

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives.

Jordi Dil Giró

Màster Universitari en Ciència de Dades
Medicina

Erola Pairó Castiñeira
Ferran Prados Carrasco

6 de Juny de 2021

© (Jordi Dil Giró)

Reservats tots els drets. Està prohibit la reproducció total o parcial d'aquesta obra per qualsevol mitjà o procediment, compresos la impressió, la reprografia, el microfilm, el tractament informàtic o qualsevol altre sistema, així com la distribució d'exemplars mitjançant lloguer i préstec, sense l'autorització escrita de l'autor o dels límits que autoritzi la Llei de Propietat Intel•lectual.

FITXA DEL TREBALL FINAL

Títol del treball:	<i>Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives.</i>
Nom de l'autor:	<i>Jordi Dil Giró</i>
Nom del consultor/a:	<i>Erola Pairó Castiñeira</i>
Nom del PRA:	<i>Ferran Prados Carrasco</i>
Data de lliurament (mm/aaaa):	<i>06/2021</i>
Titulació o programa:	<i>Màster Universitari en Ciència de Dades</i>
Àrea del Treball Final:	<i>Medicina</i>
Idioma del treball:	<i>Català</i>
Paraules clau	<i>Malalties neurodegeneratives, Expressió genètica, Alzheimer</i>

Resum del Treball

Milions de persones en el món pateixen malalties neurodegeneratives que actualment no tenen cura, i arran de l'augment de l'esperança de vida es preveu un increment en la seva morbilitat i en el nombre de morts degut a aquestes malalties. Això suposa un gran cost tant per les famílies com pel sistema sanitari.

La recerca de teràpies efectives és difícil per la seva complexitat, per la dificultat d'estudiar el cervell i el desconeixement que encara en tenim.

Cada part del cos té funcions especialitzades donades per proteïnes codificades a partir de l'expressió de gens. Tenint present que s'han publicat diversos articles científics que relacionen les malalties neurodegeneratives amb determinats gens i aquesta expressió diferencial per parts del cos, una de les aproximacions en l'estudi d'aquestes malalties és l'expressió genètica centrada en les diferents parts del cervell.

En aquest context el projecte té com objectiu estudiar l'expressió genètica cerebral i relacionar-la amb malalties neurodegeneratives a partir de l'estudi de en quines parts del cervell s'hi expressen els gens associats a les mateixes.

Això ens pot apropar a crear xarxes genètiques específiques per cadascuna de les malalties i desenvolupar millors teràpies per curar-les.

Els resultats obtinguts diferencien els teixits del cervell a partir de l'expressió genètica d'un nombre reduït de gens. En relació a la recerca de quines parts del cervell es veuen més afectades per una malaltia neurodegenerativa, en el cas de l'Alzheimer, s'han trobat sobre expressions rellevants de gens en els teixits corticals.

Abstract

Millions of people around the world suffer from neurodegenerative diseases (ND) which have no known cure, and as a result of increased life expectancy, their morbidity and the number of deaths due to these diseases are expected to increase. This is a great cost for both families and the health care system.

The research for effective therapies is difficult because of the complexity of these diseases, the difficulty of studying the brain, and the lack of knowledge we still have about them.

Each part of the body has specialized functions given by proteins encoded from gene expression. Bearing in mind that several scientific articles have been published linking neurodegenerative diseases with certain genes and this differential expression by body parts, one of the approaches to study these diseases is the genetic expression focused on the different parts of the brain.

In this context, the project aims to study the genetic expression of the brain and relate it to neurodegenerative diseases ..

The knowledge generated from this study will help us create tissue and disease specific genetic networks which could be an important step towards better treatment of ND diseases.

The results obtained differentiate brain tissues from the genetic expression of a small number of genes. In relation to the search about which parts of the brain are most affected by a neurodegenerative disease, in the case of Alzheimer's, they have been found relevant over gene expressions in the cortical tissues.

Índex

1. Introducció	6
1.1. Context i justificació del Treball.....	6
1.1.1. Impacte socioeconòmic de les malalties neurodegeneratives a Espanya	6
1.1.2. Dogma central de la biologia molecular.....	7
1.1.3. Estat de l'Art.....	8
1.1.4. Motivació	10
1.2. Objectius del Treball	11
1.3. Enfocament i mètode seguit	11
1.4. Planificació del Treball.....	12
1.4.1. Tasques	12
1.4.2. Calendari.....	15
1.5. Breu sumari de productes obtinguts.....	15
1.6. Breu descripció dels altres capítols de la memòria	15
2. Materials i Mètodes	16
2.1. Dades	19
2.2. Anàlisi Descriptiu	20
2.3. Diferenciació de Teixits.....	26
2.4. Anàlisi d'Expressió.....	29
3. Resultats	31
3.1. Diferenciació de Teixits.....	31
3.2. Anàlisi d'Expressió.....	42
4. Discussió	50
5. Conclusions	52
6. Glossari	53
7. Bibliografia	54
8. Annexes.....	57

Llista de figures

- Figura 1** Dogma central de la bioquímica molecular amb enzims
- Figura 2** Diagrama de Gantt amb la planificació del projecte
- Figura 3** Mostra del fitxer GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct
- Figura 4** Gràfica de la mitjana vs. la variància obtinguda a partir del conjunt de dades dels teixits del cervell
- Figura 5** Histograma de la distribució de gens en el conjunt de teixits
- Figura 6** Histograma de la distribució de gens en teixits de l'*Amygdala*
- Figura 7** Histograma de la distribució de gens en teixits de l'*Anterior*
- Figura 8** Histograma de la distribució de gens en teixits *Caudate*
- Figura 9** Histograma de la distribució de gens en teixits del *Cerebellar*
- Figura 10** Histograma de gens en teixits del *Cerebellum*
- Figura 11** Histograma de la distribució gens en teixits del *Cortex*
- Figura 12** Histograma de la distribució de gens en teixits del *Frontal*
- Figura 13** Histograma de la distribució de gens en teixits de l'*Hippocampus*
- Figura 14** Histograma de la distribució de gens en teixits de l'*Hypothalamus*
- Figura 15** Histograma de la distribució en log2 de gens en el teixits del *Nucleus*
- Figura 16** Histograma de la distribució de gens en teixits del *Putamen*
- Figura 17** Histograma de gens en teixits de l'*Spinal*
- Figura 18** Histograma de gens en teixits de *Substantia*
- Figura 19** Histograma de la distribució en log2 de gens en el Total de Teixits en que mitjana(TPM)>1
- Figura 20** Histograma de la distribució en log2 de gens en el Teixit Amygdala en que mitjana(TPM)>1
- Figura 21** Número de Gens Expressats, en mitjana(TPM)>1, per Teixit
- Figura 22** Flux seguit per la realització dels clústers
- Figura 23** Gens TPM. Un total de 25 components expliquen un 95% de la variància
- Figura 24** Visualització del Primer Component Principal vs el Segon Component Principal
- Figura 25** Visualització del Primer Component Principal vs el Tercer Component Principal
- Figura 26** Representació 2D emprant t-SNE (*perplexity*=30)
- Figura 27** Resultats de la matriu de confusió pel model SVM i els 1000 gens més expressats
- Figura 28** Gens codificants. Selecció dels 21 components principals explicatius d'un 95% de la variància
- Figura 29** Gens codificants. Visualització del Primer Component Principal vs el Segon Component Principal
- Figura 30** Gens codificants. Visualització del Primer Component Principal vs el Tercer Component Principal
- Figura 31** Gens codificants. Representació 2D emprant t-SNE (*perplexity*=30)
- Figura 32** Gens no mitocondrials. Selecció dels 24 components principals explicatius d'un 95% de la variància.
- Figura 33** Gens codificants. Visualització del Primer Component Principal vs el Segon Component Principal

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

- Figura 34** Gens codificants. Visualització del Primer Component Principal vs el Tercer Component Principal
- Figura 35** Gens codificants. Representació 2D emprant t-SNE (perplexity=30)
- Figura 36** Heatmap a partir dels gens sobre expressats per Teixit lligats a l'Alzheimer tenint en compte la suma de diferències entre la mitjana del teixit i la mitjana de la resta
- Figura 37** Heatmap de la sobre expressió dels Top 1000 gens del transcriptoma tenint en compte la suma de diferències entre la mitjana del teixit i la mitjana de la resta

Llista de taules

- Taula 1** Despeses mitjanes de les malalties neurodegeneratives a Espanya l'any 2015
- Taula 2** Prevalença de les principals malalties neurodegeneratives a Espanya l'any 2015
- Taula 3** Defuncions segons les causes de mort més freqüents
- Taula 4** GTExv8. Contingut de mostres, teixits i donants d'on provenen
- Taula 5** Fitxer amb tots els gens RAW lligats al cervell
- Taula 6** Dataset resultant amb 17515 gens en que la seva mitjana(TPM)>1
- Taula 7** Relació de camps destacats en el fitxer que recull el catàleg GWAS
- Taula 8** Relació de Teixits del Cervell i número de mostres de cadascun d'ells
- Taula 9** Precisió obtinguda amb el model kNN emprant els 1000 gens més expressats
- Taula 10** Precisió obtinguda amb el model SVM emprant els 1000 gens més expressats
- Taula 11** Precisió obtinguda amb el model Arbre de Decisió emprant els 1000 gens més expressats
- Taula 12** Precisió obtinguda amb el model *Random Forest* emprant els 1000 gens més expressats
- Taula 13** Comparativa de resultats obtinguts amb els diferents models amb els 1000 gens més expressats
- Taula 14** Gens Codificants: Precisió obtinguda amb el model kNN emprant els 1000 gens més expressats
- Taula 15** Gens Codificants: Precisió obtinguda amb el model SVM emprant els 1000 gens més expressats
- Taula 16** Gens Codificants: Precisió obtinguda amb el model Arbre de Decisió emprant els 1000 gens més expressats
- Taula 17** Gens Codificants: Precisió obtinguda amb el model Random Forest emprant els 1000 gens més expressats
- Taula 18** Gens Codificants: Comparativa de resultats obtinguts amb els diferents models amb els 1000 gens més expressats
- Taula 19** Gens No Mitocondrials: Precisió obtinguda amb el model kNN emprant els 1000 gens més expressats
- Taula 20** Gens No Mitocondrials: Precisió obtinguda amb el model SVM emprant els 1000 gens més expressats
- Taula 21** Gens No Mitocondrials: Precisió obtinguda amb el model Arbre de Decisió emprant els 1000 gens més expressats
- Taula 22** Gens No Mitocondrials: Precisió obtinguda amb el model Random Forest emprant els 1000 gens més expressats
- Taula 23** Gens No Mitocondrials: Comparativa de resultats obtinguts amb els diferents models amb els 1000 gens més expressats
- Taula 24** Precisions dels models obtingudes pels diferents conjunts de dades i diferent nombre de gens
- Taula 25** Número de gens lligats a l'Alzheimer sobre expressats per teixit
- Taula 26** Número de gens lligats a l'Alzheimer infra expressats per teixit

- Taula 27** Rànking de gens sobre expressats lligats a l'Alzheimer obtinguts a partir de la Suma de la diferència entre la mitjana del teixit i la mitjana de la resta
- Taula 28** Rànking de gens infra expressats lligats a l'Alzheimer obtinguts a partir de la Suma de la diferencia entre la mitjana del teixit i la mitjana de la resta.
- Taula 29** Rànking de la sobre expressió dels Top 1000 gens del transcriptoma sobre expressats per Teixit tenint en compte el seu número
- Taula 30** Rànking de la infra expressió dels Top 1000 gens del transcriptoma infra expressats per Teixit tenint en compte el seu número
- Taula 31** Rànking de la sobre expressió dels Top 1000 gens del transcriptoma sobre expressats per Teixit tenint en compte la suma de diferències entre la mitjana del teixit i la mitjana de la resta
- Taula 32** Rànking de la infra expressió dels Top 1000 gens del transcriptoma infra expressats per Teixit tenint en compte la suma de diferències entre la mitjana del teixit i la mitjana de la resta
- Taula 33** Número de casos en que el número de Gens sobre expressats respecte els GWAS és major després de 10000 iteracions

1. Introducció

1.1. Context i justificació del Treball

1.1.1. Impacte socioeconòmic de les malalties neurodegeneratives a Espanya

Milions de persones en el món pateixen malalties neurodegeneratives que actualment no tenen cura, i arran de l'augment de l'esperança de vida es preveu un increment en la seva morbilitat i mortalitat. Això suposa un gran cost tant per les famílies com pel sistema sanitari.

Garcés (2016) en el seu estudi sobre les malalties neurodegeneratives a Espanya quantificava la seva despesa total l'any 2015 al voltant dels 30.000 milions d'euros amb un cost anual mig directament suportat per les persones afectades per aquestes malalties i les seves famílies superior al 23.000€ per pacient.

Enfermedad	Afectados Número (Estimado)	Coste por paciente				Costes totales ESPAÑA			
		Directos médicos	Directos no médicos	Indirectos	Total	Directos médicos	Directos no médicos	Indirectos	Total
		Personas	(euros)		(millones de Euros)				
Alzheimer y Demencias	717.000	5.348	1.237	22.597	29.182	3.835	887	16.202	20.923
Enfermedad de Parkinson	160.000	3.988	3.325	11.487	18.800	638	532	1.838	3.008
Esclerosis Múltiple	47.000	28.964	12.370	14.252	55.586	1.361	581	670	2.613
Enf. Neuromusculares	60.000	13.829	79.312	1.030	94.171	830	4.759	62	5.650
Esc. Lat. Amiotrófica (ELA)	4.000	8.289	27.619	8.575	44.483	33	110	34	178
Totales	988.000					6.697	6.870	18.806	32.372

Taula 1: Despeses mitjanes de les malalties neurodegeneratives a Espanya l'any 2015. Font: Dades d'elaboració pròpies de l'autor a partir de diverses fonts (Garcés, 2016).

El mateix estudi situava la prevalença d'aquest tipus de malalties en el 2,08% de la població amb un total de 988.000 persones afectades.

Enfermedad	Prevalencia global	Población afectada
Alzheimer y otras demencias	1,53%	717.000
Enfermedad de Parkinson	0,34%	160.000
Esclerosis Múltiple	0,08%	47.000
Enfermedades Neuromusculares	0,12%	60.000
Esclerosis Lateral Amiotrófica (ELA)	0,008%	4.000
TOTAL AFECTADOS	2,08%	988.000

Taula 2: Prevalença de les principals malalties neurodegeneratives a Espanya l'any 2015. Font: Dades d'elaboració pròpies de l'autor a partir de diverses fonts (Garcés, 2016).

Consultant dades de l'INE (2018) s'observa com la demència i la malaltia d'Alzheimer (EA) són la quarta i sisena causa de mort més freqüent amb un total de 21.629 i 14.929 persones, respectivament.

Defunciones según las causas de muerte más frecuentes ¹ . Año 2018			
	Total	Hombres	Mujeres
Total enfermedades	427.721	216.442	211.279
Enfermedades isquémicas del corazón	31.152	18.423	12.729
Enfermedades cerebrovasculares	26.420	11.435	14.985
Cáncer de bronquios y pulmón	22.133	17.181	4.952
Demencia	21.629	7.144	14.485
Insuficiencia cardíaca	19.142	7.266	11.876
Enfermedad de Alzheimer	14.929	4.454	10.475
Enf. crónicas de las vías respiratorias inferiores (ECVRI)	14.607	10.594	4.013
Enfermedad hipertensiva	12.496	4.108	8.388
Cáncer de colon	11.265	6.690	4.575
Neumonía	10.415	5.430	4.985
Diabetes mellitus	9.921	4.407	5.514
Cáncer de páncreas	7.132	3.299	3.833
Insuficiencia renal	7.120	3.745	3.375
Cáncer de mama	6.621	87	6.534
Cáncer de próstata	5.841	5.841	0

¹Causas con peso relativo superior al 1,4%.

Taula 3: Defuncions segons les causes de mort més freqüents. Font: INE(2018)

Totes aquestes malalties són de difícil estudi doncs el seu diagnòstic ve donat per metodologies indirectes, i només es poden diagnosticar de manera concluent postmortem, a partir d'un examen del cervell.

1.1.2. Dogma central de la biologia molecular

La informació genètica és la mateixa en totes les cèl·lules. El que fa que cada cèl·lula funcioni i sigui diferent en els teixits i òrgans és l'expressió o el funcionament diferencial dels gens.

L'ADN (àcid desoxiribonucleic), és el material genètic de tots els organismes i es divideix en unitats funcionals anomenats gens. Els gens contenen la informació per crear els productes necessaris pel funcionament de les cèl·lules, principalment proteïnes.

El procés a partir del qual es codifica una proteïna es coneix com el dogma central de la biologia molecular, i consta de 2 passos que són la transcripció i la traducció.

