Author/s:
Gigante, S;Gouil, Q;Lucattini, A;Keniry, A;Beck, T;Tinning, M;Gordon, L;Woodruff, C;Speed, TP;Blewitt, ME;Ritchie, ME

Title:
Using long-read sequencing to detect imprinted DNA methylation

Date:
2019-05-01

Citation:
Gigante, S., Gouil, Q., Lucattini, A., Keniry, A., Beck, T., Tinning, M., Gordon, L., Woodruff, C., Speed, T. P., Blewitt, M. E. & Ritchie, M. E. (2019). Using long-read sequencing to detect imprinted DNA methylation. Nucleic Acids Research, 47 (8), https://doi.org/10.1093/nar/gkz107.

Persistent Link:
https://hdl.handle.net/11343/250466

Published online 22 February 2019

Nucleic Acids Research, 2019, Vol. 47, No. 8 ***e46***
doi: 10.1093/nar/gkz107

Using long-read sequencing to detect imprinted DNA methylation

Scott Gigante[1,2,*,†], Quentin Gouil[1,3,†], Alexis Lucattini[4], Andrew Keniry[1,3], Tamara Beck[1], Matthew Tinning[4], Lavinia Gordon[4], Chris Woodruff[1], Terence P. Speed[1,5], Marnie E. Blewitt[1,3,‡] and Matthew E. Ritchie[1,3,5,*,‡]

[1]The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville VIC 3052, Australia, [2]Department of Genetics, Yale University, 333 Cedar Street, New Haven CT 06510, USA, [3]Department of Medical Biology, The University of Melbourne, Melbourne VIC 3010, Australia, [4]Australian Genome Research Facility, 305 Grattan Street, Melbourne VIC 3000, Australia and [5]School of Mathematics and Statistics, The University of Melbourne, Melbourne VIC 3010, Australia

Received October 21, 2018; Revised January 14, 2019; Editorial Decision February 05, 2019; Accepted February 08, 2019

ABSTRACT

Systematic variation in the methylation of cytosines at CpG sites plays a critical role in early development of humans and other mammals. Of particular interest are regions of differential methylation between parental alleles, as these often dictate monoallelic gene expression, resulting in parent of origin specific control of the embryonic transcriptome and subsequent development, in a phenomenon known as genomic imprinting. Using long-read nanopore sequencing we show that, with an average genomic coverage of ∼10, it is possible to determine both the level of methylation of CpG sites and the haplotype from which each read arises. The long-read property is exploited to characterize, using novel methods, both methylation and haplotype for reads that have reduced basecalling precision compared to Sanger sequencing. We validate the analysis both through comparison of nanopore-derived methylation patterns with those from Reduced Representation Bisulfite Sequencing data and through comparison with previously reported data. Our analysis successfully identifies known imprinting control regions (ICRs) as well as some novel differentially methylated regions which, due to their proximity to hitherto unknown monoallelically expressed genes, may represent new ICRs.

INTRODUCTION

Methylation of the fifth carbon of cytosines (5mC or simply mC) is an epigenetic modification essential for normal mammalian development. Methylation differences between alleles contribute to establish allele-specific expression patterns. As obtaining genome-wide haplotyped methylomes with short reads remains challenging, we evaluated the ability of long-read, nanopore-based sequencing to improve allele-specific methylation analyses.

We apply the technique to the study of genomic imprinting, where differential expression of the maternal and paternal alleles in the offspring is at least partially set by the differential methylation (1–5). Imprinting is proposed to arise from the diverging interests of the maternal and paternal genes (6). In accordance with its primordial role in allocation of resources from the mother to the offspring, the placenta, along with the brain, is the organ where parental conflict results in the most pronounced imprinted expression (7–9). We thus conduct a survey of differential methylation and expression in murine embryonic placenta.

Recent studies have increased the number of genes identified as subject to imprinting in mouse to ∼200 (10–15). The cause of the differential expression between paternal and maternal alleles is only known for a subset of these genes; maternal histone marks can play a role (14), and in other cases it involves the differential methylation of adjacent regions (5). The differential methylation patterns may be established in the gametes and persist through the epigenetic reprogramming occurring after fertilization (16).

*To whom correspondence should be addressed. Tel: +61 3 9345 2856; Fax: +61 9347 0852; Email: mritchie@wehi.edu.au
Correspondence may also be addressed to Scott Gigante. Email: scott.gigante@yale.edu
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
‡The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

These differentially methylated regions (DMRs) are called primary DMRs, or imprinting control regions (ICRs). Alternatively, differential methylation may arise during development, perhaps as a downstream effect of differential expression, in which case the regions are called somatic or secondary DMRs (17).

Apart from the parent of origin of the allele, genetic differences can also be associated with differential methylation. In this case, F1 hybrids of distinct mouse strains will display DMRs between the alleles according to the strain of origin (18), and not the parent. Genetically determined DMRs can have profound effects on phenotype, for instance in humans by altering the expression of mismatch repair genes important in cancer (19). Therefore, we also investigate the link between DNA methylation and expression for strain-biased genes.

Reconstructing haplotyped methylomes necessitates the simultaneous measurement of DNA methylation and single-nucleotide polymorphisms (SNPs) differentiating the alleles. This can be achieved by deep sequencing of bisulfite-converted DNA on the Illumina platforms, although the short reads combined with the reduced complexity of the bisulfite-treated DNA make the process inefficient, meaning many regions with low SNP density remain unresolved. Long reads provided by third generation sequencing technologies can overcome the requirement of a high SNP density, while several methods allow the assessment of base modifications on native DNA (thus also avoiding the reduction in complexity associated with bisulfite conversion). These methods include: analysis of polymerase kinetics for PacBio SMRT sequencing (20), and detection of deviations in the electric signal for Oxford Nanopore sequencing, via nanopolish (21), signalAlign (22), mCaller (bioRxiv doi:10.1101/127100), Tombo (bioRxiv doi:10.1101/094672) or DeepSignal (bioRxiv doi:10.1101/385849). We note that, for the dominant eukaryotic genome base modification at 5mC, the PacBio technology requires very high coverage making it impractical for use in the analysis of mammalian genomes (23). PacBio SMRT sequencing can be combined with bisulfite treatment (SMRT-BS) to facilitate 5mC detection, but this approach is currently only available for targeted sequencing (24) and the bisulfite treatment introduces the same drawbacks noted above in addition to fragmenting the DNA. Additionally, while PacBio technology is limited to maximum read lengths of between 50 and 100 kb (25), Oxford Nanopore sequencing has no theoretical upper limit on read length and exhibits no bias in sequencing quality with read length (26), which is especially beneficial in genomic regions devoid of SNPs, or highly repetitive regions.

Here, we use the MinION and PromethION long-read nanopore sequencers to generate whole-genome haplotyped methylomes from murine embryonic placenta. With a mean coverage of 10× we successfully identify known ICRs as well as novel parent-of-origin DMRs near imprinted genes, as well as strain-specific DMRs close to both strain-biased genes and structural variants. We show the improved efficiency of this strategy over existing workflows to resolve allele-specific methylation, and highlight its utility in investigating the mechanisms of genomic imprinting.

METHODS

Animal strains and husbandry

All mice were maintained and treated in accordance with Walter and Eliza Hall Institute Animal Ethics Committee approved protocols under approval number WEHI AEC 2014.026. *Mus musculus* castaneus mice were obtained from Jackson Labs. Note that due to prior inter-crossing for transgene transmission, the female *M. musculus* domesticus C57BL/6 mice that served as dams for our study comprise 12.5% FVB/NJ genome, however for simplicity we will refer to this mouse as B6. Wild-type B6 were reciprocally mated to wild-type *Mus musculus castaneus* (Cast).

DNA and RNA extraction

Pregnant females were sacrificed at E14.5 by CO_2 asphyxiation and the embryonic portion of each placenta was dissected from the maternal portion in phosphate buffered saline, as we have done previously (27). Samples were snap frozen in buffer RLT plus (Qiagen) and DNA and RNA were later extracted from the same sample using the AllPrep DNA/RNA Mini Kit (Qiagen), according to manufacturer's instructions. Samples were sexed by polymerase chain reaction (PCR) using primers for *Otc* (X-linked gene) and *Zfy* (Y-linked gene) as previously described (28), and male samples were selected for further analysis.

Illumina sequencing

Reduced Representation Bisulfite Sequencing (RRBS) libraries were made from 100 ng of DNA purified from the embryonic layer of a male B6 × Cast E14.5 placenta using the Ovation RRBS Methyl-Seq System (NuGEN), according to the manufacturer's recommendations, which include use of the Qiagen Epitect kit for bisulfite conversion. The resultant library was sequenced on a HiSeq 2500 (Illumina) using 100 bp paired-end reads and analysed as previously described (28,29).

RNA-seq libraries were prepared from 1 μg of RNA from four B6 × Cast and four Cast × B6 samples, including the same sample as the RRBS library, using the TruSeq RNA sample preparation kit (Illumina). 75-bp paired-end sequencing was performed on a NextSeq 500 (Illumina). Reads were trimmed with *Trim Galore* v0.4.2 and mapped with *HISAT2* v2.0.5 (30) with option `--no-softclip` to the GRCm38 (mm10) mouse genome with N-masked *castaneus* SNPs. Mapped reads were haplotyped with *SNPsplit* v0.3.2 (31), and gene counts obtained by running *featureCounts* (32) on the GRCm38_v90 Ensembl annotation. Differential analysis was performed with *edgeR* (33,34) using quasi-likelihood fits (35) and controlling the false discovery rate (FDR) at 10% (36). Interactive plots were produced with *Glimma* (37).

