Overlapping euchromatin/heterochromatin- associated marks are enriched in imprinted gene regions and predict allele-specific modification

Bo Wen, Hao Wu, Hans Bjornsson, et al.

Genome Res. 2008 18: 1806-1813 originally published online September 16, 2008
Access the most recent version at doi:10.1101/gr.067587.107

Overlapping euchromatin/heterochromatin-associated marks are enriched in imprinted gene regions and predict allele-specific modification

Bo Wen,[1,4] Hao Wu,[2,4] Hans Bjornsson,[1] Roland D. Green,[3] Rafael Irizarry,[2] and Andrew P. Feinberg[1,5]

[1]*Department of Medicine and Center for Epigenetics, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA;* [2]*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA;* [3]*NimbleGen Systems, Inc., Madison, Wisconsin 53711, USA*

Most genome-level analysis treats the two parental alleles equivalently, yet diploid genomes contain two parental genomes that are often epigenetically distinct. While single nucleotide polymorphisms (SNPs) can be used to distinguish these genomes, it would be useful to develop a generalized strategy for identifying candidate genes or regions showing allele-specific differences, independent of SNPs. We have explored this problem by looking for overlapping marks in the genome related to both euchromatin (histone H3 dimethyl lysine-4 [H3K4Me2]) and heterochromatin (DNA methylation [DNAm]). "Double hits" were defined by the intersection of H3K4Me2 and DNAm. For the top 5% of marks, defined by a sliding window, imprinted gene regions were enriched for double hits 5.4-fold. When the location information of CTCF binding sites were integrated, the "triple hits" were enriched 76-fold for known imprinted genes in the regions studied. The double hits in imprinted genes were found to occur usually at the site of alternative or antisense transcripts. In addition, four of four imprinted genes tested showing double hits also showed allele-specific methylation. We suggest that overlapping euchromatin/heterochromatin marks are common and are enriched for epigenetically distinct parental chromosome regions. Furthermore, we developed a novel approach to identifying allele-specific marks that is SNP independent, by fractionating using H3K4Me2 antibodies followed by DNA methylation analysis.

[Supplemental material is available online at www.genome.org.]

Epigenetics is the study of information heritable during cell division other than the DNA sequence itself, and epigenetic alterations are critical to normal development, as well as common diseases such as cancer (Feinberg 2007). Epigenetic marks include a host of chromatin modifications affecting histones H2, H3, and H4, although most studies have focused on alterations of the amino-terminal tails of H3 and H4, such as lysine methylation and acetylation (Kouzarides 2007; Li et al. 2007). In addition, methylation of cytosine distinguishes the DNA of normal and diseased cells, which is well established in cancer (Feinberg 2007). The new field of epigenomics, or genome-wide epigenetic analysis, is beginning to afford the opportunity for analysis of chromatin modifications and DNA methylation to identify regulatory elements and disease loci without preselection of individual genes (Callinan and Feinberg 2006; Bernstein et al. 2007; Esteller 2007).

Beside the ability of epigenomics to identify tissue- or disease-specific modifications, a deeper biological question is raised by this nascent field: As we are diploid organisms, are we carrying essentially two epigenetically distinct genomes within our cells, one of maternal and one of paternal origin? A special case of this question is the identification of imprinted genes, namely those showing preferential (not necessarily monoallelic) expression of a specific parental allele. Functionally meaningful allele-specific DNA methylation (Strichman-Almashanu et al. 2002; Bergstrom et al. 2007) and histone modifications (Xin et al. 2001; Perk et al. 2002; Carr et al. 2007) have been found at imprinted genes. However, in addition to imprinted loci, it is conceivable that a large fraction of the genome is epigenetically marked and different between the two parental chromosomes. Sapienza and colleagues suggested that natural selection may favor epigenetic differences between homologous chromosomes to facilitate meiotic pairing, and that gene expression differences (imprinting) may have arisen secondarily (Pardo-Manuel de Villena et al. 2000).

Current methods to assess allele-specific modifications, including chromatin modification and DNA methylation, depend upon the presence of single nucleotide polymorphisms (SNPs) or other polymorphisms to distinguish the two parental copies. However, a SNP-independent method would be afforded by high-throughput assessment of overlapping euchromatin/heterochromatin marks. We have taken such an approach applied to the ENCODE region of the human genome (The ENCODE Project Consortium 2004), as well as a similarly sized fraction (~1%) enriched for known imprinted genes. Our goal was to identify genomic regions with coincident but functionally antagonistic euchromatin (active-associated) and heterochromatin (silencing-associated) marks, and then to ascertain whether the marks in these regions were on separate parental chromosomes and/or marked imprinted genes.

[4]These authors contributed equally to this work.
[5]Corresponding author.
E-mail afeinberg@jhu.edu; fax (410) 614-9819.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.067587.108. Freely available online through the *Genome Research* Open Access option.

18:1806–1813 ©2008 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/08; www.genome.org

Results

Experimental design

In order to test the hypothesis that colocalizing euchromatin/heterochromatin-associated marks are enriched in imprinted gene regions, we first chose for euchromatin histone H3 lysine-4 dimethylation (H3K4Me2), because it is widely used for identification of active chromatin, including in chromatin immunoprecipitation on chip array (ChIP-chip) experiments (Kim et al. 2005), and one of our goals was to identify potential control regions for allele-specific gene expression. For the heterochromatin mark, we chose DNA methylation (DNAm) rather than one of the several chromatin modifications because of the abundant studies linking DNA methylation to imprinting as well as gene regulation (Feinberg and Tycko 2004). Furthermore, in our experience, ChIP-chip using heterochromatin-associated antibodies on DNA–protein cross-linked complexes is more problematic than with euchromatin markers, and another of our goals was to develop a simple protocol that the community can use to identify novel allele-specific marks without resorting to SNPs.