La molècula d'ADN pot replicar-se de tal manera que a partir d'una molècula d'ADN en podem tenir una altra igual. A partir de l'ADN com a motlle i en el procés de transcripció obtindrem ARN (àcid ribonucleic). A partir d'aquest ARN en el procés de traducció es sintetitzaran proteïnes.

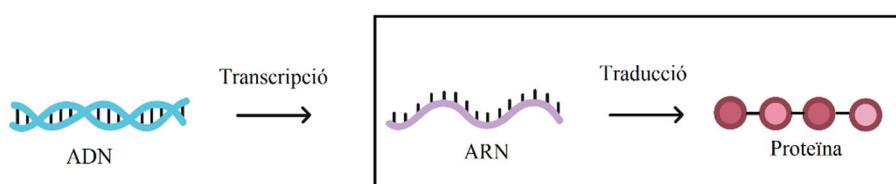


Figura 1: "Dogma central de la bioquímica molecular con enzimas". Font: Daniel Horspool ([CC BY-SA 3.0](#)). La imatge modificada està sota una llicència [CC BY-SA 3.0](#)

El 90% del nostre ADN no són gens, sinó regions que regulen quina quantitat de cada gen tenim en cada cèl·lula. Aquest procés canvia en cada cèl·lula i teixit del nostre cos. És per aquest fet que la majoria de variants en el nostre ADN no acaben afectant a la configuració d'una proteïna però si a la quantitat que d'aquesta es crea en el teixit. Aquesta regulació és la que ens pot predisposar a tenir malalties.

1.1.3. Estat de l'Art

Els importants avenços en assaigs bioquímics, eines genòmiques i l'elevada potència computacional han revolucionat la manera d'analitzar l'ADN, ARN i les proteïnes. En l'estudi del transcriptoma o els gens que es transcriuen a partir de l'ADN apareix el conjunt complert de transcripcions associats a una cèl·lula i la seva quantitat per a una etapa de desenvolupament específica o condició fisiològica. El transcriptoma reflecteix l'estat de l'expressió genètica quantificant quant ARN de cada gen tenim en una cèl·lula.

Determinar aquestes diferències en l'expressió dels gens segons els teixits als quals pertanyen és l'objectiu principal del projecte GTEx (Genotype-Tissue Expression Project), finançat pels Instituts de Salut Nord-americans (NIH), i que recull mostres de 54 tipus de teixit obtingudes a partir de mostres post-mortem pertanyents a un número aproximat de 1000 persones.

Investigadors del Consorci GTEx , han estudiat com influeix la variació del genoma en l'expressió dels gens en els diferents teixits i cèl·lules de el cos humà, i publicat una sèrie d'articles científics a Science, Cell i altres revistes científiques d'impacte, on aporten nova informació sobre la regulació de l'expressió gènica en els teixits i obren el camí a la identificació de nous biomarcadors per a diferents aspectes de les malalties humanes.

Ardlie et al. (2015) descriuen de forma acurada un conjunt de dades format per l'expressió genètica de múltiples teixits humans i avaluen la variabilitat del transcriptoma entre individus en un gran nombre de teixits proporcionant una interpretació única sobre la diversitat en la regulació de l'expressió genètica entre els mateixos. L'anàlisi facilita una visió combinada dels efectes genètics sobre l'expressió genètica en una àmplia gamma de tipus de teixits i s'indica la voluntat d'ampliar el recurs per tal de crear un conjunt de dades que canviï la nostra manera de comprendre com la variabilitat genètica té afectació en diferents teixits i sistemes biològics i, en darrera instància, en malalties complexes.

Durant la darrera dècada s'ha millorat el coneixement del paper que té la variació genètica en trets complexos i malalties humanes, especialment gràcies a estudis d'associació del genoma complert (GWAS) que han catalogat milers de variants genètiques comunes que afecten a malalties humanes i altres trets. Ara bé, els mecanismes moleculars pels quals aquesta variació genètica predisposa als individus a patir la malaltia encara es troben poc caracteritzats impedint el desenvolupament de noves teràpies. Això fa que la caracterització de l'arquitectura que regula el genoma

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

humà sigui essencial, per entendre la biologia bàsica i també per interpretar els resultats obtinguts de GWAS.

Aguet et al. (2020), partint de les dades de la versió GTExv8, observaven l'existència d'una relació entre l'expressió dels gens i l'especificitat de cada teixit assenyalant que la composició del tipus cel·lular és un factor decisiu per entendre els mecanismes de regulació gènica dels teixits humans.

Hi ha moltes malalties de les que es desconeixen a quins teixits o tipus cel·lulars afecten. La identificació d'aquests teixits i tipus de cèl·lules és fonamental per poder desenvolupar sistemes que explorin els mecanismes de regulació genètica que contribueixin al coneixement de la malaltia. En els darrers anys s'ha obtingut un major coneixement de quines parts del genoma es troben actives en teixits i tipus cel·lulars: quines parts del genoma són accessibles, quins potenciadors estan actius i quins gens s'hi expressen (Consortium et al., 2013). La combinació d'aquest tipus d'informació juntament amb les dades del GWAS possibiliten la identificació de teixits i cèl·lules afectats per malalties.

Finucane et al. (2018) realitzaren un treball per identificar teixits i tipus cel·lulars afectats per a malalties mitjançant l'anàlisi de dades d'expressió gènica i anàlisis estadístics aplicats als GWAS. En el model que van desenvolupar analitzaren com d'importants són per a l'expressió genètica les variacions que afecten a una malaltia i com aquestes es troben properes a gens expressats en un teixit. En els anàlisis específics de cervell i immunes, els enriquiments significatius són particulars pel trastorn bipolar i l'esquizofrènia. Els resultats proven que l'enfocament poligènic és de gran utilitat per treure profit de les dades d'expressió gènica per interpretar el senyal GWAS. En aquest mateix article identifiquen quins són els teixits del cervell on hi ha una major expressió dels trets lligats a malalties que l'affecten demostrant que conjunts de gens específicament expressats identifiquen els teixits i tipus de cèl·lules rellevants en l'afectació de les malalties. En la millor comprensió dels trets estudiats destaca el poder del GWAS com a font de coneixement biològic i com a utilitat en la selecció de teixits o tipus de cèl·lules *in vitro* amb el propòsit d'obtenir una millor comprensió dels mecanismes moleculars subjacents.

Clarimon et al. (2020) fan una revisió sobre l'estat de l'art respecte de com l'ús de la genètica molecular ha estat de gran ajuda per comprendre les bases moleculars de les ND més rellevants, tals com l'Alzheimer (AD) i cos de Lewy (DLB). En relació a aquestes malalties s'identifiquen una colla de gens lligats al risc de patir-les. Les noves tecnologies, entre elles GWAS, han revolucionat i accelerat la manera com s'examina la varietat del nostre ADN.

Vergouw et al. (2017) fan referència a com l'arquitectura genètica de la demència amb DLB cada vegada pren més forma i com en darrers estudis s'ha demostrat que diverses variants en el gen GBA i l'al·lel APOE ε4 són factors de risc genètics importants per a la DLB.

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

En relació a AD el gen més important que s'ha relacionat amb la malaltia és el APOE4 (Yamazaki et al., 2019). Tenim però altres estudis com ara el de (Miron, et al, 2018) on es relacionen variants del gen CDK5RAP2 amb el risc de patir aquesta malaltia.

Kamboh et al. (2012) a partir d'estudis GWAS varen detectar a banda del gen APOE nou gens/loci (CR1, BIN1, CLU, PICALM, MS4A4/MS4A6E, CD2AP, CD33, EPHA1 y ABCA7) vinculats a AD d'aparició tardana (LOAD) tot indicant que encara restaven per identificar gens de risc addicionals per LOAD.

En el cas de l'esclerosi múltiple també es posa de manifest el component genètic que aquest tipus de ND té. Per a citar algun dels treballs podem parlar de Brynedal et al. (2010) que indiquen que el gen MGAT5 altera la gravetat de l'esclerosi múltiple o el de Zhou et al. (2017) amb el seu estudi en relació a la variació genètica en el gen LRP2. En el cas de Parkinson (PD) tenim treballs com el de Chung et al. (2012) i Nalls et al. (2011).

En la malaltia de Machado-Joseph tenim el treball de Akçimen et al. (2020) on es demostra l'existència de diversos factors genètics addicionals que poden conduir a una millor comprensió de la mateixa.

Malgrat la identificació de gens lligats a malalties encara hi ha un gran desconeixement respecte dels gens causals exactes i les seves vies biològiques, que encara són en gran part desconegudes. Les dades d'expressió gènica del teixit cerebral i perifèric poden millorar la detecció de variacions reguladores lligades a les ND (Gerring et al., 2020).

La capacitat per identificar teixits i tipus de cèl·lules rellevants per la malaltia millorarà a mida que es disposin d'un major número de mostres GWAS per més fenotips i es generin dades d'expressió genètica en nous teixits i tipus de cèl·lules. Això ens permetrà avançar en la comprensió de la biologia de les malalties i servirà per futurs experiments que explorin variants i mecanismes específics.

En aquest treball lligarem el component genètic de les malalties ND amb les diferents zones específiques del cervell a partir de les seves particulars expressions genètiques. El propòsit d'aquest estudi és el de poder entendre millor la malaltia i ajudar a possibles vies d'investigació com ara la creació de xarxes genètiques específiques del teixit que ens ajudin a trobar medicines que afectin als mateixos *pathways* que la malaltia. A partir de GTExv8 i els gens associats a ND, obtinguts a partir de *papers* GWAS, veurem on poden ser localitzats aquests darrers entre els gens més expressats dels teixit cerebrals.

1.1.4. Motivació

A nivell personal, la meva motivació per emprendre aquest projecte ha estat fonamentada per l'afany d'endinsar-me en un món apassionant allunyat de la meva formació i alhora per la necessitat de mirar de fer quelcom d'utilitat per la societat. La majoria tenim algun familiar o company que malauradament han patit o pateixen

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

malalties sense cura a dia d'avui. Vull pensar que, en la mida que em sigui possible, amb el màxim d'esforç i amb l'ajuda del professorat, mirarem de fer una petita aportació en el coneixement d'aquestes malalties amb l'esperança que els resultats d'aquest treball pugui ser d'utilitat en propers projectes.

1.2. Objectius del Treball

L'objectiu principal del projecte consisteix en relacionar malalties neurodegeneratives amb els diferents teixits del cervell a partir de l'expressió genètica.

Per l'obtenció d'aquest objectiu principal es defineixen els següents objectius secundaris:

- Diferenciar els teixits del cervell emprant la seva expressió genètica per a posteriorment analitzar a quin tipus de teixit pertany una mostra.
- Trobar en quines parts del cervell es troben sobre expressats aquells gens relacionats amb malalties.

1.3. Enfocament i mètode seguit

En aquest projecte es partirà de l'estudi de l'ARN. No s'estudia l'ADN perquè aquest és el mateix en totes les cèl·lules i no té informació particular en cada teixit. S'estudiarà l'ARN i no l'ADN perquè el primer és l'expressió del segon, en el sentit que l'ADN conté informació però no realitza un paper actiu.

És a partir de l'ARN que es podrà quantificar l'expressió genètica.

Per dur a terme aquest enfocament s'empraran dades obtingudes de GTExv8 i GWAS. Les mostres de teixits que formen part del cervell s'obtindran de la Base de Dades pública GTExv8 i se n'analitzaran les seves expressions genètiques per tal de veure si és possible la diferenciació per les mateixes.

D'articles publicats en revistes d'impacte s'obtindran els gens que han estat relacionats amb cada malaltia a partir d'anàlisis GWAS.

El catàleg GWAS és una base de dades publicada en línia ,gratuïta i que recull dades d'estudis associats al genoma complet organitzats en ontologies. En aquests estudis s'hi recull la informació de les associacions entre polimorfisme d'un sol nucleòtid (SNP) i malalties. Aquest catàleg s'utilitza per identificar variants causals i comprendre els mecanismes de les malalties.

En aquest treball es cercaran gens implicats en malalties neurodegeneratives, obtinguts a partir de GWAS, en aquells teixits del cervell, obtinguts de GTExv8, a partir de l'expressió genètica dels mateixos.

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

En primer lloc, l'enfoc del treball ha estat orientat a poder distingir els diferents tipus de teixits del cervell. En aquesta primera fase del treball s'utilitza R per tal de fer: la selecció de dades dels teixits cerebrals, l'exploració de les dades i les transformacions necessàries per a la creació de datasets.

L'elecció de R ha estat per la facilitat de maneig de fitxers de mida considerable i per les funcionalitats que incorpora. A partir de les dades de teixits obtingudes de GTEx es realitza una exploració de les dades obtenint resultats de forma gràfica. En aquesta fase inicial es transformen i preparen els datasets que posteriorment s'empraran en el treball.

En una segona fase del treball, l'objectiu es veure si poden distingir-se els diferents teixits a partir de l'expressió dels gens. Per aquest efecte es realitzaran diferents mètodes de classificació amb la formació de clústers. Aquesta segona fase, al igual que la tercera, ha estat realitzada emprant Python.

En una tercera fase i a partir de GWAS s'obtenen els gens lligats a diferents ND. L'objectiu d'aquesta tercera fase es trobar en quines parts del cervell són sobre expressats aquells gens relacionats amb malalties.

Agafant com a malaltia l'Alzheimer, es cercaran tots aquells gens que hi estan relacionats obtinguts de GWAS. S'obtindran quants d'aquests s'expressen en el cervell i com ho fan en cada teixit. Per a la detecció de diferències significatives de sobre expressió o infra expressió de gens en teixits es realitzaran proves de contrast d'hipòtesi Welch t-test.

1.4. Planificació del Treball

En el transcurs de cada fase i tasca s'ha realitzat una treball continu de documentació de les tasques realitzades, codi generat i els resultats obtinguts.

1.4.1. Tasques

Objectiu I – i

1. Obtenció de les dades a partir de GTExv8 normalitzades en TPM
2. Treball amb R
3. Estudi de les dades GTExv8: camps, tipus de teixits, com són i s'entreguen les dades que conté.
4. Filtratge. Selecció a partir GTExv8 d'únicament les mostres de teixits associades al cervell.
5. Exploració de les dades i construcció d'una estructura transposada amb mostres de teixits, gens i identificador d'individu
6. Realització d'histogrames per cada teixit. Comportament *log normal* de les dades.
7. Gràfica amb la linealitat entre log mitjana i log variància dels gens en teixits.

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

8. Selecció per l'estudi dels gens que la seva expressió per Teixits compleixen el criteri $mitjana(TPM)>1$.
9. Com es distribueixen els gens en els teixits? Gràfica dels nombre de gens $mitjana(TPM)>1$ expressats per teixit

Objectiu I – ii

10. Prèvia: Obtenció de la llista de gens codificant i mitocondrials
11. Generació de fitxers:
 - a. de gens per teixits expressats en $mitjana(TPM)>1$ en ordre de magnitud
 - b. de gens per teixits expressats en $mitjana(TPM)>1$ i ordenats segons el seu CV (Coeficient de Variació)
 - c. Creació de noves versions dels fitxers tenint en compte tan sols els gens que codifiquen proteïnes
 - d. Creació de noves versions dels fitxers excluent aquells gens que són mitocondrials.

Objectiu II – i

12. Treball amb Python.
13. Realització de clústers a partir de mètodes supervisats per analitzar quina és la diferenciació de teixits del cervell emprant la seva expressió genètica.
14. Es realitzen un seguit d'estudis a partir de diferents *datasets* i amb una reducció progressiva del nombre de gens.
Els *datasets* són:
 - Top 5000, 1000, 500, 100 gens obtinguts a partir de $mitjana(TPM)>1$.
 - Top 5000, 1000, 500, 100 gens obtinguts a partir de $mitjana(TPM)>1$ seleccionant únicament aquells que codifiquen proteïnes.
 - Top 5000, 1000, 500, 100 gens obtinguts a partir de $mitjana(TPM)>1$ excluent-ne els mitocondrials.
 - Top 5000, 1000, 500, 100 gens obtinguts a partir de $mitjana(TPM)>1$ i tenint en compte el CV, seleccionant únicament aquells que codifiquen proteïnes.
 - Top 5000, 1000, 500, 100 gens obtinguts a partir de $mitjana(TPM)>1$ i tenint en compte el CV, excluent-ne els mitocondrials.
15. Tasques prèvies per cada anàlisi de clúster:
 - Com fer front a l'elevada dimensionalitat: Anàlisi PCA. Reducció de les dimensions obtingudes.
 - Exploració dels principals components obtinguts.
 - Selecció del nombre de components a emprar.
 - Visualització gràfica d'alguns del components principals per veure com es diferencien.
 - Visualització amb t-SNE en 2D dels resultats obtinguts.
16. Aplicació de diferents models: *knn*, *SVM*, *Arbre de decisió*, *Random Forest* i

obtenció de les matrius de confusió per cadascun d'ells.

