Genome-wide single-molecule analysis of long-read DNA methylation reveals heterogeneous patterns at heterochromatin

Lyndsay Kerr[1], Ramon Grima[2] and Duncan Sproul[3]

[1*]MRC Institute of Genetics and Cancer, University of Edinburgh, UK.
[2]School of Biological Sciences, University of Edinburgh, UK.
[3]MRC Human Genetics Unit and CRUK Edinburgh Centre, Institute of Genetics and Cancer, University of Edinburgh, UK.

Contributing authors: lyndsay.kerr@ed.ac.uk; ramon.grima@ed.ac.uk; d.sproul@ed.ac.uk;

Abstract

Whole-genome bisulfite sequencing is central to our current understanding of the human methylome. This technology can probe single-molecule methylation patterns, but limits analyses to reads of a few hundred basepairs. Here we analyse intra-molecular DNA methylation patterns genome-wide in single Nanopore reads of mean length 24.6kb. We show that bulk-level metrics underestimate large-scale intra-molecular heterogeneity in the methylome. Quantifying heterogeneity using the read transition score, we find that it varies significantly across the human genome. We observe the greatest heterogeneity in single-molecule patterns within heterochromatic regions and that reads originating from euchromatic regions and gene bodies have more ordered DNA methylation patterns. We also find oscillatory single-molecule DNA methylation patterns with a wavelength around 180bp that we propose are nucleosome-induced and may account for some of this heterogeneity. Our findings demonstrate the power of single-molecule analysis of long-read data to understand the structure of the human methylome.

Keywords: Nanopore, DNA methylation, epigenetics, single-molecule, heterogeneity, heterochromatin

1 Introduction

DNA methylation is an epigenetic mark commonly associated with gene repression [1]. In mammals it is prevalent on cytosines within CpG dinucleotides, with $70 - 80\%$ of CpGs being methylated in most mammalian cells [2].

High-throughput sequencing approaches have resulted in the description of many human methylomes using bisulfite conversion in combination with whole-genome sequencing (Whole-Genome Bisulfite Sequencing, WGBS) [3, 4]. Analysis of WGBS data has confirmed the pervasive nature of CpG methylation in human genomes and uncovered differences in the DNA methylation patterns between different cell types and tissues [5, 6]. They have also detailed differences in DNA methylation occurring in human disease, particularly cancer [7].

These insights have arisen from the analysis of bulk DNA methylation patterns. High-throughput bisulfite sequencing data can also be analysed at the single-read level to understand single-molecule heterogeneity in DNA methylation patterns. Analyses of single-read patterns have been used to quantify the degree of inter-molecular methylation heterogeneity in human samples [8, 9]. Further analyses have quantified differences in this heterogeneity between normal and cancer cells using information theory [10]. DNA methylation heterogeneity has been shown to change over time in cell culture [11] and in acute myeloid leukemia upon relapse [12]. Inter-molecular methylation heterogeneity has been attributed to epi-polymorphisms [8] and allele-specific methylation [13]. The mixture of cell types in a sample is also an important source of DNA methylation heterogeneity [14, 15], that could explain part of the inter-molecular heterogeneity seen in these analyses.

However, the degradation of DNA caused by bisulfite conversion [16] and the widespread use of Illumina sequencing technology limits single-molecule analyses of WGBS to short regions of a few hundred basepairs. The limited number of CpGs captured in these reads restricts analyses of intra-molecular DNA heterogeneity. Nanopore sequencing enables the generation of reads over a megabase in length [17], four orders of magnitude longer than those assayed by WGBS. DNA modifications including 5-methylcytosine can also be directly detected in these reads without chemical conversion [18]. This potentially enables the analysis of molecular heterogeneity in DNA methylation patterns at a far greater scale than previously examined.

Here we conduct a genome-wide analysis of single-molecule DNA methylation patterns in long reads derived from Nanopore sequencing in order to understand the nature of large-scale intra-molecular DNA methylation heterogeneity in the human genome. Our work demonstrates that intra-molecular DNA methylation heterogeneity is abundant in heterochromatin and is underestimated by bulk analysis of DNA methylation patterns.

2 Results

2.1 Large-scale methylation heterogeneity is underestimated by bulk analysis of methylomes

To understand the nature of large-scale heterogeneity in DNA methylation patterns, we analysed Nanopore sequencing data derived from GM24385 lymphoblastoid cells. This consisted of $6,289,480$ reads aligned to autosomes with a mean length of 24.6kb.

We first asked how large-scale intra-molecular DNA methylation heterogeneity varied across single reads. We therefore calculated the mean level of methylation for all aligned single reads along with the coefficient of variation and correlation to capture intra-molecular variation (Fig. 1**a**). To focus our attention on large-scale DNA methylation patterns, and reduce noise associated with calculating statistics from a low number of CpGs, only aligned reads with methylation called for $\geq$ 100 CpG sites were considered ($2,908,181$). All of these single-read measurements varied considerably across aligned reads suggesting that substantial large-scale intra-molecular DNA methylation heterogeneity exists in human cells (Fig. 1**b-d**).

We then asked how these single-read measurements compared to bulk-level measurements (Fig. 1**a**). To facilitate this comparison, we segmented the autosomal genome into 100kb non-overlapping windows and calculated the mean of each of the single-read statistics for each window using the reads that aligned entirely within that windows (total of $1,707,118$ reads for all windows). We compared these mean single-molecule measures to bulk-level statistics for the same windows. We calculated the bulk statistical measure for a window by first calculating the mean methylation level of each CpG in the window using all reads in which it had a defined methylation state. We then calculated the bulk statistics using these mean CpG methylation values. We hypothesised that the bulk and single-read approaches would provide similar estimates of mean methylation levels but differ in their quantification of variation. Measurements of the mean single-read methylation levels in the genomic windows were significantly correlated with bulk mean levels of DNA methylation within the same windows (Fig. 1**e**, Spearman correlation $= 0.944$, $p < 2.20 \times 10^{-16}$). While a two-sided paired t-test indicated significant difference in the bulk means compared to the single-read means, the absolute relative difference of the single-read means compared to the bulk means was small (mean absolute relative difference of 0.036). As the single-read and bulk estimates of mean methylation are mathematically equivalent, these small differences are likely to be caused by differences in the exclusion of some reads and CpGs from the two approaches. As expected, these results suggest that single-read measurements capture largely the same information regarding overall methylation levels as bulk measurements.

We next considered how single-read measurements of heterogeneity compared to their bulk counterparts. Like mean methylation levels, the mean

single-read and bulk measurements of the coefficient of variation and correlation were significantly correlated (Fig. 1**f, g**, Spearman correlation = 0.880 and 0.924, respectively, $p < 2.20 \times 10^{-16}$ in both cases). However, mean single-read coefficients of variation observed in the genomic windows were significantly greater than that of the equivalent bulk measurements (Fig. 1**f**, paired t-test $p < 2.20 \times 10^{-16}$, mean absolute relative difference of 0.332) and a linear model fitted to the two deviated from the line of identity (Fig. 1**f** solid vs. dashed line). We also observed that mean single-read correlations were significantly lower than that of the bulk measurements (Fig. 1**g**, paired t-test $p < 2.20 \times 10^{-16}$, mean absolute relative difference of 0.667) and a linear model fitted to the two deviated from the line of identity (Fig. 1**g** solid vs. dashed line). To ensure that these results were not specific to the window size used, we repeated the analysis using 50kb and 200kb genomic windows observing similar results as to when 100kb windows were used (Fig. S1).

These comparisons of single-molecule to bulk measurements suggest that the analysis of DNA methylation in single Nanopore reads captures large-scale intra-molecular DNA methylation heterogeneity that is missed by analysis of bulk DNA methylation patterns.

2.2 Intra-molecular heterogeneity varies across the human genome

Having established that significant large-scale intra-molecular DNA methylation heterogeneity exists in the human methylome, we next asked how this heterogeneity was organised across the genome.

Many measures previously used to quantify single-molecule methylation heterogeneity from short-read data [19] are difficult to apply to long reads because the number of possible states increases with read length. We therefore, measured single-read methylation heterogeneity using the read transition score (RTS; [20], Eq. 11, Fig. S2**a**). The RTS measures the probability that two neighbouring CpGs have different methylation states and ranges from 0 to 1. An RTS of 1 indicates that no neighbouring CpGs have the same methylation state and an RTS of 0 indicates that only a single methylation state is observed in the read (Fig. S2**a**).

To ask how intra-molecular heterogeneity varied across the genome, we calculated the RTS of all $2,908,181$ reads containing $\geq$ 100 CpGs that mapped to autosomes. The resulting distribution of individual read RTS was bimodal (Fig. 2**a**). We observed a significantly different unimodal distribution (Kolmogorov-Smirnov test, $p < 2.20 \times 10^{-16}$) characterised by higher RTS values when we calculated the RTS using reads simulated to have a random intra-molecular methylation pattern (Fig. 2**a**, see methods). These results suggest that the degree of intra-molecular heterogeneity might vary by location in the genome and that heterogeneous reads had methylation patterns that were non-random.