Nanopore sequencing

The B6 × Cast F1 sample was sequenced on three MinION flow cells with the 1D Sequencing Genomic Ligation (LSK108) protocol from ONT with minor adjustments:

4 μg of starting material were used for each library preparation, and for two libraries DNA was sheared to 10 kb with a Covaris G-Tube, whereas shearing was omitted for the third library (resulting in longer read lengths). Reads were basecalled with Albacore 1.2.2. The Cast × B6 F1 was sequenced on one PromethION flow cell with the 1D Sequencing Genomic Ligation (LSK109) protocol without shearing, and basecalled with Albacore 2.2.7. Nanopore reads were aligned to the same SNP-masked genome as before, using *BWA-MEM* (arXiv:1303.3997).

Haplotyping

Haplotyping is achieved through the identification of SNPs that are unique to one or the other allele. Examining only the SNPs identified as passing all filters in Keane *et al.* (38), we combine two distinct methods to confidently haplotype each read.

Basecall haplotyping. Where a read is aligned to a SNP position i on the reference genome, we assign a score S_i if the aligned base agrees with the reference haplotype, or $1 - S_i$ if the aligned base agrees with the alternate haplotype, where the score depends on the basecalling quality score q_i of the base in question as

$$S_i = 1 - e^{-0.6927-0.1203q_i},$$

where the co-efficients of the above relationship were determined empirically on successfully haplotyped reads. Bases which match neither haplotype, or which exhibit a deletion at the SNP location are excluded from the analysis. Finally, the read is assigned an aggregate haplotype value $h \in [0, 1]$ across the n informative SNP calls as follows:

$$h = \frac{1}{n}\sum_{i}^{n}\begin{cases} S_i, & \text{if base } i \text{ agrees with ref. allele;} \\ 1 - S_i, & \text{if base } i \text{ agrees with alt. allele.} \end{cases}$$

Signal-level haplotyping. For signal-level haplotyping, we use the hidden markov model (HMM) of Simpson *et al.*, implemented in *nanopolish phase-reads* (39). Briefly, the raw signal corresponding to the section of the read aligned to the reference at the SNP position is realigned using a HMM, and the likelihood of the sequence of 6-mers in this vicinity is maximized by choosing the more likely of the two possible alleles. Each read is then assigned scores according to the same rule as in *Basecall Haplotyping*, where *nanopolish* quality scores are offset by −35 in order to exhibit a similar relationship to basecall quality scores.

Combining haplotype calls. For each read with n_{base} and n_{signal} associated SNP calls and associated haplotype values h_{base} and h_{signal}, we define the haplotype calls

$$H_{\text{base}} = \text{sgn}\,(h_{\text{base}} - 0.5) \text{ and}$$
$$H_{\text{signal}} = \text{sgn}\,(h_{\text{signal}} - 0.5).$$

The calls are then combined according to the following rules, applied in order:

$$H = \begin{cases} 0, & \text{if } n_{\text{base}} < 5 \text{ and } n_{\text{signal}} < 5; & \textbf{(1)} \\ H_{\text{base}}, & \text{if } H_{\text{base}} = H_{\text{signal}}; & \textbf{(2)} \\ H_{\text{base}}, & \text{if } n_{\text{base}} > 3n_{\text{signal}} \text{ and not (1);} & \textbf{(3)} \\ H_{\text{signal}}, & \text{if } n_{\text{signal}} > 3n_{\text{base}} \text{ and not (1);} & \textbf{(4)} \\ H_{\text{base}}, & \text{if } |h_{\text{base}} - 0.5| > 3|h_{\text{signal}} - 0.5| & \textbf{(5)} \\ & \quad \text{and not (1), (3) or (4);} & \\ H_{\text{signal}}, & \text{if } |h_{\text{signal}} - 0.5| > 3|h_{\text{base}} - 0.5| & \textbf{(6)} \\ & \quad \text{and not (1), (3) or (4);} & \\ 0, & \text{otherwise.} & \textbf{(7)} \end{cases}$$

where $H = 1$ represents a read assigned to the reference haplotype, $H = -1$ represents a read assigned to the alternate haplotype and $H = 0$ represents an unassigned read. This process is shown graphically as a flowchart in Supplementary Figure S3.

Resolution of maternal recombination. Owing to the cross of an FVB-strain into the maternal line in the grandparental generation, it is necessary to resolve which section of the maternal genome was contributed by recombination from the FVB chromosome. We run the above haplotyping procedure with three possible outcomes, rather than two: mm10, FVB and CAST, with variants called by Keane *et al.* (38). The proportion of maternal (non-CAST) reads within any 100 Kb region was fitted to a recursive partition tree, which splits continuous data into a stepwise function, here representing the proportion of a contiguous section of chromosome haplotyped to FVB (Supplementary Figure S5). Fitting was performed using the R package *rpart* with parameters `minsplit=5` and `cp=0.1` (40). SNPs in sections of the chromosome with mean proportion of FVB >50% were replaced with the FVB allele for further analysis.

Methylation calling

We determined the methylation status of each CpG site on each read using *nanopolish call-methylation* (21). Briefly, *nanopolish* uses a 5-base alphabet, with 5-methylcytosine represented as M, to build a Gaussian mixture model representing every possible 6-mer with both methylated and unmethylated cytosine in a CpG context, excluding those 6-mers which contain both the methylated and unmethylated base. We ran *nanopolish* separately on reads haplotyped to the maternal and paternal chromosome, using a SNP-masked version of each chromosome to decrease bias in reads with expected deviations from the mm10 reference.

Nanopolish then assigns each section or 'event' of nanopore current to a base on the reference genome and calculates the likelihood of each 6-mer containing the CpG site being either methylated ($L_M(d_{ij})$) or unmethylated ($L_C(d_{ij})$) given the data d_{ij} for a call group i covered by a read j. Groups of consecutive CpG sites in which the distance between any two adjacent sites is <11 bases (therefore having overlap between 6-mers containing the cytosines in question) are chained into *CpG call groups*. All sites within the one CpG call group are assumed to have the same methylation status, such that each 6-mer is only considered once.

We convert these likelihoods to probabilities as follows:

$$L_M(d_{ij}) = P(d_{ij}|M) \text{ and } L_C(d_{ij}) = P(d_{ij}|C)$$

By Bayes' law,

$$\frac{P(M|d_{ij})}{P(C|d_{ij})} = \frac{P(d_{ij}|M)P(M)}{P(d_{ij}|C)P(C)}$$

and since M and C are mutually exclusive and jointly exhaustive,

$$P(C|d_{ij}) = 1 - P(M|d_{ij}).$$

Then, defining the prior probability of methylation as $P(M) = p_0$,

$$\frac{P(M|d_{ij})}{1 - P(M|d_{ij})} = \frac{p_0}{1 - p_0}\frac{L_M(d_{ij})}{L_C(d_{ij})}$$

and rearranging for $P(M|d_{ij})$,

$$P(M|d_{ij}) = \frac{1}{1 + \frac{1 - p_0}{p_0}\frac{L_C(d_{ij})}{L_M(d_{ij})}}.$$

Noting results from Decato *et al.* (41) showing methylation levels ranging from 0.433 to 0.538 for mouse placental tissue we set $p_0 = 0.5$, so finally we define the single-read, single-site probability of methylation as

$$\beta_{ij} = P(M|d_{ij}) = \frac{1}{1 + \frac{L_C(d_{ij})}{L_M(d_{ij})}}.$$

Comparison with RRBS methylation calls. Individual methylation calls on a single CpG call group are aggregated over the set of reads covering each group in order to compare with aggregate values provided by bisulfite sequencing. That is, for each CpG call group i covered by n reads, we define the call group average

$$\beta_i = \frac{1}{n}\sum_{j=1}^{n}\beta_{ij}.$$

In order to compare methylation calls between nanopore and RRBS, we must split CpG call groups defined by *nanopolish* as CpG sites separated by <11 base pairs into individual sites, including GpC sites on the reverse strand, with each site retaining the same β value as the original call group. Only those CpG sites for which both RRBS and nanopore data exist are considered.

Identification of differentially methylated regions

Following methylation detection and haplotype assignment of each read, it is possible to assign each call of methylation on the genome to one of the two haplotypes. The aggregated β methylation values for each CpG group are tested for DMRs using the DSS software (42). Briefly, DSS tests for differential methylation at single CpG-sites, using a Wald test on the co-efficients of a beta-binomial regression of count data with an 'arcsine' link function. Then, using a default P-value threshold of 10^{-5}, DSS aggregates differentially methylated sites into DMRs based on a maximum separation between sites and a minimum density and number of sites in each DMR.

To detect parent-of-origin DMRs, we perform DSS with the comparison B6♀ and Cast♀ versus Cast♂ and B6♂; to detect strain-specific DMRs, we perform DSS a second time with the comparison B6♀ and B6♂ versus Cast♂ and Cast♀.

Visualization of haplotyped methylation

Owing to the noisy nature of nanopore methylation calls, we use a loess smoothing curve to visually represent the methylation of a single nanopore read (43). Here, the smoothing parameter α is determined by

$$\alpha = 0.1 + 8 \cdot 10^{-11}(\max\{10^5 - L, 0\})^2,$$

where L is read length. This relationship was determined empirically to have minimal impact on visualization while minimizing computation time. Genomic tracks of methylation and expression were plotted with the following Bioconductor packages: *ggbio* (44), *rtracklayer* (45) and *GenomicRanges* (46).

RESULTS

Nanopore methylation calls are concordant with other technologies

We sequenced the embryonic portion of placenta derived from a male embryonic day 14.5 (E14.5) conceptus from a C57BL/6 (Black6, or B6) × *Castaneus* (Cast) F1 on the MinION platform to a depth of ∼8× (Supplementary Figure S1) and called methylation using *nanopolish* (21). The genome-wide methylation data successfully recapitulated known patterns: CpG islands (CGIs), as defined by Irizarry *et al.* (47), separated into two groups of high and low methylation (Figure 1A); the methylation level dipped at transcriptional start sites (TSSs) (Figure 1B), and the average genome-wide methylation level was ∼50%, as previously reported for placental tissue (41).