17. Selecció del model amb major precisió.

Objectiu II – ii

18. Valoració dels resultats obtinguts a partir dels diferents models i el número de gens.

Objectiu III – i

19. Obtenció del llistat GWAS de gens lligats a malalties

20. Selecció dels gens del cervell que tenen un *mitjana(TPM)>1* o bé, que agrupats per teixits, en algun d'ells la seva *mitjana(TPM)>1*

21. Hi ha gens que tan sols s'expressen en teixits únics? Anàlisi

22. Selecció dels gens obtinguts GWAS lligats a una malaltia concreta: Alzheimer entre tots els prèviament filtrats del cervell i obtinguts a partir de GTEx.

23. Existeixen diferències en l'expressió de gens en els teixits? Realització de proves de contrast d'hipòtesi *Welch t-test* doncs la variància entre els grups és diferent.

24. Anàlisi d'expressió diferencial. Gens més i menys expressats per teixit.
Elaboració de *rànkings* en funció de:

- la seva quantitat
- suma de diferències entre mitjanes del propi teixit i la resta

25. Generació de *heatmaps*

26. Són significatius els resultats obtinguts? Estudi de l'expressió dels 1000 gens amb major *mitjana(tpm)>1* en el transcriptoma complert.

27. Hi ha diferències significatives entre els resultats obtinguts i l'execució de 10.000 iteracions del mateix nombre de gens obtinguts de manera aleatòria a partir del transcriptoma complert? Realització de contrast d'hipòtesi *Welch t-test*.

Objectiu III – ii

28. Realitzar la comparació entre resultats.

29. Interpretar els resultats amb el suport d'articles científics.

Addicionals

30. Redactar la memòria definitiva

31. Enllestit: Resultats, discussió i conclusions, *abstract*, bibliografia,

32. Revisar l'estructura i la redacció dels continguts.

33. Preparar una presentació Power Point per la defensa.

34. Preparar la defensa pública.

1.4.2. Calendari

La planificació s'estructura de la següent manera seguint el diagrama de Gantt.



Figura 2: Diagrama de Gantt amb la planificació del projecte.

1.5. Breu sumari de productes obtinguts

Com resultat del treball ha estat possible diferenciar els teixits del cervell a partir de la seva expressió genètica.

S'han realitzat diferents clústers a partir de diferents models. Els millors resultats s'han obtingut de l'aplicació del model SVM al conjunt de dades format pels 5000 gens amb major CV (Coeficient de Variació). La precisió amb aquest model ha estat del 94.37%.

El model amb millors resultats globals obtinguts al disminuir el número de gens a 1000, 500 i 100 ha estat el SVM aplicat sobre el conjunt de dades format per gens codificant ordenats segons el criteri de major CV.

Amb els anàlisis realitzats s'ha observat que amb un número reduït de gens és possible dur a terme una diferenciació dels teixits.

En la recerca de quines són les part del cervell en que s'expressen els gens relacionats amb malalties, l'estudi s'ha focalitzat amb la malaltia de l'Alzheimer.

Com a resultat s'ha trobat una sobre expressió rellevant de gens en els teixits corticals.

1.6. Breu descripció dels altres capítols de la memòria

Tot seguit s'explica de forma breu la resta dels capítols:

Materials i mètodes:

- **Dades:** es detallen les dades emprades en aquest treball: format, tipus i origen.

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

- **Anàlisi estadístic:** en aquest capítol s'indiquen les accions realitzades per conèixer les dades.
- **Diferenciació de Teixits:** en aquest capítol s'explica el procés seguit per dur a terme la diferenciació de teixits a partir del models supervisats i les accions que han estat necessàries.
- **Anàlisi de l'expressió dels gens:** en aquest capítol s'explica el procés seguit en la recerca dels gens relacionats amb malalties en els diferents teixits del cervell i com han estat emprats contrastos d'hipòtesis.

Resultats:

- **Diferenciació de Teixits:** s'indiquen els resultats dels diferents clústers realitzats a partir de diferents algoritmes supervisats per veure si és possible diferenciar els teixits del cervell emprant la seva expressió genètica a partir de la creació d'un model.
- **Anàlisi de l'expressió dels gens:** s'indiquen: s'indiquen els resultats obtinguts de les proves realitzades per obtenir els gens sobre i infra expressats i els resultats en relació a les parts del cervell on s'hi expressen gens relacionats amb la malaltia de l'Alzheimer.

Discussió:

Es realitza l'anàlisi dels resultats obtinguts tot fent una valoració crítica del treball realitzat i fent propostes per treballs futurs. En relació amb els resultats obtinguts es presenten articles científics relacionats amb els mateixos.

2. Materials i Mètodes

El projecte s'ha estructurat en dos àmbits diferents, d'una banda la realització d'una colla d'algoritmes lligats a mètodes supervisats per obtenir un model que permeti la classificació dels teixits a partir de la seva expressió genètica i uns altres lligats a cercar en quins teixits són expressats aquells gens relacionats amb la malaltia. En el primer àmbit les dades s'obtenen de GTEx v8 i en el segon s'empren també dades GWAS.

S'han obtingut fitxers en que els gens han estat ordenats de més a menys en funció de la seva expressió en forma de mitjana(TPM) i posteriorment per Coeficient de Variació (CV).

El coeficient de variació, altrettament anomenat coeficient de variació de Pearson, és una mesura estadística que ens aporta informació al respecte de la dispersió relativa d'un conjunt de dades.

Aquest CV és la relació entre la desviació típica d'una mostra i la seva mitjana.

La relació de fitxers obtinguts i generats:

- **all_gensRAW.csv**

Fitxer generat amb mostres de teixits sols del Cervell i expressió dels gens en format TPM.

- **Gens_Tots_detail.csv**
Fitxer generat amb mostres de teixits sols del Cervell i expressió dels gens en format TPM complint que la seva mitjana(TPM)>1.
- **all_gensTPM_ZA.csv**
Fitxer generat amb mostres de teixits sols del Cervell i expressió dels gens en format TPM, complint que la seva mitjana(TPM)>1, ordenats de major a menor expressió.
- **all_gensTPM_noversio_ZA.csv**
Fitxer generat idèntic al 'all_gensTPM_ZA.csv' anterior però eliminant la versió en el gen.
- **vw_GENS_PROTEINES_WIKI.txt**
Fitxer obtingut amb la relació de gens que codifiquen proteïnes.
- **vw_MITOCONDRIALS_ENSEMBLE.txt**
Fitxer obtingut amb la relació de gens mitocondrials.
- **all_gensTPM_noversio_PROTEIN_ZA.csv**
Fitxer generat amb mostres de teixits sols del Cervell i expressió dels gens que codifiquen proteïnes en format TPM, complint que la seva mitjana(TPM)>1, ordenats de major a menor expressió.
- **all_gensTPM_noversio_NOMITO_ZA.csv**
Fitxer generat amb mostres de teixits sols del Cervell i expressió dels gens no mitocondrials en format TPM, complint que la seva mitjana(TPM)>1, ordenats de major a menor expressió.
- **all_gensCV_ZA.csv**
Fitxer generat amb mostres de teixits sols del Cervell i expressió dels gens en format TPM, complint que la seva mitjana(TPM)>1, ordenats de major a menor CV (Coeficient de Variació).
- **all_gensCV_noversio_ZA.csv**
Fitxer generat idèntic al 'all_gensCV_ZA.csv' anterior però eliminant la versió en el gen.
- **all_gensCV_noversio_PROTEIN_ZA.csv**
Fitxer generat amb mostres de teixits sols del Cervell i expressió dels gens que codifiquen proteïnes en format TPM, complint que la seva mitjana(TPM)>1, ordenats de major a menor CV (Coeficient de Variació).

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

- **all_gensCV_noversio_NOMITO_ZA.csv**
Fitxer generat amb mostres de teixits sols del Cervell i expressió dels gens no mitocondrials en format TPM, complint que la seva mitjana(TPM)>1, ordenats de major a menor CV (Coeficient de Variació).
- **gwas_catalog_v1.0.2-associations_e100_r2021-04-12.tsv**
Fitxer obtingut amb el catàleg GWAS de gens relacionats amb malalties humanes.
- **10000_iteracions.csv**
Fitxer generat amb el resultat del número de gens sobre expressats en teixits al fer 10000 execucions de grups de n gens aleatoris del transcriptoma.
- **10000_infraexpressatsneto.csv**
Fitxer generat amb el resultat del número de gens infra expressats en teixits al fer 10000 execucions de grups de n gens aleatoris del transcriptoma.

Relació de fitxers amb codi:

- **TFM_Part1.rmd**
Codi R utilitzat per la selecció de dades dels teixits cerebrals, la seva exploració i les transformacions necessàries per a la creació de datasets.
 - **TFM_Part1.html**
Codi R utilitzat i resultats en la selecció de dades dels teixits cerebrals, la seva exploració i les transformacions necessàries per a la creació de *datasets*.
 - **TFM_PART2 TPM-RAW1000.ipynb**
Codi *Python* utilitzat per l'estudi de diferenciació entre teixits del cervell a partir de la seva expressió genètica. Aquest és un fitxer de codi dels múltiples realitzats a fi d'exemple.
 - **TFM_PART2 TPM-RAW1000.html**
Codi *Python* utilitzat i resultats de l'estudi de diferenciació entre teixits del cervell a partir de la seva expressió genètica.
- (Altres fitxers, similars als TPM-RAW1000 i emprats en la construcció de la resta de models durant aquesta segona fase també han estat inclosos.)
- **TFM_PART3_ESTUDI_ALZHEIMER.ipynb**
Codi *Python* utilitzat en la recerca de quines són les part del cervell en que s'expressen els gens relacionats amb la malaltia de l'Alzheimer.
 - **TFM_PART3_ESTUDI_ALZHEIMER.html**
Codi *Python* utilitzat i resultats obtinguts en la recerca de quines són les part del cervell en que s'expressen els gens relacionats amb la malaltia de l'Alzheimer.

2.1. Dades

- GTEx v8:

Les dades amb l'expressió dels gens en els teixits han estat obtingudes d'un total de 17382 mostres de 54 teixits diferents obtinguts de 948 donants provinents de GTEx. Aquestes 17382 mostres contenen l'expressió en TPM de 56200 gens.

V8 Release	# Tissues	# Donors	# Samples
Total	54	948	17382

Taula 4: GTExv8. Contingut de mostres, teixits i donants d'on provenen.

```
#1.2
56200 17382
Name Description GTEX-1117F-0226-SM-5GZZ7 GTEX-1117F-0426-SM-5EGHI GTEX-1117F-0526
ENSG00000223972.5 DDX11L1 0 0 0 0 0 0 0 0 0 0.01776 0.03757 0.04667 0 0.01832 0 0
ENSG00000227232.5 WASH7P 8.764 3.861 7.349 11.07 3.306 5.389 11.99 16.95 10.04 12.5
ENSG00000278267.1 MIR6859-1 0 0 1.004 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
ENSG00000243485.5 MIR1302-2HG 0.07187 0 0 0.06761 0 0 0 0 0 0 0.06265 0.05905 0 0 0.046
ENSG00000237613.2 FAM138A 0 0 0 0 0 0 0 0.03904 0 0 0 0 0 0.0331 0 0.02599 0 0 0.030
ENSG00000268020.3 OR4G4P 0 0.056 0 0 0.0613 0.09523 0 0.0555 0.1292 0 0.03656 0
ENSG00000240361.1 OR4G11P 0.06621 0.05004 0 0 0 0 0 0.0992 0 0.1088 0 0.03267 0.138
ENSG00000186092.4 OR4F5 0 0.1025 0.07434 0 0.04233 0.05609 0.1743 0 0 0.0591 0.055
ENSG00000238009.6 RP11-34P13.7 0 0.04574 0.09953 0 0.07556 0.05006 0.03889 0.04627 0
ENSG00000233750.3 CICP27 0.03595 0.01359 0 0 0 0 0 0.02749 0 0 0 0 0.03753 0.05827 0
ENSG00000268903.1 RP11-34P13.15 3.215 0.2492 1.356 1.861 0.9263 2.387 1.695 6.681 2
```

Figura 3: Mostra del fitxer GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct

En l'estudi tindrem un total de 2641 mostres de 13 teixits del cervell amb l'expressió normalitzada en TPM de 56200 gens.

SAMPID	SMTSD	ENSG00000223972	ENSG00000227232	ENSG00000278267	ENSG00000243485	E
1 GTEX-111FC-3126-SM-5GZ22	Brain - Cortex	0.00000	4.2250	0.4912	0.07713	
2 GTEX-111FC-3326-SM-5GZYV	Brain - Cerebellum	0.00000	7.7780	0.7710	0.00000	
3 GTEX-1128S-2726-SM-5H12C	Brain - Cortex	0.01709	2.3590	0.0000	0.03413	
4 GTEX-1128S-2826-SM-5N9DI	Brain - Cerebellum	0.00000	8.9390	0.0000	0.00000	

Showing 1 to 5 of 2,641 entries, 56202 total columns

Taula 5: Fitxer amb tots els gens RAW lligats al cervell

Al mesurar l'expressió genètica, molts gens no s'expressen en un teixit. Per aquest fet existeixen molts valors propers a zero. Tots aquests valors propers al zero són soroll i seran filtrats. En aquest estudi s'eliminen tots aquells gens que tenen una mitjana(TPM)<1 en teixits del cervell.

Un cop filtrats, la mida mostra final resta formada per 17515 gens.

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

SAMPID	SMTSD	ENSG00000227232.5	ENSG00000268903.1	ENSG00000269981.1	ENSG00000279457.4
1 row names	126-SM-5GZZ	Brain - Cortex	4.2250	1.9460	2.8220
2	GTEX-11FC-3326-SM-5GZYV	Brain - Cerebellum	7.7780	5.2780	2.7690
3	GTEX-1128S-2726-SM-5H12C	Brain - Cortex	2.3590	1.4880	2.2890
4	GTEX-1128S-2826-SM-5N9DI	Brain - Cerebellum	8.9390	2.4900	5.4850
5	GTEX-117XS-3026-SM-5N9CA	Brain - Cortex	3.6930	0.8472	1.4080

Taula 6: Dataset resultant amb 17515 gens en que la seva mitjana(TPM)>1

- Relació de Gens codificant i mitocondrials:

Uns altres fitxers obtinguts han estat:

La relació de 21701 gens que codifiquen proteïnes. Aquesta relació de gens codificant s'ha obtingut de (Home | HUGO Gene Nomenclature Committee, s.f.)

La relació de 37 gens mitocondrials s'ha obtingut de (Ensembl genome browser, s.f.)

- Catàleg GWAS

El fitxer amb el catàleg GWAS ha estat obtingut directament de la pròpia web en la seva versió 1.0.2: gwas_catalog_v1.0.2-associations_e100_r2021-04-12.tsv

De tots els camps del fitxer destacar:

Nom del Camp	Descriptiu
DISEASE/TRAIT	Malaltia lligada a l'estudi
UPSTREAM_GENE_ID	Gen lligat a malaltia
DOWNSTREAM_GENE_ID	Gen lligat a malaltia
SNP_GENE_IDS	Gens lligat a malaltia
PUBMEDID	Número que identifica l'estudi a PubMed

Taula 7: Relació de camps destacats en el fitxer que recull el catàleg GWAS.

D'aquesta manera i a partir d'aquest catàleg GWAS s'obtindran els treballs i gens relacionats amb la malaltia.

2.2. Anàlisi Descriptiu

Previ anàlisi estadístic de les dades emprades s'ha realitzat una exploració general i un anàlisis descriptiu de les mateixes per obtenir-ne la distribució. Ambdues accions s'han realitzat amb el software R.

Per la visualització de les dades s'ha realitzat un anàlisi descriptiu estudiant-ne els trets més rellevants i observant la seva distribució en funció dels teixits. En la visualització d'aquests aspectes s'han realitzat gràfics i taules.

S'ha realitzat una exploració per comprovar la linealitat existent entre la mitjana en log i la variància en log: la variància augmenta amb la mitjana.

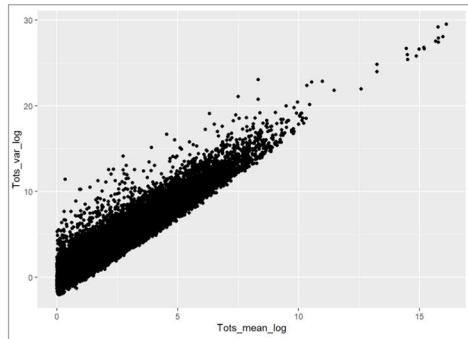


Figura 4: Gràfica del log mitjana vs. log de la variància obtinguda a partir del conjunt de dades dels teixits del cervell

Una altra prova prèvia és la realització d'histogrames en log pels diferents teixits i observar la distribució obtinguda.

