Behav Genet (2011) 41:120–124
DOI 10.1007/s10519-011-9443-8

ORIGINAL RESEARCH

An Application of the Elastic Net for an Endophenotype Analysis

Dean Palejev · Wookyeon Hwang · Nicole Landi · Maria Eastman · Stephen J. Frost · Robert K. Fulbright · Judith R. Kidd · Kenneth K. Kidd · Graeme F. Mason · W. Einar Mencl · Carolyn Yrigollen · Kenneth R. Pugh · Elena L. Grigorenko

Received: 18 December 2010 / Accepted: 23 December 2010 / Published online: 13 January 2011

Abstract We provide an illustration of an application of the elastic net to a large number of common genetic variants in the context of the search for the genetic bases of an endophenotype conceivably related to individual differences in learning. GABA concentration in the occipital cortex, a critical area for reading, was obtained in a group ($n = 76$) of children aged 6–10 years. Two extreme groups, high and low, were selected for genotyping with the 650Y Illumina array chip (Ilmn650Y). An elastic net approach was applied to the resulting SNP dataset; 100 SNPs were identified for each chromosome as "interesting" based on having the highest absolute value coefficients. The analyses highlighted chromosomes 15 and 20, which contained 55 candidate genes. The STRING partner analyses of the associated proteins pointed to a number of related genes, most notably, GABA and NTRK receptors.

Keywords Genetic association analysis · GABA concentration · Elastic net · Occipital cortex · Reading acquisition

D. Palejev · N. Landi · M. Eastman · R. K. Fulbright · J. R. Kidd · K. K. Kidd · G. F. Mason · K. R. Pugh · E. L. Grigorenko
Yale University, New Haven, CT, USA

W. Hwang
Home Entertainment Company, LG Electronics, Seoul, South Korea

N. Landi · S. J. Frost · W. E. Mencl · K. R. Pugh
Haskins Laboratories, New Haven, CT, USA

C. Yrigollen
University of California, Davis, CA, USA

E. L. Grigorenko (✉)
Child Study Center, Yale University, 230 South Fontage Road, New Haven, CT 06519-1124, USA
e-mail: elena.grigorenko@yale.edu

E. L. Grigorenko
Columbia University, New York, NY, USA

E. L. Grigorenko
Moscow State University, Moscow, Russia

Advances in the field of quantitative genetics of complex behavioral traits have generated heritability estimates from elaborate statistical models. Notably, the majority of the traits capturing the diagnosis of learning disabilities (LD) and their associated psychological processes (e.g., working memory) demonstrate statistically significant and moderately-high heritability estimates. However, translating these estimates into specific molecular-genetic findings that can explain such high heritability has been difficult. This difficulty, referred to as the puzzle of "missing heritability" (Manolio et al. 2009), may arise from a variety of problems, ranging from mistaken assumptions about the simplicity of the genetic mechanisms underlying the etiology of complex behavior traits to the limited assortment of analytical machinery that has been applied so far to the analyses of molecular-genetic data. The solutions to these problems, unfortunately, are not obvious. And, at least at this stage, the most promising approach to finding solutions is in the continuing attempts to enhance both the analytical methods with which molecular-genetic data are treated and the theoretical models capturing these data. In this brief report, we illustrate an application of an analytical approach that might be helpful in working with complex phenotypes. Specifically, we implemented the elastic net (Zou and Hastie 2005) approach. This approach allows the consideration of multiple SNPs jointly, improving both the

power and the prediction performance of genetic association analysis. Elastic net is thought to be particularly effective (Zou and Hastie 2005) at capturing correlated predictors such as SNPs that are in linkage disequilibrium. The target phenotype here is an endophenotype that is brain-based and captures the dynamics of the distribution of GABA—a major neurotransmitter that is thought to be critical to learning and memory (Reis et al. 2009).

Twenty children between the ages of 6 and 10 years participated in this research as part of a large scale longitudinal study of the neurobiological coordinates of reading ability. These 20 participants were selected from a larger pool of 76 participants (mean age = 7.68, SD = .7). These twenty represent extreme cases (top 10 [mean age = 7.6. SD = .68] and bottom 10 [mean age = 8.1, SD = .82]) with respect to GABA concentrations in the occipital cortex ($mean_{ALL}$ = .177, SD_{ALL} = .037; $mean_{HIGH}$ = .210, SD_{ALL} = .021; and $mean_{LOW}$ = .145, SD_{ALL} = .012, measured as described below). This selection approach improved the power of the association analyses, although the continuous phenotype of GABA concentration was used. The majority of the children were Caucasian, with one child of mixed race (Caucasian and African American) in the high GABA group, one Asian child in the low GABA group and one child whose parents did not report race. Participants were also similar with respect to their social-economic status (SES), with mothers having completed 6.2 (SD = .63) years of post-elementary education in the low GABA group and 6.0 (SD = .86) years of post-elementary education in the high GABA group (these did not differ significantly, $p > .1$). The participants represented a range of reading and cognitive ability. All participants donated DNA specimens, either whole blood or saliva (collected using Oragene DNA kits, DNA Genotek, Ottawa, Ontario).

