Human promoter CpG islands contain primate-specific repeats, cluster and resist reprogramming

Kommu Mohan
Birla Institute of Technology and Science
Anuhya Anne
Birla Institute of Technology and Science
Lov Kumar
National Institute of Technology
J Richard Chaillet
Chaillet@pitt.edu

University of Pittsburgh School of Medicine

Article

Keywords: DNA methylation, epigenetic reprogramming, stem cells, CpG island, human evolution, genome organization

Posted Date: March 11th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3973757/v1

Additional Declarations: No competing interests reported.

Abstract

Imperfect tandem repeats (TRs) of ≥400nt are associated with 365 human 5'-genic CpG islands (TR-CGIs). Most are clustered at chromosome ends, with a high density across chromosome 19. These genes are enriched in neurodevelopmental/behavioral disorders and show interindividual variation in methylation levels. A subset of TR-CGIs is highly methylated and remains so during reprogramming to primed iPSCs, but become unmethylated in naïve PSCs, as do imprinting control regions (ICRs). Transcript levels correlate with methylation for some TR-CGI genes. TR-CGIs occur as orthologs in primates, but the corresponding mouse promoter-CGIs are without TRs and unmethylated. Thus, non-imprinted TR-CGIs accompanied primate evolution, with unique ability to acquire methylation during embryonic development and resist reprogramming to a pluripotent stem cell state.

Introduction

About 50% of mammalian gene regulation at the transcription initiation level takes place via CpG-rich promoter regions known as CpG islands (CGIs) that are mostly unmethylated in preimplantation, embryonic development and post-natal life. The others acquire methylation during embryonic development and tend to be transcriptionally silent. A notable exception to the general rule is the parental origin-specific methylation of approximately 20 CG-rich autosomal imprinting control regions (ICRs), ~1-5 Kb in length. Methylation of these sequences is established during gametogenesis reflecting the parental sex by *de novo* methyltransferases and maintained post-fertilization by the DNA methyltransferase DNMT1 [1]. As a result, in eutherian mammals, one parental ICR allele is methylated and the counterpart unmethylated. Although there is uncertainty on the sequence requirements for establishing methylation, ICRs contain tandem repeat (TR) sequences, some of which have been experimentally shown to be essential imprinting elements [2-3]. The absence of sequence similarities between ICRs strongly suggests that repetitiveness majorly contributes to the imprinted state [4].

A second feature of monoallelic ICR methylation is its longevity. Transcription of imprinted genes in a dose-dependent manner is required for normal development in mammals, yet generally not required after birth. For example, many imprinted genes are only expressed in the placenta but monoallelic ICR methylation persists for the lifetime of the individual [5]. From this, we reasoned that CGI methylation akin to ICR methylation may occur on non-imprinted sequences, initiated as *de novo* methylation at some developmental stage post-fertilization, and thereafter maintained by DNMT1. Such sequences might differ from invariant ICR methylation because post-zygotic *de novo* methylation may be cell type-specific or differ between individuals.

Results

To test these hypotheses, we first generated and analyzed dotplots of human promoter-proximal CGIs (see Materials and Methods) and identified 365 with tandem repeats measuring ≥400bp each, containing more than two copies of a perfect or imperfect repeating unit. These repeats occur either within or

adjacent to the CGIs. Of these, 349 were autosomal that were investigated in detail in the context of DNA methylation, epigenetic reprogramming, individual-to-individual variation, and its influence on the transcript levels (Figure 1A). Dotplots of two TR-CGIs and one ICR are shown in Figure 1B. As is the case for repeats in ICRs, DNA sequence of each TR is unique. Thus, if TRs were to provide similar properties to all TR-CGIs, it must be through their repetitiveness rather than due to a sequence.

Approximately 60% of TR-CGIs are in 33 autosomal clusters, each with more than two genes, ranging from ~2.5-23 Mb in length with ~0.5-6.0 TR-CGIs/Mb (Figure 1C and Table S1). Many autosomes contain more than one cluster, predominantly at the ends of chromosomes except chromosome 19 with 51 TR-CGIs (14% of total) intermingled with *ZFP* and *ZNF* genes [6]. Clustering to chromosome ends is not seen for mouse syntenic regions (Table S1). Importantly, TR-CGI clusters do not overlap clusters of imprinted genes.

Bioinformatic analyses of the autosomal TR-CGI genes identified 25 biological processes such as chromosome condensation, chromosome separation, nuclear envelope organization, and regulation of glutamate receptor (Figure 2A). Protein-protein interaction analysis yielded a single cluster involving proteins involved in ubiquitinylation, and nuclear receptor corepressor (Figure 2B). GTex analysis confirmed significant association of TR-CGI gene expression in the amygdala (Fig. 2C and Table S2). Among the 25 diseases associated with the TR-CGI genes, 11 are neurological, neurodevelopmental, or behavioral disorders (Figure 2A) wherein there is a significant overrepresentation for epilepsy, but underrepresentation in case of schizophrenia (Figure 2D and Table S2).