Histograma de la distribució de gens en el conjunt de tots els teixits:

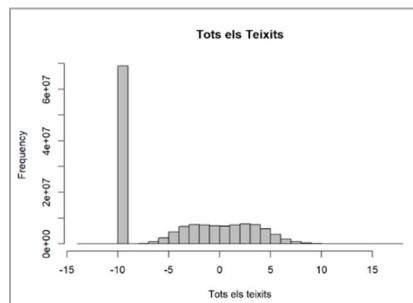


Figura 5: Histograma de la distribució en log2 de gens en el conjunt de teixits

Histograma de la distribució de gens en teixits de l'Amygdala:

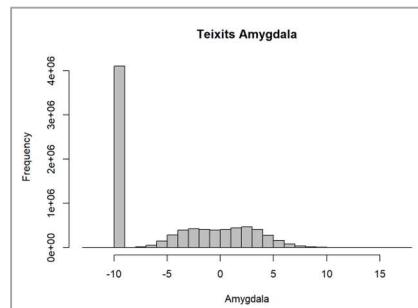


Figura 6: Histograma de la distribució en log2 de gens en teixits de l'Amygdala

Histograma de la distribució de gens en teixits de l'Anterior:

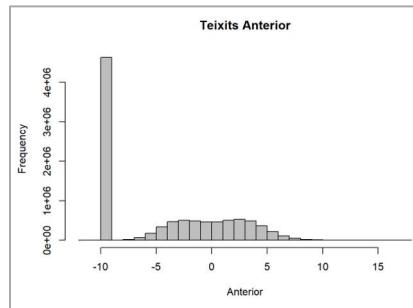


Figura 7: Histograma de la distribució en log2 de gens en teixits de *l'Anterior*

Histograma de la distribució de gens en teixits del *Caudate*:

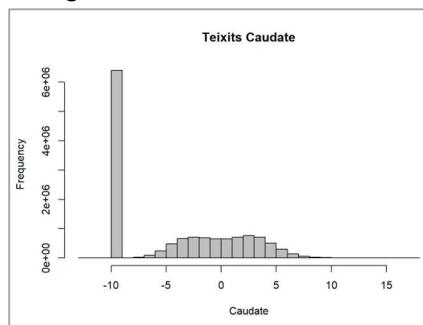


Figura 8: Histograma de la distribució en log2 de gens en teixits *Caudate*

Histograma de la distribució de gens en teixits del *Cerebellar*:

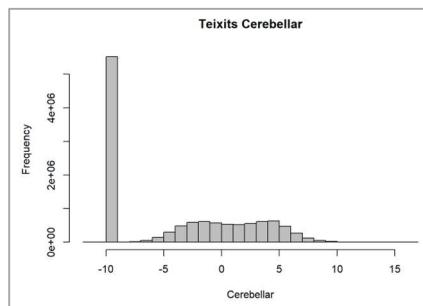


Figura 9: Histograma de la distribució en log2 de gens en teixits del *Cerebellar*

Histograma de la distribució de gens en teixits del *Cerebellum*:

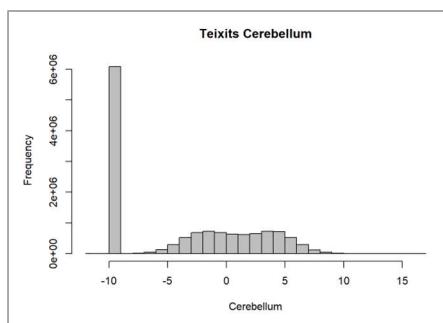


Figura 10: Histograma de gens en en log2 teixits del *Cerebellum*

Histograma de la distribució de gens en teixits del *Cortex*:

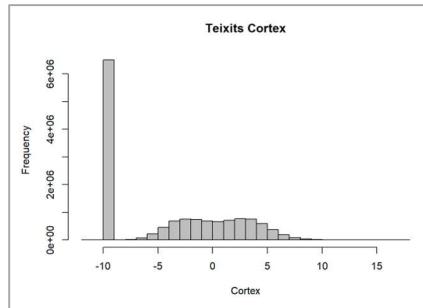


Figura 11: Histograma de la distribució en log2 gens en teixits del *Cortex*

Histograma de la distribució de gens en teixits del *Frontal*:

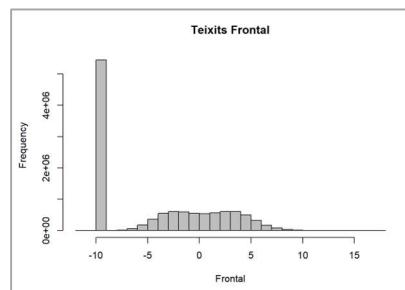


Figura 12: Histograma de la distribució en log2 de gens en teixits del *Frontal*

Histograma de la distribució de gens en teixits de l'*Hippocampus*:

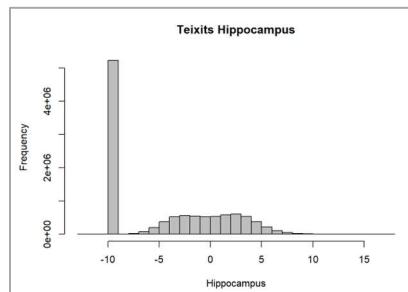


Figura 13: Histograma de la distribució en log2 de gens en teixits de l'*Hippocampus*

Histograma de la distribució de gens en teixits de l'*Hypothalamus*:

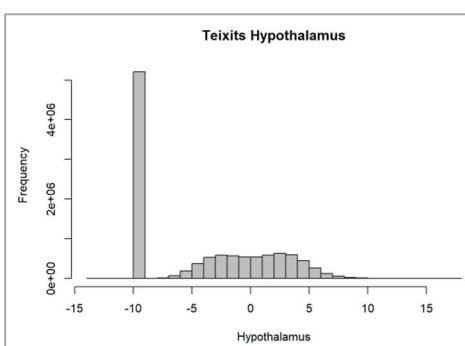


Figura 14: Histograma de la distribució en log2 de gens en teixits de l'*Hypothalamus*

Histograma de la distribució de gens en teixits del *Nucleus*:

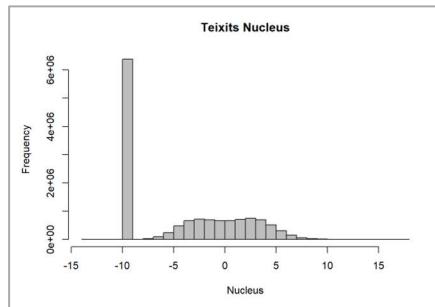


Figura 15: Histograma de la distribució en log2 de gens en el teixits del *Nucleus*

Histograma de la distribució de gens en teixits del *Putamen*:

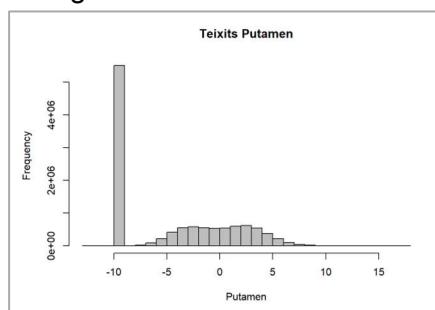


Figura 16: Histograma de la distribució en log2 de gens en teixits del *Putamen*

Histograma de la distribució de gens en teixits de *l'Spinal*

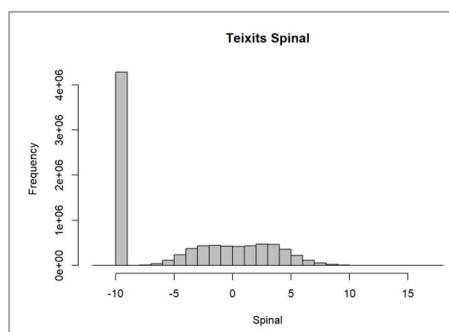


Figura 17: Histograma de la distribució en log2 de gens en teixits de Spinal

Histograma de la distribució de gens en teixits de *Substantia*

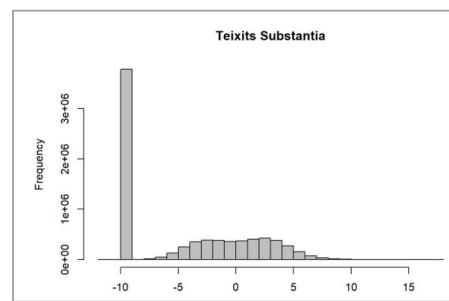


Figura 18: Histograma de la distribució en log2 de gens en teixits de *Substantia*

Per a poder realitzar el log ha calgut fer un preprocessat previ dels valors en TPM dels gens de tal manera que tots aquells que tenien valor 0 han estat normalitzats a un valor de 0.001.

En tots els histogrames s'observa la presència d'un gran valor acumulat en forma de pic a la franja esquerra.

La distribució tan sols és normal al excloure aquests valors 0. Al mesurar expressió genètica, molts gens no s'expressen en un teixit i per tant es esperable aquest pic al zero. Tanmateix valors propers al zero són soroll i per tant s'exclouen aquells que mitjana(TPM)<1 com a gens no expressats en un teixit/cervell.

Al excloure els gens on la seva mitjana(TPM)<1 s'observa la distribució log normal dels nous histogrames obtinguts:

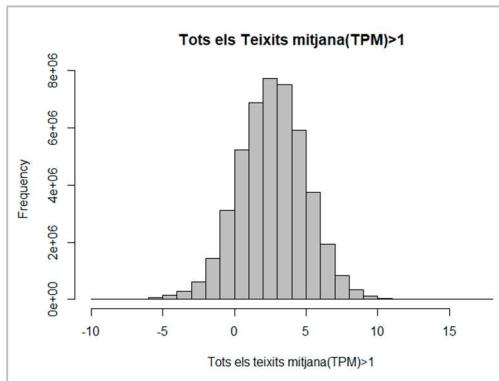


Figura 19: Histograma de la distribució en log2 de gens en el Total de Teixits en que mitjana(TPM)>1

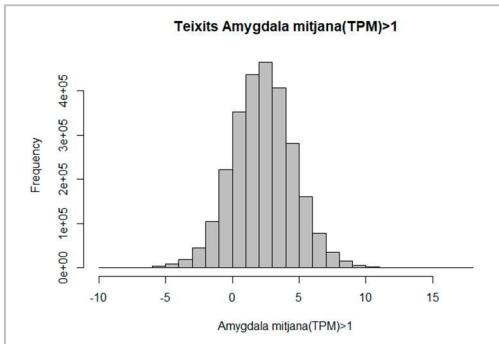


Figura 20: Histograma de la distribució en log2 de gens en el Teixit Amygdala en que mitjana(TPM)>1

En la distribució del nombre de gens expressats, en mitjana(TPM)>1 destaquen els teixits del Cerebellar i el Cerebellum com aquells que en tenen un major número.

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

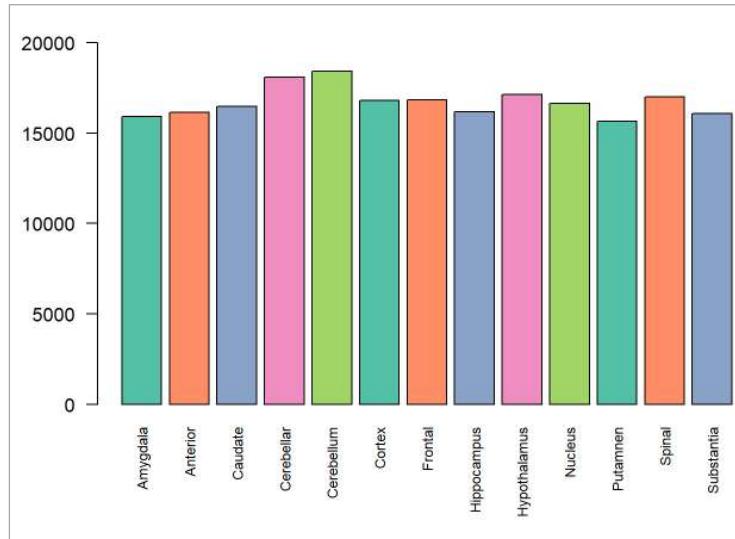


Figura 21: Número de Gens Expressats, en mitjana(TPM)>1, per Teixit

La distribució de mostres per teixit és la següent:

Teixit del Cervell	Número de Mostres
Amygdala	152
Anterior	176
Caudate	246
Cerebellar	215
Cerebellum	241
Cortex	254
Frontal	209
Hippocampus	197
Hypothalamus	202
Nucleus	246
Putamen	205
Spinal	159
Substantia	139

Taula 8: Relació de Teixits del Cervell i número de mostres de cadascun d'ells.

2.3. Diferenciació de Teixits

A partir de la matriu amb expressió genètica de teixits el propòsit es detectar característiques comunes que puguin ser discriminades per un petit número de gens i permeten diferenciar els diferents tipus de teixits.

Es crearan un seguit de classificadors aplicant diferents algoritmes supervisats que podran ser emprats posteriorment per fer la distinció de teixits a partir de la seva expressió genètica.

Es realitzaran clústers:

- kNN
- SVM

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

- Arbre de Decisió
- Random Forest

a partir de:

- diferents conjunts de dades formades per:
 - tots els gens (RAW)
 - per tan sols els codificant (PROT)
 - amb l'exclusió dels mitocondrials (NO MITO)
- Es tindrà en compte l'expressió dels gens a partir de:
 - TPM
 - CV

Es tenen en compte aquells gens amb major expressió quantificada en TPM perquè en els estudis previs (Figura 4) s'ha comprovat com la variància i la mitjana varien linealment de tal manera que els gens més expressats són aquells que tenen major variació entre teixits i seran per tant els que més poden ajudar a diferenciar-los.

Emprant CV el criteri de selecció permet tenir en compte aquells gens que potser tan sols s'expressen en un teixit i que si els volguéssim obtenir a partir de TPM no els tindríem. Amb CV es mira la variació normalitzada per la variància.

De entre tots els gens, aquells que codifiquen proteïnes són dels que millor coneixement es té de les funcions que realitzen en el nostre cos. En canvi, dels gens mitocondrials el coneixement que se'n té és molt poc i en la majoria d'estudis s'exclouen.

- Amb diferent nombre de gens més expressats ordenats de major a menor
 - 5000
 - 1000
 - 500
 - 100

Caldrà l'aplicació de mètodes concrets per cercar els gens que ens permeten obtenir aquesta discriminació i reduir l'elevada dimensionalitat present en el conjunt de dades. Per aquest propòsit s'ha plantejat la realització d'un anàlisi supervisat amb una reducció prèvia emprant PCA (Anàlisi de Components Principals). El resultat obtingut a partir de la transformació PCA ha estat visualitzat mitjançant t-SNE. t-SNE (*T-distributed Stochastic Neighbor Embedding*) és un algoritme dissenyat per la visualització de conjunts de dades amb alta dimensionalitat.

S'han elaborat models aplicant les següents tècniques supervisades: kNN, SVM, Arbre de Decisió i *Random Forest*. En aquests models s'aplica una transformació de les dades a partir del número de components principals PCA que expliquen un 95% de la variància. De entre tots els models es seleccionarà aquella amb que obtingui una major precisió.

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

Han estat definits un conjunt de dades per entrenament i test del 80% i 20% de les mostres respectivament. En els models, amb hiperparàmetres ajustats via Grid Search, s'ha realitzat una validació creuada de 10 iteracions per reduir l'efecte de l'*overfitting*. Per cadascun dels models creats a partir dels diferents datasets i per 5000, 1000, 500 i 100 gens s'han obtingut les seves respectives precisions i matrius de confusió.

S'ha emprat GridSearchCV, classe disponible a scikit-learn que permet avaluar i seleccionar de manera sistemàtica els paràmetres d'un model. Amb el model i paràmetres a provar es pot avaluar el rendiment del primer en funció dels segons mitjançant validació creuada.

La validació creuada és una tècnica que permet obtenir models més estables al permetre identificar l'existència de problemes durant l'entrenament dels mateixos com és el cas de l'aparició del sobre ajust (*overfitting*).

La validació creuada parteix del fraccionament d'un conjunt de dades en un número k de particions anomenades *folds*. Després, la validació itera entre les dades d'avaluació i entrenament k vegades d'una manera particular. En cadascuna de les iteracions s'escull un *fold* diferent com a dades d'avaluació. En aquesta iteració, la resta dels k-1 *folds* es combinen per formar les dades d'entrenament. D'aquesta manera, en cada iteració tenim $(k-1)/k$ de les dades emprades per l'entrenament i $1/k$ emprat per l'avaluació. Un cop enllestida la validació creuada, tots els exemples s'hauran utilitzat tan sols una vegada per avaluar però k-1 vegades per entrenar. Finalment s'obtenen estimacions de rendiment de tots els folds i es pot calcular, per la precisió del model, la mitjana i la desviació estàndard.

En la Figura 4 es detalla la relació de diferents anàlisis supervisats realitzats a partir dels diferents conjunts de dades i amb diferent nombre de gens.