Participants were placed supine in a 4T imaging spectrometer (Bruker Instruments, Billerica, MA). A 7-cm ^{1}H-tuned surface coil was used for all measurements. Prior to the MRS measurement, a gradient-echo MRI of the brain of each subject was obtained for subject positioning and the selection of a volume in the occipital region of the extrastriate cortex, a critical area for reading. The magnetic field homogeneity in the volume was optimized using an automated FASTERMAP shim routine (Gruetter 1993; Shen et al. 1997), and the pulse powers were calibrated. Spectroscopic localization was achieved with ISIS, outer volume suppression pulses, and selective excitation with a 68-ms spin-echo (Rothman et al. 1993). CHESS (Chemical Shift-Selective water Suppression) was used to achieve water suppression. Spectral editing separated the GABA C4 resonance at 3.0 ppm from overlapping resonances by applying a 20-ms Gauss-shaped DANTE pulse to the 1.9 ppm C3 resonance on half the scans and at 1.3 ppm, away from the C3 resonance, on the other scans.

The GABA resonance in the NMR-edited spectrum was integrated and compared with the integrated resonance of creatine obtained in the same measurement. Creatine is a stable compound in the brain with an average cortical concentration of 9.6 mM. The time-domain data were zero-filled to 8 K and apodized with a −2 Hz Lorentzian and 6 Hz Gaussian windowing before Fourier transformation. In the edited spectrum, GABA was integrated over a .30 ppm bandwidth centered at about 3.00 ppm. To reduce distortion from overlapping resonances, the intensity of the creatine resonance was integrated over a .20 ppm bandwidth. Co-edited macromolecules were measured in six people and found to be undetectable at this echo time and with the editing scheme that was used. The concentration of GABA was calculated using the formula $[GABA] = (G/Cr) \times (EE) \times (3/2) \times [Cr]$, where G is the integral in the edited spectrum, Cr is the creatine integral, EE is the correction for loss of intensity due to imperfect editing efficiency, 3/2 is the creatine to GABA proton ratio, and [Cr] is the assumed creatine concentration in the human cortex. The concentration index captured both intra and extracellular concentrations of GABA.

The participants' DNA was genotyped using Illumina 650Y (Ilmn650Y) arrays. A two-stage genetic association analysis was performed. In the preliminary stage the genetic data were standardized; 500 SNPs per chromosome with the largest absolute correlation values with the GABA concentration were chosen for subsequent analyses. The second stage consisted of the association analyses using the elastic net approach. We randomly selected 15 training data sets and 5 test data sets, which were used to calculate validation errors and test errors, respectively. Five-fold cross validations for training sets were used to select the two shrinkage parameters (Hastie et al. 2001). The first parameter for the one-norm penalty plays a role in variable selection, whereas the second parameter for the two-norm penalty captures correlated predictors at the same time (Tibshirani 1996; Zou and Hastie 2005). Five-fold cross validation split the 15 training data sets into 5 subsets evenly. One of the subsets was held for calculating a validation error, whereas the rest of the subsets were used to train an elastic net model. As a result, 5 validation data sets were obtained and 5 validation errors were calculated rotating the 5 subsets. Additionally, 8 different values were considered for the first parameter and 5 different values for the second parameter. Then the average of the 5 validation errors was calculated given each combination of the two parameters. The combination of the two parameters giving the minimum average error was chosen. Next, a validation data set was selected to give the minimum test error among the 5 validation data sets. An elastic net model was constructed based on training data that corresponded to the combination of the two parameters and the validation data

set. Finally, the top SNPs with the largest 50 positive coefficients and the bottom SNPs with the smallest 50 negative coefficients were considered as the most "interesting" SNPs. Thus, in total, the analyses generated 2,300 SNPs labeled as interesting. Of these SNPs, 901 were located within various genes and 1,399—in other regions; to annotate these genes, FunctSNP (Goodswen et al. 2010) was used. The absolute values of these coefficients are shown in Fig. 1.

