Universitat Oberta de Catalunya

uoc.edu

Associacions entre xarxes gèniques i malalties del cervell.

Josep Garcia Gutiérrez
Màster Universitari de Ciència de Dades
Medicina

Erola Pairó Castiñeira
Ferran Prados Carrasco

21 de gener de 2022

FITXA DEL TREBALL FINAL

Títol del treball:	*Associacions entre xarxes gèniques i malalties del cervell.*
Nom de l'autor:	*Josep Garcia Gutiérrez*
Nom del consultor/a:	*Erola Pairó Castiñeira*
Nom del PRA:	*Ferran Prados Carrasco*
Data de lliurament (mm/aaaa):	*01/2022*
Titulació o programa:	*Màster Universitari de Ciència de Dades*
Àrea del Treball Final:	*Medicina*
Idioma del treball:	*Català*
Paraules clau	*Malalties del cervell, expressió gènica, xarxa d'expressió gènica*

Resum del Treball:

Des de la seqüenciació del genoma humà s'han realitzat importants descobriments pel que fa a l'origen de malalties amb base genètica. GWAS (*Genome-Wide association study*) per exemple, ha identificat nombroses variacions genètiques per a diferents trets observables, tot i que el seu recorregut s'ha mostrat limitat quan les causes d'aquestes malalties tenen base poligènica i en cap cas ens informa de les relacions causals i els mecanismes biològics que hi ha al darrere.

La transcriptòmica pretén omplir aquest buit afegint informació sobre com s'expressen els gens i quines xarxes formen per regular el comportament molecular de les cèl·lules en els teixits. Identificar aquestes xarxes en presència de malalties del cervell és l'objectiu principal d'aquest treball. La metodologia ha fet ús de dos repositoris de dades, GWAS i GTEx, i un programari, SPrediXcan.

De la trentena de patologies cerebrals analitzades, han mostrat resultats significatius les següents: la malaltia d'Alzheimer, l'anorèxia nerviosa, la depressió major, l'esclerosi múltiple, l'esquizofrènia, l'insomni, la neurosi, el trastorn per dèficit d'atenció amb hiperactivitat (ADHD) i el trastorn bipolar. La mida i grau de connectivitat de les xarxes obtingudes són diversos i es faciliten diferents eines de visualització per presentar els resultats.

Abstract:

Important discoveries have been made since the sequencing of the human genome regarding the origin of genetic diseases. GWAS (Genome-Wide association study) for example, have identified numerous genetic variations for different observable traits, although its usefulness has been limited when the causes of these diseases are polygenic, and is not able to inform us about the causal relationships and the biological mechanisms behind them.

Transcriptomics aims to fill this gap by adding information about how genes are expressed and which networks regulate the molecular behaviour of cells in tissues. Identifying which networks are significant in brain disease is the main goal of this research. The methodology uses two databases, GWAS and GTEx, and one software, SPrediXcan.

Of the thirty brain pathologies analysed, the following have shown significant results: Alzheimer's disease, anorexia nervosa, major depression, multiple sclerosis, schizophrenia, insomnia, neurosis, Attention Deficit Hyperactivity Disorder (ADHD) and bipolar disorder. The size and degree of connectivity of the obtained networks are diverse and different visualisation tools are provided to present the results.

Índex

Llista de figures

1. Introducció

1.1 Context

El juny del 2000, **Bill Clinton**, aleshores president dels Estats Units presentava a la Casa Blanca els resultats del primer esborrany del **genoma humà**. L'èxit després de deu anys de treball col·lectiu de la comunitat científica convidava a l'optimisme com reflectien les paraules de l'expresident '*la Ciència Genòmica tindrà un impacte directe a les nostres vides; i encara més en la vida dels nostres fills. Revolucionarà la diagnosi, prevenció i tractament de la majoria de malalties humanes, si no totes*' [1]. El bioquímic **Francis Collins**, aleshores director de l'Agència Genòmica de l'Institut Nacional de Salut dels Estats Units, apuntava també que '*la diagnosi genètica de malalties s'assoliria en deu anys i que els tractaments començarien a aparèixer possiblement cinc anys després*' [2].

Dues dècades després, la realitat ha acabat imposant-se davant prediccions massa optimistes, tot i que seria un error obviar l'avenç en la matèria en els darrers anys. Projectes basats en el genotip com **GWAS** [3] (*Genome-Wide Association Study*), que neixen amb l'objectiu de cercar relacions entre la base genètica i els trets dels individus, han descobert milers de variants associades a diferents fenotips. A la llarga però, aquesta metodologia s'ha mostrat limitada a l'hora de descobrir l'**arquitectura gènica dels trets complexos** i comprendre els mecanismes biològics que hi ha el darrere d'aquestes associacions.

És per això que en els darrers anys apareixen nous camps d'investigació que se centren en aquestes associacions utilitzant com a base el **transcriptoma**. La transcriptòmica, és a dir, l'estudi de l'expressió gènica, ens ajuda a aprofundir més sobre les associacions amb trets complexos i les xarxes gèniques que hi actuen. A diferència del genoma, que és estàtic, el transcriptoma varia d'acord amb les condicions ambientals externes, i afegeix un component predictiu dinàmic més acurat; a priori.

En l'actualitat, les dades GWAS són ingents, mentre que les associacions basades en el transcriptoma, els anomenats **TWAS** (*Transcriptome-wide association studies*), són escasses i disperses, en gran part degut a la dificultat a l'hora de recollir dades. Projectes com **GTEx** (*Genotype-Tissue Expression*) [4] omplen aquest buit. Finançat per l'Institut Nacional de Salut dels Estats Units, recull vint mil mostres de teixits de nou-cents donants i proveeix a la comunitat científica d'una base de dades d'associacions genètiques amb trets moleculars com és l'expressió gènica.

Projectes més recents en transcriptòmica com el programari **PrediXcan** [5], aprofiten la gran quantitat de dades genòmiques existents en l'actualitat per fer prediccions d'expressió gènica i integrar la informació dels mecanismes biològics subjacents als trets complexos. És precisament amb l'ús d'aquesta

eina que el treball analitzarà de quina manera l'expressió gènica es relaciona amb **malalties del cervell** i quines són les xarxes gèniques que hi actuen.

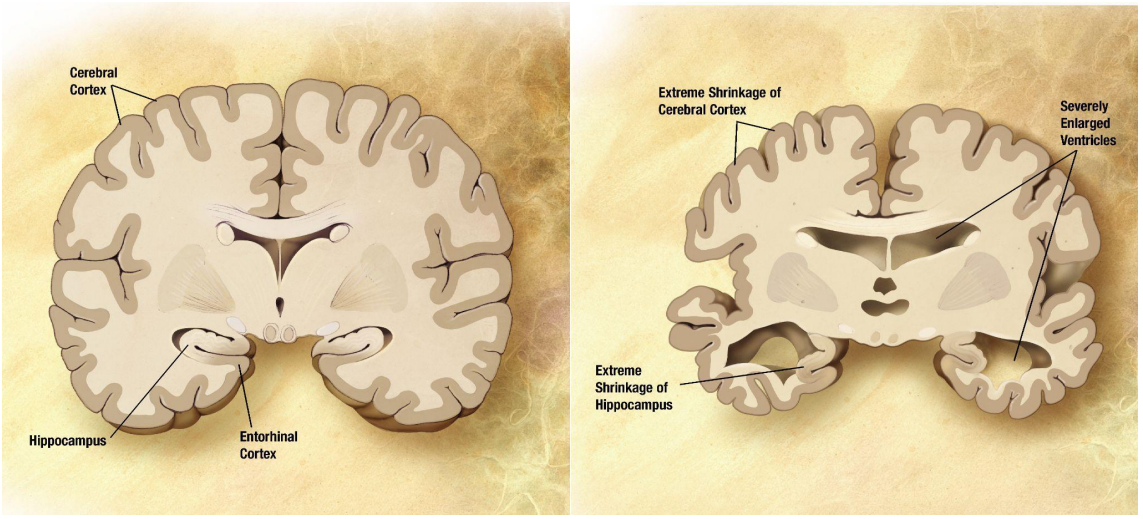

Figura 1: Comparació d'un cervell sa d'edat avançada (esquerra) amb el d'un pacient amb Alzheimer (dreta). Font Commons Wikimedia

1.2 Justificació i motivació

Malalties del cervell com les neurodegeneratives es caracteritzen per la pèrdua progressiva de l'estructura i funció de les neurones que pot desembocar en la mort cel·lular. La neurodegeneració es pot manifestar en el cervell en diferents circuits neuronals i encara no es coneix com revertir aquest procés de degeneració [6].

Patologies com la demència o l'Alzheimer apareixen sovint en el rànquing de causes de mortalitat de la població [7] com mostra l'exemple de la figura 2 l'any 2020, encapçalant la llista les morts causades per la pandèmia de la COVID-19.

Del gener al maig de l'any passat van morir a Espanya 9.284 persones de demència. Cada cas va posar punt final a una vida amb una molt que probable càrrega de patiment els darrers anys tant per al pacient com per a la família que l'envoltava. Qualsevol treball que tingui com a objectiu erradicar malalties com aquesta està més que justificat.

La motivació per l'elaboració d'aquest treball és triple. Per una banda, sempre m'han interessat les ciències naturals, la biologia i tot allò que determina el comportament humà. Tant les obres '*The Selfish Gene*', de **Richard Dawkins** [8], i '*Consciousness and the Brain*', d'**Stanislas Dehaene** [9], com les classes magistrals '*Human Behavioral Biology*' de **Robert Sapolsky**, a la Universitat d'Stanford [10], són de consum obligatori per a qualsevol persona atreta per aquesta branca científica.

	Enero a mayo		Enero y febrero		Marzo, abril y mayo	
	Valor	Variación	Valor	Variación	Valor	Variación
Total enfermedades	231.014	23,2%	78.784	-4,3%	152.230	44,8%
Covid-19 virus identificado	32.652	--	--	--	32.652	100,0%
Covid-19 sospechoso	13.032	--	--	--	13.032	100,0%
Enfermedades isquémicas del corazón	13.015	-3,6%	5.479	-9,9%	7.536	1,6%
Enfermedades cerebrovasculares	11.317	-0,3%	4.714	-2,4%	6.603	1,3%
Demencia	9.284	-4,8%	3.927	-9.5%	5.357	-1,0%

Figura 2: Defuncions segons la causa de mort més freqüent. Gener-maig del 2020. Font INE

La segona motivació són les dades. Els darrers anys, amb la voluntat de comprendre l'origen d'esdeveniments viscuts, en especial el període del crac financer 2007-08, vaig interessar-me per l'economia i particularment en l'anàlisi de dades estadística, estudis que m'han portat al màster actual de ciència de dades.
La tercera motivació és treballar en quelcom que tingui un impacte social positiu i posar el meu gra de sorra en combatre les malalties del cervell.

1.2 Objectius

L'objectiu principal del treball és l'obtenció de xarxes de gens l'expressió dels quals està relacionada amb malalties del cervell. Aquest és un enfocament potent per recollir informació biològica rellevant i per a la identificació de gens encara no associats a processos biològics explícits. Amb la detecció d'aquestes xarxes es pretén assolir objectius més a llarg termini, fora de l'abast del treball, relacionats amb l'estudi de nous gens involucrats en aquestes malalties i facilitar la recerca per a futures teràpies.

Objectius secundaris del treball consisteixen a experimentar i validar el programari PrediXcan, un seguit d'eines desenvolupades per un equip de la universitat de Chicago [5] per generar dades del transcriptoma a partir de dades GWAS.

1.3 Enfocament i metodologia

El treball consistirà a analitzar associacions basades en gens per realitzar el mapatge de trets complexos, com són les malalties del cervell, utilitzant transcriptoma imputat o de referència. Dades que s'extrauran de la base de dades del laboratori IM-Lab, l'equip que ha desenvolupat PrediXcan, i que serviran per detectar quins gens i quines xarxes genètiques són factor de risc en relació amb aquestes malalties.

Les xarxes es calcularan a partir del coeficient de correlació de Pearson de les expressions dels diferents gens significatius en cada teixit i per cada malaltia. Procés que mostra el diagrama de la figura 3 i que es detallarà en el tercer apartat de la memòria.

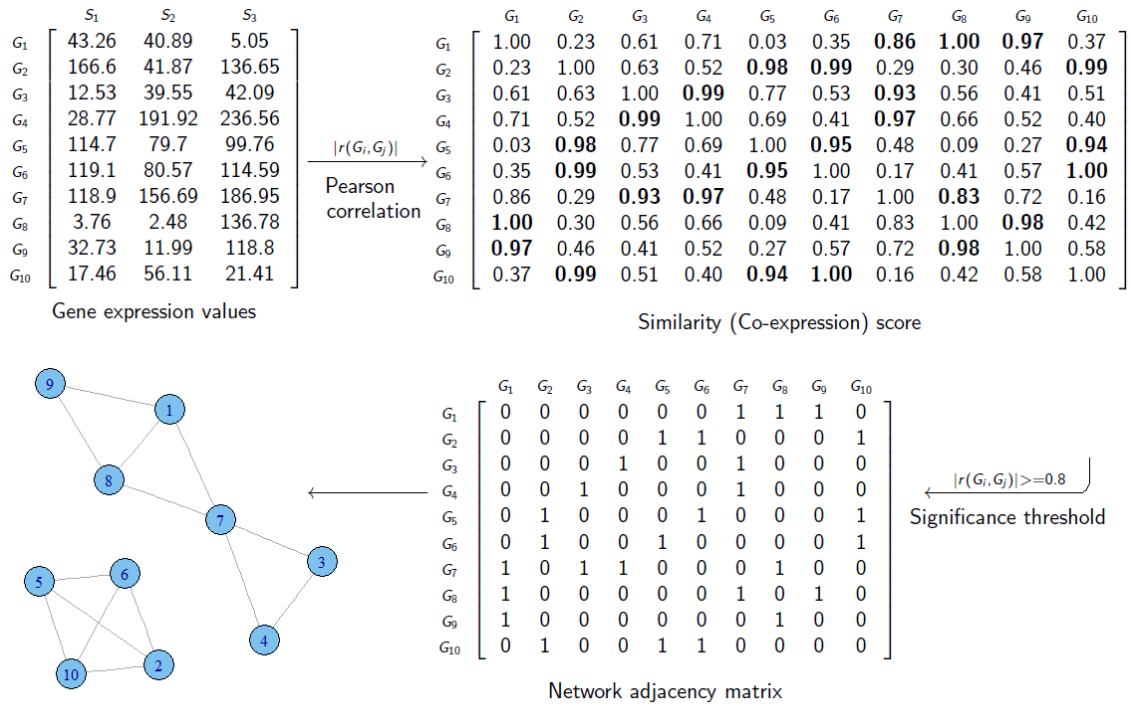

Figura 3: Metodologia per al càlcul de xarxes d'expressió gènica. Font: Wikipedia

1.4 Breu descripció del treball

Els apartats que segueixen comencen amb un breu repàs de l'estat de l'art actual fent un recorregut històric de les tècniques de seqüenciació d'ADN i ARN, i quines són les seves aplicacions en el camp biomèdic des dels seus inicis a principis de segle. Se citen els treballs més actuals que fan referència a malalties del cervell i que tenen com a punt de partida enfocaments òmics en l'àmbit del genoma, transcriptoma i xarxes gèniques.

El gruix del treball detalla el disseny de recerca i la seva implementació. Consisteix bàsicament en el recorregut de les dades, amb origen fitxers GWAS de malalties i trastorns del cervell, passant per l'obtenció dels diferents TWAS classificats per teixits cerebrals, fins a la identificació i esbós de les xarxes gèniques significatives associades a cada malaltia.

Els resultats mostren xarxes en teixits de nou malalties i trastorns del cervell: la malaltia d'Alzheimer, l'anorèxia nerviosa, la depressió major, l'esclerosi múltiple, l'esquizofrènia, l'insomni, la neurosi, el trastorn per dèficit d'atenció amb hiperactivitat (ADHD) i el trastorn bipolar. La mida, interconnexió i significació estadística d'aquestes xarxes difereixen segons la patologia. Acompanyen els

resultats diferents eines de visualització per a facilitar la identificació de la informació més rellevant.

2. Estat de l'art

Des del desxiframent del **genoma humà** i especialment en l'última dècada hi ha hagut importants avenços pel que fa a la relació entre la variació genètica i els trets complexos. En el camp biomèdic, és d'especial interès de quina manera el genotip està relacionat amb malalties i quins són els factors de risc que les desenvolupen.

2.1 GWAS

L'estudi d'associació del genoma complet (**GWAS**) [12] se centra a analitzar les variacions que existeixen al llarg de tot el genoma humà, els anomenats polimorfismes d'un sol nucleòtid (**SNPs**), i associar-les amb malalties. La metodologia consisteix a comparar el genoma d'un grup de persones que pateix la malaltia amb un grup de control.

Aquesta idea senzilla d'analitzar correlacions ja era clara a finals dels anys noranta i va ser a inicis del segle XXI quan la tecnologia ja permetia genotipar l'ADN. El problema era la dificultat d'obtenir mostres, ja que era un procediment extremadament car que tampoc aportava resultats massa convincents.
Els dubtes sobre la viabilitat de GWAS s'esvaeixen amb el treball de Robert J. Klein el 2005 [13] investigant la degeneració macular, on tot i ser una malaltia poligènica on intervenen molts gens, existeix una variant prevalent en la població europea amb un efecte molt potent.
És a partir d'aquell any que els èxits GWAS se succeeixen. En la figura 4 pot observar-se el progrés, tant pel que fa a les variants identificades com pel que fa a les mencions en la literatura acadèmica.

