-blue border color.*

The correlation between the expression of a gene in the samples across the ages and the predicted bio-age for the same samples is represented by the size of the nodes in **Figure 9B**. This correlation shows the largest value for gene HLA-DRA, up-regulated, and for gene DNAJB5, down-regulated. Regarding the biological functions within the aging signature, as described above for **Table 2**, the up-regulated genes are related with the immune system and immune-related factors (genes HLAs and CDs); and the down-regulated genes are related to neural cell transmission and vesicle trafficking (i.e., different potassium and calcium channels, and genes involved in neurotransmission).

To corroborate the previous findings obtained with the common signature of cortex and hippocampus, we performed a complementary analysis with the whole aging signature derived from cortex, that was approximately four times larger. We focused this search using the 1077 coding genes included in this signature (i.e., using only the protein coding genes included in the set of 1148 found in cortex). Within this 1077 coding genes, 411 were up-regulated and 666 were down-regulated. The results of this analysis revealed that 592 genes (a 55% of the signature) were regulated by 6 main TFs: HBP1, MEF2D, REST, TEAD4, FOSL1 and PDX1. All these TFs were found using *iRegulon* method (Janky et al., 2014) with a normalized enrichment score (NES) > 3.0, being HBP1 the one with the best enrichment (NES=4.549). Two of these TFs were part of the aging signature: HBP1 (up-regulated) and MEF2D (down-regulated); and therefore, they can be considered internal regulators associated with the signature. It is interesting to note that HBP1 (high mobility group box transcription factor 1) regulates the timing of neuronal differentiation during cortical development by controlling cell cycle progression (Watanabe et al., 2015). This regulatory factor is over-expressed with age, indicating a positive regulation of neuronal evolution in brain cortex. Similarly, MEF2D (myocyte enhancer factor 2D), is involved in control of neuronal cell differentiation and development, and it has been directly implicated in the regulation of interleukins production by microglia to protect neuronal cells from inflammation-induced death (Peters et al., 2015). As indicated above, this TF was also found as a key regulator of the aging signature of 258 genes, common to cortex and hippocampus. The repression of this gene with age might be correlated with the fate of neuronal cells that we describe below.

General function	Genes	UP/DOWN regulated	Query List	Reference List	Enrichment p.value	Similarity	Silhouette Width	Functional Terms assigned (concurrent enrichment)
Antigens presentation, Immune system recognition	CD44 , CD74 , CUBN , GBP2 , HLA-DMB , HLA-DPA1 , HLA-DPB1 , HLA-DRA , IRF2 , LRP2 , SLC26A11	UP	11 (251)	100 (34208)	2.12E-10	0.63546	0.30980	GO:0042613:MHC class II protein complex (CC); GO:0005768:endosome (CC); GO:0005765:lysosomal membrane (CC); GO:0002504:antigen processing and presentation of peptide or polysaccharide antigen via MHC class II (BP); GO:0060333:interferon-gamma-mediated signaling pathway (BP)
Response to stress and damage	ACSL5 , AQP1 , BCL2 , CASC3 , CLU , CUBN , GAB1 , IGF1R , LSS , MAOB , MGST1 , MGST2 , RPS27L , SESN1 , SGK1 , SNAP23 , TNFRSF11B , TNKS , TXLNG , TXNIP	UP	20 (251)	631 (34208)	5.48E-08	0.41177	0.25326	GO:0006979:response to oxidative stress (BP); GO:0005792:microsome (CC); GO:0031625:ubiquitin protein ligase binding (MF); GO:0031965:nuclear membrane (CC); GO:0005741:mitochondrial outer membrane (CC); GO:0007584: response to nutrient (BP); GO:0006974:response to DNA damage stimulus (BP)
Cell matrix, Cell surface, Cell adhesion	BCL2 , CAPN2 , CD44 , CD53 , CD74 , CLU , FCER1G , GFAP , HLA-DRA , IGF1R , IL17RB , ITGB4 , LPL , LRP2 , NPC2 , PPFIA1 , SORBS1 , SPP1 , TLR2 , TPP1 , VWF	UP	21 (251)	706 (34208)	7.36E-08	0.28987	0.05617	GO:0007160:cell-matrix adhesion (BP); GO:0009986:cell surface (CC); Kegg:04512:ECM-receptor interaction; GO:0009611:response to wounding (BP); Kegg:04510:Focal adhesion; GO:0005764:lysosome (CC); GO:0009897:external side of plasma membrane (CC)
Immune system activation and reponse	C1QB , C5 , CD59 , CFI , CLU , HLA-DMB , HLA-DPB1 , SERPING1 , VWF	UP	9 (251)	112 (34208)	1.52E-07	0.59192	0.40519	Kegg:04610:Complement and coagulation cascades; Kegg:05322:Systemic lupus erythematosus; Kegg:05150:Staphylococcus aureus infection; GO:0006958:complement activation, classical pathway (BP)
Insulin signaling	ACACB , ACSL5 , ARHGAP24 , EEF2K , GAB1 , IGF1R , IRF2 , LPIN1 , LPP , PHKA2 , PLSCR1 , PYGB , RHOQ , SORBS1 , STOM	UP	15 (251)	424 (34208)	6.79E-07	0.44378	0.33031	GO:0032869:cellular response to insulin stimulus (BP); GO:0008286:insulin receptor signaling pathway (BP); Kegg:04910:Insulin signaling pathway; GO:0005925:focal adhesion (CC); GO:0045121:membrane raft (CC)
Response to pathogens	BCL2 , CD74 , FCER1G , HLA-DMB , HLA-DPB1 , TLR2	UP	6 (251)	113 (34208)	1.93E-04	0.74768	0.54034	Kegg:05152:Tuberculosis; Kegg:05145:Toxoplasmosis
Synapse and neurotransmission (glutamate)	ADRA1D, ADRA2A, ARC, CACNA1G, CAMK2N1, CHRNA2, CHRNA4, DLG3, DLGAP3, ERC2 , GABBR2, GNG4, GRIK3, GRIK4, GRIN2A, GRIN3A, GRM2, HCRTR1, HOMER1, KCNIP1, KCNN1, KCNQ2, MAPK8IP1, NLGN4X, NMU, PRKCB, RIMS1, RIMS4, SEPT5, SHANK2, SLC17A7, SLC1A6, SSTR1, STX1A, SYT5, SYT6, TTYH3, ZNRF1	DOWN	38 (264)	617 (34208)	4.87E-23	0.40280	0.09959	GO:0045202:synapse (CC); GO:0045211:postsynaptic membrane (CC); GO:0005216:ion channel activity (MF); Kegg:04080:Neuroactive ligand-receptor interaction; Kegg:04724:Glutamatergic synapse; GO:0007215:glutamate signaling pathway (BP); GO:0005234:extracellular-glutamate-gated ion channel activity (MF); IPR001508:NMDA receptor; IPR001320:Ionotropic glutamate receptor; IPR019594: Glutamate receptor, L-glutamate/glycine-binding; GO:0043195:terminal button (CC)
Neurons and dendrites	ACP2 , ARC , CAMK2N1 , CXADR , DCX , DLGAP3 , ERC2 , GABBR2 , GRIN2A , GRIN3A , NLGN4X , PTGS2 , SEPT5 , SLC17A7 , SLC32A1 , SLC8A2 , STRN4 , STX1A	DOWN	18 (264)	246 (34208)	9.88E-13	0.51733	0.29743	GO:0043005:neuron projection (CC); GO:0045202:synapse (CC); GO:0019717:synaptosome (CC); GO:0043197:dendritic spine (CC)
Ion channels	CACNA1G , CACNB3 , CACNG4 , KCNF1 , KCNG1 , KCNIP1 , KCNK3 , KCNN1 , KCNQ2 , SLC24A3 , TMEM38A , TNFAIP1	DOWN	12 (264)	125 (34208)	2.83E-10	0.66012	0.64734	GO:0071805:potassium ion transmembrane transport; GO:0005267: potassium channel activity (MF); GO:0005244:voltage-gated ion channel activity (MF); GO:0008076:voltage-gated potassium channel complex (CC); GO:0005249:voltage-gated potassium channel activity (MF)
Cell-cell connection, Endocytosis-phagocytosis, Chemokine signaling	AGAP2 , AMOTL1 , ARC , ASAP1 , ASAP2 , CX3CL1 , CXADR , CXCL14 , DPYSL5 , EXPH5 , GNG4 , GPD1 , HMGCS1 , MAGI1 , MCOLN1 , MMD , MYO16 , PDGFRA , PIP5K1C , PPP2R2C , PRKCB , PRKCZ , PTGS2 , RAB11FIP4 , RASGRP1 , RASSF5 , SMAD3 , STRN4 , STX1A , SYT5 , TNFAIP1 , ZNRF1	DOWN	32 (264)	1157 (34208)	5.10E-10	0.28527	0.09443	Kegg:04144:Endocytosis; GO:0005923:tight junction (CC); GO:0005768:endosome (CC); IPR002219:Protein kinase C-like, phorbol ester/diacylglycerol binding; Kegg:04666:Fc gamma R-mediated phagocytosis; Kegg:04530:Tight junction; GO:0043234:protein complex (CC); Kegg:04062:Chemokine signaling pathway; GO:0005625:soluble fraction (CC)
Kinase signaling, Calcium signaling	ADRA1D , CACNA1G , CACNB3 , CACNG4 , CAMK4 , CCND2 , DUSP14 , DUSP3 , DUSP6 , DUSP7 , FGD1 , GRIN2A , IGF1 , ITPKA , MAPK8IP1 , NR4A1 , PAK6 , PDGFRA , PIP5K1C , PRKCB , RASGRP1 , SLC8A2	DOWN	22 (264)	618 (34208)	3.61E-09	0.28437	0.17669	Kegg:04010:MAPK signaling pathway; GO:0035335:;GO:0004725:protein tyrosine phosphatase activity (MF); IPR000387:Protein-tyrosine/Dual-specificity phosphatase; IPR020422:Dual specificity phosphatase, subgroup, catalytic domain; Kegg:04020:Calcium signaling pathway; Kegg:04510:Focal adhesion; Kegg:04810:Regulation of actin cytoskeleton

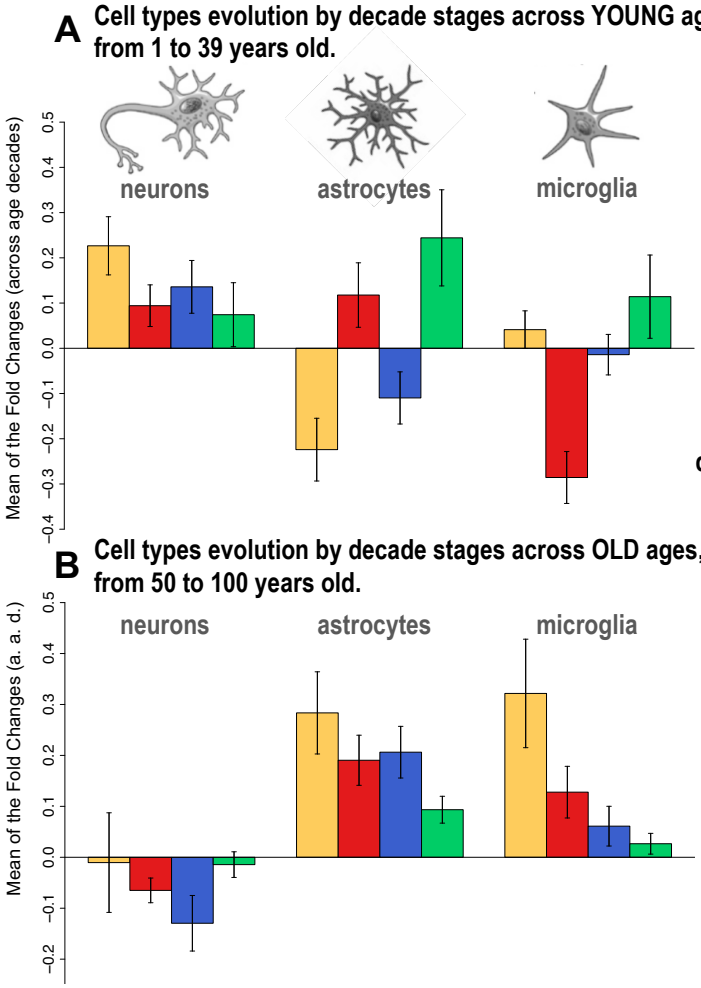
***Table 2**: Functional enrichment analysis of the 300 top genes UP-regulated and 300 top DOWN-regulated with aging in human brain cortex. The enrichment is done by concurrent (co-occurrence) method in five annotation spaces (GO-BP, GO-MF, GO-CC, KEGG and INTERPRO) using GeneTerm-Linker method.*

4.6 Loss of neurons with aging and increase in astrocytes and microglia activity

As we have four independent datasets with full transcriptomic profiles of the genes corresponding to brain cortex samples from individuals of many different ages, from young to old people; we could investigate in these samples the status and evolution of specific cell types throughout the lifespan of the individuals. To do this, we need to find subsets of specific genes that can mark specific cell types in a selective and distinguishable way.

Using single cell RNA sequencing, (Darmanis et al., 2015) identified a set of molecular markers that allow classifying cell types in the human brain. Indeed, these authors proposed a set of 21 specific genes to identify neurons, other set of 21 genes to identify astrocytes and other set to identify microglia (**Figure 10**). We used these gene sets to identify their change in expression levels between the studied age periods; calculating the average fold change of all the 21 specific genes on a cell type in young individuals in the initial decades of life (from 1 to 39 years old) (**Figure 10A**) and the average fold change in old individuals in the late decades of life (from 50 to 100 years old) (**Figure 10B**). The standard error of the mean (SEM) was calculated in the same way for the mentioned analysis (**Figure 10A,B**). It is important to note that the approach taken, dividing the data into a youth age frame (1 to <40 years) versus an elderly age frame (>50 to 100 years), is done because we want to make a solid comparison between young ages and advanced ages; not taken only into account just one stage change, but as many as we can include within these time frames for each dataset. The described approach is quite sensitive, because the calculation is done across all the decade stages within each age period frame (i.e., young ages *versus* old ages), and measures the mean change in expression of the 21 genes included in each cell type signature.

The results throughout the young ages show a clear increase in neuron related genes (i.e., in the activity of neuronal genes), with not consistent change in the levels of the genes that mark astrocytes and microglia (**Figure 10A**). On the contrary, there is a clear decrease in neuronal signal (i.e., a repression of neuronal genes) throughout the latest periods of life, with a consistent and steady increase in the levels of genes assigned to astrocytes and microglia (**Figure 10B**). An increase in neuronal genes across the young stages of life (up to 39 years), suggests that the brain increments or intensifies its activity across the first periods of life, from children to mature people. By contrast, the results show a neural decline and degradation observed in the last periods of life, mainly after 60-65 years.



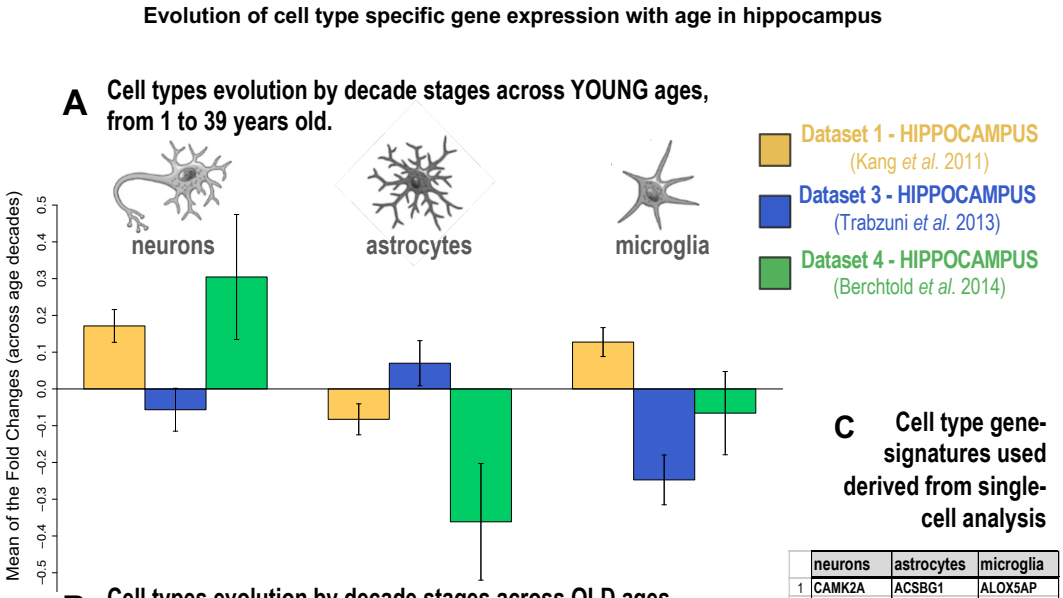
C

Cell type gene-signatures used derived from single-cell analysis

	neurons	astrocytes	microglia
1	CAMK2A	ACSBG1	ALOX5AP
2	CCK	AGT	C1QA
3	CELF4	AGXT2L1	C3
4	CHGB	AQP4	C3AR1
5	CIT	ATP13A4	CCL3
6	CKMT1B	BMPR1B	CCL4
7	DNM1	FGFR3	CD53
8	GABRA1	GFAP	CD74
9	GABRB2	GJA1	DHRS9
10	GABRG2	GJB6	FCGR3A
11	MAP7D2	GPR37L1	GPR183
12	MEG3	GPR98	IL1B
13	NRXN3	MGST1	ITGAX
14	SCG2	PRODH	LAPTM5
15	SCN2A	RANBP3L	LCP1
16	SNAP25	SDC4	LILRB4
17	SYNPR	SFXN5	MSR1
18	THY1	SLC25A18	OLR1
19	TMEM130	SLC39A12	PLEK
20	UNC80	SLC4A4	PTPRC
21	VSNL1	SLCO1C1	RGS10

***Figure 10**: Evolution of cell type specific gene expression with aging in cerebral cortex. Evolution in 4 independent datasets of RNA expression of 3 cell-type specific gene signatures derived from single-cell analysis. These signatures are used to identify the presence of 3 cell-types (neurons, astrocytes and microglia) in the human cerebral cortex throughout different age stages across young ages or across old ages. Plot (A) presents the mean of the fold changes that occur across the age decades in young individuals (from 1 to 39 years old); plot (B) presents the mean of the fold changes that occur across age decades in elderly individuals (from 50 to 100 years old). The lists of genes included in the gene signatures derived from single-cell analysis are included in panel (C).*

The same analyses were performed for the hippocampus using the 3 datasets from this region of the brain, revealing a very similar evolution of the cells signal through age stages, both across the young ages and across the old ages (shown in **Figure 11**). This confirms the initial results, suggesting that the expression changes in cortex and hippocampus with aging follow a similar pattern and affect to similar sets of genes and biological functions.



	neurons	astrocytes	microglia
1	CAMK2A	ACSBG1	ALOX5AP
2	CCK	AGT	C1QA
3	CELF4	AGXT2L1	C3
4	CHGB	AQP4	C3AR1
5	CIT	ATP13A4	CCL3
6	CKMT1B	BMPR1B	CCL4
7	DNM1	FGFR3	CD53
8	GABRA1	GFAP	CD74
9	GABRB2	GJA1	DHRS9
10	GABRG2	GJB6	FCGR3A
11	MAP7D2	GPR37L1	GPR183
12	MEG3	GPR98	IL1B
13	NRXN3	MGST1	ITGAX
14	SCG2	PRODH	LAPTM5
15	SCN2A	RANBP3L	LCP1
16	SNAP25	SDC4	LILRB4
17	SYNPR	SFXN5	MSR1
18	THY1	SLC25A18	OLR1
19	TMEM130	SLC39A12	PLEK
20	UNC80	SLC4A4	PTPRC
21	VSNL1	SLCO1C1	RGS10

***Figure 11**: Evolution of cell type specific gene expression with aging in hippocampus. Evolution, in 3 independent datasets of the expression level of 3 cell type specific gene signatures derived from single-cell analysis. These signatures are used to identify the presence of 3 cell types (neurons, astrocytes and microglia) in the human hippocampus throughout different age stages across young ages or across old ages. In this way, plot (A) presents the mean of the fold changes that occur across the age decades in young individuals (from 1 to 39 years old); and plot (B) presents the mean of the fold changes that occur across the age decades in elderly individuals (from 50 to 100 years old).*

4.7 Deep-learning Neural Network provides reliable calculation of biological age

A DLNN was built to calculate individual age based on the information provided by the aging gene expression signature that we obtained in this work. The performance of the neural network was compared with two other machine learning algorithms (random-forest, RF, and support vector machine, SVM) applying exactly the same input data for training and validation (**Figure 12A,B,C**). The regression parameters provided by each of the three predictors (DLNN, RF and SVM) indicate that the neural network (with an adjusted R^2=0.909) is the one that shows the lowest error in the calculation of the biological age with respect to the chronological age (i.e., the nominal age that is known for each person included in the validation set in this study). The R^2, coefficient of determination, is a statistical measure of how well the regression predictions approximate the real data points. Thus, for the three methods compared using the validation dataset, the regression parameters (adjusted R^2, adjusted p-values of the regression model and RMSError) allowed us to measure the correctness of each method. This showed that the biological age predictor provided by the Deep Learning Neural-Network (DLNN) method performed better than the predictors obtained using the Support-Vector-Machine (SVM) and the Random-Forest (RF) methods; since, using an equal configuration of input data, SVM and RF had R^2 values of 0.77 and 0.86, respectively.

4.7.1 Comparison of the DLNN predictor with other predictors of biological age

As far as we know, there are not many studies that use deep neural networks technique for predicting human age. A related and relevant study conducted by Putin and co-workers (Putin et al., 2016) used multiple DLNNs, stacked into an ensemble, to build a predictor of human chronological age, taking the data from common clinical blood tests. These authors trained their DLNNs using tens of thousands of samples from common blood biochemistry and cell count tests of patients undergoing routine physical examinations. The results of this work showed that the best performing DLNN, applied to the test dataset, achieved 81.5% epsilon-accuracy within a 10-year frame, r = 0.90, R^2 = 0.80 and a Mean Absolute Error (MAE) of 6.07 years in predicting chronological age. When we calculated the same parameters for our DLNN predictor of age, we obtained: 85.3 % epsilon-accuracy within a 10-year frame, r = 0.954 with adjusted R^2 = 0.91 and MAE = 5.78 years. These results indicate that the predictors based on molecular biomarkers can be more accurate than predictors derived from basic blood tests parameters.

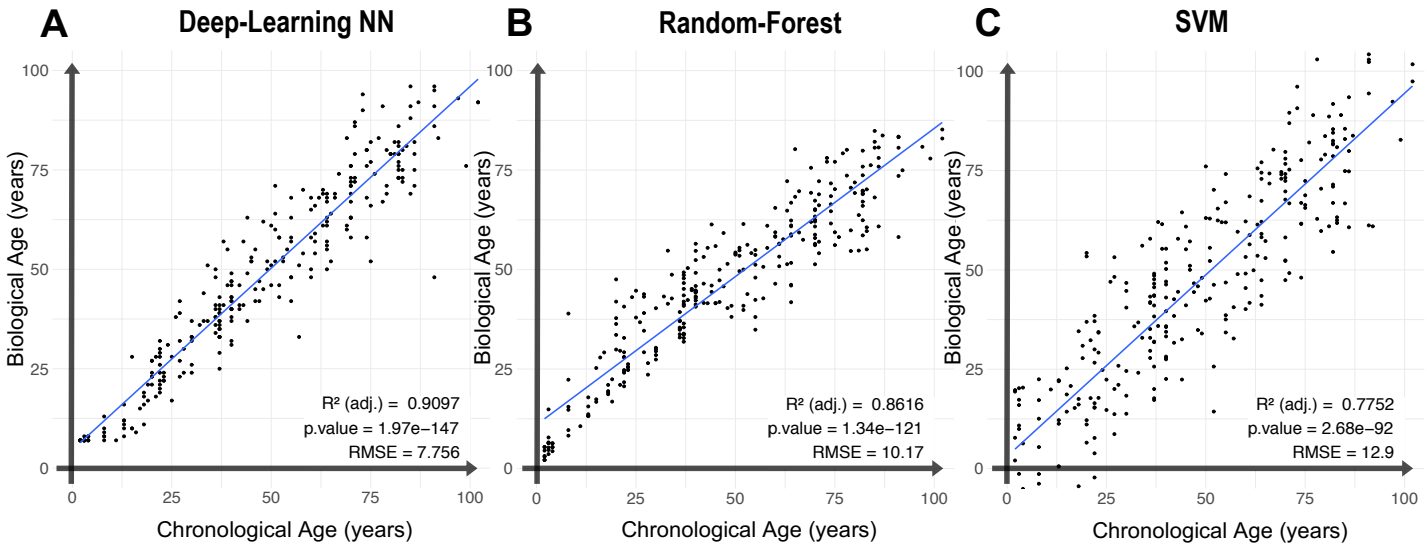
Several efforts have been made in recent years to build biological age predictors based on omic derived markers, knowing that the hallmarks of aging are based on a series of genomic and epigenomic alterations (López-Otín et al., 2013). In this way, several aging clocks able to predict human age using various biomarkers have already been proposed (Putin et al., 2016). Methylation-based markers such as epigenetic aging clocks are currently the most accurate, while transcriptomics-based age predictors have shown to be less accurate (Peters et al., 2015). In particular, Steve Horvath in 2013 developed a multi-tissue predictor of age based on DNA methylation data (Horvath, 2013). This approach was quite interesting and provided the DNA methylation age across different tissues. However, this work used a simple metric to assess the accuracy of the predictive model: a Pearson correlation between the observed and the predicted age. This can lead to erroneous interpretations since it is not the optimal measurement to compare between regression models. By contrast, we used the adjusted R^2 parameter, which calculates the square of the correlation between observed and predicted variables, adjusting the values by the degree of freedom used in the model. Moreover, the *r* correlations obtained with the epigenetic marks using the model provided by Horvath ranged from 0.98 to as low as 0.46; that showed a great variability of the prediction power and that would result, if transformed to adjusted R^2 values, in worse predictions than those provided by our DLNN based predictor of age.

Other authors also used genome-wide methylation profiles to build a predictor of biological age (Hannum et al., 2013). Again, in that work the metric used to assess the accuracy of the model differs from the one we used in our work. In fact, Hannum and collaborators wrote in their article: "The predictions were highly accurate, with a correlation between age and predicted age of 91% and an error of 4.9 years". This number (i.e., a correlation coefficient of $r = 0.91$ on the test data) corresponds to a non-adjusted R^2 of 0.83, which is quite below our adjusted R^2 of 0.91. With respect to the low RMSE of 4.9 years that its reported in their paper based on genome-wide methylation: our higher RMSE of 7.7 years, based on genome-wide expression, could be explained by a higher variability of the transcriptomic age. Indeed, this can be expected considering that the shift of the transcriptomic profile of the brain cells at different ages will always be much larger that the shifts or changes in the methylation profile. Finally, (Hannum et al., 2013) indicated in their article that: "The model included both methylome and clinical parameters, such as gender and body mass index (BMI)"; therefore, it is not a model based uniquely on biomolecular data, but includes clinical covariates. This makes their approach different to our biological age predictor, which is only based on gene expression.

4.7.2 Finding genes whose expression correlates best with biological age

As described in Materials and Methods, we calculate the correlation of the gene expression with the biological age. The list of the top 20 genes with best mean positive *r* coefficient, calculated as *Spearman* correlation, and the top 20 genes with best mean negative *r* coefficient are included in **Figure 12D,E**. In these lists we marked the genes with best correlations: (i) best positive was GMPR (guanosine monophosphate reductase); and (ii) best negative was DNAJB5 (DnaJ heat shock protein family Hsp40 member B5). We also highlighted in these tables two other genes that presented a good correlation in our study, and that have been previously related with brain aging: GFAP (glial fibrillary acidic protein) (Nichols et al., 1993) and PNOC (prepronociceptin) (Rhinn & Abeliovich, 2017). Interestingly, the correlations with our DLNN bio-age predictor for these two genes (GFAP and PNOC) are better than the ones reported in the referred studies (Rhinn & Abeliovich, 2017).

Finally, we observed that the correlation between the genes of the aging signature obtained and the chronological age was generally worse than the correlation with biological age predicted by our DLNN method. In other words, the gene expression profiles showed a lower correlation with the chrono-age, indicating that the genes are prone to measure the biological age of the sample. Moreover, it is clear that the chronological age reflects the known distance at the time when each individual was born, and, by contrast, that biological age reflects the biological state or biological situation of the brain. In this way, a negative displacement with respect to chronological age (lower predicted bio-age) would indicate that the individual is in better health conditions than expected, and a positive displacement with respect to chronological age (higher predicted bio-age) would indicate that the individual is in worse health conditions than expected. A final relevant consideration with respect to the biological age predictor proposed here is that being derived only from cerebral cortex data, it should primarily reflect brain cognitive age rather than another type of physiological age.



D Top 20 genes of best positive *Spearman* correlation with **Biological Age**

Gene Symbol ID	Mean Corr with BioAGe	Gene Description
C2CD2	0.60574	C2 calcium dependent domain containing 2 [HGNC:1266]
CD74	0.56796	CD74 molecule [HGNC:1697]
CHI3L1	0.65504	chitinase 3 like 1 [HGNC:1932]
CLU	0.55673	clusterin [HGNC:2095]
DYSF	0.54747	dysferlin [HGNC:3097]
FKBP5	0.64142	FKBP prolyl isomerase 5 [HGNC:3721]
GFAP	0.55495	glial fibrillary acidic protein [HGNC:4235]
GMPR	0.68672	guanosine monophosphate reductase [HGNC:4376]
HLA-DPA1	0.60533	major histocompatibility complex, class II, DPalpha 1 [HGNC:4938]
HLA-DPB1	0.58213	major histocompatibility complex, class II, DPbeta 1 [HGNC:4940]
HLA-DRA	0.62844	major histocompatibility complex, class II, DRalpha [HGNC:4947]
ITGB4	0.58707	integrin subunit beta 4 [HGNC:6158]
ITPKB	0.54751	inositol-trisphosphate 3-kinase B [HGNC:6179]
LPIN1	0.58197	lipin 1 [HGNC:13345]
MAOB	0.56587	monoamine oxidase B [HGNC:6834]
PLPP5	0.55346	phospholipid phosphatase 5 [HGNC:25026]
RCL1	0.54645	RNA terminal phosphate cyclase like 1 [HGNC:17687]
RPS6KA5	0.58891	ribosomal protein S6 kinase A5 [HGNC:10434]
SLC14A1	0.58430	solute carrier fam. 14 member 1 (kidd blood group)[HGNC:10918]
TPP1	0.59927	tripeptidyl peptidase 1 [HGNC:2073]

E Top 20 genes of best negative *Spearman* correlation with **Biological Age**

Gene Symbol ID	Mean Corr with BioAGe	Gene Description
ADGRB2	-0.58784	adhesion G protein-coupled receptor B2 [HGNC:944]
B4GALT2	-0.58908	beta-1,4-galactosyltransferase 2 [HGNC:925]
CACNA1G	-0.59276	calcium voltage-gated channel subunit alpha1 G [HGNC:1394]
CX3CL1	-0.60234	C-X3-C motif chemokine ligand 1 [HGNC:10647]
DNAJB5	-0.70282	DnaJ heat shock protein fam. (Hsp40) member B5 [HGNC:14887]
DPYSL4	-0.58685	dihydropyrimidinase like 4 [HGNC:3016]
EPHB3	-0.59838	EPH receptor B3 [HGNC:3394]
GPR26	-0.60415	G protein-coupled receptor 26 [HGNC:4481]
KIF21B	-0.59067	kinesin family member 21B [HGNC:29442]
MARCH4	-0.66160	membrane associated ring-CH-type finger 4 [HGNC:29269]
NREP	-0.58812	neuronal regeneration related protein [HGNC:16834]
OLFM1	-0.60750	olfactomedin 1 [HGNC:17187]
PNOC	-0.51468	prepronociceptin [HGNC:9163]
RAB11FIP4	-0.62334	RAB11 family interacting protein 4 [HGNC:30267]
RNF165	-0.59137	ring finger protein 165 [HGNC:31696]
SEMA6B	-0.58579	semaphorin 6B [HGNC:10739]
SMPD3	-0.68615	sphingomyelin phosphodiesterase 3 [HGNC:14240]
TMEM8B	-0.60885	transmembrane protein 8B [HGNC:21427]
TRIB2	-0.63471	tribbles pseudokinase 2 [HGNC:30809]
TTC9B	-0.59174	tetratricopeptide repeat domain 9B [HGNC:26395]

***Figure 12**: (A,B,C) Comparison of machine learning model performances with our model based on DLNN. (D) Top genes whose real mRNA expression signal positively correlated with the predicted bio-age. (E) Top genes whose real mRNA expression signal negatively correlated with the predicted bio-age.*

5 CONCLUSIONS

Aging can be broadly defined as the time-dependent functional decline that affects most living organisms, characterized by a progressive loss of physiological integrity, leading to impaired function, deterioration and increased vulnerability to death (López-Otín et al., 2013). In a more balanced view, aging can be formulated as the collision between destructive processes that act on cells and organs over lifetime and the positive responses that promote homeostasis recovery, vitality and longevity. However, the precise molecular elements that mark aging and the mechanisms that determine the rates of aging in organisms are not well known. Here, we have explored and analyzed the transcriptomic landscape and gene regulatory profile of the human brain linked to age and aging. Our results reveal that we can identify in the brain a consistent gene signature able to reflect the time passing and the aging. This includes not only genes that are repressed and show loss-of-functions, but also genes that indicate a positive reaction and gain-of-function, possibly as response to the damage, deterioration, and stress that time imposes on our body. As a whole, our results present a transcriptomic landscape and gene regulatory profile of the human brain linked to aging, providing the identification of particular biological signatures able to characterize and predict the brain biological age, which can differ from the chronological age of individuals. The results provide an excellent context to better understand the broad cognitive spectrum observed in the healthy aging population, and opens a new way of investigating brain diseases, especially those for which we currently do not have a clear causal clue, as it is the case of Alzheimer's disease.

CHAPTER 2

Alzheimer disease gene signature and new blood biomarkers

1 NOTE REGARDING THE EMBARGO OF THIS CHAPTER

This chapter has been called for an embargo, removing all its content due to unpublished results. These results are pending from a publication on a scientific journal and therefore have been removing from this document (see ***Anexo II***).

CHAPTER 3

Pan-cancer deep learning prediction and profiling tool for tumor samples based on transcriptomic data

1 BRIEF SUMMARY

Today, cancer diagnosis relies on several methodologies: from screening studies, imaging test, biopsy samples, laboratory test including genomic and transcriptomics, to simple clinical data and personal and familiar medical history records. But despite huge advances in those methods, some cases still pose a hard challenge and have been demonstrated to be impossible to diagnose. Among those cases, one is particularly relevant: cancers of unknown primary (CUP). This heterogeneous group of metastatic cancers are particularly difficult to diagnose, since the primary site tumor form where they originated is completely unknown, and thus diagnose and specific treatment are not possible. This presents a challenge from the data analysis point of view, and more specially, as a classification and prediction problem.

