Transcription of intragenic CpG islands influences spatiotemporal host gene pre-mRNA processing

Samuele M. Amante, Bertille Montibus, Michael Cowley, Nikolaos Barkas, Jessica Setiadi, Heba Saadeh, Joanna Giemza, Stephania Contreras-Castillo, Karin Fleischanderl, Reiner Schulz and Rebecca J. Oakey [ID]*

Department of Medical and Molecular Genetics, King's College London, Guy's Hospital, London SE1 9RT, UK

ABSTRACT

Alternative splicing (AS) and alternative polyadenylation (APA) generate diverse transcripts in mammalian genomes during development and differentiation. Epigenetic marks such as trimethylation of histone H3 lysine 36 (H3K36me3) and DNA methylation play a role in generating transcriptome diversity. Intragenic CpG islands (iCGIs) and their corresponding host genes exhibit dynamic epigenetic and gene expression patterns during development and between different tissues. We hypothesise that iCGI-associated H3K36me3, DNA methylation and transcription can influence host gene AS and/or APA. We investigate H3K36me3 and find that this histone mark is not a major regulator of AS or APA in our model system. Genomewide, we identify over 4000 host genes that harbour an iCGI in the mammalian genome, including both previously annotated and novel iCGI/host gene pairs. The transcriptional activity of these iCGIs is tissue- and developmental stage-specific and, for the first time, we demonstrate that the premature termination of host gene transcripts upstream of iCGIs is closely correlated with the level of iCGI transcription in a DNA-methylation independent manner. These studies suggest that iCGI transcription, rather than H3K36me3 or DNA methylation, interfere with host gene transcription and pre-mRNA processing genomewide and contributes to the spatiotemporal diversification of both the transcriptome and proteome.

INTRODUCTION

Between 20–25 000 protein coding genes have been identified in the human and mouse genomes that give rise to ~200 000 transcripts in tissue- and developmental stage-specific combinations (1). These transcripts can be generated via the use of alternative promoters (2) as well as co-transcriptional pre-mRNA processing mechanisms that include alternative splicing (AS) and alternative polyadenylation (APA) (3–5). Estimates based on transcriptome analyses reveal that ~90% of human transcripts undergo AS (6) and that APA occurs in at least 70% of mammalian pre-mRNAs (7,8). AS involves the differential inclusion of exons and sometimes introns in the mature mRNA. APA refers to the polyadenylation of transcripts originating from the same gene but that differ in their 3' end (5). Both AS and APA are dependent on specific sequences recognised by the cellular machinery (9,10). APA events can occur either at 3' untranslated regions (UTRs) or intragenic locations. The incidence of both 3'UTR-APA and intragenic polyadenylation (IPA) varies across tissues and cell types providing a way to diversify both the transcriptome and the proteome (11).

Epigenetic modifications of histone tail residues and cytosine bases can influence AS (12–20) and APA (21–23) in developing tissues. Imprinted genes are particularly useful models for the dissection of epigenetic gene expression regulation (22,23). There are around 130 genes in mouse and 90 genes in human that are subject to genomic imprint-

*To whom correspondence should be addressed. Tel: +44 20 7848 8147; Email: rebecca.oakey@kcl.ac.uk
Present addresses:
Samuele M. Amante, Centre for Genomics and Child Health, Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK.
Michael Cowley, Center for Human Health and the Environment and Department of Biological Sciences, North Carolina State University, Raleigh, NC 27695, USA.
Nikolaos Barkas, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.
Heba Saadeh, Department of Computer Science, KASIT, University of Jordan, Jordan.
Joanna Giemza, Oncodesign SA, 20 Rue Jean Mazen, 21000 Dijon, France.
Stephania Contreras-Castillo, INSERM U1218 ACTION, University of Bordeaux, 33076 Bordeaux, France.
Karin Fleischanderl, Greiner Bio-One GmbH, 4550 Kremsmünster, Austria.
Karin Fleischanderl, University of Applied Sciences - FH Campus Wien, Favoritenstrasse 226, 1100 Wien, Austria.
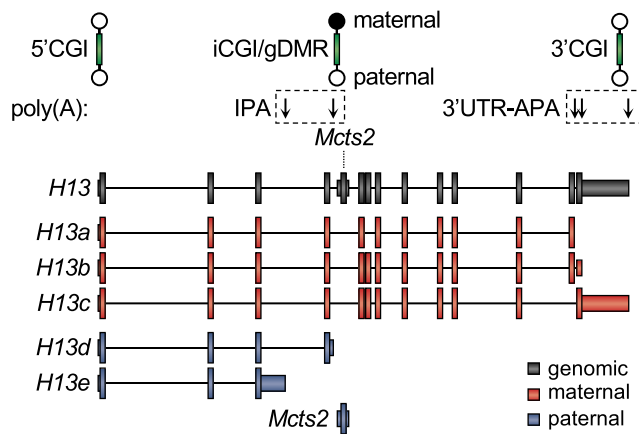
Figure 1. Schematic representation of the *Mcts2/H13* imprinted locus. Three CpG islands (CGIs) (green) are present at this locus: one is associated with the promoter of *H13* (5′CGI), a second with its 3′UTR (3′CGI) and a third, intragenic one with the promoter of *Mcts2* (iCGI). The iCGI is a germline differentially methylated region (gDMR) as it becomes methylated in oocytes. *H13a*, *H13b* and *H13c* are transcribed from the maternal (red) allele and undergo 3′UTR alternative polyadenylation (3′UTR-APA). *H13d*, *H13e* and *Mcts2* are transcribed from the paternal (blue) allele. *H13d* and *H13e* respectively retain portions of intron 4 and 3 and undergo intronic polyadenylation (IPA). Downward pointing arrows represent *H13* alternative polyadenylation sites.

ing, many of which are crucial for normal development (24). Monoallelic expression of these genes is coordinated by allele-specific DNA methylation of imprinting control regions (ICRs). Most ICRs acquire differential DNA methylation in the germline (25). Maternally methylated ICRs overlap with promoters, whereas paternal ICRs are found in intergenic regions (25). The active and silent alleles of imprinted genes share the same DNA sequence and are present within the same cellular environment, implying that allelic differences in gene expression are the consequence of epigenetic differences between the alleles (25).

Mcts2 is an imprinted, monoexonic gene that has resulted from the retrotransposition of *Mcts1* into the fourth intron of the *H13* gene, an event that occurred over 90 million years ago (26). *H13*, referred to as the host gene, is also imprinted. The ICR of *Mcts2/H13* is an intragenic CpG island (iCGI) that overlaps with the promoter of *Mcts2* and undergoes DNA methylation in the female germline (22,27). As a consequence, *Mcts2* is silent on the maternal allele that generates three *H13* transcripts (*H13a*, *H13b*, *H13c*), which undergo 3′UTR-APA downstream of the iCGI using the canonical polyadenylation sites of the *H13* gene to generate full length transcripts. In contrast, on the paternal allele, the iCGI is unmethylated and transcriptionally active (22,27). This results in two paternal *H13* transcripts (*H13d*, *H13e*) that undergo intron retention and IPA upstream of the iCGI (Figure 1). The *Mcts2/H13* locus therefore provides a paradigmatic example of how DNA methylation at iCGIs can influence AS and APA.

It has been shown that H3K36me3 coordinates tissue-specific usage of alternative exons (14,19) and prevents intron retention, possibly by facilitating the recognition of weak splice donor sites to ensure introns are correctly spliced out (16,28,29). Importantly, H3K36me3 is de-

posited along the body of actively transcribing genes (30) where it mediates the recruitment of the *de novo* DNA methylation machinery (31) to inhibit spurious transcription initiation from intragenic promoters such as iCGIs (32).

