Choice of reference genome can introduce massive bias in bisulfite sequencing data

Phillip Wulfridge[1], Ben Langmead[2,3], Andrew P. Feinberg[1,4,5,6], and Kasper D. Hansen[1,2,7,8,*]

[1]Center for Epigenetics, Johns Hopkins School of Medicine
[2]Center for Computational Biology, Johns Hopkins University
[3]Department of Computer Science, Johns Hopkins University
[4]Department of Medicine, Johns Hopkins School of Medicine
[5]Department of Biomedical Engineering, Whiting School of Engineering
[6]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health
[7]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
[8]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine

Abstract

Mapping bias can be introduced in analysis of short read sequencing data, if sequence reads are aligned to a different genome than the sample genome. Here we study mapping bias in whole-genome bisulfite sequencing using data from inbred mice. We show that the choice of reference genome used for alignment can profoundly impact the inferred methylation state, both for high and low resolution analyses. This bias can result in falsely identifying thousands of differentially methylated regions and hundreds of megabases of large-scale methylation differences. We show that the direction of these biased methylation differences can be reversed by changing the reference genome, clearly establishing mapping bias as a primary cause. We develop a strategy we call personalize-then-smooth for removing the bias by coupling alignment to personal genomes, with post-alignment smoothing. The smoothing step can be viewed as imputation, and allows a differential analysis to include methylation sites which are only present in some samples. Our results have important implications for analysis of bisulfite converted DNA.

[*] To whom correspondence should be addressed. Email: khansen@jhsph.edu

Introduction

DNA methylation is a key epigenetic mark that has become widely implicated in human development and disease (Feinberg and Vogelstein, 1983; Feinberg, Koldobskiy, and Göndör, 2016). Accurate determination of methylation at CpG dinucleotide positions across the genome is critical for understanding its association with functional regulation. Multiple techniques currently exist to perform this measurement, each with varying degrees of genomic coverage and depth. One gold-standard method is whole-genome bisulfite sequencing (WGBS), which pairs bisulfite conversion of cytosine residues with next-generation sequencing (Lister et al., 2009). At each CpG site, an aligned read is called as unmethylated if the sequence is TG (indicating bisulfite conversion) and methylated if the sequence is CG (indicating protection by the methyl group). Statistical packages such as BSmooth (Hansen, Langmead, and Irizarry, 2012) can then integrate this data across larger regions to estimate and compare overall methylation patterns between sample groups.

WGBS short-read mapping is reliant on a reference genome, from which in silico bisulfite-converted genomes are generated for use in read alignment (Lister et al., 2009; Hansen, Langmead, and Irizarry, 2012; Krueger and Andrews, 2011). Because a reference genome is used, deviations from that sequence, such as single nucleotide variants or indels, can affect alignment quality and/or methylation estimation. Notably, C/T mutation occurring at CpG sites is the most common dinucleotide mutation in the mammalian genome due to the high rate of spontaneous deamination at methylated CpG (Hodgkinson and Eyre-Walker, 2011; Coulondre et al., 1978; Bird, 1980). Such a variant would still align to a bisulfite-converted reference, but appear as an entirely unmethylated CpG site, even though the CpG site no longer exists.

Mapping bias associated with non-reference alleles has been examined in transcriptomics, particularly in studies associating genotype with expression. Degner et al. (2009) studied bias in allele-specific expression and showed it is predominantly in the direction of higher expression of the reference allele. Satya, Zavaljevski, and Reifman (2012) and Geijn et al. (2015) showed this bias can be alleviated by integrating prior knowledge of non-reference alleles present in the population, and Rozowsky et al. (2011) showed that a similar improvement can be achieved using a personalized diploid reference genome, a strategy also used in later studies (Munger et al. (2014) and Hodgkinson, Grenier, et al. (2016)). Panousis et al. (2014) studied bias in eQTL analysis, and concluded that it has little impact because the bias affects quantification locally around a sequence variant, whereas eQTL analyses consider aggregate expression over a region extending beyond the variant.

Given the prevalence of C/T mutations at CpGs in mammals, it is sensible to wonder about its effect on analysis of WGBS data, but currently, little data exists on the magnitude of its impact on analysis results. Related, tools have been developed for the joint methylation quantification and genotyping of bisulfite converted DNA (Liu et al., 2012;

Barturen et al., 2013; Gao et al., 2015), but bisulfite conversion makes this task difficult.

Here, we examine the consequences of using an improper reference genome for alignment of bisulfite converted DNA. We compare WGBS results from two inbred mouse strains: C57BL/6J, upon which the mouse reference genome is based; and the highly divergent, wild-derived CAST/EiJ. We show that the choice of reference genome has a profound impact on inferred strain differences in methylation and that this is predominately caused by CpG mutations between the strains. We show that this bias can be corrected by alignment to personal genomes, and we introduce smoothing as a method to compare methylomes in different genomic coordinate systems. Finally, we examine two other strategies for addressing this bias: alignment to a common reference followed by filtering of variable CpG sites, and genotyping using bisulfite-converted DNA.

Results

Global methylation estimates are strongly influenced by choice of reference genome

We performed whole-genome bisulfite sequencing on liver samples from two inbred mouse strains, C57BL/6J (BL6) and CAST/EiJ (CAST), with 4 mice per strain. BL6 was chosen as it is the standard laboratory strain and the basis of the mm9 reference genome; the wild-derived CAST was selected as a highly divergent strain for comparison. The purpose of the experiment was to identify strain-specific regions of differential methylation, possibly driven by genetic changes. As sequencing data was generated at relatively low coverage (Table 1), we chose to perform downstream methylation analysis using the BSmooth package (Hansen, Langmead, and Irizarry, 2012), which is designed in part to handle low-coverage data.

We first aligned sequencing reads from both strains to the standard BL6/mm9 reference genome, then computed the average methylation across all read-covered autosomal CpGs in the reference; we refer to this measurement as global methylation. When compared, the two strains showed a dramatic difference in global methylation, with the CAST strain's estimates lower by over 7.6% (Fig. 1, $p < 1.3 \times 10^{-6}$). This difference is comparable in magnitude to the level previously observed between tumor and normal colon (Hansen, Timp, et al., 2011) and associated with EBV-mediated oncogenesis (Hansen, Sabunciyan, et al., 2014), and is far larger than what we would normally expect from a comparison between strains or individuals. Note that global methylation is an average across millions of CpGs, and thus is unlikely to be affected to this extent by differences in local coverage.

We next aligned the BL6 and CAST samples to the CAST reference genome. Strikingly, the global methylation levels computed for these alignments showed a reverse relationship to the levels observed in the alignment to the BL6 reference, with BL6 now showing a global methylation lower by around 8.7% (Fig. 1, $p < 1.6 \times 10^{-6}$). As these drastically different estimates were obtained from the same sets of sequencing reads, we conclude that the choice of reference genome is by itself a major determinant of global methylation estimates. Interestingly, we also observe that aligning data from a strain to its own reference genome yields the same global methylation level in the two strains, whereas aligning a strain to the more distant reference genome results in lower global methylation. This implies that the true global methylation level in both strains is similar, and observing lower methylation is an effect of what we term mapping bias.

We observe that samples have higher alignment rate (fraction of reads aligned) when aligned to the proper strain (Table 1). However, we also observe that CAST samples aligned to the BL6 genome have higher alignment rates than BL6 samples aligned to the BL6 genome. We conclude that alignment rates could be useful when considering which reference genome to use for a given sample, but are not necessarily comparable across samples.

Mapping bias induces focal changes in DNA methylation

Having established that alignment to a reference genome of a different strain substantially impacts global methylation, we asked whether this could also result in focal changes. We previously showed that DNA methylation patterns can change at different scales (Hansen, Timp, et al., 2011; Hansen, Langmead, and Irizarry, 2012; Hansen, Sabunciyan, et al., 2014). Both in those studies and in the current study, we focused on (a) small differentially methylated regions (DMRs), ranging in size from hundreds to a few thousand bases, and (b) methylation blocks on the hundred-kilobase to megabase scale. We used the BSmooth pipeline (Hansen, Langmead, and Irizarry, 2012) to analyze the two mouse strains described previously and identified both small DMRs and methylation blocks. We repeated the analyses using the BL6 and CAST reference genomes in order to associate differential methylation events with strain differences. We assessed significance of the DMRs using a permutation scheme which controls the genome-wide family-wise error rate (gFWER), a stringent error rate for genome-wide studies (Methods).