Here we present a prediction model based on deep learning neural networks and transcriptomic data, whose aim is to aid in the diagnosis of the primary site of CUP samples, serving as a clinical diagnosis tool that is able to produce an insight of the tissue where the cancer may have been originated, open the possibility for advance treatment and drug use.

We have collected a dataset containing more than 22000 samples from both cancer patients and healthy donors (this superset is composed by the genomic data commons (GDC) and GTEx studies), with a total of 27 primary sites/tissues mapped unambiguously. Additionally, several independent datasets from cancer samples have been included to test the performance of the model. Two main models have been built using the TensorFlow free and open-source software library for machine learning: the first one uses convolutional neural networks with raw transcriptomic data as input; the last one, developed during an internship at the Institute for Computational Biomedicine at the Medical Faculty of Heidelberg University and Heidelberg University Hospital, uses feedforward neural networks with transcription factor and pathway activities, which allows the model to be more simple and lightweight, as the activities are expected to contain more

information and to be more stable across samples.

Results show a great potential for the pan-cancer prediction tool: both models achieved an accuracy of 97% and 96% on the validation dataset respectively at predicting the tissue of origin of the samples. More importantly, the convolutional neural network model achieved great accuracy when tested on external datasets of primary tumors and distant metastasis, being able to generalize the patterns found by their more intensive feature extraction layers.

2 INTRODUCTION

2.1 Cancers of unknown primary

Cancers of Unknown Primary (CUP) represent a heterogeneous group of metastatic cancer that are poorly differentiated and whose primary site is not known at the time of diagnosis. CUP poses a difficult challenge for both biomolecular understanding of causality and, consequaently, effective diagnosis and treatment. The incidence of CUP in the last decade is around 2% considering all cancer cases (Fizazi et al., 2015; Lee & Sanoff, 2020), but despite these reduced percentage, the absolute number of patients affected is high and the mechanisms underlying the carcinogenesis and progression of CUP remain elusive. Additionally, CUP can be falsely or prematurely diagnosed in patients who undergo suboptimal investigations at the time of presentation and in those in whom a primary tumour becomes detectable during the disease course after the initial diagnosis (these are around 10% of the patients with CUP).

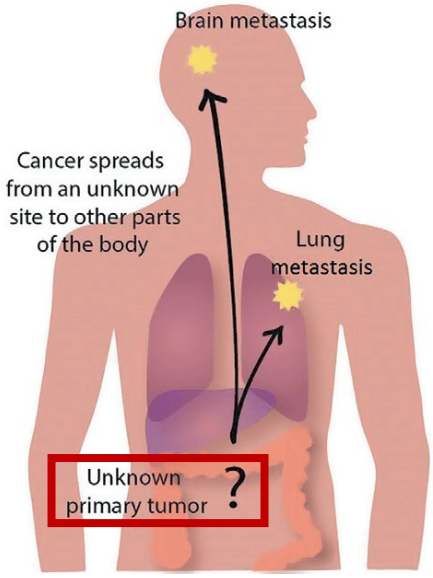


***Figure 13**: An example of a dissemination of a cancer, originating a distant metastasis. If the analysis is not able to determine the primary site, it is classified as Cancer of Unknown Primary.*

When compared to metastases originating from known primary tumours, the unique natural history of CUP is characterized by early dissemination, an aggressive clinical course, an unpredictable metastatic pattern, intrinsic treatment resistance and a dismal prognosis (Lee & Sanoff, 2020). Thus, is imperative to obtain a biomolecular signature of CUP, not just for diagnostic purposes but for understanding the mechanism of progression, resistance, and dissemination of these type of cancers.

Recent studies have demonstrated that a high majority of cancer cell lines resemble their tissue of origin at transcriptomic and methylome level (Salvadores et al., 2020), opening opportunities for drug screening and genetic screening research.

2.2 Artificial intelligence and cancer genomics

As already exposed, cancer is an extremely complex disease which is very difficult to diagnose, and thus accurate predictions of prognosis and treatment remain elusive. Advances in recent years and access to larges collections of datasets is changing this situation rapidly: a search of scientific literature shows that the number of research projects on cancer has increased exponentially, and those papers that involved machine learning and artificial intelligence techniques applied to large collection of cancer samples, has grown specially fast (Obermeyer & Emanuel, 2016), and several interesting reviews have tried to address this new trend in computational biology and medicine (Bi et al., 2019; S. Huang et al., 2020; Levine et al., 2019).

Two main targets have been selected by clinicians and researchers: cancer prognosis prediction and clinical imaging classification. Among prognosis analysis, the most common are probability of metastatic development, tumour recurrence, patient management or death after treatment. For example, AI has been recently used to predict the chances of developing brain metastasis (S. Huang et al., 2019), Ching et. al. used transcriptomic data from breast cancer samples to predict patient prognoses using neural networks that is an extension of the Cox regression model (Ching et al., 2018).

Bomane A. et al. also used transcriptomic data in addition to methylome data, but this time to predict Breast Cancer patient response to paclitaxel drug (cytotoxic-drug sensitivities) for optimizing personalized therapies in clinical practice (Bomane et al., 2019). Oh et al. used a survival recurrent network (SRN) for predicting the survival rate of gastric cancer patients compared to the Cox proportional hazard regression model, results showed a better performance than Cox models and corresponded closely with the observed survival data (Oh et al., 2019).

Diagnosis using AI based on imaging is currently under an intensive research as deep learning neural networks have shown to excel at image processing. A recent study has demonstrated to improve the diagnostic accuracy of thyroid cancer, by using convolutional networks to analyse sonographic imaging data from clinical ultrasounds (X. Li et al., 2019).

Another study was able to classify perifissural nodules of lung cancer screening images, the authors of the paper stated that this approach could be more efficient and use to reduce the number of follow-up visits. The researchers showed that the accuracy of their CNN model was close to that of human experts (Ciompi et al., 2015). It is widely believe that the use of AI on clinical image classification will help to the diagnosis of cancer in the near future, opening the possibility of large screening population studies (Topol, 2019).

Finally, and related with the objectives of this thesis chapter, metastatic identification has been tackled in recent research using AI, as an example Liu Y. et. al. developed a deep learning algorithm called LYmph Node Assistant that was able to detect metastatic breast cancer in sentinel lymph node biopsies, with the aim of improving pathologist's productivity and reducing false negative overall (Y. Liu et al., 2019).

2.3 Pan-cancer prediction and personalized biomolecular profiling

The aim of the tool that will be presented throughout this chapter is to create a sequential pipeline that, when a single sample of raw RNA-Seq expression data from a metastatic cancer is given, the application will give a personalized profile of the sample including:

- Possible primary sites with probabilities regarding the precedence of the primary tumour site.
- Possible cancer subtype (when the detected primary site is subject to subtype decomposition, such as breast or lung cancer).
- Biomolecular profile: dysregulated pathways, transcription factor activity, drug resistance… etc.

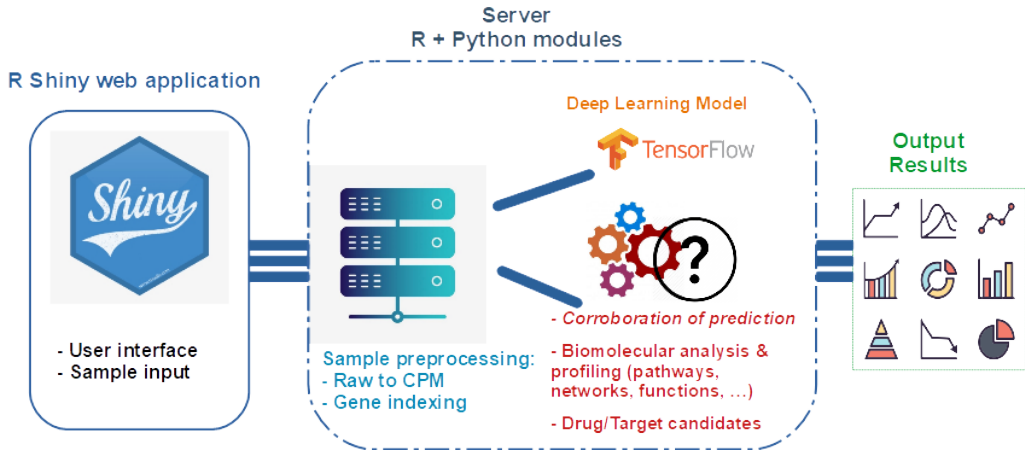


***Figure 14**: Diagram of the tool being built, with the Deep Learning model -that we discuss throughout this chapter- integrated as the key component. Ideally, this system could help at the diagnostic level, by processing a single sample of a metastatic CUP, analyzing it and showing relevant results that would help to diagnose and find the primary site of the cancer being studied.*

We have started by the developed of a deep learning neural network (DLNN) composed by multiple convolutional layers plus several fully connected layers at the output, that processes an artificially created bio-image per RNA-Seq sample using a special z-score, in which we subtract a cancer-gene-driver expression instead of the mean value of the sample.

The output gives us the probability of membership for each tissue/primary site of the sample.

3 MATERIAL AND METHODS

3.1 Samples

The data used for the Neural Network training and validation, and to be used in the subsequent parts of the pipeline, consists of approximately 22,000 RNA-Seq samples from various cancer datasets (from the Genomic Data Commons Data Portal, https://portal.gdc.cancer.gov/) as well as samples from healthy tissues (from the Genotype-Tissue Expression Data Portal, https://gtexportal.org/home/).

With these two main datasets, we aim to clearly identify the pattern of the tissue of origin and disregard the signal variation due to cancerous processes. A recent study (Salvadores et al., 2020) has identify that cancer cells strongly resemble the tissue of origin and thus, it may be possible to isolate these patterns and predict the primary site.

Genomic Data Commons (GDC) Data Portal:	GTEX consortium:
Cancer samples with diagnosed primary site • 15000 samples • 27 primary site tissues selected	**Sample from normal tissue** • 22000 samples available • 18 tissues map with GDC data tissue

***Table 3**: Compendium of datasets used for the training and validation steps of the Deep Learning models discussed along this chapter 3.*

After filtering we have a *reference dataset of approximately 22 000 samples* with balanced class groups, unambiguous tissue of origin, cancer sub-type and other clinical data.

Tissue/Primary Site	Samples
Kidney	1443
Uterus	1325
Hematopoietic and reticuloendothelial systems	3337
Brain	3426
Thyroid gland	1205
Breast	1774
Esophagus	1606
Skin	2261
Mouth and tongue and cartilage	320
Adrenal gland	554
Bronchus and lung	1857
Lymph nodes	534
Stomach	756
Bladder	424
Thymus	92
Colon	1370
Liver and intrahepatic bile ducts	683
Testis	518
Larynx	193
Ovary	569
Prostate gland	878
Pancreas	500
Connective, subcutaneous and other soft tissues	130
Rectum	88
Bones and cartilage	92
Heart, mediastinum, and pleura	118
Tonsil	41

***Table 4**: Distribution of samples of the dataset for the label 'primary site', this label is the one that has been used as the output prediction of the convolutional neural network. As it can be seen, the variability between the number of samples in the different groups is important. With "Tonsil" being the least numerous groups, having just 41 samples in total. The most represented group is "Hematopoietic and reticuloendothelial systems" with 3426 samples.*

Disease (Cancer)	Samples
Adenomas and Adenocarcinomas	5207
Cystic, Mucinous and Serous Neoplasms	699
Plasma Cell Tumors	831
Myeloid Leukemias	732
Healthy Control	11503
Ductal and Lobular Neoplasms	1424
Myomatous Neoplasms	61
Nevi and Melanomas	451
Lymphoid Leukemias	551
Squamous Cell Neoplasms	1506
Mature B-Cell Lymphomas	144
Gliomas	782
Complex Mixed and Stromal Neoplasms	269
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	504
Transitional Cell Papillomas and Carcinomas	401
Thymic Epithelial Neoplasms	120
Paragangliomas and Glomus Tumors	162
Epithelial Neoplasms, NOS	13
Nerve Sheath Tumors	7
Chronic Myeloproliferative Disorders	80
Osseous and Chondromatous Neoplasms	88
Soft Tissue Tumors and Sarcomas, NOS	32
Germ Cell Neoplasms	156
Mesothelial Neoplasms	83
Fibromatous Neoplasms	35
Lipomatous Neoplasms	11
Myelodysplastic Syndromes	7
Neuroepitheliomatous Neoplasms	82
Leukemias, NOS	104
Synovial-like Neoplasms	10
Acinar Cell Neoplasms	20
Complex Epithelial Neoplasms	18

***Table 5**: Distribution of samples of the dataset for the label 'disease'. As it can be seen, the number of samples between groups vary considerably, with the most common types of cancers like "Adenomas and Adenocarcinomas" having 5207 samples. Worth noting, the group "Healthy Control", corresponding with normal healthy tissue, is composed of 11503 samples.*

As it can be seen in **Table 5**, roughly the half of all the samples correspond with healthy controls (11503 samples). After control group, the most numerous corresponds with Adenomas and Adenocarcinomas with 5207 cases.

3.2 Convolutional Deep Learning Neural Network Model

3.2.1 Features

The selection of input features -or explanatory variables- to be used for the deep learning model is not trivial. The constrain of computational complexity forces the input space to be as reduced as possible, as the number of trainable parameters needed grows exponentially for each new input feature. Also, as the transcriptomic data by nature is prone to high levels of noise and cofounding factors, we need the selected variables to be as informative and robust as possible.

The unprocessed input data of the pipeline will be raw RNA-Seq reads, these reads are normalized for each sample using the entire library (~50 000 genes) to counts per million (CPM) (**Eq. 12**), so that distributions across samples can be directly compared.

$$CPM_i = \frac{x_i}{N} \times 10^6 \qquad \textit{Eq. 12}$$

Where:

x_i are the counts for gene i in the sample.

N is the total number of counts for all genes in the sample.

3.2.2 Bioimage

After normalization we have selected -for the first deep learning model- a representative subgroup of genes, in order to make it feasible to train and deploy the pipeline and the deep learning model.

For this purpose, we have selected 295 cancer driver genes that have been demonstrated to be tied with a large number of cancer types and subtypes (M. H. Bailey et al., 2018), additionally we have selected the whole group of human transcription factors (TF) that mapped to our sample space, composed of 1394 genes. The inclusion of these transcription factors is due to their intrinsic regulatory functions: they orchestrate gene expression such as that genes targeted by these TF are expressed in the right amount throughout the life of the cell and the organism (Vierstra et al., 2020), and by extension groups of TFs function in a coordinated way so that they control important cell states such

as: cell division, cell growth, and cell death throughout life; because of this, genetic variation in regulatory regions has been connected with diseases and diverse phenotypic traits (Vierstra et al., 2020). To summarize, the selected number of features for the first deep learning model is: 295 cancer driver genes + 1394 human transcription factors.

As it stands, with these variables we do not have sufficient information regarding the interaction of the selected genes between themselves. Here we present a solution to produce additional gene interaction information that can be exploited by convolutional neural networks: after a series of z-score transformations, we construct a matrix of numerical values that serve as an image. In this way, we can make use of neural network models that are able to process and classify images, which have been demonstrated to be among the best classifiers nowadays.

The columns of this bioimage matrix are to be computed as follow:

$$S_{z-score} = \frac{x_i - \hat{\mu}}{\hat{\sigma}} \qquad \textit{Eq. 13}$$

Where:

x_i are the counts in CPM for gene i.

$\hat{\mu}$ is the mean of a sample S in CPM.

$\hat{\sigma}$ is the standard deviation of a sample S in CPM.

This first column corresponds to the CPM sample itself standardized using the median z-score with $\hat{\mu} = 0$ and $\hat{\sigma} = 1$, so the unaltered expression data is preserved on the bioimage. The remaining columns are computed using a modified version of **Eq. 13**: instead of subtracting the mean of the sample, we subtract the value of each of the cancer driver genes, obtaining a series of vectors (295 in total) of values whose centroid is the cancer driver gene itself. By doing this we can measure how all of our feature genes are behaving in contrast to each of our selected cancer drivers.

$$S_{gene\ i\ z-score} = \frac{x_i - c_i}{\hat{\sigma}} \qquad \textit{Eq. 14}$$

Where:

x_i are the counts in CPM for gene I in sample S.

c_i is the CPM value of a cancer-driver-gene $c \in c_1, \ldots, c_{295}$ of the sample S in CPM.

$\hat{\sigma}$ is the standard deviation of the sample S in CPM.

The final dimension of this matrix computed for a single sample S is: **1689 rows** corresponding to the TF genes (1394) plus the cancer driver genes (295). The number of

columns is 296, corresponding to the median z-score transformation of the sample (1 column) and the gene-z-score transformations with al the cancer driver genes (295, one transformation of the sample S per cancer driver gene).

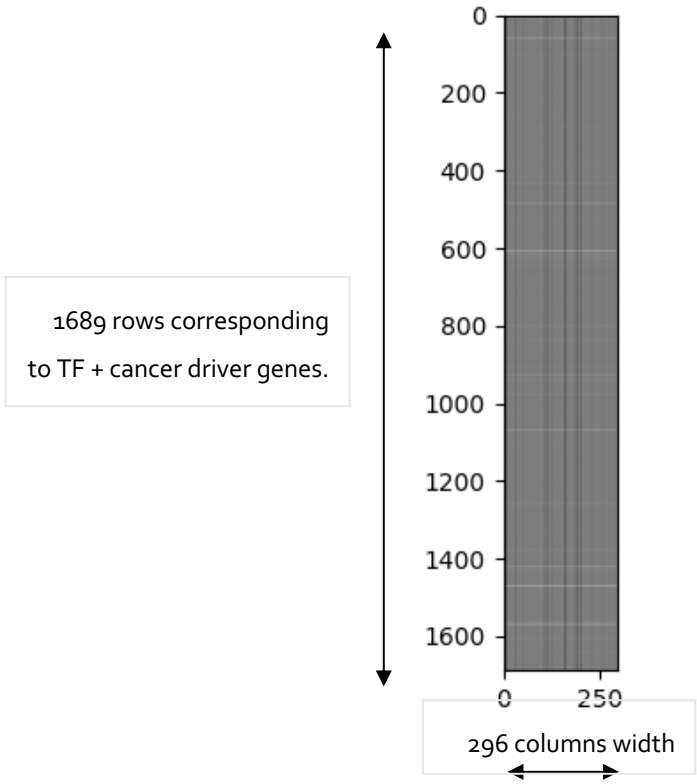


***Figure 15**: Real representation of a bioimage of a sample form the GDC dataset, the z-score values of the matrix has been upscaled to match the grey scale of a RGB image with 1 channel. Line patterns can be seen across the image, these spatial patterns are learned and used by the convolutional neural network to make the predictions.*

3.2.3 Convolutional Model Architecture

A convolutional neural network (CNN) is a class of deep learning neural network that uses shared-weight convolution kernels that scan the input matrix and the hidden layers with translation invariance characteristics (Aloysius & Geetha, 2018), which makes them especially successful in image and video processing and classification, as their unique characteristics makes them powerful feature extraction models.

CNN relies in the same mathematical components than those of the more general and classic feedforward deep neural networks (already discussed on the chapter one, section 3.6 of this thesis), such as ReLU or PReLU objective function (**Eq. 2**), stochastic

gradient descent (**Eq. 5**) and backpropagation algorithms (**Eq. 10**), among others. The main difference is the use of kernels with shared-weights and pooling layers (**Figure 16**: an example of a basic convolutional model with a 3x3 kernel. The kernel is applied over the input matrix 5x5: the green and red squares correspond with the same kernel being applied in different positions and showing the obtained result as new data points with their respective colors (usually a max-pooling layer is also applied, which means that the maximum value of the multiplication between the kernel and the matrix slice will be chosen), this 3x3 reduced output will be processed by other kernels on the next layer.), that are able to capture spatial patterns and reduce the dimensionality of the input in such a way, that for each layer a higher pattern abstraction is achieve.

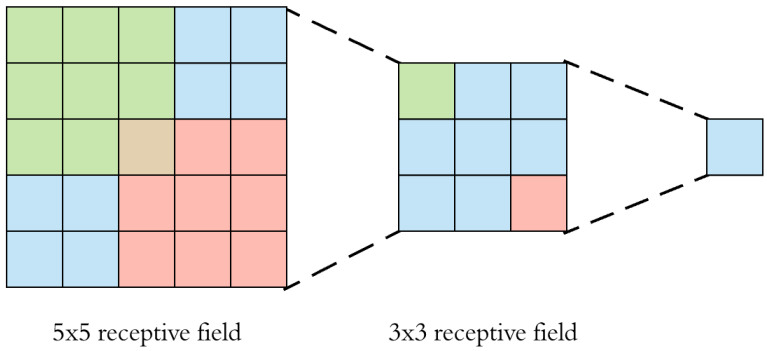


***Figure 16**: an example of a basic convolutional model with a 3x3 kernel. The kernel is applied over the input matrix 5x5: the green and red squares correspond with the same kernel being applied in different positions and showing the obtained result as new data points with their respective colors (usually a max-pooling layer is also applied, which means that the maximum value of the multiplication between the kernel and the matrix slice will be chosen), this 3x3 reduced output will be processed by other kernels on the next layer.*

The architecture which better performed and that we have chosen for our model has 9 hidden convolutional layers (**Figure 17**): the first 2 hidden layers have 32 kernels each, the following 2 have 64, the subsequent 2 layers 128 kernels, 256 kernels the following 2 hidden layers, and the last convolutional hidden layer has a total of 512 kernels, all the kernels have a dimension of 3x3. Additional max-pooling layers for dimensionality reduction and dropout layers for reduce overfitting are positioned between the mentioned hidden layers.

As our regression model is a classification model with a restricted space of labels (27 different tissues), we need the output to be a vector of probabilities associated with those observed labels, but as the convolutional model works with matrix elements, we need to connect a simple feedforward neural network with 3 hidden layers to the output of the

convolutional model (with 2064, 664 and 86 neurons each respectively). To connect them, we flatten the output of the CNN to a single vector. In this case, the output layer will have an activation function (in contrast with the DLNN described in the Chapter 1 section 3.6.2 that did not needed an activation function) that will convert the output of the model to probabilities, and also an associated cost function $J(\theta, b)$ that will calculate the error of the whole model during training, this error is used to learn from past observations and fit the parameters of the convolutional and the fully connected neural networks.

We need to define first the activation function of the output layer, as we need to calculate the probabilities associated to each layer and then compute the associated error for this probability distribution. We will use the softmax function **(Eq. 15)**: it normalizes the outputs of the neural network model so that they sum to 1, and therefore they can be directly treated as probabilities over the output.

$$\widehat{y}_i = f(x_i) = \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}} \qquad \textit{Eq. 15}$$

Once we have obtained the desired probabilities for each class, we compute the error loss of the $\widehat{Y}$ predicted probabilities and the real observed Y, for this purpose we will use the *categorical crossentropy function* which is defined as:

$$Loss = -\sum_{i=1}^{output\ size} y_i \cdot \log \widehat{y}_i \qquad \textit{Eq. 16}$$

This loss serves as a measure of how distinguishable two discrete probability distributions are from each other. As explained in chapter 1 section 3.6.2, we use stochastic gradient descent to train the model and update its parameter, and therefore we take steps proportional to the negative of the gradient. The minus sign ensures that the loss gets smaller when the distributions get closer to each other (the gradient vector can be interpreted as the direction and rate of fastest increase, thus the negative gradient vector will point to the direction and rate of fastest decrease).

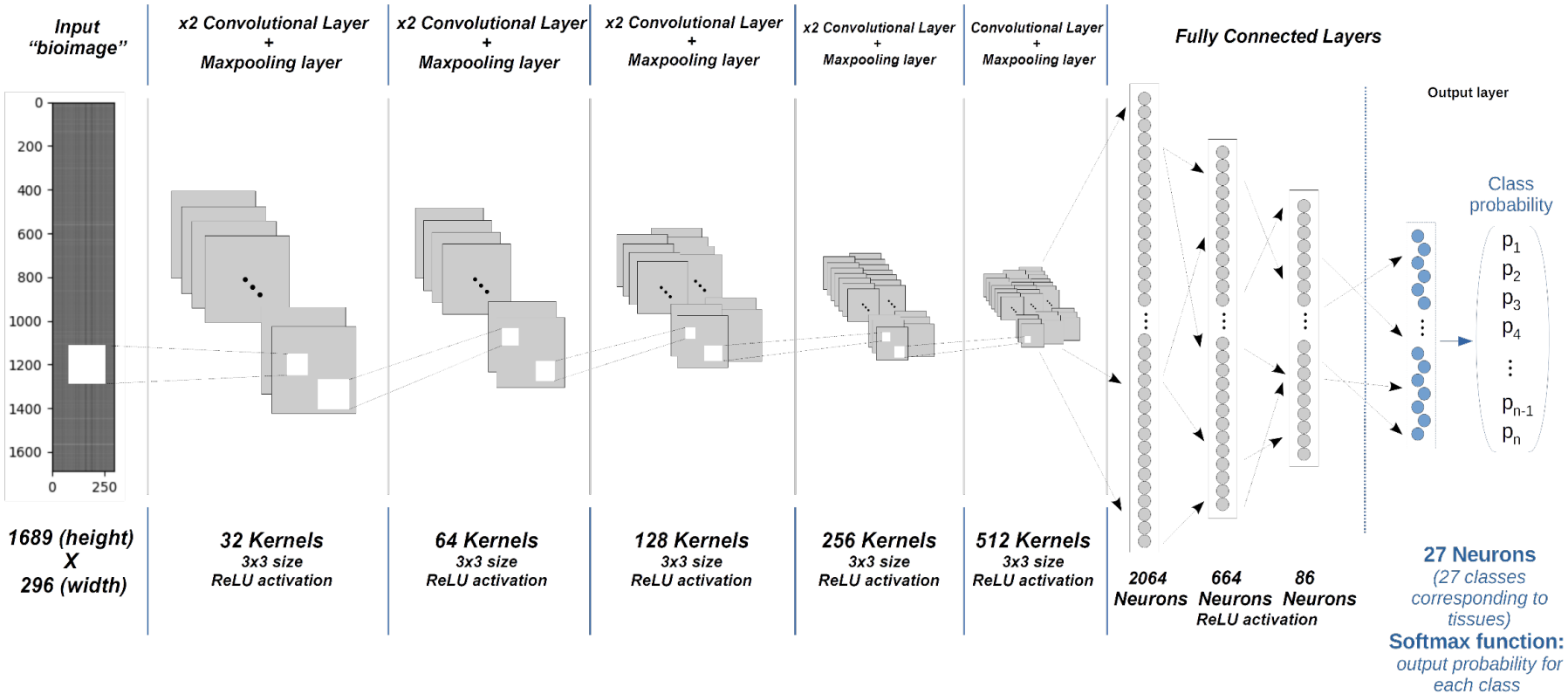


***Figure 17**: architecture used on the convolutional neural network. The model processes the input bioimage, sequential convolutional layers and max-pooling layers extract more complex features as the depth of the convolutional network increases. Finally, the kernel's outputs are flattened and connected to a feedforward neural network, which further process these extracted features until at the final layer a probability is given for each of the 27 labels corresponding with the tissues/primary sites.*

3.2.4 Training the convolutional model

For the high costly computational task of training and testing the model (train several deep learning neural network models), we have been granted access to the new Artemisa supercomputer facility (ARTificial Environment for Machine Learning and Innovation in Scientific Advanced Computing) of the Spanish National Research Council (CSIC) and the University of Valencia. All batch machines contain an NVIDIA GPU Volta V100 SMX2 with 32 GB memory each, to assist with their AI algorithms, Each GPU provides 15 TFlops for single precision (32 bits) and 7 TFlops for double precision. Furthermore, those GPU provides specific tensor computing capacity with a rate of 128 TFlops on single precision operations.

As is not possible to compute and load on RAM all the bioimages in a single batch, the data is pre-processed "on the fly" in multiple Python node workers (threads), then the batch of bioimages are forwarded to feed the model that is being run on 2 Tesla GPU's, if the GPU is busy processing a batch of bioimages the new bioimages processed by the Python nodes are buffered so the GPU always has new bioimages to train on the next batch when ready. The selected model that achieved the best accuracy was training on the Artemisa facility for 48 hours.

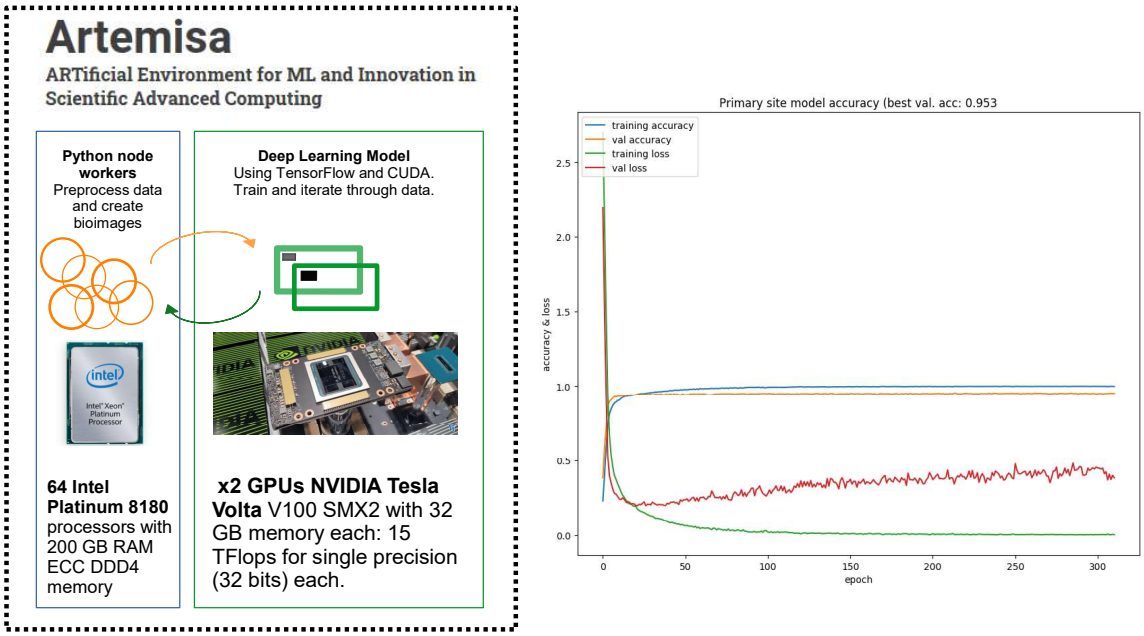


***Figure 18**: Diagram of the computing cluster, with the strategy of preprocessing the bioimages being derived on the CPU, the data is feed into the DLNN model, that is being executed on the GPU. On the right, the values of the error loss and the accuracy on both the training and the validation data is shown.*

In some preliminary analyses with a simpler model we have reached a prediction accuracy over 90% of the primary site on validation data in more than 25 different primary site tissues (the output of the model provides the probability for each class of coming from one of these primary sites), and a similar accuracy for the prediction of the type of the sample, considering these subtypes: normal solid tissue, solid tumour, normal blood, liquid tumour (related with blood cancer), normal bone marrow, tumoral bone marrow and metastasis.

3.3 Feedforward Deep Learning Neural Network Model

3.3.1 Biological Activity

As the computational complexity of the convolutional neural network model is hard to scale for additional datasets, sets of inputs and test, and as the rise of complex biological networks has helped to understand biomolecular concepts and has been able to measure biological activity (such as gene, protein or regulon activity) with a high degree of accuracy, we wanted to explore additional AI models that could make use of such biological activities, whose data is in theory simpler and more robust, and also as the goal of our pan-cancer tool aims to give deeper insights into the biology of the sample that is being studied, we were interested in embedding these network methods into our algorithm,

This new model has been built during an *PhD internship at the Institute for Computational Biomedicine at the Medical Faculty of Heidelberg University and Heidelberg University Hospital directed by Dr. Julio Sáez-Rodriguez.* This internship was supported by a *EMBO Short-Term Fellowship (fellowship id: 8927)* awarded to the author of this thesis. Saez laboratory has been focusing on biological network based on previous knowledge and causality, as a result they have implemented a series of tools that are able to produce meaningful information based on interactions using a wide range of omics data. Thus, we built a new deep learning model using two types of biological activities that could be derived from our RNA-seq dataset in a single-sample approach (since we want the tool to be used as a prediction for single samples from fresh biopsies, the processing of the data must therefore be possible using a single sample/vector of gene expression data): transcription factor activities (using DoRothEA) and pathway activities (using PROGENy).

PROGENy is a method that infer the activity on the level of pathways using gene expression by leveraging a large compendium of publicly available perturbation experiments to yield a common core of Pathway RespOnsive GENes (Schubert et al., 2018). Some advances of PROGENy compared with other methods are: (i) it can recover the effect of known driver mutations, (ii) provide or improve strong markers for drug

indications, and (iii) distinguish between oncogenic and tumor suppressor pathways for patient survival. In summary, PROGENy accurately infers pathway activity from gene expression based on previous knowledge derived from perturbation experiments. Using this tool, we are able to measure 16 pathway activities per sample.

DoRothEA is a gene set resource containing signed transcription factor (TF) target interactions, first described in (Garcia-Alonso et al., 2019). The unit that comprises a Transcription Factor and its targets is called a regulon, the regulons that are available as part of DoRothEA were curated and collected from different types of evidence, such as literature curated resources, ChIP-seq peaks, transcription factor binding site motifs and interactions inferred directly from gene expression. Each TF target interaction has a confidence level assigned based on the number of supporting evidences for such interaction. The confidence assignment comprises five levels, ranging from A (highest confidence) to E (lowest confidence). Interactions that are supported by all four lines of evidence, manually curated by experts in specific reviews, or supported both in at least two curated resources are considered to be highly reliable and were assigned an A level. For the building of our DLNN model, we selected targets that have at least C confidence. Typically, DoRothEA is coupled with the statistical method VIPER as it incorporates the mode of regulation of each TF-target interaction. In this way, through the use of DoRothEA regulons and VIPER method we were able to obtain 117 transcription factors activities per sample.

In total, we are able to produce a total of 133 input features per sample, 16 pathway activities and 117 transcription factor activities.

3.3.2 Feedforward Deep Learning Neural Network Architecture

As already explained, the focus of this new model is to test whether it is possible by using more complex and robust variables containing a sufficient amount of information, we can have a similar performance than the convolutional neural network model which is several times fold more complex. This new model (**Figure 19**) is similar to the one introduced on Chapter 1 section 3.6 (Deep-learning Neural Network method applied to biological age calculation, see **Figure 4**) with different hidden layer sizes, being the most notable differences: the output activation function, in which now we will use the softmax function (**Eq. 15**) and the loss function that we use to calculate the output error, the categorical crossentropy function (**Eq. 16**).

