Chapter 3

Development and validation of an array-based assay for the identification of DMRs

3.1 Introduction

The availability of genomic sequences of various organisms, including human, has provided an important resource in the effort of understanding biological functions. This resource has been exploited extensively by micro-array technology, including the development of DNA tiling arrays (Bertone et al., 2005; Hoheisel, 2006; Mockler et al., 2005; Yazaki et al., 2007). DNA tiling arrays represent a complete tile path of non-repetitive DNA over a locus, complex, chromosome or entire genome at various sequence resolutions. The exclusion of repetitive elements and other non-unique sequences from tiling array designs aims to reduce non-specific background signals that mask the signal resulting from specific probe hybridization. Probes used for the construction of such arrays can be partially overlapping, tiled end to end or may be spaced at regular intervals. DNA tiling arrays have enabled the discovery of novel transcribed sequences and transcription factor binding sites and, during the past few years, they have led the way for DNA methylation profiling.

DNA methylation analysis techniques, including bisulphite conversion, methylation sensitive restriction and immunoprecipitation (reviewed in (Beck and Rakyan, 2008; Weber and Schubeler, 2007; Zilberman and Henikoff, 2007) have been adapted successfully to be used in combination with DNA tilling arrays (Illingworth et al., 2008; Keshet et al., 2006; Mohn et al., 2008; Rakyan et al., 2008; Rakyan, 2008; Weber et al., 2005; Weber et al., 2007; Zhang et al., 2008; Zhang et al., 2006; Zilberman et al., 2007). The first array-based methylome has already been reported for *Arabidopsis thaliana* (Zhang et al., 2008; Zhang et al., 2006), and recently, by using Nimblegen array platforms, comprehensive genome-wide DNA methylation data have been generated for several human tissues (Rakyan et al., 2008).

My objective was to develop an unbiased tiling array-based assay for the identification of differentially methylated regions (DMRs) within the MHC region. DMRs refer to temporal or spatial patterns of DNA methylation and can be indicative of local changes in genome functionality (see section 1.3.7). To this end:

(i). I constructed and validated a tiling path micro-array covering the MHC region on chromosome 6.

(ii). I optimised and validated a protocol for the immunoprecipitation of methylated DNA fractions (MeDIP).

(iii). I tested the application of the MHC tiling array for DMR identification in combination with MeDIP.

3.2 MHC tiling array

3.2.1 Chemistry of the MHC tiling array.

The construction of the MHC tiling array was based on the 5'amino-link array surface chemistry developed at the Wellcome Trust Sanger Institute (Dhami et al., 2005) which allows single strands of DNA derived from double stranded PCR products to be retained on the surface of the micro-array slide. This is accomplished by the incorporation of a 5'-amino-linked modification at the end of one strand of a double stranded PCR product using modified primers (either forward or reverse). The 5'-amino-linked modification facilitates a covalent bond between the modified strand and the surface of the slide. Upon slide processing, the strand attached to the slide is retained, whereas the unmodified strand is removed. The single stranded DNA molecules attached at one end of the surface of the slide provide an ideal substrate for hybridization with labelled DNA samples (figure 3.1).

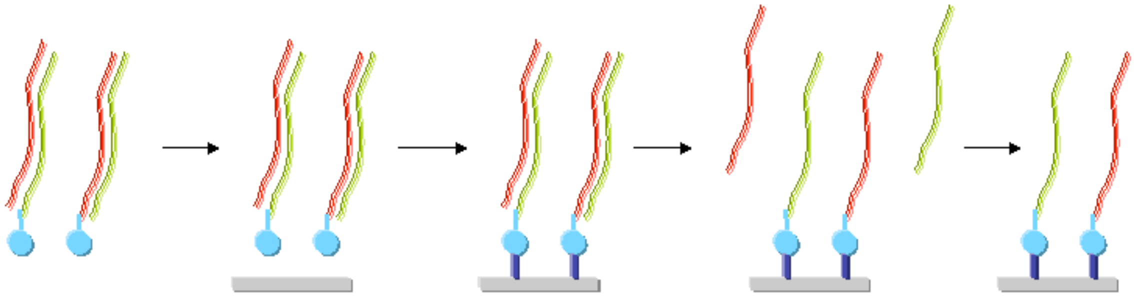

Figure 3.1. **Diagrammatic representation of processing of single-stranded array probes.** Double-stranded PCR products (red/green denote forward and reverse strands respectively) containing a 5'-amino-linked modification on one strand (blue sphere) are arrayed onto the surface of the slide (grey bar). Covalent attachment occurs via the 5'

amino-link (dark blue line) and the slide surface. Denaturation of the PCR product renders them single-stranded and suitable as hybridization substrates.

3.2.2 Generation and quality control of MHC array probes

The probes were designed to cover the entire MHC region as a minimally overlapping tile path, with appropriate controls. For this purpose I used the freely available clones from the HapMap project (The International HapMap Project, 2003). A total of 1747 overlapping plasmid clones were used to generate the array. Of those, 1662 clones (average insert size 2 kb) were picked from the HapMap chromosome 6 library and 85 clones were generated by cloning gap-spanning PCR amplicons (average insert size 332 bp). Some repeat-rich regions (about 12 kb in total) proved to be refractory to PCR amplification, and are thus missing from the array. Therefore, the total coverage represents 99.67% of the MHC region. It should be noted that at the time of the array design, the average probe size (2kb) was adequate as a comparable resolution was also used (e.g. the ENCODE project; The ENCODE project 2004). The main advantage of using the clones from the HapMap project was that all probes could be generated with the same primer set (section 2.2.9), making the process of array construction reproducible and economical.