As we expected, the choice of reference genome determined the overall direction of differences observed between strains. We identified 2,354 blocks covering 1.89 million CpGs and 347 Mb in BL6-aligned data and 4,199 blocks covering 6.42 million CpGs and 1,101 Mb in CAST-aligned data meeting our cutoff criteria (gFWER $\leq 1/18$, length > 10 kb). In both alignments, the vast majority of blocks were hypermethylated in the same strain as was used for the reference genome (98.6% when aligned to BL6, 99.3% when aligned to CAST), which is consistent with the pattern observed in global methylation. Figure 2 illustrates a large-scale methylation block present in both analyses, where the sign of the methylation difference depends on the choice of reference genome. Notably, when we compared only samples aligned to their own reference (adjusting for coordinate changes due to indels between genomes), we found that methylation levels across the block were approximately equal (Figure 2c). This observation is again consistent with our findings on global methylation, and forms the basis of our proposed solution to mapping bias, which we detail in a later section.

We found 2,865 DMRs covering 57k CpGs and 3.14 Mb of the genome that met our cutoff criteria (gFWER $\leq 1/18$, mean difference > 0.1, Methods) when samples were aligned to BL6, compared to 6,498 DMRs covering 94k CpGs and 5.72 Mb when aligned to CAST. In contrast to the methylation blocks, however, these small DMRs were not uniformly hypermethylated in the strain to which the samples were aligned; only 45.8% were hypermethylated in BL6 when aligned to BL6, and 65.4% were hypermethylated in CAST when aligned to CAST. This observation can be explained by the presence of "true" strain DMRs within the analysis which, unlike DMRs that arise from mapping bias, would not necessarily reverse direction between alignments. Figure 3 depicts an example small DMR identified in both alignments, the direction of which reverses depending on which genome was used for mapping. Again, when samples aligned to their own reference were compared, methylation levels were approximately equal (Figure 3c).

Together, these observations show that the choice of reference genome can introduce widespread focal changes in DNA methylation at both the kilobase and megabase scale, and that these changes are biased such that strains more genetically distant from the reference strain appear more hypomethylated.

Strain differences in CpG dinucleotides

We examined sequence differences between the BL6 and CAST strains, specifically at CpG locations, to see how they affect methylation estimates. Notably, a CG-to-TG (or CG-to-CA) mutation in CAST, when aligned to the BL6 reference genome, would produce unmethylated calls without inducing any alignment mismatches (i.e. essentially being ”undetectable” by standard WGBS pipelines), resulting in an excess of 0% methylated locations and the previously observed decrease in global methylation. Such mutations are in fact the most common dinucleotide mutation in mammalian genomes (Hodgkinson and Eyre-Walker, 2011).

To facilitate accurate sequence comparisons between strains, which do not share coordinate systems, we used the modmap tool (Huang et al., 2013). This tool functions similarly to UCSC's liftOver (Hinrichs et al., 2006) to convert a set of genomic locations to their corresponding coordinates in another strain, and was the method we used to compare methylation regions of self-aligned samples in the previous section (Methods).

Using modmap and the FASTA file for CAST, we extracted and tabulated the forward-strand dinucleotide sequences corresponding to the 21.3 million CG dinucleotides in BL6; the results are shown in Table 2. Approximately 19 million CGs from BL6 were retained by CAST; 1.60 million were mutated to either TG or CA, with another 0.52 million mutated to GG/CC and AG/CT (which would result in alignment errors). 100k CpGs could not be mapped from BL6 to CAST, which occurs when an indel over a CpG results in an ambiguous position after modmap conversion. We observed similar results when we performed the reverse analysis, tabulating sequences of CAST CpGs when modmapped to BL6: about 2.3 million CGs were unique to CAST with a differing or unmappable sequence in BL6 (Table 2).

The CpG mutation rate of roughly 2.1/23.1=9.1% can be compared to the genomic mutation rate of roughly 1.2% (computed by considering single nucleotide mutations as well as indels, using modmap between the two genomes, see Methods). This CpG mutation rate is not uniform across the methylome, being much lower in CpG islands (1.2%) and somewhat lower in allosomes (6.4%) and CpG island shores (5.4%). Interestingly, though the CpG mutation rate fluctuates throughout the genome, it is generally similar across the two strains for any given region (Figure 4). It is well understood that the high rate of CpG mutation is caused by the nature of DNA methylation; methylated C positions are more likely to undergo spontaneous deamination to T (Coulondre et al., 1978; Bird, 1980; Hodgkinson and Eyre-Walker, 2011).

To confirm that CpG sequence differences have a large impact on observed DNA methylation, we computed global methylation as in Figure 1, but with respect to only the 19M CpGs common between the strains (Figure 5). We found that the large differences previously observed were no longer present. Instead, we saw a similar level of global methylation across strains regardless of which genome was used for alignment. We conclude that CpG loss due to mutation, which is largely undetectable by bisulfite sequencing, is responsible for the mapping bias.

Alignment to personal genomes addresses mapping bias

Having identified mapping bias and its source, we next explored strategies for mitigating or eliminating the bias, with the goal of providing recommendations for future studies. First, we considered the strategy of using personal (i.e. strain-specific) genomes for alignment. The computational cost of this strategy is not onerous. It requires building a personal genome sequence and a corresponding index for bisulfite read alignment using a tool like BSmooth or Bismark. But the strategy incurs essentially no computational overhead at alignment time.

While the personal genome alignment strategy is easy in the alignment stage, it creates issues for downstream analysis. Following alignment, the data for each sample is in different genome coordinates. For DNA methylation data in mammals, the critical issue is to map CpG dinucleotides between the different genomes, arguably an easier task than full genome alignment. An aspect which needs special attention is the treatment of a CpG which only exists in some of the genomes. It is tempting to simplify an analysis by only considering CpGs present in all genomes. However, CpGs present in only some of the genomes will be of particular interest if the purpose is to jointly analyze the genome and the epigenome, since a loss of a CpG is a direct way for a sequence change to impact the epigenome by forcing 0% methylation at that nucleotide.

To address this problem, we propose the following strategy. Following alignment to personal genomes, place CpGs from different genomes in a common coordinate system using a conversion tool, such as modmap or liftOver. Next, use BSmooth (or another smoothing technique) to smooth the methylation data. This yields a continuous methylation curve across the whole genome. From this curve, we can infer a methylation value at any site, including at a site with a CpG shared by the two strains strains, and, crucially, at a site with a CpG in one strain but not the other. The smoothed curves can then be evaluated within the set of common CpGs; however, data from sample-specific CpGs influence the imputed methylation values at these common CpGs, and thus still contribute to the final analysis. In the following, we use the term "personalize-then-smooth" to refer to this combination of alignment and smoothing.

We reanalyzed the two mouse samples analyzed previously using our personalize-then-smooth strategy. We aligned the samples to their respective strain-specific reference genomes (i.e. BL6 to BL6 and CAST to CAST), used modmap to convert methylation calls at CpG positions into BL6 coordinates, and used BSmooth to smooth data across all CpGs present in one or both genomes. Strikingly, we identified a far fewer focal changes between strains compared to our original approach. Using the same significance cutoffs as previously, we found only 101 blocks containing 27k CpGs and covering 3.2 Mb (gFWER $\leq 1/18$, width $> 10^5$) and 976 small DMRs covering 23k CpGs and 1.13 Mb (gFWER $\leq 1/18$, mean difference > 0.1, Methods). Blocks identified were also very small (all were less than 120kb in width, with 92% smaller than 50kb), which indicates that they are likely not true large-scale blocks; additionally, blocks were no longer uniformly hypermethylated towards one strain, with only 46.5% hypermethylated in BL6. Figures 2c and 3c provide examples of respectively a biased block and biased DMR, that under the personalize-then-smooth approach is observed to have no true methylation difference. Together, these results show that our strategy largely removes false-positive blocks and DMRs produced by mapping bias.

To illustrate what drives biased DMRs, we re-examined our list of 2,865 DMRs found in the analysis where all samples were mapped to the BL6 genomes; we call these "biased DMRs". Each DMR represents a genomic region in BL6 coordinates. For each such region we determined whether the region overlapped one of the 976 DMRs identified using the personalize-then-smooth strategy (hereafter referred to as "personal" DMRs), and we computed the average difference in methylation across the two strains, using methylation quantification from personal genomes. Comparing the average methylation between the two analysis strategies, we observe that biased DMRs which do not overlap personal DMRs show disagreement between the average methylation as quantified by the two strategies, whereas biased DMRs overlapping personal DMRs show similar average methylation between the two strategies (Figure 6a,b). Additionally, for each biased DMR we computed the number of CpGs unique to both the BL6 and the CAST genomes. We observe that there are substantial differences between the number of unique CpGs in the BL6 genome for DMRs which do or do not overlap personal DMRs (Figure 6c). As a control, we see similar levels of CAST specific CpGs in the two groups. Finally, the percentage of unique CpGs is a major determinant of the methylation difference between the two analysis strategies (Figure 6d). This shows that the large amount of BL6 specific CpGs drives the bias for the biased DMRs.

