*Submitted to the Annals of Applied Statistics*

# SUPPLEMENTARY TO "SIMULTANEOUS SELECTION OF MULTIPLE IMPORTANT SINGLE NUCLEOTIDE POLYMORPHISMS IN FAMILIAL GENOME WIDE ASSOCIATION STUDIES DATA"

By Subhabrata Majumdar, Saonli Basu and Snigdhansu Chatterjee

## APPENDIX A: PROOF OF THEORETICAL RESULTS

Proof of Theorem 3.2. Define $c_{q,\infty} = q^{\text{th}}$ quantile of $\mathbb{T}_0$. Now following assumption (E1),

$$\begin{aligned} c_q(\mathbb{E}_*) &= \inf_{\boldsymbol{\theta}}\{E(\boldsymbol{\theta}, [\hat{\boldsymbol{\theta}}]) : \mathbb{F}_* \geq q\} \\ &= \inf_{\boldsymbol{\theta}}\{E(a_n(\boldsymbol{\theta} - \boldsymbol{\theta}_0), [a_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)]) : a_n(\mathbb{F}_* - \boldsymbol{\theta}_0) \geq q\} \end{aligned}$$

where $\mathbb{F}_*$ is the probability distribution function of $E(\hat{\boldsymbol{\theta}}, [\hat{\boldsymbol{\theta}}])$. Part 1 is proved following assumptions (P2) and (E3).

Now if $\mathcal{M}$ is adequate, following assumption (E1),

$$E(\hat{\boldsymbol{\theta}}_m, [\hat{\boldsymbol{\theta}}]) = E(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0, [\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0]) \tag{A.1}$$

Decompose the first argument as

$$\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta} = (\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{A.2}$$

By definition, $\hat{\theta}_{mj} - \hat{\theta}_j = 0$ if $j \in \mathcal{S}$, else equals $\theta_{0j} - \hat{\theta}_j$. Thus for the first summand in (A.2) we have

$$\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}} = O_P(1/a_n)$$

Going back to (A.1), this implies

$$|E(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0, [\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0]) - E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, [\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0])| < O_P(a_n^{-\alpha})$$

using lipschitz continuity in assumption (E2), i.e

$$|E(\hat{\boldsymbol{\theta}}_m, [\hat{\boldsymbol{\theta}}]) - E(\hat{\boldsymbol{\theta}}, [\hat{\boldsymbol{\theta}}])| < O_P(a_n^{-\alpha})$$

again using (E1). Part 2 now follows.

For part 3, we apply (E1) to get

$$E(\hat{\boldsymbol{\theta}}_m, [\hat{\boldsymbol{\theta}}]) = E(a_n(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0), [a_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)]) \tag{A.3}$$

And decompose the first argument as

$$a_n(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0) = a_n(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m) + a_n(\boldsymbol{\theta}_m - \boldsymbol{\theta}_0) \tag{A.4}$$

Since $\mathcal{M}$ is inadequate, $\theta_{mj} \neq \theta_{0j}$ when $j \notin \mathcal{S}$. So $\|a_n(\boldsymbol{\theta}_m - \boldsymbol{\theta}_0)\| \uparrow \infty$ as $a_n \uparrow \infty$. Applying (E4) now proves part 3. □

Proof of Theorem 3.3. The proof is fairly similar to that of theorem 3.2, so we give a sketch of it. For the full model, the bootstrap is consistent, i.e. $a_n(\hat{\boldsymbol{\theta}}_* - \boldsymbol{\theta}_0)$ and $(a_n/\tau_n)(\hat{\boldsymbol{\theta}}_{r*} - \hat{\boldsymbol{\theta}}_*)$ converge to same weak limit in probability, following theorems 2.2 and 2.3 in Majumdar and Chatterjee (2017). Specifically, conditions (A1)-(A6) in Majumdar and Chatterjee (2017) ensure condition (P2) in our paper through theorem 2.2 therein, following which theorem 2.3 ensures that when (A1)-(A6) are satisfied, bootstrap consistency holds. The definition of $\hat{\boldsymbol{\theta}}_m$ now means that $a_n(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m)$ and $(a_n/\tau_n)(\hat{\boldsymbol{\theta}}_{rm} - \hat{\boldsymbol{\theta}}_m)$ converge to the same weak limit in probability for any model $\mathcal{M}$. A similar approach as the proof of parts 2 and 3 of theorem 3.2 now follows, with an additional term corresponding to bootstrap estimates in (A.2) and (A.4). □