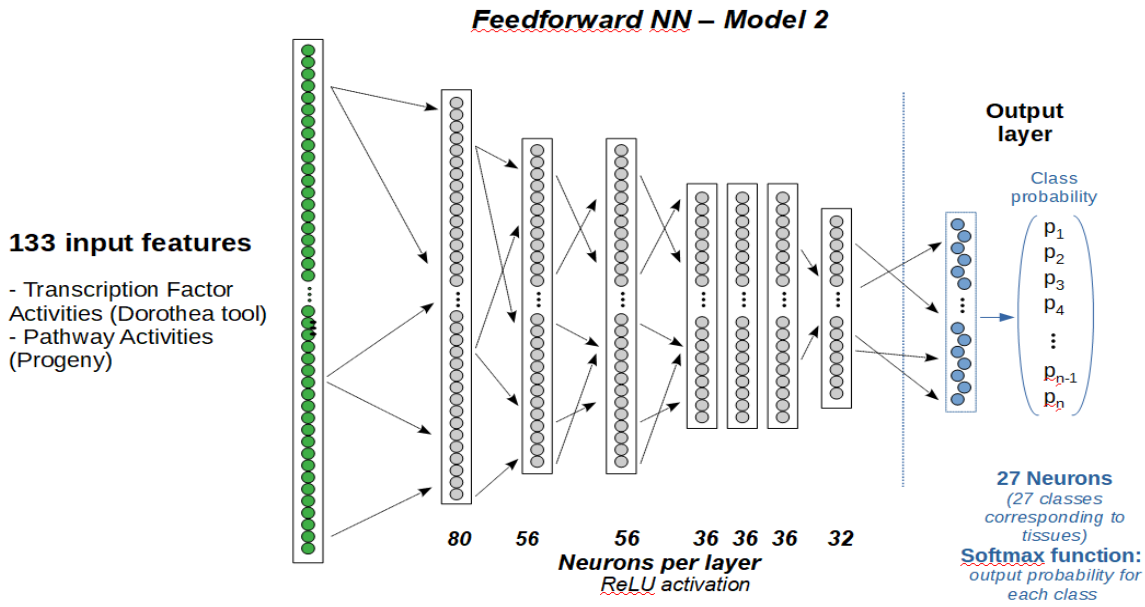


***Figure 19**: Feedforward neural network model using the 133 biological activities derived from DoRothEA and Progeny. This DLNN model is composed of just 7 hidden layers, plus the input layer and the output. The model gives the probability for a sample to come from 1 among 27 possible primary sites labels.*

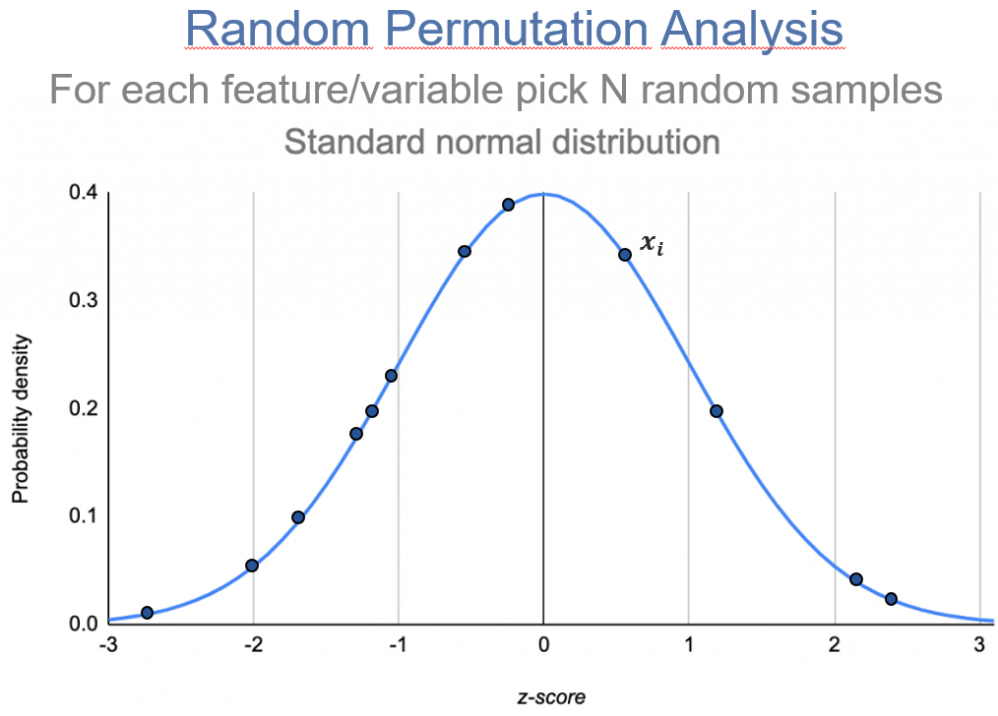
3.4 Explainable Machine Learning

One of the main problems of today's artificial intelligence is the barrier of explainability: the opaque decision systems are hard to interpret and understand, and the decisions derived from such systems ultimately affect humans' lives, economy or other important social aspects (Barredo Arrieta et al., 2020). In the last years there has been a surge of interest in trying to explain the behavior of machine learning algorithms, and more specifically in deep learning neural network models, mainly due because of the parametric space of such models: the latter space comprises hundreds of layers and millions of parameters, which makes DLNNs be considered as complex black-box models (Castelvecchi, 2016).

Here we focus on the idea of feature selection or variable weighting: the interpretability of the model can shed light how only meaningful variables infer successfully the output, acting as an empirical test of at which extent the underlying truthful causality is embedded in the model reasoning (Barredo Arrieta et al., 2020). By extending this process, we can have a collection of scores that could help identifying which variables among all

available variable space in the data are meaningful, and thus help identify causality between input variables and output observations.

One of the most common and early used approaches (Olden & Jackson, 2002) is the random permutation analysis. The idea is to permute randomly the inputs and observe the inferred outputs in order to identify which of the permuted variables has a major impact on the model inference performance.



We followed slightly different approaches for the convolutional N.N. model and the feedforward N.N. model. Since we wanted to test the permutation analysis over the validation data (samples that are not seen by the model at the training step) of the GDC and GTEx dataset (see chapter 3, section 3.1 Samples) which is composed of approximately 7800 samples. Because of the computational complexity of the convolutional model, in which we have firstly to create the bioimage for every sample (see chapter 3, section 0

Bioimage) and run the convolutional model over this image, and repeat this process for each random value for every input variable, it was unfeasible to test in a reasonable time. Because of this reason, and because the input data of the convolutional model is composed of transcriptomic RNA-Seq data counts, we decided to do a permutation analysis setting the value of each gene to 0 counts per million (CPM) at a time.

Following this approach, we were able to generate 7800 bioimages with one of its genes set to 0 CPM for each of the 1689 input genes, yielding approximately 13 million

***Figure 20**: Example of a random permutation analysis showcasing the random sampling of features along the distribution of the input variable x.*

bioimages tested. After more than 24 hours of computation on the AI cluster platform ARTEMISA (see chapter 3, section 3.2.4 Training the convolutional model) we were able to compare the performance of these permuted samples with the original performance achieved by the convolutional model. The results showed a list of genes and its respective tissues/primary sites that were affected by the permutation of those specific gene.

For the biological activity model, we were able to test a complete random permutation analysis, thanks to the lower complexity of the model. For the GDC and GTEx validation samples we generated 50 random values for each feature of each of the 7800 samples using DoRothEA and Progeny tools, the random values from each feature x_i were selected from the range $\ni \{\min(x_i) - 1, \max(x_i) + 1\}$, the decision of adding and subtracting 1 is due to the importance of the behavior of the model regarding extreme outliers which can destabilize the model when facing new data with extreme noise or patient outliers. The total number of inputs tested by this method has been around 400.000 random permuted samples.

Additionally, we tested the permutation analysis on the biological activity model in two external metastatic datasets of ovary and kidney cancer, this time with 400 permutation

of random values for each feature on each sample due to the reduced dimension of the datasets.

4 RESULTS AND DISCUSSION

4.1 Exploring the biological activity information

Because the dimension of the data has been reduced considerably, is now feasible to explore the data, and more interestingly, analyze at what extent the data is sufficiently informative for the DLNN model to successfully predict the tissue of origin of the sample.

Principal Component Analysis (PCA) was applied to the dataset with the transcription factor and pathway activities as features (133 in total). As it can be seen in **Figure 21**, the first two components are able to aggregate some of the samples by group (primary site), especially the primary sites: hematopoietic and reticuloendothelial systems, brain, testis, skin, ovary, and bronchus and lung. The other samples appear to be mixed, what could point to the data not being explanatory enough to classify the samples by its primary site of origin.

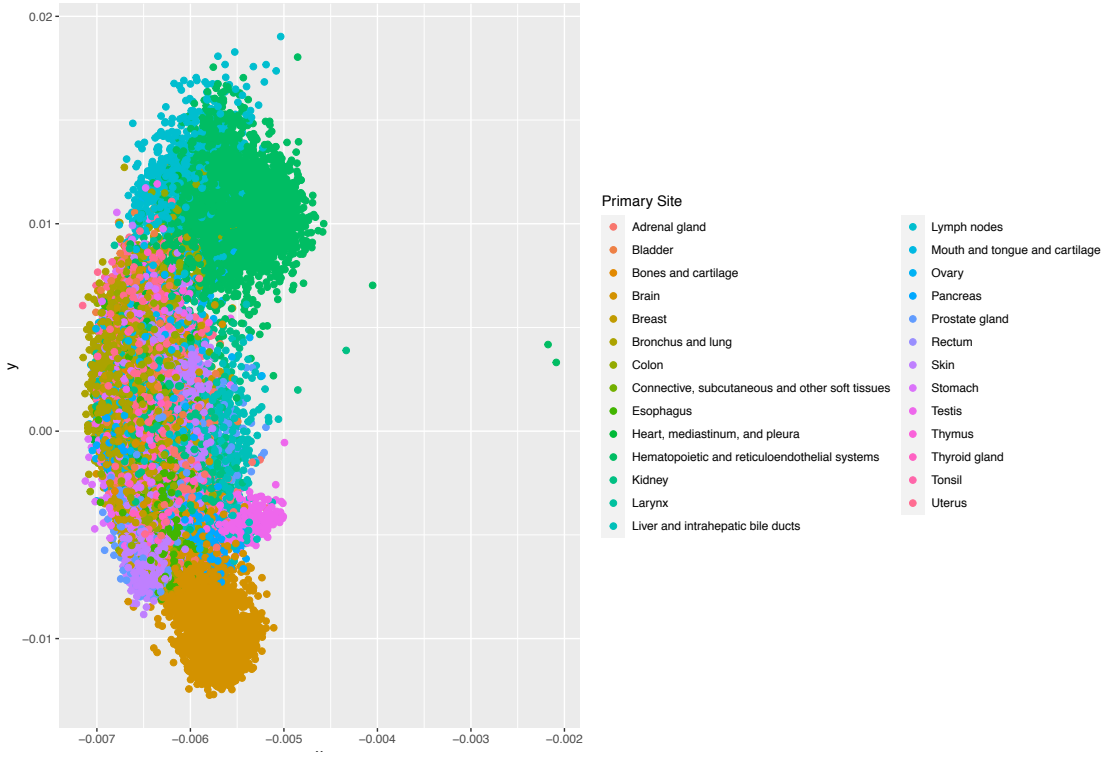


***Figure 21**: PCA applied to the dataset of the GDC and GTEx samples with the transcription factors and pathway activities as features. The x and y axis corresponds with the first and second components.*

In order to examine in depth if the transcription factor and pathway activities are sufficient to split the samples by its tissue of origin, we explored the data using UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, we applied this algorithm over the first 30 dimensions of the PCA obtaining a more detailed separation between groups as can be seen on **Figure 22**. Now, the groups that clustered together using the first 2 components of the PCA (hematopoietic and reticuloendothelial systems, brain, testis, skin, ovary, and bronchus and lung) can be seen that each split in 3 distinct groups that are close to each other, corresponding for each of these primary sites with the GTEx subgroup, the GDC control and the GDC cancerous subgroups. Although additional groups and clusters can be seen in comparison with **Figure 21**, there are still several primary sites that are not able to differentiate in separated clusters.

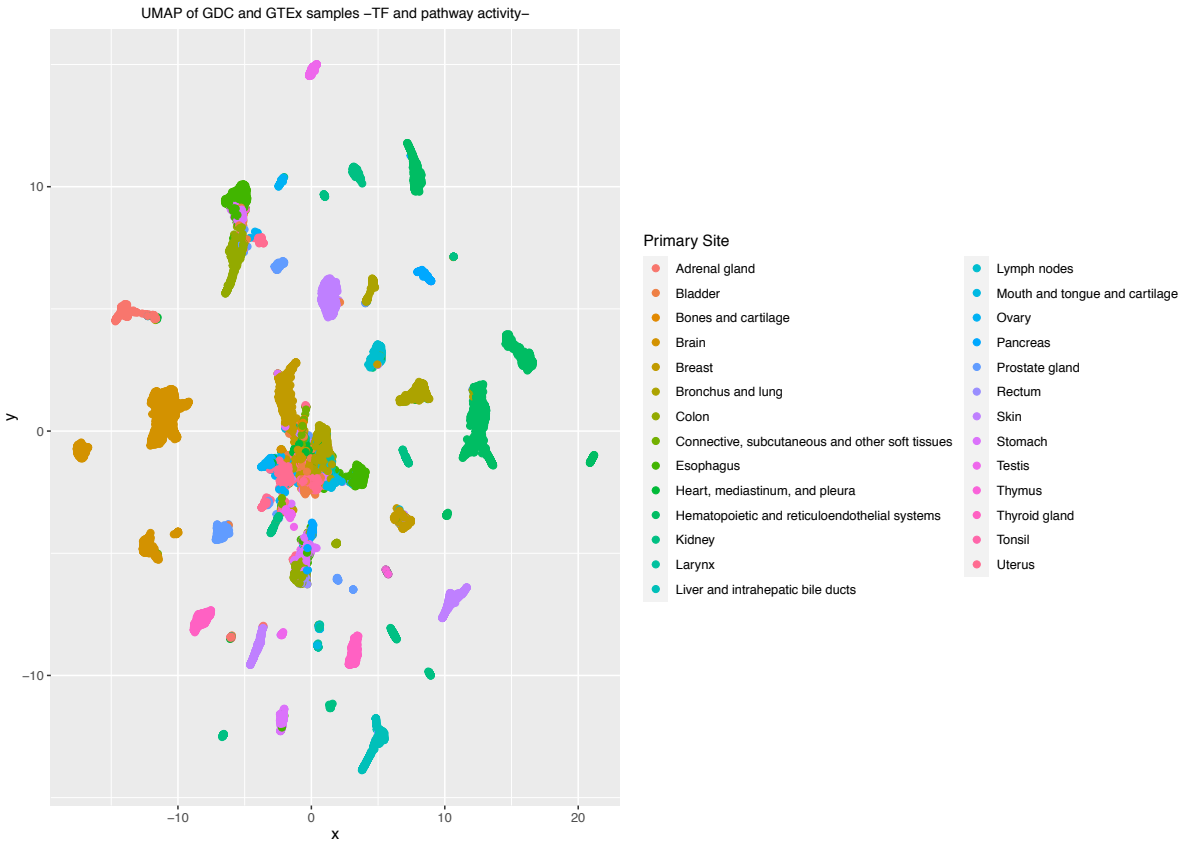


***Figure 22**: UMAP applied over the first 30 components of the PCA over the transcription factor and pathway activities. Colors correspond to each one of the 27 primary sites that the model is trained on.*

4.2 Primary site prediction with Deep Learning

After training several models with different setup of hidden layers, we come up with the best performing models, for both the convolutional neural network and the feedforward neural network. Once the model has been trained, we test the model with the validation samples, that account for the 30% of the total dataset in use, in the case of the GDC and GTEx dataset the validation consists of 7800 samples, and we construct a confusion matrix, which give us the accuracy per primary site and the global accuracy of the model.

4.2.1 Convolutional neural network accuracy

The model built using the bioimage with transcription factors and cancer driver genes reached a global accuracy on the prediction of the primary site over the validation data of 97%. Among the best primary site classification accuracies are "*Breast*" with 98,7%, "*Prostate gland*" with 99,7%, "*Testis"* with 100%, "*Hematopoietic and reticuloendothelial systems"* with 99,9%, "*Brain*" with 99,5%, "*Kidney*" with 99,3%.

Interestingly, the primary site tissues that performed the worst were mislabeled as closely related organs, physically close tissues, or organs with similar cell composition. "*Larynx*" has an accuracy of 53,4% and was labeled as "*Bronchus and lung"* 17,2% of the times, and as "*Mouth, tongue and cartilage*" 25,9% of the samples. Another important example is "*Rectum*" which has an accuracy of just 23,1% and was incorrectly mislabeled as "*Colon*" the 73,1% of the times, this example in particular is very representative, as it showcases the difficult task of finding key biomolecular differences between colon and rectum cancers. Finally, "*Heart, mediastinum and pleura*" has an accuracy of 60%, and is incorrectly predicted as "*Thymus*" the 28,6% of times (the thymus is a specialized primary lymphoid organ of the immune system, located in the upper front part of the chest, in the anterior superior mediastinum, behind the sternum, and in front of the heart).

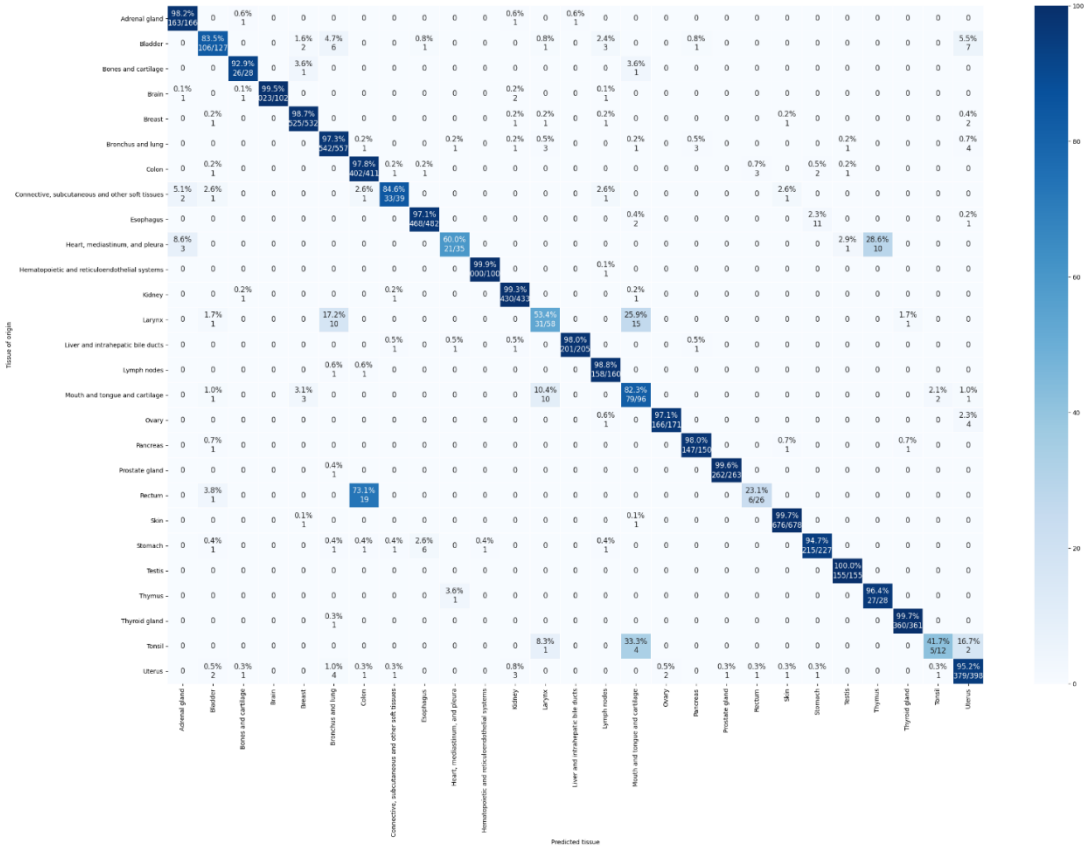


***Figure 23**: Confusion matrix of the convolutional model prediction on primary sites on the validation samples. The x axis corresponds with the predicted tissue, the y axis corresponds with the original observed tissue of the sample. There are a total of 27 tissues. The diagonal of the matrix is the exact match between the prediction and the original tissue, the more matches the best accuracy and the darkest color.*

4.2.2 Feedforward neural network based on bioactivity accuracy

The model with lower complexity that uses the Transcription Factor activities derived from DoRothEA regulons and the Pathway activities computed by Progeny reached a global accuracy on the prediction of the primary site tissue for the validation samples of 96%. Although the accuracy seems to be similar to the one obtained by the convolutional model, several key differences exist in this feedforward model the variability among tissue accuracies is much higher, with some reaching an accuracy higher than the convolutional model, meanwhile several other tissues are incorrectly labeled more than 50% of the times.

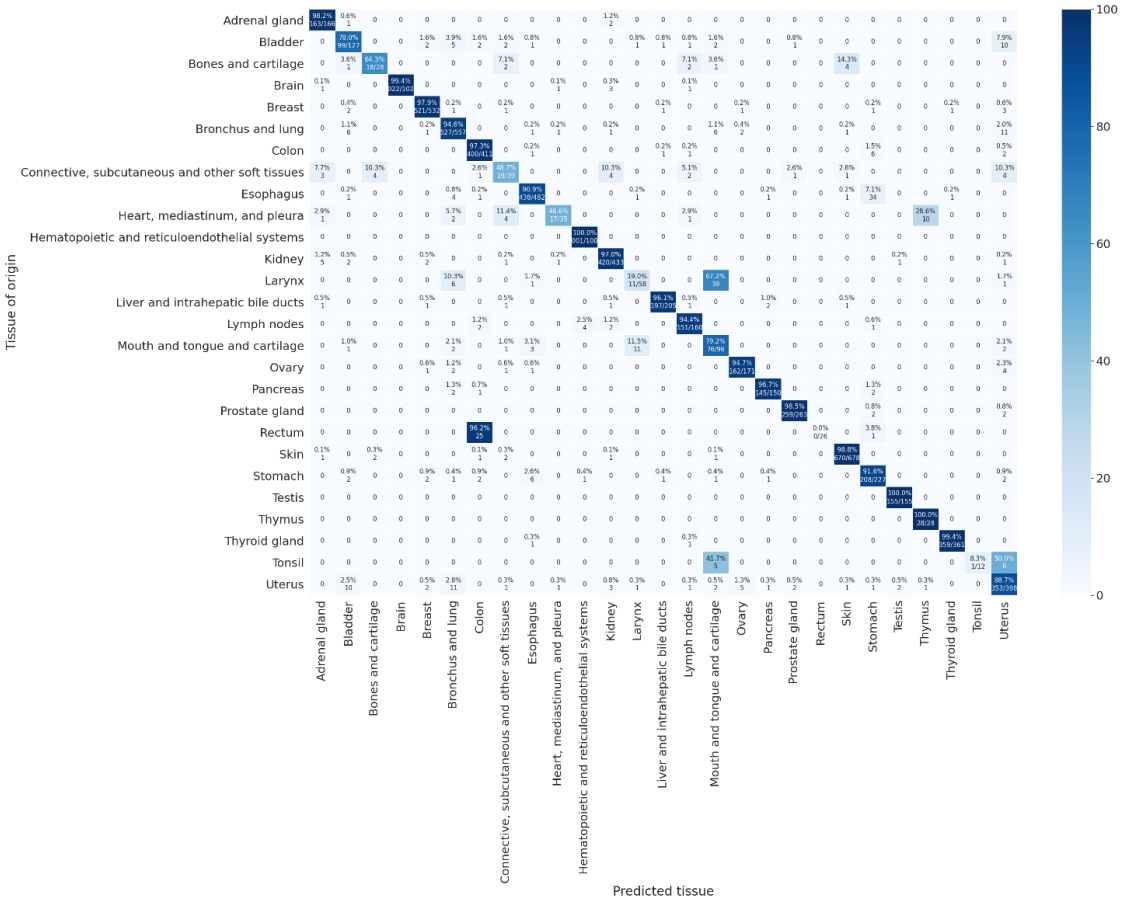


***Figure 24**: Confusion matrix of the feedforward bioactivity model prediction on primary sites on the validation samples. The x axis corresponds with the predicted tissue, the y axis corresponds with the original observed tissue of the sample. There are a total of 27 tissues. The diagonal of the matrix is the exact match between the prediction and the original tissue, the more matches the best accuracy and the darkest color.*

The global accuracy of 96% can be explained because the tissues with the major number of samples are generally very well classified.

The tissues that perform better than the convolutional model are "*Hematopoietic and reticuloendothelial systems"* with a 100% of accuracy, and "*Thymus*" with a 100% of accuracy as well. Other top performance primary sites are "*Testis*" with a 100% accuracy, "*Kidney*" with a score of 97%, "*Brain*" with 99,4%, "*Adrenal gland*" with 98,2% and "*Breast*" with 97,9% of accuracy. As in the convolutional deep learning model, some primary sites are mislabeled as closely related tissues. Especially important is the case of "*Rectum*" with a 0% of accuracy, 96.2% of the rectum samples were classified as "*Colon*".

4.3 Validation with external datasets

In order to correctly validate our deep learning models, we need to apply them to external RNA-Seq datases, RNA-Seq data, as almost all omic data, suffers from strong batch effects, technical noise, different technological platforms for sequencing, and clinical noise (conditions of the experiment, biopsy, hospital, etc.). Both the TCGA-GDC and the GTEx consortium are very popular omic datasets, since they are a collection of high-quality datasets, manually curated (despite coming from different studies, countries and researchers) reanalyzed with a standardized pipeline in order to make samples across datasets more comparable. Both datasets, specially TCGA-GDC (which contains a large collection of cancer samples) have been recognized as a huge influence (Gao et al., 2019). Because of this, we need to test our models on external datasets, and expose the deep learning algorithms to different environments, as they are prone to overfitting, and analyze if the generalization can be achieved.

The first dataset tested corresponds to the GEO dataset GSE146009 of RNA-Seq colon cancer with samples from the primary site tumor, the dataset is composed by two sub-datasets: primary colon cancer tumor from Caucasian American individuals and primary colon cancer tumor from African American individuals. It has been shown that misrepresentation of ethnicities on cancer studies is widely extended and represent a real and huge problem, since several studies during the last decade have shown that race/ethnicity have a great impact on cancer incidence, survival, drug response, and other important biomolecular aspects (Guerrero et al., 2018). Because of this, we believe that is important to add multi-ethnic samples to tackle this bias problem.

Primary Site – Bulk RNA				
DATASET	**PRIMARY SITE**	**SAMPLES**	**CONVOLUTIONAL ACCURACY**	**BIOACTIVITY ACCURACY**
GSE146009	**Colon – primary site**	35 Caucasian American	97%	28%
GSE146009	**Colon – primary site**	30 African American	93%	2%

***Table 6**: Prediction results over colon cancer with primary site tumor RNA-Seq.*

As it can be seen in **Table 6** the convolutional model has an outstanding accuracy, meanwhile the feedforward DLNN trained with the transcription factor and pathway activity fails to generalize over the two datasets.

Our main target of these tools is the deciphering of the primary site of Cancers of Unknown Primary (CUP), for this we subsequently tested the deep learning models over 3 external metastatic datasets that include distant metastatic samples. The first 2 are presented in **Table 7**, they consist of RNA-Seq from Kidney cancer (GSE157256) with primary tumor, matched controls and distant metastasis, and Ovary Cancer (GSE133296) with distant metastatic tumor samples. As our main focus is the predictive power on distant metastasis, we decided to test the deep learning models with the metastatic samples only, in this way we are able to obtain strong conclusion about the feasibility of the CUP diagnostic tool.

Metastatic – Bulk RNA				
DATASET	**PRIMARY SITE**	**SAMPLES TESTED**	**CONVOLUTIONAL ACCURACY**	**BIOACTIVITY ACCURACY**
GSE157256	**Renal Cancer HLRCC with distant metastasis**	**16 metastatic tumors**.	92%	96%
GSE133296	**Ovarian cancer: omental metastases, and non-omental metastases**	30 metastatic – from 10 patients	93%	0%

***Table 7**: Prediction results over Kidney and Ovarian cancer with distant metastatic RNA-Seq samples.*

As can be derived from **Table 7**, the convolutional deep learning model presents a strong and stable accuracy through different tissues (92% for Kidney metastatic cancer and 93% for Ovary metastatic cancer), more importantly, the fact that the deep learning model is able to generalize from local tissue samples (most of GDC and GTEx samples are from the primary tumor/local tissue) to distant metastatic samples (as a reminder, we selected only the distant metastatic samples for these 2 tested studies) aiming at the correct primary site is a strong indicator that this model could be improved and escalated to be tested in real clinical scenarios.

On the other hand, the much simpler and lighter deep learning model based on feedforward neural networks and having as input the transcription factor and pathway activity shows a mixed performance: score with Kidney metastatic dataset reaches a 96% of accuracy, that could be explained because kidney activity is more informative than in other tissues (see **Figure 22**), as could be seen in the bioactivity data exploration with PCA and UMAP (see chapter 3, section 4.1 Exploring the biological activity information).

Unfortunately, the model fails to generalize over the ovarian metastatic dataset, with an accuracy of 0%, meaning the bioactivity lacks information for segregating the ovarian tissue from the other primary sites (see **Figure 22**), as suggested by the PCA and UMAP analysis, in which the ovary samples appear to be mixed with several other different tissues (see chapter 3, section 4.1 Exploring the biological activity information). Despite the varying results of this model, is worth to highlight that some tissues show promising results, with even a higher accuracy that the convolutional model, which lead us to conclude that several transcription factor and pathway activities, and their respective regulons or pathway-composing genes, should be investigated in depth, so the convolutional model could be improved by including in its repertory these high informative biological activities.

Lastly, using the convolutional model we tested an additional RNA-Seq dataset of non-small cell lung cancer and lung metastatic carcinoma (GSE162945): in this study, paired tumor samples were collected from 10 patients with non-small cell lung cancer (NSCLC) or lung metastatic carcinoma within a week before and after SBRT (Stereotactic Body Radiation Therapy). Results show that in this case, the convolutional model is able to match some of the primary sites correctly, other metastatic samples are labeled as their site of resection of the biopsy: colorectal cancer with “Colon” and Cervical cancer with “Connective, subcutanneous and other soft tissues”, the lack of a specific label for cervical cancer let us conclude that “Connective, subcutanneous and other soft tissues” is appropriated in this context, since a biopsy of this particular site would be composed of those elements.

Sample	Cancer primary site	Cancer Biopsy tissue	Convolutional DNN Prediction	Match
Case10_A	Bronchus and lung	colorectal cancer	Colon	**Byopsy**
Case10_B	Bronchus and lung	colorectal cancer	Colon	**Byopsy**
Case11_A	Bronchus and lung	non-small-cell lung cancer	Lymph Nodes	**X**
Case11_B	Bronchus and lung	non-small-cell lung cancer	Bronchus and lung	**Match**
Case2_A	Bronchus and lung	Cervical cancer	Connective, subcutanneous and other soft tissues	**Byopsy**
Case2_B	Bronchus and lung	Cervical cancer	Connective, subcutanneous and other soft tissues	**Byopsy**
Case3_A	Bronchus and lung	non-small-cell lung cancer	Stomach	**X**
Case3_B	Bronchus and lung	non-small-cell lung cancer	Bronchus and lung	**Match**
Case4_A	Bronchus and lung	colorectal cancer	Breast	**X**
Case4_B	Bronchus and lung	colorectal cancer	Stomach	**Related**
Case5_A	Bronchus and lung	non-small-cell lung cancer	Stomach	**X**
Case5_B	Bronchus and lung	non-small-cell lung cancer	Bronchus and lung	**Match**
Case6_A	Bronchus and lung	colorectal cancer	Stomach	**Related**
Case6_B	Bronchus and lung	colorectal cancer	Stomach	**Related**
Case7_A	Bronchus and lung	non-small-cell lung cancer	Kidney	**X**
Case7_B	Bronchus and lung	non-small-cell lung cancer	Lymph Nodes	**X**
Case9_A	Bronchus and lung	non-small-cell lung cancer	Bronchus and lung	**Match**
Case9_B	Bronchus and lung	non-small-cell lung cancer	Bronchus and lung	**Match**

***Table 8**: Prediction results for the lung cancer dataset GSE162945, which contains both primary tumor and metastatic tumor RNA-Seq.*

We marked the mismatch of colorectal cancer as "*Stomach*" as a related conclusion, since as it could be seen on the validation confusion matrix (see **Figure 23**) the mislabeling of colon as stomach or other parts of the digestive system is known to happen. The distinction between this cases and other mismatches is clear: the model aims at a closely related tissue/organ, which could be studied in detail and with additional data, e.g.: the second and third most probable tissues of the prediction for the same sample by the DLNN, and other biomolecular information, pathway deregulation etc., the prediction could

be corrected for the correct primary site, or at least give enough information to the clinician to narrow the possible targets and the diagnosis.

4.4 Genes, TF and Pathways selected by permutation analysis as crucial features

Using a random permutation analysis (see chapter 3, section 3.4 Explainable Machine Learning) we were able to obtain a list of genes, Transcription factors activity and Pathways activity that are particularly important for the success of the models, in some cases the weight of the feature for the decision of the model for a particular primary site is so huge, that when substituting this feature's information by noise or extreme outliers the model prediction drops to 0.

Apart from the evident interest of these biological markers for the understanding of the biology underneath the distinction of the distant metastasis and their primary tumors, or the tissue specificity of some genes and pathway activities, this permutation analysis is important for filtering out those features that do not add information to the model, or for helping to select other features that at first could appear to not be relevant at all (because of insufficient scientific literature supporting evidence for that gene or feature to be related with the problem, in this case the marker of the primary site tumor in the metastatic sample).