Given the similarities in the structural features of ICRs and non-imprinted TR-CGIs we studied TR-CGI methylation in different tissue types, preimplantation development (blastocysts) and induced pluripotent stem cells (iPSCs). In 68 normal prefrontal cortex samples [7], 48 of TR-CGIs (~15%) showed $\geq$ 40% methylation of which ~9% are highly methylated (~75-95% methylation). In contrast, methylation levels in ICRs were ~35-75% whereas a vast majority of other promoter CGIs (~86%) were generally unmethylated (Figure 3A and Table S3). When the methylation levels of the 48 TR-CGIs were compared among the 68 prefrontal cortex samples, sequences showed lesser or greater methylation (~40%) than the average (~79%) in just six individuals (arrows in Figure 3B).

To test whether the observed differences in the three categories of sequences are also displayed in other tissues within an individual, we analyzed the data on embryonic tissues as well as primed iPSCs derived from them [8]. ICR methylation was invariant across tissues, whereas a large fraction of non-TRs & non-ICR CGIs are differentially methylated regions (DMRs). Between these two expected extremes of CGI methylation features are the 48 TR-CGIs of which a small fraction are tissue DMRs (Figure 3C and Table S4). Interestingly, the five TR-CGIs noted in Figure 3B are among these DMRs. With respect to cellular reprogramming of tissues, there is no significant difference in the proportion of TR-CGI and non-TR & non-ICR CGIs that are DMRs upon reprogramming; only four out of 48 TR-CGIs with $\geq$ 40% methylation in the tissues showed significant changes in the levels of methylation after reprogramming. However, in the case of ICRs, a surprisingly large proportion are DMRs in the derived iPSCs. This represents some degree

of instability of the allelic methylated or allelic unmethylated states. These differences between the three CGI types are apparent in the principal component analyses (Figure 3D). Further analysis of the methylated subset of TR-CGIs showed no alterations in methylation levels in motor neurons differentiated from the piPSCs [9]. Taken together, these data lead us to conclude that methylation of TR-CGIs is resistant to reprogramming into piPSCs, and, in turn, likely resistant to change with differentiation. In this feature, these TR-CGIs are like the ICRs that are also resistant to reprogramming and differentiation events.

Previous studies established that ICR methylation is not affected in primed iPSCs and primed ESCs but affected in their naïve counterparts [10-11]. To test whether the methylated TR-CGIs show a similar pattern, methylation data from terminally differentiated cell lines, their primed iPSC (piPSC) and naïve iPSC (niPSC) derivatives [10], as well as primed and naïve embryonic stem cells (ESCs) [11], were used. Consistent with the previous observation, resistance of TR-CGI and ICR methylation to reprogramming was again observed in primed iPSCs (Figures 3E, 3F; Table S5). However, methylation observed in the differentiated cells is not maintained upon reprogramming to naïve iPSCs (Figures 3E, 3F; Table S5), nor is TR-CGI methylation maintained upon conversion of primed ESCs into naïve ESCs (Table S6). Thus, both TR-CGI and ICR methylation levels in somatic tissues and differentiated cell lines are much more similar to the levels of piPSCs than niPSCs.

As expected, ICRs show intermediate levels of methylation in blastocysts (~50%; Figure 3F and Table S7), but only seven of the 48 TR-CGIs show high methylation (Figure 3E and Table S7); the remaining are unmethylated. Because the seven TR-CGIs methylated in blastocysts are not ICRs, their methylation likely initiates in early preimplantation. These data suggest peri- or post-implantation origins of a majority of TR-CGI methylation.

There is no evidence of sequence variation among humans based on the presence or absence or overall length of the TRs in the TR-CGIs. This precluded us from directly observing a TR requirement for TR-CGI methylation. However, all CGIs in mouse orthologs do not have TRs. Therefore, we indirectly addressed the association of TRs with methylation by comparing methylation levels in TR-CGIs and corresponding mouse orthologs. For the methylated TR-CGIs in *Rgpd1*, *Fam178B*, *Pcgf3*, *Dnaaf5*, *Shtn1*, *Grtp1*, *Stub1*, *Kcng2*, *Shc2*, and *Mob3A/ Izumo4,* the corresponding CGIs in both mouse embryonic fibroblasts and mouse iPSCs showed an overall unmethylated state. Only the *C5orf47* CGI showed ~50% methylation in mouse embryonic fibroblasts but was less than 10% methylated in mouse iPSCs and ESCs (Figure 3G and Table S5).

Since methylation accompanied acquisition of TRs in these orthologous regions after human – rodent divergence, we tested whether the methylated states of the TR-CGIs have a relationship with their transcript levels. We used cortex methylation data among normal individuals to test as well as analyze TR-CGIs that show interindividual variation in the levels of methylation. Fourteen TR-CGIs show $\geq$ 20% difference in methylation levels in a subset of individuals. This nature of variation and the large differences in methylation between the piPSCs and niPSCs prompted us to compare the transcript levels

with methylation levels of the TR-CGIs for these genes. Two examples for TR-CGIs, their methylation and expression levels in oral tissues are shown in Figure 4A. Similar analyses using primed and naïve ESCs are shown in Figure 4B (Table S7). Four genes show an inverse relationship between methylation and expression levels whereas one had a direct relationship. Interestingly, three of the four genes (*RGPD1, RGPD3*, and *RGPD4*) are among the eight-member *RGPD* gene family that rapidly evolved in primates via duplications of the highly conserved RAN binding protein 2 gene (*RanBP2/NUP358)* [12-13]. No such relationships could be established for the remaining TR-CGIs differentially methylated in piPSCs and niPSCs (Supplemental Data). Notably, correlations between methylation and transcript levels in iPSCs could be established for only a minority of ICRs, despite meaningful *in vivo* correlations of transcript levels to methylation for all ICRs.

