



Màster Universitari

**Anàlisi de Dades Òmiques /
Omics Data Analysis**

FACULTAT DE CIÈNCIES I TECNOLOGIA

UVIC | UVIC·UCC

Master of Science in Omics Data Analysis

Master Thesis

Evaluating the polygenicity of brain structure features using Compositional Data Analysis

by

Patricia Genius Serra

Supervisor: Natalia Vilor Tejedor, BarcelonaBeta Brain Research Center (BBRC), Centre for Genomic Regulation (CRG) and Erasmus University Medical Center Rotterdam

Co-supervisor: M. Luz Calle Rosingana, Bioscience Department, University of Vic-Central University of Catalonia

Academic tutor: M. Luz Calle Rosingana, Bioscience Department, University of Vic-Central University of Catalonia

Biosciences Department

University of Vic – Central University of Catalonia

September 15th, 2021

Acknowledgements

When I first decided to get started in the Master of Omics Data Analysis I could have never imagined such a lot of work during the whole year. I am not lying if I say that it has been an exciting but tough year. However, I have really enjoyed being part of this course, not only for the knowledge and abilities that I have acquired but also for the people, both students and teachers, that have been part of it. We have not had the opportunity to deeply meet each other, but even in remote, I have felt next to them day by day, and it is not such an easy task. So my first thanks are for all of them, for all of you.

Moreover, I would really like to dedicate some words to my supervisors, Dr. Natalia Vilor Tejedor and Dr. M.Luz Calle Rosingana, because I feel very grateful to them. I want to thank your support, guidance, time, effort, patience, proximity and kindness. I have had an incredible accompaniment and I have really enjoyed working and learning with you. Thank you very much, I really appreciate each and every single detail.

Finally, I would also like to thank all the researchers of Barcelonaβeta Brain Research Center, specially the neuroimaging group, and all the participants of the ALFA study, whose voluntary participation is essential for the research. I am very grateful to have had the opportunity to work and grow during a whole year in this group.

Abstract

Background: Imaging genetics (IG) studies aim to jointly analyse neuroimaging and genetic data with the objective of discovering new genetic variations related to brain features. Most IG studies focus on the individual analysis of brain structures. An alternative strategy is to incorporate compositional data analysis (CoDA) methods to assess the joint modulation of specific brain subregions.

Objective: The aim of this project was to investigate whether the genetic predisposition to specific neurodegenerative disorders (quantified with polygenic risk scores, PRS) was associated with the joint modulation of hippocampal subfields volumes (target regions for neurological disorders) by assessing the performance of CoDA (*Selbal* algorithm).

Methods: A total of 1,071 participants from the ALFA study with available information on genetics and neuroimaging data were included. Genetic predisposition to Alzheimer's Disease (AD), Amyotrophic Lateral Sclerosis (ALS) and Progressive Supranuclear Palsy (PSP) was estimated by calculating PRS (*PRSice v.2*). *Selbal* algorithm was applied to find the hippocampal subregions whose joint volumetric variation was most closely related to a higher genetic risk of each neurodegenerative condition. Logistic regression models were assessed to test the association between the genetic predisposition of each condition and the volumetric combination of the hippocampal subfields. Models were adjusted by sex and we also performed sex- and hemisphere-stratified models.

Results: Results showed that a compensatory increase in the average volume of CA3, CA4 and hippocampal fissure related to CA1 and hippocampal tail was significantly associated with a higher genetic risk of AD. Results also showed that a higher genetic risk of ALS was significantly related to a compensatory increase in the CA1 compared to the hippocampal fissure. Results for PSP showed that a compensatory increase in the subiculum in comparison to the parasubiculum was significantly associated with a higher genetic risk. Moreover, we found different joint volumetric modulation of hippocampal substructures associated with higher genetic risk of each condition between sex, as well as among hemispheres.

Conclusion: To our knowledge, this is the first study analysing the relationship between cognitively healthy individuals at high genetic risk of AD, ALS, and PSP and the joint volumetric variation of hippocampal subfields. Therefore, this work provides a new and innovative perspective for IG studies with the aim of improving our understanding of the effects that the genetic predisposition to neurodegenerative disorders has on brain structure modulation.

Keywords: Alzheimer's Disease; Amyotrophic Lateral Sclerosis; Compositional Data Analysis; Imaging Genetics; Progressive Supranuclear Palsy; Polygenic Risk Score; Selbal

Table of contents

1	Introduction	1
2	Material and methods	3
2.1	<i>Sample description: AlfaGeneTiCs project</i>	3
2.2	<i>Genetic data acquisition: genotyping, quality control and imputation</i>	3
2.3	<i>Acquisition of MRI and hippocampal subfields segmentation</i>	4
2.4	<i>Polygenic risk scores (PRS) computation and validation</i>	5
2.5	<i>Statistical analysis</i>	6
2.5.1	<i>Descriptive analysis</i>	6
2.5.2	<i>Compositional data: definition and methods</i>	6
2.5.3	<i>Selbal implementation in our Imaging Genetic study</i>	8
3	Results	10
3.1	<i>Descriptive analysis</i>	10
3.2	<i>Selbal algorithm results</i>	11
3.2.1	<i>Genetic risk of Alzheimer’s Disease and Hippocampal subfields modulation</i>	11
3.2.2	<i>Genetic risk of Amyotrophic Lateral Sclerosis and Hippocampal subfields modulation</i>	12
3.2.3	<i>Genetic risk of Progressive Supranuclear Palsy and Hippocampal subfields modulation</i>	13
4	Discussion and Conclusions	14
5	References	18
6	Appendix	22
6.1	<i>Supplemental Tables</i>	22
6.2	<i>Supplemental Figures</i>	28
6.3	<i>Supplemental formulas and examples</i>	34
6.4	<i>Bioinformatic pipeline (continues on next page)</i>	35

Glossary

AD: Alzheimer's Disease

alr: Additive log-ratio

ALS: Amyotrophic Lateral Sclerosis

AUC: Area Under the Curve

CA1: Cornu Ammonis 1

CA3: Cornu Ammonis 3

CA4: Cornu Ammonis 4

clr: centered log-ratio

CoDA: Compositional Data Analysis

CV: Cross-Validation

DEV: Deviance

FTD: Frontotemporal dementia

GC-ML-DG: Granular Cell and Molecular Layer of the Dentate Gyrus

GWAS: Genome-Wide Association Studies

HATA: Hippocampal Amygdala Transition Area

HWE: Hardy-Weinberg Equilibrium

IG: Imaging Genetics

ilr: isometric log-ratio

LD: Linkage Disequilibrium

MAF: Minor Allele Frequency

MFA: Multiple Factor Analysis

MRI: Magnetic Resonance Imaging

MSE: Mean Squared Error

PRS: Polygenic Risk Score

PSP: Progressive Supranuclear Palsy

QC: Quality Control

Selbal: Selection of balances

SNP: Single Nucleotide Polymorphism

1 Introduction

Many complex diseases have a genetic component and its study might provide valuable insights into the etiology of the disease. Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with complex disorders (Loos 2020). However, for neurological disorders, the number of genetic variants identified so far is considerably reduced (Buniello et al., 2019). One possible reason is the heterogeneity of the clinical diagnosis of these disorders, which can be ameliorated by the use of brain-based features (i.e. brain morphology, brain physiology, cognitive function) as intermediate phenotypes (Glahn et al., 2007; Matoba et al., 2020). Intermediate brain phenotypes derived from neuroimaging sequences not only reduce the phenotypic heterogeneity common to many neurodegenerative disorders, but also increase detection power (Hashimoto et al., 2015).

Neuroimaging studies based on neurological-related processes usually focus on specific brain structures, one of the most important being the hippocampus, both for its involvement in several neurological disorders as well as for its subfields' unique molecular properties (Flores et al., 2015; Evans et al., 2018; Vilor-Tejedor et al., 2021). Although hippocampal region is often treated as a unitary structure, the hippocampus is composed of multiple subfields with distinct molecular and functional properties (Van der Meer et al., 2020). Given the structural and functional heterogeneity of the hippocampal subfields, it is reasonable to suggest that each subfield may have distinct genetic influences and that some subfields may be more affected than others by the individual genetic variation (Elman et al., 2019).

The impact of genetic variation on these brain features can be assessed in imaging genetics (IG) studies. IG studies aim to integrate neuroimaging and genetic data with the objective of discovering new genetic variants related to brain features, which in turn explain the risk for neurodegenerative disorders. This approach can promote an understanding of the genetic basis of neurodegenerative disorders as well as provide insights into the genetic architecture of the brain that could be relevant to neurological disorders, brain development and ageing (Elliot et al., 2018, Nathoo et al., 2019, Vilor-Tejedor et al., 2018). However, jointly analysing neuroimaging and genetic data raises serious challenges for efficiently analysing large-scale data sets with many subjects.

The earliest methods developed were based on candidate genetic variants and specific brain features. This type of candidate univariate analysis is based on a standard linear regression relating a given brain feature, which is commonly the volume measurement of the brain structure (e.g. hippocampal volume) to a given genetic variant (e.g. Single Nucleotide Polymorphisms; SNP). This strategy can be extended to the full brain-wide and genome-wide data, resulting in a massive number of pairwise univariate analyses (Hibar et al., 2017; Smith et al., 2021). In this case, the multiple testing correction problem becomes evident, where very stringent corrections are applied to control for false positives along the large number of tests involved in the analysis. This

correction decreases the power of the analysis, increasing the difficulty to identify genetic variants associated with a brain feature of interest. Therefore, large samples, as well as joint consortia efforts are required.

An alternative strategy to the massive univariate approach are techniques that jointly analyse genetic data. These strategies fit regression models separately at each brain feature, considering a set of genetic markers simultaneously rather than just a single genetic marker (Dima and Breen 2015; Yao et al., 2020). Polygenic risk scores (PRS) are extremely useful in this context. PRSs combine the individual effect of each genetic variant in a single score that summarizes the genetic predisposition of each individual to a specific disorder/condition (Sugrue et al., 2019). Commonly, in IG studies, when PRSs are calculated, univariate regression models are then adjusted to assess the association between the given PRS and the specific brain feature of interest. Finally, another strategy consists of analysing both multivariate neuroimaging and genetic data to better capture the complex relationships that may exist between different biological levels (Vilor-Tejedor et al., 2018a). An example of this strategy is the application of the multiple factorial analysis (MFA) and its extensions (Vilor-Tejedor et al., 2019, Vilor-Tejedor et al., 2018b). The inclusion of the multivariate perspective provides an improvement in the statistical power and predictive capacity in IG studies. However, none of the aforementioned strategies consider the joint effect of nearby brain features in the volumetric variation of the target brain feature.

In this work we proposed a new strategy for volumetric analysis in IG studies based on the use of compositional data analysis (CoDA) methods. Compositional data consists of a set of measurements whose values are restricted by their total sum. In our context, the main brain region of interest was the hippocampal structure that can be analysed as the composition of different subfields or components (Figure 1.1).

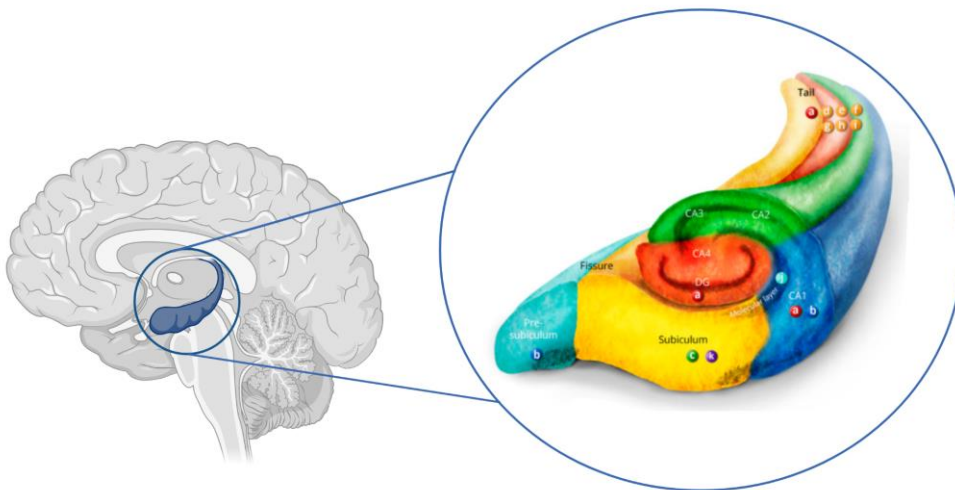


Figure 1.1. Hippocampal segmentation. Adapted from Vilor-Tejedor et al., 2021.

In CoDA, instead of exploring each component separately, the analysis focuses on the relative variation between them. *Selbal* is a recent approach in CoDA which is based on a model selection procedure that looks for the most parsimonious model capable of explaining the association between the joint effect of the selected components and a specific phenotype of interest (Calle, 2019). Thus, by applying the *Selbal* algorithm, we can select the brain components (hippocampal subfields) whose joint volumetric variation, summarized in a single score known as *balance*, is most closely associated with the genetic predisposition to several neurological diseases.

Therefore, the aim of this final master project was to investigate whether the genetic predisposition to specific neurodegenerative disorders (quantified with PRS) was associated with the joint modulation of specific hippocampal subfields volumes by assessing the performance of the *Selbal* algorithm.

