SNPs on Chips: The Hidden Genetic Code in Expression Arrays

Elzbieta Sliwerska, Fan Meng, Terence P. Speed, Edward G. Jones, William E. Bunney, Huda Akil, Stanley J. Watson, and Margit Burmeister

Gene expression microarray analysis in postmortem brains is one of the fastest growing fields of psychiatric research. Here we show that common polymorphisms (SNPs) present on probe sets can masquerade as significant "gene expression" differences. After first observing this artifact in the Catechol-O-methyl transferase (COMT) gene, we replicate the finding in two additional genes predicted to show this artifact. Many Affymetrix chips contain thousands of SNPs that are both common and in the central probe region affecting hybridization, and thus have the potential to confound expression analysis.

Key Words: Affymetrix, Catechol-O-Methyl-Transferase, gene expression, genetic association, microarray, postmortem brain

Gene expression microarray analysis in psychiatric research typically compares the expression of tens of thousands of genes between postmortem brain tissue from individuals with psychiatric illness and from control individuals. Stronger hybridization to probes or probe sets on oligonucleotide arrays is interpreted to mean increased mRNA levels in the tissue analyzed. These analyses can implicate expression changes in specific genes or whole pathways without any prior hypotheses, and thus have the potential to illuminate novel etiologies in psychiatric illness. The cause of the expression differences cannot usually be directly evaluated—the expression difference may be a secondary, "symptomatic" effect of the disorder, or could be related to a primary influence of a genotype on both expression and risk for the disorder.

In contrast, genetic studies seek to identify polymorphisms, most often SNPs, associated with these disorders in genomic DNA of affected individuals. Many recently published association studies, (for example Chumakov et al 2002; Edenberg et al 2004) found SNPs or combinations of SNPs (haplotypes) associated with the psychiatric disorder in question. In order to test whether SNP genotypes are indeed associated with expression, we genotyped coding and promoter SNPs from candidate genes for psychiatric disorders in brain samples from 65 individuals from whom we also have gene expression profiles of multiple regions available on Affymetrix chips (Bunney et al 2003; Evans et al 2004; Li et al 2004; Vawter et al 2004). These individuals included subjects with mood disorders, Schizophrenia, and controls. We then tested statistically whether these SNPs affected their gene's expression.

From the Molecular & Behavioral Neuroscience Institute (ES, HA, SW, MB), Department of Psychiatry (FM, HA, SJW, MB), and Department of Human Genetics (MB), University of Michigan, Ann Arbor, Michigan; Department of Statistics (TPS), University of California at Berkeley, Berkeley; Department of Psychiatry (EGJ), University of California Davis, Davis; Department of Psychiatry (WEB), University of California Irvine, Irvine, California.

Address reprint requests to Margit Burmeister, Ph.D., Molecular & Behavioral Neuroscience Institute, University of Michigan, 5061 BSRB, 109 Zina Pitcher Pl., Ann Arbor MI 48109-2220
E-mail: margit@umich.edu.

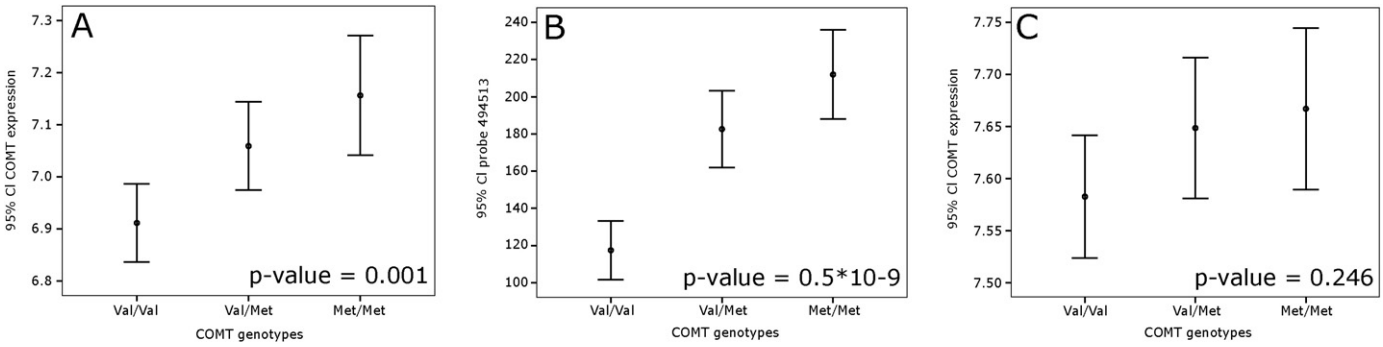
Received May 25, 2005; revised November 11, 2005; accepted January 9, 2006.