Intriguingly, we find 192 biased DMRs which do not overlap personal DMRs and which has no CpGs unique to either BL6 or CAST, suggesting that the set of biased DMRs is not entirely caused by CpG mutations (Figure 6d). One of these regions is examined in detail below.

Strain-specific DMRs are in functionally associated regions

Focusing on a possible functional role of strain-specific methylation, we investigated the overlap between our set of 976 personal DMRs and various marks suggesting functional relevance from ENCODE, as well as genomic regions of interest such as Refseq promoters and CpG islands. Specifically, ENCODE has profiled peaks for H3K4me1, H3K4me3, H3K27ac (marks enriched at enhancers and promoters), CTCF (a methylation-sensitive transcription factor), and POL2 in liver in adult BL6 mice. We quantified enrichment by computing a $\log_2$ odds ratio; we also determined the number of strain-specific DMRs overlapping the particular feature, n (Methods).

We observed strong enrichment of H3K4me1, H3K4me3, H3K27ac and CTCF at strain-associated DMRS, a moderate enrichment of CpG island shores, and a small enrichment at CpG islands and Refseq gene promoters; no enrichment was observed at POL2 peaks (Table 3). In summary 750 out of 976 personal DMRs overlapped one or more features ($\log_2(\mathrm{OR}) = 3.03, p < 2.2 \times 10^{-16}$). This strongly suggests that these DMRs are found at regions of functional relevance in liver. This should be interpreted with caution since biased DMRs, which are not personal DMRs, are also overlapping known features: 1182 out of 2121 such DMRs overlapped one or more features ($\log_2(\mathrm{OR}) = 2.12, p < 2.2 \times 10^{-16}$)

Despite large differences in methylation between strains, 94 of our 976 personal DMRs contain no CpGs that are mutated between strains, and 20 of those 94 contain no sequence variation whatsoever between strains within the DMR (10 of these 20 DMRs continue to have no sequence variation within 1kb of the DMR and 4/20 remain identical within 5kb). For both these two smaller sets of DMRs, the effect sizes of the enrichment of the various features are similar, with the exception that POL2 is depleted at DMRs over areas with no sequence variation. 70 out of 94 DMRs with no mutated CpGs overlap at least one functional feature ($\log_2(\mathrm{OR}) = 3.22, p < 2.2 \times 10^{-16}$), and 15 out of 20 DMRs with no sequence variation overlap ($\log_2(\mathrm{OR}) = 2.99, p < 2.2 \times 10^{-16}$).

Alignment to common reference genome, followed by filtering

In this section we describe an alternative to the personal genome strategy described above. We term this strategy the "shared CpG" strategy where all samples are aligned to the BL6 genome, but subsequently we only keep methylation measurements for 19M CpGs present in both genomes for input into BSmooth. By aligning to a single genome, this strategy sidesteps the issue of multiple coordinate systems, but still uses full knowledge of the CpG dinucleotides in the involved genomes. We consider this strategy for two purposes. First, it allows us to assess the impact of only considering common CpGs for analysis. Second, this strategy is a variant of the "align, then filter" strategy, which is popular in transcriptomics, where a set of known regions exhibiting mapping bias are used to remove certain loci (Lappalainen et al., 2013; Castel et al., 2015). In transcriptomics, this set of known loci can be identified using simulation together with knowledge about population-level single nucleotide variants. Such an approach is more difficult for bisulfite converted DNA, since the methylation state of a CpG influences the sequence of any read covering the CpG.

We found that this shared CpG approach removes bias to a similar extent as our personalize-then-smooth strategy. We identified only 108 methylation blocks covering 28k CpGs and 3.8 Mb meeting our cutoff criteria (gWER $\leq 1/18$, length > 10 kb), an amount similar to what we found using personalize-then-smooth. For small DMRs we found 622 in total, covering 15k CpGs and 729 kb of the genome (gFWER $\leq 1/18$, mean difference > 0.1, Methods). As expected, most (542) of these small DMRs overlapped personal DMRs, and another 69 overlapped the set of putative personal DMRs with gFWER $> 1/18$ (we use putative to indicate candidate DMRs which do not meet our criteria for a genome-wide family-wise error rate). However, there were some notable differences between the two analyses. The set of personal DMRs was markedly larger (at 976) than the set of shared CpG DMRs, and accordingly many personal DMRs were not identified by the shared CpG approach; conversely, 11 regions identified in the shared CpG analysis overlapped no personal DMRs or putative personal DMRs.

One of these 11 DMRs (Figure 7) illustrates an intriguing example of non-CpG alignment bias. This region contains no CpGs unique to either of our two genomes, which we have shown is the usual driver of bias. Nevertheless, the region displays the typical hallmarks of mapping bias, with samples extremely hypomethylated when aligned to a distant reference (Figure 7a, b) and no true differences in methylation when samples are aligned to personal genomes (Figure 7c). Notably, this bias is not removed in the shared CpG analysis, where CAST samples continue to be hypomethylated (Figure 7d). We investigated further and found that an unusually high number of sequencing reads from CAST samples aligned to this region when the BL6 reference was used. When the CAST reference was used, these same reads aligned instead to the mitochondrial genome. These fragments were all unmethylated, which would lower methylation estimates for 6 CpGs in the region and produce hypomethylation bias. Surprisingly, the exact same happens for a set of BL6 reads when aligned to the CAST genome; this is the reason for the switch in direction seen in Figures 7a,b. We concluded that non-CpG sequence variation between strains, both in this region and the mitochondrial genomes, caused misalignment and subsequent mapping bias in this particular region. This example, while obviously unusual, shows that alignment bias can occur with alignment to a distant reference, even after accounting for CpG differences.

Personal DMRs which are also shared CpG DMRs have a lower proportion of strain unique CpGs compared to personal DMRs which are not shared CpG DMRs (Figure 8). Personal DMRs which are also shared CpG DMRs are strongly enriched for functionally associated regions (434 out of 542, $\log_2(\mathrm{OR}) = 3.28, p < 2.2 \times 10^{-16}$); more so than personal DMRs which are not shared CpG DMRs (316 out of 434, $\log_2(\mathrm{OR}) = 2.67, p < 2.2 \times 10^{-16}$).

Together, these observations suggest that personalize-then-smooth has more power to identify strain differences.

Genotyping bisulfite converted data

It is possible to genotype samples using bisulfite converted data (Liu et al., 2012; Barturen et al., 2013; Gao et al., 2015). Here, we want to investigate the use of these tools in case personal genomes are not available in a given application. In general, the authors of these methods recommend against using less than 10x coverage data, suggesting that genotyping using our low-coverage data will be challenging. To overcome this, we pooled 4 genetically identical CAST replicates into one metasample with 24x coverage, aligned to the BL6 genome, and genotyped using BS-SNPer (Gao et al., 2015).

There are 2.26M CpGs which only exist in CAST, and these CpGs are well covered when the samples are aligned to the CAST genome (2.26M are covered at 1x, 2.17M are covered at 10x). BS-SNPer does not have similar performance on CpGs which are unique to the BL6 genome and CpGs which are unique to the CAST genome. BS-SNPer recovers 51% of CpGs unique to the CAST genome (where a gain of CpG needs to be identified) and 24% of the CpGs unique to the BL6 genome (where a loss of a CpG needs to be identified), see Table 4.

We used the output of BS-SNPer to compute global methylation as a proxy for bias. The BL6 metasample had a global methylation of 72.0%, very similar to our previous estimates of global methylation in BL6, whereas the CAST metasample had a global methylation of 68.2%, better than our estimates of CAST samples aligned to BL6, but not a full recovery of the unbiased global methylation. In conclusion, for our data, BS-SNPer does not rescue the bias introduced by using a wrong reference genome.