2 Material and methods

2.1 Sample description: AlfaGeneTiCs project

Individuals from the ALFA cohort (Molinuevo et al., 2016) were invited to take part in the AlfaGeneTiCs study. The AlfaGeneTiCs study is composed of 2,280 cognitively unimpaired middle-age participants (45-75 years old), most of them Alzheimer's disease (AD) patient's offspring with a high proportion of *APOE-ε4* carriers. They present available information in cognition, neuroimaging (magnetic resonance imaging sequences), lifestyle, clinical history and blood collection. For this study, 1,071 participants with available information on hippocampal subfields quantification were included.

2.2 Genetic data acquisition: genotyping, quality control and imputation

DNA samples of AlfaGeneTiCs participants were obtained from whole blood samples by applying salting out protocol. DNA was eluted in 800µl of H₂O (milliQ) and quantified using Quant-iTTM PicoGreen[®] dsDNA Assay Kit (Life Technologies). Integrity of DNA was checked in a subset of samples by running a 1% agarose gel. All the samples were within specification. DNA concentration for each sample was additionally normalized.

Genome-wide genotyping was performed using the Illumina Infinium Neuro Consortium (NeuroChip) Array (build GRCh37/hg19) (Blauwendraat et al., 2017).

Quality Control (QC) procedure of the genetic data was conducted with PLINK software (Anderson et al., 2010). The following sample quality control thresholds were applied: individuals with sample call rate of less than 98%, and exhibiting excess of heterozygosity (3 standard deviations) were excluded. Moreover, we excluded individuals showing sex discordances. Finally, individuals at higher genetic relatedness (at the level of cousin or closer) sharing proportionally more than 18.5% of alleles ($IBD > 0.185$) were also excluded.

Once QC on sample level was completed, genetic variants with minor allele frequency (MAF<1%), Hardy-Weinberg equilibrium (HWE) p-value < 10⁻⁶, and missiness rates > 5% were excluded.

Imputation of genetic variants was performed by using the Michigan imputation server (<https://imputationserver.sph.umich.edu>), using the Haplotype Reference Consortium panel (HRC r1.1 2016) (Das et al., 2016; McCarthy et al., 2016) under default parameters and according to established guidelines. Assessment and preparation of genetic data was performed in previous studies.

2.3 Acquisition of MRI and hippocampal subfields segmentation

Radiologists from the ALFA study obtained the volumes for the hippocampal subfields through high-resolution 3D-T1-weighted magnetic resonance imaging scans. The left and right hippocampus were segmented into twelve subregions: cornu ammonis region 1, 3, 4 (CA1, CA3, CA4), fimbria, granular cell and molecular layer of the dentate gyrus (GC-ML-DG), hippocampal-amygdalar region (HATA), hippocampal fissure, hippocampal tail, molecular layer, parasubiculum, presubiculum and subiculum. Volumes were also quantified globally by summing the subfields volumes of both hemispheres. Thus, a total of 36 measures were obtained: 12 for the right hemisphere, 12 for the left and 12 for the global volume of the regions. The images were pre-processed and determined using Freesurfer version 6 (Iglesias et al, 2015).

[Figure 2.3](#) shows different snapshots of the volume quantification for the 12 subfields that make up the hippocampal region for an individual in the AlfaGeneTiCs project. These snapshots were used for visual quality control criteria of hippocampal subfields segmentation performed by a specialist.

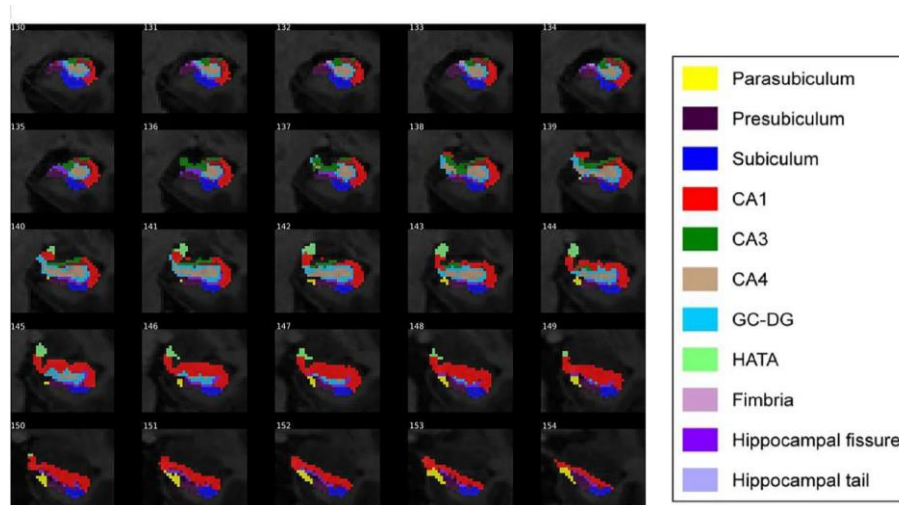


Figure 2.3. T1 images of hippocampal segmentation. From [Vilor-Tejedor et al., 2020](#).

The final sample of the study is described in ([Figure 2.4](#)).

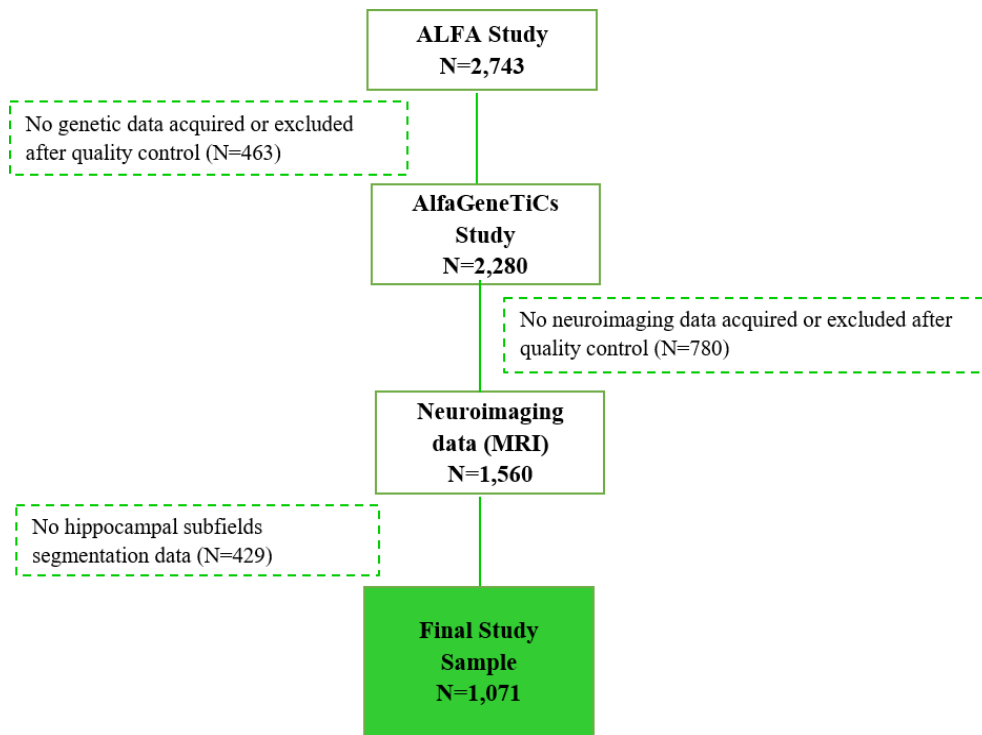


Figure 2.4. Flowchart describing the final sample size. The study sample included 1,071 individuals with available hippocampal subfields quantification and demographic data.

2.4 Polygenic risk scores (PRS) computation and validation

We used PRSice version 2 (Choi et al., 2019) which is an algorithm that computes PRSs by summing up all the SNP alleles carried by the participants weighted by the SNP allele effect size estimated in a previous GWAS, normalizing the score by the total number of alleles (Figure A11). PRSs were computed in representative genetic variants per linkage disequilibrium (LD) block (clumped variants), using a cut-off for LD of $r^2 > 0.1$ in a 250-kb window. The clumping procedure retained SNPs with the smallest p-value in each 250 kb window, and removed all those in LD ($r^2 > 0.1$). PRSs were based on the most recently published GWAS. Further details can be found in Supplementary Table A1.

We computed PRSs for 7 neurodegenerative conditions: AD, AD excluding the *APOE* gene region (chr 19: 45,409,011-45,412,650), amyotrophic lateral sclerosis (ALS), frontotemporal dementia, a meta-phenotype including frontotemporal dementia subtypes, Parkinson’s disease and progressive nuclear palsy (PSP). All PRSs were Z-standardized and dichotomized by taking as threshold the 0.8 quantile in order to compare high vs low genetic risk groups.

PRSs that did not have a continuous distribution or that presented high values both for skewness ($|S| > 3$) and kurtosis ($|K| > 10$) indexes (Kline, 2011) were excluded for the genetic association

models ([Figure A1](#), [Table A8](#)). Finally, we included PRSs for AD, ALS and PSP in the genetic association models.

2.5 Statistical analysis

2.5.1 Descriptive analysis

Descriptive analyses were performed to describe the socio demographic variables of the study (age, sex, and years of education). PRSs and hippocampal subfields volumes were also analysed. We assessed differences in the hippocampal subfields' mean volume within each sex group, and also between right and left hemispheres. We performed the same analyses, assessing the hippocampal subfields mean volume differences within each hemisphere, either for the whole sample as well as stratifying by sex.

2.5.2 Compositional data: definition and methods

Compositional data is defined by a vector of strictly positive real numbers with a constraint or non-informative total sum (Calle, 2019).

$$x = [x_1, \dots, x_D] \in R^D$$

for $x_i > 0$, $\sum_{i=1}^D x_i = k$, where k is a constant (e.g. $k=1$, $k=100, \dots$)

In a composition, the value of each component is not informative by itself and the relevant information is contained in the ratios between parts. This property implies that two proportional compositions are equally informative and this induces equivalence classes of vectors carrying the same information. Two vectors are compositionally equivalent if they are proportional.

To work with compositional methods, the data needs to accomplish different properties: permutation invariance, scale invariance and sub-compositional coherence. Permutation invariance means that the change in the order of the parts in the composition should not affect results. Scale invariance refers to the fact that multiplying all the components by a factor does not alter the results of the analysis. Sub-compositional coherence is found when results for a subset of the composition are coherent with the results for the whole composition (Calle, 2019). The simplest scale invariant function is the log-ratio between components, defined as $\log(x_i / x_j)$. Working with log-ratios is the key aspect of compositional data analysis and is known as the log-ratio approach (Calle, 2019) ([Formulas and examples section](#), Appendix).

Some data transformations that are commonly used in CoDA: additive log-ratio (*alr*), centred log-ratio (*clr*) and isometric log-ratio (*ilr*). The *alr* transformation applies a log-ratio between each component and a reference component. The *clr* scales each component by the geometric mean of the parts and the *ilr* transformation, that is associated with an orthogonal coordinate system in the

simplex. Although all these methods are applied by different authors, they present some limitations especially in the context of variable selection (Susin et al., 2020).

Selbal approach

There is a recent approach in CoDA, the selection of balances (*Selbal*), defined by a model selection procedure that searches for a sparse model that explains the response variable of interest, based on the joint change of the specific selected components of the composition (Rivera-Pinto et al, 2018).

Selbal relies on *compositional balances*, a measure that extends the log-ratio between two components to the log-ratio between two groups of components and is defined as follows: Given A and B two disjoint subcompositions (subgroups of components) of a composition, the balance between A and B is defined as the log-ratio between the geometric mean for each group:

$$\text{Balance } B_{(A,B)} = K \cdot \log(g(A)/g(B)) = K \cdot \log \frac{(\prod_{i \in IA} x_i)^{\frac{1}{k_A}}}{(\prod_{j \in IB} x_j)^{\frac{1}{k_B}}},$$

where $g(\cdot)$ is the geometric mean and K is a normalization constant

$$K = \sqrt{\frac{k_A \cdot k_B}{k_A + k_B}}$$

The balance can also be written as the difference between the arithmetic means of the log-transformed values:

$$B_{(A,B)} = \frac{1}{k_A} \cdot \sum_{i \in IA} \log x_i - \frac{1}{k_B} \cdot \sum_{j \in IB} \log x_j$$

Selbal is a joint procedure that involves both modelling and variable selection, through forward selection. The goal is to determine two sub-compositions A and B whose balance $B(A,B)$ is the most associated with the dependent variable Y after adjusting for specific covariates Z according to the following linear or logistic regression model:

$$Y = \beta_0 + \beta_1 B_{(A,B)} + \gamma Z, \text{ when Y is continuous}$$

$$\text{logit}(Y) = \beta_0 + \beta_1 B_{(A,B)} + \gamma Z, \text{ when Y is binary}$$

Selbal evaluates all possible balances composed of only two components (x_i, x_j) . Each balance is tested for association with the response variable. The optimal two-component balance is selected and at each step, a new component (i.e. subfield volume) is added to the existing balance until there is no additional variable that improves the specific optimization parameter. The maximum number of components to be included in the model is defined through a cross-validation (CV) procedure (Calle, 2019).