To gain more insights into the evolutionary origins of the TR-CGIs, we compared orthologs in different eutherians. Orthologous CGIs contained TRs only in primates. The latest evolutionary appearance varies, ranging from the presence of TRs specifically in humans, some in all apes and monkeys, whereas others only in primates (Figure 5; Supplement for Figure 1C). Almost all genomic TR-CGI clusters acquired TRs at different times during evolution. A similar pattern of evolution was also reported for orthologs of the zinc finger genes [6].

Discussion

We postulate that TR-CGI methylation, acquired after fertilization is stably maintained for the entire lifespan of the individual. Taken together, similarities in ICR and TR-CGI structures, individual-to-individual variability in TR-CGI methylation, expression in brain regions and presence of TR-CGIs in primates but not in other mammals suggest that primates usurped features of the invariant mammalian genomic (gametic) imprinting process for post-zygotic epigenetic modification of genes and for guiding embryonic development. Interestingly, these permanent TR-CGI methylation events would correspond to the hypothetical epigenetic bifurcation events creating canals in the epigenetic landscape of development proposed by Waddington [14].

There are many TR-CGIs that are unmethylated in the tissues and cell lines that were examined. The shared clustering and primate-specificity of the 365 TR-CGIs suggest a shared propensity to acquire methylation during embryogenesis. Methylated versions would have eluded detection if their *de novo* methylation occurred later in development and in specific cell types.

A fundamental feature of stable ICRs' methylated and unmethylated states is their stable inheritance in a parent- and gamete-specific manner in the zygote and the adult. Neither the methylated state of a parental ICR nor the unmethylated state of the homologous allele is affected throughout life in the somatic tissues. Any loss of a parental specific methylation mark will result in an unmethylated state of the methylated parental allele. For example, removal of DNMT1 from a single S-phase of mouse preimplantation results in a permanent loss of ICR methylation [15]. A few TR-CGIs are methylated in iPSCs and in preimplantation whereas many more are unmethylated in the blastocysts but methylated in

the iPSCs. This difference in methylated states needs to be examined in the context of *in vivo* and *in vitro* origins of pluripotency. In addition, experiments that remove methylation in TR-CGIs will help test the stability of the induced unmethylated state.

Materials and Methods

Identification of TR-CGIs

Promoter-proximal CGIs were identified from the annotated CGI track in the UCSC browser of hg19 and hg38 assemblies using visual inspection of dotplots. In case of an annotation in just one track, or different CGI lengths in the two tracks, the sole or longer CGI was used to deduce the coordinates of the other track 's CGI via sequence comparisons. The BLASTN (default) variables for generating dotplots that are then scored for presence of tandem repeats in hg19 and hg38 assemblies: somewhat similar sequences; expect threshold 0.05; word size 11; match/mismatch score 2,-3; existence: 5 extension: 2. TRs within or immediately adjacent to promoter-proximal CGIs ranged in size from <100bp to ≥400bp. We chose to limit our studies to the promoter-proximal TRs ≥400bp (defined as TR-CGIs) because of the length similarities to ICRs (example shown in Figure 1B).

Determining methylation levels in TR-CGI, ICR and non-TR & non-ICR sequences

CGI methylation values as percents or fractions were calculated from human 450K and EPIC Illumia microarrays or reduced representation bisulfite sequencing (RRBS) datasets using annotated CGI coordinates (see above for details concerning TR-CGI coordinates). For comparisons, coordinates for human imprinting control regions (ICRs; ref 11) and for the remaining (non-TR-CGI and non-ICR - associated CGIs), coordinates were derived from the manifest files of the 450K and EPIC array manifest files.

Absolute values of calculated methylation levels were used in graphs or heatmaps involving methylation levels. Cut-off values were used to define "unmethylated", intermediate" and "methylated". Minimum values were set to define significant methylation changes with reprogramming.

Generation of transcriptome and DNA methylome analyses

This work was approved by the Institutional Human Ethics Committee of BITS Pilani Hyderabad Campus. After obtaining informed consent, DNA from oral tissues of six individuals were used to generate methylation data using Infinium Human EPIC arrays. RNAs isolated from these samples were used to generate transcriptome data as previously described [16]. The method used for identification of differentially methylated regions (DMRs) was previously described [17]. Briefly, three consecutive CpG sites showing an increased methylation of ≥ 20% were taken as representing a hypermethylated DMR whereas a decreased methylation of ≤ 20% were taken as representing a hypermethylated DMR. For methylation and transcriptome comparisons, FPKM values of the genes of interest from the transcriptome data were used.