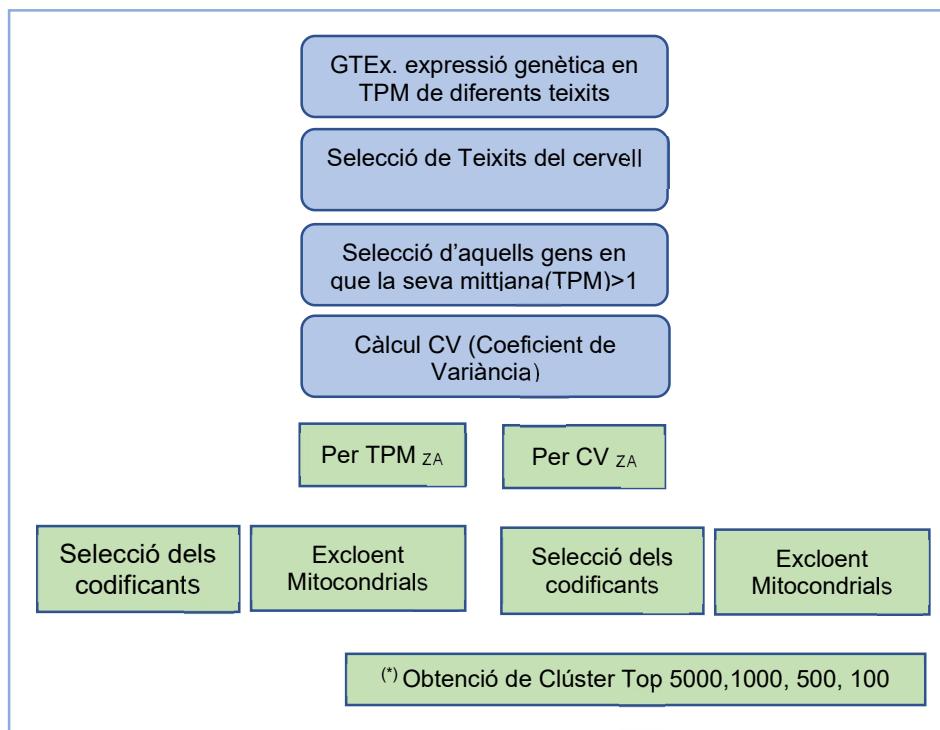


Figura 22: Flux seguit per la realització dels clústers

2.4. Anàlisi d'Expressió

El conjunt de dades està format per tots els gens associats a gens obtinguts de GWAS relacionats amb la malaltia de l'Alzheimer.

De tot aquest conjunt de dades inicial s'empraran únicament aquells gens que compleixin alguna de les següents condicions:

- Condició 1: tinguin mitjana(TPM)>1
- Condició 2: agrupats per Teixits tinguin una mitjana(TPM)>1 en algun d'ells

Pel cas de l'Alzheimer el nostre conjunt de dades estarà format per un total de 644 gens que compleixen alguna de les dues condicions indicades.

Recordar que tots els càlculs es realitzen en log normal per tal de tenir normalitat. S'aplica un log2.

Per detectar si existeixen diferències en l'expressió de gens en els teixits s'han realitzat proves de contrast d'hipòtesi Welch t-test. El propòsit de les mateixes ha estat trobar diferències, gen a gen, entre un teixit i la resta agrupada.

Les condicions per calcular un test d'hipòtesi basats en la distribució t Student són les mateixes que pel teorema del límit central:

- Independència: les observacions han de ser independents entre si.
- Normalitat: les poblacions a comparar cal que tinguin una distribució normal
- Igualtat de Variància (homocedasticitat): En aquest cas no es compleix l'homocedasticitat i per aquest fet s'utilitza el *Welch Two Sample t-test*. Aquesta correcció s'incorpora a través dels graus de llibertat permetent compensar la diferència de variàncies però amb l'inconvenient que es perd precisió.

Emprant el model Welch t-test, tenint present que la variància d'ambdós grups és diferent, s'obtenen el número de gens sobre expressats i infra expressats per teixit. D'aquests se'n fan dos rànkings:

- de teixits per número de gens sobre o infra expressats
- de teixits tenint en compte per cada gen la suma de diferències entre la mitjana en el teixit i la resta. Cal tenir en compte que tenim un diferent número de mostres per teixit.

S'han realitzat *heatmaps* amb els resultats obtinguts per veure l'agrupació dels teixits resultats.

Per veure l'existència de diferencies relacionades amb l'expressió de gens de l'alzheimer respecte dels que formen tot els transcriptoma ha calgut fer un estudi comparatiu amb 10000 execucions del mateix procés però en aquest cas escollint en cadascuna de les mateixes 644 gens diferents aleatoris.

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

Els resultats obtinguts s'han comparat amb el mateix anàlisi però aquest cop fet amb els 1000 gens associats a teixits del cervell més expressats. També s'ha repetit l'anàlisi de gens sobre expressats i infra expressats mitjançant 10.000 execucions amb grups del mateix número de gens seleccionats aleatoriament del transcriptoma. El propòsit d'aquest anàlisi és validar la significança dels resultats obtinguts.

Per cada teixit s'obté el nombre de vegades que els resultats obtinguts en cadascuna d'aquestes permutacions és superior als obtinguts en el cas dels gens lligats a l'Alzheimer.

En aquells casos en que s'acompleix:

$$\frac{\text{Nº gens expressats}}{10.000} < \frac{0.05}{13}$$

indicarà que hi ha una diferència significativa.

3. Resultats

3.1. Diferenciació de Teixits

A continuació el resultat dels models creats per diferenciar teixits emprant la seva expressió genètica.

En primer lloc es mostren els resultats obtinguts en l'estudi dels 1000 gens amb una major expressió ordenada de major a menor en base a la seva mitjana($TPM > 1$) a partir dels diferents conjunt de dades. Posteriorment es detallen els resultats obtinguts en una comparativa general tenint en compte tots els models realitzats en el total de 2641 mostres de teixit i 1000 gens.

Els resultats obtinguts en l'anàlisi dels 1000 gens més rellevants ens el diferents conjunts de dades ha estat:

- **Conjunt de dades segons TPM**

En aplicar PCA, la dimensionalitat ha passat del 1000 gens a 25 components principals.

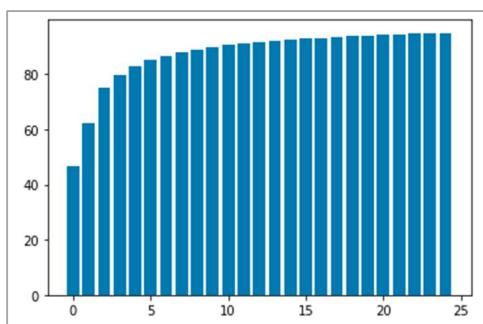


Figura 23: Gens TPM. Selecció dels 25 components principals explicatius d'un 95% de la variància

S'han realitzat gràfiques per veure la capacitat de diferenciar els teixits que tenen aquests components.

A partir dels 2 primers components ja pots observar-se una distinció entre teixits.

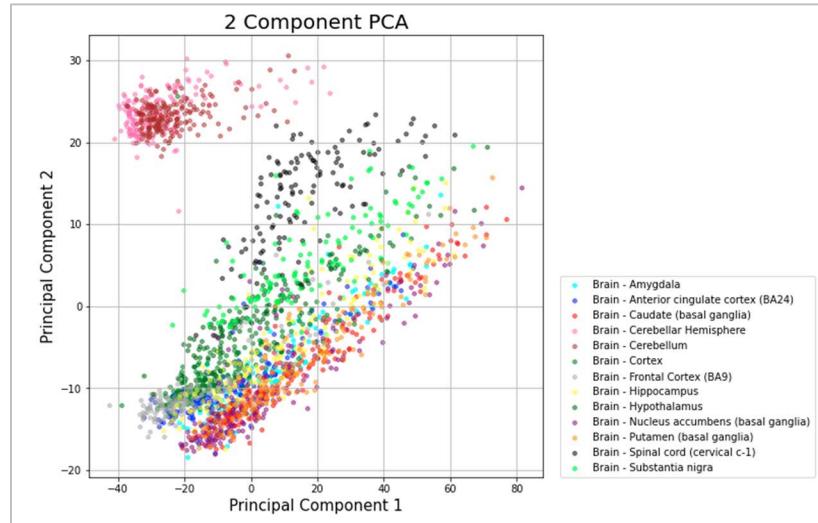


Figura 24: Visualització del Primer Component Principal vs el Segon Component Principal

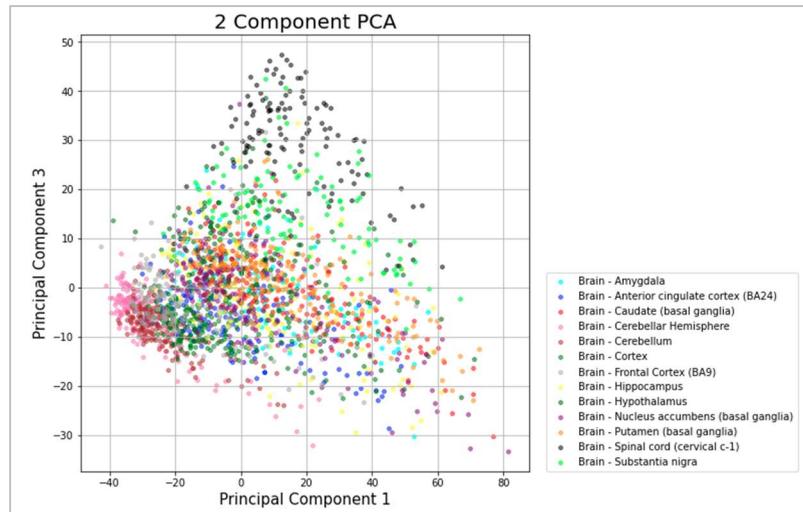


Figura 25: Visualització del Primer Component Principal vs el Tercer Component Principal

S'ha realitzat un t-sne per poder obtenir una visualització del conjunt de dades amb alta dimensionalitat.

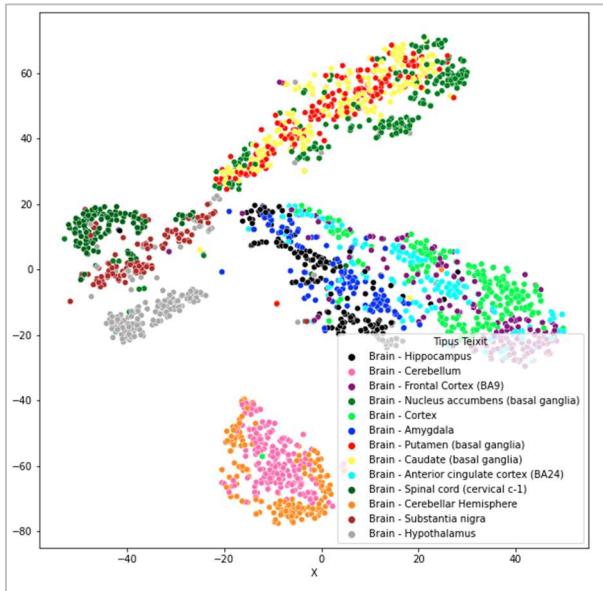


Figura 26: Representació 2D emprant t-SNE (perplexity=30)

S'hi observen les següents agrupacions:

- Cerebellar i Cerebellum
- Els teixits relacionats amb la Basal Ganglia: Nucleus, Putamen i Caudate s'agrupen.
- Els teixits relacionats amb el Cortex: Cortex, Frontal Cortex i l'Anterior Cingulate Cortex s'agrupen.
- En el mateix bloc tindríem a l'Amígda i l'Hippocampus.
- Substantia Nigra, Spinal Cord i Hipothalamus s'hi diferencien.

A partir dels components principals i del t-SNE ja es pot veure que hi ha una diferenciació entre teixits. Tot seguit s'han realitzat diferents models per veure amb quina precisió.

Amb el model kNN

```
(kNN) Millors hiperparàmetres trobats (cv)
{'n_neighbors': 8, 'weights': 'distance'} : 0.773676562639721 accuracy
(kNN) El accuracy de test és: 80.34026465028356%
```

Taula 9: Precisió obtinguda amb el model kNN emprant els 1000 gens més expressats

Amb el model SVM

```
(SVM) Millors hiperparàmetres trobats (cv)
{'C': 150, 'gamma': 0.0001} : 0.8977264598050614 accuracy
(SVM) El accuracy de test és: 92.06049149338375%
```

Taula 10: Precisió obtinguda amb el model SVM emprant els 1000 gens més expressats

Amb l'Arbre de Decisió

```
-----  
(Arbre de Decisió) Millors hiperparàmetres trobats (cv)  
-----  
{'max_depth': 550, 'min_samples_split': 2} : 0.7031118662255208 accuracy  
  
(Arbre de Decisió) El accuracy de test és: 71.83364839319471%
```

Taula 11: Precisió obtinguda amb el model Arbre de Decisió emprant els 1000 gens més expressats

Amb el Random Forest

```
-----  
(Random Forest) Millors hiperparàmetres trobats (cv)  
-----  
{'max_depth': 25, 'n_estimators': 250} : 0.8323929178216936 accuracy  
  
(Random Forest) El accuracy de test és: 86.95652173913044%
```

Taula 12: Precisió obtinguda amb el model Random Forest emprant els 1000 gens més expressats

Comparant els models s'observa com a millor precisió ha estat obtinguda amb el model SVM.

```
#####  
Model kNN Accuracy: 77.37%  
Model kNN, Accuracy de test és: 80.34%  
#####  
Model Svm Accuracy: 89.77%  
Model Svm, Accuracy de test és: 92.06%  
#####  
Model Arbre de Decisió: 70.31%  
Model Arbre de Decisió, Accuracy de test és: 71.83%  
#####  
Model Random Forest: 83.24%  
Model Random Forest, Accuracy de test és: 86.96%
```

Taula 13: Comparativa de resultats obtinguts amb els diferents models amb els 1000 gens més expressats

Per veure si s'obtenen diferents resultats tenint en compte tan sols els gens codificants.

- **Conjunt de dades segons TPM, amb tant sols aquells gens codificants**

En aplicar PCA, la dimensionalitat ha passat del 1000 gens a 21 components principals. Respecte a l'anàlisi anterior en aquest cas el nombre de components s'ha reduït.

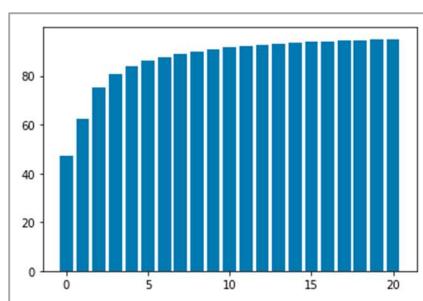


Figura 28: Gens codificants. Selecció dels 21 components principals explicatius d'un 95% de la variància.

S'han realitzat gràfiques per veure la capacitat de diferenciar els teixits que tenen aquests components.

A partir dels 2 primers components ja pots observar-se una distinció entre teixits.

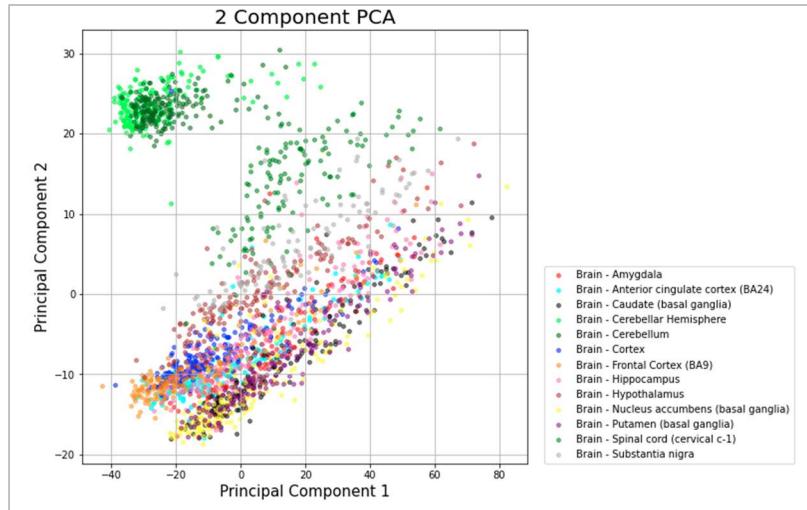


Figura 29: Gens codificant. Visualització del Primer Component Principal vs el Segon Component Principal

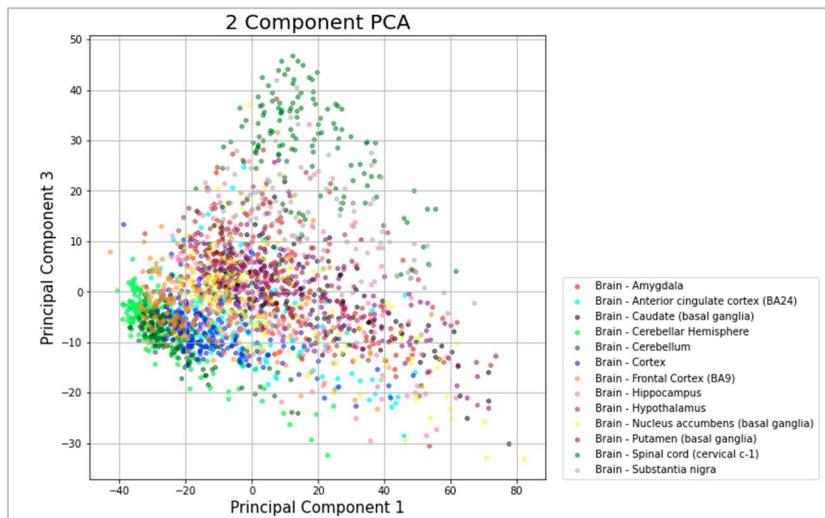


Figura 30: Gens codificant. Visualització del Primer Component Principal vs el Tercer Component Principal

S'ha realitzat un t-sne per poder obtenir una visualització del conjunt de dades amb alta dimensionalitat.