We used an immortalized lymphoblastoid cell line (GM06991) as well as freshly obtained fractionated T lymphocytes for preparation of chromatin and DNA. H3K4Me2 ChIP was performed as described (Kim et al. 2005), using formaldehyde cross-linking of DNA to chromatin and sonication to 500–1000-bp fragments, followed by reverse cross-linking and DNA extraction. DNA methylation analysis was done using methylated DNA immunoprecipitation (MeDIP). Samples were double labeled (Cy5 for enriched fraction, Cy3 for total input) and hybridized to two NimbleGen 385K arrays: a custom gene array termed the "Imprinting Array" that included 45 known imprinted genes (Supplemental Table 1), as well as 120 genes of unknown imprinting status, but within 500 kb of known imprinted genes, reasoning that the latter group would be enriched for imprinted genes because of the known association of imprinted genes with each other in clusters (Strichman-Almashanu et al. 2002), and the Encylopedia of DNA Elements (ENCODE) array (The ENCODE Project Consortium 2004), which includes 380 genes excluding the X chromosome (because of the potential for double hits due to the inactive X chromosome) and excluding the chromosome 11p15 imprinted gene domain. For the sake of this analysis, the 11p15 known imprinted gene region was included on the Imprinting Array, but it also appears on the ENCODE array by design (because of the imprinted genes), but we excluded those known genes from the ENCODE microarray analysis, as the microarray is meant to represent a set of genes unenriched for imprinted genes. The X chromosome region on ENCODE was analyzed separately.

We tested for the presence of "double hits," defined as the two marks overlapping by at least 100 bp. We further scored double hits as marking a specific gene if they occurred within 1 kb of an annotated gene based on the UCSC Known Genes track. While we cannot know with certainty that a given gene on the ENCODE array does not show imprinting or allele-specific expression, we would nevertheless expect a much higher frequency of double hits near genes on the Imprinting Array than near genes on the ENCODE array if our hypothesis is correct, i.e., that double hits identify genes showing allelic differences in their regulation.

Distribution of single hits and double hits

By using the top 5% of signal on each array (~30 Mb total) as a cutoff for each mark, we observed 1179 and 1342 H3K4Me2 hits (Supplemental Table 2) as well as 1499 and 1778 DNAm hits (Supplemental Table 3) in GM06991 and T cells, respectively. Average sizes of H3K4Me2 hits were 1934 bp (GM06991) and 1457 bp (T cells), and average sizes of DNAm hits were 1350 bp (GM06991) and 1182 bp (T cells), respectively. For H3K4Me2, ~33% hits were in promoters defined as the regions 1 kb upstream and downstream of transcription start sites (TSSs). In contrast, only ~10% DNAm hits overlapped promoters. In addition, 30%–34% H3K4Me2 hits and 15%–24% DNAm hits overlapped CpG islands.

We then looked at the overlaps of the two markers. We observed 94 double hits in GM06991 and 97 double hits in freshly isolated T cells (Supplemental Table 4), and 32 (34%) double hits were observed in both cell lines. Examining the genes containing double hits, the concordance between the two lines was 43% and 69%, for all genes and imprinted genes, respectively. This suggests that double hits are tissue specific but more generalized for imprinted genes.

Double hits are enriched in imprinted gene regions

Of the double hits, 79% (GM06991) and 70% (T cells) hits were within 1 kb of at least one annotated transcript, as defined as within 1 kb upstream or downstream of at least one transcript of the UCSC known gene track, or anywhere within that transcript. Of these double hits overlapping transcripts, 59% (GM06991) and 44% (T cells) overlapped the promoter regions, i.e., were within 1 kb of the transcriptional start site. About two-thirds (71% in GM06991 and 68% in T cells) overlapped transcripts. Furthermore, 65% (GM06991) and 53% (T cells) overlapped CpG islands. These results suggested that double hits are generally enriched in promoters and CpG islands.

We then validated the array-based double-hit identification by performing quantitative real-time PCR validation on both the ChIP and DNA methylation fractions using antibodies to H3K4Me2 and methylated DNA, respectively, comparing to input DNA. We analyzed three housekeeping genes, *GAPDH*, *RPS18*, and *RPL19*, as positive controls for H3K4Me2 (active), and two genes normally silenced in somatic cells, *GAGE5* and *HIST1H2BA*, as well as satellite 2 repeats (Sat2), as positive controls for DNA methylation. All of the 15 selected double-hit sites discovered on the arrays were confirmed by this validation (Fig. 1). Thus, for all 15 double hits tested by quantitative real-time PCR, the ratio of H3K4Me2 to input ranged from 0.44 to 2.1, similar to the three control housekeeping genes (range 0.59 to 1.0) and much greater than for the three control silenced loci (range 0.06 to 0.11). Similarly, the ratio of MeDIP to input ranged from 0.40 to 2.09, similar to the three control silenced loci (range 0.89 to 1.0) and much greater than that for the three housekeeping genes (range 0.02 to 0.10) (Fig. 1).

