Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation

Kristi Kerkel1, Alexandra Spadola2, Eric Yuan1, Jolanta Kosek1, Le Jiang1, Eldad Hod3, Kerry Li1, Vundavalli V Murty1,3, Nicole Schupf4, Eric Vilain5,6, Mitzi Morris7, Fatemeh Haghighi7 & Benjamin Tycko1,3

Allele-specific DNA methylation (ASM) is a hallmark of imprinted genes, but ASM in the larger nonimprinted fraction of the genome is less well characterized. Using methylation-sensitive SNP analysis (MSNP), we surveyed the human genome at 50K and 250K resolution, identifying ASM as recurrent genotype call conversions from heterozygosity to homozygosity when genomic DNAs were predigested with the methylation-sensitive restriction enzyme HpaII. Using independent assays, we confirmed ASM at 16 SNP-tagged loci distributed across various chromosomes. At 12 of these loci (75%), the ASM tracked strongly with the sequence of adjacent SNPs. Further analysis showed allele-specific mRNA expression at two loci from this methylation-based screen—the vanin and CYP2A6-CYP2A7 gene clusters—both implicated in traits of medical importance. This recurrent phenomenon of sequence-dependent ASM has practical implications for mapping and interpreting associations of noncoding SNPs and haplotypes with human phenotypes.

Allele-specific CpG methylation (ASM) is a hallmark of genomic imprinting, but the characteristics of ASM at nonimprinted loci are less well understood. DNA methylation tracking as a genetic trait was found in early studies of human VNTR loci1,2, and a recent study described a microarray-based approach, called MSNP, for genetic and epigenetic profiling4. Genomic DNA is digested with a non–PCR-based method, we did DNA blotting

Here, we report genome-wide surveys for ASM using MSNP at 50K and 250K resolution. The 50K dataset consisted of single XbaI and XbaI + HpaII representations from 12 normal tissue samples: peripheral blood leukocyte (PBL) DNA from four adults, kidney DNA from one adult, brain and lung DNA from one adult, DNA from three placentas (sampled at the fetal surface), and buccal cell DNA from two adults. On this 50K array, there are 39,100 class 1 SNPs (noninformative for DNA methylation, because of absence of HpaII sites within the XbaI fragment), 15,057 class 2 SNPs (informative, because of the presence of nonpolymorphic HpaII sites) and 4,328 class 3 SNPs (not reliably informative, because of polymorphic HpaII sites). In assessing the SNP-tagged loci for ASM, we first eliminated those with weak signals—below the 10th percentile in the XbaI representations—in ≥5 samples. We next asked for allele call conversions in ≥3 samples in the plus-HpaII representations, producing a list of 74 SNP-tagged loci. Lastly, we examined the hybridization quality and the selective reduction in intensity of one of the two alleles for each of these loci using the perfect match/mismatch (PM/MM) view in dChip5.

We brought forward two of the most convincing candidate loci for independent validations. SNP rs1042073, in exon 13 of the lactotransferrin (LTF) gene, showed AB→AA conversions with HpaII predigestion in five tissues (PBL, brain, kidney, lung and placenta), each obtained from a different individual, whereas no AB→BB conversions were seen (Fig. 1a). These results predicted that the C allele would be consistently more methylated than the T allele at one or both of the two HpaII sites in this XbaI fragment. As shown for several tissues from different individuals (Fig. 1), PCR/RFLP and sequencing consistently showed a reduced representation of the T allele when the DNA had been predigested with HpaII, and the same genotype dependence was seen in an expanded series (Table 1 and Supplementary Table 1 online). To validate this correlation between ASM and SNP genotype using a non–PCR-based method, we did DNA blotting

Received 2 January; accepted 13 May; published online 22 June 2008; doi:10.1038/ng.174
of genomic DNAs from 21 individuals with the 3 possible rs1042073 genotypes. ASM was readily scored as a doublet band pattern, biallelic methylation as a single upper band, and biallelic lack of methylation as a single lower band. As shown in Figure 1d, there was an invariant relationship between the rs1042073 genotype and the pattern of methylation, with C-allele homozygotes showing biallelic methylation, CT heterozygotes showing ASM and TT homozygotes showing biallelic lack of methylation at this site. Given this notably constant relationship, we sought artifactual explanations, such as the presence of a polymorphic HpaII site created by or in linkage disequilibrium with the index SNP in the same restriction fragment. However, sequencing excluded this possibility. As ASM is often found at imprinting loci, we also sought to exclude imprinting as an explanation. We addressed this question by using trios of DNA from maternal and paternal PBIs and placenta (fetal surface) from the offspring. Among the informative trios, the methylated allele in the LTF exon 13 region was paternal in origin in four but maternal in origin in one, thus excluding parental imprinting as an explanation for the ASM (Supplementary Fig. 1 online).