To further validate the accuracy of the nanopore methylation calls, we compared them to RRBS data on the same sample at sites covered by both methods. Nanopore methylation calls showed an overall similar distribution to RRBS methylation calls, albeit with a bias toward intermediate values of methylation (Figure 1C). Despite being less correlated than measurements from methylation-specific technologies (48) with a median absolute deviation of 0.18, per-site methylation was also relatively concordant between the two methods, with 85% of CpG sites being called correctly when converting average methylation values to binary calls (Figure 1D).

Sequencing of E14.5 embryonic placenta from the reciprocal cross (Cast × B6) on the PromethION platform at 12× coverage (Supplementary Figure S1) generated comparable results.

Increased haplotyping efficiency with nanopore reads

We next used high-confidence SNPs between the Cast and B6 strains to haplotype RRBS and nanopore reads. In order to mitigate the high sequencing error rates of nanopore

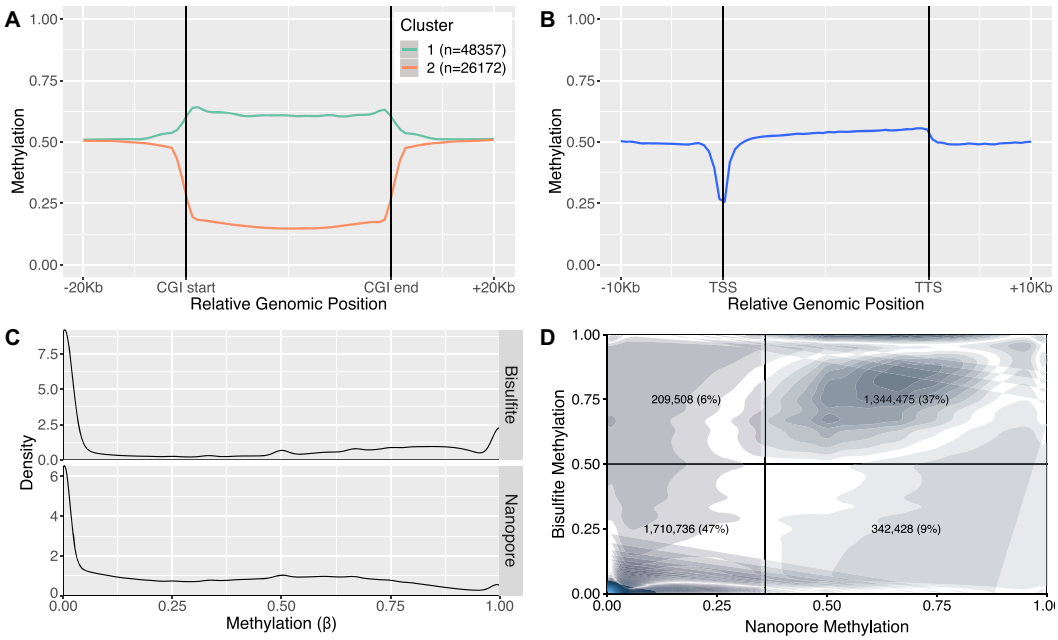

Figure 1. Nanopore methylation calls are consistent with expected results and established technologies. (**A**) Metaplot of nanopore methylation calls across CGIs, clustered in two groups of high and low methylation. (**B**) Metaplot of nanopore methylation calls across the aggregated gene bodies of all protein-coding genes recapitulated known methylation structures. (**C**) Density of methylation calls (β, the average methylation based on all reads covering that position) for sites covered by both nanopore and RRBS. (**D**) Joint density of nanopore and RRBS methylation calls for the same sites as in panel (C). Darker regions indicate regions of higher density, while lighter regions indicate regions of lower density. The density plot is split into four quadrants according to a RRBS threshold of 0.5 and a nanopore threshold of 0.36, and the percentage of sites in each quadrant is displayed.

sequencing, we used two methods of haplotype assignment, denoted the '*basecall method*', based on FASTQ data, and the '*signal method*', based on the *phase-reads* module from *nanopolish*, an HMM, which uses the raw nanopore signal to predict genotype (39). Additionally, we only assigned a haplotype to those nanopore reads with at least five high-confidence SNPs (Supplementary Figure S3). All RRBS reads overlapping at least one SNP were assigned a haplotype (31). Only 24% of the mapped RRBS reads could be assigned to one haplotype, whereas 74% of mapped nanopore reads were haplotyped in the expected proportions (Figure 2A and B): roughly half of the haplotyped reads were assigned to the maternal haplotype, and half to the paternal haplotype, albeit showing a slight bias toward the paternal haplotype (due to an increased number of split reads in regions where sections of the Cast genome has a deletion with respect to the B6 genome). The pattern of haplotype assignment was consistent across the autosomal chromosomes, while, as expected for a male sample, almost all (91%) of the reads aligned to the X chromosome were assigned to the maternal haplotype (Figure 2B). Haplotyping of the Cast × B6 cross gave similar results (Supplementary Figure S2A). The lack of maternal bias in read haplotype indicates minimal maternal contamination, which is also reflected in consistent RNA-seq library sizes (Supplementary Figure S4).

We further evaluated the accuracy of the haplotyping of the nanopore reads by sequencing the same tissue from the parents (B6 only, and Cast only). Following the same haplotyping procedure, 85.7% of the reads were correctly assigned to the relevant genotype, and 1.5% were misassigned (Supplementary Figure S2B and C).

Table 1. Accuracy and support of haplotyping methods on pure-strain reads

Method	AUROC	Accuracy (%)	Called reads
Basecall	0.976	0.963	**261 806**
Signal	0.988	0.961	199 474
Regression	**0.992**	0.972	198 684
Ad hoc	–	**0.983**	228 866

The large majority of nanopore reads showed strong agreement between the two haplotyping methods. Discrepancy between the basecall and signal methods are typically due to a low number of SNPs being scored by one or both methods, resulting in these alignments being filtered out by the haplotyping procedure (Supplementary Figure S3). However, when examining the overall predictive performance of the two methods with Area Under Receiver Operating Characteristic Curve (AUROC) on the single-strain experiments, the signal method marginally outperformed the basecall method (see Table 1). We also found that the combination of the two methods achieved a slight improvement again over the signal method, either with a logistic regression model (49) or an *ad hoc* combination of the two approaches (see 'Materials and Methods' section). The *ad hoc* method allowed classification of 30 000 additional reads over the logistic regression and signal method, both of which excluded reads for which *nanopolish* failed to produce output.

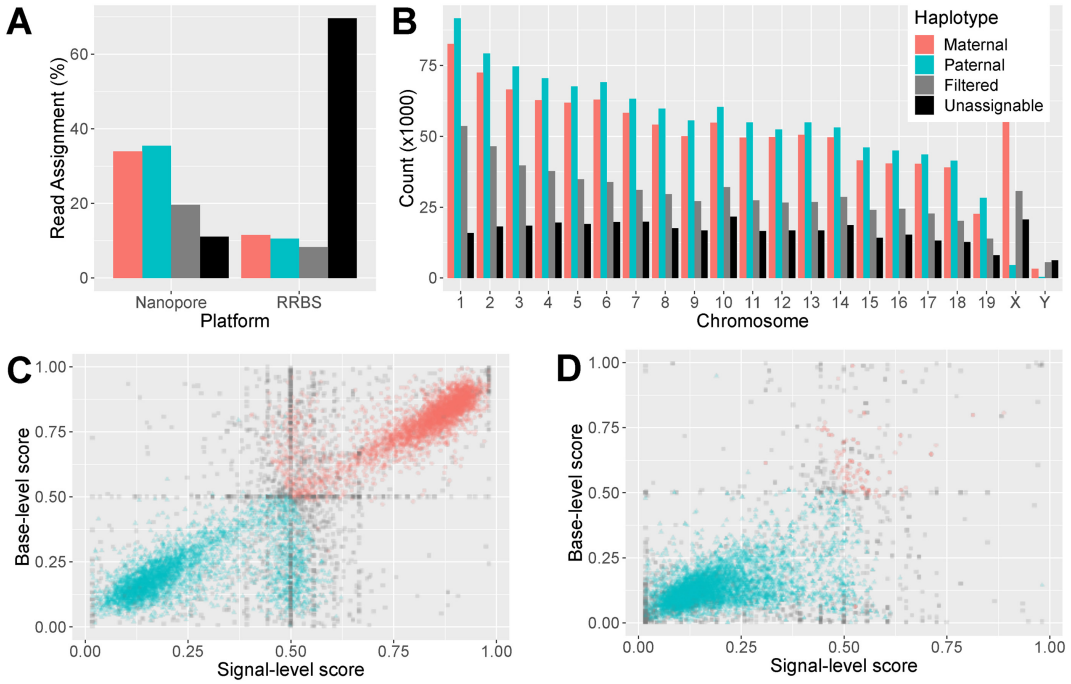

Figure 2. Accurate and efficient haplotyping of nanopore reads. (**A**) Percentages of mapped reads from RRBS and nanopore sequencing that were assigned to the B6 genome (maternal), Cast genome (paternal), or that could not be haplotyped (filtered) for the B6 × Cast F1 sample. (**B**) Percentages of mapped reads from nanopore sequencing that were assigned to each haplotype on each chromosome. (**C**) Scatter plot of haplotype scores for nanopore reads according to signal (*x*-axis) and basecall (*y*-axis) methods. Only 10 000 randomly selected reads are shown for ease of visualization. (**D**) Signal and basecall haplotype scores for reads from the sequencing of the pure parental Cast strain.