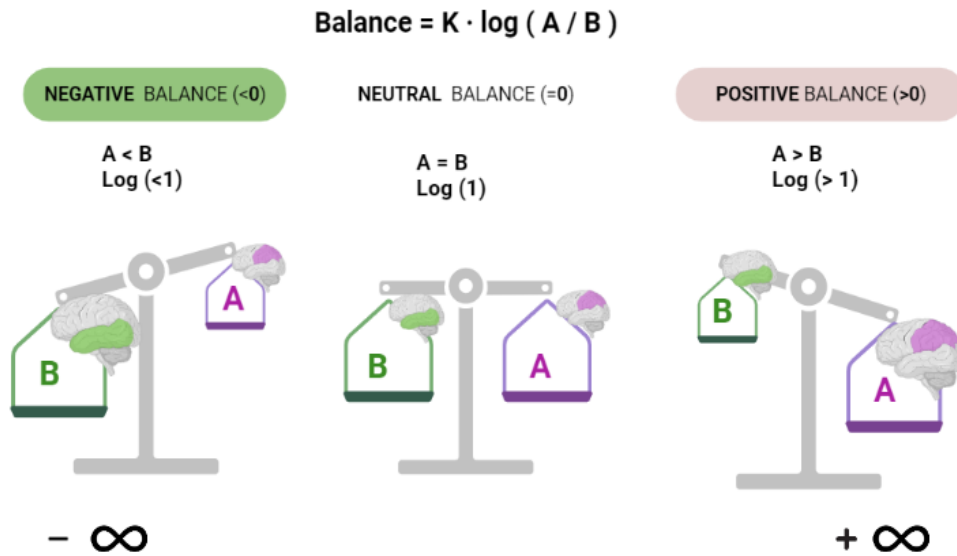


Figure 2.5. Graphical representation of the concept *balance*: possible scenarios and values that the balance can take. Positive scores involve larger average volume of subcomponents in group A compared to those in group B, while negative scores involve larger average volume of subcomponents in group B compared to those in group A.

Finally, a global balance is obtained through the selection of specific components of the composition. It results in a one-dimension measure, which can take values from $-\infty$ to $+\infty$. Thus, in our study, it can be interpreted as a score that summarizes the average log-transformed volumes of two groups of subfields volumes. When the global balance is equal 0, there is a perfect balance between both groups of components, which means that the ratio between subfields' volumes is equal 1 ($\log(1) = 0$). When the global balance takes a positive value, the average volume of the subregions in the numerator (group A) is larger in comparison to the subregions that are in the denominator (group B). When the global balance takes a negative value, the average volume of the subregions in the numerator (group A) is lower than the average volume of the subregions that are in the denominator (group B) (Figure 2.5). For further details about its application see the *Toy example* in (Susin et al., 2020).

2.5.3 *Selbal* implementation in our Imaging Genetic study

We implemented *Selbal* to the hippocampal subfields volumes composition with logistic regression models, where PRSs were defined as dichotomous outcomes (high/low genetic risk group). The selected global balance identified two groups of hippocampal subfields whose relative volume was most associated with the genetic predisposition to a specific neurological condition. Models were adjusted by sex. We also assessed sex-stratified logistic regression models.

A total of 3 different hippocampal volume compositions were defined. The first composition (C_1) was the global hippocampal region, where the components were defined by each hippocampal

substructure, without distinguishing per hemisphere. Second and third compositions were defined for the right (C_2) and left (C_3) hippocampal regions. Each one of these two compositions were defined by 12 components (12 hippocampal substructures).

More formally, compositions were defined as:

$$C_1 = (X_{1T}, \dots, X_{12T}),$$

where X_{iT} for $i \in 1, \dots, 12$, is the total (T) volume of the hippocampal subfield i .

$$C_2 = (X_{1R}, \dots, X_{12R}),$$

where X_{iR} for $i \in 1, \dots, 12$, is the right hemisphere (R) volume of the hippocampal subfield i .

$$C_3 = (X_{1L}, \dots, X_{12L}),$$

where X_{iL} for $i \in 1, \dots, 12$, is the left hemisphere (L) volume of the hippocampal subfield i .

Different models were established according to each composition:

1. How is the genetic predisposition to specific neurological conditions related to the joint volumetric variation of hippocampal substructures?

$$(M1) \text{ logit}(PRSi) = \beta_0 + \beta_1 \cdot \text{Balance}(A, B)_{TotalVolumes} + \beta_2 \cdot \text{Sex}$$

- 1.1. Is this genetic predisposition differentially affecting the joint volumetric variation of hippocampal substructures among women and men?

$$(M1.1) \text{ logit}(PRSi_{Women}) = \beta_0 + \beta_1 \cdot \text{Balance}(A, B)_{TotalVolumesWomen}$$

$$(M1.2) \text{ logit}(PRSi_{Men}) = \beta_0 + \beta_1 \cdot \text{Balance}(A, B)_{TotalVolumesMen}$$

2. How is the genetic predisposition to specific neurological conditions related to the joint volumetric variation of hippocampal substructures in the right hemisphere?

$$(M2) \text{ logit}(PRSi) = \beta_0 + \beta_1 \cdot \text{Balance}(A, B)_{RightVolumes} + \beta_2 \cdot \text{Sex}$$

2. 1. Is this genetic predisposition differentially affecting the joint volumetric variation of hippocampal substructures in the right hemisphere among women and men?

$$(M2.1) \text{logit}(PRSi_{Women}) = \beta_0 + \beta_1 \cdot \text{Balance}(A, B)_{RightVolumesWomen}$$

$$(M2.2) \text{logit}(PRSi_{Men}) = \beta_0 + \beta_1 \cdot \text{Balance}(A, B)_{RightVolumesMen}$$

3. How is the genetic predisposition to specific neurological conditions related to the joint volumetric variation of hippocampal substructures in the left hemisphere?

$$(M3) \text{logit}(PRSi) = \beta_0 + \beta_1 \cdot \text{Balance}(A, B)_{LeftVolumes} + \beta_2 \cdot \text{Sex}$$

3.1. Is this genetic predisposition differentially affecting the joint volumetric variation of hippocampal substructures in the left hemisphere between women and men?

$$(M3.1) \text{logit}(PRSi_{Women}) = \beta_0 + \beta_1 \cdot \text{Balance}(A, B)_{LeftVolumesWomen}$$

$$(M3.2) \text{logit}(PRSi_{Men}) = \beta_0 + \beta_1 \cdot \text{Balance}(A, B)_{LeftVolumesMen}$$

To avoid repetitive tables of results, in the main text we only present the results of model 1. Results of models 2 and 3, are described in the Appendix ([Figure A9-10](#)).

3 Results

3.1 Descriptive analysis

The AlfaGeneTiCs sample was defined by 62.5% of women and 37.5% of men with a similar mean age of 59 (58.7 ± 6.69 women, 58.8 ± 6.56 men) years old ([Table 3.1](#)). Years of education were significantly different between both groups.

Table 3.1. Descriptives of the main covariates of the study differentiating by sex.

	Total N=1071	Women N=670 (62.5%)	Men N=401 (37.5%)	P-value (Wilcoxon test)
Age	58.7 (± 6.64)	58.7 (± 6.69)	58.8 (± 6.56)	0.765
Education	13.5 (± 3.54)	13.3 (± 3.56)	13.9 (± 3.47)	0.004

In [Table 3.2](#) we showed the sample distribution, stratifying by sex, across high and low risk groups for each condition. From the total of individuals at high genetic risk of AD, ALS and PSP, women represented between 60-65% of them, respectively. The same was observed in the low risk group. When we considered the total of women and men in the sample, we saw that both for women and men, the number of individuals at high risk was around 20-25%, while the number of individuals at low risk was around 75-80%. Women and men were proportionally equally represented in the high and low risk groups, although women represented more than 60% of both high and low risk group (they also represented more than 60% of the total sample).

Table 3.2. Descriptives of the main covariates of the study differentiating by high and low genetic risk of each condition, and sex.

	High risk			Low risk		
	Women N=670 (62.5%)	Men N=401 (37.5%)	p-value	Women N=670 (62.5%)	Men N=401 (37.5%)	p-value
Age	58.7 (± 6.43)	58.9 (± 6.36)	0.712	59.1 (± 6.76)	59.3 (± 6.67)	0.636
Education	13.5 (± 3.53)	13.9 (± 3.45)	0.165	13.2 (± 3.56)	14.0 (± 3.47)	<0.001
Disease:			0.759			0.934
AD	133 (62%)	82 (38%)		537 (63%)	319 (37%)	
ALS	165 (63%)	96 (37%)		505 (62%)	305 (38%)	
PSP	141 (65%)	75 (35%)		529 (62%)	326 (38%)	

3.2 Selbal algorithm results

As mentioned before, in this section we only present the results of the first composition (i.e. total volumes). Results of component selection can be found in [Figures A6-A8](#). Moreover, results for the right and left hemisphere (second and third composition) can be found in [Figures A9-10](#).

3.2.1 Genetic risk of Alzheimer's Disease and Hippocampal subfields modulation

When the composition was defined by substructures within the hippocampal region, the global balance for the whole sample was defined by CA3, CA4, hippocampal fissure, CA1 and hippocampal tail ([Figure 3.1a](#)). Results showed that a 10% compensatory increase in the average volume of CA3, CA4 and hippocampal fissure compared to CA1 and hippocampal tail was significantly associated with an increased genetic risk of AD (OR=1.53 [1.22-1.92]) ([Figure 3.1d](#)). In women, the global balance was defined by CA4 and hippocampal tail ([Figure 3.1b](#)). Results showed that a 10% compensatory increase in the CA4 region with respect to the hippocampal tail was significantly associated with an increased genetic risk of AD (OR=1.45 [1.08-1.96]) ([Figure 3.1d](#)). Finally, in men, the global balance was defined by CA3 and CA1 ([Figure 3.1c](#)). Results showed that a 10% compensatory increase in the CA3 with respect to CA1

was significantly associated with an increased genetic risk of AD (OR=1.73 [1.18-2.57]) (Figure 3.1d).

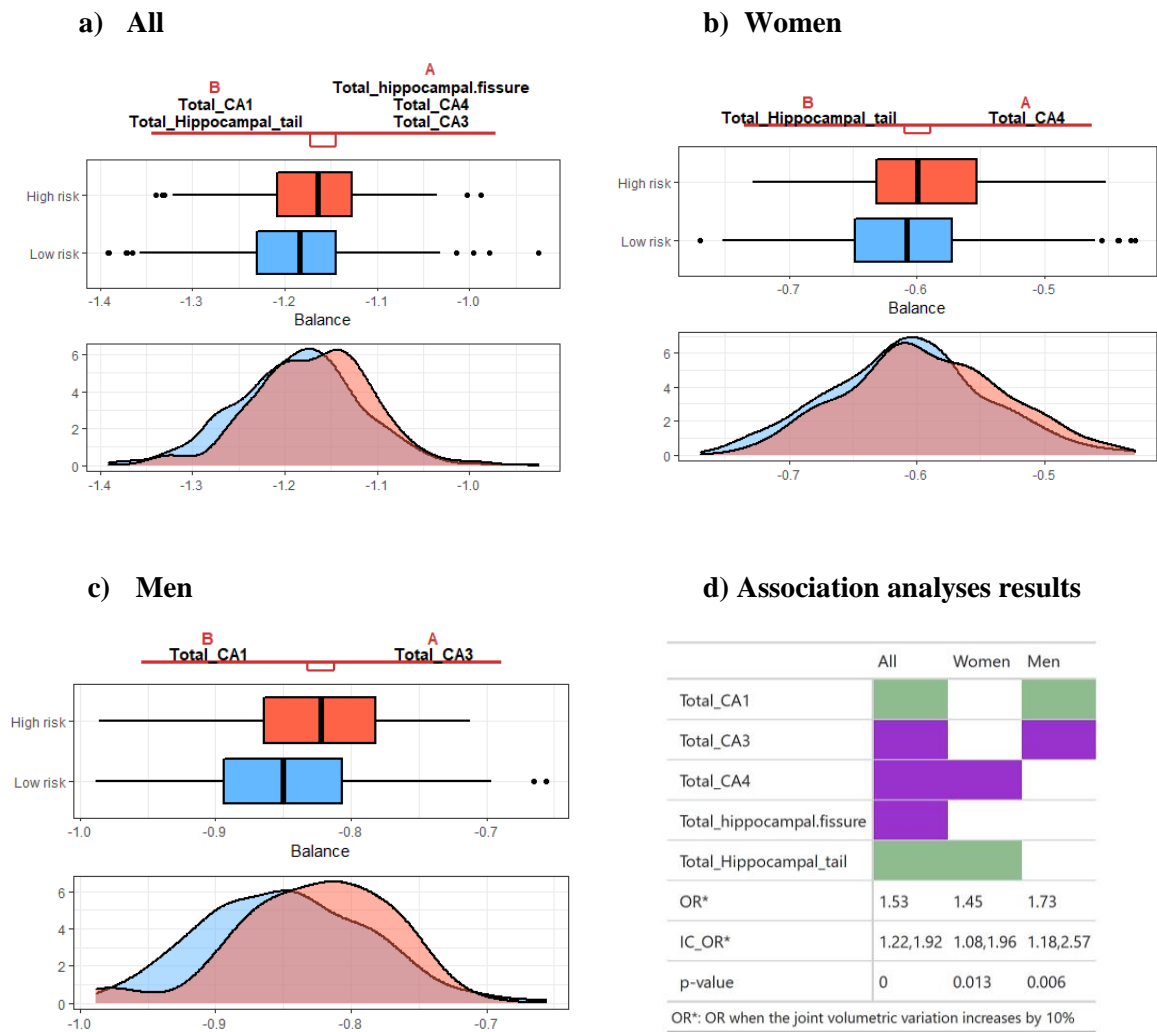


Figure 3.1. Figure 3.1a, 3.1b and 3.1c display hippocampal subfields in group A and B that define the global balance. Boxplots represent the distribution of the balance values for individuals at high (red) and low (blue) genetic risk of AD, when the sample is defined by all the individuals (Figure 3.1a), women (Figure 3.1b), and men (Figure 3.1c). The density plot is described below. Figure 3.1d the association analyses results, which include the balance as explanatory variable.