To understand whether the degree of methylation heterogeneity varied by location in the genome, we calculated the mean RTS across reads entirely

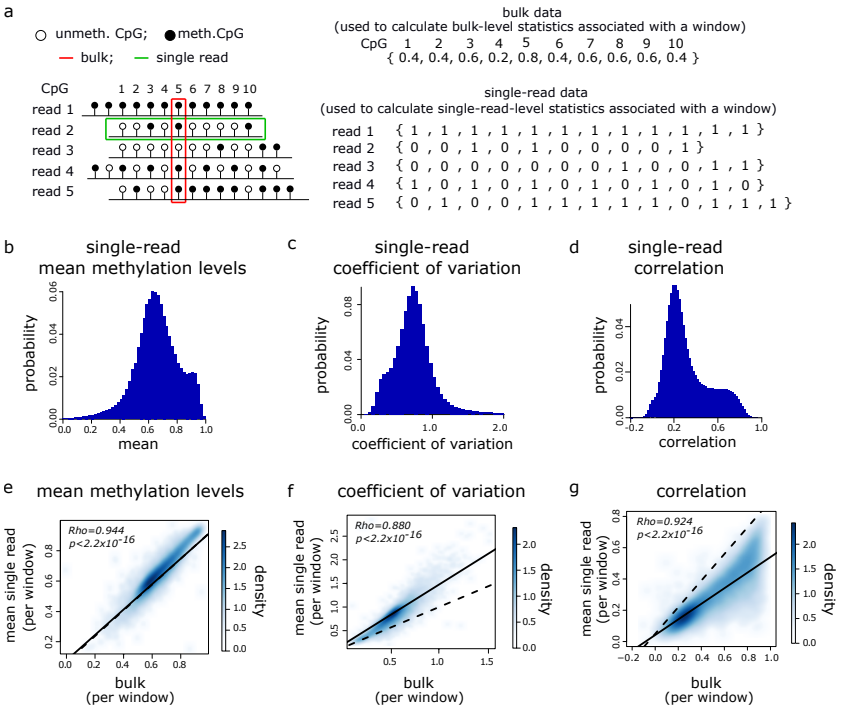

Fig. 1: Large-scale methylation heterogeneity is underestimated by bulk analysis of methylomes. **a** Left: schematic of analysis approach. Example single-read (green box) and bulk (red box) analysis for 5 hypothetical reads overlapping 10 common CpGs are shown. Right: the bulk methylation level of a CpG is its number of methylated observations divided by its total number of unmethylated and methylated observations. The red box highlights information used to calculate the bulk methylation level of CpG 5. Single-read data is obtained by creating a vector, v, for each read, where $v_i = 0, 1$ if CpG i in the read is unmethylated or methylated respectively. The green box highlights information used to construct v for read 2. **b–d** Histograms showing the distributions of single-read statistics for the $2,908,181$ reads with methylation states for ≥ 100 CpGs. (**b**) mean methylation levels, (**c**) coefficients of variation and (**d**) correlations between neighbouring CpGs. **e–g** Density scatter plots of mean single-read vs. bulk statistics in 100kb genomic windows ($n = 26,686$). (**e**) mean methylation levels, (**f**) coefficients of variation and (**g**) correlations between neighbouring CpGs. Dashed and solid lines show lines of identity and a linear model fitted to the data respectively. Spearman correlations are shown as are p-values from paired t-tests comparing bulk to mean single-read statistics.

aligned within 100kb windows across the genome (Eq. 12, Fig. 2**b**, [21]). We then compared the distribution of mean RTS to that obtained when we

assigned random reads to each genomic window (shuffled RTS). The distribution of mean RTS for genomic windows was bimodal and significantly different from the unimodal distribution observed for the shuffled RTS (Kolmogorov-Smirnov test, $p < 2.20 \times 10^{-16}$, Fig. 2**c, d**). This is consistent with some regions of the human genome possessing a greater degree of intra-molecular methylation heterogeneity than others.

To validate that intra-molecular methylation heterogeneity varies across the genome, we conducted a similar analysis using an alternative measure of intra-molecular heterogeneity, the Shannon entropy. While the Shannon entropy can measure methylation heterogeneity, it does not quantify local information on the methylation status of neighbouring CpGs as the RTS does. Although we did not observe a bimodal distribution of the Shannon entropies for individual reads (Fig. S2**b**), the Shannon entropy for individual reads and the mean Shannon entropy for 100Kb genomic windows were significantly correlated with RTS and mean RTS respectively (Spearman correlations = 0.541 and 0.870, respectively, $p < 2.20 \times 10^{-16}$ in each case, Fig. S2**c,d**). In addition, the distribution of mean Shannon Entropy scores observed for 100Kb windows was significantly different from that obtained from shuffled reads consistent with the degree of intra-molecular heterogeneity varying across the human genome (Fig. S2**e,f**, Kolmogorov-Smirnov test, $p < 2.20 \times 10^{-16}$).

Taken together, these analyses suggest that the degree of large-scale intra-molecular DNA methylation heterogeneity varies between different regions of the human genome.

2.3 Intra-molecular and inter-molecular methylation heterogeneity co-occur in the human genome

Intra-molecular methylation heterogeneity can occur in a manner that is consistent or inconsistent across reads arising from the same region of the genome. This means that a high RTS could be compatible with either high or low inter-molecular methylation heterogeneity (Fig. S3).

To understand the degree of inter-molecular heterogeneity associated with regions of the genome displaying high intra-molecular methylation heterogeneity, we computed the measure d_i. This quantifies the degree of inter-molecular heterogeneity across reads for a given CpG as $d_i = |m_i - 0.5|$ for a CpG i with methylation level m_i. This ranges from 0, which indicates maximum inter-read heterogeneity when the unmethylated and methylated states are observed equally across reads, to 0.5, when a CpG has a uniformly methylated or unmethylated state.

We then calculated d_i for every CpG in the genome before comparing the values observed for regions of the genome associated with the highest or lowest degree of intra-read methylation heterogeneity as measured by the RTS. The distribution of d_i values for the 10% of 100kb genomic windows with the lowest mean RTS were significantly skewed towards high values of d_i as compared to the background of all genomic windows (Fig. 3**a,b**, $p < 2.20 \times 10^{-16}$, Kolmogorov-Smirnov test). This indicates that these regions

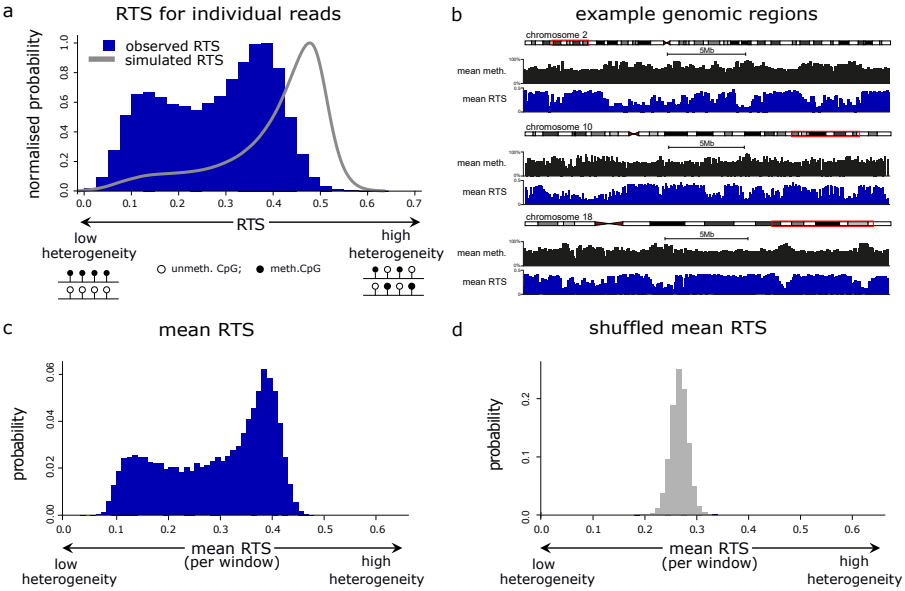

Fig. 2: Intra-molecular heterogeneity varies across the human genome. **a** Histogram showing the RTS distribution for the $2,908,181$ reads containing ≥ 100 CpGs. This is compared to the RTS calculated from reads simulated to have a random intra-molecular methylation pattern (grey line). **b** Genome browser plots showing the mean methylation levels and mean RTS in 100kb windows for selected genomic windows. **c** Histogram of the mean RTS distribution for 100kb genomic windows ($n = 26,951$), calculated using reads that align entirely within each window. **d** As for **c** but for the shuffled mean RTS.

display low inter-read heterogeneity as well as low intra-read heterogeneity. In contrast the d_i scores of CpGs found in the 10% of 100kb genomic windows with the highest mean RTS were significantly lower than the background of all genomic windows (Fig. 3**a,c**, $p < 2.20 \times 10^{-16}$, Kolmogorov-Smirnov test) suggesting that they also show a greater degree of inter-read heterogeneity.

These analyses of CpG methylation levels demonstrate that regions of the genome with a high degree of intra-molecular DNA methylation heterogeneity also posses high inter-molecular methylation heterogeneity.

2.4 Heterochromatin possesses heterogeneous methylation patterns

To understand the nature of the regions of the human genome which had the greatest degree of intra-molecular DNA methylation heterogeneity, we asked whether they were associated with particular genomic annotations or chromatin states.

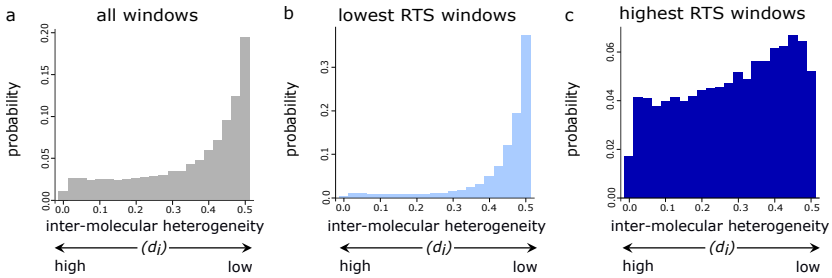

Fig. 3: Intra-molecular and inter-molecular methylation heterogeneity co-occur in the human genome. **a**–**c**: Histograms showing the distributions of d_i **a** for the 26, 169, 612 CpGs in all windows, **b** for the 3, 855, 660 CpGs in the 10% of windows with the lowest RTS and **c** for the 1, 471, 188 CpGs in the 10% of windows with the highest RTS.

We first analysed the 10% of 100kb genomic windows with the highest mean RTS (i.e. the 10% most heterogeneous windows). We observed that these genomic windows were significantly depleted in both genes and CpG islands compared to the background of all genomic windows (fold-enrichment of 0.768 and 0.011, Wilcoxon tests $p = 8.33 \times 10^{-7}$ and $p < 2.20 \times 10^{-16}$, respectively). In contrast, the least heterogeneous genomic windows (the 10% with the lowest mean RTS), were significantly enriched in genes and CpG islands (fold-enrichments of 1.41 and 2.67, respectively, Wilcoxon tests $p < 2.20 \times 10^{-16}$ in both cases).