In addition, I generated and included 43 PCR-derived clones as controls, covering: i) CpG islands of *BRCA1, GSTP1, RARB2* and *MLH1* genes as controls for studying methyl-binding domain (MBD) proteins (Ballestar et al., 2003), ii) imprinted regions (*H19, IGF2, KvDMR1, HSIGF2G, IGF2RDMR2* and *DMR0*) (Lewis and Murrell, 2004), as controls for studying imprinted regions (work in progress by Adele Murrell) iii) gene poor regions of chromosome 6; iv) matrix attachment regions of the *β-globin* gene cluster (Ottaviani et al., 2008); v) loop-associated DNA of the *PRM2* gene; vi) promoter regions of the *GAPDH* and *IRF1* genes; vii) replication origin of the *LB2* gene; vii) replication origin-lacking region of the *β-globin* locus; and viii) DNAase I-hypersensitivity sites of the *β-globin* locus control region. Controls iii to viii were used to study higher-order structure such as matrix attachment regions (MARs) within the MHC (Ottaviani D., et al., 2008) Ten genes from the *Arabidopsis thaliana* genome (spotted in replicates, distributed across the array) that can be used to assign DNA barcodes as internal controls were also included. In addition, 192 Cy3 spots were printed on each array that can be used for calibration and orientation. MHC probe coordinates and primer sets used for the generation of gap-spanning and control clones are provided in appendix tables 2.1 and 2.2. Except for the Cy3 spots, none of the other controls were used for the analysis described in this thesis, but may be useful for other types of analyses.

In order to generate strand-specific array probes, two separate PCR reactions were performed for each clone using universal M13 primers: one reaction using a 5'-amino-linked primer for the forward strand, and one using a 5'-amino-linked primer for the reverse strand.

The array probes were assessed visually by agarose gel electrophoresis (figure 3.2).

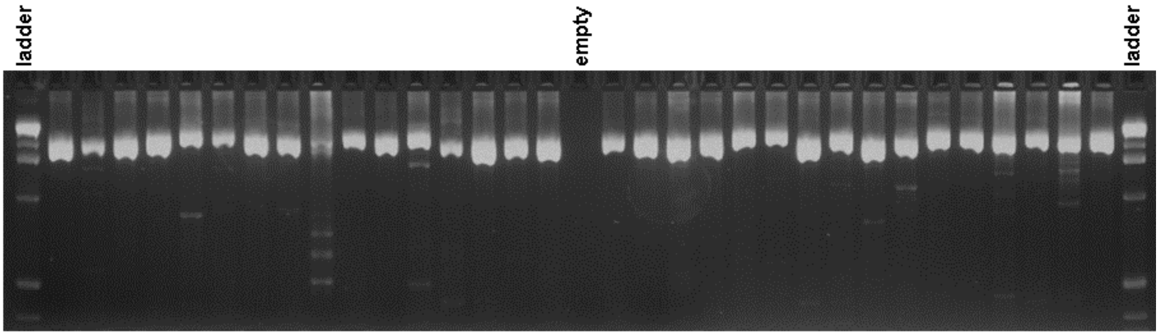

Figure 3.2. **Quality control of PCR-amplified probes.** All probes were electrophoresed on agarose gels and their bands were scored visually. Probes were electrophoresed on 2.5% agarose 1xTBE gels and visualized with ethidium bromide (a representative gel is shown).

Finally, the identity of randomly selected 240 clones (15 % of total) was confirmed by re-sequencing. Of the clones tested 7 failed to match to the expected reference sequences. From this partial analysis, I extrapolate that about 97% of the probes are correct and should be informative. Table 3.1 summarizes the characteristics of the array probes.

		No. of probes spotted on the array		
	No. of MHCclones	forward	reverse	**total**
No of MHC probes	1747	3494	3494	6988
No. of control probes	43	86	86	172
Cy3 spots				192
Arabidopsis regions				480
Total No. of array probes				**7832**

Table 3.1. **Summary of MHC tiling array probes**. The table lists the total number of probes used for the array construction. Probes were validated visually by agarose gel electrophoresis and by re-sequencing. In total 7832 probes were spotted on the array.

3.2.3 Validation of the MHC tiling array

In total 7832 probes were spotted to produce the 2kb MHC tile path array (table 3.1). Initial validation of the array was performed directly after spotting. The quality of array spots was tested using Cy3 dye and only those arrays that had less than 2% of the total spots not fluorescing or merged were used further.

The quality of array probes was further validated by performing input to input hybridization. Fragmented genomic DNA (average size 300 – 1000 bp) was labeled with Cy3 and Cy5 dyes and hybridized on the MHC tiling array. Calculation of $\log_2$ ratios (Cy3/Cy5) showed values very close to zero (figure 3.3a) indicating absence of hybridization variation within the array probes. On the other hand, when MeDIP to input hybridization was performed (see below) the range of $\log_2$ ratio was between 2 and -2 (figure 3.3b), as expected.