3.2.2 Genetic risk of Amyotrophic Lateral Sclerosis and Hippocampal subfields modulation

When the composition was defined by substructures within the hippocampal region, the global balance for the whole sample was defined by CA1 and hippocampal fissure (Figure 3.2a). Results showed that a 10% compensatory increase in the CA1 compared to the hippocampal fissure, was significantly associated with an increased genetic risk of ALS (OR=1.25 [1.02-1.54]) (Figure 3.2d). In women, the global balance included the fimbria, HATA and presubiculum (Figure 3.2b). Results showed that a 10% compensatory increase in the fimbria, with respect to the HATA and

presubiculum, was significantly related to an increased genetic risk of ALS (OR=1.18 [1.02-1.37]) (Figure 3.2d). No significant results were found in men (Figure 3.2d).

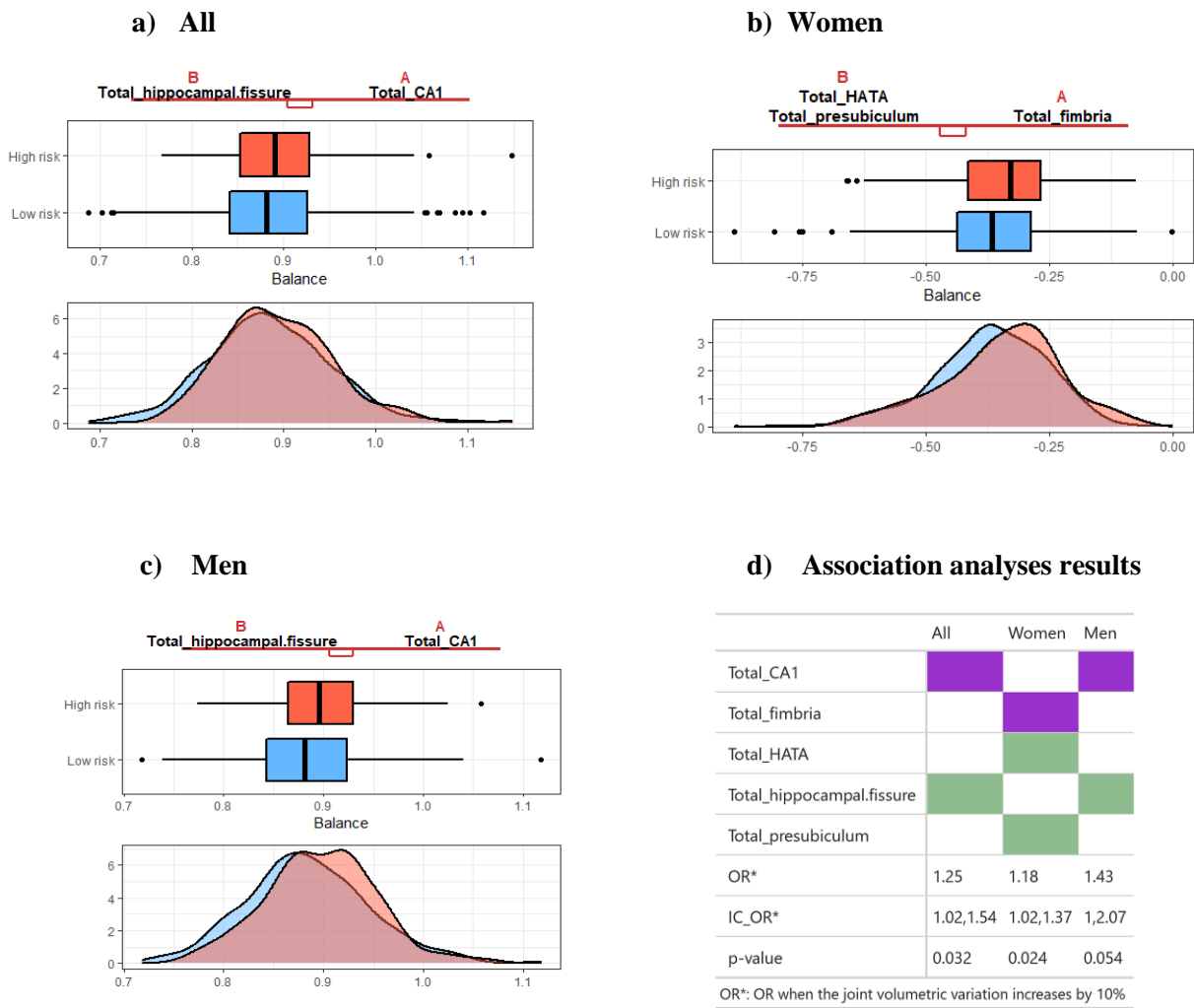


Figure 3.2. Figure 3.2a, 3.2b and 3.2c display hippocampal subfields in group A and B that define the global balance. Boxplots represent the distribution of the balance values for individuals at high (red) and low (blue) genetic risk of ALS, when the sample is defined by all the individuals (Figure 3.2a), women (Figure 3.2b), and men (Figure 3.2c). The density plot is described below. Figure 3.2d the association analyses results, which include the balance as explanatory variable.

3.2.3 Genetic risk of Progressive Supranuclear Palsy and Hippocampal subfields modulation

When the composition was defined by substructures within the hippocampal region, the global balance for the whole sample (Figure 3.3a), as well as for women (Figure 3.3b), was defined by the subiculum and the parasubiculum. Results showed that a 10% compensatory increase in the subiculum compared to the parasubiculum, was significantly associated with an increased genetic risk of PSP (whole sample; OR=1.38 [1.15-1.65], women; OR=1.34 [1.06-1.7]) (Figure 3.3d). In men, the global balance was defined by the molecular layer and the parasubiculum (Figure 3.3c). Results showed that a 10% compensatory increase in the molecular layer, compared to the

parasubiculum, was significantly associated with an increased genetic risk of PSP (OR=1.46 [1.11-1.95]) (Figure 3.4c).

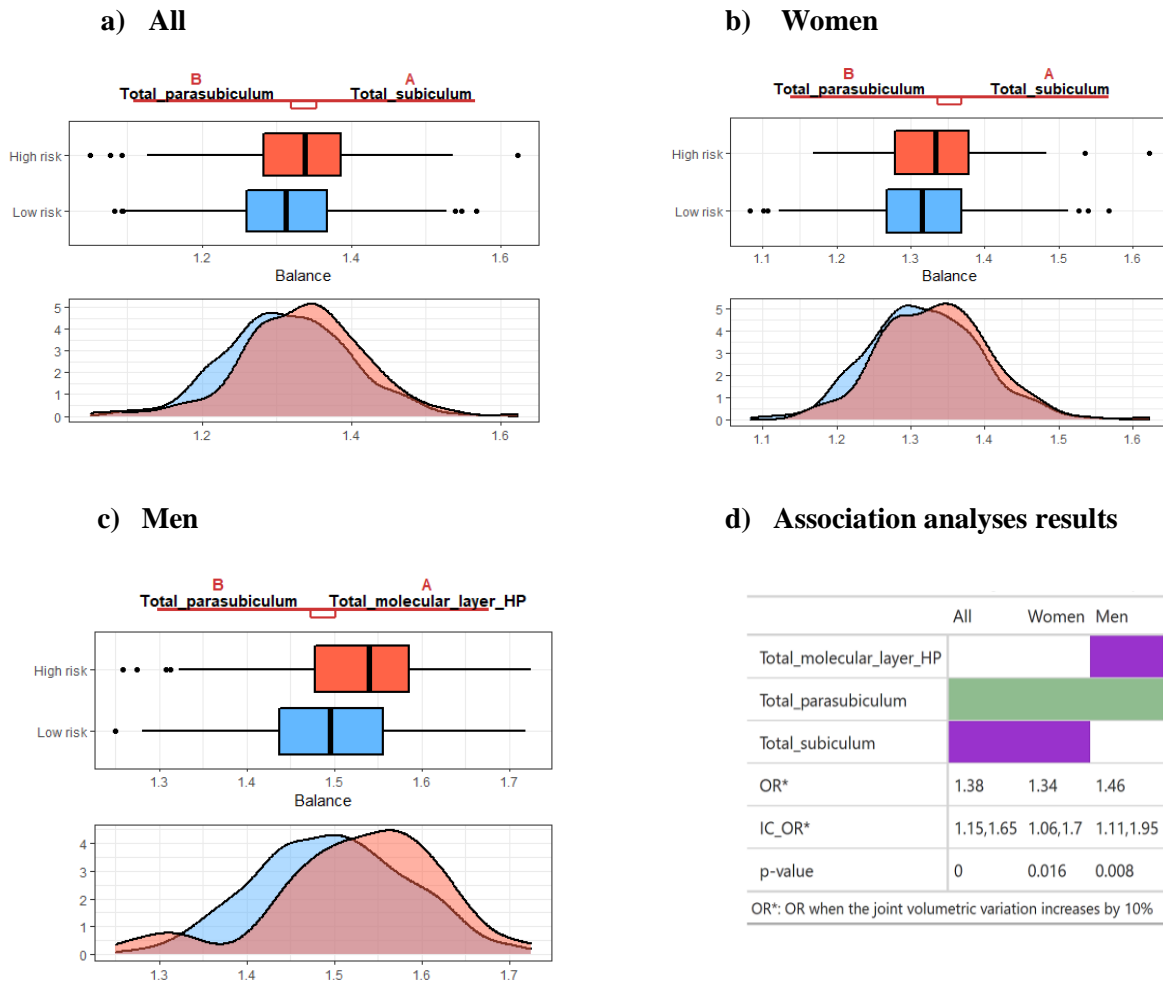


Figure 3.3. Figure 3.3a, 3.3b and 3.3c display hippocampal subfields in group A and B that define the global balance. Boxplots represent the distribution of the balance values for individuals at high (red) and low (blue) genetic risk of ALS, when the sample is defined by all the individuals (Figure 3.3a), women (Figure 3.3b), and men (Figure 3.3c). The density plot is described below. Figure 3.3d the association analyses results, which include the balance as explanatory variable.

4 Discussion and Conclusions

To our knowledge, this is the first study analysing the relationship between individuals at high genetic risk of AD, ALS, and PSP and the joint volumetric variation of hippocampal subfields in cognitively healthy individuals.

We found that a compensatory increase in the average volume of the hippocampal fissure, CA4 and CA3 compared to CA1 and hippocampal tail was associated with an increased genetic risk of AD. Although this is the first study analysing the joint volumetric variation of hippocampal

subfields in cognitively healthy individuals at high genetic risk of AD, there is some previous evidence of the individual affectation of genetic risk of AD and hippocampal subfields. For instance, a recent study performed on 17,161 UK Biobank participants showed that higher PRS-AD was individually associated with lower volumes in the bilateral whole hippocampus, HATA, hippocampal tail, right subiculum, left CA1, CA4, molecular layer, and DG (Foo et al., 2021). Murray et al., 2021 also observed associations between PRS-AD and whole hippocampal/amygdala volumes, as well as CA1 and fissure. Although we cannot directly compare the results obtained in our work with those reported in the univariate studies, the coincidence in some significant subfields (i.e. CA1, hippocampal tail, CA4) could reinforce the joint analysis of these structures.

Moreover, we found that a compensatory increase in CA1 volume compared to hippocampal fissure was associated with an increased genetic risk of ALS. Although implications of CA1 and ALS have been previously described (Machts et al., 2018), few studies have addressed the relationship between hippocampal subfields and ALS, providing inconsistent results. For instance, Christidi et al., (2019) showed that HATA and CA2/3 were the most affected subfields in ALS. Furthermore, no previous studies have been found evaluating genetic factors associated with ALS and the hippocampal region. This is in line with what is expected as it is the motor regions of the brain and not the hippocampus that tend to have a more specific involvement in ALS (Agosta et al., 2018). However, our results could suggest the involvement of the motor neuron system and the joint volumetric variation of hippocampal substructures in ALS, which is relatively in line with previous reported studies (Toyoshima et al., 2003; Anderson et al., 1995; Gómez-Pinedo et al., 2019).

Finally, we found that a compensatory increase in the subiculum volume compared to the parasubiculum was associated with an increased genetic risk of PSP. Although PSP is mainly characterized by the affectation of the basal ganglia and midbrain regions (Mimudo and Yoshira 2019), there is also scientific evidence of hippocampal degeneration in PSP individuals, affecting CA3 and subiculum subfields (Armstrong et al., 2015; Maurer et al., 2017). We did not find additional studies relating genetics of PSP and the hippocampal subregions. Nevertheless, atrophy of hippocampal subicular structures (i.e. subiculum, parasubiculum...) has been shown as a target predictor of cognitive decline and/or dementia progression as in Parkinson's disease, which is closely related to PSP (Low et al., 2019, Foo et al., 2017, Uribe et al., 2018). Moreover, specific hippocampal subfields combinations were also significantly and differentially associated with a higher genetic risk of each neurodegenerative condition according to the analysed hippocampal hemisphere.