## APPENDIX B: OUTPUTS FOR MCTFR DATA ANALYSIS

Each table gives the $90^{\text{th}}$ percentile $e$-values, which are plotted in figures 2, 3, and 4 in main paper, of SNPs analyzed in the gene. Column 'Association' is obtained from the sign of the SNP coefficient in the full model.

| SNP name | Location | $e$-value | Association |
|---|---|---|---|
| rs16859227 | 46250605 | 0.89 | + |
| rs572227 | 46251393 | 0.13 | - |
| rs534459 | 46256805 | 0.24 | + |
| rs2119183 | 46272806 | 0.92 | - |
| rs502038 | 46280318 | 0.58 | + |
| rs1808851 | 46311447 | 0.00 | + |
| rs279856 | 46317923 | 0.00 | - |
| rs3775282 | 46321863 | 0.86 | - |
| rs279841 | 46340763 | 0.75 | + |
| rs10805145 | 46358331 | 0.73 | - |
| rs13152740 | 46381221 | 0.86 | - |

Table 1

*SNPs for GABRA2, chr4, position 46243548 - 46390039; e-value cutoff 0.72*

| SNP name | Location | *e*-value | Association |
|---|---|---|---|
| rs17027299 | 99078105 | 0.84 | - |
| rs9307222 | 99101051 | 0.76 | - |
| rs10006414 | 99101401 | 0.49 | + |
| rs9994641 | 99101605 | 0.48 | + |
| rs13134014 | 99104879 | 0.75 | - |
| rs6820691 | 99105055 | 0.76 | + |
| rs6820913 | 99125659 | 0.81 | + |
| rs6532729 | 99146436 | 0.67 | - |
| rs13150538 | 99152631 | 0.49 | - |
| rs17027380 | 99157450 | 0.63 | - |
| rs17494998 | 99160699 | 0.41 | + |
| rs549467 | 99172232 | 0.81 | + |
| rs2034677 | 99187874 | 0.62 | + |
| rs12508445 | 99190653 | 0.01 | - |
| rs10003496 | 99197839 | 0.81 | + |
| rs10005811 | 99208603 | 0.02 | + |
| rs603215 | 99214851 | 0.78 | - |
| rs433146 | 99229839 | 0.87 | - |
| rs17027456 | 99235747 | 0.31 | - |
| rs17561798 | 99235941 | 0.85 | + |
| rs10516428 | 99237439 | 0.45 | - |
| rs6532731 | 99251006 | 0.80 | + |
| rs7694221 | 99260423 | 0.90 | + |
| rs10028330 | 99268949 | 0.70 | - |
| rs10022047 | 99296818 | 0.49 | + |
| rs17027523 | 99298979 | 0.05 | + |
| rs17027530 | 99303633 | 0.69 | + |
| rs3775540 | 99304544 | 0.23 | - |
| rs3756088 | 99309404 | 0.89 | - |
| rs13103626 | 99317251 | 0.75 | + |
| rs10516430 | 99337881 | 0.62 | + |
| rs9884594 | 99359318 | 0.68 | - |
| rs12503056 | 99369061 | 0.63 | + |
| rs2004316 | 99381148 | 0.43 | - |
| rs4303985 | 99399748 | 0.87 | - |
| rs4414961 | 99403784 | 0.86 | - |
| rs12509267 | 99407299 | 0.80 | + |
| rs6838913 | 99408106 | 0.84 | - |
| rs4374629 | 99411783 | 0.85 | + |
| rs4527483 | 99421741 | 0.89 | + |
| rs10009693 | 99423280 | 0.90 | - |
| rs10023791 | 99425353 | 0.88 | + |
| rs955931 | 99428163 | 0.88 | - |
| rs17027628 | 99428608 | 0.85 | - |

