DETECTING GENETIC RISK FACTORS FOR ALZHEIMER'S DISEASE IN WHOLE GENOME SEQUENCE DATA VIA LASSO SCREENING

Tao Yang[1,2], *Jie Wang*[1,2], *Qian Sun*[1,2], *Derrek P. Hibar*[3], *Neda Jahanshad*[3], *Li Liu*[2], *Yalin Wang*[1], *Liang Zhan*[3], *Paul M. Thompson*[3,*], *Jieping Ye*[4,5,*]

[1]Dept. of Computer Science and Engineering, [2]Center for Evolutionary Medicine and Informatics, The Biodesign Institute; [12]Arizona State Univ., Tempe, AZ, USA; [3]Imaging Genetics Center, Keck School of Medicine, Univ. of Southern California, Los Angeles, CA, USA; [4]Dept. of Computational Medicine and Bioinformatics; [5]Dept. of Electrical Engineering and Computer Science; [45]Univ. of Michigan, Ann Arbor, MI, USA; [*]Joint corresponding authors

ABSTRACT

Genetic factors play a key role in Alzheimer's disease (AD). The Alzheimer's Disease Neuroimaging Initiative (ADNI) whole genome sequence (WGS) data offers new power to investigate mechanisms of AD by combining entire genome sequences with neuroimaging and clinical data. Here we explore the ADNI WGS SNP (single nucleotide polymorphism) data in depth and extract approximately six million valid SNP features. We investigate imaging genetics associations using Lasso regression—a widely used sparse learning technique. To solve the large-scale Lasso problem more efficiently, we employ a highly efficient screening rule for Lasso—called dual polytope projections (DPP)—to remove irrelevant features from the optimization problem. Experiments demonstrate that the DPP can effectively identify irrelevant features and leads to a 400 × speedup. This allows us for the first time to run the compute-intensive model selection procedure called stability selection to rank SNPs that may affect the brain and AD risk.

Index Terms— Alzheimer's Disease, Whole Genome Sequence, Lasso, Lasso Screening

1. INTRODUCTION

Alzheimer's disease (AD) is the most common form of dementia; it affects more than five million Americans and is the sixth leading cause of death in United States [1]. AD is an irreversible, progressive brain disorder typically beginning with mild memory loss; later it can seriously impair an individual's ability to carry out daily activities. It has been widely recognized and emphasized that early detection of AD is beneficial. Recently, neuroimaging techniques—such as magnetic resonance imaging (MRI), computed tomography and positron emission tomography—have shown great promise for evaluating AD and tracking its progression [2].

Factors that influence AD progression are not yet fully understood, but common genetic variants are among the major risk factors [3]. Novel sequencing techniques have greatly advanced genome-wide association studies (GWAS) and whole genome sequencing (WGS) studies. Now, entire genomes can be combined with brain imaging and clinical data to facilitate the investigation of the mechanisms of AD.

Besides the well-known *APOE* genotype, recent studies [4] of the Alzheimer's Disease Neuroimaging Initiative (ADNI) GWAS data have related known AD risk genes to differences in rates of brain atrophy and biomarkers of AD in the cerebrospinal fluid. More recently, full genetic sequences have been collected for over 800 ADNI participants. The ENIGMA Consortium recently discovered six common genetic variants associated with subcortical brain volumes in a worldwide screen of over 30,000 brain MRI scans [5]. Another study—the International Genomics of Alzheimer's Project (I-GAP) study [6], with over 74,000 participants—identified genetic risk factors with statistical methods, is the largest study of AD to date. However, these studies still have limitations. First of all, GWAS studies focus on a set of selected common genetic variants rather than the entire sequence of the genome. Loci showing strongest associations with the disease, or a brain measure, are not generally the causal SNPs, as the causal loci have typically not been sequenced directly. Secondly, studies such as I-GAP are based on simple statistical models that only test associations of each SNP, one at a time, with AD-related phenotypes. In other words, these methods typically ignore potential inter-locus interactions.