Figure 2 presents (a) the number of interesting SNPs that are located within 10 kb of another interesting SNP; and (b) the number of genes that were tagged by 4 or more interesting SNPs (the range was 1–8 SNPs per gene). As expected the genes with the largest number of interesting SNPs were larger in physical size among the 584 tagged genes. However, there was no correlation between the length of the chromosome and the number of genes with 4 or more interesting SNPs. Based on the information presented in Fig. 2, chromosomes 15 and 20 were repeatedly found among the chromosomes with the highest counts in each category. In addition, the distribution of the absolute values of coefficients for interesting SNPs on chromosome 15 was significantly different from the distributions for other chromosomes as determined by the Kolmogorov–Smirnov Test. The distribution of these coefficients for chromosome 20 was significantly different from the distributions of most other chromosomes. Based on the inspection of all these sources of information, it appeared

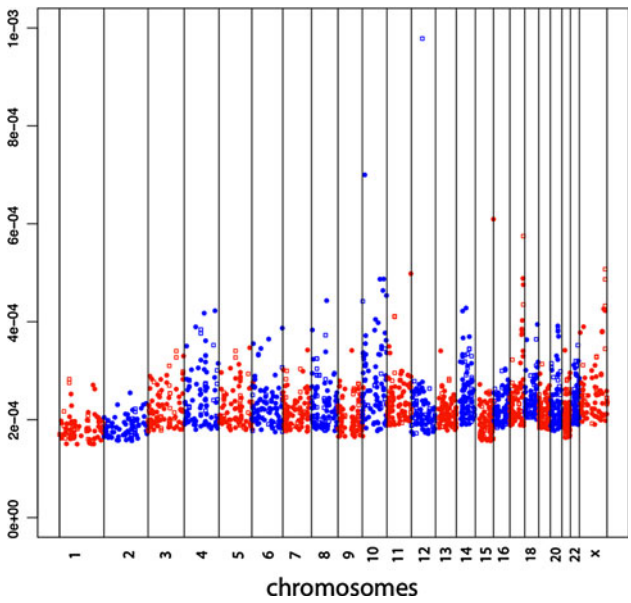

Fig. 1 The absolute values of the elastic net coefficients per chromosome. Open squares indicate SNPs that are within 10 kb of another interesting SNP; solid circles show individual SNPs that have no other SNPs within the 10 kb interval. Chromosomes 17, 19, and 21 are displayed, but their labels are omitted due to typographical constraints

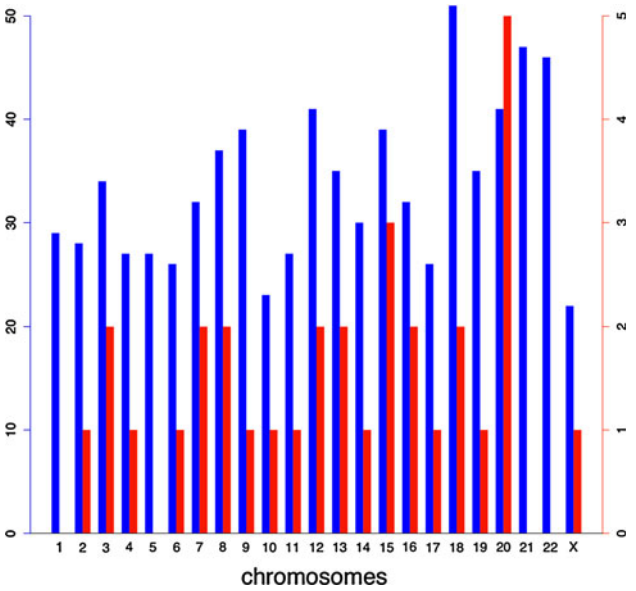

Fig. 2 *Left bar* of each pair of bars: number of interesting SNPs that are within 10 kb of another interesting SNP (marked by the scale on the *left*). *Right bar*: Number of genes with 4 + interesting SNPs (marked by the scale on the *right*). Single bar per chromosome indicates that the right bar corresponds to a value of 0

that chromosomes 15 and 20 contained the most noteworthy findings.

On chromosome 15, interesting SNPs were clustered within 29 genes, and on chromosome 20—within 26 genes (see Table 1). Two genes had 7 SNPs each captured by the elastic net: *NTRK3* (chromosome 15) and *PLCB1* (chromosome 20). Of note also is that the net encapsulated two GABA A receptor genes, *GABAB3* and *GABAG3* (chromosome 15). To investigate connections between these genes, each gene's network was analyzed with STRING (http://string-db.org/), a database integrating seven different types of evidence for physical and/or functional interactions between proteins: experimental evidence, phylogenetic profiles ('co-occurrence'), gene fusion/fission, synteny ('neighborhood'), coexpression, text mining and a data set called 'database' developed by the curators of the database (Jensen et al. 2009). When the functional partners were considered for all 55 genes, a number of observations were made. Specifically, the networks repeatedly engaged 20 GABA-related genes. Moreover, they engaged 3 members of the neurotrophic tyrosine receptor kinase (NTRK) family. Of note also is that *BDNF* appeared in the engaged networks 5 times (both on chromosomes 15 and 20).