TABLE 2

*SNPs for ADH genes, chr4, position 99070977 - 99435737; e-value cutoff 0.225*

| SNP name | Location | *e*-value | Association |
|---|---|---|---|
| rs2000371 | 154011024 | 0.39 | - |
| rs9371718 | 154011615 | 0.08 | - |
| rs12211203 | 154016936 | 0.63 | - |
| rs1937600 | 154017197 | 0.02 | - |
| rs9397637 | 154022718 | 0.00 | + |
| rs1937590 | 154036895 | 0.63 | + |
| rs12662873 | 154040810 | 0.18 | + |
| rs12661209 | 154044112 | 0.84 | - |
| rs1316368 | 154055754 | 0.00 | - |
| rs1937587 | 154060023 | 0.27 | - |
| rs6921403 | 154063906 | 0.00 | - |
| rs1937580 | 154076643 | 0.00 | + |
| rs1937645 | 154082228 | 0.00 | + |
| rs1892361 | 154099619 | 0.00 | - |
| rs1937633 | 154104857 | 0.04 | - |
| rs1937631 | 154105011 | 0.00 | - |
| rs12527197 | 154107836 | 0.02 | + |
| rs1892360 | 154111701 | 0.74 | - |
| rs1892359 | 154112042 | 0.65 | - |
| rs1892356 | 154112263 | 0.56 | + |
| rs1937622 | 154113139 | 0.54 | - |
| rs10485258 | 154113409 | 0.72 | - |
| rs1937619 | 154114583 | 0.58 | - |
| rs1748289 | 154121980 | 0.77 | - |
| rs1781619 | 154135968 | 0.64 | - |
| rs652051 | 154139344 | 0.74 | + |
| rs10485262 | 154140199 | 0.69 | - |
| rs9371312 | 154145492 | 0.81 | + |
| rs1332849 | 154151117 | 0.48 | - |
| rs9371749 | 154153369 | 0.28 | + |
| rs9285539 | 154154532 | 0.08 | + |
| rs9322439 | 154156250 | 0.07 | + |
| rs11752884 | 154159710 | 0.25 | - |
| rs4869813 | 154173845 | 0.13 | + |
| rs4870241 | 154174963 | 0.00 | - |
| rs9384156 | 154186720 | 0.13 | + |
| rs2065139 | 154192175 | 0.89 | - |
| rs689219 | 154198820 | 0.00 | - |
| rs9371761 | 154202578 | 0.20 | - |
| rs12199858 | 154204327 | 0.00 | + |
| rs9371762 | 154213973 | 0.00 | - |
| rs612450 | 154214357 | 0.00 | - |
| rs9384159 | 154219177 | 0.00 | + |
| rs6938958 | 154220427 | 0.00 | - |
| rs581564 | 154221214 | 0.00 | + |
| rs12202611 | 154237443 | 0.76 | - |
| rs4870255 | 154237937 | 0.88 | - |

TABLE 3

*SNPs for OPRM1, chr6, position 154010496 - 154246867; e-value cutoff 0.225*

| SNP name | Location | *e*-value | Association |
|---|---|---|---|
| rs10872828 | 133525348 | 0.72 | - |
| rs9419702 | 133531153 | 0.09 | - |
| rs7083395 | 133532269 | 0.77 | + |
| rs9419624 | 133534822 | 0.06 | + |
| rs7906770 | 133536902 | 0.28 | - |
| rs9419569 | 133541881 | 0.06 | + |
| rs9419629 | 133543210 | 0.06 | + |
| rs7093241 | 133556596 | 0.72 | - |
| rs9419649 | 133561098 | 0.91 | - |