CGIs are regulatory regions typically found at promoters but also at intragenic and intergenic locations. Promoter CGIs (5′CGIs) are generally unmethylated (33) but when methylated, this is usually associated with repression of transcription (34). Biochemical approaches have identified numerous CGIs within host genes (35,36). iCGIs, which are highly conserved between mouse and human, show tissue-specific patterns of DNA methylation and transcriptional initiation during development, possibly indicating involvement in the spatiotemporal regulation of host gene expression, AS and/or IPA (35–37). The DNA methylation status of these iCGIs is dependent upon their host gene transcriptional activity (33). Host gene transcription across iCGIs is required for the recruitment of *de novo* DNA methylation enzymes and the silencing of said iCGIs. However, the intrinsic capacity of iCGIs to initiate transcription negatively correlates with their sensitivity to this transcription-mediated DNA methylation mechanism (33). The hypermethylated state of iCGIs is associated with low levels of preinitiation RNA Pol II occupancy, H3K36me3 enrichment and lack of H3K4me3 (33). Conversely, iCGIs that retain an unmethylated state show increased preinitiation RNA Pol II binding, are enriched in H3K4me3 and lack H3K36me3 (33). Importantly, when two promoters are located in relatively close proximity, similarly to iCGI/host gene promoters, a transcriptional process initiating from the stronger promoter can have suppressive influence over a second transcriptional process initiating from the weaker promoter, in a phenomenon known as transcriptional interference (TI) (38–41).

Here, we sought to investigate H3K36me3 and transcription as mediators in the generation of alternative transcripts at the *Mcts2/H13* model locus and more broadly. We also sought to determine the influence of intragenic transcription and DNA methylation on host gene pre-mRNA processing. This was achieved through the identification of over 4000 host genes harbouring an iCGI in the mouse genome. The activity of these iCGIs has been found to be tissue- and developmental stage-specific and, for the first time, we demonstrate that the abundance of host gene transcripts terminating upstream of iCGIs is closely correlated with the level of iCGI transcription. These studies suggest that iCGI transcription, rather than H3K36me3 or DNA methylation, interfere with host gene transcription and pre-mRNA processing genomewide, this in turn provides a means to enable spatiotemporal diversification of both the transcriptome and proteome.

MATERIALS AND METHODS

Cell culture

NIH/3T3 cells were cultured at 37°C and 5% CO_2 in DMEM high glucose (Gibco, 41965039) supplemented with 10% FBS (Gibco, 26140079) and 1× penicillin–streptomycin (Gibco, 15070063). Cells were harvested using 1× trypsin-EDTA (Gibco, 25300054).

RNA interference

On day one, NIH/3T3 cells were seeded at 15 625 cells/cm^2 in a well of a six-well plate. After 24 h, 3.75 μl Lipofectamine 3000 (Invitrogen, L3000008) was diluted in 125 μl Opti-MEM (Gibco, 31985070). *Setd2*-specific or Scrambled siRNAs (OriGene, SR423523) were diluted in 125 μl Opti-MEM to a final concentration of 10 nM. Diluted Lipofectamine 3000 and siRNAs were mixed and incubated for 20 min at room temperature. The siRNA–lipid complex was added to the cells and incubated for 48 h. Transfections were carried out in triplicate and gene expression was assayed by RT-qPCR.

Western blot

Protein extracts were obtained by incubating 10^4 cells/μl in loading buffer (50 mM Tris–Cl pH 6.8, 2% (w/v) SDS, 0.1% (w/v) bromophenol blue, 10% (v/v) glycerol, 100 mM DTT) at 98°C for 5 min. Proteins were separated on 4–12% SDS-PAGE and transferred onto PVDF membrane. Successful transfer of proteins was confirmed by Ponceau S staining. The membranes were blocked in 5% (w/v) TBS-T BSA, incubated with the appropriate primary and secondary antibodies (Supplementary Table S1) and washed with TBS-T. Horseradish peroxidase-conjugated secondary antibody was detected by Pierce ECL Western Blotting Substrate (Thermo Scientific, 32106) coupled with iBright FL1500 Imaging System (Thermo Fisher Scientific, A44241).

RNA extraction and cDNA synthesis

Total RNA was extracted from pelleted cells using RNeasy Mini Kit (Qiagen, 74104). Membrane-bound genomic DNA was digested with RNase-Free DNase Set (Qiagen, 79254). First strand cDNA was synthesised using a Proto-Script II First Strand cDNA Synthesis Kit (New England Biolabs, E6560) using oligo d(T)$_{23}$ primers and 500 ng total RNA.

RT-qPCR

RT-qPCR was performed on a QuantStudio 6 Flex Real-Time PCR System (Applied Biosystems) in a 10 μl reaction including 1 μl cDNA, 1× TaqMan Gene Expression Master Mix (Applied Biosystems, 4369016) and appropriate TaqMan probes (Supplementary Table S2).

Statistical analysis. For every condition, three biological and two technical +RT replicates plus one technical -RT replicate were assayed. Relative gene expression was measured using the $2^{-\Delta Ct}$ method (42). Statistical significance was determined by unpaired *t*-test assuming consistent scatter and correcting for multiple comparisons using the Holm–Sidak method. Alpha was defined as equal to 0.05.

RNA-seq

Library preparation and sequencing. After RNA extraction, RNA integrity was measured with 2200 TapeStation (Agilent) using high sensitivity RNA ScreenTape and reagents (Agilent, 5067–5579, 5067–5580 and 5067–5581). TruSeq Stranded mRNA Library Prep (Illumina, 20020594) and TruSeq RNA Single Indexes Set A (Illumina, 20020492) were used to prepare two Scrambled and two *Setd2* knockdown libraries from 900 ng of total RNA per library. Libraries were validated using High Sensitivity D1000 ScreenTape system (Agilent, 5067–5584 and 5067–5585) and KAPA Library Quantification Kit (Kapa Biosystems, 07960140001). All samples were sequenced on one Illumina HiSeq 2500 lane.

Differential gene expression analysis. Raw data in the format of FastQ files were subject to quality control using FastQC (0.11.5) (43). Adapter sequences were removed using BBDuk from the BBtools kit (38.22) and the output underwent a second round of quality control by FastQC. mRNA-seq reads were quantified with Kallisto (0.44.0) (44). Differential gene expression (DGE) analysis was carried out using Sleuth (0.30.0) (45). Gene ontology analysis was carried out using PANTHER (14.0) (46) and statistical significance was calculated with the Fisher's exact test and the Bonferroni correction for multiple testing.

Alternative splicing analysis. All mRNA-seq datasets were aligned to the GRCm38 reference genome by HISAT2 (2.1.0) (47) and quality controlled using Picard (2.18.26). MAJIQ (1.1.7a) and Voila (1.1.9) were utilised to detect local splice variants (LSVs) as previously described (48).

DNA extraction and bisulfite sequencing

Genomic DNA (gDNA) was extracted form pelleted cells using DNeasy Blood and Tissue Kit (Qiagen, 69504). 500 ng of gDNA were bisulfite converted using EZ DNA Methylation-Gold Kit (Zymo Research, D5005).

Amplification and cloning. Nested PCR was performed with appropriate primers (Supplementary Table S3) using OneTaq Hot Start DNA Polymerase (New England Biolabs, M0481) and OneTaq Standard Reaction Buffer (New England Biolabs, M0481). Amplicons were cloned into a pGEM-T Easy plasmid (Promega, A1360). Ligation reactions were transformed into chemically competent cells for blue–white colony screening.

Sequencing. After colony PCR, amplicons were subject to enzymatic clean-up with ExoSAP-IT PCR Product Cleanup Reagent (Applied Biosystems, 78200.200.UL) and sequenced with T7 and SP6 primers using a BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, 4337454) in a 3730xl DNA Analyzer (Applied Biosystems).

Methylation analysis. Sanger sequencing results were uploaded to BISMA (49) and the analysis was executed using default parameters.

Identification of iCGI/host gene pairs

Strand-specific polyadenylated or non-polyadenylated RNA-seq datasets generated by the ENCODE project (50) were utilised. 30 different tissues and/or developmental

stages were available for mouse (Figure 5C) and 18 cell lines for human (Figure 5D). Genome assemblies mm9 and hg19 were screened using the Known Genes Canonical table of the UCSC genome browser in conjunction with CGI annotations from Illingworth *et al.* (36). Reads mapping upstream, across and at the iCGI were counted.