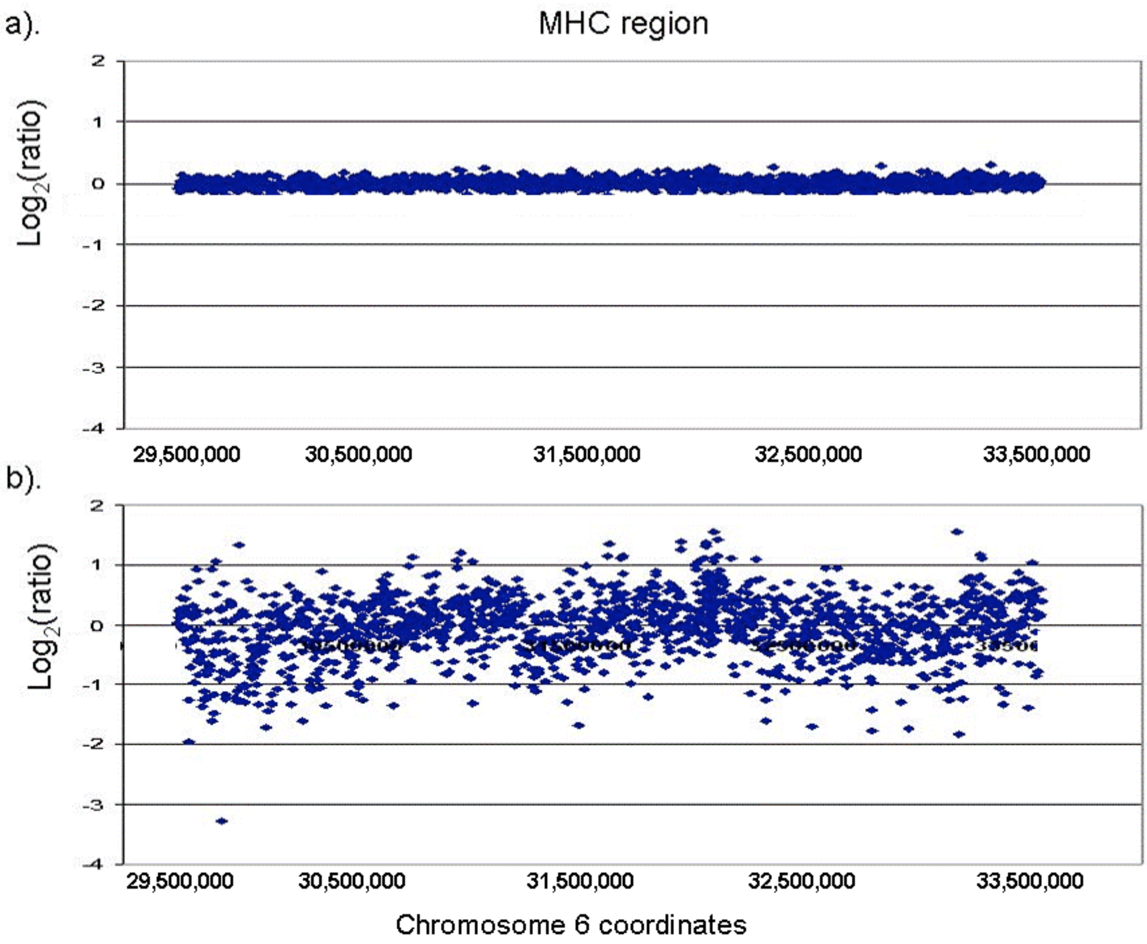

Figure 3.3. **Hybridization variation.** a) Input versus input hybridization of sperm DNA. b). MeDIP versus input hybridization on sperm DNA. Plotted are $\log_2$-transformed hybridization ratios against the linear map position of the MHC probes. Log_2 ratios of input to input hybridization are close to zero whereas ratios corresponding to MeDIP to input hybridization range from 2 to -2.

3.2.4 Repetitive elements

Compared to most commercial and custom tiling arrays, the MHC tiling array also contains repeat elements, allowing such sequences to be analysed as well if desired. Figure 3.4a shows the distribution and frequency of repeat sequences within the probes on the array. About 9% of the probes have low (0-5%) repeat content and around 11% have high (95-100%) repeat content. The majority (80%) of probes have a random repeat content ranging from 6-94%. For studies that are not designed to interrogate repeat sequences (as in the study presented here) I show that repeat sequences can be efficiently blocked by the addition of human Cot-1 DNA during hybridization (Figure 3.4b). Human Cot-1 DNA is commonly used to block non-specific hybridization in microarray screening. It is derived from placental DNA, about 50 to 300 bp in size and is enriched for repetitive DNA sequences such as the *Alu* and *Kpn* family members (Marx et al., 1976). I compared the probe intensities of the Cy5 channel for two hybridizations, one with and the other without Cot-1 DNA. In the presence of Cot-1 DNA, the intensities of repeat-containing probes are clearly reduced to the same level detected for repeat-free probes, indicating that undesired repeat signals can be blocked and that the unique parts of repeat-containing probes remain to be informative and can be kept for further analysis.

