INTRODUCTION

Mammalian autosomal genes are normally present in duplicate - one copy is inherited from the mother and one copy is inherited from the father - and both copies of each gene are usually regulated in the same manner independently of the parent of origin. A notable exception to this general rule of genetic equality observed in classical Mendelian inheritance is the phenomenon of genomic imprinting. Imprinting is the process by which the maternally- and paternally-inherited copies of the same gene become epigenetically labeled in different ways, resulting in mono-allelic expression of that gene from only one locus. Imprinted genes are typically always transcribed from one allele and silenced at the other in a parent-specific fashion, although more complex expression patterns have also been observed, such as tissue-specific imprinting.

Although imprinting only occurs in a small minority of autosomal genes, estimated to be at least 100 in mouse (Reik and Walter, 2001), it is an important biological process that is essential for normal mammalian development. Disruption of imprinting has been implicated in a number of human genetic diseases, including Prader-Willi syndrome, Angelman syndrome, and Beckwith-Wiedemann syndrome, as well as different types of cancer (Reik and Walter, 2001). It has also been suggested that defects in imprinting could be involved in many neurological disorders that exhibit parent-of-origin effects, such as autism, bipolar affective disorder, epilepsy, and schizophrenia (Reik and Walter, 2001). Furthermore, epigenetic mechanisms underlying imprinted gene expression have also been shown to be involved in the regulation of non-imprinted genes during development, indicating that knowledge of imprinting could contribute to our appreciation of more general modes of mammalian gene regulation (Ehrlich, 2003).

Investigations into the mechanisms responsible for ‘writing’ genomic imprints have demonstrated that DNA methylation is a major epigenetic modification used to distinctly label the maternally- and paternally-inherited copies of imprinted genes. In mammalian genomes, DNA methylation occurs on

cytosines at 5'-CG-3' dinucleotides (CpG), and most CpG sites in the genome are normally methylated (Bird, 2002). Since methylated CpG mutates into TpG (or, equivalently, the reverse complement CpA) at a high rate, the overall observed frequency of genomic CpG is approximately 1%, which is roughly 25% of the rate expected by chance according to the single base frequencies and an assumption of independence (Bird, 2002). However, there are some GC-rich regions of the genome that exhibit elevated frequencies of CpG around 8%. These regions, known as CpG Islands (CGIs), were first identified experimentally as 'HpaII-small fragments' because of their increased sensitivity to digestion by HpaII, a restriction enzyme which cleaves specifically at non-methylated restriction sites containing CpG (Bird, 1986). However, CGIs are now typically identified computationally based on their sequence properties, conventionally defined as regions of 200bp or more with GC-content ≥ 50% and an observed/expected CpG ratio ≥ .60 (Gardiner-Garden and Frommer, 1987). CGIs are often associated with the 5' ends of genes, and most CGIs are usually unmethylated, which accounts for their elevated rate of CpG (Bird, 1986). In contrast, many CGIs associated with imprinted genes are the sites of differential DNA methylation on only one chromosome that marks imprinted loci differently according to the parent of origin (Reik and Walter, 2001).

The importance of CGIs for imprinting has been demonstrated through a number of studies in mice and humans. Deletion of CGIs has been shown to disrupt imprinted gene regulation at several loci, including *H19/Igf2*, the PWS/AS locus, the BWS locus, and *Igf2r/Air*, indicating a central role for CGIs in imprinting (Bestor, 2000). Experiments involving bisulfite sequencing and methylation-sensitive restriction enzymes have clearly shown that differential methylation of CGIs occurs at imprinted loci (Reik and Walter, 2001). The functional importance of methylation has been established by analysis of mouse knockout mutants lacking *Dnmt1*, the main maintenance methyltransferase, and *Dnmt3L*, which encodes a non-catalytic protein in the DNA methyltransferase family with sequence similarity to the *de novo* methyltransferases *Dnmt3a* and *Dnmt3b*. Mouse embryos defective in *Dnmt1* display a loss of proper imprinted

gene expression and die in mid-gestation, implying that DNA methylation is necessary to maintain the epigenetic status of maternal and paternal genes at imprinted loci (Li *et al*, 1993). Embryos from *Dnmt3L*-/- female mice lack differential methylation of maternal DNA at imprinted regions (whereas global DNA methylation remains unperturbed) and die soon after implantation with evident growth abnormalities (Bourc'his *et al*, 2001). Together, these investigations demonstrate that genomic imprints can be established by germline-specific methylation of CGIs in imprinted regions which differentially marks the paternal and maternal copies of imprinted genes.

Imprinted genes have typically been identified in a variety of ways (Reik and Walter, 2001). Some individual imprinted genes have been fortuitously discovered by knockout experiments of genes that subsequently displayed parent-specific expression. Imprinted genes have also been identified based on their location near other known imprinted genes or in chromosomal regions associated with imprinting phenotypes.

