# BOOSTING POWER TO DETECT GENETIC ASSOCIATIONS IN IMAGING USING MULTI-LOCUS, GENOME-WIDE SCANS AND RIDGE REGRESSION

*Omid Kohannim[1], Derrek P. Hibar[1], Jason L. Stein[1], Neda Jahanshad[1],*
*Clifford R. Jack, Jr.[2], Michael W. Weiner[3,4], Arthur W. Toga[1], Paul M. Thompson[1],*
*and the Alzheimer's Disease Neuroimaging Initiative*

[1]Laboratory of Neuro Imaging, Dept. of Neurology, UCLA School of Medicine, Los Angeles, CA, [2]Mayo Clinic, Rochester, MN, [3]Depts. of Radiology, Medicine and Psychiatry, UC San Francisco, San Francisco, CA, [4]Dept. of Veterans Affairs Medical Center, San Francisco, CA

## ABSTRACT

Most algorithms used for imaging genetics examine statistical effects of each individual genetic variant, one at a time. We developed a new approach, based on ridge regression, to jointly evaluate multiple, correlated single nucleotide polymorphisms (SNPs) in genome-wide association studies (GWAS) of brain images. Our goal was to boost the power to detect gene effects on brain images. We tested our method on MRI-derived hippocampal and temporal lobe volume measures, from 740 subjects scanned by the Alzheimer's Disease Neuroimaging Initiative (ADNI). We identified two significant and one almost significant SNP for the hippocampal and temporal lobe volume phenotypes, respectively, after correcting for multiple statistical tests across the genome. Ridge regression gave more significant associations than univariate analysis. Two SNPs, near regulatory genomic regions, showed significant voxelwise effects in *post hoc*, tensor-based morphometry analyses. Genome-wide ridge regression may detect SNPs missed by univariate GWAS, by incorporating multi-SNP dependencies in the model.

***Index Terms***— Neuroimaging, MRI, Imaging Genetics, Alzheimer's Disease, Ridge Regression, Genome-Wide Association Study, GWAS

## 1. INTRODUCTION

Imaging genetics is a new, emerging field in biomedical imaging. It aims to discover specific genetic variants, such as SNPs, that account for differences in anatomy and function. Analysis of image databases may pick up gene effects more efficiently than analysis of clinical or behavioral test scores, as imaging measures have high precision and reproducibility. Many image-derived measures such as hippocampal and caudate volumes are highly heritable and may be more directly influenced by genetic variation [1]. Already, genome-wide association studies (GWAS) of large MRI datasets (e.g., ADNI; *N*=740 subjects) have identified new candidate Alzheimer's Disease (AD) risk genes including *GRIN2B* [2] and *TOMM40* [3] that are associated with regional brain volumes.

Several international efforts, such as the Enigma project (http://enigma.loni.ucla.edu) are currently searching for genetic variants that affect brain structure and function, using databases of up to 10,000 images [4]. Imaging GWAS studies, so far, only consider the independent effect of each variation in the genome. This ignores useful information from multiple SNPs in the same gene, and across the genome. Many brain measures are heritable; each SNP has a weak effect on its own, but moderate to strong effects are likely when all SNP effects are aggregated across the whole genome. To model effects of large numbers of predictors with weak effects, machine learning approaches, including penalized regression, artificial neural networks, support vector machines, and adaptive boosting methods, have been introduced for GWAS in the last few years. These so-called *multi-locus* genetic methods model the combined effect of large numbers of SNPs, to explain more of the genetic contribution to particular phenotypes [5]. These approaches have been applied to several phenotypes related to rheumatoid arthritis [6] and coronary heart disease [7], but they have not yet been applied to the analysis of brain image databases.

Ridge regression [8-10] is one of several penalized regression methods for high-dimensional data analysis. Ridge regression works in many situations where ordinary multiple regression breaks down. It handles large numbers of highly correlated predictors, such as SNPs. Malo et al. [11] showed that ridge regression tends to outperform standard, univariate regression in GWAS studies, except when only one single SNP affects the measures of interest. Ridge regression has been previously applied to the study of multiple SNPs [6,11], but not at a genome-wide level, and not in the field of imaging. Here, we apply this modified multiple regression approach to a genome-wide analysis of the baseline brain MRI data from ADNI. We set out to find SNPs associated with measures of temporal lobe volume, and hippocampal volume, based on structural MRI. We hypothesize that our novel, genome-wide, penalized regression GWAS approach would help identify new candidate SNPs associated with imaging measures, including SNPs missed using standard univariate GWAS, which tests them independently.