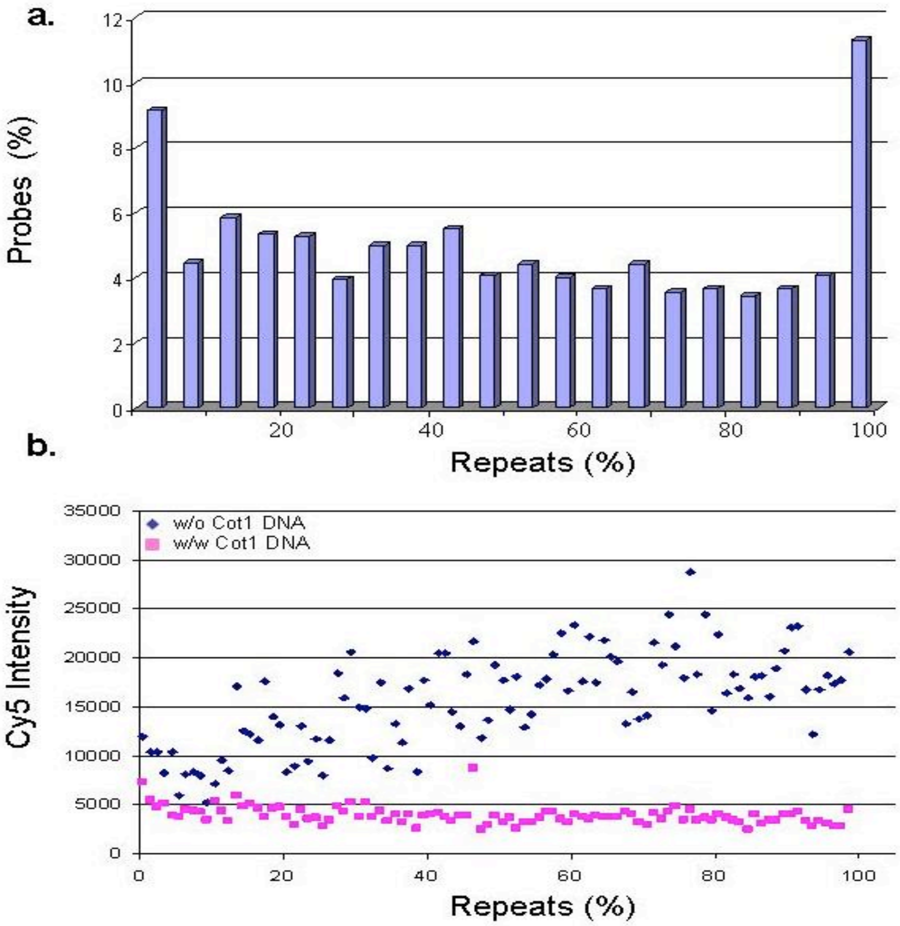

Figure 3.4. **Distribution and suppression of repeat sequences.** (a) Distribution and frequency of repeat sequences within probes on the array. (b). Suppression of repeat-specific signal using Cot1 DNA. Two independent hybridizations were carried out using genomic DNA extracted from CD8$^+$ lymphocytes. In both experiments total DNA was labelled with Cy5 dye.

Only in one of these was unlabelled Cot1 DNA added. In the hybridization without Cot1 DNA, Cy5 intensity increases almost linearly with repeat density until it reaches a plateau (around 25,000 Cy5 intensity). In the presence of Cot1 DNA, Cy5 intensity of highly repetitive probes is comparable to those of repeat-free probes.

3.3 MeDIP optimization and validation

Immunoprecipitation-based protocols for methylation analysis, methylated DNA immunoprecipitation (MeDIP) and methyl-cytosine immunoprecipitation (mCIP), were developed independently by two groups (Keshet et al., 2006; Weber et al., 2005). Both protocols are very similar and both were used as a point of reference while optimising the immunoprecipitation protocol for this study. In order to avoid confusion I refer to this technique as MeDIP in this thesis. MeDIP was first described in August 2005 and since then it has been used to generate comprehensive methylation profiles in mammals and plants as well as for DMR detection in cancer cells (Keshet et al., 2006; Rakyan, 2008; Weber et al., 2005; Zhang et al., 2006; Zilberman et al., 2007). In brief the MeDIP assay involves immunoprecipitation of methylated DNA fragments using an antibody that binds specifically to methylated cytosines. MeDIP was described in detail in section 1.3.6.2. In the below section I refer to the critical genomic DNA fragmentation step (3.3.1) and validation of MeDIP by quantitative real time qPCR (3.3.2).

3.3.1 Genomic DNA fragmentation

The first step of the MeDIP assay is genomic DNA fragmentation. Fragmentation of genomic DNA was performed by sonication. Sonication gives random, overlapping target fragments and clear interpretation of data (figure 3.5).