In both nanopore and RRBS data, the main cause of haplotyping failure is the lack of SNPs in the region covered by the read (Supplementary Figure S3). While the proportion of successfully haplotyped nanopore reads could increase with optimization for longer reads and anticipated improved sequencing accuracy, RRBS haplotyping efficiency is limited by the short read length.

Parent-of-origin and strain-biased gene expression

To investigate the correlation between differential methylation and differential gene expression, we performed RNA-seq on the same F1 placental tissue from reciprocal crosses of B6 and Cast, in quadruplicates. Maternal tissue contamination was unlikely as for each embryo maternal and paternal counts were similar (Supplementary Figure S4). We found 135 genes with a parental bias in expression (imprinted genes, 10% FDR, Figure 3A): 88 with higher expression from the maternal allele and 47 with higher expression of the paternal allele. Among the 135 genes, 53 corresponded to well-characterized imprinted genes in classic databases (50–53). A further 17 of these genes, including *Fkbp6*, *Smoc1/2*, *Gzmc/d/e/f/g*, *Zdhhc14* and *Arid1b* have been identified as imprinted in one or several recent studies (12–15). The remaining 65 genes constitute novel candidate genes with parent-biased expression in mouse placenta. The complete annotated list is reported in Additional File 1.

We also identified 4 029 genes (13% of expressed autosomal genes) with a strain bias >2-fold (5% FDR, Figure 3B and Additional File 1), evenly split between B6 dominance (2 027 genes) and Cast dominance (2 002).

Known imprinting control regions are observable by nanopore sequencing

We combined the methylation and haplotyping data from the nanopore reads to compare methylation between the parental alleles. To highlight the linkage information of the methylation data available for nanopore reads, as well as the per-site per-read data, we plotted the loess fit of the cytosine methylation levels for each read in the region of interest (Figure 4A and B).

DMRs at known ICRs (52) were readily visible and concordant with matched allele-specific RNA-seq and RRBS data (Figure 4A and B). Nanopore data recapitulated methylation differences at most known ICRs (Figure 4C), often showing extended differential methylation past the annotated ICR borders.

Nanopore sequencing reveals novel differentially methylated regions

Next, we sought to define DMRs between parental alleles as well as between strains *de novo*, using the differential methylation tool DSS (42). We ranked putative DMRs based on the area statistic. Using the DSS default threshold of 10^{-5}, we obtained a total of 933 DMRs, of which 309 were explained by parent-of-origin differences, and the remainder by strain-specific effects (Additional File 2). We then examined these DMRs, in conjunction with haplotyped RRBS and RNA sequencing data for corroborating evidence of differential methylation and differential expression, respectively, in order to find putative DMRs of interest at imprinted genes.

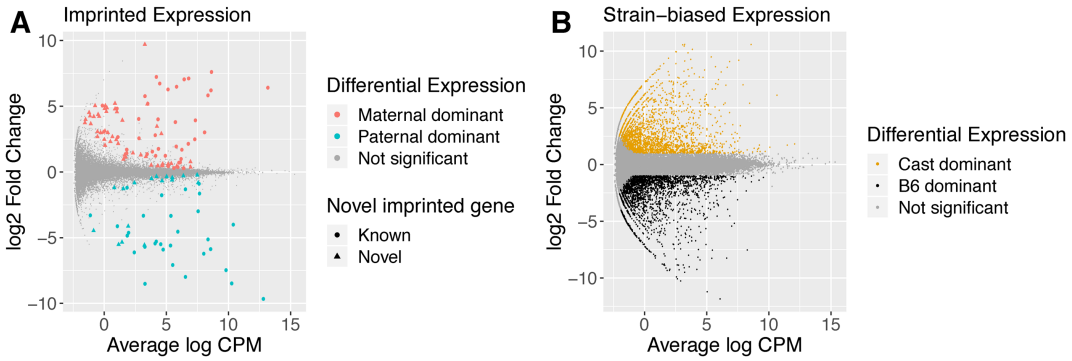

Figure 3. Differential allelic expression in mouse E14.5 embryonic placenta. (**A**) Differential expression between the maternal and paternal alleles. Genes with adjusted *P*-value < 0.1 are coloured in red when maternal expression dominates (positive log-fold change) and blue when paternal expression is greater (negative log-fold change). The shape of the point indicates whether the differentially expressed gene has previously been reported as imprinted. (**B**) Differential expression between B6 and Cast alleles. Genes with adjusted *P*-value < 0.05 and absolute $\log_2$ fold-change > 1 are coloured in black when B6 expression is higher and orange when Cast expression is higher. Interactive plots are available at bioinf.wehi.edu.au/haplotyped_methylome.

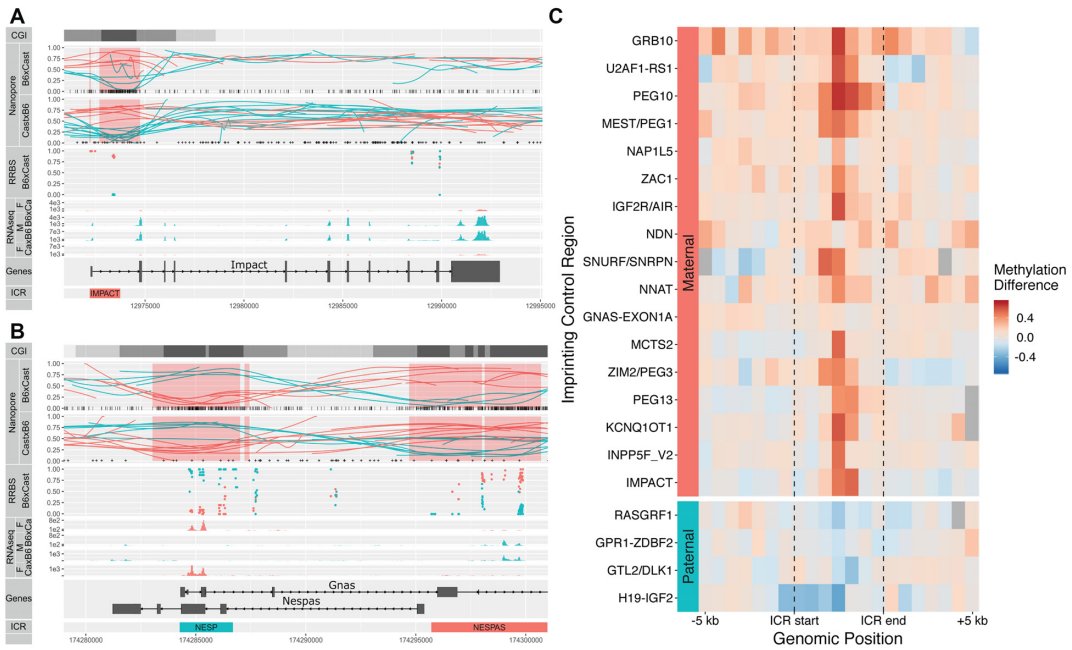

Figure 4. Nanopore allele-specific methylation captures known differential methylation at ICRs. (**A**) Allelic methylation plot of maternally imprinted gene *Impact* displays a clear DMR at its ICR. Haplotyped RRBS data shows concordance with nanopore allelic methylation. Allele-specific RNA-seq coverage plots show monoallelic paternal expression. CGIs are displayed in black, with CpG shores in dark grey and CpG Shelves in light grey. *Nanopore*: Vertical bars at the base of the *B6Cast* track denote CpG sites used for methylation calling, while '+' signs at the base of the *CastB6* track denote SNPs used for haplotyping. Highlighted red regions indicate DMRs detected by DSS. The maternal allele is shown in red and the paternal allele in cyan for all plots. (**B**) Allelic methylation plot as in **A** for the reciprocally imprinted genes *Nespas* and *Gnas*. RNA-seq gives very low expression and is not shown. (**C**) Heatmap of differences (maternal − paternal) in allelic methylation in relative-width bins along known ICRs. Regions are sorted in order of average methylation difference, with regions in the same imprinting cluster placed adjacent to each other. Regions without haplotyped calls for both alleles are shown in grey.

Of the 20 highest ranking DMRs, 15 corresponded to known ICRs. Although many of the lower-ranking DMRs are potential false-positives, they also included regions of known imprinted expression (for instance two small detected DMRs immediately adjacent to known DMRs at the IMPACT and NESP ICRs, shown in Figure 4A and B, respectively.) Thus in the absence of statistically robust DMR-finding methods for nanopore data, we kept this permissive threshold.

Five ICRs annotated in the WAMIDEX database (52) were not detected *de novo* (Table 2). INPP5F_V2 and GRB10 simply lacked coverage in the B6 × Cast sample but showed clear differential allelic methylation in the Cast × B6 sample; GNAS-EXON1A also lacked coverage in B6 × Cast but the reciprocal sample and the RRBS did not suggest differential methylation, while NDN lacked coverage in both samples. The last undetected region was GPR1-ZBDF2; however Duffié et al. (17) have shown that this region lacks important features of a bona fide ICR, and that a neighbouring maternally hypermethylated region is the true ICR. The region in question was readily detected as differ-

entially methylated from the nanopore data (DMR #229, Table 2).

In addition to ICRs, we detected numerous DMRs at imprinted genes that do not appear to be present in gametes (54–56), and which therefore likely constitute secondary DMRs (Table 2). When RRBS coverage was present, the bisulfite data corroborated the *de novo* DMR identification. Five of the secondary DMRs have been described previously, although they are not currently compiled in a database: maternal hypermethylation at the *Igf2* promoter (57), maternal hypermethylation at the placental-specific promoter of *Gab1* (58), paternal hypermethylation of the *Meg3* TSS (59), maternal hypermethylation at the *Slc38a4* TSS (60) and paternal hypermethylation at the *Igf2r* promoter (2).