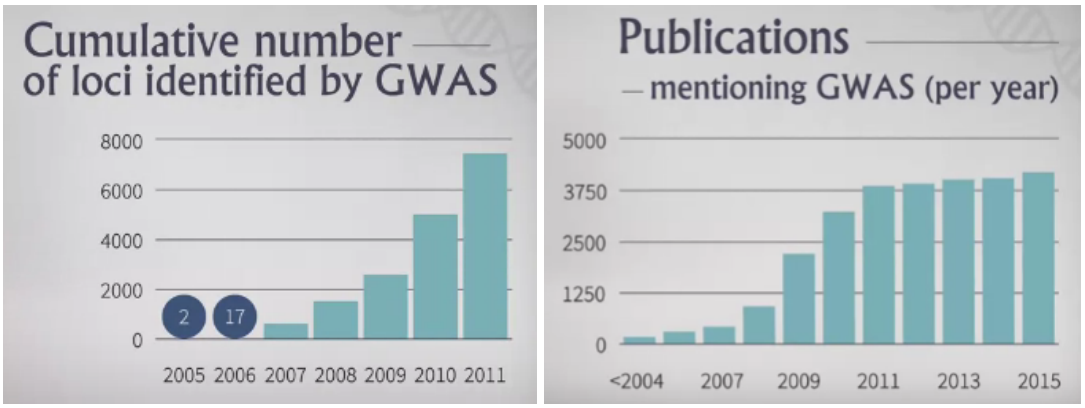

Figura 4: Evolució de variants identificades per GWAS i mencions acadèmiques.
Font: From Disease to Genes and Back | Coursera [14]

Però correlació no implica causalitat. GWAS relaciona variants amb trets complexos, però no dóna informació sobre les interaccions i les relacions

causals que intervenen entre les primeres i els segons. Pot passar que les variants detectades corresponguin a gens codificadors de proteïnes, i aquestes estiguin directament involucrades en la malaltia, però en la majoria dels casos, les variants identificades corresponen a elements reguladors d'expressió gènica que apunten a un o més gens, i que a més, poden estar ubicats molt lluny d'aquestes variants, fins i tot, en cromosomes diferents. GWAS també mostra les seves limitacions en trets poligènics on els efectes de cada gen són lleus, i l'anàlisi requereix un volum de mostres molt més gran.

El desenvolupament d'altres **ciències òmiques** complementa les anàlisis basades estrictament en el genotip com GWAS, per tenir una visió més global de les malalties i aprendre sobre quins és l'origen, les interaccions i el desenvolupament. Tècniques òmiques que van des de l'estudi dels nivells d'ARN i proteïnes en les cèl·lules, passant per la quantitat de microbiota, entre d'altres.

En aquest treball ens centrem en **xarxes gèniques** obtingudes mitjançant l'anàlisi del transcriptoma, i de quina manera, l'expressió gènica ha contribuït a identificar nous gens en el cas de trets complexos com són moltes de les **malalties cerebrals** presents en l'actualitat. Aquí, han estat clau projectes com el de **GTEx**, de l'Institut Nacional de Salut dels Estats Units, que consisteix en una base de dades de transcriptoma en teixits per a ús i estudi d'investigadors, i que té com a origen donants voluntaris. L'objectiu és identificar quines parts de l'ADN controlen el comportament dels gens (**eQTL**, *expression quantitative trait loci*), com les variants gèniques afecten l'expressió dels gens i establir bases per estudiar aquests mecanismes [15].

2.2 TWAS

Els primers intents per estudiar el transcriptoma complet arrenquen a principis dels anys noranta i, des d'aleshores, els avenços tecnològics n'han fet una disciplina general [16]. La tecnologia de **seqüenciació d'ARN** ha arribat avui en dia a tècniques modernes de nova generació per analitzar el transcriptoma (NGS *Next Generation Sequencing*). Tècniques com RNA-seq, milloren metodologies anteriors com Northern Blot, RT-PCR i Microarray, a l'hora de revelar la presència d'ARN en una cèl·lula en un moment determinat i analitzar canvis en la quantitat. També ofereix mesures més precises dels nivells de transcrits [17].

L'any 2013, l'equip de Valerio Costa (2013) [18] valorava èxits i perspectives amb l'ús de tecnologies de seqüenciació d'ARN de nova generació com **RNA-seq**. En l'article es detalla com aquesta tècnica és superior a les seves predecessores i és capaç d'identificar en un sol test potencials nous gens.

Projectes més actuals com **PrediXcan** prescindeixen de les mostres d'ARN i treballen directament amb transcriptoma imputat a partir de bases de dades de referència com Depression Genes and Networks [19], GEUVADIS [20] i el mateix GTEx. L'increment de la quantitat de base de dades actual facilita entrenar

models de referència d'expressió gènica. Aquests models s'apliquen després a les dades GWAS per identificar associacions entre SNPs, variacions en el transcriptoma i les malalties. PrediXcan gaudeix dels avantatges dels enfocaments basats en gens, a la vegada que es redueix la quantitat de mostres. Els resultats de l'equip d'Eric Gamazon (2015) [21] mostren com aquesta metodologia pot detectar gens nous, a part dels coneguts, associats amb les malalties i proporcionen nova informació sobre els mecanismes subjacents que hi actuen.

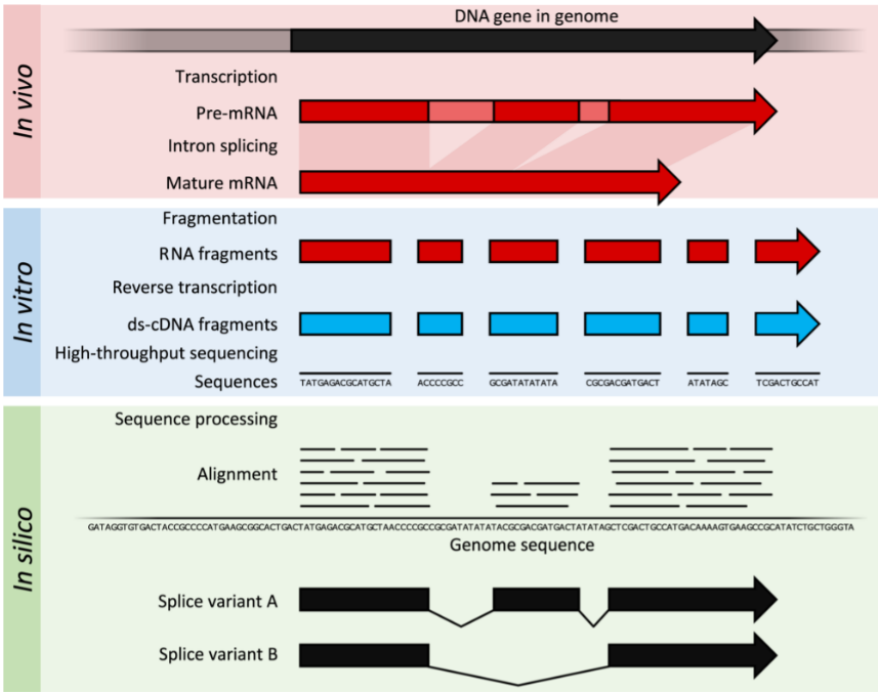

Figura 5. Resum de l'RNA-seq. Font: Transcriptomics technologies [16]

No obstant això, PrediXcan encara pateix d'algunes limitacions pel que fa a l'ús de transcriptoma imputat. Avaluacions com la de Binglan Li et al. (2018) [22] conclouen que es '*prediu amb precisió els nivells d'expressió gènica d'alguns gens, però no de tots*', i que '*incloure més imputacions eQTL a la predicció no millora la precisió*'. L'estudi apunta diverses maneres d'aprofitar el rendiment de PrediXcan per als eQTL i la recerca de malalties complexes.

Exemples d'èxit en l'estudi de malalties complexes a partir del transcriptoma en són uns quants:

- Gusev, Alexander, et al. (2016) [23] treballen amb expressió gènica imputada a partir d'una mostra relativament petita on es coneix la variació genètica i canvis en el transcriptoma, per analitzar volums de mostres molt més grans i identificar associacions d'expressió gènica i trets. Concretament, s'identifiquen 69 nous gens relacionats amb l'obesitat.
- Finucane, Hilary K., et al. (2018) [24] enfoquen la recerca sobre l'origen de malalties identificant teixits i cèl·lules rellevants mitjançant l'expressió gènica i els estudis GWAS. En les anàlisis específiques del cervell i les

malalties del sistema immunitari, aquest és un mètode especialment significatiu per casos particulars com el trastorn bipolar i l'esquizofrènia.
- Clarimon, Jordi, et al. (2020) [25] realitzen una revisió en la literatura científica sobre el paper de la transcriptòmica a l'hora de comprendre els mecanismes biològics que hi ha al darrere de malalties neurodegeneratives rellevants com l'Alzheimer i els gens rellevants que suposen un factor de risc de patir-les.
- Altres exemples d'èxit relacionats amb aquest tipus de malalties complexes són els de Vergouw, Leonie JM, et al. (2017) [26] amb els gens relacionats amb la demència amb cossos de Lewy, els treballs de Kamboh, M.I., et al. (2012) [27] i Yamazaki, Yu, et al. (2019) [28] amb l'Alzheimer, i el de Brynedal, B., et al. (2010) [29] amb l'esclerosi múltiple; per citar-ne uns quants exemples.

2.3 Xarxes gèniques

El concepte de xarxa gènica amb la qual treballarem, **xarxa de coexpressió gènica** (GCN *gene co-expression network*), apareix per primera vegada en el treball d'investigació d'Atul Butte i Isaac Kohane (1999) [30] sobre l'obtenció d'informació rellevant en mètodes no supervisats en xarxes. Aquesta tipologia de xarxes, a diferència de les xarxes de regulació gènica (GRN *gene regulatory network*), no ens informa sobre la seqüencialitat i funcionalitat en cada node de la xarxa, ens informa només que la xarxa correspon a un grup de gens que tenen funcions similars o estan involucrats en processos biològics comuns amb probables interaccions entre ells.

Destaquem en aquest camp els treballs de:

- de Jong, Simone, et al. (2012) [31] identifiquen una dotzena de xarxes associades amb l'esquizofrènia.
- Xiang, S., et al. (2018) [32] on es fa ús de les xarxes de coexpressió gènica per a la predicció de trastorns del cervell com l'autisme o la malaltia de l'Alzheimer.
- Gerring, Zachary, et al. (2019) [33] identifiquen nous gens i vies moleculars en trastorns de depressió major.

3. Disseny i implementació del treball

El disseny i metodologia del treball estan pensats per obtenir **xarxes d'expressió gènica** (GCNs, *Gene co-expression Networks*) per a malalties i trastorns del cervell.

El punt de partida seran arxius TWAS generats amb el software **MetaXcan** de l'equip **IM-Lab** de la Universitat de Chicago [34]. Aquest és un programari que a partir d'un conjunt de dades d'associació entre genotips i fenotips, i mitjançant models de predicció transcriptòmica, s'obtenen noves associacions entre expressió gènica i trets complexos.
PrediXcan combina dades de genotip i fenotip individualment mentre que **SPrediXcan** és una extensió que infereix els resultats de PrediXcan utilitzant fitxers GWAS *summary statistics*.

Tal com s'especifica en el repositori GitHub del laboratori, SPrediXcan treballa bàsicament amb tres entrades que representen els següents grups de dades:

- Un fitxer **GWAS *summary statistics***.
- Un model de predicció de transcriptoma o expressió gènica [35].
- Un fitxer amb les matrius de covariància de les variants dins de cada gen.

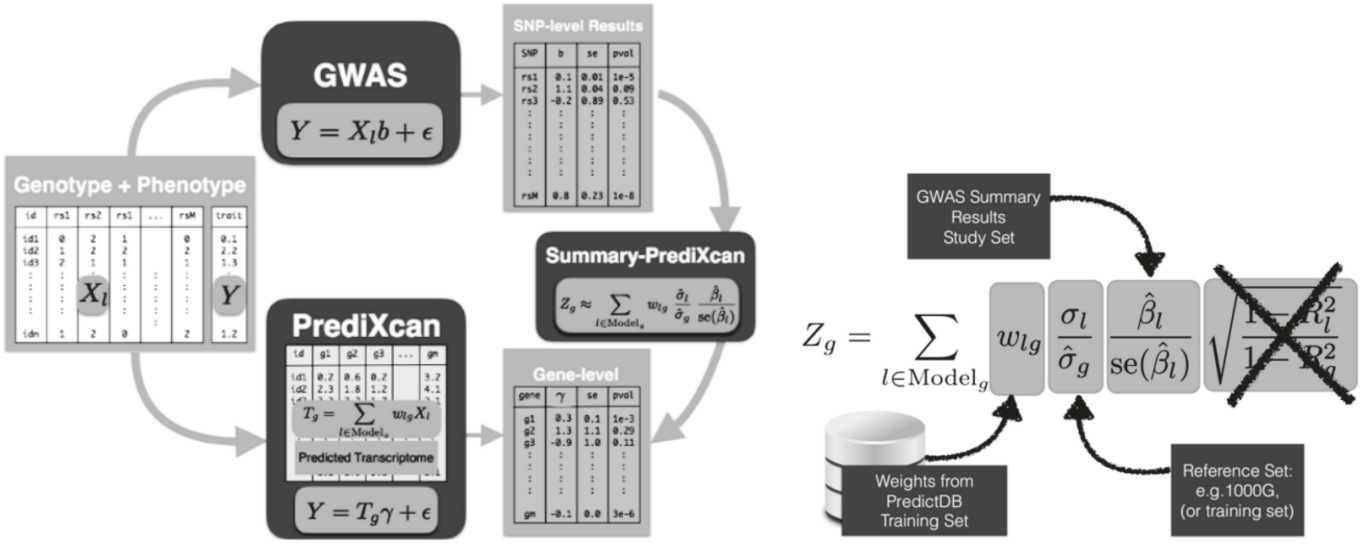

Figura 6: Mètode de càlcul per a SPrediXcan.
Font: Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics [36]

3.1 Captura de dades

"*The zoology of public GWAS data sets is more messed up than a taxonomy of the Cthulhu Mythos*" Alvaro Barbeira, Universitat de Chicago.

3.1.1 Arxiu GWAS

En l'estat de l'art vam repassar el nombre creixent de publicacions GWAS, però els estudis difereixen pel que fa al rigor i a la qualitat. El NHGRI (*National Human Genome Research Institute*), de l'Institut Nacional de Salut dels Estats Units i l'EBI (*European Bioinformatics Institute*) són els encarregats d'escollir aquells que superen un criteri de qualitat per incloure'ls al catàleg GWAS. Els considerats GWAS *summary statistics* difereixen de la resta en el fet que s'inclouen totes les associacions analitzades, i no només aquelles que han resultat significatives.

SPrediXcan requereix la instal·lació d'un entorn **Python** per executar-se i mitjançant arguments en la línia d'ordres integra tota la tipologia existent de fitxers GWAS. Les diferències apareixen bàsicament en dos aspectes: **el genoma de referència** i la nomenclatura de les variables.

El genoma de referència és una base de nucleòtids que representa una espècie sencera, en aquest cas, l'espècie humana. Per exemple, el genoma de referència actual, el GRCh38 (*Genome Reference Consortium*), està construït a partir de tretze donants diferents. Aquests genomes de referència serveixen de guia per a construir-ne de nous.

Release name	Date of release	Equivalent UCSC version
GRCh38	Dec 2013	hg38
GRCh37	Feb 2009	hg19
NCBI Build 36.1	Mar 2006	hg18
NCBI Build 35	May 2004	hg17
NCBI Build 34	Jul 2003	hg16

Figura 7: Genomes de referència. Font: Wikipedia

Els arxius del catàleg GWAS més recents estan mapejats a GRCh38 i en alguns casos, en cas de pertànyer a referències anteriors, es disposa d'una versió harmonitzada a la referència més recent.

Pel que fa a la nomenclatura de les variables aquestes poden estar etiquetades i agregades de manera molt diversa, tot i els esforços per a l'estandardització.
En el cas concret dels SNPs i el seu p-valor, per exemple, es demana una de les dues combinacions:

a) `variant_ID, p_value`
b) `chromosome, base_pair_location, p-value`

En l'annex A hi ha una ampliació del format recomanat per GWAS i la tipologia dels arxius estandarditzats i harmonitzats a l'últim genoma de referència.

3.1.2 Arxius model predictiu i matriu de covariància

Com ja es va avançar en l'apartat introductori, els models predictius d'expressió gènica faciliten la manca de dades d'associacions basades en el transcriptoma. El projecte GTEx de l'Institut Nacional dels Estats Units disposa de models predictius per a diferents teixits humans facilitant també arxius de matrius de covariància. Els teixits cerebrals amb els quals treballarem seran els següents (mantindrem la nomenclatura en anglès d'acord amb els arxius generats):

- Amygdala
- Anterior cingulate cortex
- Caudate basal ganglia
- Cerebellar hemisphere
- Cerebellum
- Cortex
- Frontal cortex
- Hippocampus
- Hypothalamus
- Nucleus accumbens basal ganglia
- Putamen basal ganglia
- Spinal cord cervical
- Substantia nigra

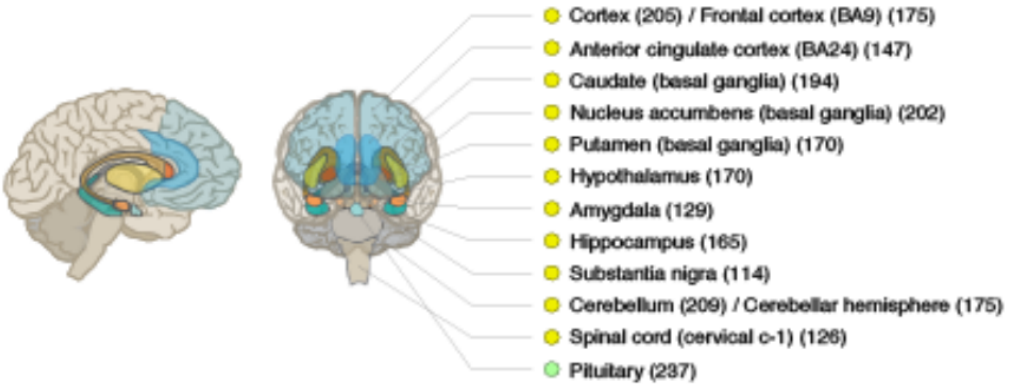

Figura 8: Teixits del cervell. Font: Sample and data types in the GTEx v8 study [37]

GTEx facilita models calculats amb dues metodologies estadístiques diferents, Elastic Net i MASHR. Aquesta última recomanada per l'equip IM-Lab i amb la que treballarem, ja que '*són parsimoniosos, biològicament informats i que utilitzen variants de mapatge fi*' [38].