In this paper, we propose a novel approach to detect genetic risk factors in the ADNI WGS data via Lasso regression, along with screening rules. We investigate the imaging genetics associations between AD imaging phenotypes (*e.g.*, volumes of entorhinal cortex and hippocampus) and SNPs in the whole genome sequences of ADNI participants. The WGS data set includes approximately 5.9 million loci and may ultimately implicate rare genetic causal variants. Rather than performing a univariate test for each SNP and phenotype individually, we use Lasso—a widely used sparse model (*e.g.*, [7, 8])—to identify the most relevant SNPs. To make it more efficient to solve large-scale Lasso problems, we employ state-of-the-art Lasso screening rules (DPP) that quickly identify irrelevant

features. Experiments on the WGS data show the speed and effectiveness of the proposed method. In association with the AD imaging phenotypes, we also present potential risk SNPs via stability selection.

2. DATA PROCESSING

2.1. ADNI WGS SNP data

The ADNI WGS data contains the entire genome sequences for 818 ADNI participants, including 128 AD patients, 415 mild cognitive impairment (MCI) subjects, 267 elderly controls, and 8 participants of uncertain diagnosis. To store statistically relevant SNPs called using Illumina's CASAVA SNP Caller, the ADNI WGS SNP data is stored in *variant call format* (VCF) [9, 10]. The VCF is a specific text file format used in bioinformatics for storing gene sequence variations. SNPs at approximately 3.7 million specific loci are recorded for each participant. The genotype information is mainly documented into two columns: MAXGT and POLY, respectively [11]. Specifically, the MAXGT column refers to the commonest genotype, while the POLY column refers to genotyping assuming the site is polymorphic. Both columns are coded according to allele values and the conditional genotype quality scores. The allele value is 0 for the reference allele, 1 for the first allele listed among the alternative alleles (ALT), 2 for the second, and so on.

2.2. SNP genotype coding

A single nucleotide polymorphism is a DNA sequence variation occurring when a single nucleotide (A, T, C, or G) in the genome differs among members of a biological species or across paired chromosomes. Based on the original ADNI WGS SNP data, in this study, we encode SNPs using the additive coding scheme [12] as follows:

- 0/0 refers to reference homozygote, coded as 0;
- 1/1 refers to alternative homozygote, coded as 2;
- 0/1, 1/0, etc., refer to heterozygote, coded as 1;
- All missing entries are treated as reference homozygotes.

2.3. Quality control

To extract high quality informative SNP features, we first apply two major quality control criteria: Minor Allele Frequency (MAF) < 0.05 and Genotype Quality (GQ) < 45.

Allele frequency is the proportion of a particular allele among all allele copies being considered. For a specific locus, the *minor allele frequency* is the frequency of the least frequent allele in the population [13]. After genotype coding, SNPs with a minor allele frequency of 5% or greater are targeted in our research. Approximate MAFs for all loci are shown in **Fig. 1**.

A GQ score is recorded for each locus by its "Phred" quality. The Phred quality score $Q = -10\log_{10} P$ is logarithmically related to the base-calling error probability P [14]. We extract GQ scores of all SNPs from all subjects. A

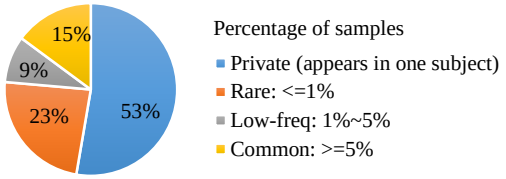

Fig. 1. The approximate MAFs among all loci.

