RESEARCH Open Access

Genome- ide di rib- ion of 5-form lc o ine in embr onic em cell i a ocia ed i h ran crip ion and depend on h mine DNA gl co la e

Eun-Ang Raiber[1†], Dario Beraldi[1,2†], Gabriella Ficz[3], Heather E Burgess[3,4], Miguel R Branco[3,4], Pierre Murat[1], David Oxley[5], Michael J Booth[1], Wolf Reik[3,4*†] and Shankar Balasubramanian[1,2,6*]

Abstract

Background: Methylation of cytosine in DNA (5mC) is an important epigenetic mark that is involved in the regulation of genome function. During early embryonic development in mammals, the methylation landscape is dynamically reprogrammed in part through active demethylation. Recent advances have identified key players involved in active demethylation pathways, including oxidation of 5mC to 5-hydroxymethylcytosine (5hmC) and 5-formylcytosine (5fC) by the TET enzymes, and excision of 5fC by the base excision repair enzyme thymine DNA glycosylase (TDG). Here, we provide the first genome-wide map of 5fC in mouse embryonic stem (ES) cells and evaluate potential roles for 5fC in differentiation.

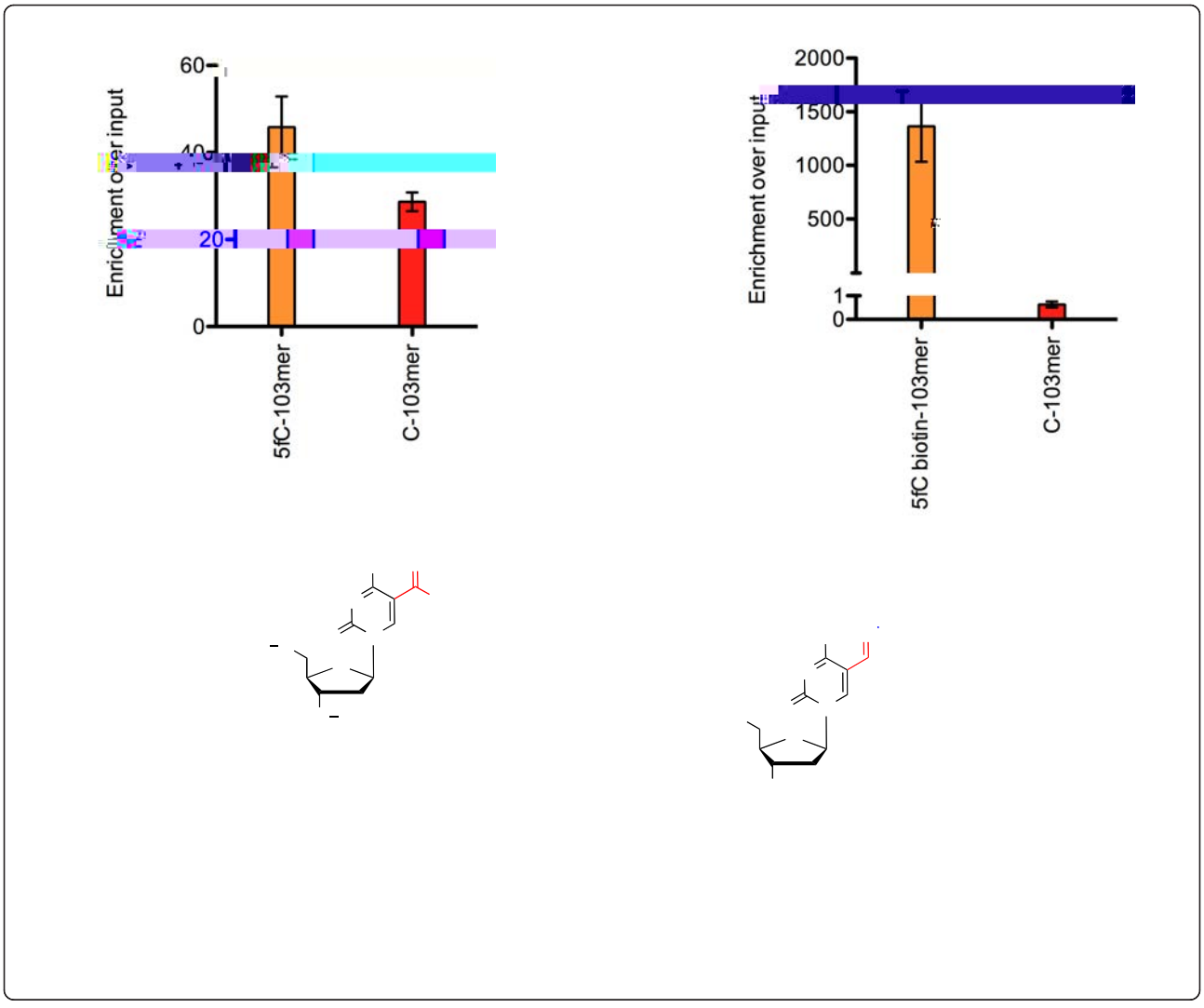
Results: Our method exploits the unique reactivity of 5fC for pulldown and high-throughput sequencing. Genome-wide mapping revealed 5fC enrichment in CpG islands (CGIs) of promoters and exons. CGI promoters in which 5fC was relatively more enriched than 5mC or 5hmC corresponded to transcriptionally active genes. Accordingly, 5fC-rich promoters had elevated H3K4me3 levels, associated with active transcription, and were frequently bound by RNA polymerase II. TDG down-regulation led to 5fC accumulation in CGIs in ES cells, which correlates with increased methylation in these genomic regions during differentiation of ES cells in wild-type and TDG knockout contexts.