Regarding sex-stratified models, we also showed differential significant volumetric variations and combinations of hippocampal subfields associated with high genetic risk of AD, ALS and PSP. However, although there is a high interest in analysing sex-differences in neurological diseases, there are few IG studies assessing these differences on hippocampal subfields. Moreover, existing studies were mostly focused on the global hippocampal volume or were based on the univariate analysis of hippocampal subfields (Hibar et al., 2018; Foley et al., 2017; Foo et al., 2021).

The application of the *Selbal* algorithm allowed us to jointly modulate multiple hippocampal subregions to discern joint volumetric compensatory relationships of these volumes in individuals at high genetic risk of AD, ALS, and PSP. This makes a clear advantage of *Selbal* over univariate brain intermediate studies, not only because of its higher statistical power but also for its easy application and biological plausibility. Another advantage of the *Selbal* approach is its robust algorithm based on the log-ratio between components. The log-ratio function avoids spurious correlations between components as it provides the same results independently of the total brain volume. Moreover, it allows working with data in which the information does not depend on the particular units in which it is expressed, as well as maintains the information unaltered even if the composition suffers a permutation of the parts.

However, the application of *Selbal* in this study also presented some limitations. When assessing the robustness of the global balance for each condition, we found that the most robust results were found for PSP, in which the global balance in all the three sample's conditions (whole sample, women and men), corresponded to the balance most frequently selected in the CV procedure, with at least 50% of times of appearance. Moreover, all components that defined the global balance were also the most frequently selected in the CV, more than 50% of times. In addition, results were not robust in specific situations for AD and ALS, where the global balance did not correspond to the balance most frequently selected in the CV or it included components with a low rate of appearance in the CV procedure. Thus, results showed that the global balance may not be the optimal in some cases. This was also stated as a limitation by the authors of the algorithm (Rivera-Pinto et al., 2018). However, although in some scenarios the component selection may not be optimal, the algorithm does guarantee a robust selection of the components providing additional biological sense in the study.

Therefore, for further studies in the field of IG, it would be interesting to apply this method to improve the understanding of brain substructures' volumetric changes associated with a high genetic risk of specific neurological disorders. Working with compositional data in IG studies allows minimising the heterogeneity in the MRI data that can occur because of the procedure and the technology applied to obtain all the specific brain structures measurements. Working with

compositions removes the effect of the total volume measurement, which can vary in minimum values across different batches, but may have an effect in further association analyses. Thus, cancelling this total effect and working with ratios between components, *Selbal* not only offers a solution to address this issue but also allows assessing the joint modulation of multiple brain subregions and its association with specific phenotypes.

To conclude, this work provides a new and innovative perspective for IG studies with the aim of improving our understanding of the effects that genetic predisposition to neurodegenerative disorders has on brain structure modulation.

5 References

1. Agosta, F., Spinelli, E. G., & Filippi, M. (2018). Neuroimaging in amyotrophic lateral sclerosis: current and emerging uses. *Expert Review of Neurotherapeutics*, 18(5), 395–406. <https://doi.org/10.1080/14737175.2018.1463160>
2. Anderson, V. E. R., Cairns, N. J., & Leigh, P. N. (1995). Involvement of the amygdala, dentate and hippocampus in motor neuron disease. *Journal of the Neurological Sciences*, 129(SUPPL.), 75–78. [https://doi.org/10.1016/0022-510X\(95\)00069-E](https://doi.org/10.1016/0022-510X(95)00069-E)
3. Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature protocols*, 5(9), 1564–1573. <https://doi.org/10.1038/nprot.2010.116>
4. Armstrong, R. A., & Cairns, N. J. (2013). Spatial patterns of the tau pathology in progressive supranuclear palsy. *Neurological Sciences*, 34(3), 337–344. <https://doi.org/10.1007/s10072-012-1006-0>
5. Armstrong, R. A., & Cairns, N. J. (2015). Comparative quantitative study of ‘signature’ pathological lesions in the hippocampus and adjacent gyri of 12 neurodegenerative disorders. *Journal of Neural Transmission*, 122(10), 1355–1367. <https://doi.org/10.1007/s00702-015-1402-8>
6. Blauwendraat, C., Faghri, F., Pihlstrom, L., Geiger, J. T., Elbaz, A., Lesage, S., Corvol, J. C., May, P., Nicolas, A., Abramzon, Y., Murphy, N. A., Gibbs, J. R., Ryten, M., Ferrari, R., Bras, J., Guerreiro, R., Williams, J., Sims, R., Lubbe, S., Hernandez, D. G., ... Scholz, S. W. (2017). NeuroChip, an updated version of the NeuroX genotyping platform to rapidly screen for variants associated with neurological diseases. *Neurobiology of aging*, 57, 247.e9–247.e13. <https://doi.org/10.1016/j.neurobiolaging.2017.05.009>
7. Choi, S. W., & O’Reilly, P. F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience*, 8(7), 1–6. <https://doi.org/10.1093/gigascience/giz082>
8. Christidi, F., Karavasilis, E., Rentzos, M., Velonakis, G., Zouvelou, V., Xirou, S., Argyropoulos, G., Papatriantafyllou, I., Pantolewn, V., Ferentinos, P., Kelekis, N., Seimenis, I., Evdokimidis, I., & Bede, P. (2019). Hippocampal pathology in amyotrophic lateral sclerosis: selective vulnerability of subfields and their associated projections. *Neurobiology of Aging*, 84, 178–188. <https://doi.org/10.1016/j.neurobiolaging.2019.07.019>
9. Dima, D., & Breen, G. (2015). Polygenic risk scores in imaging genetics: Usefulness and applications. *Journal of Psychopharmacology*, 29(8), 867–871. <https://doi.org/10.1177/0269881115584470>
10. Elliott, L. T., Sharp, K., Alfaro-Almagro, F., Shi, S., Miller, K. L., Douaud, G., Marchini, J., & Smith, S. M. (2018). Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature*, 562(7726), 210–216. <https://doi.org/10.1038/s41586-018-0571-7>

11. Elman, J. A., Panizzon, M. S., Gillespie, N. A., Hagler, D. J., Fennema-Notestine, C., Eyler, L. T., McEvoy, L. K., Neale, M. C., Lyons, M. J., Franz, C. E., Dale, A. M., & Kremen, W. S. (2019). Genetic architecture of hippocampal subfields on standard resolution MRI: How the parts relate to the whole. *Human Brain Mapping*, *40*(5), 1528–1540. <https://doi.org/10.1002/hbm.24464>
12. Evans, T. E., Adams, H. H. H., Licher, S., Wolters, F. J., van der Lugt, A., Ikram, M. K., O’Sullivan, M. J., Vernooij, M. W., & Ikram, M. A. (2018). Subregional volumes of the hippocampus in relation to cognitive function and risk of dementia. *NeuroImage*, *178*(March), 129–135. <https://doi.org/10.1016/j.neuroimage.2018.05.041>
13. Flores, Robin & La Joie, Renaud & Chételat, Gaël. (2015). Structural imaging of hippocampal subfields in healthy aging and Alzheimer's disease. *Neuroscience*. 309. [10.1016/j.neuroscience.2015.08.033](https://doi.org/10.1016/j.neuroscience.2015.08.033).
14. Foley, S. F., Tansey, K. E., Caseras, X., Lancaster, T., Bracht, T., Parker, G., Hall, J., Williams, J., & Linden, D. E. (2017). Multimodal Brain Imaging Reveals Structural Differences in Alzheimer's Disease Polygenic Risk Carriers: A Study in Healthy Young Adults. *Biological psychiatry*, *81*(2), 154–161. <https://doi.org/10.1016/j.biopsych.2016.02.033>.
15. Foo, H., Mak, E., Chander, R. J., Ng, A., Au, W. L., Sitoh, Y. Y., Tan, L. C. S., & Kandiah, N. (2016). Associations of hippocampal subfields in the progression of cognitive decline related to Parkinson’s disease. *NeuroImage: Clinical*, *14*, 37–42. <https://doi.org/10.1016/j.nicl.2016.12.008>
16. Foo, H., Thalamuthu, A., Jiang, J., Koch, F., Mather, K. A., Wen, W., & Sachdev, P. S. (2021). Associations between Alzheimer’s disease polygenic risk scores and hippocampal subfield volumes in 17,161 UK Biobank participants. *Neurobiology of Aging*, *98*, 108–115. <https://doi.org/10.1016/j.neurobiolaging.2020.11.002>
17. Hashimoto, R., Ohi, K., Yamamori, H., Yasuda, Y., Fujimoto, M., Umeda-Yano, S., Watanabe, Y., Fukunaga, M., & Takeda, M. (2015). Imaging Genetics and Psychiatric Disorders. *Current Molecular Medicine*, *15*(2), 168–175. <https://doi.org/10.2174/1566524015666150303104159>
18. Hibar, D. P., Adams, H. H. H., Jahanshad, N., Chauhan, G., Stein, J. L., Hofer, E., Renteria, M. E., Bis, J. C., Arias-Vasquez, A., Ikram, M. K., Desrivières, S., Vernooij, M. W., Abramovic, L., Alhusaini, S., Amin, N., Andersson, M., Arfanakis, K., Aribisala, B. S., Armstrong, N. J., ... Ikram, M. A. (2017). Novel genetic loci associated with hippocampal volume. *Nature Communications*, *8*. <https://doi.org/10.1038/ncomms13624>
19. Iglesias, J. E., Augustinack, J. C., Nguyen, K., Player, C. M., Player, A., Wright, M., Roy, N., Frosch, M. P., McKee, A. C., Wald, L. L., Fischl, B., & Van Leemput, K. (2015). A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI:

- Application to adaptive segmentation of in vivo MRI. *NeuroImage*, 115, 117–137. <https://doi.org/10.1016/j.neuroimage.2015.04.042>
20. Kline, R.B. (2011). Principles and practice of structural equation modeling (5th ed., pp. 3-427). New York: The Guilford Press.
 21. Liu, J., & Calhoun, V. D. (2014). A review of multivariate analyses in imaging genetics. *Frontiers in Neuroinformatics*, 8(MAR), 1–11. <https://doi.org/10.3389/fninf.2014.00029>
 22. Low, A., Foo, H., Yong, T. T., Tan, L. C. S., & Kandiah, N. (2019). Hippocampal subfield atrophy of CA1 and subicular structures predict progression to dementia in idiopathic Parkinson's disease. *Journal of Neurology, Neurosurgery and Psychiatry*, 90(6), 681–687. <https://doi.org/10.1136/jnnp-2018-319592>
 23. Luz Calle, M. (2019). Statistical analysis of metagenomics data. *Genomics and Informatics*, 17(1). <https://doi.org/10.5808/GI.2019.17.1.e6>
 24. Machts, J., Vielhaber, S., Kollwe, K., Petri, S., Kaufmann, J., & Schoenfeld, M. A. (2018). Global hippocampal volume reductions and local CA1 shape deformations in Amyotrophic Lateral Sclerosis. *Frontiers in Neurology*, 9(JUL), 1–9. <https://doi.org/10.3389/fneur.2018.00565>
 25. Matoba, N, Love, MI, Stein, JL. Evaluating brain structure traits as endophenotypes using polygenicity and discoverability. *Hum Brain Mapp.* 2020; 1– 12. <https://doi.org/10.1002/hbm.25257>
 26. Maurer, S. V., & Williams, C. L. (2017). The cholinergic system modulates memory and hippocampal plasticity via its interactions with non-neuronal cells. *Frontiers in Immunology*, 8(NOV). <https://doi.org/10.3389/fimmu.2017.01489>
 27. Mimuro, M., & Yoshida, M. (2020). Chameleons and mimics: Progressive supranuclear palsy and corticobasal degeneration. *Neuropathology*, 40(1), 57–67. <https://doi.org/10.1111/neup.12590>
 28. Molinuevo, J., et al. The ALFA project: A research platform to identify early pathophysiological features of Alzheimer's disease. *Alzheimer's & dementia*. 2(2), 82–92. (2016).
 29. Murray, A. N., Chandler, H. L., & Lancaster, T. M. (2021). Multimodal hippocampal and amygdala subfield volumetry in polygenic risk for Alzheimer's disease. *Neurobiology of Aging*, 98, 33–41. <https://doi.org/10.1016/j.neurobiolaging.2020.08.022>
 30. Nathoo, F. S., Kong, L., & Zhu, H. (2019). A Review of Statistical Methods in Imaging Genetics. *The Canadian journal of statistics = Revue canadienne de statistique*, 47(1), 108–131. <https://doi.org/10.1002/cjs.11487>
 31. Rivera-Pinto, J., Egozcue, J. J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., & Calle, M. L. (2018). Balances: a New Perspective for Microbiome Analysis. *MSystems*, 3(4), 1–12. <https://doi.org/10.1128/msystems.00053-18>