Pearson correlation coefficients

Pearson correlation coefficients (ρ) were calculated using the following formula:

$$\text{corr}\left(\log(Ui+1) - \log(Ai+1), \frac{iCGIi - \text{median}(iCGI+1)}{\text{mad}(iCGI+1)+0.1}\right)$$

This correlation coefficient measures the linear association between the transcriptional activity of the iCGI and the ratio of host gene expression upstream (U) and across (A) the iCGI calculated across available conditions (i). ρ values obtained with this formula were used to generate density plots in Figure 5B and Supplementary Figure S5.

Structural analysis of CGIs

C+G content, CpG_{obs}/CpG_{exp} ratios and CpG density were calculated for each CGI using publicly available data (36). The distribution analysis was conducted using R and an in-house script.

Processed whole genome bisulfite sequencing data analysis

Processed whole genome bisulfite sequencing data (bed-Methyl files) were downloaded from ENCODE datasets. Percentage of CpG methylation at iCGIs ($\pm$5kb) was taken into account. The heatmap was generated using the package EnrichedHeatmap (51) in R and an in-house script. Accession numbers: GSE82669_ENCFF555SRI (lung); GSE82465_ENCFF356GTX (forebrain E13.5); GSE82905_ENCFF467JWP (hindbrain E13.5); GSE83043_ENCFF326GKU (midbrain E13.5); GSE82855_ENCFF101JDR (neural tube E13.5); GSE82658_ENCFF096RYX (heart); GSE82571_ENCFF051OZQ (stomach).

ChIP-seq data analysis

H3K36me3 and H3K4me3 mouse E14.5 brain ChIP-seq data from the ENCODE project (50) were utilised for average profiling over the host gene body and flanking 3000 bp. Average host gene body length was calculated and divided into three intervals: upstream iCGI; iCGI and downstream iCGI. The logarithm of the fold enrichment over input DNA was calculated at single base resolution for each locus and then scaled to the respective average interval length.

RESULTS

Global H3K36me3 depletion does not lead to increased intron retention or intragenic polyadenylation

In mammals, SETD2 is a histone methyltransferase that can interact with the C-terminal domain of phosphorylated RNA Pol II and deposit H3K36me3 along the body of actively transcribed genes with increasingly greater occupancy towards their 3′ ends (52). Previous studies have linked H3K36me3 depletion to increased intron retention in human kidney tumours (28,29). Intron retention can be used as a proxy for IPA when measured in the context of polyadenylated mRNA since a large fraction of intron retention events map to the 3′ end of transcripts (53). For this reason, only polyadenylated mRNA was selected as a template for the synthesis of cDNA used in gene expression analyses.

To investigate the involvement of H3K36me3 in intron retention and IPA in mouse, NIH/3T3 immortalised mouse fibroblasts were co-transfected with two *Setd2*-specific siRNAs. The efficiency of the knockdown was confirmed by RT-qPCR, which showed an ∼80% reduction in *Setd2* mRNA levels in knockdown (KD) samples compared to wild type (WT) samples (Figure 2A). This was further validated by transcriptome analysis (Supplementary Figure S1). As a result, H3K36me3 was globally depleted (Figure 2B). KD cells showed subtle but significant changes in the transcription of 4266 transcripts, of which 1782 (41.77%) were downregulated and 2484 (58.23%) were upregulated (Figure 2C). Among the downregulated transcripts, 432 (10.13%) showed a $\log_2$(fold change) $\leq$–1 and 346 (8.11%) of the upregulated transcripts showed a $\log_2$(fold change) $\geq$+1 (Supplementary Figure S2A). Transcripts of canonical splicing and polyadenylation factors were readily detectable at similar levels in both WT and KD samples (Supplementary Figure S2B). Gene ontology (GO) analysis using all significantly upregulated genes revealed enrichment of biological processes associated with translation and rRNA metabolism (Figure 2D). Furthermore and consistent with previous findings (54,55), modification of histones and mRNA metabolism were also among the upregulated biological processes (Supplementary Figure S2C). GO analysis using all significantly downregulated genes in KD samples revealed that the most affected biological processes were DNA replication, cell cycle and cell migration (Figure 2E). SETD2 had previously been shown to methylate α-tubulin (56) and to act as a key regulator of DNA mismatch repair in G1 and early S phase (57). Therefore, increased mitotic and cytokinetic defects were expected. To investigate the role of H3K36me3 in AS, splicing analysis was conducted and resulted in the detection of 136 significant changes in KD samples. However, only three genes (*Ciz1*, *Kctd9* and *Rrm2*) displayed changes in intron retention and in all three cases this type of alternative splicing event was less frequent in the KD than in the WT (Supplementary Figure S3). Additionally, the percentage of intronic bases present in both WT and KD RNA-seq datasets was calculated and the result validated the splicing analysis. Examination of the mRNA-seq metrics from the final quality control step determined the percentage of bases mapping to intronic sequences in WT and KD samples to be 0.111% ($\pm$ 0.00837) and 0.0994% ($\pm$ 0.00705), respectively. Taken together, our findings suggest that H3K36me3 does not play a major role in intron retention of polyadenylated mRNA isoforms or IPA in NIH/3T3 cells. These results differ from previous studies that show increased intron retention in human kidney tumours characterised by *SETD2* mutations (28,29). Importantly, those observations are based
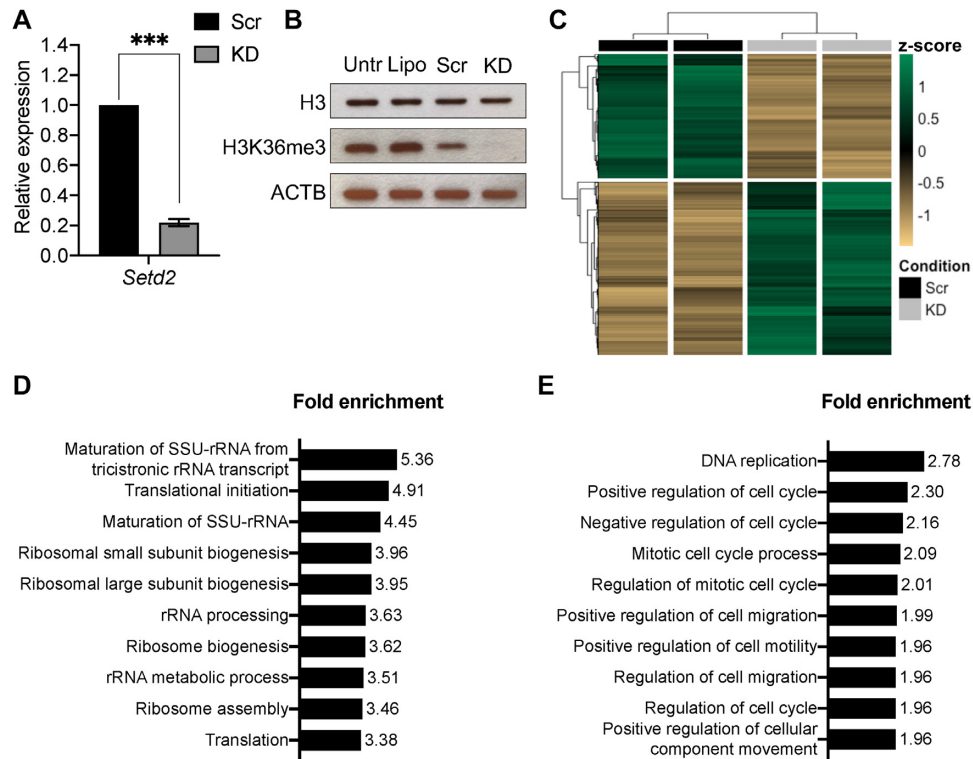
Figure 2. (**A**) *Setd2* mRNA levels assessed by RT-qPCR 48 h post transfection. All data are normalised to Ct values for *Actb*. Data are given as mean $2^{-\Delta\Delta Ct}$ values $\pm$ SEM of three independent experiments. ***$P < 0.001$ compared with Scrambled group by unpaired *t*-test. Scr, scrambled control; KD, knockdown. (**B**) Western blot of whole cell lysate showing effective depletion of H3K36me3 upon knockdown of *Setd2*. Total histone H3 levels are unaffected. ACTB was used as a loading control. Untr, untreated cells; Lipo, transfection vehicle only; Scr, scrambled control; KD, knockdown. (**C**) RNA-seq heatmap of significant differentially expressed transcripts 48 h post transfection. Values are given as row-wise standard-normalised fragments per kilobase of transcript per million mapped reads (z-score). Scr, scrambled control; KD, knockdown. (**D**) Top ten upregulated biological processes determined by GO analysis (PANTHER). See Supplementary Table S6 for a complete list of GO ID terms. (**E**) Top ten downregulated biological processes determined by GO analysis (PANTHER). See Supplementary Table S7 for a complete list of GO ID terms.

on the analysis of RNA-seq libraries generated from total RNA rather than polyadenylated mRNA, which would account for unstable and/or short-lived, intron-retaining transcripts. However, when those datasets were processed through our pipeline, the results were consistent with our findings (Supplementary Table S4) and contradicted the original studies.