## 2. METHODS

### 2.1. Structural MRI Measures

All subjects were scanned with a standard MRI protocol developed for ADNI. Hippocampal volumes were generated by an automatic segmentation method developed by our group, based on adaptive boosting [12]. Temporal lobe volumes were derived from an anatomically defined region-of-interest (ROI) on three-dimensional atrophy maps generated with tensor-based morphometry (TBM), a well-established method for mapping volumetric differences in the brain [13]. Data were available for 740 ADNI subjects (173 AD, 361 MCI, 206 controls; 438 men/302 women; mean ± SD age: 75.55 ± 6.79 years). MRI-derived measures of hippocampal volume were computed for a subset of 696 ADNI subjects (162 AD, 343 MCI, 191 controls; 405 men/ 291 women; mean ± SD age: 75.36 ± 6.77 years). These measures were adjusted for sex and age.

### 2.2. Genotypes

Genotyping procedures for ADNI are thoroughly described in [14]. As described in [15], genotypes were imputed to remove missing information and to compute the effective number of statistical tests across the genome; we also extracted SNPs that had minor allele frequencies greater than 10%, and Hardy-Weinberg equilibrium *p*-values more strict than $5.7 \times 10^{-7}$.

### 2.3. Ridge Regression

Hoerl introduced ridge regression as a variant of multiple regression [8-10]. It is designed to handle high-dimensional data, in cases where high correlations among the predictors would lead standard multiple regression methods to fail. Several variants of the same method were independently discovered in separate branches of mathematics and statistics. Tikhonov regularization [16] is a related concept, for solving inverse problems. It enforces solutions to be smooth, by minimizing a penalty function that controls the regularity of the solution.

In standard multiple regression, coefficients, $\beta_i$, are obtained by minimizing the residual sum of squares of the data, after fitting the regression model, which yields:

$$\beta = (X^t X)^{-1} X^t Y \qquad (1)$$

Here $X$ is an $n$ x $p$ matrix of $p$ predictors or SNPs and $Y$ is an $n$-dimensional vector of imaging measures obtained from $n$ subjects. In analyses such as ours, where there are many highly correlated predictors, standard multiple regression fails because the $X^t X$ matrix is highly ill-conditioned or not invertible.

Ridge regression addresses this by introducing a positive *shrinkage* parameter, $\lambda$, to obtain regression coefficients as follows:

$$\beta = (X^t X + \lambda I)^{-1} X^t Y \qquad (2)$$

Here $I$ is the $p$ x $p$ identity matrix. As its name suggests, $\lambda$ constrains the size of the regression coefficients by shrinking their variance to a specific, tunable extent. The idea of solving a regression equation with coefficients that are as small as possible (and forcing some to zero) is highly related to *compressed sensing* in computer vision [17] or "L1-norm" minimization methods in mathematics, such as Bregman splitting [18]. Coefficients are standardized after dividing them by their standard errors, which are the square roots of the diagonal elements of the variance-covariance matrix:

$$\mathrm{var}(\beta) = (X^t X + \lambda I)^{-1} X^t X (X^t X + \lambda I)^{-1} \sigma^2 \quad (3)$$

[19]. Although we implemented this more traditional form of the variance-covariance matrix for our ridge regression analyses, other studies have considered alternative forms of this matrix, such as:

$$\mathrm{var}(\beta) = \sigma^2 (X^t X + \lambda I)^{-1} \qquad (4)$$

[20]. *P*-values are then obtained from the *t*-distributed standardized coefficients, using the following formula to compute the effective number of degrees of freedom (EDF), formulated specifically for ridge regression [11]:

$$EDF = trace(X(X^t X + \lambda I)^{-1} X^t) \qquad (5)$$

As the ridge regression shrinkage parameter, $\lambda$, gets closer to zero, the model behaves more like multiple regression with a similar coefficient of determination ($R^2$). As $\lambda$ approaches infinity, the model acts more similarly to univariate regression and the $R^2$ or predictability of the model decreases. There is a need to find the optimal shrinkage parameter, offering sufficient shrinkage to allow for multi-collinearity, but not so high that the model loses its predictive ability. Several statistical methods exist to select the best shrinkage parameter for a ridge regression model. These include the Hoerl, Kennard and Baldwin estimator [21], and the Lawless and Wang estimator [22]. Here, we base our shrinkage estimation on the latter estimator (*LW* estimator, below), which, through a Bayesian approach, estimates $\lambda$ as follows:

$$\lambda = p\sigma^2 / \sum_i \varepsilon_i (Q^t \beta)_i^2 \qquad (6)$$

Here, the $\varepsilon_i$ are eigenvalues of $X^t X$, $Q$ is a $p$ x $p$ matrix, with the eigenvectors of $X^t X$ as column vectors, and $\beta$ represents the vector of standard multiple regression coefficients.

To illustrate the concept of ridge regression and test how the *LW* estimator performs with our data, we considered six SNPs in various incrementally decreasing levels of linkage disequilibrium, or correlation, with rs10845840, a SNP in the *GRIN2B* gene that our team previously identified with standard GWAS to be associated with temporal lobe volume [2]. We investigated how the ridge regression shrinkage parameter affected the power to detect the effect of the main SNP of interest, when other adjacent SNPs were added to the model. By randomly permuting the imaging data 10,000 times (i.e., assigning the images to the wrong subjects), we ensured that the reference distribution of *t* statistics was appropriate for assessing significance (results are shown in **Figure 1**).

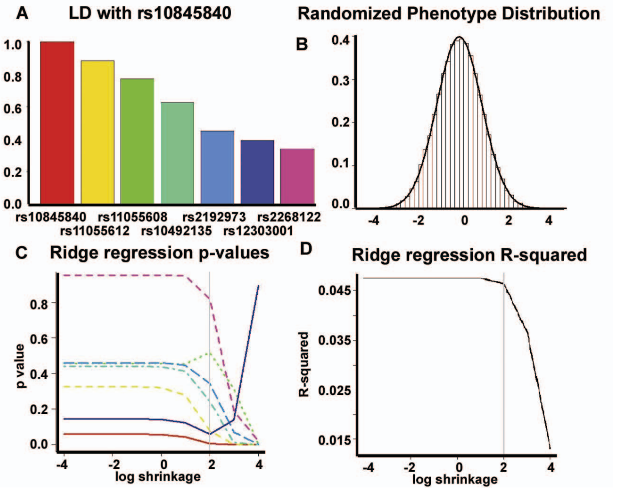

**Figure 1:** (a) The degrees of statistical correlation, across subjects, of the selected *GRIN2B* SNPs are shown with respect to rs10845840. In genetics, this is known as linkage disequilibrium (LD). (b) shows the distribution (*histogram*) of standardized ridge regression coefficients with shuffled phenotypes, along with a *t*-distribution with the same effective number of degrees of freedom (*curve*). The point of this experiment is to show that when the data are null, by construction, the *t*-statistic computed from the ridge regression formula correctly controls for false positive associations in null data. (c) Ridge regression *P*-values are graphed against the level of shrinkage. Each color represents a SNP, corresponding to panel (a). The gray line represents the optimal shrinkage determined by the *LW* estimator. (d) The regression model's $R^2$ is plotted against the degree of shrinkage; the thin gray line is as is in panel (c). With the optimized level of shrinkage identified via the *LW* estimator, there is a negligible loss in the model's predictive ability.

We performed ridge regression analyses separately for each imaging measure, for different "windows" or genomic regions of interest. In other words, we used a window-based scan of the genome, considering all nearby SNPs that are in high LD with a given SNP at the center of the window, as advocated by Malo et al. [11]. Windows were created by considering all SNPs that passed a liberal, univariate GWAS *p*-value threshold of 0.10, along with their neighboring SNPs within a fixed distance, in single ridge regression models with the optimal shrinkage parameters. We tried a range of fixed window sizes (50 Kbp, 100 Kbp, 500 Kbp, and 1 Mbp) around the SNPs of interest. Here Kbp and Mbp denote thousands or millions of base pairs on the genome.