Evolutionary appearance of TRs in CGI

Promoter-proximal CGI sequences from non-human primate and rodent species, corresponding to human TR-CGIs, were studied to determine the latest evolutionary appearance of each human TR-CGI. Given the limited number of available annotated vertebrate genome sequences, we approximated latest evolutionary appearance to all primates, new-world monkeys (NW), old-world monkeys (OW), apes or humans.

Sources of genome-wide data

Datasets used are listed in Table S8.

Declarations

Acknowledgments

Funding

Work in KNM lab was supported by grants from BITS Pilani and Centre for Human Disease

Research. AA was supported by a fellowship from BITS Pilani Hyderabad Campus.

Author contributions

Conceptualization: KNM, JRC

Design of the work: KNM, LK, JRC

Acquisition of data: KNM, AA, LK, JRC

Analysis of data: KNM, LK, JRC

Interpretation of data: KNM, LK, JRC

Writing – original draft: KNM, JRC

Additional Information

Authors declare that they have no competing interests.

Data and materials availability: All data are available in the main text or the supplementary

materials.

References

1. Eggermann, T. *et al* .Imprinting disorders. *Nat Rev Dis Primers.* **9(1)**, 33 (2023).
2. Reinhart, B., Eljanne, M. & Chaillet, J.R. Shared role for differentially methylated domains of imprinted genes. *Mol Cell Biol*. **22(7)**, 2089-98 (2002).
3. Reinhart, B., Paoloni-Giacobino, A. & Chaillet, J.R. Specific differentially methylated domain sequences direct the maintenance of methylation at imprinted genes. *Mol Cell Biol*. **26(22)**, 8347-56 (2006).
4. Paoloni-Giacobino, A., D'Aiuto, L., Cirio, M.C., Reinhart, B. & Chaillet, J.R. Conserved features of imprinted differentially methylated domains. *Gene.* **399(1)**, 33-45 (2007).
5. Kobayashi, E.H. et al. Genomic imprinting in human placentation. *Reprod Med Biol*. **21(1)**, 12490; 10.1002/rmb2.12490 (2022)
6. Imbeault, M., Helleboid, P-Y., & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature.* **543(7646)**, 550-554 (2017).
7. Smith, R.G. *et al*. Elevated DNA methylation across a 48-kb region spanning the HOXA gene cluster is associated with Alzheimer's disease neuropathology. *Alzheimers Dement.* **14(12)**, 1580-1588 (2018).
8. Roost, M.S*. et al*. DNA methylation and transcriptional trajectories during human development and reprogramming of isogenic pluripotent stem cells. *Nat Commun.* **8(1)**, 908 (2017).
9. Schulze, M. *et al*. Sporadic Parkinson's disease derived neuronal cells show disease-specific mRNA and small RNA signatures with abundant deregulation of piRNAs. *Acta Neuropathol Commun.* **6(1)**, 58 (2018).
10. Giulitti, S. *et al.* Direct generation of human naïve induced pluripotent stem cells from somatic cells in microfluidics. *Nat Cell Biol*. **21(2)**, 275-286 (2019).
11. Pastor, W.A. *et al*. Naive Human Pluripotent Cells Feature a Methylation Landscape Devoid of Blastocyst or Germline Memory*. Cell Stem Cell.* **18(3)**, 323-329 (2016).
12. Bekpen, C. & Tautz, D. Human core duplicon gene families: game changers or game players? *Brief Funct Genomics.* **18(6)**, 402-411 (2019).
13. Ciccarelli, F.D. *et al*. Complex genomic rearrangements lead to novel primate gene function. *Genome Res*. **15(3)**, 343-51 (2005).
14. Waddington, C.H. *The Strategy of the Genes:a Discussion of Some Aspects of Theoretical Biology* (Allen & Unwin, 1957).
15. Howell, C.Y*. et al*. Genomic imprinting disrupted by a maternal effect mutation in the Dnmt1 gene. *Cell.* **104(6)**, 829-38 (2001).
16. Saxena, S. *et al*. Dysregulation of schizophrenia-associated genes and genome-wide hypomethylation in neurons overexpressing DNMT1.*Epigenomics*. **13(19)**, 1539-1555 (2021).
17. Anne, A., Saxena, S. & Mohan, K.N. Genome-wide methylation analysis of post-mortem cerebellum samples supports the role of peroxisomes in autism spectrum disorder. *Epigenomics*. **14(17)**, 1015-1027 (2022).

18. Smith, Z.D. *et al*. DNA methylation dynamics of the human preimplantation embryo. *Nature*. **511**, 611-615 (2014).
19. Pasque, V. *et al.* X chromosome reactivation dynamics reveal stages of reprogramming to pluripotency. *Cell*. **159(7)**, 1681-97 (2014).