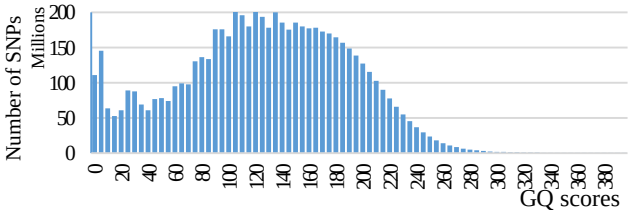

Fig. 2. Histogram of the GQ scores.

histogram of the GQ scores is shown in **Fig. 2**. We then fit the GQ score distribution using a Gaussian model $N(\mu, \sigma)$. Generally, SNPs detected via GQ scores that are less than a threshold $\theta = \mu - \sigma$ may be considered as inaccurate. Based on the fitted result, we choose $\theta = 45$ as the threshold to use. We then encode those SNPs with GQ < 45 as '-1', and remove the corresponding loci if the tag '-1' appears more than 80 times in all participants.

3. METHODS

In this paper, we investigate Alzheimer's disease genetic risk factors on the whole genome sequence data via Lasso and associated screening methods.

3.1 Lasso

Lasso is widely used linear regression technique in machine learning, data mining, bioinformatics etc. to find sparse representations [15]. Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_p] \in \Re^{N\times p}$ be the data matrix with N observations and p features, $\mathbf{y} \in \Re^{N\times 1}$ be a response vector, and $\lambda \geq 0$ be the regularization parameter. The Lasso problem takes the form of:

$$\inf_{\beta\in\Re^p} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_1. \quad (1)$$

The ℓ_1 −regularizer enforces sparsity in the optimal solution β^*. The non-zero components in β^* correspond to the relevant features in $\boldsymbol{X}$. In our experiments, $\boldsymbol{X}$ is the processed ADNI whole genome sequence SNP data matrix. Therefore, Lasso is potentially useful to locate important SNPs that are most relevant to predicting the specific phenotype. We employ the SLEP package [16] to solve the Lasso problem.

The dual problem of Lasso is equivalent to the following:

$$\inf_{\theta}\left\{\frac{1}{2}\left\|\theta - \frac{\boldsymbol{y}}{\lambda}\right\|_2^2 : |\boldsymbol{x}_i^T\theta| \leq 1, i = 1,2,\dots,p\right\}, \quad (2)$$

where θ is a dual variable. Let $\beta^*(\lambda)$ and $\theta^*(\lambda)$ be the optimal solution of problems (1) and (2), respectively. The Karush–Kuhn–Tucker (KKT) conditions are:

$$\boldsymbol{y} = \boldsymbol{X}\beta^*(\lambda) + \lambda\theta^*(\lambda), \quad (3)$$

$$\boldsymbol{x}_i^T\theta^*(\lambda) \in \begin{cases} \text{sign}([\beta^*(\lambda)]_i), & \text{if } [\beta^*(\lambda)]_i \neq 0, \\ [-1,1], & \text{if } [\beta^*(\lambda)]_i = 0, \end{cases} \quad (4)$$

where $[\cdot]_k$ denotes the k^{th} component. The KKT condition in Eq. (4) leads to

$$|\boldsymbol{x}_i^T(\theta^*(\lambda))^T| < 1 \Rightarrow [\beta^*(\lambda)]_i = 0 \Rightarrow \boldsymbol{x_i} \text{ is an inactive feature.} \quad (5)$$

Those inactive features have 0 components in β^* and thus can be removed from the optimization problem. Inspired by the SAFE rules [17], (5) can be relaxed as follows:

$$\sup_{\theta \in \boldsymbol{\Theta}} |\boldsymbol{x}_i^T \theta| < 1 \Rightarrow [\beta^*(\lambda)]_i = 0 \Rightarrow \boldsymbol{x_i} \text{ is inactive,} \quad (6)$$

where $\boldsymbol{\Theta}$ is a region that contains $\theta^*(\lambda)$. We can see that the smaller the region $\boldsymbol{\Theta}$ is, the more accurate the estimation of $\theta^*(\lambda)$—and thus more inactive features can be identified.