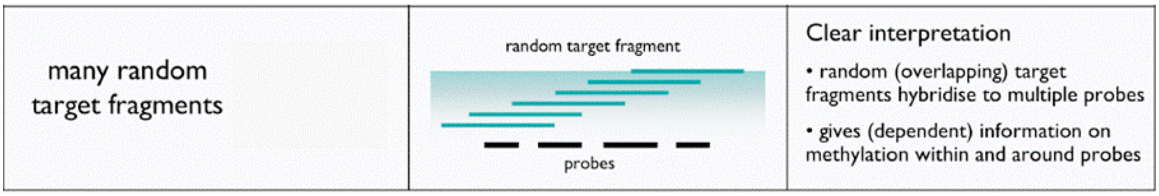
many random target fragments
random target fragment
probes
Clear interpretation
• random (overlapping) target fragments hybridise to multiple probes
• gives (dependent) information on methylation within and around probes

Figure 3.5. **Relationship between target fragments and array probes in methylation analysis.** Figure was adopted from Thorne (Thorne NP, 2006).

After sonication the size of the fragmented DNA ranges from 300 to 1000 bp (figure 3.6). It was important to keep the size range constant for all MeDIP experiments performed for this study.

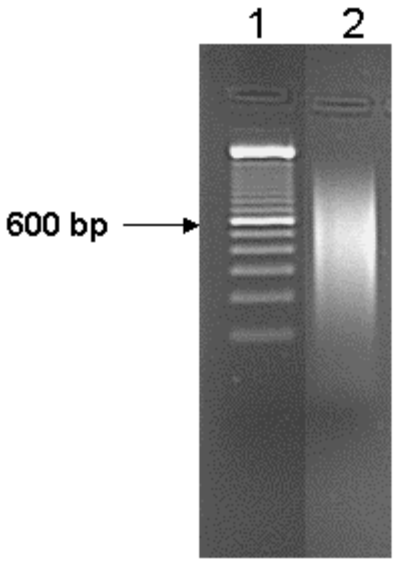

Figure 3.6. **Fragmentation of genomic DNA.** Genomic DNA was fragmented by sonication using a Virtis sonicator to generate fragments of a size range from 300 to 100bp. 500 ng of sonicated DNA was loaded onto a 2% agarose gel (lane 2). Size of fragmented DNA was estimated using 100bp ladder as a size marker (lane 1). Average size of DNA fragments is about 600bp, as indicated.

3.3.2 Validation of MeDIP

Enrichment of methylated DNA in the MeDIP fraction can be measured by qRT-PCR. I validated MeDIP by performing qRT-PCR to test the enrichment of regions with varying CpG densities for which the methylation status was known from the Human Epigenome Project (Eckhardt et al., 2006; Rakyan et al., 2004). It should be noted that CpG sites are unequally distributed within mammalian genomes and that the number of CpGs within a target region as well as their methylation status can influence target sequence enrichment in the MeDIP fraction. I showed that following

MeDIP methylated regions are enriched approximately proportionally to their CpG densities, and no significant enrichment irrespective of CpG density was observed for unmethylated regions (figure 3.7a).

Using a threshold of ≥5-fold enrichment, the MeDIP assay is therefore sensitive for regions of ≥1% CpG density. Using this threshold (actual enrichment range was 5-80 fold), I generated DNA methylation profiles of the entire MHC as described below.

In some cases (tDMRs screen, chapter 4) I had to introduce an amplification step while performing MeDIP assay due to limited starting DNA material. A ligation mediated PCR (LM-PCR) step was introduced as described previously (Oberley et al., 2004). In brief, the LM-PCR technique involves blunt ending of the fragmented DNA, ligation of a uni-directional double stranded oligo-nucleotide linker, and finally PCR amplification of the resultant DNA population after MeDIP. qRT-PCR analysis on MeDIP-LM-PCR DNA (post_LM-PCR) revealed similar enrichment for both methylated and unmethylated fractions as for non-amplified MeDIP fractions (pre_LM-PCR) (figure 3.7a,b), indicating that LM-PCR does not introduce an amplification bias. This was further supported by comparison of pre- and post-LM-PCR MeDIP-array analysis on the MHC-tile path array (see below).