Conclusions: Collectively, our data suggest that 5fC plays a role in epigenetic reprogramming within specific genomic regions, which is controlled in part by TDG-mediated excision. Notably, 5fC excision in ES cells is necessary for the correct establishment of CGI methylation patterns during differentiation and hence for appropriate patterns of gene expression during development.

Background

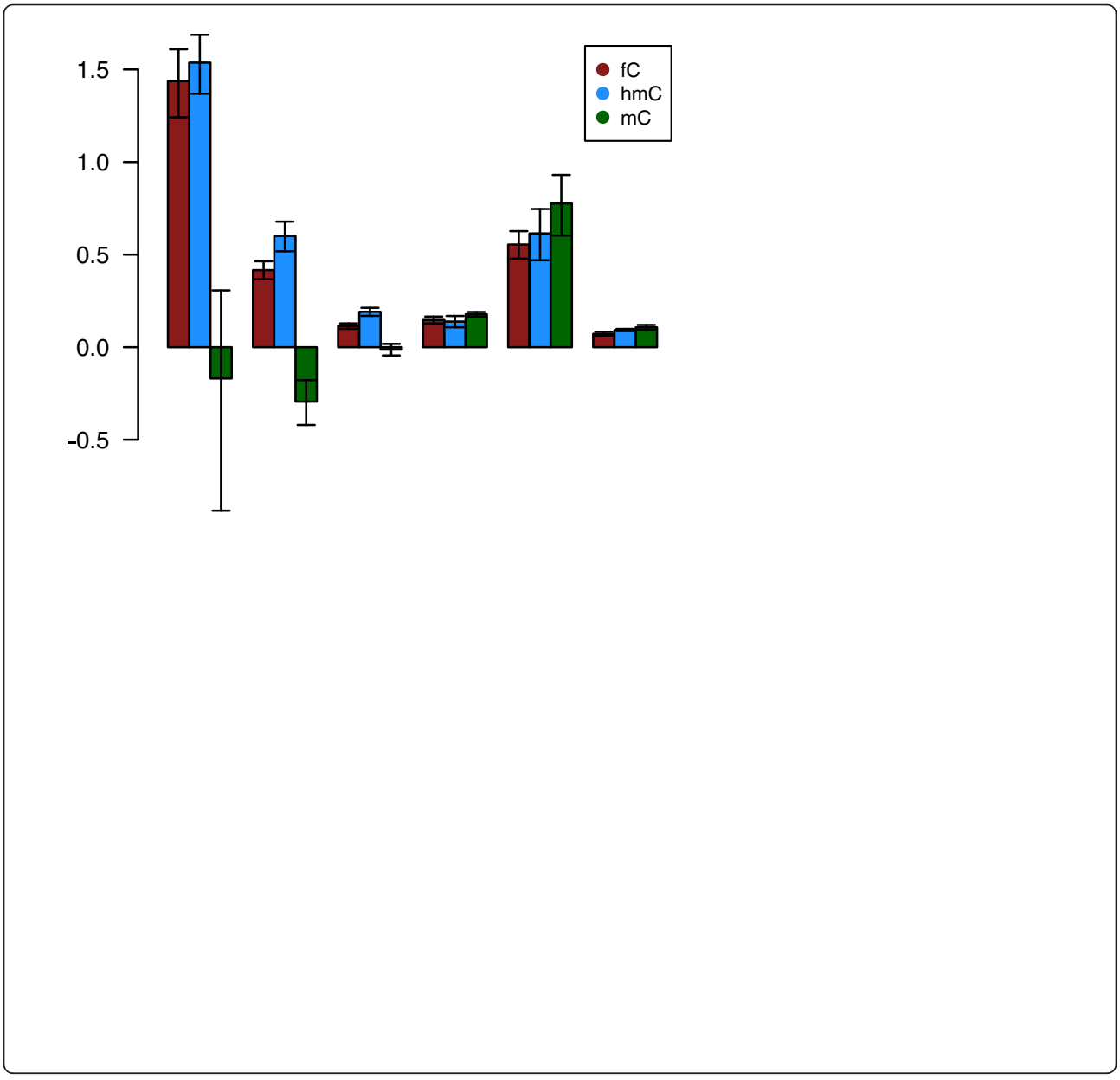
la er chromatograph and tandem liquid chromatograph - mass spectrometr . Quantification of 5fC in genomic ES cell DNA sho ed this modified base to be present at around a level of 0.02 to 0.002% of all c tosine species, hich is roughl 10- to 100-fold lo er than those of 5hmC [4,5]. In ES cells, TET1 and TET2 are highl e pressed and considered to pla roles in reprogramming 5mC and control of the differentiation potential [6,7]. 5fC levels dramaticall decrease ith ongoing differentiation, suggesting its potential involvement during epigenetic reprogramming [5]. Indeed, immunostaining of gotes that undergo global demeth lation has sho n that the appearance of 5fC and 5caC in the male pronucleus is associated ith Tet3-mediated loss of 5mC [8].

Bisulfite treatment and subsequent high throughput sequencing (BS-Seq) has been the gold standard for the detection of c tosine meth lation. This method, ho ever, does not distinguish 5mC from 5hmC or c tosine from 5fC and 5caC. Specific antibodies have been used to enrich and map 5mC (meth lated DNA immunoprecipitation (MeDIP)-Seq) and 5hmC (h dro meth lated DNA immunoprecipitation (hMeDIP)-Seq) [9]. The use of chemical labeling is an alternative method to enrich and sequence 5hmC in the genome [10,11]. The most recent breakthrough in this field came ith t o ne methods allo ing the measurement of 5hmC at single base resolution [12,13]. While various techniques for genome- ide anal sis of 5mC and 5hmC are available, there is currentl no method that allo s the positional mapping of 5fC in



