Multivariate haplotype analysis of 96 sulci opening for 15,612 UK-Biobank subjects

S. Karkar, A. Gloaguen, Y. Le Guen, Morgane Pierre-Jean, Claire Dandine-Roulland, Edith Le Floch, C. Philippe, A. Tenenhaus, V. Frouin

MULTIVARIATE HAPLOTYPE ANALYSIS OF 96 SULCI OPENING FOR 15,612 UK-BIOBANK SUJECTS

S. Karkar[1✉⋆], *A. Gloaguen*[1,2⋆], *Y. Le Guen*[1], *M. Pierre-Jean*[3], *C. Dandine-Roulland*[3], *E. Le Floch*[3], *C. Philippe*[1], *A. Tenenhaus*[2], *V. Frouin*[1]

[1]Neurospin, Institut Joliot, CEA - Université Paris-Saclay, 91191, Gif-Sur-Yvette, France
[2] École CentraleSupelec, 91191, Gif-sur-Yvette, France
[3] CNRGH, Institut Jacob, CEA - Université Paris-Saclay, 91000, Évry, France
✉ slim.karkar@cea.fr

ABSTRACT

Imaging genetic studies of large control cohorts such as UK Biobank enable to assess the range of normal variations in brain structures. Previous studies by our group have shown that the width of several cortical sulci is associated with a variant in the upstream region of KCNK2 gene even if this effect is corrected with age. Here we propose to analyze in a multivariate setup the associations between sets of genetic variants and multiple sulci widths. The genetic variants we consider are sets of SNPs of known phase called haplotypes, taken from the upstream region of KCNK2 gene. To the best of our knowledge, multivariate analysis in imaging genetics has never been used in haplotype studies. Our method was able to recover the expected association signal and uncover new associations between imaging data and genetic variants.

Index Terms— Imaging genetics, haplotype, multivariate analysis

I. INTRODUCTION

Grey matter thickness is known to shrink with aging in both diseased and normal brains [1], [2], [3]. A related effect is the cortical sulcus widening [4], [5]. The width of a sulcus can be estimated using a feature called opening [6] shown to be robustly related to grey matter thickness and which does not require spatial normalization nor regional atlas. Heritability studies pointed to a dozen sulci that appeared to be under strong genetic control [7]. Furthermore, in [8], GWAS has identified a reproducible genetic marker associated with the opening of the left, posterior, Calloso-Marginal sulcus. In this work, we used the Brainvisa cortical sulci recognition pipeline to automatically segment [9] and label [10] nearly one hundred brain sulci. GWAS use a univariate approach and as such, suffer from several drawbacks, in particular the use of an unduly conservative multiple test correction and the fact that the correlation structure of the genome is not accounted for. In the context of complex traits, where individual variant effect size is expected to be small, only SNPs that are frequent in the population can significantly be associated with the phenotype. Moreover, univariate analyses are unable to model or predict the role of a genetic variant within the genomic region. Finally, univariate approaches are inadequate in situations where a set of variants are jointly associated to multiple phenotypes (pleiotropy). Using a multiple phenotype multivariate approach, we propose to alleviate these drawbacks by simultaneously analyzing one hundred related phenotypes and to model interactions between genetic variants within the same genomic segment.

⋆Authors contributed equally

II. MATERIAL AND METHODS

We obtained 20,060 T1-weighted MRI from the January 2018 release, under UK Biobank (UKB) data appl. #25251. UKB cohort is particularly suited to study aging, with a mean participants age of 57 years, and a standard deviation of 8.2. We retained 15,612 subjects after QC protocol, British ancestry selection, and additional filtering for high heterogeneity, high missingness, first-degree relatedness and sex mismatch.

II-A. Imaging

For each selected subject, the brain mask of the T1-weighted image is obtained using SPM8 (fil.ion.ucl.ac.uk/spm). Next, individual brain images were segmented into grey matter, white matter and cerebro-spinal fluid (CSF) in BrainVisa. Finally, individual sulci were extracted using Morphologist, the sulcus identification pipeline of BrainVisa. We retained the 96 (out of 126) most sample-wide identified sulci : a sulci was retained if it was missing in less than 1000 individuals (94.6% presence rate, see [11]). For each retained sulcus and for each subject, sulcus width or opening (the average distance between both banks) was estimated as the ratio of CSF volume and surface area of the sulcus [7], [8].