To further understand the characteristics of regions displaying high methylation heterogeneity, we cross-referenced them to chromatin state data generated by the ENCODE project [22, 23]. These projects have used hidden Markov models to partition the genome into distinct chromatin states (ChromHMM). Consistent with their observed enrichment in genes and CpG islands, the least heterogeneous genomic windows were significantly enriched in ChromHMM-defined promoters, enhancers and transcribed regions from GM12878 lymphoblastoid cells (Fig. 4**a**, Wilcoxon tests with $p < 2.20 \times 10^{-16}$ in each case). In contrast, the most heterogeneous windows were significantly enriched in ChromHMM-defined heterochromatic regions (Fig. 4**b**, Wilcoxon test with $p < 2.20 \times 10^{-16}$). We observed similar enrichment results when the analysis was repeated using 50kb and 200kb genomic windows (Fig. S4**a**, **b**) suggesting that our observations were independent of the choice of genomic window size.

In some somatic cells, heterochromatin is associated with regions of intermediate methylation levels that are termed partially methylated domains (PMDs) [3, 24]. To further assess the correspondence between heterochromatin and intra-molecular DNA methylation heterogeneity, we therefore called 685 PMDs in GM24385 cells from our Nanopore DNA methylation data (Fig. S4**c**) and defined the rest of the genome as 588 non-PMDs. GM24385 PMDs showed substantial overlap with those previously defined in 11 other lymphoblastoid

cells from WGBS data (Jaccard = 0.806) [25]. Consistent with their enrichment in the ChromHMM-defined heterochromatin state, genomic windows with the highest mean RTS were significantly enriched in GM24385 PMDs (Fig. 4**c**, Wilcoxon test $p < 2.20 \times 10^{-16}$) with 89.9% of these windows overlapping PMDs by at least half the window length. Those windows with the lowest mean RTS were significantly enriched in non-PMDs (Fig. 4**c**, Wilcoxon test $p < 2.20 \times 10^{-16}$). We observed similar enrichment results when the analysis was repeated using 50kb and 200kb genomic windows (Fig. S4**d**) suggesting that overlap between intra-molecular methylation heterogeneity and PMDs was independent of the choice of genomic window size.

To examine how methylation heterogeneity associates with heterochromatin and PMDs more generally, we analysed the correlation between the mean RTS in 100Kb genomic windows and the proportion of window overlapping the chromHMM heterochromatin state or PMDs. Mean window RTS was significantly correlated with both the proportion of the window covered by the chromHMM heterochromatin state and PMDs (Fig. 4**d,e** Spearman correlation = 0.795 and 0.695 respectively, both $p < 2.20 \times 10^{-16}$). Furthermore, we also found significant correlations between a read's RTS and the proportion of the read that overlapped the heterochromatic state and PMDs (Spearman correlation = 0.708 and 0.611 respectively, both $p < 2.20 \times 10^{-16}$) confirming that the association between intra-molecular DNA heterogeneity and heterochromatin was maintained at the single-read level.

Taken together, these analyses suggest that the regions of the human genome displaying the greatest degree of large-scale intra-molecular methylation heterogeneity are heterochromatic partially methylated domains.

2.5 Oscillatory DNA patterns are observed in single reads with a high heterogeneity

Having shown that heterogeneity in DNA methylation patterns is observed in heterochromatin, we now sought to further understand the nature of intra-molecular DNA methylation heterogeneity in these regions.

In order to do so, we first defined single reads exhibiting low or high intra-molecular methylation heterogeneity by fitting the sum of two Gaussian distributions to the bimodal distribution of RTS (Fig. 2**a**, low RTS reads and high RTS reads are equivalent to the left and right peaks respectively). Consistent with our previous analyses, the mean methylation and correlation between neighbouring sites in high RTS reads were significantly lower than in the low RTS group reads (Fig. S5**a,c**, Wilcoxon tests, $p < 2.20 \times 10^{-16}$ in each case), while the coefficient of variation was significantly higher in high RTS reads (Fig. S5**b**, Wilcoxon test, $p < 2.20 \times 10^{-16}$). Furthermore, high RTS group reads were significantly enriched in heterochromatin and PMDs, while the low RTS group reads were enriched in promoters, enhancers, transcribed regions and non-PMDs (Wilcoxon tests, $p < 2.20 \times 10^{-16}$ in each case).

In order to understand the nature of DNA methylation heterogeneity within single molecules, we then calculated the correlation between DNA methylation

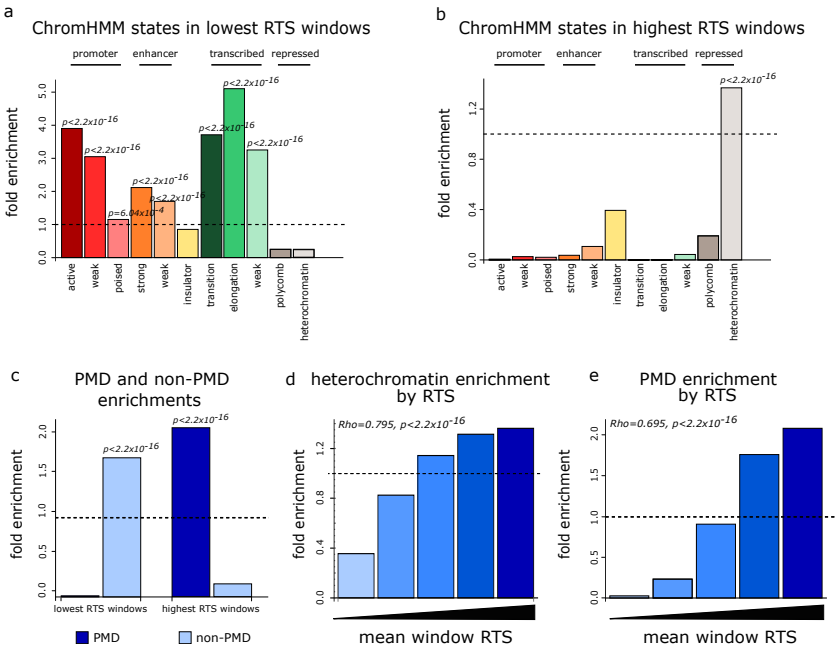

Fig. 4: Heterochromatin possesses heterogeneous methylation patterns. **a,b**: Barplots showing enrichment of the 10% least **a** and most **b** heterogeneous 100kb windows in chromHMM states defined in GM12878 lympoblastoid cells. **c**: Barplot showing the enrichment of the 10% least and most heterogeneous 100kb windows in PMDs and non-PMDs. **d**: Barplot showing enrichment of continuous heterogeneity groups of genomic windows (defined from the ranked mean RTS) in the ChromHMM heterochromatin annotation from GM12878 lymphoblastoid cells. **e**: Barplot showing enrichment of continuous heterogeneity groups of genomic windows (defined from the ranked mean RTS) in PMDs. In **d, e** the Spearman correlations (Rho) between mean window RTS and window overlap with heterochromatin or PMDs, respectively, are also shown along with the relevant p-values.

status for pairs of CpG sites within single reads located within 500bp of each other. Consistent with our observations of the overall correlation within single reads (Fig. S5**c**), mean distance-dependent correlations were higher for low RTS reads than high RTS reads across all distances analysed (Fig. 5**a**). Within both low and high RTS reads, the mean distance-dependent correlation was highest for CpGs located close to each other and reduced as the distance between CpGs increased (Fig. 5**a**). However, after initially falling, the mean distance-dependent correlation for high RTS reads then increased, showing oscillatory behaviour with a wavelength of around 180bp (Fig. 5**a**). Smaller-scale oscillations with a wavelength around 10bp were also observed in the mean distance-dependent correlation for the high RTS reads (Fig. 5**a**). Both

of these oscillatory behaviours were much less apparent in the mean distance-dependent correlation profile for low RTS reads (Fig. 5**a**).

These analyses demonstrate that the parts of the human genome showing high intra-molecular DNA methylation heterogeneity display oscillatory DNA methylation patterns within single molecules.

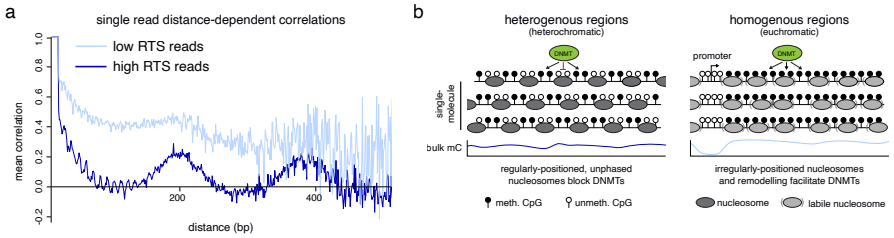

Fig. 5: Oscillatory DNA patterns are observed in single molecules with a high heterogeneity. **a**: Plot of mean single-read distance-dependent correlation between nearest-neighbour CpG sites within 500bp of each other for 730, 687 low RTS reads and 983, 188 high RTS reads. **b** Model of how intra-molecular DNA methylation heterogeneity varies across the human genome. Areas of high heterogeneity are heterochromatic and have oscillatory single-molecule DNA methylation patterns (left). We propose that these are caused by methylation of linker DNA due to nucleosomes being refractory to DNMTs [26]. Regularly positioned but un-phased nucleosomes in heterochromatin [27, 28] mean that sites of methylation are not aligned across molecules masking these patterns from bulk analyses (bulk mC). More homogeneous methylation patterns are observed in euchromatic, genic regions (right) where nucleosomes are less regularly spaced [27, 28] and more labile due to the action of nucleosome remodellers. This may increase access to nucleosomal DNA by DNMTs.

3 Discussion

Here, we have used Nanopore sequencing data to analyse large-scale intra-molecular DNA methylation heterogeneity in the human genome. Our results show that intra-molecular heterogeneity in DNA methylation patterns is abundant and underestimated by bulk-level metrics. We specifically observe a high degree of intra-molecular DNA methylation heterogeneity in heterochromatic regions and that this heterogeneity manifests as oscillatory DNA methylation patterns (Fig. 5**a**).