More detailed analyses were conducted to investigate a possible locus-specific involvement of H3K36me3 in premRNA processing. Taking advantage of previous studies on the iCGI/host gene pair *Mcts2*/*H13* (22,26,27), this locus was selected as a model system. Using publicly available ChIP-seq data from hybrid immortalised mouse fibroblasts (129/Sv × CAST/Ei) (58), H3K36me3 deposition along the *Mcts2*/*H13* locus was interrogated. This histone mark is enriched on the maternal (129/Sv) allele at the introns that are retained in the paternally expressed (CAST/Ei) *H13d* and *H13e* transcripts (Figure 3). H3K36me3 has previously been shown to compensate for weak splice donor sites via the recruitment of specific alternative splicing factors (16). When *H13* splice donor sites were scored , those present at exon three and four were found to be the weakest compared to the consensus sequence (Supplementary Table S5). It was therefore hypothesised that H3K36me3 may facilitate the recognition and usage of weak splice donor sites at this locus. If this is the case, depleting H3K36me3 would
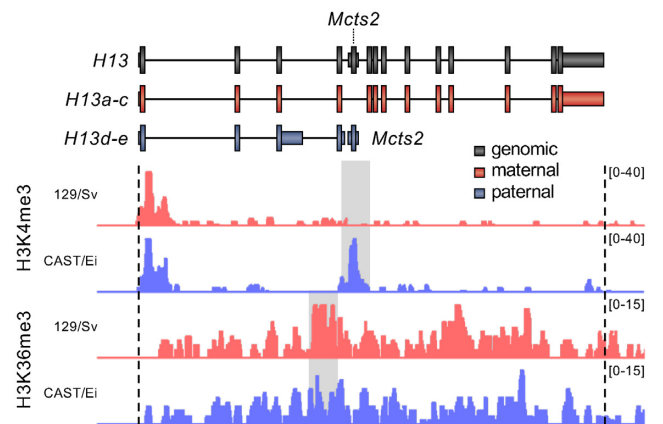

Figure 3. ChIP-seq profiles for H3K4me3 and H3K36me3 from hybrid immortalised mouse fibroblasts mapped to the maternal (129/Sv) or paternal (CAST/Ei) allele of *Mcts2*/*H13*. Genomic (black) and packed maternal (red) and paternal (blue) tracks are shown at the top. Coverage values have been normalised by input and are indicated on the y-axis. Allele-specific chromatin mark enrichments are highlighted in gray.

be expected to lead to decreased *H13a, H13b* and *H13c* isoforms and increased *H13d* and *H13e* isoforms (Figure 1). In other words, depletion of this histone mark would re-
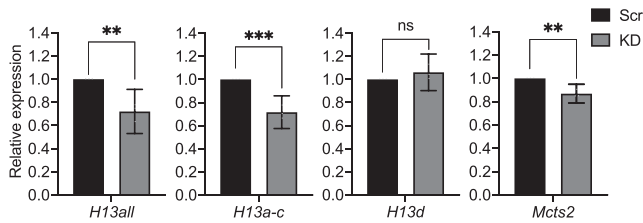
Figure 4. mRNA levels of target transcripts assessed by RT-qPCR 48 h post transfection. All data are normalised to Ct values for *Actb*. Data are given as mean $2^{-\Delta\Delta Ct}$ values $\pm$ SEM of three independent experiments. ns, $P > 0.05$; **$P = 0.01$; ***$P < 0.001$ compared with Scrambled group by unpaired *t*-test. Scr, scrambled control; KD, knockdown.

sult in an increase in intron retention and IPA. *H13a*, *H13b* and *H13c* (collectively referred to as *H13a-c*) mRNA levels measure 3′UTR-APA, whereas *H13d* and *H13e* expression are a proxy for intron retention and IPA (Figure 1). *H13e* was detectable only at very low levels by RT-qPCR and was not utilised in these expression analyses. Using the same siRNA-treated NIH/3T3 cells as above, RT-qPCR was employed to determine the effects of the knockdown on *Mcts2/H13*. At *H13*, total mRNA levels were reduced by 28% (Figure 4, *H13all*). *H13* isoforms terminating downstream of the iCGI also decreased by 28% (Figure 4, *H13a-c*). This decrease was expected since *H13a-c* constitute the vast majority of *H13* transcripts (22,27,59). Contrary to expectations however, *H13d* mRNA levels remained unchanged (Figure 4, *H13d*). Expression from the iCGI was minimally affected (Figure 4, *Mcts2*) and its DNA methylation profile was not altered (Supplementary Figure S4). Taken together, these findings recapitulated those provided by the transcriptome analysis and confirmed that global H3K36me3 depletion via siRNA knockdown approaches may not be sufficient to lead to increased intron retention or IPA.

iCGI activity influences host gene transcription and IPA

The *Mcts2/H13* locus provides a model for studying APA and the role of iCGIs in alternative transcript formation. To understand the involvement of iCGI expression in transcriptome diversity more broadly, we devised a strategy to estimate the number of loci in the genome where iCGIs reside in host gene bodies and are associated with tissue-specific gene expression patterns. Tissue-specific variation in transcription at iCGIs was determined using ENCODE RNA-seq data for 30 mouse tissues and developmental stages (50). 4033 host genes were identified genomewide that harbour iCGIs located at least 1 kb from their TSS and 500 bp from the start of their last exon. A negative control dataset was generated consisting of 1079 well-defined loci with either a corresponding protein entry in PDB or validation by RefSeq that neither harbour an iCGI nor overlap with other genes. Then, an artificial iCGI was simulated at each of the negative control loci, with its position and size randomly drawn from the normalized position and size distributions of the actual iCGIs. For each iCGI/host gene pair identified in the genome, RNA-seq reads mapping *upstream*, *across* and at the *iCGI* were counted (Figure 5A). Additionally, the ratio of reads mapping *upstream* and

across the iCGI was calculated for each gene harbouring an iCGI and used to determine the proportion of host gene transcripts terminating upstream of the iCGI or traversing the iCGI. Pearson correlation coefficients (ρ) were calculated between *upstream:across* ratios and the number of RNA-seq reads mapping to the iCGIs themselves.

The distribution of ρ values of host genes with iCGIs was found to differ substantially from that of the negative control dataset (null distribution) (Figure 5B, *upper*). The null distributions are centred around zero and have a moderate range, indicating that this approach has a sufficient level of specificity and that extreme ρ values are unlikely to occur by chance. The distributions originating from iCGI/host gene pairs are skewed to the right (Figure 5B), showing that increased iCGI transcription positively correlates with increased *upstream:across* ratios. In order to select a subset of candidate loci for further investigation, a strict cut-off value of 0.59, equal to the maximum ρ value observed in the negative data set, was imposed: 1722 (21%) iCGI/host gene pairs scored above this threshold. The sensitivity of this approach was validated by the detection of the imprinted pair *Nap1l5/Herc3* ($\rho = 0.94$) among the significant loci. The promoter of *Nap1l5* is an iCGI that, when transcriptionally active, leads to *Herc3* IPA (23). The same analysis was conducted using ENCODE RNA-seq data sets from 18 human cell lines (50) and resulted in similar findings (Figure 5B, *lower* and Supplementary Figure S5). Transcription from a large number of iCGIs positively correlates with *upstream:across* host gene transcription ratios.