The remaining secondary DMRs have not been previously characterized. Six of them overlapped the TSSs of imprinted genes: *Sfmbt2*, *Jade1*, *Ascl2*, *Cd81/R74862*, *Tssc4* and *AC158554.1*.

Other novel secondary DMRs overlapped introns rather than TSSs. *Park2*, a recently identified maternally-biased gene (13), had seven intronic DMRs, all displaying hypermethylation of the maternal allele. *Rian* displayed a DMR that had not been previously reported in mice, although its human orthologue also presents an intronic imprinted DMR (61).

In some cases, inspection of the parent-specific DMRs revealed unannotated imprinted transcription nearby, for example in the *Kcnq1* and *Igf2r* clusters (Table 2). These RNA-seq reads may be part of imprinted long non-coding RNAs, frequent at imprinted clusters.

While we have collated all the imprinted DMRs that we found to directly overlap with imprinted expression - in addition to the WAMIDEX ICRs- in Table 2, we note that other imprinted DMRs may be associated with the imprinted expression of more distant genes, or with genes that are only expressed or imprinted in specific tissues. For example, we found imprinted DMRs in the promoters of *Smoc2* (DMR #224) and *Arid1b* (DMR #863), two genes recently identified as being imprinted (and also imprinted in our data). The strong DMR #110 overlapped the TSS of *Gtsf2*, which was poorly expressed in placenta but may be imprinted in the tissues where it is expressed (in gonocytes and spermatids (62)). In the absence of chromatin conformation data or functional validation, we did not attempt to formally assign these DMRs to specific genes.

We however used the top, most reliable 400 DMRs to calculate the distance of genes to their nearest DMR depending on their expression status. We found that parentally biased genes were more likely to be proximal to parent-of-origin DMRs than unbiased genes (median distance 0.9 Mbp compared to 7.4 Mbp), whereas strain-biased genes and DMRs did not show this relationship (median distance 2.9 Mbp compared to 3.1 Mbp) The distributions of distances to the nearest DMR are shown in Figure 5, which shows a striking relationship between parentally biased genes and parent-of-origin DMRs. This result is consistent with parental bias in expression being driven necessarily by epigenetic differences, whereas differential expression between strains is mainly driven by genetic differences.

Long reads provide advantages in differential methylation analysis

Inspection of the DMRs revealed multiple advantages of our nanopore-based method of methylome haplotyping over traditional bisulfite sequencing (Figure 6).

We were able to resolve DMRs in regions of low SNP density, where there were no haplotyped RRBS reads despite the presence of a CGI (Figure 6A). The particular DMR in Figure 6A encompassed the TSS of the imprinted gene *Peg10*, and was much wider than the previously annotated ICR. The increased DMR width was a regular occurrence at ICRs (Table 2).

Our method also uncovered novel secondary DMRs at known imprinted genes such as *Jade1* (Figure 6B), as well as at previously uncharacterized imprinted transcripts such as *AC158554.1*, annotated as a long intergenic noncoding RNA (*ENSMUSG00000116295*, Figure 6C).

Another advantage provided by the long reads was apparent at the ZIM2-PEG3 ICR (Figure 6D). RRBS data from the B6Cast F1 showed that certain CG dinucleotides were highly methylated on the maternal allele (100% methylation at these positions) while others were variably methylated, resulting in averages of 25–50% methylation. Two scenarios could give rise to these intermediate values: either the variable positions are randomly unmethylated in all maternal alleles, or there exists two populations of maternal alleles, one where CG dinucleotides are methylated throughout the region and one where the variable positions are consistently unmethylated. The long nanopore reads revealed that the second scenario contributes to the observed intermediate methylation patterns: there was a mixture of cells, in some of which the maternal allele showed a contiguous loss of methylation. Although this result is well known to those who have practiced Sanger bisulfite sequencing, the haplotyped methylome derived from nanopore sequencing allowed investigation of this variability more accurately (no PCR bias) and across the whole genome.

Eight strain-specific DMRs were also found within 5 kb of a gene with strain-biased expression (Figure 6E). Most of these exhibit structural variation proximal to the DMR, although a small number exhibited changes in expression seemingly not associated with any structural variant.

One example of a structural variant between B6 and Cast associated with differential methylation can be found on Figure 6F. Upstream of the *Tap2* gene, an IAPEZ retrotransposon in the B6 genome is absent from the Cast genome (Cast reads are truncated upstream of the repeat and absent downstream), and the gene-proximal border is more highly methylated on B6 alleles than on Cast alleles. This differential methylation would be consistent with the insertion of the transposon attracting methylation that spreads to adjacent regions. We could also see at this repeat region a lot of spuriously mapping reads from both strains, suggesting that the repeat is present in multiple copies that the current assembly fails to account for.

DISCUSSION

Determining allele-specific methylation patterns in diploid or polyploid cells with short-read sequencing is hampered by the dependence on a high SNP density and the reduction

Table 2. List of known and proposed DMRs associated with imprinted genes

DMR name	Known co-ordinates	Present co-ordinates	Hypermethylated	Nearest gene	Nanopore	RRBS	RNA-seq	Note
GPR1-ZDBF2	1:63,257,407-63,264,876		p	*Zdbf2*,*Gpr1*				several transient DMRs (17)
GPR1-PLATR12/LIZ		1:63,200,250-63,200,470	m	*Zdbf2*,*Gpr1*	229	✓	✓	*Gpr1* ICR (17)
SFMBT2		2:10,371,327-10,371,731	m	*Sfmbt2*,*Gm13261*	75	✓	✓	*Sfmbt2* TSS secondary DMR
MCTS2	2:152,686,755:152,687,275	2:152,686,261-152,687,856	m	*Mcts2*	13	✓	✓	wider than annotation
NNAT	2:157,560,050-157,561,662	2:157,559,825-157,561,802	m	*Nnat*	29	✓	✓	wider than annotation
NESP	2:174,284,269-174,286,690	2:174,283,034-174,287,439	p	*Nespas*	1	✓	✓	
NESPAS	2:174,295,707-174,300,901	2:174,294,696-174,300,693	m	*Gnas*	6	✓	✓	
GNAS-EXON1A	2:174,326,930-174,329,007		m	*Gnas*			✓	
JADE1		3:41,555,359-41,556,940	m	*Jade1*	19	✓	✓	*Jade1* TSS secondary DMR
PEG10	6:4,747,209-4,747,507	6:4,746,012-4,749,480	m	*Peg10*	2		✓	wider than annotation
MEST	6:30,736,488-30,737,237	6:30,735,330-30,739,552	m	*Mest*	4	✓	✓	
NAP1L5	6:58,906,696-58,907,062	6:58,906,821-58,907,095	m	*Nap1l5*,*Herc3*	89			
NDN	7:62,348,214-62,348,695		m	*Ndn*			✓	low coverage
ZIM2	7:6,727,576-6,732,116	7:6,727,344-6,731,296	m	*Peg3*	5	✓	✓	
SNURF/SNRPN	7:60,004,992-60,005,415	7:60,003,140-60,005,295	m	*Snrpn*,*Snurf*	16;343	✓	✓	wider than annotation
INPP5F_V2	7:128,688,274-128,688,642		m	*Inpp5f*		✓		lack of coverage in B6Cast, clear DMR in CastB6
H19/IGF2	7:142,580,263-142,582,140	7:142,575,503-142,582,086	p	*H19*	17;35;54,534	✓	✓	wider than annotation
IGF2-DMR0		7:142,669,246-142,670,067	m	*Igf2*,*Igf2os*,*Gm49394*	39	✓	✓	known placenta-specific secondary DMR (57)
ASCL2		7:142,968,946-142,969,300	p	*Ascl2*	62	✓	✓	*Ascl2* TSS secondary DMR
CD81		7:143,052,956-143,053,090	p	*Cd81*,*R74862*	463;887	✓	✓	*Cd81*/*R74862* TSS secondary DMR
TSSC4		7:143,068,896-143,069,197	p	*Tssc4*,*Trpm5*,*Cd81*	153	✓	✓	*Tssc4* TSS secondary DMR
KCNQ1OT1	7:143,295,155-143,295,622	7:143,294,879-143,296,757	m	*Kcnq1ot1*	11	✓	✓	wider than annotation
KCNQ1-INTERGENIC1		7:143,438,058-143,438,341	p		205	✓	✓	secondary DMR with unannotated imprinted expression
KCNQ1-INTERGENIC2		7:143,445,526-143,445,944	p	*Gm27901*	67	✓	✓	secondary DMR with unannotated imprinted expression
CDKN1C		7:143,459,775-143,459,891	p	*Cdkn1c*,*Gm4732*	355	✓	✓	*Cdkn1c* secondary DMR
GAB1		8:80,859,569-80,859,745	m	*Gab1*	221	✓	✓	known *Gab1* placental TSS secondary DMR (58)
RASGRF1	9:89,879,568-89,879,853	9:89,879,601-89,880,045	p	*Rasgrf*	69			
ZAC1	10:13,090,470-13,091,527	10:13,090,313-13,092,161	m	*Plagl1*	12		✓	wider than annotation
U2AF1-RS1	11:22,971,842-22,972,319	11:22,971,545-22,973,999	m	*Zrsr1*	3	✓	✓	wider than annotation
GRB10	11:12,025,482-12,026,332		m	*Grb10*		✓	✓	lack of coverage in B6Cast, clear DMR in CastB6