3.2 Enhanced DPP screening rules for Lasso

The idea of "screening" has been shown to be very promising in improving the efficiency of large-scale Lasso. Briefly speaking, screening rules aim to quickly identify the inactive features, which have 0 components in the solution. Inactive features can be removed from the optimization leading to substantial savings in computational cost and memory usage.

Motivated by the idea of [17] and [18], we develop the following sequential version of screening rules for Lasso problem via the enhanced dual polytope projections (EDPP) [19]. Suppose that we are given a sequence of parameter values $\lambda_1 > \lambda_2 > \cdots > \lambda_m$. We first apply EDPP to discard inactive features for the Lasso problem at λ_1. We compute the optimal solution $\beta^*(\lambda_1)$ by solving Lasso on the reduced data matrix. Then, by Eq. (3), we can find $\theta^*(\lambda_1)$. In view of (6), if we know the dual optimal dual solution $\theta^*(\lambda_1)$, we obtain a new screening rule for problem (1) at λ_2. By repeating the above procedure, we have the sequential version of EDPP.

Let $\lambda_{max} = \|\boldsymbol{X}^T\boldsymbol{y}\|_\infty$, and let

$$\boldsymbol{x}_* = \text{argmax}_{x_i} |\boldsymbol{x}_i^T \boldsymbol{y}|, \quad (7)$$

$$\boldsymbol{v}_1(\lambda_0) = \begin{cases} \dfrac{\boldsymbol{y}}{\lambda_0} - \theta^*(\lambda_0), \text{if } \lambda_0 \in (0, \lambda_{max}), \\ \text{sign}(\boldsymbol{x}_*^T \boldsymbol{y})\boldsymbol{x}_*, \text{if } \lambda_0 = \lambda_{max}, \end{cases} \quad (8)$$

$$\boldsymbol{v}_2(\lambda, \lambda_0) = \frac{\boldsymbol{y}}{\lambda_0} - \theta^*(\lambda_0), \quad (9)$$

$$\boldsymbol{v}_2^\perp(\lambda, \lambda_0) = \boldsymbol{v}_2(\lambda, \lambda_0) - \frac{\langle \boldsymbol{v}_2(\lambda, \lambda_0), \boldsymbol{v}_1(\lambda_0) \rangle}{\|\boldsymbol{v}_1(\lambda_0)\|_2^2} \boldsymbol{v}_1(\lambda_0). \quad (10)$$

EDPP: For the Lasso problem, suppose we are given a sequence of parameter values $\lambda_{max} = \lambda_0 > \lambda_1 > \cdots > \lambda_m$. Then for any integer $0 \le k \le m$, we have $[\beta^*(\lambda_{k+1})]_i = 0$ if $\beta^*(\lambda_k)$ is known and the following holds:

$$\left| \boldsymbol{x}_i^T \left(\frac{\boldsymbol{y} - \boldsymbol{X}\beta^*(\lambda_k)}{\lambda_k} + \frac{1}{2} \boldsymbol{v}_2^\perp(\lambda_{k+1}, \lambda_k) \right) \right| < 1 - \frac{1}{2} \|\boldsymbol{v}_2^\perp(\lambda_{k+1}, \lambda_k)\|_2 \|\boldsymbol{x}_i\|_2. \quad (11)$$

The sequential version of the EDPP has several appealing features. First, in a real application, the optimal parameter value of λ is typically unknown and needs to be estimated. Second, it can help accelerate the process of stability selection. In our study, we use the DPC package [19].

4. EXPERIMENTS

We evaluate the proposed the EDPP rules for Lasso on the ADNI WGS data. Specifically, we first report the efficiency of EDPP and then investigate the potential risk SNPs related to AD imaging phenotypes by stability selection.