We propose that the most likely source of this heterogeneity within single molecules are nucleosomes. The total length of nucleosomal and inter-nucleosomal linker DNA varies in the range of $160 - 230$bp depending on the organism, cell type and region of the genome examined [29–32]. *In vitro* experiments demonstrate that the wrapping of DNA around nucleosomes is

refractory to methylation by DNMTs [33–36] so that methylation occurs at the linker DNA [37]. This suggests that the blocking of DNMTs by nucleosomes could generate the oscillations of around 180bp that we observe in reads with high intra-molecular heterogeneity (Fig. 5**b**).

Analyses of the relationship between DNA methylation and nucleosomes *in vivo* have been restricted to the estimated 8.7% of nucleosomes in the human genome that are strongly positioned and phased across molecules [38]. Linker DNA methylation is seen in between the well positioned nucleosomes around CTCF sites in human cells [26] consistent with the hypothesis that nucleosomes are refractory to DNMTs. However, most nucleosomes in the human genome are not phased between molecules [29] and it has required the application of single-molecule and single-cell approaches to probe nucleosomal organisation in the majority of the genome [27, 28]. These studies suggest that in heterochromatin, nucleosomes are regularly spaced within molecules but unphased between molecules [27, 28]. We propose that our single-molecule analysis similarly enables us to detect a widespread impact of this heterochromatic nucleosomal organisation on DNA methylation patterns despite the lack of phasing across molecules (Fig. 5**b**). In euchromatic, genic regions nucleosomes are less regularly spaced [27, 28]. Nucleosomal remodelling is also observed, for example through the action of FACT in displacing nucleosomes during transcription [39]. This could enable DNMTs to access nucleosomal DNA leading to the more uniform DNA methylation patterns we observe in euchromatin (Fig. 5**b**).

A variety of statistics have previously been used to examine heterogeneity in DNA methylation patterns. These include the proportion of discordant reads (PDR) [40], methylation haplotype load (MHL) [41] and fraction of discordant reads (FDR) [19]. However, these measures quantify inter-molecular heterogeneity and so are not directly comparable to our analysis of intra-molecular heterogeneity. An information theory approach has been used to predict probability distributions associated with single-molecule methylation patterns from WGBS data and suggested that the genome is organised into domains of disordered and ordered methylation patterns that coincide with large-scale chromatin organisation [10]. Bulk analysis of WGBS also reported disordered methylation levels in PMDs with evidence of oscillations [42]. However, although these studies are consistent with our observation of high intra-molecular heterogeneity in heterochromatin, WGBS reads are not sufficiently long to directly assess single-molecule DNA methylation patterns across the scale of multiple nucleosomes as we have done here.

4 Conclusion

Taken together our observations suggest that intra-molecular heterogeneity in DNA methylation patterns is widespread in the human genome. We propose the nature of this heterogeneity reveals a greater role for nucleosomes in

shaping DNA methylation patterns in the human genome than was previously appreciated by bulk analyses.

5 Methods

Unless otherwise specified, we performed analysis using R (v. 4.1.2), Mathematica (v. 12.3) and Bedtools (v. 2.30.0) [43]. All statistical tests were two-sided.

5.1 Preprocessing of Nanopore data

We downloaded raw Nanopore reads in fast5 format for GM24385 lymphoblastoid cells from Oxford Nanopore (amazon bucket s3:// ont-open-data/gm24385_2020_09/). We then basecalled these reads using the dna_r9.4.1_450bps_hac_prom.cfg configuration of Guppy (v. 5.0.11) [44], aligned them to the human genome (hg38 assembly) using the map-ont setting of Minimap2 (v. 2.22) [45] and called methylation state at CpGs using Nanopolish with default settings (v. 0.13.2) [18]. Methylation states were assigned to each CpG site based on the log likelihood ratio (LLR) calculated by Nanopolish. We used the LLR cut-offs used by Nanopolish scripts and estimated to have an accuracy $> 92.5\%$ [18]. Comparisons to WGBS also suggest Nanopolish methylation calls are accurate [46, 47]. Here we consider single-molecule patterns ≥ 100 CpGs in length so the influence of a small number of miscalled methylation states is likely to be minimal. CpG sites with LLR ≤ -2 were assigned an unmethylated status and a methylated status if LLR ≥ 2. For $\mid$ LLR $\mid < 2$ we deemed the CpG site to be unassignable. Using these settings, 72.6% of observations were assigned methylated or unmethylated states. Only reads mapping to canonical hg38 autosomes (chromosomes 1-22) were considered in our analyses.

5.2 Calculation of single-read statistics

We converted methylation tsv files outputted by Nanopolish to bed format using the mtsv2bedGraph.py script from the timplab/nanopore-methylation-utilities GitHub repository [48]. We used a custom R script to identify the $2,908,181$ out of $6,289,480$ reads that had defined methylation states for ≥ 100 CpGs. For each retained read, r, we discarded CpG sites that could not be assigned a methylation state and stored the methylation states of the remaining $N_r \geq 100$ CpGs in a vector $v_r = \{v_{r,1}, v_{r,2}, \ldots, v_{r,N_r}\}$, where for $i = 1, \ldots N_r$,

$$v_{r,i} = \begin{cases} 1, & \text{if CpG } i \text{ in read } r \text{ is methylated,} \\ 0, & \text{if CpG } i \text{ in read } r \text{ is unmethylated.} \end{cases} \tag{1}$$

For each read r we then used R (v. 4.0.1) to calculate the mean methylation level (μ_r), coefficient of variation (CV_r) and the correlation in methylation level between neighbouring sites (ρ_r) via

$$\mu_r = \frac{1}{N_r} \sum_{i=1}^{N_r} v_{r,i}, \tag{2}$$

$$\mathrm{CV}_r = \frac{\sigma_r}{\mu_r}, \quad \text{where} \quad \sigma_r^2 = \frac{1}{N_r - 1} \sum_{i=1}^{N_r} (v_{r,i} - \mu_r)^2, \tag{3}$$

$$\rho_r = \frac{\sum_{i=1}^{N_r-1} (v_{r,i} - \hat{\mu})(v_{r,i+1} - \tilde{\mu})}{\sqrt{\sum_{i=1}^{N_r-1} (v_{r,i} - \hat{\mu})^2 \sum_{i=2}^{N_r} (v_{r,i} - \tilde{\mu})^2}}, \tag{4}$$

where

$$\hat{\mu} = \frac{1}{N_r - 1} \sum_{i=1}^{N_r-1} v_{r,i} \quad \text{and} \quad \tilde{\mu} = \frac{1}{N_r - 1} \sum_{i=2}^{N_r} v_{r,i}.$$

5.3 Definition of genomic windows

We split canonical autosomal chromosomes into 28,760 non-overlapping genomic windows of 100kb using BEDOPS (v. 2.4.26) [49]. Note that 22 windows (corresponding to the end of each autosome) were shorter than 100kb. This process was repeated but with 50kb and 200kb windows for selected analyses.

5.4 Calculation of mean single-read statistics

We used Bedtools and custom command line scripts to extract the $1,707,118$ reads that were entirely contained within a single window and had called methylation states for $\geq$ 100 CpGs. We then calculated mean single-read statistics associated with each window using R (v. 4.0.1). In particular, if $r_1, r_2, \ldots, r_{N_w}$ were the reads that aligned entirely within a window, n, then we calculated the single-molecule mean ($\mu_{s,n}$), coefficient of variation ($\mathrm{CV}_{s,n}$) and correlation ($\rho_{s,n}$) associated with window n via

$$\mu_{s,n} = \frac{1}{N_w} \sum_{r=1}^{N_w} \mu_r, \qquad \mathrm{CV}_{s,n} = \frac{1}{N_w} \sum_{r=1}^{N_w} \mathrm{CV}_r, \qquad \rho_{s,n} = \frac{1}{N_w} \sum_{r=1}^{N_w} \rho_r. \tag{5}$$

This lead to mean single-molecule statistics being obtained for $26,952$ 100kb windows.

5.5 Calculation of bulk-level window statistics

We defined the methylation coverage for each CpG site to be the combined number of methylated and unmethylated calls associated with that CpG site

using all of the $6,289,480$ reads covering that CpG. We discarded CpG sites with < 10 methylation calls and those that overlapped the boundary between two windows (i.e. if the forward-strand cytosine belongs to one window and the guanine to another). The mean CpG methylation for each site was defined as the number of methylated calls associated with that CpG divided by the total number of calls (methylated and unmethylated) observed for that CpG. We then removed 100kb windows with < 100 retained CpG sites. We used the CpG methylation levels associated with the remaining $26,703$ windows to calculate bulk-level statistics for each window in R (v. 4.0.1). In particular, for window n with $N \geq 100$ retained CpGs with methylation levels $m_1, m_2, \ldots, m_N$, we calculated the bulk mean methylation level ($\mu_{b,n}$), coefficient of variation ($\mathrm{CV}_{b,n}$) and the correlation in methylation level between nearest-neighbour sites ($\rho_{b,n}$) via

$$\mu_{b,n} = \frac{1}{N}\sum_{i=1}^{N} m_i, \tag{6}$$

$$\mathrm{CV}_{b,n} = \frac{\sigma_{b,n}}{\mu_{b,n}}, \quad \text{where} \quad \sigma^2_{b,n} = \frac{1}{N-1}\sum_{i=1}^{N}(m_i - \mu_{b,n})^2, \tag{7}$$

$$\rho_{b,n} = \frac{\sum_{i=1}^{N-1}(m_i - \hat{\mu}_n)(m_{i+1} - \tilde{\mu}_n)}{\sqrt{\sum_{i=1}^{N-1}(m_i - \hat{\mu}_n)^2 \sum_{i=2}^{N}(m_i - \tilde{\mu}_n)^2}}, \tag{8}$$

where

$$\hat{\mu}_n = \frac{1}{N-1}\sum_{i=1}^{N-1} m_i \quad \text{and} \quad \tilde{\mu}_n = \frac{1}{N-1}\sum_{i=2}^{N} m_i.$$

5.6 Comparison of mean single-molecule and bulk-level window statistics

We used Spearman correlation tests in R to calculate correlations and associated p-values between the bulk and mean single-molecule statistics for the $26,686$ 100kb genomic windows that had sufficient bulk and single-molecule data. We also fit linear models to the bulk-level and single-molecule statistics using the base lm command in R and conducted paired t-tests to assess the difference between the bulk and single-molecule values. The average absolute relative difference between the bulk and single-molecule mean was calculated via

$$\frac{1}{26,686}\sum_{n=1}^{26,686}\frac{|\mu_{b,n} - \mu_{s,n}|}{\mu_{s,n}}. \tag{9}$$

Similarly, the average absolute relative difference between the bulk and single-molecule coefficient of variation and correlation were calculated via

$$\frac{1}{26,686}\sum_{n=1}^{26,686}\frac{|\mathrm{CV}_{b,n}-\mathrm{CV}_{s,n}|}{\mathrm{CV}_{s,n}} \qquad \text{and} \qquad \frac{1}{26,686}\sum_{n=1}^{26,686}\frac{|\rho_{b,n}-\rho_{s,n}|}{\rho_{s,n}}, \tag{10}$$

respectively.