Com passa amb els arxius GWAS, aquests gaudeixen de diferents tipologies depenent de com identifiquen cada variable i amb quina referència genòmica treballen. Per això, abans de processar res, és de vital importància que els tres arxius, GWAS + model predictiu + matriu de covariància, treballin amb dades homogènies.

3.1.3 Estructura dels arxius

Per les raons exposades i facilitar càlculs posteriors s'han desenvolupat eines amb Python per inspeccionar els arxius amb els quals es treballarà.
En el cas dels GWAS, un exemple de *summary statistics* per a la malaltia de l'Alzheimer presenta l'estructura següent:

	variant_id	p_value	chromosome	base_pair_location	effect_allele	other_allele
0	rs61769339	0.532266	1	662622	A	G
1	rs190214723	0.870407	1	693625	T	C

Figura 9: GWAS *summary statistics* de la malaltia de l'Alzheimer

Els models predictius MASHR de GTEx són bases de dades relacionals amb dues taules, `weights` i `extra`, i els següents camps:

```
weights
‘gene’, ‘rsid’, ‘varID’, ‘ref_allele’, ‘eff_allele’, ‘weight’
'ENSG00000169583.12', 'rs908836', 'chr9_136996001_G_A_b38', 'G', 'A',
-0.18060518974537

extra
‘gene’, ‘genename’, ‘gene_type’, ‘n.snps.in.model’, ‘pred.perf.R2’,
‘pred.perf.pval’, ‘pred.perf.qval’
'ENSG00000000457.13', 'SCYL3', 'protein_coding', 2, None, None, None
```

Pel que fa a la matriu de covariància, presenta el següent format:

	GENE	RSID1	RSID2	VALUE
0	ENSG00000000457.13	chr1_169894024_A_C_b38	chr1...	
1	ENSG00000000457.13	chr1_169894024_A_C_b38	chr1...	

Figura 10: Matriu de covariància model MASHR

3.2 Generació de TWAS amb SPrediXcan

El model MASHR de GTEx treballa amb arxius GWAS harmonitzats a la referència més actual, la GRCh38. Malauradament, l'harmonització del catàleg GWAS no s'entén bé amb SPrediXcan i si es vol treballar amb referències anteriors com la GRCh37, cal harmonitzar amb eines desenvolupades pel mateix equip.

En el cas de la malaltia de l'Alzheimer, per exemple, s'ha treballat amb un GWAS mapejat al genoma de referència GRCh37, el qual s'ha harmonitzat a la referència actual amb les següents línies d'ordres:

```
python summary-gwas-imputation/src/gwas_parsing.py \
-gwas_file GWAS/33589840-GCST90012877-EFO_0000249-Build37.f.tsv.gz \
-liftover liftover/hg19ToHg38.over.chain.gz \
-snp_reference_metadata reference_panel_1000G/variant_metadata.txt.gz
\
METADATA -output_column_map variant_id variant_id \
-output_column_map other_allele non_effect_allele \
-output_column_map effect_allele effect_allele \
-output_column_map beta effect_size \
-output_column_map p_value pvalue \
-output_column_map chromosome chromosome --chromosome_format \
-output_column_map base_pair_location position \
-output_column_map effect_allele_frequency frequency \
-output_order variant_id panel_variant_id chromosome position
effect_allele non_effect_allele frequency pvalue zscore effect_size
standard_error sample_size n_cases \
-output GWAS/alzheimer_harm.txt.gz
```

L'arxiu harmonitzat resultant `alzheimer_harm.txt.gz` és el que farem córrer pel SPrediXcan per a generar un TWAS per a cada teixit del cervell, tretze en total. En el cas del teixit de l'Amygdala:

```
python SPrediXcan.py --gwas_file GWAS/alzheimer_harm.txt.gz \
--snp_column panel_variant_id \
--effect_allele_column effect_allele \
--non_effect_allele_column non_effect_allele \
--beta_column effect_size \
--pvalue_column pvalue \
--model_db_path models/mashr/mashr_Brain_Amygdala.db \
--covariance models/mashr/mashr_Brain_Amygdala.txt.gz \
--keep_non_rsid --model_db_snp_key varID --throw \
--output_file results/alzheimer_amygdala.csv
```

Es mostren els camps més significatius dels fitxers SPrediXcan generats; consulteu l'annex B per al seu format complet:

	gene	gene_name	zscore	effect_size	pvalue
0	ENSG00000130203.9	APOE	29.611221	9.781687	1.071481e-192
1	ENSG00000267467.3	APOC4	13.742450	122357.623798	5.653013e-43
2	ENSG00000186567.12	CEACAM19	10.992327	0.279253	4.160585e-28
3	ENSG00000104853.15	CLPTM1	10.663039	7.920193	1.515621e-26

Figura 11: Fitxer TWAS generat amb SPrediXcan

S'han descarregat trenta arxius GWAS relacionats amb malalties i trastorns del cervell per a l'harmonització i generació de TWAS. Les fonts principals han estat el catàleg GWAS i el Psychiatric Genomic Consortium [39]. En total hem generat 390 arxius TWAS (30 malalties * 13 teixits cerebrals).

Per a cada teixit ens quedarem amb aquells gens amb p-valor igual a 0,05 corregit per **Bonferroni** [40]. Aquesta correcció es realitza quan volem mantenir la significació estadística triada per a múltiples tests. Si la volem del 95%, la correcció és tan senzilla com dividint el p-valor per la quantitat de gens predits:

p-valor corregit = 0,05 / nombre de gens

De totes les dades analitzades, han mostrat resultats significatius les següents patologies:

- Malaltia d'Alzheimer.
- ADHD (trastorn per dèficit d'atenció amb hiperactivitat).
- Anorèxia nerviosa.
- Trastorn bipolar.
- Depressió major.
- Esclerosi múltiple.
- Neurosi.
- Insomni.
- Esquizofrènia.

S'han descartat per manca de resultats les malalties i trastorns que segueixen: accident vascular cerebral, agorafòbia, aneurisma cerebral, autisme, convulsions febrils, delírium, ELA (esclerosi lateral amiotròfica), epilèpsia, hemorràgia cerebral, malaltia de Parkinson, neoplasia cerebral, OCD (trastorn obsessivocompulsiu), pànic, síndrome de Tourette, trastorn d'ansietat, trastorn mental transitori, trastorn neurològic, trastorn psicogènic i tumor cerebral.
En l'annex C hi ha la font dels estudis de totes les malalties analitzades, tant les que han mostrat resultats significatius com les descartades.

3.3 Xarxes d'expressió gènica

Per al càlcul de les xarxes d'expressió gènica ens dirigim a la base de dades GTEx i descarreguem mostres per a cada teixit del cervell.

SAMPID	SMATSSCR	SMCENTER	SMPTHNTS	SMRIN	SMTS	SMTSD	SMUBR
GTEX-1117F-0011-R10a-SM-AHZ7F	NA	B1, A1		NA	Brain	Brain - Frontal Cortex (BA9)	00098
GTEX-1117F-0011-R10b-SM-CYKQ8	NA	B1, A1		7.2	Brain	Brain - Frontal Cortex (BA9)	00098
GTEX-1117F-3226-SM-5N9CT	1	B1	2 pieces	6.2	Brain	Brain - Cortex	00018
GTEX-111FC-0011-R10a-SM-AHZ7K	NA	B1, A1		NA	Brain	Brain - Frontal Cortex (BA9)	00098
GTEX-111FC-0011-R10a-SM-CYKQ9	NA	B1, A1		8.5	Brain	Brain - Frontal Cortex (BA9)	00098
GTEX-111FC-3126-SM-5GZZ2	1	B1	2 pieces	6.1	Brain	Brain - Cortex	00018
GTEX-111FC-3326-SM-5GZYV	2	B1	2 pieces	7.1	Brain	Brain - Cerebellum	00020

Figura 12: Fitxer: GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt. Font GTEx Portal

Name	GTEX-1117F-3226-SM-5N9CT	GTEX-111FC-3126-SM-5GZZ2	GTEX-111FC-3326-SM-5GZYV	GTEX-1128S-2726-SM-5H12C	GTEX-1128S-2826-SM-5N9DI	GTEX-117XS-3026-SM-5N9CA	GTEX-117XS-3126-SM-5GIDP	GTEX-1192X-0011-R10a-SM-DO941	GTEX-1192X-0011-R5a-SM-DNZZA	GTEX-1192X-0011-R6a-SM-DNZZB	GTEX-1192X-0011-R7b-SM-DNZZC
ENSG00000223972.5	1.776e-02	0.000e+00	0.000e+00	1.709e-02	0.000e+00	0.000e+00	3.692e-02	3.947e-02	0.000e+00	0.000e+00	0.000e+00
ENSG00000227232.5	6.892e+00	4.225e+00	7.778e+00	2.359e+00	8.939e+00	3.693e+00	8.219e+00	1.542e+00	2.137e+00	1.918e+00	1.893e+00
ENSG00000278267.1	0.000e+00	4.912e-01	7.710e-01	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00
ENSG00000243485.5	0.000e+00	7.713e-02	0.000e+00	3.413e-02	0.000e+00	3.078e-02	0.000e+00	3.940e-02	5.749e-02	5.448e-02	1.037e-01
ENSG00000237613.2	0.000e+00	0.000e+00	0.000e+00	2.424e-02	0.000e+00	2.186e-02	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00

Figura 13: Fitxer GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz.
Font GTEx Portal

La captura i processament d'ambdós arxius s'ha realitzat amb **R** amb l'objectiu d'obtenir tretze llistats de gens i mostres, un per a cada teixit del cervell.

teixit	mostres	teixit	mostres
Amygdala	152	Hippocampus	197
Anterior cingulate cortex	176	Hypothalamus	202
Caudate basal ganglia	246	Nucleus accumbens basal ganglia	246
Cerebellar hemisphere	215	Putamen basal ganglia	205
Cerebellum	241	Spinal cord cervical	159
Cortex	254	Substantia nigra	139
Frontal cortex	209		

Figura 14: Taula de mostres per teixit GTEx

Creuem els fitxers TWAS generats per a cada malaltia i teixit amb les mostres obtingudes de GTEx. És a partir d'aquests gens expressats que obtenim les matrius de correlació de cada teixit amb l'objectiu de detectar quines són les xarxes gèniques expressades per a cada trastorn. Valors alts de correlació es tradueixen en mecanismes d'activació on l'expressió d'un gen augmenta amb l'augment de l'expressió de gens co-expressats.

L'estudi d'aquestes xarxes de co-expressió permeten identificar els gens que responen a la mateixa regulació del transcriptoma, els gens funcionalment relacionats o grups de gens involucrats en els mateixos processos biològics.

En el cas de la malaltia d'Alzheimer i pel teixit de l'amígdala, tenim els següents resultats:

- Gens significatius

	0	1	2	3	4	5	
ENSG00000143222.11	63.05000	48.52000	80.15000	66.72000	33.70000	37.43000	29.5400
ENSG00000143224.17	10.36000	11.13000	13.26000	9.89200	3.10300	8.02800	5.8510
ENSG00000158864.12	29.62000	31.18000	33.37000	34.02000	8.67800	18.71000	16.5500
ENSG00000203710.10	0.02234	0.08347	0.11510	0.16160	0.02012	0.12940	0.1438
ENSG00000196735.11	1.06000	5.08300	3.21600	10.28000	0.28290	0.57320	2.8640
ENSG00000179344.16	1.91100	4.99700	11.10000	9.08200	0.64800	0.59140	9.0440
ENSG00000237541.3	0.65160	3.89500	0.07299	5.60400	0.13690	0.05661	0.0513
ENSG00000232629.8	0.24480	1.37900	0.01097	3.01700	0.00000	0.12760	0.0386

Figura 15: Gens significatius per al teixit de l'amígdala per la malaltia de l'Alzheimer

- Matriu de correlació

	ENSG00000143222.11	ENSG00000143224.17	ENSG00000158864.12	ENSG00000203710.
ENSG00000143222.11	1.000000	0.737133	0.745969	0.0849
ENSG00000143224.17	0.737133	1.000000	0.852665	0.2356
ENSG00000158864.12	0.745969	0.852665	1.000000	0.4235
ENSG00000203710.10	0.084993	0.235664	0.423556	1.0000
ENSG00000196735.11	0.309236	0.382666	0.471744	0.3684
ENSG00000179344.16	0.331716	0.377517	0.463521	0.3007

Figura 16: Matriu de correlació per al teixit de l'amígdala per la malaltia de l'Alzheimer

- Heatmap

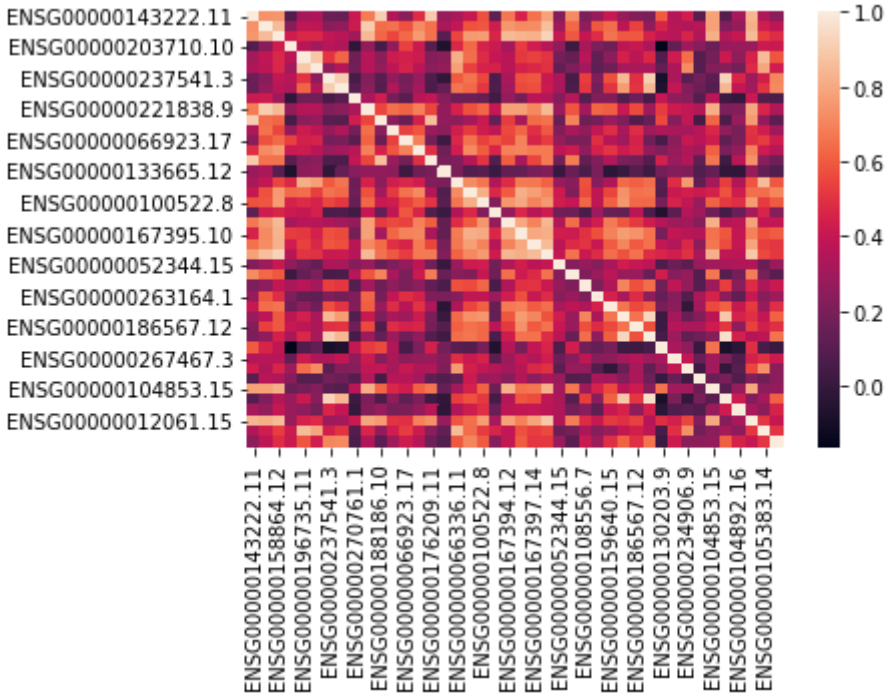

Figura 17: Heatmap per al teixit de l'amígdala per la malaltia de l'Alzheimer

En l'annex hi ha disponible l'enllaç amb la resta de resultats.

Les matrius resultants es converteixen en matrius binàries 0 i 1 quan el coeficient de correlació és inferior i superior a 0,5 respectivament (les correlacions negatives són residuals). És a partir d'aquestes matrius d'adjacència que es construeixen les xarxes de coexpressió gènica.

4 Resultats

S'han combinat tres eines de visualització diferents per mostrar els resultats del treball: gràfics de bombolles, diagrames de grafs i taules de freqüència. Per a la primera, s'ha construït un indicador que ens orienta sobre el grau de connexió de la xarxa obtinguda. El càlcul és una ràtio dels gens connectats respecte a una hipotètica connexió completa.

4.1 Gràfics de bombolles

Pel gràfic de bombolles s'ha realitzat una visualització interactiva amb Tableau [41]. En ella s'hi representen totes les xarxes gèniques dividides per malalties (color de la bombolla), per teixit (eix x), grau de connexió de la xarxa (eix y) i mida de la xarxa (mida de la bombolla).