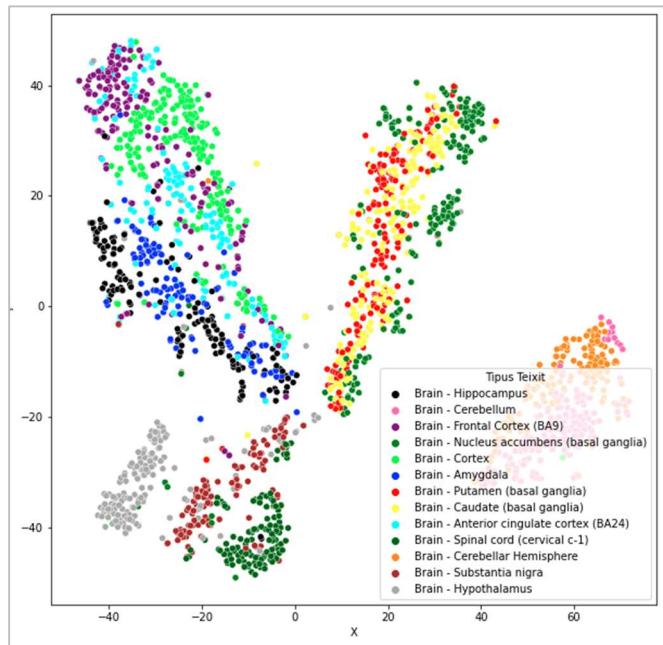


Figura 31: Gens codificant. Representació 2D emprant t-SNE (perplexity=30)

La formació dels grups és similar a l'obtinguda en el punt anterior al tenir en compte tots els gens:

- Cerebellar i Cerebellum
- Els teixits relacionats amb la Basal Ganglia: Nucleus, Putamen i Caudate s'agrupen.
- Els teixits relacionats amb el Cortex: Cortex, Frontal Cortex i l'Anterior Cingulate Cortex s'agrupen.
- En el mateix bloc tindríem a l'Amígda i l'Hippocampus.
- Substantia Nigra, Spinal Cord i Hipothalamus s'hi diferencien.

Els resultats dels diferents models ha estat:

Amb el model kNN

```
(kNN) Millors hiperparàmetres trobats (cv)
{'n_neighbors': 8, 'weights': 'distance'} : 0.7741437896807654 accuracy
(kNN) El accuracy de test és: 79.77315689981096%
```

Taula 14: Gens Codificant: Precisió obtinguda amb el model kNN emprant els 1000 gens més expressats

Amb el model SVM

```
(SVM) Millors hiperparàmetres trobats (cv)
{'C': 300, 'gamma': 0.0001} : 0.8953657337029419 accuracy
(SVM) El accuracy de test és: 92.24952741020795%
```

Taula 15: Gens Codificant: Precisió obtinguda amb el model SVM emprant els 1000 gens més expressats

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

Amb l'Arbre de Decisió

```
(Arbre de Decisió) Millors hiperparàmetres trobats (cv)
{'max_depth': 10, 'min_samples_split': 10} : 0.7154542609317713 accuracy
(Arbre de Decisió) El accuracy de test és: 72.77882797731569%
```

Taula 16: Gens Codificant: Precisió obtinguda amb el model Arbre de Decisió emprant els 1000 gens més expressats

Amb el Random Forest

```
(Random Forest) Millors hiperparàmetres trobats (cv)
{'max_depth': 25, 'n_estimators': 250} : 0.840921487972816 accuracy
(Random Forest) El accuracy de test és: 86.95652173913044%
```

Taula 17: Gens Codificant: Precisió obtinguda amb el model Random Forest emprant els 1000 gens més expressats

Comparant els models s'observa com el model amb millor precisió també ha estat el SVM.

```
#####
Model kNN Accuracy: 77.41%
Model kNN, Accuracy de test és: 79.77%
#####
Model Svm Accuracy: 89.54%
Model Svm, Accuracy de test és: 92.25%
#####
Model Arbre de Decisió: 71.55%
Model Arbre de Decisió, Accuracy de test és: 72.78%
#####
Model Random Forest: 84.09%
Model Random Forest, Accuracy de test és: 86.96%
```

Taula 18: Gens Codificant: Comparativa de resultats obtinguts amb els diferents models amb els 1000 gens més expressats

- **Conjunt de dades segons TPM, excloent els gens mitocondrials**

En aplicar PCA, la dimensionalitat ha passat del 1000 gens a 21 components principals. Respecte a l'anàlisi anterior en aquest cas el nombre de components s'ha reduït.

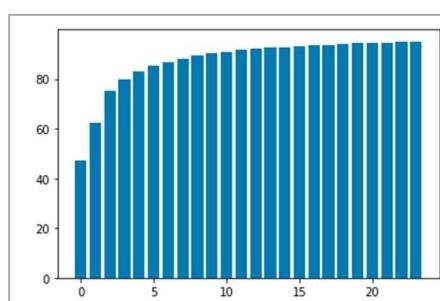


Figura 32: Gens no mitocondrials. Selecció dels 24 components principals explicatius d'un 95% de la variància.

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

El número de components necessaris augmenta al respecte del conjunt de dades amb els gens codificants però es lleugerament inferior a la primera amb tots els gens.

En les gràfiques per veure la capacitat de diferenciar els teixits que tenen aquests components es manté la capacitat de distinció entre teixits.

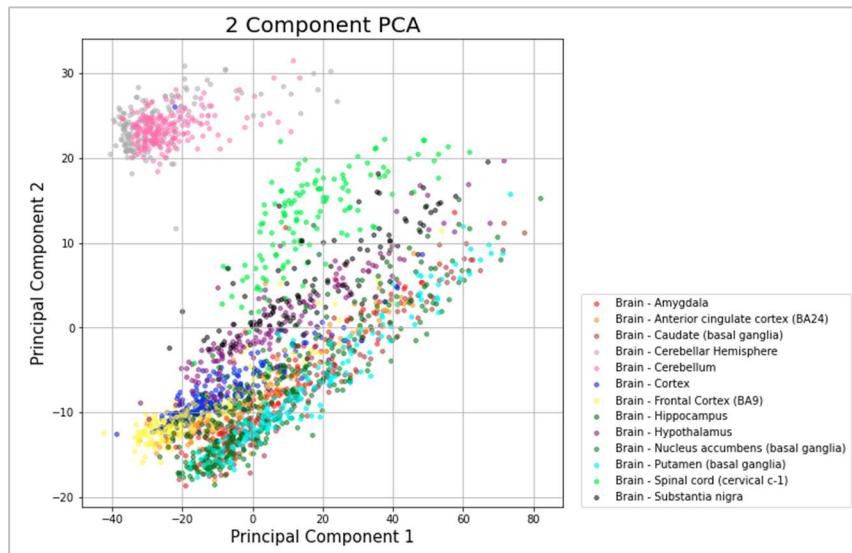


Figura 33: Gens codificants. Visualització del Primer Component Principal vs el Segon Component Principal

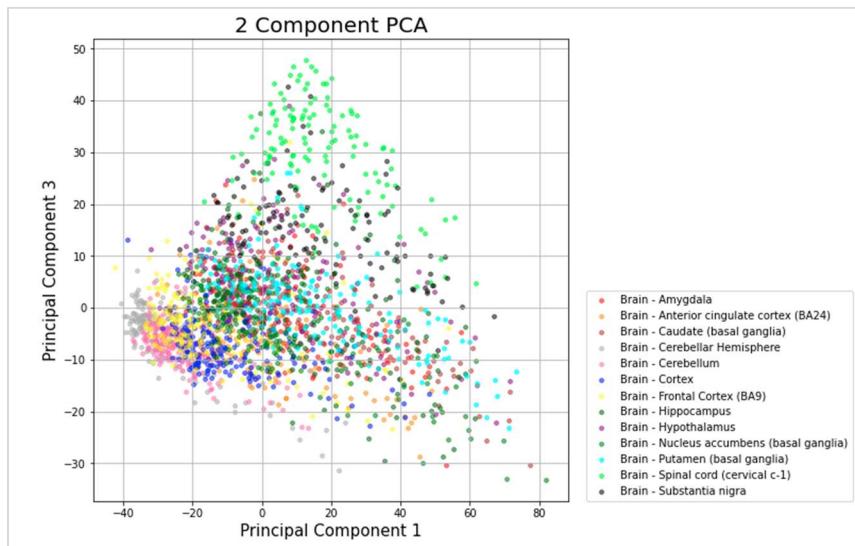


Figura 34: Gens codificants. Visualització del Primer Component Principal vs el Tercer Component Principal

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

Els resultats obtinguts a partir del t-SNE no mostren diferències respecte els anteriors

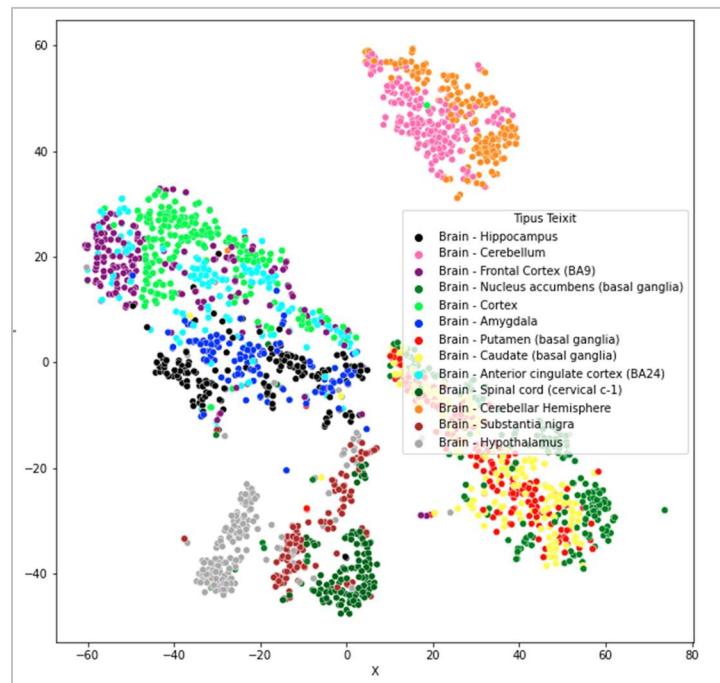


Figura 35: Gens codificant. Representació 2D emprant t-SNE (perplexity=30)

Els grups que es formen són els mateixos.

Amb el model kNN

```
(kNN) Millors hiperparàmetres trobats (cv)
-----
{'n_neighbors': 7, 'weights': 'distance'} : 0.7717785925064831 accuracy
(kNN) El accuracy de test és: 79.77315689981096%
```

Taula 19: Gens No Mitocondrials: Precisió obtinguda amb el model kNN emprant els 1000 gens més expressats

Amb el model SVM

```
(SVM) Millors hiperparàmetres trobats (cv)
-----
{'C': 1000, 'gamma': 1e-05} : 0.8963270142180095 accuracy
(SVM) El accuracy de test és: 92.62759924385632%
```

Taula 20: Gens No Mitocondrials: Precisió obtinguda amb el model SVM emprant els 1000 gens més expressats

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

Amb l'Arbre de Decisió

```
-----  
(Arbre de Decisió) Millors hiperparàmetres trobats (cv)  
-----  
{'max_depth': 10, 'min_samples_split': 10} : 0.706427166234463 accuracy  
  
(Arbre de Decisió) El accuracy de test és: 71.45557655954632%
```

Taula 21: Gens No Mitocondrials: Precisió obtinguda amb el model Arbre de Decisió emprant els 1000 gens més expressats

Amb el Random Forest

```
-----  
(Random Forest) Millors hiperparàmetres trobats (cv)  
-----  
{'max_depth': 20, 'n_estimators': 200} : 0.834758114995976 accuracy  
  
(Random Forest) El accuracy de test és: 86.01134215500946%
```

Taula 22: Gens No Mitocondrials: Precisió obtinguda amb el model Random Forest emprant els 1000 gens més expressats

Comparant els models s'observa com el model amb millor precisió ha estat també el SVM.

```
#####  
Model KNN Accuracy: 77.18%  
Model kNN, Accuracy de test és: 79.77%  
#####  
Model Svm Accuracy: 89.63%  
Model Svm, Accuracy de test és: 92.63%  
#####  
Model Arbre de Decisió: 70.64%  
Model Arbre de Decisió, Accuracy de test és: 71.46%  
#####  
Model Random Forest: 83.48%  
Model Random Forest, Accuracy de test és: 86.01%
```

Taula 23: Gens No Mitocondrials: Comparativa de resultats obtinguts amb els diferents models amb els 1000 gens més expressats

L'estudi s'ha realitzat amb diferent nombre de gens per tal de reduir la dimensionalitat. El número de mostres de teixits del que es disposa és limitat i cal evitar l'*overfitting* en els models. S'han fet models amb 5000, 1000, 500 i 100 gens per veure amb quin grup reduït de gens es podien diferenciar els teixits.

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

El quadre resum on es mostren les precisions de tots els models obtinguts a partir dels diferents ànalisis és el següent:

Conjunt de Dades	Model	Precisió del Model en %			
		Gens			
		5000	1000	500	100
TPM RAW	kNN	77,65	77,37	75,28	72,26
	SVM	90,48	89,77	88,49	84,38
	Arbre Decisió	71,22	70,31	69,56	60,7
	Random Forest	85,46	83,24	82,39	76,56
TPM PROT	kNN	76,61	77,41	75,05	73,39
	SVM	89,92	89,54	89,02	85,09
	Arbre Decisió	71,5	71,55	70,5	63,54
	Random Forest	83,86	84,09	82,34	78,65
TPM NO MITO	kNN	77,13	77,18	75,19	71,26
	SVM	90,01	89,63	88,78	84
	Arbre Decisió	70,74	70,64	68	58,48
	Random Forest	84,71	83,48	81,86	75,9
CV RAW	kNN	82,44	81,39	77,37	58,29
	SVM	94,37	92,99	90,2	77,41
	Arbre Decisió	73,91	72,54	67,85	44,98
	Random Forest	90,53	87,93	83,52	65,91
CV PROT	kNN	82,06	82,24	80,59	71,07
	SVM	94,13	93,75	92	84,52
	Arbre Decisió	74,71	73,39	70,08	58
	Random Forest	90,49	89,44	86,46	77,42
CV NO MITO	kNN	77,13	77,18	75,19	71,26
	SVM	89,87	89,63	88,73	84
	Arbre Decisió	70,64	70,41	68,28	59
	Random Forest	84,61	83,67	81,77	75,85

Taula 24: Precisions dels models obtingudes pels diferents conjunts de dades i diferent nombre de gens

3.2. Anàlisi d'Expressió

Abans de presentar els resultats de l'expressió, es mostren uns números previs sobre l'expressió del conjunt de dades original amb tots els gens.

De GWAS i relacionats amb l'Alzheimer:

- Hi ha un total de 73 estudis
- Hi ha un total de 1432 gens

Dels 52000 gens originals i fent referència a la seva expressió com a mitjana(tpm)>1:

- 1059 s'expressen tan sols en 1 teixit individual del cervell.
- 21231 s'expressen en 1 o més teixits del cervell.
- 17515 expressats en mitjana(TPM)>1 considerant tots els teixits

Amb el següent criteri de selecció aplicat:

Tots aquells que compleixin alguna de les següents condicions:

- C1: tinguin mean(TPM)>1
- C2: agrupats per Teixits tinguin un mean(TPM)>1 en algun d'ells
- 644 gens són els gens relacionats amb l'Alzheimer a partir de GWAS i presents segons el criteri de selecció aplicat

Tot seguit presenten els resultats obtinguts en la recerca de quines són les parts del cervell on s'expressen gens relacionats amb la malaltia de l'Alzheimer.