Figures

Figure 1

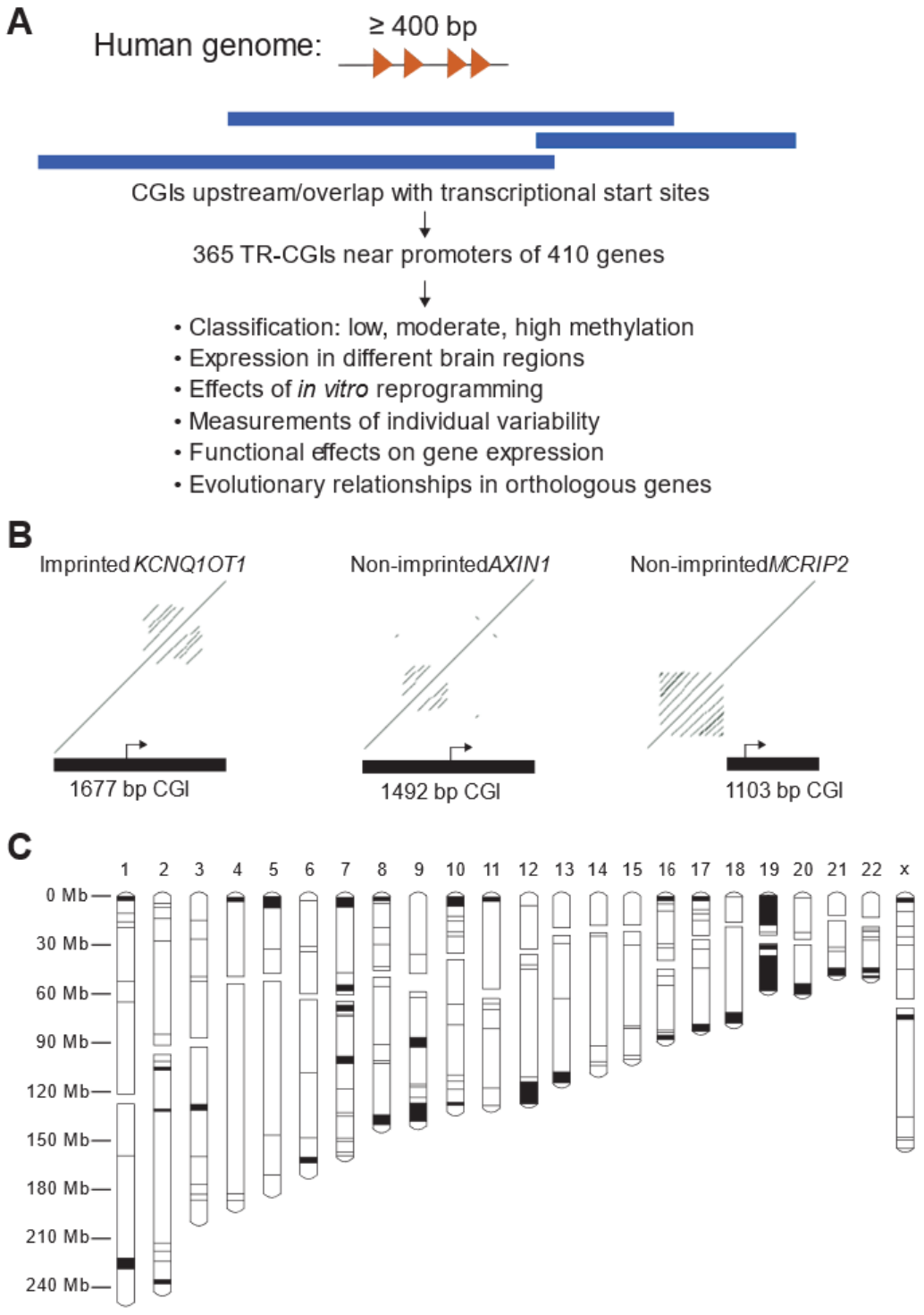


Figure 1

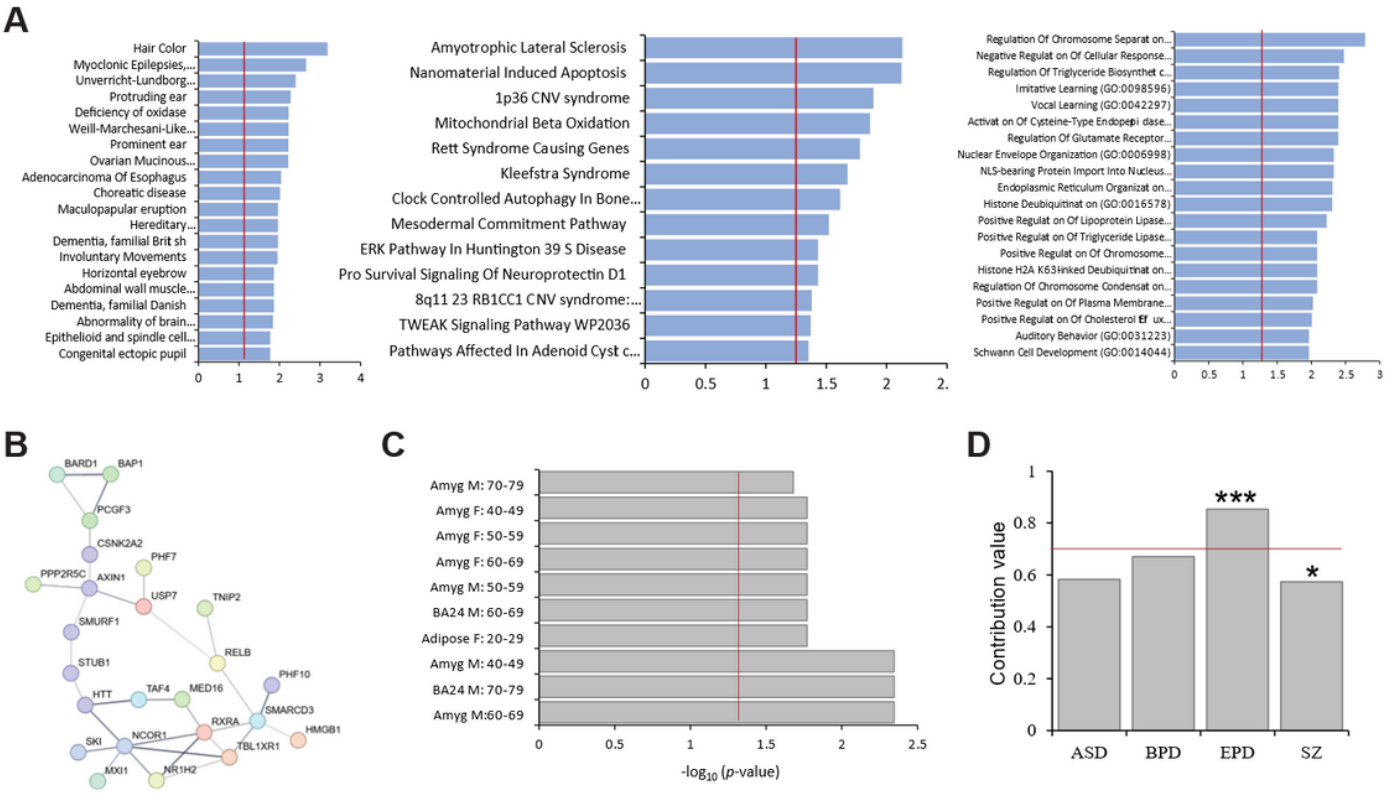


Figure 2

Bioinformatic analyses of TR-CGI genes (**A**) DisGenet Analysis showing top 20 disorders associated with the TR-CGI genes, Wikipathway analysis, and Biological Processes. Dashed red line represents a pvalue of 0.05. (**B**) Protein-protein interaction analysis of genes with TR-CGI promoters. (**C**) Proportions of genes associated with autism spectrum (ASD), bipolar (BPD), epilepsy (EPD) and schizophrenia (SZ). (**D**) GTEx-analysis of genes with TR-CGIs in promoters.