II-B. Genetics

Genotyping data in UKB (UK Biobank Axiom Array) contains $820,967$ SNPs. In such data, for a given SNP, the variant status is obtained without knowing if it lays

on the paternal or maternal chromosome for a heterozygote subject. This raises an issue when one wants to use the chain of consecutive SNPs. In the 2018 release of UK Biobank, the so-called "phased data" are available for the 500,000 subjects. With this preprocessing, the succession of SNPs alleles is inferred in contiguous small regions of maternal or paternal chromosomes. Based on the results of a GWAS [8] where SNP rs864736 is found to be associated with the opening of several sulci, we chose a genomic region of 55.8 kbp on chromosome 1, which contained all the SNPs in linkage desequilibrium with rs864736 (i.e., SNPs which are supposed to be inherited together from the parents). This region consists in 18 SNPs. Using the "phased data" of our 15,612 subjects in this region, we derived all the haplotypes with a length of 3 to 18 SNPs.

II-C. Haplotype multivariate association analysis

The interplay between neuroimaging and genetic data is uncovered using Regularized Generalized Canonical Correlation Analysis (RGCCA), a general framework for multi-block data analysis [12], [13].

The first block, denoted $\mathbf{X}_1$, is related to neuroimaging and is defined by $p_1 = 96$ sulci measured on $n = 15,612$ individuals. The second block $\mathbf{X}_2$ is related to genetic information and is defined by $p_2 = 604$ haplotypes measured on the same set of n individuals. RGCCA aims to find block components $\mathbf{y}_j = \mathbf{X}_j\mathbf{w}_j$; $j = 1, 2$ (where $\mathbf{w}_j$ is a column vector with p_j elements) summarizing the relevant information between and within the blocks. In this context, RGCCA is defined as the following optimization problem :

$$\begin{aligned} &\max_{\mathbf{w}_1,\mathbf{w}_2} \quad \mathrm{cov}^2\left(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_2\mathbf{w}_2\right) \\ &\text{s.t.} \;\; \mathbf{w}_j^\top \hat{\mathbf{S}}_{jj}\mathbf{w}_j = 1, \;\; j = 1, 2. \end{aligned} \tag{1}$$

where $\hat{\mathbf{S}}_{jj} = \tau_j\mathbf{I} + \frac{(1-\tau_j)}{n-1}\mathbf{X}_j^\top\mathbf{X}_j$ and τ_j is a scalar between 0 and 1. $\hat{\mathbf{S}}_{jj}$ can be considered as a shrinkage estimate of the true variance-covariance matrix Σ_{jj} [14]. [15] gives an analytical formula for the optimal τ_j that minimizes the mean square error between Σ_{jj} and its estimate $\hat{\mathbf{S}}_{jj}$.

II-D. Bootstrap procedure and missing data imputation

A balanced bootstrap procedure [16] is used to assess the reliability of estimated weights. For that purpose $B = 2000$ bootstrap samples are considered. Some sulci were not detected in all individuals, therefore a simple regression imputation strategy is used to avoid missing values in each bootstrap sample of sulci opening data. A simple regression model is built to predict each opening value from the covariates Age, Sex, and the 10 first components of UK Biobank-provided MDS. Residuals where reported in a new $n \times p_1$ matrix, where subjects with missing sulci (i.e. where not accounted for in the regression model) are set to 0. This procedure allows both to impute missing values and remove effects of covariates which are confounding factors in our case. Finally, each residual bootstrap sample is standardized within each block in order to make the variables comparable. To make blocks comparable, each block was divided by the square root of its number of variables [13]. The `RGCCA` package (freely available at CRAN : cran.r-project.org) was then used to yield the weights vectors $\mathbf{w}_1^b$ and $\mathbf{w}_2^b$ for each bootstrap sample $b = 1, \ldots, 2000$.