Using publicly available data (36), we determined C+G content, CpG observed over CpG expected ratio (CpG_{obs}/CpG_{exp}) and CpG density of the CGIs identified in this study. As previously shown (36), iCGIs present slightly reduced C+G content, CpG_{obs}/CpG_{exp} ratio and CpG density when compared to other CGIs (Supplementary Figure S6, *left*). In order to determine whether these features could be important in the apparent correlation between iCGI expression and *upstream:across* host gene transcription ratios, their distribution was interrogated in the context of iCGIs/host gene pairs. Thus, we observed no difference between iCGIs whose expression significantly correlates with *upstream:across* host gene transcription ratios ($\rho > 0.59$) and other iCGIs ($\rho \leq 0.59$) (Supplementary Figure S6, *right*). This indicates that these features have probably no involvement in the regulation of iCGIs expression and in their influence on host gene transcription.

Although the model *Mcts2/H13* locus is subject to genomic imprinting, we have shown through this analysis that the majority of the identified iCGI/host gene pairs (over 4000) are not imprinted, suggesting that iCGI influence on host gene polyadenylation is widespread and conserved across at least two mammalian species.

iCGIs show DNA methylation-independent spatiotemporal transcriptional activity

The transcriptional activity at iCGIs from iCGI/host gene pairs with $\rho > 0.59$ was compared across all tissues, developmental stages and cell lines. Both mouse and human datasets showed striking tissue-specific iCGI activity with
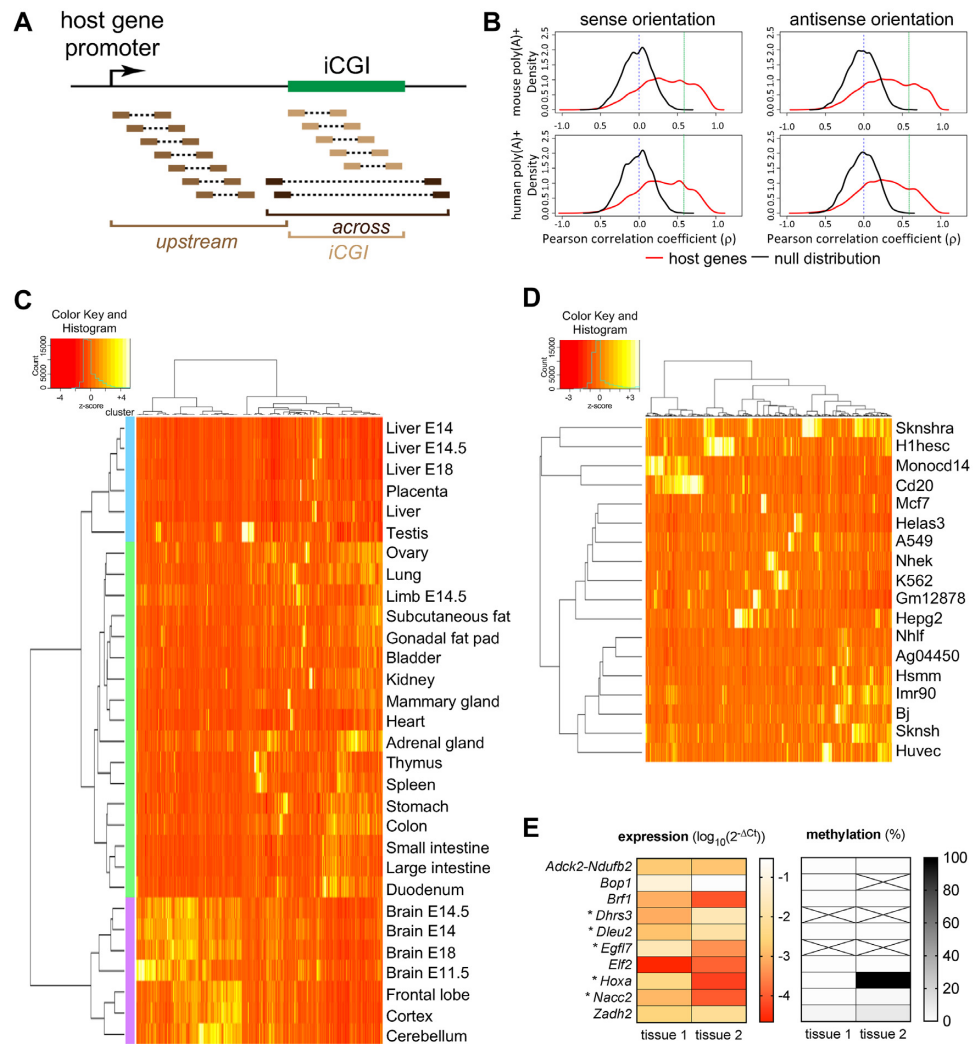
Figure 5. (**A**) Schematic representation of data collection. RNA-seq reads at iCGI/host gene pairs were divided into three groups according to the region they were mapped to: *upstream*, *across* or at the *iCGI*. (**B**) Pearson correlation coefficients (ρ) between transcription from the *iCGI* and transcription *upstream:across* the iCGI. ρ values were calculated in both sense (left) and antisense (right) orientations with respect to the host gene across 30 mouse tissues (upper) and 18 human cell lines (lower) using RNA-seq data from polyadenylated [poly(A)+] transcripts. A vertical blue dashed line is at $\rho = 0$. A strict cut-off is represented by a vertical green dashed line at $\rho = 0.59$, equal to the maximum ρ value observed in the null distribution. (**C**) RNA-seq heatmap illustrating tissue- and developmental stage-specific transcriptional activity of murine iCGIs within host genes with $\rho > 0.59$. Values are given as column-wise standard-normalised fragments per kilobase of transcript per million mapped reads (*z*-score). Tissues are from adult mice, unless specified. (**D**) RNA-seq heatmap illustrating cell type-specific transcriptional activity of human iCGIs within host genes with $\rho > 0.59$. Values are given as column-wise standard-normalised fragments per kilobase of transcript per million mapped reads (*z*-score). (**E**) Ten iCGIs from C were selected and labelled with the name of their host gene. Transcription from the iCGIs was measured by RT-qPCR in two tissues (left). All data are normalised to Ct values for *Actb*. Data are given as $\log_{10}$ of mean $2^{-\Delta Ct}$ values of three independent experiments. *, expression is consistent with RNA-seq data in C. DNA methylation was measured by sequencing of bisulfite converted genomic DNA and is given as percentage values (right). Crossed out cells indicate that DNA methylation could not be determined.

iCGIs often being highly expressed in a single tissue or cell line (Figure 5C and D). Hierarchical clustering grouped mouse tissues and developmental stages from the same organ or system within the same cluster (Figure 5C). The hierarchical clustering was reproducible when using only iCGI/host gene pairs with iCGIs fully mapping within one intron of their host genes, excluding iCGIs partially or fully overlapping with host gene exons (Supplementary Figure S7A). This is important since it is not possible to discriminate between same-strand RNA-seq reads from host genes and iCGIs when the latter fully or partially overlap with host gene exons. Additionally, when intronic iCGIs were

subjected to hierarchical clustering according to their expression level, a large number of iCGIs that showed high expression in brain tissues and developmental stages grouped together (Supplementary Figure S7A). GO term analysis using host genes harbouring these iCGIs revealed enrichment of brain-specific biological processes (Supplementary Figure S7B and Table S8), confirming that these host genes are involved in relevant cell specification processes.