This brief report intends to demonstrate the utilization of extreme groups derived from a particular endophenotype (in this case—the intra- and extra-cellular concentration of GABA in the occipital cortex), that has been shown to be associated with the trait of interest, in this case, a disruption of learning (Tabata and Kano 2010) in the context of the

Table 1 Genes tagged by the elastic net on chromosomes 15 and 20

Chromosome 15			Chromosome 20		
# of SNPs	Gene	Locus	# of SNPs	Gene	Locus
7	NTRK3	15q25.3	7	PLCB1	20p12.3
5	AKAP13	15q25.3	5	TOX2	20q13.12
4	ARNT2	15q25.1	4	EYA2	20q13.12
3	GABRG3	15q12	4	MACROD2	20p12.1
3	LINS1	15q26.3	4	PAK7	20p12.2
3	ONECUT1	15q21.3	3	BFSP1	20p12.1
2	KIAA1199	15q25.1	3	RIN2	20p11.23
2	SLC27A2	15q21.2	3	RNF24	20p13
1	ALPK3	15q25.3	2	SLC24A3	20p11.23
1	ARHGAP11A	15q13.3	2	TSHZ2	20q13.2
1	ATP10A	15q12	1	BPI	20q11.23
1	CGNL1	15q21.3	1	C20orf186	20q11.21
1	CHP	15q15.1	1	CDH4	20q13.33
1	EHD4	15q15.1	1	COL9A3	20q13.33
1	GABRB3	15q12	1	DTD1	20p11.23
1	INO80	15q15.1	1	EPB41L1	20q11.23
1	LASS3	15q26.3	1	GNRH2	20p13
1	MCTP2	15q26.2	1	HAO1	20p12.3
1	MEGF11	15q22.31	1	HNF4A	20q13.12
1	MYO5C	15q21.2	1	KIF16B	20p12.1
1	PAQR5	15q23	1	MYBL2	20q13.12
1	RYR3	15q14	1	PCSK2	20p12.1
1	SH3GL3	15q25.2	1	PTPRT	20q12-q13.11
1	SLC28A1	15q25.3	1	SPTLC3	20p12.1
1	ST8SIA2	15q26.1	1	TOP1	20q12
1	TJP1	15q13.1	1	ZHX3	20q12
1	TLN2	15q22.2			
1	USP8	15q21.2			
1	WDR72	15q21.3			

analyses of common genetic variants. The utilization of the elastic net provides an alternative to more traditional association analyses, and is particularly suitable for small samples sizes. Although the results presented here are preliminary and need further validation and interrogation, it is quite striking that the endophenotype of GABA concentration in a single area of the brain revealed a rather consistent and meaningful pattern of association with genes whose function might reasonably be biologically related to this indicator of brain activity.

Acknowledgements Preparation of this article was supported in part by the following research grants from the National Institutes of Health: HD048830, HD001994, HD053409, AA009379, GM057672, HD052120, DC007665, and TW008274. Grantees undertaking such projects are encouraged to express their professional judgment freely. Therefore, this article does not necessarily reflect the position or policies of the National Institutes of Health, and no official endorsement should be inferred. We are thankful to Ms. Mei Tan for her editorial assistance.

References

Goodswen SJ, Gondro C, Watson-Haigh NS, Kadarmideen HN (2010) FunctSNP: an R package to link SNPs to functional knowledge and dbAutoMaker: a suite of Perl scripts to build SNP databases. BMC Bioinform 11:311

Gruetter R (1993) Automatic, localized in vivo adjustment of all first- and second-order shim coils. Magn Reson Med 29:804–811

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer, New York, NY

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37: D412–D416

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Feinberg AP, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Reis HJ, Guatimosim C, Paquet M, Santos M, Ribeiro FM, Kummer A, Schenatto G, Salgado JV, Vieira LB, Teixeira AL, Palotás A (2009) Neuro-transmitters in the central nervous system and their implication in learning and memory processes. Curr Med Chem 16:796–840

Rothman DL, Petroff OA, Behar KL, Mattson RH (1993) Localized 1H NMR measurements of gamma-aminobutyric acid in human brain in vivo. Proc Nat Acad Sci 90:5662–5666

Shen J, Rycyna RE, Rothman D (1997) Improvements on an in vivo automatic shimming method. Magn Reson Med 38:834–839

Tabata T, Kano M (2010) GABAB receptor-mediated modulation of metabotropic glutamate signaling and synaptic plasticity in central neurons. Adv Pharmacol 58:149–173

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc 1:147–169

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc 67:301–320