II-E. Confidence intervals and variable selection

In order to assess the reliability of estimated weights, α-level confidence intervals ($\mathcal{C}_{k,\alpha}^j,\ j = 1, 2$) were derived for each element w_{1k} (corresp. to the k^{th} sulcus) of $\mathbf{w}_1$ and w_{2k} (corresp. to the k^{th} haplotype) of $\mathbf{w}_2$. Non-parametric ($\mathcal{C}_{k,\alpha}^{j,NP}$) and parametric ($\mathcal{C}_{k,\alpha}^{j,P}$) methods were considered to estimate these confidence intervals.

For $\mathcal{C}_{k,\alpha}^{j,NP}$, the empirical bootstrap distribution of $w_{j,k}$ was used to estimate the confidence interval. In order to be as restrictive as possible, we chose to only look at $\mathcal{C}_{k,\alpha_{min}}^{j,NP} = \left[\min_{b\in[\![1,B]\!]}(w_{j,k}^b), \max_{b\in[\![1,B]\!]}(w_{j,k}^b)\right]$, which corresponds to a level $\alpha_{min} = 2/B = 10^{-3}$.

For $\mathcal{C}_{k,\alpha}^{j,P}$, similarly to [17], we estimated the mean and variance for each element w_{1k} and w_{2k} across the bootstrap samples, from which we derived confidence intervals under the assumption that the weights estimation exhibited asymptotic normality.

Since simultaneous multiple bootstrap confidence intervals estimates were desired with an overall confidence coefficient $1 - \alpha$ for each block $\mathbf{X}_j,\ j = 1, 2$, we constructed each interval with confidence coefficient $1 - (\alpha/p_j),\ j = 1, 2$. The procedure is similar to a Bonferroni correction. For non-parametric estimation, there is no advantage in exploring levels lower than α_{min} since their confidence intervals correspond to the full distribution.

Nevertheless, non-parametric estimation still offers practical advantages as being closer to the real distribution, thus, we chose to look at three confidence intervals : $\mathcal{C}_{k,\alpha_{\min}}^{j,NP}$, $\mathcal{C}_{k,5\%,BF}^{j,P}$ and $\mathcal{C}_{k,10\%,BF}^{j,P}$. The last two confidence intervals are Bonferroni corrected as mentioned earlier. This correction is likely very conservative w.r.t. the independence of the p_j variables since many haplotypes are subsets of others. For these reasons we chose to explore the two levels of confidence $\alpha = 0.05$ and $\alpha = 0.1$.

A weight element w_{jk} is considered as relevant if zero is excluded from its confidence interval, since the probability of this weight being zero is lower than the level of confidence associated to this interval.

III. RESULTS

Figure 1 represents the weights $\mathbf{w}_j,\ j = 1, 2$ computed with RGCCA. Only relevant weights according to $\mathcal{C}_{k,\alpha_{\min}}^{j,NP}$ are shown. The ones associated to a single star were relevant for $\mathcal{C}_{k,10\%,BF}^{j,P}$, the ones with a double star were significant according to both $\mathcal{C}_{k,10\%,BF}^{j,P}$ and $\mathcal{C}_{k,5\%,BF}^{j,P}$.

