Retrotransposition and genomic imprinting

Michael Cowley and Rebecca J. Oakey

Advance Access publication date 29 June 2010

Abstract

Studies of large imprinted clusters, such as the Gnas locus, have revealed much about the significance of DNA methylation, transcription and other factors in the establishment and maintenance of imprinted gene expression. However, the complexity of such loci can make manipulating them and interpreting the results challenging. We review here a distinct class of imprinted genes, which have arisen by retrotransposition, and which have the potential to be used as models for the dissection of the fundamental features and mechanisms required for imprinting. They are also of interest in their own right, generating diversity in the transcriptome and providing raw material upon which selection can act.

Keywords: retrotransposition; retrogenes; genomic imprinting; epigenetics

IMPRINTED GENE CLASSIFICATION

One of the most striking demonstrations of the functional non-equivalence of the parental genomes comes from the study of mice with uniparental disomies (UPD) in which zygotes possess two paternal or two maternal copies of a specific chromosome or chromosomal region. For example, when both copies of proximal chromosome 11 are of maternal origin, neonates exhibit growth retardation, whilst neonatal overgrowth is apparent in mice inheriting two copies of the region through the paternal line [1]. The association of an imprinting effect with a specific gene locus was first revealed for the Igf2 gene on chromosome 7 [2]. Whilst the growth effects of Igf2 had been studied for some years previously, its imprinted status was not realised until mice with a targeted disruption of the gene were engineered. Inheritance of this mutation engendered a dwarfing phenotype, but only when the disrupted allele was transmitted through the paternal line. Molecular analyses confirmed that transcription at the locus was exclusively from the paternally inherited chromosome in the embryo and placenta, with the exception of the choroid plexus.

To date, 143 imprinted genes have been identified in mouse [3], although estimates of how many exist are wide ranging; 2114 candidates were identified by microarrays comparing gene expression in parthenogenetic and androgenetic embryos (possessing two maternal and two paternal genome copies, respectively) [4], but earlier estimates of between 100 and 200 are more consistent with current data [5]. Imprinted genes may be classified according to their function. Many are involved in the processes of growth and development, metabolism or social behaviour, with some imprinted genes spanning multiple categories. Correct expression from the maternally inherited copy of Gria2, for example, is required to establish both normal size and body proportions during development, as well as appropriate regulation of insulin signalling during postnatal life [6].

In addition to their function, imprinted genes can be classified according to their complexity of locus organization. Some imprinted genes exhibit conserved clustering in the genome, such as the Gnas locus on mouse chromosome 2. The regulation of imprinted expression from this cluster is complex and highly refined (reviewed in [7]). Oocyte-specific
methylation of two germline differentially methylated regions (DMRs) is dependent upon transcription of the protein coding, maternally expressed, gene Nesp [8]. Methylation of these two regions is hierarchical, identifying one as an imprinting control region (ICR) [9]. A third DMR is established in somatic tissues, which is methylated on the paternal chromosome [10]. Three paternally expressed genes are transcribed from the locus; these are Gnasxl, Exon 1A and Nespas. The latter is a non-coding antisense RNA whose function probably includes the silencing of Nesp expression in cis [7]. A further level of complexity is demonstrated by the Gnas gene itself, which is biallelically expressed in most tissues, but paternally silenced in others, including the proximal renal tubules, anterior pituitary gland, thyroid gland and ovary [11–14]. This suggests that tissue-specific enhancer elements may also influence expression at the locus, which are likely to be under the influence of the DMRs.

This complex regulation of the Gnas locus has been elucidated through the generation of a number of knockout mouse lines, observations from which, alongside those from other experimental systems, have been pieced together to create a comprehensive model. Whilst we have known for some time that differential methylation at Gnas is associated with imprinted expression [10], it has only recently become apparent that transcription through the locus in oocytes is necessary for differential methylation itself to become established [8].