32. Smith, S. M., Douaud, G., Chen, W., Hanayik, T., Alfaro-Almagro, F., Sharp, K., & Elliott, L. T. (2021). An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature Neuroscience*, 24(5), 737–745. <https://doi.org/10.1038/s41593-021-00826-4>
33. Susin, A., Wang, Y., Lê Cao, K.-A., & Calle, M. L. (2020). Variable selection in microbiome compositional data analysis. *NAR Genomics and Bioinformatics*, 2(2), 5–7. <https://doi.org/10.1093/nargab/lqaa029>
34. Toyoshima, Y., Piao, Y. S., Tan, C. F., Morita, M., Tanaka, M., Oyanagi, K., Okamoto, K., & Takahashi, H. (2003). Pathological involvement of the motor neuron system and hippocampal formation in motor neuron disease-inclusion dementia. *Acta Neuropathologica*, 106(1), 50–56. <https://doi.org/10.1007/s00401-003-0696-z>
35. Uribe, C., Segura, B., Baggio, H. C., Campabadal, A., Abos, A., Compta, Y., Marti, M. J., Valldeoriola, F., Bargallo, N., & Junque, C. (2018). Differential progression of regional hippocampal atrophy in aging and Parkinson’s disease. *Frontiers in Aging Neuroscience*, 10(OCT), 1–9. <https://doi.org/10.3389/fnagi.2018.00325>
36. van der Meer, D., Rokicki, J., Kaufmann, T., Córdova-Palomera, A., Moberget, T., Alnæs, D., Bettella, F., Frei, O., Doan, N. T., Sønderyby, I. E., Smeland, O. B., Agartz, I., Bertolino, A., Bralten, J., Brandt, C. L., Buitelaar, J. K., Djurovic, S., van Donkelaar, M., Dørum, E. S., ... Westlye, L. T. (2020). Brain scans from 21,297 individuals reveal the genetic architecture of hippocampal subfield volumes. *Molecular Psychiatry*, 25(11), 3053–3065. <https://doi.org/10.1038/s41380-018-0262-7>
37. Vilor-Tejedor, N., Alemany, S., Cáceres, A., Bustamante, M., Pujol, J., Sunyer, J., & González, J. R. (2018). Strategies for integrated analysis in imaging genetics studies. *Neuroscience and biobehavioral reviews*, 93, 57–70. <https://doi.org/10.1016/j.neubiorev.2018.06.013>
38. Vilor-Tejedor, N., Evans, T. E., Adams, H. H., González-de-Echávarri, J. M., Molinuevo, J. L., Guigo, R., Gispert, J. D., & Operto, G. (2021). Genetic Influences on Hippocampal Subfields. *Neurology Genetics*, 7(3), e591. <https://doi.org/10.1212/nxg.0000000000000591>
39. Vilor-Tejedor, N., Ikram, M. A., Roshchupkin, G. V., Cáceres, A., Alemany, S., Vernooij, M. W., Niessen, W. J., van Duijn, C. M., Sunyer, J., Adams, H. H., & González, J. R. (2019). Independent Multiple Factor Association Analysis for Multiblock Data in Imaging Genetics. *Neuroinformatics*, 17(4), 583–592. <https://doi.org/10.1007/s12021-019-09416-z>
40. Yao, X., Cong, S., Yan, J., Risacher, S. L., Saykin, A. J., Moore, J. H., & Shen, L. (2020). Regional imaging genetic enrichment analysis. *Bioinformatics*, 36(8), 2554–2560. <https://doi.org/10.1093/bioinformatics/btz948>

6 Appendix

6.1 Supplemental Tables

Table A1. Supplementary table: GWAS summary statistics of the phenotypes whose PRS have been computed and included in the study.

Disease/Condition Description	Sample Size	N Cases	N Controls	Ethnicity	Genomic Annotation	Base GWAS (First Author)	Link Manuscript	Access GWAS-Summary (LINK)
Alzheimer's Disease	455.258	71.880	383.378	Caucasian	GRCh37/hg19	Jansen et al., 2020	https://pubmed.ncbi.nlm.nih.gov/32029921/	https://ctg.cncr.nl/software/summary_statistics
** Alzheimer's Disease		** Alzheimer's Disease, removing the APOE region (chr 19: 45,409,011-45,412,650) for the PRSs computation						
Parkinson's Disease	1.437.688	37.688	1.400.000	Caucasian	GRCh37/hg19	Nalls et al., 2019	https://pubmed.ncbi.nlm.nih.gov/31701892/	https://drive.google.com/file/d/1FZ9UL99LAqWnyNBxxlx6qOUlfAnubIN/view?usp=sharing
Frontotemporal Dementia	4.131	1.377	2.754	Caucasian	GRCh37/hg19	Ferrari et al., 2014	https://pubmed.ncbi.nlm.nih.gov/24943344/	https://ifgcsite.wordpress.com/data-access/
** Frontotemporal Dementia		** Frontotemporal Dementia Meta-Analysis (FTD subtypes): 2154 cases and 4308 controls, N=6462						
Amyotrophic Lateral Sclerosis	36.052	12.577	23.475	Caucasian	GRCh37/hg19	van Rheenen et al., 2016	https://pubmed.ncbi.nlm.nih.gov/27455348/	http://databrowser.projectmine.com/
Progressive Supranuclear Palsy	12.308	1.646	10.662	Caucasian	GRCh37/hg19	Chen et al., 2018	https://pubmed.ncbi.nlm.nih.gov/30089514/	https://www.niaqads.org/

Table A2. Descriptive analysis of hippocampal subfields volumes.

Hemisphere	Mean	Max	Min	SD
Total				
CA1	1187.23	1819.86	872.20	127.81
CA3	359.79	525.08	240.63	43.63
CA4	452.33	648.88	298.90	44.76
Fimbria	171.32	324.50	64.23	30.68
GC-ML-DG	532.26	748.38	350.51	52.82
HATA	116.04	179.12	71.96	14.29
Hippocampal-fissure	339.98	514.15	213.66	42.56
Hippocampal_tail	1061.23	1506.08	725.59	123.81
Molecular_layer_HP	1053.14	1506.60	767.00	102.08
Parasubiculum	124.52	193.77	76.53	17.20
Presubiculum	599.26	823.97	411.87	68.47
Subiculum	800.85	1140.42	594.21	90.02
Left				
CA1	580.83	884.52	408.91	64.66
CA3	174.84	278.29	106.91	22.68
CA4	225.30	319.11	139.15	24.05
fimbria	83.56	157.92	32.90	16.86
GC-ML-DG	264.44	380.75	162.51	28.23
HATA	56.55	90.68	31.23	7.98
hippocampal-fissure	169.53	250.72	98.99	24.59
Hippocampal_tail	532.61	782.96	304.41	66.31
Molecular_layer_HP	523.25	736.22	372.33	52.86
Parasubiculum	63.20	97.98	35.27	9.85
Presubiculum	309.98	422.07	209.38	37.28
Subiculum	401.77	590.77	292.85	48.58
Right				
CA1	606.40	941.80	420.25	70.03
CA3	184.96	300.42	116.83	25.13
CA4	227.03	331.96	156.62	24.00
Fimbria	87.77	176.53	27.21	17.81
GC-ML-DG	267.82	390.45	181.47	28.22
HATA	59.49	88.43	30.02	8.05
Hippocampal-fissure	170.45	284.25	105.55	23.81
Hippocampal_tail	528.62	761.09	333.53	65.57
Molecular_layer_HP	529.89	770.38	382.54	53.08
Parasubiculum	61.32	101.14	27.47	9.48
Presubiculum	289.28	431.49	189.07	35.63
Subiculum	399.08	568.53	282.07	45.85

Table A3. Descriptive analysis of hippocampal subfields volumes by sex.

Hemisphere	Women				Men			
	Mean	Max	Min	SD	Mean	Max	Min	SD
Total								
CA1	1145.99	1468.38	872.20	110.09	1256.13	1819.86	983.01	125.77
CA3	346.83	476.82	240.63	36.41	381.45	525.08	284.56	46.08
CA4	439.18	564.22	298.90	39.78	474.29	648.88	368.84	44.03
Fimbria	164.52	288.74	64.23	27.39	182.69	324.50	87.79	32.50
GC-ML-DG	516.76	655.96	350.51	46.85	558.15	748.38	430.86	52.14
HATA	112.10	150.51	72.09	12.65	122.62	179.12	71.96	14.46
Hippocampal-fissure	328.63	480.42	213.66	40.01	358.94	514.15	257.14	39.88
Hippocampal_tail	1034.89	1488.41	725.59	115.75	1105.24	1506.08	790.00	124.49
Molecular_layer_HP	1021.08	1285.93	767.00	90.43	1106.71	1506.60	878.50	97.97
Parasubiculum	119.45	170.00	76.53	14.45	133.00	193.77	79.96	18.08
Presubiculum	579.74	789.25	411.87	61.20	631.87	823.97	428.67	67.60
Subiculum	772.01	1040.79	594.21	80.59	849.03	1140.42	629.24	84.23
Left								
CA1	559.54	754.27	408.91	55.56	616.39	884.52	459.21	63.21
CA3	169.07	236.59	106.91	19.51	184.46	278.29	120.01	24.31
CA4	219.33	297.22	139.15	22.23	235.27	319.11	163.75	23.70
fimbria	80.21	154.39	35.30	15.13	89.15	157.92	32.90	18.10
GC-ML-DG	257.38	342.55	162.51	26.03	276.23	380.75	189.04	27.84
HATA	54.71	80.39	31.75	7.11	59.63	90.68	31.23	8.40
hippocampal-fissure	163.06	250.72	98.99	23.03	180.35	249.83	117.45	23.32
Hippocampal_tail	522.63	782.96	304.41	63.82	549.28	753.66	342.97	67.11
Molecular_layer_HP	507.09	645.64	372.33	47.29	550.26	736.22	400.57	50.65
Parasubiculum	60.41	93.45	35.27	8.52	67.85	97.98	37.24	10.16
Presubiculum	300.00	405.12	209.38	34.22	326.65	422.07	227.08	36.26
Subiculum	386.92	539.74	292.85	43.68	426.58	590.77	320.77	46.19
Right								
CA1	586.45	766.47	420.25	61.38	639.74	941.80	469.44	70.97
CA3	177.76	263.79	116.83	21.00	196.98	300.42	136.26	26.83
CA4	219.85	288.63	156.62	20.89	239.03	331.96	186.12	24.10
Fimbria	84.31	144.22	27.21	16.38	93.54	176.53	39.67	18.62
GC-ML-DG	259.38	335.29	181.47	24.47	281.92	390.45	220.18	28.46
HATA	57.39	78.33	30.02	7.27	63.00	88.43	33.45	8.07
Hippocampal-fissure	165.57	234.26	105.55	22.82	178.59	284.25	115.74	23.21
Hippocampal_tail	512.26	719.94	333.53	60.05	555.96	761.09	380.46	65.37
Molecular_layer_HP	513.99	652.08	382.54	47.23	556.44	770.38	433.60	51.73
Parasubiculum	59.04	88.00	27.47	8.12	65.15	101.14	40.56	10.33
Presubiculum	279.74	394.54	189.07	31.55	305.22	431.49	196.99	36.37
Subiculum	385.10	519.93	282.07	41.57	422.45	568.53	293.67	43.10

Table A4. Wilcoxon test results, within women, assessing whether mean values of hippocampal subfields volumes were equivalent between hemispheres.

Left hemisphere	Right hemisphere	T	P-value	Mean volume (left)	Mean volume (right)	P.adj
Left_CA1	Right_CA1	167785	0.000	559.539	586.452	0.0000000
Left_CA3	Right_CA3	171261	0.000	169.073	177.756	0.0000000
Left_CA4	Right_CA4	220388	0.566	219.333	219.846	0.6174545
Left_fimbria	Right_fimbria	189733	0.000	80.207	84.312	0.0000000
Left_GC.ML.DG	Right_GC.ML.DG	214047	0.142	257.377	259.382	0.1704000
Left_HATA	Right_HATA	177792	0.000	54.709	57.387	0.0000000
Left_hippocampal.fissure	Right_hippocampal.fissure	211495	0.067	163.061	165.573	0.0893333
Left_Hippocampal_tail	Right_Hippocampal_tail	246223	0.002	522.628	512.259	0.0040000
Left_molecular_layer_HP	Right_molecular_layer_HP	205932	0.009	507.087	513.992	0.0150000
Left_parasubiculum	Right_parasubiculum	242761	0.010	60.411	59.036	0.0150000
Left_presubiculum	Right_presubiculum	298585	0.000	299.997	279.741	0.0000000
Left_subiculum	Right_subiculum	227977	0.619	386.919	385.095	0.6190000

Table A5. Wilcoxon test results, within men, assessing whether mean values of hippocampal subfields volumes were equivalent between hemispheres.

Left hemisphere	Right hemisphere	T	P-value	Mean volume (left)	Mean volume (right)	P.adj
Left_CA1	Right_CA1	65632	0.000	616.394	639.741	0.0000000
Left_CA3	Right_CA3	59125	0.000	184.465	196.985	0.0000000
Left_CA4	Right_CA4	74452	0.070	235.266	239.028	0.1050000
Left_fimbria	Right_fimbria	69368	0.001	89.152	93.535	0.0020000
Left_GC.ML.DG	Right_GC.ML.DG	72274	0.013	276.231	281.921	0.0222857
Left_HATA	Right_HATA	61809	0.000	59.626	62.998	0.0000000
Left_hippocampal.fissure	Right_hippocampal.fissure	83725	0.311	180.346	178.593	0.3110000
Left_Hippocampal_tail	Right_Hippocampal_tail	76813	0.274	549.281	555.957	0.2989091
Left_molecular_layer_HP	Right_molecular_layer_HP	75772	0.158	550.263	556.444	0.2106667
Left_parasubiculum	Right_parasubiculum	92320	0.000	67.852	65.146	0.0000000
Left_presubiculum	Right_presubiculum	107344	0.000	326.651	305.223	0.0000000
Left_subiculum	Right_subiculum	84390	0.224	426.579	422.452	0.2688000

Table A6. Wilcoxon test results, without differentiating by sex, assessing whether mean values of hippocampal subfields volumes were equivalent between hemispheres.