We obtained similar results for a second locus from the 50K survey, tagged by SNP rs9258170 in the upstream region of the HLA-F gene. There were AB→AA call conversions with HpaII predigestion in each of five samples (one kidney, one lung, one brain, one placenta, and two buccal epithelial cell preparations), and no AB→BB conversions. As shown by pre-digestion followed by PCR, ASM in the HLA-F upstream region correlated strongly, although not invariably, with the local DNA sequence, as scored by SNPs and indels (Table 1 and Supplementary Fig. 2 online). To ask whether the ASM at the HpaII sites in this region might reflect more widespread ASM at other CpG sites, we carried out bisulfite sequencing of a region near the index SNP. This procedure showed a patch of differential methylation affecting at least 17 CpG dinucleotides (Supplementary Fig. 2).

We next carried out MSNP using 250K Sty1 SNP arrays, containing 164,561 class 1 (69%), 57,583 class 2 (24%), and 15,906 class 3 SNPs (7%). Our 250K survey included six PBL samples, two normal bone marrow samples, one sample of CD34+Lin− hematopoietic cells, and three placental samples. We analyzed duplicate technical replicates for Sty1 and Sty1 + HpaII representations, and single determinations for Sty1 + MspI (60 microarrays). We asked for call conversions from AB to Sty1 to AA or BB in Sty1 + HpaII in both technical replicates in at least 2 of the 12 biological samples. To exclude loci with biallelic hypomethylation, we additionally required that the average Sty1 + HpaII intensity be reduced not more than 70% from the average Sty1-only intensity, and that the intensity in the Sty1 + MspI representation be at least 40% less than the mean signal with Sty1 + HpaII. This procedure produced a list of 58 loci as candidates for ASM. As expected, this list was depleted in class 1 SNPs (0, 0%) and enriched in class 2 (31, 53%) and class 3 (27, 47%) SNPs. The opposite analysis, asking for all class 2 SNPs with persistent AB calls and preserved intensity of both alleles in the plus-HpaII representations (reduced not more than 70% from the minus-HpaII intensity) in all heterozygous samples (requiring at least two of the samples to be heterozygous), with a signal intensity in the Sty1 + MspI representation at least 40% less than the mean signal with Sty1 + HpaII, yielded 15,454 class 2 SNPs. So within the sensitivity of this method and the validation and determination of the positive rate below, we estimate that at least 0.16% of the informative SNP-tagged loci queried by the array have recurrent ASM in the tissues examined.

As internal controls in this 250K dataset, we found consistent call conversions at SNPs in three imprinted domains: in a PBL sample at SNP rs11166318 located 0.5 kb downstream of the 3’ end of the MAGEL2 gene, in a bone marrow sample at SNP rs231907 in the KCNQ1 gene, and in the CD34+ stem cells at SNP rs2107425 upstream of the H19 gene. Monoallelic expression of MAGEL2 mRNA as a result of parental imprinting has been shown previously6,7, and DNA methylation occurs preferentially on the maternal allele at multiple sites over this imprinted domain on chromosome 15 (ref. 8), but ASM has not been previously examined in the MAGEL2 downstream region. We therefore validated the MSNP findings for this SNP-tagged Sty1 fragment using pre-digestion followed by PCR and RFLP. As expected for an imprinted locus, in which parent-of-origin, not DNA sequence, dictates which allele
ASM might be associated with allele-specific mRNA expression (ASE). expression raised the question of whether this sequence-dependent

| dbSNP ID   | Gene or region | Tissue | CT (N) | Allele C methyl. | Allele T methyl. | No ASM | Binomial test |
|------------|----------------|--------|--------|------------------|------------------|--------|---------------|
| rs1042073  | LTF exon 13    | PBL    | 49     | 48               | 0                | 1      | $P < 10^{-8}$ |
|            |                | Placenta | 33    | 31               | 0                | 2      | $P < 10^{-8}$ |
| rs6494120  | GCTN3 US       | PBL    | 48     | 44               | 0                | 4      | $P < 10^{-8}$ |
| rs12873012 | POU4F3 DS      | PBL    | 45     | 2                | 30               | 3      | $P = 0.0032$  |
| rs6969642  | DNA1B6         | PBL    | 30     | 27               | 3                | 3      | $P < 10^{-8}$ |
| rs9366927  | FG20-PIM1      | PBL    | 9      | 9                | 0                | 0      | $P = 0.002$   |
| rs9951893  | PHLPP-BCL2     | PBL    | 37     | 35               | 0                | 2      | $P < 10^{-8}$ |

**Table 1** Sequence dependence of ASM at SNP-tagged loci from the 50K and 250K MSNP screens

becomes methylated, ASM in the MAGEL2 region did not correlate
with the SNP genotype (Supplementary Fig. 3 online).