4.4.1 Feature selection on convolutional DLNN

In the below **Table 9** we present a list with some of the most representative genes (only the ones that trigger an accuracy<0 are shown) and their respective primary site tissues: the top 3 genes that, when independently set to 0 CPM, have a major impact on the model performance. The statistics are calculated as follow:

$$\boldsymbol{Sample\ mismatches_i} = \#\{true\ positives\ for\ tissue_i\ from\ original\ model\} - \#\{true\ positives\ for\ tissue_i\ from\ modified\ model\}$$

$$\boldsymbol{Accuracy\ tissue_i} = \frac{\#\{true\ positives\ for\ tissue_i\ from\ modified\ model\}}{\#\{samples\ of\ tissue_i\}}$$

Gene set to 0 CPM	Gene symbol	Sample Mismatches	Primary site	Accuracy achieved
ENSG00000143578	CREB3L4	525	Breast	0
ENSG00000143614	GATAD2B	525	Breast	0
ENSG00000143622	RIT1	525	Breast	0
ENSG00000142599	RERE	468	Esophagus	0
ENSG00000142611	PRDM16	468	Esophagus	0
ENSG00000142627	EPHA2	468	Esophagus	0
ENSG00000151612	ZNF827	676	Skin	0
ENSG00000151615	POU4F2	676	Skin	0
ENSG00000151623	NR3C2	676	Skin	0
ENSG00000167766	ZNF83	147	Pancreas	0
ENSG00000167771	RCOR2	147	Pancreas	0
ENSG00000167785	ZNF558	147	Pancreas	0
ENSG00000175387	SMAD2	1018	Brain	0.004
ENSG00000175691	ZNF77	1017	Brain	0.005
ENSG00000175395	ZNF25	1016	Brain	0.006

***Table 9**: Results of the random permutation analysis of the convolutional neural network using 7800 validation samples. Each gene is set to 0 CPM on every validation samples, then the accuracy of the prediction is analyzed.*

On breast tissue, CREB3L4, GATAD2B and RIT1 were found to be significantly important for the outcome of the prediction. CREB3L4 is a member of the CREB/ATF transcription factor family, which regulates various processes including cell proliferation, differentiation and apoptosis, by regulating gene expression through the cAMP-responsive element (Velpula et al., 2012) Recently, a paper reported that CREB3L4 was co-

upregulated in mRNA expression with MUC1 in breast cancer, the authors suggested that both genes may serve as a potential prognostic factor and therapy target for breast cancer (Jing et al., 2019). GATAD2B (GATA Zinc Finger Domain Containing 2B) is a protein coding gene known to be involved in chromatin modification and transcriptional control (Denslow & Wade, 2007), and is known to promote KRAS activity and acting as a metastasis driver in KRAS-driven lung cancer (Grzeskowiak et al., 2018), but as the best of our knowledge it has not been related to neither breast tissue not to breast cancer. RIT1 (Ras Like Without CAAX 1) is related to the Ras-MAPK signaling cascade that mediates a wide variety of cellular functions, including cell proliferation, survival, and differentiation (G.-X. Shi & Andres, 2005), and has been prominently associated with Noonan Syndrome (Aoki et al., 2013; Gos et al., 2014), but no references in the scientific literature relate RIT1 with breast biology.

Another important primary site tissue: skin, which commonly derives in metastasis, points to the importance of 3 genes that are crucial to its prediction: ZNF827, POU4F2, NR3C2. ZNF827 (Zinc Finger Protein 827) is involved in transcriptional regulation, and a recent study has suggested ZNF827 as a novel putative markers for cutaneous melanoma metastasis (Bhalla et al., 2019). POU4F2 (POU class 4 homeobox 2) is a protein coding gene of the POU-domain transcription factor family. NR3C2 (Nuclear Receptor Subfamily 3 Group C Member 2) the protein encoded by this gene acts as a receptor for mineralocorticoids (MC) and glucocorticoids (GC) (such as corticosterone or cortisol). The effect of MC is to increase ion and water transport and thus raise extracellular fluid volume and blood pressure and lower potassium levels (Arriza et al., 1987).

Lastly, we focus on pancreas primary site: pancreatic cancer is known to be extremely difficult to diagnose early, which makes it susceptible to develop metastasis. We found 3 key genes for the correct prediction of this primary site: ZNF83, RCOR2, ZNF558. ZNF83 (Zinc Finger Protein 83) may be involved in transcriptional regulation. RCOR2 (REST Corepressor 2) is a protein coding gene: Gene Ontology (GO) annotations related to this gene include DNA-binding transcription factor activity and transcription corepressor activity. Additionally, RCOR2 plays important roles in regulating ESCs (Embryonic Stem Cells) pluripotency and reprogramming somatic cells to pluripotency (Yang et al., 2011).

When inspecting the top 10 genes per primary site tissue that had the most impact on the performance of the model, we searched for genes whose weight in the inference decision were crucial in more than 1 primary site (genes whose information is not tissue-specific but rather multi tissue deterministic). Through the analysis of the results of the random permutation test we were able to select 10 genes (see **Table 11**) that appear to be relevant to several tissues.

Gene set to 0 CPM	Gene symbol	N of appearances	Primary sites affected
ENSG00000166710	B2M	4	"Hematopoietic and reticuloendothelial systems", "Ovary", "Thymus", "Uterus"
ENSG00000166823	MESP1	4	"Hematopoietic and reticuloendothelial systems", "Ovary", "Thymus", "Uterus"
ENSG00000166888	STAT6	3	"Hematopoietic and reticuloendothelial systems", "Ovary", "Thymus"
ENSG00000166925	TSC22D4	3	"Hematopoietic and reticuloendothelial systems", "Ovary", "Thymus"
ENSG00000167034	NKX3-1	3	"Hematopoietic and reticuloendothelial systems", "Prostate gland", "Thymus"
ENSG00000167074	TEF	3	"Hematopoietic and reticuloendothelial systems", "Prostate gland", "Thymus"
ENSG00000175727	MLXIP	3	"Brain", "Colon", "Stomach"
ENSG00000175745	NR2F1	3	"Brain", "Colon", "Stomach"
ENSG00000143842	SOX13	2	"Breast", "Thyroid gland"
ENSG00000143867	OSR1	2	"Breast", "Thyroid gland"

***Table 10**: genes selected by random permutation analysis whose impact on model performance affects more than 1 specific tissue. These genes are potential multi-tissue specific targets.*

Both B2M and MESP1 genes are important for the correct identification of the primary site of the sample on 4 tissues: *"Hematopoietic and reticuloendothelial systems", "Ovary", "Thymus", and "Uterus"*. B2M (β2 microglobulin) is a component of the major histocompatibility complex (MHC) class I heavy chain on the surface of nearly all nucleated cells and is involved in the presentation of peptide antigens to the immune system. A mutation in this gene is known to cause hypercatabolic hypoproteinemia (Wani et al., 2006), and because of its tissue ubiquity and its important role in the immune system, is suspected to be involved in multiple processes or act as biomarker in several cancers such as colon (Blum et al., 2008), prostate (Abdul & Hoosein, 2000), stomach (Rho et al., 2010)

and lung (Gettinger et al., 2017). MESP1 (Mesoderm posterior protein 1) is known to play a role in the epithelialization of somitic mesoderm and in the development of cardiac mesoderm (Bondue et al., 2008). Recently, MESP1 was associated with non-small cell lung cancer (NSCLC), and gene expression was found to correlate with poor prognosis in NSCLC patients, furthermore MESP1 is critical for proliferation and survival of NSCLC-derived cells (Tandon et al., 2019). To the best of our knowledge, there have not been any additional reports of MESP1 being involved with other tissues or cancers.

Our analysis suggests that STAT6 and TSC22D4 genes are important for the correct prediction of 3 primary sites: “Hematopoietic and reticuloendothelial systems”, “Ovary”, and “Thymus”. STAT6 (Signal Transducer And Activator Of Transcription 6) is a member of the STAT family of transcription factors In response to cytokines and growth factors, STAT family members are phosphorylated by the receptor associated kinases and act as transcription activators. This protein plays a central role in exerting IL4 mediated biological responses. It is found to induce the expression of BCL2L1/BCL-X(L), which is responsible for the anti-apoptotic activity of IL4 (Takeda et al., 1996). STAT6 has been related with several aspects of cancer biology: its expression is correlated with colon cancer stage and prognosis (C. G. Wang et al., 2010), in breast cancer the activation of STAT6-TP63 pathway suppress lung metastasis (Papageorgis et al., 2015), other study found that miR-135b inhibits tumor metastasis in prostate cancer by targeting STAT6 (N. Wang et al., 2016). TSC22D4 (TSC22 Domain Family Member 4) is related with hepatic metabolism (Jones et al., 2013) and diabetic hyperglycaemia and insulin resistance (Ekim Üstünel et al., 2016), but to the best of our knowledge it has not been associated with cancer processes or other tissue specificity.

4.4.2 Feature selection on bioactivity DLNN model

Using the random permutation analysis with the Transcription Factor and Pathway activity DLNN model over the validation data of the GDC and GTEx dataset, we were able to determine some important TF or Pathway activities that were essential for the model in order to correctly infer the primary site of the sample (**Figure 25**). Since this model did not reach a correct minimum accuracy in some primary sites, we decided to test the permutation analysis on tissues that achieved at least an 80% of accuracy over the validation data (see **Figure 24**). Here we discuss some of the most relevant results.

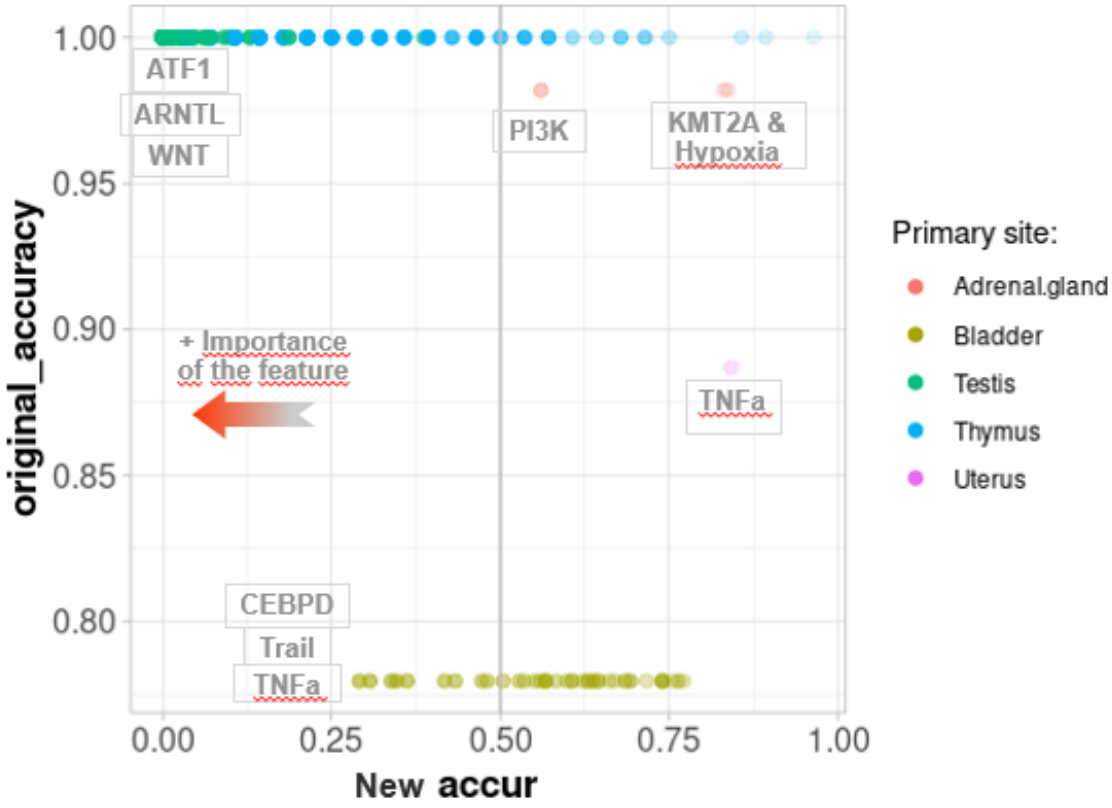


***Figure 25**: Permutation random analysis of DLNN of transcription factor and pathway activity. The y axis corresponds with the original accuracy achieved by the model on the validation data over the primary sites shown on the right labels. The x axis shows the new accuracy achieve by the model using a randomly permuting on a TF or Pathway activity (each point corresponds with a TF or a Pathway activity on a specific primary site).*

Numerous transcription factors and pathways known to be related with cancer biology are identified by the model. Interestingly, TNFa was associated to "*Uterus*" and "*Bladder*" primary sites by our model, TNFa (Tumor Necrosis Factor α) is a gene that encodes a multifunctional proinflammatory cytokine that belongs to the tumor necrosis factor (TNF) superfamily, this cytokine is mainly secreted by macrophages. Different studies have analyzed the production of this cytokine in the mouse uterus (Mamata De et al., 1992; Soboll et al., 2006). Also, high expression of TNFa cytokine has been seen on patients with bladder cancer (de Reijke et al., 1993; Raziuddin et al., 1993).

The model related both "*testis*" and "*thymus*" primary sites with 3 genes: ATF1, ARNTL, and WNT. Gene ATF1 (Activating Transcription Factor 1) influences cellular physiologic processes by regulating the expression of downstream target genes, which are

related to growth, survival, and other cellular activities. This transcription factor has been related with numerous processes on cancer biology, more precisely in colon cancer (Tian et al., 2019), hepatic carcinoma (Ding et al., 2017), and cervical cancer (Y. Shi et al., 2017; Xu et al., 2019). A paper reported that the loss of ATF1 results in reduced thymic cellularity and delayed thymic recovery in mice (Baumann et al., 2004), but not connection with testis tissue has been seen as for the present date.

ARNTL gene (Aryl Hydrocarbon Receptor Nuclear Translocator Like) is a transcriptional activator which forms a core component of the circadian clock. Interestingly, it has been found that the thymus and the testis have similar pattern of circadian clock gene expression (Alvarez & Sehgal, 2005) which could explain this random permutation results analysis on our model.

The activity of the WNT signaling pathway also was marked as indispensable for the correct characterization of testis and thymus primary sites: WNT signaling is one of the central mechanisms regulating tissue morphogenesis during embryogenesis and repair. This signaling pathway has been related with the regulation of spermatogonia and testicular tumor (Dong et al., 2015; Garcia-Moreno et al., 2019). TRAIL, the pathway involving TNF-related apoptosis-inducing ligand protein family (like the already mentioned TNFa), preferentially induces apoptosis in transformed and tumor cells, but does not appear to kill normal cells although it is expressed at a significant level in most normal tissues (Johnstone et al., 2008).

As trying to understand better the feasibility of predicting primary sites from metastatic tumors was one of our main goals, we wanted to analyze which of the transcription factor and pathway activities were essential for the correct prediction of the primary site on distant metastasis samples, for this we used the renal cancer HLRCC with distant metastasis dataset (GSE157256). By using the random permutation analysis, we were able to identify 3 genes in particular that were key for the DLNN model in order to classify distant renal metastasis as prevenient from kidney primary site.

TF	Tissue	Average Accuracy with random permutations
HIF1A	Kidney	63%
MYC	Kidney	33%
SOX2	Kidney	33%

***Table 11**: features selected by random permutation analysis on the TF and Pathway activity DLNN model over the renal cancer HLRCC with distant metastasis dataset.*

HIF1A (Hypoxia Inducible Factor 1 Subunit Alpha) is a transcription factor and master transcriptional regulator of cellular and developmental response to hypoxia (G. L. Wang et al., 1995), its rol in several types of cancer, including renal carcinoma, is well known (Morris et al., 2009; Ollerenshaw et al., 2004). MYC (proto-oncogene, bHLH transcription factor) the protein encoded by this gene is a nuclear phosphoprotein that plays a role in cell cycle progression, apoptosis, and cellular transformation. The relationship between this gene and numerous human cancers has been demonstrated, contributing to its direct tumorigenesis (Dang, 2012; Stine et al., 2015), including kidney cancer (S. T. Bailey et al., 2017; Shroff et al., 2015; Tang et al., 2009).

SOX2, also known as SRY (sex determining region Y)-box 2, is an intronless gene involved in the regulation of embryonic development and in the determination of cell fate. A paper studied the predictive potential of SOX2 of prognosis in patients with clear cell renal cell carcinoma (Gu et al., 2018), but no more relationship between this gene and kidney has been reported. Because of this reason, we wanted to determine why and at what extent SOX2 is important for the predictive power of the DLNN model over distant metastasis on kidney.

For this, we performed another random permutation analysis with additional SOX2 activity random scores, and we calculated the accuracy for each of the generated random values. The results of this analysis clearly show the existence of a sensitivity threshold when the SOX2 transcription factor activity reaches a 15.5 score, and that activity below this threshold prompts the DLNN model to unambiguously label the distant metastasis of kidney as effectively coming from kidney primary site. From this result, it can be

hypothesized that *kidney metastasis is related with a low activity of the transcription factor SOX2, and thus could act as a biomarker of renal cancer in distant metastasis.*

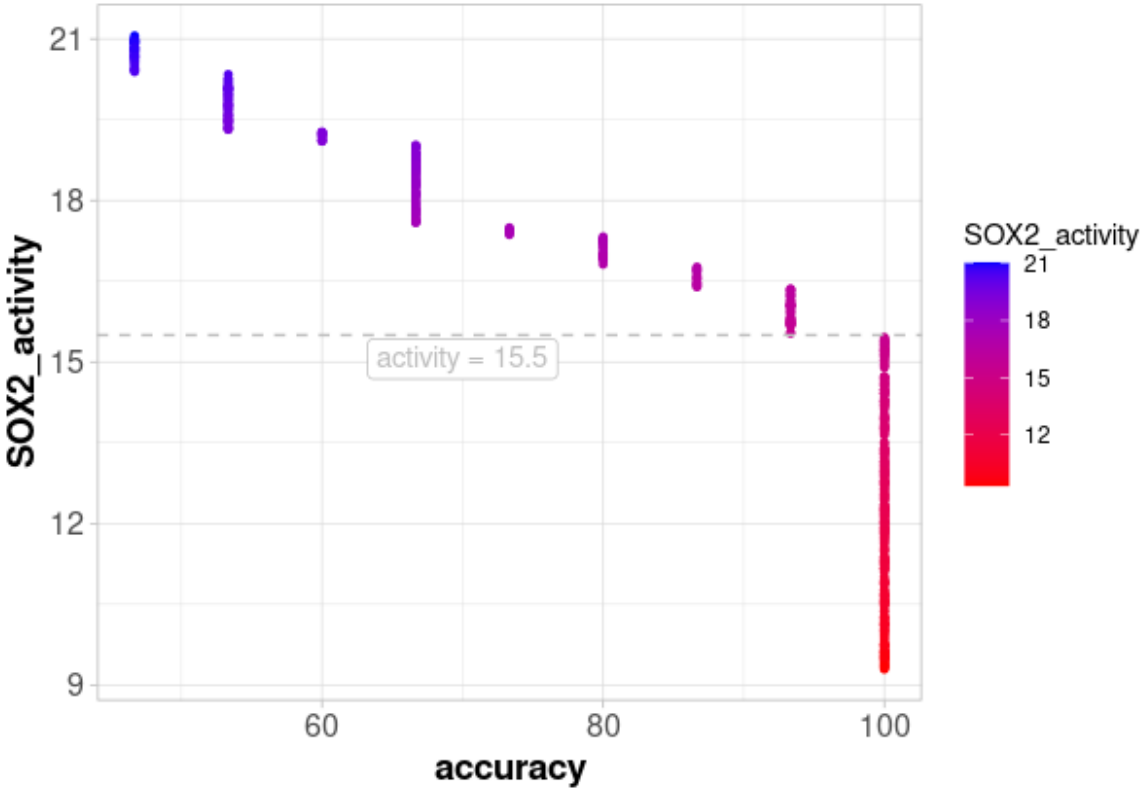


***Figure 26**: plot that shows the performance of the DLNN model with SOX2 activity random permutations. The y axis corresponds with the randomly generated SOX2 activity scores, x axis shows the model accuracy for labeling the primary site samples as "kidney" having the corresponding SOX2 permutated value.*

All these results demostrate the potencial of feature selection by interrogating the deep learning network regarding the weights in the decision making of the model: as we have seen, a majority of the results are in consonance with the biology studied. It is also important to remark that this method could help us to filter and select features that are more relevant, helping to reduce the input noise and the dimensionality of the data, it also can serve as a selector for new features that are not very well undertood or are not mentioned in the scientific literature, but that it could play an important role in the model performance, or even an important rol in the biology that is being study.

5 CONCLUSION

Cancers of Unknown Primary (CUPs) poses a difficult and deadly challenge on the future of cancer treatment and diagnosis, as more advance and successful treatments are found, the most agresive and difficult to localize and stratifie types of cancers remain, and specifically the poor-diferenciated distant metastasis. Here we show the use of deep learning neural networks as a model predictor of the primary site of these metastatic cancers; importantly, despite the lack of clear and general biomolecular markers for all the cancer types and subtypes in poor diferentiated samples, we show that it is possible to train and use effectively an AI model to infer the primary site. We show the importance of the transcription factors and literature-curated cancer genes, and also the viability and importance of patwhay and transcription factor activity (computed using the tools DoRothEA and Progeny) in some of the primary site tissues. With accuracies of 0.97 and 0.96 for the convolutional NN model and the feedforward NN model respectively, we demostrate the feasibility of using these new technologies within the diagnostic and clinical field, and the promising revolution that AI could bring to personalized medicine, giving very specifical results for each patient based on huge amounts of data.

Also, we show the potential of these models as to better understand the biology and markers that are crucial for these kind of diagnosis, not acting just pasively infering results, but also producing new insights on the problem we are studing. Remarkable is the case of the distant metastatic datasets: as we show, the models are able to generalize when facing new and external data (the convolutional model much better than the activity model), even despite of that both models have been trained with large datasets of primary tissues samples, having almost none metastatic samples to learn from. Particularly interesting is the case of the SOX2 transcription factor activity random permutation analysis, that links kidney metastasis with a low activity of SOX2.

6 GENERAL CONCLUSIONS

Throughout the chapters of this Doctoral Thesis, we have described and explained in detail several bioinformatic approaches to three different biological problems, with the three having in common the complexity and multifactorial origin of the diseases (Alzheimer and cancer) or the biological problems (brain aging, neurodegeneration, cancer origin and metastasis) that we have interrogated and analyzed.

Furthermore, we have successfully shown the value and feasibility of using artificial intelligence (AI), and more specifically deep learning neural networks (DNN), with large-scale omic data, to interrogate these complex questions. These methodologies are taking a leading role in current Personalized Medicine, not only to infer useful results, but also to generate new comprehensive insights that help to understand the complex biological problems in question. Because all of this, as a final summary of this PhD Dissertation, we propose the following **General Conclusions**:

1. We have successfully **collected and analyzed a large compendium of human brain samples**, representing individuals across all ages, from infancy to elderly people. We have successfully used **bioinformatic and statistical methods to unravel the changes that occur in the human brain with the progression of age**, shedding light into the complex process of aging at the transcriptome and biomolecular level.

2. We have **developed a deep learning neural network (DLNN)** that is able to calculate the **biological-age** of a given individual from an expression transcriptomic profile. We have shown that this **bioage model perform better than current existing models**, even than the most accurate **bioage clocks** that are based in **epigenetic features**. We have created an R package to use our algorithm which is open access (freely available in GitHub on the following URL: *https://github.com/jdelasrivas-lab/RdeeplearningNN-bioage*).

3. We have collected and normalized a **large compendium of transcriptomic Alzheimer's patients and matched control samples** from **cortex and hippocampus brain regions**. We have obtained a robust and distinctive **gene signature of the AD molecular expression profile** for these regions, which describes **key molecular changes in the AD brain** and is able to **stratify individuals** by condition. We have been able to compare the **pathological signature of AD** with the **signature of healthy brain aging**, finding important **differences between these two processes**.

4. We have analyzed **blood samples** from individuals with **Alzheimer's disease** within the framework and objectives of the **European Project *ArrestAD*** and in collaboration with six international research groups experts in the field. This work is in progress and under confidential access, but with our results we have successfully **identify risk factors and features of a specific new phenotype found in blood cells linked to AD**. We also expect that this work will help to identify new biomarkers of **late-onset Alzheimer's disease (LO-AD)** in the early stages of this neurodegenerative pathology.

5. We have successfully developed a **deep learning neural network (DLNN) model that effectively infers the primary site of a tumor sample using RNA-Seq data**, achieving an average accuracy of 97%. We have shown that this **predictive model** is able to generalize on external datasets and demonstrated that it **achieves high precision in cancer samples corresponding to distant metastases.** This work shows the great potential of using **Artificial Intelligence (AI)** technologies in real clinical scenarios.

6. We have built and analyze **gene networks based on pathway information and on transcription factor (TF) activity using deep learning models**. Furthermore, we have gained insights into **the value of these gene networks for stratification of primary cancer sites** and, more importantly, we have demonstrated the feasibility of these models used as a way to **better understand the key biological components that link metastasis and its corresponding primary tumors**.

BIBLIOGRAPHY

Abdul, M., & Hoosein, N. (2000). Changes in beta-2 microglobulin expression in prostate cancer. *Urologic Oncology*, *5*(4), 168–172. https://doi.org/10.1016/S1078-1439(00)00063-6

Aibar, S., Fontanillo, C., Droste, C., & De Las Rivas, J. (2015). Functional Gene Networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering. *Bioinformatics*, *31*(10), 1686–1688. https://doi.org/10.1093/bioinformatics/btu864

Aloysius, N., & Geetha, M. (2018). A review on deep convolutional neural networks. *Proceedings of the 2017 IEEE International Conference on Communication and Signal Processing, ICCSP 2017*, *2018-January*, 588–592. https://doi.org/10.1109/ICCSP.2017.8286426

Alvarez, J. D., & Sehgal, A. (2005). The thymus is similar to the testis in its pattern of circadian clock gene expression. *Journal of Biological Rhythms*, *20*(2), 111–121. https://doi.org/10.1177/0748730404274078

Aoki, Y., Niihori, T., Banjo, T., Okamoto, N., Mizuno, S., Kurosawa, K., Ogata, T., Takada, F., Yano, M., Ando, T., Hoshika, T., Barnett, C., Ohashi, H., Kawame, H., Hasegawa, T., Okutani, T., Nagashima, T., Hasegawa, S., Funayama, R., … Matsubara, Y. (2013). Gain-of-function mutations in RIT1 cause noonan syndrome, a RAS/MAPK pathway syndrome. *American Journal of Human Genetics*, *93*(1), 173–180. https://doi.org/10.1016/j.ajhg.2013.05.021

Ardlie, K. G., DeLuca, D. S., Segrè, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., Trowbridge, C. A., Maller, J. B., Tukiainen, T., Lek, M., Ward, L. D., Kheradpour, P., Iriarte, B., Meng, Y., Palmer, C. D., Esko, T., Winckler, W., Hirschhorn, J. N., Kellis, M., … Lockhart. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. https://doi.org/10.1126/science.1262110

Arriza, J. L., Weinberger, C., Cerelli, G., Glaser, T. M., Handelin, B. L., Housman, D. E., & Evans, R. M. (1987). Cloning of human mineralocorticoid receptor complementary DNA: Structural and functional kinship with the glucocorticoid receptor. *Science*, *237*(4812), 268–275. https://doi.org/10.1126/science.3037703

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K. S., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., … Karchin, R. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, *173*(2), 371-385.e18. https://doi.org/10.1016/j.cell.2018.02.060

Bailey, S. T., Smith, A. M., Kardos, J., Wobker, S. E., Wilson, H. L., Krishnan, B., Saito, R., Lee, H. J., Zhang, J., Eaton, S. C., Williams, L. A., Manocha, U., Peters, D. J., Pan, X., Carroll, T. J., Felsher, D. W., Walter, V., Zhang, Q., Parker, J. S., … Kim, W. Y. (2017). MYC activation cooperates with Vhl and Ink4a/Arf loss to induce clear cell renal cell carcinoma. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms15770

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Baumann, S., Kyewski, B., Bleckmann, S. C., Greiner, E., Rudolph, D., Schmid, W., Ramsay, R. G., Krammer, P. H., Schütz, G., & Mantamadiotis, T. (2004). CREB function is required for normal thymic cellularity and post-irradiation recovery. *European Journal of Immunology*, *34*(7), 1961–1971. https://doi.org/10.1002/eji.200324826

Berchtold, N. C., Coleman, P. D., Cribbs, D. H., Rogers, J., Gillen, D. L., & Cotman, C. W. (2013). Synaptic genes are extensively downregulated across multiple brain regions in normal human aging and Alzheimer's disease. *Neurobiology of Aging*, *34*(6), 1653–1661. https://doi.org/10.1016/j.neurobiolaging.2012.11.024

Berchtold, N. C., Sabbagh, M. N., Beach, T. G., Kim, R. C., Cribbs, D. H., & Cotman, C. W. (2014). Brain gene expression patterns differentiate mild cognitive impairment from normal aged and Alzheimer's disease. *Neurobiology of Aging*, *35*(9), 1961–1972. https://doi.org/10.1016/j.neurobiolaging.2014.03.031

Bhalla, S., Kaur, H., Dhall, A., & Raghava, G. P. S. (2019). Prediction and Analysis of Skin Cancer Progression using Genomics Profiles of Patients. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-52134-4

Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I. F., Mak, R. H., Tamimi, R. M., Tempany, C. M., Swanton, C., Hoffmann, U., Schwartz, L. H., Gillies, R. J., Huang, R. Y., & Aerts, H. J. W. L. (2019). Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA: A Cancer Journal for Clinicians*. https://doi.org/10.3322/caac.21552

Blasko, I., & Grubeck-Loebenstein, B. (2003). Role of the immune system in the pathogenesis, prevention and treatment of Alzheimer's disease. In *Drugs and Aging* (Vol. 20, Issue 2, pp. 101–113). https://doi.org/10.2165/00002512-200320020-00002

Blum, C., Graham, A., Yousefzadeh, M., Shrout, J., Benjamin, K., Krishna, M., Hoda, R., Hoda, R., Cole, D. J., Garrett-Mayer, E., Reed, C., Wallace, M., & Mitas, M. (2008). The expression ratio of Map7/B2M is prognostic for survival in patients with stage II colon cancer. *International Journal of Oncology*, *33*(3), 579–584. https://doi.org/10.3892/ijo_00000043

Bodenhofer, U., & Klawonn, F. (2008). Robust rank correlation coefficients on the basis of fuzzy. *Mathware & Soft Computing*, *15*(1), 5–20.

Bodenhofer, Ulrich, Krone, M., & Klawonn, F. (2013). Testing noisy numerical data for monotonic association. *Information Sciences*, *245*, 21–37. https://doi.org/10.1016/j.ins.2012.11.026

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bomane, A., Gonçalves, A., & Ballester, P. J. (2019). Paclitaxel Response Can Be Predicted With Interpretable Multi-Variate Classifiers Exploiting DNA-Methylation and miRNA Data. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.01041

Bondue, A., Lapouge, G., Paulissen, C., Semeraro, C., Iacovino, M., Kyba, M., & Blanpain, C. (2008). Mesp1 acts as a master regulator of multipotent cardiovascular progenitor specification. *Cell Stem Cell*, *3*(1), 69–84. https://doi.org/10.1016/j.stem.2008.06.009

Boudt, K., Cornelissen, J., & Croux, C. (2012). The Gaussian rank correlation estimator: Robustness properties. *Statistics and Computing*, *22*(2), 471–483. https://doi.org/10.1007/s11222-011-9237-0

Cáceres, M., Lachuer, J., Zapala, M. A., Redmond, J. C., Kudo, L., Geschwind, D. H., Lockhart, D. J., Preuss, T. M., & Barlow, C. (2003). Elevated gene expression levels distinguish human from non-human primate brains. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(22), 13030–13035. https://doi.org/10.1073/pnas.2135499100

Calderwood, S. K., Murshid, A., & Prince, T. (2009). The shock of aging: Molecular chaperones and the heat shock response in longevity and aging - A mini-review. In *Gerontology* (Vol. 55, Issue 5, pp. 550–558). https://doi.org/10.1159/000225957

Calvo, S. E., Clauser, K. R., & Mootha, V. K. (2016). MitoCarta2.0: An updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Research*, *44*(D1), D1251–D1257. https://doi.org/10.1093/nar/gkv1003

Campos-Laborie, F. J., Risueño, A., Ortiz-Estévez, M., Rosón-Burgo, B., Droste, C., Fontanillo, C., Loos, R., Sánchez-Santos, J. M., Trotter, M. W., & Rivas, J. D. Las. (2019). DECO: Decompose

heterogeneous population cohorts for patient stratification and discovery of sample biomarkers using omic data profiling. *Bioinformatics*, *35*(19), 3651–3662. https://doi.org/10.1093/bioinformatics/btz148

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*(7623), 20–23. https://doi.org/10.1038/538020a

Chen, K., & Williams, K. J. (2013). Molecular mediators for raft-dependent endocytosis of syndecan-1, a highly conserved, multifunctional receptor. *Journal of Biological Chemistry*, *288*(20), 13988–13999. https://doi.org/10.1074/jbc.M112.444737

Chiasserini, D., Biscetti, L., Eusebi, P., Salvadori, N., Frattini, G., Simoni, S., De Roeck, N., Tambasco, N., Stoops, E., Vanderstichele, H., Engelborghs, S., Mollenhauer, B., Calabresi, P., & Parnetti, L. (2017). Differential role of CSF fatty acid binding protein 3, α-synuclein, and Alzheimer's disease core biomarkers in Lewy body disorders and Alzheimer's dementia. *Alzheimer's Research and Therapy*, *9*(1). https://doi.org/10.1186/s13195-017-0276-4

Ching, T., Zhu, X., & Garmire, L. X. (2018). Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Computational Biology*, *14*(4). https://doi.org/10.1371/journal.pcbi.1006076

Christianson, H. C., & Belting, M. (2014). Heparan sulfate proteoglycan as a cell-surface endocytosis receptor. *Matrix Biology*, *35*, 51–55. https://doi.org/10.1016/j.matbio.2013.10.004

Ciompi, F., de Hoop, B., van Riel, S. J., Chung, K., Scholten, E. T., Oudkerk, M., de Jong, P. A., Prokop, M., & van Ginneken, B. (2015). Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. *Medical Image Analysis*, *26*(1), 195–202. https://doi.org/10.1016/j.media.2015.08.001

Condomitti, G., & De Wit, J. (2018). Heparan sulfate proteoglycans as emerging players in synaptic specificity. In *Frontiers in Molecular Neuroscience* (Vol. 11). https://doi.org/10.3389/fnmol.2018.00014

Corlier, F., Rivals, I., Lagarde, J., Hamelin, L., Corne, H., Dauphinot, L., Ando, K., Cossec, J. C., Fontaine, G., Dorothée, G., Malaplate-Armand, C., Olivier, J. L., Dubois, B., Bottlaender, M., Duyckaerts, C., Sarazin, M., Potier, M. C., Alnajjar-Carpentier, A., Logak, M., … Michon, A. (2015). Modifications of the endosomal compartment in peripheral blood mononuclear cells and fibroblasts from Alzheimer's disease patients. *Translational Psychiatry*, *5*, e595. https://doi.org/10.1038/tp.2015.87

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., Speed, T. P., Akil, H., Watson, S. J., & Meng, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research*, *33*(20). https://doi.org/10.1093/nar/gni179

Dang, C. V. (2012). MYC on the path to cancer. In *Cell* (Vol. 149, Issue 1, pp. 22–35). https://doi.org/10.1016/j.cell.2012.03.003

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., Gephart, M. G. H., Barres, B. A., & Quake, S. R. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(23), 7285–7290. https://doi.org/10.1073/pnas.1507125112

de Reijke, T. M., Vos, P. C. N., de Boer, E. C., Bevers, R. F. M., de Muinck Keizer, W. H., Kurth, K. H., & Schamhart, D. H. J. (1993). Cytokine production by the human bladder carcinoma cell line T24 in the presence of bacillus Calmette-Guerin (BCG). *Urological Research*, *21*(5), 349–352. https://doi.org/10.1007/BF00296835

Dede, D. S., Yavuz, B., Yavuz, B. B., Cankurtaran, M., Halil, M., Ulger, Z., Cankurtaran, E. S., Aytemir, K., Kabakci, G., & Ariogul, S. (2007). Assessment of endothelial function in Alzheimer's disease: Is Alzheimer's disease a vascular disease? *Journal of the American Geriatrics Society*, *55*(10), 1613–1617. https://doi.org/10.1111/j.1532-5415.2007.01378.x

Dembéle, D., & Kastner, P. (2003). Fuzzy C-means method for clustering microarray data. *Bioinformatics*, *19*(8), 973–980. https://doi.org/10.1093/bioinformatics/btg119

Denslow, S. A., & Wade, P. A. (2007). The human Mi-2/NuRD complex and gene regulation. In *Oncogene* (Vol. 26, Issue 37, pp. 5433–5438). https://doi.org/10.1038/sj.onc.1210611

Desikan, R. S., Fan, C. C., Wang, Y., Schork, A. J., Cabral, H. J., Cupples, L. A., Thompson, W. K., Besser, L., Kukull, W. A., Holland, D., Chen, C. H., Brewer, J. B., Karow, D. S., Kauppi, K., Witoelar, A., Karch, C. M., Bonham, L. W., Yokoyama, J. S., Rosen, H. J., … Dale, A. M. (2017). Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. *PLoS Medicine*, *14*(3). https://doi.org/10.1371/journal.pmed.1002258

Ding, G., Li, W., Liu, J., Zeng, Y., Mao, C., Kang, Y., & Shang, J. (2017). LncRNA GHET1 activated by H3K27 acetylation promotes cell tumorigenesis through regulating ATF1 in hepatocellular carcinoma. *Biomedicine and Pharmacotherapy*, *94*, 326–331. https://doi.org/10.1016/j.biopha.2017.07.046

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dong, W. L., Tan, F. Q., & Yang, W. X. (2015). Wnt signaling in testis development: Unnecessary or essential? In *Gene* (Vol. 565, Issue 2, pp. 155–165). https://doi.org/10.1016/j.gene.2015.04.066

Du, A. T., Schuff, N., Amend, D., Laakso, M. P., Hsu, Y. Y., Jagust, W. J., Yaffe, K., Kramer, J. H., Reed, B., Norman, D., Chui, H. C., & Weiner, M. W. (2001). Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. *Journal of Neurology Neurosurgery and Psychiatry*, *71*(4), 441–447. https://doi.org/10.1136/jnnp.71.4.441

Duchi, J., Hazan, E., & Singer, Y. (2010). Adaptive subgradient methods for online learning and stochastic optimization. *COLT 2010 - The 23rd Conference on Learning Theory*, 257–269.