Left hemisphere	Right hemisphere	T	P-value	Mean volumes (left)	Mean volumes (right)	P.adj
Left_CA1	Right_CA1	693032	0.000	580.827	606.404	0.000000
Left_CA3	Right_CA3	707875	0.000	174.836	184.956	0.000000
Left_CA4	Right_CA4	593907	0.154	225.298	227.028	0.184800
Left_fimbria	Right_fimbria	656324	0.000	83.556	87.765	0.000000
Left_GC.ML.DG	Right_GC.ML.DG	608817	0.014	264.436	267.821	0.021000
Left_HATA	Right_HATA	692880	0.000	56.550	59.488	0.000000
Left_hippocampal.fissure	Right_hippocampal.fissure	585807	0.391	169.533	170.448	0.391000
Left_Hippocampal_tail	Right_Hippocampal_tail	549169	0.089	532.607	528.620	0.118667
Left_molecular_layer_HP	Right_molecular_layer_HP	611918	0.007	523.253	529.887	0.012000
Left_parasubiculum	Right_parasubiculum	512047	0.000	63.197	61.324	0.000000
Left_presubiculum	Right_presubiculum	392927	0.000	309.977	289.282	0.000000
Left_subiculum	Right_subiculum	559631	0.332	401.768	399.083	0.3621818

Table A7. Wilcoxon test results, assessing whether mean values of hippocampal subfields volumes were equivalent between men and women.

Region	T	P-value	Mean volume (men)	Mean volume (women)	P.adj
Left_CA1	202310	0	616.394	559.539	0
Left_CA3	183886	0	184.465	169.073	0
Left_CA4	184963	0	235.266	219.333	0
Left_fimbria	175055	0	89.152	80.207	0
Left_GC.ML.DG	185296	0	276.231	257.377	0
Left_HATA	181724	0	59.626	54.709	0
Left_hippocampal.fissure	188785	0	180.346	163.061	0
Left_Hippocampal_tail	165786	0	549.281	522.628	0
Left_molecular_layer_HP	197689	0	550.263	507.087	0
Left_parasubiculum	191467	0	67.852	60.411	0
Left_presubiculum	189659	0	326.651	299.997	0
Left_subiculum	197650	0	426.579	386.919	0
Right_CA1	192149	0	639.741	586.452	0
Right_CA3	189981	0	196.985	177.756	0
Right_CA4	194517	0	239.028	219.846	0
Right_fimbria	173570	0	93.535	84.312	0
Right_GC.ML.DG	194289	0	281.921	259.382	0
Right_HATA	187566	0	62.998	57.387	0
Right_hippocampal.fissure	176919	0	178.593	165.573	0
Right_Hippocampal_tail	185402	0	555.957	512.259	0
Right_molecular_layer_HP	195171	0	556.444	513.992	0
Right_parasubiculum	181563	0	65.146	59.036	0
Right_presubiculum	189261	0	305.223	279.741	0
Right_subiculum	198629	0	422.452	385.095	0
Total_CA1	200554	0	1256.135	1145.992	0
Total_CA3	192244	0	381.449	346.829	0
Total_CA4	193238	0	474.294	439.179	0
Total_fimbria	180769	0	182.687	164.519	0
Total_GC.ML.DG	193197	0	558.152	516.758	0
Total_HATA	190592	0	122.624	112.097	0
Total_hippocampal.fissure	189778	0	358.939	328.634	0
Total_Hippocampal_tail	177773	0	1105.238	1034.887	0
Total_molecular_layer_HP	198618	0	1106.707	1021.080	0
Total_parasubiculum	193094	0	132.998	119.447	0
Total_presubiculum	193419	0	631.874	579.738	0
Total_subiculum	201089	0	849.031	772.014	0

Table A8. Symmetry analysis: skewness and kurtosis indexes for each PRS.

	Skewness	Kurtosis
AD	1.1618955	5.442688
AD_NOAPOE	0.3221473	3.790309
ALS	-0.3546688	1.913275
FTD	-0.7406105	2.553113
META	-0.4049775	2.561518
PKSON	8.2353697	100.743339
PSP	-0.6724381	2.627844

6.2 Supplemental Figures

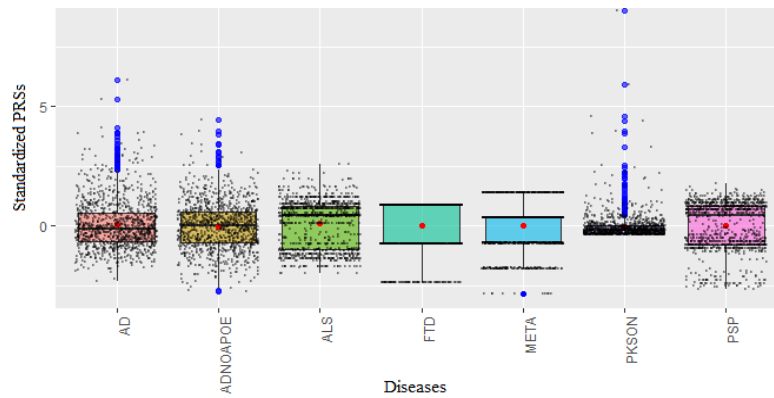


Figure A1. PRSs distribution.

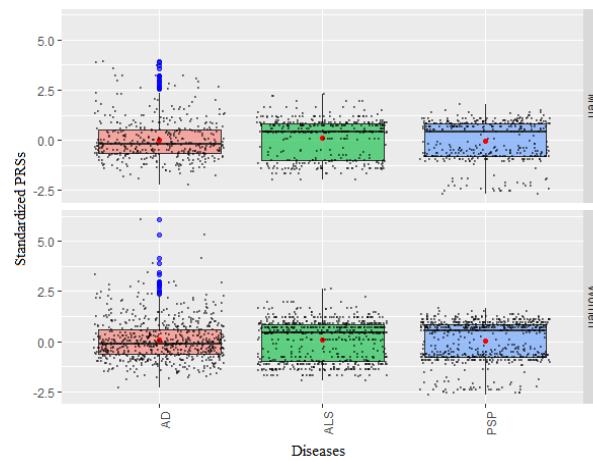


Figure A2. Included PRSs distribution by sex.

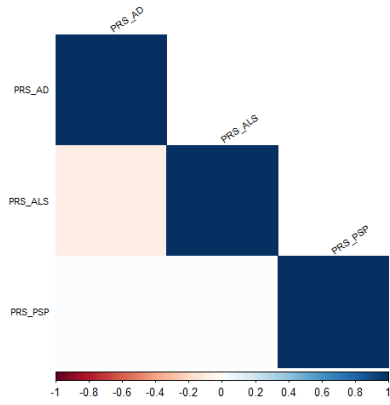


Figure A3. Included PRSs correlation.



Figure A4. PRSs distribution of Alzheimer’s Disease, Amyotrophic Lateral Sclerosis and Progressive Supranuclear Palsy by risk group.

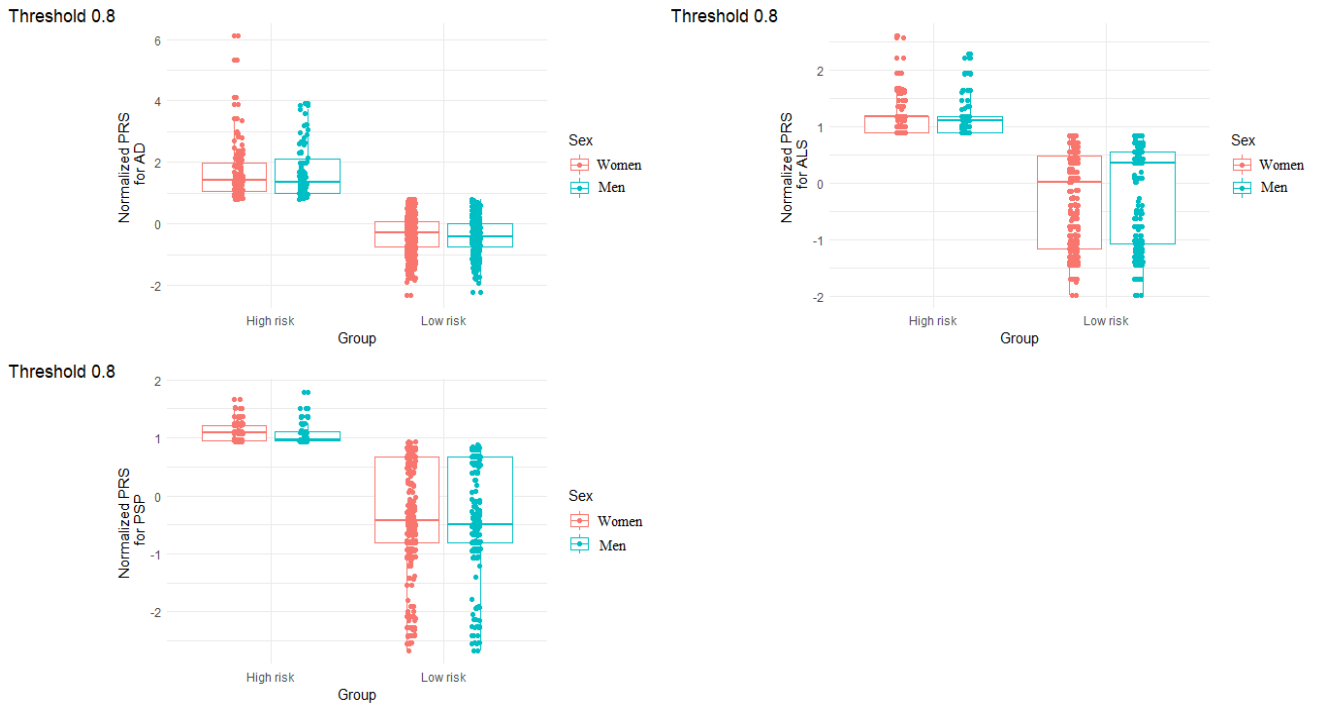


Figure A5. PRSs distribution of Alzheimer’s Disease, Amyotrophic Lateral Sclerosis and Progressive Supranuclear Palsy by risk group and sex.

a) All

	%	Global	BAL 1	BAL 2	BAL 3
Total_CA1	98				
Total_CA3	84				
Total_Hippocampal_tail	74				
Total_CA4	56				
Total_hippocampal.fissure	48				
Total_GC.ML.DG	40				
Total_subiculum	32				
FREQ	-	-	0.12	0.1	0.1

b) Women

	%	Global	BAL 1	BAL 2	BAL 3
Total_Hippocampal_tail	90				
Total_hippocampal.fissure	40				
Total_CA4	22				
Total_CA3	16				
FREQ	-	-	0.32	0.22	0.14

c) Men

	%	Global	BAL 1	BAL 2	BAL 3
Total_CA1	96				
Total_CA3	88				
Total_CA4	8				
Total_GC.ML.DG	4				
Total_molecular_layer_HP	4				
FREQ	-	-	0.88	0.08	0.04

Figure A6. CV results for AD when working with the first composition (C1): most frequent hippocampal subregions and most frequent balances selected in the CV procedure in comparison to the global balance obtained from the whole sample. Red cells indicate selected components in the numerator of the balance (group A) and blue cells indicate selected components in the denominator of the balance (group B). Empty cells indicate components not included. *FREQ*: frequency.

a) All

	%	Global	BAL 1	BAL 2	BAL 3
Total_CA1	58				
Total_hippocampal.fissure	48				
Total_fimbria	34				
Total_HATA	16				
Total_presubiculum	12				
FREQ	-	-	0.42	0.12	0.1

b) Women

	%	Global	BAL 1	BAL 2	BAL 3
Total_CA1	72				
Total_fimbria	68				
Total_CA4	52				
Total_presubiculum	30				
Total_molecular_layer_HP	20				
Total_GC.ML.DG	16				
Total_HATA	4				
FREQ	-	-	0.28	0.14	0.1

c) Men

	%	Global	BAL 1	BAL 2	BAL 3
Total_hippocampal.fissure	64				
Total_CA1	34				
Total_GC.ML.DG	26				
Total_molecular_layer_HP	22				
Total_HATA	20				
Total_subiculum	14				
FREQ	-	-	0.32	0.14	0.1

Figure A7. CV results for ALS when working with the first composition (C1): most frequent hippocampal subregions and most frequent balances selected in the CV procedure in comparison to the global balance obtained from the whole sample. Red cells indicate selected components in the numerator of the balance (group A) and blue cells indicate selected components in the denominator of the balance (group B). Empty cells indicate components not included. *FREQ*: frequency.

a) All

	%	Global	BAL 1	BAL 2	BAL 3
Total_parasubiculum	92				
Total_subiculum	72				
Total_molecular_layer_HP	16				
Total_CA1	10				
FREQ	-	-	0.64	0.16	0.1

b) Women

	%	Global	BAL 1	BAL 2	BAL 3
Total_subiculum	66				
Total_parasubiculum	52				
Total_presubiculum	38				
Total_CA4	14				
FREQ	-	-	0.5	0.16	0.12

c) Men

	%	Global	BAL 1	BAL 2	BAL 3
Total_parasubiculum	90				
Total_molecular_layer_HP	50				
Total_subiculum	44				
Total_presubiculum	4				
FREQ	-	-	0.5	0.4	0.04

Figure A8. CV results for PSP when working with the first composition (C1): most frequent hippocampal subregions and most frequent balances selected in the CV procedure in comparison to the global balance obtained from the whole sample. Red cells indicate selected components in the numerator of the balance (group A) and blue cells indicate selected components in the denominator of the balance (group B). Empty cells indicate components not included. *FREQ*: frequency.