Discussion

In this paper, we demonstrate that aligning whole-genome bisulfite sequencing reads to a divergent reference genome induces massive bias in the quantification of methylation, and this can dramatically confound comparisons between samples with different genomes. This bias drastically affects both large- and small-scale analyses. In a clear sign that the genetic differences are to blame, the direction of methylation change can be reversed by changing reference genomes. We show the bias is predominately – but not exclusively – caused by differences in CpG sites between the sample and reference genome; this is also the most common type of dinucleotide change. If large, the bias can be detected by examining the global methylation between samples. The bias can be eliminated by aligning data to personal genomes. Analyzing data mapped to personal genomes requires working in distinct coordinate systems until the quantification step. We show how smoothing can be used to impute methylation at strain-specific CpGs, allowing methylation measurements to be compared in a common coordinate system without discarding any measurements. We use the term "personalize-then-smooth" to refer to this combination of alignment and post-alignment smoothing.

We have used the term mapping bias throughout this work, because the bias is controlled by the choice of reference genome used for alignment. However, in the case of a C to T transition between the reference genome and the sample genome, the aligner is placing the read correctly, and what fails is our inference based on a combination of the aligned read and the existence of a CpG in the reference genome.

Previously, (Degner et al., 2009) showed that allele specific expression can be affected by mapping bias, whereas (Panousis et al., 2014) showed that eQTL analysis is unaffected. eQTL analysis is largely unaffected by mapping bias because the bias occurs locally around a sequence variant, whereas gene expression is averaged across a much larger region than affected by the bias. In contrast, for DNA methylation – like allele-specific expression – the quantity of interest is directly on top of a sequence variation. Furthermore, DNA methylation in mammals occur primarily at CpG dinucleotides which are also the most frequent sites of sequence variation.

We have examined three approaches for removing the bias: (1) use of personalize-then-smooth, (2) (complete) knowledge of population level variation in CpGs, and (3) use of tools which jointly genotypes and quantifies methylation in bisulfite converted DNA. The use of personalize-then-smooth is our method of choice, but it does require the availability of sample specific genomes. To examine the utility of knowledge of population level variation we considered an extreme case where complete knowledge about CpG variants in the samples under study is available. With this knowledge, mapping bias can be removed, but our results should be considered an upper bound on the effectiveness of this approach. Finally, we were unsuccessful in using BS-SNPer to remove bias; this could be caused by the use of this particular tool or the coverage of our sequence data. The later point could be addressed by generating much deeper WGBS data; whether this is cost-effective compared to performing DNA sequencing of the strains depends on the experimental design. In this context, an advantage of our smoothing approach is the ability to deal with low-coverage bisulfite sequencing data.

To detect the bias we recommend routinely examining global methylation across samples. This is easy and effective, but will only detect strong genome-wide mapping bias. Care should be taken; it is well established that global methylation is cell type dependent.

The impact of mapping bias depends critically on experimental design. Compare two scenarios where inbred mice from different strains are used. The scenario we consider in this manuscript is a direct comparison between strains, and we show that mapping bias results in biased methylation differences. A different scenario would be to compare two groups of mice, with both groups of mice being balanced among different strains; an example could be to compare young to old mice. In such an experiment mapping bias would not cause biased methylation differences, but would rather cause unnecessary between-sample variation within the two groups, with an associated loss of power. The personalize-then-smooth strategy we advocate here would address both situations, removing bias or decreasing variation.

In future work, it will be important to extend this method to the situation where the samples are outbred. Munger et al. (2014) explored this in the context of allele-specific expression and found that a personal genome strategy was also effective there. But the task of constructing the personal genome is complicated by the fact that each sample's genome is a distinct composition of founder haplotypes. This requires the additional step of inferring founder haplotypes across the subject genomes, which incurs additional computational overhead. Depending on the population under study, it may be possible to use auxillary genetic data to infer this.

In general, we believe that studies in model organisms involving multiple different strains will require the availability of strain specific genomes. This is because such studies are usually undertaken to understand the impact of genotype and therefore involves the comparison of groups of individuals with different (sometimes vastly different) genomes. This is for example the rationale behind the development of the mouse Collaborative Cross (Churchill et al., 2004). *Arabidopsis thaliana* is frequently used for the same purpose, although plants have extensive non-CpG methylation which we have not considered in our work.

In many human studies different groups of interest are composed of different individuals. We know from genome-wide association studies that individuals are rarely random samples from a background distribution, and such studies are therefore susceptible to be affected by mapping bias. But two different humans are genetically closer than the two mouse strains studied here, and in practice the impact of this bias will depend on the genetic heterogeneity of the samples and the size of the signal of interest.

Conclusion

We have shown that mapping bias can severely affect analysis of bisulfite converted DNA. We have proposed a method we call personalize-then-smooth for addressing the mapping bias, which involves alignment to personal genomes followed by smoothing of the methylation data. This method requires the availability of sample specific genomes. Future studies employing bisulfite sequencing need to carefully consider this issue.

Materials and Methods

Sample information

Liver samples from two mouse strains (C57BL/6J and CAST/EiJ, 4 mice per strain) were obtained from Jackson Laboratories. All mice were 6-week-old females; additionally, mice of the same strain were littermates.

DNA extraction and sequencing

Genomic DNA was extracted from liver using the Qiagen DNEasy kit, with an additional RNase incubation step (50 μg/sample, 30 minutes) prior to column application to remove RNA.

WGBS single indexed libraries were generated using the TruSeq DNA LT Sample Preparation Kit (Illumina) according to the manufacturer's instructions with modifications. 1μg gDNA for BL6 samples (1.34μg gDNA for CAST samples due to observed partial DNA degradation) was quantified via Qubit dsDNA BR assay (Invitrogen) and 0.8% agarose gel. 1% Unmethylated lamda DNA (cat#D1521, Promega) was spiked in for monitoring bisulfite conversion efficiency. Samples were fragmented by Covaris S2 sonicator to an average insert size of 350bp (80sec, Duty cycle 10%, Intensity 5, Cycles per burst 200). Size selection was performed using AMPure XP beads and insert sizes of 300-400bp were isolated. Samples were bisulfite converted after size selection using EZ DNA Methylation-Gold Kit (cat#D5005, Zymo) following the manufacturer's instructions. Amplification was performed following bisulfite conversion using Kapa Hifi Uracil+ (cat#KK282, Kapa Biosystems) polymerase and cycling conditions: 98degC 45s /8cycles: 98degC 15s, 65degC 30s, 72degC 30s / 72degC 1 min.

Final libraries were confirmed via 2100 Bioanalyzer (Agilent) High-Sensitivity DNA assay. Libraries were quantified by qPCR using the Library Quantification Kit for Illumina sequencing platforms (cat#KK4824, Kapa Biosystems), using 7900HT Real Time PCR System (Applied Biosystems).

Libraries were sequenced on an Illumina HiSeq2000 sequencer using 100bp paired-end runs with a control lane.

Data availability

Data are available under accession number GSE87101 in NCBI GEO. This includes alignments of the samples to both CAST and BL6 genomes, expressed in BL6 coordinates as well as alignment of the samples to the CAST genome, expressed in CAST coordinates.

Short-read alignment

Alignment and CpG read-level measurement were performed using Bismark version 0.16.1 (Krueger and Andrews, 2011) and Bowtie2 version 2.1.0 (Langmead and Salzberg, 2012). Reference genomes for BL6 and CAST were generated from their corresponding FASTA files (build 37), obtained from UNC Systems Genetics (*UNC Systems Genetics* n.d.). Sequencing reads were first trimmed using Trim Galore! version 0.3.7 (*Trim Galore! website* 2014) using default options, then aligned with Bismark using options `--bowtie2 --bam`. BAM output files were merged and sorted in preparation for methylation extraction using Samtools version 1.3 (Li et al., 2009). Read-level measurements were obtained using the Bismark methylation extractor, with options `-p --ignore 5 --ignore_r2 5 --ignore_3prime 1 --ignore_3prime_r2 1`. Read measurements were modmapped if applicable (see section below), then converted to BSseq objects in R version 3.3.0 using the bsseq package (Hansen, Langmead, and Irizarry, 2012). When creating BSseq objects, forward- and reverse-strand reads were combined for each CpG.

Coordinate mapping between strains

To facilitate direct comparison of CpGs between strains, we converted genomic coordinates using the "modmap" package from UNC Systems Genetics (Huang et al., 2013). This package functions similarly to the liftOver tool (Hinrichs et al., 2006) on the UCSC Genome Browser, and takes as input two files: 1) the user's list of genomic coordinates to be converted; 2) a strain-specific MOD file, again obtainable from UNC, describing how to convert coordinates between strains. We used this package to convert a list of all CpG locations in the CAST genome to their corresponding locations in the BL6 genome, and vice versa. CpGs containing negative positions on either strand in the modmap output (indicating the location resided within an insertion or deletion in the other strain) were discarded.