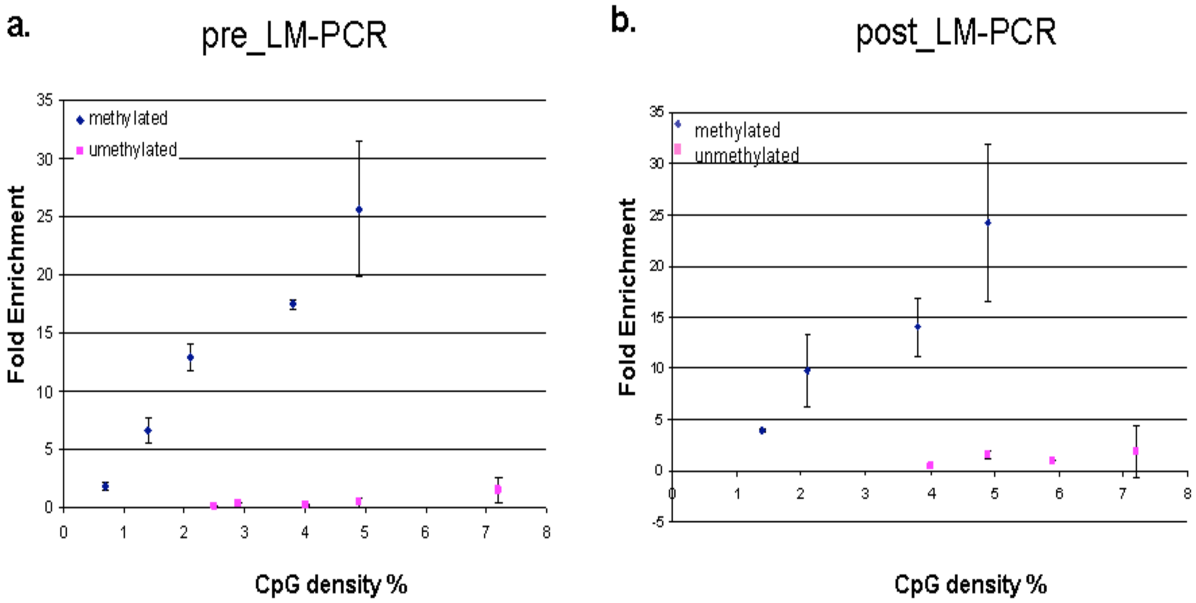

Figure 3.7. **Correlation between enrichment after MeDIP and CpG density**. Control sequences that are methylated, unmethylated or lack CpG sites were selected from HEP. MeDIP was done using liver genomic DNA. The relative enrichment of the MeDIP versus input fractions was calculated based on qRT-PCR data. Graph a. validation of MeDIP without amplification step. Graph b. Validation of MeDIP after amplification by LM-PCR. In both cases a specific and efficient enrichment of methylated over unmethylated fractions was shown. Enrichment of methylated and unmethylated fractions lies between the same range for both pre- and post-LM-PCR samples. The error bars indicate the variance of two independent measurements. Methylated amplicons display an approximately linear dependency on CpG density.

3.4 Application of the MHC tiling array for methylation analysis.

The MHC tiling array has been designed to be compatible with chromatin immunoprecipitation (ChIP), methylated DNA immunoprecipitation (MeDIP), array comparative genomic hybridization (aCGH) and expression profiling, inclusive of non-coding RNAs. In this section I demonstrate the utility of the MHC tiling path array for methylation analysis and DMR identification.

I used the array in conjunction with MeDIP. After performing MeDIP, MeDIP-enriched fractions and input fractions were differentially labelled with Cy3 and Cy5 and hybridized on the MHC tiling array. Following completion of appropriate quality control experiments DMRs were identified and validated.

In the next sections I describe how normalization of the array data was performed, how the quality of array hybridizations was tested and finally how I identified and validated DMRs.

3.4.1 Normalization of MHC tiling array data

Microarray screening can be affected by multiple sources of variation including array construction process, preparation of samples, hybridization process and the quantification of spot intensities (Repsilber and Ziegler, 2005). Normalization of array data attempts to remove such variation which might affect the outcome of the subsequent analysis.

Normalization usually applies to the $\log_2$ ratio of Cy3 and Cy5 intensities (corrected for the background) which will be: M=$\log_2$Cy3 – $\log_2$Cy5. The log-intensity of each spot is: A = ($\log_2$Cy3+$\log_2$Cy5)/2, a measure of the overall brightness of a spot. For the normalization of the MeDIP-MHC array data I applied local linear regression (loess) (Smyth and Speed, 2003) which fits a robust local regression to the relation between M and A. The normalized M values is the original one minus the loess fitted one, and thus should correct for spatial effects and for effects related to intensity. Figure 3.8 shows how data differ before and after normalization (MA plots).

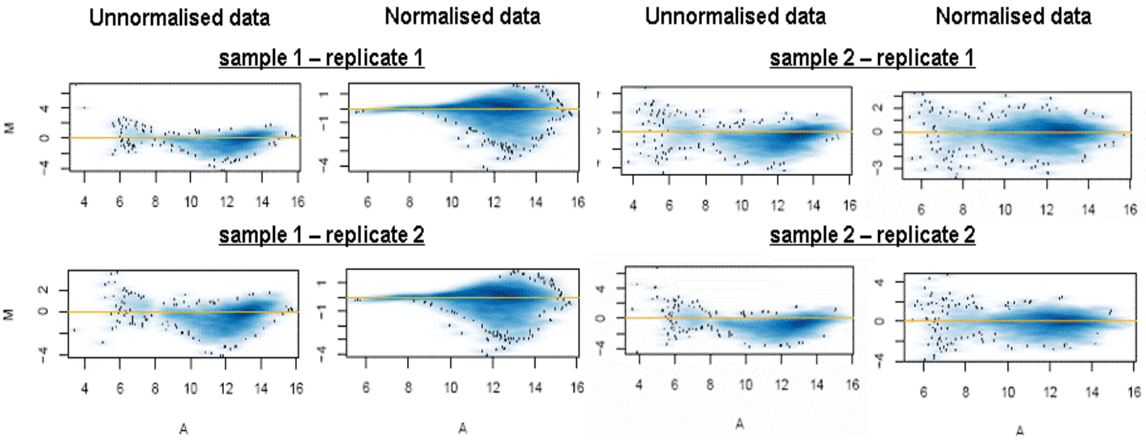

Figure 3.8. **Normalization within arrays.** Figure illustrates MA plots generated for four hybridization experiments. MeDIP-enriched and input fractions from a single sample were co-hybridized to each array. MA plots for two different samples and two biological replicates of each sample are shown. M represents the $\log_2$ ratio of the Cy3 and Cy5 intensities and A represents the $\log_2$ geometric mean of the Cy3 and Cy5 channel intensities.