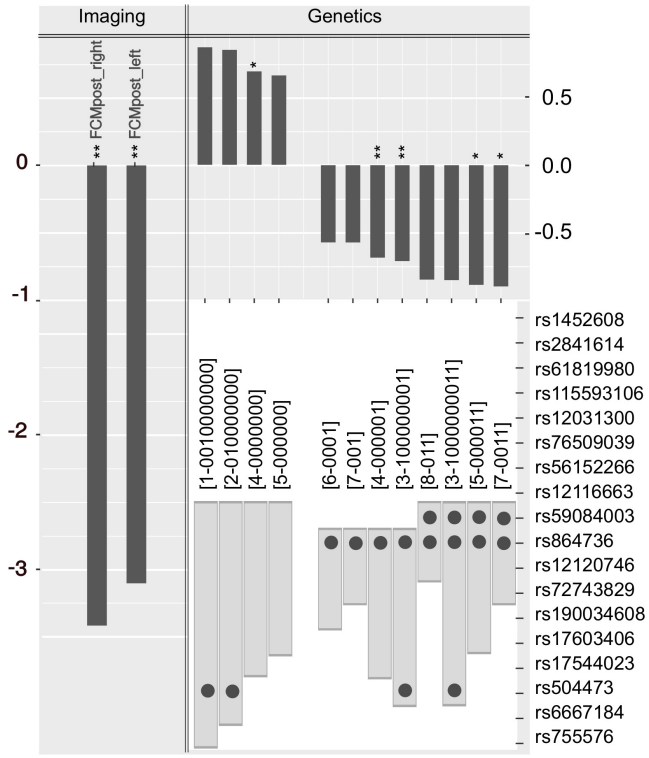

Fig. 1. (Left) : Weights $\mathbf{w}_1$ associated with the selected variables of the imaging block. Selected features were the bilateral Posterior Calloso-Marginal Sulci ; (Top, Right) : Weights $\mathbf{w}_2$ associated with the selected variables of the genetic block. All displayed variables were selected using $\mathcal{C}^{j,NP}_{k,\alpha_{\min}}$ (see text for detail) with $(**)$: variables selected with parametric confidence level $\mathcal{C}^{j,P}_{k,5\%,BF}$ and $(*)$: variables selected with parametric confidence level $\mathcal{C}^{j,P}_{k,10\%,BF}$. (Bottom, Right) : SNP composition of selected haplotypes : light grey bars show the extent of the sequence and dots indicate the location of alternative alleles (see text for details).

III-A. Selected variables for imaging block

Figure 1 (top, left) shows a barplot of the weights associated with the 2 selected features of the imaging block. The selected variables correspond to the bilateral Posterior Calloso-Marginal Sulci, of which the opening of the former was reported significantly associated with rs864736. Variables were relevant according to the three confidence intervals considered. There were 2 supplemental imaging features (right Subcallosal and Superior precentral sulci, not shown) relevant according to $\mathcal{C}^{j,P}_{k,10\%,BF}$. However, they were excluded since they were not selected using $\mathcal{C}^{j,NP}_{k,\alpha_{\min}}$.

III-B. Selected variables for genetic block

Figure 1 (top, right) depicts the weights associated with the 12 selected haplotypes (in decreasing order) according to $\mathcal{C}^{j,NP}_{k,\alpha_{\min}}$. Figure 1 (bottom, right) gives the composition of the haplotypes : selected sequences of variants are represented as a grey box. In each sequence, grey dots indicate the alternate alleles. Haplotypes are named as follows : [index of starting SNP - sequence of variants], e.g. $[4-000001]$ refers to the haplotype that starts on position #4 with 5 reference alleles and a single alternative allele at position #9. Selected haplotypes included various combinations of variants (from 3 to 10), however none of the haplotypes included variants between position 11 to 18. [4-000001] and [3-1000001] are relevant to $\mathcal{C}^{j,P}_{k,5\%,BF}$. If the confidence level is decreased to 10% using $\mathcal{C}^{j,P}_{k,10\%,BF}$, 3 additional haplotypes appear : [7-0011], [5-000011] and [4-0000000]. Only the last one has a positive weight.

III-C. Interpretation findings

We will interpret the sign of the weights using haplotype [4-000001] and Left Posterior Calloso-Marginal Sulcus (FCMpost_left) as an example. These both variables have negative weights in the model meaning that they are negatively correlated to their block component $\mathbf{y}_j,\ j = 1, 2$. However, over the $B = 2000$ bootstrap samples, correlation between $\mathbf{y}_1$ and $\mathbf{y}_2$ was always negative. To summarize, the presence of haplotype [4-000001] is associated to a lower opening for FCMpost_left : haplotypes with a negative weight have a protective effect on the sulcus opening w.r.t aging. Opposite conclusions are drawn for haplotypes with a positive weight in the model. However, interpretation of the values of these weights is not straightforward. To assess the magnitude of the association with the phenotypes we rely on explained variance in the following section.