relative enrichment to the input librar rather than the absolute 5fC levels. TET1 binding sites (data taken from [15]) ere enriched in 5hmC and 5fC, but not 5mC, hich is in accordance ith the fact that TET1 is the catal st for the generation of 5hmC and 5fC (Figure s4 in Additional file 1). The genome- ide distribution of 5fC follo ed a similar pattern to 5hmC ith enrichments in euchromatic regions, including CpG islands (CGIs), e ons and promoters (Figure 2a; Additional file 2). We also looked at the 5' UTR of LINE1 and the intracisternal A particle long terminal repeat (IAP LTR), all of hich sho ed enrichments of 5fC in contrast to the depletion in the gene bod of LINE and also other retrotransposon elements (Figure 2b; Additional file 1, Figure s5). The 5' UTR of LINE1 displa ed high levels of 5hmC, medium levels of 5fC and lo levels of 5mC. In contrast, IAP LTR had lo levels of 5hmC, medium levels of 5fC and high levels of 5mC, demonstrating that the kinetics at each o idation stage depends on the genomic conte t. It remains to be seen if these patterns are associated ith active demeth lation.

The profiles sho n in Figure 2c represent the enrichment levels of c tosine modifications for all genes separated into CGI- or non CGI-containing genes. In CGI-containing promoters there is a sharp enrichment peak of 5fC at the transcription start site and a slightl less locali ed enrichment of 5hmC ith a depletion of 5mC at the transcription start site. In contrast, the profile of non-CGI promoter regions of the reference genes sho ed a much less pronounced increase in the levels of both 5mC and 5fC upstream of the transcription start site; these then remain at a constant level throughout the gene bodies. Overall, our anal ses sho that, depending on genomic regions, e observed different distributions of 5fC, 5hmC and 5mC, hich suggests that the kinetics of processing 5mC are distinct bet een genomic regions. That 5fC is



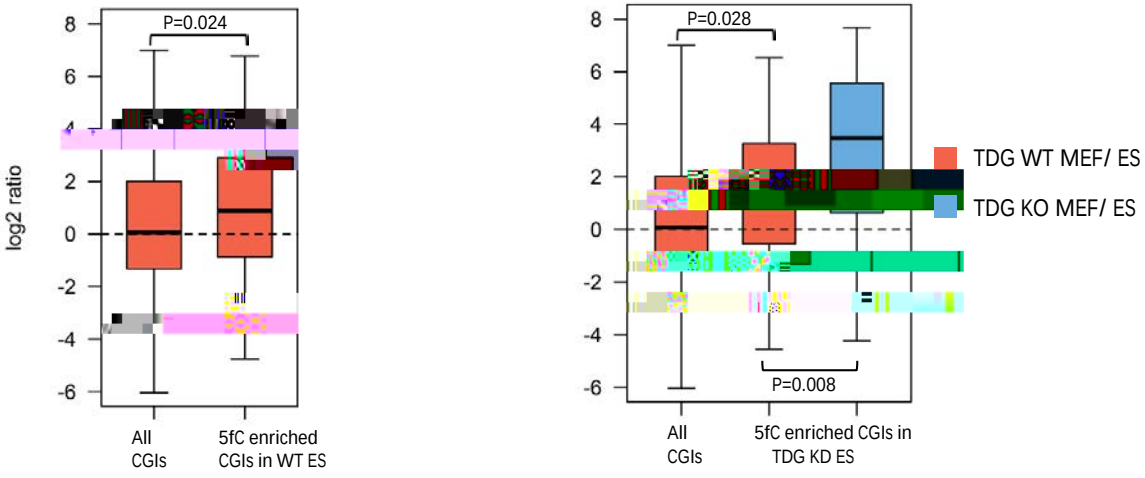
especiall enriched in CGIs also supports the role of 5fC in the maintenance of h pometh lation in these regions in ES cells.

5fC is associated with active gene expression in ES cells

We identified CGIs that sho ed a significant difference in 5fC enrichment compared to 5mC and 5hmC, and further characteri ed them using gene ontolog categories. Therefore, e associated each island ith the nearest gene ithin 5 kb and searched for overrepresented categories in this set. Gene ontolog anal sis of the 5fC-enriched genes identified path a s that ere associated ith transcription regulation (Table s4 in Additional file 1). We also e amined the correlation bet een 5fC at CGI gene promoters and their transcription levels using published gene e pression data [9]. Specificall , e compared gene e pression levels for cases here one c tosine modification in the CGI promoter region as relativel more enriched than other c tosine derivatives. We found that genes hose CGI promoters ere 5fC-rich (relative to 5mC or 5hmC) sho ed higher e pression than the overall group of CGI assigned genes (Figure 3a; Table s5 in Additional file 1). This suggests that the shift in equilibrium bet een the different c tosine modifications at promoter sites ma be linked to mechanisms that control gene activit . Consistent ith this observation, hen genes ere categori ed as lo , medium