We next selected a group of loci reported as having ASM in the 250K data, outside of known imprinted regions, for independent validations in larger series of individuals. These assays validated ASM for 14/18 candidate loci examined (true-positive rate 78%). For the minority of loci that could not be validated, it was evident from PCR that biallelic hypomethylation, with relatively greater sensitivity of the SNP array probe sets for one allele, accounted for the false-positive call conversions. The MSNP data and validations by pre-digestion followed by PCR, sizing gel analysis and/or direct sequencing, are shown for the CYP2A7 gene in Figures 2 and 3, and examples of validations of ASM at additional SNP-tagged loci are shown in Supplementary Figures 3–5 online. In the combined 50K and 250K data, the ASM correlated strongly with the allele in cis at the index SNP or closely adjacent informative SNPs for 12/16 (75%) of the validated loci (Table 1 and Supplementary Table 1).

The known relationship between DNA methylation and gene expression raised the question of whether this sequence-dependent ASM might be associated with allele-specific mRNA expression (ASE). Among four loci that we tested by comparing allelic representation in cDNA versus genomic DNA, only rare individuals showed ASE of the LTF and HLA-F genes, whereas many showed strong ASE of the CYP2A7 and VNN1 genes. Both of these positive examples are in gene clusters, on chromosome 19 and 6, respectively, which lack classical upstream CpG islands and show tissue specific expression. Although our MSNP screen was done using hematopoietic tissues, we directly assayed ASM and ASE for CYP2A7 in the main expressing organ, the liver. Comparison of allele representation in genomic and cDNA PCR products from the CYP2A7 gene showed strong ASE in 10 of 17 heterozygous human liver samples examined, with relative over-expression of the same allele in each of these 10 individuals, and sequencing across multiple intragenic SNPs revealed that the presence or absence of ASE was strongly dependent on genotype (Fig. 2b and Supplementary Table 2 online). For VNN1, we confirmed ASM by predigestion followed by RFLP analysis near the index SNP and by bisulfite sequencing across a CG-rich region, not dense enough to qualify as a CpG island, between the index SNP and the first exon (Fig. 3a,c,d). We found a consistent difference in methylation patterns between the two alleles, already detectable early in the
hematopoietic lineage (CD34+Lin− cells; Fig. 3d). By comparing cDNA and genomic PCR products in PBL, we documented a domain-wide effect, with ASE strongest for VNN1 but also evident for the other two genes (VNN3 and VNN2) in this cluster (Fig. 3b). Both of these examples are medically relevant: prior genotype-phenotype correlations for CYP2A6 (located immediately downstream of CYP2A7 and 93% identical in amino acid sequence) have shown that genes in this cytochrome P450 cluster control blood nicotine levels and may thereby influence smoking behavior, and a recent report described total VNN1 mRNA expression as a quantitative trait linked to haplotypes, and in turn to HDL cholesterol levels.10,11 Future methylation-based screens using the methods developed here will likely uncover additional examples of sequence-dependent ASE linked to important human traits.

This dependence of ASE on the local DNA sequence in cis is distinct from other epigenetic phenomena such as imprinting, allelic exclusion, random monoallelic methylation and expression12–15, and epimutation (which may be haplotype-dependent in some cases14–17). Mechanisms leading to sequence-dependent ASE will likely differ among loci: allele-specific affinity for DNA-binding proteins with downstream effects on DNA methylation is one obvious possibility and our findings, and in conjunction with prior data16,18–20, suggests that this model may apply to the CYP2 and vanin gene clusters. Direct effects of the DNA sequence on propensity for methylation are another