Recordar que el conjunt de dades final, amb els gens lligats a l'Alzheimer i que complien els criteris previs d'expressió en teixits, està format per un total de 644 gens.

Els resultats obtinguts en la quantificació del número de gens sobre expressats i infra expressats relacionats amb l'Alzheimer per teixits han estat obtinguts de 2 maneres:

- una primera tenint en compte el número de gens i
- una segona tenint en compte la suma de la diferència entre la mitjana del teixit respecte la mitjana de tots els altres per cada gen en aquells casos en que el gen es sobre expressa o infra expressa.

Resultats amb gens lligats a l'Alzheimer

- Considerant la quantitat de gens

En el cas dels sobre expressats:

	Teixit	Total
Brain - Cerebellum	376.0	
Brain - Cerebellar Hemisphere	360.0	
Brain - Frontal Cortex (BA9)	293.0	
Brain - Cortex	289.0	
Brain - Spinal cord (cervical c-1)	262.0	
Brain - Hypothalamus	166.0	
Brain - Substantia nigra	152.0	
Brain - Caudate (basal ganglia)	135.0	
Brain - Anterior cingulate cortex (BA24)	124.0	
Brain - Nucleus accumbens (basal ganglia)	104.0	
Brain - Hippocampus	83.0	
Brain - Putamen (basal ganglia)	76.0	
Brain - Amygdala	59.0	

Taula 25: Número de gens lligats a l'Alzheimer sobre expressats per teixit

En el cas dels infra expressats:

	Teixit	Total
Brain - Putamen (basal ganglia)	-395.0	
Brain - Hippocampus	-327.0	
Brain - Amygdala	-319.0	
Brain - Nucleus accumbens (basal ganglia)	-304.0	
Brain - Caudate (basal ganglia)	-293.0	
Brain - Substantia nigra	-277.0	
Brain - Anterior cingulate cortex (BA24)	-245.0	
Brain - Cerebellar Hemisphere	-230.0	
Brain - Cerebellum	-207.0	
Brain - Spinal cord (cervical c-1)	-199.0	
Brain - Hypothalamus	-173.0	
Brain - Cortex	-145.0	
Brain - Frontal Cortex (BA9)	-107.0	

Taula 26: Número de gens lligats a l'Alzheimer infra expressats per teixit

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

- Considerant la diferència entre la mitjana del teixit respecte la mitjana de tots els altres per cada gen en aquells casos en que el gen es sobre expressa o infra expressa

En el cas dels sobre expressats:

	Teixit	Total
Brain - Cerebellum	577.406161	
Brain - Cerebellar Hemisphere	574.164672	
Brain - Spinal cord (cervical c-1)	372.501016	
Brain - Frontal Cortex (BA9)	262.898648	
Brain - Cortex	230.487880	
Brain - Substantia nigra	183.415407	
Brain - Hypothalamus	172.427146	
Brain - Caudate (basal ganglia)	130.350246	
Brain - Anterior cingulate cortex (BA24)	127.904418	
Brain - Nucleus accumbens (basal ganglia)	124.248165	
Brain - Hippocampus	79.230336	
Brain - Putamen (basal ganglia)	70.767958	
Brain - Amygdala	50.628131	

Taula 27: Rànking de gens sobreexpressats lligats a l'Alzheimer obtinguts a partir de la Suma de la diferència entre la mitjana del teixit i la mitjana de la resta.

En el cas dels infra expressats:

	Teixit	Total
Brain - Cerebellar Hemisphere	384.270077	
Brain - Putamen (basal ganglia)	358.494649	
Brain - Cerebellum	322.619204	
Brain - Substantia nigra	292.725827	
Brain - Spinal cord (cervical c-1)	279.837682	
Brain - Amygdala	234.314261	
Brain - Hippocampus	221.214503	
Brain - Caudate (basal ganglia)	216.911223	
Brain - Nucleus accumbens (basal ganglia)	215.503857	
Brain - Anterior cingulate cortex (BA24)	159.440671	
Brain - Hypothalamus	106.176105	
Brain - Cortex	98.543214	
Brain - Frontal Cortex (BA9)	83.799925	

Taula 28: Rànking de gens infraexpressats lligats a l'Alzheimer obtinguts a partir de la Suma de la diferencia entre la mitjana del teixit i la mitjana de la resta.

Per poder comprovar que a partir de la selecció de gens de forma aleatòria els teixits s'agrupen de forma similar a l'obtinguda a partir dels gens lligats a l'Alzheimer posteriorment es realitza un Heatmap.

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

Pot observar-se com s'agrupen els teixits del *Cortex (Anterior, Cortex i Frontal)* i com s'agrupen els teixits de la *Ganglia (Nucleus, Caudate, Putamen)* obtenint uns resultats similars als que havíem obtingut a partir de la recerca del primer objectiu del treball en veure si podien diferenciar-se els teixits en funció de l'expressió genètica.

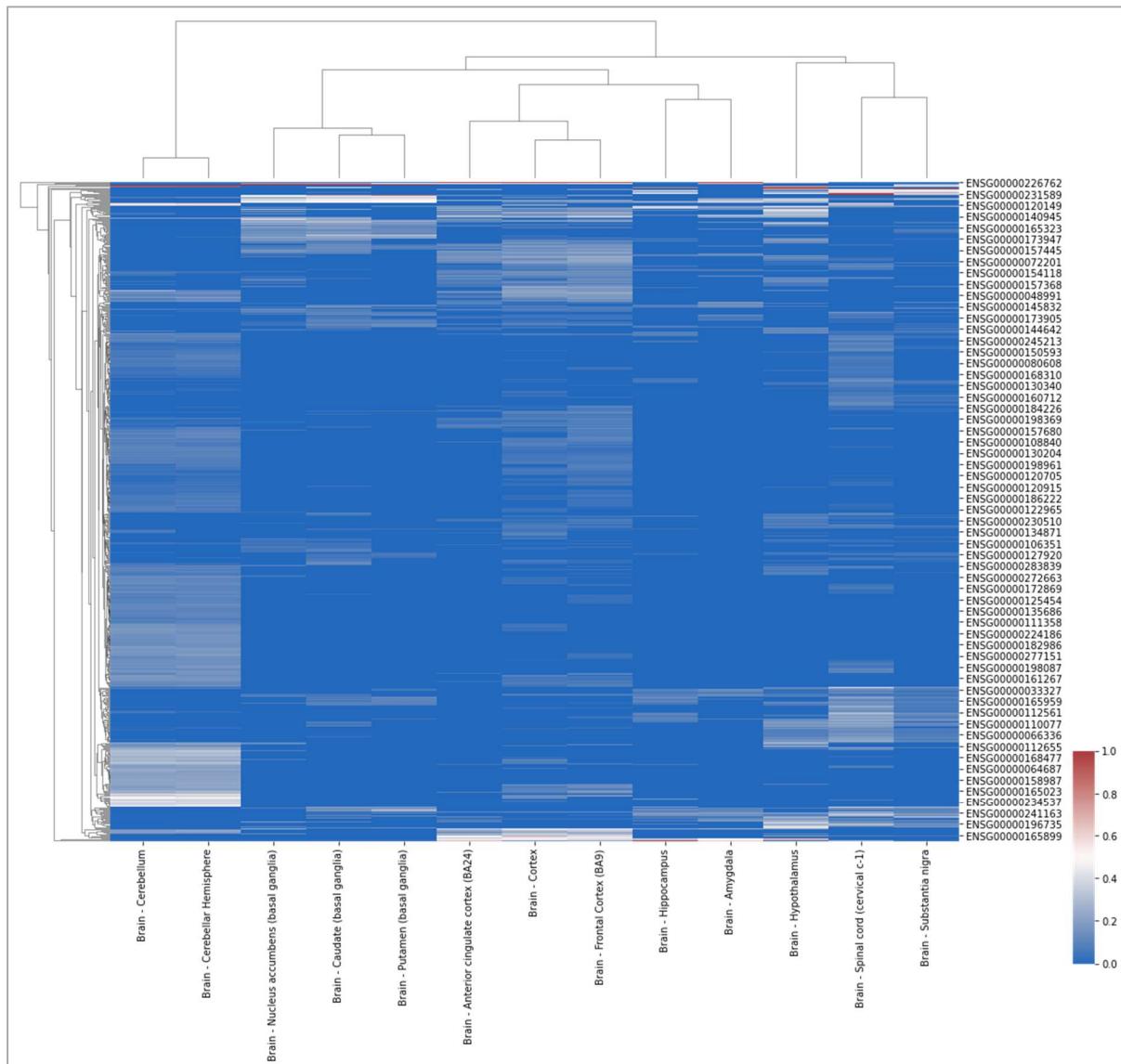


Figura 36: Heatmap a partir dels gens sobre expressats per Teixit lligats a l'Alzheimer tenint en compte la suma de diferències entre la mitjana del teixit i la mitjana de la resta

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

Estudi realitzat a partir dels 1000 gens amb major expressió en TPM del transcriptoma complert.

- Considerant la quantitat de gens sobre expressats

En el cas dels sobre expressats:

	Teixit	Total
Brain - Cerebellar Hemisphere	669.0	
Brain - Cerebellum	626.0	
Brain - Frontal Cortex (BA9)	603.0	
Brain - Cortex	463.0	
Brain - Spinal cord (cervical c-1)	375.0	
Brain - Anterior cingulate cortex (BA24)	195.0	
Brain - Substantia nigra	176.0	
Brain - Hypothalamus	133.0	
Brain - Nucleus accumbens (basal ganglia)	125.0	
Brain - Caudate (basal ganglia)	122.0	
Brain - Amygdala	107.0	
Brain - Putamen (basal ganglia)	103.0	
Brain - Hippocampus	94.0	

Taula 29: Rànking de la sobre expressió dels Top 1000 gens del transcriptoma sobre expressats per Teixit tenint en compte el seu número

En el cas dels infra expressats:

	Teixit	Total
Brain - Putamen (basal ganglia)	-686.0	
Brain - Hippocampus	-576.0	
Brain - Caudate (basal ganglia)	-549.0	
Brain - Substantia nigra	-420.0	
Brain - Amygdala	-395.0	
Brain - Spinal cord (cervical c-1)	-384.0	
Brain - Nucleus accumbens (basal ganglia)	-347.0	
Brain - Hypothalamus	-253.0	
Brain - Anterior cingulate cortex (BA24)	-237.0	
Brain - Cortex	-235.0	
Brain - Cerebellum	-233.0	
Brain - Cerebellar Hemisphere	-206.0	
Brain - Frontal Cortex (BA9)	-127.0	

Taula 30: Rànking de la infra expressió dels Top 1000 gens del transcriptoma infra expressats per Teixit tenint en compte el seu número

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

- Considerant la diferència entre la mitjana del teixit respecte la mitjana de tots els altres per cada gen en aquells casos en que el gen es sobre expressa o infra expressa

En el cas dels sobre expressats:

	Teixit	Total
Brain - Cerebellar Hemisphere	690.337007	
Brain - Cerebellum	620.399209	
Brain - Frontal Cortex (BA9)	416.243861	
Brain - Spinal cord (cervical c-1)	376.791798	
Brain - Cortex	301.403125	
Brain - Substantia nigra	142.654182	
Brain - Anterior cingulate cortex (BA24)	139.698719	
Brain - Hypothalamus	118.535711	
Brain - Nucleus accumbens (basal ganglia)	102.616742	
Brain - Caudate (basal ganglia)	95.656692	
Brain - Hippocampus	78.676550	
Brain - Putamen (basal ganglia)	77.877720	
Brain - Amygdala	74.344285	

Taula 31: Rànking de la sobre expressió dels Top 1000 gens del transcriptoma sobre expressats per Teixit tenint en compte la suma de diferències entre la mitjana del teixit i la mitjana de la resta

Els resultats són similars als anteriors obtinguts en el cas de l'Alzheimer.

En el cas dels infra expressats:

	Teixit	Total
Brain - Spinal cord (cervical c-1)	509.771483	
Brain - Putamen (basal ganglia)	454.208203	
Brain - Substantia nigra	404.965866	
Brain - Hippocampus	302.945958	
Brain - Caudate (basal ganglia)	296.945084	
Brain - Cerebellar Hemisphere	251.941079	
Brain - Amygdala	239.928700	
Brain - Cerebellum	235.829033	
Brain - Nucleus accumbens (basal ganglia)	200.030554	
Brain - Hypothalamus	173.511278	
Brain - Anterior cingulate cortex (BA24)	116.972007	
Brain - Cortex	98.994960	
Brain - Frontal Cortex (BA9)	55.421050	

Taula 32: Rànking de la infra expressió dels Top 1000 gens del transcriptoma infra expressats per Teixit tenint en compte la suma de diferències entre la mitjana del teixit i la mitjana de la resta

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

Per comprobar que agafant gens de manera aleatòria els teixits s'agrupen de manera similar a l'obtinguda a partir dels gens lligats a l'Alzheimer es realitza un Heatmap per comparar els clústers de teixits resultants.

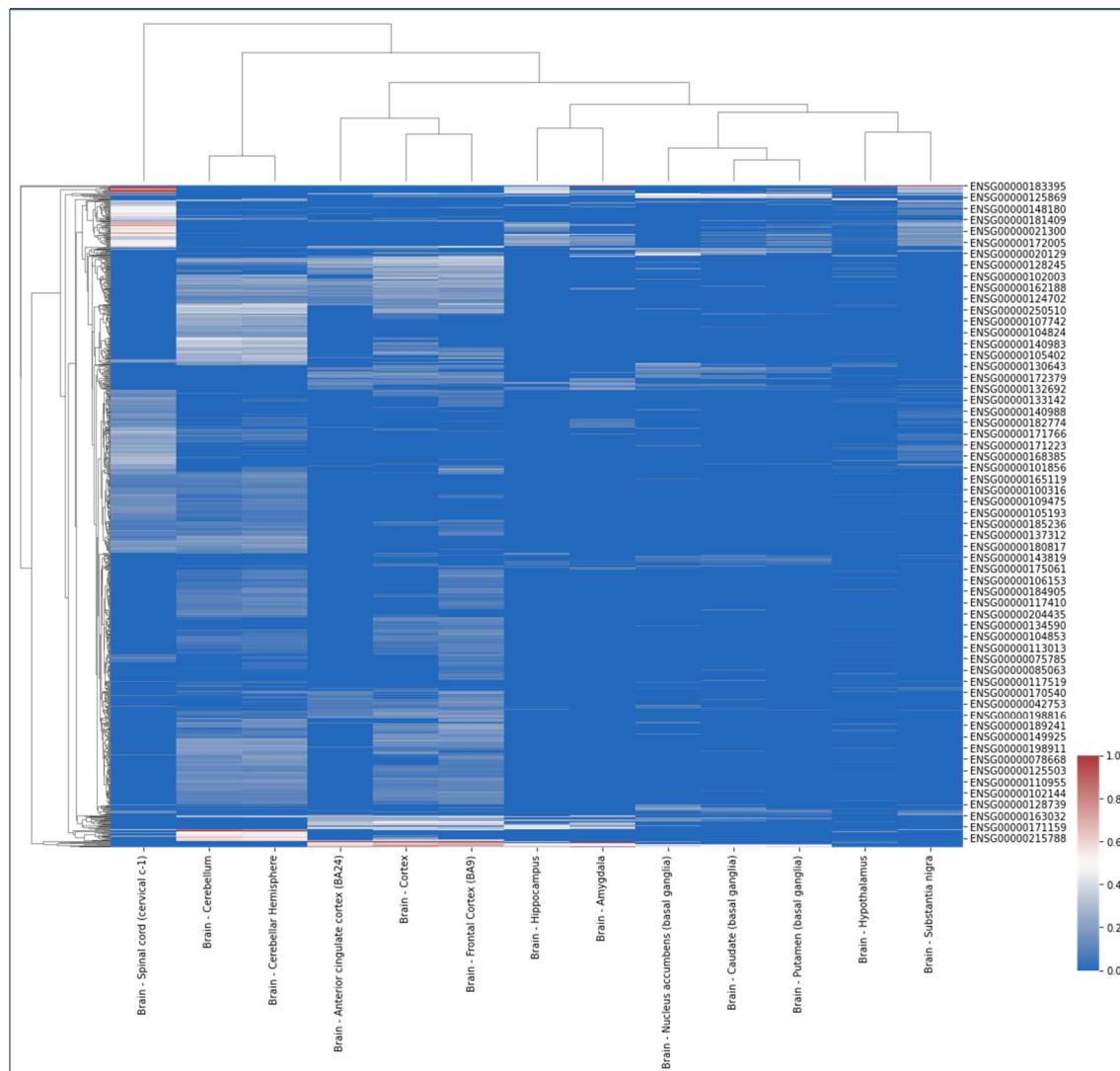


Figura 37: Heatmap de la sobre expressió dels Top 1000 gens del transcriptoma tenint en compte la suma de diferències entre la mitjana del teixit i la mitjana de la resta

Com que aquests clústers de teixits obtinguts són similars als previs a partir dels gens relacionats amb l'Alzheimer es realitzarà un estudi amb 10.000 permutacions per trobar diferències.