Genomic mutation rate

The MOD file for CAST lists 20,539,633 bp of single nucleotide variants, 6,633,124 bp of insertions, and 5,279,608 bp of deletions from the BL6 genome to CAST (including autosomes, allosomes, and chrM). The length of the BL6 genome is 2,654,911,517 bp, for an overall genomic mutation rate of (20,539,633 + 6,633,124 + 5,279,608) / 2,654,911,517 = 1.2%.

Methylation analysis

We used the BSmooth pipeline as implemented in the bsseq package from Bioconductor (Hansen, Langmead, and Irizarry, 2012), as employed previously (Hansen, Timp, et al., 2011; Hansen, Sabunciyan, et al., 2014). We only considered CpGs which had a covered of 2 or more in at least 3 out of 4 samples in each group. For our small DMR analysis, the data was smoothed using BSmooth with the following (default) parameters (ns = 70, h = 1,000, maxGap = 10^8). For our large DMR analysis, the data was smoothed using BSmooth with the following (default for this type of analysis) parameters (ns = 500, h = 20,000, maxGap = 10^8). Following smoothing, we used t-statistics to obtain putative differentially methylated regions as described previously, only analyzing CpGs where at least 3 samples in a group had at least a coverage of 2. For small DMRs we employed a t-stat cutoff of at least 4.6 (with a maxGap of 300) and for large DMRs we used a cutoff of 2 (with a maxGap of 10,000). Significance was assessed using a stringent permutation approach as described previously (Hansen, Sabunciyan, et al., 2014). Specifically, we used permutations which balanced the two strains (ie., each permutation has 2 mice from each strain in each group); there are 18 such permutations. For each DMR we calculated how many permutations we saw a better null DMR; dividing by the total number of permutations gives us the quantity we call gFWER. To compare DMRs we are searching for DMRs with many large CpG-specific t-statistics; to be precise we say one DMR is better than another if it has a greater number of CpGs as well as a greater total sum of t-statistics across all CpGs in the DMR. By comparing each putative DMR to all null DMRs in each permutation we control for multiple testing and control the familywise error rate, a stringent multiple testing error rate. The interpretation of a gWER of 1/18 for a given DMR is that in 1 out of 18 permutations do we see a bigger permutation DMR *anywhere* in the genome.

Overlap with functional regions

Regions were obtained and defined as described under external data. Given a set of DMRs as well as a class of regions, we compute the odds ratio of enrichment by considering the overlap in CpGs between the two sets of regions, accounting for the fact that not all CpGs where measured in our data. This approach naturally addresses issues of non-uniform distribution of CpGs.

Genotyping

We used BS-SNPer version 1.0 (Gao et al., 2015) to perform CpG genotyping of CAST samples. BL6 samples were also run as a control. To provide adequate coverage for BS-SNPer, Bismark-aligned BAM files from all four samples from the same strain were merged into a metasample using Samtools and used as input. BS-SNPer was run with options `--minhetfreq 0.1 --minhomfreq 0.85 --minquali 15 --mincover 10 --maxcover 1000 --minread2 2 --errorate 0.02 --mapvalue 20`.

Four main files were produced for each metasample, and used for methylation analysis of BS-SNPer output: a table of single nucleotide variants, and read-level methylation measurements in CG, CHG, and CHH contexts. Variants not marked as passing BS-SNPer's filtering criteria, or with an a frequency lower than 0.5, were excluded from analysis. Surprisingly, BS-SNPer output does not easily include methylation quantification for positions which are not CpGs in the reference genome; we obtained these estimates through looking at the CHG and CHH files, as follows. First, measurements in the CG file that overlapped a C/N variant in the C position or a G/N variant in the G position were counted as lost CpGs and discarded. Second, measurements from the CHG or CHH files that overlapped N/C or N/G variants producing a CpG position were counted as gained CpGs and added to the CG file. Global methylation was then computed from this edited measurement table.

External data

Genomic intervals for ENCODE/LICR histone (*ENCODE/LICR Histones* 2016) and TFBS tracks (*ENCODE/LICR TFBS* 2016), RefSeq genes (*Refseq genes* 2016), and CpG islands (*CpG Islands* 2016) were obtained via the UCSC Genome Browser. Histone and TFBS data were generated by the ENCODE Consortium (ENCODE Project Consortium, 2012) and the Bing Ren laboratory, and are also available at GEO accessions GSE31039 and GSE36027. ENCODE filenames as listed in the UCSC download server are provided in Table 5. Promoter regions of Refseq genes were defined as the 5-kb region flanking a gene's transcription start site. CpG shores were defined as the 2-kb regions upstream and downstream of a CpG island.

Acknowledgments

We thank Birna Berndsen for assisting with library preparation and sequencing.

Funding

Research reported in this publication was supported by National Institute of General Medical Sciences of the National Institutes of Health under award number GM118568 (to BL) and the National Cancer Institute of the National Institutes of Health under award number CA054358 (to APF).

Disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests

The authors declare that they have no competing interests.

Abbreviations

WGBS – whole-genome bisulfite sequencing; DMR – differentially methylated regions.

Bibliography

Barturen, G., A. Rueda, J. L. Oliver, and M. Hackenberg (2013). "MethylExtract: High-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data". *F1000Research* 2, p. 217. DOI: 10.12688/f1000research.2-217.v1.

Bird, A. P. (1980). "DNA methylation and the frequency of CpG in animal DNA". *Nucleic Acids Research* 8, pp. 1499–1504.

Castel, S. E., A. Levy-Moonshine, P. Mohammadi, E. Banks, and T. Lappalainen (2015). "Tools and best practices for data processing in allelic expression analysis". *Genome Biology* 16, p. 195. DOI: 10.1186/s13059-015-0762-6.

Churchill, G. A. et al. (2004). "The Collaborative Cross, a community resource for the genetic analysis of complex traits". *Nature Genetics* 36, pp. 1133–1137. DOI: 10.1038/ng1104-1133.

Coulondre, C., J. H. Miller, P. J. Farabaugh, and W. Gilbert (1978). "Molecular basis of base substitution hotspots in Escherichia coli". *Nature* 274, pp. 775–780.

CpG Islands (2016). https://genome.ucsc.edu/cgi-bin/hgTrackUi?g=cpgIslandSuper. (Visited on 07/15/2016).

Degner, J. F., J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, and J. K. Pritchard (2009). "Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data". *Bioinformatics* 25, pp. 3207–3212. DOI: 10.1093/bioinformatics/btp579.

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". *Nature* 489, pp. 57–74. DOI: 10.1038/nature11247.

ENCODE/LICR Histones (2016). https://genome.ucsc.edu/cgi-bin/hgTrackUi?g=wgEncodeLicrHistone. (Visited on 09/13/2016).

ENCODE/LICR TFBS (2016). https://genome.ucsc.edu/cgi-bin/hgTrackUi?g=wgEncodeLicrTfbs. (Visited on 09/13/2016).

Feinberg, A. P. and B. Vogelstein (1983). "Hypomethylation distinguishes genes of some human cancers from their normal counterparts". *Nature* 301, pp. 89–92.

Feinberg, A. P., M. A. Koldobskiy, and A. Göndör (2016). "Epigenetic modulators, modifiers and mediators in cancer aetiology and progression". *Nature Reviews Genetics* 17, pp. 284–299. DOI: 10.1038/nrg.2016.13.

Gao, S., D. Zou, L. Mao, H. Liu, P. Song, Y. Chen, S. Zhao, C. Gao, X. Li, Z. Gao, X. Fang, H. Yang, T. F. Ørntoft, K. D. Sørensen, and L. Bolund (2015). "BS-SNPer: SNP calling in bisulfite-seq data". *Bioinformatics* 31, pp. 4006–4008. DOI: 10.1093/bioinformatics/btv507.

Geijn, B. van de, G. McVicker, Y. Gilad, and J. K. Pritchard (2015). "WASP: allele-specific software for robust molecular quantitative trait locus discovery". *Nature Methods* 12, pp. 1061–1063. DOI: 10.1038/nmeth.3582.

Hansen, K. D., B. Langmead, and R. A. Irizarry (2012). "BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions." *Genome Biology* 13, R83. DOI: 10.1186/gb-2012-13-10-r83.