**Figure 2** Genotype-dependent ASM and ASE in the CYP2 gene cluster. (a) Hybridization data from 250K MSNP showing ASM within the StyI fragment containing SNP rs3844442 in PBL. See Table 1 for sequence dependence of the ASM at this location in the CYP2 gene cluster, queried by SNPs rs3844442 and rs34724660. (b) ASE of CYP2A7 mRNA in human livers. Liver DNAs were screened for heterozygosity at SNP rs3815710, and 17 heterozygous samples from different individuals were analyzed for ASE: 10 showed strongly preferential expression of the C allele (Liv2, Liv3 and Liv4, for example) and 7 showed equal biallelic expression or a slight preferential expression of the A allele. (c) Genomic DNA was mock-digested or digested with HpaII containing SNP rs12205095. Genomic DNA was mock-digested or digested with HpaII containing SNP rs12205095.
possibility. In practical terms, the phenomenon of sequence-dependent ASM extends the concept of haplotypes to include epigenetic marks (‘epihaplotypes’) and, as an indicator for the presence of nearby regulatory polymorphisms that confer the allelic asymmetry, it will likely be useful for fine mapping and interpreting associations of noncoding SNPs with complex genetic diseases.

**METHODS**

**MSNP procedure.** Genomic DNA from normal human tissues, obtained with institutional review board approval and anonymous to individual identifiers, was purified using the DNeasy Blood and Tissue Kit (Qiagen). For both 50K XbaI and 250K StyI MSNP, we followed the recommended protocol of the SNP array manufacturer (Affymetrix), but with the addition of an initial HpaII or MspI digestion for the plus-HpaII and plus-MspI probes. Genomic DNA (250 ng per representation) was digested with HpaII or MspI in a total volume of 12 μl (Buffer 1; New England Biolabs) for 3 h, followed by adjustment of the buffer by addition of 15.8 μl of water and 3.2 μl of 10× Buffer 3 (New England Biolabs), and a further digestion with StyI for 3 h. StyI linkers were ligated to the restriction fragments, and PCR was carried out for 30 cycles with linker primers. We carried out PCR purification using the DNA Amplification Clean-Up Kit (Glontech), and probe fragment (brief exposure to Dnase), biontin-labeling and hybridization as described in the Affymetrix protocol. We determined intensity values by normalization and model-based expression (PM-only method) using dChip.

The main determinant of the data quality is the starting material: the genomic DNA must be high molecular weight. Call rates were in the range of 89–95% with StyI and StyI + HpaII, with lower call rates of 76–88%, as predicted, in the StyI + MspI representations, in which many of the class 2 SNPs dropped out as a result of complete digestion of internal CCGG sites.

**Independent assays for ASM.** For the HpaII digestion followed by PCR, RFLP and sequencing assays, genomic DNA, 1 μg in 80 μl, was digested with 5 U of HpaII overnight, followed by an additional 5 U of HpaII for 1 h or incubated in parallel with buffer alone. Of this digested DNA, 25 ng was used for each PCR (primers in Supplementary Table 3 online). In this predigestion/PCR assay, we did a control with MspI predigestion, and verified that this procedure markedly reduce the amount of PCR product obtained under the same cycling conditions. For bisulfite conversion followed by PCR, cloning and sequencing, we treated genomic DNA with sodium bisulfite and purified the converted DNA using the CpGenome DNA modification kit (Chemicon) according to the protocol of the manufacturer. The bisulfite/PCR primers and SNP ID numbers for distinguishing the alleles in the HLA-F upstream region, two positions in the CYP2 gene cluster, and in the promoter region of the VNN1 gene are listed in Supplementary Table 3. Bisulfite PCR products were cloned in the pCR 2.1-TOPO vector (Invitrogen), and multiple clones were sequenced. For methylation-sensitive DNA (Southern) blotting of the LTF exon 13 region, we digested 3 μg genomic DNA with the indicated restriction enzymes overnight, followed by electrophoresis on 1% agarose gels, transfer to Nytran membranes (Schleicher and Schull) and hybridization of the blots with a 32P-labeled LTF probe. The probe DNA was synthesized by genomic PCR; the primer sequences are in Supplementary Table 3.

**Assays for allele-specific mRNA expression.** Reverse transcription PCR was carried out with cDNAs prepared from total PBL RNA (Protoscript First Strand cDNA Synthesis Kit, New England Biolabs), using primers specific for VNN1, VNN2, VNN3 and CYP2A7, designed to span at least one intron and include at least one exonic SNP with a high minor allele frequency. We avoided SNPs within the primer sequences and sequences that were identical among family members, conditions which put strong constraints on primer placement in these highly polymorphic gene clusters. The controls for these analyses were genomic PCR products spanning these same SNPs. The primer sequences are listed in Supplementary Table 3.