However, there are two main types of screens that have been used to systematically search for imprinted genes, both of which usually rely on comparisons between uniparental embryos containing either only paternally-derived chromosomes (androgenetic embryos) or only maternally-derived chromosomes (parthenogenetic embyos). The first type of screen searches for differences in DNA methylation patterns between androgenetic and parthenogenetic embryos after digestion with methylation-sensitive restriction enzymes using techniques such as representational difference analysis or restriction landmark genome scanning. The second type of screen for identifying imprinted genes is based on comparisons of cDNA levels between uniparental embryos of maternal and paternal origin. Recently, a large-scale screen of this type was used to identify 2,114 candidate imprinted genes by using microarrays to detect differential expression of 27,663 FANTOM2 mouse cDNAs between the total tissue of 9.5 day old parthenogenetic and androgenetic embryos (Nikaido *et al*, 2003). Although bioinformatic techniques have not been previously used to discover novel imprinted genes, previous research efforts have aimed to identify

sequence characteristics unique to imprinted regions in the human genome.

Significant differences in the guanine and cytosine, CpG, and repeat content of imprinted regions relative to control loci were recently observed in two separate studies (Greally, 2002 and Ke *et al*, 2002), both of which were based on examination of the features of large sequence windows, typically 50kb in size, spanning imprinted loci. The most striking observation was a marked decrease in the content of short interspersed transposable elements (SINEs) near imprinted genes. A possible model explaining this is that SINE accumulation in imprinted regions was selected against because non-specific methylation of SINEs could deleteriously interfere with parent-specific methylation of imprinted loci (Greally, 2002). Although direct repeat sequences have also been historically described to correlate with imprinted regions, the generality of these observations remains in question (Okamura *et al*, 2000 and Arnaud *et al*, 2003). However, no previous published investigations have directly examined the sequences of imprinted CGIs themselves in order to explore the possibility that CGIs at imprinted loci could differ in composition from control CGIs at non-imprinted loci.

Based on the clear importance of CGIs to genomic imprinting, it seems reasonable to hypothesize that significant differences in the sequence composition of imprinted and control CGIs could be detected. There are two main reasons that imprinted CGIs might differ in sequence composition from control CGIs. First, the fact that only one copy of imprinted CGIs is unmethylated might lead to observable differences in nucleotide composition when compared to CGIs for which both copies are unmethylated; after all, the normally unmethylated status of most CGIs is believed to explain their unique sequence features of GC- and CpG-richness. Secondly, it is also possible that the distribution of binding sites for trans-acting factors involved in epigenetic regulation could be enriched or reduced within imprinted CGIs, leading to detectable differences in their sequence composition. Indeed, *CTCF*, an 11-zinc finger transcription factor which regulates chromatin boundaries and may act as both a repressor and activator, has been shown to be involved directly in imprinted regulation of the *Igf2/H19* locus by binding to a differentially methylated CGI region (Schoenherr *et al*, 2003).

Additional CTCF-binding motifs have also been identified within differentially-methylated domains at other loci (Kim *et* al, 2003 and Hikichi *et* al, 2003) and it is possible that other trans-acting factors, such as those involved in the establishment of imprints, may bind specifically to imprinted CGIs to mediate epigenetic regulation of imprinted genes. We therefore decided to analyze the sequence features of CGIs from imprinted and control loci in mouse in order to look for significant characteristics which may be involved in genomic imprinting.

MATERIALS & METHODS

CpG Island Identification

CGIs are GC-rich DNA sequences that exhibit an elevated frequency of CpG dinucleotides and are often unmethylated and associated with the 5' ends of genes. To identify CGIs in mouse, the gene sequence and upstream region (50kb from the annotated start of the gene) for all 13,112 autosomal mouse genes that were defined by EnsEMBL as 'known genes' were extracted from the mus_musculus_core_9_3 database of December 2002 (www.ensembl.org). All X-linked genes were excluded from the analysis because the process of X chromosome inactivation shares many features with genomic imprinting, including epigenetic regulation and differential methylation of CpG islands (Lee, 2003). CGIs were identified in autosomal sequences using the cpgplot program (http://www.emboss.org) with default parameters. The imprinted genes *H19*, *Peg3*, and *Snrpn* all contain known differentially methylated CGIs which could not be detected using default parameters, so an alternative minimum length parameter (minlen=100) was used for these genes. Nearby cpgplot-identified islands were merged together to form a single CGI if the overall sequence properties of the resulting island satisfied the conventionally accepted criteria for CGIs (Length ≥ 200bp, GC-content ≥ 0.5, CpG Obs/Exp ratio ≥ 0.6). Duplicate and overlapping islands were then removed to yield a set of 13,665 unique CGIs.