### 2.4. Multiple Comparisons

To correct for multiple comparisons, we divided the nominal *p*-value threshold of 0.05 by an estimate of the *effective* number of statistical tests across the genome. The effective number of tests ($M_{eff}$) was calculated for each chromosome using the simpleM program, as detailed in [15]. To perform *post hoc*, exploratory tests on the top SNPs, we created voxelwise statistical maps using standard linear regression. To correct for multiple spatial comparisons, we used the standard False Discovery Rate method (FDR) [23].

## 3. RESULTS

As described in the methods, the corrected *p*-value threshold was set to be 0.05 divided by the effective number of tests (i.e., 264,889), which is 1.89 x $10^{-7}$. None of the 437,607 SNPs in our study passed this threshold in association with hippocampal volume with standard, univariate GWAS. By using our genomic scanning of SNP windows using ridge regression, we were able to identify two intergenic SNPs (rs2912975 and rs4747490) that passed the stringent "genome-wide" significance threshold (i.e., correcting for all the statistical tests across the genome). As mentioned in the methods, several genomic window sizes were used, which yield different *p*-values for their corresponding SNPs. Below, we report the most significant *p*-values for the SNPs that passed the genome-wide *p*-value threshold.

The rs2912975 polymorphism on chromosome 7 was significantly associated with hippocampal volume (*p*-value = 4.98 x $10^{-8}$), along with its neighboring SNP, with which it is almost perfectly correlated. The SNP is located close to a predicted, regulatory sequence of DNA. Univariate GWAS yielded a non-significant *p*-value of 1.19 x $10^{-2}$ for this SNP on chromosome 7. We plotted the ridge regression *p*-values from a scanning window with rs2912975 and its neighbors as an example of boosted genomic association power from ridge regression, along with the standard GWAS *p*-values for the same SNPs (**Figure 2**). In addition, another intergenic SNP, rs4747490, on chromosome 10, was significantly associated with hippocampal volume. In the temporal lobe volume analysis, our genomic scanning of SNP windows with ridge regression boosted the significance of the intergenic, likely regulatory rs2456930 SNP on chromosome 15 closer to the significance threshold.

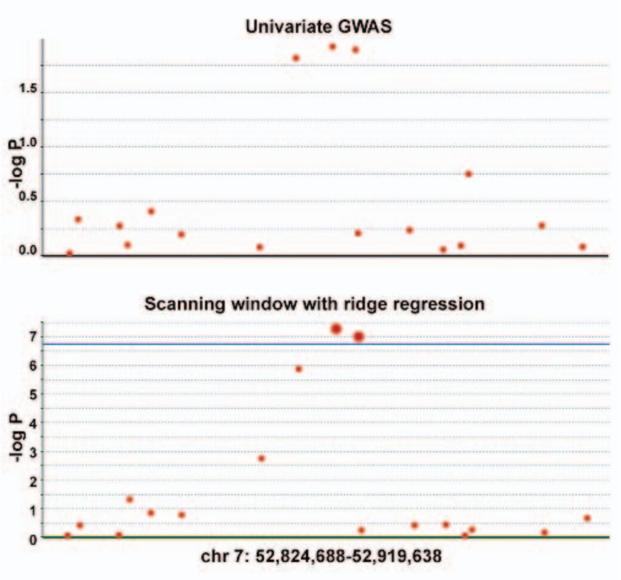

**Figure 2:** The rs2912975 SNP has a *p*-value of 1.19 x $10^{-2}$ with standard, univariate GWAS, when correlated with hippocampal volume (*top panel*). In a ridge regression model, when considered along with SNPs in its vicinity, the same SNP has a genome-wide significant *p*-value of 4.98 x $10^{-8}$ (ridge regression *p*-values for a scanning window containing this SNP are displayed in the bottom panel; the top SNP is rs2912975).