5.7 Calculation of RTS

For each read r containing methylation states for ≥ 100 CpGs, we calculated the read transition score in R (v. 4.0.1) as

$$\mathrm{RTS}_r = \frac{1}{N-1}\sum_{i=1}^{N-1}|v_{r,i+1}-v_{r,i}|. \tag{11}$$

RTS_r essentially measures the probability that two nearest-neighbour CpG sites in read r have different methylation states–hence reads which are entirely unmethylated or entirely methylated have RTS of 0, while reads where no two nearest-neighbour CpG sites share the same methylation state have RTS of 1.

Mean RTSs were then calculated for each genomic window similarly as for the other mean single-read statistics. In particular, if $\{r_1, r_2, \ldots, r_{N_w}\}$ were the reads that aligned entirely within a window, n, then we took the mean RTS associated with window n to be

$$\mathrm{RTS}_n = \frac{1}{N_w}\sum_{i=1}^{N_w}\mathrm{RTS}_{r_i}. \tag{12}$$

5.8 Generation of synthetic and shuffled RTS

To generate synthetic RTSs, we simulated the case where methylation states were randomly distributed along each read. For each read r with methylation states for ≥ 100 CpGs, we simulated a synthetic read containing the same number of CpGs, but where the methylation state of each CpG was randomly chosen from a binomial distribution with success rate equal to the mean methylation level of the read (μ_r). The RTS was then calculated for each of these synthetic reads. The RTS distribution obtained from the true reads was compared to that obtained from the synthetic reads and a Kolmogorov-Smirnov test was conducted in Mathematica to assess whether the observed methylation patterns in true reads are random or non-random.

To generate the mean shuffled RTS for genomic windows we calculated a mean RTS for each window using random reads (rather than those that align to that window). Reads were shuffled using the command-line *shuf* command. For example, if $\{r_1, r_2, \ldots, r_{N_w}\}$ were the reads aligning entirely within window n, then we replaced these by a set of random reads $\{\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_{N_w}\}$, and

calculated a mean RTS for window n via

$$\widehat{\mathrm{RTS}}_n = \frac{1}{N_w} \sum_{i=1}^{N_w} \mathrm{RTS}_{\hat{r}_i}.$$

We then compared the RTS_n and $\widehat{\mathrm{RTS}}_n$ distribution using a Kolmogorov-Smirnov test in R.

5.9 Calculation of Shannon entropy

For each read, r, with methylation states for ≥ 100 CpGs, we used R (v. 4.0.1) to calculate the Shannon entropy as

$$\mathrm{SE}_r = \begin{cases} (1-\mu_r)\log_2\big(1/(1-\mu_r)\big) + \mu_r \log_2\big(1/\mu_r\big), & \text{if } \mu_r \neq 0 \text{ and } \mu_r \neq 1, \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

If $\{r_1, r_2, \ldots, r_{N_w}\}$ are the reads that aligned entirely within a window, n, then we took the overall Shannon entropy associated with window n to be

$$\mathrm{SE}_n = \frac{1}{N_w} \sum_{i=1}^{N_w} \mathrm{SE}_{r_i}. \tag{14}$$

We calculated the Spearman correlation between RTS_r and SE_r and between RTS_n and SE_n in R. Shuffled mean Shannon entropies were also obtained for each window in a similar manner as described in Section. 5.8.

5.10 Defining windows with the lowest and highest intra-molecular heterogeneity

Using R we ranked the $26,951$ windows by RTS_n before defining the windows with the greatest intra-molecular heterogeneity as the 10% of windows with the highest mean RTS and the least heterogenous windows as the 10% of windows with the lowest mean RTS.

5.11 Comparison of intra-molecular heterogeneity levels to inter-molecular heterogeneity levels

We computed the measure $d_i = |m_i - 0.5|$, where m_i is the methylation level of CpG i for each CpG with ≥ 10 methylation calls. $d_i = 0.5$ indicates low molecule-to-molecule heterogeneity for CpG i, where as $d_i = 0$ corresponds to high molecule-to-molecule heterogeneity for CpG i. We performed Kolmogorov-Smirnov tests to compare the distribution of d_i associated with the 10% of windows with the highest RTS and the distribution associated with the rest of the genome and to compare the 10% of windows with the lowest

RTS and the distribution associated with the rest of the genome. In particular, we used the Kolmogorov-Smirnov test in R to test whether the cumulative distribution function (CDF) associated with d_i for the highest or lowest RTS windows is significantly different than that associated with the rest of the genome.

5.12 Sourcing and processing of genomic annotations

Using the command line, we derived bed files of hg38 gene locations using annotation files from Gencode (v. 41) [50]. CGI annotations were taken from [51]. Overlapping CGI intervals were merged using BEDtools (v. 2.27.1) before they were converted to hg38 positions using the UCSC browser liftover tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Non-autosomal CGIs were excluded from the analysis as were CGIs located in ENCODE blacklist regions [22] as previously described [52].

We downloaded hg19 GM12878 lymphoblastoid cell chromHMM states from UCSC (http://www.genome.ucsc.edu/cgi-bin/hgTables), [22, 23] before lifting them over to hg38 genome coordinates (https://genome.ucsc.edu/cgi-bin/hgLiftOver). To remove cases where multiple hg19 coordinates mapped to the same hg38 coordinate, we used the command line and Bedtools to remove from the analysis any regions associated with more than one annotation type in hg38 as a consequence of this liftover.

We downloaded PMD locations identified for 11 lymphoblastoid cell lines from the AllPMDs.tar.bz2 file in the bdecatoPMD_Paper_Scripts GitHub repository [25]. These coordinates were then converted from hg19 to the hg38 coordinates using Liftover (https://genome.ucsc.edu/cgi-bin/hgLiftOver). We excluded poorly mapped regions of the genome using annotations of gaps and centromeres from the UCSC browser (hg38 gap and centromere tracks). Annotations were downloaded from the UCSC table browser. Regions annotated as heterochromatin, short arm and telomeres from the gaps track were merged with the centromeres track using Bedtools merge with -d set to 10Mb. This merged file was then excluded from the PMD BED files using Bedtools subtract. Using Bedtools we also excluded PMDs shorter than 200kb.

5.13 Definition and analysis of PMDs

We called PMDs for GM24385 cells using Methpipe (v. 5.0.0) [25] on mean bulk CpG methylation levels and CpG coverage calculated from Nanopore data. We defined non-PMDs as genomic regions that did not overlap with a PMD. Poorly mapped regions of the genome were removed from the PMD and non-PMD annotations as described in Section 5.12. Using Bedtools we also excluded PMDs and non-PMDs shorter than 200kb. We compared the overlap between GM24385 PMDs to those from other lymphoblastoid cells [25] using the Bedtools Jaccard function on a BED file generated by merging the PMDs annotated in the other cell lines using Bedtools merge.

5.14 Testing enrichment of annotations in genomic windows

For each annotation type, a, we used Bedtools to compute

- $P_{a,\text{all}}$: the proportion of bases in all $26,951$ windows analysed that overlap the annotation (i.e. the sum of bases from all windows that overlap the annotation divided by the summed length of all windows);
- $P_{a,\text{least}}$: the proportion of bases in the least heterogeneous windows that overlap the annotation (i.e. the sum of bases from the least heterogeneous windows that overlap the annotation divided by the summed length of the least heterogeneous windows);
- $P_{a,\text{most}}$: the proportion of bases in the most heterogeneous windows that overlap the annotation (i.e. the sum of bases from the most heterogeneous windows that overlap the annotation divided by the summed length of the most heterogeneous windows).

We then calculated the enrichment of the least and most heterogeneous windows in annotation type a using

$$e_{a,\text{least}} = P_{a,\text{least}}/P_{a,\text{all}} \quad \text{and} \quad e_{a,\text{most}} = P_{a,\text{most}}/P_{a,\text{all}},$$

respectively. To test the significance of the enrichment, we extracted the proportion that each window overlaps annotation type a using Bedtools and compared the distribution of proportions associated with the least/most heterogeneous windows to the proportions associated with all windows using Wilcoxon tests in R.

5.15 Testing continuous relationship between RTS and annotations

To test the continuous relationship between a windows mean RTS and the proportion of overlap with genomic annotations, we ranked 100kb genomic windows by their mean RTS before analysing how RTS correlates with the percentage overlap with genomic annotations using the Spearman correlation test in R. For graphing, we split windows into five equally sized groups based on their ranked mean RTS. We then repeated the analyses in Section 5.14 to calculate the annotation enrichment level in each group.

To test the relationship between single-read RTS and genomic annotations, we used Bedtools to extracted the proportion of each read that overlapped the ChromHMM heterochromatin state or PMDs and conducted Spearman correlation tests in R.