In order to establish a role for DNA methylation in the regulation of iCGIs transcriptional activity in mouse, ten candidate iCGI/host gene pairs that showed the highest correlation between iCGI transcription and *up-*
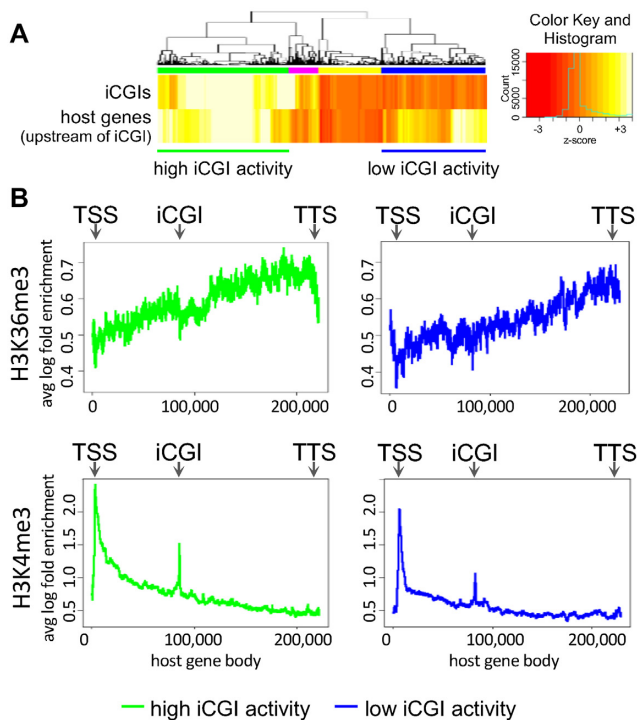
Figure 6. (**A**) RNA-seq heatmap from E14.5 mouse brain illustrating the transcriptional activity of iCGI/host gene pairs showing a ρ > 0.59 between transcription from the *iCGI* and transcription *upstream:across* of the iCGI (see main text and Figure 5A). Highly expressed iCGIs within active or inactive host genes are highlighted with green or pink bars, respectively. Moderately expressed iCGIs within active or inactive host genes are highlighted with blue or yellow bars, respectively. Values are given as column-wise standard-normalised fragments per kilobase of transcript per million mapped reads (*z*-score). (**B**) Average H3K36me3 (upper) and H3K4me3 (lower) ChIP-seq profiles for iCGI/host gene pairs from A in which the iCGIs are fully intronic. Reads are mapped to the body of active host genes harbouring intronic iCGIs that are transcribed at high (left) or low (right) levels. TSS, transcription start site; TTS, transcription termination site.

stream:across ratios were selected for further analysis. For each iCGI, two tissues were assayed that exhibited the highest and lowest measured expression levels based on RNA-seq data. For five of the iCGIs, the gene expression measured by RT-qPCR recapitulated the gene expression patterns detected in the RNA-seq datasets (Figure 5E, *left* and Supplementary Figure S8) which is consistent with expectations. The other five iCGIs did not match the RNA-seq expression patterns using RT-qPCR. These RT-qPCR assays were perfomed in tissues that were as matched as closely as possible to the descriptions in the ENCODE data repository but were not the actual tissues used by the ENCODE consortium. Therefore, there could have been slight differences in the precise age of collected embryos in developmental samples or differences in the dissected sections of tissues, possibly accounting for discordance. In some cases, strain differences may also have led to discrepant findings. To confirm that *Actb* is a reliable control gene for the RT-qPCR assays, ENCODE datasets were interrogated for *Actb* expression in tissues corresponding to the ones used here (Supplementary Figure S9). The ENCODE data revealed that *Actb* expression is similar across brain, thymus, lung, stom-

ach and placenta, confirming that this housekeeping gene is a reliable endogenous control for comparisons involving these tissues (e.g. *Adck2-Ndufb2*, brain E13.5 versus stomach adult). However, lower expression levels observed in heart, testis and cerebellum could contribute to the discrepancy detected between the RNA-seq and RT-qPCR results (e.g. *Bop1*, brain E13.5 versus heart adult).

The DNA methylation status of these candidate loci was assayed by PCR amplification of bisulfite-converted genomic DNA. At one of the concordant iCGI/host gene pairs, namely *Hoxa*, DNA methylation correlated with gene expression (Figure 5E, *right*). For the other four, no correlation between iCGI expression and DNA methylation was detected (Figure 5E, *right*) suggesting that DNA methylation may not the major regulatory factor involved. In order to validate our observations, publicly available genomewide bisulfite sequencing datasets from the ENCODE project were utilised. Although, not all the tissues in the bisulfite sequencing PCR experiments were available, data from brain, lung, heart and stomach were used. When examining the DNA methylation percentage of the iCGIs used for validation across these tissues, it was found that the majority of them are unmethylated (Supplementary Figure S10). Variation of DNA methylation levels was observed between tissues at only one of the iCGIs (*Dhrs3*), suggesting that DNA methylation is not the only mechanism involved in the regulation of expression at iCGIs.

Active intronic iCGIs present promoter-like chromatin

To determine whether these iCGIs are independent transcriptional units or a by-product of host gene transcription, ENCODE H3K36me3 and H3K4me3 ChIP-seq datasets from mouse E14.5 brain (50) were interrogated. When in promoter regions, H3K36me3 negatively affects transcription (60), whereas H3K4me3 is considered a typical promoter mark and it can be found at both transcriptionally active (61) and poised promoters (62). In order to identify appropriate subsets for comparison, hierarchical clustering was applied to host gene expression upstream of the iCGI and the expression of the iCGI itself. This provided a means to select transcriptionally active host genes with very active or less active iCGIs and exclude iCGIs within inactive host genes from further analysis (Figure 6A). Furthermore, before plotting average gene body H3K36me3 and H3K4me3 profiles, only host genes with iCGIs fully contained within introns were selected since H3K36me3 is enriched at expressed exons (12). This approach revealed a sharp decrease in H3K36me3 around highly active iCGIs and a more subtle depletion around iCGIs expressed at low levels (Figure 6B, *upper*). H3K36me3 gradually accumulates downstream of highly expressed iCGIs before decreasing again upstream of host gene transcription termination sites (TTSs), resembling the deposition of H3K36me3 along the body of actively transcribing genes (Figure 6B, *upper left*). H3K4me3 was enriched at all intronic iCGIs with a bias towards highly expressed iCGIs (Figure 6B, *lower*). Taken together, these findings suggest that the iCGIs tested may indeed be discrete tissue-specific promoters and that histone modifications could be important to regulate their expression.

DISCUSSION

Epigenetic modifications are associated with transcription and pre-mRNA processing. In some cases, the deposition of one mark is dependent upon the presence or absence of another. The same marks can differentially influence transcription depending on genomic context. For instance, DNA hypermethylation at CGI-associated promoters negatively impacts transcription (34) but facilitates it when present within gene bodies (32). One post-translational histone modification has been extensively studied in the context of pre-mRNA processing, namely H3K36me3 (14,16,28,29,55). This histone mark is deposited co-transcriptionally along the body of actively transcribing genes by the RNA Pol II-interacting histone methyltransferase SETD2 (52). Intriguingly, human kidney tumours with *SETD2* mutations are characterised by increased intron retention and altered transcription termination site usage (28,29).

We investigated the influence of H3K36me3 on intron retention and IPA in NIH/3T3 cells using siRNAs coupled with transcriptome analysis and locus-specific gene expression assays. We show that depletion of H3K36me3 leads to significant deregulation of 4266 transcripts and 136 AS events. Changes in intron retention are virtually absent between *Setd2* KD and WT samples, indicating that H3K36me3 is not a master regulator of this type of pre-mRNA processing. Locus-specific experiments on the imprinted iCGI/host gene pair *Mcts2*/*H13* are consistent with our genomewide conclusions since we did not detect changes in *H13d* transcript abundance, a proxy for intron retention, upon H3K36me3 depletion. These findings confirm the involvement of H3K36me3 in AS but differ from previous studies in human kidney tumours that associate *SETD2* mutations with increased intron retention. This could be partially explained by the use of polyadenylated RNA rather than total RNA for the preparation of RNA-seq libraries. If H3K36me3 depletion does indeed lead to increased intron retention in the nucleus, there must be efficient downstream checkpoints that prevent the export of those transcripts into the cytoplasm. However, re-analysis of previous studies shows consistent percentages of bases aligning to intronic sequences across all datasets, suggesting that the use of improved bioinformatic tools could also be the basis of these discrepancies.