Other imprinted genes exhibit a less complex organization, offering simpler models for the study of imprinting. Mcts2 is a mono-exonic gene expressed from paternal chromosome 2 [15]. It is positioned within intron 4 of H13 which is also imprinted (Figure 1A). Mcts2 has arisen through retrotransposition of an mRNA molecule originating from the Mcts1 gene on the X chromosome [15]. Through an unknown mechanism, Mcts2 has associated with a CpG island which is differentially methylated in oocytes and sperm, and is responsible for the imprinting of both Mcts2 and the ‘host’ gene H13 [16]. A similar organization has been observed for other imprinted loci, at least some of which have arisen through retrotransposition (reviewed in [17]). Understanding such retrogene–host pairs is valuable for two reasons. First, they are relatively simple loci and thus may permit manipulation and analysis in a manner which is more challenging for large imprinted clusters. As an example, consider an experiment to assess the importance of a chromatin-binding protein in the control of imprinted gene expression, using siRNA to knockdown transcripts encoding this protein of interest. Understanding the relationship between protein binding and allele-specific transcription at a complex locus is initially challenging because of the difficulties in distinguishing between the direct effects of knockdown and the indirect effects resulting from perturbed imprinting elsewhere at the locus. An understanding of the relatively simple relationship between retrogenes and their hosts permits a much easier analysis of the effects of knockdown on imprinted gene expression.

As such, retrogene–host pairs have the potential to provide excellent models for examining the fundamental molecular mechanisms responsible for the establishment of imprinting. Additionally, they may play an important role in genome evolution. Imprinting at Mcts2 is responsible for controlling alternative polyadenylation of H13 transcripts, contributing to transcript diversity and providing raw material on which selective pressures can act [16].

**RETROTRANSPOSITION**

Five murine imprinted genes are known to have arisen through retrotransposition: Mcts2, Nap1l5, U2af1-rs1, Inpp5f1v2 and Peg10 [15, 18–20]. These are different from the imprinted genes Rtl1 and Peg10 that are ‘retrotransposon-like’. This means that Rtl1 and Peg10 share sequence similarity with transposable elements, specifically the sushi-class [21], whereas the five retrogenes are derived from cellular mRNA molecules which did not possess autonomous retrotransposon activity. Here, we address briefly how autonomous and non-autonomous retrotransposition can occur.

Mobile DNA elements may be classified into two groups: DNA transposons and retrotransposons. DNA transposons are sequences encoding a transposase enzyme which permits transposition of the sequence by a ‘cut and paste’ mechanism. Thus, DNA transposons are not duplicated during a transposition event, and are relatively rare in the genome. Retrotransposons are transcribed to produce an mRNA molecule, which is subsequently reverse transcribed and re-integrated into the genome in a new location, thereby duplicating the sequence. Retrotransposons may be further classified according to the presence or absence of long terminal repeats.
The most abundant of the non-LTR retrotransposons, and most relevant to this discussion, is the long interspersed nuclear element 1 (LINE1 or L1), which alone accounts for \( \frac{20}{\%} \) of the mouse genome [22]. Most of these copies are inactive because of sequence truncations or the accumulation of mutations, but about 3000 are estimated to be retrotransposition competent [23].

Self-propagation of L1 elements is enabled by the transcription of functional proteins from two open-reading frames (ORFs) (Figure 1B). An RNA-binding protein is encoded by ORF1, whilst ORF2 encodes a protein with both endonuclease and reverse transcriptase functions. In the cytoplasm, these proteins assemble with the L1 mRNA encoding them to produce a ribonucleoprotein particle (RNP), which is transported to the nucleus where L1 integration into the genome can occur. Integration is essentially random, although the L1 endonuclease, which is responsible for initiating the process exhibits a preference for the sequence 5'-TTTTA-3' [24]. Cleavage produces a single-stranded T-rich overhang which can hybridize with the polyadenylated tail of the L1 mRNA, priming reverse transcription. Further molecular details of the integration event are described by Ding and colleagues [25].

The proteins encoded by L1 elements preferentially associate with L1 mRNA [26], making L1 elements autonomous retrotransposons. However, they are also responsible for the movement of other repeat sequences, such as B1 elements, which are non-autonomous. Further, the L1-encoded machinery can potentially be utilized for the retrotransposition of any cellular mRNA species. In the majority of cases, retrotransposition of protein-coding mRNA has generated processed pseudogenes, defined as genes with sequence similarity to a parent gene but no retention of function [25]. This usually occurs because the new gene copy cannot be transcribed due to the loss of promoter elements. Additionally, processed pseudogenes often possess remnants of 3'-polyadenylation sequences which are not functional when reverse transcribed. The
absence of a selective pressure results in neutral evolution of processed pseudogenes.