4.1 Comparison of Lasso with and without EDPP rule

In this experiment, we compare the performance of Lasso with and without the EDPP screening rule. We choose the baseline hippocampal volume to be the response vector, so the trials contain 717 subjects. We vary the number of features n_f from 0.1 million to 1 million SNPs—that are randomly selected—with a step size of 0.1 million. For each n_f, we solve the Lasso problems at a sequence of 100 parameter values equally spaced in the logarithmic scale of $\lambda/\lambda_{\max}$ from 1.0 to 0.05. The run times are reported in **Fig. 3**.

Fig. 3 shows that the solver equipped with EDPP (EDPP+Solver) gains a speedup about 406× compared to the solver without screening. Moreover, if we double the dimension of the features, the run time of the solver without screening also doubles, while the run time of the solver with EDPP screening rule only increases slightly—which is mainly due to the screening part. Thus, the experiments demonstrate that the EDPP rule is a promising approach to facilitate the Lasso solver in dealing with extremely high dimensional data.

4.2 Stability selection on WGS data

In this experiment, we explore the imaging genetics association between imaging phenotypes and SNPs from the entire ADNI WGS SNP data with 5,906,152 features. For two brain regions, the entorhinal cortex (EC) and hippocampus (HIPP), we chose the volume at baseline, and volume changes over a 24-month interval (for which there are 329 subjects with valid data) as outcomes. We employ stability selection [20]—introduced by Bühlmann *et al.*—to obtain the risk SNPs. For each outcome, we perform 100 simulations. In each simulation, we first subsample half of the samples from the original data, and then we incorporate EDPP with the solver for Lasso, to solve the Lasso problems at a sequence

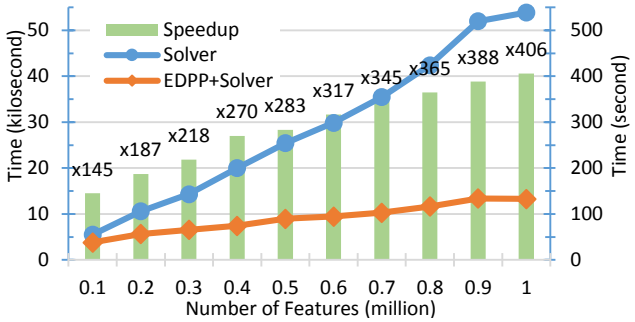

Fig. 3. Comparison of Lasso with and without the EDPP rule. Run times in units of kiloseconds are reported for each Solver (Lasso), and in units of seconds for EDPP+Solver.

of 100 parameter values equally spaced on the logarithmic scale of λ/λ_{max} from 1.0 to 0.05. The selection probabilities for each SNP are recorded and we present the top 10 selected SNPs for each outcome in **Table 1** & **Table** 2.

5. CONCLUSION

In this paper, we investigate Alzheimer's disease genetic risk factors associated with imaging phenotypes using the ADNI whole genome sequence data. Because of the sheer scale of the genomic data, we employ Lasso regression to identify the most relevant SNPs and use the enhanced DPP screening rule to identify and remove irrelevant features from the optimization. Experiments demonstrate the efficiency and effectiveness of the proposed approach.

Among the top SNPs, that survive stability selection, in **Table 1**, are several genes already implicated in AD risk or risk for other neuropsychiatric disorders. *APOE* is a top AD risk gene, and also ranks among the top predictors of baseline entorhinal and hippocampal volumes, in line with much prior work showing APOE effects on temporal lobe structures not just in old age but also in children and adolescents [21].

APOE4 carriers have a roughly 3-fold increase in AD risk per adverse copy of the gene carried, and the selection of APOE4 among the genes associated with brain measures offers strong a priori evidence of validity. APOE does not feature among the top genes that predict rates of brain tissue loss over time, although baseline data reflect the lifetime accumulation of tissue loss, which is perhaps more affected by APOE than measures over a short interval.