We then asked whether double hits were enriched in imprinted genes. If double hits do not distinguish regions of the genome with allelic differences in chromatin (null hypothesis), then there should be no difference in the frequency of double hits near genes, comparing the Imprinting and ENCODE (less 11p15 and X) arrays. However, on examining lymphoblastoid cells, we found a much higher frequency of double hits among the imprinted genes on the Imprinting Array (16/45 genes, 36%), compared with the ENCODE array (27/380, 7.1%), which was statistically significant (ratio 5.0, $P = 6.2 \times 10^{-7}$; Table 1). The results were slightly more striking when examining primary T lymphocytes, 16/45 (36%) imprinted genes on the Imprinting

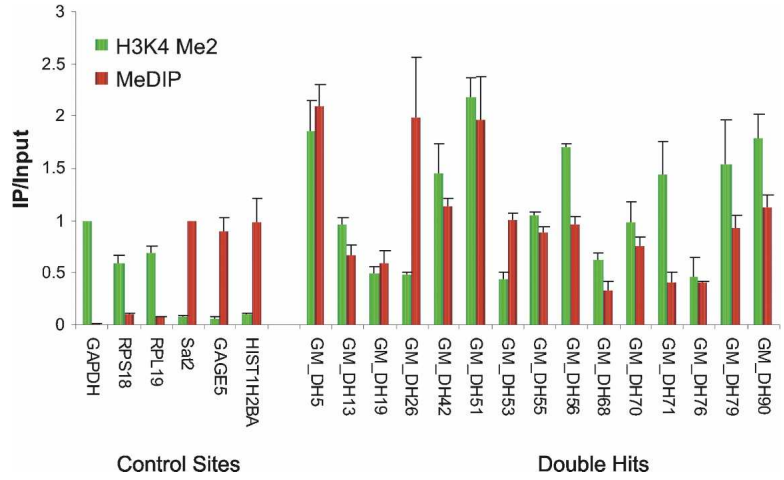

Figure 1. Validation of dual euchromatin/heterochromatin marks ("double hits"). Quantitative real-time PCR validation was performed for both chromatin immunoprecipitation (ChIP) with antibodies to histone H3 lysine-4 dimethylation (H3K4Me2), and methylated DNA immunoprecipitation (MeDIP) fractions, compared with input DNA. We analyzed three housekeeping genes, *GAPDH*, *RPS18*, and *RPL19*, as positive controls for H3K4Me2 (active), and two genes normally silenced in somatic cells, *GAGE5* and *HIST1H2BA*, as well as satellite 2 repeats (Sat2), as positive controls for DNA methylation. The *y*-axis shows the ratio of ChIP or MeDIP to input. A value of 1 represents this ratio for ChIP of *GAPDH* and MeDIP of Sat2. All 15 double-hit sites discovered on the arrays were confirmed by this validation.

Array, compared with 25/380 genes (6.5%) on the ENCODE array (ratio 5.4, $P = 2.8 \times 10^{-7}$; Table 1). Furthermore, when we added the location of known binding sites for CTCF (Kim et al. 2007), an insulator often associated with imprinted genes, the enrichment of "triple hits" in imprinted genes dramatically increased (76-fold in GM06991 and 34-fold in T cells, $P \sim 10^{-9}$), although the sensitivity was reduced (9/45 and 8/45 imprinted genes detected in GM06991 and T cells, respectively). The combined analysis identified 22 known imprinted genes between the two cell types, and the locations of these marks are summarized in Table 2. The concordance of all double hits between the two lines was 33%, but 62% for imprinted genes.

The Imprinting Array also contained 120 genes within 500 kb of known imprinted genes, but not known to be imprinted themselves. The proportion of double hits for these genes analyzing GM06991 was intermediate between that of the ENCODE and imprinted gene groups, with 20/120 (17%), a ratio of 2.3 compared with the ENCODE array ($P = 0.004$). Similar results were found with primary T cells, i.e., 19/120 (16%) genes, a ratio of 2.4 compared with ENCODE genes ($P = 0.003$). This result is consistent with the idea that some of these genes, near other imprinted genes, are likely to be imprinted themselves, and this enrichment is reflected in the number of double hits in this group of genes.

The location of double hits showed a striking relationship to imprinted gene transcripts. In 14 of 16 (73%) and 12 of 16 (75%) cases of known imprinted genes with double hits, in GM06991 and primary T cells, respectively, the double hits were present within 1 kb of the start site of an alternative transcript at the promoter or within the imprinted gene. In contrast, the frequency of double hits within 1 kb of the sites of alternative transcripts was far lower for sites not known to be associated with imprinted genes, six of 27 (22%) and five of 25 (20%) cases of control genes with double hits, in GM06991 and primary T cells, respectively (Supplemental Table 5).

For example, the known imprinted gene *MEST* on chromosome 7, normally expressed from the paternal allele, shows a double hit overlapping the start site of an alternative transcript *MESTIT1* located 4.4 kb inside the sense transcript, as well as a double hit at the *MEST* start site itself (Fig. 2A). Similarly, the imprinted gene *PEG10* on chromosome 7, normally expressed from the paternal allele, also shows a double hit at a promoter shared with the antisense transcript *SGCE* (Fig. 2B). Both are examples of triple hits, in that they also contain a CTCF binding site (Fig. 2). A third example is illustrated by the intensively studied *KCNQ1* gene, expressed from the maternal allele, which includes an imprinted untranslated RNA, *KCNQ1OT1*, expressed from the paternal allele. A triple hit lies 800 bp upstream of *KCNQ1OT1* (Fig. 2C). Another variant of the two-transcript rule is illustrated by *SNRPN* and *SNURF*, two paternally expressed imprinted genes on chromosome 15. A double hit is found at the point of overlap of the two genes, overlapping the start site of *SNURF*, which is nested within *SNRPN* (Fig. 2D). Thus, double hits mark imprinted genes, and their placement is often near the sites of multiple transcriptional start sites.