Hansen, K. D., S. Sabunciyan, B. Langmead, N. Nagy, R. Curley, G. Klein, E. Klein, D. Salamon, and A. P. Feinberg (2014). “Large-scale hypomethylated blocks associated with Epstein-Barr virus-induced B-cell immortalization.” *Genome Research* 24, pp. 177–184. DOI: 10.1101/gr.157743.113.

Hansen, K. D., W. Timp, H. C. Bravo, S. Sabunciyan, B. Langmead, O. G. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, E. Briem, K. Zhang, R. A. Irizarry, and A. P. Feinberg (2011). “Increased methylation variation in epigenetic domains across cancer types.” *Nature Genetics* 43, pp. 768–775. DOI: 10.1038/ng.865.

Hinrichs, A. S., D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent (2006). “The UCSC Genome Browser Database: update 2006.” *Nucleic Acids Research* 34, pp. D590–8. DOI: 10.1093/nar/gkj144.

Hodgkinson, A., J. C. Grenier, E. Gbeha, and P. Awadalla (2016). “A haplotype-based normalization technique for the analysis and detection of allele specific expression”. *BMC Bioinformatics* 17.1, p. 364.

Hodgkinson, A. and A. Eyre-Walker (2011). “Variation in the mutation rate across mammalian genomes”. *Nature Reviews Genetics* 12.11, pp. 756–766. DOI: 10.1038/nrg3098.

Huang, S., C.-Y. Kao, L. McMillan, and W. Wang (2013). “Transforming Genomes Using MOD Files with Applications”. *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*. BCB’13. New York, New York, USA: ACM, pp. 595–604. DOI: 10.1145/2506583.2506643.

Krueger, F. and S. R. Andrews (2011). “Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications”. *Bioinformatics* 27, pp. 1571–1572. DOI: 10.1093/bioinformatics/btr167.

Langmead, B. and S. L. Salzberg (2012). “Fast gapped-read alignment with Bowtie 2.” *Nature Methods* 9, pp. 357–359. DOI: 10.1038/nmeth.1923.

Lappalainen, T., M. Sammeth, M. R. Friedländer, P. A. C. ’t Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. J. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, Geuvadis Consortium, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Häsler, A.-C. Syvänen, G.-J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, and E. T. Dermitzakis (2013). “Transcriptome and genome sequencing uncovers functional variation in humans”. *Nature* 501, pp. 506–511. DOI: 10.1038/nature12531.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup (2009). “The Sequence Alignment/Map format and SAMtools". *Bioinformatics* 25, pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.

Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker (2009). "Human DNA methylomes at base resolution show widespread epigenomic differences". *Nature* 462, pp. 315–322. DOI: 10.1038/nature08514.

Liu, Y., K. D. Siegmund, P. W. Laird, and B. P. Berman (2012). "Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data". *Genome biology* 13, R61. DOI: 10.1186/gb-2012-13-7-r61.

Munger, S. C., N. Raghupathy, K. Choi, A. K. Simons, D. M. Gatti, D. A. Hinerfeld, K. L. Svenson, M. P. Keller, A. D. Attie, M. A. Hibbs, J. H. Graber, E. J. Chesler, and G. A. Churchill (2014). "RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations". *Genetics* 198.1, pp. 59–73.

Panousis, N. I., M. Gutierrez-Arcelus, E. T. Dermitzakis, and T. Lappalainen (2014). "Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies". *Genome Biology* 15, p. 1217. DOI: 10.1186/s13059-014-0467-2.

Refseq genes (2016). https://genome.ucsc.edu/cgi-bin/hgTrackUi?g=refGene. (Visited on 07/14/2016).

Rozowsky, J., A. Abyzov, J. Wang, P. Alves, D. Raha, A. Harmanci, J. Leng, R. Bjornson, Y. Kong, N. Kitabayashi, N. Bhardwaj, M. Rubin, M. Snyder, and M. Gerstein (2011). "AlleleSeq: analysis of allele-specific expression and binding in a network framework". *Molecular Systems Biology* 7, pp. 522–522. DOI: 10.1038/msb.2011.54.

Satya, R. V., N. Zavaljevski, and J. Reifman (2012). "A new strategy to reduce allelic bias in RNA-Seq readmapping". *Nucleic Acids Research* 40, e127. DOI: 10.1093/nar/gks425.

Trim Galore! website (2014). http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. (Visited on 07/16/2014).

UNC Systems Genetics. http://csbio.unc.edu/CCstatus/index.py?run=Pseudo.

Tables

Table 1. Number of reads and alignment statistics.

		BL6			CAST		
Sample	**nReads**	**nAligned**	**aRate**[a]	**CpG**[b]	**nAligned**	**aRate**[a]	**CpG**[b]
BL6_1	165,455,305	92,198,446	55.7	5.8	81,398,771	49.2	4.7
BL6_2	155,446,345	85,632,105	55.1	5.5	75,566,778	48.6	4.5
BL6_3	165,687,191	99,880547	60.3	6.8	87,713,251	52.9	5.5
BL6_4	172,926,402	101,892159	58.9	6.7	89,939,416	52.0	5.5
CAST_1	171,357,014	107,621,399	62.8	6.3	121,678,485	71.0	6.4
CAST_2	161,768,892	90,231,861	55.8	6.1	102,071,949	63.1	6.3
CAST_3	154,973,730	86,822,723	56.0	5.5	98,305,382	63.4	5.7
CAST_4	188,134,260	90,420,870	48.1	6.0	102,599,995	54.5	6.2

[a] alignment rate
[b] coverage of CpGs

Table 2. Number of CpGs in different strains.

	Autosomes		Allosomes	
	BL6	**CAST**	**BL6**	**CAST**
Common CG	18,140,628		916,392	
TG/CA in other strain	1,601,446	1,669,901	43,330	45,041
Lost in other strain	522,845	535,219	15,394	15,838
Unmappable	99,781	90,356	2,676	2,438
Total	20,364,700	20,436,104	977,792	979,709

Table 3. Enrichment of strain-specific DMRs in genomic features.

Feature	n	$\log_2(\mathrm{OR})$	*p*-value
Refseq promoters	197	0.87	$< 2.2 \times 10^{-16}$
CpG Islands	61	0.31	$< 3.3 \times 10^{-12}$
CpG Shores	157	1.18	$< 2.2 \times 10^{-16}$
H3K4me1	625	3.06	$< 2.2 \times 10^{-16}$
H3K4me3	182	1.86	$< 2.2 \times 10^{-16}$
H3K27ac	268	1.97	$< 2.2 \times 10^{-16}$
CTCF	136	1.64	$< 2.2 \times 10^{-16}$
Pol2	53	0.02	0.66
Any feature above	750	3.03	$< 2.2 \times 10^{-16}$

Table 4. BS-SNP results on pooled CAST samples.

BL6-unique CpGs	2,285,477
BL6-unique CpGs identified by BS-SNPer	542,642
CAST-unique CpGs	2,266,009
CAST-unique CpGs identified by BS-SNPer	1,153,611

Table 5. Filenames for ENCODE data.

Filename
`wgEncodeLicrHistoneLiverH3k27acMAdult8wksC57bl6StdPk.broadPeak`
`wgEncodeLicrHistoneLiverH3k4me1MAdult8wksC57bl6StdPk.broadPeak`
`wgEncodeLicrHistoneLiverH3k4me3MAdult8wksC57bl6StdPk.broadPeak`
`wgEncodeLicrTfbsLiverCtcfMAdult8wksC57bl6StdPk.broadPeak`
`wgEncodeLicrTfbsLiverPol2MAdult8wksC57bl6StdPk.broadPeak`

Figures

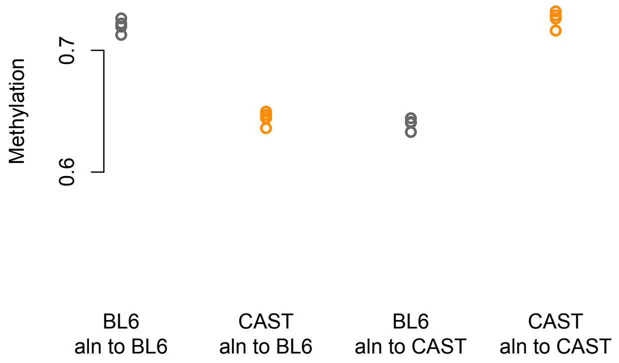


Figure 1. Global methylation. Data from each strain was aligned to two different reference genomes and global methylation across the autosomes (the average methylation across all CpGs) was computed. The two strains appear to have different levels of global methylation when both strains are aligned to the same genome, but which strain has lowest methylation depends on which genome the samples were aligned to.