Table 2. Continued

DMR name	Known co-ordinates	Present co-ordinates	Hypermethylated	Nearest gene	Nanopore	RRBS	RNA-seq	Note
GTL2/DLK1	12:109,526,740-109,528,734	12:109,527,519-109,528,845	p	*Gm27528,Gm27528*	52	✓	✓	narrower than annotation
MEG3		12:109,540,792-109,541,676	p	*Meg3,Mir1906-1,Gm27300,Gm27596*	207;776	✓	✓	known *Meg3* TSS secondary DMR (59)
MEG3-INTRON		12:109,556,071-109,556,162	m	*Meg3*	921		✓	*Meg3* intronic secondary DMR
RIAN		12:109,612,8804-109,612,962	m	*Rian,Mir1188, Mir341*	304		✓	*Rian* intronic secondary DMR
PEG13	15:72,806,335-72,811,649	15:72,809,183-72,811,180	m	*Peg13*	14	✓	✓	narrower than annotation
SLC38A4		15:97,053,880-97,056,427	m	*Slc38a4*	9	✓	✓	known TSS secondary DMR (60)
AC158554.1		15:97,166,956-97,167,257	m	*AC158554.1*	403;406	✓	✓	lincRNA TSS secondary DMR
PDE10A		17:8,772,760-8,773,118	m	*Pde10a*	474;701		✓	*Pde10a* intronic secondary DMR
PARK2		17:11,123,807-11,124,219	m	*Park2*	222;289;488;597; 782;827;852		✓	seven *Park2* intronic secondary DMRs
SLC22A2		17:12,607,783-12,608,088	m	*Slc22a2*	452		✓	*Scl22a2* intronic secondary DMR
IGF2R/AIR	17:12,741,297-12,742,707	17:12,741,160-12,742,949	m	*Igf2r,Airn*	8	✓	✓	
IGF2R-TSS		17:12,769,605-12,770,120	p	*Igf2r,Airn,Gm23833*	37	✓	✓	known *Igf2r* TSS secondary DMR (2)
SMOC2-INTERGENIC		17:14,590,437-14,590,640	m	*Smoc2*, *Thbs2*	862		✓	secondary DMR with unannotated imprinted expression
IMPACT	18:12,972,197-12,973,741	18:12,972,182-12,974,748	m	*Impact*	7	✓	✓	wider than annotation

A check mark is shown where RRBS and RNA-seq evidence supports the DMR. The DMR rank is shown when supporting nanopore data exists.

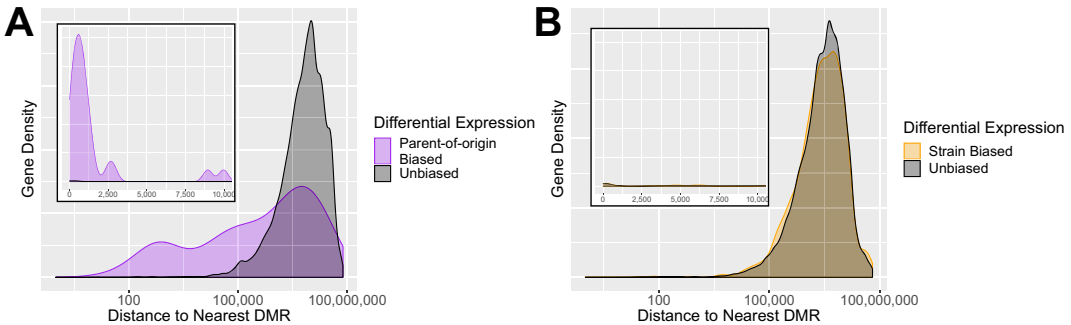

Figure 5. Proximity of differentially expressed genes to DMRs. (**A**) Distribution of distance from genes to imprinted DMRs, shown on a log-scale. Inset shows distances from 0 to 10 000 bp on a linear scale. Imprinted genes are much more frequently located within 100–100 000 bp of an imprinted DMR. (**B**) Distribution of distance from genes to strain-specific DMRs. Strain-biased genes are for the most part located no closer to a strain-specific DMR than non-differentially expressed genes, indicating the strain-specific differential expression is likely caused by other factors, such as genomic differences. In both cases, we use only DMRs ranked in the top 400.

in sequence complexity inherent to bisulfite treatment. In this study, we demonstrate the use of long-read nanopore sequencing to derive haplotyped methylomes of the embryonic portion of mouse placentae. Methylation estimates from nanopore reads are consistent with previous knowledge (Figure 1). The longer read lengths allowed most reads to overlap multiple SNPs, resulting in accurate haplotyping of 75% of the reads, a much higher proportion than comparable short-read data (Figure 2). Sequencing of native DNA not only maintains the sequence complexity that is lost in bisulfite treatment, but also has the potential to detect a variety of base modifications outside 5mC, bypassing the need for specialized chemistries such as bisulfite (for 5mC) or oxidative-bisulfite (for 5hmC) treatments. Furthermore, we are able to characterize allele-specific methylation at a relatively shallow level of genomic coverage (∼10×), which is substantially lower than the coverage required by Pacific Biosciences single-molecule sequencing to ascertain any native base modification (25×) or 5mC in particular (250×) (23). Nonetheless, a detailed comparison of the per-

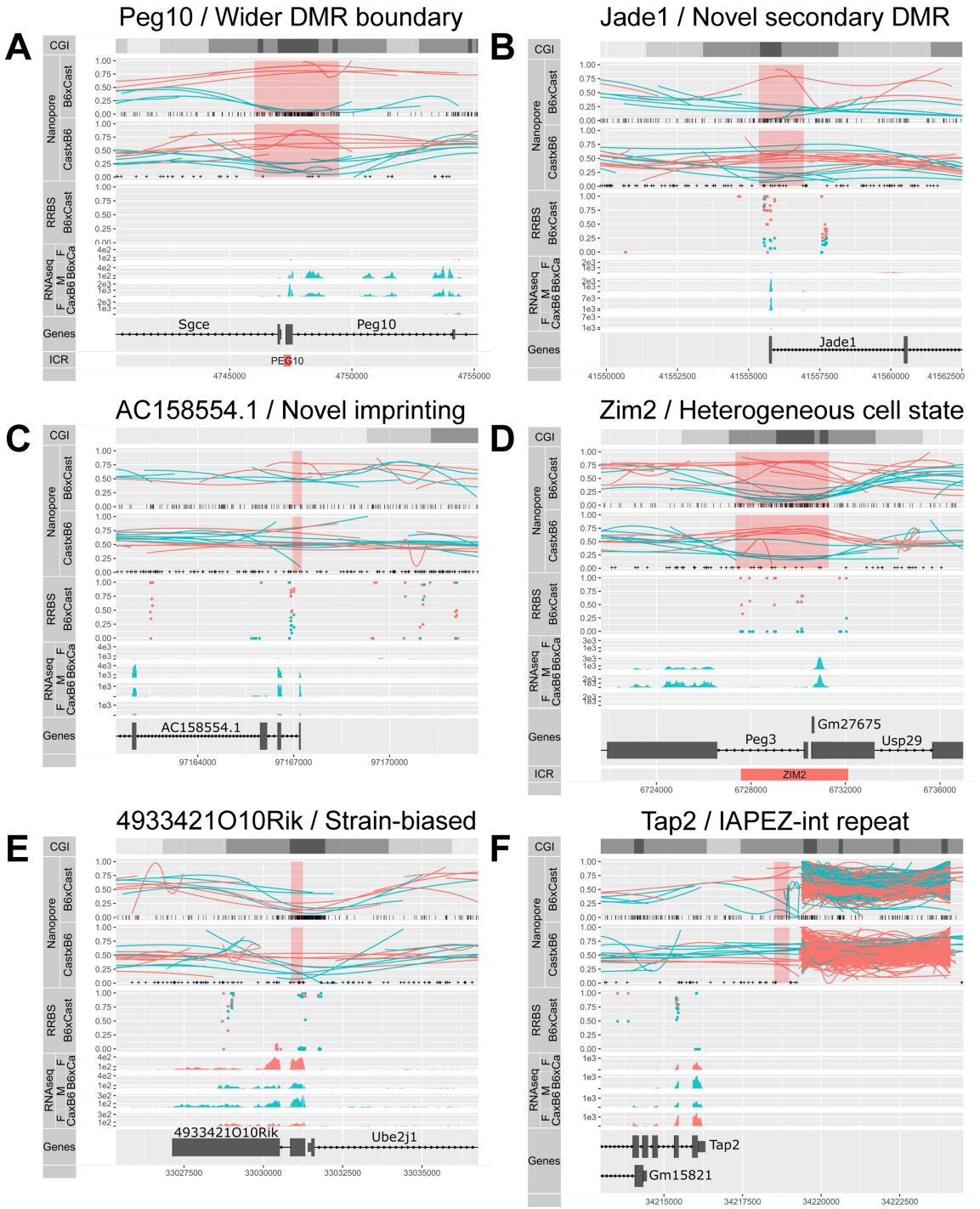

Figure 6. Examples of *de novo* DMRs and the advantages proffered by long reads. (**A**) Allelic methylation (as in Figure 4) plot of maternally imprinted gene *Peg10* displayed a clear DMR at its ICR, which was much wider than the previously annotated DMR (bottom). (**B**) Previously uncharacterized secondary DMR at the TSS of maternally imprinted gene *Jade1*. (**C**) Novel maternally imprinted gene *AC158554.1*, with imprinted methylation at its TSS. (**D**) Allelic methylation plot of maternally imprinted gene *Peg3* showed consistently high methylation across some maternal reads, and consistently low methylation across others, a conclusion that could not be drawn from the middling bisulfite methylation values. (**E**) Strain-of-origin DMR associated with the strain-biased expression of *4933421O10Rik*. (**F**) DMR associated with the omission of a IAPEZ repeat from the Cast genome, suggesting that the methylation in the flanking region was affected by the presence or absence of the repeat.

formance of nanopore and PacBio in the detection of base modifications on matched samples would be of interest.