Double hits are on different alleles of imprinted genes

Given the association of double hits with imprinted genes, one would expect the two hits, DNAm and H3K4Me2, to be on different chromosomes. We showed directly that this is the case for four gene loci, *MEST/MESTIT1*, *GNAS/GNASAS*, *SNRPN/SNURF*, and *KCNQ1/KCNQ1OT1*. For two of these loci, *MEST/MESTIT1* and *GNAS/GNASAS*, we exploited polymorphisms within the transcripts to distinguish maternal and paternal alleles. MeDIP for methylated DNA, and ChIP for H3K4Me2 were performed separately. The purified fraction was then sequenced and compared with the input DNA. For *MEST/MESTIT1*, which contained an (A/G) polymorphism in the double-hit region 1.8 kb from the

Table 1. Enrichment of double and triple chromatin marks in imprinted gene regions

Cell types	No. of genes	Double hits[a]	Ratio	*P*-value[b]	Triple hits[c]	Ratio	*P*-value[b]
GM06991							
Imprinted genes	45	16 (36%)	5.0	6.2×10^{-7}	9 (20%)	76	7.1×10^{-9}
Control genes	380	27 (7.1%)			1 (0.26%)		
Primary T cells							
Imprinted genes	45	16 (36%)	5.4	2.8×10^{-7}	8 (18%)	34	3.3×10^{-7}
Control genes	380	25 (6.6%)			2 (0.52%)		

[a]H3K4Me2 and DNAm.
[b]Numbers shown are significant by Fisher's exact test, which can be corrected for multiple testing by multiplication by four, for the four tests used.
[c]H3K4Me2, DNAm, and CTCF.

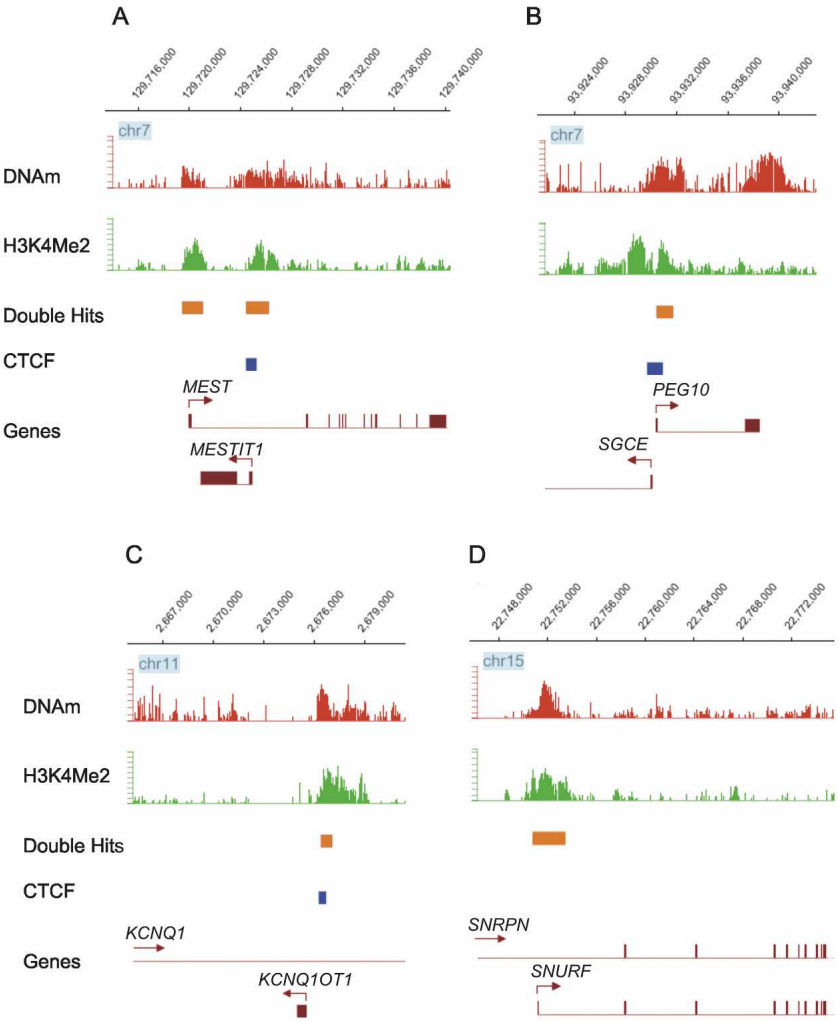

Figure 2. Examples of double hits associated with known imprinted genes. The red and green tracks show $\log_2$ ratios of IP/Input for DNAm and H3K4Me2, respectively. The orange blocks denote double hits and blue blocks denote CTCF binding sites. Genes are shown on the *bottom* in brown: blocks and lines represent exons and introns, respectively. Transcription direction is indicated by arrows. (*A*) *MEST* and *MESTIT1*: A triple hit (including CTCF) was found overlapping the alternative transcript *MESTIT1*, and a double hit was found at the *MEST* start site itself. (*B*) *PEG10* and *SGCE*: A triple hit was found at the shared promoter of these two genes. (*C*) *KCNQ1* and *KCNQ1OT1*: A triple hit was observed 800 bp upstream of *KCNQ1OT1*, with *KCNQ1*. (*D*) *SNRPN* and *SNURF*: A double hit was found at the point of overlap of these two genes at the promoter of *SNURF*.

transcriptional start site of the antisense transcript *MESTIT1*, the G allele purified with the methylated DNA fraction, and the A allele purified with the H3K4Me2 fraction (Fig. 3A). Genotyping of the parental specimens confirmed that the A allele was of paternal origin and the G allele was of maternal origin, consistent with a paternally transcribed allele of the gene. Similarly, *GNAS/GNASAS* contained a (C/T) polymorphism in the double-hit region 570 bp from the transcriptional start site of the paternally expressed antisense transcript *GNASAS*. The C allele purified with the methylated DNA fraction, and the T allele purified with the H3K4Me2 fraction (Fig. 3A). Genotyping of the parental specimens confirmed that the T allele was of paternal origin and the C allele was of maternal origin, again consistent with a paternally transcribed allele of the gene.