Since paternal *H13* transcripts are characterised by intron retention and IPA upstream of an actively transcribing iCGI, we interrogated the impact of transcription from iCGIs on host gene pre-mRNA processing more generally. We bioinformatically analyse 30 mouse tissues and developmental stages using RNA-seq datasets from the ENCODE project and identify 4033 iCGI/host gene pairs. For one in five pairs, iCGI activity is tissue- or developmental stage-specific and the level of host gene transcription upstream of the iCGI positively correlates with the level of iCGI transcription. We repeated the same analysis using ENCODE RNA-seq data from 18 human cell lines and find the same results, indicating that these observations are reproducible across two mammalian species. Additionally, we demonstrate that this effect is largely independent of DNA methylation for a small subset of iCGI/host gene pairs. Finally, we

provide evidence that these iCGIs may be discrete tissue-specific promoters since they are enriched for H3K4me3 and depleted of H3K36me3, a chromatin profile typically associated with active gene promoters (63).

Previous studies have shown that iCGIs become methylated and silenced when host gene transcription traverses them and triggers the recruitment of the *de novo* DNA methylation machinery (33). However, the methylation status of these iCGIs is dependent upon their intrinsic ability to initiate transcription, i.e. strong iCGIs can escape host gene transcription-mediated silencing (33). These iCGIs retain an unmethylated state, are enriched in H3K4me3 and lack H3K36me3 (33). Our work consolidates these findings and provides an additional layer of complexity to the cross talk between host gene promoters and iCGIs. We show for the first time that transcription from a large number of iCGIs interferes with host genes transcription in mouse and human, possibly providing a novel mechanism for spatiotemporal diversification of both transcriptome and proteome. These findings raise the question of precisely how iCGIs influence pre-mRNA processing of host gene transcripts. We speculate that iCGI activity may stimulate IPA at the expense of 3′UTR-APA in host genes with multiple polyadenylation sites. Isoforms resulting from IPA are typically as robustly expressed as full-length transcripts and therefore likely represent functional mRNAs rather than transcriptional noise (11). This is particularly relevant since IPA is important for regulating transcript diversity during differentiation and development in both physiological and pathological conditions (11). Using differentiation models, it will be important to examine the extent of the role of iCGIs in tissue- and developmental stage-specific gene expression and to understand the mechanisms involved.

DATA AVAILABILITY

The splice donor site score calculation tool is freely available at http://rulai.cshl.edu/new_alt_exon_db2/HTML/score.html.

Setd2 knockdown RNA-seq data have been deposited under accession number GSE147077.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

REFERENCES

1. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
2. Ayoubi,T.A.Y. and Van De Ven,W.J.M. (1996) Regulation of gene expression by alternative promoters. *FASEB J.*, **10**, 453–460.
3. Beyer,A.L. and Osheim,Y.N. (1988) Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev.*, **2**, 754–765.
4. Baurén,G., Belikov,S. and Wieslander,L. (1998) Transcriptional termination in the Balbiani ring 1 gene is closely coupled to 3′-end formation and excision of the 3′-terminal intron. *Genes Dev.*, **12**, 2759–2769.
5. Tian,B. and Manley,J.L. (2016) Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.*, **18**, 18–30.
6. Kalsotra,A. and Cooper,T.A. (2011) Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.*, **12**, 715–729.
7. Derti,A., Garrett-Engele,P., MacIsaac,K.D., Stevens,R.C., Sriram,S., Chen,R., Rohl,C.A., Johnson,J.M. and Babak,T. (2012) A quantitative atlas of polyadenylation in five mammals. *Genome Res.*, **22**, 1173–1183.
8. Hoque,M., Ji,Z., Zheng,D., Luo,W., Li,W., You,B., Park,J.Y., Yehia,G. and Tian,B. (2013) Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nat. Methods*, **10**, 133–139.
9. Keren,H., Lev-Maor,G. and Ast,G. (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.*, **11**, 345–355.
10. Gruber,A.J. and Zavolan,M. (2019) Alternative cleavage and polyadenylation in health and disease. *Nat. Rev. Genet.*, **20**, 599–614.
11. Singh,I., Lee,S.H., Sperling,A.S., Samur,M.K., Tai,Y.T., Fulciniti,M., Munshi,N.C., Mayr,C. and Leslie,C.S. (2018) Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nat. Commun.*, **9**, 1716.
12. Kolasinska-Zwierz,P., Down,T., Latorre,I., Liu,T., Liu,X.S. and Ahringer,J. (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat. Genet.*, **41**, 376–381.
13. Schwartz,S., Meshorer,E. and Ast,G. (2009) Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.*, **16**, 990–995.
14. Luco,R.F., Pan,Q., Tominaga,K., Blencowe,B.J., Pereira-Smith,O.M. and Misteli,T. (2010) Regulation of alternative splicing by histone modifications. *Science*, **327**, 996–1000.
15. Zhou,H.-L., Hinman,M.N., Barron,V.A., Geng,C., Zhou,G., Luo,G., Siegel,R.E. and Lou,H. (2011) Hu proteins regulate alternative splicing by inducing localized histone hyperacetylation in an RNA-dependent manner. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, E627–E635.
16. Pradeepa,M.M., Sutherland,H.G., Ule,J., Grimes,G.R. and Bickmore,W.A. (2012) Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. *PLoS Genet.*, **8**, e1002717.
17. Ameyar-Zazoua,M., Rachez,C., Souidi,M., Robin,P., Fritsch,L., Young,R., Morozova,N., Fenouil,R., Descostes,N., Andrau,J.C. *et al.* (2012) Argonaute proteins couple chromatin silencing to alternative splicing. *Nat. Struct. Mol. Biol.*, **19**, 998–1005.
18. Maunakea,A.K., Chepelev,I., Cui,K. and Zhao,K. (2013) Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res.*, **23**, 1256–1269.
19. Liu,H., Jin,T., Guan,J. and Zhou,S. (2014) Histone modifications involved in cassette exon inclusions: a quantitative and interpretable analysis. *BMC Genomics*, **15**, 1148.
20. Zhou,H.L., Luo,G., Wise,J.A. and Lou,H. (2014) Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. *Nucleic. Acids. Res.*, **42**, 701–713.
21. Spies,N., Nielsen,C.B., Padgett,R.A. and Burge,C.B. (2009) Biased chromatin signatures around polyadenylation sites and exons. *Mol. Cell*, **36**, 245–254.
22. Wood,A.J., Schulz,R., Woodfine,K., Koltowska,K., Beechey,C.V., Peters,J., Bourc'his,D. and Oakey,R.J. (2008) Regulation of alternative polyadenylation by genomic imprinting. *Genes Dev.*, **22**, 1141–1146.
23. Cowley,M., Wood,A.J., Böhm,S., Schulz,R. and Oakey,R.J. (2012) Epigenetic control of alternative mRNA processing at the imprinted Herc3/Nap1l5 locus. *Nucleic Acids Res.*, **40**, 8917–8926.
24. Morison,I.M., Ramsay,J.P. and Spencer,H.G. (2005) A census of mammalian imprinting. *Trends Genet.*, **21**, 457–465.
25. Ferguson-Smith,A.C. (2011) Genomic imprinting: the emergence of an epigenetic paradigm. *Nat. Rev. Genet.*, **12**, 565–575.
26. Mccole,R.B., Loughran,N.B., Chahal,M., Fernandes,L.P., Roberts,R.G., Fraternali,F., O'Connell,M.J. and Oakey,R.J. (2011) A case-by-case evolutionary analysis of four imprinted retrogenes. *Evolution (N.Y.)*, **65**, 1413–1427.
27. Wood,A.J., Roberts,R.G., Monk,D., Moore,G.E., Schulz,R. and Oakey,R.J. (2007) A screen for retrotransposed imprinted genes reveals an association between x chromosome homology and maternal germ-line methylation. *PLoS Genet.*, **3**, 0192–0203.
28. Simon,J.M., Hacker,K.E., Singh,D., Brannon,A.R., Parker,J.S., Weiser,M., Ho,T.H., Kuan,P.F., Jonasch,E., Furey,T.S. *et al.* (2014) Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects. *Genome Res.*, **24**, 241–250.
29. Ho,T.H., Park,I.Y., Zhao,H., Tong,P., Champion,M.D., Yan,H., Monzon,F.A., Hoang,A., Tamboli,P., Parker,A.S. *et al.* (2016) High-resolution profiling of histone h3 lysine 36 trimethylation in metastatic renal cell carcinoma. *Oncogene*, **35**, 1565–1574.
30. Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.
31. Weinberg,D.N., Papillon-Cavanagh,S., Chen,H., Yue,Y., Chen,X., Rajagopalan,K.N., Horth,C., McGuire,J.T., Xu,X., Nikbakht,H. *et al.* (2019) The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape. *Nature*, **573**, 281–286.
32. Neri,F., Rapelli,S., Krepelova,A., Incarnato,D., Parlato,C., Basile,G., Maldotti,M., Anselmi,F. and Oliviero,S. (2017) Intragenic DNA methylation prevents spurious transcription initiation. *Nature*, **543**, 72–77.
33. Jeziorska,D.M., Murray,R.J.S., De Gobbi,M., Gaentzsch,R., Garrick,D., Ayyub,H., Chen,T., Li,E., Telenius,J., Lynch,M. *et al.* (2017) DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E7526–E7535.
34. Auclair,G., Guibert,S., Bender,A. and Weber,M. (2014) Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. *Genome Biol.*, **15**, 545.
35. Maunakea,A.K., Nagarajan,R.P., Bilenky,M., Ballinger,T.J., Dsouza,C., Fouse,S.D., Johnson,B.E., Hong,C., Nielsen,C., Zhao,Y. *et al.* (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, **466**, 253–257.
36. Illingworth,R.S., Gruenewald-Schneider,U., Webb,S., Kerr,A.R.W., James,K.D., Turner,D.J., Smith,C., Harrison,D.J., Andrews,R. and Bird,A.P. (2010) Orphan CpG Islands Identify numerous conserved promoters in the mammalian genome. *PLoS Genet.*, **6**, 1001134.