Since the rs10845840 SNP in the *GRIN2B* gene is statistically significant in its association with temporal lobe volume with standard univariate GWAS [2], we expected its significance to be also detected with our genome-wide ridge regression scanning technique, possibly at a boosted level. This was not the case, however. Therefore, we looked more closely at the ridge regression windows centered at this SNP. We found that although rs10845840 was the top SNP at 50Kbp and 100Kbp scanning windows, another SNP in the *GRIN2B* gene (rs1805502) was the most significant SNP in the larger window sizes of 500Kbp and 1Mbp. In fact, in the 500-Kbp window, rs1805502 had a more significant *p*-value (7.51 x $10^{-5}$) than its univariate *p*-value of 9.08 x $10^{-5}$. The rs1805502 SNP resides in the 3'-untranslated region (UTR) of the *GRIN2B* gene, almost 200Kbp away from rs10845840. The two SNPs are not in LD with each other, i.e. they are not statistically correlated in the population ($r^2$ = 0.006), and may therefore represent two independent contributions of the *GRIN2B* gene to the temporal lobe volume phenotype. To further explore this, we considered rs1805502 along with rs10845840 in a standard, multiple regression model (with no other SNPs). Both SNPs obtained *p*-values < 0.001, with minor alleles having effects in opposite directions. All other SNPs around rs10845840, even in our largest window size of 1Mbp, had *p*-values > 0.001 when they were paired with rs10845840 in standard, multiple regression models.

In *post hoc*, exploratory tests, we evaluated more closely the effects of the genetic variants that ridge regression found to be strongly associated with our MRI-based summary measures. We performed voxel-by-voxel association studies using TBM. SNPs were coded in an additive fashion (0, 1 or 2 for the number of minor alleles). Volumetric differences at each voxel were correlated with the SNPs separately using standard regression, after adjusting for age and sex. We found that rs2456930 showed significant effects (FDR critical *p*-value of 0.023) in the temporal lobe voxels, using both temporal lobes as the search regions of interest (**Figure 3**), and at the whole-brain level, with an FDR critical *p*-value of 3.71 x $10^{-3}$ (i.e., the highest *P* value threshold that controls the FDR at the conventional 5% rate). Furthermore, the rs2912992 SNP had a significant association (with an FDR critical *p*-value of 0.0099) selectively within the left hippocampal voxels (**Figure 4**), although it did not pass FDR for both hippocampi or for the right hippocampus. Replication studies and meta-analyses in even larger samples are underway, to confirm this.

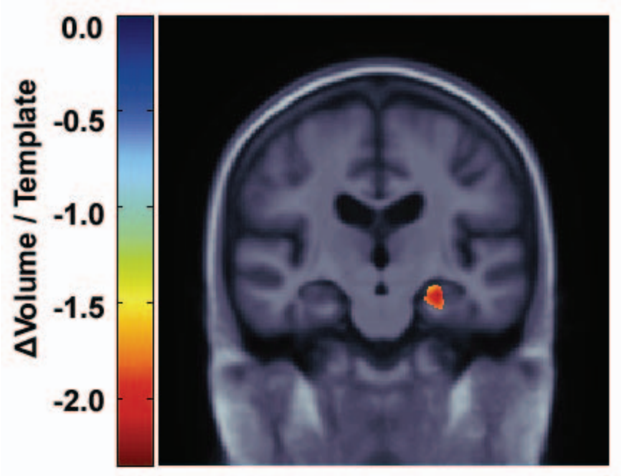

**Figure 3:** TBM reveals the profile of statistically significant effects of rs2912975 on the left hippocampus. FDR was used to correct for multiple comparisons in the left hippocampal voxels, using a binary mask. The image is in radiological convention (the right side of the image shows the left side of the subject's brain).

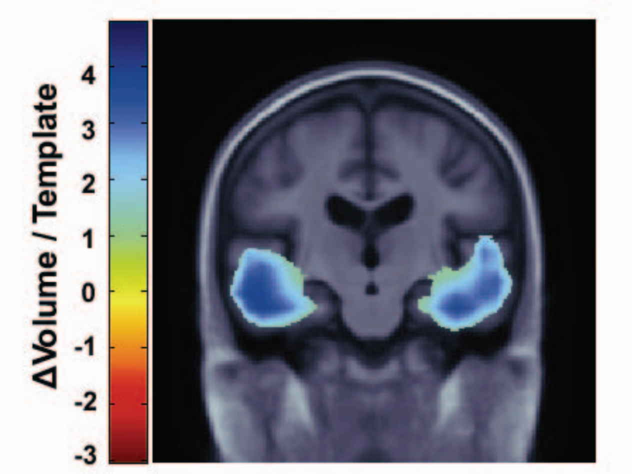

**Figure 4:** TBM reveals the profile of statistically significant effects of the rs2456930 polymorphism on the temporal lobes. FDR was used to correct for multiple comparisons in the temporal lobe voxels, using a binary mask for the temporal lobes. The image is in radiological convention (the right side of the image shows the left side of the subject’s brain).