A strong argument for allele specificity could also be made even without the use of polymorphisms, as illustrated by the analyses of *KCNQ1/KCNQ1OT1* and *SNRPN/SNURF*. We first performed ChIP with an antibody to H3K4Me2, and then performed bisulfite sequencing on the purified fraction. If one allele were H3K4 methylated and the other DNA methylated, then ChIP for H3K4Me2 would lead to enrichment of the DNA-unmethylated fraction. For both genes, we found equal representations of methylated and unmethylated DNA in the input, i.e., (C/T)G representing unconverted (methylated) and converted (to T) CpG dinucleotides, respectively. However, in the H3K4Me2-enriched fraction, only the unmethylated (TG) sequence was observed (Fig. 3B). While the latter experiment cannot rule out the presence of two equal populations of cells, half with both chromosomes with double hits and half with both chromosomes without any hits, allele specificity is the simplest explanation, and also consistent with the polymorphism data. It is also important to note that while the four genes examined here showed that the double hits were on opposite alleles, this may not be the case for all genes, as the genes studied here were imprinted.

Triple hits are a specific predictor of allele-specific expression

To test whether genes marked by double or triple hits show allele-specific expression, we tested four genes (*SERPINB10, PDIA3, SNX27*, and *GCC1*) by comparing genomic DNA and cDNA in lymphoblastoid cell lines using pyrosequencing (Fig. 4). *SERPINB10,* marked by a triple hit at its 3′ end, showed significant skewing in the cDNA compared with gDNA (*t*-test, $P < 0.01$), whereas the other three genes, which are marked by double hits in the promoters, showed a similar pattern in gDNA and cDNA.

Finally, we examined the X chromosome portion of the ENCODE array separately because one homolog of the chromosome in females is expected to be heterochromatinized. Of 41 genes, 14 showed double hits in primary T cells from a female. In contrast, six showed double hits in the primary T cells derived from a male, suggesting that even on the active X chromosome there is some DNAm at these sites, although much more in the female ($P < 0.05$). This result is consistent with a published study that examined one of the genes within the ENCODE region of the X chromosome, i.e., the *F8* gene (El-Maarri et al. 2007), and found even higher methylation on the male as on the female X chromosome.

Discussion

The major results of this study are: (1) imprinted genes are enriched for dual euchromatin/heterochromatin marks (“double

Table 2. Known imprinted genes near double and triple hits

Gene	Chromosome	Expressed allele	Double hits[a] cell type		Triple hits[b] cell type		Double hit ID[c]
			T[d]	GM[e]	T[d]	GM[e]	
TP73	1	Maternal		Yes			GM_DH2,3
HYMAI	6	Paternal		Yes			GM_DH67
PLAGL1	6	Paternal		Yes			GM_DH67
GRB10	7	Maternal	Yes				T_DH79
MEST	7	Paternal	Yes	Yes	Yes	Yes	GM_DH80, T_DH88,89
MESTIT1	7	Paternal	Yes	Yes	Yes	Yes	GM_DH80, T_DH88,89
PEG10	7	Paternal		Yes		Yes	GM_DH70
SGCE	7	Paternal		Yes		Yes	GM_DH70
CTNNA3	10	Maternal	Yes				T_DH8,9
H19	11	Maternal			Yes		T_DH18
KCNQ1	11	Maternal	Yes	Yes	Yes	Yes	GM_DH13, T_DH23
KCNQ1OT1	11	Paternal	Yes	Yes	Yes	Yes	GM_DH13, T_DH23
SLC22A18	11	Maternal	Yes				T_DH24
SLC22A18AS	11	Maternal	Yes				T_DH24
MEG3	14	Maternal	Yes	Yes	Yes	Yes	GM_DH19, T_DH36
ATP10A	15	Maternal	Yes				T_DH44
SNRPN	15	Paternal	Yes	Yes			GM_DH26, T_DH42
SNURF	15	Paternal	Yes	Yes			GM_DH26, T_DH42
GNAS	20	Maternal	Yes	Yes	Yes	Yes	GM_DH56,57,58, T_DH69,70
GNASAS	20	Paternal	Yes	Yes	Yes	Yes	GM_DH56, T_DH69
L3MBTL	20	Paternal	Yes	Yes			GM_DH55, T_DH67
NNAT	20	Paternal	Yes	Yes			GM_DH51,52, T_DH65

[a]H3K4Me2 and DNA.
[b] H3K4Me2, DNAm, and CTCF.
[c]See Supplemental Table 4.
[d]Primary T cells.
[e]GM00691.

hits"), measured here by DNA methylation (DNAm) and histone H3 lysine 4 dimethylation (H3K4Me2); (2) that the double hits at imprinted genes usually occur at the site of alternative or antisense transcripts; and (3) the double hits may represent allelic differences between the homologous chromosomes. The enrichment of double hits for imprinted genes is approximately fivefold, although this is likely an underestimate, as our control set may contain previously unknown imprinted genes. The enrichment is ~76-fold if one includes CTCF. *SERPIN10B*, the only gene marked by triple hits in the autosomal ENCODE regions of GM06991, showed substantial allelic skewing in gene expression, suggesting that triple hits are a very specific marker to predict ASE imbalance in the human genome.

In this study, we used H3K4Me2 to represent activation-associated markers; however, one could use other markers such as H3K4Me3 to search for the double hits. A limitation of these approaches is their bias toward promoters, although our intent was to examine control regions of genes. Other markers, such as H3K36Me3 would be more suitable for examination across the length of the transcripts. Most existing high-throughput analyses of allele-specific expression, including our own approach (Bjornsson et al. 2008) rely on heterozygous SNPs in either gene bodies for RNA-based assays (Pant et al. 2006; Gimelbrant et al. 2007; Pollard et al. 2008) or in promoters for ChIP-based methods (Kadota et al. 2007; McCann et al. 2007; Maynard et al. 2008), while heterozygous SNPs are generally uncommon in those regions in the genome.