Recent increases in throughput of nanopore sequencing instruments make this approach a cost-effective way of obtaining genome-wide allele-specific methylation for mammalian-sized genomes, compared to the alternatives of short-read whole-genome bisulfite sequencing, or PacBio SMRT sequencing. Thus, the approach we present is unique in its ability to characterize allele-specific single-molecule cytosine methylation state in eukaryotes, in which the 5mC modification is both common and highly relevant to transcriptional regulation.

The haplotyped methylomes for reciprocal B6 × Cast F1 samples confirm the parent-of-origin specific methylation of ICRs and provide an improved definition of their boundaries (Figure 4). By integrating the haplotyped methylomes with allele-specific expression data, we identified novel DMRs linked to imprinted genes. These are likely to constitute secondary DMRs, whose role and origin are unclear. We note that the low sequencing coverage in this study (∼10×) limits our ability to detect modest methylation differences between alleles, and is thus most suited to the detection of large differences such as those occurring at ICRs. We confirm a large number (70) of previously identified imprinted genes and propose another 65 as new candidates (Figure 3 and Additional File 1). This suggests that although the monoallelically expressed genes are now well characterized, sensitive analyses can still uncover parentally biased genes. Interestingly, though we find more maternal-dominant genes than paternal-dominant ones (88 and 47, respectively), the imbalance is much less pronounced than in Finn *et al.* (2014) (12) (96% maternal dominance). Applying long-read sequencing to the transcriptome also promises improvements in the percentage of usable data, the detection of allele-specific as well as isoform-specific differential expression and even the detection of RNA base modifications.

Our allele-specific methylation and expression data can also be used to reveal strain-biased expression of genes linked to strain-specific DMRs. The genetic divergence between the two strains accounts for most of the differences in expression, however the presence of DMRs could suggest an epigenetic component to the regulation of a subset of genes.

We foresee a number of improvements that will make the determination of haplotyped methylomes by nanopore sequencing more efficient and comprehensive in the future. First, we expect to see an expansion in the types and nucleotide contexts of base modifications characterized. Our analysis is based on Simpson *et al.* (21) (21), and is limited to 5mC at CpG sites. However that is not a limitation of the technology, as has been demonstrated by others ((22), Tombo (bioRxiv doi:10.1101/094672), mCaller (bioRxiv doi:10.1101/127100)). Second, improvements can be made in reaching true nucleotide-resolution methylation calls. Where multiple CpG sites occur within less than twice the *k*-mer length (here 6), all these sites are considered to have the same methylation state. Again, this is not a limitation of the technology, as more complete training data will allow resolution of mixed methylation states. Finally, we see opportunities for improvement in the analysis of nanopore methylation data. Instead of binary calls from bisulfite sequencing, the output of nanopore sequencing is a likelihood ratio that the site is methylated versus unmethylated. Currently, there is no method for the detection of differential methylation that accepts these continuous values as input. Additionally, the DMR detection algorithm that we used was designed originally for bisulfite data, and we expect that algorithms designed specifically to incorporate long reads and probabilistic methylation assignment would achieve greater levels of accuracy.

CONCLUSIONS

We demonstrate that long-read sequencing using nanopore technology can efficiently generate haplotyped mammalian methylomes. With no additional sample preparation than that routinely used for basic sequencing and with only a mean coverage of ∼10×, we identify differential allelic methylation throughout the genome. Combined with expression data, this improves the resolution of imprinting analyses. Our approach is widely applicable to other systems, for instance with more complex genetics, or to phase cancer mutations with methylation state and to determine the effects of structural variation on methylation.

DATA AVAILABILITY

All sequencing data are available at ENA under study accession ERP109201. Processed data can be explored via a Genome browser (63,64), interactive plots (37) and a summary page available from bioinf.wehi.edu.au/haplotyped_methylome. All analysis scripts are available on GitHub at github.com/scottgigante/haplotyped-methylome.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Jared Simpson for advice on running *nanopolish* and Stephen Wilcox for the Illumina sequencing.
Author contributions: Tissue collection was performed by T.B. and Q.G. ONT sequencing was performed by A.L. (with supervision from M.T. and L.G.) and Q.G. RRBS was performed by A.K. and T.B. and RNA-seq by Q.G. Alignment, methylation calling and haplotyping of nanopore data was performed by S.G. A.K. analysed the RRBS and Q.G. the RNA-seq. Development of visualization and DMR testing was performed by S.G., C.W. and T.P.S. Figures were generated by S.G., Q.G. and A.L. S.G., Q.G., A.K., T.P.S., M.E.B. and M.E.R. interpreted the data. S.G., Q.G., C.W., M.E.B. and M.E.R. wrote the manuscript. All authors read and approved the final version.

FUNDING

Bellberry-Viertel Senior Medical Research Fellowship (to M.E.B.); Australian National Health and Medical Research Council (NHMRC) Fellowship [GNT1104924

to M.E.R.]; NHMRC Project Grants [GNT1098290, GNT1140976 to M.E.B., M.E.R.]; Victorian State Government Operational Infrastructure Support; NHMRC Research Institute Infrastructure Support Scheme. Funding for open access charge: Walter and Eliza Hall Institute of Medical Research
Conflict of interest statement. None declared.

REFERENCES

1. Ferguson-Smith,A.C., Sasaki,H., Cattanach,B.M. and Surani,M.A. (1993) Parental-origin-specific epigenetic modification of the mouse H19 gene. *Nature*, **362**, 751–755.
2. Stöger,R., Kubicka,P., Liu,C.G., Kafri,T., Razin,A., Cedar,H. and Barlow,D.P. (1993) Maternal-specific methylation of the imprinted mouse Igf2r locus identifies the expressed locus as carrying the imprinting signal. *Cell*, **73**, 61–71.
3. Bartolomei,M.S., Webber,A.L., Brunkow,M.E. and Tilghman,S.M. (1993) Epigenetic mechanisms underlying the imprinting of the mouse H19 gene. *Genes Dev.*, **7**, 1663–1673.
4. Li,E., Beard,C. and Jaenisch,R. (1993) Role for DNA methylation in genomic imprinting. *Nature*, **366**, 362–365.
5. Ferguson-Smith,A.C. (2011) Genomic imprinting: the emergence of an epigenetic paradigm. *Nat. Rev. Genet.*, **12**, 565–575.
6. Haig,D. (2014) Coadaptation and conflict, misconception and muddle, in the evolution of genomic imprinting. *Heredity*, **113**, 96–103.
7. Frost,J.M. and Moore,G.E. (2010) The importance of imprinting in the human placenta. *PLoS Genet.*, **6**, 1–9.
8. Barlow,D.P. and Bartolomei,M.S. (2014) Genomic imprinting in mammals. *Cold Spring Harb. Perspect. Biol.*, **6**, a018382.
9. Perez,J.D., Rubinstein,N.D. and Dulac,C. (2016) New perspectives on genomic imprinting, an essential and multifaceted mode of epigenetic control in the developing and adult brain. *Annu. Rev. Neurosci.*, **39**, 347–384.
10. Wang,X., Soloway,P.D. and Clark,A.G. (2011) A survey for novel imprinted genes in the mouse placenta by mRNA-seq. *Genetics*, **189**, 109–122.
11. Okae,H., Hiura,H., Nishida,Y., Funayama,R., Tanaka,S., Chiba,H., Yaegashi,N., Nakayama,K., Sasaki,H. and Arima,T. (2012) Re-investigation and RNA sequencing-based identification of genes with placenta-specific imprinted expression. *Hum. Mol. Genet.*, **21**, 548–558.
12. Finn,E.H., Smith,C.L., Rodriguez,J., Sidow,A. and Baker,J.C. (2014) Maternal bias and escape from X chromosome imprinting in the midgestation mouse placenta. *Dev. Biol.*, **390**, 80–92.
13. Calabrese,J.M., Starmer,J., Schertzer,M.D., Yee,D. and Magnuson,T. (2015) A survey of imprinted gene expression in mouse trophoblast stem cells. *G3 (Bethesda)*, **5**, 751–759.
14. Inoue,A., Jiang,L., Lu,F., Suzuki,T. and Zhang,Y. (2017) Maternal H3K27me3 controls DNA methylation-independent imprinting. *Nature*, **547**, 419–424.
15. Andergassen,D., Dotter,C.P., Wenzel,D., Sigl,V., Bammer,P.C., Muckenhuber,M., Mayer,D., Kulinski,T.M., Theussl,H.-C., Penninger,J.M. *et al.* (2017) Mapping the mouse allelome reveals tissue-specific regulation of allelic expression. *eLife*, **6**, 1–29.
16. Eckersley-Maslin,M.A., Alda-Catalinas,C. and Reik,W. (2018) Dynamics of the epigenetic landscape during the maternal-to-zygotic transition. *Nat. Rev. Mol. Cell Biol.*, **19**, 436–450.
17. Duffié,R., Ajjan,S., Greenberg,M.V., Zamudio,N., del Arenal Escamilla,M. Iranzo J., Okamoto,I., Barbaux,S., Fauque,P. and Bourc'his,D. (2014) The Gpr1/Zdbf2 locus provides new paradigms for transient and dynamic genomic imprinting in mammals. *Genes Dev.*, **28**, 463–478.
18. Schilling,E. and El Chartouni,C. Rehli M. (2009) Allele-specific DNA methylation in mouse strains is mainly determined by cis-acting sequences. *Genome Res.*, **19**, 2028–2035.
19. Hitchins,M.P., Rapkins,R.W., Kwok,C.T., Srivastava,S., Wong,J.J., Khachigian,L.M., Polly,P., Goldblatt,J. and Ward,R.L. (2011) Dominantly inherited constitutional epigenetic silencing of MLH1 in a cancer-affected family is linked to a single nucleotide variant within the 5'UTR. *Cancer Cell*, **20**, 200–213.
20. Flusberg,B.A., Webster,D.R., Lee,J.H., Travers,K.J., Olivares,E.C., Clark,T.A., Korlach,J. and Turner,S.W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*, **7**, 461–465.
21. Simpson,J.T., Workman,R.E., Zuarte,P., David,M., Dursi,L. and Timp,W. (2017) Detecting DNA methylation using nanopore sequencing. *Nat. Methods*, **14**, 407–410.
22. Rand,A.C., Jain,M., Eizenga,J.M., Musselman-Brown,A., Olsen,H.E., Akeson,M. and Paten,B. (2017) Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods*, **14**, 411–413.
23. Pacific,Biosciences (2012) *Detecting DNA base modifications using single molecule, real-time sequencing*, Pacific Biosciences.
24. Yang,Y., Sebra,R., Pullman,B.S., Qiao,W., Peter,I., Desnick,R.J., Geyer,C.R., DeCoteau,J.F. and Scott,S.A. (2015) Quantitative and multiplexed DNA methylation analysis using long-read single-molecule real-time bisulfite sequencing (SMRT-BS). *BMC Genomics*, **16**, 350.
25. Rhoads,A. and Au,K.F. (2015) PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics*, **13**, 278–289.
26. Jain,M., Koren,S., Miga,K.H., Quick,J., Rand,A.C., Sasani,T.A., Tyson,J.R., Beggs,A.D., Dilthey,A.T., Fiddes,I.T. *et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, **36**, 338–345.
27. Mould,A.W., Pang,Z., Pakusch,M., Tonks,I.D., Stark,M., Carrie,D., Mukhopadhyay,P., Seidel,A., Ellis,J.J., Deakin,J. *et al.* (2013) Smchd1 regulates a subset of autosomal genes subject to monoallelic expression in addition to being critical for X inactivation. *Epigenetics Chromatin*, **6**, 19.
28. Keniry,A., Gearing,L.J., Jansz,N., Liu,J., Holik,A.Z., Hickey,P.F., Kinkel,S.A., Moore,D.L., Breslin,K., Chen,K. *et al.* (2016) Setdb1-mediated H3K9 methylation is enriched on the inactive X and plays a role in its epigenetic silencing. *Epigenetics Chromatin*, **9**, 16.
29. Yin,D., Ritchie,M.E., Jabbari,J.S., Beck,T., Blewitt,M.E. and Keniry,A. (2016) High concordance between Illumina HiSeq2500 and NextSeq500 for reduced representation bisulfite sequencing (RRBS). *Genom. Data*, **10**, 97–100.
30. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
31. Krueger,F. and Andrews,S.R. (2016) SNPsplit: allele-specific splitting of alignments between genomes with known SNP genotypes [version 2; referees: 3 approved]. *F1000Research*, **5**, 1479.
32. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
33. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
34. McCarthy,D.J., Chen,Y. and Smyth,G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.
35. Lun,A.T., Chen,Y. and Smyth,G.K. (2016) It's DE-licious: a recipe for differential expression analyses of RNA-seq experiments using quasi-Likelihood methods in edgeR. *Methods Mol. Biol.*, **1418**, 391–416.
36. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc: Series B*, **57**, 289–300.
37. Su,S., Law,C.W., Ah-Cann,C., Asselin-Labat,M.-L., Blewitt,M.E. and Ritchie,M.E. (2017) Glimma: interactive graphics for gene expression analysis. *Bioinformatics*, **33**, 2050–2052.
38. Keane,T.M., Goodstadt,L., Danecek,P., White,M.A., Wong,K., Yalcin,B., Heger,A., Agam,A., Slater,G., Goodson,M. *et al.* (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, **477**, 289–294.
39. Loman,N.J., Quick,J. and Simpson,J.T. (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods*, **12**, 733–735.
40. Therneau,T., Atkinson,B. and Ripley,B. (2017) rpart: Recursive Partitioning and Regression Trees. *R package version 4.1-11*. https://cran.r-project.org/package=rpart.
41. Decato,B.E., Lopez-Tello,J., Sferruzzi-Perri,A.N., Smith,A.D. and Dean,M.D. (2017) DNA methylation divergence and tissue