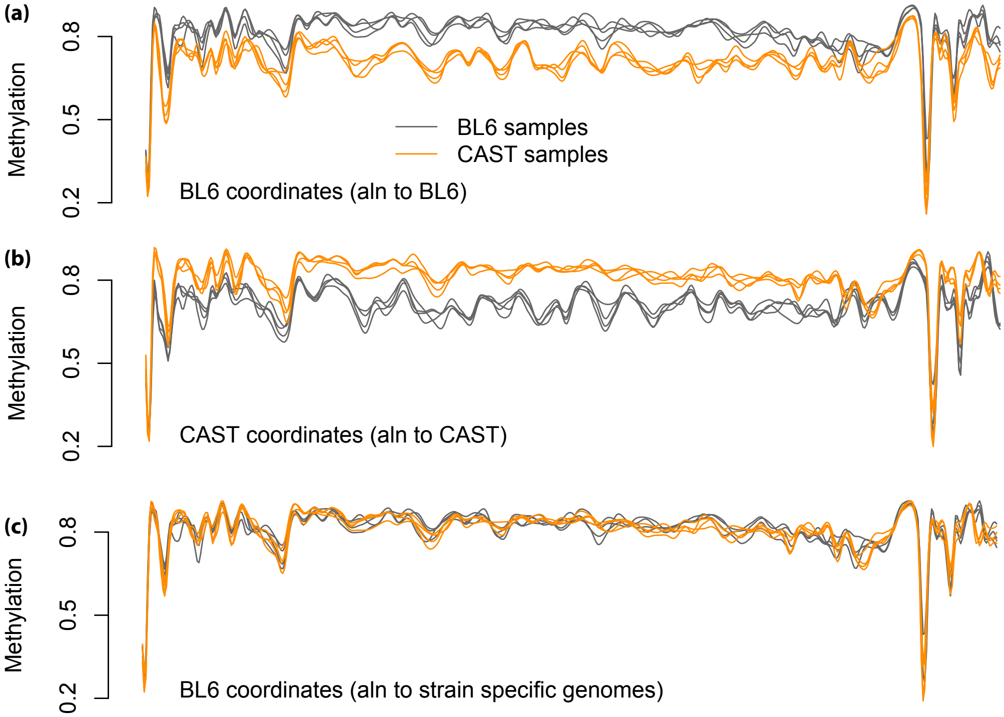


Figure 2. Mapping biases causes apparent large-scale methylation changes at the megabase level. The same 2.4 Mb genomic region is depicted in the coordinate systems of different genomes, and with different data processing. **(a,b)** The same samples were mapped to either **(a)** the BL6 or **(b)** the CAST genome. **(c)** Samples were mapped to strain-specific genomes, CpGs were put in a common coordinate system (BL6) and subsequently smoothed (Methods).

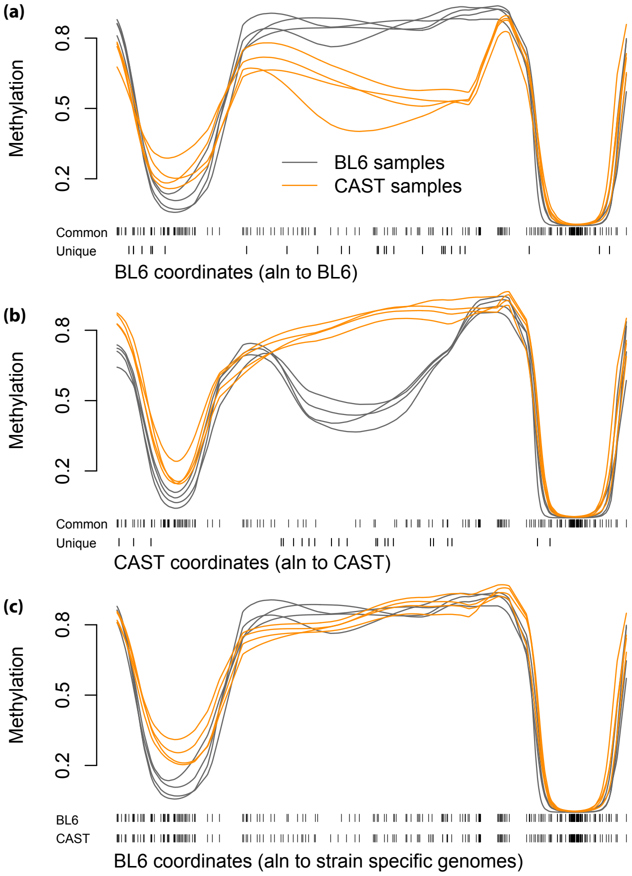
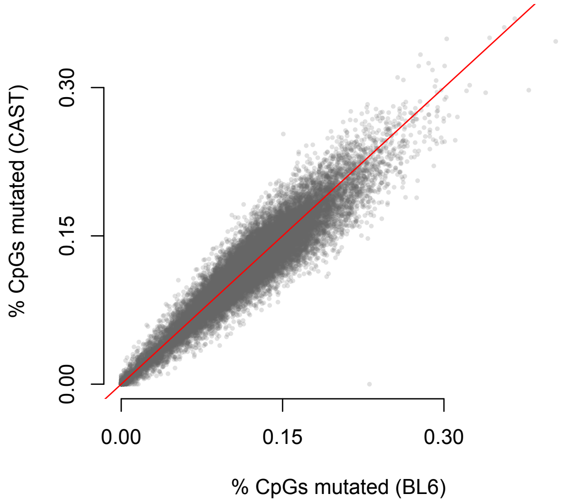
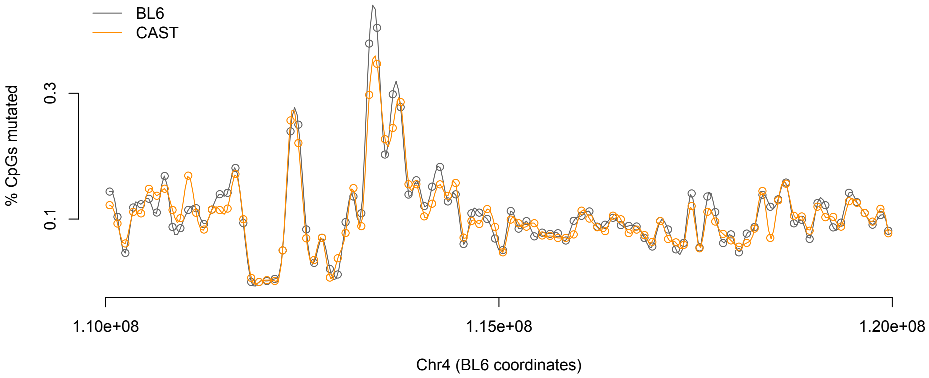


Figure 3. Mapping biases causes apparent focal methylation changes at the kilobase level. The same 13 kb genomic region is depicted in the coordinate systems of different genomes, and with different data processing. **(a,b)** The same samples were mapped to either **(a)** the BL6 or **(b)** the CAST genome, followed by smoothing. Ticks indicate CpGs either common to the two genomes (upward ticks) or unique to the genome used for mapping (downward ticks). **(c)** Samples were mapped to strain-specific genomes, CpGs were put in a common coordinate system (BL6) and subsequently smoothed (Methods). Ticks indicate CpGs in BL6 (upward ticks) or CAST (downward ticks).



(a)



(b)

Figure 4. The distributions of strain-specific CpG mutations. (a) For each 100kb bin in the mouse genome, we computed the percentage of strain-specific CpGs relative to the total number of CpGs in that strain. **(b)** Same data as in (a) but plotted across a 10 Mb region on chromosome 4.