3.4.2 MeDIP-MHC tiling array hybridization quality control

Methylation analysis across the MHC was performed by using the MHC tiling array in combination with MeDIP. I tested the reproducibility of this approach. I used DNA extracted from sperm (two biological replicates) and I performed MeDIP and the MHC array hybridization for each of the DNA samples before and after LM-PCR (see above). In both cases the correlation coefficient between biological replicates was high (R^2 >0.82) (figure 3.9a and b). In addition, I tested if LM-PCR introduces a bias in the methylation analysis. I compared hybridizations of pre– and post-LM-PCR samples (sperm DNA) and showed that LM-PCR does not have a major effect on the analysis (R^2 = 0.88) (figure 3.9c). Finally fluorochrome-reversed pairs of 2-colour labelled probes (dye swaps) were performed for LM-PCR samples. It has been shown that using standard direct labelling techniques introduce a bias to incorporation of the dye during the labelling reaction. In order to test for this I compared dye-swap hybridizations (figure 3.9d) and I found that the correlation coefficient was at least 0.72 (R^2 = 0.72).

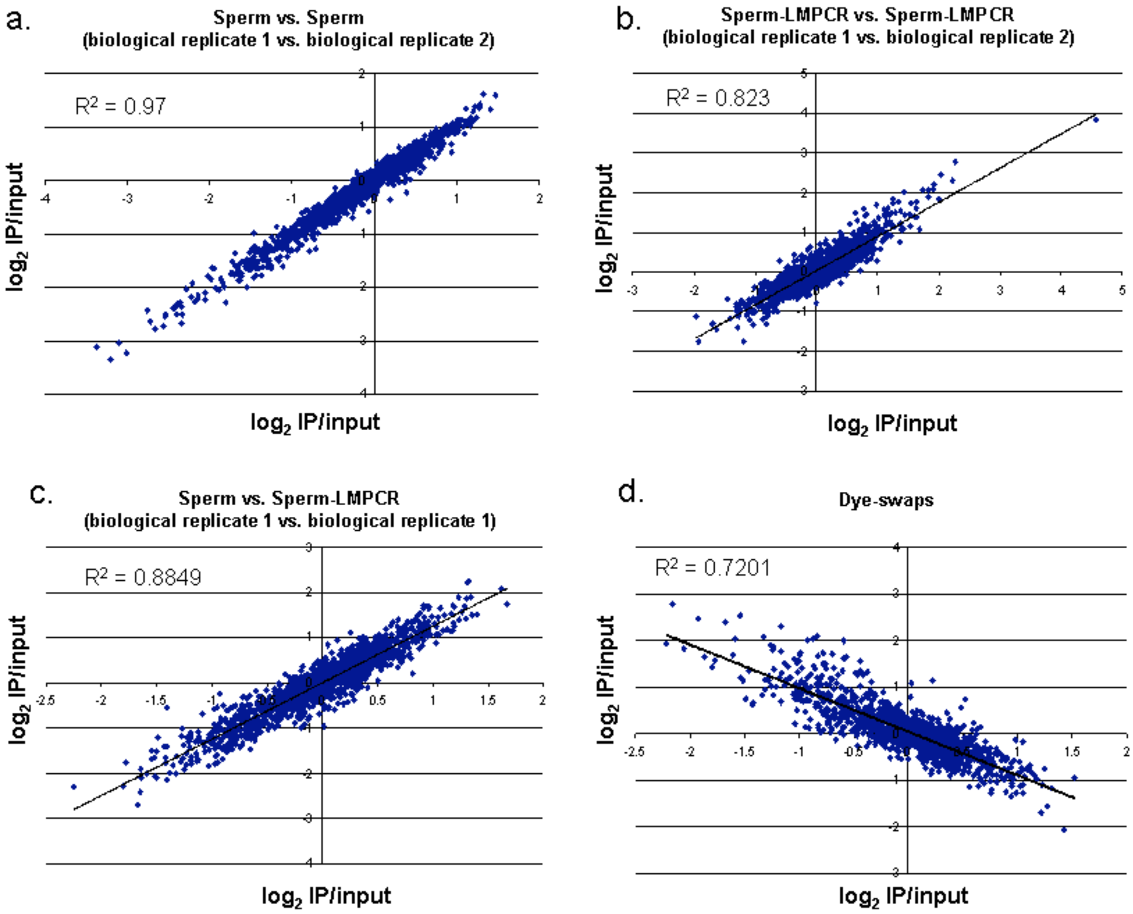

Figure 3.9. **Comparisons of MHC tiling array hybridizations.** Scatter-plots show: a). Comparison of biological replicates; b). Comparison of biological replicates after LM-PCR; c). Comparison of profiles pre- and post-LMPCR; d). Comparison of Dye swaps after LM-PCR. In all comparisons sperm DNA was used. Correlation coefficients are given for each comparison.