Resultats de l'estudi amb 10.000 permutacions

El Número de casos en que el número de Gens sobre expressats respecte els GWAS és major després de 10000 iteracions (Taula 34).

Respecte als sobre expressats:

Cortex	0
Cerebellum	10.000
Frontal	0
Caudate	0
Nucleus	35
Putamen	0
Hypothalamus	0
Spinal	276
Hippocampus	0
Anterior	0
Cerebellar	10.000
Substantia	0
Amygdala	0

Taula 33: Número de casos en que el número de Gens sobre expressats respecte els GWAS és major després de 10000 iteracions

Cerebellum i Cerebellar, en les 10.000 iteracions sempre han tingut valors de sobre expressió superiors a les obtingudes a partir dels gens lligats a l'Alzheimer. Per tant, la sobre expressió obtinguda d'aquests teixits a partir de gens de l'Alzheimer no es representativa.

En el cas d'aquest estudi de l'Alzheimer, comparant els 'Rànking 2 (Alz)' de teixits més sobre expressats amb aquests darrers resultats es descartarien:

- Cerebellum,
- Spinal,
- Cerebellar

I es seleccionarien els teixits a estudiar lligats a l'Alzheimer:

- Frontal Cortex
- Cortex
- Substantia Nigra
- Hypothalamus
- Caudate
- Anterior
- Nucleus
- Hippocampus
- Putamen
- Amygdala

4. Discussió

A partir dels resultats obtinguts dels clústers realitzats s'ha posat de manifest la distinció dels teixits cerebrals mitjançant la seva expressió genètica. Amb l'aplicació de PCA s'ha pogut gestionar l'elevada multidimensionalitat per acabar tenint uns resultats on s'ha observat que amb un nombre reduït de gens de 1000, 500 o fins i tot 100 gens els models obtinguts han estat força bons.

De tots els diferents conjunts de dades, el millor model amb una precisió del 94,37%, ha estat l'obtingut a partir dels 5000 gens amb major CV aplicant SVM. Globalment, el conjunt de dades que permet una millor distinció entre teixits és el format pels gens codificant proteïnes, escollits segons el criteri de CV.

Els primers gens amb major CV són els que tenen una major capacitat de diferenciació entre teixits. Amb 100 gens s'ha obtingut un model amb una precisió superior al 84%.

A partir de l'anàlisi de sobre expressió en el cas de l'AD s'han obtingut sobre expressions significatives en els següents teixits: *Frontal Cortex, Cortex, Substantia Nigra, Hypothalamus, Caudate, Anterior, Nucleus, Putamen, Hippocampus i Amygdala*.

Contrastant els resultats obtinguts de la sobre expressió amb d'altres estudis s'observa com l'AD afecta a múltiples teixits del cervell. Entre aquests teixits hi destaquen els corticals que són els que han presentat més sobre expressió en aquest treball. Altres estudis on s'han detectat afectacions en el teixits del *Cortex* lligats a AD són els Souza et al. (2013) que fan un estudi de recerca sobre la diagnosi de l'AD lligats a certs síndromes corticals. També Wong et al. (2014) estudien relacions entre l'AD i el *Prefrontal Cortex* en disfuncions lligades a la pèrdua de memòria. Més recentment, Fracassi et al. (2021) realitzen un estudi lligat a la progressiva neurodegeneració existent en els teixits del *Cortex cerebral* en relació a les persones afectades per AD.

Respecte de la resta de teixits on el present treball ha observat sobre expressió, els resultats són coincidents amb els de Agarwal et al. (2020) que en el seu treball observen com la *Substantia Nigra* té relació amb diferents ND com ara AD i PD.

També Tao et al. (2021) troben evidències d'atròfia en l'*Hipotàlem* en les etapes inicials de AD. Altres estudis que versen sobre aquest teixit i l'AD és el de Lyra e Silva et al. (2021) en que indiquen la necessitat d'aprofundir en la recerca d'alteracions del mateix lligats a AD.

Almeida et al. (2003) analitzen la relació entre el volum del *Caudate* i afectacions de ND com ara PD, PD+LB o AD. En el mateix i entre altres conclusions indiquen que els individus amb AD pateixen d'una reducció de tot el cervell, *Caudate* inclòs, respecte als controls i aquells que pateixen PD.

Udo et al. (2020) en seu treball sobre l'apatia i com aquesta es un símptoma comú en pacients AD detecten afectacions en el teixit del *Caudate*.

Liu et al. (2017) en el seu treball sobre l'AD fan un estudi on s'hi relaciona l'*Anterior Cingulate Cortex*.

Diferències d'expressió genètica en diferents zones del cervell i la seva relació amb malalties neurodegeneratives

de Jong et al. (2008) observen importants reduccions en el volum del *Putamen* i *Thalamus* en l'AD.

Poulin et al. (2011) en el seu treball observen com existeix una prominent atrofia de l'amígdala en etapes inicials d'AD.

5. Conclusions

L'objectiu principal d'aquest estudi era avaluar si la diferenciació entre teixits del cervell emprant la seva expressió genètica era possible. Amb els resultats obtinguts aquesta diferenciació s'ha pogut comprovar.

Un segon objectiu consistia en cercar en quines parts del cervell s'hi expressaven gens relacionats amb malalties neurodegeneratives. S'ha fet un estudi centrat en la malaltia de l'Alzheimer obtenint els teixits amb una major sobre expressió. En aquest respecte, i a partir dels resultats obtinguts, s'ha observat com l'afectació de l'Alzheimer es troba en diferents teixits del cervell i no es focalitza en un teixit concret. Entre aquests teixits destaquen els teixits corticals.

El propòsit d'avaluar altres malalties neurodegeneratives, a banda de l'Alzheimer, no s'ha realitzat per manca de temps malgrat disposar de la metodologia i tenir el pipeline desenvolupat.

A partir d'aquest estudi s'obren noves línies de treball per l'anàlisi d'altres malalties neurodegeneratives. AD és molt complexa, en aquest estudi s'han utilitzat mostres de persones donants no afectades per l'Alzheimer i fora important que els futurs estudis empressin mostres cerebrals de teixits de persones afectades per Alzheimer.

Cal tenir present, en relació a GWAS i com ha estat realitzada l'elecció dels gens lligats a les malalties, que en la seva elecció s'ha seguit el criteri de selecció per proximitat però es podria haver emprat el propi criteri de l'investigador. GWAS ens dona una idea de les probabilitats de patir una malaltia des de que naixem però existeixen altres factors ambientals que ens poden condicionar a patir-la i llarg de la vida. Aquests factors ambientals addicionals no han estat tinguts en compte.

Aquest treball mostra la rellevància que té l'estudi de les malalties neurodegeneratives com ara l'Alzheimer en relació als teixits que s'hi veuen afectats i s'encoratja a futures investigacions a seguir aquest tipus d'enfoc tenint present els punts de millora.

6. Glossari

AD:

Malaltia de l'Alzheimer.

ADN:

Àcid desoxiribonucleic

ARN:

Àcid ribonucleic

CV:

Coeficient de variació de Pearson

DLB:

Malaltia Cossos de Lewy

GWAS:

Estudis d'associació del genoma complert

GTEX:

Genotype-Tissue Expression

kNN:

k veïns més propers

OMS:

Organització Mundial de la Salut

PCA:

Anàlisis de components principals

PD:

Malaltia de Parkinson

SNP:

Polimorfisme d'un sol nucleòtid

SVM:

Support Vector Machines. Màquines de Vector Suport

TPM:

Transcripts Per Kilobase Million

7. Bibliografia

- Agarwal, D., Sandor, C., Volpato, V., Caffrey, T. M., Monzón-Sandoval, J., Bowden, R., . . . Webber, C. (2020, 8). A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders. *Nature Communications*, 11, 4183. doi:10.1038/s41467-020-17876-0
- Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., . . . Volpi, S. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509). doi:10.1126/SCIENCE.AAZ1776
- Akçimen, F., Martins, S., Liao, C., Bourassa, C. V., Catoire, H., Nicholson, G. A., . . . Rouleau, G. A. (2020). Genome-wide association study identifies genetic factors that modify age at onset in Machado-Joseph disease. *Aging*, 12(6). doi:10.18632/aging.102825
- Almeida, O. P., Burton, E. J., McKeith, I., Ghokal, A., Burn, D., & O'Brien, J. T. (2003). MRI study of caudate nucleus volume in Parkinson's disease with and without dementia with Lewy bodies and Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, 16, 57–63. doi:10.1159/000070676
- Ardlie, K. G., DeLuca, D. S., Segrè, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., . . . Lockhart. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235). doi:10.1126/science.1262110
- Brynedal, B., Wojcik, J., Esposito, F., Debailleul, V., Yaouanq, J., Martinelli-Boneschi, F., . . . Abderrahim, H. (2010). MGAT5 alters the severity of multiple sclerosis. *Journal of Neuroimmunology*, 220(1-2). doi:10.1016/j.jneuroim.2010.01.003
- Chung, S. J., Armasu, S. M., Biernacka, J. M., Anderson, K. J., Lesnick, T. G., Rider, D. N., . . . Maraganore, D. M. (2012). Genomic determinants of motor and cognitive outcomes in Parkinson's disease. *Parkinsonism and Related Disorders*, 18(7). doi:10.1016/j.parkreldis.2012.04.025
- Clarimon, J., Moreno-Grau, S., Cervera-Carles, L., Dols-Icardo, O., Sánchez-Juan, P., & Ruiz, A. (2020). Genetic architecture of neurodegenerative dementias. *Genetic architecture of neurodegenerative dementias*, 168. doi:10.1016/j.neuropharm.2020.108014
- Consortium, E. P., Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., a Davis, C., . . . Lochovsky, L. (2013). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414).
- Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258). doi:10.1038/227561a0
- de Estadística, I. N. (2018). Defunciones según la Causa de Muerte. Principales causas de muerte por grupos de enfermedades. *Defunciones según la Causa de Muerte. Principales causas de muerte por grupos de enfermedades*.
- de Jong, L. W., van der Hiele, K., Veer, I. M., Houwing, J. J., Westendorp, R. G., Bollen, E. L., . . . van der Grond, J. (2008, 12). Strongly reduced volumes of

- putamen and thalamus in Alzheimer's disease: an MRI study. *Brain: A Journal of Neurology*, 131, 3277–3285. doi:10.1093/brain/awn278
- de la Salud (OMS), O. M. (2013). OMS | Los trastornos neurológicos afectan a millones de personas en todo el mundo: informe de la OMS. *OMS | Los trastornos neurológicos afectan a millones de personas en todo el mundo: informe de la OMS*.
- de Souza, L. C., Bertoux, M., Funkiewiez, A., Samri, D., Azuar, C., Habert, M.-O., . . . Dubois, B. (2013). Frontal presentation of Alzheimer's disease: a series of patients with biological evidence by CSF biomarkers. *Dementia & Neuropsychologia*, 7, 66–74. doi:10.1590/S1980-57642013DN70100011
- Ensembl genome browser*. (s.f.). Obtenido de <https://www.ensembl.org/index.html>
- Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., . . . Price, A. L. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics*, 50(4). doi:10.1038/s41588-018-0081-4
- Fracassi, A., Marcatti, M., Zolochevska, O., Tabor, N., Woltjer, R., Moreno, S., & Tagliafata, G. (2021, 1). Oxidative Damage and Antioxidant Response in Frontal Cortex of Demented and Nondemented Individuals with Alzheimer's Neuropathology. *Journal of Neuroscience*, 41, 538–554. doi:10.1523/JNEUROSCI.0295-20.2020
- Garcés, M. (2016). Estudio sobre las enfermedades neurodegenerativas en España y su impacto económico y social. *Estudio sobre las enfermedades neurodegenerativas en España y su impacto económico y social*.
- Gerring, Z. F., Lupton, M. K., Edey, D., Gamazon, E. R., & Derkx, E. M. (2020). An analysis of genetically regulated gene expression across multiple tissues implicates novel gene candidates in Alzheimer's disease. *Alzheimer's Research and Therapy*, 12(1). doi:10.1186/s13195-020-00611-8
- Home | HUGO Gene Nomenclature Committee*. (s.f.). Obtenido de <https://www.genenames.org/>
- Kamboh, M. I., Demirci, F. Y., Wang, X., Minster, R. L., Carrasquillo, M. M., Pankratz, V. S., . . . Barmada, M. M. (2012). Genome-wide association study of Alzheimer's disease. *Translational Psychiatry*, 2. doi:10.1038/tp.2012.45
- Liu, X., Chen, W., Hou, H., Chen, X., Zhang, J., Liu, J., . . . Bai, G. (2017, 5). Decreased functional connectivity between the dorsal anterior cingulate cortex and lingual gyrus in Alzheimer's disease patients with depression. *Behavioural Brain Research*, 326, 132–138. doi:10.1016/j.bbr.2017.01.037
- Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., . . . Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *The Genotype-Tissue Expression (GTEx) project*, 45(6). doi:10.1038/ng.2653
- Lyra e Silva, N. M., Gonçalves, R. A., Pascoal, T. A., Lima-Filho, R. A., Resende, E. d., Vieira, E. L., . . . De Felice, F. G. (2021, 4). Pro-inflammatory interleukin-6 signaling links cognitive impairments and peripheral metabolic alterations in Alzheimer's disease. *Translational Psychiatry*, 11, 1–15. doi:10.1038/s41398-021-01349-z

- Miron, J., Picard, C., Nilsson, N., Frappier, J., Dea, D., Théroux, L., & Poirier, J. (2018). CDK5RAP2 gene and tau pathophysiology in late-onset sporadic Alzheimer's disease. *Alzheimer's and Dementia*, 14(6). doi:10.1016/j.jalz.2017.12.004
- Moss, D. J., Tabrizi, S. J., Mead, S., Lo, K., Pardiñas, A. F., Holmans, P., . . . Tan, L. (2017). Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. *The Lancet Neurology*, 16(9). doi:10.1016/S1474-4422(17)30161-8
- Nalls, M. A., Plagnol, V., Hernandez, D. G., Sharma, M., Sheerin, U. M., Saad, M., . . . Wood, N. W. (2011). Imputation of sequence variants for identification of genetic risks for Parkinson's disease: A meta-analysis of genome-wide association studies. *The Lancet*, 377(9766). doi:10.1016/S0140-6736(10)62345-8
- Poulin, S. P., Dautoff, R., Morris, J. C., Barrett, L. F., & Dickerson, B. C. (10 de 2011). Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. *Psychiatry research*, 194, 7–13. doi:10.1016/j.psychresns.2011.06.014
- Tao, A., Myslinski, Z., Pan, Y., Iadecola, C., Dyke, J., Chiang, G., & Ishii, M. (2021, 4). Hypothalamic Atrophy in Alzheimer's Disease (1819). *Neurology*, 96. Retrieved from https://n.neurology.org/content/96/15_Supplement/1819
- Udo, N., Hashimoto, N., Toyonaga, T., Isoyama, T., Oyanagi, Y., Narita, H., . . . Kusumi, I. (2020). Apathy in Alzheimer's Disease Correlates with the Dopamine Transporter Level in the Caudate Nuclei. *Dementia and Geriatric Cognitive Disorders Extra*, 10, 86–93. doi:10.1159/000509278
- Vergouw, L. J., van Steenoven, I., van de Berg, W. D., Teunissen, C. E., van Swieten, J. C., Bonifati, V., . . . de Jong, F. J. (2017). An update on the genetics of dementia with Lewy bodies. *An update on the genetics of dementia with Lewy bodies*, 43. doi:10.1016/j.parkreldis.2017.07.009
- Wong, S., Flanagan, E., Savage, G., Hodges, J. R., & Hornberger, M. (2014, 2). Contrasting Prefrontal Cortex Contributions to Episodic Memory Dysfunction in Behavioural Variant Frontotemporal Dementia and Alzheimer's Disease. *PLOS ONE*, 9, e87778. doi:10.1371/journal.pone.0087778
- Yamazaki, Y., Zhao, N., Caulfield, T. R., Liu, C.-C., & Bu, G. (2019, 9). Apolipoprotein E and Alzheimer disease: pathobiology and targeting strategies. *Nature Reviews Neurology*, 15, 501–518. doi:10.1038/s41582-019-0228-7
- Zhou, Y., Graves, J. S., Simpson, S., Charlesworth, J. C., Mei, I. V., Waubant, E., . . . Taylor, B. V. (2017). Genetic variation in the gene LRP2 increases relapse risk in multiple sclerosis. *Journal of Neurology, Neurosurgery and Psychiatry*, 88(10). doi:10.1136/jnnp-2017-315971

8. Annexes

Tot el codi associat al projecte està publicat a:

https://github.com/JordiDilGiro/TFM_UOC2021