IV. DISCUSSION

Previous studies by our group identified SNP rs864736 (and marginally rs59084003) as significantly associated with sulci opening and grey matter thickness for left Posterior Calloso-Marginal, Intra-Parietal and Central sulci. RGCCA was run two last times, first with only relevant variables according to $\mathcal{C}^{j,P}_{k,5\%,BF}$ (RGCCA-5%) then with variables based on $\mathcal{C}^{j,P}_{k,10\%,BF}$ (RGCCA-10%). For comparison purposes, we computed the explained variance using univariate analysis of the most significant variant (rs864736) and the sulci opening most associated with it (FCM posterior left and right). Several haplotypes had better or similar explained variance than rs864736 (see top and middle rows of Table I). Moreover, in RGCCA-10%, variance explained by the genetic block component $\mathbf{y}_2 = \mathbf{X}_2\mathbf{w}_2$ (where here, $\mathbf{X}_2$ is only composed of selected variables based on $\mathcal{C}^{j,P}_{k,10\%,BF}$) largely outperformed the single-SNP model. When we evaluated the variance of the imaging block component $\mathbf{y}_1$, as explained by the full genetic block component $\mathbf{y}_2$, both RGCCA-5% and RGCCA-10% outperformed the single-SNP model.

Several selected haplotypes are subsets of each other, and differ by one variant. If the SNP they differ from is not frequent, they represent very similar variables. To evaluate the impact of such co-linear variables, haplotype variables were excluded if the percentage of subject that differs for two haplotypes is less than 1%. In this case, only the variable for most frequent haplotype was kept. On this new dataset, we

Variants/Model	FCM-Post. left	FCM-Post. right
rs864736	0.46	0.44
$[4-000001]$	0.46	0.53
$[3-1000001]$	0.46	0.53
$[8-011]$	0.38	0.36
$[5-000011]$	0.48	0.35
$[4-0000000]$	0.48	0.35
RGCCA-5% - $\mathbf{y_2}$	0.46	0.53
RGCCA-10% - $\mathbf{y_2}$	0.75	0.69
RGCCA-5% - $\mathbf{y_1}$ & $\mathbf{y_2}$	0.59	
RGCCA-10% - $\mathbf{y_1}$ & $\mathbf{y_2}$	0.71	

Table I. Percentage of Explained Variance for univariate analysis using : (top row) single-SNP ; (five middle rows) selected variables at confidence level $\mathcal{C}^{j,P}_{k,10\%,BF}$; (RGCCA-5% and RGCCA-10%) model-derived haplotype combination and (2 bottom rows) variance of the imaging block component $\mathbf{y}_1$, explained by the genetic block component $\mathbf{y}_2$

used the same bootstrap procedure to derive a new $\mathcal{C}^{j,P}_{k,5\%,BF}$. In a similar fashion to RGCCA-5%, we then derived a third model. For this model, 3 variables in the genetic block were selected ([2-01000001], [4-000001] and [8-011]). Variables in the imaging block remained the same and their associated weights were similar to RGCCA-5%. The combination of the 3 genetic variables exhibits similar explanatory power as RGCCA-10%, with an explained variance of 0.64 and 0.7 for left and right Calosso-Marginal sulci, respectively.

Work in progress includes application of sparse versions of RGCCA [18], [19] and comparisons with the presented results.

V. CONCLUSION AND FUTURE WORKS

We proposed a multivariate model for haplotype associations with multiple quantitative traits that successfully recovered previously known associations, and gained substantial knowledge regarding the genomic region and associated sulci. We present three new findings : 1) only the genomic region located before rs864736 and rs59084003 seems to be implicated in the association ; 2) haplotype combinations are explanatory variables regarding Calloso-Marginal sulcus in both hemispheres ; and 3) an alternate allele at the third position (rs504473) seems to be associated with an antagonistic effect w.r.t rs864736 and rs59084003. Future works will extend this approach to gene clusters, gene pathways and larger intergenic regions to detect regulating patterns that interact with the observed phenotypes.