3.4.3 DMR identification and validation

The efficiency of immunoprecipitation in MeDIP depends on the density of methylated CpG sites, which vary greatly within any given mammalian genome, making it difficult to distinguish variations in enrichment from confounding CpG density effects (Weber et al., 2005; Weber et al., 2007). Hence, until recently, it has not been possible to estimate absolute methylation levels from MeDIP-array experiments. The on-going development of a novel algorithm employing a Bayesian de-convolution strategy to normalize MeDIP array data for CpG density is likely to overcome this current limitation in the near future (Down et al., 2008).

Determining absolute methylation values along the MHC region was beyond the scope of this study. I aimed to identify DMRs between samples. To this end I followed an experimental design as shown in figure 3.10. DMRs were identified by performing direct pair-wise comparisons and by applying t-test statistics. A threshold of p-value < 0.001 was employed.

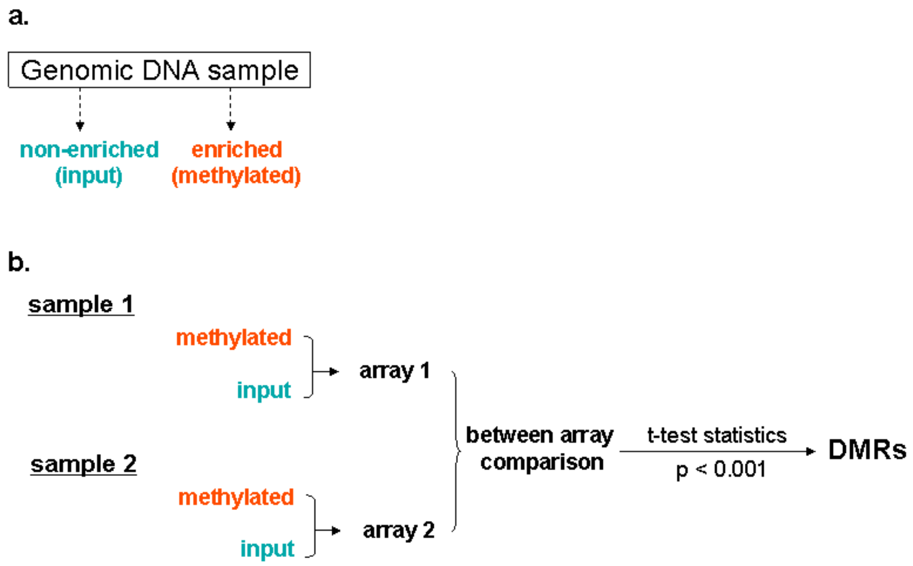

Figure 3.10. **Design of approach for calling DMRs.** (a). Target fractions after MeDIP experiment. (b). Calling of DMRs. For DMR identification, comparisons between arrays were performed. Significance of methylation differences between samples was calculated by t-test statistics. A threshold of p-value < 0.001 was used.

This approach was applied successfully for the identification of tDMRs and pDMRs (chapters 4 and 5). The approach was validated further by correlating the tDMRs identified by this study with tDMRs identified by an additional independent study, the Human Epigenome Project (HEP) (Eckhardt et al., 2006; Rakyan et al., 2004). DMRs were also validated by bisulphite sequencing. In all cases, DMR status could be confirmed, indicating that the array is suitable for DMR identification (chapter 4 and 5).

3.5 Discussion

The array reported in this chapter is the first genomic tiling array (2kb resolution) of the entire MHC. Commercially available tiling arrays usually exclude repeat sequences and therefore cover only about 50% of the genomic sequence. At the time of array design, whole-genome tiling arrays that included the MHC were constructed from P1 artificial chromosomes (PACs) and bacterial artificial chromosomes (BACs), resulting in a resolution of approximately 100 Kb. By utilizing a public clone resource, the MHC array was generated at a fraction of the costs associated with commercial arrays, albeit at lower resolution than is now achievable with these platforms. The array is compatible with standard array processing and scanning platforms and contains 7832 features. Of these 6988 correspond to the MHC region on chromosome 6. According to quality control experiments I performed, 97% of the probes can be informative for analysis of the MHC region. Upon request, the MHC array is freely available from the Microarray Facility at the Wellcome Trust Sanger Institute.

The array has been designed to be compatible with chromatin immunoprecipitation (ChIP), methylated DNA immunoprecipitation (MeDIP), array comparative genomic hybridization (aCGH) and expression profiling, inclusive of non-coding RNAs.

In this chapter I described how the array can be used for methylation analysis. To this end, MeDIP was optimised and validated showing its efficiency for immunoprecipitation of methylated genomic fragments (300-1000bp) with at least 1% CpGs.

The utility of the MHC array in conjunction with MeDIP for DMR identification was tested and validated. This approach allows DMR identification at 2kb resolution and used for the identification of tissue and phenotype specific DMRs as described in chapters 4 and 5 respectively.

3.6 Conclusion

I have generated and validated a genomic tiling array and I have demonstrated its utility for DNA methylation profiling and the identification of DMRs when combined with MeDIP. Chapters 4 and 5 describe the application of this approach for tDMR and pDMR identification respectively.