The 4th ranked gene related to hippocampal volume change is *CACNA1C*, a gene involved in calcium channel function that has already been associated with DTI measures in the hippocampus, and is among the top genes associated with anxiety, depression, and obsessive compulsive disorder [22, 23]. Intriguingly, another of the top genes related to hippocampal volume change is the beta-secretase gene, *BACE2*, is a close homolog of *BACE1*, which encodes a key enzyme involved in the cellular pathways of AD.

Rather than interpret the associations as plausible or not *post hoc*, the next step in this work is to see if the associations with regional volumes replicate in independent samples. Very little WGS data is publicly available, especially from cohorts assessed with brain imaging. Recent genome-wide association studies by the ENIGMA and CHARGE consortia identified SNPs consistently associated with hippocampal volumes in over 30,000 people worldwide [5, 24]. As such a screen for enrichment would be valuable between the WGS SNPs selected here and SNP sets that predict regional volumes in ENIGMA.

6. ACKNOWLEDGMENTS

This research was supported in part by NIH 'Big Data to Knowledge' (BD2K) Center of Excellence grant U54 EB020403, funded by a cross-NIH consortium including NIBIB and NCI.

Table 1. Top 10 SNPs associated with baseline volumes.

	EC baseline				HIPP baseline			
	Chr	Pos	RS_ID	Gene	Chr	Pos	RS_ID	Gene
Rank 1	chr6	72869836	rs201890142	RIMS1	chr10	71969989	rs12412466	PPA1
Rank 2	chr19	15136345	unknown	unknown	chr19	45411941	rs429358	APOE
Rank 3	chr1	142555416	rs6672189	unknown	chr11	11317240	rs10831576	GALNT18
Rank 4	chr19	45411941	rs429358	APOE	chr4	49147785	rs151073945	unknown
Rank 5	chr2	96630311	rs369756382	ANKRD36C	chr8	145158607	rs34173062	MAF1
Rank 6	chr2	95552943	rs199536016	LOC442028	chr6	168107162	rs71573413	unknown
Rank 7	chr2	95552986	rs200710055	LOC442028	chrX	143403234	rs4825209	unknown
Rank 8	chr1	142545571	LOC442028	unknown	chr2	231846840	rs4973360	unknown
Rank 9	chr13	94122100	rs76403280	GPC6	chr14	20710095	rs35055545	OR11H4
Rank 10	chr5	97913219	rs202036446	unknown	chr6	69775654	rs2343398	BAI3

Table 2. Top 10 SNPs associated with volume changes.

	EC changes				HIPP changes			
	Chr	Pos	RS_ID	Gene	Chr	Pos	RS_ID	Gene
Rank 1	chr7	7333294	rs1317198	unknown	chr15	23060281	rs11636690	NIPA1
Rank 2	chr10	9216956	rs1149952	unknown	chr21	42613255	rs74977559	BACE2
Rank 3	chr6	115278412	rs146156795	unknown	chr9	91366940	rs79543088	unknown
Rank 4	chr10	117854524	rs2530339	GFRA1	chr12	2342248	rs7303977	CACNA1C
Rank 5	chr8	6557130	rs2912047	LOC100507530	chr6	169666147	rs6605518	unknown
Rank 6	chr12	23117263	rs12581794	unknown	chr18	48740822	rs34794713	unknown
Rank 7	chr16	77841030	rs16946521	VAT1L	chr13	102307271	rs9518474	ITGBL1
Rank 8	chr3	175784869	rs9845573	unknown	chrX	2372296	rs7889210	DHRSX
Rank 9	chr4	7494498	rs4308363	SORCS2	chr4	174673036	rs12646029	LOC101928478
Rank 10	chr13	102616850	rs17502999	FGF14	chr4	170638416	rs149287207	CLCN3

7. REFERENCES

[1] L. E. Hebert, J. Weuve, P. A. Scherr, and D. A. Evans, "Alzheimer disease in the United States (2010–2050) estimated using the 2010 census," *Neurology*, vol. 80, pp. 1778-1783, 2013.