levels b si -fold, consistent ith its role in e cising 5fC, hereas meth lation levels sta ed constant (Figure s8 in Additional file 1). In general, e found that more than 98% of 5fC-enriched regions from TDG KD overlapped ith those found in the siRNA control. Genome- ide 5fC mapping of the TDG KD sho ed 5fC-enriched sites ere distributed ith a reduced overall coverage of the genome (5fC sites distributed over 138 Mb in contrast to 415 Mb in the control). Thus, 5fC must be present at higher levels and/or higher densit in the enriched sites for the TDG KD. This also indicates that the formation of 5fC marks at those remaining 277 Mb must be via a distinct path a that is TDG-dependent, perhaps involving TET recruitment b TDG. It can also mean that the loss of 5fC in these particular regions is TDG-independent via an alternative path a .

We then compared the enriched regions bet een TDG KD and siRNA control and found that 5% (out of 138 Mb) ere significantl more highl enriched than in the



from the absence of TDG in the pluripotent stage of the earl embr o ma promote an even higher gain of meth - lation during development. In addition, TDG ma also be acting in complementar path a s at these target CGIs to remove e cess DNA meth lation - for e ample, repairing mismatches resulting from 5mC deamination.

We also anal ed the 5fC distribution follo ing siRNA-mediated do n-regulation of TET1, hich led to a 50% decrease in genomic 5fC as measured b mass spectrome-tr (data not sho n). Due to the presence of TET2 in ES cells, hich presumabl overlaps ith TET1 in binding to chromatin in man genomic regions, e concluded that

In order to control the specificit of the reaction, the same reaction as carried out on ODN1; in the absence of the o idation step, onl the starting material as recovered (Figure s1B in Additional file 1).

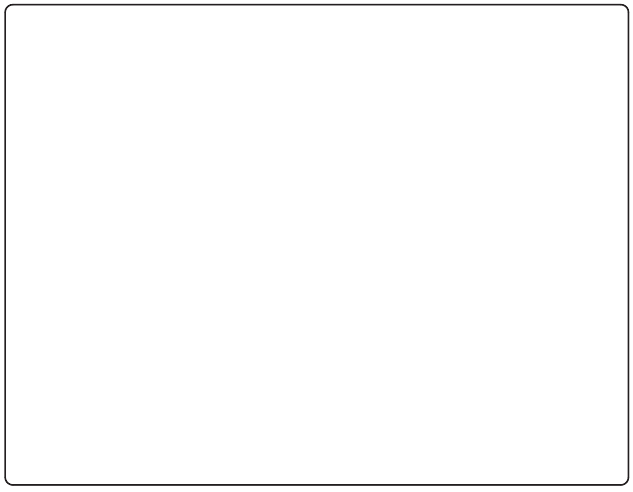
Biotin-labelling of fC in genomic DNA samples

Genomic DNA as prepared b sonicating genomic DNA e tracted from mouse embr onic stem cells J1. Genomic

siGENOME non-targeting siRNA#2 (catalogue number D-001210-02; sequence not available). Cells ere harvested after three rounds of transfection for DNA/RNA isolation.

Bioinformatics and data analysis

Reads in fastq format obtained from the Illumina sequencing pipeline have been aligned against the mouse genome (NCBI version mm9) using b a [21]



the manuscript; HB carried out sample preparation, assisted with data analysis and provided feedback on the manuscript; MB assisted with data analysis and provided feedback on the manuscript; PM assisted with sample analysis and provided feedback on the manuscript; DO assisted with sample analysis; MJB