Dunning, M. J., Smith, M. L., Ritchie, M. E., & Tavaré, S. (2007). Beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, *23*(16), 2183–2184. https://doi.org/10.1093/bioinformatics/btm311

Ekim Üstünel, B., Friedrich, K., Maida, A., Wang, X., Krones-Herzig, A., Seibert, O., Sommerfeld, A., Jones, A., Sijmonsma, T. P., Sticht, C., Gretz, N., Fleming, T., Nawroth, P. P., Stremmel, W., Rose, A. J., Berriel-Diaz, M., Blüher, M., & Herzig, S. (2016). Control of diabetic hyperglycaemia and insulin resistance through TSC22D4. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms13267

Fagerberg, L., Hallstrom, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., Asplund, A., Sjostedt, E., Lundberg, E., Szigyarto, C. A. K., Skogs, M., Ottosson Takanen, J., Berling, H., Tegel, H., Mulder, J., … Uhlen, M. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular and Cellular Proteomics*, *13*(2), 397–406. https://doi.org/10.1074/mcp.M113.035600

Fiest, K. M., Roberts, J. I., Maxwell, C. J., Hogan, D. B., Smith, E. E., Frolkis, A., Cohen, A., Kirk, A., Pearson, D., Pringsheim, T., Venegas-Torres, A., & Jetté, N. (2016). The prevalence and incidence of dementia due to Alzheimer's disease: A systematic review and meta-analysis. In *Canadian Journal of Neurological Sciences* (Vol. 43, Issue S1, pp. S51–S82). https://doi.org/10.1017/cjn.2016.36

Fizazi, K., Greco, F. A., Pavlidis, N., Daugaard, G., Oien, K., & Pentheroudakis, G. (2015). Cancers

of unknown primary site: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, *26*, vi64–vi68. https://doi.org/10.1093/annonc/mdv305

Fontanillo, C., Nogales-Cadenas, R., Pascual-Montano, A., & de Las Rivas, J. (2011). Functional analysis beyond enrichment: Non-redundant reciprocal linkage of genes and biological terms. *PLoS ONE*, *6*(9). https://doi.org/10.1371/journal.pone.0024289

Fox, N. C., Warrington, E. K., Freeborough, P. A., Hartikainen, P., Kennedy, A. M., Stevens, J. M., & Rossor, M. N. (1996). Presymptomatic hippocampal atrophy in Alzheimer's disease A longitudinal MRI study. *Brain*, *119*(6), 2001–2007. https://doi.org/10.1093/brain/119.6.2001

Futschik, M. E., & Carlisle, B. (2005). Noise-robust soft clustering of gene expression time-course data. *Journal of Bioinformatics and Computational Biology*, *3*(4), 965–988. https://doi.org/10.1142/S0219720005001375

Gao, G. F., Parker, J. S., Reynolds, S. M., Silva, T. C., Wang, L. B., Zhou, W., Akbani, R., Bailey, M., Balu, S., Berman, B. P., Brooks, D., Chen, H., Cherniack, A. D., Demchok, J. A., Ding, L., Felau, I., Gaheen, S., Gerhard, D. S., Heiman, D. I., … Noble, M. S. (2019). Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. *Cell Systems*, *9*(1), 24-34.e10. https://doi.org/10.1016/j.cels.2019.06.006

Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., & Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, *29*(8), 1363–1375. https://doi.org/10.1101/gr.240663.118

Garcia-Moreno, S. A., Lin, Y. T., Futtner, C. R., Salamone, I. M., Capel, B., & Maatouk, D. M. (2019). CBX2 is required to stabilize the testis pathway by repressing wnt signaling. *PLoS Genetics*, *15*(5). https://doi.org/10.1371/journal.pgen.1007895

Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). Affy - Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, *20*(3), 307–315. https://doi.org/10.1093/bioinformatics/btg405

Gettinger, S., Choi, J., Hastings, K., Truini, A., Datar, I., Sowell, R., Wurtz, A., Dong, W., Cai, G., Melnick, M. A., Du, V. Y., Schlessinger, J., Goldberg, S. B., Chiang, A., Sanmamed, M. F., Melero, I., Agorreta, J., Montuenga, L. M., Lifton, R., … Politi, K. (2017). Impaired HLA class I antigen processing and presentation as a mechanism of acquired resistance to immune checkpoint inhibitors in lung cancer. *Cancer Discovery*, *7*(12), 1420–1435. https://doi.org/10.1158/2159-8290.CD-17-0593

Goedert, M., Jakes, R., Spillantini, M. G., Hasegawa, M., Smith, M. J., & Crowther, R. A. (1996).

Assembly of microtubule-associated protein tau into Alzheimer-like filaments induced by sulphated glycosaminoglycans. *Nature*, *383*(6600), 550–553. https://doi.org/10.1038/383550a0

Goodman, L. A., & Kruskal, W. H. (1954). Measures of Association for Cross Classifications*. *Journal of the American Statistical Association*, *49*(268), 732–764. https://doi.org/10.1080/01621459.1954.10501231

Gos, M., Fahiminiya, S., Poznański, J., Klapecki, J., Obersztyn, E., Piotrowicz, M., Wierzba, J., Posmyk, R., Bal, J., & Majewski, J. (2014). Contribution of RIT1 mutations to the pathogenesis of Noonan syndrome: Four new cases and further evidence of heterogeneity. *American Journal of Medical Genetics, Part A*, *164*(9), 2310–2316. https://doi.org/10.1002/ajmg.a.36646

Grammas, P. (2011). Neurovascular dysfunction, inflammation and endothelial activation: Implications for the pathogenesis of Alzheimer's disease. In *Journal of Neuroinflammation* (Vol. 8, p. 26). https://doi.org/10.1186/1742-2094-8-26

Grzeskowiak, C. L., Kundu, S. T., Mo, X., Ivanov, A. A., Zagorodna, O., Lu, H., Chapple, R. H., Tsang, Y. H., Moreno, D., Mosqueda, M., Eterovic, K., Fradette, J. J., Ahmad, S., Chen, F., Chong, Z., Chen, K., Creighton, C. J., Fu, H., Mills, G. B., … Scott, K. L. (2018). In vivo screening identifies GATAD2B as a metastasis driver in KRAS-driven lung cancer. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-04572-3

Gu, W., Wang, B., Wan, F., Wu, J., Lu, X., Wang, H., Zhu, Y., Zhang, H., Shi, G., Dai, B., & Ye, D. (2018). SOX2 and SOX12 are predictive of prognosis in patients with clear cell renal cell carcinoma. *Oncology Letters*, *15*(4), 4564–4570. https://doi.org/10.3892/ol.2018.7828

Guerrero, S., López-Cortés, A., Indacochea, A., García-Cárdenas, J. M., Zambrano, A. K., Cabrera-Andrade, A., Guevara-Ramírez, P., González, D. A., Leone, P. E., & Paz-y-Miño, C. (2018). Analysis of Racial/Ethnic Representation in Select Basic and Applied Cancer Research Studies. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-32264-x

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S. V., Klotzle, B., Bibikova, M., Fan, J. B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T., & Zhang, K. (2013). Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell*, *49*(2), 359–367. https://doi.org/10.1016/j.molcel.2012.10.016

Hawrylycz, M. J., Lein, E. S., Guillozet-Bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., Van De Lagemaat, L. N., Smith, K. A., Ebbert, A., Riley, Z. L., Abajian, C., Beckmann, C. F., Bernard,

A., Bertagnolli, D., Boe, A. F., Cartagena, P. M., Mallar Chakravarty, M., Chapin, M., Chong, J., … Jones, A. R. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, *489*(7416), 391–399. https://doi.org/10.1038/nature11405

Hawrylycz, M., Miller, J. A., Menon, V., Feng, D., Dolbeare, T., Guillozet-Bongaarts, A. L., Jegga, A. G., Aronow, B. J., Lee, C. K., Bernard, A., Glasser, M. F., Dierker, D. L., Menche, J., Szafer, A., Collman, F., Grange, P., Berman, K. A., Mihalas, S., Yao, Z., … Lein, E. (2015). Canonical genetic signatures of the adult human brain. *Nature Neuroscience*, *18*(12), 1832–1844. https://doi.org/10.1038/nn.4171

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision*, *2015 International Conference on Computer Vision*, *ICCV 2015*, 1026–1034. https://doi.org/10.1109/ICCV.2015.123

Hernandez, D. G., Nalls, M. A., Moore, M., Chong, S., Dillman, A., Trabzuni, D., Gibbs, J. R., Ryten, M., Arepalli, S., Weale, M. E., Zonderman, A. B., Troncoso, J., O'Brien, R., Walker, R., Smith, C., Bandinelli, S., Traynor, B. J., Hardy, J., Singleton, A. B., & Cookson, M. R. (2012). Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiology of Disease*, *47*(1), 20–28. https://doi.org/10.1016/j.nbd.2012.03.020

Holmes, B. B., DeVos, S. L., Kfoury, N., Li, M., Jacks, R., Yanamandra, K., Ouidja, M. O., Brodsky, F. M., Marasa, J., Bagchi, D. P., Kotzbauer, P. T., Miller, T. M., Papy-Garcia, D., & Diamond, M. I. (2013). Heparan sulfate proteoglycans mediate internalization and propagation of specific proteopathic seeds. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(33), E3138–E3147. https://doi.org/10.1073/pnas.1301440110

Honma, N., Saji, S., Mikami, T., Yoshimura, N., Mori, S., Saito, Y., Murayama, S., & Harada, N. (2017). Estrogen-Related Factors in the Frontal Lobe of Alzheimer's Disease Patients and Importance of Body Mass Index. *Scientific Reports*, *7*(1), 726. https://doi.org/10.1038/s41598-017-00815-3

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, *14*(10). https://doi.org/10.1186/gb-2013-14-10-r115

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. https://doi.org/10.1038/nprot.2008.211

Huang, S., Yang, J., Fong, S., & Zhao, Q. (2019). Mining prognosis index of brain metastases using artificial intelligence. *Cancers*, *11*(8). https://doi.org/10.3390/cancers11081140

Huang, S., Yang, J., Fong, S., & Zhao, Q. (2020). Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. In *Cancer Letters* (Vol. 471, pp. 61–71). https://doi.org/10.1016/j.canlet.2019.12.007

Huynh, M. B., Villares, J., Sepúlveda Díaz, J. E., Christiaans, S., Carpentier, G., Ouidja, M. O., Sissoeff, L., Raisman-Vozari, R., & Papy-Garcia, D. (2012). Glycosaminoglycans from aged human hippocampus have altered capacities to regulate trophic factors activities but not Aβ42 peptide toxicity. *Neurobiology of Aging*, *33*(5), 1005.e11-1005.e22. https://doi.org/10.1016/j.neurobiolaging.2011.09.030

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, *4*(2), 249–264. https://doi.org/10.1093/biostatistics/4.2.249

Ittner, L. M., & Götz, J. (2011). Amyloid-β and tau - A toxic pas de deux in Alzheimer's disease. *Nature Reviews Neuroscience*, *12*(2), 67–72. https://doi.org/10.1038/nrn2967

Janky, R., Verfaillie, A., Imrichová, H., van de Sande, B., Standaert, L., Christiaens, V., Hulselmans, G., Herten, K., Naval Sanchez, M., Potier, D., Svetlichnyy, D., Kalender Atak, Z., Fiers, M., Marine, J. C., & Aerts, S. (2014). iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections. *PLoS Computational Biology*, *10*(7). https://doi.org/10.1371/journal.pcbi.1003731

Jing, X., Liang, H., Hao, C., Yang, X., & Cui, X. (2019). Overexpression of MUC1 predicts poor prognosis in patients with breast cancer. *Oncology Reports*, *41*(2), 801–810. https://doi.org/10.3892/or.2018.6887

Johnson, M. B., Kawasawa, Y. I., Mason, C. E., Krsnik, Ž., Coppola, G., Bogdanović, D., Geschwind, D. H., Mane, S. M., State, M. W., & Šestan, N. (2009). Functional and Evolutionary Insights into Human Brain Development through Global Transcriptome Analysis. *Neuron*, *62*(4), 494–509. https://doi.org/10.1016/j.neuron.2009.03.027

Johnstone, R. W., Frew, A. J., & Smyth, M. J. (2008). The TRAIL apoptotic pathway in cancer onset, progression and therapy. In *Nature Reviews Cancer* (Vol. 8, Issue 10, pp. 782–798). https://doi.org/10.1038/nrc2465

Jones, A., Friedrich, K., Rohm, M., Schäfer, M., Algire, C., Kulozik, P., Seibert, O., Müller-Decker,

K., Sijmonsma, T., Strzoda, D., Sticht, C., Gretz, N., Dallinga-Thie, G. M., Leuchs, B., Kögl, M., Stremmel, W., Diaz, M. B., & Herzig, S. (2013). TSC22D4 is a molecular output of hepatic wasting metabolism. *EMBO Molecular Medicine*, *5*(2), 294–308. https://doi.org/10.1002/emmm.201201869

Kadavath, H., Hofele, R. V., Biernat, J., Kumar, S., Tepper, K., Urlaub, H., Mandelkow, E., & Zweckstetter, M. (2015). Tau stabilizes microtubules by binding at the interface between tubulin heterodimers. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(24), 7501–7506. https://doi.org/10.1073/pnas.1504081112

Kandimalla, K. K., Scott, O. G., Fulzele, S., Davidson, M. W., & Poduslo, J. F. (2009). Mechanism of Neuronal versus endothelial cell uptake of Alzheimer's disease amyloid β protein. *PLoS ONE*, *4*(2). https://doi.org/10.1371/journal.pone.0004627

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M. M., Pletikos, M., Meyer, K. A., Sedmak, G., Guennel, T., Shin, Y., Johnson, M. B., Krsnik, Ž., Mayer, S., Fertuzinhos, S., Umlauf, S., Lisgo, S. N., Vortmeyer, A., … Šestan, N. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, *478*(7370), 483–489. https://doi.org/10.1038/nature10523

Kienlen-Campard, P., Tasiaux, B., Van Hees, J., Li, M., Huysseune, S., Sato, T., Fei, J. Z., Aimoto, S., Courtoy, P. J., Smith, S. O., Constantinescu, S. N., & Octave, J. N. (2008). Amyloidogenic processing but not Amyloid Precursor Protein (APP) intracellular C-terminal domain production requires a precisely oriented APP dimer assembled by transmembrane GXXXG motifs. *Journal of Biological Chemistry*, *283*(12), 7733–7744. https://doi.org/10.1074/jbc.M707142200

Lee, M. S., & Sanoff, H. K. (2020). Cancer of unknown primary. *The BMJ*, *371*. https://doi.org/10.1136/bmj.m4050

Levine, A. B., Schlosser, C., Grewal, J., Coope, R., Jones, S. J. M., & Yip, S. (2019). Rise of the Machines: Advances in Deep Learning for Cancer Diagnosis. In *Trends in Cancer* (Vol. 5, Issue 3, pp. 157–169). https://doi.org/10.1016/j.trecan.2019.02.002

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-323

Li, J. P., & Kusche-Gullberg, M. (2016). Heparan Sulfate: Biosynthesis, Structure, and Function. In *International Review of Cell and Molecular Biology* (Vol. 325, pp. 215–273).

https://doi.org/10.1016/bs.ircmb.2016.02.009

Li, X., Zhang, S., Zhang, Q., Wei, X., Pan, Y., Zhao, J., Xin, X., Qin, C., Wang, X., Li, J., Yang, F., Zhao, Y., Yang, M., Wang, Q., Zheng, Z., Zheng, X., Yang, X., Whitlow, C. T., Gurcan, M. N., … Chen, K. (2019). Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *The Lancet Oncology*, *20*(2), 193–201. https://doi.org/10.1016/S1470-2045(18)30762-9

Liu, S., Liu, X., Li, S., Huang, X., Qian, H., Jin, K., & Xiang, M. (2020). Foxn4 is a temporal identity factor conferring mid/late-early retinal competence and involved in retinal synaptogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(9), 5016–5027. https://doi.org/10.1073/pnas.1918628117

Liu, Y., Kohlberger, T., Norouzi, M., Dahl, G. E., Smith, J. L., Mohtashamian, A., Olson, N., Peng, L. H., Hipp, J. D., & Stumpe, M. C. (2019). Artificial intelligence–based breast cancer nodal metastasis detection insights into the black box for pathologists. *Archives of Pathology and Laboratory Medicine*, *143*(7), 859–868. https://doi.org/10.5858/arpa.2018-0147-OA

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2013). The hallmarks of aging. In *Cell* (Vol. 153, Issue 6, p. 1194). https://doi.org/10.1016/j.cell.2013.05.039

Lorenzi, I., Oeljeklaus, S., Aich, A., Ronsör, C., Callegari, S., Dudek, J., Warscheid, B., Dennerlein, S., & Rehling, P. (2018). The mitochondrial TMEM177 associates with COX20 during COX2 biogenesis. *Biochimica et Biophysica Acta - Molecular Cell Research*, *1865*(2), 323–333. https://doi.org/10.1016/j.bbamcr.2017.11.010

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

Lu, T., Pan, Y., Kao, S. Y., Li, C., Kohane, I., Chan, J., & Yankner, B. A. (2004). Gene regulation and DNA damage in the ageing human brain. *Nature*, *429*(6994), 883–891. https://doi.org/10.1038/nature02661

Maïza, A., Chantepie, S., Vera, C., Fifre, A., Huynh, M. B., Stettler, O., Ouidja, M. O., & Papy-Garcia, D. (2018). The role of heparan sulfates in protein aggregation and their potential impact on neurodegeneration. In *FEBS Letters* (Vol. 592, Issue 23, pp. 3806–3818). https://doi.org/10.1002/1873-3468.13082

Mamata De, Sanford, T. R., & Wood, G. W. (1992). Interleukin-1, interleukin-6, and tumor necrosis factor α are produced in the mouse uterus during the estrous cycle and are induced by

estrogen and progesterone. *Developmental Biology*, *151*(1), 297–305. https://doi.org/10.1016/0012-1606(92)90234-8

Marr, R., & Hafez, D. (2014). Amyloid beta and Alzheimer's Disease: The role of neprilysin-2 in amyloid-beta clearance. *Frontiers in Aging Neuroscience*, *6*(JUL). https://doi.org/10.3389/fnagi.2014.00187

Martinez, P., Denys, A., Delos, M., Sikora, A. S., Carpentier, M., Julien, S., Pestel, J., & Allain, F. (2015). Macrophage polarization alters the expression and sulfation pattern of glycosaminoglycans. *Glycobiology*, *25*(5), 502–513. https://doi.org/10.1093/glycob/cwu137

Mcgeer, P. L., Akiyama, H., Itagaki, S., & Mcgeer, E. G. (1989). Immune System Response in Alzheimer's Disease. *Canadian Journal of Neurological Sciences / Journal Canadien Des Sciences Neurologiques*, *16*(S4), 516–527. https://doi.org/10.1017/S0317167100029863

Mitsou, I., Multhaupt, H. A. B., & Couchman, J. R. (2017). Proteoglycans, ion channels and cell-matrix adhesion. In *Biochemical Journal* (Vol. 474, Issue 12, pp. 1965–1979). https://doi.org/10.1042/BCJ20160747

Mochizuki, H., Yoshida, K., Shibata, Y., & Kimata, K. (2008). Tetrasulfated disaccharide unit in heparan sulfate: Enzymatic formation and tissue distribution. *Journal of Biological Chemistry*, *283*(45), 31237–31245. https://doi.org/10.1074/jbc.M801586200

Moreira, P. I., Carvalho, C., Zhu, X., Smith, M. A., & Perry, G. (2010). Mitochondrial dysfunction is a trigger of Alzheimer's disease pathophysiology. In *Biochimica et Biophysica Acta - Molecular Basis of Disease* (Vol. 1802, Issue 1, pp. 2–10). https://doi.org/10.1016/j.bbadis.2009.10.006

Moreira, P. I., Siedlak, S. L., Wang, X., Santos, M. S., Oliveira, C. R., Tabaton, M., Nunomura, A., Szweda, L. I., Aliev, G., Smith, M. A., Zhu, X., & Perry, G. (2007). Erratum: Increased autophagic degradation of mitochondria in Alzheimer disease (Autophagy). In *Autophagy* (Vol. 3, Issue 6, pp. 614–615). https://doi.org/10.4161/auto.4872

Morris, M. R., Hughes, D. J., Tian, Y. M., Ricketts, C. J., Lau, K. W., Gentle, D., Shuib, S., Serrano-Fernandez, P., Lubinski, J., Wiesener, M. S., Pugh, C. W., Latif, F., Ratcliffe, P. J., & Maher, E. R. (2009). Mutation analysis of hypoxia-inducible factors HIF1A and HIF2A in renal cell carcinoma. *Anticancer Research*, *29*(11), 4337–4343.

Mu, Y., & Gage, F. H. (2011). Adult hippocampal neurogenesis and its role in Alzheimer's disease. In *Molecular Neurodegeneration* (Vol. 6, Issue 1, p. 85). https://doi.org/10.1186/1750-1326-6-85

Nichols, N. R., Day, J. R., Laping, N. J., Johnson, S. A., & Finch, C. E. (1993). GFAP mRNA increases

with age in rat and human brain. *Neurobiology of Aging*, *14*(5), 421–429. https://doi.org/10.1016/0197-4580(93)90100-P

Niu, H., Álvarez-Álvarez, I., Guillén-Grima, F., & Aguinaga-Ontoso, I. (2017). Prevalence and incidence of Alzheimer's disease in Europe: A meta-analysis. *Neurología (English Edition)*, *32*(8), 523–532. https://doi.org/10.1016/j.nrleng.2016.02.009

O'Brien, R. J., & Wong, P. C. (2011). Amyloid precursor protein processing and alzheimer's disease. *Annual Review of Neuroscience*, *34*, 185–204. https://doi.org/10.1146/annurev-neuro-061010-113613

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, *375*(13), 1216–1219. https://doi.org/10.1056/nejmp1606181

Oh, S. E., Choi, M. G., & Seo, S. W. (2019). ASO Author Reflections: Use of the Survival Recurrent Network for Prediction of Overall Survival in Patients with Gastric Cancer. In *Annals of Surgical Oncology* (Vol. 26, pp. 539–540). https://doi.org/10.1245/s10434-018-7044-y

Olden, J. D., & Jackson, D. A. (2002). Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, *154*(1–2), 135–150. https://doi.org/10.1016/S0304-3800(02)00064-9

Oldham, M. C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., & Geschwind, D. H. (2008). Functional organization of the transcriptome in human brain. *Nature Neuroscience*, *11*(11), 1271–1282. https://doi.org/10.1038/nn.2207

Ollerenshaw, M., Page, T., Hammonds, J., & Demaine, A. (2004). Polymorphisms in the hypoxia inducible factor-1α gene (HIF1A) are associated with the renal cell carcinoma phenotype. *Cancer Genetics and Cytogenetics*, *153*(2), 122–126. https://doi.org/10.1016/j.cancergencyto.2004.01.014

Ori, A., Wilkinson, M. C., & Fernig, D. G. (2011). A systems biology approach for the investigation of the heparin/heparan sulfate interactome. *Journal of Biological Chemistry*, *286*(22), 19892–19904. https://doi.org/10.1074/jbc.M111.228114

Papageorgis, P., Ozturk, S., Lambert, A. W., Neophytou, C. M., Tzatsos, A., Wong, C. K., Thiagalingam, S., & Constantinou, A. I. (2015). Targeting IL13Ralpha2 activates STAT6-TP63 pathway to suppress breast cancer lung metastasis. *Breast Cancer Research*, *17*(1). https://doi.org/10.1186/s13058-015-0607-y

Patey, S. J., Edwards, E. A., Yates, E. A., & Turnbull, J. E. (2006). Heparin derivatives as inhibitors

of BACE-1, the Alzheimer's β-secretase, with reduced activity against factor Xa and other proteases. *Journal of Medicinal Chemistry*, *49*(20), 6129–6132. https://doi.org/10.1021/jm0512210

Peters, M. J., Joehanes, R., Pilling, L. C., Schurmann, C., Conneely, K. N., Powell, J., Reinmaa, E., Sutphin, G. L., Zhernakova, A., Schramm, K., Wilson, Y. A., Kobes, S., Tukiainen, T., Ramos, Y. F., Göring, H. H. H., Fornage, M., Liu, Y., Gharib, S. A., Stranger, B. E., … Singleton, A. B. (2015). The transcriptional landscape of age in human peripheral blood. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms9570

Poirier, K., Lebrun, N., Broix, L., Tian, G., Saillour, Y., Boscheron, C., Parrini, E., Valence, S., Pierre, B. Saint, Oger, M., Lacombe, D., Geneviève, D., Fontana, E., Darra, F., Cances, C., Barth, M., Bonneau, D., Bernadina, B. D., N'Guyen, S., … Chelly, J. (2013). Mutations in TUBG1, DYNC1H1, KIF5C and KIF2A cause malformations of cortical development and microcephaly. *Nature Genetics*, *45*(6), 639–647. https://doi.org/10.1038/ng.2613

Putin, E., Mamoshina, P., Aliper, A., Korzinkin, M., Moskalev, A., Kolosov, A., Ostrovskiy, A., Cantor, C., Vijg, J., & Zhavoronkov, A. (2016). Deep biomarkers of human aging: Application of deep neural networks to biomarker development. *Aging*, *8*(5), 1021–1033. https://doi.org/10.18632/aging.100968

Rasmussen, J., & Langerman, H. (2019). Alzheimer's Disease – Why We Need Early Diagnosis . *Degenerative Neurological and Neuromuscular Disease*, *Volume 9*, 123–130. https://doi.org/10.2147/dnnd.s228939

Raziuddin, S., Masihuzzaman, M., Shetty, S., & Ibrahim, A. (1993). Tumor necrosis factor alpha production in schistosomiasis with carcinoma of urinary bladder. *Journal of Clinical Immunology*, *13*(1), 23–29. https://doi.org/10.1007/BF00920632

Rhinn, H., & Abeliovich, A. (2017). Differential Aging Analysis in Human Cerebral Cortex Identifies Variants in TMEM106B and GRN that Regulate Aging Phenotypes. *Cell Systems*, *4*(4), 404-415.e5. https://doi.org/10.1016/j.cels.2017.02.009

Rho, H. W., Lee, B. C., Choi, E. S., Choi, I. J., Lee, Y. S., & Goh, S. H. (2010). Identification of valid reference genes for gene expression studies of human stomach cancer by reverse transcription-qPCR. *BMC Cancer*, *10*. https://doi.org/10.1186/1471-2407-10-240

Riley, J. W., Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A., & Williams, R. M. (1949). The American Soldier: Adjustment During Army Life. In *American Sociological Review* (Vol. 14, Issue 4). https://doi.org/10.2307/2087216

Risueño, A., Fontanillo, C., Dinger, M. E., & De Las Rivas, J. (2010). GATExplorer: Genomic and Transcriptomic Explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs. *BMC Bioinformatics*, *11*, 221. https://doi.org/10.1186/1471-2105-11-221

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Rössler, M., Zarski, R., Bohl, J., & Ohm, T. G. (2002). Stage-dependent and sector-specific neuronal loss in hippocampus during alzheimer's disease. *Acta Neuropathologica*, *103*(4), 363–369. https://doi.org/10.1007/s00401-001-0475-7

Salvadores, M., Fuster-Tormo, F., & Supek, F. (2020). Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns. *Science Advances*, *6*(27). https://doi.org/10.1126/sciadv.aba1862

Sandwall, E., O'Callaghan, P., Zhang, X., Lindahl, U., Lannfelt, L., & Li, J. P. (2010). Heparan sulfate mediates amyloid-beta internalization and cytotoxicity. *Glycobiology*, *20*(5), 533–541. https://doi.org/10.1093/glycob/cwp205

Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., Garnett, M. J., Blüthgen, N., & Saez-Rodriguez, J. (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-017-02391-6

Sepe, F. N., Chiasserini, D., & Parnetti, L. (2018). Role of FABP3 as biomarker in Alzheimer's disease and synucleinopathies. *Future Neurology*, *13*(4), 199–207. https://doi.org/10.2217/fnl-2018-0003

Sepulveda-Diaz, J. E., Alavi Naini, S. M., Huynh, M. B., Ouidja, M. O., Yanicostas, C., Chantepie, S., Villares, J., Lamari, F., Jospin, E., Van Kuppevelt, T. H., Mensah-Nyagan, A. G., Raisman-Vozari, R., Soussi-Yanicostas, N., & Papy-Garcia, D. (2015). HS3ST2 expression is critical for the abnormal phosphorylation of tau in Alzheimer's disease-related tau pathology. *Brain*, *138*(5), 1339–1354. https://doi.org/10.1093/brain/awv056

Sharma, P., Srivastava, P., Seth, A., Tripathi, P. N., Banerjee, A. G., & Shrivastava, S. K. (2019). Comprehensive review of mechanisms of pathogenesis involved in Alzheimer's disease and potential therapeutic strategies. In *Progress in Neurobiology* (Vol. 174, pp. 53–89). https://doi.org/10.1016/j.pneurobio.2018.12.006

Shi, G.-X., & Andres, D. A. (2005). Rit Contributes to Nerve Growth Factor-Induced Neuronal

Differentiation via Activation of B-Raf-Extracellular Signal-Regulated Kinase and p38 Mitogen-Activated Protein Kinase Cascades. *Molecular and Cellular Biology*, *25*(2), 830–846. https://doi.org/10.1128/mcb.25.2.830-846.2005

Shi, Y., Wang, W., Yang, B., & Tian, H. (2017). ATF1 and RAS in exosomes are potential clinical diagnostic markers for cervical cancer. *Cell Biochemistry and Function*, *35*(7), 477–483. https://doi.org/10.1002/cbf.3307

Shroff, E. H., Eberlin, L. S., Dang, V. M., Gouw, A. M., Gabay, M., Adam, S. J., Bellovin, D. I., Trand, P. T., Philbrick, W. M., Garcia-Ocana, A., Casey, S. C., Li, Y., Dang, C. V., Zare, R. N., & Felsher, D. W. (2015). MYC oncogene overexpression drives renal cell carcinoma in a mouse model through glutamine metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(21), 6539–6544. https://doi.org/10.1073/pnas.1507228112

Shworak, N. W., Liu, J., Petros, L. M., Zhang, L., Kobayashi, M., Copeland, N. G., Jenkins, N. A., & Rosenberg, R. D. (1999). Multiple isoforms of heparan sulfate D-glucosaminyl 3-O-sulfotransferase: Isolation, characterization, and expression of human cDNAs and identification of distinct genomic loci. *Journal of Biological Chemistry*, *274*(8), 5170–5184. https://doi.org/10.1074/jbc.274.8.5170

Šimić, G., Babić Leko, M., Wray, S., Harrington, C., Delalle, I., Jovanov-Milošević, N., Bažadona, D., Buée, L., de Silva, R., Giovanni, G. Di, Wischik, C., & Hof, P. R. (2016). Tau protein hyperphosphorylation and aggregation in alzheimer's disease and other tauopathies, and possible neuroprotective strategies. In *Biomolecules* (Vol. 6, Issue 1, pp. 2–28). https://doi.org/10.3390/biom6010006

Sims, R., Hill, M., & Williams, J. (2020). The multiplex model of the genetics of Alzheimer's disease. In *Nature Neuroscience* (Vol. 23, Issue 3, pp. 311–322). https://doi.org/10.1038/s41593-020-0599-5

Snow, A. D., Mar, H., Nochlin, D., Sekiguchi, R. T., Kimata, K., Koike, Y., & Wight, T. N. (1990). Early accumulation of heparan sulfate in neurons and in the beta-amyloid protein-containing lesions of Alzheimer's disease and Down's syndrome. *American Journal of Pathology*, *137*(5), 1253–1270. https://doi.org/10.1016/0197-4580(90)90765-r

Soboll, G., Shen, L., & Wira, C. R. (2006). Expression of Toll-like receptors (TLR) and responsiveness to TLR agonists by polarized mouse uterine epithelial cells in culture. *Biology of Reproduction*, *75*(1), 131–139. https://doi.org/10.1095/biolreprod.106.050690

Stine, Z. E., Walton, Z. E., Altman, B. J., Hsieh, A. L., & Dang, C. V. (2015). MYC, metabolism, and

cancer. In *Cancer Discovery* (Vol. 5, Issue 10, pp. 1024–1039). https://doi.org/10.1158/2159-8290.CD-15-0507

Su, J. H., Cummings, B. J., & Cotman, C. W. (1992). Localization of heparan sulfate glycosaminoglycan and proteoglycan core protein in aged brain and Alzheimer's disease. *Neuroscience*, *51*(4), 801–813. https://doi.org/10.1016/0306-4522(92)90521-3

Suzuki, Y., Sa, Q., Ochiai, E., Mullins, J., Yolken, R., & Halonen, S. K. (2013). Cerebral Toxoplasmosis. Pathogenesis, Host Resistance and Behavioural Consequences. In *Toxoplasma Gondii: The Model Apicomplexan - Perspectives and Methods: Second Edition* (pp. 755–796). https://doi.org/10.1016/B978-0-12-396481-6.00023-4

Takeda, H., Tanaka, T., Shi, W., Matsumoto, M., Minami, M., Kashiwamura, S. I., Nakanishi, K., Yoshida, N., Kishimoto, T., & Akira, S. (1996). Essential role of Stat6 in IL-4 signalling. *Nature*, *380*(6575), 627–630. https://doi.org/10.1038/380627a0

Tanahashi, H., & Tabira, T. (1999). Molecular cloning of human Fe65L2 and its interaction with the Alzheimer's β-amyloid precursor protein. *Neuroscience Letters*, *261*(3), 143–146. https://doi.org/10.1016/S0304-3940(98)00995-1

Tandon, N., Goller, K., Wang, F., Soibam, B., Gagea, M., Jain, A. K., Schwartz, R. J., & Liu, Y. (2019). Aberrant expression of embryonic mesendoderm factor MESP1 promotes tumorigenesis. *EBioMedicine*, *50*, 55–66. https://doi.org/10.1016/j.ebiom.2019.11.012

Tang, S. W., Chang, W. H., Su, Y. C., Chen, Y. C., Lai, Y. H., Wu, P. T., Hsu, C. I., Lin, W. C., Lai, M. K., & Lin, J. Y. (2009). MYC pathway is activated in clear cell renal cell carcinoma and essential for proliferation of clear cell renal cell carcinoma cells. *Cancer Letters*, *273*(1), 35–43. https://doi.org/10.1016/j.canlet.2008.07.038

Thacker, B. E., Xu, D., Lawrence, R., & Esko, J. D. (2014). Heparan sulfate 3-O-sulfation: A rare modification in search of a function. *Matrix Biology*, *35*, 60–72. https://doi.org/10.1016/j.matbio.2013.12.001

Tian, J., Chang, J., Gong, J., Lou, J., Fu, M., Li, J., Ke, J., Zhu, Y., Gong, Y., Yang, Y., Zou, D., Peng, X., Yang, N., Mei, S., Wang, X., Zhong, R., Cai, K., & Miao, X. (2019). Systematic Functional Interrogation of Genes in GWAS Loci Identified ATF1 as a Key Driver in Colorectal Cancer Modulated by a Promoter-Enhancer Interaction. *American Journal of Human Genetics*, *105*(1), 29–47. https://doi.org/10.1016/j.ajhg.2019.05.004

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. In *Nature Medicine* (Vol. 25, Issue 1, pp. 44–56). https://doi.org/10.1038/s41591-

018-0300-7

Trabzuni, D., Ramasamy, A., Imran, S., Walker, R., Smith, C., Weale, M. E., Hardy, J., & Ryten, M. (2013). Widespread sex differences in gene expression and splicing in the adult human brain. *Nature Communications*, *4*. https://doi.org/10.1038/ncomms3771

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I. M., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A. K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., … Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220). https://doi.org/10.1126/science.1260419

Vagnucci, A. H., & Li, W. W. (2003). Alzheimer's disease and angiogenesis. In *Lancet* (Vol. 361, Issue 9357, pp. 605–608). https://doi.org/10.1016/S0140-6736(03)12521-4

Van Der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2625.