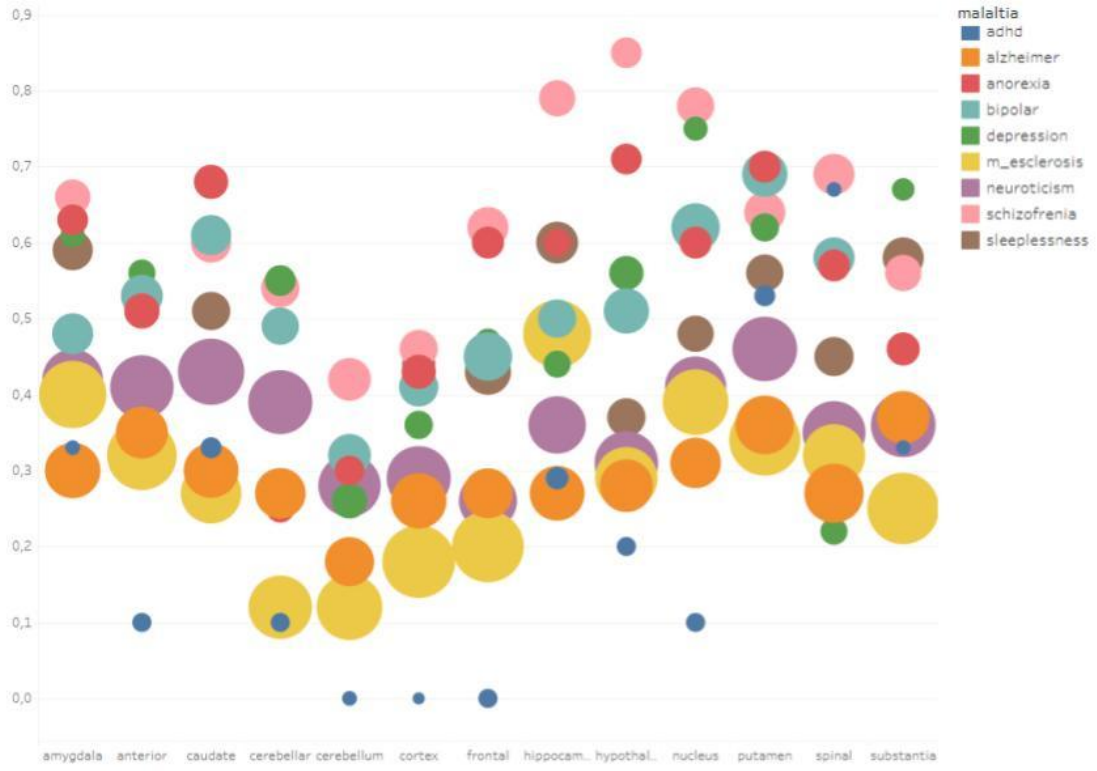

Figura 18: Gràfic de bombolles de les xarxes gèniques obtingudes.
Font: elaboració pròpia amb Tableau.

Al portal de Tableau, l'eina permet inspeccionar per cada bombolla els seus valors de grau de connexió de la xarxa i mida total de gens. També permet filtrar per malalties i teixits.

Tot seguit es mostra cada malaltia per separat afegint a la llegenda de les bombolles el nombre de gens, o mida de la xarxa:

ADHD

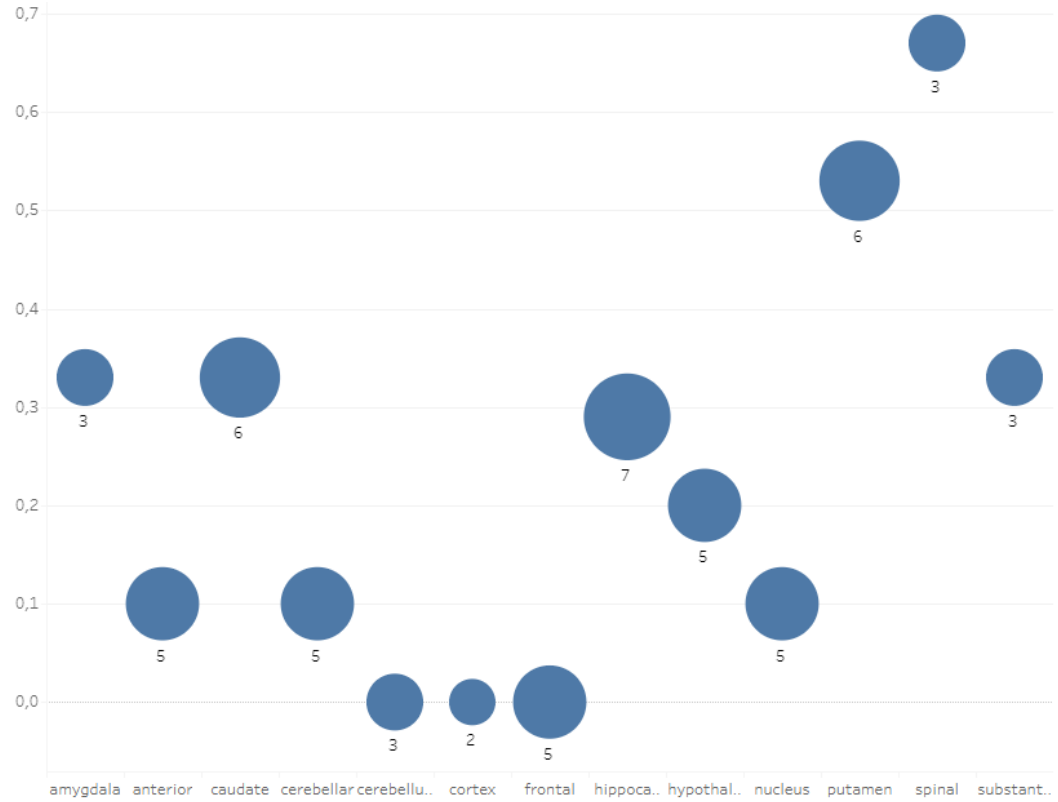

Figura 19: Gràfic de bombolles de les xarxes gèniques obtingudes per ADHD.
Font: elaboració pròpia amb Tableau.

Malaltia de l'Alzheimer

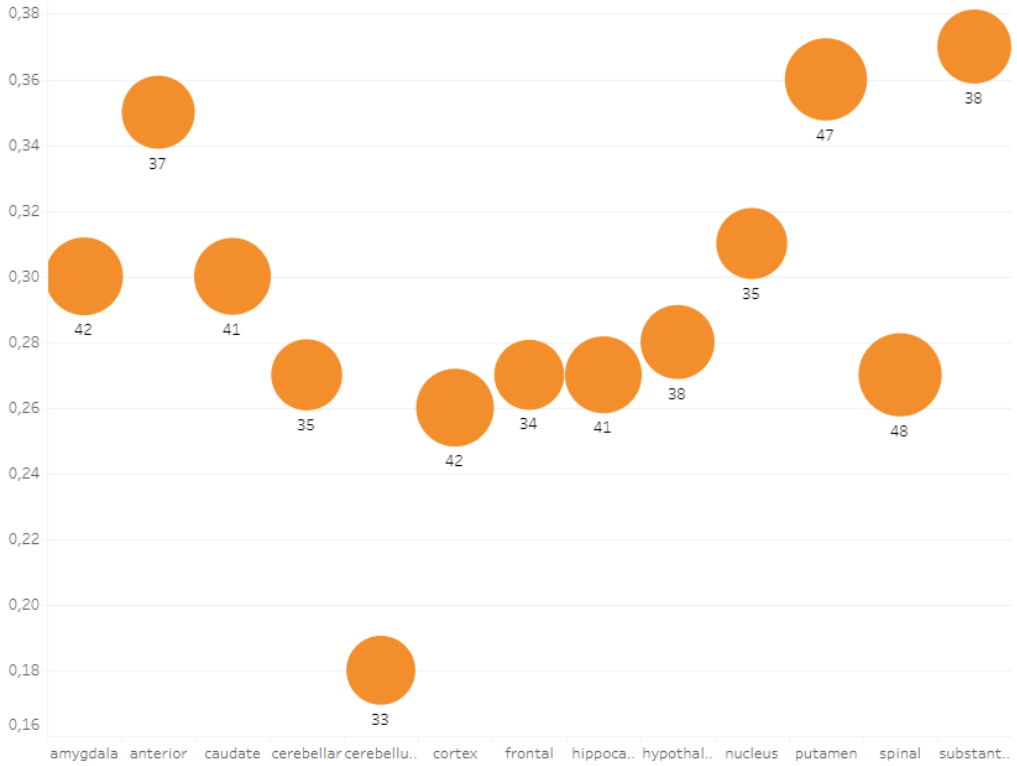

Figura 20: Gràfic de bombolles de les xarxes gèniques obtingudes per la malaltia de l'Alzheimer. Font: elaboració pròpia amb Tableau.

Anorèxia nerviosa

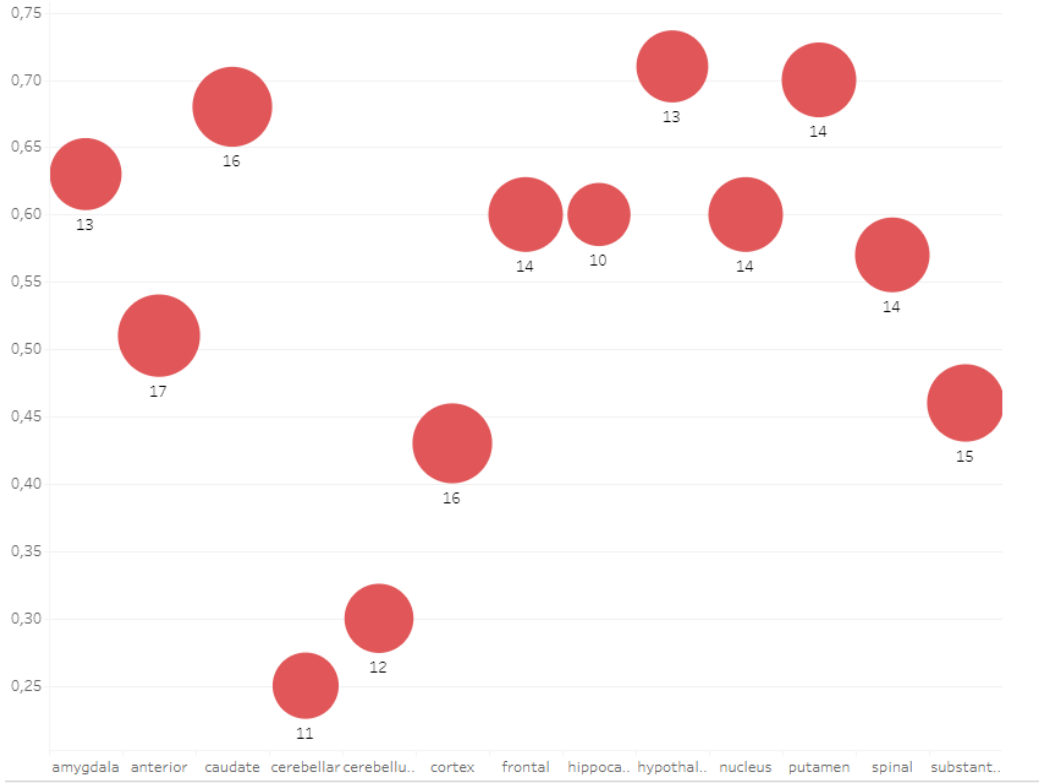

Figura 21: Gràfic de bombolles de les xarxes gèniques obtingudes per l'anorèxia nerviosa.
Font: elaboració pròpia amb Tableau.

Trastorn bipolar

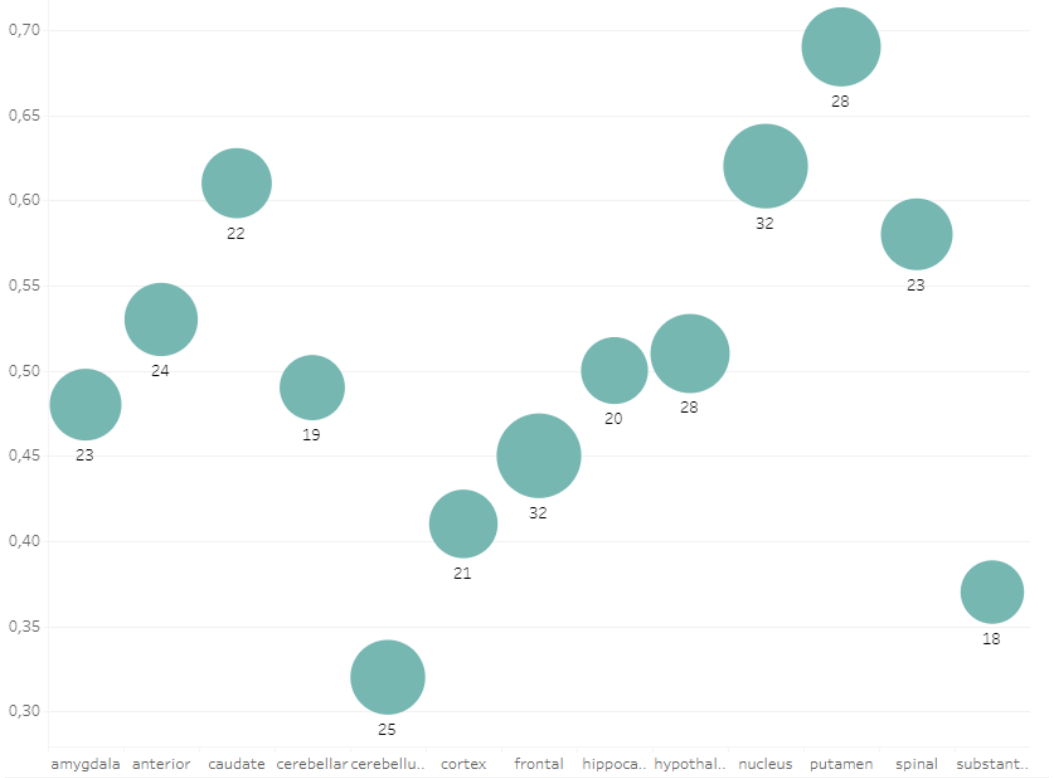

Figura 22: Gràfic de bombolles de les xarxes gèniques obtingudes pel trastorn bipolar.
Font: elaboració pròpia amb Tableau.

Depressió major

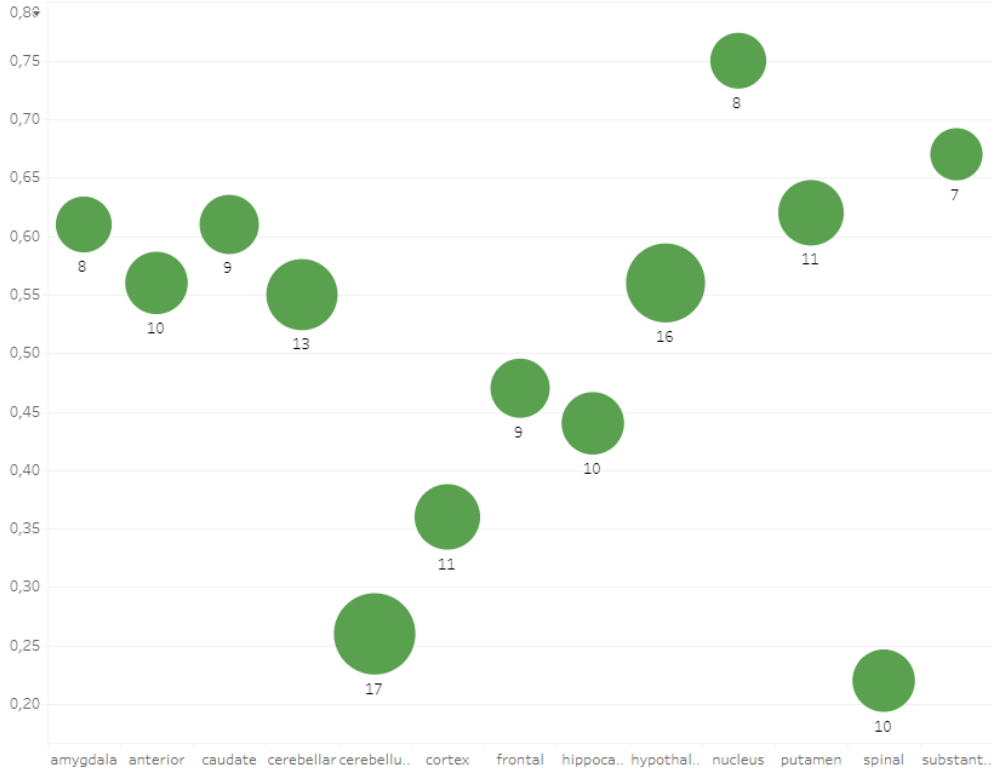

Figura 23: Gràfic de bombolles de les xarxes gèniques obtingudes per la depressió major.
Font: elaboració pròpia amb Tableau.

Esclerosi múltiple

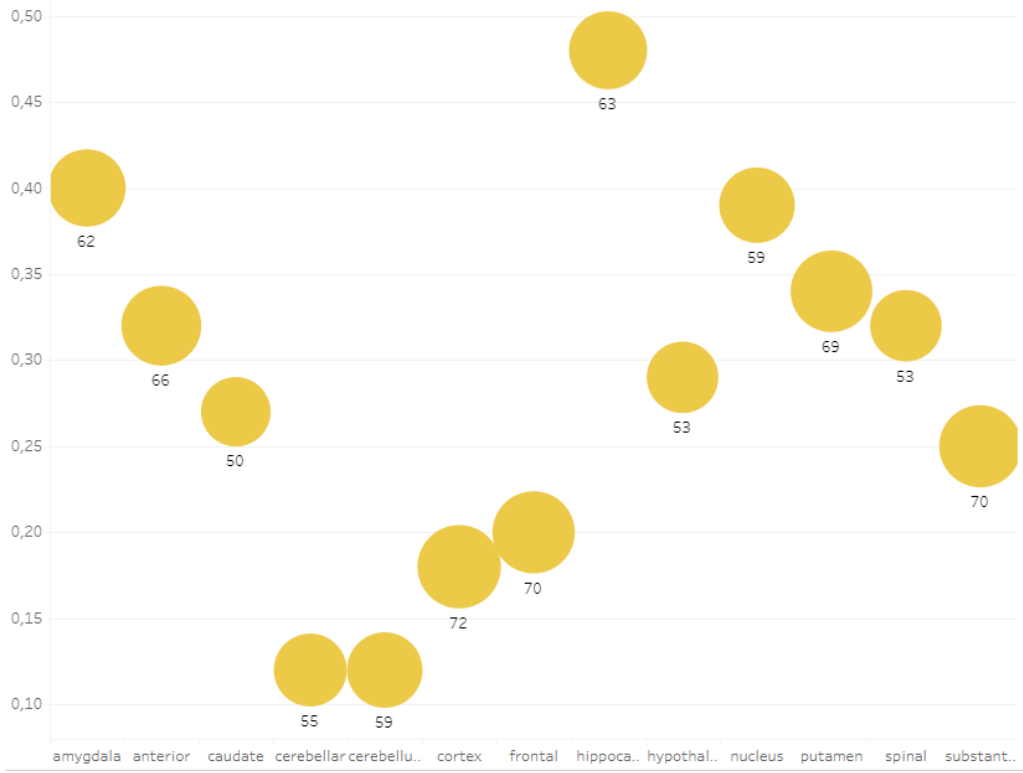

Figura 24: Gràfic de bombolles de les xarxes gèniques obtingudes per l'esclerosi múltiple.
Font: elaboració pròpia amb Tableau.