A recent report used an approach comparing H3K4Me3 and H3K9Me3, i.e., using double ChIP rather than comparing ChIP to DNAm. Those authors did not compare imprinted to nonimprinted gene regions; rather, they enriched by ChIP and performed shotgun sequencing, and 17 of the top 20 regions enriched for both marks mapped to known imprinted genes, so it is likely that this approach will also identify loci with allele-specific gene expression (Mikkelsen et al. 2007).

An important general advantage of the approach described in this report is that one can distinguish allele-specific marks even in the absence of a SNP by fractionation by H3K4 ChIP

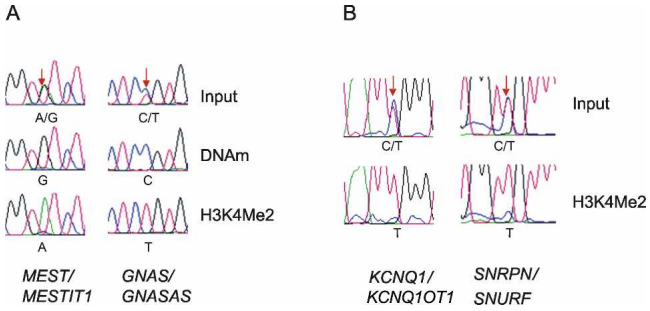

Figure 3. H3K4Me2 and DNAm double hits are allele specific. (*A*) Identification by SNP genotyping. The two marks are on separate alleles of the imprinted genes *MEST* and *GNAS*. H3K4Me2 ChIP and MeDIP were conducted on CEPH line GM10847, which was heterozygous for a SNP within the double-hit regions. The alleles are distinguished with polymorphisms, A/G (rs2301335) for *MEST/MESTIT1* and C/T (rs6026560) for *GNAS/GNASAS*. Note that the A and G alleles are purified separately by ChIP and MeDIP for *MEST/MESTIT1*, and that the T and C alleles are purified separately by ChIP and MeDIP for *GNAS/GNASAS*. The paternal genotype is A/A for *MEST/MESTIT1* and T/T for *GNAS/GNASAS*, consistent with paternal expression of the two genes. (*B*) Identification by serial analysis. Euchromatin was purified by H3K4Me2 ChIP, and the product was analyzed for DNA methylation by bisulfite sequencing. Methylation was assessed using a pseudopolymorphism at CpG sites, TG for unmethylated sites converted by bisulfite, and CG for methylated sites that resist conversion. For both the *SNRPN/SNURF* and *KCNQ1/KCNQ1OT1* double hits, the input DNA is half methylated, while the euchromatin fraction is unmethylated. Arrows indicate polymorphic sites for *A* and CpG sites for *B*.

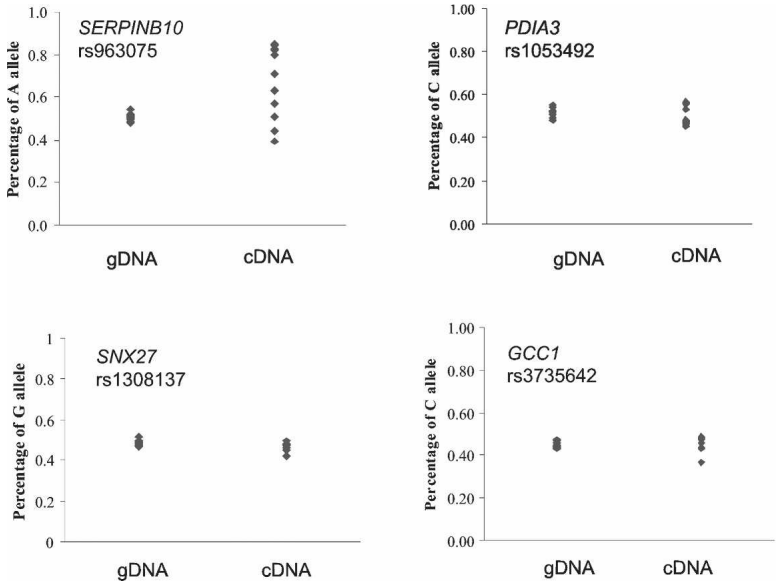

Figure 4. Quantitative allele-specific expression (ASE) analysis of the genes with double hits. ASE was tested by comparing the skewing of gDNA and cDNA using pyrosequencing. The *y*-axis represents the percentage of total expression level of one allele of the heterozygous SNPs. Each diamond represents the measurement of one individual. For *PDIA3*, *SNX27*, and *GCC1*, the expression ratios were not significantly different between gDNA and cDNA. For *SERPINB10*, the expression ratios of cDNA were significantly skewed over gDNA (*t*-test, $P < 0.01$), indicating ASE in this gene.

followed by DNA methylation analysis, as shown in Figure 3B. In contrast, allele-specific binding of RNA polymerase II, with SNP analysis using the Illumina Human Hap300 genotyping array, revealed only five known imprinted genes (Maynard et al. 2008).

Several recent studies have shown a substantial fraction (5%–20%) of genes with significant allelic skewing or monoallelic expression (Lo et al. 2003; Gimelbrant et al. 2007; Serre et al. 2008), which is consistent with our observation that ~7% of the autosomal genes in the ENCODE region contain or are within 1 kb of double hits. A striking difference in our study between known imprinted genes and control genes with double hits was the much greater frequency in the former of proximity of the double hit to an alternative start site of transcription, whether for the same gene, or another gene in sense or antisense orientation. This result resonates with an earlier finding from our group that we could not explain at the time. We had compared the sequence of the imprinted gene domain on chromosome 11p15 with the syntenic region on mouse chromosome 7 and found the frequent occurrence of two CpG islands within or near imprinted genes (Strichman-Almashanu et al. 2002). Indeed, CpG islands were found in the promoters of 14 out of 16 imprinted genes associated with alternative transcripts marked by double hits.