COMT—Hybridization Difference is Due to Genotype, not Expression Difference

In order to obtain genotype information in conjunction with gene expression data, we extracted genomic DNA from postmortem brains using standard procedures. We genotyped about 50 SNPs in candidate genes using either ABI Taqman or PCR-RFLP. We then tested whether each SNP was associated with its gene's expression level, using RMA for normalization and background correction of the expression data (Irizarry et al 2003). For common SNPs (with at least 4 samples homozygous for the minor allele), association was tested by Analysis of Variance (ANOVA), using the program SPSS, version 13.0. For rarer alleles, a T-test of the RMA-normalized expression values between homozygotes for the common variant and those containing at least one allele of the rare variant was performed. We found that one association, between the common Val108Met allele of COMT and its expression level, measured by COMT probe set 208817_at, stood out not only as significant [dorsolateral prefrontal cortex (DLPFC), p=.001; see Figure 1A], but also was consistent in all other brain regions [e.g., Anterior Cingulate (AnCg) p=.009].

While this result was intriguing, a second probe set for COMT, 208818_s_at, on the same array did not show this association. To determine the cause of this discrepancy, we investigated the expression difference at the probe level of the array. For each gene, Affymetrix arrays typically have 11 perfect match oligonucleotides (probes) that are 25 bases in length. Some genes are represented by more than one probe set of 11 probes. Typical MAS5 (Microarray Suite v.5) or RMA (Robust Microarray Analysis) analysis integrates in various ways over these 11 different probes. When examining the sequence of the probes for COMT, we found that among the 11 probes in the first probe set for COMT, 208817_at, two probes contained the SNP in question, with the A, the allele encoding Met, encoded in both cases (see Figure 1D). In one probe, the A was at position 14 of the 25mer, near the center. This probe thus presents a near-central mismatch probe when mRNAs encoding the Val allele are hybridized, and is expected to hybridize significantly less to Val-encoding mRNA than to Met-encoding mRNA. The other probe had the A allele near the end of the probe (3 bases from the end), not expected to affect hybridization efficiency, and indeed, this probe alone showed no significant "expression" difference (data not shown). However, when only hybridization to the probe with the central A allele is considered, we observe an approximately 2 fold "expression" difference (Figure 1B), which is highly significant ($p=10^{-9}$). In contrast, when only the 9 probes not containing the SNP are considered, there is no significant difference between

0006-3223/07/$32.00
doi:10.1016/j.biopsych.2006.01.023



D

Sequence	Probe Coordinates		Probe Interrogation Position	Strandedness
CGGATGGTGGATTTCGCTGGC**A**TGA	94	243	649	Antisense
GGATTTCGCTGGC**A**TGAAGGACAAG	494	513	657	Antisense
GAAGGACAAGGTCACCCTTGTGGTT	689	331	672	Antisense
AAGGTCACCCTTGTGGTTGGAGCGT	295	149	679	Antisense
TGGTTGGAGCGTCCCAGGACATCAT	207	559	692	Antisense
ATCATCCCCCAGCTGAAGAAGAAGT	461	23	712	Antisense
TGGTCTTCCTCGACCACTGGAAGGA	380	559	758	Antisense
TCGACCACTGGAAGGACCGGTACCT	408	593	767	Antisense
CACTGGAAGGACCGGTACCTGCCGG	703	177	772	Antisense
GTGGTGGACGGCCTGGAGAAGGCCA	493	483	955	Antisense
GAAGGCCATCTACAAGGGCCCAGGC	407	347	972	Antisense