5.16 Derivation of high and low-RTS reads by mixture modelling

Using Mathematica we fit the sum of two Gaussian distributions, with means μ_1, μ_2 ($\mu_1 < \mu_2$) and standard deviations σ_1, σ_2, to the bimodal distribution

for the read RTS (Fig. 2**a**). Of the $2,908,181$ reads containing methylation information for a minimum of 100 CpG sites, we defined the $791,654$ reads that satisfied $|\mathrm{RTS}_r - \mu_1| \leq \sigma_1$ to be the "low RTS group" and the $1,157,194$ that satisfied $|\mathrm{RTS}_r - \mu_2| \leq \sigma_2$ to be the "high RTS group", where RTS_r is the RTS associated with each read. We discarded the remaining reads.

We visualised the distributions of statistics for the low and high RTS reads, as calculated in Section 5.2, and conducted Wilcoxon tests in R to assess the differences in the statistics between the two groups. Enrichment analyses were also conducted for the low and high RTS read groups similarly to Section 5.14.

5.17 Calculating distance-dependent correlations in low and high RTS reads

For each read in the low and high RTS groups, we extracted the methylation states of pairs of "nearest-neighbour" CpGs, along with the distances in bases between these neighbouring CpGs. In particular, for a read r covering N_r CpG sites we created three vectors $u_r = \{u_{r,1}, u_{r,2}, \ldots, u_{r,N_r-1}\}$, $w_r = \{w_{r,1}, w_{r,2}, \ldots, w_{r,N_r-1}\}$, $d_r = \{d_{r,1}, d_{r,2}, \ldots, d_{r,N_r-1}\}$ where, for $i = 1, 2, \ldots, N_r - 1$

$$u_{r,i} = \begin{cases} 1, & \text{if CpG } i \text{ in read } r \text{ is methylated,} \\ 0, & \text{if CpG } i \text{ in read } r \text{ is unmethylated,} \\ \mathrm{NA}, & \text{if CpG } i \text{ in read } r \text{ is unassigned,} \end{cases}$$

$$w_{r,i} = \begin{cases} 1, & \text{if CpG } i+1 \text{ in read } r \text{ is methylated,} \\ 0, & \text{if CpG } i+1 \text{ in read } r \text{ is unmethylated,} \\ \mathrm{NA}, & \text{if CpG } i+1 \text{ in read } r \text{ is unassigned,} \end{cases}$$

and $d_{r,i}$ is the distance between CpG i and CpG $i+1$ in the read. If the length of the set $\{i : u_{r,i} \neq \mathrm{NA} \text{ and } w_{r,i} \neq \mathrm{NA}\}$ was less than 100, then we discarded the read. Otherwise, for each distance $d = \{1, 2, \ldots, 500\}$, we extracted the elements $\{i : u_{r,i} \neq \mathrm{NA} \text{ and } w_{r,i} \neq \mathrm{NA} \text{ and } d_{r,i} = d\}$ of u_r and w_r to obtain subvectors $\hat{u}_{r,d}$ and $\hat{w}_{r,d}$ of equal length L, giving information regarding the methylation state of neighbouring CpGs a distance d apart. If $L \geq 2$, then we used R (v. 4.0.1) to compute the correlation between neighbouring sites on read r a distance of d apart via

$$\rho_{r,d} = \frac{\sum_{i=1}^{L} (\hat{u}_{r,d,i} - \mu_{u,r,d})(\hat{w}_{r,d,i} - \mu_{w,r,d})}{\sqrt{\sum_{i=1}^{L} (\hat{u}_{r,d,i} - \mu_{u,r,d})^2 \sum_{i=1}^{L} (\hat{w}_{r,d,i} - \mu_{w,r,d})^2}},$$

where

$$\mu_{u,r,d} = \frac{1}{L} \sum_{i=1}^{L} \hat{u}_{r,d,i} \quad \text{and} \quad \mu_{w,r,d} = \frac{1}{L} \sum_{i=1}^{L} \hat{w}_{r,d,i}.$$

For each $d = \{2, 3, \ldots, 500\}$ we took the mean correlation $\rho_{r,d}$ over all low RTS group reads to obtain mean single-molecule distance-dependent correlations associated with the low RTS group. We obtained mean single-molecule distance-dependent correlations associated with the high RTS group analogously.

Author contributions. LK conducted the processing and analyses presented in the manuscript. DS and RG supervised the study. LK, DS and RG planned the study and wrote the manuscript.

Acknowledgements. We thank Chris Ponting, Marcus Wilson, Andreas Kapourani and members of the Sproul and Grima lab for helpful discussions about the manuscript. We thank Ioannis Kafetzopoulos for providing the annotation of poorly mapped regions of the human genome.

LK is a cross-disciplinary post-doctoral fellow supported by funding from the University of Edinburgh and Medical Research Council (MC_UU_00009/2). RG is supported by Leverhulme Trust research awards (RPG-2018-423 and RPG-2020-327). DS is a Cancer Research UK Career Development fellow (reference C47648/A20837), and work in his laboratory is also supported by an MRC university grant to the MRC Human Genetics Unit.

Data accessibility. Raw data used is available from Oxford Nanopore Technologies (amazon bucket s3://ont-open-data/gm24385_2020_09/).

Competing interests. The authors declare no competing interests.

References

[1] Schübeler, D.: Function and information content of dna methylation. Nature **517**(7534), 321–326 (2015)

[2] Suzuki, M.M., Bird, A.: Dna methylation landscapes: provocative insights from epigenomics. Nature reviews genetics **9**(6), 465–476 (2008)

[3] Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., *et al.*: Human dna methylomes at base resolution show widespread epigenomic differences. nature **462**(7271), 315–322 (2009)

[4] Urich, M.A., Nery, J.R., Lister, R., Schmitz, R.J., Ecker, J.R.: Methylc-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nature protocols **10**(3), 475–483 (2015)

[5] Ziller, M.J., Gu, H., Müller, F., Donaghey, J., Tsai, L.T.-Y., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., *et al.*: Charting a dynamic dna methylation landscape of the human genome. Nature **500**(7463), 477–481 (2013)

[6] Schultz, M.D., He, Y., Whitaker, J.W., Hariharan, M., Mukamel, E.A., Leung, D., Rajagopal, N., Nery, J.R., Urich, M.A., Chen, H., *et al.*: Human body epigenome maps reveal noncanonical dna methylation variation. Nature **523**(7559), 212–216 (2015)

[7] Berman, B.P., Weisenberger, D.J., Aman, J.F., Hinoue, T., Ramjan, Z., Liu, Y., Noushmehr, H., Lange, C.P., van Dijk, C.M., Tollenaar, R.A., *et al.*: Regions of focal dna hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina–associated domains. Nature genetics **44**(1), 40 (2012)

[8] Landan, G., Cohen, N.M., Mukamel, Z., Bar, A., Molchadsky, A., Brosh, R., Horn-Saban, S., Zalcenstein, D.A., Goldfinger, N., Zundelevich, A., *et al.*: Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nature genetics **44**(11), 1207–1214 (2012)

[9] Sheffield, N.C., Pierron, G., Klughammer, J., Datlinger, P., Schönegger, A., Schuster, M., Hadler, J., Surdez, D., Guillemot, D., Lapouble, E., *et al.*: Dna methylation heterogeneity defines a disease spectrum in ewing sarcoma. Nature medicine **23**(3), 386–395 (2017)

[10] Jenkinson, G., Pujadas, E., Goutsias, J., Feinberg, A.P.: Potential energy landscapes identify the information-theoretic nature of the epigenome. Nature genetics **49**(5), 719–729 (2017)

[11] Shipony, Z., Mukamel, Z., Cohen, N.M., Landan, G., Chomsky, E., Zeliger, S.R., Fried, Y.C., Ainbinder, E., Friedman, N., Tanay, A.: Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. Nature **513**(7516), 115–119 (2014)

[12] Li, S., Garrett-Bakelman, F.E., Chung, S.S., Sanders, M.A., Hricik, T., Rapaport, F., Patel, J., Dillon, R., Vijay, P., Brown, A.L., *et al.*: Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. Nature medicine **22**(7), 792–799 (2016)

[13] Abante, J., Fang, Y., Feinberg, A., Goutsias, J.: Detection of haplotype-dependent allele-specific dna methylation in wgbs data. Nature communications **11**(1), 1–13 (2020)

[14] Bergstedt, J., Azzou, S.A.K., Tsuo, K., Jaquaniello, A., Urrutia, A., Rotival, M., Lin, D.T., MacIsaac, J.L., Kobor, M.S., Albert, M.L., *et al.*: The immune factors driving dna methylation variation in human blood. Nature communications **13**(1), 1–20 (2022)

[15] Luo, C., Keown, C.L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J.R., Sandoval, J.P., *et al.*: Single-cell methylomes

identify neuronal subtypes and regulatory elements in mammalian cortex. Science **357**(6351), 600–604 (2017)

[16] Tanaka, K., Okamoto, A.: Degradation of dna by bisulfite treatment. Bioorganic & medicinal chemistry letters **17**(7), 1912–1915 (2007)

[17] Payne, A., Holmes, N., Rakyan, V., Loose, M.: Bulkvis: a graphical viewer for oxford nanopore bulk fast5 files. Bioinformatics **35**(13), 2193–2198 (2019)

[18] Simpson, J.T., Workman, R.E., Zuzarte, P., David, M., Dursi, L., Timp, W.: Detecting dna cytosine methylation using nanopore sequencing. Nature methods **14**(4), 407–410 (2017)

[19] Scherer, M., Nebel, A., Franke, A., Walter, J., Lengauer, T., Bock, C., Müller, F., List, M.: Quantitative comparison of within-sample heterogeneity scores for dna methylation data. Nucleic acids research **48**(8), 46–46 (2020)

[20] Hetzel, S., Giesselmann, P., Reinert, K., Meissner, A., Kretzmer, H.: Rlm: fast and simplified extraction of read-level methylation metrics from bisulfite sequencing data. Bioinformatics **37**(21), 3934–3935 (2021)

[21] Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P.: Integrative genomics viewer. Nature biotechnology **29**(1), 24–26 (2011)