Beside being a useful tool for imprinted or monoallelically expressed gene discovery, the data presented here provide strong support for Sapienza's idea that the two homologs of autosomes are broadly epigenetically distinct for functions not simply related to genomic imprinting, perhaps to assist chromosomal segregation (Pardo-Manuel de Villena et al. 2000). Thus, there are many double hits not found within imprinted genes, and most appear to be different in their proximity to promoters from imprinted genes. Perhaps the establishment of transcriptional start sites near the double hits permits genomic imprinting to arise on the background of epigenetically distinct chromosomes, providing a potential mechanism for Sapienza's hypothesis.

Methods

Cell culture

Human total T cells were isolated from peripheral blood using the Dynal T Cell Negative Isolation Kit (Invitrogen), then cultured in complete medium (PRMI-1640, 20 U/mL penicillin, 20 µg/mL streptomycin, 10% FBS, and 100 U/mL recombinant human IL-2 [Sigma]) for 24 h. The cells were stimulated with 1 µg/mL PHA (Murex) for 40 h and then in complete medium without PHA for four more days. The immortalized lymphoblastoid cell lines GM06991 and GM10847 were cultured in RPMI 1640 medium with 15% FBS.

ChIP-chip

ChIP was performed as described by Kim et al. (2005), but with the following conditions specific to these experiments: sonication to predominant sizes of 500 to 1000 bp; using 200 µg of chromatin for each ChIP with a commercial antibody specific to H3K4Me2 (Abcam); and amplification using the WGA2 kit (Sigma) according to the manufacturer's instruction. Labeling of ChIP and input DNA, hybridization, and scanning were conducted using NimbleGen standard protocols in their service laboratory in Iceland. The ENCODE array was from NimbleGen.

MeDIP on chip

Methylated DNA immunoprecipitation (MeDIP) was performed as described (Weber et al. 2005). Amplification, labeling, hybridization, and scanning were performed the same as for ChIP-chip described above.

Microarray data analysis

We first quantile normalized the Cy5 (IP) and Cy3 (input) raw intensities using a method similar to the one described in Bolstad et al. (2003). The raw intensities in all arrays were made to have the same probe intensity distribution, which we refer to as the reference distribution. In this application we constructed a reference distribution by averaging the Cy3 intensities from all arrays, since Cy3 is the control channel. We then computed log ratios (log Cy5/Cy3).

The next step was to find peaks. On tiling arrays, signals from nearby probes need to be combined to make calls for peaks. The easiest way to combine signals is to average them, so a kernel-smoothing (weighted moving average) approach was used. It was known that if the kernel had the same shape as the true signal it would provide the best sensitivity (Buck and Lieb 2004). After random shearing, the DNA segments are 500–1000 bp in size, and we expect the log ratios around a mark to form a triangular peak with length equal to the average DNA segment length (Buck and Lieb 2004). We therefore used a triangular kernel with a window size of 20 probes (~600 bp) on the log ratios. The probes on the array were segmented by repeat masking during the array design. We ran the kernel smoothing on each segment separately.

The final step was to provide a data-driven definition of a true peak. We used a simple threshold approach in which regions

with the smoothed log ratios above a predetermined constant was defined as a peak. The 95th quantile of the moving averages proved to be a reasonable threshold. We also required a minimum of 200 bp and five probes with smoothed log-ratios above the threshold to consider the region a peak. If a H3K4Me2 peak and a DNAm bump overlapped for at least 100 bp, it was deemed a double hit. The double-hit regions were defined as the union of the matching H3K4Me2 and DNAm peaks.

If two double hits were close to each other (<1 kb), we merged them. A triple hit was defined as a CTCF binding site existing within 1 kb of a double hit. A gene within 1 kb of a double hit was deemed to be a gene with double hits. All of the analyses were performed using the R package. Scripts are available upon request.

Quantitative real-time PCR

Quantitative real-time PCR was performed using SYBR Green PCR master mix (Applied Biosystems) on an ABI 7700 sequence detector. We analyzed three housekeeping genes, *GAPDH*, *RSP18*, and *RPL19*, as positive controls for H3K4Me2 (euchromatin), and two genes normally silenced in somatic cells, *GAGE5* and *HIST1H2BA*, as well as satellite 2 repeats (Sat2), as positive controls for DNA methylation. Fifteen double-hit sites were selected for the validation of array results. Primers were designed using Primer3 software (Rozen and Skaletsky 2000) and the sequences are provided in Supplemental Table 6. A total of 2 ng of IP (ChIP or MeDIP) and input DNA were used for each PCR reaction, and PCR was conducted in triplicates for each sample. The enrichment of IP over input was calculated by difference in the numbers of threshold cycle (Ct) between IP and input, assuming a twofold change by one cycle.

Genotyping

To explore allele specificity of double hits, two sites with heterozygous SNPs in GM10847 were genotyped. H3K4Me2, MeDIP, and input DNA were amplified and sequenced using the BigDye sequencing kit on an ABI 3100 sequencer (Applied Biosystems). Primer sequences are provided in Supplemental Table 6.

Bisulfite sequencing

The products from one H3K4Me2 ChIP and 1 µg of input DNA were bisulfite converted using the EpiTect Bisulfite Kit (QIAGEN) following the manufacturer's instruction. Nested PCR was performed using the primers whose sequences are provided in Supplemental Table 6. PCR products were sequenced using the BigDye sequencing kit on an ABI 3100 sequencer (Applied Biosystems).

Allele-specific expression (ASE) analysis

Quantitative ASE analysis was conducted on a PSQ HS96 Pyrosequencer (Biotage) based on the manufacturer's instructions. Primers are provided in Supplemental Table 6.

Acknowledgments

We thank Tae Hoon Kim and Bing Ren for training on ChIP-chip experimental methods, and Sheri Brandenburg for reading the manuscript. This work was supported by NIH grant P50HG003233 to A.P.F.