**Accession codes.** NCBI GEO: data have been deposited with accession code GSE11409.

**Note:** Supplementary information is available on the Nature Genetics website.

**ACKNOWLEDGMENTS**

This work was supported by grants to B.T. from the US National Institutes of Health, the March of Dimes and the Leukemia and Lymphoma Society in collaboration with the Douglas Kroll Research Program.

**AUTHOR CONTRIBUTIONS**

Study design: K.K. and B.T. MSNP, molecular validations and analysis of ASE: K.K., A.S., E.Y., L.I., L.J., E.H., K.L. and V.V.M. Bioinformatic analyses: E.H., M.M. and F.H. Statistical analyses: N.S. and E.H. Provision of biological samples: A.S., V.V.M., N.S., E.V. and B.T. Interpretation and writing the paper: K.K., N.S., E.V. and B.T.

Published online at http://www.nature.com/naturegenetics/

Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Chandler, L.A., Ghazi, H., Jones, P.A., Boukamp, P. & Fusenig, N.E. Allele-specific methylation of the human c-Ha-ras-1 gene. Cell 50, 711–717 (1987).
2. Silva, A.J. & White, R. Inheritance of allelic blueprints for methylation patterns. Cell 54, 145–152 (1988).
3. Yamada, Y. et al. A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. Genome Res. 14, 247–266 (2004).
4. Yuan, E. et al. A single nucleotide polymorphism chip-based method for combined genetic and epigenetic profiling: validation in decitabine therapy and tumornormal comparisons. Cancer Res. 66, 3443–3451 (2006).
5. Lin, M. et al. d ChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 20, 1233–1240 (2004).
6. Bocaccio, I. et al. The Human MAGEL2 gene and its mouse homologue are paternally expressed and mapped to the Prader-Willi region. Hum. Mol. Genet. 8, 2497–2505 (1999).
7. Lee, S. et al. Expression and imprinting of MAGEL2 suggest a role in Prader-willi syndrome and the homologous murine imprinting phenotype. Hum. Mol. Genet. 9, 1813–1819 (2000).
8. Salton, S. et al. Clinical spectrum and molecular diagnosis of Angelman and Prader-Willi syndrome patients with an imprinting mutation. Am. J. Med. Genet. 68, 195–206 (1997).
9. Tyndale, R.F. Genetics of alcohol and tobacco use in humans. Ann. Med. 35, 94–121 (2003).
10. Goring, H.H. et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat. Genet. 39, 1208–1216 (2007).
11. Gimelbrant, A., Hutchinson, J.N., Thompson, B.R. & Chess, A. Widespread monoallelic expression on human autosomes. Science 318, 1136–1140 (2007).
12. Ohlisson, R., Tycko, B. & Sapienza, C. Monoallelic expression: ‘there can only be one’. Trends Genet. 14, 435–438 (1998).
13. Wang, J., Valo, Z., Smith, D. & Singer-Sam, J. Monoallelic expression of multiple genes in the CNS. PLOS ONE 2, e1293 (2007).
14. Chan, T.L. et al. Heritable germine epimutation of MSH2 in a family with hereditary nonpolyposis colorectal cancer. Nat. Genet. 38, 1178–1183 (2006).
15. Zogel, C. et al. Identification of cis- and trans-acting factors possibly modifying the risk of epimutations on chromosome 15. Eur. J. Hum. Genet. 14, 752–758 (2006).
16. Suter, C.M. & Martin, D.I. Inherited epimutation or a haplotypic basis for the propensity to silence? Nat. Genet. 39, 573; author reply 576 (2007).
17. Murrell, A. et al. An association between variants in the IGF2 gene and Beckwith-Wiedemann syndrome: interaction between genotype and epigenotype. Hum. Mol. Genet. 13, 247–255 (2004).
18. von Richter, O. et al. Polymorphic NF-Y dependent regulation of human nicotine C-oxidase (CYP2A6). Pharmacogenetics 14, 369–379 (2004).
19. Pitarque, M. et al. Identification of a single nucleotide polymorphism in the TATA box of the CYP2A6 gene: impairment of its promoter activity. Biochem. Biophys. Res. Commun. 284, 495–460 (2001).
20. Pitarque, M. et al. A nicotine C-oxidase gene (CYP2A6) polymorphism important for promoter activity, Hum. Mutat. 23, 258–266 (2004).
21. Cokus, S.J. et al. Shotgun bisulfite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452, 215–219 (2008).