[22] Consortium, E.P., *et al.*: An integrated encyclopedia of dna elements in the human genome. Nature **489**(7414), 57 (2012)

[23] Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., *et al.*: Mapping and analysis of chromatin state dynamics in nine human cell types. Nature **473**(7345), 43–49 (2011)

[24] Hovestadt, V., Jones, D.T., Picelli, S., Wang, W., Kool, M., Northcott, P.A., Sultan, M., Stachurski, K., Ryzhova, M., Warnatz, H.-J., *et al.*: Decoding the regulatory landscape of medulloblastoma using dna methylation sequencing. Nature **510**(7506), 537–541 (2014)

[25] Decato, B.E., Qu, J., Ji, X., Wagenblast, E., Knott, S.R., Hannon, G.J., Smith, A.D.: Characterization of universal features of partially methylated domains across tissues and species. Epigenetics & chromatin **13**(1), 1–14 (2020)

[26] Kelly, T.K., Liu, Y., Lay, F.D., Liang, G., Berman, B.P., Jones, P.A.: Genome-wide mapping of nucleosome positioning and dna methylation

within individual dna molecules. Genome research **22**(12), 2497–2506 (2012)

[27] Baldi, S., Krebs, S., Blum, H., Becker, P.B.: Genome-wide measurement of local nucleosome array regularity and spacing by nanopore sequencing. Nature structural & molecular biology **25**(9), 894–901 (2018)

[28] Lai, B., Gao, W., Cui, K., Xie, W., Tang, Q., Jin, W., Hu, G., Ni, B., Zhao, K.: Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. Nature **562**(7726), 281–285 (2018)

[29] Baldi, S., Korber, P., Becker, P.B.: Beads on a string—nucleosome array arrangements and folding of the chromatin fiber. Nature structural & molecular biology **27**(2), 109–118 (2020)

[30] Van Holde, K.E.: Chromatin. Springer, ??? (2012)

[31] Teif, V.B., Vainshtein, Y., Caudron-Herger, M., Mallm, J.-P., Marth, C., Höfer, T., Rippe, K.: Genome-wide nucleosome positioning during embryonic stem cell development. Nature structural & molecular biology **19**(11), 1185–1192 (2012)

[32] Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z., Sidow, A.: Determinants of nucleosome organization in primary human cells. Nature **474**(7352), 516–520 (2011)

[33] Felle, M., Hoffmeister, H., Rothammer, J., Fuchs, A., Exler, J.H., Längst, G.: Nucleosomes protect dna from dna methylation in vivo and in vitro. Nucleic acids research **39**(16), 6956–6969 (2011)

[34] Schrader, A., Gross, T., Thalhammer, V., Längst, G.: Characterization of dnmt1 binding and dna methylation on nucleosomes and nucleosomal arrays. PloS one **10**(10), 0140076 (2015)

[35] Robertson, A.K., Geiman, T.M., Sankpal, U.T., Hager, G.L., Robertson, K.D.: Effects of chromatin structure on the enzymatic and dna binding functions of dna methyltransferases dnmt1 and dnmt3a in vitro. Biochemical and biophysical research communications **322**(1), 110–118 (2004)

[36] Takeshima, H., Suetake, I., Shimahara, H., Ura, K., Tate, S.-i., Tajima, S.: Distinct dna methylation activity of dnmt3a and dnmt3b towards naked and nucleosomal dna. Journal of biochemistry **139**(3), 503–515 (2006)

[37] Takeshima, H., Suetake, I., Tajima, S.: Mouse dnmt3a preferentially methylates linker dna and is inhibited by histone h1. Journal of molecular biology **383**(4), 810–821 (2008)

[38] Gaffney, D.J., McVicker, G., Pai, A.A., Fondufe-Mittendorf, Y.N., Lewellen, N., Michelini, K., Widom, J., Gilad, Y., Pritchard, J.K.: Controls of nucleosome positioning in the human genome. PLoS genetics **8**(11), 1003036 (2012)

[39] Orphanides, G., LeRoy, G., Chang, C.-H., Luse, D.S., Reinberg, D.: Fact, a factor that facilitates transcript elongation through nucleosomes. Cell **92**(1), 105–116 (1998)

[40] Landau, D.A., Clement, K., Ziller, M.J., Boyle, P., Fan, J., Gu, H., Stevenson, K., Sougnez, C., Wang, L., Li, S., *et al.*: Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. Cancer cell **26**(6), 813–825 (2014)

[41] Guo, S., Diep, D., Plongthongkum, N., Fung, H.-L., Zhang, K., Zhang, K.: Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma dna. Nature genetics **49**(4), 635–642 (2017)

[42] Gaidatzis, D., Burger, L., Murr, R., Lerch, A., Dessus-Babus, S., Schübeler, D., Stadler, M.B.: Dna sequence explains seemingly disordered methylation levels in partially methylated domains of mammalian genomes. PLoS genetics **10**(2), 1004143 (2014)

[43] Quinlan, A.R., Hall, I.M.: Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6), 841–842 (2010)

[44] Technologies, O.N.: Guppy protocol: modified base calling. https://community.nanoporetech.com/docs/prepare/library_prep_protocols/Guppy-protocol/v/gpb_2003_v1_revam_14dec2018/modified-base-calling, Accessed 11 Nov 2022

[45] Li, H.: Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**(18), 3094–3100 (2018)

[46] Liu, Y., Rosikiewicz, W., Pan, Z., Jillette, N., Wang, P., Taghbalout, A., Foox, J., Mason, C., Carroll, M., Cheng, A., *et al.*: Dna methylation-calling tools for oxford nanopore sequencing: a survey and human epigenome-wide evaluation. Genome biology **22**(1), 1–33 (2021)

[47] Akbari, V., Garant, J.-M., O'Neill, K., Pandoh, P., Moore, R., Marra, M.A., Hirst, M., Jones, S.J.: Megabase-scale methylation phasing using nanopore long reads and nanomethphase. Genome biology **22**(1), 1–21 (2021)

[48] Lee, I., Razaghi, R., Gilpatrick, T., Molnar, M., Gershman, A., Sadowski, N., Sedlazeck, F.J., Hansen, K.D., Simpson, J.T., Timp, W.: Simultaneous

profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Nature Methods **17**(12), 1191–1199 (2020)

[49] Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., *et al.*: Bedops: high-performance genomic feature operations. Bioinformatics **28**(14), 1919–1920 (2012)

[50] Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., *et al.*: Gencode 2021. Nucleic acids research **49**(D1), 916–923 (2021)

[51] Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R., Bird, A.P.: Orphan cpg islands identify numerous conserved promoters in the mammalian genome. PLoS genetics **6**(9), 1001134 (2010)

[52] Masalmeh, R.H.A., Taglini, F., Rubio-Ramon, C., Musialik, K.I., Higham, J., Davidson-Smith, H., Kafetzopoulos, I., Pawlicka, K.P., Finan, H.M., Clark, R., *et al.*: De novo dna methyltransferase activity in colorectal cancer is directed towards h3k36me3 marked cpg islands. Nature communications **12**(1), 1–13 (2021)

Supplementary Information

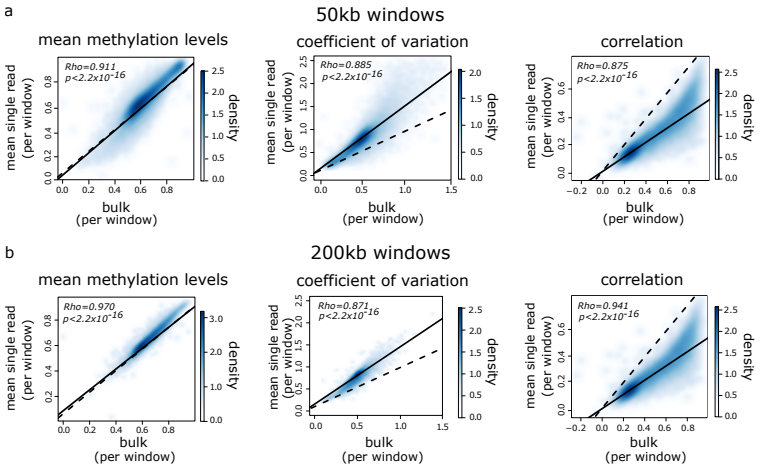

Fig. S1: Large-scale methylation heterogeneity is underestimated by bulk analysis of methylomes independently of the window size used. Density scatter plots of statistics of mean single-read vs. bulk statistics for 50kb genomic windows ($n = 52,839$) **a** and 200kb genomic windows ($n = 13,438$) **b**. Left: mean methylation level, middle: coefficient of variation, right: correlation between neighbouring CpG sites. Dashed and solid lines show lines of identity and a linear model fitted to the data respectively. Spearman correlations are shown, as are p-values from paired t-tests comparing bulk to single-read statistics.

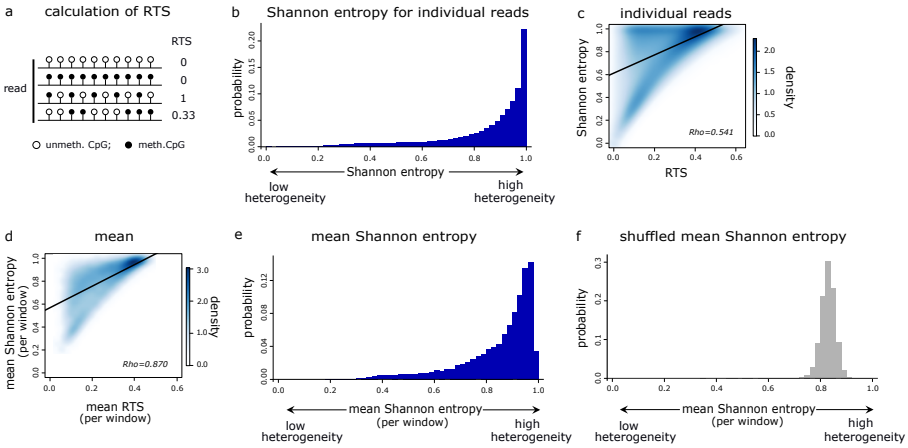

Fig. S2: Intra-molecular heterogeneity varies across the human genome. **a** Schematic of RTS values calculated for hypothetical 10 CpG reads. **b** Histogram showing the Shannon Entropy distribution for the $2,908,181$ individual reads containing at least 100 CpGs. **c** Density scatter plot of the Shannon entropy vs. RTS for individual reads. **d** Density scatter plot of the mean Shannon entropy vs. mean RTS for 100kb genomic windows. In **c, d** the solid line shows a linear model fitted to the data and the Spearman correlation coefficient is shown in the bottom right. **e** Histogram showing the distribution of mean Shannon entropies for 100kb windows ($n = 26,951$) calculated using reads that align entirely within these windows. **f** Histogram showing the mean shuffled Shannon entropy in 100kb windows ($n = 26,951$), calculated using random reads rather than those that align to each window.