TABLE 4

*SNPs for CYP2E1, chr10, position 133520406 - 133561220; e-value cutoff 0.72*

| SNP name | Location | *e*-value | Association |
|---|---|---|---|
| rs7398343 | 111774068 | 0.34 | - |
| rs7297186 | 111778178 | 0.36 | + |
| rs3803167 | 111785586 | 0.00 | + |
| rs10219736 | 111788402 | 0.00 | - |
| rs16941437 | 111793039 | 0.00 | - |
| rs3742004 | 111798553 | 0.75 | + |

TABLE 5

*SNPs for ALDH2, chr12, position 111766887 - 111817529; e-value cutoff 0.72*

| SNP name | Location | *e*-value | Association |
|---|---|---|---|
| rs4646312 | 19948337 | 0.41 | - |
| rs165656 | 19948863 | 0.22 | - |
| rs165722 | 19949013 | 0.24 | + |
| rs2239393 | 19950428 | 0.50 | + |
| rs4680 | 19951271 | 0.60 | + |
| rs4646316 | 19952132 | 0.81 | - |
| rs165774 | 19952561 | 0.72 | - |
| rs174699 | 19954458 | 0.07 | + |
| rs165599 | 19956781 | 0.58 | - |
| rs165728 | 19957023 | 0.02 | - |
| rs165815 | 19959473 | 0.00 | + |
| rs5993891 | 19959746 | 0.04 | - |
| rs887199 | 19961955 | 0.04 | - |
| rs2239395 | 19962203 | 0.07 | + |
| rs2518824 | 19962963 | 0.59 | + |

TABLE 6

*SNPs for COMT, chr22, position 19941607 - 19969975; e-value cutoff 0.72*

| SNP name | Location | *e*-value | Association |
|---|---|---|---|
| rs27072 | 1394522 | 0.87 | + |
| rs40184 | 1395077 | 0.78 | - |
| rs11564771 | 1398797 | 0.80 | - |
| rs11133767 | 1401580 | 0.79 | + |
| rs6869645 | 1404548 | 0.82 | + |
| rs3776512 | 1407116 | 0.84 | + |
| rs6347 | 1411412 | 0.83 | - |
| rs27048 | 1412645 | 0.90 | - |
| rs2042449 | 1416646 | 0.63 | + |
| rs13161905 | 1417212 | 0.72 | - |
| rs2735917 | 1420268 | 0.92 | + |
| rs464049 | 1423905 | 0.21 | - |
| rs460700 | 1429969 | 0.00 | - |
| rs460000 | 1432825 | 0.00 | + |
| rs4975646 | 1433401 | 0.88 | - |
| rs403636 | 1438354 | 0.78 | - |
| rs2617605 | 1442521 | 0.89 | + |
| rs6350 | 1443199 | 0.93 | + |

TABLE 7

*SNPs for SLC6A3, chr5, position 1392790 - 1445430; e-value cutoff 0.72*

| SNP name | Location | *e*-value | Association |
|---|---|---|---|
| rs16967029 | 30195292 | 0.79 | + |
| rs2051810 | 30195841 | 0.84 | - |
| rs11658318 | 30206059 | 0.72 | - |
| rs8079471 | 30218317 | 0.64 | + |
| rs3760454 | 30222002 | 0.90 | + |