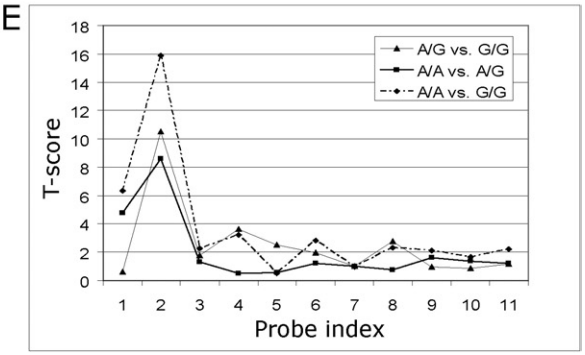


Figure 1. COMT gene expression levels across 3 different COMT genotypes groups—Val/Val, Val/Met, Met/Met. COMT genotyping was performed using a TaqMan assay designed by Applied Biosystems. The graph shows mean gene expression levels for each group. ANOVA was performed to determine the significance of the difference between the Val/Val and the Met/Met genotype groups. **(A)** Expression levels based on all 11 probes of probe set 208817_at; data were normalized using RMA. There is a significant difference between genotypes. **(B)** Signal intensities for the one probe that contains the A allele (encoding Met) near the probe's center. The difference between genotypes becomes much more significant. **(C)** COMT expression levels, RMA normalized, measured using the 9 probes not containing the SNP shows no significant difference between genotypes. **(D)** Sequences of all 11 probes from the 208817_at probe set. The SNP "A" allele is shown in bold. **(E)** Expression-normalized hybridization signal illustrates the effect of SNP-containing probes. The ratio of the signal from each probe was divided by the signal of all probes, and was calculated separately for each genotype. The graph illustrates, with T-scores, a highly significant difference in this expression-normalized signal between all three genotype pairs [see Dai et al. (2005) for further details of this calculation].

expression of COMT and genotype (Figure 1C, $p=0.2$). We conclude that the apparent expression difference originally found was largely an artifact due to the presence of the SNP on one of the probes on the chips, not a true expression difference. This result was unexpected since typically the hybridization of 11 oligonucleotides is averaged before analysis, and RMA down-weights results from atypically hybridizating probes. However, while the twofold difference with one probe was reduced to only a 1.18 fold difference overall, it remained significant and showed a misleading expression difference, and thus is not fully compensated for.

It should also be pointed out that Zhu et al. (2004) reported that the Met allele is expressed at a higher level than the Val allele, using more sensitive technologies not affected by the artifact described here. It may thus be possible that the nonsignificant 1.05 fold difference seen in Figure 1C may indeed be a real, though very small, difference not reliably significant with microarray data.

SNPs on Chips are Common

Next we asked whether the observed misleading "expression" difference caused by a SNP on the probe set on the chip could be a common problem with expression chips. Sequences for oligonucleotide probes on Affymetrix HG-U95Av2, HG-U133A, HG-U133B and HG-U133 Plus 2.0 were downloaded from the Affymetrix Support website. Sequences of known human SNP sites were obtained from dbSNP, build 124. Allele-specific probes were identified through sequence alignments, and the location of the allele-specific base on the corresponding probe was recorded. Table 1 outlines the presence of SNPs in dbSNP on the different human Affymetrix arrays and their frequency. Because a SNP in the central region is much more likely to affect hybridization efficiency than a SNP near the end of the probe (e.g., Mei et al 2003), we also recorded these separately. In addition, SNPs that are common are more likely to affect results than rare SNPs. We find that there are several thousand probes

Table 1. Frequency of SNPs on Affymetrix Probes on Different Types of Human Chips. All Numbers in Italics Refer Only to SNPs in the Central 15 bp Region

Chip Type	HG-U95Av2	HG-U133A	HG-U133B	HG-U133 Plus
Total allele-specific probe	19498	245939	18437	50461
Allele in central 15 bp region	12490	16463	11856	32784
SNP with known frequency	*2503*	*4318*	*3379*	*8738*
Minor allele frequency>.1	*1286*	*2511*	*2189*	*5445*
Minor allele frequency>.2	*851*	*1717*	*1449*	*3677*
Bi-allelic probe pair	*381*	*459*	*280*	*1085*