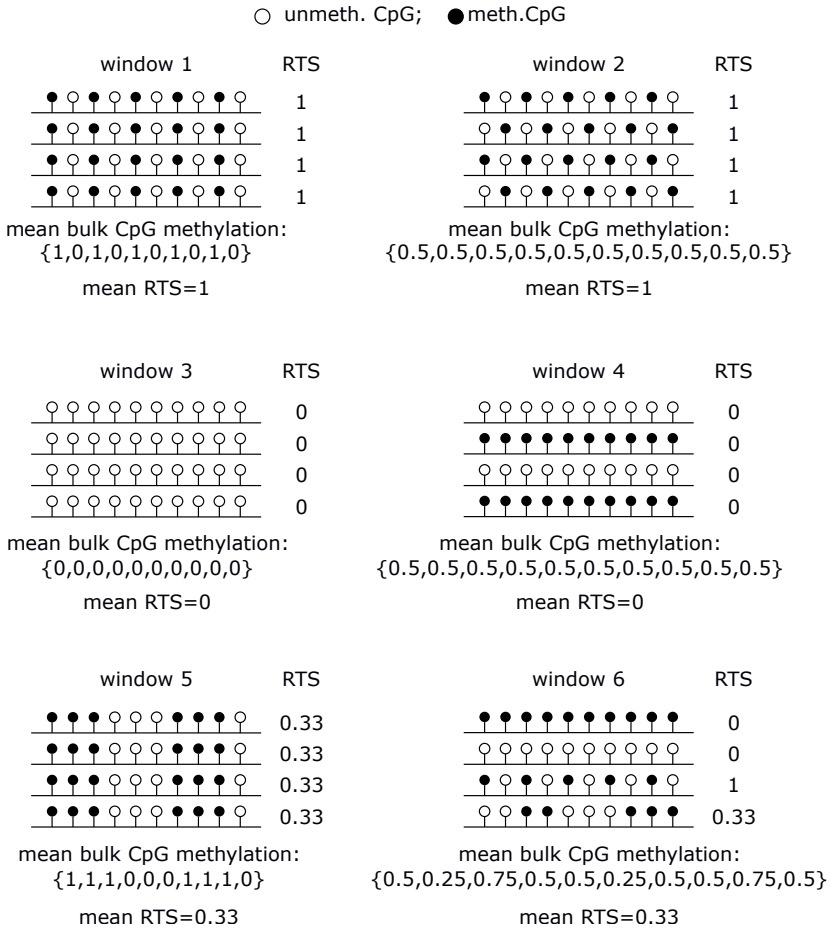

Fig. S3: Equivalent levels of intra-molecular heterogeneity can be associated with different levels of inter-molecular heterogeneity. Shown are six hypothetical examples of genomic windows composed of 10 CpG sites and covered by four reads. For each example, the RTS associated with each read is shown to the right of the read and the mean RTS for the window is shown below each example. The mean bulk methylation of each individual CpG is indicated below the reads. Each pair of examples: windows 1 and 2, windows 3 and 4, windows 5 and 6, show situations where the same mean window RTS is associated with differing degrees of inter-molecular heterogeneity.

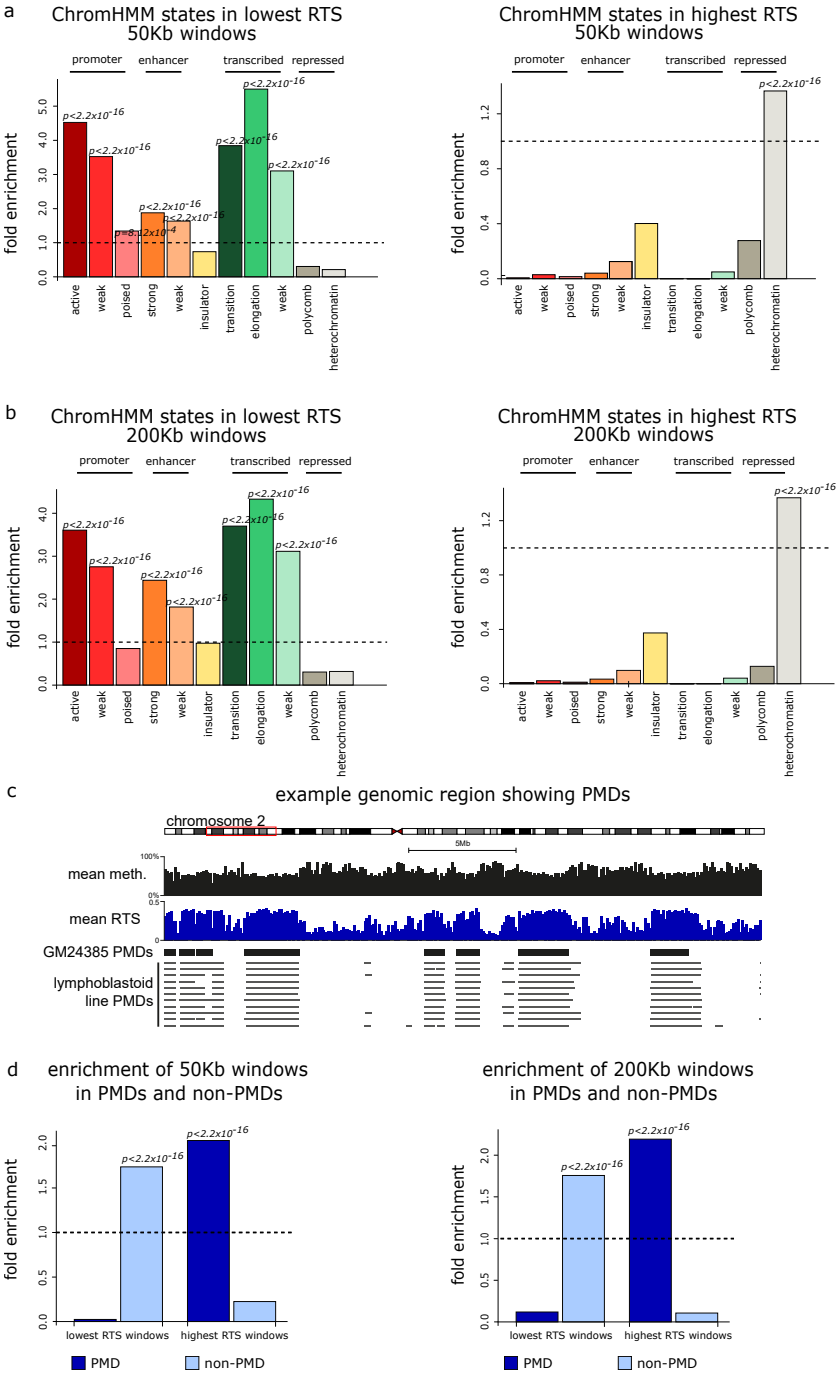
a
ChromHMM states in lowest RTS 50Kb windows
ChromHMM states in highest RTS 50Kb windows
promoter
enhancer
transcribed
repressed
fold enrichment
active
weak
poised
strong
weak
insulator
transition
elongation
weak
polycomb
heterochromatin
p<2.2x10^-16
p=8.2x10^-4
b
ChromHMM states in lowest RTS 200Kb windows
ChromHMM states in highest RTS 200Kb windows
c
example genomic region showing PMDs
chromosome 2
5Mb
mean meth.
mean RTS
GM24385 PMDs
lymphoblastoid line PMDs
d
enrichment of 50Kb windows in PMDs and non-PMDs
enrichment of 200Kb windows in PMDs and non-PMDs
lowest RTS windows
highest RTS windows
PMD
non-PMD

Fig. S4

Fig. S4: Heterogeneous genomic windows are enriched in heterochromatin and PMDs independently of the window size used. **a**: Barplots showing enrichments of the least and most heterogeneous 50kb windows in GM12878 chromHMM states. **b**: Barplots showing enrichment of the least and most heterogeneous 200kb windows in GM12878 chromHMM states. **c** Genome browser plot showing a representative genomic region with the mean DNA methylation and mean RTS in 100kb windows alongside GM24385 PMD locations and those previously identified in 11 other lymphoblastoid cell lines [25]. **d**: Barplots showing enrichment of the least and most heterogeneous 50kb (left) and 200kb windows (right) in PMDs and non-PMDs. Shown are significant p-values from Wilcoxon tests.

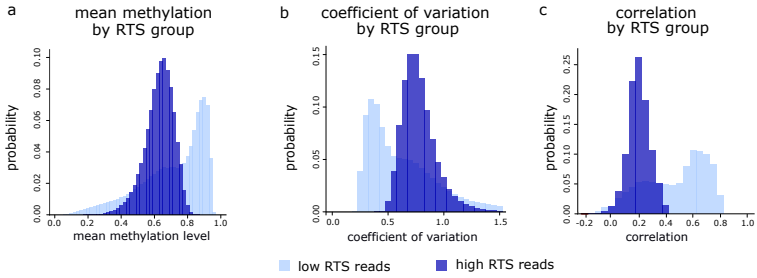

Fig. S5: Low and high RTS reads possess different statistical properties. Histograms showing the distribution of **a** mean methylation levels, **b** coefficients of variation and, **c** correlations between neighbouring sites associated with the $791,654$ low RTS reads and the $1,157,194$ high RTS reads.