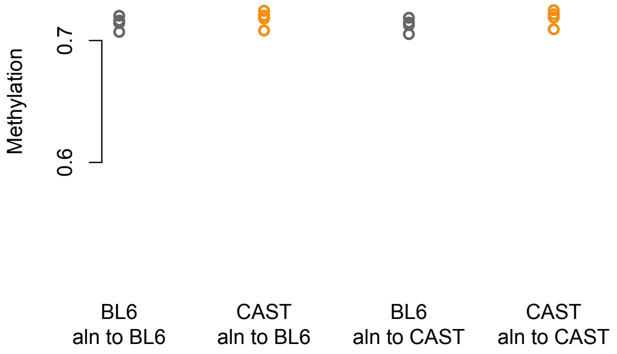
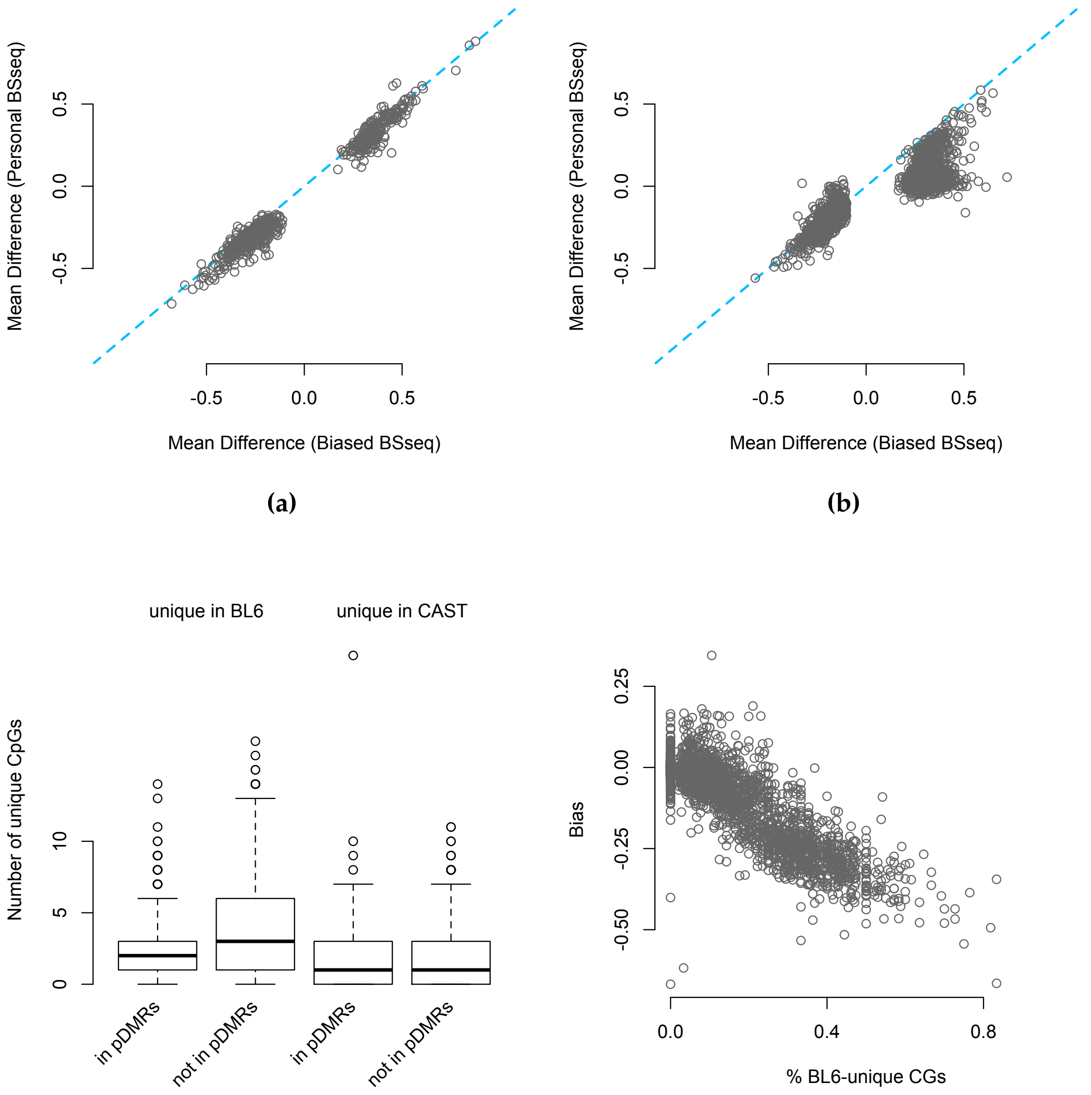


Figure 5. Global methylated computed on CpGs present in both strains. Like Figure 1 but computed using only the 19 million CpGs present in both strains. There is now little to no difference between the different strains and the different genomes used for alignment.



(a)

(b)

(c)

(d)

Figure 6. The effect of using personal genomes on methylation differences. (a,b) For each of 2,865 DMRs obtained from a biased analysis where all samples were aligned to BL6, we computed methylation differences across the region using the biased analysis as well as alignment to personal genomes. We compare these estimates, stratified by whether the DMR is also found using personal genomes (a) or not (b). **(c)** Number of unique CpGs in the two strain genomes inside biased DMRs, stratified by whether the DMR is also found using personal genomes or not. **(d)** Bias in methylation (defined as difference between the mean difference obtained by the two alignment strategies) as a function of the number of BL6 unique CpGs in the DMR.

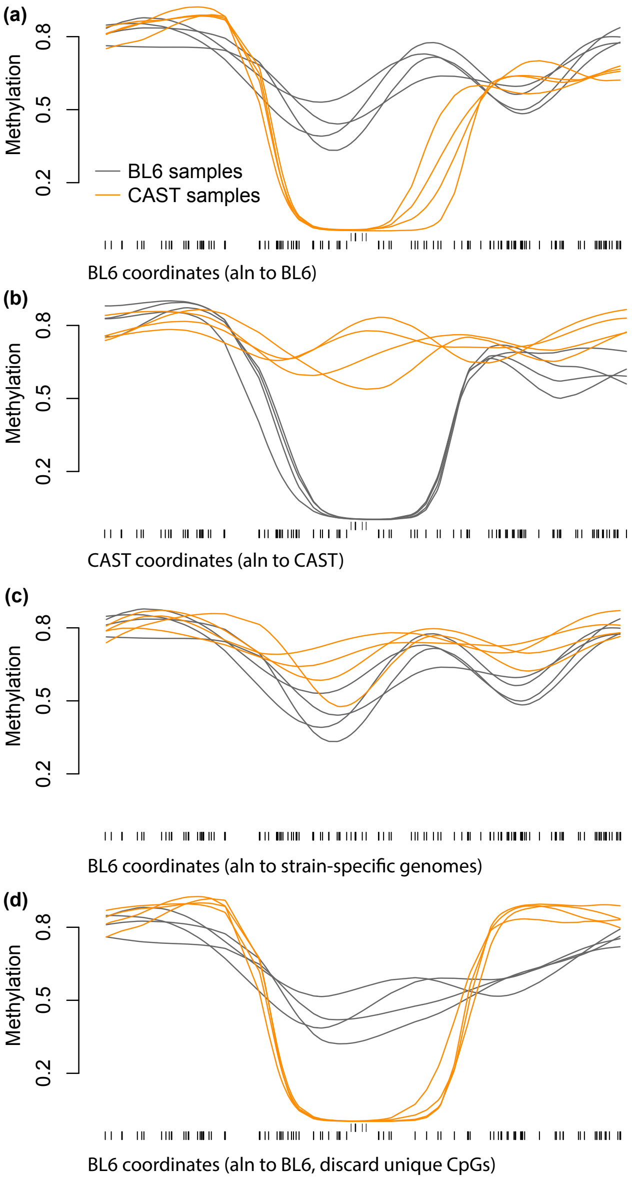


Figure 7. An example of alignment bias not caused by CpG mutations. A 11kb genomic region containing no CpGs unique to either of the two mouse strains, depicted using different alignment strategies. **(a)** Alignment to the BL6 genome. **(b)** Alignment to the CAST genome. **(c)** Alignment to personal genomes. **(d)** Alignment to the BL6 genome, followed by discarding any CpGs which are unqiue to either of the strains. Tickmarks indicate the position of CpGs; upwards ticks marks are CpGs with average coverage greater than 100.

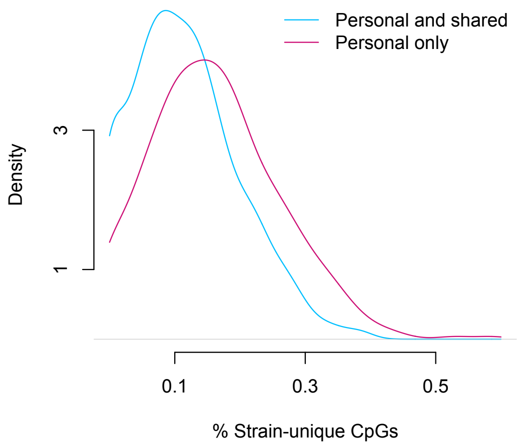


Figure 8. The distribution of strain unique CpGs in personal DMRs. Personal DMRs were split into two groups depending on whether they overlap shared DMRs or not. Depicted is the distribution of strain unique CpGs in a DMR, in the two groups.