TABLE 8

*SNPs for SLC6A4, chr17, position 30194319 - 30236002; e-value cutoff 0.63*

| SNP name | Location | *e*-value | Association |
|---|---|---|---|
| rs2514229 | 113410000 | 0.87 | - |
| rs11214654 | 113410917 | 0.86 | + |
| rs7937641 | 113415976 | 0.63 | - |
| rs12222458 | 113417603 | 0.73 | - |
| rs10736470 | 113418371 | 0.73 | - |
| rs12576506 | 113419869 | 0.85 | + |
| rs10750025 | 113424042 | 0.66 | + |
| rs7952106 | 113424558 | 0.70 | - |
| rs4373974 | 113430486 | 0.88 | - |
| rs4130345 | 113436487 | 0.88 | - |
| rs7123697 | 113440331 | 0.78 | + |
| rs6589386 | 113443753 | 0.75 | + |
| rs4132966 | 113451589 | 0.86 | + |
| rs7940164 | 113451765 | 0.90 | - |
| rs4245155 | 113457324 | 0.92 | - |
| rs11607834 | 113461680 | 0.92 | - |
| rs12280220 | 113469219 | 0.93 | - |

TABLE 9
*SNPs for DRD2, chr11, position 113409595 - 113475691; e-value cutoff 0.63*

## APPENDIX C: DISCUSSION ON GENE-SPECIFIC FINDINGS IN THE MCTFR DATA

*GABRA2:* As seen in the plots, the first two SNPs detected are close to two separate exons. The 4th and 5th detected SNPs, rs1808851 and rs279856, are at perfect LD with rs279858 in the larger 7188-individual dataset (Irons, 2012). This SNP had not been genotyped in our sample, but is the marker in GABRA2 that is most frequently associated in the literature with alcohol abuse (Cui et al., 2012). Interestingly, a single SNP RFGLS analysis of the same twin studies data that used Bonferroni correction on marginal $p$-values to detect SNPs had missed these SNPs (Irons, 2012). This highlights the advantage of our approach.

*ADH genes:* Multiple studies have associated rs1229984 in the ADH1B gene (position 99318162 of chromosome 4) with alcohol dependence (`https://www.snpedia.com/index.php/Rs1229984`), which as seen in the plot of ADH2 is close to an exon region. Our data does not contain this marker, but detects one SNP 20 kb upstream of this, rs17027523. Another SNP, rs3775540 at position 99304544 has an $e$-value of 0.226, so narrowly misses detection. This is close to rs1229984, and also rs1042026 at position 99307309, which Macgregor et al. (2008) found to be strongly associated with alcohol consumption.

The SNP rs17027523 is interesting: it resides in the uncharacterized long non-coding RNA gene LOC100507053. One previous study (Gelernter et al.,

2014; Xu et al., 2015) found significant associations for 5 SNPs in this gene with alcohol consumption for African American population through single-SNP analysis on non-familial GWAS data. Notably, their analysis found a much stronger evidence of the association in African-American part of the sample than the European American part, while our findings are entirely from a Caucasian sample.

*OPRM1:* Many of the SNPs analyzed in this gene have very low $e$-values, and tend to cluster together. The minor allele of the SNP rs1799971 (chr 6, position 154039662) has been associated with stronger alcohol cravings (https://www.snpedia.com/index.php/Rs1799971), and we detect rs12662873 at position 154040810.

*CYP2E1:* Five of the 9 SNPs studied are detected through our analysis. Four of them are within 10 kb of one another (base pairs 133534822 to 133543210 in chr 10). In the analysis of Lind et al. (2012) rs4646976 at 133534223 position was most associated with a measure of breath alcohol concentration: this is within our detected region. This study had also detected rs4838767 in the promoter region of CYP2E1 (position 133520114) associated with multiple alcohol consumption measures. We detect rs9419702 at position 133531153.

*ALDH2:* All 6 SNPs we study are close to exons, and 5 get picked up by the $e$-value procedure. While all five are at a lesser base pair position than the well-known SNP rs671 (https://www.snpedia.com/index.php/Rs671, position 111803962), one of the SNPs we analyze (rs16941437) is within 10 kb upstream of this SNP.