This method relies on a critical variable selection procedure based on the bootstrap folds. This procedure has shown to be sensitive to strongly co-linear variables, therefore we intend to propose several developments that could enhance this step. First, using a tree-like representation of haplotypes, we could regularize or combine variables, thus allowing us to keep more observations for the model estimation. Second, using block sparsity and regularization, multivariate procedures such as sparse group-lasso could better account for co-linearity of the variables. In the context of imaging genetics, we argue that insights provided by multivariate approaches are key in uncovering the complex interactions between genes, structure and function.

VI. REFERENCES

[1] Y. Ge *et al.*, "Age-related total gray matter and white matter changes in normal adult brain. part ii: Quantitative magnetization transfer ratio histogram analysis," *American Journal of Neuroradiology*, vol. 23, no. 8, pp. 1334–1341, 2002.

[2] A. M. Fjell *et al.*, "Structural brain changes in aging: Courses, causes and cognitive consequences," *Reviews in the Neurosciences*, vol. 21, no. 3, pp. 187–222, 2010.

[3] S. N. Lockhart *et al.*, "Structural imaging measures of brain aging," *Neuropsychology review*, vol. 24, no. 3, pp. 271–289, 2014.

[4] P. Kochunov *et al.*, "Age-related morphology trends of cortical sulci," *Human Brain Mapping*, vol. 26, pp. 210–220, 2005.

[5] X. Shen *et al.*, "Variation in longitudinal trajectories of cortical sulci in normal elderly," *NeuroImage*, vol. 166, pp. 1 – 9, 2018.

[6] D. Riviere *et al.*, "Brainvisa: an extensible software environment for sharing multimodal neuroimaging data and processing tools," *NeuroImage*, vol. 47, p. S163, 2009.

[7] Y. Le Guen *et al.*, "Genetic influence on the sulcal pits: On the origin of the first cortical folds," *Cerebral Cortex*, vol. 28, no. 6, pp. 1922–1933, 2018.

[8] Y. Le Guen *et al.*, "eQTL of KCNK2 regionally influences the brain sulcal widening: evidence from 15,597 uk biobank participants with neuroimaging data," *Brain Structure and Function*, Dec 2018.

[9] C. Fischer *et al.*, "Morphologist 2012: the new morphological pipeline of brainvisa," in *Proc. HBM*, 2012.

[10] M. Perrot *et al.*, "Cortical sulci recognition and spatial normalization," *Medical Image Analysis*, vol. 15, no. 4, pp. 529 – 550, 2011.

[11] L. Borne *et al.*, "A patch-based segmentation approach with high level representation of the data for cortical sulci recognition," in *Patch-Based Techniques in Medical Imaging*, ser. Lecture Notes in Computer Science, W. Bai *et al.*, Eds., vol. 11075. Springer, Cham, 2018.

[12] A. Tenenhaus *et al.*, "Regularized Generalized Canonical Correlation Analysis," *Psychometrika*, vol. 76, pp. 257–284, 2011.

[13] M. Tenenhaus *et al.*, "Regularized generalized canonical correlation analysis: A framework for sequential multiblock component methods," *Psychometrika*, vol. in press, 2017.

[14] O. Ledoit *et al.*, "A well conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, pp. 365–411, 2004.

[15] J. Schäfer *et al.*, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical applications in genetics and molecular biology*, vol. 4(1):Article32, 2005.

[16] J. R. Gleason, "Algorithms for balanced bootstrap simulations," *The American Statistician*, vol. 42, no. 4, pp. 263–266, 1988.

[17] A. Tenenhaus *et al.*, "Kernel Generalized Canonical Correlation Analysis," *Computational Statistics and Data Analysis*, vol. 90, no. C, pp. 114–131, Oct. 2015.

[18] A. Tenenhaus *et al.*, "Variable selection for generalized canonical correlation analysis." *Biostatistics (Oxford, England)*, vol. 15, no. 3, pp. 569–83, 2014.

[19] T. Löfstedt *et al.*, "A general multiblock method for structured variable selection," 2016.