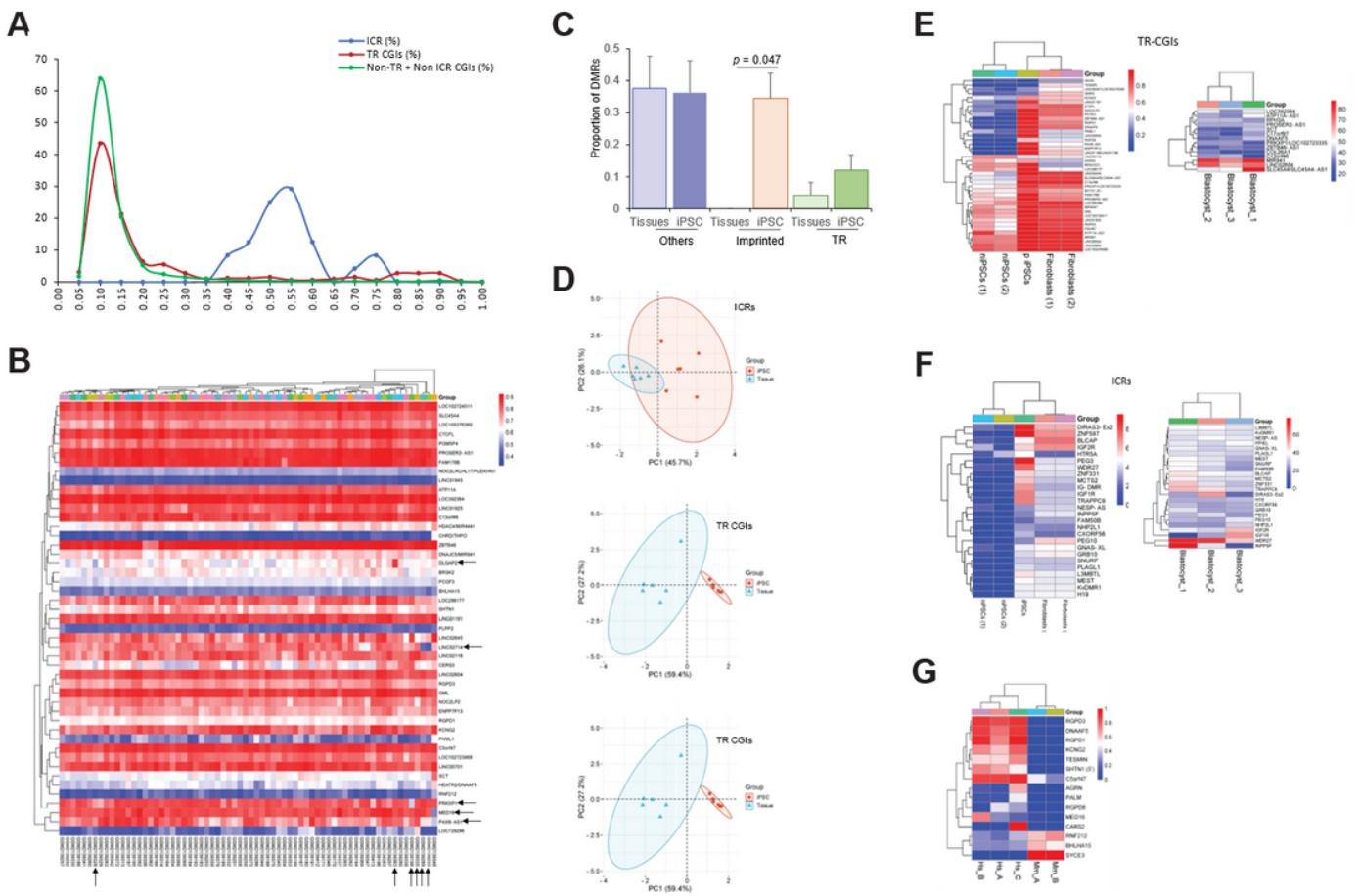


Figure 3

Developmental features of TR-CGI methylation (**A**) Methylation levels (X-axis) of three CGI categories in human cerebellum: imprint control regions (ICR), TR-CGIs, and Non-TR-CGI & non-ICRs (Non-TR & non-ICR). Y-axis: percentages of CGIs at different methylation levels. 0.0 and 1.00 correspond to 0% and 100% methylation, respectively. Arrows and horizontal line indicate significant differences in the proportion of TR-CGIs and Non-TR + non-ICRs with the indicated methylation values. (**B**) Heatmap of methylation levels of 47 TR-CGIs in 68 frontal cortex specimens [7]. (**C**) Proportion of the three categories of CGIs that are differentially methylated regions (DMRs) in embryonic tissues and iPSCs derived from them. (**D**) PCA of methylation levels observed in imprinted ICRs and TR-CGIs in tissues and iPSCs (GSE). (**E-F**) Heatmap of methylation levels of (**E**) TR-CGIs and (**F**) ICRs between naïve pluripotent cells, primed pluripotent cells, parental fibroblasts (left panels, ref 10), and three pools of blastocysts (right panels, ref 18). (**G**) Comparisons of human TR-CGI methylation levels in fibroblasts [10] with those in mouse CGI orthologs in fibroblasts [19].