specialization in the developing mouse placenta. *Mol. Biol. Evol.*, **34**, 1702–1712.
42. Park,Y. and Wu,H. (2016) Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*, **32**, 1446–1453.
43. Cleveland,W.S., Grosse,E. and Shyu,W.M. (1992) Local regression models. In: Chambers,JM and Hastie,T (eds). *Statistical models in S*. Chapman and Hall, NY, Vol. **2**, pp. 309–376.
44. Yin,T., Cook,D. and Lawrence,M. (2012) ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol.*, **13**, R77.
45. Lawrence,M., Gentleman,R. and Carey,V. (2009) rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*, **25**, 1841–1842.
46. Lawrence,M., Huber,W., Pagès,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M. and Carey,V. (2013) Software for Computing and Annotating Genomic Ranges. *PLoS Comp Biol.*, **9**, e1003118.
47. Irizarry,R.A., Ladd-Acosta,C., Wen,B., Wu,Z., Montano,C., Onyango,P., Cui,H., Gabo,K., Rongione,M., Webster,M. *et al.* (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.*, **41**, 178–186.
48. Walker,D.L., Bhagwate,A.V., Baheti,S., Smalley,R.L., Hilker,C.A., Sun,Z. and Cunningham,J.M. (2015) DNA methylation profiling: comparison of genome-wide sequencing methods and the Infinium Human Methylation 450 Bead Chip. *Epigenomics*, **7**, 1287–1302.
49. Friedman,J., Hastie,T. and Tibshirani,R. (2009) glmnet: lasso and elastic-net regularized generalized linear models. *R Package Version 2.0-16*. https://cran.r-project.org/package=glmnet.
50. Morison,I.M., Ramsay,J.P. and Spencer,H.G. (2005) A census of mammalian imprinting. *Trends Genet.*, **21**, 457–465.
51. Jirtle,R. (2001) *World Wide Web Site - Geneimprint Imprinted genes database*. http://www.geneimprint.com/site/genes-by-species.
52. Schulz,R., Woodfine,K., Menheniott,T.R., Bourc'his,D., Bestor,T. and Oakey,R.J. (2008) WAMIDEX: a web atlas of murine genomic imprinting and differential expression. *Epigenetics*, **3**, 89–96.
53. Williamson,C., Blake,A., Thomas,S., Beechey,C., Hancock,J., Cattanach,B. and Peters,J. (2013) *World Wide Web Site - Mouse Imprinting Data and References*. MRC Harwell, Oxfordshire, https://www.mousebook.org/introduction-data-imprinting-genes.
54. Kobayashi,H., Sakurai,T., Imai,M., Takahashi,N., Fukuda,A., Yayoi,O., Sato,S., Nakabayashi,K., Hata,K., Sotomaru,Y. *et al.* (2012) Contribution of intragenic DNA methylation in mouse gametic DNA methylomes to establish oocyte-specific heritable marks. *PLoS Genet.*, **8**, e1002440.
55. Smallwood,S.A., Tomizawa,S.I., Krueger,F., Ruf,N., Carli,N., Segonds-Pichon,A., Sato,S., Hata,K., Andrews,S.R. and Kelsey,G. (2011) Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nat Genet.*, **43**, 811–814.
56. Smith,Z.D., Chan,M.M., Mikkelsen,T.S., Gu,H., Gnirke,A., Regev,A. and Meissner,A. (2012) A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature*, **484**, 339–344.
57. Moore,T., Constancia,M., Zubair,M., Bailleul,B., Feil,R., Sasaki,H. and Reik,W. (1997) Multiple imprinted sense and antisense transcripts, differential methylation and tandem repeats in a putative imprinting control region upstream of mouse Igf2. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 12509–12514.
58. Okae,H., Matoba,S., Nagashima,T., Mizutani,E., Inoue,K., Ogonuki,N., Chiba,H., Funayama,R., Tanaka,S., Yaegashi,N. *et al.* (2014) RNA sequencing-based identification of aberrant imprinting in cloned mice. *Hum. Mol. Genet.*, **23**, 992–1001.
59. da Rocha,S.T., Edwards,C.A., Ito,M., Ogata,T. and Ferguson-Smith,A.C. (2008) Genomic imprinting at the mammalian Dlk1-Dio3 domain. *Trends Genet.*, **24**, 306–316.
60. Smith,R.J., Dean,W., Konfortova,G. and Kelsey,G. (2003) Identification of novel imprinted genes in a genome-wide screen for maternal methylation. *Genome Res.*, **13**, 558–569.
61. Court,F., Tayama,C., Romanelli,V., Martin-Trujillo,A., Iglesias-Platas,I., Okamura,K., Sugahara,N., Simon,C., Moore,H., Harness,J.V. *et al.* (2014) Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. *Genome Res.*, **24**, 554–569.
62. Takemoto,N., Yoshimura,T., Miyazaki,S., Tashiro,F. and Miyazaki,J. (2016) Gtsf1l and Gtsf2 are specifically expressed in gonocytes and spermatids but are not essential for spermatogenesis. *PLoS One*, **11**, 1–12.
63. Buels,R., Yao,E., Diesh,C.M., Hayes,R.D., Munoz-Torres,M., Helt,G., Goodstein,D.M., Elsik,C.G., Lewis,S.E., Stein,L. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.
64. Hofmeister,B.T. and Schmitz,R.J. (2018) Enhanced JBrowse plugins for epigenomics data visualization. *BMC Bioinformatics*, **19**, 159.