Van Horssen, J., Wesseling, P., Van Den Heuvel, L. P. W. J., De Waal, R. M. W., & Verbeek, M. M. (2003). Heparan sulphate proteoglycans in Alzheimer's disease and amyloid-related disorders. In *Lancet Neurology* (Vol. 2, Issue 8, pp. 482–492). https://doi.org/10.1016/S1474-4422(03)00484-8

Vassar, R., Bennett, B. D., Babu-Khan, S., Kahn, S., Mendiaz, E. A., Denis, P., Teplow, D. B., Ross, S., Amarante, P., Loeloff, R., Luo, Y., Fisher, S., Fuller, J., Edenson, S., Lile, J., Jarosinski, M. A., Biere, A. L., Curran, E., Burgess, T., … Citron, M. (1999). β-Secretase cleavage of Alzheimer's amyloid precursor protein by the transmembrane aspartic protease BACE. *Science*, *286*(5440), 735–741. https://doi.org/10.1126/science.286.5440.735

Velpula, K. K., Rehman, A. A., Chigurupati, S., Sanam, R., Kishore Inampudi, K., & Akila, C. S. (2012). Computational analysis of human and mouse CREB3L4 Protein. *Bioinformation*, *8*(12), 574–577. https://doi.org/10.6026/97320630008574

Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., Rynes, E., Reynolds, A., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Kaul, R., Meuleman, W., & Stamatoyannopoulos, J. A. (2020). Global reference mapping of human transcription factor footprints. *Nature*, *583*(7818), 729–736. https://doi.org/10.1038/s41586-020-2528-x

Wang, C. G., Ye, Y. J., Yuan, J., Liu, F. F., Zhang, H., & Wang, S. (2010). EZH2 and STAT6 expression profiles are correlated with colorectal cancer stage and prognosis. *World Journal*

of Gastroenterology, *16*(19), 2421–2427. https://doi.org/10.3748/wjg.v16.i19.2421

Wang, C., Zhang, F., Jiang, S., Siedlak, S. L., Shen, L., Perry, G., Wang, X., Tang, B., & Zhu, X. (2016). Estrogen receptor-α is localized to neurofibrillary tangles in Alzheimer's disease. *Scientific Reports*, *6*, 20352. https://doi.org/10.1038/srep20352

Wang, G. L., Jiang, B. H., Rue, E. A., & Semenza, G. L. (1995). Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O2 tension. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(12), 5510–5514. https://doi.org/10.1073/pnas.92.12.5510

Wang, N., Tao, L., Zhong, H., Zhao, S., Yu, Y., Yu, B., Chen, X., Gao, J., & Wang, R. (2016). MiR-135b inhibits tumour metastasis in prostate cancer by targeting STAT6. *Oncology Letters*, *11*(1), 543–550. https://doi.org/10.3892/ol.2015.3970

Wang, X., Wang, W., Li, L., Perry, G., Lee, H. gon, & Zhu, X. (2014). Oxidative stress and mitochondrial dysfunction in Alzheimer's disease. In *Biochimica et Biophysica Acta - Molecular Basis of Disease* (Vol. 1842, Issue 8, pp. 1240–1247). https://doi.org/10.1016/j.bbadis.2013.10.015

Wani, M. A., Haynes, L. D., Kim, J., Bronson, C. L., Chaudhury, C., Mohanty, S., Waldmann, T. A., Robinson, J. M., & Anderson, C. L. (2006). Familial hypercatabolic hypoproteinemia caused by deficiency of the neonatal Fc receptor, FcRn, due to a mutant β2-microglobulin gene. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(13), 5084–5089. https://doi.org/10.1073/pnas.0600548103

Watanabe, N., Kageyama, R., & Ohtsuka, T. (2015). Hbp1 regulates the timing of neuronal differentiation during cortical development by controlling cell cycle progression. *Development (Cambridge)*, *142*(13), 2278–2290. https://doi.org/10.1242/dev.120477

Weller, J., & Budson, A. (2018). Current understanding of Alzheimer's disease diagnosis and treatment. In *F1000Research* (Vol. 7, p. 1161). https://doi.org/10.12688/f1000research.14506.1

West, M. J., Coleman, P. D., Flood, D. G., & Troncoso, J. C. (1994). Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. *The Lancet*, *344*(8925), 769–772. https://doi.org/10.1016/S0140-6736(94)92338-8

Whitlock, M. C. (2005). Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology*, *18*(5), 1368–1373. https://doi.org/10.1111/j.1420-9101.2005.00917.x

Williamson, R., & Sutherland, C. (2011). Neuronal Membranes are Key to the Pathogenesis of Alzheimer's Disease: the Role of Both Raft and Non-Raft Membrane Domains. *Current Alzheimer Research*, *999*(999), 1–9. https://doi.org/10.2174/156721121222605205o

Wilsie, L. C., Gonzales, A. M., & Orlando, R. A. (2006). Syndecan-1 mediates internalization of apoE-VLDL through a low density lipoprotein receptor-related protein (LRP)-independent, non-clathrin-mediated pathway. *Lipids in Health and Disease*, *5*(23). https://doi.org/10.1186/1476-511X-5-23

Wirz, K. T. S., Keitel, S., Swaab, D. F., Verhaagen, J., & Bossers, K. (2014). Early molecular changes in Alzheimer disease: Can we catch the disease in its presymptomatic phase? In *Journal of Alzheimer's Disease* (Vol. 38, Issue 4, pp. 719–740). https://doi.org/10.3233/JAD-130920

Xu, Y., Zhou, W., Zhang, C., Liu, X., Lv, J., Li, X., Zhao, L., Li, W., Li, J., Ren, Y., & Ou, R. (2019). Long non-coding RNA RP11-552M11.4 favors tumorigenesis and development of cervical cancer via modulating miR-3941/ATF1 signaling. *International Journal of Biological Macromolecules*, *130*, 24–33. https://doi.org/10.1016/j.ijbiomac.2019.02.083

Yabe, T., Hata, T., He, J., & Maeda, N. (2005). Developmental and regional expression of heparan sulfate sulfotransferase genes in the mouse brain. *Glycobiology*, *15*(10), 982–993. https://doi.org/10.1093/glycob/cwi090

Yang, P., Wang, Y., Chen, J., Li, H., Kang, L., Zhang, Y., Chen, S., Zhu, B., & Gao, S. (2011). RCOR2 is a subunit of the LSD1 complex that regulates ESC property and substitutes for SOX2 in reprogramming somatic cells to pluripotency. *Stem Cells*, *29*(5), 791–801. https://doi.org/10.1002/stem.634

Zaykin, D. V. (2011). Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of Evolutionary Biology*, *24*(8), 1836–1841. https://doi.org/10.1111/j.1420-9101.2011.02297.x

Zhang, Q., Sidorenko, J., Couvy-Duchesne, B., Marioni, R. E., Wright, M. J., Goate, A. M., Marcora, E., Huang, K. lin, Porter, T., Laws, S. M., Masters, C. L., Bush, A. I., Fowler, C., Darby, D., Pertile, K., Restrepo, C., Roberts, B., Robertson, J., Rumble, R., … Visscher, P. M. (2020). Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture. *Nature Communications*, *11*(1), 4799. https://doi.org/10.1038/s41467-020-18534-1

Zhao, L. N., Long, H., Mu, Y., & Chew, L. Y. (2012). The toxicity of amyloid β oligomers. In *International Journal of Molecular Sciences* (Vol. 13, Issue 6, pp. 7303–7327). https://doi.org/10.3390/ijms13067303

LIST OF PUBLICATIONS

List of publications and scientific output associated to this PhD thesis work.

Scientific record: *https://scholar.google.com/citations?user=c8x96qEAAAAJ*

ORCID ID: *0000-0002-5054-8635*

1 Main publications associated to this PhD:

Gonzalez-Velasco Óscar, Papy-Garcia Dulce, Gael Duaron, Sanchez-Santos Jose Manuel, De Las Rivas Javier. ***Transcriptomic landscape, gene signatures and regulatory profile of aging in the human brain***. BBA Gene regulatory mechanism. 2020 ;1863(6):194491. doi: 10.1016/j.bbagrm.2020.194491

2 Additional publications:

Gonzalez-Velasco Óscar, De Las Rivas Javier, Lacal Jesús. ***Proteomic and Transcriptomic Profiling Identifies Early Developmentally Regulated Proteins in Dictyostelium discoideum***. Cells (2019) 8(10):1187. doi: 10.3390/cells8101187. PMID:31581556.

ACKNOWLEDGEMENTS

Agradecimientos

En primer lugar, quiero agradecer especialmente a mi director de tesis, el Dr. Javier De Las Rivas, por darme la maravillosa oportunidad de trabajar en su laboratorio y, con ello, abrirme las puertas al extraordinario mundo de la ciencia (*casi* siempre extraordinario), siempre soñé con poder trabajar en ésta una de mis grandes pasiones (*el que trabaja en lo que ama nunca tendrá que volver a trabajar*…), y el poder aprender un poco cada día del mundo que nos rodea. Por esto, por guiarme en mi carrera investigadora y por el apoyo y la confianza en mi trabajo diario, gracias.

También quiero agradecer a mi co-director de tesis, el Dr. José Manuel Sánchez, por su ayuda inestimable, por enseñarme el poder (y los *significativos peligros*) de la estadística y las matemáticas, y por su incansable buen humor que hace más llevadero todo alrededor.

Gracias también a todo el personal del Centro de Investigación del Cáncer y la Universidad de Salamanca (las -muchas veces- invisibles personas que hacen que toda la normalidad del día a día sea *realmente normal*: administración, conserjería, limpieza, servicio técnico, mantenimiento, y compañeros en general).

También agradecer al magnífico laboratorio del Dr. Julio Sáez-Rodríguez y a todo su enorme equipo, en el *Institute for Computational Biomedicine at the Medical Faculty of Heidelberg University and Heidelberg University Hospital*, por a pesar de estar en medio de la pandemia del siglo haberme acogido y haberme hecho sentir como uno más. *Vielen Dank für Ihre Freundlichkeit.*

¡Cuántos años! ¡Cuánta gente! Cuántos momentos inolvidables…

El laboratorio 19… Cabe una España entera dentro de él, de norte a sur. Gracias a los que están: Fernando, Marina, Natalia, Alberto, Enrique, Diego y Elena (segueix picant igual de fort i trobaràs el teu camí), a los que estuvieron (Curro, Santi, Conrad, Mónica) y a todos los que pasaron alguna vez por nuestro 19 de forma fugaz.

A la gente del laboratorio 17: Alba, Antonio, Óscar, Cris, Luís (la pandemia nos quitó nuestros cafés, pero no nuestras cervezas), a Chema y Carmen, a los ex-17 Sara y Arturo (aún conservo la canción que grabamos: saxo vs guitarra). Las chicas del 4: Helena, Patri, Eva, Elena. A las gentes del CIC, Víctor, Ignacio y mucha más gente que sería imposible mencionar aquí (a cambio os invito a darme un tortazo, ¡sabéis dónde encontrarme!).

A mis amigos de casi toda la vida, ellos saben quienes son.

A Jesús Lacal por sus consejos y su amistad.

Para mi pequeña gran familia:

A mi madre: jamás vi nunca una persona tan luchadora.

A mi padre. *Sé* que estaría orgulloso: con eso es suficiente para seguir sonriendo.

A mi hermana Cristina y Jacob, os quiero más de lo que sé expresar.

A mi tío abuelo, Manolo, por tanto cariño y sabiduría. Por ser ejemplo.

A mi abuela, que me ha visto crecer.

Ya sólo falta un corazón.

Por último, gracias a el corazón que me acompaña por este camino (vida y travesía), el latido que camina junto a mi: Milena.

ANEXO I

Resumen en Español

Tesis Doctoral

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

Análisis bioinformático y uso de deep learning en grandes conjuntos de datos transcriptómicos: estudios de envejecimiento, Alzheimer's, neurodegeneración y cáncer

Óscar González Velasco

SUPERVISORES

Javier De Las Rivas Sanz, Ph.D.

José Manuel Sánchez Santos, Ph.D.

Salamanca, España.

2021

Dr. Javier De Las Rivas Sanz, con D.N.I. 15949000H, Investigador Científico del Consejo Superior de Investigaciones Científicas (CSIC), director del grupo de Bioinformática y Genómica Funcional en el Centro de Investigación del Cáncer (CiC-IBMCC), y profesor del Programa de Doctorado y del Máster de Biología y Clínica del Cáncer de dicho Centro y de la Universidad de Salamanca (USAL).

Y el **Dr. José Manuel Sánchez Santos**, con D.N.I. 07870414K, Profesor Titular de Universidad del Departamento de Estadística, Facultad de Ciencias de la Universidad de Salamanca (USAL).

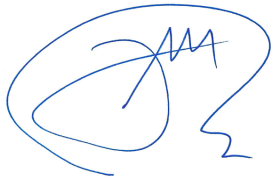
Certifican:

Que han dirigido la Tesis Doctoral titulada "***Bioinformatic analysis and deep learning on large-scale human transcriptomic data: studies on aging, Alzheimer's neurodegeneration and cancer***" realizada por D. **Óscar González Velasco**, dentro del Programa de Doctorado *Biociencias: Biología y Clínica del Cáncer y Medicina Traslacional* del Centro de Investigación del Cáncer (CiC-IBMCC, CSIC/USAL).

Y AUTORIZAN

La presentación de la misma, considerando que reúne las condiciones de originalidad y contenidos requeridos para optar al grado de Doctor por la Universidad de Salamanca.

En Salamanca, a 1 de Abril de 2021



Dr. Javier De Las Rivas Sanz
Director

Dr. José Manuel Sánchez Santos
Codirector

Para la realización de esta Tesis Doctoral, el doctorando Óscar González Velasco obtuvo en concurso público una ***ayuda destinada a financiar la Contratación Predoctoral de Personal Investigador***, cofinanciadas por el **Fondo Social Europeo (FSE)** y convocadas por la **Junta de Castilla y León** (ORDEN EDU/310/2015, de 18 de diciembre de 2017).

Además, la investigación de esta Tesis Doctoral ha sido realizada gracias a los fondos proporcionados a varios Proyectos de Investigación competitivos concedidos al **Grupo del Dr. Javier De Las Rivas** (laboratorio 19) en el **Centro de Investigación del Cáncer (CiC-IBMCC, CSIC/USAL)**. En concreto, se pueden citar los Proyectos Nacionales de la AES del Instituto de Salud Carlos III (ISCiii) PI15/00328 y PI18/00591; y el Proyecto Europeo Horizon 2020 *ArrestAD* ref. 737390 (https://cordis.europa.eu/project/id/737390), iniciado en 2017 que terminará en Diciembre de 2021.

Durante el tiempo de trabajo en esta Tesis Doctoral se logró realizar una **estancia de investigación** durante tres meses (de Enero a Abril de 2021) en Heidelberg (Alemania), gracias a la concesión de una beca **EMBO Short-Term Fellowship** (ref. 8927) para trabajar en el laboratorio dirigido por el **Dr. Julio Saez-Rodriguez** (**Full Professor of Biomedical Informatics and Data Analysis**, https://saezlab.org/) del **Institute for Computational Biomedicine** de la **Medical Faculty,** perteneciente a la **Universidad de Heidelberg** y al **Heidelberg University Hospital**. Como se indica, para la realización de dicha estancia se obtuvo de modo competitivo una beca internacional de la EMBO.

Finalmente, con los méritos anteriores y basado en el trabajo de investigación realizado en los últimos años, esta Tesis Doctoral opta a la **Mención de Doctorado Internacional** otorgado por parte de la **Universidad de Salamanca**, y por ello se presenta escrita en **inglés** en su totalidad, adjuntando también un **resumen en castellano**.

OBJETIVOS GENERALES

Esta tesis doctoral, titulada "***Análisis bioinformático y uso de deep learning en grandes conjuntos de datos transcriptómicos: estudios de envejecimiento, Alzheimer's, neurodegeneración y cáncer***", está enfocada en dos importantes enfermedades muy heterogéneas y complejas: la primera corresponde a la patología neurodegenerativa más prevalente, **la enfermedad de Alzheimer (EA)**, (que incluye un estudio preliminar sobre el deterioro cognitivo del cerebro humano debido al envejecimiento); y la segunda corresponde a cáncer, enfocándose en el análisis de **Cánceres de Origen Primario Desconocido (Cancers of Unknown Primary o CUP)** que representan un grupo heterogéneo de cánceres metastásicos que tienen en común que todos están poco diferenciados y cuyo origen primario es desconocido en el momento del diagnóstico. Como resultado, el tratamiento es muy complicado.

Ambos grupos de enfermedades comparten algunas similitudes: son poligénicas y multicausales, sensibles a múltiples factores externos e internos; son por lo tanto de estructura compleja, con un alto grado de interacciones y posibles reguladores biológicos. Para abordar estos problemas necesitamos desarrollar y aplicar **algoritmos y métodos bioinformáticos robustos para el análisis de grandes cohortes de datos ómicos** humanos derivados de pacientes con estas enfermedades. Postulamos que un análisis y metaanálisis integrativo estadístico y computacional robusto de grandes cohortes (miles de muestras) de diferentes tipos de datos ómicos (principalmente transcriptómica, genómica, proteómica e interactómica) y la búsqueda de relaciones entre entidades biomoleculares (es decir, genes y proteínas) permitirá la identificación de firmas biomoleculares claras y robustas, rutas moleculares asociadas y funciones biológicas que están desreguladas en los pacientes que padecen estas complejas enfermedades.

Resumen de los principales objetivos de la tesis doctoral:

Dentro de este contexto temático, a lo largo de esta Tesis Doctoral se han utilizado y desarrollado diversos métodos bioinformáticos y estadísticos para el análisis, integración y descubrimiento de firmas biomoleculares asociadas a estas enfermedades. Por ello, esta disertación se ha organizado en tres capítulos diferentes dispuestos para abordar los tres OBJETIVOS principales de nuestro trabajo científico, de la siguiente manera:

Objetivo 1: Determinación de la firma transcriptómica asociada con el envejecimiento en el cerebro humano y su relación con la neurodegeneración y el deterioro o deterioro cognitivo.

a. Integración de una gran colección de **muestras transcriptómicas** (obtenidas con microarrays de expresión de alta densidad y con RNA-Seq) de biopsias de cerebro humano sano, provenientes de diferentes estudios sobre las regiones del **hipocampo, la corteza y el cerebelo**, y que cubren distintas edades desde la niñez hasta la senescencia. Diseño de un método bioinformático para descubrir de manera sólida patrones estadísticamente significativos de cambio en la expresión debido al factor de edad dentro de múltiples conjuntos de datos e infiera un patrón común a lo largo del meta-análisis de los resultados obtenidos.

b. Desarrollo de un método de red neuronal de aprendizaje profundo e implementación de un paquete R para calcular la edad biológica (bioage) por individuo en base a la firma transcriptómica obtenida asociada con el envejecimiento en el cerebro humano. Comparación entre la edad biológica y la edad cronológica en diferentes individuos.

Este Objetivo, que incluye una introducción al tema, una descripción detallada de nuestros resultados y una discusión específica, se presenta en el CAPÍTULO I: Panorama transcriptómico del envejecimiento en el cerebro humano.

Objetivo 2: Elaboración de perfiles integradores de datos transcriptómicos de muestras cerebrales de grandes cohortes de pacientes con enfermedad de Alzheimer y búsqueda de nuevos biomarcadores encontrados en sangre.

a. Integración y análisis de una gran colección de muestras transcriptómicas (obtenidas con microarrays de expresión de alta densidad y RNA-Seq) de biopsias de cerebro humano de pacientes con enfermedad de Alzheimer, principalmente de regiones provenientes de hipocampo, corteza y cerebelo. Diseño de un pipeline y framework bioinformático para patrones robustos y significativos de cambio en la expresión génica debido a la enfermedad de Alzheimer, e identificación de un patrón común mediante un meta-análisis de los resultados obtenidos, comparando la firma de Alzheimer con la firma obtenida en el análisis del envejecimiento en cerebro humano.

b. Análisis del perfil transcriptómico de nuevas muestras de sangre generadas de pacientes con Alzheimer y donantes sanos, obtenidas con microarrays y RNA-Seq, utilizando tanto muestras ***bulk RNA-Seq*** como muestras ***single-cell RNA-Seq***, para encontrar firmas de genes clave y

desregulación asignada de manera significativa a la enfermedad de Alzheimer.

Este Objetivo, que incluye una introducción al tema, una descripción detallada de nuestros resultados y una discusión específica, se presenta en el CAPÍTULO II: Firma de expresión génica de la enfermedad de Alzheimer y nuevos biomarcadores de EA encontrados en muestras de sangre. Todo este trabajo se ha llevado a cabo como parte de un esfuerzo científico de colaboración con varios grupos de investigación europeos dentro del Proyecto Europeo Horizonte 2020 ArrestAD (https://cordis.europa.eu/project/id/737390), que todavía está en progreso en al momento de redactar esta tesis (hasta diciembre de 2021) y que no ha publicado sus resultados. Por tal motivo, todos los resultados correspondientes a este capítulo son confidenciales y se mantendrán bajo embargo temporal.

Objetivo 3: Construcción de una herramienta de aprendizaje profundo basada en datos de pan-cáncer y tejidos normales a gran escala para la identificación y predicción del origen primario de muestras tumorales y para el perfil transcriptómico de subtipos de cáncer.

a. Desarrollo e implementación de un modelo predictivo de red neuronal de aprendizaje profundo avanzado (DLNN) con datos pan-cáncer, dirigido al diagnóstico de cánceres de sitio primario desconocido (CUP) (es decir, tumores metastásicos de origen desconocido). Integración de miles de conjuntos de datos transcriptómicos que incluyen muestras de diferentes tipos de cáncer en combinación con datos transcriptómicos de tejidos humanos sanos para entrenar y construir este predictor.

b. Desarrollo de una herramienta de elaboración de perfiles biomoleculares utilizando análisis de redes de genes para evaluar la predicción de la red neuronal y proporcionar más información sobre la firma molecular de muestras de cáncer, así como las vías alteradas y las actividades de factor de transcripción (TF) mejoradas o desreguladas.

Este Objetivo, que incluye una introducción al tema, una descripción detallada de nuestros resultados y una discusión específica, se presenta en el CAPÍTULO III: Herramienta de predicción y elaboración de perfiles de aprendizaje profundo de pancáncer para muestras tumorales basadas en datos transcriptómicos.

CAPÍTULO I

Perfil Transcriptómico del Envejecimiento en el Cerebro Humano

1 SUMARIO DEL CAPÍTULO

Las características biológicas del envejecimiento humano, que conducen a una mayor susceptibilidad a sufrir enfermedades, siguen siendo poco conocidas. A lo largo de este capítulo presentamos un análisis transcriptómico del cerebro humano asociado a la influencia de la edad, dicho análisis proviene la integración de cuatro cohortes independientes de datos de expresión genómica de 2202 muestras de cerebro (de las regiones de cortex, hipocampo y cerebelo) de individuos de diferentes edades (desde niños pequeños de 5 a 10 años, a ancianos de hasta 100 años) categorizados por décadas. El estudio proporciona una firma de 1148 genes detectados en el cortex, 874 genes en el hipocampo y 657 genes en el cerebelo, que presentan cambios de expresión diferenciales significativos con la edad, utilizando la correlación gamma como métrica estadística. Las firmas muestran una gran superposición significativa de 258 genes entre el cortex y el hipocampo, y 63 genes comunes entre las tres regiones del cerebro. Utilizando estas firmas, realizamos un análisis de enriquecimiento funcional y un análisis de actividad de tipos celulares del sistema nervioso central, que proporciona información biológica sobre la compleja firma del envejecimiento.

Finalmente, desarrollamos y entrenamos una red neuronal (Deep Learning), utilizando el descenso de gradiente estocástico como algoritmo de optimización, y los genes obtenidos y que relacionados con la edad como variables de entrada, desarrollando de esa manera un paquete en R con el objetivo principal de construir un predictor de edad biológica basado en la firma de envejecimiento.

2 RESULTADOS Y DISCUSIÓN

2.1 Descripción general de las muestras transcriptómicas de las tres regiones de cerebro usando t-SNE

Para construir un patrón transcriptómico y un perfil de regulación genética del cerebro humano relacionado con el envejecimiento, logramos integrar una gran colección de conjuntos de datos transcriptómicos de muestras post mortem humanas de cerebros de individuos sanos con diferentes edades, normalizados individualmente (**Table 1**).

Antes de comenzar el estudio de los cambios en la expresión génica con la edad, realizamos un análisis de toda la señal transcriptómica en las muestras recogidas en diferentes conjuntos de datos. Con cada uno de los 4 conjuntos de datos recopilados en el compendio (**Table 1**) que incluyen muestras de las tres regiones estudiadas en el cerebro, aplicamos el algoritmo de aprendizaje automático t-SNE para la visualización de cada conjunto de datos (**Figure 5**). El tSNE muestran una clara separación entre las dos regiones del cerebro que incluyen los respectivos conjuntos de datos. Adicionalmente, en los resultados que se muestran en la **Figure 5**, la segregación de los grupos de edad en estos análisis no es tan clara, probablemente porque todos los datos de expresión génica de las muestras revelan la variación de muchos factores posibles.

2.2 Análisis integrativo del patrón de expresión de los genes en cerebro sobre grupos de edad

Estos análisis encontraron genes que presentan una regulación positiva o negativa significativa y consistente con las etapas de edad en los 4 conjuntos de datos independientes. Los análisis se realizaron individualmente, seleccionando las muestras de cada conjunto de datos y cada región del cerebro, y luego combinando los resultados de todos los conjuntos de datos para cada una de las regiones: corteza, hipocampo o cerebelo. De esta manera, los análisis proporcionaron una firma de 1148 genes para la corteza, 874 genes para el hipocampo y 657 genes para el cerebelo. Todos estos genes mostraron una correlación gamma significativa con la edad.

2.3 Perfiles transcriptómicos del cortex cerebral e hipocampo con la edad

Realizamos un análisis completo de las trayectorias de expresión a lo largo de las etapas de edad de cada gen, en cada uno de los 4 conjuntos de datos en las 3 regiones cerebrales estudiadas. Con este enfoque, obtuvimos perfiles (como los presentados en la **¡Error! No se encuentra el origen de la referencia.**), que son independientes para cada c onjunto de datos y cada región del cerebro. Los genes más importantes siguieron dos perfiles principales: uno que aumenta constantemente la expresión durante las etapas progresivas de la vida y otro en el que la expresión disminuye a lo largo de las etapas de la edad. Uno de los genes regulados positivamente con correlación más significativa en los datos del transcriptoma de la corteza es HLA-DPA1 (**¡Error! No se encuentra el origen d e la referencia.**A, B, C, D), que corresponde a una proteína miembro del complejo principal de histocompatibilidad (MHC clase II, DP alfa 1) y que juega un papel central en el sistema inmunológico al presentar péptidos antigénicos derivados de proteínas extracelulares. Nuestro perfil transcriptómico también reveló que varios otros genes HLA se regulaban positivamente con la edad: HLA-DMB, HLA-DPB1, HLA-DPB2 y HLA-DRA. Con respecto a los genes regulados negativamente con la edad, presentamos en la **¡Error! No se encuentra el origen de la referencia.** (E, F, G, H) el perfil en los 4 c onjuntos de datos del gen que mostró la regulación negativa más significativa: la heat shock protein DNAJB5 (miembro de la familia HSP40).

Finalmente, vale la pena indicar que la firma genética encontrada asociada con el envejecimiento en la corteza incluye más genes regulados negativamente (692) que genes regulados positivamente (456), revelando que el gen o efecto genómico del envejecimiento, al menos en la corteza cerebral, parece más inclinado hacia una pérdida de función que hacia una ganancia de función. Sin embargo, esto no puede tomarse como una observación general, ya que los cambios en el hipocampo no siguen esta misma tendencia. Después de este primer análisis, realizamos un análisis similar de los datos obtenidos del hipocampo y el cerebelo. La **Figure 7** presenta los perfiles de expresión producidos para 6 genes del hipocampo que mostraron una alta correlación gamma con las etapas de la edad (utilizando el conjunto de datos del hipocampo con el mayor número de muestras: 121). Nuevamente, el análisis de los perfiles de expresión a lo largo de las 7 décadas mostró algunos genes con una clara tendencia de sobreexpresión, desde jóvenes hasta edades avanzadas, y otro grupo de genes que mostró una represión significativa con la edad.

Datasets	Authors / Reference	Data Public	GEO GSE_ID	Transcriptomic Platform Used	Tissue Type	Total number of samples	0	1	2	3	4	5	6	7	8	9	Number of AGE stages
Dataset 1	**Kang et al.** *Nature (2011)*	**2011**	**GSE25219**	Affymetrix Human Exon 1.0 ST Array	***Cortex*** [1]	**429**	**D0i D0ii** (51) (63)	**D1** (88)	**D23** (99)		**D45** (64)		**D6>** (64)		–	–	**6**
"	"	"	**GSE25219**	Affymetrix Human Exon 1.0 ST Array	***Hippocampus***	**35**	**D0i D0ii** (4) (5)	**D1** (8)	**D23** (7)		**D45** (5)		**D6>** (6)		–	–	**6**
"	"	"	**GSE25219**	Affymetrix Human Exon 1.0 ST Array	***Cerebellum***	**34**	**D0i D0ii** (5) (6)	**D1** (7)	**D23** (8)		**D45** (4)		**D6>** (4)		–	–	**6**
Dataset 2	**Hernandez et al.** *Neurobiol Dis (2012)*	**2012**	**GSE36192**	Illumina HumanHT-12 V3.0 Expression Beadchip Array	***Cortex*** [2]	**453**	**D0** (13)	**D1** (67)	**D2** (44)	**D3** (46)	**D4** (77)	**D5** (56)	**D6** (34)	**D7** (35)	**D8** (52)	**D9** (29)	**10**
"	"	"	**GSE36192**	Illumina HumanHT-12 V3.0 Expression Beadchip Array	***Cerebellum***	**454**	**D0** (13)	**D1** (67)	**D2** (44)	**D3** (47)	**D4** (78)	**D5** (55)	**D6** (34)	**D7** (35)	**D8** (52)	**D9** (29)	**10**
Dataset 3	**Trabzuni et al.** *Nat Commun (2013)*	**2013**	**GSE46706**	Affymetrix Human Exon 1.0 ST Array	***Cortex*** [3]	**374**	–	**D12** (31)		**D3** (26)	**D4** (59)	**D5** (77)	**D6** (71)	**D7** (51)	**D89** (59)		**7**
"	"	"	**GSE46706**	Affymetrix Human Exon 1.0 ST Array	***Hippocampus***	**121**	–	**D12** (11)		**D3** (9)	**D4** (20)	**D5** (27)	**D6** (23)	**D7** (17)	**D89** (14)		**7**
"	"	"	**GSE46706**	Affymetrix Human Exon 1.0 ST Array	***Cerebellum***	**130**	–	**D12** (11)		**D3** (9)	**D4** (21)	**D5** (25)	**D6** (24)	**D7** (19)	**D89** (21)		**7**
Dataset 4	**Berchtold et al.** *Neurobiol Aging (2013)*	**2014**	**GSE48350**	Affymetrix Human Genome U133 Plus 2.0 Array	***Cortex*** [4]	**129**	–	–	**D2** (21)	**D3** (13)	**D4** (25)	**D56** (14)		**D7** (19)	**D8** (17)	**D9** (20)	**7**
"	"	"	**GSE48350**	Affymetrix Human Genome U133 Plus 2.0 Array	***Hippocampus***	**43**	–	–	**D2** (6)	**D3** (4)	**D4** (7)	**D56** (5)		**D7** (5)	**D8** (9)	**D9** (7)	**7**
					TOTAL number of samples	**2202**	**AGE stages in consecutive decades *** (Number of samples in each stage)										

Tabla 12: Conjuntos de datos de expresión utilizados en el perfil transcriptómico integrativo de muestras de cerebro humano de diferentes edades, desde niños hasta personas mayores. La serie de datos originales se descargaron de la plataforma GEO (correspondientes a los ID de GSE: GSE25219; GSE36192; GSE46706; GSE48350).