Trastorn neuròtic

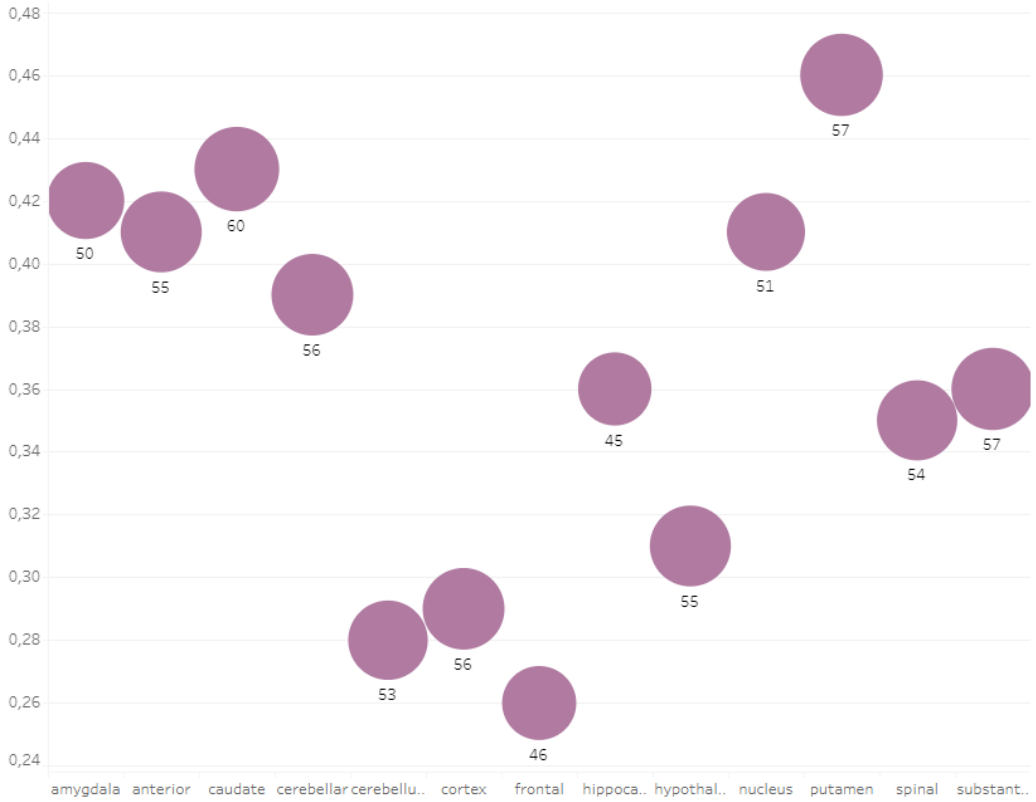

Figura 25: Gràfic de bombolles de les xarxes gèniques obtingudes pel trastorn neuròtic.
Font: elaboració pròpia amb Tableau.

Esquizofrènia

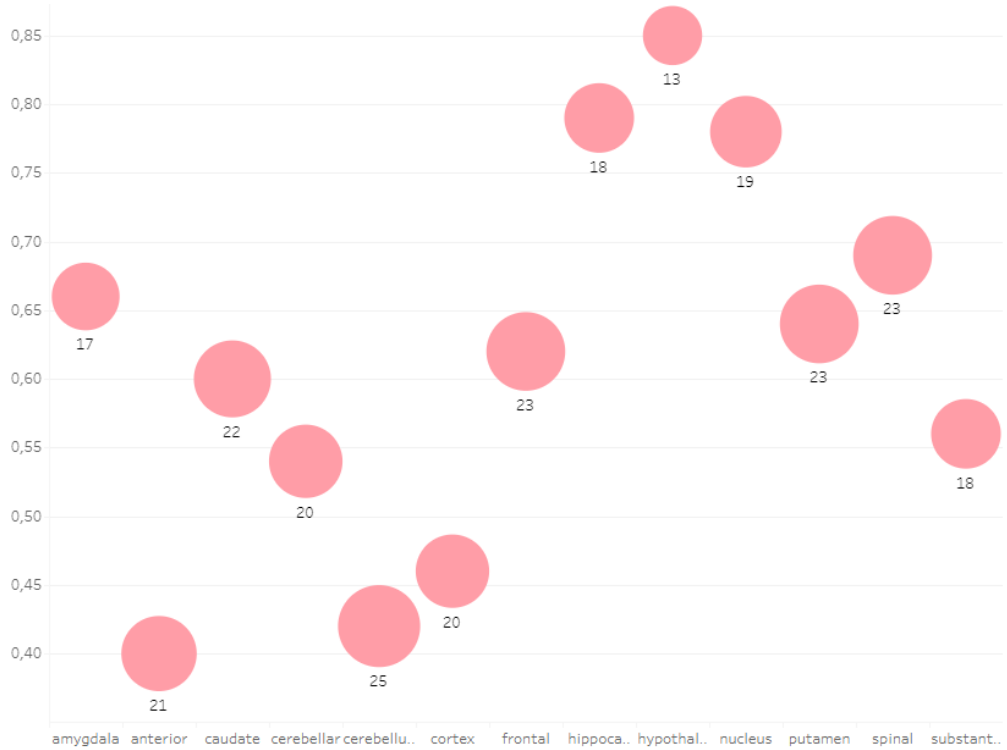

Figura 26: Gràfic de bombolles de les xarxes gèniques obtingudes per l'esquizofrènia.
Font: elaboració pròpia amb Tableau.

Insomni

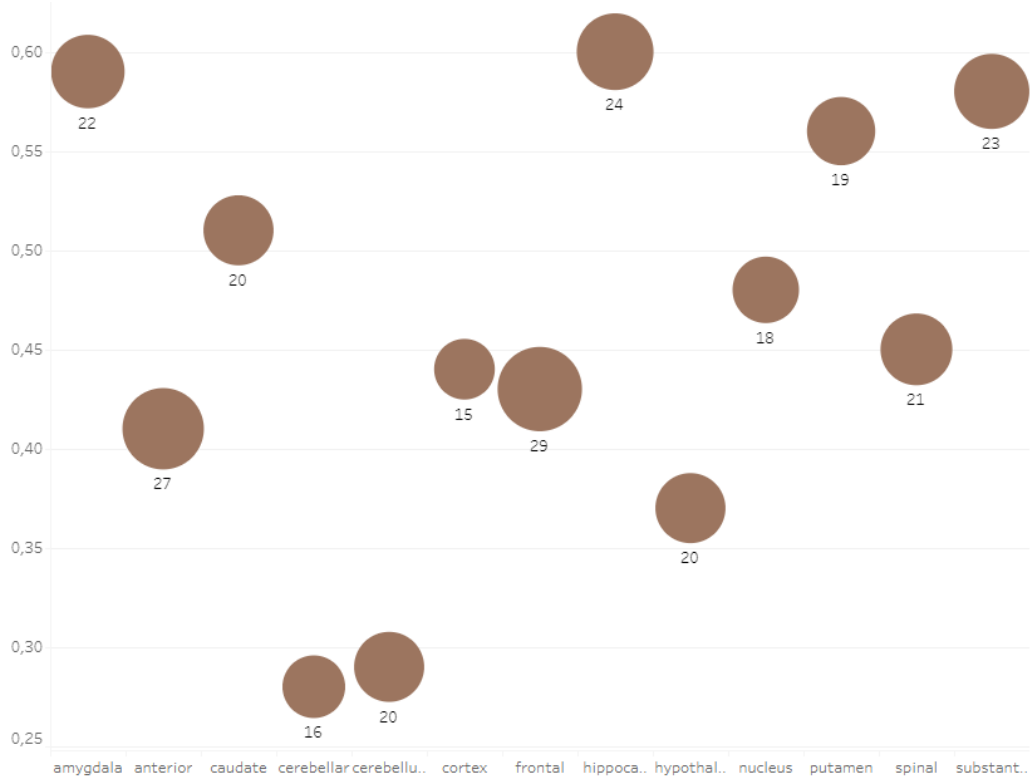

Figura 27: Gràfic de bombolles de les xarxes gèniques obtingudes per l'imsomni.
Font: elaboració pròpia amb Tableau.

4.2 Diagrames de grafs

En segon lloc, per la visualització en grafs, s'han representat algunes xarxes representatives amb Cytoscape [42], una eina de codi obert per visualitzar xarxes gèniques d'interacció molecular. S'han escollit teixits amb una major interacció gènica respecte a la resta.

Trastorn per dèficit d'atenció amb hiperactivitat
Teixit: hippocampus

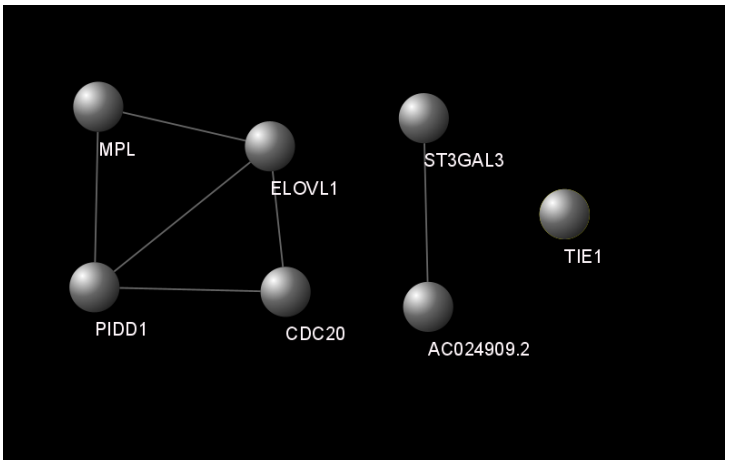

Figura 28: Xarxa gènica d'ADHD pel teixit hippocampus.
Font: elaboració pròpia amb Cytoscape.

Malaltia de l'Alzheimer
teixit: putamen

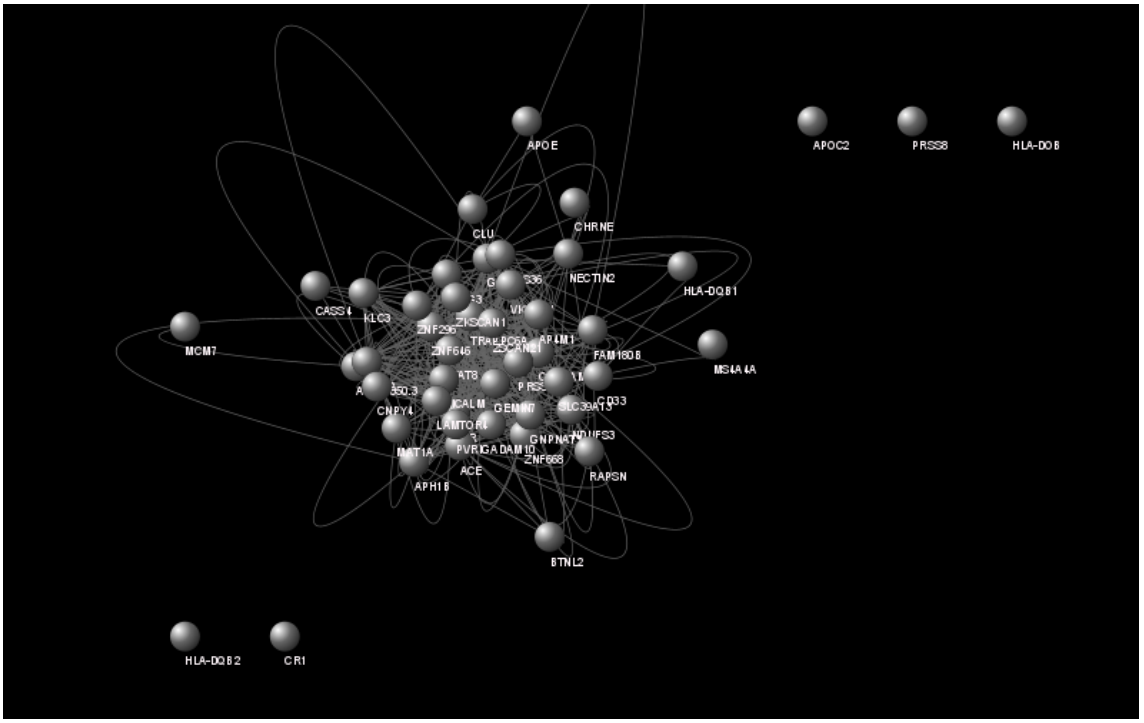

Figura 29: Xarxa gènica de la malaltia de l'Alzheimer pel teixit putamen.
Font: elaboració pròpia amb Cytoscape.

Anorèxia nerviosa
teixit: caudate

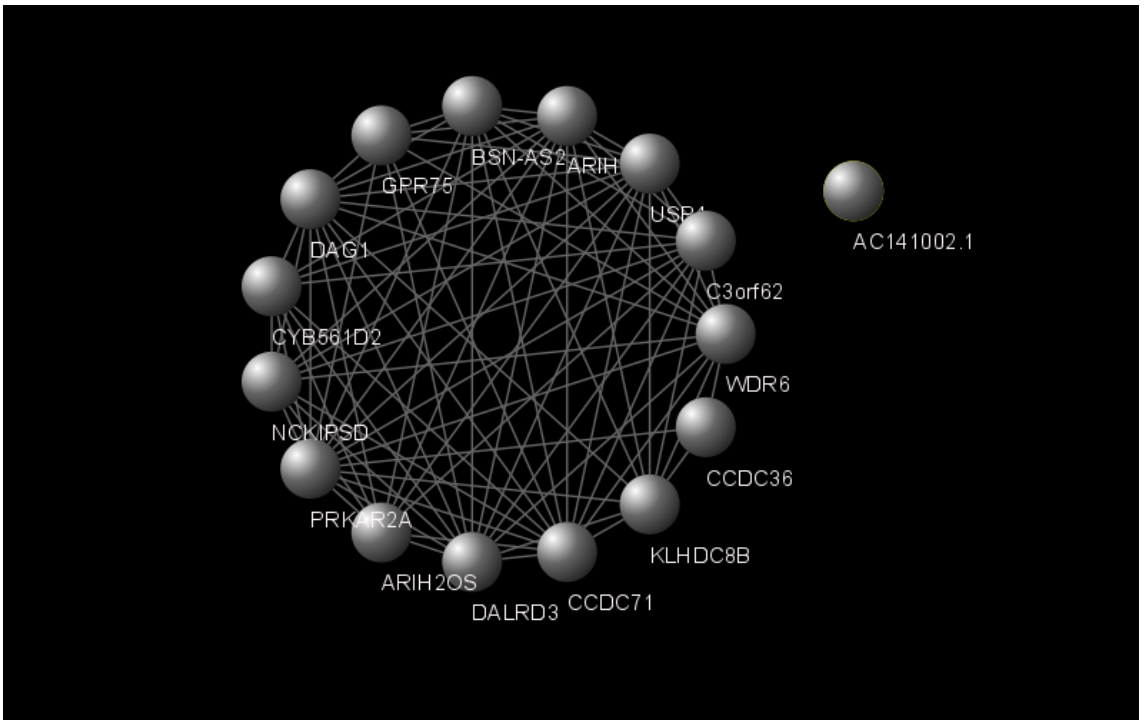

Figura 30: Xarxa gènica de l'anorèxia pel teixit caudate.
Font: elaboració pròpia amb Cytoscape.

Trastorn bipolar
teixit: frontal

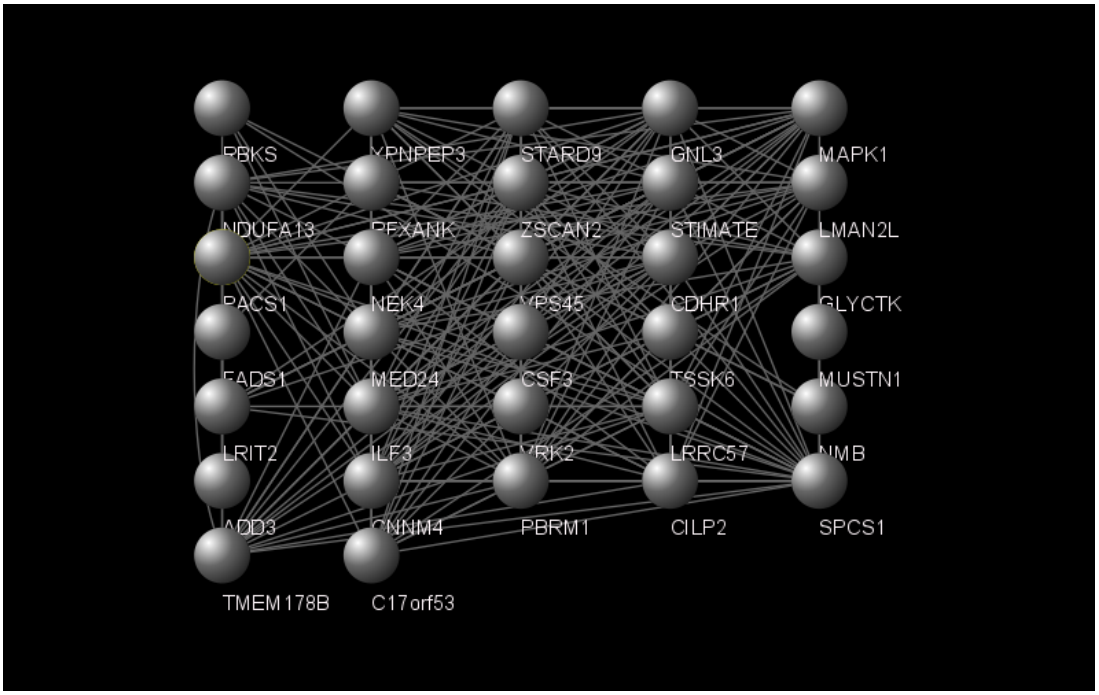

Figura 31: Xarxa gènica del trastorn bipolar pel teixit frontal.
Font: elaboració pròpia amb Cytoscape.