## 4. CONCLUSION

We applied a novel, genome-wide, ridge regression approach to study the association of multiple SNPs with imaging phenotypes. We identified three SNPs with significant or near genome-wide significant effects on MRI-derived hippocampal volume and temporal lobe volume measures. In some but not all cases, associations were boosted in power relative to standard, univariate genome-wide association analyses. In addition, two of the three SNPs we identified had significant voxelwise effects in *post hoc* analyses and are located near regulatory DNA sequences, making them potentially important genetic variants for influencing brain structure in large populations.

## 5. ACKNOWLEDGMENTS

This study was supported by NIH grants EB008281 HD050735, RR013642, AG036535 (to PT), and the UCLA Medical Scientist Training Program (to OK).

## 6. REFERENCES

[1] P.M. Thompson et al., “Imaging genomics,” *Current Opinion in Neurology* 23:368-373, 2010.

[2] J.L. Stein et al., “Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in Alzheimer's disease,” *NeuroImage* 51(2):542-554, 2010.

[3] S.G. Potkin et al., “Hippocampal Atrophy as a Quantitative Trait in a Genome-Wide Association Study Identifying Novel Susceptibility Genes for Alzheimer’s Disease,” *PLoS One* 4(8):e6501, 2009.

[4] The ENIGMA Consortium (2011). “Genome-Wide Association Meta-Analysis of Hippocampal Volume: Results from the ENIGMA Consortium”, *Organization for Human Brain Mapping meeting*, Quebec City, Canada, June 2011.

[5] S. Szymczak et al., “Machine Learning in Genome-Wide Association Studies,” *Genetic Epidemiology* 33:S51-S57, 2009.

[6] Y.V. Sun et al., “Identification of correlated genetic variants jointly associated with rheumatoid arthritis using ridge regression,” *BMC Proceedings* 3:S67, 2009.

[7] Y. Kim et al., “Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects,” *BMC Proceedings* 3:S64, 2009.

[8] A.E. Hoerl, “Optimum solution of many variable equations,” *Chemical Engineering Progress* 55:67-78, 1959.

[9] A.E. Hoerl, “Application of ridge analysis to regression problems” *Chemical Engineering Progress* 58:54-59, 1962.

[10] A.E. Hoerl and R.W. Kennard, “Ridge regression: Biased estimation for non-orthogonal problems,” *Technometrics* 12:55-67, 1970.

[11] N. Malo et al., “Accommodating Linkage Disequilibrium in Genetic-Association Analysis via Ridge Regression,” *American Journal of Human Genetics* 82:375-385, 2008.

[12] J.H. Morra et al., “Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease mild cognitive impairment and elderly controls,” *NeuroImage* 43:59-68, 2008.

[13] X. Hua et al., “3D characterization of brain atrophy in Alzheimer’s disease and mild cognitive impairment using tensor-based morphometry,” *NeuroImage* 41:19-34, 2008.

[14] A.J. Saykin et al., “Alzheimer’s Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims progress and plans,” *Alzheimer’s Dement* 6(3):265-273, 2010.

[15] J.L. Stein et al., “Voxelwise genome-wide association study (vGWAS),” *Neuroimage* 53(3):1160-1174, 2010.

[16] A.N. Tikhonov and V.Y. Arsenin, *Solutions of ill-posed problems*. Winston, Washington, 1977.

[17] D.L. Donoho, “Compressed sensing,” *IEEE Trans Inf Theory*. 52:1289-1306, 2006.

[18] T. Goldstein and S. Osher, “The split Bregman method for L1 regularized problems,” *UCLA CAM report* 8-29, 2008.

[19] S. Chatterjee S and B. Price, *Regression analysis by example*. Wiley, New York, 1977.

[20] A.M. Halawa and M.Y. El Bassiouni, “Tests of regression coefficients under ridge regression models,” *J Statist Comput Stimul* 65:341-356, 2000.

[21] A.E. Hoerl et al., “Ridge regression: Some simulations,” *Comm Statist Theory Methods* 4:105-123, 1975.

[22] J.F. Lawless and P. Wang, “A simulation study of ridge and other regression estimators,” *Communications in Statistics - Theory and Methods* 5(4):307-323, 1976.

[23] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing,” *J R Statist Soc B* 57(1):289-300, 1995.