37. Illingworth,R., Kerr,A., DeSousa,D., Jørgensen,H., Ellis,P., Stalker,J., Jackson,D., Clee,C., Plumb,R., Rogers,J. *et al.* (2008) A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.*, **6**, 0037–0051.

38. Callen,B.P., Shearwin,K.E. and Egan,J.B. (2004) Transcriptional interference between convergent promoters caused by elongation over the promoter. *Mol. Cell*, **14**, 647–656.

39. Martens,J.A., Laprade,L. and Winston,F. (2004) Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. *Nature*, **429**, 571–574.

40. Wang,P., Yang,J., Ishihama,A. and Pittard,A.J. (1998) Demonstration that the TyrR protein and RNA polymerase complex formed at the divergent P3 promoter inhibits binding of RNA polymerase to the major promoter, P1, of the aroP gene of Escherichia coli. *J. Bacteriol.*, **180**, 5466–5472.

41. Boussadia,O., Amiot,F., Cases,S., Triqueneaux,G., Jacquemin-Sablon,H. and Dautry,F. (1997) Transcription of unr (upstream of N-ras) down-modulates N-ras expression in vivo. *FEBS Lett.*, **420**, 20–24.

42. Livak,K.J. and Schmittgen,T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. *Methods*, **25**, 402–408.

43. Andrews,S., Krueger,F., Seconds-Pichon,A., Biggins,F. and Wingett,S. (2015) FastQC. A quality control tool for high throughput sequence data. Babraham Bioinformatics. *Babraham Inst.*, **1**, 1.

44. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.

45. Pimentel,H., Bray,N.L., Puente,S., Melsted,P. and Pachter,L. (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods*, **14**, 687–690.

46. Mi,H., Muruganujan,A., Huang,X., Ebert,D., Mills,C., Guo,X. and Thomas,P.D. (2019) Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat. Protoc.*, **14**, 703–721.

47. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.

48. Vaquero-Garcia,J., Barrera,A., Gazzara,M.R., Gonzalez-Vallinas,J., Lahens,N.F., Hogenesch,J.B., Lynch,K.W. and Barash,Y. (2016) A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife*, **5**, e11752.

49. Rohde,C., Zhang,Y., Reinhardt,R. and Jeltsch,A. (2010) BISMA - fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences. *BMC Bioinformatics*, **11**, 230.

50. Feingold,E.A., Good,P.J., Guyer,M.S., Kamholz,S., Liefer,L., Wetterstrand,K., Collins,F.S., Gingeras,T.R., Kampa,D., Sekinger,E.A. *et al.* (2004) The ENCODE (ENCyclopedia of DNA Elements) Project. *Science*, **306**, 636–640.

51. Gu,Z., Eils,R., Schlesner,M. and Ishaque,N. (2018) EnrichedHeatmap: an R/Bioconductor package for comprehensive visualization of genomic signal associations. *BMC Genomics*, **19**, 234.

52. Edmunds,J.W., Mahadevan,L.C. and Clayton,A.L. (2008) Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. *EMBO J.*, **27**, 406–420.

53. Galante,P.A.F., Sakabe,N.J., Kirschbaum-Slager,N. and De Souza,S.J. (2004) Detection and evaluation of intron retention events in the human transcriptome. *RNA*, **10**, 757–765.

54. Zhang,Y., Xie,S., Zhou,Y., Xie,Y., Liu,P., Sun,M., Xiao,H., Jin,Y., Sun,X., Chen,Z. *et al.* (2014) H3K36 histone methyltransferase Setd2 is required for murine embryonic stem cell differentiation toward endoderm. *Cell Rep.*, **8**, 1989–2002.

55. Zhu,K., Lei,P.J., Ju,L.G., Wang,X., Huang,K., Yang,B., Shao,C., Zhu,Y., Wei,G., Fu,X.D. *et al.* (2017) SPOP-containing complex regulates SETD2 stability and H3K36me3-coupled alternative splicing. *Nucleic Acids Res.*, **45**, 92–105.

56. Park,I.Y., Powell,R.T., Tripathi,D.N., Dere,R., Ho,T.H., Blasius,T.L., Chiang,Y.C., Davis,I.J., Fahey,C.C., Hacker,K.E. *et al.* (2016) Dual chromatin and cytoskeletal remodeling by SETD2. *Cell*, **166**, 950–962.

57. Li,F., Mao,G., Tong,D., Huang,J., Gu,L., Yang,W. and Li,G.M. (2013) The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSα. *Cell*, **153**, 590–600.

58. Yildirim,E., Sadreyev,R.I., Pinter,S.F. and Lee,J.T. (2012) X-chromosome hyperactivation in mammals via nonlinear relationships between chromatin states and transcription. *Nat. Struct. Mol. Biol.*, **19**, 56–62.

59. Huang,Z., Han,Z., Zhang,F., He,H., Yu,S. and Wu,Q. (2014) Spatiotemporal expression of retrogene-host pair Mcts2/H13 in mouse embryo, and Mcts2 has no influence on H13 transcription pattern in NIH/3T3 cells. *Acta Histochem.*, **116**, 312–318.

60. Strahl,B.D., Grant,P.A., Briggs,S.D., Sun,Z.-W., Bone,J.R., Caldwell,J.A., Mollah,S., Cook,R.G., Shabanowitz,J., Hunt,D.F. *et al.* (2002) Set2 is a nucleosomal histone H3-selective methyltransferase that mediates transcriptional repression. *Mol. Cell. Biol.*, **22**, 1298–1306.

61. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigó,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

62. Bernstein,B.E., Mikkelsen,T.S., Xie,X., Kamal,M., Huebert,D.J., Cuff,J., Fry,B., Meissner,A., Wernig,M., Plath,K. *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.

63. Barth,T.K. and Imhof,A. (2010) Fast signals and slow marks: the dynamics of histone modifications. *Trends Biochem. Sci.*, **35**, 618–626.