a) AD				b) ALS				c) PSP			
	All	Women	Men		All	Women	Men		All	Women	Men
Right_CA1				Right_CA1				Right_molecular_layer_HP			
Right_CA4				Right_fimbria				Right_parasubiculum			
Right_Hippocampal_tail				Right_GC.ML.DG				Right_presubiculum			
OR*	1.55	1.43	2.2	Right_HATA				Right_subiculum			
IC_OR*	1.16,2.08	1.12,1.85	1.37,3.6	Right_hippocampal.fissure				OR*	1.3	1.38	1.33
p-value	0.003	0.005	0.001	Right_Hippocampal_tail				IC_OR*	1.13,1.51	1.09,1.76	1.06,1.68
OR*: OR when the joint volumetric variation increases by 10%				Right_subiculum				p-value	0	0.008	0.014
				OR*	1.12	1.27	1.26	OR*: OR when the joint volumetric variation increases by 10%			
				IC_OR*	0.89,1.42	1.07,1.5	0.97,1.64				
				p-value	0.333	0.007	0.084				
				OR*: OR when the joint volumetric variation increases by 10%							

Figure A9. *Selbal* algorithm results for the second composition (C2): right hippocampal volume. Results of the logistic regression models, both general and stratifying by sex. The global balance is defined by the ratio between volumes in group A (dark violet; numerator) and volumes in group B (green; denominator).

a) AD				b) ALS				c) PSP			
	All	Women	Men		All	Women	Men		All	Women	Men
Left_CA1				Left_CA1				Left_CA3			
Left_CA3				Left_CA4				Left_CA4			
Left_hippocampal.fissure				Left_HATA				Left_hippocampal.fissure			
Left_Hippocampal_tail				Left_hippocampal.fissure				Left_molecular_layer_HP			
OR*	1.24	1.2	1.58	OR*	1.25	1.47	1.36	Left_parasubiculum			
IC_OR*	1.07,1.43	1.02,1.42	1.12,2.25	IC_OR*	1.06,1.48	1.09,2	0.98,1.88	Left_presubiculum			
p-value	0.004	0.031	0.01	p-value	0.008	0.012	0.066	Left_subiculum			
OR*: OR when the joint volumetric variation increases by 10%				OR*: OR when the joint volumetric variation increases by 10%				OR*: OR when the joint volumetric variation increases by 10%			
								OR*	1.27	1.27	1.48
								IC_OR*	1.08,1.49	1.02,1.59	1.12,1.98
								p-value	0.004	0.036	0.007

Figure A10. *Selbal* algorithm results for the third composition (C3): left hippocampal volume. Results of the logistic regression models, both general and stratifying by sex. The global balance is defined by the ratio between volumes in group A (dark violet; numerator) and volumes in group B (green; denominator).

6.3 Supplemental formulas and examples

How the effect of the total measurement is cancelled when working with ratios?

Example when working with a composition defined by 3 components.

$$\text{Composition: } C = (x_1, x_2, x_3)$$

$$\text{Total measurement: } x_1 + x_2 + x_3 = \sum_{j=1}^3 x_j$$

$$\text{Proportions: } \frac{x_1}{\sum_1^3 x_j}, \frac{x_2}{\sum_1^3 x_j}, \frac{x_3}{\sum_1^3 x_j}$$

$$\text{Logarithms: } \log\left(\frac{x_1}{\sum_1^3 x_j}\right), \log\left(\frac{x_2}{\sum_1^3 x_j}\right), \log\left(\frac{x_3}{\sum_1^3 x_j}\right)$$

$$\text{Log - ratio between components } x_1 - x_3: \frac{\log\left(\frac{x_1}{\sum_1^3 x_j}\right)}{\log\left(\frac{x_2}{\sum_1^3 x_j}\right)}$$

$$= \log(x_1) - \log(\sum_1^3 x_j) - (\log(x_2) - \log(\sum_1^3 x_j))$$

$$= \log(x_1) - \log(\sum_1^3 x_j) - \log(x_2) + \log(\sum_1^3 x_j)$$

$$= \log(x_1) - \log(x_2)$$

$$= \log\left(\frac{x_1}{x_2}\right)$$

Selbal results: define the OR for a n-units change when the explanatory variable is in the logarithmic scale.

Logistic model: $\text{logit}(Y) = \beta_0 + \beta \cdot x$, where $x = K \cdot \ln(z_1) - \ln(z_2) = K \cdot \ln\left(\frac{z_1}{z_2}\right)$ (balance).

For one-unit increase of x, the OR is defined by the $\exp(\beta)$.

For n-units increase of x, the OR is defined by the $\exp(\beta \cdot n)$

A n-units increase of x means that $x + n = K \cdot (\ln(z_1) - \ln(z_2) + n) = K \cdot \left(\ln\left(\frac{z_1}{z_2}\right) + n\right)$

Defining n in the logarithmic scale, we can see that $x + n = K \cdot (\ln(z_1) - \ln(z_2) + \ln(e^n)) = K \cdot$

$$\left(\ln\left(\frac{z_1}{z_2}\right) + \ln(e^n)\right) = K \cdot \left(\ln\left(\frac{z_1}{z_2}\right) \cdot \ln(e^n)\right) = K \cdot \left(\ln\left(\frac{z_1 \cdot e^n}{z_2}\right)\right).$$

Example when we want to observe a 10% increase.

- 10% of x = 0.1 · x
- a 10% increase = x + 0.1 · x = 1.1
- $e^n = 1.1, n = \ln(1.1)$
- $OR = \exp(n \cdot \beta) = \exp(\ln(1.1) \cdot \beta)$

6.4 Bioinformatic pipeline (continues on next page)

In Figure A11, we can observe the workflow needed to compute PRSs.

PRSs computation: PRSice version 1.2

Files needed to compute PRSs	<p>PRSice v1.2</p> <ul style="list-style-type: none"> Script PRSice_v2.R: a wrapper for the PRSice executable and for plotting PRSice_linux <p>Execution scripts (.sh)</p>	<p>Target data: Raw genotype data of target phenotype</p> <p>Files:</p> <div style="border: 1px solid black; padding: 2px; display: flex; justify-content: space-around;"> .BIM .BED .FAM </div>																														
Execution code: main steps and parameters	<pre> Rscript /nfs/users2/rg/nvlortejedor/PRS_Project/scripts/PRSice_v2.R --dir . \ --prsice /nfs/users2/rg/nvlortejedor/PRS_Project/scripts/PRSice_linux \ --base base_data.txt \ --target /nfs/users2/rg/nvlortejedor/ALFA-GWAS/HRC_Imputation/ALFABatch1_N=918/ALFAB1_impQC.rs \ --thread 16 \ --clump-r2 0.1 \ --fastscore T \ --no-regress \ --bar-levels 0.0000005,0.000005,0.0001,0.001,0.01,0.05,0.1,0.2,0.5,1 \ --perm 10000 \ --all-score \ --out base_data_gwas_results </pre>	<p>Target: input file containing genetic data from the study sample.</p> <p>Thread: limited to available core in the system.</p> <p>Clump-r2: threshold for clumping (by default correlation of 0.1)</p> <p>Fast score: if TRUE, only calculates PRSs in thresholds indicated in bar-levels</p> <p>No regress: we do not perform any regression model between the phenotype and the PRS. Simply output all PRSs.</p> <p>Bar-levels: thresholds of significance.</p> <p>Perm: number of permutations to calculate the empirical p-value.</p> <p>All-score: calculate PRSs for all thresholds</p> <p>Out: the output file name.</p>																														
Input files	<div style="display: flex; align-items: center;"> <table border="1" style="font-size: 8px; border-collapse: collapse;"> <thead> <tr> <th>CHR</th> <th>BP</th> <th>A1</th> <th>A2</th> <th>SNP</th> <th>P</th> <th>BETA</th> <th>SE</th> <th></th> <th></th> </tr> </thead> <tbody> <tr> <td>1</td> <td>732809</td> <td>T</td> <td>C</td> <td>rs12131618</td> <td>0.2571</td> <td>0.088426</td> <td>0.078921</td> <td></td> <td></td> </tr> <tr> <td>1</td> <td>768448</td> <td>G</td> <td>A</td> <td>rs12562034</td> <td>0.1702</td> <td>0.088242</td> <td>0.064331</td> <td></td> <td></td> </tr> </tbody> </table> </div> <p>The most recent GWAS for each phenotype</p>	CHR	BP	A1	A2	SNP	P	BETA	SE			1	732809	T	C	rs12131618	0.2571	0.088426	0.078921			1	768448	G	A	rs12562034	0.1702	0.088242	0.064331			<div style="border: 1px solid red; padding: 5px;"> $PRS_j = \sum_i \frac{S_i \times G_{ij}}{M_j}$ <p> S_i : summary statistic for the effective allele (effect size GWAS) G_{ij} : number of the effective allele observed in the sample M_j : number of alleles included in the PRS of the j^{th} individual </p> </div>
CHR	BP	A1	A2	SNP	P	BETA	SE																									
1	732809	T	C	rs12131618	0.2571	0.088426	0.078921																									
1	768448	G	A	rs12562034	0.1702	0.088242	0.064331																									
Output file	<pre> base_data_gwas_results_SCORES_AT_ALL_THRESHOLDS_imputed_ALLBATCHES_AD1 Bloc de notes Archivo Editar Formato Ver Ayuda FID IID Pt 5e-08 Pt 5e-06 Pt 0.0001 Pt 0.001 Pt 0.01 Pt 0.05 Pt 0.1 Pt 0.2 Pt 0.5 Pt 1 1 202186250009_R01C01 -0.00136644944 -0.000501005068 -0.0041713093 -0.00354366839 -0.00155197258 -0.00129608675 -0.00101401154 -0.000736673537 -0.00047 2 202186250009_R01C02 0.000118842426 0.000328767454 -0.00365220309 -0.00344406876 -0.00152186325 -0.00128896865 -0.00100632723 -0.000732243049 -0.00047 3 202186250009_R02C01 0.00153474934 0.00125096285 -0.00392095538 -0.00343064771 -0.00157423791 -0.00129506667 -0.00101391272 -0.000736409055 -0.00047 </pre> <p>PRSs calculated under different thresholds (SNPs significance obtained in the GWAS)</p> <p>PRSs selected for the analyses: those calculated under the threshold 5.10^{-8}</p>																															

Figure A11. Workflow to compute PRSs: from the files needed to compute PRSs to the output file.

In Figure A12, we can observe the workflow needed to apply the compositional method: from the dichotomisation of the PRSs to the application of the *Selbal* algorithm.


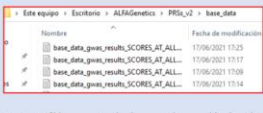








Step	Aim	Input	Output
1. Analysis PRSs			
Select PRSs  3.2.2-MergePRS_TFM_5e8.R	<ul style="list-style-type: none"> Select the PRSs of interest, under a specific threshold of significance, together with the "ID" column to be able to merge with other information of the subjects. Standardize the scores . Merge with other information of interest (covariates). 	All files within the directory "base_data" that contains a specific pattern indicating that it is a .txt file for a concrete PRS.  A .txt file containing other clinical and epidemiological information: age, sex, APOE status, volumes measurements.	 20210720_ALFA_PRS_TFM_5e-8_COVSVolHip.txt A .txt file which merges all the information: hippocampal subfields volumes, covariables and desired standardized PRSs.
Dichotomise PRSs  4.6-Function_GenerateBinaryPRS.R	<ul style="list-style-type: none"> Define two functions to dichotomise the PRSs and analyse the results <p>Functions</p> <p>a) <i>DichotomisePRS</i></p> <p>b) <i>AnalysisPRSbin</i></p>	<p>DichotomisePRS...</p> <p><i>PRSp</i> = path where the .txt of interest is located.</p> <p><i>Thresvalue</i> = threshold to dichotomise.</p> <p><i>Txtpath</i> = path for the new generated .txt file.</p> <p>AnalysisPRSbin</p> <p><i>Diseasebin</i> = label PRS_disease_bin (e.g. "PRS_AD_bin").</p> <p><i>Txtpath</i> = path .txt file of interest with binary PRSs.</p> <p><i>Thresvalue</i> = threshold to dichotomise (0.8)</p> <p><i>Disease</i> = label "disease" (e.g. "AD").</p> <p><i>DiseasePRS</i> = label "PRS_disease" (e.g. "PRS_AD").</p>	 20210720_ALFA_PRS_TFM_BIN_5e-8_COVSVolHip.txt  A list containing different elements: Boxplot PRSs, boxplot stratifying by sex, number of individuals within each PRS group....
2. Application selbal algorithm			
Apply selbal  ApplySelbal.R	Define a function to apply the selbal algorithm in different cases, depending on whether we stratify by sex, define different compositions, adjust by covariates..	The .txt file containing binary PRSs together with all the covariates information.	  A list containing different elements: accuracy graph, number of variables included in the balance, variables which are included in the numerator or denominator, mean AUC, ROC plot, and other plots. Results saved in .RData objects.
3. Analysis selbal results			
Analysis selbal results  AnalyseSelbal.R	Analyse the results of the selbal application and create tables and graphical visualisations to describe the results.	.Rdata objects saved from the application of the selbal algorithm.	Tables and graphics summarising the results.

Figure A12. Workflow describing the scripts and steps needed to apply the compositional method

R scripts are available under request in https://github.com/PatiGenius/FMP_Omics.