Contrasting DNA methylation platforms: technical report for EU FP7 RADIANT deliverable 10.2

Mirco Menigatti, Charity Law, Mark D. Robinson (UZH)

DNA methylation (DNAme), the heritable addition of a methyl group to cytosine residues, is a vital component of the epigenetic regulatory machinery, and plays a role in maintaining genome stability, imprinting, and is known to be deregulated in cancer and other diseases (Laird, 2010). There is wide interest in profiling DNAme in patient cohorts for many purposes: identifying genes whose epigenetic state are commonly interrupted in cancer, the discovery of markers that relate to disease prognosis and to associate DNAme aberrations with progression of disease.

Present techniques for interrogating DNAme fall into three categories: methylation-specific enzyme digestion (ED), affinity enrichment (AE) and chemical treatment with bisulphite (BS), in combination with a microarray or high-throughput sequencing readout; some of these techniques have been used in combination (e.g. ED+BS, commonly known as RRBS; see (Laird, 2010)). There is also wide interest in distinguishing DNAme in the process of sequencing itself. DNAme information comes at low (~100-200 base pair) or high resolution (individual CpG sites) and the costs vary widely. Each platform has its own limitations (Bock, 2012; Laird, 2010; Robinson, Statham, Speed, & Clark, 2010). For example, ED studies remain at the mercy of the location and number of restriction enzyme sites, the sensitivity of AE approaches depends on CpG density and comprehensive sequencing-based BS methods are costly and require significant computing infrastructure. Furthermore, many research groups establish a method in their laboratory with limited guidance of what platform to choose, and due to the resource investment, are not easily convinced to change.

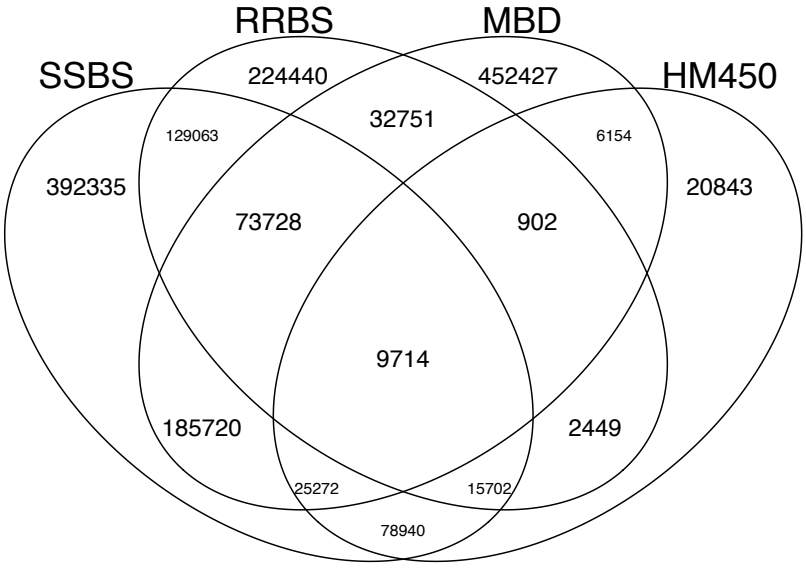
Here, we aim to give ourselves and other researchers the basis for choosing a DNAme profiling approach, bearing in mind that the scientific aim of the study in an important component of that decision.

The table below gives a concise summary of the popular methods, with some of the advantages and disadvantages of each. For example, BS-based methods are considered to be the gold standard, because alleles from the cell population of interest are interrogated at base resolution and as a byproduct, can reveal information about copy number variation (CNV) and single nucleotide polymorphisms (SNP). However, these methods are also cost limiting. In the case of whole genome BS sequencing (WGBS), this is akin to sequencing an entire human genome, but with a lot of degeneracy introduced because of the BS

treatment. So, typically to get a comprehensive view of DNAme status, higher coverage than standard genome sequencing is required. However, in terms of maximizing the sequencing efficiency, some have noted that there is more to be gained by lower coverage sequencing across *replicates* (Hansen et al., 2011). Compared to untreated genome sequencing, BS-treated is largely devoid of cytosines (BS converts unmethylated cytosines to uracil, which are read as thymine after PCR) and mapping algorithms typically remove cytosines from both the reads and the reference genome *in silico*. As a result, degeneracy (3-base genome instead of 4-base genome) is introduced and mappability is decreased in many regions of the genome. Therefore, researchers are encouraged to collect long and/or paired-end reads where possible, noting that this may add to the overall cost. Researchers should also note BS treatment of DNA does not discriminate between methylcytosine and hydroxymethylcytosine. An intermediate solution is to capture regions of interest before (BS treatment and) sequencing, analogous to exome sequencing as opposed to full genome sequencing. Instead of selecting coding regions, researchers can select regulatory regions (e.g. CpG island) according to either the standard kits available (e.g. Agilent SureSelect MethylSeq) or according to a custom capture. This allows the researcher to focus the sequencing capacity on regions of the genome that are of interest. An entirely separate class of methods to profile methylation is analogous to chromatin immunoprecipitation enrichment, whereby fragments that are methylated (methyl or hydroxymethyl, depending on how they are captured) are selected, sequenced and mapped back to a reference genome for identification. Unfortunately, these assays are very sensitive to the method used for capture. We recently showed that this can be accurately normalized by running a fully methylated sample on the same capture platform (Riebler et al., 2014). In this framework, other features, such as CNV can be readily accounted for. In early 2011, Illumina have released an updated beadchip (an extension to the HumanMethylation27 array) that gives single base resolution at 450,000 CpG (effectively ~400,000 sites after filtering multimapping probes and those with SNPs in their body; e.g. (Price et al., 2013)) sites for a modest cost. While this represents “low” coverage (~2%) of the over 28 million human CpG sites, there is considerable coverage spanning CpG islands, CpG shores, and sites around known cancer-related genes. In addition, the platform is cost-effective and readily scales to larger studies (e.g. epigenome-wide association studies), which the majority of other methods do not.