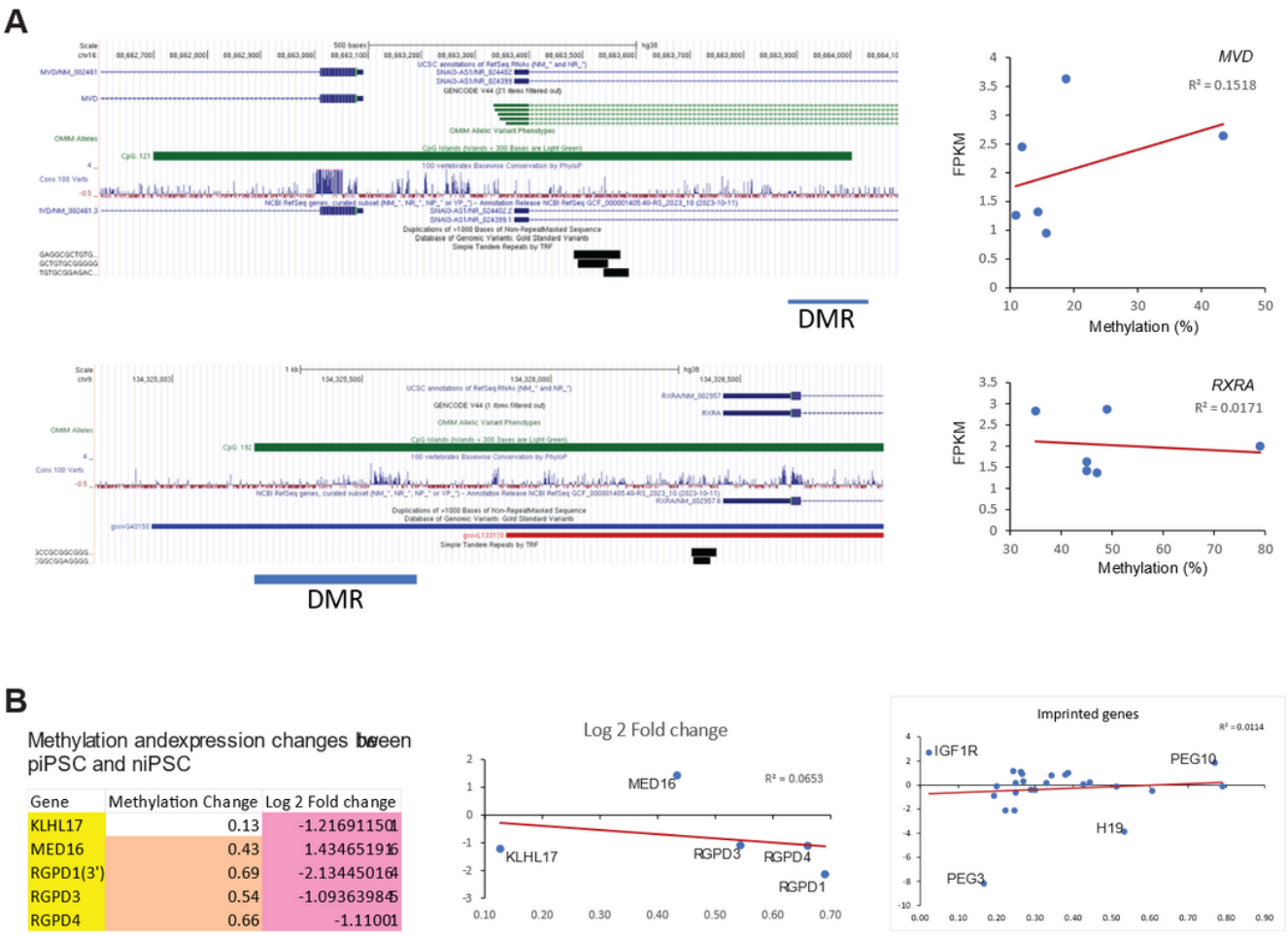
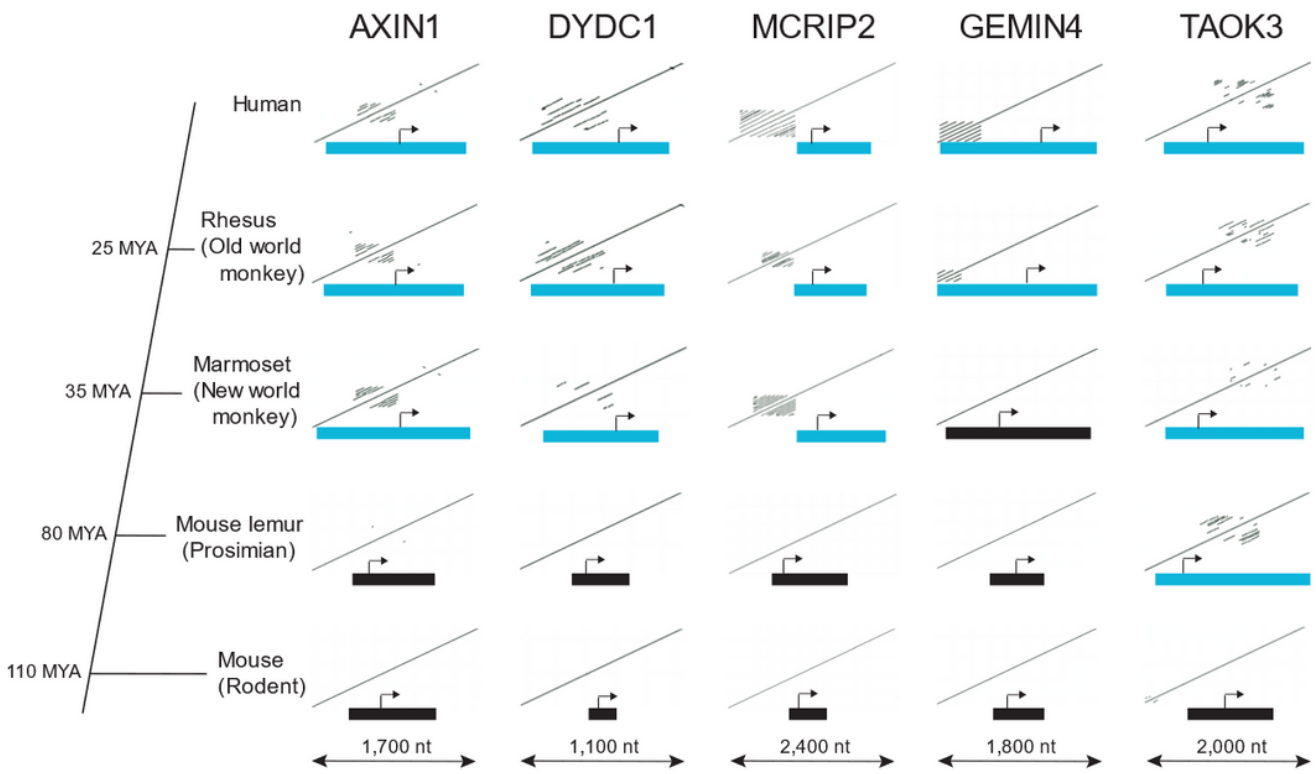


Figure 4

Effects of TR-CGI methylation on gene expression Plots of methylation levels (x-axis) and transcript levels (y-axis) for TR-CGI genes in (**A**) Normal oral tissues and (**B**) Naïve and primed human embryonic stem cells [11].



8

Figure 5

Evolutionary origins of TRs in TR-CGI genes. For each gene, dotplots of CGIs and surrounding sequences from five species are displayed in a column. Blue rectangles are TR-CGIs and black rectangles are CGIs without TRs. Arrows are transcriptional start sites and directions of transcription.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalTablestitles.docx
- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx
- TableS5.xlsx
- TableS6.xlsx
- TableS7.xlsx
- TableS8.xlsx