*COMT:* The SNP rs4680 has long been associated with schizophrenia and substance abuse, including alcoholism. A case-control study (Voisey et al., 2011) associated rs4680 and rs165774 with alcohol dependence through a SNP-wise chi-squared test, and had these two SNPs in high LD in their study population. Compared to this, in our simultaneous model of all COMT polymorphisms, the more well-known rs4680 has a below threshold $e$-value.

*SLC6A3:* Our analysis does not detect rs27072, which has been associated with alcohol withdrawal symptoms (https://www.snpedia.com/index.php/Rs27072).

Finally, most $e$-values for the last 3 genes, i.e. SLC6A3, SLC6A4 and DRD2, are large: indicating weak SNP signals. We found this observation interesting, because variants of these genes have known interaction effects behind alcohol withdrawal-induced seizure (Karpyak et al., 2010) and bipolar disorder (Wang et al., 2014), as well as additive effect on the susceptibility to smoking addiction (Erblich et al., 2005).

# REFERENCES


Cui, W. Y., Seneviratne, C., Gu, J. and Li, M. D. (2012). Genetics of GABAergic signaling in nicotine and alcohol dependence. *Hum. Genet.* **131** 843–855. doi: 10.1007/s00439-011-1108-4.

Erblich, J. A., Lerman, C., Self, D. W. et al. (2005). Effects of dopamine D2 receptor (DRD2) and transporter (SLC6A3) polymorphisms on smoking cue-induced cigarette craving among African-American smokers. *Mol. Psychiatry* **10** 407–414.

Gelernter, J., Kranzler, H. R., Sherva, R., Almasy, L. et al. (2014). Genome-wide association study of alcohol dependence: significant findings in African- and European-Americans including novel risk loci. *Mol. Psychiatry* **19** 41–49. doi:10.1038/mp.2013.145.

Irons, D. E. (2012). Characterizing specific genetic and environmental influences on alcohol use PhD thesis, University of Minnesota.

Karpyak, V. M., Biernacka, J. M., Weg, M. W. et al. (2010). Interaction of SLC6A4 and DRD2 polymorphisms is associated with a history of delirium tremens. *Addict. Biol.* **15** 23–34. doi: 10.1111/j.1369-1600.2009.00183.x.

Lind, P. A., Macgregor, S., Heath, A. C. and Madden, P. A. F. (2012). Association between *in vivo* alcohol metabolism and genetic variation in pathways that metabolize the carbon skeleton of ethanol and NADH reoxidation in the Alcohol Challenge Twin Study. *Alcohol Clin. Exp. Res.* **36** 2074–2085. doi:10.1111/j.1530-0277.2012.01829.x.

Macgregor, S., Lind, P. A., Bucholtz, K. K. et al. (2008). Associations of ADH and ALDH2 gene variation with self report alcohol reactions, consumption and dependence: an integrated analysis. *Hum. Mol. Genet.* **18** 580–593.

Majumdar, S. and Chatterjee, S. (2017). Fast and General Model Selection using Data Depth and Resampling. https://arxiv.org/abs/1706.02429.

Voisey, J., Swagell, C. D., Hughes, I. P. et al. (2011). A novel SNP in COMT is associated with alcohol dependence but not opiate or nicotine dependence: a case control study. *Behav. Brain Funct.* **7**. doi: 10.1186/1744-9081-7-51.

Wang, T. Y., Lee, S. Y., Chen, S. L. et al. (2014). Gender-specific association of the SLC6A4 and DRD2 gene variants in bipolar disorder. *Int. J. Neuropsychopharmacol.* **17** 211–222. doi: 10.1017/S1461145713001296.

Xu, K., Kranzler, H. R., Sherva, R., Sartor, C. E. et al. (2015). Genomewide Association Study for Maximum Number of Alcoholic Drinks in European Americans and African Americans. *Alcohol Clin. Exp. Res.* **39** 1137–1147. doi: 10.1111/acer.12751.