Retrotransposition can also generate a novel gene which is transcribed and translated, producing a functional protein. These are termed retrogenes [25]. Transcription may occur because of the retention of a cryptic promoter within the retrotransposed sequence. Alternatively, integration may have occurred downstream of a pre-existing promoter element, or a promoter evolved later.

RETROGENES AND IMPRINTING

Retrogenes are relatively rare in mammalian genomes; ~105 have been discovered in the mouse [27]. It is, therefore, initially surprising that at least five of them are imprinted. Of additional interest is that four of these genes share three common features: derivation from a parental gene on the X chromosome, a location within the intron of a host gene and a 5′-CpG island [15]. Indeed, these features alone were used to identify the imprinted gene *Mcts2* using a bioinformatic approach [15]. The CpG island is necessary for the establishment of allele-specific expression; the ablation of DNA methylation at the *Mcts2* locus in embryos derived from *Dnmt3L*−/− mothers resulted in biallelic expression of *Mcts2* and loss of allele-specific polyadenylation of the host gene *H19* [16]. Derivation from a parental gene on the X chromosome correlates with the over-representation of functional retrogenes arising from this chromosome [27] and might reflect the functional significance of the genes (discussed below). Finally, the intronic location might indicate the importance of transcription through the locus for the establishment of differential methylation, as demonstrated for *Nesp* at the *Gnas* locus [8]. *Peg12* is not located within an intron, but has retrotransposed into a previously imprinted domain on chromosome 7, presumably accounting for its imprinted expression status [20].

Whilst these three features appear to be necessary for imprinting of retrogene–host pairs, they may not be sufficient; several retrogenes with all three features are biallelically expressed (A. Oates et al., in preparation). Previous studies have suggested that L1 elements are more prevalent in imprinted regions [28], and thus might contribute to the establishment of differential expression. This might occur because repetitive elements are methylated and silenced by the cell as part of a defence against their activity. Indeed, there are some common mechanisms utilized for methylation establishment at imprinted loci and repetitive elements, although the enzymes involved differ depending on the loci or repetitive elements in question. For example, during male germ cell development, the *de novo* DNA methyltransferase Dnmt3a is required for the correct methylation of the *H19* and *Dkl1-Gtl2* DMRs, as well as for B1 elements [29]. However, both Dnmt3a and Dnmt3b are necessary for establishing methylation at the *Rag2fl* locus and L1 repeat elements [29]. The relative importance of the *de novo* DNA methyltransferases in germ cell development also differs between males and females. Disruption of Dnmt3L, for example, results in moderate to severe hypomethylation of all DMRs and repetitive elements in prospermatogonia [29], whilst repetitive elements are unaffected by the loss of Dnmt3L in oocytes [30]. The association between L1 elements and imprinted loci might result from this partial conservation of methylation mechanism; for example, L1 elements may help to recruit the machinery required for establishing methylation at DMRs, although there is currently no experimental evidence in support of this. Despite the reported increased prevalence of L1 elements in imprinted regions [28], no differences in the number of nearby repeat elements were found for imprinted versus non-imprinted retrogenes sharing the three common features (A. Oates et al., in preparation), suggesting that this is not part of the mechanism responsible for retrogene imprinting.