Esclerosi múltiple
teixit: hippocampus

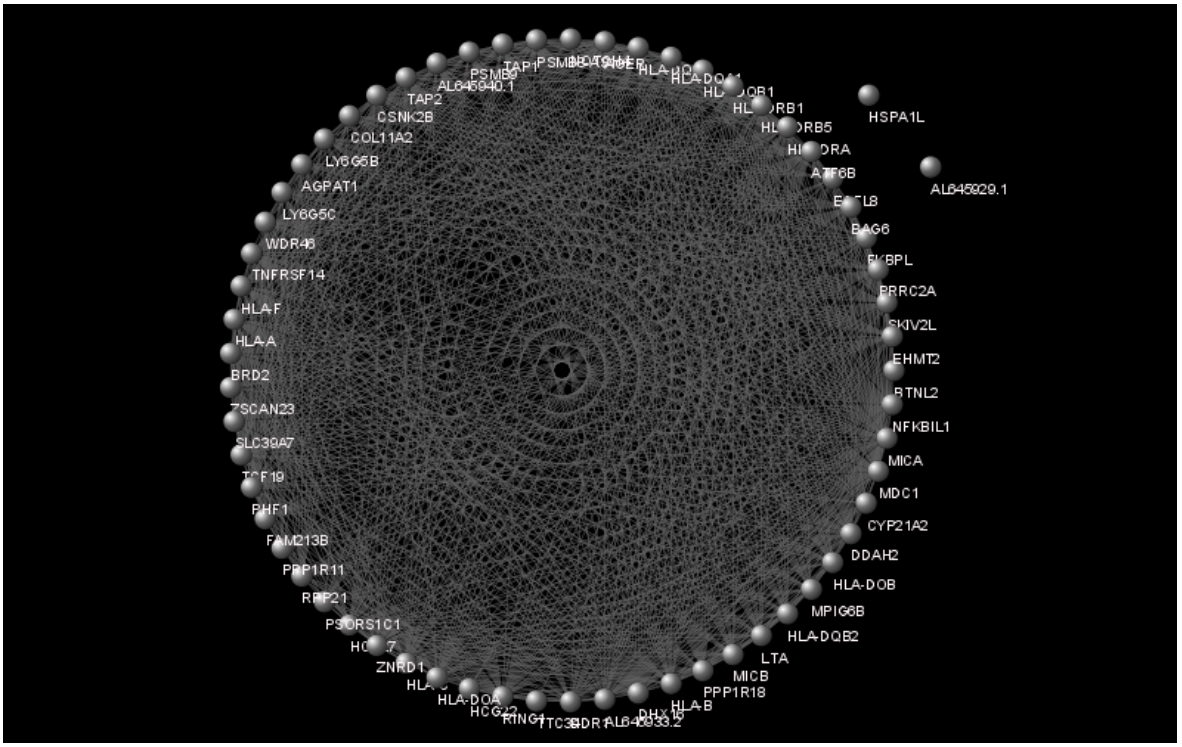

Figura 32: Xarxa gènica de l'esclerosi múltiple pel teixit hippocampus.
Font: elaboració pròpia amb Cytoscape.

Esquizofrènia
teixit: spinal

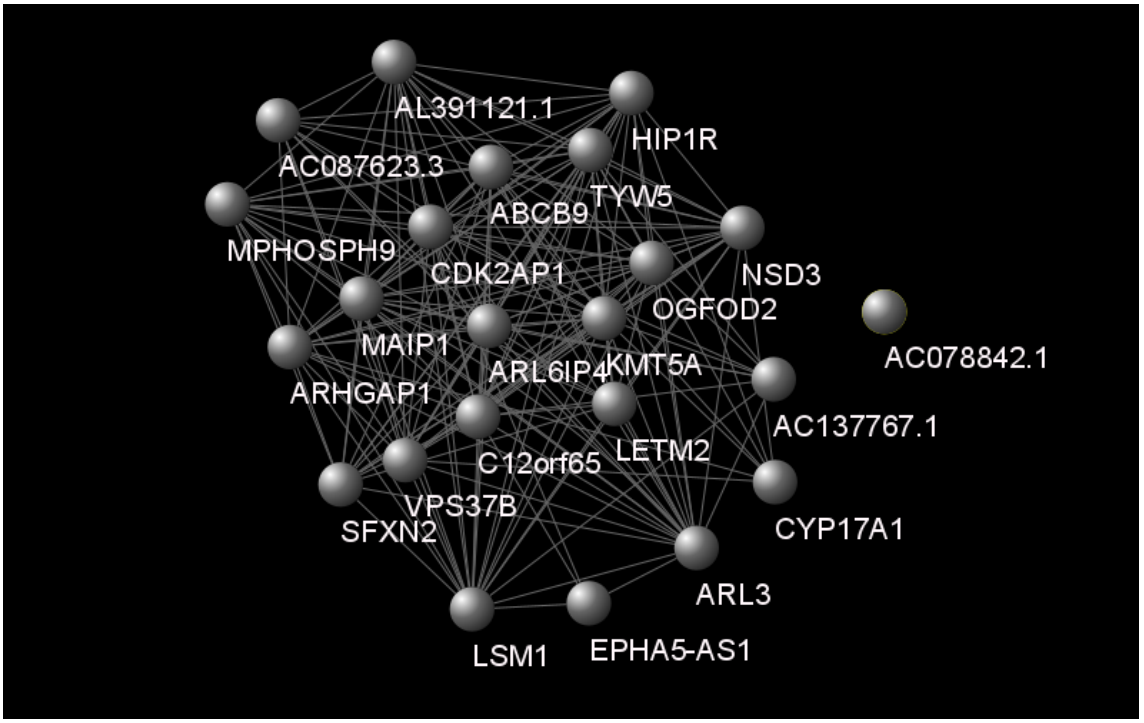

Figura 33: Xarxa gènica de l'esquizofrènia pel teixit spinal.
Font: elaboració pròpia amb Cytoscape.

4.3 Taules de freqüència

En tercer lloc, es faciliten unes taules amb el llistat dels gens més ben connectats a cada xarxa. Els gens venen amb la seva nomenclatura HUGO i el total de connexions.

Trastorn per dèficit d'atenció amb hiperactivitat

amygdala		ST3GAL3	1	AC024909.2	1
MED8	1	**hippocampus**		**putamen**	
AC024909.2	1	ELOVL1	3	MPL	4
anterior		PIDD1	3	PIDD1	4
TIE1	1	MPL	2	CDC20	3
PNPLA2	1	CDC20	2	ELOVL1	3
caudate		ST3GAL3	1	ST3GAL3	2
MED8	4	AC024909.2	1	**spinal**	
ST3GAL3	2	**hypothalamus**		MED8	2
PANO1	2	MED8	2	MPL	1
MPL	1	ST3GAL3	1	ST3GAL3	1
ELOVL1	1	AC010201.2	1	**substantia**	
cerebellar		**nucleus**		TIE1	1
MED8	1	DUSP6	1	MED8	1

Figura 34: Taula amb els gens ordenats per nombre de connexions per adhd.
Font: elaboració pròpia.

Malaltia de l'Alzheimer

amygdala		cerebellar		frontal		nucleus		substantia	
PICALM	26	BCKDK	19	SLC39A13	18	GRN	22	SLC39A13	25
ZNF646	26	GRN	19	VKORC1	18	ZNF646	21	ZNF668	24
VKORC1	26	SLC39A13	18	AP4M1	17	BCKDK	21	KAT8	24
GNPNAT1	25	PPOX	17	TSPAN14	17	SLC39A13	19	GNPNAT1	23
KAT8	24	PICALM	17	ZNF646	17	SPI1	18	ZKSCAN1	22
NDUFS2	22	ZNF668	17	KAT8	16	VKORC1	18	PICALM	22
AP4M1	21	CLPTM1	17	BIN1	14	KAT8	17	ZNF646	22
PPOX	20	PSMC6	16	ACE	14	ZNF668	16	ZNF285	22
ZNF668	20	ZNF646	16	PICALM	13	PPP1R37	16	LAMTOR4	21
ERCC1	20	PVR	16	APH1B	13	FCER1G	14	STYX	21
anterior		**cerebellum**		**hippocampus**		**putamen**			
VKORC1	25	CASTOR3	14	VKORC1	23	GRN	32		
ZNF646	23	BCKDK	14	SLC39A13	22	SLC39A13	31		
PVR	22	GNPNAT1	13	BCKDK	21	PRSS53	31		
ACP2	21	ZNF668	13	ZNF646	20	AP4M1	30		
AP4M1	20	PVR	13	HLA-DQA1	19	APH1B	30		
SLC39A13	20	NDUFS2	12	AP4M1	19	KAT8	30		
KAT8	20	PPP1R37	12	ADAM10	19	ADAM10	29		
CLPTM1	20	AGFG2	11	APH1B	19	ZNF646	29		
GNPNAT1	19	KAT8	11	KAT8	19	ZNF668	28		
ZNF668	19	BIN1	10	NDUFS2	18	VKORC1	28		
caudate		**cortex**		**hypothalamus**		**spinal**			
SLC39A13	24	SLC39A13	25	KAT8	24	BCKDK	32		
BCKDK	24	ZNF646	23	SLC39A13	21	CD2AP	25		
KAT8	23	AP4M1	22	ZNF646	21	SLC39A13	25		
GRN	23	MEPCE	22	AP4M1	19	ZNF646	25		
AP4M1	22	VKORC1	22	PSMC6	19	CLPTM1	25		
ZNF646	22	KAT8	22	ZNF668	19	PSMC3	24		
PRSS53	22	BCKDK	21	VKORC1	19	NDUFS2	23		
VKORC1	21	PPP1R37	20	BCKDK	17	VKORC1	23		
ZNF668	20	TOMM40	19	PPP1R37	17	ZNF668	21		
PVR	20	BIN1	18	GNPNAT1	16	LAMTOR4	20		

Figura 35: Taula amb els gens ordenats per nombre de connexions per la malaltia de l'Alzheimer. Font: elaboració pròpia.

Anorèxia nerviosa

amygdala		cerebellar		frontal		nucleus		substantia	
ASB3	11	NICN1	6	USP4	12	PRKAR2A	12	DALRD3	11

CCDC71	11	PRKAR2A	5	PRKAR2A	11	WDR6	11	USP4	11
PRKAR2A	10	NCKIPSD	4	WDR6	11	CCDC71	11	P4HTM	10
WDR6	10	C3orf62	4	C3orf62	11	NCKIPSD	10	NCKIPSD	9
KLHDC8B	10	WDR6	3	P4HTM	9	DALRD3	10	C3orf62	9
P4HTM	8	AC141002.1	2	DALRD3	9	C3orf62	10	NICN1	9
C3orf62	8	CCDC71	2	KLHDC8B	9	ARIH2OS	9	ARIH2	8
RHOA	8	BSN-AS2	1	CELSR3	8	P4HTM	8	WDR6	8
NCKIPSD	7	ZMYND10	1	NCKIPSD	8	RHOA	8	CYB561D2	8
CHAC2	5	AC008280.3	0	CCDC71	8	BSN-AS2	8	BSN-AS2	5
anterior		**cerebellum**		**hippocampus**		**putamen**			
CCDC71	13	NCKIPSD	6	QRICH1	8	SLC26A6	12		
CYB561D2	13	ARIH2	6	WDR6	7	ASB3	11		
PRKAR2A	12	CYB561D2	6	CCDC71	7	NCKIPSD	11		
WDR6	12	PRKAR2A	5	C3orf62	7	WDR6	11		
C3orf62	12	KLHDC8B	5	SLC26A6	6	CCDC71	11		
CELSR3	9	WDR6	4	NCKIPSD	6	C3orf62	11		
NCKIPSD	9	NDUFAF3	3	ARIH2OS	5	USP4	11		
P4HTM	9	CCDC71	3	KLHDC8B	5	APEH	11		
KLHDC8B	9	C3orf62	2	CAMKV	2	ARIH2OS	10		
RHOA	9	AC141002.1	0	CCDC36	1	CYB561D2	10		
caudate		**cortex**		**hypothalamus**		**spinal**			
PRKAR2A	14	PRKAR2A	11	ASB3	10	ASB3	10		
ARIH2	13	C3orf62	10	PRKAR2A	10	NCKIPSD	10		
DALRD3	13	ARIH2	9	ARIH2	10	PRKAR2A	10		
CCDC71	13	CCDC71	9	P4HTM	10	WDR6	10		
USP4	13	NCKIPSD	8	WDR6	10	CCDC71	10		
DAG1	13	DALRD3	8	CCDC71	10	C3orf62	10		
WDR6	12	NDUFAF3	8	C3orf62	10	APEH	10		
C3orf62	12	CYB561D2	8	USP4	10	ACYP2	9		
BSN-AS2	11	WDR6	7	BSN-AS2	10	ARIH2OS	9		
NCKIPSD	10	KLHDC8B	7	SLC26A6	9	CYB561D2	9		

Figura 36: Taula amb els gens ordenats per nombre de connexions per l'anorèxia.
Font: elaboració pròpia.

Trastorn bipolar

amygdala		**cerebellar**		**frontal**		**nucleus**		**substantia**	
ILF3	19	GNL3	14	MED24	23	ILF3	28	VPS45	13
VPS45	17	XPNPEP1	14	ILF3	23	DIP2A	28	WDR73	12
PACS1	17	XPNPEP3	14	VPS45	22	PBRM1	27	CNNM4	11
WDR73	14	VPS45	13	CNNM4	22	LRRC57	27	PBRM1	10
GATAD2A	14	CNNM4	12	GNL3	22	CNNM4	26	GLYCTK	9
XPNPEP3	14	PBRM1	12	PACS1	22	SPCS1	26	TENM4	8
LMAN2L	13	NEK4	12	ZSCAN2	22	XPNPEP1	26	TSSK6	8

NEK4	13	TMEM178B	12	SPCS1	21	VPS45	25	XPNPEP3	8
FADS1	13	SNAP23	12	STIMATE	21	NEK4	25	LMAN2L	7
TMEM56	12	ADD3	11	PBRM1	20	LMAN2L	24	TMEM178B	7
anterior		**cerebellum**		**hippocampus**		**putamen**			
ILF3	20	TMEM178B	15	VPS45	17	LRRC57	25		
VPS45	18	XPNPEP1	15	MRPS33	17	VPS45	24		
XPNPEP1	18	VPS45	14	LMAN2L	14	CNNM4	24		
ZSCAN2	18	NEK4	14	GNL3	13	GNL3	24		
TWF2	17	LRRC57	13	WDR73	12	GLT8D1	24		
PACS1	17	GNL3	12	NDUFA13	12	LMAN2L	23		
PBRM1	16	HDAC5	12	TMEM178B	10	TWF2	23		
GNL3	16	XPNPEP3	12	PACS1	10	PBRM1	23		
CDHR1	16	PDCD10	11	GLYCTK	9	NEK4	22		
NEK4	15	NDUFA13	11	NEK4	9	TMEM178B	22		
caudate		**cortex**		**hypothalamus**		**spinal**			
LRRC57	20	MED24	16	ILF3	22	VPS45	18		
TRPC4AP	19	XPNPEP1	15	VPS45	20	LMAN2L	18		
VPS45	18	VPS45	14	PBRM1	20	TWF2	17		
PBRM1	18	GNL3	13	GNL3	20	GNL3	17		
NEK4	18	LRRC57	13	SPCS1	20	NEK4	17		
CNNM4	16	CNNM4	11	NEK4	20	XPNPEP1	17		
PACS1	16	PACS1	11	MRPS33	19	TMEM258	17		
LMAN2L	15	NEK4	10	PACS1	19	PACS1	17		
GNL3	15	MRPS33	10	LRRC57	19	MAPK1	17		
TMEM178B	13	CDHR1	9	ZNF584	19	PBRM1	16		

Figura 37: Taula amb els gens ordenats per nombre de connexions pel trastorn bipolar.
Font: elaboració pròpia.