[2] Alzheimer's Disease Association and M. Albert, *The use of MRI and PET for clinical diagnosis of dementia and investigation of cognitive impairment: a consensus report*: Alzheimer's Association, 2003.

[3] Y. Huang and L. Mucke, "Alzheimer mechanisms and therapeutic strategies," *Cell*, vol. 148, pp. 1204-1222, 2012.

[4] K. Bettens, K. Sleegers, and C. Van Broeckhoven, "Genetic insights in Alzheimer's disease," *The Lancet Neurology*, vol. 12, pp. 92-104, 2013.

[5] J. L. Stein, S. E. Medland, A. A. Vasquez, D. P. Hibar, R. E. Senstad, A. M. Winkler, *et al.*, "Identification of common variants associated with human hippocampal and intracranial volumes," *Nature Genetics*, vol. 44, pp. 552-561, 2012.

[6] J.-C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez, *et al.*, "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease," *Nature Genetics*, 2013.

[7] O. Kohannim, D. P. Hibar, J. L. Stein, N. Jahanshad, X. Hua, P. Rajagopalan, *et al.*, "Discovery and replication of gene influences on brain structure using LASSO regression," *Frontiers in Neuroscience*, vol. 6, 2012.

[8] D. P. Hibar, O. Kohannim, J. L. Stein, M.-C. Chiang, and P. M. Thompson, "Multilocus genetic analysis of brain images," *Frontiers in Genetics*, vol. 2, 2011.

[9] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, *et al.*, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, pp. 2156-2158, 2011.

[10] ADNI, "Introduction and Procedures for Accessing Data from Whole Genome Sequencing of ADNI," 2013.

[11] Illumina, "Instructions for using BaseSpace (15044182 E)," 2014.

[12] P. D. Sasieni, "From genotypes to genes: doubling the sample size," *Biometrics*, pp. 1253-1261, 1997.

[13] International HapMap Consortium, "A haplotype map of the human genome," *Nature*, vol. 437, pp. 1299-1320, 2005.

[14] B. Ewing and P. Green, "Base-calling of automated sequencer traces using phred. II. Error probabilities," *Genome Research*, vol. 8, pp. 186-194, 1998.

[15] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267-288, 1996.

[16] J. Liu, S. Ji, and J. Ye, "SLEP: Sparse learning with efficient projections," *Arizona State University*, vol. 6, 2009.

[17] L. El Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination in sparse supervised learning," *CoRR*, 2010.

[18] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, *et al.*, "Strong rules for discarding predictors in lasso - type problems," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, pp. 245-266, 2012.

[19] J. Wang, P. Wonka, and J. Ye, "Lasso Screening Rules via Dual Polytope Projection," *Journal of Machine Learning Research (to appear)*, 2014.

[20] N. Meinshausen and P. Bühlmann, "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, pp. 417-473, 2010.

[21] P. Shaw, J. P. Lerch, J. C. Pruessner, K. N. Taylor, A. B. Rose, D. Greenstein, *et al.*, "Cortical morphology in children and adolescents with different apolipoprotein E gene polymorphisms: an observational study," *The Lancet Neurology*, vol. 6, pp. 494-500, 2007.

[22] NCBI. *CACNA1C*. Available: http://www.ncbi.nlm.nih.gov/gene/775

[23] J. Strohmaier, M. Amelang, L. A. Hothorn, S. H. Witt, V. Nieratschker, D. Gerhard, *et al.*, "The psychiatric vulnerability gene CACNA1C and its sex-specific relationship with personality traits, resilience factors and depressive symptoms in the general population," *Molecular Psychiatry*, vol. 18, pp. 607-613, 2012.

[24] D. P. Hibar, J. L. Stein, M. Renteria, A. Arias-Vasquez, S. Desrivieres, N. Jahanshad, *et al.*, "Common genetic variants influence human subcortical brain structures," *Nature (in press)*, 2014.