An alternative explanation for the imprinting of a subset of retrogenes might be dosage control during meiotic sex chromosome inactivation (MSCI). This occurs during spermatogenesis and describes the silencing of transcription from most genes on the sex chromosomes [31], resulting from their exclusion to sex bodies during the pairing of homologous chromosomes in pachytene. Thus, any retrogenes arising from the sex chromosomes which are beneficial during spermatogenesis might be positively selected for and more likely to retain a function after retrotransposing into an autosome. This compensation theory would account for the over-representation of functional retrogenes arising from the X chromosome [27]. One such example is that of the parent gene *Utp14a* on the X chromosome and its associated retrogene *Utp14b* on chromosome 1. Whilst *Utp14b* does not possess a promoter sequence similar to its parent, it has integrated into the intron of a host gene *Acsl3*, the promoter of
which drives the expression of one transcript variant [32]. Other Utp14b transcript variants appear to have evolved later as a consequence of a novel promoter arising at the locus, which drives expression specifically within the germ cells of the testis [33]. Utp14a expression is greatly reduced in pachytene spermatids, with a concomitant increase in total Utp14b transcript levels, suggestive of functional compensation during MSCI. Consistent with this, male mice with a frameshift mutation in Utp14b exhibit reduced fertility due to failure of spermatogenesis [32, 34]. Given that in somatic cells the parent gene is normally expressed from a single X chromosome, an important consideration is how the correct expression dosage of the retrogene can be achieved during MSCI. Imprinting is one mechanism by which retrogene expression could effectively be halved. Whilst there is no evidence that Utp14b is imprinted, this retrogene has utilized the host promoter and acquired other novel promoters, which might be weaker than that of its parent gene Utp14a. Mcts2, on the other hand, has retained its parental promoter, and thus imprinting might be necessary for appropriate dosage control.

Consistent with this hypothesis, Mcts2 has retained 94% amino acid sequence identity with its parent gene Mcts1 and is up-regulated during pachytene when Mcts1 is silenced ([35]; M. Cowley and R.J. Oakey, unpublished data). However, the strong purifying selection exhibited by Mcts2 is not mirrored by Nap1l5, U2af1-rs1 or Inpp5f_v2, which have followed distinct evolutionary pathways, suggesting that parent gene compensation during MSCI may not apply to these genes (R. McCole and R.J. Oakey, submitted for publication).

**FUTURE DIRECTIONS**

To fully appreciate the evolutionary significance of imprinted retrogenes, our future studies must focus on both elucidating the mechanisms responsible for establishing and maintaining imprinting, and addressing the functional consequences of their monoallelic expression. For example, how does imprinting of Mcts2 result in allele-specific alternative polyadenylation of the host gene H13? Certainly a transcriptional interference model is likely, but other epigenetic factors may be involved. Recently, specific histone modifications have been associated with the control of alternative splicing of some human genes [36]. It is conceivable that the imprinted expression of Mcts2 might direct allele-specific enrichment of such histone modifications and might be part of the mechanism responsible for H13 allele-specific alternative polyadenylation. Studying these retrogene–host pairs may help to further our understanding of the relationship between transcription and imprinting.

Whilst studies of imprinted retrogene–host pairs have predominantly focused on the mouse genome, there is evidence that this type of locus organization extends to humans. The human orthologues of three imprinted retrogenes in the retrogene–host pair configuration are imprinted, namely MCTS2, INPP5F_V2 and NAPIL5 [15], although their importance in influencing host gene expression has not yet been explored. Interestingly, a processed pseudogene located in intron 2 of the human retinoblastoma 1 (RBI) gene has recently been identified, which is methylated on the maternal chromosome and acts as a promoter for an alternative RBI transcript arising from the paternal chromosome [37]. This builds on the studies of mouse retrogenes, suggesting that pseudogenes can also control imprinted expression of their hosts. This understanding may lead to the discovery of other imprinted loci with similar configurations.

In parallel, investigating the functions of proteins encoded by the retrogenes may indicate the reason(s) for their imprinted status. This may also provide an indication of the relevance of this mechanism to genome evolution. One clear outcome of imprinted retrogene insertion into the intron of a host is the generation of host transcript diversity. In the case of H13, the parent-of-origin-specific transcripts would be expected to give rise to different protein products [16], but it is currently not clear if all are functionally relevant. Careful manipulation of the retrogene–host pairs in knockout mouse models will help to address this issue.

**CONCLUSIONS**

Whilst the traditional view has been that most imprinted genes are organized into clusters, the imprinted retrogene–host pairs, of which there are at least four examples in the mouse genome, represent a distinct class. Their genomic organization is simpler, providing an excellent model for the dissection of the fundamental features and mechanisms required for the establishment of imprinting.
References