Depressió major

amygdala		**cerebellar**		**frontal**		**nucleus**		**substantia**	
ZKSCAN4	6	BAG5	10	NEGR1	6	FLOT1	7	HIST1H2BN	5
PTPN1	6	PTPN1	10	LRFN5	6	ZKSCAN4	6	PGBD1	5
HIST1H2BN	5	PGBD1	9	ZKSCAN4	5	ZSCAN26	6	PTPN1	5
BTN3A2	4	TRIM26	9	TRMT61A	5	CLP1	6	NEGR1	4
ZSCAN16	4	NEGR1	8	ZSCAN26	4	LRFN5	6	BTN3A2	4
FLOT1	4	ZKSCAN4	8	AL121821.1	3	NEGR1	5	ZSCAN16	4
IER3	4	FLOT1	8	CKB	3	CCDC152	3	LRFN5	1
NEGR1	1	ZSCAN9	7	BTN3A2	1	BTN3A2	3		
		ZSCAN16	6	MICB	1				
		HCG11	5						
anterior		**cerebellum**		**hippocampus**		**putamen**			
ZKSCAN4	7	PGBD1	9	FLOT1	7	ZKSCAN4	9		
FLOT1	7	FLOT1	9	HIST1H2BN	5	TUBB	9		

HIST1H2BN	6	KLC1	9	ZKSCAN4	5	HIST1H2BN	8		
TRIM26	6	TRIM26	7	ZSCAN12	5	ZSCAN26	8		
BTN3A2	5	C6orf48	6	ZSCAN26	4	ZNF165	7		
ZSCAN16	5	NEGR1	5	NEGR1	3	TRIM26	7		
NEGR1	4	HCG11	5	BTN3A2	3	CKB	7		
LRFN5	4	ZSCAN9	5	LRFN5	3	NEGR1	6		
AL021807.1	3	TUBB	5	AL021807.1	2	CCDC152	5		
CKB	3	CKB	5			BTN3A2	2		
caudate		**cortex**		**hypothalamus**		**spinal**			
NEGR1	6	BAG5	7	DDX39B	14	PGBD1	5		
ZKSCAN4	6	FLOT1	6	TRIM26	12	ZKSCAN4	4		
ZSCAN26	6	NEGR1	5	PTPN1	12	PTPN1	4		
PGBD1	6	LRFN5	5	ZKSCAN4	11	HIST1H2BN	3		
FLOT1	6	ZSCAN26	4	FLOT1	11	CKB	3		
CKB	6	CKB	4	BAG5	11	NEGR1	1		
BAG5	6	PTPN1	4	HIST1H2BN	9	BTN3A2	0		
BTN3A2	1	AL121821.1	3	ZNF165	8	HIST1H1B	0		
MICB	1	ZSCAN16	1	ZSCAN16	8	ZSCAN31	0		
		CLIC1	1	CTTNBP2	8	PPP1R18	0		

Figura 38: Taula amb els gens ordenats per nombre de connexions per la depressió.
Font: elaboració pròpia.

Esclerosi múltiple

amygdala		**cerebellar**		**frontal**		**nucleus**		**substantia**	
PBX2	52	BAG6	16	CLIC1	29	BRD2	48	DDX39B	35
MSH5	49	STK19	16	CD40	26	DDX39B	47	HLA-A	33
STK19	46	ATF6B	16	TRIM26	25	ABCF1	43	BAG6	33
DDAH2	41	PHF1	16	HLA-F	24	DHX16	43	MSH5	32
AGER	40	NEU1	15	PSMB8	24	MDC1	39	ATF6B	32
C6orf47	39	HLA-B	13	HLA-B	23	HLA-F	37	RPL5	31
HLA-DRB1	39	PPT2	13	ATF6B	23	C2	36	WDR46	31
EGFL8	38	TAP2	13	HLA-DRA	23	TNXB	35	CD40	30
HLA-DPB1	37	ATP6V1G2	12	HLA-DRB5	23	PHF1	35	BCL10	28
HLA-DRA	35	AIF1	12	HLA-DRB1	23	SKIV2L	34	HLA-F	27
anterior		**cerebellum**		**hippocampus**		**putamen**			
BRD2	47	VPS52	20	SLC39A7	56	HLA-F	48		
TNXB	39	ABCF1	19	DHX16	54	BRD2	44		
ZNRD1	35	DDX39B	19	BRD2	51	SKIV2L	43		
TAP2	35	BAG6	19	NFKBIL1	50	AGPAT1	43		
AGER	34	ATF6B	18	DDAH2	48	BAG6	42		
NOTCH4	34	PPT2	18	ZNRD1	46	CSNK2B	42		
PPT2	33	VARS2	17	MDC1	46	RING1	42		
HLA-DRB1	32	ATP6V1G2	17	RING1	45	TNXB	41		

PSMB8-AS1	32	NFKBIL1	17	MICA	44	C6orf47	40		
LTB	31	RNF5	17	AGER	44	WDR46	40		
caudate		**cortex**		**hypothalamus**		**spinal**			
DDX39B	28	DDX39B	29	PPT2	34	DDX39B	35		
PHF1	27	CSNK2B	27	TRIM26	31	HLA-A	33		
SKIV2L	26	ATF6B	26	BRD2	30	BAG6	33		
ATF6B	26	BRD2	26	ATF6B	28	MSH5	32		
STK19	25	PHF1	26	DDR1	26	ATF6B	32		
BRD2	23	PPT2	25	C4B	26	RPL5	31		
VARS2	22	PBX2	25	HLA-DRB1	25	WDR46	31		
PRRC2A	22	EHMT2	23	BAG6	24	CD40	30		
MPV17L2	22	WDR46	23	PHF1	24	BCL10	28		
AGPAT1	21	FAM69A	22	VARS2	23	HLA-F	27		

Figura 39: Taula amb els gens ordenats per nombre de connexions per l'esclerosi múltiple.
Font: elaboració pròpia.

Trastorn neuròtic

amygdala		**cerebellar**		**frontal**		**nucleus**		**substantia**	
FAM120A	39	TNKS	35	CSNK1G1	25	SBF2	36	CSNK1G1	38
SLC25A17	36	STK24	35	ATF6B	24	SLC39A13	36	EIF4G3	36
ATF6B	35	DLST	35	AGO2	24	FNBP4	36	ATF6B	36
MIEF1	35	ZC3H7B	35	FAM120AOS	24	ATF6B	35	SLC39A13	36
SLC39A13	34	CAMTA1	34	MED24	24	NEIL2	34	L3MBTL2	36
RCN1	33	AREL1	34	ARHGAP27	23	KBTBD4	34	HIVEP1	35
PLEKHM1	33	MAPT	34	NMT2	22	CSNK1G1	34	KBTBD4	35
STX4	32	MIEF1	34	SLC39A13	22	SETD1A	34	ZC3H7B	35
KBTBD4	31	RASSF1	33	PLEKHM1	22	PLEKHM1	34	MTMR9	34
SETD1A	31	MTMR9	33	KANSL1	22	ZC3H7B	34	FDFT1	34
anterior		**cerebellum**		**hippocampus**		**putamen**			
CSNK1G1	39	PAX6	30	ATF6B	29	FNBP4	43		
FAM120A	38	SLC39A13	29	MTCH2	29	SLC39A13	42		
ZDHHC5	38	YLPM1	29	NMT2	28	PLEKHM1	42		
DLST	38	DLST	29	NPRL2	27	NPRL2	41		
SETD1A	38	SLC12A5	29	MTMR9	27	STK24	41		
PLEKHM1	38	MIEF1	29	NEIL2	27	ZNF646	41		
NCOA6	37	ZC3H7B	29	FAM120A	27	MAPT	41		
NMT2	36	PCCB	28	FNBP4	27	NCOA6	41		
L3MBTL2	35	KBTBD4	28	PLEKHM1	27	VWA7	40		
MTMR9	34	PLCL2	27	RCN1	26	CSNK1G1	40		
caudate		**cortex**		**hypothalamus**		**spinal**			
FAM120A	43	FBXO38	35	MED24	34	DLST	35		
FNBP4	43	NCOA6	34	SLC39A13	32	ZDHHC5	34		
YLPM1	43	CSNK1G1	33	CSNK1G1	32	ZNF646	34		

PLEKHM1	43	ATF6B	32	HP1BP3	31	ATF6B	33		
ZDHHC5	42	SLC39A13	32	MSL2	31	DDB2	33		
SETD1A	41	ZNF646	32	FAM120A	31	SLC39A13	33		
ATF6B	40	YLPM1	31	PLEKHM1	31	FNBP4	33		
MTMR9	40	EXD2	30	PIGU	31	CSNK1G1	33		
ZC3H7B	40	SETD1A	30	ATF6B	30	ERI1	32		
SLC39A13	39	KBTBD4	29	SETD1A	30	NEIL2	32		

Figura 40: Taula amb els gens ordenats per nombre de connexions pel trastorn neuròtic.
Font: elaboració pròpia.

Insomni

amygdala		**cerebellar**		**frontal**		**nucleus**		**substantia**	
CCDC71	17	COX19	8	IPO9	20	IPO9	14	SLC39A13	20
CGGBP1	17	SLC39A13	8	PCDH1	20	UBE2W	13	NMT1	19
FAM120A	17	HEXIM1	8	CNNM2	20	RBM6	12	RBM6	18
ZNF143	17	RBM6	7	CGGBP1	19	HEXIM1	12	UBE2W	18
IPO9	16	UBE2W	7	UBE2W	19	LONRF1	11	HEXIM1	18
UBE2W	16	MIR9-3HG	7	KCTD10	19	CPEB2	10	CGGBP1	17
CNNM2	16	PSMC3	6	NMT1	19	NAB2	10	COX19	17
NMT1	16	ZNF420	5	ZNF420	19	SKOR1	10	NADK	16
ZNF420	16	MST1R	4	RBM6	18	SMAD5	9	SMAD5	16
ZNF585A	16	CCDC36	2	SLC39A13	18	PCDH1	9	CNNM2	16
anterior		**cerebellum**		**hippocampus**		**putamen**			
CNNM2	18	CNNM2	11	NMT1	20	FAM120AOS	15		
NMT1	18	MAP2K5	11	CNNM2	19	KLC2	15		
ZNF585A	18	NMT1	11	PSMC3	19	NADK	14		
IPO9	17	PSMC3	10	IPO9	18	RBM6	14		
UBE2W	17	KLHDC8B	9	UBE2W	18	CGGBP1	14		
SLC39A13	17	RBM5	9	SLC39A13	18	NMT1	14		
RBM5	16	STAU2	9	ZNF420	18	MIR9-3HG	13		
ZNF143	16	RBM6	8	RBM5	17	DCAKD	13		
PSMC3	16	SKOR1	8	LONRF1	17	CCDC71	12		
KLHDC8B	15	ZNF585A	7	ZNF143	17	UBE2W	12		
caudate		**cortex**		**hypothalamus**		**spinal**			
NMT1	16	ZNF420	11	UBE2W	13	CGGBP1	14		
SLC39A13	15	LONRF1	10	CGGBP1	12	UBE2W	14		
CNNM2	14	NMT1	10	NMT1	12	DCAKD	14		
HEXIM1	14	RBM6	8	RBM6	11	UBA7	13		
IPO9	13	CGGBP1	8	CNNM2	11	LONRF1	13		
PCDH1	13	PSMC3	8	ZNF143	11	ZNF143	13		
ZNF420	13	UBE2W	7	ZNF420	11	MAP2K5	13		
UBE2W	12	SLC39A13	7	TDRKH	10	PML	13		
ZNF143	12	CPEB2	5	MST1R	8	NMT1	13		

PSMC3	11	SMAD5	5	LIN28B-AS1	8	HEXIM1	13		

Figura 41: Taula amb els gens ordenats per nombre de connexions per l'insomni.
Font: elaboració pròpia.

Esquizofrènia

amygdala		cerebellar		frontal		nucleus		substantia	
NPRL2	15	ARL3	15	KMT5A	20	TYW5	17	NSD3	14
CNNM2	14	CKAP5	15	RAD54L2	19	IFRD2	17	C12orf65	14
ABCB9	14	OGFOD2	15	LETM2	19	CYB561D2	17	SBNO1	14
SBNO1	14	RASSF1	14	C12orf65	19	PCCB	17	KMT5A	14
KMT5A	14	ABCB9	14	NPRL2	18	KMT5A	17	TYW5	13
WBP1L	13	KMT5A	14	PCCB	18	NPRL2	16	NPRL2	13
MPHOSPH9	13	TYW5	13	NSD3	17	LSM1	16	LSM1	12
TYW5	12	CNNM2	13	MFSD13A	16	LETM2	16	RASSF1	11
CYB561D2	12	PITPNM2	13	CNNM2	16	CNNM2	16	ARHGAP1	11
ZNF408	12	C12orf65	13	MPHOSPH9	16	RASSF1	15	LSMEM2	10
anterior		**cerebellum**		**hippocampus**		**putamen**			
NPRL2	13	ABCB9	17	NSD3	17	IFRD2	19		
MSL2	13	OGFOD2	17	WBP1L	17	PCCB	19		
NSD3	13	ARL3	16	KMT5A	17	KMT5A	19		
CNNM2	13	DENR	16	NPRL2	16	NPRL2	18		
ABCB9	13	KMT5A	16	PCCB	16	CISD2	18		
TYW5	12	MAIP1	15	TYW5	15	LETM2	18		
SFXN2	12	RASSF1	15	CYB561D2	15	CNNM2	18		
KMT5A	12	PCCB	15	LSM1	15	VPS37B	18		
EPHA5	11	LETM2	15	C12orf65	15	SBNO1	18		
TMTC1	11	WBP1L	15	CNNM2	14	TYW5	17		
caudate		**cortex**		**hypothalamus**		**spinal**			
NPRL2	17	C12orf65	14	CYB561D2	12	TYW5	20		
NSD3	17	KMT5A	14	PCCB	12	LSM1	20		
CNNM2	17	MAIP1	13	KMT5A	12	ARL3	20		
TMTC1	17	NPRL2	13	RASSF1	11	KMT5A	20		
DENR	17	NSD3	13	LSM1	11	NSD3	19		
RASSF1	16	MFSD13A	13	LETM2	11	C12orf65	19		
RAD54L2	16	CNNM2	13	ARHGAP1	11	MAIP1	18		
PCCB	16	OGFOD2	13	TYW5	10	HIP1R	18		
ABCB9	16	MSL2	12	NPRL2	10	ABCB9	18		
OGFOD2	16	ABCB9	12	ARL6IP4	10	OGFOD2	18		

Figura 42: Taula amb els gens ordenats per nombre de connexions per l'esquizofrènia.
Font: elaboració pròpia.

4.4 Discussió

Les xarxes del trastorn ADHD són les més petites del grup de malalties significatives analitzades. En el cerebellum, cortex i frontal, els gens no estan connectats i el nombre de connexions en la resta de teixits no supera la desena. S'ha escollit el teixit hippocampus per representar amb Cytoscape (figura 28) perquè és la xarxa més nombrosa, 7 gens, tot i que el grau de connexió és baix, no arriba al 30%.

L'esclerosi múltiple i el trastorn neuròtic són a la part alta pel que fa a la mida de xarxes. Les de la primera malaltia contenen de seixanta a setanta gens en gairebé tots els teixits, tot i que el grau de connexió no és gaire alt, al voltant del 35% de mitjana sense superar mai el 50% (el valor més alt és del 48% en el teixit del hippocampus, figura 32). És de les malalties amb valors de grau de connexió més dispersos.
El trastorn neuròtic mostra xarxes de dimensió lleugerament inferiors, una cinquantena de gens, però més ben connectats entre ells.

La malaltia de l'Alzheimer té associada una distribució de xarxes bastant homogènia, tant en la mida com en el grau de connexió; una quarantena de gens i connectivitat que no arriba al 30%. Els valors màxims d'ambdós paràmetres els trobem en el teixit putamen, xarxa representada en la figura 29.

Les mides de les xarxes de la malaltia de l'esquizofrènia són modestes, una vintena de gens, però el grau de connexió entre ells és dels més elevats de les malalties estudiades, superant el 75% en teixits com el hippocampus, el hypothalamus (85%) i el nucleus. En el teixit spinal per exemple, representada en la figura 33, s'identifica una xarxa de 23 gens amb un 69% de connexió.

El trastorn bipolar mostra valors moderats de mida i connexió, amb mitjanes d'una trentena de gens i un 50% de connectivitat, amb valors que s'enfilen fins a un 69% de connexió.
Les tres patologies restants, l'insomni, la depressió major i l'anorèxia nerviosa, tenen associades xarxes amb poques desenes de gens i amb graus de connexió diversos. Destaquem el teixit caudate de l'anorèxia (16 gens i 68% de connexió, figura 30), i la xarxa del nucleus en la depressió, 8 gens connectats en un 75%.

Segueix una breu descripció dels gens més ben connectats en les xarxes dibuixades a Cytoscape:

Alzheimer (putamen):

- GRN (granulin precursor) és un gen codificador de proteïnes. Les malalties associades amb aquest gen inclouen degeneracions del lòbul frontal i temporal.

- SLC39A13 (Solute Carrier Family 39 Member 13), associat amb proteïnes transportadores de zinc.

Anorèxia nerviosa (caudate):

- PRKAR2A (Pretein Kinase CAMP-Dependent Type II Regulatory Subunit Alpha). cAMP és una molècula de senyalització important per a una varietat de funcions cel·lulars. Exerceix els seus efectes activant la proteïna-cinasa.

Trastorn bipolar (frontal):

- MED24 (Mediator Complex Subunit 24), gen que codifica un complex coactivador transcripcional necessari per a l'expressió de gairebé tots els gens.
- ILF3 (Interleukin Enhancer Binding Factor 3), responsable de codificar una proteïna d'unió d'ARN de doble cadena (dsRNA) amb altres proteïnes dsRNA, petits ARN no codificants i ARNm per regular l'expressió gènica i estabilitzar els ARNm.

Esclerosi múltiple (hippocampus):

- SLC39A7 (Solute Carrier Family 39 Member 7), la proteïna que codifica aquest gen és la responsable de transportar zinc des de l'aparell de Golgi i el reticle endoplasmàtic fins al citoplasma.
- DHX16 (DEAH-Box Helicase 16), proteïnes implicades en nombrosos processos cel·lulars que tenen a veure amb l'alteració de l'estructura secundària d'ARN.