Method	Advantages	Disadvantages	Resolution
Whole genome BS	Highest coverage, SNPs and CNV as bonus, single molecule	$$$$ (inefficient), serious computing resources, does not distinguish hydroxymethylation	1bp
Reduced representation / SureSelect BS	All advantages of whole genome BS with better efficiency, single molecule, choice over regions covered (custom capture or choice of restriction enzyme)	Laborious capture protocol, additional steps, medium computing resources	1bp
Affinity Capture (MeDIP/MBD)	$, efficient (lower depth, shorter reads required), data capture (especially combined w/ SssI control)	Low resolution, bias in capture efficiency (e.g. CpG density dependent)	~100bp
450k array	$, easy, scales to large numbers of samples, data processing is straightforward, CNV as bonus	"Low" coverage	1bp

A recent study, using 42 whole-genome BS sequencing datasets, highlighted that only a modest proportion of CpG sites (~25% or 7M CpG sites) are regulated or exhibit changes across 30 diverse cell types (Ziller et al., 2013). In particular, see Figure 1b of the Ziller et al. manuscript, which shows that the "CpG-wise DNAme level differences", the degree to which the CpG changes across cell types, stays near zero for a large proportion of CpG sites. However, it required the very expensive WGBS to know this. As an analysis of interest, we compare a set of the popular available platforms for their ability to capture exactly these regions. At this stage, we do not consider the *sensitivity* of such platforms, as this depends on many factors, such as sequencing depth, the purity of the cell populations, the quality of the input DNA and so on. Of the over 7M



CpG sites highlighted in the Ziller et al. manuscript to be *variable* across various cell types, the Venn diagram gives the number of sites that are assayed by the various platforms as well as their overlaps with other platforms. The table below highlights that a large amount of the changing CpG sites are missed by using platforms other than WGBS. Interestingly, the targeted BS sequencing (denoted as SSBS) approach provides the best coverage of the changing CpG sites. Notably, this platform is much more efficient than WGBS and could be designed to get maximal coverage of these regions, within the limits of the capture technology (e.g. probe design). Only 28.7% of the current real estate (default capture design) is dedicated to capturing the changing regions. AE methods, such as MBD-seq (denoted as MBD) can be done at a much lower cost because the depth of sequencing is not required to be very high. However, the platform is only able to interrogate ~10% of the desired sites. Reduced representation methods (RRBS) and the Illumina 450k array (HM450) are able to interrogate even less of this of these regions of interest.

Method	**CpG sites covered**	**% of Ziller regions covered**	**% of platforms' CpG coverage**
SSBS (SureSelect)	910,474	12.3	28.7
MBD	786,668	10.6	33.7
RRBS	488,749	6.6	23.8
HM450	159,976	2.2	33.2

References:

Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*. doi:10.1038/nrg3273

Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., … Feinberg, A. P. (2011). Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, (June). doi:10.1038/ng.865

Laird, P. W. (2010). Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, *11*(3), 191–203. doi:10.1038/nrg2732

Price, M. E., Cotton, A. M., Lam, L. L., Farré, P., Emberly, E., Brown, C. J., … Kobor, M. S. (2013). Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics & Chromatin*, *6*(1), 4. doi:10.1186/1756-8935-6-4

Riebler, A., Menigatti, M., Song, J. Z., Statham, A. L., Stirzaker, C., Mahmud, N., … Robinson, M. D. (2014). BayMeth: improved DNA methylation quantification for affinity capture sequencing data using a flexible Bayesian approach. *Genome Biology*, *15*(2), R35. doi:10.1186/gb-2014-15-2-r35

Robinson, M. D., Statham, A. L., Speed, T. P., & Clark, S. J. (2010). Protocol matters : which methylome are you actually studying ? *Epigenomics*, *2*(4), 587–598. doi:10.2217/epi.10.36

Ziller, M. J., Gu, H., Muller, F., Donaghey, J., Tsai, L. T., Kohlbacher, O., … Meissner, A. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature*, *500*(7463), 477–481. doi:10.1038/nature12433