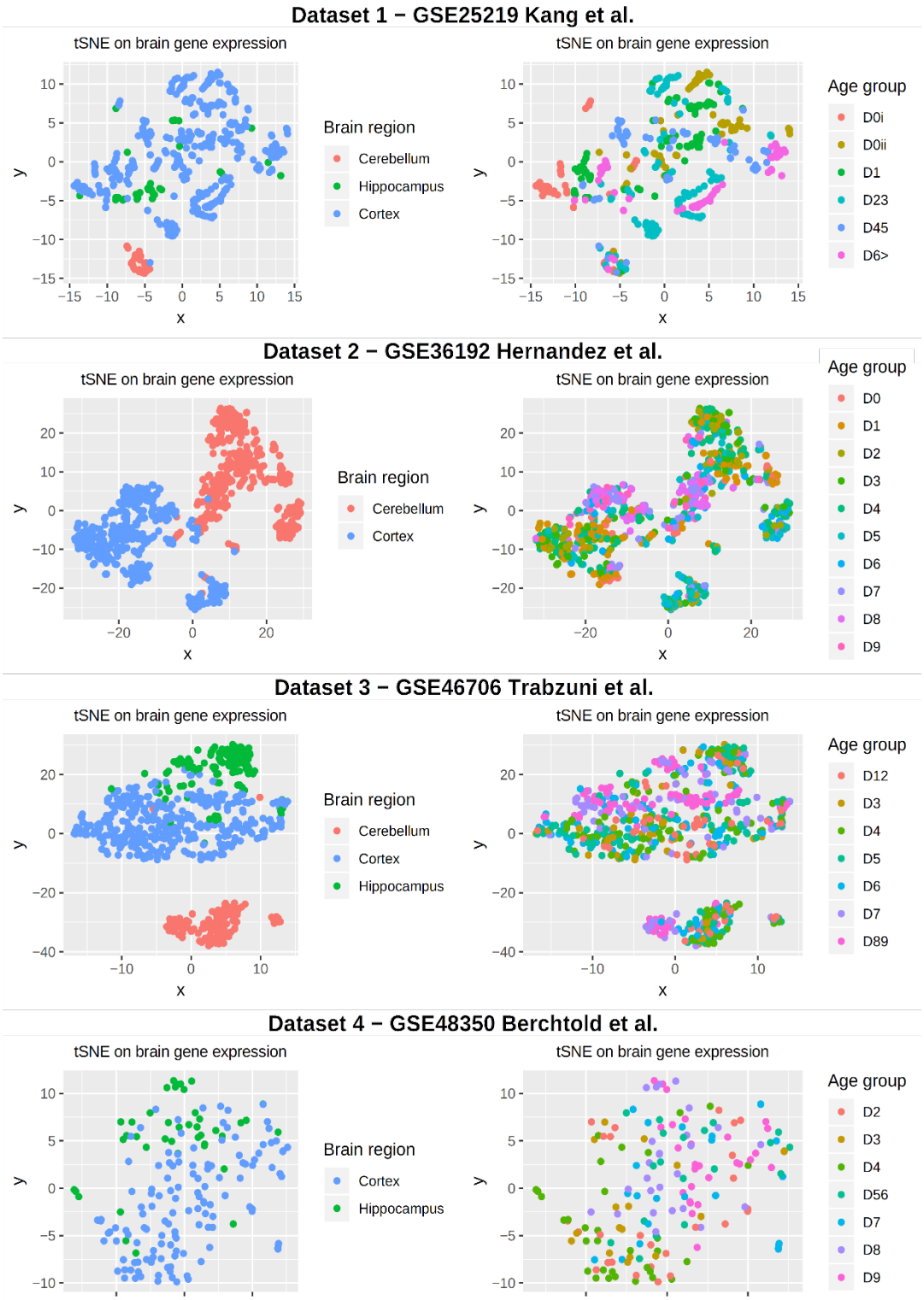


Figura 27: algoritmo tSNE aplicado sobre cada uno de los cuatro conjuntos de datos de envejecimiento; el código de colores en la columna izquierda pertenece a la región del cerebro de donde proviene la muestra, mientras que la columna de la derecha se define por el grupo de edad del individuo. Los resultados generales muestran que la diferencia espacial es más fuerte que la diferencia de envejecimiento a nivel de transcriptoma.

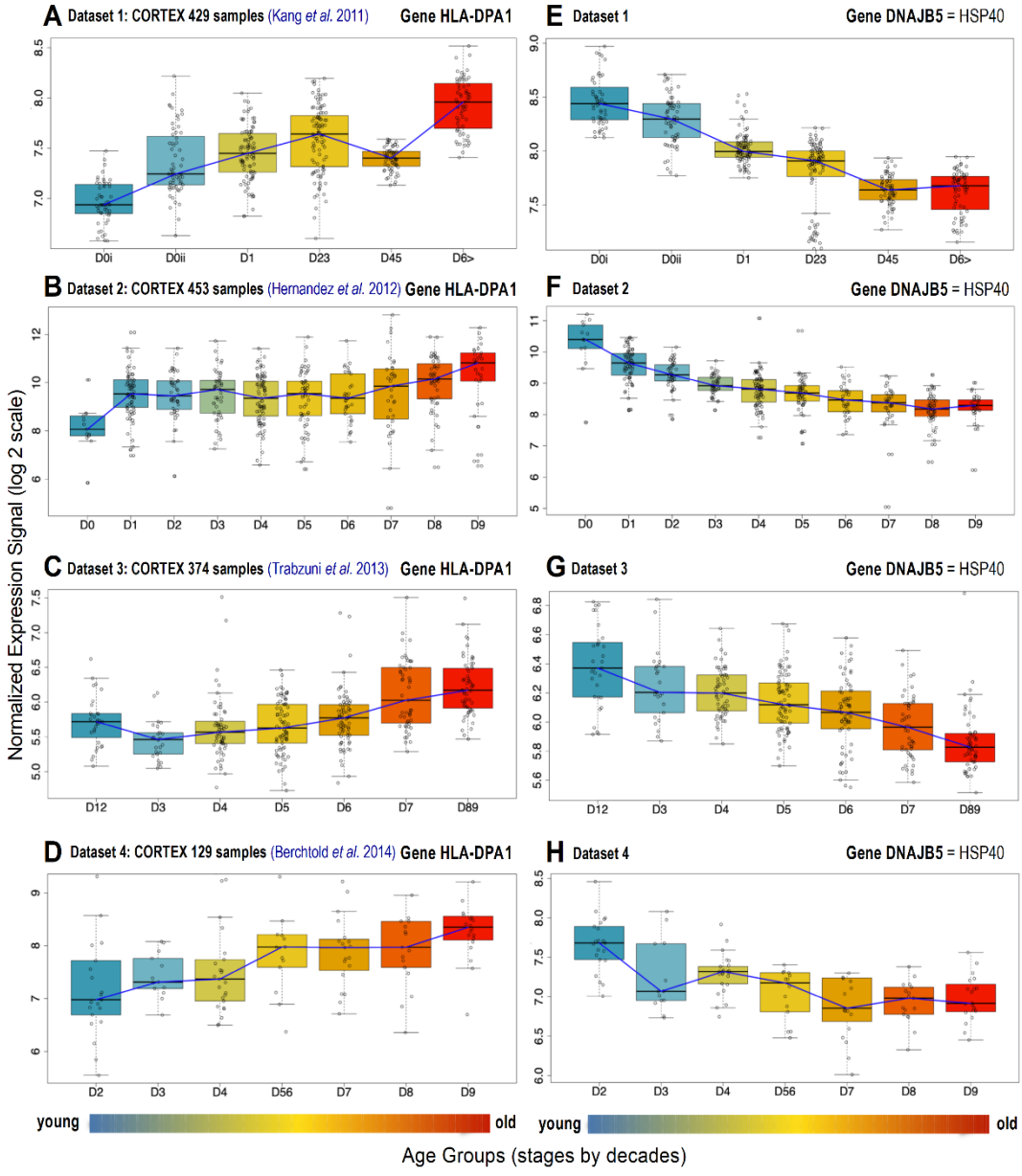


Figura 28: Perfiles de expresión a lo largo de las etapas de la edad en cuatro conjuntos de datos de la corteza cerebral humana: ejemplo de 2 genes (HLA-DPA1 y DNAJB5) que muestran una regulación positiva significativa con el envejecimiento.

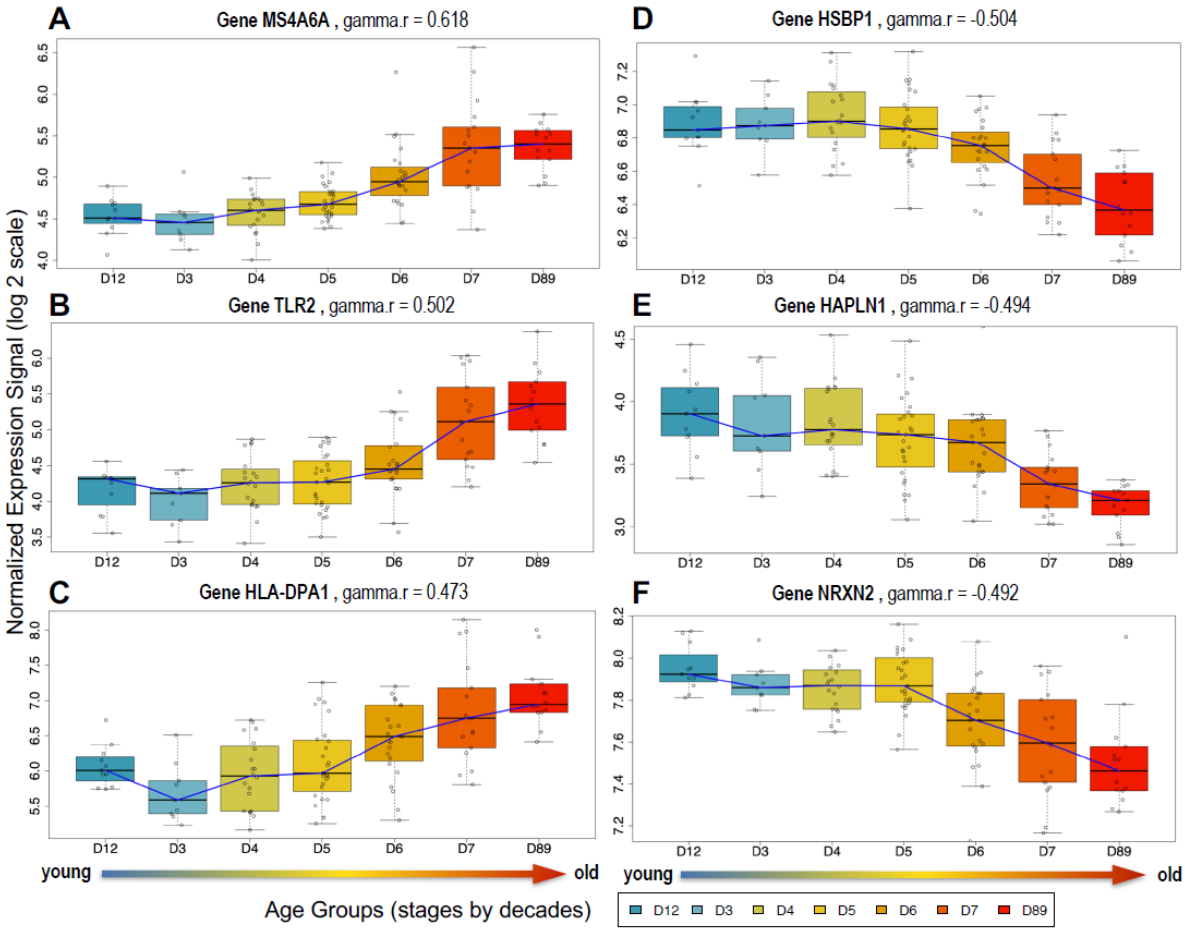


Figura 29: ***Perfiles de expresión génica a lo largo del envejecimiento en el hipocampo humano.*** *Se muestran seis genes en diferentes etapas de edad (desde individuos jóvenes hasta ancianos) en 121 muestras de hipocampo humano (conjunto de datos 3). Los gráficos A, B y C corresponden a 3 genes regulados positivamente: MS4A6A, TLR2 y HLA-DPA1. Los gráficos D, E y F corresponden a 3 genes regulados negativamente: HSBP1, HAPLN1 y NRXN2. El coeficiente gamma r de cada gen se indica en la etiqueta.*

General function	Genes	UP/DOWN regulated	Query List	Reference List	Enrichment p.value	Similarity	Silhouette Width	Functional Terms assigned (concurrent enrichment)
Antigens presentation, Immune system recognition	CD44 , CD74 , CUBN , GBP2 , HLA-DMB , HLA-DPA1 , HLA-DPB1 , HLA-DRA , IRF2 , LRP2 , SLC26A11	UP	11 (251)	100 (34208)	2.12E-10	0.63546	0.30980	GO:0042613:MHC class II protein complex (CC); GO:0005768:endosome (CC); GO:0005765:lysosomal membrane (CC); GO:0002504:antigen processing and presentation of peptide or polysaccharide antigen via MHC class II (BP); GO:0060333:interferon-gamma-mediated signaling pathway (BP)
Response to stress and damage	ACSL5 , AQP1 , BCL2 , CASC3 , CLU , CUBN , GAB1 , IGF1R , LSS , MAOB , MGST1 , MGST2 , RPS27L , SESN1 , SGK1 , SNAP23 , TNFRSF11B , TNKS , TXLNG , TXNIP	UP	20 (251)	631 (34208)	5.48E-08	0.41177	0.25326	GO:0006979:response to oxidative stress (BP); GO:0005792:microsome (CC); GO:0031625:ubiquitin protein ligase binding (MF); GO:0031965:nuclear membrane (CC); GO:0005741:mitochondrial outer membrane (CC); GO:0007584: response to nutrient (BP); GO:0006974:response to DNA damage stimulus (BP)
Cell matrix, Cell surface, Cell adhesion	BCL2 , CAPN2 , CD44 , CD53 , CD74 , CLU , FCER1G , GFAP , HLA-DRA , IGF1R , IL17RB , ITGB4 , LPL , LRP2 , NPC2 , PPFIA1 , SORBS1 , SPP1 , TLR2 , TPP1 , VWF	UP	21 (251)	706 (34208)	7.36E-08	0.28987	0.05617	GO:0007160:cell-matrix adhesion (BP); GO:0009986:cell surface (CC); Kegg:04512:ECM-receptor interaction; GO:0009611:response to wounding (BP); Kegg:04510:Focal adhesion; GO:0005764:lysosome (CC); GO:0009897:external side of plasma membrane (CC)
Immune system activation and reponse	C1QB , C5 , CD59 , CFI , CLU , HLA-DMB , HLA-DPB1 , SERPING1 , VWF	UP	9 (251)	112 (34208)	1.52E-07	0.59192	0.40519	Kegg:04610:Complement and coagulation cascades; Kegg:05322:Systemic lupus erythematosus; Kegg:05150:Staphylococcus aureus infection; GO:0006958:complement activation, classical pathway (BP)
Insulin signaling	ACACB , ACSL5 , ARHGAP24 , EEF2K , GAB1 , IGF1R , IRF2 , LPIN1 , LPP , PHKA2 , PLSCR1 , PYGB , RHOQ , SORBS1 , STOM	UP	15 (251)	424 (34208)	6.79E-07	0.44378	0.33031	GO:0032869:cellular response to insulin stimulus (BP); GO:0008286:insulin receptor signaling pathway (BP); Kegg:04910:Insulin signaling pathway; GO:0005925:focal adhesion (CC); GO:0045121:membrane raft (CC)
Response to pathogens	BCL2 , CD74 , FCER1G , HLA-DMB , HLA-DPB1 , TLR2	UP	6 (251)	113 (34208)	1.93E-04	0.74768	0.54034	Kegg:05152:Tuberculosis; Kegg:05145:Toxoplasmosis
Synapse and neurotransmission (glutamate)	ADRA1D, ADRA2A, ARC, CACNA1G, CAMK2N1, CHRNA2, CHRNA4, DLG3, DLGAP3, ERC2 , GABBR2, GNG4, GRIK3, GRIK4, GRIN2A, GRIN3A, GRM2, HCRTR1, HOMER1, KCNIP1, KCNN1, KCNQ2, MAPK8IP1, NLGN4X, NMU, PRKCB, RIMS1, RIMS4, SEPT5, SHANK2, SLC17A7, SLC1A6, SSTR1, STX1A, SYT5, SYT6, TTYH3, ZNRF1	DOWN	38 (264)	617 (34208)	4.87E-23	0.40280	0.09959	GO:0045202:synapse (CC); GO:0045211:postsynaptic membrane (CC); GO:0005216:ion channel activity (MF); Kegg:04080:Neuroactive ligand-receptor interaction; Kegg:04724:Glutamatergic synapse; GO:0007215:glutamate signaling pathway (BP); GO:0005234:extracellular-glutamate-gated ion channel activity (MF); IPR001508:NMDA receptor; IPR001320:Ionotropic glutamate receptor; IPR019594: Glutamate receptor, L-glutamate/glycine-binding; GO:0043195:terminal button (CC)
Neurons and dendrites	ACP2 , ARC , CAMK2N1 , CXADR , DCX , DLGAP3 , ERC2 , GABBR2 , GRIN2A , GRIN3A , NLGN4X , PTGS2 , SEPT5 , SLC17A7 , SLC32A1 , SLC8A2 , STRN4 , STX1A	DOWN	18 (264)	246 (34208)	9.88E-13	0.51733	0.29743	GO:0043005:neuron projection (CC); GO:0045202:synapse (CC); GO:0019717:synaptosome (CC); GO:0043197:dendritic spine (CC)
Ion channels	CACNA1G , CACNB3 , CACNG4 , KCNF1 , KCNG1 , KCNIP1 , KCNK3 , KCNN1 , KCNQ2 , SLC24A3 , TMEM38A , TNFAIP1	DOWN	12 (264)	125 (34208)	2.83E-10	0.66012	0.64734	GO:0071805:potassium ion transmembrane transport; GO:0005267: potassium channel activity (MF); GO:0005244:voltage-gated ion channel activity (MF); GO:0008076:voltage-gated potassium channel complex (CC); GO:0005249:voltage-gated potassium channel activity (MF)
Cell-cell connection, Endocytosis-phagocytosis, Chemokine signaling	AGAP2 , AMOTL1 , ARC , ASAP1 , ASAP2 , CX3CL1 , CXADR , CXCL14 , DPYSL5 , EXPH5 , GNG4 , GPD1 , HMGCS1 , MAGI1 , MCOLN1 , MMD , MYO16 , PDGFRA , PIP5K1C , PPP2R2C , PRKCB , PRKCZ , PTGS2 , RAB11FIP4 , RASGRP1 , RASSF5 , SMAD3 , STRN4 , STX1A , SYT5 , TNFAIP1 , ZNRF1	DOWN	32 (264)	1157 (34208)	5.10E-10	0.28527	0.09443	Kegg:04144:Endocytosis; GO:0005923:tight junction (CC); GO:0005768:endosome (CC); IPR002219:Protein kinase C-like, phorbol ester/diacylglycerol binding; Kegg:04666:Fc gamma R-mediated phagocytosis; Kegg:04530:Tight junction; GO:0043234:protein complex (CC); Kegg:04062:Chemokine signaling pathway; GO:0005625:soluble fraction (CC)
Kinase signaling, Calcium signaling	ADRA1D , CACNA1G , CACNB3 , CACNG4 , CAMK4 , CCND2 , DUSP14 , DUSP3 , DUSP6 , DUSP7 , FGD1 , GRIN2A , IGF1 , ITPKA , MAPK8IP1 , NR4A1 , PAK6 , PDGFRA , PIP5K1C , PRKCB , RASGRP1 , SLC8A2	DOWN	22 (264)	618 (34208)	3.61E-09	0.28437	0.17669	Kegg:04010:MAPK signaling pathway; GO:0035335:;GO:0004725:protein tyrosine phosphatase activity (MF); IPR000387:Protein-tyrosine/Dual-specificity phosphatase; IPR020422:Dual specificity phosphatase, subgroup, catalytic domain; Kegg:04020:Calcium signaling pathway; Kegg:04510:Focal adhesion; Kegg:04810:Regulation of actin cytoskeleton

Tabla 13 Análisis de enriquecimiento funcional de los mejores 300 genes regulados positivamente y los top 300 regulados negativamente con el envejecimiento en la corteza cerebral humana. El enriquecimiento se realiza mediante el método concurrente (co-ocurrencia) en cinco espacios de anotación (GO-BP, GO-MF, GO-CC, KEGG e INTERPRO) utilizando el método GeneTerm-Linker.

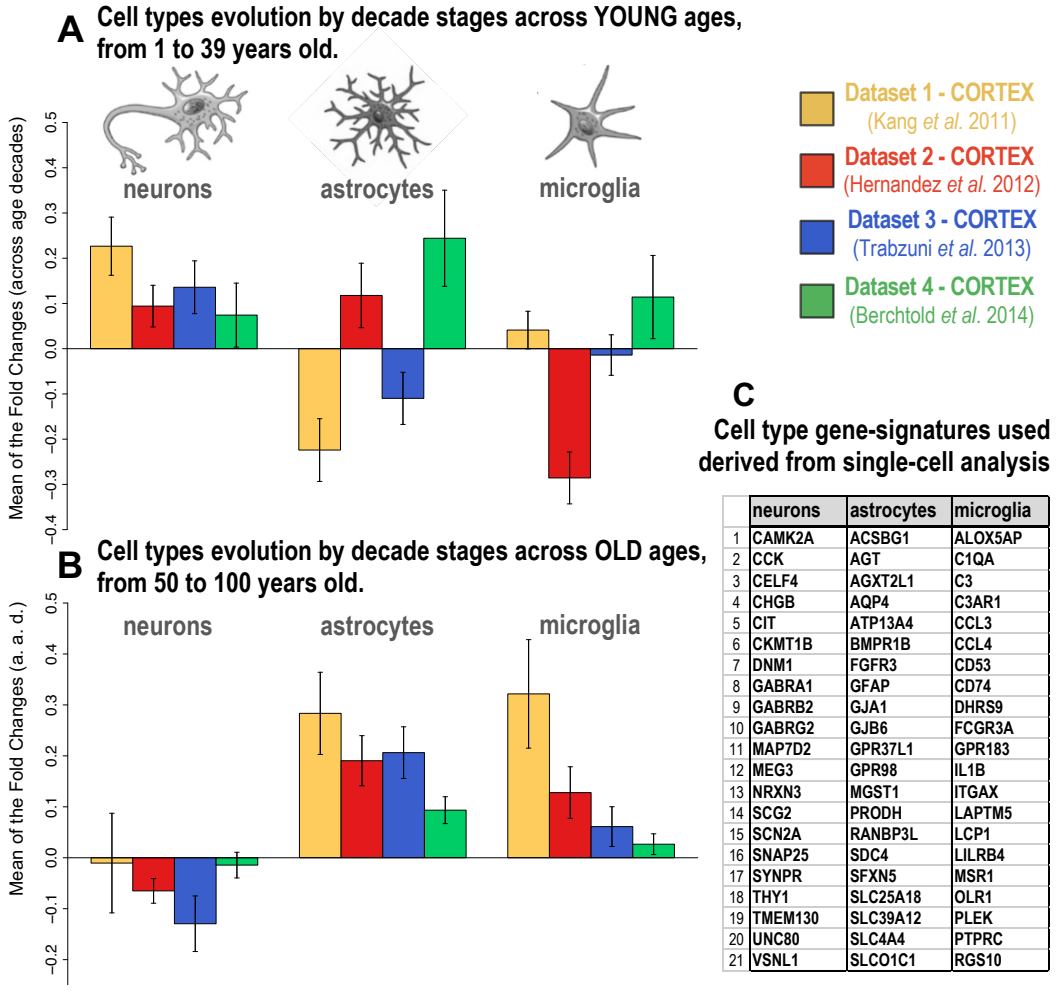
2.4 Pérdida de función neuronal con la edad e incremento de actividad en astrocitos y microglía

Debido a que tenemos cuatro conjuntos de datos independientes, con perfiles transcriptómicos completos de los genes correspondientes a muestras de corteza cerebral de individuos de muchas edades diferentes, desde jóvenes hasta ancianos, podemos analizar el estado y evolución de tipos celulares específicos a lo largo de la vida de los individuos. Para hacer esto, es necesaria la obtención de una firma de genes específicos que puedan marcar tipos de células de una manera selectiva y distinguible.

Utilizando tecnología single cell RNA-Seq, (Darmanis et al., 2015) se identificó un conjunto de marcadores moleculares que permiten clasificar los tipos de células en el cerebro humano. De hecho, estos autores propusieron un conjunto de 21 genes específicos para identificar neuronas, otro conjunto de 21 genes para identificar astrocitos y otro conjunto para identificar microglia. Usamos estos conjuntos de genes para identificar su cambio en los niveles de expresión entre los períodos de edad estudiados; calculamos el Fold Change estadístico promedio de los 21 genes específicos en un tipo de célula en individuos jóvenes en las primeras décadas de vida (de 1 a 39 años) y en individuos de edad avanzada en las últimas décadas de vida (de 50 a 39 años a 100 años de edad). El error estándar de la media (SEM) se calculó de la misma forma para el análisis mencionado.

Los resultados a lo largo del conjunto de edades jóvenes muestran un claro aumento en los genes relacionados con las neuronas (es decir, en la actividad de los genes neuronales), sin cambios consistentes en los niveles de los genes que marcan los astrocitos y la microglía Figura 30. Por el contrario, hay una clara disminución de la señal neuronal (es decir, una represión de los genes neuronales) a lo largo de los últimos períodos de la vida, con un aumento constante de los niveles de los genes asignados a los astrocitos y la microglía. Un aumento en los genes neuronales en las etapas jóvenes de la vida (hasta los 39 años) sugiere que el cerebro incrementa o intensifica su actividad durante los primeros períodos de la vida, desde los niños hasta las personas maduras. Por el contrario, los resultados muestran un declive y deterioro neuronal observado en los últimos períodos de la vida, principalmente después de los 60-65 años. Se realizaron los mismos análisis para el hipocampo utilizando los 3 conjuntos de datos de esta región del cerebro, revelando una evolución muy similar de la señal de las células a través de las etapas de edad, tanto en edades tempranas como en edades avanzadas. Esto confirma

los resultados iniciales, lo que sugiere que los cambios de expresión en la corteza y el hipocampo con el envejecimiento siguen un patrón similar, y que afectan por tanto a un conjunto similar de genes y de funciones biológicas.



C

Cell type gene-signatures used derived from single-cell analysis

	neurons	astrocytes	microglia
1	CAMK2A	ACSBG1	ALOX5AP
2	CCK	AGT	C1QA
3	CELF4	AGXT2L1	C3
4	CHGB	AQP4	C3AR1
5	CIT	ATP13A4	CCL3
6	CKMT1B	BMPR1B	CCL4
7	DNM1	FGFR3	CD53
8	GABRA1	GFAP	CD74
9	GABRB2	GJA1	DHRS9
10	GABRG2	GJB6	FCGR3A
11	MAP7D2	GPR37L1	GPR183
12	MEG3	GPR98	IL1B
13	NRXN3	MGST1	ITGAX
14	SCG2	PRODH	LAPTM5
15	SCN2A	RANBP3L	LCP1
16	SNAP25	SDC4	LILRB4
17	SYNPR	SFXN5	MSR1
18	THY1	SLC25A18	OLR1
19	TMEM130	SLC39A12	PLEK
20	UNC80	SLC4A4	PTPRC
21	VSNL1	SLCO1C1	RGS10

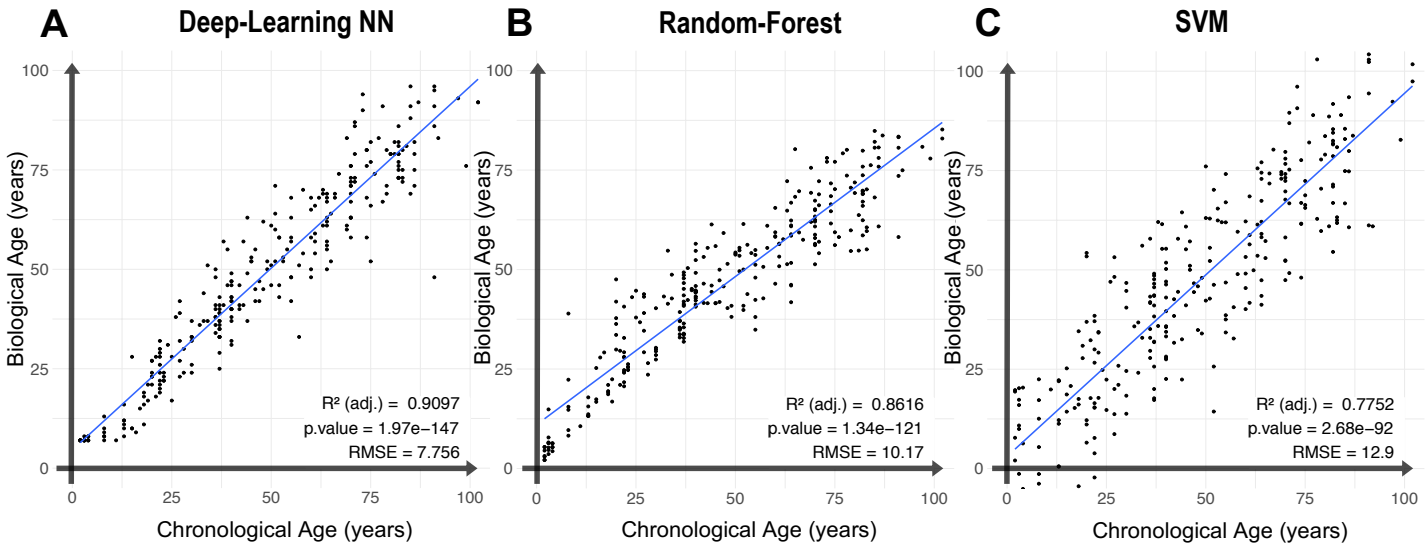
Figura 30: Evolución de la expresión génica específica a un tipo celular en el envejecimiento en la corteza cerebral. Evolución en 4 conjuntos de datos independientes de expresión de ARN de 3 firmas genéticas específicas de tipo celular derivadas de análisis single cell. Estas firmas se utilizan para identificar la presencia de 3 tipos de células (neuronas, astrocitos y microglía) en la corteza cerebral humana a lo largo de diferentes etapas de edad a través de edades tempranas o en edades avanzadas. El gráfico (A) presenta la media de los cambios que ocurren a lo largo de las décadas de edad en individuos jóvenes (de 1 a 39 años); El gráfico (B) presenta la media de los cambios que ocurren a lo largo de las décadas en personas de edad avanzada (de 50 a 100 años). Las listas de genes incluidas en las firmas de genes son derivadas del análisis single cell y que se incluyen en el panel (C).

2.5 El uso de redes neuronales Deep-learning proporciona un cálculo preciso de la edad biológica

Se construyó una red neuronal profunda (DLNN) para calcular la edad individual en función de la información proporcionada por la firma de expresión génica del envejecimiento obtenida en este trabajo y que ha sido explicada en los párrafos anteriores. El rendimiento de la red neuronal se comparó con otros dos algoritmos de machine learning (Random Forest, RF y Support Vector Machines, SVM) aplicando exactamente los mismos datos de entrada para el entrenamiento y la validación (**Figure 12**A, B, C). Los parámetros de regresión proporcionados por cada uno de los tres predictores (DLNN, RF y SVM) indican que nuestra red neuronal (con un R2 ajustado = 0,909) es la que presenta menor error en el cálculo de la edad biológica con respecto a la cronológica.

La lista de los 20 genes principales con el mejor coeficiente r positivo medio, calculado como correlación de Spearman, y los 20 genes principales con el mejor coeficiente r negativo medio, se incluyen en la **Figure 12**D, E. En estas listas marcamos los genes con mejores correlaciones: (i) el gen con mejor correlación positiva fue GMPR (guanosina monofosfato reductasa); y (ii) el mejor con correlación negativo fue DNAJB5 (miembro B5 de Hsp40 de la familia de proteínas heat shock protein DnaJ). También destacamos en estas tablas otros dos genes que presentaron una buena correlación en nuestro estudio, y que han sido previamente relacionados con el envejecimiento cerebral: GFAP (proteína ácida fibrilar glial) (Nichols et al., 1993) y PNOC (prepronociceptina) (Rhinn & Abeliovich, 2017). Curiosamente, las correlaciones con nuestro predictor de bioedad para estos dos genes (GFAP y PNOC) son mejores que las reportadas en los estudios referidos (Rhinn & Abeliovich, 2017).

Finalmente, observamos que la correlación directa entre los genes de la firma de envejecimiento obtenida y la edad cronológica de cada donante fue generalmente peor que la correlación con la edad biológica predicha por nuestro método DLNN. La edad biológica refleja el estado biológico o la situación biológica del cerebro. De esta forma, un desplazamiento negativo con respecto a la edad cronológica (menor bioedad pronosticada) indicaría que el individuo se encuentra en mejores condiciones de salud de lo esperado, y un desplazamiento positivo con respecto a la edad cronológica (mayor bioedad pronosticada) indicaría que el individuo se encuentra en peores condiciones de salud de lo esperado, en este caso refiriéndonos a salud cognitiva.