References

Bergstrom, R., Whitehead, J., Kurukuti, S., and Ohlsson, R. 2007. CTCF regulates asynchronous replication of the imprinted H19/Igf2 domain. *Cell Cycle* **6:** 450–454.
Bernstein, B.E., Meissner, A., and Lander, E.S. 2007. The mammalian epigenome. *Cell* **128:** 669–681.
Bjornsson, H.T., Albert, T.J., Ladd-Acosta, C.M., Green, R.D., Rongione, M.A., Middle, C.M., Irizarry, R.A., Broman, K.W., and Feinberg, A.P. 2008. SNP-specific array-based allele-specific expression analysis. *Genome Res.* **18:** 771–779.
Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19:** 185–193.
Buck, M.J. and Leib, J.D. 2004. ChIP-chip: Considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83:** 349–360.
Carr, M.S., Yevtodiyenko, A., Schmidt, C.L., and Schmidt, J.V. 2007. Allele-specific histone modifications regulate expression of the Dlk1-Gtl2 imprinted domain. *Genomics* **89:** 280–290.
El-Maarri, O., Becker, T., Junen, J., Manzoor, S.S., Diaz-Lacava, A., Schwaab, R., Wienker, T., and Oldenburg, J. 2007. Gender specific differences in levels of DNA methylation at selected loci from human total blood: A tendency toward higher methylation levels in males. *Hum. Genet.* **122:** 505–514.
The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640.
Esteller, M. 2007. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat. Rev. Genet.* **8:** 286–298.
Feinberg, A.P. 2007. Phenotypic plasticity and the epigenetics of human disease. *Nature* **447:** 433–440.
Feinberg, A.P. and Tycko, B. 2004. The history of cancer epigenetics. *Nat. Rev. Cancer* **4:** 143–153.
Gimelbrant, A., Hutchinson, J.N., Thompson, B.R., and Chess, A. 2007. Widespread monoallelic expression on human autosomes. *Science* **318:** 1136–1140.
Kadota, M., Yang, H.H., Hu, N., Wang, C., Hu, Y., Taylor, P.R., Buetow, K.H., and Lee, M.P. 2007. Allele-specific chromatin immunoprecipitation studies show genetic influence on chromatin state in human genome. *PLoS Genet.* **3:** e81. doi: 10.1371/journal.pgen.0030081.
Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. 2005. A high-resolution map of active promoters in the human genome. *Nature* **436:** 876–880.
Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. 2007. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128:** 1231–1245.
Kouzarides, T. 2007. Chromatin modifications and their function. *Cell* **128:** 693–705.
Li, B., Carey, M., and Workman, J.L. 2007. The role of chromatin during transcription. *Cell* **128:** 707–719.
Lo, H.S., Wang, Z., Hu, Y., Yang, H.H., Gere, S., Buetow, K.H., and Lee, M.P. 2003. Allelic variation in gene expression is common in the human genome. *Genome Res.* **13:** 1855–1862.
Maynard, N.D., Chen, J., Stuart, R.K., Fan, J.B., and Ren, B. 2008. Genome-wide mapping of allele-specific protein-DNA interactions in human cells. *Nat. Methods* **5:** 307–309.
McCann, J.A., Muro, E.M., Palmer, C., Palidwor, G., Porter, C.J., Andrade-Navarro, M.A., and Rudnicki, M.A. 2007. ChIP on SNP-chip for genome-wide analysis of human histone H4 hyperacetylation. *BMC Genomics* **8:** 322. doi: 10.1186/1471-2164-8-322.
Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553–560.
Pant, P.V., Tao, H., Beilharz, E.J., Ballinger, D.G., Cox, D.R., and Frazer, K.A. 2006. Analysis of allelic differential expression in human white blood cells. *Genome Res.* **16:** 331–339.
Pardo-Manuel de Villena, F., de la Casa-Esperon, E., and Sapienza, C. 2000. Natural selection and the function of genome imprinting: Beyond the silenced minority. *Trends Genet.* **16:** 573–579.
Perk, J., Makedonski, K., Lande, L., Cedar, H., Razin, A., and Shemer, R. 2002. The imprinting mechanism of the Prader-Willi/Angelman regional control center. *EMBO J.* **21:** 5807–5814.
Pollard, K.S., Serre, D., Wang, X., Tao, H., Grundberg, E., Hudson, T.J., Clark, A.G., and Frazer, K. 2008. A genome-wide approach to identifying novel-imprinted genes. *Hum. Genet.* **122:** 625–634.
Rozen, S. and Skaletsky, H. 2000. Primer3 on the WWW for general

users and for biologist programmers. *Methods Mol. Biol.* **132:** 365–386.
Serre, D., Gurd, S., Ge, B., Sladek, R., Sinnett, D., Harmsen, E., Bibikova, M., Chudin, E., Barker, D.L., Dickinson, T., et al. 2008. Differential allelic expression in the human genome: A robust approach to identify genetic and epigenetic *cis*-acting mechanisms regulating gene expression. *PLoS Genet.* **4:** e1000006. doi: 10.1371/journal.pgen.1000006.
Strichman-Almashanu, L.Z., Lee, R.S., Onyango, P.O., Perlman, E., Flam, F., Frieman, M.B., and Feinberg, A.P. 2002. A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. *Genome Res.* **12:** 543–554.
Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L., and Schubeler, D. 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* **37:** 853–862.
Xin, Z., Allis, C.D., and Wagstaff, J. 2001. Parent-specific complementary patterns of histone H3 lysine 9 and H3 lysine 4 methylation at the Prader-Willi syndrome imprinting center. *Am. J. Hum. Genet.* **69:** 1389–1394.

Received March 28, 2008; accepted in revised form September 4, 2008.