Esquizofrènia (spinal):

- TYW5 (TRNA-YW Synthesizing Protein 5), les malalties associades a aquest gen inclouen apnea del son i l'ELA.
- LSM1 (Homolog MRNA Degradation Associated), gen codificador de la família de proteïnes RNA-*binding*.

5. Conclusions

S'han identificat xarxes d'expressió gènica en els diferents teixits del cervell en nou dels trenta trastorns del cervell analitzats inicialment: la malaltia d'Alzheimer, l'anorèxia nerviosa, la depressió major, l'esclerosi múltiple, l'esquizofrènia, l'insomni, la neurosi, el trastorn per dèficit d'atenció amb hiperactivitat (ADHD) i el trastorn bipolar. Les xarxes difereixen en mida i grau de connexió depenent de la malaltia.

El procediment recorre el procediment estàndard de captura, processament i visualització de dades com es mostra a la figura 43. El punt de partida, arxius GWAS *Summary Statistics*, recullen variants del genoma i la seva associació amb malalties. La dimensió dels registres varia en funció de l'estudi escollit oscil·lant entre centenars de milers de variants fins a desenes de milions. El motor SPrediXcan genera tretze arxius TWAS, un per a cada teixit del cervell, per cada malaltia d'entrada. Les 30 malalties originals generen 390 fitxers amb desenes de milers de registres d'expressió gènica i com es relacionen amb la malaltia. El volum de dades es redueix quan corregim la significació estadística, passant de 30 malalties a 9 i amb només unes quantes desenes de gens rellevants.
Les xarxes dels gens resultants es construeixen a partir de l'expressió gènica de les mostres de GTEx. La part final del treball s'ha focalitzat en mostrar tota la informació obtinguda de forma intel·ligible, incloent-hi tres visualitzacions diferents.

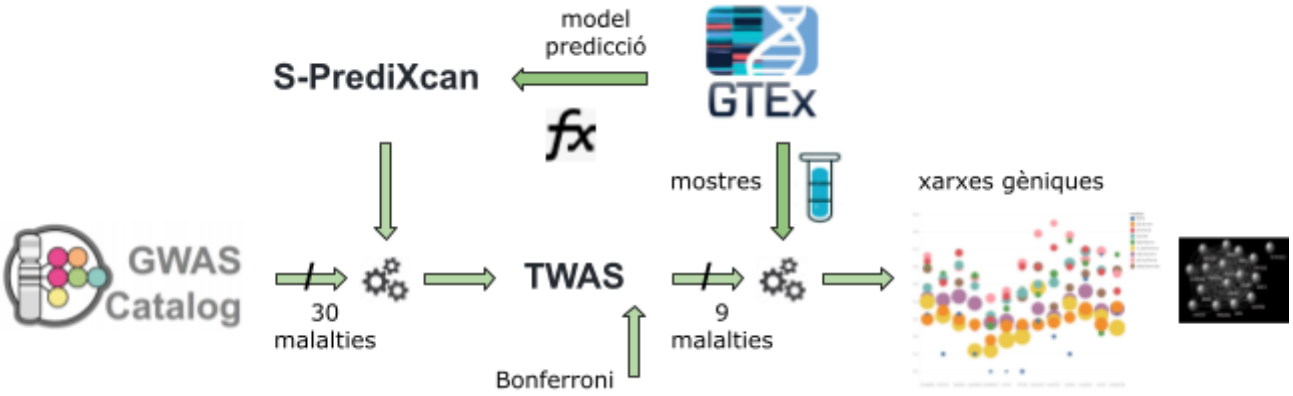

Figura 43: Metodologia per a l'obtenció de xarxes gèniques de malalties del cervell.
Font: elaboració pròpia.

Algunes de les limitacions de la metodologia escollida ja s'han insinuat en capítols anteriors quan parlàvem del programari emprat, SPrediXcan, concretament pel que fa a la precisió en els nivells d'expressió d'alguns gens.

Una altra de les limitacions és en el càlcul de la xarxa d'expressió gènica. El coeficient de correlació escollit ha estat el de Pearson i, juntament amb altres metodologies com el de la Informació Mútua (MI), el coeficient de correlació de

Spearman i la Distància Euclidiana, són dels més utilitzats, però totes tenen avantatges i desavantatges. En el cas que ens ocupa, el càlcul només és capaç de detectar relacions lineals i és sensible als *outliers*. També treballa assumint que la distribució de les expressions dels gens és normal.

Un altre criteri arbitrari ha estat considerar gens connectats amb valors de coeficient de correlació R^2 superior o igual a 0,5. Pot repetir-se l'exercici amb llindars més exigents, o usar altres mètodes que tenen en compte el nombre de mostres o la distribució de correlacions.

En darrer lloc, la tipologia de xarxa calculada mostra connexions entre grups de gens, però no ens diu res sobre quina és la causalitat o quin és l'ordre d'esdeveniments en el procés bioquímic de la cèl·lula. En cas de voler recollir informació sobre interaccions entre els diferents gens que componen la xarxa, són més adients les xarxes de regulació gènica (GRN *Gene Regulatory Network*).

Com qualsevol treball de recerca acabat de publicar, és aviat per copsar la rellevància dels resultats obtinguts i si les xarxes publicades juguen un paper significatiu en l'etiologia d'aquestes malalties. Investigacions en el futur, sempre incert, resoldran l'abast i èxit dels objectius proposats.

6. Glossari

ADN (Àcid desoxiribonucleic): molècula que conté la informació genètica.

ARN (Àcid ribonucleic): molècula que participa en la codificació i descodificació de gens.

Cytoscape: programari per a la visualització de grafs.

eQTL (*expression quantitative trait loci)*: associació entre variació genètica i la seva expressió cel·lular.

Expressió gènica: procés pel qual a partir d'un gen se sintetitza el producte gènic corresponent, normalment una proteïna.

Fenotip: característica observable d'un individu.

Genoma de referència: ADN determinat que representa una espècie.

Genotip: conjunt de gens que forma tota la cadena d'ADN.

GRCh38 (*Genome Reference Consortium*): genoma de referència actual.

GTEx (*Genotype-Tissue Expression*): projecte per estudiar l'expressió i regulació de gens en diferents teixits de l'organisme.

GWAS (*Genome-wide association study*): associacions entre variacions genètiques i malalties.

Heat map (mapa de calor): tècnica de visualització de dades que mesura la magnitud d'un fenomen en colors en dues dimensions.

MASHR: model de predicció per a generar fitxers TWAS.

Matriu de correlació: matriu que mostra els valors de correlació de Pearson dels diferents elements que la conformen.

Òmiques: conjunt de tècniques per l'anàlisi de la totalitat de dades d'un cert camp d'estudi.

RNA-seq: tècnica de seqüenciació que mesura la quantitat d'ARN en una mostra.

SNP (*Single Nucleotide Polymorphism*): variació en una base de l'ADN entre individus.

SPrediXcan: programari per generar transcriptoma imputat a partir de GWAS.

Transcriptoma: conjunt d'expressió gènica present dins la cèl·lula.

Tret complex: característiques que no responen al comportament d'un sol gen.

TWAS (Transcriptome-*wide association study*): associacions entre les variacions del transcriptoma i trets complexos.

Tableau: programari per a la visualització de dades.

Xarxa d'expressió gènica (GCN, *Gene co-expression Network*): grup de gens que s'expressen de forma conjunta.

7. Bibliografia

[1] "June 2000 White House Event - National Human Genome Research" 26 de juny 2000, June 2000 White House Event

[2] "A Decade Later, Genetic Map Yields Few New Cures." 12 de juny 2010, A Decade Later, Human Genome Project Yields Few New Cures

[3] "Genome-Wide Association Studies (GWAS)." Genome-Wide Association Studies (GWAS) Play Audio

[4] GTEx Consortium, et al. "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans." *Science* 348.6235 (2015): 648-660.

[5] "MetaXcan software and manuscript - GitHub." hakyimlab/MetaXcan: MetaXcan software and manuscript

[6] "What is a Neurodegenerative Disease? | JPND." What?

[7] "Estadística de defunciones según la causa de muerte. Últimos datos." INEbase / Sociedad /Salud /Estadística de defunciones según la causa de muerte / Últimos datos

[8] Dawkins, Richard, and Nicola Davis. *The selfish gene*. Macat Library, 2017.

[9] Dehaene, Stanislas. *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Penguin, 2014.

[10] Human Behavioral Biology with Prof. Robert Sapolsky at Stanford University - YouTube

[11] Gamazon, Eric R., et al. "A gene-based association method for mapping traits using reference transcriptome data." *Nature genetics* 47.9 (2015): 1091-1098.

[12] GWAS Catalog - EMBL-EBI." GWAS Catalog

[13] Klein, Robert J., et al. "Complement factor H polymorphism in age-related macular degeneration." *Science* 308.5720 (2005): 385-389.

[14] From Disease to Genes and Back | Coursera." From Disease to Genes and Back.

[15] "GTEx Portal." The GTEx Project.

[16] Lowe, Rohan, et al. "Transcriptomics technologies." *PLoS computational biology* 13.5 (2017): e1005457.

[17] Wang, Zhong, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics." *Nature reviews genetics* 10.1 (2009): 57-63.

[18] Costa, Valerio, et al. "RNA-Seq and human complex diseases: recent accomplishments and future perspectives." *European Journal of Human Genetics* 21.2 (2013): 134-142.

[19] Battle, Alexis, et al. "Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals." *Genome research* 24.1 (2014): 14-24.

[20] Geuvadis | IGSR data collection - 1000 Genomes." Geuvadis | IGSR data collection

[21] Gamazon, Eric R., et al. "A gene-based association method for mapping traits using reference transcriptome data." *Nature genetics* 47.9 (2015): 1091-1098.

[22] Li, Binglan, et al. "Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression." *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*. 2018.

[23] Gusev, Alexander, et al. "Integrative approaches for large-scale transcriptome-wide association studies." *Nature genetics* 48.3 (2016): 245-252.

[24] Finucane, Hilary K., et al. "Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types." *Nature genetics* 50.4 (2018): 621-629.

[25] Clarimon, Jordi, et al. "Genetic architecture of neurodegenerative dementias." *Neuropharmacology* 168 (2020): 108014.

[26] Vergouw, Leonie JM, et al. "An update on the genetics of dementia with Lewy bodies." *Parkinsonism & related disorders* 43 (2017): 1-8.

[27] Kamboh, M. I., et al. "Genome-wide association study of Alzheimer's disease." *Translational psychiatry* 2.5 (2012): e117-e117.

[28] Yamazaki, Yu, et al. "Apolipoprotein E and Alzheimer disease: pathobiology and targeting strategies." *Nature Reviews Neurology* 15.9 (2019): 501-518.

[29] Brynedal, B., et al. "MGAT5 alters the severity of multiple sclerosis." *Journal of neuroimmunology* 220.1-2 (2010): 120-124.

[30] Butte, Atul J., and Isaac S. Kohane. "Unsupervised knowledge discovery in medical databases using relevance networks." *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 1999.

[31] de Jong, Simone, et al. "A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes." *PloS one* 7.6 (2012): e39498.

[32] Xiang, Shunian, et al. "Condition-specific gene co-expression network mining identifies key pathways and regulators in the brain tissue of Alzheimer's disease patients." *BMC medical genomics* 11.6 (2018): 39-51.

[33] Gerring, Zachary F., et al. "A gene co-expression network-based analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression." PLoS genetics 15.7 (2019): e1008245.

[34] IM-Lab: Home

[35] PredictDB: Home

[36] Barbeira, Alvaro N., et al. "Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics." Nature communications 9.1 (2018): 1-20.

[37] Figure 1. Sample and data types in the GTEx v8 study. (A) Illustration...

[38] hakyimlab/MetaXcan: MetaXcan software and manuscript

[39] Psychiatric Genomics Consortium

[40] Bonferroni Correction -- from Wolfram MathWorld

[41] Tableau Software

[42] Cytoscape

8. Annexos

Codi del treball: https://github.com/JosepGarciaG/TFM

Visualització de les xarxes gèniques: Tableau

Repositori de dades: Zenodo.

A Formats catàleg GWAS

Els arxius estan estructurats en:
Directori FTP amb:
- arxiu original
- carpeta harmonized:
  - arxiu format standard (nom de columnes)
  - arxiu format harmonitzat (mapejat

Font: GWAS Catalog

A1. Arxiu format estàndard

The standard format summary statistics file:

- Is saved as .tsv
- Contains mandatory fields in the following combinations:

```
a) variant_ID and p-value
OR
b) chromosome, base_pair_location and p-value
```

Additional standard column headings are listed below.

Standard format file headings

```
'variant_id' = variant ID
'p-value' = p-value
'chromosome' = chromosome
'base_pair_location' = base pair location
'odds_ratio' = odds ratio
'ci_lower' = lower 95% confidence interval
'ci_upper' = upper 95% confidence interval
'beta' = beta
'standard_error' = standard error
'effect_allele' = effect allele
'other_allele' = other allele
'effect_allele_frequency' = effect allele frequency
```

Note that the headers in the formatted file are not limited to the above headers.

A2. Arxiu harmonitzat

Arxiu mapejat a l'últim genoma de referència (GRCh38)

Harmonised file headings (not all may be present in file):

```
'variant_id' = variant ID
'p-value' = p-value
'chromosome' = chromosome
'base_pair_location' = base pair location
'odds_ratio' = odds ratio
'ci_lower' = lower 95% confidence interval
'ci_upper' = upper 95% confidence interval
'beta' = beta
'standard_error' = standard error
'effect_allele' = effect allele
'other_allele' = other allele
'effect_allele_frequency' = effect allele frequency
'hm_variant_id' = harmonised variant ID
'hm_odds_ratio' = harmonised odds ratio
'hm_ci_lower' = harmonised lower 95% confidence interval
'hm_ci_upper' =  harmonised lower 95% confidence interval
'hm_beta' = harmonised beta
'hm_effect_allele' = harmonised effect allele
'hm_other_allele' = harmonised other allele
'hm_effect_allele_frequency' = harmonised effect allele frequency
'hm_code = harmonisation code (to lookup in 'Harmonisation Code Table')
```

Font: NHGRI-EBI GWAS Catalog

B Tipologia TWAS generats per SPrediXcan

Where each row is a gene's association result:

- `gene` : a gene's id: as listed in the Tissue Transcriptome model. Ensemble Id for most gene model releases. Can also be a intron's id for splicing model releases.
- `gene_name` : gene name as listed by the Transcriptome Model, typically HUGO for a gene. It can also be an intron's id.
- `zscore` : S-PrediXcan's association result for the gene, typically HUGO for a gene.
- `effect_size` : S-PrediXcan's association effect size for the gene. Can only be computed when `beta` from the GWAS is used.
- `pvalue` : P-value of the aforementioned statistic.
- `pred_perf_r2` : (cross-validated) R2 of tissue model's correlation to gene's measured transcriptome (prediction performance). Not all model families have this (e.g. MASHR).
- `pred_perf_pval` : pval of tissue model's correlation to gene's measured transcriptome (prediction performance). Not all model families have this (e.g. MASHR).
- `pred_perf_qval` : qval of tissue model's correlation to gene's measured transcriptome (prediction performance). Not all model families have this (e.g. MASHR).
- `n_snps_used` : number of snps from GWAS that got used in S-PrediXcan analysis
- `n_snps_in_cov` : number of snps in the covariance matrix
- `n_snps_in_model` : number of snps in the model
- `var_g` : variance of the gene expression, calculated as `W' * G * W` (where `W` is the vector of SNP weights in a gene's model, `W'` is its transpose, and `G` is the covariance matrix)

Font: hakyimlab/MetaXcan: MetaXcan software and manuscript

C Documentació GWAS analitzats

C1 Malalties amb resultats significatius

Alzheimer's disease	Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes
ADHD	Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder
Anorexia nervosa	Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervos
Bipolar disorder	Genome-wide association study identifies 30 loci associated with bipolar disorder
Major depression	A mega-analysis of genome-wide association studies for major depressive disorder
Multiple sclerosis	Widespread dose-dependent effects of RNA expression and splicing on complex diseases and traits
Neuroticisme	Widespread dose-dependent effects of RNA expression and splicing on complex diseases and traits
Slepplessness	Widespread dose-dependent effects of RNA expression and splicing on complex diseases and traits
Squizofrenia	Comparative genetic architectures of schizophrenia in East Asian and European populations

C2 Malalties descartades

Epilepsy	A cross-population atlas of genetic associations for 220 human phenotypes
Intracerebral hemorrhage	A cross-population atlas of genetic associations for 220 human phenotypes
Ischemic stroke	A cross-population atlas of genetic associations for 220 human phenotypes
Mental or behavioural disorder	Genome-wide association study of psychiatric and substance use comorbidity in Mexican individuals
Parkinson's disease	A cross-population atlas of genetic associations for 220 human phenotypes
Febrile seizures	A cross-population atlas of genetic associations for 220 human phenotypes
Brain tumor	A cross-population atlas of genetic associations for 220 human phenotypes
OCD	Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis

Anxiety disorders	Meta-analysis of genome-wide association studies of anxiety disorders
Panic disorder	Genome-wide association study of panic disorder reveals genetic overlap with neuroticism and depression
Tourette's syndrome	Interrogating the Genetic Determinants of Tourette's Syndrome and Other Tic Disorders Through Genome-Wide Association Studies
Cerebral aneurysm	A cross-population atlas of genetic associations for 220 human phenotypes
ASD	Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia
Amyotrophic lateral sclerosis	Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis

Del catàleg prepublished de GWAS: Agorophobia, delirium dementia, neoplasm brain, neurological disorders, paranoid disorders, personality disorders, psychogenic disorders, transient mental disorders
Studies with available summary statistics - GWAS Catalog