D Top 20 genes of best positive *Spearman* correlation with **Biological Age**

Gene Symbol ID	Mean Corr with BioAGe	Gene Description
C2CD2	0.60574	C2 calcium dependent domain containing 2 [HGNC:1266]
CD74	0.56796	CD74 molecule [HGNC:1697]
CHI3L1	0.65504	chitinase 3 like 1 [HGNC:1932]
CLU	0.55673	clusterin [HGNC:2095]
DYSF	0.54747	dysferlin [HGNC:3097]
FKBP5	0.64142	FKBP prolyl isomerase 5 [HGNC:3721]
GFAP	0.55495	glial fibrillary acidic protein [HGNC:4235]
GMPR	0.68672	guanosine monophosphate reductase [HGNC:4376]
HLA-DPA1	0.60533	major histocompatibility complex, class II, DPalpha 1 [HGNC:4938]
HLA-DPB1	0.58213	major histocompatibility complex, class II, DPbeta 1 [HGNC:4940]
HLA-DRA	0.62844	major histocompatibility complex, class II, DRalpha [HGNC:4947]
ITGB4	0.58707	integrin subunit beta 4 [HGNC:6158]
ITPKB	0.54751	inositol-trisphosphate 3-kinase B [HGNC:6179]
LPIN1	0.58197	lipin 1 [HGNC:13345]
MAOB	0.56587	monoamine oxidase B [HGNC:6834]
PLPP5	0.55346	phospholipid phosphatase 5 [HGNC:25026]
RCL1	0.54645	RNA terminal phosphate cyclase like 1 [HGNC:17687]
RPS6KA5	0.58891	ribosomal protein S6 kinase A5 [HGNC:10434]
SLC14A1	0.58430	solute carrier fam. 14 member 1 (kidd blood group)[HGNC:10918]
TPP1	0.59927	tripeptidyl peptidase 1 [HGNC:2073]

E Top 20 genes of best negative *Spearman* correlation with **Biological Age**

Gene Symbol ID	Mean Corr with BioAGe	Gene Description
ADGRB2	-0.58784	adhesion G protein-coupled receptor B2 [HGNC:944]
B4GALT2	-0.58908	beta-1,4-galactosyltransferase 2 [HGNC:925]
CACNA1G	-0.59276	calcium voltage-gated channel subunit alpha1 G [HGNC:1394]
CX3CL1	-0.60234	C-X3-C motif chemokine ligand 1 [HGNC:10647]
DNAJB5	-0.70282	DnaJ heat shock protein fam. (Hsp40) member B5 [HGNC:14887]
DPYSL4	-0.58685	dihydropyrimidinase like 4 [HGNC:3016]
EPHB3	-0.59838	EPH receptor B3 [HGNC:3394]
GPR26	-0.60415	G protein-coupled receptor 26 [HGNC:4481]
KIF21B	-0.59067	kinesin family member 21B [HGNC:29442]
MARCH4	-0.66160	membrane associated ring-CH-type finger 4 [HGNC:29269]
NREP	-0.58812	neuronal regeneration related protein [HGNC:16834]
OLFM1	-0.60750	olfactomedin 1 [HGNC:17187]
PNOC	-0.51468	prepronociceptin [HGNC:9163]
RAB11FIP4	-0.62334	RAB11 family interacting protein 4 [HGNC:30267]
RNF165	-0.59137	ring finger protein 165 [HGNC:31696]
SEMA6B	-0.58579	semaphorin 6B [HGNC:10739]
SMPD3	-0.68615	sphingomyelin phosphodiesterase 3 [HGNC:14240]
TMEM8B	-0.60885	transmembrane protein 8B [HGNC:21427]
TRIB2	-0.63471	tribbles pseudokinase 2 [HGNC:30809]
TTC9B	-0.59174	tetratricopeptide repeat domain 9B [HGNC:26395]

Figura 31: (A,B,C) Comparación del rendimiento del modelo de aprendizaje automático con nuestro modelo basado en DLNN. (D) Principales genes cuya señal de expresión de ARNm se correlacionó positivamente con la bioedad predicha. (E) Principales genes cuya señal de expresión de ARNm real se correlacionó negativamente con la bioedad predicha.

3 CONCLUSIONES

El envejecimiento se puede definir en términos generales como la disminución funcional dependiente del tiempo, que afecta a la mayoría de los organismos vivos, caracterizada por una pérdida progresiva de la integridad fisiológica, que conduce a una función deteriorada y una mayor vulnerabilidad a la muerte (López-Otín et al., 2013). Desde una perspectiva más equilibrada, el envejecimiento se puede formular como la colisión entre los procesos destructivos que actúan sobre las células y los órganos durante la vida y las respuestas positivas que promueven la recuperación de la homeostasis, la vitalidad y la longevidad. Sin embargo, a día de hoy, todavía no son bien conocidos los elementos moleculares precisos que marcan el envejecimiento y los mecanismos que determinan las tasas de envejecimiento en los organismos. En este estudio, hemos explorado y analizado la firma transcriptómica y el perfil regulador de los genes en el cerebro humano, relacionando estos cambios con la edad y el envejecimiento. Nuestros resultados revelan la identificación precisa en el cerebro de una firma genética consistente capaz de reflejar el paso del tiempo y el envejecimiento. Esto incluye no solo genes que están reprimidos y que muestran pérdida de funciones biológicas, sino también genes que indican una reacción positiva y una ganancia de función, posiblemente como respuesta al daño, deterioro y estrés que el tiempo impone a nuestro cuerpo. En conjunto, nuestros resultados presentan un paisaje transcriptómico y un perfil regulador genético del cerebro humano vinculado al envejecimiento, proporcionando la identificación de firmas biológicas particulares capaces de caracterizar y predecir la edad biológica del cerebro, que puede diferir de la edad cronológica de los individuos. Los resultados brindan un excelente contexto para comprender mejor el amplio espectro cognitivo observado en la población que envejece de forma sana, y abre una nueva forma de investigar las enfermedades neurodegenerativas, especialmente aquellas para las que actualmente no tenemos una pista causal clara, como es el caso de Enfermedad de Alzheimer.

CAPÍTULO 2

Firma genética de la enfermedad de Alzheimer y nuevos biomarcadores en sangre

1 NOTA SOBRE EL EMBARGO DE ESTE CAPÍTULO

Este capítulo se enmarca en el proyecto europeo "*ArrestAD (3-O-sulfated heparan sulfate translocation in altered membrane biology: A new strategy for early population screening and halting Alzheimer's neurodegeneration)*" (para más información, consultar: https://cordis.europa.eu/project/id/737390) dentro del programa European Horizon 2020. Este proyecto se está realizando internacionalmente y de forma colaborativa entre varios grupos de investigación europeos, siendo el coordinador principal la universidad de Paris XII VAL DE MARNE. Este proyecto tiene una financiación global de casi 4 millones de €.

Este estudio tiene como misión principal la caracterización de nuevos marcadores de la enfermedad de Alzheimer en sangre, así como de la caracterización biomolecular de un fenotipo celular asociado a la enfermedad. Dichos resultados son totalmente novedosos y de especial relevancia, ya que actualmente se desconocen marcadores fiables para dicha enfermedad, además la posibilidad de su detección en sangre puede prometer un gran avance en el tratamiento de la enfermedad. Adicionalmente, los datos muestran la caracterización de un fenotipo celular desconocido hasta ahora, y que podría ayudar a entender mejor las características biomoleculares de la enfermedad. Los resultados más relevantes de este proyecto están aún por publicar (se espera que sean publicados a lo largo del año 2021 en distintos artículos y en revistas científicas de alto impacto).

Por todo ello, y para asegurar la confidencialidad de todos los resultados de dicho proyecto internacional que se lleva desarrollando durante 4 años, se solicitó a la comisión correspondiente la no publicación del capítulo 2.

2 CONCLUSIONES

A modo de resumen, podemos citar las conclusiones globales a las que hemos llegado a lo largo de los diferentes análisis de este capítulo. Estos objetivos son de carácter general y han sido modificados para omitir los detalles confidenciales más sensibles.

Hemos recopilado y normalizado un gran compendio de datos transcriptómicos provenientes de pacientes con Alzheimer y muestras de control, de las regiones cerebrales de cortex e hipocampo. Hemos obtenido una firma genética robusta y distintiva del perfil de expresión molecular de AD para estas regiones, que describe cambios moleculares clave en el cerebro de los enfermos de AD y es capaz de estratificar a los individuos por condición. Hemos podido comparar la firma patológica del Alzheimer con la firma del envejecimiento cerebral saludable, encontrando diferencias importantes entre estos dos procesos.

Hemos analizado muestras de sangre de personas con enfermedad de Alzheimer en el marco y objetivos del Proyecto Europeo ArrestAD y en colaboración con seis grupos de investigación internacionales expertos en la materia. Este trabajo está en progreso y bajo acceso confidencial, pero con nuestros resultados hemos identificado con éxito los factores de riesgo y las características de un nuevo fenotipo específico que se encuentra en las células sanguíneas vinculadas a Alzheimer. También esperamos que este trabajo ayude a identificar nuevos biomarcadores de la enfermedad de Alzheimer de inicio tardío (LO-AD) en las primeras etapas de esta patología neurodegenerativa.

CAPÍTULO 3

Herramienta de predicción y generación de perfiles para tumores pan-cancer mediante uso de deep learning y datos transcriptómicos

1 SUMARIO DEL CAPÍTULO

Hoy en día, el diagnóstico del cáncer está basado en distintas metodologías: desde estudios de detección, pruebas de imagen, muestras de biopsia, pruebas de laboratorio, incluidas pruebas genómicas y transcriptómicas, hasta datos clínicos más rutinarios y registros de antecedentes médicos personales y familiares. Pero a pesar de los enormes avances en esos métodos, algunos casos aún representan un enorme desafío, mostrándose extremadamente difíciles de diagnosticar. Entre estos casos, uno es particularmente relevante: los conocidos como cánceres de origen primario desconocido (Cancers of Unknown Primary o CUP). Este grupo heterogéneo de cánceres metastásicos es particularmente difícil de diagnosticar, ya que debido a la poca especifidad de las metástasis hace completamente imposible el averiguar el tumor primario y el sitio donde se ha originado, y por lo tanto no es posible realizar un diagnóstico o tratamiento específico. Estos casos particulares presentan además un desafío desde el punto de vista del análisis de datos, más específicamente, como un problema de clasificación y predicción.

En este capítulo de la tesis presentamos un modelo de predicción basado en redes neuronales de aprendizaje profundo y datos transcriptómicos, cuyo objetivo es ayudar en el diagnóstico del tumor primario de las muestras de CUP, sirviendo como una herramienta de diagnóstico clínico que es capaz de producir información relevante al tejido donde se originó el cáncer, abriendo de este modo la posibilidad de un tratamiento específico y el uso de medicinas dirigidas.

Hemos recopilado un conjunto de datos que contiene más de 22000 muestras tanto de pacientes con cáncer como de donantes sanos (este superconjunto está compuesto por los estudios genómicos muy comúnmente usados GDC y GTEx), con un total de 27 sitios / tejidos primarios del cuerpo humano mapeados sin ambigüedades. Además, se han realizado pruebas y análisis con varios conjuntos de datos independientes de muestras de cáncer para probar el rendimiento del modelo. Se han creado dos modelos

principales utilizando la biblioteca de software de código abierto TensorFlow para el aprendizaje automático: el primero utiliza redes neuronales convolucionales con datos transcriptómicos sin procesar como entrada; el último, desarrollado durante una estancia doctoral en el Instituto de Biomedicina Computacional de la Facultad de Medicina de la Universidad de Heidelberg y el Hospital Universitario de Heidelberg, utiliza redes neuronales feedforward con actividades biológicas, concretamente de factores de transcripción y rutas de señalización, lo que permite que el modelo sea más simple y ligero, ya que se espera que las actividades contengan más información y sean más estables a lo largo de todas las muestras.

Los resultados muestran un gran potencial para nuestra herramienta de predicción pan-cáncer: ambos modelos lograron una precisión del 97% y 96% en el conjunto de datos de validación, respectivamente, en la predicción del tejido de origen de las muestras. Más importante aún, el modelo de red neuronal convolucional logró una gran precisión en conjuntos de datos externos de tumores primarios y metástasis distantes, pudiendo generalizar los patrones encontrados por sus capas de extracción teniendo excelentes resultados de predicción en escenarios más cercanos a la clínica.

2 RESULTADOS Y DISCUSIÓN

2.1 Explorando la información de actividad biológica

Debido a que la dimensión de los datos se ha reducido considerablemente, ahora es factible explorarlos usando nuevas técnicas y, lo que es más interesante, analizar en qué medida los datos son suficientemente informativos para que el modelo DLNN pueda predecir con éxito el tejido de origen de la muestra.

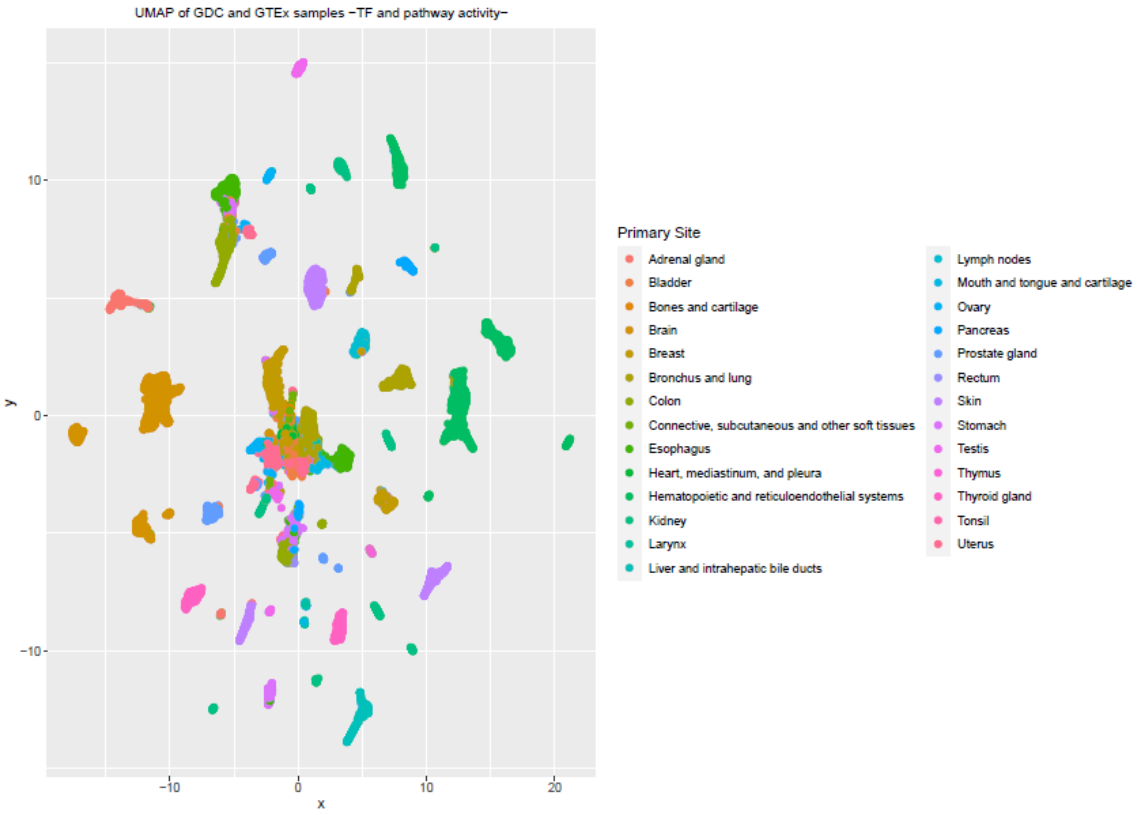


Figura 32: resultados del UMAP sobre los primeros 30 componentes del PCA sobre los factores de transcripción y las actividades de las vías de señalización. Los colores se corresponden con cada uno de los 27 tejidos principales en los que se entrena el modelo.

2.2 Predicción del tejido primario usando Deep Learning

Después de entrenar varios modelos con diferentes configuraciones de capas ocultas, obtenemos los modelos de mejor rendimiento, tanto para la red neuronal convolucional como para la red neuronal feedforward.

2.2.1 Precisión de la red neuronal convolucional

El modelo construido utilizando la bioimagen con factores de transcripción y genes relacionados con cáncer alcanzó una precisión global en la predicción del tejido primario sobre los datos de validación del 97%. Entre las mejores precisiones de clasificación de sitios primarios se encuentran "Mama" con 98,7%, "Glándula prostática" con 99,7%, "Testículo" con 100%, "Sistemas hematopoyéticos y reticuloendotelial" con 99,9%, "Cerebro" con 99,5%, y “Riñón” con 99,3%.

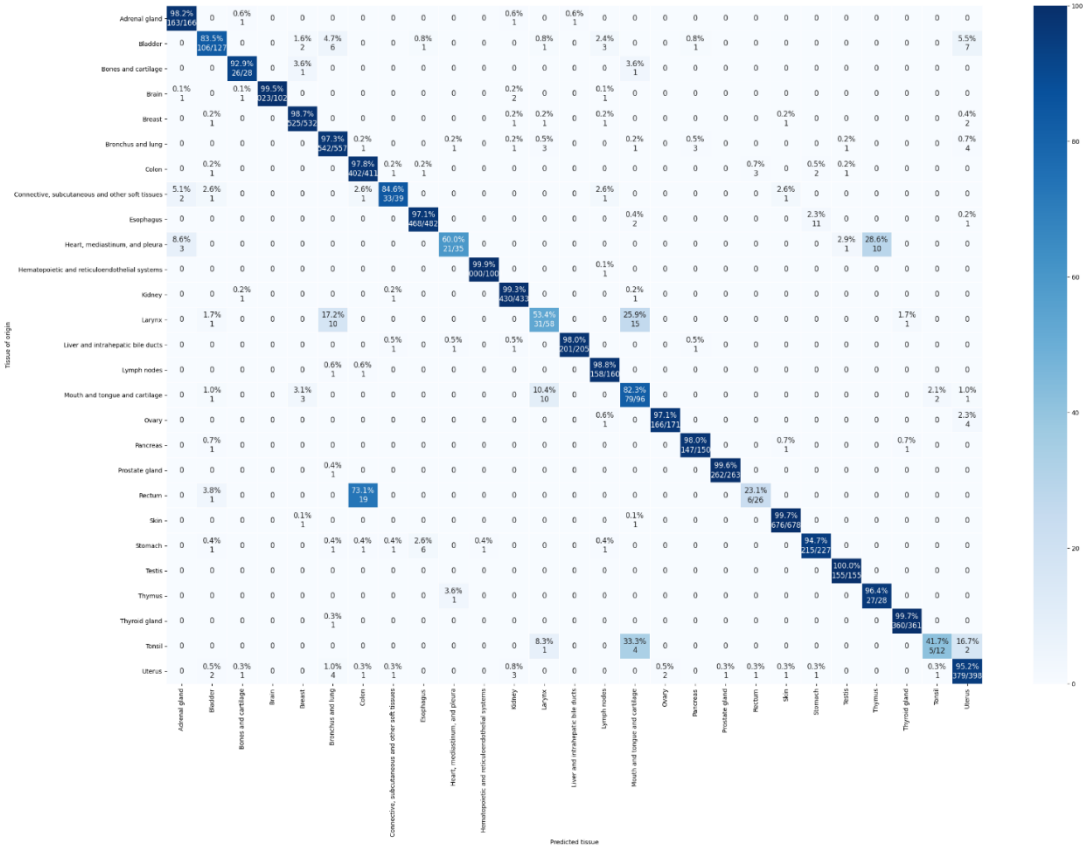


Figura 33: Matriz de confusión de los resultados de la predicción del modelo convolucional en tejidos primarios en las muestras de validación. El eje x corresponde al tejido predicho, el eje y corresponde al tejido original observado de la muestra. Hay un total de 27 tejidos.

2.2.2 Precisión de la red neuronal basada en bioactividad

El modelo con menor complejidad que utiliza las actividades del factor de transcripción derivadas de los regulones DoRothEA y las actividades de las vías de señalización calculadas por Progeny alcanzó una precisión global en la predicción del tejido primario para las muestras de validación del 96%. Aunque la precisión parece ser similar a la obtenida por el modelo convolucional, existen varias diferencias clave en este modelo feedforward, la variabilidad entre las precisiones de los tejidos es mucho mayor, incluso algunas alcanzan una precisión mayor que el modelo convolucional, mientras que varios otros tejidos están incorrectamente etiquetados más del 50% de las veces.

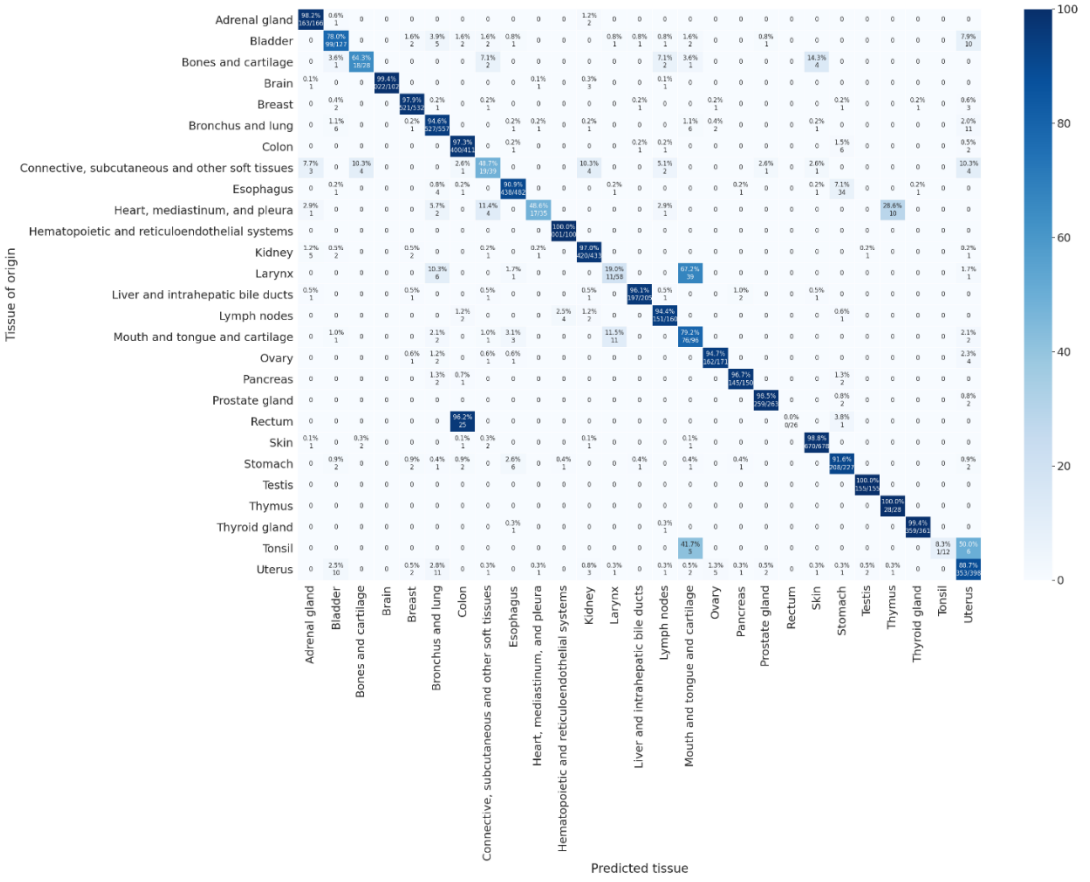


Figure 34: Matriz de confusión de los resultados de predicción del modelo de bioactividad feedforward en tejidos primarios en las muestras de validación. El eje x corresponde al tejido predicho, el eje y corresponde al tejido original observado de la muestra. Hay un total de 27 tejidos. La diagonal de la matriz es la coincidencia exacta entre la predicción y el tejido original.

2.3 Validación con datasets externos

Para validar correctamente nuestros modelos, necesitamos probarlos en conjuntos de datos externos y exponer los algoritmos de aprendizaje profundo a diferentes entornos, ya que son propensos a sobreajustarse, y analizar si se puede lograr la generalización sobre datos reales.

Primary Site – Bulk RNA				
DATASET	**PRIMARY SITE**	**SAMPLES**	**CONVOLUTIONAL ACCURACY**	**BIOACTIVITY ACCURACY**
GSE146009	**Colon – primary site**	35 Caucasian American patients	97%	28%
GSE146009	**Colon – primary site**	30 African American patients	93%	2%

***Tabla 14**: Resultados de la predicción sobre los datasets RNA-Seq correspondientes con cáncer de colon primario.*

El primer conjunto de datos probado corresponde al dataset GSE146009 de cáncer de colon RNA-Seq, con muestras del tumor primario.

Como puede verse en la **Table 6**, el modelo convolucional tiene una precisión sobresaliente, mientras que el DLNN feedforward entrenado con el factor de transcripción y la actividad de la vía no se generaliza en los dos conjuntos de datos.

Posteriormente probamos los modelos de aprendizaje profundo en 3 conjuntos de datos metastásicos externos que incluyen muestras metastásicas distantes. Los 2 primeros se presentan en la **Table 7** y consisten en RNA-Seq de cáncer de riñón (GSE157256) con tumor primario, controles emparejados y metástasis a distancia, y cáncer de ovario (GSE133296) con muestras de tumor metastásico a distancia.

Metastatic – Bulk RNA				
DATASET	**PRIMARY SITE**	**SAMPLES TESTED**	**CONVOLUTIONAL ACCURACY**	**BIOACTIVITY ACCURACY**
GSE157256	**Renal Cancer HLRCC with distant metastasis**	**16 metastatic tumors**.	92%	96%
GSE133296	**Ovarian cancer: omental metastases, and non-omental metastases**	30 metastatic – from 10 patients	93%	0%

***Tabla 15**: Resultados de la predicción sobre los datasets de RNA-Seq de cáncer de riñón y ovario con muestras de metástasis distantes.*

Como puede deducirse de la Table 7, el modelo de aprendizaje profundo convolucional presenta una gran precisión, además de estable a través de diferentes tejidos (92% para cáncer metastásico de riñón y 93% para cáncer metastásico de ovario). Por otro lado, el modelo de aprendizaje profundo mucho más simple y liviano basado en redes neuronales feedforward y que tiene como entrada el factor de transcripción y la actividad de la vía muestra un rendimiento mixto la puntuación con el conjunto de datos metastásicos de Riñón alcanza un 96% de precisión. Desafortunadamente, el modelo no generaliza sobre el conjunto de datos de metástasis ováricas, con una precisión del 0%, lo que significa que la bioactividad carece de información para segregar el tejido ovárico de los otros sitios primarios.

2.4 Selección de variables en el modelo convolucional DLNN

En la siguiente Table 9 presentamos una lista con algunos de los genes más representativos (solo se muestran los que desencadenan una precisión <0) y sus respectivos tejidos de sitio primario: los 3 genes principales que, cuando se establecen

independientemente en 0 CPM, tienen un impacto importante en el rendimiento del modelo.

Gene set to 0 CPM	Gene symbol	Sample Mismatches	Primary site	Accuracy achieved
ENSG00000143578	CREB3L4	525	Breast	0
ENSG00000143614	GATAD2B	525	Breast	0
ENSG00000143622	RIT1	525	Breast	0
ENSG00000142599	RERE	468	Esophagus	0
ENSG00000142611	PRDM16	468	Esophagus	0
ENSG00000142627	EPHA2	468	Esophagus	0
ENSG00000151612	ZNF827	676	Skin	0
ENSG00000151615	POU4F2	676	Skin	0
ENSG00000151623	NR3C2	676	Skin	0
ENSG00000167766	ZNF83	147	Pancreas	0
ENSG00000167771	RCOR2	147	Pancreas	0
ENSG00000167785	ZNF558	147	Pancreas	0
ENSG00000175387	SMAD2	1018	Brain	0.004
ENSG00000175691	ZNF77	1017	Brain	0.005
ENSG00000175395	ZNF25	1016	Brain	0.006

***Tabla 16**: Resultados del análisis de permutación aleatoria de la red neuronal convolucional utilizando 7800 muestras de validación. Cada gen se establece en 0 CPM en cada muestra de validación, después se calcula la precisión de la predicción.*

2.5 Selección de variables en el modelo de bioactividades DLNN

Como tratar de comprender mejor la viabilidad de predecir los sitios primarios de los tumores metastásicos era uno de nuestros principales objetivos, analizamos cuáles de los factores de transcripción y las actividades de las vías de señalización son esenciales para la predicción correcta en metástasis distantes, para esto utilizamos el dataset de cáncer renal con un conjunto de datos de metástasis a distancia (GSE157256). Mediante el uso del análisis de permutación aleatoria, pudimos identificar 3 genes en particular que fueron clave para el modelo DLNN, con el fin de clasificar la metástasis renal distante como proveniente del tejido primario del riñón:

TF	Tissue	Average Accuracy with random permutations
HIF1A	Kidney	63%
MYC	Kidney	33%
SOX2	Kidney	33%

Realizamos otro análisis de permutaciones aleatorias para la actividad del factor de transcripción SOX2: los resultados de este análisis muestran claramente la existencia de un umbral de sensibilidad cuando la actividad del factor de transcripción SOX2 alcanza una puntuación de 15,5, y esa actividad por debajo de este umbral hace que el modelo DLNN etiquete inequívocamente la metástasis distante del riñón como procedente efectivamente del sitio primario del riñón. A partir de este resultado, se puede plantear la hipótesis de que la metástasis renal está relacionada con una baja actividad del factor de transcripción SOX2, por lo que podría actuar como biomarcador de cáncer renal en metástasis distantes.

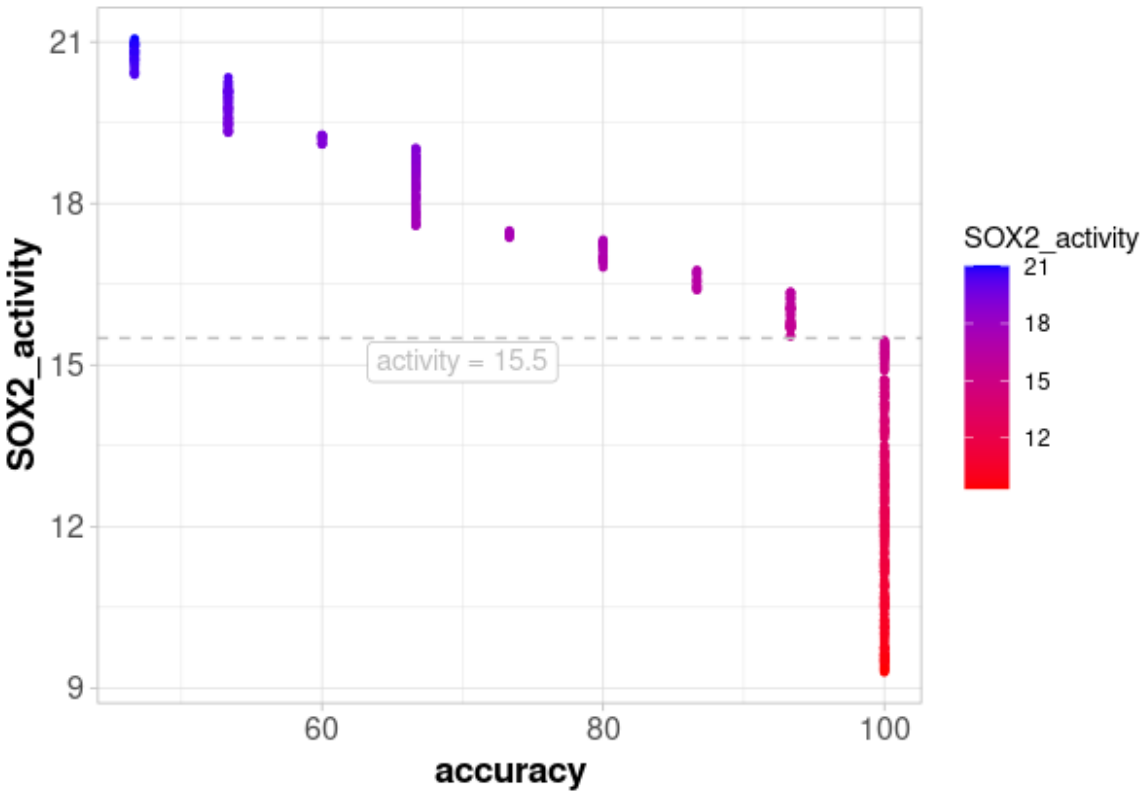


***Figura 35**: gráfico que muestra el rendimiento del modelo DLNN con permutaciones aleatorias de actividad SOX2. El eje y se corresponde con las puntuaciones de actividad de SOX2 generadas aleatoriamente, el eje x muestra la precisión del modelo para etiquetar las muestras del sitio primario como "riñón" con el correspondiente valor permutado de SOX2.*

3 CONCLUSIONES

Los cánceres de origen primario desconocido (CUP) plantean un desafío extremadamente difícil para el futuro del tratamiento y el diagnóstico del cáncer, a medida que se encuentran tratamientos más avanzados y exitosos, los tipos de cáncer más agresivos y difíciles de localizar y estratificar permanecen, y específicamente las metástasis distantes poco diferenciadas. En este capítulo hemos mostramos el uso de redes neuronales de aprendizaje profundo (Deep Learning) como un modelo de predicción del tejido primario de estos cánceres metastásicos. Es importante destacar que, a pesar de la falta de marcadores biomoleculares generales y robustos para los tipos y subtipos de cáncer en muestras metastásicas pobremente diferenciadas, somos capaces de demostrar que es posible entrenar y usar de manera efectiva un modelo de IA para inferir el tejido primario de dicho cancer. Mostramos también la importancia de los factores de transcripción y oncogenes provenientes de la literatura científica, y también la viabilidad e importancia de la actividad de los factores de transcripción y patwhays (calculada utilizando las herramientas DoRothEA y Progeny) para algunos de los tejidos primarios. Con precisiones de 0.97 y 0.96 para el modelo convolucional y el modelo feedforward de bioacticidad respectivamente, demostramos la viabilidad de usar estas nuevas tecnologías dentro del campo diagnóstico y clínico, y la prometedora revolución que la IA podría traer a la medicina personalizada, dando resultados muy específicos para cada paciente haciendo uso de grandes cantidades de datos.

Además, mostramos el potencial de estos modelos para comprender mejor la biología y los marcadores que son cruciales para este tipo de diagnóstico, no realizando únicamente la predicción de la muestra, sino también para producir nuevos conocimientos sobre el problema que estamos estudiando. Destacable es el caso de los conjuntos de datos metastásicos distantes: como mostramos, los modelos son capaces de generalizar al enfrentarse a datos nuevos y externos, incluso a pesar de que ambos modelos han sido entrenados con conjuntos de datos principalmente de muestras de tejidos primarios, incluyendo prácticamente ninguna muestra metastásica de la que aprender. Particularmente interesante es el caso del análisis de permutación aleatoria de la actividad del factor de transcripción SOX2, que vincula la metástasis renal con una baja actividad de SOX2.

ANEXO II

Nota Sobre el Embargo del Capítulo II

NOTA SOBRE EL EMBARGO DEL CAPÍTULO II:

Alzheimer disease gene signature and new blood biomarkers

(Firma genética de la enfermedad de Alzheimer y nuevos biomarcadores en sangre)

Este capítulo se enmarca en el proyecto europeo "*ArrestAD (3-O-sulfated heparan sulfate translocation in altered membrane biology: A new strategy for early population screening and halting Alzheimer's neurodegeneration)*" (para más información, consultar: https://cordis.europa.eu/project/id/737390) dentro del programa European Horizon 2020. Este proyecto se está realizando internacionalmente y de forma colaborativa entre varios grupos de investigación europeos, siendo el coordinador principal la universidad de Paris XII VAL DE MARNE. Este proyecto tiene una financiación global de casi 4 millones de €.

Este estudio tiene como misión principal la caracterización de nuevos marcadores de la enfermedad de Alzheimer en sangre, así como de la caracterización biomolecular de un fenotipo celular asociado a la enfermedad. Dichos resultados son totalmente novedosos y de especial relevancia, ya que actualmente se desconocen marcadores fiables para dicha enfermedad, además la posibilidad de su detección en sangre puede prometer un gran avance en el tratamiento de la enfermedad. Adicionalmente, los datos muestran la caracterización de un fenotipo celular desconocido hasta ahora, y que podría ayudar a entender mejor las características biomoleculares de la enfermedad. Los resultados más relevantes de este proyecto están aún por publicar (se espera que sean publicados a lo largo del año 2021 en distintos artículos y en revistas científicas de alto impacto).

Por todo ello, y para asegurar la confidencialidad de todos los resultados de dicho proyecto internacional que se lleva desarrollando durante 4 años, se solicitó a la comisión correspondiente la no publicación del capítulo 2.